



Guida per l'utente serverless di Amazon EMR

# Amazon EMR



# Amazon EMR: Guida per l'utente serverless di Amazon EMR

Copyright © 2025 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

I marchi e l'immagine commerciale di Amazon non possono essere utilizzati in relazione a prodotti o servizi che non siano di Amazon, in una qualsiasi modalità che possa causare confusione tra i clienti o in una qualsiasi modalità che denigri o discrediti Amazon. Tutti gli altri marchi non di proprietà di Amazon sono di proprietà delle rispettive aziende, che possono o meno essere associate, collegate o sponsorizzate da Amazon.

---

# Table of Contents

Cos'è Amazon EMR Serverless? .....	1
Concetti .....	1
Versione di rilascio .....	1
Applicazione .....	2
Esecuzione del processo .....	3
Worker .....	3
Capacità preinizializzata .....	3
EMR Studio .....	4
Prerequisiti per iniziare .....	5
Iscriviti per un Account AWS .....	5
Crea un utente con accesso amministrativo .....	6
Concessione delle autorizzazioni .....	7
Concessione dell'accesso programmatico .....	9
Configura il AWS CLI .....	10
Aprire la console .....	11
Nozioni di base .....	12
Autorizzazioni .....	12
Storage .....	12
Carichi di lavoro interattivi .....	12
Creare un ruolo Job Runtime .....	13
Guida introduttiva alla console .....	18
Fase 1: creazione di un'applicazione .....	18
Fase 2: Inviare un job run o un carico di lavoro interattivo .....	19
Passaggio 3: Visualizza l'interfaccia utente e i registri dell'applicazione .....	23
Fase 4: pulizia .....	23
Iniziare da AWS CLI .....	23
Fase 1: creazione di un'applicazione .....	23
Fase 2: Inviare un job run .....	24
Fase 3: Esamina l'output .....	27
Fase 4: pulizia .....	28
Interazione e configurazione di un'applicazione EMR Serverless .....	30
Stati dell'applicazione .....	30
Utilizzo della console EMR Studio .....	31
Creazione di un'applicazione .....	32

Elenca le applicazioni dalla console EMR Studio .....	33
Gestione delle applicazioni dalla console EMR Studio .....	33
Utilizzando il AWS CLI .....	34
Configurazione di un'applicazione .....	35
Il comportamento delle applicazioni .....	35
Capacità preinizializzata per lavorare con un'applicazione in EMR Serverless .....	37
Configurazione predefinita dell'app .....	40
Personalizzazione di un'immagine .....	46
Prerequisiti .....	36
Fase 1: Creare un'immagine personalizzata da immagini di base EMR Serverless .....	48
Passaggio 2: convalida l'immagine localmente .....	49
Fase 3: carica l'immagine nel tuo repository Amazon ECR .....	50
Passaggio 4: creare o aggiornare un'applicazione con immagini personalizzate .....	50
Fase 5: consentire a EMR Serverless di accedere all'archivio di immagini personalizzato .....	52
Considerazioni e limitazioni .....	52
Configurazione dell'accesso VPC per le applicazioni EMR Serverless per la connessione ai dati .....	53
Crea applicazione .....	53
Configura l'applicazione .....	57
Procedure consigliate per la pianificazione delle sottoreti .....	57
Opzioni di architettura .....	59
Utilizzo dell'architettura x86_64 .....	59
Utilizzo dell'architettura arm64 (Graviton) .....	59
Avvia nuove app con Graviton .....	60
Converti le app esistenti in Graviton .....	60
Considerazioni .....	61
Concorrenza e attesa dei lavori .....	62
Principali vantaggi della concorrenza e dell'accodamento .....	62
Guida introduttiva alla concorrenza e alla messa in coda .....	62
Considerazioni sulla concorrenza e sull'accodamento .....	63
Caricamento dei dati .....	65
Prerequisiti .....	65
Nozioni di base su S3 Express One Zone .....	66
Esecuzione di processi .....	68
Stati delle esecuzioni di processi .....	68
Annullamento del job run con periodo di grazia .....	70

Periodo di grazia per i lavori in batch .....	70
Periodo di grazia per i lavori di streaming .....	72
Considerazioni .....	52
Utilizzo della console EMR Studio .....	75
Invio di un processo .....	75
Visualizza esecuzioni dei processi .....	77
Usando il AWS CLI .....	78
Utilizzo di dischi ottimizzati per la riproduzione casuale .....	79
Vantaggi principali .....	80
Nozioni di base .....	80
Lavori di streaming per l'elaborazione di dati in streaming continuo .....	84
Considerazioni e limitazioni .....	86
Nozioni di base .....	86
Connettori di streaming .....	87
Gestione dei log .....	89
Utilizzo delle configurazioni Spark quando si eseguono job EMR Serverless .....	90
Parametri Spark .....	90
Proprietà Spark .....	94
Esempi Spark .....	99
Utilizzo delle configurazioni Hive quando si eseguono job EMR Serverless .....	100
Parametri Hive .....	100
Proprietà Hive .....	102
Esempi di Hive .....	116
Resilienza del processo .....	117
Monitoraggio di un processo con una policy di ripetizione .....	121
Registrazione con politica di riprova .....	121
Configurazione Metastore per EMR Serverless .....	121
Utilizzo del AWS Glue Data Catalog come metastore .....	122
Utilizzo di un metastore Hive esterno .....	126
Utilizzo della gerarchia AWS multicatalogo di Glue su EMR Serverless .....	131
Considerazioni sull'utilizzo di un metastore esterno .....	133
Accesso a S3 su più account .....	133
Prerequisiti .....	133
Utilizza una policy sui bucket S3 .....	134
Usa un ruolo presunto .....	135
Esempi di ruoli presunti .....	137

Risoluzione degli errori .....	142
Errore: Job non riuscito poiché l'account ha raggiunto il limite di servizio sulla vCPU massima che può utilizzare contemporaneamente. ....	142
Errore: Job non riuscito perché l'applicazione ha superato le impostazioni di MaximumCapacity. ....	142
Errore: Job non riuscito a causa dell'impossibilità di allocare Worker poiché l'applicazione ha superato la capacità massima. ....	143
Errore: l'accesso a S3 è negato. Controlla le autorizzazioni di accesso S3 del ruolo Job Runtime sulle risorse S3 richieste. ....	143
Errore ModuleNotFoundError: nessun modulo denominato<module>. Consulta la guida per l'utente su come utilizzare le librerie python con EMR Serverless. ....	143
Errore: impossibile assumere il ruolo di esecuzione <role name>perché non esiste o non è configurato con la relazione di trust richiesta. ....	143
Esecuzione di carichi di lavoro interattivi .....	144
Panoramica .....	144
Prerequisiti .....	144
Autorizzazioni .....	145
Configurazione .....	146
Considerazioni .....	146
Esecuzione di carichi di lavoro interattivi tramite l'endpoint Apache Livy .....	147
Prerequisiti .....	148
Autorizzazioni richieste .....	148
Nozioni di base .....	149
Considerazioni .....	156
Registrazione di log e monitoraggio .....	158
Archiviazione dei registri .....	158
Storage gestito .....	159
Amazon S3 .....	160
Amazon CloudWatch .....	161
Registri rotanti .....	164
Crittografia dei log .....	166
Storage gestito .....	166
Bucket Amazon S3 .....	166
Amazon CloudWatch .....	167
Autorizzazioni richieste .....	167
Configurazione di Log4j2 .....	171

Log4j2 e Spark .....	171
Monitoraggio .....	175
Applicazioni e lavori .....	175
Metriche del motore Spark .....	185
Parametri di utilizzo .....	189
Automazione con EventBridge .....	190
Esempi di eventi EMR Serverless EventBridge .....	191
Applicazione di tag alle risorse .....	196
Che cos'è un tag? .....	196
Applicazione di tag alle risorse .....	196
Limitazioni relative all'etichettatura .....	197
Lavorare con i tag .....	198
Tutorial .....	200
Utilizzo di Java 17 .....	200
JAVA_HOME .....	200
spark-defaults .....	201
Usare Hudi .....	202
Utilizzo di Iceberg .....	203
Usare le librerie Python .....	204
Utilizzo delle funzionalità native di Python .....	204
Creazione di un ambiente virtuale Python .....	204
Configurazione dei PySpark lavori per l'utilizzo delle librerie Python .....	206
Usare diverse versioni di Python .....	207
Utilizzo di Delta Lake OSS .....	208
Amazon EMR versioni 6.9.0 e successive .....	208
Amazon EMR versioni 6.8.0 e precedenti .....	210
Invio di lavori da Airflow .....	211
Utilizzo delle funzioni definite dall'utente di Hive .....	213
Utilizzo di immagini personalizzate .....	215
Usa una versione Python personalizzata .....	215
Usa una versione Java personalizzata .....	216
Crea un'immagine di data science .....	216
Elaborazione di dati geospaziali con Apache Sedona .....	217
Informazioni sulle licenze per l'utilizzo di immagini personalizzate .....	217
Utilizzo di Spark su Amazon Redshift .....	218
Avvio di un'applicazione Spark .....	218

Autenticazione su Amazon Redshift .....	220
Lettura e scrittura su Amazon Redshift .....	222
Considerazioni .....	224
Connessione a DynamoDB .....	225
Fase 1: Caricamento su Amazon S3 .....	225
Fase 2: Creare una tabella Hive .....	226
Fase 3: Copia su DynamoDB .....	227
Fase 4: Interrogazione da DynamoDB .....	229
Configurazione dell'accesso multi-account .....	231
Considerazioni .....	233
Sicurezza .....	235
Best practice di sicurezza .....	236
Applicazione del principio del privilegio minimo .....	236
Isolamento di un codice dell'applicazione non attendibile .....	236
Autorizzazioni per il controllo degli accessi basato su ruoli (RBAC) .....	236
Protezione dei dati .....	236
Crittografia a riposo .....	237
Crittografia in transito .....	240
Identity and Access Management (IAM) .....	240
Destinatari .....	241
Autenticazione con identità .....	242
Gestione dell'accesso con policy .....	245
Come funziona EMR Serverless con IAM .....	248
Uso di ruoli collegati ai servizi .....	255
Ruoli Job Runtime per Amazon EMR Serverless .....	260
Politiche di accesso degli utenti .....	263
Policy per il controllo degli accessi basato su tag .....	267
Policy basate sull'identità .....	270
Aggiornamenti alle policy .....	273
Risoluzione dei problemi .....	274
Utilizzo di EMR Serverless con Lake Formation .....	276
Accesso non filtrato a Lake Formation per EMR Serverless .....	276
Lake Formation per FGAC .....	283
Crittografia tra lavoratori .....	301
Abilitazione della crittografia Mutual-TLS su EMR Serverless .....	301
Secrets Manager per la protezione dei dati .....	302

Come funzionano i segreti .....	302
Creazione di un segreto .....	303
Specificare i riferimenti segreti .....	303
Concedi l'accesso al segreto .....	306
Ruota il segreto .....	307
S3 Access Grants per il controllo dell'accesso ai dati .....	308
Panoramica .....	308
Avvia un'applicazione .....	308
Considerazioni .....	310
CloudTrail per la registrazione .....	310
Informazioni EMR Serverless in CloudTrail .....	310
Informazioni sulle voci dei file di registro EMR Serverless .....	311
Convalida della conformità .....	313
Resilienza .....	314
Sicurezza dell'infrastruttura .....	314
Analisi della configurazione e delle vulnerabilità .....	315
Punti finali e quote .....	316
Endpoint di servizio .....	316
Quote del servizio .....	320
Limiti di API .....	321
Altre considerazioni .....	52
Versioni di rilascio .....	325
EMR Serverless7.9.0 .....	326
EMR Serverless7,80 .....	326
EMR Serverless7,7,0 .....	326
EMR Serverless7,6,0 .....	327
EMR Serverless7,5,0 .....	327
EMR Serverless7,40 .....	327
EMR Serverless7,3,0 .....	328
EMR Serverless7.2.0 .....	328
EMR Serverless7.1.0 .....	329
EMR Serverless7,0,0 .....	329
EMR Serverless6,15,0 .....	329
EMR Serverless6.14.0 .....	330
EMR Serverless6,13,0 .....	330
EMR Serverless6,12,0 .....	331

---

EMR Serverless6.11,0 .....	331
EMR Serverless6.10.0 .....	331
EMR Serverless6.9.0 .....	332
EMR Serverless6.8.0 .....	333
EMR Serverless6.7.0 .....	333
Modifiche specifiche del motore .....	334
EMR Serverless6.6.0 .....	334
Cronologia dei documenti .....	336
.....	cccxxxviii

# Cos'è Amazon EMR Serverless?

Amazon EMR Serverless è un'opzione di implementazione per Amazon EMR che fornisce un ambiente di runtime serverless. Ciò semplifica il funzionamento delle applicazioni di analisi che utilizzano i più recenti framework open source, come Apache Spark e Apache Hive. Con EMR Serverless, non è necessario configurare, ottimizzare, proteggere o gestire i cluster per eseguire applicazioni con questi framework.

EMR Serverless consente di evitare un approvvigionamento eccessivo o insufficiente delle risorse per i processi di elaborazione dei dati. EMR Serverless determina automaticamente le risorse di cui l'applicazione ha bisogno, ottiene queste risorse per elaborare i lavori e rilascia le risorse al termine dei processi. Nei casi d'uso in cui le applicazioni richiedono una risposta in pochi secondi, come l'analisi interattiva dei dati, è possibile preinizializzare le risorse di cui l'applicazione ha bisogno al momento della creazione dell'applicazione.

Con EMR Serverless, continuerai a ottenere i vantaggi di Amazon EMR, come la compatibilità open source, la concorrenza e le prestazioni di runtime ottimizzate per i framework più diffusi.

EMR Serverless è adatto ai clienti che desiderano semplificare il funzionamento delle applicazioni che utilizzano framework open source. Offre un avvio rapido dei processi, una gestione automatica della capacità e un controllo diretto dei costi.

## Concetti

In questa sezione, trattiamo i termini e i concetti di EMR Serverless che compaiono nella nostra Guida per l'utente EMR Serverless.

### Versione di rilascio

Una versione di Amazon EMR è un insieme di applicazioni open source dell'ecosistema dei big data. Ogni versione include diverse applicazioni, componenti e funzionalità per i big data che l'utente seleziona per l'implementazione e la configurazione di EMR Serverless in modo che possano eseguire le applicazioni. Quando si crea un'applicazione, è necessario specificarne la versione di rilascio. Scegli la versione di rilascio di Amazon EMR e la versione del framework open source che desideri utilizzare nella tua applicazione. Per ulteriori informazioni sulle versioni preliminari, consulta.

[Versioni di rilascio di Amazon EMR Serverless](#)

## Applicazione

Con EMR Serverless, è possibile creare una o più applicazioni EMR Serverless che utilizzano framework di analisi open source. Per creare un'applicazione, è necessario specificare i seguenti attributi:

- La versione di rilascio di Amazon EMR per la versione del framework open source che desideri utilizzare. Per determinare la tua versione di rilascio, consulta [Versioni di rilascio di Amazon EMR Serverless](#).
- Il runtime specifico che desideri venga utilizzato dall'applicazione, ad esempio Apache Spark o Apache Hive.

Dopo aver creato un'applicazione, puoi inviare lavori di elaborazione dati o richieste interattive all'applicazione.

Ogni applicazione EMR Serverless viene eseguita su un Amazon Virtual Private Cloud (VPC) sicuro, completamente diverso dalle altre applicazioni. Inoltre, puoi utilizzare le policy AWS Identity and Access Management (IAM) per definire quali utenti e ruoli possono accedere all'applicazione. È inoltre possibile specificare limiti per controllare e tenere traccia dei costi di utilizzo sostenuti dall'applicazione.

Prendi in considerazione la possibilità di creare più applicazioni quando devi fare quanto segue:

- Utilizza diversi framework open source
- Utilizza versioni diverse di framework open source per diversi casi d'uso
- Esegui test A/B durante l'aggiornamento da una versione all'altra
- Mantieni ambienti logici separati per scenari di test e produzione
- Fornisci ambienti logici separati per diversi team con controlli dei costi e monitoraggio dell'utilizzo indipendenti
- Separa line-of-business le diverse applicazioni

EMR Serverless è un servizio regionale che semplifica il modo in cui i carichi di lavoro vengono eseguiti su più zone di disponibilità in una regione. Per ulteriori informazioni su come utilizzare le applicazioni con EMR Serverless, vedere [Interazione e configurazione di un'applicazione EMR Serverless](#)

## Esecuzione del processo

L'esecuzione di un processo è una richiesta inviata a un'applicazione EMR Serverless che l'applicazione esegue in modo asincrono e ne tiene traccia fino al completamento. Esempi di lavori includono una query HiveQL che invii a un'applicazione Apache Hive o uno script di elaborazione dati che invii a PySpark un'applicazione Apache Spark. Quando invii un lavoro, devi specificare un ruolo di runtime, creato in IAM, che il job utilizza per accedere alle AWS risorse, come gli oggetti Amazon S3. È possibile inviare più richieste di esecuzione di un processo a un'applicazione e ogni esecuzione di lavoro può utilizzare un ruolo di runtime diverso per accedere AWS alle risorse. Un'applicazione EMR Serverless inizia a eseguire i lavori non appena li riceve ed esegue più richieste di lavoro contemporaneamente. Per ulteriori informazioni su come EMR Serverless esegue i job, vedere.

[Esecuzione di processi](#)

## Worker

Un'applicazione EMR Serverless utilizza internamente i lavoratori per eseguire i carichi di lavoro. Le dimensioni predefinite di questi worker si basano sul tipo di applicazione e sulla versione di rilascio di Amazon EMR. Quando pianifichi l'esecuzione di un lavoro, puoi sostituire queste dimensioni.

Quando si invia un lavoro, EMR Serverless calcola le risorse necessarie all'applicazione per il lavoro e pianifica i lavoratori. EMR Serverless suddivide i carichi di lavoro in attività, scarica immagini, provvede e organizza i lavoratori e li disattiva al termine del lavoro. EMR Serverless aumenta o riduce automaticamente i lavoratori in base al carico di lavoro e al parallelismo richiesti in ogni fase del lavoro. Questa scalabilità automatica elimina la necessità di stimare il numero di lavoratori necessari all'applicazione per eseguire i carichi di lavoro.

## Capacità preinizializzata

EMR Serverless offre una funzionalità di capacità preinizializzata che mantiene gli operatori inizializzati e pronti a rispondere in pochi secondi. Questa capacità crea in modo efficace un pool di lavoratori accogliente per un'applicazione. Per configurare questa funzionalità per ogni applicazione, impostate il `initial-capacity` parametro di un'applicazione. Quando si configura la capacità preinizializzata, i lavori possono iniziare immediatamente in modo da poter implementare applicazioni iterative e lavori urgenti. Per ulteriori informazioni sui lavoratori preinizializzati, consulta.

[Configurazione di un'applicazione quando si lavora con EMR Serverless](#)

## EMR Studio

EMR Studio è la console utente che puoi utilizzare per gestire le tue applicazioni EMR Serverless. Se nel tuo account non esiste un EMR Studio quando crei la tua prima applicazione EMR Serverless, ne creiamo automaticamente uno per te. Puoi accedere a EMR Studio dalla console Amazon EMR oppure puoi attivare l'accesso federato dal tuo provider di identità (IdP) tramite IAM o IAM Identity Center. In questo modo, gli utenti possono accedere a Studio e gestire le applicazioni EMR Serverless senza accesso diretto alla console Amazon EMR. Per ulteriori informazioni su come le applicazioni EMR Serverless funzionano con EMR Studio, vedere e. [Creazione di un'applicazione EMR Serverless dalla console EMR Studio](#) [Esecuzione di processi dalla console EMR Studio](#)

# Prerequisiti per iniziare a usare EMR Serverless

Questa sezione descrive i prerequisiti amministrativi per l'esecuzione di EMR Serverless. Questi includono la configurazione dell'account e la gestione delle autorizzazioni.

## Argomenti

- [Iscriviti per un Account AWS](#)
- [Crea un utente con accesso amministrativo](#)
- [Concessione delle autorizzazioni](#)
- [Installa e configura il AWS CLI](#)
- [Aprire la console](#)

## Iscriviti per un Account AWS

Se non ne hai uno Account AWS, completa i seguenti passaggi per crearne uno.

Per iscriverti a un Account AWS

1. Apri la <https://portal.aws.amazon.com/billing/registrazione>.
2. Segui le istruzioni online.

Nel corso della procedura di registrazione riceverai una telefonata, durante la quale sarà necessario inserire un codice di verifica attraverso la tastiera del telefono.

Quando ti iscrivi a un Account AWS, Utente root dell'account AWS viene creato un. L'utente root dispone dell'accesso a tutte le risorse e tutti i Servizi AWS nell'account. Come best practice di sicurezza, assegna l'accesso amministrativo a un utente e utilizza solo l'utente root per eseguire [attività che richiedono l'accesso di un utente root](#).

AWS ti invia un'email di conferma dopo il completamento della procedura di registrazione. In qualsiasi momento, puoi visualizzare l'attività corrente del tuo account e gestirlo accedendo a <https://aws.amazon.com/> e scegliendo Il mio account.

## Crea un utente con accesso amministrativo

Dopo esserti registrato Account AWS, proteggi Utente root dell'account AWS AWS IAM Identity Center, abilita e crea un utente amministrativo in modo da non utilizzare l'utente root per le attività quotidiane.

Proteggi i tuoi Utente root dell'account AWS

1. Accedi [AWS Management Console](#) come proprietario dell'account scegliendo Utente root e inserendo il tuo indirizzo Account AWS email. Nella pagina successiva, inserisci la password.

Per informazioni sull'accesso utilizzando un utente root, consulta la pagina [Signing in as the root user](#) della Guida per l'utente di Accedi ad AWS .

2. Abilita l'autenticazione a più fattori (MFA) per l'utente root.

Per istruzioni, consulta [Abilitare un dispositivo MFA virtuale per l'utente Account AWS root \(console\)](#) nella Guida per l'utente IAM.

Crea un utente con accesso amministrativo

1. Abilita Centro identità IAM.

Per istruzioni, consulta [Abilitazione di AWS IAM Identity Center](#) nella Guida per l'utente di AWS IAM Identity Center .

2. In IAM Identity Center, assegna l'accesso amministrativo a un utente.

Per un tutorial sull'utilizzo di IAM Identity Center directory come fonte di identità, consulta [Configurare l'accesso utente con l'impostazione predefinita IAM Identity Center directory](#) nella Guida per l'AWS IAM Identity Center utente.

Accesso come utente amministratore

- Per accedere con l'utente IAM Identity Center, utilizza l'URL di accesso che è stato inviato al tuo indirizzo e-mail quando hai creato l'utente IAM Identity Center.

Per informazioni sull'accesso utilizzando un utente IAM Identity Center, consulta [AWS Accedere al portale di accesso](#) nella Guida per l'Accedi ad AWS utente.

## Assegna l'accesso a ulteriori utenti

1. In IAM Identity Center, crea un set di autorizzazioni conforme alla best practice dell'applicazione di autorizzazioni con il privilegio minimo.

Segui le istruzioni riportate nella pagina [Creazione di un set di autorizzazioni](#) nella Guida per l'utente di AWS IAM Identity Center .

2. Assegna al gruppo prima gli utenti e poi l'accesso con autenticazione unica (Single Sign-On).

Per istruzioni, consulta [Aggiungere gruppi](#) nella Guida per l'utente di AWS IAM Identity Center .

## Concessione delle autorizzazioni

Negli ambienti di produzione, ti consigliamo di utilizzare politiche più dettagliate. Per esempi di tali politiche, vedere. [Esempi di policy di accesso degli utenti per EMR Serverless](#) Per ulteriori informazioni sulla gestione degli accessi, consulta la sezione [Gestione degli accessi per AWS le risorse](#) nella IAM User Guide.

Per gli utenti che devono iniziare a usare EMR Serverless in un ambiente sandbox, utilizza una policy simile alla seguente:

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "EMRStudioCreate",
      "Effect": "Allow",
      "Action": [
        "elasticmapreduce:CreateStudioPresignedUrl",
        "elasticmapreduce:DescribeStudio",
        "elasticmapreduce:CreateStudio",
        "elasticmapreduce:ListStudios"
      ],
      "Resource": "*"
    },
    {
      "Sid": "EMRServerlessFullAccess",
      "Effect": "Allow",
      "Action": [
        "emr-serverless:*"
      ],
    }
  ]
}
```

```

        "Resource": "*"
    },
    {
        "Sid": "AllowEC2ENICreationWithEMRTags",
        "Effect": "Allow",
        "Action": [
            "ec2:CreateNetworkInterface"
        ],
        "Resource": [
            "arn:aws:ec2:*:*:network-interface/*"
        ],
        "Condition": {
            "StringEquals": {
                "aws:CalledViaLast": "ops.emr-serverless.amazonaws.com"
            }
        }
    },
    {
        "Sid": "AllowEMRServerlessServiceLinkedRoleCreation",
        "Effect": "Allow",
        "Action": "iam:CreateServiceLinkedRole",
        "Resource": "arn:aws:iam:*:*:role/aws-service-role/*"
    }
]
}

```

Per fornire l'accesso, aggiungi autorizzazioni agli utenti, gruppi o ruoli:

- Utenti e gruppi in: AWS IAM Identity Center

Crea un set di autorizzazioni. Segui le istruzioni riportate nella pagina [Create a permission set](#) (Creazione di un set di autorizzazioni) nella Guida per l'utente di AWS IAM Identity Center .

- Utenti gestiti in IAM tramite un provider di identità:

Crea un ruolo per la federazione delle identità. Segui le istruzioni riportate nella pagina [Create a role for a third-party identity provider \(federation\)](#) della Guida per l'utente IAM.

- Utenti IAM:

- Crea un ruolo che l'utente possa assumere. Segui le istruzioni riportate nella pagina [Create a role for an IAM user](#) della Guida per l'utente IAM.

- (Non consigliato) Collega una policy direttamente a un utente o aggiungi un utente a un gruppo di utenti. Segui le istruzioni riportate nella pagina [Aggiunta di autorizzazioni a un utente \(console\)](#) nella Guida per l'utente IAM.

## Concessione dell'accesso programmatico

Gli utenti necessitano di un accesso programmatico se desiderano interagire con utenti AWS esterni a. AWS Management Console Il modo per concedere l'accesso programmatico dipende dal tipo di utente che accede. AWS

Per fornire agli utenti l'accesso programmatico, scegli una delle seguenti opzioni.

Quale utente necessita dell'accesso programmatico?	Per	Come
Identità della forza lavoro (Utenti gestiti nel centro identità IAM)	Utilizza credenziali temporane e per firmare le richieste programmatiche a AWS CLI,, AWS SDKs o. AWS APIs	<p>Segui le istruzioni per l'interfaccia che desideri utilizzare.</p> <ul style="list-style-type: none"> <li>• Per la AWS CLI, vedere <a href="#">Configurazione dell'uso AWS IAM Identity Center nella AWS CLI Guida per l'utente</a>.AWS Command Line Interface</li> <li>• Per AWS SDKs gli strumenti e AWS APIs, consulta <a href="#">l'autenticazione di IAM Identity Center</a> nella Guida di riferimento AWS SDKs and Tools.</li> </ul>
IAM	Utilizza credenziali temporane e per firmare le richieste programmatiche a AWS CLI, AWS SDKs, o. AWS APIs	Seguendo le istruzioni riportate in <a href="#">Utilizzo delle credenziali temporanee con le AWS risorse nella Guida per l'utente IAM</a> .

Quale utente necessita dell'accesso programmatico?	Per	Come
IAM	(Non consigliato) Utilizza credenziali a lungo termine per firmare richieste programmatiche a AWS CLI,, AWS SDKs o. AWS APIs	<p>Segui le istruzioni per l'interfaccia che desideri utilizzare.</p> <ul style="list-style-type: none"> <li>• Per la AWS CLI, consulta <a href="#">Autenticazione tramite credenziali utente IAM nella Guida per l'utente.AWS Command Line Interface</a></li> <li>• Per gli strumenti AWS SDKs e gli strumenti, consulta <a href="#">Autenticazione tramite credenziali a lungo termine</a> nella Guida di riferimento agli strumenti e agli AWS SDKs strumenti.</li> <li>• Per AWS APIs, consulta la sezione <a href="#">Gestione delle chiavi di accesso per gli utenti IAM</a> nella Guida per l'utente IAM.</li> </ul>

## Installa e configura il AWS CLI

Se si desidera utilizzare EMR Serverless APIs, è necessario installare l'ultima versione di (). AWS Command Line Interface AWS CLI Non è necessario AWS CLI utilizzare EMR Serverless dalla console EMR Studio e puoi iniziare senza la CLI seguendo i passaggi indicati. [Guida introduttiva a EMR Serverless dalla console](#)

Per configurare il AWS CLI

1. Per installare la versione più recente di AWS CLI per macOS, Linux o Windows, vedi [Installazione o aggiornamento della versione più recente di. AWS CLI](#)

2. Per configurare AWS CLI e configurare in modo sicuro l'accesso a Servizi AWS, incluso EMR Serverless, vedere Configurazione [rapida](#) con. `aws configure`
3. Per verificare la configurazione, immettere il seguente DataBrew comando al prompt dei comandi.

```
aws emr-serverless help
```

AWS CLI i comandi utilizzano l'impostazione predefinita Regione AWS della configurazione, a meno che non venga impostata con un parametro o un profilo. Per impostare Regione AWS un parametro, è possibile aggiungere il `--region` parametro a ciascun comando.

Per impostare Regione AWS un profilo, aggiungi prima un profilo denominato nel `~/.aws/config` file o nel `%UserProfile%/.aws/config` file (per Microsoft Windows). Segui la procedura descritta in [Profili denominati per AWS CLI](#). Successivamente, imposta le tue Regione AWS e le altre impostazioni con un comando simile a quello illustrato nell'esempio seguente.

```
[profile emr-serverless]
aws_access_key_id = ACCESS-KEY-ID-OF-IAM-USER
aws_secret_access_key = SECRET-ACCESS-KEY-ID-OF-IAM-USER
region = us-east-1
output = text
```

## Aprire la console

La maggior parte degli argomenti relativi alla console in questa sezione parte dalla console Amazon [EMR](#). Se non hai già effettuato l'accesso al tuo Account AWS, accedi, quindi apri la [console Amazon EMR](#) e passa alla sezione successiva per iniziare a usare Amazon EMR.

# Guida introduttiva ad Amazon EMR Serverless

Questo tutorial ti aiuta a iniziare a usare EMR Serverless quando distribuisce un carico di lavoro Spark o Hive di esempio. Potrai creare, eseguire ed eseguire il debug della tua applicazione. Mostriamo le opzioni predefinite nella maggior parte di questo tutorial.

Prima di avviare un'applicazione EMR Serverless, completare le seguenti attività.

## Argomenti

- [Concedere le autorizzazioni per l'utilizzo di EMR Serverless](#)
- [Preparazione dello storage per EMR Serverless](#)
- [Crea un EMR Studio per eseguire carichi di lavoro interattivi](#)
- [Creare un ruolo Job Runtime](#)
- [Guida introduttiva a EMR Serverless dalla console](#)
- [Iniziare da AWS CLI](#)

## Concedere le autorizzazioni per l'utilizzo di EMR Serverless

Per utilizzare EMR Serverless, è necessario un utente o un ruolo IAM con una policy allegata che conceda le autorizzazioni per EMR Serverless. Per creare un utente e allegare la policy appropriata a quell'utente, segui le istruzioni riportate in [Concessione delle autorizzazioni](#)

## Preparazione dello storage per EMR Serverless

In questo tutorial, utilizzerai un bucket S3 per archiviare i file di output e i log del carico di lavoro Spark o Hive di esempio che eseguirai utilizzando un'applicazione EMR Serverless. Per creare un bucket, segui le istruzioni in [Creazione di un bucket](#) nella Guida per l'utente della console di Amazon Simple Storage Service. Sostituisci ogni ulteriore riferimento a *amzn-s3-demo-bucket* con il nome del bucket appena creato.

## Crea un EMR Studio per eseguire carichi di lavoro interattivi

Se desideri utilizzare EMR Serverless per eseguire query interattive tramite notebook ospitati in EMR Studio, devi specificare un bucket S3 e il ruolo di servizio [minimo per EMR Serverless per creare un workspace](#). Per la procedura di configurazione, consulta [Configurare un EMR Studio](#) nella Amazon

EMR Management Guide. Per ulteriori informazioni sui carichi di lavoro interattivi, consulta. [Esegui carichi di lavoro interattivi con EMR Serverless tramite EMR Studio](#)

## Creare un ruolo Job Runtime

I job eseguiti in EMR Serverless utilizzano un ruolo di runtime che fornisce autorizzazioni granulari per risorse specifiche Servizi AWS e in fase di esecuzione. In questo tutorial, un bucket S3 pubblico ospita i dati e gli script. Il bucket *amzn-s3-demo-bucket* memorizza l'output.

Per impostare un ruolo di job runtime, è necessario innanzitutto creare un ruolo di runtime con una policy di fiducia in modo che EMR Serverless possa utilizzare il nuovo ruolo. Successivamente, collega la politica di accesso S3 richiesta a quel ruolo. I seguenti passaggi ti guidano attraverso il processo.

### Console

1. Accedi alla console IAM all'indirizzo <https://console.aws.amazon.com/iam/>.
2. Nel pannello di navigazione a sinistra seleziona Ruoli.
3. Scegliere Crea ruolo.
4. Per il tipo di ruolo, scegli Custom trust policy e incolla la seguente policy di fiducia. Ciò consente ai lavori inviati alle tue applicazioni Amazon EMR Serverless di accedere ad altri per tuo Servizi AWS conto.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Principal": {
        "Service": "emr-serverless.amazonaws.com"
      },
      "Action": "sts:AssumeRole"
    }
  ]
}
```

5. Scegli Avanti per accedere alla pagina Aggiungi autorizzazioni, quindi scegli Crea policy.
6. La pagina Crea policy si apre in una nuova scheda. Incolla il codice JSON della policy qui sotto.

**⚠ Important**

Sostituisci *amzn-s3-demo-bucket* la policy seguente con il nome effettivo del bucket creato in [Preparazione dello storage per EMR Serverless](#). Questa è una politica di base per l'accesso a S3. Per altri esempi di ruoli di job runtime, consulta [Ruoli Job Runtime per Amazon EMR Serverless](#).

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "ReadAccessForEMRSamples",
      "Effect": "Allow",
      "Action": [
        "s3:GetObject",
        "s3:ListBucket"
      ],
      "Resource": [
        "arn:aws:s3::*.elasticmapreduce",
        "arn:aws:s3::*.elasticmapreduce/*"
      ]
    },
    {
      "Sid": "FullAccessToOutputBucket",
      "Effect": "Allow",
      "Action": [
        "s3:PutObject",
        "s3:GetObject",
        "s3:ListBucket",
        "s3:DeleteObject"
      ],
      "Resource": [
        "arn:aws:s3::amzn-s3-demo-bucket",
        "arn:aws:s3::amzn-s3-demo-bucket/*"
      ]
    },
    {
      "Sid": "GlueCreateAndReadDataCatalog",
      "Effect": "Allow",
      "Action": [
```

```

        "glue:GetDatabase",
        "glue:CreateDatabase",
        "glue:GetDataBases",
        "glue:CreateTable",
        "glue:GetTable",
        "glue:UpdateTable",
        "glue>DeleteTable",
        "glue:GetTables",
        "glue:GetPartition",
        "glue:GetPartitions",
        "glue:CreatePartition",
        "glue:BatchCreatePartition",
        "glue:GetUserDefinedFunctions"
    ],
    "Resource": ["*"]
}
]
}

```

7. Nella pagina Rivedi la politica, inserisci un nome per la tua politica, ad esempio `EMRServerlessS3AndGlueAccessPolicy`.
8. Aggiorna la pagina Allega criteri di autorizzazione e scegli `EMRServerlessS3AndGlueAccessPolicy`.
9. Nella pagina Nome, rivedi e crea, in Nome ruolo, inserisci un nome per il tuo ruolo, ad esempio `EMRServerlessS3RuntimeRole`. Per creare questo ruolo IAM, scegli `Crea ruolo`.

## CLI

1. Crea un file denominato `emr-serverless-trust-policy.json` contenente la policy di attendibilità da utilizzare per il ruolo IAM. Il file deve contenere la seguente politica.

```

{
  "Version": "2012-10-17",
  "Statement": [{
    "Sid": "EMRServerlessTrustPolicy",
    "Action": "sts:AssumeRole",
    "Effect": "Allow",
    "Principal": {
      "Service": "emr-serverless.amazonaws.com"
    }
  }]
}

```

```
}

```

2. Creare un ruolo IAM denominato `EMRServerlessS3RuntimeRole`. Utilizza la politica di fiducia che hai creato nel passaggio precedente.

```
aws iam create-role \
  --role-name EMRServerlessS3RuntimeRole \
  --assume-role-policy-document file://emr-serverless-trust-policy.json

```

Annotate l'ARN nell'output. L'ARN del nuovo ruolo viene utilizzato durante l'invio del lavoro, denominato successivamente. *job-role-arn*

3. Crea un file denominato `emr-sample-access-policy.json` che definisca la policy IAM per il tuo carico di lavoro. Ciò fornisce l'accesso in lettura allo script e ai dati archiviati nei bucket S3 pubblici e l'accesso in lettura/scrittura a. *amzn-s3-demo-bucket*

#### Important

*amzn-s3-demo-bucket* Sostituiscilo nella policy seguente con il nome effettivo del bucket creato in.. [Preparazione dello storage per EMR Serverless](#) Questa è una politica di base per l'accesso a AWS Glue e S3. Per altri esempi di ruoli in fase di job runtime, consulta [Ruoli Job Runtime per Amazon EMR Serverless](#).

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "ReadAccessForEMRSamples",
      "Effect": "Allow",
      "Action": [
        "s3:GetObject",
        "s3:ListBucket"
      ],
      "Resource": [
        "arn:aws:s3::*.elasticmapreduce",
        "arn:aws:s3::*.elasticmapreduce/*"
      ]
    },
    {
      "Sid": "FullAccessToOutputBucket",

```

```

    "Effect": "Allow",
    "Action": [
        "s3:PutObject",
        "s3:GetObject",
        "s3:ListBucket",
        "s3:DeleteObject"
    ],
    "Resource": [
        "arn:aws:s3:::amzn-s3-demo-bucket",
        "arn:aws:s3:::amzn-s3-demo-bucket/*"
    ]
},
{
    "Sid": "GlueCreateAndReadDataCatalog",
    "Effect": "Allow",
    "Action": [
        "glue:GetDatabase",
        "glue:CreateDatabase",
        "glue:GetDataBases",
        "glue:CreateTable",
        "glue:GetTable", Understanding default application behavior,
including auto-start and auto-stop, as well as maximum capacity and worker
configurations for configuring an application with &EMRServerless;.
        "glue:UpdateTable",
        "glue>DeleteTable",
        "glue:GetTables",
        "glue:GetPartition",
        "glue:GetPartitions",
        "glue:CreatePartition",
        "glue:BatchCreatePartition",
        "glue:GetUserDefinedFunctions"
    ],
    "Resource": ["*"]
}
]
}

```

4. Crea una policy IAM denominata `EMRServerlessS3AndGlueAccessPolicy` con il file di policy che hai creato nel passaggio 3. Prendi nota dell'ARN nell'output, poiché utilizzerai l'ARN della nuova politica nella fase successiva.

```

aws iam create-policy \
  --policy-name EMRServerlessS3AndGlueAccessPolicy \

```

```
--policy-document file://emr-sample-access-policy.json
```

Annota l'ARN della nuova politica nell'output. Lo sostituirai *policy-arn* nel passaggio successivo.

5. Allega la policy IAM EMRServerlessS3AndGlueAccessPolicy al ruolo EMRServerlessS3RuntimeRole di job runtime.

```
aws iam attach-role-policy \  
  --role-name EMRServerlessS3RuntimeRole \  
  --policy-arn policy-arn
```

## Guida introduttiva a EMR Serverless dalla console

Questa sezione descrive l'utilizzo di EMR Serverless, inclusa la creazione di un EMR Studio. Descrive inoltre come inviare le esecuzioni dei job e visualizzare i log.

Passaggi da completare

- [Fase 1: Creare un'applicazione EMR Serverless](#)
- [Fase 2: Inviare un job run o un carico di lavoro interattivo](#)
- [Passaggio 3: Visualizza l'interfaccia utente e i registri dell'applicazione](#)
- [Fase 4: pulizia](#)

### Fase 1: Creare un'applicazione EMR Serverless

Crea una nuova applicazione con EMR Serverless come segue.

1. [Accedi AWS Management Console e apri la console Amazon EMR su https://console.aws.amazon.com/emr.](https://console.aws.amazon.com/emr)
2. Nel riquadro di navigazione a sinistra, scegli EMR Serverless per accedere alla pagina di destinazione EMR Serverless.
3. Per creare o gestire applicazioni EMR Serverless, è necessaria l'interfaccia utente di EMR Studio.
  - Se hai già un EMR Studio nel Regione AWS luogo in cui desideri creare un'applicazione, seleziona Gestisci applicazioni per accedere a EMR Studio o seleziona lo studio che desideri utilizzare.

- Se non disponi di uno studio EMR nel Regione AWS luogo in cui desideri creare un'applicazione, scegli Inizia, quindi scegli Crea e avvia Studio. EMR Serverless crea per te un EMR Studio in modo che tu possa creare e gestire applicazioni.
4. Nell'interfaccia utente di Create studio che si apre in una nuova scheda, inserisci il nome, il tipo e la versione di rilascio dell'applicazione. Se desideri eseguire solo processi batch, seleziona Usa le impostazioni predefinite solo per i processi batch. Per i carichi di lavoro interattivi, seleziona Usa le impostazioni predefinite per i carichi di lavoro interattivi. Con questa opzione puoi anche eseguire lavori in batch su applicazioni abilitate all'interattività. Se necessario, è possibile modificare queste impostazioni in un secondo momento.

Per ulteriori informazioni, consulta [Creare uno studio](#).

5. Seleziona Crea applicazione per creare la tua prima applicazione.

Passa alla sezione successiva [Fase 2: Inviare un job run o un carico di lavoro interattivo](#) per inviare un job run o un carico di lavoro interattivo.

## Fase 2: Inviare un job run o un carico di lavoro interattivo

### Spark job run

In questo tutorial, utilizziamo uno PySpark script per calcolare il numero di occorrenze di parole uniche in più file di testo. Un bucket S3 pubblico di sola lettura memorizza sia lo script che il set di dati.

Per eseguire un job Spark

1. Carica lo script di esempio `wordcount.py` nel tuo nuovo bucket con il seguente comando.

```
aws s3 cp s3://us-east-1.elasticmapreduce/emr-containers/samples/wordcount/scripts/wordcount.py s3://amzn-s3-demo-bucket/scripts/
```

2. Il completamento [Fase 1: Creare un'applicazione EMR Serverless](#) porta alla pagina dei dettagli dell'applicazione in EMR Studio. Qui, scegli l'opzione Invia lavoro.
3. Nella pagina Invia offerta di lavoro, completa quanto segue.
  - Nel campo Nome, inserisci il nome con cui vuoi chiamare il job run.
  - Nel campo Runtime role, inserisci il nome del ruolo in cui hai creato [Creare un ruolo Job Runtime](#).

- Nel campo Posizione dello script, inserisci `s3://amzn-s3-demo-bucket/scripts/wordcount.py` come URI S3.
- Nel campo Argomenti dello script, inserisci `["s3://amzn-s3-demo-bucket/emr-serverless-spark/output"]`.
- Nella sezione delle proprietà di Spark, scegli Modifica come testo e inserisci le seguenti configurazioni.

```
--conf spark.executor.cores=1 --conf spark.executor.memory=4g --  
conf spark.driver.cores=1 --conf spark.driver.memory=4g --conf  
spark.executor.instances=1
```

4. Per avviare l'esecuzione del lavoro, scegli Invia lavoro.
5. Nella scheda Job run, dovresti vedere il tuo nuovo job eseguito con lo stato Running.

## Hive job run

In questa parte del tutorial, creiamo una tabella, inseriamo alcuni record ed eseguiamo una query di aggregazione del conteggio. Per eseguire il job Hive, devi prima creare un file che contenga tutte le query Hive da eseguire come parte di un singolo job, carica il file su S3 e specifica questo percorso S3 all'avvio del job Hive.

### Per eseguire un job Hive

1. Crea un file chiamato `hive-query.q1` che contenga tutte le query che desideri eseguire nel tuo job Hive.

```
create database if not exists emrserverless;  
use emrserverless;  
create table if not exists test_table(id int);  
drop table if exists Values__Tmp__Table__1;  
insert into test_table values (1),(2),(2),(3),(3),(3);  
select id, count(id) from test_table group by id order by id desc;
```

2. Carica `hive-query.q1` nel tuo bucket S3 con il seguente comando.

```
aws s3 cp hive-query.q1 s3://amzn-s3-demo-bucket/emr-serverless-hive/query/hive-  
query.q1
```

3. Il completamento [Fase 1: Creare un'applicazione EMR Serverless](#) porta alla pagina dei dettagli dell'applicazione in EMR Studio. Qui, scegli l'opzione Invia lavoro.
4. Nella pagina Invia offerta di lavoro, completa quanto segue.
  - Nel campo Nome, inserisci il nome con cui vuoi chiamare il job run.
  - Nel campo Runtime role, inserisci il nome del ruolo in cui hai creato [Creare un ruolo Job Runtime](#).
  - Nel campo Posizione dello script, inserisci `s3://amzn-s3-demo-bucket/emr-serverless-hive/query/hive-query.q1` come URI S3.
  - Nella sezione delle proprietà di Hive, scegli Modifica come testo e inserisci le seguenti configurazioni.

```
--hiveconf hive.log.explain.output=false
```

- Nella sezione Configurazione del lavoro, scegli Modifica come JSON e inserisci il seguente codice JSON.

```
{
  "applicationConfiguration":
  [{
    "classification": "hive-site",
    "properties": {
      "hive.exec.scratchdir": "s3://amzn-s3-demo-bucket/emr-
serverless-hive/hive/scratch",
      "hive.metastore.warehouse.dir": "s3://amzn-s3-demo-bucket/emr-
serverless-hive/hive/warehouse",
      "hive.driver.cores": "2",
      "hive.driver.memory": "4g",
      "hive.tez.container.size": "4096",
      "hive.tez.cpu.vcores": "1"
    }
  ]
}
```

5. Per avviare l'esecuzione del processo, scegli Invia lavoro.
6. Nella scheda Job run, dovresti vedere il tuo nuovo job eseguito con lo stato Running.

## Interactive workload

Con Amazon EMR 6.14.0 e versioni successive, puoi utilizzare notebook ospitati in EMR Studio per eseguire carichi di lavoro interattivi per Spark in EMR Serverless. Per ulteriori informazioni, tra cui autorizzazioni e prerequisiti, consulta [Esegui carichi di lavoro interattivi con EMR Serverless tramite EMR Studio](#)

Dopo aver creato l'applicazione e impostato le autorizzazioni richieste, segui i seguenti passaggi per eseguire un notebook interattivo con EMR Studio:

1. Vai alla scheda Aree di lavoro in EMR Studio. Se devi ancora configurare una posizione di archiviazione Amazon S3 e il [ruolo del servizio EMR Studio](#), seleziona il pulsante Configura studio nel banner nella parte superiore dello schermo.
2. Per accedere a un notebook, seleziona un Workspace o crea un nuovo Workspace. Usa Avvio veloce per aprire il tuo spazio di lavoro in una nuova scheda.
3. Vai alla scheda appena aperta. Seleziona l'icona Compute dalla barra di navigazione a sinistra. Seleziona EMR Serverless come tipo di elaborazione.
4. Seleziona l'applicazione interattiva che hai creato nella sezione precedente.
5. Nel campo Runtime role, inserisci il nome del ruolo IAM che l'applicazione EMR Serverless può assumere per l'esecuzione del job. Per ulteriori informazioni sui ruoli di runtime, consulta [Job runtime roles](#) nella Amazon EMR Serverless User Guide.
6. Seleziona Allega. Questa operazione potrebbe richiedere fino a un minuto. La pagina verrà aggiornata una volta allegata.
7. Scegliete un kernel e avviate un notebook. Puoi anche sfogliare taccuini di esempio su EMR Serverless e copiarli nel tuo spazio di lavoro. Per accedere ai taccuini di esempio, accedi al `{...}` menu nella barra di navigazione a sinistra e sfoglia i taccuini che contengono il nome del file del taccuino. `serverless`
8. Nel notebook, puoi accedere al collegamento al registro dei driver e a un collegamento all'interfaccia utente di Apache Spark, un'interfaccia in tempo reale che fornisce metriche per monitorare il tuo lavoro. Per ulteriori informazioni, consulta [Monitoraggio delle applicazioni e dei job EMR Serverless](#) nella Amazon EMR Serverless User Guide.

Quando colleghi un'applicazione a un'area di lavoro di Studio, l'avvio dell'applicazione si attiva automaticamente se non è già in esecuzione. È inoltre possibile preavviare l'applicazione e tenerla pronta prima di collegarla all'area di lavoro.

## Passaggio 3: Visualizza l'interfaccia utente e i registri dell'applicazione

Per visualizzare l'interfaccia utente dell'applicazione, identificate innanzitutto il job eseguito.

Un'opzione per l'interfaccia utente Spark o l'interfaccia utente Hive Tez è disponibile nella prima riga di opzioni per l'esecuzione del processo, in base al tipo di lavoro. Seleziona l'opzione appropriata.

Se hai scelto l'interfaccia utente Spark, scegli la scheda Executors per visualizzare i log dei driver e degli executors. Se hai scelto l'interfaccia utente di Hive Tez, scegli la scheda Tutte le attività per visualizzare i log.

Una volta che lo stato di esecuzione del processo viene visualizzato come Riuscito, puoi visualizzare l'output del processo nel tuo bucket S3.

### Fase 4: pulizia

Sebbene l'applicazione che hai creato dovrebbe interrompersi automaticamente dopo 15 minuti di inattività, ti consigliamo comunque di rilasciare risorse che non intendi utilizzare più.

Per eliminare l'applicazione, vai alla pagina Elenco applicazioni. Seleziona l'applicazione che hai creato e scegli Azioni → Stop per interrompere l'applicazione. Dopo che l'applicazione è nello STOPPED stato, seleziona la stessa applicazione e scegli Azioni → Elimina.

Per altri esempi di esecuzione dei job Spark e Hive, consulta [Utilizzo delle configurazioni Spark quando si eseguono job EMR Serverless](#) e [Utilizzo delle configurazioni Hive quando si eseguono job EMR Serverless](#)

## Iniziare da AWS CLI

Inizia a usare EMR Serverless AWS CLI con i comandi per creare un'applicazione, eseguire lavori, controllare l'output dell'esecuzione del processo ed eliminare le tue risorse.

### Fase 1: Creare un'applicazione EMR Serverless

Usa il `emr-serverless create-application` comando per creare la tua prima applicazione EMR Serverless. Devi specificare il tipo di applicazione e l'etichetta di rilascio di Amazon EMR associata alla versione dell'applicazione che desideri utilizzare. Il nome dell'applicazione è facoltativo.

#### Spark

Per creare un'applicazione Spark, esegui il comando seguente.

```
aws emr-serverless create-application \  
  --release-label emr-6.6.0 \  
  --type "SPARK" \  
  --name my-application
```

## Hive

Per creare un'applicazione Hive, esegui il comando seguente.

```
aws emr-serverless create-application \  
  --release-label emr-6.6.0 \  
  --type "HIVE" \  
  --name my-application
```

Annotate l'ID dell'applicazione restituito nell'output. Utilizzerai l'ID per avviare la candidatura e durante l'invio del lavoro, in seguito denominato *application-id*

Prima di passare a [Fase 2: Invia un job run alla tua applicazione EMR Serverless](#), assicurati che la tua candidatura abbia raggiunto CREATED lo stato con l'[get-application](#) API.

```
aws emr-serverless get-application \  
  --application-id application-id
```

EMR Serverless crea lavoratori per soddisfare i lavori richiesti. Per impostazione predefinita, questi vengono creati su richiesta, ma è anche possibile specificare una capacità preinizializzata impostando il `initialCapacity` parametro al momento della creazione dell'applicazione. È inoltre possibile limitare la capacità massima totale che un'applicazione può utilizzare con il `maximumCapacity` parametro. Per ulteriori informazioni su queste opzioni, consulta [Configurazione di un'applicazione quando si lavora con EMR Serverless](#).

## Fase 2: Invia un job run alla tua applicazione EMR Serverless

Ora la tua applicazione EMR Serverless è pronta per eseguire i lavori.

### Spark

In questo passaggio, utilizziamo uno PySpark script per calcolare il numero di occorrenze di parole uniche in più file di testo. Un bucket S3 pubblico di sola lettura memorizza sia lo script che

il set di dati. L'applicazione invia il file di output e i dati di registro dal runtime Spark `/output` e alle `/logs` directory del bucket S3 che hai creato.

Per eseguire un job Spark

1. Usa il seguente comando per copiare lo script di esempio che eseguiremo nel tuo nuovo bucket.

```
aws s3 cp s3://us-east-1.elasticmapreduce/emr-containers/samples/wordcount/scripts/wordcount.py s3://amzn-s3-demo-bucket/scripts/
```

2. Nel comando seguente, sostituiscilo *application-id* con l'ID dell'applicazione. Sostituiscilo *job-role-arn* con il ruolo di runtime ARN in cui hai creato. [Creare un ruolo Job Runtime](#) *job-run-name* Sostituiscilo con il nome con cui vuoi chiamare il job run. Sostituisci tutte *amzn-s3-demo-bucket* le stringhe con il bucket Amazon S3 che hai creato e `/output` aggiungile al percorso. Questo crea una nuova cartella nel bucket in cui EMR Serverless può copiare i file di output dell'applicazione.

```
aws emr-serverless start-job-run \
  --application-id application-id \
  --execution-role-arn job-role-arn \
  --name job-run-name \
  --job-driver '{
    "sparkSubmit": {
      "entryPoint": "s3://amzn-s3-demo-bucket/scripts/wordcount.py",
      "entryPointArguments": ["s3://amzn-s3-demo-bucket/emr-serverless-
spark/output"],
      "sparkSubmitParameters": "--conf spark.executor.cores=1
--conf spark.executor.memory=4g --conf spark.driver.cores=1 --conf
spark.driver.memory=4g --conf spark.executor.instances=1"
    }
  }'
```

3. Annotate l'ID di esecuzione del lavoro restituito nell'output. Sostituiscilo *job-run-id* con questo ID nei passaggi seguenti.

## Hive

In questo tutorial, creiamo una tabella, inseriamo alcuni record ed eseguiamo una query di aggregazione del conteggio. Per eseguire il job Hive, devi prima creare un file che contenga tutte

le query Hive da eseguire come parte di un singolo job, carica il file su S3 e specifica questo percorso S3 all'avvio del job Hive.

Per eseguire un lavoro Hive

1. Crea un file chiamato `hive-query.q1` che contenga tutte le query che desideri eseguire nel tuo job Hive.

```
create database if not exists emrserverless;
use emrserverless;
create table if not exists test_table(id int);
drop table if exists Values__Tmp__Table__1;
insert into test_table values (1),(2),(2),(3),(3),(3);
select id, count(id) from test_table group by id order by id desc;
```

2. Carica `hive-query.q1` nel tuo bucket S3 con il seguente comando.

```
aws s3 cp hive-query.q1 s3://amzn-s3-demo-bucket/emr-serverless-hive/query/hive-
query.q1
```

3. Nel comando seguente, sostituiscilo *application-id* con il tuo ID dell'applicazione. Sostituiscilo *job-role-arn* con il ruolo di runtime ARN in cui hai creato. [Creare un ruolo Job Runtime](#) Sostituisci tutte *amzn-s3-demo-bucket* le stringhe con il bucket Amazon S3 che hai creato e `/output` aggiungi `/logs` e al percorso. Questo crea nuove cartelle nel bucket, dove EMR Serverless può copiare i file di output e di registro dell'applicazione.

```
aws emr-serverless start-job-run \
  --application-id application-id \
  --execution-role-arn job-role-arn \
  --job-driver '{
    "hive": {
      "query": "s3://amzn-s3-demo-bucket/emr-serverless-hive/query/hive-
query.q1",
      "parameters": "--hiveconf hive.log.explain.output=false"
    }
  }' \
  --configuration-overrides '{
    "applicationConfiguration": [{
      "classification": "hive-site",
      "properties": {
        "hive.exec.scratchdir": "s3://amzn-s3-demo-bucket/emr-serverless-
hive/hive/scratch",
```

```

        "hive.metastore.warehouse.dir": "s3://amzn-s3-demo-bucket/emr-
serverless-hive/hive/warehouse",
        "hive.driver.cores": "2",
        "hive.driver.memory": "4g",
        "hive.tez.container.size": "4096",
        "hive.tez.cpu.vcores": "1"
    }
}],
"monitoringConfiguration": {
    "s3MonitoringConfiguration": {
        "logUri": "s3://amzn-s3-demo-bucket/emr-serverless-hive/logs"
    }
}
}'

```

- Annotate l'ID di esecuzione del lavoro restituito nell'output. Sostituiscilo *job-run-id* con questo ID nei passaggi seguenti.

## Fase 3: Esaminate l'output dell'esecuzione del job

Il completamento dell'esecuzione del processo dovrebbe in genere richiedere 3-5 minuti.

### Spark

Puoi controllare lo stato del tuo job Spark con il seguente comando.

```

aws emr-serverless get-job-run \
  --application-id application-id \
  --job-run-id job-run-id

```

Con la destinazione dei log impostata su `s3://amzn-s3-demo-bucket/emr-serverless-spark/logs`, puoi trovare i log relativi a questo specifico job in cui viene eseguito. `s3://amzn-s3-demo-bucket/emr-serverless-spark/logs/applications/application-id/jobs/job-run-id`

Per le applicazioni Spark, EMR Serverless invia i registri degli eventi ogni 30 secondi nella cartella nella destinazione `sparklogs` dei registri S3. Al termine del processo, i log di runtime di Spark relativi al driver e agli esecutori vengono caricati in cartelle denominate in modo appropriato in base al tipo di lavoratore, ad esempio `driver` `executor`. L'output del lavoro viene caricato su. PySpark `s3://amzn-s3-demo-bucket/output/`

## Hive

Puoi controllare lo stato del tuo lavoro in Hive con il seguente comando.

```
aws emr-serverless get-job-run \  
  --application-id application-id \  
  --job-run-id job-run-id
```

Con la destinazione del registro impostata su `s3://amzn-s3-demo-bucket/emr-serverless-hive/logs`, puoi trovare i log relativi a questo specifico processo in cui viene eseguito. `s3://amzn-s3-demo-bucket/emr-serverless-hive/logs/applications/application-id/jobs/job-run-id`

Per le applicazioni Hive, EMR Serverless carica continuamente il driver Hive nella cartella e i registri HIVE\_DRIVER delle attività Tez nella cartella della destinazione TEZ\_TASK del registro S3. Dopo che l'esecuzione del processo raggiunge SUCCEEDED lo stato, l'output della tua query Hive diventa disponibile nella posizione Amazon S3 che hai specificato `monitoringConfiguration` nel campo di `configurationOverrides`

## Fase 4: pulizia

Quando hai finito di lavorare con questo tutorial, valuta la possibilità di eliminare le risorse che hai creato. Ti consigliamo di rilasciare risorse che non intendi riutilizzare.

### Elimina la tua applicazione

Per eliminare un'applicazione, utilizzare il seguente comando.

```
aws emr-serverless delete-application \  
  --application-id application-id
```

### Elimina il tuo bucket di log S3

Per eliminare il bucket di registrazione e output S3, usa il seguente comando. Sostituisci `amzn-s3-demo-bucket` con il nome effettivo del bucket S3 creato in.. [Preparazione dello storage per EMR Serverless](#)

```
aws s3 rm s3://amzn-s3-demo-bucket --recursive
```

```
aws s3api delete-bucket --bucket amzn-s3-demo-bucket
```

## Elimina il ruolo Job Runtime

Per eliminare il ruolo di runtime, scollega la policy dal ruolo. È quindi possibile eliminare sia il ruolo che la politica.

```
aws iam detach-role-policy \  
  --role-name EMRServerlessS3RuntimeRole \  
  --policy-arn policy-arn
```

Per eliminare il ruolo, utilizzare il comando seguente.

```
aws iam delete-role \  
  --role-name EMRServerlessS3RuntimeRole
```

Per eliminare la politica allegata al ruolo, utilizzare il comando seguente.

```
aws iam delete-policy \  
  --policy-arn policy-arn
```

Per altri esempi di esecuzione dei job Spark e Hive, consulta [Utilizzo delle configurazioni Spark quando si eseguono job EMR Serverless](#) e [Utilizzo delle configurazioni Hive quando si eseguono job EMR Serverless](#)

# Interazione e configurazione di un'applicazione EMR Serverless

Questa sezione spiega come interagire con la tua applicazione Amazon EMR Serverless con AWS CLI. Descrive inoltre la configurazione di un'applicazione, l'esecuzione di personalizzazioni e le impostazioni predefinite per i motori Spark e Hive.

## Argomenti

- [Stati dell'applicazione](#)
- [Creazione di un'applicazione EMR Serverless dalla console EMR Studio](#)
- [Interazione con l'applicazione EMR Serverless su AWS CLI](#)
- [Configurazione di un'applicazione quando si lavora con EMR Serverless](#)
- [Personalizzazione di un'immagine EMR Serverless](#)
- [Configurazione dell'accesso VPC per le applicazioni EMR Serverless per la connessione ai dati](#)
- [Opzioni di architettura Serverless Amazon EMR](#)
- [Concorrenza dei job e accodamento per un'applicazione EMR Serverless](#)

## Stati dell'applicazione

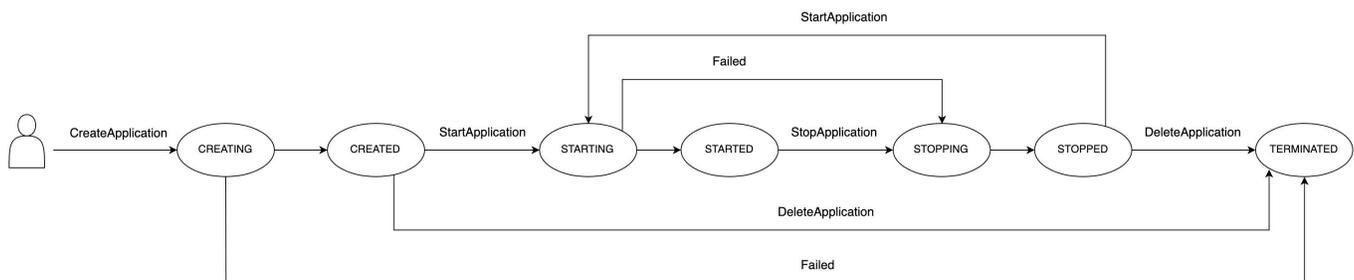
Quando si crea un'applicazione con EMR Serverless, l'esecuzione dell'applicazione entra nello stato CREATING. Dopodiché passa per i seguenti stati fino a quando non ha esito positivo (codice di uscita 0) o negativo (codice di uscita diverso da zero).

Le applicazioni possono avere i seguenti stati:

Stato	Descrizione
Creating (Creazione in corso)	L'applicazione è in fase di preparazione e non è ancora pronta per l'uso.
Creato	L'applicazione è stata creata ma non ha ancora fornito capacità. È possibile modificare l'applicazione per modificare la configurazione iniziale della capacità.

Stato	Descrizione
Avvio di	L'applicazione si sta avviando e sta fornendo capacità.
Avviato	L'applicazione è pronta per accettare nuovi lavori. L'applicazione accetta offerte di lavoro solo quando si trova in questo stato.
Stopping (In arresto)	Tutti i lavori sono stati completati e l'applicazione sta esaurendo la sua capacità.
Arrestate	L'applicazione è arrestata e nessuna risorsa è in esecuzione sull'applicazione. È possibile modificare l'applicazione per cambiarne la configurazione iniziale della capacità.
Terminata	L'applicazione è stata chiusa e non appare nell'elenco delle applicazioni.

Il diagramma seguente mostra la traiettoria degli stati delle applicazioni EMR Serverless.



## Creazione di un'applicazione EMR Serverless dalla console EMR Studio

Dalla console EMR Studio, è possibile creare, visualizzare e gestire applicazioni EMR Serverless. Per accedere alla console EMR Studio, segui le istruzioni in [Guida introduttiva dalla console](#).

## Creazione di un'applicazione

Con la pagina Crea applicazione, è possibile creare un'applicazione EMR Serverless seguendo questi passaggi.

1. Nel campo Nome, inserisci il nome con cui vuoi chiamare l'applicazione.
2. Nel campo Tipo, scegli Spark o Hive come tipo di applicazione.
3. Nel campo Versione release, scegli il numero di release EMR.
4. Nelle opzioni Architettura, scegli l'architettura del set di istruzioni da utilizzare. Per ulteriori informazioni, consulta [Opzioni di architettura Serverless Amazon EMR](#).
  - arm64 — Architettura ARM a 64 bit; per usare i processori Graviton
  - x86\_64 — architettura x86 a 64 bit; per utilizzare processori basati su x86
5. Esistono due opzioni di configurazione dell'applicazione per i campi rimanenti: impostazioni predefinite e impostazioni personalizzate. Questi campi sono opzionali.

**Impostazioni predefinite:** le impostazioni predefinite consentono di creare rapidamente un'applicazione con capacità preinizializzata. Ciò include un driver e un executor per Spark e un driver e un Tez Task per Hive. Le impostazioni predefinite non abilitano la connettività di rete al tuo VPC. L'applicazione è configurata per arrestarsi se inattiva per 15 minuti e si avvia automaticamente all'invio del lavoro.

**Impostazioni personalizzate:** le impostazioni personalizzate consentono di modificare le seguenti proprietà.

- Capacità preinizializzata: il numero di driver ed esecutori o di Hive Tez Task worker e le dimensioni di ciascun lavoratore.
- Limiti delle applicazioni: la capacità massima di un'applicazione.
- Comportamento dell'applicazione: il comportamento di avvio e arresto automatico dell'applicazione.
- Connessioni di rete: connettività di rete alle risorse VPC.
- Tag: tag personalizzati che è possibile assegnare all'applicazione.

Per ulteriori informazioni sulla capacità preinizializzata, sui limiti delle applicazioni e sul comportamento delle applicazioni, consulta. [Configurazione di un'applicazione quando si lavora](#)

con [EMR Serverless](#) Per ulteriori informazioni sulla connettività di rete, vedere. [Configurazione dell'accesso VPC per le applicazioni EMR Serverless per la connessione ai dati](#)

6. Per creare l'applicazione, scegli Crea applicazione.

## Elenca le applicazioni dalla console EMR Studio

È possibile visualizzare tutte le applicazioni EMR Serverless esistenti dalla pagina Elenco applicazioni. È possibile scegliere il nome di un'applicazione per accedere alla pagina dei dettagli dell'applicazione.

## Gestione delle applicazioni dalla console EMR Studio

È possibile eseguire le seguenti azioni su un'applicazione dalla pagina Elenco applicazioni o dalla pagina Dettagli di un'applicazione specifica.

### Avvia l'applicazione

Scegliete questa opzione per avviare manualmente un'applicazione.

### Interrompi l'applicazione

Scegliete questa opzione per interrompere manualmente un'applicazione. Un'applicazione non deve avere processi in esecuzione per poter essere interrotta. Per ulteriori informazioni sulle transizioni di stato delle applicazioni, consulta [Stati dell'applicazione](#).

### Configurare l'applicazione

Modifica le impostazioni opzionali per un'applicazione dalla pagina Configura applicazione. È possibile modificare la maggior parte delle impostazioni dell'applicazione. Ad esempio, puoi modificare l'etichetta di rilascio per un'applicazione per aggiornarla a una versione diversa di Amazon EMR oppure puoi cambiare l'architettura da x86\_64 a arm64. Le altre impostazioni opzionali sono le stesse presenti nella sezione Impostazioni personalizzate nella pagina Crea applicazione. Per ulteriori informazioni sulle impostazioni dell'applicazione, vedere [Creazione di un'applicazione](#).

### Eliminare l'applicazione

Scegliete questa opzione per eliminare manualmente un'applicazione. È necessario interrompere un'applicazione per eliminarla. Per ulteriori informazioni sulle transizioni di stato delle applicazioni, consulta [Stati dell'applicazione](#).

## Interazione con l'applicazione EMR Serverless su AWS CLI

Da AWS CLI, è possibile creare, descrivere ed eliminare singole applicazioni. Puoi anche elencare tutte le tue applicazioni in modo da poterle visualizzare a colpo d'occhio. Questa sezione descrive come eseguire queste azioni. Per ulteriori operazioni applicative, come l'avvio, l'arresto e l'aggiornamento dell'applicazione, consulta l'[EMR Serverless API Reference](#). Per esempi di come utilizzare l'API EMR Serverless utilizzando il AWS SDK per Java, consulta gli [esempi di Java](#) nel nostro repository. GitHub Per esempi di come utilizzare l'API EMR Serverless utilizzando il, AWS SDK for Python (Boto) consulta Esempi in [Python](#) nel nostro repository. GitHub

Per creare un'applicazione, usa `create-application` È necessario specificare SPARK o HIVE come `application-type`. Questo comando restituisce l'ARN, il nome e l'ID dell'applicazione.

```
aws emr-serverless create-application \  
--name my-application-name \  
--type 'application-type' \  
--release-label release-version
```

Per descrivere un'applicazione, usala `get-application` e forniscila. `application-id` Questo comando restituisce le configurazioni relative allo stato e alla capacità dell'applicazione.

```
aws emr-serverless get-application \  
--application-id application-id
```

Per elencare tutte le tue applicazioni, chiama `list-applications` Questo comando restituisce le stesse proprietà `get-application` ma include tutte le applicazioni.

```
aws emr-serverless list-applications
```

Per eliminare la tua applicazione, chiama `delete-application` e fornisci il tuo `application-id`.

```
aws emr-serverless delete-application \  
--application-id application-id
```

# Configurazione di un'applicazione quando si lavora con EMR Serverless

Con EMR Serverless, puoi configurare le applicazioni che utilizzi. Ad esempio, è possibile impostare la capacità massima fino a cui un'applicazione può scalare, configurare la capacità preinizializzata per mantenere il conducente e gli operatori pronti a rispondere e specificare un set comune di configurazioni di runtime e monitoraggio a livello di applicazione. Le pagine seguenti descrivono come configurare le applicazioni quando si utilizza EMR Serverless.

## Argomenti

- [Comprensione del comportamento delle applicazioni in EMR Serverless](#)
- [Capacità preinizializzata per lavorare con un'applicazione in EMR Serverless](#)
- [Configurazione predefinita dell'applicazione per EMR Serverless](#)

## Comprensione del comportamento delle applicazioni in EMR Serverless

Questa sezione descrive il comportamento di invio dei lavori, la configurazione della capacità per la scalabilità e le impostazioni di configurazione del worker per EMR Serverless.

### Comportamento predefinito dell'applicazione

**Avvio automatico:** per impostazione predefinita, un'applicazione è configurata per l'avvio automatico all'invio del lavoro. È possibile disattivare questa funzionalità.

**Arresto automatico:** per impostazione predefinita, un'applicazione è configurata per l'arresto automatico quando è inattiva per 15 minuti. Quando un'applicazione passa allo STOPPED stato, rilascia qualsiasi capacità preinizializzata configurata. È possibile modificare la quantità di tempo di inattività prima dell'arresto automatico di un'applicazione oppure disattivare questa funzionalità.

### Capacità massima

È possibile configurare la capacità massima fino alla quale un'applicazione può scalare. È possibile specificare la capacità massima in termini di CPU, memoria (GB) e disco (GB).

#### Note

Si consiglia di configurare la capacità massima in modo che sia proporzionale alle dimensioni dei lavoratori supportate moltiplicando il numero di lavoratori per le loro dimensioni. Ad

esempio, se desideri limitare l'applicazione a 50 worker con 2 vCPUs, 16 GB per la memoria e 20 GB per il disco, imposta la capacità massima su 100 vCPUs, 800 GB per la memoria e 1000 GB per il disco.

## Configurazioni dei lavoratori supportate

La tabella seguente mostra le configurazioni e le dimensioni dei lavoratori supportate che è possibile specificare per EMR Serverless. È possibile configurare diverse dimensioni per driver ed esecutori in base alle esigenze del carico di lavoro.

CPU	Memoria	Archiviazione temporanea predefinita
1 vCPU	Minimo 2 GB, massimo 8 GB, con incrementi di 1 GB	20 GB - 200 GB
2 vCPU	Minimo 4 GB, massimo 16 GB, con incrementi di 1 GB	20 GB - 200 GB
4 vCPU	Minimo 8 GB, massimo 30 GB, con incrementi di 1 GB	20 GB - 200 GB
8 vCPU	Minimo 16 GB, massimo 60 GB, con incrementi di 4 GB	20 GB - 200 GB
16 vCPU	Minimo 32 GB, massimo 120 GB, con incrementi di 8 GB	20 GB - 200 GB

CPU: ogni lavoratore può avere 1, 2, 4, 8 o 16 CPUs v.

Memoria: ogni worker dispone di memoria, specificata in GB, entro i limiti elencati nella tabella precedente. I job Spark hanno un sovraccarico di memoria, il che significa che la memoria che usano è superiore alle dimensioni del contenitore specificate. Questo sovraccarico è specificato con le proprietà `e.spark.driver.memoryOverhead` e `spark.executor.memoryOverhead`. L'overhead ha un valore predefinito del 10% della memoria del contenitore, con un minimo di 384 MB. È necessario considerare questo sovraccarico quando si scelgono le dimensioni dei lavoratori.

Ad esempio, se scegli 4 v CPUs per la tua istanza di lavoro e una capacità di storage preinizializzata di 30 GB, dovresti impostare un valore di circa 27 GB come memoria executor per il tuo job Spark. Ciò massimizza l'utilizzo della capacità preinizializzata. La memoria utilizzabile sarebbe di 27 GB, più il 10% di 27 GB (2,7 GB), per un totale di 29,7 GB.

Disco: è possibile configurare ogni lavoratore con dischi di archiviazione temporanei con una dimensione minima di 20 GB e un massimo di 200 GB. Paghi solo lo spazio di archiviazione aggiuntivo oltre i 20 GB configurato per lavoratore.

## Capacità preinizializzata per lavorare con un'applicazione in EMR Serverless

EMR Serverless offre una funzionalità opzionale che mantiene il conducente e il personale preinizializzati e pronti a rispondere in pochi secondi. In questo modo si crea in modo efficace un pool di lavoratori accogliente per un'applicazione. Questa funzionalità è chiamata capacità preinizializzata. Per configurare questa funzionalità, è possibile impostare il `initialCapacity` parametro di un'applicazione sul numero di lavoratori che si desidera preinizializzare. Con la capacità dei lavoratori preinizializzata, i lavori iniziano immediatamente. Questa soluzione è ideale quando si desidera implementare applicazioni iterative e lavori urgenti.

La capacità preinizializzata consente a un gruppo ristretto di lavoratori di essere pronti all'avvio di lavori e sessioni in pochi secondi. Pagherai per i worker preinizializzati assegnati anche quando l'applicazione è inattiva, pertanto consigliamo di abilitarla per i casi d'uso che traggono vantaggio dai tempi di avvio rapidi e di dimensionarla per un utilizzo ottimale delle risorse. Le applicazioni EMR Serverless si chiudono automaticamente quando sono inattive. Si consiglia di mantenere attiva questa funzionalità quando si utilizzano worker preinizializzati per evitare addebiti imprevisti.

Quando invii un lavoro, se `initialCapacity` sono disponibili lavoratori di, il lavoro utilizza tali risorse per avviarne l'esecuzione. Se tali lavoratori sono già utilizzati da altri lavori o se il lavoro richiede più risorse di quelle disponibili `initialCapacity`, l'applicazione richiede e ottiene lavoratori aggiuntivi, fino ai limiti massimi di risorse impostati per l'applicazione. Al termine dell'esecuzione, un processo rilascia i worker utilizzati e il numero di risorse disponibili per l'applicazione torna a essere lo stesso `initialCapacity`. Un'applicazione mantiene le `initialCapacity` risorse anche dopo che i job hanno terminato l'esecuzione. L'applicazione rilascia risorse in eccesso oltre il periodo in `initialCapacity` cui i job non ne hanno più bisogno per l'esecuzione.

La capacità preinizializzata è disponibile e pronta all'uso all'avvio dell'applicazione. La capacità preinizializzata diventa inattiva quando l'applicazione viene arrestata. Un'applicazione passa allo `STARTED` stato solo se la capacità preinizializzata richiesta è stata creata ed è pronta per

l'uso. Per tutto il tempo in cui l'applicazione è nello `STARTED` stato, EMR Serverless mantiene la capacità preinizializzata disponibile per l'uso o l'uso da parte di lavori o carichi di lavoro interattivi. La funzionalità ripristina la capacità dei contenitori rilasciati o guasti. Ciò mantiene il numero di lavoratori specificato dal `InitialCapacity` parametro. Lo stato di un'applicazione senza capacità preinizializzata può cambiare immediatamente da `CREATED` a `STARTED`.

È possibile configurare l'applicazione in modo che rilasci la capacità preinizializzata se non viene utilizzata per un determinato periodo di tempo, con un valore predefinito di 15 minuti. Una candidatura interrotta si avvia automaticamente quando invii un nuovo lavoro. È possibile impostare queste configurazioni di avvio e arresto automatiche quando si crea l'applicazione oppure modificarle quando l'applicazione si trova in uno `STOPPED` stato `CREATED` o `STOPPED`.

È possibile modificare i `InitialCapacity` conteggi e specificare configurazioni di calcolo come CPU, memoria e disco per ogni lavoratore. Poiché non è possibile apportare modifiche parziali, è necessario specificare tutte le configurazioni di calcolo quando si modificano i valori. È possibile modificare le configurazioni solo quando l'applicazione è nello stato `STOPPED` o `CREATED`.

#### Note

Per ottimizzare l'uso delle risorse da parte dell'applicazione, consigliamo di allineare le dimensioni dei container alle dimensioni dei dipendenti con capacità preinizializzata. Ad esempio, se configuri la dimensione dell'esecutore Spark su 2 CPU e la memoria su 8 GB, ma la capacità preinizializzata del worker è 4 CPU con 16 GB di memoria, gli esecutori Spark utilizzano solo la metà delle risorse dei lavoratori quando vengono assegnati a questo lavoro.

## Personalizzazione della capacità preinizializzata per Spark e Hive

Puoi personalizzare ulteriormente la capacità preinizializzata per i carichi di lavoro eseguiti su specifici framework di big data. Ad esempio, quando un carico di lavoro viene eseguito su Apache Spark, puoi specificare quanti lavoratori iniziano come driver e quanti come esecutori. Allo stesso modo, quando si utilizza Apache Hive, è possibile specificare quanti lavoratori si avviano come driver Hive e quanti devono eseguire le attività Tez.

### Configurazione di un'applicazione che esegue Apache Hive con capacità preinizializzata

La seguente richiesta API crea un'applicazione che esegue Apache Hive basata sulla release `emr-6.6.0` di Amazon EMR. L'applicazione inizia con 5 driver Hive preinizializzati, ciascuno con 2

vCPU e 4 GB di memoria, e 50 task worker Tez preinizializzati, ciascuno con 4 vCPU e 8 GB di memoria. Quando le query Hive vengono eseguite su questa applicazione, utilizzano innanzitutto i worker preinizializzati e iniziano l'esecuzione immediatamente. Se tutti i worker preinizializzati sono occupati e vengono inviati più lavori Hive, l'applicazione può scalare fino a un totale di 400 vCPU e 1024 GB di memoria. Facoltativamente, è possibile omettere la capacità per il lavoratore o per il lavoratore. DRIVER TEZ\_TASK

```
aws emr-serverless create-application \
  --type "HIVE" \
  --name my-application-name \
  --release-label emr-6.6.0 \
  --initial-capacity '{
    "DRIVER": {
      "workerCount": 5,
      "workerConfiguration": {
        "cpu": "2vCPU",
        "memory": "4GB"
      }
    },
    "TEZ_TASK": {
      "workerCount": 50,
      "workerConfiguration": {
        "cpu": "4vCPU",
        "memory": "8GB"
      }
    }
  }' \
  --maximum-capacity '{
    "cpu": "400vCPU",
    "memory": "1024GB"
  }'
```

### Configurazione di un'applicazione che esegue Apache Spark con capacità preinizializzata

La seguente richiesta API crea un'applicazione che esegue Apache Spark 3.2.0 basata sulla versione 6.6.0 di Amazon EMR. L'applicazione inizia con 5 driver Spark preinizializzati, ciascuno con 2 vCPU e 4 GB di memoria, e 50 executor preinizializzati, ciascuno con 4 vCPU e 8 GB di memoria. Quando i job Spark vengono eseguiti su questa applicazione, utilizzano innanzitutto i worker preinizializzati e iniziano a essere eseguiti immediatamente. Se tutti i worker preinizializzati sono occupati e vengono inviati più job Spark, l'applicazione può scalare fino a un totale di 400 vCPU e 1024 GB di memoria. Facoltativamente, puoi omettere la capacità per il o il. DRIVER EXECUTOR

**Note**

Spark aggiunge un sovraccarico di memoria configurabile, con un valore predefinito del 10%, alla memoria richiesta per driver ed esecutori. Affinché i job utilizzino worker preinizializzati, la configurazione della memoria con capacità iniziale deve essere maggiore della memoria richiesta dal job e dall'overhead.

```
aws emr-serverless create-application \  
  --type "SPARK" \  
  --name my-application-name \  
  --release-label emr-6.6.0 \  
  --initial-capacity '{  
    "DRIVER": {  
      "workerCount": 5,  
      "workerConfiguration": {  
        "cpu": "2vCPU",  
        "memory": "4GB"  
      }  
    },  
    "EXECUTOR": {  
      "workerCount": 50,  
      "workerConfiguration": {  
        "cpu": "4vCPU",  
        "memory": "8GB"  
      }  
    }  
  }' \  
  --maximum-capacity '{  
    "cpu": "400vCPU",  
    "memory": "1024GB"  
  }'
```

## Configurazione predefinita dell'applicazione per EMR Serverless

È possibile specificare un set comune di configurazioni di runtime e monitoraggio a livello di applicazione per tutti i lavori inviati nella stessa applicazione. Ciò riduce il sovraccarico aggiuntivo associato alla necessità di inviare le stesse configurazioni per ogni lavoro.

È possibile modificare le configurazioni nei seguenti momenti:

- [Dichiarare le configurazioni a livello di applicazione al momento dell'invio del lavoro.](#)
- [Sostituisci le configurazioni predefinite durante l'esecuzione del lavoro.](#)

Le sezioni seguenti forniscono maggiori dettagli e un esempio per un ulteriore contesto.

## Dichiarazione delle configurazioni a livello di applicazione

È possibile specificare le proprietà di registrazione a livello di applicazione e di configurazione di runtime per i lavori inviati tramite l'applicazione.

### **monitoringConfiguration**

Per specificare le configurazioni di registro per i lavori inviati con l'applicazione, utilizza il campo. [monitoringConfiguration](#) Per ulteriori informazioni sulla registrazione per EMR Serverless, vedere. [Archiviazione dei registri](#)

### **runtimeConfiguration**

Per specificare proprietà di configurazione in fase di esecuzione `spark-defaults`, ad esempio, fornire un oggetto di configurazione nel campo. `runtimeConfiguration` Ciò influisce sulle configurazioni predefinite per tutti i lavori inviati con l'applicazione. Per ulteriori informazioni, consulta [Parametro di sovrascrittura della configurazione Hive](#) e [La configurazione di Spark sovrascrive il parametro.](#)

Le classificazioni di configurazione disponibili variano in base alla specifica release EMR Serverless. Ad esempio, le classificazioni per Log4j personalizzate `spark-executor-log4j2` sono disponibili solo con le versioni `spark-driver-log4j2 6.8.0` e successive. Per un elenco delle proprietà specifiche dell'applicazione, vedere e. [Proprietà del lavoro Spark](#) [Proprietà del lavoro di Hive](#)

È inoltre possibile configurare le [proprietà di Apache Log4j2](#), [AWS Secrets Manager per la protezione dei dati](#), e il runtime di [Java 17](#) a livello di applicazione.

Per trasmettere i segreti di Secrets Manager a livello di applicazione, allega la seguente policy agli utenti e ai ruoli che devono creare o aggiornare applicazioni EMR Serverless con segreti.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
```

```

        "Sid": "SecretsManagerPolicy",
        "Effect": "Allow",
        "Action": [
            "secretsmanager:GetSecretValue",
            "secretsmanager:DescribeSecret",
            "kms:Decrypt"
        ],
        "Resource": "arn:aws:secretsmanager:your-secret-arn"
    }
]
}

```

Per ulteriori informazioni sulla creazione di policy personalizzate per i segreti, consulta [Esempi di policy di autorizzazione AWS Secrets Manager nella Guida per l'AWS Secrets Manager utente](#).

### Note

runtimeConfiguration Ciò che specifichi a livello di applicazione viene applicationConfiguration mappato all'[StartJobRunAPI](#).

## Dichiarazione di esempio

L'esempio seguente mostra come dichiarare le configurazioni predefinite con. create-application

```

aws emr-serverless create-application \
  --release-label release-version \
  --type SPARK \
  --name my-application-name \
  --runtime-configuration '[
    {
      "classification": "spark-defaults",
      "properties": {
        "spark.driver.cores": "4",
        "spark.executor.cores": "2",
        "spark.driver.memory": "8G",
        "spark.executor.memory": "8G",
        "spark.executor.instances": "2",

"spark.hadoop.javax.jdo.option.ConnectionDriverName": "org.mariadb.jdbc.Driver",

```

```

        "spark.hadoop.java.jdbc.option.ConnectionURL": "jdbc:mysql://db-host:db-
port/db-name",
        "spark.hadoop.java.jdbc.option.ConnectionUserName": "connection-user-
name",
        "spark.hadoop.java.jdbc.option.ConnectionPassword":
"EMR.secret@SecretID"
    }
},
{
    "classification": "spark-driver-log4j2",
    "properties": {
        "rootLogger.level": "error",
        "logger.IdentifierForClass.name": "classpathForSettingLogger",
        "logger.IdentifierForClass.level": "info"
    }
}
]' \
--monitoring-configuration '{
    "s3MonitoringConfiguration": {
        "logUri": "s3://amzn-s3-demo-logging-bucket/logs/app-level"
    },
    "managedPersistenceMonitoringConfiguration": {
        "enabled": false
    }
}'

```

## Sovrascrivere le configurazioni durante l'esecuzione di un lavoro

È possibile specificare sostituzioni di configurazione per la configurazione dell'applicazione e la configurazione di monitoraggio con l'API. [StartJobRun](#) EMR Serverless unisce quindi le configurazioni specificate a livello di applicazione e a livello di processo per determinare le configurazioni per l'esecuzione del lavoro.

Il livello di granularità al momento dell'unione è il seguente:

- [ApplicationConfiguration](#)- Tipo di classificazione, ad esempio. spark-defaults
- [MonitoringConfiguration](#)- Tipo di configurazione, ad esempio s3MonitoringConfiguration.

**Note**

La priorità delle configurazioni fornite in [StartJobRun](#) sostituisce le configurazioni fornite a livello di applicazione.

Per ulteriori informazioni sulla classificazione delle priorità, vedere e. [Parametro di sovrascrittura della configurazione Hive](#) [La configurazione di Spark sovrascrive il parametro](#)

Quando si avvia un lavoro, se non si specifica una configurazione particolare, questa verrà ereditata dall'applicazione. Se si dichiarano le configurazioni a livello di processo, è possibile eseguire le seguenti operazioni:

- Sostituisci una configurazione esistente: fornisci lo stesso parametro di configurazione nella `StartJobRun` richiesta con i valori di override.
- Aggiungi una configurazione aggiuntiva: aggiungi il nuovo parametro di configurazione nella `StartJobRun` richiesta con i valori che desideri specificare.
- Rimuovere una configurazione esistente: per rimuovere una configurazione di runtime dell'applicazione, fornisci la chiave per la configurazione che desideri rimuovere e passa una dichiarazione vuota `{}` per la configurazione. Non è consigliabile rimuovere le classificazioni che contengono parametri necessari per l'esecuzione di un processo. Ad esempio, se si tenta di rimuovere le [proprietà richieste per un lavoro Hive](#), il processo avrà esito negativo.

Per rimuovere una configurazione di monitoraggio dell'applicazione, utilizzate il metodo appropriato per il tipo di configurazione pertinente:

- **cloudWatchLoggingConfiguration**- Per rimuovere `cloudWatchLogging`, passate il flag `enabled as false`.
- **managedPersistenceMonitoringConfiguration**- Per rimuovere le impostazioni di persistenza gestita e tornare allo stato abilitato predefinito, passa una dichiarazione vuota `{}` per la configurazione.
- **s3MonitoringConfiguration**- Per rimuovere `s3MonitoringConfiguration`, passa una dichiarazione vuota `{}` per la configurazione.

Esempio: override

L'esempio seguente mostra diverse operazioni che è possibile eseguire durante l'invio di un lavoro all'indirizzo. `start-job-run`

```

aws emr-serverless start-job-run \
  --application-id your-application-id \
  --execution-role-arn your-job-role-arn \
  --job-driver '{
    "sparkSubmit": {
      "entryPoint": "s3://us-east-1.elasticmapreduce/emr-containers/samples/
wordcount/scripts/wordcount.py",
      "entryPointArguments": ["s3://amzn-s3-demo-destination-bucket1/
wordcount_output"]
    }
  }' \
  --configuration-overrides '{
    "applicationConfiguration": [
      {
        // Override existing configuration for spark-defaults in the
application
        "classification": "spark-defaults",
        "properties": {
          "spark.driver.cores": "2",
          "spark.executor.cores": "1",
          "spark.driver.memory": "4G",
          "spark.executor.memory": "4G"
        }
      },
      {
        // Add configuration for spark-executor-log4j2
        "classification": "spark-executor-log4j2",
        "properties": {
          "rootLogger.level": "error",
          "logger.IdentifierForClass.name": "classpathForSettingLogger",
          "logger.IdentifierForClass.level": "info"
        }
      },
      {
        // Remove existing configuration for spark-driver-log4j2 from the
application
        "classification": "spark-driver-log4j2",
        "properties": {}
      }
    ],
    "monitoringConfiguration": {
      "managedPersistenceMonitoringConfiguration": {
        // Override existing configuration for managed persistence

```

```
        "enabled": true
    },
    "s3MonitoringConfiguration": {
        // Remove configuration of S3 monitoring
    },
    "cloudWatchLoggingConfiguration": {
        // Add configuration for CloudWatch logging
        "enabled": true
    }
}
}'
```

Al momento dell'esecuzione del lavoro, verranno applicate le seguenti classificazioni e configurazioni in base alla classificazione di priorità prioritaria descritta in and. [Parametro di sovrascrittura della configurazione Hive](#) [La configurazione di Spark sovrascrive il parametro](#)

- La classificazione `spark-defaults` verrà aggiornata con le proprietà specificate a livello di mansione. Ai fini di questa classificazione `StartJobRun` verranno prese in considerazione solo le proprietà incluse in.
- La classificazione `spark-executor-log4j2` verrà aggiunta all'elenco di classificazioni esistente.
- La classificazione `spark-driver-log4j2` verrà rimossa.
- Le configurazioni di `managedPersistenceMonitoringConfiguration` verranno aggiornate con le configurazioni a livello di processo.
- Le configurazioni di `s3MonitoringConfiguration` verranno rimosse.
- Le configurazioni di `cloudWatchLoggingConfiguration` verranno aggiunte alle configurazioni di monitoraggio esistenti.

## Personalizzazione di un'immagine EMR Serverless

A partire da Amazon EMR 6.9.0, puoi utilizzare immagini personalizzate per impacchettare le dipendenze delle applicazioni e gli ambienti di runtime in un unico contenitore con Amazon EMR Serverless. Ciò semplifica la gestione delle dipendenze del carico di lavoro e rende i pacchetti più portabili. La personalizzazione dell'immagine EMR Serverless offre i seguenti vantaggi:

- Installa e configura pacchetti ottimizzati per i carichi di lavoro. Questi pacchetti potrebbero non essere ampiamente disponibili nella distribuzione pubblica degli ambienti di runtime Amazon EMR.

- Integra EMR Serverless con gli attuali processi di compilazione, test e implementazione all'interno dell'organizzazione, inclusi sviluppo e test locali.
- Applica processi di sicurezza consolidati, come la scansione delle immagini, che soddisfano i requisiti di conformità e governance all'interno dell'organizzazione.
- Consente di utilizzare le proprie versioni di JDK e Python per le applicazioni.

EMR Serverless fornisce immagini che è possibile utilizzare come base per la creazione di immagini personalizzate. L'immagine di base fornisce i jar, la configurazione e le librerie essenziali per l'interazione dell'immagine con EMR Serverless. Puoi trovare l'immagine di base nella [Amazon ECR Public Gallery](#). Usa l'immagine che corrisponde al tipo di applicazione (Spark o Hive) e alla versione di rilascio. Ad esempio, se crei un'applicazione su Amazon EMR versione 6.9.0, utilizza le seguenti immagini.

Tipo	Immagine
Spark	<code>public.ecr.aws/emr-serverless/spark/emr-6.9.0:latest</code>
Hive	<code>public.ecr.aws/emr-serverless/hive/emr-6.9.0:latest</code>

## Prerequisiti

Prima di creare un'immagine personalizzata EMR Serverless, completare questi prerequisiti.

1. Crea un repository Amazon ECR nello stesso Regione AWS che usi per avviare le applicazioni EMR Serverless. Per creare un repository privato Amazon ECR, consulta [Creazione di un repository privato](#).
2. Per concedere agli utenti l'accesso al tuo repository Amazon ECR, aggiungi le seguenti policy agli utenti e ai ruoli che creano o aggiornano applicazioni EMR Serverless con immagini da questo repository.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "ECRRepositoryListGetPolicy",
```

```

    "Effect": "Allow",
    "Action": [
        "ecr:GetDownloadUrlForLayer",
        "ecr:BatchGetImage",
        "ecr:DescribeImages"
    ],
    "Resource": "ecr-repository-arn"
  }
]
}

```

Per ulteriori esempi di policy basate sull'identità di Amazon ECR, consulta [Esempi di policy basate sull'identità di Amazon Elastic Container Registry](#).

## Fase 1: Creare un'immagine personalizzata da immagini di base EMR Serverless

Innanzitutto, crea un [Dockerfile](#) che inizi con un'FROMistruzione che utilizza l'immagine di base preferita. Dopo l'FROMistruzione, puoi includere qualsiasi modifica che desideri apportare all'immagine. L'immagine di base imposta automaticamente USER suhadoop. Questa impostazione potrebbe non disporre delle autorizzazioni per tutte le modifiche incluse. Come soluzione alternativa, imposta suroot, modifica USER l'immagine e quindi reimpostala su. USER hadoop:hadoop Per visualizzare esempi di casi d'uso comuni, consulta [Utilizzo di immagini personalizzate con EMR Serverless](#).

```

# Dockerfile
FROM public.ecr.aws/emr-serverless/spark/emr-6.9.0:latest

USER root
# MODIFICATIONS GO HERE

# EMRS will run the image as hadoop
USER hadoop:hadoop

```

Dopo aver creato il Dockerfile, crea l'immagine con il seguente comando.

```

# build the docker image
docker build . -t aws-account-id.dkr.ecr.region.amazonaws.com/my-repository[:tag]or[@digest]

```

## Passaggio 2: convalida l'immagine localmente

EMR Serverless fornisce uno strumento offline in grado di controllare staticamente l'immagine personalizzata per convalidare file di base, variabili di ambiente e correggere le configurazioni dell'immagine. Per informazioni su come installare ed eseguire lo strumento, consulta [l'Amazon EMR Serverless Image CLI](#). GitHub

Dopo aver installato lo strumento, esegui il seguente comando per convalidare un'immagine:

```
amazon-emr-serverless-image \
validate-image -r emr-6.9.0 -t spark \
-i aws-account-id.dkr.ecr.region.amazonaws.com/my-repository:tag/@digest
```

Dovresti vedere un output simile al seguente.

```
Amazon EMR Serverless - Image CLI
Version: 0.0.1
... Checking if docker cli is installed
... Checking Image Manifest
[INFO] Image ID: 9e2f4359cf5beb466a8a2ed047ab61c9d37786c555655fc122272758f761b41a
[INFO] Created On: 2022-12-02T07:46:42.586249984Z
[INFO] Default User Set to hadoop:hadoop : PASS
[INFO] Working Directory Set to : PASS
[INFO] Entrypoint Set to /usr/bin/entrypoint.sh : PASS
[INFO] HADOOP_HOME is set with value: /usr/lib/hadoop : PASS
[INFO] HADOOP_LIBEXEC_DIR is set with value: /usr/lib/hadoop/libexec : PASS
[INFO] HADOOP_USER_HOME is set with value: /home/hadoop : PASS
[INFO] HADOOP_YARN_HOME is set with value: /usr/lib/hadoop-yarn : PASS
[INFO] HIVE_HOME is set with value: /usr/lib/hive : PASS
[INFO] JAVA_HOME is set with value: /etc/alternatives/jre : PASS
[INFO] TEZ_HOME is set with value: /usr/lib/tez : PASS
[INFO] YARN_HOME is set with value: /usr/lib/hadoop-yarn : PASS
[INFO] File Structure Test for hadoop-files in /usr/lib/hadoop: PASS
[INFO] File Structure Test for hadoop-jars in /usr/lib/hadoop/lib: PASS
[INFO] File Structure Test for hadoop-yarn-jars in /usr/lib/hadoop-yarn: PASS
[INFO] File Structure Test for hive-bin-files in /usr/bin: PASS
[INFO] File Structure Test for hive-jars in /usr/lib/hive/lib: PASS
[INFO] File Structure Test for java-bin in /etc/alternatives/jre/bin: PASS
[INFO] File Structure Test for tez-jars in /usr/lib/tez: PASS
-----
Overall Custom Image Validation Succeeded.
-----
```

## Fase 3: carica l'immagine nel tuo repository Amazon ECR

Invia la tua immagine Amazon ECR al tuo repository Amazon ECR con i seguenti comandi. Assicurati di disporre delle autorizzazioni IAM corrette per inviare l'immagine al tuo repository. Per ulteriori informazioni, consulta [Pushing an image](#) nella Amazon ECR User Guide.

```
# login to ECR repo
aws ecr get-login-password --region region | docker login --username AWS --password-
stdin aws-account-id.dkr.ecr.region.amazonaws.com

# push the docker image
docker push aws-account-id.dkr.ecr.region.amazonaws.com/my-repository:tag/@digest
```

## Passaggio 4: creare o aggiornare un'applicazione con immagini personalizzate

Scegli la AWS Management Console scheda o la AWS CLI scheda in base al modo in cui desideri avviare l'applicazione, quindi completa i seguenti passaggi.

### Console

1. [Accedere alla console EMR Studio all'indirizzo https://console.aws.amazon.com/emr](https://console.aws.amazon.com/emr). Accedi alla tua applicazione o crea una nuova applicazione seguendo le istruzioni riportate in [Creare un'applicazione](#).
2. Per specificare immagini personalizzate quando si crea o si aggiorna un'applicazione EMR Serverless, selezionare Impostazioni personalizzate nelle opzioni di configurazione dell'applicazione.
3. Nella sezione Impostazioni personalizzate dell'immagine, selezionare la casella di controllo Usa l'immagine personalizzata con questa applicazione.
4. Incolla l'URI dell'immagine Amazon ECR nel campo URI dell'immagine. EMR Serverless utilizza questa immagine per tutti i tipi di worker per l'applicazione. In alternativa, puoi scegliere Immagini personalizzate diverse e incollare immagini Amazon ECR diverse URIs per ogni tipo di lavoratore.

### CLI

- Crea un'applicazione con il `image-configuration` parametro. EMR Serverless applica questa impostazione a tutti i tipi di lavoratore.

```
aws emr-serverless create-application \
--release-label emr-6.9.0 \
--type SPARK \
--image-configuration '{
  "imageUri": "aws-account-id.dkr.ecr.region.amazonaws.com/my-repository:tag/
@digest"
}'
```

Per creare un'applicazione con impostazioni di immagine diverse per ogni tipo di lavoratore, utilizzare il `worker-type-specifications` parametro.

```
aws emr-serverless create-application \
--release-label emr-6.9.0 \
--type SPARK \
--worker-type-specifications '{
  "Driver": {
    "imageConfiguration": {
      "imageUri": "aws-account-id.dkr.ecr.region.amazonaws.com/my-
repository:tag/@digest"
    }
  },
  "Executor" : {
    "imageConfiguration": {
      "imageUri": "aws-account-id.dkr.ecr.region.amazonaws.com/my-
repository:tag/@digest"
    }
  }
}'
```

Per aggiornare un'applicazione, utilizzate il `image-configuration` parametro. EMR Serverless applica questa impostazione a tutti i tipi di lavoratore.

```
aws emr-serverless update-application \
--application-id application-id \
--image-configuration '{
  "imageUri": "aws-account-id.dkr.ecr.region.amazonaws.com/my-repository:tag/
@digest"
}'
```

## Fase 5: consentire a EMR Serverless di accedere all'archivio di immagini personalizzato

Aggiungi la seguente policy sulle risorse al repository Amazon ECR per consentire al responsabile del servizio EMR Serverless di utilizzare `getdescribe`, e `download` le richieste provenienti da questo repository.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "Emr Serverless Custom Image Support",
      "Effect": "Allow",
      "Principal": {
        "Service": "emr-serverless.amazonaws.com"
      },
      "Action": [
        "ecr:BatchGetImage",
        "ecr:DescribeImages",
        "ecr:GetDownloadUrlForLayer"
      ],
      "Condition": {
        "StringEquals": {
          "aws:SourceArn": "arn:aws:emr-serverless:region:aws-account-id:/
applications/application-id"
        }
      }
    }
  ]
}
```

Come best practice di sicurezza, aggiungi una chiave di `aws:SourceArn` condizione alla policy del repository. La chiave di condizione globale IAM `aws:SourceArn` garantisce che EMR Serverless utilizzi il repository solo per l'ARN di un'applicazione. Per ulteriori informazioni sulle politiche del repository Amazon ECR, consulta [Creazione di un repository privato](#).

## Considerazioni e limitazioni

Quando lavori con immagini personalizzate, considera quanto segue:

- Usa l'immagine di base corretta che corrisponda al tipo (Spark o Hive) e all'etichetta di rilascio (ad esempio `emr-6.9.0`) dell'applicazione.
- EMR Serverless ignora le nostre `[ENTRYPOINT]` istruzioni nel `[CMD]` file Docker. Usa istruzioni comuni nel file Docker, ad esempio, `e. [COPY] [RUN] [WORKDIR]`
- Non è necessario modificare le variabili di ambiente `JAVA_HOMESPARK_HOME`, `HIVE_HOME`, `TEZ_HOME` quando si crea un'immagine personalizzata.
- Le immagini personalizzate non possono superare i 10 GB di dimensione.
- Se modifichi file binari o jar nelle immagini di base di Amazon EMR, ciò potrebbe causare errori nell'avvio dell'applicazione o del processo.
- Il repository Amazon ECR deve trovarsi nello stesso Regione AWS che usi per lanciare le applicazioni EMR Serverless.

## Configurazione dell'accesso VPC per le applicazioni EMR Serverless per la connessione ai dati

Puoi configurare applicazioni EMR Serverless per connetterti ai tuoi archivi di dati all'interno del tuo VPC, come cluster Amazon Redshift, database Amazon RDS o bucket Amazon S3 con endpoint VPC. La tua applicazione EMR Serverless dispone di connettività in uscita agli archivi dati all'interno del tuo VPC. Per impostazione predefinita, EMR Serverless blocca l'accesso in entrata alle applicazioni per migliorare la sicurezza.

### Note

È necessario configurare l'accesso VPC se si desidera utilizzare un database metastore Hive esterno per l'applicazione. [Per informazioni su come configurare un metastore Hive esterno, consulta Configurazione di Metastore.](#)

## Crea applicazione

Nella pagina Crea applicazione, è possibile scegliere impostazioni personalizzate e specificare il VPC, le sottoreti e i gruppi di sicurezza che le applicazioni EMR Serverless possono utilizzare.

## VPCs

Scegli il nome del cloud privato virtuale (VPC) che contiene i tuoi archivi dati. La pagina Crea applicazione elenca tutti quelli VPCs che hai scelto Regione AWS.

## Sottoreti

Scegli le sottoreti all'interno del VPC che contiene il tuo data store. La pagina Crea applicazione elenca tutte le sottoreti per gli archivi dati nel tuo VPC. Sono supportate sia le sottoreti pubbliche che quelle private. Puoi passare sottoreti private o pubbliche alle tue applicazioni. La scelta di disporre di una sottorete pubblica o privata comporta alcune considerazioni di cui tenere conto.

Per le sottoreti private:

- Le tabelle di routing associate non devono avere gateway Internet.
- Per la connettività in uscita a Internet, se necessario, configura i percorsi in uscita utilizzando un gateway NAT. [Per configurare un gateway NAT, vedi Gateway NAT.](#)
- Per la connettività Amazon S3, configura un gateway NAT o un endpoint VPC. Per configurare un endpoint VPC S3, [consulta Creare un endpoint gateway.](#)
- Se configuri un endpoint VPC S3 e alleggi una policy endpoint per controllare l'accesso, devi seguire le istruzioni in [Logging for EMR Serverless with managed storage per fornire le autorizzazioni a EMR Serverless per archiviare](#) e servire i log delle applicazioni.
- Per la connettività con altri Servizi AWS dispositivi esterni al VPC, ad esempio Amazon DynamoDB, configura gli endpoint VPC o un gateway NAT. Per configurare gli endpoint VPC per Servizi AWS, consulta [Lavorare con gli endpoint VPC.](#)

### Note

Quando configuri un'applicazione Amazon EMR Serverless in una sottorete privata, ti consigliamo di configurare anche gli endpoint VPC per Amazon S3. Se la tua applicazione EMR Serverless si trova in una sottorete privata senza endpoint VPC per Amazon S3, potresti incorrere in costi aggiuntivi per il gateway NAT associati al traffico S3. Questo perché il traffico tra l'applicazione EMR e Amazon S3 non rimarrà all'interno del VPC quando gli endpoint VPC non sono configurati.

Per le sottoreti pubbliche:

- Questi hanno un percorso verso un Internet Gateway.
- È necessario garantire configurazioni adeguate dei gruppi di sicurezza per controllare il traffico in uscita.

I lavoratori possono connettersi agli archivi dati all'interno del tuo VPC tramite il traffico in uscita. Per impostazione predefinita, EMR Serverless blocca l'accesso in entrata ai lavoratori. Questo serve a migliorare la sicurezza.

Quando si utilizza AWS Config, EMR Serverless crea un record di elementi dell'interfaccia di rete elastica per ogni lavoratore. Per evitare i costi legati a questa risorsa, prendi in considerazione la possibilità di disattivarla. `AWS::EC2::NetworkInterface` AWS Config

#### Note

Ti consigliamo di selezionare più sottoreti in più zone di disponibilità. Questo perché le sottoreti scelte determinano le zone di disponibilità disponibili per l'avvio di un'applicazione EMR Serverless. Ogni lavoratore utilizza un indirizzo IP nella sottorete in cui viene avviato. Assicurati che le sottoreti specificate abbiano indirizzi IP sufficienti per il numero di worker che intendi avviare. Per ulteriori informazioni sulla pianificazione delle sottoreti, vedere. [the section called “Procedure consigliate per la pianificazione delle sottoreti”](#)

#### Considerazioni e limitazioni per le sottoreti

- EMR Serverless con sottoreti pubbliche non supporta Lake Formation. AWS
- Il traffico in entrata non è supportato per le sottoreti pubbliche.

#### Gruppi di sicurezza

Scegli uno o più gruppi di sicurezza in grado di comunicare con i tuoi archivi di dati. La pagina Crea applicazione elenca tutti i gruppi di sicurezza nel tuo VPC. EMR Serverless associa questi gruppi di sicurezza a interfacce di rete elastiche collegate alle sottoreti VPC.

#### Note

Si consiglia di creare un gruppo di sicurezza separato per le applicazioni EMR Serverless. EMR Serverless non consente di creare, aggiornare, avviare un'applicazione se i gruppi di

sicurezza hanno porte aperte alla rete Internet pubblica su 0.0.0.0/0 o nell'intervallo: :/0. Ciò offre maggiore sicurezza e isolamento e rende più efficiente la gestione delle regole di rete. Ad esempio, questo blocca il traffico imprevisto verso i lavoratori con indirizzi IP pubblici. Per comunicare con i cluster Amazon Redshift, ad esempio, puoi definire le regole del traffico tra i gruppi di sicurezza Serverless di Redshift ed EMR, come illustrato nell'esempio seguente.

### Example Esempio: comunicazione con i cluster Amazon Redshift

1. Aggiungi una regola per il traffico in entrata al gruppo di sicurezza Amazon Redshift da uno dei gruppi di sicurezza EMR Serverless.

Tipo	Protocollo	Intervallo porte	Origine
Tutte le regole TCP	TCP	5439	emr-serverless-security-group

2. Aggiungi una regola per il traffico in uscita da uno dei gruppi di sicurezza EMR Serverless. Ci sono due modi per farlo. Innanzitutto, è possibile aprire il traffico in uscita verso tutte le porte.

Tipo	Protocollo	Intervallo porte	Destinazione
Tutto il traffico	TCP	ALL	0.0.0.0/0

In alternativa, puoi limitare il traffico in uscita ai cluster Amazon Redshift. Ciò è utile solo quando l'applicazione deve comunicare con i cluster Amazon Redshift e nient'altro.

Tipo	Protocollo	Intervallo porte	Origine
Tutte le regole TCP	TCP	5439	redshift-security-group

## Configura l'applicazione

È possibile modificare la configurazione di rete per un'applicazione EMR Serverless esistente dalla pagina Configura applicazione.

### Visualizza i dettagli dell'esecuzione del processo

Nella pagina dei dettagli del Job run, è possibile visualizzare la sottorete utilizzata dal job per un'esecuzione specifica. Si noti che un processo viene eseguito solo in una sottorete selezionata dalle sottoreti specificate.

## Procedure consigliate per la pianificazione delle sottoreti

AWS le risorse vengono create in una sottorete che è un sottoinsieme di indirizzi IP disponibili in un Amazon VPC. Ad esempio, un VPC con maschera di rete /16 ha fino a 65.536 indirizzi IP disponibili che possono essere suddivisi in più reti più piccole utilizzando maschere di sottorete. Ad esempio, è possibile suddividere questo intervallo in due sottoreti, ognuna delle quali utilizza la maschera /17 e 32.768 indirizzi IP disponibili. Una sottorete si trova all'interno di una zona di disponibilità e non può estendersi su più zone.

Le sottoreti devono essere progettate tenendo conto dei limiti di scalabilità delle applicazioni EMR Serverless. Ad esempio, se un'applicazione richiede 4 vCPU worker ed è possibile scalare fino a 4.000 vCPU, l'applicazione richiederà al massimo 1.000 worker per un totale di 1.000 interfacce di rete. Si consiglia di creare sottoreti su più zone di disponibilità. Ciò consente a EMR Serverless di riprovare il lavoro o di fornire capacità preinizializzata in una zona di disponibilità diversa nell'improbabile eventualità di un guasto in una zona di disponibilità. Pertanto, ogni sottorete in almeno due zone di disponibilità deve avere più di 1.000 indirizzi IP disponibili.

Sono necessarie sottoreti con una dimensione della maschera inferiore o uguale a 22 per effettuare il provisioning di 1.000 interfacce di rete. Qualsiasi maschera superiore a 22 non soddisferà il requisito. Ad esempio, una subnet mask di /23 fornisce 512 indirizzi IP, mentre una maschera di /22 fornisce 1024 e una maschera di /21 fornisce 2048 indirizzi IP. Di seguito è riportato un esempio di 4 sottoreti con maschera /22 in un VPC di /16 netmask che possono essere allocate a diverse zone di disponibilità. Esiste una differenza di cinque tra gli indirizzi IP disponibili e quelli utilizzabili perché i primi quattro indirizzi IP e l'ultimo indirizzo IP in ogni sottorete sono riservati da AWS

ID sottorete	Indirizzo di sottorete	Maschera di sottorete	Intervallo di indirizzi IP	Indirizzi IP disponibili	Indirizzi IP utilizzabili
1	10.0.0.0	255,255,252,0/22	10.0.0.0 - 10.0.3.255	1,024	1.019
2	10,04,0	255,255,252,0/22	10.0.4.0 - 10.0.7.255	1,024	1.019
3	10,08.0	255,255,252,0/22	10.0.4.0 - 10.0.7.255	1,024	1.019
4	10,012,0	255,255,252,0/22	10.0.12.0 - 10.0.15.255	1,024	1.019

Dovresti valutare se il tuo carico di lavoro è più adatto per lavoratori di grandi dimensioni. L'utilizzo di lavoratori di dimensioni maggiori richiede un minor numero di interfacce di rete. Ad esempio, l'utilizzo di worker a 16 vCPU con un limite di scalabilità delle applicazioni di 4.000 vCPU richiederà al massimo 250 lavoratori per un totale di 250 indirizzi IP disponibili per il provisioning delle interfacce di rete. Per effettuare il provisioning di 250 interfacce di rete sono necessarie sottoreti in più zone di disponibilità con una dimensione della maschera inferiore o uguale a 24. Qualsiasi maschera di dimensioni superiori a 24 offre meno di 250 indirizzi IP.

Se condividi sottoreti tra più applicazioni, ogni sottorete deve essere progettata tenendo conto dei limiti di scalabilità collettivi di tutte le applicazioni. Ad esempio, se hai 3 applicazioni che richiedono 4 vCPU worker e ciascuna può scalare fino a 4000 vCPU con una quota di servizio basata su 12.000 vCPU a livello di account, ogni sottorete richiederà 3000 indirizzi IP disponibili. In caso contrario, è possibile provare ad aumentare il numero di indirizzi IP disponibili. Tale operazione può essere effettuata associando i blocchi di instradamento interdominio senza classi (CIDR) secondari al VPC. Per ulteriori informazioni, consulta [Associare blocchi IPv4 CIDR aggiuntivi al tuo VPC](#) nella Amazon VPC User Guide.

Puoi utilizzare uno dei tanti strumenti disponibili online per generare rapidamente definizioni di sottorete e rivedere la gamma di indirizzi IP disponibili.

# Opzioni di architettura Serverless Amazon EMR

L'architettura del set di istruzioni dell'applicazione Amazon EMR Serverless determina il tipo di processori utilizzati dall'applicazione per eseguire il processo. Amazon EMR offre due opzioni di architettura per l'applicazione: x86\_64 e arm64. EMR Serverless si aggiorna automaticamente alle istanze di ultima generazione non appena sono disponibili, in modo che le applicazioni possano utilizzare le istanze più recenti senza richiedere ulteriori sforzi da parte dell'utente.

## Argomenti

- [Utilizzo dell'architettura x86\\_64](#)
- [Utilizzo dell'architettura arm64 \(Graviton\)](#)
- [Lancio di nuove applicazioni con supporto Graviton](#)
- [Configurazione delle applicazioni esistenti per utilizzare Graviton](#)
- [Considerazioni sull'utilizzo di Graviton](#)

## Utilizzo dell'architettura x86\_64

L'architettura x86\_64 è anche nota come x86 64-bit o x64. x86\_64 è l'opzione predefinita per le applicazioni EMR Serverless. Questa architettura utilizza processori basati su x86 ed è compatibile con la maggior parte degli strumenti e delle librerie di terze parti.

La maggior parte delle applicazioni è compatibile con la piattaforma hardware x86 e può essere eseguita correttamente sull'architettura x86\_64 predefinita. Tuttavia, se l'applicazione è compatibile con ARM a 64 bit, è possibile passare ad arm64 per utilizzare i processori Graviton per migliorare prestazioni, potenza di calcolo e memoria. L'esecuzione di istanze sull'architettura arm64 costa meno rispetto a quando si eseguono istanze di uguali dimensioni sull'architettura x86.

## Utilizzo dell'architettura arm64 (Graviton)

AWS I processori Graviton sono progettati su misura AWS con core ARM Neoverse a 64 bit e sfruttano l'architettura arm64 (nota anche come Arch64 o ARM a 64 bit). La linea di processori AWS Graviton disponibile su EMR Serverless include processori Graviton3 e Graviton2. Questi processori offrono un rapporto prezzo/prestazioni superiore per i carichi di lavoro Spark e Hive rispetto ai carichi di lavoro equivalenti eseguiti sull'architettura x86\_64. EMR Serverless utilizza automaticamente i processori di ultima generazione quando disponibili senza alcuno sforzo da parte dell'utente per l'aggiornamento alla generazione di processori di ultima generazione.

## Lancio di nuove applicazioni con supporto Graviton

Utilizzate uno dei seguenti metodi per avviare un'applicazione che utilizza l'architettura arm64.

### AWS CLI

Per avviare un'applicazione che utilizza i processori Graviton da AWS CLI, specifica ARM64 come `architecture` parametro nell'`create-applicationAPI`. Fornisci i valori appropriati per la tua applicazione negli altri parametri.

```
aws emr-serverless create-application \  
  --name my-graviton-app \  
  --release-label emr-6.8.0 \  
  --type "SPARK" \  
  --architecture "ARM64" \  
  --region us-west-2
```

### EMR Studio

Per avviare un'applicazione utilizzando i processori Graviton di EMR Studio, scegli arm64 come opzione Architettura quando crei o aggiorni un'applicazione.

## Configurazione delle applicazioni esistenti per utilizzare Graviton

Puoi configurare le tue applicazioni Amazon EMR Serverless esistenti per utilizzare l'architettura Graviton (arm64) con l'SDK o EMR Studio. AWS CLI

Per convertire un'applicazione esistente da x86 a arm64

1. Conferma di utilizzare la versione principale più recente di [AWS CLI/SDK che supporta](#) il parametro. `architecture`
2. Verificate che non vi siano lavori in esecuzione, quindi arrestate l'applicazione.

```
aws emr-serverless stop-application \  
  --application-id application-id \  
  --region us-west-2
```

3. Per aggiornare l'applicazione per utilizzare Graviton, specifica ARM64 il `architecture` parametro nell'`update-applicationAPI`.

```
aws emr-serverless update-application \  
--application-id application-id \  
--architecture 'ARM64' \  
--region us-west-2
```

4. Per verificare che l'architettura della CPU dell'applicazione sia aggiornata ARM64, utilizza l'get-applicationAPI.

```
aws emr-serverless get-application \  
--application-id application-id \  
--region us-west-2
```

5. Quando sei pronto, riavvia l'applicazione.

```
aws emr-serverless start-application \  
--application-id application-id \  
--region us-west-2
```

## Considerazioni sull'utilizzo di Graviton

Prima di avviare un'applicazione EMR Serverless utilizzando arm64 per il supporto Graviton, confermare quanto segue.

### Compatibilità con le librerie

Quando selezioni Graviton (arm64) come opzione di architettura, assicurati che i pacchetti e le librerie di terze parti siano compatibili con l'architettura ARM a 64 bit. Per informazioni su come impacchettare le librerie Python in un ambiente virtuale Python compatibile con l'architettura selezionata, vedere. [Utilizzo delle librerie Python con EMR Serverless](#)

Per saperne di più su come configurare un carico di lavoro Spark o Hive per utilizzare ARM a 64 bit, consulta il repository [AWS Graviton](#) Getting Started su GitHub. Questo repository contiene risorse essenziali che possono aiutarti a iniziare con Graviton basato su ARM.

# Concorrenza dei job e accodamento per un'applicazione EMR Serverless

A partire dalla versione 7.0.0 di Amazon EMR e successive, puoi specificare il timeout della coda di esecuzione dei job e la configurazione della concorrenza per la tua applicazione. Quando specifichi questa configurazione, Amazon EMR Serverless inizia mettendo in coda il lavoro e inizia l'esecuzione in base all'utilizzo simultaneo dell'applicazione. Ad esempio, se la contemporanea esecuzione di un processo è pari a 10, sull'applicazione vengono eseguiti solo dieci processi alla volta. I job rimanenti vengono messi in coda fino al termine di uno dei job in esecuzione. Se il timeout della coda viene raggiunto prima, il job scade. Per ulteriori informazioni, vedere [Job run states](#).

## Principali vantaggi della concorrenza e dell'accodamento

La concorrenza e l'accodamento dei lavori offrono i seguenti vantaggi quando sono necessari molti invii di lavoro:

- Aiuta a controllare i processi di esecuzione simultanei per utilizzare in modo efficiente i limiti di capacità a livello di applicazione.
- La coda può contenere una serie improvvisa di invii di lavori, con un'impostazione di timeout configurabile.

## Guida introduttiva alla concorrenza e alle code

Le procedure seguenti mostrano un paio di modi diversi per implementare la concorrenza e l'accodamento.

### Usando il AWS CLI

1. Crea un'applicazione Amazon EMR Serverless con timeout di coda ed esecuzioni di job simultanee:

```
aws emr-serverless create-application \  
--release-label emr-7.0.0 \  
--type SPARK \  
--scheduler-configuration '{"maxConcurrentRuns": 1, "queueTimeoutMinutes": 30}'
```

2. Aggiorna un'applicazione per modificare il timeout e la concorrenza della coda dei lavori:

```
aws emr-serverless update-application \  
--application-id application-id \  
--scheduler-configuration '{"maxConcurrentRuns": 5, "queueTimeoutMinutes": 30}'
```

### Note

È possibile aggiornare l'applicazione esistente per abilitare la concorrenza e l'accodamento dei lavori. A tale scopo, l'applicazione deve avere un'etichetta di rilascio emr-7.0.0 o successiva.

## Usando il AWS Management Console

Nei passaggi seguenti viene illustrato come iniziare a utilizzare la concorrenza e l'accodamento dei lavori, utilizzando: AWS Management Console

1. Vai a EMR Studio e scegli di creare un'applicazione con etichetta di rilascio EMR-7.0.0 o superiore.
2. In Opzioni di configurazione dell'applicazione, seleziona l'opzione Usa impostazioni personalizzate.
3. In Configurazioni aggiuntive c'è una sezione per le impostazioni del Job Run. Seleziona l'opzione Abilita la concorrenza dei lavori per abilitare la funzionalità.
4. Una volta selezionata, è possibile selezionare sia Esecuzioni di job simultanee che Queue timeout per configurare rispettivamente il numero di esecuzioni di job simultanee e il timeout della coda. Se non si inseriscono valori per queste impostazioni, vengono utilizzati i valori predefiniti.
5. Scegliete Crea applicazione e l'applicazione verrà creata con questa funzionalità abilitata. Per verificare, vai alla dashboard, seleziona l'applicazione e controlla nella scheda delle proprietà per determinare se la funzionalità è abilitata.

Dopo la configurazione, puoi inviare lavori con questa funzionalità abilitata.

## Considerazioni sulla concorrenza e sull'accodamento

Quando implementate la concorrenza e l'accodamento, tenete in considerazione quanto segue:

- Job concurrency and queuing è supportato dalla release 7.0.0 e successive di Amazon EMR.
- La concorrenza e l'accodamento dei job sono abilitati per impostazione predefinita su Amazon EMR versione 7.3.0 e successive.
- È possibile aggiornare la concorrenza per un'applicazione nello stato STARTED.
- L'intervallo valido per maxConcurrentRuns è compreso tra 1 e 1000 e queueTimeoutMinutes è compreso tra 15 e 720.
- Lo stato IN CODA può essere impostato su un massimo di 2000 lavori per account.
- La concorrenza e l'accodamento si applicano ai lavori in batch e in streaming. Non può essere utilizzato per lavori interattivi. Per ulteriori informazioni, consulta [Eseguire carichi di lavoro interattivi con EMR Serverless tramite EMR Studio](#).

# Trasferisci i dati in S3 Express One Zone con EMR Serverless

Con le versioni 7.2.0 e successive di Amazon EMR, puoi usare EMR Serverless con la classe di storage [Amazon S3 Express One Zone](#) per migliorare le prestazioni durante l'esecuzione di job e carichi di lavoro. S3 Express One Zone è una classe di storage Amazon S3 a zona singola ad alte prestazioni che offre un accesso ai dati coerente a una cifra in millisecondi per la maggior parte delle applicazioni sensibili alla latenza. Al momento del suo rilascio, S3 Express One Zone offre lo storage di oggetti cloud con la latenza più bassa e le prestazioni più elevate in Amazon S3.

## Prerequisiti

- **Autorizzazioni S3 Express One Zone:** quando S3 Express One Zone esegue inizialmente un'azione come o su un oggetto S3GET, LIST la classe di storage chiama per tuo conto. PUT CreateSession La tua policy IAM deve consentire l'autorizzazione in modo che s3express:CreateSession S3A il connettore può richiamare l>CreateSessionAPI. Per un esempio di policy con questa autorizzazione, consulta [Nozioni di base su S3 Express One Zone](#).
- **S3A connettore:** per configurare Spark per accedere ai dati da un bucket Amazon S3 che utilizza la classe di storage S3 Express One Zone, devi utilizzare il connettore Apache Hadoop S3A. Per utilizzare il connettore, assicurati che tutti gli S3 URIs utilizzino lo s3a schema. In caso contrario, puoi modificare l'implementazione del file system che utilizzi per gli schemi s3 e s3n.

Per modificare lo schema s3, specifica le seguenti configurazioni del cluster:

```
[
  {
    "Classification": "core-site",
    "Properties": {
      "fs.s3.impl": "org.apache.hadoop.fs.s3a.S3AFileSystem",
      "fs.AbstractFileSystem.s3.impl": "org.apache.hadoop.fs.s3a.S3A"
    }
  }
]
```

Per modificare lo schema s3n, specifica le seguenti configurazioni del cluster:

```
[
  {
    "Classification": "core-site",
    "Properties": {
      "fs.s3n.impl": "org.apache.hadoop.fs.s3a.S3AFileSystem",
      "fs.AbstractFileSystem.s3n.impl": "org.apache.hadoop.fs.s3a.S3A"
    }
  }
]
```

## Nozioni di base su S3 Express One Zone

Segui questi passaggi per iniziare a usare S3 Express One Zone.

1. [Crea un endpoint VPC](#). Aggiungi l'endpoint `com.amazonaws.us-west-2.s3express` all'endpoint VPC.
2. Segui [Guida introduttiva ad Amazon EMR Serverless](#) per creare un'applicazione con etichetta di rilascio Amazon EMR 7.2.0 o successiva.
3. [Configura la tua applicazione](#) per utilizzare l'endpoint VPC appena creato, un gruppo di sottoreti privato e un gruppo di sicurezza.
4. Aggiungi l'`CreateSession` autorizzazione al tuo ruolo di esecuzione del lavoro.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Resource": "*",
      "Action": [
        "s3express:CreateSession"
      ]
    }
  ]
}
```

5. Esegui il tuo lavoro. Tieni presente che devi utilizzare lo S3A schema per accedere ai bucket S3 Express One Zone.

```
aws emr-serverless start-job-run \
```

```
--application-id <application-id> \  
--execution-role-arn <job-role-arn> \  
--name <job-run-name> \  
--job-driver '{  
  "sparkSubmit": {  
  
    "entryPoint": "s3a://<DOC-EXAMPLE-BUCKET>/scripts/wordcount.py",  
    "entryPointArguments":["s3a://<DOC-EXAMPLE-BUCKET>/emr-serverless-spark/output"],  
    "sparkSubmitParameters": "--conf spark.executor.cores=4  
--conf spark.executor.memory=8g --conf spark.driver.cores=4  
--conf spark.driver.memory=8g --conf spark.executor.instances=2  
--conf spark.hadoop.fs.s3a.change.detection.mode=none  
--conf spark.hadoop.fs.s3a.endpoint.region={<AWS_REGION>}  
--conf spark.hadoop.fs.s3a.select.enabled=false  
--conf spark.sql.sources.fastS3PartitionDiscovery.enabled=false  
  }'  
'
```

# Esecuzione di processi

Dopo aver fornito la candidatura, è possibile inviare offerte di lavoro all'applicazione. Questa sezione spiega come utilizzare AWS CLI per eseguire questi lavori. Questa sezione identifica anche i valori predefiniti per ogni tipo di applicazione disponibile su EMR Serverless.

## Argomenti

- [Stati delle esecuzioni di processi](#)
- [Annullamento dell'esecuzione di un processo EMR Serverless con periodo di grazia](#)
- [Esecuzione di processi dalla console EMR Studio](#)
- [Esecuzione di lavori da AWS CLI](#)
- [Utilizzo di dischi ottimizzati per la riproduzione casuale](#)
- [Lavori di streaming per l'elaborazione di dati in streaming continuo](#)
- [Utilizzo delle configurazioni Spark quando si eseguono job EMR Serverless](#)
- [Utilizzo delle configurazioni Hive quando si eseguono job EMR Serverless](#)
- [Resilienza EMR Serverless Job](#)
- [Configurazione Metastore per EMR Serverless](#)
- [Accesso ai dati S3 in un altro AWS account da EMR Serverless](#)
- [Risoluzione degli errori in EMR Serverless](#)

## Stati delle esecuzioni di processi

Quando invii un'esecuzione di lavoro a una coda di processi Serverless di Amazon EMR, l'esecuzione del processo entra nello stato. SUBMITTED Lo stato di un processo passa da «a» SUBMITTED RUNNING fino a raggiungere FAILED, SUCCESS o. CANCELLING

Le esecuzioni di processi possono avere i seguenti stati:

Stato	Descrizione
Inviato	Lo stato iniziale del processo quando si invia un'esecuzione di lavoro a EMR Serverless. Il lavoro attende di essere pianificato per l'applica

Stato	Descrizione
	zione. EMR Serverless inizia a stabilire le priorità e a pianificare l'esecuzione del lavoro.
In coda	L'esecuzione del processo attende in questo stato quando la contemporanea esecuzione e del processo a livello di applicazione è completamente occupata. Per ulteriori informazioni sull'accodamento e sulla concorrenza, vedere. <a href="#">Concorrenza dei job e accodamento per un'applicazione EMR Serverless</a>
Pending (In attesa)	Lo scheduler sta valutando l'esecuzione del processo per stabilire le priorità e pianificare l'esecuzione dell'applicazione.
Pianificato	EMR Serverless ha pianificato l'esecuzione del lavoro per l'applicazione e sta allocando risorse per eseguire il lavoro.
In esecuzione	EMR Serverless ha allocato le risorse di cui il lavoro inizialmente necessita e il lavoro è in esecuzione nell'applicazione. Nelle applicazioni Spark, ciò significa che il processo del driver Spark si trova nello stato <code>running</code> .
Failed (Non riuscito)	EMR Serverless non è riuscito a inviare il job run all'applicazione oppure è stato completato senza successo. <code>StateDetails</code> Per ulteriori informazioni su questo errore di lavoro, vedere.
Riuscito	L'esecuzione del processo è stata completata correttamente.

Stato	Descrizione
Annullamento in corso	L'CancelJobRun API ha richiesto l'annullamento dell'esecuzione del lavoro oppure l'esecuzione del processo è scaduta. EMR Serverless sta tentando di annullare il processo nell'applicazione e di rilasciare le risorse.
Annullato	L'esecuzione del processo è stata annullata correttamente e le risorse utilizzate sono state rilasciate.

## Annullamento dell'esecuzione di un processo EMR Serverless con periodo di grazia

Nei sistemi di elaborazione dati, le interruzioni improvvise possono portare a sprechi di risorse, operazioni incomplete e potenziali incongruenze nei dati. Amazon EMR Serverless consente di specificare un periodo di prova quando si annullano le esecuzioni dei job. Questa funzionalità consente di dedicare tempo alla pulizia e al completamento adeguati dei lavori in corso prima della cessazione del lavoro.

Quando si annulla l'esecuzione di un lavoro, è possibile specificare un periodo di grazia (in secondi) utilizzando il parametro `shutdownGracePeriodInSeconds` durante il quale il lavoro può eseguire le operazioni di pulizia prima dell'interruzione finale. Il comportamento e le impostazioni predefinite variano tra i processi in batch e quelli in streaming.

### Periodo di grazia per i lavori in batch

Per i lavori in batch, EMR Serverless consente di implementare operazioni di pulizia personalizzate che vengono eseguite durante il periodo di prova. È possibile registrare queste operazioni di pulizia come parte dell'hook di spegnimento della JVM nel codice dell'applicazione.

#### Comportamento predefinito

Il comportamento predefinito per l'arresto è quello di non avere un periodo di tolleranza. Consiste nelle due azioni seguenti:

- Cessazione immediata

- Le risorse vengono rilasciate immediatamente

## Opzioni di configurazione

È possibile specificare le impostazioni che comportano un arresto regolare:

- Intervallo valido per il periodo di tolleranza dello spegnimento: 15-1800 secondi (opzionale)
- Interruzione immediata (senza alcun periodo di grazia): 0 secondi

## Abilita lo spegnimento graduale

Per implementare Graceful Shutdown per i lavori in batch, procedi nel seguente modo:

1. Aggiungi shutdown hook nel codice dell'applicazione contenente una logica di spegnimento personalizzata.

### Example in Scala

```
import org.apache.hadoop.util.ShutdownHookManager

// Register shutdown hook with priority (second argument)
// Higher priority hooks run first
ShutdownHookManager.get().addShutdownHook(() => {
  logger.info("Performing cleanup operations...")
}, 100)
```

### Uso di [ShutdownHookManager](#)

### Example in PySpark

```
import atexit

def cleanup():
    # Your cleanup logic here
    print("Performing cleanup operations...")

# Register the cleanup function
atexit.register(cleanup)
```

2. Specificate un periodo di tolleranza quando annullate il lavoro per consentire l'esecuzione degli hook aggiunti sopra

## Esempio

```
# Default (immediate termination)
aws emr-serverless cancel-job-run \
  --application-id APPLICATION_ID \
  --job-run-id JOB_RUN_ID

# With 5-minute grace period
aws emr-serverless cancel-job-run \
  --application-id APPLICATION_ID \
  --job-run-id JOB_RUN_ID \
  --shutdown-grace-period-in-seconds 300
```

## Periodo di grazia per i lavori di streaming

In Spark Structured Streaming, dove i calcoli implicano la lettura o la scrittura su fonti di dati esterne, arresti improvvisi possono portare a risultati indesiderati. I processi di streaming elaborano i dati in microbatch e l'interruzione di queste operazioni a metà processo può comportare un'elaborazione duplicata nei tentativi successivi. Ciò si verifica quando l'ultimo checkpoint del microbatch precedente non è stato scritto, causando la rielaborazione degli stessi dati al riavvio del processo di streaming. Tale elaborazione duplicata non solo comporta uno spreco di risorse informatiche, ma può anche influire sulle operazioni aziendali, il che rende fondamentale evitare arresti improvvisi.

EMR Serverless offre un supporto integrato per un corretto spegnimento tramite un listener di query in streaming. Ciò garantisce il corretto completamento dei microbatch in corso prima della fine del lavoro. Il servizio gestisce automaticamente lo shutdown corretto tra i microbatch per le applicazioni di streaming, assicurando che il microbatch corrente completi l'elaborazione, i checkpoint siano scritti correttamente e che il contesto di streaming venga terminato in modo corretto senza dover importare nuovi dati durante il processo di spegnimento.

### Comportamento predefinito

- Il periodo di grazia di 120 secondi è abilitato per impostazione predefinita.
- Il listener di query in streaming integrato gestisce lo spegnimento automatico.

### Opzioni di configurazione

- Intervallo valido per il periodo di tolleranza dello spegnimento: 15-1800 secondi (opzionale)

- Terminazione immediata: 0 secondi

## Abilita Graceful Shutdown

Per implementare Graceful Shutdown per i lavori di streaming:

Specificate un periodo di tolleranza al momento dell'annullamento del lavoro per consentire il completamento del microbatch in corso.

### Esempio

```
# Default graceful shutdown (120 seconds)
aws emr-serverless cancel-job-run \
  --application-id APPLICATION_ID \
  --job-run-id JOB_RUN_ID

# Custom grace period (e.g. 300 seconds)
aws emr-serverless cancel-job-run \
  --application-id APPLICATION_ID \
  --job-run-id JOB_RUN_ID \
  --shutdown-grace-period-in-seconds 300

# Immediate Termination
aws emr-serverless cancel-job-run \
  --application-id APPLICATION_ID \
  --job-run-id JOB_RUN_ID \
  --shutdown-grace-period-in-seconds 0
```

## Aggiungi ganci di spegnimento personalizzati (opzionale)

Sebbene EMR Serverless gestisca per impostazione predefinita lo spegnimento graduale tramite il suo listener di query di streaming integrato, è possibile implementare opzionalmente una logica di spegnimento personalizzata per singole query di streaming. EMR Serverless registra il suo grazioso listener di spegnimento con priorità 60 (utilizzo). ShutdownHookManager Poiché gli hook con priorità più alta vengono eseguiti per primi, è possibile registrare le operazioni di pulizia personalizzate con una priorità superiore a 60 per garantire che vengano eseguite prima dell'inizio del processo di spegnimento di EMR Serverless.

Per aggiungere un hook personalizzato, fate riferimento al primo esempio di questo argomento che mostra come aggiungere un hook di spegnimento nel codice dell'applicazione. Qui, 100 è la priorità, che è maggiore di 60. Quindi tale gancio di spegnimento verrà eseguito per primo.

### Note

Gli hook di spegnimento personalizzati sono opzionali e non necessari per una corretta funzionalità di spegnimento, che viene gestita automaticamente da EMR Serverless.

## Costi del periodo di grazia e durata del Batch

Se viene utilizzato il valore predefinito per il periodo di grazia (120 secondi):

- Se la durata del batch è inferiore a 120 secondi, ti verrà addebitato solo il tempo effettivo necessario per completare il batch.
- Se la durata del batch supera i 120 secondi, ti verrà addebitato il periodo di prova massimo (120 secondi), ma la query potrebbe non essere chiusa correttamente poiché verrà interrotta forzatamente.

Per ottimizzare i costi e garantire un arresto corretto:

- Per lotti di durata superiore a 120 secondi: valuta la possibilità di aumentare il periodo di tolleranza in modo che corrisponda alla durata del batch
- Per lotti di durata inferiore a 120 secondi: non è necessario modificare il periodo di prova, in quanto ti verrà addebitato solo il tempo di elaborazione effettivo

## Considerazioni

### Comportamento del periodo di grazia

- Il periodo di grazia prevede il completamento degli hook di spegnimento registrati.
- Il lavoro termina non appena il gancio di spegnimento termina, anche se è ben prima del periodo di grazia.
- Se le operazioni di pulizia superano il periodo di prova, il lavoro verrà interrotto con forza.

### Comportamento del servizio

- L'arresto durante il periodo di prova è disponibile solo per i lavori nello stato RUNNING.
- Le successive richieste di annullamento durante lo stato CANCELLING vengono ignorate.

- Se EMR Serverless non riesce ad avviare l'arresto del periodo di prova a causa di errori interni del servizio:
  - Il servizio riproverà per un massimo di 2 minuti.
  - Se i nuovi tentativi non hanno esito positivo, il processo verrà interrotto con forza.

## Fatturazione

I lavori vengono fatturati in base alle risorse di elaborazione utilizzate fino alla chiusura completa del lavoro, incluso il tempo impiegato durante il periodo di prova.

## Esecuzione di processi dalla console EMR Studio

È possibile inviare esecuzioni di job alle applicazioni EMR Serverless e visualizzarle dalla console EMR Studio. Per creare o accedere alla tua applicazione EMR Serverless sulla console EMR Studio, segui le istruzioni in [Guida introduttiva](#) dalla console.

## Invio di un processo

Nella pagina Invia lavoro, è possibile inviare un lavoro a un'applicazione EMR Serverless come segue.

### Spark

1. Nel campo Nome, inserisci un nome per l'esecuzione del lavoro.
2. Nel campo Runtime role, inserisci il nome del ruolo IAM che l'applicazione EMR Serverless può assumere per l'esecuzione del job. Per ulteriori informazioni sui ruoli di runtime, consulta [Ruoli Job Runtime per Amazon EMR Serverless](#)
3. Nel campo Posizione dello script, inserisci la posizione Amazon S3 per lo script o il JAR che desideri eseguire. Per i lavori Spark, lo script può essere un file Python `.py` () o un file JAR `.jar` ().
4. Se la posizione dello script è un file JAR, inserisci il nome della classe che è il punto di ingresso per il lavoro nel campo Classe principale.
5. (Facoltativo) Inserite i valori per i campi rimanenti.
  - Argomenti dello script: inserisci gli argomenti che desideri passare allo script JAR o Python principale. Il codice legge questi parametri. Separare ogni argomento dell'array con una virgola.

- Proprietà Spark: espandi la sezione delle proprietà Spark e inserisci qualsiasi parametro di configurazione Spark in questo campo.

 Note

Se specifichi le dimensioni del driver e dell'executor Spark, devi tenere conto del sovraccarico di memoria. Specificate i valori del sovraccarico di memoria nelle proprietà `spark.driver.memoryOverhead` e `spark.executor.memoryOverhead`. Il sovraccarico di memoria ha un valore predefinito del 10% della memoria del contenitore, con un minimo di 384 MB. La memoria dell'esecutore e il sovraccarico di memoria insieme non possono superare la memoria di lavoro. Ad esempio, il massimo `spark.executor.memory` per un worker da 30 GB deve essere di 27 GB.

- Job configuration: specifica qualsiasi configurazione del lavoro in questo campo. È possibile utilizzare queste configurazioni di lavoro per sovrascrivere le configurazioni predefinite per le applicazioni.
- Impostazioni aggiuntive: attiva o disattiva il AWS Glue Data Catalog come metastore e modifica le impostazioni del registro dell'applicazione. Per ulteriori informazioni sulle configurazioni dei metastore, consulta [Configurazione Metastore per EMR Serverless](#). Per ulteriori informazioni sulle opzioni di registrazione delle applicazioni, consulta [Archiviazione dei registri](#).
- Tag: assegna tag personalizzati all'applicazione.

6. Seleziona Submit job (Invia processo).

## Hive

1. Nel campo Nome, inserisci un nome per l'esecuzione del job.
  2. Nel campo Runtime role, inserisci il nome del ruolo IAM che l'applicazione EMR Serverless può assumere per l'esecuzione del job.
  3. Nel campo Posizione dello script, inserisci la posizione Amazon S3 per lo script o il JAR che desideri eseguire. Per i lavori Hive, lo script deve essere un file Hive (`hive`).sql.
  4. (Facoltativo) Immettete i valori per i campi rimanenti.
- Posizione dello script di inizializzazione: immettere la posizione dello script che inizializza le tabelle prima dell'esecuzione dello script Hive.

- **Proprietà Hive:** espandi la sezione delle proprietà di Hive e inserisci qualsiasi parametro di configurazione Hive in questo campo.
- **Configurazione del lavoro:** specifica qualsiasi configurazione del lavoro. È possibile utilizzare queste configurazioni di lavoro per sovrascrivere le configurazioni predefinite per le applicazioni. Per i lavori Hive, `hive.exec.scratchdir` e `hive.metastore.warehouse.dir` sono proprietà obbligatorie nella configurazione. `hive-site`

```
{
  "applicationConfiguration": [
    {
      "classification": "hive-site",
      "configurations": [],
      "properties": {
        "hive.exec.scratchdir": "s3://DOC-EXAMPLE_BUCKET/hive/scratch",
        "hive.metastore.warehouse.dir": "s3://DOC-EXAMPLE_BUCKET/hive/warehouse"
      }
    }
  ],
  "monitoringConfiguration": {}
}
```

- **Impostazioni aggiuntive:** attiva o disattiva il AWS Glue Data Catalog come metastore e modifica le impostazioni del registro dell'applicazione. Per ulteriori informazioni sulle configurazioni dei metastore, consulta. [Configurazione Metastore per EMR Serverless](#) Per ulteriori informazioni sulle opzioni di registrazione delle applicazioni, consulta. [Archiviazione dei registri](#)
- **Tag:** assegna qualsiasi tag personalizzato all'applicazione.

5. Seleziona Submit job (Invia processo).

## Visualizza esecuzioni dei processi

Dalla scheda Job run nella pagina Dettagli di un'applicazione, è possibile visualizzare le esecuzioni dei job ed eseguire le seguenti azioni per le esecuzioni di job.

Annula processo: per annullare l'esecuzione di un processo in questo RUNNING stato, scegliete questa opzione. Per ulteriori informazioni sulle transizioni Job Run, consulta [Stati delle esecuzioni di processi](#).

Clona processo: per clonare un processo precedente eseguito e inviarlo nuovamente, scegli questa opzione.

## Esecuzione di lavori da AWS CLI

È possibile creare, descrivere ed eliminare singoli lavori su AWS CLI. Puoi anche elencare tutti i tuoi lavori per visualizzarli a colpo d'occhio.

Per inviare un nuovo lavoro, usa `start-job-run`. Fornisci l'ID dell'applicazione che desideri eseguire, insieme alle proprietà specifiche del lavoro. Per gli esempi di Spark, consulta [Utilizzo delle configurazioni Spark quando si eseguono job EMR Serverless](#). Per gli esempi di Hive, vedi [Utilizzo delle configurazioni Hive quando si eseguono job EMR Serverless](#). Questo comando restituisce il tuo `application-id`, ARN e `new-job-id`.

Ogni esecuzione di un processo ha una durata di timeout impostata. Se l'esecuzione del job supera questa durata, EMR Serverless la annulla automaticamente. Il timeout predefinito è di 12 ore. Quando si avvia l'esecuzione del job, è possibile configurare questa impostazione di timeout su un valore che soddisfi i requisiti del job. Configura il valore con la `executionTimeoutMinutes` proprietà.

```
aws emr-serverless start-job-run \  
  --application-id application-id \  
  --execution-role-arn job-role-arn \  
  --execution-timeout-minutes 15 \  
  --job-driver '{  
    "hive": {  
      "query": "s3://amzn-s3-demo-bucket/scripts/create_table.sql",  
      "parameters": "--hiveconf hive.exec.scratchdir=s3://amzn-s3-demo-bucket/  
hive/scratch --hiveconf hive.metastore.warehouse.dir=s3://amzn-s3-demo-bucket/hive/  
warehouse"  
    }  
  }' \  
  --configuration-overrides '{  
    "applicationConfiguration": [{  
      "classification": "hive-site",  
      "properties": {  
        "hive.client.cores": "2",
```

```
        "hive.client.memory": "4GIB"
    }
  ]}
}'
```

Per descrivere un lavoro, usa `aws emr-serverless get-job-run`. Questo comando restituisce le configurazioni specifiche del job e la capacità impostata per il nuovo job.

```
aws emr-serverless get-job-run \
--job-run-id job-id \
--application-id application-id
```

Per elencare i tuoi lavori, usa `aws emr-serverless list-job-runs`. Questo comando restituisce un insieme abbreviato di proprietà che include il tipo di lavoro, lo stato e altri attributi di alto livello. Se non desideri visualizzare tutti i tuoi lavori, puoi specificare il numero massimo di lavori che desideri visualizzare, fino a 50. L'esempio seguente specifica che desideri vedere i tuoi ultimi due job eseguiti.

```
aws emr-serverless list-job-runs \
--max-results 2 \
--application-id application-id
```

Per annullare un lavoro, usa `aws emr-serverless cancel-job-run`. Fornisci il nome `application-id` e il nome `job-id` del lavoro che desideri annullare.

```
aws emr-serverless cancel-job-run \
--job-run-id job-id \
--application-id application-id
```

Per ulteriori informazioni su come eseguire i job da AWS CLI, vedere [EMR Serverless API Reference](#).

## Utilizzo di dischi ottimizzati per la riproduzione casuale

Con le versioni 7.1.0 e successive di Amazon EMR, puoi utilizzare dischi ottimizzati per lo shuffle quando esegui job Apache Spark o Hive per migliorare le prestazioni per carichi di lavoro a uso intensivo di I/O. Rispetto ai dischi standard, i dischi ottimizzati per lo shuffle offrono IOPS (operazioni di I/O al secondo) più elevati per uno spostamento più rapido dei dati e una latenza ridotta durante le operazioni di shuffle. I dischi ottimizzati per lo shuffle consentono di collegare dischi di dimensioni fino a 2 TB per lavoratore, in modo da poter configurare la capacità appropriata per i requisiti del carico di lavoro.

## Vantaggi principali

I dischi ottimizzati per lo shuffle offrono i seguenti vantaggi.

- Prestazioni IOPS elevate: i dischi ottimizzati per lo shuffle offrono IOPS più elevati rispetto ai dischi standard, il che consente uno shuffling dei dati più efficiente e rapido durante i job Spark e Hive e altri carichi di lavoro che richiedono un uso intensivo dello shuffle.
- Dimensioni del disco più grandi: i dischi ottimizzati per Shuffle supportano dimensioni del disco da 20 GB a 2 TB per lavoratore, in modo da poter scegliere la capacità appropriata in base ai carichi di lavoro.

## Nozioni di base

Consulta i passaggi seguenti per utilizzare dischi ottimizzati per lo shuffle nei flussi di lavoro.

### Spark

1. Creare un'applicazione EMR Serverless release 7.1.0 con il seguente comando.

```
aws emr-serverless create-application \  
  --type "SPARK" \  
  --name my-application-name \  
  --release-label emr-7.1.0 \  
  --region <AWS_REGION>
```

2. Configura il tuo job Spark per includere i parametri `spark.emr-serverless.driver.disk.type` e/o per l'esecuzione con dischi ottimizzati `spark.emr-serverless.executor.disk.type` per lo shuffle. Puoi utilizzare uno o entrambi i parametri, a seconda del caso d'uso.

```
aws emr-serverless start-job-run \  
  --application-id application-id \  
  --execution-role-arn job-role-arn \  
  --job-driver '{  
    "sparkSubmit": {  
      "entryPoint": "/usr/lib/spark/examples/jars/spark-examples.jar",  
      "entryPointArguments": ["1"],  
      "sparkSubmitParameters": "--class org.apache.spark.examples.SparkPi  
      --conf spark.executor.cores=4  
      --conf spark.executor.memory=20g
```

```

        --conf spark.driver.cores=4
        --conf spark.driver.memory=8g
        --conf spark.executor.instances=1
        --conf spark.emr-serverless.executor.disk.type=shuffle_optimized"
    }
}'

```

Per maggiori informazioni, consulta [Spark job properties](#).

## Hive

1. Creare un'applicazione EMR Serverless release 7.1.0 con il seguente comando.

```

aws emr-serverless create-application \
  --type "HIVE" \
  --name my-application-name \
  --release-label emr-7.1.0 \
  --region <AWS_REGION>

```

2. Configura il tuo job Hive per includere i parametri `hive.driver.disk.type` e/o per l'esecuzione con dischi ottimizzati `hive.tez.disk.type` per lo shuffle. È possibile utilizzare uno o entrambi i parametri, a seconda del caso d'uso.

```

aws emr-serverless start-job-run \
  --application-id application-id \
  --execution-role-arn job-role-arn \
  --job-driver '{
    "hive": {
      "query": "s3://<DOC-EXAMPLE-BUCKET>/emr-serverless-hive/query/hive-
query.q1",
      "parameters": "--hiveconf hive.log.explain.output=false"
    }
  }' \
  --configuration-overrides '{
    "applicationConfiguration": [{
      "classification": "hive-site",
      "properties": {
        "hive.exec.scratchdir": "s3://<DOC-EXAMPLE-BUCKET>/emr-
serverless-hive/hive/scratch",
        "hive.metastore.warehouse.dir": "s3://<DOC-EXAMPLE-BUCKET>/emr-
serverless-hive/hive/warehouse",
        "hive.driver.cores": "2",

```

```

        "hive.driver.memory": "4g",
        "hive.tez.container.size": "4096",
        "hive.tez.cpu.vcores": "1",
        "hive.driver.disk.type": "shuffle_optimized",
        "hive.tez.disk.type": "shuffle_optimized"
    }
  ]
}'

```

Per ulteriori informazioni, consulta [Hive job properties](#).

## Configurazione di un'applicazione con capacità preinizializzata

Guarda i seguenti esempi per creare applicazioni basate sulla release 7.1.0 di Amazon EMR. Queste applicazioni hanno le seguenti proprietà:

- 5 driver Spark preinizializzati, ciascuno con 2 vCPU, 4 GB di memoria e 50 GB di disco ottimizzato per lo shuffle.
- 50 executor preinizializzati, ciascuno con 4 vCPU, 8 GB di memoria e 500 GB di disco ottimizzato per lo shuffle.

Quando questa applicazione esegue i job Spark, utilizza innanzitutto i worker preinizializzati e quindi ridimensiona i worker su richiesta fino alla capacità massima di 400 vCPU e 1024 GB di memoria.

Facoltativamente, puoi omettere la capacità per uno o. DRIVER EXECUTOR

## Spark

```

aws emr-serverless create-application \
  --type "SPARK" \
  --name <my-application-name> \
  --release-label emr-7.1.0 \
  --initial-capacity '{
    "DRIVER": {
      "workerCount": 5,
      "workerConfiguration": {
        "cpu": "2vCPU",
        "memory": "4GB",
        "disk": "50GB",
        "diskType": "SHUFFLE_OPTIMIZED"
      }
    }
  }

```

```

    },
    "EXECUTOR": {
      "workerCount": 50,
      "workerConfiguration": {
        "cpu": "4vCPU",
        "memory": "8GB",
        "disk": "500GB",
        "diskType": "SHUFFLE_OPTIMIZED"
      }
    }
  } \
  --maximum-capacity '{
    "cpu": "400vCPU",
    "memory": "1024GB"
  }'

```

## Hive

```

aws emr-serverless create-application \
  --type "HIVE" \
  --name <my-application-name> \
  --release-label emr-7.1.0 \
  --initial-capacity '{
    "DRIVER": {
      "workerCount": 5,
      "workerConfiguration": {
        "cpu": "2vCPU",
        "memory": "4GB",
        "disk": "50GB",
        "diskType": "SHUFFLE_OPTIMIZED"
      }
    },
    "EXECUTOR": {
      "workerCount": 50,
      "workerConfiguration": {
        "cpu": "4vCPU",
        "memory": "8GB",
        "disk": "500GB",
        "diskType": "SHUFFLE_OPTIMIZED"
      }
    }
  } \
  --maximum-capacity '{

```

```
"cpu": "400vCPU",  
"memory": "1024GB"  
'
```

## Lavori di streaming per l'elaborazione di dati in streaming continuo

Un processo di streaming in EMR Serverless è una modalità di lavoro che consente di analizzare ed elaborare i dati di streaming quasi in tempo reale. Questi processi di lunga durata analizzano i dati in streaming ed elaborano continuamente i risultati man mano che i dati arrivano. I lavori in streaming sono più adatti per attività che richiedono l'elaborazione dei dati in tempo reale, come analisi quasi in tempo reale, rilevamento delle frodi e motori di suggerimenti. I job di streaming EMR Serverless offrono ottimizzazioni, come la resilienza integrata dei processi, il monitoraggio in tempo reale, la gestione avanzata dei log e l'integrazione con i connettori di streaming.

Di seguito sono riportati alcuni casi d'uso con i lavori di streaming:

- **Analisi quasi in tempo reale:** i lavori di streaming in Amazon EMR Serverless ti consentono di elaborare i dati in streaming quasi in tempo reale, in modo da poter eseguire analisi in tempo reale su flussi di dati continui, come dati di log, dati di sensori o dati clickstream per ricavare informazioni e prendere decisioni tempestive sulla base delle informazioni più recenti.
- **Rilevamento delle frodi:** puoi utilizzare i processi di streaming per eseguire il rilevamento delle frodi quasi in tempo reale nelle transazioni finanziarie, nelle operazioni con carte di credito o nelle attività online quando analizzi i flussi di dati e identifichi modelli o anomalie sospetti man mano che si verificano.
- **Motori di raccomandazione:** i processi di streaming possono elaborare i dati sulle attività degli utenti e aggiornare i modelli di suggerimenti. In questo modo si aprono possibilità di consigli personalizzati e in tempo reale basati su comportamenti e preferenze.
- **Analisi dei social media:** le offerte di lavoro in streaming possono elaborare i dati dei social media, come tweet, commenti e post, in modo che le organizzazioni possano monitorare le tendenze, l'analisi del sentiment e gestire la reputazione del marchio quasi in tempo reale.
- **Analisi dell'Internet of Things (IoT):** i job in streaming possono gestire e analizzare flussi di dati ad alta velocità provenienti da dispositivi IoT, sensori e macchinari connessi, in modo da poter eseguire il rilevamento delle anomalie, la manutenzione predittiva e altri casi d'uso di analisi IoT.
- **Analisi clickstream:** i job di streaming possono elaborare e analizzare i dati clickstream provenienti da siti Web o applicazioni mobili. Le aziende che utilizzano tali dati possono eseguire analisi

per saperne di più sul comportamento degli utenti, personalizzare le esperienze degli utenti e ottimizzare le campagne di marketing.

- Monitoraggio e analisi dei log: i processi di streaming possono anche elaborare i dati di registro da server, applicazioni e dispositivi di rete. Ciò consente il rilevamento delle anomalie, la risoluzione dei problemi e lo stato e le prestazioni del sistema.

## Principali vantaggi

I lavori di streaming in EMR Serverless forniscono automaticamente la resilienza del lavoro, che è una combinazione dei seguenti fattori:

- Riprova automatica: EMR Serverless riprova automaticamente tutti i lavori che non sono riusciti senza alcun input manuale da parte dell'utente.
- Resilienza della zona di disponibilità (AZ): EMR Serverless commuta automaticamente i job di streaming su una zona di disponibilità integra se la zona di disponibilità originale presenta problemi.
- Gestione dei registri:
  - Rotazione dei log: per una gestione più efficiente dello storage su disco, EMR Serverless ruota periodicamente i log per lunghi lavori di streaming. In questo modo si evita l'accumulo di log che potrebbe consumare tutto lo spazio su disco.
  - Compattazione dei log: consente di gestire e ottimizzare in modo efficiente i file di registro in modalità di persistenza gestita. La compattazione migliora anche l'esperienza di debug quando si utilizza il server di cronologia Spark gestito.

## Fonti di dati e data sink supportati

EMR Serverless funziona con una serie di sorgenti di dati di input e data sink di output:

- Fonti di dati di input supportate: Amazon Kinesis Data Streams, Amazon Managed Streaming for Apache Kafka e cluster Apache Kafka autogestiti. Per impostazione predefinita, le versioni di Amazon EMR 7.1.0 e successive includono il connettore [Amazon Kinesis Data Streams](#), quindi non è necessario creare o scaricare pacchetti aggiuntivi.
- Dissipatori di dati di output supportati: tabelle AWS Glue Data Catalog, Amazon S3, Amazon Redshift, MySQL, PostgreSQL Oracle, Oracle, Microsoft SQL, Apache Iceberg, Delta Lake e Apache Hudi.

## Considerazioni e limitazioni

Quando utilizzi i lavori di streaming, tieni presente le seguenti considerazioni e limitazioni.

- I lavori di streaming sono supportati con le [versioni 7.1.0 e successive di Amazon EMR](#).
- EMR Serverless prevede che i processi di streaming durino a lungo, quindi non è possibile impostare il timeout di esecuzione per limitare il tempo di esecuzione del lavoro.
- [I job di streaming sono compatibili solo con il motore Spark, che si basa sul framework di streaming strutturato](#).
- EMR Serverless riprova a tempo indeterminato i processi di streaming e non è possibile personalizzare il numero massimo di tentativi. La prevenzione degli errori viene inclusa automaticamente per interrompere il nuovo tentativo di lavoro se il numero di tentativi falliti supera una soglia impostata su una finestra oraria. La soglia predefinita è di cinque tentativi falliti nell'arco di un'ora. È possibile configurare questa soglia in modo che sia compresa tra 1 e 10 tentativi. Per ulteriori informazioni, consulta [Job resiliency](#).
- I processi di streaming hanno punti di controllo per salvare lo stato di esecuzione e l'avanzamento, quindi EMR Serverless può riprendere il processo di streaming dal checkpoint più recente. Per ulteriori informazioni, consulta [Recovery from failures with Checkpointing](#) nella documentazione di Apache Spark.

## Guida introduttiva allo streaming di lavori

Consulta le seguenti istruzioni per scoprire come iniziare a utilizzare i lavori di streaming.

1. Segui [Guida introduttiva ad Amazon EMR Serverless per creare un'applicazione](#). Tieni presente che la tua applicazione deve eseguire [Amazon EMR versione 7.1.0](#) o successiva.
2. Una volta che l'applicazione è pronta, imposta il mode parametro per STREAMING inviare un lavoro in streaming, simile all'esempio seguente AWS CLI .

```
aws emr-serverless start-job-run \  
--application-id <APPLICATION_ID> \  
--execution-role-arn <JOB_EXECUTION_ROLE> \  
--mode 'STREAMING' \  
--job-driver '{  
  "sparkSubmit": {  
    "entryPoint": "s3://<streaming script>",  
    "entryPointArguments": ["s3://<DOC-EXAMPLE-BUCKET-OUTPUT>/output"],  
    "sparkSubmitParameters": "--conf spark.executor.cores=4
```

```

        --conf spark.executor.memory=16g
        --conf spark.driver.cores=4
        --conf spark.driver.memory=16g
        --conf spark.executor.instances=3"
    }
}'

```

## Connettori di streaming supportati

I connettori di streaming facilitano la lettura dei dati da una fonte di streaming e possono anche scrivere dati su un sink di streaming.

I seguenti sono i connettori di streaming supportati:

### Connettore Amazon Kinesis Data Streams

Il connettore [Amazon Kinesis Data Streams](#) per Apache Spark consente di creare applicazioni e pipeline di streaming che consumano e scrivono dati da Amazon Kinesis Data Streams. Il connettore supporta un maggiore consumo di fan-out con una velocità di trasmissione dedicata di lettura fino a 2 MB/secondo per shard. Per impostazione predefinita, Amazon EMR Serverless 7.1.0 e versioni successive includono il connettore, quindi non è necessario creare o scaricare pacchetti aggiuntivi. [Per ulteriori informazioni sul connettore, consulta la spark-sql-kinesis-connector pagina su. GitHub](#)

Di seguito è riportato un esempio di come avviare l'esecuzione di un processo con la dipendenza del connettore Kinesis Data Streams.

```

aws emr-serverless start-job-run \
--application-id <APPLICATION_ID> \
--execution-role-arn <JOB_EXECUTION_ROLE> \
--mode 'STREAMING' \
--job-driver '{
  "sparkSubmit": {
    "entryPoint": "s3://<Kinesis-streaming-script>",
    "entryPointArguments": ["s3://<DOC-EXAMPLE-BUCKET-OUTPUT>/output"],
    "sparkSubmitParameters": "--conf spark.executor.cores=4
      --conf spark.executor.memory=16g
      --conf spark.driver.cores=4
      --conf spark.driver.memory=16g
      --conf spark.executor.instances=3
      --jars /usr/share/aws/kinesis/spark-sql-kinesis/lib/spark-streaming-
sql-kinesis-connector.jar"
  }
}'

```

```
}
}'
```

Per connettersi a Kinesis Data Streams, è necessario configurare l'applicazione EMR Serverless con accesso VPC e utilizzare un endpoint VPC per consentire l'accesso privato oppure utilizzare un gateway NAT per ottenere l'accesso pubblico. Per ulteriori informazioni, consulta [Configurazione dell'accesso VPC](#). È inoltre necessario assicurarsi che il ruolo di job runtime disponga delle autorizzazioni di lettura e scrittura necessarie per accedere ai flussi di dati richiesti. Per ulteriori informazioni su come configurare un ruolo Job runtime, consulta [Job runtime roles for Amazon EMR Serverless](#). [Per un elenco completo di tutte le autorizzazioni richieste, consulta la spark-sql-kinesis-connector pagina su. GitHub](#)

## Connettore Apache Kafka

Il connettore Apache Kafka per lo streaming strutturato di Spark è un connettore open source della community Spark ed è disponibile in un repository Maven. Questo connettore consente alle applicazioni di streaming strutturato Spark di leggere e scrivere dati su Apache Kafka e Amazon Managed Streaming for Apache Kafka autogestiti. Per ulteriori informazioni sul connettore, consulta la [Structured Streaming + Kafka Integration Guide](#) nella documentazione di Apache Spark.

L'esempio seguente mostra come includere il connettore Kafka nella richiesta di esecuzione del lavoro.

```
aws emr-serverless start-job-run \
--application-id <APPLICATION_ID> \
--execution-role-arn <JOB_EXECUTION_ROLE> \
--mode 'STREAMING' \
--job-driver '{
  "sparkSubmit": {
    "entryPoint": "s3://<Kafka-streaming-script>",
    "entryPointArguments": ["s3://<DOC-EXAMPLE-BUCKET-OUTPUT>/output"],
    "sparkSubmitParameters": "--conf spark.executor.cores=4
      --conf spark.executor.memory=16g
      --conf spark.driver.cores=4
      --conf spark.driver.memory=16g
      --conf spark.executor.instances=3
      --packages org.apache.spark:spark-sql-
kafka-0-10_2.12:<KAFKA_CONNECTOR_VERSION>"
  }
}'
```

La versione del connettore Apache Kafka dipende dalla versione EMR Serverless in uso e dalla versione Spark corrispondente. [Per trovare la versione corretta di Kafka, consulta la Guida all'integrazione di Structured Streaming + Kafka.](#)

Per utilizzare Amazon Managed Streaming for Apache Kafka con l'autenticazione IAM, devi includere un'altra dipendenza per consentire al connettore Kafka di connettersi ad Amazon MSK con IAM. [Per ulteriori informazioni, consulta il repository su. `aws-msk-iam-auth` GitHub](#) È inoltre necessario assicurarsi che il ruolo di job runtime disponga delle autorizzazioni IAM necessarie. L'esempio seguente mostra come utilizzare il connettore con l'autenticazione IAM.

```
aws emr-serverless start-job-run \
--application-id <APPLICATION_ID> \
--execution-role-arn <JOB_EXECUTION_ROLE> \
--mode 'STREAMING' \
--job-driver '{
  "sparkSubmit": {
    "entryPoint": "s3://<Kafka-streaming-script>",
    "entryPointArguments": ["s3://<DOC-EXAMPLE-BUCKET-OUTPUT>/output"],
    "sparkSubmitParameters": "--conf spark.executor.cores=4
      --conf spark.executor.memory=16g
      --conf spark.driver.cores=4
      --conf spark.driver.memory=16g
      --conf spark.executor.instances=3
      --packages org.apache.spark:spark-sql-
kafka-0-10_2.12:<KAFKA_CONNECTOR_VERSION>,software.amazon.msk:aws-msk-iam-
auth:<MSK_IAM_LIB_VERSION>"
  }
}'
```

Per utilizzare il connettore Kafka e la libreria di autenticazione IAM di Amazon MSK, è necessario configurare l'applicazione EMR Serverless con accesso VPC. Le sottoreti devono avere accesso a Internet e utilizzare un gateway NAT per accedere alle dipendenze Maven. Per ulteriori informazioni, consulta [Configurazione dell'accesso VPC](#). Le sottoreti devono disporre di connettività di rete per accedere al cluster Kafka. Questo vale indipendentemente dal fatto che il cluster Kafka sia autogestito o che utilizzi Amazon Managed Streaming for Apache Kafka.

## Gestione dei registri dei lavori in streaming

I job di streaming supportano la rotazione dei log per i log delle applicazioni Spark e i log degli eventi e la compattazione dei log per i log degli eventi Spark. Questo ti aiuta a gestire le tue risorse in modo efficace.

## Rotazione del registro

I lavori di streaming supportano la rotazione dei log per i log delle applicazioni Spark e i log degli eventi. La rotazione dei log impedisce che i lavori di streaming lunghi generino file di registro di grandi dimensioni che potrebbero occupare tutto lo spazio disponibile su disco. La rotazione dei log consente di risparmiare spazio su disco e previene gli errori dei lavori dovuti allo scarso spazio su disco. Per ulteriori informazioni, vedere [Rotazione dei log](#).

## Compattazione dei tronchi

I lavori di streaming supportano anche la compattazione dei log dei log degli eventi Spark ogni volta che è disponibile la registrazione gestita. [Per maggiori dettagli sulla registrazione gestita, consulta Registrazione con storage gestito](#). I processi di streaming possono durare a lungo e la quantità di dati sugli eventi può accumularsi nel tempo e aumentare notevolmente le dimensioni dei file di registro. Lo Spark History Server legge e carica questi eventi in memoria per l'interfaccia utente dell'applicazione Spark. Questo processo può causare latenze e costi elevati, soprattutto se i log degli eventi archiviati in Amazon S3 sono molto grandi.

La compattazione dei log riduce le dimensioni del registro degli eventi, quindi lo Spark History Server non deve caricare più di 1 GB di registri degli eventi in qualsiasi momento. Per ulteriori informazioni, consulta [Monitoraggio e strumentazione](#) nella documentazione di Apache Spark.

# Utilizzo delle configurazioni Spark quando si eseguono job EMR Serverless

È possibile eseguire i job Spark su un'applicazione con il `type` parametro impostato su `SPARK`. I lavori devono essere compatibili con la versione Spark compatibile con la versione di rilascio di Amazon EMR. Ad esempio, quando esegui job con Amazon EMR release 6.6.0, il job deve essere compatibile con Apache Spark 3.2.0. Per informazioni sulle versioni delle applicazioni per ogni versione, consulta [Versioni di rilascio di Amazon EMR Serverless](#)

## Parametri del job Spark

Quando usi l'[StartJobRunAPI](#) per eseguire un job Spark, puoi specificare i seguenti parametri.

### Parametri obbligatori

- [Ruolo di runtime del job Spark](#)
- [Parametro Spark Job Driver](#)

- [La configurazione di Spark sovrascrive il parametro](#)
- [Ottimizzazione dinamica dell'allocazione delle risorse Spark](#)

## Ruolo di runtime del job Spark

**executionRoleArn** Da utilizzare per specificare l'ARN per il ruolo IAM utilizzato dall'applicazione per eseguire i job Spark. Questo ruolo deve contenere le seguenti autorizzazioni:

- Leggi dai bucket S3 o da altre fonti di dati in cui risiedono i tuoi dati
- Leggi dai bucket o dai prefissi S3 in cui risiede lo script o il file JAR PySpark
- Scrivi nei bucket S3 dove intendi scrivere l'output finale
- Scrivi i log in un bucket o prefisso S3 che specifica `S3MonitoringConfiguration`
- Accesso alle chiavi KMS se utilizzi le chiavi KMS per crittografare i dati nel tuo bucket S3
- Accesso al AWS Glue Data Catalog se usi SparkSQL

Se il tuo job Spark legge o scrive dati da o verso altre fonti di dati, specifica le autorizzazioni appropriate in questo ruolo IAM. Se non fornisci queste autorizzazioni al ruolo IAM, il lavoro potrebbe fallire. Per ulteriori informazioni, consulta [Ruoli Job Runtime per Amazon EMR Serverless](#) e [Archiviazione dei registri](#).

## Parametro Spark Job Driver

**jobDriver** Utilizzato per fornire input al lavoro. Il parametro job driver accetta solo un valore per il tipo di processo che si desidera eseguire. Per un job Spark, il valore del parametro è `sparkSubmit`. Puoi usare questo tipo di lavoro per eseguire Scala, Java PySpark, SparkR e qualsiasi altro lavoro supportato tramite Spark submit. I job Spark hanno i seguenti parametri:

- **sparkSubmitParameters**— Questi sono i parametri Spark aggiuntivi che desideri inviare al job. Utilizzate questo parametro per sovrascrivere le proprietà predefinite di Spark, come la memoria del driver o il numero di esecutori, come quelle definite negli argomenti `or. --conf --class`
- **entryPointArguments**— Questa è una serie di argomenti che vuoi passare al tuo file JAR o Python principale. Dovrai gestire la lettura di questi parametri con il tuo codice di `entrypoint`. Separare ogni argomento dell'array con una virgola.
- **entryPoint**— Questo è il riferimento in Amazon S3 al file JAR o Python principale che desideri eseguire. Se stai usando un JAR Scala o Java, specifica la classe di ingresso principale nell'argomento `sparkSubmitParameters` using the `--class`.

Per ulteriori informazioni, consulta [Avvio delle applicazioni con spark-submit](#).

## La configurazione di Spark sovrascrive il parametro

Utilizzato **configurationOverrides** per sovrascrivere le proprietà di configurazione a livello di monitoraggio e a livello di applicazione. Questo parametro accetta un oggetto JSON con i due campi seguenti:

- **monitoringConfiguration**- Utilizza questo campo per specificare l'URL di Amazon S3 (`s3MonitoringConfiguration`) in cui desideri che il job EMR Serverless memorizzi i log del tuo job Spark. Assicurati di aver creato questo bucket con lo stesso Account AWS che ospita l'applicazione e nello stesso luogo in cui è in esecuzione il processo. Regione AWS
- **applicationConfiguration**— Per sovrascrivere le configurazioni predefinite per le applicazioni, puoi fornire un oggetto di configurazione in questo campo. Puoi utilizzare una sintassi abbreviata per fornire la configurazione oppure fare riferimento all'oggetto di configurazione in un file JSON. Gli oggetti di configurazione sono composti da una classificazione, proprietà e configurazioni nidificate opzionali. Le proprietà sono costituite dalle impostazioni che desideri ignorare in un dato file. Puoi specificare diverse classificazioni per più applicazioni in un singolo oggetto JSON.

### Note

Le classificazioni di configurazione disponibili variano in base alla specifica release EMR Serverless. Ad esempio, le classificazioni per Log4j personalizzate `spark-executor-log4j2` sono disponibili solo con le versioni `spark-driver-log4j2 6.8.0` e successive.

Se usi la stessa configurazione in un'applicazione override e nei parametri di invio di Spark, i parametri di invio di Spark hanno la priorità. Le configurazioni hanno la priorità seguente, dalla più alta alla più bassa:

- Configurazione fornita da EMR Serverless al momento della creazione. `SparkSession`
- Configurazione fornita come parte dell'`sparkSubmitParameters--conf` argomento.
- La configurazione fornita come parte dell'applicazione ha la precedenza quando si avvia un lavoro.
- La configurazione fornita come parte della configurazione `runtimeConfiguration` quando si crea un'applicazione.
- Configurazioni ottimizzate utilizzate da Amazon EMR per il rilascio.

- Configurazioni open source predefinite per l'applicazione.

Per ulteriori informazioni sulla dichiarazione delle configurazioni a livello di applicazione e sulla sovrascrittura delle configurazioni durante l'esecuzione del processo, consulta [Configurazione predefinita dell'applicazione per EMR Serverless](#)

## Ottimizzazione dinamica dell'allocazione delle risorse Spark

Utilizzalo `dynamicAllocationOptimization` per ottimizzare l'utilizzo delle risorse in EMR Serverless. L'impostazione di questa proprietà `true` nella classificazione della configurazione Spark indica a EMR Serverless di ottimizzare l'allocazione delle risorse degli esecutori per allineare meglio la velocità con cui Spark richiede e annulla gli esecutori con la velocità con cui EMR Serverless crea e rilascia i lavoratori. In questo modo, EMR Serverless riutilizza in modo più ottimale i lavoratori tra le fasi, con conseguente riduzione dei costi di esecuzione di lavori con più fasi mantenendo le stesse prestazioni.

Questa proprietà è disponibile in tutte le versioni di rilascio di Amazon EMR.

Di seguito è riportato un esempio di classificazione della configurazione `dynamicAllocationOptimization`.

```
[
  {
    "Classification": "spark",
    "Properties": {
      "dynamicAllocationOptimization": "true"
    }
  }
]
```

Considerate quanto segue se utilizzate l'ottimizzazione dinamica dell'allocazione:

- Questa ottimizzazione è disponibile per i job Spark per i quali hai abilitato l'allocazione dinamica delle risorse.
- Per ottenere la massima efficienza in termini di costi, consigliamo di configurare una scalabilità superiore limitata ai lavoratori utilizzando l'impostazione a livello di job `spark.dynamicAllocation.maxExecutors` o l'impostazione della capacità massima a livello di [applicazione in base al carico di lavoro](#).

- Potresti non vedere un miglioramento dei costi nei lavori più semplici. Ad esempio, se il job viene eseguito su un piccolo set di dati o termina l'esecuzione in un'unica fase, Spark potrebbe non aver bisogno di un numero maggiore di executor o di più eventi di scalabilità.
- I lavori con una sequenza composta da una fase grande, fasi più piccole e poi ancora una fase grande potrebbero subire una regressione nella fase di esecuzione del processo. Poiché EMR Serverless utilizza le risorse in modo più efficiente, potrebbe portare a un minor numero di lavoratori disponibili per fasi più grandi, con conseguente maggiore autonomia.

## Proprietà del lavoro Spark

La tabella seguente elenca le proprietà Spark opzionali e i relativi valori predefiniti che puoi sostituire quando invii un lavoro Spark.

Chiave	Descrizione	Valore predefinito
<code>spark.archives</code>	Un elenco di archivi separati da virgole che Spark estrae nella directory di lavoro di ogni esecutore. I tipi di file supportati includono, e. <code>.jar</code> <code>.tar.gz</code> <code>.tgz</code> <code>.zip</code> Per specificare il nome della directory da estrarre, aggiungilo o # dopo il nome del file che desideri estrarre. Ad esempio <code>file.zip#directory</code> .	NULL
<code>spark.authenticate</code>	Opzione che attiva l'autenticazione delle connessioni interne di Spark.	TRUE
<code>spark.driver.cores</code>	Il numero di core utilizzati dal driver.	4
<code>spark.driver.extraJavaOptions</code>	Opzioni Java aggiuntive per il driver Spark.	NULL

Chiave	Descrizione	Valore predefinito
<code>spark.driver.memory</code>	La quantità di memoria utilizzata dal driver.	14 G
<code>spark.dynamicAllocation.enabled</code>	Opzione che attiva l'allocazione dinamica delle risorse. Questa opzione aumenta o riduce il numero di esecutori registrati con l'applicazione, in base al carico di lavoro.	TRUE
<code>spark.dynamicAllocation.executorIdleTimeout</code>	Il periodo di tempo in cui un executor può rimanere inattivo prima che Spark lo rimuova. Questo vale solo se attivi l'allocazione dinamica.	anni '60
<code>spark.dynamicAllocation.initialExecutors</code>	Il numero iniziale di esecutori da eseguire se si attiva l'allocazione dinamica.	3
<code>spark.dynamicAllocation.maxExecutors</code>	Il limite superiore per il numero di esecutori se si attiva l'allocazione dinamica.	Per 6.10.0 e versioni successive, <i>infinity</i> Per 6.9.0 e versioni precedenti, <i>100</i>
<code>spark.dynamicAllocation.minExecutors</code>	Il limite inferiore per il numero di esecutori se si attiva l'allocazione dinamica.	0
<code>spark.emr-serverless.allocation.batch.size</code>	Il numero di contenitori da richiedere in ogni ciclo di allocazione degli esecutori. C'è un intervallo di un secondo tra ogni ciclo di allocazione.	20

Chiave	Descrizione	Valore predefinito
<code>spark.emr-serverless.driver.disk</code>	Il disco del driver Spark.	20G
<code>spark.emr-serverless.driverEnv.</code> <b>[KEY]</b>	Opzione che aggiunge variabili di ambiente al driver Spark.	NULL
<code>spark.emr-serverless.executor.disk</code>	Il disco Spark Executor.	20G
<code>spark.emr-serverless.memoryOverheadFactor</code>	Imposta il sovraccarico di memoria da aggiungere alla memoria del contenitore del driver e dell'executor.	0.1
<code>spark.emr-serverless.driver.disk.type</code>	Il tipo di disco collegato al driver Spark.	Standard
<code>spark.emr-serverless.executor.disk.type</code>	Il tipo di disco collegato agli esecutori Spark.	Standard
<code>spark.executor.cores</code>	Il numero di core utilizzati da ogni executor.	4
<code>spark.executor.extraJavaOptions</code>	Opzioni Java aggiuntive per l'esecutore Spark.	NULL
<code>spark.executor.instances</code>	Il numero di contenitori Spark Executor da allocare.	3
<code>spark.executor.memory</code>	La quantità di memoria utilizzata da ogni executor.	14 G
<code>spark.executorEnv.</code> <b>[KEY]</b>	Opzione che aggiunge variabili di ambiente agli esecutori Spark.	NULL

Chiave	Descrizione	Valore predefinito
<code>spark.files</code>	Un elenco di file separati da virgole da inserire nella directory di lavoro di ogni executor. È possibile accedere ai percorsi di questi file nell'executore con <code>SparkFile s.get( <i>fileName</i> )</code>	NULL
<code>spark.hadoop.hive.metastore.client.factory.class</code>	La classe di implementazione del metastore Hive.	NULL
<code>spark.jars</code>	Jar aggiuntivi da aggiungere al classpath di runtime del driver e degli executor.	NULL
<code>spark.network.crypto.enabled</code>	Opzione che attiva la crittografia RPC basata su AES. Ciò include il protocollo di autenticazione aggiunto in Spark 2.2.0.	FALSE
<code>spark.sql.warehouse.dir</code>	La posizione predefinita per i database e le tabelle gestiti.	Il valore di <code>\$PWD/spark-warehouse</code>
<code>spark.submit.pyFiles</code>	Un elenco separato da virgole di <code>.zip.egg</code> , o <code>.py</code> file da inserire nelle app per PYTHONPATH Python.	NULL

La tabella seguente elenca i parametri di invio Spark predefiniti.

Chiave	Descrizione	Valore predefinito
<code>archives</code>	Un elenco di archivi separati da virgole che Spark estrae nella directory di lavoro di ogni esecutore.	NULL
<code>class</code>	La classe principale dell'applicazione (per le app Java e Scala).	NULL
<code>conf</code>	Una proprietà di configurazione arbitraria di Spark.	NULL
<code>driver-cores</code>	Il numero di core utilizzati dal driver.	4
<code>driver-memory</code>	La quantità di memoria utilizzata dal driver.	14 G
<code>executor-cores</code>	Il numero di core utilizzati da ogni esecutore.	4
<code>executor-memory</code>	La quantità di memoria utilizzata dall'esecutore.	14 G
<code>files</code>	Un elenco di file separati da virgole da inserire nella directory di lavoro di ogni esecutore. È possibile accedere ai percorsi di questi file nell'esecutore con <code>SparkFile s.get( <i>fileName</i> )</code>	NULL
<code>jars</code>	Un elenco di jar separati da virgole da includere nei	NULL

Chiave	Descrizione	Valore predefinito
	percorsi di classe del driver e dell'executor.	
num-executors	Il numero di esecutori da avviare.	3
py-files	Un elenco separato da virgole di .zip.egg, o .py file da inserire nelle app per PYTHONPATH Python.	NULL
verbose	Opzione che attiva un output di debug aggiuntivo.	NULL

## Esempi Spark

L'esempio seguente mostra come utilizzare l'StartJobRunAPI per eseguire uno script Python. Per un end-to-end tutorial che utilizza questo esempio, vedi [Guida introduttiva ad Amazon EMR Serverless](#). Puoi trovare altri esempi su come eseguire PySpark lavori e aggiungere dipendenze Python nel repository [EMR Serverless Samples](#). GitHub

```
aws emr-serverless start-job-run \
  --application-id application-id \
  --execution-role-arn job-role-arn \
  --job-driver '{
    "sparkSubmit": {
      "entryPoint": "s3://us-east-1.elasticmapreduce/emr-containers/samples/
wordcount/scripts/wordcount.py",
      "entryPointArguments": ["s3://amzn-s3-demo-destination-bucket/
wordcount_output"],
      "sparkSubmitParameters": "--conf spark.executor.cores=1 --conf
spark.executor.memory=4g --conf spark.driver.cores=1 --conf spark.driver.memory=4g --
conf spark.executor.instances=1"
    }
  }'
```

L'esempio seguente mostra come utilizzare l'StartJobRunAPI per eseguire uno Spark JAR.

```
aws emr-serverless start-job-run \  
  --application-id application-id \  
  --execution-role-arn job-role-arn \  
  --job-driver '{  
    "sparkSubmit": {  
      "entryPoint": "/usr/lib/spark/examples/jars/spark-examples.jar",  
      "entryPointArguments": ["1"],  
      "sparkSubmitParameters": "--class org.apache.spark.examples.SparkPi --conf  
spark.executor.cores=4 --conf spark.executor.memory=20g --conf spark.driver.cores=4 --  
conf spark.driver.memory=8g --conf spark.executor.instances=1"  
    }  
  }'
```

## Utilizzo delle configurazioni Hive quando si eseguono job EMR Serverless

È possibile eseguire lavori Hive su un'applicazione con il `type` parametro impostato su `HIVE`. I lavori devono essere compatibili con la versione di Hive compatibile con la versione di rilascio di Amazon EMR. Ad esempio, quando esegui lavori su un'applicazione con Amazon EMR release 6.6.0, il processo deve essere compatibile con Apache Hive 3.1.2. Per informazioni sulle versioni dell'applicazione per ogni versione, consulta [Versioni di rilascio di Amazon EMR Serverless](#)

### Parametri del lavoro Hive

Quando utilizzi l'[StartJobRunAPI](#) per eseguire un job Hive, devi specificare i seguenti parametri.

Parametri obbligatori

- [Ruolo di runtime del job Hive](#)
- [Parametro Hive Job Driver](#)
- [Parametro di sovrascrittura della configurazione Hive](#)

### Ruolo di runtime del job Hive

**executionRoleArn** Da utilizzare per specificare l'ARN per il ruolo IAM utilizzato dall'applicazione per eseguire i job Hive. Questo ruolo deve contenere le seguenti autorizzazioni:

- Leggi dai bucket S3 o da altre fonti di dati in cui risiedono i tuoi dati

- Leggi dai bucket o dai prefissi S3 in cui risiedono il file di query Hive e il file di query init
- Leggi e scrivi nei bucket S3 in cui risiedono la directory Hive Scratch e la directory di magazzino Hive Metastore
- Scrivi nei bucket S3 in cui intendi scrivere l'output finale
- Scrivi i log in un bucket o prefisso S3 che specifica `S3MonitoringConfiguration`
- Accesso alle chiavi KMS se utilizzi le chiavi KMS per crittografare i dati nel tuo bucket S3
- Accesso al AWS Glue Data Catalog

Se il tuo job Hive legge o scrive dati da o verso altre fonti di dati, specifica le autorizzazioni appropriate in questo ruolo IAM. Se non fornisci queste autorizzazioni al ruolo IAM, il tuo lavoro potrebbe fallire. Per ulteriori informazioni, consulta [Ruoli Job Runtime per Amazon EMR Serverless](#).

## Parametro Hive Job Driver

**jobDriver** Utilizzato per fornire input al lavoro. Il parametro job driver accetta solo un valore per il tipo di processo che si desidera eseguire. Quando si specifica `hive` come tipo di lavoro, EMR Serverless passa una query Hive al parametro. `jobDriver` I job Hive hanno i seguenti parametri:

- **query**— Questo è il riferimento in Amazon S3 al file di query Hive che desideri eseguire.
- **parameters**— Queste sono le proprietà di configurazione aggiuntive di Hive che desideri sovrascrivere. Per sovrascrivere le proprietà, passale a questo parametro come. `--hiveconf property=value` Per sovrascrivere le variabili, passale a questo parametro come. `--hivevar key=value`
- **initQueryFile**— Questo è il file di query init Hive. Hive esegue questo file prima della query e può usarlo per inizializzare le tabelle.

## Parametro di sovrascrittura della configurazione Hive

Utilizzato **configurationOverrides** per sovrascrivere le proprietà di configurazione a livello di monitoraggio e di applicazione. Questo parametro accetta un oggetto JSON con i due campi seguenti:

- **monitoringConfiguration**— Utilizza questo campo per specificare l'URL di Amazon S3 (`s3MonitoringConfiguration`) in cui desideri che il job EMR Serverless memorizzi i log del tuo processo Hive. Assicurati di creare questo bucket con lo stesso Account AWS che ospita l'applicazione e nello stesso luogo in cui è in esecuzione il processo. Regione AWS

- **applicationConfiguration**— È possibile fornire un oggetto di configurazione in questo campo per sovrascrivere le configurazioni predefinite per le applicazioni. Puoi utilizzare una sintassi abbreviata per fornire la configurazione oppure fare riferimento all'oggetto di configurazione in un file JSON. Gli oggetti di configurazione sono composti da una classificazione, proprietà e configurazioni nidificate opzionali. Le proprietà sono costituite dalle impostazioni che desideri ignorare in un dato file. Puoi specificare diverse classificazioni per più applicazioni in un singolo oggetto JSON.

#### Note

Le classificazioni di configurazione disponibili variano in base alla specifica release EMR Serverless. Ad esempio, le classificazioni per Log4j personalizzate `spark-executor-log4j2` sono disponibili solo con le versioni `spark-driver-log4j2 6.8.0` e successive.

Se si passa la stessa configurazione in un'applicazione override e nei parametri Hive, i parametri Hive hanno la priorità. L'elenco seguente classifica le configurazioni dalla priorità più alta alla priorità più bassa.

- Configurazione fornita come parte dei parametri Hive con. `--hiveconf property=value`
- La configurazione fornita come parte dell'applicazione ha la precedenza quando si avvia un lavoro.
- La configurazione fornita come parte della configurazione `runtimeConfiguration` quando si crea un'applicazione.
- Configurazioni ottimizzate che Amazon EMR assegna per il rilascio.
- Configurazioni open source predefinite per l'applicazione.

Per ulteriori informazioni sulla dichiarazione delle configurazioni a livello di applicazione e sulla sovrascrittura delle configurazioni durante l'esecuzione del processo, vedere. [Configurazione predefinita dell'applicazione per EMR Serverless](#)

## Proprietà del lavoro di Hive

La tabella seguente elenca le proprietà obbligatorie che è necessario configurare quando si invia un lavoro Hive.

Impostazione	Descrizione
<code>hive.exec.scratchdir</code>	La posizione Amazon S3 in cui EMR Serverless crea file temporanei durante l'esecuzione del job Hive.
<code>hive.metastore.warehouse.dir</code>	La posizione Amazon S3 dei database per le tabelle gestite in Hive.

La tabella seguente elenca le proprietà opzionali di Hive e i relativi valori predefiniti che puoi sostituire quando invii un lavoro Hive.

Impostazione	Descrizione	Valore predefinito
<code>fs.s3.customAWSCredentialsProvider</code>	Il provider di AWS credenziali che desideri utilizzare.	<code>com.amazonaws.auth.DefaultAWSCredentialsProviderChain</code>
<code>fs.s3a.aws.credentials.provider</code>	Il provider di AWS credenziali che desideri utilizzare con un file system S3A.	<code>com.amazonaws.auth.DefaultAWSCredentialsProviderChain</code>
<code>hive.auto.convert.join</code>	Opzione che attiva la conversione automatica dei join comuni in mapjoin, in base alla dimensione del file di input.	TRUE
<code>hive.auto.convert.join.noconditionaltask</code>	Opzione che attiva l'ottimizzazione quando Hive converte un common join in un mapjoin in base alla dimensione del file di input.	TRUE
<code>hive.auto.convert.join.noconditionaltask.size</code>	Un join si converte direttamente in un mapjoin di dimensioni inferiori a questa dimensione.	Il valore ottimale viene calcolato in base alla memoria delle attività Tez

Impostazione	Descrizione	Valore predefinito
<code>hive.cbo.enable</code>	Opzione che attiva ottimizzazioni basate sui costi con il framework Calcite.	TRUE
<code>hive.cli.tez.session.async</code>	Opzione per avviare una sessione Tez in background durante la compilazione della query Hive. Se impostato su <code>false</code> , Tez AM si avvia dopo la compilazione della query Hive.	TRUE
<code>hive.compute.query.using.stats</code>	Opzione che attiva Hive per rispondere a determinate domande con statistiche memorizzate nel metastore. Per le statistiche di base, imposta su <code>hive.stats.autogather</code> TRUE. Per una raccolta di interrogazioni più avanzata, <code>analyze table queries</code> esegui.	TRUE
<code>hive.default.fileformat</code>	Il formato di file predefinito per le CREATE TABLE istruzioni. È possibile sovrascriverlo esplicitamente se lo si specifica <code>STORED AS [FORMAT]</code> nel comando. CREATE TABLE	TEXTFILE
<code>hive.driver.cores</code>	Il numero di core da utilizzare per il processo del driver Hive.	2
<code>hive.driver.disk</code>	La dimensione del disco per il driver Hive.	20G

Impostazione	Descrizione	Valore predefinito
<code>hive.driver.disk.type</code>	Il tipo di disco per il driver Hive.	Standard
<code>hive.tez.disk.type</code>	La dimensione del disco per i Tez worker.	Standard
<code>hive.driver.memory</code>	La quantità di memoria da utilizzare per ogni processo del driver Hive. Hive CLI e Tez Application Master condividono questa memoria in parti uguali con il 20% di spazio di crescita.	6 G
<code>hive.emr-serverless.launch.env.[ KEY]</code>	Opzione per impostare la variabile di <b>KEY</b> ambiente in tutti i processi specifici di Hive, come il driver Hive, Tez AM e l'attività Tez.	
<code>hive.exec.dynamic.partition</code>	Opzioni che attivano le partizioni dinamiche in DML/DDL.	TRUE
<code>hive.exec.dynamic.partition.mode</code>	Opzione che specifica se si desidera utilizzare la modalità rigorosa o la modalità non rigorosa. In modalità rigorosa, è necessario specificare almeno una partizione statica in caso di sovrascrittura accidentale di tutte le partizioni. In modalità non rigorosa, tutte le partizioni possono essere dinamiche.	strict

Impostazione	Descrizione	Valore predefinito
<code>hive.exec.max.dynamic.partitions</code>	Il numero massimo di partizioni dinamiche create da Hive in totale.	1000
<code>hive.exec.max.dynamic.partitions.per.node</code>	Numero massimo di partizioni dinamiche che Hive crea in ogni nodo mapper e reducer.	100
<code>hive.exec.orc.split.strategy</code>	Prevede uno dei seguenti valori: <code>BI</code> , <code>ETL</code> , <code>HYBRID</code> . Questa non è una configurazione a livello utente. <code>BI</code> specifica che si desidera dedicare meno tempo alla generazione frazionata rispetto all'esecuzione delle query. <code>ETL</code> specifica che si desidera dedicare più tempo alla generazione frazionata. <code>HYBRID</code> specifica una scelta delle strategie di cui sopra in base all'euristica.	HYBRID
<code>hive.exec.reducers.bytes.per.reducer</code>	La dimensione di ogni riduttore. L'impostazione predefinita è 256 MB. Se la dimensione di input è 1G, il job utilizza 4 riduttori.	256000000
<code>hive.exec.reducers.max</code>	Il numero massimo di riduttori.	256

Impostazione	Descrizione	Valore predefinito
<code>hive.exec.stagingdir</code>	Il nome della directory che memorizza i file temporanei creati da Hive all'interno delle posizioni delle tabelle e nella posizione della directory scratch specificata nella proprietà <code>hive.exec.scratchdir</code>	<code>.hive-staging</code>
<code>hive.fetch.task.conversion</code>	Prevede uno dei seguenti valori: <code>NONEMINIMAL</code> , o <code>MORE</code> . Hive può convertire le interrogazioni di selezione in una singola attività. <code>FETCH</code> riduce al minimo la latenza.	<code>MORE</code>
<code>hive.groupby.position.alias</code>	Opzione che fa sì che Hive utilizzi un alias di posizione della colonna nelle istruzioni <code>GROUP BY</code>	<code>FALSE</code>
<code>hive.input.format</code>	Il formato di input predefinito. Imposta su <code>HiveInputFormat</code> se riscontri problemi con <code>CombineHiveInputFormat</code> .	<code>org.apache.hadoop.hive.q1.io.CombineHiveInputFormat</code>
<code>hive.log.explain.output</code>	Opzione che attiva le spiegazioni dell'output esteso per qualsiasi query nel registro di Hive.	<code>FALSE</code>
<code>hive.log.level</code>	Il livello di registrazione di Hive.	<code>INFO</code>

Impostazione	Descrizione	Valore predefinito
<code>hive.mapred.reduce.tasks.speculative.execution</code>	Opzione che attiva il lancio speculativo dei riduttori. Supportato solo con Amazon EMR 6.10.x e versioni precedenti.	TRUE
<code>hive.max-task-containers</code>	Il numero massimo di contenitori simultanei. La memoria del mappatore configurata viene moltiplicata per questo valore per determinare la memoria disponibile utilizzata per il calcolo diviso e la priorità delle attività.	1000
<code>hive.merge.mapfiles</code>	Opzione che causa l'unione di file di piccole dimensioni alla fine di un lavoro basato sulla sola mappa.	TRUE
<code>hive.merge.size.per.task</code>	La dimensione dei file uniti alla fine del lavoro.	256000000
<code>hive.merge.tezfiles</code>	Opzione che attiva l'unione di file di piccole dimensioni alla fine di un Tez DAG.	FALSE
<code>hive.metastore.client.factory.class</code>	Il nome della classe factory che produce oggetti che implementano l'interfaccia. <code>IMetaStoreClient</code>	<code>com.amazonaws.glue.catalog.metastore.AWSGlueDataCatalogHiveClientFactory</code>

Impostazione	Descrizione	Valore predefinito
<code>hive.metastore.glue.catalogid</code>	Se il AWS Glue Data Catalog funge da metastore ma viene eseguito in un ambiente Account AWS diverso dai lavori, l'ID del Account AWS luogo in cui i lavori sono in esecuzione.	NULL
<code>hive.metastore.uris</code>	L'URI parsimonioso utilizzato dal client metastore per connettersi al metastore remoto.	NULL
<code>hive.optimize.ppd</code>	Opzione che attiva il pushdown dei predicati.	TRUE
<code>hive.optimize.ppd.storage</code>	Opzione che attiva il pushdown dei predicati verso i gestori di archiviazione.	TRUE
<code>hive.orderby.position.alias</code>	Opzione che fa sì che Hive utilizzi un alias di posizione della colonna nelle istruzioni. ORDER BY	TRUE
<code>hive.prewarm.enabled</code>	Opzione che attiva il preriscaldamento del contenitore per Tez.	FALSE
<code>hive.prewarm.numcontainers</code>	Il numero di contenitori da preriscaldare per Tez.	10

Impostazione	Descrizione	Valore predefinito
<code>hive.stats.autogather</code>	Opzione che fa sì che Hive raccolga automaticamente le statistiche di base durante il comando. <code>INSERT OVERWRITE</code>	TRUE
<code>hive.stats.fetch.column.stats</code>	Opzione che disattiva il recupero delle statistiche delle colonne dal metastore. Il recupero delle statistiche sulle colonne può essere costoso quando il numero di colonne è elevato.	FALSE
<code>hive.stats.gather.num.threads</code>	Il numero di thread utilizzati dai comandi <code>partialscan</code> e <code>noscan analysis</code> per le tabelle partizionate. Questo vale solo per i formati di file che implementano <code>StatsProvidingRecordReader</code> (come ORC).	10
<code>hive.strict.checks.cartesian.product</code>	Opzioni che attivano rigorosi controlli cartesiani di join. Questi controlli non consentono un prodotto cartesiano (un <code>cross join</code> ).	FALSE
<code>hive.strict.checks.type.safety</code>	Opzione che attiva rigorosi controlli di sicurezza tipografici e disattiva il confronto tra entrambi <code>string</code> e <code>bigint double</code>	TRUE

Impostazione	Descrizione	Valore predefinito
<code>hive.support.quote d.identifiers</code>	Si aspetta un valore di NONE o COLUMN. NONE implica che solo i caratteri alfanumerici e i caratteri di sottolineatura siano validi negli identificatori. COLUMN implica che i nomi delle colonne possono contenere qualsiasi carattere.	COLUMN
<code>hive.tez.auto.redu cer.parallelism</code>	Opzione che attiva la funzione di parallelismo del riduttore automatico Tez. Hive stima ancora le dimensioni dei dati e imposta le stime di parallelismo. Tez campiona le dimensioni di output dei vertici della sorgente e regola le stime in fase di esecuzione, se necessario.	TRUE
<code>hive.tez.container .size</code>	La quantità di memoria da utilizzare per il processo di operazione Tez.	6144
<code>hive.tez.cpu.vcores</code>	Il numero di core da utilizzare per ogni attività Tez.	2
<code>hive.tez.disk.size</code>	La dimensione del disco per ogni contenitore di attività.	20G
<code>hive.tez.input.for mat</code>	Il formato di input per la generazione di split in Tez AM.	<code>org.apache.hadoop. hive ql.io.HiveInp utFormat</code>

Impostazione	Descrizione	Valore predefinito
<code>hive.tez.min.partition.factor</code>	Limite inferiore di riduttori specificato da Tez quando si attiva il parallelismo del riduttore automatico.	0.25
<code>hive.vectorized.execution.enabled</code>	Opzione che attiva la modalità vettoriale di esecuzione delle query.	TRUE
<code>hive.vectorized.execution.reduce.enabled</code>	Opzione che attiva la modalità vettoriale del lato ridotto dell'esecuzione di una query.	TRUE
<code>javax.jdo.option.ConnectionDriverName</code>	Il nome della classe del driver per un metastore JDBC.	<code>org.apache.derby.jdbc.EmbeddedDriver</code>
<code>javax.jdo.option.ConnectionPassword</code>	La password associata a un database metastore.	NULL
<code>javax.jdo.option.ConnectionURL</code>	La stringa di connessione JDBC per un metastore JDBC.	<code>jdbc:derby;;databaseName=metastore_db;create=true</code>
<code>javax.jdo.option.ConnectionUserName</code>	Il nome utente associato a un database metastore.	NULL
<code>mapreduce.input.fileinputformat.split.maxsize</code>	La dimensione massima di una divisione durante il calcolo della suddivisione quando il formato di input è <code>org.apache.hadoop.hive.q1.io.CombineHiveInputFormat</code> . Il valore 0 indica l'assenza di limiti.	0

Impostazione	Descrizione	Valore predefinito
<code>tez.am.dag.cleanup.on.completion</code>	Opzione che attiva la pulizia dei dati casuali al termine del DAG.	TRUE
<code>tez.am.emr-serverless.launch.env.[ KEY]</code>	Opzione per impostare la variabile di <code>KEY</code> ambiente nel processo Tez AM. Per Tez AM, questo valore sostituisce il valore <code>hive.emr-serverless.launch.env.[ KEY]</code>	
<code>tez.am.log.level</code>	Il livello di registrazione root che EMR Serverless passa all'app master Tez.	INFO
<code>tez.am.sleep.time.before.exit.millis</code>	EMR Serverless dovrebbe inviare gli eventi ATS dopo questo periodo di tempo successivo alla richiesta di spegnimento AM.	0
<code>tez.am.speculation.enabled</code>	Opzione che causa l'avvio speculativo di attività più lente. Ciò può aiutare a ridurre la latenza del lavoro quando alcune attività sono più lente a causa di macchine danneggiate o lente. Supportato solo con Amazon EMR 6.10.x e versioni precedenti.	FALSE

Impostazione	Descrizione	Valore predefinito
<code>tez.am.task.max.failed.attempts</code>	Il numero massimo di tentativi che possono fallire per una determinata attività prima che l'attività fallisca. Questo numero non include i tentativi terminati manualmente.	3
<code>tez.am.vertex.cleanup.height</code>	Una distanza alla quale, se tutti i vertici dipendenti sono completi, Tez AM eliminerà i dati relativi allo shuffle dei vertici. Questa funzione è disattivata quando il valore è 0. Le versioni 6.8.0 e successive di Amazon EMR supportano questa funzionalità.	0
<code>tez.client.asynchronous-stop</code>	Opzione che fa sì che EMR Serverless invii gli eventi ATS prima che termini il driver Hive.	FALSE
<code>tez.grouping.max-size</code>	Il limite massimo di dimensione (in byte) di una divisione raggruppata. Questo limite impedisce suddivisioni eccessivamente grandi.	1073741824
<code>tez.grouping.min-size</code>	Il limite di dimensione inferiore (in byte) di una divisione raggruppata. Questo limite impedisce troppe suddivisioni di piccole dimensioni.	16777216

Impostazione	Descrizione	Valore predefinito
<code>tez.runtime.io.sort.mb</code>	La dimensione del soft buffer quando Tez ordina l'output viene ordinata.	Il valore ottimale viene calcolato in base alla memoria delle attività di Tez
<code>tez.runtime.unordered.output.buffer.size-mb</code>	La dimensione del buffer da usare se Tez non scrive direttamente su disco.	Il valore ottimale viene calcolato in base alla memoria delle attività di Tez
<code>tez.shuffle-vertex-manager.max-src-fraction</code>	La frazione di attività di origine che deve essere completata prima che EMR Serverless pianifichi tutte le attività per il vertice corrente (in caso di connessione). ScatterGather Il numero di attività pronte per la pianificazione sul vertice corrente varia linearmente tra <code>e.min-fraction</code> e <code>e.max-fraction</code> Il valore predefinito è il valore predefinito o, a seconda del valore maggiore. <code>tez.shuffle-vertex-manager.min-src-fraction</code>	0.75
<code>tez.shuffle-vertex-manager.min-src-fraction</code>	La frazione di attività di origine che deve essere completata prima che EMR Serverless pianifichi le attività per il vertice corrente (in caso di connessione). ScatterGather	0.25

Impostazione	Descrizione	Valore predefinito
<code>tez.task.emr-serverless.launch.env.[ <i>KEY</i>]</code>	Opzione per impostare la variabile di <i>KEY</i> ambiente nel processo di attività Tez. Per le attività Tez, questo valore ha la precedenza sul valore. <code>hive.emr-serverless.launch.env.[ <i>KEY</i>]</code>	
<code>tez.task.log.level</code>	Il livello di registrazione principale che EMR Serverless passa alle attività Tez.	INFO
<code>tez.yarn.ats.event.flush.timeout.millis</code>	Il periodo massimo di attesa che AM deve attendere prima che gli eventi vengano eliminati prima dello spegnimento.	300000

## Esempi di lavoro in Hive

Il seguente esempio di codice mostra come eseguire una query Hive con l'API StartJobRunAPI.

```
aws emr-serverless start-job-run \
  --application-id application-id \
  --execution-role-arn job-role-arn \
  --job-driver '{
    "hive": {
      "query": "s3://amzn-s3-demo-bucket/emr-serverless-hive/query/hive-
query.q1",
      "parameters": "--hiveconf hive.log.explain.output=false"
    }
  }' \
  --configuration-overrides '{
    "applicationConfiguration": [{
      "classification": "hive-site",
      "properties": {
```

```

        "hive.exec.scratchdir": "s3://amzn-s3-demo-bucket/emr-serverless-hive/
hive/scratch",
        "hive.metastore.warehouse.dir": "s3://amzn-s3-demo-bucket/emr-
serverless-hive/hive/warehouse",
        "hive.driver.cores": "2",
        "hive.driver.memory": "4g",
        "hive.tez.container.size": "4096",
        "hive.tez.cpu.vcores": "1"
    }
}
}'

```

È possibile trovare altri esempi di come eseguire i job Hive nel repository [EMR Serverless GitHub Samples](#).

## Resilienza EMR Serverless Job

Le versioni 7.1.0 e successive di EMR Serverless includono il supporto per la resilienza dei job, in modo da riprovare automaticamente tutti i job non riusciti senza alcun input manuale da parte dell'utente. Un altro vantaggio della resilienza dei processi è che EMR Serverless sposta i job run in diverse zone di disponibilità (AZ) in caso di problemi in una zona di disponibilità.

Per abilitare la resilienza del lavoro, imposta la politica di riprova per il tuo lavoro. Una politica di riprova assicura che EMR Serverless riavvii automaticamente un processo in caso di errore in qualsiasi momento. Le politiche di riprova sono supportate sia per i processi in batch che per quelli in streaming, quindi è possibile personalizzare la resilienza dei processi in base al caso d'uso. La tabella seguente confronta i comportamenti e le differenze di resilienza dei processi tra processi in batch e in streaming.

	Processi batch	Lavori in streaming
Comportamento predefinito	Non riesegue il lavoro.	Riprova sempre a eseguire il processo man mano che l'applicazione crea checkpoint durante l'esecuzione del lavoro.
Punto di riprova	I processi Batch non hanno punti di controllo, quindi EMR	I processi di streaming supportano i checkpoint

	Processi batch	Lavori in streaming
	Serverless riesegue sempre il lavoro dall'inizio.	t, quindi puoi configurare la query di streaming per salvare lo stato di runtime e l'avanzamento verso una posizione di checkpoint in Amazon S3. EMR Serverless riprende l'esecuzione del job dal checkpoint. Per ulteriori informazioni, consulta <a href="#">Recovery from failure with Checkpointing</a> nella documentazione di Apache Spark.
Numero massimo di tentativi di nuovo tentativo	Consente un massimo di 10 tentativi.	I job di streaming dispongono di un sistema integrato di prevenzione degli errori, pertanto l'applicazione smette di riprovare i lavori se questi continuano a fallire dopo un'ora. Il numero predefinito di tentativi entro un'ora è di cinque tentativi. È possibile configurare questo numero di tentativi in modo che sia compreso tra 1 o 10. Non è possibile personalizzare il numero massimo di tentativi. Il valore 1 indica che non ci sono nuovi tentativi.

Quando EMR Serverless tenta di eseguire nuovamente un processo, lo indicizza anche con un numero di tentativo, in modo da poter tenere traccia del ciclo di vita di un processo in tutti i suoi tentativi.

È possibile utilizzare le operazioni dell'API EMR Serverless o AWS CLI per modificare la resilienza del lavoro o visualizzare le informazioni relative alla resilienza del lavoro. Per ulteriori informazioni, consulta la guida alle [API EMR Serverless](#).

Per impostazione predefinita, EMR Serverless non esegue nuovamente i processi batch. Per abilitare i nuovi tentativi per i processi batch, configura il `maxAttempts` parametro quando si avvia l'esecuzione di un processo batch. Il `maxAttempts` parametro è applicabile solo ai processi batch. L'impostazione predefinita è 1, il che significa che non viene eseguito nuovamente il processo. I valori accettati sono compresi tra 1 e 10.

L'esempio seguente mostra come specificare un numero massimo di 10 tentativi all'avvio di un job run.

```
aws emr-serverless start-job-run
--application-id <APPLICATION_ID> \
--execution-role-arn <JOB_EXECUTION_ROLE> \
--mode 'BATCH' \
--retry-policy '{
  "maxAttempts": 10
}' \
--job-driver '{
  "sparkSubmit": {
    "entryPoint": "/usr/lib/spark/examples/jars/spark-examples-does-not-
exist.jar",
    "entryPointArguments": ["1"],
    "sparkSubmitParameters": "--class org.apache.spark.examples.SparkPi"
  }
}'
```

EMR Serverless riprova a tempo indeterminato i processi di streaming se falliscono. Per evitare che si verifichino problemi dovuti a errori ripetuti e irrecuperabili, utilizza il controllo di prevenzione degli errori per lo streaming di nuovi tentativi di lavoro. `maxFailedAttemptsPerHour` Questo parametro consente di specificare il numero massimo di tentativi falliti consentiti con un'ora prima che EMR Serverless interrompa i nuovi tentativi. L'impostazione predefinita è cinque. I valori accettati sono compresi tra 1 e 10.

```
aws emr-serverless start-job-run
--application-id <APPLICATION_ID> \
--execution-role-arn <JOB_EXECUTION_ROLE> \
--mode 'STREAMING' \
--retry-policy '{
```

```

    "maxFailedAttemptsPerHour": 7
  }' \
  --job-driver '{
    "sparkSubmit": {
      "entryPoint": "/usr/lib/spark/examples/jars/spark-examples-does-not-
exist.jar",
      "entryPointArguments": ["1"],
      "sparkSubmitParameters": "--class org.apache.spark.examples.SparkPi"
    }
  }'

```

Puoi anche utilizzare le altre operazioni API Job Run per ottenere informazioni sui lavori. Ad esempio, è possibile utilizzare il `attempt` parametro con l'`GetJobRun` operazione per ottenere dettagli su uno specifico tentativo di lavoro. Se non si include il `attempt` parametro, l'operazione restituisce informazioni sull'ultimo tentativo.

```

aws emr-serverless get-job-run \
  --job-run-id job-run-id \
  --application-id application-id \
  --attempt 1

```

L'`ListJobRunAttempts` operazione restituisce informazioni su tutti i tentativi relativi all'esecuzione di un processo.

```

aws emr-serverless list-job-run-attempts \
  --application-id application-id \
  --job-run-id job-run-id

```

L'`GetDashboardForJobRun` operazione crea e restituisce un URL che è possibile utilizzare per accedere all'applicazione UIs per l'esecuzione di un processo. Il `attempt` parametro consente di ottenere un URL per un tentativo specifico. Se non si include il `attempt` parametro, l'operazione restituisce informazioni sull'ultimo tentativo.

```

aws emr-serverless get-dashboard-for-job-run \
  --application-id application-id \
  --job-run-id job-run-id \
  --attempt 1

```

## Monitoraggio di un processo con una policy di ripetizione

Il supporto per la resilienza dei lavori aggiunge anche il nuovo evento EMR Serverless job run retry. EMR Serverless pubblica questo evento a ogni nuovo tentativo di lavoro. È possibile utilizzare questa notifica per tenere traccia dei nuovi tentativi di lavoro. Per ulteriori informazioni sugli eventi, consulta [Amazon EventBridge events](#).

## Registrazione con politica di riprova

Ogni volta che EMR Serverless riprova un job, il tentativo genera il proprio set di log. Per garantire che EMR Serverless possa inviare correttamente questi log ad Amazon S3 e Amazon senza CloudWatch sovrascriverli, EMR Serverless aggiunge un prefisso al formato del percorso di log S3 e CloudWatch al nome del flusso di log per includere il numero del tentativo del processo.

Di seguito è riportato un esempio dell'aspetto del formato.

```
'/applications/<applicationId>/jobs/<jobId>/attempts/<attemptNumber>/'.
```

Questo formato garantisce che EMR Serverless pubblichi tutti i log per ogni tentativo di lavoro nella propria posizione designata in Amazon S3 e CloudWatch [Per maggiori dettagli, consulta Archiviazione dei log](#).

### Note

EMR Serverless utilizza questo formato di prefisso solo con tutti i processi di streaming e tutti i processi batch con riprova abilitata.

## Configurazione Metastore per EMR Serverless

Un metastore Hive è una posizione centralizzata che memorizza le informazioni strutturali sulle tabelle, inclusi schemi, nomi delle partizioni e tipi di dati. Con EMR Serverless, puoi mantenere i metadati di questa tabella in un metastore che ha accesso ai tuoi lavori.

Hai due opzioni per un metastore Hive:

- Il catalogo dati AWS Glue
- Un metastore esterno di Apache Hive

## Utilizzo del AWS Glue Data Catalog come metastore

Puoi configurare i tuoi job Spark e Hive per utilizzare il AWS Glue Data Catalog come metastore. Consigliamo questa configurazione quando hai bisogno di un metastore persistente o un metastore condiviso da diverse applicazioni, servizi o Account AWS. Per ulteriori informazioni sul Data Catalog, consulta [Populating the AWS Glue Data Catalog](#). Per informazioni sui prezzi di AWS Glue, consulta i [prezzi di AWS Glue](#).

Puoi configurare il tuo job EMR Serverless per utilizzare il AWS Glue Data Catalog nella Account AWS stessa applicazione o in un'altra Account AWS.

### Configurazione del AWS Glue Data Catalog

Per configurare il Data Catalog, scegli il tipo di applicazione EMR Serverless che desideri utilizzare.

#### Spark

Quando utilizzi EMR Studio per eseguire i tuoi lavori con le applicazioni EMR Serverless Spark, il AWS Glue Data Catalog è il metastore predefinito.

Quando si utilizza SDKs o AWS CLI, è possibile impostare la `spark.hadoop.hive.metastore.client.factory.class` configurazione su `com.amazonaws.glue.catalog.metastore.AWSGlueDataCatalogHiveClientFactory` nei parametri dell'esecuzione del `sparkSubmit` processo. L'esempio seguente mostra come configurare il Data Catalog con AWS CLI.

```
aws emr-serverless start-job-run \
  --application-id application-id \
  --execution-role-arn job-role-arn \
  --job-driver '{
    "sparkSubmit": {
      "entryPoint": "s3://amzn-s3-demo-bucket/code/pyspark/
extreme_weather.py",
      "sparkSubmitParameters": "--conf
spark.hadoop.hive.metastore.client.factory.class=com.amazonaws.glue.catalog.metastore.AWSGlueDataCatalogHiveClientFactory
--conf spark.driver.cores=1 --conf spark.driver.memory=3g --conf
spark.executor.cores=4 --conf spark.executor.memory=3g"
    }
  }'
```

In alternativa, puoi impostare questa configurazione quando ne crei una nuova `SparkSession` nel tuo codice Spark.

```
from pyspark.sql import SparkSession

spark = (
    SparkSession.builder.appName("SparkSQL")
    .config(
        "spark.hadoop.hive.metastore.client.factory.class",
        "com.amazonaws.glue.catalog.metastore.AWSGlueDataCatalogHiveClientFactory",
    )
    .enableHiveSupport()
    .getOrCreate()
)

# we can query tables with SparkSQL
spark.sql("SHOW TABLES").show()

# we can also them with native Spark
print(spark.catalog.listTables())
```

## Hive

Per le applicazioni EMR Serverless Hive, il Data Catalog è il metastore predefinito. In altre parole, quando si eseguono lavori su un'applicazione EMR Serverless Hive, Hive registra le informazioni sui metastore nel Data Catalog nello stesso modo in cui viene eseguita l'applicazione. Account AWS Non è necessario un cloud privato virtuale (VPC) per utilizzare il Data Catalog come metastore.

Per accedere alle tabelle dei metastore di Hive, aggiungi le politiche AWS Glue richieste descritte in [Configurazione delle autorizzazioni IAM per Glue](#). AWS

## Configurazione dell'accesso tra account per EMR Serverless AWS e Glue Data Catalog

Per configurare l'accesso tra più account per EMR Serverless, è necessario prima accedere a quanto segue: Account AWS

- AccountA— E Account AWS in cui è stata creata un'applicazione EMR Serverless.

- AccountB— Un Account AWS che contiene un AWS Glue Data Catalog a cui desideri che i tuoi job EMR Serverless possano accedere.
1. Assicurati che un amministratore o un'altra identità autorizzata AccountB alleggi una politica delle risorse al Data Catalog in AccountB. Questa politica concede autorizzazioni AccountA specifiche per diversi account per eseguire operazioni sulle risorse del catalogo. AccountB

```
{
  "Version" : "2012-10-17",
  "Statement" : [ {
    "Effect" : "Allow",
    "Principal": {
      "AWS": [
        "arn:aws:iam::accountA:role/job-runtime-role-A"
      ]
    },
    "Action" : [
      "glue:GetDatabase",
      "glue:CreateDatabase",
      "glue:GetDataBases",
      "glue:CreateTable",
      "glue:GetTable",
      "glue:UpdateTable",
      "glue>DeleteTable",
      "glue:GetTables",
      "glue:GetPartition",
      "glue:GetPartitions",
      "glue:CreatePartition",
      "glue:BatchCreatePartition",
      "glue:GetUserDefinedFunctions"
    ],
    "Resource": ["arn:aws:glue:region:AccountB:catalog"]
  } ]
}
```

2. Aggiungi una policy IAM al ruolo di job runtime EMR Serverless in AccountA modo che quel ruolo possa accedere alle risorse del Data Catalog in AccountB

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
```

```

    "Action": [
      "glue:GetDatabase",
      "glue:CreateDatabase",
      "glue:GetDataBases",
      "glue:CreateTable",
      "glue:GetTable",
      "glue:UpdateTable",
      "glue>DeleteTable",
      "glue:GetTables",
      "glue:GetPartition",
      "glue:GetPartitions",
      "glue:CreatePartition",
      "glue:BatchCreatePartition",
      "glue:GetUserDefinedFunctions"
    ],
    "Resource": ["arn:aws:glue:region:AccountB:catalog"]
  }
]
}

```

3. Inizia il tuo job run. Questo passaggio è leggermente diverso a seconda del tipo AccountA di applicazione EMR Serverless.

### Spark

Impostate la `spark.hadoop.hive.metastore.glue.catalogid` proprietà nella `hive-site` classificazione come illustrato nell'esempio seguente. Sostituisci *AccountB-catalog-id* con l'ID del Data Catalog in AccountB.

```

aws emr-serverless start-job-run \
--application-id "application-id" \
--execution-role-arn "job-role-arn" \
--job-driver '{
  "sparkSubmit": {
    "query": "s3://amzn-s3-demo-bucket/hive/scripts/create_table.sql",
    "parameters": "--hiveconf hive.exec.scratchdir=s3://amzn-s3-demo-bucket/hive/scratch --hiveconf hive.metastore.warehouse.dir=s3://amzn-s3-demo-bucket/hive/warehouse"
  }
}' \
--configuration-overrides '{
  "applicationConfiguration": [{
    "classification": "hive-site",

```

```

        "properties": {
            "spark.hadoop.hive.metastore.glue.catalogid": "AccountB-catalog-id"
        }
    ]}
}'

```

## Hive

Imposta la `hive.metastore.glue.catalogid` proprietà nella `hive-site` classificazione come illustrato nell'esempio seguente. Sostituisci `AccountB-catalog-id` con l'ID del Data Catalog in AccountB.

```

aws emr-serverless start-job-run \
--application-id "application-id" \
--execution-role-arn "job-role-arn" \
--job-driver '{
    "hive": {
        "query": "s3://amzn-s3-demo-bucket/hive/scripts/create_table.sql",
        "parameters": "--hiveconf hive.exec.scratchdir=s3://amzn-s3-demo-bucket/hive/scratch --hiveconf hive.metastore.warehouse.dir=s3://amzn-s3-demo-bucket/hive/warehouse"
    }
}' \
--configuration-overrides '{
    "applicationConfiguration": [{
        "classification": "hive-site",
        "properties": {
            "hive.metastore.glue.catalogid": "AccountB-catalog-id"
        }
    }]
}'

```

## Considerazioni sull'utilizzo del AWS Glue Data Catalog

Puoi aggiungere elementi ausiliari `ADD JAR` negli `JARs script` di Hive. Per ulteriori considerazioni, consulta [Considerazioni sull'utilizzo di AWS Glue Data Catalog](#).

## Utilizzo di un metastore Hive esterno

Puoi configurare i job EMR Serverless Spark e Hive per connetterti a un metastore Hive esterno, come Amazon Aurora o Amazon RDS for MySQL. Questa sezione descrive come configurare un

metastore Amazon RDS Hive, configurare il VPC e configurare i job EMR Serverless per utilizzare un metastore esterno.

## Crea un metastore Hive esterno

1. [Crea un Amazon Virtual Private Cloud \(Amazon VPC\) con sottoreti private seguendo le istruzioni in Creare un VPC.](#)
2. Crea la tua applicazione EMR Serverless con il tuo nuovo Amazon VPC e le sottoreti private. Quando si configura un'applicazione EMR Serverless con un VPC, viene prima fornita un'interfaccia di rete elastica per ogni sottorete specificata. Quindi collega il gruppo di sicurezza specificato a quell'interfaccia di rete. Ciò consente il controllo dell'accesso all'applicazione. Per maggiori dettagli su come configurare il tuo VPC, consulta [Configurazione dell'accesso VPC per le applicazioni EMR Serverless per la connessione ai dati](#)
3. Crea un database MySQL o Aurora PostgreSQL in una sottorete privata nel tuo Amazon VPC. Per informazioni su come creare un database Amazon RDS, consulta [Creazione di un'istanza database Amazon RDS.](#)
4. [Modifica il gruppo di sicurezza del tuo database MySQL o Aurora per consentire le connessioni JDBC dal tuo gruppo di sicurezza EMR Serverless seguendo i passaggi descritti in Modificare un'istanza database Amazon RDS.](#) Aggiungi una regola per il traffico in entrata al gruppo di sicurezza RDS da uno dei tuoi gruppi di sicurezza EMR Serverless.

Tipo	Protocollo	Intervallo porte	Origine
Tutte le regole TCP	TCP	3306	emr-serverless-security-group

## Configura le opzioni Spark

### Usare JDBC

Per configurare la tua applicazione EMR Serverless Spark per connettersi a un metastore Hive basato su un'istanza Amazon RDS for MySQL o Amazon Aurora MySQL, usa una connessione JDBC. `mariadb-connector-java.jar--jars` Inserisci `spark-submit` i parametri del tuo job run.

```
aws emr-serverless start-job-run \
```

```

--application-id "application-id" \
--execution-role-arn "job-role-arn" \
--job-driver '{
    "sparkSubmit": {
        "entryPoint": "s3://amzn-s3-demo-bucket/scripts/spark-jdbc.py",
        "sparkSubmitParameters": "--jars s3://amzn-s3-demo-bucket/mariadb-
connector-java.jar
        --conf
spark.hadoop.javax.jdo.option.ConnectionDriverName=org.mariadb.jdbc.Driver
        --conf spark.hadoop.javax.jdo.option.ConnectionUserName=<connection-user-
name>
        --conf spark.hadoop.javax.jdo.option.ConnectionPassword=<connection-
password>
        --conf spark.hadoop.javax.jdo.option.ConnectionURL=<JDBC-Connection-
string>
        --conf spark.driver.cores=2
        --conf spark.executor.memory=10G
        --conf spark.driver.memory=6G
        --conf spark.executor.cores=4"
    }
}' \
--configuration-overrides '{
    "monitoringConfiguration": {
        "s3MonitoringConfiguration": {
            "logUri": "s3://amzn-s3-demo-bucket/spark/logs/"
        }
    }
}'

```

Il seguente esempio di codice è uno script endpoint Spark che interagisce con un metastore Hive su Amazon RDS.

```

from os.path import expanduser, join, abspath
from pyspark.sql import SparkSession
from pyspark.sql import Row
# warehouse_location points to the default location for managed databases and tables
warehouse_location = abspath('spark-warehouse')
spark = SparkSession \
    .builder \
    .config("spark.sql.warehouse.dir", warehouse_location) \
    .enableHiveSupport() \
    .getOrCreate()
spark.sql("SHOW DATABASES").show()

```

```
spark.sql("CREATE EXTERNAL TABLE `sampledb`.`sparknyctaxi`(`dispatching_base_num`
  string, `pickup_datetime` string, `dropoff_datetime` string, `pulocationid` bigint,
  `dolocationid` bigint, `sr_flag` bigint) STORED AS PARQUET LOCATION 's3://<s3 prefix>/
nyctaxi_parquet/'")
spark.sql("SELECT count(*) FROM sampledb.sparknyctaxi").show()
spark.stop()
```

## Utilizzo del servizio Thrift

Puoi configurare la tua applicazione EMR Serverless Hive per connettersi a un metastore Hive basato su un'istanza Amazon RDS for MySQL o Amazon Aurora MySQL. A tale scopo, esegui un server Thrift sul nodo master di un cluster Amazon EMR esistente. Questa opzione è ideale se disponi già di un cluster Amazon EMR con un server Thrift che desideri utilizzare per semplificare le configurazioni di lavoro EMR Serverless.

```
aws emr-serverless start-job-run \
  --application-id "application-id" \
  --execution-role-arn "job-role-arn" \
  --job-driver '{
    "sparkSubmit": {
      "entryPoint": "s3://amzn-s3-demo-bucket/thriftscript.py",
      "sparkSubmitParameters": "--jars s3://amzn-s3-demo-bucket/mariadb-
connector-java.jar
      --conf spark.driver.cores=2
      --conf spark.executor.memory=10G
      --conf spark.driver.memory=6G
      --conf spark.executor.cores=4"
    }
  }' \
  --configuration-overrides '{
    "monitoringConfiguration": {
      "s3MonitoringConfiguration": {
        "logUri": "s3://amzn-s3-demo-bucket/spark/logs/"
      }
    }
  }'
```

Il seguente esempio di codice è uno script entrypoint (`thriftscript.py`) che utilizza il protocollo thrift per connettersi a un metastore Hive. Nota che la `hive.metastore.uris` proprietà deve essere impostata per la lettura da un metastore Hive esterno.

```
from os.path import expanduser, join, abspath
```

```

from pyspark.sql import SparkSession
from pyspark.sql import Row
# warehouse_location points to the default location for managed databases and tables
warehouse_location = abspath('spark-warehouse')
spark = SparkSession \
    .builder \
    .config("spark.sql.warehouse.dir", warehouse_location) \
    .config("hive.metastore.uris", "thrift://thrift-server-host:thrift-server-port") \
    .enableHiveSupport() \
    .getOrCreate()
spark.sql("SHOW DATABASES").show()
spark.sql("CREATE EXTERNAL TABLE sampledb.`sparknyctaxi`(`dispatching_base_num`
  string, `pickup_datetime` string, `dropoff_datetime` string, `pulocationid` bigint,
  `dolocationid` bigint, `sr_flag` bigint) STORED AS PARQUET LOCATION 's3://<s3 prefix>/
nyctaxi_parquet/'")
spark.sql("SELECT * FROM sampledb.sparknyctaxi").show()
spark.stop()

```

## Configura le opzioni Hive

### Usare JDBC

Se desideri specificare una posizione di database Hive esterna su un'istanza Amazon RDS MySQL o Amazon Aurora, puoi sovrascrivere la configurazione predefinita del metastore.

#### Note

In Hive, puoi eseguire più scritture su tabelle metastore contemporaneamente. Se condivide le informazioni sui metastore tra due job, assicuratevi di non scrivere contemporaneamente sulla stessa tabella di metastore a meno che non scriviate su partizioni diverse della stessa tabella di metastore.

Imposta le seguenti configurazioni nella classificazione per attivare il metastore Hive esterno `hive-site`.

```

{
  "classification": "hive-site",
  "properties": {
    "hive.metastore.client.factory.class":
    "org.apache.hadoop.hive.ql.metadata.SessionHiveMetaStoreClientFactory",

```

```

    "javax.jdo.option.ConnectionDriverName": "org.mariadb.jdbc.Driver",
    "javax.jdo.option.ConnectionURL": "jdbc:mysql://db-host:db-port/db-name",
    "javax.jdo.option.ConnectionUserName": "username",
    "javax.jdo.option.ConnectionPassword": "password"
  }
}

```

### Utilizzo di un server parsimonioso

Puoi configurare la tua applicazione EMR Serverless Hive per connettersi a un metastore Hive basato su Amazon RDS for MySQL o Amazon Aurora My. SQLInstance A tale scopo, esegui un server Thrift sul nodo principale di un cluster Amazon EMR esistente. Questa opzione è ideale se disponi già di un cluster Amazon EMR che esegue un server Thrift e desideri utilizzare le configurazioni di lavoro EMR Serverless.

Imposta le seguenti configurazioni nella `hive-site` classificazione in modo che EMR Serverless possa accedere al thrift metastore remoto. Si noti che è necessario impostare la `hive.metastore.uris` proprietà per la lettura da un metastore Hive esterno.

```

{
  "classification": "hive-site",
  "properties": {
    "hive.metastore.client.factory.class":
    "org.apache.hadoop.hive.ql.metadata.SessionHiveMetaStoreClientFactory",
    "hive.metastore.uris": "thrift://thrift-server-host:thrift-server-port"
  }
}

```

### Utilizzo della gerarchia AWS multicatalogo di Glue su EMR Serverless

È possibile configurare le applicazioni EMR Serverless in modo che funzionino con la gerarchia multicatalogo AWS Glue. L'esempio seguente mostra come utilizzare EMR-S Spark con la gerarchia multicatalogo AWS Glue.

Per ulteriori informazioni sulla gerarchia multicatalogo, consulta [Lavorare con una gerarchia multicatalogo in AWS Glue Data Catalog with Spark su Amazon EMR](#).

### Utilizzo di Redshift Managed Storage (RMS) con Iceberg e Glue Data Catalog AWS

Di seguito viene illustrato come configurare Spark per l'integrazione con un AWS Glue Data Catalog con Iceberg:

```
aws emr-serverless start-job-run \
  --application-id application-id \
  --execution-role-arn job-role-arn \
  --job-driver '{
    "sparkSubmit": {
      "entryPoint": "s3://amzn-s3-demo-bucket/myscript.py",
      "sparkSubmitParameters": "--conf spark.sql.catalog.nfgac_rms =
org.apache.iceberg.spark.SparkCatalog
      --conf spark.sql.catalog.rms.type=glue
      --conf spark.sql.catalog.rms.glue.id=Glue RMS catalog ID
      --conf spark.sql.defaultCatalog=rms
      --conf
spark.sql.extensions=org.apache.iceberg.spark.extensions.IcebergSparkSessionExtensions"
    }
  }'
```

Una query di esempio da una tabella del catalogo, dopo l'integrazione:

```
SELECT * FROM my_rms_schema.my_table
```

## Utilizzo di Redshift Managed Storage (RMS) con l'API REST di Iceberg e Glue Data Catalog AWS

Di seguito viene mostrato come configurare Spark per funzionare con il catalogo REST di Iceberg:

```
aws emr-serverless start-job-run \
  --application-id application-id \
  --execution-role-arn job-role-arn \
  --job-driver '{
"sparkSubmit": {
"entryPoint": "s3://amzn-s3-demo-bucket/myscript.py",
  "sparkSubmitParameters": "
  --conf spark.sql.catalog.rms=org.apache.iceberg.spark.SparkCatalog
  --conf spark.sql.catalog.rms.type=rest
  --conf spark.sql.catalog.rms.warehouse=Glue RMS catalog ID
  --conf spark.sql.catalog.rms.uri=Glue endpoint URI/iceberg
  --conf spark.sql.catalog.rms.rest.sigv4-enabled=true
  --conf spark.sql.catalog.rms.rest.signing-name=glue
  --conf
spark.sql.extensions=org.apache.iceberg.spark.extensions.IcebergSparkSessionExtensions"
  }
}
```

```
}'
```

Un esempio di query da una tabella del catalogo:

```
SELECT * FROM my_rms_schema.my_table
```

## Considerazioni sull'utilizzo di un metastore esterno

- Puoi configurare database compatibili con Mariadb JDBC come metastore. Esempi di questi database sono RDS per Mariadb, MySQL e Amazon Aurora.
- I metastore non vengono inizializzati automaticamente. [Se il tuo metastore non è inizializzato con uno schema per la tua versione di Hive, usa lo strumento Hive Schema.](#)
- EMR Serverless non supporta l'autenticazione Kerberos. Non è possibile utilizzare un server metastore Thrift con autenticazione Kerberos con i job EMR Serverless Spark o Hive.
- È necessario configurare l'accesso VPC per utilizzare la gerarchia multicatalogo.

## Accesso ai dati S3 in un altro AWS account da EMR Serverless

Puoi eseguire job Serverless di Amazon EMR da un AWS account e configurarli per accedere ai dati nei bucket Amazon S3 che appartengono a un altro account. AWS Questa pagina descrive come configurare l'accesso tra account a S3 da EMR Serverless.

I lavori eseguiti su EMR Serverless possono utilizzare una policy di bucket S3 o un ruolo presunto per accedere ai dati in Amazon S3 da un altro account. AWS

### Prerequisiti

Per configurare l'accesso tra più account per Amazon EMR Serverless, devi completare le attività mentre sei connesso a due account: AWS

- **AccountA**— Questo è l' AWS account su cui hai creato un'applicazione Amazon EMR Serverless. Prima di configurare l'accesso tra più account, devi avere a disposizione quanto segue in questo account:
  - Un'applicazione Serverless Amazon EMR in cui eseguire lavori.
  - Un ruolo di esecuzione del lavoro che dispone delle autorizzazioni necessarie per eseguire i lavori nell'applicazione. Per ulteriori informazioni, consulta [Ruoli Job Runtime per Amazon EMR Serverless.](#)

- **AccountB**— Questo è l' AWS account che contiene il bucket S3 a cui desideri accedere per i tuoi job Serverless di Amazon EMR.

## Utilizza una policy sui bucket S3 per accedere ai dati S3 tra account diversi

Per accedere al bucket S3 in account B from account A, collega la seguente policy al bucket S3 in account B.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "Example permissions 1",
      "Effect": "Allow",
      "Principal": {
        "AWS": "arn:aws:iam::AccountA:root"
      },
      "Action": [
        "s3:ListBucket"
      ],
      "Resource": [
        "arn:aws:s3::bucket_name_in_AccountB"
      ]
    },
    {
      "Sid": "Example permissions 2",
      "Effect": "Allow",
      "Principal": {
        "AWS": "arn:aws:iam::AccountA:root"
      },
      "Action": [
        "s3:PutObject",
        "s3:GetObject",
        "s3:DeleteObject"
      ],
      "Resource": [
        "arn:aws:s3::bucket_name_in_AccountB/*"
      ]
    }
  ]
}
```

Per ulteriori informazioni sull'accesso a più account S3 con le policy dei bucket S3, consulta [l'Esempio 2: Il proprietario del bucket concede le autorizzazioni per i bucket tra account nella Guida per l'utente di Amazon Simple Storage Service](#).

## Usa un ruolo presunto per accedere ai dati S3 tra più account

Un altro modo per configurare l'accesso tra account per Amazon EMR Serverless consiste nell'eseguire `AssumeRole` l'azione di AWS Security Token Service (`sts.amazonaws.com`). AWS STS è un servizio web globale che consente di richiedere credenziali temporanee con privilegi limitati per gli utenti. Puoi effettuare chiamate API verso EMR Serverless e Amazon S3 con le credenziali di sicurezza temporanee con cui crei `AssumeRole`.

I passaggi seguenti illustrano come utilizzare un ruolo presunto per accedere ai dati S3 tra account diversi da EMR Serverless:

1. Crea un bucket Amazon S3, *cross-account-bucket*, in AccountB. Per ulteriori informazioni, consulta [Creare un bucket](#) nella Guida per l'utente di Amazon Simple Storage Service. Se si desidera avere un accesso multi-account a DynamoDB, è anche possibile creare una tabella DynamoDB in AccountB. Per ulteriori informazioni, consulta [Creazione di una tabella DynamoDB nella Amazon DynamoDB Developer Guide](#).
2. Crea un ruolo IAM `Cross-Account-Role-B` in AccountB per accedere a *cross-account-bucket*.
  - a. Accedi AWS Management Console e apri la console IAM all'indirizzo. <https://console.aws.amazon.com/iam/>
  - b. Scegli Roles (Ruoli), quindi crea un nuovo ruolo: `Cross-Account-Role-B`. Per ulteriori informazioni su come creare ruoli IAM, consulta [Creazione di ruoli IAM](#) nella Guida per l'utente di IAM.
  - c. Crea una policy IAM che specifichi le autorizzazioni per `Cross-Account-Role-B` per accedere al bucket S3 *cross-account-bucket*, come dimostra la seguente istruzione di policy. Quindi, allega la policy IAM a `Cross-Account-Role-B`. Per ulteriori informazioni, consulta [Creazione di policy IAM](#) nella Guida per l'utente di IAM.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": "s3:*",
```

```

    "Resource": [
      "arn:aws:s3:::cross-account-bucket",
      "arn:aws:s3:::cross-account-bucket/*"
    ]
  }
]
}

```

Se hai bisogno dell'accesso a DynamoDB, crea una policy IAM che specifichi le autorizzazioni per accedere alla tabella DynamoDB tra account. Quindi, allega la policy IAM a Cross-Account-Role-B. Per ulteriori informazioni, consulta [Amazon DynamoDB: consente l'accesso a una tabella specifica](#) nella IAM User Guide.

Di seguito è riportata una politica per consentire l'accesso alla tabella DynamoDBCrossAccountTable.

```

{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": "dynamodb:*",
      "Resource": "arn:aws:dynamodb:MyRegion:AccountB:table/CrossAccountTable"
    }
  ]
}

```

### 3. Modifica la relazione di fiducia per il ruolo Cross-Account-Role-B.

- a. Per configurare la relazione di trust per il ruolo, scegli la scheda Trust Relationships nella console IAM per il ruolo Cross-Account-Role-B che hai creato nel passaggio 2.
- b. Seleziona Edit Trust Relationship (Modifica relazione di fiducia).
- c. Aggiungi il seguente documento di policy. Ciò consente Job-Execution-Role-A AccountA di assumere il Cross-Account-Role-B ruolo.

```

{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Principal": {

```

```

    "AWS": "arn:aws:iam::AccountA:role/Job-Execution-Role-A"
  },
  "Action": "sts:AssumeRole"
}
]
}

```

4. Concedi Job-Execution-Role-A AccountA il AWS STS AssumeRole permesso di assumereCross-Account-Role-B.

- a. Nella console IAM per l' AWS accountAccountA, selezionaJob-Execution-Role-A.
- b. Aggiungi la seguente istruzione di policy a Job-Execution-Role-A per autorizzare l'operazione AssumeRole nel ruolo Cross-Account-Role-B.

```

{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": "sts:AssumeRole",
      "Resource": "arn:aws:iam::AccountB:role/Cross-Account-Role-B"
    }
  ]
}

```

## Esempi di ruoli presunti

Puoi utilizzare un singolo ruolo assunto per accedere a tutte le risorse S3 in un account oppure, con Amazon EMR 6.11 e versioni successive, puoi configurare più ruoli IAM da assumere quando accedi a diversi bucket S3 tra account diversi.

### Argomenti

- [Accedi alle risorse S3 con un solo ruolo](#)
- [Accedi alle risorse S3 con più ruoli presunti](#)

## Accedi alle risorse S3 con un solo ruolo

### Note

Quando configuri un lavoro per utilizzare un singolo ruolo assunto, tutte le risorse S3 del job utilizzano quel ruolo, incluso lo `entryPoint` script.

Se desideri utilizzare un singolo ruolo assunto per accedere a tutte le risorse S3 nell'account B, specifica le seguenti configurazioni:

1. Specificare la configurazione EMRFS per. `fs.s3.customAWSCredentialsProvider`  
`spark.hadoop.fs.s3.customAWSCredentialsProvider=com.amazonaws.emr.AssumeRoleAW`
2. Per Spark, usa `spark.emr-serverless.driverEnv.ASSUME_ROLE_CREDENTIALS_ROLE_ARN` e specifica le variabili `spark.executorEnv.ASSUME_ROLE_CREDENTIALS_ROLE_ARN` di ambiente su driver ed executor.
3. Per Hive `hive.emr-serverless.launch.env.ASSUME_ROLE_CREDENTIALS_ROLE_ARN` e `tez.am.emr-serverless.launch.env.ASSUME_ROLE_CREDENTIALS_ROLE_ARN`, usa e specifica le variabili `tez.task.emr-serverless.launch.env.ASSUME_ROLE_CREDENTIALS_ROLE_ARN` di ambiente nei contenitori di attività Hive driver, Tez application master e Tez.

Gli esempi seguenti mostrano come utilizzare un ruolo presunto per avviare un processo EMR Serverless con accesso tra account.

### Spark

L'esempio seguente mostra come utilizzare un ruolo assunto per avviare un job EMR Serverless Spark eseguito con accesso multiaccount a S3.

```
aws emr-serverless start-job-run \
  --application-id application-id \
  --execution-role-arn job-role-arn \
  --job-driver '{
    "sparkSubmit": {
      "entryPoint": "entrypoint_location",
```

```

    "entryPointArguments": [":argument_1:", ":argument_2:"],
    "sparkSubmitParameters": "--conf spark.executor.cores=4 --conf
spark.executor.memory=20g --conf spark.driver.cores=4 --conf spark.driver.memory=8g
--conf spark.executor.instances=1"
  }
}' \
--configuration-overrides '{
  "applicationConfiguration": [{
    "classification": "spark-defaults",
    "properties": {
      "spark.hadoop.fs.s3.customAWSCredentialsProvider":
"spark.hadoop.fs.s3.customAWSCredentialsProvider=com.amazonaws.emr.AssumeRoleAWSCredentials
      "spark.emr-serverless.driverEnv.ASSUME_ROLE_CREDENTIALS_ROLE_ARN":
"arn:aws:iam::AccountB:role/Cross-Account-Role-B",
      "spark.executorEnv.ASSUME_ROLE_CREDENTIALS_ROLE_ARN":
"arn:aws:iam::AccountB:role/Cross-Account-Role-B"
    }
  }]
}'

```

## Hive

L'esempio seguente mostra come utilizzare un ruolo presunto per avviare un job EMR Serverless Hive eseguito con accesso multiaccount a S3.

```

aws emr-serverless start-job-run \
  --application-id application-id \
  --execution-role-arn job-role-arn \
  --job-driver '{
    "hive": {
      "query": "query_location",
      "parameters": "hive_parameters"
    }
  }' \
  --configuration-overrides '{
    "applicationConfiguration": [{
      "classification": "hive-site",
      "properties": {
        "fs.s3.customAWSCredentialsProvider":
"com.amazonaws.emr.serverless.credentialsprovider.AssumeRoleAWSCredentialsProvider",
        "hive.emr-serverless.launch.env.ASSUME_ROLE_CREDENTIALS_ROLE_ARN":
"arn:aws:iam::AccountB:role/Cross-Account-Role-B",

```

```

        "tez.am.emr-serverless.launch.env.ASSUME_ROLE_CREDENTIALS_ROLE_ARN":
    "arn:aws:iam::AccountB:role/Cross-Account-Role-B",
        "tez.task.emr-
serverless.launch.env.ASSUME_ROLE_CREDENTIALS_ROLE_ARN":
    "arn:aws:iam::AccountB:role/Cross-Account-Role-B"
    }
    ]]
}'

```

## Accedi alle risorse S3 con più ruoli presunti

Con le versioni 6.11.0 e successive di EMR Serverless, puoi configurare più ruoli IAM da assumere quando accedi a diversi bucket tra account. Se desideri accedere a diverse risorse S3 con diversi ruoli assunti nell'account B, utilizza le seguenti configurazioni all'avvio del job run:

1. Specificare la configurazione EMRFS per `fs.s3.customAWSCredentialsProvider` `com.amazonaws.emr.serverless.credentialsprovider.BucketLevelAssumeRoleCredenti`
2. Specificate la configurazione EMRFS `fs.s3.bucketLevelAssumeRoleMapping` per definire la mappatura dal nome del bucket S3 al ruolo IAM nell'account B da assumere. Il valore deve essere nel formato di `bucket1->role1;bucket2->role2`

Ad esempio, è possibile utilizzare `arn:aws:iam::AccountB:role/Cross-Account-Role-B-1` per accedere al bucket `bucket1` e utilizzare `arn:aws:iam::AccountB:role/Cross-Account-Role-B-2` per accedere al `bucket2` bucket. Gli esempi seguenti mostrano come avviare l'esecuzione di un job EMR Serverless con accesso tra più account tramite più ruoli presunti.

### Spark

L'esempio seguente mostra come utilizzare più ruoli presunti per creare un'esecuzione di job EMR Serverless Spark.

```

aws emr-serverless start-job-run \
  --application-id application-id \
  --execution-role-arn job-role-arn \
  --job-driver '{
    "sparkSubmit": {
      "entryPoint": "entrypoint_location",
      "entryPointArguments": [":argument_1:", ":argument_2:"],

```

```

        "sparkSubmitParameters": "--conf spark.executor.cores=4 --conf
spark.executor.memory=20g --conf spark.driver.cores=4 --conf spark.driver.memory=8g
--conf spark.executor.instances=1"
    }
}' \
--configuration-overrides '{
  "applicationConfiguration": [{
    "classification": "spark-defaults",
    "properties": {
      "spark.hadoop.fs.s3.customAWSCredentialsProvider":
"com.amazonaws.emr.serverless.credentialsprovider.BucketLevelAssumeRoleCredentialsProvider"
      "spark.hadoop.fs.s3.bucketLevelAssumeRoleMapping":
"bucket1->arn:aws:iam::AccountB:role/Cross-Account-Role-B-1;bucket2-
>arn:aws:iam::AccountB:role/Cross-Account-Role-B-2"
    }
  }]
}'

```

## Hive

Gli esempi seguenti mostrano come utilizzare più ruoli presunti per creare un'esecuzione di job EMR Serverless Hive.

```

aws emr-serverless start-job-run \
  --application-id application-id \
  --execution-role-arn job-role-arn \
  --job-driver '{
    "hive": {
      "query": "query_location",
      "parameters": "hive_parameters"
    }
  }' \
  --configuration-overrides '{
    "applicationConfiguration": [{
      "classification": "hive-site",
      "properties": {
        "fs.s3.customAWSCredentialsProvider":
"com.amazonaws.emr.serverless.credentialsprovider.AssumeRoleAWSCredentialsProvider",
        "fs.s3.bucketLevelAssumeRoleMapping": "bucket1-
>arn:aws:iam::AccountB:role/Cross-Account-Role-B-1;bucket2-
>arn:aws:iam::AccountB:role/Cross-Account-Role-B-2"
      }
    }]
  }'

```

```
}'
```

## Risoluzione degli errori in EMR Serverless

Utilizza le seguenti informazioni per diagnosticare e risolvere problemi comuni che potresti riscontrare quando lavori con Amazon EMR Serverless.

### Argomenti

- [Errore: Job non riuscito poiché l'account ha raggiunto il limite di servizio sulla vCPU massima che può utilizzare contemporaneamente.](#)
- [Errore: Job non riuscito perché l'applicazione ha superato le impostazioni di MaximumCapacity.](#)
- [Errore: Job non riuscito a causa dell'impossibilità di allocare Worker poiché l'applicazione ha superato la capacità massima.](#)
- [Errore: l'accesso a S3 è negato. Controlla le autorizzazioni di accesso S3 del ruolo Job Runtime sulle risorse S3 richieste.](#)
- [Errore ModuleNotFoundError: nessun modulo denominato <module>. Consulta la guida per l'utente su come utilizzare le librerie python con EMR Serverless.](#)
- [Errore: impossibile assumere il ruolo di esecuzione <role name> perché non esiste o non è configurato con la relazione di trust richiesta.](#)

**Errore: Job non riuscito poiché l'account ha raggiunto il limite di servizio sulla vCPU massima che può utilizzare contemporaneamente.**

Questo errore indica che EMR Serverless non è riuscito a inviare il lavoro poiché l'account ha superato la capacità massima. Aumentare la capacità massima dell'account. Verifica i limiti del servizio nelle quote di servizio [EMR Serverless](#).

**Errore: Job non riuscito perché l'applicazione ha superato le impostazioni di MaximumCapacity.**

Questo errore indica che EMR Serverless non è riuscito a inviare il lavoro poiché l'applicazione ha superato la capacità massima configurata. Aumentare la capacità massima dell'applicazione.

**Errore: Job non riuscito a causa dell'impossibilità di allocare Worker poiché l'applicazione ha superato la capacità massima.**

Questo errore indica che il processo non è stato completato. Impossibile allocare i lavoratori perché l'applicazione ha superato le impostazioni di MaximumCapacity.

**Errore: l'accesso a S3 è negato. Controlla le autorizzazioni di accesso S3 del ruolo Job Runtime sulle risorse S3 richieste.**

Questo errore indica che il tuo job non ha accesso alle tue risorse S3. Verifica che il ruolo Job Runtime sia autorizzato ad accedere alle risorse S3 che il job deve utilizzare. Per ulteriori informazioni sui ruoli di runtime, consulta [Ruoli Job Runtime per Amazon EMR Serverless](#).

**Errore ModuleNotFoundError: nessun modulo denominato <module>.**  
**Consulta la guida per l'utente su come utilizzare le librerie python con EMR Serverless.**

Questo errore indica che un modulo Python non era disponibile per il job Spark. Verifica che le librerie Python dipendenti siano disponibili per il lavoro. Per ulteriori informazioni su come impacchettare le librerie Python, vedere. [Utilizzo delle librerie Python con EMR Serverless](#)

**Errore: impossibile assumere il ruolo di esecuzione <role name> perché non esiste o non è configurato con la relazione di trust richiesta.**

Questo errore indica che il ruolo di esecuzione del lavoro specificato per il lavoro non esiste o che il ruolo non ha una relazione di trust per le autorizzazioni EMR Serverless. Per verificare l'esistenza del ruolo IAM e verificare che la policy di fiducia del ruolo sia stata impostata correttamente, consulta le istruzioni riportate in. [Ruoli Job Runtime per Amazon EMR Serverless](#)

# Esegui carichi di lavoro interattivi con EMR Serverless tramite EMR Studio

Con le applicazioni interattive EMR Serverless, puoi eseguire carichi di lavoro interattivi per Spark con EMR Serverless utilizzando notebook ospitati in EMR Studio.

## Panoramica

Un'applicazione interattiva è un'applicazione EMR Serverless con funzionalità interattive abilitate. Con le applicazioni interattive Amazon EMR Serverless, puoi eseguire carichi di lavoro interattivi con notebook Jupyter gestiti in Amazon EMR Studio. Questo aiuta i data engineer, i data scientist e gli analisti di dati a utilizzare EMR Studio per eseguire analisi interattive con set di dati in archivi di dati come Amazon S3 e Amazon DynamoDB.

I casi d'uso per le applicazioni interattive in EMR Serverless includono quanto segue:

- I data engineer utilizzano l'esperienza IDE di EMR Studio per creare uno script ETL. Lo script acquisisce i dati dall'ambiente locale, li trasforma per l'analisi e li archivia in Amazon S3.
- I data scientist utilizzano i notebook per esplorare i set di dati e addestrare modelli di apprendimento automatico (ML) per rilevare anomalie nei set di dati.
- Gli analisti di dati esplorano i set di dati e creano script che generano report giornalieri per aggiornare applicazioni come i dashboard aziendali.

## Prerequisiti

Per utilizzare carichi di lavoro interattivi con EMR Serverless, è necessario soddisfare i seguenti requisiti:

- Le applicazioni interattive EMR Serverless sono supportate con Amazon EMR 6.14.0 e versioni successive.
- Per accedere all'applicazione interattiva, eseguire i carichi di lavoro inviati ed eseguire notebook interattivi da EMR Studio, sono necessari permessi e ruoli specifici. Per ulteriori informazioni, consulta [Autorizzazioni richieste per i carichi di lavoro interattivi](#).

## Autorizzazioni richieste per i carichi di lavoro interattivi

Oltre alle [autorizzazioni di base necessarie per accedere a EMR Serverless](#), è necessario configurare autorizzazioni aggiuntive per l'identità o il ruolo IAM:

Per accedere alla tua applicazione interattiva

Configura le autorizzazioni utente e Workspace per EMR Studio. Per ulteriori informazioni, consulta [Configurare le autorizzazioni utente di EMR Studio](#) nella Amazon EMR Management Guide.

Per eseguire i carichi di lavoro inviati con EMR Serverless

Imposta un ruolo di job runtime. Per ulteriori informazioni, consulta [Creare un ruolo Job Runtime](#).

Per eseguire i taccuini interattivi da EMR Studio

Aggiungi le seguenti autorizzazioni aggiuntive alla policy IAM per gli utenti di Studio:

- **emr-serverless:AccessInteractiveEndpoints**- Concede l'autorizzazione per accedere e connettersi all'applicazione interattiva specificata come. Resource Questa autorizzazione è necessaria per collegarsi a un'applicazione EMR Serverless da un EMR Studio Workspace.
- **iam:PassRole**- Concede l'autorizzazione ad accedere al ruolo di esecuzione IAM che intendi utilizzare quando ti colleghi a un'applicazione. È richiesta l'`PassRole` autorizzazione appropriata per collegarsi a un'applicazione EMR Serverless da un EMR Studio Workspace.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "EMRServerlessInteractiveAccess",
      "Effect": "Allow",
      "Action": "emr-serverless:AccessInteractiveEndpoints",
      "Resource": "arn:aws:emr-serverless:Region:account:/applications/*"
    },
    {
      "Sid": "EMRServerlessRuntimeRoleAccess",
      "Effect": "Allow",
      "Action": "iam:PassRole",
      "Resource": "arn:aws:iam:Region:account:role/interactive-execution-role-ARN",
      "Condition": {
        "StringLike": {
          "iam:PassedToService": "emr-serverless.amazonaws.com"
        }
      }
    }
  ]
}
```

```
}  
  }  
} ]  
}
```

## Configurazione di applicazioni interattive

Utilizza i seguenti passaggi di alto livello per creare un'applicazione EMR Serverless con funzionalità interattive di Amazon EMR Studio in AWS Management Console

1. Segui i passaggi indicati per creare un'applicazione [Guida introduttiva ad Amazon EMR Serverless](#).
2. Quindi, avvia un'area di lavoro da EMR Studio e collegala a un'applicazione EMR Serverless come opzione di elaborazione. Per ulteriori informazioni, vedere la scheda Carico di lavoro interattivo nella fase 2 della documentazione introduttiva di [EMR Serverless](#).

Quando si collega un'applicazione a Studio Workspace, l'avvio dell'applicazione si attiva automaticamente se non è già in esecuzione. È inoltre possibile preavviare l'applicazione e tenerla pronta prima di collegarla al Workspace.

## Considerazioni sulle applicazioni interattive

- Le applicazioni interattive EMR Serverless sono supportate con Amazon EMR 6.14.0 e versioni successive.
- EMR Studio è l'unico client integrato con le applicazioni interattive EMR Serverless. Le seguenti funzionalità di EMR Studio non sono supportate con le applicazioni interattive EMR Serverless: collaborazione Workspace, SQL Explorer ed esecuzione programmata di notebook.
- Le applicazioni interattive sono supportate solo per il motore Spark.
- Le applicazioni interattive supportano i kernel Python 3 PySpark e Spark Scala.
- È possibile eseguire fino a 25 notebook simultanei su una singola applicazione interattiva.
- Non esiste un'interfaccia endpoint o API che supporti notebook Jupyter ospitati autonomamente con applicazioni interattive.

- Per un'esperienza di avvio ottimizzata, consigliamo di configurare la capacità preinizializzata per driver ed esecutori e di preavviare l'applicazione. Quando preavvii l'applicazione, ti assicuri che sia pronta quando desideri collegarla al tuo Workspace.

```
aws emr-serverless start-application \  
--application-id your-application-id
```

- Per impostazione predefinita, `autoStopConfig` è abilitato per le applicazioni. Questa operazione chiude l'applicazione dopo 30 minuti di inattività. Puoi modificare questa configurazione come parte della tua `requestcreate-application.update-application`
- Quando si utilizza un'applicazione interattiva, si consiglia di configurare una capacità preinizializzata di kernel, driver ed esecutori per far funzionare i notebook. Ogni sessione interattiva Spark richiede un kernel e un driver, quindi EMR Serverless mantiene un kernel worker preinizializzato per ogni driver preinizializzato. Per impostazione predefinita, EMR Serverless mantiene una capacità preinizializzata di un kernel worker per l'intera applicazione anche se non si specifica alcuna capacità preinizializzata per i driver. Ogni kernel worker utilizza 4 vCPU e 16 GB di memoria. Per informazioni aggiornate sui prezzi, consulta la pagina dei [prezzi di Amazon EMR](#).
- È necessario disporre di una quota di servizi vCPU sufficiente per Account AWS eseguire carichi di lavoro interattivi. Se non esegui carichi di lavoro compatibili con Lake Formation, ti consigliamo almeno 24 vCPU. In tal caso, consigliamo almeno 28 vCPU.
- EMR Serverless interrompe automaticamente i kernel dai notebook se sono rimasti inattivi per più di 60 minuti. EMR Serverless calcola il tempo di inattività del kernel dall'ultima attività completata durante la sessione del notebook. Al momento non è possibile modificare l'impostazione del timeout di inattività del kernel.
- Per abilitare Lake Formation con carichi di lavoro interattivi, imposta la configurazione `spark.emr-serverless.lakeformation.enabled` su `true` sotto la `spark-defaults` classificazione nell'`runtime-configuration` oggetto quando [crei un'applicazione EMR Serverless](#). Per ulteriori informazioni sull'abilitazione di Lake Formation in EMR Serverless, consulta [Enabling Lake Formation in Amazon EMR](#).

## Esegui carichi di lavoro interattivi con EMR Serverless tramite un endpoint Apache Livy

Con le versioni 6.14.0 e successive di Amazon EMR, puoi creare e abilitare un endpoint Apache Livy mentre crei un'applicazione EMR Serverless ed eseguire carichi di lavoro interattivi tramite notebook

ospitati autonomamente o con un client personalizzato. Un endpoint Apache Livy offre i seguenti vantaggi:

- Puoi connetterti in modo sicuro a un endpoint Apache Livy tramite i notebook Jupyter e gestire i carichi di lavoro Apache Spark con l'interfaccia REST di Apache Livy.
- Utilizza le operazioni dell'API REST di Apache Livy per applicazioni Web interattive che utilizzano i dati dei carichi di lavoro Apache Spark.

## Prerequisiti

Per utilizzare un endpoint Apache Livy con EMR Serverless, è necessario soddisfare i seguenti requisiti:

- Completa i passaggi descritti in [Guida introduttiva ad Amazon EMR Serverless](#).
- Per eseguire carichi di lavoro interattivi tramite gli endpoint Apache Livy, sono necessari determinati permessi e ruoli. Per ulteriori informazioni, consulta Autorizzazioni [richieste](#) per carichi di lavoro interattivi.

## Autorizzazioni richieste

Oltre alle autorizzazioni necessarie per accedere a EMR Serverless, devi aggiungere anche le seguenti autorizzazioni al tuo ruolo IAM per accedere a un endpoint Apache Livy ed eseguire applicazioni:

- `emr-serverless:AccessLivyEndpoints`— concede l'autorizzazione per accedere e connettersi all'applicazione abilitata per Livy specificata come `Resource`. È necessaria questa autorizzazione per eseguire le operazioni dell'API REST disponibili dall'endpoint Apache Livy.
- `iam:PassRole`— concede l'autorizzazione ad accedere al ruolo di esecuzione IAM durante la creazione della sessione Apache Livy. EMR Serverless utilizzerà questo ruolo per eseguire i carichi di lavoro.
- `emr-serverless:GetDashboardForJobRun`— concede l'autorizzazione a generare l'interfaccia utente di Spark Live e i link ai registri dei driver e fornisce l'accesso ai log come parte dei risultati della sessione di Apache Livy.

```
{  
  "Version": "2012-10-17",
```

```

"Statement": [{
  "Sid": "EMRServerlessInteractiveAccess",
  "Effect": "Allow",
  "Action": "emr-serverless:AccessLivyEndpoints",
  "Resource": "arn:aws:emr-serverless:<AWS_REGION>:account:/applications/*"
},
{
  "Sid": "EMRServerlessRuntimeRoleAccess",
  "Effect": "Allow",
  "Action": "iam:PassRole",
  "Resource": "execution-role-ARN",
  "Condition": {
    "StringLike": {
      "iam:PassedToService": "emr-serverless.amazonaws.com"
    }
  }
},
{
  "Sid": "EMRServerlessDashboardAccess",
  "Effect": "Allow",
  "Action": "emr-serverless:GetDashboardForJobRun",
  "Resource": "arn:aws:emr-serverless:<AWS_REGION>:account:/applications/*"
}
]
}

```

## Nozioni di base

Per creare un'applicazione compatibile con Apache Livy ed eseguirla, segui questi passaggi.

1. Per creare un'applicazione compatibile con Apache Livy, esegui il comando seguente.

```

aws emr-serverless create-application \
--name my-application-name \
--type 'application-type' \
--release-label <Amazon EMR-release-version>
--interactive-configuration '{"livyEndpointEnabled": true}'

```

2. Dopo che EMR Serverless ha creato l'applicazione, avviala per rendere disponibile l'endpoint Apache Livy.

```

aws emr-serverless start-application \

```

```
--application-id application-id
```

Utilizzate il seguente comando per verificare lo stato della vostra applicazione. Una volta raggiunto lo stato `STARTED`, puoi accedere all'endpoint Apache Livy.

```
aws emr-serverless get-application \  
--region <AWS_REGION> --application-id >application_id<
```

3. Utilizza il seguente URL per accedere all'endpoint:

```
https://_<application-id>_.livy.emr-serverless-  
services._<AWS_REGION>_.amazonaws.com
```

Una volta che l'endpoint è pronto, puoi inviare carichi di lavoro in base al tuo caso d'uso. È necessario firmare ogni richiesta all'endpoint con [il SIGv4 protocollo](#) e inserire un'intestazione di autorizzazione. È possibile utilizzare i seguenti metodi per eseguire carichi di lavoro:

- Client HTTP: è necessario inviare le operazioni dell'API degli endpoint Apache Livy con un client HTTP personalizzato.
- Kernel Sparkmagic: è necessario eseguire localmente il kernel sparkmagic e inviare query interattive con i notebook Jupyter.

## Client HTTP

Per creare una sessione di Apache Livy, devi inserire il parametro del corpo della richiesta. `emr-serverless.session.executionRoleArn` conf L'esempio seguente è un esempio POST / sessions di richiesta.

```
{  
  "kind": "pyspark",  
  "heartbeatTimeoutInSeconds": 60,  
  "conf": {  
    "emr-serverless.session.executionRoleArn": "<executionRoleArn>"  
  }  
}
```

La tabella seguente descrive tutte le operazioni dell'API Apache Livy disponibili.

Operazione API	Descrizione
GET /sessions	Restituisce un elenco di tutte le sessioni interattive attive.
POST /sessioni	Crea una nuova sessione interattiva tramite spark o pyspark.
GET /sessions/ < > <i>sessionId</i>	Restituisce le informazioni sulla sessione.
GET /sessions/ < >/state <i>sessionId</i>	Restituisce lo stato della sessione.
ELIMINA /sessions/ < > <i>sessionId</i>	Interrompe ed elimina la sessione.
GET /sessions/ < >/statements <i>sessionId</i>	Restituisce tutte le istruzioni di una sessione.
POST /sessions/ < >/dichiarazioni <i>sessionId</i>	Esegue un'istruzione in una sessione.
GET /sessions/ < >/statements/< > <i>sessionId</i> <i>statementId</i>	Restituisce i dettagli dell'istruzione specificata in una sessione.
POST /sessions/ < >/statements/< >/cancel <i>sessionId</i> <i>statementId</i>	Annulla l'istruzione specificata in questa sessione.

### Invio di richieste all'endpoint Apache Livy

Puoi anche inviare richieste direttamente all'endpoint Apache Livy da un client HTTP. In questo modo è possibile eseguire in remoto il codice per i casi d'uso al di fuori di un notebook.

Prima di iniziare a inviare richieste all'endpoint, assicurati di aver installato le seguenti librerie:

```
pip3 install botocore awscli requests
```

Di seguito è riportato un esempio di script Python per inviare richieste HTTP direttamente a un endpoint:

```
from botocore import crt
import requests
from botocore.awsrequest import AWSRequest
```

```
from botocore.credentials import Credentials
import botocore.session
import json, pprint, textwrap

endpoint = 'https://<application_id>.livy.emr-serverless-
services-<AWS_REGION>.amazonaws.com'
headers = {'Content-Type': 'application/json'}

session = botocore.session.Session()
signer = crt.auth.CrtS3SigV4Auth(session.get_credentials(), 'emr-serverless',
    '<AWS_REGION>')

### Create session request

data = {'kind': 'pyspark', 'heartbeatTimeoutInSecond': 60, 'conf': { 'emr-
serverless.session.executionRoleArn': 'arn:aws:iam::123456789012:role/role1'}}

request = AWSRequest(method='POST', url=endpoint + "/sessions", data=json.dumps(data),
    headers=headers)

request.context["payload_signing_enabled"] = False

signer.add_auth(request)

prepped = request.prepare()

r = requests.post(prepped.url, headers=prepped.headers, data=json.dumps(data))

pprint.pprint(r.json())

### List Sessions Request

request = AWSRequest(method='GET', url=endpoint + "/sessions", headers=headers)

request.context["payload_signing_enabled"] = False

signer.add_auth(request)

prepped = request.prepare()

r2 = requests.get(prepped.url, headers=prepped.headers)
pprint.pprint(r2.json())
```

```
### Get session state

session_url = endpoint + r.headers['location']

request = AWSRequest(method='GET', url=session_url, headers=headers)

request.context["payload_signing_enabled"] = False

signer.add_auth(request)

prepped = request.prepare()

r3 = requests.get(prepped.url, headers=prepped.headers)

pprint.pprint(r3.json())

### Submit Statement

data = {
    'code': "1 + 1"
}

statements_url = endpoint + r.headers['location'] + "/statements"

request = AWSRequest(method='POST', url=statements_url, data=json.dumps(data),
    headers=headers)

request.context["payload_signing_enabled"] = False

signer.add_auth(request)

prepped = request.prepare()

r4 = requests.post(prepped.url, headers=prepped.headers, data=json.dumps(data))

pprint.pprint(r4.json())

### Check statements results

specific_statement_url = endpoint + r4.headers['location']
```

```
request = AWSRequest(method='GET', url=specific_statement_url, headers=headers)

request.context["payload_signing_enabled"] = False

signer.add_auth(request)

prepped = request.prepare()

r5 = requests.get(prepped.url, headers=prepped.headers)

pprint.pprint(r5.json())

### Delete session

session_url = endpoint + r.headers['location']

request = AWSRequest(method='DELETE', url=session_url, headers=headers)

request.context["payload_signing_enabled"] = False

signer.add_auth(request)

prepped = request.prepare()

r6 = requests.delete(prepped.url, headers=prepped.headers)

pprint.pprint(r6.json())
```

## Kernel Sparkmagic

Prima di installare sparkmagic, assicurati di aver configurato AWS le credenziali nell'istanza in cui desideri installare sparkmagic

1. [Installa sparkmagic seguendo i passaggi di installazione.](#) Nota che devi solo eseguire i primi quattro passaggi.
2. Il kernel sparkmagic supporta autenticatori personalizzati, quindi puoi integrare un autenticatore con il kernel sparkmagic in modo che ogni richiesta venga firmata. SIGv4
3. Installa l'autenticatore personalizzato EMR Serverless.

```
pip install emr-serverless-customauth
```

4. Ora fornisci il percorso dell'autenticatore personalizzato e l'URL dell'endpoint Apache Livy nel file json di configurazione sparkmagic. Usa il seguente comando per aprire il file di configurazione.

```
vim ~/.sparkmagic/config.json
```

Di seguito è riportato un config.json file di esempio.

```
{
  "kernel_python_credentials" : {
    "username": "",
    "password": "",
    "url": "https://<application-id>.livy.emr-serverless-
services.<AWS_REGION>.amazonaws.com",
    "auth": "Custom_Auth"
  },

  "kernel_scala_credentials" : {
    "username": "",
    "password": "",
    "url": "https://<application-id>.livy.emr-serverless-
services.<AWS_REGION>.amazonaws.com",
    "auth": "Custom_Auth"
  },
  "authenticators": {
    "None": "sparkmagic.auth.customauth.Authenticator",
    "Basic_Access": "sparkmagic.auth.basic.Basic",
    "Custom_Auth":
    "emr_serverless_customauth.customauthenticator.EMRServerlessCustomSigV4Signer"
  },
  "livy_session_startup_timeout_seconds": 600,
  "ignore_ssl_errors": false
}
```

5. Avvia Jupyter lab. Dovrebbe utilizzare l'autenticazione personalizzata che hai impostato nell'ultimo passaggio.
6. Puoi quindi eseguire i seguenti comandi del notebook e il tuo codice per iniziare.

```
%info //Returns the information about the current sessions.
```

```
%%configure -f //Configure information specific to a session. We supply
  executionRoleArn in this example. Change it for your use case.
{
  "driverMemory": "4g",
  "conf": {
    "emr-serverless.session.executionRoleArn":
    "arn:aws:iam::123456789012:role/JobExecutionRole"
  }
}
```

```
<your code>//Run your code to start the session
```

Internamente, ogni istruzione richiama tutte le operazioni dell'API Apache Livy tramite l'URL dell'endpoint Apache Livy configurato. È quindi possibile scrivere le istruzioni in base al caso d'uso.

## Considerazioni

Prendi in considerazione le seguenti considerazioni quando esegui carichi di lavoro interattivi tramite endpoint Apache Livy.

- EMR Serverless mantiene l'isolamento a livello di sessione utilizzando il caller principal. Il principale chiamante che crea la sessione è l'unico che può accedere a tale sessione. Per un isolamento più granulare, puoi configurare un'identità di origine quando assumi le credenziali. In questo caso, EMR Serverless applica l'isolamento a livello di sessione in base sia al principale chiamante che all'identità di origine. Per ulteriori informazioni sull'identità di origine, vedere [Monitoraggio e controllo](#) delle azioni intraprese con i ruoli presunti.
- Gli endpoint Apache Livy sono supportati con le versioni EMR Serverless 6.14.0 e successive.
- Gli endpoint Apache Livy sono supportati solo per il motore Apache Spark.
- Gli endpoint Apache Livy supportano Scala Spark e. PySpark
- Per impostazione predefinita, autoStopConfig è abilitato nelle tue applicazioni. Ciò significa che le applicazioni si chiudono dopo 15 minuti di inattività. Puoi modificare questa configurazione come parte della tua create-application update-application richiesta.
- È possibile eseguire fino a 25 sessioni simultanee su una singola applicazione abilitata per endpoint Apache Livy.
- Per una migliore esperienza di avvio, si consiglia di configurare la capacità preinizializzata per driver ed esecutori.

- È necessario avviare manualmente l'applicazione prima di connettersi all'endpoint Apache Livy.
- È necessario disporre di una quota di servizi vCPU sufficiente per Account AWS eseguire carichi di lavoro interattivi con l'endpoint Apache Livy. Consigliamo almeno 24 vCPU.
- Il timeout di sessione predefinito di Apache Livy è di 1 ora. Se non esegui istruzioni per un'ora, Apache Livy elimina la sessione e rilascia il driver e gli esecutori. Non puoi modificare questa configurazione.
- Solo le sessioni attive possono interagire con un endpoint Apache Livy. Una volta terminata, annullata o terminata la sessione, non è possibile accedervi tramite l'endpoint Apache Livy.

# Registrazione di log e monitoraggio

Il monitoraggio è una parte importante del mantenimento dell'affidabilità, della disponibilità e delle prestazioni delle applicazioni e dei lavori EMR Serverless. È necessario raccogliere i dati di monitoraggio da tutte le parti delle soluzioni EMR Serverless in modo da poter eseguire più facilmente il debug di un errore multipunto, se si verifica uno.

## Argomenti

- [Archiviazione dei registri](#)
- [Registri rotanti](#)
- [Crittografia dei log](#)
- [Configurazione delle proprietà di Apache Log4j2 per Amazon EMR Serverless](#)
- [Monitoraggio EMR Serverless](#)
- [Automazione di EMR Serverless con Amazon EventBridge](#)

## Archiviazione dei registri

Per monitorare l'avanzamento del lavoro su EMR Serverless e risolvere i problemi relativi ai processi, è possibile scegliere in che modo EMR Serverless archivia e gestisce i registri delle applicazioni. Quando invii un job run, puoi specificare storage gestito, Amazon S3 e Amazon CloudWatch come opzioni di registrazione.

Con CloudWatch, puoi specificare i tipi e le posizioni di registro che desideri utilizzare o accettare i tipi e le posizioni predefiniti. Per ulteriori informazioni sui CloudWatch log, vedere [the section called "Amazon CloudWatch"](#). Per quanto riguarda lo storage gestito e la registrazione S3, la tabella seguente mostra le posizioni dei log e la disponibilità dell'interfaccia utente che puoi aspettarti se scegli [lo storage gestito](#), i bucket [Amazon S3](#) o entrambi.

Opzione	Registri degli eventi	Registri di container	Interfaccia utente dell'applicazione
Archiviazione gestita	Archiviata in uno storage gestito	Archiviata in un sistema di storage gestito	Supportato

Opzione	Registri degli eventi	Registri di container	Interfaccia utente dell'applicazione
Storage gestito e bucket S3	Archiviati in entrambi i posti	Memorizzato in un bucket S3	Supportato
Bucket Amazon S3	Memorizzato nel bucket S3	Memorizzato nel bucket S3	Non supportato <sup>1</sup>

<sup>1</sup> Si consiglia di mantenere selezionata l'opzione di archiviazione gestita. In caso contrario, non è possibile utilizzare l'applicazione integrata UIs.

## Registrazione per EMR Serverless con storage gestito

Per impostazione predefinita, EMR Serverless archivia i log delle applicazioni in modo sicuro nello storage gestito di Amazon EMR per un massimo di 30 giorni.

### Note

Se disattivi l'opzione predefinita, Amazon EMR non può risolvere i tuoi lavori per tuo conto.

Per disattivare questa opzione da EMR Studio, deseleziona la casella di controllo Consenti di AWS conservare i log per 30 giorni nella sezione Impostazioni aggiuntive della pagina Invia lavoro.

Per disattivare questa opzione da AWS CLI, utilizza la `managedPersistenceMonitoringConfiguration` configurazione quando invii l'esecuzione di un lavoro.

```
{
  "monitoringConfiguration": {
    "managedPersistenceMonitoringConfiguration": {
      "enabled": false
    }
  }
}
```

Se la tua applicazione EMR Serverless si trova in una sottorete privata con endpoint VPC per Amazon S3 e alleghi una policy di endpoint per controllare l'accesso, devi aggiungere le seguenti

autorizzazioni per consentire a EMR Serverless di archiviare e servire i log delle applicazioni. Sostituisci Resource con i AppInfo bucket della tabella delle regioni disponibili in [Politiche di esempio per sottoreti private che accedono ad Amazon S3](#).

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "EMRServerlessManagedLogging",
      "Effect": "Allow",
      "Principal": "*",
      "Action": [
        "s3:PutObject",
        "s3:PutObjectAcl"
      ],
      "Resource": [
        "arn:aws:s3:::prod.AWS_REGION.appinfo.src",
        "arn:aws:s3:::prod.AWS_REGION.appinfo.src/*"
      ],
      "Condition": {
        "StringEquals": {
          "aws:PrincipalServiceName": "emr-serverless.amazonaws.com",
          "aws:SourceVpc": "vpc-XXXX" // Optional
        }
      }
    }
  ]
}
```

Inoltre, puoi utilizzare la chiave `aws:SourceVpc` condition per assicurarti che la richiesta passi attraverso il VPC a cui è collegato l'endpoint VPC.

## Registrazione per EMR Serverless con bucket Amazon S3

Prima che i tuoi lavori possano inviare dati di log ad Amazon S3, devi includere le seguenti autorizzazioni nella politica di autorizzazione per il ruolo di esecuzione del lavoro. Sostituisci `amzn-s3-demo-logging-bucket` con il nome del bucket di accesso.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
```

```

    "Effect": "Allow",
    "Action": [
      "s3:PutObject"
    ],
    "Resource": [
      "arn:aws:s3:::amzn-s3-demo-logging-bucket/*"
    ]
  }
]
}

```

Per configurare un bucket Amazon S3 per archiviare i log di AWS CLI, utilizza la `s3MonitoringConfiguration` configurazione all'avvio di un job. A tale scopo, fornisci quanto segue `--configuration-overrides` nella configurazione.

```

{
  "monitoringConfiguration": {
    "s3MonitoringConfiguration": {
      "logUri": "s3://amzn-s3-demo-logging-bucket/logs/"
    }
  }
}

```

Per i processi batch che non hanno i nuovi tentativi abilitati, EMR Serverless invia i log al seguente percorso:

```
'/applications/<applicationId>/jobs/<jobId>'
```

Le versioni 7.1.0 e successive di EMR Serverless supportano i tentativi di riprovare per processi di streaming e processi batch. Se si esegue un lavoro con i nuovi tentativi abilitati, EMR Serverless aggiunge automaticamente un numero di tentativo al prefisso del percorso di registro, in modo da poter distinguere e tracciare meglio i log.

```
'/applications/<applicationId>/jobs/<jobId>/attempts/<attemptNumber>/'
```

## Registrazione per EMR Serverless con Amazon CloudWatch

Quando invii un lavoro a un'applicazione EMR Serverless, puoi scegliere Amazon CloudWatch come opzione per archiviare i log delle tue applicazioni. Ciò consente di utilizzare funzionalità di analisi dei

CloudWatch log come CloudWatch Logs Insights e Live Tail. Puoi anche trasmettere i log da altri sistemi, CloudWatch ad esempio OpenSearch per ulteriori analisi.

EMR Serverless fornisce la registrazione in tempo reale dei registri dei driver. È possibile visualizzare i log in tempo reale con la funzionalità CloudWatch live tail o tramite CloudWatch i comandi di coda della CLI.

Per impostazione predefinita, la CloudWatch registrazione è disabilitata per EMR Serverless. Per abilitarlo, consulta la configurazione in. [AWS CLI](#)

### Note

Amazon CloudWatch pubblica i log in tempo reale, quindi richiede più risorse ai lavoratori. Se scegli una capacità di manodopera ridotta, l'impatto sulla durata del lavoro potrebbe aumentare. Se abiliti CloudWatch la registrazione, ti consigliamo di scegliere una maggiore capacità di lavoro. È anche possibile che la pubblicazione dei log rallenti se la frequenza delle transazioni al secondo (TPS) è troppo bassa per. PutLogEvents La configurazione di CloudWatch limitazione è globale per tutti i servizi, incluso EMR Serverless. Per ulteriori informazioni, vedi [Come posso determinare la limitazione nei miei registri?](#) CloudWatch su [re:post.AWS](#)

## Autorizzazioni richieste per la registrazione con CloudWatch

Prima che i tuoi lavori possano inviare dati di log ad Amazon CloudWatch, devi includere le seguenti autorizzazioni nella politica di autorizzazione per il ruolo Job Runtime.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "logs:DescribeLogGroups"
      ],
      "Resource": [
        "arn:aws:logs:Regione AWS:111122223333:*"
      ]
    },
    {
```

```

    "Effect": "Allow",
    "Action": [
      "logs:PutLogEvents",
      "logs:CreateLogGroup",
      "logs:CreateLogStream",
      "logs:DescribeLogStreams"
    ],
    "Resource": [
      "arn:aws:logs:Regione AWS:111122223333:log-group:my-log-group-name:*"
    ]
  }
]
}

```

## AWS CLI

Per configurare Amazon per CloudWatch archiviare i log per EMR Serverless da, utilizza AWS CLI la configurazione `cldWatchLoggingConfiguration` all'avvio di un job. A tale scopo, fornisci le seguenti modifiche di configurazione. Facoltativamente, puoi anche fornire il nome del gruppo di log, il nome del prefisso del flusso di log, i tipi di log e una chiave di crittografia ARN.

Se non specifichi i valori opzionali, CloudWatch pubblica i log in un gruppo di log predefinito/ `aws/emr-serverless`, con il flusso di log predefinito. `/applications/applicationId/jobs/jobId/worker-type`

Le versioni 7.1.0 e successive di EMR Serverless supportano i tentativi di riprovare per processi di streaming e processi batch. Se sono stati abilitati i nuovi tentativi per un lavoro, EMR Serverless aggiunge automaticamente un numero di tentativo al prefisso del percorso di registro, in modo da poter distinguere e tracciare meglio i log.

```
'/applications/<applicationId>/jobs/<jobId>/attempts/<attemptNumber>/worker-type'
```

Di seguito viene mostrata la configurazione minima richiesta per attivare Amazon CloudWatch Logging con le impostazioni predefinite per EMR Serverless:

```

{
  "monitoringConfiguration": {
    "cloudWatchLoggingConfiguration": {
      "enabled": true
    }
  }
}

```

```
}

```

L'esempio seguente mostra tutte le configurazioni obbligatorie e opzionali che puoi specificare quando attivi Amazon CloudWatch Logging per EMR Serverless. I logTypes valori supportati sono elencati anche sotto questo esempio.

```
{
  "monitoringConfiguration": {
    "cloudWatchLoggingConfiguration": {
      "enabled": true, // Required
      "logGroupName": "Example_logGroup", // Optional
      "logStreamNamePrefix": "Example_logStream", // Optional
      "encryptionKeyArn": "key-arn", // Optional
      "logTypes": {
        "SPARK_DRIVER": ["stdout", "stderr"] //List of values
      }
    }
  }
}
```

Per impostazione predefinita, EMR Serverless pubblica solo i log dei driver su cui stdout e stderr. CloudWatch Se desideri altri log, puoi specificare un ruolo del contenitore e i tipi di log corrispondenti con il campo. logTypes

L'elenco seguente mostra i tipi di worker supportati che è possibile specificare per la logTypes configurazione:

#### Spark

- SPARK\_DRIVER : ["STDERR", "STDOUT"]
- SPARK\_EXECUTOR : ["STDERR", "STDOUT"]

#### Hive

- HIVE\_DRIVER : ["STDERR", "STDOUT", "HIVE\_LOG", "TEZ\_AM"]
- TEZ\_TASK : ["STDERR", "STDOUT", "SYSTEM\_LOGS"]

## Registri rotanti

Amazon EMR Serverless può ruotare i log delle applicazioni Spark e i log degli eventi. La rotazione dei log aiuta a risolvere il problema dei processi di lunga durata che generano file di registro di

grandi dimensioni che possono occupare tutto lo spazio su disco. La rotazione dei log consente di risparmiare spazio su disco e riduce il numero di errori dei processi, poiché non rimane più spazio sul disco.

La rotazione dei log è abilitata di default ed è disponibile solo per i job Spark.

### Registri degli eventi di Spark

#### Note

La rotazione dei log degli eventi Spark è disponibile su tutte le etichette di rilascio di Amazon EMR.

Invece di generare un singolo file di registro degli eventi, EMR Serverless ruota il registro degli eventi a intervalli di tempo regolari e rimuove i file di registro degli eventi più vecchi. La rotazione dei log non influisce sui log caricati nel bucket S3.

### Registri delle applicazioni Spark

#### Note

La rotazione dei log delle applicazioni Spark è disponibile su tutte le etichette di rilascio di Amazon EMR.

EMR Serverless ruota anche i log delle applicazioni Spark per driver ed esecutori, come i file `e.stdout` e `e.stderr`. È possibile accedere ai file di registro più recenti scegliendo i collegamenti di registro in Studio utilizzando i collegamenti Spark History Server e Live UI. I file di registro sono le versioni troncate dei log più recenti. Per visualizzare i log ruotati precedenti, è necessario specificare una posizione Amazon S3 durante l'archiviazione dei log. Per ulteriori informazioni, consulta [Logging for EMR Serverless con bucket Amazon S3](#).

Puoi trovare i file di log più recenti nella seguente posizione. EMR Serverless aggiorna i file ogni 15 secondi. Questi file possono variare da 0 MB a 128 MB.

```
<example-S3-logUri>/applications/<application-id>/jobs/<job-id>/SPARK_DRIVER/stderr.gz
```

La seguente posizione contiene i file ruotati più vecchi. Ogni file è di 128 MB.

```
<example-S3-logUri>/applications/<application-id>/jobs/<job-id>/SPARK_DRIVER/archived/
stderr_<index>.gz
```

Lo stesso comportamento si applica anche agli esecutori Spark. Questa modifica è applicabile solo alla registrazione di S3. La rotazione dei log non introduce alcuna modifica ai flussi di log caricati su Amazon CloudWatch.

Le versioni 7.1.0 e successive di EMR Serverless supportano i tentativi di riprovare per lo streaming e i job in batch. Se hai abilitato i tentativi di riprovare con il tuo processo, EMR Serverless aggiunge un prefisso al percorso di registro per tali lavori in modo da poter tracciare e distinguere meglio i log l'uno dall'altro. Questo percorso contiene tutti i log ruotati.

```
 '/applications/<applicationId>/jobs/<jobId>/attempts/<attemptNumber>/'.
```

## Crittografia dei log

### Crittografia dei log EMR Serverless con storage gestito

Per crittografare i log nello storage gestito con la tua chiave KMS, usa la `managedPersistenceMonitoringConfiguration` configurazione quando invii un job run.

```
{
  "monitoringConfiguration": {
    "managedPersistenceMonitoringConfiguration" : {
      "encryptionKeyArn": "key-arn"
    }
  }
}
```

### Crittografia dei log EMR Serverless con bucket Amazon S3

Per crittografare i log nel tuo bucket Amazon S3 con la tua chiave KMS, usa la configurazione `s3MonitoringConfiguration` quando invii un job run.

```
{
  "monitoringConfiguration": {
    "s3MonitoringConfiguration": {
      "logUri": "s3://amzn-s3-demo-logging-bucket/logs/",
      "encryptionKeyArn": "key-arn"
    }
  }
}
```

```

    }
  }
}

```

## Crittografia dei log EMR Serverless con Amazon CloudWatch

Per crittografare i log in Amazon CloudWatch con la tua chiave KMS, usa la `cloudWatchLoggingConfiguration` configurazione quando invii un job run.

```

{
  "monitoringConfiguration": {
    "cloudWatchLoggingConfiguration": {
      "enabled": true,
      "encryptionKeyArn": "key-arn"
    }
  }
}

```

## Autorizzazioni richieste per la crittografia dei log

In questa sezione

- [Autorizzazioni utente richieste](#)
- [Autorizzazioni delle chiavi di crittografia per Amazon S3 e storage gestito](#)
- [Autorizzazioni per le chiavi di crittografia per Amazon CloudWatch](#)

### Autorizzazioni utente richieste

L'utente che invia il lavoro o visualizza i log dell'applicazione UIs deve disporre delle autorizzazioni per utilizzare la chiave. Puoi specificare le autorizzazioni nella politica delle chiavi KMS o nella politica IAM per l'utente, il gruppo o il ruolo. Se l'utente che invia il lavoro non dispone delle autorizzazioni della chiave KMS, EMR Serverless rifiuta l'invio dell'esecuzione del lavoro.

### Esempio di politica chiave

La seguente politica chiave fornisce le autorizzazioni per `kms:GenerateDataKey` e `kms:Decrypt`:

```

{
  "Effect": "Allow",
  "Principal":{

```

```

    "AWS": "arn:aws:iam::111122223333:user/user-name"
  },
  "Action": [
    "kms:GenerateDataKey",
    "kms:Decrypt"
  ],
  "Resource": "*"
}

```

## Policy IAM di esempio

La seguente politica IAM fornisce le autorizzazioni per `kms:GenerateDataKey` e `kms:Decrypt`

```

{
  "Version": "2012-10-17",
  "Statement": {
    "Effect": "Allow",
    "Action": [
      "kms:GenerateDataKey",
      "kms:Decrypt"
    ],
    "Resource": "key-arn"
  }
}

```

Per avviare l'interfaccia utente Spark o Tez, devi concedere ai tuoi utenti, gruppi o ruoli le autorizzazioni per accedere all'`emr-serverless:GetDashboardForJobRunAPI` come segue:

```

{
  "Version": "2012-10-17",
  "Statement": {
    "Effect": "Allow",
    "Action": [
      "emr-serverless:GetDashboardForJobRun"
    ]
  }
}

```

## Autorizzazioni delle chiavi di crittografia per Amazon S3 e storage gestito

Quando crittografi i log con la tua chiave di crittografia nello storage gestito o nei bucket S3, devi configurare le autorizzazioni delle chiavi KMS come segue.

Il `emr-serverless.amazonaws.com` principale deve disporre delle seguenti autorizzazioni nella politica per la chiave KMS:

```
{
  "Effect": "Allow",
  "Principal": {
    "Service": "emr-serverless.amazonaws.com"
  },
  "Action": [
    "kms:Decrypt",
    "kms:GenerateDataKey"
  ],
  "Resource": "*"
  "Condition": {
    "StringLike": {
      "aws:SourceArn": "arn:aws:emr-serverless:region:aws-account-id:/
applications/application-id"
    }
  }
}
```

Come best practice di sicurezza, ti consigliamo di aggiungere una chiave di `aws:SourceArn` condizione alla politica delle chiavi KMS. La chiave di condizione globale IAM `aws:SourceArn` aiuta a garantire che EMR Serverless utilizzi la chiave KMS solo per l'ARN di un'applicazione.

Il ruolo di job runtime deve disporre delle seguenti autorizzazioni nella sua policy IAM:

```
{
  "Version": "2012-10-17",
  "Statement": {
    "Effect": "Allow",
    "Action": [
      "kms:GenerateDataKey",
      "kms:Decrypt"
    ],
    "Resource": "key-arn"
  }
}
```

## Autorizzazioni per le chiavi di crittografia per Amazon CloudWatch

Per associare l'ARN della chiave KMS al tuo gruppo di log, utilizza la seguente politica IAM per il ruolo di job runtime.

```
{
  "Version": "2012-10-17",
  "Statement": {
    "Effect": "Allow",
    "Action": [
      "logs:AssociateKmsKey"
    ],
    "Resource": [
      "arn:aws:logs:Regione AWS:111122223333:log-group:my-log-group-name:"
    ]
  }
}
```

Configura la politica delle chiavi KMS per concedere le autorizzazioni KMS ad Amazon: CloudWatch

```
{
  "Version": "2012-10-17",
  "Id": "key-default-1",
  "Statement": [
    {
      "Effect": "Allow",
      "Principal": {
        "Service": "logs.Regione AWS.amazonaws.com"
      },
      "Action": [
        "kms:Decrypt",
        "kms:GenerateDataKey"
      ],
      "Resource": "*",
      "Condition": {
        "ArnLike": {
          "kms:EncryptionContext:aws:logs:arn": "arn:aws:logs:Regione AWS:111122223333:"
        }
      }
    }
  ]
}
```

# Configurazione delle proprietà di Apache Log4j2 per Amazon EMR Serverless

Questa pagina descrive come configurare le [proprietà personalizzate di Apache Log4j 2.x](#) per i lavori EMR Serverless su `StartJobRun`. Se desideri configurare le classificazioni Log4j a livello di applicazione, consulta [Configurazione predefinita dell'applicazione per EMR Serverless](#).

## Configurazione delle proprietà di Spark Log4j2 per Amazon EMR Serverless

Con le versioni 6.8.0 e successive di Amazon EMR, puoi personalizzare le proprietà di [Apache Log4j 2.x](#) per specificare configurazioni di log dettagliate. Ciò semplifica la risoluzione dei problemi dei job Spark su EMR Serverless. Per configurare queste proprietà, usa le classificazioni `spark-driver-log4j2` e `spark-executor-log4j2`.

### Argomenti

- [Classificazioni Log4j2 per Spark](#)
- [Esempio di configurazione Log4j2 per Spark](#)
- [Log4j2 negli esempi di job Spark](#)
- [Considerazioni su Log4j2 per Spark](#)

## Classificazioni Log4j2 per Spark

Per personalizzare le configurazioni dei log di Spark, usa le seguenti classificazioni con [applicationConfiguration](#). Per configurare le proprietà di Log4j 2.x, usa quanto segue [properties](#).

### **spark-driver-log4j2**

Questa classificazione imposta i valori nel `log4j2.properties` file per il driver.

### **spark-executor-log4j2**

Questa classificazione imposta i valori nel `log4j2.properties` file per l'esecutore.

## Esempio di configurazione Log4j2 per Spark

L'esempio seguente mostra come inviare un job Spark con per personalizzare le configurazioni di Log4j2 `applicationConfiguration` per il driver e l'esecutore Spark.

Per configurare le classificazioni Log4j a livello di applicazione anziché al momento dell'invio del lavoro, consulta. [Configurazione predefinita dell'applicazione per EMR Serverless](#)

```
aws emr-serverless start-job-run \
  --application-id application-id \
  --execution-role-arn job-role-arn \
  --job-driver '{
    "sparkSubmit": {
      "entryPoint": "/usr/lib/spark/examples/jars/spark-examples.jar",
      "entryPointArguments": ["1"],
      "sparkSubmitParameters": "--class org.apache.spark.examples.SparkPi --conf
spark.executor.cores=4 --conf spark.executor.memory=20g --conf spark.driver.cores=4 --
conf spark.driver.memory=8g --conf spark.executor.instances=1"
    }
  }'
  --configuration-overrides '{
    "applicationConfiguration": [
      {
        "classification": "spark-driver-log4j2",
        "properties": {
          "rootLogger.level": "error", // will only display Spark error logs
          "logger.IdentifierForClass.name": "classpath for setting logger",
          "logger.IdentifierForClass.level": "info"
        }
      },
      {
        "classification": "spark-executor-log4j2",
        "properties": {
          "rootLogger.level": "error", // will only display Spark error logs
          "logger.IdentifierForClass.name": "classpath for setting logger",
          "logger.IdentifierForClass.level": "info"
        }
      }
    ]
  }'
```

## Log4j2 negli esempi di job Spark

I seguenti esempi di codice mostrano come creare un'applicazione Spark mentre iniziizzi una configurazione Log4j2 personalizzata per l'applicazione.

## Python

### Example - Utilizzo di Log4j2 per un lavoro Spark con Python

```
import os
import sys

from pyspark import SparkConf, SparkContext
from pyspark.sql import SparkSession

app_name = "PySparkApp"
if __name__ == "__main__":
    spark = SparkSession\
        .builder\
        .appName(app_name)\
        .getOrCreate()

    spark.sparkContext._conf.getAll()
    sc = spark.sparkContext
    log4jLogger = sc._jvm.org.apache.log4j
    LOGGER = log4jLogger.LogManager.getLogger(app_name)

    LOGGER.info("pyspark script logger info")
    LOGGER.warn("pyspark script logger warn")
    LOGGER.error("pyspark script logger error")

    // your code here

    spark.stop()
```

Per personalizzare Log4j2 per il driver quando esegui un job Spark, puoi usare la seguente configurazione:

```
{
  "classification": "spark-driver-log4j2",
  "properties": {
    "rootLogger.level": "error", // only display Spark error logs
    "logger.PySparkApp.level": "info",
    "logger.PySparkApp.name": "PySparkApp"
  }
}
```

## Scala

### Example - Usare Log4j2 per un job Spark con Scala

```
import org.apache.log4j.Logger
import org.apache.spark.sql.Session

object ExampleClass {
  def main(args: Array[String]): Unit = {
    val spark = SparkSession
      .builder
      .appName(this.getClass.getName)
      .getOrCreate()

    val logger = Logger.getLogger(this.getClass);
    logger.info("script logging info logs")
    logger.warn("script logging warn logs")
    logger.error("script logging error logs")

    // your code here
    spark.stop()
  }
}
```

Per personalizzare Log4j2 per il driver quando esegui un job Spark, puoi usare la seguente configurazione:

```
{
  "classification": "spark-driver-log4j2",
  "properties": {
    "rootLogger.level": "error", // only display Spark error logs
    "logger.ExampleClass.level": "info",
    "logger.ExampleClass.name": "ExampleClass"
  }
}
```

## Considerazioni su Log4j2 per Spark

Le seguenti proprietà Log4j2.x non sono configurabili per i processi Spark:

- `rootLogger.appenderRef.stdout.ref`

- `appender.console.type`
- `appender.console.name`
- `appender.console.target`
- `appender.console.layout.type`
- `appender.console.layout.pattern`

[Per informazioni dettagliate sulle proprietà di Log4j2.x che puoi configurare, consulta il file su `log4j2.properties.template` GitHub](#)

## Monitoraggio EMR Serverless

Questa sezione illustra i modi in cui è possibile monitorare le applicazioni e i lavori Serverless di Amazon EMR.

### Argomenti

- [Monitoraggio delle applicazioni e dei lavori EMR Serverless](#)
- [Monitora i parametri di Spark con Amazon Managed Service for Prometheus](#)
- [Metriche di utilizzo EMR Serverless](#)

## Monitoraggio delle applicazioni e dei lavori EMR Serverless

Con Amazon CloudWatch Metrics for EMR Serverless, puoi ricevere parametri in CloudWatch 1 minuto e CloudWatch accedere a dashboard per visualizzare near-real-time le operazioni e le prestazioni delle tue applicazioni EMR Serverless.

EMR Serverless invia metriche a ogni minuto. CloudWatch EMR Serverless emette queste metriche a livello di applicazione, nonché di mansione, tipo di lavoratore e livelli. `capacity-allocation-type`

Per iniziare, utilizza il modello di CloudWatch dashboard EMR Serverless fornito nell'archivio [EMR GitHub Serverless](#) e distribuiscilo.

### Note

I [carichi di lavoro interattivi EMR Serverless](#) hanno solo il monitoraggio a livello di applicazione abilitato e hanno una nuova dimensione di tipo di lavoratore, `Spark_Kernel`

Per monitorare ed eseguire il debug dei carichi di lavoro interattivi, è possibile visualizzare i log e l'interfaccia utente di Apache Spark direttamente dall'area di lavoro di EMR Studio.

## Monitoraggio di parametri

### Important

Stiamo ristrutturando la visualizzazione delle nostre metriche per aggiungere `ApplicationName` e `JobName` come dimensioni. Le metriche precedenti ed esistenti sono ancora disponibili, ma consigliamo di utilizzare le nuove metriche.

### Dimensioni attuali

La tabella seguente descrive le dimensioni EMR Serverless disponibili all'interno dello spazio dei nomi. `AWS/EMR Serverless`

### Dimensioni per le metriche EMR Serverless

Dimensione	Descrizione	
<code>ApplicationId</code>	Filtri per tutte le metriche di un'applicazione EMR Serverless utilizzando l'ID dell'applicazione.	
<code>ApplicationName</code>	Filtri per tutte le metriche di un'applicazione EMR Serverless utilizzando il nome. Se il nome non viene fornito o contiene caratteri non ASCII, viene pubblicato come [Non specificato].	
<code>JobId</code>	Filtri per tutte le metriche di un server EMR senza l'ID di esecuzione del lavoro.	

Dimensione	Descrizione	
JobName	Filtri per tutte le metriche di un job EMR Serverless eseguito utilizzando il nome. Se il nome non viene fornito o contiene caratteri non ASCII, viene pubblicato come [Non specificato].	
WorkerType	Filtri per tutte le metriche di un determinato tipo di lavoratore. Ad esempio, puoi filtrare per SPARK_DRIVER e SPARK_EXECUTORS per i job Spark.	
CapacityAllocation Type	Filtri per tutte le metriche di un determinato tipo di allocazione della capacità. Ad esempio, puoi filtrare per la capacità preinizializzata e PreInitCapacity OnDemandCapacity per tutto il resto.	

## Monitoraggio a livello di applicazione

Puoi monitorare l'utilizzo della capacità a livello di applicazione EMR Serverless con i parametri di Amazon CloudWatch. Puoi anche configurare una vista singola per monitorare l'utilizzo della capacità delle applicazioni in una dashboard. CloudWatch

## Metriche delle applicazioni EMR Serverless

Parametro	Descrizione	Unità	Dimensione
MaxCPUAllowed	La CPU massima consentita per l'applicazione.	VPCU	ApplicationId , ApplicationName
MaxMemoryAllowed	La memoria massima in GB consentita per l'applicazione.	Gigabyte (GB)	ApplicationId , ApplicationName
MaxStorageAllowed	Lo spazio di archiviazione massimo in GB consentito per l'applicazione.	Gigabyte (GB)	ApplicationId , ApplicationName
CPUAllocated	Il numero totale di v CPUs allocato.	VPCU	ApplicationId , ApplicationName , WorkerType , CapacityAllocationType
IdleWorkerCount	Il numero totale di lavoratori inattivi.	Conteggio	ApplicationId , ApplicationName , WorkerType , CapacityAllocationType
MemoryAllocated	Memoria totale in GB allocata.	Gigabyte (GB)	ApplicationId , ApplicationName , WorkerType , CapacityAllocationType
PendingCreationWorkerCount	Il numero totale di lavoratori in attesa di creazione.	Conteggio	ApplicationId , ApplicationName , WorkerType

Parametro	Descrizione	Unità	Dimensione
			e ,CapacityAllocationType
RunningWorkerCount	Il numero totale di lavoratori utilizzati dall'applicazione.	Conteggio	ApplicationId , ApplicationName , WorkerType ,CapacityAllocationType
StorageAllocated	Lo spazio di archiviazione totale su disco in GB allocato.	Gigabyte (GB)	ApplicationId , ApplicationName , WorkerType ,CapacityAllocationType
TotalWorkerCount	Il numero totale di lavoratori disponibili.	Conteggio	ApplicationId , ApplicationName , WorkerType ,CapacityAllocationType

## Monitoraggio a livello di mansione

Amazon EMR Serverless invia i seguenti parametri a livello di processo ogni minuto. Amazon CloudWatch Puoi visualizzare i valori delle metriche per le esecuzioni aggregate dei processi in base allo stato di esecuzione del processo. L'unità per ciascuna metrica è il conteggio.

### Metriche a livello di job EMR Serverless

Parametro	Descrizione	Dimensione
SubmittedJobs	Il numero di lavori in uno stato Inviato.	ApplicationId , ApplicationName
PendingJobs	Il numero di lavori in uno stato In sospeso.	ApplicationId , ApplicationName

Parametro	Descrizione	Dimensione
ScheduledJobs	Il numero di lavori in uno stato Pianificato.	ApplicationId , ApplicationName
RunningJobs	Il numero di lavori in uno stato In esecuzione.	ApplicationId , ApplicationName
SuccessJobs	Il numero di lavori in uno stato di successo.	ApplicationId , ApplicationName
FailedJobs	Il numero di lavori in uno stato Non riuscito.	ApplicationId , ApplicationName
CancellingJobs	Il numero di lavori in uno stato di annullamento.	ApplicationId , ApplicationName
CancelledJobs	Il numero di lavori in uno stato annullato.	ApplicationId , ApplicationName

È possibile monitorare le metriche specifiche del motore sia per i lavori EMR Serverless in esecuzione che per quelli completati con un'applicazione specifica del motore. UIs Quando si visualizza l'interfaccia utente per un processo in esecuzione, viene visualizzata l'interfaccia utente dell'applicazione in tempo reale con aggiornamenti in tempo reale. Quando si visualizza l'interfaccia utente di un lavoro completato, viene visualizzata l'interfaccia utente persistente dell'app.

### Esecuzione di processi

Per eseguire i job EMR Serverless, è possibile visualizzare un'interfaccia in tempo reale che fornisce metriche specifiche del motore. Puoi utilizzare l'interfaccia utente di Apache Spark o l'interfaccia utente di Hive Tez per monitorare ed eseguire il debug dei lavori. Per accedervi UIs, usa la console EMR Studio o richiedi un endpoint URL sicuro con. AWS Command Line Interface

### Lavori completati

Per i job EMR Serverless completati, puoi utilizzare lo Spark History Server o l'interfaccia utente Persistent Hive Tez per visualizzare i dettagli, le fasi, le attività e le metriche dei job Spark o Hive eseguiti. Per accedervi UIs, usa la console EMR Studio o richiedi un endpoint URL sicuro con. AWS Command Line Interface

## Monitoraggio a livello di Job Worker

Amazon EMR Serverless invia ad Amazon i seguenti parametri a livello di job worker disponibili nel `AWS/EMRServerless` namespace e nel gruppo di metrici `Job Worker Metrics` CloudWatch. Amazon EMR Serverless raccoglie punti dati dai singoli lavoratori durante le esecuzioni dei lavori a livello di mansione, tipo di lavoratore e livello. `capacity-allocation-type` È possibile utilizzarlo `ApplicationId` come dimensione per monitorare più lavori che appartengono alla stessa applicazione.

### Metriche EMR Serverless a livello di job worker

Parametro	Descrizione	Unità	Dimensione
<code>WorkerCpuAllocated</code>	Il numero totale di core vCPU allocati per i lavoratori in un job run.	VPCU	JobId, JobName, ApplicationId, ApplicationName, WorkerType e CapacityAllocationType
<code>WorkerCpuUsed</code>	Il numero totale di core vCPU utilizzati dai lavoratori in un job run.	VPCU	JobId, JobName, ApplicationId, ApplicationName, WorkerType e CapacityAllocationType
<code>WorkerMemoryAllocated</code>	Memoria totale in GB allocata per i lavoratori durante l'esecuzione	Gigabyte (GB)	JobId, JobName, ApplicationId,

Parametro	Descrizione	Unità	Dimensione
	one di un processo.		ApplicationName , WorkerType e CapacityAllocationType
WorkerMemoryUsed	Memoria totale in GB utilizzata dai lavoratori durante l'esecuzione di un processo.	Gigabyte (GB)	JobId, JobName, ApplicationId , ApplicationName , WorkerType e CapacityAllocationType
WorkerEphemeralStorageAllocated	Il numero di byte di storage temporaneo allocato ai lavoratori durante l'esecuzione di un job.	Gigabyte (GB)	JobId, JobName, ApplicationId , ApplicationName , WorkerType e CapacityAllocationType

Parametro	Descrizione	Unità	Dimensione
<code>WorkerEphemeralStorageUsed</code>	Il numero di byte di storage temporane o utilizzati dai lavoratori durante l'esecuzione di un processo.	Gigabyte (GB)	JobId, JobName, ApplicationId , ApplicationName , WorkerType e CapacityAllocationType
<code>WorkerStorageReadBytes</code>	Il numero di byte letti dallo storage dai lavoratori durante l'esecuzione di un job.	Byte	JobId, JobName, ApplicationId , ApplicationName , WorkerType e CapacityAllocationType
<code>WorkerStorageWriteBytes</code>	Il numero di byte scritti nello storage dai lavoratori durante l'esecuzione di un job.	Byte	JobId, JobName, ApplicationId , ApplicationName , WorkerType e CapacityAllocationType

I passaggi seguenti descrivono come visualizzare i vari tipi di metriche.

## Console

Per accedere all'interfaccia utente dell'applicazione con la console

1. Accedi alla tua applicazione EMR Serverless su EMR Studio con le istruzioni in [Guida introduttiva](#) dalla console.
2. Per visualizzare l'applicazione UIs e i registri specifici del motore per un processo in esecuzione:
  - a. Scegli un lavoro con uno stato. RUNNING
  - b. Seleziona il lavoro nella pagina dei dettagli della candidatura o vai alla pagina dei dettagli del lavoro relativa al tuo lavoro.
  - c. Nel menu a discesa Display UI, scegli Spark UI o Hive Tez UI per accedere all'interfaccia utente dell'applicazione per il tuo tipo di lavoro.
  - d. Per visualizzare i log del motore Spark, vai alla scheda Executors nell'interfaccia utente Spark e scegli il link Logs per il driver. Per visualizzare i log del motore Hive, scegli il link Logs per il DAG appropriato nell'interfaccia utente di Hive Tez.
3. Per visualizzare l'applicazione e i registri specifici del motore per un lavoro completato: UIs
  - a. Scegli un lavoro con uno stato. SUCCESS
  - b. Seleziona il lavoro nella pagina dei dettagli della candidatura o vai alla pagina dei dettagli del lavoro.
  - c. Nel menu a discesa Display UI, scegli Spark History Server o Persistent Hive Tez UI per accedere all'interfaccia utente dell'applicazione per il tuo tipo di lavoro.
  - d. Per visualizzare i log del motore Spark, vai alla scheda Executors nell'interfaccia utente Spark e scegli il link Logs per il driver. Per visualizzare i log del motore Hive, scegli il link Logs per il DAG appropriato nell'interfaccia utente di Hive Tez.

## AWS CLI

Per accedere all'interfaccia utente dell'applicazione con il AWS CLI

- Per generare un URL da utilizzare per accedere all'interfaccia utente dell'applicazione sia per i lavori in esecuzione che per quelli completati, chiama l'GetDashboardForJobRunAPI.

```
aws emr-serverless get-dashboard-for-job-run /  
--application-id <application-id> /
```

```
--job-run-id <job-id>
```

L'URL generato è valido per un'ora.

## Monitora i parametri di Spark con Amazon Managed Service for Prometheus

Con le versioni 7.1.0 e successive di Amazon EMR, puoi integrare EMR Serverless con Amazon Managed Service for Prometheus per raccogliere i parametri di Apache Spark per lavori e applicazioni EMR Serverless. Questa integrazione è disponibile quando si invia un lavoro o si crea un'applicazione utilizzando la AWS console, l'API EMR Serverless o il. AWS CLI

### Prerequisiti

Prima di poter fornire i parametri Spark ad Amazon Managed Service for Prometheus, devi completare i seguenti prerequisiti.

- [Crea un'area di lavoro Amazon Managed Service per Prometheus](#). Questo Workspace funge da endpoint di acquisizione. Prendi nota dell'URL visualizzato per Endpoint - URL di scrittura remota. È necessario specificare l'URL quando si crea l'applicazione EMR Serverless.
- Per concedere l'accesso delle tue offerte di lavoro ad Amazon Managed Service for Prometheus a scopo di monitoraggio, aggiungi la seguente politica al tuo ruolo di esecuzione del lavoro.

```
{
  "Sid": "AccessToPrometheus",
  "Effect": "Allow",
  "Action": ["aps:RemoteWrite"],
  "Resource": "arn:aws:aps:<AWS_REGION>:<AWS_ACCOUNT_ID>:workspace/<WORKSPACE_ID>"
}
```

### Configurazione

Per utilizzare la AWS console per creare un'applicazione integrata con Amazon Managed Service for Prometheus

1. Consulta la sezione [Guida introduttiva ad Amazon EMR Serverless](#) per creare un'applicazione.

2. Durante la creazione di un'applicazione, scegli Usa impostazioni personalizzate, quindi configura l'applicazione specificando le informazioni nei campi che desideri configurare.
3. In Application logs and metrics, scegli Deliver engine metrics to Amazon Managed Service for Prometheus, quindi specifica l'URL di scrittura remota.
4. Specificate le altre impostazioni di configurazione desiderate, quindi scegliete Crea e avvia l'applicazione.

Usa la nostra AWS CLI API serverless EMR

Puoi anche utilizzare l'API AWS CLI o EMR Serverless per integrare la tua applicazione EMR Serverless con Amazon Managed Service for Prometheus quando esegui i comandi o. `create-application start-job-run`

`create-application`

```
aws emr-serverless create-application \
--release-label emr-7.1.0 \
--type "SPARK" \
--monitoring-configuration '{
  "prometheusMonitoringConfiguration": {
    "remoteWriteUrl": "https://aps-workspaces.<AWS_REGION>.amazonaws.com/
workspaces/<WORKSPACE_ID>/api/v1/remote_write"
  }
}'
```

`start-job-run`

```
aws emr-serverless start-job-run \
--application-id <APPLICATION_ID> \
--execution-role-arn <JOB_EXECUTION_ROLE> \
--job-driver '{
  "sparkSubmit": {
    "entryPoint": "local:///usr/lib/spark/examples/src/main/python/pi.py",
    "entryPointArguments": ["10000"],
    "sparkSubmitParameters": "--conf spark.dynamicAllocation.maxExecutors=10"
  }
}' \
--configuration-overrides '{
  "monitoringConfiguration": {
    "prometheusMonitoringConfiguration": {
```

```

        "remoteWriteUrl": "https://aps-workspaces.<AWS_REGION>.amazonaws.com/
workspaces/<WORKSPACE_ID>/api/v1/remote_write"
    }
}
}'

```

L'inclusione `prometheusMonitoringConfiguration` nel comando indica che EMR Serverless deve eseguire il job Spark con un agente che raccoglie le metriche Spark e le scrive sul tuo endpoint per Amazon Managed Service for Prometheus. Puoi quindi utilizzare le metriche Spark in Amazon Managed Service for Prometheus per la visualizzazione, gli avvisi e l'analisi.

## Proprietà di configurazione avanzate

EMR Serverless utilizza un componente all'interno di Spark denominato `PrometheusServlet` per raccogliere le metriche Spark e traduce i dati sulle prestazioni in dati compatibili con Amazon Managed Service for Prometheus. Per impostazione predefinita, EMR Serverless imposta i valori predefiniti in Spark e analizza le metriche del driver e dell'esecutore quando invii un lavoro utilizzando `PrometheusMonitoringConfiguration`.

La tabella seguente descrive tutte le proprietà che puoi configurare quando invii un job Spark che invia metriche ad Amazon Managed Service for Prometheus.

Proprietà Spark	Valore predefinito	Descrizione
<code>spark.metrics.conf</code> <code>.*.sink.prometheusServlet.class</code>	<code>org.apache.spark.metrics.sink.PrometheusServlet</code>	La classe utilizzata da Spark per inviare metriche ad Amazon Managed Service for Prometheus. Per ignorare il comportamento predefinito, specifica la tua classe personalizzata.
<code>spark.metrics.conf</code> <code>.*.source.jvm.class</code>	<code>org.apache.spark.metrics.source.JvmSource</code>	La classe utilizzata da Spark per raccogliere e inviare metriche cruciali dalla macchina virtuale Java sottostante. Per interrompere la raccolta delle metriche JVM,

Proprietà Spark	Valore predefinito	Descrizione
		disabilita questa proprietà impostandola su una stringa vuota, ad esempio. "" Per ignorare il comportamento predefinito, specificate la vostra classe personalizzata.
<code>spark.metrics.conf.driver.sink.prometheusServlet.path</code>	<code>/metrics/prometheus</code>	L'URL distinto che Amazon Managed Service for Prometheus utilizza per raccogliere le metriche dal driver. Per ignorare il comportamento predefinito, specifica il tuo percorso. Per interrompere la raccolta delle metriche dei driver, disattiva questa proprietà impostandola su una stringa vuota, ad esempio. ""
<code>spark.metrics.conf.executor.sink.prometheusServlet.path</code>	<code>/metrics/executor/prometheus</code>	L'URL distinto che Amazon Managed Service for Prometheus utilizza per raccogliere le metriche dell'esecutore. Per ignorare il comportamento predefinito, specifica il tuo percorso. Per interrompere la raccolta delle metriche degli esecutori, disabilitate questa proprietà impostandola su una stringa vuota, ad esempio. ""

[Per ulteriori informazioni sulle metriche Spark, consulta Metriche di Apache Spark.](#)

## Considerazioni e limitazioni

Quando utilizzi Amazon Managed Service for Prometheus per raccogliere metriche da EMR Serverless, considera le seguenti considerazioni e limitazioni.

- Il supporto per l'utilizzo di Amazon Managed Service for Prometheus con EMR Serverless è disponibile solo in Regioni AWS nei paesi in [cui Amazon](#) Managed Service for Prometheus è generalmente disponibile.
- Far sì che l'agente raccolga i parametri Spark su Amazon Managed Service for Prometheus richiede più risorse da parte dei lavoratori. Se si sceglie un worker di dimensioni inferiori, ad esempio un lavoratore vCPU, il tempo di esecuzione del lavoro potrebbe aumentare.
- Il supporto per l'utilizzo di Amazon Managed Service for Prometheus con EMR Serverless è disponibile solo per le versioni 7.1.0 e successive di Amazon EMR.
- Amazon Managed Service for Prometheus deve essere distribuito nello stesso account in cui esegui EMR Serverless per raccogliere i parametri.

## Metriche di utilizzo EMR Serverless

Puoi utilizzare i parametri di CloudWatch utilizzo di Amazon per fornire visibilità sulle risorse utilizzate dal tuo account. Utilizza queste metriche per visualizzare l'utilizzo del servizio su CloudWatch grafici e dashboard.

Le metriche di utilizzo di EMR Serverless corrispondono a Service Quotas. È possibile configurare gli allarmi che avvisano quando l'uso si avvicina a una quota di servizio. Per ulteriori informazioni, consulta [Service Quotas e Amazon CloudWatch alarms](#) nella Service Quotas User Guide.

Per ulteriori informazioni sulle quote dei servizi EMR Serverless, vedere. [Punti finali e quote per EMR Serverless](#)

## Metriche di utilizzo delle quote di servizio per EMR Serverless

EMR Serverless pubblica le seguenti metriche di utilizzo delle quote di servizio nel namespace. AWS/Usage

Parametro	Descrizione
ResourceCount	Il numero totale della risorsa specificata in esecuzione sul tuo account. La risorsa è definita dalle <a href="#">dimensioni</a> associate alla metrica.

## Dimensioni per le metriche di utilizzo delle quote di servizio EMR Serverless

È possibile utilizzare le seguenti dimensioni per perfezionare le metriche di utilizzo pubblicate da EMR Serverless.

Dimensione	Valore	Descrizione
Service	EMR serverless	Il nome del file che contiene la risorsa. Servizio AWS
Type	Risorsa	Il tipo di entità segnalata da EMR Serverless.
Resource	VPCU	Il tipo di risorsa monitorata da EMR Serverless.
Class	Nessuno	La classe di risorse monitorata da EMR Serverless.

## Automazione di EMR Serverless con Amazon EventBridge

È possibile Amazon EventBridge utilizzarlo per automatizzare Servizi AWS e rispondere automaticamente agli eventi di sistema, ad esempio problemi di disponibilità delle applicazioni o modifiche delle risorse. EventBridge offre un flusso quasi in tempo reale di eventi di sistema che descrivono i cambiamenti nelle AWS risorse. Puoi compilare regole semplici che indichino quali eventi sono considerati di interesse per te e quali azioni automatizzate intraprendere quando un evento corrisponde a una regola. Con EventBridge, puoi automaticamente:

- Invocare una funzione AWS Lambda
- Inoltrare un evento ad Amazon Kinesis Data Streams

- Attiva una macchina a stati AWS Step Functions
- Notifica un argomento Amazon SNS o una coda Amazon SQS

Ad esempio, se lo utilizzi EventBridge con EMR Serverless, puoi attivare una AWS Lambda funzione quando un processo ETL ha esito positivo o notificare un argomento di Amazon SNS quando un processo ETL fallisce.

EMR Serverless emette quattro tipi di eventi:

- Eventi di modifica dello stato dell'applicazione: eventi che generano ogni cambiamento di stato di un'applicazione. Per ulteriori informazioni sugli stati delle applicazioni, vedere [Stati dell'applicazione](#).
- Eventi di modifica dello stato del job run: eventi che generano ogni modifica di stato di un job run. Per ulteriori informazioni su, vedere [Stati delle esecuzioni di processi](#).
- Eventi di ripetizione del processo: eventi che generano ogni nuovo tentativo di un processo eseguito dalle versioni 7.1.0 e successive di Amazon EMR Serverless.
- Eventi di aggiornamento sull'utilizzo delle risorse del lavoro: gli eventi che emettono aggiornamenti sull'utilizzo delle risorse per un processo vengono eseguiti a intervalli prossimi a 30 minuti.

## Esempi di eventi EMR Serverless EventBridge

Gli eventi segnalati da EMR Serverless hanno un valore `aws.emr-serverless` assegnato a `source`, come illustrato negli esempi seguenti.

Evento di modifica dello stato dell'applicazione

L'evento di esempio seguente mostra un'applicazione nello CREATING stato.

```
{
  "version": "0",
  "id": "9fd3cf79-1ff1-b633-4dd9-34508dc1e660",
  "detail-type": "EMR Serverless Application State Change",
  "source": "aws.emr-serverless",
  "account": "123456789012",
  "time": "2022-05-31T21:16:31Z",
  "region": "us-east-1",
  "resources": [],
  "detail": {
    "applicationId": "00f1cb5c6anuij25",
    "applicationName": "3965ad00-8fba-4932-a6c8-ded32786fd42",
```

```

    "arn": "arn:aws:emr-serverless:us-east-1:111122223333:/
applications/00f1cb3c6anuij25",
    "releaseLabel": "emr-6.6.0",
    "state": "CREATING",
    "type": "HIVE",
    "createdAt": "2022-05-31T21:16:31.547953Z",
    "updatedAt": "2022-05-31T21:16:31.547970Z",
    "autoStopConfig": {
      "enabled": true,
      "idleTimeout": 15
    },
    "autoStartConfig": {
      "enabled": true
    }
  }
}

```

## Evento di modifica dello stato di Job run

L'evento di esempio seguente mostra l'esecuzione di un processo che passa da SCHEDULED uno RUNNING stato all'altro.

```

{
  "version": "0",
  "id": "00df3ec6-5da1-36e6-ab71-20f0de68f8a0",
  "detail-type": "EMR Serverless Job Run State Change",
  "source": "aws.emr-serverless",
  "account": "123456789012",
  "time": "2022-05-31T21:07:42Z",
  "region": "us-east-1",
  "resources": [],
  "detail": {
    "jobRunId": "00f1cbn5g4bb0c01",
    "applicationId": "00f1982r1uukb925",
    "arn": "arn:aws:emr-serverless:us-east-1:123456789012:/
applications/00f1982r1uukb925/jobruns/00f1cbn5g4bb0c01",
    "releaseLabel": "emr-6.6.0",
    "state": "RUNNING",
    "previousState": "SCHEDULED",
    "createdBy": "arn:aws:sts::123456789012:assumed-role/
TestRole-402dcef3ad14993c15d28263f64381e4cda34775/6622b6233b6d42f59c25dd2637346242",
    "updatedAt": "2022-05-31T21:07:42.299487Z",
    "createdAt": "2022-05-31T21:07:25.325900Z"
  }
}

```

```
}
}
```

## Evento Job run Retry

Di seguito è riportato un esempio di evento Job Run Retry.

```
{
  "version": "0",
  "id": "00df3ec6-5da1-36e6-ab71-20f0de68f8a0",
  "detail-type": "EMR Serverless Job Run Retry",
  "source": "aws.emr-serverless",
  "account": "123456789012",
  "time": "2022-05-31T21:07:42Z",
  "region": "us-east-1",
  "resources": [],
  "detail": {
    "jobRunId": "00f1cbn5g4bb0c01",
    "applicationId": "00f1982r1uukb925",
    "arn": "arn:aws:emr-serverless:us-east-1:123456789012:/
applications/00f1982r1uukb925/jobruns/00f1cbn5g4bb0c01",
    "releaseLabel": "emr-6.6.0",
    "createdBy": "arn:aws:sts::123456789012:assumed-role/
TestRole-402dcef3ad14993c15d28263f64381e4cda34775/6622b6233b6d42f59c25dd2637346242",
    "updatedAt": "2022-05-31T21:07:42.299487Z",
    "createdAt": "2022-05-31T21:07:25.325900Z",
    //Attempt Details
    "previousAttempt": 1,
    "previousAttemptState": "FAILED",
    "previousAttemptCreatedAt": "2022-05-31T21:07:25.325900Z",
    "previousAttemptEndedAt": "2022-05-31T21:07:30.325900Z",
    "newAttempt": 2,
    "newAttemptCreatedAt": "2022-05-31T21:07:30.325900Z"
  }
}
```

## Aggiornamento sull'utilizzo delle risorse Job

L'evento di esempio seguente mostra l'aggiornamento finale sull'utilizzo delle risorse per un processo che è passato allo stato terminale dopo l'esecuzione.

```
{
  "version": "0",
```

```

    "id": "00df3ec6-5da1-36e6-ab71-20f0de68f8a0",
    "detail-type": "EMR Serverless Job Resource Utilization Update",
    "source": "aws.emr-serverless",
    "account": "123456789012",
    "time": "2022-05-31T21:07:42Z",
    "region": "us-east-1",
    "resources": [
      "arn:aws:emr-serverless:us-east-1:123456789012:/applications/00f1982r1uukb925/
jobruns/00f1cbn5g4bb0c01"
    ],
    "detail": {
      "applicationId": "00f1982r1uukb925",
      "jobRunId": "00f1cbn5g4bb0c01",
      "attempt": 1,
      "mode": "BATCH",
      "createdAt": "2022-05-31T21:07:25.325900Z",
      "startedAt": "2022-05-31T21:07:26.123Z",
      "calculatedFrom": "2022-05-31T21:07:42.299487Z",
      "calculatedTo": "2022-05-31T21:07:30.325900Z",
      "resourceUtilizationFinal": true,
      "resourceUtilizationForInterval": {
        "vCPUHour": 0.023,
        "memoryGBHour": 0.114,
        "storageGBHour": 0.228
      },
      "billedResourceUtilizationForInterval": {
        "vCPUHour": 0.067,
        "memoryGBHour": 0.333,
        "storageGBHour": 0
      },
      "totalResourceUtilization": {
        "vCPUHour": 0.023,
        "memoryGBHour": 0.114,
        "storageGBHour": 0.228
      },
      "totalBilledResourceUtilization": {
        "vCPUHour": 0.067,
        "memoryGBHour": 0.333,
        "storageGBHour": 0
      }
    }
  }
}

```

Il campo `StartedAt` sarà presente solo nel caso in cui il lavoro sia passato allo stato di esecuzione.

# Applicazione di tag alle risorse

Puoi assegnare i tuoi metadati a ciascuna risorsa utilizzando i tag per aiutarti a gestire le tue risorse EMR Serverless. Questa sezione fornisce una panoramica delle funzioni dei tag e mostra come creare tag.

## Argomenti

- [Che cos'è un tag?](#)
- [Tagging delle risorse](#)
- [Limitazioni relative all'etichettatura](#)
- [Utilizzo dei tag utilizzando l'API AWS CLI Serverless di Amazon EMR](#)

## Che cos'è un tag?

Un tag è un'etichetta che si assegna a una AWS risorsa. Ogni tag consiste di una chiave e di un valore, entrambi personalizzabili. I tag consentono di classificare le AWS risorse in base ad attributi quali scopo, proprietario e ambiente. Se disponi di un numero elevato di risorse, puoi individuare rapidamente una risorsa specifica in base ai tag assegnati. Ad esempio, puoi definire un set di tag per le tue applicazioni Amazon EMR Serverless per aiutarti a monitorare il proprietario e il livello di stack di ciascuna applicazione. Consigliamo di definire un set coerente di chiavi di tag per ciascun tipo di risorsa.

I tag non vengono assegnati in automatico alle risorse. Dopo aver aggiunto un tag a una risorsa, puoi modificare il valore di un tag o rimuoverlo dalla risorsa in qualsiasi momento. I tag non hanno alcun significato semantico per Amazon EMR Serverless e vengono interpretati rigorosamente come stringhe di caratteri. Se aggiungi un tag con la stessa chiave di un tag esistente a una risorsa specifica, il nuovo valore sovrascrive quello precedente.

Se usi IAM, puoi controllare quali utenti del tuo AWS account sono autorizzati a gestire i tag. Per esempi di policy di controllo dell'accesso basate su tag, consulta [Policy per il controllo degli accessi basato su tag](#).

## Tagging delle risorse

È possibile contrassegnare applicazioni ed esecuzioni di lavori nuove o esistenti. Se utilizzi l'API Serverless di Amazon EMR o un AWS SDK AWS CLI, puoi applicare tag a nuove risorse utilizzando

il `tags` parametro sull'azione API pertinente. Puoi applicare tag a risorse esistenti utilizzando l'operazione `API TagResource`.

Puoi utilizzare alcune operazioni per la creazione di risorse per specificare tag per una risorsa durante la sua creazione. In questo caso, se i tag non possono essere applicati durante la creazione della risorsa, la risorsa non viene creata. Mediante questo meccanismo, le risorse a cui desideri applicare tag al momento della creazione vengono create con tag specifici o non vengono create affatto. Se tagghi le risorse al momento della creazione, non è necessario eseguire script di tagging personalizzati dopo aver creato una risorsa.

La tabella seguente descrive le risorse Serverless di Amazon EMR che possono essere taggate.

Risorsa	Supporta tag	Supporta la propagazione di tag	Supporta l'etichettatura alla creazione (Amazon EMR Serverless API, AWS CLI e SDK)	API per la creazione (i tag possono essere aggiunti durante la creazione)
Applicazione	Sì	No. I tag associati a un'applicazione non si propagano alle esecuzioni di job inviate a tale applicazione.	Sì	<code>CreateApplication</code>
Esecuzione del processo	Sì	No	Sì	<code>StartJobRun</code>

## Limitazioni relative all'etichettatura

Le seguenti limitazioni di base si applicano ai tag:

- Ogni risorsa può avere un massimo di 50 tag creati dall'utente.

- Per ciascuna risorsa, ogni chiave del tag deve essere univoca e ogni chiave del tag può avere un solo valore.
- La lunghezza massima delle chiavi è 128 caratteri Unicode in UTF-8.
- Il valore massimo è 256 caratteri Unicode in UTF-8.
- I caratteri consentiti sono lettere, numeri, spazi rappresentabili in UTF-8 e i seguenti caratteri: `_.:/=+ - @`.
- Una chiave di tag non può essere una stringa vuota. Un valore di tag può essere una stringa vuota, ma non nullo.
- I valori e le chiavi dei tag rispettano la distinzione tra maiuscole e minuscole.
- Non utilizzare `AWS`: alcuna combinazione maiuscola o minuscola, ad esempio un prefisso per chiavi o valori. Essi sono riservati per l'utilizzo esclusivo di AWS.

## Utilizzo dei tag utilizzando l'API AWS CLI Serverless di Amazon EMR

Utilizza i seguenti AWS CLI comandi o le operazioni API Serverless di Amazon EMR per aggiungere, aggiornare, elencare ed eliminare i tag per le tue risorse.

Risorsa	Supporta tag	Supporta la propagazione di tag
Aggiunta o sovrascrittura di uno o più tag	<code>tag-resource</code>	TagResource
Elencazione dei tag associati a una risorsa	<code>list-tags-for-resource</code>	ListTagsForResource
Eliminazione di uno o più tag	<code>untag-resource</code>	UntagResource

I seguenti esempi mostrano come aggiungere o rimuovere tag alle o dalle risorse utilizzando la AWS CLI.

Etichetta un'applicazione esistente

Il comando seguente contrassegna un'applicazione esistente.

```
aws emr-serverless tag-resource --resource-arn resource_ARN --tags team=devs
```

## Rimuovere i tag da un'applicazione esistente

Il comando seguente elimina un tag da un'applicazione esistente.

```
aws emr-serverless untag-resource --resource-arn resource_ARN --tag-keys tag_key
```

## Elenca i tag per una risorsa

Il comando seguente elenca i tag associati a una risorsa esistente.

```
aws emr-serverless list-tags-for-resource --resource-arn resource_ARN
```

# Tutorial per EMR Serverless

Questa sezione descrive i casi d'uso comuni quando si lavora con applicazioni EMR Serverless. Ciò include una varietà di strumenti tra cui Hudi e Iceberg per lavorare su set di dati di grandi dimensioni e utilizzare le librerie Python e Python per inviare lavori Spark.

## Argomenti

- [Utilizzo di Java 17 con Amazon EMR Serverless](#)
- [Utilizzo di Apache Hudi con EMR Serverless](#)
- [Utilizzo di Apache Iceberg con EMR Serverless](#)
- [Utilizzo delle librerie Python con EMR Serverless](#)
- [Utilizzo di diverse versioni di Python con EMR Serverless](#)
- [Utilizzo di Delta Lake OSS con EMR Serverless](#)
- [Invio di lavori EMR Serverless da Airflow](#)
- [Utilizzo delle funzioni Hive definite dall'utente con EMR Serverless](#)
- [Utilizzo di immagini personalizzate con EMR Serverless](#)
- [Utilizzo dell'integrazione di Amazon Redshift per Apache Spark su Amazon EMR Serverless](#)
- [Connessione a DynamoDB con Amazon EMR Serverless](#)

## Utilizzo di Java 17 con Amazon EMR Serverless

Con le versioni 6.11.0 e successive di Amazon EMR, puoi configurare i job EMR Serverless Spark per utilizzare il runtime Java 17 per la Java Virtual Machine (JVM). Utilizza uno dei seguenti metodi per configurare Spark con Java 17.

### **JAVA\_HOME**

Per sovrascrivere l'impostazione JVM per EMR Serverless 6.11.0 e versioni successive, è possibile fornire l'impostazione alle relative classificazioni e all'ambiente. `JAVA_HOME spark.emr-serverless.driverEnv spark.executorEnv`

`x86_64`

Imposta le proprietà richieste per specificare Java 17 come configurazione per il driver e gli `JAVA_HOME` esecutori Spark:

```
--conf spark.emr-serverless.driverEnv.JAVA_HOME=/usr/lib/jvm/java-17-amazon-
corretto.x86_64/
--conf spark.executorEnv.JAVA_HOME=/usr/lib/jvm/java-17-amazon-corretto.x86_64/
```

## arm\_64

Imposta le proprietà richieste per specificare Java 17 come JAVA\_HOME configurazione per il driver e gli esecutori Spark:

```
--conf spark.emr-serverless.driverEnv.JAVA_HOME=/usr/lib/jvm/java-17-amazon-
corretto.aarch64/
--conf spark.executorEnv.JAVA_HOME=/usr/lib/jvm/java-17-amazon-corretto.aarch64/
```

## spark-defaults

In alternativa, è possibile specificare Java 17 nella spark-defaults classificazione per sovrascrivere l'impostazione JVM per EMR Serverless 6.11.0 e versioni successive.

## x86\_64

Specificare Java 17 nella classificazione: spark-defaults

```
{
  "applicationConfiguration": [
    {
      "classification": "spark-defaults",
      "properties": {
        "spark.emr-serverless.driverEnv.JAVA_HOME" : "/usr/lib/jvm/java-17-
amazon-corretto.x86_64/",
        "spark.executorEnv.JAVA_HOME": "/usr/lib/jvm/java-17-amazon-
corretto.x86_64/"
      }
    }
  ]
}
```

## arm\_64

Specificare Java 17 nella spark-defaults classificazione:

```
{
```

```
"applicationConfiguration": [  
  {  
    "classification": "spark-defaults",  
    "properties": {  
      "spark.emr-serverless.driverEnv.JAVA_HOME" : "/usr/lib/jvm/java-17-  
amazon-corretto.aarch64/",  
      "spark.executorEnv.JAVA_HOME": "/usr/lib/jvm/java-17-amazon-  
corretto.aarch64/"  
    }  
  }  
]  
}
```

## Utilizzo di Apache Hudi con EMR Serverless

Questa sezione descrive l'utilizzo di Apache Hudi con le applicazioni EMR Serverless. Hudi è un framework di gestione dei dati che semplifica l'elaborazione dei dati.

Per utilizzare Apache Hudi con applicazioni EMR Serverless

1. Imposta le proprietà Spark richieste nell'esecuzione del job Spark corrispondente.

```
spark.jars=/usr/lib/hudi/hudi-spark-bundle.jar  
spark.serializer=org.apache.spark.serializer.KryoSerializer
```

2. Per sincronizzare una tabella Hudi con il catalogo configurato, designa AWS Glue Data Catalog come metastore o configura un metastore esterno. EMR Serverless supporta la modalità di sincronizzazione per hms le tabelle Hive per i carichi di lavoro Hudi. EMR Serverless attiva questa proprietà come impostazione predefinita. Per ulteriori informazioni su come configurare il metastore, consulta. [Configurazione Metastore per EMR Serverless](#)

### Important

EMR Serverless non supporta HIVEQL né offre opzioni di modalità di sincronizzazione per JDBC le tabelle Hive per gestire i carichi di lavoro Hudi. [Per saperne di più, consulta Modalità di sincronizzazione.](#)

Quando si utilizza il AWS Glue Data Catalog come metastore, è possibile specificare le seguenti proprietà di configurazione per il job Hudi.

```
--conf spark.jars=/usr/lib/hudi/hudi-spark-bundle.jar,  
--conf spark.serializer=org.apache.spark.serializer.KryoSerializer,  
--conf  
spark.hadoop.hive.metastore.client.factory.class=com.amazonaws.glue.catalog.metastore.AWSG
```

[Per ulteriori informazioni sulle versioni di Apache Hudi di Amazon EMR, consulta la cronologia delle versioni di Hudi.](#)

## Utilizzo di Apache Iceberg con EMR Serverless

Questa sezione descrive come utilizzare Apache Iceberg con le applicazioni EMR Serverless. Apache Iceberg è un formato di tabella che aiuta a lavorare con set di dati di grandi dimensioni nei data lake.

Per utilizzare Apache Iceberg con applicazioni EMR Serverless

1. Imposta le proprietà Spark richieste nell'esecuzione del job Spark corrispondente.

```
spark.jars=/usr/share/aws/iceberg/lib/iceberg-spark3-runtime.jar
```

2. Designate il AWS Glue Data Catalog come metastore o configurate un metastore esterno. Per ulteriori informazioni sulla configurazione del metastore, consulta [Configurazione Metastore per EMR Serverless](#)

Configura le proprietà del metastore che desideri utilizzare per Iceberg. Ad esempio, se desideri utilizzare il AWS Glue Data Catalog, imposta le seguenti proprietà nella configurazione dell'applicazione.

```
spark.sql.catalog.dev.warehouse=s3://amzn-s3-demo-bucket/EXAMPLE-PREFIX/  
spark.sql.extensions=org.apache.iceberg.spark.extensions.IcebergSparkSessionExtensions  
spark.sql.catalog.dev=org.apache.iceberg.spark.SparkCatalog  
spark.sql.catalog.dev.catalog-impl=org.apache.iceberg.aws.glue.GlueCatalog  
spark.hadoop.hive.metastore.client.factory.class=com.amazonaws.glue.catalog.metastore.AWSG
```

Quando utilizzi AWS Glue Data Catalog come metastore, puoi specificare le seguenti proprietà di configurazione per il tuo job Iceberg.

```
--conf spark.jars=/usr/share/aws/iceberg/lib/iceberg-spark3-runtime.jar,  
--conf  
  spark.sql.extensions=org.apache.iceberg.spark.extensions.IcebergSparkSessionExtensions,  
--conf spark.sql.catalog.dev=org.apache.iceberg.spark.SparkCatalog,  
--conf spark.sql.catalog.dev.catalog-impl=org.apache.iceberg.aws.glue.GlueCatalog,  
--conf spark.sql.catalog.dev.warehouse=s3://amzn-s3-demo-bucket/EXAMPLE-PREFIX/  
--conf  
  spark.hadoop.hive.metastore.client.factory.class=com.amazonaws.glue.catalog.metastore.AWSGlueDataCatalogHiveMetastore
```

[Per ulteriori informazioni sulle versioni di Apache Iceberg di Amazon EMR, consulta la cronologia delle versioni di Iceberg.](#)

## Utilizzo delle librerie Python con EMR Serverless

Quando esegui PySpark lavori su applicazioni Amazon EMR Serverless, puoi impacchettare diverse librerie Python come dipendenze. Per fare ciò, puoi usare le funzionalità native di Python, creare un ambiente virtuale o configurare direttamente i tuoi PySpark lavori per utilizzare le librerie Python. Questa pagina descrive ogni approccio.

### Utilizzo delle funzionalità native di Python

Quando imposti la seguente configurazione, puoi usarla PySpark per caricare file Python (.py), pacchetti Python compressi (.egg) e file .zip negli esecutori Spark.

```
--conf spark.submit.pyFiles=s3://amzn-s3-demo-bucket/EXAMPLE-PREFIX/<.py|.egg|.zip  
file>
```

Per maggiori dettagli su come usare gli ambienti virtuali Python per i PySpark lavori, vedi [Uso delle funzionalità PySpark native](#).

### Creazione di un ambiente virtuale Python

Per impacchettare più librerie Python per un PySpark lavoro, puoi creare ambienti virtuali Python isolati.

1. Per creare l'ambiente virtuale Python, usa i seguenti comandi. L'esempio mostrato installa i pacchetti `scipy` e `matplotlib` in un pacchetto di ambiente virtuale e copia l'archivio in una posizione Amazon S3.

 Important

È necessario eseguire i seguenti comandi in un ambiente Amazon Linux 2 simile con la stessa versione di Python utilizzata in EMR Serverless, ovvero Python 3.7.10 per Amazon EMR release 6.6.0. È possibile trovare un Dockerfile di esempio nel repository [EMR Serverless Samples](#). [GitHub](#)

```
# initialize a python virtual environment
python3 -m venv pyspark_venvsource
source pyspark_venvsource/bin/activate

# optionally, ensure pip is up-to-date
pip3 install --upgrade pip

# install the python packages
pip3 install scipy
pip3 install matplotlib

# package the virtual environment into an archive
pip3 install venv-pack
venv-pack -f -o pyspark_venv.tar.gz

# copy the archive to an S3 location
aws s3 cp pyspark_venv.tar.gz s3://amzn-s3-demo-bucket/EXAMPLE-PREFIX/

# optionally, remove the virtual environment directory
rm -fr pyspark_venvsource
```

2. Invia il job Spark con le proprietà impostate per utilizzare l'ambiente virtuale Python.

```
--conf spark.archives=s3://amzn-s3-demo-bucket/EXAMPLE-PREFIX/
pyspark_venv.tar.gz#environment
--conf spark.emr-serverless.driverEnv.PYSPARK_DRIVER_PYTHON=./environment/bin/
python
--conf spark.emr-serverless.driverEnv.PYSPARK_PYTHON=./environment/bin/python
--conf spark.executorEnv.PYSPARK_PYTHON=./environment/bin/python
```

Nota che se non sovrascrivi il binario Python originale, la seconda configurazione nella sequenza di impostazioni precedente sarà. `--conf spark.executorEnv.PYSPARK_PYTHON=python`

Per ulteriori informazioni su come utilizzare gli ambienti virtuali Python per i PySpark lavori, vedere [Using Virtualenv](#). Per altri esempi su come inviare lavori Spark, vedi. [Utilizzo delle configurazioni Spark quando si eseguono job EMR Serverless](#)

## Configurazione dei PySpark lavori per l'utilizzo delle librerie Python

[Con le versioni 6.12.0 e successive di Amazon EMR, puoi configurare direttamente i job EMR Serverless PySpark per utilizzare le più diffuse librerie Python di data science come panda e senza alcuna configurazione aggiuntiva. NumPyPyArrow](#)

Gli esempi seguenti mostrano come impacchettare ogni libreria Python per un PySpark lavoro.

### NumPy

NumPy è una libreria Python per il calcolo scientifico che offre array e operazioni multidimensionali per matematica, ordinamento, simulazione casuale e statistiche di base. NumPyPer utilizzarla, esegui il seguente comando:

```
import numpy
```

### pandas

pandas è una libreria Python basata su NumPy. La libreria pandas fornisce ai data scientist strutture di dati e strumenti di analisi [DataFrame](#) dei dati. Per usare pandas, esegui il seguente comando:

```
import pandas
```

### PyArrow

PyArrow è una libreria Python che gestisce i dati colonnari in memoria per migliorare le prestazioni lavorative. PyArrow si basa sulla specifica di sviluppo multilingue di Apache Arrow, che è un modo standard per rappresentare e scambiare dati in un formato colonnare. Per utilizzarlo PyArrow, esegui il seguente comando:

```
import pyarrow
```

## Utilizzo di diverse versioni di Python con EMR Serverless

Oltre allo use case in [Utilizzo delle librerie Python con EMR Serverless](#), puoi anche utilizzare ambienti virtuali Python per lavorare con versioni di Python diverse rispetto alla versione inclusa nella release Amazon EMR per la tua applicazione Amazon EMR Serverless. Per fare ciò, devi creare un ambiente virtuale Python con la versione di Python che desideri utilizzare.

Per inviare un lavoro da un ambiente virtuale Python

1. Crea il tuo ambiente virtuale con i comandi nell'esempio seguente. Questo esempio installa Python 3.9.9 in un pacchetto di ambiente virtuale e copia l'archivio in una posizione Amazon S3.

### Important

Se utilizzi le versioni 7.0.0 e successive di Amazon EMR, devi eseguire i comandi in un ambiente Amazon Linux 2023 simile a quello che usi per le tue applicazioni EMR Serverless.

Se utilizzi la versione 6.15.0 o precedente, devi eseguire i seguenti comandi in un ambiente Amazon Linux 2 simile.

```
# install Python 3.9.9 and activate the venv
yum install -y gcc openssl-devel bzip2-devel libffi-devel tar gzip wget make
wget https://www.python.org/ftp/python/3.9.9/Python-3.9.9.tgz && \
tar xzf Python-3.9.9.tgz && cd Python-3.9.9 && \
./configure --enable-optimizations && \
make altinstall

# create python venv with Python 3.9.9
python3.9 -m venv pyspark_venv_python_3.9.9 --copies
source pyspark_venv_python_3.9.9/bin/activate

# copy system python3 libraries to venv
cp -r /usr/local/lib/python3.9/* ./pyspark_venv_python_3.9.9/lib/python3.9/

# package venv to archive.
# Note that you have to supply --python-prefix option
# to make sure python starts with the path where your
# copied libraries are present.
# Copying the python binary to the "environment" directory.
```

```
pip3 install venv-pack
venv-pack -f -o pyspark_venv_python_3.9.9.tar.gz --python-prefix /home/hadoop/
environment

# stage the archive in S3
aws s3 cp pyspark_venv_python_3.9.9.tar.gz s3://<path>

# optionally, remove the virtual environment directory
rm -fr pyspark_venv_python_3.9.9
```

## 2. Imposta le tue proprietà per utilizzare l'ambiente virtuale Python e invia il job Spark.

```
# note that the archive suffix "environment" is the same as the directory where you
  copied the Python binary.
--conf spark.archives=s3://amzn-s3-demo-bucket/EXAMPLE-PREFIX/
pyspark_venv_python_3.9.9.tar.gz#environment
--conf spark.emr-serverless.driverEnv.PYSPARK_DRIVER_PYTHON=./environment/bin/
python
--conf spark.emr-serverless.driverEnv.PYSPARK_PYTHON=./environment/bin/python
--conf spark.executorEnv.PYSPARK_PYTHON=./environment/bin/python
```

Per ulteriori informazioni su come utilizzare gli ambienti virtuali Python per i PySpark lavori, vedere [Using Virtualenv](#). Per altri esempi su come inviare lavori Spark, vedi [Utilizzo delle configurazioni Spark quando si eseguono job EMR Serverless](#)

## Utilizzo di Delta Lake OSS con EMR Serverless

### Amazon EMR versioni 6.9.0 e successive

#### Note

Amazon EMR 7.0.0 e versioni successive utilizzano Delta Lake 3.0.0, che rinomina il file in `delta-core.jar` `delta-spark.jar`. Se utilizzi Amazon EMR 7.0.0 o versioni successive, assicurati di `delta-spark.jar` specificarlo nelle configurazioni.

Amazon EMR 6.9.0 e versioni successive includono Delta Lake, quindi non devi più impacchettare Delta Lake personalmente o fornire il flag per i tuoi lavori EMR `--packages Serverless`.

1. Quando invii lavori EMR Serverless, assicurati di avere le seguenti proprietà di configurazione e di includere i seguenti parametri nel campo. `sparkSubmitParameters`

```
--conf spark.jars=/usr/share/aws/delta/lib/delta-core.jar,/usr/share/aws/delta/lib/delta-storage.jar
--conf spark.sql.extensions=io.delta.sql.DeltaSparkSessionExtension
--conf
spark.sql.catalog.spark_catalog=org.apache.spark.sql.delta.catalog.DeltaCatalog
```

2. Crea un locale `delta_sample.py` per testare la creazione e la lettura di una tabella Delta.

```
# delta_sample.py
from pyspark.sql import SparkSession

import uuid

url = "s3://amzn-s3-demo-bucket/delta-lake/output/%s/" % str(uuid.uuid4())
spark = SparkSession.builder.appName("DeltaSample").getOrCreate()

## creates a Delta table and outputs to target S3 bucket
spark.range(5).write.format("delta").save(url)

## reads a Delta table and outputs to target S3 bucket
spark.read.format("delta").load(url).show
```

3. Utilizzando AWS CLI, carica il `delta_sample.py` file nel tuo bucket Amazon S3. Quindi utilizzare il `start-job-run` comando per inviare un lavoro a un'applicazione EMR Serverless esistente.

```
aws s3 cp delta_sample.py s3://amzn-s3-demo-bucket/code/

aws emr-serverless start-job-run \
  --application-id application-id \
  --execution-role-arn job-role-arn \
  --name emr-delta \
  --job-driver '{
    "sparkSubmit": {
      "entryPoint": "s3://amzn-s3-demo-bucket/code/delta_sample.py",
      "sparkSubmitParameters": "--conf spark.jars=/usr/share/
aws/delta/lib/delta-core.jar,/usr/share/aws/delta/lib/delta-storage.jar --
conf spark.sql.extensions=io.delta.sql.DeltaSparkSessionExtension --conf
spark.sql.catalog.spark_catalog=org.apache.spark.sql.delta.catalog.DeltaCatalog"
```

```
}  
}'
```

Per usare le librerie Python con Delta Lake, puoi aggiungere la `delta-core` libreria [impacchettandola come dipendenza](#) o [usandola come immagine personalizzata](#).

In alternativa, puoi usare `SparkContext.addPyFile` per aggiungere le librerie Python dal file `delta-core` JAR:

```
import glob  
from pyspark.sql import SparkSession  
  
spark = SparkSession.builder.getOrCreate()  
spark.sparkContext.addPyFile(glob.glob("/usr/share/aws/delta/lib/delta-core_*.jar")[0])
```

## Amazon EMR versioni 6.8.0 e precedenti

Se utilizzi Amazon EMR 6.8.0 o versioni precedenti, segui questi passaggi per utilizzare Delta Lake OSS con le tue applicazioni EMR Serverless.

1. Per creare una versione open source di [Delta Lake](#) compatibile con la versione di Spark sulla tua applicazione Amazon EMR Serverless, accedi a GitHub Delta e segui [le](#) istruzioni.
2. Carica le librerie Delta Lake in un bucket Amazon S3 nel tuo Account AWS
3. Quando invii lavori EMR Serverless nella configurazione dell'applicazione, includi i file JAR di Delta Lake che ora sono nel tuo bucket.

```
--conf spark.jars=s3://amzn-s3-demo-bucket/jars/delta-core_2.12-1.1.0.jar
```

4. Per assicurarti di poter leggere e scrivere da una tabella Delta, esegui un test di esempio.  
PySpark

```
from pyspark import SparkConf, SparkContext  
from pyspark.sql import HiveContext, SparkSession  
  
import uuid  
  
conf = SparkConf()  
sc = SparkContext(conf=conf)  
sqlContext = HiveContext(sc)
```

```
url = "s3://amzn-s3-demo-bucket/delta-lake/output/1.0.1/%s/" %
str(uuid.uuid4())

## creates a Delta table and outputs to target S3 bucket
session.range(5).write.format("delta").save(url)

## reads a Delta table and outputs to target S3 bucket
session.read.format("delta").load(url).show
```

## Invio di lavori EMR Serverless da Airflow

L'Amazon Provider di Apache Airflow fornisce operatori EMR Serverless. Per ulteriori informazioni sugli operatori, consulta [Amazon EMR Serverless Operators](#) nella documentazione di Apache Airflow.

Puoi usarlo `EmrServerlessCreateApplicationOperator` per creare un'applicazione Spark o Hive. Puoi anche utilizzarla `EmrServerlessStartJobOperator` per avviare uno o più lavori con la tua nuova applicazione.

Per utilizzare l'operatore con Amazon Managed Workflows for Apache Airflow (MWAA) con Airflow 2.2.2, aggiungi la riga seguente al `requirements.txt` file e aggiorna l'ambiente MWAA per utilizzare il nuovo file.

```
apache-airflow-providers-amazon==6.0.0
boto3>=1.23.9
```

Tieni presente che il supporto EMR Serverless è stato aggiunto alla versione 5.0.0 del provider Amazon. La release 6.0.0 è l'ultima versione compatibile con Airflow 2.2.2. È possibile utilizzare versioni successive con Airflow 2.4.3 su MWAA.

Il seguente esempio abbreviato mostra come creare un'applicazione, eseguire più job Spark e quindi arrestare l'applicazione. Un esempio completo è disponibile nell'archivio [EMR Serverless Samples](#). GitHub Per ulteriori dettagli sulla `sparkSubmit` configurazione, vedere. [Utilizzo delle configurazioni Spark quando si eseguono job EMR Serverless](#)

```
from datetime import datetime

from airflow import DAG
from airflow.providers.amazon.aws.operators.emr import (
    EmrServerlessCreateApplicationOperator,
```

```

    EmrServerlessStartJobOperator,
    EmrServerlessDeleteApplicationOperator,
)

# Replace these with your correct values
JOB_ROLE_ARN = "arn:aws:iam::account-id:role/emr_serverless_default_role"
S3_LOGS_BUCKET = "amzn-s3-demo-bucket"

DEFAULT_MONITORING_CONFIG = {
    "monitoringConfiguration": {
        "s3MonitoringConfiguration": {"logUri": f"s3://amzn-s3-demo-bucket/logs/"}
    },
}

with DAG(
    dag_id="example_endtoend_emr_serverless_job",
    schedule_interval=None,
    start_date=datetime(2021, 1, 1),
    tags=["example"],
    catchup=False,
) as dag:
    create_app = EmrServerlessCreateApplicationOperator(
        task_id="create_spark_app",
        job_type="SPARK",
        release_label="emr-6.7.0",
        config={"name": "airflow-test"},
    )

    application_id = create_app.output

    job1 = EmrServerlessStartJobOperator(
        task_id="start_job_1",
        application_id=application_id,
        execution_role_arn=JOB_ROLE_ARN,
        job_driver={
            "sparkSubmit": {
                "entryPoint": "local:///usr/lib/spark/examples/src/main/python/
pi_fail.py",
            }
        },
        configuration_overrides=DEFAULT_MONITORING_CONFIG,
    )

    job2 = EmrServerlessStartJobOperator(

```

```

    task_id="start_job_2",
    application_id=application_id,
    execution_role_arn=JOB_ROLE_ARN,
    job_driver={
        "sparkSubmit": {
            "entryPoint": "local:///usr/lib/spark/examples/src/main/python/pi.py",
            "entryPointArguments": ["1000"]
        }
    },
    configuration_overrides=DEFAULT_MONITORING_CONFIG,
)

delete_app = EmrServerlessDeleteApplicationOperator(
    task_id="delete_app",
    application_id=application_id,
    trigger_rule="all_done",
)

(create_app >> [job1, job2] >> delete_app)

```

## Utilizzo delle funzioni Hive definite dall'utente con EMR Serverless

Le funzioni definite dall'utente di Hive (UDFs) consentono di creare funzioni personalizzate per elaborare record o gruppi di record. In questo tutorial, utilizzerai un UDF di esempio con un'applicazione Amazon EMR Serverless preesistente per eseguire un processo che genera un risultato di query. Per informazioni su come configurare un'applicazione, consulta [Guida introduttiva ad Amazon EMR Serverless](#)

Per utilizzare un UDF con EMR Serverless

1. Vai alla pagina UDF [GitHub](#) per un esempio. Clona il repository e passa al ramo git che desideri utilizzare. Aggiorna il `pom.xml` file maven-compiler-plugin nel repository per avere una fonte. Aggiorna anche la configurazione della versione java di destinazione su `1.8`. Esegui `mvn package -DskipTests` per creare il file JAR che contiene il tuo campione UDFs.
2. Dopo aver creato il file JAR, caricalo nel tuo bucket S3 con il seguente comando.

```
aws s3 cp brickhouse-0.8.2-JS.jar s3://amzn-s3-demo-bucket/jars/
```

3. Crea un file di esempio per utilizzare una delle funzioni UDF di esempio. Salva questa query con nome `udf_example.q` e caricala nel tuo bucket S3.

```
add jar s3://amzn-s3-demo-bucket/jars/brickhouse-0.8.2-JS.jar;
CREATE TEMPORARY FUNCTION from_json AS 'brickhouse.udf.json.FromJsonUDF';
select from_json('{"key1":[0,1,2], "key2":[3,4,5,6], "key3":[7,8,9]}', map("",
  array(cast(0 as int))));
select from_json('{"key1":[0,1,2], "key2":[3,4,5,6], "key3":[7,8,9]}', map("",
  array(cast(0 as int))))["key1"][2];
```

#### 4. Invia il seguente lavoro Hive.

```
aws emr-serverless start-job-run \
  --application-id application-id \
  --execution-role-arn job-role-arn \
  --job-driver '{
    "hive": {
      "query": "s3://amzn-s3-demo-bucket/queries/udf_example.q",
      "parameters": "--hiveconf hive.exec.scratchdir=s3://amzn-s3-demo-bucket/
emr-serverless-hive/scratch --hiveconf hive.metastore.warehouse.dir=s3://'$BUCKET'/
emr-serverless-hive/warehouse"
    }
  }' --configuration-overrides '{
  "applicationConfiguration": [{
    "classification": "hive-site",
    "properties": {
      "hive.driver.cores": "2",
      "hive.driver.memory": "6G"
    }
  }],
  "monitoringConfiguration": {
    "s3MonitoringConfiguration": {
      "logUri": "s3://amzn-s3-demo-bucket/logs/"
    }
  }
}'
```

#### 5. Usa il `get-job-run` comando per controllare lo stato del tuo lavoro. Attendi che lo stato cambi in SUCCESS.

```
aws emr-serverless get-job-run --application-id application-id --job-run-id job-id
```

#### 6. Scarica i file di output con il seguente comando.

```
aws s3 cp --recursive s3://amzn-s3-demo-bucket/logs/applications/application-id/
jobs/job-id/HIVE_DRIVER/ .
```

Il stdout.gz file è simile al seguente.

```
{"key1": [0,1,2], "key2": [3,4,5,6], "key3": [7,8,9]}
2
```

## Utilizzo di immagini personalizzate con EMR Serverless

### Argomenti

- [Usa una versione Python personalizzata](#)
- [Usa una versione Java personalizzata](#)
- [Crea un'immagine di data science](#)
- [Elaborazione di dati geospaziali con Apache Sedona](#)
- [Informazioni sulle licenze per l'utilizzo di immagini personalizzate](#)

## Usa una versione Python personalizzata

Puoi creare un'immagine personalizzata per usare una versione diversa di Python. Per usare la versione 3.10 di Python per i job Spark, ad esempio, esegui il seguente comando:

```
FROM public.ecr.aws/emr-serverless/spark/emr-6.9.0:latest

USER root

# install python 3
RUN yum install -y gcc openssl-devel bzip2-devel libffi-devel tar gzip wget make
RUN wget https://www.python.org/ftp/python/3.10.0/Python-3.10.0.tgz && \
tar xzf Python-3.10.0.tgz && cd Python-3.10.0 && \
./configure --enable-optimizations && \
make altinstall

# EMRS will run the image as hadoop
USER hadoop:hadoop
```

Prima di inviare il job Spark, imposta le proprietà per utilizzare l'ambiente virtuale Python, come segue.

```
--conf spark.emr-serverless.driverEnv.PYSPARK_DRIVER_PYTHON=/usr/local/bin/python3.10
--conf spark.emr-serverless.driverEnv.PYSPARK_PYTHON=/usr/local/bin/python3.10
--conf spark.executorEnv.PYSPARK_PYTHON=/usr/local/bin/python3.10
```

## Usa una versione Java personalizzata

L'esempio seguente mostra come creare un'immagine personalizzata per utilizzare Java 11 per i job Spark.

```
FROM public.ecr.aws/emr-serverless/spark/emr-6.9.0:latest

USER root

# install JDK 11
RUN sudo amazon-linux-extras install java-openjdk11

# EMRS will run the image as hadoop
USER hadoop:hadoop
```

Prima di inviare il job Spark, imposta le proprietà di Spark per utilizzare Java 11, come segue.

```
--conf spark.executorEnv.JAVA_HOME=/usr/lib/jvm/java-11-
openjdk-11.0.16.0.8-1.amzn2.0.1.x86_64
--conf spark.emr-serverless.driverEnv.JAVA_HOME=/usr/lib/jvm/java-11-
openjdk-11.0.16.0.8-
```

## Crea un'immagine di data science

L'esempio seguente mostra come includere pacchetti Python comuni per la scienza dei dati, come Pandas e NumPy

```
FROM public.ecr.aws/emr-serverless/spark/emr-6.9.0:latest

USER root

# python packages
RUN pip3 install boto3 pandas numpy
```

```
RUN pip3 install -U scikit-learn==0.23.2 scipy
RUN pip3 install sk-dist
RUN pip3 install xgboost

# EMR Serverless will run the image as hadoop
USER hadoop:hadoop
```

## Elaborazione di dati geospaziali con Apache Sedona

L'esempio seguente mostra come creare un'immagine per includere Apache Sedona per l'elaborazione geospaziale.

```
FROM public.ecr.aws/emr-serverless/spark/emr-6.9.0:latest

USER root

RUN yum install -y wget
RUN wget https://repo1.maven.org/maven2/org/apache/sedona/sedona-core-3.0_2.12/1.3.0-incubating/sedona-core-3.0_2.12-1.3.0-incubating.jar -P /usr/lib/spark/jars/
RUN pip3 install apache-sedona

# EMRS will run the image as hadoop
USER hadoop:hadoop
```

## Informazioni sulle licenze per l'utilizzo di immagini personalizzate

È possibile creare immagini personalizzate con EMR Serverless per eseguire attività specifiche o utilizzare versioni specifiche di un pacchetto software. La modifica e la distribuzione di immagini personalizzate possono essere soggette a regole e termini di licenza. Il testo della licenza viene visualizzato nella sottosezione che segue.

### Licenze che si applicano alle immagini personalizzate

Copyright Amazon.com e sue affiliate; tutti i diritti riservati. Questo software è AWS contenuto ai sensi [AWS del Contratto con il cliente](#) e non può essere distribuito senza autorizzazione. Oltre alle autorizzazioni previste dalla licenza di [proprietàAWS intellettuale, il AWS Licenziante](#) concede all'utente le seguenti autorizzazioni aggiuntive:

È consentito creare, copiare e utilizzare derivati del AWS contenuto a condizione che siano soddisfatte le seguenti condizioni:

- L'utente non modifica il AWS Contenuto stesso e gli eventuali derivati sono strettamente il risultato dell'aggiunta di nuovi contenuti da parte dell'Utente.
- Le riproduzioni interne devono conservare la suddetta nota sul copyright.
- La distribuzione esterna, in formato sorgente o binario, con o senza modifiche, non è consentita dai termini di questa licenza.

Per ulteriori informazioni sull'utilizzo di immagini personalizzate, vedere [Utilizzo di immagini personalizzate con EMR Serverless](#).

## Utilizzo dell'integrazione di Amazon Redshift per Apache Spark su Amazon EMR Serverless

Con Amazon EMR rilascio 6.9.0 e successivi, ogni immagine del rilascio include un connettore tra [Apache Spark](#) e Amazon Redshift. Con questo connettore, puoi utilizzare Spark su Amazon EMR Serverless per elaborare i dati archiviati in Amazon Redshift. L'integrazione si basa sul [connettore spark-redshift open source](#). Per Amazon EMR Serverless, l'integrazione [Amazon Redshift per Apache Spark è inclusa come integrazione](#) nativa.

### Argomenti

- [Avvio di un'applicazione Spark con l'integrazione Amazon Redshift per Apache Spark](#)
- [Autenticazione con l'integrazione di Amazon Redshift per Apache Spark](#)
- [Lettura e scrittura da e su Amazon Redshift](#)
- [Considerazioni e limitazioni relative all'utilizzo del connettore Spark](#)

## Avvio di un'applicazione Spark con l'integrazione Amazon Redshift per Apache Spark

Per utilizzare l'integrazione con EMR Serverless 6.9.0, devi passare le dipendenze Spark-Redshift richieste con il tuo job Spark. `--jars`Da utilizzare per includere le librerie relative al connettore Redshift. Per vedere le altre posizioni dei file supportate dall'opzione `--jars`, consulta la sezione [Advanced Dependency Management](#) (Gestione avanzata delle dipendenze) nella documentazione di Apache Spark.

- `spark-redshift.jar`

- `spark-avro.jar`
- `RedshiftJDBC.jar`
- `minimal-json.jar`

Le versioni 6.10.0 e successive di Amazon EMR non richiedono la dipendenza `minimal-json.jar` e installano automaticamente le altre dipendenze su ciascun cluster per impostazione predefinita. Gli esempi seguenti mostrano come avviare un'applicazione Spark con l'integrazione di Amazon Redshift per Apache Spark.

#### Amazon EMR 6.10.0 +

Avvia un job Spark su Amazon EMR Serverless con l'integrazione Amazon Redshift per Apache Spark su EMR Serverless versione 6.10.0 e successive.

```
spark-submit my_script.py
```

#### Amazon EMR 6.9.0

Per avviare un job Spark su Amazon EMR Serverless con l'integrazione Amazon Redshift per Apache Spark su EMR Serverless release 6.9.0, utilizza l'opzione come illustrato nell'esempio seguente. `--jars` Come vedrai, i percorsi elencati con l'opzione `--jars` sono i percorsi predefiniti per i file JAR.

```
--jars
  /usr/share/aws/redshift/jdbc/RedshiftJDBC.jar,
  /usr/share/aws/redshift/spark-redshift/lib/spark-redshift.jar,
  /usr/share/aws/redshift/spark-redshift/lib/spark-avro.jar,
  /usr/share/aws/redshift/spark-redshift/lib/minimal-json.jar
```

```
spark-submit \
  --jars /usr/share/aws/redshift/jdbc/RedshiftJDBC.jar,/usr/share/aws/redshift/
spark-redshift/lib/spark-redshift.jar,/usr/share/aws/redshift/spark-redshift/lib/
spark-avro.jar,/usr/share/aws/redshift/spark-redshift/lib/minimal-json.jar \
  my_script.py
```

## Autenticazione con l'integrazione di Amazon Redshift per Apache Spark

### Utilizzalo AWS Secrets Manager per recuperare le credenziali e connetterti ad Amazon Redshift

Puoi autenticarti in modo sicuro su Amazon Redshift archiviando le credenziali in Secrets Manager e facendo in modo che il job Spark `GetSecretValue` chiami l'API per recuperarle:

```
from pyspark.sql import SQLContextimport boto3

sc = # existing SparkContext
sql_context = SQLContext(sc)

secretsmanager_client = boto3.client('secretsmanager',
    region_name=os.getenv('AWS_REGION'))
secret_manager_response = secretsmanager_client.get_secret_value(
    SecretId='string',
    VersionId='string',
    VersionStage='string'
)
username = # get username from secret_manager_response
password = # get password from secret_manager_response
url = "jdbc:redshift://redshifthost:5439/database?user=" + username + "&password="
    + password

# Access to Redshift cluster using Spark
```

### Autenticazione su Amazon Redshift con un driver JDBC

Impostazione di nome utente e password all'interno dell'URL di JDBC

Puoi autenticare un job Spark in un cluster Amazon Redshift specificando il nome e la password del database Amazon Redshift nell'URL JDBC.

#### Note

Se passi le credenziali del database nell'URL, anche chiunque abbia accesso all'URL può accedere alle credenziali. Questo metodo non è generalmente consigliato perché non è un'opzione sicura.

Se la sicurezza non è un problema per la tua applicazione, puoi utilizzare il formato seguente per impostare il nome utente e la password nell'URL di JDBC:

```
jdbc:redshift://redshifthost:5439/database?user=username&password=password
```

## Usa l'autenticazione basata su IAM con il ruolo di esecuzione dei job di Amazon EMR Serverless

A partire dalla versione 6.9.0 di Amazon EMR Serverless, il driver Amazon Redshift JDBC 2.1 o superiore viene integrato nell'ambiente. Con il driver JDBC 2.1 e versioni successive, è possibile specificare l'URL di JDBC evitando di includere il nome utente e la password non elaborati.

È invece possibile specificare uno schema `jdbc:redshift:iam://`. Ciò ordina al driver JDBC di utilizzare il ruolo di esecuzione del lavoro EMR Serverless per recuperare automaticamente le credenziali. Per ulteriori informazioni, consulta la sezione [Configure a JDBC or ODBC connection to use IAM credentials](#) (Configurazione di una connessione JDBC oppure ODBC per l'utilizzo delle credenziali IAM) nella Guida alla gestione di Amazon Redshift. Un esempio di questo URL è:

```
jdbc:redshift:iam://examplecluster.abc123xyz789.us-west-2.redshift.amazonaws.com:5439/  
dev
```

Le seguenti autorizzazioni sono necessarie per il ruolo di esecuzione del lavoro quando sono soddisfatte le condizioni fornite:

Autorizzazione	Condizioni richieste per il ruolo di esecuzione di processo
<code>redshift:GetClusterCredentials</code>	Obbligatoria affinché il driver JDBC recuperi le credenziali da Amazon Redshift
<code>redshift:DescribeCluster</code>	Obbligatoria se specifichi il cluster Amazon Redshift e Regione AWS nell'URL di JDBC anziché nell'endpoint
<code>redshift-serverless:GetCredentials</code>	Obbligatoria affinché il driver JDBC recuperi le credenziali da Amazon Redshift Serverless
<code>redshift-serverless:GetWorkgroup</code>	Obbligatorio se utilizzi Amazon Redshift Serverless e specifichi l'URL in termini di nome e regione del gruppo di lavoro

## Connessione ad Amazon Redshift all'interno di un altro VPC

Quando configuri un cluster Amazon Redshift o un gruppo di lavoro Amazon Redshift Serverless con provisioning sotto un VPC, devi configurare la connettività VPC per l'applicazione Amazon EMR Serverless per accedere alle risorse. Per ulteriori informazioni su come configurare la connettività VPC su un'applicazione EMR Serverless, vedere. [Configurazione dell'accesso VPC per le applicazioni EMR Serverless per la connessione ai dati](#)

- Se il cluster Amazon Redshift o il gruppo di lavoro Serverless Amazon Redshift fornito sono accessibili al pubblico, puoi specificare una o più sottoreti private con un gateway NAT collegato quando crei applicazioni EMR Serverless.
- Se il cluster Amazon Redshift o il gruppo di lavoro Serverless Amazon Redshift fornito non sono accessibili al pubblico, devi creare un endpoint VPC gestito da Amazon Redshift per il tuo cluster Amazon Redshift come descritto in. [Configurazione dell'accesso VPC per le applicazioni EMR Serverless per la connessione ai dati](#) In alternativa, puoi creare il tuo gruppo di lavoro Amazon Redshift Serverless come descritto in Connessione ad [Amazon Redshift Serverless nella Amazon Redshift Management Guide](#). È necessario associare il cluster o il sottogruppo alle sottoreti private specificate al momento della creazione dell'applicazione EMR Serverless.

### Note

Se utilizzi l'autenticazione basata su IAM e le tue sottoreti private per l'applicazione EMR Serverless non dispongono di un gateway NAT collegato, devi anche creare un endpoint VPC su tali sottoreti per Amazon Redshift o Amazon Redshift Serverless. In questo modo, il driver JDBC può recuperare le credenziali.

## Lettura e scrittura da e su Amazon Redshift

I seguenti esempi di codice consentono PySpark di leggere e scrivere dati di esempio da e verso un database Amazon Redshift con un'API di origine dati e con SparkSQL.

### Data source API

PySpark Utilizzalo per leggere e scrivere dati di esempio da e verso un database Amazon Redshift con API di origine dati.

```
import boto3
```

```

from pyspark.sql import SQLContext

sc = # existing SparkContext
sql_context = SQLContext(sc)

url = "jdbc:redshift:iam://redshifthost:5439/database"
aws_iam_role_arn = "arn:aws:iam::account-id:role/role-name"

df = sql_context.read \
    .format("io.github.spark_redshift_community.spark.redshift") \
    .option("url", url) \
    .option("dbtable", "table-name") \
    .option("tempdir", "s3://path/for/temp/data") \
    .option("aws_iam_role", "aws-iam-role-arn") \
    .load()

df.write \
    .format("io.github.spark_redshift_community.spark.redshift") \
    .option("url", url) \
    .option("dbtable", "table-name-copy") \
    .option("tempdir", "s3://path/for/temp/data") \
    .option("aws_iam_role", "aws-iam-role-arn") \
    .mode("error") \
    .save()

```

## SparkSQL

PySpark Utilizzalo per leggere e scrivere dati di esempio da e verso un database Amazon Redshift con SparkSQL.

```

import boto3
import json
import sys
import os
from pyspark.sql import SparkSession

spark = SparkSession \
    .builder \
    .enableHiveSupport() \
    .getOrCreate()

url = "jdbc:redshift:iam://redshifthost:5439/database"
aws_iam_role_arn = "arn:aws:iam::account-id:role/role-name"

```

```

bucket = "s3://path/for/temp/data"
tableName = "table-name" # Redshift table name

s = f"""CREATE TABLE IF NOT EXISTS {table-name} (country string, data string)
  USING io.github.spark_redshift_community.spark.redshift
  OPTIONS (dbtable '{table-name}', tempdir '{bucket}', url '{url}', aws_iam_role
    '{aws-iam-role-arn}'); """

spark.sql(s)

columns = ["country" ,"data"]
data = [("test-country", "test-data")]
df = spark.sparkContext.parallelize(data).toDF(columns)

# Insert data into table
df.write.insertInto(table-name, overwrite=False)
df = spark.sql(f"SELECT * FROM {table-name}")
df.show()

```

## Considerazioni e limitazioni relative all'utilizzo del connettore Spark

- Si consiglia di attivare SSL per la connessione JDBC da Spark su Amazon EMR ad Amazon Redshift.
- Come best practice, è consigliabile gestire le credenziali per il cluster Amazon Redshift in AWS Secrets Manager . [AWS Secrets Manager Per un esempio, consulta Utilizzo per recuperare le credenziali per la connessione ad Amazon Redshift.](#)
- Si consiglia di passare un ruolo IAM con il parametro `aws_iam_role` per il parametro di autenticazione di Amazon Redshift.
- Il parametro `tempformat` attualmente non supporta il formato Parquet.
- L'URI `tempdir` indica una posizione Amazon S3. Questa directory temporanea non viene pulita in automatico e quindi potrebbe generare costi aggiuntivi.
- Prendi in considerazione i seguenti consigli per Amazon Redshift:
  - Si consiglia di bloccare l'accesso pubblico al cluster Amazon Redshift.
  - Si consiglia di attivare la [registrazione di log di verifica di Amazon Redshift.](#)
  - Si consiglia di attivare la [crittografia dei dati inattivi di Amazon Redshift.](#)
- Prendi in considerazione i seguenti consigli per Amazon S3:

- Si consiglia di [bloccare l'accesso pubblico ai bucket Amazon S3](#).
- Si consiglia di utilizzare la [crittografia lato server di Amazon S3](#) per crittografare i bucket Amazon S3 utilizzati.
- Si consiglia di utilizzare le [policy del ciclo di vita di Amazon S3](#) per definire le regole di conservazione del bucket Amazon S3.
- Amazon EMR verifica sempre il codice importato dall'open source nell'immagine. Per motivi di sicurezza, non supportiamo i seguenti metodi di autenticazione da Spark ad Amazon S3:
  - Impostazione delle chiavi di AWS accesso nella classificazione della configurazione `hadoop-env`
  - Codifica delle chiavi di AWS accesso nell'URI `tempdir`

Per ulteriori informazioni sull'utilizzo del connettore e dei parametri supportati, consulta le seguenti risorse:

- [Amazon Redshift integration for Apache Spark](#) (Integrazione di Amazon Redshift per Apache Spark) nella Guida alla gestione di Amazon Redshift
- Il [repository della community spark-redshift](#) su Github

## Connessione a DynamoDB con Amazon EMR Serverless

In questo tutorial, carichi un sottoinsieme di dati dal [Board on Geographic Names degli Stati Uniti su](#) un bucket Amazon S3 e poi usi Hive o Spark su Amazon EMR Serverless per copiare i dati in una tabella Amazon DynamoDB su cui eseguire query.

### Fase 1: caricare i dati in un bucket Amazon S3

Per creare un bucket Amazon S3, segui le istruzioni in [Creazione di un bucket nella Guida per l'utente della console](#) di Amazon Simple Storage Service. Sostituisci i riferimenti a `amzn-s3-demo-bucket` con il nome del bucket appena creato. Ora la tua applicazione EMR Serverless è pronta per eseguire i lavori.

1. Scarica l'archivio di dati di esempio `features.zip` con il seguente comando.

```
wget https://docs.aws.amazon.com/amazondynamodb/latest/developerguide/samples/features.zip
```

## 2. Estrai i `features.txt` file dall'archivio e visualizza le prime poche righe del file:

```
unzip features.zip
head features.txt
```

Il risultato dovrebbe essere simile al seguente.

```
1535908|Big Run|Stream|WV|38.6370428|-80.8595469|794
875609|Constable Hook|Cape|NJ|40.657881|-74.0990309|7
1217998|Gooseberry Island|Island|RI|41.4534361|-71.3253284|10
26603|Boone Moore Spring|Spring|AZ|34.0895692|-111.410065|3681
1506738|Missouri Flat|Flat|WA|46.7634987|-117.0346113|2605
1181348|Minnow Run|Stream|PA|40.0820178|-79.3800349|1558
1288759|Hunting Creek|Stream|TN|36.343969|-83.8029682|1024
533060|Big Charles Bayou|Bay|LA|29.6046517|-91.9828654|0
829689|Greenwood Creek|Stream|NE|41.596086|-103.0499296|3671
541692|Button Willow Island|Island|LA|31.9579389|-93.0648847|98
```

I campi in ogni riga qui indicano un identificatore univoco, un nome, un tipo di elemento naturale, lo stato, la latitudine in gradi, la longitudine in gradi e l'altezza in piedi.

## 3. Carica i tuoi dati su Amazon S3

```
aws s3 cp features.txt s3://amzn-s3-demo-bucket/features/
```

## Fase 2: Creare una tabella Hive

Usa Apache Spark o Hive per creare una nuova tabella Hive che contenga i dati caricati in Amazon S3.

### Spark

Per creare una tabella Hive con Spark, esegui il seguente comando.

```
import org.apache.spark.sql.Session

val sparkSession = SparkSession.builder().enableHiveSupport().getOrCreate()

sparkSession.sql("CREATE TABLE hive_features \
  (feature_id BIGINT, \
  feature_name STRING, \
```

```
feature_class STRING, \
state_alpha STRING, \
prim_lat_dec DOUBLE, \
prim_long_dec DOUBLE, \
elev_in_ft BIGINT) \
ROW FORMAT DELIMITED \
FIELDS TERMINATED BY '|' \
LINES TERMINATED BY '\n' \
LOCATION 's3://amzn-s3-demo-bucket/features';")
```

Ora hai una tabella Hive popolata con i dati del file. `features.txt` Per verificare che i dati siano presenti nella tabella, esegui una query SQL Spark come mostrato nell'esempio seguente.

```
sparkSession.sql(
  "SELECT state_alpha, COUNT(*) FROM hive_features GROUP BY state_alpha;")
```

## Hive

Per creare una tabella Hive con Hive, esegui il comando seguente.

```
CREATE TABLE hive_features
(feature_id          BIGINT,
feature_name        STRING ,
feature_class       STRING ,
state_alpha         STRING,
prim_lat_dec        DOUBLE ,
prim_long_dec       DOUBLE ,
elev_in_ft          BIGINT)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '|'
LINES TERMINATED BY '\n'
LOCATION 's3://amzn-s3-demo-bucket/features';
```

Ora hai una tabella Hive che contiene i dati del file. `features.txt` Per verificare che i dati siano presenti nella tabella, esegui una query HiveQL, come illustrato nell'esempio seguente.

```
SELECT state_alpha, COUNT(*) FROM hive_features GROUP BY state_alpha;
```

## Fase 3: Copiare i dati su DynamoDB

Usa Spark o Hive per copiare i dati in una nuova tabella DynamoDB.

## Spark

[Per copiare i dati dalla tabella Hive creata nel passaggio precedente in DynamoDB, segui i passaggi 1-3 in Copiare i dati su DynamoDB.](#) Questo crea una nuova tabella DynamoDB chiamata `Features`. È quindi possibile leggere i dati direttamente dal file di testo e copiarli nella tabella DynamoDB, come illustrato nell'esempio seguente.

```
import com.amazonaws.services.dynamodbv2.model.AttributeValue
import org.apache.hadoop.dynamodb.DynamoDBItemWritable
import org.apache.hadoop.dynamodb.read.DynamoDBInputFormat
import org.apache.hadoop.io.Text
import org.apache.hadoop.mapred.JobConf
import org.apache.spark.SparkContext

import scala.collection.JavaConverters._

object EmrServerlessDynamoDbTest {

  def main(args: Array[String]): Unit = {

    jobConf.set("dynamodb.input.tableName", "Features")
    jobConf.set("dynamodb.output.tableName", "Features")
    jobConf.set("dynamodb.region", "region")

    jobConf.set("mapred.output.format.class",
"org.apache.hadoop.dynamodb.write.DynamoDBOutputFormat")
    jobConf.set("mapred.input.format.class",
"org.apache.hadoop.dynamodb.read.DynamoDBInputFormat")

    val rdd = sc.textFile("s3://amzn-s3-demo-bucket/ddb-connector/")
      .map(row => {
        val line = row.split("\\|")
        val item = new DynamoDBItemWritable()

        val elevInFt = if (line.length > 6) {
          new AttributeValue().withN(line(6))
        } else {
          new AttributeValue().withNULL(true)
        }

        item.setItem(Map(
          "feature_id" -> new AttributeValue().withN(line(0)),
          "feature_name" -> new AttributeValue(line(1)),
```

```

        "feature_class" -> new AttributeValue(line(2)),
        "state_alpha" -> new AttributeValue(line(3)),
        "prim_lat_dec" -> new AttributeValue().withN(line(4)),
        "prim_long_dec" -> new AttributeValue().withN(line(5)),
        "elev_in_ft" -> elevInFt)
        .asJava)
    (new Text(""), item)
    })
    rdd.saveAsHadoopDataset(jobConf)
}
}

```

## Hive

Per copiare i dati dalla tabella Hive creata nel passaggio precedente in DynamoDB, segui le istruzioni in [Copiare](#) i dati su DynamoDB.

## Fase 4: Interrogare i dati da DynamoDB

Usa Spark o Hive per interrogare la tua tabella DynamoDB.

### Spark

Per interrogare i dati dalla tabella DynamoDB creata nel passaggio precedente, puoi utilizzare Spark SQL o l'API Spark. MapReduce

Example — Interroga la tua tabella DynamoDB con Spark SQL

La seguente query SQL Spark restituisce un elenco di tutti i tipi di funzionalità in ordine alfabetico.

```

val dataframe = sparkSession.sql("SELECT DISTINCT feature_class \
FROM ddb_features \
ORDER BY feature_class;")

```

La seguente query SQL Spark restituisce un elenco di tutti i lake che iniziano con la lettera M.

```

val dataframe = sparkSession.sql("SELECT feature_name, state_alpha \
FROM ddb_features \
WHERE feature_class = 'Lake' \
AND feature_name LIKE 'M%' \
ORDER BY feature_name;")

```

La seguente query SQL Spark restituisce un elenco di tutti gli stati con almeno tre funzionalità superiori a un miglio.

```
val dataframe = sparkSession.dql("SELECT state_alpha, feature_class, COUNT(*) \
  FROM ddb_features \
  WHERE elev_in_ft > 5280 \
  GROUP by state_alpha, feature_class \
  HAVING COUNT(*) >= 3 \
  ORDER BY state_alpha, feature_class;")
```

**Example —** Interroga la tua tabella DynamoDB con l'API Spark MapReduce

La seguente MapReduce query restituisce un elenco di tutti i tipi di feature in ordine alfabetico.

```
val df = sc.hadoopRDD(jobConf, classOf[DynamoDBInputFormat], classOf[Text],
  classOf[DynamoDBItemWritable])
  .map(pair => (pair._1, pair._2.getItem))
  .map(pair => pair._2.get("feature_class").getS)
  .distinct()
  .sortBy(value => value)
  .toDF("feature_class")
```

La seguente MapReduce query restituisce un elenco di tutti i laghi che iniziano con la lettera M.

```
val df = sc.hadoopRDD(jobConf, classOf[DynamoDBInputFormat], classOf[Text],
  classOf[DynamoDBItemWritable])
  .map(pair => (pair._1, pair._2.getItem))
  .filter(pair => "Lake".equals(pair._2.get("feature_class").getS))
  .filter(pair => pair._2.get("feature_name").getS.startsWith("M"))
  .map(pair => (pair._2.get("feature_name").getS,
  pair._2.get("state_alpha").getS))
  .sortBy(_._1)
  .toDF("feature_name", "state_alpha")
```

La seguente MapReduce query restituisce un elenco di tutti gli stati con almeno tre caratteristiche superiori a un miglio.

```
val df = sc.hadoopRDD(jobConf, classOf[DynamoDBInputFormat], classOf[Text],
  classOf[DynamoDBItemWritable])
  .map(pair => pair._2.getItem)
  .filter(pair => pair.get("elev_in_ft").getN != null)
  .filter(pair => Integer.parseInt(pair.get("elev_in_ft").getN) > 5280)
```

```
.groupBy(pair => (pair.get("state_alpha").getS, pair.get("feature_class").getS))
.filter(pair => pair._2.size >= 3)
.map(pair => (pair._1._1, pair._1._2, pair._2.size))
.sortBy(pair => (pair._1, pair._2))
.toDF("state_alpha", "feature_class", "count")
```

## Hive

Per interrogare i dati dalla tabella DynamoDB creata nel passaggio precedente, segui le istruzioni in [Interrogare i dati nella](#) tabella DynamoDB.

## Configurazione dell'accesso multi-account

Per configurare l'accesso tra più account per EMR Serverless, completare i seguenti passaggi. Nell'esempio, AccountA è l'account in cui hai creato l'applicazione Amazon EMR Serverless ed AccountB è l'account in cui si trova Amazon DynamoDB.

1. Crea una tabella DynamoDB in AccountB. Per ulteriori informazioni, consulta [Fase 1: Creare una tabella](#).
2. Crea un ruolo Cross-Account-Role-B IAM in AccountB grado di accedere alla tabella DynamoDB.
  - a. Accedi AWS Management Console e apri la console IAM all'indirizzo. <https://console.aws.amazon.com/iam/>
  - b. Scegli Ruoli e crea un nuovo ruolo chiamato Cross-Account-Role-B. Per ulteriori informazioni su come creare ruoli IAM, consulta [Creazione di ruoli IAM](#) nella guida per l'utente.
  - c. Crea una policy IAM che conceda le autorizzazioni per accedere alla tabella DynamoDB tra account. Quindi, allega la policy IAM a Cross-Account-Role-B.

Di seguito è riportata una politica che concede l'accesso a una tabella DynamoDB.  
CrossAccountTable

```
{"Version": "2012-10-17",
  "Statement": [
    {"Effect": "Allow",
      "Action": "dynamodb:*",
      "Resource": "arn:aws:dynamodb:region:AccountB:table/
CrossAccountTable"}
```

```

    }
  ]
}

```

- d. Modifica la relazione di fiducia per il ruolo `Cross-Account-Role-B`.

Per configurare la relazione di trust per il ruolo, scegli la scheda `Trust Relationships` nella console IAM per il ruolo che hai creato nel passaggio 2: `Cross-Account-Role-B`

Seleziona `Modifica relazione di fiducia`, quindi aggiungi il seguente documento politico. Questo documento consente `Job-Execution-Role-A` AccountA di assumere questo `Cross-Account-Role-B` ruolo.

```

{"Version": "2012-10-17",
 "Statement": [
  {"Effect": "Allow",
   "Principal": {"AWS": "arn:aws:iam::AccountA:role/Job-Execution-Role-A"},
   "Action": "sts:AssumeRole"}
 ]
}

```

- e. Concedi `Job-Execution-Role-A` AccountA con `STS Assume role` le autorizzazioni da assumere `Cross-Account-Role-B`.

Nella console IAM per Account AWS AccountA, seleziona `Job-Execution-Role-A`. Aggiungi la seguente istruzione di policy a `Job-Execution-Role-A` per autorizzare l'operazione `AssumeRole` nel ruolo `Cross-Account-Role-B`.

```

{"Version": "2012-10-17",
 "Statement": [
  {"Effect": "Allow",
   "Action": "sts:AssumeRole",
   "Resource": "arn:aws:iam::AccountB:role/Cross-Account-Role-B"}
 ]
}

```

- f. Imposta la `dynamodb.customAWSCredentialsProvider` proprietà con il valore come `com.amazonaws.emr.AssumeRoleAWSCredentialsProvider` nella classificazione del

sito principale. Imposta la variabile di ambiente `ASSUME_ROLE_CREDENTIALS_ROLE_ARN` con il valore ARN di `Cross-Account-Role-B`

3. Esegui il job Spark o Hive utilizzando `Job-Execution-Role-A`

## Considerazioni

Nota questi comportamenti e limitazioni quando utilizzi il connettore DynamoDB con Apache Spark o Apache Hive.

### Considerazioni sull'utilizzo del connettore DynamoDB con Apache Spark

- Spark SQL non supporta la creazione di una tabella Hive con l'opzione `storage-handler`. Per ulteriori informazioni, consulta [Specificare il formato di archiviazione per le tabelle Hive](#) nella documentazione di Apache Spark.
- Spark SQL non supporta il funzionamento con lo storage handler `STORED BY`. Se desideri interagire con una tabella DynamoDB tramite una tabella Hive esterna, usa Hive per creare prima la tabella.
- Per tradurre una query in una query DynamoDB, il connettore DynamoDB utilizza il pushdown dei predicati. Predicate pushdown filtra i dati in base a una colonna mappata alla chiave di partizione di una tabella DynamoDB. Il predicate pushdown funziona solo quando si utilizza il connettore con Spark SQL e non con l'API MapReduce.

### Considerazioni sull'utilizzo del connettore DynamoDB con Apache Hive

#### Ottimizzazione del numero massimo di mappatori

- Se si utilizza la `SELECT` query per leggere i dati da una tabella Hive esterna mappata a DynamoDB, il numero di attività di mappa su EMR Serverless viene calcolato come il throughput di lettura totale configurato per la tabella DynamoDB, diviso per il throughput per task di mappa. La velocità effettiva predefinita per attività di mappa è 100.
- Il job Hive può utilizzare il numero di attività di mappatura oltre al numero massimo di contenitori configurati per applicazione EMR Serverless, a seconda del throughput di lettura configurato per DynamoDB. Inoltre, una query Hive a esecuzione prolungata può consumare tutta la capacità di lettura assegnata alla tabella DynamoDB. Ciò ha un impatto negativo sugli altri utenti.

- È possibile utilizzare la `dynamodb.max.map.tasks` proprietà per impostare un limite superiore per le attività di mappatura. È inoltre possibile utilizzare questa proprietà per ottimizzare la quantità di dati letti da ogni attività della mappa in base alla dimensione del contenitore dell'attività.
- È possibile impostare la `dynamodb.max.map.tasks` proprietà a livello di query Hive o nella `hive-site` classificazione del `start-job-run` comando. Questo valore deve essere maggiore o uguale a 1. Quando Hive elabora la tua query, il job Hive risultante utilizza solo i valori di `dynamodb.max.map.tasks` quando legge dalla tabella DynamoDB.

### Ottimizzazione della velocità di scrittura per attività

- Il throughput di scrittura per attività su EMR Serverless viene calcolato come il throughput di scrittura totale configurato per una tabella DynamoDB, diviso per il valore della proprietà `mapreduce.job.maps`. Per Hive, il valore predefinito di questa proprietà è 2. Pertanto, le prime due attività nella fase finale del processo Hive possono consumare tutta la velocità di scrittura. Ciò comporta una riduzione delle scritture di altre attività nello stesso lavoro o in altri lavori.
- Per evitare la limitazione della scrittura, è possibile impostare il valore della `mapreduce.job.maps` proprietà in base al numero di attività nella fase finale o alla velocità di scrittura che si desidera allocare per attività. Imposta questa proprietà nella `mapred-site` classificazione del `start-job-run` comando su EMR Serverless.

# Sicurezza

Per, la sicurezza del cloud ha la massima priorità. AWS In quanto AWS cliente, è possibile trarre vantaggio da un'architettura di data center e di rete progettata per soddisfare i requisiti delle aziende più esigenti a livello di sicurezza.

La sicurezza è una responsabilità condivisa tra AWS e l'utente. Il [modello di responsabilità condivisa](#) descrive questo come sicurezza del cloud e sicurezza nel cloud:

- Sicurezza del cloud: AWS è responsabile della protezione dell'infrastruttura che esegue AWS i servizi nel AWS cloud. AWS fornisce, inoltre, i servizi che puoi utilizzare in modo sicuro. I revisori di terze parti testano e verificano regolarmente l'efficacia della sicurezza come parte dei [programmi di conformitàAWS](#). Per ulteriori informazioni sui programmi di conformità che si applicano a Amazon EMR Serverless, consulta [AWS Servizi coperti dal programma di conformità](#).
- Sicurezza nel cloud: la tua responsabilità è determinata dal AWS servizio che utilizzi. Inoltre, sei responsabile anche di altri fattori, tra cui la riservatezza dei dati, i requisiti dell'azienda e le leggi e le normative applicabili.

Questa documentazione aiuta a comprendere come applicare il modello di responsabilità condivisa quando si utilizza Amazon EMR Serverless. Gli argomenti in questo capitolo mostrano come configurare Amazon EMR Serverless e utilizzare altri AWS servizi per soddisfare gli obiettivi di sicurezza e conformità.

## Argomenti

- [Best practice di sicurezza per Amazon EMR Serverless](#)
- [Protezione dei dati](#)
- [Identity and Access Management \(IAM\) in Amazon EMR Serverless](#)
- [Utilizzo di EMR Serverless con Lake Formation](#)
- [Crittografia tra lavoratori](#)
- [Secrets Manager per la protezione dei dati con EMR Serverless](#)
- [Utilizzo di Amazon S3 Access Grants con EMR Serverless](#)
- [Registrazione delle chiamate API Serverless di Amazon EMR utilizzando AWS CloudTrail](#)
- [Convalida della conformità per Amazon EMR Serverless](#)
- [Resilienza in Amazon EMR Serverless](#)

- [Sicurezza dell'infrastruttura in Amazon EMR Serverless](#)
- [Analisi della configurazione e delle vulnerabilità in Amazon EMR Serverless](#)

## Best practice di sicurezza per Amazon EMR Serverless

Amazon EMR Serverless offre una serie di funzionalità di sicurezza da prendere in considerazione durante lo sviluppo e l'implementazione delle proprie politiche di sicurezza. Le seguenti best practice sono linee guida generali e non rappresentano una soluzione di sicurezza completa. Poiché queste best practice potrebbero non essere appropriate o sufficienti per l'ambiente, gestiscile come considerazioni utili anziché prescrizioni.

### Applicazione del principio del privilegio minimo

EMR Serverless fornisce una policy di accesso granulare per le applicazioni che utilizzano ruoli IAM, come i ruoli di esecuzione. Si consiglia di concedere ai ruoli di esecuzione solo il set minimo di privilegi richiesti dal processo, ad esempio la copertura dell'applicazione e l'accesso alla destinazione del log. Si consiglia inoltre di verificare i processi per le autorizzazioni su base regolare e a qualsiasi modifica del codice dell'applicazione.

### Isolamento di un codice dell'applicazione non attendibile

EMR Serverless crea l'isolamento completo della rete tra i lavori appartenenti a diverse applicazioni EMR Serverless. Nei casi in cui si desidera l'isolamento a livello di processo, si consiglia di isolare i lavori in diverse applicazioni EMR Serverless.

### Autorizzazioni per il controllo degli accessi basato su ruoli (RBAC)

Gli amministratori devono controllare rigorosamente le autorizzazioni RBAC (Role-based access control) per le applicazioni EMR Serverless.

### Protezione dei dati

Il [modello di responsabilità AWS condivisa](#) si applica alla protezione dei dati in Amazon EMR Serverless. Come descritto in questo modello, AWS è responsabile della protezione dell'infrastruttura globale che gestisce tutto il AWS cloud. L'utente è responsabile di mantenere il controllo sui contenuti ospitati su questa infrastruttura. Questo contenuto include le attività di configurazione e gestione della sicurezza per i AWS servizi che utilizzi. Per ulteriori informazioni sulla privacy dei dati, consulta

[Domande frequenti sulla privacy dei dati](#). Per informazioni sulla protezione dei dati in Europa, consulta [il modello di responsabilità AWS condivisa e il post sul blog sul GDPR](#) sul AWS Security Blog.

Ai fini della protezione dei dati, ti consigliamo di proteggere le credenziali degli AWS account e di configurare singoli account con AWS Identity and Access Management (IAM). In questo modo, a ogni utente verranno assegnate solo le autorizzazioni necessarie per svolgere il proprio lavoro. Ti suggeriamo, inoltre, di proteggere i dati nei seguenti modi:

- Utilizza l'autenticazione a più fattori (MFA) con ogni account.
- Utilizza SSL/TLS per comunicare con le risorse. AWS È consigliabile TLS 1.2 o versioni successive.
- Configura l'API e la registrazione delle attività degli utenti con. AWS CloudTrail
- Utilizza soluzioni di AWS crittografia, insieme a tutti i controlli di sicurezza predefiniti all'interno AWS dei servizi.
- Utilizza i servizi di sicurezza gestiti avanzati, ad esempio Amazon Macie, che aiutano a individuare e proteggere i dati personali archiviati in Amazon S3.
- Utilizza le opzioni di crittografia Serverless di Amazon EMR per crittografare i dati a riposo e in transito.
- Se hai bisogno di moduli crittografici convalidati FIPS 140-2 per l'accesso AWS tramite un'interfaccia a riga di comando o un'API, utilizza un endpoint FIPS. Per ulteriori informazioni sugli endpoint FIPS disponibili, consulta il [Federal Information Processing Standard \(FIPS\) 140-2](#).

Consigliamo di non inserire mai informazioni identificative sensibili, ad esempio i numeri di account dei clienti, in campi a formato libero come un campo Nome. Ciò include quando lavori con Amazon EMR Serverless o altri AWS servizi utilizzando la console, l'API o. AWS CLI AWS SDKs Tutti i dati che inserisci in Amazon EMR Serverless o in altri servizi potrebbero essere raccolti per essere inclusi nei log di diagnostica. Quando fornisci un URL a un server esterno, non includere informazioni sulle credenziali nell'URL per convalidare la tua richiesta a tale server.

## Crittografia a riposo

La crittografia dei dati consente di impedire agli utenti non autorizzati di leggere dati su un cluster e sui sistemi di archiviazione di dati associati. Sono inclusi i dati salvati su supporti persistenti, noti come dati a riposo, e i dati che possono essere intercettati quando circolano in rete, noti come dati in transito.

La crittografia dei dati richiede chiavi e certificati. Puoi scegliere tra diverse opzioni, tra cui chiavi gestite da AWS Key Management Service, chiavi gestite da Amazon S3 e chiavi e certificati di fornitori personalizzati da te forniti. Quando lo utilizzi AWS KMS come fornitore di chiavi, vengono addebitati costi per l'archiviazione e l'uso delle chiavi di crittografia. Per ulteriori informazioni, consulta [Prezzi di AWS KMS](#).

Prima di specificare le opzioni di crittografia, scegli i sistemi di gestione di chiavi e certificati che intendi utilizzare. Quindi, crea le chiavi e i certificati per i provider personalizzati specificati come parte delle impostazioni di crittografia.

## Crittografia dei dati a riposo per dati EMRFS in Amazon S3

Ogni applicazione EMR Serverless utilizza una versione di release specifica, che include EMRFS (EMR File System). La crittografia di Amazon S3 funziona con gli oggetti del file system EMR (EMRFS) letti da e scritti su Amazon S3. Puoi specificare la crittografia lato server (SSE) o la crittografia lato client (CSE) di Amazon S3 come modalità di crittografia predefinita quando abiliti la crittografia a riposo. Facoltativamente, è possibile specificare diversi metodi di crittografia per singoli bucket utilizzando override di crittografia per bucket. Indipendentemente dall'abilitazione o meno della crittografia di Amazon S3, il protocollo Transport Layer Security (TLS) esegue la crittografia degli oggetti EMRFS in transito tra i nodi cluster EMR e Amazon S3. Se utilizzi Amazon S3 CSE con chiavi gestite dal cliente, il ruolo di esecuzione utilizzato per eseguire lavori in un'applicazione EMR Serverless deve avere accesso alla chiave. Per informazioni approfondite sulla crittografia di Amazon S3, [consulta Protection data using](#) encryption nella Amazon Simple Storage Service Developer Guide.

### Note

Quando si utilizza AWS KMS, vengono addebitati costi per l'archiviazione e l'uso delle chiavi di crittografia. Per ulteriori informazioni, consulta [Prezzi di AWS KMS](#).

## Crittografia lato server di Amazon S3

Quando configuri la crittografia lato server Amazon S3, Amazon S3 esegue la crittografia dei dati a livello di oggetto, tramite la scrittura dei dati su disco, e ne esegue la decrittografia al momento dell'accesso. Per ulteriori informazioni su SSE, consulta la sezione [Protezione dei dati utilizzando la crittografia lato server](#) nella Amazon Simple Storage Service Developer Guide.

Puoi scegliere tra due diversi sistemi di gestione delle chiavi quando specifichi SSE in Amazon EMR Serverless:

- SSE-S3: Amazon S3 gestisce le chiavi per te. Non è richiesta alcuna configurazione aggiuntiva su EMR Serverless.
- SSE-KMS - Si utilizza una AWS KMS key configurazione con politiche adatte per EMR Serverless. Non è richiesta alcuna configurazione aggiuntiva su EMR Serverless.

Per utilizzare AWS KMS la crittografia per i dati che scrivi su Amazon S3, hai due opzioni quando usi l'`StartJobRunAPI`. Puoi abilitare la crittografia per tutto ciò che scrivi su Amazon S3 oppure puoi abilitare la crittografia per i dati che scrivi in un bucket specifico. Per ulteriori informazioni sull'`StartJobRunAPI`, vedere l'[EMR Serverless API Reference](#).

Per attivare AWS KMS la crittografia per tutti i dati che scrivi su Amazon S3, usa i seguenti comandi quando chiami l'`StartJobRunAPI`.

```
--conf spark.hadoop.fs.s3.enableServerSideEncryption=true
--conf spark.hadoop.fs.s3.serverSideEncryption.kms.keyId=<kms_id>
```

Per attivare AWS KMS la crittografia dei dati che scrivi in un bucket specifico, usa i seguenti comandi quando chiami l'`StartJobRunAPI`.

```
--conf spark.hadoop.fs.s3.bucket.<amzn-s3-demo-bucket1>.enableServerSideEncryption=true
--conf spark.hadoop.fs.s3.bucket.<amzn-s3-demo-bucket1>.serverSideEncryption.kms.keyId=<kms-id>
```

SSE con chiavi fornite dal cliente (SSE-C) non è disponibile per l'uso con EMR Serverless.

## Crittografia lato client Amazon S3

Con la crittografia lato client di Amazon S3, la crittografia e la decrittografia di Amazon S3 avvengono nel client EMRFS disponibile in ogni versione di Amazon EMR. Gli oggetti vengono crittografati prima di essere caricati su Amazon S3 e decrittati dopo il loro download. Il provider specificato fornisce la chiave di crittografia utilizzata dal client. Il client può utilizzare le chiavi fornite da AWS KMS (CSE-KMS) o una classe Java personalizzata che fornisce la chiave principale lato client (CSE-C). Le specifiche di crittografia sono leggermente diverse tra CSE-KMS e CSE-C, a seconda del provider specificato e dei metadati dell'oggetto da decrittare o crittografare. Se utilizzi Amazon S3 CSE con chiavi gestite dal cliente, il ruolo di esecuzione utilizzato per eseguire lavori in un'applicazione EMR Serverless deve avere accesso alla chiave. Potrebbero essere applicati costi KMS aggiuntivi. Per

ulteriori informazioni su queste differenze, consulta la sezione [Protezione dei dati utilizzando la crittografia lato client](#) nella Amazon Simple Storage Service Developer Guide.

## Crittografia del disco locale

I dati archiviati nello storage temporaneo sono crittografati con chiavi di proprietà del servizio utilizzando l'algoritmo crittografico AES-256 standard del settore.

## Gestione delle chiavi

È possibile configurare KMS per ruotare in automatico le chiavi KMS. Questa impostazione effettua una rotazione delle chiavi una volta all'anno e allo stesso tempo salva le vecchie chiavi a tempo indeterminato, in modo che i dati possano ancora essere decrittati. Per ulteriori informazioni, consulta [Rotazione delle chiavi cliente principali](#).

## Crittografia in transito

Le seguenti funzionalità di crittografia specifiche dell'applicazione sono disponibili con Amazon EMR Serverless:

- Spark
  - Per impostazione predefinita, la comunicazione tra i driver Spark e gli esecutori è autenticata e interna. La comunicazione RPC tra driver ed esecutori è crittografata.
- Hive
  - La comunicazione tra il metastore AWS Glue e le applicazioni EMR Serverless avviene tramite TLS.

Dovresti consentire solo connessioni crittografate su HTTPS (TLS) utilizzando [le policy IAM di aws: SecureTransport condition](#) on Amazon S3 bucket.

## Identity and Access Management (IAM) in Amazon EMR Serverless

AWS Identity and Access Management (IAM) è un software Servizio AWS che aiuta un amministratore a controllare in modo sicuro l'accesso alle risorse. AWS Gli amministratori IAM controllano chi può essere autenticato (effettuato l'accesso) e autorizzato (disporre delle autorizzazioni) a utilizzare le risorse Serverless di Amazon EMR. IAM è un software Servizio AWS che puoi utilizzare senza costi aggiuntivi.

### Argomenti

- [Destinatari](#)
- [Autenticazione con identità](#)
- [Gestione dell'accesso con policy](#)
- [Come funziona EMR Serverless con IAM](#)
- [Utilizzo di ruoli collegati ai servizi per EMR Serverless](#)
- [Ruoli Job Runtime per Amazon EMR Serverless](#)
- [Esempi di policy di accesso degli utenti per EMR Serverless](#)
- [Policy per il controllo degli accessi basato su tag](#)
- [Esempi di policy basate sull'identità per EMR Serverless](#)
- [Amazon EMR Serverless: aggiornamenti alle policy gestite AWS](#)
- [Risoluzione dei problemi relativi all'identità e all'accesso senza server di Amazon EMR](#)

## Destinatari

Il modo in cui utilizzi AWS Identity and Access Management (IAM) varia a seconda del lavoro svolto in Amazon EMR Serverless.

**Utente del servizio:** se utilizzi il servizio Amazon EMR Serverless per svolgere il tuo lavoro, l'amministratore ti fornisce le credenziali e le autorizzazioni necessarie. Man mano che utilizzi più funzionalità Serverless di Amazon EMR per svolgere il tuo lavoro, potresti aver bisogno di autorizzazioni aggiuntive. La comprensione della gestione dell'accesso ti consente di richiedere le autorizzazioni corrette all'amministratore. Se non riesci ad accedere a una funzionalità di Amazon EMR Serverless, consulta [Risoluzione dei problemi relativi all'identità e all'accesso senza server di Amazon EMR](#)

**Amministratore del servizio:** se sei responsabile delle risorse Amazon EMR Serverless della tua azienda, probabilmente hai pieno accesso ad Amazon EMR Serverless. Spetta a te determinare a quali funzionalità e risorse Amazon EMR Serverless devono accedere gli utenti del servizio. Devi inviare le richieste all'amministratore IAM per cambiare le autorizzazioni degli utenti del servizio. Esamina le informazioni contenute in questa pagina per comprendere i concetti di base relativi a IAM. Per ulteriori informazioni su come la tua azienda può utilizzare IAM con Amazon EMR Serverless, consulta [Identity and Access Management \(IAM\) in Amazon EMR Serverless](#)

**Amministratore IAM:** se sei un amministratore IAM, potresti voler saperne di più su come scrivere policy per gestire l'accesso ad Amazon EMR Serverless. Per visualizzare esempi di policy basate

sull'identità di Amazon EMR Serverless che puoi utilizzare in IAM, consulta. [Esempi di policy basate sull'identità per EMR Serverless](#)

## Autenticazione con identità

L'autenticazione è il modo in cui accedi utilizzando le tue credenziali di identità. AWS Devi essere autenticato (aver effettuato l' Utente root dell'account AWS accesso AWS) come utente IAM o assumendo un ruolo IAM.

Puoi accedere AWS come identità federata utilizzando le credenziali fornite tramite una fonte di identità. AWS IAM Identity Center Gli utenti (IAM Identity Center), l'autenticazione Single Sign-On della tua azienda e le tue credenziali di Google o Facebook sono esempi di identità federate. Se accedi come identità federata, l'amministratore ha configurato in precedenza la federazione delle identità utilizzando i ruoli IAM. Quando accedi AWS utilizzando la federazione, assumi indirettamente un ruolo.

A seconda del tipo di utente, puoi accedere al AWS Management Console o al portale di AWS accesso. Per ulteriori informazioni sull'accesso a AWS, vedi [Come accedere al tuo Account AWS nella Guida per l'Accedi ad AWS utente](#).

Se accedi a AWS livello di codice, AWS fornisce un kit di sviluppo software (SDK) e un'interfaccia a riga di comando (CLI) per firmare crittograficamente le tue richieste utilizzando le tue credenziali. Se non utilizzi AWS strumenti, devi firmare tu stesso le richieste. Per ulteriori informazioni sul metodo consigliato per la firma delle richieste, consulta [Signature Version 4 AWS per le richieste API](#) nella Guida per l'utente IAM.

A prescindere dal metodo di autenticazione utilizzato, potrebbe essere necessario specificare ulteriori informazioni sulla sicurezza. Ad esempio, ti AWS consiglia di utilizzare l'autenticazione a più fattori (MFA) per aumentare la sicurezza del tuo account. Per ulteriori informazioni, consulta [Autenticazione a più fattori](#) nella Guida per l'utente di AWS IAM Identity Center e [Utilizzo dell'autenticazione a più fattori \(MFA\)AWS in IAM](#) nella Guida per l'utente IAM.

## Account AWS utente root

Quando si crea un account Account AWS, si inizia con un'identità di accesso che ha accesso completo a tutte Servizi AWS le risorse dell'account. Questa identità è denominata utente Account AWS root ed è accessibile effettuando l'accesso con l'indirizzo e-mail e la password utilizzati per creare l'account. Si consiglia vivamente di non utilizzare l'utente root per le attività quotidiane. Conserva le credenziali dell'utente root e utilizzale per eseguire le operazioni che solo l'utente root

può eseguire. Per un elenco completo delle attività che richiedono l'accesso come utente root, consulta la sezione [Attività che richiedono le credenziali dell'utente root](#) nella Guida per l'utente IAM.

## Identità federata

Come procedura consigliata, richiedi agli utenti umani, compresi gli utenti che richiedono l'accesso come amministratore, di utilizzare la federazione con un provider di identità per accedere Servizi AWS utilizzando credenziali temporanee.

Un'identità federata è un utente dell'elenco utenti aziendale, di un provider di identità Web AWS Directory Service, della directory Identity Center o di qualsiasi utente che accede utilizzando le Servizi AWS credenziali fornite tramite un'origine di identità. Quando le identità federate accedono Account AWS, assumono ruoli e i ruoli forniscono credenziali temporanee.

Per la gestione centralizzata degli accessi, consigliamo di utilizzare AWS IAM Identity Center. Puoi creare utenti e gruppi in IAM Identity Center oppure puoi connetterti e sincronizzarti con un set di utenti e gruppi nella tua fonte di identità per utilizzarli su tutte le tue applicazioni. Account AWS Per ulteriori informazioni su IAM Identity Center, consulta [Cos'è IAM Identity Center?](#) nella Guida per l'utente di AWS IAM Identity Center .

## Utenti e gruppi IAM

Un [utente IAM](#) è un'identità interna Account AWS che dispone di autorizzazioni specifiche per una singola persona o applicazione. Ove possibile, consigliamo di fare affidamento a credenziali temporanee invece di creare utenti IAM con credenziali a lungo termine come le password e le chiavi di accesso. Tuttavia, se si hanno casi d'uso specifici che richiedono credenziali a lungo termine con utenti IAM, si consiglia di ruotare le chiavi di accesso. Per ulteriori informazioni, consulta la pagina [Rotazione periodica delle chiavi di accesso per casi d'uso che richiedono credenziali a lungo termine](#) nella Guida per l'utente IAM.

Un [gruppo IAM](#) è un'identità che specifica un insieme di utenti IAM. Non è possibile eseguire l'accesso come gruppo. È possibile utilizzare gruppi per specificare le autorizzazioni per più utenti alla volta. I gruppi semplificano la gestione delle autorizzazioni per set di utenti di grandi dimensioni. Ad esempio, potresti avere un gruppo denominato IAMAdminse concedere a quel gruppo le autorizzazioni per amministrare le risorse IAM.

Gli utenti sono diversi dai ruoli. Un utente è associato in modo univoco a una persona o un'applicazione, mentre un ruolo è destinato a essere assunto da chiunque ne abbia bisogno. Gli utenti dispongono di credenziali a lungo termine permanenti, mentre i ruoli forniscono credenziali

temporanee. Per ulteriori informazioni, consulta [Casi d'uso per utenti IAM](#) nella Guida per l'utente IAM.

## Ruoli IAM

Un [ruolo IAM](#) è un'identità interna all'utente Account AWS che dispone di autorizzazioni specifiche. È simile a un utente IAM, ma non è associato a una persona specifica. Per assumere temporaneamente un ruolo IAM in AWS Management Console, puoi [passare da un ruolo utente a un ruolo IAM \(console\)](#). Puoi assumere un ruolo chiamando un'operazione AWS CLI o AWS API o utilizzando un URL personalizzato. Per ulteriori informazioni sui metodi per l'utilizzo dei ruoli, consulta [Utilizzo di ruoli IAM](#) nella Guida per l'utente IAM.

I ruoli IAM con credenziali temporanee sono utili nelle seguenti situazioni:

- **Accesso utente federato:** per assegnare le autorizzazioni a una identità federata, è possibile creare un ruolo e definire le autorizzazioni per il ruolo. Quando un'identità federata viene autenticata, l'identità viene associata al ruolo e ottiene le autorizzazioni da esso definite. Per ulteriori informazioni sulla federazione dei ruoli, consulta [Create a role for a third-party identity provider \(federation\)](#) nella Guida per l'utente IAM. Se utilizzi IAM Identity Center, configura un set di autorizzazioni. IAM Identity Center mette in correlazione il set di autorizzazioni con un ruolo in IAM per controllare a cosa possono accedere le identità dopo l'autenticazione. Per informazioni sui set di autorizzazioni, consulta [Set di autorizzazioni](#) nella Guida per l'utente di AWS IAM Identity Center.
- **Autorizzazioni utente IAM temporanee:** un utente IAM o un ruolo può assumere un ruolo IAM per ottenere temporaneamente autorizzazioni diverse per un'attività specifica.
- **Accesso multi-account:** è possibile utilizzare un ruolo IAM per permettere a un utente (un principale affidabile) con un account diverso di accedere alle risorse nell'account. I ruoli sono lo strumento principale per concedere l'accesso multi-account. Tuttavia, con alcuni Servizi AWS, è possibile allegare una policy direttamente a una risorsa (anziché utilizzare un ruolo come proxy). Per informazioni sulle differenze tra ruoli e policy basate su risorse per l'accesso multi-account, consulta [Accesso a risorse multi-account in IAM](#) nella Guida per l'utente IAM.
- **Accesso a più servizi:** alcuni Servizi AWS utilizzano le funzionalità di altri Servizi AWS. Ad esempio, quando effettui una chiamata in un servizio, è normale che quel servizio esegua applicazioni in Amazon EC2 o archivi oggetti in Amazon S3. Un servizio può eseguire questa operazione utilizzando le autorizzazioni dell'entità chiamante, utilizzando un ruolo di servizio o utilizzando un ruolo collegato al servizio.

- **Sessioni di accesso inoltrato (FAS):** quando utilizzi un utente o un ruolo IAM per eseguire azioni AWS, sei considerato un principale. Quando si utilizzano alcuni servizi, è possibile eseguire un'operazione che attiva un'altra operazione in un servizio diverso. FAS utilizza le autorizzazioni del principale che chiama un Servizio AWS, combinate con la richiesta Servizio AWS per effettuare richieste ai servizi downstream. Le richieste FAS vengono effettuate solo quando un servizio riceve una richiesta che richiede interazioni con altri Servizi AWS o risorse per essere completata. In questo caso è necessario disporre delle autorizzazioni per eseguire entrambe le azioni. Per i dettagli delle policy relative alle richieste FAS, consulta [Forward access sessions](#).
- **Ruolo di servizio:** un ruolo di servizio è un [ruolo IAM](#) che un servizio assume per eseguire operazioni per tuo conto. Un amministratore IAM può creare, modificare ed eliminare un ruolo di servizio dall'interno di IAM. Per ulteriori informazioni, consulta la sezione [Create a role to delegate permissions to an Servizio AWS](#) nella Guida per l'utente IAM.
- **Ruolo collegato al servizio:** un ruolo collegato al servizio è un tipo di ruolo di servizio collegato a un Servizio AWS. Il servizio può assumere il ruolo per eseguire un'azione per tuo conto. I ruoli collegati al servizio vengono visualizzati nel tuo account Account AWS e sono di proprietà del servizio. Un amministratore IAM può visualizzare le autorizzazioni per i ruoli collegati ai servizi, ma non modificarle.
- **Applicazioni in esecuzione su Amazon EC2:** puoi utilizzare un ruolo IAM per gestire le credenziali temporanee per le applicazioni in esecuzione su un' EC2 istanza e che AWS CLI effettuano richieste AWS API. È preferibile archiviare le chiavi di accesso all'interno dell' EC2 istanza. Per assegnare un AWS ruolo a un' EC2 istanza e renderlo disponibile per tutte le sue applicazioni, create un profilo di istanza collegato all'istanza. Un profilo di istanza contiene il ruolo e consente ai programmi in esecuzione sull' EC2 istanza di ottenere credenziali temporanee. Per ulteriori informazioni, consulta [Utilizzare un ruolo IAM per concedere le autorizzazioni alle applicazioni in esecuzione su EC2 istanze Amazon](#) nella IAM User Guide.

## Gestione dell'accesso con policy

Puoi controllare l'accesso AWS creando policy e collegandole a AWS identità o risorse. Una policy è un oggetto AWS che, se associato a un'identità o a una risorsa, ne definisce le autorizzazioni. AWS valuta queste politiche quando un principale (utente, utente root o sessione di ruolo) effettua una richiesta. Le autorizzazioni nelle policy determinano l'approvazione o il rifiuto della richiesta. La maggior parte delle politiche viene archiviata AWS come documenti JSON. Per ulteriori informazioni sulla struttura e sui contenuti dei documenti delle policy JSON, consulta [Panoramica delle policy JSON](#) nella Guida per l'utente IAM.

Gli amministratori possono utilizzare le policy AWS JSON per specificare chi ha accesso a cosa. In altre parole, quale principale può eseguire operazioni su quali risorse e in quali condizioni.

Per impostazione predefinita, utenti e ruoli non dispongono di autorizzazioni. Per concedere agli utenti l'autorizzazione a eseguire operazioni sulle risorse di cui hanno bisogno, un amministratore IAM può creare policy IAM. L'amministratore può quindi aggiungere le policy IAM ai ruoli e gli utenti possono assumere i ruoli.

Le policy IAM definiscono le autorizzazioni relative a un'operazione, a prescindere dal metodo utilizzato per eseguirla. Ad esempio, supponiamo di disporre di una policy che consente l'operazione `iam:GetRole`. Un utente con tale policy può ottenere informazioni sul ruolo dall' AWS Management Console AWS CLI, dall'o dall' AWS API.

## Policy basate sull'identità

Le policy basate su identità sono documenti di policy di autorizzazione JSON che è possibile allegare a un'identità (utente, gruppo di utenti o ruolo IAM). Tali policy definiscono le operazioni che utenti e ruoli possono eseguire, su quali risorse e in quali condizioni. Per informazioni su come creare una policy basata su identità, consulta [Definizione di autorizzazioni personalizzate IAM con policy gestite dal cliente](#) nella Guida per l'utente IAM.

Le policy basate su identità possono essere ulteriormente classificate come policy inline o policy gestite. Le policy inline sono integrate direttamente in un singolo utente, gruppo o ruolo. Le politiche gestite sono politiche autonome che puoi allegare a più utenti, gruppi e ruoli nel tuo Account AWS. Le politiche gestite includono politiche AWS gestite e politiche gestite dai clienti. Per informazioni su come scegliere tra una policy gestita o una policy inline, consulta [Scelta fra policy gestite e policy inline](#) nella Guida per l'utente IAM.

## Policy basate sulle risorse

Le policy basate su risorse sono documenti di policy JSON che è possibile collegare a una risorsa. Esempi di policy basate sulle risorse sono le policy di attendibilità dei ruoli IAM e le policy dei bucket Amazon S3. Nei servizi che supportano policy basate sulle risorse, gli amministratori dei servizi possono utilizzarli per controllare l'accesso a una risorsa specifica. Quando è collegata a una risorsa, una policy definisce le operazioni che un principale può eseguire su tale risorsa e a quali condizioni. È necessario [specificare un principale](#) in una policy basata sulle risorse. I principali possono includere account, utenti, ruoli, utenti federati o. Servizi AWS

Le policy basate sulle risorse sono policy inline che si trovano in tale servizio. Non puoi utilizzare le policy AWS gestite di IAM in una policy basata sulle risorse.

## Liste di controllo degli accessi (ACLs)

Le liste di controllo degli accessi (ACLs) controllano quali principali (membri dell'account, utenti o ruoli) dispongono delle autorizzazioni per accedere a una risorsa. ACLs sono simili alle politiche basate sulle risorse, sebbene non utilizzino il formato del documento di policy JSON.

Amazon S3 e Amazon VPC sono esempi di servizi che supportano. AWS WAF ACLs Per ulteriori informazioni ACLs, consulta la [panoramica della lista di controllo degli accessi \(ACL\)](#) nella Amazon Simple Storage Service Developer Guide.

## Altri tipi di policy

AWS supporta tipi di policy aggiuntivi e meno comuni. Questi tipi di policy possono impostare il numero massimo di autorizzazioni concesse dai tipi di policy più comuni.

- **Limiti delle autorizzazioni:** un limite delle autorizzazioni è una funzionalità avanzata nella quale si imposta il numero massimo di autorizzazioni che una policy basata su identità può concedere a un'entità IAM (utente o ruolo IAM). È possibile impostare un limite delle autorizzazioni per un'entità. Le autorizzazioni risultanti sono l'intersezione delle policy basate su identità dell'entità e i relativi limiti delle autorizzazioni. Le policy basate su risorse che specificano l'utente o il ruolo nel campo `Principal` sono condizionate dal limite delle autorizzazioni. Un rifiuto esplicito in una qualsiasi di queste policy sostituisce l'autorizzazione. Per ulteriori informazioni sui limiti delle autorizzazioni, consulta [Limiti delle autorizzazioni per le entità IAM](#) nella Guida per l'utente IAM.
- **Politiche di controllo del servizio (SCPs):** SCPs sono politiche JSON che specificano le autorizzazioni massime per un'organizzazione o un'unità organizzativa (OU) in. AWS Organizations AWS Organizations è un servizio per il raggruppamento e la gestione centralizzata di più di proprietà dell' Account AWS azienda. Se abiliti tutte le funzionalità di un'organizzazione, puoi applicare le politiche di controllo del servizio (SCPs) a uno o tutti i tuoi account. L'SCP limita le autorizzazioni per le entità presenti negli account dei membri, inclusa ciascuna di esse. Utente root dell'account AWS Per ulteriori informazioni su Organizations and SCPs, consulta [le politiche di controllo dei servizi](#) nella Guida AWS Organizations per l'utente.
- **Politiche di controllo delle risorse (RCPs):** RCPs sono politiche JSON che puoi utilizzare per impostare le autorizzazioni massime disponibili per le risorse nei tuoi account senza aggiornare le politiche IAM allegate a ciascuna risorsa di tua proprietà. L'RCP limita le autorizzazioni per le risorse negli account dei membri e può influire sulle autorizzazioni effettive per le identità, incluse le Utente root dell'account AWS, indipendentemente dal fatto che appartengano o meno all'organizzazione. Per ulteriori informazioni su Organizations e RCPs, incluso un elenco di

Servizi AWS tale supporto RCPs, vedere [Resource control policies \(RCPs\)](#) nella Guida per l'AWS Organizations utente.

- Policy di sessione: le policy di sessione sono policy avanzate che vengono trasmesse come parametro quando si crea in modo programmatico una sessione temporanea per un ruolo o un utente federato. Le autorizzazioni della sessione risultante sono l'intersezione delle policy basate su identità del ruolo o dell'utente e le policy di sessione. Le autorizzazioni possono anche provenire da una policy basata su risorse. Un rifiuto esplicito in una qualsiasi di queste policy sostituisce l'autorizzazione. Per ulteriori informazioni, consulta [Policy di sessione](#) nella Guida per l'utente IAM.

## Più tipi di policy

Quando più tipi di policy si applicano a una richiesta, le autorizzazioni risultanti sono più complicate da comprendere. Per scoprire come si AWS determina se consentire o meno una richiesta quando sono coinvolti più tipi di policy, consulta la [logica di valutazione delle policy](#) nella IAM User Guide.

## Come funziona EMR Serverless con IAM

Prima di utilizzare IAM per gestire l'accesso ad Amazon EMR Serverless, scopri quali funzionalità IAM sono disponibili per l'uso con Amazon EMR Serverless.

Funzionalità IAM che puoi utilizzare con EMR Serverless

Funzionalità IAM	Supporto Serverless per Amazon EMR
<a href="#">Policy basate su identità</a>	Sì
<a href="#">Policy basate su risorse</a>	No
<a href="#">Azioni di policy</a>	Sì
<a href="#">Risorse relative alle policy</a>	Sì
<a href="#">Chiavi di condizione delle policy</a>	No
<a href="#">ACLs</a>	No
<a href="#">ABAC (tag nelle policy)</a>	Sì
<a href="#">Credenziali temporanee</a>	Sì

Funzionalità IAM	Supporto Serverless per Amazon EMR
<a href="#">Autorizzazioni del principale</a>	Sì
<a href="#">Ruoli di servizio</a>	No
<a href="#">Ruoli collegati al servizio</a>	Sì

Per avere una visione di alto livello di come EMR Serverless e AWS altri servizi funzionano con la maggior parte delle funzionalità IAM, [AWS consulta i servizi che funzionano con IAM](#) nella IAM User Guide.

## Policy basate sull'identità per EMR Serverless

Supporta le policy basate su identità: sì

Le policy basate su identità sono documenti di policy di autorizzazione JSON che è possibile allegare a un'identità (utente, gruppo di utenti o ruolo IAM). Tali policy definiscono le operazioni che utenti e ruoli possono eseguire, su quali risorse e in quali condizioni. Per informazioni su come creare una policy basata su identità, consulta [Definizione di autorizzazioni personalizzate IAM con policy gestite dal cliente](#) nella Guida per l'utente IAM.

Con le policy basate su identità di IAM, è possibile specificare quali operazioni e risorse sono consentite o respinte, nonché le condizioni in base alle quali le operazioni sono consentite o respinte. Non è possibile specificare l'entità principale in una policy basata sull'identità perché si applica all'utente o al ruolo a cui è associato. Per informazioni su tutti gli elementi utilizzabili in una policy JSON, consulta [Guida di riferimento agli elementi delle policy JSON IAM](#) nella Guida per l'utente di IAM.

### Esempi di policy basate sull'identità per EMR Serverless

Per visualizzare esempi di policy basate sull'identità di Amazon EMR Serverless, consulta. [Esempi di policy basate sull'identità per EMR Serverless](#)

## Policy basate sulle risorse all'interno di EMR Serverless

Supporta le policy basate su risorse: no

Le policy basate su risorse sono documenti di policy JSON che è possibile collegare a una risorsa. Esempi di policy basate sulle risorse sono le policy di attendibilità dei ruoli IAM e le policy dei bucket

Amazon S3. Nei servizi che supportano policy basate sulle risorse, gli amministratori dei servizi possono utilizzarli per controllare l'accesso a una risorsa specifica. Quando è collegata a una risorsa, una policy definisce le operazioni che un principale può eseguire su tale risorsa e a quali condizioni. È necessario [specificare un principale](#) in una policy basata sulle risorse. I principali possono includere account, utenti, ruoli, utenti federati o. Servizi AWS

Per consentire l'accesso multi-account, puoi specificare un intero account o entità IAM in un altro account come principale in una policy basata sulle risorse. L'aggiunta di un principale multi-account a una policy basata sulle risorse rappresenta solo una parte della relazione di trust. Quando il principale e la risorsa sono diversi Account AWS, un amministratore IAM dell'account affidabile deve inoltre concedere all'entità principale (utente o ruolo) l'autorizzazione ad accedere alla risorsa. L'autorizzazione viene concessa collegando all'entità una policy basata sull'identità. Tuttavia, se una policy basata su risorse concede l'accesso a un principale nello stesso account, non sono richieste ulteriori policy basate su identità. Per ulteriori informazioni, consulta [Accesso a risorse multi-account in IAM](#) nella Guida per l'utente IAM.

## Azioni politiche per EMR Serverless

Supporta le operazioni di policy: si

Gli amministratori possono utilizzare le policy AWS JSON per specificare chi ha accesso a cosa. In altre parole, quale principale può eseguire operazioni su quali risorse, e in quali condizioni.

L'elemento `Actions` di una policy JSON descrive le operazioni che è possibile utilizzare per consentire o negare l'accesso a un criterio. Le azioni politiche in genere hanno lo stesso nome dell'operazione AWS API associata. Ci sono alcune eccezioni, ad esempio le operazioni di sola autorizzazione che non hanno un'operazione API corrispondente. Esistono anche alcune operazioni che richiedono più operazioni in una policy. Queste operazioni aggiuntive sono denominate operazioni dipendenti.

Includi le operazioni in una policy per concedere le autorizzazioni a eseguire l'operazione associata.

Per visualizzare un elenco di azioni EMR Serverless, consulta [Azioni, risorse e chiavi di condizione per Amazon EMR Serverless](#) nel Service Authorization Reference.

Le azioni politiche in EMR Serverless utilizzano il seguente prefisso prima dell'azione.

```
emr-serverless
```

Per specificare più operazioni in una sola istruzione, occorre separarle con la virgola.

```
"Action": [  
  "emr-serverless:action1",  
  "emr-serverless:action2"  
]
```

Per visualizzare esempi di policy basate sull'identità di Amazon EMR Serverless, consulta. [Esempi di policy basate sull'identità per EMR Serverless](#)

## Risorse relative alle policy per EMR Serverless

Supporta le risorse di policy: sì

Gli amministratori possono utilizzare le policy AWS JSON per specificare chi ha accesso a cosa. In altre parole, quale principale può eseguire operazioni su quali risorse, e in quali condizioni.

L'elemento JSON `Resource` della policy specifica l'oggetto o gli oggetti ai quali si applica l'operazione. Le istruzioni devono includere un elemento `Resource` o un elemento `NotResource`. Come best practice, specifica una risorsa utilizzando il suo [nome della risorsa Amazon \(ARN\)](#). È possibile eseguire questa operazione per operazioni che supportano un tipo di risorsa specifico, note come autorizzazioni a livello di risorsa.

Per le azioni che non supportano le autorizzazioni a livello di risorsa, ad esempio le operazioni di elenco, utilizza un carattere jolly (\*) per indicare che l'istruzione si applica a tutte le risorse.

```
"Resource": "*"
```

Per visualizzare un elenco dei tipi di risorse Serverless di Amazon EMR e relativi ARNs, consulta [Resources defined by Amazon EMR Serverless](#) nel Service Authorization Reference. Per sapere quali azioni è possibile specificare l'ARN di ogni risorsa, consulta [Azioni, risorse e chiavi di condizione per Amazon EMR Serverless](#).

Per visualizzare esempi di policy basate sull'identità di Amazon EMR Serverless, consulta. [Esempi di policy basate sull'identità per EMR Serverless](#)

## Chiavi delle condizioni delle policy per EMR Serverless

Supporta le chiavi di condizione delle policy specifiche del servizio No

Gli amministratori possono utilizzare le policy AWS JSON per specificare chi ha accesso a cosa. In altre parole, quale principale può eseguire operazioni su quali risorse, e in quali condizioni.

L'elemento `Condition`(o blocco `Condition`) consente di specificare le condizioni in cui un'istruzione è in vigore. L'elemento `Condition` è facoltativo. È possibile compilare espressioni condizionali che utilizzano [operatori di condizione](#), ad esempio uguale a o minore di, per soddisfare la condizione nella policy con i valori nella richiesta.

Se specifichi più elementi `Condition` in un'istruzione o più chiavi in un singolo elemento `Condition`, questi vengono valutati da AWS utilizzando un'operazione AND logica. Se si specificano più valori per una singola chiave di condizione, AWS valuta la condizione utilizzando un'operazione logica. OR Tutte le condizioni devono essere soddisfatte prima che le autorizzazioni dell'istruzione vengano concesse.

È possibile anche utilizzare variabili segnaposto quando specifichi le condizioni. Ad esempio, è possibile autorizzare un utente IAM ad accedere a una risorsa solo se è stata taggata con il relativo nome utente IAM. Per ulteriori informazioni, consulta [Elementi delle policy IAM: variabili e tag](#) nella Guida per l'utente di IAM.

AWS supporta chiavi di condizione globali e chiavi di condizione specifiche del servizio. Per visualizzare tutte le chiavi di condizione AWS globali, consulta le chiavi di [contesto delle condizioni AWS globali nella Guida](#) per l'utente IAM.

Per visualizzare un elenco di chiavi di condizione di Amazon EMR Serverless e per scoprire quali azioni e risorse è possibile utilizzare una chiave di condizione, consulta [Azioni, risorse e chiavi di condizione per Amazon EMR Serverless](#) nel Service Authorization Reference.

Tutte le EC2 azioni di Amazon supportano le chiavi `aws:RequestedRegion` e `ec2:Region` condition. Per ulteriori informazioni, consulta [Esempio: limitazione dell'accesso a una regione specifica](#).

## Elenchi di controllo degli accessi (ACLs) in EMR Serverless

Supporti ACLs: no

Le liste di controllo degli accessi (ACLs) controllano quali principali (membri dell'account, utenti o ruoli) dispongono delle autorizzazioni per accedere a una risorsa. ACLs sono simili alle politiche basate sulle risorse, sebbene non utilizzino il formato del documento di policy JSON.

## Controllo degli accessi basato sugli attributi (ABAC) con EMR Serverless

Supporta ABAC (tag nelle policy)	Sì
----------------------------------	----

Il controllo dell'accesso basato su attributi (ABAC) è una strategia di autorizzazione che definisce le autorizzazioni in base agli attributi. In, questi attributi sono chiamati AWS tag. Puoi allegare tag a entità IAM (utenti o ruoli) e a molte AWS risorse. L'assegnazione di tag alle entità e alle risorse è il primo passaggio di ABAC. In seguito, vengono progettate policy ABAC per consentire operazioni quando il tag dell'entità principale corrisponde al tag sulla risorsa a cui si sta provando ad accedere.

La strategia ABAC è utile in ambienti soggetti a una rapida crescita e aiuta in situazioni in cui la gestione delle policy diventa impegnativa.

Per controllare l'accesso basato su tag, fornisci informazioni sui tag nell'[elemento condizione](#) di una policy utilizzando le chiavi di condizione `aws:ResourceTag/key-name`, `aws:RequestTag/key-name` o `aws:TagKeys`.

Se un servizio supporta tutte e tre le chiavi di condizione per ogni tipo di risorsa, il valore per il servizio è Yes (Sì). Se un servizio supporta tutte e tre le chiavi di condizione solo per alcuni tipi di risorsa, allora il valore sarà Parziale.

Per ulteriori informazioni su ABAC, consulta [Definizione delle autorizzazioni con autorizzazione ABAC](#) nella Guida per l'utente IAM. Per visualizzare un tutorial con i passaggi per l'impostazione di ABAC, consulta [Utilizzo del controllo degli accessi basato su attributi \(ABAC\)](#) nella Guida per l'utente di IAM.

## Utilizzo di credenziali temporanee con EMR Serverless

Supporta le credenziali temporanee: sì

Alcune Servizi AWS non funzionano quando si accede utilizzando credenziali temporanee. Per ulteriori informazioni, incluse quelle che Servizi AWS funzionano con credenziali temporanee, consulta la sezione relativa alla [Servizi AWS compatibilità con IAM nella IAM User Guide](#).

Stai utilizzando credenziali temporanee se accedi AWS Management Console utilizzando qualsiasi metodo tranne nome utente e password. Ad esempio, quando accedete AWS utilizzando il link Single

Sign-On (SSO) della vostra azienda, tale processo crea automaticamente credenziali temporanee. Le credenziali temporanee vengono create in automatico anche quando accedi alla console come utente e poi cambi ruolo. Per ulteriori informazioni sullo scambio dei ruoli, consulta [Passaggio da un ruolo utente a un ruolo IAM \(console\)](#) nella Guida per l'utente IAM.

È possibile creare manualmente credenziali temporanee utilizzando l'API or. AWS CLI AWS È quindi possibile utilizzare tali credenziali temporanee per accedere. AWS AWS consiglia di generare dinamicamente credenziali temporanee anziché utilizzare chiavi di accesso a lungo termine. Per ulteriori informazioni, consulta [Credenziali di sicurezza provvisorie in IAM](#).

## Autorizzazioni principali multiservizio per EMR Serverless

Supporta l'inoltro delle sessioni di accesso (FAS): sì

Quando utilizzi un utente o un ruolo IAM per eseguire azioni AWS, sei considerato un principale. Quando si utilizzano alcuni servizi, è possibile eseguire un'operazione che attiva un'altra operazione in un servizio diverso. FAS utilizza le autorizzazioni del principale che chiama un Servizio AWS, in combinazione con la richiesta Servizio AWS per effettuare richieste ai servizi downstream. Le richieste FAS vengono effettuate solo quando un servizio riceve una richiesta che richiede interazioni con altri Servizi AWS o risorse per essere completata. In questo caso è necessario disporre delle autorizzazioni per eseguire entrambe le azioni. Per i dettagli delle policy relative alle richieste FAS, consulta [Forward access sessions](#).

## Ruoli di servizio per EMR Serverless

Supporta i ruoli di servizio	No
------------------------------	----

## Ruoli collegati ai servizi per EMR Serverless

Supporta i ruoli collegati ai servizi	Sì
---------------------------------------	----

Per ulteriori informazioni su come creare e gestire i ruoli collegati ai servizi, consulta [Servizi AWS supportati da IAM](#). Trova un servizio nella tabella che include un Yes nella colonna Service-linked role (Ruolo collegato ai servizi). Scegli il collegamento Sì per visualizzare la documentazione relativa al ruolo collegato ai servizi per tale servizio.

## Utilizzo di ruoli collegati ai servizi per EMR Serverless

[Amazon EMR Serverless utilizza ruoli collegati ai servizi AWS Identity and Access Management \(IAM\)](#). Un ruolo collegato ai servizi è un tipo unico di ruolo IAM collegato direttamente a EMR Serverless. I ruoli collegati ai servizi sono predefiniti da EMR Serverless e includono tutte le autorizzazioni richieste dal servizio per chiamare altri servizi per conto dell'utente. AWS

Un ruolo collegato al servizio semplifica la configurazione di EMR Serverless perché non è necessario aggiungere manualmente le autorizzazioni necessarie. EMR Serverless definisce le autorizzazioni dei suoi ruoli collegati ai servizi e, se non diversamente definito, solo EMR Serverless può assumerne i ruoli. Le autorizzazioni definite includono la policy di attendibilità e la policy delle autorizzazioni che non può essere allegata a nessun'altra entità IAM.

È possibile eliminare un ruolo collegato ai servizi solo dopo aver eliminato le risorse correlate. Ciò protegge le risorse EMR Serverless perché non è possibile rimuovere inavvertitamente l'autorizzazione ad accedere alle risorse.

Per informazioni su altri servizi che supportano i ruoli collegati ai servizi, consulta [AWS Servizi compatibili con IAM e cerca i servizi con](#) Sì nella colonna Ruoli collegati ai servizi. Scegli Sì in corrispondenza di un link per visualizzare la documentazione relativa al ruolo collegato al servizio per tale servizio.

### Autorizzazioni di ruolo collegate ai servizi per EMR Serverless

EMR Serverless utilizza il ruolo collegato al servizio denominato `AWSServiceRoleForAmazonEMRServerless` per consentirgli di effettuare chiamate per conto dell'utente. AWS APIs

Il ruolo `AWSServiceRoleForAmazonEMRServerless` collegato al servizio prevede che i seguenti servizi assumano il ruolo:

- `ops.emr-serverless.amazonaws.com`

La politica di autorizzazione dei ruoli denominata `AmazonEMRServerlessServiceRolePolicy` consente a EMR Serverless di completare le seguenti azioni sulle risorse specificate.

**Note**

Il contenuto della policy gestita cambia, pertanto la politica mostrata qui potrebbe non essere aggiornata. Visualizza la maggior parte delle up-to-date politiche di [Amazon EMRServerless ServiceRolePolicy](#) nel AWS Management Console.

- Operazione: `ec2:CreateNetworkInterface`
- Operazione: `ec2>DeleteNetworkInterface`
- Operazione: `ec2:DescribeNetworkInterfaces`
- Operazione: `ec2:DescribeSecurityGroups`
- Operazione: `ec2:DescribeSubnets`
- Operazione: `ec2:DescribeVpcs`
- Operazione: `ec2:DescribeDhcpOptions`
- Operazione: `ec2:DescribeRouteTables`
- Operazione: `cloudwatch:PutMetricData`

Di seguito è riportata la `AmazonEMRServerlessServiceRolePolicy` politica completa.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "EC2PolicyStatement",
      "Effect": "Allow",
      "Action": [
        "ec2:CreateNetworkInterface",
        "ec2>DeleteNetworkInterface",
        "ec2:DescribeNetworkInterfaces",
        "ec2:DescribeSecurityGroups",
        "ec2:DescribeSubnets",
        "ec2:DescribeVpcs",
        "ec2:DescribeDhcpOptions",
        "ec2:DescribeRouteTables"
      ],
      "Resource": "*"
    }
  ],
}
```

```

    {
      "Sid": "CloudWatchPolicyStatement",
      "Effect": "Allow",
      "Action": [
        "cloudwatch:PutMetricData"
      ],
      "Resource": [
        "*"
      ],
      "Condition": {
        "StringEquals": {
          "cloudwatch:namespace": [
            "AWS/EMRServerless",
            "AWS/Usage"
          ]
        }
      }
    }
  ]
}

```

La seguente politica di fiducia è associata a questo ruolo per consentire al principale EMR Serverless di assumere questo ruolo.

```

{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Principal": {
        "Service": [
          "ops.emr-serverless.amazonaws.com"
        ]
      },
      "Action": "sts:AssumeRole"
    }
  ]
}

```

Per consentire a un'entità IAM (come un utente, un gruppo o un ruolo) di creare, modificare o eliminare un ruolo collegato ai servizi devi configurare le relative autorizzazioni. Per ulteriori informazioni, consulta [Autorizzazioni del ruolo collegato ai servizi](#) nella Guida per l'utente di IAM.

## Creazione di un ruolo collegato ai servizi per EMR Serverless

Non hai bisogno di creare manualmente un ruolo collegato ai servizi. Quando si crea una nuova applicazione EMR Serverless (AWS Management Console utilizzando EMR Studio), o l'AWS CLI API AWS, EMR Serverless crea automaticamente il ruolo collegato al servizio. Per consentire a un'entità IAM (come un utente, un gruppo o un ruolo) di creare, modificare o eliminare un ruolo collegato ai servizi devi configurare le relative autorizzazioni.

Per creare il ruolo collegato al servizio utilizzando IAM AWSService RoleForAmazon EMRServerless

Aggiungi la seguente dichiarazione alla politica di autorizzazione per l'entità IAM che deve creare il ruolo collegato al servizio.

```
{
  "Effect": "Allow",
  "Action": [
    "iam:CreateServiceLinkedRole"
  ],
  "Resource": "arn:aws:iam::*:role/aws-service-role/ops.emr-serverless.amazonaws.com/AWSServiceRoleForAmazonEMRServerless*",
  "Condition": {"StringLike": {"iam:AWSServiceName": "ops.emr-serverless.amazonaws.com"}}
}
```

Se elimini questo ruolo collegato ai servizi, è possibile ricrearlo seguendo lo stesso processo utilizzato per ricreare il ruolo nell'account. Quando crei una nuova applicazione EMR Serverless, EMR Serverless crea nuovamente il ruolo collegato al servizio per te.

Puoi anche utilizzare la console IAM per creare un ruolo collegato ai servizi con lo use case EMR Serverless. Nella AWS CLI o nell'AWS API, crea un ruolo collegato al servizio con il nome del servizio. `ops.emr-serverless.amazonaws.com` Per ulteriori informazioni, consulta [Creazione di un ruolo collegato ai servizi](#) nella Guida per l'utente di IAM. Se elimini il ruolo collegato ai servizi, è possibile utilizzare lo stesso processo per crearlo nuovamente.

## Modifica di un ruolo collegato ai servizi per EMR Serverless

EMR Serverless non consente di modificare il ruolo AWSService RoleForAmazon EMRServerless collegato al servizio perché diverse entità potrebbero fare riferimento al ruolo. Non è possibile modificare la policy IAM AWS di proprietà utilizzata dal ruolo collegato al servizio EMR Serverless, poiché contiene tutte le autorizzazioni necessarie di EMR Serverless. È possibile tuttavia modificarne la descrizione utilizzando IAM.

## Per modificare la descrizione del ruolo collegato al servizio utilizzando IAM AWSService RoleForAmazon EMRServerless

Aggiungi la seguente istruzione alla policy delle autorizzazioni per l'entità IAM che deve modificare la descrizione di un ruolo collegato ai servizi.

```
{
  "Effect": "Allow",
  "Action": [
    "iam: UpdateRoleDescription"
  ],
  "Resource": "arn:aws:iam::*:role/aws-service-role/ops.emr-serverless.amazonaws.com/AWSServiceRoleForAmazonEMRServerless*",
  "Condition": {"StringLike": {"iam:AWSServiceName": "ops.emr-serverless.amazonaws.com"}}
}
```

Per ulteriori informazioni, consulta [Modifica di un ruolo collegato ai servizi](#) nella Guida per l'utente di IAM.

## Eliminazione di un ruolo collegato al servizio per EMR Serverless

Se non è più necessario utilizzare una funzionalità o un servizio che richiede un ruolo collegato al servizio, ti consigliamo di eliminare il ruolo. In questo modo non si ha un'entità inutilizzata che non viene monitorata o gestita attivamente. Tuttavia, è necessario eliminare tutte le applicazioni EMR Serverless in tutte le regioni prima di poter eliminare il ruolo collegato al servizio.

### Note

Se il servizio EMR Serverless utilizza il ruolo quando si tenta di eliminare le risorse associate al ruolo, l'eliminazione potrebbe non riuscire. In questo caso, attendi alcuni minuti e quindi ripeti l'operazione.

Per eliminare il ruolo collegato al AWSService RoleForAmazon EMRServerless servizio utilizzando IAM

Aggiungi la seguente dichiarazione alla politica di autorizzazione per l'entità IAM che deve eliminare un ruolo collegato al servizio.

```
{
  "Effect": "Allow",
  "Action": [
    "iam:DeleteServiceLinkedRole",
    "iam:GetServiceLinkedRoleDeletionStatus"
  ],
  "Resource": "arn:aws:iam::*:role/aws-service-role/ops.emr-serverless.amazonaws.com/AWSServiceRoleForAmazonEMRServerless*",
  "Condition": {"StringLike": {"iam:AWSServiceName": "ops.emr-serverless.amazonaws.com"}}
}
```

Per eliminare manualmente il ruolo collegato ai servizi mediante IAM

Utilizza la console IAM AWS CLI, l'o l' AWS API per eliminare il ruolo collegato al AWSService RoleForAmazon EMRServerless servizio. Per ulteriori informazioni, consulta [Eliminazione del ruolo collegato al servizio](#) nella Guida per l'utente di IAM.

## Regioni supportate per i ruoli collegati ai servizi EMR Serverless

EMR Serverless supporta l'utilizzo di ruoli collegati ai servizi in tutte le regioni in cui il servizio è disponibile. Per ulteriori informazioni, consulta [AWS Regioni ed endpoint](#).

## Ruoli Job Runtime per Amazon EMR Serverless

È possibile specificare le autorizzazioni dei ruoli IAM che l'esecuzione di un job EMR Serverless può assumere quando si chiamano altri servizi per conto dell'utente. Ciò include l'accesso ad Amazon S3 per qualsiasi fonte di dati, destinazione e altre AWS risorse come i cluster Amazon Redshift e le tabelle DynamoDB. Per ulteriori informazioni su come creare un ruolo, consulta. [Creare un ruolo Job Runtime](#)

### Esempi di politiche di runtime

È possibile allegare una policy di runtime, come la seguente, a un ruolo di job runtime. La seguente politica di esecuzione dei processi consente di:

- Accesso in lettura ai bucket Amazon S3 con esempi EMR.
- Accesso completo ai bucket S3.
- Crea e leggi l'accesso a AWS Glue Data Catalog.

Per aggiungere l'accesso ad altre AWS risorse come DynamoDB, dovrai includere le relative autorizzazioni nella policy al momento della creazione del ruolo di runtime.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "ReadAccessForEMRSamples",
      "Effect": "Allow",
      "Action": [
        "s3:GetObject",
        "s3:ListBucket"
      ],
      "Resource": [
        "arn:aws:s3::*.elasticmapreduce",
        "arn:aws:s3::*.elasticmapreduce/*"
      ]
    },
    {
      "Sid": "FullAccessToS3Bucket",
      "Effect": "Allow",
      "Action": [
        "s3:PutObject",
        "s3:GetObject",
        "s3:ListBucket",
        "s3:DeleteObject"
      ],
      "Resource": [
        "arn:aws:s3:::amzn-s3-demo-bucket",
        "arn:aws:s3:::amzn-s3-demo-bucket/*"
      ]
    },
    {
      "Sid": "GlueCreateAndReadDataCatalog",
      "Effect": "Allow",
      "Action": [
        "glue:GetDatabase",
        "glue:CreateDatabase",
        "glue:GetDataBases",
        "glue:CreateTable",
        "glue:GetTable",
        "glue:UpdateTable",
        "glue:DeleteTable",

```

```

        "glue:GetTables",
        "glue:GetPartition",
        "glue:GetPartitions",
        "glue:CreatePartition",
        "glue:BatchCreatePartition",
        "glue:GetUserDefinedFunctions"
    ],
    "Resource": ["*"]
}
]
}

```

## Passa i privilegi di ruolo

Puoi allegare le policy di autorizzazione IAM al ruolo di un utente per consentire all'utente di passare solo ruoli approvati. Ciò consente agli amministratori di controllare quali utenti possono passare ruoli di job runtime specifici ai job EMR Serverless. Per ulteriori informazioni sull'impostazione delle autorizzazioni, vedere [Concessione a un utente delle autorizzazioni per passare un ruolo a un servizio. AWS](#)

Di seguito è riportato un esempio di policy che consente di passare un ruolo di job runtime al principale del servizio EMR Serverless.

```

{
  "Effect": "Allow",
  "Action": "iam:PassRole",
  "Resource": "arn:aws:iam::1234567890:role/JobRuntimeRoleForEMRServerless",
  "Condition": {
    "StringLike": {
      "iam:PassedToService": "emr-serverless.amazonaws.com"
    }
  }
}

```

## Politiche di autorizzazione gestite associate ai ruoli di runtime

Quando invii le esecuzioni di job a EMR serverless tramite la console EMR Studio, c'è un passaggio in cui scegli un ruolo Runtime da associare alla tua applicazione. A ciascuna selezione nella console sono associate politiche gestite sottostanti di cui è importante tenere presente. Le tre selezioni sono le seguenti:

1. Tutti i bucket: quando scegli questa opzione, specifica la politica FullAccess AWS gestita di [AmazonS3](#), che fornisce l'accesso completo a tutti i bucket.
2. Bucket specifici: specifica l'identificatore Amazon Resource Name (ARN) di ogni bucket scelto. Non è inclusa una politica gestita sottostante.
3. Nessuna: non sono incluse le autorizzazioni relative alle policy gestite.

Ti consigliamo di aggiungere bucket specifici. Se scegli tutti i bucket, tieni presente che imposta l'accesso completo per tutti i bucket.

## Esempi di policy di accesso degli utenti per EMR Serverless

È possibile impostare policy granulari per gli utenti in base alle azioni che si desidera che ciascun utente esegua quando interagisce con le applicazioni EMR Serverless. Le seguenti politiche sono esempi che potrebbero aiutare a configurare le autorizzazioni corrette per gli utenti. Questa sezione si concentra solo sulle politiche EMR Serverless. Per esempi di politiche utente di EMR Studio, vedere Configurazione delle autorizzazioni utente di [EMR Studio](#). Per informazioni su come allegare le policy agli utenti IAM (principali), consulta [Managing IAM policy nella IAM User Guide](#).

### Politica per i Power User

Per concedere tutte le azioni richieste per EMR Serverless, crea e allega una `AmazonEMRServerlessFullAccess` policy all'utente, ruolo o gruppo IAM richiesto.

Di seguito è riportato un esempio di policy che consente agli utenti esperti di creare e modificare applicazioni EMR Serverless, nonché di eseguire altre azioni come l'invio e il debug di lavori. Rivela tutte le azioni che EMR Serverless richiede per altri servizi.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "EMRServerlessActions",
      "Effect": "Allow",
      "Action": [
        "emr-serverless:CreateApplication",
        "emr-serverless:UpdateApplication",
        "emr-serverless>DeleteApplication",
        "emr-serverless:ListApplications",
        "emr-serverless:GetApplication",
        "emr-serverless:StartApplication",
```

```

        "emr-serverless:StopApplication",
        "emr-serverless:StartJobRun",
        "emr-serverless:CancelJobRun",
        "emr-serverless:ListJobRuns",
        "emr-serverless:GetJobRun"
    ],
    "Resource": "*"
}
]
}

```

Quando abiliti la connettività di rete al tuo VPC, le applicazioni EMR Serverless creano interfacce di rete elastiche EC2 Amazon ENIs () per comunicare con le risorse VPC. La seguente politica garantisce che tutti EC2 ENIs i nuovi vengano creati solo nel contesto delle applicazioni EMR Serverless.

#### Note

Consigliamo vivamente di impostare questo criterio per garantire che gli utenti non possano creare EC2 ENIs se non nel contesto del lancio di applicazioni EMR Serverless.

```

{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "AllowEC2ENICreationWithEMRTags",
      "Effect": "Allow",
      "Action": [
        "ec2:CreateNetworkInterface"
      ],
      "Resource": [
        "arn:aws:ec2:*:*:network-interface/*"
      ],
      "Condition": {
        "StringEquals": {
          "aws:CalledViaLast": "ops.emr-serverless.amazonaws.com"
        }
      }
    }
  ]
}

```

Se desideri limitare l'accesso EMR Serverless a determinate sottoreti, puoi etichettare ogni sottorete con una condizione di tag. Questa policy IAM garantisce che le applicazioni EMR Serverless possano creare solo EC2 ENIs all'interno di sottoreti consentite.

```
{
  "Sid": "AllowEC2ENICreationInSubnetAndSecurityGroupWithEMRTags",
  "Effect": "Allow",
  "Action": [
    "ec2:CreateNetworkInterface"
  ],
  "Resource": [
    "arn:aws:ec2:*:*:subnet/*",
    "arn:aws:ec2:*:*:security-group/*"
  ],
  "Condition": {
    "StringEquals": {
      "aws:ResourceTag/KEY": "VALUE"
    }
  }
}
```

### Important

Se sei un amministratore o un utente esperto che crea la tua prima applicazione, devi configurare le tue politiche di autorizzazione per consentirti di creare un ruolo collegato al servizio EMR Serverless. Per ulteriori informazioni, consulta [Utilizzo di ruoli collegati ai servizi per EMR Serverless](#).

La seguente policy IAM ti consente di creare un ruolo collegato al servizio EMR Serverless per il tuo account.

```
{
  "Sid": "AllowEMRServerlessServiceLinkedRoleCreation",
  "Effect": "Allow",
  "Action": "iam:CreateServiceLinkedRole",
  "Resource": "arn:aws:iam::*:account-id:role/aws-service-role/ops.emr-serverless.amazonaws.com/AWSServiceRoleForAmazonEMRServerless"
}
```

## Politica del tecnico dei dati

Di seguito è riportato un esempio di policy che consente agli utenti le autorizzazioni di sola lettura sulle applicazioni EMR Serverless, nonché la possibilità di inviare ed eseguire il debug dei lavori. Tieni presente che poiché questa policy non nega esplicitamente le operazioni, un'altra dichiarazione di policy può essere utilizzata per concedere l'accesso alle operazioni specificate.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "EMRServerlessActions",
      "Effect": "Allow",
      "Action": [
        "emr-serverless:ListApplications",
        "emr-serverless:GetApplication",
        "emr-serverless:StartApplication",
        "emr-serverless:StartJobRun",
        "emr-serverless:CancelJobRun",
        "emr-serverless:ListJobRuns",
        "emr-serverless:GetJobRun"
      ],
      "Resource": "*"
    }
  ]
}
```

## Utilizzo di tag per il controllo degli accessi

È possibile utilizzare le condizioni dei tag per un controllo granulare degli accessi. Ad esempio, puoi limitare gli utenti di un team in modo che possano inviare lavori solo alle applicazioni EMR Serverless contrassegnate con il nome del team.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "EMRServerlessActions",
      "Effect": "Allow",
      "Action": [
        "emr-serverless:ListApplications",
        "emr-serverless:GetApplication",
```

```
        "emr-serverless:StartApplication",
        "emr-serverless:StartJobRun",
        "emr-serverless:CancelJobRun",
        "emr-serverless:ListJobRuns",
        "emr-serverless:GetJobRun"
    ],
    "Resource": "*",
    "Condition": {
        "StringEquals": {
            "aws:ResourceTag/Team": "team-name"
        }
    }
}
]
```

## Policy per il controllo degli accessi basato su tag

È possibile utilizzare le condizioni della policy basata sull'identità per controllare l'accesso alle applicazioni e le esecuzioni di lavoro in base ai tag.

Gli esempi seguenti mostrano diversi scenari e modi di utilizzare gli operatori di condizione con le chiavi di condizione EMR Serverless. Queste dichiarazioni di policy IAM sono concepite unicamente per scopi dimostrativi e non devono essere utilizzate negli ambienti di produzione. Esistono vari modi di combinare dichiarazioni di policy per concedere o negare autorizzazioni in base alle tue esigenze. Per ulteriori informazioni sulla pianificazione e sul test di policy IAM, consulta la [Guida per l'utente IAM](#).

### Important

Negare esplicitamente l'autorizzazione per operazioni di assegnazione di tag è una possibilità da tenere in debita considerazione. Ciò impedisce agli utenti di concedersi personalmente autorizzazioni tramite tag di una risorsa che non avevi intenzione di accordare. Se le operazioni di assegnazione di tag per una risorsa non vengono negate, un utente può modificare i tag e aggirare l'intenzione delle policy basate su tag. Per un esempio di policy che nega le operazioni di tag, consulta [Negazione dell'accesso per aggiungere ed eliminare tag](#).

Gli esempi seguenti illustrano le politiche di autorizzazione basate sull'identità utilizzate per controllare le azioni consentite con le applicazioni EMR Serverless.

## Autorizzazione di operazioni solo su risorse con specifici valori di tag

Nel seguente esempio di policy, l'operatore di `StringEquals` condizione cerca di corrispondere al valore del `dev` reparto tag. Se il reparto tag non è stato aggiunto all'applicazione o non contiene il valore `dev`, la politica non si applica e le azioni non sono consentite da questa politica. Se nessun'altra dichiarazione politica consente le azioni, l'utente può lavorare solo con applicazioni che hanno questo tag con questo valore.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "emr-serverless:GetApplication"
      ],
      "Resource": "*",
      "Condition": {
        "StringEquals": {
          "emr-serverless:ResourceTag/department": "dev"
        }
      }
    }
  ]
}
```

Puoi anche specificare più valori di tag utilizzando un operatore di condizione. Ad esempio, per consentire azioni sulle applicazioni in cui il `department` tag contiene il valore `dev` oppure `test`, è possibile sostituire il blocco delle condizioni nell'esempio precedente con il seguente.

```
"Condition": {
  "StringEquals": {
    "emr-serverless:ResourceTag/department": ["dev", "test"]
  }
}
```

## Richiesta dell'assegnazione di tag alla creazione di una risorsa

Nell'esempio seguente, il tag deve essere applicato durante la creazione dell'applicazione.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "emr-serverless:CreateApplication"
      ],
      "Resource": "*",
      "Condition": {
        "StringEquals": {
          "emr-serverless:RequestTag/department": "dev"
        }
      }
    }
  ]
}
```

La seguente dichiarazione politica consente a un utente di creare un'applicazione solo se l'applicazione ha un department tag, che può contenere qualsiasi valore.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "emr-serverless:CreateApplication"
      ],
      "Resource": "*",
      "Condition": {
        "Null": {
          "emr-serverless:RequestTag/department": "false"
        }
      }
    }
  ]
}
```

## Negazione dell'accesso per aggiungere ed eliminare tag

Questa politica impedisce a un utente di aggiungere o rimuovere tag su applicazioni EMR Serverless con un `department` tag il cui valore non è `dev`.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Deny",
      "Action": [
        "emr-serverless:TagResource",
        "emr-serverless:UntagResource"
      ],
      "Resource": "*",
      "Condition": {
        "StringNotEquals": {
          "emr-serverless:ResourceTag/department": "dev"
        }
      }
    }
  ]
}
```

## Esempi di policy basate sull'identità per EMR Serverless

Per impostazione predefinita, gli utenti e i ruoli non dispongono dell'autorizzazione per creare o modificare risorse Amazon EMR Serverless. Inoltre, non possono eseguire attività utilizzando AWS Management Console, AWS Command Line Interface (AWS CLI) o AWS API. Per concedere agli utenti l'autorizzazione a eseguire operazioni sulle risorse di cui hanno bisogno, un amministratore IAM può creare policy IAM. L'amministratore può quindi aggiungere le policy IAM ai ruoli e gli utenti possono assumere i ruoli.

Per informazioni su come creare una policy basata su identità IAM utilizzando questi documenti di policy JSON di esempio, consulta [Creazione di policy IAM \(console\)](#) nella Guida per l'utente IAM.

Per dettagli sulle azioni e sui tipi di risorse definiti da Amazon EMR Serverless, incluso il formato di ARNs per ogni tipo di risorsa, consulta [Azioni, risorse e chiavi di condizione per Amazon EMR Serverless](#) nel Service Authorization Reference.

### Argomenti

- [Best practice per le policy](#)
- [Consentire agli utenti di visualizzare le loro autorizzazioni](#)

## Best practice per le policy

### Note

EMR Serverless non supporta le policy gestite, quindi la prima pratica elencata di seguito non si applica.

Le policy basate sull'identità determinano se qualcuno può creare, accedere o eliminare le risorse Serverless di Amazon EMR nel tuo account. Queste azioni possono comportare costi aggiuntivi per l' Account AWS. Quando crei o modifichi policy basate su identità, segui queste linee guida e raccomandazioni:

- Inizia con le policy AWS gestite e passa alle autorizzazioni con privilegi minimi: per iniziare a concedere autorizzazioni a utenti e carichi di lavoro, utilizza le policy gestite che concedono le autorizzazioni per molti casi d'uso comuni. AWS Sono disponibili nel tuo Account AWS Ti consigliamo di ridurre ulteriormente le autorizzazioni definendo politiche gestite dai AWS clienti specifiche per i tuoi casi d'uso. Per ulteriori informazioni, consulta [Policy gestite da AWS](#) o [Policy gestite da AWS per le funzioni dei processi](#) nella Guida per l'utente IAM.
- Applica le autorizzazioni con privilegio minimo: quando imposti le autorizzazioni con le policy IAM, concedi solo le autorizzazioni richieste per eseguire un'attività. È possibile farlo definendo le azioni che possono essere intraprese su risorse specifiche in condizioni specifiche, note anche come autorizzazioni con privilegi minimi. Per ulteriori informazioni sull'utilizzo di IAM per applicare le autorizzazioni, consulta [Policy e autorizzazioni in IAM](#) nella Guida per l'utente IAM.
- Condizioni d'uso nelle policy IAM per limitare ulteriormente l'accesso: per limitare l'accesso a operazioni e risorse è possibile aggiungere una condizione alle tue policy. Ad esempio, è possibile scrivere una condizione di policy per specificare che tutte le richieste devono essere inviate utilizzando SSL. Puoi anche utilizzare le condizioni per concedere l'accesso alle azioni del servizio se vengono utilizzate tramite uno specifico Servizio AWS, ad esempio AWS CloudFormation. Per ulteriori informazioni, consulta la sezione [Elementi delle policy JSON di IAM: condizione](#) nella Guida per l'utente IAM.
- Utilizzo di IAM Access Analyzer per convalidare le policy IAM e garantire autorizzazioni sicure e funzionali: IAM Access Analyzer convalida le policy nuove ed esistenti in modo che aderiscano alla

sintassi della policy IAM (JSON) e alle best practice di IAM. IAM Access Analyzer offre oltre 100 controlli delle policy e consigli utili per creare policy sicure e funzionali. Per ulteriori informazioni, consulta [Convalida delle policy per il Sistema di analisi degli accessi IAM](#) nella Guida per l'utente IAM.

- Richiedi l'autenticazione a più fattori (MFA): se hai uno scenario che richiede utenti IAM o un utente root nel Account AWS tuo, attiva l'MFA per una maggiore sicurezza. Per richiedere la MFA quando vengono chiamate le operazioni API, aggiungi le condizioni MFA alle policy. Per ulteriori informazioni, consulta [Protezione dell'accesso API con MFA](#) nella Guida per l'utente IAM.

Per maggiori informazioni sulle best practice in IAM, consulta [Best practice di sicurezza in IAM](#) nella Guida per l'utente di IAM.

## Consentire agli utenti di visualizzare le loro autorizzazioni

Questo esempio mostra in che modo è possibile creare una policy che consente agli utenti IAM di visualizzare le policy inline e gestite che sono collegate alla relativa identità utente. Questa policy include le autorizzazioni per completare questa azione sulla console o utilizzando programmaticamente l'API o. AWS CLI AWS

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "ViewOwnUserInfo",
      "Effect": "Allow",
      "Action": [
        "iam:GetUserPolicy",
        "iam:ListGroupsWithUser",
        "iam:ListAttachedUserPolicies",
        "iam:ListUserPolicies",
        "iam:GetUser"
      ],
      "Resource": ["arn:aws:iam::*:user/${aws:username}"]
    },
    {
      "Sid": "NavigateInConsole",
      "Effect": "Allow",
      "Action": [
        "iam:GetGroupPolicy",
        "iam:GetPolicyVersion",
```

```

        "iam:GetPolicy",
        "iam:ListAttachedGroupPolicies",
        "iam:ListGroupPolicies",
        "iam:ListPolicyVersions",
        "iam:ListPolicies",
        "iam:ListUsers"
    ],
    "Resource": "*"
}
]
}

```

## Amazon EMR Serverless: aggiornamenti alle policy gestite AWS

Visualizza i dettagli sugli aggiornamenti delle politiche AWS gestite per Amazon EMR Serverless da quando questo servizio ha iniziato a tracciare queste modifiche. [Per ricevere avvisi automatici sulle modifiche a questa pagina, iscriviti al feed RSS nella pagina della cronologia dei documenti Serverless di Amazon EMR.](#)

Modifica	Descrizione	Data
Amazon EMRServerlessServiceRolePolicy : aggiornamento a una politica esistente	<a href="#">Amazon EMR Serverless ha aggiunto il nuovo SidCloudWatchPolicyStatement e EC2PolicyStatement alla politica di Amazon. EMRServerlessServiceRolePolicy</a>	25 gennaio 2024
Amazon EMRServerlessServiceRolePolicy : aggiornamento a una politica esistente	Amazon EMR Serverless ha aggiunto nuove autorizzazioni per consentire ad Amazon EMR Serverless di pubblicare e parametri aggregati degli account per l'utilizzo di vCPU	20 aprile 2023

Modifica	Descrizione	Data
	nello spazio dei nomi. "AWS/Usage"	
Amazon EMR Serverless ha iniziato a tracciare le modifiche	Amazon EMR Serverless ha iniziato a tracciare le modifiche per le sue AWS policy gestite.	20 aprile 2023

## Risoluzione dei problemi relativi all'identità e all'accesso senza server di Amazon EMR

Utilizza le seguenti informazioni per aiutarti a diagnosticare e risolvere i problemi più comuni che potresti riscontrare quando lavori con Amazon EMR Serverless e IAM.

### Argomenti

- [Non sono autorizzato a eseguire un'azione in Amazon EMR Serverless](#)
- [Non sono autorizzato a eseguire iam: PassRole](#)
- [Voglio consentire a persone esterne al mio AWS account di accedere alle mie risorse Amazon EMR Serverless](#)
- [Non riesco ad aprire il server di cronologia Live UI/Spark da EMR Studio per eseguire il debug del mio lavoro o si verifica un errore API quando cerco di ottenere i log utilizzando get-dashboard-for-job-run](#)

### Non sono autorizzato a eseguire un'azione in Amazon EMR Serverless

Se ti AWS Management Console dice che non sei autorizzato a eseguire un'azione, devi contattare l'amministratore per ricevere assistenza. L'amministratore è la persona da cui si sono ricevuti il nome utente e la password.

Il seguente esempio di errore si verifica quando l'utente mateojackson prova a utilizzare la console per visualizzare i dettagli relativi a una risorsa *my-example-widget* fittizia, ma non dispone di autorizzazioni `emr-serverless:GetWidget` fittizie.

```
User: arn:aws:iam::123456789012:user/mateojackson is not authorized to perform: emr-serverless:GetWidget on resource: my-example-widget
```

In questo caso, Mateo richiede al suo amministratore di aggiornare le policy per poter accedere alla risorsa *my-example-widget* utilizzando l'operazione `emr-serverless:GetWidget`.

## Non sono autorizzato a eseguire iam: PassRole

Se ricevi un messaggio di errore indicante che non sei autorizzato a eseguire l'`iam:PassRole` operazione, le tue policy devono essere aggiornate per consentirti di trasferire un ruolo ad Amazon EMR Serverless.

Alcuni Servizi AWS consentono di trasferire un ruolo esistente a quel servizio invece di creare un nuovo ruolo di servizio o un ruolo collegato al servizio. Per eseguire questa operazione, è necessario disporre delle autorizzazioni per trasmettere il ruolo al servizio.

Il seguente errore di esempio si verifica quando un utente IAM denominato `marymajor` tenta di utilizzare la console per eseguire un'azione in Amazon EMR Serverless. Tuttavia, l'azione richiede che il servizio disponga delle autorizzazioni concesse da un ruolo di servizio. Mary non dispone delle autorizzazioni per passare il ruolo al servizio.

```
User: arn:aws:iam::123456789012:user/marymajor is not authorized to perform:
iam:PassRole
```

In questo caso, le policy di Mary devono essere aggiornate per poter eseguire l'operazione `iam:PassRole`.

Se hai bisogno di aiuto, contatta il tuo AWS amministratore. L'amministratore è la persona che ti ha fornito le credenziali di accesso.

## Voglio consentire a persone esterne al mio AWS account di accedere alle mie risorse Amazon EMR Serverless

È possibile creare un ruolo con il quale utenti in altri account o persone esterne all'organizzazione possono accedere alle tue risorse. È possibile specificare chi è attendibile per l'assunzione del ruolo. Per i servizi che supportano politiche basate sulle risorse o liste di controllo degli accessi (ACLs), puoi utilizzare tali policy per consentire alle persone di accedere alle tue risorse.

Per ulteriori informazioni, consulta gli argomenti seguenti:

- Per sapere se Amazon EMR Serverless supporta queste funzionalità, consulta. [Identity and Access Management \(IAM\) in Amazon EMR Serverless](#)

- Per sapere come fornire l'accesso alle tue risorse attraverso Account AWS le risorse di tua proprietà, consulta [Fornire l'accesso a un utente IAM in un altro Account AWS di tua proprietà](#) nella IAM User Guide.
- Per scoprire come fornire l'accesso alle tue risorse a terze parti Account AWS, consulta [Fornire l'accesso a soggetti Account AWS di proprietà di terze parti](#) nella Guida per l'utente IAM.
- Per informazioni su come fornire l'accesso tramite la federazione delle identità, consulta [Fornire l'accesso a utenti autenticati esternamente \(Federazione delle identità\)](#) nella Guida per l'utente IAM.
- Per informazioni sulle differenze di utilizzo tra ruoli e policy basate su risorse per l'accesso multi-account, consulta [Accesso a risorse multi-account in IAM](#) nella Guida per l'utente IAM.

Non riesco ad aprire il server di cronologia Live UI/Spark da EMR Studio per eseguire il debug del mio lavoro o si verifica un errore API quando cerco di ottenere i log utilizzando **get-dashboard-for-job-run**

Se utilizzi lo storage gestito EMR Serverless per la registrazione e la tua applicazione EMR Serverless si trova in una sottorete privata con endpoint VPC per Amazon S3 e alleggi una policy di endpoint per controllare l'accesso, devi aggiungere le autorizzazioni menzionate in Logging for [EMR Serverless con storage gestito nella tua politica VPC all'endpoint gateway S3 per EMR Serverless a](#) archivia e servi i registri delle applicazioni.

## Utilizzo di EMR Serverless con Lake Formation

Scopri come utilizzare Lake Formation per fornire un controllo granulare degli accessi.

### Accesso non filtrato a Lake Formation per EMR Serverless

Con le versioni 7.8.0 e successive di Amazon EMR, puoi sfruttare AWS Lake Formation with Glue Data Catalog, dove il ruolo di job runtime dispone di autorizzazioni complete per le tabelle senza le limitazioni del controllo granulare degli accessi. Questa funzionalità consente di leggere e scrivere su tabelle protette da Lake Formation dal batch EMR Serverless Spark e dai lavori interattivi. Consulta le seguenti sezioni per saperne di più su Lake Formation e su come utilizzarlo con EMR Serverless.

### Utilizzo di Lake Formation con accesso completo ai tavoli

È possibile accedere alle tabelle del catalogo Glue Data protette da AWS Lake Formation dai job EMR Serverless Spark o dalle sessioni interattive in cui il ruolo di runtime del job ha accesso

completo alla tabella. Non è necessario abilitare AWS Lake Formation sull'applicazione EMR Serverless. Quando un job Spark è configurato per Full Table Access (FTA), le credenziali di AWS Lake Formation vengono utilizzate per leggere/scrivere i dati S3 per le tabelle registrate di Lake AWS Formation, mentre le credenziali del ruolo di runtime del job verranno utilizzate per leggere/scrivere tabelle non registrate con Lake Formation. AWS

### Important

Non abilitare AWS Lake Formation per il controllo granulare degli accessi. Un job non può eseguire contemporaneamente Full Table Access (FTA) e Fine-Grained Access Control (FGAC) sullo stesso cluster o applicazione EMR.

## Passaggio 1: abilitare l'accesso completo alla tabella in Lake Formation

Per utilizzare la modalità Full Table Access (FTA), devi consentire ai motori di query di terze parti di accedere ai dati senza la convalida dei tag di sessione IAM in AWS Lake Formation. Per abilitarla, segui i passaggi indicati in [Integrazione delle applicazioni per l'accesso completo alla tabella](#).

### Note

Quando si accede alle tabelle con più account, è necessario abilitare l'accesso completo alla tabella sia negli account produttore che in quelli consumatori. Allo stesso modo, quando si accede alle tabelle interregionali, questa impostazione deve essere abilitata sia nelle regioni produttrici che in quelle di consumo.

## Fase 2: Configurazione delle autorizzazioni IAM per il ruolo Job Runtime

Per l'accesso in lettura o scrittura ai dati sottostanti, oltre alle autorizzazioni di Lake Formation, un ruolo di job runtime richiede l'autorizzazione `lakeformation:GetDataAccess` IAM. Con questa autorizzazione, Lake Formation concede la richiesta di credenziali temporanee per accedere ai dati.

Di seguito è riportato un esempio di politica su come fornire le autorizzazioni IAM per accedere a uno script in Amazon S3, caricare i log su S3, le autorizzazioni dell'API AWS Glue e l'autorizzazione per accedere a Lake Formation.

```
{
  "Version": "2012-10-17",
  "Statement": [
```

```
{
  "Sid": "ScriptAccess",
  "Effect": "Allow",
  "Action": [
    "s3:GetObject"
  ],
  "Resource": [
    "arn:aws:s3:::*.amzn-s3-demo-bucket/scripts"
  ]
},
{
  "Sid": "LoggingAccess",
  "Effect": "Allow",
  "Action": [
    "s3:PutObject"
  ],
  "Resource": [
    "arn:aws:s3:::amzn-s3-demo-bucket/logs/*"
  ]
},
{
  "Sid": "GlueCatalogAccess",
  "Effect": "Allow",
  "Action": [
    "glue:Get*",
    "glue:Create*",
    "glue:Update*"
  ],
  "Resource": [
    "*"
  ]
},
{
  "Sid": "LakeFormationAccess",
  "Effect": "Allow",
  "Action": [
    "lakeformation:GetDataAccess"
  ],
  "Resource": [
    "*"
  ]
}
]
```

```
}
```

## Passaggio 2.1 Configurare i permessi di Lake Formation

- I job Spark che leggono dati da S3 richiedono l'autorizzazione Lake Formation SELECT.
- I lavori Spark che scrivono/eliminano dati in S3 richiedono l'autorizzazione Lake Formation ALL (SUPER).
- I lavori Spark che interagiscono con il catalogo Glue Data richiedono l'autorizzazione DESCRIBE, ALTER, DROP, a seconda dei casi.

Per ulteriori informazioni, consulta [Concessione delle autorizzazioni sulle risorse di Data Catalog](#).

Passaggio 3: inizializza una sessione Spark per l'accesso completo alla tabella utilizzando Lake Formation

### Prerequisiti

AWS Glue Data Catalog deve essere configurato come metastore per accedere alle tabelle di Lake Formation.

Imposta le seguenti impostazioni per configurare il catalogo Glue come metastore:

```
--conf spark.sql.catalogImplementation=hive
--conf
  spark.hive.metastore.client.factory.class=com.amazonaws.glue.catalog.metastore.AWSGlueDataCatalogHiveMetastore
```

Per ulteriori informazioni sull'attivazione di Data Catalog per EMR Serverless, vedere [Configurazione Metastore](#) per EMR Serverless.

Per accedere alle tabelle registrate con AWS Lake Formation, è necessario impostare le seguenti configurazioni durante l'inizializzazione di Spark per configurare Spark per utilizzare le credenziali di Lake Formation AWS .

### Hive

```
--conf
  spark.hadoop.fs.s3.credentialsResolverClass=com.amazonaws.glue.accesscontrol.AWSLakeFormationCredentialsResolver
--conf spark.hadoop.fs.s3.useDirectoryHeaderAsFolderObject=true
--conf spark.hadoop.fs.s3.folderObject.autoAction.disabled=true
```

```
--conf spark.sql.catalog.skipLocationValidationOnCreateTable.enabled=true
--conf spark.sql.catalog.createDirectoryAfterTable.enabled=true
--conf spark.sql.catalog.dropDirectoryBeforeTable.enabled=true
```

## Iceberg

```
--conf spark.sql.catalog.spark_catalog=org.apache.iceberg.spark.SparkSessionCatalog
--conf spark.sql.catalog.spark_catalog.warehouse=S3_DATA_LOCATION
--conf spark.sql.catalog.spark_catalog.client.region=REGION
--conf spark.sql.catalog.spark_catalog.type=glue
--conf spark.sql.catalog.spark_catalog.glue.account-id=ACCOUNT_ID
--conf spark.sql.catalog.spark_catalog.glue.lakeformation-enabled=true
--conf spark.sql.catalog.dropDirectoryBeforeTable.enabled=true
```

- `spark.hadoop.fs.s3.credentialsResolverClass=com.amazonaws.glue.accesscontrol.AWSCredentialsProvider`  
Configurare EMR Filesystem (EMRFS) per utilizzare le credenziali di Lake AWS Formation S3 per le tabelle registrate di Lake Formation. Se la tabella non è registrata, utilizza le credenziali del ruolo di runtime del job.
- `spark.hadoop.fs.s3.useDirectoryHeaderAsFolderObject=true`  
configura EMRFS per utilizzare Content Type Header `application/x-directory` invece del suffisso `$folder$` durante la creazione di cartelle S3. Questo è necessario per leggere le tabelle di Lake Formation, poiché le credenziali di Lake Formation non consentono la lettura delle cartelle delle tabelle con il suffisso `$folder$`.
- `spark.sql.catalog.skipLocationValidationOnCreateTable.enabled=true`:  
Configura Spark per saltare la convalida del vuoto della posizione della tabella prima della creazione. Ciò è necessario per le tabelle registrate di Lake Formation, poiché le credenziali di Lake Formation per verificare la posizione vuota sono disponibili solo dopo la creazione della tabella Glue Data Catalog. Senza questa configurazione, le credenziali del ruolo di runtime del job convalideranno la posizione della tabella vuota.
- `spark.sql.catalog.createDirectoryAfterTable.enabled=true`: configura Spark per creare la cartella Amazon S3 dopo la creazione della tabella nel metastore Hive. Questo è necessario per le tabelle registrate di Lake Formation, poiché le credenziali di Lake Formation per creare la cartella S3 sono disponibili solo dopo la creazione della tabella Glue Data Catalog.
- `spark.sql.catalog.dropDirectoryBeforeTable.enabled=true`: Configura Spark per eliminare la cartella S3 prima dell'eliminazione della tabella nel metastore Hive. Ciò è necessario per le tabelle registrate di Lake Formation, poiché le credenziali di Lake Formation per eliminare la cartella S3 non sono disponibili dopo l'eliminazione della tabella dal Glue Data Catalog.

- `spark.sql.catalog.<catalog>.glue.lakeformation-enabled=true`: Configura il catalogo Iceberg per utilizzare le credenziali di AWS Lake Formation S3 per le tabelle registrate di Lake Formation. Se la tabella non è registrata, usa le credenziali di ambiente predefinite.

## Configura la modalità di accesso completo alla tabella in SageMaker Unified Studio

Per accedere alle tabelle registrate di Lake Formation dalle sessioni interattive di Spark nei JupyterLab notebook, è necessario utilizzare la modalità di autorizzazione alla compatibilità. Usa il comando `magic %%configure` per configurare la configurazione di Spark. Scegli la configurazione in base al tipo di tabella:

### For Hive tables

```
%%configure -f
{
  "conf": {
    "spark.hadoop.fs.s3.credentialsResolverClass":
"com.amazonaws.glue.accesscontrol.AWSLakeFormationCredentialResolver",
    "spark.hadoop.fs.s3.useDirectoryHeaderAsFolderObject": true,
    "spark.hadoop.fs.s3.folderObject.autoAction.disabled": true,
    "spark.sql.catalog.skipLocationValidationOnCreateTable.enabled": true,
    "spark.sql.catalog.createDirectoryAfterTable.enabled": true,
    "spark.sql.catalog.dropDirectoryBeforeTable.enabled": true
  }
}
```

### For Iceberg tables

```
%%configure -f
{
  "conf": {
    "spark.sql.catalog.spark_catalog":
"org.apache.iceberg.spark.SparkSessionCatalog",
    "spark.sql.catalog.spark_catalog.warehouse": "S3_DATA_LOCATION",
    "spark.sql.catalog.spark_catalog.client.region": "REGION",
    "spark.sql.catalog.spark_catalog.type": "glue",
    "spark.sql.catalog.spark_catalog.glue.account-id": "ACCOUNT_ID",
    "spark.sql.catalog.spark_catalog.glue.lakeformation-enabled": true,
    "spark.sql.catalog.dropDirectoryBeforeTable.enabled": true,
  }
}
```

Sostituisci i segnaposto:

- S3\_DATA\_LOCATION: Il percorso del tuo bucket S3
- REGION: AWS regione (ad es. us-east-1)
- ACCOUNT\_ID: L'ID del tuo account AWS

 Note

È necessario impostare queste configurazioni prima di eseguire qualsiasi operazione Spark sul notebook.

Operazioni supportate

Queste operazioni utilizzeranno le credenziali di AWS Lake Formation per accedere ai dati della tabella.

- CREATE TABLE
- ALTER TABLE
- INSERT INTO
- INSERT OVERWRITE
- UPDATE
- MERGE INTO
- DELETE FROM
- ANALIZZA LA TABELLA
- TABELLA DI RIPARAZIONE
- DROP TABLE
- Query su origini dati Spark
- Scritture di origini dati Spark

 Note

Le operazioni non elencate sopra continueranno a utilizzare le autorizzazioni IAM per accedere ai dati delle tabelle.

## Considerazioni

- Se una tabella Hive viene creata utilizzando un lavoro per cui non è abilitato l'accesso completo alla tabella e non viene inserito alcun record, le letture o scritture successive da un lavoro con accesso completo alla tabella avranno esito negativo. Questo perché EMR Spark senza accesso completo alla tabella aggiunge il `$folder$` suffisso al nome della cartella della tabella. Per risolvere questo problema, puoi:
  - Inserire almeno una riga nella tabella da un lavoro in cui l'FTA non è abilitato.
  - Configura il lavoro che non ha FTA abilitato in modo che non utilizzi il `$folder$` suffisso nel nome della cartella in S3. Ciò può essere ottenuto impostando la configurazione di Spark.  
`spark.hadoop.fs.s3.useDirectoryHeaderAsFolderObject=true`
  - Crea una cartella S3 nella posizione della tabella `s3://path/to/table/table_name` utilizzando la console AWS S3 o la CLI AWS S3.
- Full Table Access funziona esclusivamente con EMR Filesystem (EMRFS). Il file system S3A non è compatibile.
- L'accesso completo alla tabella è supportato per le tabelle Hive e Iceberg. Il supporto per le tabelle Hudi e Delta non è stato ancora aggiunto.
- I lavori che fanno riferimento alle tabelle con le regole Lake Formation Fine-Grained Access Control (FGAC) o Glue Data Catalog Views avranno esito negativo. Per interrogare una tabella con una regola FGAC o una Glue Data Catalog View, è necessario utilizzare la modalità FGAC. È possibile abilitare la modalità FGAC seguendo i passaggi descritti nella AWS documentazione: Utilizzo di [EMR Serverless with Lake AWS Formation](#) per un controllo granulare degli accessi.
- L'accesso completo alla tabella non supporta Spark Streaming.
- Quando si scrive Spark DataFrame su una tabella Lake Formation, è supportata solo la modalità APPEND: `df.write.mode("append").saveAsTable(table_name)`

## Utilizzo di EMR Serverless per un controllo granulare degli AWS Lake Formation accessi

### Panoramica

Con le versioni 7.2.0 e successive di Amazon EMR, puoi sfruttare AWS Lake Formation per applicare controlli di accesso granulari alle tabelle di Data Catalog supportate da S3. Questa funzionalità consente di configurare i controlli di accesso a livello di tabella, riga, colonna e cella per read le query all'interno dei job Amazon EMR Serverless Spark. Per configurare un controllo granulare degli

accessi per i processi batch e le sessioni interattive di Apache Spark, usa EMR Studio. Consulta le seguenti sezioni per saperne di più su Lake Formation e su come utilizzarlo con EMR Serverless.

L'utilizzo di Amazon EMR Serverless AWS Lake Formation comporta costi aggiuntivi. Per ulteriori informazioni, consulta [Prezzi di Amazon EMR](#).

## Come funziona EMR Serverless con AWS Lake Formation

L'utilizzo di EMR Serverless con Lake Formation ti consente di applicare un livello di autorizzazioni su ogni lavoro Spark per applicare il controllo delle autorizzazioni di Lake Formation quando EMR Serverless esegue i lavori. EMR Serverless utilizza i profili di [risorse Spark per creare due profili per eseguire i lavori](#) in modo efficace. Il profilo utente esegue il codice fornito dall'utente, mentre il profilo di sistema applica le politiche di Lake Formation. Per ulteriori informazioni, consulta [Cos'è AWS Lake Formation](#) e [Considerazioni](#) e limitazioni.

Quando utilizzi la capacità preinizializzata con Lake Formation, ti consigliamo di avere almeno due driver Spark. Ogni job abilitato per Lake Formation utilizza due driver Spark, uno per il profilo utente e uno per il profilo di sistema. Per ottenere prestazioni ottimali, è necessario utilizzare il doppio del numero di driver per i lavori abilitati a Lake Formation rispetto a chi non utilizza Lake Formation.

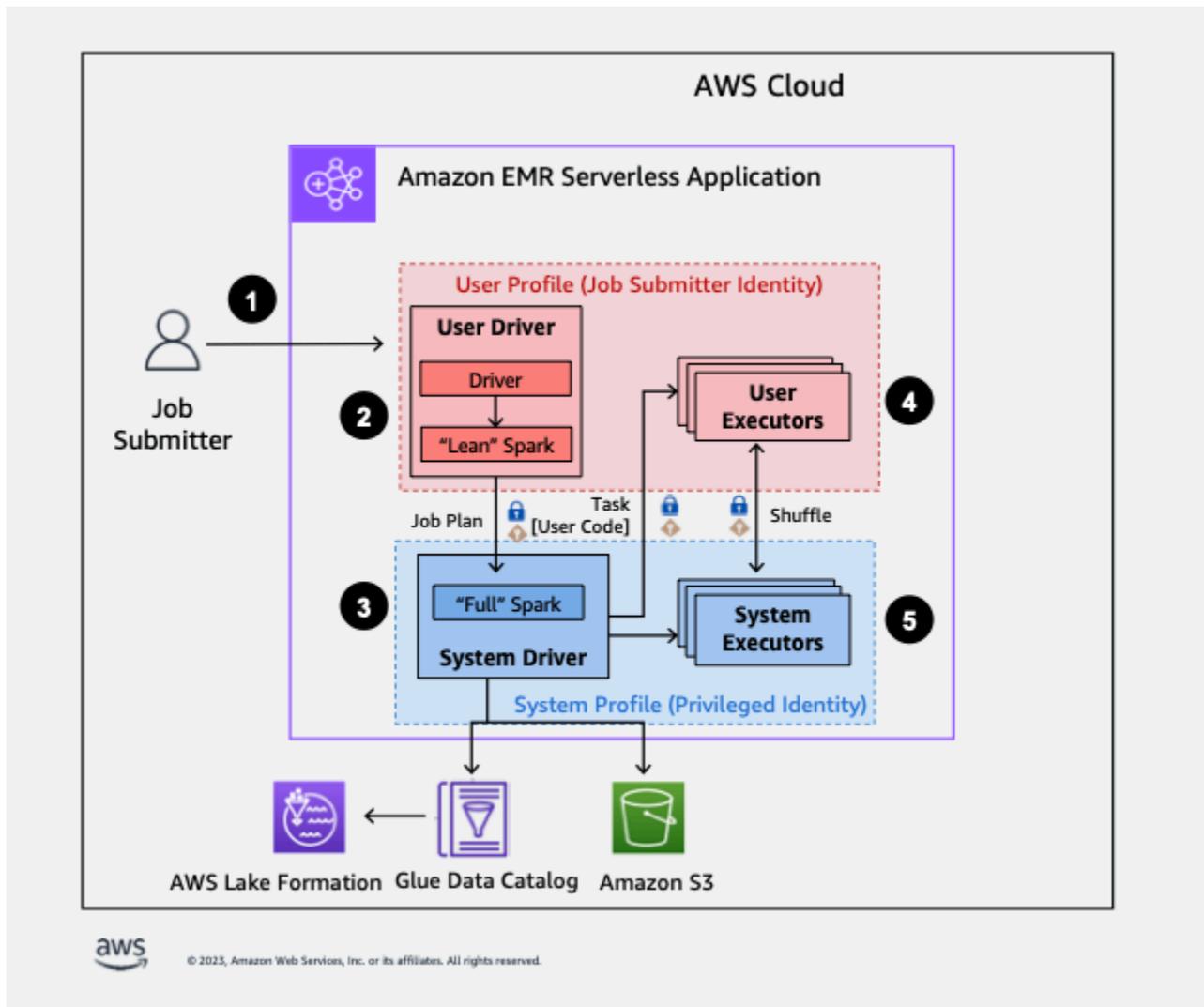
Quando esegui i job Spark su EMR Serverless, devi anche considerare l'impatto dell'allocazione dinamica sulla gestione delle risorse e sulle prestazioni del cluster. La configurazione `spark.dynamicAllocation.maxExecutors` del numero massimo di executor per profilo di risorsa si applica sia agli executor utente che a quelli di sistema. Se si configura tale numero in modo che sia uguale al numero massimo consentito di executor, l'esecuzione del job potrebbe bloccarsi a causa di un tipo di executor che utilizza tutte le risorse disponibili, impedendo all'altro executor di eseguire i job di job.

Per non esaurire le risorse, EMR Serverless imposta il numero massimo predefinito di executor per profilo di risorsa al 90% del valore. `spark.dynamicAllocation.maxExecutors`

È possibile sovrascrivere questa configurazione quando si specifica `spark.dynamicAllocation.maxExecutorsRatio` un valore compreso tra 0 e 1. Inoltre, è possibile configurare le seguenti proprietà per ottimizzare l'allocazione delle risorse e le prestazioni complessive:

- `spark.dynamicAllocation.cachedExecutorIdleTimeout`
- `spark.dynamicAllocation.shuffleTracking.timeout`
- `spark.cleaner.periodicGC.interval`

Di seguito è riportata una panoramica di alto livello su come EMR Serverless ottiene l'accesso ai dati protetti dalle politiche di sicurezza di Lake Formation.



1. Un utente invia un job Spark a un'applicazione EMR AWS Lake Formation Serverless abilitata.
2. EMR Serverless invia il lavoro a un driver utente ed esegue il lavoro nel profilo utente. Il driver utente esegue una versione snella di Spark che non è in grado di avviare attività, richiedere esecutori, accedere a S3 o al Glue Catalog. Costruisce un piano di lavoro.
3. EMR Serverless imposta un secondo driver chiamato driver di sistema e lo esegue nel profilo di sistema (con un'identità privilegiata). EMR Serverless configura un canale TLS crittografato tra i due driver per la comunicazione. Il driver utente utilizza il canale per inviare i piani di lavoro al driver di sistema. Il driver di sistema non esegue il codice inviato dall'utente. Esegue Spark completo e comunica con S3 e il Data Catalog per l'accesso ai dati. Richiede esecutori e compila il Job Plan in una sequenza di fasi di esecuzione.

4. EMR Serverless esegue quindi le fasi sugli executor con il driver utente o il driver di sistema. Il codice utente in qualsiasi fase viene eseguito esclusivamente sugli executor dei profili utente.
5. Le fasi che leggono i dati dalle tabelle del Data Catalog protette da AWS Lake Formation o che applicano filtri di sicurezza sono delegate agli executor di sistema.

## Abilitazione di Lake Formation in Amazon EMR

Per abilitare Lake Formation, è necessario impostare `spark.emr-serverless.lakeformation.enabled` la `true` `spark-defaults` sottoclassificazione per il parametro di configurazione di runtime durante la [creazione di un'applicazione EMR Serverless](#).

```
aws emr-serverless create-application \  
  --release-label emr-7.9.0 \  
  --runtime-configuration '{  
    "classification": "spark-defaults",  
    "properties": {  
      "spark.emr-serverless.lakeformation.enabled": "true"  
    }  
  }' \  
  --type "SPARK"
```

Puoi anche abilitare Lake Formation quando crei una nuova applicazione in EMR Studio. Scegli Use Lake Formation per un controllo granulare degli accessi, disponibile in Configurazioni aggiuntive.

La [crittografia tra lavoratori](#) è abilitata per impostazione predefinita quando si utilizza Lake Formation con EMR Serverless, quindi non è necessario abilitare nuovamente la crittografia tra lavoratori.

### Attivazione dei lavori di Lake Formation per Spark

Per abilitare Lake Formation per i singoli job Spark, imposta su `spark.emr-serverless.lakeformation.enabled true` durante l'utilizzo `spark-submit`.

```
--conf spark.emr-serverless.lakeformation.enabled=true
```

## Autorizzazioni IAM per il ruolo Job Runtime

Le autorizzazioni di Lake Formation controllano l'accesso alle risorse di AWS Glue Data Catalog, alle sedi Amazon S3 e ai dati sottostanti in tali sedi. Le autorizzazioni IAM controllano l'accesso a

Lake Formation and AWS Glue APIs e alle risorse. Sebbene tu possa avere l'autorizzazione Lake Formation per accedere a una tabella nel Data Catalog (SELECT), l'operazione fallisce se non disponi dell'autorizzazione IAM sull'operazione `glue:Get*` API.

Di seguito è riportato un esempio di politica su come fornire le autorizzazioni IAM per accedere a uno script in S3, caricare i log su S3, le autorizzazioni dell'API AWS Glue e l'autorizzazione per accedere a Lake Formation.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "ScriptAccess",
      "Effect": "Allow",
      "Action": [
        "s3:GetObject",
        "s3:ListBucket"
      ],
      "Resource": [
        "arn:aws:s3::*.amzn-s3-demo-bucket/scripts",
        "arn:aws:s3::*.amzn-s3-demo-bucket/*" ]
    },
    {
      "Sid": "LoggingAccess",
      "Effect": "Allow",
      "Action": [
        "s3:PutObject"
      ],
      "Resource": [
        "arn:aws:s3:::amzn-s3-demo-bucket/logs/*"
      ]
    },
    {
      "Sid": "GlueCatalogAccess",
      "Effect": "Allow",
      "Action": [
        "glue:Get*",
        "glue:Create*",
        "glue:Update*"
      ],
      "Resource": ["*"]
    },
    {
```

```

        "Sid": "LakeFormationAccess",
        "Effect": "Allow",
        "Action": [
            "lakeformation:GetDataAccess"
        ],
        "Resource": ["*"]
    }
]
}

```

## Configurazione delle autorizzazioni di Lake Formation per il ruolo Job Runtime

Innanzitutto, registra la posizione del tuo tavolo Hive con Lake Formation. Quindi crea le autorizzazioni per il tuo ruolo di job runtime nella tabella desiderata. Per maggiori dettagli su Lake Formation, vedi [What is AWS Lake Formation?](#) nella Guida per gli AWS Lake Formation sviluppatori.

Dopo aver configurato le autorizzazioni di Lake Formation, puoi inviare lavori Spark su Amazon EMR Serverless. [Per ulteriori informazioni sui job Spark, consulta gli esempi di Spark.](#)

### Invio di un job run

Dopo aver completato la configurazione delle sovvenzioni Lake Formation, puoi [inviare lavori Spark su EMR](#) Serverless. Per eseguire i job Iceberg, devi fornire le seguenti proprietà. `spark-submit`

```

--conf spark.sql.catalog.spark_catalog=org.apache.iceberg.spark.SparkSessionCatalog
--conf spark.sql.catalog.spark_catalog.warehouse=<S3_DATA_LOCATION>
--conf spark.sql.catalog.spark_catalog.glue.account-id=<ACCOUNT_ID>
--conf spark.sql.catalog.spark_catalog.client.region=<REGION>
--conf spark.sql.catalog.spark_catalog.glue.endpoint=https://
glue.<REGION>.amazonaws.com

```

## Supporto per il formato a tabella aperta

La versione 7.2.0 di Amazon EMR include il supporto per il controllo granulare degli accessi basato su Lake Formation. EMR Serverless supporta i tipi di tabelle Hive e Iceberg. La tabella seguente descrive tutte le operazioni supportate.

Operazioni	Hive	Iceberg
Comandi DDL	Solo con le autorizzazioni dei ruoli IAM	Solo con le autorizzazioni dei ruoli IAM

Operazioni	Hive	Iceberg
Query incrementali	Non applicabile	Completamente supportato
Query temporali	Non applicabile a questo formato di tabella	Completamente supportato
Tabelle dei metadati	Non applicabile a questo formato di tabella	Supportato, ma alcune tabelle sono nascoste. Per ulteriori informazioni, consulta <a href="#">Considerazioni e limitazioni</a> .
DML INSERT	Solo con autorizzazioni IAM	Solo con autorizzazioni IAM
AGGIORNAMENTO DML	Non applicabile a questo formato di tabella	Solo con autorizzazioni IAM
DML DELETE	Non applicabile a questo formato di tabella	Solo con autorizzazioni IAM
Operazioni di lettura	Completamente supportato	Completamente supportato
Stored procedure	Non applicabile	Supportato con le eccezioni di <code>register_table</code> emigrate. Per ulteriori informazioni, consulta <a href="#">Considerazioni e limitazioni</a> .

## Lavori di debug

### Note

Con questa funzionalità, è possibile visualizzare i log stdout e stderr per i worker dei profili di sistema che possono contenere informazioni sensibili e non filtrate. La seguente autorizzazione deve essere utilizzata solo per accedere ai dati non di produzione. Per le applicazioni create per essere utilizzate con lavori di produzione, consigliamo vivamente di

aggiungere queste autorizzazioni solo agli amministratori o agli utenti con accesso elevato ai dati.

Con EMR-7.3.0 e versioni successive, EMR Serverless abilita la funzionalità di autodebug per i processi batch abilitati per Lake Formation. A tale scopo, utilizzate il nuovo parametro Logs nell'API. accessSystemProfile [GetDashboardForJobRun](#) Se accessSystemProfileLogs è impostato su true, è possibile visualizzare i log stdout e stderr per i worker del profilo di sistema, che possono essere utilizzati per il debug di un processo batch EMR Serverless abilitato per Lake Formation.

```
aws emr-serverless get-dashboard-for-job-run \
  --application-id application-id
  --job-run-id job-run-id
  --access-system-profile-logs
```

## Autorizzazioni richieste

Il principale che desidera eseguire il debug dei processi batch abilitati per Lake Formation `GetDashboardForJobRun` deve disporre delle seguenti autorizzazioni aggiuntive:

```
{
  "Sid": "AccessSystemProfileLogs",
  "Effect": "Allow",
  "Action": [
    "emr-serverless:GetDashboardForJobRun",
    "emr-serverless:AccessSystemProfileLogs",
    "glue:GetDatabases",
    "glue:SearchTables"
  ],
  "Resource": [
    "arn:aws:emr-serverless:region:account-id:/applications/applicationId/jobruns/jobid",
    "arn:aws:glue:region:account-id:catalog",
    "arn:aws:glue:region:account-id:database/*",
    "arn:aws:glue:region:account-id:table/*/*"
  ]
}
```

## Considerazioni

I log dei profili di sistema per il debug sono visibili per i lavori che accedono a database o tabelle in Lake Formation all'interno dello stesso account del lavoro. Non sono visibili nei seguenti scenari:

- Se il catalogo dati gestito utilizzando le autorizzazioni di Lake Formation ha database e tabelle tra account
- Se il catalogo dati gestito utilizzando le autorizzazioni di Lake Formation contiene collegamenti alle risorse

## Utilizzo delle viste del Glue Data Catalog

È possibile creare e gestire viste nel AWS Glue Data Catalog da utilizzare con EMR Serverless. Queste sono note comunemente come viste del AWS Glue Data Catalog. Queste viste sono utili perché supportano più motori di query SQL, quindi puoi accedere alla stessa vista su AWS servizi diversi, come EMR Serverless e Amazon Amazon Athena Redshift.

Creando una vista nel Data Catalog, puoi utilizzare la concessione di risorse e i controlli di accesso basati su tag per concedere l'accesso AWS Lake Formation ad essa. Utilizzando questo metodo di controllo degli accessi, non è necessario configurare un accesso aggiuntivo alle tabelle a cui hai fatto riferimento durante la creazione della vista. Questo metodo di concessione delle autorizzazioni si chiama *definer semantics* e queste viste sono chiamate *definer views*. Per ulteriori informazioni sul controllo degli accessi in Lake Formation, consulta [Concessione e revoca delle autorizzazioni sulle risorse del Data Catalog](#) nella AWS Lake Formation Developer Guide.

Le visualizzazioni del catalogo dati sono utili per i seguenti casi d'uso:

- Controllo granulare degli accessi: è possibile creare una visualizzazione che limiti l'accesso ai dati in base alle autorizzazioni necessarie all'utente. Ad esempio, puoi utilizzare le viste nel Catalogo dati per impedire ai dipendenti che non lavorano nel reparto delle risorse umane di visualizzare le informazioni di identificazione personale (PII).
- Definizione completa della vista: applicando filtri alla visualizzazione nel Data Catalog, ti assicuri che i record di dati disponibili in una vista del Data Catalog siano sempre completi.
- Sicurezza avanzata: la definizione della query utilizzata per creare la vista deve essere completa. Questo vantaggio significa che le visualizzazioni del Data Catalog sono meno suscettibili ai comandi SQL di attori malintenzionati.
- Condivisione semplice dei dati: condividi i dati con altri AWS account senza spostare i dati. Per ulteriori informazioni, consulta [Condivisione dei dati tra account in Lake Formation](#).

## Creazione di una vista di Catalogo Dati

Esistono diversi modi per creare una visualizzazione del catalogo dati. Questi includono l'utilizzo di AWS CLI o Spark SQL. Seguono alcuni esempi.

### Using SQL

Di seguito viene illustrata la sintassi per la creazione di una vista del catalogo dati. Nota il tipo di MULTI DIALECT visualizzazione. Ciò distingue la vista del catalogo dati dalle altre viste. Il SECURITY predicato è specificato come. DEFINER Ciò indica una vista del catalogo dati con DEFINER semantica.

```
CREATE [ OR REPLACE ] PROTECTED MULTI DIALECT VIEW [IF NOT EXISTS] view_name
[(column_name [COMMENT column_comment], ...) ]
[ COMMENT view_comment ]
[TBLPROPERTIES (property_name = property_value, ... )]
SECURITY DEFINER
AS query;
```

Di seguito è riportato un esempio di CREATE dichiarazione, che segue la sintassi:

```
CREATE PROTECTED MULTI DIALECT VIEW catalog_view
SECURITY DEFINER
AS
SELECT order_date, sum(totalprice) AS price
FROM source_table
GROUP BY order_date
```

È inoltre possibile creare una vista in modalità dry-run, utilizzando SQL, per testare la creazione della vista, senza creare effettivamente la risorsa. L'utilizzo di questa opzione comporta una «esecuzione a secco» che convalida l'input e, se la convalida ha esito positivo, restituisce il JSON dell'oggetto della tabella AWS Glue che rappresenterà la vista. In questo caso, la visualizzazione effettiva non viene creata.

```
CREATE [ OR REPLACE ] PROTECTED MULTI DIALECT VIEW view_name
SECURITY DEFINER
[ SHOW VIEW JSON ]
AS view-sql
```

## Using the AWS CLI

### Note

Quando si utilizza il comando CLI, l'SQL utilizzato per creare la vista non viene analizzato. Ciò può causare un caso in cui la vista viene creata, ma le query non hanno esito positivo. Assicuratevi di testare la sintassi SQL prima di creare la vista.

Utilizzate il seguente comando CLI per creare una vista:

```
aws glue create-table --cli-input-json '{
  "DatabaseName": "database",
  "TableInput": {
    "Name": "view",
    "StorageDescriptor": {
      "Columns": [
        {
          "Name": "col1",
          "Type": "data-type"
        },
        ...
        {
          "Name": "col_n",
          "Type": "data-type"
        }
      ],
      "SerdeInfo": {}
    },
    "ViewDefinition": {
      "SubObjects": [
        "arn:aws:glue:aws-region:aws-account-id:table/database/referenced-table1",
        ...
        "arn:aws:glue:aws-region:aws-account-id:table/database/referenced-tableN",
      ],
      "IsProtected": true,
      "Representations": [
        {
          "Dialect": "SPARK",
          "DialectVersion": "1.0",
          "ViewOriginalText": "Spark-SQL",
          "ViewExpandedText": "Spark-SQL"
        }
      ]
    }
  }
}
```

```

    }
  ]
}
}'

```

## Operazioni di visualizzazione supportate

I seguenti frammenti di comandi mostrano vari modi di lavorare con le viste del catalogo dati:

- **CREA VISTA**

Crea una vista del catalogo di dati. Di seguito è riportato un esempio che mostra la creazione di una vista da una tabella esistente:

```

CREATE PROTECTED MULTI DIALECT VIEW catalog_view
SECURITY DEFINER AS SELECT * FROM my_catalog.my_database.source_table

```

- **MODIFICA VISUALIZZAZIONE**

Sintassi disponibile:

- ALTER VIEW view\_name [FORCE] ADD DIALECT AS query
- ALTER VIEW view\_name [FORCE] UPDATE DIALECT AS query
- ALTER VIEW view\_name DROP DIALECT

È possibile utilizzare l'FORCE ADD DIALECT opzione per forzare l'aggiornamento dello schema e degli oggetti secondari secondo il nuovo dialetto del motore. Tieni presente che questa operazione può causare errori di query se non lo utilizzi anche FORCE per aggiornare altri dialetti del motore. Di seguito viene mostrato un esempio:

```

ALTER VIEW catalog_view FORCE ADD DIALECT
AS
SELECT order_date, sum(totalprice) AS price
FROM source_table
GROUP BY orderdate;

```

Quanto segue mostra come modificare una vista per aggiornare il dialetto:

```

ALTER VIEW catalog_view UPDATE DIALECT AS

```

```
SELECT count(*) FROM my_catalog.my_database.source_table;
```

- DESCRIVI LA VISTA

Sintassi disponibile per descrivere una vista:

- `SHOW COLUMNS {FROM|IN} view_name [{FROM|IN} database_name]`— Se l'utente dispone delle autorizzazioni AWS Glue and Lake Formation necessarie per descrivere la vista, può elencare le colonne. Di seguito sono riportati un paio di comandi di esempio per la visualizzazione delle colonne:

```
SHOW COLUMNS FROM my_database.source_table;
SHOW COLUMNS IN my_database.source_table;
```

- `DESCRIBE view_name`— Se l'utente dispone delle autorizzazioni AWS Glue and Lake Formation necessarie per descrivere la vista, può elencare le colonne della vista insieme ai relativi metadati.
- ELIMINA VISTA

Sintassi disponibile:

- `DROP VIEW [ IF EXISTS ] view_name`

L'esempio seguente mostra un'istruzione che verifica l'esistenza di una vista prima di eliminarla:

```
DROP VIEW IF EXISTS catalog_view;
```

- MOSTRA CREA VISTA

- `SHOW CREATE VIEW view_name`— Mostra l'istruzione SQL che crea la vista specificata. Di seguito è riportato un esempio che mostra la creazione di una vista del catalogo di dati:

```
SHOW CREATE TABLE my_database.catalog_view;
CREATE PROTECTED MULTI DIALECT VIEW my_catalog.my_database.catalog_view (
  net_profit,
  customer_id,
  item_id,
  sold_date)
TBLPROPERTIES (
  'transient_lastDdlTime' = '1736267222')
SECURITY DEFINER AS SELECT * FROM
```

```
my_database.store_sales_partitioned_lf WHERE customer_id IN (SELECT customer_id
from source_table limit 10)
```

- MOSTRA VISUALIZZAZIONI

Elenca tutte le viste del catalogo, ad esempio viste regolari, visualizzazioni multidialettali (MDV) e MDV senza dialetto Spark. La sintassi disponibile è la seguente:

- `SHOW VIEWS [{ FROM | IN } database_name] [LIKE regex_pattern]:`

Di seguito viene illustrato un comando di esempio per mostrare le viste:

```
SHOW VIEWS IN marketing_analytics LIKE 'catalog_view*';
```

Per ulteriori informazioni sulla creazione e la configurazione delle viste del catalogo dati, consulta Building [AWS Glue Data Catalog views](#) nella Developer Guide. AWS Lake Formation

### Interrogazione di una vista di Catalogo Dati

Dopo aver creato una vista del catalogo dati, puoi interrogarla utilizzando un job Amazon EMR Serverless Spark con controllo granulare degli accessi abilitato AWS Lake Formation . Il ruolo di job runtime deve disporre dell'autorizzazione Lake Formation per la visualizzazione Data Catalog. Non è necessario concedere l'accesso alle tabelle sottostanti a cui si fa riferimento nella vista.

Dopo aver impostato tutto, puoi interrogare la tua vista. Ad esempio, dopo aver creato un'applicazione EMR Serverless in EMR Studio, è possibile eseguire la seguente query per accedere a una vista.

```
SELECT * from my_database.catalog_view LIMIT 10;
```

Una funzione utile è la `invoker_principal` Restituisce l'identificatore univoco del ruolo EMRS Job Runtime. Questo può essere usato per controllare l'output della vista, in base al principio di invocazione. Puoi usarlo per aggiungere una condizione nella tua vista che perfeziona i risultati della query, in base al ruolo chiamante. Il ruolo di job runtime deve disporre dell'autorizzazione all'azione `LakeFormation:GetDataLakePrincipal` IAM per utilizzare questa funzione.

```
select invoker_principal();
```

È possibile aggiungere questa funzione a una WHERE clausola, ad esempio, per perfezionare i risultati delle query.

## Considerazioni e limitazioni

Quando si creano viste del catalogo dati, si applica quanto segue:

- Puoi creare viste del catalogo dati solo con Amazon EMR 7.6 e versioni successive.
- Il defintore della vista del catalogo dati deve avere SELECT accesso alle tabelle di base sottostanti a cui si accede dalla vista. La creazione della vista Data Catalog non riesce se una tabella base specifica ha dei filtri Lake Formation imposti sul ruolo di definizione.
- Le tabelle di base non devono avere l'autorizzazione del IAMAllowedPrincipals data lake in Lake Formation. Se presente, si verifica l'errore Multi Dialect views può fare riferimento solo a tabelle senza le autorizzazioni IAMAllowed Principal.
- La posizione Amazon S3 della tabella deve essere registrata come posizione del data lake Lake Formation. Se la tabella non è registrata, si verifica l'errore Le viste multidialettali possono fare riferimento solo alle tabelle gestite da Lake Formation. Per informazioni su come registrare le sedi Amazon S3 in Lake Formation, consulta [Registrazione di una sede Amazon S3](#) nella Developer Guide. AWS Lake Formation
- Puoi creare solo viste di PROTECTED Data Catalog. UNPROTECTED le visualizzazioni non sono supportate.
- Non è possibile fare riferimento alle tabelle di un altro AWS account in una definizione di visualizzazione del catalogo dati. Inoltre, non puoi fare riferimento a una tabella nello stesso account che si trova in una regione separata.
- Per condividere i dati tra un account o un'area geografica, l'intera visualizzazione deve essere condivisa tra account e regioni, utilizzando i link alle risorse di Lake Formation.
- Le funzioni definite dall'utente (UDFs) non sono supportate.
- È possibile utilizzare viste basate sulle tabelle Iceberg. Sono supportati anche i formati a tabella aperta Apache Hudi e Delta Lake.
- Non puoi fare riferimento ad altre viste nelle viste del Catalogo dati.
- Lo schema di visualizzazione di AWS Glue Data Catalog viene sempre memorizzato in lettere minuscole. Ad esempio, se si utilizza un'istruzione DDL per creare una vista del Glue Data Catalog con una colonna denominataCastle, la colonna creata nel Glue Data Catalog verrà composta in minuscolo, to. castle Se poi si specifica il nome della colonna in una query DML come Castle oCASTLE, EMR Spark renderà il nome in minuscolo per eseguire la query. Tuttavia, l'intestazione della colonna viene visualizzata utilizzando il maiuscolo/minuscolo specificato nella query.

Se si desidera che una query abbia esito negativo nel caso in cui il nome di colonna specificato nella query DML non corrisponda al nome della colonna nel Glue Data Catalog, è possibile impostare `spark.sql.caseSensitive=true`.

## Considerazioni e limitazioni

Considerate le seguenti considerazioni e limitazioni quando utilizzate Lake Formation con EMR Serverless.

### Note

Quando abiliti Lake Formation per un job Spark su EMR Serverless, il job avvia un driver di sistema e un driver utente. Se hai specificato la capacità preinizializzata all'avvio, i driver forniti dalla capacità preinizializzata e il numero di driver di sistema è uguale al numero di driver utente specificato. Se si sceglie la capacità On Demand, EMR Serverless avvia un driver di sistema oltre a un driver utente. Per stimare i costi associati al tuo lavoro EMR Serverless with Lake Formation, utilizza il [Calcolatore dei prezzi AWS](#)

[Amazon EMR Serverless with Lake Formation è disponibile in tutte le regioni serverless EMR supportate.](#)

- Amazon EMR Serverless supporta il controllo granulare degli accessi tramite Lake Formation solo per le tabelle Apache Hive e Apache Iceberg. I formati Apache Hive includono Parquet, ORC e XSV.
- Le applicazioni abilitate per Lake Formation non supportano l'utilizzo di immagini [EMR Serverless](#) personalizzate.
- Non puoi smettere di lavorare `DynamicResourceAllocation` per Lake Formation.
- Puoi usare Lake Formation solo con i job Spark.
- EMR Serverless with Lake Formation supporta solo una singola sessione Spark durante un job.
- EMR Serverless with Lake Formation supporta solo le query tabellari tra account condivise tramite link alle risorse.
- Quanto segue non è supportato:
  - Set di dati distribuiti resilienti (RDD)
  - Streaming Spark

- Scrivi con le autorizzazioni concesse da Lake Formation
- Controllo degli accessi per le colonne annidate
- EMR Serverless blocca le funzionalità che potrebbero compromettere il completo isolamento dei driver di sistema, tra cui:
  - UDTs, Hive e qualsiasi funzione UDFs definita dall'utente che coinvolga classi personalizzate
  - Origini dati personalizzate
  - Fornitura di vasetti aggiuntivi per l'estensione, il connettore o il metastore Spark
  - Comando `ANALYZE TABLE`
- Per applicare i controlli di accesso `EXPLAIN PLAN` e le operazioni DDL, ad esempio non esporre informazioni riservate `DESCRIBE TABLE`.
- EMR Serverless limita l'accesso ai registri Spark dei driver di sistema sulle applicazioni abilitate per Lake Formation. Poiché il driver di sistema viene eseguito con autorizzazioni elevate, gli eventi e i registri generati dal driver di sistema possono includere informazioni riservate. Per impedire a utenti o codici non autorizzati di accedere a questi dati sensibili, EMR Serverless disabilita l'accesso ai registri dei driver di sistema.

I log dei profili di sistema vengono sempre conservati nello storage gestito: si tratta di un'impostazione obbligatoria che non può essere disabilitata. Questi registri vengono archiviati in modo sicuro e crittografati utilizzando una chiave KMS gestita dal cliente o una chiave KMS gestita. AWS

[Se la tua applicazione EMR Serverless si trova in una sottorete privata con endpoint VPC per Amazon S3 e alleggi una policy di endpoint per controllare l'accesso, prima che i tuoi job possano inviare i dati di log a AWS Managed Amazon S3, devi includere le autorizzazioni dettagliate in Managed storage nella tua policy VPC all'endpoint gateway S3.](#) Per la risoluzione dei problemi relativi alle richieste, contatta l'assistenza. AWS

- Se hai registrato una posizione in una tabella con Lake Formation, il percorso di accesso ai dati passa attraverso le credenziali archiviate di Lake Formation indipendentemente dall'autorizzazione IAM per il ruolo di job runtime EMR Serverless. Se configuri erroneamente il ruolo registrato con la posizione della tabella, i lavori inviati che utilizzano il ruolo con l'autorizzazione S3 IAM per la posizione della tabella avranno esito negativo.
- La scrittura su una tabella Lake Formation utilizza l'autorizzazione IAM anziché le autorizzazioni concesse da Lake Formation. Se il tuo ruolo di job runtime dispone delle autorizzazioni S3 necessarie, puoi utilizzarlo per eseguire operazioni di scrittura.

Di seguito sono riportate considerazioni e limitazioni relative all'utilizzo di Apache Iceberg:

- È possibile utilizzare Apache Iceberg solo con il catalogo delle sessioni e non con i cataloghi con nomi arbitrari.
- Le tabelle Iceberg registrate in Lake Formation supportano solo le tabelle di `metadati`, `history`, `metadata_log_entries`, `snapshots`, `files`, `manifests`, e `refs`. Amazon EMR nasconde le colonne che potrebbero contenere dati sensibili, ad esempio `partitions`, `path` e `summaries`. Questa limitazione non si applica alle tabelle Iceberg che non sono registrate in Lake Formation.
- Le tabelle che non vengono registrate in Lake Formation supportano tutte le stored procedure Iceberg. Le migrate procedure `register_table` and non sono supportate per nessuna tabella.
- Ti consigliamo di utilizzare Iceberg `DataFrameWriter V2` anziché `V1`.

## Risoluzione dei problemi

Consulta le seguenti sezioni per le soluzioni di risoluzione dei problemi.

### Registrazione

EMR Serverless utilizza i profili di risorse Spark per suddividere l'esecuzione del lavoro. EMR Serverless utilizza il profilo utente per eseguire il codice fornito, mentre il profilo di sistema applica le politiche di Lake Formation. È possibile accedere ai registri delle attività eseguite come profilo utente.

[Per ulteriori informazioni sul debug dei job abilitati per Lake Formation, consulta \*Debugging jobs\*.](#)

### Interfaccia utente live e Spark History Server

L'interfaccia utente Live e lo Spark History Server contengono tutti gli eventi Spark generati dal profilo utente e gli eventi oscurati generati dal driver di sistema.

Puoi vedere tutte le attività dei driver utente e di sistema nella scheda `Executors`. Tuttavia, i link di registro sono disponibili solo per il profilo utente. Inoltre, alcune informazioni vengono cancellate dall'interfaccia utente Live, ad esempio il numero di record di output.

### Job non riuscito con permessi insufficienti per Lake Formation

Assicurati che il tuo ruolo di job runtime disponga delle autorizzazioni per eseguire `SELECT` e `DESCRIBE` sulla tabella a cui stai accedendo.

## Job con esecuzione RDD non riuscita

EMR Serverless attualmente non supporta operazioni RDD (Resilient Distributed Dataset) sui job abilitati per Lake Formation.

## Impossibile accedere ai file di dati in Amazon S3

Assicurati di aver registrato la posizione del data lake in Lake Formation.

## Eccezione di convalida della sicurezza

EMR Serverless ha rilevato un errore di convalida della sicurezza. Contatta l'AWS assistenza per ricevere assistenza.

## Condivisione del catalogo dati e delle tabelle di AWS Glue tra account

Puoi condividere database e tabelle tra account e continuare a utilizzare Lake Formation. Per ulteriori informazioni, consulta [Condivisione dei dati tra account in Lake Formation](#) e [Come posso condividere AWS Glue Data Catalog e tabelle utilizzando più account?](#) AWS Lake Formation.

## Crittografia tra lavoratori

Con le versioni 6.15.0 e successive di Amazon EMR, puoi abilitare la comunicazione crittografata Mutual-TLS tra i lavoratori durante le esecuzioni dei job Spark. Se abilitato, EMR Serverless genera e distribuisce automaticamente un certificato univoco per ogni lavoratore assegnato nell'ambito delle esecuzioni dei job. Quando questi lavoratori comunicano per scambiare messaggi di controllo o trasferire dati shuffle, stabiliscono una connessione TLS reciproca e utilizzano i certificati configurati per verificare l'identità reciproca. Se un lavoratore non è in grado di verificare un altro certificato, l'handshake TLS fallisce e EMR Serverless interrompe la connessione tra i due certificati.

Se utilizzi Lake Formation con EMR Serverless, la crittografia Mutual-TLS è abilitata per impostazione predefinita.

## Abilitazione della crittografia Mutual-TLS su EMR Serverless

Per abilitare la crittografia TLS reciproca sulla tua applicazione Spark, imposta su `true` durante la `spark.ssl.internode.enabled` creazione dell'applicazione [EMR Serverless](#). Se utilizzi la AWS console per creare un'applicazione EMR Serverless, scegli Usa impostazioni personalizzate, quindi espandi Configurazione dell'applicazione e inserisci le tue `runtimeConfiguration`

```
aws emr-serverless create-application \  
--release-label emr-6.15.0 \  
--runtime-configuration '{  
  "classification": "spark-defaults",  
  "properties": {"spark.ssl.internode.enabled": "true"}  
}' \  
--type "SPARK"
```

Se desideri abilitare la crittografia TLS reciproca per le singole esecuzioni di job Spark, imposta su `true` durante `spark.ssl.internode.enabled` l'utilizzo. `spark-submit`

```
--conf spark.ssl.internode.enabled=true
```

## Secrets Manager per la protezione dei dati con EMR Serverless

AWS Secrets Manager è un servizio di archiviazione segreto che puoi utilizzare per proteggere le credenziali del database, le chiavi API e altre informazioni segrete. Quindi, nel codice, puoi sostituire le credenziali hardcoded con una chiamata API a Secrets Manager. Questo aiuta a garantire che il segreto non possa essere compromesso da qualcuno che esamina il codice, perché il segreto non è presente. Per una panoramica, consulta la [Guida per l'utente di AWS Secrets Manager](#).

Secrets Manager crittografa i segreti utilizzando AWS Key Management Service le chiavi. Per ulteriori informazioni, vedere [Crittografia e decrittografia segrete](#) nella Guida per l'AWS Secrets Manager utente.

È possibile configurare Secrets Manager in modo che ruoti automaticamente i segreti in base a una pianificazione specificata. In questo modo puoi sostituire i segreti a lungo termine con altri a breve termine, contribuendo a ridurre notevolmente il rischio di compromissione. Per ulteriori informazioni, consulta [Ruotare AWS Secrets Manager i segreti nella Guida](#) per l'AWS Secrets Manager utente.

Amazon EMR Serverless si integra AWS Secrets Manager in modo da poter archiviare i dati in Secrets Manager e utilizzare l'ID segreto nelle configurazioni.

## In che modo EMR Serverless utilizza i segreti

Quando si archiviano i dati in Secrets Manager e si utilizza l'ID segreto nelle configurazioni per EMR Serverless, non si trasmettono dati di configurazione sensibili a EMR Serverless in testo semplice per poi esporli a fonti esterne. APIs Se si indica che una coppia chiave-valore contiene un ID segreto per

un segreto archiviato in Secrets Manager, EMR Serverless recupera il segreto quando invia i dati di configurazione ai worker per i job in esecuzione.

Per indicare che una coppia chiave-valore per una configurazione contiene un riferimento a un segreto archiviato in Secrets Manager, aggiungete l'`EMR.secret@`annotazione al valore di configurazione. Per qualsiasi proprietà di configurazione con annotazione ID segreta, EMR Serverless chiama Secrets Manager e risolve il segreto al momento dell'esecuzione del job.

## Come creare un segreto

Per creare un segreto, segui i passaggi in [Creare un AWS Secrets Manager segreto](#) nella Guida per l'AWS Secrets Manager utente. Nel passaggio 3, scegli il campo Testo normale per inserire il tuo valore sensibile.

## Fornisci un segreto in una classificazione di configurazione

Gli esempi seguenti mostrano come fornire un segreto in una classificazione di configurazione in `StartJobRun`. Se si desidera configurare le classificazioni per Secrets Manager a livello di applicazione, vedere [Configurazione predefinita dell'applicazione per EMR Serverless](#).

Negli esempi, *SecretName* sostituisilo con il nome del segreto da recuperare. Includi il trattino seguito dai sei caratteri che Secrets Manager aggiunge alla fine dell'ARN segreto. Per ulteriori informazioni, consulta [Come creare un segreto](#).

In questa sezione

- [Specificate i riferimenti segreti - Spark](#)
- [Specificare i riferimenti segreti - Hive](#)

### Specificate i riferimenti segreti - Spark

Example — Specificate i riferimenti segreti nella configurazione del metastore Hive esterno per Spark

```
aws emr-serverless start-job-run \  
  --application-id "application-id" \  
  --execution-role-arn "job-role-arn" \  
  --job-driver '{  
    "sparkSubmit": {  
      "entryPoint": "s3://amzn-s3-demo-bucket/scripts/spark-jdbc.py",  
      "sparkSubmitParameters": "--jars s3://amzn-s3-demo-bucket/mariadb-  
connector-java.jar
```

```

        --conf
spark.hadoop.javax.jdo.option.ConnectionDriverName=org.mariadb.jdbc.Driver
        --conf spark.hadoop.javax.jdo.option.ConnectionUserName=connection-user-
name
        --conf
spark.hadoop.javax.jdo.option.ConnectionPassword=EMR.secret@SecretName
        --conf spark.hadoop.javax.jdo.option.ConnectionURL=jdbc:mysql://db-host:db-
port/db-name
        --conf spark.driver.cores=2
        --conf spark.executor.memory=10G
        --conf spark.driver.memory=6G
        --conf spark.executor.cores=4"
    }
}' \
--configuration-overrides '{
    "monitoringConfiguration": {
    "s3MonitoringConfiguration": {
        "logUri": "s3://amzn-s3-demo-bucket/spark/logs/"
    }
}
}'
}'

```

Example — Specificate i riferimenti segreti per la configurazione del metastore Hive esterno nella classificazione **spark-defaults**

```

{
    "classification": "spark-defaults",
    "properties": {
        "spark.hadoop.javax.jdo.option.ConnectionDriverName": "org.mariadb.jdbc.Driver"
        "spark.hadoop.javax.jdo.option.ConnectionURL": "jdbc:mysql://db-host:db-
port/db-name"
        "spark.hadoop.javax.jdo.option.ConnectionUserName": "connection-user-name"
        "spark.hadoop.javax.jdo.option.ConnectionPassword":
        "EMR.secret@SecretName",
    }
}

```

## Specificare i riferimenti segreti - Hive

Example — Specificate i riferimenti segreti nella configurazione del metastore Hive esterno per Hive

```
aws emr-serverless start-job-run \
```

```

--application-id "application-id" \
--execution-role-arn "job-role-arn" \
  --job-driver '{
    "hive": {
      "query": "s3://amzn-s3-demo-bucket/emr-serverless-hive/query/hive-query.q1",
      "parameters": "--hiveconf hive.exec.scratchdir=s3://amzn-s3-demo-bucket/emr-
serverless-hive/hive/scratch
                    --hiveconf hive.metastore.warehouse.dir=s3://amzn-s3-demo-bucket/
emr-serverless-hive/hive/warehouse
                    --hiveconf javax.jdo.option.ConnectionUserName=username
                    --hiveconf
javax.jdo.option.ConnectionPassword=EMR.secret@SecretName
                    --hiveconf
hive.metastore.client.factory.class=org.apache.hadoop.hive.ql.metadata.SessionHiveMetaStoreCli
                    --hiveconf
javax.jdo.option.ConnectionDriverName=org.mariadb.jdbc.Driver
                    --hiveconf javax.jdo.option.ConnectionURL=jdbc:mysql://db-host:db-
port/db-name"
    }
  }' \
  --configuration-overrides '{
    "monitoringConfiguration": {
      "s3MonitoringConfiguration": {
        "logUri": "s3://amzn-s3-demo-bucket"
      }
    }
  }'

```

Example — Specificare i riferimenti segreti per la configurazione del metastore Hive esterno nella classificazione **hive-site**

```

{
  "classification": "hive-site",
  "properties": {
    "hive.metastore.client.factory.class":
"org.apache.hadoop.hive.ql.metadata.SessionHiveMetaStoreClientFactory",
    "javax.jdo.option.ConnectionDriverName": "org.mariadb.jdbc.Driver",
    "javax.jdo.option.ConnectionURL": "jdbc:mysql://db-host:db-port/db-name",
    "javax.jdo.option.ConnectionUserName": "username",
    "javax.jdo.option.ConnectionPassword": "EMR.secret@SecretName"
  }
}

```

## Concedi l'accesso a EMR Serverless per recuperare il segreto

Per consentire a EMR Serverless di recuperare il valore segreto da Secrets Manager, aggiungi la seguente dichiarazione di policy al tuo segreto quando lo crei. È necessario creare il segreto con la chiave KMS gestita dal cliente per consentire a EMR Serverless di leggere il valore segreto. Per ulteriori informazioni, consulta [Autorizzazioni per la chiave KMS](#) nella Guida per l'utente di AWS Secrets Manager .

Nella seguente politica, sostituiscilo *applicationId* con l'ID della tua applicazione.

### Politica delle risorse per il segreto

È necessario includere le seguenti autorizzazioni nella politica delle risorse per il secret in per consentire AWS Secrets Manager a EMR Serverless di recuperare i valori segreti. Per garantire che solo un'applicazione specifica possa recuperare questo segreto, è possibile specificare facoltativamente l'ID dell'applicazione EMR Serverless come condizione nella policy.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "secretsmanager:GetSecretValue",
        "secretsmanager:DescribeSecret"
      ],
      "Principal": {
        "Service": [
          "emr-serverless.amazonaws.com"
        ]
      },
      "Resource": [
        "*"
      ],
      "Condition": {
        "StringEquals": {
          "aws:SourceArn": "arn:aws:emr-serverless:Regione AWS:aws_account_id:/
applications/applicationId"
        }
      }
    }
  ]
}
```

```
}
```

Crea il tuo segreto con la seguente politica per la chiave gestita dal cliente AWS Key Management Service ():AWS KMS

Politica per la chiave gestita dal cliente AWS KMS

```
{
  "Sid": "Allow EMR Serverless to use the key for decrypting secrets",
  "Effect": "Allow",
  "Principal": {
    "Service": [
      "emr-serverless.amazonaws.com"
    ]
  },
  "Action": [
    "kms:Decrypt",
    "kms:DescribeKey"
  ],
  "Resource": "*",
  "Condition": {
    "StringEquals": {
      "kms:ViaService": "secretsmanager.Regione AWS.amazonaws.com"
    }
  }
}
```

## Ruotare il segreto

La rotazione avviene quando si aggiorna periodicamente un segreto. Puoi AWS Secrets Manager configurare la rotazione automatica del segreto secondo una pianificazione da te specificata. In questo modo, puoi sostituire i segreti a lungo termine con segreti a breve termine. Questo aiuta a ridurre il rischio di compromessi. EMR Serverless recupera il valore segreto da una configurazione annotata quando il processo passa allo stato di esecuzione. Se tu o un processo aggiorni il valore segreto in Secrets Manager, devi inviare un nuovo lavoro in modo che il lavoro possa recuperare il valore aggiornato.

**Note**

I lavori che sono già in esecuzione non possono recuperare un valore segreto aggiornato. Ciò potrebbe causare un fallimento del lavoro.

## Utilizzo di Amazon S3 Access Grants con EMR Serverless

### Panoramica di S3 Access Grants per EMR Serverless

Con le versioni 6.15.0 e successive di Amazon EMR, Amazon S3 Access Grants fornisce una soluzione di controllo degli accessi scalabile che puoi utilizzare per aumentare l'accesso ai tuoi dati Amazon S3 da EMR Serverless. Se hai una configurazione di autorizzazioni complessa o di grandi dimensioni per i dati S3, puoi utilizzare Access Grants per dimensionare le autorizzazioni relative ai dati S3 per utenti, ruoli e applicazioni.

Usa S3 Access Grants per aumentare l'accesso ai dati di Amazon S3 oltre alle autorizzazioni concesse dal ruolo di runtime o ai ruoli IAM collegati alle identità con accesso alla tua applicazione EMR Serverless.

Per ulteriori informazioni, consulta [Gestione dell'accesso con S3 Access Grants per Amazon EMR](#) nella Guida alla gestione di Amazon EMR e [Gestione dell'accesso con S3 Access Grants](#) nella Guida per l'utente di Amazon Simple Storage Service.

Questa sezione descrive come avviare un'applicazione EMR Serverless che utilizza S3 Access Grants per fornire l'accesso ai dati in Amazon S3. Per conoscere i passaggi per utilizzare S3 Access Grants con altre implementazioni Amazon EMR, consulta la seguente documentazione:

- [Utilizzo di S3 Access Grants con Amazon EMR](#)
- [Utilizzo di S3 Access Grants con Amazon EMR su EKS](#)

### Avvia un'applicazione EMR Serverless con S3 Access Grants per la gestione dei dati

Puoi abilitare S3 Access Grants su EMR Serverless e avviare un'applicazione Spark. Quando l'applicazione effettua una richiesta di dati S3, Amazon S3 fornisce credenziali temporanee che rientrano nell'ambito del bucket, del prefisso o dell'oggetto specifico.

1. Imposta un ruolo di esecuzione del lavoro per la tua applicazione EMR Serverless. Includi le autorizzazioni IAM necessarie per eseguire i job Spark e utilizzare S3 Access Grants e: `s3:GetDataAccess` `s3:GetAccessGrantsInstanceForPrefix`

```
{
  "Effect": "Allow",
  "Action": [
    "s3:GetDataAccess",
    "s3:GetAccessGrantsInstanceForPrefix"
  ],
  "Resource": [
    //LIST ALL INSTANCE ARNS THAT THE ROLE IS ALLOWED TO QUERY
    "arn:aws_partition:s3:Region:account-id1:access-grants/default",
    "arn:aws_partition:s3:Region:account-id2:access-grants/default"
  ]
}
```

#### Note

Se specifichi ruoli IAM per l'esecuzione del lavoro che dispongono di autorizzazioni aggiuntive per accedere direttamente a S3, gli utenti saranno in grado di accedere ai dati consentiti dal ruolo anche se non dispongono dell'autorizzazione di S3 Access Grants.

2. Avvia la tua applicazione EMR Serverless con un'etichetta di release Amazon EMR 6.15.0 o successiva e la `spark-defaults` classificazione, come illustrato nell'esempio seguente. Sostituisci i valori in *red text* con i valori appropriati per il tuo scenario di utilizzo.

```
aws emr-serverless start-job-run \
  --application-id application-id \
  --execution-role-arn job-role-arn \
  --job-driver '{
    "sparkSubmit": {
      "entryPoint": "s3://us-east-1.elasticmapreduce/emr-containers/samples/
wordcount/scripts/wordcount.py",
      "entryPointArguments": ["s3://amzn-s3-demo-destination-bucket1/
wordcount_output"],
      "sparkSubmitParameters": "--conf spark.executor.cores=1 --conf
spark.executor.memory=4g --conf spark.driver.cores=1 --conf spark.driver.memory=4g
--conf spark.executor.instances=1"
    }
  }' \
  --configuration-overrides '{
```

```
"applicationConfiguration": [{  
  "classification": "spark-defaults",  
  "properties": {  
    "spark.hadoop.fs.s3.s3AccessGrants.enabled": "true",  
    "spark.hadoop.fs.s3.s3AccessGrants.fallbackToIAM": "false"  
  }  
}]  
'
```

## S3 Access concede considerazioni con EMR Serverless

Per informazioni importanti su supporto, compatibilità e comportamento quando usi Amazon S3 Access Grants con EMR Serverless, [consulta le considerazioni su S3 Access Grants con Amazon EMR nella Amazon EMR Management Guide](#).

## Registrazione delle chiamate API Serverless di Amazon EMR utilizzando AWS CloudTrail

Amazon EMR Serverless è integrato con AWS CloudTrail un servizio che fornisce un registro delle azioni intraprese da un utente, ruolo o servizio AWS in EMR Serverless. CloudTrail acquisisce tutte le chiamate API per EMR Serverless come eventi. Le chiamate acquisite includono chiamate dalla console EMR Serverless e chiamate in codice alle operazioni dell'API EMR Serverless. Se crei un trail, puoi abilitare la distribuzione continua di CloudTrail eventi a un bucket Amazon S3, inclusi gli eventi per EMR Serverless. Se non configuri un percorso, puoi comunque visualizzare gli eventi più recenti nella CloudTrail console in Cronologia eventi. Utilizzando le informazioni raccolte da CloudTrail, è possibile determinare la richiesta effettuata a EMR Serverless, l'indirizzo IP da cui è stata effettuata la richiesta, chi ha effettuato la richiesta, quando è stata effettuata e dettagli aggiuntivi.

Per ulteriori informazioni CloudTrail, consulta la Guida per l'[AWS CloudTrail utente](#).

## Informazioni EMR Serverless in CloudTrail

CloudTrail è abilitato sul tuo Account AWS quando crei l'account. Quando si verifica un'attività in EMR Serverless, tale attività viene registrata in un CloudTrail evento insieme ad altri eventi di AWS servizio nella cronologia degli eventi. Puoi visualizzare, cercare e scaricare gli eventi recenti in Account AWS. Per ulteriori informazioni, vedere [Visualizzazione degli eventi con CloudTrail la cronologia degli eventi](#).

Per una registrazione continua degli eventi nel tuo Account AWS, compresi gli eventi per EMR Serverless, crea un percorso. Un trail consente di CloudTrail inviare file di log a un bucket Amazon S3. Per impostazione predefinita, quando si crea un percorso nella console, questo sarà valido in tutte le Regioni AWS. Il trail registra gli eventi di tutte le regioni della AWS partizione e consegna i file di log al bucket Amazon S3 specificato. Inoltre, puoi configurare altri AWS servizi per analizzare ulteriormente e agire in base ai dati sugli eventi raccolti nei log. CloudTrail Per ulteriori informazioni, consulta gli argomenti seguenti:

- [Panoramica della creazione di un percorso](#)
- [CloudTrail servizi e integrazioni supportati](#)
- [Configurazione delle notifiche Amazon SNS per CloudTrail](#)
- [Ricezione di file di CloudTrail registro da più regioni](#) e [ricezione di file di CloudTrail registro da più account](#)

Tutte le azioni EMR Serverless vengono registrate CloudTrail e documentate nell'[EMR](#) Serverless API Reference. Ad esempio, le chiamate a `StartJobRun` e le `CancelJobRun` azioni generano `CreateApplication` voci nei file di registro. CloudTrail

Ogni evento o voce di log contiene informazioni sull'utente che ha generato la richiesta. Le informazioni di identità consentono di determinare quanto segue:

- Se la richiesta è stata effettuata con credenziali utente root o AWS Identity and Access Management (IAM).
- Se la richiesta è stata effettuata con le credenziali di sicurezza temporanee per un ruolo o un utente federato.
- Se la richiesta è stata effettuata da un altro AWS servizio.

Per ulteriori informazioni, consulta [Elemento CloudTrail userIdentity](#).

## Informazioni sulle voci dei file di registro EMR Serverless

Un trail è una configurazione che consente la distribuzione di eventi come file di log in un bucket Amazon S3 specificato dall'utente. CloudTrail i file di registro contengono una o più voci di registro. Un evento rappresenta una singola richiesta proveniente da qualsiasi fonte e include informazioni sull'azione richiesta, la data e l'ora dell'azione, i parametri della richiesta e così via. CloudTrail i file di registro non sono una traccia ordinata dello stack delle chiamate API pubbliche, quindi non vengono visualizzati in un ordine specifico.

L'esempio seguente mostra una voce di CloudTrail registro che illustra l'CreateApplicationazione.

```
{
  "eventVersion": "1.08",
  "userIdentity": {
    "type": "AssumedRole",
    "principalId": "AIDACKCEVSQ6C2EXAMPLE:admin",
    "arn": "arn:aws:sts::012345678910:assumed-role/Admin/admin",
    "accountId": "012345678910",
    "accessKeyId": "AKIAIOSFODNN7EXAMPLE",
    "sessionContext": {
      "sessionIssuer": {
        "type": "Role",
        "principalId": "AIDACKCEVSQ6C2EXAMPLE",
        "arn": "arn:aws:iam::012345678910:role/Admin",
        "accountId": "012345678910",
        "userName": "Admin"
      },
      "webIdFederationData": {},
      "attributes": {
        "creationDate": "2022-06-01T23:46:52Z",
        "mfaAuthenticated": "false"
      }
    }
  },
  "eventTime": "2022-06-01T23:49:28Z",
  "eventSource": "emr-serverless.amazonaws.com",
  "eventName": "CreateApplication",
  "awsRegion": "us-west-2",
  "sourceIPAddress": "203.0.113.0",
  "userAgent": "PostmanRuntime/7.26.10",
  "requestParameters": {
    "name": "my-serverless-application",
    "releaseLabel": "emr-6.6",
    "type": "SPARK",
    "clientToken": "0a1b234c-de56-7890-1234-567890123456"
  },
  "responseElements": {
    "name": "my-serverless-application",
    "applicationId": "1234567890abcdef0",
    "arn": "arn:aws:emr-serverless:us-west-2:555555555555:/
applications/1234567890abcdef0"
  }
}
```

```
  },
  "requestID": "890b8639-e51f-11e7-b038-EXAMPLE",
  "eventID": "874f89fa-70fc-4798-bc00-EXAMPLE",
  "readOnly": false,
  "eventType": "AwsApiCall",
  "managementEvent": true,
  "recipientAccountId": "012345678910",
  "eventCategory": "Management"
}
```

## Convalida della conformità per Amazon EMR Serverless

La sicurezza e la conformità di EMR Serverless vengono valutate da revisori di terze parti nell'ambito di più programmi di AWS conformità, tra cui:

- System and Organization Controls (SOC)
- Payment Card Industry Data Security Standard (PCI DSS)
- Programma federale di gestione dei rischi e delle autorizzazioni (FedRAMP) Moderato
- Health Insurance Portability and Accountability Act (HIPAA)

AWS fornisce un elenco frequentemente aggiornato di AWS servizi nell'ambito di specifici programmi di conformità nella pagina [AWS Services in Scope by Compliance Program](#).

I report di audit di terze parti possono essere scaricati utilizzando AWS Artifact. Per ulteriori informazioni, consulta [Scaricamento dei report in AWS Artifact](#).

Per ulteriori informazioni sui programmi di AWS conformità, consulta Programmi di [AWS conformità](#).

La responsabilità della conformità quando si utilizza EMR Serverless è determinata dalla sensibilità dei dati, dagli obiettivi di conformità dell'organizzazione e dalle leggi e dai regolamenti applicabili. Se l'uso di EMR Serverless è soggetto alla conformità a standard come HIPAA, PCI o FedRAMP Moderate, fornisce risorse per aiutare a: AWS

- [Guide introduttive su sicurezza e conformità che illustrano le](#) considerazioni sull'architettura e i passaggi per implementare ambienti di base incentrati sulla sicurezza e la conformità su AWS
- [AWS Le guide alla conformità dei clienti](#) possono aiutarti a comprendere il modello di responsabilità condivisa attraverso la lente della conformità. Le guide riassumono le migliori pratiche per la protezione Servizi AWS e mappano le linee guida per i controlli di sicurezza su più framework (tra

cui il National Institute of Standards and Technology (NIST), il Payment Card Industry Security Standards Council (PCI) e l'International Organization for Standardization (ISO)).

- [AWS Config](#) è utile per valutare il livello di conformità delle configurazioni delle risorse con pratiche interne, linee guida e regolamenti del settore.
- [AWS Compliance Resources](#) è una raccolta di cartelle di lavoro e guide applicabili al tuo settore e alla tua località.
- [AWS Security Hub](#) ti offre una visione completa del tuo stato di sicurezza interno AWS e ti aiuta a verificare la tua conformità agli standard e alle migliori pratiche del settore della sicurezza.
- [AWS Audit Manager](#)— questo vi Servizio AWS aiuta a controllare continuamente AWS l'utilizzo per semplificare la gestione del rischio e la conformità alle normative e agli standard di settore.

## Resilienza in Amazon EMR Serverless

L'infrastruttura AWS globale è costruita attorno AWS a regioni e zone di disponibilità. AWS Le regioni forniscono più zone di disponibilità fisicamente separate e isolate, collegate con reti a bassa latenza, ad alto throughput e altamente ridondanti. Con le zone di disponibilità, puoi progettare e gestire applicazioni e database che eseguono automaticamente il failover tra zone di disponibilità senza interruzioni. Le zone di disponibilità sono più disponibili, tolleranti ai guasti e scalabili rispetto alle infrastrutture a data center singolo o multiplo tradizionali.

[Per ulteriori informazioni su AWS regioni e zone di disponibilità, consulta Global Infrastructure.AWS](#)

Oltre all'infrastruttura AWS globale, Amazon EMR Serverless offre l'integrazione con Amazon S3 tramite EMRFS per supportare le tue esigenze di resilienza e backup dei dati.

## Sicurezza dell'infrastruttura in Amazon EMR Serverless

In quanto servizio gestito, Amazon EMR è protetto dalla sicurezza di rete AWS globale. Per informazioni sui servizi di AWS sicurezza e su come AWS protegge l'infrastruttura, consulta [AWS Cloud Security](#). Per progettare il tuo AWS ambiente utilizzando le migliori pratiche per la sicurezza dell'infrastruttura, vedi [Infrastructure Protection](#) in Security Pillar AWS Well-Architected Framework.

Utilizzi chiamate API AWS pubblicate per accedere ad Amazon EMR attraverso la rete. I client devono supportare quanto segue:

- Transport Layer Security (TLS). È richiesto TLS 1.2 ed è consigliato TLS 1.3.

- Suite di cifratura con Perfect Forward Secrecy (PFS), ad esempio Ephemeral Diffie-Hellman (DHE) o Elliptic Curve Ephemeral Diffie-Hellman (ECDHE). La maggior parte dei sistemi moderni, come Java 7 e versioni successive, supporta tali modalità.

Inoltre, le richieste devono essere firmate utilizzando un ID chiave di accesso e una chiave di accesso segreta associata a un principale IAM. In alternativa, è possibile utilizzare [AWS Security Token Service](#) (AWS STS) per generare le credenziali di sicurezza temporanee per sottoscrivere le richieste.

## Analisi della configurazione e delle vulnerabilità in Amazon EMR Serverless

AWS gestisce le attività di sicurezza di base come l'applicazione di patch al sistema operativo guest (OS) e al database, la configurazione del firewall e il disaster recovery. Queste procedure sono state riviste e certificate dalle terze parti appropriate. Per ulteriori dettagli, consulta le seguenti risorse :

- [Convalida della conformità per Amazon EMR Serverless](#)
- [Modello di responsabilità condivisa](#)
- [Amazon Web Services: panoramica dei processi di sicurezza](#) (whitepaper)

# Punti finali e quote per EMR Serverless

## Endpoint di servizio

Per connettersi a livello di codice a un Servizio AWS, si utilizza un endpoint. Un endpoint è l'URL del punto di ingresso per un servizio Web. AWS Oltre agli AWS endpoint standard, alcuni Servizi AWS offrono endpoint FIPS in regioni selezionate. La tabella seguente elenca gli endpoint di servizio per EMR Serverless. Per ulteriori informazioni, consulta la pagina relativa agli [endpoint Servizio AWS](#).

Nome Regione	Regione	Endpoint	Protocollo
Stati Uniti orientali (Ohio)	us-east-2 (limitato alle seguenti zone di disponibilità: use2-az1, use2-az2 e use2-az3)	emr-serverless.us-east-2.amazonaws.com	HTTPS
Stati Uniti orientali (Virginia settentrionale)	us-east-1 (limitato alle seguenti zone di disponibilità: use1-az1, use1-az2, use1-az4, use1-az5, use1-az6)	emr-serverless.us-east-1.amazonaws.com emr-serverless-fips.us-east-1.amazonaws.com	HTTPS
Stati Uniti occidentali (California settentrionale)	us-west-1	emr-serverless.us-west-1.amazonaws.com	HTTPS
US West (Oregon)	us-west-2	emr-serverless.us-west-2.amazonaws.com	HTTPS

Nome Regione	Regione	Endpoint	Protocollo
		emr-serverless-fips.us-west-2.amazonaws.com	
Africa (Città del Capo)	af-south-1	emr-serverless.af-south-1.amazonaws.com	HTTPS
Asia Pacifico (Hong Kong)	ap-east-1	emr-serverless.ap-east-1.amazonaws.com	HTTPS
Asia Pacifico (Giacarta)	ap-southeast-3	emr-serverless.ap-southeast-3.amazonaws.com	HTTPS
Asia Pacifico (Mumbai)	ap-south-1	emr-serverless.ap-south-1.amazonaws.com	HTTPS
Asia Pacifico (Osaka-Locale)	ap-northeast-3	emr-serverless.ap-northeast-3.amazonaws.com	HTTPS

Nome Regione	Regione	Endpoint	Protocollo
Asia Pacifico (Seoul)	ap-northeast-2	emr-serverless.ap-northeast-2.amazonaws.com	HTTPS
Asia Pacifico (Singapore)	ap-southeast-1	emr-serverless.ap-southeast-1.amazonaws.com	HTTPS
Asia Pacifico (Sydney)	ap-southeast-2	emr-serverless.ap-southeast-2.amazonaws.com	HTTPS
Asia Pacifico (Tokyo)	ap-northeast-1	emr-serverless.ap-northeast-1.amazonaws.com	HTTPS
Canada (Centrale)	ca-central-1 (limitato alle seguenti zone di disponibilità: cac1-az1 ecac1-az2)	emr-serverless.ca-central-1.amazonaws.com	HTTPS
Europa (Francoforte)	eu-central-1	emr-serverless.eu-central-1.amazonaws.com	HTTPS

Nome Regione	Regione	Endpoint	Protocollo
Europa (Irlanda)	eu-west-1	emr-serverless.eu-west-1.amazonaws.com	HTTPS
Europa (Londra)	eu-west-2	emr-serverless.eu-west-2.amazonaws.com	HTTPS
Europa (Milano)	eu-south-1	emr-serverless.eu-south-1.amazonaws.com	HTTPS
Europa (Parigi)	eu-west-3	emr-serverless.eu-west-3.amazonaws.com	HTTPS
Europa (Spagna)	eu-south-2	emr-serverless.eu-south-2.amazonaws.com	HTTPS
Europa (Stoccolma)	eu-north-1	emr-serverless.eu-north-1.amazonaws.com	HTTPS
Medio Oriente (Bahrein)	me-south-1	emr-serverless.me-south-1.amazonaws.com	HTTPS

Nome Regione	Regione	Endpoint	Protocollo
Medio Oriente (Emirati Arabi Uniti)	me-central-1	emr-serverless.me-central-1.amazonaws.com	HTTPS
Sud America (San Paolo)	sa-east-1	emr-serverless.sa-east-1.amazonaws.com	HTTPS
Cina (Pechino)	cn-north-1 (limitato alle seguenti zone di disponibilità: cnn1-az1, cnn1-az2)	emr-serverless.cn-north-1.amazonaws.com.cn	HTTPS
AWS GovCloud (Stati Uniti orientali)	us-gov-east-1	emr-serverless.us-gov-east-1.amazonaws.com	HTTPS
AWS GovCloud (Stati Uniti occidentali)	us-gov-west-1	emr-serverless.us-gov-west-1.amazonaws.com	HTTPS

## Quote del servizio

Le quote di servizio, note anche come limiti, sono il numero massimo di risorse o operazioni di servizio che è Account AWS possibile utilizzare. EMR Serverless raccoglie i parametri di utilizzo delle quote di servizio ogni minuto e li pubblica nel namespace. AWS/Usage

**Note**

AWS I nuovi account potrebbero avere quote iniziali inferiori che possono aumentare nel tempo. Amazon EMR Serverless monitora l'utilizzo degli account all'interno di ciascun account Regione AWS, quindi aumenta automaticamente le quote in base all'utilizzo.

La tabella seguente elenca le quote di servizio per EMR Serverless. Per ulteriori informazioni, consulta la pagina relativa alle [quote di Servizio AWS](#).

Nome	Limite predefinito	Modificabile?	Descrizione
Numero massimo di v simultanei per account CPUs	16	Sì	Il numero massimo di v CPUs che può essere eseguito contemporaneamente per l'account nella versione corrente. Regione AWS
Numero massimo di lavori in coda per account	2000	Sì	Il numero massimo di lavori in coda per l'account nel periodo corrente. Regione AWS

## Limiti di API

Di seguito vengono descritti i limiti API per regione per il tuo Account AWS

Risorsa	Quota predefinita
<a href="#">ListApplications</a>	10 transazioni al secondo. Burst di 50 transazioni al secondo.

Risorsa	Quota predefinita
<a href="#">CreateApplication</a>	1 transazione al secondo. Una raffica di 25 transazioni al secondo.
<a href="#">DeleteApplication</a>	1 transazione al secondo. Una raffica di 25 transazioni al secondo.
<a href="#">GetApplication</a>	10 transazioni al secondo. Burst di 50 transazioni al secondo.
<a href="#">UpdateApplication</a>	1 transazione al secondo. Una raffica di 25 transazioni al secondo.
<a href="#">ListJobRuns</a>	1 transazione al secondo. Una raffica di 25 transazioni al secondo.
<a href="#">StartJobRun</a>	1 transazione al secondo. Una raffica di 25 transazioni al secondo.
<a href="#">GetDashboardForJobRun</a>	1 transazione al secondo. Una raffica di 2 transazioni al secondo.
<a href="#">CancelJobRun</a>	1 transazione al secondo. Una raffica di 25 transazioni al secondo.
<a href="#">GetJobRun</a>	10 transazioni al secondo. Burst di 50 transazioni al secondo.
<a href="#">StartApplication</a>	1 transazione al secondo. Una raffica di 25 transazioni al secondo.
<a href="#">StopApplication</a>	1 transazione al secondo. Una raffica di 25 transazioni al secondo.

## Altre considerazioni

L'elenco seguente contiene altre considerazioni su EMR Serverless.

- EMR Serverless è disponibile nelle seguenti versioni: Regioni AWS
  - Stati Uniti orientali (Ohio)
  - Stati Uniti orientali (Virginia settentrionale)
  - Stati Uniti occidentali (California settentrionale)
  - Stati Uniti occidentali (Oregon)
  - Africa (Città del Capo)
  - Asia Pacifico (Hong Kong)
  - Asia Pacifico (Giacarta)
  - Asia Pacifico (Mumbai)
  - Asia Pacifico (Osaka)
  - Asia Pacifico (Seul)
  - Asia Pacifico (Singapore)
  - Asia Pacifico (Sydney)
  - Asia Pacifico (Tokyo)
  - Canada (Centrale)
  - Europa (Francoforte)
  - Europa (Irlanda)
  - Europa (Londra)
  - Europa (Milano)
  - Europa (Parigi)
  - Europa (Spagna)
  - Europa (Stoccolma)
  - Medio Oriente (Bahrein)
  - Medio Oriente (Emirati Arabi Uniti)
  - Sud America (San Paolo)
- ~~AWS GovCloud (Stati Uniti orientali)~~
- AWS GovCloud (Stati Uniti occidentali)

Per un elenco degli endpoint associati a queste regioni, vedere. [Endpoint di servizio](#)

- Il timeout predefinito per l'esecuzione di un processo è di 12 ore. Puoi modificare questa impostazione con la `executionTimeoutMinutes` proprietà nell'`startJobRunAPI` o nell' AWS SDK. È possibile impostare `executionTimeoutMinutes` su 0 se si desidera che l'esecuzione del processo non scada mai. Ad esempio, se si dispone di un'applicazione di streaming, è possibile impostare `executionTimeoutMinutes` su 0 per consentire l'esecuzione continua del processo di streaming.
- La `billedResourceUtilization` proprietà nell'`getJobRunAPI` mostra la vCPU, la memoria e lo storage aggregati fatturati per AWS l'esecuzione del processo. Le risorse fatturate includono un utilizzo minimo di 1 minuto per i lavoratori, oltre a uno storage aggiuntivo di oltre 20 GB per lavoratore. Queste risorse non includono l'utilizzo per lavoratori inattivi preinizializzati.
- Senza la connettività VPC, un job può accedere ad alcuni Servizio AWS endpoint nello stesso. Regione AWS Questi servizi includono Amazon S3, AWS Glue, AWS Lake Formation, Amazon CloudWatch Logs, AWS KMS AWS Security Token Service, Amazon DynamoDB e. AWS Secrets Manager Puoi abilitare la connettività VPC per accedere ad altri Servizi AWS tramite [AWS PrivateLink](#), ma non è necessario farlo. Per accedere a servizi esterni, puoi creare la tua applicazione con un VPC.
- EMR Serverless non supporta HDFS. I dischi locali sui lavoratori sono lo storage temporale che EMR Serverless utilizza per mescolare ed elaborare i dati durante le esecuzioni dei job.

# Versioni di rilascio di Amazon EMR Serverless

Una versione di Amazon EMR è un insieme di applicazioni open source dell'ecosistema dei big data. Ogni versione include applicazioni, componenti e funzionalità per big data selezionati per consentire la distribuzione e la configurazione di Amazon EMR Serverless durante l'esecuzione del lavoro.

Con Amazon EMR 6.6.0 e versioni successive, puoi implementare EMR Serverless. Questa opzione di distribuzione non è disponibile con le versioni precedenti di Amazon EMR. Quando invii il tuo lavoro, devi specificare una delle seguenti versioni supportate.

## Argomenti

- [EMR Serverless7.9.0](#)
- [EMR Serverless7,80](#)
- [EMR Serverless7,7,0](#)
- [EMR Serverless7,6,0](#)
- [EMR Serverless7,5,0](#)
- [EMR Serverless7,40](#)
- [EMR Serverless7,3,0](#)
- [EMR Serverless7.2.0](#)
- [EMR Serverless7.1.0](#)
- [EMR Serverless7,0,0](#)
- [EMR Serverless6,15,0](#)
- [EMR Serverless6.14.0](#)
- [EMR Serverless6,13,0](#)
- [EMR Serverless6,12,0](#)
- [EMR Serverless6.11,0](#)
- [EMR Serverless6.10.0](#)
- [EMR Serverless6.9.0](#)
- [EMR Serverless6.8.0](#)
- [EMR Serverless6.7.0](#)
- [EMR Serverless6.6.0](#)

## EMR Serverless 7.9.0

La tabella seguente elenca le versioni delle applicazioni disponibili con EMR Serverless 7.9.0.

Applicazione	Versione
Apache Spark	3.5.5
Apache Hive	3.1.3
Apache Tez	0,10,2

## EMR Serverless 7,80

La tabella seguente elenca le versioni delle applicazioni disponibili con EMR Serverless 7.8.0.

Applicazione	Versione
Apache Spark	3.5.4
Apache Hive	3.1.3
Apache Tez	0,10,2

## EMR Serverless 7,7,0

La tabella seguente elenca le versioni delle applicazioni disponibili con EMR Serverless 7.7.0.

Applicazione	Versione
Apache Spark	3.5.3
Apache Hive	3.1.3
Apache Tez	0,10,2

## EMR Serverless 7,6,0

La tabella seguente elenca le versioni delle applicazioni disponibili con EMR Serverless 7.6.0.

Applicazione	Versione
Apache Spark	3.5.3
Apache Hive	3.1.3
Apache Tez	0,10,2

## EMR Serverless 7,5,0

La tabella seguente elenca le versioni delle applicazioni disponibili con EMR Serverless 7.5.0.

Applicazione	Versione
Apache Spark	3.5.2
Apache Hive	3.1.3
Apache Tez	0.10.2

## EMR Serverless 7,4,0

La tabella seguente elenca le versioni delle applicazioni disponibili con EMR Serverless 7.4.0.

Applicazione	Versione
Apache Spark	3.5.2
Apache Hive	3.1.3
Apache Tez	0.10.2

## EMR Serverless 7,3,0

La tabella seguente elenca le versioni delle applicazioni disponibili con EMR Serverless 7.3.0.

Applicazione	Versione
Apache Spark	3.5.1
Apache Hive	3.1.3
Apache Tez	0,10,2

### Note di rilascio di EMR Serverless 7.3.0

- Job concurrency e accodamento con EMR Serverless: la concorrenza e l'accodamento dei job sono abilitati per impostazione predefinita quando crei una nuova applicazione EMR Serverless su Amazon EMR versione 7.3.0 o successiva. Per ulteriori informazioni, consulta [the section called “Concorrenza e attesa dei lavori”](#) la sezione che descrive in dettaglio come iniziare a utilizzare la concorrenza e l'accodamento e contiene anche un elenco di considerazioni sulle funzionalità.

## EMR Serverless 7.2.0

La tabella seguente elenca le versioni delle applicazioni disponibili con EMR Serverless 7.2.0.

Applicazione	Versione
Apache Spark	3.5.1
Apache Hive	3.1.3
Apache Tez	0,10,2

### Note di rilascio di EMR Serverless 7.2.0

- Lake Formation con EMR Serverless: ora puoi utilizzarlo AWS Lake Formation per applicare controlli di accesso granulari alle tabelle del Data Catalog supportate da S3. Questa funzionalità consente di configurare i controlli di accesso a livello di tabella, riga, colonna e cella per le query di

lettura all'interno dei job EMR Serverless Spark. Per ulteriori informazioni, consultare [the section called “Lake Formation per FGAC”](#) e [the section called “Considerazioni”](#).

## EMR Serverless7.1.0

La tabella seguente elenca le versioni delle applicazioni disponibili con EMR Serverless 7.1.0.

Applicazione	Versione
Apache Spark	3.5.0
Apache Hive	3.1.3
Apache Tez	0.10.2

## EMR Serverless7,0,0

La tabella seguente elenca le versioni delle applicazioni disponibili con EMR Serverless 7.0.0.

Applicazione	Versione
Apache Spark	3.5.0
Apache Hive	3.1.3
Apache Tez	0.10.2

## EMR Serverless6,15,0

La tabella seguente elenca le versioni delle applicazioni disponibili con EMR Serverless 6.15.0.

Applicazione	Versione
Apache Spark	3.4.1
Apache Hive	3.1.3

Applicazione	Versione
Apache Tez	0.10.2

### Note di rilascio di EMR Serverless 6.15.0

- **Supporto TLS:** con le versioni 6.15.0 e successive di Amazon EMR Serverless, puoi abilitare la comunicazione crittografata Mutual-TLS tra i lavoratori che eseguono i job Spark. Se abilitato, EMR Serverless genera automaticamente un certificato univoco per ogni lavoratore che fornisce nell'ambito di un job run che i lavoratori utilizzano durante l'handshake TLS per autenticarsi a vicenda e stabilire un canale crittografato per elaborare i dati in modo sicuro. [Per ulteriori informazioni sulla crittografia Mutual-TLS, vedere Crittografia tra lavoratori.](#)

## EMR Serverless6.14.0

La tabella seguente elenca le versioni delle applicazioni disponibili con EMR Serverless 6.14.0.

Applicazione	Versione
Apache Spark	3.4.1
Apache Hive	3.1.3
Apache Tez	0.10.2

## EMR Serverless6,13,0

La tabella seguente elenca le versioni delle applicazioni disponibili con EMR Serverless 6.13.0.

Applicazione	Versione
Apache Spark	3.4.1
Apache Hive	3.1.3
Apache Tez	0.10.2

## EMR Serverless 6.12.0

La tabella seguente elenca le versioni delle applicazioni disponibili con EMR Serverless 6.12.0.

Applicazione	Versione
Apache Spark	3.4.0
Apache Hive	3.1.3
Apache Tez	0.10.2

## EMR Serverless 6.11.0

La tabella seguente elenca le versioni delle applicazioni disponibili con EMR Serverless 6.11.0.

Applicazione	Versione
Apache Spark	3.3.2
Apache Hive	3.1.3
Apache Tez	0.10.2

### Note di rilascio di EMR Serverless 6.11.0

- [Accedi alle risorse S3 da altri account](#) - Con le versioni 6.11.0 e successive, puoi configurare più ruoli IAM da assumere quando accedi ai bucket Amazon S3 in account diversi da EMR Serverless.  
AWS

## EMR Serverless 6.10.0

La tabella seguente elenca le versioni delle applicazioni disponibili con EMR Serverless 6.10.0.

Applicazione	Versione
Apache Spark	3.3.1
Apache Hive	3.1.3
Apache Tez	0.10.2

### Note di rilascio di EMR Serverless 6.10.0

- Per le applicazioni EMR Serverless con release 6.10.0 o superiore, il valore predefinito per la proprietà è `spark.dynamicAllocation.maxExecutors infinity`. Le versioni precedenti hanno come impostazione predefinita `100`. Per ulteriori informazioni, consulta [Proprietà del lavoro Spark](#).

## EMR Serverless 6.9.0

La tabella seguente elenca le versioni delle applicazioni disponibili con EMR Serverless la versione 6.9.0.

Applicazione	Versione
Apache Spark	3.3.0
Apache Hive	3.1.3
Apache Tez	0.10.2

### Note di rilascio di EMR Serverless 6.9.0

- L'integrazione di Amazon Redshift per Apache Spark è inclusa in Amazon EMR rilascio 6.9.0 e successivi. In precedenza uno strumento open source, l'integrazione nativa è un connettore Spark che è possibile utilizzare per creare applicazioni Apache Spark in grado di leggere e scrivere dati in Amazon Redshift e Amazon Redshift Serverless. Per ulteriori informazioni, consulta [Utilizzo dell'integrazione di Amazon Redshift per Apache Spark su Amazon EMR Serverless](#).

- La release 6.9.0 di EMR Serverless aggiunge il supporto per l'architettura AWS Graviton2 (arm64). È possibile utilizzare il `architecture` parametro per scegliere l'architettura arm64. `create-application` `update-application` APIs Per ulteriori informazioni, consulta [Opzioni di architettura Serverless Amazon EMR](#).
- Ora puoi esportare, importare, interrogare e unire tabelle Amazon DynamoDB direttamente dalle tue applicazioni EMR Serverless Spark e Hive. Per ulteriori informazioni, consulta [Connessione a DynamoDB con Amazon EMR Serverless](#).

## Problemi noti

- Se utilizzi l'integrazione di Amazon Redshift per Apache Spark e disponi di un'indicazione temporale `time`, `timez`, `timestamp` o `timestamptz` con una precisione di microsecondi in formato Parquet, il connettore arrotonda i valori temporali al valore in millisecondi più vicino. Come soluzione alternativa, utilizza il parametro `unload_s3_format` del formato di scaricamento del testo.

## EMR Serverless6.8.0

La tabella seguente elenca le versioni delle applicazioni disponibili con EMR Serverless la versione 6.8.0.

Applicazione	Versione
Apache Spark	3.3.0
Apache Hive	3.1.3
Apache Tez	0.9.2

## EMR Serverless6.7.0

La tabella seguente elenca le versioni delle applicazioni disponibili con EMR Serverless 6.7.0.

Applicazione	Versione
Apache Spark	3.2.1

Applicazione	Versione
Apache Hive	3.1.3
Apache Tez	0.9.2

## Modifiche, miglioramenti e problemi risolti specifici del motore

La tabella seguente elenca una nuova funzionalità specifica del motore.

Modifica	Descrizione
Funzionalità	Tez scheduler ora supporta la priorità dell'attività Tez anziché la priorità del contenitore

## EMR Serverless 6.6.0

La tabella seguente elenca le versioni delle applicazioni disponibili con EMR Serverless 6.6.0.

Applicazione	Versione
Apache Spark	3.2.0
Apache Hive	3.1.2
Apache Tez	0.9.2

### Note di rilascio iniziali di EMR Serverless

- EMR Serverless supporta la classificazione della configurazione Spark. `spark-defaults` Questa classificazione modifica i valori nel file XML di Spark. `spark-defaults.conf` Le classificazioni di configurazione consentono di personalizzare le applicazioni. Per ulteriori informazioni, consulta la sezione [Configurazione delle applicazioni](#).
- EMR Serverless supporta le classificazioni `hive-site` di configurazione Hive, e `tez-site`, `emrfs-site`, `core-site` Questa classificazione può modificare rispettivamente i valori nel file di Hive, nel `hive-site.xml` file di Tez, nelle impostazioni EMRFS di Amazon EMR o nel `tez-`

`site.xml` file di Hadoop. `core-site.xml` Le classificazioni di configurazione consentono di personalizzare le applicazioni. Per ulteriori informazioni, consulta la sezione [Configurazione delle applicazioni](#).

Modifiche, miglioramenti e problemi risolti specifici del motore

- La tabella seguente elenca i backport di Hive e Tez.

Modifiche a Hive e Tez

Modifica	Descrizione
Backport	<a href="#">TEZ-4430</a> : problema risolto con la proprietà <code>tez.task.launch.cmd-opts</code>
Backport	<a href="#">HIVE-25971</a> : Risolti i ritardi di chiusura delle attività Tez dovuti all'apertura del pool di thread memorizzato nella cache

## Cronologia dei documenti

La tabella seguente descrive le modifiche importanti alla documentazione dall'ultima versione di EMR Serverless. Per ulteriori informazioni sugli aggiornamenti di questa documentazione, puoi abbonarti a un feed RSS.

Modifica	Descrizione	Data
<a href="#">Nuova versione</a>	<a href="#">EMR Serverless7.2.0</a>	25 luglio 2024
<a href="#">Nuova versione</a>	<a href="#">EMR Serverless7.1.0</a>	17 aprile 2024
<a href="#">Aggiornamento a una politica esistente.</a>	Aggiunta la nuova politica <code>Sid CloudWatchPolicyStatement</code> e <code>EC2PolicyStatement</code> alla <a href="#">EMRServerlessServiceRolePolicy</a> politica di Amazon.	25 gennaio 2024
<a href="#">Nuova versione</a>	<a href="#">EMR Serverless7,0,0</a>	29 dicembre 2023
<a href="#">Nuova versione</a>	<a href="#">EMR Serverless6,15,0</a>	17 novembre 2023
<a href="#">Nuova caratteristica</a>	<a href="#">Configura più ruoli IAM da assumere quando accedi ai bucket Amazon S3 in account diversi da EMR Serverless (6.11 e versioni successive)</a>	18 ottobre 2023
<a href="#">Nuova versione</a>	<a href="#">EMR Serverless6.14.0</a>	17 ottobre 2023
<a href="#">Nuova caratteristica</a>	<a href="#">Configurazione predefinita dell'applicazione per EMR Serverless</a>	25 settembre 2023
<a href="#">Aggiornamento alle proprietà Hive predefinite</a>	Sono stati aggiornati i valori predefiniti per <code>hive.driver.disk</code> , <code>hive.tez</code> .	12 settembre 2023

	<p><code>disk.size</code> <code>hive.tez.auto.reducer.parallelism</code>, e le proprietà del <a href="#">lavoro <code>tez.grouping.min-size</code></a> <a href="#">Hive</a>.</p>	
<a href="#">Nuova versione</a>	<a href="#">EMR Serverless6,13,0</a>	11 settembre 2023
<a href="#">Nuova versione</a>	<a href="#">EMR Serverless6,12,0</a>	21 luglio 2023
<a href="#">Nuova versione</a>	<a href="#">EMR Serverless6.11,0</a>	8 giugno 2023
<a href="#">Aggiornamento della politica relativa ai ruoli collegati ai servizi</a>	<p>È stato aggiornato il ruolo <a href="#">AmazonEMRServerlessServiceRolePolicy</a> <code>_SLR</code> per pubblicare l'utilizzo a livello di account nel namespace. "AWS/Usage"</p>	20 aprile 2023
<a href="#">EMR Serverless disponibilità generale (GA)</a>	Questa è la prima versione pubblica di EMR Serverless.	1 giugno 2022

Le traduzioni sono generate tramite traduzione automatica. In caso di conflitto tra il contenuto di una traduzione e la versione originale in Inglese, quest'ultima prevarrà.