



Verwenden Sie Apache Iceberg auf AWS

AWS Präskriptive Leitlinien



AWS Präskriptive Leitlinien: Verwenden Sie Apache Iceberg auf AWS

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

Die Handelsmarken und Handelsaufmachung von Amazon dürfen nicht in einer Weise in Verbindung mit nicht von Amazon stammenden Produkten oder Services verwendet werden, durch die Kunden irregeführt werden könnten oder Amazon in schlechtem Licht dargestellt oder diskreditiert werden könnte. Alle anderen Handelsmarken, die nicht Eigentum von Amazon sind, gehören den jeweiligen Besitzern, die möglicherweise zu Amazon gehören oder nicht, mit Amazon verbunden sind oder von Amazon gesponsert werden.

Table of Contents

Einführung	1
Moderne Datenseen	3
Fortgeschrittene Anwendungsfälle in modernen Data Lakes	3
Einführung in Apache Iceberg	4
AWS Unterstützung für Apache Iceberg	5
Erste Schritte mit Iceberg-Tabellen in Athena SQL	8
Eine unpartitionierte Tabelle erstellen	8
Eine partitionierte Tabelle erstellen	9
Erstellen einer Tabelle und Laden von Daten mit einer einzigen CTAS-Anweisung	9
Daten einfügen, aktualisieren und löschen	10
Iceberg-Tabellen abfragen	11
Anatomie der Eisbergtabelle	12
Arbeiten mit Iceberg in Amazon EMR	14
Versions- und Funktionskompatibilität	14
Erstellen eines Amazon EMR-Clusters mit Iceberg	14
Entwicklung von Iceberg-Anwendungen in Amazon EMR	15
Verwenden von Amazon EMR Studio-Notizbüchern	15
Iceberg-Jobs in Amazon EMR ausführen	16
Bewährte Methoden für Amazon EMR	21
Ich arbeite mit Iceberg in AWS Glue	23
Verwenden Sie die native Iceberg-Integration	23
Verwenden Sie eine benutzerdefinierte Iceberg-Version	24
Verwenden Sie einen benutzerdefinierten Konnektor	24
Bringen Sie Ihre eigenen JAR-Dateien mit	26
Spark-Konfigurationen für Iceberg in AWS Glue	26
Bewährte Methoden für AWS Glue Jobs	28
Arbeiten mit Iceberg-Tabellen mithilfe von Spark	30
Iceberg-Tabellen erstellen und schreiben	30
Verwenden von Spark SQL	30
Verwenden der API DataFrames	31
Daten in Iceberg-Tabellen aktualisieren	32
Daten in Iceberg-Tabellen werden aktualisiert	33
Löschen von Daten in Iceberg-Tabellen	33
Lesen von Daten	34

Zeitreisen nutzen	34
Verwendung inkrementeller Abfragen	35
Zugreifen auf Metadaten	36
Arbeiten mit Iceberg-Tabellen mithilfe von Firehose	38
Arbeiten mit Iceberg-Tabellen mithilfe von Athena SQL	39
Versions- und Funktionskompatibilität	39
Unterstützung der Iceberg-Tabellenspezifikationen	39
Unterstützung für Iceberg-Funktionen	39
Mit Iceberg-Tabellen arbeiten	40
Migration vorhandener Tabellen zu Iceberg	42
Direkte Migration	42
Vollständige Datenmigration	47
Auswahl einer Migrationsstrategie	48
Bewährte Methoden zur Optimierung von Iceberg-Workloads	51
Allgemeine bewährte Methoden	51
Optimierung der Leseleistung	52
Partitionierung	53
Optimieren der Dateigrößen	55
Optimieren Sie die Spaltenstatistiken	57
Wählen Sie die richtige Aktualisierungsstrategie	58
Verwenden Sie die STD-Komprimierung	58
Legen Sie die Sortierreihenfolge fest	58
Optimierung der Schreibleistung	61
Stellen Sie den Tabellenverteilungsmodus ein	61
Wählen Sie die richtige Aktualisierungsstrategie	61
Wählen Sie das richtige Dateiformat	62
Optimierung des Speichers	63
Aktivieren Sie S3 Intelligent-Tiering	63
Archivieren oder löschen Sie historische Snapshots	64
Löschen Sie verwaiste Dateien	67
Pflege von Tabellen mithilfe von Komprimierung	68
Verdichtung von Eisbergen	68
Optimieren des Verdichtungsverhaltens	70
Verdichtung mit Spark auf Amazon EMR ausführen oder AWS Glue	71
Verdichtung mit Amazon Athena ausführen	72
Empfehlungen für die Ausführung der Komprimierung	72

Verwenden von Iceberg-Workloads in Amazon S3	74
Vermeiden Sie Hot-Partitionierung (HTTP 503-Fehler)	74
Verwenden Sie die Wartungsoperationen von Iceberg, um ungenutzte Daten freizugeben	74
Replizieren Sie Daten zwischen AWS-Regionen	75
Überwachung von Iceberg-Workloads	77
Überwachung auf Tabellenebene	77
Überwachung auf Datenbankebene	79
Präventive Wartung	81
Governance und Zugriffskontrolle	82
Referenzarchitekturen	83
Nächtliche Batch-Erfassung	83
Data Lake, der Batch-Erfassung und Erfassung nahezu in Echtzeit kombiniert	84
Ressourcen	85
Mitwirkende	86
Dokumentverlauf	88
Glossar	89
#	89
A	90
B	93
C	95
D	98
E	103
F	105
G	107
H	108
I	110
L	112
M	113
O	118
P	121
Q	124
R	124
S	127
T	131
U	133
V	134

W	134
Z	135
.....	CXXXVII

Apache Iceberg verwenden auf AWS

Amazon Web Services ([Mitwirkende](#))

Februar 2025 ([Verlauf der Dokumente](#))

Apache Iceberg ist ein Open-Source-Tabellenformat, das die Tabellenverwaltung vereinfacht und gleichzeitig die Leistung verbessert. AWS Analysedienste wie Amazon EMR AWS Glue, Amazon Athena und Amazon Redshift bieten native Unterstützung für Apache Iceberg, sodass Sie auf einfache Weise transaktionale Data Lakes auf Amazon Simple Storage Service (Amazon S3) aufbauen können. AWS

Dieser technische Leitfaden bietet Anleitungen für die ersten Schritte mit Apache Iceberg auf verschiedenen AWS-Services Geräten und enthält bewährte Methoden und Empfehlungen für die skalierbare Ausführung von Apache Iceberg bei AWS gleichzeitiger Kosten- und Leistungsoptimierung.

Dieses Handbuch richtet sich an alle, die Apache Iceberg verwenden AWS, von Anfängern, die schnell mit Apache Iceberg beginnen möchten, bis hin zu fortgeschrittenen Benutzern, die ihre bestehenden Apache Iceberg-Workloads optimieren und optimieren möchten. AWS

In diesem Handbuch:

- [Moderne Data Lakes](#)
- [Erste Schritte mit Iceberg-Tabellen in Athena SQL](#)
- [Arbeiten mit Iceberg in Amazon EMR](#)
- [Zusammenarbeit mit Iceberg in AWS Glue](#)
- [Arbeiten mit Iceberg-Tabellen mithilfe von Spark](#)
- [Arbeiten mit Iceberg-Tabellen mithilfe von Amazon Data Firehose](#)
- [Arbeiten mit Iceberg-Tabellen mithilfe von Athena SQL](#)
- [Migrieren vorhandener Tabellen nach Iceberg](#)
- [Bewährte Methoden zur Optimierung von Iceberg-Workloads](#)
- [Überwachung von Iceberg-Workloads](#)
- [Verwaltung und Zugriffskontrolle](#)
- [Referenzarchitekturen](#)
- [Ressourcen](#)

- [Mitwirkende](#)

Moderne Datenseen

Fortgeschrittene Anwendungsfälle in modernen Data Lakes

Data Lakes bieten eine der besten Optionen für die Speicherung von Daten in Bezug auf Kosten, Skalierbarkeit und Flexibilität. Sie können einen Data Lake verwenden, um große Mengen strukturierter und unstrukturierter Daten kostengünstig aufzubewahren und diese Daten für verschiedene Arten von Analyse-Workloads zu verwenden, von Business Intelligence-Berichten über Big-Data-Verarbeitung, Echtzeitanalysen, maschinelles Lernen und generative künstliche Intelligenz (KI), um bessere Entscheidungen zu treffen.

Trotz dieser Vorteile wurden Data Lakes ursprünglich nicht mit datenbankähnlichen Funktionen konzipiert. Ein Data Lake bietet keine Unterstützung für die Verarbeitungssemantik Atomicity, Consistency, Isolation, Durability (ACID), die Sie möglicherweise benötigen, um Ihre Daten mithilfe einer Vielzahl verschiedener Technologien effektiv zu optimieren und zu verwalten, und zwar in großem Umfang für Hunderte oder Tausende von Benutzern. Data Lakes bieten keine native Unterstützung für die folgenden Funktionen:

- Durchführung effizienter Aktualisierungen und Löschungen auf Datensatzebene, wenn sich Daten in Ihrem Unternehmen ändern
- Verwaltung der Abfrageleistung, wenn Tabellen auf Millionen von Dateien und Hunderttausende von Partitionen anwachsen
- Sicherstellung der Datenkonsistenz bei mehreren gleichzeitigen Schreib- und Lesegeräten
- Verhinderung von Datenbeschädigungen, wenn Schreibvorgänge während des Vorgangs fehlschlagen
- Weiterentwicklung von Tabellenschemas im Laufe der Zeit, ohne Datensätze (teilweise) neu zu schreiben

Diese Herausforderungen haben sich in Anwendungsfällen wie der Handhabung von Change Data Capture (CDC) oder in Anwendungsfällen, die den Datenschutz, das Löschen von Daten und die Aufnahme von Streaming-Daten betreffen, besonders ausgeprägt, was zu suboptimalen Tabellen führen kann.

Data Lakes, die traditionelle Tabellen im HIVE-Format verwenden, unterstützen Schreibvorgänge nur für ganze Dateien. Dies macht die Implementierung von Aktualisierungen und Löschungen schwierig,

zeitaufwändig und kostspielig. Darüber hinaus sind Parallelitätskontrollen und -garantien erforderlich, die in ACID-konformen Systemen angeboten werden, um die Integrität und Konsistenz der Daten zu gewährleisten.

Um diese Herausforderungen zu bewältigen, bietet Apache Iceberg zusätzliche datenbankähnliche Funktionen, die den Optimierungs- und Verwaltungsaufwand von Data Lakes vereinfachen und gleichzeitig die Speicherung auf kostengünstigen Systemen wie [Amazon Simple Storage Service \(Amazon S3\)](#) unterstützen.

Einführung in Apache Iceberg

Apache Iceberg ist ein Open-Source-Tabellenformat, das Funktionen in Data-Lake-Tabellen bereitstellt, die in der Vergangenheit nur in Datenbanken oder Data Warehouses verfügbar waren. Es ist auf Skalierbarkeit und Leistung ausgelegt und eignet sich hervorragend für die Verwaltung von Tabellen mit einer Größe von mehr als Hunderten von Gigabyte. Einige der Hauptmerkmale von Iceberg-Tabellen sind:

- Löschen, aktualisieren und zusammenführen. Iceberg unterstützt Standard-SQL-Befehle für Data Warehousing zur Verwendung mit Data-Lake-Tabellen.
- Schnelle Scanplanung und erweiterte Filterung. Iceberg speichert Metadaten wie Statistiken auf Partitions- und Spaltenebene, die von Engines verwendet werden können, um die Planung und Ausführung von Abfragen zu beschleunigen.
- Vollständige Entwicklung des Schemas. Iceberg unterstützt das Hinzufügen, Löschen, Aktualisieren oder Umbenennen von Spalten ohne Nebenwirkungen.
- Entwicklung der Partition. Sie können das Partitionslayout einer Tabelle aktualisieren, wenn sich das Datenvolumen oder die Abfragemuster ändern. Iceberg unterstützt das Ändern der Spalten, auf die eine Tabelle partitioniert ist, oder das Hinzufügen von Spalten zu oder das Entfernen von Spalten aus zusammengesetzten Partitionen.
- Versteckte Partitionierung. Diese Funktion verhindert, dass unnötige Partitionen automatisch gelesen werden. Dadurch müssen Benutzer die Partitionierungsdetails der Tabelle nicht verstehen oder ihren Abfragen zusätzliche Filter hinzufügen.
- Rollback der Version. Benutzer können Probleme schnell korrigieren, indem sie zu einem Zustand vor der Transaktion zurückkehren.
- Zeitreise. Benutzer können eine bestimmte frühere Version einer Tabelle abfragen.
- Serialisierbare Isolierung. Tabellenänderungen sind atomar, sodass Leser keine teilweisen oder nicht festgeschriebenen Änderungen sehen.

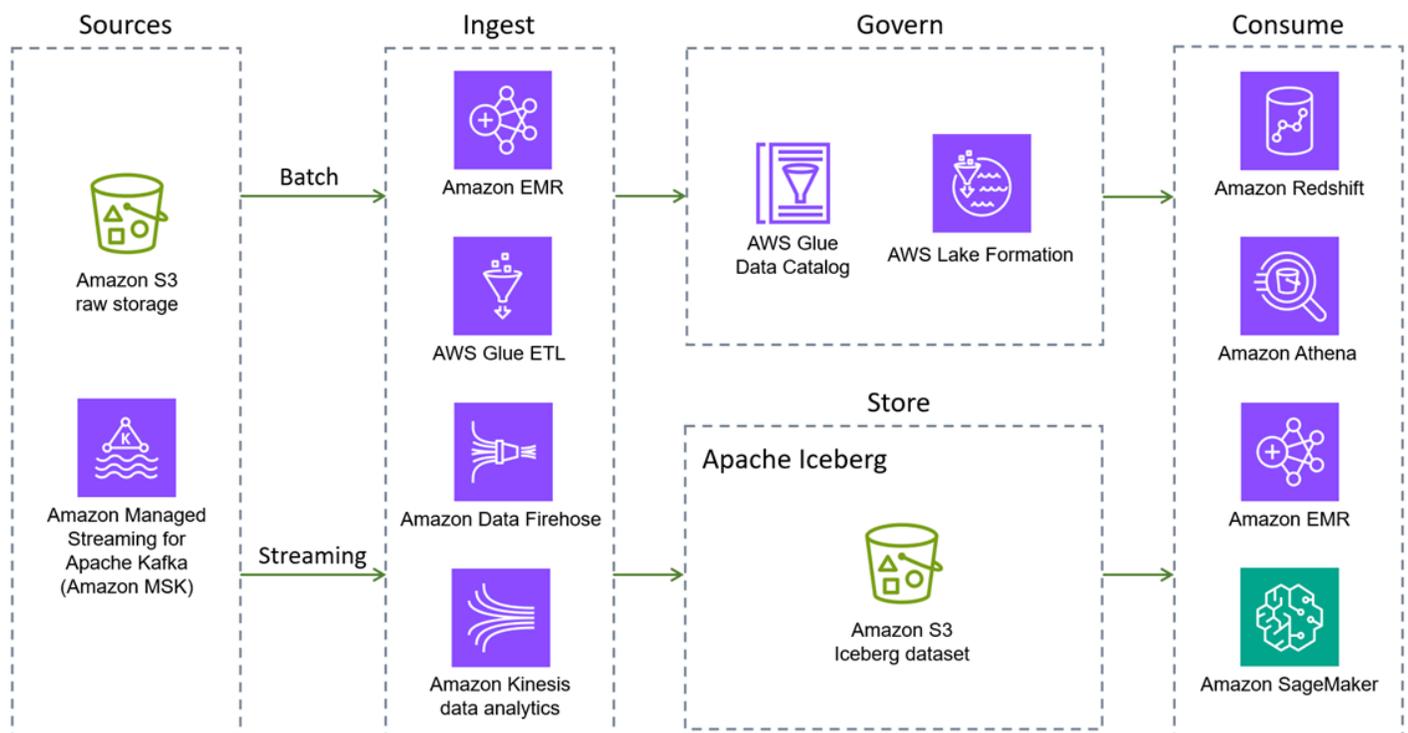
- Gleichzeitige Autoren. Iceberg verwendet optimistische Parallelität, um den Erfolg mehrerer Transaktionen zu ermöglichen. Im Falle von Konflikten muss einer der Autoren die Transaktion erneut versuchen.
- Dateiformate öffnen. Iceberg unterstützt mehrere Open-Source-Dateiformate, darunter [Apache Parquet](#), [Apache Avro](#) und [Apache ORC](#).

Zusammenfassend lässt sich sagen, dass Data Lakes, die das Iceberg-Format verwenden, von Transaktionskonsistenz, Geschwindigkeit, Skalierbarkeit und Schemaentwicklung profitieren.

[Weitere Informationen zu diesen und anderen Iceberg-Funktionen finden Sie in der Apache Iceberg-Dokumentation.](#)

AWS Unterstützung für Apache Iceberg

Apache Iceberg wird von beliebigen Open-Source-Datenverarbeitungs-Frameworks AWS-Services wie [Amazon EMR](#), [Amazon Athena](#), [Amazon Redshift](#) und unterstützt. [AWS Glue](#) Das folgende Diagramm zeigt eine vereinfachte Referenzarchitektur eines Data Lakes, der auf Iceberg basiert.



Im Folgenden finden Sie native AWS-Services Iceberg-Integrationen. Es gibt weitere AWS-Services, die entweder indirekt oder durch Paketierung der Iceberg-Bibliotheken mit Iceberg interagieren können.

- Amazon S3 ist aufgrund seiner Beständigkeit, Verfügbarkeit, Skalierbarkeit, Sicherheit, Compliance und Auditfunktionen der beste Ort für den Aufbau von Data Lakes. Iceberg wurde für die nahtlose Interaktion mit Amazon S3 konzipiert und entwickelt und bietet Unterstützung für viele Amazon S3 S3-Funktionen, die in der [Iceberg-Dokumentation](#) aufgeführt sind.
- Amazon EMR ist eine Big-Data-Lösung für Datenverarbeitung im Petabyte-Bereich, interaktive Analysen und maschinelles Lernen unter Verwendung von Open-Source-Frameworks wie Apache Spark, Flink, Trino und Hive. Amazon EMR kann auf benutzerdefinierten Amazon Elastic Compute Cloud (Amazon EC2) -Clustern, Amazon Elastic Kubernetes Service (Amazon EKS) oder Amazon EMR AWS Outposts Serverless ausgeführt werden.
- Amazon Athena ist ein serverloser, interaktiver Analysedienst, der auf Open-Source-Frameworks basiert. Er unterstützt Open-Table- und Dateiformate und bietet eine vereinfachte, flexible Möglichkeit, Petabyte an Daten dort zu analysieren, wo sie sich befinden. Athena bietet native Unterstützung für Lese-, Zeitreise-, Schreib- und DDL-Abfragen für Iceberg und verwendet den Metastore AWS Glue Data Catalog for the Iceberg.
- Amazon Redshift ist ein Cloud-Data Warehouse im Petabyte-Bereich, das sowohl clusterbasierte als auch serverlose Bereitstellungsoptionen unterstützt. Amazon Redshift Spectrum kann externe Tabellen abfragen, die bei Amazon S3 registriert AWS Glue Data Catalog und dort gespeichert sind. Redshift Spectrum unterstützt auch das Iceberg-Speicherformat.
- AWS Glue ist ein serverloser Datenintegrationsdienst, der es einfacher macht, Daten aus verschiedenen Quellen für Analysen, maschinelles Lernen (ML) und Anwendungsentwicklung zu finden, aufzubereiten, zu verschieben und zu integrieren. AWS Glue 3.0 und spätere Versionen unterstützen das Iceberg-Framework für Data Lakes. Sie können AWS Glue damit Lese- und Schreibvorgänge an Iceberg-Tabellen in Amazon S3 durchführen oder mit Iceberg-Tabellen arbeiten, indem Sie den verwenden. AWS Glue Data Catalog Zusätzliche Operationen wie Einfügen, Aktualisieren, Spark-Abfragen und Spark-Schreibvorgänge werden ebenfalls unterstützt.
- AWS Glue Data Catalog bietet einen mit Hive-Metastores kompatiblen Datenkatalogdienst, der Iceberg-Tabellen unterstützt.
- AWS-Glue-Crawler bietet Automatisierungen zur Registrierung von Iceberg-Tabellen in der. AWS Glue Data Catalog
- Amazon Data Firehose ist ein vollständig verwalteter Service für die Bereitstellung von Echtzeit-Streaming-Daten an Ziele wie Amazon S3, Amazon Redshift, Amazon OpenSearch Service, Amazon OpenSearch Serverless, Splunk, Apache Iceberg-Tabellen und alle benutzerdefinierten HTTP- oder HTTP-Endpunkte unterstützter Drittanbieter wie Datadog, Dynatrace, MongoDB, New Relic, Coralogix und Elastic. LogicMonitor Mit Firehose müssen Sie keine Anwendungen schreiben oder Ressourcen verwalten. Sie konfigurieren Ihre Datenproduzenten zum Senden von Daten

an Firehose. Die Daten werden dann automatisch an das angegebene Ziel geliefert. Sie können Firehose auch so konfigurieren, dass die Daten vor der Bereitstellung transformiert werden.

- Amazon SageMaker AI unterstützt die Speicherung von Feature-Sets im Amazon SageMaker AI Feature Store mithilfe des Iceberg-Formats.
- AWS Lake Formation bietet grobe und detaillierte Zugriffsberechtigungen für den Zugriff auf Daten, einschließlich Iceberg-Tabellen, die von Athena oder Amazon Redshift verwendet werden. Weitere Informationen zur Unterstützung von Berechtigungen für Iceberg-Tabellen finden Sie in der [Dokumentation zu Lake Formation](#).

AWS bietet eine breite Palette von Diensten, die Iceberg unterstützen, aber das Abdecken all dieser Dienste würde den Rahmen dieses Leitfadens sprengen. Die folgenden Abschnitte behandeln Spark (Batch- und strukturiertes Streaming) auf Amazon EMR und AWS Glue Amazon Athena SQL. Der folgende Abschnitt bietet einen kurzen Überblick über die Iceberg-Unterstützung in Athena SQL.

Erste Schritte mit Apache Iceberg-Tabellen in Amazon Athena SQL

Amazon Athena bietet integrierte Unterstützung für Apache Iceberg. Sie können Iceberg ohne zusätzliche Schritte oder Konfiguration verwenden, mit Ausnahme der Einrichtung der Servicevoraussetzungen, die im Abschnitt [Erste Schritte](#) der Athena-Dokumentation beschrieben sind. Dieser Abschnitt bietet eine kurze Einführung in das Erstellen von Tabellen in Athena. Weitere Informationen finden Sie unter [Arbeiten mit Apache Iceberg-Tabellen mithilfe von Athena SQL](#) weiter unten in diesem Handbuch.

Sie können Iceberg-Tabellen AWS mithilfe verschiedener Engines erstellen. Diese Tabellen funktionieren problemlos. AWS-Services Um Ihre ersten Iceberg-Tabellen mit Athena SQL zu erstellen, können Sie den folgenden Boilerplate-Code verwenden.

```
CREATE TABLE <table_name> (  
    col_1 string,  
    col_2 string,  
    col_3 bigint,  
    col_ts timestamp)  
PARTITIONED BY (col_1, <<<partition_transform>>(col_ts))  
LOCATION 's3://<bucket>/<folder>/<table_name>/'  
TBLPROPERTIES (  
    'table_type' = 'ICEBERG'  
)
```

Die folgenden Abschnitte enthalten Beispiele für die Erstellung einer partitionierten und unpartitionierten Iceberg-Tabelle in Athena. Weitere Informationen finden Sie in der Iceberg-Syntax in der [Athena-Dokumentation](#).

Eine unpartitionierte Tabelle erstellen

Die folgende Beispielanweisung passt den SQL-Standardcode an, um eine unpartitionierte Iceberg-Tabelle in Athena zu erstellen. [Sie können diese Anweisung dem Abfrage-Editor in der Athenaconsole hinzufügen, um die Tabelle zu erstellen.](#)

```
CREATE TABLE athena_iceberg_table (  
    color string,
```

```
date string,  
name string,  
price bigint,  
product string,  
ts timestamp)  
LOCATION 's3://DOC_EXAMPLE_BUCKET/ice_warehouse/iceberg_db/athena_iceberg_table/'  
TBLPROPERTIES (  
  'table_type' = 'ICEBERG'  
)
```

step-by-step Anweisungen zur Verwendung des Abfrage-Editors finden Sie unter [Erste Schritte](#) in der Athena-Dokumentation.

Eine partitionierte Tabelle erstellen

[Die folgende Anweisung erstellt eine partitionierte Tabelle auf der Grundlage des Datums, wobei das Konzept der versteckten Partitionierung von Iceberg verwendet wird.](#) Sie verwendet die `day()` Transformation, um tägliche Partitionen unter Verwendung des `dd-mm-yyyy` Formats aus einer Zeitstempelspalte abzuleiten. Iceberg speichert diesen Wert nicht als neue Spalte im Datensatz. Stattdessen wird der Wert spontan abgeleitet, wenn Sie Daten schreiben oder abfragen.

```
CREATE TABLE athena_iceberg_table_partitioned (  
  color string,  
  date string,  
  name string,  
  price bigint,  
  product string,  
  ts timestamp)  
PARTITIONED BY (day(ts))  
LOCATION 's3://DOC_EXAMPLE_BUCKET/ice_warehouse/iceberg_db/athena_iceberg_table/'  
TBLPROPERTIES (  
  'table_type' = 'ICEBERG'  
)
```

Erstellen einer Tabelle und Laden von Daten mit einer einzigen CTAS-Anweisung

In den partitionierten und unpartitionierten Beispielen in den vorherigen Abschnitten wurden die Iceberg-Tabellen als leere Tabellen erstellt. Sie können Daten in die Tabellen laden, indem Sie

die Anweisung `OR` verwenden. `INSERT MERGE` Alternativ können Sie eine `CREATE TABLE AS SELECT` (CTAS) Anweisung verwenden, um Daten in einem einzigen Schritt zu erstellen und in eine Iceberg-Tabelle zu laden.

CTAS ist die beste Methode in Athena, um eine Tabelle zu erstellen und Daten in einer einzigen Anweisung zu laden. Das folgende Beispiel zeigt, wie CTAS verwendet wird, um eine Iceberg-Tabelle (`iceberg_ctas_table`) aus einer vorhandenen Hive/Parquet-Tabelle (`hive_table`) in Athena zu erstellen.

```
CREATE TABLE iceberg_ctas_table WITH (  
  table_type = 'ICEBERG',  
  is_external = false,  
  location = 's3://DOC_EXAMPLE_BUCKET/ice_warehouse/iceberg_db/iceberg_ctas_table/'  
) AS  
SELECT * FROM "iceberg_db"."hive_table" limit 20  
---  
SELECT * FROM "iceberg_db"."iceberg_ctas_table" limit 20
```

Weitere Informationen zu CTAS finden Sie in der [Athena CTAS-Dokumentation](#).

Daten einfügen, aktualisieren und löschen

Athena unterstützt verschiedene Arten, Daten mithilfe der Anweisungen, `INSERT INTO`, `UPDATE` und `DELETE FROM` in eine Iceberg-Tabelle zu schreiben. `MERGE INTO`

Hinweis: `UPDATE`, `MERGE INTO`, und `DELETE FROM` verwenden Sie den merge-on-read Ansatz mit Positionslösungen. Der copy-on-write Ansatz wird derzeit in Athena SQL nicht unterstützt.

Die folgende Anweisung verwendet beispielsweise das Hinzufügen von Daten `INSERT INTO` zu einer Iceberg-Tabelle:

```
INSERT INTO "iceberg_db"."ice_table" VALUES (  
  'red', '222022-07-19T03:47:29', 'PersonNew', 178, 'Tuna', now()  
)  
  
SELECT * FROM "iceberg_db"."ice_table"  
where color = 'red' limit 10;
```

Beispielausgabe:

Results (1)							
#	color	date	name	price	product	ts	
1	red	222022-07-19T03:47:29	PersonNew	178	Tuna	2023-10-11 11:35:01.298000 UTC	

Weitere Informationen finden Sie in der [Athena-Dokumentation](#).

Iceberg-Tabellen abfragen

Sie können reguläre SQL-Abfragen für Ihre Iceberg-Tabellen ausführen, indem Sie Athena SQL verwenden, wie im vorherigen Beispiel dargestellt.

Zusätzlich zu den üblichen Abfragen unterstützt Athena auch Zeitreiseabfragen für Iceberg-Tabellen. Wie bereits erwähnt, können Sie vorhandene Datensätze durch Aktualisierungen oder Löschungen in einer Iceberg-Tabelle ändern. Daher ist es praktisch, Zeitreiseabfragen zu verwenden, um anhand eines Zeitstempels oder einer Snapshot-ID in ältere Versionen Ihrer Tabelle zurückzuschauen.

Die folgende Anweisung aktualisiert beispielsweise einen Farbwert für Person5 und zeigt dann einen früheren Wert vom 4. Januar 2023 an:

```
UPDATE ice_table SET color='new_color' WHERE name='Person5'

SELECT * FROM "iceberg_db"."ice_table" FOR TIMESTAMP AS OF TIMESTAMP '2023-01-04
12:00:00 UTC'
```

Beispielausgabe:

Results (15)							
#	color	date	name	price	product	ts	
1	cyan	222022-07-19T03:47:29	Person5	353	Keyboard	2023-01-03 10:15:52.268000 UTC	
2	lime	222022-07-19T03:47:29	Person1	833	Towels	2023-01-03 10:15:52.268000 UTC	
3	turquoise	222022-07-19T03:47:29	Person1	1319	Shirt	2023-01-03 10:15:52.268000 UTC	
4	blue	222022-07-19T03:47:29	Person3	163	Sausages	2023-01-03 10:15:52.268000 UTC	

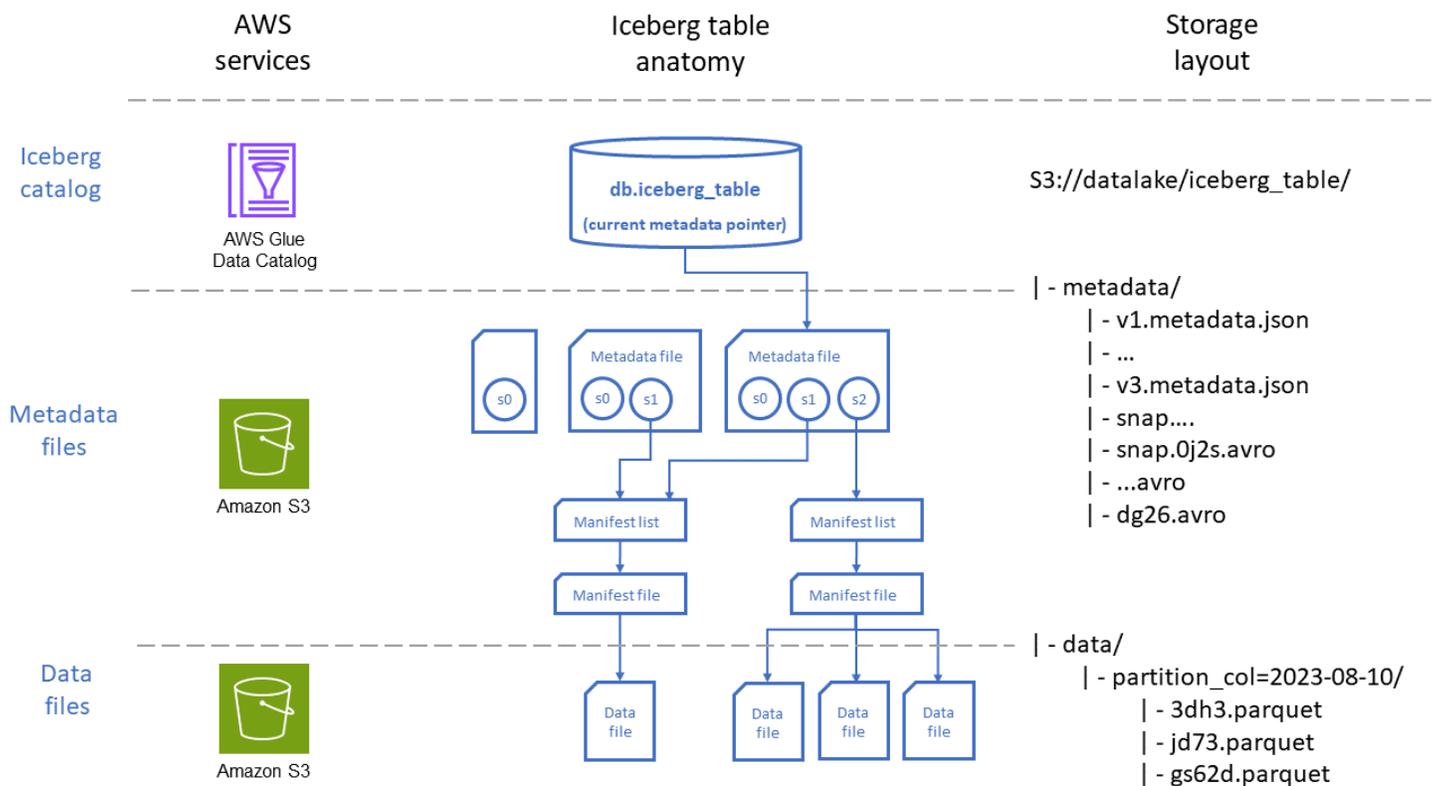
Syntax und weitere Beispiele für Zeitreiseabfragen finden Sie in der [Athena-Dokumentation](#).

Anatomie der Eisbergtabelle

Nachdem wir nun die grundlegenden Schritte der Arbeit mit Iceberg-Tischen behandelt haben, wollen wir uns eingehender mit den komplizierten Details und dem Design eines Iceberg-Tisches befassen.

Um die [zuvor in diesem Handbuch beschriebenen](#) Funktionen nutzen zu können, wurde Iceberg mit hierarchischen Ebenen von Daten- und Metadateien entworfen. Diese Ebenen verwalten Metadaten auf intelligente Weise, um die Planung und Ausführung von Abfragen zu optimieren.

Das folgende Diagramm zeigt die Organisation einer Iceberg-Tabelle aus zwei Perspektiven: die zum Speichern der Tabelle AWS-Services verwendete und die Dateiplatzierung in Amazon S3.



Wie in der Abbildung dargestellt, besteht eine Eisbergtabelle aus drei Hauptebenen:

- **Iceberg-Katalog:** AWS Glue Data Catalog lässt sich nativ in Iceberg integrieren und ist in den meisten Anwendungsfällen die beste Option für Workloads, die auf ausgeführt werden. AWS Dienste, die mit Iceberg-Tabellen interagieren (z. B. Athena), verwenden den Katalog, um die aktuelle Snapshot-Version der Tabelle zu finden, entweder um Daten zu lesen oder zu schreiben.
- **Metadatenebene:** Metadateien, nämlich die Manifestdateien und die Manifestlistendateien, verfolgen Informationen wie das Schema der Tabellen, die Partitionsstrategie und den Speicherort der Datendateien sowie Statistiken auf Spaltenebene wie Mindest- und Höchstbereiche für die

Datensätze, die in jeder Datendatei gespeichert sind. Diese Metadatendateien werden in Amazon S3 innerhalb des Tabellenpfads gespeichert.

- Manifestdateien enthalten einen Datensatz für jede Datendatei, einschließlich Speicherort, Format, Größe, Prüfsumme und anderer relevanter Informationen.
- Manifestlisten bieten einen Index der Manifestdateien. Wenn die Anzahl der Manifestdateien in einer Tabelle zunimmt, trägt die Aufteilung dieser Informationen in kleinere Unterabschnitte dazu bei, die Anzahl der Manifestdateien zu reduzieren, die durch Abfragen gescannt werden müssen.
- Metadatendateien enthalten Informationen über die gesamte Iceberg-Tabelle, einschließlich der Manifestlisten, Schemas, Partitionsmetadaten, Snapshot-Dateien und anderer Dateien, die zur Verwaltung der Tabellenmetadaten verwendet werden.
- Datenschicht: Diese Ebene enthält die Dateien mit den Datensätzen, für die Abfragen ausgeführt werden. Diese Dateien können in verschiedenen Formaten gespeichert werden, darunter [Apache Parquet](#), [Apache Avro](#) und [Apache ORC](#).
- Datendateien enthalten die Datensätze für eine Tabelle.
- Mit gelöschten Dateien werden Lösch- und Aktualisierungsvorgänge auf Zeilenebene in einer Iceberg-Tabelle kodiert. [In Iceberg gibt es zwei Arten von Löschdateien, wie in der Iceberg-Dokumentation beschrieben](#). Diese Dateien werden durch Operationen unter Verwendung des merge-on-read Modus erstellt.

Arbeiten mit Apache Iceberg in Amazon EMR

Amazon EMR bietet Datenverarbeitung im Petabyte-Bereich, interaktive Analysen und maschinelles Lernen in der Cloud mithilfe von Open-Source-Frameworks wie Apache Spark, Apache Hive, Flink und Trino.

Note

In diesem Handbuch wird Apache Spark als Beispiele verwendet.

Amazon EMR unterstützt mehrere Bereitstellungsoptionen: Amazon EMR auf Amazon EC2, Amazon EMR auf Amazon EKS, Amazon EMR Serverless und Amazon EMR auf AWS Outposts. Informationen zur Auswahl einer Bereitstellungsoption für Ihren Workload finden Sie in den [häufig gestellten Fragen zu Amazon EMR](#).

Versions- und Funktionskompatibilität

Amazon EMR Version 6.5.0 und spätere Versionen unterstützen Apache Iceberg nativ. Eine Liste der unterstützten Iceberg-Versionen für jede Amazon EMR-Version finden Sie im [Iceberg-Versionsverlauf](#) in der Amazon EMR-Dokumentation. Lesen Sie auch die [Überlegungen und Einschränkungen zur Verwendung von Iceberg auf Amazon EMR](#), um zu erfahren, welche Iceberg-Funktionen in Amazon EMR auf verschiedenen Frameworks unterstützt werden.

Wir empfehlen Ihnen, die neueste Amazon EMR-Version zu verwenden, um von der neuesten unterstützten Iceberg-Version zu profitieren. Bei den Codebeispielen und Konfigurationen in diesem Abschnitt wird davon ausgegangen, dass Sie die Amazon EMR-Version emr-6.9.0 verwenden.

Erstellen eines Amazon EMR-Clusters mit Iceberg

Um einen Amazon EMR-Cluster auf Amazon EC2 mit installiertem Iceberg zu erstellen, folgen Sie den Anweisungen in der [Amazon EMR-Dokumentation](#).

Insbesondere sollte Ihr Cluster mit der folgenden Klassifizierung konfiguriert werden:

```
[{  
  "Classification": "iceberg-defaults",
```

```
"Properties": {
  "iceberg.enabled": "true"
}
}]
```

Ab Amazon EMR 6.6.0 können Sie auch Amazon EMR Serverless oder Amazon EMR auf Amazon EKS als Bereitstellungsoptionen für Ihre Iceberg-Workloads verwenden.

Entwicklung von Iceberg-Anwendungen in Amazon EMR

Um den Spark-Code für Ihre Iceberg-Anwendungen zu entwickeln, können Sie [Amazon EMR Studio](#) verwenden, eine webbasierte integrierte Entwicklungsumgebung (IDE) für vollständig verwaltete Jupyter-Notebooks, die auf Amazon EMR-Clustern ausgeführt werden.

Verwenden von Amazon EMR Studio-Notizbüchern

Sie können Spark-Anwendungen interaktiv in Amazon EMR Studio Workspace-Notebooks entwickeln und diese Notebooks mit Ihren Amazon EMR auf Amazon EC2 EC2-Clustern oder Amazon EMR auf Amazon EKS-verwalteten Endpunkten verbinden. [Anweisungen zur Einrichtung von EMR Studio für Amazon EMR auf Amazon EC2 und Amazon EMR auf Amazon EKS finden Sie in der AWS-Service Dokumentation.](#)

Gehen Sie wie folgt vor, um Iceberg in EMR Studio zu verwenden:

1. Starten Sie einen Amazon EMR-Cluster mit aktiviertem Iceberg, wie unter [Verwenden Sie einen Cluster mit](#) installiertem Iceberg beschrieben.
2. Richten Sie ein EMR Studio ein. Anweisungen finden Sie unter [Amazon EMR Studio einrichten](#).
3. Öffnen Sie ein EMR Studio Workspace-Notizbuch und führen Sie den folgenden Code als erste Zelle im Notizbuch aus, um Ihre Spark-Sitzung für die Verwendung von Iceberg zu konfigurieren:

```
%%configure -f
{
  "conf": {
    "spark.sql.catalog.<catalog_name>": "org.apache.iceberg.spark.SparkCatalog",
    "spark.sql.catalog.<catalog_name>.warehouse": "s3://YOUR-BUCKET-NAME/YOUR-
FOLDER-NAME/",
    "spark.sql.catalog.<catalog_name>.catalog-impl":
"org.apache.iceberg.aws.glue.GlueCatalog",
    "spark.sql.catalog.<catalog_name>.io-impl":
"org.apache.iceberg.aws.s3.S3FileIO",
```

```
"spark.sql.extensions":  
  "org.apache.iceberg.spark.extensions.IcebergSparkSessionExtensions"  
  }  
}
```

Wobei:

- `<catalog_name>` ist der Name Ihres Iceberg Spark-Sitzungskatalogs. Ersetzen Sie ihn durch den Namen Ihres Katalogs und denken Sie daran, die Verweise in allen Konfigurationen, die mit diesem Katalog verknüpft sind, zu ändern. In Ihrem Code sollten Sie dann wie folgt auf Ihre Iceberg-Tabellen mit dem vollqualifizierten Tabellennamen, einschließlich des Namens des Spark-Sitzungskatalogs, verweisen:

```
<catalog_name>.<database_name>.<table_name>
```

- `<catalog_name>.warehouse` verweist auf den Amazon S3 S3-Pfad, in dem Sie Ihre Daten und Metadaten speichern möchten.
 - Um den Katalog zu einem zu machen AWS Glue Data Catalog, setzen Sie `<catalog_name>.catalog-impl` auf `org.apache.iceberg.aws.glue.GlueCatalog`. Dieser Schlüssel ist erforderlich, um auf eine Implementierungsklasse für jede benutzerdefinierte Katalogimplementierung zu verweisen. Im Abschnitt [Allgemeine bewährte Methoden weiter](#) unten in diesem Handbuch werden die verschiedenen von Iceberg unterstützten Kataloge beschrieben.
 - Verwenden Sie `org.apache.iceberg.aws.s3.S3FileIO` als, um `<catalog_name>.io-impl` den mehrteiligen Amazon S3 S3-Upload für hohe Parallelität zu nutzen.
4. Sie können jetzt wie bei jeder anderen Spark-Anwendung mit der interaktiven Entwicklung Ihrer Spark-Anwendung für Iceberg im Notizbuch beginnen.

Weitere Informationen zur Konfiguration von Spark für Apache Iceberg mithilfe von Amazon EMR Studio finden Sie im Blogbeitrag [Build a high-performance, ACID-compliant, evolving data lake using Apache Iceberg on Amazon EMR](#).

Iceberg-Jobs in Amazon EMR ausführen

Nachdem Sie den Spark-Anwendungscode für Ihren Iceberg-Workload entwickelt haben, können Sie ihn auf jeder Amazon EMR-Bereitstellungsoption ausführen, die Iceberg unterstützt (siehe Häufig gestellte Fragen zu [Amazon EMR](#)).

Wie bei anderen Spark-Jobs können Sie Arbeit an einen Amazon EMR on Amazon EC2 EC2-Cluster senden, indem Sie Schritte hinzufügen oder Spark-Jobs interaktiv an den Master-Knoten senden. Informationen zum Ausführen eines Spark-Jobs finden Sie auf den folgenden Amazon EMR-Dokumentationsseiten:

- Eine Übersicht über die verschiedenen Optionen zum Einreichen von Arbeit an einen Amazon EMR on Amazon EC2 EC2-Cluster und detaillierte Anweisungen für jede Option finden Sie unter [Arbeit an einen Cluster einreichen](#).
- Informationen zu Amazon EMR auf Amazon EKS finden Sie unter [Spark-Jobs ausführen mit StartJobRun](#).
- Informationen zu Amazon EMR Serverless finden Sie unter Jobs [ausführen](#).

Die folgenden Abschnitte enthalten ein Beispiel für jede Amazon EMR-Bereitstellungsoption.

Amazon EMR auf Amazon EC2

Sie können die folgenden Schritte ausführen, um den Iceberg Spark-Job einzureichen:

1. Erstellen Sie die Datei `emr_step_iceberg.json` mit dem folgenden Inhalt auf Ihrer Workstation:

```
[{
  "Name": "iceberg-test-job",
  "Type": "spark",
  "ActionOnFailure": "CONTINUE",
  "Args": [
    "--deploy-mode",
    "client",
    "--conf",

    "spark.sql.extensions=org.apache.iceberg.spark.extensions.IcebergSparkSessionExtensions",
    "--conf",
    "spark.sql.catalog.<catalog_name>=org.apache.iceberg.spark.SparkCatalog",
    "--conf",
    "spark.sql.catalog.<catalog_name>.catalog-
impl=org.apache.iceberg.aws.glue.GlueCatalog",
    "--conf",
    "spark.sql.catalog.<catalog_name>.warehouse=s3://YOUR-BUCKET-NAME/YOUR-
FOLDER-NAME/",
    "--conf",
    "spark.sql.catalog.<catalog_name>.io-
impl=org.apache.iceberg.aws.s3.S3FileIO",
```

```

    "s3://YOUR-BUCKET-NAME/code/iceberg-job.py"
  ]
}]

```

2. Ändern Sie die Konfigurationsdatei für Ihren speziellen Spark-Job, indem Sie die fett hervorgehobenen Iceberg-Konfigurationsoptionen anpassen.
3. Reichen Sie den Schritt mit AWS Command Line Interface (AWS CLI) ein. Führen Sie den Befehl in dem Verzeichnis aus, in dem sich die `emr_step_iceberg.json` Datei befindet.

```
aws emr add-steps --cluster-id <cluster_id> --steps file://emr_step_iceberg.json
```

Amazon EMR Serverless

Um einen Iceberg Spark-Job an Amazon EMR Serverless zu senden, verwenden Sie: AWS CLI

1. Erstellen Sie die Datei `emr_serverless_iceberg.json` mit dem folgenden Inhalt auf Ihrer Workstation:

```

{
  "applicationId": "<APPLICATION_ID>",
  "executionRoleArn": "<ROLE_ARN>",
  "jobDriver": {
    "sparkSubmit": {
      "entryPoint": "s3://YOUR-BUCKET-NAME/code/iceberg-job.py",
      "entryPointArguments": [],
      "sparkSubmitParameters": "--jars /usr/share/aws/iceberg/lib/iceberg-spark3-runtime.jar"
    }
  },
  "configurationOverrides": {
    "applicationConfiguration": [{
      "classification": "spark-defaults",
      "properties": {
        "spark.sql.extensions":
"org.apache.iceberg.spark.extensions.IcebergSparkSessionExtensions",
        "spark.sql.catalog.<catalog_name>":
"org.apache.iceberg.spark.SparkCatalog",
        "spark.sql.catalog.<catalog_name>.catalog-impl":
"org.apache.iceberg.aws.glue.GlueCatalog",
        "spark.sql.catalog.<catalog_name>.warehouse": "s3://YOUR-BUCKET-NAME/YOUR-FOLDER-NAME/"
      }
    }
  ]
}

```

```

"spark.sql.catalog.<catalog_name>.io-impl":
"org.apache.iceberg.aws.s3.S3FileIO",
"spark.jars": "/usr/share/aws/iceberg/lib/iceberg-spark3-runtime.jar",

"spark.hadoop.hive.metastore.client.factory.class": "com.amazonaws.glue.catalog.metastore.AWSGlueCatalogMetastoreClientFactory",
    }
}],
"monitoringConfiguration": {
    "s3MonitoringConfiguration": {
        "logUri": "s3://YOUR-BUCKET-NAME/emr-serverless/logs/"
    }
}
}
}

```

2. Ändern Sie die Konfigurationsdatei für Ihren speziellen Spark-Job, indem Sie die fett hervorgehobenen Iceberg-Konfigurationsoptionen anpassen.
3. Senden Sie den Job mit dem AWS CLI. Führen Sie den Befehl in dem Verzeichnis aus, in dem sich die `emr_serverless_iceberg.json` Datei befindet:

```
aws emr-serverless start-job-run --cli-input-json file://emr_serverless_iceberg.json
```

So reichen Sie mithilfe der EMR Studio-Konsole einen Iceberg Spark-Job an Amazon EMR Serverless ein:

1. Folgen Sie den Anweisungen in der [Amazon EMR Serverless-Dokumentation](#).
2. Verwenden Sie für die Jobkonfiguration die für Spark bereitgestellte Iceberg-Konfiguration für Spark AWS CLI und passen Sie die hervorgehobenen Felder für Iceberg an. Eine ausführliche Anleitung finden Sie unter [Using Apache Iceberg with EMR Serverless](#) in der Amazon EMR-Dokumentation.

Amazon EMR auf Amazon EKS

Um einen Iceberg Spark-Job an Amazon EMR auf Amazon EKS zu senden, verwenden Sie: AWS CLI

1. Erstellen Sie die Datei `emr_eks_iceberg.json` mit dem folgenden Inhalt auf Ihrer Workstation:

```
{
```

```

"name": "iceberg-test-job",
"virtualClusterId": "<VIRTUAL_CLUSTER_ID>",
"executionRoleArn": "<ROLE_ARN>",
"releaseLabel": "emr-6.9.0-latest",
"jobDriver": {
  "sparkSubmitJobDriver": {
    "entryPoint": "s3://YOUR-BUCKET-NAME/code/iceberg-job.py",
    "entryPointArguments": [],
    "sparkSubmitParameters": "--jars local:///usr/share/aws/iceberg/lib/
iceberg-spark3-runtime.jar"
  }
},
"configurationOverrides": {
  "applicationConfiguration": [{
    "classification": "spark-defaults",
    "properties": {
      "spark.sql.extensions":
"org.apache.iceberg.spark.extensions.IcebergSparkSessionExtensions",
      "spark.sql.catalog.<catalog_name>":
"org.apache.iceberg.spark.SparkCatalog",
      "spark.sql.catalog.<catalog_name>.catalog-impl":
"org.apache.iceberg.aws.glue.GlueCatalog",
      "spark.sql.catalog.<catalog_name>.warehouse": "s3://YOUR-BUCKET-NAME/
YOUR-FOLDER-NAME/",
      "spark.sql.catalog.<catalog_name>.io-impl":
"org.apache.iceberg.aws.s3.S3FileIO",
      "spark.hadoop.hive.metastore.client.factory.class":
      "com.amazonaws.glue.catalog.metastore.AWSGlueDataCatalogHiveClientFactory"
    }
  }],
  "monitoringConfiguration": {
    "persistentAppUI": "ENABLED",
    "s3MonitoringConfiguration": {
      "logUri": "s3://YOUR-BUCKET-NAME/emr-serverless/logs/"
    }
  }
}
}

```

2. Ändern Sie die Konfigurationsdatei für Ihren Spark-Job, indem Sie die fett hervorgehobenen Iceberg-Konfigurationsoptionen anpassen.
3. Senden Sie den Job mit dem AWS CLI Führen Sie den folgenden Befehl in dem Verzeichnis aus, in dem sich die `emr_eks_iceberg.json` Datei befindet:

```
aws emr-containers start-job-run --cli-input-json file://emr_eks_iceberg.json
```

Eine ausführliche Anleitung finden Sie unter [Using Apache Iceberg with Amazon EMR on EKS](#) in der Dokumentation zu Amazon EMR on EKS.

Bewährte Methoden für Amazon EMR

Dieser Abschnitt enthält allgemeine Richtlinien für die Optimierung von Spark-Jobs in Amazon EMR, um das Lesen und Schreiben von Daten in Iceberg-Tabellen zu optimieren. Iceberg-spezifische Best Practices finden Sie im Abschnitt „Bewährte Methoden“ weiter [unten in diesem Handbuch](#).

- Verwenden Sie die neueste Version von Amazon EMR — Amazon EMR bietet mit der Amazon EMR Spark-Laufzeit sofort Spark-Optimierungen. AWS verbessert die Leistung der Spark-Runtime-Engine mit jeder neuen Version.
- Ermitteln Sie die optimale Infrastruktur für Ihre Spark-Workloads — Spark-Workloads benötigen möglicherweise unterschiedliche Hardwaretypen für unterschiedliche Jobmerkmale, um eine optimale Leistung zu gewährleisten. Amazon EMR [unterstützt verschiedene Instance-Typen](#) (z. B. rechenoptimiert, speicheroptimiert, universell einsetzbar und speicheroptimiert), um alle Arten von Verarbeitungsanforderungen abzudecken. Wenn Sie neue Workloads integrieren, empfehlen wir, einen Benchmark mit allgemeinen Instance-Typen wie M5 oder M6g durchzuführen. Überwachen Sie das Betriebssystem (OS) und die YARN-Metriken von Ganglia und Amazon, CloudWatch um die Systemengpässe (CPU, Speicher, Speicher und I/O) bei Spitzenlast zu ermitteln und geeignete Hardware auszuwählen.
- Tune `spark.sql.shuffle.partitions` — Legen Sie die `spark.sql.shuffle.partitions` Eigenschaft auf die Gesamtzahl der virtuellen Kerne (vCores) in Ihrem Cluster oder auf ein Vielfaches dieses Werts fest (in der Regel das 1- bis 2-fache der Gesamtzahl der vCores). Diese Einstellung wirkt sich auf die Parallelität von Spark aus, wenn Sie Hash- und Bereichspartitionierung als Schreibverteilungsmodus verwenden. Vor dem Schreiben wird ein Shuffle angefordert, um die Daten zu organisieren, wodurch die Ausrichtung der Partitionen gewährleistet wird.
- Verwaltungsskalierung aktivieren — Für fast alle Anwendungsfälle empfehlen wir, die verwaltete Skalierung und die dynamische Zuweisung zu aktivieren. Wenn Sie jedoch einen Workload haben, der ein vorhersehbares Muster aufweist, empfehlen wir Ihnen, die automatische Skalierung und die dynamische Zuweisung zu deaktivieren. Wenn die verwaltete Skalierung aktiviert ist, empfehlen wir die Verwendung von Spot-Instances, um die Kosten zu senken. Verwenden Sie

Spot-Instances für Task-Knoten anstelle von Core- oder Master-Knoten. Wenn Sie Spot-Instances verwenden, verwenden Sie Instance-Flotten mit mehreren Instance-Typen pro Flotte, um die Spot-Verfügbarkeit sicherzustellen.

- Verwenden Sie nach Möglichkeit Broadcast-Join — Der Broadcast-Join (mapside) ist der optimalste Join, sofern eine Ihrer Tabellen klein genug ist, um in den Speicher Ihres kleinsten Knotens zu passen (in der Reihenfolge von MB) und Sie einen Equi (=) -Join durchführen. Alle Join-Typen mit Ausnahme von vollständigen Outer-Joins werden unterstützt. Ein Broadcast-Join überträgt die kleinere Tabelle als Hashtabelle an alle Worker-Knoten im Speicher. Nachdem die kleine Tabelle übertragen wurde, können Sie keine Änderungen mehr daran vornehmen. Da sich die Hashtabelle lokal in der Java Virtual Machine (JVM) befindet, kann sie anhand der Join-Bedingung mithilfe eines Hash-Joins problemlos mit der großen Tabelle zusammengeführt werden. Broadcast-Joins bieten aufgrund des minimalen Shuffle-Overheads eine hohe Leistung.
- Optimieren Sie den Garbage-Collector — Wenn die Garbage-Collection-Zyklen (GC) langsam sind, sollten Sie erwägen, für eine bessere Leistung vom standardmäßigen parallel Garbage-Collector auf G1GC umzusteigen. Um die GC-Leistung zu optimieren, können Sie die GC-Parameter fein abstimmen. Um die GC-Leistung zu verfolgen, können Sie sie mithilfe der Spark-Benutzeroberfläche überwachen. Idealerweise sollte die GC-Zeit weniger als oder gleich 1 Prozent der gesamten Aufgabenlaufzeit sein.

Arbeiten mit Apache Iceberg in AWS Glue

[AWS Glue](#) ist ein serverloser Datenintegrationsdienst, der es einfacher macht, Daten aus verschiedenen Quellen für Analysen, maschinelles Lernen (ML) und Anwendungsentwicklung zu entdecken, aufzubereiten, zu verschieben und zu integrieren. Eine der Kernfunktionen von AWS Glue ist die Fähigkeit, Extraktions-, Transformations- und Ladevorgänge (ETL) auf einfache und kostengünstige Weise durchzuführen. Auf diese Weise können Sie Ihre Daten kategorisieren, bereinigen, anreichern und zuverlässig zwischen verschiedenen Datenspeichern und Datenströmen verschieben.

[AWS Glue Jobs](#) kapseln Skripten, die Transformationslogik mithilfe einer [Apache Spark](#) - oder Python-Laufzeit definieren. AWS Glue Jobs können sowohl im Batch- als auch im Streaming-Modus ausgeführt werden.

Wenn Sie Iceberg-Jobs in erstellen AWS Glue, können Sie je nach Version von AWS Glue entweder die native Iceberg-Integration oder eine benutzerdefinierte Iceberg-Version verwenden, um Iceberg-Abhängigkeiten an den Job anzuhängen.

Verwenden Sie die native Iceberg-Integration

AWS Glue Die Versionen 3.0 und 4.0 unterstützen nativ transaktionale Data-Lake-Formate wie Apache Iceberg, Apache Hudi und Linux Foundation Delta Lake in for Spark. AWS Glue Diese Integrationsfunktion vereinfacht die Konfigurationsschritte, die erforderlich sind, um mit der Verwendung dieser Frameworks in zu beginnen. AWS Glue

Um die Iceberg-Unterstützung für Ihren AWS Glue Job zu aktivieren, legen Sie den Job fest: Wählen Sie den Tab Jobdetails für Ihren AWS Glue Job, scrollen Sie unter Erweiterte Eigenschaften zu Job-Parametern und setzen Sie den Schlüssel `--dataLake-formats` und seinen Wert auf `iceberg`.

Wenn Sie einen Job mit einem Notizbuch erstellen, können Sie den Parameter in der ersten Notebookzelle konfigurieren, indem Sie die `%%configure` Magie wie folgt verwenden:

```
%%configure
{
  "--conf" : <job-specific Spark configuration discussed later>,
  "--dataLake-formats" : "iceberg"
}
```

Verwenden Sie eine benutzerdefinierte Iceberg-Version

In einigen Situationen möchten Sie möglicherweise die Kontrolle über die Iceberg-Version für den Job behalten und sie in Ihrem eigenen Tempo aktualisieren. Ein Upgrade auf eine neuere Version kann beispielsweise den Zugriff auf neue Funktionen und Leistungsverbesserungen freischalten. Um eine bestimmte Iceberg-Version mit zu verwenden AWS Glue, können Sie einen benutzerdefinierten Connector oder Ihre eigene JAR-Datei verwenden.

Verwenden Sie einen benutzerdefinierten Konnektor

AWS Glue unterstützt Konnektoren, bei denen es sich um optionale Codepakete handelt, die den Zugriff auf Datenspeicher in unterstützen AWS Glue Studio. Sie können [einen Connector abonnieren](#) oder einen benutzerdefinierten Connector erstellen. AWS Marketplace

Note

AWS Marketplace bietet den [Apache Iceberg-Konnektor für AWS Glue](#). Wir empfehlen jedoch, stattdessen einen benutzerdefinierten Konnektor zu verwenden, um die Kontrolle über die Iceberg-Versionen zu behalten.

Gehen Sie beispielsweise wie folgt vor, um einen Kunden-Connector für Iceberg Version 0.13.1 zu erstellen:

1. Laden Sie die Dateien `iceberg-spark-runtime-3.1_2.12-0.13.1.jar` und `url-connection-client-2.17.161.jar` in einen Amazon S3 S3-Bucket hoch. Sie können diese Dateien aus den jeweiligen Apache Maven-Repositorys herunterladen.
2. Erstellen Sie auf der [AWS Glue Studio Konsole](#) einen benutzerdefinierten Spark-Connector:
 - a. Wählen Sie im Navigationsbereich Datenverbindungen aus. (Wenn Sie die ältere Navigation verwenden, wählen Sie Konnektoren, Benutzerdefinierten Konnektor erstellen.)
 - b. Wählen Sie im Feld Konnektoren die Option Benutzerdefinierten Konnektor erstellen aus.
 - c. Gehen Sie auf der Seite Benutzerdefinierten Connector erstellen wie folgt vor:
 - Geben Sie den Pfad zu den JAR-Dateien in Amazon S3 an.
 - Geben Sie einen Namen für den Connector ein.
 - Wählen Sie Spark als Steckertyp.

- Geben Sie als Klassenname den vollqualifizierten Klassennamen der Datenquelle (oder seinen Alias) an, den Sie beim Laden der Spark-Datenquelle mit dem `format` Operator verwenden.
- (Optional) Geben Sie eine Beschreibung des Konnektors an.

3. Wählen Sie Konnektor erstellen.

Wenn Sie mit Konnektoren arbeiten AWS Glue, müssen Sie eine Verbindung für den Konnektor herstellen. Eine Verbindung enthält die Eigenschaften, die für die Verbindung mit einem bestimmten Datenspeicher erforderlich sind. Sie verwenden die Verbindung mit Ihren Datenquellen und Datenzielen im ETL-Auftrag. Connectors und Verbindungen ermöglichen zusammen den Zugriff auf die Datenspeicher.

Gehen Sie wie folgt vor, um mithilfe des von Ihnen erstellten benutzerdefinierten Iceberg-Connectors eine Verbindung herzustellen:

1. Wählen Sie auf der [AWS Glue Studio Konsole](#) Ihren benutzerdefinierten Iceberg-Connector aus.
2. Folgen Sie den Anweisungen, um die Details anzugeben, z. B. Ihre VPC und andere Netzwerkkonfigurationen, die für den Job erforderlich sind, und wählen Sie dann Verbindung erstellen aus.

Sie können die Verbindung jetzt in Ihrem AWS Glue ETL-Job verwenden. Je nachdem, wie Sie den Job erstellen, gibt es verschiedene Möglichkeiten, die Verbindung mit Ihrem Job zu verknüpfen:

- Wenn Sie mithilfe von einem visuellen Auftrag erstellen AWS Glue Studio, können Sie die Verbindung aus der Verbindungsliste auf der Registerkarte Datenquelleneigenschaften — Konnektor auswählen.
- Wenn Sie den Job in einem Notizbuch entwickeln, verwenden Sie die `%connections` Magie, um den Verbindungsnamen festzulegen:

```
%glue_version 3.0

%connections <name-of-the iceberg-connection>

%%configure
{
  "--conf" : "job-specific Spark configurations, to be discussed later",
  "--datalake-formats" : "iceberg"
```

```
}
```

- Wenn Sie den Job mithilfe des Skripteditors erstellen, geben Sie die Verbindung auf der Registerkarte Auftragsdetails unter Erweiterte Eigenschaften, Zusätzliche Netzwerkverbindungen an.

Weitere Informationen zu den Verfahren in diesem Abschnitt finden Sie AWS Glue Studio in der AWS Glue Dokumentation [unter Verwenden von Konnektoren und Verbindungen mit](#).

Bringen Sie Ihre eigenen JAR-Dateien mit

In AWS Glue können Sie auch mit Iceberg arbeiten, ohne einen Connector verwenden zu müssen. Dieser Ansatz ist nützlich, wenn Sie die Kontrolle über die Iceberg-Version behalten und sie schnell aktualisieren möchten. Um diese Option zu verwenden, laden Sie die erforderlichen Iceberg-JAR-Dateien in einen S3-Bucket Ihrer Wahl hoch und verweisen Sie auf die Dateien in Ihrem AWS Glue Job. Wenn Sie beispielsweise mit Iceberg 1.0.0 arbeiten, lauten die erforderlichen JAR-Dateien `iceberg-spark-runtime-3.0_2.12-1.0.0.jar`, `undurl-connection-client-2.15.40.jar`, `bundle-2.15.40.jar`. Sie können die zusätzlichen JAR-Dateien im Klassenpfad auch priorisieren, indem Sie den `--user-jars-first` Parameter `true` für den Job auf setzen.

Spark-Konfigurationen für Iceberg in AWS Glue

In diesem Abschnitt werden die Spark-Konfigurationen beschrieben, die für die Erstellung eines AWS Glue ETL-Jobs für einen Iceberg-Datensatz erforderlich sind. Sie können diese Konfigurationen festlegen, indem Sie den `--conf` Spark-Schlüssel mit einer kommagetrennten Liste aller Spark-Konfigurationsschlüssel und `-werte` verwenden. Sie können die `%%configure` Magie in einem Notizbuch oder im Bereich Job-Parameter der AWS Glue Studio Konsole verwenden.

```
%glue_version 3.0

%connections <name-of-the iceberg-connection>

%%configure
{
  "--conf" : "spark.sql.extensions=org.apache.iceberg.spark.extensions...",
  "--datalake-formats" : "iceberg"
}
```

Konfigurieren Sie die Spark-Sitzung mit den folgenden Eigenschaften:

- `<catalog_name>` ist der Name Ihres Iceberg Spark-Sitzungskatalogs. Ersetzen Sie ihn durch den Namen Ihres Katalogs und denken Sie daran, die Verweise in allen Konfigurationen, die mit diesem Katalog verknüpft sind, zu ändern. In Ihrem Code sollten Sie dann wie folgt auf Ihre Iceberg-Tabellen mit dem vollqualifizierten Tabellennamen, einschließlich des Namens des Spark-Sitzungskatalogs, verweisen: `<catalog_name>.<database_name>.<table_name>`.
- `<catalog_name>.<warehouse>` verweist auf den Amazon S3 S3-Pfad, in dem Sie Ihre Daten und Metadaten speichern möchten.
- Um den Katalog zu einem zu machen AWS Glue Data Catalog, setzen Sie `<catalog_name>.catalog-impl=org.apache.iceberg.aws.glue.GlueCatalog`. Dieser Schlüssel ist erforderlich, um auf eine Implementierungsklasse für jede benutzerdefinierte Katalogimplementierung zu verweisen. Informationen zu Katalogen, die von Iceberg unterstützt werden, finden Sie im [??? Abschnitt Allgemeine bewährte Methoden](#) weiter unten in diesem Handbuch.
- Verwenden Sie `org.apache.iceberg.aws.s3.S3FileIO` als, um `<catalog_name>.io-impl` den mehrteiligen Amazon S3 S3-Upload für hohe Parallelität zu nutzen.

Wenn Sie beispielsweise einen Katalog aufgerufen `habenglue_iceberg`, können Sie Ihren Job mithilfe mehrerer `--conf` Schlüssel wie folgt konfigurieren:

```
%%configure
{
  "--datalake-formats" : "iceberg",
  "--conf" :
  "spark.sql.extensions=org.apache.iceberg.spark.extensions.IcebergSparkSessionExtensions",
  "--conf" : "spark.sql.catalog.glue_iceberg=org.apache.iceberg.spark.SparkCatalog",
  "--conf" : "spark.sql.catalog.glue_iceberg.warehouse=s3://<your-warehouse-dir>/",
  "--conf" : " spark.sql.catalog.glue_iceberg.catalog-
impl=org.apache.iceberg.aws.glue.GlueCatalog ",
  "--conf" : " spark.sql.catalog.glue_iceberg.io-
impl=org.apache.iceberg.aws.s3.S3FileIO
}
```

Alternativ können Sie Code verwenden, um die obigen Konfigurationen wie folgt zu Ihrem Spark-Skript hinzuzufügen:

```
spark = SparkSession.builder\
```

```
.config("spark.sql.extensions","org.apache.iceberg.spark.extensions.IcebergSparkSessionExtensions")\
    .config("spark.sql.catalog.glue_iceberg",
"org.apache.iceberg.spark.SparkCatalog")\
    .config("spark.sql.catalog.glue_iceberg.warehouse","s3://<your-warehouse-dir>/")\
    .config("spark.sql.catalog.glue_iceberg.catalog-impl",
"org.apache.iceberg.aws.glue.GlueCatalog") \
    .config("spark.sql.catalog.glue_iceberg.io-impl",
"org.apache.iceberg.aws.s3.S3FileIO") \
    .getOrCreate()
```

Bewährte Methoden für AWS Glue Jobs

Dieser Abschnitt enthält allgemeine Richtlinien für die Optimierung von Spark-Jobs, AWS Glue um das Lesen und Schreiben von Daten in Iceberg-Tabellen zu optimieren. Iceberg-spezifische Best Practices finden Sie im Abschnitt „Bewährte [Methoden](#)“ weiter unten in diesem Handbuch.

- Verwenden Sie die neueste Version von AWS Glue und führen Sie nach Möglichkeit Upgrades durch — Neue Versionen von AWS Glue bieten Leistungsverbesserungen, kürzere Startzeiten und neue Funktionen. Sie unterstützen auch neuere Spark-Versionen, die möglicherweise für die neuesten Iceberg-Versionen erforderlich sind. Eine Liste der verfügbaren AWS Glue Versionen und der Spark-Versionen, die sie unterstützen, finden Sie in der [AWS Glue Dokumentation](#).
- AWS Glue Arbeitsspeicher optimieren — Folgen Sie den Empfehlungen im AWS Blogbeitrag [Optimieren Sie die Speicherverwaltung in AWS Glue](#).
- AWS Glue Auto Scaling verwenden — Wenn Sie Auto Scaling aktivieren, passt es die Anzahl der AWS Glue Worker AWS Glue automatisch dynamisch an Ihre Arbeitslast an. Dies trägt dazu bei, die Kosten Ihrer AWS Glue Arbeit bei Spitzenauslastung zu senken, da AWS Glue die Anzahl der Mitarbeiter reduziert wird, wenn die Arbeitslast gering ist und die Mitarbeiter untätig sitzen. Um AWS Glue Auto Scaling zu verwenden, geben Sie eine maximale Anzahl von Mitarbeitern an, auf die Ihr AWS Glue Job skaliert werden kann. Weitere Informationen finden Sie AWS Glue in der AWS Glue Dokumentation [unter Using Auto Scaling for](#).
- Verwenden Sie benutzerdefinierte Konnektoren oder fügen Sie Bibliotheksabhängigkeiten hinzu — die AWS Glue native Integration für Iceberg eignet sich am besten für den Einstieg in Iceberg. Für Produktionsworkloads empfehlen wir jedoch, benutzerdefinierte Container zu verwenden oder Bibliotheksabhängigkeiten hinzuzufügen (wie weiter [oben in diesem Handbuch](#) beschrieben), um die volle Kontrolle über die Iceberg-Version zu erhalten. Dieser Ansatz hilft Ihnen, bei Ihren Jobs von den neuesten Iceberg-Funktionen und Leistungsverbesserungen zu profitieren. AWS Glue

- Aktivieren Sie die Spark-Benutzeroberfläche für die Überwachung und das Debuggen — Sie können die [Spark-Benutzeroberfläche](#) auch verwenden, AWS Glue um Ihren Iceberg-Job zu überprüfen, indem Sie die verschiedenen Phasen eines Spark-Jobs in einem gerichteten azyklischen Graphen (DAG) visualisieren und die Jobs im Detail überwachen. Die Spark-Benutzeroberfläche bietet eine effektive Möglichkeit, Iceberg-Jobs sowohl zu beheben als auch zu optimieren. So können Sie beispielsweise Engpass-Phasen identifizieren, in denen es zu häufigem Durcheinander oder zu viel Festplattenverlust kommt, um Optimierungsmöglichkeiten zu identifizieren. Weitere Informationen finden Sie in der Dokumentation unter [Überwachen von Aufträgen mithilfe der Apache Spark-Weboberfläche](#). AWS Glue

Arbeiten mit Apache Iceberg-Tabellen mithilfe von Apache Spark

Dieser Abschnitt bietet einen Überblick über die Verwendung von Apache Spark für die Interaktion mit Iceberg-Tabellen. Bei den Beispielen handelt es sich um Standardcode, der auf Amazon EMR oder ausgeführt werden kann. AWS Glue

Hinweis: Die primäre Schnittstelle für die Interaktion mit Iceberg-Tabellen ist SQL, daher kombinieren die meisten Beispiele Spark SQL mit der API. DataFrames

Iceberg-Tabellen erstellen und schreiben

Sie können Spark SQL und Spark verwenden, um Daten DataFrames zu Iceberg-Tabellen zu erstellen und hinzuzufügen.

Verwenden von Spark SQL

Verwenden Sie standardmäßige Spark-SQL-Anweisungen wie `CREATE TABLE` und `INSERT INTO`, um einen Iceberg-Datensatz zu schreiben.

Unpartitionierte Tabellen

Hier ist ein Beispiel für die Erstellung einer unpartitionierten Iceberg-Tabelle mit Spark SQL:

```
spark.sql(f"""
    CREATE TABLE IF NOT EXISTS {CATALOG_NAME}.{DB_NAME}.{TABLE_NAME}_nopartitions (
        c_customer_sk          int,
        c_customer_id          string,
        c_first_name           string,
        c_last_name            string,
        c_birth_country        string,
        c_email_address        string)
    USING iceberg
    OPTIONS ('format-version'='2')
    """)
```

Verwenden Sie eine Standardanweisung, um Daten in eine unpartitionierte Tabelle einzufügen:
`INSERT INTO`

```
spark.sql(f"""
INSERT INTO {CATALOG_NAME}.{DB_NAME}.{TABLE_NAME}_nopartitions
SELECT c_customer_sk, c_customer_id, c_first_name, c_last_name, c_birth_country,
       c_email_address
FROM another_table
""")
```

Partitionierte Tabellen

Hier ist ein Beispiel für die Erstellung einer partitionierten Iceberg-Tabelle mit Spark SQL:

```
spark.sql(f"""
CREATE TABLE IF NOT EXISTS {CATALOG_NAME}.{DB_NAME}.{TABLE_NAME}_withpartitions (
    c_customer_sk          int,
    c_customer_id         string,
    c_first_name          string,
    c_last_name           string,
    c_birth_country       string,
    c_email_address       string)
USING iceberg
PARTITIONED BY (c_birth_country)
OPTIONS ('format-version'='2')
""")
```

Um Daten mit Spark SQL in eine partitionierte Iceberg-Tabelle einzufügen, führen Sie eine globale Sortierung durch und schreiben dann die Daten:

```
spark.sql(f"""
INSERT INTO {CATALOG_NAME}.{DB_NAME}.{TABLE_NAME}_withpartitions
SELECT c_customer_sk, c_customer_id, c_first_name, c_last_name, c_birth_country,
       c_email_address
FROM another_table
ORDER BY c_birth_country
""")
```

Verwenden der API DataFrames

Um einen Iceberg-Datensatz zu schreiben, können Sie die `DataFrameWriterV2` API verwenden.

Verwenden Sie die Funktion `df.writeTo(t)`, um eine Iceberg-Tabelle zu erstellen und Daten in diese zu schreiben. Wenn die Tabelle existiert, verwenden Sie die `.append()` Funktion. Ist dies

nicht `.create()`. der Fall `.createOrReplace()`, verwenden Sie die folgenden Beispiele use, was einer `.create()` Variante entspricht `CREATE OR REPLACE TABLE AS SELECT`.

Unpartitionierte Tabellen

So erstellen und füllen Sie eine unpartitionierte Iceberg-Tabelle mithilfe der API:

`DataFrameWriterV2`

```
input_data.writeTo(f"{CATALOG_NAME}.{DB_NAME}.{TABLE_NAME}_nopartitions") \  
    .tableProperty("format-version", "2") \  
    .createOrReplace()
```

So fügen Sie mithilfe der API Daten in eine bestehende unpartitionierte Iceberg-Tabelle ein:

`DataFrameWriterV2`

```
input_data.writeTo(f"{CATALOG_NAME}.{DB_NAME}.{TABLE_NAME}_nopartitions") \  
    .append()
```

Partitionierte Tabellen

Um eine partitionierte Iceberg-Tabelle mithilfe der `DataFrameWriterV2` API zu erstellen und aufzufüllen, können Sie eine lokale Sortierung verwenden, um Daten aufzunehmen:

```
input_data.sortWithinPartitions("c_birth_country") \  
    .writeTo(f"{CATALOG_NAME}.{DB_NAME}.{TABLE_NAME}_withpartitions") \  
    .tableProperty("format-version", "2") \  
    .partitionedBy("c_birth_country") \  
    .createOrReplace()
```

Um mithilfe der `DataFrameWriterV2` API Daten in eine partitionierte Iceberg-Tabelle einzufügen, können Sie eine globale Sortierung verwenden, um Daten aufzunehmen:

```
input_data.orderBy("c_birth_country") \  
    .writeTo(f"{CATALOG_NAME}.{DB_NAME}.{TABLE_NAME}_withpartitions") \  
    .append()
```

Daten in Iceberg-Tabellen aktualisieren

Das folgende Beispiel zeigt, wie Daten in einer Iceberg-Tabelle aktualisiert werden. In diesem Beispiel werden alle Zeilen geändert, die eine gerade Zahl in der `c_customer_sk` Spalte haben.

```
spark.sql(f"""
UPDATE {CATALOG_NAME}.{db.name}.{table.name}
SET c_email_address = 'even_row'
WHERE c_customer_sk % 2 == 0
""")
```

Dieser Vorgang verwendet die copy-on-write Standardstrategie, sodass alle betroffenen Datendateien neu geschrieben werden.

Daten in Iceberg-Tabellen werden aktualisiert

Daten aktualisieren bezieht sich auf das Einfügen neuer Datensätze und das Aktualisieren vorhandener Datensätze in einer einzigen Transaktion. Um Daten in eine Iceberg-Tabelle hochzuladen, verwenden Sie die Anweisung. SQL `MERGE INTO`

Im folgenden Beispiel wird der Inhalt der Tabelle `{UPSERT_TABLE_NAME}` innerhalb der Tabelle verschoben: `{TABLE_NAME}`

```
spark.sql(f"""
MERGE INTO {CATALOG_NAME}.{DB_NAME}.{TABLE_NAME} t
USING {UPSERT_TABLE_NAME} s
  ON t.c_customer_id = s.c_customer_id
  WHEN MATCHED THEN UPDATE SET t.c_email_address = s.c_email_address
  WHEN NOT MATCHED THEN INSERT *
""")
```

- Wenn ein Kundendatensatz, der sich in `{UPSERT_TABLE_NAME}` befindet, bereits `{TABLE_NAME}` mit demselben vorhanden ist `c_customer_id`, überschreibt der `{UPSERT_TABLE_NAME}` `c_email_address` Datensatzwert den vorhandenen Wert (Aktualisierungsvorgang).
- Wenn ein Kundendatensatz, der `{UPSERT_TABLE_NAME}` sich in befindet, nicht existiert `{TABLE_NAME}`, wird der `{UPSERT_TABLE_NAME}` Datensatz hinzugefügt `{TABLE_NAME}` (Vorgang einfügen).

Löschen von Daten in Iceberg-Tabellen

Um Daten aus einer Iceberg-Tabelle zu löschen, verwenden Sie den `DELETE FROM` Ausdruck und geben Sie einen Filter an, der den zu löschenden Zeilen entspricht.

```
spark.sql(f"""
DELETE FROM {CATALOG_NAME}.{db.name}.{table.name}
WHERE c_customer_sk % 2 != 0
""")
```

Wenn der Filter auf eine gesamte Partition zutrifft, führt Iceberg eine reine Metadaten-Löschung durch und belässt die Datendateien an Ort und Stelle. Andernfalls werden nur die betroffenen Datendateien neu geschrieben.

Die Methode `delete` verwendet die Datendateien, die von der `WHERE` Klausel betroffen sind, und erstellt eine Kopie davon ohne die gelöschten Datensätze. Anschließend wird ein neuer Tabellen-Snapshot erstellt, der auf die neuen Datendateien verweist. Daher sind die gelöschten Datensätze immer noch in den älteren Snapshots der Tabelle vorhanden. Wenn Sie beispielsweise den vorherigen Snapshot der Tabelle abrufen, werden Ihnen die Daten angezeigt, die Sie gerade gelöscht haben. Informationen zum Entfernen nicht benötigter alter Snapshots mit den zugehörigen Datendateien zu Säuberungszwecken finden Sie im Abschnitt [Dateien mithilfe der Komprimierung verwalten weiter unten in diesem Handbuch](#).

Lesen von Daten

Sie können den aktuellen Status Ihrer Iceberg-Tabellen in Spark sowohl mit Spark SQL als auch lesen. DataFrames

Beispiel mit Spark SQL:

```
spark.sql(f"""
SELECT * FROM {CATALOG_NAME}.{db.name}.{table.name} LIMIT 5
""")
```

Beispiel mit der DataFrames API:

```
df = spark.table(f"{CATALOG_NAME}.{DB_NAME}.{TABLE_NAME}").limit(5)
```

Zeitreisen nutzen

Jeder Schreibvorgang (Einfügen, Aktualisieren, Hochsetzen, Löschen) in einer Iceberg-Tabelle erstellt einen neuen Snapshot. Sie können diese Snapshots dann für Zeitreisen verwenden, um in die Vergangenheit zu reisen und den Status einer Tabelle in der Vergangenheit zu überprüfen.

Informationen zum Abrufen des Verlaufs von Snapshots für Tabellen mithilfe von Werten `snapshot-id` und Zeitangaben finden Sie im Abschnitt [Zugreifen auf Metadaten](#) weiter unten in diesem Handbuch.

Die folgende Zeitreiseabfrage zeigt den Status einer Tabelle auf der Grundlage eines bestimmten `snapshot-id` Status an.

Verwenden von Spark SQL:

```
spark.sql(f"""
SELECT * FROM {CATALOG_NAME}.{DB_NAME}.{TABLE_NAME} VERSION AS OF {snapshot_id}
""")
```

Mithilfe der DataFrames API:

```
df_1st_snapshot_id = spark.read.option("snapshot-id", snapshot_id) \
    .format("iceberg") \
    .load(f"{CATALOG_NAME}.{DB_NAME}.{TABLE_NAME}") \
    .limit(5)
```

Die folgende Zeitreiseabfrage zeigt den Status einer Tabelle auf der Grundlage des letzten Snapshots, der vor einem bestimmten Zeitstempel erstellt wurde, in Millisekunden () an. `as-of-timestamp`

Verwenden von Spark SQL:

```
spark.sql(f"""
SELECT * FROM dev.{db.name}.{table.name} TIMESTAMP AS OF '{snapshot_ts}'
""")
```

Mithilfe der DataFrames API:

```
df_1st_snapshot_ts = spark.read.option("as-of-timestamp", snapshot_ts) \
    .format("iceberg") \
    .load(f"dev.{DB_NAME}.{TABLE_NAME}") \
    .limit(5)
```

Verwendung inkrementeller Abfragen

Sie können Iceberg-Snapshots auch verwenden, um angehängte Daten inkrementell zu lesen.

Hinweis: Derzeit unterstützt dieser Vorgang das Lesen von Daten aus Snapshots. `append` Das Abrufen von Daten aus Operationen wie `replaceoverwrite`, oder wird nicht unterstützt. `delete` Darüber hinaus werden inkrementelle Lesevorgänge in der Spark-SQL-Syntax nicht unterstützt.

Im folgenden Beispiel werden alle Datensätze abgerufen, die an eine Iceberg-Tabelle zwischen dem Snapshot `start-snapshot-id` (exklusiv) und `end-snapshot-id` (inklusive) angehängt wurden.

```
df_incremental = (spark.read.format("iceberg")
    .option("start-snapshot-id", snapshot_id_start)
    .option("end-snapshot-id", snapshot_id_end)
    .load(f"glue_catalog.{DB_NAME}.{TABLE_NAME}")
)
```

Zugreifen auf Metadaten

Iceberg bietet Zugriff auf seine Metadaten über SQL. Sie können auf die Metadaten für jede beliebige Tabelle (`<table_name>`) zugreifen, indem Sie den Namespace abfragen. `<table_name>.<metadata_table>` Eine vollständige Liste der Metadatatabellen finden Sie in der [Iceberg-Dokumentation unter Tabellen](#) überprüfen.

Das folgende Beispiel zeigt, wie Sie auf die Iceberg-Historien-Metadatatabelle zugreifen können, in der der Verlauf der Commits (Änderungen) für eine Iceberg-Tabelle angezeigt wird.

Verwenden von Spark SQL (mit der `%%sql` Magie) von einem Amazon EMR Studio-Notizbuch aus:

```
Spark.sql(f"""
SELECT * FROM {CATALOG_NAME}.{DB_NAME}.{TABLE_NAME}.history LIMIT 5
""")
```

Mithilfe der DataFrames API:

```
spark.read.format("iceberg").load("{CATALOG_NAME}.{DB_NAME}."
{TABLE_NAME}.history").show(5, False)
```

Beispielausgabe:

Type:

Table

Pie

Scatter

Line

Area

Bar

made_current_at	snapshot_id	parent_id	is_current_ancestor
2023-01-09 02:50:17.547000+00:00	7501027970051178613	6598755163776233735	True
2023-01-12 05:39:29.567000+00:00	7069175828427777019	7501027970051178613	True
2023-01-12 05:39:58.807000+00:00	5173022175861138222	7069175828427777019	True
2023-01-12 05:40:18.499000+00:00	3703414997660223390	5173022175861138222	True
2023-01-12 05:40:41.827000+00:00	3807904412292252460	3703414997660223390	True

Arbeiten mit Apache Iceberg-Tabellen mithilfe von Amazon Data Firehose

Amazon Data Firehose ist ein serverloser Service ohne Code für die Bereitstellung von Datenströmen aus über 20 Quellen wie AWS WAF Logs, Amazon CloudWatch Logs, Amazon Kinesis Data Streams und Amazon Managed Streaming for Apache Kafka (Amazon MSK) an Ziele wie Amazon S3, Amazon Redshift, Snowflake und Splunk. AWS IoT

Sie können Firehose verwenden, um Streaming-Daten direkt an Apache Iceberg-Tabellen in Amazon S3 zu übertragen. Mit Firehose können Sie Datensätze aus einem einzelnen Stream in verschiedene Apache Iceberg-Tabellen weiterleiten und automatisch Einfüge-, Aktualisierungs- und Löschvorgänge auf Datensätze in den Tabellen anwenden. Firehose garantiert die exakte Einmallieferung an die Tische von Iceberg. Diese Funktion erfordert die Verwendung von AWS Glue Data Catalog

Firehose kann Streaming-Daten auch direkt an Amazon S3-Tabellen liefern. Diese Tabellen bieten Speicher, der für umfangreiche Analyse-Workloads optimiert ist, und enthalten Funktionen, die die Abfrageleistung kontinuierlich verbessern und die Speicherkosten für tabellarische Daten senken.

Informationen darüber, wie Sie einen Firehose-Stream einrichten, um Daten an Apache Iceberg-Tabellen zu liefern, finden Sie unter Den [Firehose-Stream einrichten in der Firehose-Dokumentation](#).

Arbeiten mit Apache Iceberg-Tabellen mithilfe von Amazon Athena SQL

Amazon Athena bietet integrierte Unterstützung für Apache Iceberg und erfordert keine zusätzlichen Schritte oder Konfigurationen. Dieser Abschnitt bietet einen detaillierten Überblick über die unterstützten Funktionen und allgemeine Anleitungen zur Verwendung von Athena für die Interaktion mit Iceberg-Tabellen.

Versions- und Funktionskompatibilität

Unterstützung der Iceberg-Tabellenspezifikationen

Die Apache Iceberg-Tabellenspezifikation spezifiziert, wie sich Iceberg-Tabellen verhalten sollen. Athena unterstützt das Tabellenformat Version 2, sodass jede Iceberg-Tabelle, die Sie mit der Konsole, der CLI oder dem SDK erstellen, grundsätzlich diese Version verwendet.

Wenn Sie eine Iceberg-Tabelle verwenden, die mit einer anderen Engine wie Apache Spark auf Amazon EMR oder erstellt wurde AWS Glue, stellen Sie sicher, dass Sie die Tabellenformatversion mithilfe der [Tabelleneigenschaften](#) festlegen. Als Referenz finden Sie den Abschnitt [Erstellen und Schreiben von Iceberg-Tabellen weiter](#) oben in diesem Handbuch.

Unterstützung für Iceberg-Funktionen

Sie können Athena verwenden, um aus Eisberg-Tabellen zu lesen und in sie zu schreiben. Wenn Sie Daten mithilfe der DELETE FROM AnweisungenUPDATE, und ändernMERGE INTO, unterstützt Athena nur merge-on-read den Modus. Diese Eigenschaft kann nicht geändert werden. Um Daten mit zu aktualisieren oder zu löschen copy-on-write, müssen Sie andere Engines wie Apache Spark auf Amazon EMR oder AWS Glue verwenden. In der folgenden Tabelle wird die Unterstützung der Iceberg-Funktionen in Athena zusammengefasst.

		DDL-Unterstützung		DML-Unterstützung		AWS Lake Formation aus Sicherheitsgründen (optional)
	Tabellenformat	Create table	Schemaentwicklung	Lesen von Daten	Daten schreiben	Zugriffskontrolle für Zeilen und Spalten
Amazon Athena	Version 2	✓	✓	✓	X Copy-on-write	✓
					✓ Merge-on-read	✓

Note

- Athena unterstützt keine inkrementellen Abfragen.
- In Athena verwenden Aktualisierungs-, Lösch- und Zusammenführungsoperationen immer standardmäßig Merge on Read (MoR), unabhängig von den Copy-on-Write-Einstellungen (CoW) in den Tabelleneigenschaften, da CoW nicht unterstützt wird.

Mit Iceberg-Tabellen arbeiten

Einen schnellen Einstieg in die Verwendung von Iceberg in Athena finden Sie im Abschnitt [Erste Schritte mit Iceberg-Tabellen in Athena SQL weiter oben](#) in diesem Handbuch.

In der folgenden Tabelle sind Einschränkungen und Empfehlungen aufgeführt.

Szenario	Einschränkung	Empfehlung
Tabelle DDL-Generierung	Iceberg-Tabellen, die mit anderen Engines erstellt wurden, können Eigenschaften haben, die in Athena nicht verfügbar sind. Für diese Tabellen ist es nicht möglich, die DDL zu generieren.	Verwenden Sie die entsprechende Anweisung in der Engine, die die Tabelle erstellt hat (z. B. die <code>SHOW CREATE TABLE</code> Anweisung für Spark).
Zufällige Amazon S3 S3-Präfixe in Objekten, die in eine Iceberg-Tabelle geschrieben werden	In Iceberg-Tabellen, die mit Athena erstellt wurden, ist die <code>write.object-storage.enabled</code> Eigenschaft standardmäßig aktiviert.	Um dieses Verhalten zu deaktivieren und die volle Kontrolle über die Eigenschaften von Iceberg-Tabellen zu erlangen, erstellen Sie eine Iceberg-Tabelle mit einer anderen Engine wie Spark on Amazon EMR oder AWS Glue
Inkrementelle Abfragen	Wird derzeit in Athena nicht unterstützt.	Um inkrementelle Abfragen zu verwenden, um inkrementelle Datenerfassungspipelines zu aktivieren, verwenden Sie Spark auf Amazon EMR oder AWS Glue

Migration vorhandener Tabellen zu Apache Iceberg

Um Ihre aktuellen Athena- oder AWS Glue Tabellen (auch bekannt als Hive-Tabellen) in das Iceberg-Format zu migrieren, können Sie entweder die direkte oder die vollständige Datenmigration verwenden:

- Bei der direkten Migration werden die Metadateien von Iceberg zusätzlich zu den vorhandenen Datendateien generiert.
- Bei der vollständigen Datenmigration wird die Iceberg-Metadatenebene erstellt und außerdem vorhandene Datendateien aus der Originaltabelle in die neue Iceberg-Tabelle umgeschrieben.

Die folgenden Abschnitte bieten einen Überblick über die für die Migration APIs verfügbaren Tabellen sowie Anleitungen zur Auswahl einer Migrationsstrategie. Weitere Informationen zu diesen beiden Strategien finden Sie im Abschnitt [Tabellenmigration](#) in der Iceberg-Dokumentation.

Direkte Migration

Durch die direkte Migration entfällt die Notwendigkeit, alle Datendateien neu zu schreiben. Stattdessen werden Iceberg-Metadateien generiert und mit Ihren vorhandenen Datendateien verknüpft. Iceberg bietet drei Optionen für die Implementierung einer In-Place-Migration:

- Verwenden Sie das snapshot Verfahren, wie in den Abschnitten [Snapshot-Tabelle](#) und [Spark-Verfahren: Snapshot](#) in der Iceberg-Dokumentation beschrieben.
- Verwenden Sie das `add_files` Verfahren, wie in den Abschnitten [Dateien hinzufügen](#) und [Spark-Verfahren: add_files](#) in der Iceberg-Dokumentation beschrieben.
- Verwenden Sie das `migrate` Verfahren, wie in den Abschnitten [Tabelle migrieren](#) und [Spark-Verfahren: Migrieren](#) in der Iceberg-Dokumentation beschrieben.

Derzeit funktioniert das Migrationsverfahren nicht direkt mit dem AWS Glue Data Catalog— es funktioniert nur mit dem Hive-Metastore. Wenn Sie das `migrate` Verfahren anstelle von `snapshot` oder verwenden möchten `add_files`, können Sie einen temporären Amazon EMR-Cluster mit dem Hive Metastore (HMS) verwenden. Für diesen Ansatz ist Iceberg Version 1.2 oder höher erforderlich.

Nehmen wir an, Sie möchten die folgende Hive-Tabelle erstellen:

The screenshot shows the AWS Athena console interface. On the left, there's a sidebar with 'Data source' set to 'AwsDataCatalog' and 'Database' set to 'iceberg_db'. Below that, a list of tables and views is shown, including 'hive_table', 'table_w_rowgroupsize_134217728', 'table_w_rowgroupsize_16777216', and 'upsert_batch'. The main area displays a query execution for 'Query 24'. The SQL query is 'SELECT * FROM \"iceberg_db\".\"hive_table\" limit 10;'. The query status is 'Completed' with a run time of 356 ms and 0.06 KB of data scanned. The results table has one row with 'id' 1 and 'data' 'a'.

Sie können diese Hive-Tabelle erstellen, indem Sie diesen Code in der Athena-Konsole ausführen:

```
CREATE EXTERNAL TABLE 'hive_table'(
  'id' bigint,
  'data' string)
USING parquet
LOCATION
  's3://datalake-xxxx/aws_workshop/iceberg_db/hive_table'

INSERT INTO iceberg_db.hive_table VALUES (1, 'a')
```

Wenn Ihre Hive-Tabelle partitioniert ist, fügen Sie die Partitionsanweisung hinzu und fügen Sie die Partitionen gemäß den Hive-Anforderungen hinzu.

```
ALTER TABLE default.placeholder_table_for_migration ADD
PARTITION (date = '2023-10-10')
```

Schritte:

1. Erstellen Sie einen Amazon EMR-Cluster, ohne die AWS Glue Data Catalog Integration zu aktivieren — aktivieren Sie also nicht die Kontrollkästchen für Hive- oder Spark-Tabellenmetadaten. Das liegt daran, dass Sie für diese Problemumgebung den nativen Hive Metastore (HMS) verwenden werden, der im Cluster verfügbar ist.

AWS Glue Data Catalogue settings

Use the AWS Glue Data Catalog to provide an external metastore for your application.

- Use for Hive table metadata
- Use for Spark table metadata

- Konfigurieren Sie die Spark-Sitzung so, dass sie die Iceberg Hive-Katalogimplementierung verwendet.

```
"spark.sql.extensions": "org.apache.iceberg.spark.extensions.IcebergSparkSessionExtensions",
"spark.sql.catalog.spark_catalog": "org.apache.iceberg.spark.SparkSessionCatalog",
"spark.sql.catalog.spark_catalog.type": "hive",
```

- Stellen Sie sicher, dass Ihr Amazon EMR-Cluster nicht mit dem verbunden ist, AWS Glue Data Catalog indem Sie `show databases` oder `show tables` ausführen.

```
[2]: %%sql
show databases
```

Last executed at 2023-07-05 12:24:26 in 35.03s

Starting Spark application

ID	YARN Application ID	Kind	State	Spark UI	Driver log	User	Current session?
1	application_1686667730124_0002	pyspark	idle	Link	Link	None	✓

SparkSession available as 'spark'.

Type:

namespace

default

- Registrieren Sie die Hive-Tabelle im Hive-Metastore Ihres Amazon EMR-Clusters und verwenden Sie dann das Iceberg-Verfahren. `migrate`

Register the Hive table in this local HMS catalog pointing to the S3 location where the files from the original Hive tables exist

```
%%sql -q
CREATE TABLE default.placeholder_hive_table (id bigint NOT NULL, data string)
USING parquet
LOCATION 's3://datalake-743490154766/aws_workshop/iceberg_db/hive_table/'
```

Last executed at 2023-07-05 12:55:19 in 3.25s

```
%%sql
select * from default.placeholder_hive_table limit 5
```

Last executed at 2023-07-05 12:57:13 in 7.43s

Type: Table Pie Scatter Line Area Bar

id	data
1	a

Once the Hive table is registered in this local HMS catalog, you can use Iceberg's migrate procedure.

```
spark.sql("CALL spark_catalog.system.migrate('default.placeholder_hive_table')")
```

Last executed at 2023-07-05 13:00:06 in 3.27s

► Spark Job Progress

DataFrame[migrated_files_count: bigint]

```
%%sql
show tables from default
```

Last executed at 2023-07-05 13:00:49 in 7.42s

Type: Table Pie Scatter Line Area Bar

namespace	tableName	isTemporary
default	placeholder_hive_table	False
default	placeholder_hive_table_backup_	False

Bei diesem Verfahren werden die Iceberg-Metadatendateien am selben Speicherort wie die Hive-Tabelle erstellt.

5. Registrieren Sie die migrierte Iceberg-Tabelle in der AWS Glue Data Catalog
6. Wechseln Sie zurück zu einem Amazon EMR-Cluster, für den die AWS Glue Data Catalog Integration aktiviert ist.

AWS Glue Data Catalogue settings

Use the AWS Glue Data Catalog to provide an external metastore for your application.

- Use for Hive table metadata
- Use for Spark table metadata

7. Verwenden Sie die folgende Iceberg-Konfiguration in der Spark-Sitzung.

```
"spark.sql.extensions": "org.apache.iceberg.spark.extensions.IcebergSparkSessionExtensions",
  "spark.sql.catalog.glue_catalog": "org.apache.iceberg.spark.SparkCatalog",
  "spark.sql.catalog.glue_catalog.warehouse": "s3://datalake-xxxx/
aws_workshop",
  "spark.sql.catalog.glue_catalog.catalog-impl":
"org.apache.iceberg.aws.glue.GlueCatalog",
  "spark.sql.catalog.glue_catalog.io-impl":
"org.apache.iceberg.aws.s3.S3FileIO",
```

Sie können diese Tabelle jetzt von Amazon EMR oder Athena AWS Glue abfragen.

```

%%sql
show tables from iceberg_db
Last executed at 2023-07-05 13:10:50 in 7.44s

Type: Table Pie Scatter Line Area Bar

namespace      tableName      isTemporary
iceberg_db      hive_table      False
iceberg_db      table_w_rowgroupsize_134217728      False
iceberg_db      table_w_rowgroupsize_16777216      False
iceberg_db      upsert_batch      False
iceberg_db      ws_webpage_pk_partitioned_140gb_trino      False

%%bash
aws s3 ls s3://datalake-743490154766/aws_workshop/iceberg_db/hive_table/metadata/
Last executed at 2023-07-05 13:10:20 in 488ms
2023-07-05 12:00:07      2239 00000-12a20051-6a3f-4b46-bae1-985f6df254db.metadata.json
2023-07-05 12:00:07      5802 b3d40480-0cb9-4cea-a4af-94c40a123689-m0.avro
2023-07-05 12:00:07      3781 snap-6104693268717769849-1-b3d40480-0cb9-4cea-a4af-94c40a123689.avro

metadata_file = "s3://datalake-743490154766/aws_workshop/iceberg_db/hive_table/metadata/00000-12a20051-6a3f-4b46-bae1-985f6df254db.metadata.json"
Last executed at 2023-07-05 13:11:46 in 49ms

spark.sql(f"CALL glue_catalog.system.register_table('iceberg_db.migrated_iceberg_table',{metadata_file})")
Last executed at 2023-07-05 13:12:27 in 3.32s

> Spark Job Progress

DataFrame[current_snapshot_id: bigint, total_records_count: bigint, total_data_files_count: bigint]

%%sql -q
alter table glue_catalog.iceberg_db.migrated_iceberg_table SET TBLPROPERTIES('format-version'='2')
Last executed at 2023-07-05 13:12:33 in 2.24s

%%sql
select * from glue_catalog.iceberg_db.migrated_iceberg_table limit 5
Last executed at 2023-07-05 13:12:44 in 7.42s

Type: Table Pie Scatter Line Area Bar

id data
1 a

```

Vollständige Datenmigration

Bei der vollständigen Datenmigration werden sowohl die Datendateien als auch die Metadaten neu erstellt. Dieser Ansatz dauert länger und erfordert zusätzliche Rechenressourcen im Vergleich zur direkten Migration. Diese Option trägt jedoch zur Verbesserung der Tabellenqualität bei: Sie können die Daten validieren, Schema- und Partitionsänderungen vornehmen, die Daten sortieren usw. Verwenden Sie eine der folgenden Optionen, um eine vollständige Datenmigration zu implementieren:

- Verwenden Sie die `CREATE TABLE ... AS SELECT (CTAS)`-Anweisung in Spark auf Amazon EMR oder AWS Glue Athena. Sie können die Partitionspezifikation und die

Tabelleneigenschaften für die neue Iceberg-Tabelle mithilfe der UND-Klauseln festlegen.

`PARTITIONED BY TBLPROPERTIES` Sie können das Schema und die Partitionierung für die neue Tabelle an Ihre Bedürfnisse anpassen, anstatt sie einfach von der Quelltable zu erben.

- Lesen Sie aus der Quelltable und schreiben Sie die Daten als neue Iceberg-Tabelle, indem Sie Spark auf Amazon EMR verwenden oder AWS Glue (siehe [Erstellen einer Tabelle](#) in der Iceberg-Dokumentation).

Auswahl einer Migrationsstrategie

Beachten Sie bei der Auswahl der besten Migrationsstrategie die Fragen in der folgenden Tabelle.

Frage	Empfehlung
Was ist das Datendateiformat (z. B. CSV oder Apache Parquet)?	<ul style="list-style-type: none"> • Ziehen Sie eine direkte Migration in Betracht, wenn Ihr Tabellendateiformat Parquet, ORC oder Avro ist. • Verwenden Sie für andere Formate wie CSV, JSON usw. die vollständige Datenmigration.
Möchten Sie das Tabellenschema aktualisieren oder konsolidieren?	<ul style="list-style-type: none"> • Wenn Sie das Tabellenschema mithilfe der systemeigenen Funktionen von Iceberg weiterentwickeln möchten, sollten Sie eine direkte Migration in Betracht ziehen. Sie können beispielsweise Spalten nach der Migration umbenennen. (Das Schema kann in der Iceberg-Metadatenebene geändert werden.) • Wenn Sie ganze Spalten aus Datendateien löschen möchten, empfehlen wir Ihnen, die vollständige Datenmigration zu verwenden.
Würde die Tabelle von einer Änderung der Partitionsstrategie profitieren?	<ul style="list-style-type: none"> • Wenn der Partitionierungsansatz von Iceberg Ihren Anforderungen entspricht (z. B. werden neue Daten mithilfe des neuen Partition

Frage

Würde es für die Tabelle von Vorteil sein, die Strategie für die Sortierreihenfolge hinzuzufügen oder zu ändern?

Empfehlung

layouts gespeichert, während die vorhandenen Partitionen unverändert bleiben), sollten Sie eine direkte Migration in Betracht ziehen.

- Wenn Sie versteckte Partitionen in Ihrer Tabelle verwenden möchten, sollten Sie eine vollständige Datenmigration in Betracht ziehen. Weitere Informationen zu versteckten Partitionen finden Sie im Abschnitt [Bewährte Methoden](#).
- Um die Sortierreihenfolge Ihrer Daten hinzuzufügen oder zu ändern, muss der Datensatz neu geschrieben werden. In diesem Fall sollten Sie die vollständige Datenmigration in Betracht ziehen.
- Bei großen Tabellen, bei denen es unerschwinglich teuer ist, alle Tabellenpartitionen neu zu schreiben, sollten Sie die direkte Migration in Betracht ziehen und die Komprimierung (mit aktivierter Sortierung) für die Partitionen ausführen, auf die am häufigsten zugegriffen wird.

Frage

Enthält die Tabelle viele kleine Dateien?

Empfehlung

- Um kleine Dateien zu größeren Dateien zusammenzuführen, muss der Datensatz neu geschrieben werden. In diesem Fall sollten Sie die vollständige Datenmigration in Betracht ziehen.
- Bei großen Tabellen, bei denen es unerschwinglich teuer ist, alle Tabellenpartitionen neu zu schreiben, sollten Sie die direkte Migration in Betracht ziehen und die Komprimierung (mit aktivierter Sortierung) für die Partitionen ausführen, auf die am häufigsten zugegriffen wird.

Bewährte Methoden zur Optimierung von Apache Iceberg-Workloads

Iceberg ist ein Tabellenformat, das entwickelt wurde, um das Data Lake-Management zu vereinfachen und die Workload-Leistung zu verbessern. In verschiedenen Anwendungsfällen können unterschiedliche Aspekte wie Kosten, Leseleistung, Schreibleistung oder Datenspeicherung priorisiert werden. Daher bietet Iceberg Konfigurationsoptionen, um diese Kompromisse zu bewältigen. Dieser Abschnitt bietet Einblicke in die Optimierung und Feinabstimmung Ihrer Iceberg-Workloads, um Ihre Anforderungen zu erfüllen.

Themen

- [Allgemeine bewährte Methoden](#)
- [Optimierung der Leseleistung](#)
- [Optimierung der Schreibleistung](#)
- [Optimierung des Speichers](#)
- [Pflege von Tabellen mithilfe von Komprimierung](#)
- [Verwenden von Iceberg-Workloads in Amazon S3](#)

Allgemeine bewährte Methoden

Unabhängig von Ihrem Anwendungsfall empfehlen wir Ihnen, diese allgemeinen bewährten Methoden zu befolgen AWS, wenn Sie Apache Iceberg auf verwenden.

- Verwenden Sie Version 2 des Iceberg-Formats.

Athena verwendet standardmäßig Version 2 des Iceberg-Formats.

Wenn Sie Spark auf Amazon EMR verwenden oder AWS Glue Iceberg-Tabellen erstellen, geben Sie die Formatversion an, wie in der [Iceberg-Dokumentation](#) beschrieben.

- Verwenden Sie das AWS Glue Data Catalog als Ihren Datenkatalog.

Athena verwendet AWS Glue Data Catalog standardmäßig das.

Wenn Sie Spark auf Amazon EMR verwenden oder AWS Glue mit Iceberg arbeiten, fügen Sie Ihrer Spark-Sitzung die folgende Konfiguration hinzu, um den AWS Glue Glue-Datenkatalog zu

verwenden. Weitere Informationen finden Sie im Abschnitt [Spark-Konfigurationen für Iceberg in AWS Glue](#) weiter oben in diesem Handbuch.

```
"spark.sql.catalog.<your_catalog_name>.catalog-impl":  
"org.apache.iceberg.aws.glue.GlueCatalog"
```

- Verwenden Sie den AWS Glue Data Catalog AS-Lock-Manager.

Athena verwendet standardmäßig den AWS Glue Data Catalog AS-Lock-Manager für Iceberg-Tabellen.

Wenn Sie Spark auf Amazon EMR verwenden oder AWS Glue mit Iceberg arbeiten, stellen Sie sicher, dass Sie Ihre Spark-Sitzungskonfiguration so konfigurieren, dass sie den AWS Glue Data Catalog AS-Lock-Manager verwendet. Weitere Informationen finden Sie unter [Optimistic Locking](#) in der Iceberg-Dokumentation.

- Verwenden Sie die Zstandard-Komprimierung (ZSTD).

Der Standard-Komprimierungscodec von Iceberg ist gzip, der mithilfe der Tabelleneigenschaft geändert werden kann. `write.<file_type>.compression-codec` Athena verwendet ZSTD bereits als Standard-Komprimierungscodec für Iceberg-Tabellen.

Im Allgemeinen empfehlen wir die Verwendung des ZSTD-Komprimierungscodecs, da er ein Gleichgewicht zwischen GZIP und Snappy herstellt und eine gute Lese-/Schreibleistung bietet, ohne die Komprimierungsrate zu beeinträchtigen. Darüber hinaus können die Komprimierungsstufen an Ihre Bedürfnisse angepasst werden. Weitere Informationen finden Sie unter [ZSTD-Komprimierungsstufen in Athena in](#) der Athena-Dokumentation.

Snappy bietet zwar insgesamt die beste Lese- und Schreibleistung, hat aber ein niedrigeres Komprimierungsverhältnis als GZIP und ZSTD. Wenn Sie der Leistung Priorität einräumen — auch wenn das bedeutet, größere Datenmengen in Amazon S3 zu speichern — ist Snappy möglicherweise die optimale Wahl.

Optimierung der Leseleistung

In diesem Abschnitt werden Tabelleneigenschaften beschrieben, die Sie unabhängig von der Engine anpassen können, um die Leseleistung zu optimieren.

Partitionierung

Wie bei Hive-Tabellen verwendet Iceberg Partitionen als primäre Indizierungsebene, um das Lesen unnötiger Metadaten- und Datendateien zu vermeiden. Spaltenstatistiken werden auch als sekundäre Indizierungsebene berücksichtigt, um die Abfrageplanung weiter zu verbessern, was zu einer besseren Gesamtausführungszeit führt.

Ihre Daten partitionieren

Um die Datenmenge zu reduzieren, die bei der Abfrage von Iceberg-Tabellen gescannt wird, wählen Sie eine ausgewogene Partitionsstrategie, die Ihren erwarteten Lesemustern entspricht:

- Identifizieren Sie Spalten, die häufig in Abfragen verwendet werden. Dies sind ideale Partitionierungskandidaten. Wenn Sie beispielsweise normalerweise Daten von einem bestimmten Tag abfragen, wäre ein natürliches Beispiel für eine Partitionsspalte eine Datumsspalte.
- Wählen Sie eine Partitionsspalte mit niedriger Kardinalität, um zu vermeiden, dass zu viele Partitionen erstellt werden. Zu viele Partitionen können die Anzahl der Dateien in der Tabelle erhöhen, was sich negativ auf die Abfrageleistung auswirken kann. Als Faustregel gilt, dass „zu viele Partitionen“ als ein Szenario definiert werden können, in dem die Datengröße in den meisten Partitionen weniger als das Zwei- bis Fünffache des von `target-file-size-bytes` festgelegten Werts beträgt.

Note

Wenn Sie normalerweise Filter für eine Spalte mit hoher Kardinalität verwenden (z. B. eine `id` Spalte, die Tausende von Werten haben kann), verwenden Sie die Funktion für versteckte Partitionierung von Iceberg mit Bucket-Transformationen, wie im nächsten Abschnitt erklärt.

Verwenden Sie versteckte Partitionierung

Wenn Ihre Abfragen häufig nach einer Ableitung einer Tabellenspalte filtern, verwenden Sie versteckte Partitionen, anstatt explizit neue Spalten zu erstellen, die als Partitionen verwendet werden können. Weitere Informationen zu dieser Funktion finden Sie in der [Iceberg-Dokumentation](#).

Verwenden Sie beispielsweise in einem Datensatz mit einer Zeitstempelspalte (z. B. `2023-01-01 09:00:00`) Partitionstransformationen, um den Datumsanteil aus dem Zeitstempel zu extrahieren

und diese Partitionen im Handumdrehen zu erstellen, anstatt eine neue Spalte mit dem analysierten Datum zu erstellen (z. B. 2023-01-01).

Die häufigsten Anwendungsfälle für versteckte Partitionierung sind:

- Partitionierung nach Datum oder Uhrzeit, wenn die Daten eine Zeitstempelspalte haben. Iceberg bietet mehrere Transformationen, um die Datums- oder Uhrzeitteile eines Zeitstempels zu extrahieren.
- Partitionierung nach einer Hash-Funktion einer Spalte, wenn die partitionierende Spalte eine hohe Kardinalität aufweist und zu vielen Partitionen führen würde. Die Bucket-Transformation von Iceberg gruppiert mehrere Partitionswerte zu weniger versteckten (Bucket-) Partitionen, indem Hash-Funktionen für die Partitionierungsspalte verwendet werden.

Einen Überblick über alle [verfügbaren Partitionstransformationen](#) finden Sie unter Partitionstransformationen in der Iceberg-Dokumentation.

Spalten, die für versteckte Partitionierung verwendet werden, können durch die Verwendung regulärer SQL-Funktionen wie `year()` `month()` Prädikate können auch mit Operatoren wie `BETWEEN AND`

Note

Iceberg kann keine Partitionsbereinigung für Funktionen durchführen, die einen anderen Datentyp ergeben, z. B. `substring(event_time, 1, 10) = '2022-01-01'`

Verwenden Sie Partition Evolution

Verwenden Sie die [Partitionsentwicklung von Iceberg](#), wenn die bestehende Partitionsstrategie nicht optimal ist. Wenn Sie beispielsweise stündliche Partitionen wählen, die sich als zu klein herausstellen (jeweils nur ein paar Megabyte), sollten Sie eine Umstellung auf tägliche oder monatliche Partitionen in Betracht ziehen.

Sie können diesen Ansatz verwenden, wenn die beste Partitionsstrategie für eine Tabelle zunächst unklar ist und Sie Ihre Partitionierungsstrategie verfeinern möchten, wenn Sie mehr Erkenntnisse gewinnen. Eine weitere effektive Anwendung der Partitionsentwicklung besteht darin, dass sich die Datenmengen ändern und die aktuelle Partitionierungsstrategie im Laufe der Zeit an Effektivität verliert.

Anweisungen zur Weiterentwicklung von Partitionen finden Sie unter [ALTER TABLE SQL-Erweiterungen](#) in der Iceberg-Dokumentation.

Optimieren der Dateigrößen

Zur Optimierung der Abfrageleistung gehört die Minimierung der Anzahl kleiner Dateien in Ihren Tabellen. Für eine gute Abfrageleistung empfehlen wir generell, Parquet- und ORC-Dateien größer als 100 MB zu verwenden.

Die Dateigröße wirkt sich auch auf die Abfrageplanung für Iceberg-Tabellen aus. Mit zunehmender Anzahl von Dateien in einer Tabelle nimmt auch die Größe der Metadatendateien zu. Größere Metadatendateien können zu einer langsameren Abfrageplanung führen. Wenn die Tabellengröße zunimmt, sollten Sie daher die Dateigröße erhöhen, um die exponentielle Erweiterung der Metadaten zu verringern.

Verwenden Sie die folgenden bewährten Methoden, um Dateien mit der richtigen Größe in Iceberg-Tabellen zu erstellen.

Legen Sie die Größe der Zieldatei und der Zeilengruppe fest

Iceberg bietet die folgenden wichtigen Konfigurationsparameter zur Optimierung des Layouts der Datendatei. Wir empfehlen, dass Sie diese Parameter verwenden, um die Zieldateigröße und die Zeilengruppen- oder Strike-Größe festzulegen.

Parameter	Standardwert	Kommentar
<code>write.target-file-size-bytes</code>	512 MB	Dieser Parameter gibt die maximale Dateigröße an, die Iceberg erstellen wird. Bestimmte Dateien können jedoch mit einer kleineren Größe als diesem Limit geschrieben werden.
<code>write.parquet.row-group-size-bytes</code>	128 MB	Sowohl Parquet als auch ORC speichern Daten in Blöcken, sodass Engines bei einigen Vorgängen vermeiden können, die gesamte Datei zu lesen.

Parameter	Standardwert	Kommentar
<code>write.orc.stripe-size-bytes</code>	64 MB	
<code>write.distribution-mode</code>	Keine, für Iceberg Version 1.1 und niedriger Hash, beginnend mit Iceberg Version 1.2	Iceberg fordert Spark auf, Daten zwischen seinen Aufgaben zu sortieren, bevor sie in den Speicher geschrieben werden.

- Basierend auf Ihrer erwarteten Tabellengröße sollten Sie sich an diese allgemeinen Richtlinien halten:
 - Kleine Tabellen (bis zu einigen Gigabyte) — Reduzieren Sie die Zieldateigröße auf 128 MB. Reduzieren Sie auch die Zeilengruppen- oder Stripe-Größe (z. B. auf 8 oder 16 MB).
 - Mittlere bis große Tabellen (von einigen Gigabyte bis zu Hunderten von Gigabyte) — Die Standardwerte sind ein guter Ausgangspunkt für diese Tabellen. Wenn Ihre Abfragen sehr selektiv sind, passen Sie die Zeilengruppen- oder Stripe-Größe an (z. B. auf 16 MB).
 - Sehr große Tabellen (Hunderte von Gigabyte oder Terabyte) — Erhöhen Sie die Zieldateigröße auf 1024 MB oder mehr und erwägen Sie, die Zeilengruppen- oder Stripe-Größe zu erhöhen, wenn Ihre Abfragen normalerweise große Datenmengen abrufen.
- Um sicherzustellen, dass Spark-Anwendungen, die in Iceberg-Tabellen schreiben, Dateien mit der entsprechenden Größe erstellen, setzen Sie die `write.distribution-mode` Eigenschaft entweder auf `oder` oder `hash range`. Eine ausführliche Erklärung des Unterschieds zwischen diesen Modi finden Sie in der [Iceberg-Dokumentation unter Writing Distribution Modes](#).

Dies sind allgemeine Richtlinien. Wir empfehlen Ihnen, Tests durchzuführen, um die am besten geeigneten Werte für Ihre spezifischen Tabellen und Workloads zu ermitteln.

Führen Sie eine regelmäßige Komprimierung durch

Die Konfigurationen in der vorherigen Tabelle legen eine maximale Dateigröße fest, die Schreibaufgaben erzeugen können, garantieren jedoch nicht, dass Dateien diese Größe haben. Führen Sie die Komprimierung regelmäßig durch, um kleine Dateien zu größeren Dateien zusammenzufassen, um die richtigen Dateigrößen sicherzustellen. Eine ausführliche Anleitung

zum Ausführen der Komprimierung finden Sie weiter unten in diesem [Handbuch unter Iceberg-Komprimierung](#).

Optimieren Sie die Spaltenstatistiken

Iceberg verwendet Spaltenstatistiken, um Dateien zu bereinigen. Dadurch wird die Abfrageleistung verbessert, da die Datenmenge, die durch Abfragen gescannt wird, reduziert wird. Um von Spaltenstatistiken zu profitieren, stellen Sie sicher, dass Iceberg Statistiken für alle Spalten sammelt, die häufig in Abfragefiltern verwendet werden.

Standardmäßig sammelt Iceberg Statistiken nur für die [ersten 100 Spalten in jeder Tabelle](#), wie in der Tabelleneigenschaft definiert. `write.metadata.metrics.max-inferred-column-defaults`. Wenn Ihre Tabelle mehr als 100 Spalten hat und Ihre Abfragen häufig auf Spalten außerhalb der ersten 100 Spalten verweisen (z. B. wenn Sie Abfragen haben, die nach Spalte 132 filtern), stellen Sie sicher, dass Iceberg Statistiken zu diesen Spalten sammelt. Es gibt zwei Möglichkeiten, dies zu erreichen:

- Wenn Sie die Iceberg-Tabelle erstellen, ordnen Sie die Spalten neu an, sodass die Spalten, die Sie für Abfragen benötigen, in den angegebenen Spaltenbereich fallen `write.metadata.metrics.max-inferred-column-defaults` (Standard ist 100).

Hinweis: Wenn Sie keine Statistiken für 100 Spalten benötigen, können Sie die `write.metadata.metrics.max-inferred-column-defaults` Konfiguration auf einen gewünschten Wert (z. B. 20) anpassen und die Spalten neu anordnen, sodass die Spalten, die Sie zum Lesen und Schreiben von Abfragen benötigen, in die ersten 20 Spalten auf der linken Seite des Datensatzes fallen.

- Wenn Sie in Abfragefiltern nur wenige Spalten verwenden, können Sie die allgemeine Eigenschaft für die Erfassung von Metriken deaktivieren und selektiv einzelne Spalten auswählen, für die Statistiken erfasst werden sollen, wie in diesem Beispiel gezeigt:

```
.tableProperty("write.metadata.metrics.default", "none")
.tableProperty("write.metadata.metrics.column.my_col_a", "full")
.tableProperty("write.metadata.metrics.column.my_col_b", "full")
```

Hinweis: Spaltenstatistiken sind am effektivsten, wenn die Daten nach diesen Spalten sortiert sind. Weitere Informationen finden Sie im Abschnitt „[Sortierreihenfolge festlegen](#)“ weiter unten in diesem Handbuch.

Wählen Sie die richtige Aktualisierungsstrategie

Verwenden Sie eine copy-on-write Strategie zur Optimierung der Leseleistung, wenn langsamere Schreibvorgänge für Ihren Anwendungsfall akzeptabel sind. Dies ist die von Iceberg verwendete Standardstrategie.

Copy-on-write führt zu einer besseren Leseleistung, da Dateien leseoptimiert direkt in den Speicher geschrieben werden. Im Vergleich merge-on-read zu dauert jeder Schreibvorgang jedoch länger und verbraucht mehr Rechenressourcen. Dies stellt einen klassischen Kompromiss zwischen Lese- und Schreiblatenz dar. In der Regel copy-on-write ist dies ideal für Anwendungsfälle, in denen die meisten Aktualisierungen in denselben Tabellenpartitionen gespeichert sind (z. B. für tägliche Batchladevorgänge).

Copy-on-write Konfigurationen (`write.update.modewrite.delete.mode`, `undwrite.merge.mode`) können auf Tabellenebene oder unabhängig voneinander auf der Anwendungsseite festgelegt werden.

Verwenden Sie die STD-Komprimierung

Sie können den von Iceberg verwendeten Komprimierungscodec mithilfe der Tabelleneigenschaft ändern. `write.<file_type>.compression-codec` Wir empfehlen, den ZSTD-Komprimierungscodec zu verwenden, um die Gesamtleistung von Tabellen zu verbessern.

Standardmäßig verwenden die Iceberg-Versionen 1.3 und früher die GZIP-Komprimierung, die im Vergleich zu ZSTD eine langsamere Lese-/Schreibleistung bietet.

Hinweis: Einige Engines verwenden möglicherweise andere Standardwerte. Dies ist bei [Iceberg-Tabellen der Fall, die mit Athena oder Amazon EMR Version 7.x erstellt wurden](#).

Legen Sie die Sortierreihenfolge fest

Um die Leseleistung bei Iceberg-Tabellen zu verbessern, empfehlen wir, dass Sie Ihre Tabelle nach einer oder mehreren Spalten sortieren, die häufig in Abfragefiltern verwendet werden. Durch Sortierung in Kombination mit den Spaltenstatistiken von Iceberg kann das Bereinigen von Dateien erheblich effizienter gestaltet werden, was zu schnelleren Lesevorgängen führt. Die Sortierung reduziert auch die Anzahl der Amazon S3 S3-Anfragen für Abfragen, die die Sortierspalten in Abfragefiltern verwenden.

Sie können eine hierarchische Sortierreihenfolge auf Tabellenebene festlegen, indem Sie eine DDL-Anweisung (Data Definition Language) mit Spark ausführen. Die verfügbaren Optionen finden Sie in

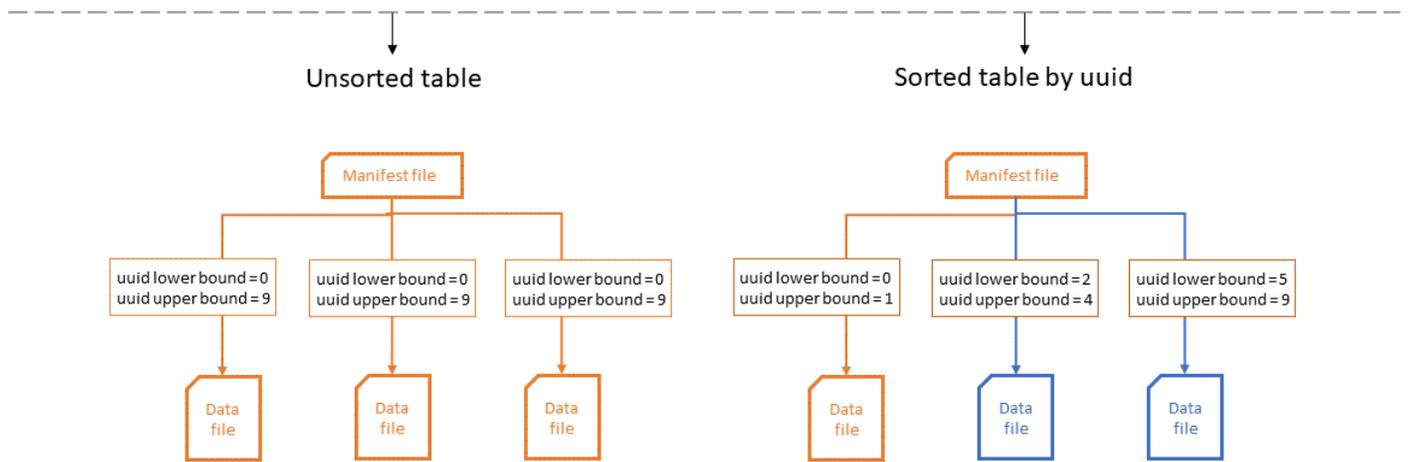
der [Iceberg-Dokumentation](#). Nachdem Sie die Sortierreihenfolge festgelegt haben, wenden Autoren diese Sortierung auf nachfolgende Datenschreibvorgänge in der Iceberg-Tabelle an.

In Tabellen, die nach Datum (yyyy-mm-dd) partitioniert sind, nach denen die meisten Abfragen filtern uuid, können Sie beispielsweise die DDL-Option verwenden, `Write Distributed By Partition Locally Ordered` um sicherzustellen, dass Spark Dateien mit nicht überlappenden Bereichen schreibt.

Das folgende Diagramm zeigt, wie sich die Effizienz von Spaltenstatistiken verbessert, wenn Tabellen sortiert werden. In diesem Beispiel muss die sortierte Tabelle nur eine einzige Datei öffnen und profitiert optimal von der Partition und Datei von Iceberg. In der unsortierten Tabelle uuid kann jede Datei potenziell in jeder beliebigen Datendatei vorkommen, sodass die Abfrage alle Datendateien öffnen muss.

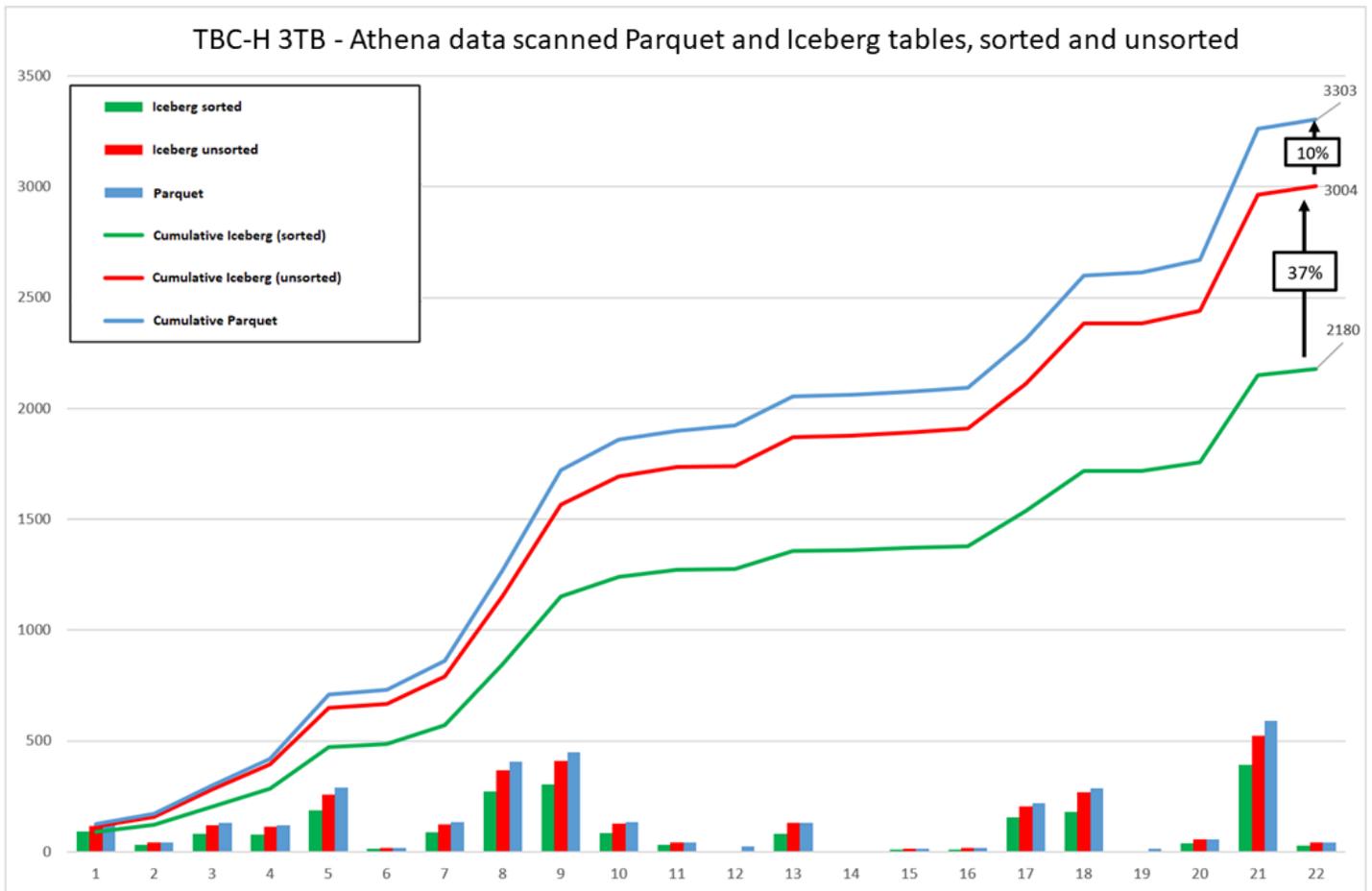
Query example:

```
SELECT * FROM Table
WHERE date > 2022-02-05 AND date < 2022-02-10 AND uuid = 1
```



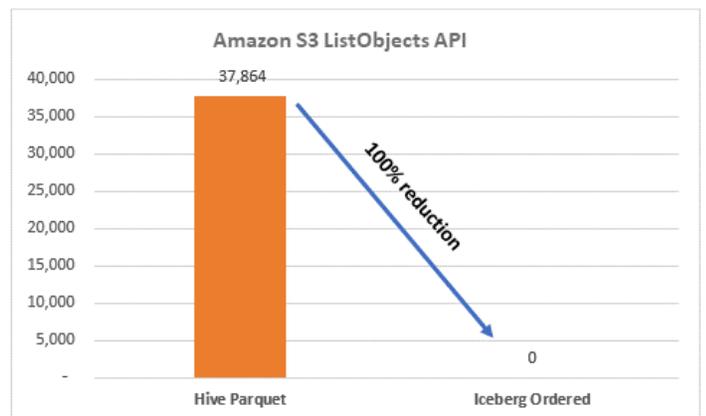
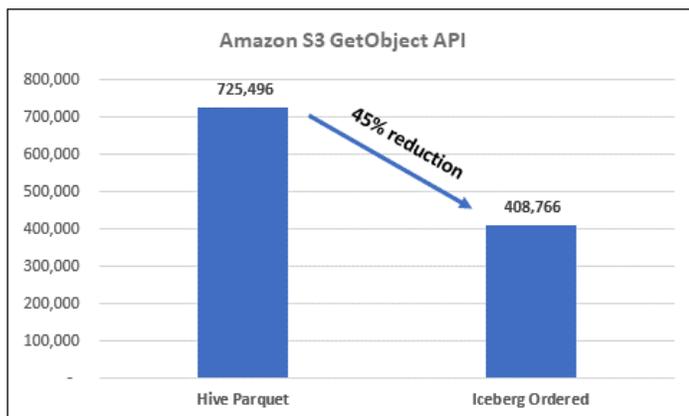
Das Ändern der Sortierreihenfolge hat keine Auswirkungen auf bestehende Datendateien. Sie können die Iceberg-Komprimierung verwenden, um die Sortierreihenfolge auf diese anzuwenden.

Die Verwendung von nach Iceberg sortierten Tabellen kann die Kosten für Ihre Arbeitslast senken, wie in der folgenden Grafik dargestellt.



Diese Grafiken fassen die Ergebnisse der Durchführung des TPC-H-Benchmarks für Hive-Tabellen (Parquet) im Vergleich zu sortierten Iceberg-Tabellen zusammen. Bei anderen Datensätzen oder Workloads können die Ergebnisse jedoch anders ausfallen.

TPC-H 3TB - 22 queries



Optimierung der Schreibleistung

In diesem Abschnitt werden Tabelleneigenschaften beschrieben, die Sie anpassen können, um die Schreibleistung auf Iceberg-Tabellen unabhängig von der Engine zu optimieren.

Stellen Sie den Tabellenverteilungsmodus ein

Iceberg bietet mehrere Schreibverteilungsmodi, die definieren, wie Schreibdaten auf Spark-Aufgaben verteilt werden. Einen Überblick über die verfügbaren Modi finden Sie unter [Writing Distribution Modes](#) in der Iceberg-Dokumentation.

Für Anwendungsfälle, in denen der Schreibgeschwindigkeit Priorität eingeräumt wird, insbesondere bei Streaming-Workloads, legen Sie die Einstellung auf fest. `write.distribution-mode none`. Dadurch wird sichergestellt, dass Iceberg kein zusätzliches Spark-Shuffling anfordert und dass Daten geschrieben werden, sobald sie in Spark-Aufgaben verfügbar sind. Dieser Modus eignet sich besonders für Spark Structured Streaming-Anwendungen.

Note

Wenn Sie den Schreibverteilungsmodus auf `none` einstellen, werden in der `none` Regel zahlreiche kleine Dateien erzeugt, was die Leseleistung beeinträchtigt. Wir empfehlen eine regelmäßige Komprimierung, um diese kleinen Dateien aus Gründen der Abfrageleistung in Dateien mit der richtigen Größe zu konsolidieren.

Wählen Sie die richtige Aktualisierungsstrategie

Verwenden Sie eine `merge-on-read` Strategie zur Optimierung der Schreibleistung, wenn langsamere Lesevorgänge mit den neuesten Daten für Ihren Anwendungsfall akzeptabel sind.

Wenn Sie Iceberg verwenden `merge-on-read`, schreibt Iceberg Aktualisierungen und Löschungen als separate kleine Dateien in den Speicher. Wenn die Tabelle gelesen wird, muss der Leser diese Änderungen mit den Basisdateien zusammenführen, um die neueste Ansicht der Daten zurückzugeben. Dies führt zu Leistungseinbußen bei Lesevorgängen, beschleunigt aber das Schreiben von Aktualisierungen und Löschungen. In der Regel `merge-on-read` ist es ideal für das Streaming von Workloads mit Updates oder Jobs mit wenigen Updates, die auf viele Tabellenpartitionen verteilt sind.

Sie können merge-on-read Konfigurationen (`write.update.modewrite.delete.mode`, `undwrite.merge.mode`) auf Tabellenebene oder unabhängig voneinander auf der Anwendungsseite festlegen.

Die Verwendung merge-on-read erfordert eine regelmäßige Komprimierung, um zu verhindern, dass sich die Leseleistung im Laufe der Zeit verschlechtert. Durch die Komprimierung werden Aktualisierungen und Löschungen mit vorhandenen Datendateien abgeglichen, um einen neuen Satz von Datendateien zu erstellen, wodurch die Leistungseinbußen beim Lesen vermieden werden. [Standardmäßig werden bei der Komprimierung von Iceberg gelöschte Dateien nicht zusammengeführt, es sei denn, Sie ändern die Standardeinstellung der `delete-file-threshold` Eigenschaft auf einen kleineren Wert \(siehe die Iceberg-Dokumentation\)](#). Weitere Informationen zur Komprimierung finden Sie im Abschnitt [Eisberg-Komprimierung](#) weiter unten in diesem Handbuch.

Wählen Sie das richtige Dateiformat

Iceberg unterstützt das Schreiben von Daten in den Formaten Parquet, ORC und Avro. Parquet ist das Standardformat. Parquet und ORC sind spaltenförmige Formate, die eine hervorragende Leseleistung bieten, aber im Allgemeinen langsamer zu schreiben sind. Dies stellt den typischen Kompromiss zwischen Lese- und Schreibleistung dar.

Wenn die Schreibgeschwindigkeit für Ihren Anwendungsfall wichtig ist, z. B. bei Streaming-Workloads, sollten Sie erwägen, im Avro-Format zu schreiben, indem Sie Avro in den `write-format` Writer-Optionen auf setzen. Da Avro ein zeilenbasiertes Format ist, bietet es schnellere Schreibzeiten auf Kosten einer langsameren Leseleistung.

Um die Leseleistung zu verbessern, sollten Sie regelmäßig komprimieren, um kleine Avro-Dateien zusammenzuführen und in größere Parquet-Dateien umzuwandeln. Das Ergebnis des Verdichtungsvorgangs wird durch die `write.format.default` Tabelleneinstellung bestimmt. Das Standardformat für Iceberg ist Parquet. Wenn Sie also in Avro schreiben und dann die Komprimierung ausführen, wandelt Iceberg die Avro-Dateien in Parquet-Dateien um. Ein Beispiel:

```
spark.sql(f"""
  CREATE TABLE IF NOT EXISTS glue_catalog.{DB_NAME}.{TABLE_NAME} (
    Col_1 float,
    <<<...other columns...>>
    ts timestamp)
  USING iceberg
  PARTITIONED BY (days(ts))
  OPTIONS (
```

```

    'format-version'='2',
    write.format.default='parquet')
    """)

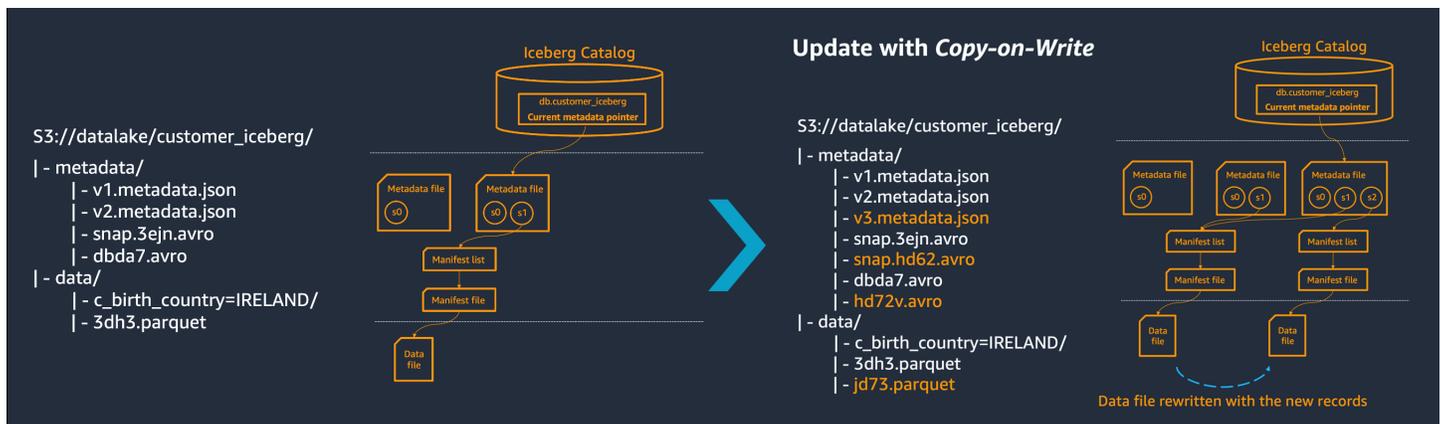
query = df \
    .writeStream \
    .format("iceberg") \
    .option("write-format", "avro") \
    .outputMode("append") \
    .trigger(processingTime='60 seconds') \
    .option("path", f"glue_catalog.{DB_NAME}.{TABLE_NAME}") \
    .option("checkpointLocation", f"s3://{BUCKET_NAME}/checkpoints/iceberg/")

    .start()

```

Optimierung des Speichers

Durch das Aktualisieren oder Löschen von Daten in einer Iceberg-Tabelle erhöht sich die Anzahl der Kopien Ihrer Daten, wie in der folgenden Abbildung dargestellt. Das Gleiche gilt für die laufende Komprimierung: Sie erhöht die Anzahl der Datenkopien in Amazon S3. Das liegt daran, dass Iceberg die allen Tabellen zugrunde liegenden Dateien als unveränderlich behandelt.



Folgen Sie den bewährten Methoden in diesem Abschnitt, um die Speicherkosten zu verwalten.

Aktivieren Sie S3 Intelligent-Tiering

Verwenden Sie die [Amazon S3 S3-Speicherklasse Intelligent-Tiering](#), um Daten automatisch auf die kostengünstigste Zugriffsebene zu verschieben, wenn sich die Zugriffsmuster ändern. Diese Option hat keinen betrieblichen Mehraufwand oder Auswirkungen auf die Leistung.

Hinweis: Verwenden Sie nicht die optionalen Stufen (wie Archive Access und Deep Archive Access) in S3 Intelligent-Tiering mit Iceberg-Tabellen. Informationen zur Archivierung von Daten finden Sie in den Richtlinien im nächsten Abschnitt.

Sie können [Amazon S3 S3-Lebenszyklusregeln](#) auch verwenden, um Ihre eigenen Regeln für das Verschieben von Objekten in eine andere Amazon S3 S3-Speicherklasse wie S3 Standard-IA oder S3 One Zone-IA festzulegen (siehe [Unterstützte Übergänge und zugehörige Einschränkungen](#) in der Amazon S3 S3-Dokumentation).

Archivieren oder löschen Sie historische Snapshots

Für jede festgeschriebene Transaktion (Einfügen, Aktualisieren, Zusammenführen, Komprimieren) zu einer Iceberg-Tabelle wird eine neue Version oder ein Snapshot der Tabelle erstellt. Im Laufe der Zeit häufen sich die Anzahl der Versionen und die Anzahl der Metadateien in Amazon S3 an.

Das Speichern von Snapshots einer Tabelle ist für Funktionen wie Snapshot-Isolation, Tabellen-Rollback und Zeitreiseabfragen erforderlich. Die Speicherkosten steigen jedoch mit der Anzahl der Versionen, die Sie behalten.

In der folgenden Tabelle werden die Entwurfsmuster beschrieben, die Sie implementieren können, um die Kosten auf der Grundlage Ihrer Datenaufbewahrungsanforderungen zu verwalten.

Entwurfsmuster	Lösung	Anwendungsfälle
Lösche alte Schnappschüsse	<ul style="list-style-type: none"> • Verwenden Sie die VACUUM-Anweisung in Athena, um alte Schnappschüsse zu entfernen. Für diesen Vorgang fallen keine Rechenkosten an. • Alternativ können Sie Spark auf Amazon EMR verwenden oder AWS Glue Snapshots entfernen. Weitere Informationen finden Sie unter expire_sn 	<p>Bei diesem Ansatz werden Snapshots gelöscht, die nicht mehr benötigt werden, um die Speicherkosten zu senken. Sie können je nach Ihren Datenaufbewahrungsanforderungen konfigurieren, wie viele Snapshots aufbewahrt werden sollen oder wie lange.</p> <p>Mit dieser Option werden die Snapshots dauerhaft gelöscht. Ein Rollback oder</p>

Entwurfsmuster	Lösung	Anwendungsfälle
Legen Sie Aufbewahrungsrichtlinien für bestimmte Snapshots fest	<p data-bbox="623 212 971 289">apsshots in der Iceberg-Dokumentation.</p> <p data-bbox="591 338 987 751">1. Verwenden Sie Tags, um bestimmte Snapshots zu markieren und eine Aufbewahrungsrichtlinie in Iceberg zu definieren. Weitere Informationen finden Sie unter Historische Tags in der Iceberg-Dokumentation.</p> <p data-bbox="623 800 1024 1115">Sie können beispielsweise einen Snapshot pro Monat für ein Jahr aufbewahren, indem Sie die folgende SQL-Anweisung in Spark auf Amazon EMR verwenden:</p>	<p data-bbox="1068 212 1507 289">eine Zeitreise zu abgelaufenen Snapshots sind nicht möglich.</p> <p data-bbox="1068 338 1507 1230">Dieses Muster ist hilfreich bei der Einhaltung geschäftlicher oder rechtlicher Anforderungen, nach denen Sie den Status einer Tabelle zu einem bestimmten Zeitpunkt in der Vergangenheit anzeigen müssen. Indem Sie Aufbewahrungsrichtlinien für bestimmte markierte Snapshots festlegen, können Sie andere (nicht markierte) Snapshots entfernen, die erstellt wurden. Auf diese Weise können Sie die Anforderungen an die Datenaufbewahrung erfüllen, ohne jeden einzelnen erstellen Snapshot behalten zu müssen.</p>

```
ALTER TABLE glue_catalog.db.table
CREATE TAG 'EOM-01' AS
OF VERSION 30 RETAIN
365 DAYS
```

Entwurfsmuster

Archivieren Sie alte Schnappschüsse

Lösung

1. Verwenden Sie Amazon S3 S3-Tags, um Objekte mit Spark zu markieren . (Amazon S3 S3-Tags unterscheiden sich von Iceberg-Tags. Weitere Informationen finden Sie in der [Iceberg-Dokumentation](#).) Zum Beispiel:

```
spark.sql.catalog.  
my_catalog.s3.delete-enabled=false and  
\  
spark.sql.catalog.  
my_catalog.s3.delete.tags.my_key=to_archive
```

2. Verwenden Sie Spark auf Amazon EMR oder AWS Glue um [Snapshots zu entfernen](#). Wenn Sie die Einstellungen im Beispiel verwenden, kennzeichnet dieses Verfahren Objekte und trennt sie von den Metadaten der Iceberg-Tabelle, anstatt sie aus Amazon S3 zu löschen.
3. Verwenden Sie S3-Lebenszyklusregeln, um Objekte, die als gekennzeichnet sind `to_archive` , in eine der [S3 Glacier-Speicherklassen](#) zu übertragen.

Anwendungsfälle

Dieses Muster ermöglicht es Ihnen, alle Tabellenversionen und Snapshots zu geringeren Kosten aufzubewahren.

Sie können keine Zeitreise unternehmen oder zu archivierten Snapshots zurückkehren, ohne diese Versionen zuerst als neue Tabellen wiederherzustellen. Dies ist in der Regel für Prüfungszwecke akzeptabel.

Sie können diesen Ansatz mit dem vorherigen Entwurfsmuster kombinieren und Aufbewahrungsrichtlinien für bestimmte Snapshots festlegen.

Entwurfsmuster	Lösung	Anwendungsfälle
----------------	--------	-----------------

4. Um archivierte Daten abzufragen:

- [Stellen Sie die archivierten Objekte](#) wieder her.
- Verwenden Sie die [Prozedur register_table](#) in Iceberg, um den Snapshot als Tabelle im Katalog zu registrieren.

Eine ausführliche Anleitung finden Sie im AWS Blogbeitrag [Verbessern Sie die betriebliche Effizienz von Apache Iceberg-Tabellen, die auf Amazon S3 S3-Datenseen basieren](#).

Löschen Sie verwaiste Dateien

In bestimmten Situationen können Iceberg-Anwendungen fehlschlagen, bevor Sie Ihre Transaktionen festschreiben. Dadurch verbleiben Datendateien in Amazon S3. Da kein Commit vorgenommen wurde, werden diese Dateien keiner Tabelle zugeordnet, sodass Sie sie möglicherweise asynchron bereinigen müssen.

Um diese Löschungen zu handhaben, können Sie die [VACUUM-Anweisung](#) in Amazon Athena verwenden. Diese Anweisung entfernt Schnappschüsse und löscht auch verwaiste Dateien. Dies ist sehr kosteneffizient, da Athena die Rechenkosten für diesen Vorgang nicht in Rechnung stellt. Außerdem müssen Sie keine zusätzlichen Operationen einplanen, wenn Sie die VACUUM Anweisung verwenden.

Alternativ können Sie Spark auf Amazon EMR verwenden oder AWS Glue das `remove_orphan_files` Verfahren ausführen. Dieser Vorgang ist mit Rechenkosten verbunden und muss unabhängig geplant werden. Weitere Informationen finden Sie in der [Iceberg-Dokumentation](#).

Pflege von Tabellen mithilfe von Komprimierung

Iceberg enthält Funktionen, mit denen Sie [Tabellenverwaltungsoperationen durchführen können, nachdem Sie Daten in die Tabelle](#) geschrieben haben. Einige Wartungsvorgänge konzentrieren sich auf die Optimierung von Metadateien, während andere die Clusterung der Daten in den Dateien verbessern, sodass Abfrage-Engines die für die Beantwortung von Benutzeranfragen erforderlichen Informationen effizient finden können. Dieser Abschnitt konzentriert sich auf Optimierungen im Zusammenhang mit der Komprimierung.

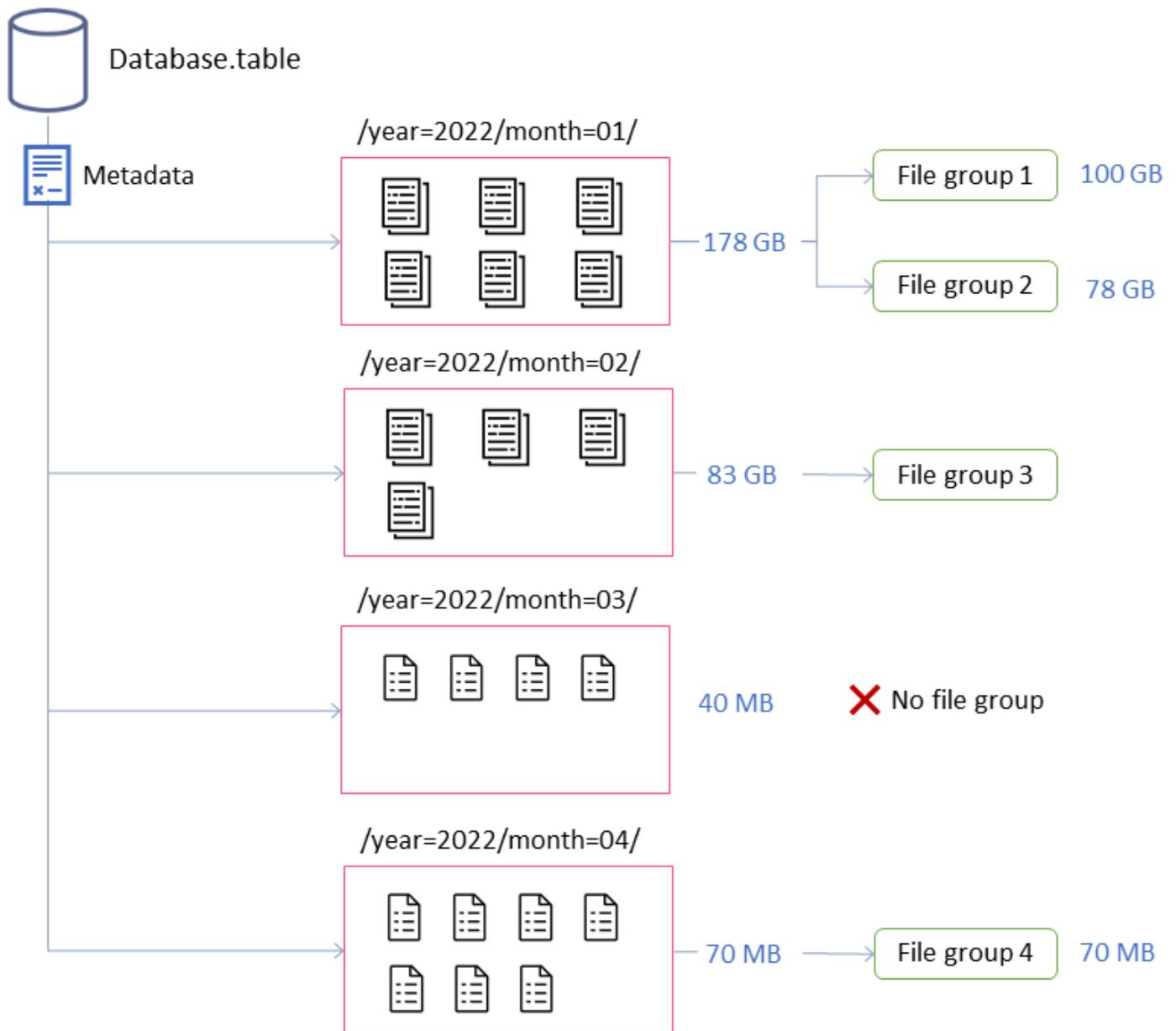
Verdichtung von Eisbergen

In Iceberg können Sie die Komprimierung verwenden, um vier Aufgaben auszuführen:

- Kombinieren kleiner Dateien zu größeren Dateien, die in der Regel über 100 MB groß sind. Diese Technik wird als Mülltonnenverpackung bezeichnet.
- Zusammenführen von gelöschten Dateien mit Datendateien. Löschdateien werden durch Aktualisierungen oder Löschungen generiert, die diesen Ansatz verwenden. `merge-on-read`
- (Neu-) Sortierung der Daten gemäß Abfragemustern. Daten können ohne jegliche Sortierreihenfolge oder mit einer Sortierreihenfolge geschrieben werden, die für Schreibvorgänge und Aktualisierungen geeignet ist.
- Clustering der Daten mithilfe von raumfüllenden Kurven zur Optimierung für unterschiedliche Abfragemuster, insbesondere für die Sortierung in Z-Reihenfolge.

Auf AWS, können Sie Tabellenkomprimierungs- und Wartungsvorgänge für Iceberg über Amazon Athena oder mithilfe von Spark in Amazon EMR oder ausführen. AWS Glue

Wenn Sie die Komprimierung mithilfe der Prozedur [rewrite_data_files](#) ausführen, können Sie mehrere Regler einstellen, um das Verdichtungsverhalten zu steuern. Das folgende Diagramm zeigt das Standardverhalten beim Packen von Behältern. Das Verständnis der Kompaktierung von Behältern ist für das Verständnis der Implementierungen der hierarchischen Sortierung und der Sortierung in Z-Reihenfolge von entscheidender Bedeutung, da es sich dabei um Erweiterungen der Schnittstelle zum Packen von Behältern handelt, die auf ähnliche Weise funktionieren. Der Hauptunterschied besteht in dem zusätzlichen Schritt, der zum Sortieren oder Clustern der Daten erforderlich ist.



In diesem Beispiel besteht die Iceberg-Tabelle aus vier Partitionen. Jede Partition hat eine andere Größe und eine unterschiedliche Anzahl von Dateien. Wenn Sie eine Spark-Anwendung starten, um die Komprimierung auszuführen, erstellt die Anwendung insgesamt vier Dateigruppen, die verarbeitet werden sollen. Eine Dateigruppe ist eine Iceberg-Abstraktion, die eine Sammlung von Dateien darstellt, die von einem einzelnen Spark-Job verarbeitet werden. Das heißt, die Spark-Anwendung, die die Komprimierung ausführt, erstellt vier Spark-Jobs zur Verarbeitung der Daten.

Optimieren des Verdichtungsverhaltens

Die folgenden wichtigen Eigenschaften steuern, wie Datendateien für die Komprimierung ausgewählt werden:

- [MAX_FILE_GROUP_SIZE_BYTES](#) legt das Datenlimit für eine einzelne Dateigruppe (Spark-Job) standardmäßig auf 100 GB fest. Diese Eigenschaft ist besonders wichtig für Tabellen ohne Partitionen oder Tabellen mit Partitionen, die sich über Hunderte von Gigabyte erstrecken. Wenn Sie dieses Limit festlegen, können Sie die Abläufe unterteilen, um die Arbeit zu planen und Fortschritte zu erzielen, und gleichzeitig eine Erschöpfung der Ressourcen im Cluster verhindern.

Hinweis: Jede Dateigruppe ist separat sortiert. Wenn Sie also eine Sortierung auf Partitionsebene durchführen möchten, müssen Sie dieses Limit an die Partitionsgröße anpassen.

- [MIN_FILE_SIZE_BYTES](#) oder [MIN_FILE_SIZE_DEFAULT_RATIO](#) sind standardmäßig auf [75 Prozent der auf Tabellenebene festgelegten Zieldateigröße](#) eingestellt. Wenn eine Tabelle beispielsweise eine Zielgröße von 512 MB hat, ist jede Datei, die kleiner als 384 MB ist, in der Gruppe der zu komprimierenden Dateien enthalten.
- [MAX_FILE_SIZE_BYTES](#) oder [MAX_FILE_SIZE_DEFAULT_RATIO](#) sind standardmäßig [auf 180 Prozent der Zieldateigröße eingestellt](#). Wie bei den beiden Eigenschaften, die Mindestdateigrößen festlegen, werden diese Eigenschaften verwendet, um Kandidatendateien für den Komprimierungsauftrag zu identifizieren.
- [MIN_INPUT_FILES](#) gibt die Mindestanzahl der zu komprimierenden Dateien an, wenn die Größe einer Tabellenpartition kleiner als die Zieldateigröße ist. Der Wert dieser Eigenschaft wird verwendet, um zu bestimmen, ob es sich lohnt, die Dateien anhand der Anzahl der Dateien zu komprimieren (standardmäßig 5).
- [DELETE_FILE_THRESHOLD](#) gibt die Mindestanzahl von Löschvorgängen für eine Datei an, bevor sie in die Komprimierung aufgenommen wird. Sofern Sie nichts anderes angeben, werden bei der Komprimierung keine Löschdateien mit Datendateien kombiniert. Um diese Funktion zu aktivieren, müssen Sie mithilfe dieser Eigenschaft einen Schwellenwert festlegen. Dieser Schwellenwert ist spezifisch für einzelne Datendateien. Wenn Sie ihn also auf 3 setzen, wird eine Datendatei nur dann neu geschrieben, wenn drei oder mehr Löschdateien darauf verweisen.

Diese Eigenschaften geben Aufschluss über die Bildung der Dateigruppen im vorherigen Diagramm.

Die beschriftete Partition `month=01` umfasst beispielsweise zwei Dateigruppen, da sie die maximale Größenbeschränkung von 100 GB überschreitet. Im Gegensatz dazu enthält die `month=02` Partition eine einzelne Dateigruppe, da sie weniger als 100 GB groß ist. Die `month=03` Partition erfüllt nicht

die standardmäßige Mindestanforderung für Eingabedateien von fünf Dateien. Infolgedessen wird sie nicht komprimiert. Schließlich werden die Dateien komprimiert, obwohl die `month=04` Partition nicht genügend Daten enthält, um eine einzelne Datei mit der gewünschten Größe zu bilden, da die Partition mehr als fünf kleine Dateien enthält.

Sie können diese Parameter für die Ausführung von Spark auf Amazon EMR oder AWS Glue festlegen. Für Amazon Athena können Sie ähnliche Eigenschaften verwalten, indem Sie die [Tabelleneigenschaften](#) verwenden, die mit dem Präfix `beginnenoptimize_` beginnen.

Verdichtung mit Spark auf Amazon EMR ausführen oder AWS Glue

In diesem Abschnitt wird beschrieben, wie Sie die Größe eines Spark-Clusters für die Ausführung des Komprimierungsprogramms von Iceberg richtig dimensionieren. Im folgenden Beispiel wird Amazon EMR Serverless verwendet, aber Sie können dieselbe Methode in Amazon EMR on Amazon EC2 oder Amazon EKS oder in verwenden. AWS Glue

Sie können die Korrelation zwischen Dateigruppen und Spark-Jobs nutzen, um die Cluster-Ressourcen zu planen. Um die Dateigruppen sequentiell zu verarbeiten und dabei die maximale Größe von 100 GB pro Dateigruppe zu berücksichtigen, können Sie die folgenden [Spark-Eigenschaften](#) festlegen:

- `spark.dynamicAllocation.enabled = FALSE`
- `spark.executor.memory = 20 GB`
- `spark.executor.instances = 5`

Wenn Sie die Komprimierung beschleunigen möchten, können Sie horizontal skalieren, indem Sie die Anzahl der Dateigruppen erhöhen, die parallel komprimiert werden. Sie können Amazon EMR auch mithilfe manueller oder dynamischer Skalierung skalieren.

- Manuelles Skalieren (z. B. um den Faktor 4)
 - `MAX_CONCURRENT_FILE_GROUP_REWRITES= 4` (unser Faktor)
 - `spark.executor.instances= 5` (im Beispiel verwendeter Wert) x 4 (unser Faktor) = 20
 - `spark.dynamicAllocation.enabled = FALSE`
- Dynamische Skalierung
 - `spark.dynamicAllocation.enabled= TRUE` (Standard, keine Aktion erforderlich)

- [MAX_CONCURRENT_FILE_GROUP_REWRITES = N](#) (richten Sie diesen Wert `anspark.dynamicAllocation.maxExecutors`, der standardmäßig 100 ist; basierend auf den Executor-Konfigurationen im Beispiel können Sie ihn auf 20 setzen) N

Dies sind Richtlinien zur Unterstützung der Clustergröße. Sie sollten jedoch auch die Leistung Ihrer Spark-Jobs überwachen, um die besten Einstellungen für Ihre Workloads zu finden.

Verdichtung mit Amazon Athena ausführen

[Athena bietet eine Implementierung des Verdichtungsprogramms von Iceberg als verwaltete Funktion über die OPTIMIZE-Anweisung an.](#) Sie können diese Anweisung verwenden, um die Komprimierung auszuführen, ohne die Infrastruktur auswerten zu müssen.

Diese Anweisung gruppiert kleine Dateien mithilfe des Bin-Packing-Algorithmus zu größeren Dateien und führt gelöschte Dateien mit vorhandenen Datendateien zusammen. Um die Daten mithilfe von hierarchischer Sortierung oder Sortierung in Z-Reihenfolge zu clustern, verwenden Sie Spark auf Amazon EMR oder AWS Glue

Sie können das Standardverhalten der OPTIMIZE Anweisung bei der Tabellenerstellung ändern, indem Sie Tabelleneigenschaften in der CREATE TABLE Anweisung übergeben, oder nach der Tabellenerstellung, indem Sie die Anweisung verwenden. ALTER TABLE Standardwerte finden Sie in der [Athena-Dokumentation](#).

Empfehlungen für die Ausführung der Komprimierung

Anwendungsfall

Die Komprimierung von Bin Packing wird nach einem Zeitplan ausgeführt

Empfehlung

- Verwenden Sie die OPTIMIZE Anweisung in Athena, wenn Sie nicht wissen, wie viele kleine Dateien Ihre Tabelle enthält. Das Athena-Preismodell basiert auf den gescannten Daten. Wenn also keine Dateien komprimiert werden müssen, fallen mit diesen Vorgängen keine Kosten an. Um Timeouts an Athena-Tabellen zu vermeiden, führen Sie das Programm OPTIMIZE auf Basis aus. `per-table-partition`

Anwendungsfall

Die Bin-Packing-Komprimierung wird auf der Grundlage von Ereignissen ausgeführt

Die Komprimierung wird ausgeführt, um Daten zu sortieren

Führen Sie die Komprimierung aus, um die Daten mithilfe der Sortierung in Z-Reihenfolge zu clustern

Die Komprimierung wird auf Partitionen ausgeführt, die aufgrund spät eingehender Daten möglicherweise von anderen Anwendungen aktualisiert werden

Empfehlung

- Verwenden Sie Amazon EMR oder AWS Glue mit dynamischer Skalierung, wenn Sie erwarten, dass große Mengen kleiner Dateien komprimiert werden.
- Verwenden Sie Amazon EMR oder AWS Glue mit dynamischer Skalierung, wenn Sie erwarten, dass große Mengen kleiner Dateien komprimiert werden.
- Verwenden Sie Amazon EMR oder AWS Glue, weil das Sortieren ein teurer Vorgang ist und möglicherweise Daten auf die Festplatte übertragen werden müssen.
- Verwenden Sie Amazon EMR oder AWS Glue, da die Sortierung in Z-Reihenfolge ein sehr teurer Vorgang ist und möglicherweise Daten auf die Festplatte übertragen werden müssen.
- Verwenden Sie Amazon EMR oder AWS Glue. Aktivieren Sie die Iceberg-Eigenschaft [PARTIAL_PROGRESS_ENABLED](#). Wenn Sie diese Option verwenden, teilt Iceberg die Komprimierungsausgabe in mehrere Commits auf. Wenn es zu einer Kollision kommt (d. h. wenn die Datendatei während der Komprimierung aktualisiert wird), reduziert diese Einstellung die Kosten für Wiederholungsversuche, indem sie auf den Commit beschränkt wird, der die betroffene Datei enthält. Andernfalls müssen Sie möglicherweise alle Dateien neu komprimieren.

Verwenden von Iceberg-Workloads in Amazon S3

In diesem Abschnitt werden Iceberg-Eigenschaften beschrieben, mit denen Sie die Interaktion von Iceberg mit Amazon S3 optimieren können.

Vermeiden Sie Hot-Partitionierung (HTTP 503-Fehler)

Einige Data Lake-Anwendungen, die auf Amazon S3 ausgeführt werden, verarbeiten Millionen oder Milliarden von Objekten und verarbeiten Petabyte an Daten. Dies kann dazu führen, dass Präfixe ein hohes Datenvolumen empfangen und diese in der Regel anhand von HTTP-503-Fehlern (Service nicht verfügbar) erkannt werden. Verwenden Sie die folgenden Iceberg-Eigenschaften, um dieses Problem zu vermeiden:

- Auf `hash` oder `range` so einstellen `write.distribution-mode`, dass Iceberg große Dateien schreibt, was zu weniger Amazon S3 S3-Anfragen führt. Dies ist die bevorzugte Konfiguration und sollte für die meisten Fälle geeignet sein.
- Wenn Sie aufgrund eines immensen Datenvolumens in Ihren Workloads weiterhin 503-Fehler feststellen, können Sie dies `true` in `write.object-storage.enabled` Iceberg einstellen. Dadurch wird Iceberg angewiesen, Objektnamen zu hashen und die Last auf mehrere, zufällige Amazon S3 S3-Präfixe zu verteilen.

Weitere Informationen zu diesen Eigenschaften finden [Sie unter Eigenschaften schreiben](#) in der Iceberg-Dokumentation.

Verwenden Sie die Wartungsoperationen von Iceberg, um ungenutzte Daten freizugeben

Um Iceberg-Tabellen zu verwalten, können Sie die Iceberg-Kern-API, Iceberg-Clients (wie Spark) oder verwaltete Dienste wie Amazon Athena verwenden. [Um alte oder unbenutzte Dateien aus Amazon S3 zu löschen, empfehlen wir, nur Iceberg native APIs zu verwenden, um Schnappschüsse, alte Metadatendateien und verwaiste Dateien zu entfernen.](#)

Die Verwendung von Amazon S3 APIs über Boto3, das Amazon S3 S3-SDK oder das AWS Command Line Interface (AWS CLI) oder die Verwendung anderer, nicht von Iceberg stammender Methoden zum Überschreiben oder Entfernen von Amazon S3 S3-Dateien für eine Iceberg-Tabelle führt zu Tabellenbeschädigungen und Abfragefehlern.

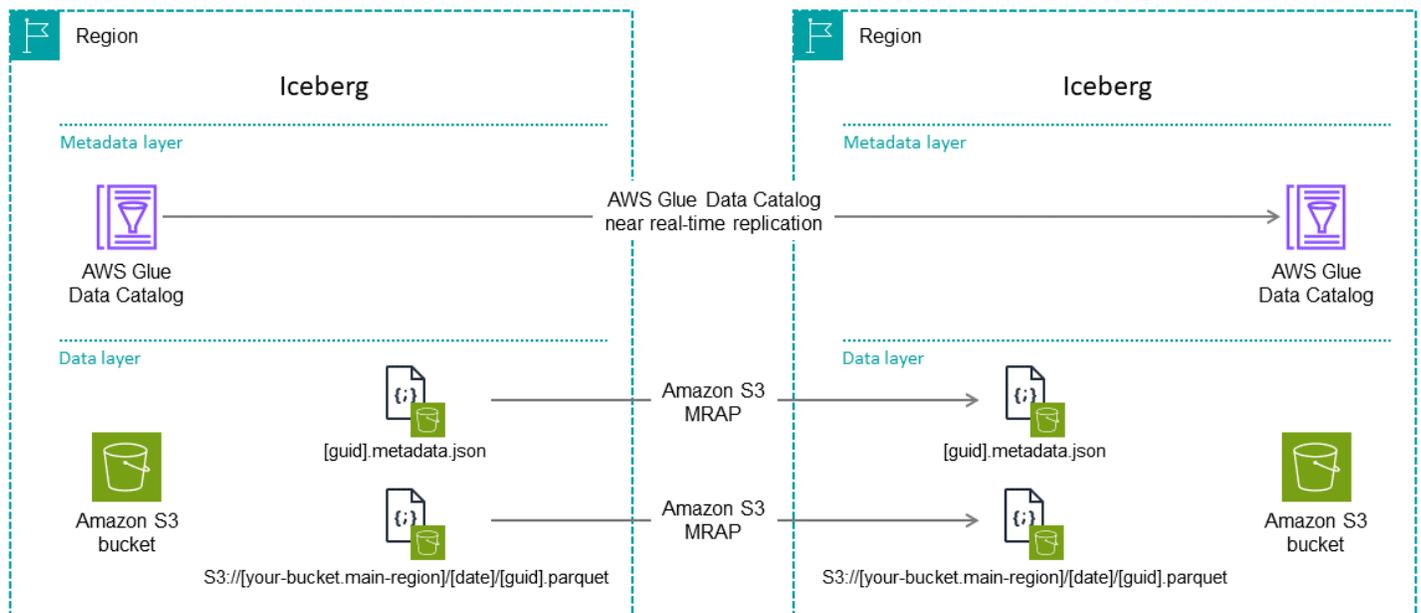
Replizieren Sie Daten zwischen AWS-Regionen

Wenn Sie Iceberg-Tabellen in Amazon S3 speichern, können Sie die in Amazon S3 integrierten Funktionen wie [Cross-Region Replication \(CRR\)](#) und [Multi-Region Access Points \(MRAP\)](#) verwenden, um Daten über mehrere AWS-Regionen hinweg zu replizieren. MRAP bietet einen globalen Endpunkt für Anwendungen, um auf S3-Buckets zuzugreifen, die sich in mehreren befinden. AWS-Regionen Iceberg unterstützt keine relativen Pfade, aber Sie können MRAP verwenden, um Amazon S3 S3-Operationen durchzuführen, indem Sie Buckets Access Points zuordnen. MRAP lässt sich auch nahtlos in den regionsübergreifenden Replikationsprozess von Amazon S3 integrieren, was zu einer Verzögerung von bis zu 15 Minuten führt. Sie müssen sowohl Daten- als auch Metadaten Dateien replizieren.

⚠ Important

Derzeit funktioniert die Iceberg-Integration mit MRAP nur mit Apache Spark. Wenn Sie ein Failover auf die Sekundärseite durchführen müssen AWS-Region, müssen Sie planen, Benutzeranfragen an eine Spark-SQL-Umgebung (wie Amazon EMR) in der Failover-Region umzuleiten.

Die CRR- und MRAP-Funktionen helfen Ihnen beim Aufbau einer regionsübergreifenden Replikationslösung für Iceberg-Tabellen, wie in der folgenden Abbildung dargestellt.



So richten Sie diese regionsübergreifende Replikationsarchitektur ein:

1. Erstellen Sie Tabellen mithilfe des MRAP-Speicherorts. Dadurch wird sichergestellt, dass Iceberg-Metadatendateien auf den MRAP-Speicherort und nicht auf den physischen Bucket-Speicherort verweisen.
2. Replizieren Sie Iceberg-Dateien mithilfe von Amazon S3 MRAP. MRAP unterstützt Datenreplikation mit einem Service Level Agreement (SLA) von 15 Minuten. Iceberg verhindert, dass Lesevorgänge bei der Replikation zu Inkonsistenzen führen.
3. Stellen Sie die Tabellen in der sekundären Region AWS Glue Data Catalog zur Verfügung. Sie können aus zwei Optionen wählen:
 - Richten Sie eine Pipeline für die Replikation von Iceberg-Tabellenmetadaten mithilfe AWS Glue Data Catalog der Replikation ein. Dieses Tool ist im [Replikationsrepository GitHub Glue Catalog und Lake Formation Permissions](#) verfügbar. Dieser ereignisgesteuerte Mechanismus repliziert Tabellen in der Zielregion auf der Grundlage von Ereignisprotokollen.
 - Registrieren Sie die Tabellen in der sekundären Region, wenn ein Failover erforderlich ist. Für diese Option können Sie das vorherige Hilfsprogramm oder die [Iceberg-Prozedur register_table](#) verwenden und auf die neueste Datei verweisen. metadata.json

Überwachung von Apache Iceberg-Workloads

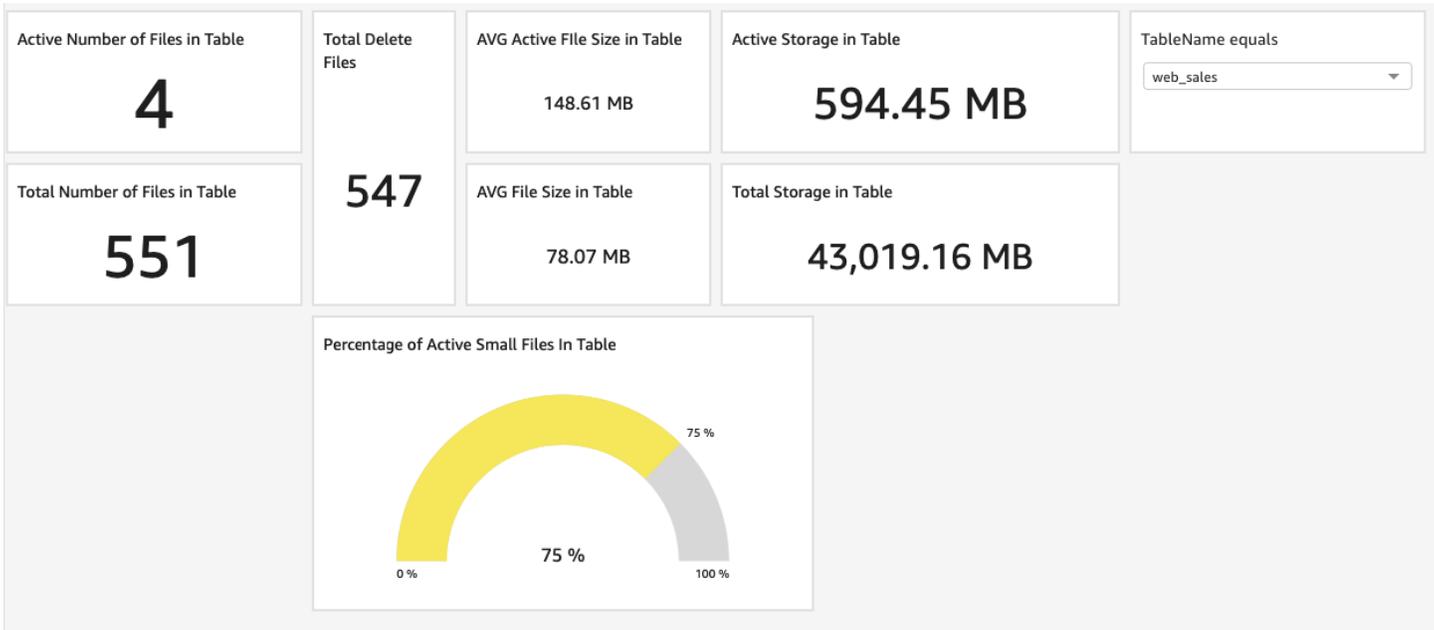
[Um Iceberg-Workloads zu überwachen, haben Sie zwei Möglichkeiten: die Analyse von Metadatentabellen oder die Verwendung von Metrik-Reportern.](#) Metrik-Reporter wurden in Iceberg Version 1.2 eingeführt und sind nur für REST- und JDBC-Kataloge verfügbar.

Wenn Sie Iceberg-Tabellen verwenden AWS Glue Data Catalog, können Sie Einblicke in den Zustand Ihrer Iceberg-Tabellen gewinnen, indem Sie zusätzlich zu den von Iceberg bereitgestellten Metadatentabellen eine Überwachung einrichten.

Die Überwachung ist für das Leistungsmanagement und die Fehlerbehebung von entscheidender Bedeutung. Wenn beispielsweise eine Partition in einer Iceberg-Tabelle einen bestimmten Prozentsatz kleiner Dateien erreicht, kann Ihr Workload einen Komprimierungsjob starten, um die Dateien zu größeren Dateien zu konsolidieren. Dadurch wird verhindert, dass Abfragen über ein akzeptables Maß hinaus verlangsamt werden.

Überwachung auf Tabellenebene

Der folgende Bildschirm zeigt ein Dashboard zur Tabellenüberwachung, das in erstellt wurde. QuickSight Dieses Dashboard fragt Iceberg-Metadatentabellen mithilfe von Spark SQL ab und erfasst detaillierte Messwerte wie die Anzahl der aktiven Dateien und den Gesamtspeicher. Diese Informationen werden dann für betriebliche Zwecke in AWS Glue Tabellen gespeichert. Schließlich wird ein QuickSight Dashboard, wie in der folgenden Abbildung gezeigt, mithilfe von Amazon Athena erstellt. Diese Informationen helfen Ihnen, spezifische Probleme in Ihren Systemen zu identifizieren und zu beheben.



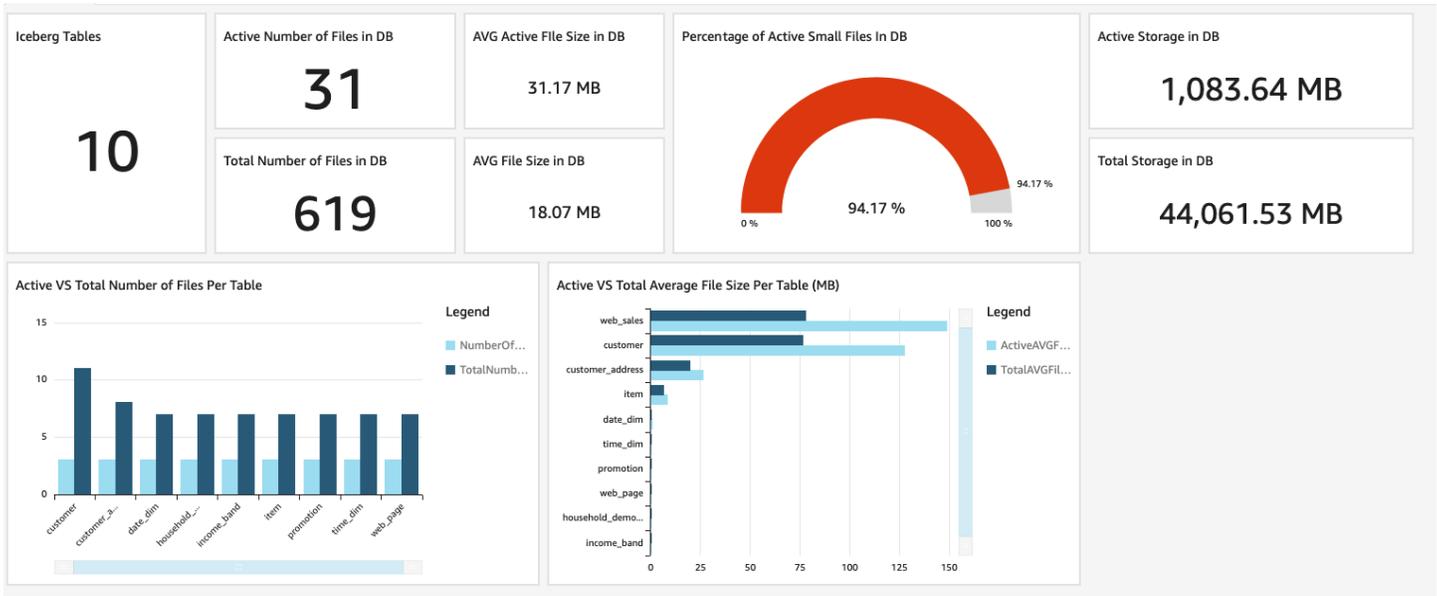
Das QuickSight Beispiel-Dashboard erfasst die folgenden wichtigen Leistungsindikatoren (KPIs) für eine Iceberg-Tabelle:

KPI	Beschreibung	Abfrage
Anzahl der Dateien	Die Anzahl der Dateien in der Iceberg-Tabelle (für alle Snapshots)	<pre>select count(*) from <catalog.database. table_name>.all_files</pre>
Anzahl der aktiven Dateien	Die Anzahl der aktiven Dateien im letzten Snapshot der Iceberg-Tabelle	<pre>select count(*) from <catalog.database. table_name>.files</pre>
Durchschnittliche Dateigröße	Die durchschnittliche Dateigröße in Megabyte für alle Dateien in der Iceberg-Tabelle	<pre>select avg(file_ size_in_bytes)/100 0000 from <catalog.database. table_name>.all_files</pre>
Durchschnittliche Größe der aktiven Datei	Die durchschnittliche Dateigröße in Megabyte für die	<pre>select avg(file_ size_in_bytes)/100 0000</pre>

KPI	Beschreibung	Abfrage
	aktiven Dateien in der Iceberg-Tabelle	<pre>from <catalog.database.table_name>.files</pre>
Prozentsatz kleiner Dateien	Der Prozentsatz der aktiven Dateien, die kleiner als 100 MB sind	<pre>select cast(sum(case when file_size _in_bytes < 100000000 then 1 else 0 end)*100/ count(*) as decimal(1 0,2)) from <catalog.database.table_name>.files</pre>
Gesamtpeichergröße	Die Gesamtgröße aller Dateien in der Tabelle, ausgenommen verwaiste Dateien und Amazon S3 S3-Objektversionen (falls aktiviert)	<pre>select sum(file_ size_in_bytes)/100 0000 from <catalog.database.table_name>.all_files</pre>
Gesamtgröße des aktiven Speichers	Die Gesamtgröße aller Dateien in den aktuellen Snapshots einer bestimmten Tabelle	<pre>select sum(file_ size_in_bytes)/100 0000 from <catalog.database.table_name>.files</pre>

Überwachung auf Datenbankebene

Das folgende Beispiel zeigt ein Überwachungs-Dashboard, das erstellt wurde, QuickSight um einen Überblick über die Datenbankebene KPIs für eine Sammlung von Iceberg-Tabellen zu bieten.



Dieses Dashboard sammelt Folgendes: KPIs

KPI	Beschreibung	Abfrage
Anzahl der Dateien	Die Anzahl der Dateien in der Iceberg-Datenbank (für alle Snapshots)	Dieses Dashboard verwendet die Abfragen auf Tabellenebene aus dem vorherigen Abschnitt und konsolidiert die Ergebnisse.
Anzahl der aktiven Dateien	Die Anzahl der aktiven Dateien in der Iceberg-Datenbank (basierend auf den letzten Snapshots der Iceberg-Tabellen)	
Durchschnittliche Dateigröße	Die durchschnittliche Dateigröße in Megabyte für alle Dateien in der Iceberg-Datenbank	
Durchschnittliche Größe der aktiven Datei	Die durchschnittliche Dateigröße in Megabyte für alle aktiven Dateien in der Iceberg-Datenbank	

KPI	Beschreibung	Abfrage
Prozentsatz kleiner Dateien	Der Prozentsatz der aktiven Dateien, die kleiner als 100 MB sind, in der Iceberg-Datenbank	
Gesamtspeichergröße	Die Gesamtgröße aller Dateien in der Datenbank, ausgenommen verwaiste Dateien und Amazon S3 S3-Objektversionen (falls aktiviert)	
Gesamtgröße des aktiven Speichers	Die Gesamtgröße aller Dateien in den aktuellen Snapshots aller Tabellen in der Datenbank	

Präventive Wartung

Durch die Einrichtung der in den vorherigen Abschnitten erläuterten Überwachungsfunktionen können Sie die Tabellenpflege aus einem präventiven und nicht aus einem reaktiven Blickwinkel betrachten. Sie können beispielsweise die Metriken auf Tabellen- und Datenbankebene verwenden, um Aktionen wie die folgenden zu planen:

- Verwenden Sie die Bin-Packing-Komprimierung, um kleine Dateien zu gruppieren, wenn eine Tabelle N kleine Dateien erreicht.
- Verwenden Sie die Bin-Packing-Komprimierung, um gelöschte Dateien zusammenzuführen, wenn eine Tabelle N Löschdateien in einer bestimmten Partition erreicht.
- Entfernen Sie kleine Dateien, die bereits komprimiert wurden, indem Sie Snapshots entfernen, wenn der Gesamtspeicher X-mal höher ist als der aktive Speicher.

Verwaltung und Zugriffskontrolle für Apache Iceberg auf AWS

Apache Iceberg lässt sich integrieren AWS Lake Formation , um die Datenverwaltung zu vereinfachen. Diese Integration ermöglicht es Data Lake-Administratoren, Iceberg-Tabellen Zugriffsberechtigungen auf Zellebene zuzuweisen. Ein Beispiel für das Abfragen von Iceberg-Tabellen mithilfe von Amazon Athena und AWS Lake Formation finden Sie im AWS Blogbeitrag [Interagieren Sie mit Apache Iceberg-Tabellen mithilfe von Amazon Athena und kontoübergreifenden, fein abgestuften Berechtigungen mithilfe von](#) AWS Lake Formation

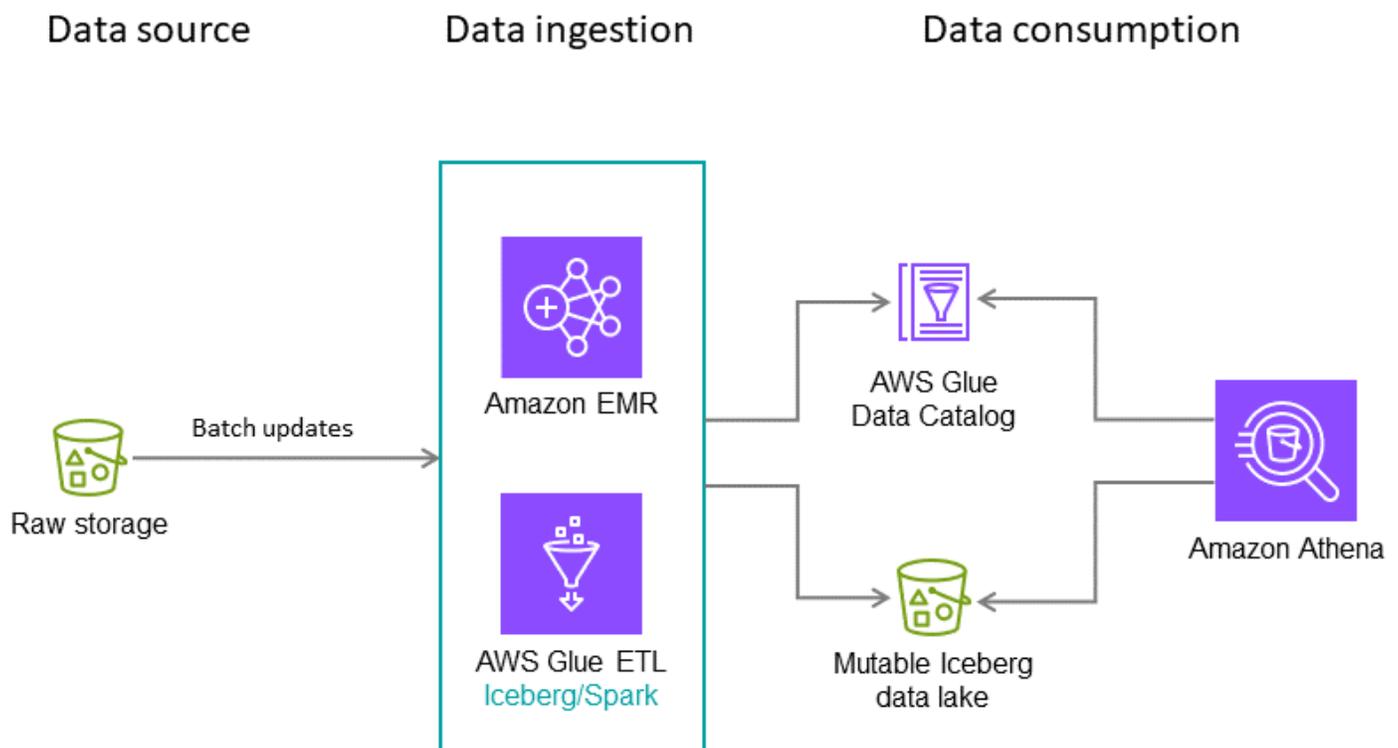
Referenzarchitekturen für Apache Iceberg auf AWS

Dieser Abschnitt enthält Beispiele für die Anwendung von Best Practices in verschiedenen Anwendungsfällen wie Batch-Ingestion und einem Data Lake, der Batch- und Streaming-Datenaufnahme kombiniert.

Nächtliche Batch-Erfassung

Nehmen wir für diesen hypothetischen Anwendungsfall an, dass Ihre Iceberg-Tabelle jede Nacht Kreditkartentransaktionen aufnimmt. Jeder Stapel enthält nur inkrementelle Aktualisierungen, die mit der Zieltabelle zusammengeführt werden müssen. Mehrmals pro Jahr werden vollständige historische Daten empfangen. Für dieses Szenario empfehlen wir die folgende Architektur und Konfigurationen.

Hinweis: Dies ist nur ein Beispiel. Die optimale Konfiguration hängt von Ihren Daten und Anforderungen ab.



Empfehlungen:

- Dateigröße: 128 MB, da Apache Spark-Aufgaben Daten in 128-MB-Blöcken verarbeiten.

- Schreibtyp: copy-on-write Wie bereits weiter oben in diesem Handbuch beschrieben, trägt dieser Ansatz dazu bei, dass Daten leseoptimiert geschrieben werden.
- Partitionsvariablen: Jahr/Monat/Tag. In unserem hypothetischen Anwendungsfall fragen wir am häufigsten aktuelle Daten ab, obwohl wir gelegentlich vollständige Tabellenscans für die Daten der letzten zwei Jahre durchführen. Das Ziel der Partitionierung besteht darin, schnelle Lesevorgänge auf der Grundlage der Anforderungen des Anwendungsfalls zu ermöglichen.
- Sortierreihenfolge: Zeitstempel
- Datenkatalog: AWS Glue Data Catalog

Data Lake, der Batch-Erfassung und Erfassung nahezu in Echtzeit kombiniert

Sie können einen Data Lake auf Amazon S3 bereitstellen, der Batch- und Streaming-Daten über Konten und Regionen hinweg gemeinsam nutzt. Ein Architekturdiagramm und weitere Informationen finden Sie im AWS Blogbeitrag [Build a transactional data lake using Apache Iceberg and Cross-account data shares using AWS Lake Formation](#) Amazon Athena. AWS Glue

Ressourcen

- [Verwendung des Iceberg-Frameworks in AWS Glue\(Dokumentation\)](#) AWS Glue
- [Iceberg](#) (Amazon EMR-Dokumentation)
- [Verwenden von Apache Iceberg-Tabellen](#) (Amazon Athena Athena-Dokumentation)
- [Amazon S3 S3-Dokumentation](#)
- [QuickSight Dokumentation](#)
- [Replikation von Glue Catalog und Lake Formation Permissions](#) (GitHub Repository)
- [Dokumentation zu Apache Iceberg](#)
- [Apache Spark-Dokumentation](#)

Mitwirkende

Die folgenden Personen haben diesen Leitfaden AWS verfasst, mitverfasst und rezensiert.

Mitwirkende

- Carlos Rodrigues, Lösungsarchitekt, Big Data
- Imtiaz (Taz) Sayed, Lösungsarchitekt, Technischer Leiter, Analytik
- Shana Schipers, Lösungsarchitektin für Big Data
- Prashant Singh, Softwareentwicklungsingenieur, Amazon EMR
- Stefano Sandona, Lösungsarchitekt, Big Data
- Arun A K, Lösungsarchitekt für Big Data und ETL
- Francisco Morillo, Lösungsarchitekt, Streaming
- Suthan Phillips, Analytics-Architekt, Amazon EMR
- Sercan Karaoglu, Lösungsarchitekt
- Yonatan Dolan, Analysespezialist
- Guy Bachar, Lösungsarchitekt
- Sofia Zilberman, Lösungsarchitektin, Streaming
- Ismail Makhoulouf, Lösungsarchitekt, Analytik
- Dan Stair, spezialisierter Lösungsarchitekt
- Sakti Mishra, Lösungsarchitekt

Rezensenten

- Rick Sears, Geschäftsführer, Amazon EMR
- Linda O'Connor, EMR-Spezialistin bei Amazon
- Ian Meyers, Direktor von Amazon EMR
- Vinita Ananth, Direktorin, Amazon EMR-Produktmanagement
- Jason Berkowitz, Produktmanager, AWS Lake Formation
- Mahesh Mishra, Produktmanager, Amazon Redshift
- Vladimir Zlatkin, Manager, Lösungsarchitektur, Big Data
- Karthik Prabhakar, Analytics-Architekt, Amazon EMR

- Jack Ye, Softwareentwicklungsingenieur, Amazon EMR
- Vijay Jain, Produktmanager
- Anupriti Warade, Produktmanager, Amazon S3
- Molly Brown, Geschäftsführerin, AWS Lake Formation
- Ajit Tandale, Lösungsarchitekt, Daten
- Gwen Chen, Produktmarketing-Managerin

Dokumentverlauf

In der folgenden Tabelle werden wichtige Änderungen in diesem Leitfaden beschrieben. Um Benachrichtigungen über zukünftige Aktualisierungen zu erhalten, können Sie einen [RSS-Feed](#) abonnieren.

Änderung	Beschreibung	Datum
Ergänzungen	Es wurde ein neuer Abschnitt über die Arbeit mit Iceberg-Tabellen mithilfe von Amazon Data Firehose hinzugefügt.	20. Februar 2025
Erste Veröffentlichung	—	30. April 2024

AWS Glossar zu präskriptiven Leitlinien

Die folgenden Begriffe werden häufig in Strategien, Leitfäden und Mustern von AWS Prescriptive Guidance verwendet. Um Einträge vorzuschlagen, verwenden Sie bitte den Link Feedback geben am Ende des Glossars.

Zahlen

7 Rs

Sieben gängige Migrationsstrategien für die Verlagerung von Anwendungen in die Cloud. Diese Strategien bauen auf den 5 Rs auf, die Gartner 2011 identifiziert hat, und bestehen aus folgenden Elementen:

- Faktorwechsel/Architekturwechsel – Verschieben Sie eine Anwendung und ändern Sie ihre Architektur, indem Sie alle Vorteile cloudnativer Feature nutzen, um Agilität, Leistung und Skalierbarkeit zu verbessern. Dies beinhaltet in der Regel die Portierung des Betriebssystems und der Datenbank. Beispiel: Migrieren Sie Ihre lokale Oracle-Datenbank auf die Amazon Aurora PostgreSQL-kompatible Edition.
- Plattformwechsel (Lift and Reshape) – Verschieben Sie eine Anwendung in die Cloud und führen Sie ein gewisses Maß an Optimierung ein, um die Cloud-Funktionen zu nutzen. Beispiel: Migrieren Sie Ihre lokale Oracle-Datenbank zu Amazon Relational Database Service (Amazon RDS) für Oracle in der AWS Cloud
- Neukauf (Drop and Shop) – Wechseln Sie zu einem anderen Produkt, indem Sie typischerweise von einer herkömmlichen Lizenz zu einem SaaS-Modell wechseln. Beispiel: Migrieren Sie Ihr CRM-System (Customer Relationship Management) zu Salesforce.com.
- Hostwechsel (Lift and Shift) – Verschieben Sie eine Anwendung in die Cloud, ohne Änderungen vorzunehmen, um die Cloud-Funktionen zu nutzen. Beispiel: Migrieren Sie Ihre lokale Oracle-Datenbank zu Oracle auf einer EC2 Instanz in der AWS Cloud
- Verschieben (Lift and Shift auf Hypervisor-Ebene) – Verlagern Sie die Infrastruktur in die Cloud, ohne neue Hardware kaufen, Anwendungen umschreiben oder Ihre bestehenden Abläufe ändern zu müssen. Sie migrieren Server von einer lokalen Plattform zu einem Cloud-Dienst für dieselbe Plattform. Beispiel: Migrieren Sie eine Microsoft Hyper-V Anwendung zu AWS.
- Beibehaltung (Wiederaufgreifen) – Bewahren Sie Anwendungen in Ihrer Quellumgebung auf. Dazu können Anwendungen gehören, die einen umfangreichen Faktorwechsel erfordern und

die Sie auf einen späteren Zeitpunkt verschieben möchten, sowie ältere Anwendungen, die Sie beibehalten möchten, da es keine geschäftliche Rechtfertigung für ihre Migration gibt.

- Außerbetriebnahme – Dekommissionierung oder Entfernung von Anwendungen, die in Ihrer Quellumgebung nicht mehr benötigt werden.

A

ABAC

Siehe [attributbasierte](#) Zugriffskontrolle.

abstrahierte Dienste

Weitere Informationen finden Sie unter [Managed Services](#).

ACID

Siehe [Atomarität, Konsistenz, Isolierung und Haltbarkeit](#).

Aktiv-Aktiv-Migration

Eine Datenbankmigrationsmethode, bei der die Quell- und Zieldatenbanken synchron gehalten werden (mithilfe eines bidirektionalen Replikationstools oder dualer Schreibvorgänge) und beide Datenbanken Transaktionen von miteinander verbundenen Anwendungen während der Migration verarbeiten. Diese Methode unterstützt die Migration in kleinen, kontrollierten Batches, anstatt einen einmaligen Cutover zu erfordern. Es ist flexibler, erfordert aber mehr Arbeit als eine [aktiv-passive](#) Migration.

Aktiv-Passiv-Migration

Eine Datenbankmigrationsmethode, bei der die Quell- und Zieldatenbanken synchron gehalten werden, aber nur die Quelldatenbank Transaktionen von verbindenden Anwendungen verarbeitet, während Daten in die Zieldatenbank repliziert werden. Die Zieldatenbank akzeptiert während der Migration keine Transaktionen.

Aggregatfunktion

Eine SQL-Funktion, die mit einer Gruppe von Zeilen arbeitet und einen einzelnen Rückgabewert für die Gruppe berechnet. Beispiele für Aggregatfunktionen sind SUM und MAX.

AI

Siehe [künstliche Intelligenz](#).

AIOps

Siehe [Operationen im Bereich künstliche Intelligenz](#).

Anonymisierung

Der Prozess des dauerhaften Löschens personenbezogener Daten in einem Datensatz. Anonymisierung kann zum Schutz der Privatsphäre beitragen. Anonymisierte Daten gelten nicht mehr als personenbezogene Daten.

Anti-Muster

Eine häufig verwendete Lösung für ein wiederkehrendes Problem, bei dem die Lösung kontraproduktiv, ineffektiv oder weniger wirksam als eine Alternative ist.

Anwendungssteuerung

Ein Sicherheitsansatz, bei dem nur zugelassene Anwendungen verwendet werden können, um ein System vor Schadsoftware zu schützen.

Anwendungsportfolio

Eine Sammlung detaillierter Informationen zu jeder Anwendung, die von einer Organisation verwendet wird, einschließlich der Kosten für die Erstellung und Wartung der Anwendung und ihres Geschäftswerts. Diese Informationen sind entscheidend für [den Prozess der Portfoliofindung und -analyse](#) und hilft bei der Identifizierung und Priorisierung der Anwendungen, die migriert, modernisiert und optimiert werden sollen.

künstliche Intelligenz (KI)

Das Gebiet der Datenverarbeitungswissenschaft, das sich der Nutzung von Computertechnologien zur Ausführung kognitiver Funktionen widmet, die typischerweise mit Menschen in Verbindung gebracht werden, wie Lernen, Problemlösen und Erkennen von Mustern. Weitere Informationen finden Sie unter [Was ist künstliche Intelligenz?](#)

Operationen mit künstlicher Intelligenz (AIOps)

Der Prozess des Einsatzes von Techniken des Machine Learning zur Lösung betrieblicher Probleme, zur Reduzierung betrieblicher Zwischenfälle und menschlicher Eingriffe sowie zur Steigerung der Servicequalität. Weitere Informationen zur Verwendung in der AWS Migrationsstrategie finden Sie im [Operations Integration Guide](#). AIOps

Asymmetrische Verschlüsselung

Ein Verschlüsselungsalgorithmus, der ein Schlüsselpaar, einen öffentlichen Schlüssel für die Verschlüsselung und einen privaten Schlüssel für die Entschlüsselung verwendet. Sie können den

öffentlichen Schlüssel teilen, da er nicht für die Entschlüsselung verwendet wird. Der Zugriff auf den privaten Schlüssel sollte jedoch stark eingeschränkt sein.

Atomizität, Konsistenz, Isolierung, Haltbarkeit (ACID)

Eine Reihe von Softwareeigenschaften, die die Datenvalidität und betriebliche Zuverlässigkeit einer Datenbank auch bei Fehlern, Stromausfällen oder anderen Problemen gewährleisten.

Attributbasierte Zugriffskontrolle (ABAC)

Die Praxis, detaillierte Berechtigungen auf der Grundlage von Benutzerattributen wie Abteilung, Aufgabenrolle und Teamname zu erstellen. Weitere Informationen finden Sie unter [ABAC AWS](#) in der AWS Identity and Access Management (IAM-) Dokumentation.

autoritative Datenquelle

Ein Ort, an dem Sie die primäre Version der Daten speichern, die als die zuverlässigste Informationsquelle angesehen wird. Sie können Daten aus der maßgeblichen Datenquelle an andere Speicherorte kopieren, um die Daten zu verarbeiten oder zu ändern, z. B. zu anonymisieren, zu redigieren oder zu pseudonymisieren.

Availability Zone

Ein bestimmter Standort innerhalb einer AWS-Region, der vor Ausfällen in anderen Availability Zones geschützt ist und kostengünstige Netzwerkkonnektivität mit niedriger Latenz zu anderen Availability Zones in derselben Region bietet.

AWS Framework für die Cloud-Einführung (AWS CAF)

Ein Framework mit Richtlinien und bewährten Verfahren, das Unternehmen bei der Entwicklung eines effizienten und effektiven Plans für den erfolgreichen Umstieg auf die Cloud unterstützt. AWS CAF unterteilt die Leitlinien in sechs Schwerpunktbereiche, die als Perspektiven bezeichnet werden: Unternehmen, Mitarbeiter, Unternehmensführung, Plattform, Sicherheit und Betrieb. Die Perspektiven Geschäft, Mitarbeiter und Unternehmensführung konzentrieren sich auf Geschäftskompetenzen und -prozesse, während sich die Perspektiven Plattform, Sicherheit und Betriebsabläufe auf technische Fähigkeiten und Prozesse konzentrieren. Die Personalperspektive zielt beispielsweise auf Stakeholder ab, die sich mit Personalwesen (HR), Personalfunktionen und Personalmanagement befassen. Aus dieser Perspektive bietet AWS CAF Leitlinien für Personalentwicklung, Schulung und Kommunikation, um das Unternehmen auf eine erfolgreiche Cloud-Einführung vorzubereiten. Weitere Informationen finden Sie auf der [AWS -CAF-Webseite](#) und dem [AWS -CAF-Whitepaper](#).

AWS Workload-Qualifizierungsrahmen (AWS WQF)

Ein Tool, das Workloads bei der Datenbankmigration bewertet, Migrationsstrategien empfiehlt und Arbeitsschätzungen bereitstellt. AWS WQF ist in () enthalten. AWS Schema Conversion Tool AWS SCT Es analysiert Datenbankschemas und Codeobjekte, Anwendungscode, Abhängigkeiten und Leistungsmerkmale und stellt Bewertungsberichte bereit.

B

schlechter Bot

Ein [Bot](#), der Einzelpersonen oder Organisationen stören oder ihnen Schaden zufügen soll.

BCP

Siehe [Planung der Geschäftskontinuität](#).

Verhaltensdiagramm

Eine einheitliche, interaktive Ansicht des Ressourcenverhaltens und der Interaktionen im Laufe der Zeit. Sie können ein Verhaltensdiagramm mit Amazon Detective verwenden, um fehlgeschlagene Anmeldeversuche, verdächtige API-Aufrufe und ähnliche Vorgänge zu untersuchen. Weitere Informationen finden Sie unter [Daten in einem Verhaltensdiagramm](#) in der Detective-Dokumentation.

Big-Endian-System

Ein System, welches das höchstwertige Byte zuerst speichert. Siehe auch [Endianness](#).

Binäre Klassifikation

Ein Prozess, der ein binäres Ergebnis vorhersagt (eine von zwei möglichen Klassen). Beispielsweise könnte Ihr ML-Modell möglicherweise Probleme wie „Handelt es sich bei dieser E-Mail um Spam oder nicht?“ vorhersagen müssen oder „Ist dieses Produkt ein Buch oder ein Auto?“

Bloom-Filter

Eine probabilistische, speichereffiziente Datenstruktur, mit der getestet wird, ob ein Element Teil einer Menge ist.

Blau/Grün-Bereitstellung

Eine Bereitstellungsstrategie, bei der Sie zwei separate, aber identische Umgebungen erstellen. Sie führen die aktuelle Anwendungsversion in einer Umgebung (blau) und die neue

Anwendungsversion in der anderen Umgebung (grün) aus. Mit dieser Strategie können Sie schnell und mit minimalen Auswirkungen ein Rollback durchführen.

Bot

Eine Softwareanwendung, die automatisierte Aufgaben über das Internet ausführt und menschliche Aktivitäten oder Interaktionen simuliert. Manche Bots sind nützlich oder nützlich, wie z. B. Webcrawler, die Informationen im Internet indexieren. Einige andere Bots, sogenannte böartige Bots, sollen Einzelpersonen oder Organisationen stören oder ihnen Schaden zufügen.

Botnetz

Netzwerke von [Bots](#), die mit [Malware](#) infiziert sind und unter der Kontrolle einer einzigen Partei stehen, die als Bot-Herder oder Bot-Operator bezeichnet wird. Botnetze sind der bekannteste Mechanismus zur Skalierung von Bots und ihrer Wirkung.

branch

Ein containerisierter Bereich eines Code-Repositorys. Der erste Zweig, der in einem Repository erstellt wurde, ist der Hauptzweig. Sie können einen neuen Zweig aus einem vorhandenen Zweig erstellen und dann Feature entwickeln oder Fehler in dem neuen Zweig beheben. Ein Zweig, den Sie erstellen, um ein Feature zu erstellen, wird allgemein als Feature-Zweig bezeichnet. Wenn das Feature zur Veröffentlichung bereit ist, führen Sie den Feature-Zweig wieder mit dem Hauptzweig zusammen. Weitere Informationen finden Sie unter [Über Branches](#) (GitHub Dokumentation).

Zugang durch Glasbruch

Unter außergewöhnlichen Umständen und im Rahmen eines genehmigten Verfahrens ist dies eine schnelle Methode für einen Benutzer, auf einen Bereich zuzugreifen AWS-Konto, für den er in der Regel keine Zugriffsrechte besitzt. Weitere Informationen finden Sie unter dem Indikator [Implementation break-glass procedures](#) in den AWS Well-Architected-Leitlinien.

Brownfield-Strategie

Die bestehende Infrastruktur in Ihrer Umgebung. Wenn Sie eine Brownfield-Strategie für eine Systemarchitektur anwenden, richten Sie sich bei der Gestaltung der Architektur nach den Einschränkungen der aktuellen Systeme und Infrastruktur. Wenn Sie die bestehende Infrastruktur erweitern, könnten Sie Brownfield- und [Greenfield](#)-Strategien mischen.

Puffer-Cache

Der Speicherbereich, in dem die am häufigsten abgerufenen Daten gespeichert werden.

Geschäftsfähigkeit

Was ein Unternehmen tut, um Wert zu generieren (z. B. Vertrieb, Kundenservice oder Marketing). Microservices-Architekturen und Entwicklungsentscheidungen können von den Geschäftskapazitäten beeinflusst werden. Weitere Informationen finden Sie im Abschnitt [Organisiert nach Geschäftskapazitäten](#) des Whitepapers [Ausführen von containerisierten Microservices in AWS](#).

Planung der Geschäftskontinuität (BCP)

Ein Plan, der die potenziellen Auswirkungen eines störenden Ereignisses, wie z. B. einer groß angelegten Migration, auf den Betrieb berücksichtigt und es einem Unternehmen ermöglicht, den Betrieb schnell wieder aufzunehmen.

C

CAF

Weitere Informationen finden Sie unter [Framework für die AWS Cloud-Einführung](#).

Bereitstellung auf Kanaren

Die langsame und schrittweise Veröffentlichung einer Version für Endbenutzer. Wenn Sie sich sicher sind, stellen Sie die neue Version bereit und ersetzen die aktuelle Version vollständig.

CCoE

Weitere Informationen finden Sie [im Cloud Center of Excellence](#).

CDC

Siehe [Erfassung von Änderungsdaten](#).

Erfassung von Datenänderungen (CDC)

Der Prozess der Nachverfolgung von Änderungen an einer Datenquelle, z. B. einer Datenbanktabelle, und der Aufzeichnung von Metadaten zu der Änderung. Sie können CDC für verschiedene Zwecke verwenden, z. B. für die Prüfung oder Replikation von Änderungen in einem Zielsystem, um die Synchronisation aufrechtzuerhalten.

Chaos-Technik

Absichtliches Einführen von Ausfällen oder Störungsereignissen, um die Widerstandsfähigkeit eines Systems zu testen. Sie können [AWS Fault Injection Service \(AWS FIS\)](#) verwenden, um Experimente durchzuführen, die Ihre AWS Workloads stressen, und deren Reaktion zu bewerten.

CI/CD

Siehe [Continuous Integration und Continuous Delivery](#).

Klassifizierung

Ein Kategorisierungsprozess, der bei der Erstellung von Vorhersagen hilft. ML-Modelle für Klassifikationsprobleme sagen einen diskreten Wert voraus. Diskrete Werte unterscheiden sich immer voneinander. Beispielsweise muss ein Modell möglicherweise auswerten, ob auf einem Bild ein Auto zu sehen ist oder nicht.

clientseitige Verschlüsselung

Lokale Verschlüsselung von Daten, bevor das Ziel sie AWS-Service empfängt.

Cloud-Exzellenzzentrum (CCoE)

Ein multidisziplinäres Team, das die Cloud-Einführung in der gesamten Organisation vorantreibt, einschließlich der Entwicklung bewährter Cloud-Methoden, der Mobilisierung von Ressourcen, der Festlegung von Migrationszeitplänen und der Begleitung der Organisation durch groß angelegte Transformationen. Weitere Informationen finden Sie in den [CCoE-Beiträgen](#) im AWS Cloud Enterprise Strategy Blog.

Cloud Computing

Die Cloud-Technologie, die typischerweise für die Ferndatenspeicherung und das IoT-Gerätemanagement verwendet wird. Cloud Computing ist häufig mit [Edge-Computing-Technologie](#) verbunden.

Cloud-Betriebsmodell

In einer IT-Organisation das Betriebsmodell, das zum Aufbau, zur Weiterentwicklung und Optimierung einer oder mehrerer Cloud-Umgebungen verwendet wird. Weitere Informationen finden Sie unter [Aufbau Ihres Cloud-Betriebsmodells](#).

Phasen der Einführung der Cloud

Die vier Phasen, die Unternehmen bei der Migration in der Regel durchlaufen AWS Cloud:

- Projekt – Durchführung einiger Cloud-bezogener Projekte zu Machbarkeitsnachweisen und zu Lernzwecken
- Fundament — Tätigen Sie grundlegende Investitionen, um Ihre Cloud-Einführung zu skalieren (z. B. Einrichtung einer landing zone, Definition eines CCo E, Einrichtung eines Betriebsmodells)

- Migration – Migrieren einzelner Anwendungen
- Neuentwicklung – Optimierung von Produkten und Services und Innovation in der Cloud

Diese Phasen wurden von Stephen Orban im Blogbeitrag [The Journey Toward Cloud-First & the Stages of Adoption](#) im AWS Cloud Enterprise Strategy-Blog definiert. Informationen darüber, wie sie mit der AWS Migrationsstrategie zusammenhängen, finden Sie im Leitfaden zur Vorbereitung der [Migration](#).

CMDB

Siehe [Datenbank für das Konfigurationsmanagement](#).

Code-Repository

Ein Ort, an dem Quellcode und andere Komponenten wie Dokumentation, Beispiele und Skripts gespeichert und im Rahmen von Versionskontrollprozessen aktualisiert werden. Zu den gängigen Cloud-Repositorys gehören GitHub oder Bitbucket Cloud. Jede Version des Codes wird Zweig genannt. In einer Microservice-Struktur ist jedes Repository einer einzelnen Funktionalität gewidmet. Eine einzelne CI/CD-Pipeline kann mehrere Repositorien verwenden.

Kalter Cache

Ein Puffer-Cache, der leer oder nicht gut gefüllt ist oder veraltete oder irrelevante Daten enthält. Dies beeinträchtigt die Leistung, da die Datenbank-Instance aus dem Hauptspeicher oder der Festplatte lesen muss, was langsamer ist als das Lesen aus dem Puffercache.

Kalte Daten

Daten, auf die selten zugegriffen wird und die in der Regel historisch sind. Bei der Abfrage dieser Art von Daten sind langsame Abfragen in der Regel akzeptabel. Durch die Verlagerung dieser Daten auf leistungsschwächere und kostengünstigere Speicherstufen oder -klassen können Kosten gesenkt werden.

Computer Vision (CV)

Ein Bereich der [KI](#), der maschinelles Lernen nutzt, um Informationen aus visuellen Formaten wie digitalen Bildern und Videos zu analysieren und zu extrahieren. Amazon SageMaker AI bietet beispielsweise Bildverarbeitungsalgorithmen für CV.

Drift in der Konfiguration

Bei einer Arbeitslast eine Änderung der Konfiguration gegenüber dem erwarteten Zustand. Dies kann dazu führen, dass der Workload nicht mehr richtlinienkonform wird, und zwar in der Regel schrittweise und unbeabsichtigt.

Verwaltung der Datenbankkonfiguration (CMDB)

Ein Repository, das Informationen über eine Datenbank und ihre IT-Umgebung speichert und verwaltet, inklusive Hardware- und Softwarekomponenten und deren Konfigurationen. In der Regel verwenden Sie Daten aus einer CMDB in der Phase der Portfolioerkennung und -analyse der Migration.

Konformitätspaket

Eine Sammlung von AWS Config Regeln und Abhilfemaßnahmen, die Sie zusammenstellen können, um Ihre Konformitäts- und Sicherheitsprüfungen individuell anzupassen. Mithilfe einer YAML-Vorlage können Sie ein Conformance Pack als einzelne Entität in einer AWS-Konto AND-Region oder unternehmensweit bereitstellen. Weitere Informationen finden Sie in der Dokumentation unter [Conformance Packs](#). AWS Config

Kontinuierliche Bereitstellung und kontinuierliche Integration (CI/CD)

Der Prozess der Automatisierung der Quell-, Build-, Test-, Staging- und Produktionsphasen des Softwareveröffentlichungsprozesses. CI/CD is commonly described as a pipeline. CI/CD kann Ihnen helfen, Prozesse zu automatisieren, die Produktivität zu steigern, die Codequalität zu verbessern und schneller zu liefern. Weitere Informationen finden Sie unter [Vorteile der kontinuierlichen Auslieferung](#). CD kann auch für kontinuierliche Bereitstellung stehen. Weitere Informationen finden Sie unter [Kontinuierliche Auslieferung im Vergleich zu kontinuierlicher Bereitstellung](#).

CV

Siehe [Computer Vision](#).

D

Daten im Ruhezustand

Daten, die in Ihrem Netzwerk stationär sind, z. B. Daten, die sich im Speicher befinden.

Datenklassifizierung

Ein Prozess zur Identifizierung und Kategorisierung der Daten in Ihrem Netzwerk auf der Grundlage ihrer Kritikalität und Sensitivität. Sie ist eine wichtige Komponente jeder Strategie für das Management von Cybersecurity-Risiken, da sie Ihnen hilft, die geeigneten Schutz- und Aufbewahrungskontrollen für die Daten zu bestimmen. Die Datenklassifizierung ist ein Bestandteil

der Sicherheitssäule im AWS Well-Architected Framework. Weitere Informationen finden Sie unter [Datenklassifizierung](#).

Datendrift

Eine signifikante Variation zwischen den Produktionsdaten und den Daten, die zum Trainieren eines ML-Modells verwendet wurden, oder eine signifikante Änderung der Eingabedaten im Laufe der Zeit. Datendrift kann die Gesamtqualität, Genauigkeit und Fairness von ML-Modellvorhersagen beeinträchtigen.

Daten während der Übertragung

Daten, die sich aktiv durch Ihr Netzwerk bewegen, z. B. zwischen Netzwerkressourcen.

Datennetz

Ein architektonisches Framework, das verteilte, dezentrale Dateneigentum mit zentraler Verwaltung und Steuerung ermöglicht.

Datenminimierung

Das Prinzip, nur die Daten zu sammeln und zu verarbeiten, die unbedingt erforderlich sind. Durch Datenminimierung im AWS Cloud können Datenschutzrisiken, Kosten und der CO2-Fußabdruck Ihrer Analysen reduziert werden.

Datenperimeter

Eine Reihe präventiver Schutzmaßnahmen in Ihrer AWS Umgebung, die sicherstellen, dass nur vertrauenswürdige Identitäten auf vertrauenswürdige Ressourcen von erwarteten Netzwerken zugreifen. Weitere Informationen finden Sie unter [Aufbau eines Datenperimeters](#) auf AWS

Vorverarbeitung der Daten

Rohdaten in ein Format umzuwandeln, das von Ihrem ML-Modell problemlos verarbeitet werden kann. Die Vorverarbeitung von Daten kann bedeuten, dass bestimmte Spalten oder Zeilen entfernt und fehlende, inkonsistente oder doppelte Werte behoben werden.

Herkunft der Daten

Der Prozess der Nachverfolgung des Ursprungs und der Geschichte von Daten während ihres gesamten Lebenszyklus, z. B. wie die Daten generiert, übertragen und gespeichert wurden.

betreffene Person

Eine Person, deren Daten gesammelt und verarbeitet werden.

Data Warehouse

Ein Datenverwaltungssystem, das Business Intelligence wie Analysen unterstützt. Data Warehouses enthalten in der Regel große Mengen historischer Daten und werden in der Regel für Abfragen und Analysen verwendet.

Datenbankdefinitionssprache (DDL)

Anweisungen oder Befehle zum Erstellen oder Ändern der Struktur von Tabellen und Objekten in einer Datenbank.

Datenbankmanipulationssprache (DML)

Anweisungen oder Befehle zum Ändern (Einfügen, Aktualisieren und Löschen) von Informationen in einer Datenbank.

DDL

Siehe [Datenbankdefinitionssprache](#).

Deep-Ensemble

Mehrere Deep-Learning-Modelle zur Vorhersage kombinieren. Sie können Deep-Ensembles verwenden, um eine genauere Vorhersage zu erhalten oder um die Unsicherheit von Vorhersagen abzuschätzen.

Deep Learning

Ein ML-Teilbereich, der mehrere Schichten künstlicher neuronaler Netzwerke verwendet, um die Zuordnung zwischen Eingabedaten und Zielvariablen von Interesse zu ermitteln.

defense-in-depth

Ein Ansatz zur Informationssicherheit, bei dem eine Reihe von Sicherheitsmechanismen und -kontrollen sorgfältig in einem Computernetzwerk verteilt werden, um die Vertraulichkeit, Integrität und Verfügbarkeit des Netzwerks und der darin enthaltenen Daten zu schützen. Wenn Sie diese Strategie anwenden AWS, fügen Sie mehrere Steuerelemente auf verschiedenen Ebenen der AWS Organizations Struktur hinzu, um die Ressourcen zu schützen. Ein defense-in-depth Ansatz könnte beispielsweise Multi-Faktor-Authentifizierung, Netzwerksegmentierung und Verschlüsselung kombinieren.

delegierter Administrator

In AWS Organizations kann ein kompatibler Dienst ein AWS Mitgliedskonto registrieren, um die Konten der Organisation und die Berechtigungen für diesen Dienst zu verwalten. Dieses Konto

wird als delegierter Administrator für diesen Service bezeichnet. Weitere Informationen und eine Liste kompatibler Services finden Sie unter [Services, die mit AWS Organizations funktionieren](#) in der AWS Organizations -Dokumentation.

Bereitstellung

Der Prozess, bei dem eine Anwendung, neue Feature oder Codekorrekturen in der Zielumgebung verfügbar gemacht werden. Die Bereitstellung umfasst das Implementieren von Änderungen an einer Codebasis und das anschließende Erstellen und Ausführen dieser Codebasis in den Anwendungsumgebungen.

Entwicklungsumgebung

Siehe [Umgebung](#).

Detektivische Kontrolle

Eine Sicherheitskontrolle, die darauf ausgelegt ist, ein Ereignis zu erkennen, zu protokollieren und zu warnen, nachdem ein Ereignis eingetreten ist. Diese Kontrollen stellen eine zweite Verteidigungslinie dar und warnen Sie vor Sicherheitsereignissen, bei denen die vorhandenen präventiven Kontrollen umgangen wurden. Weitere Informationen finden Sie unter [Detektivische Kontrolle](#) in Implementierung von Sicherheitskontrollen in AWS.

Abbildung des Wertstroms in der Entwicklung (DVSM)

Ein Prozess zur Identifizierung und Priorisierung von Einschränkungen, die sich negativ auf Geschwindigkeit und Qualität im Lebenszyklus der Softwareentwicklung auswirken. DVSM erweitert den Prozess der Wertstromanalyse, der ursprünglich für Lean-Manufacturing-Praktiken konzipiert wurde. Es konzentriert sich auf die Schritte und Teams, die erforderlich sind, um durch den Softwareentwicklungsprozess Mehrwert zu schaffen und zu steigern.

digitaler Zwilling

Eine virtuelle Darstellung eines realen Systems, z. B. eines Gebäudes, einer Fabrik, einer Industrieanlage oder einer Produktionslinie. Digitale Zwillinge unterstützen vorausschauende Wartung, Fernüberwachung und Produktionsoptimierung.

Maßtabelle

In einem [Sternschema](#) eine kleinere Tabelle, die Datenattribute zu quantitativen Daten in einer Faktentabelle enthält. Bei Attributen von Dimensionstabellen handelt es sich in der Regel um Textfelder oder diskrete Zahlen, die sich wie Text verhalten. Diese Attribute werden häufig zum Einschränken von Abfragen, zum Filtern und zur Kennzeichnung von Ergebnismengen verwendet.

Katastrophe

Ein Ereignis, das verhindert, dass ein Workload oder ein System seine Geschäftsziele an seinem primären Einsatzort erfüllt. Diese Ereignisse können Naturkatastrophen, technische Ausfälle oder das Ergebnis menschlichen Handelns sein, z. B. unbeabsichtigte Fehlkonfigurationen oder ein Malware-Angriff.

Disaster Recovery (DR)

Die Strategie und der Prozess, die Sie verwenden, um Ausfallzeiten und Datenverluste aufgrund einer [Katastrophe](#) zu minimieren. Weitere Informationen finden Sie unter [Disaster Recovery von Workloads unter AWS: Wiederherstellung in der Cloud im AWS Well-Architected Framework](#).

DML

Siehe Sprache zur [Datenbankmanipulation](#).

Domainorientiertes Design

Ein Ansatz zur Entwicklung eines komplexen Softwaresystems, bei dem seine Komponenten mit sich entwickelnden Domains oder Kerngeschäftsziele verknüpft werden, denen jede Komponente dient. Dieses Konzept wurde von Eric Evans in seinem Buch Domaingesteuertes Design: Bewältigen der Komplexität im Herzen der Software (Boston: Addison-Wesley Professional, 2003) vorgestellt. Informationen darüber, wie Sie domaingesteuertes Design mit dem Strangler-Fig-Muster verwenden können, finden Sie unter [Schrittweises Modernisieren älterer Microsoft ASP.NET \(ASMX\)-Webservices mithilfe von Containern und Amazon API Gateway](#).

DR

Siehe [Disaster Recovery](#).

Erkennung von Driften

Verfolgung von Abweichungen von einer Basiskonfiguration. Sie können es beispielsweise verwenden, AWS CloudFormation um [Abweichungen bei den Systemressourcen zu erkennen](#), oder Sie können AWS Control Tower damit [Änderungen in Ihrer landing zone erkennen](#), die sich auf die Einhaltung von Governance-Anforderungen auswirken könnten.

DVSM

Siehe [Abbildung des Wertstroms in der Entwicklung](#).

E

EDA

Siehe [explorative Datenanalyse](#).

EDI

Siehe [elektronischer Datenaustausch](#).

Edge-Computing

Die Technologie, die die Rechenleistung für intelligente Geräte an den Rändern eines IoT-Netzwerks erhöht. Im Vergleich zu [Cloud Computing](#) kann Edge Computing die Kommunikationslatenz reduzieren und die Reaktionszeit verbessern.

elektronischer Datenaustausch (EDI)

Der automatisierte Austausch von Geschäftsdokumenten zwischen Organisationen. Weitere Informationen finden Sie unter [Was ist elektronischer Datenaustausch](#).

Verschlüsselung

Ein Rechenprozess, der Klartextdaten, die für Menschen lesbar sind, in Chiffretext umwandelt.

Verschlüsselungsschlüssel

Eine kryptografische Zeichenfolge aus zufälligen Bits, die von einem Verschlüsselungsalgorithmus generiert wird. Schlüssel können unterschiedlich lang sein, und jeder Schlüssel ist so konzipiert, dass er unvorhersehbar und einzigartig ist.

Endianismus

Die Reihenfolge, in der Bytes im Computerspeicher gespeichert werden. Big-Endian-Systeme speichern das höchstwertige Byte zuerst. Little-Endian-Systeme speichern das niedrigwertigste Byte zuerst.

Endpunkt

[Siehe](#) Service-Endpunkt.

Endpunkt-Services

Ein Service, den Sie in einer Virtual Private Cloud (VPC) hosten können, um ihn mit anderen Benutzern zu teilen. Sie können einen Endpunktdienst mit anderen AWS-Konten oder AWS Identity and Access Management (IAM AWS PrivateLink -) Prinzipalen erstellen und diesen

Berechtigungen gewähren. Diese Konten oder Prinzipale können sich privat mit Ihrem Endpunktservice verbinden, indem sie Schnittstellen-VPC-Endpunkte erstellen. Weitere Informationen finden Sie unter [Einen Endpunkt-Service erstellen](#) in der Amazon Virtual Private Cloud (Amazon VPC)-Dokumentation.

Unternehmensressourcenplanung (ERP)

Ein System, das wichtige Geschäftsprozesse (wie Buchhaltung, [MES](#) und Projektmanagement) für ein Unternehmen automatisiert und verwaltet.

Envelope-Verschlüsselung

Der Prozess der Verschlüsselung eines Verschlüsselungsschlüssels mit einem anderen Verschlüsselungsschlüssel. Weitere Informationen finden Sie unter [Envelope-Verschlüsselung](#) in der AWS Key Management Service (AWS KMS) -Dokumentation.

Umgebung

Eine Instance einer laufenden Anwendung. Die folgenden Arten von Umgebungen sind beim Cloud-Computing üblich:

- **Entwicklungsumgebung** – Eine Instance einer laufenden Anwendung, die nur dem Kernteam zur Verfügung steht, das für die Wartung der Anwendung verantwortlich ist. Entwicklungsumgebungen werden verwendet, um Änderungen zu testen, bevor sie in höhere Umgebungen übertragen werden. Diese Art von Umgebung wird manchmal als Testumgebung bezeichnet.
- **Niedrigere Umgebungen** – Alle Entwicklungsumgebungen für eine Anwendung, z. B. solche, die für erste Builds und Tests verwendet wurden.
- **Produktionsumgebung** – Eine Instance einer laufenden Anwendung, auf die Endbenutzer zugreifen können. In einer CI/CD-Pipeline ist die Produktionsumgebung die letzte Bereitstellungsumgebung.
- **Höhere Umgebungen** – Alle Umgebungen, auf die auch andere Benutzer als das Kernentwicklungsteam zugreifen können. Dies kann eine Produktionsumgebung, Vorproduktionsumgebungen und Umgebungen für Benutzerakzeptanztests umfassen.

Epics

In der agilen Methodik sind dies funktionale Kategorien, die Ihnen helfen, Ihre Arbeit zu organisieren und zu priorisieren. Epics bieten eine allgemeine Beschreibung der Anforderungen und Implementierungsaufgaben. Zu den Sicherheitsthemen AWS von CAF gehören beispielsweise Identitäts- und Zugriffsmanagement, Detektivkontrollen, Infrastruktursicherheit,

Datenschutz und Reaktion auf Vorfälle. Weitere Informationen zu Epics in der AWS - Migrationsstrategie finden Sie im [Leitfaden zur Programm-Implementierung](#).

ERP

Siehe [Enterprise Resource Planning](#).

Explorative Datenanalyse (EDA)

Der Prozess der Analyse eines Datensatzes, um seine Hauptmerkmale zu verstehen. Sie sammeln oder aggregieren Daten und führen dann erste Untersuchungen durch, um Muster zu finden, Anomalien zu erkennen und Annahmen zu überprüfen. EDA wird durchgeführt, indem zusammenfassende Statistiken berechnet und Datenvisualisierungen erstellt werden.

F

Faktentabelle

Die zentrale Tabelle in einem [Sternschema](#). Sie speichert quantitative Daten über den Geschäftsbetrieb. In der Regel enthält eine Faktentabelle zwei Arten von Spalten: Spalten, die Kennzahlen enthalten, und Spalten, die einen Fremdschlüssel für eine Dimensionstabelle enthalten.

schnell scheitern

Eine Philosophie, die häufige und inkrementelle Tests verwendet, um den Entwicklungslebenszyklus zu verkürzen. Dies ist ein wichtiger Bestandteil eines agilen Ansatzes.

Grenze zur Fehlerisolierung

Dabei handelt es sich um eine Grenze AWS Cloud, z. B. eine Availability Zone AWS-Region, eine Steuerungsebene oder eine Datenebene, die die Auswirkungen eines Fehlers begrenzt und die Widerstandsfähigkeit von Workloads verbessert. Weitere Informationen finden Sie unter [Grenzen zur AWS Fehlerisolierung](#).

Feature-Zweig

Siehe [Zweig](#).

Features

Die Eingabedaten, die Sie verwenden, um eine Vorhersage zu treffen. In einem Fertigungskontext könnten Feature beispielsweise Bilder sein, die regelmäßig von der Fertigungslinie aus aufgenommen werden.

Bedeutung der Feature

Wie wichtig ein Feature für die Vorhersagen eines Modells ist. Dies wird in der Regel als numerischer Wert ausgedrückt, der mit verschiedenen Techniken wie Shapley Additive Explanations (SHAP) und integrierten Gradienten berechnet werden kann. Weitere Informationen finden Sie unter [Interpretierbarkeit von Modellen für maschinelles Lernen mit AWS](#).

Featuretransformation

Daten für den ML-Prozess optimieren, einschließlich der Anreicherung von Daten mit zusätzlichen Quellen, der Skalierung von Werten oder der Extraktion mehrerer Informationssätze aus einem einzigen Datenfeld. Das ermöglicht dem ML-Modell, von den Daten profitieren. Wenn Sie beispielsweise das Datum „27.05.2021 00:15:37“ in „2021“, „Mai“, „Donnerstag“ und „15“ aufschlüsseln, können Sie dem Lernalgorithmus helfen, nuancierte Muster zu erlernen, die mit verschiedenen Datenkomponenten verknüpft sind.

Eingabeaufforderung mit wenigen Klicks

Bereitstellung einer kleinen Anzahl von Beispielen, die die Aufgabe und das gewünschte Ergebnis veranschaulichen, bevor das [LLM](#) aufgefordert wird, eine ähnliche Aufgabe auszuführen. Bei dieser Technik handelt es sich um eine Anwendung des kontextbezogenen Lernens, bei der Modelle anhand von Beispielen (Aufnahmen) lernen, die in Eingabeaufforderungen eingebettet sind. Bei Aufgaben, die spezifische Formatierungs-, Argumentations- oder Fachkenntnisse erfordern, kann die Eingabeaufforderung mit wenigen Handgriffen effektiv sein. [Siehe auch Zero-Shot Prompting](#).

FGAC

Siehe [detaillierte Zugriffskontrolle](#).

Feinkörnige Zugriffskontrolle (FGAC)

Die Verwendung mehrerer Bedingungen, um eine Zugriffsanfrage zuzulassen oder abzulehnen.

Flash-Cut-Migration

Eine Datenbankmigrationsmethode, bei der eine kontinuierliche Datenreplikation durch [Erfassung von Änderungsdaten](#) verwendet wird, um Daten in kürzester Zeit zu migrieren, anstatt einen schrittweisen Ansatz zu verwenden. Ziel ist es, Ausfallzeiten auf ein Minimum zu beschränken.

FM

Siehe [Fundamentmodell](#).

Fundamentmodell (FM)

Ein großes neuronales Deep-Learning-Netzwerk, das mit riesigen Datensätzen generalisierter und unbeschrifteter Daten trainiert wurde. FMs sind in der Lage, eine Vielzahl allgemeiner Aufgaben zu erfüllen, z. B. Sprache zu verstehen, Text und Bilder zu generieren und Konversationen in natürlicher Sprache zu führen. Weitere Informationen finden Sie unter [Was sind Foundation-Modelle](#).

G

generative KI

Eine Untergruppe von [KI-Modellen](#), die mit großen Datenmengen trainiert wurden und mit einer einfachen Textaufforderung neue Inhalte und Artefakte wie Bilder, Videos, Text und Audio erstellen können. Weitere Informationen finden Sie unter [Was ist Generative KI](#).

Geoblocking

Siehe [geografische Einschränkungen](#).

Geografische Einschränkungen (Geoblocking)

Bei Amazon eine Option CloudFront, um zu verhindern, dass Benutzer in bestimmten Ländern auf Inhaltsverteilungen zugreifen. Sie können eine Zulassungsliste oder eine Sperrliste verwenden, um zugelassene und gesperrte Länder anzugeben. Weitere Informationen finden Sie in [der Dokumentation unter Beschränkung der geografischen Verteilung Ihrer Inhalte](#). CloudFront

Gitflow-Workflow

Ein Ansatz, bei dem niedrigere und höhere Umgebungen unterschiedliche Zweige in einem Quellcode-Repository verwenden. Der Gitflow-Workflow gilt als veraltet, und der [Trunk-basierte Workflow](#) ist der moderne, bevorzugte Ansatz.

goldenes Bild

Ein Snapshot eines Systems oder einer Software, der als Vorlage für die Bereitstellung neuer Instanzen dieses Systems oder dieser Software verwendet wird. In der Fertigung kann ein Golden Image beispielsweise zur Bereitstellung von Software auf mehreren Geräten verwendet werden und trägt zur Verbesserung der Geschwindigkeit, Skalierbarkeit und Produktivität bei der Geräteherstellung bei.

Greenfield-Strategie

Das Fehlen vorhandener Infrastruktur in einer neuen Umgebung. Bei der Einführung einer Neuausrichtung einer Systemarchitektur können Sie alle neuen Technologien ohne Einschränkung der Kompatibilität mit der vorhandenen Infrastruktur auswählen, auch bekannt als [Brownfield](#). Wenn Sie die bestehende Infrastruktur erweitern, könnten Sie Brownfield- und Greenfield-Strategien mischen.

Integritätsschutz

Eine allgemeine Regel, die dazu beiträgt, Ressourcen, Richtlinien und die Einhaltung von Vorschriften in allen Unternehmenseinheiten zu regeln (OUs). Präventiver Integritätsschutz setzt Richtlinien durch, um die Einhaltung von Standards zu gewährleisten. Sie werden mithilfe von Service-Kontrollrichtlinien und IAM-Berechtigungsgrenzen implementiert. Detektivischer Integritätsschutz erkennt Richtlinienverstöße und Compliance-Probleme und generiert Warnmeldungen zur Abhilfe. Sie werden mithilfe von AWS Config, AWS Security Hub, Amazon GuardDuty AWS Trusted Advisor, Amazon Inspector und benutzerdefinierten AWS Lambda Prüfungen implementiert.

H

HEKTAR

Siehe [Hochverfügbarkeit](#).

Heterogene Datenbankmigration

Migrieren Sie Ihre Quelldatenbank in eine Zieldatenbank, die eine andere Datenbank-Engine verwendet (z. B. Oracle zu Amazon Aurora). Eine heterogene Migration ist in der Regel Teil einer Neuarchitektur, und die Konvertierung des Schemas kann eine komplexe Aufgabe sein. [AWS bietet AWS SCT](#), welches bei Schemakonvertierungen hilft.

hohe Verfügbarkeit (HA)

Die Fähigkeit eines Workloads, im Falle von Herausforderungen oder Katastrophen kontinuierlich und ohne Eingreifen zu arbeiten. HA-Systeme sind so konzipiert, dass sie automatisch ein Failover durchführen, gleichbleibend hohe Leistung bieten und unterschiedliche Lasten und Ausfälle mit minimalen Leistungseinbußen bewältigen.

historische Modernisierung

Ein Ansatz zur Modernisierung und Aufrüstung von Betriebstechnologiesystemen (OT), um den Bedürfnissen der Fertigungsindustrie besser gerecht zu werden. Ein Historian ist eine Art von Datenbank, die verwendet wird, um Daten aus verschiedenen Quellen in einer Fabrik zu sammeln und zu speichern.

Holdout-Daten

Ein Teil historischer, beschrifteter Daten, der aus einem Datensatz zurückgehalten wird, der zum Trainieren eines Modells für [maschinelles](#) Lernen verwendet wird. Sie können Holdout-Daten verwenden, um die Modellleistung zu bewerten, indem Sie die Modellvorhersagen mit den Holdout-Daten vergleichen.

Homogene Datenbankmigration

Migrieren Sie Ihre Quelldatenbank zu einer Zieldatenbank, die dieselbe Datenbank-Engine verwendet (z. B. Microsoft SQL Server zu Amazon RDS für SQL Server). Eine homogene Migration ist in der Regel Teil eines Hostwechsels oder eines Plattformwechsels. Sie können native Datenbankserviceprogramme verwenden, um das Schema zu migrieren.

heiße Daten

Daten, auf die häufig zugegriffen wird, z. B. Echtzeitdaten oder aktuelle Transaktionsdaten. Für diese Daten ist in der Regel eine leistungsstarke Speicherebene oder -klasse erforderlich, um schnelle Abfrageantworten zu ermöglichen.

Hotfix

Eine dringende Lösung für ein kritisches Problem in einer Produktionsumgebung. Aufgrund seiner Dringlichkeit wird ein Hotfix normalerweise außerhalb des typischen DevOps Release-Workflows erstellt.

Hypercare-Phase

Unmittelbar nach dem Cutover, der Zeitraum, in dem ein Migrationsteam die migrierten Anwendungen in der Cloud verwaltet und überwacht, um etwaige Probleme zu beheben. In der Regel dauert dieser Zeitraum 1–4 Tage. Am Ende der Hypercare-Phase überträgt das Migrationsteam in der Regel die Verantwortung für die Anwendungen an das Cloud-Betriebsteam.

|

IaC

Sehen Sie sich [Infrastruktur als Code](#) an.

Identitätsbasierte Richtlinie

Eine Richtlinie, die einem oder mehreren IAM-Prinzipalen zugeordnet ist und deren Berechtigungen innerhalb der AWS Cloud Umgebung definiert.

Leerlaufanwendung

Eine Anwendung mit einer durchschnittlichen CPU- und Arbeitsspeicherauslastung zwischen 5 und 20 Prozent über einen Zeitraum von 90 Tagen. In einem Migrationsprojekt ist es üblich, diese Anwendungen außer Betrieb zu nehmen oder sie On-Premises beizubehalten.

IIoT

Siehe [Industrielles Internet der Dinge](#).

unveränderliche Infrastruktur

Ein Modell, das eine neue Infrastruktur für Produktionsworkloads bereitstellt, anstatt die bestehende Infrastruktur zu aktualisieren, zu patchen oder zu modifizieren. [Unveränderliche Infrastrukturen sind von Natur aus konsistenter, zuverlässiger und vorhersehbarer als veränderliche Infrastrukturen](#). Weitere Informationen finden Sie in der Best Practice [Deploy using immutable infrastructure](#) im AWS Well-Architected Framework.

Eingehende (ingress) VPC

In einer Architektur AWS mit mehreren Konten ist dies eine VPC, die Netzwerkverbindungen von außerhalb einer Anwendung akzeptiert, überprüft und weiterleitet. Die [AWS Security Reference Architecture](#) empfiehlt, Ihr Netzwerkkonto mit eingehendem und ausgehendem Datenverkehr und Inspektion einzurichten, VPCs um die bidirektionale Schnittstelle zwischen Ihrer Anwendung und dem Internet im weiteren Sinne zu schützen.

Inkrementelle Migration

Eine Cutover-Strategie, bei der Sie Ihre Anwendung in kleinen Teilen migrieren, anstatt eine einziges vollständiges Cutover durchzuführen. Beispielsweise könnten Sie zunächst nur einige Microservices oder Benutzer auf das neue System umstellen. Nachdem Sie sich vergewissert haben, dass alles ordnungsgemäß funktioniert, können Sie weitere Microservices oder Benutzer

|

schrittweise verschieben, bis Sie Ihr Legacy-System außer Betrieb nehmen können. Diese Strategie reduziert die mit großen Migrationen verbundenen Risiken.

Industrie 4.0

Ein Begriff, der 2016 von [Klaus Schwab](#) eingeführt wurde und sich auf die Modernisierung von Fertigungsprozessen durch Fortschritte in den Bereichen Konnektivität, Echtzeitdaten, Automatisierung, Analytik und KI/ML bezieht.

Infrastruktur

Alle Ressourcen und Komponenten, die in der Umgebung einer Anwendung enthalten sind.

Infrastructure as Code (IaC)

Der Prozess der Bereitstellung und Verwaltung der Infrastruktur einer Anwendung mithilfe einer Reihe von Konfigurationsdateien. IaC soll Ihnen helfen, das Infrastrukturmanagement zu zentralisieren, Ressourcen zu standardisieren und schnell zu skalieren, sodass neue Umgebungen wiederholbar, zuverlässig und konsistent sind.

industrielles Internet der Dinge (T) Ilo

Einsatz von mit dem Internet verbundenen Sensoren und Geräten in Industriesektoren wie Fertigung, Energie, Automobilindustrie, Gesundheitswesen, Biowissenschaften und Landwirtschaft. Weitere Informationen finden Sie unter [Aufbau einer digitalen Transformationsstrategie für das industrielle Internet der Dinge \(IIoT\)](#).

Inspektions-VPC

In einer Architektur AWS mit mehreren Konten eine zentralisierte VPC, die Inspektionen des Netzwerkverkehrs zwischen VPCs (in demselben oder unterschiedlichen AWS-Regionen), dem Internet und lokalen Netzwerken verwaltet. In der [AWS Security Reference Architecture](#) wird empfohlen, Ihr Netzwerkkonto mit eingehendem und ausgehendem Datenverkehr sowie Inspektionen einzurichten, VPCs um die bidirektionale Schnittstelle zwischen Ihrer Anwendung und dem Internet im weiteren Sinne zu schützen.

Internet of Things (IoT)

Das Netzwerk verbundener physischer Objekte mit eingebetteten Sensoren oder Prozessoren, das über das Internet oder über ein lokales Kommunikationsnetzwerk mit anderen Geräten und Systemen kommuniziert. Weitere Informationen finden Sie unter [Was ist IoT?](#)

Interpretierbarkeit

Ein Merkmal eines Modells für Machine Learning, das beschreibt, inwieweit ein Mensch verstehen kann, wie die Vorhersagen des Modells von seinen Eingaben abhängen. Weitere Informationen finden Sie unter Interpretierbarkeit von [Modellen für maschinelles Lernen](#) mit AWS

IoT

Siehe [Internet der Dinge](#).

IT information library (ITIL, IT-Informationsbibliothek)

Eine Reihe von bewährten Methoden für die Bereitstellung von IT-Services und die Abstimmung dieser Services auf die Geschäftsanforderungen. ITIL bietet die Grundlage für ITSM.

T service management (ITSM, IT-Servicemanagement)

Aktivitäten im Zusammenhang mit der Gestaltung, Implementierung, Verwaltung und Unterstützung von IT-Services für eine Organisation. Informationen zur Integration von Cloud-Vorgängen mit ITSM-Tools finden Sie im [Leitfaden zur Betriebsintegration](#).

BIS

Weitere Informationen finden Sie in der [IT-Informationsbibliothek](#).

ITSM

Siehe [IT-Servicemanagement](#).

L

Labelbasierte Zugangskontrolle (LBAC)

Eine Implementierung der Mandatory Access Control (MAC), bei der den Benutzern und den Daten selbst jeweils explizit ein Sicherheitslabelwert zugewiesen wird. Die Schnittmenge zwischen der Benutzersicherheitsbeschriftung und der Datensicherheitsbeschriftung bestimmt, welche Zeilen und Spalten für den Benutzer sichtbar sind.

Landing Zone

Eine landing zone ist eine gut strukturierte AWS Umgebung mit mehreren Konten, die skalierbar und sicher ist. Dies ist ein Ausgangspunkt, von dem aus Ihre Organisationen Workloads und Anwendungen schnell und mit Vertrauen in ihre Sicherheits- und Infrastrukturmgebung starten

und bereitstellen können. Weitere Informationen zu Landing Zones finden Sie unter [Einrichtung einer sicheren und skalierbaren AWS -Umgebung mit mehreren Konten..](#)

großes Sprachmodell (LLM)

Ein [Deep-Learning-KI-Modell](#), das anhand einer riesigen Datenmenge vorab trainiert wurde. Ein LLM kann mehrere Aufgaben ausführen, z. B. Fragen beantworten, Dokumente zusammenfassen, Text in andere Sprachen übersetzen und Sätze vervollständigen. [Weitere Informationen finden Sie unter Was sind LLMs](#)

Große Migration

Eine Migration von 300 oder mehr Servern.

SCHWARZ

Weitere Informationen finden Sie unter [Label-basierte Zugriffskontrolle](#).

Geringste Berechtigung

Die bewährte Sicherheitsmethode, bei der nur die für die Durchführung einer Aufgabe erforderlichen Mindestberechtigungen erteilt werden. Weitere Informationen finden Sie unter [Geringste Berechtigungen anwenden](#) in der IAM-Dokumentation.

Lift and Shift

Siehe [7 Rs](#).

Little-Endian-System

Ein System, welches das niedrigwertigste Byte zuerst speichert. Siehe auch [Endianness](#).

LLM

Siehe [großes Sprachmodell](#).

Niedrigere Umgebungen

Siehe [Umgebung](#).

M

Machine Learning (ML)

Eine Art künstlicher Intelligenz, die Algorithmen und Techniken zur Mustererkennung und zum Lernen verwendet. ML analysiert aufgezeichnete Daten, wie z. B. Daten aus dem Internet der

Dinge (IoT), und lernt daraus, um ein statistisches Modell auf der Grundlage von Mustern zu erstellen. Weitere Informationen finden Sie unter [Machine Learning](#).

Hauptzweig

Siehe [Filiale](#).

Malware

Software, die entwickelt wurde, um die Computersicherheit oder den Datenschutz zu gefährden. Malware kann Computersysteme stören, vertrauliche Informationen durchsickern lassen oder sich unbefugten Zugriff verschaffen. Beispiele für Malware sind Viren, Würmer, Ransomware, Trojaner, Spyware und Keylogger.

verwaltete Dienste

AWS-Services für die die Infrastrukturebene, das Betriebssystem und die Plattformen AWS betrieben werden, und Sie greifen auf die Endgeräte zu, um Daten zu speichern und abzurufen. Amazon Simple Storage Service (Amazon S3) und Amazon DynamoDB sind Beispiele für Managed Services. Diese werden auch als abstrakte Dienste bezeichnet.

Manufacturing Execution System (MES)

Ein Softwaresystem zur Nachverfolgung, Überwachung, Dokumentation und Steuerung von Produktionsprozessen, bei denen Rohstoffe in der Fertigung zu fertigen Produkten umgewandelt werden.

MAP

Siehe [Migration Acceleration Program](#).

Mechanismus

Ein vollständiger Prozess, bei dem Sie ein Tool erstellen, die Akzeptanz des Tools vorantreiben und anschließend die Ergebnisse überprüfen, um Anpassungen vorzunehmen. Ein Mechanismus ist ein Zyklus, der sich im Laufe seiner Tätigkeit selbst verstärkt und verbessert. Weitere Informationen finden Sie unter [Aufbau von Mechanismen](#) im AWS Well-Architected Framework.

Mitgliedskonto

Alle AWS-Konten außer dem Verwaltungskonto, die Teil einer Organisation in sind. AWS Organizations Ein Konto kann jeweils nur einer Organisation angehören.

DURCHEINANDER

Siehe [Manufacturing Execution System](#).

Message Queuing-Telemetrietransport (MQTT)

[Ein leichtes machine-to-machine \(M2M\) -Kommunikationsprotokoll, das auf dem Publish/Subscribe-Muster für IoT-Geräte mit beschränkten Ressourcen basiert.](#)

Microservice

Ein kleiner, unabhängiger Dienst, der über genau definierte Kanäle kommuniziert APIs und in der Regel kleinen, eigenständigen Teams gehört. Ein Versicherungssystem kann beispielsweise Microservices beinhalten, die Geschäftsfunktionen wie Vertrieb oder Marketing oder Subdomains wie Einkauf, Schadenersatz oder Analytik zugeordnet sind. Zu den Vorteilen von Microservices gehören Agilität, flexible Skalierung, einfache Bereitstellung, wiederverwendbarer Code und Ausfallsicherheit. Weitere Informationen finden Sie unter [Integration von Microservices mithilfe serverloser Dienste](#). AWS

Microservices-Architekturen

Ein Ansatz zur Erstellung einer Anwendung mit unabhängigen Komponenten, die jeden Anwendungsprozess als Microservice ausführen. Diese Microservices kommunizieren mithilfe von Lightweight über eine klar definierte Schnittstelle. APIs Jeder Microservice in dieser Architektur kann aktualisiert, bereitgestellt und skaliert werden, um den Bedarf an bestimmten Funktionen einer Anwendung zu decken. Weitere Informationen finden Sie unter [Implementierung von Microservices](#) auf. AWS

Migration Acceleration Program (MAP)

Ein AWS Programm, das Beratung, Unterstützung, Schulungen und Services bietet, um Unternehmen dabei zu unterstützen, eine solide betriebliche Grundlage für die Umstellung auf die Cloud zu schaffen und die anfänglichen Kosten von Migrationen auszugleichen. MAP umfasst eine Migrationsmethode für die methodische Durchführung von Legacy-Migrationen sowie eine Reihe von Tools zur Automatisierung und Beschleunigung gängiger Migrationsszenarien.

Migration in großem Maßstab

Der Prozess, bei dem der Großteil des Anwendungsportfolios in Wellen in die Cloud verlagert wird, wobei in jeder Welle mehr Anwendungen schneller migriert werden. In dieser Phase werden die bewährten Verfahren und Erkenntnisse aus den früheren Phasen zur Implementierung einer Migrationsfabrik von Teams, Tools und Prozessen zur Optimierung der Migration von Workloads durch Automatisierung und agile Bereitstellung verwendet. Dies ist die dritte Phase der [AWS - Migrationsstrategie](#).

Migrationsfabrik

Funktionsübergreifende Teams, die die Migration von Workloads durch automatisierte, agile Ansätze optimieren. Zu den Teams in der Migrationsabteilung gehören in der Regel Betriebsabläufe, Geschäftsanalysten und Eigentümer, Migrationsingenieure, Entwickler und DevOps Experten, die in Sprints arbeiten. Zwischen 20 und 50 Prozent eines Unternehmensanwendungsportfolios bestehen aus sich wiederholenden Mustern, die durch einen Fabrik-Ansatz optimiert werden können. Weitere Informationen finden Sie in [Diskussion über Migrationsfabriken](#) und den [Leitfaden zur Cloud-Migration-Fabrik](#) in diesem Inhaltssatz.

Migrationsmetadaten

Die Informationen über die Anwendung und den Server, die für den Abschluss der Migration benötigt werden. Für jedes Migrationsmuster ist ein anderer Satz von Migrationsmetadaten erforderlich. Beispiele für Migrationsmetadaten sind das Zielsubnetz, die Sicherheitsgruppe und AWS das Konto.

Migrationsmuster

Eine wiederholbare Migrationsaufgabe, in der die Migrationsstrategie, das Migrationsziel und die verwendete Migrationsanwendung oder der verwendete Migrationservice detailliert beschrieben werden. Beispiel: Rehost-Migration zu Amazon EC2 mit AWS Application Migration Service.

Migration Portfolio Assessment (MPA)

Ein Online-Tool, das Informationen zur Validierung des Geschäftsszenarios für die Migration auf das bereitstellt. AWS Cloud MPA bietet eine detaillierte Portfoliobewertung (richtige Servergröße, Preisgestaltung, Gesamtbetriebskostenanalyse, Migrationskostenanalyse) sowie Migrationsplanung (Anwendungsdatenanalyse und Datenerfassung, Anwendungsgruppierung, Migrationspriorisierung und Wellenplanung). Das [MPA-Tool](#) (Anmeldung erforderlich) steht allen AWS Beratern und APN-Partnerberatern kostenlos zur Verfügung.

Migration Readiness Assessment (MRA)

Der Prozess, bei dem mithilfe des AWS CAF Erkenntnisse über den Cloud-Bereitschaftsstatus eines Unternehmens gewonnen, Stärken und Schwächen identifiziert und ein Aktionsplan zur Schließung festgestellter Lücken erstellt wird. Weitere Informationen finden Sie im [Benutzerhandbuch für Migration Readiness](#). MRA ist die erste Phase der [AWS - Migrationsstrategie](#).

Migrationsstrategie

Der Ansatz, der verwendet wurde, um einen Workload auf den AWS Cloud zu migrieren. Weitere Informationen finden Sie im Eintrag [7 Rs](#) in diesem Glossar und unter [Mobilisieren Sie Ihr Unternehmen, um umfangreiche Migrationen zu beschleunigen](#).

ML

Siehe [maschinelles Lernen](#).

Modernisierung

Umwandlung einer veralteten (veralteten oder monolithischen) Anwendung und ihrer Infrastruktur in ein agiles, elastisches und hochverfügbares System in der Cloud, um Kosten zu senken, die Effizienz zu steigern und Innovationen zu nutzen. Weitere Informationen finden Sie unter [Strategie zur Modernisierung von Anwendungen in der AWS Cloud](#).

Bewertung der Modernisierungsfähigkeit

Eine Bewertung, anhand derer festgestellt werden kann, ob die Anwendungen einer Organisation für die Modernisierung bereit sind, Vorteile, Risiken und Abhängigkeiten identifiziert und ermittelt wird, wie gut die Organisation den zukünftigen Status dieser Anwendungen unterstützen kann. Das Ergebnis der Bewertung ist eine Vorlage der Zielarchitektur, eine Roadmap, in der die Entwicklungsphasen und Meilensteine des Modernisierungsprozesses detailliert beschrieben werden, sowie ein Aktionsplan zur Behebung festgestellter Lücken. Weitere Informationen finden Sie unter [Evaluierung der Modernisierungsbereitschaft von Anwendungen in der AWS Cloud](#).

Monolithische Anwendungen (Monolithen)

Anwendungen, die als ein einziger Service mit eng gekoppelten Prozessen ausgeführt werden. Monolithische Anwendungen haben verschiedene Nachteile. Wenn ein Anwendungs-Feature stark nachgefragt wird, muss die gesamte Architektur skaliert werden. Das Hinzufügen oder Verbessern der Feature einer monolithischen Anwendung wird ebenfalls komplexer, wenn die Codebasis wächst. Um diese Probleme zu beheben, können Sie eine Microservices-Architektur verwenden. Weitere Informationen finden Sie unter [Zerlegen von Monolithen in Microservices](#).

MPA

Siehe [Bewertung des Migrationsportfolios](#).

MQTT

Siehe [Message Queuing-Telemetrietransport](#).

Mehrklassen-Klassifizierung

Ein Prozess, der dabei hilft, Vorhersagen für mehrere Klassen zu generieren (wobei eines von mehr als zwei Ergebnissen vorhergesagt wird). Ein ML-Modell könnte beispielsweise fragen: „Ist dieses Produkt ein Buch, ein Auto oder ein Telefon?“ oder „Welche Kategorie von Produkten ist für diesen Kunden am interessantesten?“

veränderbare Infrastruktur

Ein Modell, das die bestehende Infrastruktur für Produktionsworkloads aktualisiert und modifiziert. Für eine verbesserte Konsistenz, Zuverlässigkeit und Vorhersagbarkeit empfiehlt das AWS Well-Architected Framework die Verwendung einer [unveränderlichen Infrastruktur](#) als bewährte Methode.

O

OAC

[Siehe Origin Access Control.](#)

OAI

Siehe [Zugriffsidentität von Origin.](#)

COM

Siehe [organisatorisches Change-Management.](#)

Offline-Migration

Eine Migrationsmethode, bei der der Quell-Workload während des Migrationsprozesses heruntergefahren wird. Diese Methode ist mit längeren Ausfallzeiten verbunden und wird in der Regel für kleine, unkritische Workloads verwendet.

OI

Siehe [Betriebsintegration.](#)

OLA

Siehe Vereinbarung auf [operativer Ebene.](#)

Online-Migration

Eine Migrationsmethode, bei der der Quell-Workload auf das Zielsystem kopiert wird, ohne offline genommen zu werden. Anwendungen, die mit dem Workload verbunden sind, können während

der Migration weiterhin funktionieren. Diese Methode beinhaltet keine bis minimale Ausfallzeit und wird in der Regel für kritische Produktionsworkloads verwendet.

OPC-UA

Siehe [Open Process Communications — Unified Architecture](#).

Offene Prozesskommunikation — Einheitliche Architektur (OPC-UA)

Ein machine-to-machine (M2M) -Kommunikationsprotokoll für die industrielle Automatisierung. OPC-UA bietet einen Interoperabilitätsstandard mit Datenverschlüsselungs-, Authentifizierungs- und Autorisierungsschemata.

Vereinbarung auf Betriebsebene (OLA)

Eine Vereinbarung, in der klargestellt wird, welche funktionalen IT-Gruppen sich gegenseitig versprechen zu liefern, um ein Service Level Agreement (SLA) zu unterstützen.

Überprüfung der Betriebsbereitschaft (ORR)

Eine Checkliste mit Fragen und zugehörigen bewährten Methoden, die Ihnen helfen, Vorfälle und mögliche Ausfälle zu verstehen, zu bewerten, zu verhindern oder deren Umfang zu reduzieren. Weitere Informationen finden Sie unter [Operational Readiness Reviews \(ORR\)](#) im AWS Well-Architected Framework.

Betriebstechnologie (OT)

Hardware- und Softwaresysteme, die mit der physischen Umgebung zusammenarbeiten, um industrielle Abläufe, Ausrüstung und Infrastruktur zu steuern. In der Fertigung ist die Integration von OT- und Informationstechnologie (IT) -Systemen ein zentraler Schwerpunkt der [Industrie 4.0-Transformationen](#).

Betriebsintegration (OI)

Der Prozess der Modernisierung von Abläufen in der Cloud, der Bereitschaftsplanung, Automatisierung und Integration umfasst. Weitere Informationen finden Sie im [Leitfaden zur Betriebsintegration](#).

Organisationspfad

Ein Pfad, der von erstellt wird und in AWS CloudTrail dem alle Ereignisse für alle AWS-Konten in einer Organisation protokolliert werden. AWS Organizations Diese Spur wird in jedem AWS-Konto , der Teil der Organisation ist, erstellt und verfolgt die Aktivität in jedem Konto. Weitere Informationen finden Sie in der CloudTrail Dokumentation unter [Einen Trail für eine Organisation erstellen](#).

Organisatorisches Veränderungsmanagement (OCM)

Ein Framework für das Management wichtiger, disruptiver Geschäftstransformationen aus Sicht der Mitarbeiter, der Kultur und der Führung. OCM hilft Organisationen dabei, sich auf neue Systeme und Strategien vorzubereiten und auf diese umzustellen, indem es die Akzeptanz von Veränderungen beschleunigt, Übergangsprobleme angeht und kulturelle und organisatorische Veränderungen vorantreibt. In der AWS Migrationsstrategie wird dieses Framework aufgrund der Geschwindigkeit des Wandels, der bei Projekten zur Cloud-Einführung erforderlich ist, als Mitarbeiterbeschleunigung bezeichnet. Weitere Informationen finden Sie im [OCM-Handbuch](#).

Ursprungszugriffskontrolle (OAC)

In CloudFront, eine erweiterte Option zur Zugriffsbeschränkung, um Ihre Amazon Simple Storage Service (Amazon S3) -Inhalte zu sichern. OAC unterstützt alle S3-Buckets insgesamt AWS-Regionen, serverseitige Verschlüsselung mit AWS KMS (SSE-KMS) sowie dynamische PUT und DELETE Anfragen an den S3-Bucket.

Ursprungszugriffsidentität (OAI)

In CloudFront, eine Option zur Zugriffsbeschränkung, um Ihre Amazon S3 S3-Inhalte zu sichern. Wenn Sie OAI verwenden, CloudFront erstellt es einen Principal, mit dem sich Amazon S3 authentifizieren kann. Authentifizierte Principals können nur über eine bestimmte Distribution auf Inhalte in einem S3-Bucket zugreifen. CloudFront Siehe auch [OAC](#), das eine detailliertere und verbesserte Zugriffskontrolle bietet.

ORR

Weitere Informationen finden Sie unter [Überprüfung der Betriebsbereitschaft](#).

NICHT

Siehe [Betriebstechnologie](#).

Ausgehende (egress) VPC

In einer Architektur AWS mit mehreren Konten eine VPC, die Netzwerkverbindungen verarbeitet, die von einer Anwendung aus initiiert werden. Die [AWS Security Reference Architecture](#) empfiehlt die Einrichtung Ihres Netzwerkkontos mit eingehendem und ausgehendem Datenverkehr sowie Inspektion, VPCs um die bidirektionale Schnittstelle zwischen Ihrer Anwendung und dem Internet im weiteren Sinne zu schützen.

P

Berechtigungsgrenze

Eine IAM-Verwaltungsrichtlinie, die den IAM-Prinzipalen zugeordnet ist, um die maximalen Berechtigungen festzulegen, die der Benutzer oder die Rolle haben kann. Weitere Informationen finden Sie unter [Berechtigungsgrenzen](#) für IAM-Entitäts in der IAM-Dokumentation.

persönlich identifizierbare Informationen (PII)

Informationen, die, wenn sie direkt betrachtet oder mit anderen verwandten Daten kombiniert werden, verwendet werden können, um vernünftige Rückschlüsse auf die Identität einer Person zu ziehen. Beispiele für personenbezogene Daten sind Namen, Adressen und Kontaktinformationen.

Personenbezogene Daten

Siehe [persönlich identifizierbare Informationen](#).

Playbook

Eine Reihe vordefinierter Schritte, die die mit Migrationen verbundenen Aufgaben erfassen, z. B. die Bereitstellung zentraler Betriebsfunktionen in der Cloud. Ein Playbook kann die Form von Skripten, automatisierten Runbooks oder einer Zusammenfassung der Prozesse oder Schritte annehmen, die für den Betrieb Ihrer modernisierten Umgebung erforderlich sind.

PLC

Siehe [programmierbare Logiksteuerung](#).

PLM

Siehe [Produktlebenszyklusmanagement](#).

policy

Ein Objekt, das Berechtigungen definieren (siehe [identitätsbasierte Richtlinie](#)), Zugriffsbedingungen spezifizieren (siehe [ressourcenbasierte Richtlinie](#)) oder die maximalen Berechtigungen für alle Konten in einer Organisation definieren kann AWS Organizations (siehe [Dienststeuerungsrichtlinie](#)).

Polyglotte Beharrlichkeit

Unabhängige Auswahl der Datenspeichertechnologie eines Microservices auf der Grundlage von Datenzugriffsmustern und anderen Anforderungen. Wenn Ihre Microservices über dieselbe

Datenspeichertechnologie verfügen, kann dies zu Implementierungsproblemen oder zu Leistungseinbußen führen. Microservices lassen sich leichter implementieren und erzielen eine bessere Leistung und Skalierbarkeit, wenn sie den Datenspeicher verwenden, der ihren Anforderungen am besten entspricht. Weitere Informationen finden Sie unter [Datenpersistenz in Microservices aktivieren](#).

Portfoliobewertung

Ein Prozess, bei dem das Anwendungsportfolio ermittelt, analysiert und priorisiert wird, um die Migration zu planen. Weitere Informationen finden Sie in [Bewerten der Migrationsbereitschaft](#).

predicate

Eine Abfragebedingung, die `true` oder zurückgibt `false`, was üblicherweise in einer Klausel vorkommt. WHERE

Prädikat Pushdown

Eine Technik zur Optimierung von Datenbankabfragen, bei der die Daten in der Abfrage vor der Übertragung gefiltert werden. Dadurch wird die Datenmenge reduziert, die aus der relationalen Datenbank abgerufen und verarbeitet werden muss, und die Abfrageleistung wird verbessert.

Präventive Kontrolle

Eine Sicherheitskontrolle, die verhindern soll, dass ein Ereignis eintritt. Diese Kontrollen stellen eine erste Verteidigungslinie dar, um unbefugten Zugriff oder unerwünschte Änderungen an Ihrem Netzwerk zu verhindern. Weitere Informationen finden Sie unter [Präventive Kontrolle](#) in Implementierung von Sicherheitskontrollen in AWS.

Prinzipal

Eine Entität AWS, die Aktionen ausführen und auf Ressourcen zugreifen kann. Bei dieser Entität handelt es sich in der Regel um einen Root-Benutzer für eine AWS-Konto, eine IAM-Rolle oder einen Benutzer. Weitere Informationen finden Sie unter Prinzipal in [Rollenbegriffe und -konzepte](#) in der IAM-Dokumentation.

Datenschutz von Natur aus

Ein systemtechnischer Ansatz, der den Datenschutz während des gesamten Entwicklungsprozesses berücksichtigt.

Privat gehostete Zonen

Ein Container, der Informationen darüber enthält, wie Amazon Route 53 auf DNS-Abfragen für eine Domain und deren Subdomains innerhalb einer oder mehrerer VPCs Domains antworten

soll. Weitere Informationen finden Sie unter [Arbeiten mit privat gehosteten Zonen](#) in der Route-53-Dokumentation.

proaktive Steuerung

Eine [Sicherheitskontrolle](#), die den Einsatz nicht richtlinienkonformer Ressourcen verhindern soll. Diese Steuerelemente scannen Ressourcen, bevor sie bereitgestellt werden. Wenn die Ressource nicht mit der Steuerung konform ist, wird sie nicht bereitgestellt. Weitere Informationen finden Sie im [Referenzhandbuch zu Kontrollen](#) in der AWS Control Tower Dokumentation und unter [Proaktive Kontrollen](#) unter Implementierung von Sicherheitskontrollen am AWS.

Produktlebenszyklusmanagement (PLM)

Das Management von Daten und Prozessen für ein Produkt während seines gesamten Lebenszyklus, vom Design, der Entwicklung und Markteinführung über Wachstum und Reife bis hin zur Markteinführung und Markteinführung.

Produktionsumgebung

Siehe [Umgebung](#).

Speicherprogrammierbare Steuerung (SPS)

In der Fertigung ein äußerst zuverlässiger, anpassungsfähiger Computer, der Maschinen überwacht und Fertigungsprozesse automatisiert.

schnelle Verkettung

Verwendung der Ausgabe einer [LLM-Eingabeaufforderung](#) als Eingabe für die nächste Aufforderung, um bessere Antworten zu generieren. Diese Technik wird verwendet, um eine komplexe Aufgabe in Unteraufgaben zu unterteilen oder um eine vorläufige Antwort iterativ zu verfeinern oder zu erweitern. Sie trägt dazu bei, die Genauigkeit und Relevanz der Antworten eines Modells zu verbessern und ermöglicht detailliertere, personalisierte Ergebnisse.

Pseudonymisierung

Der Prozess, bei dem persönliche Identifikatoren in einem Datensatz durch Platzhalterwerte ersetzt werden. Pseudonymisierung kann zum Schutz der Privatsphäre beitragen. Pseudonymisierte Daten gelten weiterhin als personenbezogene Daten.

publish/subscribe (pub/sub)

Ein Muster, das asynchrone Kommunikation zwischen Microservices ermöglicht, um die Skalierbarkeit und Reaktionsfähigkeit zu verbessern. In einem auf Microservices basierenden [MES](#) kann ein Microservice beispielsweise Ereignismeldungen in einem Kanal veröffentlichen,

den andere Microservices abonnieren können. Das System kann neue Microservices hinzufügen, ohne den Veröffentlichungsservice zu ändern.

Q

Abfrageplan

Eine Reihe von Schritten, wie Anweisungen, die für den Zugriff auf die Daten in einem relationalen SQL-Datenbanksystem verwendet werden.

Abfrageplanregression

Wenn ein Datenbankserviceoptimierer einen weniger optimalen Plan wählt als vor einer bestimmten Änderung der Datenbankumgebung. Dies kann durch Änderungen an Statistiken, Beschränkungen, Umgebungseinstellungen, Abfrageparameter-Bindungen und Aktualisierungen der Datenbank-Engine verursacht werden.

R

RACI-Matrix

Siehe [verantwortlich, rechenschaftspflichtig, konsultiert, informiert \(RACI\)](#).

LAPPEN

Siehe [Erweiterte Generierung beim Abrufen](#).

Ransomware

Eine bösartige Software, die entwickelt wurde, um den Zugriff auf ein Computersystem oder Daten zu blockieren, bis eine Zahlung erfolgt ist.

RASCI-Matrix

Siehe [verantwortlich, rechenschaftspflichtig, konsultiert, informiert \(RACI\)](#).

RCAC

Siehe [Zugriffskontrolle für Zeilen und Spalten](#).

Read Replica

Eine Kopie einer Datenbank, die nur für Lesezwecke verwendet wird. Sie können Abfragen an das Lesereplikat weiterleiten, um die Belastung auf Ihrer Primärdatenbank zu reduzieren.

neu strukturieren

Siehe [7 Rs.](#)

Recovery Point Objective (RPO)

Die maximal zulässige Zeitspanne seit dem letzten Datenwiederherstellungspunkt. Damit wird festgelegt, was als akzeptabler Datenverlust zwischen dem letzten Wiederherstellungspunkt und der Serviceunterbrechung gilt.

Wiederherstellungszeitziel (RTO)

Die maximal zulässige Verzögerung zwischen der Betriebsunterbrechung und der Wiederherstellung des Dienstes.

Refaktorisierung

Siehe [7 Rs.](#)

Region

Eine Sammlung von AWS Ressourcen in einem geografischen Gebiet. Jeder AWS-Region ist isoliert und unabhängig von den anderen, um Fehlertoleranz, Stabilität und Belastbarkeit zu gewährleisten. Weitere Informationen finden [Sie unter Geben Sie an, was AWS-Regionen Ihr Konto verwenden kann.](#)

Regression

Eine ML-Technik, die einen numerischen Wert vorhersagt. Zum Beispiel, um das Problem „Zu welchem Preis wird dieses Haus verkauft werden?“ zu lösen Ein ML-Modell könnte ein lineares Regressionsmodell verwenden, um den Verkaufspreis eines Hauses auf der Grundlage bekannter Fakten über das Haus (z. B. die Quadratmeterzahl) vorherzusagen.

rehosten

Siehe [7 Rs.](#)

Veröffentlichung

In einem Bereitstellungsprozess der Akt der Förderung von Änderungen an einer Produktionsumgebung.

umziehen

Siehe [7 Rs.](#)

neue Plattform

Siehe [7 Rs.](#)

Rückkauf

Siehe [7 Rs.](#)

Ausfallsicherheit

Die Fähigkeit einer Anwendung, Störungen zu widerstehen oder sich von ihnen zu erholen. [Hochverfügbarkeit](#) und [Notfallwiederherstellung](#) sind häufig Überlegungen bei der Planung der Ausfallsicherheit in der. AWS Cloud Weitere Informationen finden Sie unter [AWS Cloud Resilienz](#).

Ressourcenbasierte Richtlinie

Eine mit einer Ressource verknüpfte Richtlinie, z. B. ein Amazon-S3-Bucket, ein Endpunkt oder ein Verschlüsselungsschlüssel. Diese Art von Richtlinie legt fest, welchen Prinzipalen der Zugriff gewährt wird, welche Aktionen unterstützt werden und welche anderen Bedingungen erfüllt sein müssen.

RACI-Matrix (verantwortlich, rechenschaftspflichtig, konsultiert, informiert)

Eine Matrix, die die Rollen und Verantwortlichkeiten aller an Migrationsaktivitäten und Cloud-Operationen beteiligten Parteien definiert. Der Matrixname leitet sich von den in der Matrix definierten Zuständigkeitstypen ab: verantwortlich (R), rechenschaftspflichtig (A), konsultiert (C) und informiert (I). Der Unterstützungstyp (S) ist optional. Wenn Sie Unterstützung einbeziehen, wird die Matrix als RASCI-Matrix bezeichnet, und wenn Sie sie ausschließen, wird sie als RACI-Matrix bezeichnet.

Reaktive Kontrolle

Eine Sicherheitskontrolle, die darauf ausgelegt ist, die Behebung unerwünschter Ereignisse oder Abweichungen von Ihren Sicherheitsstandards voranzutreiben. Weitere Informationen finden Sie unter [Reaktive Kontrolle](#) in Implementieren von Sicherheitskontrollen in AWS.

Beibehaltung

Siehe [7 Rs.](#)

zurückziehen

Siehe [7 Rs.](#)

Retrieval Augmented Generation (RAG)

Eine [generative KI-Technologie](#), bei der ein [LLM](#) auf eine maßgebliche Datenquelle verweist, die sich außerhalb seiner Trainingsdatenquellen befindet, bevor eine Antwort generiert wird. Ein RAG-Modell könnte beispielsweise eine semantische Suche in der Wissensdatenbank oder in benutzerdefinierten Daten einer Organisation durchführen. Weitere Informationen finden Sie unter [Was ist RAG](#).

Drehung

Der Vorgang, bei dem ein [Geheimnis](#) regelmäßig aktualisiert wird, um es einem Angreifer zu erschweren, auf die Anmeldeinformationen zuzugreifen.

Zugriffskontrolle für Zeilen und Spalten (RCAC)

Die Verwendung einfacher, flexibler SQL-Ausdrücke mit definierten Zugriffsregeln. RCAC besteht aus Zeilenberechtigungen und Spaltenmasken.

RPO

Siehe [Recovery Point Objective](#).

RTO

Siehe [Ziel der Wiederherstellungszeit](#).

Runbook

Eine Reihe manueller oder automatisierter Verfahren, die zur Ausführung einer bestimmten Aufgabe erforderlich sind. Diese sind in der Regel darauf ausgelegt, sich wiederholende Operationen oder Verfahren mit hohen Fehlerquoten zu rationalisieren.

S

SAML 2.0

Ein offener Standard, den viele Identitätsanbieter (IdPs) verwenden. Diese Funktion ermöglicht föderiertes Single Sign-On (SSO), sodass sich Benutzer bei den API-Vorgängen anmelden AWS Management Console oder die AWS API-Operationen aufrufen können, ohne dass Sie einen Benutzer in IAM für alle in Ihrer Organisation erstellen müssen. Weitere Informationen zum SAML-2.0.-basierten Verbund finden Sie unter [Über den SAML-2.0-basierten Verbund](#) in der IAM-Dokumentation.

SCADA

Siehe [Aufsichtskontrolle und Datenerfassung](#).

SCP

Siehe [Richtlinie zur Dienstkontrolle](#).

Secret

Interne AWS Secrets Manager, vertrauliche oder eingeschränkte Informationen, wie z. B. ein Passwort oder Benutzeranmeldedaten, die Sie in verschlüsselter Form speichern. Es besteht aus dem geheimen Wert und seinen Metadaten. Der geheime Wert kann binär, eine einzelne Zeichenfolge oder mehrere Zeichenketten sein. Weitere Informationen finden Sie unter [Was ist in einem Secrets Manager Manager-Geheimnis?](#) in der Secrets Manager Manager-Dokumentation.

Sicherheit durch Design

Ein systemtechnischer Ansatz, der die Sicherheit während des gesamten Entwicklungsprozesses berücksichtigt.

Sicherheitskontrolle

Ein technischer oder administrativer Integritätsschutz, der die Fähigkeit eines Bedrohungsakteurs, eine Schwachstelle auszunutzen, verhindert, erkennt oder einschränkt. Es gibt vier Haupttypen von Sicherheitskontrollen: [präventiv](#), [detektiv](#), [reaktionsschnell](#) und [proaktiv](#).

Härtung der Sicherheit

Der Prozess, bei dem die Angriffsfläche reduziert wird, um sie widerstandsfähiger gegen Angriffe zu machen. Dies kann Aktionen wie das Entfernen von Ressourcen, die nicht mehr benötigt werden, die Implementierung der bewährten Sicherheitsmethode der Gewährung geringster Berechtigungen oder die Deaktivierung unnötiger Feature in Konfigurationsdateien umfassen.

System zur Verwaltung von Sicherheitsinformationen und Ereignissen (security information and event management – SIEM)

Tools und Services, die Systeme für das Sicherheitsinformationsmanagement (SIM) und das Management von Sicherheitsereignissen (SEM) kombinieren. Ein SIEM-System sammelt, überwacht und analysiert Daten von Servern, Netzwerken, Geräten und anderen Quellen, um Bedrohungen und Sicherheitsverletzungen zu erkennen und Warnmeldungen zu generieren.

Automatisierung von Sicherheitsreaktionen

Eine vordefinierte und programmierte Aktion, die darauf ausgelegt ist, automatisch auf ein Sicherheitsereignis zu reagieren oder es zu beheben. Diese Automatisierungen dienen als

[detektive](#) oder [reaktionsschnelle](#) Sicherheitskontrollen, die Sie bei der Implementierung bewährter AWS Sicherheitsmethoden unterstützen. Beispiele für automatisierte Antwortaktionen sind das Ändern einer VPC-Sicherheitsgruppe, das Patchen einer EC2 Amazon-Instance oder das Rotieren von Anmeldeinformationen.

Serverseitige Verschlüsselung

Verschlüsselung von Daten am Zielort durch denjenigen AWS-Service, der sie empfängt.

Service-Kontrollrichtlinie (SCP)

Eine Richtlinie, die eine zentrale Steuerung der Berechtigungen für alle Konten in einer Organisation in ermöglicht AWS Organizations. SCPs Definieren Sie Leitplanken oder legen Sie Grenzwerte für Aktionen fest, die ein Administrator an Benutzer oder Rollen delegieren kann. Sie können sie SCPs als Zulassungs- oder Ablehnungslisten verwenden, um festzulegen, welche Dienste oder Aktionen zulässig oder verboten sind. Weitere Informationen finden Sie in der AWS Organizations Dokumentation unter [Richtlinien zur Dienststeuerung](#).

Service-Endpunkt

Die URL des Einstiegspunkts für einen AWS-Service. Sie können den Endpunkt verwenden, um programmgesteuert eine Verbindung zum Zielservice herzustellen. Weitere Informationen finden Sie unter [AWS-Service -Endpunkte](#) in der Allgemeine AWS-Referenz.

Service Level Agreement (SLA)

Eine Vereinbarung, in der klargestellt wird, was ein IT-Team seinen Kunden zu bieten verspricht, z. B. in Bezug auf Verfügbarkeit und Leistung der Services.

Service-Level-Indikator (SLI)

Eine Messung eines Leistungsaspekts eines Dienstes, z. B. seiner Fehlerrate, Verfügbarkeit oder Durchsatz.

Service-Level-Ziel (SLO)

Eine Zielkennzahl, die den Zustand eines Dienstes darstellt, gemessen anhand eines [Service-Level-Indikators](#).

Modell der geteilten Verantwortung

Ein Modell, das die Verantwortung beschreibt, mit der Sie gemeinsam AWS für Cloud-Sicherheit und Compliance verantwortlich sind. AWS ist für die Sicherheit der Cloud verantwortlich, während Sie für die Sicherheit in der Cloud verantwortlich sind. Weitere Informationen finden Sie unter [Modell der geteilten Verantwortung](#).

SIEM

Siehe [Sicherheitsinformations- und Event-Management-System](#).

Single Point of Failure (SPOF)

Ein Fehler in einer einzelnen, kritischen Komponente einer Anwendung, der das System stören kann.

SLA

Siehe [Service Level Agreement](#).

SLI

Siehe [Service-Level-Indikator](#).

ALSO

Siehe [Service-Level-Ziel](#).

split-and-seed Modell

Ein Muster für die Skalierung und Beschleunigung von Modernisierungsprojekten. Sobald neue Features und Produktversionen definiert werden, teilt sich das Kernteam auf, um neue Produktteams zu bilden. Dies trägt zur Skalierung der Fähigkeiten und Services Ihrer Organisation bei, verbessert die Produktivität der Entwickler und unterstützt schnelle Innovationen. Weitere Informationen finden Sie unter [Schrittweiser Ansatz zur Modernisierung von Anwendungen in der AWS Cloud](#)

SPOTTEN

Siehe [Single Point of Failure](#).

Sternschema

Eine Datenbank-Organisationsstruktur, die eine große Faktentabelle zum Speichern von Transaktions- oder Messdaten und eine oder mehrere kleinere dimensionale Tabellen zum Speichern von Datenattributen verwendet. Diese Struktur ist für die Verwendung in einem [Data Warehouse](#) oder für Business Intelligence-Zwecke konzipiert.

Strangler-Fig-Muster

Ein Ansatz zur Modernisierung monolithischer Systeme, bei dem die Systemfunktionen schrittweise umgeschrieben und ersetzt werden, bis das Legacy-System außer Betrieb

genommen werden kann. Dieses Muster verwendet die Analogie einer Feigenrebe, die zu einem etablierten Baum heranwächst und schließlich ihren Wirt überwindet und ersetzt. Das Muster wurde [eingeführt von Martin Fowler](#) als Möglichkeit, Risiken beim Umschreiben monolithischer Systeme zu managen. Ein Beispiel für die Anwendung dieses Musters finden Sie unter [Schrittweises Modernisieren älterer Microsoft ASP.NET \(ASMX\)-Webservices mithilfe von Containern und Amazon API Gateway](#).

Subnetz

Ein Bereich von IP-Adressen in Ihrer VPC. Ein Subnetz muss sich in einer einzigen Availability Zone befinden.

Aufsichtskontrolle und Datenerfassung (SCADA)

In der Fertigung ein System, das Hardware und Software zur Überwachung von Sachanlagen und Produktionsabläufen verwendet.

Symmetrische Verschlüsselung

Ein Verschlüsselungsalgorithmus, der denselben Schlüssel zum Verschlüsseln und Entschlüsseln der Daten verwendet.

synthetisches Testen

Testen eines Systems auf eine Weise, die Benutzerinteraktionen simuliert, um potenzielle Probleme zu erkennen oder die Leistung zu überwachen. Sie können [Amazon CloudWatch Synthetics](#) verwenden, um diese Tests zu erstellen.

Systemaufforderung

Eine Technik, mit der einem [LLM](#) Kontext, Anweisungen oder Richtlinien zur Verfügung gestellt werden, um sein Verhalten zu steuern. Systemaufforderungen helfen dabei, den Kontext festzulegen und Regeln für Interaktionen mit Benutzern festzulegen.

T

tags

Schlüssel-Wert-Paare, die als Metadaten für die Organisation Ihrer Ressourcen dienen. AWS Mit Tags können Sie Ressourcen verwalten, identifizieren, organisieren, suchen und filtern. Weitere Informationen finden Sie unter [Markieren Ihrer AWS -Ressourcen](#).

Zielvariable

Der Wert, den Sie in überwachtem ML vorhersagen möchten. Dies wird auch als Ergebnisvariable bezeichnet. In einer Fertigungsumgebung könnte die Zielvariable beispielsweise ein Produktfehler sein.

Aufgabenliste

Ein Tool, das verwendet wird, um den Fortschritt anhand eines Runbooks zu verfolgen. Eine Aufgabenliste enthält eine Übersicht über das Runbook und eine Liste mit allgemeinen Aufgaben, die erledigt werden müssen. Für jede allgemeine Aufgabe werden der geschätzte Zeitaufwand, der Eigentümer und der Fortschritt angegeben.

Testumgebungen

[Siehe Umgebung.](#)

Training

Daten für Ihr ML-Modell bereitstellen, aus denen es lernen kann. Die Trainingsdaten müssen die richtige Antwort enthalten. Der Lernalgorithmus findet Muster in den Trainingsdaten, die die Attribute der Input-Daten dem Ziel (die Antwort, die Sie voraussagen möchten) zuordnen. Es gibt ein ML-Modell aus, das diese Muster erfasst. Sie können dann das ML-Modell verwenden, um Voraussagen für neue Daten zu erhalten, bei denen Sie das Ziel nicht kennen.

Transit-Gateway

Ein Netzwerk-Transit-Hub, über den Sie Ihre Netzwerke VPCs und Ihre lokalen Netzwerke miteinander verbinden können. Weitere Informationen finden Sie in der Dokumentation unter [Was ist ein Transit-Gateway](#). AWS Transit Gateway

Stammbasierter Workflow

Ein Ansatz, bei dem Entwickler Feature lokal in einem Feature-Zweig erstellen und testen und diese Änderungen dann im Hauptzweig zusammenführen. Der Hauptzweig wird dann sequentiell für die Entwicklungs-, Vorproduktions- und Produktionsumgebungen erstellt.

Vertrauenswürdiger Zugriff

Gewährung von Berechtigungen für einen Dienst, den Sie angeben, um Aufgaben in Ihrer Organisation AWS Organizations und in deren Konten in Ihrem Namen auszuführen. Der vertrauenswürdige Service erstellt in jedem Konto eine mit dem Service verknüpfte Rolle, wenn diese Rolle benötigt wird, um Verwaltungsaufgaben für Sie auszuführen. Weitere Informationen

finden Sie in der AWS Organizations Dokumentation [unter Verwendung AWS Organizations mit anderen AWS Diensten](#).

Optimieren

Aspekte Ihres Trainingsprozesses ändern, um die Genauigkeit des ML-Modells zu verbessern. Sie können das ML-Modell z. B. trainieren, indem Sie einen Beschriftungssatz generieren, Beschriftungen hinzufügen und diese Schritte dann mehrmals unter verschiedenen Einstellungen wiederholen, um das Modell zu optimieren.

Zwei-Pizzen-Team

Ein kleines DevOps Team, das Sie mit zwei Pizzen ernähren können. Eine Teamgröße von zwei Pizzen gewährleistet die bestmögliche Gelegenheit zur Zusammenarbeit bei der Softwareentwicklung.

U

Unsicherheit

Ein Konzept, das sich auf ungenaue, unvollständige oder unbekannte Informationen bezieht, die die Zuverlässigkeit von prädiktiven ML-Modellen untergraben können. Es gibt zwei Arten von Unsicherheit: Epistemische Unsicherheit wird durch begrenzte, unvollständige Daten verursacht, wohingegen aleatorische Unsicherheit durch Rauschen und Randomisierung verursacht wird, die in den Daten liegt. Weitere Informationen finden Sie im Leitfaden [Quantifizieren der Unsicherheit in Deep-Learning-Systemen](#).

undifferenzierte Aufgaben

Diese Arbeit wird auch als Schwerstarbeit bezeichnet. Dabei handelt es sich um Arbeiten, die zwar für die Erstellung und den Betrieb einer Anwendung erforderlich sind, aber dem Endbenutzer keinen direkten Mehrwert bieten oder keinen Wettbewerbsvorteil bieten. Beispiele für undifferenzierte Aufgaben sind Beschaffung, Wartung und Kapazitätsplanung.

höhere Umgebungen

Siehe [Umgebung](#).

V

Vacuuming

Ein Vorgang zur Datenbankwartung, bei dem die Datenbank nach inkrementellen Aktualisierungen bereinigt wird, um Speicherplatz zurückzugewinnen und die Leistung zu verbessern.

Versionskontrolle

Prozesse und Tools zur Nachverfolgung von Änderungen, z. B. Änderungen am Quellcode in einem Repository.

VPC-Peering

Eine Verbindung zwischen zwei VPCs , die es Ihnen ermöglicht, den Verkehr mithilfe privater IP-Adressen weiterzuleiten. Weitere Informationen finden Sie unter [Was ist VPC-Peering?](#) in der Amazon-VPC-Dokumentation.

Schwachstelle

Ein Software- oder Hardwarefehler, der die Sicherheit des Systems beeinträchtigt.

W

Warmer Cache

Ein Puffer-Cache, der aktuelle, relevante Daten enthält, auf die häufig zugegriffen wird. Die Datenbank-Instance kann aus dem Puffer-Cache lesen, was schneller ist als das Lesen aus dem Hauptspeicher oder von der Festplatte.

warme Daten

Daten, auf die selten zugegriffen wird. Bei der Abfrage dieser Art von Daten sind mäßig langsame Abfragen in der Regel akzeptabel.

Fensterfunktion

Eine SQL-Funktion, die eine Berechnung für eine Gruppe von Zeilen durchführt, die sich in irgendeiner Weise auf den aktuellen Datensatz beziehen. Fensterfunktionen sind nützlich für die Verarbeitung von Aufgaben wie die Berechnung eines gleitenden Durchschnitts oder für den Zugriff auf den Wert von Zeilen auf der Grundlage der relativen Position der aktuellen Zeile.

Workload

Ein Workload ist eine Sammlung von Ressourcen und Code, die einen Unternehmenswert bietet, wie z. B. eine kundenorientierte Anwendung oder ein Backend-Prozess.

Workstream

Funktionsgruppen in einem Migrationsprojekt, die für eine bestimmte Reihe von Aufgaben verantwortlich sind. Jeder Workstream ist unabhängig, unterstützt aber die anderen Workstreams im Projekt. Der Portfolio-Workstream ist beispielsweise für die Priorisierung von Anwendungen, die Wellenplanung und die Erfassung von Migrationsmetadaten verantwortlich. Der Portfolio-Workstream liefert diese Komponenten an den Migrations-Workstream, der dann die Server und Anwendungen migriert.

WURM

Sehen [Sie einmal schreiben, viele lesen](#).

WQF

Siehe [AWS Workload-Qualifizierungsrahmen](#).

einmal schreiben, viele lesen (WORM)

Ein Speichermodell, das Daten ein einziges Mal schreibt und verhindert, dass die Daten gelöscht oder geändert werden. Autorisierte Benutzer können die Daten so oft wie nötig lesen, aber sie können sie nicht ändern. Diese Datenspeicherinfrastruktur gilt als [unveränderlich](#).

Z

Zero-Day-Exploit

Ein Angriff, in der Regel Malware, der eine [Zero-Day-Sicherheitslücke](#) ausnutzt.

Zero-Day-Sicherheitslücke

Ein unfehlbarer Fehler oder eine Sicherheitslücke in einem Produktionssystem. Bedrohungsakteure können diese Art von Sicherheitslücke nutzen, um das System anzugreifen. Entwickler werden aufgrund des Angriffs häufig auf die Sicherheitsanfälligkeit aufmerksam.

Zero-Shot-Aufforderung

Bereitstellung von Anweisungen für die Ausführung einer Aufgabe an einen [LLM](#), jedoch ohne Beispiele (Schnappschüsse), die ihm als Orientierungshilfe dienen könnten. Der LLM muss sein

vortrainiertes Wissen einsetzen, um die Aufgabe zu bewältigen. Die Effektivität von Zero-Shot Prompting hängt von der Komplexität der Aufgabe und der Qualität der Aufforderung ab. [Siehe auch Few-Shot-Prompting.](#)

Zombie-Anwendung

Eine Anwendung, deren durchschnittliche CPU- und Arbeitsspeichernutzung unter 5 Prozent liegt. In einem Migrationsprojekt ist es üblich, diese Anwendungen außer Betrieb zu nehmen.

Die vorliegende Übersetzung wurde maschinell erstellt. Im Falle eines Konflikts oder eines Widerspruchs zwischen dieser übersetzten Fassung und der englischen Fassung (einschließlich infolge von Verzögerungen bei der Übersetzung) ist die englische Fassung maßgeblich.