Scelta di un database AWS vettoriale per i casi d'uso di RAG

AWS Linee guida prescrittive

AWS Linee guida prescrittive: Scelta di un database AWS vettoriale per i casi d'uso di RAG

Table of Contents

Introduzione 1
Destinatari principali 1
Panoramica dei vettori 3
Panoramica dei database vettoriali 5
Opzioni del database vettoriale 7
Opzioni di database vettoriali individuali 7
Amazon Kendra 7
OpenSearch Servizio Amazon 8
Amazon RDS per PostgreSQL con pgvector 9
Amazon DocumentDB 9
Amazon MemoryDB 10
Analisi di Amazon Neptune 11
Amazon S3 Vectors 11
Opzione di servizio gestito 12
Scelta del database vettoriale giusto 13
Confronto tra database vettoriali 15
Database vettoriali individuali 15
Servizio gestito — Amazon Bedrock Knowledge Bases 19
Scelta tra opzioni individuali e gestite 22
Confronti e considerazioni sui costi 23
Casi d'uso di database vettoriali 27
Gestione della conoscenza con Amazon Kendra 27
Analisi in tempo reale con Serverless OpenSearch 28
Risorse e passaggi successivi 29
Resources 29
AWS post sul blog 30
AWS documentazione di servizio 30
Altre risorse AWS 30
Altre risorse 30
Cronologia dei documenti 32
..... xxxiii

Scelta di un database AWS vettoriale per i casi d'uso di RAG

Mayuri Shinde, Ivan Cui e Anand Bukkapatnam Tirumala, Amazon Web Services

Marzo 2026 (storia del documento)

I database vettoriali stanno diventando sempre più importanti per le organizzazioni che implementano applicazioni di intelligenza artificiale generativa. Questi database archiviano e gestiscono i vettori, che sono rappresentazioni numeriche di dati che consentono l'elaborazione di testo, immagini e altri contenuti in modo da catturarne il significato e le relazioni.

Quando le organizzazioni esplorano le opzioni relative ai database vettoriali AWS, devono comprendere le funzionalità, i compromessi e le migliori pratiche per le diverse soluzioni. Questa guida ti aiuta a confrontare gli archivi vettoriali di uso comune AWS e a prendere decisioni informate su quali opzioni si adattano meglio alle tue esigenze o ai tuoi casi d'uso specifici. Che stiate implementando Retrieval Augmented Generation (RAG), creando sistemi di raccomandazione o sviluppando altre applicazioni di intelligenza artificiale, questa guida fornisce un framework per aiutarvi a valutare e scegliere una soluzione di database vettoriale.

Destinatari principali

Questa guida è destinata alle persone che ricoprono i seguenti ruoli:

- Scienziati dei dati e ingegneri del machine learning (ML) che utilizzano database vettoriali per archiviare e recuperare dati ad alta dimensione per modelli ML.
- Ingegneri dei dati che progettano e implementano pipeline di dati che includono database vettoriali per l'archiviazione e l'elaborazione di dati ad alta dimensione.
- MLOps ingegneri che utilizzano database vettoriali come parte della pipeline ML per archiviare e fornire output di modelli o rappresentazioni intermedie.
- Ingegneri del software che integrano database vettoriali in applicazioni che richiedono sistemi di ricerca o raccomandazione per analogie.
- DevOps ingegneri responsabili della distribuzione e della manutenzione dei database vettoriali negli ambienti di produzione, garantendo scalabilità e affidabilità.
- Ricercatori di intelligenza artificiale che utilizzano database vettoriali per archiviare e analizzare grandi set di dati di incorporamenti o vettori di funzionalità.

- Responsabili di prodotto di intelligenza artificiale che devono comprendere le funzionalità e i limiti dei database vettoriali per prendere decisioni informate sulle caratteristiche e sull'architettura del prodotto.

Panoramica dei vettori

I vettori sono rappresentazioni numeriche che aiutano le macchine a comprendere ed elaborare i dati. Nell'intelligenza artificiale generativa, hanno due scopi principali:

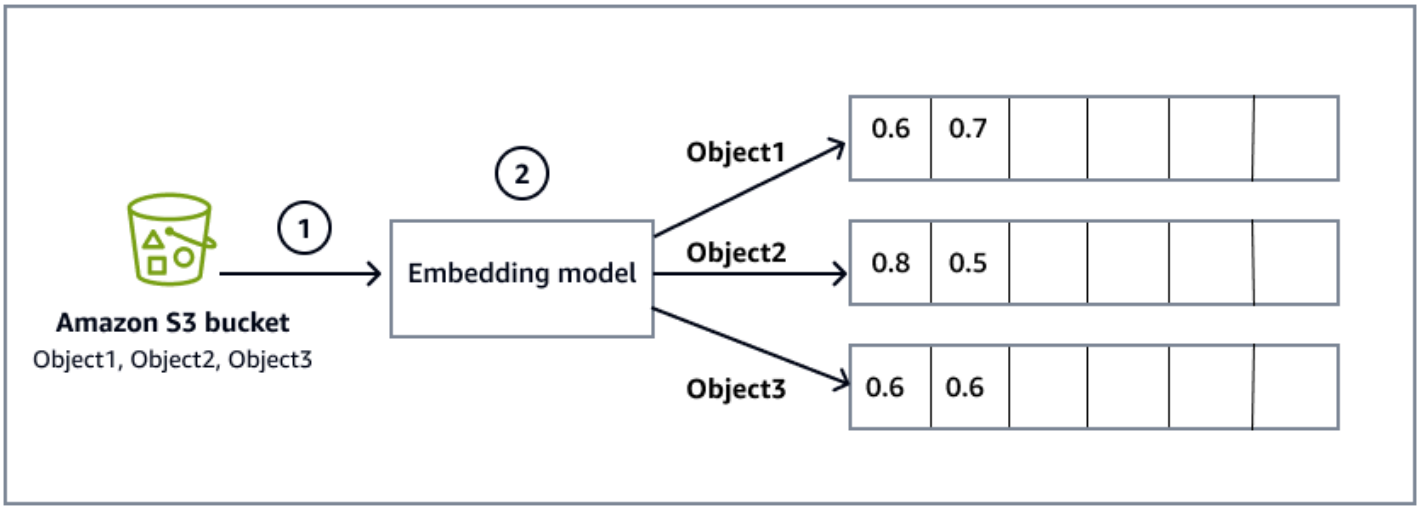
- Rappresentano spazi latenti che acquisiscono la struttura dei dati in forma compressa
- Creazione di incorporamenti per dati, come parole, frasi e immagini

I modelli di incorporamento come Word2Vec GloVee Amazon Titan Text Embeddings convertono i dati in vettori tramite un processo chiamato incorporamento. Questi modelli di incorporamento possono fare quanto segue:

- Impara dal contesto per rappresentare le parole come vettori
- Posiziona parole simili più vicine tra loro nello spazio vettoriale
- Consenti alle macchine di elaborare i dati in uno spazio continuo

Il diagramma seguente fornisce una panoramica di alto livello del processo di incorporamento:

1. Un bucket Amazon Simple Storage Service (Amazon S3) Simple Storage Service (Amazon S3) contiene file che sono le fonti di dati da cui il sistema legge ed elabora le informazioni. Il bucket Amazon S3 viene specificato durante la configurazione della knowledge base di Amazon Bedrock, che include anche la sincronizzazione dei dati con la knowledge base.
2. Il modello di incorporamento converte i dati grezzi dai file oggetto nel bucket Amazon S3 in incorporamenti vettoriali. Ad esempio, `Object1` viene convertito in un vettore `[0.6, 0.7, ...]` che ne rappresenta il contenuto in uno spazio multidimensionale.



Gli incorporamenti di testi sono fondamentali per l'elaborazione del linguaggio naturale (NLP) perché svolgono le seguenti funzioni:

- Cattura le relazioni semantiche tra le parole
- Abilita la generazione di testo contestualmente rilevante
- Potenzia modelli linguistici di grandi dimensioni (LLM) per produrre risposte simili a quelle umane

Panoramica dei database vettoriali

Un database vettoriale è un sistema specializzato che archivia e interroga vettori ad alta dimensione in modo efficiente. Questi database sono fondamentali per le applicazioni Retrieval Augmented Generation (RAG).

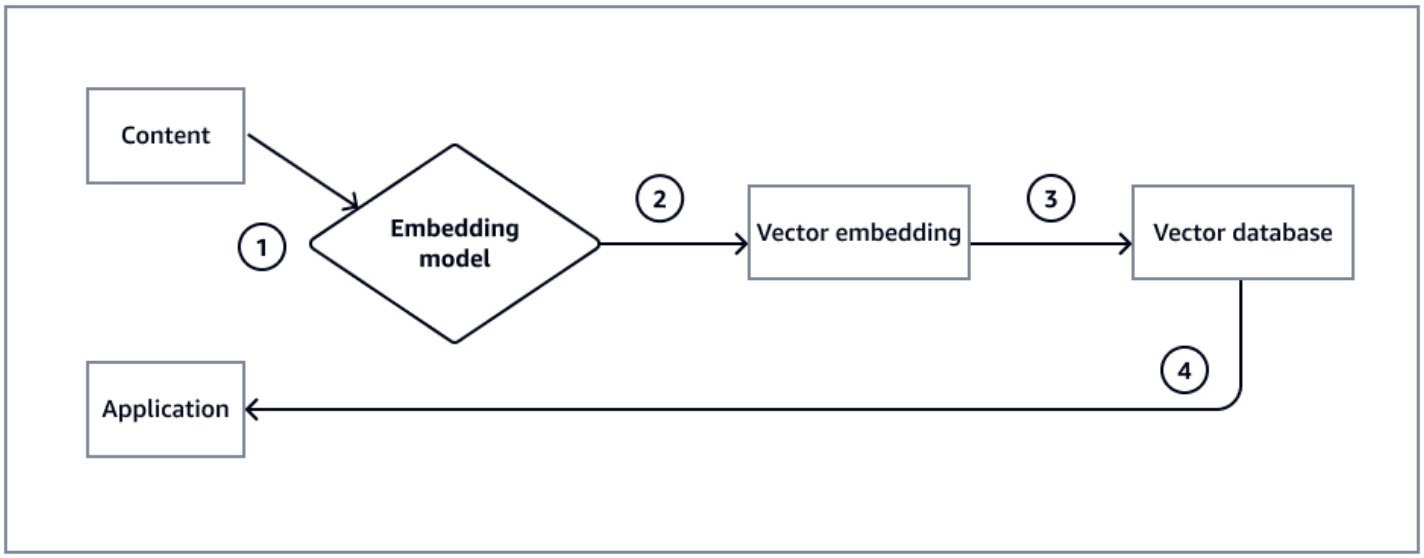
I database vettoriali gestiscono la conversione e l'archiviazione dei dati nei seguenti modi:

- Gli oggetti (come audio, immagini e file di testo) vengono convertiti in vettori utilizzando modelli di incorporamento.
- I vettori vengono archiviati in formati di dati specializzati.
- I database vettoriali consentono ricerche rapide di similarità.

I database vettoriali offrono diversi vantaggi chiave rispetto ai database tradizionali, il che li rende particolarmente adatti per le moderne sfide legate ai dati. Sono ottimizzati specificamente per le operazioni vettoriali e gestiscono dati ad alta dimensione in modo efficiente. Sono inoltre specializzati nelle ricerche di similarità che i database tradizionali non riescono a gestire. Oltre a queste funzionalità principali, i database vettoriali sono progettati per soddisfare le esigenze in continua evoluzione delle applicazioni ML e di intelligenza artificiale generativa. Eccellono nello storage vettoriale su larga scala e utilizzano il calcolo distribuito per bilanciare i carichi di lavoro su più nodi. Ciò offre scalabilità e prestazioni man mano che i volumi di dati crescono.

Il diagramma seguente mostra un'implementazione RAG:

1. Il contenuto, come documenti, PDF o file di testo, viene inserito nel modello di incorporamento come dati grezzi per l'elaborazione.
2. Il modello di incorporamento trasforma i dati grezzi in vettori numerici, che rappresentano il significato semantico del contenuto.
3. Gli incorporamenti vettoriali generati vengono archiviati in un database vettoriale ottimizzato per l'archiviazione e il recupero di vettori ad alta dimensione.
4. Le applicazioni possono ora interrogare il database vettoriale in risposta a casi d'uso come la ricerca semantica e la raccomandazione di contenuti.



La scelta di un database vettoriale inappropriato per una soluzione RAG può portare a difficoltà e limitazioni significative, tra cui:

- Scarse prestazioni delle query
- Colli di bottiglia in termini di scalabilità
- Sfide legate all'ingestione dei dati
- Mancanza di funzionalità avanzate, come il filtraggio e il posizionamento
- Difficoltà di integrazione con altri sistemi
- Problemi di persistenza e durabilità
- Problemi di concorrenza e coerenza in ambienti con più utenti
- Costi di licenza più elevati o vincolo al fornitore
- Supporto e risorse limitati per la community
- Potenziali rischi per la sicurezza e la conformità

Opzioni del database vettoriale

AWS offre una vasta gamma di soluzioni di database vettoriali per supportare diversi casi d'uso e requisiti nelle applicazioni di intelligenza artificiale generativa. Queste opzioni possono essere ampiamente suddivise in singoli servizi di database e offerte di servizi gestiti, ciascuno con caratteristiche e vantaggi distinti. La comprensione di queste opzioni è fondamentale per le organizzazioni che desiderano implementare in modo efficace le funzionalità di ricerca vettoriale mantenendo prestazioni, scalabilità ed efficienza dei costi ottimali.

Per ulteriori informazioni sulle soluzioni di database vettoriali, consulta le seguenti sezioni:

- Opzioni di database vettoriali individuali
- Opzione di servizio gestito
- Scelta del database vettoriale giusto

Opzioni di database vettoriali individuali

Le opzioni di database vettoriali individuali AWS includono Amazon Kendra OpenSearch , AmazonService, Amazon RDS per PostgreSQL con pgvector,Amazon MemoryDB, Amazon DocumentDB, Amazon DocumentDB,Amazon Neptune Analytics e Amazon S3 Vector. (Un'estensione open source, pgvector aggiunge la possibilità di archiviare e cercare incorporamenti vettoriali generati da ML.) Queste soluzioni offrono approcci diversi alla ricerca vettoriale, consentendo alle organizzazioni di scegliere in base all'infrastruttura esistente, ai requisiti tecnici e ai casi d'uso specifici.

Amazon Kendra

Amazon Kendra è un servizio di ricerca intelligente di livello aziendale che utilizza l'elaborazione del linguaggio naturale e algoritmi avanzati di apprendimento automatico per restituire risposte specifiche alle domande di ricerca dai tuoi dati. Amazon Kendra semplifica l'implementazione della funzionalità di ricerca, rendendola una soluzione di backend efficace per applicazioni di intelligenza artificiale generativa.

Altre caratteristiche chiave di Amazon Kendra includono:

- Connessioni native a oltre 40 fonti di dati
- Funzionalità integrate di preparazione dei dati

- Configurazione rapida che non richiede competenze tecniche approfondite

I vantaggi di Amazon Kendra includono i seguenti:

- Elaborazione automatizzata dei dati (suddivisione in blocchi, ingestione, recupero)
- Potenti opzioni di personalizzazione:
  - Ricerca sfaccettata
  - Analisi della ricerca
  - Ottimizzazione della pertinenza della ricerca
- Accesso programmatico semplice tramite AWS SDK for Python (Boto3)

Per ulteriori informazioni, consulta i vantaggi di Amazon Kendra nella documentazione di Amazon Kendra.

OpenSearch Servizio Amazon

Amazon OpenSearch Service è un servizio gestito che ti aiuta a distribuire, gestire e scalare i cluster OpenSearch di servizi in. Cloud AWS

Le funzionalità principali di OpenSearch Service includono quanto segue:

- Motore di ricerca e analisi open source
- Architettura distribuita
- Elaborazione dati in tempo reale

Alcuni vantaggi dell'utilizzo del OpenSearch servizio includono quanto segue:

- Scalabilità orizzontale
- RESTful Supporto API
- Gestisce dati strutturati e non strutturati
- Analisi dei dati in tempo reale
- Adatto a diverse dimensioni di implementazione

Per ulteriori informazioni, consulta le funzionalità di Amazon OpenSearch Service nella documentazione del OpenSearch servizio.

Amazon RDS per PostgreSQL con pgvector

Amazon RDS per PostgreSQL con AWS pgvector combina il servizio di database relazionale gestito con l'estensione di elaborazione vettoriale di PostgreSQL. Questa combinazione consente alle organizzazioni di archiviare e interrogare vettori ad alta dimensione mantenendo Amazon RDS. La soluzione è particolarmente adatta per applicazioni di intelligenza artificiale generativa che richiedono operazioni vettoriali in tempo reale senza il sovraccarico di gestione dell'infrastruttura di database.

I vantaggi principali di Amazon RDS for PostgreSQL con pgvector includono:

- Elevata disponibilità
- Failover automatico
- pay-per-useConveniente ()
- Monitoraggio integrato
- Integrazione di dati vettoriali in tempo reale

Per ulteriori informazioni, consulta i vantaggi di Amazon RDS nella documentazione di Amazon RDS.

Amazon DocumentDB

Amazon DocumentDB (compatibile con MongoDB) è un database di documenti che offre funzionalità di ricerca vettoriale native nella versione 5.0 e successive. Combina la flessibilità dell'archiviazione di documenti basata su JSON con la ricerca vettoriale, supportando i metodi di indicizzazione Hierarchical Navigable Small World (HNSW) e Inverted File Flat (). IVFFlat

Le funzionalità principali di Amazon DocumentDB includono quanto segue:

- Archivia e indicizza vettori fino a 2.000 dimensioni (fino a 16.000 dimensioni senza indicizzazione)
- Tempi di risposta in millisecondi per le ricerche di similarità vettoriale
- Support per le metriche della distanza tra prodotti euclidei, coseno e scalari
- Perfetta integrazione con le applicazioni esistenti compatibili con MongoDB

Usa Amazon DocumentDB nelle seguenti situazioni:

- Per applicazioni che utilizzano già APIs MongoDB e che necessitano di funzionalità di ricerca vettoriale

- Per casi d'uso che richiedono strutture di dati di documenti flessibili combinate con la ricerca semantica
- Per scenari che richiedono sia le tradizionali interrogazioni sui documenti che le ricerche di similarità vettoriale
- Per applicazioni che forniscono consigli sui prodotti, personalizzazione, assistenti di chat e rilevamento delle frodi

Per ulteriori informazioni, consulta Vector search for Amazon DocumentDB nella documentazione di Amazon DocumentDB.

Amazon MemoryDB

Amazon MemoryDB è un database in memoria compatibile con Redis che offre le prestazioni di ricerca vettoriale più veloci tra i database vettoriali più diffusi su. AWS Fornisce latenze di query inferiori al millisecondo con durabilità Multi-Availability Zone.

Le funzionalità principali di MemoryDB includono quanto segue:

- Archivia i dati delle applicazioni e milioni di vettori in un unico database
- Tempi di risposta alle interrogazioni e agli aggiornamenti a una cifra in millisecondi
- Tassi di richiamo più elevati alle prestazioni più elevate su AWS
- Support per un massimo di 32.768 dimensioni per vettore
- Funzionalità di ricerca semantica e memorizzazione nella cache in tempo reale

Usa MemoryDB nelle seguenti situazioni:

- Per applicazioni in tempo reale che richiedono una latenza estremamente bassa (inferiore a 10 ms)
- Per carichi di lavoro ad alta velocità con milioni di richieste al giorno
- Per casi d'uso come motori di raccomandazione in tempo reale, memorizzazione nella cache semantica e rilevamento di anomalie
- Per applicazioni che richiedono sia funzionalità di archiviazione dati in memoria che di ricerca vettoriale

Per ulteriori informazioni, consulta Ricerca vettoriale nella documentazione di MemoryDB.

Analisi di Amazon Neptune

Amazon Neptune Analytics è un motore di analisi dei grafici che offre funzionalità di ricerca vettoriale native, che lo rendono ideale per i casi d'uso di Graph Retrieval Augmented Generation (GraphRag). Combina la ricerca per similarità vettoriale con algoritmi e attraversamenti di grafici.

Le funzionalità principali di Neptune Analytics includono quanto segue:

- Analizza decine di miliardi di relazioni in pochi secondi
- Combina la ricerca vettoriale con algoritmi grafici (ricerca del percorso, rilevamento della comunità, centralità)
- Support per applicazioni GraphRag con conoscenze topologiche
- Fino a 80 volte più veloce rispetto alle soluzioni analitiche grafiche esistenti
- Integrazione con Amazon Bedrock per GraphRag completamente gestito

Usa Neptune Analytics nelle seguenti situazioni:

- Per applicazioni GraphRag che richiedono grafici della conoscenza con incorporamenti vettoriali
- Per casi d'uso che richiedono l'attraversamento di relazioni complesse insieme alla somiglianza vettoriale
- Per applicazioni che richiedono risposte AI spiegabili con contesto relazionale
- Per scenari quali visualizzazioni a 360 gradi dei clienti, reti di rilevamento delle frodi e scoperta delle conoscenze

Per ulteriori informazioni, consulta la documentazione di Amazon Neptune Analytics.

Amazon S3 Vectors

Amazon S3 Vectors è il primo archivio di oggetti nel cloud AWS con funzionalità native di archiviazione vettoriale e di interrogazione. Fornisce uno storage vettoriale personalizzato e ottimizzato in termini di costi per applicazioni di intelligenza artificiale che richiedono una scalabilità enorme.

Le funzionalità principali di Amazon S3 Vectors includono quanto segue:

- Archiviazione per un massimo di 2 miliardi di vettori per indice con supporto per un massimo di 10.000 indici per bucket vettoriale

- Latenza delle query inferiore a 100 ms ottimizzata per l'archiviazione a lungo termine e modelli di accesso poco frequenti
- Riduzione dei costi fino al 90% per le operazioni vettoriali rispetto ai database vettoriali specializzati
- Architettura serverless con scalabilità automatica e durabilità del 99,25% (11 9s)

Usa Amazon S3 Vectors nelle seguenti situazioni:

- Per applicazioni che richiedono lo storage di miliardi di vettori a costi minimi
- Per carichi di lavoro che tollerano una latenza delle query inferiore al secondo (100 ms o più) anziché inferiore a 10 ms
- Per casi d'uso di conservazione vettoriale e archiviazione a lungo termine
- Per applicazioni RAG con schemi di recupero poco frequenti
- Per le organizzazioni che danno priorità ai vantaggi economici dello storage rispetto alla latenza ultra bassa

Amazon S3 Vectors si integra nativamente con le Knowledge Base di Amazon Bedrock e funziona bene in architetture a più livelli con Amazon Service. OpenSearch Puoi usare Amazon S3 Vectors per la conservazione a freddo e usare OpenSearch Service per le hot query.

Per ulteriori informazioni, consulta Lavorare con vettori S3 e bucket vettoriali nella documentazione di Amazon S3.

Opzione di servizio gestito

Amazon Bedrock Knowledge Bases rappresenta l'approccio AWS completamente gestito all'implementazione di database vettoriali. La flessibilità del servizio nelle opzioni di storage, combinata con le sue funzionalità di gestione automatizzata, lo rende particolarmente utile per le organizzazioni che desiderano implementare RAG senza gestire infrastrutture complesse.

Con Amazon Bedrock Knowledge Bases, puoi creare, gestire e interrogare basi di conoscenza che migliorano i tuoi modelli di base utilizzando RAG. Questo servizio semplifica il complesso processo di implementazione di RAG gestendo l'intera pipeline di inserimento, vettorizzazione e recupero dei dati.

I vantaggi principali delle Knowledge Base di Amazon Bedrock includono:

- Elaborazione semplificata dei dati

  - Inserimento e suddivisione automatici dei dati
  - Estrazione di testo integrata da più formati di file
  - Generazione gestita di incorporamenti vettoriali
  - Estrazione e indicizzazione automatiche dei metadati
- Implementazione RAG semplificata
  - Strategie di recupero preconfigurate
  - Ottimizzazione automatica della finestra contestuale
  - Ottimizzazione della pertinenza integrata
  - Funzionalità di ricerca semantica pronte all'uso
- Sicurezza e governance
  - Controlli integrati AWS Identity and Access Management (IAM)
  - Crittografia dei dati a riposo e in transito
  - Supporto per VPC
  - Registrazione di audit con AWS CloudTrail

Amazon Bedrock Knowledge Bases supporta diverse opzioni di archiviazione vettoriale, tra cui:

- Amazon Aurora PostgreSQL con pgvector
- Analisi di Amazon Neptune
- Amazon EMR Serverless
- Amazon S3 Vectors
- Pigna
- Redis Enterprise Cloud

Questo servizio gestito gestisce l'inserimento, la vettorizzazione e il recupero automatizzati dei dati. Questo semplifica le implementazioni RAG.

Per informazioni dettagliate su ogni archivio vettoriale supportato, consulta la documentazione di Amazon Bedrock Knowledge Bases.

Scelta del database vettoriale giusto

Seleziona il tuo database vettoriale in base a questi fattori decisionali chiave:

- Se hai bisogno di un database di documenti compatibile con MongoDB con ricerca vettoriale, scegli Amazon DocumentDB. Questa soluzione è ideale quando l'applicazione utilizza APIs MongoDB e si desidera aggiungere funzionalità di ricerca semantica senza gestire un'infrastruttura vettoriale separata.
- Se hai bisogno di una latenza ultra bassa per applicazioni in tempo reale, scegli Amazon MemoryDB. Ciò fornisce le prestazioni di ricerca vettoriale più veloci con tempi di risposta inferiori al millisecondo AWS . È ideale per i motori di raccomandazione in tempo reale e le applicazioni ad alto rendimento.
- Se hai bisogno di rappresentazioni della conoscenza basate su grafici con ricerca vettoriale, scegli Amazon Neptune Analytics. Questa è la soluzione ideale per le applicazioni GraphRag che devono attraversare relazioni complesse ed eseguire query basate su grafici insieme a ricerche vettoriali, fornendo risposte di intelligenza artificiale spiegabili.
- Se hai bisogno di combinare query relazionali con la ricerca vettoriale, scegli Amazon Aurora PostgreSQL con pgvector. Questa opzione è ideale quando l'applicazione richiede sia operazioni SQL tradizionali che ricerche di similarità vettoriale all'interno dello stesso database.
- Se hai bisogno di query ad alta velocità con latenza inferiore a 10 ms, scegli Amazon Service. OpenSearch Eccelle nella gestione di query ad alta frequenza e applicazioni in tempo reale e include recenti miglioramenti nell'accelerazione della GPU.
- Se hai bisogno di archiviare miliardi di vettori a costi contenuti, scegli Amazon S3 Vectors. Questa opzione offre un risparmio sui costi fino al 90% ed è ideale per applicazioni con schemi di recupero poco frequenti (da minuti a ore tra le query) che possono tollerare una latenza inferiore a 100 ms.
- Se hai bisogno di una ricerca di testo completo oltre alla ricerca vettoriale, scegli Amazon OpenSearch Service. Questa opzione combina potenti funzionalità di ricerca full-text con la ricerca vettoriale in un'unica piattaforma.

Confronto tra database vettoriali

AWS offre diversi approcci all'implementazione delle funzionalità di ricerca vettoriale, dai database vettoriali individuali alle Amazon Bedrock Knowledge Bases, un servizio completamente gestito. Nel valutare queste opzioni, le organizzazioni devono considerare vari aspetti, tra cui architettura, scalabilità, capacità di integrazione, caratteristiche prestazionali e caratteristiche di sicurezza.

Database vettoriali individuali

La tabella seguente fornisce una panoramica delle funzionalità chiave di diverse soluzioni di database vettoriali AWS individuali, con particolare attenzione alle architetture, alle capacità di scalabilità, alle integrazioni delle fonti di dati e alle caratteristiche prestazionali.

Funzionalità	Amazon Kendra	OpenSearch Servizio Amazon	Amazon RDS per SQLwith Postgre pgvector	Amazon DocumentDB	Amazon MemoryDB	Analisi di Amazon Neptune	Amazon S3 Vectors
Caso d'uso principale	Ricerca aziendale e RAG	Ricerca e analisi distribuite	DB relazionale con supporto vettoriale	DB di documenti con ricerca vettoriale	Ricerca vettoriale in memoria in tempo reale	Analisi grafica con ricerca vettoriale	Archiviazione vettoriale ottimizzata in termini di costi
Architecture	Servizio completamente gestito	Cluster distribuito	Database relazionale	Orientato ai documenti	Database in memoria	Motore di analisi grafica	Storage di oggetti senza server
Modello di dati	Basato su documenti	Documenti JSON	Tabelle relazionali	Documenti JSON	Valore-chiave con JSON	Grafico delle proprietà	Archiviazione di oggetti

Dimensioni vettoriali	Gestito automaticamente	Fino a 16.000	Configurabile	Fino a 2.000 (indicizzati); 16.000 (non indicizzati)	Fino a 32.768	Configurabile	Fino a 4.096
Metodi di indicizzazione	Automatica	HNSW, IVF	HNSW, IVFFlat	NSW, IVFFlat	HNSW	Grafico e vettore nativi	Automatica
Metriche della distanza	Automatica	Coseno, euclideo, prodotto a punti	Coseno, euclideo, prodotto interno	Coseno, euclideo, prodotto a punti	Coseno, euclideo, prodotto interno	Coseno, euclideo	Coseno, euclideo
Latenza delle interrogazioni	Frazioni di secondo	Meno di 10 ms (accelerato da GPU)	10-100 ms	Millisecondi	Meno di millisecondo	Frazioni di secondo	Meno di 100 ms
Modello di ridimensionamento	Automatica	Orizzontale (aggiungi nodi)	Repliche verticali e lette	Orizzontale (aggiungi istanze)	Verticale e repliche	Automatica	Automatico (senza server)
Numero massimo di vettori	Gestiti	Miliardi (a seconda del cluster)	Milioni (a seconda dell'istanza)	Milioni per collezione	Milioni per database	Miliardi	2 miliardi per indice; 10.000 indici per bucket

Throughput	Elevata	Molto alto (migliaia di QPS)	Media	Elevata	Molto alto (milioni di richieste al giorno)	Elevata	Medio (ottimizzato per richieste poco frequenti)
Durabilità dei dati	99,999999999% (11 9 s)	Configurabile con repliche	99,99% (Multi-AZ)	99,99% (Multi-AZ)	99,99% (Multi-AZ)	99,99%	99,999999999% (11 9 s)
Modello di coerenza	Eventuale	Eventuale (configurabile)	Forte (ACIDO)	Eventuale	Forte	Forte	Forte
Funzionalità aggiuntive	40 o più connettori dati, NLP	Ricerca nel testo completo, analisi, dashboard	Interrogazioni SQL, transazioni ACID	Compatibilità con le API MongoDB	Compatibilità con le API Redis, memorizzazione nella cache	Algoritmi grafici, traversi	Integrazione con Amazon S3 e politiche del ciclo di vita
Modello tariffario	Pagamento in base alle richieste e allo storage	Ore e spazio di archiviazione delle istanze	Ore e archiviazione delle istanze	Ore e archiviazione delle istanze	Ore e archiviazione delle istanze	Unità di capacità e storage	Archiviazione, interrogazioni e trasferimento dati

Ottimizzazione dei costi	Basato sull'utilizzo	Istanze riservate, auto-scaling	Istanze riservate, Aurora Serverless	Istanze riservate	Istanze riservate	Dimensionamento automatico	Fino al 90% di risparmio rispetto alle soluzioni specializzate DBs
Ideale per	Ricerca aziendale con configurazione minima	Query ad alta velocità e bassa latenza	Carichi di lavoro SQL e vettoriali ibridi	App compatibili con MongoDB che necessitano di vettori	App in tempo reale con latenza ultra bassa	GraphRag e grafici della conoscenza	Storage a lungo termine ed economico
Modello di interrogazione ideale	Ricerche aziendali frequenti	Interrogazioni in tempo reale ad alta frequenza	Interrogazioni SQL e vettoriali miste	Interrogazioni documentali con ricerca semantica	Milioni di richieste al giorno	Rappresenta graficamente le traversate con la ricerca vettoriale	Domande poco frequenti (da minuti a ore)
Complessità della configurazione	Bassa (completamente gestita)	Medio (configurazione del cluster)	Medio (configurazione dell'estensione)	Medio (configurazione del cluster)	Medio (configurazione del cluster)	Basso (completamente gestito)	Basso (senza server)

È richiesta l'esperienza del team	Minima	OpenSearch o Elasticsearch	PostgreSQL, SQL	MongoDB	Redis	Database grafici	Amazon S3, concetti vettoriali di base

Servizio gestito — Amazon Bedrock Knowledge Bases

Amazon Bedrock Knowledge Bases offre una soluzione completamente gestita con diverse opzioni di storage vettoriale. La tabella seguente confronta queste opzioni di storage.

Funzionalità	Immagine vettoriale e Aurora Poster SQLwith	Analisi di Neptune	OpenSearch Servizio serverless	Vettori Amazon S3	Pigna	RedisEnterprise Cloud
Caso d'uso principale	DB relazionale con RAG vettoriale	Ricerca vettoriale basata su grafici per GraphRag	Gestione della conoscenza RAG	RAG vettoriale ottimizzato in termini di costi	Ricerca vettoriale ad alte prestazioni	Ricerca vettoriale in memoria
Architecture	Relazionale completamente gestito	Analisi grafica completamente gestita	Completamente gestita senza server	Storage di oggetti senza server	Cloud ibrido completamente gestito	In-memory completamente gestito
Modello di dati	Tabelle relazionali	Grafico delle proprietà	Documenti JSON	Archiviazione di oggetti	Vettori creati appositamente	Valore-chiave con vettori
Archiviazione vettoriale	Tramite l'estensi	Vettori grafici nativi	Tramite motore	Storage vettoriale	Database vettoriale nativo	Vettori in memoria

	one pgvector		OpenSearch	Amazon S3 nativo		
Integrazione con Amazon Bedrock	Nativo	Nativo	Nativo	Nativo	Nativo	Nativo
Ingestione automatica	Sì (tramite Amazon Bedrock)	Sì (tramite Amazon Bedrock)	Sì (tramite Amazon Bedrock)	Sì (tramite Amazon Bedrock)	Sì (tramite Amazon Bedrock)	Sì (tramite Amazon Bedrock)
Vettorizzazione automatica	Sì (tramite Amazon Bedrock)	Sì (tramite Amazon Bedrock)	Sì (tramite Amazon Bedrock)	Sì (tramite Amazon Bedrock)	Sì (tramite Amazon Bedrock)	Sì (tramite Amazon Bedrock)
Dimensionamento	Scalabilità automatica (Aurora Serverless)	Ridimensionamento automatico del grafico	Automatico senza server	Automatico (miliardi di vettori)	Pod con ridimensionamento automatico	Cluster con scalabilità automatica
Prestazioni delle query	Alto per valori relazionali o vettoriali	Alto per i vettori grafici	Elevata	Medio (latenza pari o superiore a 100 ms)	Molto alto	Molto alto
Vettori massimi	Milioni (a seconda dell'istanza)	Miliardi	Miliardi	2 miliardi per indice	Miliardi	Milioni (dipende dalla memoria)
Funzionalità aggiuntive	Interrogazioni SQL, transazioni ACID	Algoritmi grafici, traversi	Ricerca nel testo completo, analisi	Ciclo di vita di Amazon S3, suddivisione in più livelli	Filtraggio dei metadati, namespace	Strutture dati Redis, memorizzazione nella cache

Ottimizzazione dei costi	Moderato (Aurora Serverless)	Moderato (unità di capacità)	Alto (senza server, pay-per-use)	Molto alto (fino al 90% di risparmio)	Moderato (prezzi basati su pod)	Basso (prezzo in memoria premium)
Ideale per	Carichi di lavoro SQL/vector ibridi	Grafici della conoscenza connessi	Testo completo con ricerca vettoriale	Vettori a lungo termine ad accesso raro	Ricerca vettoriale in tempo reale su larga scala	Esigenze di latenza estremamente basse
Modello di interrogazione ideale	Interrogazioni SQL e vettoriali miste	Traverse grafiche con vettori	Ricerche frequenti con analisi	Recupero poco frequente (da minuti a ore)	Interrogazioni in tempo reale ad alta frequenza	Milioni di richieste al secondo
Configurazione con Amazon Bedrock	Semplice (gestito da Amazon Bedrock)	Semplice (gestito da Amazon Bedrock)	Semplice (gestito da Amazon Bedrock)	Semplice (gestito da Amazon Bedrock)	Semplice (gestito da Amazon Bedrock)	Semplice (gestito da Amazon Bedrock)
Residenza dei dati	Regioni AWS	Regioni AWS	Regioni AWS	Regioni AWS	Multi-cloud (AWS e altri)	Multi-cloud (AWS e altri)
Modello tariffario	Orari e spazio di archiviazione delle istanze	Unità di capacità e storage	Elaborazione e archiviazione (senza server)	Archiviazione, interrogazioni e trasferimento	Orari e spazio di archiviazione del pod	Ore e spazio di archiviazione dei nodi

Scelta tra opzioni individuali e gestite

Considerazione	Scegli DB vettoriali individuali	Scegli Amazon Bedrock Knowledge Base (gestite)
Implementazione RAG	Vuoi il pieno controllo sulla pipeline RAG	Volete un RAG completam ente gestito con una configura zione minima
Personalizzazione	Sono necessarie una logica di recupero e una preelabor azione personalizzate	I pattern RAG standard soddisfano le tue esigenze
Infrastruttura esistente	Il database è già distribuito	Stai iniziando da zero o desideri una gestione semplific ata
Esperienza del team	Il tuo team ha esperienz a nell'amministrazione dei database	Preferisci concentrarti sulla logica delle applicazioni, non sull'infrastruttura
Complessità dell'integrazione	È necessaria una profonda integrazione con i sistemi esistenti	Desideri un'integrazione rapida con i modelli Amazon Bedrock
Sovraccarico operativo	È possibile gestire le operazioni del database	Vuoi AWS gestire le operazion i
Struttura dei costi	Preferisci la tariffazione diretta dei database	Preferisci i prezzi unificati di Amazon Bedrock
È ora di commercializzare	Hai tempo per l'impleme ntazione personalizzata	Hai bisogno di un'implem entazione rapida

Confronti e considerazioni sui costi

Comprendere la struttura dei costi delle diverse opzioni di database vettoriali è essenziale per prendere decisioni di implementazione informate. La tabella seguente riporta alcune considerazioni chiave sui costi per varie soluzioni di database vettoriali, inclusi database individuali e servizi gestiti. Ogni opzione presenta fattori di prezzo distinti che possono influire sul costo totale di proprietà (TCO), dai pay-as-you-go modelli all'infrastruttura e ai costi operativi.

Database vettoriale	Modello di costo	Considerazioni sui costi
Amazon Kendra	Pagamento in base al consumo, in base alle domande	I costi possono variare in base al numero di query e alla quantità di dati indicizza ti. Potrebbero essere applicati costi aggiuntivi per l'archivi azione e il trasferimento dei dati. Per ulteriori informazioni, consulta i prezzi di Amazon Kendra.
OpenSearch Servizio Amazon	Pagamento in base al consumo, in base alle ore di utilizzo delle istanze e allo spazio di archiviazione	I costi includono ore di istanza, storage (volumi Amazon EBS), trasferimento dati e UltraWarm storage opzionale . Le istanze riservate possono offrire risparmi fino al 30%. Le istanze accelerate da GPU offrono un miglior rapporto prezzo/prestazioni per carichi di lavoro vettoriali. Per ulteriori informazioni, consulta i prezzi OpenSearch di Amazon Service.
Open source OpenSearch	Open source, senza costi diretti (non è necessario pagare per scaricare o utilizzar	I costi includono l'infrastruttura (come server e storage) e i costi operativi (come

	e il software e non sono previsti costi di licenza)	manutenzione e monitoraggio). Le organizzazioni devono disporre di un budget per il personale necessario alla gestione e alla manutenzione dell'infrastruttura.
Amazon RDS per PostgreSQL con pgvector	Pagamento in base al consumo, in base all'utilizzo	I costi includono i tipi di istanze di database, lo storage, il trasferimento dei dati e i backup. Oltre il piano AWS gratuito, possono essere applicati costi aggiuntivi per il trasferimento dei dati, i tipi di istanze e lo storage. Per ulteriori informazioni, consulta Prezzi di Amazon RDS.
Amazon DocumentDB	Pagamento in base al consumo, in base alle ore di utilizzo dell'istanza e allo spazio di archiviazione	I costi includono ore di istanza, archiviazione (GB al mese), I/O richieste, archiviazione di backup e trasferimento dei dati. I cluster elastici consentono la scalabilità dinamica. Istanze riservate disponibili per l'ottimizzazione dei costi. Per ulteriori informazioni, consulta i prezzi di Amazon DocumentDB.

Amazon MemoryDB	Pagamento in base al consumo, in base alle ore dei nodi e allo storage dei dati	I costi includono le ore dei nodi (per tipo di nodo), l'archivi azione dei dati (GB all'ora), l'archiviazione delle istantane e e il trasferimento dei dati. I nodi riservati possono offrire risparmi fino al 55%. Ottimizza to per carichi di lavoro ad alta velocità e bassa latenza. Per ulteriori informazioni, consulta i prezzi di Amazon MemoryDB.
Analisi di Amazon Neptune	Pagamento in base al consumo, in base alle unità di capacità	I costi includono le unità di capacità Neptune NCUs (), lo storage (GB al mese) e il trasferimento dei dati. Scalabili tà automatica basata sul carico di lavoro senza impegni iniziali. È richiesto un minimo di 128. NCUs Per ulteriori informazioni, consulta i prezzi di Amazon Neptune.
Amazon S3 Vectors	Pagamento in base al consumo, in base allo spazio di archiviazione e alle richieste	I costi includono lo storage (GB al mese), le richieste PUT e GET, la gestione degli indici vettoriali e il trasferimento dei dati. Offre un risparmio sui costi fino al 90% rispetto ai database vettoriali specializ zati. Amazon S3 Intelligent-Tiering e policy del ciclo di vita disponibili per un'ulteriore ottimizzazione. Per ulteriori informazioni, consulta Prezzi di Amazon S3.

Basi di conoscenza di Amazon Bedrock	Pagamento in base al consumo, in base all'utilizzo	I costi possono variare in base all'utilizzo della knowledge base e di servizi aggiuntivi come Amazon OpenSearch Serverless. Potrebbero essere applicati costi aggiuntivi per l'archiviazione e il trasferim ento dei dati e funzional ità aggiuntive. Per ulteriori informazioni sui prezzi, consulta i prezzi OpenSearch di Amazon Service.

Casi d'uso di database vettoriali

Gli esempi seguenti evidenziano come diverse opzioni di database vettoriali possano essere utilizzate efficacemente per migliorare la gestione delle conoscenze, migliorare l'efficienza operativa e fornire migliori risultati di business. Questi casi d'uso illustrano le applicazioni pratiche delle soluzioni di database vettoriali discusse in precedenza in questa guida e forniscono informazioni sulle loro prestazioni e sui vantaggi nel mondo reale.

Gestione della conoscenza con Amazon Kendra

Problema del cliente: uno dei maggiori appaltatori generali del Giappone stava affrontando un calo di personale esperto. L'azienda aveva bisogno di un modo per trasferire in modo efficiente le conoscenze e le competenze del personale esperto alle nuove generazioni. Avevano bisogno di una soluzione per acquisire e diffondere complesse conoscenze di ingegneria edile ed esperienze passate.

AWS soluzione — Per risolvere questo problema, il cliente si è rivolto ad Amazon Kendra, una soluzione di intelligenza artificiale in grado di gestire in modo rapido e preciso la propria base di conoscenze interna e consentire query in linguaggio naturale. Con Amazon Kendra, i dipendenti possono ora trovare le informazioni di cui hanno bisogno molto più velocemente, migliorando la produttività e facilitando il trasferimento delle conoscenze da personale esperto a personale più giovane.

Impatto: implementando un chatbot di intelligenza artificiale generativo basato su Amazon Kendra, l'azienda ha creato una piattaforma di conoscenza unificata. Il chatbot consente ai dipendenti di accedere rapidamente alle conoscenze tecniche e alle esperienze passate in materia di ingegneria edile. Questa soluzione ha notevolmente migliorato l'efficienza del trasferimento delle conoscenze e dei processi decisionali all'interno dell'organizzazione, contribuendo a garantire che le preziose competenze siano preservate e facilmente accessibili a tutti i dipendenti. Il costo di questa soluzione può variare in base all'utilizzo e alla configurazione. Per una stima dettagliata dei costi, consulta il Calcolatore dei prezzi AWS. Per la stima dei costi dei database vettoriali, consulta la sezione Confronto dei costi e considerazioni di questa guida o consulta i prezzi di Amazon Kendra.

Per informazioni su altri casi d'uso dei clienti, consulta Clienti Amazon Kendra.

Analisi in tempo reale con Serverless OpenSearch

Problema del cliente: un importante fornitore di servizi finanziari ha dovuto affrontare la sfida di gestire un enorme ecosistema di dati. Ha elaborato 300 milioni di autorizzazioni e 90 miliardi di transazioni all'anno, per un totale di circa 1,1 petabyte (PB) di dati. Il sistema esistente, che serviva 300.000 utenti che richiedevano l'accesso a oltre 6.000 report, necessitava di una modernizzazione per fornire coerenza globale e consentire il processo decisionale in tempo reale.

AWS soluzione: l'architettura della soluzione utilizzava modelli di base disponibili tramite Amazon Bedrock (inclusi Anthropic, Sonnet 3, Sonnet 3.5 e Haiku) per l'elaborazione del linguaggio naturale. Il cliente ha scelto OpenSearch Serverless come database vettoriale per la sua scalabilità superiore e la capacità di gestire l'enorme volume di dati in modo efficiente. Questa architettura ha consentito l'elaborazione senza interruzioni di query complesse e la generazione dinamica di report.

Impatto: l'implementazione ha consentito un aumento della produttività del 50% eliminando la necessità di generare manualmente oltre 100 dashboard di business intelligence. Gli utenti possono ora generare report tramite query in linguaggio naturale con tempi di risposta compresi tra 20 e 40 secondi. Il costo di questa soluzione può variare a seconda dell'utilizzo e della configurazione. Per una stima dettagliata dei costi, consulta il Calcolatore dei prezzi AWS. Per la stima dei costi dei database vettoriali, consulta la sezione Confronto dei costi e considerazioni di questa guida o consulta i prezzi di Amazon OpenSearch Service. Per informazioni su altri casi d'uso dei clienti, consulta Amazon OpenSearch Serverless.

Risorse e passaggi successivi

Dopo aver esaminato questa guida, considera le seguenti azioni per passare dalla comprensione all'implementazione:

1. Valuta le tue esigenze attuali:
   - Valuta l'infrastruttura e le competenze di database esistenti.
   - Documenta i tuoi requisiti specifici di ricerca vettoriale.
   - Definisci i tuoi obiettivi di prestazioni, scalabilità e costi.
2. Scegliete una delle seguenti opzioni per testare le opzioni del database vettoriale:
   - Opzione 1: impostate un proof of concept utilizzando la vostra soluzione di database vettoriale preferita.
   - Opzione 2: sperimenta con set di dati di esempio in Amazon Bedrock Knowledge Bases. Prova l'esperienza di creazione rapida per una Knowledge Base di Amazon Bedrock. Per un esempio, consulta Quick create an Aurora PostgreSQL Knowledge Base per Amazon Bedrock nella documentazione di Aurora.
3. Consulta le risorse aggiuntive.
4. Fatti aiutare da un esperto:
   - Contatta il tuo Account AWS team o i AWS Solutions Architects per una guida all'implementazione.
   - Interagisci con AWS partner specializzati in database vettoriali.
5. Pianifica la tua implementazione in produzione:
   - Crea una strategia di migrazione se passi da database esistenti.
   - Sviluppa un piano di scalabilità per la soluzione prescelta.
   - Progetta le tue procedure di monitoraggio e manutenzione.

Resources

Le seguenti risorse possono aiutarti nella scelta di un database vettoriale.

AWS post sul blog

- Accelera lo sviluppo di applicazioni di intelligenza artificiale generativa con Amazon Bedrock Knowledge Bases, Quick Create e Amazon Aurora Serverless.
- Spiegazione delle funzionalità del database vettoriale di Amazon OpenSearch Service
- Approfondisci gli archivi di dati vettoriali utilizzando Amazon Bedrock Knowledge Bases
- Sfrutta pgvector e Amazon Aurora PostgreSQL per l'elaborazione del linguaggio naturale, i chatbot e l'analisi dei sentimenti

AWS documentazione di servizio

- Scelta di un servizio di AWS database
- Come funzionano le knowledge base di Amazon Bedrock
- Documentazione di Neptune Analytics
- Panoramica di Amazon Web Services: database
- Utilizzo di Aurora PostgreSQL come Knowledge Base per Amazon Bedrock
- Lavorare con Amazon Aurora PostgreSQL
- Amazon DocumentDB
- Amazon MemoryDB
- Amazon S3 Vectors

Altre risorse AWS

- Basi di conoscenza di Amazon Bedrock
- Database vettoriali e incorporamenti
- Database vettoriali per applicazioni di intelligenza artificiale generativa
- Cosa sono gli Embeddings nel Machine Learning?

Altre risorse

- Informazioni su PostgreSQL
- documentazione pgvector

- Pinecone come base di conoscenza per Amazon Bedrock
- Redis Enterprise Cloud su AWS

Cronologia dei documenti

La tabella seguente descrive le modifiche significative a questa guida, Scelta di un AWS database vettoriale per i casi d'uso RAG. Per ricevere notifiche sugli aggiornamenti futuri, puoi abbonarti a un feed RSS.

Modifica	Descrizione	Data
Aggiunto Servizi AWS	Abbiamo aggiunto informazioni sull'utilizzo di Amazon DocumentDB, Amazon MemoryDB, Amazon S3 Vectors e Amazon Neptune Analytics.	30 marzo 2026
Pubblicazione iniziale	—	6 marzo 2025

Le traduzioni sono generate tramite traduzione automatica. In caso di conflitto tra il contenuto di una traduzione e la versione originale in Inglese, quest'ultima prevarrà.