Choosing an AWS vector database for RAG use cases

AWS Prescriptive Guidance

AWS Prescriptive Guidance: Choosing an AWS vector database for RAG use cases

Table of Contents

Choosing an AWS vector database for RAG use cases

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

March 2026 (document history)

Vector databases are becoming increasingly important for organizations that implement generative AI applications. These databases store and manage *vectors*, which are numerical representations of data that enable processing of text, images, and other content in ways that capture their meaning and relationships.

As organizations explore vector database options on AWS, they need to understand the capabilities, trade-offs, and best practices for different solutions. This guide helps you compare commonly used vector stores on AWS and make informed decisions about which options best suit your specific needs or use case. Whether you're implementing Retrieval Augmented Generation (RAG), building recommendation systems, or developing other AI applications, this guide provides a framework to help you evaluate and choose a vector database solution.

Intended audience

This guide is intended for people in the following roles:

- Data scientists and machine learning (ML) engineers who use vector databases to store and retrieve high-dimensional data for ML models.
- Data engineers who design and implement data pipelines that include vector databases for storing and processing high-dimensional data.
- MLOps engineers who use vector databases as part of the ML pipeline to store and serve model outputs or intermediate representations.
- Software engineers who integrate vector databases into applications that require similarity search or recommendation systems.
- DevOps engineers who are responsible for deploying and maintaining vector databases in production environments, ensuring scalability and reliability.
- AI researchers who use vector databases to store and analyze large datasets of embeddings or feature vectors.
- AI product managers who need to understand the capabilities and limitations of vector databases to make informed decisions about product features and architecture.

Overview of vectors

Vectors are numerical representations that help machines understand and process data. In generative AI, they serve two key purposes:

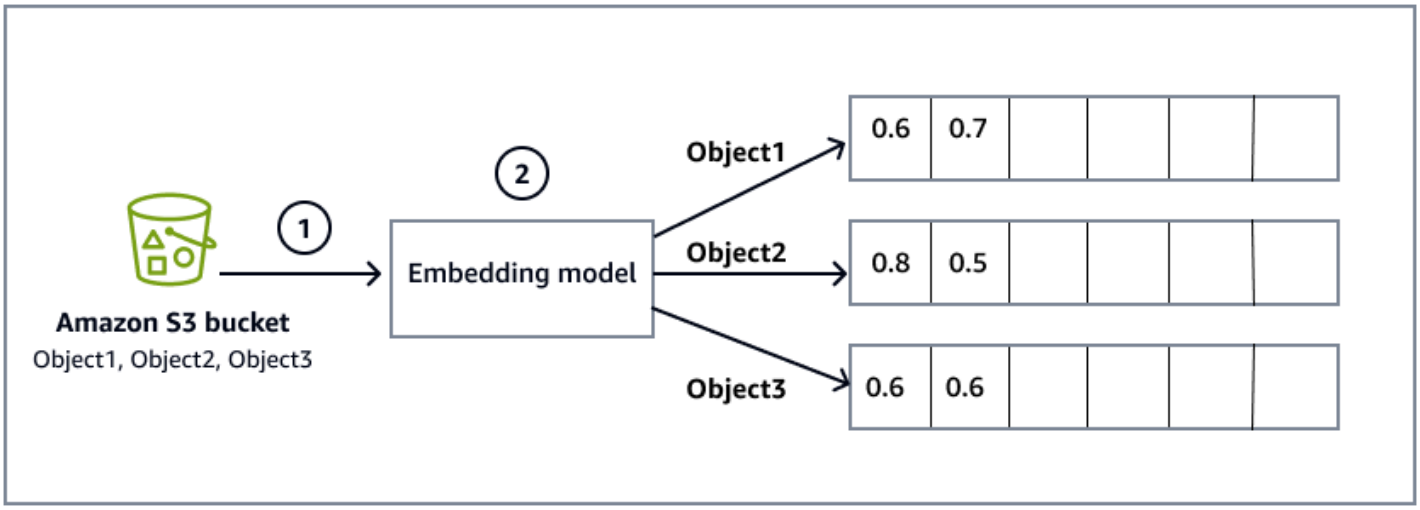
- Representing latent spaces that capture data structure in compressed form
- Creating embeddings for data, such as words, sentences, and images

Embedding models like Word2Vec, GloVe, and Amazon Titan Text Embeddings convert data into vectors through a process called *embedding*. These embedding models can do the following:

- Learn from context to represent words as vectors
- Place similar words closer together in vector space
- Enable machines to process data in a continuous space

The following diagram provides a high-level overview of the embedding process:

1. An Amazon Simple Storage Service (Amazon S3) bucket contains files that are the data sources from which the system will read and process information. The Amazon S3 bucket is specified during the Amazon Bedrock knowledge base configuration, which also includes syncing data with the knowledge base.
2. The embedding model converts the raw data from the object files in the Amazon S3 bucket into vector embeddings. For example, `Object1` is converted into a vector `[0.6, 0.7, ...]` that represents its content in a multi-dimensional space.



Word embeddings are crucial for natural language processing (NLP) because they do the following:

- Capture semantic relationships between words
- Enable generation of contextually relevant text
- Power large language models (LLMs) to produce human-like responses

Overview of vector databases

A vector database is a specialized system that stores and queries high-dimensional vectors efficiently. These databases are fundamental for Retrieval Augmented Generation (RAG) applications.

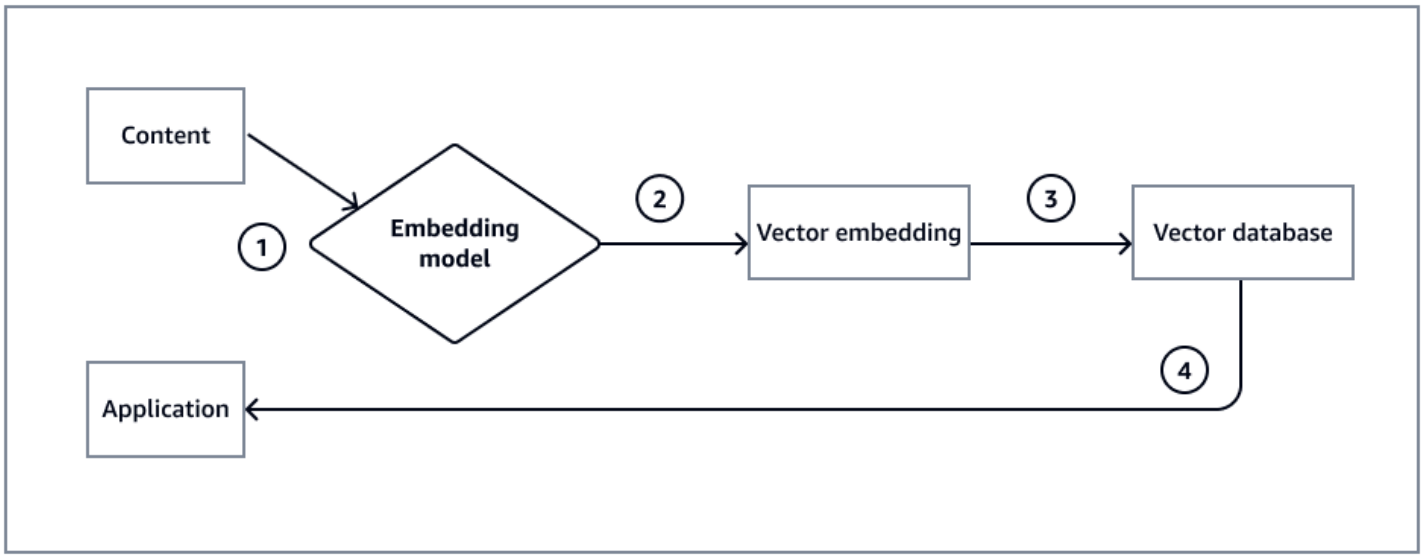
Vector databases handle data conversion and storage in the following ways:

- Objects (such as audio, images, and text files) are converted to vectors by using embedding models.
- Vectors are stored in specialized data formats.
- Vector databases enable rapid similarity searches.

Vector databases offer several key advantages over traditional databases, making them particularly well-suited for modern data challenges. They are specifically optimized for vector operations and handle high-dimensional data efficiently. They also specialize in similarity searches that traditional databases struggle with. Beyond these core capabilities, vector databases are built to meet the evolving demands of ML and generative AI applications. They excel at large-scale vector storage and use distributed computing to balance workloads across multiple nodes. This provides scalability and performance as data volumes grow.

The following diagram shows a RAG implementation:

1. Content, such as documents, PDFs, or text files, is fed into the embedding model as raw data for processing.
2. The embedding model transforms the raw data into numerical vectors, which represent the semantic meaning of the content.
3. The generated vector embeddings are stored in a vector database that is optimized for the storage and retrieval of high-dimensional vectors.
4. Applications can now query the vector database in response to use cases such as semantic search and content recommendation.



Choosing an inappropriate vector database for a RAG solution can lead to significant struggles and limitations including the following:

- Poor query performance
- Scalability bottlenecks
- Data ingestion challenges
- Lack of advanced features, such as filtering and ranking
- Integration difficulties with other systems
- Persistence and durability concerns
- Concurrency and consistency issues in environments with multiple users
- Higher licensing costs or vendor lock-in
- Limited community support and resources
- Potential security and compliance risks

Vector database options

AWS offers a diverse range of vector database solutions to support different use cases and requirements in generative AI applications. These options can be broadly categorized into individual database services and managed service offerings, each with distinct characteristics and advantages. Understanding these options is crucial for organizations looking to implement vector search capabilities effectively while maintaining optimal performance, scalability, and cost efficiency.

For more information about vector database solutions, see the following sections:

- Individual vector database options
- Managed service option
- Choosing the right vector database

Individual vector database options

The individual vector database options on AWS include Amazon Kendra, Amazon OpenSearch Service, Amazon RDS for PostgreSQL with pgvector, Amazon MemoryDB, Amazon DocumentDB, Amazon Neptune Analytics, and Amazon S3 Vector. (An open-source extension, pgvector adds the ability to store and search ML-generated vector embeddings.) These solutions offer different approaches to vector search, allowing organizations to choose based on their existing infrastructure, technical requirements, and specific use cases.

Amazon Kendra

Amazon Kendra is an enterprise-grade intelligent search service that uses natural language processing and advanced machine learning algorithms to return specific answers to search questions from your data. Amazon Kendra simplifies the implementation of search functionality, making it an effective backend solution for generative AI applications.

Other key features of Amazon Kendra include the following:

- Native connections to over 40 data sources
- Built-in data preparation capabilities
- Quick setup that doesn't require deep technical expertise

Benefits of Amazon Kendra include the following

- Automated data processing (chunking, ingestion, retrieval)
- Powerful customization options:
  - Facet search
  - Search analytics
  - Tuning search relevance
- Simple programmatic access through the AWS SDK for Python (Boto3)

For more information, see Benefits of Amazon Kendra in the Amazon Kendra documentation.

Amazon OpenSearch Service

Amazon OpenSearch Service is a managed service that helps you deploy, operate, and scale OpenSearch Service clusters in the AWS Cloud.

Core capabilities of OpenSearch Service include the following:

- Open-source search and analytics engine
- Distributed architecture
- Real-time data processing

Some advantages of using OpenSearch Service include the following:

- Horizontal scalability
- RESTful API support
- Handles structured and unstructured data
- Real-time data analysis
- Suitable for various deployment sizes

For more information, see Features of Amazon OpenSearch Service in the OpenSearch Service documentation.

Amazon RDS for PostgreSQL with pgvector

Amazon RDS for PostgreSQL with pgvector combines the AWS managed relational database service with PostgreSQL's vector processing extension. This combination enables organizations to store and query high-dimensional vectors while maintaining Amazon RDS. The solution is particularly suitable for generative AI applications that require real-time vector operations without the overhead of managing database infrastructure.

Key benefits of Amazon RDS for PostgreSQL with pgvector include the following:

- High availability
- Automatic failover
- Cost-effective (pay-per-use)
- Built-in monitoring
- Real-time vector data integration

For more information, see Advantages of Amazon RDS in the Amazon RDS documentation.

Amazon DocumentDB

Amazon DocumentDB (with MongoDB compatibility) is a document database that offers native vector search capabilities in version 5.0 and later. It combines the flexibility of JSON-based document storage with vector search, supporting both hierarchical navigable small world (HNSW) and Inverted File Flat (IVFFlat) indexing methods.

Core capabilities of Amazon DocumentDB include the following:

- Store and index vectors up to 2,000 dimensions (up to 16,000 dimensions without indexing)
- Millisecond response times for vector similarity searches
- Support for euclidean, cosine, and dot product distance metrics
- Seamless integration with existing MongoDB-compatible applications

Use Amazon DocumentDB in the following situations:

- For applications that are already using MongoDB APIs and that need vector search capabilities
- For use cases that require flexible document data structures combined with semantic search

- For scenarios that need both traditional document queries and vector similarity searches
- For applications that provide product recommendations, personalization, chat assistants, and fraud detection

For more information, see Vector search for Amazon DocumentDB in the Amazon DocumentDB documentation.

Amazon MemoryDB

Amazon MemoryDB is a Redis-compatible, in-memory database that delivers the fastest vector search performance among popular vector databases on AWS. It provides sub-millisecond query latencies with multi-Availability Zone durability.

Core capabilities of MemoryDB include the following:

- Store application data and millions of vectors in a single database
- Single-digit millisecond query and update response times
- Highest recall rates at the fastest performance on AWS
- Support for up to 32,768 dimensions per vector
- Real-time semantic search and caching capabilities

Use MemoryDB in the following situations:

- For real-time applications that require ultra-low latency (sub-10ms)
- For high-throughput workloads with millions of requests per day
- For use cases such as real-time recommendation engines, semantic caching, and anomaly detection
- For applications that need both in-memory data store and vector search capabilities

For more information, see Vector search in the MemoryDB documentation.

Amazon Neptune Analytics

Amazon Neptune Analytics is a graph analytics engine that offers native vector search capabilities, making it ideal for Graph Retrieval Augmented Generation (GraphRAG) use cases. It combines vector similarity search with graph traversals and algorithms.

Core capabilities of Neptune Analytics include the following:

- Analyze tens of billions of relationships within seconds
- Combine vector search with graph algorithms (path finding, community detection, centrality)
- Support for GraphRAG applications with topological knowledge
- Up to 80 times faster than existing graph analytical solutions
- Integration with Amazon Bedrock for fully managed GraphRAG

Use Neptune Analytics in the following situations:

- For GraphRAG applications that require knowledge graphs with vector embeddings
- For use cases that require traversing complex relationships alongside vector similarity
- For applications that require explainable AI responses with relationship context
- For scenarios such as customer 360 views, fraud detection networks, and knowledge discovery

For more information, see the Amazon Neptune Analytics documentation.

Amazon S3 Vectors

Amazon S3 Vectors is the first cloud object store in AWS with native vector storage and query capabilities. It provides purpose-built, cost-optimized vector storage for AI applications that require massive scale.

Core capabilities of Amazon S3 Vectors include the following:

- Storage for up to 2 billion vectors per index with support for up to 10,000 indexes per vector bucket
- Sub-100 ms query latency that is optimized for long-term storage and infrequent access patterns
- Up to 90% cost reduction for vector operations compared to specialized vector databases
- Serverless architecture with automatic scaling and 99.999999999% (11 9s) durability

Use Amazon S3 Vectors in the following situations:

- For applications that require storage of billions of vectors at minimal cost

- For workloads that tolerate sub-second query latency (100 ms or more) rather than sub-10 ms
- For long-term vector retention and archival use cases
- For RAG applications with infrequent retrieval patterns
- For organizations that prioritize storage economics over ultra-low latency

Amazon S3 Vectors integrates natively with Amazon Bedrock Knowledge Bases and works well in tiered architectures with Amazon OpenSearch Service. You can use Amazon S3 Vectors for cold storage and use OpenSearch Service for hot queries.

For more information, see Working with S3 Vectors and vector buckets in the Amazon S3 documentation.

Managed service option

Amazon Bedrock Knowledge Bases represents the AWS fully managed approach to vector database implementation. The service's flexibility in storage options, combined with its automated management features, makes it particularly valuable for organizations seeking to implement RAG without managing complex infrastructure.

With Amazon Bedrock Knowledge Bases, you can create, maintain, and query knowledge bases that enhance your foundation models using RAG. This service simplifies the complex process of implementing RAG by managing the entire data ingestion, vectorization, and retrieval pipeline.

Key benefits of Amazon Bedrock Knowledge Bases include the following:

- Simplified data processing
  - Automatic data ingestion and chunking
  - Built-in text extraction from multiple file formats
  - Managed vector embeddings generation
  - Automatic metadata extraction and indexing
- Streamlined RAG implementation
  - Pre-configured retrieval strategies
  - Automatic context window optimization
  - Built-in relevancy tuning
  - Semantic search capabilities out of the box

- Security and governance
  - Integrated AWS Identity and Access Management (IAM) controls
  - Data encryption at rest and in transit
  - VPC support
  - Audit logging with AWS CloudTrail

Amazon Bedrock Knowledge Bases supports multiple vector store options, including:

- Amazon Aurora PostgreSQL with pgvector
- Amazon Neptune Analytics
- Amazon EMR Serverless
- Amazon S3 Vectors
- Pinecone
- Redis Enterprise Cloud

This managed service handles automated data ingestion, vectorization, and retrieval. This simplifies RAG implementations.

For detailed information about each supported vector store, see the Amazon Bedrock Knowledge Bases documentation.

Choosing the right vector database

Select your vector database based on these key decision factors:

- **If you need MongoDB-compatible document database with vector search** – Choose Amazon DocumentDB. This is ideal when your application uses MongoDB APIs and you want to add semantic search capabilities without managing separate vector infrastructure.
- **If you need ultra-low latency for real-time applications** – Choose Amazon MemoryDB. This provides the fastest vector search performance on AWS with sub-millisecond response times. It's ideal for real-time recommendation engines and high-throughput applications.
- **If you need graph-based knowledge representations with vector search** – Choose Amazon Neptune Analytics. This is best for GraphRAG applications that need to traverse complex relationships and perform graph-based queries alongside vector searches, providing explainable AI responses.

- **If you need to combine relational queries with vector search** – Choose Amazon Aurora PostgreSQL with pgvector. This option is ideal when your application requires both traditional SQL operations and vector similarity searches within the same database.
- **If you require high-throughput queries with sub-10 ms latency** – Choose Amazon OpenSearch Service. It excels at handling high-frequency queries and real-time applications and includes recent GPU acceleration improvements.
- **If you need to store billions of vectors cost-effectively** – Choose Amazon S3 Vectors. This option provides up to 90% cost savings and is ideal for applications with infrequent retrieval patterns (minutes to hours between queries) that can tolerate sub-100 ms latency.
- **If you need full-text search alongside vector search** – Choose Amazon OpenSearch Service. This option combines powerful full-text search capabilities with vector search in a single platform.

Vector database comparison

AWS provides multiple approaches to implementing vector search capabilities, ranging from individual vector databases to Amazon Bedrock Knowledge Bases, which is a fully managed service. When evaluating these options, organizations must consider various aspects including architecture, scalability, integration capabilities, performance characteristics, and security features.

Individual vector databases

The following table provides an overview of key features of several AWS individual vector database solutions, focusing on their architectures, scaling capabilities, data source integrations, and performance characteristics.

Feature	Amazon Kendra	Amazon OpenSearch Service	Amazon RDS for PostgreSQL with pgvector	Amazon DocumentDB	Amazon MemoryDB	Amazon Neptune Analytics	Amazon S3 Vectors
Primary use case	Enterprise search and RAG	Distributed search and analytics	Relational DB with vector support	DocumentDB with vector search	Real-time in-memory vector search	Graph analytics with vector search	Cost-optimized vector storage
Architecture	Fully managed	Distributed cluster	Relational database	Document-oriented	In-memory database	Graph analytics engine	Serverless object storage
Data model	Document-based	JSON documents	Relational tables	JSON documents	Key-value with JSON	Property graph	Object storage
Vector dimensions	Managed automatically	Up to 16,000	Configurable	Up to 2,000 (indexed)	Up to 32,768	Configurable	Up to 4,096

				; 16,000 (unindexed)			
Indexing methods	Automatic	HNSW, IVF	HNSW, IVFFlat	HNSW, IVFFlat	HNSW	Native graph and vector	Automatic
Distance metrics	Automatic	Cosine, Euclidean, dot product	Cosine, Euclidean, inner product	Cosine, Euclidean, dot product	Cosine, Euclidean, inner product	Cosine, Euclidean	Cosine, Euclidean
Query latency	Sub-second	Sub-10 ms (GPU-accelerated)	10-100 ms	Millisecond	Sub-millisecond	Sub-second	Sub-100 ms
Scaling model	Automatic	Horizontal (add nodes)	Vertical and read replicas	Horizontal (add instances)	Vertical and replicas	Automatic	Automatic (serverless)
Maximum vectors	Managed	Billions (cluster-dependent)	Millions (instance-dependent)	Millions per collection	Millions per database	Billions	2 billion per index; 10,000 indexes per bucket
Throughput	High	Very high (thousands of QPS)	Medium	High	Very high (millions of requests per day)	High	Medium (optimized for infrequent queries)

Data durability	99.999999999% (11 9s)	Configurable with replicas	99.99% (Multi-AZ)	99.99% (Multi-AZ)	99.99% (Multi-AZ)	99.99%	99.999999999% (11 9s)
Consistency model	Eventual	Eventual (configurable)	Strong (ACID)	Eventual	Strong	Strong	Strong
Additional capabilities	40 or more data connectors, NLP	Full-text search, analytics, dashboards	SQL queries, ACID transactions	MongoDB API compatibility	Redis API compatibility, caching	Graph algorithms, traversals	Amazon S3 integration, lifecycle policies
Pricing model	Pay per query and storage	Instance hours and storage	Instance hours and storage	Instance hours and storage	Instance hours and storage	Capacity units and storage	Storage, queries, and data transfer
Cost optimization	Usage-based	Reserved instances, auto-scaling	Reserved instances, Aurora Serverless	Reserved instances	Reserved instances	Auto-scaling	Up to 90% savings vs specialized DBs
Best for	Enterprise search with minimal setup	High-throughput, low-latency queries	Hybrid SQL and vector workloads	MongoDB-compatible apps needing vectors	Real-time, ultra-low latency apps	GraphRAG and knowledge graphs	Long-term, cost-effective storage

Ideal query pattern	Frequent enterprise searches	High-frequency real-time queries	Mixed SQL and vector queries	Document queries with semantic search	Millions of requests per day	Graph traversals with vector search	Infrequent queries (minutes to hours)
Setup complexity	Low (fully managed)	Medium (cluster configuration)	Medium (extension setup)	Medium (cluster configuration)	Medium (cluster configuration)	Low (fully managed)	Low (serverless)
Team expertise required	Minimal	OpenSearch or Elasticsearch	PostgreSQL, SQL	MongoDB	Redis	Graph databases	Amazon S3, basic vector concepts

Managed service – Amazon Bedrock Knowledge Bases

Amazon Bedrock Knowledge Bases provides a fully managed solution with multiple vector storage options. The following table compares these storage options.

Feature	**Aurora PostgreSQL with pgvector**	**Neptune Analytics**	**OpenSearch Service Serverless**	**Amazon S3 vectors**	**Pinecone**	**RedisEnterprise Cloud**
Primary use case	Relational DB with vector RAG	Graph-based vector search for GraphRAG	Knowledge management RAG	Cost-optimized vector RAG	High-performance vector search	In-memory vector search
Architecture	Fully managed relational	Fully managed graph analytics	Fully managed serverless	Serverless object storage	Fully managed hybrid cloud	Fully managed in-memory

Data model	Relational tables	Property graph	JSON documents	Object storage	Purpose-built vectors	Key-value with vectors
Vector storage	Through pgvector extension	Native graph vectors	Through OpenSearch engine	Native Amazon S3 vector storage	Native vector database	In-memory vectors
Amazon Bedrock integration	Native	Native	Native	Native	Native	Native
Automatic ingestion	Yes (via Amazon Bedrock)	Yes (via Amazon Bedrock)	Yes (via Amazon Bedrock)	Yes (via Amazon Bedrock)	Yes (via Amazon Bedrock)	Yes (via Amazon Bedrock)
Automatic vectorization	Yes (via Amazon Bedrock)	Yes (via Amazon Bedrock)	Yes (via Amazon Bedrock)	Yes (via Amazon Bedrock)	Yes (via Amazon Bedrock)	Yes (via Amazon Bedrock)
Scaling	Auto-scaling (Aurora Serverless)	Automatic graph scaling	Automatic serverless	Automatic (billions of vectors)	Auto-scaling pods	Auto-scaling clusters
Query performance	High for relational or vector	High for graph vectors	High	Medium (100 ms or more latency)	Very high	Very high
Maximum vectors	Millions (instance-dependent)	Billions	Billions	2 billion per index	Billions	Millions (memory-dependent)

Additional capabilities	SQL queries, ACID transactions	Graph algorithms, traversals	Full-text search, analytics	Amazon S3 lifecycle, tiering	Metadata filtering, namespaces	Redis data structures, caching
Cost optimization	Moderate (Aurora Serverless)	Moderate (capacity units)	High (serverless, pay-per-use)	Very high (up to 90% savings)	Moderate (pod-based pricing)	Low (in-memory premium)
Best for	Hybrid SQL/ vector workloads	Connected knowledge graphs	Full-text with vector search	Long-term, infrequent-access vectors	Real-time vector search at scale	Ultra-low latency needs
Ideal query pattern	Mixed SQL and vector queries	Graph traversals with vectors	Frequent searches with analytics	Infrequent retrieval (minutes to hours)	High-frequency real-time queries	Millions of requests per second
Setup with Amazon Bedrock	Simple (managed by Amazon Bedrock)	Simple (managed by Amazon Bedrock)	Simple (managed by Amazon Bedrock)	Simple (managed by Amazon Bedrock)	Simple (managed by Amazon Bedrock)	Simple (managed by Amazon Bedrock)
Data residency	AWS Regions	AWS Regions	AWS Regions	AWS Regions	Multi-cloud (AWS and others)	Multi-cloud (AWS and others)
Pricing model	Instance hours and storage	Capacity units and storage	Compute and storage (serverless)	Storage, queries, and transfer	Pod hours and storage	Node hours and storage

Choosing between individual and managed options

Consideration	Choose individual vector DB	Choose Amazon Bedrock Knowledge Bases (managed)
RAG implementation	You want full control over RAG pipeline	You want fully managed RAG with minimal setup
Customization	You need custom retrieval logic and preprocessing	Standard RAG patterns meet your needs
Existing infrastructure	You already have the database deployed	You're starting fresh or want simplified management
Team expertise	Your team has database administration expertise	You prefer to focus on application logic, not infrastru cture
Integration complexity	You need deep integration with existing systems	You want quick integration with Amazon Bedrock models
Operational overhead	You can manage database operations	You want AWS to handle operations
Cost structure	You prefer direct database pricing	You prefer unified Amazon Bedrock pricing
Time to market	You have time for custom implementation	You need rapid deployment

Cost comparisons and considerations

Understanding the cost structure of different vector database options is essential for making informed implementation decisions. The following table outlines some key cost considerations for various vector database solutions, including both individual databases and managed services. Each option has distinct pricing factors that can impact total cost of ownership (TCO), from pay-as-you-go models to infrastructure and operational costs.

Vector database	Cost model	Cost considerations
Amazon Kendra	Pay as you go, based on queries	Costs can vary based on the number of queries and the amount of data indexed. Additional charges can apply for data storage and data transfer. For more informati on, see Amazon Kendra pricing.
Amazon OpenSearch Service	Pay as you go, based on instance hours and storage	Costs include instance hours, storage (Amazon EBS volumes), data transfer, and optional UltraWarm storage. Reserved Instances can provide up to 30% savings. GPU-accelerated instances offer better price-performance for vector workloads. For more information, see Amazon OpenSearch Service pricing.
Open-source OpenSearch	Open-source, no direct cost (you don't have to pay to download or use the software, and there are no license costs)	Costs include infrastructure (such as servers and storage) and operational costs (such as maintenance and monitoring). Organizations need to budget

		for personnel to manage and maintain the infrastructure.
Amazon RDS for PostgreSQL with pgvector	Pay as you go, based on usage	Costs include database instance types, storage, data transfer, and backups. Additional charges can apply for data transfer, instance types, and storage beyond the AWS Free Tier. For more information, see Amazon RDS pricing.
Amazon DocumentDB	Pay as you go, based on instance hours and storage	Costs include instance hours, storage (GB-month), I/O requests, backup storage, and data transfer. Elastic clusters enable dynamic scaling. Reserved Instances available for cost optimizat ion. For more information, see Amazon DocumentDB pricing.
Amazon MemoryDB	Pay as you go, based on node hours and data storage	Costs include node hours (per node type), data storage (GB-hour), snapshot storage, and data transfer. Reserved Nodes can provide up to 55% savings. Optimized for high-throughput, low-latency workloads. For more information, see Amazon MemoryDB pricing.

Amazon Neptune Analytics	Pay as you go, based on capacity units	Costs include Neptune Capacity Units (NCUs), storage (GB-month), and data transfer. Auto-scaling based on workload with no upfront commitments. Minimum 128 NCUs required. For more information, see Amazon Neptune pricing.
Amazon S3 Vectors	Pay as you go, based on storage and requests	Costs include storage (GB-month), PUT and GET requests, vector index management, and data transfer. Provides up to 90% cost savings compared to specialized vector databases. Amazon S3 Intelligent-Tiering and lifecycle policies available for additional optimization. For more information, see Amazon S3 pricing.
Amazon Bedrock Knowledge Bases	Pay as you go, based on usage	Costs can vary based on the usage of the knowledge base and additional services such as Amazon OpenSearch Serverless. Additional charges can apply for data storage, data transfer, and additional features. For more informati on about pricing, see Amazon OpenSearch Service pricing.

Vector database use cases

The following examples highlight how different vector database options can be used effectively to enhance knowledge management, improve operational efficiency, and deliver better business outcomes. These use cases illustrate practical applications of the vector database solutions discussed earlier in this guide and provide insights into their real-world performance and benefits.

Knowledge management with Amazon Kendra

Customer problem – One of the largest general contractors in Japan was facing a decline in experienced personnel. The company needed a way to transfer the knowledge and skills of the experience personnel to the younger generation efficiently. They required a solution to capture and disseminate complex construction engineering knowledge and past experiences.

AWS solution – To address this problem, the customer turned to Amazon Kendra, an AI solution that could quickly and accurately handle their internal knowledge base and allow natural language queries. With Amazon Kendra, employees can now find the information they need much faster, improving productivity and facilitating knowledge transfer from experienced personnel to younger staff.

Impact – By implementing a generative AI chatbot powered by Amazon Kendra, the company created a unified knowledge platform. The chatbot allows employees to quickly access technical knowledge and past experiences on construction engineering. This solution has significantly improved the efficiency of knowledge transfer and decision-making processes within the organization, helping to ensure that valuable expertise is preserved and easily accessible to all employees. The cost of this solution may vary depending on your usage and configuration. For a detailed cost estimate, see the AWS Pricing Calculator. For vector database cost estimation, see the Cost comparison and considerationssection of this guide or see Amazon Kendra pricing.

For information about other customer use cases, see Amazon Kendra customers.

Real-time analytics with OpenSearch Serverless

Customer problem – A leading financial services provider faced the challenge of managing an enormous data ecosystem. It processed 300 million authorizations and 90 billion transactions annually, accumulating to approximately 1.1 petabytes (PB) of data. The existing system, serving

300,000 users who required access to over 6,000 reports, needed modernization to provide global consistency and enable real-time decision-making.

AWS solution – The solution architecture used foundation models available through Amazon Bedrock (including Anthropic, Sonnet 3, Sonnet 3.5, and Haiku) for natural language processing. The customer chose OpenSearch Serverless as the vector database for its superior scalability and ability to handle the massive data volume efficiently. This architecture enabled seamless processing of complex queries and dynamic report generation.

Impact – The implementation achieved a 50 percent increase in productivity by eliminating the need for manual generation of over 100 business intelligence dashboards. Users can now generate reports through natural language queries with response times of between 20-40 seconds. The cost of this solution may vary depending on your usage and configuration. For a detailed cost estimate, see the AWS Pricing Calculator. For vector database cost estimation, see the Cost comparison and considerationssection of this guide or see Amazon OpenSearch Service pricing. For information about other customer use cases, see Amazon OpenSearch Serverless.

Next steps and resources

After reviewing this guide, consider the following actions to move from understanding to implementation:

1. Evaluate your current needs:
   - Assess your existing database infrastructure and expertise.
   - Document your specific vector search requirements.
   - Define your performance, scaling, and cost targets.
2. Choose one of the following options to test vector database options:
   - **Option 1:** Set up a proof of concept using your preferred vector database solution.
   - **Option 2:** Experiment with sample datasets in Amazon Bedrock Knowledge Bases. Try the quick-create experience for an Amazon Bedrock Knowledge Base. For an example, see Quick create an Aurora PostgreSQL Knowledge Base for Amazon Bedrock in the Aurora documentation.
3. Review additional resources.
4. Get expert help:
   - Contact your AWS account team or AWS Solutions Architects for implementation guidance.
   - Engage with AWS Partners that specialize in vector databases.
5. Plan your production deployment:
   - Create a migration strategy if moving from existing databases.
   - Develop a scaling plan for your chosen solution.
   - Design your monitoring and maintenance procedures.

Resources

The following resources can help you in choosing a vector database.

AWS blog posts

- Accelerate your generative AI application development with Amazon Bedrock Knowledge Bases Quick Create and Amazon Aurora Serverless
- Amazon OpenSearch Service's vector database capabilities explained

- Dive deep into vector data stores using Amazon Bedrock Knowledge Bases
- Leverage pgvector and Amazon Aurora PostgreSQL for Natural Language Processing, Chatbots and Sentiment Analysis

AWS service documentation

- Choosing an AWS database service
- How Amazon Bedrock knowledge bases work
- Neptune Analytics documentation
- Overview of Amazon Web Services: Databases
- Using Aurora PostgreSQL as a Knowledge Base for Amazon Bedrock
- Working with Amazon Aurora PostgreSQL
- Amazon DocumentDB
- Amazon MemoryDB
- Amazon S3 Vectors

Other AWS resources

- Amazon Bedrock Knowledge Bases
- Vector Databases & Embeddings
- Vector Databases for generative AI applications
- What are Embeddings in Machine Learning?

Other resources

- About PostgreSQL
- pgvector documentation
- Pinecone as a Knowledge Base for Amazon Bedrock
- Redis Enterprise Cloud on AWS

Document history

The following table describes significant changes to this guide, Choosing an AWS vector database for RAG use cases. If you want to be notified about future updates, you can subscribe to an RSS feed.

Change	Description	Date
Added AWS services	We added information about using Amazon DocumentDB, Amazon MemoryDB, Amazon S3 Vectors, and Amazon Neptune Analytics.	March 30, 2026
Initial publication	—	March 6, 2025

AWS Prescriptive Guidance glossary

The following are commonly used terms in strategies, guides, and patterns provided by AWS Prescriptive Guidance. To suggest entries, please use the **Provide feedback** link at the end of the glossary.

Numbers

7 Rs

Seven common migration strategies for moving applications to the cloud. These strategies build upon the 5 Rs that Gartner identified in 2011 and consist of the following:

- Refactor/re-architect – Move an application and modify its architecture by taking full advantage of cloud-native features to improve agility, performance, and scalability. This typically involves porting the operating system and database. Example: Migrate your on-premises Oracle database to the Amazon Aurora PostgreSQL-Compatible Edition.
- Replatform (lift and reshape) – Move an application to the cloud, and introduce some level of optimization to take advantage of cloud capabilities. Example: Migrate your on-premises Oracle database to Amazon Relational Database Service (Amazon RDS) for Oracle in the AWS Cloud.
- Repurchase (drop and shop) – Switch to a different product, typically by moving from a traditional license to a SaaS model. Example: Migrate your customer relationship management (CRM) system to Salesforce.com.
- Rehost (lift and shift) – Move an application to the cloud without making any changes to take advantage of cloud capabilities. Example: Migrate your on-premises Oracle database to Oracle on an EC2 instance in the AWS Cloud.
- Relocate (hypervisor-level lift and shift) – Move infrastructure to the cloud without purchasing new hardware, rewriting applications, or modifying your existing operations. You migrate servers from an on-premises platform to a cloud service for the same platform. Example: Migrate a Microsoft Hyper-V application to AWS.
- Retain (revisit) – Keep applications in your source environment. These might include applications that require major refactoring, and you want to postpone that work until a later time, and legacy applications that you want to retain, because there's no business justification for migrating them.

- Retire – Decommission or remove applications that are no longer needed in your source environment.

A

ABAC

See attribute-based access control.

abstracted services

See managed services.

ACID

See atomicity, consistency, isolation, durability.

active-active migration

A database migration method in which the source and target databases are kept in sync (by using a bidirectional replication tool or dual write operations), and both databases handle transactions from connecting applications during migration. This method supports migration in small, controlled batches instead of requiring a one-time cutover. It's more flexible but requires more work than active-passive migration.

active-passive migration

A database migration method in which the source and target databases are kept in sync, but only the source database handles transactions from connecting applications while data is replicated to the target database. The target database doesn't accept any transactions during migration.

aggregate function

A SQL function that operates on a group of rows and calculates a single return value for the group. Examples of aggregate functions include `SUM` and `MAX`.

AI

See artificial intelligence.

AIOps

See artificial intelligence operations.

anonymization

The process of permanently deleting personal information in a dataset. Anonymization can help protect personal privacy. Anonymized data is no longer considered to be personal data.

anti-pattern

A frequently used solution for a recurring issue where the solution is counter-productive, ineffective, or less effective than an alternative.

application control

A security approach that allows the use of only approved applications in order to help protect a system from malware.

application portfolio

A collection of detailed information about each application used by an organization, including the cost to build and maintain the application, and its business value. This information is key to the portfolio discovery and analysis process and helps identify and prioritize the applications to be migrated, modernized, and optimized.

artificial intelligence (AI)

The field of computer science that is dedicated to using computing technologies to perform cognitive functions that are typically associated with humans, such as learning, solving problems, and recognizing patterns. For more information, see What is Artificial Intelligence?

artificial intelligence operations (AIOps)

The process of using machine learning techniques to solve operational problems, reduce operational incidents and human intervention, and increase service quality. For more information about how AIOps is used in the AWS migration strategy, see the operations integration guide.

asymmetric encryption

An encryption algorithm that uses a pair of keys, a public key for encryption and a private key for decryption. You can share the public key because it isn't used for decryption, but access to the private key should be highly restricted.

atomicity, consistency, isolation, durability (ACID)

A set of software properties that guarantee the data validity and operational reliability of a database, even in the case of errors, power failures, or other problems.

attribute-based access control (ABAC)

The practice of creating fine-grained permissions based on user attributes, such as department, job role, and team name. For more information, see ABAC for AWS in the AWS Identity and Access Management (IAM) documentation.

authoritative data source

A location where you store the primary version of data, which is considered to be the most reliable source of information. You can copy data from the authoritative data source to other locations for the purposes of processing or modifying the data, such as anonymizing, redacting, or pseudonymizing it.

Availability Zone

A distinct location within an AWS Region that is insulated from failures in other Availability Zones and provides inexpensive, low-latency network connectivity to other Availability Zones in the same Region.

AWS Cloud Adoption Framework (AWS CAF)

A framework of guidelines and best practices from AWS to help organizations develop an efficient and effective plan to move successfully to the cloud. AWS CAF organizes guidance into six focus areas called perspectives: business, people, governance, platform, security, and operations. The business, people, and governance perspectives focus on business skills and processes; the platform, security, and operations perspectives focus on technical skills and processes. For example, the people perspective targets stakeholders who handle human resources (HR), staffing functions, and people management. For this perspective, AWS CAF provides guidance for people development, training, and communications to help ready the organization for successful cloud adoption. For more information, see the AWS CAF website and the AWS CAF whitepaper.

AWS Workload Qualification Framework (AWS WQF)

A tool that evaluates database migration workloads, recommends migration strategies, and provides work estimates. AWS WQF is included with AWS Schema Conversion Tool (AWS SCT). It analyzes database schemas and code objects, application code, dependencies, and performance characteristics, and provides assessment reports.

B

bad bot

A bot that is intended to disrupt or cause harm to individuals or organizations.

BCP

See business continuity planning.

behavior graph

A unified, interactive view of resource behavior and interactions over time. You can use a behavior graph with Amazon Detective to examine failed logon attempts, suspicious API calls, and similar actions. For more information, see Data in a behavior graph in the Detective documentation.

big-endian system

A system that stores the most significant byte first. See also endianness.

binary classification

A process that predicts a binary outcome (one of two possible classes). For example, your ML model might need to predict problems such as “Is this email spam or not spam?" or "Is this product a book or a car?"

bloom filter

A probabilistic, memory-efficient data structure that is used to test whether an element is a member of a set.

blue/green deployment

A deployment strategy where you create two separate but identical environments. You run the current application version in one environment (blue) and the new application version in the other environment (green). This strategy helps you quickly roll back with minimal impact.

bot

A software application that runs automated tasks over the internet and simulates human activity or interaction. Some bots are useful or beneficial, such as web crawlers that index information on the internet. Some other bots, known as *bad bots*, are intended to disrupt or cause harm to individuals or organizations.

botnet

Networks of bots that are infected by malware and are under the control of a single party, known as a *bot herder* or *bot operator*. Botnets are the best-known mechanism to scale bots and their impact.

branch

A contained area of a code repository. The first branch created in a repository is the *main branch*. You can create a new branch from an existing branch, and you can then develop features or fix bugs in the new branch. A branch you create to build a feature is commonly referred to as a *feature branch*. When the feature is ready for release, you merge the feature branch back into the main branch. For more information, see About branches (GitHub documentation).

break-glass access

In exceptional circumstances and through an approved process, a quick means for a user to gain access to an AWS account that they don't typically have permissions to access. For more information, see the Implement break-glass procedures indicator in the AWS Well-Architected guidance.

brownfield strategy

The existing infrastructure in your environment. When adopting a brownfield strategy for a system architecture, you design the architecture around the constraints of the current systems and infrastructure. If you are expanding the existing infrastructure, you might blend brownfield and greenfield strategies.

buffer cache

The memory area where the most frequently accessed data is stored.

business capability

What a business does to generate value (for example, sales, customer service, or marketing). Microservices architectures and development decisions can be driven by business capabilities. For more information, see the Organized around business capabilities section of the Running containerized microservices on AWS whitepaper.

business continuity planning (BCP)

A plan that addresses the potential impact of a disruptive event, such as a large-scale migration, on operations and enables a business to resume operations quickly.

C

CAF

See AWS Cloud Adoption Framework.

canary deployment

The slow and incremental release of a version to end users. When you are confident, you deploy the new version and replace the current version in its entirety.

CCoE

See Cloud Center of Excellence.

CDC

See change data capture.

change data capture (CDC)

The process of tracking changes to a data source, such as a database table, and recording metadata about the change. You can use CDC for various purposes, such as auditing or replicating changes in a target system to maintain synchronization.

chaos engineering

Intentionally introducing failures or disruptive events to test a system's resilience. You can use AWS Fault Injection Service (AWS FIS) to perform experiments that stress your AWS workloads and evaluate their response.

CI/CD

See continuous integration and continuous delivery.

classification

A categorization process that helps generate predictions. ML models for classification problems predict a discrete value. Discrete values are always distinct from one another. For example, a model might need to evaluate whether or not there is a car in an image.

client-side encryption

Encryption of data locally, before the target AWS service receives it.

Cloud Center of Excellence (CCoE)

A multi-disciplinary team that drives cloud adoption efforts across an organization, including developing cloud best practices, mobilizing resources, establishing migration timelines, and leading the organization through large-scale transformations. For more information, see the CCoE posts on the AWS Cloud Enterprise Strategy Blog.

cloud computing

The cloud technology that is typically used for remote data storage and IoT device management. Cloud computing is commonly connected to edge computing technology.

cloud operating model

In an IT organization, the operating model that is used to build, mature, and optimize one or more cloud environments. For more information, see Building your Cloud Operating Model.

cloud stages of adoption

The four phases that organizations typically go through when they migrate to the AWS Cloud:

- Project – Running a few cloud-related projects for proof of concept and learning purposes
- Foundation – Making foundational investments to scale your cloud adoption (e.g., creating a landing zone, defining a CCoE, establishing an operations model)
- Migration – Migrating individual applications
- Re-invention – Optimizing products and services, and innovating in the cloud

These stages were defined by Stephen Orban in the blog post The Journey Toward Cloud-First & the Stages of Adoption on the AWS Cloud Enterprise Strategy blog. For information about how they relate to the AWS migration strategy, see the migration readiness guide.

CMDB

See configuration management database.

code repository

A location where source code and other assets, such as documentation, samples, and scripts, are stored and updated through version control processes. Common cloud repositories include GitHub or Bitbucket Cloud. Each version of the code is called a *branch*. In a microservice structure, each repository is devoted to a single piece of functionality. A single CI/CD pipeline can use multiple repositories.

cold cache

A buffer cache that is empty, not well populated, or contains stale or irrelevant data. This affects performance because the database instance must read from the main memory or disk, which is slower than reading from the buffer cache.

cold data

Data that is rarely accessed and is typically historical. When querying this kind of data, slow queries are typically acceptable. Moving this data to lower-performing and less expensive storage tiers or classes can reduce costs.

computer vision (CV)

A field of AI that uses machine learning to analyze and extract information from visual formats such as digital images and videos. For example, Amazon SageMaker AI provides image processing algorithms for CV.

configuration drift

For a workload, a configuration change from the expected state. It might cause the workload to become noncompliant, and it's typically gradual and unintentional.

configuration management database (CMDB)

A repository that stores and manages information about a database and its IT environment, including both hardware and software components and their configurations. You typically use data from a CMDB in the portfolio discovery and analysis stage of migration.

conformance pack

A collection of AWS Config rules and remediation actions that you can assemble to customize your compliance and security checks. You can deploy a conformance pack as a single entity in an AWS account and Region, or across an organization, by using a YAML template. For more information, see Conformance packs in the AWS Config documentation.

continuous integration and continuous delivery (CI/CD)

The process of automating the source, build, test, staging, and production stages of the software release process. CI/CD is commonly described as a pipeline. CI/CD can help you automate processes, improve productivity, improve code quality, and deliver faster. For more information, see Benefits of continuous delivery. CD can also stand for *continuous deployment*. For more information, see Continuous Delivery vs. Continuous Deployment.

CV

See computer vision.

D

data at rest

Data that is stationary in your network, such as data that is in storage.

data classification

A process for identifying and categorizing the data in your network based on its criticality and sensitivity. It is a critical component of any cybersecurity risk management strategy because it helps you determine the appropriate protection and retention controls for the data. Data classification is a component of the security pillar in the AWS Well-Architected Framework. For more information, see Data classification.

data drift

A meaningful variation between the production data and the data that was used to train an ML model, or a meaningful change in the input data over time. Data drift can reduce the overall quality, accuracy, and fairness in ML model predictions.

data in transit

Data that is actively moving through your network, such as between network resources.

data mesh

An architectural framework that provides distributed, decentralized data ownership with centralized management and governance.

data minimization

The principle of collecting and processing only the data that is strictly necessary. Practicing data minimization in the AWS Cloud can reduce privacy risks, costs, and your analytics carbon footprint.

data perimeter

A set of preventive guardrails in your AWS environment that help make sure that only trusted identities are accessing trusted resources from expected networks. For more information, see Building a data perimeter on AWS.

data preprocessing

To transform raw data into a format that is easily parsed by your ML model. Preprocessing data can mean removing certain columns or rows and addressing missing, inconsistent, or duplicate values.

data provenance

The process of tracking the origin and history of data throughout its lifecycle, such as how the data was generated, transmitted, and stored.

data subject

An individual whose data is being collected and processed.

data warehouse

A data management system that supports business intelligence, such as analytics. Data warehouses commonly contain large amounts of historical data, and they are typically used for queries and analysis.

database definition language (DDL)

Statements or commands for creating or modifying the structure of tables and objects in a database.

database manipulation language (DML)

Statements or commands for modifying (inserting, updating, and deleting) information in a database.

DDL

See database definition language.

deep ensemble

To combine multiple deep learning models for prediction. You can use deep ensembles to obtain a more accurate prediction or for estimating uncertainty in predictions.

deep learning

An ML subfield that uses multiple layers of artificial neural networks to identify mapping between input data and target variables of interest.

defense-in-depth

An information security approach in which a series of security mechanisms and controls are thoughtfully layered throughout a computer network to protect the confidentiality, integrity, and availability of the network and the data within. When you adopt this strategy on AWS, you add multiple controls at different layers of the AWS Organizations structure to help secure resources. For example, a defense-in-depth approach might combine multi-factor authentication, network segmentation, and encryption.

delegated administrator

In AWS Organizations, a compatible service can register an AWS member account to administer the organization's accounts and manage permissions for that service. This account is called the *delegated administrator* for that service. For more information and a list of compatible services, see Services that work with AWS Organizations in the AWS Organizations documentation.

deployment

The process of making an application, new features, or code fixes available in the target environment. Deployment involves implementing changes in a code base and then building and running that code base in the application's environments.

development environment

See environment.

detective control

A security control that is designed to detect, log, and alert after an event has occurred. These controls are a second line of defense, alerting you to security events that bypassed the preventative controls in place. For more information, see Detective controls in *Implementing security controls on AWS*.

development value stream mapping (DVSM)

A process used to identify and prioritize constraints that adversely affect speed and quality in a software development lifecycle. DVSM extends the value stream mapping process originally designed for lean manufacturing practices. It focuses on the steps and teams required to create and move value through the software development process.

digital twin

A virtual representation of a real-world system, such as a building, factory, industrial equipment, or production line. Digital twins support predictive maintenance, remote monitoring, and production optimization.

dimension table

In a star schema, a smaller table that contains data attributes about quantitative data in a fact table. Dimension table attributes are typically text fields or discrete numbers that behave like text. These attributes are commonly used for query constraining, filtering, and result set labeling.

disaster

An event that prevents a workload or system from fulfilling its business objectives in its primary deployed location. These events can be natural disasters, technical failures, or the result of human actions, such as unintentional misconfiguration or a malware attack.

disaster recovery (DR)

The strategy and process you use to minimize downtime and data loss caused by a disaster. For more information, see Disaster Recovery of Workloads on AWS: Recovery in the Cloud in the AWS Well-Architected Framework.

DML

See database manipulation language.

domain-driven design

An approach to developing a complex software system by connecting its components to evolving domains, or core business goals, that each component serves. This concept was introduced by Eric Evans in his book, *Domain-Driven Design: Tackling Complexity in the Heart of Software* (Boston: Addison-Wesley Professional, 2003). For information about how you can use domain-driven design with the strangler fig pattern, see Modernizing legacy Microsoft ASP.NET (ASMX) web services incrementally by using containers and Amazon API Gateway.

DR

See disaster recovery.

drift detection

Tracking deviations from a baselined configuration. For example, you can use AWS CloudFormation to detect drift in system resources, or you can use AWS Control Tower to detect changes in your landing zone that might affect compliance with governance requirements.

DVSM

See development value stream mapping.

E

EDA

See exploratory data analysis.

EDI

See electronic data interchange.

edge computing

The technology that increases the computing power for smart devices at the edges of an IoT network. When compared with cloud computing, edge computing can reduce communication latency and improve response time.

electronic data interchange (EDI)

The automated exchange of business documents between organizations. For more information, see What is Electronic Data Interchange.

encryption

A computing process that transforms plaintext data, which is human-readable, into ciphertext.

encryption key

A cryptographic string of randomized bits that is generated by an encryption algorithm. Keys can vary in length, and each key is designed to be unpredictable and unique.

endianness

The order in which bytes are stored in computer memory. Big-endian systems store the most significant byte first. Little-endian systems store the least significant byte first.

endpoint

See service endpoint.

endpoint service

A service that you can host in a virtual private cloud (VPC) to share with other users. You can create an endpoint service with AWS PrivateLink and grant permissions to other AWS accounts or to AWS Identity and Access Management (IAM) principals. These accounts or principals can connect to your endpoint service privately by creating interface VPC endpoints. For more

information, see Create an endpoint service in the Amazon Virtual Private Cloud (Amazon VPC) documentation.

enterprise resource planning (ERP)

A system that automates and manages key business processes (such as accounting, MES, and project management) for an enterprise.

envelope encryption

The process of encrypting an encryption key with another encryption key. For more information, see Envelope encryption in the AWS Key Management Service (AWS KMS) documentation.

environment

An instance of a running application. The following are common types of environments in cloud computing:

- development environment – An instance of a running application that is available only to the core team responsible for maintaining the application. Development environments are used to test changes before promoting them to upper environments. This type of environment is sometimes referred to as a *test environment*.
- lower environments – All development environments for an application, such as those used for initial builds and tests.
- production environment – An instance of a running application that end users can access. In a CI/CD pipeline, the production environment is the last deployment environment.
- upper environments – All environments that can be accessed by users other than the core development team. This can include a production environment, preproduction environments, and environments for user acceptance testing.

epic

In agile methodologies, functional categories that help organize and prioritize your work. Epics provide a high-level description of requirements and implementation tasks. For example, AWS CAF security epics include identity and access management, detective controls, infrastructure security, data protection, and incident response. For more information about epics in the AWS migration strategy, see the program implementation guide.

ERP

See enterprise resource planning.

exploratory data analysis (EDA)

The process of analyzing a dataset to understand its main characteristics. You collect or aggregate data and then perform initial investigations to find patterns, detect anomalies, and check assumptions. EDA is performed by calculating summary statistics and creating data visualizations.

F

fact table

The central table in a star schema. It stores quantitative data about business operations. Typically, a fact table contains two types of columns: those that contain measures and those that contain a foreign key to a dimension table.

fail fast

A philosophy that uses frequent and incremental testing to reduce the development lifecycle. It is a critical part of an agile approach.

fault isolation boundary

In the AWS Cloud, a boundary such as an Availability Zone, AWS Region, control plane, or data plane that limits the effect of a failure and helps improve the resilience of workloads. For more information, see AWS Fault Isolation Boundaries.

feature branch

See branch.

features

The input data that you use to make a prediction. For example, in a manufacturing context, features could be images that are periodically captured from the manufacturing line.

feature importance

How significant a feature is for a model's predictions. This is usually expressed as a numerical score that can be calculated through various techniques, such as Shapley Additive Explanations (SHAP) and integrated gradients. For more information, see Machine learning model interpretability with AWS.

feature transformation

To optimize data for the ML process, including enriching data with additional sources, scaling values, or extracting multiple sets of information from a single data field. This enables the ML model to benefit from the data. For example, if you break down the “2021-05-27 00:15:37” date into “2021”, “May”, “Thu”, and “15”, you can help the learning algorithm learn nuanced patterns associated with different data components.

few-shot prompting

Providing an LLM with a small number of examples that demonstrate the task and desired output before asking it to perform a similar task. This technique is an application of in-context learning, where models learn from examples (*shots*) that are embedded in prompts. Few-shot prompting can be effective for tasks that require specific formatting, reasoning, or domain knowledge. See also zero-shot prompting.

FGAC

See fine-grained access control.

fine-grained access control (FGAC)

The use of multiple conditions to allow or deny an access request.

flash-cut migration

A database migration method that uses continuous data replication through change data capture to migrate data in the shortest time possible, instead of using a phased approach. The objective is to keep downtime to a minimum.

FM

See foundation model.

foundation model (FM)

A large deep-learning neural network that has been training on massive datasets of generalized and unlabeled data. FMs are capable of performing a wide variety of general tasks, such as understanding language, generating text and images, and conversing in natural language. For more information, see What are Foundation Models.

G

generative AI

A subset of AI models that have been trained on large amounts of data and that can use a simple text prompt to create new content and artifacts, such as images, videos, text, and audio. For more information, see What is Generative AI.

geo blocking

See geographic restrictions.

geographic restrictions (geo blocking)

In Amazon CloudFront, an option to prevent users in specific countries from accessing content distributions. You can use an allow list or block list to specify approved and banned countries. For more information, see Restricting the geographic distribution of your content in the CloudFront documentation.

Gitflow workflow

An approach in which lower and upper environments use different branches in a source code repository. The Gitflow workflow is considered legacy, and the trunk-based workflow is the modern, preferred approach.

golden image

A snapshot of a system or software that is used as a template to deploy new instances of that system or software. For example, in manufacturing, a golden image can be used to provision software on multiple devices and helps improve speed, scalability, and productivity in device manufacturing operations.

greenfield strategy

The absence of existing infrastructure in a new environment. When adopting a greenfield strategy for a system architecture, you can select all new technologies without the restriction of compatibility with existing infrastructure, also known as brownfield. If you are expanding the existing infrastructure, you might blend brownfield and greenfield strategies.

guardrail

A high-level rule that helps govern resources, policies, and compliance across organizational units (OUs). *Preventive guardrails* enforce policies to ensure alignment to compliance standards. They are implemented by using service control policies and IAM permissions boundaries.

Detective guardrails detect policy violations and compliance issues, and generate alerts for remediation. They are implemented by using AWS Config, AWS Security Hub CSPM, Amazon GuardDuty, AWS Trusted Advisor, Amazon Inspector, and custom AWS Lambda checks.

H

HA

See high availability.

heterogeneous database migration

Migrating your source database to a target database that uses a different database engine (for example, Oracle to Amazon Aurora). Heterogeneous migration is typically part of a re-architecting effort, and converting the schema can be a complex task. AWS provides AWS SCT that helps with schema conversions.

high availability (HA)

The ability of a workload to operate continuously, without intervention, in the event of challenges or disasters. HA systems are designed to automatically fail over, consistently deliver high-quality performance, and handle different loads and failures with minimal performance impact.

historian modernization

An approach used to modernize and upgrade operational technology (OT) systems to better serve the needs of the manufacturing industry. A *historian* is a type of database that is used to collect and store data from various sources in a factory.

holdout data

A portion of historical, labeled data that is withheld from a dataset that is used to train a machine learning model. You can use holdout data to evaluate the model performance by comparing the model predictions against the holdout data.

homogeneous database migration

Migrating your source database to a target database that shares the same database engine (for example, Microsoft SQL Server to Amazon RDS for SQL Server). Homogeneous migration is typically part of a rehosting or replatforming effort. You can use native database utilities to migrate the schema.

hot data

Data that is frequently accessed, such as real-time data or recent translational data. This data typically requires a high-performance storage tier or class to provide fast query responses.

hotfix

An urgent fix for a critical issue in a production environment. Due to its urgency, a hotfix is usually made outside of the typical DevOps release workflow.

hypercare period

Immediately following cutover, the period of time when a migration team manages and monitors the migrated applications in the cloud in order to address any issues. Typically, this period is 1–4 days in length. At the end of the hypercare period, the migration team typically transfers responsibility for the applications to the cloud operations team.

I

IaC

See infrastructure as code.

identity-based policy

A policy attached to one or more IAM principals that defines their permissions within the AWS Cloud environment.

idle application

An application that has an average CPU and memory usage between 5 and 20 percent over a period of 90 days. In a migration project, it is common to retire these applications or retain them on premises.

IIoT

See industrial Internet of Things.

immutable infrastructure

A model that deploys new infrastructure for production workloads instead of updating, patching, or modifying the existing infrastructure. Immutable infrastructures are inherently more consistent, reliable, and predictable than mutable infrastructure. For more information, see the Deploy using immutable infrastructure best practice in the AWS Well-Architected Framework.

inbound (ingress) VPC

In an AWS multi-account architecture, a VPC that accepts, inspects, and routes network connections from outside an application. The AWS Security Reference Architecture recommends setting up your Network account with inbound, outbound, and inspection VPCs to protect the two-way interface between your application and the broader internet.

incremental migration

A cutover strategy in which you migrate your application in small parts instead of performing a single, full cutover. For example, you might move only a few microservices or users to the new system initially. After you verify that everything is working properly, you can incrementally move additional microservices or users until you can decommission your legacy system. This strategy reduces the risks associated with large migrations.

Industry 4.0

A term that was introduced by Klaus Schwab in 2016 to refer to the modernization of manufacturing processes through advances in connectivity, real-time data, automation, analytics, and AI/ML.

infrastructure

All of the resources and assets contained within an application's environment.

infrastructure as code (IaC)

The process of provisioning and managing an application's infrastructure through a set of configuration files. IaC is designed to help you centralize infrastructure management, standardize resources, and scale quickly so that new environments are repeatable, reliable, and consistent.

industrial Internet of Things (IIoT)

The use of internet-connected sensors and devices in the industrial sectors, such as manufacturing, energy, automotive, healthcare, life sciences, and agriculture. For more information, see Building an industrial Internet of Things (IIoT) digital transformation strategy.

inspection VPC

In an AWS multi-account architecture, a centralized VPC that manages inspections of network traffic between VPCs (in the same or different AWS Regions), the internet, and on-premises networks. The AWS Security Reference Architecture recommends setting up your Network account with inbound, outbound, and inspection VPCs to protect the two-way interface between your application and the broader internet.

Internet of Things (IoT)

The network of connected physical objects with embedded sensors or processors that communicate with other devices and systems through the internet or over a local communication network. For more information, see What is IoT?

interpretability

A characteristic of a machine learning model that describes the degree to which a human can understand how the model's predictions depend on its inputs. For more information, see Machine learning model interpretability with AWS.

IoT

See Internet of Things.

IT information library (ITIL)

A set of best practices for delivering IT services and aligning these services with business requirements. ITIL provides the foundation for ITSM.

IT service management (ITSM)

Activities associated with designing, implementing, managing, and supporting IT services for an organization. For information about integrating cloud operations with ITSM tools, see the operations integration guide.

ITIL

See IT information library.

ITSM

See IT service management.

L

label-based access control (LBAC)

An implementation of mandatory access control (MAC) where the users and the data itself are each explicitly assigned a security label value. The intersection between the user security label and data security label determines which rows and columns can be seen by the user.

landing zone

A landing zone is a well-architected, multi-account AWS environment that is scalable and secure. This is a starting point from which your organizations can quickly launch and deploy workloads and applications with confidence in their security and infrastructure environment. For more information about landing zones, see Setting up a secure and scalable multi-account AWS environment.

large language model (LLM)

A deep learning AI model that is pretrained on a vast amount of data. An LLM can perform multiple tasks, such as answering questions, summarizing documents, translating text into other languages, and completing sentences. For more information, see What are LLMs.

large migration

A migration of 300 or more servers.

LBAC

See label-based access control.

least privilege

The security best practice of granting the minimum permissions required to perform a task. For more information, see Apply least-privilege permissions in the IAM documentation.

lift and shift

See 7 Rs.

little-endian system

A system that stores the least significant byte first. See also endianness.

LLM

See large language model.

lower environments

See environment.

M

machine learning (ML)

A type of artificial intelligence that uses algorithms and techniques for pattern recognition and learning. ML analyzes and learns from recorded data, such as Internet of Things (IoT) data, to generate a statistical model based on patterns. For more information, see Machine Learning.

main branch

See branch.

malware

Software that is designed to compromise computer security or privacy. Malware might disrupt computer systems, leak sensitive information, or gain unauthorized access. Examples of malware include viruses, worms, ransomware, Trojan horses, spyware, and keyloggers.

managed services

AWS services for which AWS operates the infrastructure layer, the operating system, and platforms, and you access the endpoints to store and retrieve data. Amazon Simple Storage Service (Amazon S3) and Amazon DynamoDB are examples of managed services. These are also known as *abstracted services*.

manufacturing execution system (MES)

A software system for tracking, monitoring, documenting, and controlling production processes that convert raw materials to finished products on the shop floor.

MAP

See Migration Acceleration Program.

mechanism

A complete process in which you create a tool, drive adoption of the tool, and then inspect the results in order to make adjustments. A mechanism is a cycle that reinforces and improves itself as it operates. For more information, see Building mechanisms in the AWS Well-Architected Framework.

member account

All AWS accounts other than the management account that are part of an organization in AWS Organizations. An account can be a member of only one organization at a time.

MES

See manufacturing execution system.

Message Queuing Telemetry Transport (MQTT)

A lightweight, machine-to-machine (M2M) communication protocol, based on the publish/subscribe pattern, for resource-constrained IoT devices.

microservice

A small, independent service that communicates over well-defined APIs and is typically owned by small, self-contained teams. For example, an insurance system might include microservices that map to business capabilities, such as sales or marketing, or subdomains, such as purchasing, claims, or analytics. The benefits of microservices include agility, flexible scaling, easy deployment, reusable code, and resilience. For more information, see Integrating microservices by using AWS serverless services.

microservices architecture

An approach to building an application with independent components that run each application process as a microservice. These microservices communicate through a well-defined interface by using lightweight APIs. Each microservice in this architecture can be updated, deployed, and scaled to meet demand for specific functions of an application. For more information, see Implementing microservices on AWS.

Migration Acceleration Program (MAP)

An AWS program that provides consulting support, training, and services to help organizations build a strong operational foundation for moving to the cloud, and to help offset the initial cost of migrations. MAP includes a migration methodology for executing legacy migrations in a methodical way and a set of tools to automate and accelerate common migration scenarios.

migration at scale

The process of moving the majority of the application portfolio to the cloud in waves, with more applications moved at a faster rate in each wave. This phase uses the best practices and lessons learned from the earlier phases to implement a *migration factory* of teams, tools, and processes to streamline the migration of workloads through automation and agile delivery. This is the third phase of the AWS migration strategy.

migration factory

Cross-functional teams that streamline the migration of workloads through automated, agile approaches. Migration factory teams typically include operations, business analysts and owners,

migration engineers, developers, and DevOps professionals working in sprints. Between 20 and 50 percent of an enterprise application portfolio consists of repeated patterns that can be optimized by a factory approach. For more information, see the discussion of migration factories and the Cloud Migration Factory guide in this content set.

migration metadata

The information about the application and server that is needed to complete the migration. Each migration pattern requires a different set of migration metadata. Examples of migration metadata include the target subnet, security group, and AWS account.

migration pattern

A repeatable migration task that details the migration strategy, the migration destination, and the migration application or service used. Example: Rehost migration to Amazon EC2 with AWS Application Migration Service.

Migration Portfolio Assessment (MPA)

An online tool that provides information for validating the business case for migrating to the AWS Cloud. MPA provides detailed portfolio assessment (server right-sizing, pricing, TCO comparisons, migration cost analysis) as well as migration planning (application data analysis and data collection, application grouping, migration prioritization, and wave planning). The MPA tool (requires login) is available free of charge to all AWS consultants and APN Partner consultants.

Migration Readiness Assessment (MRA)

The process of gaining insights about an organization's cloud readiness status, identifying strengths and weaknesses, and building an action plan to close identified gaps, using the AWS CAF. For more information, see the migration readiness guide. MRA is the first phase of the AWS migration strategy.

migration strategy

The approach used to migrate a workload to the AWS Cloud. For more information, see the 7 Rs entry in this glossary and see Mobilize your organization to accelerate large-scale migrations.

ML

See machine learning.

modernization

Transforming an outdated (legacy or monolithic) application and its infrastructure into an agile, elastic, and highly available system in the cloud to reduce costs, gain efficiencies, and take advantage of innovations. For more information, see Strategy for modernizing applications in the AWS Cloud.

modernization readiness assessment

An evaluation that helps determine the modernization readiness of an organization's applications; identifies benefits, risks, and dependencies; and determines how well the organization can support the future state of those applications. The outcome of the assessment is a blueprint of the target architecture, a roadmap that details development phases and milestones for the modernization process, and an action plan for addressing identified gaps. For more information, see Evaluating modernization readiness for applications in the AWS Cloud.

monolithic applications (monoliths)

Applications that run as a single service with tightly coupled processes. Monolithic applications have several drawbacks. If one application feature experiences a spike in demand, the entire architecture must be scaled. Adding or improving a monolithic application's features also becomes more complex when the code base grows. To address these issues, you can use a microservices architecture. For more information, see Decomposing monoliths into microservices.

MPA

See Migration Portfolio Assessment.

MQTT

See Message Queuing Telemetry Transport.

multiclass classification

A process that helps generate predictions for multiple classes (predicting one of more than two outcomes). For example, an ML model might ask "Is this product a book, car, or phone?" or "Which product category is most interesting to this customer?"

mutable infrastructure

A model that updates and modifies the existing infrastructure for production workloads. For improved consistency, reliability, and predictability, the AWS Well-Architected Framework recommends the use of immutable infrastructure as a best practice.

O

OAC

See origin access control.

OAI

See origin access identity.

OCM

See organizational change management.

offline migration

A migration method in which the source workload is taken down during the migration process. This method involves extended downtime and is typically used for small, non-critical workloads.

OI

See operations integration.

OLA

See operational-level agreement.

online migration

A migration method in which the source workload is copied to the target system without being taken offline. Applications that are connected to the workload can continue to function during the migration. This method involves zero to minimal downtime and is typically used for critical production workloads.

OPC-UA

See Open Process Communications - Unified Architecture.

Open Process Communications - Unified Architecture (OPC-UA)

A machine-to-machine (M2M) communication protocol for industrial automation. OPC-UA provides an interoperability standard with data encryption, authentication, and authorization schemes.

operational-level agreement (OLA)

An agreement that clarifies what functional IT groups promise to deliver to each other, to support a service-level agreement (SLA).

operational readiness review (ORR)

A checklist of questions and associated best practices that help you understand, evaluate, prevent, or reduce the scope of incidents and possible failures. For more information, see Operational Readiness Reviews (ORR) in the AWS Well-Architected Framework.

operational technology (OT)

Hardware and software systems that work with the physical environment to control industrial operations, equipment, and infrastructure. In manufacturing, the integration of OT and information technology (IT) systems is a key focus for Industry 4.0 transformations.

operations integration (OI)

The process of modernizing operations in the cloud, which involves readiness planning, automation, and integration. For more information, see the operations integration guide.

organization trail

A trail that's created by AWS CloudTrail that logs all events for all AWS accounts in an organization in AWS Organizations. This trail is created in each AWS account that's part of the organization and tracks the activity in each account. For more information, see Creating a trail for an organization in the CloudTrail documentation.

organizational change management (OCM)

A framework for managing major, disruptive business transformations from a people, culture, and leadership perspective. OCM helps organizations prepare for, and transition to, new systems and strategies by accelerating change adoption, addressing transitional issues, and driving cultural and organizational changes. In the AWS migration strategy, this framework is called *people acceleration*, because of the speed of change required in cloud adoption projects. For more information, see the OCM guide.

origin access control (OAC)

In CloudFront, an enhanced option for restricting access to secure your Amazon Simple Storage Service (Amazon S3) content. OAC supports all S3 buckets in all AWS Regions, server-side encryption with AWS KMS (SSE-KMS), and dynamic PUT and DELETE requests to the S3 bucket.

origin access identity (OAI)

In CloudFront, an option for restricting access to secure your Amazon S3 content. When you use OAI, CloudFront creates a principal that Amazon S3 can authenticate with. Authenticated principals can access content in an S3 bucket only through a specific CloudFront distribution. See also OAC, which provides more granular and enhanced access control.

ORR

See operational readiness review.

OT

See operational technology.

outbound (egress) VPC

In an AWS multi-account architecture, a VPC that handles network connections that are initiated from within an application. The AWS Security Reference Architecture recommends setting up your Network account with inbound, outbound, and inspection VPCs to protect the two-way interface between your application and the broader internet.

P

permissions boundary

An IAM management policy that is attached to IAM principals to set the maximum permissions that the user or role can have. For more information, see Permissions boundaries in the IAM documentation.

personally identifiable information (PII)

Information that, when viewed directly or paired with other related data, can be used to reasonably infer the identity of an individual. Examples of PII include names, addresses, and contact information.

PII

See personally identifiable information.

playbook

A set of predefined steps that capture the work associated with migrations, such as delivering core operations functions in the cloud. A playbook can take the form of scripts, automated runbooks, or a summary of processes or steps required to operate your modernized environment.

PLC

See programmable logic controller.

PLM

See product lifecycle management.

policy

An object that can define permissions (see identity-based policy), specify access conditions (see resource-based policy), or define the maximum permissions for all accounts in an organization in AWS Organizations (see service control policy).

polyglot persistence

Independently choosing a microservice's data storage technology based on data access patterns and other requirements. If your microservices have the same data storage technology, they can encounter implementation challenges or experience poor performance. Microservices are more easily implemented and achieve better performance and scalability if they use the data store best adapted to their requirements.

portfolio assessment

A process of discovering, analyzing, and prioritizing the application portfolio in order to plan the migration. For more information, see Evaluating migration readiness.

predicate

A query condition that returns `true` or `false`, commonly located in a `WHERE` clause.

predicate pushdown

A database query optimization technique that filters the data in the query before transfer. This reduces the amount of data that must be retrieved and processed from the relational database, and it improves query performance.

preventative control

A security control that is designed to prevent an event from occurring. These controls are a first line of defense to help prevent unauthorized access or unwanted changes to your network. For more information, see Preventative controls in *Implementing security controls on AWS*.

principal

An entity in AWS that can perform actions and access resources. This entity is typically a root user for an AWS account, an IAM role, or a user. For more information, see *Principal* in Roles terms and concepts in the IAM documentation.

privacy by design

A system engineering approach that takes privacy into account through the whole development process.

private hosted zones

A container that holds information about how you want Amazon Route 53 to respond to DNS queries for a domain and its subdomains within one or more VPCs. For more information, see Working with private hosted zones in the Route 53 documentation.

proactive control

A security control designed to prevent the deployment of noncompliant resources. These controls scan resources before they are provisioned. If the resource is not compliant with the control, then it isn't provisioned. For more information, see the Controls reference guide in the AWS Control Tower documentation and see Proactive controls in *Implementing security controls on AWS*.

product lifecycle management (PLM)

The management of data and processes for a product throughout its entire lifecycle, from design, development, and launch, through growth and maturity, to decline and removal.

production environment

See environment.

programmable logic controller (PLC)

In manufacturing, a highly reliable, adaptable computer that monitors machines and automates manufacturing processes.

prompt chaining

Using the output of one LLM prompt as the input for the next prompt to generate better responses. This technique is used to break down a complex task into subtasks, or to iteratively refine or expand a preliminary response. It helps improve the accuracy and relevance of a model's responses and allows for more granular, personalized results.

pseudonymization

The process of replacing personal identifiers in a dataset with placeholder values. Pseudonymization can help protect personal privacy. Pseudonymized data is still considered to be personal data.

publish/subscribe (pub/sub)

A pattern that enables asynchronous communications among microservices to improve scalability and responsiveness. For example, in a microservices-based MES, a microservice can publish event messages to a channel that other microservices can subscribe to. The system can add new microservices without changing the publishing service.

Q

query plan

A series of steps, like instructions, that are used to access the data in a SQL relational database system.

query plan regression

When a database service optimizer chooses a less optimal plan than it did before a given change to the database environment. This can be caused by changes to statistics, constraints, environment settings, query parameter bindings, and updates to the database engine.

R

RACI matrix

See responsible, accountable, consulted, informed (RACI).

RAG

See Retrieval Augmented Generation.

ransomware

A malicious software that is designed to block access to a computer system or data until a payment is made.

RASCI matrix

See responsible, accountable, consulted, informed (RACI).

RCAC

See row and column access control.

read replica

A copy of a database that's used for read-only purposes. You can route queries to the read replica to reduce the load on your primary database.

re-architect

See 7 Rs.

recovery point objective (RPO)

The maximum acceptable amount of time since the last data recovery point. This determines what is considered an acceptable loss of data between the last recovery point and the interruption of service.

recovery time objective (RTO)

The maximum acceptable delay between the interruption of service and restoration of service.

refactor

See 7 Rs.

Region

A collection of AWS resources in a geographic area. Each AWS Region is isolated and independent of the others to provide fault tolerance, stability, and resilience. For more information, see Specify which AWS Regions your account can use.

regression

An ML technique that predicts a numeric value. For example, to solve the problem of "What price will this house sell for?" an ML model could use a linear regression model to predict a house's sale price based on known facts about the house (for example, the square footage).

rehost

See 7 Rs.

release

In a deployment process, the act of promoting changes to a production environment.

relocate

See 7 Rs.

replatform

See 7 Rs.

repurchase

See 7 Rs.

resiliency

An application's ability to resist or recover from disruptions. High availability and disaster recovery are common considerations when planning for resiliency in the AWS Cloud. For more information, see AWS Cloud Resilience.

resource-based policy

A policy attached to a resource, such as an Amazon S3 bucket, an endpoint, or an encryption key. This type of policy specifies which principals are allowed access, supported actions, and any other conditions that must be met.

responsible, accountable, consulted, informed (RACI) matrix

A matrix that defines the roles and responsibilities for all parties involved in migration activities and cloud operations. The matrix name is derived from the responsibility types defined in the matrix: responsible (R), accountable (A), consulted (C), and informed (I). The support (S) type is optional. If you include support, the matrix is called a *RASCI matrix*, and if you exclude it, it's called a *RACI matrix*.

responsive control

A security control that is designed to drive remediation of adverse events or deviations from your security baseline. For more information, see Responsive controls in *Implementing security controls on AWS*.

retain

See 7 Rs.

retire

See 7 Rs.

Retrieval Augmented Generation (RAG)

A generative AI technology in which an LLM references an authoritative data source that is outside of its training data sources before generating a response. For example, a RAG model might perform a semantic search of an organization's knowledge base or custom data. For more information, see What is RAG.

rotation

The process of periodically updating a secret to make it more difficult for an attacker to access the credentials.

row and column access control (RCAC)

The use of basic, flexible SQL expressions that have defined access rules. RCAC consists of row permissions and column masks.

RPO

See recovery point objective.

RTO

See recovery time objective.

runbook

A set of manual or automated procedures required to perform a specific task. These are typically built to streamline repetitive operations or procedures with high error rates.

S

SAML 2.0

An open standard that many identity providers (IdPs) use. This feature enables federated single sign-on (SSO), so users can log into the AWS Management Console or call the AWS API operations without you having to create user in IAM for everyone in your organization. For more information about SAML 2.0-based federation, see About SAML 2.0-based federation in the IAM documentation.

SCADA

See supervisory control and data acquisition.

SCP

See service control policy.

secret

In AWS Secrets Manager, confidential or restricted information, such as a password or user credentials, that you store in encrypted form. It consists of the secret value and its metadata.

The secret value can be binary, a single string, or multiple strings. For more information, see What's in a Secrets Manager secret? in the Secrets Manager documentation.

security by design

A system engineering approach that takes security into account through the whole development process.

security control

A technical or administrative guardrail that prevents, detects, or reduces the ability of a threat actor to exploit a security vulnerability. There are four primary types of security controls: preventative, detective, responsive, and proactive.

security hardening

The process of reducing the attack surface to make it more resistant to attacks. This can include actions such as removing resources that are no longer needed, implementing the security best practice of granting least privilege, or deactivating unnecessary features in configuration files.

security information and event management (SIEM) system

Tools and services that combine security information management (SIM) and security event management (SEM) systems. A SIEM system collects, monitors, and analyzes data from servers, networks, devices, and other sources to detect threats and security breaches, and to generate alerts.

security response automation

A predefined and programmed action that is designed to automatically respond to or remediate a security event. These automations serve as detective or responsive security controls that help you implement AWS security best practices. Examples of automated response actions include modifying a VPC security group, patching an Amazon EC2 instance, or rotating credentials.

server-side encryption

Encryption of data at its destination, by the AWS service that receives it.

service control policy (SCP)

A policy that provides centralized control over permissions for all accounts in an organization in AWS Organizations. SCPs define guardrails or set limits on actions that an administrator can delegate to users or roles. You can use SCPs as allow lists or deny lists, to specify which services or actions are permitted or prohibited. For more information, see Service control policies in the AWS Organizations documentation.

service endpoint

The URL of the entry point for an AWS service. You can use the endpoint to connect programmatically to the target service. For more information, see AWS service endpoints in *AWS General Reference*.

service-level agreement (SLA)

An agreement that clarifies what an IT team promises to deliver to their customers, such as service uptime and performance.

service-level indicator (SLI)

A measurement of a performance aspect of a service, such as its error rate, availability, or throughput.

service-level objective (SLO)

A target metric that represents the health of a service, as measured by a service-level indicator.

shared responsibility model

A model describing the responsibility you share with AWS for cloud security and compliance. AWS is responsible for security *of* the cloud, whereas you are responsible for security *in* the cloud. For more information, see Shared responsibility model.

SIEM

See security information and event management system.

single point of failure (SPOF)

A failure in a single, critical component of an application that can disrupt the system.

SLA

See service-level agreement.

SLI

See service-level indicator.

SLO

See service-level objective.

split-and-seed model

A pattern for scaling and accelerating modernization projects. As new features and product releases are defined, the core team splits up to create new product teams. This helps scale your

organization's capabilities and services, improves developer productivity, and supports rapid innovation. For more information, see Phased approach to modernizing applications in the AWS Cloud.

SPOF

See single point of failure.

star schema

A database organizational structure that uses one large fact table to store transactional or measured data and uses one or more smaller dimensional tables to store data attributes. This structure is designed for use in a data warehouse or for business intelligence purposes.

strangler fig pattern

An approach to modernizing monolithic systems by incrementally rewriting and replacing system functionality until the legacy system can be decommissioned. This pattern uses the analogy of a fig vine that grows into an established tree and eventually overcomes and replaces its host. The pattern was introduced by Martin Fowler as a way to manage risk when rewriting monolithic systems. For an example of how to apply this pattern, see Modernizing legacy Microsoft ASP.NET (ASMX) web services incrementally by using containers and Amazon API Gateway.

subnet

A range of IP addresses in your VPC. A subnet must reside in a single Availability Zone.

supervisory control and data acquisition (SCADA)

In manufacturing, a system that uses hardware and software to monitor physical assets and production operations.

symmetric encryption

An encryption algorithm that uses the same key to encrypt and decrypt the data.

synthetic testing

Testing a system in a way that simulates user interactions to detect potential issues or to monitor performance. You can use Amazon CloudWatch Synthetics to create these tests.

system prompt

A technique for providing context, instructions, or guidelines to an LLM to direct its behavior. System prompts help set context and establish rules for interactions with users.

T

tags

Key-value pairs that act as metadata for organizing your AWS resources. Tags can help you manage, identify, organize, search for, and filter resources. For more information, see Tagging your AWS resources.

target variable

The value that you are trying to predict in supervised ML. This is also referred to as an *outcome variable*. For example, in a manufacturing setting the target variable could be a product defect.

task list

A tool that is used to track progress through a runbook. A task list contains an overview of the runbook and a list of general tasks to be completed. For each general task, it includes the estimated amount of time required, the owner, and the progress.

test environment

See environment.

training

To provide data for your ML model to learn from. The training data must contain the correct answer. The learning algorithm finds patterns in the training data that map the input data attributes to the target (the answer that you want to predict). It outputs an ML model that captures these patterns. You can then use the ML model to make predictions on new data for which you don't know the target.

transit gateway

A network transit hub that you can use to interconnect your VPCs and on-premises networks. For more information, see What is a transit gateway in the AWS Transit Gateway documentation.

trunk-based workflow

An approach in which developers build and test features locally in a feature branch and then merge those changes into the main branch. The main branch is then built to the development, preproduction, and production environments, sequentially.

trusted access

Granting permissions to a service that you specify to perform tasks in your organization in AWS Organizations and in its accounts on your behalf. The trusted service creates a service-linked role in each account, when that role is needed, to perform management tasks for you. For more information, see Using AWS Organizations with other AWS services in the AWS Organizations documentation.

tuning

To change aspects of your training process to improve the ML model's accuracy. For example, you can train the ML model by generating a labeling set, adding labels, and then repeating these steps several times under different settings to optimize the model.

two-pizza team

A small DevOps team that you can feed with two pizzas. A two-pizza team size ensures the best possible opportunity for collaboration in software development.

U

uncertainty

A concept that refers to imprecise, incomplete, or unknown information that can undermine the reliability of predictive ML models. There are two types of uncertainty: *Epistemic uncertainty* is caused by limited, incomplete data, whereas *aleatoric uncertainty* is caused by the noise and randomness inherent in the data.

undifferentiated tasks

Also known as *heavy lifting*, work that is necessary to create and operate an application but that doesn't provide direct value to the end user or provide competitive advantage. Examples of undifferentiated tasks include procurement, maintenance, and capacity planning.

upper environments

See environment.

V

vacuuming

A database maintenance operation that involves cleaning up after incremental updates to reclaim storage and improve performance.

version control

Processes and tools that track changes, such as changes to source code in a repository.

VPC peering

A connection between two VPCs that allows you to route traffic by using private IP addresses. For more information, see What is VPC peering in the Amazon VPC documentation.

vulnerability

A software or hardware flaw that compromises the security of the system.

W

warm cache

A buffer cache that contains current, relevant data that is frequently accessed. The database instance can read from the buffer cache, which is faster than reading from the main memory or disk.

warm data

Data that is infrequently accessed. When querying this kind of data, moderately slow queries are typically acceptable.

window function

A SQL function that performs a calculation on a group of rows that relate in some way to the current record. Window functions are useful for processing tasks, such as calculating a moving average or accessing the value of rows based on the relative position of the current row.

workload

A collection of resources and code that delivers business value, such as a customer-facing application or backend process.

workstream

Functional groups in a migration project that are responsible for a specific set of tasks. Each workstream is independent but supports the other workstreams in the project. For example, the portfolio workstream is responsible for prioritizing applications, wave planning, and collecting migration metadata. The portfolio workstream delivers these assets to the migration workstream, which then migrates the servers and applications.

WORM

See write once, read many.

WQF

See AWS Workload Qualification Framework.

write once, read many (WORM)

A storage model that writes data a single time and prevents the data from being deleted or modified. Authorized users can read the data as many times as needed, but they cannot change it. This data storage infrastructure is considered immutable.

Z

zero-day exploit

An attack, typically malware, that takes advantage of a zero-day vulnerability.

zero-day vulnerability

An unmitigated flaw or vulnerability in a production system. Threat actors can use this type of vulnerability to attack the system. Developers frequently become aware of the vulnerability as a result of the attack.

zero-shot prompting

Providing an LLM with instructions for performing a task but no examples (*shots*) that can help guide it. The LLM must use its pre-trained knowledge to handle the task. The effectiveness of zero-shot prompting depends on the complexity of the task and the quality of the prompt. See also few-shot prompting.

zombie application

An application that has an average CPU and memory usage below 5 percent. In a migration project, it is common to retire these applications.