Replatform options for Oracle Database on AWS

AWS Prescriptive Guidance

AWS Prescriptive Guidance: Replatform options for Oracle Database on AWS

Table of Contents

Replatform options for Oracle Database on AWS

Song Hu, Mohit Montu, and Rajeev Pardipuram, Amazon Web Services (AWS)

September 2024 (document history)

Migrating Oracle Database from on premises to Amazon Web Services (AWS) requires an optimal strategy. There are seven common strategies, which are often referred to as the 7 Rs:

- Retire
- Retain
- Rehost
- Relocate
- Repurchase
- Replatform
- Refactor/re-architect

Choosing the optimal migration strategy depends on business need, resource requirements, time, and financial constraints. For more information, see Overview of the 7 Rs of Database Migration and Determining the R Type for Migration.

Replatform is a good candidate to move workloads quickly to AWS. It offers the following benefits:

- Introduces some level of optimization to take advantage of cloud capabilities
- Reduces the amount of time spent on managing the database
- Avoids the need to redesign or rewrite applications

After you select replatform as your migration strategy, the next steps are to evaluate different options that are available and to select the most appropriate one.

This guide walks through different options available for replatforming Oracle databases by using Amazon Relational Database Service (Amazon RDS). The guide discusses advantages and limitations that you can compare against your specific workload to identify the most appropriate approach for your business. The comparison is divided into three phases based on the AWS migration process best practices:

- Phase 1: Assess
- Phase 2: Mobilize
- Phase 3: Migrate and modernize

This guide is intended for database administrators, solutions architects, and operations or infrastructure managers who are planning to migrate on-premises Oracle databases to AWS.

Objectives

Replatforming Oracle Database to AWS provides the following benefits.

Balance of risk and improvement

Replatforming is more cost-effective, faster, and carries less risk than refactoring. It also enhances automation and improves application performance, security, and scalability more than rehosting.

Lower cost

Replatforming provides flexibility in payment options offered by AWS, which are pay-as-you-go, On-Demand Instances, and Reserved Instances. AWS provides various levels of discount based on use cases, and you pay only for what you use, which can reduce both fixed and variable costs

For Oracle Database Standard Edition 2 (SE2), AWS also provides the License Included model with Amazon RDS. The price includes Oracle licenses as part of a pay-as-you-go subscription model, and you don't need to purchase the licenses separately.

When running Oracle workloads on AWS, the Amazon RDS instance size can be scaled up and down dynamically according to load fluctuation. This can further reduce cost because you can provision compute power as needed.

For more information about pricing, see Amazon RDS for Oracle pricing.

Enhanced automation

Replatforming provides a higher level of automation on maintenance tasks, such as backup, storage scaling, logging, and monitoring, which minimizes human errors. Staff productivity can also be improved by focusing on more valuable tasks, such as business development, performance tuning, and schema optimization.

Increased agility

Provisioning Oracle databases in an on-premises environment is time-consuming and can take weeks to months. By replatforming to AWS, you can complete the same task within minutes to a

couple of hours. Replatforming also gives you the flexibility to delete a full stack of database when it's no longer needed, and to stop paying for it. That isn't an option in an on-premises environment.

Better cloud maturity

Replatforming helps align with a cloud-first approach and grows cloud maturity over time. It builds the foundation for future database and application modernization by doing the following:

- Offloading unstructured data to Amazon Simple Storage Service (Amazon S3)
- Migrating data warehouse functions to Amazon Redshift
- Migrating transactional functions to open source database engines such as Amazon Aurora PostgreSQL-Compatible Edition or Amazon Aurora MySQL-Compatible Edition to save licensing cost and reduce operational overhead

Replatforming options for Oracle Database

When you replatform Oracle Database from on premises to managed database services on AWS, you have the following options:

- Amazon RDS for Oracle
- Amazon RDS Custom for Oracle

The following sections list important features of these options.

Amazon RDS for Oracle

Amazon RDS for Oracle is a managed database service that simplifies the provisioning and management of Oracle databases on AWS. It has the following key advantages:

- Provides a console for setting up, operating, managing and scaling Oracle database deployments.
- Automates time-consuming database administration tasks, including provisioning, software patching, monitoring, hardware scaling, and failure detection.
- Automates the backup and recovery process reliably and efficiently.
- Provides high availability with a Multi-AZ deployment. The primary instance and a synchronous secondary instance can be used to switch over during planned events, and automatically fail over during unplanned events.
- Provides provisioning of read replica databases to enhance availability, performance, and reliability.

Amazon RDS for Oracle supports both Enterprise Edition (EE) and Standard Edition 2 (SE2). Oracle Database EE offers enterprise-level features but is significantly more expensive than SE2 in license cost. It also requires the Bring Your Own License (BYOL) licensing model. Applications with no or minimum usage of EE features are good candidates for downgrading to Oracle Database SE2 to reduce total cost of ownership (TCO). For more information, see Evaluate downgrading Oracle databases to Standard Edition 2 on AWS.

Amazon RDS Custom for Oracle

Amazon RDS Custom for Oracle is a managed database service for legacy, custom, and packaged applications that require access to the database administrator privilege and underlying operating system. It has the following key features:

- Automates setup, operation, and scaling of databases in the AWS Cloud
- Provides root access to the operating system of the underlying Amazon Elastic Compute Cloud (Amazon EC2) instance, and database access as built-in SYS and SYSTEM user
- Provides the capability to configure settings, install patches, and enable native features manually to meet the dependent application and database requirements
- Provides support on legacy Oracle Database versions (12.1, 12.2, and 18c)

Phase 1: Assess

The assess phase focuses on collecting and analyzing information about the source Oracle database. It's the fundamental part of migration because all subsequent phases are based on the data points that are collected in this phase. The analysis result of this phase is the input of the remaining phases. It determines the most appropriate choice for the replatforming option, migration tool, and approach.

You can use the following tools to assess the source Oracle database when preparing for migration to AWS.

Oracle diagnostic support scripts

Oracle diagnostic support scripts analyze an on-premises Oracle database. These scripts have the following characteristics:

- Oracle diagnostic scripts are all written to run using the SQL*Plus command-line utility. A user account with permissions to query Oracle dictionary views is necessary to prepare the report.
- The scripts collect information related to the Oracle database configuration and database objects.
- The scripts produce an HTML report of multiple sections that include database size, schema size, large binary object (LOB) information, redo log, and archive log information.
- The report can assist with deciding the migration strategy.

Oracle Automatic Workload Repository

Oracle Automatic Workload Repository (AWR) is a native Oracle tool with the following characteristics:

- Oracle AWR collects, processes, and maintains database performance statistics.
- This information is gathered at regular intervals or on-demand. It can be displayed in both reports and views.
- AWR produces reports of CPU, memory, I/O, and other critical information. The reports help you to understand the nature of the workload running in the database and the resources required in the AWS Cloud.

Collecting statistics

By using these tools, you can collect statistics about Oracle Database configuration, usage, and performance. To achieve a successful migration, you also need to understand the complexity, compatibility, and dependency of the database. This includes information about the operating system, network, application, and business requirements.

The following list contains the most common preparation tasks:

- Identify the recovery time objective (RTO), recovery point objective (RPO), and service-level agreement (SLA) requirements for the Oracle database.
- Check the network connectivity between the on-premises environment and AWS. Make sure that it provides enough bandwidth for fast transfers of data between on premises and AWS.
- Determine the amount of downtime available for migration. This helps you to choose an online or an offline migration approach.
- Check the chipset endian platform of the database workload. AWS supports x86-x64 little-endian platforms. Other platforms, such as Sun SPARC, HP Tru64, or IBM Z series-based big-endian platforms, require cross-platform migration.
- AWS supports Linux (32-bit and 64-bit) and Windows operating systems. It doesn't support Solaris, HP-UX, or IBM AIX operating systems, which are commonly used for Oracle databases. Migrating Oracle databases from these operating systems requires platform conversion.
- Review current architecture and auditing or compliance needs to make sure that all requirements can be met after migrating to AWS.
- Understand the limitations of different replatforming options:
  - Check the edition and version of the Oracle database software to make sure it's supported.
  - Determine the input/output operations per second (IOPS) and throughput of the database.
  - Check the current database size and storage growth pattern.
- If you are migrating Oracle Database Enterprise Edition, identify which Enterprise Edition features are actually used by the application. This is important when evaluating the option of downgrading Enterprise Edition to Standard Edition 2 (SE2).
- Collect details of the current license agreement for Oracle databases.
- Check for application dependencies. If the Oracle database supports legacy, custom, or packaged applications, the application will need access to the database administrator privilege and underlying operating system.

Phase 2: Mobilize

In the mobilize phase, you determine the most appropriate replatform option for your specific Oracle database. You evaluate all replatforming options against the data that you collected during the assess phase. The evaluation process compares all options from many different aspects.

The topics in this section detail each item, and the data is consolidated in a comparison table at the end. The comparison table lists key differences in a multidimensional view to assist you with making the final decision.

Topics

Oracle license

On AWS, there are two licensing models for running Oracle databases:

- Bring Your Own License (BYOL)
- License Included

Under the BYOL model, you can use your existing on-premises Oracle Database licenses on AWS. To run a DB instance under the BYOL model, you must have the appropriate licenses for software and support. Under this model, you continue to use your active Oracle Support account, and you contact Oracle Support directly for Oracle database-specific service requests. If you have an

active AWS Support account, you can contact AWS Support for issues with infrastructure, such as operating system, storage, network, and hardware.

In the License Included model, you don't need to purchase Oracle licenses separately. The Oracle database software has been licensed by AWS. In this model, AWS Support, if purchased and active, can be contacted for both Amazon RDS service requests and Oracle Database–specific service requests.

Another advantage of the License Included model is that cost is incurred only for the hours that the database is running. This is especially cost-effective for non-production environments where databases don't need to run all day, every day.

The License Included model is supported only on Amazon RDS for Oracle Database SE2. It isn't available on Amazon RDS Custom for Oracle.

License model	Amazon RDS for Oracle	Amazon RDS Custom for Oracle
Bring Your Own License	Yes	Yes
License Included (SE2 only)	Yes	No

Editions and versions

In addition to choosing your licensing model, you need to choose the edition that supports your database requirements. Amazon RDS for Oracle supports the following options:

- Oracle Database Enterprise Edition (EE) is the most common choice for production workloads in many large organizations and enterprises. EE provides key features for mission-critical applications, including Active Data Guard and Oracle Partitioning.
- Oracle Database Standard Edition 2 (SE2) is an affordable database and supports a variety of use cases, from small business single-server environments to highly distributed branch environments. SE2 can be licensed on servers with a maximum of two sockets. However, the core counts per two-socket server can increase over time without impacting your license obligation.

With Oracle Database SE2, your license costs remain the same regardless of the number of cores in the socket. Currently AWS supports up to 16 virtual CPUs (vCPUs).

From the aspect of license cost, Oracle Database SE2 is much less expensive than EE. If your application uses very few or none of the EE features, consider downgrading from EE to SE2. For more information, see the Evaluate downgrading Oracle databases to Standard Edition 2 on AWS guide.

For more details about the availability of features, options, and management packs in different editions, see the Oracle documentation.

Each replatform option supports different Oracle editions. The following table lists the latest support information.

Editions and versions	Amazon RDS for Oracle	Amazon RDS Custom for Oracle
Enterprise Edition	Yes	Yes
Standard Edition 2	Yes	Yes
Versions	19c 21c	12.1.0.2 12.2.0.1 18c 19c

Oracle multi-tenant architecture

Oracle multi-tenant architecture enables an Oracle database to function as a container database (CDB), and includes multiple pluggable databases (PDBs). By consolidating multiple Oracle databases into a single instance, the multi-tenant architecture reduces both cost and management efforts:

- Both Amazon RDS for Oracle and Amazon RDS Custom for Oracle support multi-tenant architecture in Oracle Database Enterprise Edition (EE) and Standard Edition 2 (SE2).
- Amazon RDS for Oracle supports Oracle multi-tenant architecture in versions 19c and 21c. Amazon RDS Custom for Oracle supports the architecture in version 19c only.
- Amazon RDS for Oracle also supports Oracle single-tenant architecture in versions 19c and 21c. Amazon RDS Custom for Oracle currently does not support single-tenant architecture.
- With EE, an Amazon RDS for Oracle CDB instance supports up to 30 PDBs, depending on the licenses. Amazon RDS Custom for Oracle doesn't restrict the number of PDBs that you can create.
- In SE2, both Amazon RDS for Oracle and Amazon RDS Custom for Oracle support up to 3 PDBs per CDB.

For more information, see the AWS documentation for Amazon RDS for Oracle and Amazon RDS Custom for Oracle.

Tenancy configuration	Amazon RDS for Oracle	Amazon RDS Custom for Oracle
Multi-tenant–supported edition	EE & SE2	EE & SE2
Multi-tenant–supported version	19c, 21c	19c
Single-tenant architecture	Yes	No
Multi-tenant architecture	Yes	Yes
Number of PDBs per CDB in EE	Up to 30	No restriction

Tenancy configuration	Amazon RDS for Oracle	Amazon RDS Custom for Oracle
Number of PDBs per CDB in SE2	Up to 3	Up to 3

Resource availability

Your replatforming choice might depend on the AWS Region you are using and the resources required by your business. Both Amazon RDS for Oracle and Amazon RDS Custom for Oracle use AWS services, but not all of the services are available in all AWS Regions. AWS services also vary in supported engine versions and instance classes. Amazon RDS for Oracle provides more choices in AWS Regions and instance classes than Amazon RDS Custom for Oracle. This is because Amazon RDS Custom for Oracle is still in the process of expanding.

It's also important to consider scaling needs. The AWS BYOL model is based on CPU cores. After you create an Amazon RDS for Oracle instance, you cannot change the DB instance class to a different number of cores unless the change is agreed to by the Oracle license policy. However, the AWS License Included model gives you the flexibility to dynamically change the number of cores by scaling the instance class up and down.

Resource availability	Amazon RDS for Oracle	Amazon RDS Custom for Oracle
AWS Region	Most	Limited
DB instance class	Most	Limited
CPU scalability	License Included model	Not available

Storage capacity

Amazon RDS for Oracle supports the following AWS storage types:

- General Purpose solid-state drive (SSD): gp2, gp3
- Provisioned IOPS SSD: io1, io2
- Magnetic

The storage types differ in performance characteristics and price. You can tailor storage performance and cost to the needs of your database workload.

Amazon RDS Custom for Oracle supports SSD storage type gp2, gp3 and io1. Magnetic storage isn't supported.

The maximum IOPS and throughput per RDS instance depends on the selected storage type and instance class. For more information, see Amazon RDS DB instance storage.

Amazon RDS for Oracle provides storage autoscaling that can automatically scale storage capacity in response to growing database workloads, with zero downtime. Amazon RDS storage autoscaling continuously monitors storage consumption. Capacity scales up automatically when actual utilization approaches provisioned storage capacity. There is no additional cost for enabling the storage autoscaling feature. You pay only for the storage that is provisioned.

Amazon RDS Custom for Oracle doesn't support storage autoscaling. You must manually provision storage.

Storage features	Amazon RDS for Oracle	Amazon RDS Custom for Oracle
Storage type	All	gp2, gp3, io1
Maximum storage size	64 TiB	64 TiB
Maximum IOPS per instance	256,000	256,000

Storage features	Amazon RDS for Oracle	Amazon RDS Custom for Oracle
Maximum throughput per instance	16,000 MiB/s	4,000 MiB/s
Storage autoscaling	Yes	No

Privileged access

Amazon RDS for Oracle is fully managed. To deliver a managed service experience, it does not allow access to the underlying host, and it restricts access to some procedures and objects that require high-level privileges.

Amazon RDS Custom for Oracle grants access to the database administrator privilege and underlying operating system. You can perform operations as root user at the operating system level, and as SYS or SYSTEM user at the database level. For legacy, custom, and packaged applications, you can customize the operating system and Amazon RDS Custom for Oracle database environment by doing the following:

- Install a custom database, and operating system patches, and packages.
- Configure specific database settings.
- Configure file systems to share files directly with their applications.

Access	Amazon RDS for Oracle	Amazon RDS Custom for Oracle
Access to operating system	No	Yes
Access to built-in Oracle users (for example,	No	Yes

Access	Amazon RDS for Oracle	Amazon RDS Custom for Oracle
SYS, SYSTEM)		

Patching and upgrading

One benefit of Amazon RDS for Oracle is the ease of maintenance. AWS does all the undifferentiated heavy lifting work behind the scenes so that your attention can be on applications and users. You can enable maintenance options during setup. Amazon RDS for Oracle will then automatically apply operating system (OS) patching, Oracle database patching, and minor database version upgrades in a predefined maintenance window.

With Amazon RDS Custom for Oracle, because you have database administrator privileges and operating system root access, you're responsible for patching and upgrade activities instead of AWS.

Responsibility	Amazon RDS for Oracle	Amazon RDS Custom for Oracle
Automatic OS patching	Yes	No
Automatic Oracle patching	Yes	No
Automatic minor Oracle version upgrades	Yes	No

High availability

Amazon RDS for Oracle supports Multi-AZ deployment, which automatically creates a standby instance in a different Availability Zone. A Multi-AZ deployment provides automatic failover during planned maintenance and unplanned disruptions.

Amazon RDS Custom for Oracle doesn't support Multi-AZ deployment. As an alternative, you can use a replica to manually build a high availability (HA) solution. Depending on the design, you can implement both synchronous and asynchronous data replication. For more information, see the Build high availability for Amazon RDS Custom for Oracle using read replicas blog post.

Both Amazon RDS for Oracle and Amazon RDS Custom for Oracle support up to five managed read replicas. You can automatically create the read replicas from the AWS Management Console or by using AWS Command Line Interface (AWS CLI).

With Amazon RDS Custom for Oracle, you can also create your own manually configured external Oracle replicas. This offers you the flexibility to host replicas on Amazon EC2 instances in the same or another AWS Region, and also in an on-premises environment. External replicas don't count toward the instance limit within the AWS account. They also lie outside the RDS Custom support perimeter. For more information about the support perimeter, see RDS Custom support perimeter.

HA support	Amazon RDS for Oracle	Amazon RDS Custom for Oracle
Multi-AZ deployment	Yes	No
Standby replication	Synchronous	Asynchronous or synchronous
AWS managed automatic failover	Yes	No
Automatic creation of read replica	Yes	Yes
Maximum managed read replicas	5	5
AWS managed	Yes	No

HA support	Amazon RDS for Oracle	Amazon RDS Custom for Oracle
cross-Region read replica		
Modification of AWS managed read replica	No	Yes
Creation of self-managed read replica	No	Yes

Backup and recovery

Amazon RDS for Oracle and Amazon RDS Custom for Oracle both provide automatic backup and point-in-time recovery (PITR), which are benefits of managed services. When Multi-AZ deployment is enabled in Amazon RDS for Oracle, backup is automatically taken from the standby instance, and there is no I/O impact to the primary instance.

Amazon RDS Custom for Oracle does not support Multi-AZ deployment, and automatic backup takes place on the primary instance.

Backup and recovery options	Amazon RDS for Oracle	Amazon RDS Custom for Oracle
Automatic backup	Yes	Yes
Automatic PITR	Yes	Yes
Automatic backup from	Yes	No

Backup and recovery options	Amazon RDS for Oracle	Amazon RDS Custom for Oracle
standby instance		

Performance monitoring

AWS provides a number of features and services to monitor the performance of Oracle database instances. They cover a variety of aspects from the hypervisor level, to the operating system, to inside the database.

Automatic monitoring

Amazon RDS for Oracle and Amazon RDS Custom for Oracle both provide automatic monitoring at the hypervisor level. By default, Amazon RDS automatically sends metrics data to Amazon CloudWatch in 60-second periods. Data points are available for 15 days.

Enhanced Monitoring

Enhanced Monitoring for Amazon RDS provides deeper visibility into operating system metrics and process information. You can configure it to collect at an interval of 1, 5, 10, 15, 30, or 60 seconds. The information can be visualized on the AWS Management Console, and you can customize the metrics and dashboard specifically to your business needs. For more information, see OS metrics in Enhanced Monitoring and Amazon RDS FAQs for Enhanced Monitoring.

Enhanced Monitoring is currently not supported on Amazon RDS Custom for Oracle.

Performance Insights

Performance Insights expands Amazon RDS monitoring features even further inside the database instance to help you analyze your database performance. With the Performance Insights dashboard, you can visualize the Oracle database load and filter by waits, SQL statements, hosts, or users. For more information, see Monitoring DB load with Performance Insights on Amazon RDS.

Amazon RDS Custom for Oracle does not support Performance Insights.

Oracle Enterprise Manager

Oracle Enterprise Manager (OEM) is the Oracle native monitoring solution. It uses Management Agent running on the database host to push database monitoring and performance metrics data to a centralized Oracle Manager Server (OMS). It's your responsibility to install, configure, and manage the entire OEM system.

Both Amazon RDS for Oracle and Amazon RDS Custom for Oracle support installing OEM Management Agent.

Performance monitoring options

The following table compares performance monitoring options for Amazon RDS for Oracle and Amazon RDS Custom for Oracle.

Performance monitoring option	Amazon RDS for Oracle	Amazon RDS Custom for Oracle
Automatic monitoring	Yes	Yes
Enhanced Monitoring	Yes	No
Performance Insights	Yes	No
OEM Management Agent	Yes	Yes

Mobilize phase comparison table

Based on the thorough analysis, Amazon RDS for Oracle and Amazon RDS Custom for Oracle are similar in many ways but different in some areas.

The following table lists the key differences between Amazon RDS for Oracle and Amazon RDS Custom for Oracle. The table provides you with a comprehensive summary for the entire evaluation process to help you make a final decision.

Feature	Amazon RDS for Oracle	Amazon RDS Custom for Oracle
License Included (SE2 only)	Yes	No
Version	19c 21c	12.1.0.2 12.2.0.1 18c 19c
Multi-tenant supported version	19c, 21c	19c
Single-tenant configuration	Yes	No
Number of PDBs per CDB in EE	Up to 30	No restriction
AWS Region	Most	Limited
DB instance class	Most	Limited
CPU scalability	License Included model	Not available
Storage type	All	gp2, gp3, io1

Feature	Amazon RDS for Oracle	Amazon RDS Custom for Oracle
Maximum throughput per instance	16,000 MiB/s	4,000 MiB/s
Automatic storage scaling	Yes	No
Access to operating system	No	Yes
Access to built-in Oracle users (for example, SYS, SYSTEM)	No	Yes
Automatic operating system patching	Yes	No
Automatic Oracle Database patching	Yes	No
Automatic Oracle Database minor-version upgrade	Yes	No

Feature	Amazon RDS for Oracle	Amazon RDS Custom for Oracle
Automatic backup from standby database	Yes	No
Multi-AZ deployment	Yes	No
Standby replication	Synchronous	Asynchronous or synchronous
AWS managed automatic failover	Yes	No
AWS managed cross-Region read replica	Yes	No
Modification of AWS managed read replica	No	Yes
Creation of self-managed read replica	No	Yes
Enhanced Monitoring	Yes	No
Performance Insights	Yes	No

Phase 3: Migrate and modernize

In the migrate and modernize phase, you perform the actual database migration by using suitable tools and the appropriate approach. Your selection of tools and approach is based on business requirements such as the following:

- Timeline
- Database size
- Consistency
- Available network bandwidth between the on-premises environment and AWS
- Afforded downtime

The following topics evaluate different migration tools and approaches that are available to help you decide which ones to use.

Topics

After you successfully migrate your Oracle database to AWS, you can continue to modernize your database by migrating to open source database engines, such as Amazon Aurora PostgreSQL-Compatible Edition or Amazon Aurora MySQL-Compatible Edition. For more information about database modernization, see AWS Prescriptive Guidance for re-architecting Oracle Database.

Migration tools

The following tools are listed in order of logical migration to physical migration.

Oracle Data Pump

Oracle Data Pump is a native tool that comes with Oracle Database. It provides the ability to export and import data and metadata from or to Oracle databases. You can use Oracle Data Pump at the database, tablespace, schema, and object level. Oracle Data Pump supports flexible data extraction options, parallelism, compression, and encryption.

Oracle Data Pump is commonly used to migrate Oracle databases because it provides a high level of compatibility. Oracle Data Pump is an especially suitable option for migrations to different database editions, versions, and endian platforms. Oracle Data Pump is also often used along with other tools, such as AWS Database Migration Service (AWS DMS) and Oracle Recovery Manager (Oracle RMAN), to build comprehensive solutions for complex use cases.

AWS DMS

AWS Database Migration Service (AWS DMS) is a managed service that helps move data to AWS securely. AWS DMS provides both one-time full database copy and change data capture (CDC) technology. The CDC feature can keep the source and target database in sync and minimize downtime during the migration. To migrate large databases, you can use AWS DMS together with other AWS services, such as Amazon S3, AWS Direct Connect, or AWS Snow Family devices.

Oracle GoldenGate

Oracle GoldenGate is a tool that Oracle offers to collect, replicate, and manage transactional data between databases. It provides CDC by interpreting Oracle database transaction logs. Similar to AWS DMS, Oracle GoldenGate is a common option for migrating Oracle Database. For more information, see Using Oracle GoldenGate with Amazon RDS for Oracle.

Oracle GoldenGate is not part of Oracle Database and requires a separate license from Oracle.

Oracle Recovery Manager

Oracle Recovery Manager (RMAN) is a tool provided by Oracle to perform and manage Oracle database backups and restorations. You can use RMAN to back up an Oracle database from on premises and then restore it to an Oracle instance on AWS. RMAN is a physical-level tool that works on data files and log files instead of schemas and objects.

You can use Oracle RMAN with Amazon RDS Custom for Oracle. RMAN is usually combined with other AWS services, such as Direct Connect, AWS DataSync, and Amazon S3, to form an end-to-end migration solution.

Oracle Data Guard

Oracle Data Guard is a built-in feature of Oracle Database that maintains a physical copy of the database and keeps it in sync. It provides the capability to switch over the roles between primary and standby databases, which can minimize downtime during the migration.

Oracle Data Guard can't be directly used with Amazon RDS for Oracle or Amazon RDS Custom for Oracle for migration. Instead, Oracle Data Guard is usually used with AWS services such as Amazon EC2, Direct Connect, or AWS DMS to build a complete migration solution. For example, you can build a physical standby on an EC2 instance using Oracle Data Guard. Then you can use AWS DMS or Oracle Data Pump to migrate data to the target RDS for Oracle instance.

Migration approach

There are two approaches for migrating Oracle Database from on premises to AWS: offline migration and online migration.

Offline migration

You can use the offline migration approach when your application can afford a planned downtime. In this approach, the source database is taken offline at the beginning of the migration period, and then it's migrated over to the target database on AWS. After the migration is complete, validation and verification checks are performed in the target database to ensure data consistency. When all checks are passed, you perform a cutover by connecting the application to the target database.

Offline migration usually consists of fewer steps, has simpler architecture, and is more cost effective.

Online migration

Use the online migration approach when your application requires minimal or near zero downtime. In this scenario, the source database is migrated in multiple steps to AWS. Initially, the data in the source database is copied to the target database while the source database is still running. In subsequent steps, all changes from the source database are propagated to the target database online. When the source and target databases are in sync, they are ready for cutover. During the cutover, the application switches its connections over to the target database, leaving no connections to the source database.

Online migration achieves less downtime, but it requires more steps, resources, and effort, and it's more expensive.

Migration phase comparison table

The following table provides a summary of suitable migration scenarios for each tool to help you choose the option that best meets your business requirements.

Tool	Online migration	Offline migration	Amazon RDS for Oracle	Amazon RDS Custom for Oracle
Oracl Data Pump	No	Yes	Yes	Yes
AWS DMS	Yes	Yes	Yes	Yes
Oracl Golde e	Yes	No	Yes	Yes
Oracl Recov Mana (RMA	No	Yes	No	Yes
Oracl Data Guard	Yes	No	No	No

Next steps

Now that you have selected replatform as your migration strategy and determined the appropriate migration tools and approach, you are ready for the final migration. AWS provides a collection of prescriptive guidance developed by AWS technology experts and the global community of AWS Partners. The guidance provides step-by-step instructions for database migration. For more information, see AWS Prescriptive Guidance for Oracle Database. You can use the filters in the left panel to further narrow the results based on your criteria.

Resources

References

- Amazon RDS for Oracle
- Amazon RDS Custom
- AWS DMS documentation
- Using an Oracle database as a source for AWS DMS

Guides and patterns

- Migration strategy for relational databases (strategy)
- Migrating Oracle databases to the AWS Cloud (guide)
- Evaluate downgrading Oracle databases to Standard Edition 2 on AWS (guide)
- Replatform Oracle Database Enterprise Edition to Standard Edition 2 on Amazon RDS for Oracle (pattern)
- Migrate an on-premises Oracle database to Amazon RDS for Oracle (pattern)
- Migrate an on-premises Oracle database to Amazon RDS for Oracle using Oracle Data Pump (pattern)
- Migrate an on-premises Oracle database to Amazon RDS for Oracle by using direct Oracle Data Pump Import over a database link (pattern)
- Migrate an Oracle database to Amazon RDS for Oracle by using Oracle GoldenGate flat file adapters (pattern)

Whitepapers

- Strategies for Migrating Oracle Databases to AWS
- Best Practices for Running Oracle Database on AWS

Document history

The following table describes significant changes to this guide. If you want to be notified about future updates, you can subscribe to an RSS feed.

Change	Description	Date
Initial publication	—	September 30, 2024

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

A2A (Agent-to-Agent)

A stateful protocol for agent-to-agent collaboration supporting task delegation and state transfer.

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.

Agent

An AI system that can autonomously reason, plan, and take actions using tools to achieve goals.

Agent Ops

Operational practices for building, testing, deploying, and running AI agents in production at scale.

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.

Citizen Developer

A business user who creates AI applications using no-code/low-code platforms without specialized technical skills.

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.

FM gateway

A centralized intermediary that controls and normalizes access to foundation models. Also known as an *LLM gateway*.

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.

guardrails (AI)

Safety mechanisms that filter, validate, and constrain agent inputs and outputs to help ensure responsible and safe AI behavior.

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.

human-in-the-loop (HitL)

A workflow pattern where agent execution pauses for human review and approval at critical decision points.

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.

MCP

See Model Context Protocol.

Model Context Protocol (MCP)

A stateless protocol for agent-to-tool communication.

MCP server

A service that exposes one or more tools through the Model Context Protocol.

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.

Shadow AI

Unauthorized AI applications built or used outside of governed channels within an organization.

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.

tool

A function or API that an agent can invoke to perform operations in external systems.

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.