Building a high availability and disaster recovery architecture with native and hybrid methods for Microsoft SQL Server databases on Amazon EC2

AWS Prescriptive Guidance

AWS Prescriptive Guidance: Building a high availability and disaster recovery architecture with native and hybrid methods for Microsoft SQL Server databases on Amazon EC2

Table of Contents

Introduction **1**
SQL Server on Amazon EC2 single-node architecture **2**
Instance types 4
Storage 5
Amazon EBS and Amazon S3 considerations 6
SQL Server on Amazon FSx for Windows File Server 8
HA/DR options and considerations **10**
Managing HA/DR resources in AWS Backup 11
Using AWS DMS for HA/DR 11
Using AWS Application Migration Service for DR 14
Additional considerations 14
Disaster recovery scenarios **16**
Availability Zone failure 16
Region failure 17
Common use cases 18
SQL Server on Amazon EC2 architecture diagrams **22**
Two-node HA/DR architecture with Always On availability group cluster (single-Region, Multi-AZ) 22
Three-node HA/DR architecture (single-Region, Multi-AZ) 23
Four-node HA/DR architecture with Always On distributed availability group cluster (multi-Region, Multi-AZ) 24
Three-node HA/DR architecture with single availability group (multi-Region) 25
Three-node HA/DR architecture with log shipping (multi-Region) 26
Restore options **27**
Using Amazon S3 27
Using AWS DataSync and Amazon FSx 28
Using Amazon S3 File Gateway 29
Next steps and resources **31**
Appendix: Amazon EBS SSD storage types **33**
Document history **35**
Glossary **36**
..... 36
A 37
B 40

C 42
D 45
E 49
F 51
G 53
H 54
I 55
L 58
M 59
O 63
P 66
Q 68
R 69
S 72
T 76
U 77
V 78
W 78
Z 79

Building a high availability and disaster recovery architecture with native and hybrid methods for Microsoft SQL Server databases on Amazon EC2

Ram Yellapragada and Alysia Tran, Amazon Web Services (AWS)

February 2022 (document history)

Microsoft SQL Server has many native options to support high availability (HA) and disaster recovery (DR), to help ensure business continuity for your database workloads. This guide outlines an ideal configuration for SQL Server on Amazon Elastic Compute Cloud (Amazon EC2) in the Amazon Web Services (AWS) Cloud. Rehosting SQL Server on Amazon EC2 provides a self-managed system where you can retain full control over database operations and configuration.

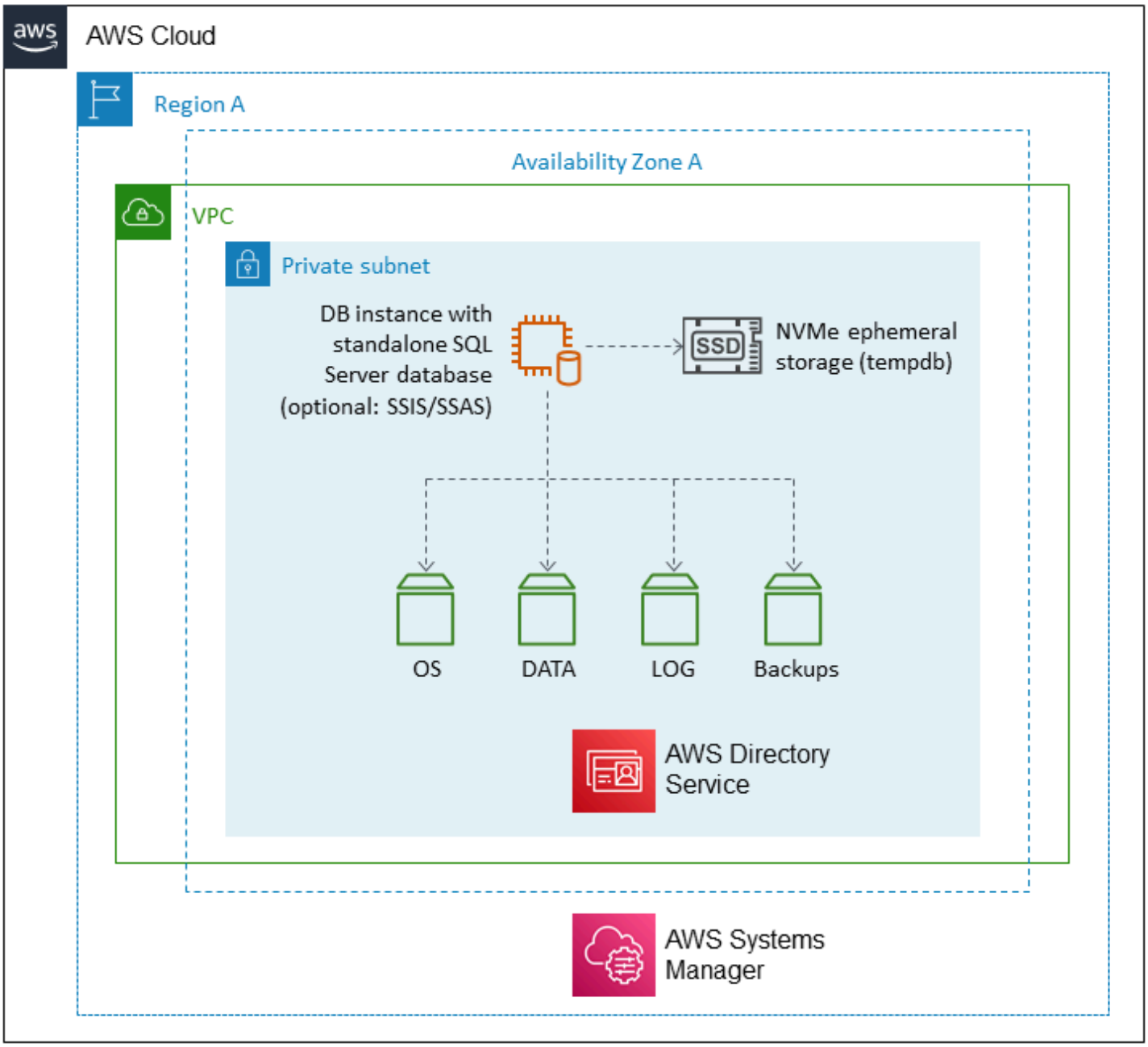
The guide discusses SQL Server hybrid HA/DR options that include various AWS services and infrastructure, and provides guidance on infrastructure components and settings, including instance classes, storage options, configuration, and HA/DR setup. This document also explains how a given HA/DR strategy might fit into an example use case that has specific recovery time objective (RTO) and recovery point objective (RPO) requirements, and covers a few recovery scenarios, including relevant architecture diagrams. This guide doesn't provide solutions designed for specific applications or requirements. It presents some HA/DR options based on RTO and RPO, so you can choose an architecture that matches your requirements.

In addition, as a sizing exercise, the guide defines HA/DR options for a typical SQL Server online transaction processing (OLTP) workload and provides a side-by-side comparison of these options. For a discussion on rehosting SQL Server on AWS, see the section Amazon EC2 for SQL Server in the guide *Migrating Microsoft SQL Server databases to the AWS Cloud*. For information about other migration options, see the section SQL Server database migration strategies in that guide. For additional reading, see the Next steps and resources section.

SQL Server on Amazon EC2 single-node architecture

The following diagram illustrates a recommended architecture for a single-node SQL Server on Amazon Elastic Compute Cloud (Amazon EC2) before adding support for high availability (HA) and disaster recovery (DR).

In this architecture, the SQL Server database is deployed to an EC2 instance, using an Amazon Machine Image (AMI) for SQL Server and separate volumes for OS, DATA, LOG, and backups. Non-volatile memory express (NVMe) storage is attached directly to the EC2 instance and used for the SQL Server tempdb database. AWS Directory Service is used to set up Windows authentication for the SQL Server database. You can also use AWS Systems Manager to detect and install SQL Server patches and updates.



The following table summarizes the recommendations for configuring this architecture. These recommendations are discussed in detail in the sections that follow.

Instance type/AMI	• Amazon Elastic Block Store (Amazon EBS)-optimized instance type for performance • NVMe for instance storage (temporary)

	• Amazon EC2 AMIs for SQL Server
SQL Server edition	• SQL Server Developer edition (non-production) • SQL Server Standard and Enterprise editions (production)
Storage type	• Amazon EBS • NVMe (tempdb) (gp2/io1/io2)
Volumes	• OS • DATA • LOG • tempdb • Scratch space for storing and downloading backups
DR options	• Amazon EC2 • Amazon EBS snapshots • SQL Server native backups

Instance types

AWS offers a selection of instance classes for your SQL Server workloads. You can choose among compute optimized, memory optimized, storage optimized, general purpose, and other types, depending on the expected workload on the database server, version, HA/DR options, cores required, and licensing considerations. We recommend that you choose Amazon EBS-optimized instance types for SQL Server. These offer the best throughput with attached EBS volumes in a dedicated network, which is critical for SQL Server workloads that might have heavy data access requirements. For standard database workloads, you can run memory optimized instance classes such as R5, R5b, R5d, and R5n. You can also include either instance storage or NVMe storage. These are both ideal for tempdb and offer balanced performance for database workloads.

For critical workloads, the high-performance z1d instance is optimized for workloads that carry high licensing costs, such as SQL Server. The z1d instance is built with a custom Intel Xeon Scalable

processor that delivers a sustained all-core turbo frequency of up to 4.0 GHz, which is significantly faster than other instances. For workloads that need faster sequential processing, you can run fewer cores with a z1d instance and get the same or better performance than other instances with more cores.

Amazon also provides dedicated AMIs for SQL Server on Microsoft Windows Server to help you host the latest SQL Server editions on Amazon EC2.

Storage

Some instance types offer NVMe instance store volumes. NVMe is a temporary (epheremal) storage option. This storage is directly attached to the EC2 instance. Although NVMe storage is temporary and data is lost on reboot, it offers the most optimal performance. Therefore, it is suitable for the SQL Server tempdb database, which has high I/O and random data access patterns. There is no additional charge for using an NVMe instance store for tempdb.

Amazon EBS is a durable storage solution that meets SQL Server's requirements for fast, available storage. Microsoft recommends keeping the data and log volumes separate for optimal performance. The reasons for this separation include the following:

- Different data access methods. Data volumes use online transaction processing (OLTP) random data access, whereas log volumes use serial access.
- Better recovery options. The loss of one volume doesn't affect the other volume, and helps in the recovery of data.
- Different workload types. Data volumes are for OLTP workloads, whereas log volumes target online analytic processing (OLAP) workloads.
- Different performance requirements. Data and log volumes have different IOPS and latency requirements, minimum throughput rates, and similar performance benchmarks.

To select the right Amazon EBS volume type, you should analyze your database access methods, IOPS, and throughput. Collect metrics both during standard working hours and during peak usage. SQL Server uses extents to store data. The atomic unit of storage in SQL Server is a page, which is 8 KB in size. Eight physically contiguous pages make up an extent, which is 64 KB in size. Therefore, on a SQL Server machine, the NTFS allocation unit size for hosting SQL database files (including tempdb) should be 64 KB.

The choice of EBS volume depends on the workload—that is, whether the database is read-intensive or write-intensive, requires high IOPS, archive storage, and similar considerations. The following table shows a sample configuration.

Amazon EBS resource	Type	Description
OS disk	`gp3`	General-purpose storage.
DATA disk	`io1/io2`	Write-intensive storage.
LOG disk	`gp3` or `io2`	General-purpose storage for intensive workloads.
Backup disk	`st1`	Less expensive archive storage. For better performance, backups can also be stored on a faster disk if they're copied to Amazon Simple Storage Service (Amazon S3) regularly.

Amazon EBS and Amazon S3 considerations

The following table shows a comparison of Amazon EBS and Amazon S3 for storage. Use this information to understand the differences between the two services and to choose the best approach for your use case.

Service	Availability	Durability	Notes
Amazon EBS	• All EBS volume types offer durable snapshot capabilities and are designed for 99.999% availability.	• EBS volume data is replicated across multiple servers in a single Availability Zone to prevent the loss of data from the failure	• An Amazon EBS–optimized instance uses an optimized configuration stack and provides additional, dedicated

Service	Availability	Durability	Notes
	• You can use snapshots to provision new instances in different AWS Regions in case of a disaster.	of any single component. • EBS volumes are designed for an annual failure rate (AFR) of between 0.1 and 0.2 percent, where failure refers to a complete or partial loss of the volume, depending on the size and performan ce of the volume.	bandwidth for Amazon EBS I/O. This optimization provides the best performance for your EBS volumes by minimizin g contention between Amazon EBS I/O and other traffic from your instance. • Fast snapshot restores are supported for up to 50 snapshots at the same time. You must enable this feature explicitly on a per-snapshot basis. • An Amazon EBS–optimized instance offers full provision ed performance at initialization, so no warmup time is involved.

Service	Availability	Durability	Notes

Amazon S3	• Highly available.
	• Designed for 99.99% availability over a given year.
	• Multiple storage classes are available, such as S3 Standard and S3 Standard-Infrequent Access ((S3 Standard-IA)). You can move backup files to a storage class based on a retention period.
		• Amazon S3, Amazon Glacier, and S3 Glacier Deep Archive are designed for 99.999999 999% (11 nines) of durability. Both Amazon S3 and Amazon Glacier offer reliable backup of data, with object replication across at least three, geographically dispersed Availabil ity Zones.	• You can use Amazon S3 for long-term SQL Server file-level backups (includin g full backups and transaction logs).
			• Amazon S3 supports:
			  • Replication time control (RTC)
			  • Cross-Reg ion replicati on through S3 Lifecycle management and AWS Backup
			  • Intelligent tiering
			• Amazon S3 provides the least expensive storage. Cross-Region data transfer costs apply.

SQL Server on Amazon FSx for Windows File Server

Amazon FSx for Windows File Server provides fast performance with baseline throughput up to 2 GB/second per file system, hundreds of thousands of IOPS, and consistent sub-millisecond latencies. To provide the right performance for your SQL Server instances, you can choose a throughput level that is independent of your file system size. Higher levels of throughput capacity also come with higher levels of IOPS that the file server can serve to the SQL Server instances

accessing it. The storage capacity determines not only how much data you can store, but also how many I/O operations per second (IOPS) you can perform on the storage—each GB of storage provides 3 IOPS. You can provision each file system to be up to 64 TiB in size (compared with 16 TiB for Amazon EBS). You can also use Amazon FSx systems as a file share witness for Windows Server Failover Cluster (WSFC) deployments.

HA/DR options and considerations

Although the possibility of an AWS Availability Zone or Region going completely offline is extremely rare, we recommend a multi-pronged approach to backup and recovery in the event of a disaster for redundancy and to minimize data loss. Backup and recovery processes should include the appropriate level of granularity to meet the the recovery time objective (RTO) and recovery point objective (RPO) for the workload and to support business processes, and are often dependent on the application. In the case of databases, AWS also supports all Microsoft recommendations for SQL Server setup and configuration for high availability and disaster recovery (HA/DR). Different editions of SQL Server support various HA/DR options, and you should consider special cases such as very large databases (VLDBs) on a case-by-case basis. As with any DR configuration, testing is essential to ensure that each application meets its service-level agreements (SLAs) for HA/DR. For your test/development environment, consider using SQL Server Developer edition, which is free but comes with limitations.

For a use case that requires an RPO of 15 minutes and an RTO of 4 hours, you can consider a combination of the following HA/DR options:

- **SQL Server native HA/DR options with a warm standby (database level)** – For illustrations of some of these architectures, see the SQL Server on Amazon EC2 architecture diagrams section later in this guide.
  - Two-node, Multi-AZ in a single Region (synchronous-commit mode) or in multiple Regions (asynchronous-commit mode, basic availability group)
  - Three-node (or more), Multi-AZ in multiple Regions (synchronous-commit and asynchronous-commit modes)
  - Two-node, Multi-AZ and log shipping in multiple Regions (with log backups every 5 minutes)
- **SQL Server native backups to Amazon S3 (database level, DR only)** – Full backups (one time daily)
  - Differential backups (every 2–4 hours).
  - Log backups (every 5–10 minutes).
  - Backups need to be taken and copied to Amazon Simple Storage Service (Amazon S3) by using custom scripting or an option such as a File Gateway for efficient backup and transfer.
  - If you have hundreds of databases, you can continue to use your existing backup tools (such as Commvault or Litespeed) to efficiently manage backups and store them directly in Amazon S3.

- Use Amazon S3 Cross-Region Replication (CRR) with S3 Replication Time Control (RTC) to control and monitor object replication within an SLA of 15 minutes.
- For compliance and cost savings, you can also use S3 Lifecycle management to move and store older backups for long-term storage.
- If you take SQL Server native backups and move them to Amazon S3 regularly, in the event of a disaster, backups will be readily available in the target Region. This eliminates the need to transfer backups or restore snapshots.
- We recommend using SQL Native Backup Compression to reduce file sizes.

- **AWS snapshots (instance and volume level, DR only)**
  - Amazon Elastic Compute Cloud (Amazon EC2) Amazon Machine Image (AMI) backups to rebuild databases from scratch
  - Amazon Elastic Block Store (Amazon EBS) volume snapshots to attach EBS volumes to Amazon EC2

Managing HA/DR resources in AWS Backup

AWS Backup is a fully managed service that offers the ability to create backup plans and schedules, and assign AWS resources that are involved in HA/DR configuration—such as Amazon EBS volumes to create snapshots and Amazon EC2 AMIs—to these backup plans. You can also use AWS Backup to schedule multi-Region copies of these EBS snapshots. For optimal usage, AWS Backup requires an efficient tagging mechanism for resources to be in place. AWS Backup also supports application-consistent backups through the Windows Volume Shadow Copy Service (VSS), which you can use for SQL Server. For storage-level protection, we recommend using EBS snapshots. Initial EBS snapshots are full, and subsequent snapshots are incremental. Although EBS snapshots offer storage-level protection, they do not replace SQL Server file-based native backups that offer point-in-time recovery.

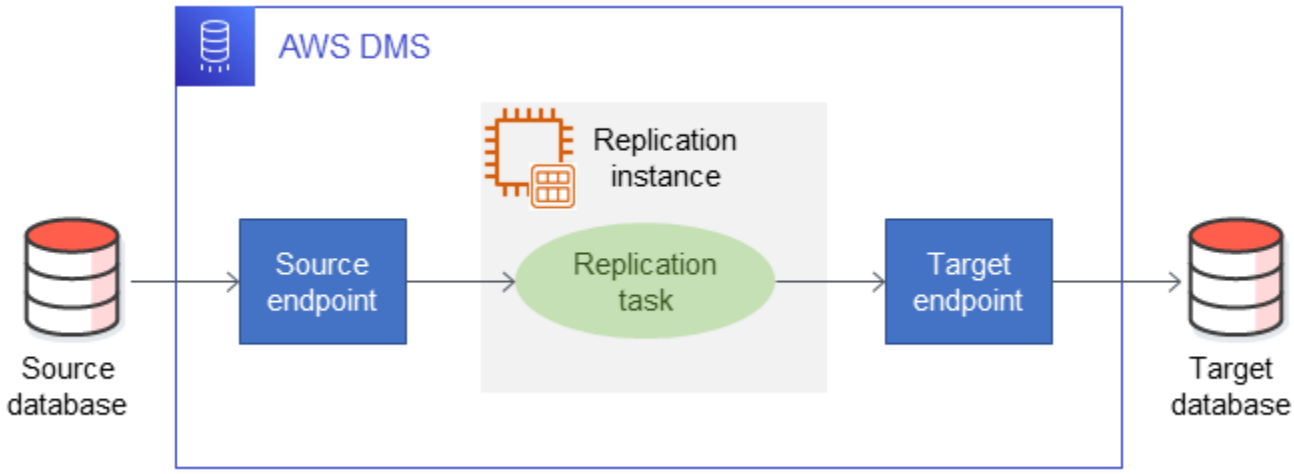
Using AWS DMS for HA/DR

If you're looking for an alternative to SQL Server Always On options for replication or if you have heterogeneous source and target databases, either in a hybrid setup or in AWS, you can use AWS Database Migration Service (AWS DMS) in the following ways.

If you use AWS DMS with SQL Server in a self-managed context (hosted on Amazon EC2 or on premises), it supports one-time and ongoing replication in two modes: by using MS-REPLICATION

(to capture changes to tables that have primary keys) and MS-CDC (to capture changes to tables that don't have primary keys). However, if you use Amazon Relational Database Service (Amazon RDS) as a source for AWS DMS, only MS-CDC is supported. AWS DMS offers a range of source and target endpoints, supports heterogeneous database engines, and offers fine-grained control over the replication process. You can also use the AWS Schema Conversion Tool (AWS SCT) with AWS DMS for heterogeneous database migrations. AWS SCT automates schema-level changes and also produces reports for migration readiness and planning.

You add source and target databases as end points in AWS DMS, as illustrated in the following diagram. This service implements a logical replication process by using either MS-REPLICATION or MS-CDC. If you have a hybrid setup, you can configure AWS DMS for ongoing replication between on premises and AWS. During the cutover, the AWS DMS migration task can be stopped and the application will be able to connect to the database that is already in sync with the on-premises database without further delay. Using AWS DMS for SQL Server as a source has a few limitations, which are outlined in the AWS DMS documentation.

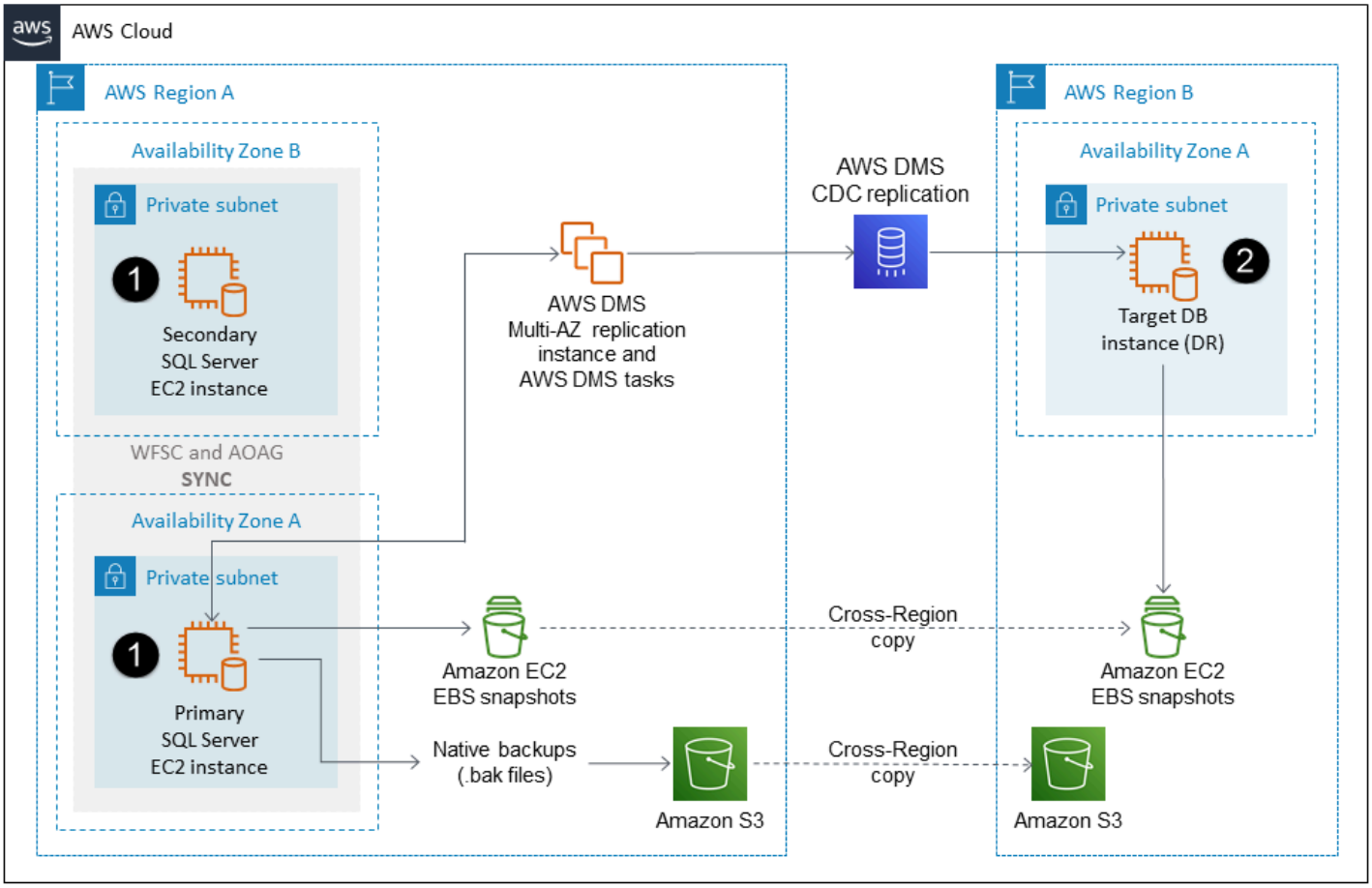


Consider using AWS DMS instead of native HA/DR methods in the following scenarios:

- When you want to save on licensing costs. For example, if you're using an advanced version such as SQL Server Enterprise edition only for its Always On options, you might consider setting up AWS DMS instead, because it can provide a logical replication option without the cost of an Enterprise edition license.
- When you have heterogeneous sources and targets. SQL Server versions on primary and disaster recovery nodes do not need to match (within AWS DMS limitations), which provides significant flexibility.

- To avoid the overhead of Windows, SQL Server clustering, and distributed availability group setup and management. AWS DMS offers a straightforward setup and easy management of replication tasks.
- For business use cases such as near real-time transfer (depending on replication instance, network configuration, and data volume), data masking, selective filtering, schema/table mapping (homogeneous and heterogeneous), pre-migration assessments, and JSON support.
- To easily duplicate, stop, and start tasks as needed based on log sequence numbers (LSNs), timestamps, and similar options.

The following diagram shows an alternative approach to how AWS DMS can provide replication support. In this configuration, the source is a SQL Server Always On availability group cluster, and AWS DMS uses the change data capture (CDC) option to continuously replicate data to a target in a different AWS Region. For the most optimal performance, is critical to ensure that the replication instance is right-sized and remains in the source Region.



The source and target engines do not have to match. In the diagram, the primary and secondary nodes marked as (1) can be a SQL Server cluster in a Single-AZ or Multi-AZ configuration. Or the source can be a single SQL Server node that supports MS-CDC or MS-REPLICATION.

The target DB instance, marked as (2) in the diagram, can be any version of SQL Server on Amazon RDS, Amazon EC2, or any other heterogeneous target. It doesn't have to match the primary and secondary instances or support Always On availability groups. For example, the source can be a SQL Server Always On availability group cluster, and the target can be Amazon Aurora PostgreSQL-Compatible Edition.

Using AWS Application Migration Service for DR

We recommend using the AWS Application Migration Service for lift-and-shift migrations to AWS. Application Migration Service continuously replicates your machines (including the operating system, system state configuration, databases, applications, and files) into a low-cost staging area in your target AWS account and preferred Region. In the case of a disaster, you can use Application Migration Service to automatically launch thousands of your machines in their fully provisioned state in minutes.

Additional considerations

The following list identifies the possible bottlenecks you should consider when you design an HA/DR strategy.

- Bandwidth, latency, network complexity, and connectivity in a multi-Region node setup.
- Size of the Amazon EBS or Amazon EC2 snapshots, and the time it takes to copy them over by using AWS Backup.
  - Amazon EBS and Amazon EC2 snapshots are stored in Amazon S3 by using AWS Backup.
  - An EBS snapshot does not replicate to the target Region in Amazon S3 until the current snapshot is completed. The duration of replication also depends on the size of the volume.
  - When the snapshot is complete, the duration of time to copy snapshots can be as little as 15 minutes for 99.99% of the objects. However, thorough testing is required for specific use cases and critical large volumes.
- Time required to restore EBS volumes in the target Availability Zone and Region.
- Time required to restore Amazon EC2 images in the target Availability Zone and Region.

- If building from scratch, time required to provision infrastructure for the Amazon EC2 image or restored EBS snapshots in the target Availability Zone and Region.
- If restoring from scratch, time required to restore SQL Server native full, differential, and log backups in the target Availability Zone and Region.
- Application and external dependencies that need to be available across Regions.
- Limitations on file sizes for volumes and for uploading to Amazon S3.

Disaster recovery scenarios

This section provides examples of a single Availability Zone or AWS Region failure, and discusses options for disaster recovery (DR). The examples assume a recovery point objective (RPO) of 15 minutes and a recovery time objective (RTO) of 4 hours.

Availability Zone failure

You can use one of the following options to recover from a single Availability Zone failure within the given parameters (RPO of 15 minutes, RTO of 4 hours).

- Provision the application recovery by using the most recent Amazon Elastic Compute Cloud (Amazon EC2) image backup, and connect to the existing warm standby database instance through an Always On availability group deployment or log shipping.
  - A SQL Server Always On availability group setup for DR with two or more nodes provides automatic failover to the secondary node through synchronous-commit or asynchronous-commit mode, so the database is available immediately. For an HA setup, both nodes are available for read operations. This option meets both RTO and RPO requirements comfortably. In SQL Server Standard edition, using basic availability groups is also an option, but it's limited to two nodes, because an availability group can include only one database. However, you can set up multiple availability groups within one Region or across Regions. This setup provides cost savings, because there is no additional cost for the secondary node, which isn't accessible for read operations. SQL Server Enterprise edition provides full functionality and failover for all databases within a single availability group. For examples of this option, see the following architecture diagrams:
    - Two-node HA/DR architecture with Always On availability group cluster (single-Region, Multi-AZ)
    - Three-node HA/DR architecture (single-Region, Multi-AZ)
    - Four-node HA/DR architecture with Always On distributed availability group cluster (multi-Region, Multi-AZ)
    - Three-node HA/DR architecture with single availability group (multi-Region)
- SQL Server log shipping as a DR solution requires a manual failover to a standby server and depends on the frequency of log backups. This is one of the least expensive DR options. SQL Server editions for the primary and log-shipped DR site do not need to match. This

option meets the RPO (using transaction log backups every 5 minutes and RTO, but requires maintenance through manual, custom scripts. For an example of this option, see the following architecture diagram:

  - Three-node HA/DR architecture with log shipping (multi-Region)

- If you have an application such as an SQL Server Reporting Services (SSRS) application that has a scaled-out deployment, the load balancer can redirect all traffic to the secondary node.
- You can use Amazon EC2 base AMIs for the application and database server to provision the infrastructure. Databases can be restored in a new Availability Zone, depending on their size and backup frequency, from the most recent native backups (full backup, differential backup, or transaction log backups every 5 minutes) or by using EBS snapshots. This option meets the RPO and RTO requirements but requires custom scripting. You must also consider the time required to provision the infrastructure, and meeting RPO and RTO requirements can be challenging.
- Amazon EC2 images (including EBS volumes) for both applications and the database server can be restored in a new Availability Zone. RPO can be challenging, depending on the most recent backup, but this option can be combined with the most recent transaction logs to meet requirements. This option supports Windows Volume Shadow Copy Service (VSS) snapshots.

Region failure

You can use one of the following options to recover from a single AWS Region failure within the given parameters (RPO of 15 minutes, RTO of 4 hours).

- You can use Amazon EC2 base Amazon Machine Images (AMIs) for the application and database server to provision the infrastructure. Databases can be restored in a new Region, depending on their size and backup frequency, from the most recent native backups (full backup, differential backup, or transaction log backups every 5 minutes). This option meets the RPO and RTO requirements but requires custom scripting.
  - SQL Server log shipping as a DR solution requires a manual failover to a standby server and depends on the frequency of log backups. This is one of the least expensive DR options. SQL Server editions for the primary and log-shipped DR site do not need to match. This option meets the RPO (by using transaction log backups every 5 minutes) and RTO, but requires maintenance through manual, custom scripts. Large databases require long restoration times.
- You can use an Amazon EC2 AMI for both the application and the database server and restore it to a target in a new Region. RPO depends on the size and frequency of the backups.

  - The most recent application images can be restored by using an AMI. You can use recent native differential or transaction log backups every 5 minutes to bring the database up to date to meet the RPO.
  - RTO depends on the size and time to transfer and restore the snapshots to the new Region, if the source isn't already in sync with the target.
- The solution with the least downtime is to restore the application backup image and have a warm standby SQL Server node in a remote Region by using a two-node, three-node, or four-node availability group setup (basic, classic, or distributed) and to connect to the standby database server after a failover. The synchronous-commit mode replica meets the RPO requirements, whereas asynchronous-commit mode replica might be delayed depending on the volume of transactions. You can use a distributed availability group configuration to scale out database nodes in a new Region, if needed. This configuration also reduces complexity because it uses two independently availability groups instead of a single availability group spread across Regions in either synchronous-commit or asynchronous-commit mode, and meets both RTO and RPO requirements comfortably. Alternatively, using SQL Server basic availability groups in the Standard edition is also an option. However, it has limitations because it supports only up to two nodes, and only one database can be in a single availability group although multiple availability groups are supported. You can set up SQL Server Standard edition within one Region or across Regions. This edition provides cost savings because it doesn't charge for the secondary node, which isn't accessible for read operations. SQL Server Enterprise edition provides full functionality, and supports the failover of all databases as a single availability group failover.

Common use cases

As a sizing exercise, 80% of SQL Server applications running on Amazon EC2 that have a normal online transaction processing (OLTP) workload can be grouped into one of three categories based on how critical they are:

- SQL Server HA/DR with SQL Server backups, using two synchronous-commit replicas and one asynchronous-commit mode replica
- AWS Backup HA/DR with SQL Server backups, using an Amazon EC2 AMI for both the application and the database, and Amazon EBS storage
- AWS Backup HA/DR with SQL Server backups, using an Amazon EC2 base AMI for the database server, an Amazon EC2 image for the application, and Amazon EBS snapshots

The following table provides details about each category.

	SQL Server HA/ DR with SQL Server backups	AWS Backup HA/ DR with AMIs, EBS storage, and SQL Server backups	AWS Backup HA/ DR with AMIs, EBS snapshots, and SQL Server backups
Restore process in case of a disaster	• Restore Amazon EC2 base AMI for the application from AWS Backup • Fail over to the standby instance in the Region (in the case of Availability Zone failure) or to the cross-Region instance (in case of Region failure) • Meets RPO and RTO requirements	• Restore Amazon EC2 images from backups for both the application and the database • Provides both in-Region and cross-Region support • Apply most recent SQL Server differential and transaction log backups (every 15 minutes) to meet RPO and RTO requirements for the database	• Restore Amazon EC2 image from backup for the application • Restore Amazon EC2 base AMI for the database server • Restore EBS snapshots (if any) • Cluster has to be rebuilt • Provides both in-Region and cross-Region support • Apply most recent differential and transaction log backups to the database to meet RPO requirements, but RTO might not be met
Primary resources	• Three SQL Server Enterprise edition licenses (passive HA and DR nodes license is free	• One SQL Server license (any edition).	• One SQL Server license (any edition).

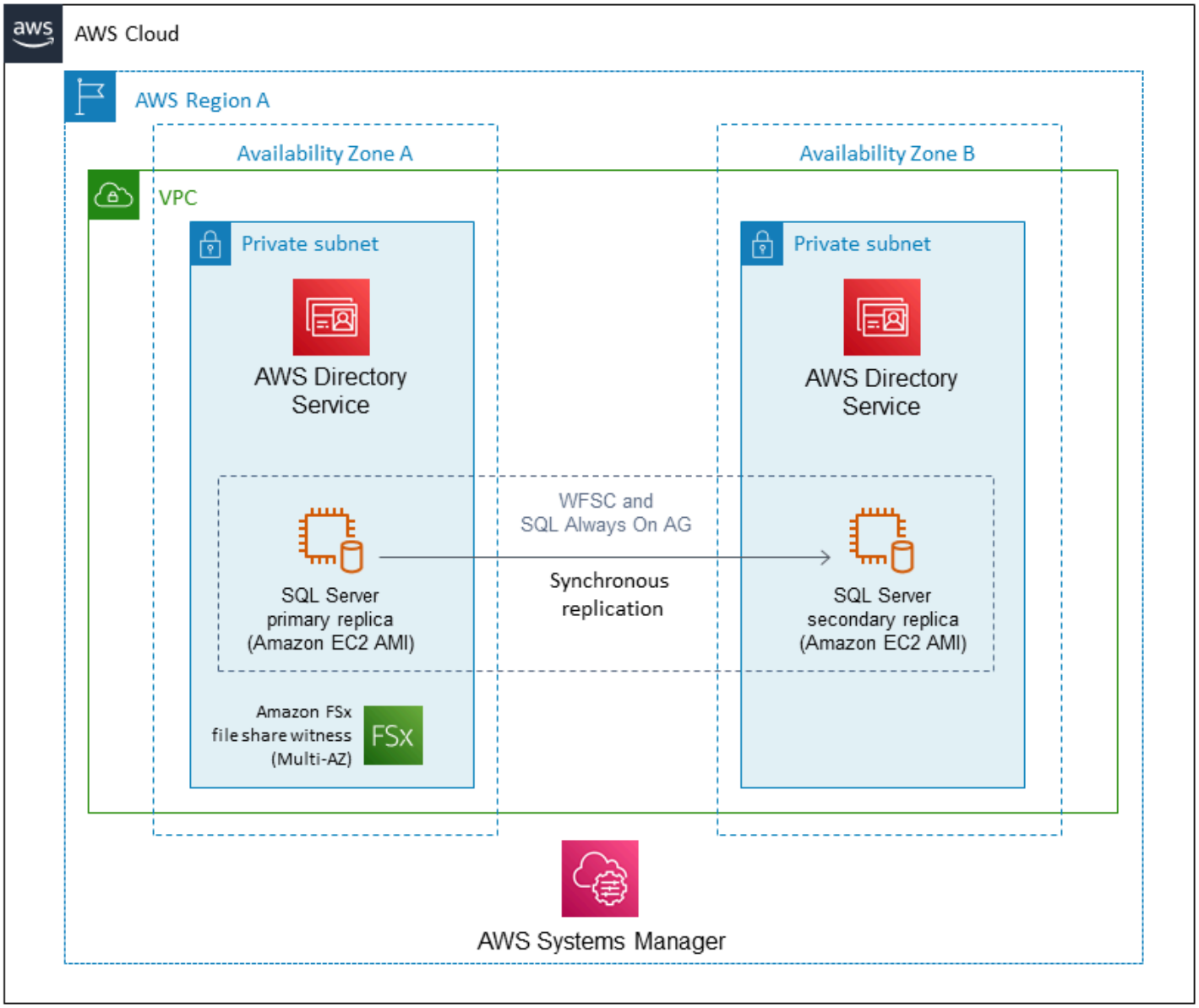
	SQL Server HA/ DR with SQL Server backups	AWS Backup HA/ DR with AMIs, EBS storage, and SQL Server backups	AWS Backup HA/ DR with AMIs, EBS snapshots, and SQL Server backups
	if you have an existing Software Assurance licensing agreement with Microsoft; see announcement) • Amazon EC2 backup space on Amazon Simple Storage Service (Amazon S3) • Cross-Region data transfer	• Amazon EC2 backup space on Amazon S3 • SQL Server backups (differential and log files) on Amazon S3 • Cross-Region data transfer	• Amazon EC2 backup space on Amazon S3 • SQL Server backups (differential and log files) on Amazon S3 • Cross-Region data transfer
HA/DR	Offers HA and DR	Offers DR only	Offers DR only
RPO	Failover is handled by SQL Server availability group (DR is manual)	Manual or custom scripted	Manual or custom scripted
RTO	Seconds to minutes	Minutes to hours	Multiple hours
Risk of missing SLAs	Low	Medium	High
Manageability	Simple	Medium	Medium
Scaling	Simple	Medium	Medium

	SQL Server HA/ DR with SQL Server backups	AWS Backup HA/ DR with AMIs, EBS storage, and SQL Server backups	AWS Backup HA/ DR with AMIs, EBS snapshots, and SQL Server backups
File size limitations for uploads to Amazon S3 or cross-Region transfer	N/A – Handled in synchronous-commit mode or asynchronous-commit mode to a warm standby	Yes	Yes
Data loss	Near zero (depends on the workload and infrastructure provisioned)	Depends on the frequency of Amazon EC2 backup images and SQL Server backups	Depends on the frequency of Amazon EC2 backup images or EBS snapshots and SQL Server backups
Cost	Medium	Low – medium	Low – medium

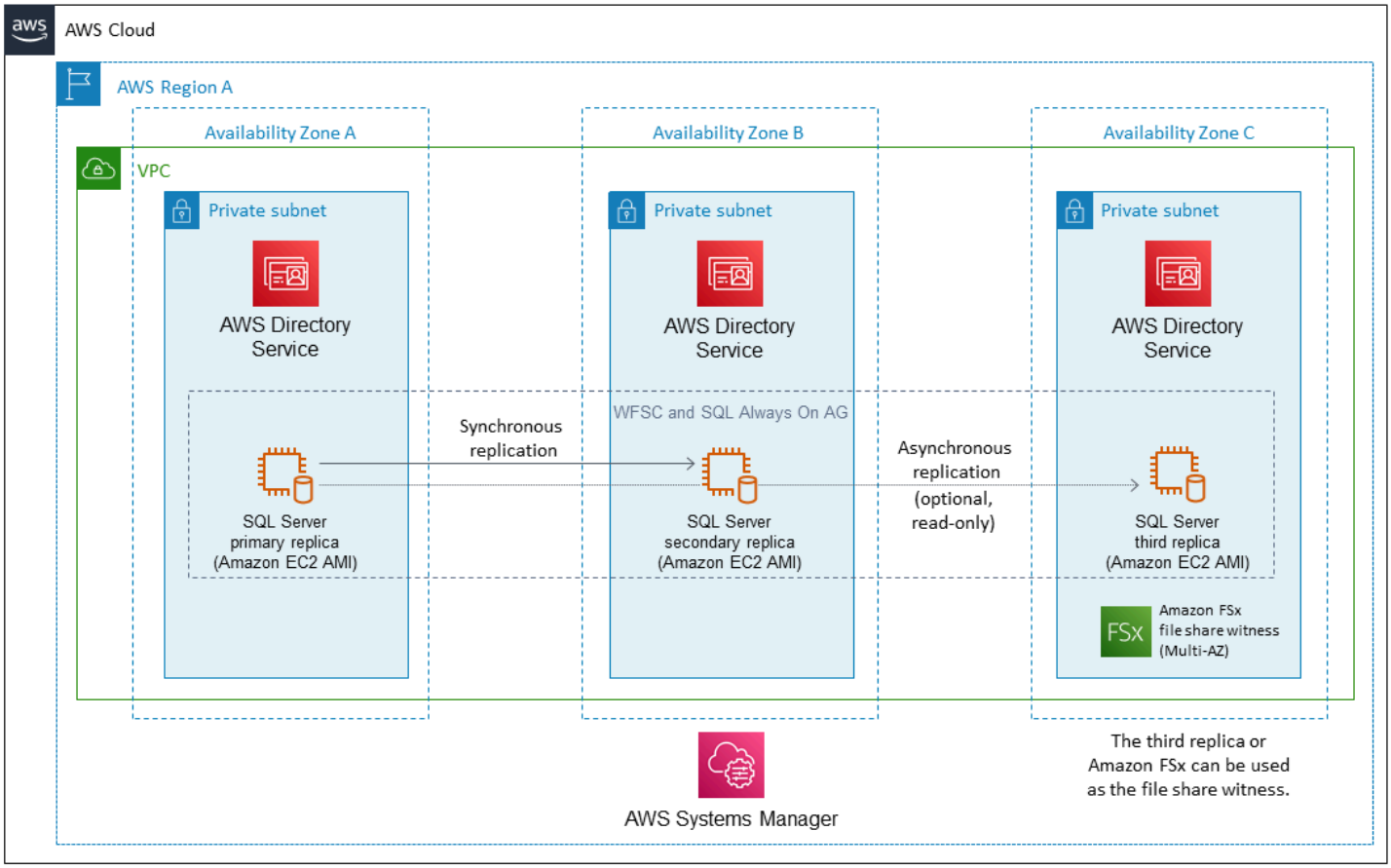
SQL Server on Amazon EC2 architecture diagrams

This section provides architecture diagrams that illustrate the HA/DR strategies described in previous sections.

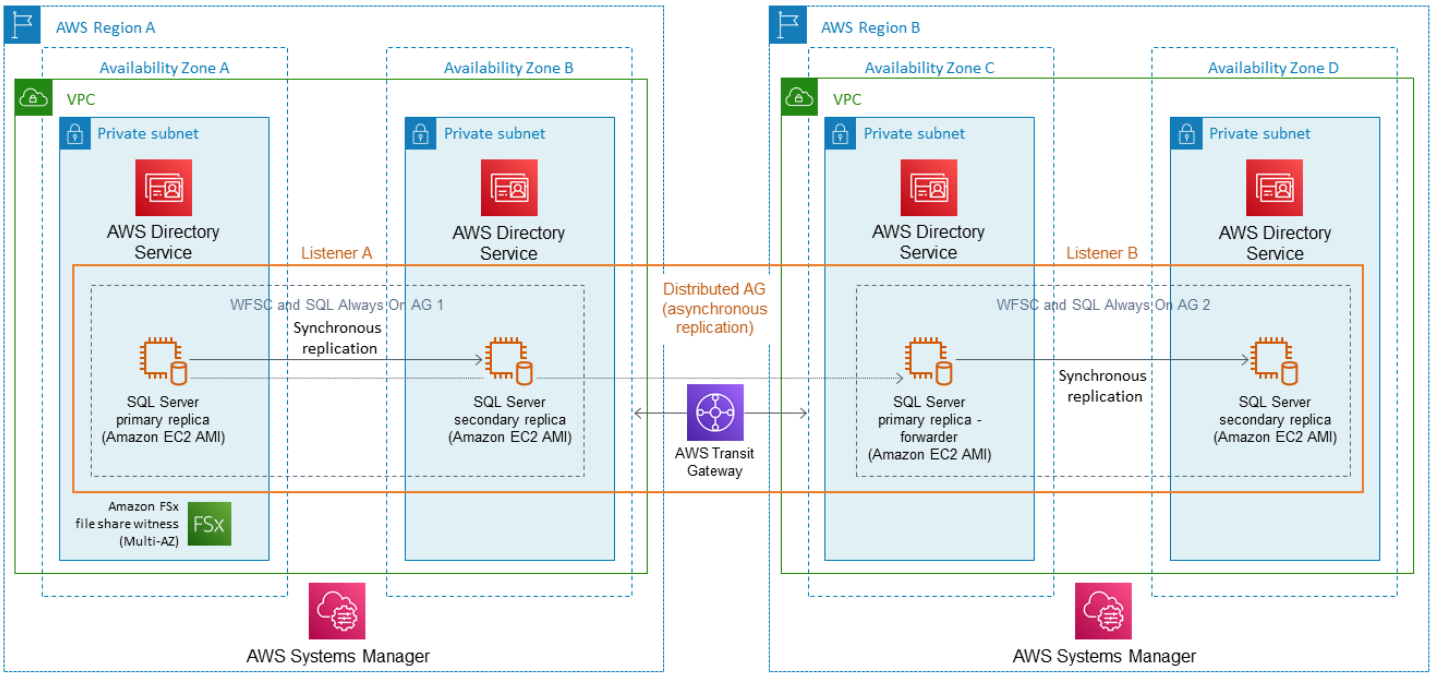
Two-node HA/DR architecture with Always On availability group cluster (single-Region, Multi-AZ)



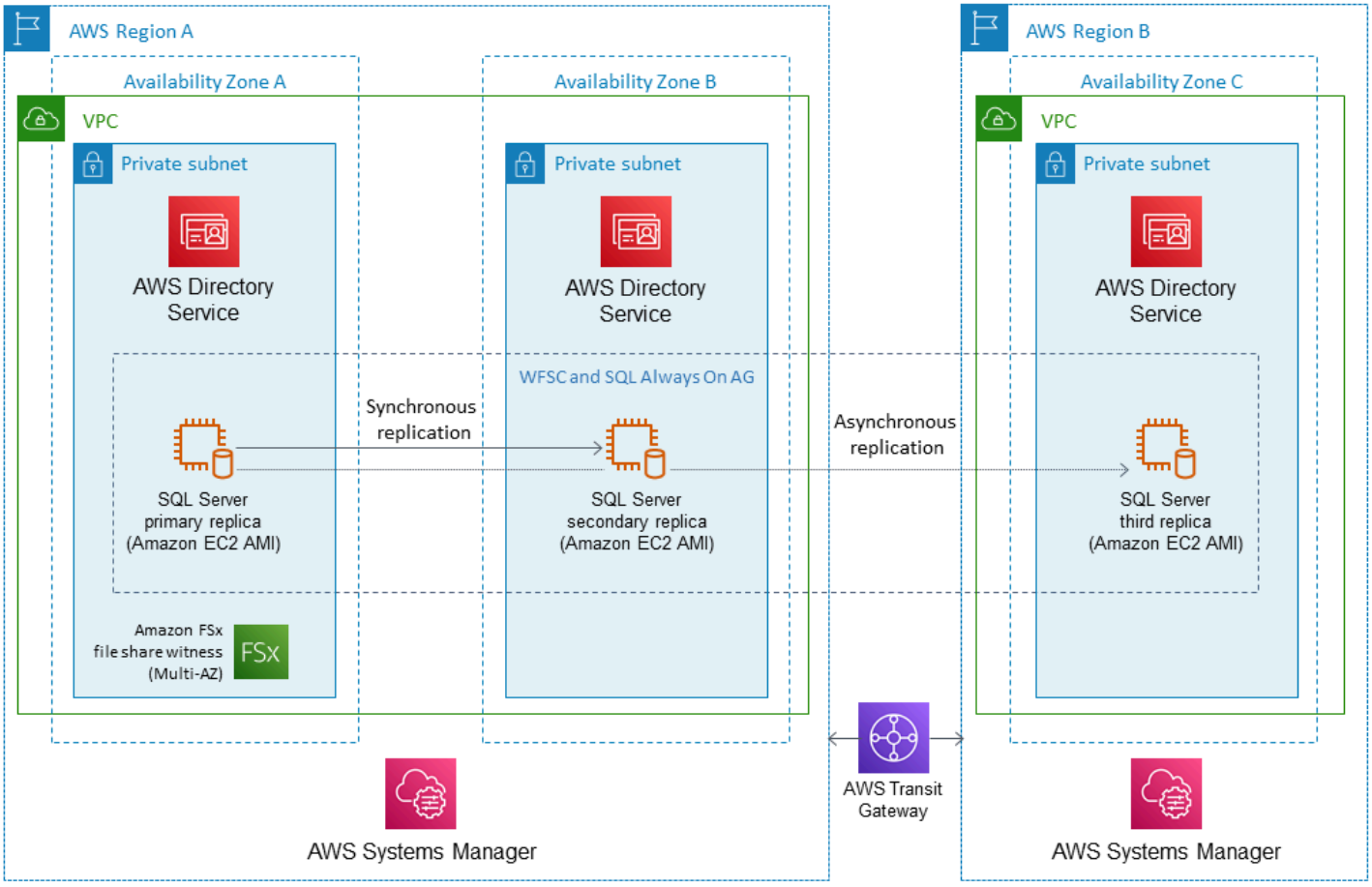
Three-node HA/DR architecture (single-Region, Multi-AZ)



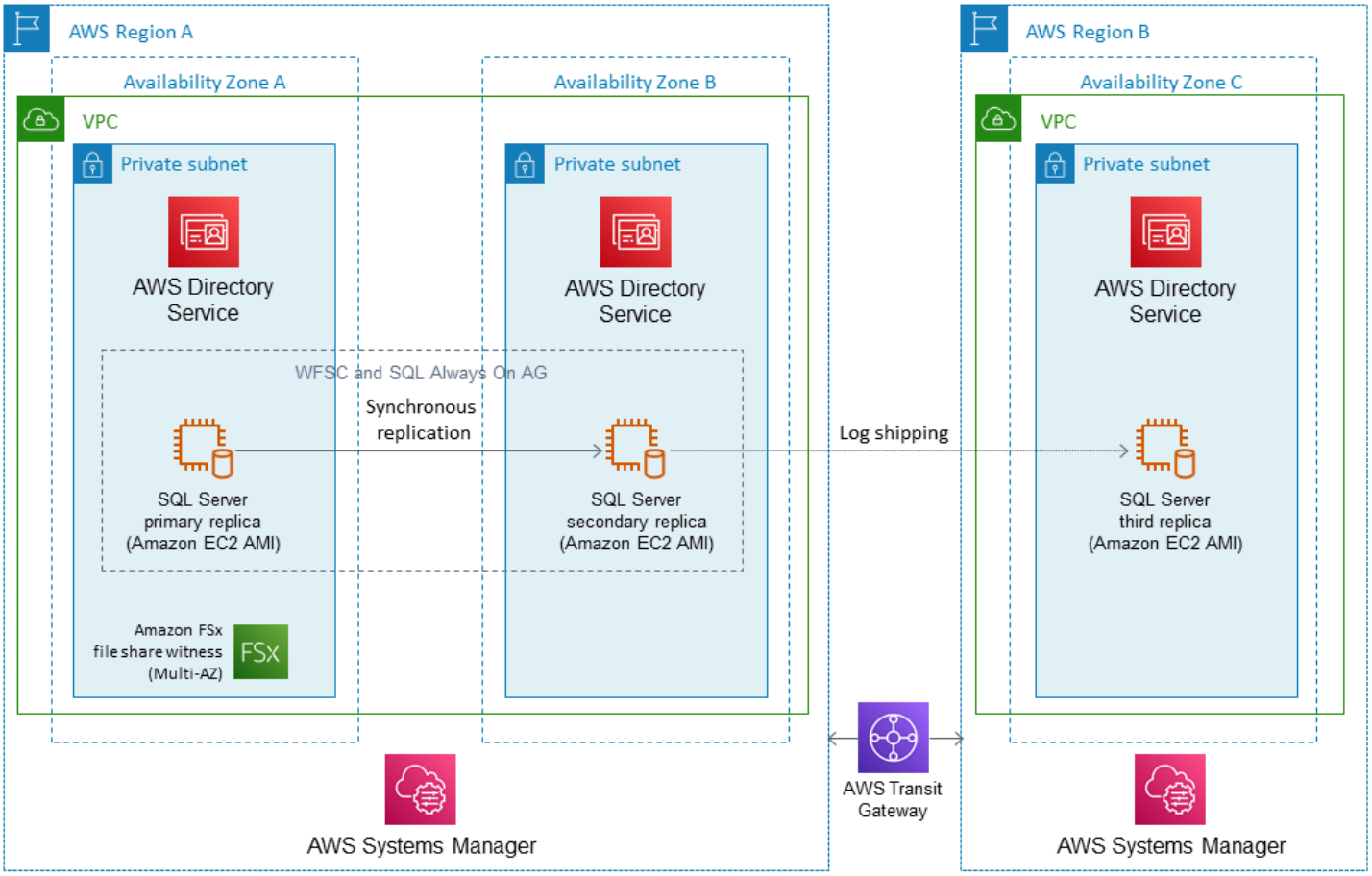
Four-node HA/DR architecture with Always On distributed availability group cluster (multi-Region, Multi-AZ)



Three-node HA/DR architecture with single availability group (multi-Region)



Three-node HA/DR architecture with log shipping (multi-Region)

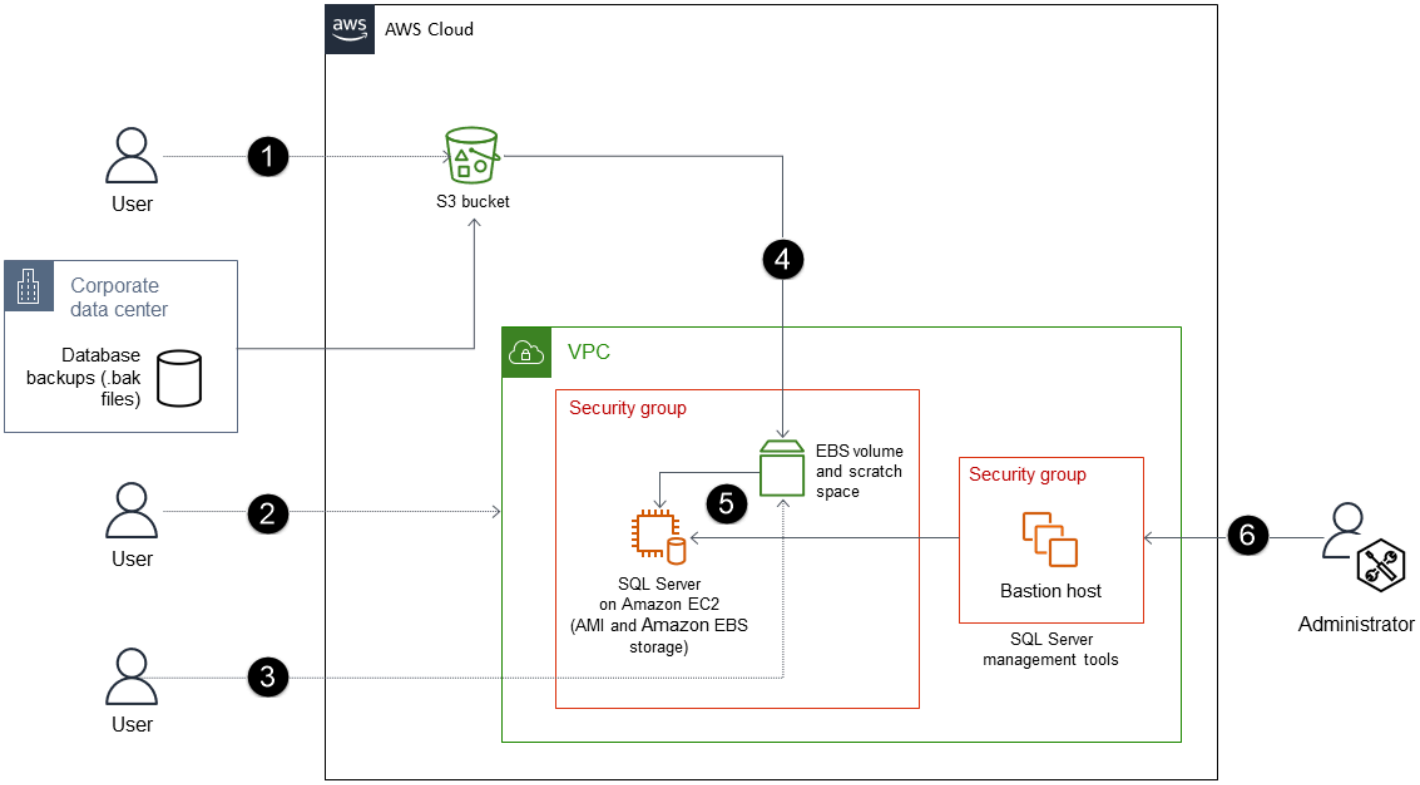


Restore options

The following sections provide two database restore options for SQL Server on Amazon Elastic Compute Cloud (Amazon EC2), when your backups are on premises.

Using Amazon S3

This SQL Server database restore approach uses Amazon Simple Storage Service (Amazon S3) commands for the AWS Command Line Interface (AWS CLI) or the Amazon S3 API to upload the backup files directly to an S3 bucket.



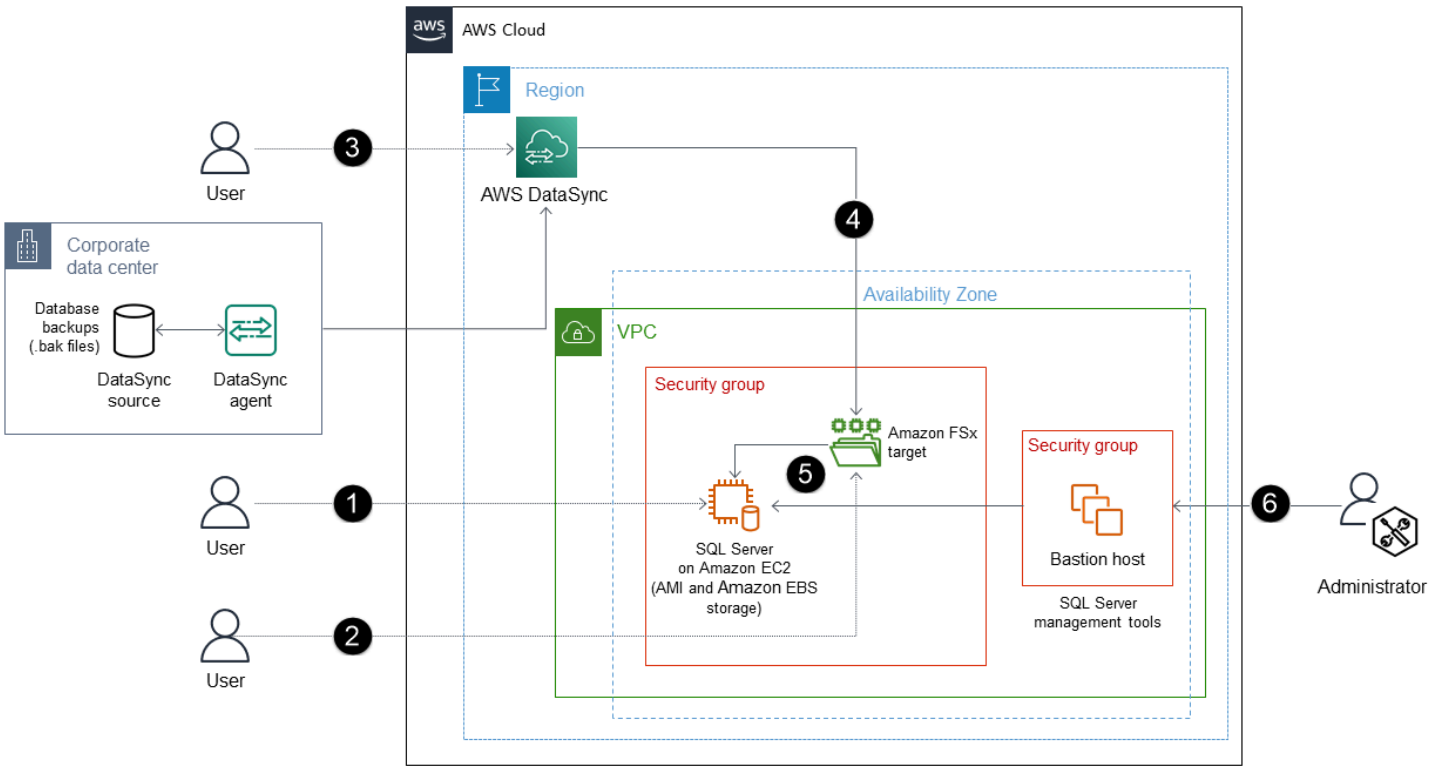
The process consists of these steps:

1. Create an S3 bucket (or use an existing bucket) to store the backup files, and transfer backup (.bak) files from your on-premises database to the S3 bucket by using the AWS CLI or Amazon S3 API.
2. Deploy SQL Server on an EBS-optimized EC2 instance, using a SQL Server Amazon Machine Image (AMI). This AMI must contain EBS volumes that are configured with an OS partition, a DATA partition, a LOG partition, tempdb (NVMe) storage, and scratch space.

3. (Optional) Attach a non-root EBS volume to the EC2 instance.
4. Copy the backup files to the non-root EBS volume.
5. Restore the backup files from the EBS volume to SQL Server on the EC2 instance.
6. Use SQL Server management tools to manage your database.

Using AWS DataSync and Amazon FSx

This SQL Server database restore approach uses AWS DataSync to transfer the backup files to Amazon FSx for Windows File Server.



The process consists of these steps:

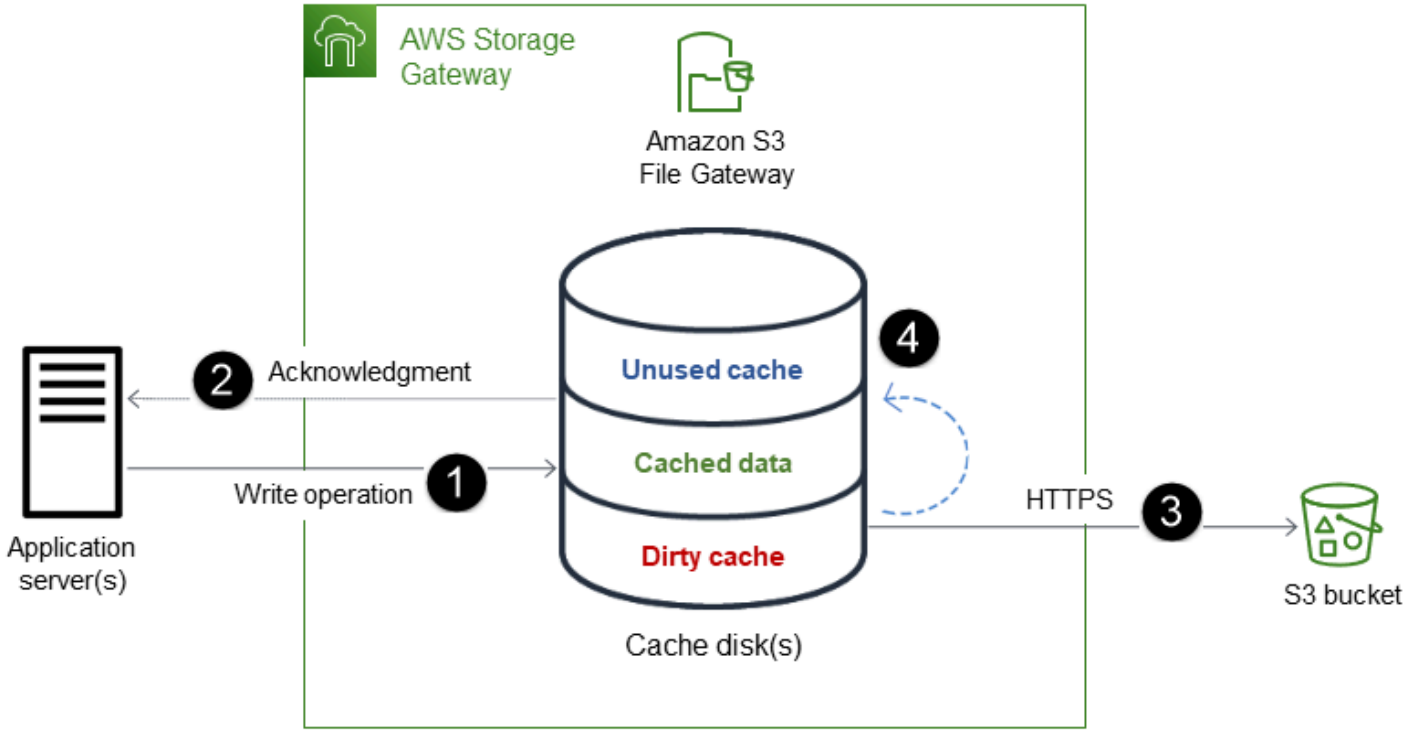
1. Deploy SQL Server on an EBS-optimized EC2 instance with attached NVMe, using an AMI that contains EBS volumes configured with OS, DATA, LOG, and tempdb. (For example, you can use the memory optimized `r5d.large` instance class.)
2. Use FSx for Windows File Server to create a file server. This can be used as a temporary storage location to download SQL Server backup (.bak) files from your on-premises environment.
3. Create an DataSync endpoint and agent for the Amazon FSx file server.

4. DataSync automates data synchronization between your on-premises storage and the Amazon FSx file server without requiring Amazon S3.
5. Restore the backup files from the Amazon FSx file server to SQL Server on the EC2 instance.
6. Use SQL Server management tools to manage your database.

> **Note**
>
> Amazon EC2 offers Microsoft SQL Server on Microsoft Windows Server AMIs for multiple SQL Server editions.

Using Amazon S3 File Gateway

You can use Amazon S3 File Gateway to store native SQL Server backups to Amazon S3, as illustrated in the following diagram. Alternatively, there are tools such as Commvault and LiteSpeed that help you manage file-level backups at scale and store them directly in Amazon S3. You can also use a tool such as SIOS DataKeeper for backup/recovery and DR configuration.



The process consists of these steps:

1. Data is written on the file gateway's local cache disk.
2. After the data is safely persisted to the local cache, the file gateway acknowledges the completion of the write operation to the client application.
3. The file gateway transfers data to the S3 bucket asynchronously. It optimizes data transfer and uses HTTPS to encrypt data in transit.
4. After data is uploaded to the S3 bucket, it stays in the file gateway's local cache until it is evicted.

Next steps and resources

This guide covered best practices for quick disaster recovery of SQL Server databases. The recommendations include using images to restore the application instance and to use native SQL methods to restore the database, or, preferably, to fail over the database. In contrast to large database restores that can take hours, using Amazon Elastic Compute Cloud (Amazon EC2) Amazon Machine Image (AMI) backups in combination with the most recent transaction logs helps you meet your recovery point objective (RPO) and recovery time objective (RTO) requirements while keeping your overall costs low. The optimal approach depends on the size of your database, the number and nature of backups, and the frequency of transaction log backups for which a disaster recovery strategy needs to be designed. See the following links for more information, best practices, Quick Start guides, and prescriptive guidance on migrating and hosting SQL Server on Amazon EC2.

Documentation

- Best practices and recommendations for SQL Server clustering on Amazon EC2 (Amazon EC2 documentation)
- Amazon EC2 instance store (Amazon EC2 documentation)
- Replicating objects (Amazon S3 documentation)
- Amazon EBS fast snapshot restore (Amazon EC2 documentation)
- SQL Server with Always On replication on the AWS Cloud (Quick Start reference deployment)
- Amazon EBS volume types (Amazon EC2 documentation)
- Using FSx for Windows File Server with Microsoft SQL Server (Amazon FSx documentation)
- What is AWS Backup? (AWS Backup documentation)
- AWS Windows AMIs (Amazon EC2 documentation)

AWS Prescriptive Guidance

- Amazon EC2 backup and recovery with snapshots and AMIs

Blog posts and news

- Easily store your SQL Server backups in Amazon S3 using File Gateway
- Monitor data transfer costs related to Amazon S3 Replication

- Multi-region SQL Server deployment using distributed availability groups
- Field Notes: Building a Multi-Region Architecture for SQL Server using FCI and Distributed Availability Groups
- Amazon EC2 now offers Microsoft SQL Server on Microsoft Windows Server 2022 AMIs

SQL Server documentation

- Editions and supported features of SQL Server

Appendix: Amazon EBS SSD storage types

Amazon Elastic Block Store (Amazon EBS) provides the following solid state drive (SSD) backed volumes. For the most recent information, see Amazon EBS volume types in the Amazon EC2 documentation.

	General Purpose SSD		Provisioned IOPS SSD		
Volume type	`gp3`	`gp2`	`io2 Block Express`[1]	`io2`	`io1`
Durability	99.8% – 99.9% durability (0.1% – 0.2% annual failure rate)	99.8% – 99.9% durability (0.1% – 0.2% annual failure rate)	99.999% durability (0.001% annual failure rate)	99.999% durability (0.001% annual failure rate)	99.8% – 99.9% durability (0.1% – 0.2% annual failure rate)
Use cases	• Low-latency interactive apps • Development and test environments	• Low-latency interactive apps • Development and test environments	Workloads that require: • Sub-millisecond latency • Sustained IOPS performance • More than 64,000 IOPS or 1,000 MiB/s of throughput	• Workloads that require sustained IOPS performance or more than 16,000 IOPs • I/O-intensive database workloads	• Workloads that require sustained IOPS performance or more than 16,000 IOPs • I/O-intensive database workloads

	General Purpose SSD		Provisioned IOPS SSD		
Volume size	1 GiB – 16 TiB	1 GiB – 16 TiB	4 GiB – 64 TiB	4 GiB – 16 TiB	4 GiB – 16 TiB
Maximum IOPS per volume (16 KiB I/O)	16,000	16,000	256,000	64,000[2]	64,000[2]
Maximum throughput per volume	1,000 MiB/s	250 MiB/s[3]	4,000 MiB/s	1,000 MiB/s[2]	1,000 MiB/s[2]
Amazon EBS Multi-attach	Not supported	Not supported	Supported	Supported	Supported
Boot volume	Supported	Supported	Supported	Supported	Supported

[1] `io2 Block Express` volumes are supported with R5b instances only. `io2` volumes attached to an R5b instance during or after launch automatically run on Block Express. For more information, see io2 Block Express volumes in the Amazon EC2 documentation.

[2] Maximum IOPS and throughput are guaranteed only on instances built on the Nitro System provisioned with more than 32,000 IOPS. Other instances guarantee up to 32,000 IOPS and 500 MiB/s. `io1` volumes that were created before December 6, 2017 and that have not been modified since creation might not reach full performance unless you modify the volume.

[3] The throughput limit is between 128 MiB/s and 250 MiB/s, depending on the volume size. Volumes smaller than or equal to 170 GiB deliver a maximum throughput of 128 MiB/s. Volumes larger than 170 GiB but smaller than 334 GiB deliver a maximum throughput of 250 MiB/s if burst credits are available. Volumes larger than or equal to 334 GiB deliver 250 MiB/s regardless of burst credits. gp2 volumes that were created before December 3, 2018 and that have not been modified since creation might not reach full performance unless you modify the volume.

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	—	February 28, 2022

AWS Prescriptive Guidance glossary

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

Numbers

7 Rs

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

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

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

A

ABAC

See attribute-based access control.

abstracted services

See managed services.

ACID

See atomicity, consistency, isolation, durability.

active-active migration

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

active-passive migration

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

aggregate function

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

AI

See artificial intelligence.

AIOps

See artificial intelligence operations.

anonymization

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

anti-pattern

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

application control

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

application portfolio

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

artificial intelligence (AI)

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

artificial intelligence operations (AIOps)

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

asymmetric encryption

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

atomicity, consistency, isolation, durability (ACID)

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

attribute-based access control (ABAC)

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

authoritative data source

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

Availability Zone

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

AWS Cloud Adoption Framework (AWS CAF)

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

AWS Workload Qualification Framework (AWS WQF)

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

B

bad bot

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

BCP

See business continuity planning.

behavior graph

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

big-endian system

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

binary classification

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

bloom filter

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

blue/green deployment

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

bot

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

botnet

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

branch

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

break-glass access

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

brownfield strategy

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

buffer cache

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

business capability

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

business continuity planning (BCP)

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

C

CAF

See AWS Cloud Adoption Framework.

canary deployment

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

CCoE

See Cloud Center of Excellence.

CDC

See change data capture.

change data capture (CDC)

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

chaos engineering

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

CI/CD

See continuous integration and continuous delivery.

classification

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

client-side encryption

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

Cloud Center of Excellence (CCoE)

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

cloud computing

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

cloud operating model

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

cloud stages of adoption

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

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

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

CMDB

See configuration management database.

code repository

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

cold cache

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

cold data

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

computer vision (CV)

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

configuration drift

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

configuration management database (CMDB)

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

conformance pack

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

continuous integration and continuous delivery (CI/CD)

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

CV

See computer vision.

D

data at rest

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

data classification

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

data drift

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

data in transit

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

data mesh

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

data minimization

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

data perimeter

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

data preprocessing

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

data provenance

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

data subject

An individual whose data is being collected and processed.

data warehouse

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

database definition language (DDL)

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

database manipulation language (DML)

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

DDL

See database definition language.

deep ensemble

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

deep learning

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

defense-in-depth

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

delegated administrator

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

deployment

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

development environment

See environment.

detective control

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

development value stream mapping (DVSM)

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

digital twin

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

dimension table

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

disaster

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

disaster recovery (DR)

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

DML

See database manipulation language.

domain-driven design

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

DR

See disaster recovery.

drift detection

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

DVSM

See development value stream mapping.

E

EDA

See exploratory data analysis.

EDI

See electronic data interchange.

edge computing

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

electronic data interchange (EDI)

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

encryption

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

encryption key

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

endianness

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

endpoint

See service endpoint.

endpoint service

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

enterprise resource planning (ERP)

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

envelope encryption

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

environment

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

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

epic

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

ERP

See enterprise resource planning.

exploratory data analysis (EDA)

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

F

fact table

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

fail fast

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

fault isolation boundary

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

feature branch

See branch.

features

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

feature importance

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

feature transformation

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

few-shot prompting

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

FGAC

See fine-grained access control.

fine-grained access control (FGAC)

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

flash-cut migration

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

FM

See foundation model.

foundation model (FM)

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

G

generative AI

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

geo blocking

See geographic restrictions.

geographic restrictions (geo blocking)

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

Gitflow workflow

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

golden image

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

greenfield strategy

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

guardrail

A high-level rule that helps govern resources, policies, and compliance across organizational units (OUs). *Preventive guardrails* enforce policies to ensure alignment to compliance standards. They are implemented by using service control policies and IAM permissions boundaries. *Detective guardrails* detect policy violations and compliance issues, and generate alerts for remediation. They are implemented by using AWS Config, AWS Security Hub CSPM, Amazon GuardDuty, AWS Trusted Advisor, Amazon Inspector, and custom AWS Lambda checks.

H

HA

See high availability.

heterogeneous database migration

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

high availability (HA)

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

historian modernization

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

holdout data

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

homogeneous database migration

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

hot data

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

hotfix

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

hypercare period

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

I

IaC

See infrastructure as code.

identity-based policy

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

idle application

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

IIoT

See industrial Internet of Things.

immutable infrastructure

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

inbound (ingress) VPC

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

incremental migration

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

Industry 4.0

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

infrastructure

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

infrastructure as code (IaC)

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

industrial Internet of Things (IIoT)

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

inspection VPC

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

Internet of Things (IoT)

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

interpretability

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

IoT

See Internet of Things.

IT information library (ITIL)

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

IT service management (ITSM)

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

ITIL

See IT information library.

ITSM

See IT service management.

L

label-based access control (LBAC)

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

landing zone

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

large language model (LLM)

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

large migration

A migration of 300 or more servers.

LBAC

See label-based access control.

least privilege

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

lift and shift

See 7 Rs.

little-endian system

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

LLM

See large language model.

lower environments

See environment.

M

machine learning (ML)

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

main branch

See branch.

malware

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

managed services

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

manufacturing execution system (MES)

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

MAP

See Migration Acceleration Program.

mechanism

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

member account

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

MES

See manufacturing execution system.

Message Queuing Telemetry Transport (MQTT)

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

microservice

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

microservices architecture

An approach to building an application with independent components that run each application process as a microservice. These microservices communicate through a well-defined interface by using lightweight APIs. Each microservice in this architecture can be updated, deployed,

and scaled to meet demand for specific functions of an application. For more information, see Implementing microservices on AWS.

Migration Acceleration Program (MAP)

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

migration at scale

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

migration factory

Cross-functional teams that streamline the migration of workloads through automated, agile approaches. Migration factory teams typically include operations, business analysts and owners, migration engineers, developers, and DevOps professionals working in sprints. Between 20 and 50 percent of an enterprise application portfolio consists of repeated patterns that can be optimized by a factory approach. For more information, see the discussion of migration factories and the Cloud Migration Factory guide in this content set.

migration metadata

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

migration pattern

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

Migration Portfolio Assessment (MPA)

An online tool that provides information for validating the business case for migrating to the AWS Cloud. MPA provides detailed portfolio assessment (server right-sizing, pricing, TCO

comparisons, migration cost analysis) as well as migration planning (application data analysis and data collection, application grouping, migration prioritization, and wave planning). The MPA tool (requires login) is available free of charge to all AWS consultants and APN Partner consultants.

Migration Readiness Assessment (MRA)

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

migration strategy

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

ML

See machine learning.

modernization

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

modernization readiness assessment

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

monolithic applications (monoliths)

Applications that run as a single service with tightly coupled processes. Monolithic applications have several drawbacks. If one application feature experiences a spike in demand, the entire architecture must be scaled. Adding or improving a monolithic application's features

also becomes more complex when the code base grows. To address these issues, you can use a microservices architecture. For more information, see Decomposing monoliths into microservices.

MPA

See Migration Portfolio Assessment.

MQTT

See Message Queuing Telemetry Transport.

multiclass classification

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

mutable infrastructure

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

O

OAC

See origin access control.

OAI

See origin access identity.

OCM

See organizational change management.

offline migration

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

OI

See operations integration.

OLA

See operational-level agreement.

online migration

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

OPC-UA

See Open Process Communications - Unified Architecture.

Open Process Communications - Unified Architecture (OPC-UA)

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

operational-level agreement (OLA)

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

operational readiness review (ORR)

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

operational technology (OT)

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

operations integration (OI)

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

organization trail

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

organizational change management (OCM)

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

origin access control (OAC)

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

origin access identity (OAI)

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

ORR

See operational readiness review.

OT

See operational technology.

outbound (egress) VPC

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

P

permissions boundary

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

personally identifiable information (PII)

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

PII

See personally identifiable information.

playbook

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

PLC

See programmable logic controller.

PLM

See product lifecycle management.

policy

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

polyglot persistence

Independently choosing a microservice's data storage technology based on data access patterns and other requirements. If your microservices have the same data storage technology, they can encounter implementation challenges or experience poor performance. Microservices are more

easily implemented and achieve better performance and scalability if they use the data store best adapted to their requirements.

portfolio assessment

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

predicate

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

predicate pushdown

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

preventative control

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

principal

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

privacy by design

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

private hosted zones

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

proactive control

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

AWS Control Tower documentation and see Proactive controls in *Implementing security controls on AWS*.

product lifecycle management (PLM)

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

production environment

See environment.

programmable logic controller (PLC)

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

prompt chaining

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

pseudonymization

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

publish/subscribe (pub/sub)

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

Q

query plan

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

query plan regression

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

R

RACI matrix

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

RAG

See Retrieval Augmented Generation.

ransomware

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

RASCI matrix

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

RCAC

See row and column access control.

read replica

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

re-architect

See 7 Rs.

recovery point objective (RPO)

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

recovery time objective (RTO)

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

refactor

See 7 Rs.

Region

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

regression

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

rehost

See 7 Rs.

release

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

relocate

See 7 Rs.

replatform

See 7 Rs.

repurchase

See 7 Rs.

resiliency

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

resource-based policy

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

responsible, accountable, consulted, informed (RACI) matrix

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

responsive control

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

retain

See 7 Rs.

retire

See 7 Rs.

Retrieval Augmented Generation (RAG)

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

rotation

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

row and column access control (RCAC)

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

RPO

See recovery point objective.

RTO

See recovery time objective.

runbook

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

S

SAML 2.0

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

SCADA

See supervisory control and data acquisition.

SCP

See service control policy.

secret

In AWS Secrets Manager, confidential or restricted information, such as a password or user credentials, that you store in encrypted form. It consists of the secret value and its metadata. The secret value can be binary, a single string, or multiple strings. For more information, see What's in a Secrets Manager secret? in the Secrets Manager documentation.

security by design

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

security control

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

security hardening

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

security information and event management (SIEM) system

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

security response automation

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

server-side encryption

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

service control policy (SCP)

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

service endpoint

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

service-level agreement (SLA)

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

service-level indicator (SLI)

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

service-level objective (SLO)

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

shared responsibility model

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

SIEM

See security information and event management system.

single point of failure (SPOF)

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

SLA

See service-level agreement.

SLI

See service-level indicator.

SLO

See service-level objective.

split-and-seed model

A pattern for scaling and accelerating modernization projects. As new features and product releases are defined, the core team splits up to create new product teams. This helps scale your organization's capabilities and services, improves developer productivity, and supports rapid

innovation. For more information, see Phased approach to modernizing applications in the AWS Cloud.

SPOF

See single point of failure.

star schema

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

strangler fig pattern

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

subnet

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

supervisory control and data acquisition (SCADA)

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

symmetric encryption

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

synthetic testing

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

system prompt

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

T

tags

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

target variable

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

task list

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

test environment

See environment.

training

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

transit gateway

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

trunk-based workflow

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

trusted access

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

tuning

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

two-pizza team

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

U

uncertainty

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

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.