LEVVEL

AWS Disaster Recovery Strategies

Modern Strategies to Protect Your Business

By Chris Madison

LEVVEL

AWS Disaster Recovery Strategies

Modern Strategies to Protect Your Business

Amazon Web Services (AWS) is a superior choice for organizations seeking to reduce technology costs by moving on-premise and collocated compute resources to the cloud. The primary benefit is that AWS entails a much lower total cost of ownership (TCO) than self-hosted and collocated data centers, especially when considering servers, network infrastructure, hardware, software, operating systems, power costs, cooling costs, and a variety of other components. The low TCO of AWS not only makes it an excellent choice for any organization seeking to reduce their capital and operational expenses, it also serves as a valuable disaster recovery (DR) platform.

Disaster recovery is the essential business continuity and operational component that addresses an organization's technical infrastructure and ability to survive an outage. Regardless of the cause, a disaster, to a business, is any event that results in the loss of data, limits the ability to satisfy customer requests, or that disrupts income. The goal of disaster recovery is to salvage the technical and operational assets needed to run a business after natural or man-made disasters—with minimal manual input or nonstandard processes.

The primary measures of disaster recovery mechanisms are:

- **Recovery Time Objective (RTO)**: This metric is used to define the service level of time to recovery or when a system must be available after a failure. For example, an eight-hour RTO indicates that the application must be available eight hours after failure occurs.

- **Recovery Point Objective (RPO)**: RPO defines the amount of time that data is lost prior to the event. A one-hour RPO indicates that application operational data must be available one hour or older from the point of failure.

From a business goals perspective, typical trade-offs are analyzed in terms of down time, the longevity of data loss, and the cost of the disaster recovery solution. That is, as RTO or RPO decrease, the disaster recovery solution typically becomes costlier.

Geographic Isolation in AWS

Disasters are typically categorized as induced by nature or man. Man-made disasters are generally confined to a smaller geographical area than natural disasters (the Chernobyl disaster of 1986 is a notable exception). Natural disasters include floods, hurricanes, and earthquakes, and impact large areas.

AWS is designed around the idea of regions and Availability Zones. Regions are separate and completely independent geographic areas where AWS provides hosting facilities. Each region contains two or more Availability Zones for fault tolerance. Availability Zones are located within the same region, but in different locations (e.g., floodplains, fault zones) to survive local natural and man-made disasters. Availability Zones are connected to each other within a region and provide inter-regional communication capabilities.

The AWS region and availability zone concepts lay the foundation for an effective, fault-tolerant disaster recovery platform.

Disaster Recovery Strategies

Due to the variability of RTO and RPO requirements, there are a variety of disaster recovery strategies employed at the enterprise level. Referencing the diagram below, on the left are solutions (Backup & Restore, Pilot Light) that tend to have high RTO and RPO and thereby lower cost. As one progresses to the right of the diagram (Warm Standby, Multi-Site), RTO and RPO are minimal, but have a greater solution cost.

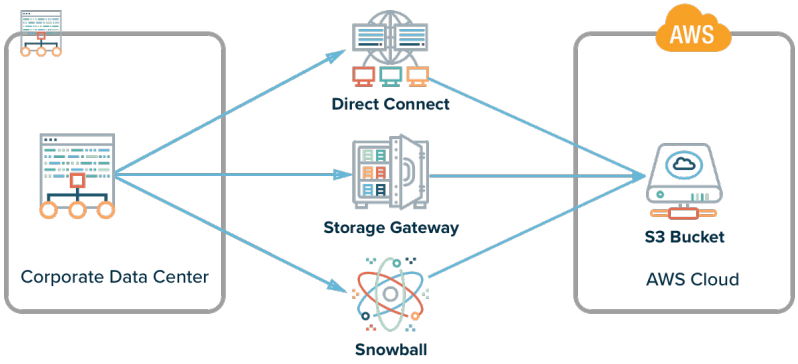
The diagram above identifies four general strategies for disaster recovery solutions and illustrates the relative cost size of each extreme.

Backup & Restore

The Backup & Restore strategy is a common DR pattern that is typically cheap but accompanied by high RPO and RTO. The canonical example is data being backed up onto tape and that tape archive being stored for some amount of time; these tapes are stored off-site in a facility specifically designed to maintain digital archives.

AWS provides a variety of low-cost options to support this DR strategy. Generally, the AWS Simple Storage Solution (S3) is used to store objects in the cloud. S3 is a powerful and low-cost storage solution with 11 9s durability (99.999999999% durable storage solution). In other terms, one may lose one object every 10 million years on average. As a companion service, AWS Glacier offers low-cost archiving of infrequently accessed data. S3 lifecycle operations may automatically archive data from an S3 bucket into AWS Glacier and enforce corporate disposition policies for compliance purposes.

Besides direct Internet and VPN connectivity to AWS to access S3 storage buckets, corporate data centers have several other options. These include Direct Connect, the AWS Storage Gateway, and, for extremely large data sets, Snowball.



- **Direct Connect**: Provides a direct, dedicated connection from a data center to AWS via a controllable telco infrastructure. In many cases, Direct Connect provides a more reliable and consistent network experience over standard Internet connectivity. Direct Connect provides improved connectivity, reliability, and capability for storage solutions over internet-based solutions alone.

- **Storage Gateway**: The storage gateway connects the corporate data center to cloud-based storage solutions—namely, S3 and Glacier. There are three Storage Gateway solutions:
    - **Gateway-cached volumes**: This solution mounts to on-premise compute resources as an ISCSI disk. The Storage Gateway caches frequently accessed data on-premises. However, all data is stored in an S3 bucket.
    - **Gateway-stored volumes**: The stored volumes solution stores data on-premises for low latency access and replicates all data to S3. Replication may be synchronous or asynchronous.
    - **Gateway-virtual tape library**: The virtual tape library (VTL) stores tape archives in S3 and Glacier, depending on lifecycle configuration.
- **Snowball**: Snowball is a batch cloud transfer solution for very large data sets. Snowball reduces the cost and time necessary to transfer data sets in the petabyte range. AWS sends the customer a Snowball device, the customer loads data onto the Snowball device and ships it back to AWS, and AWS loads the data directly into the customer's S3 environment.

A relatively new solution from AWS is the File Gateway solution. Similar to Storage Gateway, File Gateway is deployed on-premise as a virtual machine. The File Gateway integrates with the corporate data center through NFS. Corporate compute resources mount the NFS file system to store and retrieve files. Stored files are replicated to AWS and stored in S3 where lifecycle policies manage the files' disposition.

S3 buckets are regional assets, meaning that the S3 bucket is located and addressable through a specific AWS region. To increase durability and survivability of critical business data, S3 supports copying data between buckets in different regions through Cross-Region Replication. S3 buckets may be configured to automatically and asynchronously copy new objects across regions. Some customers leverage Cross-Region Replication to ensure data survivability across multiple regions.

Within the context of the Backup & Recovery strategy of disaster recovery, AWS offers several solutions, ranging from direct internet connectivity to an S3 bucket to using Snowball to transfer petabytes worth of information into the cloud.

Pilot Light Strategy

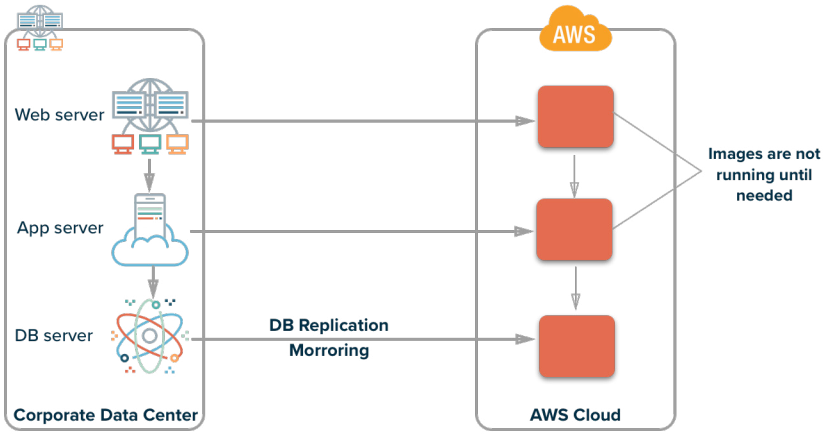
The Pilot Light strategy takes disaster recovery a little further, maintaining a small footprint in the AWS environment so that business operations may commence in AWS in the advent of a corporate data center failure.
Metaphorically, a pilot light keeps a gas appliance (e.g., a water heater) primed for use. When water needs to be heated, the water heater uses the pilot light to ignite the furnace. Similarly, the basic level of application capability is duplicated into the cloud. When disaster strikes, the disaster recovery environment can be brought online by leveraging the small amount of information stored in the cloud. In this case, however, information extends beyond the traditional object or database storage mechanisms to also include snapshots of virtual images and similar elements of the application.

At a minimum, application data must be replicated to the AWS disaster recovery environment. Databases must be replicated or mirrored. Data may be replicated to EC2 instances or AWS database services, depending on the type of database in use. In the prior case, EC2 may be used for database software that is managed by the customer. The customer maintains the EC2 instance (patches the operating system and database software) and sets up replication between the data center and the cloud instance. If the customer is leveraging AWS RDS, DynamoDB, or Redshift, AWS manages the database instance, while the customer sets up replication between the corporate database to the AWS-based DR database.

However, many applications have application-specific logic and support software stacks that must also be replicated in the cloud-based DR environment. Maintaining up-to-date application stacks in AWS allows for the DR environment to be brought online much more quickly than when building application stacks from scratch.

There are two common approaches to creating application stacks in AWS. The first is to create custom images in AWS by building the application stack on a base image (AMI) available from the AWS Marketplace. Once the image is constructed, it may be referenced when manually initializing the environment or through automation in CloudFormation templates. Note that these images must be kept up-to-date with corporate data center operating system, patches, and application updates. Otherwise, the DR environment may not function as expected.

The second approach is to migrate VMWare images directly into AWS. One method to do that is to create Elastic Block Store (EBS) images from VMFS artifacts. VMFS stores block images of VMWare images and snapshots. Coupled with the Storage Gateway, VMWare images and snapshots may be replicated to the AWS DR environment. These images and snapshots may be stored as EBS file system snapshots and from there, EC2 images may be created from those EBS artifacts.



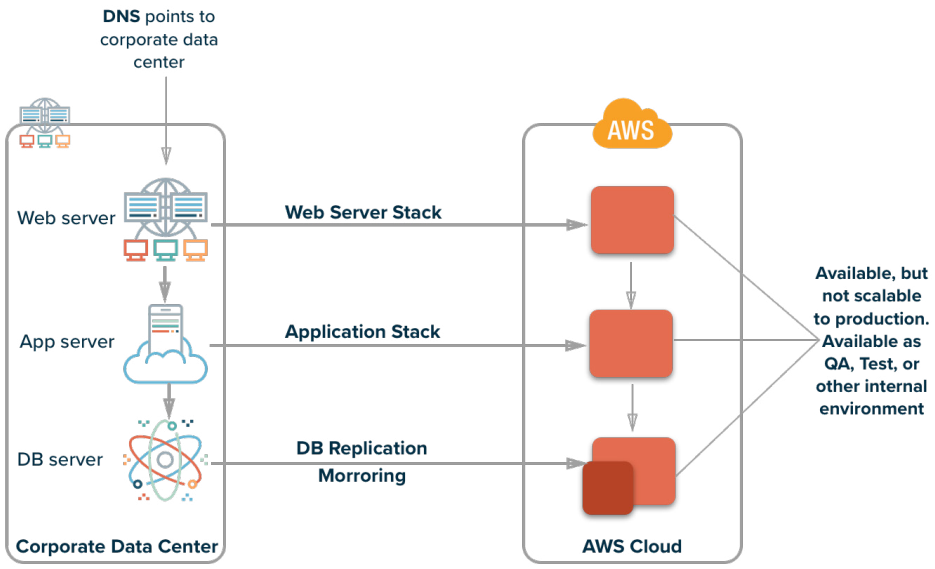
When the DR environment is needed, the images can be spun up and the environment prepared to assume the role of the corporate data center. Domain name service (DNS) can be manually or automatically pointed to the DR environment. Horizontal scaling can be achieved using Elastic Load Balancers and Auto Scaling Groups to right-size the fleet to meet demand. The database may require vertical scaling to handle the product load being placed on it.

Finally, the AWS-based Disaster Recovery environment should be hosted in a different geographical region from the corporate data center. Inter-regional deployment of a disaster recovery environment is a best practice.

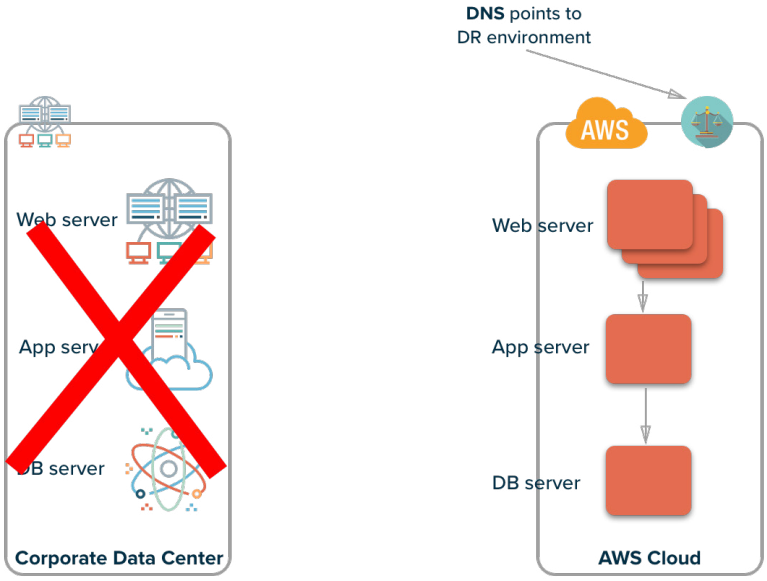
Warm Standby Strategy

The Warm Standby strategy takes the Pilot Light strategy a step further and maintains a fully functional environment in the DR environment. However, fully functional does not imply the DR environment is sized to handle production-level traffic. Because the DR environment is fully functional, it is commonly used as a QA, testing, or training environment by the organization.

The warm standby strategy is conceptually outlined in the diagram below. The production environment is hosted in the corporate data center, and DNS directs production traffic to the data center. The DR environment, running in AWS, maintains a running environment that matches production in terms of application software versions and patches. However, the DR environment is not designed to handle production-level traffic.



When the disaster recovery environment is required to assume production-level traffic, DNS may be switched over to the DR environment. Prior to assuming production-level traffic, the environment must be modified to scale. The database should be scaled vertically using larger EC2 instance types in order for it to properly handle additional transactional traffic. Web and application servers should be scaled horizontally (horizontal scaling is preferred over vertical scaling in AWS).



Scaling horizontally includes the use of Application Load Balancers and Auto Scaling Groups (ASGs). Application Load Balancers distribute incoming HTTP/HTTPS traffic across all available targets (instances). Application Load Balancers also monitor the health of each target and will only send requests to healthy hosts.

An ASG maintains a logical collection of EC2 instances that have similar attributes and functions (e.g., web server, application server). An ASG ensures a minimum number of EC2 instances are available and provides the capability to elastically scale by adding and removing EC2 instances to match traffic patterns. ASGs manage EC2 instance lifecycles by user-defined rules. For example, a scaling rule might define that a new EC2 instance be added to the group when CPU traffic exceeds 75%; another rule removes EC2 instances when average CPU use is below 55%.

Application Load Balancers and ASGs integrate together. The ASG will register and deregister EC2 instances as they are available or are removed from service.

Multi-Site Strategy

Commonly known as the ‘active-active’ configuration, the Multi-Site strategy provides a fully functional and scalable disaster recovery environment that operates synchronously with the corporate data center. DNS may be configured to route a portion of production traffic through both the DR and production environments. In short, the DR environment performs much the same as the production environment.

However, there are some differences. First, the application servers may point to a single environment for database access. Applications leverage data stores for transactional and persistent data necessary to function. That being said, it is best to keep the production and DR databases synchronized, as this allows the applications to leverage one database and fail over to the DR environment if the production database fails.

Scalability is the next environment change to tackle. While running in parallel to the production environment, the DR environment may not be configured to assume full production load. In such cases, Elastic Load Balancers, Auto Scaling Groups, and other mechanisms may be leveraged to scale out the DR environment.

Summary

This paper has briefly discussed four general disaster recovery strategies and how these strategies may be implemented on an AWS platform. Additionally, the paper has covered several AWS services that may contribute to a cloud-based disaster recovery platform. In all, the theme of this paper is that AWS is an extremely powerful and low-cost solution that demonstrates superior TCO for disaster recovery solutions.

About Levvel

Levvel helps clients transform their business with strategic consulting and technical execution services. We work with your IT organization, product groups, and innovation teams to design and deliver on your technical priorities.

Levvel's cloud experts combine decades of traditional architecture, development, security, and infrastructure experience with a complete mastery of available and emerging cloud offerings. Our client-centric approach focuses first on understanding your business needs and goals, then selecting the right cloud technology to make you efficient, agile, and scalable. We tailor custom solutions to fit within your business processes, simultaneously reducing TCO and downtime while increasing productivity, security, ROI, and speed to market.

If you are interested in understanding how AWS may contribute to your business through a disaster recovery infrastructure with reduced expenses, contact Levvel at hello@levvel.io.

hello@levvel.io
980.278.3065