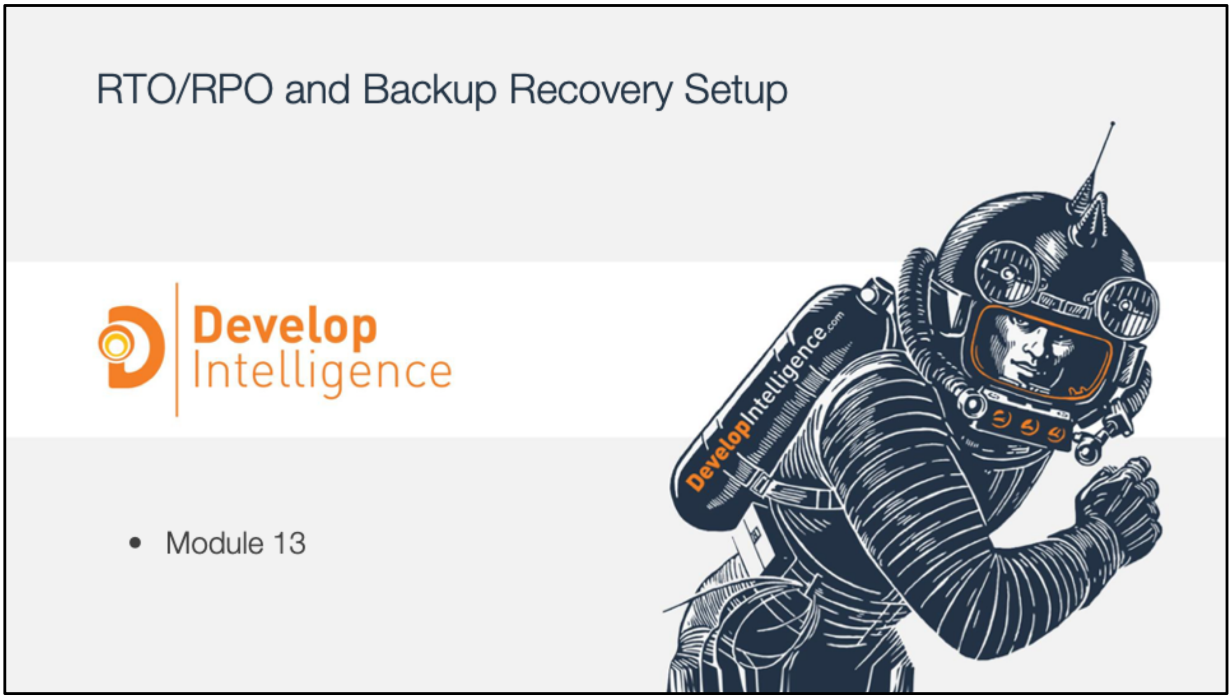
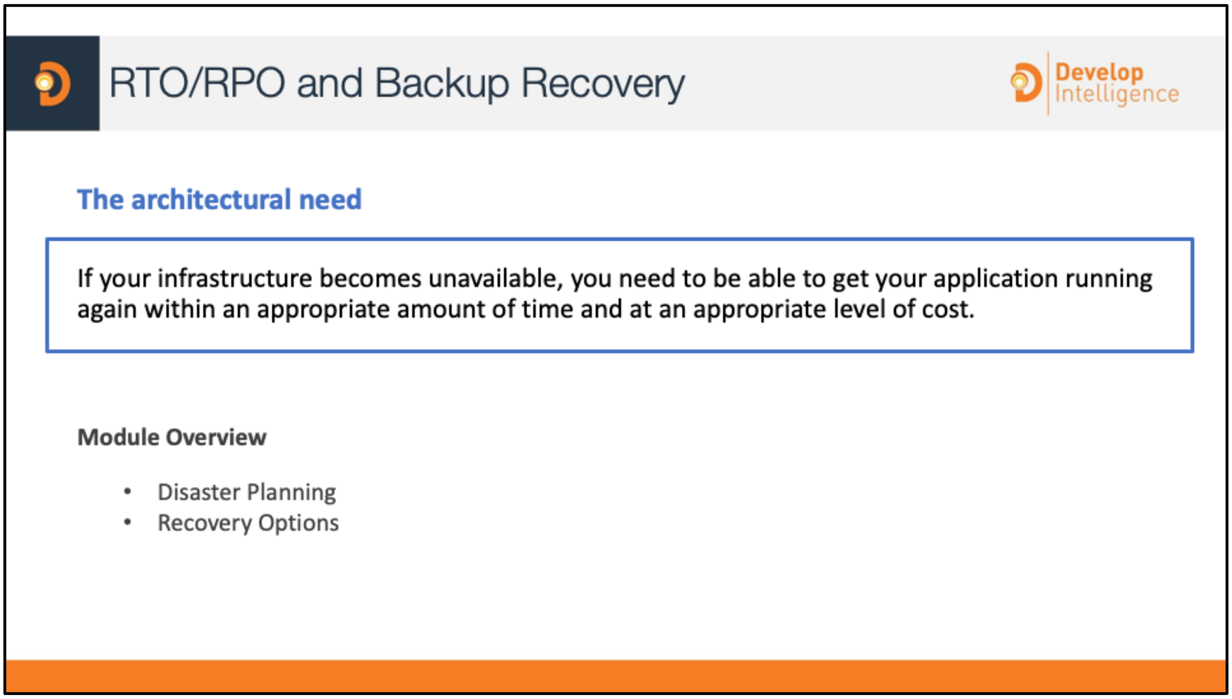


RTO/RPO and Backup Recovery Setup

Develop Intelligence

- Module 13

DevelopIntelligence.com



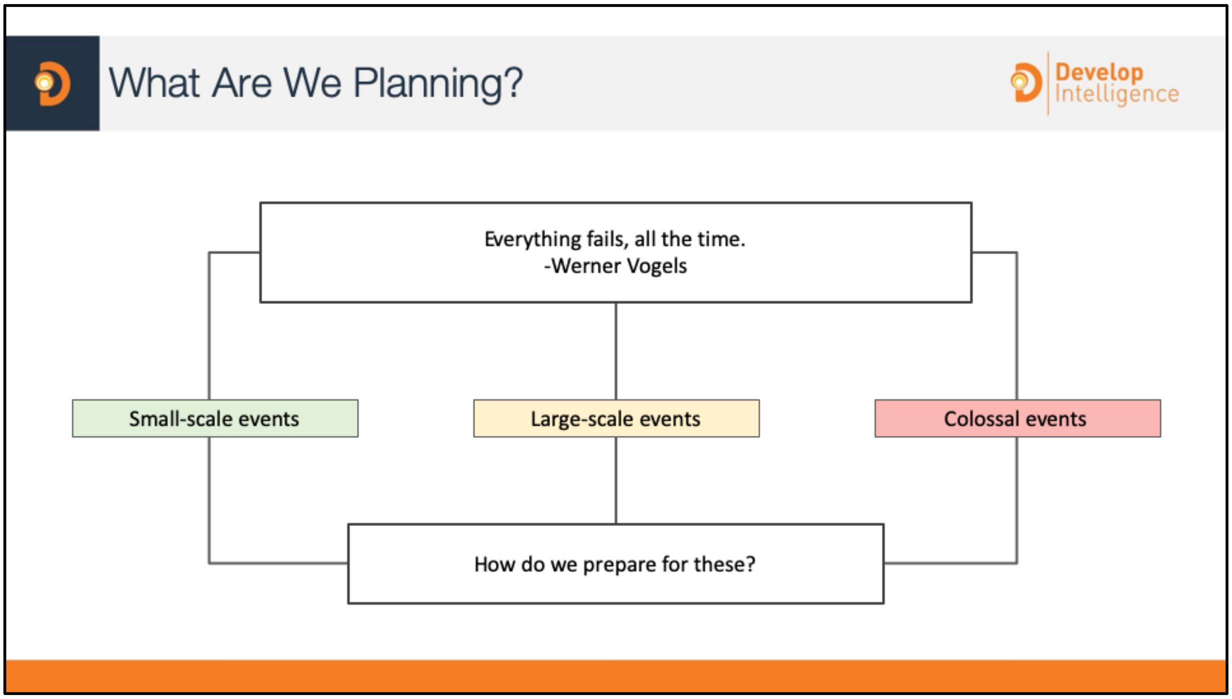
RTO/RPO and Backup Recovery

The architectural need

If your infrastructure becomes unavailable, you need to be able to get your application running again within an appropriate amount of time and at an appropriate level of cost.

Module Overview

- Disaster Planning
- Recovery Options



What Are We Planning?

Everything fails, all the time.
-Werner Vogels

Small-scale events

Large-scale events

Colossal events

How do we prepare for these?

What kind of disaster are you planning for?

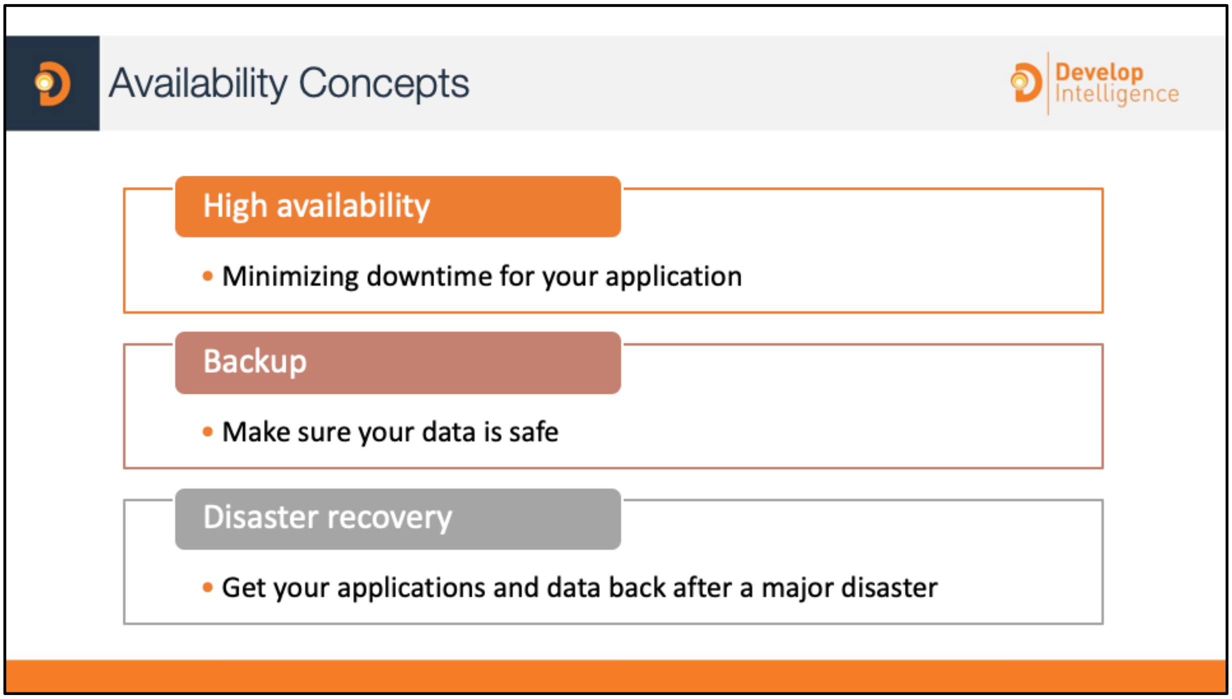
- A small-scale event where you simply need to get a restoration and backup?
- A larger-scale event where multiple resources are impacted?
- A colossal scale event where multiple people and resources will be impacted?

Disaster recovery (DR) is about preparing for and recovering from a disaster. Any event that has a negative impact on a company's business continuity or finances could be termed a disaster. This includes hardware or software failure, a network outage, a power outage, physical damage to a building like fire or flooding, human error, or some other significant event.

To minimize the impact of a disaster, companies invest time and resources to plan and prepare, to train employees, and to document and update processes. The amount of investment for DR planning for a particular system can vary dramatically depending on the cost of a potential outage.

Companies that have traditional physical environments typically must duplicate their infrastructure to ensure the availability of spare capacity in the event of a disaster. The infrastructure needs to be procured, installed, and maintained so that it is ready to support the anticipated capacity requirements. During normal operations, the infrastructure typically is under-utilized or over-provisioned.

With AWS, your company can scale up its infrastructure on an as-needed, pay-as-you-go basis. You get access to the same highly secure, reliable, and fast infrastructure that Amazon uses to run its own global network of websites. AWS also gives you the flexibility to quickly change and optimize resources during a DR event, which can result in significant cost savings.



Availability Concepts

Develop Intelligence

High availability

- Minimizing downtime for your application

Backup

- Make sure your data is safe

Disaster recovery

- Get your applications and data back after a major disaster

Production systems typically come with defined or implicit objectives in terms of uptime. A system is **highly available** when it can withstand the failure of an individual or multiple components (e.g., hard disks, servers, network links etc.).

High availability provides redundancy and fault tolerance. Its goal is to ensure this service is always available even in the event of a failure.

Backup is critical to protect data and to ensure business continuity. At the same time, it can be a challenge to implement well. The pace at which data is generated is growing exponentially. The density and durability of local disk is not benefiting from the same growth rate. The enterprise backup has become its own industry.

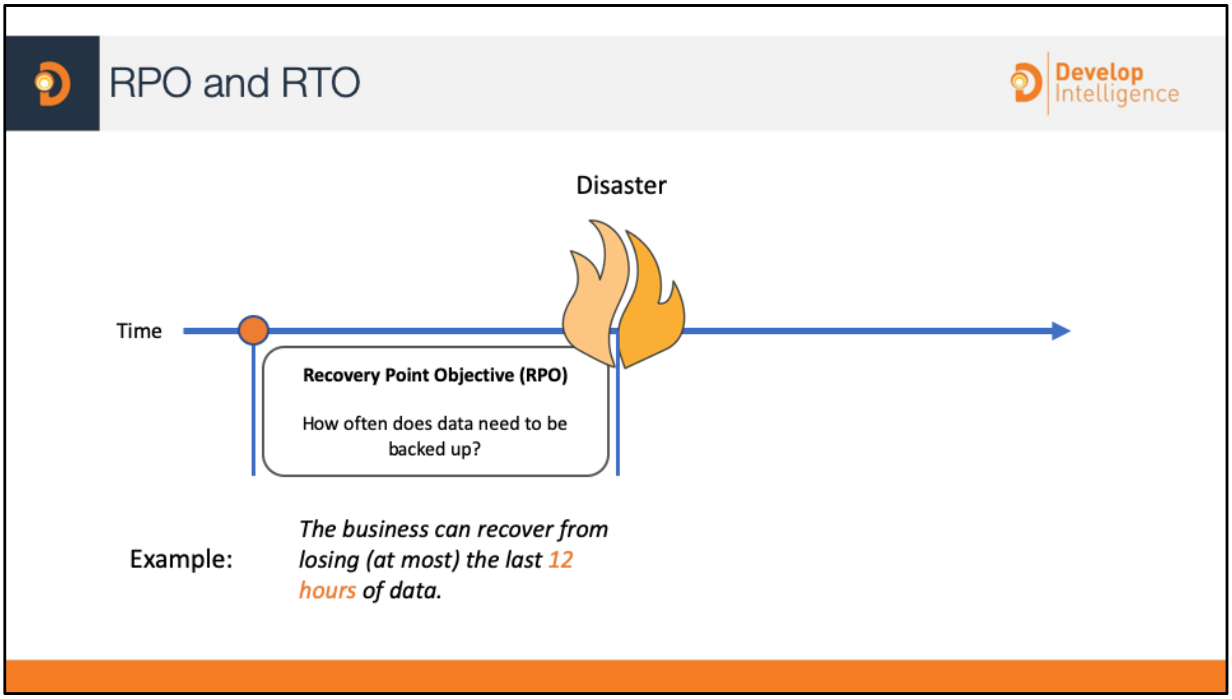
Data is generated on an arbitrarily large number of endpoints; laptops, desktops, servers, virtual machines, and now mobile devices, that is, the problem is distributed in nature. Current backup software is very centralized – the general model is to collect data from many devices and store it in single place. Sometimes a copy of that stored data is also sent to tape. The centralized approach has the potential to overwhelm the backup target during recovery from a disaster and result in broken recovery SLAs.

Enterprise backup scenarios used to look like this: If you wanted high performance data access, it had to live on disk. If you wanted cost-effective archival storage, it had to live on tape. If you wanted to archive off-site, you had to physically deliver your archival tapes to another location. Recovery from local disk was fine, unless you needed something from a tape, and it might have been a while if that tape wasn't on site.

The cloud has changed things. Backup software can write to the cloud without any changes to the backup software itself. (This will be discussed later.)

Disaster recovery (DR) is about preparing for and recovering from a disaster. A *disaster* is any event that has a negative impact on a company's business continuity or finances—including hardware or software failure, a network outage, a power outage, physical damage to a building like fire or flooding, human error, or some other significant event.

To minimize the impact of a disaster, companies invest time and resources to plan and prepare, to train employees, and to document and update processes. The amount of investment for DR planning for a particular system can vary dramatically depending on the cost of a potential outage. Companies that have traditional physical environments typically must duplicate their infrastructure to ensure the availability of spare capacity in the event of a disaster. The infrastructure needs to be procured, installed, and maintained so that it is ready to support the anticipated capacity requirements. During normal operations, the infrastructure typically is under-utilized or over-provisioned.



RPO and RTO

Disaster

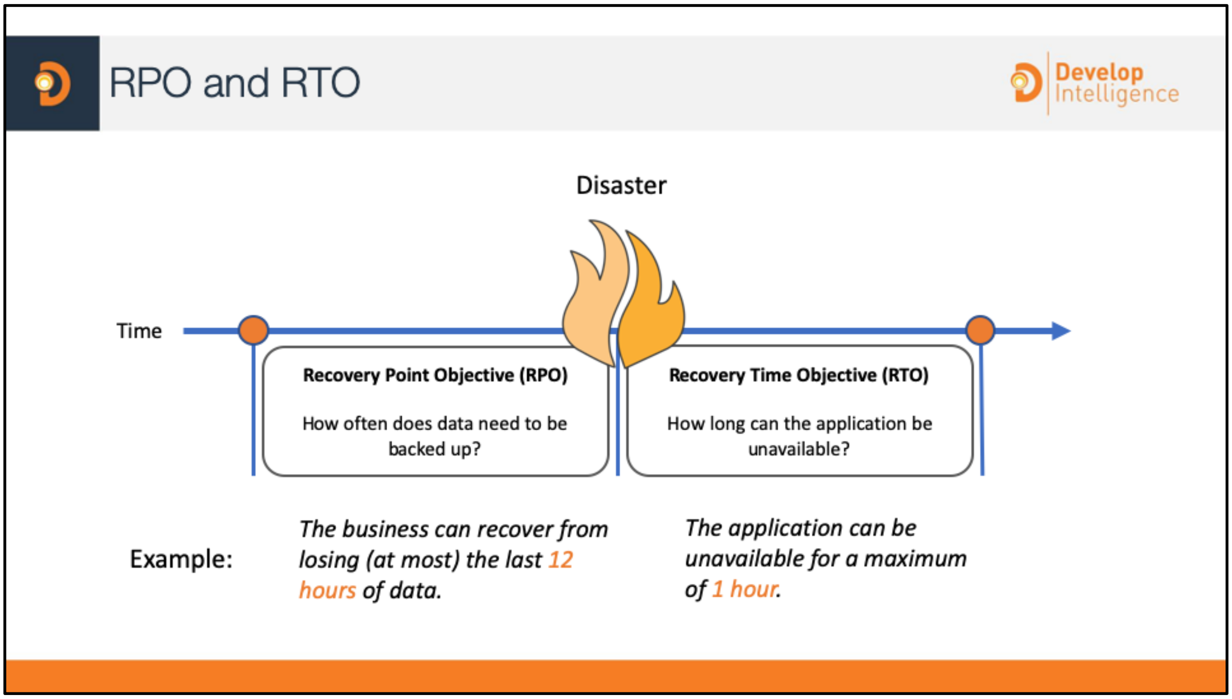
Time

Recovery Point Objective (RPO)

How often does data need to be backed up?

Example: *The business can recover from losing (at most) the last 12 hours of data.*

Recovery point objective (RPO) is the acceptable amount of data loss measured in time. For example, if a disaster occurs at 12:00 PM (noon) and the RPO is one hour, the system should recover all data that was in the system before 11:00 AM. Data loss will span only one hour, between 11:00 AM and 12:00 PM (noon).



RPO and RTO

Disaster

Time

Recovery Point Objective (RPO)

How often does data need to be backed up?

Recovery Time Objective (RTO)

How long can the application be unavailable?

Example:

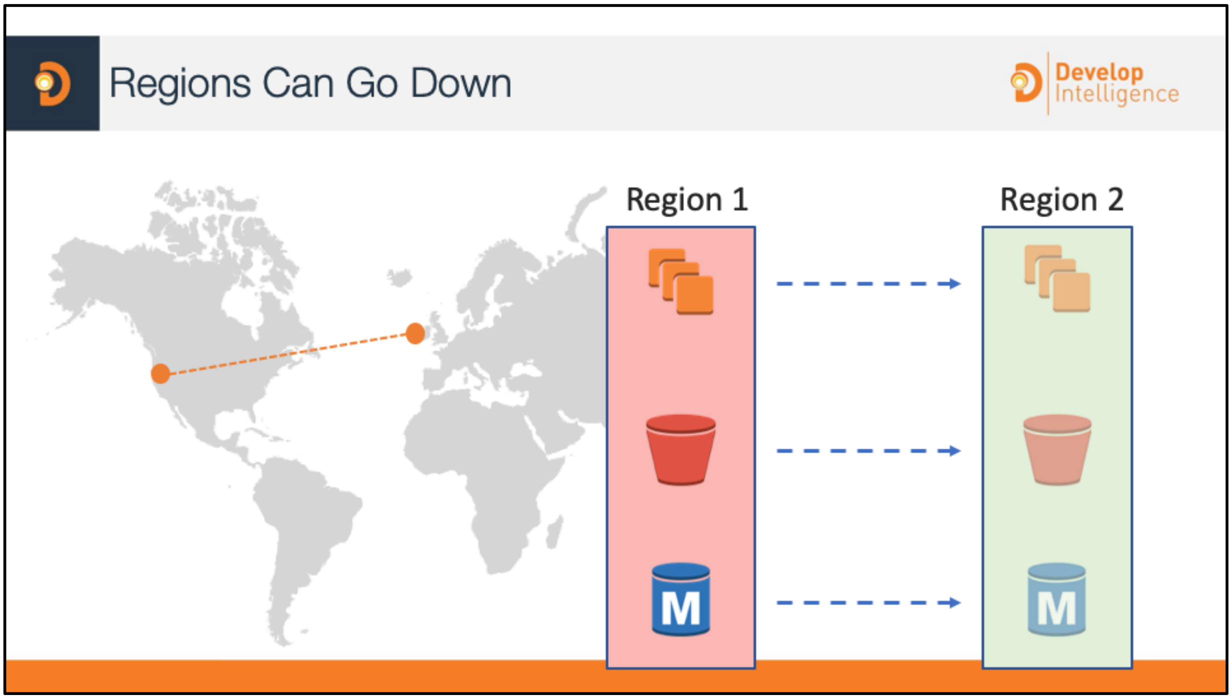
The business can recover from losing (at most) the last 12 hours of data.

The application can be unavailable for a maximum of 1 hour.

Recovery time objective (RTO) is the time it takes after a disruption to restore a business process to its service level, as defined by the operational level agreement (OLA). For example, if a disaster occurs at 12:00 PM (noon) and the RTO is eight hours, the DR process should restore the business process to the acceptable service level by 8:00 PM.

A company typically decides on an acceptable RPO and RTO based on the financial impact to the business when systems are unavailable. The company determines financial impact by considering many factors, such as the loss of business and damage to its reputation due to downtime and the lack of systems availability.

IT organizations then plan solutions to provide cost-effective system recovery based on the RPO within the timeline and the service level established by the RTO.



Regions Can Go Down

AWS is available in multiple regions around the globe, so you can choose the most appropriate location for your DR site, in addition to the site where your system is fully deployed.

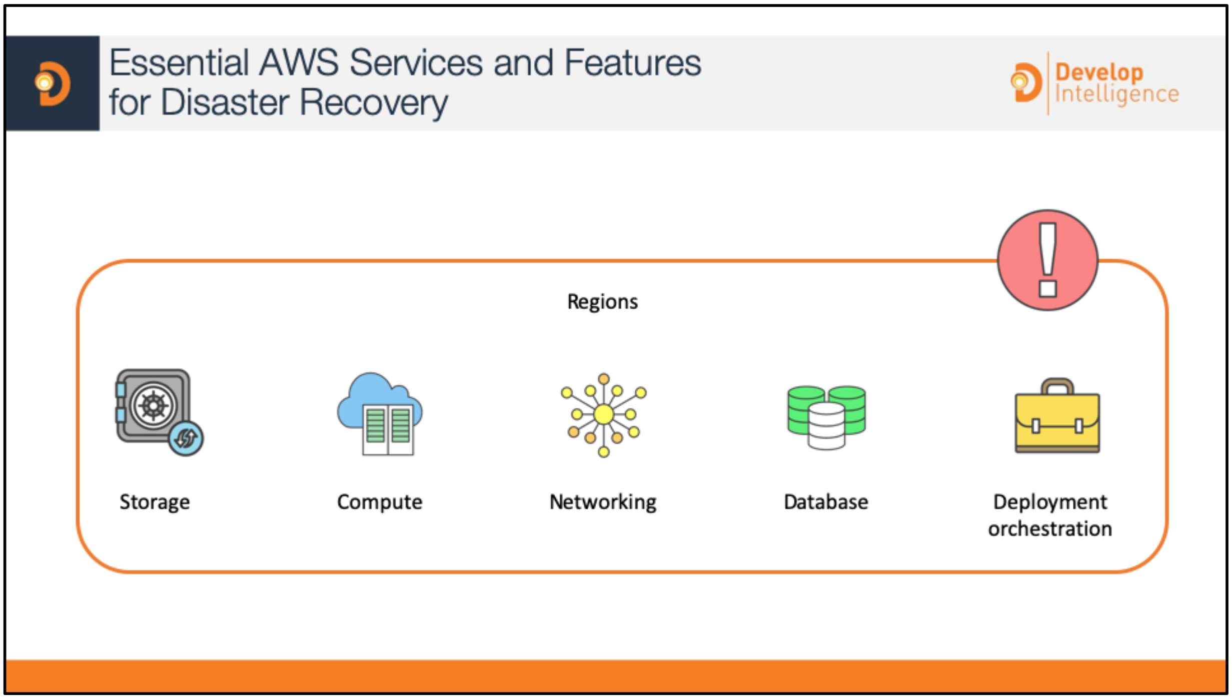
It’s highly unlikely for a region to be unavailable. But if some very large-scale event impacts a region—for instance, a meteor strike—it is within the realm of possibility.

AWS maintains a page that inventories current services offered by region (products and services by region). AWS maintains a strict region isolation policy so that any large-scale event in one region will not impact any other region. We encourage our customers to take a similar approach to their multi-region strategy. Each region should be able to be taken offline with no impact to any other region.

If you have an AWS Direct Connect (DX) circuit to any AWS Region in the United States, it will provide you with access to all regions in the US, including AWS GovCloud, without that traffic going through the public internet.

Also consider how applications are deployed. If you deploy to each region separately, you can isolate that region in case of disaster, and transfer all your traffic to another region.

If you are deploying new applications and infrastructure rapidly, you may want to have an active-active region. Let’s say you deploy something that causes a region's applications to be unavailable or misbehaving. You can remove the region from the active record set in Route 53, identify the root cause, and roll back the change before re-enabling the region.



Essential AWS Services and Features for Disaster Recovery

Regions

Storage

Compute

Networking

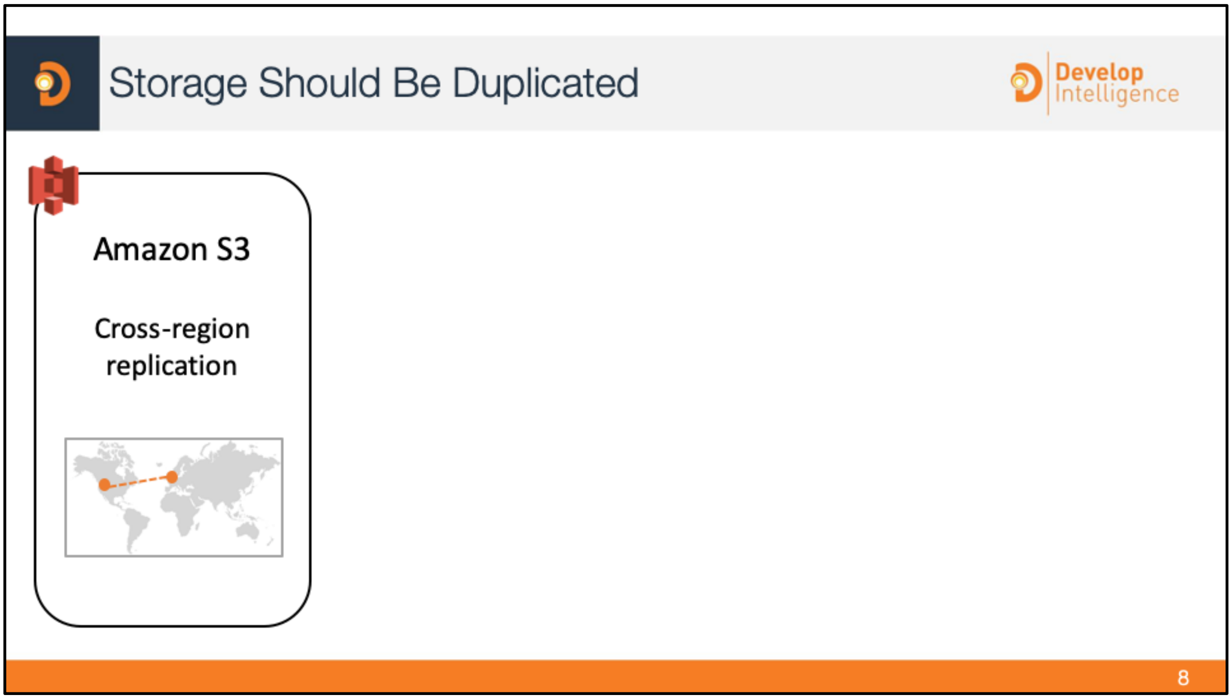
Database

Deployment orchestration

Before discussing the various approaches to disaster recovery, it is important to review the AWS services and features that are the most relevant to it. This section provides a summary.

When planning for DR, it is important to consider the use of services and features that support data migration and durable storage, because they enable you to restore backed-up, critical data to AWS when disaster strikes. For some of the scenarios that involve either a scaled-down or a fully scaled deployment of your system in AWS, compute resources will be required as well.

During a disaster, you need to either spin up new resources or failover to existing pre-configured resources. These resources not only include code and content, but other pieces such as DNS entries, network firewall rules, and virtual machines/instances.



Storage Should Be Duplicated

Amazon S3

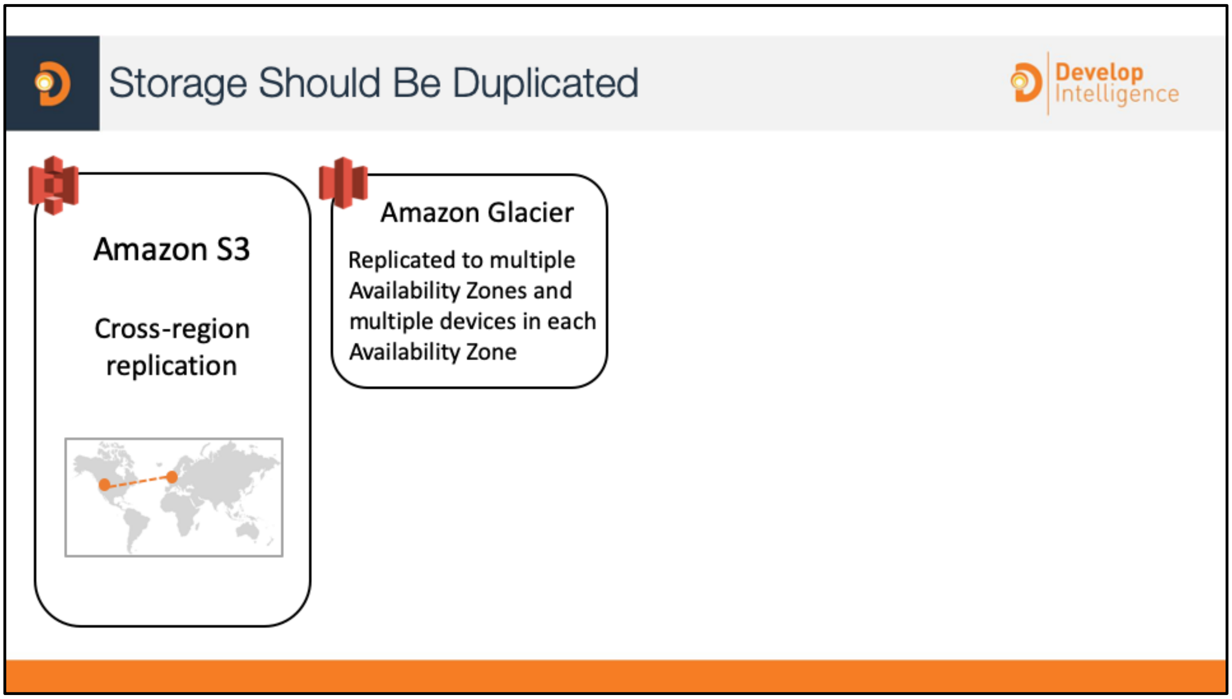
Cross-region replication

8

AWS offers many different ways of storing your data. Each service has different capabilities, so that you can match the right service with the right need for each of your systems.

Amazon S3 provides a highly durable storage infrastructure designed for mission critical and primary data storage. Objects are redundantly stored on multiple devices across multiple facilities within a region, designed to provide a durability of 99.999999999% (119s). AWS provides further protection for data retention and archiving through versioning in Amazon S3, AWS MFA, bucket policies, and AWS IAM. Cross-region replication is a bucket-level configuration that enables automatic, asynchronous copying of objects across buckets in different AWS Regions. These buckets are called *source* bucket and *destination* bucket, and they can be owned by different AWS accounts.

To activate this feature, you add a replication configuration to your source bucket to direct Amazon S3 to replicate objects according to the configuration.



Storage Should Be Duplicated

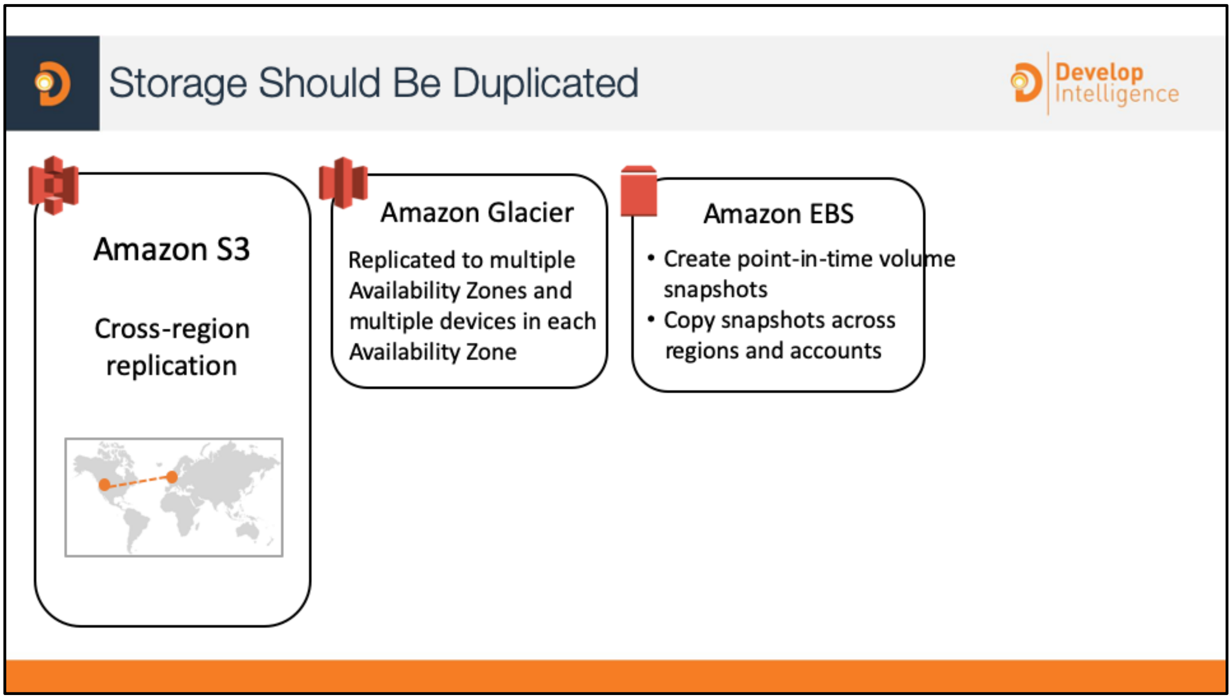
Amazon S3

Cross-region replication

Amazon Glacier

Replicated to multiple Availability Zones and multiple devices in each Availability Zone

Amazon Glacier provides extremely low-cost storage for data archiving and backup. Objects (or *archives*, as they are known in Amazon Glacier) are optimized for infrequent access, for which retrieval times of several hours are adequate. Amazon Glacier is designed for the same durability as Amazon S3. Although you need to maintain your own index of data you upload to Amazon Glacier, an inventory of all archives in each of your vaults is maintained for disaster recovery or occasional reconciliation purposes. The vault inventory is updated approximately once a day. You can request a vault inventory as either a JSON or CSV file and will contain details about the archives within your vault including the size, creation date and the archive description (if you provided one during upload). The inventory will represent the state of the vault at the time of the most recent inventory update.



Storage Should Be Duplicated

Develop Intelligence

Amazon S3

Cross-region replication

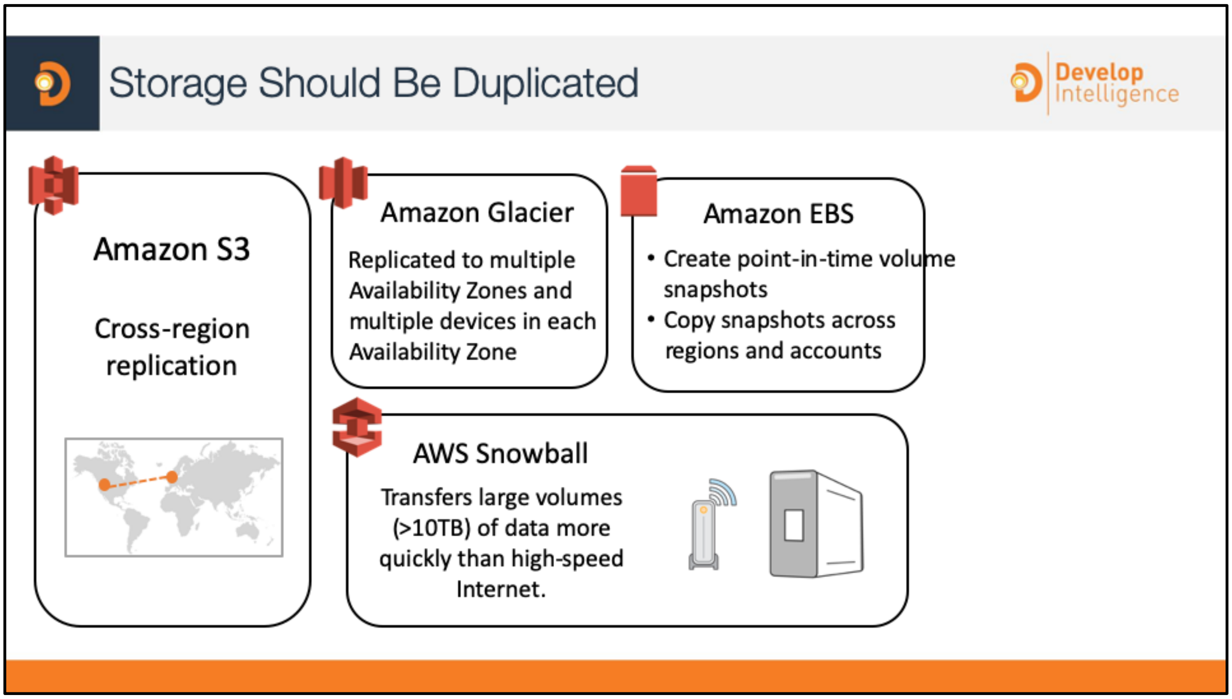
Amazon Glacier

Replicated to multiple Availability Zones and multiple devices in each Availability Zone

Amazon EBS

- Create point-in-time volume snapshots
- Copy snapshots across regions and accounts

Amazon EBS gives you the ability to create point-in-time snapshots of data volumes. You can use the snapshots as the starting point for new Amazon EBS volumes, and you can protect your data for long-term durability because snapshots are stored within Amazon S3. After a volume is created, you can attach it to a running Amazon EC2 instance. Amazon EBS volumes provide off-instance storage that persists independently from the life of an instance and is replicated across multiple servers in an Availability Zone to prevent the loss of data from the failure of any single component. After you've created a snapshot and it has finished copying to Amazon S3 (when the snapshot status is completed), you can copy it from one AWS region to another, or within the same region. Amazon S3 server-side encryption (256-bit AES) protects a snapshot's data in-transit during a copy operation. The snapshot copy receives an ID that is different than the ID of the original snapshot.



Storage Should Be Duplicated

Develop Intelligence

Amazon S3

Cross-region replication

Amazon Glacier

Replicated to multiple Availability Zones and multiple devices in each Availability Zone

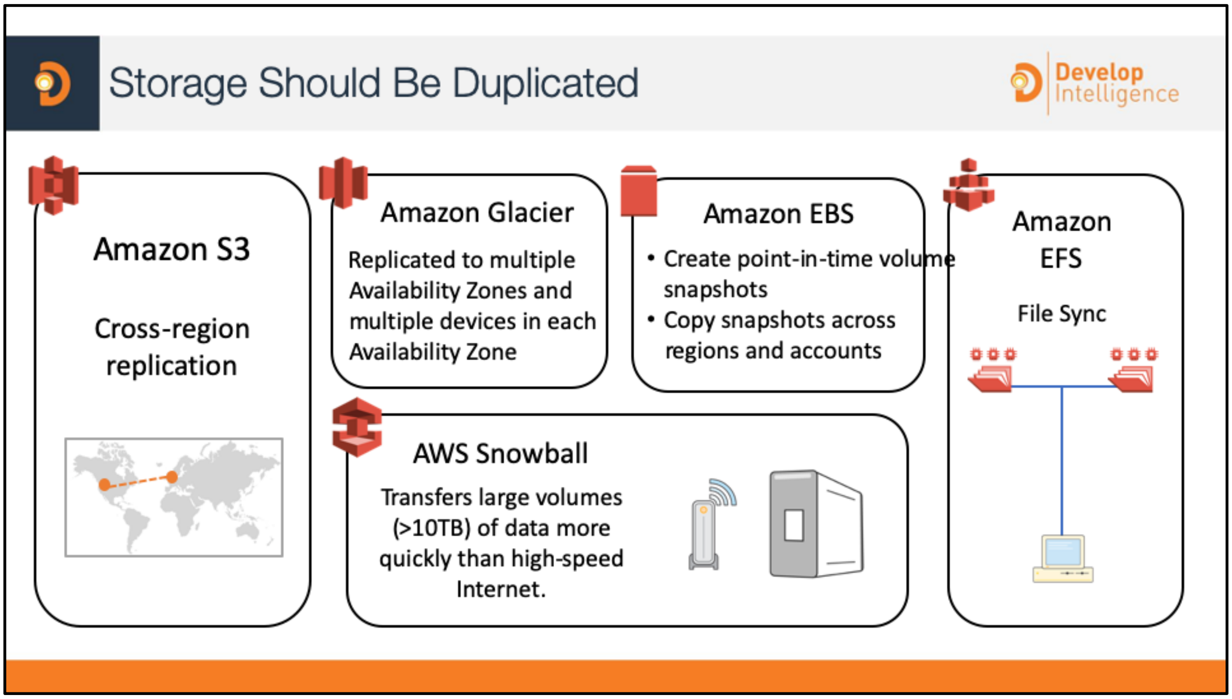
Amazon EBS

- Create point-in-time volume snapshots
- Copy snapshots across regions and accounts

AWS Snowball

Transfers large volumes (>10TB) of data more quickly than high-speed Internet.

AWS Snowball is a data transport solution that accelerates moving terabytes to petabytes of data into and out of AWS using storage devices designed to be secure for physical transport. Using Snowball helps to eliminate challenges that can be encountered with large-scale data transfers including high network costs, long transfer times, and security concerns. In the event that you need to quickly retrieve a large quantity of data stored in Amazon S3, Snowball devices can help retrieve the data much quicker than high-speed internet.



Storage Should Be Duplicated

Amazon S3

Cross-region replication

Amazon Glacier

Replicated to multiple Availability Zones and multiple devices in each Availability Zone

Amazon EBS

- Create point-in-time volume snapshots
- Copy snapshots across regions and accounts

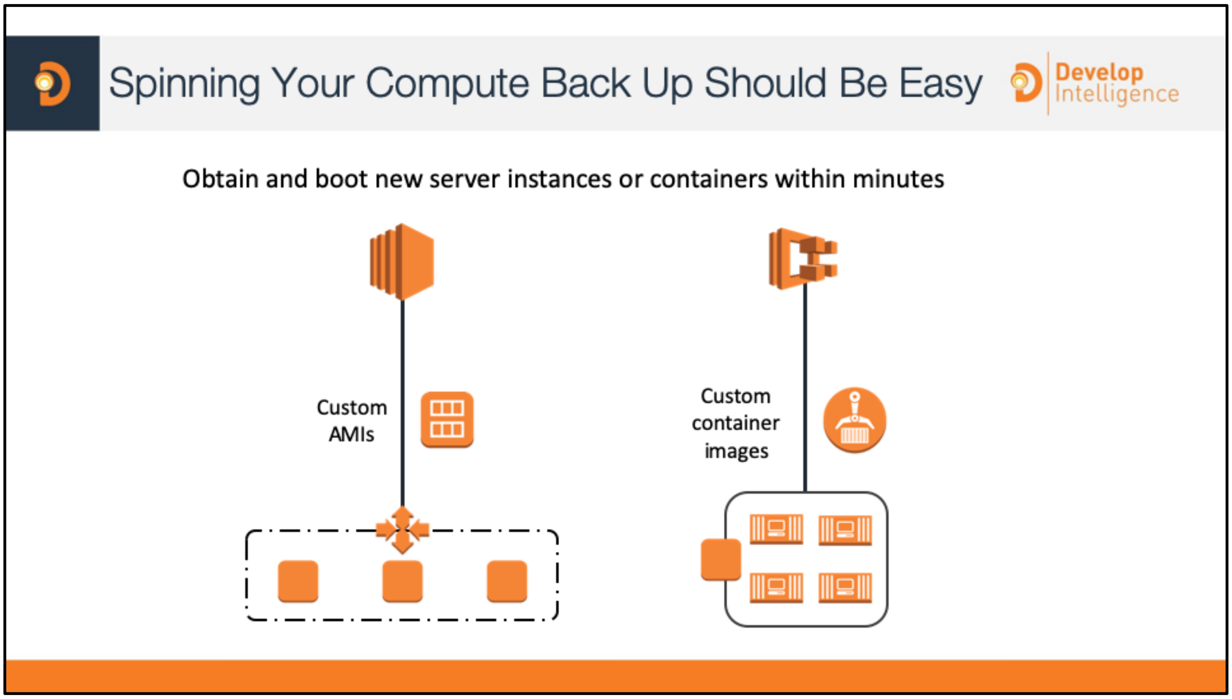
Amazon EFS

File Sync

AWS Snowball

Transfers large volumes (>10TB) of data more quickly than high-speed Internet.

Use **Amazon EFS** File Sync to efficiently and securely sync files from on-premises or in-cloud file systems to Amazon Elastic File System (Amazon EFS) at speeds of up to 5x faster than standard Linux copy tools. EFS File Sync securely and efficiently copies files over the internet or a DX connection.



Spinning Your Compute Back Up Should Be Easy

Obtain and boot new server instances or containers within minutes

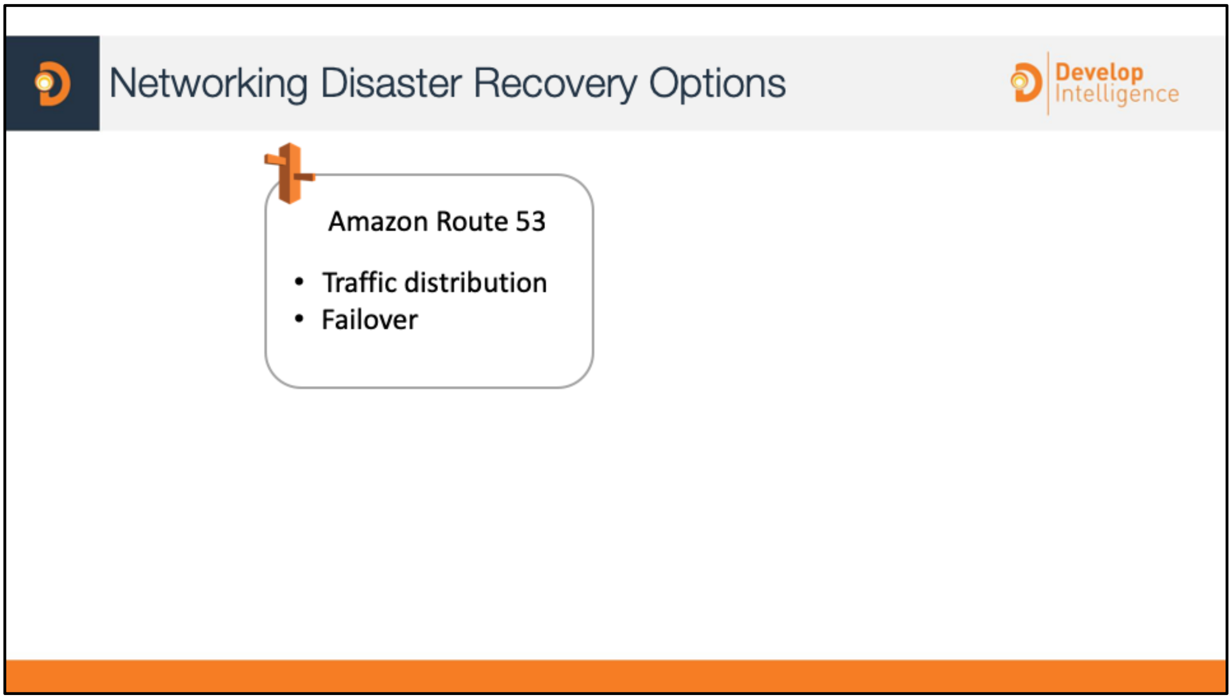
Custom AMIs

Custom container images

In the context of DR, it's critical to be able to rapidly create virtual machines that you control. By launching instances in separate Availability Zones, you can protect your applications from the failure of a single location.

You can arrange for automatic recovery of an EC2 instance when a system status check of the underlying hardware fails. The instance will be rebooted (on new hardware if necessary) but will retain its Instance Id, IP Address, Elastic IP Addresses, EBS Volume attachments, and other configuration details. In order for the recovery to be complete, you'll need to make sure that the instance automatically starts up any services or applications as part of its initialization process.

Amazon Machine Images (AMIs) are preconfigured with operating systems, and some preconfigured AMIs might also include application stacks. You can also configure your own AMIs. In the context of DR, AWS strongly recommends that you configure and identify your own AMIs so that they can launch as part of your recovery procedure. Such AMIs should be preconfigured with your operating system of choice plus appropriate pieces of the application stack.



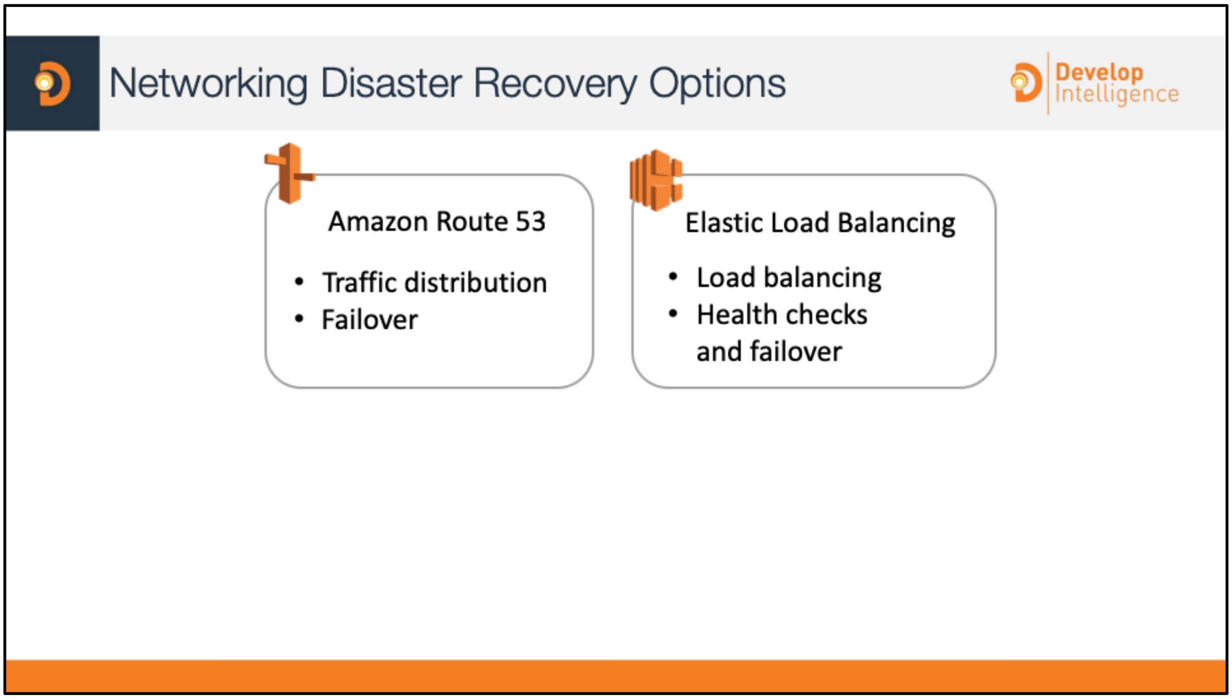
Networking Disaster Recovery Options

Amazon Route 53

- Traffic distribution
- Failover

When you are dealing with a disaster, it’s very likely that you will have to modify network settings as your system is failing over to another site. AWS offers several services and features that enable you to manage and modify network settings, such as Amazon Route 53, ELB, Amazon VPC, and DX.

Amazon Route 53 includes a number of global load balancing capabilities (which can be effective when you are dealing with DR scenarios such as DNS endpoint health checks) and the ability to failover between multiple endpoints and even static websites hosted in Amazon S3.



Networking Disaster Recovery Options

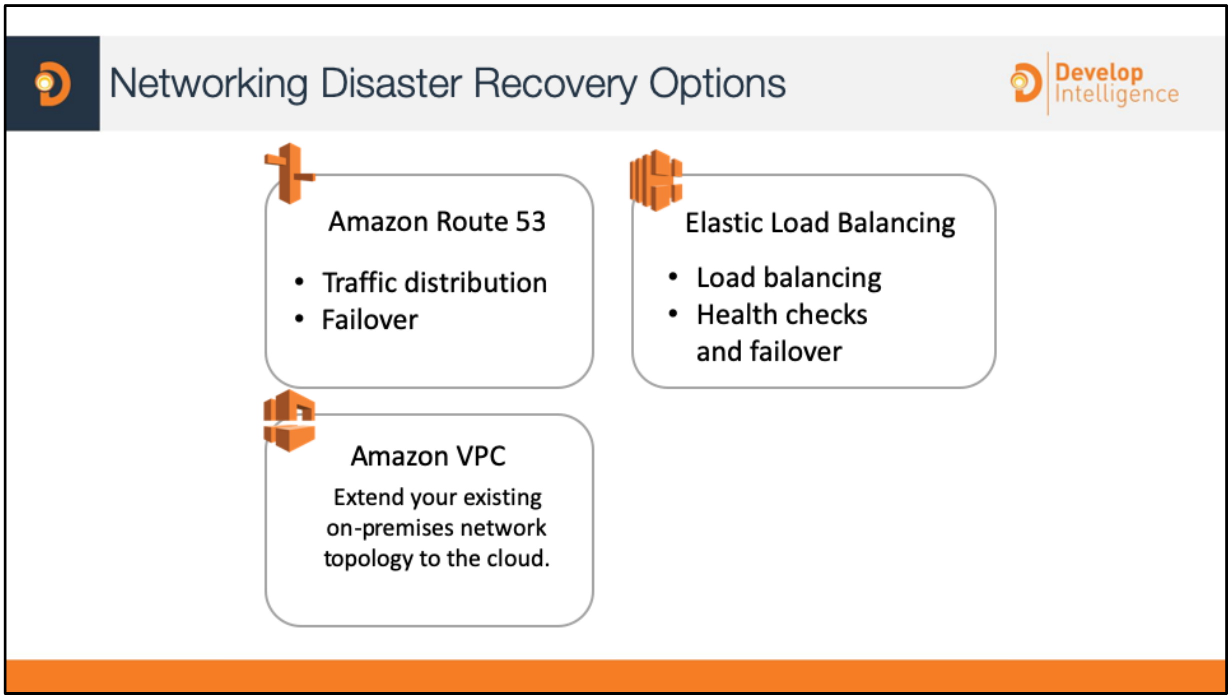
Amazon Route 53

- Traffic distribution
- Failover

Elastic Load Balancing

- Load balancing
- Health checks and failover

ELB automatically distributes incoming application traffic across multiple Amazon EC2 instances. It enables you to achieve even greater fault tolerance in your applications by seamlessly providing the load-balancing capacity that is needed in response to incoming application traffic. Just as you can pre-allocate Elastic IP addresses, you can pre-allocate your load balancer so that its DNS name already known, which can simplify the execution of your DR plan.



Networking Disaster Recovery Options

Develop Intelligence

Amazon Route 53

- Traffic distribution
- Failover

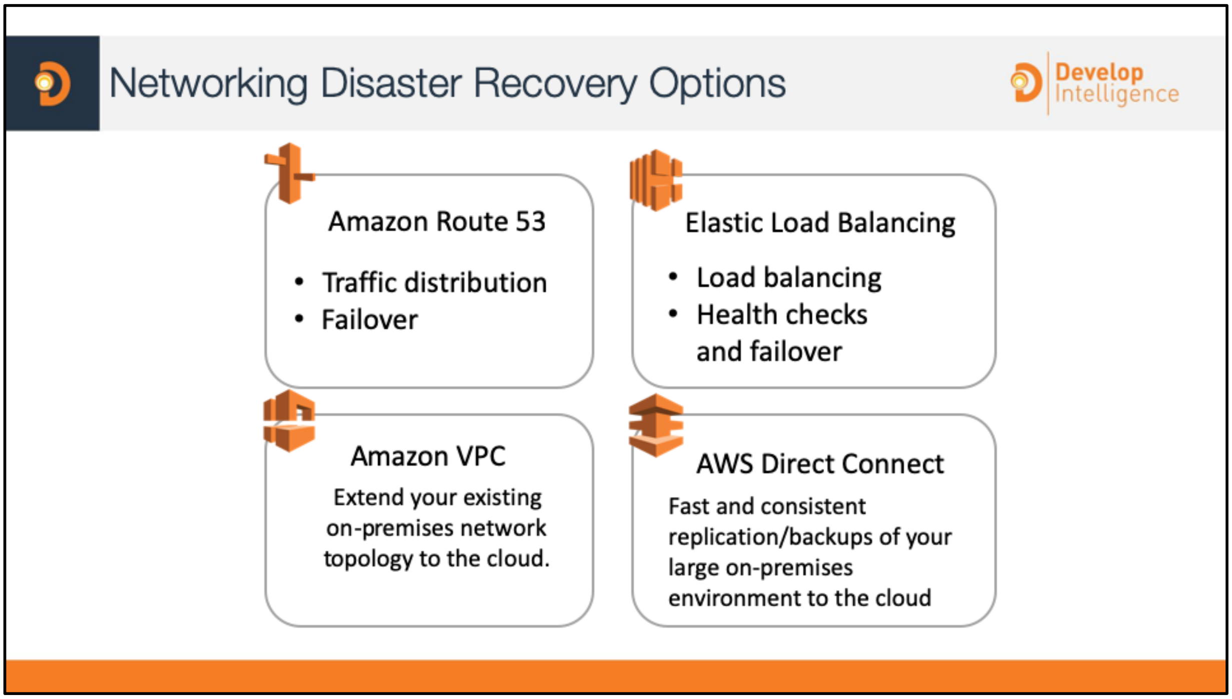
Elastic Load Balancing

- Load balancing
- Health checks and failover

Amazon VPC

Extend your existing on-premises network topology to the cloud.

In the context of DR, you can use **Amazon VPC** to extend your existing network topology to the cloud. This can be especially appropriate when recovering enterprise applications that are typically on the internal network.



Networking Disaster Recovery Options

Amazon Route 53

- Traffic distribution
- Failover

Elastic Load Balancing

- Load balancing
- Health checks and failover

Amazon VPC

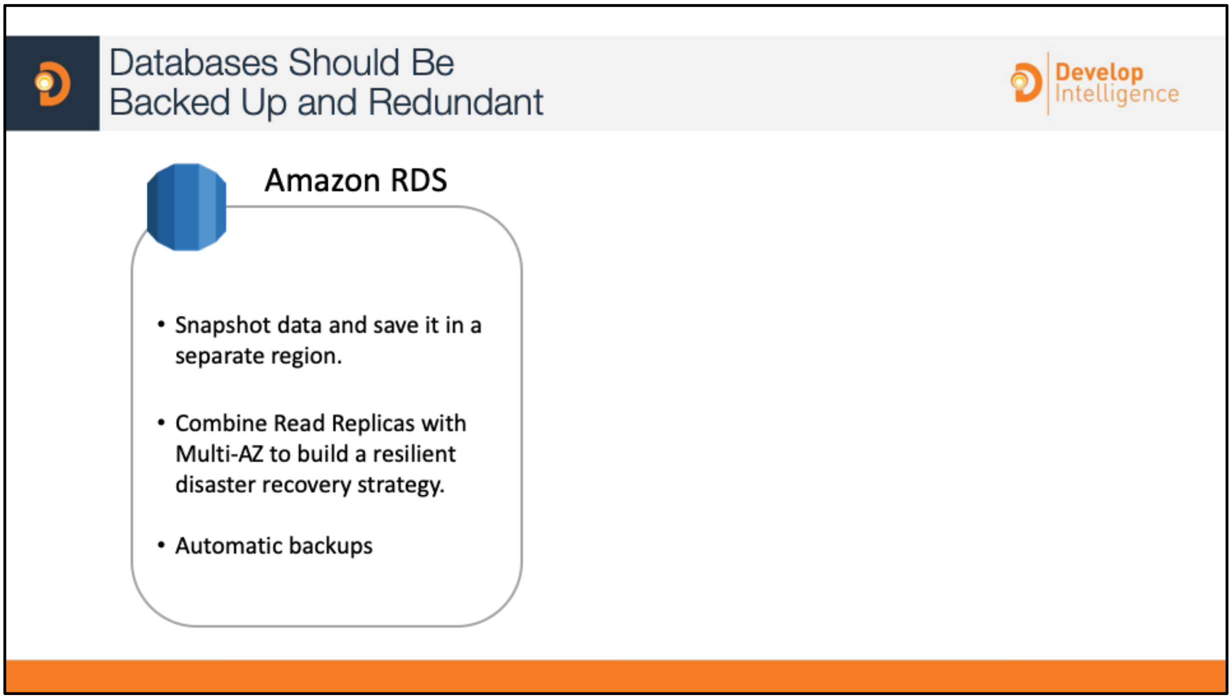
Extend your existing on-premises network topology to the cloud.

AWS Direct Connect

Fast and consistent replication/backups of your large on-premises environment to the cloud

AWS Direct Connect (DX) makes it easy to set up a dedicated network connection from your premises to AWS. In many cases, this can reduce your network costs, increase bandwidth throughput, and provide a more consistent network experience than internet-based connections.

For information on using AWS Direct Connect for high resiliency for critical workloads, see https://aws.amazon.com/directconnect/resiliency-recommendation/



Databases Should Be Backed Up and Redundant

Develop Intelligence

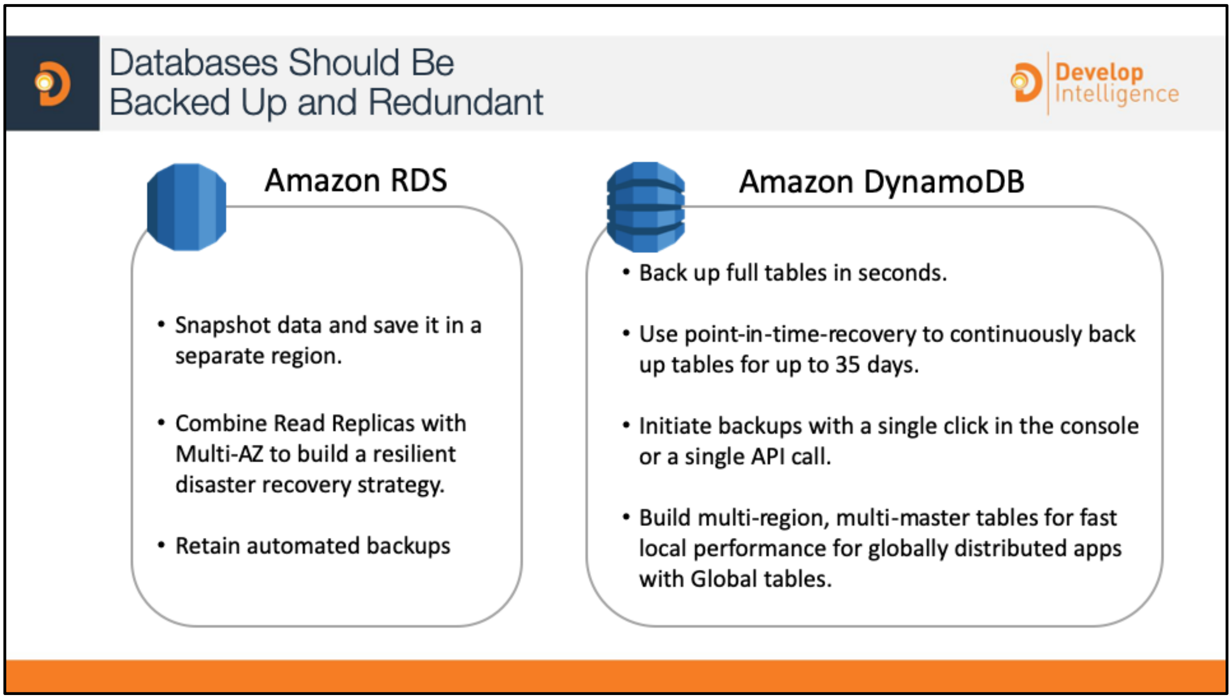
Amazon RDS

- Snapshot data and save it in a separate region.
- Combine Read Replicas with Multi-AZ to build a resilient disaster recovery strategy.
- Automatic backups

For your database needs, consider using these AWS services: Amazon RDS, Amazon DynamoDB, and Amazon Redshift.

You can use **Amazon RDS** either in the preparation phase for DR to hold your critical data in a database that is already running, or in the recovery phase to run your production database. When you want to look at multiple regions, Amazon RDS gives you the ability to snapshot data from one region to another, and also to have a read replica running in another region. Using Amazon RDS, you can share a manual DB snapshot or DB cluster snapshot. You can share a manual snapshot with up to 20 other AWS accounts. You can also share an unencrypted manual snapshot as public, which makes the snapshot available to all AWS accounts. Take care when sharing a snapshot as public so that none of your private information is included in any of your public snapshots.

Amazon RDS Read Replicas for MySQL and MariaDB now support Multi-AZ deployments. Combining Read Replicas with Multi-AZ enables you to build a resilient disaster recovery strategy and simplify your database engine upgrade process. Amazon RDS Read Replicas enable you to create one or more read-only copies of your database instance within the same AWS Region or in a different AWS Region. Updates made to the source database are then asynchronously copied to your Read Replicas. In addition to providing scalability for read-heavy workloads, Read Replicas can be promoted to become a standalone database instance when needed.



Databases Should Be Backed Up and Redundant

Develop Intelligence

Amazon RDS

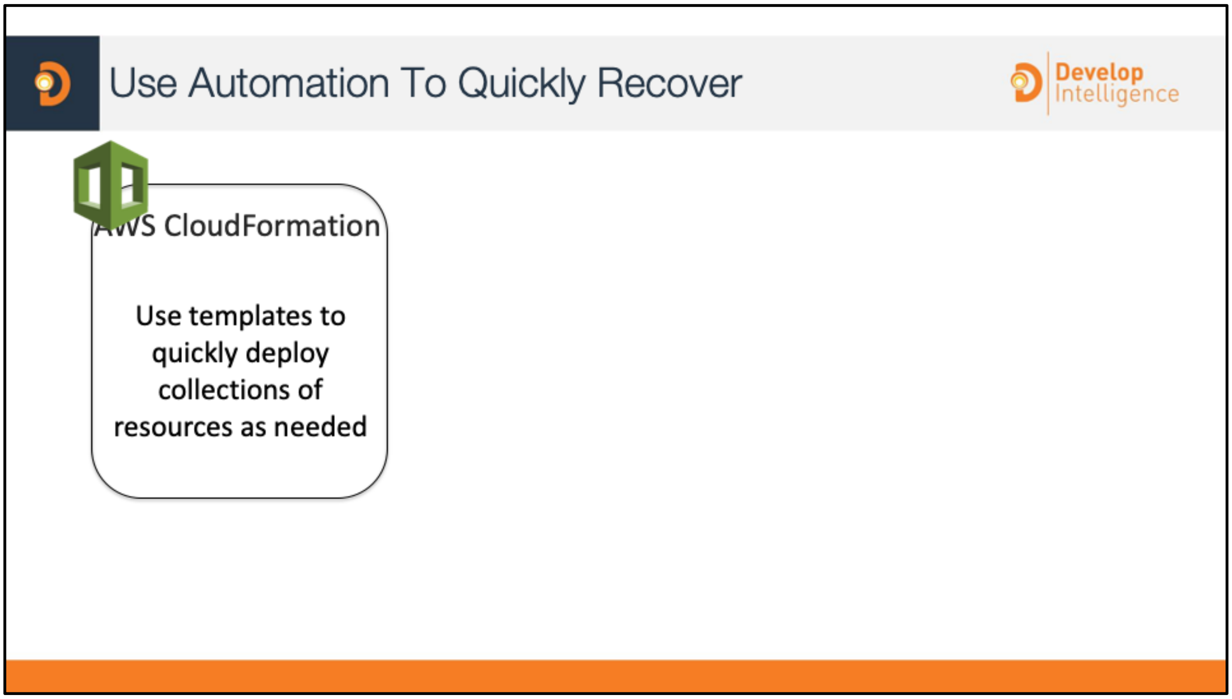
- Snapshot data and save it in a separate region.
- Combine Read Replicas with Multi-AZ to build a resilient disaster recovery strategy.
- Retain automated backups

Amazon DynamoDB

- Back up full tables in seconds.
- Use point-in-time-recovery to continuously back up tables for up to 35 days.
- Initiate backups with a single click in the console or a single API call.
- Build multi-region, multi-master tables for fast local performance for globally distributed apps with Global tables.

You can use **Amazon DynamoDB** in the preparation phase to copy data to DynamoDB in another region or to Amazon S3. During the recovery phase of DR, you can scale up seamlessly in a matter of minutes with a single click or API call.
Global Tables builds on the DynamoDB global footprint to provide you with a fully managed, multi-region, and multi-master database that provides fast, local, read and write performance for massively scaled, global applications. Global Tables replicates your Amazon DynamoDB tables automatically across your choice of AWS regions.

Global Tables eliminates the difficult work of replicating data between regions and resolving update conflicts, enabling you to focus on your application's business logic. In addition, Global Tables enables your applications to stay highly available even in the unlikely event of isolation or degradation of an entire region.



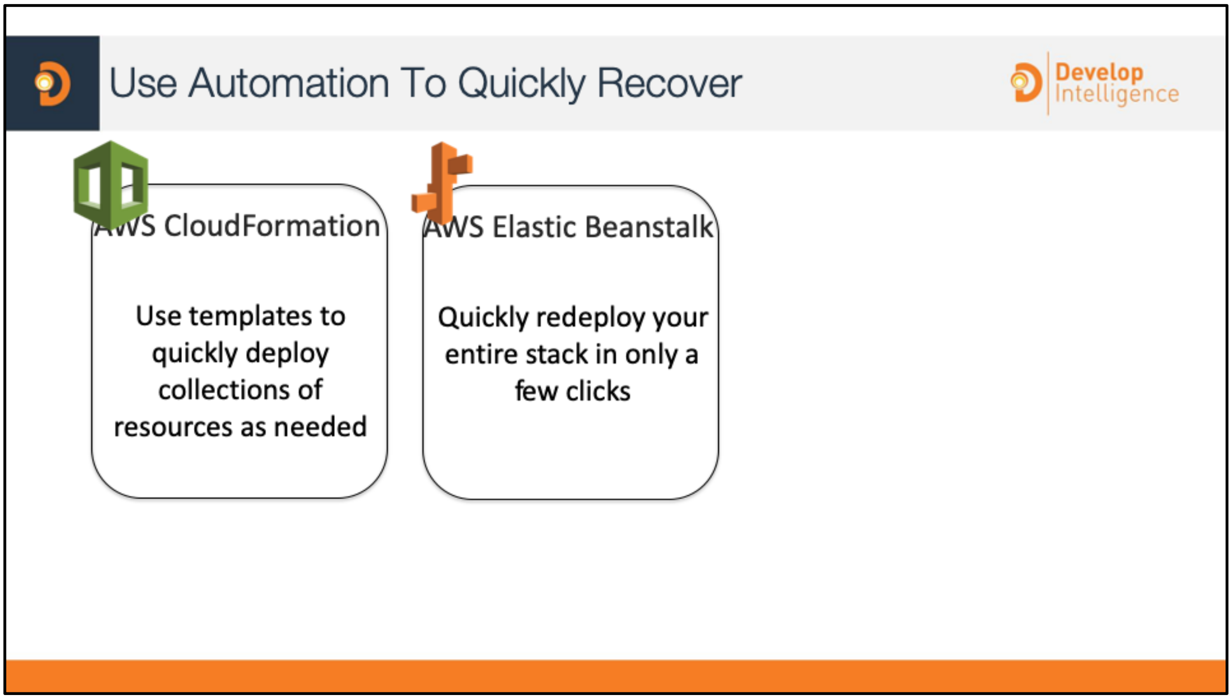
Use Automation To Quickly Recover

AWS CloudFormation

Use templates to quickly deploy collections of resources as needed

AWS CloudFormation allows you to model your entire infrastructure in a text file. This template becomes the single source of truth for your infrastructure. This helps you to standardize infrastructure components used across your organization, enabling configuration compliance and faster troubleshooting.

AWS CloudFormation provisions your resources in a safe, repeatable manner, allowing you to build and rebuild your infrastructure and applications, without having to perform manual actions or write custom scripts. CloudFormation takes care of determining the right operations to perform when managing your stack, and rolls back changes automatically if errors are detected.



Use Automation To Quickly Recover

Develop Intelligence

AWS CloudFormation

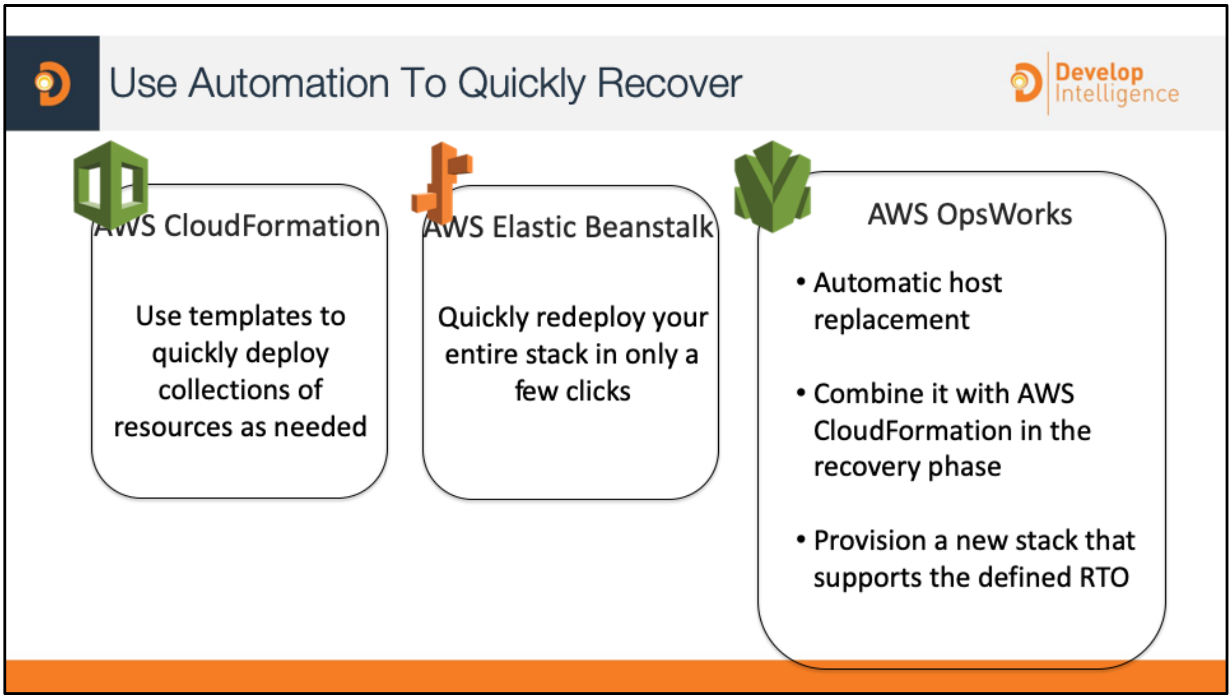
Use templates to quickly deploy collections of resources as needed

AWS Elastic Beanstalk

Quickly redeploy your entire stack in only a few clicks

You can use the **AWS Elastic Beanstalk** to upload an updated source bundle and deploy it to your AWS Elastic Beanstalk environment, or redeploy a previously uploaded version.

You can deploy a previously uploaded version of your application to any of its environments.



Use Automation To Quickly Recover

AWS CloudFormation

Use templates to quickly deploy collections of resources as needed

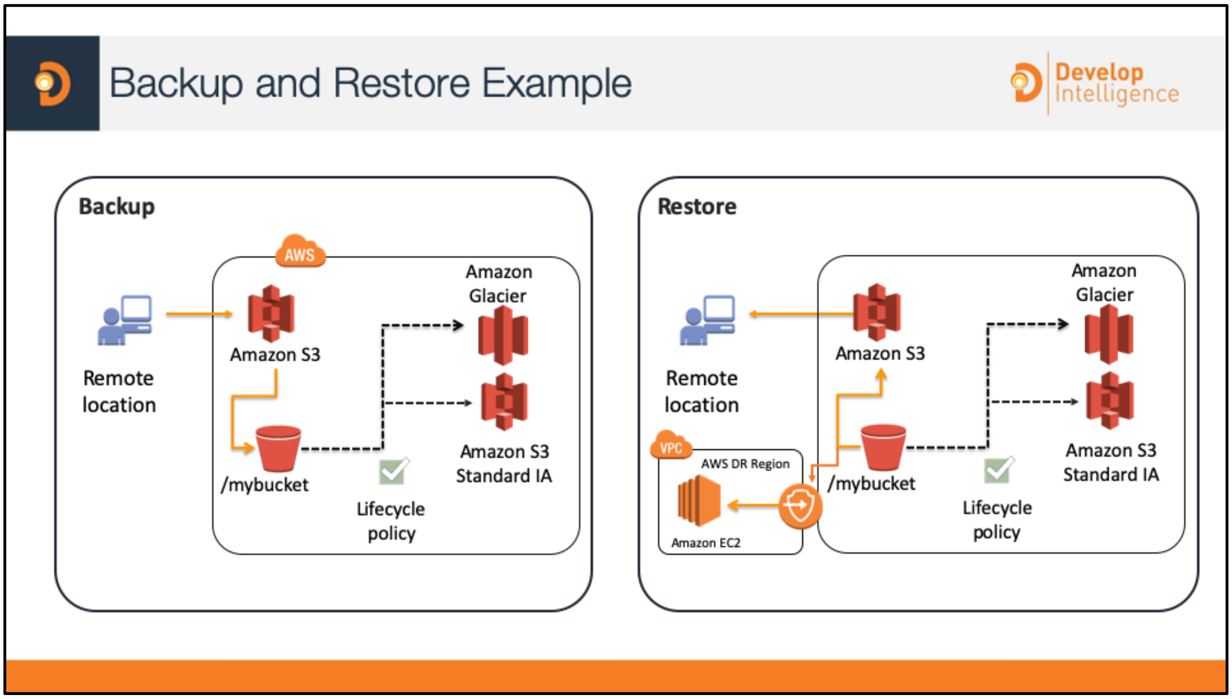
AWS Elastic Beanstalk

Quickly redeploy your entire stack in only a few clicks

AWS OpsWorks

- Automatic host replacement
- Combine it with AWS CloudFormation in the recovery phase
- Provision a new stack that supports the defined RTO

AWS OpsWorks is an application management service that makes it easy to deploy and operate applications of all types and sizes. You can define your environment as a series of layers, and configure each layer as a tier of your application. AWS OpsWorks has automatic host replacement, so in the event of an instance failure it will be automatically replaced. You can use AWS OpsWorks in the preparation phase to template your environment, and you can combine it with AWS CloudFormation in the recovery phase. You can quickly provision a new stack from the stored configuration that supports the defined RTO.

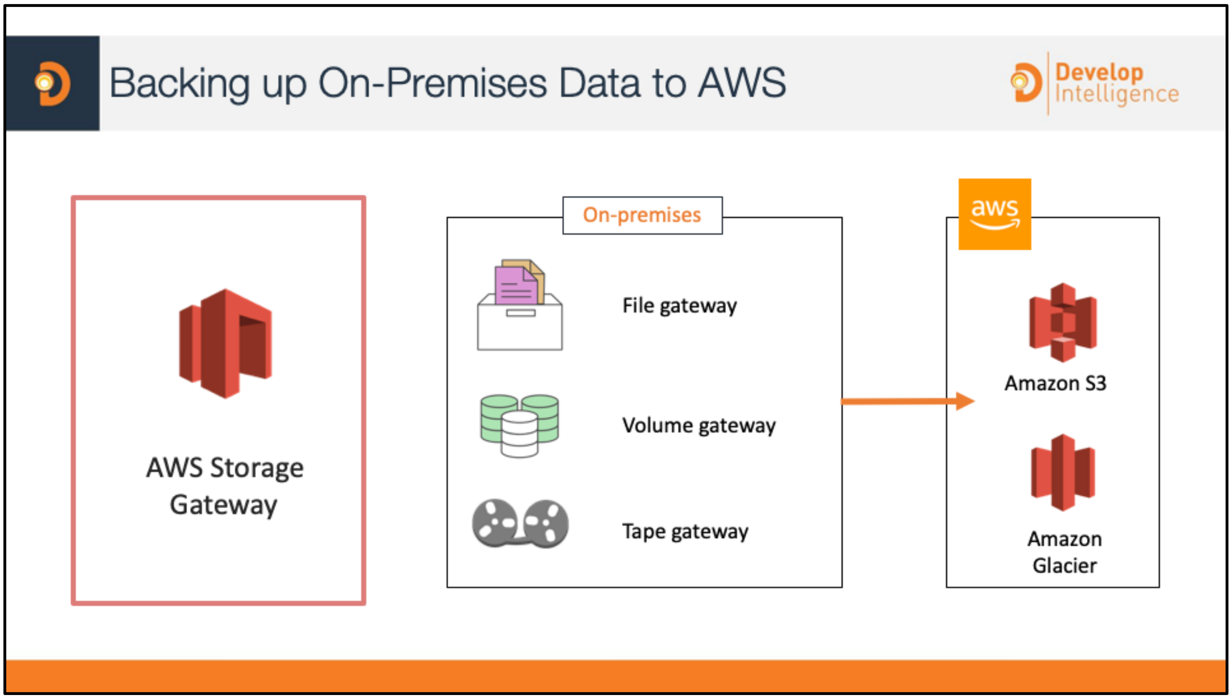


Backup and Restore Example

In most traditional environments, data is backed up to tape and sent offsite regularly. If you use this method, it can take a long time to restore your system in the event of a disruption or disaster.

Amazon S3 is an ideal destination for backup data that might be needed quickly to perform a restore. Transferring data to and from Amazon S3 is typically done through the network and is therefore accessible from any location. There are many commercial and open-source backup solutions that integrate with Amazon S3. For example:

- You can use AWS Snowball to transfer very large data sets by shipping storage devices directly to AWS.
- For longer-term data storage where retrieval times of several hours are adequate, there is Amazon Glacier, which has the same durability model as Amazon S3. Amazon Glacier and Amazon S3 can be used in conjunction to produce a tiered backup solution.



Backing up On-Premises Data to AWS

AWS Storage Gateway

On-premises

File gateway

Volume gateway

Tape gateway

aws

Amazon S3

Amazon Glacier

AWS Storage Gateway connects an on-premises software appliance with cloud-based storage to provide seamless and highly secure integration between your on-premises IT environment and the AWS storage infrastructure. The service allows you to securely store data in the AWS cloud for scalable and cost-effective storage. Storage Gateway supports industry-standard storage protocols that work with your existing applications while securely storing all of your data encrypted in Amazon S3 or Amazon Glacier.

With AWS Storage Gateway, you get an extension of AWS management services locally; the service is also integrated with Amazon CloudWatch, AWS CloudTrail, AWS KMS, AWS IAM, and etc.

AWS Storage Gateway supports three storage interfaces: file, volume, and tape. Each gateway you have can provide one type of interface.

The **file gateway** enables you to store and retrieve objects in Amazon S3 using the NFS and SMB file protocols. Objects written through file gateway can be directly accessed in S3.

The **volume gateway** provides block storage to your applications using the iSCSI protocol. Data on the volumes is stored in Amazon S3. To access your iSCSI volumes in AWS, you can take EBS snapshots which can be used to create EBS volumes.

The **tape gateway** provides your backup application with an iSCSI virtual tape library (VTL) interface, consisting of a virtual media changer, virtual tape drives, and virtual tapes. Virtual tape data is stored in Amazon S3 or can be archived to Amazon Glacier.

To back up your on-premises data to the AWS cloud, you can choose between two common approaches:

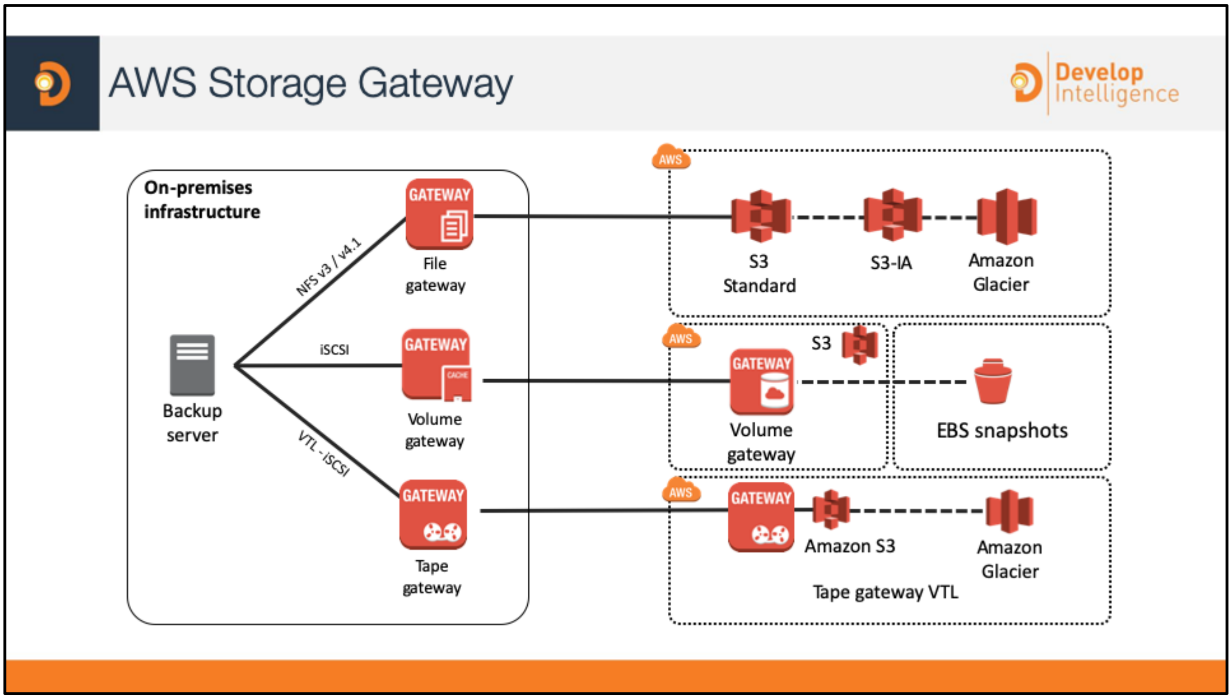
- Writing backup data directly to Amazon S3 by making API calls to the AWS service.
- Writing or retrieving backup data through secure HTTP PUT and GET requests directly across the Internet. Here, the endpoint itself makes a direct connection with Amazon S3 to write data and retrieve data.

Gateway-Virtual Tape Library (VTL)
You can have a limitless collection of virtual tapes. Each virtual tape can be stored in a virtual tape library backed by Amazon S3 or a virtual tape shelf backed by Amazon Glacier.

Gateway-Cached Volumes
You can store your primary data in Amazon S3 and retain your frequently accessed data locally. Gateway-cached volumes provide substantial cost savings on primary storage, minimize the need to scale your storage on-premises, and retain low-latency access to your frequently accessed data.

Gateway-Stored Volumes
If you need low-latency access to your entire data set, you can configure your on-premises data gateway to store your primary data locally and asynchronously back up point-in-time snapshots of this data to Amazon S3.

AWS Storage Gateway Hardware Appliance
The AWS Storage Gateway Hardware Appliance is a hardware appliance that provides AWS Storage Gateway software that is preinstalled on a third-party server that can be installed on-premises. AWS Storage Gateway Hardware Appliance can be managed from the Hardware page on the AWS Management Console.
https://docs.aws.amazon.com/storagegateway/latest/userguide/HardwareAppliance.html



AWS Storage Gateway

On-premises infrastructure

Backup server

NFS v3 / v4.1

iSCSI

VTL - iSCSI

GATEWAY

File gateway

GATEWAY

CACHE

Volume gateway

GATEWAY

Tape gateway

AWS

S3 Standard

S3-IA

Amazon Glacier

AWS

S3

GATEWAY

Volume gateway

EBS snapshots

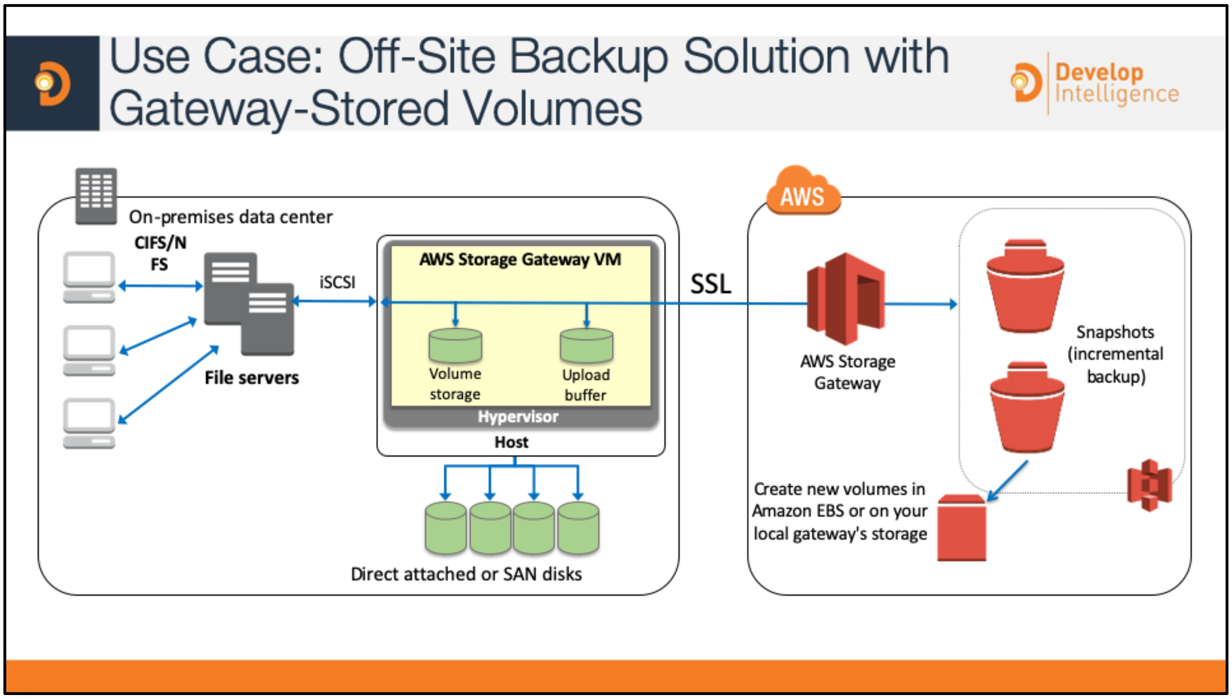
AWS

GATEWAY

Amazon S3

Amazon Glacier

Tape gateway VTL



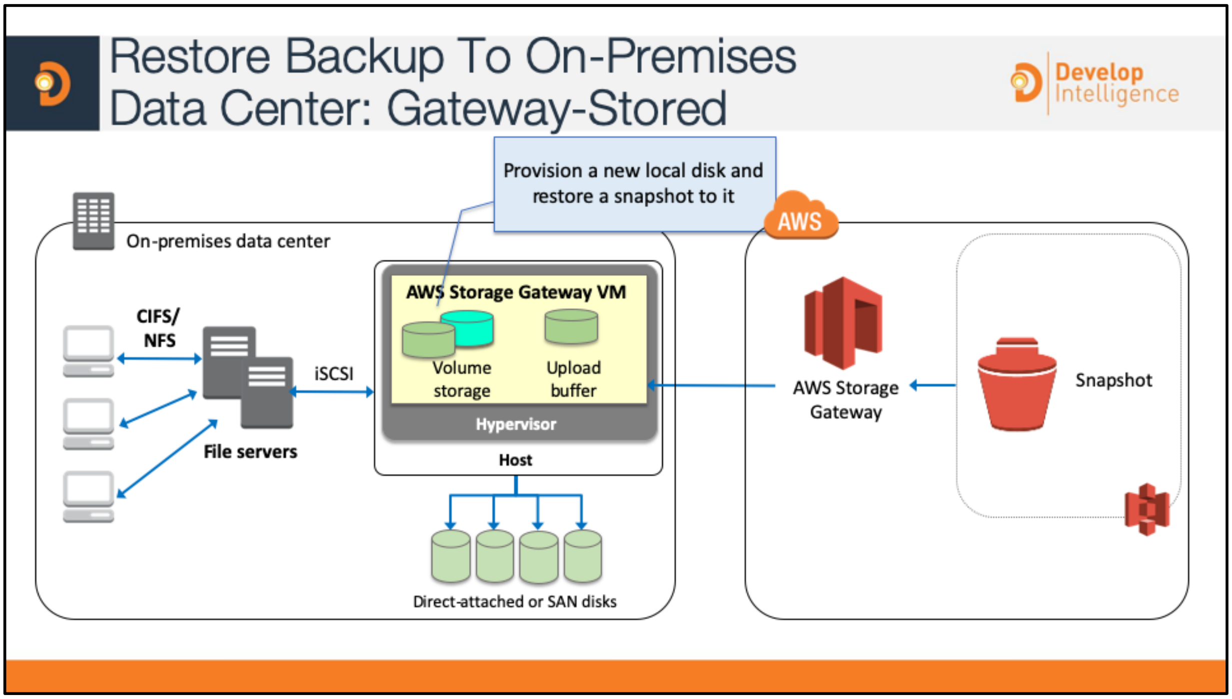
Use Case: Off-Site Backup Solution with Gateway-Stored Volumes

After you've installed the AWS Storage Gateway software appliance—the virtual machine (VM)—on a host in your data center and activated it, you can create gateway *storage volumes* and map them to on-premises direct-attached storage (DAS) or storage area network (SAN) disks. You can start with either new disks or disks already holding data. You can then mount these storage volumes to your on-premises application servers as iSCSI devices. As your on-premises applications write data to and read data from a gateway's storage volume, this data is stored and retrieved from the volume's assigned disk.

To prepare data for upload to Amazon S3, your gateway also stores incoming data in a staging area, referred to as an *upload buffer*. You can use on-premises DAS or SAN disks for working storage. Your gateway uploads data from the upload buffer over an encrypted Secure Sockets Layer (SSL) connection to the AWS Storage Gateway service running in the AWS cloud. The service then stores the data encrypted in Amazon S3.

You can take incremental backups, called *snapshots*, of your storage volumes. The gateway stores these snapshots in Amazon S3 as Amazon EBS snapshots. When you take a new snapshot, only the data that has changed since your last snapshot is stored. You can initiate snapshots on a scheduled or one-time basis. When you delete a snapshot, only the data not needed for any other snapshot is removed.

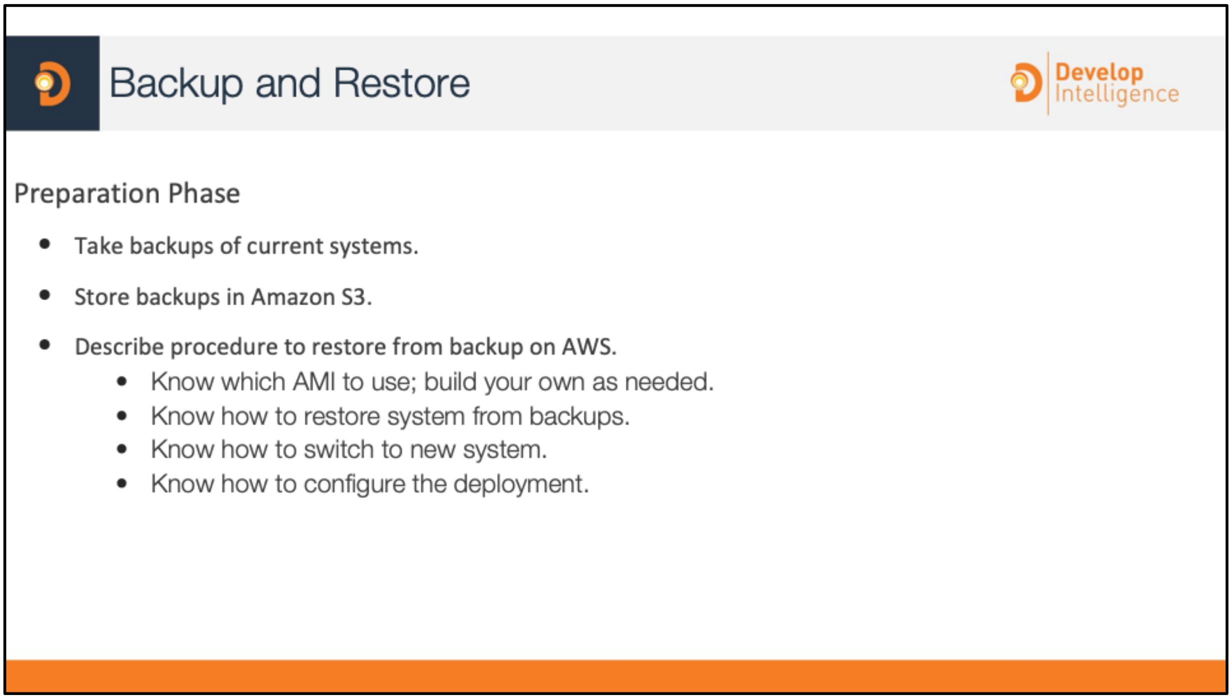
You can restore an Amazon EBS snapshot to an on-premises gateway storage volume if you need to recover a backup of your data. You can also use the snapshot as a starting point for a new Amazon EBS volume, which you can then attach to an Amazon Elastic Compute Cloud (Amazon EC2) instance.



Restore Backup To On-Premises Data Center: Gateway-Stored

For gateway-stored volumes, your volume data is stored on-premises. In this case, snapshots provide durable, off-site backups in Amazon S3. For example, if a local disk allocated as a storage volume crashes, you can provision a new local disk and restore a snapshot to it during the volume creation process. (For more information on this approach, see Adding a Storage Volume at http://docs.aws.amazon.com/storagegateway/latest/userguide/ApplicationStorageVolumesStored-Adding.html).

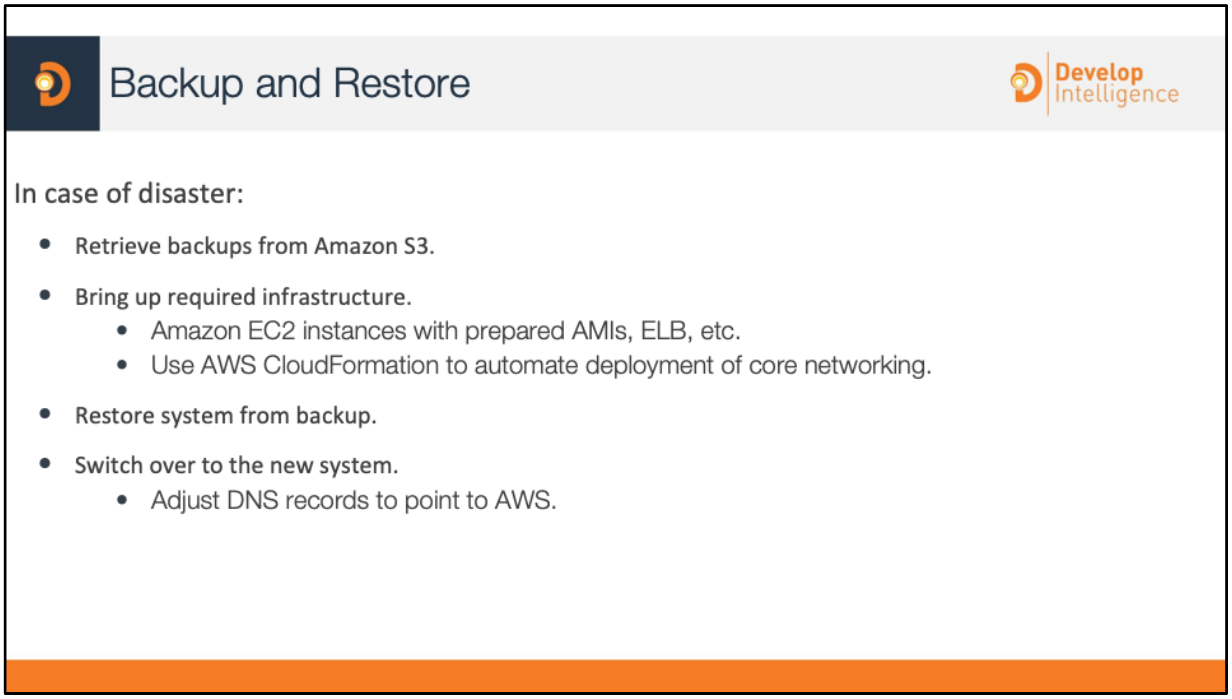
After you initiate a snapshot restore to a gateway-stored volume, snapshot data is downloaded in the background. This functionality means that after you create a volume from a snapshot, there is no need to wait for all of the data to transfer from Amazon S3 to your volume before your application can start accessing the volume and all of its data. If your application accesses a piece of data that has not yet been loaded, the gateway immediately downloads the requested data from Amazon S3. The gateway then continues loading the rest of the volume's data in the background.



Backup and Restore

Preparation Phase

- Take backups of current systems.
- Store backups in Amazon S3.
- Describe procedure to restore from backup on AWS.
  - Know which AMI to use; build your own as needed.
  - Know how to restore system from backups.
  - Know how to switch to new system.
  - Know how to configure the deployment.

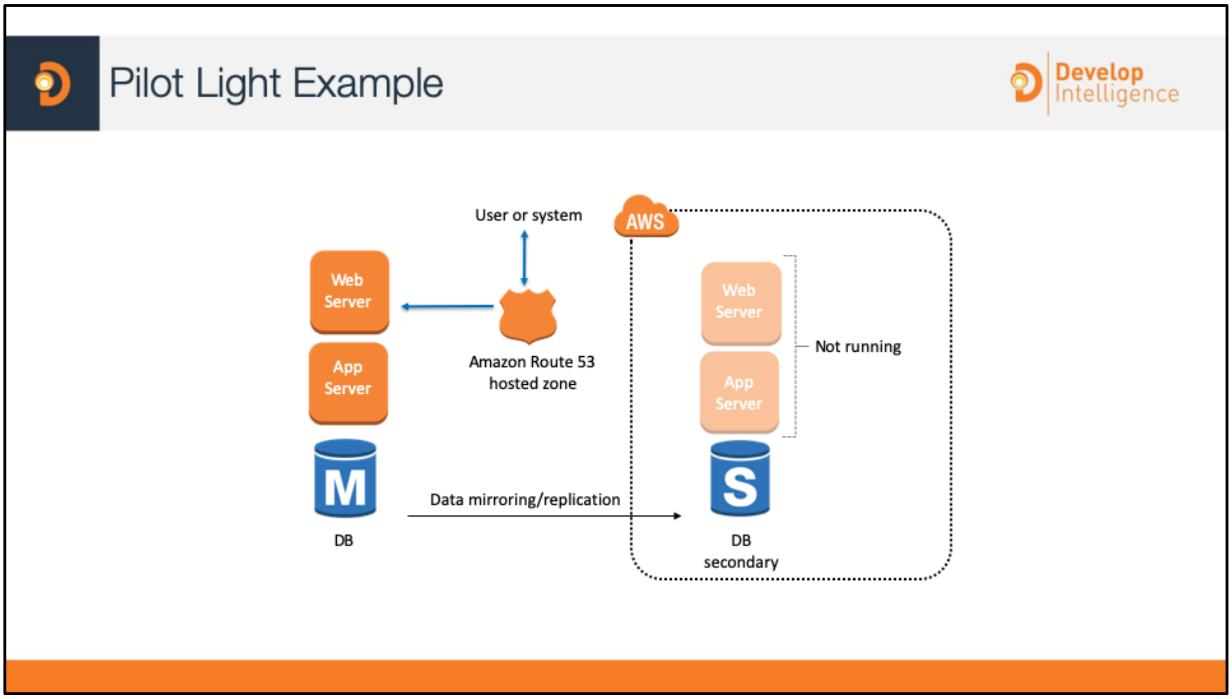


Backup and Restore

Develop Intelligence

In case of disaster:

- Retrieve backups from Amazon S3.
- Bring up required infrastructure.
  - Amazon EC2 instances with prepared AMIs, ELB, etc.
  - Use AWS CloudFormation to automate deployment of core networking.
- Restore system from backup.
- Switch over to the new system.
  - Adjust DNS records to point to AWS.

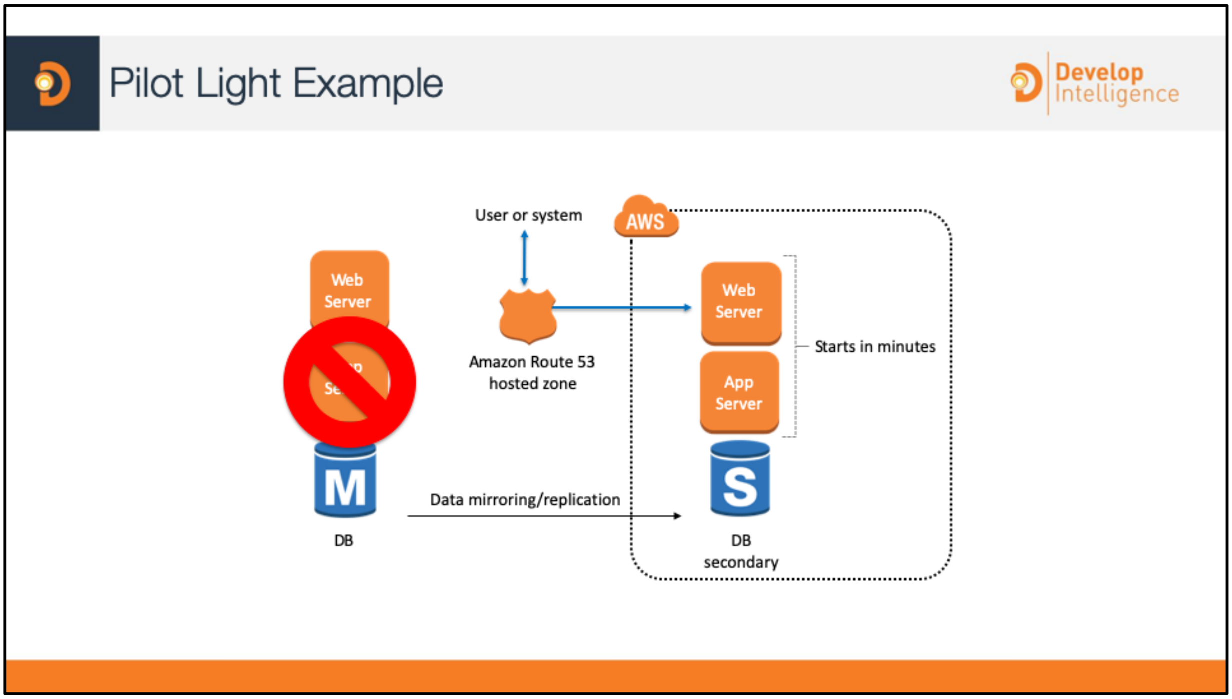


Pilot Light Example

This pattern is relatively inexpensive to implement. In the preparation phase of DR, it is important to consider the use of services and features that support data migration and durable storage, because they enable you to restore backed-up, critical data to AWS when disaster strikes. For some of the scenarios that involve either a scaled-down or a fully scaled deployment of your system in AWS, compute resources will be required as well.

When reacting to a disaster, it is essential to either quickly commission compute resources to run your system in AWS or to orchestrate the failover to already running resources in AWS. The essential infrastructure pieces include DNS, networking features, and various Amazon EC2 features.

In the preparation phase, in which you need to have your regularly changing data replicated to the pilot light, the small core around which the full environment will be started in the recovery phase. Your less frequently updated data, such as operating systems and applications, can be periodically updated and stored as AMIs.



Pilot Light Example

User or system

AWS

Web Server

Amazon Route 53 hosted zone

Web Server

App Server

Starts in minutes

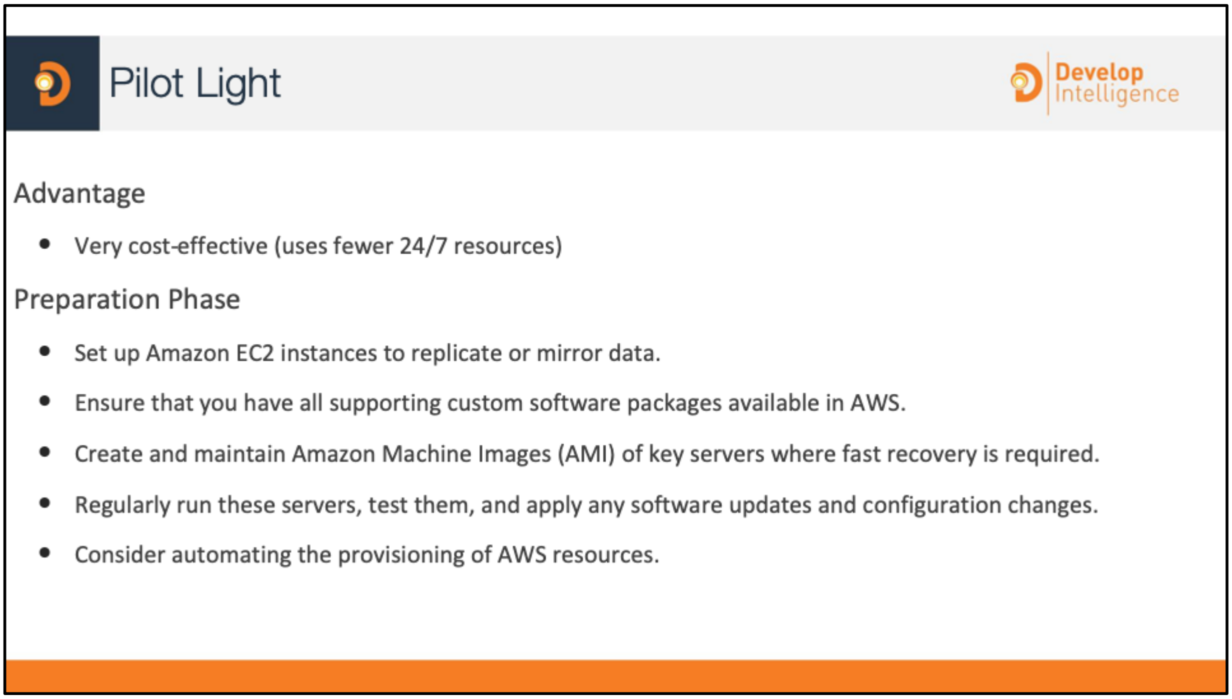
Data mirroring/replication

M

DB

S

DB secondary



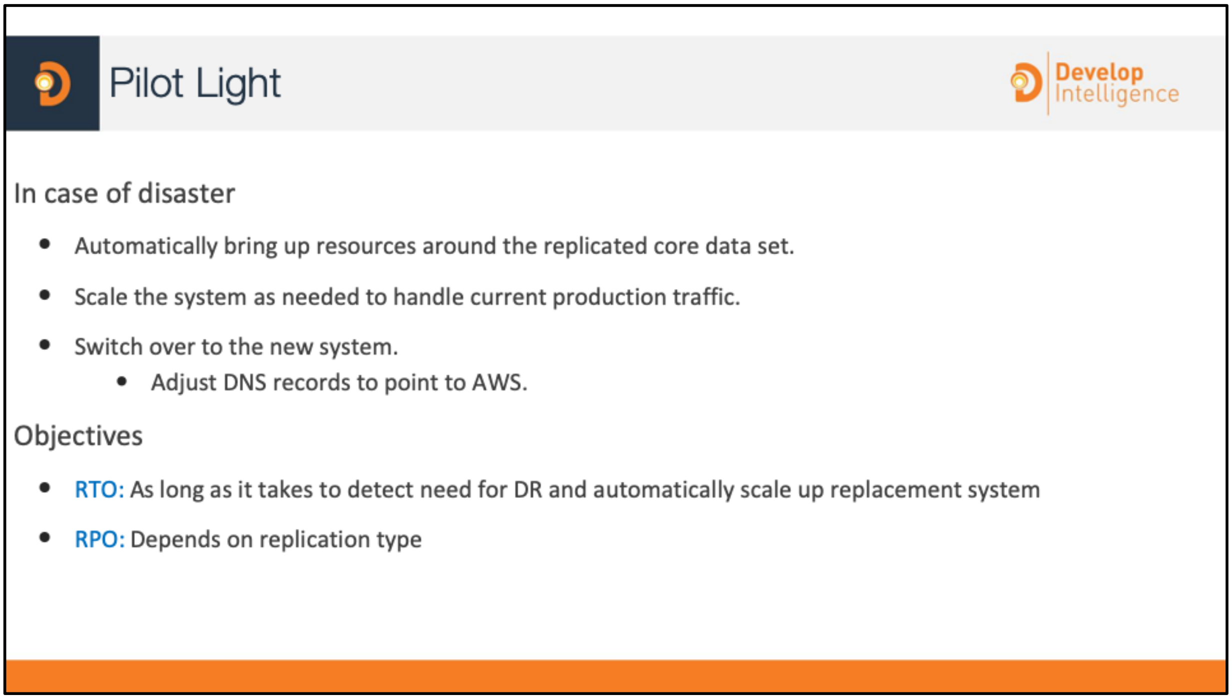
Pilot Light

Advantage

- Very cost-effective (uses fewer 24/7 resources)

Preparation Phase

- Set up Amazon EC2 instances to replicate or mirror data.
- Ensure that you have all supporting custom software packages available in AWS.
- Create and maintain Amazon Machine Images (AMI) of key servers where fast recovery is required.
- Regularly run these servers, test them, and apply any software updates and configuration changes.
- Consider automating the provisioning of AWS resources.



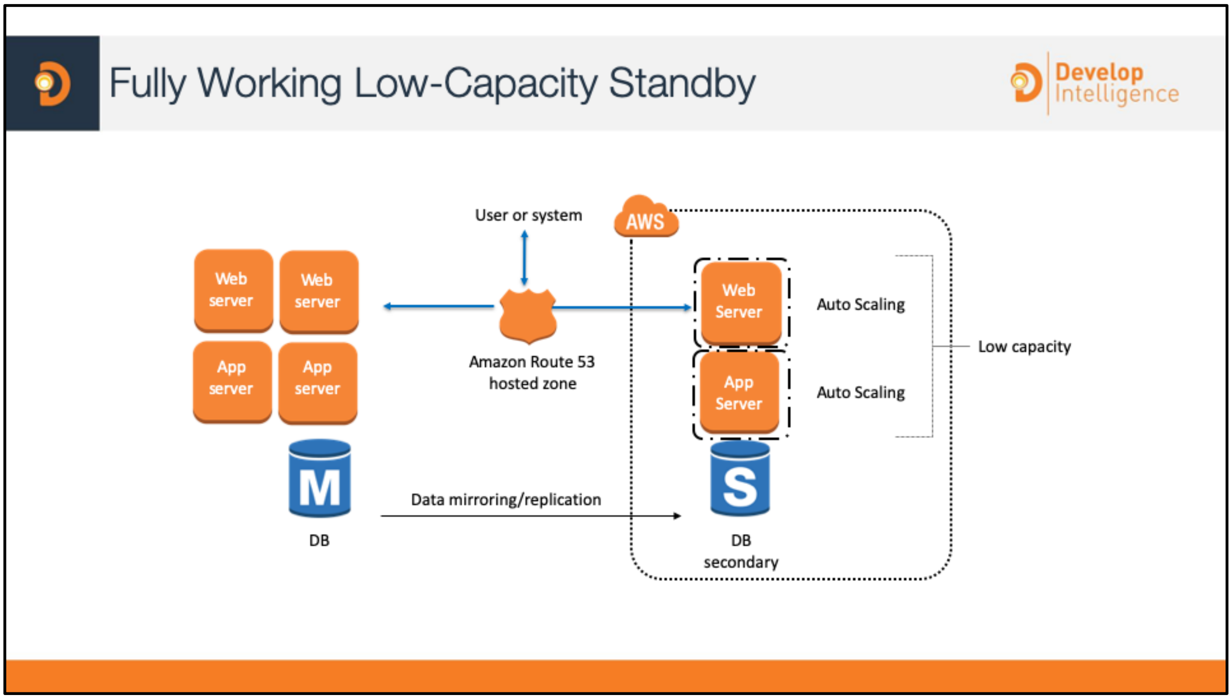
Pilot Light

In case of disaster

- Automatically bring up resources around the replicated core data set.
- Scale the system as needed to handle current production traffic.
- Switch over to the new system.
    - Adjust DNS records to point to AWS.

Objectives

- RTO: As long as it takes to detect need for DR and automatically scale up replacement system
- RPO: Depends on replication type



Fully Working Low-Capacity Standby

Develop Intelligence

User or system

AWS

Web server

Web server

App server

App server

M

DB

Amazon Route 53 hosted zone

Web Server

App Server

Auto Scaling

Auto Scaling

Low capacity

Data mirroring/replication

S

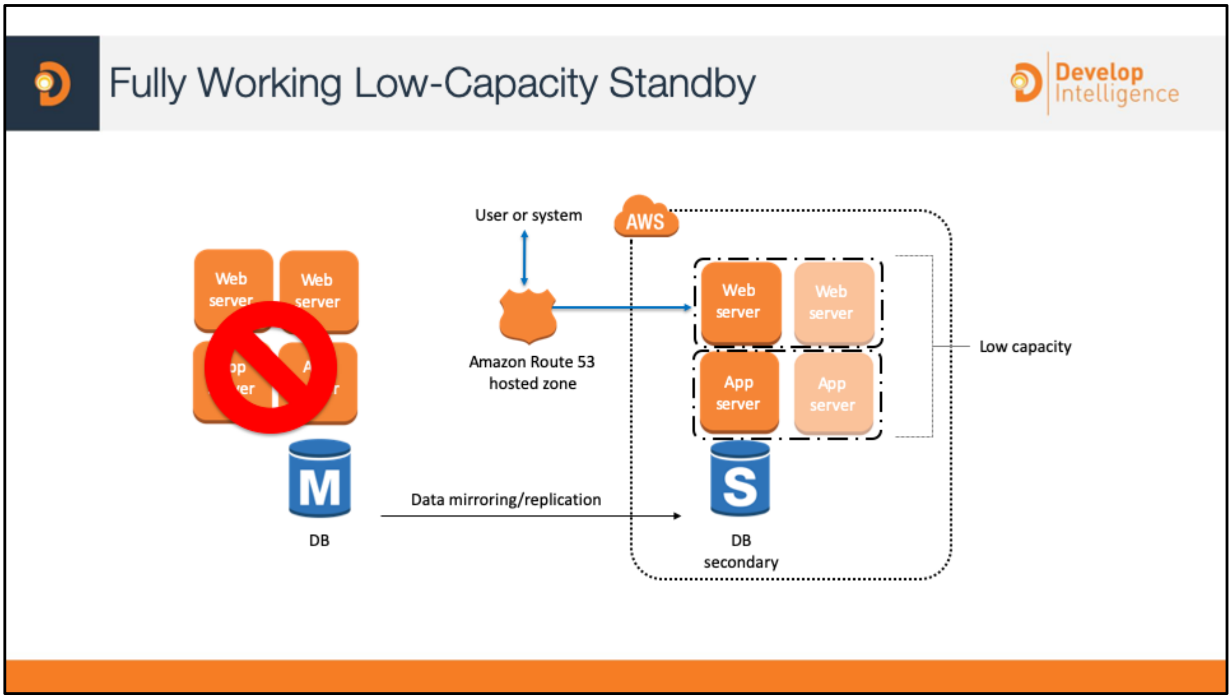
DB secondary

Low capacity standby is like the next level of Pilot Light. The term *warm standby* is used to describe a DR scenario in which a scaled-down version of a fully functional environment is always running in the cloud. A warm standby solution extends the pilot light elements and preparation. It further decreases the recovery time because some services are always running. By identifying your business-critical systems, you can fully duplicate these systems on AWS and have them always on.

These servers can be running on a minimum-sized fleet of Amazon EC2 instances on the smallest sizes possible. This solution is not scaled to take a full production load, but it is fully functional. It can be used for non-production work, such as testing, quality assurance, and internal use.

In a disaster, the system is scaled up quickly to handle the production load. In AWS, this can be done by adding more instances to the load balancer and by resizing the small capacity servers to run on larger Amazon EC2 instance types. As stated in the preceding section, horizontal scaling is preferred over vertical scaling.

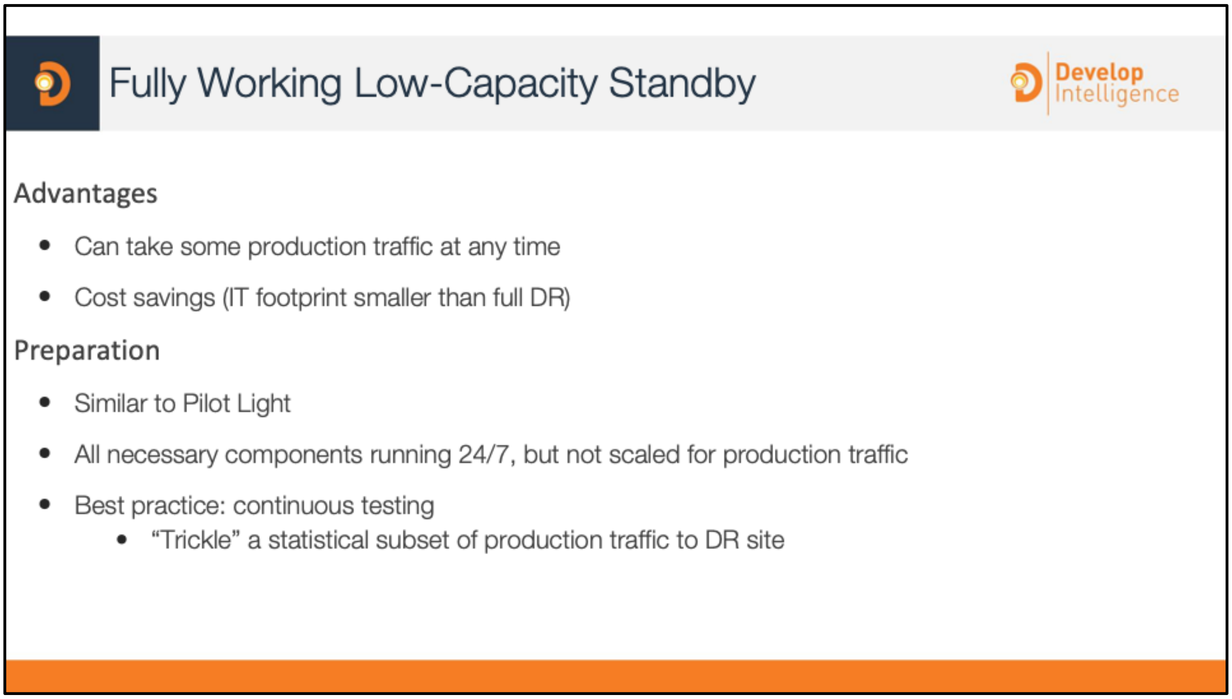
In the diagram above there are two systems running: the main system and a low-capacity system running on AWS. Use Amazon Route 53 to distribute requests between the main system and the cloud system.



Fully Working Low-Capacity Standby

Develop Intelligence

If the primary environment is unavailable, Amazon Route 53 switches over to the secondary system, which is designed to automatically scale its capacity up in the event of a failover from the primary system.



Fully Working Low-Capacity Standby

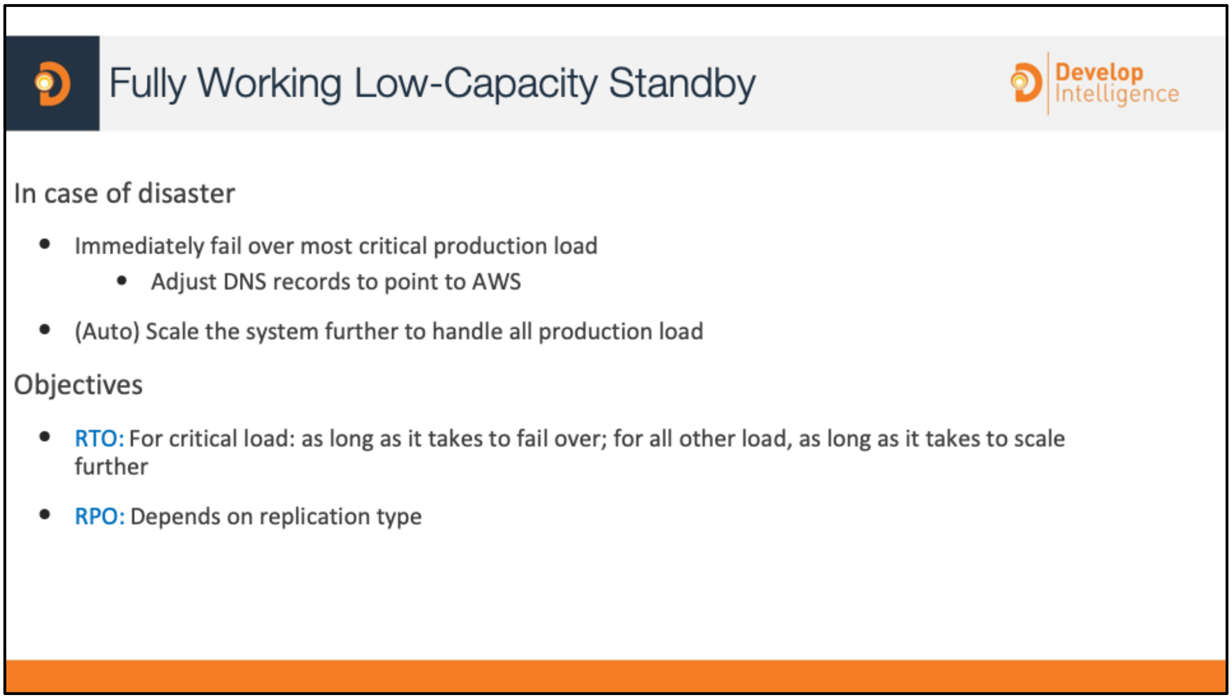
Advantages

- Can take some production traffic at any time
- Cost savings (IT footprint smaller than full DR)

Preparation

- Similar to Pilot Light
- All necessary components running 24/7, but not scaled for production traffic
- Best practice: continuous testing
    - “Trickle” a statistical subset of production traffic to DR site

This pattern is more expensive because active systems are running.



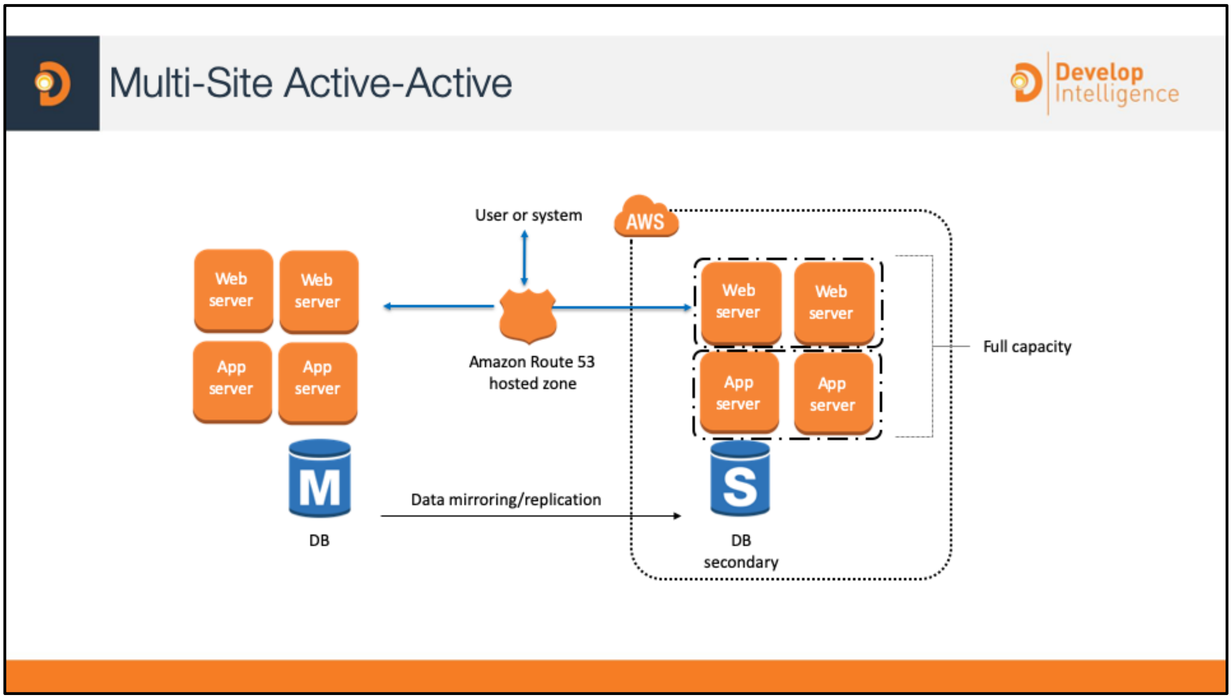
Fully Working Low-Capacity Standby

In case of disaster

- Immediately fail over most critical production load
  - Adjust DNS records to point to AWS
- (Auto) Scale the system further to handle all production load

Objectives

- RTO: For critical load: as long as it takes to fail over; for all other load, as long as it takes to scale further
- RPO: Depends on replication type



Multi-Site Active-Active

Develop Intelligence

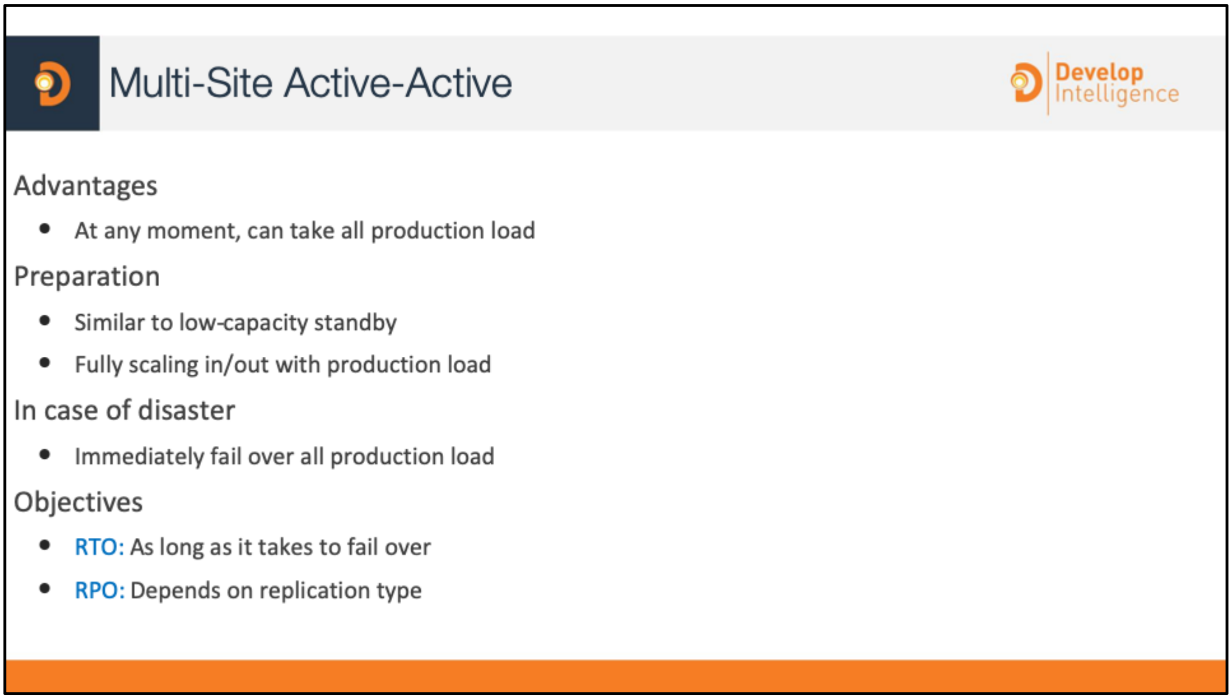
The next level of disaster recovery is to have a fully functional system running in AWS at the same time as the on-premises systems.

A multi-site solution runs in AWS as well as on your existing on-site infrastructure, in an active-active configuration. The data replication method that you employ will be determined by the recovery point that you choose.

You can use a DNS service that supports weighted routing, such as Amazon Route 53, to route production traffic to different sites that deliver the same application or service. A proportion of traffic will go to your infrastructure in AWS, and the remainder will go to your on-site infrastructure.

In an on-site disaster situation, you can adjust the DNS weighting and send all traffic to the AWS servers. The capacity of the AWS service can be rapidly increased to handle the full production load. You can use Amazon EC2 Auto Scaling to automate this process. You might need some application logic to detect the failure of the primary database services and cut over to the parallel database services running in AWS.

The cost of this scenario is determined by how much production traffic is handled by AWS during normal operation. In the recovery phase, you pay only for what you use for the duration that the DR environment is required at full scale. You can further reduce cost by purchasing Amazon EC2 Reserved Instances for your “always on” AWS servers.



Multi-Site Active-Active

Advantages

- At any moment, can take all production load

Preparation

- Similar to low-capacity standby
- Fully scaling in/out with production load

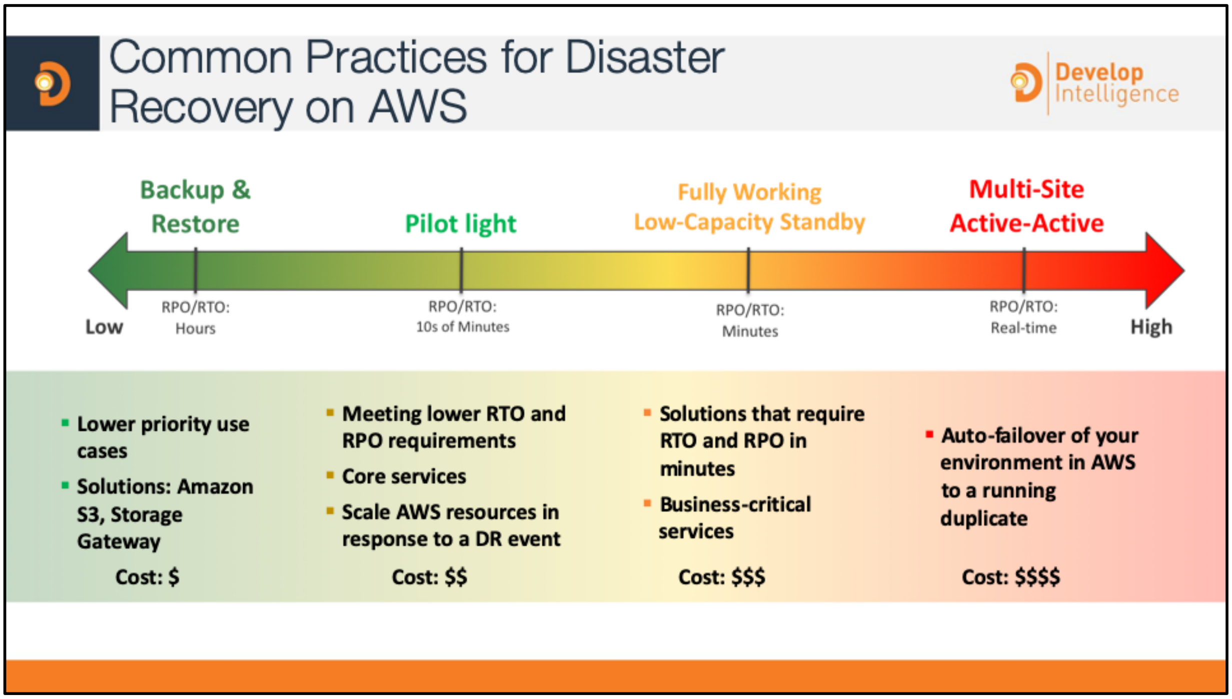
In case of disaster

- Immediately fail over all production load

Objectives

- RTO: As long as it takes to fail over
- RPO: Depends on replication type

This pattern potentially has the least downtime of all. It does have more costs associated with it, because more systems are running.



Common Practices for Disaster Recovery on AWS

Develop Intelligence

Backup & Restore — RPO/RTO: Hours

Pilot light — RPO/RTO: 10s of Minutes

Fully Working Low-Capacity Standby — RPO/RTO: Minutes

Multi-Site Active-Active — RPO/RTO: Real-time

Low

High

- Lower priority use cases
- Solutions: Amazon S3, Storage Gateway

Cost: $

- Meeting lower RTO and RPO requirements
- Core services
- Scale AWS resources in response to a DR event

Cost: $$

- Solutions that require RTO and RPO in minutes
- Business-critical services

Cost: $$$

- Auto-failover of your environment in AWS to a running duplicate

Cost: $$$$

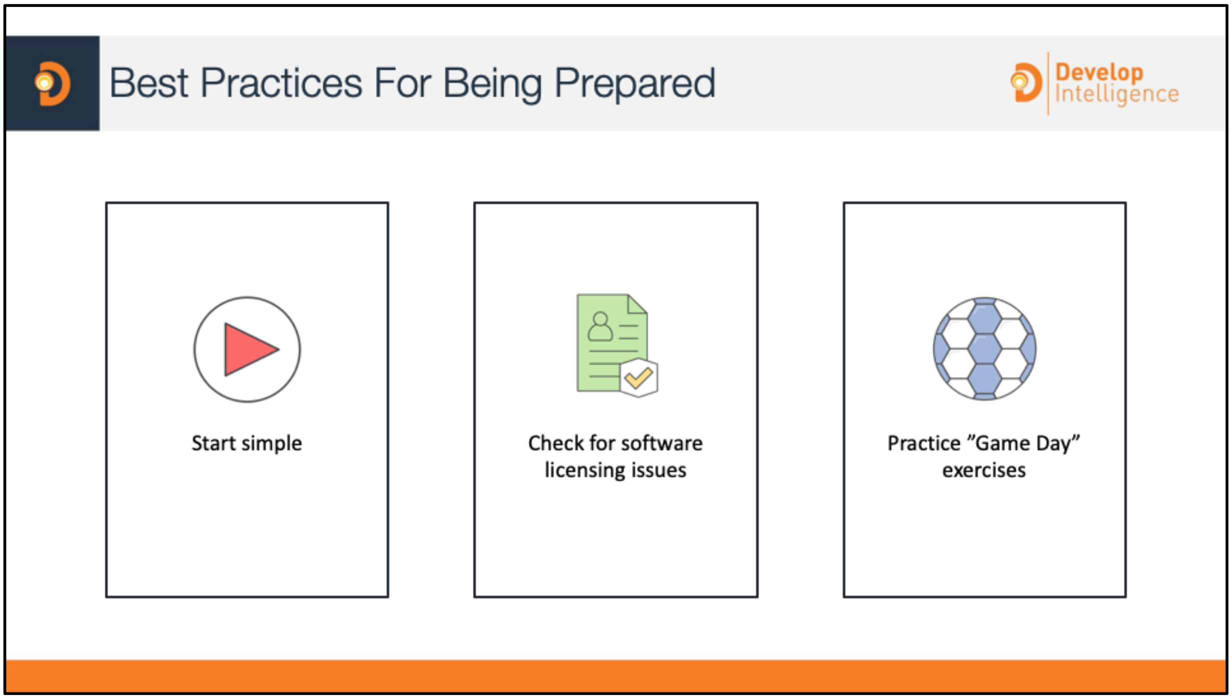
Applications can be placed on a spectrum of complexity. Business continuity ensures that critical business functions continue to operate or recover quickly despite serious disasters.

The next slides outline four DR scenarios that highlight the use of AWS and compare AWS with traditional DR methods (sorted from highest to lowest RTO/RPO), as follows:

- Backup and Restore
- Pilot Light
- Fully Working Low-Capacity Standby
- Multi-Site Active-Active

The figure above shows a spectrum for the four scenarios, arranged by how quickly a system can be available to users after a DR event.

AWS enables you to cost-effectively operate each of these DR strategies. It’s important to note that these are just examples of possible approaches, and variations and combinations of these are possible. If your application is already running on AWS, then multiple regions can be employed and the same DR strategies will still apply.



Best Practices For Being Prepared

Start simple

Check for software licensing issues

Practice ”Game Day” exercises

Start simple and work your way up.

- Backups in AWS are a first step.
- Incrementally improve RTO/RPO as a continuous effort.

Check for any software licensing issues.

Exercise your DR solution

- Practice "Game Day" exercises. These exercises test critical systems going offline or even entire regions. What if an entire fleet were to crash?
- Ensure that backups, snapshots, AMIs, etc. are working.
- Monitor your monitoring system.