What you will learn

At the core of the lesson

You will learn how to:

- Describe the AWS Global Infrastructure and its features
- Identify the difference between AWS Regions, Availability Zones, and Points of Presence

Key terms:
Elastic infrastructure
Scalable infrastructure
Fault-tolerant

2

aws re/start

In this module, you will review the AWS Global Infrastructure and its features. You will also learn how to identify the difference between AWS Regions, Availability Zones, and Points of Presence.

AWS Global Infrastructure

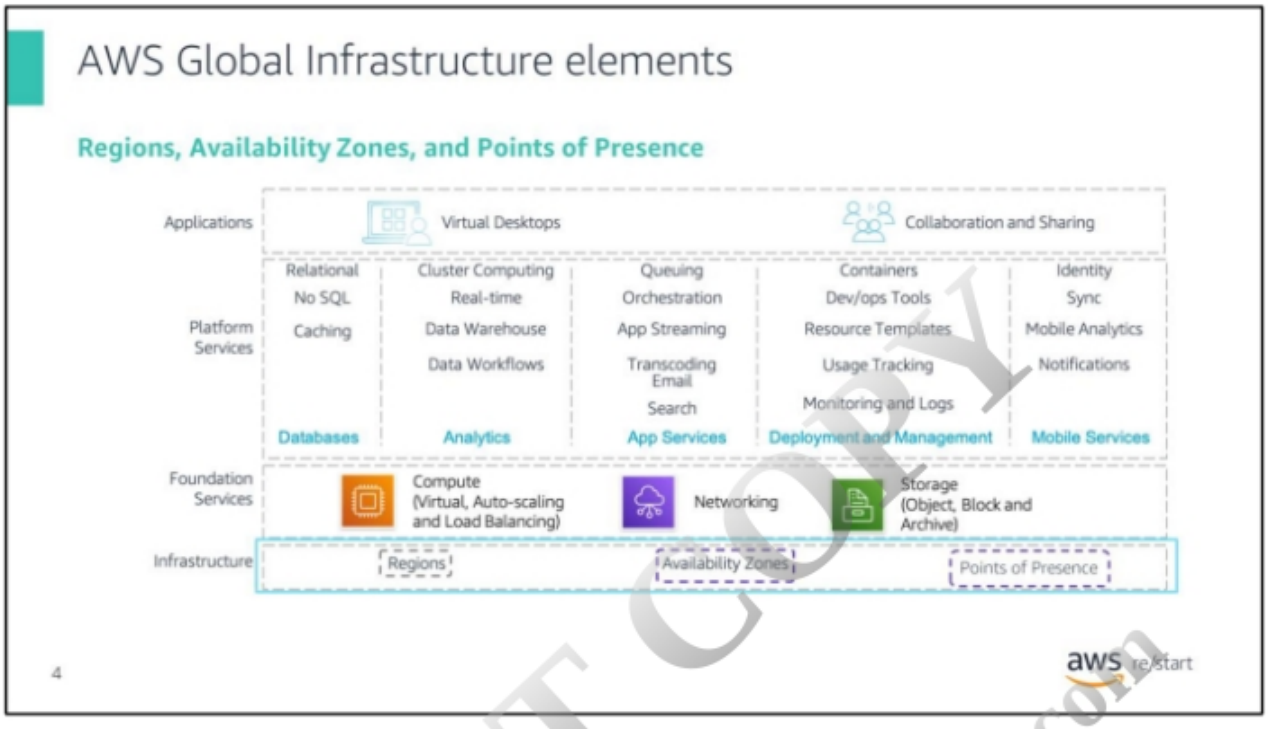
The AWS Global Infrastructure is designed and built to deliver a **flexible**, **reliable**, **scalable**, and **secure** cloud computing environment with high-quality **global network performance**.

3

aws re/start

The diagram shows the 24 current AWS Regions, in addition to a few Regions that will become available soon (as of August 2020).

To learn more about the current AWS Regions, refer to the Global Infrastructure page.

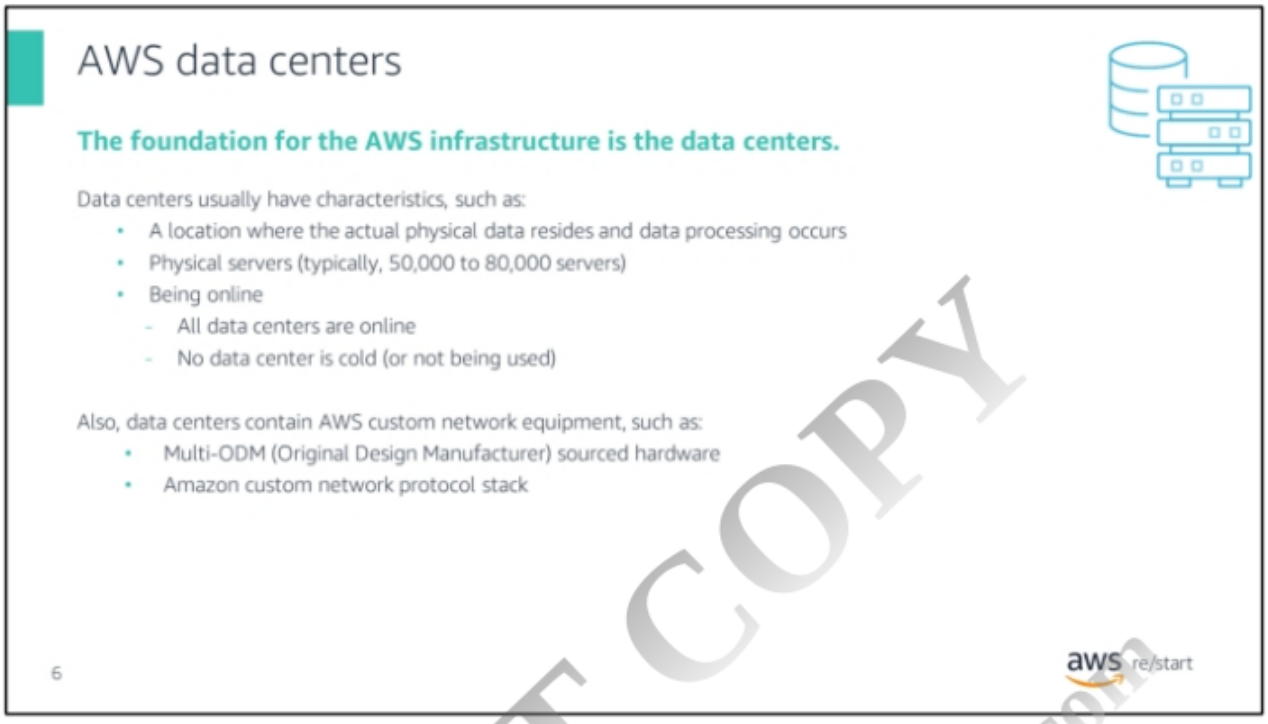


As discussed earlier, AWS provides a broad set of services—such as compute, storage options, networking, and databases—delivered as an on-demand utility that is available in seconds, with pay-as-you-go pricing. All of these services reside on the AWS Global Infrastructure.

The AWS Global Infrastructure can be broken down into three elements: Regions, Availability Zones, and Points of Presence.

Next, you will take an in-depth look at the AWS Global Infrastructure and learn about these elements.

The educator might now choose to conduct a live demonstration of the AWS Global Infrastructure tool. This resource provides an interactive way to learn about the AWS Global Infrastructure. The remaining slides in this section cover many of the same topics and go into greater detail on some topics.



The foundation for the AWS infrastructure is the data centers. A data center is a location where the actual physical data resides and data processing occurs. AWS data centers are built in clusters in various global regions.

Data centers are securely designed with several factors in mind.

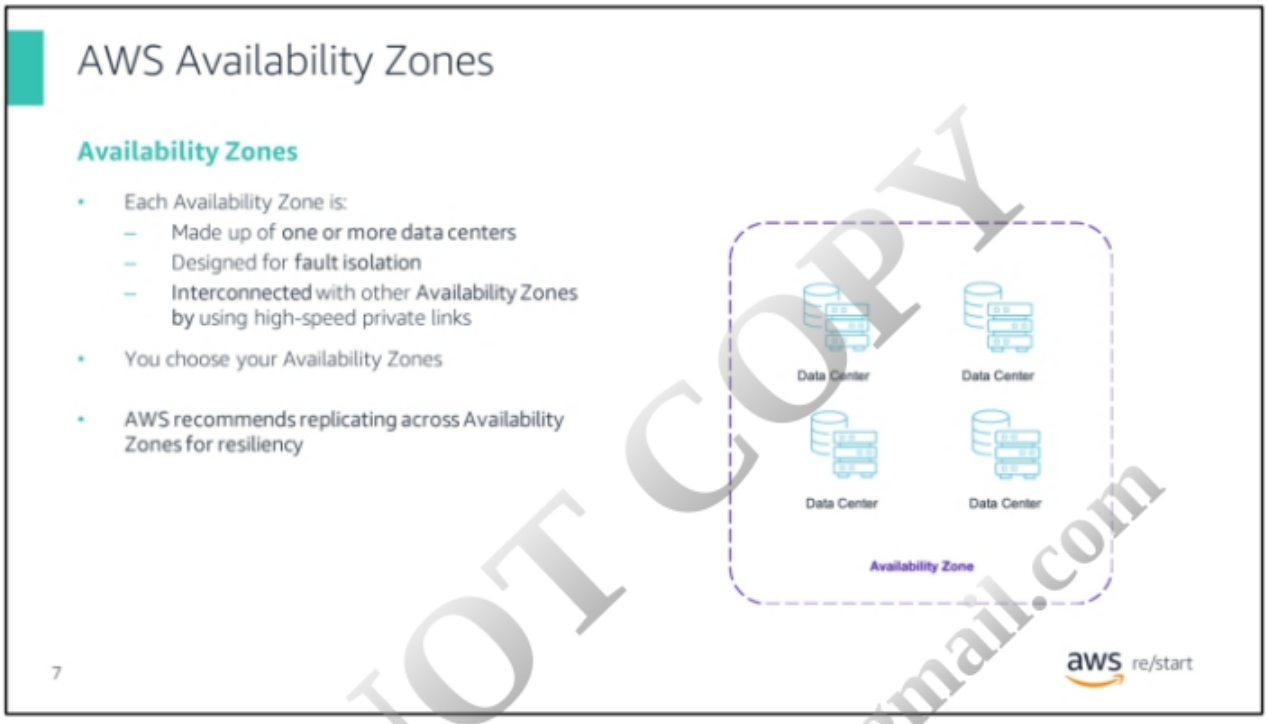
- Each location is carefully evaluated to mitigate environmental risk
- Data centers have a redundant design that anticipates and tolerates failure while maintaining service levels
- To ensure availability, critical system components are backed up across multiple isolated locations that are known as Availability Zones
- To ensure capacity, AWS continuously monitors service usage to deploy infrastructure to support availability commitments and requirements
- Data center locations are not disclosed and all access to them is restricted
- In case of failure, automated processes move customer data traffic away from the affected area

A single data center typically houses 50,000 to 80,000 physical servers.

All data centers are online and serving customers, so no data center is cold.

AWS uses custom, multi-ODM sourced network equipment. An Original Design Manufacturer (or ODM) designs and manufactures products based on specifications from a second company. The second company then rebrands the products for sale.

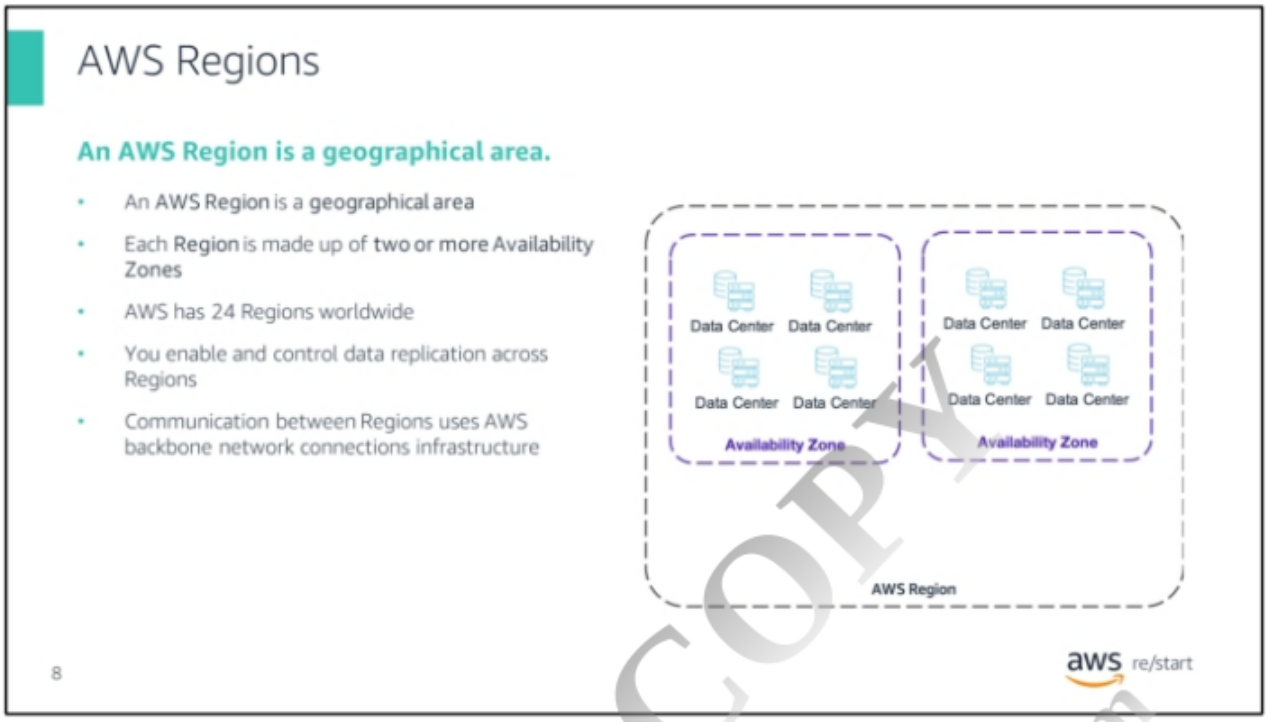
To learn more about AWS data center security, refer to the AWS Data Centers page.



Availability Zones consist of one or more discrete data centers that are designed for fault isolation. They each have redundant power, networking, and connectivity resources that are housed in separate facilities. They are interconnected with other Availability Zones by using high-speed private links. Some Availability Zones have as many as six data centers. However, no data center can be part of two Availability Zones.

Each Availability Zone is designed as an independent failure zone. Availability Zones are physically separated in a typical metropolitan region. They are located in lower-risk flood plains with specific flood-zone categorization that varies by Region. In addition to having a discrete uninterruptible power supply and onsite backup generation facilities, they are each fed via different grids from independent utilities to further reduce single points of failure. Availability Zones are all redundantly connected to multiple tier-1 transit providers. Availability Zones in a Region are connected through low-latency links.

You are responsible for selecting the Availability Zones where your systems will reside. Systems can span across multiple Availability Zones. AWS recommends replicating across Availability Zones for resiliency. You should design your systems to survive temporary or prolonged failure of an Availability Zone if a disaster occurs. Distributing applications across multiple Availability Zones enables them to remain resilient in most failure situations, including natural disasters or system failures.



The AWS Cloud infrastructure is built around Regions and Availability Zones.

An AWS Region is a physical geographical location in the world where AWS has multiple Availability Zones. To achieve fault tolerance and stability, Regions are isolated from each other. Resources in one Region are not automatically replicated to other Regions. Each AWS Region contains two or more Availability Zones. As of August 2020, AWS has 24 Regions worldwide.

When you store data in a specific Region, it's not replicated outside that Region. AWS never moves your data out of the Region that you put it in. It's your responsibility to replicate data across Regions, if your business needs require it. AWS provides information about the country, and—where applicable—the state where each Region resides. You are responsible for selecting the Region to store data in, based on your compliance and network latency requirements.

Consider these additional details. If you are using cloud computing services, you can easily deploy your application in multiple Regions. For instance, you can have an application in a Region that's nearest to your headquarters, such as San Diego on the West Coast of the US. You could then also have a deployable application in a Region in the East Coast of the US. Say that your largest customer base is located in Virginia. With a few clicks, you can deploy in the US

East Region to provide a better experience for your customers who are located there. You will reduce latency and increase agility for your organization within minutes and with minimal cost.

Some Regions have restricted access. For example, the isolated AWS GovCloud (US) Region is designed so that U.S. government agencies and customers can move sensitive workloads into the cloud by addressing their specific regulatory and compliance requirements.