# Reading 1.3: AWS Global Infrastructure

Infrastructure, like data centers and networking connectivity, still exists as the foundation of every cloud application. In AWS, this physical infrastructure makes up the AWS Global Infrastructure, in the form of Availability Zones and Regions.

## Regions



Regions are geographic locations worldwide where AWS hosts its data centers. AWS Regions are named after the location where they reside. For example, in the United States, there is a Region in Northern Virginia called the Northern Virginia Region and a Region in Oregon called the Oregon Region. There are Regions in Asia Pacific, Canada, Europe, the Middle East, and South America, and AWS continues to expand to meet the needs of its customers.

Each AWS Region is associated with a geographical name and a Region code.

**US East (N. Virginia)**  us-east-1

US East (Ohio)  us-east-2

US West (N. California)  us-west-1

US West (Oregon)  us-west-2

Africa (Cape Town)  af-south-1

Asia Pacific (Hong Kong)  ap-east-1

Asia Pacific (Mumbai)  ap-south-1

Asia Pacific (Osaka-Local)  ap-northeast-3

Asia Pacific (Seoul)  ap-northeast-2

Asia Pacific (Singapore)  ap-southeast-1

Asia Pacific (Sydney)  ap-southeast-2

Asia Pacific (Tokyo)  ap-northeast-1

Canada (Central)  ca-central-1

Europe (Frankfurt)  eu-central-1

Europe (Ireland)  eu-west-1

Europe (London)  eu-west-2

Europe (Milan)  eu-south-1

Europe (Paris)  eu-west-3

Europe (Stockholm)  eu-north-1

Middle East (Bahrain)  me-south-1

South America (São Paulo)  sa-east-1

Here are a few examples of Region codes:

- us-east-1: This is the first Region created in the east of the US. The geographical name for this Region is N. Virginia.

- ap-northeast-1: The first Region created in the northeast of Asia Pacific. The geographical name for this Region is Tokyo.

AWS Regions are independent from one another. This means that your data is not replicated from one Region to another, without your explicit consent and authorization.

# Choose the Right AWS Region

Consider four main aspects when deciding which AWS Region to host your applications and workloads: latency, price, service availability, and compliance.

**Latency.** If your application is sensitive to latency, choose a Region that is close to your user base. This helps prevent long wait times for your customers. Synchronous applications such as gaming, telephony, WebSockets, and IoT are significantly affected by higher latency, but even asynchronous workloads, such as ecommerce applications, can suffer from an impact on user connectivity.

**Price.** Due to the local economy and the physical nature of operating data centers, prices may vary from one Region to another. The pricing in a Region can be impacted by internet connectivity, prices of imported pieces of equipment, customs, real estate, and more. Instead of charging a flat rate worldwide, AWS charges based on the financial factors specific to the location.

**Service availability.** Some services may not be available in some Regions. The AWS documentation provides a table containing the Regions and the available services within each one.

**Data compliance.** Enterprise companies often need to comply with regulations that require customer data to be stored in a specific geographic territory. If applicable, you should choose a Region that meets your compliance requirements.

# Availability Zones
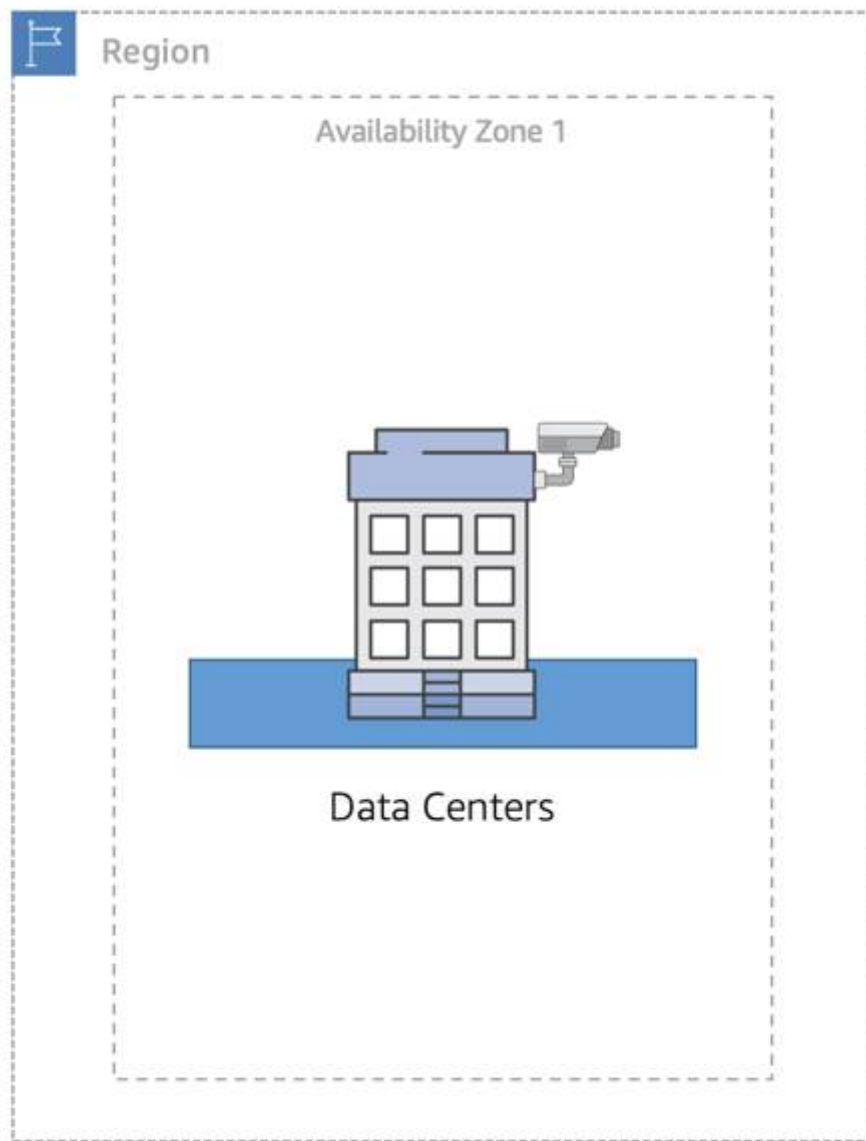
Inside every Region is a cluster of Availability Zones (AZ). An AZ consists of one or more data centers with redundant power, networking, and connectivity. These data centers operate in discrete facilities with undisclosed locations. They are connected using redundant high-speed and low-latency links.

AZs also have a code name. Since they're located inside Regions, they can be addressed by appending a letter to the end of the Region code name. For example:

- us-east-1a: an AZ in us-east-1 (Northern Virginia Region)
- sa-east-1b: an AZ in sa-east-1 (São Paulo Region in South America)

If you see that a resource exists in us-east-1c, you know this resource is located in AZ c of the us-east-1 Region.

## Scope AWS Services

Depending on the AWS Service you use, your resources are either deployed at the AZ, Region, or Global level. Each service is different, so you need to understand how the scope of a service may affect your application architecture.

When you operate a Region-scoped service, you only need to select the Region you want to use. If you are not asked to specify an individual AZ to deploy the service in, this is an indicator that the service operates on a

Region-scope level. For region-scoped services, AWS automatically performs actions to increase data durability and availability.

On the other hand, some services ask you to specify an AZ. With these services, you are often responsible for increasing the data durability and high availability of these resources.

## Maintain Resiliency

To keep your application available, you need to maintain high availability and resiliency. A well-known best practice for cloud architecture is to use Region-scoped, managed services. These services come with availability and resiliency built in.

When that is not possible, make sure the workload is replicated across multiple AZs. At a minimum, you should use two AZs. If one entire AZ fails, your application will have infrastructure up and running in at least a second AZ to take over the traffic.
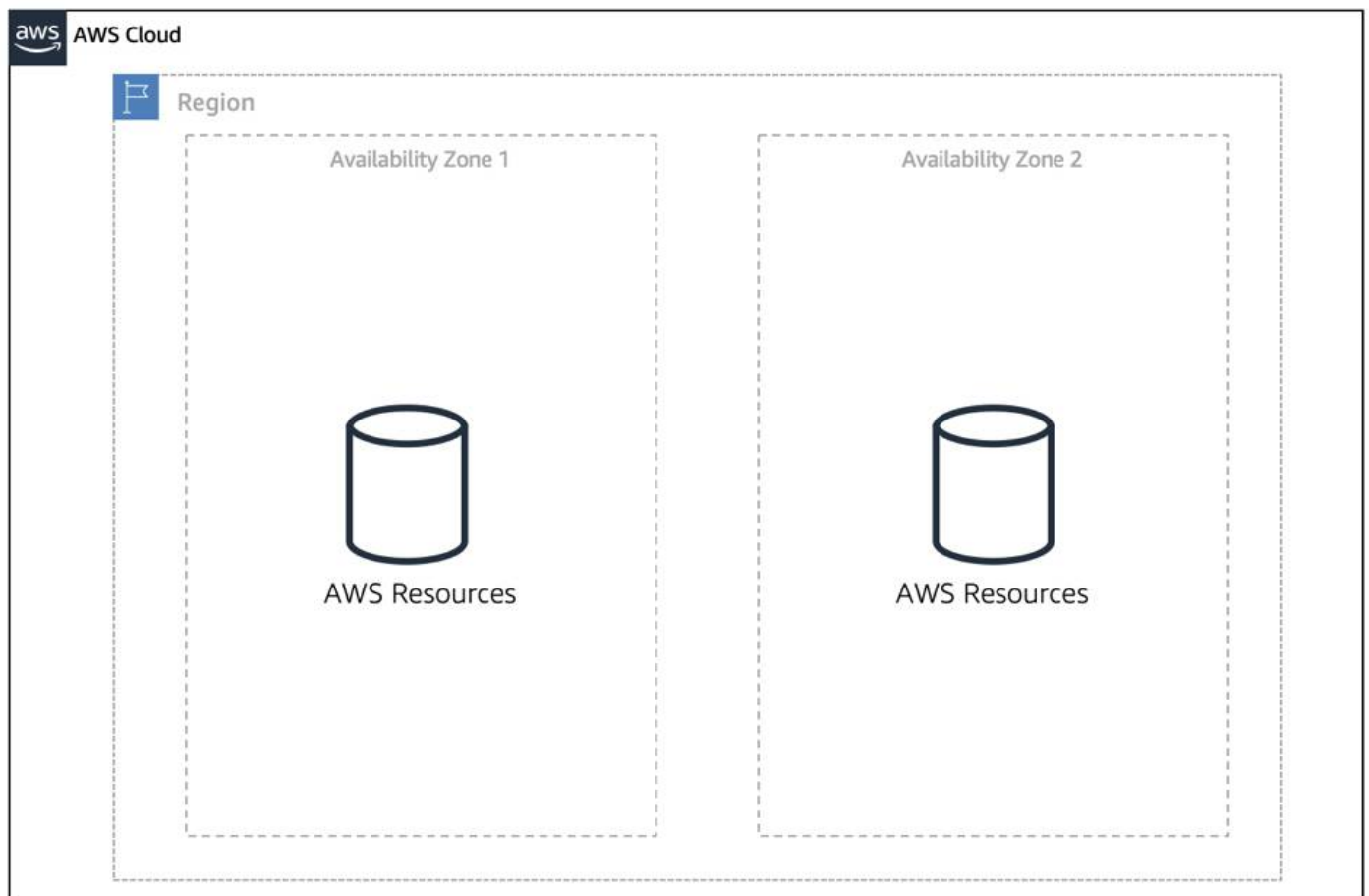


**Resources:**

- *External Site:* AWS: Global Infrastructure
- *External Site:* AWS: AWS Global Infrastructure (dynamic site)
- *External Site:* AWS: AWS Global Infrastructure Documentation
- *External Site:* AWS: AWS Regions and Availability Zones

- *External Site:* AWS: AWS service endpoints
- *External Site:* AWS: AWS Regional Services