MODULE 3 · STUDY NOTES

AWS Global Infrastructure Overview

Regions, Availability Zones, data centers, edge locations — and the service categories built on top of them

The cloud is not a metaphor — it is a planet-wide machine. This module maps that machine from the bottom up: data centers grouped into isolated Availability Zones, Availability Zones grouped into geographic Regions, and a separate mesh of edge locations pushing content close to users. Then it switches lenses from *where AWS runs* to *what AWS offers*: 23 service categories, of which this course works through seven — storage, compute, database, networking and content delivery, security, cost management, and management and governance — and the core services inside each. Both halves feed the certification exam directly: the Region / AZ / edge distinction and the "which service does X?" one-liners are Cloud Practitioner staples.

LEARNING OBJECTIVES

- Identify the difference between AWS Regions, Availability Zones, and edge locations
- Identify AWS service and service categories

3.1 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. It is organized as a strict hierarchy: **Regions** are geographical areas; each Region contains two or more **Availability Zones (AZs)**; each AZ consists of one or more discrete **data centers**. A parallel global network of **edge locations** sits outside that hierarchy, caching content close to end users. AWS had 22 Regions when this course was written and continually updates its footprint — the interactive AWS Global Infrastructure Map (the subject of this module's educator-led demo) always shows the current and planned picture.

TABLE 3.1 The four layers of AWS infrastructure

LAYER	WHAT IT IS	WHO CHOOSES IT
Region	Geographical area with full redundancy and connectivity to the network; typically two or more Availability Zones	You — per workload, weighing four factors
Availability Zone	Fully isolated partition of the AWS infrastructure; one or more discrete data centers (typically three), designed for fault isolation	You — AWS recommends spanning more than one
Data center	Facility where the data physically resides and data processing occurs	Nobody — the AZ is the most granular level a customer can specify
Edge location	Point of Presence that caches content near end users for CloudFront and Route 53	Automatic — requests route to the nearest one

Regions are isolated from one another to achieve fault tolerance and stability. Resources in one Region are *not* automatically replicated to other Regions, and data stored in a specific Region is not replicated outside it — if your business needs cross-Region copies, replicating them is **your responsibility**. When Regions do need to talk, the

communication travels over AWS's own backbone network infrastructure. Two administrative wrinkles are worth knowing: Regions introduced **before March 20, 2019** are enabled by default, while later ones — such as Asia Pacific (Hong Kong) and Middle East (Bahrain) — are **disabled by default** and must be enabled from the AWS Management Console. And some Regions carry restricted access: an Amazon AWS (China) account reaches only the Beijing and Ningxia Regions, and the isolated **AWS GovCloud (US)** Region exists so US government agencies and customers can move sensitive workloads into the cloud while meeting their specific regulatory and compliance requirements.

> **COMMON PITFALL**
>
> Keep the two replication statements straight. *Across Regions*: never automatic — you set it up if you need it. *Across Availability Zones*: what AWS actively recommends for resiliency, and the inter-AZ network is fast enough to do it synchronously.

Selecting a Region comes down to four factors. First, **data governance and legal requirements**: local laws might require that certain information be kept within geographical boundaries — the EU Data Protection Directive is the deck's example — which can restrict the Regions where you may offer content or services. Second, **proximity to customers**: all else being equal, run applications and store data as close as possible to the users and systems that access them, to reduce latency (the CloudPing website tests latency between your location and every AWS Region). Third, **services available within the Region** — not all services are available in all Regions. Fourth, **cost**, which varies by Region even for identical resources.

> **WORKED EXAMPLE – SAME INSTANCE, TWO PRICES**
>
> As of the course's writing, running an On-Demand `t3.medium` Linux EC2 instance costs **$0.0416 per hour** in US East (Ohio) but **$0.0544 per hour** in Asia Pacific (Tokyo) — roughly 31% more for the same virtual machine. Region choice is a pricing decision, not just a technical one.

Availability Zones are each Region's multiple, isolated locations. Each AZ lets you operate applications and databases that are more highly available, fault-tolerant, and scalable than a single data center could ever be. When an application is partitioned across AZs, it is better isolated and protected from issues such as lightning, tornadoes, and earthquakes. You are responsible for selecting the AZs where your systems reside; systems can span multiple AZs, and AWS recommends replicating across them — design your systems to survive the temporary or prolonged failure of an entire Availability Zone.

> **AVAILABILITY ZONE FACTS**
>
> - Fully isolated partition of the AWS Global Infrastructure, with its **own power infrastructure**.
> - One or more data centers per AZ (**typically three**); at full scale, hundreds of thousands of servers.
> - Physically separated by **many kilometers** from other AZs — yet all AZs in a Region are within **100 km** of each other.
> - Interconnected by **high-bandwidth, low-latency networking over fully redundant, dedicated fiber** — high enough throughput for **synchronous replication** between AZs.
> - You choose your AZs; AWS recommends replicating data and resources across AZs for resiliency.

Data centers are the foundation of the whole structure — the place where data actually resides and data processing occurs. A typical AWS data center houses **50,000 to 80,000 physical servers** and has redundant power, networking, and connectivity, housed in a separate facility. Customers never deploy “into” a data center: the Availability Zone is the most granular level of specification a customer can make. Although rare, failures that affect the availability of instances in the same location do occur — which is exactly why hosting everything in a single location is a design mistake.

HOW AWS DESIGNS DATA CENTERS FOR SECURITY

- Every location is carefully evaluated to **mitigate environmental risk**; locations are not disclosed, and all access is restricted.
- Redundant design **anticipates and tolerates failure** while maintaining service levels; critical system components are backed up across multiple AZs.
- AWS continuously monitors service usage to deploy infrastructure ahead of **capacity** and availability commitments.
- In case of failure, **automated processes** move data traffic away from the affected area.
- Network equipment is custom, sourced from multiple **original device manufacturers (ODMs)** that build to AWS's specifications.

Points of Presence (PoPs) are AWS's fourth kind of location: a global network of **edge locations** plus a much smaller number of **Regional edge caches**, sited in most major cities around the world. They exist to cut latency. **Amazon CloudFront** — a content delivery network (CDN) — distributes content to end users with reduced latency, and **Amazon Route 53** — a Domain Name System (DNS) service — resolves names; requests to either service are **automatically routed to the nearest edge location**. AWS Shield and AWS WAF also run at the PoPs. By continuously measuring internet connectivity, performance, and computing, the PoPs find the best way to route requests and deliver a near-real-time user experience. **Regional edge caches**, used by default with CloudFront, hold content that is not accessed frequently enough to remain in an edge location — absorbing it there provides an alternative to fetching it all the way from the origin server.

COMMON PITFALL

The module's sample exam question asks which infrastructure component CloudFront uses to ensure low-latency delivery. The tempting wrong answer is "Availability Zones" — the correct answer is **edge locations**. AZs make applications *resilient*; edge locations make content delivery *fast*.

TABLE 3.2 Features of the AWS Global Infrastructure

FEATURE	WHAT IT MEANS
Elasticity & scalability	Resources dynamically adjust to increases or decreases in capacity requirements; the infrastructure rapidly adapts to accommodate growth
Fault tolerance	Continues operating properly in the presence of a failure — component redundancy is built in
High availability	High level of operational performance with minimized downtime — and no human intervention required

SECTION KEY TAKEAWAYS

- The AWS Global Infrastructure consists of Regions and Availability Zones.
- Your choice of Region is typically driven by **compliance requirements** or the need to **reduce latency**.
- Each Availability Zone is physically separate from the others and has redundant power, networking, and connectivity.
- Edge locations and Regional edge caches improve performance by **caching content closer to users**.

3.2 AWS services and service category overview

The infrastructure exists to host services. AWS pictures its offerings as a stack: the Global Infrastructure — Regions, Availability Zones, and edge locations — at the bottom; **foundational services** (compute, networking, storage) directly on top; **platform services** (databases, analytics, app services, deployment and management, mobile services) above those; and end-user **applications** such as virtual desktops and collaboration tools at the top. Every layer is delivered as an on-demand utility: available in seconds, with pay-as-you-go pricing.

AWS organizes its products into **23 product and service categories**, each consisting of one or more services. This course does not attempt to introduce every service; it focuses on the most widely used ones, the best introduction to the AWS Cloud, and the services most likely to appear on the **AWS Certified Cloud Practitioner exam**. Seven categories get that focus: **Compute · Cost Management · Database · Management and Governance · Networking and Content Delivery · Security, Identity, and Compliance · Storage**. The AWS Cloud Products page organizes every product by these categories, and each product page carries a detailed description and its benefits. The tables that follow give the one-line job of every service the course covers — precisely what "which service does X?" exam questions test.

TABLE 3.3 Storage services

SERVICE	ONE-LINE JOB
Amazon S3	Object storage offering scalability, data availability, security, and performance — store any amount of data for websites, mobile apps, backup/restore, archive, enterprise applications, IoT devices, and big data analytics
Amazon EBS	High-performance **block** storage designed for use with EC2, for throughput- and transaction-intensive workloads: databases, enterprise apps, containers, big data analytics engines, file systems, media workflows
Amazon EFS	Scalable, fully managed, elastic **NFS file system** for AWS Cloud services and on-premises resources — scales on demand to petabytes, growing and shrinking automatically as files are added and removed
Amazon S3 Glacier	Secure, durable, extremely **low-cost S3 storage class** for data archiving and long-term backup — designed for 11 9s of durability and stringent regulatory compliance

TABLE 3.4 Compute services

SERVICE	ONE-LINE JOB
Amazon EC2	Resizable compute capacity as **virtual machines** in the cloud
Amazon EC2 Auto Scaling	Automatically **adds or removes EC2 instances** according to conditions you define
Amazon ECS	Highly scalable, high-performance **container orchestration** service that supports Docker containers
Amazon ECR	Fully managed **Docker container registry** for storing, managing, and deploying container images
AWS Elastic Beanstalk	**Deploys and scales web applications** on familiar servers such as Apache and Microsoft IIS
AWS Lambda	**Run code without provisioning or managing servers** — pay only for the compute time consumed; no charge when your code isn't running
Amazon EKS	Deploy, manage, and scale containerized applications using **Kubernetes** on AWS
AWS Fargate	Compute engine for ECS that runs **containers without managing servers or clusters**

TABLE 3.5 Database services

SERVICE	ONE-LINE JOB
Amazon RDS	Set up, operate, and scale a **relational database** — automates hardware provisioning, database setup, patching, and backups
Amazon Aurora	MySQL- and PostgreSQL-**compatible** relational database — up to **5× faster** than standard MySQL and **3× faster** than standard PostgreSQL
Amazon DynamoDB	**Key-value and document** database with single-digit-millisecond performance at any scale; built-in security, backup and restore, and in-memory caching
Amazon Redshift	**Analytic queries** against petabytes stored locally in Redshift and exabytes stored directly in S3 — fast performance at any scale

TABLE 3.6 Networking and content delivery services

SERVICE	ONE-LINE JOB
Amazon VPC	Provision **logically isolated sections** of the AWS Cloud
Elastic Load Balancing	Automatically **distributes incoming application traffic** across multiple targets: EC2 instances, containers, IP addresses, and Lambda functions
Amazon CloudFront	Fast **CDN** — securely delivers data, videos, applications, and APIs globally with low latency and high transfer speeds
AWS Transit Gateway	Connects VPCs and on-premises networks to a **single gateway**
Amazon Route 53	Scalable cloud **DNS** — translates names like `www.example.com` into numeric IP addresses like `192.0.2.1`
AWS Direct Connect	**Dedicated private network connection** from your data center or office to AWS — can reduce network costs and increase bandwidth throughput
AWS VPN	**Secure private tunnel** from your network or device to the AWS global network

TABLE 3.7 Security, identity, and compliance services

SERVICE	ONE-LINE JOB
AWS IAM	**Manage access** to AWS services and resources securely — create and manage users and groups; allow and deny access via permissions
AWS Organizations	**Restrict what services and actions** are allowed in your accounts
Amazon Cognito	Add user **sign-up, sign-in, and access control** to web and mobile apps
AWS Artifact	On-demand access to AWS **security and compliance reports** and select online agreements
AWS KMS	**Create and manage encryption keys**; control the use of encryption across AWS services and your applications
AWS Shield	Managed **DDoS protection** that safeguards applications running on AWS

TABLE 3.8 Cost management services

SERVICE	ONE-LINE JOB
AWS Cost and Usage Report	The **most comprehensive** set of AWS cost and usage data available, including metadata about services, pricing, and reservations
AWS Budgets	Set **custom budgets** that alert you when costs or usage exceed — or are forecasted to exceed — the budgeted amount
AWS Cost Explorer	**Visualize, understand, and manage** your AWS costs and usage over time

TABLE 3.9 Management and governance services

SERVICE	ONE-LINE JOB
AWS Management Console	**Web-based user interface** for accessing your AWS account
AWS Config	Track **resource inventory and changes**
Amazon CloudWatch	**Monitor** resources and applications
AWS Auto Scaling	Scale **multiple resources** to meet demand
AWS CLI	**Unified command line tool** to manage AWS services
AWS Trusted Advisor	Optimize **performance and security**
AWS Well-Architected Tool	**Review and improve** your workloads
AWS CloudTrail	Track **user activity and API usage**

The module closes with a hands-on **AWS Management Console clickthrough** in the sandbox environment. The console's **Services** menu groups every service by category (EC2 appears under Compute, and so on), and the Region selector in the top-right corner scopes what you see: switching from N. Virginia to EU (London) reveals that Region's own VPCs and subnets. The activity's five questions distill into facts worth keeping:

CONSOLE ACTIVITY – THE FIVE ANSWERS

Q1 IAM appears under the **Security, Identity, & Compliance** category. **Q2** Amazon VPC appears under **Networking & Content Delivery**. **Q3** Subnets exist at the level of the **Availability Zone**. **Q4** VPCs exist at the **Region** level. **Q5** Of EC2, IAM, Lambda, and Route 53: **IAM and Route 53 are global; EC2 and Lambda are Regional**.

COMMON PITFALL

Global vs Regional trips people constantly. The console's Region selector affects **Regional** services — EC2, Lambda, and VPC resources created in one Region don't appear in another. **Global** services (IAM, Route 53) look identical from every Region. And within a Region, remember the scoping: **VPC = Region level, subnet = Availability Zone level**.

Module summary

- The AWS Global Infrastructure delivers a flexible, reliable, scalable, and secure cloud with high-quality global network performance — and the footprint grows continually.
- Hierarchy: Region (geographical area) → two or more Availability Zones (fully isolated partitions) → data centers (50,000–80,000 servers each). Customer specification stops at the AZ.
- Regions are isolated from one another for fault tolerance and stability; data never leaves a Region unless you replicate it; inter-Region traffic uses the AWS backbone. Post-March 2019 Regions are disabled by default; China and GovCloud (US) are restricted.
- Choose a Region on four factors: data governance and legal requirements, proximity to customers (latency), services available in the Region, and cost — prices vary by Region.

- AZs have their own power, sit many kilometers apart (all within 100 km), and are linked by fully redundant dedicated fiber fast enough for synchronous replication — replicate across AZs for resiliency.
- Points of Presence = edge locations plus a smaller number of Regional edge caches; CloudFront and Route 53 automatically answer from the nearest edge location; Regional edge caches absorb infrequently accessed content.
- The infrastructure is elastic and scalable, fault-tolerant (built-in redundancy), and highly available with minimal to no human intervention.
- AWS has 23 service categories; the course covers seven — Storage, Compute, Database, Networking & Content Delivery, Security/Identity/Compliance, Cost Management, Management & Governance. Know each core service's one-line job — and that IAM and Route 53 are global while EC2 and Lambda are Regional.

Review questions

1. **Explain the relationship between Regions, Availability Zones, and data centers.**

   Answer: A Region is a geographical area containing two or more Availability Zones; each AZ is a fully isolated partition of the infrastructure made of one or more discrete data centers (typically three); a data center is the facility where data resides and processing occurs. Customers can specify down to the AZ — never an individual data center.

2. **Why doesn't AWS automatically replicate your data to other Regions, and what does that mean for you?**

   Answer: Regions are deliberately isolated from one another for fault tolerance and stability, and keeping data inside its Region supports data-governance requirements. If your business needs cross-Region copies, setting up that replication is your responsibility.

3. **Your users are in Frankfurt and regulations require customer data to stay in the EU. Walk through the four Region-selection factors.**

   Answer: Governance/legal: EU rules (like the Data Protection Directive) require an EU Region. Latency: pick the Region closest to the users — test with CloudPing. Services: confirm the services you need are available in that Region. Cost: compare prices, since the same instance costs different amounts in different Regions.

4. **How do Availability Zones make applications more available than a single data center ever could?**

   Answer: Each AZ is a fault-isolated partition with its own power, physically separated by many kilometers (yet within 100 km), so a local disaster — lightning, tornado, earthquake — hits at most one AZ. The AZs are linked by fully redundant dedicated fiber that supports synchronous replication, so an application partitioned across AZs keeps running when one fails.

5. **Trace what happens when a user requests content through CloudFront.**

   Answer: The request is automatically routed to the nearest edge location and served from cache with low latency. Content not accessed frequently enough to stay at the edge is held in a Regional edge cache, which serves it without going back to the origin server. (Route 53 DNS lookups are answered from the nearest edge location the same way.)

6. **Distinguish elasticity, scalability, fault tolerance, and high availability as features of the infrastructure.**

   Answer: Elastic: capacity dynamically adjusts to increases or decreases in requirements. Scalable: the infrastructure adapts to accommodate growth. Fault-tolerant: it continues operating properly through a component failure because redundancy is built in. Highly available: it delivers high operational performance and minimized downtime with no human intervention.

7. Which service category would you look in for each: Lambda, DynamoDB, Route 53, Budgets, CloudTrail, EFS?

Answer: Lambda: Compute. DynamoDB: Database. Route 53: Networking & Content Delivery. Budgets: Cost Management. CloudTrail: Management & Governance. EFS: Storage.

8. In the console activity, why did switching Regions change the VPCs and subnets you saw, but not your IAM users?

Answer: VPC, EC2, and Lambda are Regional — each Region has its own resources, with VPCs scoped to the Region and subnets scoped to a single Availability Zone. IAM (like Route 53) is global, so its users, groups, and policies apply account-wide no matter which Region the console selector shows.