CloudNinja.pro

ASSOCIATE · SAA-C03

AWS Solutions Architect Associate

SAA-C03 Cheat Sheet

CloudNinja · cloudninja.pro · Free — no signup required

A dense, scannable reference covering every major service and concept on the SAA-C03 exam. Organised by architecture domain — not alphabetically — so you build mental models, not just memorise facts.

EXAM SNAPSHOT

- **65 questions** — 50 scored + 15 unscored
- **130 minutes** total time
- **720 / 1000** to pass
- **$150 USD** exam fee
- **Secure** — 30%
- **Resilient** — 26%
- **High-Performing** — 24%
- **Cost-Optimised** — 20%

FREE SAA-C03 practice exam →

Closed or retired services. AWS has closed several services to new customers that older SAA-C03 questions still name, including Amazon Kendra, Simple AD, CloudTrail Lake, Application Discovery Service, Migration Hub and Timestream for LiveAnalytics. Amazon Pinpoint, Elastic Transcoder and QLDB are retired. They stay in this sheet with a DEPRECATED or RETIRED tag: know what they did, but pick the replacement for a new design (list on the last page).

1 · Networking & Content Delivery

Amazon VPC

A logically isolated virtual network you define inside AWS. You control IP ranges (CIDR blocks), subnets, route tables, and gateways. Every new AWS account gets a default VPC in each Region.

Component	What It Does
Subnet	A range of IPs inside a VPC. Public subnets have a route to an Internet Gateway; private subnets do not.
Internet Gateway (IGW)	Allows resources in public subnets to reach the internet. One per VPC, horizontally scaled by AWS.
NAT Gateway	Lets private-subnet instances reach the internet for updates without exposing them to inbound traffic. Deploy one per AZ for HA.
Route Table	Rules that determine where network traffic is directed. Each subnet is associated with exactly one route table.
Elastic IP	Static public IPv4 address. Remap between instances. One free per running instance; charged when unattached.
Security Group	Stateful virtual firewall at the instance/ENI level. Allow rules only — no explicit deny. Return traffic is automatically allowed.
Network ACL (NACL)	Stateless firewall at the subnet level. Supports allow AND deny rules. Evaluated in rule-number order. Both inbound and outbound rules needed.
VPC Peering	Private connectivity between two VPCs (same or different accounts/Regions). Not transitive — A↔B and B↔C does NOT give A↔C.
VPC Endpoints	Private connection to AWS services without traversing the internet. Gateway endpoints (S3, DynamoDB) are free; Interface endpoints (most other services) use PrivateLink and cost per hour + per GB.
VPC Flow Logs	Capture IP traffic info for network interfaces in your VPC. Publish to CloudWatch Logs, S3, or Amazon Data Firehose. Useful for troubleshooting and security analysis.

Exam Tip: Security Groups are stateful (return traffic auto-allowed). NACLs are stateless (you must add both inbound AND outbound rules). This is one of the most frequently tested distinctions.

Elastic Load Balancing (ELB)

Distributes incoming traffic across multiple targets (EC2, containers, IPs, Lambda). Three types:

Type	Layer	Best For	Key Features
Application LB (ALB)	Layer 7 (HTTP/S)	Web apps, microservices, containers	Path-based and host-based routing, WebSocket, gRPC, sticky sessions, native WAF integration, authentication (OIDC/Cognito)
Network LB (NLB)	Layer 4 (TCP/UDP/TLS)	Ultra-low latency, static IPs, gaming, IoT	Millions of requests/sec, static IP per AZ, preserves source IP, TLS termination, PrivateLink compatible
Gateway LB (GWLB)	Layer 3 (IP)	Third-party virtual appliances (firewalls, IDS)	Transparent network gateway + load balancer, uses GENEVE encapsulation

- **Cross-zone load balancing:** Distributes traffic evenly across all registered targets in all AZs. Enabled by default for ALB; optional for NLB.

- **Connection draining (deregistration delay):** Gives in-flight requests time to complete before removing a target. Default 300s.

> **Exam Tip:** If the question mentions 'static IP' or 'whitelisting IPs' — think NLB. If it mentions 'URL path routing' or 'host-based routing' — think ALB. If it mentions 'inline security appliances' — think GWLB.

Amazon CloudFront

Global CDN with 600+ edge locations. Caches content close to users for low latency. Supports static and dynamic content, WebSocket, and video streaming.

- **Origin types:** S3 bucket, ALB/NLB, EC2, custom HTTP server, MediaPackage (MediaStore was shut down in Nov 2025)
- **Origin Access Control (OAC)** restricts S3 access so users must go through CloudFront — replaces legacy OAI
- **Signed URLs:** one file, one URL. **Signed Cookies:** multiple files, one cookie. Use for premium/paywalled content
- **Field-level encryption:** protect sensitive POST fields at the edge before they reach your origin
- **Cache behaviours:** match URL path patterns to different origins, TTL settings, and viewer protocols

Edge Compute: Lambda@Edge vs. CloudFront Functions

Feature	CloudFront Functions	Lambda@Edge
Runtime	JavaScript only	Node.js, Python
Execution location	218+ edge locations	13 Regional edge caches
Max execution time	1 ms	5 s (viewer) / 30 s (origin)
Max memory	2 MB	128 – 10,240 MB
Network / file access	No	Yes
Best for	Header manipulation, URL rewrites, cache key normalisation	Complex logic, origin selection, A/B testing, authentication
Event types	Viewer request / viewer response only	All 4: viewer req/resp + origin req/resp

> **Exam Tip:** CloudFront + S3 with OAC is the standard pattern for serving static websites securely. If a question says 'users worldwide report slow page loads' — CloudFront is almost always the answer.

Amazon Route 53

Highly available DNS service. Registers domains, routes traffic, and health-checks endpoints.

Routing Policy	Use Case
Simple	Single resource, no health checks. Returns all values in random order if multiple records.
Weighted	Split traffic by percentage (e.g. 80/20 blue-green deployment).
Latency-based	Route to the Region with lowest latency for the user.
Failover	Active-passive. Route to standby when primary fails health check.
Geolocation	Route based on user's geographic location (continent, country, US state).
Geoproximity	Route based on geographic distance; use bias values to shift traffic between Regions.
Multivalue Answer	Return up to 8 healthy records; client picks one. Not a substitute for a load balancer.
IP-based	Route based on client's source IP CIDR. Useful for ISP-specific routing.

- **Health checks:** Monitor endpoint health (HTTP, HTTPS, TCP). Calculated health checks combine multiple checks. CloudWatch alarm-based checks for private resources.

- **Alias records:** Route 53 extension — point to AWS resources (ALB, CloudFront, S3 website, API Gateway) at the zone apex (no CNAME at root). Free of charge.

> **Exam Tip:** Failover routing + health checks is the standard answer for 'automatic DNS failover.' Latency-based routing is for 'route users to the fastest Region.'

Hybrid Connectivity

Service	What It Does	Key Details
AWS Direct Connect	Dedicated private network link from on-premises to AWS	1 Gbps or 10 Gbps dedicated; 50 Mbps–10 Gbps hosted. Consistent latency, no internet. Takes weeks to provision. Not encrypted by default — add VPN on top for encryption.
AWS Site-to-Site VPN	Encrypted tunnel over public internet to your VPC	Quick to set up (minutes). Up to 1.25 Gbps per tunnel. Uses IPSec. Two tunnels for redundancy.
AWS Client VPN	Managed VPN for individual users to connect to AWS or on-prem	OpenVPN-based. For remote workers accessing AWS resources.
AWS Transit Gateway	Hub-and-spoke network connecting VPCs, VPNs, and Direct Connect	Simplifies complex topologies. Supports transitive routing (unlike VPC peering). Cross-Region peering. Thousands of VPCs via a single gateway.
AWS PrivateLink	Private access to services over AWS network without internet exposure	Interface VPC endpoints. Access your own or third-party services privately. Preferred when you need to expose a service to customers without VPC peering.

> **Exam Tip:** 'Consistent network performance' + 'private connection' = Direct Connect. 'Quick setup' + 'encrypted' = VPN. Need both consistency AND encryption? = Direct Connect + VPN. 'Connect hundreds of VPCs' = Transit Gateway.

AWS Global Accelerator

Routes traffic through the AWS global network (not the public internet) to the optimal Regional endpoint. Provides two static anycast IPs. Instant regional failover. Good for non-HTTP workloads (gaming, IoT, VoIP) or when you need static IP addresses.

> **Exam Tip:** Global Accelerator = static IPs + AWS backbone routing + TCP/UDP. CloudFront = HTTP/S content caching at the edge. Both improve global performance but in different ways.

2 · Security, Identity & Compliance

AWS IAM

Global service (not Region-specific) that controls who can do what in your AWS account. Follows the principle of least privilege — grant only the permissions needed.

Concept	Explanation
Users	Individual identities with long-term credentials. Map to a real person or application.
Groups	Collections of users. Attach policies to groups, not individual users, for easier management. Cannot be nested.
Roles	Temporary credentials assumed by users, services, or external identities. No passwords — use STS tokens. EC2 instance profiles, Lambda execution roles, cross-account access.
Policies	JSON documents defining permissions (Effect, Action, Resource, Condition). Identity-based (attached to user/group/role) or resource-based (attached to the resource itself, e.g. S3 bucket policy).

Concept	Explanation
AWS STS	Security Token Service — issues temporary credentials when a role is assumed. AssumeRole, AssumeRoleWithSAML, AssumeRoleWithWebIdentity.
Service Control Policies (SCPs)	Organization-level guardrails that limit what member accounts CAN do. SCPs don't grant permissions — they restrict them. Applied via AWS Organizations.
Permission Boundaries	IAM feature to set the maximum permissions a role or user can have. Useful for delegated admin — let developers create roles but cap their permissions.

- **Policy evaluation logic:** Explicit Deny always wins → SCPs → Resource-based policies → Permission boundaries → Identity-based policies. If no explicit Allow, default is Deny.

Exam Tip: When a question says 'an EC2 instance needs to access S3' — the answer is an IAM role attached via an instance profile, NEVER access keys stored on the instance. Similarly, Lambda functions use execution roles.

AWS IAM Identity Center (formerly SSO)

Central place to manage workforce access to multiple AWS accounts and business applications. Supports SAML 2.0 federation with external IdPs (Active Directory, Okta, etc.). Assign permission sets to users/groups across your Organization.

AWS KMS & Encryption

Managed service for creating and controlling encryption keys used to encrypt your data.

Key Type	Details
AWS owned keys	Used by AWS services internally. No visibility, no cost, no management.
AWS managed keys	Created and managed by AWS for integrated services (e.g. aws/s3). Free. You cannot manage rotation or policies. Visible in your account.
Customer managed keys (CMKs)	You create, manage, and set policies. Automatic annual rotation available. $1/month per key. Full control over access.
AWS CloudHSM	Dedicated hardware security modules you control entirely. FIPS 140-2 Level 3. You manage your own keys. Use when you need full key control for compliance.

- **Encryption at rest:** KMS integrates with S3, EBS, RDS, EFS, DynamoDB, Redshift, and more. Enable it with a single checkbox or API parameter.
- **Encryption in transit:** Use TLS/SSL. AWS Certificate Manager (ACM) provides free public TLS certificates for ALB, CloudFront, and API Gateway.
- **Envelope encryption:** KMS encrypts a data key; the data key encrypts your data. Required when encrypting data larger than 4 KB directly.
- **Key policies + grants:** Key policy is the primary access control for CMKs. Grants allow temporary, granular access to specific operations without modifying the key policy.

Secrets Management

Service	Best For	Key Details
AWS Secrets Manager	Database credentials, API keys, OAuth tokens	Automatic rotation with built-in Lambda functions for RDS, Redshift, DocumentDB. Cross-account access. Costs $0.40/secret/month.
SSM Parameter Store	Configuration values, feature flags, non-rotating secrets	Free for Standard tier (up to 10,000 params). SecureString uses KMS. No built-in rotation — use Lambda. Advanced tier for >10K params or larger values.

Exam Tip: If the question mentions 'automatic credential rotation' — Secrets Manager. If it mentions 'store configuration data cheaply' — Parameter Store.

Threat Detection & Protection

Service	What It Does
Amazon GuardDuty	Intelligent threat detection using ML. Analyses VPC Flow Logs, DNS logs, CloudTrail events, S3 data events, EKS audit logs. No agents to install. Detects compromised instances, crypto mining, unusual API calls.
Amazon Inspector	Automated vulnerability scanning for EC2 instances, Lambda functions, and container images in ECR. Checks for CVEs and network reachability. Continuous scanning.
Amazon Macie	Uses ML to discover and protect sensitive data (PII, financial data) in S3 buckets. Alerts on public buckets or unencrypted sensitive data.
AWS WAF	Web Application Firewall on ALB, CloudFront, API Gateway, AppSync. Block SQL injection, XSS, rate limiting, geo-blocking. Uses Web ACLs with rules. AWS Managed Rules for common threats.
AWS Shield	DDoS protection. Standard is free and automatic on all AWS resources. Advanced ($3K/month) adds 24/7 DDoS Response Team, cost protection, and enhanced detection for ELB, CloudFront, Route 53, Global Accelerator.
AWS Firewall Manager	Centrally manage WAF, Shield Advanced, Security Groups, and Network Firewall rules across accounts in an Organization.
AWS Security Hub	Central security dashboard. Aggregates findings from GuardDuty, Inspector, Macie, Firewall Manager, IAM Access Analyzer. Runs automated compliance checks (CIS, PCI DSS, AWS Foundational). Single pane of glass for security posture.

Amazon Cognito

Authentication and authorisation for web and mobile apps.

- **User Pools:** Sign-up/sign-in, MFA, social/SAML/OIDC federation. Returns JWT tokens (ID, access, refresh). Think 'authentication' — who is this person? Supports custom auth flows with Lambda triggers (pre-signup, post-auth, token generation).
- **Identity Pools:** Exchange tokens (from User Pools, Google, Facebook, SAML, OIDC) for temporary AWS credentials via STS. Think 'authorisation' — what can this person do in AWS? Supports unauthenticated (guest) access with limited IAM role.

> **Exam Tip:** 'Web app needs Google login' = Cognito User Pool (social federation). 'Mobile app needs to upload to S3 directly' = Cognito Identity Pool (temporary AWS creds). Many apps use BOTH: User Pool for auth, Identity Pool for AWS resource access.

AWS Directory Service

- **AWS Managed Microsoft AD:** Full Microsoft Active Directory on AWS. Trust relationships with on-prem AD. For enterprises needing AD-aware workloads (SQL Server, .NET, SharePoint).
- **AD Connector:** Proxy that redirects directory requests to your existing on-prem AD. No caching, no directory data on AWS. For SSO and EC2 domain join.
- **Simple AD** DEPRECATED **:** Samba-based, basic AD-compatible directory. Closed to new customers (June 2026): use AWS Managed Microsoft AD, or AD Connector for an existing directory.

AWS Resource Access Manager (RAM)

Share AWS resources across accounts or within your Organization. Common shared resources: Transit Gateways, subnets, License Manager configurations, Route 53 Resolver rules, Aurora DB clusters. Avoids resource duplication and reduces cost.

3 · Compute

Amazon EC2

Virtual servers in the cloud. Choose OS, instance type, storage, and networking. Full control over the environment.

Instance Families

Family	Optimised For	Examples
General Purpose (M, T)	Balanced CPU/memory. Web servers, small DBs, dev/test.	t3.micro, m6i.large
Compute Optimised (C)	CPU-intensive: batch processing, HPC, ML inference, gaming.	c6i.xlarge
Memory Optimised (R, X, z)	In-memory databases, big data analytics, real-time processing.	r6i.2xlarge, x2idn.large
Storage Optimised (I, D, H)	High sequential I/O: data warehousing, distributed file systems.	i3.large, d3.xlarge
Accelerated Computing (P, G, Inf, Trn)	GPU/FPGA workloads: ML training, video encoding, graphics.	p4d.24xlarge, g5.xlarge

Purchasing Options

Option	Discount	Commitment	Best For
On-Demand	None (baseline)	None	Short-term, spiky, unpredictable workloads
Reserved Instances (RI)	Up to 72%	1 or 3 years	Steady-state workloads (databases, baseline servers)
Savings Plans	Up to 72%	1 or 3 years ($/hr)	Flexible across instance families, Regions, or compute services (Lambda, Fargate)
Spot Instances	Up to 90%	None — can be interrupted	Fault-tolerant batch jobs, CI/CD, data processing, HPC
Dedicated Hosts	Varies	On-Demand or Reserved	Compliance, BYOL (bring your own licence), socket/core licensing
Dedicated Instances	Varies	On-Demand pricing	Hardware isolation without host-level visibility
Capacity Reservations	None (On-Demand price)	None	Guarantee capacity in a specific AZ without commitment

Placement Groups

- **Cluster:** Instances packed close together in one AZ. Lowest latency for HPC, tightly coupled workloads.
- **Spread:** Each instance on distinct hardware. Max 7 per AZ. Critical instances that must not share failure domains.
- **Partition:** Instances spread across logical partitions (each on separate racks). Large distributed workloads like HDFS, Cassandra, Kafka.

Instance Store vs. EBS

Feature	Instance Store	EBS
Persistence	Ephemeral — data lost on stop/terminate/failure	Persistent — survives stop/start
Performance	Very high IOPS (physically attached NVMe/SSD)	Good, but over network (io2 Block Express for highest)
Snapshots	Not supported	Snapshots to S3, cross-Region copy
Use case	Caches, buffers, temp data, scratch space	Boot volumes, databases, persistent storage

Exam Tip: Spot + Auto Scaling + multiple instance types = cost-effective fault-tolerant compute. If a question says 'most cost-effective for interruptible workload' — Spot. 'Predictable baseline' — Reserved or Savings Plan. 'Highest IOPS for temporary data' — Instance Store.

EC2 Auto Scaling

Automatically adjusts the number of EC2 instances based on demand. Uses launch templates to define instance configuration.

- **Scaling policies:** Target tracking (keep CPU at 50%), step scaling (add 2 when CPU > 80%), simple scaling (one adjustment per alarm), scheduled scaling (time-based), predictive scaling (ML-based forecast).
- **Cooldown period** prevents rapid scale in/out oscillation. Default 300 seconds.
- **Health checks:** EC2 status checks (default) or ELB health checks for web-facing apps.
- **Warm pools:** pre-initialised instances in a stopped state, ready to join the group quickly. Reduces scale-out time for apps with long boot.

AWS Lambda

Serverless compute — upload your code, and AWS runs it in response to events. No servers to manage. Pay per invocation + execution duration (per 1ms).

Parameter	Limit / Detail
Timeout	Max 15 minutes per invocation
Memory	128 MB to 10,240 MB (CPU scales with memory)
Deployment package	50 MB zipped, 250 MB unzipped (use layers or container images for larger)
Concurrent executions	1,000 per Region (soft limit, can be increased)
Triggers	API Gateway, S3 events, DynamoDB Streams, SQS, SNS, EventBridge, Kinesis, CloudWatch Events, ALB, Cognito
VPC access	Can run inside a VPC for private resource access (RDS, ElastiCache). Uses Hyperplane ENIs — no cold start penalty since 2019.
Provisioned Concurrency	Pre-initialise execution environments to eliminate cold starts. Use for latency-sensitive workloads.
Lambda Layers	Share common code/libraries across functions. Up to 5 layers per function.
Lambda@Edge	Run Lambda at CloudFront edge locations. Viewer/origin request/response events (see CloudFront section).

Exam Tip: Lambda is the default 'serverless compute' answer. But if the workload runs >15 minutes, needs persistent local storage, or requires GPU — use Fargate or EC2 instead.

Containers on AWS

Service	What It Does
Amazon ECS	AWS-native container orchestration. Task definitions specify containers, CPU/memory, IAM roles. Uses EC2 or Fargate launch types.
Amazon EKS	Managed Kubernetes. Choose EKS if you already use Kubernetes or need Kubernetes-specific features. Also supports EC2 or Fargate.
AWS Fargate	Serverless compute engine for ECS and EKS. No EC2 instances to manage. Specify CPU/memory per task. Good for variable, bursty container workloads.
Amazon ECR	Managed container image registry. Stores, scans (via Inspector), and encrypts Docker images. Integrates with ECS, EKS, and Lambda.

Exam Tip: 'Migrate from on-premises Kubernetes' = EKS. 'Run containers without managing servers' = Fargate. 'Docker containers with simple orchestration' = ECS.

AWS Elastic Beanstalk

PaaS for deploying web apps. Upload your code, and Beanstalk handles capacity provisioning, load balancing, auto-scaling, and health monitoring. You still have full access to the underlying EC2 instances. Supports Java, .NET, Node.js, Python, Ruby, Go, Docker. Free — you pay only for the resources created.

AWS Batch

Fully managed batch processing at any scale. Automatically provisions the optimal compute resources (EC2 or Fargate) based on queue volume and job requirements. Supports Spot for cost savings. No need to install or manage batch computing software.

- **Use case:** Large-scale data processing, ML model training, financial modelling, genomics, rendering, ETL workloads that need to run thousands of jobs.
- **Components:** Job definitions (container image, vCPU, memory), job queues (priority ordering), compute environments (managed or unmanaged).

> **Exam Tip:** 'Run thousands of batch jobs efficiently' = AWS Batch. 'Short serverless event-driven compute' = Lambda. 'Long-running container without managing servers' = Fargate.

4 · Storage

Amazon S3

Object storage with virtually unlimited capacity. Objects stored in buckets. Each object can be up to 5 TB. 11 nines (99.999999999%) durability. Globally unique bucket names.

Storage Classes

Class	Use Case	Availability	Retrieval
S3 Standard	Frequently accessed data	99.99%	Instant
S3 Intelligent-Tiering	Unknown or changing access patterns	99.9%	Instant (auto-moves between tiers)
S3 Standard-IA	Infrequent access, rapid retrieval needed	99.9%	Instant, per-GB retrieval fee
S3 One Zone-IA	Infrequent, re-creatable data (thumbnails, replicas)	99.5% (1 AZ)	Instant, per-GB retrieval fee
S3 Glacier Instant Retrieval	Archive data accessed once per quarter	99.9%	Milliseconds
S3 Glacier Flexible Retrieval	Archives with minutes-to-hours retrieval	99.9%	Expedited (1-5 min), Standard (3-5 hr), Bulk (5-12 hr)
S3 Glacier Deep Archive	Lowest cost, regulatory/compliance archives	99.9%	Standard (12 hr), Bulk (48 hr)

Key Features

- **Versioning:** Keep multiple versions of an object. Protects against accidental deletion (delete markers, not permanent). Needed for cross-Region replication.
- **Lifecycle policies:** Automatically transition objects between storage classes or expire them after N days. Combine with Intelligent-Tiering for unknown patterns.
- **Replication:** Cross-Region (CRR) for compliance/latency or Same-Region (SRR) for log aggregation. Requires versioning on both source and destination. Can replicate to different account.
- **S3 Transfer Acceleration:** Uses CloudFront edge locations to speed uploads from distant locations.
- **Multipart upload:** Required for objects >5 GB, recommended for >100 MB. Parallel uploads improve throughput.

- **S3 Select / Glacier Select** DEPRECATED **:** Query subsets of data using SQL without retrieving the entire object. Closed to new customers since **July 25, 2024** — existing customers can continue using it, but no new features. Modern alternative: Athena on S3.
- **Pre-signed URLs:** Time-limited URLs granting temporary access to private objects. Great for sharing files with external users without making bucket public.
- **Bucket policies vs. ACLs:** Use bucket policies (JSON) for most access control. ACLs are legacy — use only for cross-account object ownership edge cases. Block Public Access setting overrides all.
- **S3 Object Lock:** WORM (Write Once Read Many) model. Governance mode (privileged users can override) or Compliance mode (nobody can delete, not even root). Required for regulatory compliance.

S3 Event Notifications

Trigger actions when objects are created, deleted, or restored. Destinations: SNS, SQS, Lambda, or EventBridge. Common patterns: thumbnail generation on upload (Lambda), audit logging on delete (SQS), real-time search indexing (Lambda → OpenSearch).

- **EventBridge integration:** Advanced filtering, multiple targets, and replay capability. Recommended over direct SNS/SQS for new designs.

> **Exam Tip:** S3 + CloudFront OAC for static websites. S3 lifecycle policies for cost optimisation. S3 Object Lock in Compliance mode for regulatory requirements that 'no one should be able to delete.' S3 event notifications + Lambda for processing uploads.

Amazon EBS (Elastic Block Store)

Persistent block storage for EC2. Attached to a single instance at a time (except multi-attach io1/io2). Survives instance stop/start. Snapshots stored in S3.

Volume Type	Type	Max IOPS	Max Throughput	Use Case
gp3 (General Purpose)	SSD	16,000	1,000 MB/s	Boot volumes, dev/test, virtual desktops. Baseline 3,000 IOPS free.
gp2 (General Purpose)	SSD	16,000	250 MB/s	Legacy. Burstable IOPS tied to volume size (3 IOPS/GiB).
io2 Block Express	SSD	256,000	4,000 MB/s	Mission-critical databases (Oracle, SAP). Sub-millisecond latency.
io1 / io2	SSD	64,000	1,000 MB/s	High-performance databases needing sustained IOPS. io1 is previous generation — use io2 for 100× better durability at same price.
st1 (Throughput Opt.)	HDD	500	500 MB/s	Big data, data warehouses, log processing. Cannot be boot volume.
sc1 (Cold HDD)	HDD	250	250 MB/s	Lowest cost. Infrequent access, archival. Cannot be boot volume.

- **Snapshots:** Incremental backups to S3. Copy across Regions for DR. Can create AMIs from snapshots. EBS Fast Snapshot Restore for instant full-performance.
- **Encryption:** AES-256 using KMS. Encrypt at creation or copy an unencrypted snapshot to create an encrypted copy. All data in flight between EC2 and EBS is encrypted.
- **Multi-attach (io1/io2):** Attach a single volume to up to 16 Nitro instances in the same AZ. For clustered applications that manage concurrent writes.

> **Exam Tip:** gp3 is the default choice for most workloads (cheaper than gp2 with independently tuneable IOPS). io2/io2 Block Express for databases needing guaranteed IOPS. st1 for sequential throughput. Remember: HDD volumes cannot be boot volumes.

Amazon EFS (Elastic File System)

Managed NFS file system. Shared across multiple EC2 instances and AZs simultaneously. Auto-scales from GB to PB. Supports Linux only (NFS v4.1). Standard and Infrequent Access tiers with lifecycle management.

- **Performance modes:** General Purpose (default, low latency) or Max I/O (higher throughput for HPC). Throughput modes: Bursting, Provisioned, or Elastic (recommended).
- **EFS One Zone:** Single-AZ option at ~47% lower cost. For dev/test or data that doesn't need multi-AZ resilience.

Exam Tip: Multiple EC2 instances need shared file storage? EFS. Windows workloads? FSx for Windows. High-performance computing (HPC)? FSx for Lustre. Single-instance block storage? EBS.

Amazon FSx

Variant	Protocol	Use Case
FSx for Windows File Server	SMB, NTFS, AD integration	Windows applications, SharePoint, SQL Server, home directories. Multi-AZ for HA.
FSx for Lustre	Lustre (POSIX)	HPC, ML training, video processing. Can integrate with S3 as data repository. Scratch (temporary) or Persistent (long-term).
FSx for NetApp ONTAP	NFS, SMB, iSCSI	Hybrid environments, workloads needing multi-protocol access. Data deduplication and compression.
FSx for OpenZFS	NFS	Linux workloads migrating from on-prem ZFS/NAS. Snapshots, clones, data compression.

AWS Storage Gateway

Hybrid storage bridge between on-premises environments and AWS cloud storage.

Type	Interface	Use Case
S3 File Gateway	NFS / SMB	Store files as S3 objects with local cache. Good for backups, media workflows.
FSx File Gateway DEPRECATED	SMB	Cached on-prem access to FSx for Windows. Closed to new customers (Oct 2024) → access FSx for Windows File Server directly over Direct Connect / VPN.
Volume Gateway (Cached)	iSCSI	Primary data in S3, frequently accessed data cached locally.
Volume Gateway (Stored)	iSCSI	Primary data on-premises, async backup to S3 as EBS snapshots.
Tape Gateway	iSCSI (VTL)	Replace physical tape backups. Data stored in S3 Glacier / Deep Archive.

Exam Tip: 'On-premises app needs NFS access to S3' = S3 File Gateway. 'Replace tape backup' = Tape Gateway. 'Low-latency local access to S3 data' = Volume Gateway (Cached).

AWS Snow Family DEPRECATED

Important: AWS has significantly reduced the Snow Family. Effective **November 7, 2025**, Snowball Edge is available to **existing customers only** — new customers are directed to **AWS DataSync** (online transfer) or **AWS Data Transfer Terminal** (secure physical upload locations). Snowmobile and Snowcone are fully retired. The SAA-C03 exam may still reference these services — learn to recognise them in a scenario, but recommend DataSync / Data Transfer Terminal for new designs.

Device	Status	Details
Snowball Edge Storage Optimised (210 TB)	Current; existing customers only	NVMe storage, up to 1.5 GB/s transfer. Large offline data migrations.

Device	Status	Details
Snowball Edge Compute Optimised (AMD EPYC Gen2)	Current; existing customers only	104 vCPUs, 416 GB RAM, 28 TB NVMe SSD. Edge computing, ML inference.
Snowball Edge 80 TB / 42 TB / with GPU DEPRECATED	Retired Nov 12, 2024	Previous-generation models. No longer orderable. May still appear on exam questions.
Snowcone (8 TB HDD / 14 TB SSD) DEPRECATED	Retired Nov 12, 2024 (support ended Nov 12, 2025)	Portable, rugged edge device. Fully discontinued. Use DataSync or Data Transfer Terminal.
Snowmobile (100 PB truck) DEPRECATED	Retired April 2024	Exabyte-scale migration service. Permanently discontinued.

Exam Tip: For a modern answer on 'bulk offline data transfer,' favour **AWS DataSync** (online) or **AWS Data Transfer Terminal** (physical drop-off). Snow Family may still be a correct answer on older SAA-C03 question banks but flag it mentally as legacy.

AWS Backup

Centrally manage and automate backups across AWS services: EC2, EBS, RDS, Aurora, DynamoDB, EFS, FSx, Storage Gateway, S3, and more. Supports cross-Region and cross-account backup.

- **Backup plans:** Define frequency, retention, lifecycle (transition to cold storage), and destination vault.
- **Backup Vault Lock:** WORM model for backup data. Compliance mode prevents anyone from deleting backups before retention expires. Supports regulatory requirements (SEC, FINRA).
- **AWS Backup Audit Manager:** Evaluate backup compliance with controls like 'all RDS instances have daily backups' or 'all backups are encrypted.'

Exam Tip: 'Centralise backup management across multiple AWS services' = AWS Backup. 'Ensure backups cannot be deleted for compliance' = Backup Vault Lock in compliance mode.

5 · Database

Amazon RDS

Managed relational database supporting MySQL, PostgreSQL, MariaDB, Oracle, SQL Server, and IBM Db2. Handles backups, patching, scaling, and HA. You do NOT get SSH access to the underlying instance.

- **Multi-AZ:** Synchronous standby replica in another AZ for automatic failover. Not for read scaling — for HA only. Failover typically 60-120 seconds.
- **Multi-AZ Cluster (new):** Two readable standbys in different AZs. Provides both HA and read scaling. Transaction commit latency ~35ms. Available for MySQL and PostgreSQL.
- **Read Replicas:** Asynchronous copies for read scaling. Up to 15 replicas (5 for non-Aurora). Cross-Region supported. Can be promoted to standalone DB.
- **Automated backups:** Daily snapshots + transaction logs. Point-in-time recovery up to 5 minutes ago. Retained 1-35 days.
- **RDS Proxy:** Managed connection pooler. Reduces DB connection overhead for Lambda or apps with many short-lived connections. Improves failover time by 66%.
- **RDS Custom:** Full OS and database customization for SQL Server (and Oracle). SSH access, custom patches. For legacy apps that need it. RDS Custom for Oracle DEPRECATED reaches end of support on 31 March 2027.

Exam Tip: Multi-AZ = high availability (automatic failover). Read Replicas = read scaling (manual promotion). They serve DIFFERENT purposes and are frequently tested together. RDS Proxy is the answer for 'Lambda functions exhausting database connections.'

Amazon Aurora

AWS-built relational engine compatible with MySQL (5x faster) and PostgreSQL (3x faster). Storage auto-scales up to 128 TB. Replicates 6 copies across 3 AZs automatically.

- **Aurora Replicas:** Up to 15 read replicas with ms-level replication lag. Auto-failover to replicas (priority tiers).
- **Aurora Serverless v2:** Auto-scales capacity in fine-grained increments (0.5 ACU steps). Ideal for variable/unpredictable workloads. Does NOT scale to zero (minimum 0.5 ACU).
- **Aurora Global Database:** Cross-Region replication with <1 second lag. RPO <1s, RTO <1 minute for disaster recovery. Up to 5 secondary Regions.
- **Aurora Cloning:** Create a copy of the database using copy-on-write protocol. Fast and space-efficient. Useful for testing against production data.
- **Backtrack:** Rewind the database to a specific point in time without restoring from backup. Aurora MySQL only. Faster than traditional restore.

Amazon DynamoDB

Fully managed NoSQL key-value and document database. Single-digit millisecond performance at any scale. Serverless — no instances to manage.

Feature	Details
Capacity Modes	On-Demand: pay-per-request, auto-scales instantly. Provisioned: specify RCU/WCU, cheaper for predictable workloads. Auto Scaling available.
Partition Key	Required. Determines which partition stores the item. Choose high-cardinality keys for even distribution.
Sort Key	Optional. Combined with partition key to form a composite primary key. Enables range queries.
Global Secondary Index (GSI)	Alternate partition key + optional sort key. Eventually consistent only. Has its own provisioned throughput. Can be added after table creation.
Local Secondary Index (LSI)	Same partition key, different sort key. Must be created at table creation. Supports strong consistency. Max 5 per table.
DynamoDB Streams	Ordered log of item-level changes. Triggers Lambda functions for event-driven processing. 24-hour retention.
DynamoDB Accelerator (DAX)	In-memory cache for DynamoDB. Microsecond read latency. Transparent — no code changes.
Global Tables	Multi-Region, multi-active replication. Active-active writes in any Region. Requires Streams enabled.
TTL	Automatically deletes expired items at no cost. Great for session data, temporary records.
PartiQL	SQL-compatible query language for DynamoDB. Simplifies queries without changing your data model.

Exam Tip: 'Serverless, single-digit ms, key-value' = DynamoDB. 'Need SQL and complex joins' = RDS/Aurora. 'Microsecond reads on DynamoDB' = DAX. 'Multi-Region active-active writes' = DynamoDB Global Tables.

Amazon ElastiCache

In-memory data store for microsecond-latency caching. Two engines:

Engine	Best For	Key Differences
Redis	Complex data types, pub/sub, sorted sets, geospatial, persistence	Single-threaded. Supports replication, clustering, Multi-AZ with auto-failover. Backup/restore. Auth tokens (encryption in transit + at rest).

Engine	Best For	Key Differences
Memcached	Simple key-value caching, multithreaded performance	Multi-threaded. No replication/persistence. Simpler, pure caching layer. Supports auto-discovery.

- **Caching strategies:** Lazy loading (cache on read miss, may serve stale), Write-through (cache on write, never stale but higher write latency), TTL (set expiry to control staleness).

Exam Tip: Redis when you need persistence, replication, or advanced data structures. Memcached when you need the simplest possible cache. Both reduce database load for read-heavy workloads.

Purpose-Built Databases

Service	Type	Use Case
Amazon Redshift	Columnar data warehouse	OLAP, analytics, BI. Redshift Spectrum queries S3 data directly. Redshift Serverless for variable workloads. Concurrency Scaling for burst reads.
Amazon DocumentDB	Document DB (MongoDB compatible)	Content management, catalogs, user profiles. Managed MongoDB workloads.
Amazon Neptune	Graph database	Social networks, knowledge graphs, fraud detection, recommendation engines. Supports Gremlin and SPARQL.
Amazon Keyspaces	Wide-column (Cassandra compatible)	Migrating from Apache Cassandra. Serverless, auto-scales.
Amazon Timestream for InfluxDB	Time series database	IoT sensor data, application metrics, DevOps monitoring. Managed InfluxDB. Timestream for LiveAnalytics DEPRECATED is closed to new customers.
Amazon QLDB RETIRED	Immutable ledger	**Fully retired July 31, 2025.** AWS recommends Aurora PostgreSQL (with audit logging) as the replacement. QLDB may still appear on older SAA-C03 items as the answer for 'cryptographically verifiable ledger' — recognise it, but do not choose it for new designs.
Amazon MemoryDB for Redis	Redis-compatible durable database	Ultra-fast primary database with Redis API. Multi-AZ durability. Unlike ElastiCache, designed as a database not just a cache.

6 · High Availability & Disaster Recovery

Disaster Recovery Strategies

Four strategies, ordered from cheapest/slowest to most expensive/fastest:

Strategy	RPO	RTO	Cost	How It Works
Backup & Restore	Hours	Hours	Lowest	Regular backups to S3/Glacier. Restore from backup when disaster strikes. Slowest recovery.
Pilot Light	Minutes	Tens of minutes	Low	Core components (DB) kept running in DR Region. Other resources (EC2, ECS) are OFF but pre-configured. Scale up when activated.
Warm Standby	Seconds-Minutes	Minutes	Medium	Scaled-down but fully functional copy of production in DR Region. Scale up to production level during failover.
Active-Active (Multi-Site)	Near zero	Near zero	Highest	Full production in multiple Regions. Route 53 distributes traffic. Instant failover.

- **RPO (Recovery Point Objective):** Maximum acceptable data loss measured in time. Lower RPO = more frequent backups/replication.
- **RTO (Recovery Time Objective):** Maximum acceptable downtime. Lower RTO = faster recovery = more expensive.

Exam Tip: Match DR strategy to business RTO/RPO requirements and budget. Most exam questions describe a tolerable downtime and data loss — map those to the right strategy.

Multi-AZ Patterns

- **EC2:** Deploy across multiple AZs behind an ALB/NLB with Auto Scaling.
- **RDS:** Multi-AZ deployment for automatic failover (synchronous standby).
- **Aurora:** 6 storage copies across 3 AZs by default. Up to 15 replicas in multiple AZs.
- **EFS:** Automatically stores data redundantly across AZs.
- **S3:** Data replicated across ≥3 AZs by default (Standard class).
- **DynamoDB:** Data replicated across 3 AZs within a Region automatically.
- **ElastiCache Redis:** Multi-AZ with automatic failover to read replicas.

Cross-Region Patterns

- **Aurora Global Database** — <1s replication lag, promote secondary Region in <1 min.
- **DynamoDB Global Tables** — active-active multi-Region writes.
- **S3 Cross-Region Replication** — asynchronous object-level replication with versioning.
- **Route 53 failover** — DNS-based failover between Regions using health checks.
- **CloudFormation StackSets** — deploy infrastructure across multiple Regions/accounts simultaneously.

7 · Application Integration & Serverless

Messaging & Queuing

Service	Pattern	Key Details
Amazon SQS	Queue (point-to-point)	Standard: unlimited throughput, at-least-once delivery, best-effort ordering. FIFO: exactly-once, strict ordering, 300 msg/s (3,000 with batching). Visibility timeout prevents duplicate processing. Dead-letter queues for failed messages. Long polling reduces empty responses.
Amazon SNS	Pub/Sub (fan-out)	Topic-based. One message published to a topic reaches ALL subscribers. Subscribers: SQS, Lambda, HTTP, email, SMS, mobile push. SNS + SQS fan-out is a classic decoupling pattern. FIFO topics pair with FIFO queues.
Amazon EventBridge	Event bus	Serverless event router. Rules match events and route to targets. Schema registry. Supports third-party SaaS events. Archive and replay events. More powerful than CloudWatch Events (its predecessor).
Amazon MQ	Message broker	Managed ActiveMQ / RabbitMQ. For migrating on-premises apps that use standard protocols (AMQP, MQTT, STOMP, JMS). If building new — use SQS/SNS instead.

Exam Tip: SQS for decoupling components (buffering). SNS for fan-out to multiple consumers. SNS → SQS fan-out when each consumer needs its own queue. EventBridge for complex event routing with rules. MQ only for legacy migration.

Amazon API Gateway

Fully managed service to create, publish, and secure REST and WebSocket APIs at any scale.

- **REST API:** Full-featured (caching, request validation, WAF integration, usage plans, API keys). Regional, edge-optimized, or private.

- **HTTP API:** Simpler, faster, cheaper. Best for Lambda/HTTP proxy integrations. No caching or usage plans. Supports JWT authorizers natively.
- Integrates with Lambda (most common), HTTP backends, AWS services directly (Step Functions, SQS, DynamoDB, etc.).
- **Throttling:** 10,000 requests/second across all APIs per account (default). Configurable per stage/method.
- **Canary deployments:** route a percentage of traffic to a new stage for safe rollout.

AWS Step Functions

Serverless workflow orchestration. Visual state machine to coordinate Lambda functions, ECS tasks, human approvals, and other AWS services. Standard Workflows (up to 1 year, exactly-once) or Express Workflows (up to 5 minutes, at-least-once, high throughput).

- **States:** Task (do work), Choice (branching), Parallel, Map (iterate), Wait, Pass, Succeed, Fail.
- **Direct service integrations:** 200+ AWS services without Lambda. Reduces cost and latency.

AWS AppSync

Managed GraphQL and Pub/Sub API service. Real-time data synchronisation across web and mobile clients. Resolvers connect to DynamoDB, Lambda, RDS, HTTP, OpenSearch, and EventBridge. Offline support via client-side caching and conflict resolution.

> **Exam Tip:** 'Real-time data sync between mobile/web clients' or 'GraphQL API' = AppSync. 'REST API for Lambda' = API Gateway.

8 · Analytics & Data

Service	What It Does	When to Use
Amazon Athena	Serverless SQL queries on S3 data	Ad-hoc analysis of logs, CSV/JSON/Parquet in S3. Pay per query ($5/TB scanned). Use columnar formats (Parquet/ORC) to reduce costs. Integrates with Glue Data Catalog.
Amazon Kinesis Data Streams	Real-time data streaming	Ingest and process streaming data (clickstreams, IoT, logs) in real time. You manage shard capacity. Retention 1-365 days. On-demand mode auto-scales.
Amazon Data Firehose	Managed delivery stream	Load streaming data into S3, Redshift, OpenSearch, Splunk, HTTP endpoint. Near real-time (60s min buffer). Handles transformation via Lambda. Fully managed — no shards. (Renamed from Kinesis Data Firehose in Feb 2024.)
AWS Glue	Serverless ETL + Data Catalog	Discover, prepare, and transform data. Glue Crawlers auto-detect schema. Central metadata catalog used by Athena, Redshift Spectrum, EMR. Glue DataBrew for visual data prep.
Amazon EMR	Managed Hadoop/Spark clusters	Big data processing at scale. Apache Spark, Hive, Presto, HBase. Use Spot for cost savings on transient clusters. EMR Serverless for simpler deployments.
AWS Lake Formation	Build data lakes on S3	Centralised permissions, fine-grained access control (row/column level) over data lake. Simplifies security across Athena, Glue, EMR, Redshift.
Amazon OpenSearch Service	Search and log analytics	Full-text search, log analysis (ELK stack replacement). Pairs with Kinesis/Firehose for log ingestion. Dashboards included.
Amazon MSK	Managed Apache Kafka	Streaming data for apps already using Kafka. Alternative to Kinesis if Kafka compatibility is needed. MSK Serverless available.
Amazon Quick Sight	Serverless BI / visualisation (formerly QuickSight)	Dashboards, reports, embedded analytics. SPICE in-memory engine. Connects to RDS, S3, Athena, Redshift, and more.

Service	What It Does	When to Use
Amazon Managed Flink	Real-time stream analytics	SQL or Java/Python on streaming data from Kinesis or MSK. Complex analytics like tumbling/sliding windows, anomaly detection. Formerly Kinesis Data Analytics.

Exam Tip: 'Query S3 directly without loading data' = Athena. 'Real-time streaming ingestion' = Kinesis. 'Deliver streams to S3/Redshift with zero management' = Firehose. 'ETL jobs + schema discovery' = Glue. 'Real-time SQL on streams' = Managed Flink.

9 · Migration & Transfer

Service	What It Does
AWS DMS (Database Migration Service)	Migrate databases to AWS with minimal downtime. Supports homogeneous (MySQL→MySQL) and heterogeneous (Oracle→Aurora) migrations. Use with SCT (Schema Conversion Tool) for heterogeneous migrations. Continuous replication supported.
AWS Application Migration Service (MGN)	Lift-and-shift migration for servers. Agent-based replication to AWS. Replaces the older Server Migration Service. Automated testing and cutover.
AWS DataSync	Fast online data transfer for NFS, SMB, HDFS, S3, EFS, FSx. Agent-based, automated, scheduled. 10x faster than open-source tools. Ideal for ongoing sync or one-time migration. Recommended for new customers in place of Snow Family.
AWS Data Transfer Terminal	Physical upload centres (launched 2025, initial locations: Los Angeles, New York City). Book a time slot, bring your own storage media, upload directly to S3/EFS over a dedicated high-speed link. Modern replacement for Snowball for one-off bulk physical uploads.
AWS Transfer Family	Managed SFTP, FTPS, FTP, AS2 server backed by S3 or EFS. For partners/customers who must use file transfer protocols.
AWS Snow Family LEGACY	Physical devices for offline data transfer. Effective Nov 7, 2025, Snowball Edge is closed to new customers. Snowmobile and Snowcone retired. See Storage section for detail.
AWS Transform	Agentic AI migration and modernisation service: discovery and assessment (agentless or agent-based), dependency mapping, wave planning and migration tracking. Takes over from the two services below.
AWS Application Discovery Service DEPRECATED	Discover on-premises servers, dependencies, and utilisation for migration planning. Closed to new customers since 7 Nov 2025 → AWS Transform.
AWS Migration Hub DEPRECATED	Central dashboard to track migration progress across MGN, DMS, and other tools. Closed to new customers since 7 Nov 2025 → AWS Transform.

Migration Strategies (7 Rs)

- **Rehost (lift-and-shift):** Move as-is to EC2. Fastest, uses MGN.
- **Replatform (lift-tinker-shift):** Minor optimisations (e.g. move to RDS instead of self-managed DB).
- **Repurchase:** Move to SaaS (e.g. CRM to Salesforce).
- **Refactor / Re-architect:** Redesign using cloud-native features (Lambda, DynamoDB, containers).
- **Retire:** Decommission servers no longer needed.
- **Retain:** Keep on-premises for now (not ready to migrate).
- **Relocate:** Move VMware workloads to VMware Cloud on AWS.

Exam Tip: 'Migrate database with minimal downtime' = DMS. 'Lift and shift servers' = Application Migration Service (MGN). 'Sync NFS/SMB to S3/EFS continuously' = DataSync. 'Bulk offline upload for a new customer' = Data Transfer Terminal (or DataSync online). 'Plan a migration with dependency mapping' = AWS Transform (older questions: Application Discovery Service, now closed to new customers).

10 · Management, Monitoring & Governance

Service	Purpose
Amazon CloudWatch	Metrics, logs, alarms, dashboards. Collect and track metrics from AWS services and on-prem. CloudWatch Alarms trigger Auto Scaling, SNS notifications, or Lambda. CloudWatch Logs for centralised log analysis. Logs Insights for SQL-like queries. Contributor Insights for top-N analysis.
AWS CloudTrail	API activity logging. Records every API call (who, what, when, from where). Management events on by default. Data events (S3 object-level, Lambda invocations) optional. Store in S3, analyse with Athena, or send to CloudWatch Logs for Logs Insights queries. CloudTrail Lake DEPRECATED is closed to new customers (31 May 2026) → CloudWatch.
AWS Config	Resource configuration tracking and compliance. Records config changes over time. Config Rules evaluate compliance (e.g. 'all EBS volumes must be encrypted'). Remediation actions via SSM Automation. Conformance packs for standards (PCI, HIPAA).
AWS CloudFormation	Infrastructure as Code (IaC). Define AWS resources in JSON/YAML templates. Stack = collection of resources. Change sets preview changes. Drift detection. StackSets for multi-account/Region deployments. Nested stacks for modularity.
AWS Organizations	Multi-account management. Consolidated billing across accounts. SCPs to restrict permissions org-wide. OUs (Organizational Units) for hierarchical structure. Tag policies and backup policies.
AWS Control Tower	Automated multi-account setup with guardrails. Landing Zone with pre-configured accounts, SSO, logging. Preventive guardrails (SCPs) and detective guardrails (Config rules). Account Factory for standardised account provisioning.
AWS Systems Manager	Operational management. Session Manager for SSH-less access to EC2 (no open port 22). Patch Manager for automated patching. Parameter Store for config. Run Command for remote execution. Automation for runbooks. Inventory for software/config tracking.
AWS Trusted Advisor	Best practice recommendations across cost, performance, security, fault tolerance, service limits. Full checks with Business/Enterprise Support plans.
AWS Compute Optimizer	ML-based recommendations for right-sizing EC2, EBS, Lambda, Fargate, and Auto Scaling groups.

Exam Tip: CloudTrail = WHO did WHAT (API audit). CloudWatch = HOW are my resources PERFORMING (metrics/logs). Config = WHAT changed in my resource CONFIGURATION over time. These three are frequently tested together.

AWS X-Ray

Distributed tracing service for analysing and debugging microservices and serverless applications. Traces requests as they travel through your application — from API Gateway through Lambda, ECS, EC2, to DynamoDB and other downstream services. Generates a visual service map showing latency, errors, and bottlenecks.

Exam Tip: 'Debug latency issues in a distributed microservices application' or 'find which service is causing slow responses' = X-Ray.

Developer Tools

Service	What It Does
AWS CodeCommit	Managed private Git repository with IAM-based access. *Note:* AWS closed CodeCommit to new customers in July 2024, then **returned it to full General Availability** on November 24, 2025 after customer feedback — new signups are again accepted.
AWS CodeBuild	Managed build service. Compiles code, runs tests, produces deployment-ready artifacts. Pay per build minute. Uses buildspec.yml.
AWS CodeDeploy	Automated deployment to EC2, Lambda, or ECS. Supports in-place (EC2), blue/green (EC2, ECS, Lambda), canary, and linear deployment strategies. Uses appspec.yml.
AWS CodePipeline	Continuous delivery pipeline. Orchestrates source, build, test, and deploy stages. Integrates with CodeCommit, CodeBuild, CodeDeploy, and third-party tools.

Exam Tip: CodePipeline orchestrates the full CI/CD pipeline. CodeBuild compiles and tests. CodeDeploy handles deployment strategies (blue/green, canary). Know the difference.

11 · Cost Optimisation

Cost Management Tools

Tool	What It Does
AWS Cost Explorer	Visualise, understand, and forecast your AWS spending. Filter by service, tag, account. Shows Reserved Instance and Savings Plan recommendations.
AWS Budgets	Set custom cost and usage thresholds. Alerts via email/SNS when you exceed or forecast to exceed. Can trigger actions (e.g. apply SCP to stop provisioning).
AWS Cost and Usage Report (CUR)	Most granular billing dataset. Hourly line items by resource. Export to S3 for analysis with Athena / Quick Sight.
Cost Allocation Tags	Tag resources to attribute costs to projects, teams, or environments. AWS-generated tags (aws:createdBy) and user-defined tags. Must be activated in billing console.
AWS Pricing Calculator	Estimate the cost of AWS services before you deploy. Model architectures and compare options.

Key Cost Optimisation Strategies

- **Right-sizing:** Use Compute Optimizer and Cost Explorer to identify over-provisioned instances. Downsize or switch families.
- **Purchasing options:** Reserved Instances or Savings Plans for predictable workloads (up to 72% savings). Spot Instances for fault-tolerant batch jobs (up to 90%).
- **S3 lifecycle policies:** Move infrequently accessed data to IA or Glacier tiers. Intelligent-Tiering for unknown patterns.
- **Serverless:** Lambda, Fargate, Aurora Serverless — pay only for what you use. Eliminates idle capacity.
- **Data transfer:** Keep traffic within the same AZ when possible (free). Use VPC endpoints to avoid NAT Gateway data processing charges. CloudFront reduces origin egress.
- **Auto Scaling:** Scale in during low demand to avoid paying for idle resources.
- **Consolidation:** AWS Organizations for consolidated billing. Volume discounts apply across all accounts.
- **Graviton instances:** ARM-based processors (M7g, C7g, R7g). Up to 40% better price-performance than x86. Compatible with most Linux workloads.

Exam Tip: Most cost questions have a 'MOST cost-effective' constraint. Check: Can you use Spot? Can you use Savings Plans? Can you move to serverless? Can you use a cheaper storage class? Can you reduce data transfer costs with VPC endpoints? Can you use Graviton?

12 · Well-Architected Framework — Quick Reference

Pillar	Focus	Key Services
Operational Excellence	Run and monitor systems, continuously improve processes	CloudFormation, Config, CloudWatch, Systems Manager, X-Ray, CodePipeline
Security	Protect information, systems, and assets	IAM, KMS, WAF, Shield, GuardDuty, CloudTrail, VPC, Security Hub, Macie
Reliability	Recover from failure, meet demand	Auto Scaling, Multi-AZ, Route 53, S3, Backup, CloudWatch Alarms, Aurora Global DB
Performance Efficiency	Use resources efficiently as demand changes	Auto Scaling, Lambda, CloudFront, ElastiCache, Aurora, Global Accelerator, Graviton
Cost Optimization	Avoid unnecessary costs	Cost Explorer, Budgets, Savings Plans, Spot, S3 tiers, Compute Optimizer, Graviton
Sustainability	Minimise environmental impact	Graviton instances, serverless, right-sizing, managed services, Region selection

13 · Additional Services — Quick Reference

Service	Category	What It Does
AWS Amplify	Front-End	Build, deploy, and host full-stack web and mobile apps. Git-based CI/CD. Managed hosting with CloudFront.
Amazon AppFlow	Integration	No-code integration between SaaS apps (Salesforce, SAP, Slack) and AWS services (S3, Redshift). Scheduled or event-triggered flows.
AWS CloudShell	Developer	Browser-based shell with AWS CLI pre-configured. Free, no setup. Useful for quick CLI tasks.
Amazon SES	Messaging	Scalable email sending service. Transactional emails, marketing campaigns. Pay per email sent.
Amazon Pinpoint RETIRED	Messaging	Multi-channel engagement campaigns. End of support 30 Oct 2026 → Amazon Connect outbound campaigns; SMS, voice and push via AWS End User Messaging; email via SES.
AWS Wavelength	Edge	Run workloads on 5G carrier networks for ultra-low latency to mobile devices. Deploy EC2 in Wavelength Zones.
AWS Outposts	Hybrid	Run AWS infrastructure on-premises. Fully managed racks with EC2, EBS, S3, RDS, ECS, EKS. For data residency or local processing needs.
AWS Local Zones	Edge	Extend a Region to a metro area for single-digit-ms latency. Like a mini-AZ close to users.
Amazon Textract	AI/ML	Extract text, tables, and forms from scanned documents. Goes beyond simple OCR.
Amazon Rekognition	AI/ML	Image and video analysis: object detection, facial analysis, content moderation.
Amazon Transcribe	AI/ML	Speech-to-text. Real-time and batch. Medical-specific model available.
Amazon Comprehend	AI/ML	NLP: sentiment analysis, entity recognition, PII detection, language detection.
Amazon Translate	AI/ML	Neural machine translation between 75+ languages.
Amazon Polly	AI/ML	Text-to-speech with natural-sounding voices. SSML support.

Service	Category	What It Does
Amazon Lex	AI/ML	Build conversational bots (same tech as Alexa). Integrates with Lambda for fulfilment.
Amazon Kendra DEPRECATED	AI/ML	Intelligent enterprise search powered by ML. Closed to new customers → Amazon Bedrock Knowledge Bases.
Amazon SageMaker	AI/ML	Build, train, and deploy ML models at scale. JumpStart for pre-trained models. Studio for visual ML development.
Amazon Bedrock	AI/ML	Access foundation models (Claude, Titan, Llama) via API. Fine-tune with your data. Guardrails for safety.
Amazon SageMaker AI	AI/ML	Build, train, and deploy ML models at scale. JumpStart for pre-trained models. Studio for visual ML development.
Amazon Quick	Analytics	AI workspace over company data that includes Quick Sight BI: dashboards, natural-language Q&A, embedded analytics, SPICE in-memory engine.
AWS Data Exchange	Analytics	Find, subscribe to, and use third-party data sets (financial, healthcare, weather, etc.) directly in S3.
Amazon Kinesis Video Streams	Media	Ingest and process video from connected devices. Integrate with Rekognition Video for real-time analysis.
Amazon Elastic Transcoder RETIRED	Media	Media transcoding. Shut down 13 Nov 2025 → AWS Elemental MediaConvert.
Amazon ECS Anywhere	Containers	Run ECS tasks on your own on-prem or other-cloud servers using the same ECS control plane.
Amazon EKS Anywhere	Containers	Deploy and operate Kubernetes clusters on your own infrastructure (vSphere, bare metal) with AWS-supported tooling.
Amazon EKS Distro	Containers	Open-source Kubernetes distribution used by EKS. Build your own custom Kubernetes platform with the same components AWS runs.
AWS Device Farm	Front-end	Test mobile and web apps on real devices in the AWS cloud. Continuous test runs, automation, remote interactive sessions.
AWS Serverless Application Repository	Compute	Publish and discover reusable serverless applications (SAM templates). Deploy into your account with one click.
AWS Management Console	Management	Browser-based UI for AWS. Mobile app available. Switch account/Region via header; federated SSO via IAM Identity Center.
AWS Health Dashboard	Management	Service Health (global) and Personal Health (account-specific) events. Pipe into ops via EventBridge.
AWS Well-Architected Tool	Management	Self-assessment against the six Well-Architected pillars. Generates improvement plans you can track over time.
Amazon Managed Grafana	Observability	Fully managed Grafana. Data sources include CloudWatch, Prometheus, X-Ray, OpenSearch, third-party. IAM Identity Center / SAML auth.
Amazon Managed Service for Prometheus	Observability	Fully managed Prometheus-compatible TSDB. PromQL queries. Ingest via `remote_write` or the EKS scraper.
AWS Network Firewall	Security	Stateful deep-packet-inspection firewall for VPCs. Suricata-compatible rules. Deploy in a central inspection VPC and route spokes through it.
AWS License Manager	Management	Track and enforce software licences (BYOL) across AWS and on-prem. License rules prevent over-consumption.
AWS Service Catalog	Management	Curated portfolios of approved CloudFormation-backed products end-users can self-service. Admins control what's offered and who can launch it.

14 · Retired or Closed to New Customers

Older questions and study material still name these. Know what they did; for a new design, pick the replacement.

Service	Status → What to use instead
Amazon Kendra DEPRECATED	Closed to new customers → Amazon Bedrock Knowledge Bases.
AWS Snow Family LEGACY	Snowball Edge closed to new customers (7 Nov 2025); Snowcone and Snowmobile retired → DataSync online, AWS Data Transfer Terminal.
S3 Select · Glacier Select DEPRECATED	Closed to new customers since 25 Jul 2024 → Amazon Athena on S3.
Amazon QLDB RETIRED	Shut down 31 Jul 2025 → Aurora PostgreSQL with audit logging.
Timestream for LiveAnalytics DEPRECATED	Closed to new customers → Timestream for InfluxDB.
RDS Custom for Oracle DEPRECATED	End of support 31 Mar 2027 → RDS for Oracle, or Oracle on EC2.
Simple AD DEPRECATED	Closed to new customers → AWS Managed Microsoft AD, or AD Connector.
CloudTrail Lake DEPRECATED	Closed to new customers since 31 May 2026 → CloudWatch (Logs Insights), or trail logs in S3 queried with Athena. CloudTrail itself is current.
Application Discovery Service · Migration Hub DEPRECATED	Closed to new customers since 7 Nov 2025 → AWS Transform.
Amazon Pinpoint RETIRED	End of support 30 Oct 2026 → Amazon Connect outbound campaigns, AWS End User Messaging, SES.
Amazon Elastic Transcoder RETIRED	Shut down 13 Nov 2025 → AWS Elemental MediaConvert.
FSx File Gateway DEPRECATED	Closed to new customers (Oct 2024) → FSx for Windows File Server accessed directly.
AWS Elemental MediaStore RETIRED	Shut down 12 Nov 2025 → S3 or MediaPackage as the CloudFront origin.
Kinesis Data Analytics for SQL RETIRED	Shut down 27 Jan 2026 → Amazon Managed Service for Apache Flink.

cloudninja.pro — Free AWS certification practice exams, study guides, and cheat sheets.
Pair this cheat sheet with our free SAA-C03 practice exam for the best results.