AWS Solutions Architect Associate Cheat Sheet

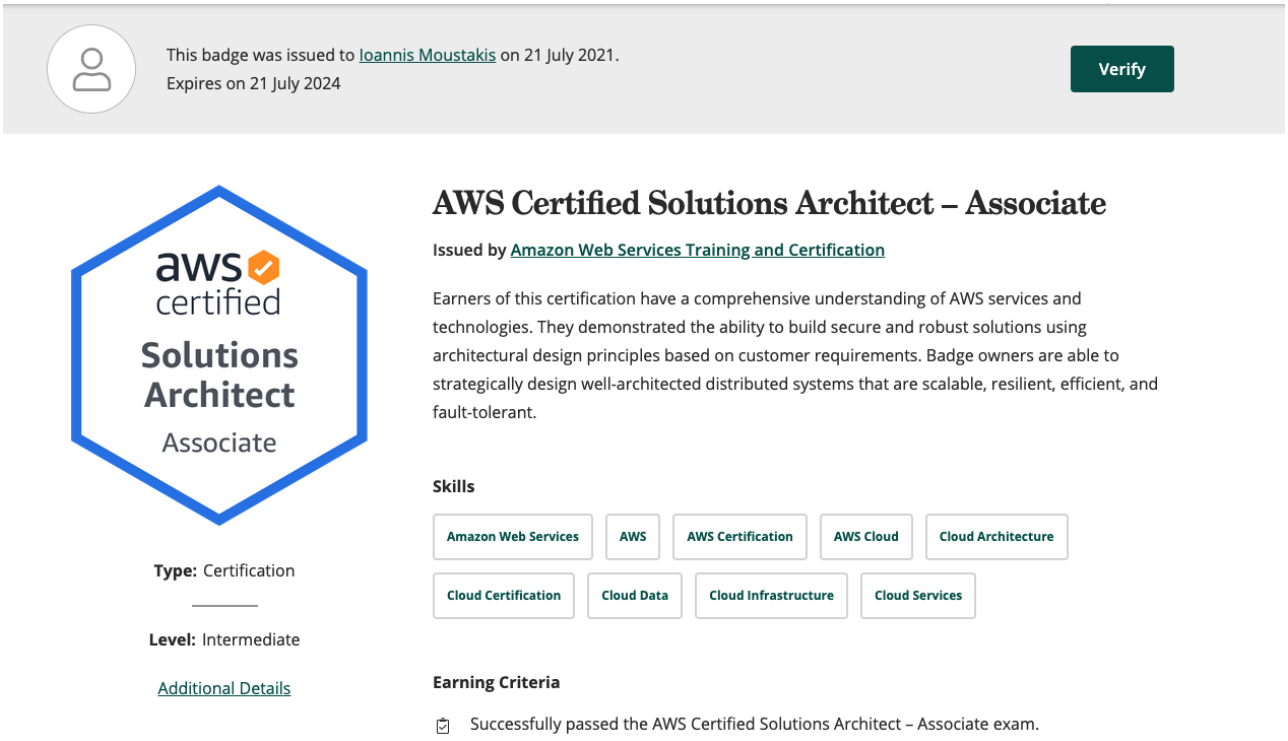
AWS Solutions Architect Associate study notes, exam pointers, and cheat sheet.

Published Sun, Jul 25, 2021 by Ioannis Moustakis

Estimated reading time: 34 min

This material was gathered during my preparation for the `AWS Solutions Architect Associate Exam`. I created and curated this cheatsheet with useful information that will be handy to review before taking the exam.

Gathered all the topics and details that I struggled with and I believe this cheatsheet greatly helped me pass this certification and get my badge.



Note that the most important thing to pass this exam is to `get a good overall understanding of the basic AWS services` first. Use these notes as `complementary material and not complete study` material for the exam.

This cheat sheet doesn't include all the information you will need to pass the exam, `easy or obvious topics are skipped`. It is rather focused on gathering `tricky, hidden & more difficult` information and questions.

AWS frequently changes information, configuration, and options of different services so `some of the content might become outdated at some point`. Make sure to cross-check and validate the information you are getting from online sources with the `official AWS Documentation and FAQs` before your exam.

OK enough with the disclaimers, let's get to it.

EC2

- Dedicated (Instances): No other customers will share the hardware. May share hardware with other instances of ONLY your account.
- (Dedicated) Hosts: Book an entire physical server and have full control of EC2 instance placement.
- You can only change the tenancy of an instance from dedicated to host, or from host to dedicated after you've launched it.
- Good EC2 combo -> reserved instances for baseline + on-demand & spot for peaks.

Userdata

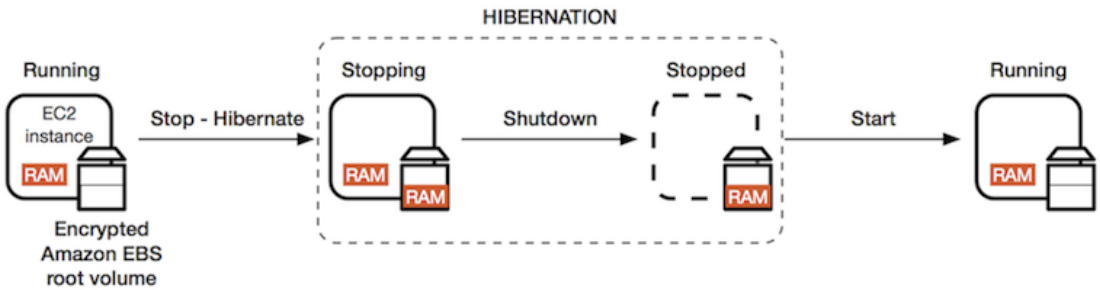
- Executed as `root` by default.

Hibernate

- Hibernation saves the contents from the instance memory (RAM) to your Amazon EBS root volume. When you start your instance: The Amazon EBS root volume is restored to its previous state The RAM contents are reloaded.

Overview of hibernation

The following diagram shows a basic overview of the hibernation process.



- To use hibernation, the root volume must be an encrypted EBS volume.
- When the instance state is stopping, you will not be billed if it is preparing to stop however, you will still be billed if it is just preparing to hibernate.

Spot instances

- A Spot Instance request is either one-time or persistent. If the spot request is persistent, the request is opened again after your Spot Instance is interrupted.

- Spot blocks are Spot Instances with a `defined duration & are designed not to be interrupted`.
- If your Spot Instance request is disabled and has an associated stopped Spot Instance, `canceling the request does not terminate the instance`.

Placement groups

- It is recommended that you launch the number of instances that you need in the placement group in a single launch request and that you use the same instance type for all instances in the placement group. If you try to add more instances to the placement group later, or if you try to launch more than one instance type in the placement group, you increase your chances of getting an insufficient capacity error.
- If you receive a capacity error when launching an instance in a placement group that already has running instances, stop and start all of the instances in the placement group, and try the launch again. Restarting the instances may migrate them to hardware that has the capacity for all the requested instances.

Spread

- Maximum of 7 running instances per Availability Zone per group.
- Recommended for applications that have a `small number of critical instances that should be kept separate from each other`.
- Spread placement groups provide access to `distinct racks`, and are therefore suitable for mixing instance types or launching instances over time.

Cluster

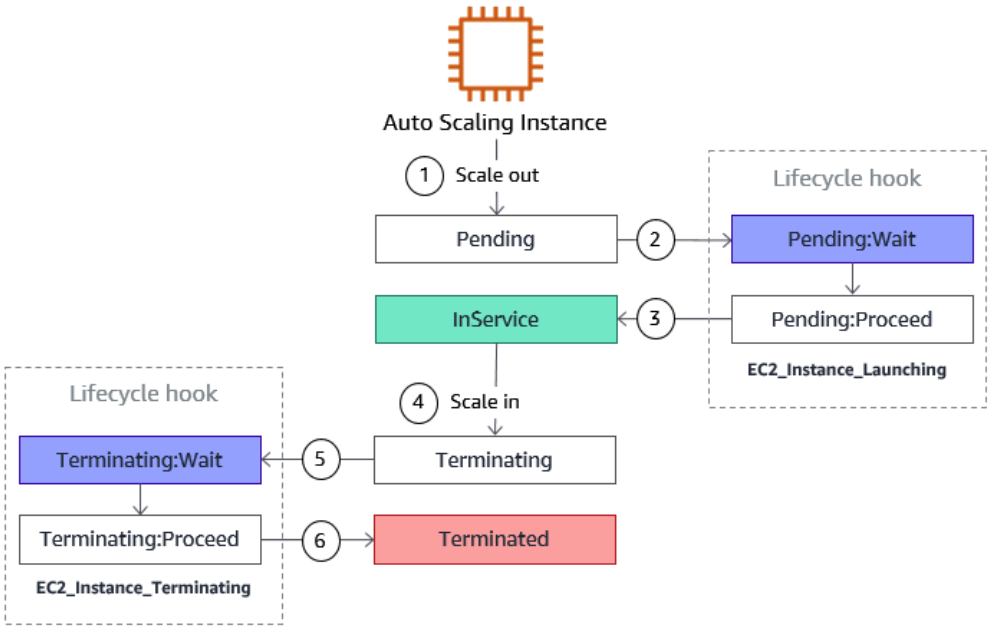
- Higher per-flow throughput limit of up to 10 Gbps for TCP/IP traffic and are placed in the same high-bisection bandwidth segment of the network.

Partition

- Spreads your instances across logical partitions such that groups of `instances in one partition do not share the underlying hardware with groups of instances in different partitions`.
- Used by `large distributed and replicated workloads`.

Autoscaling

- Lifecycle hooks enable you to perform custom actions as the Auto Scaling group launches or terminates instances.



- Lifecycle hooks `put the instance into wait state until the script or timeout period ends`.
- With launch templates, you can provision capacity across multiple instance types using both On-Demand Instances and Spot Instances.
- You can put an instance that is in the InService state into the Standby state, update some software or troubleshoot the instance, and then return the instance to service.
- Auto Scaling doesn't terminate an instance that came into service based on EC2 status checks and ELB health checks `until the health check grace period expires`.
- Cooldown period: It ensures that the Auto Scaling group does not launch or terminate additional EC2 instances before the previous scaling activity takes effect(default 300s).
- Amazon EC2 Auto Scaling `does not immediately terminate instances with an Impaired status`.
- By default, Amazon EC2 Auto Scaling doesn't use the results of ELB health checks to determine an instance's health status when the group's health check configuration is set to EC2.

- When there are multiple policies in force at the same time, Auto Scaling `chooses the policy that provides the largest capacity for both scale-out and scale-in`.
- The default value for the instance placement tenancy is null and the instance tenancy is controlled by the tenancy attribute of the VPC. If you set the Launch Configuration Tenancy to default and the VPC Tenancy is set to dedicated, then the instances have dedicated tenancy. If you set the Launch Configuration Tenancy to dedicated and the VPC Tenancy is set to default, then again the instances have dedicated tenancy.
- If you have an EC2 Auto Scaling group (ASG) with running instances and you `choose to delete the ASG, the instances will be terminated and the ASG will be deleted`.
- Rebalancing AZs launches new instances before terminating the old ones.
- Auto Scaling creates a new scaling activity for terminating the unhealthy instance and then terminates it. Later, another scaling activity launches a new instance to replace the terminated instance.

S3

- S3 standard: There is no minimum storage duration charge and no retrieval fee (use case: if you want to keep data for a few days only)
- Object-level permissions: `For actions inside the bucket` (e.g. GetObject), add `/*` after arn, -> arn:aws:s3:::test/*
- With bucket policies, you can grant users within your AWS Account or other AWS Accounts access to your Amazon S3 resources.
- The AWS S3 sync command uses the CopyObject APIs to copy objects between S3 buckets.
- By default, S3 replication only supports copying new Amazon S3 objects after it is enabled.
- `Max upload 5GB per time`, for more use multi-part upload. If the object to upload is `> 100 MB`, you should `consider using multipart uploads`.
- Amazon S3 delivers `strong read-after-write consistency` automatically.
- You can increase your read or write performance by parallelizing reads with prefixes.
- Once you version-enable a bucket, it can never return to an unversioned state. Versioning can only be suspended once it has been

enabled.

- No S3 data transfer charges `when data is transferred in` from the internet.
- Also with S3TA, `you pay only for transfers that are accelerated`.
- Using the Range HTTP header in a GET Object request, you can fetch a byte-range from an object, transferring only the specified portion. `A byte-range request is a perfect way to get the beginning of a file`.
- You can place a retention period on an object version. Different versions of a single object can have different retention modes and periods.
- `Max object size 5TB`.
- For replication must enable versioning in source and destination.
- By default, an S3 object is owned by the AWS account that uploaded it, even in a bucket in a different account. To get full access to the object, the object owner must explicitly grant the bucket owner access. You can create a bucket policy to require external users to grant bucket-owner-full-control when uploading objects so the bucket owner can have full access to the objects.
- Object lock: store objects as locked(only on versioned buckets).
- `Metadata`, which can be included with the object, `is not encrypted` while being stored on Amazon S3. Therefore, AWS recommends that customers not place sensitive information in Amazon S3 metadata.
- S3 event notification allows destinations: `SQS standard, Lambda, SNS`.
- Allowed names for S3 website endpoints: http://bucket-name.s3-website.Region.amazonaws.com & http://bucket-name.s3-website-Region.amazonaws.com
- S3 Select scan a subset of an object by specifying a range of bytes to query based on the bucket's name and the object's key.
- With S3 Select, you can use simple structured query language (SQL) statements to filter the contents of an Amazon S3 object and retrieve just the subset of data that you need. CSV, JSON, or Apache Parquet format.
- S3 can publish notifications for the following events: `New object-created events, Object removal events, Restore object events, Reduced Redundancy Storage (RRS) object lost events, Replication events`.

- `To encrypt an object at the time of upload`, you need to add a header called `x-amz-server-side-encryption`. To enforce object encryption, create an S3 bucket policy that denies any S3 Put request that does not include the x-amz-server-side-encryption header.
- To enable S3 website: a) An S3 bucket that is configured to host a static website. The bucket must have the same name as your domain or subdomain b) a registered domain name c) Route 53 as the DNS service for the domain.
- S3 server access logs provide detailed records for the requests that are made to an S3 bucket.
- 3,500 requests per second to add data and 5,500 requests per second to retrieve data.
- You can have an S3 bucket that has different objects stored in S3 Standard, S3 Intelligent-Tiering, S3 Standard-IA, and S3 One Zone-IA.

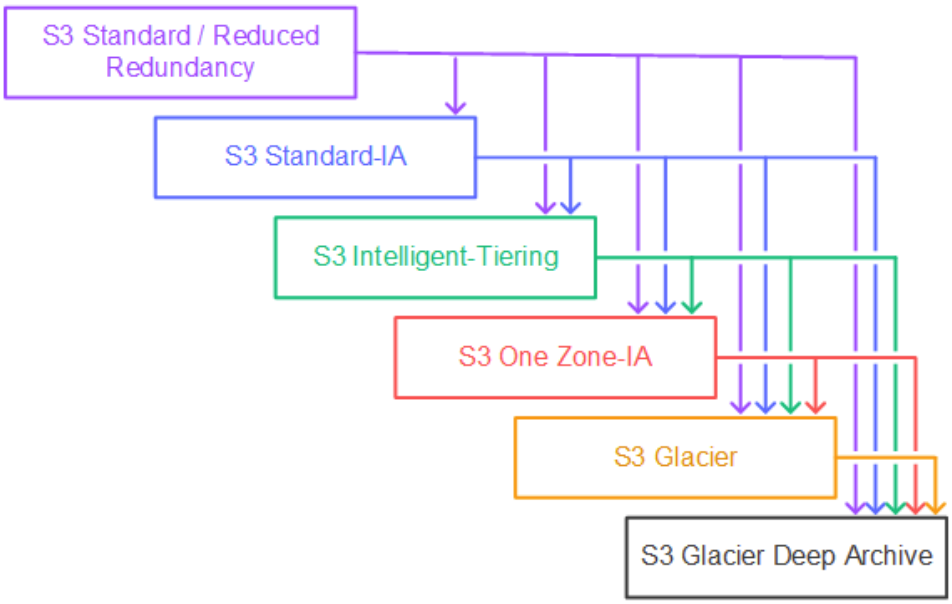
S3 IA

- S3 One Zone-IA is for data that is accessed less frequently but requires rapid access when needed.
- The `minimum storage duration is 30 days` before you can transition objects from S3 Standard to S3 Standard IA or One Zone-IA.(This limitation does not apply to Intelligent Tiering, Glacier, and Glacier Deep Archive)

S3 Lifecycle Transitions

Supported lifecycle transitions - `waterfall model` :

- The S3 Standard storage class to any other storage class.
- Any storage class to the S3 Glacier or S3 Glacier Deep Archive storage classes.
- The S3 Standard-IA storage class to the S3 Intelligent-Tiering or S3 One Zone-IA storage classes.
- The S3 Intelligent-Tiering storage class to the S3 One Zone-IA storage class.
- The S3 Glacier storage class to the S3 Glacier Deep Archive storage class.



- `Encrypted objects remain encrypted throughout the storage class transition process`.

Glacier

- Glacier supports encryption by default for both data at rest as well as in-transit.
- The minimal storage duration period is `90 days for the S3 Glacier` storage class and `180 days for S3 Glacier Deep Archive`.
- Data can be stored directly in Amazon S3 Glacier Deep Archive.

Snowball

- Snowball Edge storage optimised: 80TB 40 vCPUs, 1 TB of SATA SSD storage, and up to 40 Gb network connectivity.
- You can't directly copy data from Snowball Edge devices into AWS Glacier.
- `For data < 10PB or distributed in multiple locations`.
- Snowball Edge compute optimised(52 vCPUs, 42 TB of usable block or object storage, and an optional GPU).
- Snowball Edge possibility for storage clustering.
- AWS OpsHub is a graphical user interface you can use to manage your AWS Snowball devices.

Snowmobile

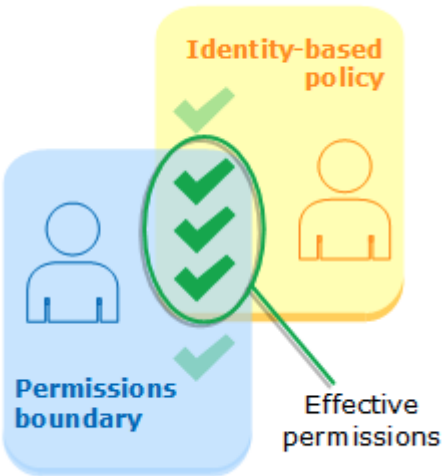
- Each Snowmobile has a total capacity of up to 100 petabytes.

- `For data > 10PB in a single location`.

	AWS Snowcone	AWS Snowball Edge Storage Optimized	AWS Snowball Edge Compute Optimized	AWS Snowmobile
Usage Scenario	Edge computing, Data transfer, Edge storage	Data transfer, Edge storage	Edge computing, Data transfer	Data transfer
Usable HDD Storage	8 TB	80 TB	42 TB	100 PB
Usable SSD Storage	No	1 TB	7.68 TB	No
Usable vCPUs	2 vCPUs	40 vCPUs	52 vCPUs	N/A
Usable Memory	4 GB	80 GB	208 GB	N/A
GPU	No	No	nVidia V100 (optional)	No
Onboard Computing Options	AWS IoT Greengrass Amazon EC2 AMIs	AWS IoT Greengrass Amazon EC2 AMIs	AWS IoT Greengrass Amazon EC2 AMIs	N/A
DataSync	Yes	No	No	No
Transfers via NFS	Yes	Yes	Yes	Yes
Transfers via S3 API	No	Yes	Yes	No
Network Interfaces	2x 1/10 Gbit - RJ45	2x 10 Gbit – RJ45 1x 25 Gbit – SFP+ 1x 100 Gbit – QSFP28	2x 10 Gbit – RJ45 1x 25 Gbit – SFP+ 1x 100 Gbit – QSFP28	6x 40 Gbit
Device Size	9 inches long, 6 inches wide, and 3 inches tall (227 mm x 148.6 mm x 82.65 mm)	28.3 inches long, 10.6 inches wide, and 15.5 inches tall (548 mm x 320 mm x 501 mm)	28.3 inches long, 10.6 inches wide, and 15.5 inches tall (548 mm x 320 mm x 501 mm)	N/A
Device Weight	4.5 lbs. (2.1 kg)	49.7 lbs. (22.3 kg)	49.7 lbs. (22.3 kg)	N/A
Encryption	Yes, 256-bit	Yes, 256-bit	Yes, 256-bit	Yes, 256-bit
Portability	Battery-based Operation	No	No	No
Wireless	Wi-Fi	No	No	No
Storage Clustering	No	Yes, 5-10 nodes	Yes, 5-10 nodes	N/A

IAM

- Permissions Boundary to `limit max access of users`. They can only be applied to roles or users, not IAM groups.



- IAM Policy Evaluation Logic: if there is an explicit deny, the final decision is deny for the resource.
- When you assume a role, you give up your original permissions and take the permissions of the assigned role.
- When using a resource-based policy the principal doesn't have to give up his permissions.
- In a policy condition: aws:RequestedRegion represents the target of the API call.
- You can share an AMI with another account.
- Trust Policy: only IAM resource-based policy.
- If you got your certificate from a third-party CA, import the certificate into ACM or upload it to the IAM certificate store.
- With `web identity federation` , you don't need to create custom sign-in code or manage your own user identities. Instead, users of your app can sign in using a well-known external identity provider (IdP), such as Login with Amazon, Facebook, Google, or any other OpenID Connect (OIDC)-compatible IdP.

Security Token Service(STS)

- Temporary security credentials that can control access to your AWS resources.

AWS Organizations

- It does not offer federation capability.
- To migrate an account to another Organization: remove member account, send an invite to new Org, Accept the invite to the new Org from the member account.
- SCPs offer central control over the maximum available permissions for all accounts in your organization, allowing you to ensure your accounts stay within your organization's access control guidelines.
- SCPs affect all users and roles in the attached accounts, including the root user.
- SCPs do not affect any service-linked role.

VPC

- VPN connection: Virtual Private Gateway endpoint on the AWS VPC side - Customer Gateway on the on-premises side.

- You can't have a VPC with only a public subnet and AWS Site-to-Site VPN.
- Private IPs allowed ranges: 10.0.0.0/8 (10.0.0.0 - 10.255.255.255), 172.16.0.0/12 (172.16.0.0 - 172.31.255.255), 192.168.0.0/16(192.168.0.0 - 192.168.255.255)
- AWS `reserves 5 Ip addresses in each subnet`.
- Shared services VPC, which provides access to services required by workloads in each of the VPCs. This might include directory services or VPC endpoints. Sharing resources from a central location instead of building them in each VPC may reduce administrative overhead and cost.
- Use AZ ID to uniquely identify the Availability Zones across the two AWS Accounts.
- By default, non-default subnets have the `IPv4 public addressing(assign public IP) attribute set to false`, and default subnets have this attribute set to true.
- You cannot disable IPv4 support for your VPC and subnets since this is the default IP addressing system for Amazon VPC and Amazon EC2.
- Every subnet that you create is automatically associated with the main route table for the VPC.
- Allowed block size in VPC is between a /16 netmask (65,536 IP addresses) and /28 netmask.
- While primary ENIs cannot be detached from an instance, secondary ENIs can be detached and attached to a different instance.

Security Groups

- If nothing is defined in a security group then all access is blocked.

NACL

- NACLs are stateless so outbound rules have to be evaluated again.
- Defined at Subnet level.
- Should allow outbound traffic from ephemeral ports.
- NACL rules are evaluated starting with the lowest numbered rule. As soon as a rule matches traffic, it's applied immediately regardless of any higher-numbered rule that may contradict it.

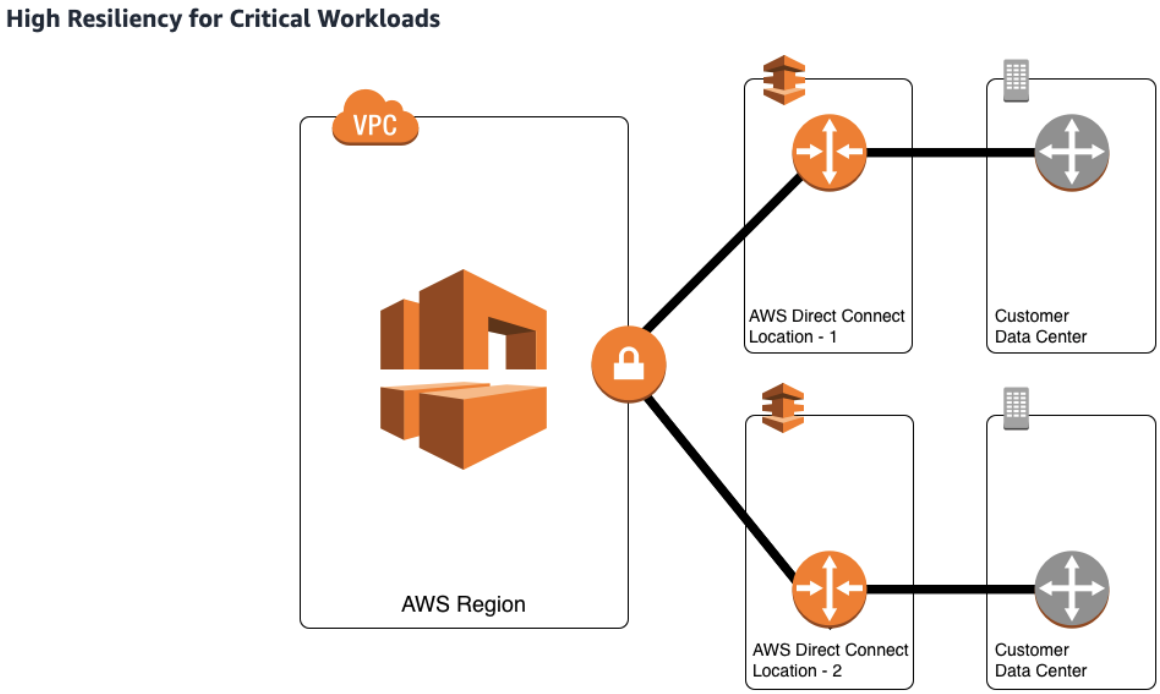
Cloudhub

- Multiple AWS Site-to-Site VPN connections, you can provide secure communication between sites using the AWS VPN CloudHub including

Direct Connect connections.
- Supports IP Multicast.
- Low-cost primary or secondary network connectivity between locations, only for VPNs.

Direct Connect

- Maximum resilience is achieved by separate connections terminating on separate devices in more than one location.



- Dedicated connection 1-10 Gbps.
- Hosted connection 50Mbps -10Gbps, add or remove capacity on demand.
- `Data in transit not encrypted` , but private.

Transit Gateway

- Network transit hub that you can use to interconnect your virtual private clouds (VPC) and on-premises networks.
- AWS Transit Gateway also enables you to scale the IPsec VPN throughput with equal-cost multi-path (ECMP) routing support over multiple VPN tunnels.

NAT Instance

- Can be used as a bastion, supports security groups, supports port-forwarding, must disable ec2 flag source/destination check.

Nat Gateway

- Only for IPv4.
- Set up in a public subnet.
- In a specific AZ and can only be used by instances in other subnets.

Egress-only Internet Gateway:

- nat for ipv6.

Route53

- Routing policy multi-value supports up to 8 healthy records for each multi-value query.
- To integrate an external domain to route53, update the nameservers on the 3rd party registrar with your public hosted zone.
- To resolve any DNS queries `for resources in the AWS VPC from the on-premises network`, you can create an `inbound endpoint` on Route 53 Resolver, and then DNS resolvers on the on-premises network can forward DNS queries to Route 53 Resolver via this endpoint.
- To resolve DNS queries `for any resources in the on-premises network from the AWS VPC`, you can create an `outbound endpoint` on Route 53 Resolver, and then Route 53 Resolver can conditionally forward queries to resolvers on the on-premises network via this endpoint.
- Cannot create a CNAME record for the top node of the DNS namespace. So, if you register the DNS name mpla.com the zone apex is mpla.com You can't create a CNAME record for mpla.com but you can create an alias record for mpla.com that routes traffic to www.mpla.com.
- Route 53 `doesn't charge for alias queries` to AWS resources but Route 53 does `charge for CNAME queries`.
- For each VPC that you want to associate with the Route 53 hosted zone, change the following VPC settings to true: `enableDnsHostnames`, `enableDnsSupport`.
- You configure active-active failover using any routing policy (or combination of routing policies) other than failover, and you configure active-passive failover using the failover routing policy.

- `Active-Active Failover` when you want all of your resources to be available the majority of the time.
- `Active-Passive Failover` when you want a primary resource or group of resources to be available the majority of the time and you want a secondary resource or group of resources to be on standby in case all the primary resources become unavailable.

EBS

- By default, the root volume for an AMI backed by Amazon EBS is deleted when the instance terminates.
- For an encrypted EBS volume data stored at rest on the volume, data moving between the volume and the instance, snapshots created from the volume, and volumes created from those snapshots are all encrypted.
- GP2: system boot volumes, 1GB - 16TB, max IOPS 16,000, if you add 1TB you get +3000IOPS, for low latency interactive apps.
- io1/io2: 4GB-16TB, max 64,000 IOPS, `50:1` IOPS:GB ratio.
- io2 Block Express volumes, Provisioned IOPS (PIOPS) up to `256,000`, with an IOPS:GiB ratio of `1,000:1`, for submillisecond latency for > 64,000 IOPS or 1000 MB/s throughput.
- Throughput optimised HDD(st1): `max throughput 500 MB/s – max 500 IOPS`, Big data, log processing, data warehouses.
- Cold HDD(scl): max throughput 250 MB/s - max 250 IOPS, throughput-oriented storage that is infrequently accessed, low storage cost scenarios.
- Amazon EBS Multi-Attach enables you to attach a single Provisioned IOPS SSD (io1 or io2) volume to multiple instances with Nitro system that are in the same Availability Zone
- Throughput Optimized HDD (`st1`) and Cold HDD (`sc1`) volume types `cannot be used for boot volumes`.

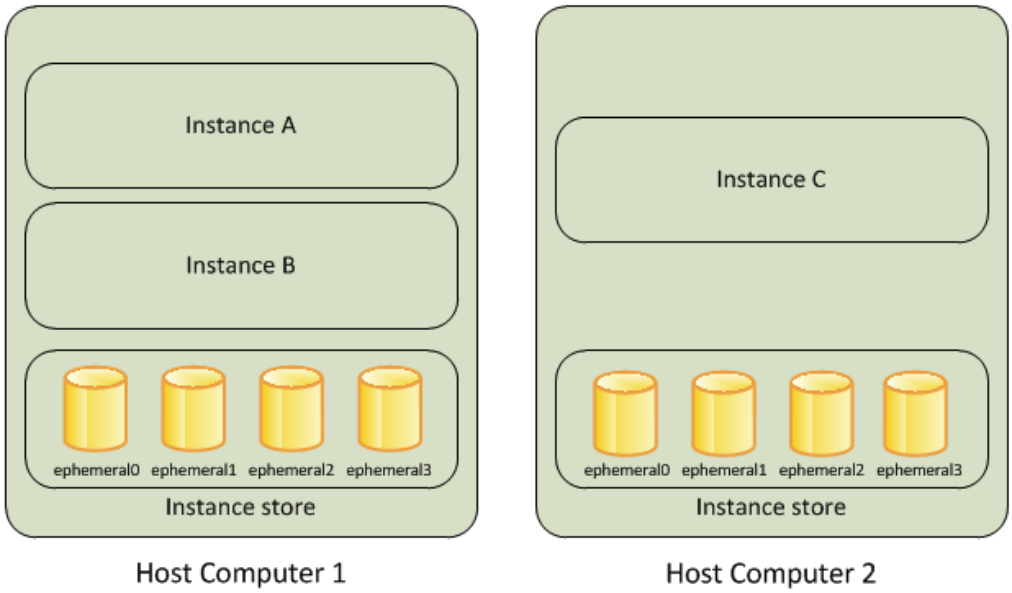
	General Purpose SSD		Provisioned IOPS SSD		
Volume type	gp3	gp2	io2Block Express ‡	io2	io1
Durability	99.8% - 99.9% durability (0.1% - 0.2% annual failure rate)	99.8% - 99.9% durability (0.1% - 0.2% annual failure rate)	99.999% durability (0.001% annual failure rate)	99.999% durability (0.001% annual failure rate)	99.8% - 99.9% durability (0.1% - 0.2% annual failure rate)
Use cases	• Low-latency interactive apps • Development and test environments		Workloads that require: • Sub-millisecond latency • Sustained IOPS performance • More than 64,000 IOPS or 1,000 MiB/s of throughput	• Workloads that require sustained IOPS performance or more than 16,000 IOPS • I/O-intensive database workloads	
Volume size	1 GiB - 16 TiB		4 GiB - 64 TiB	4 GiB - 16 TiB	
Max IOPS per volume (16 KiB I/O)	16,000		256,000	64,000 †	
Max throughput per volume	1,000 MiB/s	250 MiB/s *	4,000 MiB/s	1,000 MiB/s †	
Amazon EBS Multi-attach	Not supported		Not supported	Supported	
Boot volume	Supported				

- Locked to AZ, to attach to other AZ you have to snapshot it.
- Copying an unencrypted snapshot allows encryption.
- When copying an AMI to another region, automatically creates the underlying EBS snapshot also in the new region.
- RAID 0 to increase performance.
- RAID 1 to increase fault tolerance.
- If the instance is already running, you can set `DeleteOnTermination` to False using the `command line` for the root EBS volume.
- An in-progress snapshot is not affected by ongoing reads and writes to the volume hence, you can still use the EBS volume normally.
- Enforce the encryption of the new EBS volumes and snapshot copies that you create with `Encryption by Default` feature(no effect on existing EBS volumes or snapshots). If you enable it for a Region, you cannot disable it for individual volumes or snapshots in that Region.
- When you enable encryption by default, you can launch an instance only if the instance type supports EBS encryption.
- Amazon EBS does not support asymmetric CMKs.

Instance Store

- Temporary block-level storage for your instance.
- Ideal for temporary storage of information that changes frequently, such as buffers, caches, scratch data, and other temporary content, or

for data that is replicated across a fleet of instances, such as a load-balanced pool of web servers



- For high I/O performance, instance store volumes are a better option.
- You cant resize the instance store.

EFS

- Control which EC2 instances can access your EFS file system with security group rules and IAM policies.
- 1000s on concurrent NFS clients, 10Gbs throughput.
- Use EFS Access Points to manage application access.
- Max I/O performance mode is used to scale to higher levels of aggregate throughput and operations per second - tradeoff of slightly higher latencies.
- General Purpose performance mode is ideal for latency-sensitive use cases.
- `POSIX` compliant.
- Provisioned Throughput mode: for applications with high throughput to storage (MiB/s per TiB) ratios, or with requirements greater than those allowed by the Bursting Throughput mode.
- Bursting Throughput mode: designed to burst to high throughput levels for periods of time.
- Higher price point than EBS.
- Maximum days for the EFS lifecycle policy is 90.

Amazon FSx for Lustre:

- Run the world's most popular high-performance file system.
- For machine learning, high-performance computing (`HPC`), video processing, and financial modeling.
- Ability to both process the 'hot data' in a parallel and distributed fashion as well as easily store the 'cold data' on Amazon S3.

RDS

- `Multi A-Z synchronous replication` across AZs. Replication between the primary and standby instances does not incur additional data transfer charges.
- `Read replicas asynchronous replication` across AZs or cross-region, up to 5.

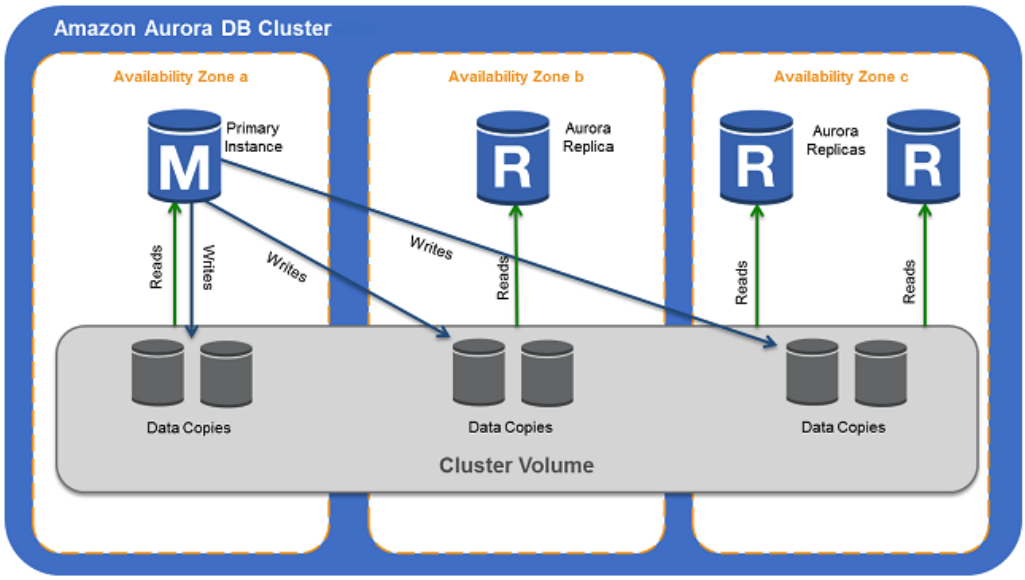
Multi-AZ deployments	Multi-Region deployments	Read replicas
Main purpose is high availability	Main purpose is disaster recovery and local performance	Main purpose is scalability
Non-Aurora: synchronous replication; Aurora: asynchronous replication	Asynchronous replication	Asynchronous replication
Non-Aurora: only the primary instance is active; Aurora: all instances are active	All regions are accessible and can be used for reads	All read replicas are accessible and can be used for readscaling
Non-Aurora: automated backups are taken from standby; Aurora: automated backups are taken from shared storage layer	Automated backups can be taken in each region	No backups configured by default
Always span at least two Availability Zones within a single region	Each region can have a Multi-AZ deployment	Can be within an Availability Zone, Cross-AZ, or Cross-Region
Non-Aurora: database engine version upgrades happen on primary; Aurora: all instances are updated together	Non-Aurora: database engine version upgrade is independent in each region; Aurora: all instances are updated together	Non-Aurora: database engine version upgrade is independent from source instance; Aurora: all instances are updated together
Automatic failover to standby (non-Aurora) or read replica (Aurora) when a problem is detected	Aurora allows promotion of a secondary region to be the master	Can be manually promoted to a standalone database instance (non-Aurora) or to be the primary instance (Aurora)

- Backups every 5min, ability to restore at any point in time.
- Supports storage autoscaling.
- IAM database authentication works with `MySQL and PostgreSQL` . Use an authentication token with a lifetime of 15 minutes.
- RDS provides metrics in real-time for the operating system (OS) that your DB instance runs on with `Enhanced Monitoring` (RDS processes, RDS child processes, OS processes).

- To encrypt unencrypted RDS database: create a snapshot of your DB instance, and then create an encrypted copy of that snapshot, restore DB from encrypted snapshot, terminate previous DB.
- `Upgrades to the database engine level require downtime`. Even if your RDS DB instance uses a Multi-AZ deployment, both the primary and standby DB instances are upgraded at the same time. This causes downtime until the upgrade is complete, and the duration of the downtime varies based on the size of your DB instance.
- RDS applies OS updates by performing maintenance on the standby, then promoting the standby to primary, and finally performing maintenance on the old primary, which becomes the new standby.
- Maximum backup retention period for automated backup is 35 days.

Aurora

- Auto-scales `up to 128 TB` per database instance.
- Aurora cluster: one Primary DB instance - `up to 15 replicas(read-only)`.



- You can specify the failover priority for Aurora Replicas, each Read Replica is associated with a priority tier (0-15). Aurora will promote the Read Replica that has the highest priority (the lowest numbered tier). If two or more Aurora Replicas share the same priority, then Amazon RDS promotes the replica that is the largest in size.

- Aurora `Global Database` is designed for `globally distributed applications`, allowing a single Amazon Aurora database to span multiple AWS regions, `sub-second data access in any region`.
- Storage automatically grows in increments of 10GB.
- In a multi-master cluster, all DB instances can perform write operations(scale writes, avoid downtime for writes) - `continuous availability` for applications where you can't afford even brief downtime for database write operations.
- Using endpoints, you can map each connection to the appropriate instance or group of instances based on your use case. For clusters with DB instances of different capacities or configurations, you can connect to custom endpoints associated with different subsets of DB instances.
- Reader endpoint automatically performs load-balancing among all the Aurora Replicas.
- For diagnosis or tuning, you can connect to a specific instance endpoint to examine details about a specific DB instance.
- If you are running Aurora Serverless and the DB instance or AZ becomes unavailable, Aurora will automatically recreate the DB instance in a different AZ.
- If you have a single instance, Aurora will attempt to create a new DB Instance in the same Availability Zone as the original instance. This replacement of the original instance is done on a best-effort basis and may not succeed.

DynamoDB

- DynamoDB Accelerator (DAX) is a fully managed, highly available, in-memory cache for Amazon DynamoDB that delivers `up to a 10 times performance improvement—from milliseconds to microseconds`.
- Tables must have provisioned read and write capacity units RCU, WRC.
- DynamoDB Streams allow changes in DynamoDB to be streamed to other services(read by Lambda etc, 24h retention on streams).
- Global Tables support multi-region replication, low latency, disaster recovery. Must first enable Streams.
- Can only query on primary key, sort key, or indexes.
- `All DynamoDB tables are encrypted`. There is no option to enable or disable encryption for new or existing tables. By default, all DynamoDB tables are encrypted under an AWS owned customer master key (CMK), which do not write to CloudTrail logs.

- If the shard iterator expires immediately before you can use it, this might indicate that the DynamoDB table used by Kinesis does not have enough capacity to store the lease data. To solve increase the write capacity assigned to the shard table.

Elasticache

- For `sub-millisecond latency caching`, ElastiCache is the best choice.

Memcached

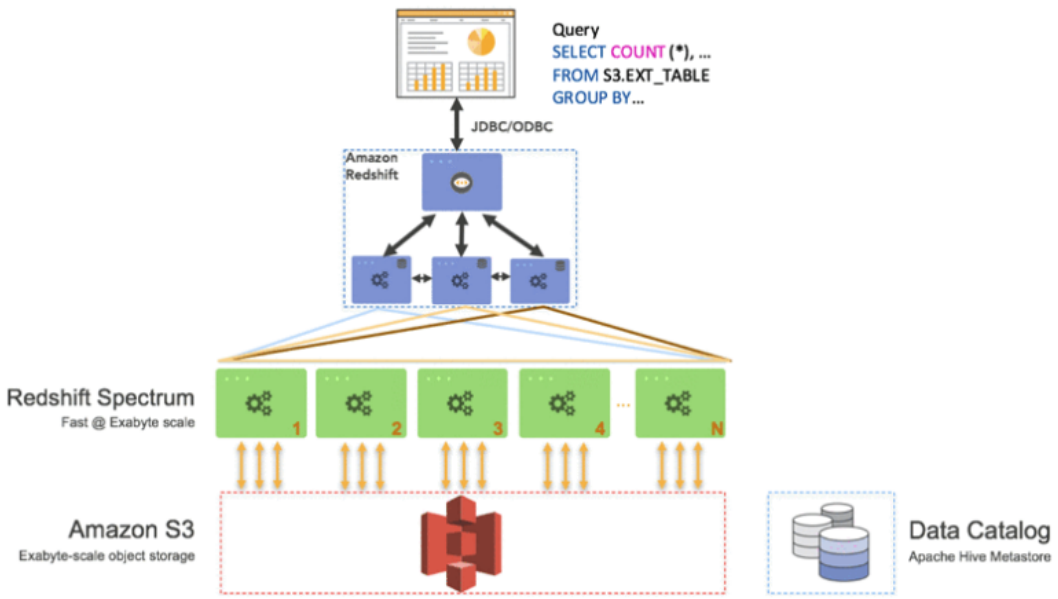
- Supports `multithreaded architecture`.

Redis

- Redis `HIPAA compliant`, supports replication, high availability, and cluster sharding.
- In-memory data store that provides `sub-millisecond latency`.
- IAM Auth is not supported by ElastiCache.
- Redis AUTH(enable Redis to require a token (password) before allowing clients to execute commands, thereby improving data security).

Redshift

- With Spectrum, you can efficiently query and retrieve structured and semistructured data from files in Amazon S3 without having to load the data into Amazon Redshift tables.



- For OLAP: online analytical processing.
- Redshift enhanced VPC routing, copy/unload goes through VPC.
- Possibility to copy snapshots for a cluster to another region for DR.

Cloudwatch

- Metrics belong to namespaces, Dimension is an attribute of a metric, Up to 10 dimensions per metric.
- Automatically recover ec2: If your instance has a public IPv4 address, it `retains the public IPv4 address` after recovery. During instance recovery, the instance is migrated during an instance reboot, and `any data that is in-memory is lost`.
- You can use CloudWatch Events to run Amazon ECS tasks when certain AWS events occur.

EventBridge

- Recommended when you want to build an application that reacts to events from SaaS applications and/or AWS services. Only event-based service that integrates directly with third-party SaaS partners.

Encryption/Secrets

- Key Policies: control access to keys, you cannot control access without them.
- Automatic key rotation: CMK every one year.
- SSE-KMS is a service that combines secure, highly available hardware and software to provide a key management system scaled for the cloud. When you use server-side encryption with AWS KMS (SSE-KMS), you can specify a customer-managed CMK that you have already created. SSE-KMS provides you with an audit trail that shows when your CMK was used and by whom.
- `Deleting a customer master key (CMK) has enforced a waiting period`, you schedule key deletion(`minimum of 7 days up to a maximum of 30 days(default)`)
- SSE-C - With Server-Side Encryption with Customer-Provided Keys (SSE-C), you manage the encryption keys and Amazon S3 manages the encryption, as it writes to disks and decryption when you access your objects.
- SSE-S3 - When you use Server-Side Encryption with Amazon S3-Managed Keys (SSE-S3), each object is encrypted with a unique key.

Uses 256-bit Advanced Encryption Standard (AES-256).
- Client-side encryption when there is a proprietary encryption algorithm.

Secrets Manager

- Nice RDS integration.
- Force secret rotation every X days.

SSM Parameter Store

- Allow assigning TTL to a parameter(expiration date) to force update/delete of sensitive data.

CloudHSM

- Dedicated hardware, you manage your own encryption keys.
- Good option to use with SSE-C.
- It is possible to lose keys that were created since the most recent daily backup if the CloudHSM cluster that you are using fails and you are not using two or more HSMs.

Kinesis

Kinesis Data Streams

- `Default data retention 1 day`, can go up to 7.
- `1MB/sec/shard ingest capacity`.
- By default, the `2MB/second/shard output` is shared between all of the applications consuming data from the stream.
- Use enhanced fan-out if you have multiple consumers retrieving data from a stream in parallel, automatically scales throughput with the number of shards.
- Ability for multiple apps to consume the same stream concurrently.
- Ability to consume records in the `same order a few hours later`.
- Routing related records to the same record processor. For example, counting and aggregation are simpler when all records for a given key are routed to the same record processor.

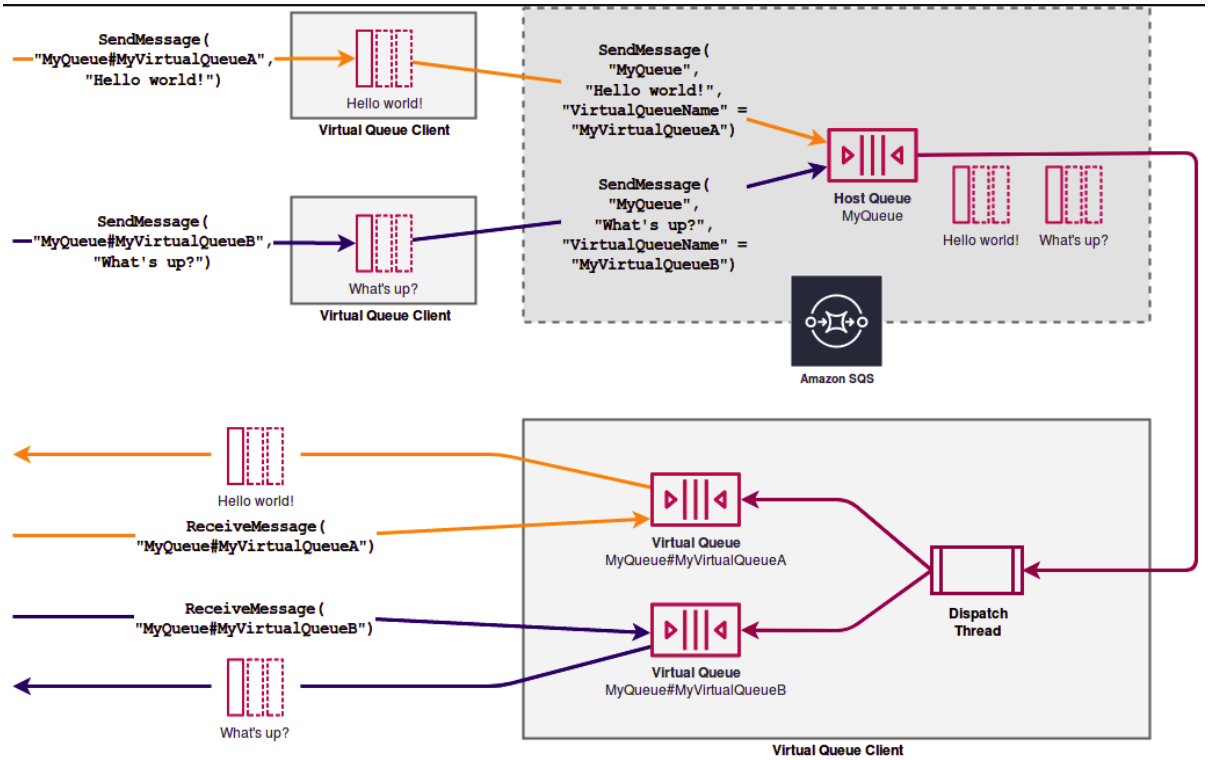
Kinesis Firehose

- `Automatically scales to match the throughput` of your data and requires no ongoing administration. Auto-scaling solution, as there is

no need to provision any shards like Kinesis Data Streams.
- Kinesis Agent cannot write to a Kinesis Firehose for which the delivery stream source is already set as Kinesis Data Streams.
- Data into Amazon S3, Amazon Redshift, Amazon Elasticsearch Service, generic HTTP endpoints and Datadog, New Relic, MongoDB, Splunk.
- Load streaming data into Redshift for near real-time analytics.

SQS

- When you need `messaging semantics` (ack/fail) and `visibility timeout` (default 30s).
- Dynamically increasing concurrency/throughput at read time.
- FIFO queues support up to `3,000 messages` (batch 10 messages per operation- max) per second with batching(`300 without`), have an 80-character queue name limit.
- Message retention `4 days default, 14 days max` .
- Limit 256kb per message sent.
- To scale to `same number of consumers as producers` , send data with a `Group ID attribute` .
- Delay queues let you postpone the delivery of new messages to a queue for several seconds. `The default (minimum) delay for a queue is 0 seconds` . The maximum is 15 minutes.
- You can use message timers to set an initial invisibility period for a message added to a queue. Default delay for a message is 0 seconds. The maximum is 15 minutes.
- Temporary queues help you save development time and deployment costs when using common message patterns such as request-response. To better support short-lived, lightweight messaging destinations, AWS recommends Amazon SQS Temporary Queue Client. The key concept behind the client is the `Virtual Queue` . Virtual queues let you multiplex many low-traffic queues onto a single SQS queue.



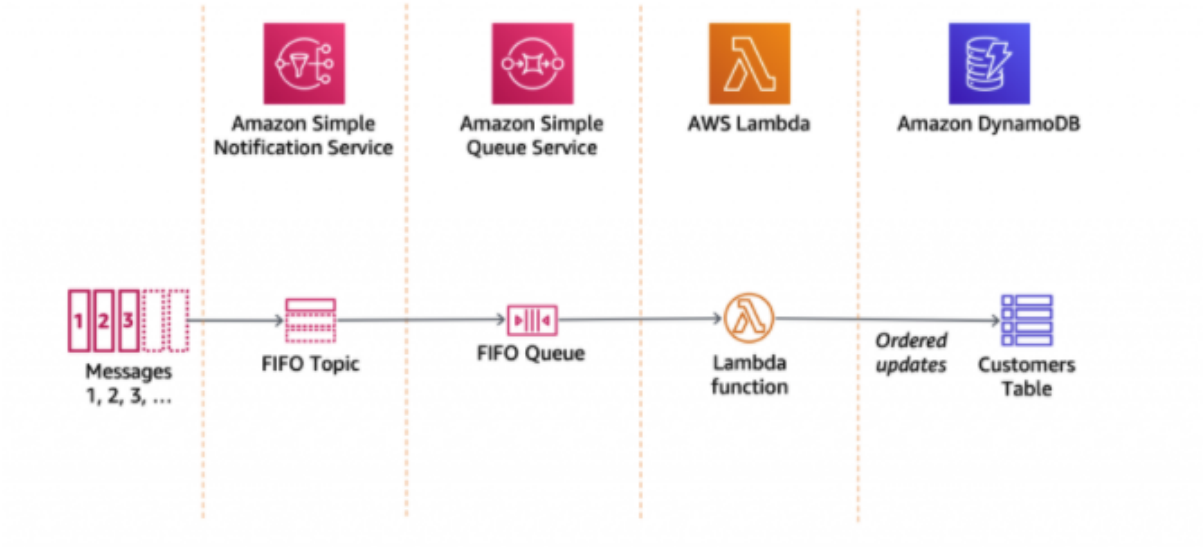
- AWS recommends using `separate queues to provide prioritization of work`.
- A single SQS message queue can contain an unlimited number of messages. However, there is a 120,000 quota for the number of inflight messages for a standard queue and 20,000 for a FIFO queue. Messages `are inflight after they have been received from the queue by a consuming component, but have not yet been deleted` from the queue.
- Standard queues provide `at-least-once` delivery, which means that each message is delivered at least once.
- FIFO queues provide `exactly once` processing, which means that each message is delivered once and remains available until a consumer processes it and deletes it. `Duplicates are not introduced` into the queue.
- An Amazon SQS message can contain up to 10 metadata attributes.
- By default, `ReceiveMessageWaitTimeSeconds` is zero which means it is using Short polling. If it is set to a value greater than zero, then it is Long polling.

SNS

- Event producers send events to 1 topic, we can have many subs.
- 100000 topics limit
- Use SNS `message filtering` to assign a filter policy to the topic subscription, and the subscriber will `only receive a message that`

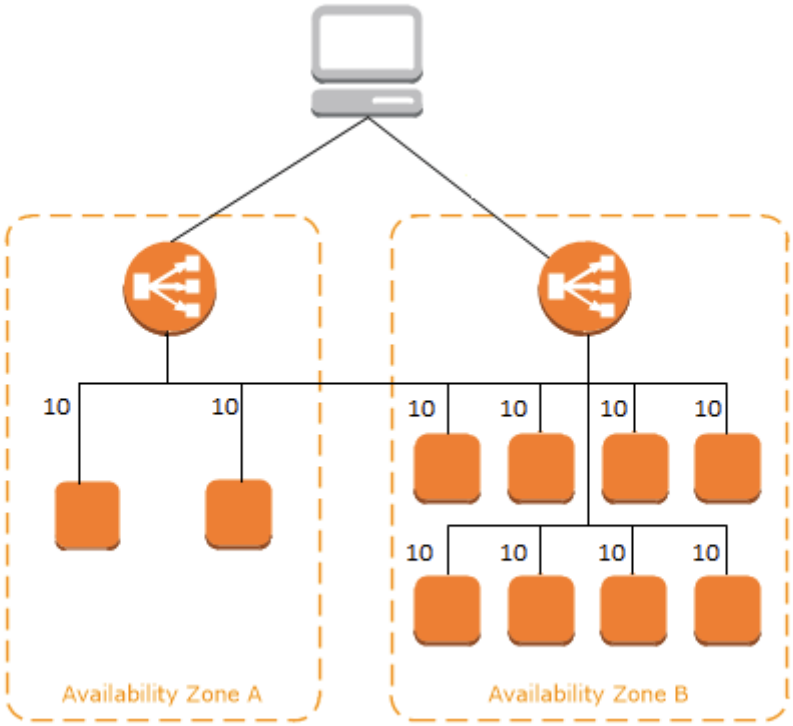
`they are interested in`.

- SNS FIFO for strict message ordering and deduplicated message delivery to one or more subscribers



LoadBalancers

- LBs can scale but `not instantaneously`.
- Elastic Load Balancing stops sending requests to targets that are deregistering. By default, Elastic Load Balancing waits 300s(can be set between 1s to 3600s) seconds before completing the deregistration process, which can `help in-flight requests to the target to complete` (connection drain).
- When `cross-zone load balancing is enabled`, each load balancer node `distributes traffic` across the registered targets in all enabled Availability Zones `evenly`.



- By default, cross-zone load balancing is enabled for Application Load Balancer and disabled for Network Load Balancer
- ELB cannot distribute incoming traffic for targets deployed in different regions.
- Access logging is an optional feature of Elastic Load Balancing that is disabled by default. Use to analyze traffic patterns and troubleshoot issues.

Application LoadBalancer

- ALB targets with `instance ID` route to `primary private IP in primary NIC` , targets using IP addresses route to any private IP from one or more NICs.
- Host-based Routing: You can route a client request based on the Host field of the HTTP header allowing you to route to multiple domains from the same load balancer.
- Path-based Routing: You can route a client request based on the URL path of the HTTP header.
- HTTP header-based routing: You can route a client request based on the value of any standard or custom HTTP header.
- HTTP method-based routing: You can route a client request based on any standard or custom HTTP method.
- Query string parameter-based routing: You can route a client request based on the query string or query parameters.

- Source IP address CIDR-based routing: You can route a client request based on source IP address CIDR from where the request originates.
- ALB not registered any targets with the target groups -> 503 error.
- Use Cognito Authentication via Cognito User Pools for your ALB.
- With SNI support AWS makes it easy to use more than one certificate with the same ALB.
- You can host multiple TLS secured applications, each with its own TLS certificate, behind a single ALB. In order to use SNI, all you need to do is bind multiple certificates to the same secure listener on your load balancer. ALB will automatically choose the optimal TLS certificate for each client.
- ALBs support Weighted Target Groups routing.

Network LoadBalancer

- NLB traffic is routed using the private IP address.
- Network LB has no security groups, lets traffic passing by.
- With Network Load Balancer (NLB), you can offload the decryption/encryption of TLS traffic from your application servers to the NLB.

Classic LoadBalancer

- CLB does not support Server Name Indication (SNI).

Lambda

- Supports `1000 concurrent executions per AWS account per region` , contact support to raise the limit if needed.
- Supported languages: C#/.NET, GO, node.js, Python, Java, Ruby.
- lamda@edge: deploy lambda to each region alongside your CloudFront CDN.
- You can set your memory from 128MB to 10,240MB
- If your Lambda function accesses a VPC, you must make sure that your VPC has sufficient ENI or subnect IPs capacity to support the scale requirements of your Lambda function.

Step Functions

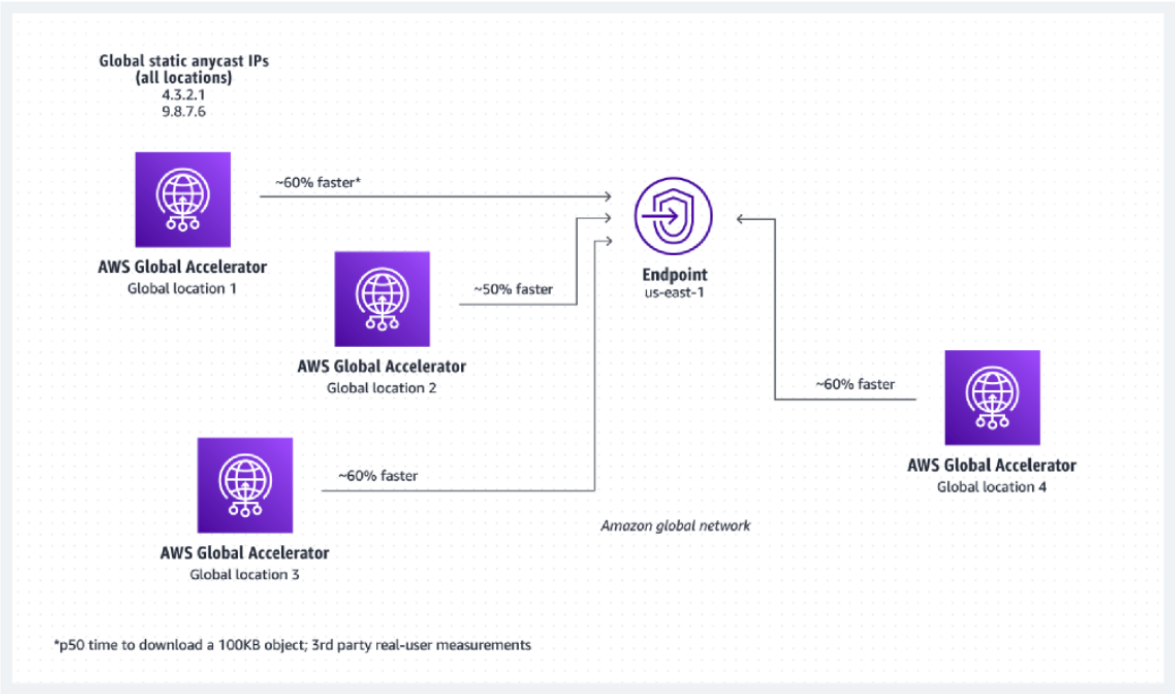
- Serverless workflows orchestration.

Cloudfront

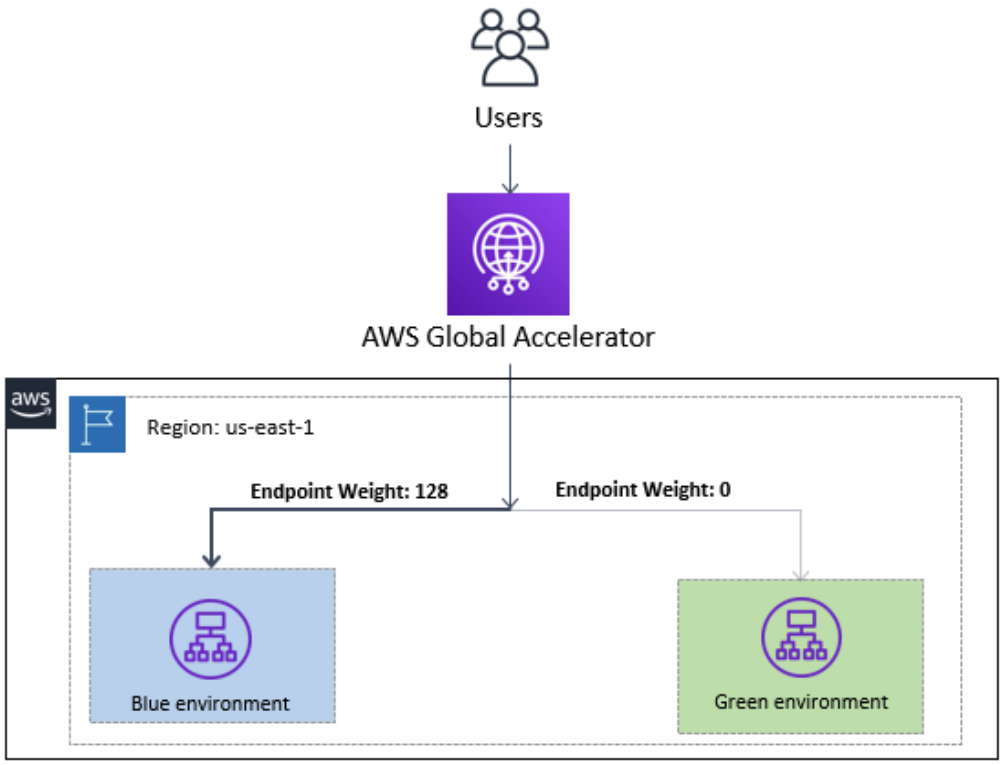
- Delivering data out of CloudFront can be `more cost-effective than delivering it from S3` directly to your users.
- Use CloudFront to `improve application performance` to serve static content from S3.
- Dynamic content does not flow through regional edge caches, but goes directly to the origin - Proxy methods PUT/POST/PATCH/OPTIONS/DELETE go directly to the origin.
- Preferred to handle spikes in traffic over GA.
- You `cannot directly integrate Cognito User Pools` with CloudFront distribution as you have to create a separate Lambda@Edge function to accomplish the authentication via Cognito User Pools.
- CloudFront can route to `multiple origins based on the content type`.
- Field-level encryption: The sensitive information provided by your users is `encrypted at the edge` (You can't encrypt all of the data in a request with field-level encryption; you must specify individual fields to encrypt).
- CloudFront `signed cookies` -> provide access to `multiple restricted files`.
- CloudFront `signed URLs` -> access to `one file`.
- Can also use an EC2 instance or a custom origin in configuring CloudFront.
- The `Cache-Control` and `Expires` headers control how long objects stay in the cache. The `Cache-Control max-age` directive lets you specify how long (in seconds) you want an object to remain in the cache before CloudFront gets the object again from the origin server. The minimum expiration time CloudFront supports is 0 seconds for web distributions and 3600 seconds for RTMP distributions.

Global Accelerator(GA)

- Directs traffic to optimal endpoints over the AWS global network.
- Improves the availability and performance of your internet applications.
- Two static anycast IP addresses that act as a fixed entry point to your application endpoints.



- Good fit for non-HTTP use cases, such as gaming (UDP), IoT (MQTT), or Voice over IP.
- Uses endpoint weights to determine the proportion of traffic that is directed to endpoints in an endpoint group(can be used in blue/green deployments).



AWS Global Accelerator endpoint weights

WAF

- Use AWS WAF to `block or allow requests based on conditions` that you specify, such as the IP addresses.
- Geographic `(Geo) Match Conditions` in AWS WAF to restrict application access based on the geographic location of your viewers - choose the countries from which AWS WAF should allow access.
- Protects against `SQL injection and Cross-Site Scripting`.
- `Rate based` rules(DDoS protection).

Firewall Manager

- Centrally configure and manage firewall rules across your accounts and applications in AWS Organizations.
- You can centrally configure AWS WAF rules, AWS Shield Advanced protection, Amazon Virtual Private Cloud (VPC) security groups, AWS Network Firewalls, and Amazon Route 53 Resolver DNS Firewall rules across accounts and resources in your organization. It does `not support Network ACLs` as of today.

AWS Shield

- `DDoS`, protection against `SYN/UDP floods`, reflection attacks, and other layer/3 & layer 4 attacks.

EMR

- Cloud big data platform for processing vast amounts of data using Apache Spark, Apache Hive, Apache HBase, Apache Flink, Apache Hudi, and Presto, Hadoop.

Beanstalk

- Easy-to-use service for deploying and scaling web applications and services developed with Java, .NET, PHP, Node.js, Python, Ruby, Go, and Docker.
- Automatically handles the deployment, from capacity provisioning, load balancing, auto-scaling to application health monitoring.
- You retain `full control over the AWS resources powering your application and can access the underlying resources at any time`.
- Application files are stored in S3. The server log files can also optionally be stored in S3 or in CloudWatch Logs.

CloudFormation

- StackSet extends the functionality of stacks by enabling you to create, update, or delete stacks `across multiple accounts and regions` with a single operation.
- Use the `CreationPolicy` attribute when you want to wait on resource configuration actions before stack creation proceeds.

Cognito

User pools

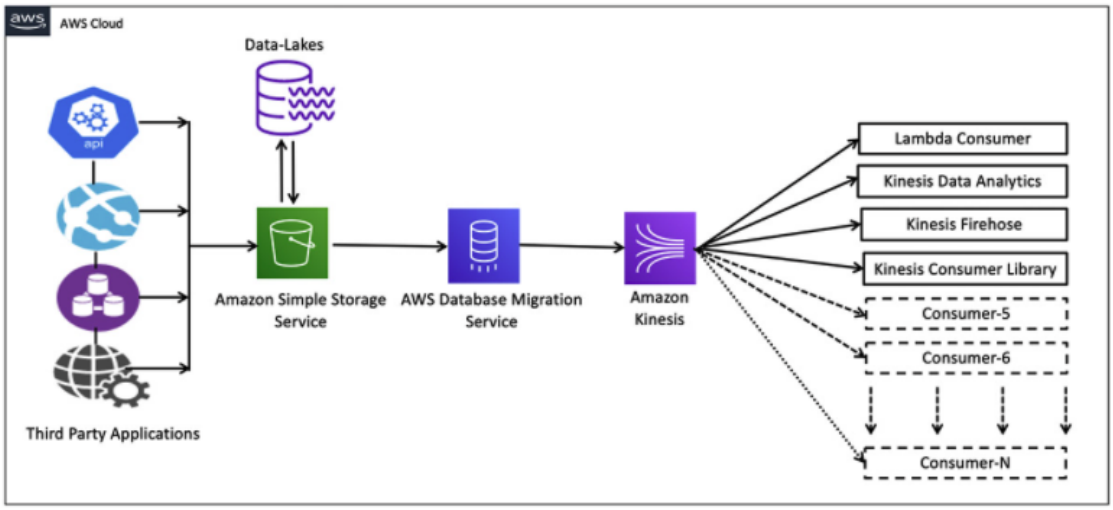
- Provide built-in user management e.g. sign-in and register functionality for apps.

Identity pools

- Provide temporary credentials for AWS access to users.

AWS Database Migration Service

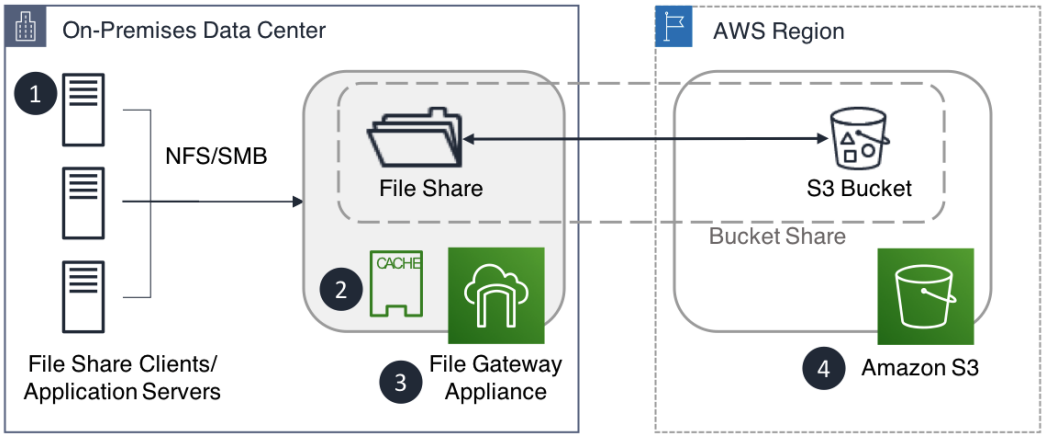
- Seamlessly migrate data from supported sources to relational databases, data warehouses, streaming platforms, and other data stores in AWS cloud. (e.g. quickly move data from S3 to Kinesis data streams, not only for DBs).



Storage Gateway

File Gateway

A *bucket share* consists of a file share hosted from a file gateway across a single Amazon S3 bucket. The file gateway virtual machine appliance currently supports up to 10 bucket shares.



- SMB or NFS access to data in S3 with local caching.

Volume Gateway

- Present cloud-based iSCSI block storage volumes to your on-premises applications.

Tape Gateway

- Supports archiving directly to Glacier and Glacier Deep Archive.

DataSync

- Move large data from on-premise to AWS.
- Can move data `directly to Glacier or Glacier Deep Archive`.

AppSync

- Store and sync data across mobile and web apps in real-time.

CloudTrail

- By default, CloudTrail event log files are encrypted using Amazon S3 server-side encryption (SSE).

Xray

- AWS X-Ray helps developers analyze and `debug` production, distributed applications, such as those built using a microservices architecture.

- End-to-end view of requests `as they travel through your application`, and shows a map of your application's underlying components.
- The X-Ray agent can assume a role to publish data into an account different from the one in which it is running.

GuardDuty

- Threat detection that enables you to continuously monitor and protect your AWS accounts, workloads, and data stored in Amazon S3.
- Analyses AWS `CloudTrail Events, Amazon VPC Flow Logs, and DNS Logs`.
- `Disabling the service` in the general settings `deletes all the remaining data`.

Macie

- Discover and protect your sensitive data on Amazon S3.

Inspector

- Helps you check for unintended network accessibility of your Amazon EC2 instances and for vulnerabilities on those EC2 instances.

Recognition

- Automate your image and video analysis with machine learning.

VPC Endpoints

- When you create a VPC endpoint, you can attach an endpoint policy that controls access to the service to which you are connecting.

Gateway Endpoints

- GE is a gateway that you specify as a `target for a route in your route table for traffic destined to a supported AWS service`.
- S3 & DynamoDB only.

Interface Endpoints

- An elastic network interface with a private IP address from the IP address range of your subnet that serves as an entry point for traffic destined to a supported service.

Elastic Network Adapter(ENA)

- Enhanced networking capabilities with network speeds of up to 100 Gbps.
- Supports Windows.

Elastic Fabric Adapter (EFA)

- Network device that you can attach to your Amazon EC2 instance to accelerate `High-Performance Computing (HPC)` and machine learning applications.
- ENA with added capabilities.
- Doesn't support Windows.
- EFA support can be enabled either at the launch of the instance or added to a stopped instance. EFA devices cannot be attached to a running instance.

API Gateway

- Rest APIs - stateful client-server communication.
- Websocket APIs - stateless full-duplex communication.
- All of the APIs created with Amazon API Gateway expose HTTPS endpoints only

SWF Simple Workflow Service

- Use if you need: `external signals to intervene`, or child processes to `return values to parent processes`.
- For decoupled architectures.
- Provides useful guarantees around task assignments. It ensures that a task is never duplicated and is assigned only once.

AWS Backup

- Centralized backup service.
- A backup plan is a policy expression that defines when and how you want to back up your AWS resources.

AWS Batch

- Multi-node parallel jobs.

AWS ParallelCluster

- Cluster management tool to deploy HPC, automate creation of vpc, subnet, cluster type etc.

AD Connector

- If you only need to allow your on-premises users to log in to AWS applications and services with their Active Directory credentials.

AWS Managed Microsoft AD

- Configure a trust relationship between AWS Managed Microsoft AD in the AWS Cloud and your existing on-premises Microsoft Active Directory, providing users and groups with access to resources in either domain, using single sign-on (SSO).

Data Transfer

- No charge for inbound data transfer across all services in all Regions.
- Data transfer from AWS to the internet is charged per service, with rates specific to the originating Region.
- There is a charge for data transfer across Regions.
- Data transfer within the same Availability Zone is free.
- Data transfer over a VPC peering connection that stays within an Availability Zone is free. Data transfer over a VPC peering connection that crosses Availability Zones will incur a data transfer charge for ingress/egress traffic. If the VPCs are peered across Regions, standard inter-Region data transfer charges will apply.
- Data processing charges apply for each GB sent from a VPC, Direct Connect, or VPN to Transit Gateway.
- Direct Connect & VPN also incur charges for data flowing out of AWS.

Important ports:

- FTP: 21
- SSH: 22
- SFTP: 22 (same as SSH)
- HTTP: 80
- HTTPS: 443
- RDP: TCP 3389 and UDP 3389

RDS Databases ports:

- PostgreSQL: 5432

- MySQL: 3306
- Oracle RDS: 1521
- MSSQL Server: 1433
- MariaDB: 3306 (same as MySQL)
- Aurora: 5432 (if PostgreSQL compatible) or 3306 (if MySQL compatible)

Disaster Recovery in AWS

- RPO: Recovery Point Objective -> how much data loss we are willing to recover
- RTO: Recovery Time Objective -> downtime between disaster and RTO