IAM 101 and S3

identity access management offers the following features.

* Centralised control of your AWS account.

* Shared access to your AWS account.

* Granular permissions. So you can say, okay, I want people to be able to access this service, but I don't want people to be able to access that service.

* Identity Federation (including Active Directory, Facebook, LinkedIn ... etc)

* Multifactor authentication.

* Provides temporary access for users or devices and services where necessary

* Allows you to set up your own password rotation policy.

* Integrates with many different AWS services

* It supports PCI DSS compliance.

* PCI DSS compliance just is basically a compliant framework that if you're taking credit card details, you need to be compliant with the framework. So IAM supports PCI DSS.

https://docs.aws.amazon.com/IAM/latest/UserGuide/intro-structure.html

Key Terminology for IAM:

1. Users: End users such as people, employees of an organisation ... etc

2. Groups: A collection of users. So each user in the group will inherit the permissions of the group.

3. Policies: policies are made up of documents, called policy documents. These documents are in a format called JSON and they give permission as to what a User/Group is able to do.

4. Roles: You can create roles and assign them to AWS resources

Exam Tips:

* IAM is universal (Global). It doesn't apply to regions at this time.

* The "root account" is simply the account created when first setup your AWS account. It has complete Admin access.

* New Users have NO permissions when first created.

* New Users are assigned Access Key ID & Secret Access Keys when first created.

* These are not the same as a password. You cannot use the Access key ID & Secret Access to login to the console. You can use this to access AWS via the APIs and Command Line, however.

* You can get to view these once. If you lose them, you have to regenerate them. So, save them in a secure location.

* Always setup Multifactor Authentication on your root account.

* You can create and customise your own password rotation policies.

S3 Exam Tips:

1. Remember that S3 is Object-based: i.e allows you to upload files.

2. Files can be 0 Bytes to 5 TB.

3. There is unlimited storage.

4. Files are stored in Buckets.

5. S3 is a universal namespace. That is, names must be unique globally. Example: https://s3-eu-west-1.amazonaws.com/acloudguru

6. Not suitable to install operating systems on S3 due to it being object based (not block storage - EBS), can only be used to store files

7. Successful upload will generate a HTTP 200 status code.

8. You can turn on MFA delete to avoid accidental delete.

9. The key fundamentals of S3 are:

    - Key (This is simply the name of the object)

    - Value (This is simply the data and is made up of a sequence of bytes).

10. S3 Model:

    - Read after Write consistency for PUTS of new objects (if you write a new files and read it immediately afterwards, you will be able to view the data)

    - Eventual Consistency for overwrite PUTS and DELETES (can take sometime to propagate) (If you update AN EXISTING file or delete a file and read it immediately, you may get the older version, or you may not. Basically changes to objects can take a little bit of time to propagate.)

11. S3 Storage Classes

    1. S3 Standard: 99.99% availability, 99.99999999999% durability stored redundantly across multiple devices in multiple facilities and is designed to sustain the loss of 2 facilities concurrently.

2. S3 - IA (Infrequently Accessed): 99.9% availability, For the data that is accessed less frequently, but requires rapid access when needed. Lower fee than S3, but you are charged a retrieval fee.

3. S3 - Intelligent Tiering: 99.9% availability, Designed to optimise costs by automatically moving data to the most cost-effective access tier, without performance impact or operational overhead.

4. S3 One Zone IA (also called S3 RRS): 99.5% availability, For where you want a lower-cost option for infrequently accessed data, but do not require the multiple availability zone data resilience.

5. S3 Glacier: S3 Glacier is a secure, durable and low-cost storage class for data archiving. You can reliably store any amount of data at costs that are competitive with or cheaper than on-premises solutions. Retrieval times are configurable from minutes to hours.

6. S3 Glacier Deep Archive: S3 Glacier Deep Archive is Amazon's lowest-cost storage class where a retrieval time of 12 hours is acceptable.

7. S3 Outposts for on-premises object storage to meet data residency needs.

S3 Bucket Exam Tips:

12. Bucket names share a common name space, you can't have the same bucket name as the others.

13. When you view Buckets you view them globally but you can have buckets in individual regions.

14. We can use cross region replication to replicate buckets automatically to different regions.

15. We can change storage class and encryption on the fly.

16. Transfer acceleration

17. Restricting Bucket Access:

    1. Bucket Policies - Applies across the bucket.

    2. Object Policies - Applies to individual files.

    3. IAM Policies to Users & Groups - Applies to Users & Groups.

18. By default Buckets are not public.

19. You can use bucket policies to make entire S3 buckets public.

20. You can use S3 bucket to host static websites but can't host dynamic websites or websites which require database, for ex: Wordpress...etc.

21. Control access to buckets using either a bucket ACL or bucket policies.

22. When we change from Allow to Deny on a policy OR create a policy with DENY, it's called an explicit DENY and it would always overwrite the allow in any other policy.

23. By default, all permissions are implicitly denied.

* Important: Read S3 FAQ's: https://aws.amazon.com/s3/faqs/

S3 Pricing Tier: What drives the price?

What makes up the cost?

1. Storage
2. Requests and Data Retrievals
3. Data Transfer
4. Management and Replication

S3 Pricing (Very Important) Exam Tips:

Understand how to get the best value out of s3

1. S3 Standard - Avoid S3 Standard, use S3 - Intelligent Tiering
2. S3 IA
3. S3 - Intelligent Tiering
4. S3 One Zone - IA (No redundancy, if the zone fails, we loose the data)
5. S3 Glacier - For Archival services
6. S3 Glacier Deep Archive - For Archival services

S3 Security & Encryption:

1. By default, all newly created buckets are PRIVATE, you can setup access control to your buckets using;

a. Bucket Policies

b. ACL'S

2. S3 buckets can be configured to create access logs which log all requests made to the S3 bucket.

This can be sent to another bucket or even another bucket in another account.

S3 Encryption (Very important - will be tested on this in the exam)

1. Encryption In Transit is achieved by

    a. SSL/TLS (HTTPS)

2. Encryption at Rest (Server Side) is achieved by

    a. Serve-Side

        1. S3 Managed Keys - SSE - S3

        2. AWS Key Management Service, Managed Keys - SSE-KMS

        3. Serve-Side Encryption with Customer Provided Keys - SSE-C

    b. Client-Side Encryption

        4. Done by the client

Lab: We can encrypt the files in S3 by using AES-256 or AWS-KMS.

Note: AWS-KMS is beyond the scope of this exam

S3 Encryption Demo

1. Stores all versions of an object (including all writes and even if you delete an object)

2. Great backup tool.

3. Once enabled, Versioning cannot be disabled, only be suspended. You have to delete the bucket and create a new one.

4. Integrates with Lifecycle rules.

5. Versioning's MFA Delete capability, which uses multi-factor authentication, can be used to provide an additional layer of security.

S3 Versioning Lab

Imp:

1. With versioning enabled, when we upload a newer version of the file - the newer file's permission need to be made public. Older versions permissions remain the same.

2. From Architecture perspective - based on the requirements, versioning need to be enabled as the size of S3 bucket will increase exponentially.

Or make a decision to enable lifecycle policy to retire old versions quickly.

S3 Versioning Exam Tips:

1. Stores all versions of an object (including all writes and even if you delete an object)

2. Great backup tool.

3. Once enabled, Versioning cannot be disabled, only suspended

4. Integrates with Lifecycle rules.

5. Versioning's MFA Delete capability, which uses multi-factor authentication, can be used to provide an additional layer of security.

S3 LifeCycle Management Lab

Note:

1. Transitioning small objects to Glacier or Glacier Deep Archive will increase costs

Before creating a lifecycle rule that transitions small objects to Glacier or Glacier Deep Archive, consider how many objects will be transitioned and how long you plan to keep the objects. Lifecycle request charges for these objects will increase your costs.

2. Expire current versions of objects: Min days must be greater than 60

S3 LifeCycle Management - Exam Tips

1. Automates moving your objects between the different storage tiers.

2. Can be used in conjunction with versioning.

3. Can be applied to current and previous versions.

S3 Object Lock and Glacier Vault Lock

1. S3 Object Lock: Store objects using a write-once-read-many (WORM) model to help you prevent objects from being deleted or overwritten for a fixed amount of time or indefinitely.

2. You can use S3 Object lock to meet regulatory requirements that require WORM storage, or add an extra layer of protection against object changes and deletion.

3. To enable object lock, it must be first enabled at the bucket level. Amazon S3 currently does not support enabling object lock after a bucket has been created. To enable object lock for this bucket, contact customer support.

4. Comes in different modes:

a. Governance Mode: Users can't overwrite or delete an object version or alter its lock settings unless they have special permissions.

With governance mode, you protect objects against being deleted by most users, but you can still grant some user permissions to alter the retention settings or delete the object if unnecessary.

b. Compliance Mode: A protected object version can't be overwritten or deleted by any user, including the root user in your AWS account. When an object is locked in compliance mode, its retention mode can't be changed and its retention period can't be shortened. Compliance mode ensures an object version can't be overwritten or deleted for the duration of the retention period.

5. Retention Period: A retention period protects an object version for a fixed amount of time. When you place a retention period on an object version, Amazon S3 stores a timestamp in the object version's metadata to indicate when the retention period expires. After the retention period expired, the object version can be overwritten or deleted unless you also placed a legal hold on the object version.

6. Legal Hold: S3 Object lock also enables you to okay a legal hold on an object version. Like a retention period, a legal prevents object version from being overwritten or deleted. However, a legal hold doesn't have an associated retention period and remains in effect until removed. Legal holds can be freely placed and removed by any user who has the s3:PutObjectLegalHold permission.

Glacier Vault Lock: It allows you to easily deploy and enforce compliance controls for individual S3 Glacier vaults with a Vault Lock policy. You can specify controls, such as WORM, in a Vault Lock policy from future edits. Once locked, the policy can no longer be changed.

S3 Object Lock and Glacier Vault Lock - Exam Tips

1. Use S3 Object lock to store objects using a write once, read many (WORM) model.

2. Object locks can be on individual objects or applied across the bucket as a whole.

3. Object locks come in two models: Governance mode and Compliance mode.

4. With governance mode, Users can't overwrite or delete an object version or alter its lock settings unless they have special permissions.

5. With Compliance Mode: A protected object version can't be overwritten or deleted by any user, including the root user in your AWS account.

6. S3 Glacier Vault Lock allows you to easily deploy and enforce compliance controls for individual S3 Glacier vaults with a Vault Lock policy. You can specify controls, such as WORM, in a Vault Lock policy from future edits. Once locked, the policy can no longer be changed.

S3 Performance

1. S3 Prefix:

mybucketname/folder1/subfolder1/myfile.jpg > /folder1/subfolder1 is the prefix

mybucketname/folder2/subfolder1/myfile.jpg > /folder2/subfolder1 is the prefix

mybucketname/folder3/myfile.jpg > /folder3 is the prefix

mybucketname/folder4/subfolder4/myfile.jpg > /folder4/subfolder4 is the prefix

Why is prefix important?

1. S3 is all about performance, S3 has extremely low latency. You can get the first byte out of S3 within 100-200 milliseconds

You can also achieve a high number of requests: 3,500 PUT/COPY/POST/DELETE and 5,500 GET/HEAD requests per second per prefix

2. You can get better performance by spreading your reads across different prefixes. For Example, If you are using two prefixes, you can achieve 11,000 requests per second.

3. If we used all four prefixes in the last example, you would achieve 22,000 requests per second.

2. S3 Limitations when using KMS

1. If you are using SSE-KMS to encrypt your objects in S3, you must keep in mind the KMS limits.

2. When you upload a file, you will call GenerateDataKey in the KMS API.

3. When you download a file, you will call Decrypt in the KMS API.

4. Uploading/Downloading will count towards the KMS quota.

5. Region-specific, however, it's either 5,500, 10,000 or 30,000 requests our second.

6. Currently, you cannot request a quota increase for KMS.

3. S3 Performance Uploads

Multipart Uploads:

1. Recommended for files over 100MB

2. Required for files over 5GB

3. Parallelise uploads (increase efficiency)

4. S3 Performance Downloads:

S3 Byte-Range Downloads:

1. Parallelise downloads by specifying byte ranges.

2. If there's a failure in the download, it's only for a specific byte range.

S3 Byte-Range Fetches:

1. Can be used to speed up downloads

2. Can be used to just download partial amounts of the file (eg., header information)

S3 Performance - Exam Tips:

1. mybucketname/folder1/subfolder1/myfile.jpg > /folder1/subfolder1 is the prefix

2. You can also achieve a high number of requests: 3,500 PUT/COPY/POST/DELETE and 5,500 GET/HEAD requests per second per prefix

3. You can get better performance by spreading your reads across different prefixes. For Example, If you are using two prefixes, you can achieve 11,000 requests per second.

4. If you are using SSE-KMS to encrypt your objects in S3, you must keep in mind the KMS limits.

a. Uploading/Downloading will count towards the KMS quota.

b. Region-specific, however, it's either 5,500, 10,000 or 30,000 requests our second.

c. Currently, you cannot request a quota increase for KMS.

5. Use multipart uploads to increase performance when uploading files to S3.

6. Should be used for any files over 100MB and must be used for any file over 5GB.

7. Use S3 byte-range fetches to increase performance when downloading files to S3.

S3 Select and Glacier Select:

S3 Select: S3 Select enables applications to retrieve only a subset of data from an object by using simple SQL expressions.

By using S3 Select to retrieve only the data needed by your application, you can achieve drastic performance increases - in many cases, you can get as much as 400% improvement (up to 80% cheaper)

Glacier Select: Some companies in highly regulated industries - e.g., financial services, healthcare, and others - write data directly to Amazon Glacier to satisfy compliance needs like SEC Rule 17a-4 or HIPAA. Many S3 users have lifecycle policies designed to save on storage costs by moving their data into Glacier when they no longer need to access to on a regular basis.

Glacier Select allows you to run SQL queries against Glacier directly.

S3 Select and Glacier Select - Exam Tips

1. Remember that S3 Select is used to retrieve only a subset of data from an object by using simple SQL expressions.

2. Get data by rows or columns using simple SQL expressions.

3. Save money on data transfer and increase speed.

AWS organisations & Consolidated Billing

1. AWS organisations is an account management service that enables you to consolidate multiple AWS accounts into an organisation that you create and centrally manage.

2. We can have a group within a group and a user can belong to more than one group as they inherit from the super groups.

3. Policy will trickle down to all the other accounts and OU's underneath it.

4. Can do consolidated billing: more that you use, the less that you pay.

5. Paying accounts is independent. Cannot access resources of the other accounts.

6. All linked accounts are independent.

Advantages of Consolidated Billing:

1. One bill per AWS account.

2. Very easy to track charges and allocate costs.

3. Volume pricing discount.

AWS organisations & Consolidated Billing - Exam Tips

1. Always enable multi-factor authentication on root account.

2. Always use a strong and complex password on root account.

3. Paying account should be used for billing purposes only. Do no deploy resources into the paying account.

4. Enable/Disable AWS services using Service Control Policies (SCP) either on OU or on individual accounts.

Lab - Sharing S3 buckets across accounts - Exam Tips:

3 different ways to share S3 buckets across accounts

1. Using Bucket policies & IAM (applies across the entire bucket). Programmatic access only.

2. Using Bucket ACLs & IAM (individual objects). Programmatic access only.

3. Cross-account IAM Roles. Programmatic and Console access.

Lab - AWS Cross-Region Replication - Demo and Exam Tips

1. Versioning must be enabled on both the source and destination buckets.

2. Files in an existing bucket are not replicated automatically.

3. Delete markers are not replicated.

4. Deleting individual versions or delete markets will not be replicated.

5. Understand what Cross Region Replication is at a high level

S3 Transfer Acceleration

S3 Transfer Acceleration utilised the CLoudFront Edge Network to accelerate your uploads to S3. Instead of uploading to your S3 bucket, you can use a distinct URL to upload directly to an edge location which will then transfer that file to S3. You will get a distinct URL to upload to sri-s3-accelarate.amazonaws.com

AWS DataSync: Exam Tips

1. Used to move large amounts of data from on-premises to AWS.

2. Used with NFS and SMB compatible file systems (On premise).

3. Replication can be done hourly, daily or weekly.

4. Install the DataSync agent to start the replication.

5. Can be used to replicate EFS to EFS.

6. AWS DataSync securely connects to Amazon S3, Amazon EFS or Amazon Fsx (Windows File Server) to copy data and metadata to and from AWS.

CloudFront

CloudFront is a content delivery network (CDN) is a system of distributed servers (network) that deliver webpages and other web content to a user based on the geographic locations of the user, the origin of the webpage and a content delivery server.

CloudFront - Key Terminology:

1. Edge location: This is the location where content will be cached. This is separate to an AWS Region/AZ.

2. Origin - This is the origin of all the files that the CDN will distribute. This can be an S3 Bucket, an EC2 Instance, an Elastic Load Balancer or Route 53.

3. Distribution - This is the name given the CDN which consists of a collection of Edge locations.

Amazon CloudFront can be used to deliver your entire website, including dynamic, static, streaming and interactive content using a global network of edge locations.

Requests for your content are automatically routed to the nearest edge location, so content is delivered with the best possible performance.

When the first user queries for a file, it gets downloaded from the server. The second user gets a cached copy from the Edge Location instead of downloading it again from the server. The file has Time to live defined usually 48 hours.

2 types of distribution:

Web Distribution - Typically used for Websites

RTMP - Used for Media Streaming

CloudFront Exam Tips

1. Edge location: This is the location where content will be cached. This is separate to an AWS Region/AZ.

2. Origin - This is the origin of all the files that the CDN will distribute. This can be an S3 Bucket, an EC2 Instance, an Elastic Load Balancer or Route 53.

3. Distribution - This is the name given the CDN which consists of a collection of Edge locations.

2 types of distribution:

Web Distribution - Typically used for Websites

RTMP - Used for Media Streaming

Note:

1. Edge Locations are not just READ only - you can write them too. (ie Put an object on to them).

2. Objects are cached for the life of TTL (Time to live)

3. You can clear cached objects, but you will be charged.

4. We can restrict access using signed URL's (Example: Netflix - Option in AWS CloudFront is "Restrict Viewer access (Use Signed URL's or Signed cookies)")

Create a CloudFront Distribution - Demo

1. Exam Tip: We use Create invalidation to invalidate an object in CloudFront. Example: We pushed out some data but it is not showing up correctly, in order to deal with this we use Invalidations.

2. We need to disable before we delete a CloudFront distribution.

CloudFront Signed URLs and Cookies vs S3 Signed URL

CloudFront Signed URL:

1. A signed URL is for individual files, 1 files = 1 URL.
2. A signed cookie is for multiple file, 1 cookie = multiple files.
3. When we create a signed URL or signed cookie, we attach a policy.

   The policy can include:

   a. URL expiration.

   b. IP ranges

   c. Trusted Signers (which AWS accounts can create signed URL's)
4. Can have different origins. Does not have to be EC2.
5. Key-pair is account wide and managed by the root user.
6. Can utilise caching features.
7. Can filter by date, path, IP address, expiration, etc.

S3 Signed URL:

1. Issues a request as the IAM user who creates the pre-signed URL.
2. Limited lifetime.

Exam Tips:

1. Use signed URLs/cookies when you want to secure content so that only the people you authorise are able to access it.
2. A signed URL is for individual files, 1 files = 1 URL.
3. A signed cookie is for multiple file, 1 cookie = multiple files.
4. If your origin is EC2, then use CloudFront.
5. If your origin is S3, then use S3 signed URL instead of CloudFront Signed URL.

Snowball

What is Snowball?

1. AWS Snowball is a PB-Scale data transport solution that uses secure appliances to transfer large amounts of data in and out of the AWS cloud. Think of it as a gigantic disk to move your data into AWS. Using Snowball addresses common challenges with large-scale data transfers including high network costs, long transfer times, and security concerns.

Transferring data with snowball is simple, fast. Secure and can be as little as one-fifth the cost of high-speed internet.

2. AWS Snowball Edge is a 100TB data transfer device with on-board storage and compute capabilities. You can use Snowball Edge to move large amounts of data into and out of AWS, as a temporary storage tier for large local datasets, or to support local workloads in remote or offline locations.

3. AWS Snowmobile is an Exabyte-scale data transfer service used to move extremely large amounts of data to AWS. You can transfer up to 100PB per Snowmobile, a 45-foot long ruggedised shipping container, pulled by a semi-trailer truck. Snowmobile makes it easy to move massive volumes of data to the cloud, including video libraries, image repositories, or even a complete data centre migration. Transferring data with Snowmobile is secure, fast and cost effective.

- What determines price for Snowball?

1. Service fee per job

   - Snowball 50 TB: $200

   - Snowball 80 TB: $250

2. Snowball uses multiple layers of security designed to protect your data including tamper-resistant enclosures, 256-bit encryption, and an industry-standard Trusted Platform Module (TPM) designed to ensure both security and full chain-of-custody of your data. Once the data transfer job has been processed and verified, AWS performs a software erasure of the Snowball appliance.

3. Daily Charge

   - First 10 days are free, after that it's $15 a day.

4. Data transfer

- Data transfer in to S3 is free. Data transfer out is not.

5. Snowball can Import to S3 and Export from S3.

- When should I use Snowball

Available Internet Connection Theoretical Min. No of days to transfer 100TB at 80% n/w utilisation When to Consider AWS Import/Export Snowball?

T3 (44.736 Mbps)	269 days	2TB or more
100 Mbps	120 days	5TB or more
1000 Mbps	12 days	60TB or more

Storage Gateway

1. AWS Storage Gateway is a service that connects an on-premises software appliance with cloud-based storage to provide seamless and secure integration between an organisation's on-premise IT environment and AWS's storage infrastructure. The service enables you to securely store data to the AWS cloud for scalable and cost effective storage.

2. AWS Storage Gateway's software appliance is available for download as a VM image that you install on a host in your datacenter. StorageGateway supports either VMWare ESXi or Microsoft Hyper-V. Once you've installed your gateway and associated it with your AWS account through the activation process, you can use the AWS management console to create the storage gateway option that is right for you.

3. Three different types of Storage gateway:

a. File Gateway (NFS & SMB): Files are stored as objects in your S3 buckets, accessed through a Network File System (NFS) mount point. Ownership, permissions and timestamps are durably stored in S3 in the user-metadata of the object associated with the file. Once objects are transferred to S3, they can be managed as native S3 objects, and bucket policies such as versioning, lifecycle management and cross-region replication apply directly to objects stored in your bucket.

b. Volume Gateway (iSCSI)

1. Stored Volumes:

- The volume interface presents your applications with disk volumes using the iSCSI block protocol.

- Data written to these volumes can be asynchronously backed up as point-in-time snapshots of your volumes, and stored in the cloud as Amazon EBS snapshots.

- Snapshots are incremental backups that capture only changed blocks. All snapshot storage is also compressed to minimise your storage charges.

- Stored Volumes let you store your primary data locally, while asynchronously backing up the data to AWS. Stored volumes provide your on-premises applications with low-latency access to their entire datasets, while providing durable, off-site backups. You can create storage volumes and mount them as iSCSI devices from your on-premises application servers.

Data written to your stored volumes in stored on your on-premises storage hardware. This data is asynchronously backed up to Amazon Simple Storage Service (Amazon S3) in the form of Amazon Elastic Block Store (Amazon EBS) snapshots. 1 GB - 16 TB in size for Stored Volumes.

2. Cached Volumes: Cached Volumes let you use Amazon Simple Storage Service (Amazon S3) as your primary data storage while retaining frequently accessed data locally in your storage. Cached volumes minimise the need to scale your on-premises storage infrastructure, while still providing your applications with low-latency access to their frequently accessed data,

You can create storage volumes up to 32TB in size and attach to them as iSCSI devices from your on-premises application servers. Your gateway stores data that you write to these volumes in Amazon S3 and retains recently read data in your on-premises storage gateways cache and upload buffer storage. 1GB - 32TB in size for Cached Volumes.

c. Tape Gateway (VTL): Tape Gateway offers a durable, cost-effective solution to archive your data in the AWS Cloud. The VTL interface it provides lets you leverage your existing tape-based backup application infrastructure to store data on virtual tape cartridges that you create on your tape gateway, Each tape gateway is preconfigured with a media changer and tape drives, which are available to your existing client backup applications as iSCSI devices. You add tape cartridges as you need to archive your data. Supported by NetBackup, Backup Exec, Veeam etc.

Exam Tips:

1. Three different types of Storage gateway:

a. File Gateway (NFS & SMB): For flat files, stored directly on S3

b. Volume Gateway (iSCSI)

1. Stored Volumes: Entire Dataset is stored on site and is asynchronously backed up to S3.

2. Cached Volumes: Entire Dataset is stored on S3 and the most frequently accessed data is cached on site.

c. Tape Gateway (VTL)

Athena Vs Macie (Exam Tips)

- What is Athena?

  - Interactive query service which enables you to analyse and query data located in S3 using standard SQL

  - Serverless, nothing to provision, pay per query / per TB scanned

  - No need to set up complex Extract/Transform/Load (ETL) process

  - Works directly with data stored in S3

- Athena can be used for

  1. Can be used to query log files stores in S3. Ex: ELB Logs, S3 access logs ... etc

  2. Generate business reports on data stored in S3.

  3. Analyse AWS cost and usage reports

  4. Run queries on click-stream data.

- What is Macie?

  - Security service which uses machine learning and NLP (natural language processing) to discover, classify and protect sensitive data stored in S3

- Uses AI to recognise if your S3 objects contain sensitive data such as personal identification information (PII).

- Dashboards, reporting and alerts

- Works directly with data stored in S3

- Can also analyse CloudTrail logs

- Great for PCI-DSS and preventing ID theft.

IAM Summary

- IAM is universal, it does not apply to regions at this time.

- The root account is simply the account that's created when you first set up your AWS account and it has complete administrator access.

- New users have no permissions when first created and you'll find this is a theme within Amazon. It's called least privilege.

- So, whenever you create a new user, that user's not going to have any rights or any privileges until you grant them privileges. Likewise when we look at S3, when we create our bucket,

it's locked down, it's not public and making objects public is not that easy. You have to go through a process to do it. So, that's a common theme within Amazon Web Services.

- New users are assigned an access key ID and secret access key when first created.

- These are not the same as a password. You cannot use the access key ID and secret access key to log into the console.

You use it to access AWS via the APIs and the Command Lines, however.

- You only get to view your access key ID and secret access key once. If you lose them you have to regenerate them. So, make sure you save them in a secure location.

- Always setup multi-factor authentication on your root account and you can also create and customize your own password rotation policies.

S3 Summary

- S3 is object based. i.e. allows you to upload files.

- Files can be zero bytes all the way up to 5 terabytes

- There is unlimited storage

- Files are stored in buckets

- S3 is a universal namespace. That is, bucket names must be unique.

- https://s3-eu-west-1.amazonaws.com/acloudguru

- Not suitable to install an operating system on or a database or anything like that.

- Successful uploads will generate HTTP 200 status code.

- By default all newly created buckets are PRIVATE.

- You set up access control to your bucket using

a. Bucket policies and bucket policies are bucket wide

b. Access control lists and these can go down to the individual files or objects in your bucket.

- S3 buckets can be configured to create access logs which logs all requests made to the S3 bucket

and these can be sent to another bucket in the same AWS account or even another bucket in another AWS account.

- The key fundamentals of S3 are:

a. key (This is simply the name of the object.)

b. value (This is simply the data is made up a sequence of bytes). so some sometimes people refer to S3 as a key value pair.

c. Version ID (Important for versioning)

d. Metadata (Data about data you are storing) and we do that through tags and then you get some sub resources such as access control lists and then torrents as well

- Read after write consistency of puts of new objects

- Eventual consistency for overwrite puts and deletes and this can take some time to propagate.

- Exam Tips:

1. S3 Standard: 99.99% availability, 99.99999999999% durability, stored redundantly across multiple devices, and is designed to sustain the loss of 2 facilities

2. S3 - IA (Infrequently accessed): For data that is accessed less frequently, but requires rapid access when needed. Lower fee than S3, but you are charged a retrieval fee.

3. S3 One Zone IA: For where you want a lower cost option for infrequently accessed data, but do not require the multiple availability zone.

4. S3 Intelligent Tiering: Designed to optimise costs by automatically moving data to the most cost-effective access tier, without performance impact or operational overhead.

5. S3 Glacier: S3 Glacier is a secure, durable and a low-cost storage class for data archiving. Retrieval times configurable from mins to hours.

6. S3 Glacier Deep Archive: S3 Glacier Deep Archive is Amazon S3's lowest-cost storage class where a retrieval time of 12 hours is acceptable.

7. You can use bucket policies to make entire S3 buckets public.

8. You can use S3 to host STATIC websites (such as .html). Websites that require database connections such as Wordpress etc cannot be hosted on S3.

9. S3 Scales automatically to meet your demand. Many enterprises will put static websites on S3 when they think there is going to be a large number of requests (such as for a movie preview for example)

- Understand how to get the best value out of S3

1. S3 Standard (Availability: 99.99%)

2. S3 - IA (Availability: 99.9%)

3. S3 - Intelligent Tiering (Availability: 99.9%)

4. S3 One Zone - IA (Don't use this if the data is crucial) (Availability: 99.5%)

5. S3 Glacier - Data Archival (Availability: 99.99%)

6. S3 Glacier Deep Archive (Availability: 99.99%)

- Encryption in transit is achieved by using SSL/TLS.

- S3 also has encryption at rest (Server Side) is achieved by three different ways.

1. S3 managed keys - SSE-S3: This is where S3 just handle all our encryption for us and we don't have to worry about them.

2. AWS key management service or KMS: This is where we can start using keys from the KMS service.

3. Server side encryption with customer provided keys - SSE-C: This is where you provide your keys and you manage the encryption and the actual you know maintenance of those keys.

4. Client-side encryption: This is where you encrypt the objects and then you upload them to S3.

- AWS Organizations: Some best practices with AWS Organizations.

1. Always enable multi-factor authentication on root or master account.

2. Always use strong and complex passwords on root account.

3. Paying account should be used for billing purposes only. Do not deploy resources into the paying account, into the root account or the master account,

4. Enable and disable AWS services using service control policies (SCPs) either on organisational units or on individual accounts.

- three different ways to share S3 buckets across accounts.

1. Using bucket policies and IAM (applies across the entire bucket). Programmatic access only.

2. Using bucket ACLs and IAM (individual objects). Programmatic access only.

3. Cross account IAM Roles, Programmatic and Console access

- Cross Region Replication

1. Versioning to be enabled on both the source and the destination buckets.

2. Files in an existing bucket are not replicated automatically.

3. All subsequent updated files will be replicated automatically.

4. Delete markers are not replicated

5. Deleting individual versions or delete markers will also not be replicated.

6. Understand what cross region replication is at a high level.

- Lifecycle policies

1. Automates moving your objects between the different storage tiers.

2. Used in conjunction with versioning.

3. Can be applied to current versions as well as previous versions.

- S3 Transfer Accelerations.

So, we have our users, they're all around the world. We have our edge locations. Our users will upload their files to the edge locations first and then those files will go over the AWS backbone network to S3. And we saw how mostly it can improve speed and performance. So, if you do need to increase the performance of your, you know, of your users being able to upload files to S3, look at S3 Transfer Acceleration.

- CloudFront.

1. Edge location - This is the location where the content is going to be cached and it's separate to an AWS region or availability zone.

2. Origin - This is the origin of all our files that the CDN will distribute and this can either be an S3 bucket, an EC2 instance, an elastic load balancer or Route 53.

3. Distribution - This is simply the name given to the CDN which consists of a collection of edge locations.

4. We have two different types of distributions.

a. Web distributions - This is typically used for websites

b. RTMP - this is used for Adobe media and it's used for media streaming.

5. Edge locations are not read-only, you can write to them as well, i.e. put an object to them.

6. Objects are cached for the time to live or TTL (Time To Live) and that value is always in seconds

7. You can clear cache objects by invalidating them but you will be charged.

- Snowball: Understand what Snowball is

1. It's a big disk that you can use to move your data in and out of the AWS cloud.

2. Snowball can be imported into S3. So, you can import data into S3.

3. You can also use Snowball to move large amounts of data out of S3.

- Storage Gateway.

1. File Gateway - This is used for flat files that are stored directly on S3 and that's NFS.

2. Volume gateway - That's iSCSI and we have two different types of volume gateways.

a. Stored Volumes - Entire dataset is stored on site so it's literally a 100% copy (asynchronously) of your data being stored on-site and then it's backed up to S3.

b. Cached Volume - This is where the entire data set is stored on S3 and only the most frequently accessed data is cached on site.

3. Gateway virtual tape library - This is used for backups and works with really popular backup applications like NetBackup, Backup Exec, Veeam ... etc.

- Athena (This is a very popular exam topic.)

  1. Athena is an interactive query service.

  2. It allows you to query data located in S3 using standard SQL.

  3. It's serverless

  4. Commonly used to analyse log data stored in S3

- Macie

  1. Macie uses AI to analyse data in S3 and helps to identify personally identifiable information or PII.

  2. It can also be used to analyse CloudTrail logs for suspicious API activity.

  3. It includes dashboards, reports and alerting

  4. it's great for PCI-DSS compliance as well as preventing ID theft.

EC2 101 Elastic Compute Cloud

1. Amazon Elastic Compute Cloud (Amazon EC2) is a web service that provides resizable compute capacity in the cloud. It's just a virtual server (or servers) in the cloud.

2. Amazon EC2 reduces the time required to obtain and boot new server instances to minutes, allowing you to quickly scale capacity, both up and down, as your computing requirements change.

EC2 Pricing Models:

1. On Demand: Allows you to pay a fixed rate by the hour (or by the second) with no commitment.

2. Reserved: Provides you with a capacity reservation, and offer significant discount on the hourly charge for an instance. Contracts are 1 - 3 year terms. Higher discount with upfront payments and longer contracts.

3. Spot: Enables you to bid whatever price you want for instance capacity, providing for even greater savings if your applications have flexible start and end times.

4. Dedicated Hosts: Physical EC2 server dedicated for your use. Dedicated hosts can help reduce the costs by allowing you to use your existing server-bound software licenses.

- On Demand pricing is useful for:

    1. Users that want the low cost and flexibility of Amazon EC2 without any up-front payment for long-term commitment.

    2. Applications with short term, spiky or unpredictable workloads that cannot be interrupted

    3. Applications being developed or tested on Amazon EC2 for the first time.

- Reserved Pricing is useful for:

    1. Applications with steady or predictable usage.

    2. Applications that require reserved capacity.

    3. Users able to make upfront payments to reduce their total computing costs even further.

- Reserved Pricing Types:

    1. Standard Reserved Instances: These offer up to 75% off on demand instances. The more you pay up front and the longer the contract, the greater the discount.

    2. Convertible Reserved Instances: These offer up to 54% off on demand capability to change the attributed of the RI as long as the exchange results in creation of Reserved Instances of equal or greater value.

    3. Scheduled Reserved Instances: These are available to launch within the time windows you reserve. This option allows you to match your capacity reservation to a predictable recurring schedule that only requires a fraction of a day, a week or a month.

- Spot pricing is useful for:

1. Applications that have flexible start and end times.

2. Applications that are only feasible at very low compute prices.

3. Users with urgent computing needs for large amounts of additional capacity.

- Dedicated Hosts pricing is useful for:

1. Useful for regulatory requirements that may not support multi-tenant virtualisation.

2. Great for licensing which doesn't support multi-tenancy or cloud.

3. Can be purchased On-Demand (hourly)

4. Can be purchased as a Reservation for up to 70% off the On-Demand price.

- EC2 instance classes FIGHTDRMCPXZ

1. F1 - For FPGA (Filed Programmable gate array) - Genomics research, financial analytics, real-time video processing, big data etc

2. I3 - For IOPS (High Speed Storage) - NoSQL DBs, Data Warehousing etc

3. G3 - Graphics Intensive (Video Encoding/3D Application Streaming)

4. H1 - High Disk Throughput (MapReduced-based networks, distributed file systems such as HDFS and MapR-FS)

5. T3 - Low Cost and Cheap general purpose (think T2 micro) (Web Servers/Small DB's)

6. D2 - Dense Storage (FileServers/Data Warehousing/Hadoop)

7. R5 - RAM, Memory optimised (Memory Intensive Apps/DBs)

8. M5 - Main choice for general purpose apps (Application Servers)

9. C5 - Compute Optimised (CPU Intensive Apps/DBs)

10. P3 - Graphics/General Purpose GPU (think Pics) (Machine Leaning, Bit Coin Mining etc)

11. X1 - Extreme Memory (SAP HANA/Apache Spark etc)

12. Z1D - Extreme Memory and CPU (Ideal for electronic design automation (EDA) and certain relational database workloads with high per-core licensing costs.)

13. A1 - Arm-based workloads (Scale-out workloads such as web servers)

14. U-6tb1 - Bare Metal (Bare metal capabilities that eliminate virtualisation overhead)

EC2 Exam Tips:

1. EC2 is a compute based service. It is not server less. It's a Server!

2. Use a private key to connect to EC2.

3. Common Ports: Linux (port 22), Microsoft - RDP (port 3389), HTTP (80) and HTTPS (443)

4. Security groups are virtual firewalls. To let everything in 0.0.0.0/0. To let a single IP address in X.X.X.X/32 (32 means this ip address)

5. Always design for failure. Have one EC2 instance in each availability zone.

6. Root device volumes can be encrypted now (a popular exam topic)

7. Termination protection is turned off by default, you must turn it on.

8. On an EBS-backed instance, the default action is for the root EBS volume to be deleted when the instance is terminated. Any additional EBS volumes by default won't be deleted.

9. EBS Root volumes of your DEFAULT AMI's CAN be encrypted. You can also use a third party tool (such as bit locker etc) to encrypt the root volume, or this can be done when creating AMI's (lab to follow) in the AWS console or using the API.

10. Additional volumes can be encrypted as well.

- EC2 Exam Tips: Amazon Elastic Compute Cloud (Amazon EC2) is a web service that provided resizable compute capacity in the cloud. Amazon EC2 reduces the time required to obtain and boot new server instances to minutes, allowing you to quickly scale capacity, both up and down, as your computing requirements change.

1. On Demand: Allows you to pay a fixed rate by the hour (or by the second) with no commitment.

2. Reserved: Provides you with a capacity reservation and offer a significant discount on the hourly charge for an instance. Contract Terms are 1 year or 3 year terms.

3. Spot: Enables you to bid whatever price you want for instance capacity, providing for even greater savings if your application have flexible start and end times.

Imp Note: If the spot instance is terminated by Amazon EC2, you will not be charged for a partial hour of usage. However, if you terminate the instance yourself, you will be charged for any hour in which the instance ran.

4. Dedicated Hosts: Physical EC2 server dedicated for your use. Dedicated Hosts can help reduce costs by allowing you to use your existing server-bound software licenses.

Security Groups Exam Tips:

1. All inbound traffic is blocked by default - so we enable some IP and ports using Security Groups.

2. All outbound traffic is allowed.

3. Security Groups are STATEFUL, when you create an inbound rule and an outbound rule is automatically created.

4. NACL's are STATELESS, when you create an inbound rule and an outbound rule is not automatically created.

5. You CANNOT block specific IP's/Port's using Security Groups instead use Network Access Control Lists.

6. You can have any number of EC2 instances within a security group.

7. You can have multiple Security Groups attached/assigned to EC2 instances.

8. Changes to Security Groups take effect immediately.

9. You can specify allows rule, but not deny rules.

EBS 101

Amazon Elastic Block Storage (EBS) provides persistent block storage volumes for use with Amazon EC2 instances in the AWS cloud.

Each Amazon EBS volume is automatically replicated within it's Availability Zone to protect you from component failure, offering high availability and durability.

- 5 different types of EBS Storage:

SSD:

1. General purpose SSD (GP2) - balances prices and performance for a wide variety of workloads

2. Provisioned IOPS SSD (IO1) - higher-performance SSD volume for mission-critical low-latency or high throughput workloads.

Magnetic/HDD:

1. Throughput Optimised Hard Disk Drive (ST1) - Low cost HDD volume designed for frequently accessed, throughput-intensive workloads.

2. Cold Hard Disk Drive (SC1) - Lowest cost HDD volume designed for less frequently accessed workloads (File Servers)

3. Magnetic - Previous generation and some point will probably be phased out.

- Compare EBS Types (Exam Tip: Important: API names are important for the exam)

https://aws.amazon.com/ebs/features/

EBS Volumes & Snapshots - Lab (Exam Tips)

1. Important: Where-ever you have EC2 instance it's EBS volume will be in the same region.

2. Important: When we terminate EC2 instance, it removes EBS volumes automatically.

3. Volumes exist on EBS. Think of EBS as a virtual hard disk.

4. Snapshots exist on S3. Think of snapshots as a photograph of the disk.

5. Snapshots are point in time copies of Volumes.

6. Snapshots are incremental - this means that only the blocks that have changed since your last snapshot are moved to S3.

7. If this is your first snapshot, it may take some time to create.

8. To create a snapshot for Amazon EBS volumes that serve as a root devices, you should stop the instance before taking the snapshot.

9. However you can take a snap while the instance is running.

10. You can create AMI's from both Volumes and Snapshots.

11. You can change EBS volume sizes on the fly, including changing the size and storage type.

12. Volumes will be in the same Availability Zone as the EC2 instance.

13. To move an EC2 volume from one AZ to another, take a snapshot of it, create an AMI from the snapshot and then use the AMI to launch the EC2 instance in a new AZ.

14. To move an EC2 volume from one region to another, take a snapshot of it, create an AMI from the snapshot and then copy the AMI from one region to other. Then use the copied AMI to launch the new EC2 instance in the new region.

15. Virtualisation Type: PV or HVM, use Hardware-assisted virtualisation as it will give lot more different EC2 instance types.

Linux Amazon Machine Images use one of two types of virtualisation: paravirtual (PV) or hardware virtual machine (HVM). The main differences between PV and HVM AMIs are the way in which they boot and whether they can take advantage of special hardware extensions (CPU, network, and storage) for better performance.

For the best performance, we recommend that you use current generation instance types and HVM AMIs when you launch your instances

16. Snapshots of encrypted volumes are encrypted automatically.

17. Volumes restored from encrypted snapshots are encrypted automatically.

18. You can share snapshots, but only if they are unencrypted.

19. These snapshots can be shared with other AWS accounts or made public.

AMI Types (EBS vs. Instance Store) - Important for the exam

1. You can select your AMI based on:

    1. Region (see Regions and AZ's)

    2. Operating System.

    3. Architecture (32-bit or 64-bit)

    4. Launch Permissions

    5. Storage for the Root Device (Root Device Volume)

        5.1 Instance Store (EPHEMERAL STORAGE)

        5.2 EBS Backed Volumes

2. All AMI's are categorised as either backed by Amazon EBS or backed by Instance Store

2.1 For EBS Volumes: The root device for an instance launched from the Ami is an Amazon EBS volume created from an Amazon EBS snapshot

2.2 For Instance Store Volumes: The root device for an instance launched from the AMI is an instance store volume created from a template stored in Amazon S3.

AMI Types (EBS vs. Instance Store) - Exam Tips

1. Instance Store Volumes are sometimes called Ephemeral Storage (For some reason if the underlying hypervisor is stopped then we are going to loose all our data)

2. You cannot see the volumes of Instance Store EC2 under volumes because its Instance Store.

3. Instance Store volumes cannot be stopped. If the underlying host fails, you will lose your data.

4. EBS backed instances can be stopped. You will not lose the data on this instance if it is stopped.

5. You can reboot both, you will not loose data.

6. By default, both ROOT volumes will be deleted on termination. However, with EBS volumes, you can tell AWS to keep the root device volume.

ENI vs. ENA vs. EFA (Important for the exam - Scenario based questions)

1. ENI: Elastic Network Interface - essentially a virtual network card.

1.1 An ENI is simply virtual network card on your EC2 instance, when you provision an EC2 instance, it's going to have a ENI attached to it automatically, and then you can add additional ones.

1.2 It basically allows a primary private IPv4 address, from the IPv4 address range of your VPC.

1.3 it also allows one or more secondary private IPv4 addresses from the IPv4 address range of your VPC.

1.4 With an ENI you get one elastic IP address, per private IPv4 address

1.5 You get one public IPv4 address

1.6 You get one or more IPv6 addresses.

1.7 One or more security groups

1.8 A MAC address source

1.9 A source/destination check flag

2.0 A description of what the ENI is.

Scenarios for Network Interfaces:

1. You might have multiple ENIs if you want to create a management network and you wanted to have, that separate to your production network.

2. You can also, have an additional ENI if you're using network and security appliances in your VPC.

3. It also allows you by having multiple ENIs you can create dual-homed instances with workloads or roles on distinct subnets.

4. You might have your production subnet, and then you might have your database subnet, and you might want to segregate that by using multiple ENIs, and it allows you to create low budget high availability solutions.

2. EN: Enhanced Networking uses what's called single root I/O virtualisation or SR-IOV, to provide high performance networking capabilities to unsupported instance types.

ENA is a subset of Enhanced networking

2.1 Is uses single root IO virtualisation, or SR-IOV, to provide high performance networking capabilities on supported instance types.

SR-IOV is a method of device virtualisation, that provides higher IO performance and lower CPU utilisation, when compared to traditional network interfaces.

So it's just a way of speeding up your network essentially.

2.2 Enhanced Networking provides higher bandwidth, higher packets per second performance consistently lower it into instance latencies, and there's no additional charge for using Enhanced Networking, but your EC2 instance does have to support it.

2.3 Use Enhanced Networking where you want good network performance.

2.4 Depending on your instance type, Enhance Networking can be enabled using two methodologies,

2.4.1 Elastic Network Adapter or ENA, which supports network speeds of up to 100 gigabits per second for supported instance types.

Or

Intel 82599 Virtual Function, or VF interface, which supports network speeds of up to 10 gigabits per second, for supported instance types.

And this is typically used on older instances.

Tip: In any scenario, question that you get in your exam, you probably want to choose ENA (100 Gbps) over VF (10 Gbps)

3. Elastic Fabric Adapter: A network device that you attach to your EC2 instance to accelerate High Performance Compute, so HPC and machine learning applications.

3.1 An Elastic Fabric Adapter (EFA) is a network device that you can attach to your EC2 instance to accelerate High Performance Computing or HPC, and machine learning applications.

3.2 It's really important, if you get a scenario question and they're talking about ENI versus ENA versus EFA, and they're talking about HPC and machine learning, then you want to choose an Elastic Fabric Adapter.

3.3 Elastic Fabric Adapters, provides lower and more consistent latency and higher throughputs than TCP transport, traditionally used in cloud based, HPC systems.

3.4 EFA can use OS-bypass. OS-bypass enables high performance compute and machine learning applications to bypass the operating system kernel and to communicate directly with the EFA device. It makes it a lot faster with a lot lower latency. However, it's not supported on Windows currently, it's only supported with Linux.

ENI vs. ENA vs. EFA - Exam Tips

In the exam you will be given different scenario questions and you'll be asked to choose whether you should be using an ENI enhance networking or an Elastic Fabric Adapter

1. ENI

For basic networking, perhaps you need a separate management workload to your production network or a separate logging network and you need to do this at a low cost.

In this scenario, just use multiple ENI's for each network.

2. Enhanced Network

For when you need speeds between 10 gigabits per second and 100 gigabits per second anywhere where you need reliable high throughput.

3. Elastic Fabric Adapter

For when you need to accelerate, High Performance Computing, HPC and machine learning applications or if you need to do an OS bypass.

If you see a scenario question mentioning HPC, or machine learning or asking about OS-bypass, then you want to choose an Elastic Fabric Adapter.

Encrypted Root Device Volumes & Snapshots - LAB (Exam Tips)

1. Snapshots of encrypted volumes are encrypted automatically.

2. Volumes resorted from encrypted snapshots are encrypted automatically.

3. You can share snapshots, but only if they are unencrypted.

4. These snapshots can be shared with other AWS accounts or made public.

5. You can now encrypt root device volumes upon creation of the EC2 instance.

6. If for some reason, if we didn't encrypt the root device volume then the process to encrypt is as follows

6.1 Create a Snapshot of the unencrypted root device volume

6.2 Create a copy of the Snapshot and select the encrypt option

6.3 Create an AMI from the encrypted Snapshot.

6.4 Use that AMI to launch new encrypted instances.

Spot Instances & Spot Fleets

1. AMAZON EC2 Spot Instances let you take advantage of unused EC2 capacity in the AWS cloud.

2. Spot instances are available at up to a 90% discount compared to On-Demand prices.

3. You can use Spot Instances for various stateless, fault-tolerant or flexible applications, such as

3.1 Big data and analytics

3.2 Containerised workloads

3.3 CI/CD

3.4 Web Services

3.5 High-performance computing (HPC)

3.6 Image and media rendering

3.7 and other test and development workloads

4. Spot Instances can be used with flexible workloads (can be terminated)

5. Spot Instances are not good for

5.1 Critical applications/Persistent workloads.

5.2 Critical Jobs

5.3 Databases

6. To use Spot Instances, you must first decide on your maximum Spot price. The instance will be provisioned as long as the Spot price is BELOW your maximum Spot price.

7. The hourly Spot price varies depending on capacity and region.

8. If the Spot price goes above your maximum, you have two minutes to choose whether to stop or terminate your instance.

9. You may also use a Spot Block to stop your Spot Instances from being terminated even if the Spot price goes over your max Spot price. You can set Spot blocks for between one to six hours currently.

10. How to terminate Spot Instances: https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/spot-requests.html

Spot Fleets

1. Spot fleets and a spot fleet is just a collection of spot instances and optionally on-demand instances.

2. The spot fleet attempts to launch the number of spot and on-demand instances to meet the target capacity that you specified in the spot fleet requests. The request for spot instances is fulfilled if there's available capacity and if the maximum price that you specified in the request exceeds the current spot price. The spot fleet also attempts to maintain its target capacity fleet,

if your spot instances are interrupted. So it will relaunch those instances.

3. Spot fleets, will try and match the get capacity with your price restraints

3.1 Set up different launch pools and define things like your EC2 instance type, your operating system and Availability Zone

3.2 You can have multiple pools and the fleet will choose the best way to implement depending on the strategy that you define

3.3 Spot fleets will stop launching instances once reach your price threshold or your capacity desire.

4 Strategies: You can have the following different strategies that you can have with spot fleets.

4.1 capacityOptimized: The spot instances comes from the pool with optimal capacity for the number of instances launching. So you're basically guaranteeing that you have a certain amount of capacity.

4.2 lowestPrice: The spot instances come from the pool with the lowest price and that's the default strategy.

4.3 Diversified: The spot instances distributed across all your different pools that you defined

4.4 InstancePoolsToUseCount: The spot instances are distributed across the number of spot instance pools that you specify and the parameter is only valid when used in combination with the lowest price. So this is kind of like a combination of diversified with lowest price, but with diversified, it's using all the pools.

With instance pools to use count you define which pools that you want. It will launch in those pools at the lowest price.

Exam Tips:

1. Spot instances can save you up to 90% of on-demand instances.

2. Useful for any type of computing where you don't need persistent storage.So ephemeral computing, for example

3. You can block spot instances from terminating by using spot block

4. A Spot Fleet is a collection of spot instances and optionally on-demand instances

EC2 Hibernate:

We have learned so far we can stop and terminate EC2 instances. If we stop the instance, the data is kept on the disk (EBS) and will remain on the disk until the EC2 instance is started.

If the instance is terminated, then by default the root device volume will also be terminated.

When we start our EC2 instance, the following happens:

1. OS boots up

2. User data script is run (bootstrap scripts)

3. Applications start (can take some time)

EC2 Hibernate: When you hibernate an EC2 instance, the OS is told to perform hibernation (suspend-to-disk). Hibernation saves the content from the instance memory (RAM) to your Amazon EBS root volume. We persist the instance's Amazon EBS root volume and any attached Amazon EBS data volumes.

When you start your instance out of Hibernation:

1. The Amazon EBS root volume is restored to its previous state

2. The RAM contents are reloaded

3. The processes that were previously running on the instance are resumed.

4. Previously attached data volumes are reattached and the instance retains its instance ID.

With EC2 Hibernate, the instance boots much faster. The OS does not need to reboot because the in-memory state (RAM) is preserved.

This is useful for:

1. Long-running processes

2. Services that take time to initialise

Important: https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/Hibernate.html

Exam Tips:

1. EC2 Hibernate preserves the in-memory RAM on persistent storage (EBS)

2. Much faster to boot up because you do not need to reload the OS

3. Instance RAM must be less than 150 GB

4. Instance families include C3, C4, C5, M3, M4, M5, R3, R4 and R5

5. Available for Windows, Amazon Linux 2 AMI, and Ubuntu

6. Instances can't be hibernated for more than 60 days.

7. Available for On-Demand instances and Reserved Instances.

8. To hibernate an instance, it must first be enabled for hibernation. To enable hibernation, you must do it while launching the instance.

Important: You can't enable or disable hibernation for an instance after you launch it.

CloudWatch 101

1. CloudWatch is to monitor performance

2. Monitors

2.1 EC2 Instances

2.2 Autoscaling Groups

2.3 Elastic Load Balancers

2.4 Route53 Health Checks

3. Storage and Content Delivery

3.1 EBS Volumes

3.2 Storage Gateways

3.3 CloudFront

4. CloudWatch and Ec2: Important: It can monitor host level metrics such as CPU, Network, Status check and Disk

CloudWatch 101 - Exam Tips:

1. CloudWatch is used for monitoring performance.

2. CloudWatch can monitor most of AWS as well as your applications that run on AWS

3. CloudWatch with EC2 will monitor events every 5 mins by default.

4. You can have 1 min interval by turning on detailed monitoring

5. You can create CloudWatch alarms which trigger not monitoring.

6. CloudWatch is all about performance.

7. CloudTrail is all about auditing.

8. Standard Monitoring is 5 mins

9. Detailed monitoring is 1 min with additional cost.

10. Dashboards: Creates dashboards to see what is happening with your AWS environment

11. Alarms: Allows you to set Alarms that notify you when particular thresholds are hit.

12. Events: CloudWatch Events helps you to respond to state changes in your AWS resources.

13. Logs - CloudWatch Logs helps you to aggregate, monitor and store logs.

CloudTrail vs CloudWatch

- CloudWatch monitors performance.

- CloudTrail monitors API calls in the AWS platform.

CloudTrail (Exam Tips)

1. Per AWS account and is enabled per region.

2. Can consolidate logs using S3 bucket:

2.1 Turn on CloudTrail in paying account.

2.2 Create a bucket policy that allows cross-account access.

2.3 Turn on CloudTrail in the other accounts and use the bucket in the paying account.

Using the command line: Exam Tips:

1. You can interact with AWS in 3 different ways

a. Using the console

b. Using the command line interface (CLI)

c. Using the SDK's

2. You will need to set up access in IAM.

3. Command themselves are not in the exam but some basic commands will be useful to know for real life.

Using Roles: Exam Tips

1. Roles are much more secure than using access key id's and secret access keys and are easier to manage.

2. Adding/Modifying policies to roles are instantaneous. You can apply roles to EC2 instances anytime. When you do this, the change takes place immediately.

3. Roles are universal. You do not need to specify what region they are in, similar to users.

IAM Management Roles - Lab (Exam Tips)

1. Roles are more secure than storing your access key and secret access key on individual EC2 instances.

2. Roles are easier to manage.

3. Roles can be assigned to an EC2 instance after it is created using both the console & command line.

4. Roles are universal.

EFS Lab (REPEAT)

EFS stands for Elastic File System and it's a file storage service for Amazon's Elastic Compute Cloud or EC2 instances.

Important: So it's similar to EBS, except where you've got EBS, you can only mount your virtual disk to one EC2 instance and that's just the way it works.

You cannot have an EC2 instance, two EC2 instances sharing an EBS volume.

However, you can have them sharing an EFS volume. So EFS is an easy way to use and provide a simple interface that allows you to create and configure file systems

quickly and easily and with the EFS, storage capacity is elastic, growing and shrinking automatically as you add and remove files so your applications have the storage

they need when they need it.

So it is actually really cool. Basically if you provision an EFS instance,

it will just grow automatically so you could put like one terabyte file on there and then add another terabyte, you don't need to pre-provision storage like you do with EBS.

So EFS is a way, it's great for basically file servers, it's a great way to share files between different EC2 instances. So as always, the best way to learn EFS is to go in and start using it.

What we're going to do is we're going to create two little EC2 instances, we're going to run a bootstrap script on them to install some tools as well as Apache and then what we're going to do is we're going to go and mount this, our var dub-dub-dub HTML directory to an EFS mount point and then that way, we only need one copy of our website or our website will be stored on EFS and when we go and update or make changes to it we're just that those changes are replicated automatically across all our EC2 instances. So if you've got the time, please join me in the AWS console.

For the Lab:

```
#!/bin/bash

yum update -y

yum install http -y

service httpd start

chkconfig httpd on

yum install -y amazon-efs-utils
```

Exam Tips:

1. Supports the Network File System version 4 (NFSv4) protocol
2. You only pay for the storage you use (no pre-provisioning required)
3. Can scale up to the petabytes
4. Can support thousands of concurrent NFS connections.
5. Data is stored across multiple AZ's within a region
6. Read After Write Consistency.

Amazon FSx for Windows and Amazon FSx for Lustre

1. Amazon Fsx for Windows File Server provides a fully managed native Microsoft Windows file system so you can easily move your Windows-based applications that require file storage to AWS.

Amazon FSx is built on Windows Server.

How is Windows FSx different to EFS

- Windows FSx:

1. A managed Windows Server that runs Windows Server Message Block (SMB)-based file services.

2. Designed for Windows and Windows applications.

3. Supports AD users, access control lists, groups and security policies, along with Distributed File System (DFS) namespaces and replication.

- EFS:

1. A managed NFS filer for EC2 instances based on Network File System (NFS) version 4.

2. One of the first network file sharing protocols native to Unix and Linux.

2. Amazon FSx for Lustre is a fully managed file system that is optimised for compute-intensive workloads, such as high-performance computing, machine learning, media data processing workflows, and electronic design automation (EDA).

With Amazon FSx, you can launch and run a Lustre file system that can process massive data sets at up to hundreds of gigabytes per second of throughput, millions of IOPS and sub-millisecond latencies.

How is Amazon FSx for Lustre different to EFS

- Lustre FSx:

1. Designed specifically for fast processing of workloads such as machine learning, high performance computing (HOC), video processing, financial modelling and electronic design automation (EDA).

2. Lets you launch and run a file system that provides sub-millisecond access to your data and allows you to read and write data at speeds of up to hundreds of gigabytes per second of throughput and millions of IOPS.

- EFS:

1. A managed NFS filer for EC2 instances based on Network File System (NFS) version 4.

2. One of the first network file sharing protocols native to Unix and Linux.

* Exam Tips: In the exam you'll be given different scenarios and asked to choose whether you should use an EFS, FSx for Windows or FSx for Lustre

1. EFS:

When you need distributed, highly resilient storage for Linux instances and Linux-based applications.

2. Amazon FSx for Windows:

When you need centralised storage for Windows-based applications such as Sharepoint, Microsoft SQL Server, Workspaces, IIS Web Server or any other native Microsoft Application.

3. Amazon FSx for Lustre:

When you need high-speed, high-capacity distributed storage. This will be for applications that do High CPU, financial modelling etc. Remember that FSx for Lustre can store data directly to S3.

EC2 Placement Groups

Three types of Placement groups:

1. Clustered Placement Groups: A cluster placement group is grouping of instances within a single AZ.

Placement groups are recommended for applications that need low latency, high n/w throughput, or both.

Only certain instances can be launched in to a Clustered Placement Group.

2. Spread Placement Groups: A spread placement group is a group of instances that are each placed on distinct underlying hardware.

Spread placement groups are recommended for applications that have a small number of critical instances that should be kept separate from each other.

THINK INDIVIDUAL INSTANCES.

3. Partitioned Placement Groups: When using partition placement groups, Amazon divides each group into logical segments called partitions.

Amazon EC2 ensures that each partition with a placement group has its own set of racks.

No two partitions within a placement group share the same racks, allowing you to isolate the impact of h/w failure within your application.

THINK MULTIPLE INSTANCES.

Use case: HDFS, HBase, Hadoop, Cassandra and Kafka

EC2 Placement Groups - Exam Tips

1. A Clustered placement group can't span multiple AZ's.

2. A spread placement and partitioned can.

3. The name you specify for a placement group must be unique within your AWS account.

4. Only certain types of instances can be launched in a placement group (Compute Optimised, GPU, Memory Optimised, Storage Optimised)

5. AWS recommend homogeneous instances within clustered placement groups.

6. You can't merge placement groups.

7. You can move an existing instance into a placement group. Before you move the instance, the instance must be in the stopped state.

You can move or remove an instance using the AWS CLI or AWS SDK, you can't do it via the console yet.

HPC on AWS - Exam Tips:

It's never been easier to get started with high-performance compute (HPC) that in any other time in history - and AWS is the perfect place to perform it.

You can create a large number of resources in almost no time. You only pay for the resources yo use - and once finished, you can destroy the resources.

HPC is used for industries such as genomics, finance and financial risk modelling, machine learning, weather prediction and even autonomous driving.

What are the different services we can use to achieve HPC on AWS?

1. Data Transfer:

a. Snowball, Snowmobile (terabytes/petabytes worth of data)

b. AWS Data Sync to store S3, EFS, Fsx for windows ... etc

c. AWS Direct Connect: Is a cloud service solution that makes it easy to establish a dedicated network connection from your premises to AWS.

Using AWS Direct Connect, you can establish private connectivity between AWS and your data centre, office or colocation environment - which, in many cases,

can reduce your network costs, increase bandwidth throughput and provide a more consistent network experience than internet-base connections.

2. Compute and networking:

a. EC2 instances that are GPU or CPU optimised

b. EC2 fleets (Spot Instances or Spot Fleets)

c. Placement Groups (Cluster Placement groups)

d. Enhanced networking

1. It uses single root I/O virtualisation (SR-IOV) to provide high-performance n/w capabilities on supported instance types.

SR-IOV is a method of device virtualisation that provides high I/O performance and lower CPU utilisation when compared to traditional virtualised n/w interfaces.

2. Enhanced networking provides higher bandwidth, higher packet per second (PPS) performance, and consistency lower inter-instance latencies.

There is no additional charge for using enhanced networking.

3. Use where you want for network performance.

e. Elastic Network Adapters

2. Depending on your instance type, Enhance Networking can be enabled using two methodologies,

1.1 Elastic Network Adapter or ENA, which supports network speeds of up to 100 gigabits per second for supported instance types.

Or

Intel 82599 Virtual Function, or VF interface, which supports network speeds of up to 10 gigabits per second, for supported instance types.

And this is typically used on older instances.

Tip: In any scenario, question that you get in your exam, you probably want to choose ENA (100 Gbps) over VF (10 Gbps),

f. Elastic Fabric Adapters

3. Elastic Fabric Adapter: A network device that you attach to your EC2 instance to accelerate High Performance Compute, so HPC and machine learning applications.

3.1 An Elastic Fabric Adapter (EFA) is a network device that you can attach to your EC2 instance to accelerate High Performance Computing or HPC, and machine learning applications.

3.2 It's really important, if you get a scenario question and they're talking about ENI versus EFA versus EFA, and they're talking about HPC and machine learning, then you want to choose an Elastic Fabric Adapter.

3.3 Elastic Fabric Adapters, provides lower and more consistent latency and higher throughputs than TCP transport, traditionally used in cloud based, HPC systems.

3.4 EFA can use OS-bypass. OS-bypass enables high performance compute and machine learning applications to bypass the operating system kernel and to communicate directly with the EFA device. It makes it a lot faster with a lot lower latency. However, it's not supported on Windows currently, it's only supported with Linux.

3. Storage: What are the storage services that allow us to achieve HPC on AWS

3.1 Instance-attached storage:

3.1.1 EBS: Scale up to 64K IOPS with Provisioned IOPS (PIOPS)

3.1.2 Instance Store: Scale to millions of IOPS; low latency

3.2 Network Storage:

3.2.1 Amazon S3: Distributed object-based storage; not a file system

3.2.2 Amazon EFS: Scale IOPS based on total size, or use provisioned IOPS

3.2.3 Amazon FSx for Lustre: HPC-optimised distributed file system; millions of IOPS, which is also backed by S3.

4. Orchestration and automation:

4.1 AWS Batch

4.1.1 AWS Batch enables developers, scientists and engineers to easily and efficiently run hundreds of thousands of batch computing jobs on AWS.

4.1.2 AWS Batch supports multi-node parallel jobs, which allows your to run a single job that spans multiple EC2 instances.

4.1.3 You can easily schedule jobs and launch EC2 instances according to your needs.

4.2 AWS Parallel Cluster

4.2.1 Open-source cluster management tools that makes it easy for you to deploy and manage HPC clusters on AWS.

4.2.2 Parallel Cluster uses a simple text file to model and provision all the resources needed for your HPC applications in an automated and secure manner.

4.2.3 Automate creation of VPC, subnet, cluster type and instance types.

AWS WAF

1. AWS WAF is a web application firewall that lets you monitor the HTTP and HTTPS requests that are forwarded to Amazon CloudFront, an application load balancer (Layer 7) or API Gateway.

2. AWS WAF also lets you control access to your content (WAF can see Query String parameters, info sent to web servers)

3. You can configure conditions such as what IP addresses are allowed to make this request or what query string parameters need to be passed for the request to be allowed

4. Then the application load balancer or CloudFront or API Gateway will either allow this content to be received or to give a HTTP 403 status code.

At its most basic level, AWS WAF allows 3 different behaviours

1. Allow all request except the ones you specify.

2. Block all request except the ones you specify.

3. Count the requests that match the properties you specify.

Extra protection against web attacks using conditions you specify. You can define conditions by using characteristics of web requests such as:

1. IP addresses that requests originate from.

2. Country that requests originate from.

3. Values in request headers.

4. Strings that appear in requests, either specific strings or string that match regular expressions

5. Length of requests.

6. Presence of SQL code that is likely to be malicious (known as SQL injection)

7. Presence of a script that is likely to be malicious (known as cross-site scripting)

In the exam you will be given different scenarios and you will be asked how to block malicious IP addresses

a. Use AWS WAF

b. Use Network ACL's - Covered under VPC section of this course.

Database 101

1. Relational databases on AWS - RDS: (Exam Tips)

    1.1 SQL Server

    1.2 Oracle

    1.3 Postgres

    1.4 MySQL

1.5 Aurora

1.6 MariaDB

2. Exam Tips: RDS has two features

2.1 Multi-AZ - For DR

2.2 Read Replicas - For performance (Allows 5 copies of read replica)

3. Non Relational Databases (JSON/No SQL) called DynamoDB

3.1 Collection = Table, Document = Row and "Key Value pairs" = Fields.

3.2 The columns in the table can vary, this will not affect other rows in the DB

4. OLTP (Online transaction processing) will differ from OLAP (Online analytics processing) in terms of the type of query you will run.

5. AWS data warehouse database called Redshift, uses a different type of architecture from a DB perspective and infrastructure layer.

6. ElastiCache is a web service that makes it easy to deploy, operate and scale as in-memory cache in the cloud. The service improves the performance of web applications

by allowing you to retrieve the information fast, managed, in-memory cache instead of relying entirely on slower disk-based databases.

6.1 ElastiCache supports two open-source in-memory caching engines

6.1.1 Memcached

6.1.2 Redis

7. RDS runs on Virtual Machines

8. You cannot log in to these OS however.

9. Patching of the RDS OS and DB is Amazon's responsibility.

10. RDS is NOT Server-less. (Aurora is the only one, which is server-less)

RDS - Back Ups, Mutli-AZ & Read Replicas

1. Automated Backups allow you to recover your database to any point in time within a "retention period".

The retention period can be between one and 35 days. Automated Backups will take a full daily snapshot and will also store transaction logs throughout the day.

When you do a recovery, AWS will first choose the most recent daily back up , and then apply transaction logs relevant to that day. This allows you to do a point in time

recovery down to a second, within the retention period.

2. Automated Backups are enabled by default. The backup data is stored in S3 and you get free storage space equal to the size of your database. So if you have an RDS instance of 10GB, you will get 10GB worth of storage

3. Backups are taken within a defined window, During the backup window, storage I/O may be suspended while your data is being backed up and you may experience elevated latency.

4. DB Snapshots are done manually (i.e they are user initiated.) They are stored even after you delete the original RDS instance, unlike automated backups.

5. Whenever you restore either an Automatic Backup or a manual Snapshot, the restored version of the database will be a new RDS instance with a new DNS endpoint.

6. Encryption at rest is supported for MySQL, Oracle, SQL Server, PostgreSQL, MariaDB & Aurora. Encryption is done using the AWS Key Management Service (KMS) Service.

Once your RDS instance is encrypted, as are its automated backups, read replicas, and snapshots.

7. Multi-AZ allows you to have an exact copy of your production database in another AZ. AWS handles the replication for you, so when your production database is written to, this write will automatically synchronised to the stand by database.

In the event of planned database maintenance, DB Instance failure, or an AZ failure, Amazon RDS will automatically failover to the standby so that database operations can resume quickly without administrative intervention.

8. Multi-AZ is for DR only. It's NOT primarily used for improving performance. For perfomance improvement, you need Read replicas.

9. Multi-AZ is available for the following databases

- SQL Server
- Oracle
- MySQL Server
- PostgreSQL
- MariaDB
- Aurora is fault tolerant based on it's architecture.

10. A Read Replica allows you to have a read-only copy of your production database. This is achieved by using Asynchronous replication from the primary RDS instance to the read replica.

You use read replicas primarily for very read-heavy database workloads.

11. In the exam: How can you improve performance of the DB

1. By using Read Replicas
2. Using Elasti Cache

12. Read Replicas are available for the following databases

- SQL Server
- Oracle
- MySQL Server
- PostgreSQL
- MariaDB
- Aurora

13. Things to know about Read Replicas:

- Used for scaling, not for DR!
- Must have automatic backups turned on in order to deploy a read replica.
- You can have up to 5 read replica copies of any database.
- You can have read replicas of read replicas (but watch out for latency)
- Each read replica will have its own DNS end point.
- You can have read replicas that have Multi-AZ.
- You can create read replicas of Multi-AZ source databases.
- Read replicas can be promoted to be their own databases. This breaks the replication.
- You can have a read replica in a second region.

Exam Tips:

1. There are two different types of Backups for RDS

a. Automated Backups

b. Database Snapshots

2. Read Replicas

- Can be Multi-AZ.
- Used to increase performance.
- Must have backups turned ON.
- Can be in different regions.
- Can be MySQL, PostgreSQL, MariaDB, Oracle, Aurora and MSSQL.
- Can be promoted to master, this will break the Read Replica

3. Multi-AZ

- Used for DR.
- You can force a fail-over from AZ to another by rebooting the RDS instance.

4. Encryption at rest is supported for MySQL, Oracle, SQL Server, PostgreSQL, MariaDB & Aurora.

- Encryption is done using the AWS Key Management Service (KMS). Once your RDS instance is encrypted, the data stored at rest in the underlying storage is encrypted,

As are its automated backups, read replicas, and snapshots.

DynamoDB (NoSQL DB) - Exam Tips

1. Amazon DynamoDB is a fast and flexible NoSQL database service for all applications that need consistent, single-digit millisecond latency at any scale.

It is fully managed database and supports both document and key-value data models. Its flexible data model and reliable performance make it a great fit for mobile, web, gaming, ad-tech, IoT, and many other applications.

2. The basics of DynamoDB are as follows:

- Stored on SSD Storage.
- Spread across 3 geographically distinct data centers.
- Eventual Consistent Reads (Default).

3. Eventual Consistent Read (Default): Consistency across all copies of data is usually reached within a second. Repeating a read after a short time should return the updated data. (Best Read Performance)

4. Strongly Consistent Reads: A strongly consistent read returns a result that reflects all writes that received a successful response prior to the read.

Advanced DynamoDB

1. DynamoDB Accelerator (DAX)

- Fully Managed, highly available, in-memory cache
- 10x performance improvement
- Reduces request time from milliseconds to microseconds - even under load.
- No need for developers to manage caching logic

- Compatible with DynamoDB API calls.

- DAX not only gives read performance improve but also gives write performance improvement.

- DAX provides multiple "all-or-nothing" operations such as Financial transactions or fulfilling orders.

- Two underlying reads/writes per transaction - prepare/commit, so DynamoDB is going to consume more of the resources.

- Up to 25 items or 4 MB of data.

On Demand Capacity:

- Pay-per-request pricing

- Balance cost and performance

- No min capacity

- No charge for read/write - only storage and backups

- You might think that this is better but you pay more per request than with provisioned capacity

- Use for new product launches.

On-Demand Backup and Restore

- Full backups at any time

- Zero impact on table performance or availability

- Consistent within seconds and retained until deleted

- Operates within same region as the source table.

Point-in-Time recovery

- Protects against accidental writes or deletes

- Restore to any point in the last 35 days

- Incremental backups

- Not enabled by default.

- Latest restorable: five mins in the past

Streams

- Time-ordered sequence of item-level changes in a table

- Stored for 24hours

- Inserts, updates and deletes.

- Streams records are organised in Shards

- Combine with Lambda functions for functionality like stored procs.

Manage Multi-Master, Multi-Region Replication (Global Tables)

- Globally distributed applications

- Based on DynamoDB Streams

- Multi-region redundancy for DR or HA

- No application rewrites and fully managed by AWS.

- Replication latency under one second

Demo of Global Tables replication between two Regions

Database Migration Service

Source: Aurora, S3, DB2, MariaDB, AzureDb, SQL Server, MongoDB, MySQL, Oracle, PostgreSQl and Sybase (at the time of the recording, DynamoDB isn't in the source, check AWS documentation for the latest source DB list)

Destination: Aurora, DocumentDB, DynamoDB, Kinesis, Redshift, S3, ElasticSearch, Kafka, MariaDB, MySQL, Oracle, PostgreSQl and Sybase

At the time of migration , source DB is completely operational

- Security

    - Encryption at rest using KMS

    - Site-to-Site VPN

    - Direct Connect (Dx)

    - IAM policies and roles

- Fine-grained access (use IAM policy to allow access to certain attributes of the DynamoDB table items)

- Use CloudWatch and CloudTrail to monitor

- VPC endpoints: For EC2 instances to access DynamoDB

Redshift

1. Amazon Redshift is a fast and powerful, fully managed, petabyte-scale data warehouse service in the cloud.

Customers can start small for just $0.25 per hour with no commitments or upfront costs and scale to a petabyte or more for $1000 per terabyte per year, less than a tenth of most other data warehousing solutions.

OLTP vs OLAP:

- OLAP transaction example: Net Profit for EMEA and Pacific for the digital radio product, Pulls in large number of records

- Data Warehousing databases use different type of architecture both from a database perspective and infrastructure layer.

Redshift can be configured as follows:

- Single Node (160Gb)

- Multi-Node

1. Leader Node (manages client connections and receives queries)

2. Compute Node (store data and perform queries and computations). Up to 128 Compute Nodes.

Redshift uses Advanced compression:

- Columnar data stores can be compressed much more than row-based data stores because similar data is stored sequentially on disk.

- Amazon Redshift employs multiple compression techniques and can often achieve significant compression relative to traditional relational data stores.

In addition, Amazon Redshift doesn't require indexes or materialised views, and so uses less space than traditional relational database systems.

When loading data into an empty table, Amazon Redshift automatically samples your data and selects the most appropriate compression scheme.

Massive Parallel Processing (MPP):

- Amazon Redshift automatically distributes data and query load across all nodes. Amazon Redshift makes it easy to add nodes to your data warehouse and enables you to maintain fast query performance as your data warehouse grows.

Backups:

- Enabled by default with a 1 day retention period.
- Maximum retention period is 35 days.
- Redshift always attempts to maintain at least three copies of your data (the original and replica on the compute nodes and a backup in Amazon S3).
- Redshift can also asynchronously replicate your snapshots to S3 in another region for DR.

Pricing

- Compute Node Hours (total no of hours you run across all your compute nodes for the billing period. You are billed for 1 unit per node per hour,

  so a 3-node data warehouse cluster running persistently for an entire month would incur 2,160 instance hours. You will not be charged for leader node hours; only compute nodes will incur charges.)
- Backup
- Data Transfer (only within a VPC, not outside of it)

Security Considerations

- Encrypted in transit using SSL
- Encrypted at rest using AES-256 encryption
- By default Redshift takes care of key management
  - Manage your own keys through HSM
  - AWS Key Management Service

Availability

- Currently only in 1 AZ (check AWS to confirm for the latest)
- Can restore snapshots to new AZs in the event of an outage

Exam Tips:

- Redshift is used for Business Intelligence.
- Available in only 1 AZ.
- Enabled by default with a 1 day retention period.
- Maximum retention period is 35 days.
- Redshift always attempts to maintain at least three copies of your data (the original and replica on the compute nodes and a backup in Amazon S3).
- Redshift can also asynchronously replicate your snapshots to S3 in another region for DR.

Aurora

1. Amazon Aurora is a MySQL and PostgresSQL-compatible relational database engine that combines the speed and availability of high-end commercial databases with the simplicity and cost-effectiveness of open source databases.

2. Amazon Aurora provides up to 5x better performance than MySQL and 3x better performance than PostgreSQL databases at a much lower price point, whilst delivering similar performance and availability.

3. Things to know about Aurora

- Start with 10Gb, Scales in 10 GB increments to 64 TB (Storage Autoscaling)
- Compute resources can scale up tp 32vCPUS and 244GB of Memory
- 2 copies of your data is contained in each availability zone, with minimum of 3 availability zones. 6 copies of your data.

4. Scaling Aurora

- Aurora is designed to transparently handle the loss of up to two copies of data without affecting database write availability and up to three copies without affecting read availability.

- Aurora storage is also self-healing. Data blocks and disks are continuously scanned for errors and repaired automatically.

5. Three types of Aurora Replicas are available:

   1. Aurora Replicas (currently 15)

   2. MySQL Read Replicas (currently 5)

   3. PostgreSQL (currently 1)

6. What kind of replicas does Aurora support? on https://aws.amazon.com/rds/aurora/faqs/

7. Backups with Aurora

   - Automated backups are always enabled on Amazon Aurora DB Instances. Backups do not impact performance.

   - You can also take snapshots with Aurora. This also does not impact on performance.

   - You can share Aurora snapshots with other AWS accounts

8. Amazon Aurora Serverless is an on-demand, autoscaling configuration for the MYSQL-compatible and PostgreSQL-compatible editions of Amazon Aurora.

   - An Aurora Serverless DB cluster automatically starts up, shuts down, and scales capacity up or down based on your application's needs.

9. Amazon Aurora Serverless provides a relatively simple cost-effective option for infrequent, intermittent, or unpredictable workloads.

Exam Tips:

- 2 copies of your data is contained in each availability zone, with minimum of 3 availability zones. 6 copies of your data.

- You can share Aurora snapshots with other AWS accounts

- Three types of Aurora Replicas are available:

  1. Aurora Replicas (currently 15)
  2. MySQL Read Replicas (currently 5)
  3. PostgreSQL (currently 1)

- Automated failover is only available with Aurora Replicas

- Automated backups are always enabled on Amazon Aurora DB Instances. Backups do not impact performance.

- You can also take snapshots with Aurora. This also does not impact on performance.

- You can share Aurora snapshots with other AWS accounts

- Use Amazon Aurora Serverless if you want a simple cost-effective option for infrequent, intermittent, or unpredictable workloads.

Elasti Cache - Exam Tips

1. Elasti Cache is a web service that makes it easy to deploy, operate and scale an in-memory cache in the cloud.

The service improves the performance of web applications by allowing you to retrieve information from fast, managed, in-memory caches, instead of relying entirely on slower disk-based databases

2. Elasti Cache supports two open-source in-memory caching engines

- Memcached
- Redis

3. Memcached vs Redis: https://aws.amazon.com/elasticache/redis-vs-memcached/ (not required for SAA exam)

Multithreaded architecture is the only difference, which Memcached supports. Everything else is supported by Redis.

Exam Question: How to improve DB performance

1. To use Read Replicas

2. To use Elasti Cache

4. Redis is Multi-AZ

5. You can do backups and restores of Redis.

Database Migration Service

1. AWS DMS is a cloud service that makes it easy to migrate relational databases, data warehouses, NoSQL databases and other types of data stores.

You can use AWS DMS to migrate your data into the AWS Cloud, between on-premises instances (through an AWS cloud setup), or between combinations of cloud and on-premises setups.

2. How does DMS work?

- At its most basic level, AWS DMS is a server in the AWS cloud that runs replication software.

1. Create a source and a target endpoints.

2. Schedule/Run a Replication Task (Replication Instance - VM) to move the data

- AWS DMS creates the tables and associated primary keys if they don't exist on the target.

- You can pre-create the target tables manually, or you can use AWS Schema Conversion Tool (SCT) to create some or all of the target tables, indexes, views, triggers, etc.

3. Types of DMS Migrations:

- Supports homogenous migrations: (no need SCT)

- Oracle to Oracle

- Support Heterogeneous migrations: (need SCT)

- MS SQL to Amazon Aurora

4. Sources and Targets: The source can either be on-premises or inside AWS itself or another cloud provider such as Azure.

https://docs.aws.amazon.com/dms/latest/userguide/CHAP_Source.html

https://docs.aws.amazon.com/dms/latest/userguide/CHAP_Target.html

Note:

AWS DMS doesn't support migration across AWS Regions for the following target endpoint types:

Amazon DynamoDB

Amazon Elasticsearch Service

Amazon Kinesis Data Streams

Caching Strategies on AWS

Caching is a balancing act between up-to-date, accurate information and latency.

1. The following services have caching capabilities

- CloudFront
- API Gateway
- ElastiCache - Memcached and Redis
- DynamoDB Accelerator (DAX)

Elastic Map Reduce (EMR) Overview - Exam Tips

1. Amazon EMR is the industry-leading cloud big data platform for processing vast amounts of data using open-source tools such as Apache Spark, Apache Hive, Apache HBase, Apache Flink, Apache Hudi and Presto.

With EMR, you can run petabyte-scale analysis at less than half the cost of traditional on-premises solutions and over three times faster than standard Apache Spark.

2. The central component of Amazon EMR is the cluster. A cluster is a collection of Amazon EC2 instances. Each instance in the cluster is called a node.

Each node has a role within the cluster, referred to as the node type.

Amazon EMR also installs different software components on each node type, giving each node a role in a distributed application like Apache Hadoop.

3. The node types in Amazon EMR are as follows:

- Master node:

A node that manages the cluster. The master node tracks the status of tasks and monitors the health of the cluster. Every cluster has a master node.

By default - log data is stored on the master node.

- Core node:

A node with software components that runs tasks and stores data in the Hadoop Distributed File System (HDFS) on your cluster.

Mutli-node clusters have at least one core node.

- Task node: A node with software components that only runs tasks and does not store data in HDFS. Task nodes are optional.

Exam Question: With EMR, if we lose the master node, we lose everything. So how we save the logs of /mnt/var/log

- You can configure a cluster to periodically archive the log files stored on the master node to Amazon S3. This ensures the log files are available after the cluster terminates, whether this is through normal shutdown or due to an error.

- Amazon EMR archives the log files to Amazon S3 at five-minute intervals.

- We can set this up only when we create the cluster and not after creating the cluster.

Database Summary

1. Relational databases on AWS - RDS:

    1.1 SQL Server

    1.2 Oracle

    1.3 PostgreSQL

    1.4 MySQL

    1.5 Aurora

    1.6 MariaDB

2. DynamoDB

3. Redshift OLAP

4. Elasticache

    4.1 Memcached

    4.2 Redis

5. Remember the following points:

    5.1 RDS runs on virtual machines

    5.2 You cannot log in to these operating systems however

    5.3 Patching of the RDS OS and DB is Amazon's responsibility

    5.4 RDS is not serverless

    5.5 Aurora Serverless and DynamoDB are serverless

6. There are two different types of Backups for RDS:

    6.1 Automated Backups

    6.2 Database Snapshots

7. Read Replicas

7.1 Can be Multi-AZ

7.2 Used to increase performance

7.3 Must have backups turned on.

7.4 Can be in different regions

7.5 Can be MySQL, PostgreSQL, MariaDB, Oracle and Aurora

7.6 Can be promoted to master, this will break the Read Replica.

8. MultiAZ

8.1 Used for DR

8.2 You can force a failover from one AZ to another by rebooting the RDS instance.

9. Encryption at rest is supported for MySQL, Oracle, SQL Server, PostgreSQL, MariaDB & Aurora.

Encryption is done using the AWS Key Management Service (KMS). Once your RDS instance is encrypted, the data stored at rest in the underlying storage is encrypted,

As are its automated backups, read replicas, and snapshots.

10. DynamoDB

- Stored on SSD Storage.

- Spread across 3 geographically distinct data centers.

- Eventual Consistent Reads (Default): Consistency across all copies of data is usually reached within a second. Repeating a read after a short time should return the updated data. (Best Read Performance)

- Strongly Consistent Reads: A strongly consistent read returns a result that reflects all writes that received a successful response prior to the read.

11. Redshift

- Redshift is used for Business Intelligence.

- Available in only 1 AZ.

- Enabled by default with a 1 day retention period.

- Maximum retention period is 35 days.

- Redshift always attempts to maintain at least three copies of your data (the original and replica on the compute nodes and a backup in Amazon S3).

- Redshift can also asynchronously replicate your snapshots to S3 in another region for DR.

12. Aurora

- 2 copies of your data is contained in each availability zone, with minimum of 3 availability zones. 6 copies of your data.

- You can share Aurora snapshots with other AWS accounts

- Three types of Aurora Replicas are available:

  1. Aurora Replicas (currently 15)

  2. MySQL Read Replicas (currently 5)

  3. PostgreSQL (currently 1)

- Automated failover is only available with Aurora Replicas

- Automated backups are always enabled on Amazon Aurora DB Instances. Backups do not impact performance.

- You can also take snapshots with Aurora. This also does not impact on performance.

- You can share Aurora snapshots with other AWS accounts

- Use Amazon Aurora Serverless if you want a simple cost-effective option for infrequent, intermittent, or unpredictable workloads.

13. Elasticache

- Use Elasticache to increase database and web application performance.

- Redis is Multi-AZ

- You can do backups and restore of Redis

- If you need to scale horizontally, use Memcached

Advanced IAM

AWS Directory Service

- Family of managed services
- Connect AWS resources with on-premises AS
- Standalone directory in the cloud
- Use existing corporate credentials
- SSO to any domain-joined EC2 instance

What is Active Directory?

- On-premises directory service
- Hierarchical database of users, groups, computers - trees and forests
- Group Policies
- LDAP and DNS
- Kerberos, LDAP and NTLM authentication
- Highly available

AWS Managed Microsoft AD

- AD domain controllers (DCs) running Windows server
- Reachable by application in your VPC
- Add DCs for HA and performance
- Exclusive access to DCs
- Extend existing AD to on-premises using AD Trust
- AWS manages
    1. Mutli-AZ deployment
    2. Patch, monitor, recover
    3. Instance rotation
    4. Snapshot and restore
- Customer manages
    1. Users, Groups and GPOs
    2. Standard AD tools
    3. Scale out DCs
    4. Trusts (resource forest)
    5. Certificate authorities (LDAPS)
    6. Federation

- Simple AD
    1. Standalone managed directory
    2. Basic AD features

3. Small: <= 500; Large <= 5K users

4. Easier to manage EC2

5. Linux workloads that need LDAP

6. Does not support trusts (can't join on-premises AD)

- AD Connector

1. Directory gateway (proxy) for on-premises AD

2. Avoid caching information in the cloud

3. Allow on-premises users to log in to AWS using AD

4. Join EC2 instances to your existing AD domain

5. Scale across multiple AD Connectors.

- Cloud Directory

1. Directory-based store for developers

2. Multiple hierarchies with hundreds of millions of objects

3. Use cases: org charts, course catalogs, device registries

4. Full managed service

- Amazon Cognito User Pools

1. Managed user directory for SaaS applications

2. Sign-up and Sign-in for web or mobile

3. Works with social media identities

Exam Tips:

AD Compatible:

- Managed Microsoft AD (a.k.a. Directory Service for Microsoft AD)

- AD Connector

- Simple AD

These allow to logon to Amazon workspaces and QuickSight with your AD credentials

Exam Question: Logging into workspaces or QuickSight with AD credentials, think about the AD compatible services here.

Not AD Compatible

- Cloud Directory

- Cognito User pools

If you are a developer and you don't need AD, you can use Cloud Directory to create directories that organise and manage hierarchical information and

Cognito user pools work with mobile and web

IAM Policies - this lesson cover IAM boundaries and how to IAM evaluates policies

* Amazon Resource Name (ARN)

- ARN's are unique and all ARNs begin with arn:partition:service:region:account_id

- and end with:

resource

resource_type/resource

resource_type/resource/qualifier

resource_type/resource:qualifier

resource_type:resource

resource_type:resource:qualifier

Examples:

arn:aws:iam::123456789012:user/sri <-- :: region is omitted because IAM is global arn:aws:s3:::my_awesome_bucket/image.png <-- ::: - no region, no account id needed to identify an object in S3. all objects in S3 are globally unique

arn:aws:dynamodb:us-east-1:123456789012:table/orders

arn:aws:ec2:us-east-1:123456789012:instance/* - wildcard. EC2 is a regional service

* IAM Policies

- JSON document that defines permissions
- Identity policy
- Resource policy
- No effect until attached to an identity or a resource
- A policy is a list of statements
- Each statement matches an AWS API request
- Statement has Sid, Effect: Allow/Deny, Action is the API and Resource is the action is against.
- Inline policy is just limited to the role and not available outside that role, typically it's not a best practise to use.

Exam Tips:

- Anything not explicitly allowed == implicitly denied
- Explicit deny > everything else (overrides any other allows)
- Only attached policies have effect
- AWS joins all applicable policies. (When we use multiple policies)
- AWS-managed vs. Customer-managed

Permission Boundaries

- Used to delegate administration to other users.
- Prevent privilege escalation or unnecessarily broad permissions
- Control max permissions an IAM policy can grant

- Use Cases

1. Developers creating roles for LAMBDA functions
2. Application owners creating roles for EC2 instances
3. Admins creating ad hoc users

AWS Resource Access Manager (RAM)

- AWS Resource Access Manager (RAM) allows resource sharing between accounts (individual accounts or AWS organisation)

- Which AWS resources can I share using RAM (Check AWS for the latest: https://docs.aws.amazon.com/ram/latest/userguide/shareable.html)

1. AWS App Mesh
2. Amazon Aurora
3. AWS Certificate Manager Private Certificate Authority
4. AWS CodeBuild
5. Amazon EC2
6. EC2 Image Builder
7. AWS Glue
8. AWS License Manager
9. AWS Network Firewall
10. AWS Outposts
11. AWS Resource Groups
12. Amazon Route 53
13. Amazon VPC

Example: Create a Resource using account 1 and the account 2 need to accept the resource sharing before they can start using the resource.

AWS Single Sign-On

- Single Sign-On (SSO) service helps centrally manage access to AWS accounts and business applications.

* AD and SAML Integration:

- On-premises AD -> AWS SSO (Security Assertion Markup Language - SAML) - lets you access the apps or OU or SAML 2.0 enabled applications.

All activity is logged in CloudTrail

Exam Tip: If you see SAML in the question, look for SSO in the answers

Advanced IAM Summary(Exam Tips)

- AD
- Connect AWS resources with on-premises AD
- SSO to any domain-joined EC2 instance
- AWS Managed Microsoft AD
- AD Trust - Extend existing AD to on-premises AD using AD Trust
- AWS vs customer responsibility
- Simple AD (trusts are not supported)
- You can use AD Connector
- Cloud Directory (has nothing to do with AD)
- Cognito user pools (also has nothing to do with AD)
- AD vs Non-AD Compatible services
- IAM Policies
    - ARN
    - IAM Policy Structure
    - Effect/Action/Resource
    - Identity vs. resource policies

- Policy evaluation logic (when we have multiple policies assigned to a role, DENY always overrides the ALLOW)

- AWS Managed (can't be edited by users) vs. Customer managed policies

- Permission boundaries (do not deny or allow permission on their own, they defined max permission an identity can have)

- Resource Access Manager (RAM)

    - Resource Sharing between accounts

    - Individual accounts and AWS Organisations

    - Types of resources you can share (remember not all AWS services are available in resource access manager)

- SSO

    - Centrally manage access

    Exam Scenarios for SSO:

        - G Suite, Office 365, Salesforce

        - Use existing identities

        - Account-level permissions

        - SAML, If you see SAML in the question, look for SSO in the answers.

DNS 101

Route53 Exam Tips

- ELB's do not have a pre-defined IPv4 addresses, you resolve them using a DNS name

- Understand the difference between an Alias Record and a CNAME.

- Given the choice, always choose an Alias Record over a CNAME.

- Common DNS Types
    - SOA Records
    - NS Records
    - A Records
    - CNAMES
    - MX Records
    - PTR Records (Reverse of A Record)
    - You can buy domain names directly with AWS.
    - It can take up to 3 days to register depending on the circumstances.

Route 53 Routing Policies

The following Routing Policies are available with Route 53

- Simple Routing
- Weighted Routing
- Latency-based Routing
- Failover Routing
- Geolocation Routing
- Geoproximity Routing (Traffic Flow Only)
- Multivalue Answer Routing

Route 53 Simple Routing Policy - LAB

- You can only have one record with multiple IP addresses. If you specify multiple values in a record, Route 53 returns all values to the user in a random order.

You can't have any health checks.

Route 53 Weighted Routing Policy - LAB

- Allows you to split your traffic based on different weights assigned.

For Example: you can set 10% of your traffic to go to US-EAST-1 and 90% to EU-WEST-1.

Health Checks

- You can set health checks on individual record sets.

- If a recordset fails a health check it will be removed from Route53 until it passes the health check.

- You can also set SNS notifications to alert you if a health check is failed.

Route 53 Latency-Based Routing Policy - LAB

- Allows you to route your traffic based on the lowest network latency for your end user (ie which region will give them the fastest response time)

- To use latency-based routing, you create a latency resource record set for the Amazon EC2 (or ELB) resource in each region that hosts your website.

When Amazon Route 53 receives a query for your site, it selects the latency resource record set for the region that gives the user the lowest latency.

Route 53 then responds with the value associated with that resource record set.

Route 53 Failover Routing Policy - LAB

- Failover Routing policies are used when you want to create an active/passive set-up. For Example: you may want your primary site to be in EU-WEST-2 and your secondary DR site in AP-SOUTHEAST-2

- Route 53 will monitor the health of your primary site using a health check.

- A health check monitors the health of your end points.

Route 53 Geo-Location Routing Policy - LAB

- Geo-Location Routing Policy lets you choose where your traffic will be sent based on the geographic location of your users (ie the location from which DNS queries originate).

For Example: You might want all queries from Europe to be routed to a fleet of EC2 instances that are specifically configured for European customers. These servers may have the local language of European customers and all prices are displayed in Euros.

Route 53 Geoproximity Routing Policy - LAB (beyond the scope of SAA exam and SAP exam)

- Geoproximity Routing lets Amazon Route 53 route traffic to your resources based on the geographic location of your users and your resources. You can also optionally choose to route more traffic or less to a given resource by specifying a value, known as a bias. A bias expands or shrinks the size of the geographic region from which traffic is routed to a resource.

- To use Geoproximity Routing, you must use Route 53 traffic flow.

Route 53 Multivalue Answer Routing Policy - LAB

- Multivalue Answer Routing lets you configure Amazon Route 53 to return multiple values, such as IP addresses for your web servers, in response to DNS queries.

You can specify multiple value for almost any record, but multivalue answer routing also lets you check the health of each resource, so Route 53 returns only values from healthy resources.

- This is similar to simple routing however it allows you to put health checks on each record set.

DNS Summary/Route 53 Exam Tips:

- ELBs do not have pre-defined IPv4 addresses, you resolve them using a DNS name.

- Understand the difference between an Alias Record a CNAME.

- Given the choice, always choose an Alias Record over a CNAME.

Exam Question: You need to map your naked domain name or your zone apex record to an S3 bucket, what should you be using a CNAME or an Alias record?

Answer: Alias record.

- Common DNS Types

    - SOA Records
    - NS Records
    - A Records
    - CNAMES
    - MX Records
    - PTR Records (Reverse of A Record)

- Routing Policies

  - Simple Routing (No health check)

  - Weighted Routing

  - Latency-based Routing

  - Failover Routing

  - Geolocation Routing

  - Geoproximity Routing (Traffic Flow Only)

  - Multivalue Answer Routing (Similar to Simple Routing with health check)

- You can set health checks on individual record sets.

- If a recordset fails a health check it will be removed from Route53 until it passes the health check.

- You can set SNS notification to alert you if a health check is failed.

- Alias Records have special functions that are not present in other DNS servers. Their main function is to provide special functionality and integration into AWS services. Unlike CNAME records, they can also be used at the Zone Apex, where CNAME records cannot. Alias Records can also point to AWS Resources that are hosted in other accounts by manually entering the ARN

- Important

The DNS protocol does not allow you to create a CNAME record for the top node of a DNS namespace, also known as the zone apex. For example, if you register the DNS name example.com, the zone apex is example.com. You cannot create a CNAME record for example.com, but you can create CNAME records for www.example.com, newproduct.example.com, and so on.

In addition, if you create a CNAME record for a subdomain, you cannot create any other records for that subdomain. For example, if you create a CNAME for www.example.com, you cannot create any other records for which the value of the Name field is www.example.com.

- You have an enterprise solution that operates Active-Active with facilities in Regions US-West and India. Due to growth in the Asian market you have been directed by the CTO to ensure that only traffic in Asia (between Turkey and Japan) is directed to the India Region. Which of these will deliver that result?

Answers:

Route 53 - Geolocation routing policy

Route 53 - Geoproximity routing policy

The instruction from the CTO is clear that that the division is based on geography. Latency based routing will approximate geographic balance only when all routes and traffic evenly supported which is rarely the case due to infrastructure and day night variations. You cannot combine blacklisting and whitelisting in CloudFront. Weighted routing is randomized and will not respect Geo boundaries. Geolocation is based on national boundaries and will meet the needs well. Geoproximity is based on Latitude & Longitude and will also provide a good approximation with potentially less configuration.

- You have created a new subdomain for your popular website, and you need this subdomain to point to an Elastic Load Balancer using Route53. Which DNS record set type (or DNS extension type) could you create? (Choose 2).

Answers:

CNAME

Alias

Incorrect ones

MX, AAAA, A

CNAME maps to the host name

An alias could be created for the ELB. Alias records provide a Route 53–specific extension to DNS functionality

- Your company hosts 10 web servers all serving the same web content in AWS. They want Route 53 to serve traffic to random web servers. Which routing policy will meet this requirement, and provide the best resiliency?

Answer: Multivalue Routing

The R53 Simple policy will provide a list of multiple IP addresses in random order. However it is less feature rich that the Multivalue Routing policy which is why AWS released the newer solution.

https://docs.aws.amazon.com/Route53/latest/DeveloperGuide/routing-policy.html

Round Robin would be the traditional solution to this problem. However, a) it is not offered and as answer, and b) R53 Multivalue offers a smarter service. R53 Multivalue lets you responds to DNS queries with up to eight IP address of 'healthy' targets. Plus it will give a different set of 8 to different DNS resolvers. The choice of which to use is left to the requesting service effectively creating a form or randomisation. The R53 Simple policy will provide a list of multiple instance, but Multivalue is the AWS preferred option for this type of service. https://docs.aws.amazon.com/Route53/latest/DeveloperGuide/routing-policy.html https://docs.aws.amazon.com/Route53/latest/DeveloperGuide/dns-failover-simple-configs.html

- Which of the following Route 53 policies allow you to route data to a second resource if the first is unhealthy, and route data to resources that have better performance?

Answers: Failover Routing and Latency-based Routing

Incorrect: Geolocation Routing and Latency-based Routing, Geoproximity Routing and Geolocation Routing, Multivalue Routing and Simple Routing

Failover Routing and Latency-based Routing are the only two correct options, as they consider routing data based on whether the resource is healthy or whether one set of resources is more performant than another. Any answer containing location based routing (Geoproximity and Geolocation) cannot be correct in this case, as these types only consider where the client or resources are located before routing the data. They do not take into account whether a resource is online or slow. Simple Routing can also be discounted as it does not take into account the state of the resources.

- True or False: There is a limit to the number of domain names that you can manage using Route 53.

Answer: True and False. With Route 53, there is a default limit of 50 domain names. However, this limit can be increased by contacting AWS support.

VPC

VPC Overview

- Amazon VPC lets you provision a logically isolated section of the AWS Clous where you can launch AWS resources in a virtual network that you define.

You have complete control over your virtual networking environment, including selection of your own IP address range, creation of subnets, and configuration of route tables and network gateways

- Additionally, you can create a Hardware Virtual Private Network (VPN) connection between your corporate datacenter and your VPC and leverage the AWS cloud as extension of your corporate datacenter.

- Network ACL's are Stateless, they facilitate allow and deny. When you open up a port inbound, it doesn't automatically open up a port on outbound.

- Security Group are Stateful, they only facilitate allow (no deny)

- Internet Assigned Number Authority actually has 3 different sets of IP Addresses that are reserver for private IP address range

    - 10.0.0.0 - 10.255.255.255 (10/8 prefix)

    - 172.16.0.0 - 172.31.255.255 (172.16/12 prefix)

    - 192.168.0.0 - 192.168.255.255 (192.168/16 prefix)

- What can we do with a VPC?

    - Launch instances into a subnet of your choosing

    - Assign custom IP address range in each subnet

    - Configure route tables between subnets

    - Create internet gateway and attach it to our VPC

    - Much better security control over your AWS resources

    - Instance security groups

    - Subnet network access control lists (ACL's)

-Default VPC vs Custom VPC

    - Default VPC is user friendly, allowing you to immediately deploy instances.

    - All subnets in a default VPC have a route out to the internet.

    - Each EC2 instance has both a public and private IP address.

    - If you delete a default VPC, you can recover it now. Try not to delete it.

- VPC Peering

    - Allows you to connect one VPC with another via a direct network route using private IP addresses.

    - Instances behave as if they were on the same private network.

    - You can peer VPC's with other AWS accounts as well as with other VPCs in the same account

    - Peering is in a star configuration: ie 1 central VPC peers with 4 others. NO TRANSITIVE PEERING (Meaning: VPCA peered to VPCB and VPCC, so VPCB can't talk to VPCC through VPCA, instead a VPC peering need to be established between VPCB and VPCC for the communication)

- You can peer between Regions now.

Exam Tips

- Think of a VPC as a logical datacenter in AWS
- Consists of IGW's (or Virtual Private Gateways), Route Tables, Network ACL's, Subnets and Security Groups
- You cannot have a Subnet stretched over multiple AZ's, However an AZ can have multiple subnets.
- Security Groups are Stateful, Network ACL's are Stateless.
- NO TRANSITIVE PEERING (Meaning: VPCA peered to VPCB and VPCC, so VPCB can't talk to VPCC through VPCA, instead a VPC peering need to be established between VPCB and VPCC for the communication)

Remember the following

- When you create a VPC, it will create a default Route Table, Network Access Control List(NACL) and a default Security Group.
- It won't create any subnets, nor will it create a default internet gateway.
- US-East-1A in your AWS account can be completely different availability zone to US-East-1A in another AWS account. The AZ's are randomised.
- Amazon always reserve 5 IP addresses within your subnets (First 4 IPs and the last IP).
- You can only have 1 internet gateway attached one VPC.
- Security Groups can't span VPC's.

NAT Instances and NAT Gateways - Demo

NAT Instance Exam Tips

- When creating a NAT instance, Disable Source/Destination Check on the instance.
- NAT instances must be in a public subnet

- There must be a route out of the private subnet to the NAT instance in order for this to work.

- The amount of traffic that NAT instances can support depends on the instance size. If you are bottlenecking, increase the instance size.

- You can create high availability using Autoscaling Groups, multiple subnets in different AZ's and a script to automate a failover but it is tedious

- NAT instances are behind a security group.

NAT Gateway Exam Tips

- Redundant inside the AZ

- Preferred by the enterprise

- Starts at 5Gbps and scales currently to 45Gbps

- No need to patch

- Not associated with Security Groups

- Automatically assigned a public ip address

- Remember to update your route tables.

- No need to disable Source/Destination Checks

- If you have resources in multiple AZ's and they share one NAT gateway, in the event that the NAT gateway's AZ is down, resources in the other AZ lose internet access, create a NAT gateway in each AZ and configure your routing to ensure that resources use the NAT gateway in the same AZ.

Network Access Control Lists vs. Security Groups - Demo

- Your VPC automatically comes with a default ACL, and by default it allows all inbound and outbound traffic.

- You can create custom network ACL's. By default, each custom network ACL denies all inbound and outbound traffic.

- Each subnet in your VPC must be associate with a network ACL, if you don't explicitly associate a subnet with a network ACL, the subnet is automatically associated with the default network ACL.

- Block IP addresses using network ACL's and not security groups.

- You can associate a network ACL with multiple subnets; however a subnet can be associated with only one network ACL at a time.

When you associate a network ACL with a subnet, the previous association is removed.

- Network ACL's contain a numbered list of rules that is evaluated in order starting with the lowest numbered rule.

- Network ACL's have a seperate inbound and outbound rules, and each rule can either allow or deny traffic.

- Network ACL's are stateless; responses to allowed inbound traffic are subject to the rules for outbound traffic (and vice versa.)

VPC Flow Logs - Demo

- VPC Flow Logs is a feature that enables you to capture information about the IP traffic going to and from network interfaces in your VPC.

Flow log data is stored using Amazon CloudWatch Logs/S3. After you've created a flow log, you can view and retrieve its data in Amazon CloudWatch Logs/S3.

- Flow Logs can be created at 3 levels

1. VPC
2. Subnet Level
3. Network Interface Level

Exam Tips:

1. You cannot enable flow logs for VPC's that are peered with your VPC unless the peer VPC is in your account.

2. You can tag flow logs.

3. After you have created a flow log, you cannot change its configuration; for example: you can't associate a different IAM role with the flow log.

4. Not all IP Traffic is monitored

- Traffic generated by instances when they contact the Amazon DNS server. If you use your own DNS server, then all traffic to that DNS server is logged.

- Traffic generated by a windows instance for Amazon Windows license activation.

- Traffic to and fro from 169.254.169.254 for instance metadata.

- DHCP traffic.

- Traffic to the reserved IP address for the default VPC router.

Bastion Hosts

- A bastion host is a special purpose computer on a network specifically designed and configured to withstand attacks. The computer generally hosts a single application, for example a proxy server, and all other services are removed or limited to reduce the threat to the computer.

It is hardened in this manner primarily due to its location and purpose, which is either on the outside of a firewall or in a demilitarised zone (DMZ) and usually involves access from untrusted networks or computers.

Exam Tips:

- A NAT Gateway or a NAT instance is used to provide internet traffic to EC2 instances in a private subnets.

- A Bastion host is used to securely administer EC2 instances (using SSH or RDP). Bastions are called Jump Boxes in Australia.

- You cannot use a NAT Gateway as a Bastion host.

Direct Connect

- AWS Direct connect is a cloud service solution that makes it easy to establish a dedicated network connection from your on-premises to AWS.

Using AWS Direct connect, you can establish a private connectivity between AWS and your datacenter, office or colocation environment, which in many cases can reduce your network costs, increase bandwidth, and provide a more consistent network experience than internet-based connections.

- Direct Connect directly connects your data center to AWS

- Useful for high throughput workloads (ie lots of network traffic)

- or If you need a stable and reliable secure connection.

Exam question: A VPN connection keeps dropping out because the amount of throughput, and what kinds of things could you do to solve that?

Answer: Direct Connect

Setting Up Direct Connect

- Steps to setting up Direct Connect
  - Create a virtual interface in the Direct Connect Console. This is a public Virtual Interface.
  - Go to the VPC console and then to the VPN connections. Create a Customer Gateway
  - Create a Virtual Private Gateway
  - Attach the Virtual Private Gateway to the desired VPC.
  - Select VPN Connections and create a new VPN Connection.
  - Select the Virtual Private Gateway and the Customer Gateway
  - Once the VPN is available, setup the VPN on the customer gateway and firewall.

Global Accelerator

- AWS Global Accelerator is a service which you create accelerators to improve availability and performance of your applications for local and global users.
- Global Accelerator directs traffic to optimal endpoints over the AWS Global network. This improves the availability and performance of your internet applications that are used by a global audience.
- By default Global Accelerator gives you two static IP addresses that you associate with your accelerator.

  Alternatively, you can bring your own.
- AWS Global Accelerator includes the following components or bring your own

  1. Static IP addresses - Provides two static addresses

  2. Accelerator - An accelerator directs traffic to optimal endpoints over the AWS global network to improve the availability and performance of your internet applications.

3. DNS Name - Global Accelerator assigns each accelerator a default DNS name - similar to somename.domain.com - that points to the static IP addresses that Global Accelerator assigns to you.

Depending on the use case, you can use your accelerator's static IP address or DNS name to route traffic to your accelerator, or set up DNS records to route traffic using your own custom domain name.

4. Network Zone - A network zone services the static IP addresses for your accelerator from a unique IP subnet. Similar to an AWS AZ, a network zone is an isolated unit with its own set of physical infrastructure.

When you configure an accelerator, by default, Global Accelerator allocates two IPv4 addresses for it. If one IP address from a network zone becomes unavailable due to IP address blocking by certain client networks or network disruptions, client applications can retry on the healthy static IP address from the other isolated network zone.

5. Listener - A listener processes inbound connections from clients to Global Accelerator, based on the port (or port range) and protocol that you configure. Global Accelerator supports both TCP and UDP protocols.

Each Listener has one or more endpoints groups associated with it, and traffic is forwarded to endpoints in one of the groups.

You associate endpoint groups with listeners by specifying the Regions that you want to distribute traffic to. Traffic is distributed to optimal endpoints within the endpoint groups associated with the listener.

6. Endpoint Group

- Each endpoint group is associated with a specific AWS region

- Endpoint groups include one or more endpoints in the Region.

- You can increase or reduce the percentage of traffic that would be otherwise directed to an endpoint group by adjusting a setting called a traffic dial.

- The traffic dial lets you easily do performance testing or blue/green deployment testing for new releases across different AWS regions, for example.

7. Endpoint

- Endpoints can be Network Load Balancers, Application Load Balancers, EC2 instances or Elastic IP addresses.

- An application Load Balancer endpoint can be an internet-facing or internal. Traffic is routed to endpoints based on configuration options that you choose, such as endpoint weights.

- For each endpoint, you can configure weights, which are numbers that you can use to specify the proportion of traffic to route to each one.

This can be useful, for example, to do performance testing within a Region.

Exam Tips:

- AWS Global Accelerator is a service which you create accelerators to improve availability and performance of your applications for local and global users.

- Global Accelerator directs traffic to optimal endpoints over the AWS Global network. This improves the availability and performance of your internet applications that are used by a global audience.

- By default Global Accelerator gives you two static IP addresses that you associate with your accelerator or bring your own.

- You can increase or reduce the percentage of traffic that would be otherwise directed to an endpoint group by adjusting a setting called a traffic dial.

- You can control traffic using traffic dials. This is done within the endpoint group.

- You can control weighting to individual end points using weights.

VPC Endpoints (Exam Tips)

- A VPC endpoint enables you to privately connect your VPC to supported AWS services and VPC endpoint services powered by private link without requiring an internet gateway and Nat device, a VPN connection, or an AWS Direct Connect connection.

Instances in your VPC do not require public IP addresses to communicate with resources in the service. So traffic between your VPC and other services does not leave the

Amazon network.

Endpoints are virtual devices, and they are horizontally scaled, redundant, and highly available VPC components that allow communication between instances in your VPC and services without imposing availability, risk, or bandwidth constraints on your network traffic.

- Two types of VPC endpoints.

  1. Interface Endpoints

  2. Gateway Endpoints

Exam Tips:

- An Interface endpoint is an elastic network interface with a private IP address that serves as an entry point for traffic destined to a supported service.

  Many services such as Amazon API Gateway, AWS Cloud Watch, Amazon CloudFormation..etc are supported and many more will be added.

- Currently Gateway Endpoints Support

  1. Amazon S3

  2. Amazon DynamoDB

VPC Private Link

- Sharing applications across VPC's - To open our applications up to other VPC's, we can either

  1. Open the VPC up to the internet, However the disadvantages are

     a. Security considerations; everything in the public subnet is public

     b. A lot more to manage

- Use VPC Peering:

a. You will have to create and manage many different peering relationships

b. The whole network is accessible. This isn't good if you have multiple applications within your VPC.

So Amazon invented Prive Link to open your services in a VPC to another VPC using a Private Link

- The best way to expose a service VPC to tens, hundreds or thousands of customer VPC's

- Doesn't require VPC peering, no route tables, NAT, IGW's ..etc

- Requires a Network Load Balancer on the service VPC and an ENI on the customer VPC.

Exam Tips:

- If you see a scenario question asking about peering VPC tens, hundreds or thousands of customer VPC's, think of AWS Private Link.

- Doesn't require VPC peering, no route tables, NAT, IGW's ..etc

- Requires a Network Load Balancer on the service VPC and an ENI on the customer VPC.

VPC Transit Gateway

- Allows you to have transitive peering between thousands of VPC's and on-premises data centers.

- Works on hub-and-spoke model.

- Works on regional basis, but you can have it across multiple regions.

- You can use it across multiple AWS accounts using RAM (Resource Access Manager)

- You can use route tables to limit how VPCs talk to one another

- Works with Direct Connect as well as VPN connections.

- Support IP Multicast (not supported by any other AWS service)

Exam Tips:

- If you're given a scenario question where it's talking about how you can simplify n/w topology, may be hundreds of VPN connections coming in, direct connect connections,

a whole bunch of VPC peering going on, and you also need to support Multicast, think about Transit Gateway, which uses Hub and Spoke model.

AWS VPN CloudHub

- If you have multiple sites, each with its own VPN connection, you can use AWS VPN CloudHub to connect those sites together.

- Hub and Spoke model

- Low cost, easy to manage.

- It operates over the public internet, but all traffic between the customer gateway and the AWS VPN CloudHub is encrypted.

Exam Tips:

- If you're given a scenario question what's a good way to manage your multiple sites with VPN definitely consider AWS VPN CloudHub.

AWS Network Costs

- Use private IP addresses over public IP addresses to save on costs. This then utilizes the AWS backbone network.

- If you want to cut all network costs, group your EC2 instances in the same AZ and use private IP addresses. This will be cost-free, but make sure to keep in mind single point of failures.

VPC Summary

- Think of a VPC as a logical datacenter in AWS

- Consists of IGW's (or Virtual Private Gateways), Route Tables, Network ACL's, Subnets and Security Groups

- You cannot have a Subnet stretched over multiple AZ's, However an AZ can have multiple subnets.

- Security Groups are Stateful, Network ACL's are Stateless.

- NO TRANSITIVE PEERING (Meaning: VPCA peered to VPCB and VPCC, so VPCB can't talk to VPCC through VPCA, instead a VPC peering need to be established between VPCB and VPCC for the communication)

- When you create a VPC, it will create a default Route Table, Network Access Control List(NACL) and a default Security Group.

- It won't create any subnets, nor will it create a default internet gateway.

- US-East-1A in your AWS account can be completely different availability zone to US-East-1A in another AWS account. The AZ's are randomised.

- Amazon always reserve 5 IP addresses within your subnets (First 4 IP's and the last IP).

- You can only have 1 internet gateway attached one VPC.

- Security Groups can't span VPC's.

NAT Instance Exam Tips

- When creating a NAT instance, Disable Source/Destination Check on the instance.

- NAT instances must be in a public subnet

- There must be a route out of the private subnet to the NAT instance in order for this to work.

- The amount of traffic that NAT instances can support depends on the instance size. If you are bottlenecking, increase the instance size.

- You can create high availability using Autoscaling Groups, multiple subnets in different AZ's and a script to automate a failover but it is tedious

- NAT instances are behind a security group.

NAT Gateway Exam Tips

- Redundant inside the AZ

- Preferred by the enterprise

- Starts at 5Gbps and scales currently to 45Gbps

- No need to patch

- Not associated with Security Groups

- Automatically assigned a public ip address

- Remember to update your route tables.

- No need to disable Source/Destination Checks

- If you have resources in multiple AZ's and they share one NAT gateway, in the event that the NAT gateway's AZ is down, resources in the other AZ lose internet access, create a NAT gateway in each AZ and configure your routing to ensure that resources use the NAT gateway in the same AZ.

Network Access Control Lists vs. Security Groups - Demo

- Your VPC automatically comes with a default ACL, and by default it allows all inbound and outbound traffic.

- You can create custom network ACL's. By default, each custom network ACL denies all inbound and outbound traffic.

- Each subnet in your VPC must be associate with a network ACL, if you don't explicitly associate a subnet with a network ACL, the subnet is automatically associated with the default network ACL.

- Block IP addresses using network ACL's and not security groups.

- You can associate a network ACL with multiple subnets; however a subnet can be associated with only one network ACL at a time.

When you associate a network ACL with a subnet, the previous association is removed.

- Network ACL's contain a numbered list of rules that is evaluated in order starting with the lowest numbered rule.

- Network ACL's have a seperate inbound and outbound rules, and each rule can either allow or deny traffic.

- Network ACL's are stateless; responses to allowed inbound traffic are subject to the rules for outbound traffic (and vice versa.)

VPC's and ELB's

- Important Exam Tip: You need at least two public subnets to deploy an internet facing load balancer.

VPC Flow Logs:

1. You cannot enable flow logs for VPC's that are peered with your VPC unless the peer VPC is in your account.

2. You can tag flow logs.

3. After you have create a flow log, you cannot change its configuration; for example: you can't associate a different IAM role with the flow log.

4. Not all IP Traffic is monitored

- Traffic generated by instances when they contact the Amazon DNS server. If you use your own DNS server, than all traffic to that DNS server is logged.

- Traffic generated by a windows instance for Amazon Windows license activation.

- Traffic to and fro from 169.254.169.254 for instance metadata.

- DHCP traffic.

- Traffic to the reserved IP address for the default VPC router.

Bastion Host

- A NAT Gateway or a NAT instance is used to provide internet traffic to EC2 instances in a private subnets.

- A Bastion host is used to securely administer EC2 instances (using SSH or RDP). Bastions are called Jump Boxes in Australia.

- You cannot use a NAT Gateway as a Bastion host.

Direct Connect

- AWS Direct connect is a cloud service solution that makes it easy to establish a dedicated network connection from your on-premises to AWS.

Using AWS Direct connect, you can establish a private connectivity between AWS and your datacenter, office or colocation environment, which in many cases can reduce your network costs, increase bandwidth, and provide a more consistent network experience than internet-based connections.

- Direct Connect directly connects your data center to AWS

- Useful for high throughput workloads (ir lots of network traffic)

- or If you need a stable and reliable secure connection.

Exam question: A VPN connection keeps dropping out because the amount of throughput, and what kinds of things could you do to solve that?

Answer: Direct Connect

- Steps to setting up Direct Connect

  - Create a virtual interface in the Direct Connect Console. This is a public Virtual Interface.

  - Go to the VPC console and then to the VPN connections. Create a Customer Gateway

  - Create a Virtual Private Gateway

  - Attach the Virtual Private Gateway to the desired VPC.

  - Select VPN Connections and create a new VPN Connection.

  - Select the Virtual Private Gateway and the Customer Gateway

  - Once the VPN is available, setup the VPN on the customer gateway and firewall.

Global Accelerator

- AWS Global Accelerator is a service which you create accelerators to improve availability and performance of your applications for local and global users.

- Global Accelerator directs traffic to optimal endpoints over the AWS Global network. This improves the availability and performance of your internet applications that are used by a global audience.

- By default Global Accelerator gives you two static IP addresses that you associate with your accelerator or bring your own.

- You can increase or reduce the percentage of traffic that would be otherwise directed to an endpoint group by adjusting a setting called a traffic dial.

- You can control traffic using traffic dials. This is done within the endpoint group.

- You can control weighting to individual end points using weights.

VPC Endpoints

- A VPC endpoint enables you to privately connect your VPC to supported AWS services and VPC endpoint services powered by private link without requiring an internet gateway and Nat device, a VPN connection, or an AWS Direct Connect connection.

- Instances in your VPC do not require public IP addresses to communicate with resources in the service. So traffic between your VPC and other services does not leave the Amazon network.

- Endpoints are virtual devices, and they are horizontally scaled, redundant, and highly available VPC components that allow communication between instances in your VPC and services without imposing availability, risk, or bandwidth constraints on your network traffic.

- Two types of VPC endpoints.

1. Interface Endpoints
2. Gateway Endpoints

Exam Tips:

- An Interface endpoint is an elastic network interface with a private IP address that serves as an entry point for traffic destined to a supported service.

Many services such as Amazon API Gateway, AWS Cloud Watch, Amazon CloudFormation..etc are supported and many more will be added.

- Currently Gateway Endpoints Support

1. Amazon S3
2. Amazon DynamoDB

VPC Private Link

- If you see a scenario question asking about peering VPC's tens, hundreds or thousands of customer VPC's, think of AWS Private Link.
- Doesn't require VPC peering, no route tables, NAT, IGW's ..etc
- Requires a Network Load Balancer on the service VPC and an ENI on the customer VPC.

VPC Transit Gateway

- Allows you to have transitive peering between thousands of VPC's and on-premises data centers.
- Works on hub-and-spoke model.
- Works on regional basis, but you can have it across multiple regions.
- You can use it across multiple AWS accounts using RAM (Resource Access Manager)
- You can use route tables to limit how VPCs talk to one another
- Works with Direct Connect as well as VPN connections.
- Support IP Multicast (not supported by any other AWS service)

Exam Tips:

- If you're given a scenario question where it's talking about how you can simplify n/w topology, may be hundreds of VPN connections coming in, direct connect connections, a whole bunch of VPC peering going on, and you also need to support Multicast, think about Transit Gateway, which uses Hub and Spoke model.

AWS VPN CloudHub

- If you have multiple sites, each with its own VPN connection, you can use AWS VPN CloudHub to connect those sites together.

- Hub and Spoke model

- Low cost, easy to manage.

- It operates over the public internet, but all traffic between the customer gateway and the AWS VPN CloudHub is encrypted.

Exam Tips:

- If you're given a scenario question what's a good way to manage your multiple sites with VPN definitely consider AWS VPN CloudHub.

AWS Network Costs

- Use private IP addresses over public IP addresses to save on costs. This then utilizes the AWS backbone network.

- If you want to cut all network costs, group your EC2 instances in the same AZ and use private IP addresses. This will be cost-free, but make sure to keep in mind single point of failures.

HA Architecture (10 questions on Load balancers)

Elastic Load Balancers - Exam Tips

- Application Load Balancer (HTTP and HTTPS) - Layer 7 aware (make intelligent decisions/routing).

- Network Load Balancer (TCP, TLS and UDP) - Extreme Performance/Static IP Addresses.

- Classic Load Balancer (PREVIOUS GENERATION for HTTP, HTTPS, and TCP) - Test & Dev to keep costs low.

- 504 Error means that the gateway has timed out. This means that the application not responding within the idle timeout period.

- Troubleshoot the application to figure out App/DB layer to scale it out or up.

- If you need the IPv4 address of your end user, look for the X-Forwarded-For header.

Load Balancers and Health Checks

- Instances monitored by ELB are reported as; InService, or OutofService

- Health Checks check the instance health by talking to it.

- Load Balancers have their own DNS name. You are never given an IP address.

- Read the ELB FAQ for all Load Balancers.

- Want to deep dive on application load balancers? Check out our deep dive course!

Advanced Load Balancers Theory

1. Sticky Sessions:

Classic Load Balancer routes each request independently to the registered EC2 instance with the smallest load.

Sticky sessions allow you to bind a user's session to a specific EC2 instance. This ensures that all requests from the user during the session are sent to the same instance.

You can enable Sticky Sessions for Application Load Balancers as well, but the traffic will be sent to the Target Group level.

Exam Scenario 1: A user trying to visit a website behind a classic load balancer and essentially what's happening is it's just sending all the traffic to one EC2 instance.

Answer: Disable Sticky session.

Exam Scenario 2: If you have got an EC2 instance or an application, where you're writing to an EC2 instance like local disk, then of course you would want to enable Sticky session.

2. Cross Zone Load Balancing:

Enables you to load balance across multiple AZ's

Exam Scenario 1:

With No Cross Zone Load Balancing, we got a user and we are using Route 53 for our DNS, which is splitting of our traffic 50/50 and sending the requests to EC2's in two diff AZ's

Each AC has a Load Balancer, The first AZ has 4 EC2 instances and the second has only one EC2 instance.

Because we don't have Cross Zone Load Balancing enabled - First AZ will split 50% to 4 instances and the second AZ receives 50% on 1 instance.

When we enable Cross Zone Load Balancing: The Load balancer will distribute the load evenly among instances on both AZ's.

Exam Scenario 2:

We got a user and we are using Route 53 for our DNS, which is sending all the requests (100%) to a Load Balancer in AZ1, The first AZ1 has 4 EC2 instances and the second has only one EC2 instance.

Route 53's 100% traffic is sent to the only load balancer in US-EAST-1A and no traffic is being sent to US-EAST-1B.

In this scenario, we enable Cross Zone Load Balancing to distribute the traffic evenly between US-EAST-1A and US-EAST-1B

3. Path Patterns:

You can create a listener with rules to forward requests based on the URL path. This is known as path-based routing.

If you are running microservices, you can route traffic to multiple back-end services using path-based routing.

For Example: you can route general requests to one target group and requests to render images to another target group

Exam Scenario 1:

We got a user and we are using Route 53 for our DNS, which is sending all the requests (100%) to a Load Balancer in AZ1, The first AZ1 has 4 EC2 instances and the second has only one EC2 instance.

www.myurl.com should go to AZ1 and www.myurl.com/images should go to the media instances in AZ2. In this instance, we enable Path Patterns.

Auto Scaling (3 to 4 questions)

Auto Scaling group: Min size, Desired Capacity, Max Size and Scale out as needed

Auto Scaling has 3 components

1. Groups: Logical component, Webserver group or Application group or DB group ... etc.

2. Configuration Templates: Group uses a launch template or a launch configuration as a configuration template for its EC2 instances.

You can specify information such as the AMI ID, instance type, key pair, security groups and block device mapping for your instances.

3. Scaling Options: Scaling Options provides several ways for you to scale your Auto Scaling groups.

For Example: you can configure a group to scale based on the occurence of specified conditions (dynamic scaling) or on a schedule.

What are my scaling options: (Exam Tips)

1. Maintain current instance levels at all times - Performs periodic health checks and launches a new one when an instance is unhealthy.

2. Scale Manually - the most basic way to scale your resources where you specify Min size, Desired Capacity, Max Size of your Auto Scaling group.

Amazon EC2 Auto scaling manages the process of creating or terminating instances to maintain the updated capacity.

3. Scale based on a schedule: based on date and time and predictable schedule.

4. Scaled based on a demand: A more advanced way to scale your resources using scaling policies, which let you define parameters that control the scaling process.

5. Use predictive scaling: You can use Amazon EC2 Auto Scaling with Amazon Auto Scaling to scale resources across multiple services.

AWS Auto Scaling can help you maintain optimal availability and performance by combing predictive scaling and dynamic scaling (proactive and reactive, respectively)

to scale your Amazon EC2 capacity faster.

Launch Configurations and Auto Scaling Groups - Demo

- Create a Launch Configuration and then use Launch Configuration during Auto Scaling Group creation
- During Auto Scaling - there is warmup of 300 secs for new instances

HA Architecture

- Always Design for failure
- Use Multiple AZ's and Multiple Regions where ever you can.
- Know the difference between Multi-AZ and Read Replicas for RDS.
- Know the difference between scaling out and scaling up
- Read the question carefully and always consider the cost element.

- Know the different S3 storage classes.

High Availability with Bastion Hosts

- Scenario 1: Two hosts in two seperate AZ's. Uses a Network Load Balancer with static IP addresses and health checks to fail over from one host to another.

- Can't use Application Load Balancer, at it's a layer 7 and you need to use layer 4

- Scenario 2: One host in one AZ behind an Auto Scaling Group with health checks and a fixed Elastic IP.

If the host fails, the health check will fail and the Auto Scaling group will provision a new EC2 instance in a seperate AZ.

You can use a user data script to provision the same EIP to the new host. This is the cheapeast option but it is not 100% fault tolerant.

On-Premises Strategies with AWS

- You need to be aware of what high-level AWS services you can use on-premises for the exam:

  - Database Migration Service (DMS)
  - Server Migration Service (SMS)
  - AWS Application Discovery Service
  - VM Import/Export
  - Download Amazon Linux 2 and an ISO

Applications

SQS

- Amazon SQS is a web service that gives you access to a message queue that can be used to store messages while waiting for a computer to process them

- it's a distributed queue system that enables web service applications to quickly and reliably queue messages that one component in the application generates to be consumed by another component

- A queue is a temporary repository for messages that are awaiting processing.

- Using Amazon SQS, you can decouple the components of an application so they run independently, easing message management between components

- Any component of a distributed application can store messages in a fail-safe queue.

- Messages can contain up to 256KB of text in any format. Any component can later retrieve the messages programmatically using the Amazon SQS API.

- Think decoupling infrastructure/micro-services means SQS.

- 256KB is not a hard limit and it can go up tp 2GB, however it will be stored in S3

- The queue acts as a buffer between the component producing and saving data, and the component receiving the data for processing.

- This means the queue resolves issues that arise if the producer is producing work faster than the consumer can process it, or if the producer or consumer are only intermittently connected to the network.

- There are 2 types of queues

1. Standard Queues (default): A standard queue lets you have a nearly-unlimited number of transactions per second. Standard queues guarantee that a message is delivered at least once.

Occasionally (because of the highly-distributed architecture that allows high throughput), more than one copy of a message might be delivered out of order.

However, standard queues provide best-effort ordering which ensures that messages are generally delivered in the same order as they are sent

2. FIFO: The FIFO complement the standard queue.

The most important feature of this queue type are FIFO delivery and exactly-once processing: the order in which messages are sent and received is strictly preserved and a message is delivered once and remains available until a consumer processes and deleted it; duplicates are not introduced into the queue.

FIFO queues also support message groups that allow multiple ordered message groups within a single queue.

FIFO queues are limited to 300 transactions per second (TPS), but have all the capabilities of standard queues.

Exam Tips:

- SQS is pull based, not push based.

- Messages are 256KB in size

- Messages can be kept in the queue from 1 min to 14 days; the default retention period is 14 days.

- Visibility timeout is the amount of time that the message is invisible in the SQS queue after a reader picks up that message.

Provided that the job processed before the visibility timeout expires, the message will then be deleted from the queue.

If the job has not processed within that time, the message will become visible again and another reader will process it.

This could result in the same message being delivered twice.

- Visibility timeout maximum is 12 hours.

Exam Scenario: A message is being delivered twice, what could be the reason?

Answer: Visibility timeout is not long enough.

- SQS guarantees that your messages will be processed at least once.

- Amazon SQS long polling is a way to retrieve messages from your Amazon SQS queues. While the regular short polling returns immediately (even if the message queue being polled is empty) , long polling doesn't return a response until a message arrives in the message queue, or the long poll times out.

Exam Scenario: Your EC2 instances are constantly polling while the queue is empty, so this could increase your costs. How can you reduce the costs?

Answer: Use long polling

- Any time you see a scenario based question about "decoupling your infrastructure" - think SQS.

Simple Workflow Service

- Amazon Workflow Service (Amazon SWF) is a web service that makes it easy to coordinate work across distributed application components.

SWF enables applications for a range of use cases, including media processing, web application back-ends, business process workflows and analytics pipelines, to be designed as a coordination of tasks.

- Tasks represent invocations of various processing steps in an application which can be performed by executable code, web service calls, human actions and scripts.

Exam Scenario: Any human interaction required in the service, think of SWF

SWF vs SQS:

- SQS has a retention period of 14 days; with SWF, workflow executions can last up to 1 year.

- Amazon SWF is a task-oriented API. whereas Amazon SQS offers a message-oriented API.

- Amazon SWF ensures that a task is assigned only once and is never duplicated. With Amazon SQS, you need to handle duplicated messages and may also need to ensure that a message is processed only once.

- Amazon SWF keeps track of all the tasks and events in an application, With Amazon SQS, you need to implement your own application-level tracking, especially if your application uses multiple queues

SWF Actors:

1. Workflow Starters: An application that can initiate (start) a workflow. Could be your e-commerce website following the placement of order or a mobile app searching for business.

2. Deciders: Control the flow of activity tasks in a workflow execution. If something has finished (or failed) in a workflow, a Decider decides what to do next.

3. Activity Workers: Carry out the activity tasks.

Simple Notification Service

- Amazon Simple Notification Service (Amazon SNS) is a web service that makes it easy to set up, operate and send notifications from the cloud.

- It provides developers with highly scalable, flexible and cost-effective capability to publish messages from an application and immediately deliver them to subscribers or other applications.

- Push notifications to Apple, Google, Fire OS and Windows devices as well as Android devices in China with Baidu Cloud Push.

- Besides pushing cloud notifications directly to mobile devices, Amazon SNS can also deliver notification by SMS text message or email to Amazon SQS or to any HTTP endpoint.

- SNS allows you to group multiple recipients using topics. A topic is an "access point for allowing recipients to dynamically subscribe to identical copies of the same

- One topic can support deliveries to multiple endpoint types - for example, you can group together iOS, Android and SMS recipients,

When you publish once to a topic, SNS delivers appropriately formatted copies of your message to each subscriber.

Exam Tips:

- Instantaneous, push-based delivery (no pull)
- Simple APIs and easy integration with applications
- Flexible message delivery over multiple transport protocols
- Inexpensive, pay-as-you-go model with no up-front costs.
- Web-based AWS Management Console offers the simplicity of a point-and-click interface.

SNS vs SQS

- Both Messaging Services with AWS
- SNS - Push
- SQS - Poll (Pulls)

Elastic Transcoder - Exam Tips

- Media Transcoder in the cloud.

- Convert Media files from their original source format in different formats that will play on smartphones, tables, PCS etc.

- Provides transcoding presets for popular output formats, which means that you don't need to guess about which settings work best on particular devices.

- Pay based on the minutes that you transcode and the resolution at which you transcode.

API Gateway - Exam Tips (5 to 10 questions)

- API Gateway is a fully managed service that makes it easy for developers to publish, maintain, monitor and secure APIs at any scale.

- API Gateway has caching capabilities to increase performance

- API Gateway is low cost and scales automatically

- You can throttle API Gateway to prevent attacks

- You can log results to CloudWatch

- If you are using Javascript/AJAX that uses multiple domains with API Gateway, ensure that you have enable CORS on API Gateway.

- CORS is enforced by the client (Browser)

Kinesis 101 (Streaming Data)

- Amazon Kinesis is a platform on AWS to send your streaming data to. Kinesis makes it easy to load and analyse streaming data, and also providing the ability to build your own custom applications for your business needs.

- 3 different types of Kinesis (Exam Tips)

1. Kinesis Streams: Stores data in Shards. (Retention from 24 hours to 7 days). EC2 instances read this data and store it in S3, DynamoDB, RedShift and EMR.

Kinesis Stream consists of Shards (not imp for the exam)

- 5 transactions per sec for reads, up to a max total data read rate of 2 MB per sec abd up to 1000 records per sec for writes, up to a max total data write of 1 MB per sec (including partition keys)

- The data capacity of your stream is a function of the number of shards that you specify for the stream. The total capacity of the stream is the sum of the capacities of its shards.

2. Kinesis Firehose: Kinesis Firehose has no data persistence, however it processes the data using Lambda and stores it in S3, DynamoDB, RedShift, EMR, Amazon Elasticsearch Service, and Splunk, enabling near real-time analytics with existing business intelligence tools and dashboards you're already using today.

3. Kinesis Analytics: Works with Kinesis Streams and Kinesis Analytics to analyse the data on the fly.

- Exam Scenario:

You must choose the most relevant service.

Web Identity Federation & Cognito (Exam Tips)

- Web Identity Federation lets you give your users access to AWS resources after they have successfully authenticated with a web-based identity provider such as Apple, Google and Facebook. Following successful authentication, the user receives an authentication code from the Web ID provider, which they can trade for temporary AWS security credentials.

- Amazon Cognito provides Web Identity Federation with the following features

- Sign-up and sign-in to your apps

- Access for guest users

- Acts as an Identity Broker between your application and Web ID providers, so you don't need to write any additional code.

- Synchronises user data for multiple devices

- Recommended for all mobile applications AWS services.

- The user authenticates first with the Web ID Provider and receives an authentication token, which is exchanged for temporary AWS credentials allowing them to assume an IAM role.

- Cognito is an Identity Broker which handles interaction between your applications and the Web ID provide (You don't need to write your own code to do this)

- User pool is user based, It handles things like user registration, authentication and account recovery

- Identity pools authorise access to your AWS resources.

- Exam question about difference between User pool and Identity pools

Event Processing Patterns - Exam Tips

- Understand the pub/sub pattern - facilitated by SNS
- DLQ - SNS, SQS, Lambda
- Fanout pattern - SNS
- S3 event notifications - Which events trigger; which services consume

Applications Summary

- SQS is a way to de-couple your infrastructure

- SQS is pull based not push based
- Messages are 256 KB in size
- Messages can be kept in the queue from 1 min to 14 days; the default retention period is 4 days.
- Standard SQS and FIFO SQS
- Standard order is not guaranteed and messages can be delivered more than once.

- FIFO order is strictly maintained and messages are delivered only once.

- Visibility timeout is the amount of time that the message is invisible in the SQS queue after a reader picks up that message.

Provided that the job processed before the visibility timeout expires, the message will then be deleted from the queue.

If the job has not processed within that time, the message will become visible again and another reader will process it.

This could result in the same message being delivered twice.

- Visibility timeout maximum is 12 hours.

- SQS guarantees that your messages will be processed at least once.

- Amazon SQS long polling is a way to retrieve messages from your Amazon SQS queues. While the regular short polling returns immediately (even if the message queue being polled is empty) , long polling doesn't return a response until a message arrives in the message queue, or the long poll times out.

SWF vs SQS:

- SQS has a retention period of 14 days; with SWF, workflow executions can last up to 1 year.

- Amazon SWF is a task-oriented API. whereas Amazon SQS offers a message-oriented API.

- Amazon SWF ensures that a task is assigned only once and is never duplicated. With Amazon SQS, you need to handle duplicated messages and may also need to ensure that a message is processed only once.

- Amazon SWF keeps track of all the tasks and events in an application, With Amazon SQS, you need to implement your own application-level tracking, especially if your application uses multiple queues

SWF Actors:

1. Workflow Starters: An application that can initiate (start) a workflow. Could be your e-commerce website following the placement of order or a mobile app searching for business.

2. Deciders: Control the flow of activity tasks in a workflow execution. If something has finished (or failed) in a workflow, a Decider decides what to do next.

3. Activity Workers: Carry out the activity tasks.

SNS Benefits:

- Instantaneous, push-based delivery (no pull)
- Simple APIs and easy integration with applications
- Flexible message delivery over multiple transport protocols
- Inexpensive, pay-as-you-go model with no up-front costs.
- Web-based AWS Management Console offers the simplicity of a point-and-click interface.

SNS vs SQS

- Both Messaging Services with AWS
- SNS - Push
- SQS - Poll (Pulls)

Elastic Transcoder - Exam Tips

- Media Transcoder in the cloud.
- Convert Media files from their original source format in different formats that will play on smartphones, tables, PCS etc.
- Provides transcoding presets for popular output formats, which means that you don't need to guess about which settings work best on particular devices.
- Pay based on the minutes that you transcode and the resolution at which you transcode.

API Gateway - Exam Tips (5 to 10 questions)

- API Gateway is at a fully managed service that makes it easy for developers to publish, maintain, monitor and secure APIs at any scale.
- API Gateway has caching capabilities to increase performance
- API Gateway is low cost and scales automatically
- You can throttle API Gateway to prevent attacks
- You can log results to CloudWatch
- If you are using Javascript/AJAX that uses multiple domains with API Gateway, ensure that you have enable CORS on API Gateway.

- CORS is enforced by the client (Browser)

Kinesis 101 (Streaming Data): 3 different types of Kinesis (Exam Tips)

1. Kinesis Streams: Stores data in Shards. (Retention from 24 hours to 7 days). EC2 instances read this data and store it in S3, DynamoDB, RedShift and EMR.

Kinesis Stream consists of Shards (not imp for the exam)

- 5 transactions per sec for reads, up to a mac total data read rate of 2 MB per sec abd up to 1000 records per sec for writes, up to a mac total data write of 1 MB per sec (including partition keys)

- The data capacity of your stream is a function of the number of shards that you specify for the stream. The total capacity of the stream is the sum of the capacities of its shards.

2. Kinesis Firehose: Kinesis Firehose has no data persistence, however it processes the data using Lambda and stores it in S3, DynamoDB, RedShift, EMR, Amazon Elasticsearch Service, and Splunk, enabling near real-time analytics with existing business intelligence tools and dashboards you’re already using today.

3. Kinesis Analytics: Works with Kinesis Streams and Kinesis Analytics to analyse the data on the fly.

Refer to Web Identity Federation & Cognito (Exam Tips)

Event Processing Patterns - Exam Tips

- Understand the pub/sub pattern - facilitated by SNS
- DLQ - SNS, SQS, Lambda
- Fanout pattern - SNS
- S3 event notifications - Which events trigger; which services consume

Security

Reducing Security Threats

- NACL - IP or range of IP's, hard to maintain. so use WAF
- Security Groups at the instance level
- AWS WAF - SQL Injection or cross site scripting
- You can attach WAF to CloudFront or NACL

Key Management Service (KMS)

- Regional secure key management and encryption and decryption
  - Manages customer master keys (CMKs)
  - Ideal for S3 objects, database passwords and API keys stored in.
  - Encrypt and decrypt data up to 4KB in size
  - Integrated with most AWS services.
  - Pay per API call
  - Audit capability using CloudTrail - logs delivered to S3.
  - FIPS 140-2 Level 2
  - Level 3 is CloudHSM
  - Types of CMK's
    - AWS Managed CMK: Free; used by default if you pick encryption in most AWS services. Only that service can use them directly. Dedicated to my account.
    - AWS Owned CMK: Used by AWS on a shared basis across many accounts; you typically won't see these. Not dedicated to my account.
    - Customer Managed CMK: Allows key rotation; controlled via key policies and can be enabled/disabled. Dedicated to my account.

Symmetric vs Asymmetric CMKs

Symmetric:

- Default, same key used for encryption and decryption
- AES-256
- Never leave AWS unencrypted
- Must call the KMS API to use

- AWS services integrated with KMS use symmetric CMKs
- Encrypt, decrypt and re-encrypt data
- Generate data keys, data key pairs and random byte strings
- Import your own key material.

Asymmetric:

- Mathematically related public/private key pair, used by SSL's
- RSA and elliptic-curve cryptography (ECC)
- Private key never leaves AWS encrypted
- Must call the KMIs API to use private key
- Download the public key and use outside AWS
- Used outside AWS by users who can't call KMS APIs
- AWS services integrated with KMS do not support asymmetric CMKs
- Sign messages and verify signatures

CloudHSM

- Dedicated hardware security module (HSM)
  - FIPS 140-2 level 3 (Exam Scenario: if you see FIPS 140-2 level 3 - answer is CloudHSM)
  - Level 2 is KMS (need to show evidence of tampering)
  - Difference between KMS and CloudHSM - We manage our keys
  - Difference between KMS (Multi Tenant) and CloudHSM (Single Tenant, dedicate h/w, Mutli-AZ cluster)
  - Runs within a VPC in your account
  - Industry-standard APIs - no AWS APIs
  - PKCS#11
  - Java Cryptography Extensions (JCE)
  - Microsoft CryptoNG (CNG)

- Keep your keys safe - irretrievable if lost!

- CloudHSM will operate inside its own VPC dedicated to CloudHSM

- CloudHSM will project to ENI of customer VPC

- CloudHSM is not highly available by default, so we need to provision one HSM per subnet

Exam Tips:

- Regulatory compliance requirements

  - FIPS 140-2 level 3 (Exam Scenario: if you see FIPS 140-2 level 3 - answer is CloudHSM)

System Manager Parameters Store

- Component of AWS Systems Manager (SSM)
- Secure serverless storage of configuration and secrets
  - Passwords
  - Database Connection Strings
  - License Codes
  - API Keys
- Values can be encrypted (KMS) or plain text
- Separate data from source control
- Store parameters in hierarchies (Up to 15 levels deep)
- Track versions
- Set TTL to expire values such as passwords

Secrets Manager (Exam Tips: System Manager Parameters Store vs Secrets Manager)

- Similar to System Manager Parameters Store

- Charge per secret and per 10,000 API calls, so this might add up for large organisations, where as System Manager Parameters Store is free

- Automatically rotate secrets

- Apply the new key/passwords in RDS for you
- Generate random secrets

AWS Shield

- Protects against distributed denial-of-service (DDoS) attacks

AWS Shield Standard vs AWS Shield Advanced

AWS Shield Standard:

- Automatically enabled for all customers at no cost
- Protects against common layer 3 and 4 attacks
    - SYN/UDP floods, SYN is TCP and UDP is UDP
    - Reflection attacks
- Stopped a 2.3 Tbps DDOS attack for three days in Feb 2020

AWS Shield Advanced

- $3K per month per org
- Enhanced protection for EC2, ELB, CloudFront, Global Accelerator, Route 53
- Business and Enterprise support customers get 24 X 7 access to the DDoS Response Team (DRT)
- DDoS cost protection (similar to insurance)

Web Application Firewall

- WAF that lets you monitor HTTP(S) requests to CloudFront, ALB or API Gateway
- Control access to content
- Configure filtering rules to allow/deny traffic:
    - IP addresses
    - Query string parameters
- SQL query injection
- Blocked traffic returns HTTP 403 Forbidden code.

- WAF allows three different behaviours;
    - Allow all requests, except the ones you specify
    - Block all requests, except the ones you specify
    - Count the requests that match the properties you specify

- Request Properties
    - Originating IP address
    - Originating country
    - Request Size
    - Values in request headers
    - Strings in request matching regex patterns.
    - SQL Code injection
    - Cross-site scripting (XSS)

- AWS Firewall Manager
    - Centrally configure and manage firewall rules across an AWS Organisation
        - WAF Rules
            - ALB
            - API Gateway
            - CloudFront distributions

        - AWS Shield Advanced protections:
            - ALB
            - ELB Classic
            - EIP
            - CloudFront distributions

        - Enable security groups for EC2 and ENIs

AWS WAF & AWS Shield (Exam Tips)

- What us AWS WAF? To prevent Cross scripting or SQL injection

1. AWS WAF is a web application firewall that helps protect your web application from common web exploits that could affect application availability, compromise security, or consume excessive resources

- What is AWS Shield - to prevent DDOS

1. AWS Shield is a managed DDOS protection service that safeguards web applications running on AWS. AWS Shield provided always-on detection and automatic inline mitigations that minimise application downtime and latency, so there is no need to engage AWS support to benefit from DDOS protection.

- There are two tiers of AWS shield - Standard and Advanced.
- Advanced costs $3K per month per org.
- Only Advanced offers automated application layer monitoring.
- What determines price for Lambda?

1. Request Pricing
    - Free Tier: 1 million requests per month
    - $0.20 per 1 million requests thereafter
2. Duration Pricing
    - 400,000 GB-seconds per month free, up to 3.2 million seconds of compute time*
    - $0.00001667 for every GB-second used thereafter
3. Additional Charges
    - You may incur additional charges if your lambda functions uses other AWS services or transfers data. For example, If your lambda function reads and writes data to or from Amazon S3, you will be billed for the read/write requests and the data stored in Amazon S3

Serverless

Lambda - Exam Tips

- Lambda scales out (not up) automatically

- Lambda is serverless

- Lambda functions are independent, 1 event = 1 functions

- Know what services are serverless (Aurora Serverless (Only RDS), DynamoDB, S3, Lambda, API Gateway are serverless. EC2 and RDS are not serverless).

- Lambda functions can trigger other lambda functions, 1 event can = x functions if functions can trigger other functions.

- Architectures can get extremely complicated, AWS X-ray allows you to debug what is happening.

- Lambda can do things globally, you can use it to back up S3 buckets to other S3 buckets etc

- Know Lambda triggers, RDS can't trigger Lambda (check AWS to confirm) - Important for the Exam

Serverless Application Model (SAM)

- CloudFormation extension optimised for serverless applications

- New Types: functions, APIs, tables

- Supports anything that CloudFormation supports

- Run serverless application locally using docker

- Package and deploy using CodeDeploy

Elastic Container Service (ECS)

What are Containers and Docker?

- A container is a package that contains an application, libraries, runtime and tools required to run it

- Run on a container engine like Docker

- Provides the isolation benefits of virtualisation with less overhead and faster starts than VMs

- Containerised applications are portable and offer a consistent environment. Which allocate memory and CPU per container.

What is ECS?

- Managed container orchestration service.
- Create clusters to manage fleets of container deployments
- EC2 manages EC2 or Fargate instances
- Schedules containers for optimal placement
- Defines rules for CPU and memory requirements
- Monitors resource utilisation
- Deploy, update and roll back
- Free... for real!, However we pay for the resources that are provisioned.
- Integrates with VPC, security groups, EBS volumes and ELB
- CloudTrail and CloudWatch

ECS Components:

- Cluster: Logical collection of ECS resources - either ECS EC2 instances or Fargate instances

- Task Defintion: Defines your application. Similar to a Dockerfile but for running containers in ECS. Can contain multiple containers.

- Container Definition: Inside a task defintion, it defines the individual container a task uses. Controls CPU and memory allocation and port mappings.

- Task: Single running copy of any containers defined by a task definition. One working copy of an application (e.g., DB and web containers)

- Service: Allows tasks definitions to be scaled by adding tasks. Defines minimum and maximum values.

- Registry: Storage for container images (e.g, Elastic Container Registry (ECR) or Docket Hub). Used to download images to create containers.

Fargate

- Serverless container engine
- Eliminates need to provision and manage servers
- Works with both ECS and EKS
- Each workload runs in its own kernel
- Isolation and security
- Choose EC2 instead if
    - Compliance requirements
    - Require broader customisation
    - Applications require GPU

EKS

- Elastic Kubernetes Service
- K8s is open-source software that lets you deploy and manage containerised applications at scale
- Same toolset on-premises and in cloud
- Containers are grouped in pods.
- Like ECS, supports both EC2 and Fargate
- Why use EKS?
    - Already using K8s
    - Want to migrate to AWS

ECR

- Managed Docker container registry
- Store, manage and deploy images
- Integrated with ECS and EKS

- Works with on-premises deployments
- Highly available
- Integrated with IAM
- Pay for storage and data transfer

ECS + ELB

- Distribute traffic evenly across tasks in your service
- Supports ALB, NLB, CLB
- Use ALB to route HTTP/HTTPS (layer 7) traffic
- Use NLB or CLB to route TCP (layer 4) traffic
- Supported by both EC2 and Fargate launch types.
- ALB allows:
    - Dynamic host port mapping
    - Path-based routing
    - Priority rules
- ALB is recommended over NLB or CLB

ECS Security (Imp for the Exam)

- Instance Roles vs Task Roles
    - Instance Roles - applies policy to all tasks running on EC2 instance (for ex: S3 access), which is not ideal for security.
    - Task Roles - Applies policies per task (for ex: Role A: S3 only, Role B: S3 + DynamoDB). which is ideal for security.

Serverless Summary

- Traditional vs Serverless Architecture
- Go for Serveless Model with API Gateway, Lambda and DynamoDB
- Lambda Exam Tips:
    - Lambda scales out (not up) dynamically
    - Lambda functions are independent, 1 event = 1 functions

- Lambda is serverless

- Know what services are serverless! RDS is not serverless with the exception of Aurora Serverless

- Lambda functions can trigger other lambda functions, 1 event can = x functions if functions trigger other functions.

- Architectures can get extremely complicated, AWS X-ray allows you to debug what is happening.

- Lambda can do things globally, you can use it to back up S3 buckets to other S3 buckets etc.

- Know your triggers and know what can't be triggered from Lambda.

- Which of the following services can invoke a Lambda function synchronously?

Answer: ALB, Cognito, Lex, Alexa, API Gateway, CloudFront, and Kinesis Data Firehose are all valid direct (synchronous) triggers for Lambda functions. S3 is one of the valid asynchronous triggers.

- Source: https://docs.aws.amazon.com/lambda/latest/dg/lambda-services.html

Services That Lambda Reads Events From

Amazon Kinesis

Amazon DynamoDB

Amazon Simple Queue Service

Services That Invoke Lambda Functions Synchronously

Elastic Load Balancing (Application Load Balancer)

Amazon Cognito

Amazon Lex

Amazon Alexa

Amazon API Gateway

Amazon CloudFront (Lambda@Edge)

Amazon Kinesis Data Firehose

AWS Step Functions

Services That Invoke Lambda Functions Asynchronously

Amazon Simple Storage Service

Amazon Simple Notification Service

Amazon Simple Email Service

AWS CloudFormation

Amazon CloudWatch Logs

Amazon CloudWatch Events

AWS CodeCommit

AWS Config

AWS IoT Events