Cloud Guru – AWS Certified Solutions Architect Associate (SAA-C02) Notes

Chapter 1 – Introduction

Exam Guide

- Exam Questions
    - 1 correct response and 3 incorrect responses (distractors)
    - 2 correct responses out of 5 response options
- Curriculums
    - Design Resilient Architectures
        - Multi-tier architecture
        - High available or fault-tolerance architectures
        - Decoupling mechanisms using AWS services
    - Design High-Performing Architectures
        - Identify elastic and scalable compute solutions
        - Select high-performing and scalable solutions for a workload
        - Select high-performing networking solutions
        - Choose high-performing database solutions
    - Secure Applications and Architectures
        - Design secure access to AWS resources
        - Design secure application tiers
        - Select appropriate data security options
    - Design Cost-Optimized Architectures
        - Identify cost-effective storage solutions
        - Identify cost-effective compute databases
        - Design cost-optimized network architectures

Chapter 4 – AWS Fundamentals

AWS Global Infrastructure

- Regions
    - A geographical area which consists of 2 or more availability zones
    - 24 Regions
- Availability Zones
    - Are data centers each with redundant power, networking, and connectivity
        - Building filled with servers
    - May be several data centers, but because they are close together, they are counted as one
    - 77 AZ
- Edge Locations
    - Are endpoints for AWS that are used for caching content.

- Typically consist of CloudFront, Amazon's CDN

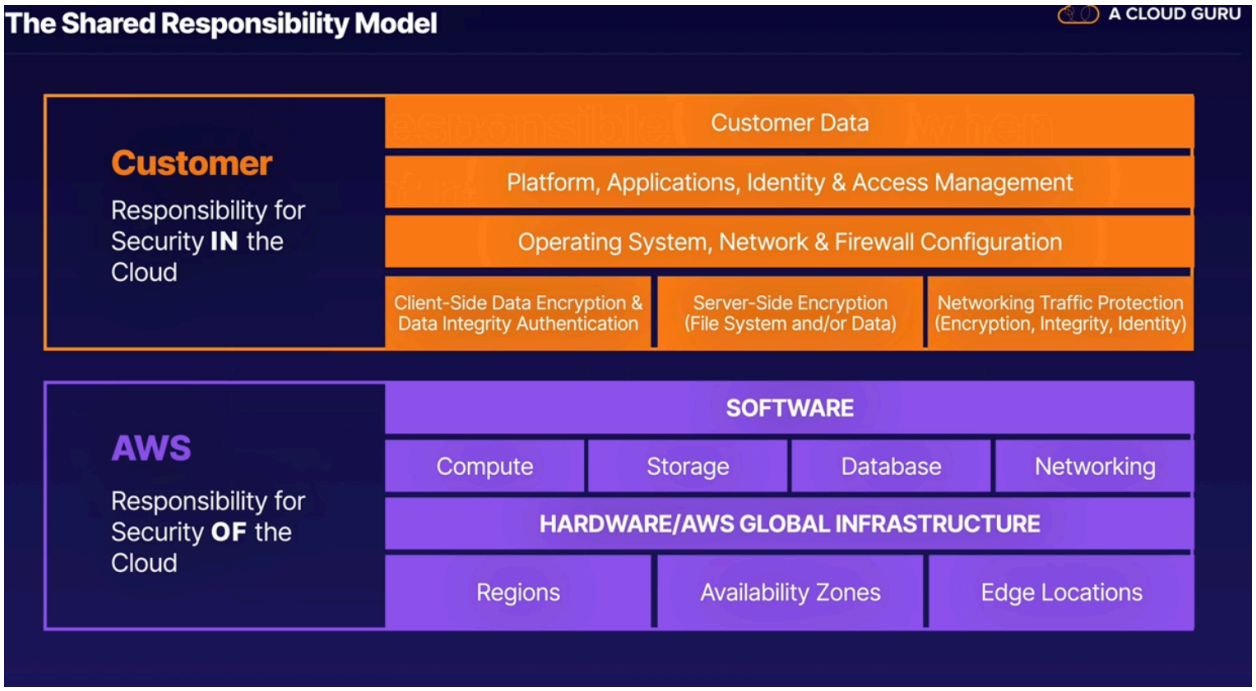
AWS Global Infrastructure Services to know for the Exam:

- Analytics
- Management and Governance
- Migration and Transfer
- Compute
- Storage
- Network and Content Delivery
- Application Integration
- Security, Identity ,and Compliance
- Containers
- AWS Cost Management
- Databases

Who owns What in the Cloud?

You vs Amazon:

- The Shared Responsibility Model



Read the shared Responsibility Model before the exam.

Exam Tip: Ask yourself, can you do this yourself in the AWS Management Console?

- If yes, you are likely responsible
    - Examples: Security groups, IAM users, patching EC2 operating systems, patching databases running on EC2, etc
- If not, AWS is likely responsible
    - Examples: Management of data centers, security cameras, cabling, patching RDS, operating Systems, etc

Compute, Storage, Databases, and Networking

- Compute: Services to crunch the data and powerful power to calculate
    - EC2
    - Lambda
    - Elastic Beanstalk
- Storage: Safe place to save your information
    - S3
    - EBS
    - EFS
- Databases: Reliable way to store and retrieve your information
    - RDS
    - DynamoDB
    - Redshift
- Networking: A way for compute, storage, and databases to communicate
    - VPCs
    - Direct Connect
    - Route 53
    - API Gateway
    - AWS Global Accelerator

Know the above services for the exam.

The Well-Architected Framework

Read the Well-architected framework.

The five pillars of a well-architected framework.

- Operational Excellence
    - Focuses on running and monitoring systems to deliver business value and continually improve processes and procedures
- Security
    - Focuses on protecting information and systems
- Performance Efficiency
    - Focuses on using IT and computing resources efficiently
- Reliability

    - Focuses on ensuring a workload performs its intended function correctly and consistently as its expected to
- Cost Optimization
    - Focuses on avoiding unnecessary costs

Securing the Root Account

What is IAM?

- IAM allows you to manage users and their level of access to the AWS console.
    - Create users and grant permissions to those users
    - Create groups and roles
    - Control access to AWS resources

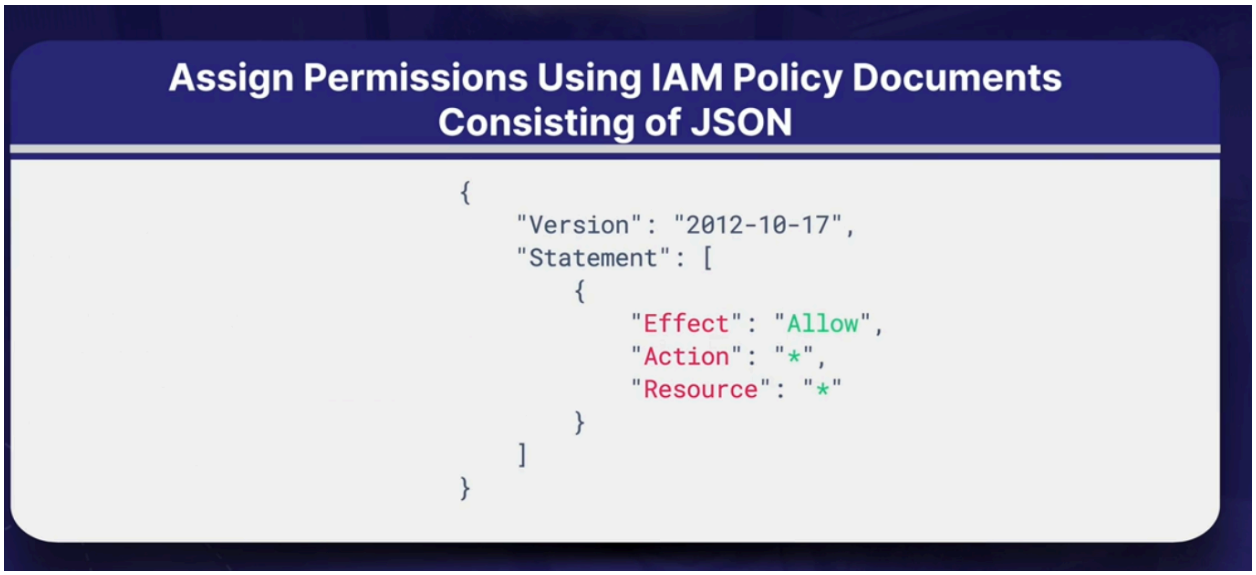
The root account is the email address you used to sign up for AWS. The root account has full admin access.

4 Steps to Secure Root:

- Enable MFA
- Create an admin group for your admins, assign appropriate permissions to this group
- Create user accounts for your admins
- Add your users to the admin group

Controlling User's Actions with IAM Policy Documents

- We control permissions using the policy documents, which are made up of JSON documents.



Permanent IAM Credentials

The Building Blocks of IAM

- Users
    - A physical person
- Groups
    - Functions such as admin, developer, etc which contain users
- Roles
    - Internal usage within AWS

Best practices:

- Have users inherit permissions from groups.
- The principle of least privilege for users and groups.

One person equals one user.

Key Points for Exam:

- The default user has no default permissions.
- IAM Universal, it does not apply to regions at the time
- The root account is the first account used when created, has complete admin access. Do not use this account for day to day.
- Access key ID and secret access keys are not the same as usernames and passwords. You can only view them once.
- Always setup password rotations
- IAM Federation, combine existing user accounts with AWS, Microsoft Active Directory
- Identity Federation, uses SAML standard

Chapter 5 – Simple Storage Service (S3)

S3 Overview

What is S3?

- Object Storage: Provides secure, durable, highly scalable object storage
- Scalable: S3 allows you to store and retrieve any amount of data from anywhere
- Simple: Amazon S3 is easy to use, with a simple web service interface

S3 is Object-Based Storage

- Manages data as objects rather than file systems or blocks
- Upload any file type (photos, videos, code, documents, text files)

S3:

- Unlimited storage
- Up to 5TB per object
- S3 files in buckets (similar to folders)
- Data is spread across multiple devices and facilities to ensure availability and durability (99.95 – 99.99%)
- Designed for durability (the chance of a file to not get corrupted)

Working with S3 Buckets:

- Universal Namespace (has to be globally unique)
- URL https://bucket-name.s3.Region.amazonaws.com/key-name
- When a file upload is complete an HTTP 200 code appears

Storage of S3 – Key-Value Store:

- Key: Name of the object
- Version ID: Important for storing multiple versions in the same object
- Value: The data itself
- Metadata: Data about the data

S3 Versions:

- S3 Standard
    - Data is stored in >= 3 AZs
    - Designed for Frequent Access
    - Suitable for Most Workloads

Characteristics of S3:

- Tiered Storage
- Lifecyle Management: Define rules to automatically transition objects to a cheaper storage option or delete objects after a set period of time
- Versioning: All versions of an object are stored and can be retrieved, including deleted objects.

Securing Your Data:

- Server-side Encryption
    - All objects will be encrypted when stored in bucket
- Access Control Lists (ACLs)
    - Define which AWS accounts or groups are granted access and type of access.
- Bucket Policies
    - S3 bucket policies specify what actions are allowed or denied

Strong Read-After-Write Consistency:

- After overwrite or upload of new object, a read request immediately receives the latest version

Securing Your S3 Bucket

- Buckets are private by default. You have to allow public access to both the bucket and its objects
- HTTP 200 code will appear if upload is successful

Object ACLs vs Bucket Policies

- Object ACLs
    - Work on an individual object level
- Bucket Policy
    - Work on an entire bucket level

Hosting a Static Website in S3

- You can use S3 to host static websites (think HTML pages)
- S3 scales automatically to meet demand

Tips for Exam:

- Bucket Policies: Make entire buckets public using bucket policies
- Static Content: Use S3 to host static content only
- Automatic Scaling: S3 scaling automatically with demand

Versioning Objects in S3

- Versioning: Having multiple versions of an object.
- Advantages:
    - Have all version
    - Great backup tool
    - Cannot be disabled (only suspended)
    - Lifecycle Rules
    - Supports MFA
        - Deletion of objects can only happen with two factors of authentication

Tips for Exam:

- All versions are stored, even if you delete an object
- Great backup tool
- Can’t be disabled
- Can be integrated with lifecycle rules
- Supports MFA

- Once versioning is turned off a deleted object can't be restored

S3 Storage Classes

- S3 Standard
    - Data is stored redundantly across multiple devices in multiple facilities (>=3 AZs)
    - Designed for frequent access
    - Default storage class
    - Uses included big data, websites
    - Always highest cost
- S3 Standard-Infrequent Access (Standard-IA)
    - Data is rapidly accessed
    - Pay access to the data
    - Used for long-term storage, backups, disaster recovery
- S3 One Zone-Infrequent Access
    - Similar to S3-IA, but stored redundantly within a single AZ
    - Costs 20% less than S3 Standard-IA
    - Good for long-lived, infrequently accessed, non-critical data
- Glacier
    - Optimized for data that is accessed infrequently
    - You pay each time you access your data
    - Data archiving
    - Retrieval time of 1 minute to 12 hours
    - Typical retrieval time is 3 – 5 hours
- Glacier Archive
    - Default retrieval time of more than 12 hours
- S3 Intelligent-Tiering
    - Automatically moves your data to the most cost-effective tier
    - Cost-optimized

Tip for Exam: Know storage class and use case only

Storage Class	Availability and Durability	AZ(s)	Use Case
S3 Standard	99.99% Availability 11 9's Durability	>=3	Suitable for most workloads (e.g., websites, content distribution, mobile and gaming applications, and big data analytics)
S3 Standard-Infrequent Access	99.9% Availability 11 9's Durability	>=3	Long-term, infrequently accessed critical data (e.g., backups, data store for disaster recovery files, etc.)
S3 One Zone-Infrequent Access	99.5% Availability 11 9's Durability	1	Long-term, infrequently accessed, non-critical data
S3 Glacier	99.99% Availability 11 9's Durability	>=3	Long-term data archiving that occasionally needs to be accessed within a few hours or minutes
S3 Glacier Deep Archive	99.99% Availability 11 9's Durability	>=3	Rarely accessed data archiving with a default retrieval time of 12 hours (e.g., financial records for regulatory purposes)
S3 Intelligent-Tiering	99.9% Availability 11 9's Durability	>=3	Unknown or unpredictable access patterns

Lifecycle Management with S3

- Automates moving your objects between the different storage tiers, maximizing cost effectiveness
- You can use lifecycle management to move different *current or previous* versions

S3 Object Lock and Glacier Vault

S3 Object Lock

- Used to store objects using a write once, read many (WORM) model. Help prevent objects from being deleted or modified for a fixed amount of time
- Used to meet regulatory requirements that require WORM storage and add an extra layer of protection
- Can be applied across entire buck or individual object
- Two Modes:
    - Governance mode: Users can't overwrite or delete an object version or alter its lock settings unless they have special permissions
    - Compliance mode: A protected object version can't be overwritten or deleted by any user, including the AWS root account
- Retention Periods:
    - Protects an object version for a fixed period of time.
    - After retention period expires, the object version can be overwritten or deleted unless a legal hold is placed
- Legal Hold
    - Prevent an object version from being overwritten or deleted.
    - Remains in affect until explicitly removed from user.

Glacier Vault Lock

- Easily deploy and enforce compliance controls for individual S3 Glacier vaults with a vault lock policy. Can specify controls such as WORM or lock policy from future edits

Encrypting S3 Objects

Encryption in Transit

- SSL/TLS
- HTTPS

Encryption at Rest: Server-Side

- SSE-S3 (AES 256-bit encryption)
- SSE-KMS (Key Management Service)
- SSE-C (Customer-provided keys)

Encryption at Rest: Client-Side

- You encrypt the files yourself before uploading them to S3

Encryption of S3 Objects

- Console
- You can encrypt using a bucket Policy

Optimizing S3 Performance

S3 Prefixes:

- Folders and subfolders inside the bucket
    - The more prefixes you have, the more efficient S3 performance 3500 per second PUT/COPY/POST/DELETE
    - 5,500 per second GET/HEAD
- SSE-KMS enabled will count toward KMS quota
    - Region specific (5,500 10,00 and 30,000)

Multipart Uploads:

- Required for files over 5 GB
- Over 100 MB, it's recommended
- Parallelize uploads (increases efficiency)
    - Break the file into segments of bytes

S3 Downloads

- Parallelize downloads. You can specify byte ranges.
- Can be used to download only portions of the files

Backing up Data with S3 Replication

S3 Replication

- You can replicate objects from one bucket to another
    - Versioning must be enabled on both source and destination buckets
- Objects already existing in bucket will not be replicated automatically
- Delete markers are not replicated by default

Chapter 6 – Elastic Compute Cloud (EC2)

EC2 Overview

- A secure, resizable compute capacity in the cloud
- A VM, hosted in AWS instead of a personal data center
- Complete control of your own instances
- Pay for what you use. Scale and shrink when needed

EC2 Pricing Options

- On-Demand: Pay by hour or second
- Reserved: Reserved capacity 1 or 3-year contract 72% discount on hourly charge
- Spot: Purchase unused capacity at a discount of up to 90%. Prices will fluctuate with supply and demand. Applications with start and end times
- Dedicated: Physical EC2 server dedicated for your use. Most expensive option. Used for compliance and licensing-based requirements

On-Demand Instances

- Flexible: Low cost and flexibility of Amazon Ec2 without upfront payment
- Short-term: Applications with short-term or unpredictable workloads
- Testing: Applications being developed or tested on Amazon EC2 for the first time

Reserved Instances (RI)

- Predictable Usage: Application with a steady state or predictable usage
- Specific Capacity Requirements: Applications that require reserved capacity
- Payment up front: Reduce the total computing costs for the future
- Standard RIs (Reserved Instances): 72% off the on-demand price
- Convertible RIs: Up to 54% off the on-demand price. Has the option to change to a different RI of equal or greater value
- Scheduled RIs: Launch within the time window you define. Match your capacity reservation to a predictable recurring schedule a fraction a day, week, or month
    - Operate on regional level (one region only)

Spot Instances

- Applications that have flexible start and end times
- Applications that are only feasible at very low compute prices
- Users with an urgent need for large amounts of additional computing capacity

Dedicated Hosts

- Compliance: Regulatory requirements that may not support multi-tenant virtualization
- Licensing: Does not support multi-tenancy or cloud deployments
- On-Demand: Can be purchased on-demand (hourly)
- Reserved: Can be purchased as a reservation for up to 70% off the on-demand price

AWS Command Line

- You can interact with AWS services using the command line

Exam Tips:

- Use Groups
    - Create IAM groups and assign your users to groups
- Least Privilege

3 CLI Exam Tips:

- Secret Access Key
    - You will only see this key once. If lost, a new pair must be generated
- Don't Share Key Pairs
    - Each developer should have their own access key ID and secret
- Supports Linux, Windows, or MacOS

Using Roles

IAM Roles:

- Role is an identity you can create in IAM that has specific permissions
- A role is similar to a user, it has specific permission policies which determine what the identity can and cannot do in AWS
- Can be assumable by anyone
- Roles are temporary
    - Provides you with temporary security credentials
- Can be used in user, service, architecture accounts

Exam Tips:

- Preferred Option (security perspective)
- Avoid hard-coded credentials
- Can update and attach policies which will take immediate effect
- You can attach and detach roles without having to stop EC2 instances

Security Groups & Bootstrap Script

Security Groups

- Virtual firewalls for your EC2 instance.
- Defaults: Inbound (everything blocked) Outbound (everything allows)
- To let everything in: 0.0.0.0/0
- In order to use SSH/RDP/HTTP/HTTPS, the correct ports needs to be opened
- Changes to security groups take effect immediately
- Can have multiple security groups attached to EC2 instances
- Can have any number of EC2 instance within a security group

Bootstrap Scripts

- A script that runs when the instance first runs. Runs at root level
- Adds more time to boot instance, allows you to automate the installation of applications

EC2 Metadata

- Data about your EC2 instance (data about data)
    - E.x. private IP address, public IP address, hostname, etc

Retrieve Metadata:

- Curl http://ip-address/latest/meta-data/

Exam

- User data is bootstrap scripts
- You can use bootstrap scripts to access metadata

EC2 Networking

3 Types of EC2 Virtual Cards

- Elastic Network Interface (ENI): Day-to-day networking
- Enhanced Networking (EN): Uses single root I/O for high performance
- Elastic Fabric Adapter (EFA): Accelerates High Performance Computing (HPC) and machine learning applications

ENI Use Cases

- Use for network and security appliances in your VPC
- Create a low-budget, high-availability solution

- Create a management network

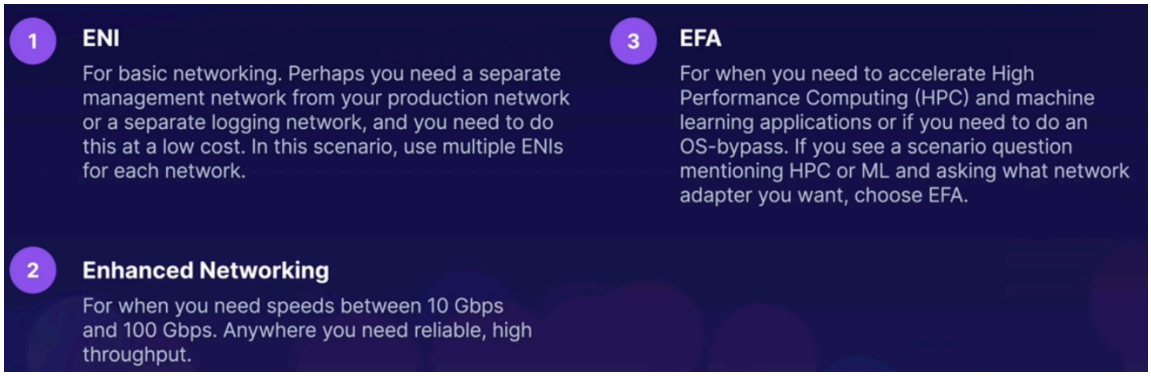
EN

- Adapter Types
    - Elastic Network Adaptor (ENA)
        - Support up to 100 Gbps
    - Intel 82599 Virtual Function (VF) interface
        - Support up to 10 Gpbs
    - Always Choose ENA over Intel

EFA Use Cases

- Best for high performance computing
- OS-bypass enabled

Exam Tip



Optimizing EC2 Placement Types

3 Types

- Cluster Placement
- Spread Placement
- Partition Placement

Cluster Placement

- Grouping of instances within a single Availability Zone
- Best for application which need low network latency, high network throughput

Spread Placement

- Are placed on different underlying hardware
- Used for individual instances

Partition Placement

- Each partition has its own rack.

- Use cases Multiple EC2 Instances, HBase, HDFS, Cassandra

Exam Tip

- AWS recommends the same placement groups
- Can't merge placement groups
- Cluster placement group can't span multiple AZ
- Can move an existing instance with CLI, AWS SDK, must be in stop state

Solving Licensing Issues with Dedicated Hosts

Dedicated Hosts are best for:

- Compliance Demands
- Licensing (does not support multi-tenancy)
- On-Demand
- Reserved

Exam Tip

- Any questions that talk about special licensing or compliance requirements, think Dedicated

Timing Workloads with Spot Instances and Spot Fleets

Spot Instances

- Let you take advantage of unused EC2 capacity in the AWS Cloud
- Can save up to 90% compared to On-demand prices
- Used for stateless or fault-tolerant applications
    - Big data, CI/CD, containerized, **high-performance**, etc
- Not good for
    - Critical jobs, databases

Spot Prices

- Must decide what the maximum Spot price is
- Depends on the capacity and region
- The instance will be provisioned while it's below your maximum
- If price rises above maximum, instances can be stopped or terminated

Spot Block

- Won't stop block spot price even if it rises above maximum price
- Can last 1 – 6 hours
-

Terminating Spot Instances

- How do you terminate spot instances under a persistent spot request?
    - Cancel spot request, then terminate instances

Spot Fleets

- Is a collection of Spot Instances and On-demand instances
- Will try to match the target capacity with your price restraints
- You can setup different launch pools with different strategies

Chapter 7 – Elastic Block Storage (EBS) and Elastic File System (EFS)

EBS Overview

- Storage volumes you can attach to your EC2 instance
    - Can attach database, applications, etc
- Essentially a virtual hard disk in the cloud
- Designed for mission critical data
- Automatically replicated within a single Availability Zone
- Highly scalable, increase capacity and change volume type with no downtime

EBS Volume Types

General Purpose SSD

- General Purpose SSD (gp2 or gp3)
- Gp2
    - Up to 16,000 IOPS
- Gp3
    - Suitable for high performance applications
    - Baseline 300 IOPS
- Good for boot volumes

Provisioned IOPS SSD

- io1 & io2
- Good for I/O-intensive apps or large databases
- More than 16,000 IOPS per volume

Won't be tested on the differences between gp2 vs gp3 / io1 vs io2

Thoughput Optimized Hard Disk Drive (HDD) (st1)

- st1
    - Best for frequently accessed, throughput-intensive workloads. Big data, data warehouses, log processing

Cold HDD (sc1)

- sc1
    - Lowest cost option. Good for fewer scans (File server), performance is not a factor, less access

IOPS

- Measures the number of read and write operations per second
- Important metric for quick transactions, low-latency apps

- Ability to read and write very quickly

Throughput

- Measures number of bits read or written per second
- Important for large datasets, large I/O sizes, complex queries
- Exam: Remember Big Data

Volumes and Snapshots

- Volumes: Exist on EBS, they are virtual hard disks.
- Snapshots: Snapshots exist on S3. They are point-in-time images with an incremental (only the data that has been changed since your last snapshot is moved to S3). The first snapshot may take some time to create

3 Tips for Snapshots

- For a consistent snapshot, it is recommended to stop the instance and take a snapshot.
- Encrypted EBS volumes will have encrypted snapshots
- *To move a snapshot from one region to another, you must copy the snapshot into the region*

EBS Volumes

- EBS volumes will always be in the same AZ as EC2
- You can resize the volume without stopping or restarting the instance
- You can switch volume types without stopping or restarting
- Not encrypted by default

Protecting EBS Volumes with Encryption

- Data key uses the AES-256 algorithm
    - Keys can be managed by your CMK or through AWS KMS

What happens when an EBS volume is encrypted?

- Data at rest and data in transit is encrypted
- Snapshots are encrypted
- All volumes created from the snapshot are encrypted

EBS Encryption

- Little to no latency
- Copying an unencrypted snapshot allows encryption
- Snapshots are encrypted volumes are encrypted

4 Steps to Encrypt an Unencrypted Volume

- Create a snapshot of the unencrypted device
- Create a copy of the snapshot and select encrypt option
- Create AMI from encrypted snapshot
- Use AMI to launch a new encrypted instance

EC2 Hibernation

- EC2 Hibernation: The operating system is put into hibernation (suspend-to-disk). Hibernation saves all RAM contents to the Amazon EBS root volume

When your instances it out of hibernation:

- Amazon EBS volume is restored to previous state
- RAM contents are reloaded
- The processes that were previously running on the instance are resumed
- Previously attached data volumes are reattached

EC2 Hibernation:

- The instance boots much faster
- Must be less than 150 GB
- Instance families include all the C, M, and R
- Available for Windows, Amazon Linux, and Ubuntu
- Can't be hibernated for more than 60 days
- Available for on-demand and reserved instances

Elastic Files System (EFS) Overview

- Managed Network File System (NFS)
    - Way to share files across multiple EC2 instance
- Can work with EC2 instance in multiple AZ
- Highly available and scalable, is expensive

EFS Use Cases

- Content Management, easily share content among instances
- Web Servers

EFS Overview

- Uses NFSv4 protocol
- Only Linux-based AMI
- File system which scales automatically
- Encryption at rest using KMS
- Pay per use
- 1000s of concurrent connections of EC2 instances
- Scales to the petabytes

EFS Performance Options

- General Purpose: Used for web servers
- Max I/O: Big data and media processing

Storage Tiers

- Standard: For frequently accessed files

- Infrequently Accessed: For files not frequently accessed

What to Remember about EFS

- Pay for use
- Can scale up to petabytes
- Can support thousands of NFS connections
- Data is stored in multiple AZ within a region
- Read-after-write consistency

FSx Overview

FSx for Windows

- Fully managed native Microsoft Windows file system to easily move Windows-based applications into AWS
- FSx runs on Windows Server Message Block (SMB)
- Designed for Windows
- Supports AD, Windows services

FSx for Lustre Performance

- Fully managed file system that is optimized for compute-intensive workloads, high performance computing, machine learning, etc
- Can store data directly on S3

EFS vs FSx

- EFS is used for distributed, highly resilient storage for Linux instances and Linux-based applications

Amazon Machine Images: EBS vs Instance Stores

What is AMI?

- Amazon Machine Image (AMI): Provides the information required to launch an instance. Blueprint for EC2 instance
- 5 Things to Base an AMI on:
    - Region
    - OS
    - Architecture
    - Launch permissions
    - Storage for root device

Amazon EBS vs Instance Stores

- Amazon EBS: The root device for an instance is launched from the AMI is an Amazon EBS volume Instance Store: The root device for an instance is launched from the AMI is an instance store volume created from a template stored in S3

Instance Store Volumes

- Will lose data if the host fails or if the instance is deleted

- Cannot be stopped
- Sometimes referred to as ephemeral storage
- Can reboot without losing data

EBS-backed Volume

- Can be stopped
- Can reboot without losing data
- EBS volumes can be kept while deleting the AMI
- By default, root volume will be deleted, but can be kept

AWS Backup

- Backup allows you to consolidate your backups across multiple AWS services, EC2, EBS, EFS, Amazon FSx
- Backup can be used with AWS organizations to back up multiple AWS accounts
- Provides a central management interface
- Provides automation backup schedule and retention policies
- Improved Compliance can be enforced while backups can be encrypted at rest and in transit
    - Makes it easy for auditing since services are consolidated in one interface

Chapter 8 – Databases

Relational Database Services (RDS)

- Relational Database Structure
    - Tables
    - Rows
    - Columns
- Relational Database Engines
    - SQL Server
    - Oracle
    - MySQL
    - PostgreSQL
    - MariaDB
    - Amazon Aurora

Advantages of RDS

- Up and running in minutes
- Failover capacity
- Multi-AZ
- Automated Backups

When do you use RDS databases?

- Generally used for online transactions processing (OLTP)

OLTP vs OLAP (Online Analytical Processing)

- OLTP: Processes data from transactions in real time (customer orders, banking transactions, payments, and more)
- OLAP: Processes complex queries to analyze historical data
    - Redshift AWS Service is optimized for data warehousing

Multi-AZ RDS

- RDS creates exact copy of production database in a different AZ
    - SQL Server, MySQL, MariaDB, Oracle, PostgreSQL, Aurura
- RDS will automatically fail over to the standby so database operations can resume quickly
- Used for Disaster Recovery, can't use in tandem with production system

Increasing Read Performance with Read Replicas

What is a read replica?

- A read-only copy of your primary database. Scales read performance
- Boosts performance
- Can be cross an AZ
- Cross region
- Each read replica has its own DNS endpoint
- Can be promoted to be its own database
- Requires Automatic Backups to be turned on
- Multiple read replicas are supported

What is Amazon Aurora?

- Amazon Aurora: MySQL and PostgreSQL compatible relational database engine. Combines speeds and availability of high-end commercial databases with simplicity and cost-effectiveness
- 5x better performance than MySQL
- 3x better performance than PostgreSQL
- Proprietary Amazon database

Things to know about Aurora:

- Start with 10GB, scales in 10 GB
- 2 copies of data are contained in each AZ, minimum of 3 AZ. 6 copies of data
- Transparently handle the loss of up to 2 copies of data
- Continuously scanned for errors

3 Types of Aurora Replicas:

- Aurora Replicas: have 15 replicas with Aurora. Automatic failover is only available with Aurora replicas.
- MySQL Replicas: have 5 read replicas with Aurora MySQL
- PostgreSQL Replicas: have 5 read replicas with Aurora PostgreSQL

Backups with Aurora

- Automated backups are always enabled
- You can create and share snapshots with Aurora with other AWS accounts

Aurora Serverless

- Serverless DB cluster automatically starts up, shuts down, and scales capacity up or down based on your applications need
- Cost-effective for infrequent access or unpredictable workloads

DynamoDB Overview

- Non-relational (NoSQL) Amazon DB for applications needs consistent, single-digit millisecond latency at any scale

Facts about DynamoDB

- Spread across 3 geographical distinct data centers
- Eventually Consistent reads (default)
- Strongly Consistent reads
- Stored on SSD storage

Eventually consistent reads vs Strongly

- Eventually: Copies of data is usually reached within a second. Best for read performance
- Strongly: Returns a result that reflects all writes prior to the read

DynamoDB Accelerator (DAX)

- Fully managed, highly available, in-memory cache
- 10x performance improvement
- Reduces to microseconds

On-demand capacity

- Pay-per-request pricing

Security

- Encryption at rest using KMS
- Can connect site-to-site VPN
- Direct Connect (DX)
- IAM policies and roles
    - Fine-grained access
- CloudWatch and CloudTrail
- VPC endpoints

When do you use DynamoDB transactions?

- ACID: Atomic, consistent, isolated, durable
    - Methodology in DB
- Atmoic: All changes to the data are performed successfully
- Consistent: Data must be in a consistent state before and after the transaction

- Isolated: No other process can change the data while the transactions is running
- Durable: The changes made by a transactions must persist
- ACID means “All or nothing”

ACID with DynamoDB

- Transactions provide developers atomicity, consistency, isolation, and durability (ACID) across 1 or more tables within a single AWS account and region
- *Must enable DynamoDB transactions with ACID*

Use cases

- Financial transactions
- Fulfilling orders

Things to remember

- If you see any scenario question with ACID, think DynamoDB transactions
- All-or-nothing
- DynamoDB transactions provide across 1 or more tables within a single AWS account and region

Saving Your Data with DynamoDB Backups

- Full backups at any time
- Consistency within seconds and retained until deleted
- No performance impact
- Manual backups are available

Point-in-Time Recovery (PITR)

- Protects against accidental writes or deletes
- Not enabled by default
- Restore to any point in the last 35 days through incremental backups
- Latest restorable time is 5 minutes in the past.

Taking Your Data Global using DynamoDB streams and Global Tables

DynamoDB Streams

- Time-ordered sequences of item-level changes in a table
    - First in first out change
- Stored for 24 hours
- Inserts, updates, and deletes
- Combine with Lambda functions

Global Tables

- Managed multi-master, multi-region replication
- Globally distributed applications
- *Based on DynamoDB streams – Enable DynamoDB streams*
- Multi-region redundancy for disaster recovery or high availability

- Replication latency is under 1 second

Chapter 9 – Virtual Private Cloud (VPC)

Build your own VPC the day before the exam

VPC Introduction

- A virtual data center in the cloud
- Are logically isolated part of the AWS cloud, where you can define your own network
- Complete control over your virtual network
- Consists of route tables, VPG, network access control lists, subnets, and security groups
- 1 subnet is always in 1 AZ
- Fully customizable network
    - Leverage multiple layers of security, including security groups, subnets, etc

VPNs

- You can create a hardware Virtual Private Network connection between a corporate data center and the VPC to leverage AWS cloud

What can you do with VPCs?

- Launch instances into a subnet of your choosing
- Customize IP addresses
- Customize route tables
- Create internet gateway
- More control over your AWS resources
- Access control lists (use network access control lists to block specific IP addresses)

Default VPC vs Custom

Default

- Default VPC is user friendly
- All subnets in default VPC route to internet
- Each EC2 instance has both a public and private IP address

Custom

- Fully customizable
- Takes time to setup

Using a NAT Gateway for the Internet

- Network address translation (NAT) gateway to enable instances in a private subnet to connect to the internet, while preventing instances from being publicly accessible

Facts to Remember about NAT

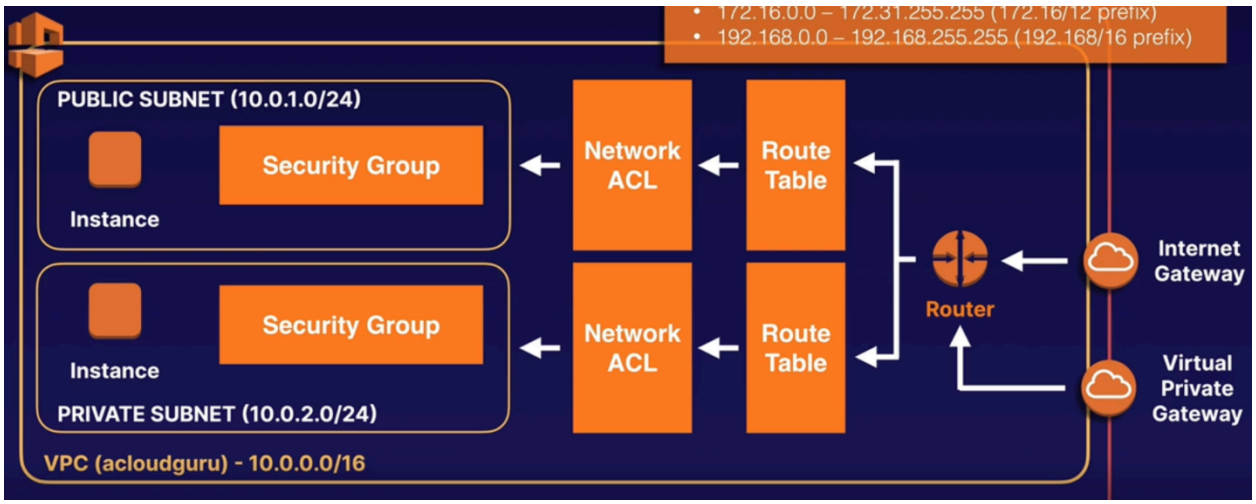
- Redundant inside the AZ
- No need to patch

- Not associated with security groups
- Automatically assigned a public IP address
- Starts at 5 Gbps, scales to 45 Gbps (not quizzed)

High Availability with NAT Gateways

- If you have resources in multiple AZ and they share a NAT gateway, if the AZ goes down in NAT gateway, resources in other AZ will lose Internet access
    - Create an AZ independent architecture, create a NAT gateway in each AZ, configure routing to ensure resource use the NAT gateway in the same AZ

Security Groups – A refresh



- Security Groups are virtual firewalls for EC2 instances
- *Security groups are stateful*
    - Responses to allowed inbound traffic are allowed to flow out, regardless of outbound rule

Controlling Subnet Traffic with Network ACLs

Network ACL

- Optional layer of security, acts as a firewall for controlling traffic in and out of one more subnets
- First line of defense
- VPC will comes with a default network ACL, all traffic is allowed outbound and inbound
- Custom network ACL denies all inbound and outbound traffic until rules are added
- Each subnet in a VPC must be associated with a network ACL
- You can block IP addresses using network ACLS, NOT security groups

Assigning Network ACLs in subnets

- You can associate network ACLs with multiple subnets
- A subnet can only be associated with 1 network ACL
- Network ACLs are a numbered list of rules, evaluated in order, starting with lowest number first
- Have separate inbound and outbound rules

- *Network ACLs are stateless*, responses to allowed inbound traffic are subject to rules for outbound traffic

Private Communications Using VPC Endpoints

- Virtual Private Cloud (VPC) Endpoints: Enables you to privately connect your VPC to supported AWS services. Is powered by PowerLink without requiring an internet gateway, NAT device, VPN connection or AWS Direct Connect
    - Does not require a public IP address to communicate with other AWS resources
    - Does not leave the Amazon network
- Endpoints are scaled, redundant, highly available VPC

2 types of Endpoints:

- Interface Endpoint: Elastic network interface with a private ip address, serves as an entry point for traffic headed to a supported service
- Gateway Endpoint: Similar to ANT gateways, a virtual device you provision. Supports connection to S3 and DynamoDB

Network Privacy with AWS PrivateLink

Sharing Applications Across VPCs

- Open VPC up to the Internet
    - Security considerations; everything is public in subnet
- Use VPC peering
    - Create and manage many different peering relationships
    - Whole network will be accessible.

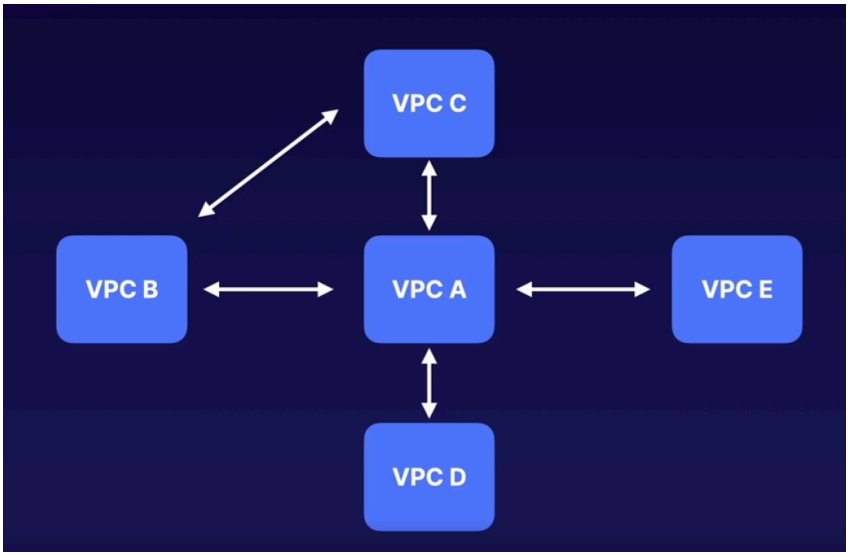
Using PrivateLink in VPC

- The best way to expose a service VPC to hundreds or thousands of customers VPC
- Doesn't require VPC peering; no route tables, NAT gateways, etc
- Requires a Network Load Balancer on the service VPC and an Elastic Network Interface (ENI) on the customer VPC

Building Solutions across VPCs with Peering

- Sometimes you may need to have several VPCs for different environments, and you could need to connect these VPCs. V
- VPC Peering: Allows you to connect 1 VPC with another via a direct network using a private IP address. Instances behave as if they were on the same network
    - You can peer between regions
    - Cannot perform transitive peering.
    - Can't have overlapping CIDER addresses

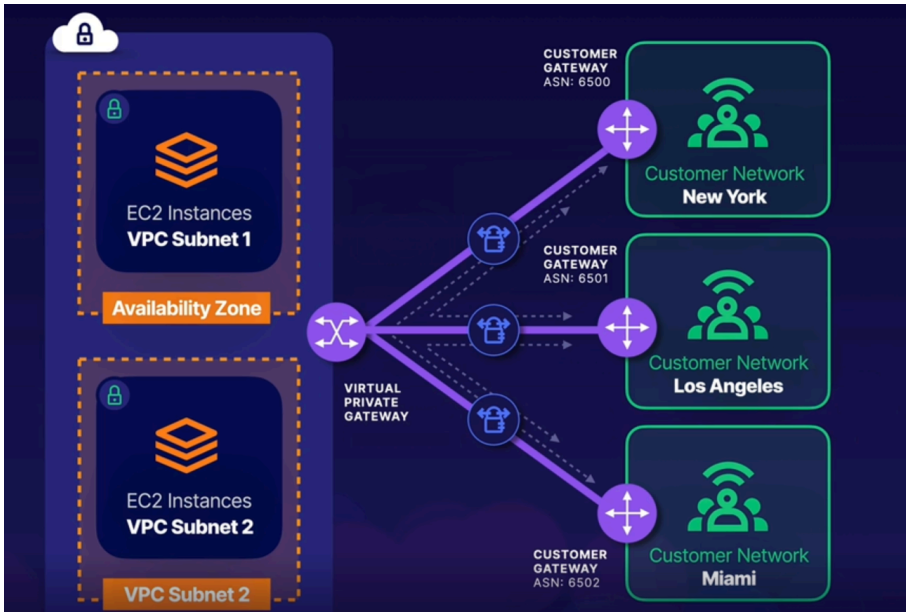
Transitive Peering:



VPC B can’t talk to VPC C through VPC A (transitive peering)

Securing Your Network with VPN CloudHub

- If you have multiple sites, each with its own VPN connection, you can use AWS VPN CloudHub to connect those sites together
- Operates on Hub and spoke model
- Low cost and easy to manage,
- Routed over public internet, but is encrypted between customer gateway and the AWS VPN CloudHub is encrypted



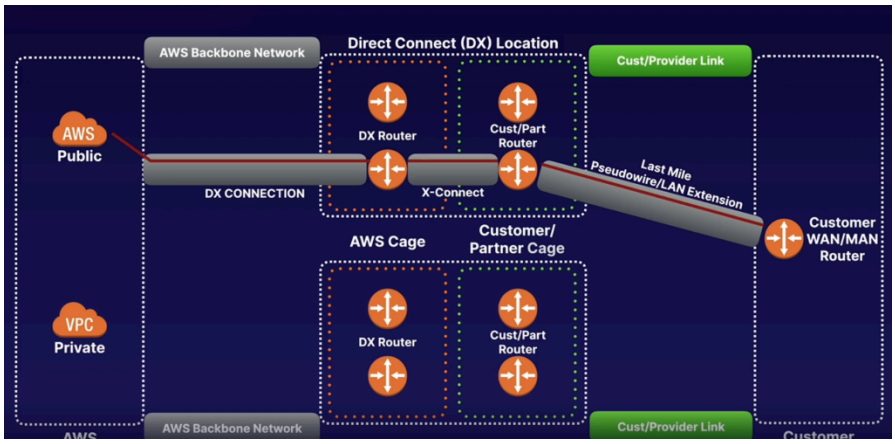
New York can talk to Miami through CloudHub.

Connecting On-Premises with Direct Connect

- Direct Connect: Cloud service solution which makes it easy to establish a dedicated network connection from your premises to AWS. Useful for high-throughput workloads

2 Types of Direct Connect

- Dedicated Connection: A physical Ethernet connection associated with a single customer
- Hosted Connection: A physical Ethernet connection that is an AWS Direct Connect Partner, who provisions the connection.



VPNs vs Direct Connect

- VPN allows private communication, but traverses public internet, can be slow
- Direct connect is fast, secure, reliable, and massive throughput

Simplifying Networks with Transit Gateway

- Transit Gateway: Connects VPCs and on-premise networks through a central hub. Simplifies your network
    - Allows for transitive peering between VPCs and on-premises data centers
    - Hub-and-spoke model
    - Works on a regional basis, but can span across multiple regions

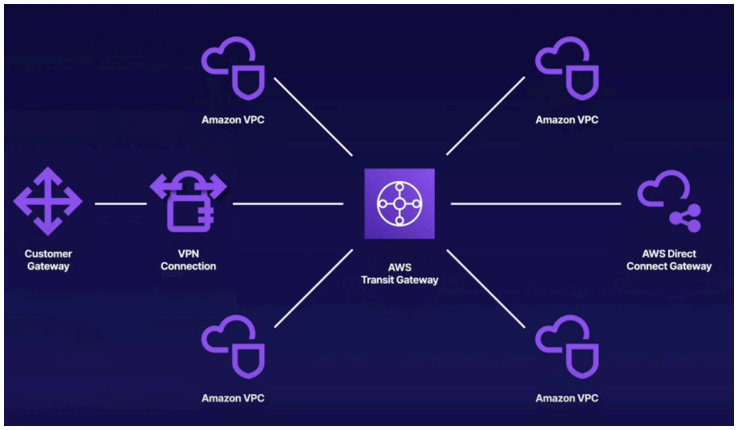
Transit Gateway Exam Tips

- Route tables to limit how VPCs talk to one another
- Works with Direct Connect as well as VPN connections
- Supports IP multicast

Normal topology (complex)



Transit Gateway (makes topology simple)



Chapter 10 - Route 53

Route 53 Overview

What is DNS?

- DNS is used to convert human-friendly domain names into Internet Protocol addresses

IPv4 vs IPv6

- IPv4 address are running out
- IPv6 address were created to solve the depletion issues
- Route 53 supports IPv4 & IPv6

Top-Level Domain

- The last word in a domain name represents the top-level domain. Second word is the second domain name
- Controlled by IANA in a root zone database, a database of all TLDs

Domain Registrars

- A registrar is an authority that can assign domain names under one or more top-level domains.

DNS Record Types:

- Start of Authority (SOA) Record: Name of the server supplied the data for the zone, admin of the zone, current version of the data file, and default number of seconds for TTL
    - Name Server (NS) Record: Used by top-level domain servers to direct traffic to the content DNS server that contains the authoritative DNS records



- [A] Record: Used to translate the domain name to IP address
- Time to Live Record (TTL): Length that a DNS record is cached on the resolving server or user's local PC is equal to the value of the TTL in seconds
- Canonical Name (CNAME): Resolve one domain to another. (m.acloud.guru -> acloud.guru)
    - Can't be for naked domains (acloud.guru)
    - Can be used for subdomains (m.acloud.guru)
- Alias Records (AWS Ecosystem Record): Map resource record in hosted zone to load balancer, CloudFront distributions, or S3 buckets
    - Can be used for naked domain names (acloud.guru)
    - *Always choose alias record over CNAME record for AWS*

Health Checks

- Set health checks on individual record sets
- If the record set fails, it will be removed from Route 53 until it passes the health check

Route 53 Traffic Flow

- You can use Route 53 traffic flow to build a routing system that uses a combination of geographic location, latency, and availability to route traffic from your users to your cloud or on-premises endpoints
    - Can build from template or scratch

7 Routing Policies in Route 53

- Simple Routing

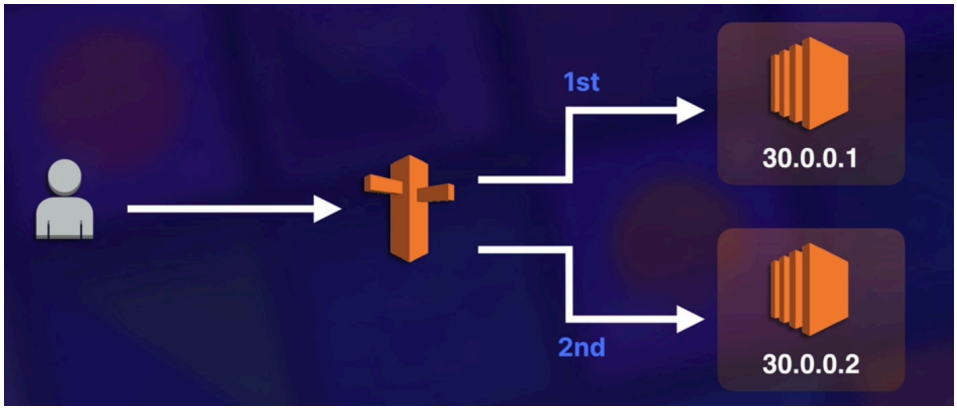
- Weighted Routing
- Latency-Based Routing
- Failover Routing
- Geolocation Routing
- Geoproximity Routing (Traffic-Flow Only) (won't be tested)
- Multivalue Answer Routing

Register a Domain Name

- You can buy a domain name directly with AWS
- It can take up to 3 days to register, depending on the circumstances

Simple Routing Policy

- If simple routing policy is chosen, there can be one record with multiple IP addresses.
    - If there are multiple values in the record, Route 53 returns all values to the user at random



Failover Routing Policy

- Failover routing polices are used when you want to create an active/passive set up
- Best used for primary site and disaster recovery site
- Route53 will monitor the health of your primary site using a health check

Geolocation Routing Policy

- Lets you choose where your traffic will be sent based on geographic location of your users (location of DNS queries originate)

Geo proximity Routing Policy

- Lets Route 53 route traffic to a resource based on the geographic location of users and resources
- Can optionally choose to route more or less traffic to a given resource, known as bias
- Must use Route 53 traffic flow
- *Won't see this exam*

Latency Based Routing Policy

- Allows traffic based on the lowest network latency for the end user (ex region)

- Create a latency resource record set for the EC2 resource in each region that hosts your website
    - Route 53 will respond with correct IP address of user

Multivalue Answering Routing Policy

- Configure to return multiple values, such as IP addresses for web server in response to DNS
- Also allows you to monitor health checks of endpoints (sends to the healthy network)
- Simple routing with health checks

Weighted Routing Policy

- Allows you to split traffic based on different weights assigned
    - Ex 10% of traffic go to us-east-1, 90% go to eu-west-1

Chapter 11 – Elastic Load Balancing (ELB)

What is Elastic Load Balancing?

- Automatically distributes incoming application traffic across multiple targets. Can be done in multiple AZs

3 Types of Load Balancers

- Application
    - Best suited for load balancing of HTTP and HTTPS. Layer 7 and application-aware.
    - Intelligent load balancer
- Network
    - Operates at layer 4. Capable of handling millions of requests per second, while maintaining ultra-low latencies.
    - Performance-based load balancer
- Classic
    - Legacy load balancer. You can balance HTTP/HTTPS applications and use Layer 7-specific feature.
    - Best for test and development

Health Checks

- Health checks periodically send requests to load balancers to test their status
- `InService`= Healthy state
- `OutOfService` = Unhealthy state
- Load balancer routes requests only to healthy state

Using Application Load Balancers

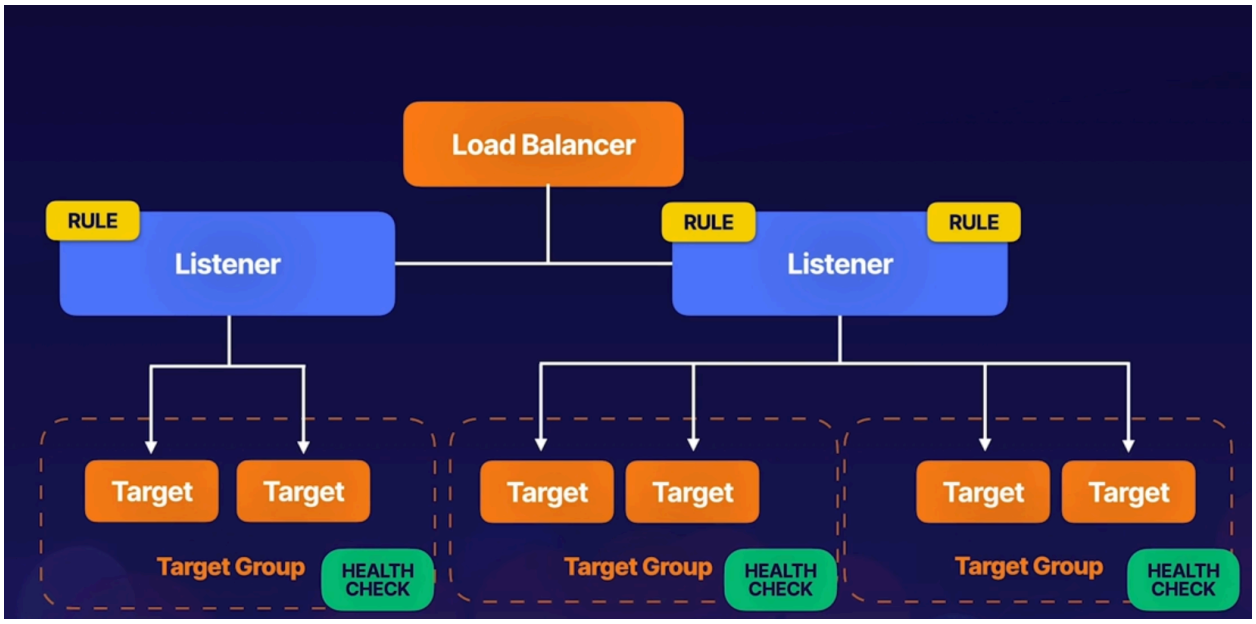
Layer 7 Load Balancing

- Application Load Balancer functions at application layer
- After the load balancer receives a request, it evaluates the listener rules in priority order to determine which rule to apply and then selects a target from the target group for the rule action
- By default, Load Balancers will have dynamic IP addresses. Can be set to static IP address

Application Load Balancer Components

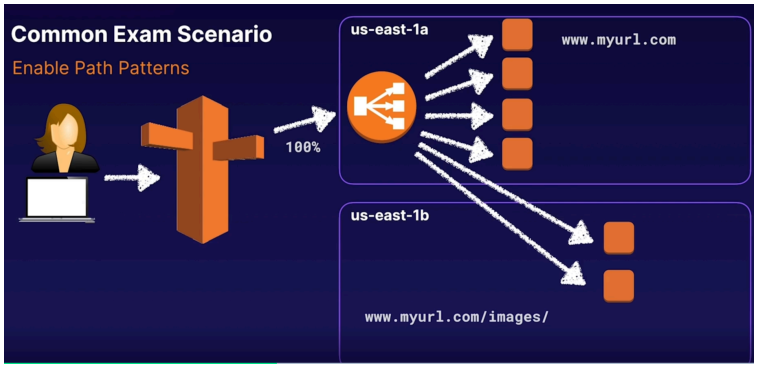
- Listeners
    - Checks for connection requests from clients using the protocol and port configured
    - Define rules that determine how the load balancer routes requests to its registered target
    - Each rule consists of a priority, one or more actions, and one or more conditions
- Rules
    - When a condition for a rule is met, then actions are performed.
    - You must define a default rule for each listener. d
- Target Groups
    - Routes requests to one or more registered targets (such as EC2 instances), using the protocol and port number you specify

Load Balancing Example



Path-Based Routing

- Common Exam Scenario
- Load Balancer can do intelligent routing based on paths



Limitations of Application Load Balancers

- Only supports of HTTP and HTTPS

HTTPS Load Balancing

- Must deploy one SSL/TLS server certificate on load balancer
- Decrypts requests from clients before sending them to targets. Actual decryption is done on the load balancer itself

Extreme Performance with Network Load Balancer

Layer 4 Load Balancing

- *Operates at layer 4 of OSI*
- *Handle millions of requests per second*
- *Use where other protocols not supported by Application*

How it works

- After receiving a connection request, it selects a target from the garget group
- It then attempts to open a TCP connection to the selected target on the port specified in the listener configuration

Listeners

- Checks for connection requests using the protocol and port you configure
- Then forwards the request to the target group

Target Groups

- Routes requests to one or more registered targets via port and protocols

Supported

- TCP, TLS, UDP, TCP_UDP

Encryption

- *Offload the work of encryption and decryption to your load balancer*
- *Must deploy exactly one SSL/TLS certification on the load balancer*

Classic Load Balancer

- Legacy load balancers
- Can load balance HTTP/HTTPS applications and use Layer 7-specific features such as X-Forwarded and sticky sessions
- Use strict Layer 4 load balancing for applications that rely on TCP

`X-Forward-For` Header

- When traffic is sent from a local balancer, the server access logs contain the IP address of the proxy or load balancer
- To see the original IP address of client, the X-Forwarded-For request header is used.
- Popular exam topic

Gateway Timeouts

- If application times out (stops responding), Load Balancer responses with a 504 error

Getting "Stuck" with Sticky Sessions

What are sticky sessions?

- Classic Load Balancer route 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
    - Ensures all requests from a user during the session are sent to the same instance
- Can enable sticky sessions for Application Load Balancer, but traffic will be sent at the target group level
    - If you want one target group, then you must use on target group per EC2 instance

Leaving the Load Balancer with Deregistration Delay

What is deregistration delay?

- Allows Load Balancers to keep existing connections open if the EC2 instances are de-registered or become unhealthy
    - Enables the load balancer to complete in-flight requests made to instance that are de-registered
- Can disable if you want to immediately close connections

Chapter 12 – Monitoring

CloudWatch Overview

What is CloudWatch?

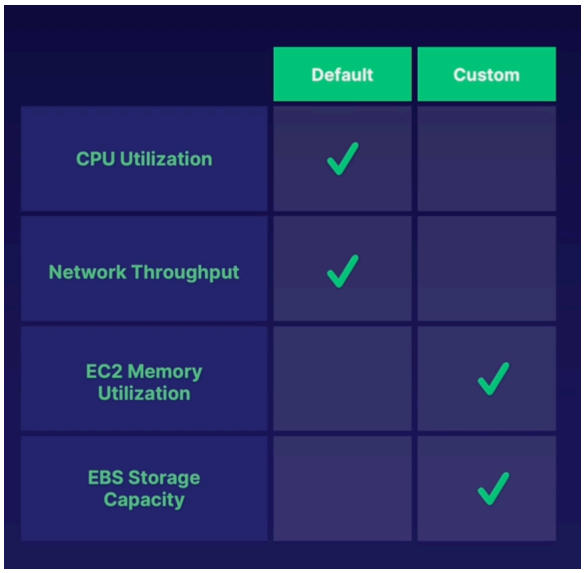
- Is a monitoring and observability platform that was designed to give us insight into our AWS architecture
- Allows us to monitor multiple levels of applications and identify potential issues

CloudWatch Features

- System Metrics
    - Metrics that you get out of the box (resource utilization, CPU, etc)
- Application Metrics
    - Get information from inside your EC2 instances (applications use (apache))
- Alarms
    - Alert you when something goes wrong

Metrics

- Default
    - Provided out of the box, do not require additional configuration
- Custom
    - Metrics will need to be provided by using CloudWatch agent installed



What You Need to Know:

- Anytime you think of monitoring, think CloudWatch
- Alarms are not set by default
- AWS can't see past the hypervisor level for EC2
- The more managed a service is, the more checks you get out of the box
- Standard is 5-minute whereas detailed is 1-minute

Application Monitoring with CloudWatch Logs

What is CloudWatch Logs?

- A tool that allows you to monitor, store, access log files from a variety of sources
- Give you the ability to query your logs and look for potential issues
- Logs go to CloudWatch Logs (unless you need real-time solution)
- Agent based, not automatic by default

3 CloudWatch Log Terms

1. Log Event
    a. A record of what happened, containing a timestamp and data
2. Log Stream
    a. A collection of log events from the same source creates a log stream, a continuous set of logs from a single instance
3. Log Group
    a. A collection of log streams (ex. Group all Apache web server logs across hosts together)

CloudWatch Log Features

- Filter Patterns
    - You can look for particular terms in a log (such as a 404 error)
- CloudWatch Logs Insights
    - Allows you to query SQL-like commands, an interactive solution
- Alarms
    - Once you have identified your trends or patterns, its time for an alert

4 Questions to Ask in the Exam

- What is the best tool to monitor with?
- Is that metric available by default?
- Where can I find those logs?
- Do I need to adjust my alarm threshold?

What to know for the Exam

- CloudWatch is the main tool for anything alarm related
- Not everything should not go through CloudWatch
- Know the intervals (metric deliver every 5-min (free), detailed monitoring every 1-min)
- CloudWatch Logs is the place for logs
- SQL? Think CloudWatch Log Insights
- Real-time solution, Kinesis is used

Chapter 13 – High Availability and Scaling

Horizontal vs Vertical Scaling

Vertical Scaling

- Scaling up vertically (changing disk drive size, RAM, CPU) on the same EC2 instance

Horizontal Scaling

- Spinning up new or multiple EC2 instances with the same settings
- Best for fault tolerance and high availability

3 W's of Scaling

- What do we scale?
    - Decide what resource to scale

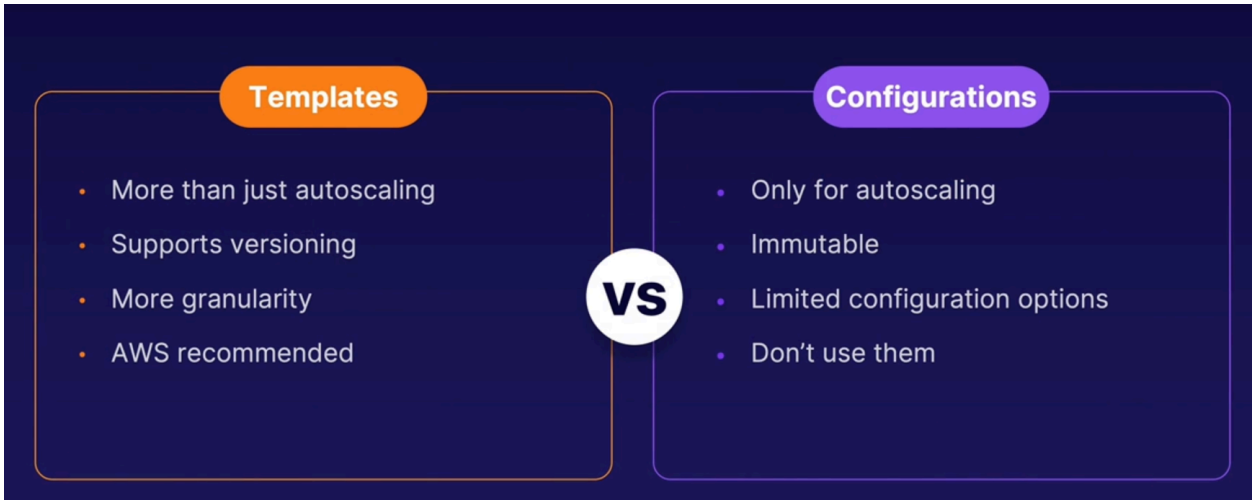
- Where do we scale?
    - Where does the scaling go? Database? Webserver?
- When do we scale?
    - How do we need more? CloudWatch alarms can tell us when it's time to add resources

What Are Launch Templates and Launch Configurations?

What are we scaling?

- Auto Scaling is only for EC2
- Use a launch template
- Launch Template: A launch template specifies all the needed settings that go into building out an EC2 instance. It is a collection of settings you can configure so you don't have to walk through the EC2 wizard over and over.
- Always use a launch template

Templates vs Configurations



What goes into a launch template?

- AMI
- EC2 Instance State
- Security Groups
- Network information
- Must change User Data in Launch Template

Scaling EC2 Instances with Autoscaling

- Where are we going to scale?
    - Autoscaling Groups
- Auto Scaling Group: Contains a collection of EC2 instances that are treated as a collective group for purposes of scaling and management

Steps to Auto Scaling

1. Define Your Template
    a. Have to pick from your available launch templates or launch configurations
2. Networking and Purchasing
    a. Pick network space and purchasing options. Using multiple AZs for high availability
3. ELB Configuration
    a. EC2 instances can be registered behind a load balancer. Auto scaling group can be set to respect the load balancer health checks
4. Set Scaling Policies
    a. Minimum, maximum, and desired capacity needs to be set to ensure you don't have too many or too few resources
5. Notifications
    a. SNS can act as a notification tool, allowing you to let someone know when scaling events occurs

Autoscaling Setting Capacity Limits

1. Minimum: Allows lowest number of EC2 instances you have online (2 is recommended)
2. Maximum: Set highest number of EC2 instances
3. Desired: How many instances do you want right now?

Things to Remember

- Auto Scaling and High Availability go together. Use multiple AZ. Utilize load balancers
- Auto scaling networking
- Use ELB health check
- Enable SNS notifications
- Balancing: Balance your EC2 instances across the AZ

Diving Deeper into Auto Scaling

When do we scale?

- Scaling policies

Example of Scaling Policies

Scaling Out

- Add 10 instances when memory is between 60 – 80%
- Add 15 instances when memory is between 80 – 100%

Scaling In

- Terminate 10 instances when memory is between 40 – 20%
- Terminate 15 instances when memory is between 20 – 0%

Periods

Warm-Up Period

- Stop instances from being places behind the load balancer
- Bake AMI for smaller provision time

Cooldown Period

- Pauses auto scaling for a set period

Avoid Thrashing

- You want to create instances quickly and spin them down slowly

Scaling Types

- Reactive Scaling
    - You're playing catchup. Once the load is there, you measure it and then determine if you need to create more resources
- Scheduled Scaling
    - If there is a predictable workload, create a scaling event to get your resources ready to go
- Predictive Scaling
    - Uses machine learning algorithms to determine when you'll need to scale. Reevaluated every 24 hours
- Steady State Groups
    - Allows us to create a situation where the failure of a legacy codebase or resource that can not be scaled automatically (set min, max, and desired to 1)

Things to Remember

- Cost implications of Auto scaling
    - Use EC2 Reserved-Instance for minimum count of EC2 instances
- Scale out aggressively
- Scale in conservatively
- Keep an eye out of provisioning time
- CloudWatch is your number 1 tool for alerting Auto scaling you need more or less

Scaling Relational Databases

4 Ways to Scale (Scale RDS)

1. Vertical Scaling
    a. Resizing the database from one size to another
2. Scaling Storage
    a. Storage can be resized, can't go down (best service is Aurora)
3. Read Replicas
    a. Create read-only copies of data can help spread the workload (read heavy workload)
4. Aurora Serverless
    a. Offload scaling to AWS. Excel with unpredictable workloads

Things to Remember

- Scaling vs Refactoring
    - Refactoring and changing to DynamoDB is a viable scaling option
- Read Replicas
    - Read-heavy workload = Read Replicas (have to update code)
- RDS storage only goes up (Aurora scales in 10 GB increments)
- Vertical Scaling
- Turn on Multi-AZ in production environments
- Aurora when a situation calls for a relational database

Scaling Non-Relational Databases

DynamoDB

- Scaling is simplified
- Set read and write speed and add your data
- Provisioned Scaling
    - Generally predictable workload
    - Most cost-effective
- On-Demand
    - Sporadic workload
    - Simply select on-demand
    - Less cost-effective (pay per read and write)

Things to Remember

- Keep cost in mind when scaling DynamoDB
    - Predictable -> Provisioned
    - Unpredictable -> On-demand
- It takes 24 hours per table switch

4 Questions to Ask in the Exam

- Is it highly available?
- Which is appropriate for the situation: horizontal or vertical?
- Is it cost effective?
- Would switching databases fix the problem?

Chapter 14 – Decoupling Workflows

Decoupling Workflows Overview

Tight Coupling: One EC2 instance talking directly to another EC2 instance. Simple to develop, no fault tolerance. Never want to tightly couple resources

Loose Coupling: One EC2 instance directly talking to an ELB, behind the ELB is multiple EC2 instances.

3 Services

- Simple Queue Service (SQS)

    - Fully managed message queuing service that enables you to decouple and scale microservices, distributed systems, and serverless applications
- Simple Notification Services (SNS)
    - SNS is a fully managed messaging service for both application-to-application and application-to-person for communication
- API Gateway
    - Fully managed service that makes it easy for developers to create, publish, maintain, monitor, and secure APIs at any scale. It's a front door

Messaging with SQS

What is poll-based messaging?

- Producer (front-end) creates a message, adds it to a queue, then the receiver (back-end) adds to the queue
- Short Polling: The instantaneous connection and termination to see if there are any messages in SQS
- Long Polling: The set amount of time a connection stays alive in the SQS service

SQS

- Is a messaging queue that allows asynchronous processing of work. One resource will write a message to an SQS queue, then another resource will retrieve that message from SQS
    - Asynchronous: Not a direct connection, uses the queue as a buffer

SQS Settings (*Important for Exam*)

- Delivery Delay
    - Queue hides the message for a set period of time
- Message Size
    - Up to 256 KB of text in any format (JSON, YAML, etc)
- Encryption
    - Messages encrypted in transit, but not at-rest
- Message Retention
    - Default is 4 days, up to 14 days. Purged from queue after
- Long vs short
    - Long polling isn't the default, but it should be
- Queue Depth
    - Can be a trigger for autoscaling
- Visibility Timeout
    - Message is locked in queue for 30 seconds by default, even when pulled by backend EC2 instance. It is private once downloaded by back-end.
    - If EC2 instance is successful, it deletes the message after set timeout (default is 30 seconds)
    - If EC2 instance does not successful finish the task outlined in message in the set period of time (30 seconds), the SQS message is public again for all other EC2 instances to see.

Sidelining Messages with Dead-Letter Queues

Dead-Letter Queues (DLQ)

- Another SQS queue to sideline the problematic messages in the SQS queue
- Its important to setup CloudWatch alarm to monitor queue depth
- Useable with SNS, create SQS DLQ for SNS topics
- Same retention windows (max of 14 days)

Ordered Messages with SQS First-in First-out (FIFO)

SQS Message Ordering

- Standard: Messages may be coming in and out at different orders
    - Can order messages with SQS standard, customized by you. Lots of work.
- FIFO: Guarantees ordering. First in, gets processed first
    - Message Group ID
        - Ensures messages are processed one by one

Standard Ordering

- Best-effort ordering
- Duplicates messages
- Nearly unlimited transactions per second

FIFO

- Guaranteed ordering
- No message duplication
- 300 messages per second

Delivering Messages with SNS

Push-Based Messaging

- Direct messages are sent to various AWS services
- SNS is a push-based messaging system
- Proactively deliver messages to endpoints subscribed to it
- Can be used to alert a system or a person

SNS Settings

- Subscribers
    - What will receive data from the topic? (Email, HTTP, Kinesis Data Firehouse, SQS, Lambda, SMS, application endpoint)
- Message Size
    - 256 KB
- DLQ Support
    - Messages that fail are delivered to SQS DLQ
- FIFO or Standard
    - FIFO only supports SQS as a subscriber
- Encryption

    - Messages are encrypted transit by default, but can add at-rest
- Access Policy
    - A resource policy can be added to a topic

Things to Remember:

- SNS = Alert
- Push = SNS
- SNS and CloudWatch are best friends

Fronting Applications with API Gateway

What is API Gateway?

- Fully managed service that allows you to easily publish, create, maintain, monitor, and secure your API. Allows you to put a safe in the “front door” on your application
- API is the front door of an application

Reasons to Use an API Gateway

- Security: Protect your endpoints by attaching a web application firewall (WAF)
- Stop Abuse: Users can easily implement DDoS protection and rate limiting
- Ease of Use: API Gateway is simple to set up

Not on Exam:

- API Gateway processing dynamic content
- S3 store static content

Things to Remember

- Anytime the exam talks about creating or managing an API, think API gateway
- DDoS protection by using WAF
- Versioning
- No baking credentials in the code

Things to Remember in the Exam

- Is it a synchronous or asynchronous workload?
- What type of decoupling makes sense?
- Does the order of messages matter?
- What type of application load will you see?

SQS

- SQS can duplicate messages only once. Multiple duplicates messages could be a Visibility Timeout issue
- SQS are not bidirectional. If you need to return to the instance that sent a message, you will need a second queue
- Know the defaults of SQS
- Does order matter? Think SQS FIFO

- SNS delivers Cloudwatch alarms

Chapter 15- Big Data

Exploring Large Redshift Databases

What is Big Data?

- Volume: From terabytes to petabytes
- Variety: Range in sources and formats
- Velocity: Require speed, data needs to be collected, stored, processed, and analyzed within a short period of time

What is Redshift?

- Fully managed, petabyte-scale data warehouse service in the cloud
- Very large relational database
- Don't use Redshift in place of RDS
- Single AZ-deployments. Not highly available

Redshift Cloud Data Warehouse

- Size: Incredibly big, hold up to 16 PB of data
- Relational: Standard SQL and business intelligence (BI)
- Usage: It is not a replacement for standard RDS database

Things to Remember

- Used for BI purposes and applications

Processing Data with Elastic MapReduce (EMR)

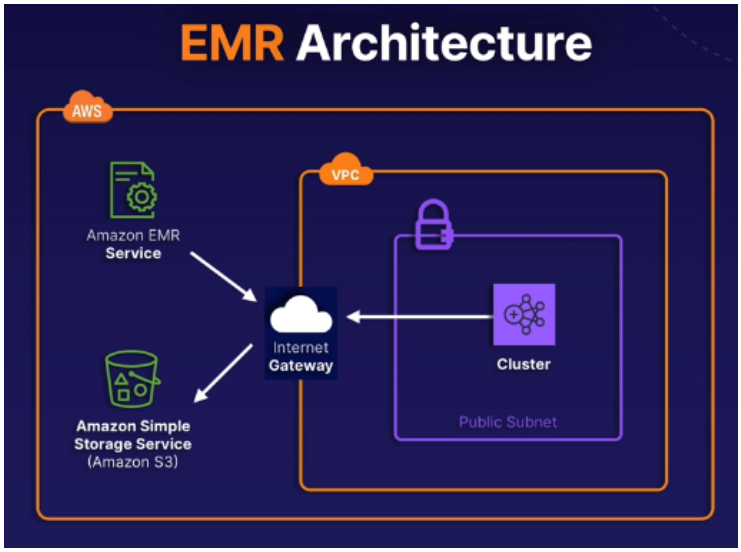
What is ETL?

- ETL: Extract Transform Load

What is EMR?

- Helps with ETL process
- EMR is a managed big data platform which allows you to process vast amounts of data using open-source tools
- Fleet of EC2 instances is managed by open-source tools
- Used to transform data
- Lives inside a VPC

EMR Architecture



Streaming Data with Kinesis

What is Kinesis?

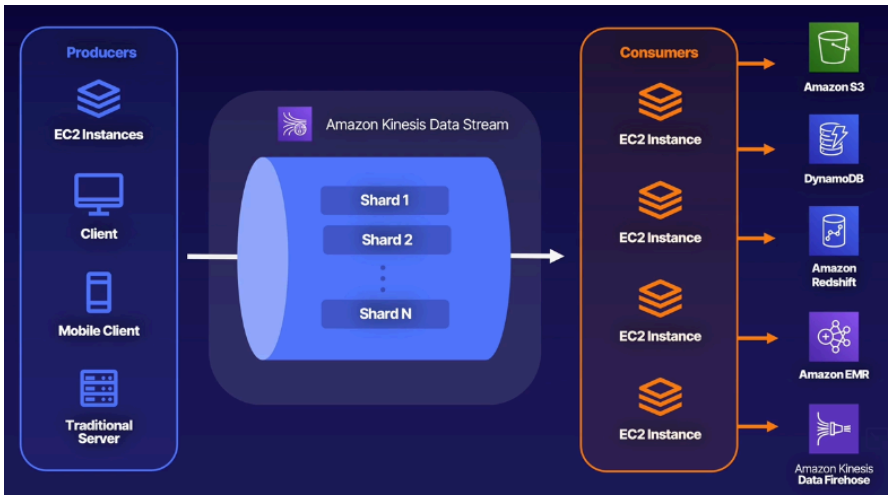
- Allows you to ingest, process, and analyze real-time streaming data
- Huge data highway connected to your AWS account
- Kinesis is the only service with real-time response

2 Versions

- Data Streams: Real-time streaming for ingesting data
    - Does not automatically scale
- Data Firehose: Near real time, plug and play with Aws architecture
    - Automatically scales

Data Streams vs Data Firehose

Data Stream



Data Firehose



Kinesis Data Analytics

- If you want to analyze data (transform data), you can tie Data Analytics into a Kinesis pipeline
- Analyze data using standard SQL
    - Easy to Use
    - No Servers
    - Cost Effective

When looking for broker, which do you pick?

- SQS: Simple, doesn't require configuration. ***Doesn't offer real-time*** *message delivery*
- Kinesis: A bit more complicated to configure, used for big data applications. *Provides* ***real-time*** *communication*
    - *o* Real-Time -> Think Data Stream
    - *o* Near Real-time -> Think Data Firehose

Amazon Athena and AWS Glue

What is Athena?

- An interactive query service that makes an easy to analyze data in S3 and SQL
- Directly query data in your S3 bucket without loading it into a database
- "Serverless SQL"

What is Glue?

- A serverless data integration service that makes it easy to discover, prepare, and combine data. Perform ETL workloads without managing underlying servers
- Can help design a schema for your data
- "Serverless ETL"

Using Them Together Architecture



Things to Remember

- Scenario that's looking for a serverless SQL solution, Athena is the best choice, the only choice that allows you query in an S3 bucket
- Both are managed serverless services

Visualizing Data with QuickSight

What is QuickSight?

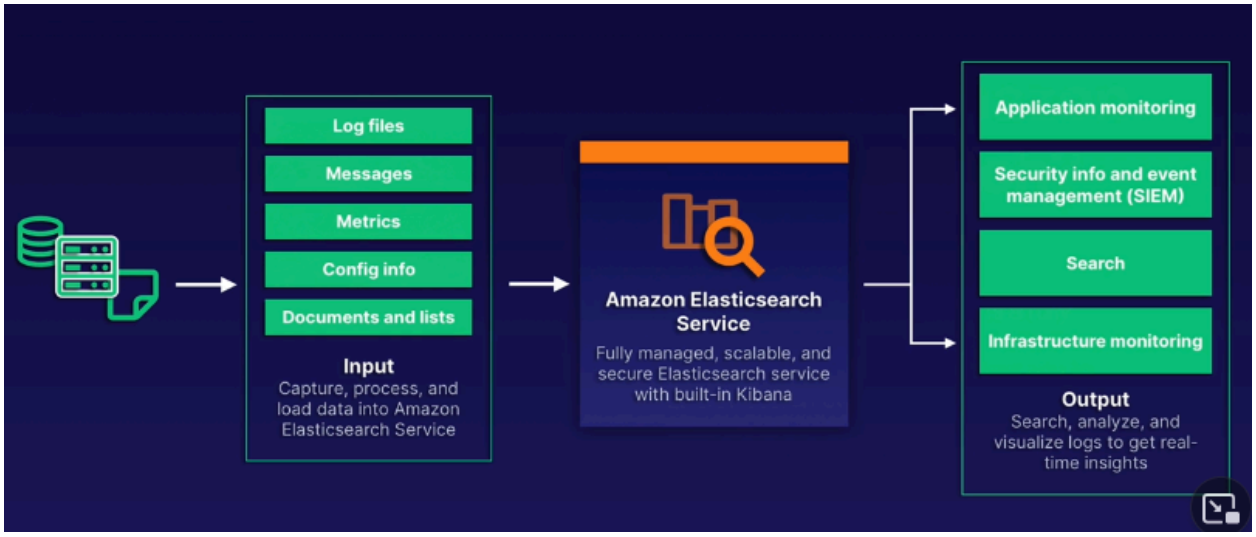
- Fully managed business intelligence (BI) data visualization services
- Allows you to visualize data and create dashboards
- Interpreting, visualize BI Data – think QuickSight

Analyzing Data with Elasticsearch

What is Elasticsearch?

- Fully managed version of the open-source application Elasticsearch
- Allows you to quickly search over your stored data and analyze data you get back
- "A search engine" for data
- Used in an Elasticsearch, Logstash, and Kibana (ELK) stack
- Fully managed service
- Third party log solution that is not an Amazon proprietary tool

What Can It Do?



Things to Remember

- Elasticsearch = ELK = Logs
- Used as a Distractor as the exam. CloudWatch Logs are often used

4 Questions to Ask in the Exam

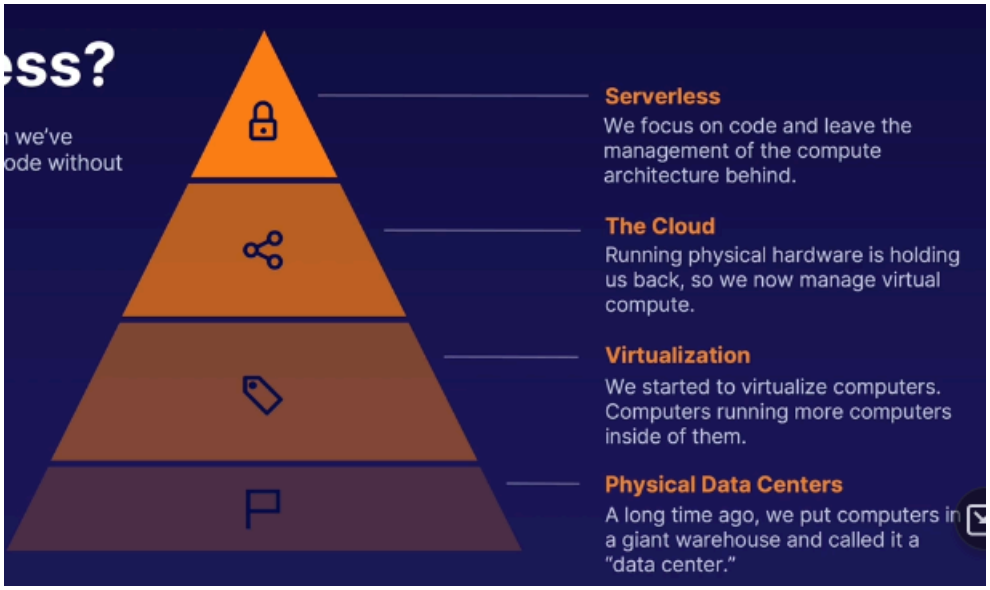
- What kind of database works?
- How much data do you have?
- Is serverless a requirement?
- How do we optimize cost?

Chapter 16 – Serverless Architecture

Serverless Overview

History of Computing

- Serverless?
    - Doesn't mean we have figured out how to run code without computers
- Serverless
    - Focus on the code, leave management of the compute architecture behind
    - Focus on answers that move away from unmanaged AWS architecture like EC2



Serverless Benefits

- Ease of Use
- Event Based
    - Compute resource can be brought online in response to an event happening
- Billing Model
    - "Pay as you go"

2 Main Serverless Services

- Lambda: Write your code, build your function and launch
- Fargate: Run containers in Fargate, no servers, no patching, no mess

Computing with Lambda

What is Lambda?

- Serverless compute service that lets you run code without provisioning or managing the underlying servers. It's like running code without computers.

Lambda Functions

- Build an environment to help run your code.

Building a Lambda Function

1. Runtime: Pick from an available runtime or bring your own. Your environment your code will run in
2. Permissions: If a Lambda function needs to make an AWS API call, you will need to attach a role
3. Networking: Optionally define a VPC. But can run in a "small like EC2 instance"
4. Resources: Defining the amount of memory will allocate how CPU and RAM your code gets

5. Trigger: What's going to trigger your Lambda function? Defining a trigger will kick Lambda off if the event occurs

Things to Remember

- Maximum runtime is 15 minutes
- Maximum 10 GB of RAM
- If there is a need to automatically remove entries or start and stop EC2 instances, think Lambda
- Used for microservices, small and lightweight functions
- Plays a major role in AWS, easily integrated into many services
- Lambda functions can run inside or outside a VPC
    - If it needs to make a private API or database calls, put in VPC

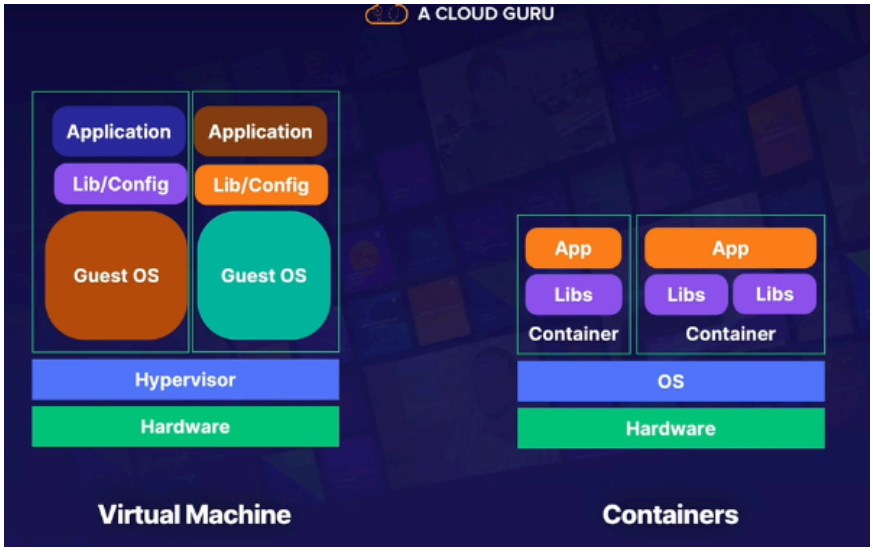
Container Overview

What is a container?

- Is a standard unit of software that packages up code and all its dependences, so the application runs quickly and reliably from one computing environment to another

Why containerize?

- Cut down the Guest OS, virtualizing the Gust OS instead of the hardware itself



Container Terminology

- Dockerfile: Text document that contains all the commands or instructions that will be used to build an image. What applications to download, where to files go, etc?
- Image: Immutable file that contains the code, libraries, dependencies, and configuration files needed to run an application
- Registry: Stores Docker images for distribution. Can be private or public
- Container: A running copy of the image that has been created

Things to Remember:

- Containers are generally seen as more flexible
- Easy to use (migrate on-prem to AWS, vise versa)
- Dev is prod and prod is dev

Running Containers in ECS and EKS

Problems with Containers

- What happens when you have 100s of containers? How do you manage them?

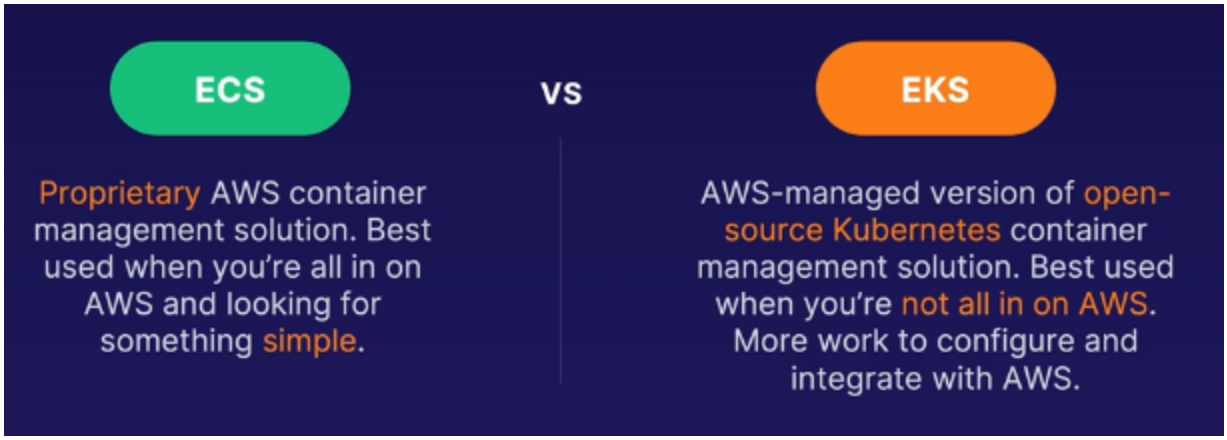
Why ECS?

- Elastic Container Service: Management of containers at scale
- Has ELB Integration: Go online and offline
- Role Integration: Have individual roles attached to them, making security a breeze

Using Kubernetes

- Kubernetes is an open-source management and orchestration platform. An open-source alternative
- Can be used on-prem and in the cloud
- AWS has a managed version of Kubernetes Service, Elastic Kubernetes Service

ECS vs EKS



Things to Remember

- ECS is preferred
    - Only if it needs an on-prem or Kubernetes
- Favor AWS-designed services

Removing Servers with Fargate

Problems with Servers

- Patching, updating EC2 instances is time consuming

What is Fargate?

- Serverless compute engine for containers that work with ECS and EKS
- AWS own and manage infrastructure
- Requires use of ECS or EKS
- Linux-only workload

EC2 vs Fargate

EC2

- Responsible for the underlying OS
- EC2 is a better pricing model
- Excels for long-running containers (24 hours, 365 days a year)
- Multiple containers share the same host

Fargate

- No operating System
- Pay based on resources allocated and time ran
- Excels in short-running tasks
- Isolated environment

How does Lambda Compare?

- Fargate has more consistent workloads
- Allows docker use across the organization and greater level of control by developers
- Lambda is great for unpredictable or inconsistent workloads
- Perfect for applications with a single function

Things to Remember

- Lambda
    - Lightweight functions, run very quickly
        - Ex responding to an event,
- Fargate
    - Containers that don't need all the time
- EC2
    - Containers that need to run all the time, best pricing option
- Fargate doesn't work by itself. ECS and EKS is required

Amazon EventBridge (CloudWatch Events)

What is EventBridge?

- Is a severless event bus. It allows you to pass events from a source to an endpoint. It's the glue that holds your serverless application. "CloudWatch Events 2.0"

Creating a Rule

1. Define Pattern: Do you want the rule to be invoked based on the event happening?

2. Select Event Bus: Is this going to be AWS-based? A custom event? Or a partner?
3. Select Your Target: What happens when this event kicks off? Does Lambda function? Post to SQS queue. Send an email?
4. Tag: Tag everything
5. Sit Back: View what happens

Things to Remember

- EventBridge holds together a serverless application and Lambda function. Any API call that happens in AWS can alert a Lambda function, or a verity of different endoints, that something has happened
- Very common use is triggering Lambda functions when an AWS API call happens
- Fastest way to respond to an API call

4 Questions to Ask

- Is the application right for containers? (use for microservices)
- Do you need those servers?
- Is the application AWS specific?
- How long does the code need to run?

Things to Remember for the Exam

- Lambda loves roles, always a attach a role to the function
- You will commonly see questions asking what can kick off a Lambda function
- Any AWS API call can trigger to kick off an EventBridge
- Open-source = Kubernetes
- Fargate needs ECS or EKS
- Containers are flexible
- Know the basics

Chapter 17 – Security

DDoS Overview

- An attack that attempts to make your website or application unavailable to your end users
    - Can be achieved through large packet floods, using reflection and amplification techniques

DDoS Types

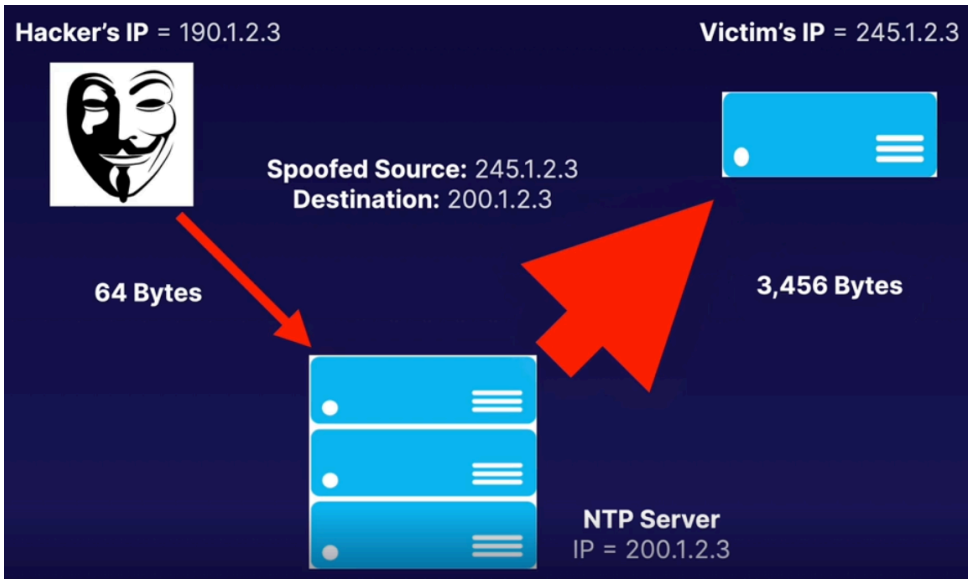
Layer 4 DDoS Attack

- The Process: Establish a TCP connection over a 3-way handshake. The client sends a SYN packet to a server, the server replies with a SYN-ACK, and then the client responds with a ACK.
- SYN Flood attack: Sends an overwhelming amount of SYN packets and then ignores the SYN-ACK packets. This causes the server to use up resources waiting for the ACK message.

Amplification Attack

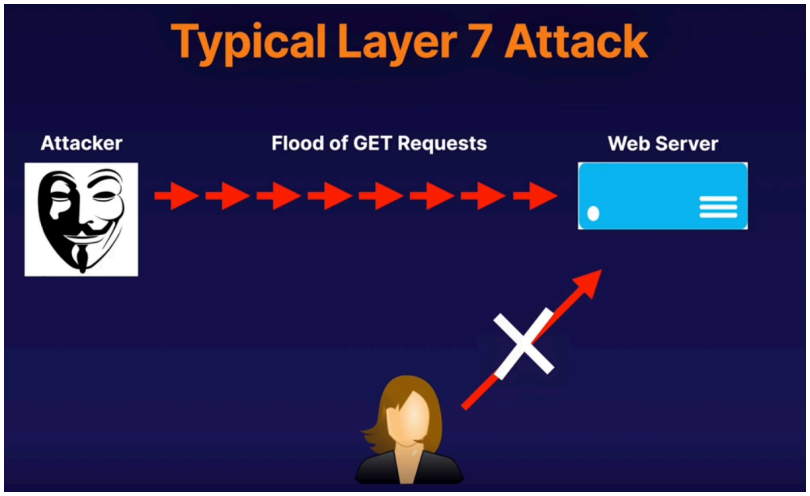
- Can include NTP, SSDP, DNS, CharGEN, SNMP attack.

- Amplification Attacks: A spoofed IP address is used to query a server. Server will respond to that request with a greater payload than the initial request



Layer 7 Attack

- Receives a flood of GET or POST requests, large number of compromised computers



Logging API Calls with CloudTrail

What is CloudTrail?

- Increases visibility into your user and resource activity by recording AWS Management Console actions and API calls
- You can identify which users and accounts, the source IP address, and when API calls occurred
- Logs in S3
- Monitoring and logging what goes on in your AWS account

CloudTrail Logging

- Metadata around API calls
- The identity of API caller
- The time of API
- Source IP address
- Request parameters
- Response elements returned by the service

CloudTrail

- After-the-fact incident investigation
- Near real-time intrusion detection
- Industry and regulatory compliance
- “CCTV” for your AWS account

Protecting Applications Using Shield

What is Shield?

- Free DDoS protection. All customers on ELB, CloudFront, and Route 53
- *Prevents Layer 3 and Layer 4 attacks only*
- *Layer 7 Application attacks are prevented from a WAF*

Shield Advanced

- Provides enhanced protections against more sophisticated attacks
- Offers always-on, flow-based monitoring of network traffic and active application monitoring
- Gives you access to 24/7 to DDoS Response Team (DRT) to help manage and mitigate application-layer DDoS Attacks
- Protects your AWS Bill
- 3,000 USB per month

Filtering Traffic with AWS WAF

What is a WAF?

- Web application firewall that lets you monitor HTTP and HTTPS requests that are forwarded to Amazon CloudFront or Application Load Balancer
- Configure what IP addresses are allowed to make a request or query string parameters
- Will give a 403 status code
- ***Operates at Layer 7***

Layer 7 Behaviors

- Allow all request expect the ones you specify
- Block all requests except the ones you specify
- Count the requests that match the properties you specify

Define Conditions

- IP address
- Country Origin
- Values in request headers
- Presence of SQL code (SQL Injection)
- Presence of a script (cross-site scripting)
- Strings that appear in requests (regex patterns)

Guarding Your Network with GuardDuty

What is GuardDuty?

- Threat detection service that uses machine learning to continuously monitor for malicious behavior. Centralized threat detection
    - Unusual API calls, malicious IP
    - Attempts to disable CloudTrial Logging
    - Unauthorized deployments
    - Compromised instances
    - Reconnaissance by would-be attackers
    - Port scanning and failed logins

GuardDuty Features

- Alerts appear in GuardDuty console and CloudWatch Events
- Receives feeds from third parties like Proofpoint and CrowdStrike
- Monitors CloudTrail logs, VPC Flow Logs, and DNS logs
- Updates a database of known malicious domains using external feeds from third parties
- Can be used to trigger Lambda function to address a threat

Threat Detection

- It takes 7 – 14 days to create a baseline
- 30 Day free trail

Monitoring S3 Buckets with Macie

What is Macie?

- Personally Identifiable Information (PII): Personal data used to establish an individual's identity.
    - Home address, email address, SSN
- Macie uses machine learning and pattern matching to discover sensitive data stored in S3
    - Uses AI to recognize if your S3 objects contain sensitive data
    - Alerts you on unencrypted and public buckets
    - Great from frameworks like HIPPA and GDPR
    - Can automate remediation actions

Macie Alerts

- Alerts are sent Amazon EventBridge
- Can be integrated with AWS Security Hub
- Can be integrated with other AWS services

Securing Operating Systems with Inspector

What is Amazon Inspector?

- *Automated security assessment service that helps improve the security and compliance of applications deployed on AWS*
    - *Assess applications for vulnerabilities and deviations from best practices. Host and network assessments are performed*
- Assessment Findings: A list with a security findings prioritized by security level
    - Network Assessments: Network configuration analysis.
        - Inspector agent not required.
    - Host Assessments: Vulnerable Software (CVE), host hardening (CIS), and security best practices
        - Inspector agent is required

How does Inspector work?

- Create assessment target
- Install agents on EC2 instances
- Create assessment template
- Perform assessment run
- Review findings against rules

Managing Encryption Keys with KMS and CloudHSM

What is AWS KMS?

- Managed service that makes it easy for you to create and control encryption keys
    - Integrates with EBS, S3, and RDS
- Provides with centralized control over the life cycle and permissions of the keys

Customer Master Key (CMK)

- Logical representation of the master key. Start using the KMS service by requesting a CMK
    - Contains the key material used to encrypt and decrypt data
    - The lifecycle of the CMK can be controlled

Hardware Security Module (HSM)

- Is a physical computing device that safeguards and managed digital keys and performs encryption and decryption

3 Ways to Generate a CMK

- AWS creates the CMK for you. Generated within HSMs managed by AWS
- Import key material from your own key management infrastructure
- Have it used by AWS CloudHSM cluster, part of custom key store feature

Key Rotation

- Can choose to automatically rotate CMKs every year, provided those keys were generated within AWS KMS HSMS

    - Not supported for imported keys, asymmetric keys, or keys generated in an AWS CloudHSM cluster

Policies

- Policies are document that describe who has access to what AWS KMS
    - Policies attached to an IAM identity are referred to as identity-based policies
    - Policies attached to other kinds of resources are referred to as resource-based policies

Key Policies

- All KMS CMKs must have a resource-based key policy

3 Ways to Control Permission

- Use a key policy
- Use IAM policies in combination with the key policy
- Use grants in combination with the key policy

CloudHSM

- Cloud-based HSM used to easily generate and use your own encryption keys on the AWS Cloud
    - Physical device, entirely dedicated to you

KMS vs CloudHSM

KMS

- Shared tenancy of underlying hardware
- Automatic key rotation
- Automatic key generation

CloudHSM

- Dedicated HSM to you
- Full control of underlying hardware
- Full control of users, groups, keys
- No automatic key rotation

Storing Your Secrets in Secrets Manager

What is Secrets Manager?

- Service that securely stores, encrypts, and rotates DB credentials and other secrets
    - Encryption in transit and at rest using KMS
    - Automatic rotation of credentials
    - Apply find-grained access control using IAM policies
    - Highly scalable

Applications use the Secrets Manager API

Storage

- RDS credentials
- Non-RDS database
- SSH keys, API keys, etc

If enabled rotation, Secrets Manger immediately rotates the secrets once for testing.

- If credentials are hard-coded into your applications, this is not recommended
- Ensure applications instances are configured before enabling credential rotation

Storing Your Secrets in Parameter Store

- Parameter Store: Capability of AWS Systems Manger that provides secure, hierarchical storage for configuration data management and secrets management.
    - Passwords, DB strings, AMI IDs, license codes
    - Parameter Store is free

2 Limits of Parameter Store

- Limit of 10,000 parameters
- No rotation of keys

Temporarily Sharing S3 Objects Using Presigned URLs and Cookies

All objects in S3 are private by default. Object owner can optionally share objects using a presigned URL with temporary access

- Must provide security credentials, bucket name, object key, and indicate the HTTP method and the expiration date
- Anyone who receives the presigned URL can then have access the object. You can share one file in a private bucket

Presigned Cookies

- Provide access to multiple restricted files
- Cookies will be saved on user's computer, and can browse the entire contents of the restricted contents

Advanced IAM Policy Documents

Amazon Resource Names (ARNs)

- `arn:partition:service:region:account_id`



IAM Policies

- JSON document that defines permissions
- Identity Policy (users)
- Resource Policy (S3, RDS, etc)
- No effect until attached
- Each statement matches an AWS API request



```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "SpecificTable",
            "Effect": "Allow",
            "Action": [
                "dynamodb:BatchGet*",
                "dynamodb:DescribeStream",
                "dynamodb:DescribeTable",
                "dynamodb:Get*",
                "dynamodb:Query",
                "dynamodb:Scan",
                "dynamodb:BatchWrite*",
                "dynamodb:CreateTable",
                "dynamodb:Delete*",
                "dynamodb:Update*",
                "dynamodb:PutItem"
            ],
            "Resource": "arn:aws:dynamodb:*:*:table/MyTable"
        }
    ]
}
```

Permission Boundaries

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

Things to Remember

- Not explicitly allowed == implicitly denied
- Explicit deny > everything else
- Only attached policies have effect
- AWS joins all applicable policies

AWS Certificate Manager

- Allows you to create, manager, and deploy public and private SSL certs for use with other AWS services
    - Integrates into ELB, CloudFront, API Gateway
- You do not have to pay for SSL certificates
- Automated Renewals and Deployment
- Easy to setup

Chapter 18 – Automation

Why Use Automation?

- Manual builds are a gamble. You'll build it correctly, but you're more likely to make some costly errors

Benefits of Automation

- Time Savings
- Security
- Consistency
- Immutable

Automation Solutions in AWS

- CloudFormation
    - Infrastructure as code
- Elastic Beanstalk
    - Deployment and scaling for web applications and services
- System Manager
    - Easily patch, update, manage, and configure EC2 instances along with on-prem architecture

Things to Remember

- *Replace a manual step with automated tools in AWS*

CloudFormation

CloudFormation Overview

Step 1: Write Code

- CloudFormation is a declarative programming language. It supports JSON or YAML

Step 2: Deploy

- Upload your template to CloudFormation. AWS will automatically

Things to Remember:

- CloudFormation is perfect for creating immutable architecture (consistency)
- Parameters: Questions / Values users fill in to customize the CloudFormation Template (store AMI or snapshot IDs here)
- Mappings: Values that fill themselves in (store AMI or snapshot IDs here)
- Resources: Where resources go
- *Hard-coded values and resources IDs can be the reason for templates to fail*
- If there is an error, CloudFormation rolls back to the last known good state
- It's just API calls

- When a CloudFormation stack is deleted or modified (IaC), the physical infrastructure will be deleted or modified

Elastic Beanstalk

What is Paas?

- PaaS (Platform as a Service): Single-stop application deployment model. You bring your code, and the provider builds everything for you, deploys your application, and then manages it going forward
    - If you have a web application you have written, everything else can be managed through Elastic Beanstalk

Elastic Beanstalk

- *Your one stop shop for everything AWS related*
    - Bring your own code, supports Windows, containerized, and linux applications
- Automation: Elastic Beanstalk automates all your deployments. You can add templates to your environment
- Deployment: It will handle your deployments for you.
- Management: Will handle building out the insides of your EC2 instances for you. You don't have to log into the hosts yourself
- Elastic Breanstalk is not severless

Systems Manager

What is Systems Manager?

- Suite of tools designed to let you view, control, and automate both your AWS architecture and on-premises resources
    - Free set of tools
- Primary method used in configuring the Inside of an EC2 instance

Features of Systems Manager

- Automation Documents (also known as Run Books): Can be used to control your instances or AWS resources
- Run Command: Execute commands on your hosts
- Patch Manager: Manages your application versions
- Parameter Store: Securely store your secret values
- Hybrid Activations: Control your on-prem architecture
- Sessions Manager: Remotely connect and interact with your architecture

Cross-Account or Cross-Region Automation Troubles

- Do not hard code IDs or soft limits

Chapter 19 - Caching

Caching Overview

Where to Cache

- External content (images, videos, and static content)
- Internal content (databases)

AWS Caching Options

- CloudFront (external CDN)
- ElastiCache
- DAX
- Global Accelerator

Things to Remember

- AWS likes caching, choose answers that have caching
- Speed benefits
- Internal vs External
- Caches go in front of everything

Global Caching with CloudFront

- Fast CDN services that securely delivers data, videos, applications, and APIs to customers globally. It helps reduce latency and provide higher transfer speed using AWS edge locations
    - Instead of storing static content in one location, you can spread your contents to the 100+ edge locations

Important CloudFront Settings

- Security (defaults to HTTPS connection). Implements secure connection
- Global Distribution. Can't pick specific countries. Just general areas
- Endpoint Support (and non-AWS endpoints)
- Expiring Content (TTL) (before it gets refreshed)

Things to Remember

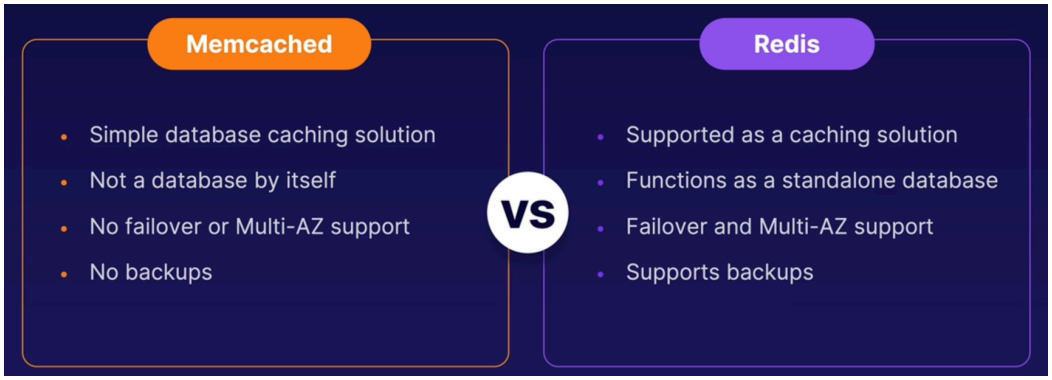
- CloudFront is Fast
- On-Site Support and AWS
- Can be used for blocking individual countries, but WAF is better
- Cannot pick specific edge locations for your distributions

Caching Your Data with ElastiCache and DAX

What is ElastiCache?

- AWS managed version of two open-source technologies Memcached and Redis. Avoid a lot of common issues using ElastiCache

Memcached vs Redis

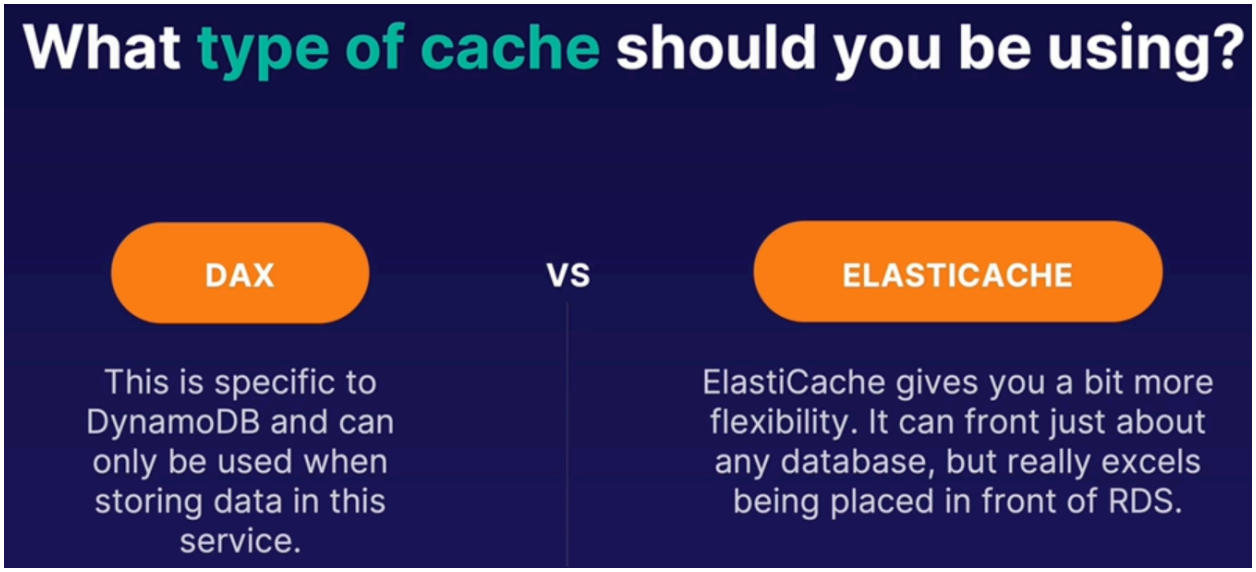


- Traditionally set in front of a relational database

DynamoDB Accelerator (DAX)

- In-Memory Cache: DAX can reduce DynamoDB response times from milliseconds to microseconds
- Lives inside the VPC
- You are in control, size of node, count for cluster, TTL for data, and maintenance windows for changes and updates

What type of cache should you be using?



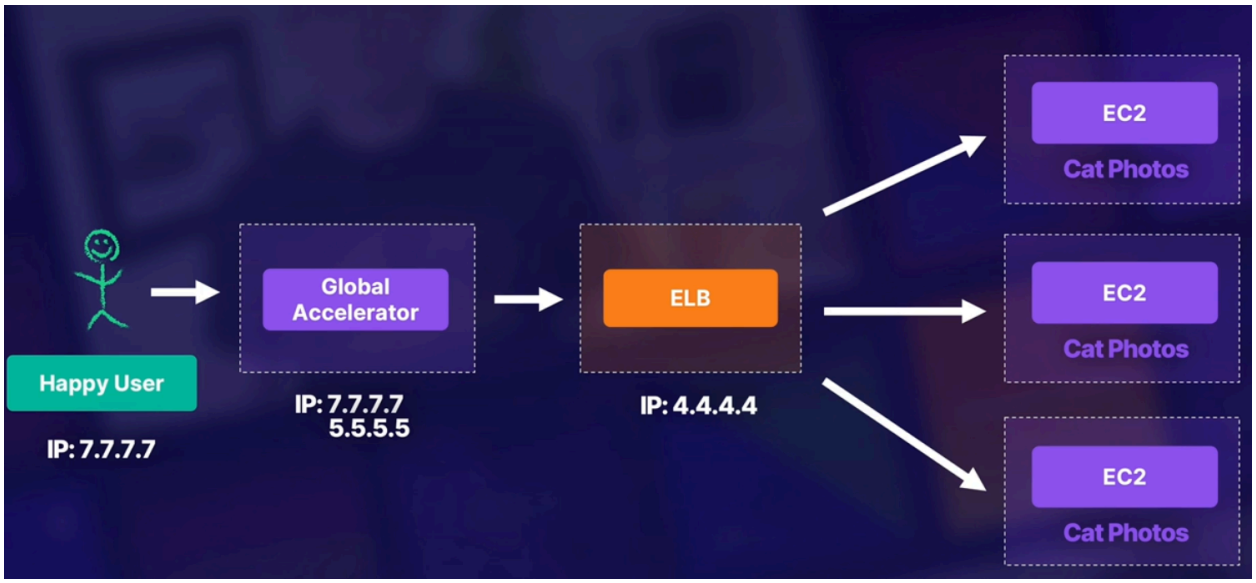
Things to Remember

- Favor answers that include a database caching solution
- Always Cache
- Redis can be more than just a cache, it can be a NOSQL database

Fixing IP Caching with Global Accelerator

What is Global Accelerator?

- A networking service that sends your user's traffic through AWS's global network infrastructure. It can increase performance and help deal with IP caching
    - 2 static IP addresses you're assigned in front of an ELB
- Problem with ELB
    - If ELB IP address changes, the user will still be using the old IP address instead of the new one. Global Accelerator sits in front of the ELB with only two IP addresses for the user to remember or cache



3 Top Features

- Mask Complex Architecture: Use the same IPs no matter what
- Speeds Things Up, Traffic is routed through AWS's global network infrastructure
- Weighted Pools, Test out new features with weighted groups

Things to Remember

- Edge locations help with speed
- Weighting (different countries, regions, etc)
- 2 Static IPs (will not change)

4 Questions to Ask in the Exam

- Can it be cached?
- What kind of cache should be used?
- How does the content get updated? (setting those TTLs)
- Does it add anything besides speed? (speed and security)

What to Know for the Exam

- If "in-memory database", think DynamoDB and Redis
- Redis has more features than Memcached
- Backups are only supported by Redis. Not Memcached and DAX

Chapter 20 - Governance

Managing Accounts with Organizations

What is organizations?

- Organizations is a free governance tool that allows you to create and manage multiple AWS accounts. You can control your accounts from a single location rather than jumping from account to account

Key Features

- Logging Account: Best practice to create a specific account dedicated to logging. CloudTrail supports logs aggregation
- Programmatic Creation: Easily create and story new AWS accounts
- Reserved Instances: Can be shared across all accounts
- Consolidated Billing: The primary accounts pays the bills
- Service Control Policies: SCPS can limits users' permissions

Service Control Policies (SCP)

- They are the ultimate way to restrict permissions and even apply to the root account. The ultimate enforcer
- Allow statement in IAM policy document means only the possibly to allow
- SCP only take permissions away

Things to Remember

- Scenarios about wanting to enture logs are centralized and no one can edit or delete them
    - Make sure to select Organizations to centralize and SCPs to restrict anyone from making changes to them
- *SCPs are the final say*
- Billing made easy
- Designate a single AWS account to hold your logs. *Centralize logs are always the right answer*
    - *o* Have all CloudTrail logs placed into the central location
- Sharing RIs
- *SCPs are the only way to restrict root accounts*
- *Isolating workloads into separate accounts is a great way to add more layers of security and control*

Sharing Resources with AWS RAM

What is AWS RAM?

- Resource Access Manager (RAM) is a free service that allows you to share AWS resources with other accounts and within your organization
- Easily share resources rather than having to create duplicate copies in your different accounts

What can be shared?

- Transit gateways
- VPC subnet
- License Manager
- Route 53 Resolver
- Dedicated Hosts

RAM vs VPC peering

- Sharing resources within same region, use RAM
- Sharing across regions? Use VPC peering

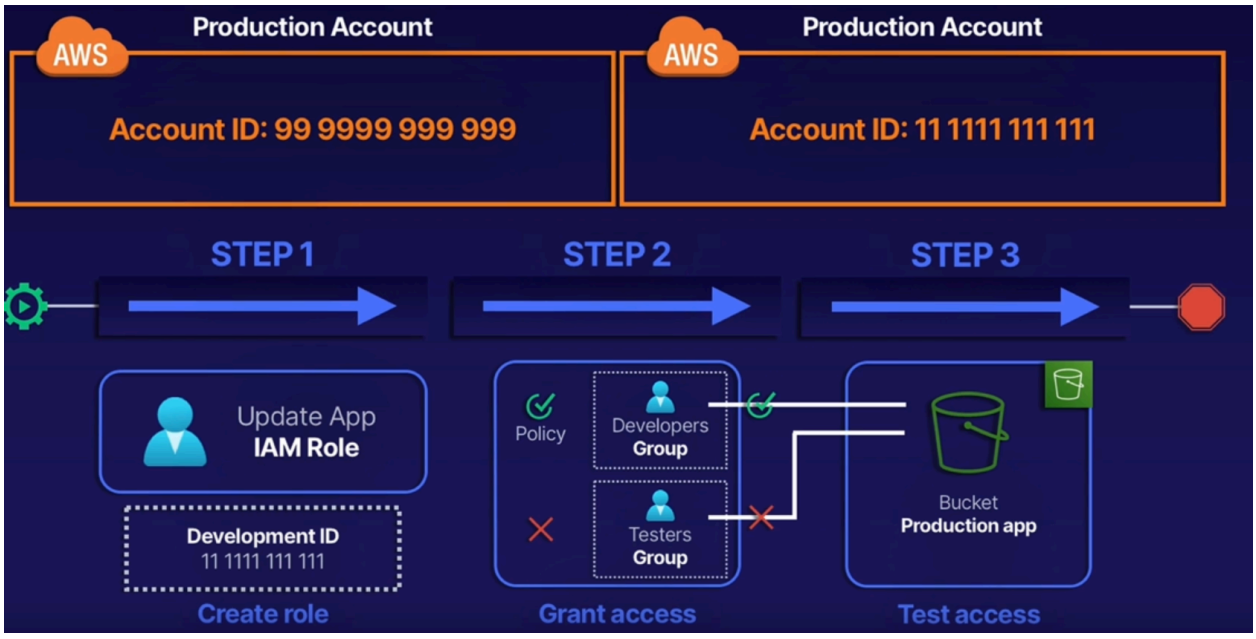
AWS RAM

- Sharing: Means you don't have to duplicate them
- VPC Peering: Excels when you connect 2 separate networks
- Organizations: RAM easily allows organizations to share architecture

Setting Up Cross-Account Role Access

Cross-Account Role Access

- *As the number of AWS accounts you manage increases, you can set up a cross-account access*
- Duplicating IAM accounts create a security vulnerability
- Gives you the ability to set up temporary access you can easily control



Things to Remember

- Anytime credentials are mention, scan the answers looking for roles
- It is preferred to create cross-account roles rather than IAM accounts
- Auditing: temporary access. No permanent credentials
- Roles are temporary. Not permanent.

Inventory Management with AWS Config

What is Config?

- Config is an inventory management and control tool
- It allows you to show the history of your infrastructure along with creating rules to make sure it conforms with best practices

State of Your Architecture: Config

- You can easily discover what architecture you have in your account. You can query by resource type, tag, and even see deleted infrastructure
- Enforce: created to flag when something is going wrong. When a rule is violated, you can be alerted or even have it automatically fixed
- Learn: Learn the history of your environment

Things to Remember

- You should be looking for Config if the question lays out any type of *standard that needs to be managed across all accounts*
- Config = Standards
- You can track deleted resources
- You can use Automation documents or Lambda function to enforce standards – *Automate the Response*
- *Analyze what changed (history of date, time, etc)*

Offloading Active Directory to Directory Service

What is Directory Service?

- AWS Directory Service is a fully managed version of AD. It allows you to offload the painful parts of keeping AD online to Aws while still giving you full control

Why Use Directory Service?

- *Managed Microsoft AD: Entire AD suite. Easily build out AD in AWS*
- *AD Connector: Creates a tunnel between AWS and your on-prem AD*
- Simple AD: Standalone directory powered by Linux Samba, AD compatible server

Authentication

- Make sure you're using AWS SSO for internal user management and Cognito for external

Things to Remember

- Managed Microsoft AD (migrate everything into Aws)
- AD Connector (used to keep users and data AD on-prem)
- Whenever possible, use Directory service over EC2 instances

Exploring with Cost Explorer

Why do we budget?

- Money

What is Cost Explorer?

- Easy-to-use tool that allows you to visualize your cloud costs. *You can generate reports based on a variety of factors, including resource tag*

Cost Explorer Features

- Service: Easily break down costs on service-by-service basis
- Time: What was your bill last month?
- Filter: Where is the spend coming from?
- Can be Predictive

Using AWS Budgets

- AWS Budgets allows organization to easily plan and set expectations around cloud costs
- You can easily track your ongoing spend

4 Types

- Cost Budget
- Usage Budget
- Reservation Budget
- Saving Plans Budget

Auditing with Trusted Advisor

What is Trusted Advisor?

- Fully managed best-practice audit tool. Will scan 5 different parts of your account and look for places where you could improve your adoption of the recommended best practices by AWS
    - It's free but you'll need a Business or Enterprise Support plan
    - Can solve the problem for you

5 Questions Trusted Advisor Asks

- Cost Optimization
- Performance
- Security
- Fault Tolerance
- Service Limits

Things to Remember

- *Automate a Response, use Lambda or CloudWatch Events*
- *Use SNS to let users know*
- Business or Enterprise support plan
- *Will not fix the problems for you (Lambda, etc)*
    - Automate using EventBridge (CloudWatch Events) to kick off Lambda to solve the problem for you

4 Questions to Ask in the Exam

- Can it be centralized?
- How do we standardize?
- How do we enforce the standards?
- Are the users internal or external?

Chapter 21 – Migrating Data with AWS Snow Family

Migrating Data with AWS Snow Family

How do we Migrate Data?

- Internet: Using your existing connection, but very slow and could be a security risk
- Direct Connect: Faster, more secure. But not practical.
- Physical: What if you bundled your data and physical delivered it to AWS?

What is the Snow Family?

- A sack of hard drives AWS sends to you

Snowcone

- Smallest device in the snow family
    - 8 TB of storage, 4 GB of memory, and 2 vCPUs
    - Easily migrate data to AWS after you've processed it
    - IoT sensor integration
    - Perfect for edge computing where power and space are constrained

Snowball Edge

- Bigger device
    - 48 TB to 81 TB in storage
    - Storage, Compute, and GPU flavors
    - Varying amount of CPU and RAM
    - Perfect for one-time migration

Snowmobile

- Biggest device
    - 100 PB of storage
    - Exabyte-scale data center migration
    - You won't need to know this

You won't need to know specifics

Things to Remember

- Make sure to pay attention to restrictions in situation
    - Slow Internet, security, having large amount of data
- You can ship data back to you

Storage Gateway

What is Storage Gateway?

- *Hybrid cloud storage* service that helps you merge on-prem resources with the cloud
- It can help with one-time migration or a long-term paring of your architecture with AWS for backups

Types of Gateways

File Gateway

- Caching of local files (file share)
- NFS or SMB mount
- Keep local copy of recently used files
- Exam scenario: Don't have enough on-prem storage space, set up a file gateway
- Helps with migrations to AWS
- Simply VMs you run in on-prem environment

Volume Gateway

- iSCSI mount
- Cached or stored mode
- Create EBS snapshots
- Easy way to migrate or backup all Volumes

Tape Gateway

- Replace physical backup tapes into type gateway
- Tricks back trip gateway into thinking it's backing up to a physical tape gateway
    - Won't cover it as much as the other two in the exam

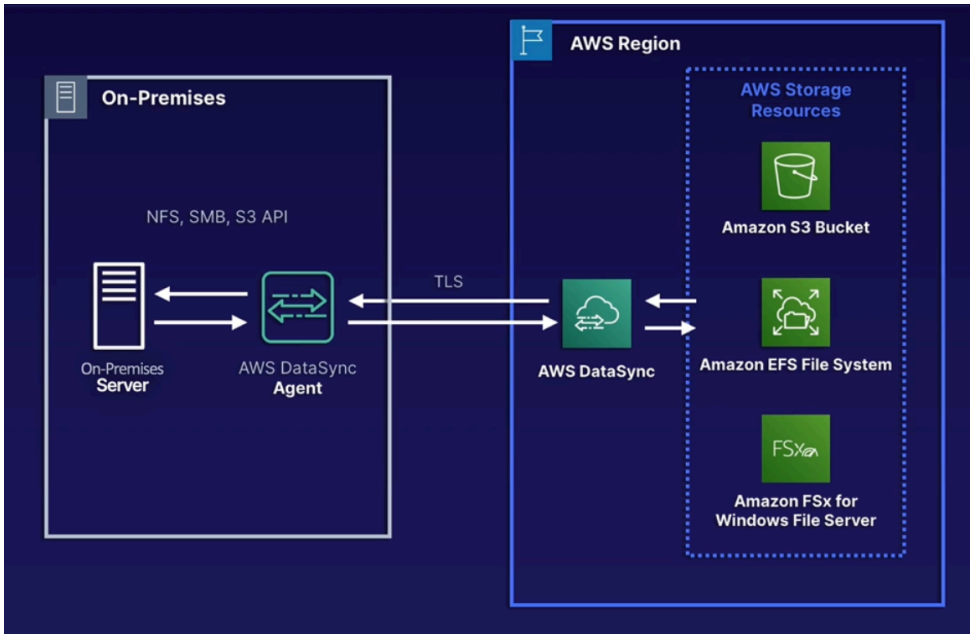
AWS DataSync

What is DataSync?

- *Agent-based solution* for migrating on-prem storage to AWS.
- Easily move data between NFSA and SMB shares and AWS storage solutions
- *Good for one-time migration*

Using DataSync

- Setup datasync agent on the on-prem server. Then setup how DataSync can send the data. *Supported locations include S3, EFS, and FSx*
    - All data is encrypted



DataSync Service communications set up with On-prem server with storage

Storage Gateway vs DataSync

- DataSync is good for a one-time migration
- Storage Gateway is great for a continuous sync

AWS Transfer Family

What is the AWS Transfer Family?

- Allows you to easily move files in out of S3 or EFS using SFTP, FTPS, or FTP
    - FTP is supported within the VPC and not over the public Internet
- *Excels with older or legacy applications*

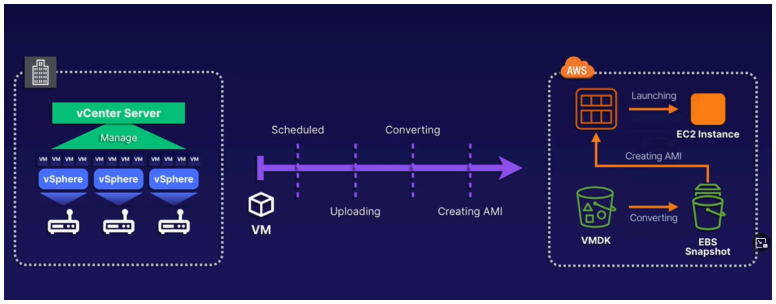
How Does It Transfer?

- "Tricks" the endpoint on the receiving end (endpoint where data is being uploaded by users) into storing for Amazon S3 and Amazon EFS.
    - Swaps this through the DNS entry

Moving to the Cloud with Migration Hub

What is AWS Migration Hub?

- Gives you a single place to track the process of your application migration to AWS. It integrates with Server Migration Service (SMS) and Database (DMS)
    - Organizational tool, used to

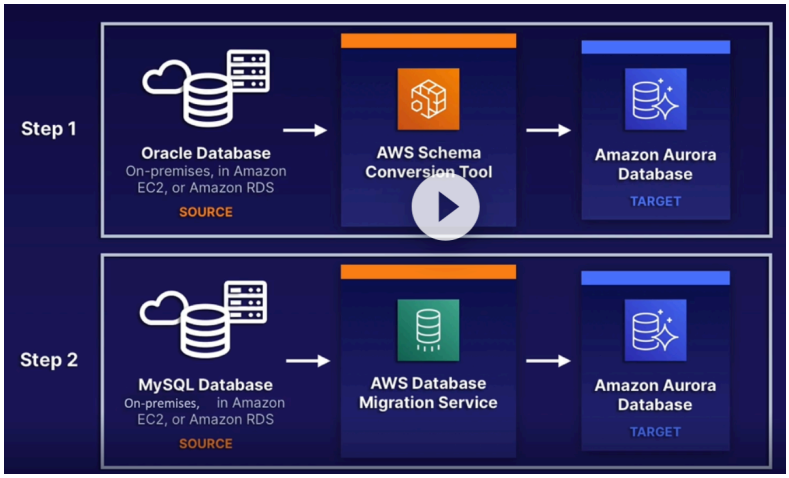


Server Migration Service (SMS)

- Take your VMware architecture into an AMI (AWS compliant)

Database Migration Service (DMS)

- Used to take an on-prem DB (typically Oracle and MySQL) and convert into Amazon DBs (Aurora)
    - Can be used to combine multiple DBs int one
- Uses the AWS Schema Conversion Tool and Database Migration Service to accomplish transport of data



Things to Remember

- DMS is going to be used to move databases
- SMS will be the tool of choice for moving servers
- Migration Hub provides an overview you need
- Focus on answers that include a complete migration

Migration Exam Tips

4 Questions

1. Where are we going?
2. How do we get there?
3. Is it all at once?
4. Is it a partial migration?

Snow Family

- Snowball are often on how much data you're moving. Snowballs are perfect for moving terabytes of data.
    - Snowcone and snowmobile Mostly used as a distractor
- Snowball is best used in situations where you have slow Internet and require security

3 Exam Tips for Storage Gateway

- Good for Hybrid Architecture
- Out of Space? (keep data backed-up)
- It's a VM on-prem

DataSync

- Agent-based solution
- Best for one-time migrations and file shares
- EFS and FSx are both viable location for DataSync

Transfer Family

- Best for legacy file transfers
- Give ability to read and write from S3 and EFS

Migration Hub

- Organizational tool. Organize and track migration steps
- DMS, go-to tool for DB migration. Good for on-prem to cloud, or RDS to other DBs
- SMS is the best tool for VMs and servers