(2022) AWS Solutions Architect Associate

Sunday, July 18, 2021 3:15 PM

- High Performance Computing
  - Importing data to AWS
    - Snowball, Snowmobile
    - AWS DataSync
    - Direct Connect
  - Compute and Network
    - EC2 (GPU/CPU optimized)
    - EC2 Fleets
    - Placement Groups (cluster placement groups)
    - Enhanced Networking
    - Elastic Network Adapters
    - Elastic Fabric Adapters
  - Storage
    - Instance-attached storage
      - EBS -> up to 64,000 IOPS
      - Instance store
    - Network attached storage
      - S3
      - EFS
      - FSx for Lustre
  - Orchestration and Automation
    - AWS Batch
    - AWS ParallelCluster

- Key Terms
  - Local Zone
    - Complements an existing AWS region by placing compute/storage/network close to population centers
  - Edge Location
    - Used by Amazon CloudFront for content delivery and Amazon Route 53 for DNS geo-load balancing
  - Regional Edge Cache

        - Used by Amazon CloudFront by default to provide a larger cache which allows for longer storage after cached content has expired from standard edge locations
    - ARN
        - arn:PARTITION:SERVICE:REGION:ACCOUNT_ID:RESOURCE_NAME

- AWS Support Plans
    - Developer
        - Support is provided by email during business hours
        - General Guidance < 24 hours
        - System Impaired < 12 hours

    - Business
        - Support is provided 24/7 by phone, email, and chat
        - General Guidance < 24 hours
        - System Impaired < 12 hours
        - Production systems impaired < 4 hours
        - Production system down < 1 hour

    - Enterprise On-Ramp
        - Support is provided 24/7 by phone, email, and chat
        - General Guidance < 24 hours
        - System Impaired < 12 hours
        - Production systems impaired < 4 hours
        - Production system down < 1 hour
        - Business-critical system down < 30 minutes

    - Enterprise
        - Support is provided 24/7 by phone, email, and chat
        - General Guidance < 24 hours
        - System Impaired < 12 hours
        - Production systems impaired < 4 hours
        - Production system down < 1 hour
        - Business-critical system down < 15 minutes

- AWS Organizations
    - General
        - Account management service enabling you to organize multiple AWS

accounts for centralized management
        - Accounts are organized under Ous
        - Use Service Control Policies (SCPs) for authorization, AI Opt opt-out policies, Backup Policies, and Tag Policies
    - Consolidated Billing
        - Consolidate billing for member AWS accounts to a central management AWS account
        - Share volume pricing, reservations, and savings plans
    - Best Practices
        - When using consolidated billing, do not deploy resources into the payment account

- AWS Cloud Shell
    - Browser-based shell which supports bash, PowerShell, and Z Shell
    - 1GB/user of persistent storage in $HOME within each region

- AWS Budget
    - Create for an AWS account with a fixed or monthly budget
    - Set alerts based on thresholds which can then send via SNS, ChatBot, and email

- AWS Savings Plan
    - Less overhead than Ris since RI requires coordination of RI purchase and exchanges
    - Commit to use an amount of compute (dollars per hour) over 1 or 3 year span
    - Usage with savings plan charged at cheaper rate and anything over gets charged at on-demand
    - Types
        - Compute Plan
            - Apply to any region, instance family, OS, tenancy including EMR, ECS, and EKS
        - EC2 Instance Plan
            - Apply to specific instance family within a region and allows for switching OS

- AWS Capacity Reservation
    - Ensure you always have access to EC2 capacity when you need it
    - Separate from discount like RI or Savings Plan

    - Set for a specific AZ, number of instances, instance type, tenancy, platform, and OS

- Key Ports
    - Postgress / Aurora - 5432
    - MySQL / MariaDB - 3306
    - Oracle - 1521
    - Redis - 6379

- AWS Resource Access Manager (RAM)
    - Feature which allows you to share AWS resources between accounts
    - Resources are included in a share which is then shared with one or more accounts
    - Invitation/Acceptance pattern

- Amazon Event Bridge
    - General
        - Allows events to be received from CloudWatch Events as well as 3rd party events and custom events
        - Will replace CloudWatch Events

- AWS CloudWatch Alarms
    - General
        - Alarms for high resolution custom metrics can be triggered at 10 sec, 30 sec, and 60 sec

- AWS CloudTrail
    - General
        - Enabled by default
        - Trails can send logs to S3 or CloudWatch Logs
        - Trails can be by region or all regions
        - CloudTrail Insights analyzes CloudTrail log entries for suspicious behavior
        - 90 days stored in Event History

- Amazon CloudWatch Logs
    - General
        - Components - Group, stream
        - Expiration can be set

- Destinations
  - S3
  - Kinesis Data Streams
  - Kinesis Data Firehose
  - Lambda
  - ElasticSearch

- Amazon CloudWatch Metrics
  - General
    - Built-in metrics are delivered every 5 minutes by default but can be delivered at 1 minute by using detailed monitoring
    - CPU, NetworkIn/Out, Disk Read/Write

  - Custom Metrics
    - Metrics accepts data points two weeks into the past and 2 hours into the future
    - Standard custom metrics are received every minute but can be higher resolution 1/5/10/30 seconds
- **EC2 General**
  - EC2 General Facts
    - By default, the root EBS volume is deleted when an instance is terminated
    - By default, termination protection is off but you can turn it on in the instance settings
    - Both root and additional volumes for an EC2 instance can be encrypted (this includes default AMIs)

  - EC2 Health Checks
    - Performed every minute
    - System health checks validate the health of the physical host and failures can be dealt with by stopping and starting the instance to move it to a new host
    - Instance health checks validate the health of the instance itself and failures can be dealt with by restarting the instance

  - EC2 User Data
    - On Linux EC2 instances shell or cloud-init and Windows PowerShell or command prompt commands that are run when an instance is first started to bootstrap it

started to bootstrap it
    - Data is base64 encoded prior to be submitted to the API
    - There is a new feature where an EC2 instance can be customized to run the user data at each boot

- EC2 Launch Template Parameters
    - Instance type
    - AMI ID
    - Key pair
    - Security groups
    - Tenancy (dedicated or default)
    - Subnet

- EC2 Launch Template
    - Launch templates support versioning which can be configured with different parameters and permissions while launch configurations do not
    - Launch templates are used with on-demand instances, spot instances, autoscale groups, and fleets

- EC2 Instance Classes Codes
    - T, Mac, M -> General Purpose
    - A -> ARM-based
    - C -> Compute optimized
    - R, X, High Memory, Z -> Memory optimized
    - P, G, F -> CPU optimized
    - I, D, H -> Storage optimized

- EC2 Instance Classes
    - General Purpose
        - Web servers, code repos
    - Compute Optimized
        - Batch processing, media transcoding, high performance web servers, HPC, modeling and ML
    - Memory Optimized
        - High performance relational/non-relational databases
        - Distributed web scale caches
        - In-memory databases optimized for business intelligence
        - Real time processing of big unstructed data
    - Storage Optimized

        - Storage Optimized
            - OLTP
            - Relational and NO SQL DBs
            - Redis
            - Data warehouses
            - Distributed file systems

- EC2 Launch Types
    - On-Demand
        - Linux instances are billed per second after first minute while other OS such as Windows as billed per hour
    - Standard RI
        - Up to 75% discount to on-demand
        - Reserve 1 or 3 years
        - Pay in full immediately, pay partial up front and rest monthly, or pay monthly
        - Aligned to a specific instance type
    - Convertible RI
        - Allow for changing instance type
        - Provide a smaller discount from standard Ris (up to 54%)
    - Scheduled RI
        - Reserve an instance for a discount for a specific time window
    - Spot instances
        - Provide up to 90% discount
        - Instances can be terminated anytime after your max price is exceeded
        - Good for batch jobs, image processing, data analysis, distributed workloads, and workloads with a flexible start and end time
    - Dedicated Hosts
        - Run on dedicated physical hardware
        - 3 year commitment
        - Good for leveraging existing per CPU licensed software or for heavy compliance workloads
    - Dedicated Instances
        - Run on dedicated physical hardware but you don't get insight into sockets, cores, and host id
        - Good for compliance workloads but not CPU licensed software because of the lack of visibilty into the physical host

- EC2 Placement Groups
  - Three types include cluster, spread, and partition
  - Used to influence placement of interdependent EC2 instances
  - An instance can only belong to one placement group

- EC2 Cluster Placement Group
  - EC2 instances are placed close to each other within an AZ to maximize throughput and minimize latency
  - Up to 10Gbps single flow traffic

- EC2 Spread Placement Group
  - Ensures machines are deployed to different racks
  - Can span AZ
  - Maximum of 7 instances per AZ

- EC2 Partition Placement Groups
  - Ensures machines within the same partition do not share the same rack
  - Can span AZ
  - Maximum of 7 partitions per AZ
  - Use case is HDFS, Cassandra, HBase

- EC2 Hibernate
  - RAM saved to EBS root volume
  - No charge while instance is hibernated
  - RAM must be less than 150GB and cannot be hibernated for more than 60 days-
  - Root volume must be encrypted

- EC2 Nitro
  - AWS new hypervisor
  - Supports higher speed EBS volumes which can be up to 64,000 IOPS (pre-Nitro is limited to 32,000 IOPs)
  - Generation 5 and above

- EC2 vCPU
  - Each vCPU represents a CPU thread on the host
  - vCPUs can be adjusted when launching the EC2 instance to reduce CPUs (for licensing) or reduce threads (better performance for HPC)

- EC2 Instance Profile
  - Container for an IAM Role
  - Each instance profile can contain only one role
  - Instance profiles are created automatically when creating an IAM Role if using the console, otherwise they must be created separately

- EC2 Attach/Replace IAM Role
  - EC2 instances without an existing attached IAM Role can have the role attached in both the stopped and running state
  - EC2 instances with an existing attached IAM Role must be in the stopped state to replace the role
  - Attaching

- **AMI (Amazon Machine Image)**
  - AMI General
    - Regional
    - Can be copied across regions
    - Can be shared publicly (if not encrypted) or with specific AWS accounts (if not encrypted or encrypted with CMK)
    - Public, private, or marketplace

  - Create an AMI
    - Start an EC2 instance and customize
    - Stop the EC2 instance
    - Create an image (AMI) from the EC2 instance
    - Launch a new EC2 instance from the created AMI

- **EC2 Spot Instances**
  - EC2 Spot Instances General Facts
    - Spot instances are interrupted due to the price going above your maximum price, AWS needs to reclaim capacity, or a constraint such as launch group or AZ
    - You receive a two minute warning before an interruption occurs
    - If AWS interrupts you are not charged for the partial hour

- EC2 Spot Instances Request Parameters
  - Number of instances
  - Instance types
  - Availability zones
  - Maximum price
  - Start time and expiration time (optional)
  - Launch specifications / launch template
  - Request type which is either maintain or request

- EC2 Spot Instance Request Types - One-Time vs Persistent
  - One-time request remains active until you cancel it, it expires, or AWS fulfills it
  - Persistent request remains until you cancel it or it expires
  - Persistent request allows for hibernation and stopping spot instance in addition to termination (which is default)

- **<u>Spot and EC2 Fleets</u>**
  - Spot Capacity Pool
    - Set of unused EC2 instances of the same instance type and operating system in the same availability zone
    - Used by spot fleets and EC2 fleets

  - Spot and EC2 Fleets - General Facts
    - Use fleets to launch EC2 instances of different instance types across multiple availability zones within a given region
    - Use a combination of on-demand, spot, and reserved instances
    - Create a request with the total capacity (on-demand vs spot instances), launch specifications, and maximum price (for spot instances)
    - Fleets will launch instances until the desired capacity is reached or the maximum price level is hit
    - Optionally instead of number of instances you can specify number of vCPUs or memory

  - Spot Fleet Allocation Strategies
    - lowestPrice (default)
    - Diversification
    - CapacityOptimized

- Spot Fleet Capacity Optimized Strategy
  - Instances are pulled from the pools with the most capacity reducing the risk of interruption
  - Allows for prioritization of specific instance types which are "best effort"
  - Ideal for big data, HPC, ML, media transcoding

- Spot Fleet Lowest Price Strategy
  - Instances are pulled from the pools that will result in the lowest price based upon the fleet requirements
  - The default option
  - Option for InstancePoolsToUseCount to specify how many pools to use to add diversification to the request mitigating the risk of interruptions

- Spot Fleet On-Demand Instance Strategies
  - Lowest Price (Default)
  - Prioritization

- Spot Fleet Capacity Re-Balancing
  - Fleet will automatically spin up instances in other pools when it receives a rebalancing notification that specific instances at at risk for interruption
  - Available for maintain request type only

- EC2 Fleet Request Types
  - Instant
  - Request
  - Maintain

- EC2 Fleet vs Spot Fleet
  - EC2 Fleets offer the additional request type of instant
  - EC2 Fleets can contain only on-demand instances, only spot instances, or both
  - EC2 Fleets support capacity reservations

- **<u>EC2 Autoscaling Groups (ASG)</u>**
  - Autoscaling Group (ASG) General
    - Automatically scale in or out for EC2 instances

        - Optional registration with ELB
        - Works across AZ
        - Setting min/max with desired capacity and no other conditions will keep it at desired capacity
        - Health checks can be simple EC2 health checks, ELB health checks, or custom
        - Instances are removed by AZ with most instances and instance launched with oldest launch template

    - ASG Launch Template
        - Launch templates allow for ASGs to contain both spot and on-demand instances
        - Modifying the launch template assigned to an ASG affects only new instances not existing instances

    - ASG Rebalancing
        - Rebalancing occurs based upon rebalancing to distribute across AZ or capacity rebalancing (for spot instances in danger of interruption)
        - Allows AWS to exceed max instances by 1 instance or 10%, whichever is greater

    - ASG Attributes
        - Launch configuration / Launch Template
        - VPC and AZ/Subnets
        - No load balancer, attach to new/existing load balancer
        - Health check type and grace period
        - Minimum, maximum, desired capacity
        - Scaling policies
        - Optional scale-in protection

    - ASG Scaling Options
        - Maintain specific number
        - Manually scale
        - Scale on schedule
        - Scale on Demand
        - Predictive scaling

    - ASG Scaling Policies

        - Dynamic scaling
        - Predictive scaling
        - Scheduled scaling

    - ASG Dynamic Scaling Policies
        - Includes Simple, Step, and Target Tracking
        - Scales on capacity units or % of the ASG
        - Scaling cooldown only applies to Simple Scaling Policies while Step and Target Tracking Policies have a warmup counter

    - ASG Dynamic Scaling Policies - Simple
        - Scales based upon CloudWatch Alarm going into the alarm state based upon monitoring of a metric and will not scale again until the cooldown and health checks (if applicable) are completed
        - Create CloudWatch Alarm separately from the scaling policy
        - Has a scaling cooldown which is by default 300 seconds

    - ASG Dynamic Scaling Policies - Step
        - Scales based upon the specific range of values of a metric being monitored by a CloudWatch Alarm and there can be multiple levels of scale depending on the threshold of the metric
        - Create the CloudWatch Alarm separately from the scaling policy
        - A warmup metric can be included (300 seconds by default) before the instance is counted

    - ASG Dynamic Scaling Policies - Target Tracking
        - Set a minimum and maximum threshold and the AWS will try to keep as many instances as needed to align with this metric
        - CloudWatch Alarms are created automatically by the ASG
        - A warmup metric can be included (300 seconds by default) before the instance is counted

    - ASG Lifecycle Hooks
        - Perform actions during launch or termination
        - Autoscaling:EC2_INSTANCE_LAUNCHING and autoscale:EC2 _INSTANCE_TERMINATING

    - ASG Troubleshooting
        - Detaching an instance from an ASG removes it from the ASG and allows

        - Detaching an instance from an ASG removes it from the ASG and allows you to further troubleshoot
        - Suspending an instance keeps it in the ASG but allows you to troubleshoot it before returning to the ASG
- **S3 (Simple Storage Service)**
    - General
        - Concepts: Bucket, Prefix, Object
        - Bucket name is globally unique
        - BUCKET_NAME.s3.amazonaws.com
        - Supports REST API and SOAP API (deprecated)
        - Read-after-write consistency for new objects and eventual consistency for updates
        - Multi-part uploads should be used for objects >100MB
        - Max size object is 5TB

    - S3 Object Tags
        - Key/value pairs applied to an S3 object
        - 10 maximum tags per object

    - Optimizing S3 Performance
        - Partitions are created based upon the prefix used for an object
        - If bucket is receiving more than 300 GET or 100 PUT/LIST/DELETE then partitioning should be done to improve performance
        - Each prefix can achieve 3,500 PUT/COPY/POST/DELETE and 5,500 GET/HEAD

    - Optimizing S3 Upload/Download
        - For uploads recommended to use multi-part upload for files >100MB and required for files >5GB
        - For downloads use S3 Byte Range Fetches to parallelize downloads or to grab just the header of a file

    - S3 Select / Glacier Select
        - Download portion of an object (such as CSV) using SQL queries on either data stored in S3 or S3 Glacier

    - S3 Static Website Hosting
        - Endpoint of http://BUCKET_NAME/UNIQUE_REGION_NAME.amazonaws.com

http://BUCKET_NAME/UNIQUE_REGION_NAME.amazonaws.com
    - Enabled as property of the bucket and requires the bucket have anonymous access
    - Server-side scripting IS NOT supported

- S3 Object Encryption
    - Encryption performed on individual object or entire bucket
    - S3 SSE-S3
        - Objects encrypted by a data encryption key which is generated per object and is wrapped by a key managed by AWS
    - S3 SSE-C
        - Objects encrypted by a data encryption key you provide when you upload or download
    - S3 SSE-KMS
        - Objects encrypted by a data encryption key which is generated per object and is wrapped by a CMK stored in KMS

- S3 Object Lock
    - WORM for S3
    - Two modes: Compliance Mode and Governance Mode
    - Compliance mode locks the object from being modified by anyone (including root) for a time period
    - Governance mode locks the object from being modified but those with special permissions can modify
    - Legal hold locks the object indefinitely and requires s3:PutObjectLegalHold permission to unlock it
    - Applied to individual objects or applied across the entire bucket

- S3 Access Control
    - IAM Policy
        - Use when providing access to an IAM User, Group, or Role
    - Bucket Policy
        - Use when providing access to other AWS accounts
    - ACL
        - One ACL per bucket and each object

- S3 Transfer Acceleration

- S3 Transfer Acceleration
  - Utilizes AWS edge locations to accelerate upload speeds
  - User uploads to an edge location which is then synchronized to the AWS data center
  - Utilizes a special link

- S3 Storage Classes
  - S3 Standard
    - Designed for 11x9 durability and 99.99% availability with an SLA of 99.99%
    - Synchronously replicated across multiple availability zones within a region

  - S3 Standard IA
    - Designed for 11x9 durability and 99.9% availability with an SLA of 99.9%
    - Minimum storage duration of 30 days

  - S3 One Zone IA
    - Designed for 11x9 durability and 99.5% availability with an SLA of 99%
    - Stored in a single availability zone

  - S3 Intelligent Tiering
    - Designed for 11x9 durability and 99.9% availability with an SLA of 99.9%
    - Minimum storage duration of 30 days
    - Automatically moves data between storage classes

  - S3 Glacier Flexible Retrieval
    - Use case is for backups or disaster recovery
    - Objects restored are written to Standard IA
    - Minimum storage of 90 days and object size of 128KB
    - Designed for 11x9 durability and 99.99% availability with an SLA of 99.9%
    - Restore Modes
      - Standard -> 3-5 hours
      - Expedited -> 1-5 minutes
      - Bulk -> 5-12 hours

    - S3 Glacier Instant Retrieval
        - Use case is for rarely accessed data that needs milliseconds retrieval
        - Objects restored are written to Standard IA
        - Minimum storage of 90 days and object size of 128KB
        - Designed for 11x9 durability, 99.9% availability, and 99% SLA

    - S3 Glacier Deep Archive
        - Use case is for data accessed once or twice a year
        - Objects restored are written to Standard One Zone IA
        - Minimum storage of 180 days and object size of 40KB
        - Restore Modes
            - Standard -> 12 hours
            - Bulk -> 48 hours

- S3 Glacier Terminology
    - Endpoint -> http://REGION_ENDPOINT/ACCOUNT_ID/vaults/VAULT_NAME/archives/ARCHIVE_ID
    - Archive -> objects stored in Glacier
    - Vault -> similar to bucket but contains archives
    - Inventory -> cold index of a vault updated every 24 hours
    - Job -> requires to restore an archive

- S3 Versioning
    - Enabled on the bucket level and cannot be disabled only suspended
    - Each object gets a version ID assigned to it which distinguishes it from other objects with the same key
    - Each version of an object has a separate ACL such that making one version of an object public won't make new versions public
    - Delete markers are inserted when an object is deleted
    - Required for MFA Delete
    - Required for CRR/SRR
    - Required for object lock

- S3 Permanently Delete a Versioned Object
    - Specify the version ID when issuing the DELETE command

- S3 Restore an Object
  - Remove the delete marker for the object

- S3 Signed URL
  - Used to grant one-time access to a single object in S3 with a specific start and expiration time
  - Created by an authenticated user with access to the S3 object

- S3 MFA Delete
  - Enabled on a bucket by the bucket owner
  - Requires MFA for modifying versions or deleting and object version
  - Requires the bucket have versioning enabled

- S3 Lifecycle Management
  - Migrate between storage classes or expire the object
  - Created a bucket level and apply to all objects in a bucket or a subset of objects by a specific prefix or object tags
  - Can be used in combination with versioning to apply to current and non-current versionsq

- S3 Cross-Region Replication (CRR) and Same Region Replication (SRR)
  - Asynchronous replication of S3 data between storage classes, availability zones, regions, or accounts
  - Requires versioning be enabled on both source and destination buckets
  - Only objects modified AFTER CRR/SRR is enabled are replicated
  - Delete markers ARE NOT replicated by default but can be enabled for replication
  - Prior versions of objects ARE NOT replicated

- S3 Replication Control
  - SLA-backed replication time for CRR/SRR
  - 99.9% of objects replicate in less than 15 minutes SLA but designed for 99.99%

- S3 Static Web Hosting
  - Endpoint of http://BUCKET_NAME.s3-websites.REGION.amazonaws.com
  - Content must be publicly readable using ACL or bucket policy

- **EBS (Elastic Block Storage)**
  - Instance Store vs EBS Store
    - Instance store storage is directly attached to the host machine and provides better performance but all data is lost if the instance is stopped or terminated
    - EBS store is attached over the network and persists after the instance is stopped or terminated (unless configured to be deleted when the instance terminates)

  - EBS General
    - By default EBS root volume is deleted when instance is terminated but data volumes are not
    - EBS volumes are fixed to an availability zone
    - EBS volumes can be attached to a single EC2 instance until using Multi-Attach EBS volumes which are Provisioned IOPS only
    - General, Provisioned, Throughput Optimized, Cold HDD, Magnetic
    - EBS volumes can have the size and storage class changed on the fly

  - EBS Multi-Attach Volumes
    - Volume can be attached to one or more instances at the same time in the same AZ granting read/write access to the volume
    - Supported on Provisioned IOPs only volumes
    - Use case is for clustered filesystem

  - EBS General Purpose
    - SSD
    - Gp2, gp3
    - 1GiB - 16TiB
    - 16,000 IOPS
    - 1,000 MiB throughput
    - Can be a boot volume
    - Use case is applications requiring low latency, VDI, and Dev+Test

  - EBS General Purpose GP2 vs GP3
    - Gp3 has a baseline of 3,000 IOPS and throughput of 125MB/s and these can be set INDEPENDENTLY
    - Maximum of 16,000 IOPS and throughput of 1,000MB/s
    - Gp2 can BURST IOPS up to 3,000

    - Gp2 size of the volume and IOPS are linked

- EBS Provisioned IOPS
    - SSD
    - Io1, io2
    - 4GiB - 16TiB
    - 64,000 IOPS (Nitro and non-Nitro is 32,000 IOPS)
    - 1,000 MiB throughput
    - Can be a boot volume
    - Use Case: Databases

- EBS Provisioned IOPS io1 vs io2
    - io2 has better durability and higher IOPs than io1
    - Both options allow you to provision the IOPS independently of the storage size
    - io2 supports Block Express which allows for 4GB - 64TB, sub-millisecond latency and IOPS of 256,000

- EBS Throughput Optimized
    - HDD
    - St1
    - 125GiB - 16TiB
    - 500 IOPS
    - 500 MiB throughput
    - Use Cases - Datawarehouse, log processing, big data

- EBS Cold HDD (sc1)
    - HDD
    - 125GiB - 16TiB
    - 250 IOPS
    - 250 MiB throughput

- Magnetic (standard)
    - HDD
    - 1GiB - 1TiB
    - 200 IOPS
    - 90 MiB throughput
    - Can be a boot volume

- EBS Snapshots
    - Recommended to detach volume before creating snapshot
    - Can be copied across regions or restored to a different availability zone
    - Snapshots are incremental with first snapshot being largest and additional snapshots being sized only for incremental changes
    - Can be shared publicly if not encrypted
    - Can be shared with other AWS accounts in encrypted format if it is encrypted using a CMK and the target account has appropriate permissions over the key
    - Multi-Volume snapshots create a crash-consistent snapshot of all volumes attached to an instance

- EBS Volume Encryption
    - Uses AES256 symmetric encryption
    - Snapshots created from encrypted volume are encrypted by default
    - Volumes created from encrypted snapshots are encrypted by default

- Encrypt Existing Unencrypted EBS Volumes
    - Option 1
        - Create snapshot of unencrypted volume which will be unencrypted
        - Create copy of snapshot, select to encrypt, resulting in encrypted snapshot
        - Create new volume from encrypted snapshot resulting in encrypted volume
        - Attach encrypted volume to existing EC2 instance
    - Option 2
        - Create snapshot of unencrypted volume which will be unencrypted
        - Create a new volume from the unencrypted snapshot, select to encrypt, resulting in encrypted volume
        - Attach encrypted volume to existing EC2 instance

- **<u>EFS (Elastic File Storage)</u>**
    - EFS General
        - Distributed file share available across AZ
        - NFS 4.1, POSIX
        - Security Group associated to resource is used to control network access

        - Encryption is available with KMS
        - Scales automatically
        - Offers lifecycle management options with standard or infrequently accessed
        - Offers intelligent tiering

    - EFS Storage Classes
        - EFS Standard
        - EFS Standard IA
        - EFS One Zone
        - EFS One Zone IA

    - EFS Performance Modes
        - Set at creation time
        - General Purpose -> latency sensitive use cases
        - Max I/O -> higher latency but better throughput

    - EFS Throughput Mode
        - Bursting -> throughput grows with share size
        - Provisioned -> throughput is set and independent of the size of the share

- **<u>Amazon FSx for Lustre</u>**
    - General
        - Used when require high-speed and high-capacity distributed storage
        - Use cases are HPC and financial modeling
        - Capable of storing data directly in S3

- **<u>AMI (Amazon Machine Image)</u>**
    - AMI General
        - Regional
        - Can be copied across regions
        - Can be shared publicly (if not encrypted) or with specific AWS accounts (if not encrypted or encrypted with CMK)
        - Public, private, or marketplace

    - Create an AMI
        - Start an EC2 instance and customize
        - Stop the EC2 instance

    - Create an image (AMI) from the EC2 instance
    - Launch a new EC2 instance from the created AMI

- **AWS Storage Gateway**
    - AWS Storage Gateway Types
        - File Gateway
        - FSx Gateway
        - Volume Gateway
        - Tape Gateway

    - AWS File Gateway
        - Data stored in S3
        - Supports access from on-premises devices using SMB and NFS
        - Maximum file size of 5TB

    - AWS FSx Gateway
        - Data stored in AWS FSx for Windows Servers
        - Supports access from on-premises devices using SMB

    - AWS Volume Gateway
        - Data stored in S3 as EBS snapshots
        - Supports access over iSCSI
        - Max volume size of 32TB
        - Modes
            - Cached
            - Stored

    - AWS Tape Gateway
        - Data stored in S3, S3 Glacier, and S3 Glacier Deep Archive
        - Supports access over iSCSI VTL

- **AWS Snowball**
    - AWS Snowball Edge
        - Replaces legacy AWS Snowball devices
        - Offers compute, network, and storage to run EC2 or Lambdas
        - Two flavors of storage optimized and compute optimized
        - 80TB allowable storage

- Provides an S3 Endpoint

- **AWS Migration Hub**
  - Central location to review discovered data about servers/applications and perform and monitor migrations
  - Integrates with AWS SMS, DMS, and MGN

- **AWS Application Migration Service (MGN)**
  - Supports both agent and agentless (VMWare-only) migrations
  - Agent supports continuous data protection (CDP) and is preferred over agentless

- **AWS Server Migration Service (SMS)**
  - Legacy service used to perform migration of on-premises servers to AWS
  - Discontinued as of 3/1/2022

- **AWS Application Discovery Service**
  - Agent or agentless discovery tool that discovers servers, gathers performance data, information on processes, and TCP connections

- **AWS Data Migration Services (DMS)**
  - AWS DMS General
    - Migrate between on-premises and AWS and back
    - Supports heterogenous and homogenous migrations
    - One-time migration or on-going sync
    - AWS Schema Conversation Tool is used for heterogenous migrations
    - Components are source and destination endpoint where one endpoint must be in AWS

  - AWS DMS Sources
    - On-Prem - Oracle, MS SQL, MySQL, MariaDB, PostgreSQL, MongoDB, SAP, IBM DB2
    - AWS - RDS, S3, DocumentDB
    - Azure - Azure SQL

  - AWS DMS Targets
    - On-Prem - Oracle SQL, MySQL, MariaDB, PostgreSQL, SAP, Redis
    - AWS - RDS, RedShift, DynamoDB, S3, ElastiCache for Redis, DocumentDB, Neptune

- **VM Import/Export**
  - Import/Export from/to EC2/AMI from/to VM

- **AWS Data Sync**
  - AWS Data Sync General Facts
    - Agent-based tool that is installed on a machine
    - Replication can occur hourly, daily, or weekly
    - Use cases include data migration from on-premises, archiving old data to Glacier, backing up data to AWS, or moving data into the cloud for cloud-processing

  - AWS Data Sync - Supported Origins and Destinations
    - NFS file servers
    - SMB file servers
    - Hadoop (HDFS) file servers
    - Amazon S3
    - Amazon EFS
    - Amazon FSx for Windows Server
    - Amazon FSx for Lustre
    - On-premises self-managed object storage
    - AWS Snowcone/Snowball/Snowmobile

- **Amazon SQS**
  - General
    - Message queue service
    - 256KB max message size unless storing in S3 which allows up to 2GB
    - Default retention of 4 days with a maximum of 14 days

  - Queue Types
    - Standard (Default)
      - Guarantees messages are delivered at least once but messages may be out of order
      - High throughput
    - FIFO
      - Guarantees messages are delivered once and in order
      - Multiple message groups are supported in a single queue
      - 300 transactions per second throughput

- Visibility Timeout
  - Amount of time message is invisible in a queue after a message is picked up
  - If the consumer does not delete the message the message becomes visible once again and could be processed twice
  - Default of 30 seconds and a max of 12 hours

- Short Polling vs Long Polling
  - Short polling is the default
  - Long polling can be a method to reduce costs because the application connecting to the queue will keep an active connection with the queue until a message is received or the timeout is reached which could reduce the number of times the application is connecting to the queue which would reduce costs
  - Receive message wait time setting for the queue

- **Amazon SNS**
  - General
    - Push-based Publisher/subscriber model
    - Standard and FIFO topics
    - Publishing endpoints include Lambda, SQS, HTTP Endpoint, SMS, Email
    - 256KB max message size unless storing on S3 then get 2GB
    - Send to Kinesis Data Firehose for archiving to a destination such as S3

  - Components
    - Topic
    - Subscription
    - Subscription Filter Policy
    - Redrive Policy (Dead Letter) - Messages that can't be delivered can be send to SQS queue

- **Simple Workflow Services (SWF)**
  - Web service used for coordinating work across application components and can include automated and manual tasks
  - Workflows can run for up to 1 year and run once and are never duplicated

- **Amazon API Gateway**
  - General

- General
    - Create, maintain, publish, monitor, secure APIs
    - Integrate with CloudWatch Logs for access logging for APIs
    - Supports integration with AWS WAF
    - Support for Oauth / Open ID Connect
    - Endpoints can be given custom domain name or use default domain name of APP_ID.execute-api.REGION.amazonaws.com

- Capabilities
    - Authentication with AWS IAM, Lambda, Cognito
    - Developer portal
    - Monitor and throttle requests
    - Mutual TLS
    - CORS

- Components
    - API
    - Stage
    - Route - direct incoming messages to integrations and have a method and path
    - Integration - connect route to backend resource with a method, URI, timeout
    - API key

- **Kinesis Data Streams**
    - General
        - Producer/consumer model where producers store streaming data in multiple shards which provides fixed unit of capacity
        - Data can be stored for up to 35 days (24 hours by default)
    - Shards
        - Read -> 5 transactions/sec up to total of 2MB
        - Write -> 1,000 transactions/sec up to total of 1MB
        - More shards equate to more capacity within a stream
        - Data distributed to shards based upon a partitioning key

- **Kinesis Data Firehose**
    - General
        - Producers stream data and Kinesis Data Firehose can optionally transform the data with Lambdas before it is delivered to the destination

the data with Lambda before it is delivered to the destination
        - Destinations include S3, RedShift, HTTP Endpoint
        - Records can be up to 1,000KB

- **Kinesis Data Analytics**
    - General
        - Integrate with streaming data services like Kinesis Data Firehose to get immediate insights on the data

- **Lambda**
    - General
        - Billed for the request, duration, memory with the first 1M requests being free
        - Max duration of 15 minutes
    - Languages
        - Node.js, Java, Python, C#, Go, PowerShell
    - Triggers
        - AWS Config
        - Amazon SNS
        - Amazon SQS
        - Amazon SES
        - Amazon S3
        - AWS Secrets Manager
        - DynamoDB
        - Application Load Balancer
        - Amazon EFS
        - CloudWatch
        - CloudFront (Lambda@Edge)
        - CodeCommit/CodePipelines
        - Kinesis Data Firehose

- **General**
    - AWS Networking Costs
        - Charged for traffic egressing out of VPC
        - Charged for inter-AZ traffic (cheapest option)
        - Charged for inter-region traffic

- VPC Endpoints
  - Gateway, Interface, Gateway Load Balancer
  - Gateway Endpoints are supported for S3 and DynamoDB and configure an efficient route to public prefix lists for those services
  - Interface Endpoints create an ENI in the VPC

- Transit Gateway
  - Regional resource used to connect multiple AWS network resources
  - Connect VPCs, VPNs, Direct Connect Gateway
  - Supports multi-cast

- AWS VPN CloudHub
  - Deployed in a VPC
  - Allows multiple VPN Sites to transitively communicate with each other through the Virtual Private Gateway deployed in the VPC

- Security Groups
  - Aligned with a region and VPC combination
  - By default all incoming denied and outgoing allowed
  - Stateful firewall
  - Allow rules ONLY
  - Attached to the network interface of an instance ONLY
  - Three-tuple -> Source (Or Destination for outbound) CIDR/Prefix List/Security Group, Port, Protocol
  - Max of 5 security groups per instance
  - Filter on destination ports only
  - Security groups can reference security groups in other VPCs if the VPCs are peered
  - Process all rules before allowing traffic

- NACL (Network Address Control List)
  - Stateless
  - Child of a VPC and can be associated to a subnet
  - New NACLs are by default deny all inbound/outbound while default NACL is allow all inbound/outbound
  - Subnet can only be associated with one NACL
  - Allows for blocking of specific IP addresses which Security Groups do not
  - Process rules in order

- Elastic IP
  - Static public IP
  - 5 per AWS account

- Elastic Network interface (ENI)
  - Use case is basic networking
  - One primary IPv4 from VPC and one or more secondary IPv4 from VPC
  - One public IPv4 and one or more IPv6
  - One Elastic IPv4 per private IPv4

  - Elastic Network Adapter (ENA) vs Virtual Function (VF)
    - Use case is for 10Gbps or 100Gbps and reliable high throughput
    - Enhanced networking for EC2 instances
    - ENA supports up to 100 Gbps
    - VF supports up to 10 Gbps for supported instances (older instances)

  - Elastic Fabric Adapter (ENF)
    - Use cases are HPC and ML
    - Provides lower and more consistent latency and higher throughput
    - Supports OS-bypass allowing HPC / ML applications to bypass OS kernal and hit EFA directly

- **<u>Amazon VPC (Virtual Private Cloud)</u>**
  - General
    - Largest CIDR block is /16 and smallest is /28
    - Each new VPC contains a default route table, default security group, and NACL

  - Subnets
    - Subnets are pinned to an availability zone
    - Amazon reserves first four IPs and last IP in each subnet
    - Subnets are not configured by default to assign public IPv4 to EC2 instances deployed in the subnet

  - Default VPC
    - All subnets have a route to the Internet
    - All EC2 instances have both public and private IPs

        - All EC2 instances have both public and private IP's

    - Internet Gateway
        - Associate one per VPC
        - Standard Internet Gateway and Egress-only Internet Gateway
        - Egress-only Internet Gateway is for IPv6 outgoing only

    - NAT Instance (Deprecated)
        - Single EC2 instance launched from a marketplace AMI
        - Launched into a subnet that has an Internet Gateway
        - Requires source and destination check to be disabled on the EC2 instance
        - Other instances are configured with a default route with the next hop to the NAT instance ENI

    - NAT Gateway
        - Deployed into single AZ but redundant in AZ
        - Comes in both public and private NAT
        - Public requires an Elastic IP
        - Deploy one to each AZ and route appropriately to relative NAT Gateway in the relative AZ

    - VPC Flow Logs
        - Metadata about traffic to and from a VPC
        - Written to S3 or CloudWatch Logs
        - Create at the VPC, Subnet, or ENI level
        - Log all traffic, accepted traffic, or rejected traffic
        - Does not log DHCP traffic, DNS queries to AWS DNS resolver, traffic to instance metadata endpoint, Windows License traffic, or traffic to AWS reserved IPs

- **AWS WAF (Web Application Firewall)**
    - AWS WAF - General
        - Layer 7 security service for CloudFront, Application Load Balancer, and API Gateway
        - Allows three behaviors: Allows all except what you specify, Block all except what you specify, Count requests that match what you specify

- AWS WAF - Protections
  - IP
  - Country
  - Headers
  - Regex matches in request
  - Length of request
  - SQL Injection
  - XSS

- **CloudFront**
  - Cloud Front - General
    - Components: distribution, origin, specifications
    - Supports AWS WAF
    - Content cached for 24 hours by default but can be invalidated for an extra cost

  - CloudFront - Signed URL/Signed Cookies
    - Use case is to protect premium paid content or to enforce specific rules based on IP or expiration
    - Policies associated with distribution can enforce conditions such as IP address, start/expiration time
    - Either trusted key groups or an AWS account can be used as a trusted signer
    - Public-private key pair used for verification

  - CloudFront - Origin Access Identity (OAI)
    - Unique CloudFront identity that is associated with a distribution for S3 origins
    - S3 bucket policy must allow the OAI access to the bucket
    - Use case is to restrict access direct to an S3 bucket

  - CloudFront - Restricting Access To CloudFront Content
    - OAI method can be used
    - Signed cookies and signed URLs can be used
    - Signed cookies and signed URLs use public/private key pairs
    - Trusted Key Groups are collections of trusted public keys while trusted signers are trusted AWS accounts that require root access to administer
    - Signed cookies and signed URLs allow you to enforce IP restrictions and

start/expiry times

- **Route 53**
  - Route 53 - General
    - Supports registration of domain names which can take up to three days to process
    - Alias record allows for alias at zone apex

  - Route 53 Routing Policies
    - Simple
      - Single record set with multiple values which returns records using round robin
    - Weighted
      - Multiple record sets each with a weight where higher weight equals more frequency
    - Latency
      - Multiple record sets where record is associated with a region and the region with the lowest latency to the end user's DNS will be returned
    - Failover
      - Multiple record sets each associated with a health check and one set as primary and one as secondary
    - Geolocation
      - Multiple record sets associated with a region and region closest to end user's DNS will be returned
    - Geoproximity
      - Available in Traffic Flow only
      - End users are routed via location but with option to include a bias which can increase or decrease the rate of return of the record
    - Multivalue Answer
      - Multiple record sets which behave similar to simple in that records are returned round robin but also includes a health check for each record set

  - Route 53 Health Checks
    - One health check per record set
    - Monitor an endpoint, status of another healthcheck (calculated healthcheck), or CloudWatch alarm (use case could be for private

resource)
        - Endpoint can be IP/DNS, supports hostname if IP, port, and path
        - Check can run every 10 or 30 seconds with a threshold
        - Option to pick which regions health checks originate from

- **AWS Direct Connect**
    - General
        - Direct connectivity to AWS services and VPCs
        - Private, Public, and Transit VIFs
        - 1Gbps or 10Gbps

    - Direct Connect Gateway
        - Global resource which allows connectivity to Virtual Private Gateway or Transit Gateway and resolves the challenge around connectivity across multiple regions with a single VIF
        - No support for transitive routing so one VPGW cannot connect to another VPGW through the Direct Connect Gateway

- **AWS Global Accelerator**
    - General
        - Global resource that optimizes performance, minimizes latency, and improves availability for globally deployed workloads
        - Cold potato routing solution where user is directed to nearest AWS edge location and is kept on AWS backbone
        - Customer is provided with two anycast IP addresses deployed in separate network zones (similar to AZ)
        - Benefit over Route 53 is the usage of anycast IP addresses avoids issues stale DNS records due to DNS cache
        - Place CloudFront in front if you require CDN
    - Features
        - Either uses ELB health checks or you can define your own health checks for EC2 and ENI use cases
        - Client affinity is supported with two-tuple load balancing algorithm
        - Users are load balanced with geo-proximity load balancing
        - Use traffic dial to maximize or minimize traffic to a specific region
        - Use weights to distribute traffic within a region
    - Components
        - Two anycast IP addresses
        - TCP/UDP listeners pointing to one or more endpoint groups

- TCP/UDP listeners pointing to one or more endpoint groups
    - Endpoint group associated to a region and containing one or more endpoints

- **Amazon RDS (Relational Database Service)**
    - RDS Supported Databases
        - MySQL
        - PostgreSQL
        - Microsoft SQL
        - Oracle
        - MariaDB
        - Aurora

    - RDS General
        - Optional automatic minor upgrades during a user-defined maintenance window
        - Supports user created snapshots which are retained as long as user wants
        - Storage can be magnetic, general purpose, or provisioned IOPS
        - Private by default and public optionally while be protected by Security Group
        - Restoring a backup or snapshot creates a new RDS instance which has a new DNS name

    - RDS Authentication and Authorization
        - IAM policies handle authorization at control plane
        - All RDS instances support username and password
        - Postgres, MySQL, and Aurora support IAM authentication and authorization at the data plane

    - RDS Automated Backups
        - Automatic full backups performed every day
        - Automatic transaction log backups every 5 minutes
        - Backups can be retained for up to 35 days with 7 as default
        - Customer can specify backup window
        - Set to 0 days to disable automated backup

    - RDS Snapshots
        - Performed manually

- Retained after the RDS instance is deleted

- RDS Storage Scaling
  - Automatically scales database storage
  - Three conditions must be true
    - Less than 10% free space of allocated storage
    - Low storage must last 5 minutes
    - 6 hours must have passed since last scale
  - Scales until you hit maximum threshold which you define

- RDS Read Replicas
  - Use case is to improve performance by distributing read activities
  - Asynchronous replication
  - Create read-only replicas in the same AZ or in another region
  - Must update connection string of application to use read-replicas
  - SELECT-only SQL statements
  - No charge for intra-AZ traffic when using this feature
  - Read-replicas can be enabled for Multi-AZ
  - Max of 5 read replicas for all offerings except for Aurora
  - Backups must be enabled

- RDS Multi AZ
  - Use case is for disaster recovery
  - Standby replica is provisioned in another AZ
  - Synchronous replication between master and standby
  - One DNS name for user and AWS manages cutover in the backend

- RDS Encryption
  - Supports encryption of entire database using KMS
  - TDE supported for Oracle and MS SQL
  - Master must be encrypted for read-replicas to be encrypted
  - Encryption MUST be defined at launch time

- RDS Encrypt an Existing RDS instance
  - Create a snapshot of the unencrypted RDS instance which will result in an unencrypted snapshot
  - Create a copy of the snapshot and select to encrypt which will create an encrypted snapshot

        - Create a new RDS instance from the encrypted snapshot which will create an encrypted instance
        - Point applications to the new RDS instance

- Aurora - General
    - Compatible with Postgres and MySQL
    - Increments by 10GB up to 128TB of storage
    - Provides up to 15 read-replicas which are automatically load balanced behind a single endpoint
    - Writer endpoint, reader endpoint, and custom endpoints
    - Supports ML through SQL queries when integrated with SageMaker or Comprehend

- Aurora Database Features
    - One writer and multiple readers
    - One writer and multiple readers plus parallel querying
    - Multiple writers
    - Serverless

- Aurora Custom Endpoints
    - Option to create custom endpoints with specific DB instances in the cluster
    - Use case is to create endpoints for different applications, some of which may require a larger instance type

- Aurora Autoscaling
    - Automatically adjust the number of read replicas based on a scaling policy
    - Scaling policy dictates min/max replicas and is triggered based on CloudWatch metrics and target tracking
    - Custom cooldown is available (300 seconds by default)

- Aurora Backtrack
    - Rewind the database without requiring the use of backups and restores
    - Enable at creation of Aurora cluster or when restoring a cluster
    - Changes can be reverted up to 72 hours (24 hours def)
    - Available for MySQL-compatible Aurora databases

- Aurora Multi-Master (Multiple Writers)

        - All DB instances in the cluster are writeable nodes with unique endpoints
        - Provides the capability of immediately failing over to a new writeable instance with no downtime
        - Max of four DB instances and must be in the same region
        - Available for MySQL-compatible Aurora databases

    - Aurora Parallel Query
        - Optimizes long-running queries by utilizing the shared storage of Aurora
        - Available for MySQL-compatible Aurora databases

    - Aurora Global Databases
        - Primary region contains the read/write DB instance
        - Each DB instances has its own endpoint
        - Up to 5 secondary regions with 15 read replicas per region
        - <1 second replication time
        - <1 minute RTO

    - Aurora Serverless
        - Define minimum and maximum resources
        - Excellent for unpredictable workloads
        - Pay per second

- **Amazon DynamoDB**
    - DynamoDB General Facts
        - Replicated within three AZs within a given region
        - Maximum item size of both attribute name and value is 400KB
        - On-Demand and Provisioned deployment models
        - Tables can be configured as standard or IA
        - Tables and global secondary indexes support autoscaling RCU/WCU

    - DynamoDB Consistency Models
        - Eventually Consistency (Default)
        - Strongly Consistent
        - ACID Transaction

- DynamoDB Eventually Consistent Consistency Model
  - Default model
  - Provides maximum read throughput when compared to other models
  - Consistency is usually reached in 1 second

- DynamoDB Strongly Consistent Consistency Model
  - Returns result that reflects that all writes have completed successfully
  - Guarantees consistency in 1 second

- DynamoDB Provisioned Capacity Model
  - Pay per hour for RCU/WCU
  - Pay for data storage
  - Pay for data transfer out of region

- DynamoDB On-Demand Capacity Model
  - Use case is for a new application where the demand is unknown
  - Could be more costly for ACID transactions
  - RCU/WCU is automatically scaled
  - Pay for each request, data storage, and traffic leaving the region

- DynamoDB Accelerator (DAX)
  - Fully managed, HA, in-memory cache that sits between application and Dynamo
  - Supports both reads and writes
  - Reduces request time to microseconds

- DynamoDB Transactions
  - Use case example is financial transaction where both the debit and credit must occur
  - Two underlining reads and writes with one for preparing and one for the commit
  - Single request can be 25 items or 4MB

- DynamoDB On-Demand Backup/Restore
  - Full backups can be performed at any time without performance impact and are completed within seconds
  - Retained until explicitly deleted
  - Backups are stored and can be restored only within the same AWS

account and region

- DynamoDB Point-In-Time-Recovery (PITR)
  - Protects against accidental writes or deletes
  - Recovery from any point within the last 35 days
  - RPO of 5 minutes
  - NOT enabled by default

- DynamoDB Streams
  - Use cases are messaging/notifications for application or aggregating data between DynamoDB instances
  - Time-recorded sequence of item-level changes in a table
  - Stored for 24-hours
  - Stream records are organized into shards
  - Combine with a Lambda to create something similar to a stored procedure

- DynamoDB Global Table
  - Multi-master and multi-region
  - Uses DynamoDB Streams under the hood
  - Use this for DR or HA
  - Replication latency is under 1 second

- **Amazon ElastiCache**
  - ElastiCache General
    - Supports MemCached and Redis
    - MemCached uses sharding and is non-persistent
    - Redis supports backups, high availability (multi-AZ), persistent data
    - Supports encryption via KMS and Security Groups for network control
    - Use cases include cached database queries or session stores

  - ElastiCache Redis
    - Add authentication to data plane using Redis Auth
    - Selected sets order records and add uniqueness (such as for leaderboards)

  - ElastiCache Patterns
    - Lazy Loading -> Application queries cached store first and only queries

database if something isn't cached which can result in stale records
    - Write-through -> Application adds or updates the cache when the database is modified resulting in no stale records
    - Session Store -> using cached store to store user sessions to make an application stateless

- **(NOTES) Amazon RedShift**
    - General
        - Scales up to 128GB
        - Deploys into a single AZ
        - Leader nodes and compute nodes
        - RedShift Spectrum queries data in S3
        - Enhanced VPC Routing allows COPY/UNLOAD through VPC

    - Disaster Recovery
        - Snapshots can be done automatically ever 8 hours or 5GBs or manually and are held for a retention period you set
        - Snapshots can be automatically configured to be copied to another region

    - Loading Data
        - Manually copied from S3 to RedShift
        - Kinesis Data Firehose copies from S3 to RedShift
        - EC2 instance running JDBC driver

- **Amazon EMR**
    - EMR General
        - Big data platform
        - Components are the cluster, none, and node type
        - Node types include master, core, and task nodes

    - EMR Node Types
        - Master Node - manages the cluster and tracks status of the tasks
        - Core Node - runs software components that execute tasks and store data in HDFS
        - Task Node - runs software components that executes tasks but does not store data in HDFS

    - EMR Logs

        - Logs stored only on master node so if the master node is lost all logs are lost
        - Logs can be configured to archive to S3 at 5 minute interval
- **AWS IAM**
    - AWS STS
        - Used to generate short-term credentials for trusted entities to grant access to AWS resources
        - Credentials last minutes to hours
        - Credentials are not stored with the user and are generated dynamically for the user
        - Service is used for IAM Roles and federated users

    - IAM Entities
        - Root User (Account Owner)
            - One per AWS account and associated with email used to create the AWS account

        - IAM User
            - Represents human or non-human entity within an AWS account
            - Assigned a unique username and password which is used for access to AWS Console
            - Allowed up to two access keys for programmatic access keys

        - IAM Group
            - Contains one or more IAM Users
            - Used to group IAM users by function then assign function a set of permissions

        - IAM Roles
            - Entity without credentials and is assumed by human or non-human
            - Assumed by federated user, IAM User, AWS resource, and other entities
            - AWS STS provides a temporary set of credentials for use of the role

    - IAM Policy
        - General
            - Denies in any policy take precedent over Allow
            - Cross account access only works within an AWS partition (such as

commercial, government, China)
    - AWS stores up to 5 versions of a customer managed policy
    - Customize the alias for an account to modify the URL of https://ALIAS.signin.aws.amazon.com/console

- Policy Structure
    - Version
    - Statement (array)
        - SID (optional, quick description)
        - Effect
        - Principal
        - Action
        - Resource
        - Condition (optional)

- Identity-based Policy
    - Attached to IAM User, Group, or Role
    - AWS-managed policies, customer-managed policies, and inline policies
    - AWS-managed policies cannot be modified, are updated automatically by AWS, and consist of job-function and task-related policies
    - Customer-managed policies exist within a single AWS account and are managed by the account owner
    - Inline policies are embedded in an single IAM entity and share the lifecycle of the IAM entity

- Permissions Boundary
    - Sets maximum permissions that an identity-based policy can grant to an IAM entity
    - Does not grant access to do something but rather sets the boundary of what an IAM entity can do
    - Changes effective permissions of the identity-based policy but NOT resource-based policies
    - Assigned to an IAM entity

- Multifactor Support
    - Supports Virtual MFA devices, U2F Security Keys, Hardware MFA
    - Enable through the AWS Console -> IAM -> User's Security

- Enable through the AWS console -> IAM -> User's Security Credential tab

- Password Policy
  - Options
    - Length
    - Strength (upper case, lower case, number, symbol)
    - Expiration
    - Expiration requires admin to reset the password
    - Allow users to change their password
    - Password history (1-24)
  - Default
    - Password length of 8 characters
    - Minimum of three (upper case, lower case, number, symbol)
    - Cannot be the same as account name or email address

- Policy Evaluation in a Single AWS Account
  - All actions are DENIED by default (except root account)
  - All policies are evaluated for any Denies and action is evaluated against those Denies and is stopped if it matches any
  - Service Control Policies (SCPs) are evaluated
  - Resource Policies are evaluated (action is allowed if resource policies allows it)
  - Permission boundaries are processed
  - Session policies are processed
  - Identity-based policies are processed

- Cross Account Access Facts
  - Resource-based policies can allow all IAM entities in a trusted AWS account by including just the AWS account ID or only specific entities by including the entity's full ARN
  - Common practice is to attach resource policy (called a trust policy in this case) to a role an allow the trusted account to assume the role; this role is granted permissions within the trusting account
  - If using a resource policy in a trusting account allows an IAM entity from a trusted AWS account the IAM entity from the trusted account can still access resources in the trusted account allowing for use cases like copying data across AWS accounts
  - Principals specified in resource policies include IAM Users, IAM

Roles, Federated Users, Assumed-role sessions, and AWS Services

        - Cross Account Policy Evaluation Process
            - □ IAM entity goes through standard single-account evaluation process (everything that can effect identity-based policy)
            - □ Resource policy of trusting account is evaluated and any policies that limit access to the resource (session policies)
            - □ Both the trusted evaluation and trusting evaluations must allow the access

    - IAM Credential Report
        - Generated in IAM section of the AWS Console
        - Creates a CSV which lists each IAM User, their password information (expiration, last changed, etc) MFA status, and access keys and age

    - IAM Access Advisor
        - Lists services and permissions in a service an entity has access to
        - Indicates which policy is providing that access
        - Displays the last time the user accessed a service and permissions
        - Use to remove unneeded permissions during the access certification process

    - IAM Access Analyzer
        - Use case is to identify resources in an account which are accessible by an external entity
        - During creation you specify a zone of trust which can be account or organization
        - Findings are instances where a resources can be accessed by entities outside the trust zone of the analyzer instance
        - Additionally Analyzer can perform grammar and best practice checks on IAM Policies
        - Additionally Analyzer can review the actions performed by an entity using CloudTrail data across a specific range to suggest an IAM Policy for that entity

- **<u>Amazon Cognito</u>**
    - User Pools

        - User directories used for sign in/sign up and can store user identity and credentials authoritatively or integrated with a 3rd party IdP
    - Identity Pools
        - Authorization repository for mapping unauthenticated and authenticated users to AWS IAM Roles
- Secrets Manager
    - Charged per secret and 10,000 API calls
    - Automation rotation of secrets
    - Generate random secrets for use in CloudFormation templates

- AWS Shield
    - Standard is provided for free and offers basic DDoS protection for public facing resources
    - Advanced is available for EC2, ELB, CloudFront, Global Accelerator, and Route 53
    - Provides cost protection and access to IRT