Link: tinyurl.com/SAA-notes

The capability of EC2:

1. renting virtual machines(EC2)
2. storing data on virtual drives(EBS)
3. distributing load across machines(ELB)
4. Scaling the services using an auto-scaling group(ASG)

EC2 sizing and configuration options

1. OS(Linus, Windows or Mac OS)
2. Compute power & cores(CPU)
3. RAM
4. Storage space(Network-attached(EBS & EFS), hardware(EC2 Instance Sore))
5. Network card(speed of the card, public IP address)
6. Firewall rules(Security group)
7. Bootstrap script(EC2 User Data(runs with the root user)

Bootstrapping means launching commands when a machine starts

Security groups

1. firewalls on EC2 instances
2. can be attached to multiple instances
3. locked down to a region/VPC combination
4. is a firewall outside of EC2
5. it's good to maintain one separate security group for SSH access
6. time out is 100% a security group issue
7. connection refused is an application error or its not launched
8. By default all inbound traffic is blocked, all outbound traffic is authorized
9. cannot restrict access based on the user's geographic location.

Classic Ports

1. 22 = SSH (Secure Shell) - log into a linux instance
2. 21 = FTP (File Transfer Protocol) - upload files into a file share
3. 22 = SFTP (Secure File Transfer Protocol) - upload files using SSH
4. 80 = HTTP - access unsecured websites
5. 443 = HTTPS - access secured websites
6. 3389 = RDP (Remote Desktop Protocol) - log into a Windows instance

Dedicated instances V.S. Dedicated Host

- Dedicated instances means u have ur own instances on ur hardware while dedicated hosts mean u can access the physical server and it allows u to check low-level hardware

Spot Fleets

- allow us to automatically request Spot instances with the lowest price
- Spot Fleets = set of Spot instances + optional on-demand instance

Types of EC2 Instances

- General Purpose EC2 instances
    - great for a diversity of workloads that require a balance between computing, memory, and network.
- Storage Optimized EC2 instances
    - great for workloads requiring high, sequential read/write access to large data sets on local storage.
- Memory Optimized EC2 instances
    - great for workloads requiring large data sets in memory.
- Compute Optimized EC2 instances
    - great for compute-intensive workloads requiring high-performance processors (e.g., batch processing, media transcoding, high-performance computing, scientific modeling & machine learning, and dedicated gaming servers).

Elastic IP

- An elastic IP is a public IPV4 IP you own as long as you don't delete it
- it’s a bad idea to use that, it's better to use a random public IP and register a DNS name to it, or the best practice is to use a Load Balancer without a public IP
- you can have up to 5 elastic ip(can ask Amazon to increase the limit)
- Can be attach to one instance at a time

Placement Groups

- controls how ec2 instances are placed
- strategies:
    - **cluster** - cluster instances into a low-latency group in a single AZ(high performance and high risk)
        - use case: big data job that needs to complete fast, application that needs extremely low latency and high network throughput
    - **spread** - spreads instances across underlying hardware (max 7 instances per group per AZ for critical applications)
        - use case: high availability and isolated from failure
    - **partition** - spreads instances across many different partitions within an AZ. Scales to 100s of EC2 instances(EC2 instances could get access to the partition information as metadata)
        - use case: HDFS Hbase, Cassandra, Kafka

ENI(Elastic Network Interface)

- logical component in a VPC that represents a virtual network card
- Attributes:
    1. One primary private IPv4, one or more secondary IPv4
    2. One ElasticIP(IPv4) per private IPV4
    3. One Public IPv4
    4. One or more security groups
    5. A MAC address
    6. can be created independently and be moved between EC2 instances for failover
    7. Bound to a specific AZ

EC2 Hibernate

- When a hibernating happens, the RAM is encrypted and written to a file in the root EBS volume. When rebooted, the OS would start faster.
- The EC2 Instance Root Volume type must be an EBS volume and must be encrypted to ensure the protection of sensitive content.
- use case:
  - Long-running processing
  - Save the RAM state
  - Services that take time to initialize

EBS

- is a network drive you can attach to instances while they run
- it allows ur instances to persist data even after termination
- EBS can be mounted to multiple instances
  - EBS is bound to a specific AZ
  - can be moved across AZ by snapshot
- have a provisioned(pre-setup) capacity
- Delete on Termination attribute
  - by default, the root EBS would be deleted if an EC2 instance is terminated
  - by default, any other attached EBS volume is not deleted
- EBS backups use IO and you shouldn’t run them while your application is handling a lot of traffic

EBS Snapshots

- is a backup of ur EBS volume
- unnecessary but recommended
- EBS can be copied across AZ or even Region using Snapshot
- EBS Snapshot Archive
  - move a Snapshot to an "archive tier" thats 75% cheaper
  - takes 24-72 hours to restore
- EBS Snapshots Recycle Bin
  - deleted snapshots are put in a recycle bin
  - the retention can be specified from 1 day to 1 year
- Fast Snapshot Resore(FSR)
  - Force full initialization of snapshot to have no latency on the first use(costly)
  - allows you to quickly restore EBS snapshots into new EBS volumes
  - use case: disaster recovery, backup restore, and TEST/DEV environment

EBS Encryption

- When you create an encrypted EBS volume, you get the following:
  - Data at rest is encrypted inside the volume
  - All the data in flight moving between the instance and the volume is encrypted
  - All snapshots are encrypted
  - All volumes created from the snapshot are encrypted
  - Encryption and decryption are handled transparently (you have nothing to do)
- Encryption has a minimal impact on latency
- EBS Encryption leverages keys from KMS (AES-256)
- Copying an unencrypted snapshot allows encryption
- Steps to encrypt an unencrypted EBS volume
  1. Create an EBS snapshot of the volume

2. Encrypt the EBS snapshot (using copy)
3. Create new EBS volume from the snapshot (the volume will also be encrypted)
4. Now you can attach the encrypted volume to the original instance

AMI(Amazon Machine Image)

- are a customization of an EC2 instance
- are built for a specific region(can be copied across regions)
    - public AMI, provided by AWS
    - your own AMI, maintain by you
    - AWS Marketplace AMI（sold by others
- build an AMI will also create a EBS snapshot

EC2 Instance Store

- used for high-performance hardware disk since EBS volumes are network drives and are limited by network latency
- offers better I/O performance
- loses storage if they are stopped
- Use case: buffer / cache / scratch data / temporary content
- However if hardware fails then all data is loss
- need backup and replication

When creating EC2 instances, you can only use the following EBS volume types as boot volumes: gp2, gp3, io1, io2, and Magnetic (Standard).

EBS volume types

- General Purpose SSD(gp3/gp2)
    - balance of price and performance
    - gp2: IO increases if the disk size increases
    - use case
        - low-latency interactive apps
        - development and test environment
    - GP3 allows 16000 IOPS and is cheaper than gp2
- provisioned IOPS SSD(io1/io2)
    - Highest-performance SSD volume for mission-critical low-latency or high-throughput workloads
    - io1: can increase IO independently
    - use case
        - Critical business applications with sustained IOPS performance or applications that need more than 16,000 IOPS
        - Great for databases workloads (sensitive to storage perf and consistency)
    - IOPS > 32000, use EC2 nitro with io1/io2

- Hard Disk Drives (HDD)(st1/sc1)
    - Cannot be a boot volume
    - use case:
        - high throughput and lowest price
        - bigdata, datawarehouse or log processing

    - st1 (HDD): Low cost HDD volume designed for frequently accessed, throughput- intensive workloads
    - sc1 (HDD): Lowest cost HDD volume designed for less frequently accessed workloads

EBS Multi-attach - io1/io2 family

- means to attach the same EBS volume to multiple EC2 instances in the same AZ
- Each instance has full read & write permissions to the high-performance volume
- Must use a file system that's cluster-aware(not XFS, EXT4)
- usecase:
    - achieve higher application availability in clustered Linux applications(ex: Teradata)
    - Applications must manage concurrent write operations
- Allows up to 16 EC2 instances at a time

EFS(Elastic File System)

- can be mounted with up to 100 EC2 instances in multiple AZs
- highly available, scalable but expensive. No capacity planning since pay by use
- use cases: content management, web serving, data sharing and Wordpress
- uses NFSv4.1 protocol
- uses security group to control access to EFS
- only compatible with Linux based AMI(not windows) with POSIX FS
- Encryption at rest with KMS
- EFS storage tiers:
    - Standard: for frequently accessed files
    - Infrequent access(EFS-IA): cost to retrieve files with a lower price to store
- EFS Availability and durability
    - Standard: Multi-AZ, for prod
    - One Zone: One-AZ, for dev. Backups are enabled by default. This option is compatible with IA(EFS One Zone-IA)

AWS Network Firewall

- is a managed firewall service that provides filtering for both inbound and outbound network traffic. It allows you to create rules for traffic inspection and filtering, which can help protect your production VPC.

AWS Firewall Manager

- is a security management service that helps you to centrally configure and manage firewalls across your accounts.

S3

- buckets must have a globally unique name across all regions and all accounts
- S3 buckets are created in a region
- There are no S3 data transfer charges when data is transferred in from the internet
- objects
    - every object(file) has a key, which is the full path
    - object values are the contents of the body
    - max object size is 5TB(5000GB)
    - if uploading more than 5 GB, must use multi-part upload
    - Metadata(list of text key / value pairs - system or user metadata)

  - Tags(Unicode key / value pair - up to 10 - for security or lifecycle)
  - Version ID(if versioning is enabled)
- S3 Security
  - User-based
    - IAM policies, controls which API calls should be allowed for a specific user from IAM
  - Resource-based
    - bucket policies - bucket-wide rules from the S3 console - allows cross account
    - object ACL - finer grain(can be disabled)
    - bucket ACL - less common(can be disabled)
- S3 Bucket Policies - JSON based policies
  - Resources: buckets and objects
  - Effect: Allow / Deny
  - Actions: Set of API to Allow or Deny
  - Principal: The account or user to apply the policy to
  - use case
    - Grant public access to the bucket
    - Force objects to be encrypted at upload Grant access to another account (Cross account access must use bucket policy)
- S3 Versioning
  - enabled at the bucket level
  - ability to restore a version and easy roll back
  - default version id is "null"
  - suspending versionsing does not delete the previous versions
- S3 Replication(Cross-Region Replication & Same-Region Replication)
  - requires versioning enabled in the source and destination buckets
  - buckets can be in different aws accounts
  - copying is asynchronous
  - proper IAM permissions to S3 is needed
  - Use case
    - CRR: compliance, lower latency access, replication across accounts
    - SRR: log aggregation, live replication between PROD and TEST
  - After replication is enabled, only new objects would be replicated. Existing objects should use S3 Batch Replication to be replicated
  - For DELETE operations
    - Can replicate delete markers with an optional setting
    - Deletions with a version ID are not replicated to avoid malicious deletes
  - No chaining of replication(if A copy to B, B copy C, A does not copy to C)
- S3 Durability and Availability
  - Durability:
    - High durability (99.999999999%, 11 9’s) of objects across multiple AZ
    - If you store 10,000,000 objects with Amazon S3, you can on average expect to incur a loss of a single object once every 10,000 years
    - Same for all storage classes
  - Availability:
    - Measures how readily available a service is
    - varies depending on storage class
    - Eg: S3 standard has 99.99% availability = not available 53 minutes a year
- S3 analytics

    - provide storage class analysis which help u decide when to transition objects to the right storage class
    - Recommendations for Standard and Standard IA(does not work for one-zone IA or Glacier)
- S3 Requester pays
    - the requester (must be an authenticated AWS user) would pay for networking cost while the bucket owner still needs to pay for storage cost
- S3 event notifications
    - when an event happens, send notifications to SQS, SNS, Lambda or Amazon EventBridge
    - Amazon EventBridge offers
        - advanced filtering options with JSON rules(metadata, object size, name)
        - Multiple Destinations - ex Step Functions, Kinesis Streams / Firehose
        - EventBridge Capabilities - Archive, Replay Events and reliable delivery
    - object name filtering possible(e.g. *.jpg)
    - create as many "S3 events" as u want
    - usually takes seconds but can take 1 min+
- S3 Baseline Performance
    - 3500 put/copy/post/delete or 5500 get/head requests per second per prefix in a bucket
- S3 Multi-part upload
    - recommend for 100 MB files, must for 5GB+ files
    - can help parallelize uploads
- S3 Transfer Acceleration
    - transfer your data to an nearby AWS edge location and then AWS forward the data to the bucket in the target region using AWS private network
    - Compatible with multi-part upload
    - Empowers both upload and downlaod
    - you pay only for transfers that are accelerated
- S3 Byte-range Fetches
    - Parallelize GETs by requesting specific byte ranges
    - use case: speed up downloads or only need partial data(eg. the head of a file)
- S3 Select & Glacier Select
    - Retrieve less data using simple SQL by performing server-side filtering
    - can filter by rows & columns
    - less network transfer and less CPU cost client-size
- S3 Batch Operations
    - perform bulk operations on existing S3 objects with a single request
    - A job consists of a list of objects, the action to perform and optional parameters
    - S3 Batch Operations manages retries, tracks progress, sends completion notifications and generate reports
    - Steps:
        - use S3 Inventory to get a object list
        - use S3 Select to filter ur objects
        - use S3 Batch Operations to process object
- S3 Security
    - Server-Side Encryption (**SSE**)
        - Server-Side Encryption with Amazon S3-Managed Keys (**SSE-S3**)
            - Encrypts S3 objects using keys managed, and owned by AWS
            - Uses AES-256
            - The default option of all newly created objects in S3

      - Can’t set key rotation since it’s managed be S3
  - Server-Side Encryption with KMS Keys stored in AWS KMS (**SSE-KMS**)
    - Leverage AWS KMS to manage encryption keys
    - Use case: user control + audit key usage using CloudTrail
    - Disadvantages:
      - limited by KMS API call quota
      - upload data uses GenerateDataKey()
      - download data uses Decrypt()
  - Server-Side Encryption with Customer-Provided Keys (**SSE-C**)
    - When you want to manage your own encryption keys
  - Client-Side Encryption
    - data must be encrypted before uploading to and after taking from S3
- Encryption in transit(SSL/TLS)
- Force Encryption
  - uses policies to deny PUT requests without a security header x-amz-server-side-encryption
  - Another way is to use the “default encryption” option in S3
  - Note: Bucket Policies are evaluated before “default encryption”
- Cross-Origin Resource Sharing (CORS)
  - Origin = scheme (protocol) + host (domain) + port
  - CORS is a web browser security that allows you to enable images or assets or files being retrieved from one S3 bucket in case the request is originating from another origin
- MFA Delete
  - To use MFA Delete, Versioning must be enabled on the bucket
  - Only the root account can enable/disable MFA Delete
  - MFA is required when:
    - Permanently delete an object version
    - Suspend Versioning on the bucket
- S3 Access Logs
  - you can log any requests made to S3 into another(to avoid loop) S3 bucket for analysis
- S3 Pre-Signed URLs
  - Users given a pre-signed URL inherits the permissions of the user that generated the URL for GET / PUT
  - Use case: grant temporary access
  - Examples:
    - Allow only logged-in users to download a premium video from your S3 bucket
    - Allow an ever-changing list of users to download files by generating URLs dynamically
    - Allow temporarily a user to upload a file to a precise location in your S3 bucket
- S3 Glacier Vault Lock
  - when an object is moved into a Glacier Vault, NO ONE can change or delete it
  - use case: compliance and data retention
- S3 Object Lock (versioning must be enabled)
  - blocks an object from deletion for a specific predefined retention period
  - Retention mode - **Compliance**:
    - Object versions can't be overwritten or deleted by any user, including the root user
    - Objects retention modes can't be changed, and retention periods can't be shortened
  - Retention mode - **Governance**:
    - Some users have special permissions to change the retention or delete the object
  - Legal Hold:

        - protect the object indefinitely, independent from retention period
        - can be freely placed and removed using the s3:PutObjectLegalHold IAM permission
- S3 Access Points
    - Each Access Point gets its own DNS and policy to limit who can access it
    - A specific IAM user / group
    - One policy per Access Point => Easier to manage than complex bucket policies
- S3 Object Lambda
    - Use AWS Lambda Functions to change the object before it is retrieved by the caller application
    - Use Cases:
        - Edit personally identifiable information for analytics or non-production environments
        - Convert across data formats, such as converting XML to JSON
        - Resize and watermark images on the fly using caller-specific details, such as the user who requested the object

CloudFront

- is a Content Delivery Network, which improves read performance by caching content at edge locations
- if the required content is not cached in an edge location, it has to be fetched from the original src
- provides DDoS protection by integration with Shield and Web Application Firewall(WAF)
- Can expose external HTTPS and can talk to internal HTTPS backends
- CloudFront Geo Restriction
    - you can restrict users from which country can access your distribution by Whitelist or Blacklist
    - use case: copyright laws
- Price Classes
    - Price Class All: all regions - best performance
    - Price Class 200: most regions but excludes the most expensive ones
    - Price Class 100: only the least expensive regions
- Cache Invalidation
    - when you update your origin and want all cached content in global edge locations to be refreshed
    - you can force an entire or partial cache refresh
    - you can invalidate all files(*) or a special path (/images/*)

AWS Global Accelerator

- Leverages the AWS internal network to route to your application
- Works with Elastic IP, EC2 instances, ALB, NLB, public or private
- Consistent Performance
    - Intelligent routing to lowest latency and fast regional failover
    - No issue with client cache (because the IP doesn't change)
    - Internal AWS network
- Health Checks
    - Global Accelerator performs a health check of your applications
    - Helps make your application global (failover less than 1 minute for unhealthy)
    - Great for disaster recovery (thanks to the health checks)
- Security
    - only 2 external IP need to be whitelisted
    - DDoS protection thanks to AWS Shield

AWS Global Accelerator V.S. CloudFront

- They both use the AWS global network and its edge locations around the world
- Both services integrate with AWS Shield for DDoS protection
- CloudFront
    - Improves performance for both cacheable content (such as images and videos)
    - Dynamic content (such as API acceleration and dynamic site delivery)
    - Content is served at the edge
- Global Accelerator
    - Improves performance for a wide range of applications over TCP or UDP
    - Proxying packets at the edge to applications running in one or more AWS Regions
    - Good fit for non-HTTP use cases, such as gaming (UDP), IoT (MQTT), or Voice over IP
    - Good for HTTP use cases that require static IP addresses
    - Good for HTTP use cases that required deterministic, fast regional failover

CloudFront V.S. S3 Cross Region Replication

- CloudFront
    - CloudFront uses global edge network
    - files are cached for a TTL(like a day)
    - greate for static content that must be available everywhere
- S3 Cross Region Replication
    - must be setup for each region you want replication
    - files are updated in near real-time
    - read only
    - great for dynamic content that needs to be available at low-latency regions

Trusted Advisor

- Provide recommendation on: Cost optimization, Performance, Security, Fault tolerance and Service limits
- Support Plans
    - 7 Core checks for Basic and Developer Support plan
        - S3 Bucket Permissions
        - Security Groups - Specific Ports Unrestricted
        - IAM Use (one IAM user minimum)
        - MFA on Root Account
        - EBS public Snapshots
        - RDS Public Snapshots
        - Service Limits
    - Full checks for Business and Enterprise Support plan
        - Ability to set CloudWatch alarms when reaching limits
        - Programmatic access using AWS support API

Amazon SQS - Standard Queue

- a fully managed service to decouple applications
- attributes:
    - unlimited throughput
    - unlimited number of messages in queue
    - Default retention of messages: 4 days, maximum of 14 days
    - low latency(< 10 ms on publish and receive)

    - limitation of 256 KB per message sent
    - Can have duplicate messages
    - can have out of order messages
- Producer produces messages to SQS using SendMessage() API in SDK, the message will be persisted in SQS until a consumer deletes it
- Consumers(running on EC2 instances, servers or AWS Lambda) poll (up to 10) messages from SQS using ReceiveMessage() API, then process the message and delete the messages from the queue using the DeleteMessage() API
- Security:
    - Encryption
        - In-flight encryption using HTTPS API
        - At-rest encryption using KMS keys
        - Client-side encryption if the client wants to perform encryption/decryption itself
    - Access Controls
        - IAM policies to regulate access to the SQS API
    - SQS Access Policies (similar to S3 bucket policies)
        - Useful for cross-account access to SQS queues
        - Useful for allowing other services (SNS, S3…) to write to an SQS queue
- Message Visibility Timeout
    - A message is invisible to other consumers if it was polled by a consumer already. However, there's a limit of time when this message is invisible
    - By default, the time limit is 30 seconds. If 30s later the message has still not being processed, it will become visible again, which might cause duplicated processing
    - A consumer could call the ChangeMessageVisibility() API to get more time
- Long Polling
    - A consumer can wait for messages in an empty queue
    - The wait time can be 1s to 20s(20s is preferable)
    - Decreases the number of API calls and increase the efficiency and reduce latency
    - Long polling can be enabled at the queue level or at the API level using WaitTimeSeconds()

Amazon SQS - FIFO Queue

- Limited throughput: 300 msg/s without batching, 3000 msg/s with batch(10 msg in one batch)
- Exactly-once send capability (by removing duplicates)
- Messages are processed in order by the consumer

Amazon SNS

- An event producer sends message to one SNS topic, then unlimited event listeners listen to the SNS topic notification and all get the same message(or filtered message upon setting)
- 12,500,000 subscriptions per topic and 100,000 topics limit
- Security
    - Same with SQS Security
- SNS + SQS: Fan Out
    - Push once in SNS, receive in all SQS queues that are subscribers
    - can add more SQS subscribers over time
    - make sure SQS queue access policy allows for SNS to write
- FIFO
    - can only have SQS FIFO queues as subscribers
    - same as SQS FIFO

- Message FIltering
    - JSON policy used to filter messages sent to SNS topic's subscriptions
    - If a subscription doesn't have a filter policy, it receives every message
    -

Dead-letter queues can be used by other queues (source queues) as a target for messages that can't be processed (consumed) successfully.

Kinesis

- Makes it easy to collect, process, and analyze streaming data in real-time
- Ingest real-time data such as: Application logs, Metrics, Website clickstreams, IoT telemetry data
- Kinesis Data Streams: capture, process, and store data streams
- Kinesis Data Firehose: load data streams into AWS data stores
- Kinesis Data Analytics: analyze data streams with SQL or Apache Flink
- Kinesis Video Streams: capture, process, and store video streams

Kinesis Data Streams

- Retention between 1 day to 365 days
- Ability to reprocess (replay) data
- Once data is inserted in Kinesis, it can't be deleted (immutability)
- Data that shares the same partition goes to the same shard (ordering)
- Capacity Modes
    - Provisioned mode:
        - You choose the number of shards provisioned, scale manually or using API
        - Each shard gets 1MB/s in (or 1000 records per second)
        - Each shard gets 2MB/s out (classic or enhanced fan-out consumer)
        - You pay per shard provisioned per hour
    - On-demand mode:
        - No need to provision or manage the capacity
        - Default capacity provisioned (4 MB/s in or 4000 records per second)
        - Scales automatically based on observed throughput peak during the last 30 days
        - Pay per stream per hour & data in/out per GB
- Security
    - Control access / authorization using IAM policies
    - Encryption in flight using HTTPS endpoints
    - Encryption at rest using KMS
    - You can implement encryption/decryption of data on client side (harder)
    - VPC Endpoints available for Kinesis to access within VPC
    - Monitor API calls using CloudTrail

Kinesis Data Firehose

- Fully Managed Service for data ingestion, no administration, automatic scaling, serverless
- Desitinations:
    - AWS: Redshift / Amazon S3 / ElasticSearch
    - 3rd party partner: Splunk / MongoDB / DataDog / NewRelic / …
    - Custom: send to any HTTP endpoint
- Pay for data going through Firehose
- Near Real Time

    - 60 seconds latency minimum for non full batches
    - Or minimum 1 MB of data at a time
- Can send failed or all data to a backup S3 bucket
- You cannot set up multiple consumers for Kinesis Data Firehose delivery streams as it can dump data in a single data repository at a time, so this option is incorrect.

Amazon MQ

- A managed message broker service for RabbitMQ and ActiveMQ
- Amazon MQ doesn't "scale" as much as SQS / SNS
- Amazon MQ runs on servers, can run in Multi-AZ with failover
- Amazon MQ has both queue feature (~SQS) and topic features (~SNS)

Amazon RDS

- a managed DB service for DB uses SQL as a query language.
    - Automated provisioning, OS patching
    - Continuous backups and restore to specific timestamp (Point in Time Restore)!
    - Monitoring dashboards
    - Read replicas for improved read performance
    - Multi AZ setup for DR (Disaster Recovery)
    - Maintenance windows for upgrades
    - Scaling capability (vertical and horizontal)
    - Storage backed by EBS (gp2 or io1)
    - BUT you can't SSH into your instances
- RDS - Storage Auto Scaling
    - When RDS detects you are running out of free database storage, it scales automatically
    - You have to set Maximum Storage Threshold (maximum limit for DB storage)
    - Automatically modify storage if:
        - Free storage is less than 10% of allocated storage
        - Low-storage lasts at least 5 minutes
        - 6 hours have passed since last modification
    - Useful for applications with unpredictable workloads
- RDS read replicas
    - can have up to 5 read replicas within AZ, cross AZ or cross Region
    - Async replication and eventually consistent
    - Replicas can be promoted to their own DB
    - Applications must update the connection string to leverage read replicas
    - No data transfer network cost for same region replication
- RDS Multi AZ(Disaster Recovery）
    - Sync replication
    - Increase availability
    - One DNS name - automatic app failover to standby
    - Multi-AZ replication is free
    - Note:The Read Replicas can be setup as Multi AZ for Disaster Recovery (DR)
- RDS - From Single-AZ to Multi-AZ
    - No downtime(no need to stop the DB)
    - simply click "modify" for the database
    - underline steps:

        - A snapshot is taken
        - A new DB is restored from the snapshot in a new AZ
        - Synchronization is established between the two databases
- RDS Custom
    - Managed Oracle and Microsoft SQL Server Database with OS and database customization
    - full admin access to the underlying OS and the database
        - Configure settings
        - Install patches
        - Enable native features
        - Access the underlying EC2 Instance using SSH or System Manager Session Manager
- RDS Backups
    - Automated backups
        - Daily full backup of the database (during the maintenance window)
        - Transaction logs are backed-up by RDS every 5 minutes => ability to restore to any point in time (from oldest backup to 5 minutes ago)
        - 1 to 35 days of retention, set 0 to disable automated backups
    - Manual DB Snapshots
        - Manually triggered by the user
        - Retention of backup for as long as you want
    - Trick: in a stopped RDS database, you will still pay for storage. If you plan on stopping it for a long time, you should snapshot & restore instead
- RDS Proxy
    - fully managed database proxy for RDS
    - Allows apps to pool and share DB connections established with the database
    - Improving database efficiency by reducing the stress on database resources (e.g., CPU, RAM) and minimize open connections (and timeouts)
    - Reduced RDS & Aurora failover time by up 66%
    - Enforce IAM Authentication for DB, and securely store credentials in AWS Secrets Manager
    - RDS Proxy is never publicly accessible (must be accessed from VPC)

Amazon Aurora

- Supports both Postgres and MySQL, 5x performance improvement over MySQL on RDS and 3x performance improvement over Postgres on RDS
- Aurora costs 20% more than RDS
- can have 15 replicas while MySQL has 5 and the replication process is faster. Support CRR
- Failover is instantaneous. High Availability native
- Aurora storage automatically grows in increments of 10GB, up to 128 TB.
- Read and write mode:
    - client talks to a consistent Writer Endpoint, which points to the current Master DB.
    - client also talks to a Reader Endpoint, which balances the connection to a RR
    - Master and Read Replicas share the same storage Volume that supports
        - Replication + Self Healing + Auto Expanding
- Custom Endpoints
    - define a subset of RR as a Custom Endpoint
    - use case: Run analytical queries on specific replicas like larger ones
    - The Reader Endpoint is generally not used after defining Custom Endpoints
- Aurora Serverless
    - Automated database instantiation and auto-scaling based on actual usage

    - Good for infrequent, intermittent or unpredictable workloads
    - No capacity planning needed, no particular instance type required
    - Pay per second, can be more cost-effective
- Aurora Multi-Master
    - Every node does both R/W for high HA
- Global Aurora
    - Aurora Cross Region Read Replicas:
        - Useful for disaster recovery
        - Simple to put in place
    - Aurora Global Database (recommended):
        - 1 Primary Region (read / write)
        - Up to 5 secondary (read-only) regions, replication lag is less than 1 second and up to 16 Read Replicas per secondary region
        - Helps for decreasing latency
    - Promoting another region (for disaster recovery) has an RTO of < 1 minute
    - Typical cross-region replication takes less than 1 second
- Aurora Machine Learning
    - Enables you to add ML-based predictions to your applications via SQL
    - Supported services
        - Amazon SageMaker (use with any ML model)
        - Amazon Comprehend (for sentiment analysis)
    - Use cases: fraud detection, ads targeting, sentiment analysis, product recommendations
- Aurora Backups
    - Automated backups
        - 1 to 35 days (cannot be disabled)
        - point-in-time recovery in that timeframe
    - Manual DB Snapshots
        - Manually triggered by the user
        - Retention of backup for as long as you want
- Aurora DB Cloning
    - Create a new Aurora DB Cluster from an existing one, faster than snapshot & restore
    - Very fast & cost-effective
    - Useful to create a “staging” database from a “production” database without impacting the production database

RDS & Aurora Restore options

- Restoring a RDS / Aurora backup or a snapshot creates a new database
- Restoring MySQL RDS database from S3
    - Create a backup of your on-premises database
    - Store it on Amazon S3 (object storage)
    - Restore the backup file onto a new RDS instance running MySQL
- Restoring MySQL Aurora cluster from S3
    - Create a backup of your on-premises database using Percona XtraBackup
    - Store the backup file on Amazon S3
    - Restore the backup file onto a new Aurora cluster running MySQL

RDS & Aurora Security

- At-rest encryption:

    - Database master & replicas encryption using AWS KMS – must be defined as launch time
    - If the master is not encrypted, the read replicas cannot be encrypted
    - To encrypt an un-encrypted database, go through a DB snapshot & restore as encrypted
- In-flight encryption:
    - TLS-ready by default, use the AWS TLS root certificates client-side
- IAM Authentication: IAM roles to connect to your database (instead of username/password)
- Security Groups: Control Network access to your RDS / Aurora DB
- No SSH available except on RDS Custom
- Audit Logs can be enabled and sent to CloudWatch Logs for longer retention

Amazon ElastiCache

- managed Redis or Memcached
- Caches are in-memory databases with really high performance, low latency
- Helps reduce load off of databases for read intensive workloads by DB cache
- Helps make your application stateless by User Session Store
- AWS takes care of OS maintenance / patching, optimizations, setup, configuration, monitoring, failure recovery and backups
- Using ElastiCache involves heavy application code changes
- Redis VS Memcached
    - Redis: Multi AZ + Read Replicas + Backup and restore = persistent data
    - Memcached: no replication, no backup and restore, data store in shrading, multi-thread

Scalability & High Availability

- Scalability
    - an application/system can handle greater loads by adapting
    - Vertical Scalability means increasing the size of the instance (hardware limit) - common in DB
    - Horizontal Scalability means adding more instances - common in Auto Scaling Group
- High Availability
    - running your application / system in at least 2 data centers (== Availability Zones)

AWS Machine Learning

- Recognition
    - Find objects, people, text, scenes in images and videos
    - use case: face detection, labeling, celebrity recognition
- Transcribe: audio to text (ex: subtitles)
- Polly: text to audio
    - Customize the pronunciation of words with Pronunciation lexicons
        - ■ Stylized words: St3ph4ne => “Stephane”
        - ■ Acronyms: AWS => “Amazon Web Services”
    - Upload the lexicons and use them in the SynthesizeSpeech operation •
    - Generate speech from plain text or from documents marked up with Speech Synthesis Markup Language (SSML) – enables more customization
- Translate: translations
- Lex: build conversational bots – chatbots
    - Automatic Speech Recognition (ASR) to convert speech to text like Alexa
- Connect: cloud contact center
- Comprehend: natural language processing
- SageMaker: machine learning for every developer and data scientist

- Forecast: build highly accurate forecasts
- Kendra: ML-powered search engine
    - Fully managed document search service powered by Machine Learning
- Personalize: real-time personalized recommendations
- Textract: detect text and data in documents

Other Services:

- CloudFormation
    - a declarative way(Infrastructure as code) of outlining your AWS Infrastructure, for any resources (most of them are supported)
- Simple Email Service
    - Fully managed service to send emails securely, globally and at scale
- PinPoint
    - Scalable 2-way (outbound/inbound) marketing communications service
    - Supports email, SMS, push, voice, and in-app messaging
- Systems Manager
    - Session Manager
        - Allows you to start a secure shell on your EC2 and on-premises servers
        - No SSH access, bastion hosts, or SSH keys needed
        - No port 22 needed (better security)
        - Supports Linux, macOS, and Windows
        - Send session log data to S3 or CloudWatch Logs
    - Run Command
        - Execute a document (= script) or just run a command
        - Run command across multiple instances (using resource groups)
        - No need for SSH
    - Patch Manager
        - Automates the process of patching managed instances
        - OS updates, applications updates, security updates
        - Supports EC2 instances and on-premises servers
    - Maintenance Windows
        - Defines a schedule for when to perform actions on your instances
        - Example: OS patching, updating drivers, installing software
    - Automation
        - Simplifies common maintenance and deployment tasks of EC2 instances and other AWS resources
        - Examples: restart instances, create an AMI, EBS snapshot
- Cost Explorer
    - Visualize, understand, and manage your AWS costs and usage over time
    - Create custom reports that analyze cost and usage data
    - Analyze your data at a high level: total costs and usage across all accounts
    - Or Monthly, hourly, resource level granularity
    - Choose an optimal Savings Plan (to lower prices on your bill)
    - Forecast usage up to 12 months based on previous usage
- Amazon AppFlow
    - Fully managed integration service that enables you to securely transfer data between Software-as-a-Service (SaaS) applications and AWS

        - Sources: Salesforce, SAP, Zendesk, Slack, and ServiceNow
        - Destinations: AWS services like Amazon S3, Amazon Redshift or nonAWS such as SnowFlake and Salesforce

Elastic Load Balancer

- An Elastic Load Balancer is a managed load balancer
- AWS guarantees that it will be working
- AWS takes care of upgrades, maintenance, high availability
- AWS provides only a few configuration knobs
- Health Checks
    - enable the load balancer to know if instances it forwards traffic to are available to reply to requests
    - The health check is done on a port and a route (/health is common)
    - If the response is not 200 (OK), then the instance is unhealthy
- Classic Load Balancer (v1, deprecated)
    - Supports TCP (Layer 4), HTTP & HTTPS (Layer 7)
    - Health checks are TCP or HTTP based
    - Fixed hostname XXX.region.elb.amazonaws.com
    - we'd need multiple Classic Load Balancer per application
- Application Load Balancer (v2)
    - Application Layer 7 (HTTP)
    - Load balancing to multiple HTTP applications across machines (target groups)
        - Routing based on path in URL (example.com/users & example.com/posts)
        - Routing based on hostname in URL (one.example.com & other.example.com)
        - Routing based on Query String, Headers (example.com/users?id=123&order=false)
    - Load balancing to multiple applications on the same machine (ex: containers)
        - ALB are a great fit for micro services & container-based application
        - Has a port mapping feature to redirect to a dynamic port in ECS
    - Support for HTTP/S and WebSocket
    - Support redirects (from HTTP to HTTPS for example)
    - Target Groups
        - EC2 instances (can be managed by an Auto Scaling Group) – HTTP
        - ECS tasks (managed by ECS itself) – HTTP
        - Lambda functions – HTTP request is translated into a JSON event
        - IP Addresses – must be private IPs
    - ALB can route to multiple target groups and Health checks are at the target group level
    - Fixed hostname (XXX.region.elb.amazonaws.com)
    - The application servers don't see the IP of the client directly
        - The true IP of the client is inserted in the header X-Forwarded-For. We can also get Port (X-Forwarded-Port) and proto (X-Forwarded-Proto)
        - When receive a connection from a client, the ALB would terminate the connection and build a new connection with the target group using the Load Balancer IP(Private IP)
- Network Load Balancer (v2)
    - Support transport Layer(Layer 4) and forward TCP & UDP traffic to your instances
    - Handle millions of request per seconds with less latency ~100 ms (vs 400 ms for ALB)
    - NLB has one static IP per AZ, and supports assigning Elastic IP (helpful for whitelisting specific IP)
    - Target groups

        - EC2 instances
        - IP Addresses – must be private IPs
        - Application Load Balancer
    - Health Checks support the TCP, HTTP and HTTPS Protocols
- Gateway Load Balancer
    - Operates at Layer 3 (Network Layer) – IP Packets
    - Deploy, scale, and manage a fleet of 3rd party network virtual appliances in AWS
        - When receiving a connection from a client, the GLB would route traffic to a series of third-party Security Virtual Appliances to inspect or filter the traffic, then GLB receives the response traffic and directs it to Application.
    - Combines the following functions:
        - Transparent Network Gateway – single entry/exit for all traffic
        - Load Balancer – distributes traffic to your virtual appliances
    - Uses the GENEVE protocol on port 6081
    - Target Groups
        - EC2 instances
        - IP Addresses – must be private IPs
- Sticky Sessions(Session Affinity)
    - It is possible to implement stickiness so that the same client is always redirected to the same instance behind a load balancer
    - This works for Classic Load Balancers & Application Load Balancers
    - Use case: make sure the user doesn't lose his session data
    - Enabling stickiness may bring imbalance to the load over the backend EC2 instances
        - E.g. A client sends millions of requests but only one instance would receive them
- Cross-Zone Load Balancing
    - each load balancer instance distributes evenly across all registered instances in all AZ
        - E.g. AZ1 has 2 instances, AZ2 has 8 instances, all instances receive 10% traffic
    - Application Load Balancer
        - Always on (can't be disabled)
        - No charges for inter AZ data
    - Network Load Balancer
        - Disabled by default
        - You pay charges ($) for inter AZ data if enabled
    - Classic Load Balancer
        - Disabled by default
        - No charges for inter AZ data if enabled
- SSL – Server Name Indication (SNI)
    - SNI solves the problem of loading multiple SSL certificates on one web server (to serve multiple websites)
    - It's a "newer" protocol, and requires the client to indicate the hostname of the target server in the initial SSL handshake, the server will then find the correct certificate, or return the default one
    - Only works for ALB & NLB (newer generation), CloudFront
- SSL Certificates
    - An SSL Certificate allows traffic between your clients and your load balancer to be encrypted in transit (in-flight encryption)
    - The load balancer uses an X.509 certificate (SSL/TLS server certificate)
    - You can manage certificates using ACM (AWS Certificate Manager)

    - You can create upload your own certificates alternatively
    - HTTPS listener:
        - You must specify a default certificate
        - You can add an optional list of certs to support multiple domains
        - Clients can use SNI (Server Name Indication) to specify the hostname they reach
        - Ability to specify a security policy to support older versions of SSL / TLS (legacy clients)
    - Classic Load Balancer (v1)
        - Support only one SSL certificate
        - Must use multiple CLB for multiple hostname with multiple SSL certificates
    - Application Load Balancer (v2) and Network Load Balancer (v2)
        - Supports multiple listeners with multiple SSL certificates
        - Uses Server Name Indication (SNI) to make it work
- Connection Draining(CLB) or Deregistration Delay(ALB & NLB)
    - Time to complete “in-flight requests” while the instance is de-registering or unhealthy
    - Stops sending new requests to the EC2 instance which is de-registering
    - Between 1 to 3600 seconds (default: 300 seconds) Can be disabled (set value to 0)

Auto Scaling Group

- ASG are free (you only pay for the underlying EC2 instances)
- The goal of an Auto Scaling Group (ASG) is to:
    - Scale out (add EC2 instances) to match an increased load
    - Scale in (remove EC2 instances) to match a decreased load
    - Ensure we have a minimum and a maximum number of EC2 instances running
    - Automatically register new instances to a load balancer
    - Re-create an EC2 instance in case a previous one is terminated (ex: if unhealthy)
- Launch Template(Launch Configurations was deprecated) specifies
    - attributes of EC2 instances
    - Min Size / Max Size / Initial Capacity
    - Scaling Policies
- It is possible to scale an ASG based on CloudWatch alarms, metrics such as Average CPU are computed for the overall ASG instances
- Dynamic Scaling Policies
    - Target Tracking Scaling
        - Most simple and easy to set-up
        - Example: I want the average ASG CPU to stay at around 40%
    - Simple / Step Scaling
        - When a CloudWatch alarm is triggered (example CPU > 70%), then add 2 units
        - When a CloudWatch alarm is triggered (example CPU < 30%), then remove 1
    - Scheduled Actions
        - Anticipate a scaling based on known usage patterns
        - Example: increase the min capacity to 10 at 5 pm on Fridays
- Predictive Scaling
    - continuously forecast load and schedule scaling ahead
- Scaling Cooldowns
    - After a scaling activity happens, you are in the cooldown period (default 300 seconds)

Route 53

- Domain Name System(DNS) translates the human friendly hostnames into the machine IP addresses
  - www.google.com => 172.217.18.36
- Route 53 is a highly available, scalable, fully managed and Authoritative DNS
  - Authoritative = the customer (you) can update the DNS records
- Ability to check the health of your resources
- Records
  - controls how you want to route traffic for a domain
  - contains:
    - Domain/subdomain Name – e.g., example.com
    - Record Type – e.g., A or AAAA
      - A – maps a hostname to IPv4
      - AAAA – maps a hostname to IPv6
      - CNAME – maps a hostname to another hostname
      - NS – Name Servers for the Hosted Zone
    - Value – e.g., 12.34.56.78
    - Routing Policy – how Route 53 responds to queries
    - TTL(Time To Live) – amount of time the record cached at DNS Resolvers
- Hosted Zones
  - A container for records that define how to route traffic to a domain and its subdomains
  - public hosted zone
    - contains records that specify how to route traffic on the Internet
  - private hosted zone
    - contain records that specify how you route traffic within one or more VPCs
  - charges $0.5 per month per hosted zone
- Alias Records
  - Maps a hostname to an AWS resource
  - Alias Record is always of type A/AAAA for AWS resources (IPv4 / IPv6)
  - You can't set the TTL
  - You cannot set an ALIAS record for an EC2 DNS name
- Health Checks
  - HTTP Health Checks are only for public resources
  - Health Check => Automated DNS Failover
  - Health Checks are integrated with CloudWatch metrics
  - Configure your router/firewall to allow incoming requests from Route 53 Health Checkers
  - Calculated Health Checks
    - Combine the results of multiple Health Checks into a single Health Check
    - You can use OR, AND, or NOT
    - Can monitor up to 256 Child Health Checks
    - Specify how many of the health checks need to pass to make the parent pass
  - Health Checkers are outside the VPC, to health check private endpoints
    - You can create a CloudWatch Metric and associate a CloudWatch Alarm, then create a Health Check that checks the alarm itself
- Routing Policies
  - Defines how Route53 responds to DNS queries
  - Simple
    - return all records for a single resource
    - if multiple values of the same record are returned, the client will choose a random one
    - can't be associated with Health Checks

    - Weighted
        - control traffic by weight
        - can be associated with Health Checks
        - Assign a weight of 0 to a record to stop sending traffic to a resource • If all records have we
    - Failover
        - When an active instance failed the health check, the standby instance will failover and become active
    - Latency based
        - Redirect to the resource that has the least latency close to user
        - Latency is based on traffic between users and AWS Regions
        - Can be associated with Health Checks
    - Geolocation
        - This routing is based on user location
        - Can be associated with Health Checks
    - Multi-Value Answer
        - Use when routing traffic to multiple resources
        - Each resource receives a separate DNS response, and Route 53 responds to DNS queries with multiple IP addresses.
        - Can be associated with Health Checks (return only values for healthy resources)
    - Geoproximity
        - Route traffic to your resources based on the geographic location of users and resources
        - Ability to shift more traffic to resources based on the defined bias

Virtual Private Cloud(VPC)

- Classless Inter-Domain Routing(CIDR)
    - a method for allocating IP addresses
    - used in Security Groups rules and AWS networking in general
    - A CIDR consists of two components
        - Base IP: Represents an IP contained in the range (XX.XX.XX.XX)
            - Example: 10.0.0.0, 192.168.0.0, …
        - Subnet Mask: Defines how many bits can change in the IP
            - Example: /0, /24, /32
    - Up to 5 CIDER in one VPC
    - Your VPC CIDR should NOT overlap with your other networks
- Subnet (IPv4)
    - AWS reserves 5 IP addresses (first 4 & last 1) in each subnet
    - Example: if you need 29 IP addresses, use subnet mask /26(since 2^6 = 64, 64 - 5 > 29)
- Internet Gateway(IGW)
    - Allows resources (e.g., EC2 instances) in a VPC connect to the Internet
    - It scales horizontally and is highly available and redundant
    - Must be created separately from a VPC
    - One VPC can only be attached to one IGW and vice versa
- Bastion Hosts
    - We can use a Bastion Host(in a public subnet) to SSH into our private EC2 instances
    - Bastion Host security group must allow inbound from the internet on port 22 from restricted CIDR, for example the public CIDR of your corporation

    - Security Group of the EC2 Instances must allow the Security Group of the Bastion Host, or the private IP of the Bastion host
- NAT Instances
    - gives Internet access to EC2 instances in private subnets.
    - Old, must be setup in a public subnet, disable Source / Destination check flag
- NAT Gateway
    - managed by AWS, provides scalable Internet access to private EC2 instances, IPv4 only
- NAT Gateway > NAT Instance(since higher performance, higher availability, fully managed)
- Security Groups
    - are stateful, which means if in traffic was allowed, out traffic will be automatically allowed
- Network Access Control List(NACL)
    - are stateless, which means in and out both need to be checked
    - rules have a number between 1 and 32766, the lower of the number, the higher the priority
    - Default NACL
        - Accepts everything inbound/outbound with the subnets it's associated with
        - Do NOT modify the Default NACL, instead create custom NACLs
- Ephemeral Ports
    - For any two endpoints to establish a connection, they must use ports
    - Clients connect to a defined port, and expect a response on this port
- VPC Peering
    - Privately connect two VPCs using AWS' network
    - not transitive(means if A and B has VPC Peering, B and C has VPC Peering, A and C don't)
    - You must update route tables in each VPC's subnets to ensure EC2 instances can communicate with each other
    - You can create VPC Peering connection between VPCs in different AWS accounts/regions
    - You can reference a security group in a peered VPC (works cross accounts – same region)
- VPC Endpoints
    - Every AWS service is publicly exposed (public URL)
    - VPC Endpoints (powered by AWS PrivateLink) allow you to connect to AWS services using a private network instead of using the public Internet
    - They're redundant and scale horizontally
    - Types:
        - Interface Endpoints (powered by PrivateLink)
            - Provisions an ENI (private IP address) as an entry point (must attach a Security Group)
            - Supports most AWS services
            - $ per hour + $ per GB of data processed
        - Gateway Endpoints
            - Provisions a gateway and must be used as a target in a route table (does not use security groups)
            - Supports both S3 and DynamoDB
            - Free
        - Gateway or Interface Endpoint for S3?
            - Gateway is most likely going to be preferred all the time at the exam
            - Unless, access is required from on premises (Site to Site VPN or Direct Connect), a different VPC or a different region
- VPC Flow Logs
    - Capture information about IP traffic going into your interfaces

    - Helps to monitor & troubleshoot connectivity issues
    - Flow logs data can go to S3 / CloudWatch Logs
    - Query VPC flow logs using Athena on S3 or CloudWatch Logs Insights
- Site-to-Site VPN
    - Virtual Private Gateway (VGW)
        - VPN concentrator on the AWS side of the VPN connection
        - VGW is created and attached to the VPC from which you want to create the Site-to-Site VPN connection
    - Customer Gateway (CGW)
        - Software application or physical device on customer side of the VPN connection
    - Site-to-Site VPN ECMP
        - ECMP = Equal-cost multi-path routing
        - Routing strategy to allow to forward a packet over multiple best path
        - Use case: create multiple Site-To-Site VPN connections to increase the bandwidth of your connection to AWS
- AWS VPN CloudHub
    - Provide secure communication between multiple sites, if you have multiple VPN connections
    - Low-cost hub-and-spoke model for primary or secondary network connectivity between different locations (VPN only)
    - It's a VPN connection so it goes over the public Internet
- Direct Connect (DX)
    - Provides a dedicated private connection from a remote network to your VPC
    - Dedicated connection must be setup between your Data Center and AWS Direct Connect locations
    - Access public resources (S3) and private (EC2) on same connection
    - Supports both IPv4 and IPv6
    - Direct Connect Gateway
        - If you want to set up a Direct Connect to one or more VPC in many different regions (same account), you must use a Direct Connect Gateway
    - takes more than 1 month to build a new connection
    - data in transit is not encrypted but is private
        - AWS Direct Connect + VPN provides an IPsec-encrypted private connection
    - In case Direct Connect fails, you can set up a backup Direct Connect connection (expensive), or a Site-to-Site VPN connection
- Transit Gateway
    - For having transitive peering between thousands of VPC and on-premises, hub-and-spoke (star) connection
    - Regional resource, can work cross-region
    - Share cross-account contents using Resource Access Manager (RAM)
- Traffic Mirroring
    - Allows you to capture and inspect network traffic in your VPC
    - Route the traffic to security appliances that you manage
    - Source and Target can be in the same VPC or different VPCs (VPC Peering)
    - Use cases: content inspection, threat monitoring, troubleshooting, …

Security Group vs. NACLs:

- Security Group
    - instance level

    - supports allow rules only
    - stateful
    - all rules are evaluated
    - applies to a single instance
- NACLs:
    - subnet level
    - supports allow and deny rules
    - stateless(return traffic must be checked as well)
    - rules are evaluated in order and first match wins
    - applies to all instances in a subnet

IAM credential reports & IAM access Advisor

- IAM credential reports(account-level)
    - lists all accounts' users and the status of their various credentials

- IAM access Advisor(user-level)
    - shows the service permissions granted to a user and when those services were last accessed. This information can be used to revise policies

Important ports:

- FTP: 21
- SSH: 22
- SFTP: 22 (same as SSH)
- HTTP: 80
- HTTPS: 443

RDS Databases ports:

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

Simple SQL query - Athena
Complex SQL query - Redshift

AWS Glue provides both visual and code-based interfaces to make data integration easier. Users can easily find and access data using the AWS Glue Data Catalog. Data engineers and ETL (extract, transform, and load) developers can visually create, run, and monitor ETL workflows with a few clicks in AWS Glue Studio.

VPC endpoint allows you to connect to AWS services using a private network instead of using the public Internet

To implement password rotation lifecycles, use AWS Secrets Manager. You can rotate, manage, and retrieve database credentials, API keys, and other secrets throughout their lifecycle using Secrets Manager.

Amazon Cognito is for authentication, authorization and user management

To remediate Cross-site scripting(XSS) attacks, we need WAF(web application firewall)

DDoS - Shield - Load Balancer, CloudFront, Route53
XSS and SQL Inject - WAF - CloudFront, ALB, API Gateway

S3 Glacier Instant 90 days
S3 Standard IA 30 days
S3 One Zone IA 30 days

A route table contains a set of rules, called routes, that are used to determine where network traffic from your subnet or gateway is directed. The route table in the instance's subnet should have a route defined to the Internet Gateway.

DynamoDB Accelerator (DAX) - Amazon DynamoDB Accelerator (DAX) is a fully managed, highly available, in-memory cache for DynamoDB that delivers up to a 10x performance improvement

AM users or AWS services can assume a role to obtain temporary security credentials that can be used to make AWS API calls.

Amazon FSx for Lustre makes it easy and cost-effective to launch and run the world's most popular high-performance file system. It is used for workloads such as machine learning, high-performance computing (HPC), video processing, and financial modeling. The open-source Lustre file system is designed for applications that require fast storage

when the new AMI is copied from region A into region B, it also creates a snapshot in region B because AMIs are based on the underlying snapshots.

Spot blocks can only be used for a span of up to 6 hours,

Lambda can only work for 15 mins

API Gateway supports stateless RESTful APIs and stateful WebSocket APIs.

As each Snowball Edge Storage Optimized device can handle 80TB of data,
Each Snowmobile has a total capacity of up to 100 petabytes.

You can't transition from the following:

1. Any storage class to the S3 Standard storage class.
2. Any storage class to the Reduced Redundancy Storage (RRS) class.
3. The S3 Intelligent-Tiering storage class to the S3 Standard-IA storage class.

4. The S3 One Zone-IA storage class to the S3 Intelligent-Tiering, S3 Standard-IA, or S3 Glacier Instant Retrieval storage classes.

You can connect to Amazon EFS file systems from EC2 instances in other AWS regions using an inter-region VPC peering connection, and from on-premises servers using an AWS VPN connection.

In the event of a failover, Amazon Aurora will promote the Read Replica that has the highest priority (the lowest numbered tier). If two or more Aurora Replicas share the same priority, then Amazon RDS promotes the replica that is largest in size. If two or more Aurora Replicas share the same priority and size, then Amazon Aurora promotes an arbitrary replica in the same promotion tier.

To delete a CMK in AWS KMS you schedule key deletion. You can set the waiting period from a minimum of 7 days up to a maximum of 30 days. The default waiting period is 30 days.

the minimum storage duration is 30 days before you can transition objects from S3 Standard to S3 One Zone-IA or S3 Standard-IA,

ECS with EC2 launch type is charged based on EC2 instances and EBS volumes used. ECS with Fargate launch type is charged based on vCPU and memory resources that the containerized application requests

In a stopped RDS database, you will still pay for storage. If you plan on stopping it for a long time, you should snapshot & restore instead

AWS Transfer Family securely scales your recurring business-to-business file transfers to AWS Storage services using SFTP, FTPS, FTP, and AS2 protocols.

Manage fine-grained access control using AWS Lake Formation

in most cases NAT instances are a bad thing because they are customer managed while NAT gateways are AWS Managed.

Only Network Load Balancer provides both static DNS name and static IP. Application Load Balancer provides a static DNS name but it does NOT provide a static IP. The reason being that AWS wants your Elastic Load Balancer to be accessible using a static endpoint, even if the underlying infrastructure that AWS manages changes.

DynamoDB Streams enable DynamoDB to get a changelog and use that changelog to replicate data across replica tables in other AWS Regions.

When using an Application Load Balancer to distribute traffic to your EC2 instances, the IP address you'll receive requests from will be the ALB's private IP addresses. To get the client's IP address, ALB adds an additional header called "X-Forwarded-For" contains the client's IP address.

Network Load Balancer has one static IP address per AZ and you can attach an Elastic IP address to it. Application Load Balancers and Classic Load Balancers have a static DNS name.

the NLB supports HTTP health checks as well as TCP and HTTPS

The following cookie names are reserved by the ELB (AWSALB, AWSALBAPP, AWSALBTG).

structured data means no sql

FSx File Gateway only provides S3 or EFS storage, neither of which offer both NFS and SMB access.

Amazon FSx for NetApp ONTAP provides highly performant file storage that is accessible via both NFS and SMB.

A subnet must reside within a single Availability Zone.

AWS DataSync is a data transfer service that can copy large amounts of data between on-premises storage and Amazon FSx for Windows File Server at high speeds. It allows you to control the amount of bandwidth used during data transfer.

DataSync uses agents at the source and destination to automatically copy files and file metadata over the network. This optimizes the data transfer and minimizes the impact on your network bandwidth.

DataSync allows you to schedule data transfers and configure transfer rates to suit your needs. You can transfer 30 TB within 5 days while controlling bandwidth usage.

DataSync can resume interrupted transfers and validate data to ensure integrity. It provides detailed monitoring and reporting on the progress and performance of data transfers.

S3 Storage Lens is a fully managed S3 storage analytics solution that provides a comprehensive view of object storage usage, activity trends, and recommendations to optimize costs. Storage Lens allows you to analyze object access patterns across all of your S3 buckets and generate detailed metrics and reports.

identity-based policy used for role and group

You can only have one VPC Peering per VPC pair.

You can attach one virtual private gateway to a VPC at a time