# AWS SAA Notes

## Regiones

Para elegir una reunión hay que pensar en si se debe cumplir con temas de compliance, la proximidad de la región con los usuarios, que los servicios que se vayan a usar estén disponibles en la región y los precios.

Las regiones tienen mínimo 3 y máximo 6 AZ. Las AZ's cuentan mínimo con un data center y están aisladas una de la otra.

## IAM

Los grupos solo pueden contener usuarios, no otros grupos.
Un usuario puede pertenecer a más de un grupo.
A los usuarios y grupos se les asigna un JSON que contiene las políticas de permisos.
Una in line policy es una política que solo se le asigna a un solo usuario.
Eliminar permisos de la cuenta root no es posible porque la cuenta root tiene todos los permisos de manera predeterminada.
IAM consta de Sid, Effect, Principal, Action, Resource y Condition.

### Password Policy

Se puede obligar a que las contraseñas tengan ciertas características específicas.
Se puede habilitar la opción para que los usuarios puedan cambiar sus contraseñas.
Se puede configurar un password expiration.
Se puede configurar un prevent password re-use.

### IAM Roles

Los IAM Roles solo existen para los servicios de AWS.
Los IAM Roles tienen una política de permiso asignada que le permite a un servicio hacer cosas con otro servicio.

### IAM Security Tools

IAM Credentials Report es un reporte de toda la cuenta que lista los usuarios y el estado de sus credenciales.
IAM Access Advisor es a nivel de usuario y muestra los permisos que tiene y la última vez que se usaron. Puede usarse para ver si ya un usuario no necesita un permiso y modificar su política de permisos.

### Organizations

Una sola bill
Manejar múltiples cuentas
Beneficios en precios de servicios
Se comparten las instancias reservadas y Saving Plans entre todas las cuentas

**Ventajas:**

- Más seguro
- Tags para billing
- CloudTrail en todas las cuentas y mandar logs a la cuenta central de S3
- CloudWatch Logs a cuenta central
- Service Control Policies:
    - IAM para OU o Cuentas para restringir a usuarios y roles
    - Hay que definir allow para todo lo que se quiera tener acceso

**Identity Center**

SSO para todas las cuentas en Organizations y otras aplicaciones en la nube siempre y cuando se tenga SAML 2.0
EC2 Windows Instances
Multi-account permissions
Attribute Based Access Control

**Control Tower**

Easy way to govern a secure and compliant multi-account AWS environment basado en las mejores prácticas
Preventive Guardrail con SCPs
Detective Guardrail con AWS Config

## EC2

General purpose (M, T):

- Cargas de trabajo, servidor web o repo
- Buen balance de red, computo y memoria

Compute optimized (C):

- Cargas que necesiten buen procesador
- Batches; Media transcoding; High performance web servers; Scientific modeling y ML; Gaming servers

Memory optimized (R, U, X, Z):

- Cargas que corren grandes data sets en la memoria
- High performance DBs; In-memory DB for BI; Apps con procesamiento de datos en tiempo real de datos no estructurados

Storage optimized (D, I):

- Cargas con números grandes de escritura y lectura secuenciales en almacenamiento local
- OLTP systems; DBs relacionales y no-relacionales; Redis; Data warehousing apps; Distributed file systems

Accelerating computing (DL, F, G, Inf, P, Trn, VT):

- Cargas que necesiten aceleradores de hardware
- Cálculos matemáticos; procesamiento de gráficos; funciones; patrones de datos

High-performance computing (HPC):

- Best price performance for running HPC workloads at scale on AWS
- Deep learning; simulaciones complejas

**Security Groups**

Están sujetos a la región en que se crean
Vive afuera de la EC2. Si el tráfico es bloqueado, la EC2 no ve el tráfico

**Purchasing Options**
**On-demand**

Short workload
Pagas por lo que usas en segundos (Linux, Windows)
Pagas por hora para todos los demás
Es el más caro pero no necesitas pagar por adelantado
No hay contratos de largo plazo
Ideal para cargas de trabajo cortas y continuas

**Reserved (1-3 años)**

Hasta +- 70% de descuento en relación con las on-demand
Cuando se reserva la instancia se debe especificar: tipo, región, tenancy y OS
El descuento es mayor si se reserva la instancia por más tiempo y dependiendo del método de pago, si se paga todo en seguida hay más descuento

Un buen uso de estas reservas es una base datos
Puedes comprar y vender instancias reservadas en el Marketplace

Convertible Reserved Instances te deja cambiar las especificaciones, pero el descuento es menor (+-66%)

**Saving plans (1-3 años)**

Hasta +- 70% de descuento
Aquí lo que uno especifica es la cantidad de dinero que se va a gastar uno por hora por 1 o 3 años
Si te pasas de lo pactado, se te factura como on-demand
Está sujeto a la región y familia de la instancia que se pacte, o sea, puedes variar el tamaño, OS y tenancy

**Spot Instances**

Hasta +- 90% de descuento en relación las on-demand
Las puedes perder en cualquier momento si el precio por el que estás pagando la instancia es inferior al precio real del spot
Es la más **cost-efficient**
Ideal para cargas de trabajo como batches, análisis de datos, procesamiento de imágenes, cargas de trabajo distribuidas y cargas de trabajo con un flex start y end time
No cargas críticas ni bases de datos
Si el precio máximo que definiste a pagar es superado por el precio spot, tienes 2 minutos para parar o matar las instancias
Solo se puede cancelar requests de Spot Instances si está abierto, activo o deshabilitado
Cancelar un request no mata las instancias
Spot Fleet: nos permite hacer request de Spot Instances con el menor precio posible automáticamente

**Dedicated Host**

Un servidor de EC2s solo para ti
Ideal para compliance y si ya tienes licenciamiento de herramientas
Se puede on-demand o por reserved
Es la **más cara**

**Dedicated Instances**

Hardware solo para ti
Comparten hardware con otras instancias tuyas
No hay control de donde se levantan las instancias

**Capacity Reservations**

Reservar on-demand en una AZ específica
Disponibilidad siempre
No contrato, no descuento
Te cobran on-demand estés corriendo o no instancias porque la idea es que reservas la capacidad y la disponibilidad inmediata de EC2s
Para cargas de trabajo cortas que no deban ser interrumpidas y también necesiten correr en una AZ específica

**IP Charges**

$0.005 por hora por IPv4 pública

**Placement Groups**

Usado para controlar donde quieres que se levanten las instancias

Clusters: baja latencia en la misma AZ, mejor networking, 10GBps
Spread: distribuidas en diferentes AZs, están en hardware diferente. Limitado a 7 instancias por AZs por placement group
Partition: 7 particiones por AZ, múltiples AZs en la misma región, hasta 100 instancias. Instancias de una partición no comparten hardware con instancias en otras particiones

HDFS, HBase, Cassandra, Kafka

**ENI**

Componente lógico en la VPC que representa una tarjeta de red, es lo que le da acceso a la red a las instancias
Son independientes de las instancias y se puede mover de una a otra para failover
Están sujetas a AZ

**EC2 Hibernate**

El estado de la RAM es preservado porque se escribe en un archivo del root EBS, para esto el EBS tiene que estar encriptado
La instancia inicia más rápido porque el OS no se reinicia
La EC2 no puede hibernar por más de 60 días

**Elastic Block Storage Volume**

Network drive que se puede montar en instancias mientras corren para persistir data, incluso si la instancia es terminada. Es como una USB
Solo se puede montar a una sola instancia
Está sujeta a la AZ, para moverlo de una AZ a otra se puede hacer un snapshot para usarlo en la otra AZ
Tienen un poquito de latencia porque es un dispositivo de red
Capacidad en tamaño (GB) e IOPS, se puede modificar la capacidad en el tiempo

**EBS Snapshots**

Es cross-AZ
Se puede mover un Snapshot a Snapshot Archive, es 75% más barato
Tarda 24-72 para restaurar
Existe el recycle bin para poder recuperar un Snapshot si es eliminado por accidente y se puede configurar para que retenga lo eliminado hasta un máximo de 1 año
Fast Snapshot Restore si se necesita inmediatamente la primera vez que se usa, es caro, pero útil si el snapshot es bastante pesado

**EBS Volume Types**

General purpose:

- Cost-effective, low-latency
- 1 GiB - 16 TiB
- gp3:
    - Base de 3000 IOPS y throughput de 125 MiB/s
    - Hasta 16000 IOPS y throughput de 1000 MiB/s, se puede aumentar uno de los dos o los dos
- gp2:
    - Hasta 3000 IOPS
    - El tamaño del disco e IOPS están corelacionados, entre más volumen más IOPS hasta 16000 IOPS

Provisioned IOPS:

- Apps que necesitan IOPS sostenido o mucho más de 16000
- Ideal para bases de datos
- Soporta EBS Multi-attach
- io1 (4 GiB - 16 TiB):
    - Máximo 64000 IOPS para Nitro EC2 y máximo 32000 para otros tipos
    - Se puede aumentar el IOPS independiente del tamaño del volumen
- io2 Block Express (4 GiB - 16 TiB):
    - Latencia en los sub milisegundos
    - Máximo 256000 IOPS; ratio 1000:1 IOPS:GiB

Hard Disk Drive:

- No se puede usar para bootear
- 125 GiB - 16 TiB
- st1:
    - Throughput optimized
    - Big Data, Data Warehouse, Log Processing
    - Max throughput 500 MiB/s, max IOPS 500
- sc1:
    - Cold HDD

- Para archivar data a la que no se accede tan frecuentemente
- Máximo 250 MiB/s, max IOPS 250

**EBS Multi-attach (io1 e io2)**

Montar el mismo EBS a múltiples EC2s en la misma AZ
Todas las instancias pueden leer y escribir con la capacidad del volumen
Aplicaciones que necesitas escritura concurrente
Hasta 16 instancias al mismo tiempo
El file system tiene que ser *cluster-aware*

**EBS Encryption**

Data encriptada at rest en el volumen
Data inflight entre la instancia y el volumen está encriptada
Todos los snapshots están encriptados
No se maneja nada del tema de la encriptación
La encriptación tiene un impacto mínimo en la latencia
Usa llaves del KMS (AES-256)
Se puede encriptar un volumen desencriptado

**Amazon EFS**

Network file system que se puede montar en múltiples EC2s
Se puede usar cross-az
Alta disponibilidad, escalable automáticamente hasta petabytes, **caro**
Pagas por lo que usas
Usado para manejar contenido, servidores web, data share, wordpress
Usa NFSv4.1 protocol
Usa security group para dar acceso al EFS
Solo es para AMI basadas en Linux
Encryption at rest con KMS
POSIX file system

**AMI**

Son imágenes de EC2 que están customizadas
Primero se construye en una región y luego puede ser copiada de una región a otra
Está el AMI Marketplace donde se consiguen AMIs hechas por otras personas

**Instance Store**

Tienen un disco físico conectado al servidor
Mejor para IOPS
Es efímero

## High Availability and Scalability: ELB y ASG

**Elastic Load Balancer**

Dividir carga entre instancias
Un solo punto de acceso a la aplicación (DNS)
Stickiness con cookies
HA

Separar tráfico público del privado
Tiene health checks para verificar si hay errores en las instancias

Existen 3 tipos:
- Application Load Balancer (Capa 7): HTTP, HTTPS, WebSocket
- Network Load Balancer (Capa 4): TCP, TLS, UDP
- Gateway Load Balancer (Capa 3): IP

**Application Load Balancer**

Usa target groups para balancer cargas
Balancea cargas que van a contenedores
Redirige HTTP a HTTPS
Ideal para aplicación basadas en microservicios y contenedores
La aplicación no ve directamente la IP del cliente

Routing tables a target groups diferentes:
- Routing basado en path "/ejemplo"
- Routing basado en hostname "one.ejemplo.com" "dos.ejemplo.com"
- Query string, headers

**Network Load Balancer**

Redirige y balancea tráfico TCP y UDP a las instancias
Millions requests per sec
Ultra-low latency
Una IP static por AZ
Redirige a IP privadas
Health checks en TCP, HTTP y HTTPS

**Gateway Load Balancer**

Despliega, escala y maneja aplicaciones de terceros en AWS
Usa el GENEVE protocol en el puerto 6081

**Sticky Sessions**

Sirve para que el cliente siempre vaya a la misma instancia detrás del balanceador de cargas
La cookie usada para la stickiness se expira

**Draining Connection**

Estado de una instancia que hace que los usuarios que ya estén conectados a esa instancia tengan un periodo para terminar su trabajo y luego son redirigidos a otra.
Los usuarios nuevos desde que la instancia está en draining son redirigidos a instancias que no estén en draining
El periodo por default es 300 segundos, pero se puede ajustar entre 1 y 3600 segundos. Si se pone en 0 segundos cuenta como desactivado
Mantenerlo en un valor bajo es bueno para requests cortas

**Auto-scaling Group**

Escala y desescala el número de instancias dependiendo de la carga

Asegura que tengamos un mínimo y máximo de instancias corriendo
Registra automáticamente nuevas instancias en los load balancers
Si hay una instancia unhealthy, levantará una nueva para reemplazarla
Se puede escalar basado en CloudWatch alarms

**ASG Scaling Policies**

Dynamic Scaling:
- Target Tracking Scaling:
  - Simple
  - Mencionas qué es lo que quieres como mantener el uso de CPU en 40%
- Simple Scaling
  - Cuando CPU mayor que 70%, agrega 2 EC2s
  - Cuando CPUT menor que 30%, mata 2 EC2s

Scheduled Scaling:
- Anticipación a un patrón de uso
- Yo sé que de 9 a 5 hay más personas usando la aplicación, agrego 2 EC2s en ese horario

Predictive Scaling: predice la carga y ajusta las instancias a la predicción

Scaling Cooldown es un periodo en el cual no se acciona ninguna regla de escalado sino hasta que se termine el mismo.
Por default es de 300 segundos

## RDS + Aurora + ElastiCache

Una RDS puede escalar automáticamente
Puedes poner un límite en el escalamiento en caso de no querer que escale infinito

**Read Replicas**

Hasta 15 read replicas
Misma AZ, cross-AZ o cross-region
La replicación es asíncrona
Cualquiera de las replicas puede ser designada como la principal para que sea esta la que escriba

El disaster recovery es la replicación síncrona de la base de datos a otra base de datos
La base de datos del failover y la que está en live están bajo el mismo DNS name

**Amazon Aurora**

Soporta PostgreSQL y MySQL
Desempeño 5 veces mejor que MySQL y 3 veces mejor que PostgreSQL
Incrementa automáticamente de 10 GB en 10 GB hasta 128 TB
Hasta 15 replicas
Failover instantáneo, HA nativo
Es 20% más caro que otros RDS
Hace 6 copias across 3 AZs
Self-healing
Cross-region replication

El writer endpoint te conecta a la instancia de base de datos que se encarga de escribir, que es la master
El reader endpoint te conecta a las read replicas y balancea la carga

Aurora Global Database:
- 1 Primary Region (W/R)
- Hasta 5 Secondary Regions (R), replicación es de menos de 1 sec
- Hasta 16 read replicas per Secondary Region
- RTO < 1 min
- Cross-region replication toma menos de 1 sec

Babelfish sirve para que una aplicación con motor de base de datos como MSSQL pueda usarse en Amazon Aurora que usa PostgreSQL y no se necesita cambiar nada

**RDS/Aurora Backup**

Para restaurar: hacer una copia de la DB, almacenarla en S3 y hacer el backup desde ese archivo
La diferencia es que para Aurora se debe usar Percona XtraBackup para hacer la copia de DB

Amazon Database Cloning, ideal para clonar una base de datos sin interrumpir, ni impactar, la que se esté clonando de form rápida y costo efectiva

**Amazon RDS Proxy**

Junta todas las conexiones de las bases de datos al proxy
Mejora la eficiencia de las bases de datos reduciendo el estrés directo a las mismas y minimiza las conexiones abiertas
Serverless, auto-scaling, HA, multi-AZ
Reduce el tiempo de failover en un 66%
Necesita autenticación IAM y guarda las credenciales en Secrets Manager
Es privado, no es accesible desde la red pública

**Amazon ElastiCache**

Redis:
- Multi-AZ
- Read replicas
- Data durability usando AOF
- Backup y restore
- Soporta sets y sorted sets
- IAM Auth

Memcached:
- Sharding, multi-nodo
- No HA
- No persistencia
- Backup y restore (serverless)
- Multithreaded architecture

## Route 53

Existen muchos tipos de records, pero los 4 principales son:
- A: mapea un hostname a una IPv4

- AAAA: mapea un hostname a una IPv6
- CNAME: mapea un hostname a otro hostname
    - el hostname a mapear debe ser de tipo A o AAAA
- NS: name servers for the Hosted Zone
    - controls how traffic is routed to a domain and its subdomains
    - public only contains records that specify how to route traffic on the internet
    - private only contains records that specify how to route traffic within one or more VPCs
    - cuesta $0.50 mensual por Hosted Zone

**Health Checks**

HTTP health checks son solo para recursos públicos
HTTP, HTTPS, TCP
Se debe configurar para permitir el trafico de los requests de los health checks

DNS Failover Automático:
- Health checks que monitorean un endpoint
- Health checks que monitorean otros health checks
- Health checks que monitoren alarmas de Cloud Watch (puede servir para recursos privados)

Calculated Health Checks:
- Un health check padre monitorea hasta 256 health checks
- Combina los resultados
- OR, AND, NOT
- Se puede especificar cuantos health checks deben pasar para que el padre pase

**TTL**

Guarda la respuesta del DNS en un cache para no tener que andar haciendo la petición para que resuelva la dirección
Para todos excepto Alias records es mandatorio setear un TTL

**CNAME**

Apunta un hostname a otro hostname
Solo sirve para dominios con prefijos, o sea xxx.google.com y no con google.com

**Alias**

Apunta un hostname a un recurso de AWS
Sirve para cualquier dominio
Gratis
Health check
Automáticamente reconoce si la IP del recurso cambia
Funciona para el top node de un DNS namespace
Es de tipo A o AAAA
No se puede setear un TTL
No se puede setear un Alias para un DNS de una EC2

**Routing Policies**

Simple:
- Routea tráfico a un recurso

- Múltiples valores en un solo record
- Si responden multiples valores, el cliente escoge uno al azar
- No health checks

Weighted:
- Controla el % de requests que van a cada recurso
- DNS records tienen que tener el mismo nombre y tipo
- Health checks
- Si 0 entonces no manda tráfico a ese recurso
- Si todos los recursos tienen 0, entonces la cantidad distribuida entre los recursos es la misma

Latency-based:
- Redirige al recurso más cerca del usuario
- La latencia es basada entre el usuario y las regiones
- Health checks
- Failover

Geolocation:
- Basado en la ubicación del usuario
- Por continente, país o estado en EE.UU.
- Health checks

Geoproximity:
- Basado en la ubicación de los usuarios y los recursos

IP-based routing:
- Basado en la IP de los clientes
- Provees CIDR y a dónde se quiere redirigir el tráfico de esas IPs

## Amazon S3

Nombre único globalmente
Los buckets son creados en regiones
El tamaño máximo de un archivo es de 5TB
Para archivos de más de 5GB se tiene que usar multi-part upload y recomendable para archivos de más de 100MB
Tags para seguridad y lifecycle
Version ID, si el versionado está activado

**Seguridad**

Usuario:
- Políticas IAM: cuales API calls son permitidas para un usuario IAM

Recurso:
- Política Bucket: reglas directamente al bucket, cross-account
- Object ACL: política directa al objeto
- Bucket ACL:

**Encriptación de S3**

Existen 4 métodos:

Server-Side Encryption:
- SSE con Amazon S3-Managed Keys - default:
    - Administrada por AWS y propiedad de AWS
    - AES-256
    - Header = "x-amz-server-side-encryption”:”AES256”
- SSE con KMS Keys en KMS:
    - Control sobre la key
    - Auditar su uso con CloudTrail
    - Header = "x-amz-server-side-encryption”:”aws:kms”
    - La API call desde S3 para usar la llave cuenta en las requests
- SSE con llave propia

Client-Side Encryption: la encriptación sucede del lado del cliente, el objeto se envía ya encriptado al bucket

**S3 Glacier Vault Lock**

Se crea un archivo y luego se pone en un Glacier con Vault Lock y solo se podrá leer, pero no modificar.
Cuando se crea una política de Vault Lock no se le va a poder hacer cambios después
Ideal para compliance y retención

**S3 Object Lock**

Debe estar activado el versionamiento de objetos
Funciona a nivel del objeto
Compliance mode: no se puede sobreescribir ni borrar el objeto y no se pueden modificar los periodos, ni modos de retención
Governance mode: la mayoría de usuarios no puede sobreescribir ni borrar el objeto, tampoco alterar los ajustes de lock y otros usuarios si pueden hacerlo
Retention period: proteger un objeto por un periodo de tiempo, puede ser extendido
Legal Hold: proteger indefinidamente un objeto, es independiente del periodo de retención

**S3 Access Points:** puertas para acceder a diferentes S3 buckets
**S3 VPC Origin:** acceso a un S3 bucket dentro de la VPC, necesita un VPC Endpoint
**S3 Object Lambda:** para modificar un objeto antes de que sea retrieved por quien lo esté llamando, necesita un S3 Access Point y un Object Lambda Access Point

**Clases**

Standard:
- 99.99% disponibilidad
- Data usada frecuentemente
- Baja latencia alto throughput
- Big data analytics, mobile y gaming apps, content distribution, etc

Infrequent Access:
- Frecuentemente menos usada, pero acceso rápido
- Más barato que standard
    - Standard - IA:

        - 99.9% disponibilidad
        - DR, backups
    - One Zone - IA:
        - 99.99999999999% durabilidad en una sola AZ
        - 99.5% disponibilidad
        - Backup secundario

Glacier:

- Barato
- Pagas por el almacenamiento y el retrieval
    - Instant Retrieval:
        - Milisecond retrieval
        - Mínimo almacenar por 90 días
    - Flexible Retrieval:
        - Expedited (1-5 min), Standard (3-5 horas), Bulk (5-12 horas,free)
        - Mínimo almacenar 90 días
    - Deep Archive:
        - Standard (12 horas), Bulk (48 horas)
        - Mínimo almacenar por 180 días

Intelligent Tiering: cobro mensual por optimización de localización objetos según el uso que le den

Requester-pays: quien hace el request al objeto en el bucket es quien paga por el tráfico

## CloudFront

Es un content delivery network
Mejora las lecturas y el contenido es cacheado en edge
Mejora la experiencia de usuario
DDoS protection, se puede integrar con Shield y WAF
Cachean por un TTL
Ideal para contenido estático

El origen del contenido puede venir de: S3, aplicaciones dentro de una VPC y custom origin por HTTP

**Price Classes**

100: Norte América
200: Norte América y Europa
All: todos los Edge Locations

**Global Accelerator**

Funciona para Elastic IP, EC2, ALB, NLB, públicos o privados
Baja latencia y failover rápido, evita saltos entre servidores para llegar al origen del servidor de la aplicación
Red interna de AWS
Health checks
DDoS protection por el Shield
Capa 4 (TCP/UDP)

## Storage

### AWS Snowball

Seguro, portable, procesar data at edge, migrar data a AWS o fuera de AWS
Petabytes
Se pueden correr EC2s y Lambdas
Para importar data de Snowball a Glacier debe ser a través de S3

### FSX

3rd party high performance file systems on AWS
Administrado totalmente por el usuario

### FSX Windows

Sistema de archivo de Windows administrado por el usuario
Soporta SMB protocol y NTFS
Microsoft AD integration, ACLs, user quotas
Se puede montar en instancias con Linux
Soporta Microsoft Distributed File System Namespaces para agrupar los file systems
Escala hasta 10 GB/s, millones de IOPS, 100s PB de datos
Almacenamiento en SSD y HDD
Se puede acceder desde on-premise con VPN o Direct Connect
Multi-AZ
Backup diario en S3

### FSX Lustre

File system paralelo para large-scale computing
Usado para ML y HPC
Video processing, financial modeling, electronic design automation
Escala hasta 100s GB/s, millones de IOPS, latencia sub-ms
Almacenamiento en SSD y HDD
Integración con S3 para leer y escribir archivos
Se puede acceder desde on-premise con VPN o Direct Connect

### Storage Gateway

Puente entre datos on-premise y en la nube
Ideal para DR, backup & restore
A veces en la nube está la data fría y on-premise la caliente o viceversa

### S3 File Gateway

Acceso a todos los tipos de bucket excepto Glacier
Se comunica con el servidor on-premise con NFS o SMB y el tráfico desde on-premise a nube es HTTPS
La data más usada se almacena en cache en el gateway
Permisos al gateway con IAM para acceso a buckets
SMB protocol se puede integrar con AD para autenticación

**Data Sync**
Mover data de/a:

- On-premise/otras nubes-AWS (necesita agente)
- AWS-AWS (no necesita agente)

Se puede sincronizar a:

- S3
- EFS
- FSx

Replicación puede ser programada por hora, diariamente o semanalmente
Los permisos de los archivos y metadatos

## SQS

Servicio de colas ideal para desacoplar servicios/aplicaciones
Throughput ilimitado, mensajes ilimitados en la cola
Por defecto un mensaje permanece 4 días en la cola, y puede permanecer hasta 14 días, o hasta que sea consumido y eliminado por el consumidor usando DeleteMessage API
Baja latencia (10ms o menos)
Máximo mensajes de 256KB
Puede tener mensajes duplicados
Puede tener mensajes desorganizados
**Polling** messages tiene capacidad de hacerlo hasta con 10 mensajes
Encryption in-flight con HTTPS API, at-rest con KMS y client-side
Políticas IAM para regular el acceso a SQS
Políticas de acceso SQS para cross-account y otorgar permiso a otros servicios de usar SQS

**Message Visibility Timeout**

Cuando un mensaje es polled por un consumidor se vuelve invisible, por defecto es de 30 segundos y se puede aumentar usando ChangeMessageVisibility API

**Long Polling**

Cuando un consumidor va a buscar mensajes y puede esperar un rato a que aparezcan para consumir esos
Ideal para reducir las API calls para incrementar la eficiencia y reducir la latencia
Puede ser entre 1 y 20 segundos, idealmente 20 segundos

**FIFO Queues**

First In First Out
Throughput limitado sin batch a 300 msg/s y con batches a 3000 msg/s

## SNS

Hasta 12.500.000 suscriptores por tópico
Límite de 100.000 tópicos
Encryption in-flight con HTTPS API, at-rest con KMS y client-side
Políticas IAM para regular el acceso a SQS
Políticas de acceso SQS para cross-account y otorgar permiso a otros servicios de usar SQS

## Kinesis Data Streams

Recopila y almacena datos en tiempo real
Retención de hasta 365 días
Reprocesa data
Data no puede ser borrada
Datos hasta de 1MB
Encryption in-flight con HTTPS API, at-rest con KMS

### Capacity Modes

Provisioned:

- Eliges cuantos shards
- Cada shard tiene 1MB/s in y 2MB/s out
- Escalado manual
- Pagas por cada shard provisionado por hora

On-demand:

- No aprovisionas ni administras capacidad
- Default 4MB/s
- Escala automáticamente basado en los picos de los últimos 30 días
- Paga por stream por hora y la data que entra y sale en GB

### Data Firehose

Servicio para cargar data desde algún lado hacia otro
Recibe records de hasta 1MB
La data puede ser transformada por medio de Lambdas
Puede hacer escrituras en batches
Puede ser enviada a un S3 la data para backup
Casi tiempo real

## Amazon MQ

Ideal para migrar aplicaciones que ya tengan su sistema de colas con RabbitMQ y ActiveMQ
No escala tanto como SQS y SNS
Multi-AZ, failover
Tiene las funciones de SNS y SQS integrado

## ECS

Docker containers = ECS Tasks

### EC2

El usuario provee y mantiene la infraestructura
Un agente de ECS debe correr en cada EC2

### Fargate

Serverless
Solo se crean las ECS Tasks

**Auto-scaling**

Aumentar o disminuir el número de Tasks basado en:

- CPU
- RAM
- ALB requests

Target tracking: basado en métricas de CloudWatch
Step scaling: basado en CloudWatch Alarm
Scheduled scaling: basado en un lapso de tiempo

**EC2 Scaling**

Agregado de EC2s en ECS usando ASG o automáticamente usando ECS Cluster Capacity Provider

## EKS

**Managed Node Groups**

Crea y administra los nodos de EC2
Los nodos hacen parte de un ASG administrado por EKS
Puede ser con instancias on-demand o spot

**Self-Managed Nodes**

Los nodos los crea el usuario y los registra a EKS, administrados por un ASG
Se pueden usar AMIs, hay AMIs optimizadas para EKS
Puede ser con instancias on-demand o spot

**Fargate**

No es necesario administrar nada, serverless

## App2Container

CLI para migrar y modernizar Java y .NET web apps a contenedores de docker
Lift and shift
Deploy a ECS, EKS y App Runner

## Lambda

Funciones virtuales, serverless
Hasta 15 minutos de tiempo de ejecución
On-demand
Escalamiento automático
Hasta 10GB de RAM por función
$0.20 por 1 millón de requests
4KB para environment variables
Hasta 10GB de disco
Tiene un máximo de 1000 lambdas concurrentes, pero se puede reservar más capacidad
La cuota de lambdas concurrentes es para todos los servicios, o sea son 1000 al mismo tiempo para todos los servicios

Las lambdas viven en un VPC de AWS, o sea no viven dentro del VPC del usuario por ende no se pueden comunicar con recursos dentro de VPC a menos que le montes una ENI
Se pueden invocar lambdas dentro de PostgreSQL y Aurora

**Lambda SnapStart:** mejora 10 veces el funcionamiento de lamba para .NET, Java y Python sin ningún costo

## DynamoDB

Millones de requests por segundo
Trillones de filas
100s de TB de almacenamiento
Rápida y consistente
IAM para autenticación y administración
Escala automáticamente
No hay que hacerle mantenimiento
Existe el tipo **Standard** para información a la que se accede frecuentemente e **Infrequent Access** para la que no

DynamoDB esta hecha de tablas
Cada tabla tiene su **Primary Key**
Un objeto tiene un tamaño máximo de 400KB
Ideal para esquemas que evolucionan rápidamente

**Stream Processing:** Registro de modificaciones a objetos; existe DynamoDB Streams y Kinesis Data Streams

### Provisioned Mode

Se especifica el número de writes/reads por segundos
Planear la capacidad de antemano
Pagar por el aprovisionamiento de unidades para escribir y leer
Se puede habilitar autoscaling para las unidades

### On-demand Mode

Escalamiento automático
Pagas por lo que usas, caro
Ideal para cargas de trabajo impredecibles

### DynamoDB Accelerator (DAX)

Cache para DynamoDB
Ideal para minimizar el impacto de lectura a la DB
Microsegundos para data cacheada
5 minutos de TTL por defecto

### Global Tables

Accesibilidad a las tablas con baja latencia desde multiples regiones
Activa-Activa, o sea se puede escribir y leer en cualquiera
DynamoDB streams debe estar habilitado

## Cognito

Otorga identidades a usuarios para que interactúen con nuestra app
Cognito User Pools:

- Funcionalidad de sign in
- Integrable con API Gateway y ALB
- Serverless
- Username/Email y Password
- Reset password
- Email & Phone number verification
- MFA

Cognito Identity Pools (Federated Identity):

- Credenciales para que usuarios tengan acceso directo a recursos de AWS
- Se puede aplicar políticas de IAM para seguridad y control
- Autenticación por medio de apps de terceros (gmail, facebook, instagram, etc)

## Data & Analytics

### Athena

Servicio de query serverless para analizar data almacenada en S3
$5.00 por TB
Normalmente usado en conjunto con QuickSight
Usar columnar-data para minimizar costos
Usar archivos de más de 128MB para minimizar overhead

### Redshift

Basada PostgreSQL
Online analytical process (OLAP) usado para análisis y data warehouses
10 veces mejor que otros data warehouses
Escala a PBs
Columnar-data
Parallel query engine
Se integra con Tableau y QuickSight
Multi-AZ
Snaphosts incrementales cada 8 horas o cada 5GBs
Point-in-time recovery
Se puede configurar para que copie snapshots a otras regiones y levantar un cluster de Redshift nuevo a partir del mismo

**Redshift Spectrum:** query data en S3 sin cargarla, Redshift cluster habilitado

### OpenSearch

Para buscar cualquier cosa en una base de datos
Viene con un dashboard

### Elastic MapReduce

Analiza y procesa Big Data
Apache Spark, HBase, Apache Flink, Presto

No hay que configurar, ni aprovisionar nada
Autoscaling e integración con spot instances

**QuickSight**

Serverless service de BI de ML para crear dashboards interactivos
Rápido, escala automáticamente, precio por sesión
Integrado con RDS, Aurora, Athena, Redshift, S3, OpenSearch, Timestream
Si se importa data a QuickSight se puede usar in-memory computation con SPICE

## Machine Learning

**Rekognition**

Encuentra cosas con ML
Análisis y búsqueda facial

**Transcribe**

Speech-to-Text
Deep Learning y automatic speech recognition
Remueve automáticamente información personal sensible
Identifica automáticamente el idioma para audios que tengan muchos idiomas

**Polly**

Text-to-Speech
Lexicon para determinar como debe leer cosas en específico

**Comprenhend**

Procesa lenguaje natural
Análisis de textos
Medical: protege datos de pacientes

**SageMaker**

Construcción de modelos ML

**Kendra**

Buscador de información en documentos
Natural Language search capabilities
Aprende con la interacción
Fine-tune

## Security & Encryption

**Key Management Service**

Administra las llaves
Integrado con IAM para autorización
Se puede auditar el uso de la llaves con CloudTrail
$1 menual por llave

$0.03 por cada 10000 API calls
Rotación automática
Políticas de llaves para control de acceso

**Tipos de llaves**

Simétrica (AES-256):

- Una sola llave para encriptar y desencriptar
- No se tiene acceso directo a las llaves, solo por API calls

Asimétrica (RSA y ECC key pair):

- Llave pública para encriptar y privada para desencriptar
- La pública es descargable, pero la privada desencriptada es inaccesible

**Multi-region Keys**

Replicar una llave de una región a otras regiones (son idénticas a la original)
No son globales, cada réplica es independiente
Se puede encriptar en una región y desencriptar en otra con las réplicas

**Secrets Manager**

Almacenar secretos
Se puede habilitar rotación cada X días
Se puede automatizar la generación de secretos en rotación usando Lambda
Los secretos son encriptados con KMS
Se pueden replicar a otras regiones

**Certificate Manager**

Aprovisionar, administrar y desplegar certificados TLS
Soporta públicos y privados
No cobran por certificados públicos
Se renueva automáticamente
Integración con ELBs, CloudFront y APIs en API Gateway

Para integrar ACM con ELBs, los certificados deben estar en la misma región que el ELB

**Web Application Firewall**

Protege aplicaciones web en la capa 7, HTTP
Integrable con ALB, API Gateway, CloudFront, AppSync GraphQL API, Cognito User Pool
Protege contra inyección SQL y Cross-Site Scripting
Block countries
Rate-based rules para evitar DDoS
Se pueden hacer grupos de reglas reusables

**Shield**

Protege contra DDoS
Gratis para todos los usuarios
Protege contra ataques en la capa 3 y 4

**Advanced Shield**

Mitigación opcional de DDoS - $3000
Protege contra ataques más sofisticados
24/7 con el equipo de respuesta de AWS
Protege contra costos por picos altos en caso de ataque DDoS
Se integra con WAF para mitigar ataques de capa 7

**Firewall Manager**

Para manejar las reglas en todas las cuentas de AWS en una Organización
Las reglas que se creen desde aquí son aplicadas a todos los recursos nuevos que se vayan creando
Ideal para compliance

**GuardDuty**

Intelligent Threat discovery based on ML algorithms
Detección de anomalías y archivos de terceros sospechosos
Protege contra CryptoCurrency attacks

**Inspector**

Automated Security Assesments (escaneo continuo)
EC2, ECR, Lambda
Reporta al Security Hub

**Macie**

Servicio para seguridad y privacidad de datos que usa ML para proteger data sensible en AWS
Si descubre PII, te alerta sobre ello

## Networking

**VPC**

Máximo 5 VPCs por región (puede aumentar)
Máximo 5 CIDRs por VPC:

- Min size /28
- Max size /16

10.0.0.0
172.16.0.0
192.168.0.0

**Subnet**

Las primeras 4 y la última IPs están reservadas:

- Address
- Router
- DNS
- Una para uso futuro
- Broadcast

**Internet Gateway**

Permite acceso a internet
Escala horizontalmente y tiene alta disponibilidad y redundancia
Solo se puede asociar a una VPC
Se deben configurar las route tables para habilitar el acceso

**Bastion Hosts**

Se usa como punto de acceso a instancias en una subnet privada

**NAT Instances (Outdated)**

Permite el acceso a internet a instancias en subnets privadas
Debe estar en una subnet pública
Necesita IP elástica
Necesita configuración de route tables

**NAT Gateway**

Administrada por AWS
Pay per hour por uso y ancho de banda
Es creada en una AZ y necesita IP elástica
No puede ser usada por una EC2 en la misma subnet, debe estar en otra subnet
Requiere un IGW
Resiliente en una sola AZ
Para fail-tolerance se necesitan otras NGW en otras AZs

**NACL**

Es como un stateless firewall
Una por subnet
El usuario define las reglas, la regla con un valor menor tiene prioridad

**VPC Peering**

Conectar dos VPCs internamente sin exponer el tráfico a la red pública
Se comportarían como si estuviesen en la misma red
Sus CIDRs no deben overlapearse
Hay que actualizar las route tables para asegurarse de la comunicación
Se puede desde diferentes cuentas y regiones

**VPC Endpoints**

Recurso que permite comunicar servicios de AWS por medio de la red privada de AWS

**Interface Endpoints**

Aprovisiona una ENI como punto de entrada
Soporta la mayoría de servicios de AWS
Pago por hora y GB procesado

**Gateway Endpoints**

Gateway que se usa como un target en una route table
Gratis
Acceso a S3 y DynamoDB

**VPC Flow Logs**

Captura la información de tráfico por IP en las interfaces
Se puede hacer query a los logs usando Athena en S3 o CloudWatch Logs

**Networking costs**

Tráfico de entrada gratis
Tráfico por la red pública $0.02
Tráfico por la red privada $0.01
Tráfico entre regiones $0.02