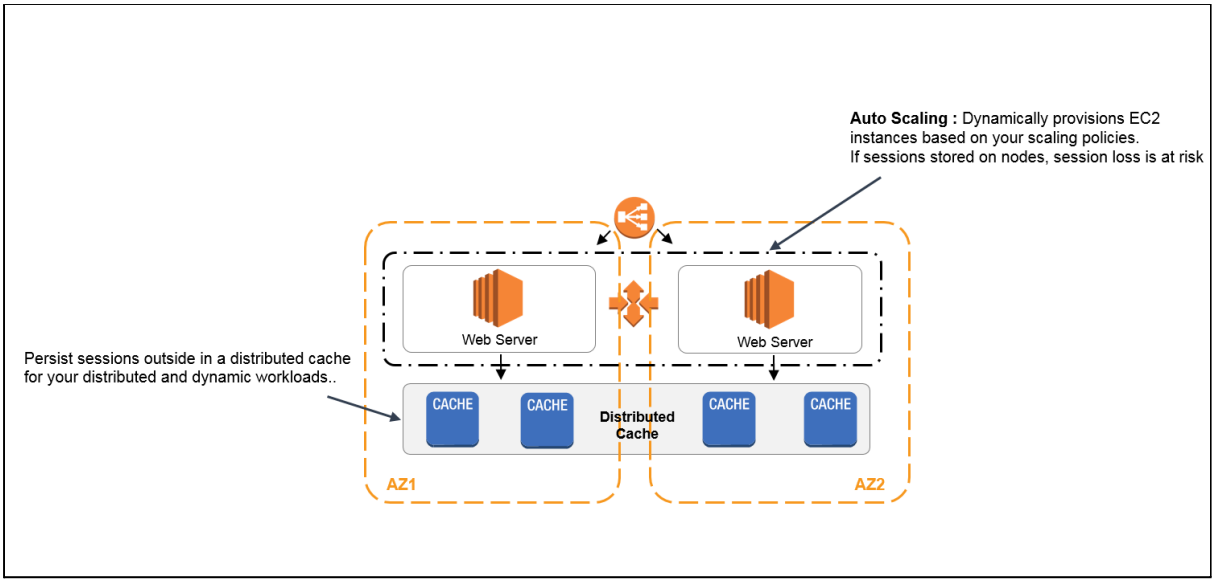
# Flashcards

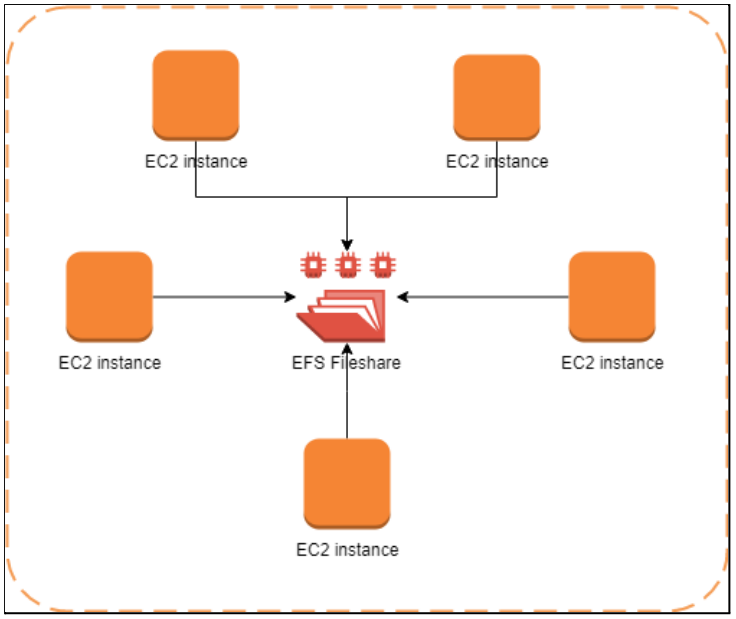
## 1) Sessions

- **Sticky Sessions**
  - Think of ALB and ElastiCache. If it says about storing session data in a database at all, then the choice would be DynamoDB.
  - The particular **web server** that is managing that **INDIVIDUAL user's session**
- **ElastiCache in-memory caching**
  - **User's session SHARED** among the fleet of instances




## 2) EFS

- **Distributed**
- **Lowest-latency** access to their data
- **Multiple EC2 instances connect** (allows concurrent connection from **multiple EC2** instances hosted on **multiple AZs**)
- **File system**
- **NFS**
- **File Storage**
- **POSIX**-compliant share file storage
- Scalable performance



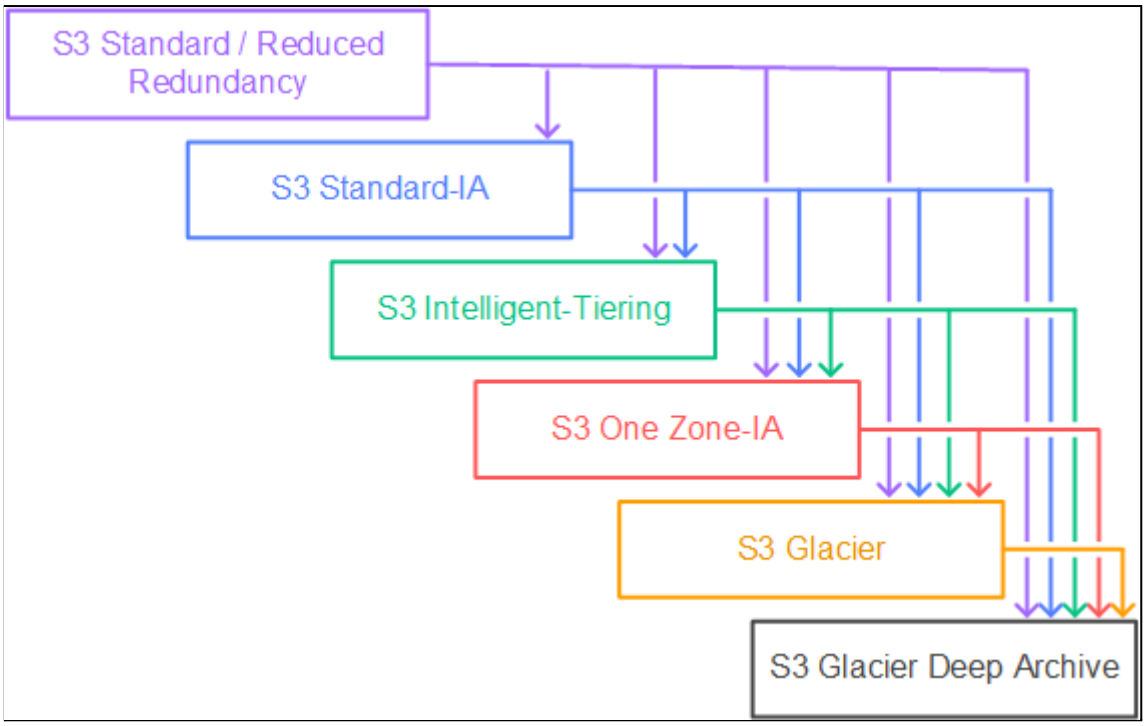

## 3) S3

Key: **Durable storage**
Transferation: **Enable Transfer Acceleration** in the destination bucket and upload the collected data **using Multipart Upload.**

### Encryption

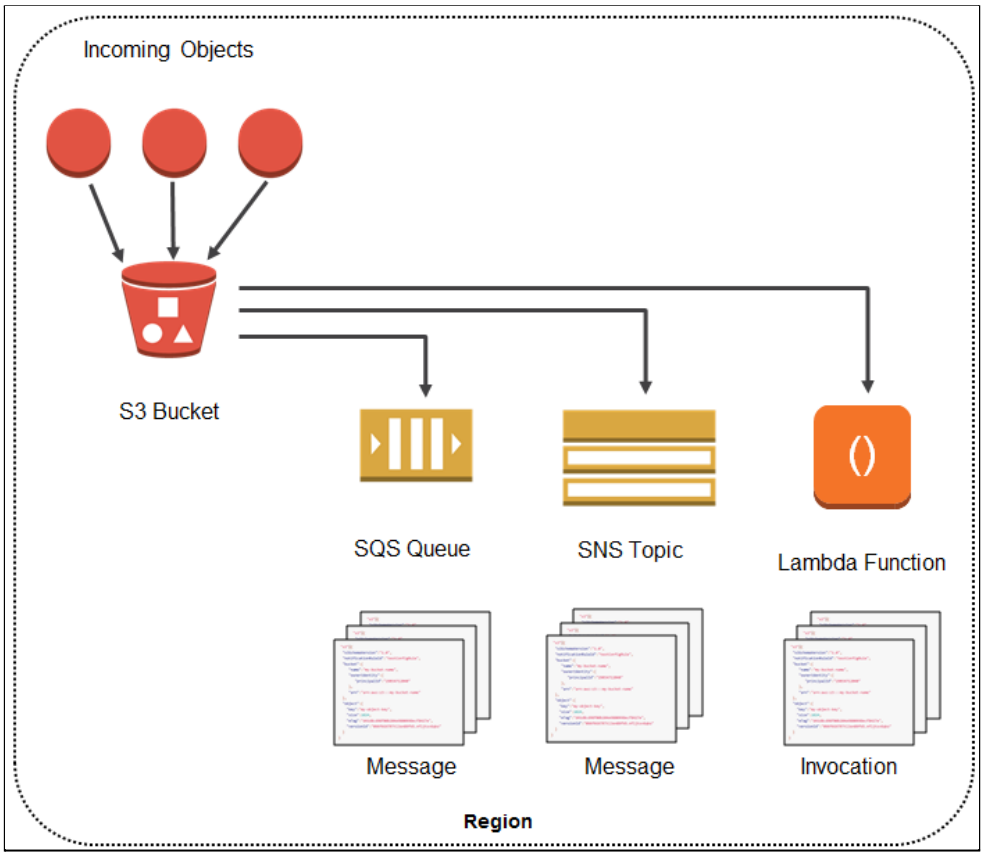
- **Transit**
  - SSL/TLS
  - **Before sending** the data to Amazon S3 over HTTPS, **encrypt the data locally first using your own encryption keys.**
- **Rest (Server Side)**
  - S3 Managed Keys - **SSE-S3** (**AWS managed the keys**)
    - Enable **Server-Side Encryption** on an S3 bucket to make use of **AES-256** encryption.
  - AWS Key Management Service, Managed Keys - **SSE-KMS** (**AWS/Customer**)
  - Server Side Encryption With Customer Provided Keys - **SSE-C** (**Customer**)
- **Client Side Encryption**
  - Use **S3 client-side encryption with a client-side master key**.
    - Both the master keys and the unencrypted data should never be sent to AWS

## Storages




## S3 Event Notification

- SQS: Message
- SNS: Message
- Lambda Function: Invocation



### Retrieval

- **Expedited retrievals**
    - Allow you to quickly access your data when occasional urgent requests for a subset of archives are required. For all but **the largest archives (250 MB+)**, data accessed using Expedited retrievals are typically made **available within 1–5 minutes.**
- **Provisioned capacity**
    - Ensures that your retrieval capacity for expedited retrievals is **available when you need it.** Each unit of capacity provides that at least three expedited retrievals **can be performed every five minutes and provides up to 150 MB/s of retrieval throughput.**

### Policies

#### ACLs

- When **objects** are **NOT owned by bucket owner**
- Individual objects and S3

#### Bucket Policy

- Bucket level

### S3 Access logs

S3 buckets can be configured to create access logs which log all requests made to the S3 bucket. This can be sent to another bucket and even another bucket in another account.

### Athena

It is interactive query service which enables you to analyze and query data located in S3 using standard SQL

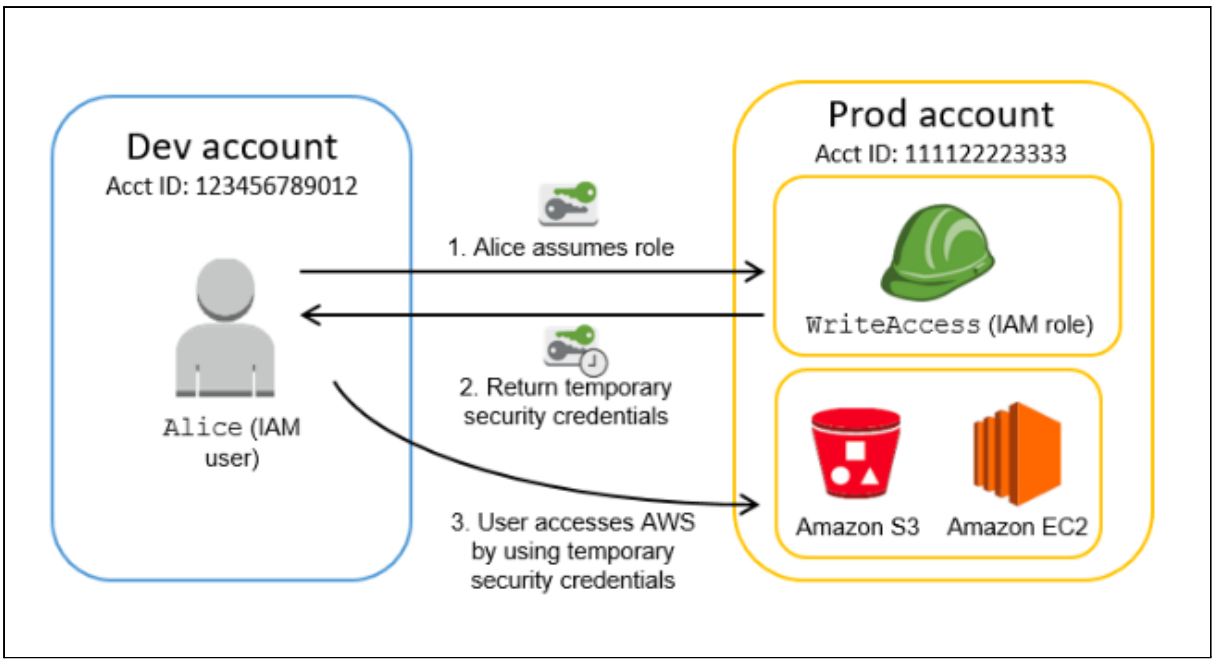
- Serverless
- Works directly with data stored in S3

### Macie

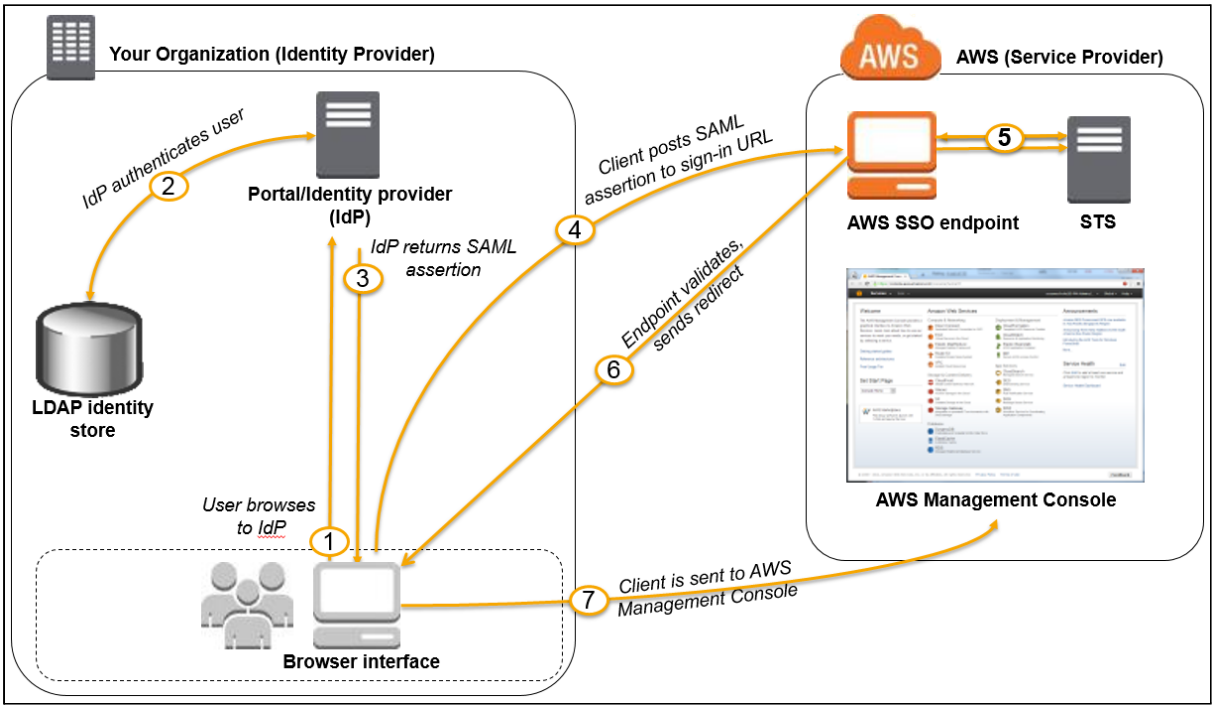
Security service which uses Machine Learning and NLP to discover, classify and protect sensitive data stored in S3

## 4) IAM

- With **Web Identity Federation** you **don't need** to **create** custom **sign-in** code or **manage** your **own** user **identities** (Amazon, Facebook, Google or any **OpenID** Connect -OIDC-)
- AWS Security Token Service (STS) - Temporary security credentials




- **Create a set of Access Keys** for the user and attach the necessary permissions in order to **send API requests** to AWS resources
- Develop an on-premises custom identity broker application and use STS to issue short-lived AWS credentials when your identity store is not compatible with SAML 2.0
- SSO with SAML
- SAML 2.0-Based Federation by using MS AD Federation Service (AD FS)




## Web Identity Federation

Users can authenticate with web-based identity providers like Amazon, Facebook or Google.

### Cognito

- Federation allows users to authenticate with a Web Identity Provider
- The user authenticates first with the Web ID Provider and receives an authentication token, which is exchanged for temporary AWS credentials allowing them to assume an IAM role.
- User pool is user based (user registration, authentication and account recovery)
- Identity pool authorize access to your AWS resources

## 5) EC2

- **Spot**
    - Can be interrupted
- **On-Demand**
    - Cannot be interrupted and needed for sometime
- **Reserved Instance**
    - Cannot be interrupted and needed for a specific time (9pm-4am) periodically for 1 year or more- scheduled
    - Cannot be interrupted and needed constantly for 1 year to 3 year
    - You can sell your RI on marketplace
- **Dedicated Instance**
    - Need a server for compliance and auditing, or a dedicated hardware
- **Convertible Instance**
    - Can only be converted to another convertible instance
- **AWS CLI EC2**
    - The **describe-instances** command shows the status of the EC2 instances including the recently terminated instances. It also returns a **StateReason** of why the instance was terminated.
- **Limitations**
    - There is a vCPU-based On-Demand Instance limit per region which is why subsequent requests failed. Just submit the limit increase form to AWS and retry the failed requests once approved (20)

### ENI vs ENA vs EFA

- **ENI**
    - Elastic Network Interface - essentially a virtual **network card**
- **EN (Enhanced Network)**
    - **Enhanced Networking** - uses single root I/O virtualization (SR-IOV) to provide **high-performance networking** capabilities on supported instance types.
        - ENA: 100 Gbps
        - VF: 10 Gbps (older instances)
- **EFA (Elastic Fabric Adaptor)**
    - A network device that you can attach to your EC2 to accelerate **High Performance Computing (HPC) and machine learning** applications
    - It can use **OS-bypass**
    - **NOT supported with Windows** currently, only Linux

### Placement Strategies

- **Cluster**
  - This strategy enables workloads to achieve the low-latency network performance necessary for tightly-coupled node-to-node communication that is typical of HPC applications.
  - If you **receive a capacity error** when launching an instance in a placement group that already has running instances, **stop and start all of the instances in the placement group, and try the launch again**.

- **Partition**
  - Groups of instances in one partition do not share the underlying hardware with groups of instances in different partitions. This strategy is typically used by large distributed and replicated workloads, such as **Hadoop, Cassandra, and Kafka.**
- **Spread**
  - Strictly places a small group of instances across distinct underlying hardware to reduce correlated failures.

## 6) EBS

- **General Purpose SSD (16,000 IOPS) - gp2 - Small/Random**
  - Balances
  - Boot
- **Provisioned IOPS SSD (64,000 IOPS) - io1 - Small/Random**
  - Mission-Critical Apps
  - DB
- **Throughput Optimized HDD (500 IOPS) - st1 - Large/Sequential**
  - Frequently accessed
  - Big Data & Data Warehouse
- **Cold HDD (250 IOPS) - sc1 - Large/Sequential**
  - Less frequently accessed
  - File Servers

| Solid-State Drives (SSD) | | | Hard disk Drives (HDD) | | |
|---|---|---|---|---|---|
| Volume Type | **General Purpose SSD** | **Provisioned IOPS SSD** | **Throughput Optimized HDD** | **Cold HDD** | **EBS Magnetic** |
| Description | General purpose SSD volume that balances price and performance for a wide variety of transactional workloads | Highest-performance SSD volume designed for mission-critical applications | Low cost HDD volume designed for frequently accessed, throughput-intensive workloads | Lowest cost HDD volume designed for less frequently accessed workloads | Previous generation HDD |
| Use Cases | **Most Work Loads** | **Databases** | **Big Data & Data Warehouses** | **File Servers** | **Workloads where data is infrequently accessed** |
| API Name | gp2 | io1 | st1 | sc1 | Standard |
| Volume Size | 1 GiB - 16 TiB | 4 GiB - 16 TiB | 500 GiB - 16 TiB | 500 GiB - 16 TiB | 1 GiB-1 TiB |
| Max. IOPS**/ Volume | 16,000 | 64,000 | 500 | 250 | 40-200 |

### Snapshots

- Exist on S3
- Snapshots are **incremental**
- You can **take a snap while the instance is running** also change volume sizes on the fly
- **Volumes** will **ALWAYS** be in the **AZ as the EC2 instance**
- Use **Amazon Data Lifecycle Manager (DLM) to automate the creation of EBS snapshots**
- To move an EC2 volume from one AZ to another, take a snapshot of it, create an AMI from the snapshot and then use the AMI to launch the EC2 instance in a new AZ
- To move an EC2 volume from one region to another, take a snapshot of it, create an AMI from the snapshot and then copy the AMI from one region to the other. Then use the copied AMI to launch the new EC2 instance in the new region.

## 8) Elastic Load Balancing (ELB)

- ALB - TLS termination capabilities, path-based routing, host-based routing, and bi-directional communication channels using WebSockets, dynamic host mapping
- ALB, NLB - IP address for distributing traffic
- CLB - instance IDs for distributing traffic
- Go over **access logs for ELB**, don't confuse them with server access logging of S3
    - IP
    - Latencies
    - Request paths
    - Server responses
- **ALB - cross zone load balancing enabled by default,** NLB not on by default, CLB - on by default from management console else disabled by default
- CloudFront and ELB support **Perfect Forward Secrecy**
- **ELB tests status** with **health checks** (example: HTTP/S)
- ELB uses **path conditions** that **forward** requests to different target groups based on the URL in the request.
- DNS
    - Create an A record aliased to the load balancer DNS name

## 9) Auto Scaling group

### Components

- Groups
    - Web Server, Application or Database groups
- Configuration Templates
    - Launch template
    - Launch configuration
    - AMI ID, instance type, key pair, security groups and block device mapping for your instances
- Scaling Options
    - Dynamic
    - Schedule

### Scaling Policies

- **Simple**: Increase or decrease instances based on a **single adjustment**
- **Scheduled**: Scaling schedule for **predictable load changes**
- **Target tracking**:  Increase or decrease instances based on a **target value for specific metric** (example: you select a temperature and the thermostat does the rest)
- **Step**: Increase or decrease instances based on a set of **scaling adjustments** (step adjustments - the size of the alarm breach)

### Allocation strategy

1. Oldest Launch Configuration
2. Oldest Launch Template
3. Closest to next billing hour
4. Choose the AZ with the most number of instances
5. Random

### Health Check

- **Unhealthy instance** - First terminate then launch a new instance. AZ rebalancing - other way round
- If you configure the Auto Scaling group to use Elastic Load Balancing health checks, it considers the instance unhealthy if it fails either the EC2 status checks or the load balancer health checks.
- If you attach multiple load balancers to an Auto Scaling group, all of them must report that the instance is healthy in order for it to consider the instance healthy. If one load balancer reports an instance as unhealthy, the Auto Scaling group replaces the instance, even if other load balancers report it as healthy.

### Cooldown Period

- It ensures that the Auto Scaling group does not launch or terminate additional EC2 instances before the previous scaling activity takes effect.
- Its default value is 300 seconds.
- It is a configurable setting for your Auto Scaling Group

### Launch Configuration

You can specify your launch configuration with multiple Auto Scaling groups. However, you can only specify **one launch configuration for an Auto Scaling group at a time** and **you can’t modify a launch configuration after you’ve created it**

## 10) Encryption

Data at rest, data-in-transit for all services. S3 and RDS encryption to understand in depth is a must. In **Lambda**, use **encryption helpers along with KMS** for securing environment variables

### CloudHSM

**FIPS 140-2 Level 3**

- PKCS#11
- Java Cryptography Extensions (JCE)
- Microsoft CryptoNG (CNG).

HSM zeroized, which means that the encryption keys on it have been wiped.

- The keys are lost permanently if you did not have a copy.

## 11) API Gateway

Throttling and caching

## 12) Network Filtering

- Security group (SG)
    - Default VPC - inbound (only the same SG), outbound (all IP)
    - New VPC - inbound (all deny), outbound (all IP)
    - SG - All rules are evaluated until a permit is encountered or continues until the implicit deny.

- NACL
    - Default - Allow all traffic (inbound and outbound)
    - New - Deny all traffic (inbound and outbound)
    - NACLs only apply to traffic that is ingress or egress to the subnet not to traffic within the subnet

## 13) VPC

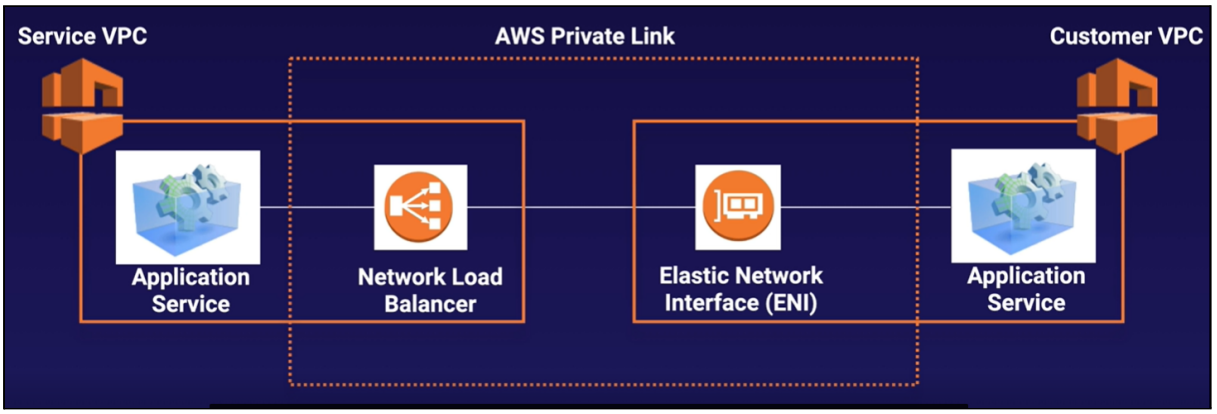
### VPC Endpoints

A VPC endpoint enables you to privately connect your VPC to supported AWS service and BPC endpoint service powered by PrivateLink

- Interface Endpoints
    - Elastic Network Interface with a private IP address that serves as an entry point for traffic destined to a supported service
- Gateway Endpoints
    - Amazon S3
    - DynamoDB
    - S3 and DynamoDB have Gateway endpoints (route table entry), other services have interface endpoints (entry via Private IP)

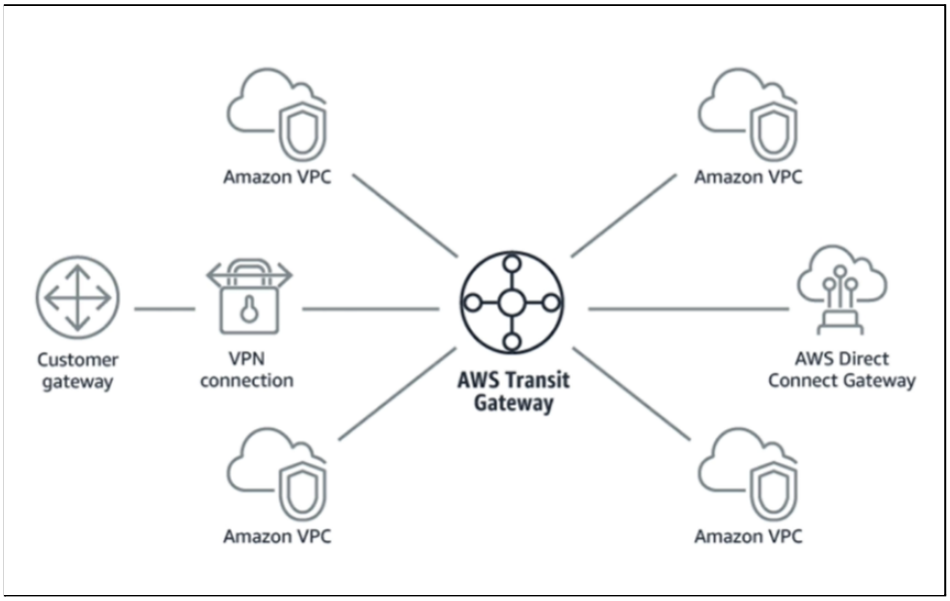
## AWS Private Link

- The best way to expose a service VPC to tens, hundreds, or thousands of customer VPCs
- Doesn't require VPC Peering; no route tables, NAT, IGWs, etc.
- Requires a NLB on the service VPC and an ENI on the customer VPC



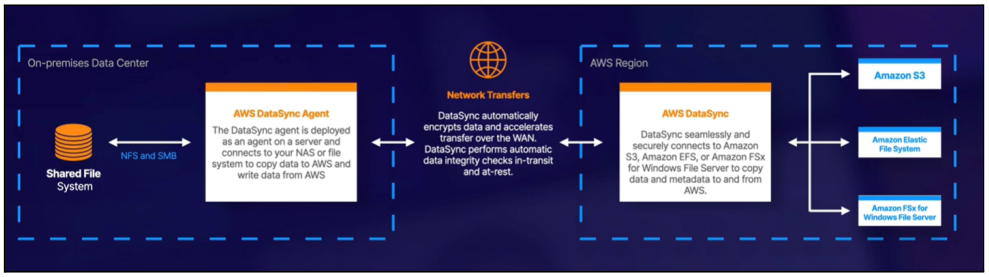

## AWS Transit Gateway

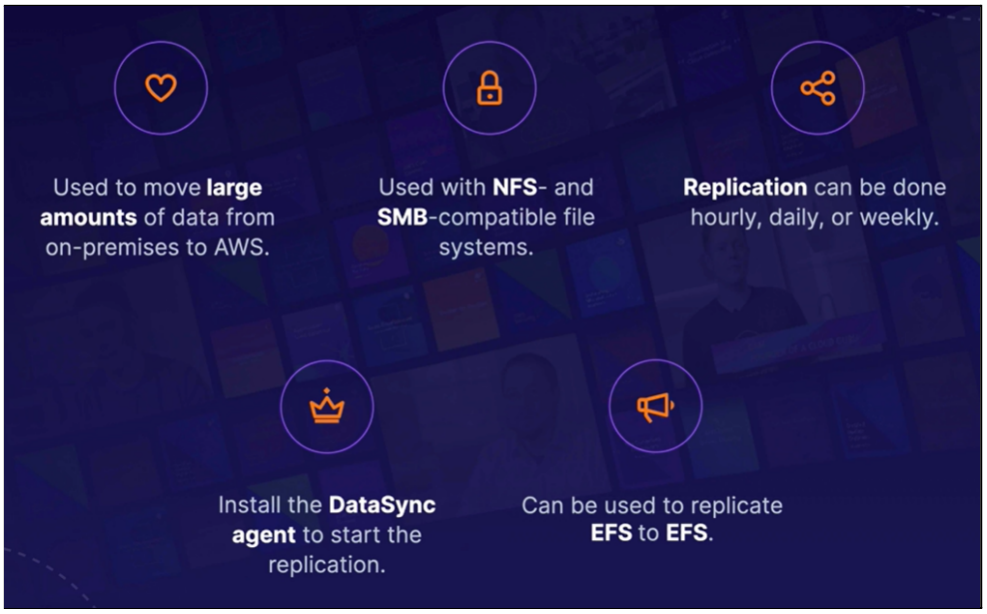
- Allows you to have transitive peering between thousands of VPCs and on-premises data centers
- Work on a hub-and-spoke model
- Works on a regional basis, but you can have it across multiple regions
- You can use it across multiple AWS accounts using RAM
- You can use route tables to limit how VPCs talk to one another
- **Works with Direct Connect as well as VPN connections**
- **Support IP multicast**
- **Simplify network topology**



## AWS DataSync

- Moving data from on-premises data center to AWS
- It used to move large amounts of data from on-premise to AWS
- It's used with NFS and SMB compatible file systems




## VPC Considerations

**To enable access to or from the Internet for instances in a VPC subnet, you must do the following:**

- Attach an internet gateway to your VPC.
- Ensure that your subnet's route table points to the Internet gateway.
- Ensure that instances in your subnet have a globally unique IP address (public IPv4 address, Elastic IP address, or IPv6 address).
- Ensure that your network access control lists and security groups allow the relevant traffic to flow to and from your instance.

**The following VPC peering connection configurations are NOT supported.**

- Overlapping CIDR Blocks
- Transitive Peering
- Edge to Edge Routing Through a Gateway or Private Connection

**VPC Interface Endpoint doesn't** support the **S3** service. **S3 and DynamoDB use VPC Gateway.**

## 14) Kinesis

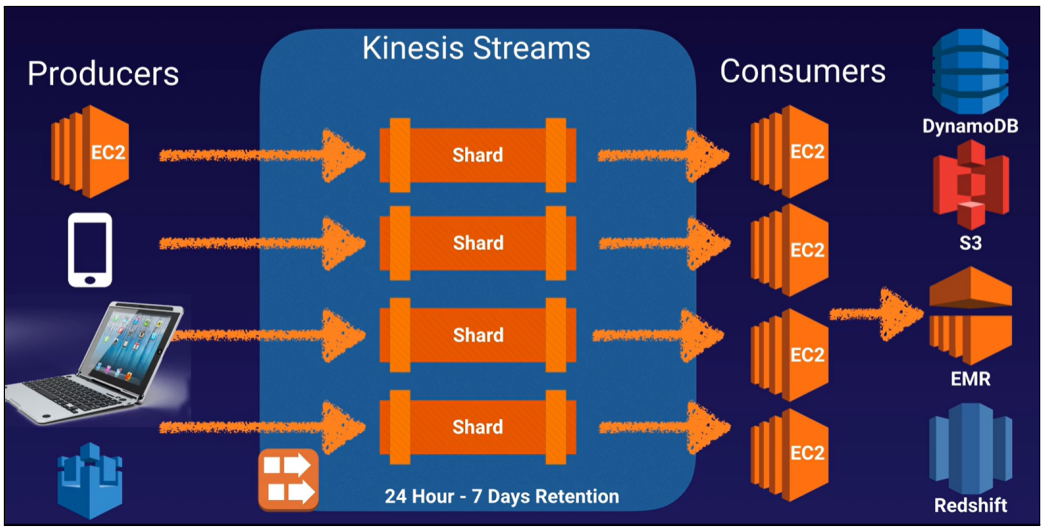
**Streaming Data** is data that is generated continuously by thousands of data sources

- Purchases from online stores
- Stock prices
- Game data
- Social network data
- Geospatial data
- IoT sensor data

Kinesis is a platform on AWS to send your streaming data to. It analyzes streaming data, and also provides the ability for you to build your own custom applications for your business.

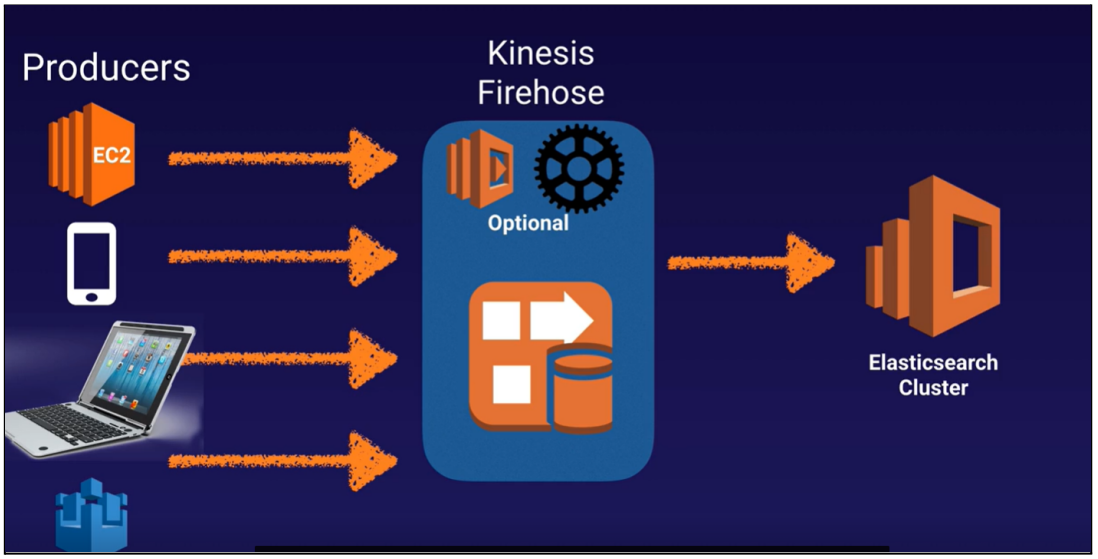
### Kinesis Streams

- **Producers**: laptop, mobile phones, EC2 and so on (senders)
- **Consumers**: EC2 (analyzers)
- **Storage**: DynamoDB, S3, EMR, Redshift
- **Data is persistence**
    - It stores data for 24 hours (it can store it up to seven days)
- **Shards are containers of data** (social data, iot data and so on)
- **Real time click streaming**, can store data in DynamoDB, Redshift, S3, EMR, Kinesis Firehose



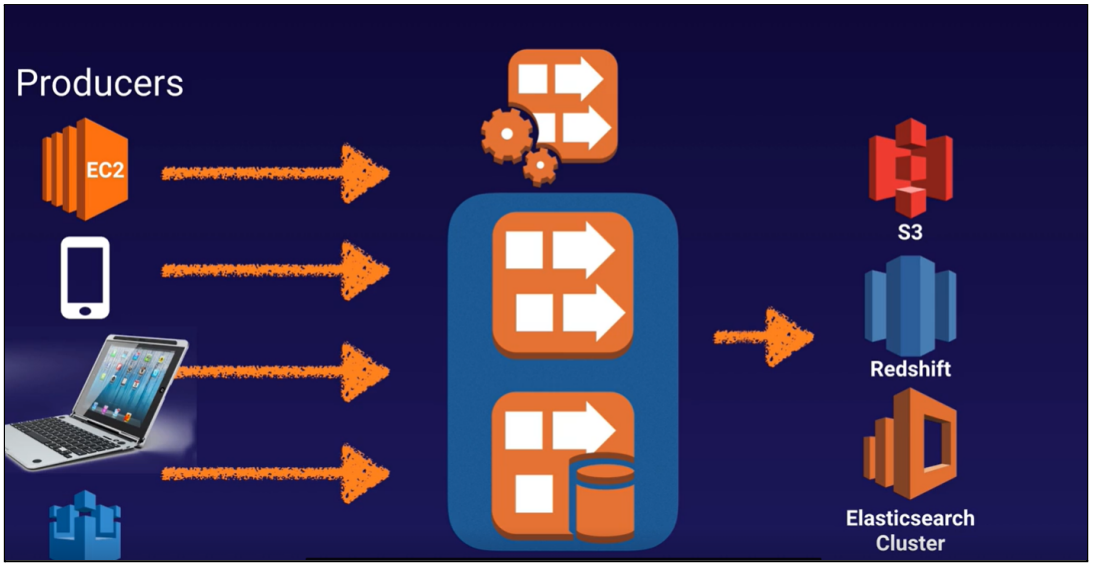
## Kinesis Firehose

- **Producers**: laptop, mobile phones, EC2 and so on (senders)
- **Consumers**: EC2 (analyzers)
- There are not persistence storage
- It can have origins as S3, Redshift, Elasticsearch, Splunk.
- Optional: Lambda function inside in your Kinesis Firehose
    - Storage
        - S3 > Redshift
        - Elasticsearch Cluster




## Kinesis Analytics

- Analyzes data on the fly
- Analyzes inside both Kinesis Stream and Kinesis Firehose
- Storage:
    - S3
    - Redshift
    - Elasticsearch Cluster



## 15) SQS

SQS is a web service that gives you access to a message queue that can be used to store messages while waiting for a computer to process them.

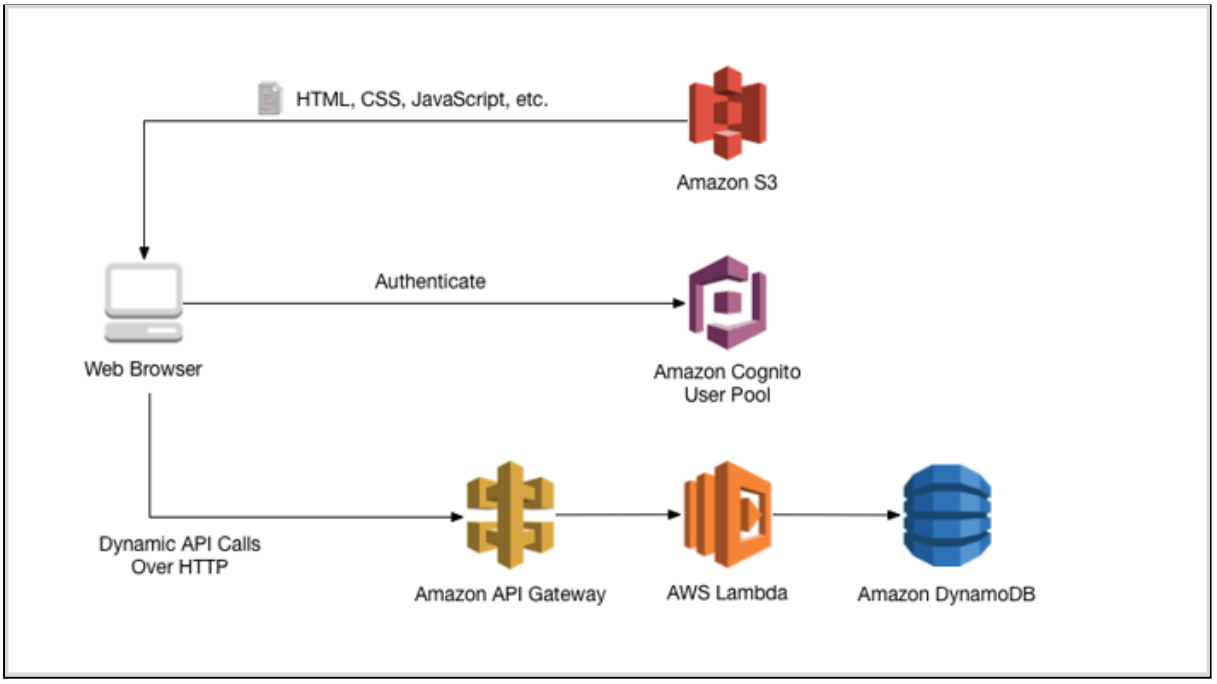
- **Messages are 256 KB in size**
- **Messages can be kept** in the queue from **1 minute to 14 days**; the **default** retention period is **4 days**)
- **Visibility Timeout** is a period of time during which Amazon SQS **prevents** other consuming components from receiving and processing a message (the **default** visibility timeout for a message is **30 seconds** and the **maximum** is **12 hours**)
    - To prevent other consumers from processing the message again, Amazon SQS sets a visibility timeout
- **ReceiveMessageWaitTimeSeconds** is the queue attribute that determines whether you are using Short or Long polling (**default 0** - which means it is using **Short polling**)
    - **Long polling reduces** the number of **empty resources**, helps reduce your **cost** and eliminates **false empty responses**
- **SQS automatically deletes the messages** that have been in a queue for more than **the maximum message retention period (default 4 days** and **maximum 14 days using the SetQueueAttributes).**

## 16) CloudFront

- **SSL Traffic**
    - **Generate** an **SSL certificate with AWS Certificate Manager** and **create** a CloudFront web **distribution**. **Associate** the **certificate with** your web **distribution** and **enable** the support **Server Name Indication (SNI)**
- **Create an Origin Access Identity (OAI)** when you **serve content** that is stored in **S3 but not publicly accessible from S3 directly**
- **Control how long your objects stay in a CloudFront cache with Cache-Control and Expires headers**
    - CloudFront **supports** is **0 seconds for web distributions** and **3600 seconds for RTMP distributions**
- **Signed URLs**
    - RTMP distribution
    - Restrict access to **individual** files (installation download for you application)
    - Your users are using a client
- **Signed Cookies**
    - Access to **multiple** restricted files
    - You don't want to change your current URLs
- Use **Amazon ElastiCache** for the website's **in-memory data store or cache**
- **Edge Location** helps **deliver high availability, scalability and performance** of your application for all of your customers from **anywhere in the world**

## 17) Lambda

- Supports the synchronous and asynchronous invocations
- You pay only for what you use
- The **default timeout is 3 seconds** and **the maximum** execution duration per request is **900 seconds (15 minutes)**
- Use **encryption helpers along with KMS** for securing environment variables
- Lambda Edge is a feature of CloudFront that lets you run code closer to users of your application which improves performance and reduces latency
- Use API Gateway and Lambda for scalable and cost-effective

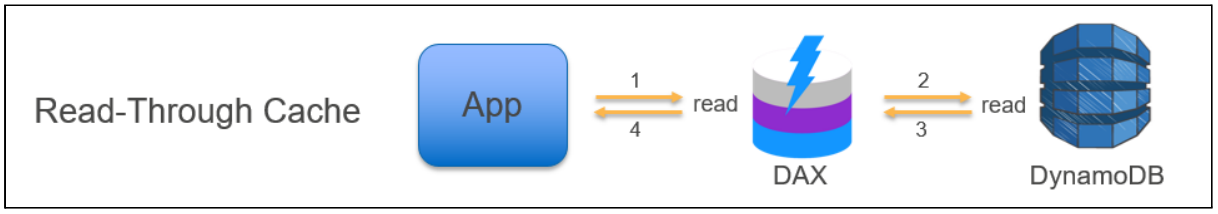



## 18) RDS (OLTP)

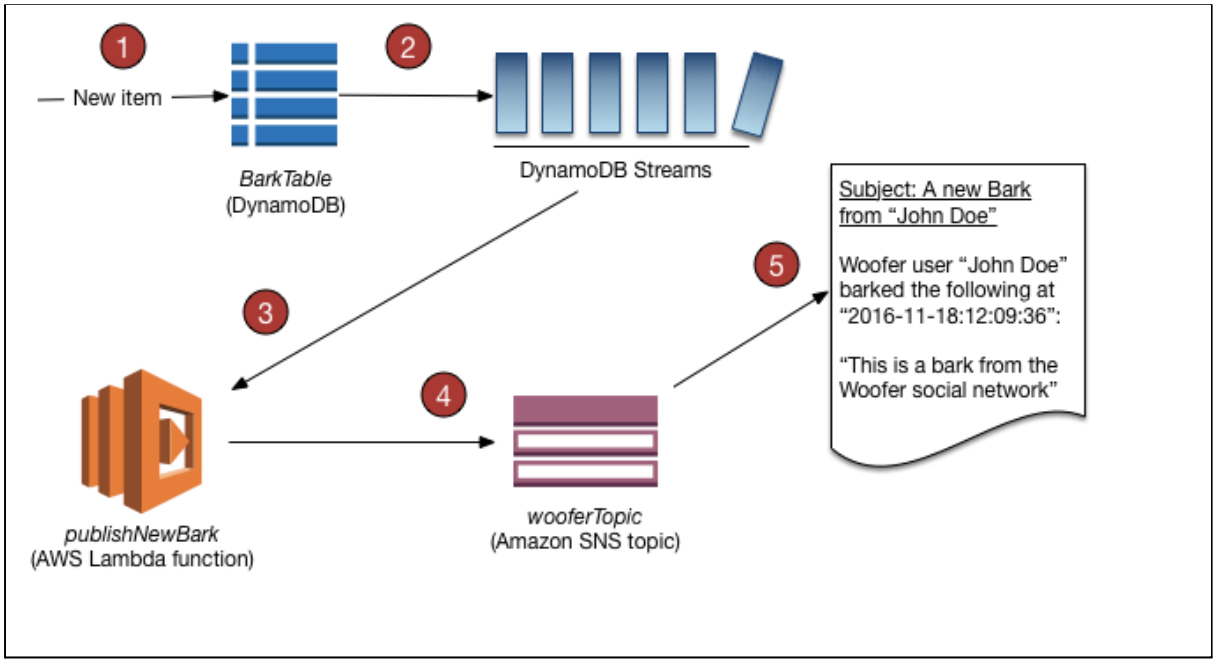
- Amazon RDS **Read Replicas** provide enhanced performance and durability for database (DB) instances (**asynchronously**)
- **Secure Communication**
  - Download the **Amazon RDS Root CA certificate**. Import the certificate to your servers and configure your application to use SSL to encrypt the connection to RDS.
  - Force all connections to your DB instance to use **SSL** by setting the **rds.force_ssl parameter to true**. Once done, reboot your DB instance.
  - Enable the **IAM DB Authentication.**
    - With this authentication method, you **don't need to use a password when you connect to a DB instance**. Instead, you use an authentication token.
  - **Failover**
    - Storage failure on primary
    - Loss of availability in primary Availability Zone
  - **Multi-AZ Failover**
    - The canonical name record (**CNAME**) is switched from the primary to standby instance.

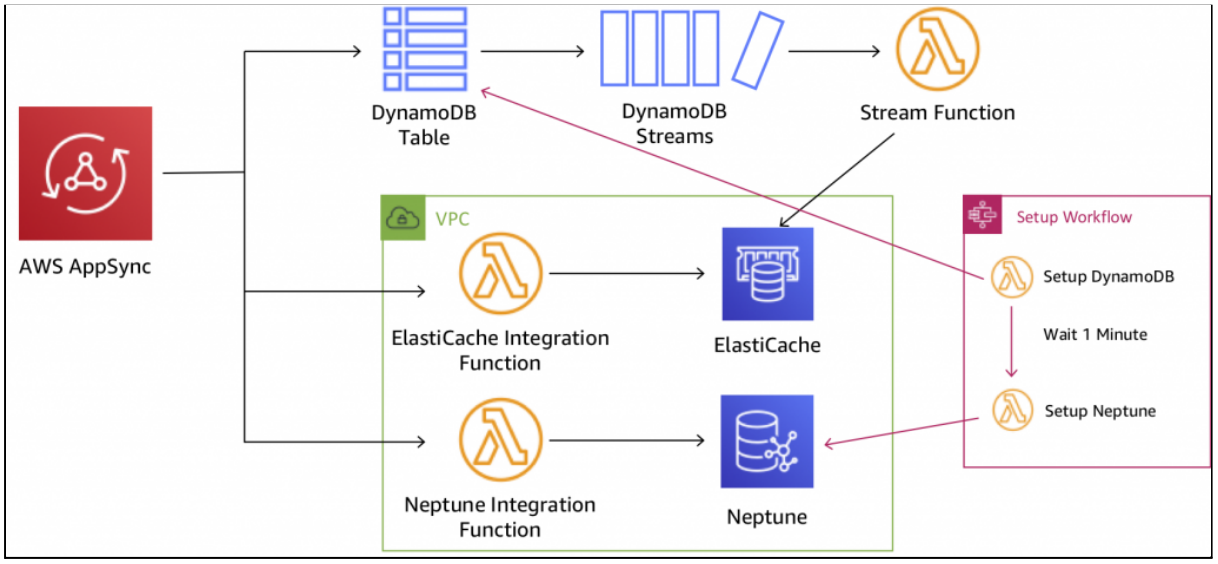
## 19) DynamoDB (NoSQL)

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




- Use DynamoDB for schema changes
- If you enable DynamoDB Streams on a table you can associate the stream ARN with a Lambda function that you write



- The partition key portion of a table's primary key determines the logical partitions in which a table's data is stored.
    - The more distinct partition key values, the more those requests will be spread across the partition space

#### Index

##### GSI

A global secondary index is considered "global" because queries on the index can **span all of the data in the base table, across all partition**s. A global secondary index is stored in its own partition space away from the base table and scales separately from the base table.

##### LSI

An index that has the same partition key as the base table, but a **different sort key**. A local secondary index is "local" in the sense that every partition of a local secondary index is scoped to a base table partition that has the same partition key value.

## 20) Redshift (OLAP)

- Amazon data warehousing solution
- Online analytical processing data
- Business Intelligence or Data Warehousing

## 21) Elasticache

Speed up performance of existing databases (frequent identical queries)

- Memcached
- Redis

## 22) Storage Gateway

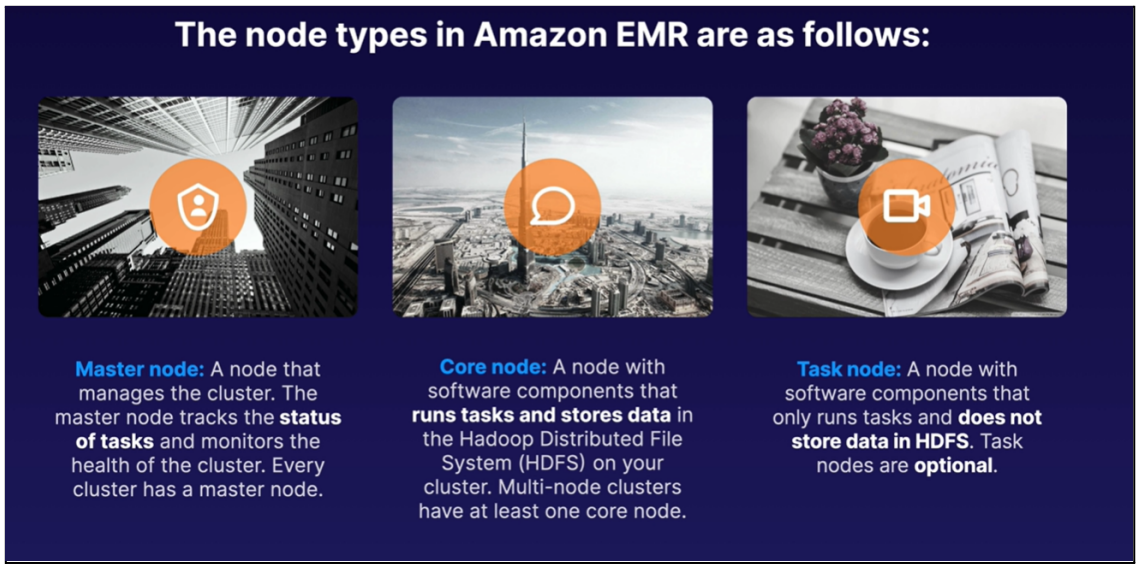
- **File Gateway**
    - **S3**
    - You can store and retrieve files directly using the **NFS** version 3 or 4.1 protocol.
    - You can store and retrieve files directly using the **SMB** file system version, 2 and 3 protocol.
    - You can access your data directly in **S3** from any AWS Cloud application or service.
    - You can manage your **S3** data using lifecycle policies, cross-region replication, and versioning.
- **Volume Gateway**
    - **SCSI**
    - Type
        - **Cached**
            - You store your data in S3 and retain a copy of **frequently accessed** data subsets locally

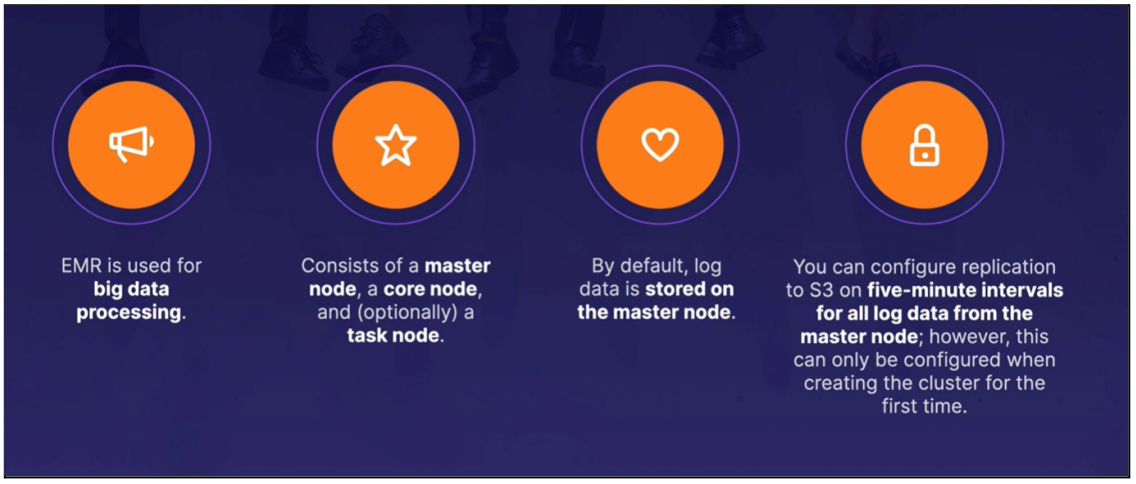
- **Stored**
    - If you need **low-latency access** to your entire dataset, first configure your on-premises gateway to store all your data locally.

## 23) EMR (Elastic MapReduce)

It is a cloud **big data** platform for processing vast amounts of data using open-source tools such as Apache Spark, Apache Hive, Apache HBase, Apache  Flink, Apache Hudi and Presto (faster than standard Apache Spark).

- Component: Cluster
    - It is a collection of EC2 instances
    - Each intaces in the cluster is a node




## 24) Global Accelerator

- It is a service in which you create accelerators to **improve availability and performance of your applications for local and global users**.
- Global Accelerator provides you with **two static IP addresses**
- You can control traffic using traffic dials. This is done within the endpoint group

## 25) Elastic Beanstalk

You can quickly deploy and manage applications in the AWS Cloud without worrying about the infrastructure that runs those applications. You simply upload your application and it automatically handles the details of capacity provisioning, load balancing, scaling, and application health monitor

- It will grow out your infrastructure

## 26) SNS

It is a web service that make it easy to set up, operate and send notifications from the cloud

- Instantaneous, push-based delivery (no polling)
- Simple APIs and easy integration with applications
- Flexible message delivery over multiple transport protocols

### SNS Topic

Access point for allowing recipients to dynamically subscribe for identical copies of the same notification.

### SNS vs SQS

- SNS - Push
- SQS - Polls