# AWS Solution Architect Associate

Version : C03

Domain 3

Task 5

## Determine high-performing data ingestion and transformation solutions

# Data Analytics and Visualization Services

AWS offers a robust suite of data analytics and visualization services designed to help businesses manage, analyze and visualize their data effectively. Key services include :

- ➔ **Amazon Athena** : Interactive query service that makes it easy to analyze data in Amazon S3 using SQL
- ➔ **Amazon EMR** : Big data platform for processing vast amounts of data using open-source tools such as Apache Hadoop, Spark and HBase
- ➔ **Amazon QuickSight** : Scalable, serverless, machine learning-powered business intelligence service
- ➔ **Amazon Kinesis** : Platform for real-time processing of streaming data at massive scale
- ➔ **Amazon CloudSearch** : Simple\Cost-effective service to set up, manage, and scale a search solution
- ➔ **Amazon OpenSearch Service** : Service to deploy, operate and scale OpenSearch clusters
- ➔ **AWS Glue** : ETL service to prepare and load data for analytics
- ➔ **Amazon Redshift** : Fast, scalable data warehouse that makes it simple and cost-effective to analyze all data using standard SQL and existing business intelligence tools
- ➔ **AWS Data Pipeline** : Web service that process and move data between different AWS compute and storage services & on-premises data sources at specified intervals
- ➔ **Amazon DataZone** : Data management and governance service that understand and control data usage
- ➔ **AWS Lake Formation** : Service for setting up a secure data lake

# Data Ingestion patterns

AWS provides various data ingestion patterns to meet diverse needs and use cases. Primary patterns :

➔ **Homogeneous Data Ingestion :**
This pattern is used when the source and target data stores are of the same type, typically relational databases. It involves migrating or replicating data from on-premises relational databases to AWS services like Amazon RDS or Amazon Aurora

➔ **Heterogeneous Data Ingestion :**
This pattern applies when the source and target data stores are different types, requiring transformation during the ingestion process. For example, ingesting data from a relational database to a NoSQL database like Amazon DynamoDB

➔ **Data Lake Ingestion :**
AWS offers various services and capabilities to ingest diverse data types into a data lake built on Amazon S3. This method supports batch processing, streaming data, and real-time data ingestion to consolidate data from multiple sources

# Data transfer services and usecases

| AWS Service | Details | Usecases |
|---|---|---|
| **AWS DataSync** | To automate data transfer between on-premises & AWS storage services | Analysis, backup or archiving data in the cloud<br>Migrating active data sets |
| **AWS Snowball** | Petabyte-scale data transport solution, Uses secure appliances to transfer huge data | Large-scale data migrations |
| **AWS Snowmobile** | Exabyte-scale data transfer service using a 45-foot long ruggedized shipping container | Migrating massive data sets up to 100 PB per Snowmobile |
| **AWS Storage Gateway** | Hybrid cloud storage service providing on-premises applications access to virtually unlimited cloud storage | Backup and restore functions.<br>Tiered storage capabilities. |
| **Amazon S3 Transfer Acceleration** | Accelerates content transfers to\from S3 using CloudFront's globally distributed edge locations | Speeding up the transfer of data to S3 across long distances |
| **AWS Transfer Family** | For file transfers directly into and out of Amazon S3 using SFTP, FTPS and FTP | Securely transferring files for internal and external stakeholders |

# Streaming data services and usecases

Streaming data services in AWS include Amazon Kinesis, which offers various services for collecting, processing and analyzing streaming data.
Different use cases :

➔ **Data analysis** : Real-time processing of data streams for immediate insights
➔ **IoT applications** : Handling large volumes of data from sensors and devices in real-time
➔ **Financial analysis** : Processing financial transactions and market data in real-time for timely decision-making
➔ **Real-time recommendations** : Providing personalized recommendations based on user behavior and preferences
➔ **Service guarantees** : Monitoring service health and performance metrics in real-time to ensure SLA compliance
➔ **Media and gaming** : Handling real-time data streams for interactive media and gaming applications

# Secure access to ingestion access points

To ensure secure access to ingestion access points in AWS:

➔ **Use SSL Encryption**
  - Implement Secure Socket Layer (SSL) encryption to protect data while it's in transit to AWS ingestion points like Amazon S3 and Kinesis

➔ **Leverage AWS Transfer Family**
  - AWS Transfer Family supports Secure Shell (SSH) File Transfer Protocol (FTP), FTP Secure (FTPS), and FTP for data ingestion
  - Utilize these secure transfer protocols to ingest data securely

# Data transformation services and usecases

| AWS Services | Description | Use-cases |
|---|---|---|
| **AWS Glue** | Serverless data integration service for automating the ETL process. | Data cleaning, enrichment and normalization |
| **AWS Lambda** | Execute code in response to events for data transformations | Real-time data processing before loading into data stores like Amazon Redshift |
| **AWS EMR** | Cloud big data platform for running large-scale distributed data processing jobs. | Transforming data into Parquet format in data lakes, large-scale data analysis |
| **Amazon SageMaker Data Wrangler** | Interface for data scientists to prepare data for machine learning models. | Data cleaning, transformation, and featurization for ML |
| **AWS Data Pipeline** | Web service to process and move data between different AWS compute and storage services | Data transformation during data transfer |
| **AWS Glue DataBrew** | Visual data preparation tool that allows users to clean and normalize data | Common data preparation transformations such as normalization, deduplication and data type conversion |

# Building and securing data lakes

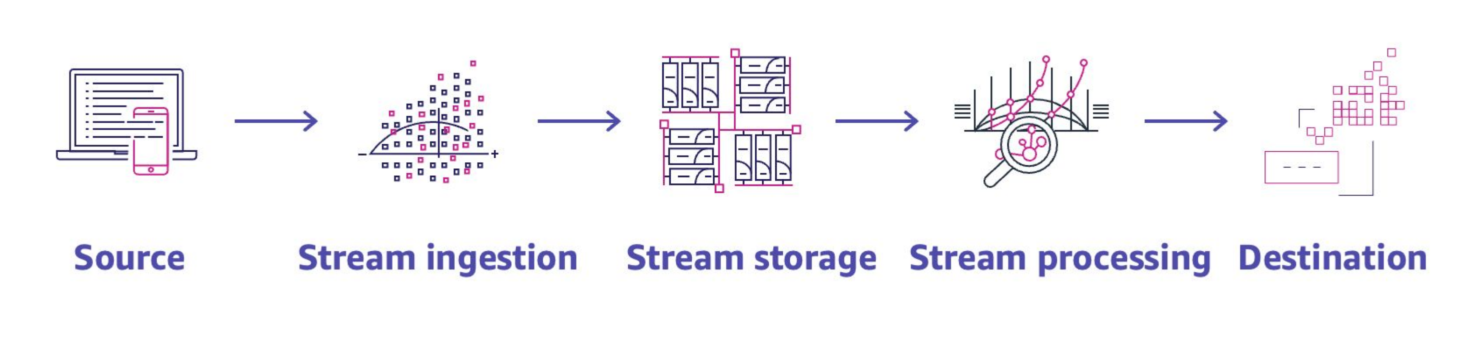
Data lakes in AWS can be built and secured effectively using AWS Lake Formation which simplifies and automates many tasks involved in creating and managing data lakes -

- **Building Data Lakes**
  - Utilize AWS Lake Formation to streamline the process of building data lakes, allowing you to centralize diverse data sources into a unified repository
- **Securing Data Lakes**
  - AWS Lake Formation enables you to enforce security policies consistently across various data sources within the data lake
  - It supports fine-grained access controls, ensuring data security
- **Encryption**
  - Ensure data security by encrypting data within the data lake
  - AWS Lake Formation supports encryption to safeguard data at rest and in transit
- **Managing Data**
  - AWS provides guidelines for managing data within data lakes, emphasizing the importance of organizing and cataloging data effectively for improved accessibility and governance

# Designing Data Streaming Architectures

Streaming data architecture can be designed as a stack of five logical layers, Each layer is composed of multiple purpose-built components that address specific requirements

- ➔ **Source** - Includes data sources like sensors, social media, IoT devices, log files generated by using web and mobile applications, mobile devices that generates semi-structured and unstructured data
- ➔ **Stream storage** - Responsible for providing scalable and cost-effective components to store streaming data. It can be stored in the order it was received for a set duration of time and can be replayed indefinitely
- ➔ **Stream ingestion** - Responsible for ingesting data into the stream storage layer. It provides the ability to collect data from tens of thousands of data sources and ingest in near real-time
- ➔ **Stream processing** - Responsible for transforming data into a consumable state through data validation, cleanup, normalization, transformation and enrichment
- ➔ **Destination** - It can be an event driven application, data lake, data warehouse, database or an OpenSearch



# Visualization Strategies

AWS offers a variety of services and tools to facilitate data visualization, helping users to create meaningful insights from their data.Key visualization strategies -

- ➔ **Amazon QuickSight** - Business Analytics , Real time data monitoring , Interactive reporting
- ➔ **AWS Glue Databrew** - Data Cleaning , Transformation for visualization , Preparation of data for ML models
- ➔ **Amazon Athena** - Ad Hoc data analysis , Quick insights from S3 data , Integration with QuickSight for visualization
- ➔ **AWS SageMaker Studio** - Data exploration , Model training Insights , Feature engineering
- ➔ **Amazon Managed Grafana** - Monitoring Infrastructure , Tracking application performance and Operational Insights
- ➔ **AWS CloudWatch Dashboards** - Real Time operational monitoring , Alert visualization , Metric tracking

# Compute Options for data processing

- **Amazon EC2**
  - Provides resizable compute capacity in the cloud, allowing users to scale up or down as needed
  - Suitable for applications requiring significant control over the computing environment
- **AWS Lambda**
  - Serverless compute service that automatically manages infrastructure needed to run the code
  - Ideal for data processing tasks that gets executed in response to events
- **Amazon EMR(Elastic MapReduce)**
  - Managed cluster platform that simplifies running big data frameworks (Ex-Apache Hadoop and Spark)
  - Perfect for large-scale data processing and analysis
- **AWS Fargate**
  - Serverless compute engine for containers that works with Amazon ECS and Amazon EKS
  - Run containers without managing the underlying infrastructure
  - Suitable for containerized applications and microservices
- **Amazon Redshift**
  - Data warehouse service that can handle petabyte-scale data
  - Used for running complex queries and data analytics
- **Amazon Glue**
  - ETL (Extract, Transform and Load) service , Prepare and load data for analytics
  - Useful for data integration and preparation

# AWS Solution Architect Associate

Version : C03

Domain 3

Task 5

## The END