

## **AWS Overview & Architecture**

Intro to Cloud Computing, AWS Global Infrastructure, Regions, AZs, Core Services (Compute, Storage, Networking), AWS Shared Responsibility Model

## **IAM**

Users, Groups, Roles, Policies, Best Practices (Least Privilege), MFA, Policy Types (Managed, Inline), Service Control Policies (SCPs)

## **S3 for Data Engineering**

Buckets, Objects, Storage Classes (Standard, IA, Intelligent-Tiering, Glacier), Lifecycle Policies, Versioning, Security (Bucket Policies, ACLs), Data Organization (Prefixes, Partitioning), Performance Optimization

## **S3 Advanced**

Pre-signed URLs, Access Control Lists (ACLs), Cross-Region Replication, S3 Select, S3 Object Lambda, S3 Event Notifications

## **Lambda**

Intro to Serverless, Lambda Functions, Event Triggers, Deployment (using CLI, Console, SAM), Function Packaging (Zip files, Container Images), Lambda Layers  
API Gateway Integration, Environment Variables, Layers, Concurrency Control, Asynchronous Invocation, Error Handling, Monitoring and Logging

## **SQS**

Queues (Standard, FIFO), Message Processing, Visibility Timeout, Dead-Letter Queues, Message Batching, Long Polling

## **SNS**

Topics, Subscriptions, Message Filtering, Fan-out Pattern, Message Attributes, Delivery Retry Policies, Monitoring

## **CloudWatch**

Metrics, Logs, Alarms, Dashboards, Monitoring Lambda, SQS, EC2, Custom Metrics, Log Insights

## **EMR**

EMR Architecture, Cluster Creation (using Console, CLI, CloudFormation), Running Spark on EMR, Hadoop Ecosystem (HDFS, YARN)  
Cluster Configuration (Instance Types, Core/Task Nodes), Spot Instances, Security (IAM Roles, Security Groups), EMRFS, Hive, Pig

## **AWS Glue**

Glue Components (Crawler, Job, Workflow), Data Catalog, DataBrew, Glue Studio

## **Glue ETL**

Building ETL Pipelines with Glue, Data Transformations (using PySpark or Scala), Job Scheduling, Dynamic Frames, Error Handling

## **Kinesis**

Kinesis Streams, Shards, Data Ingestion, Kinesis Firehose, Kinesis Data Analytics, Kinesis Data Streams vs. Firehose Real-time Data Processing with Kinesis, Kinesis Data Streams vs. Firehose, Shard Management, Data Serialization (Avro, Parquet)

## **Data Modeling (Conceptual)**

Data Modeling Concepts, Entity-Relationship Diagrams (ERDs), Normalization (1NF, 2NF, 3NF), Data Vault Modeling

## **Data Modeling (Practical)**

Schema Design for Data Warehouses and Data Lakes, Star Schema, Snowflake Schema, Data Lake Schema Design Patterns

## **Airflow**

Intro to Airflow, DAGs, Operators (BashOperator, PythonOperator, EmailOperator), Scheduling, Variables and Connections, Task Dependencies Advanced DAG Concepts, Branching, SubDAGs, Task Dependencies, XComs, Trigger Rules, Backfilling Sensors, Operators for different data stores (S3, Snowflake, Postgres), Best practices for Airflow DAG development, Monitoring and Alerting, Security

## **Athena**

Querying Data in S3 using SQL, Data Partitioning, Performance Optimization, CTAS, Views, Data Definition Language (DDL)