AWS ML Associate — Section 1.1: Collect, Ingest, and Store Data

This section covers how to bring data into AWS, store it securely, and prepare it for ML workflows in SageMaker. Key areas include sources, ingestion methods, storage options, cataloging/querying, governance, and SageMaker consumption.

# Data Sources

• Batch files: CSV, Parquet, JSON, logs

• Databases: on‑prem, Amazon RDS/Aurora (SQL), DynamoDB (NoSQL)

• Streaming: IoT devices, clickstreams, apps

• External: SFTP partners, SaaS exports

# Ingestion Tools

• Kinesis Data Streams / Firehose — real-time streaming to S3/Redshift

• AWS Glue — ETL, schema discovery, job orchestration

• Database Migration Service (DMS) — move relational DBs into AWS

• AWS Transfer Family — managed SFTP/FTP for file ingestion

• Snowball — bulk offline ingestion

# Storage (Landing & Curated)

• Amazon S3 — primary ML data lake, versioning, lifecycle mgmt

• Amazon Redshift — data warehouse, Spectrum queries on S3

• Amazon RDS/Aurora — relational queries

• Amazon DynamoDB — NoSQL key-value for features

# Catalog & Query

• Glue Data Catalog — schemas, partitions, metadata

• Athena / Redshift Spectrum — SQL queries on S3

• Lake Formation — fine-grained permissions

# Security & Governance

• IAM roles, bucket policies — access control

• Encryption — SSE-S3 (managed), SSE-KMS (customer keys)

• Network isolation — VPC endpoints, PrivateLink

# SageMaker Consumption

• Training input channels from S3 — Pipe (stream) vs File (copy)

• Processing jobs / Data Wrangler — feature engineering & prep

• Ground Truth — managed data labeling service

• Model Registry — manage versions, deploy for batch/real-time inference

# Exam Tips

• Default landing zone is S3; always think S3 → Glue → Athena/Redshift → SageMaker

• Kinesis/Firehose for streams; DMS for DB migrations; Glue for ETL

• Prefer Pipe mode for large datasets in SageMaker

• Always consider cost (S3 cheapest), scalability, and encryption by default