AWS Certified Machine Learning Engineer - Associate Exam Guide

# Section 1.1 — Ingest and Store Data (In-Depth)

This section covers how to collect, ingest, and store data in AWS for ML workloads, with emphasis on formats, ingestion methods, storage, cataloging, security, and SageMaker integration.

## 1. Data Formats

• Structured: CSV, TSV, Parquet, ORC (efficient columnar), Avro

• Semi-structured: JSON, XML

• ML-specific: RecordIO (used by MXNet, optimized for sequential read/write), TFRecords

Why it matters: Columnar formats (Parquet/ORC) reduce query time and storage, while RecordIO/TFRecords are optimized for ML pipelines.

## 2. Storage Options

• Amazon S3 → Core data lake, cheap, durable, integrates with all ML services

• Amazon EFS → NFS-style, concurrent access, model/data sharing

• Amazon FSx → Managed file systems (Lustre = high-performance, integrates with S3)

• Amazon RDS/Aurora → Structured relational data

• Amazon DynamoDB → NoSQL, key-value/document-based storage

## 3. Data Ingestion Methods

Batch ingestion:

• AWS Glue → ETL service

• AWS DMS → migrate relational/NoSQL data

• AWS Snowball/Snowmobile → petabyte-scale physical transfer

Streaming ingestion:

• Amazon Kinesis Data Streams → real-time streaming (logs, IoT, clicks)

• Amazon Kinesis Firehose → delivers to S3/Redshift/ES directly

• Amazon MSK (Managed Kafka) → Kafka-compatible ingestion

• Apache Flink on Kinesis Data Analytics → real-time transformations

Other ingestion:

• AWS Transfer Family → SFTP/FTPS/FTP ingestion into S3

• API Gateway + Lambda → custom ingestion

## 4. Cataloging & Querying

• AWS Glue Data Catalog → central metadata repository

• Amazon Athena → SQL queries directly on S3 data

• Redshift Spectrum → query S3 data from Redshift

• AWS Lake Formation → secure data lake with fine-grained access control

## 5. Security & Governance

• IAM → least-privilege access control

• Encryption: SSE-S3, SSE-KMS, TLS in transit

• Networking: VPC endpoints, PrivateLink

• Data compliance: handle PII/PHI, residency requirements

## 6. SageMaker Integration

• Input channels: Pipe mode (streaming, large datasets) vs File mode (downloads full dataset)

• Data Wrangler → ingest, clean, prepare data visually

• Feature Store → central store for features ensuring train/inference consistency

• Ground Truth → labeling service

• Model Registry → track lineage and reproducibility

## ✅ Exam Tips

• S3 is always the hub → most workflows start with S3

• Pipe mode is preferred for large datasets

• Use Parquet/ORC for frequent queries

• Feature Store ensures feature consistency

• Kinesis Firehose is default for continuous streaming ingestion

• Always encrypt & enforce least privilege with IAM