AWS Certified Machine Learning Engineer - Associate Exam Guide

# Section 1.1 — Ingest and Store Data

This section focuses on the ability to collect, ingest, and store data in AWS for machine learning workloads. Candidates should understand supported data formats, ingestion pipelines, storage options, and cataloging tools.

## Knowledge

• Common data formats: CSV, JSON, Parquet, ORC, Avro, RecordIO

• AWS storage: Amazon S3, Amazon EFS, Amazon FSx

• Streaming ingestion: Amazon Kinesis, Amazon MSK (Kafka), Apache Flink

• Storage trade-offs: cost, performance, structure

## Skills

• Extract data from sources (S3, EBS, EFS, RDS, DynamoDB) using appropriate AWS features

• Choose data formats aligned to access patterns

• Ingest data into SageMaker Data Wrangler and SageMaker Feature Store

• Merge multi-source data with AWS Glue or Apache Spark on EMR

• Troubleshoot ingestion and capacity/performance bottlenecks

• Make initial storage decisions balancing cost, performance, and scalability