AWS ML Associate — Section 3.1: Deploy Models

This section covers deploying trained models into production environments using SageMaker and related AWS services. It focuses on real-time, batch, asynchronous, and edge deployments.

# Deployment Options

• Real-time inference — low-latency predictions via SageMaker Endpoints

• Batch inference — run predictions on large datasets using SageMaker Batch Transform

• Asynchronous inference — handle long-running requests efficiently

• Edge deployment — SageMaker Neo compiles and optimizes models for edge devices

# Exam Tips

• Use real-time endpoints for millisecond-latency predictions

• Use batch transform for offline, large-scale predictions

• Asynchronous inference is for requests that take minutes to process

• SageMaker Neo = optimization/compilation for edge deployments