Lesson Overview

Note: Multi-AZ is a term specific to AWS. AZ stands for availability zones.

At this point, we should feel comfortable with training a machine learning model and making use of it to produce inferences. But this may have been done in an exploratory environment or may have required a significant amount of manual effort to fully execute.

In this lesson, we will learn how to leverage Amazon Sagemaker to:

- **Create Training Jobs**
  - To produce not just one, but a series of trained models, we will want to parameterize as much as possible with regards to the input datasets and output artifacts in S3, and which instances to use.
- **Deploy Model Endpoints**
  - After training a model, we deploy the model to a persistent endpoint to listen for new inference requests using Sagemaker hosting services.
- **Process Batch Transforms**
  - For cases when we have an *entire* dataset to be inferred by the trained model, we create a batch transform job using Sagemaker and S3.

- **Create Processing Jobs**
  - To simplify and automate the pre-processing of inputs or post-processing of outputs, we can leverage processing jobs within Sagemaker to manage much of the extracting, transforming, and loading involved.

- **Debug SageMaker**
  - All of these processes mentioned are tough to initially implement. We can leverage tools in AWS to help us debug AWS SageMaker services.