Importing Libraries

Spark comes equipped with a selection of libraries, including Spark SQL, Spark Streaming, and MLlib. However, if you want to use a custom library with a Spark application (such as a compression library or Magellan), you can use one of the following two spark-submit script options:

- The --jars option transfers associated .jar files to the cluster.
  Specify a list of comma-separated .jar files.
- The --packages option pulls files directly from Spark packages.
  This approach requires an internet connection.

Spark Machine Learning Library (MLlib)

- MLlib in Spark's machine learning (ML) library. Its goal is to make practical machine learning scalable and easy.
  - **ML Algorithms:** common learning algorithms such as classification, regression, clustering, and collaborative filtering
  - **Featurization:** feature extraction, transformation, dimensionality reduction, and selection
  - **Pipelines:** tools for constructing, evaluating, and tuning ML Pipelines
  - **Persistence:** saving and load algorithms, models, and Pipelines
  - **Utilities:** linear algebra, statistics, data handling, etc.
- DataFrame-based API is primary API
- The MLlib RDD-based API is now in maintenance mode.
- As of Spark 2.0, the RDD-based APIs in the spark.mllib package have entered maintenance mode.
- The primary Machine Learning API for Spark is now the DataFrame-based API in the spark.ml package.

Why is MLlib switching to the DataFrame-based API?

- DataFrames provide a more user-friendly API than RDDs. The many benefits of DataFrames include Spark Datasources, SQL/DataFrame queries, Tungsten and Catalyst optimizations, and uniform APIs across languages.
- The DataFrame-based API for MLlib provides a uniform API across ML algorithms and across multiple languages.
- DataFrames facilitate practical ML Pipelines, particularly feature transformations. See the Pipelines guide for details.