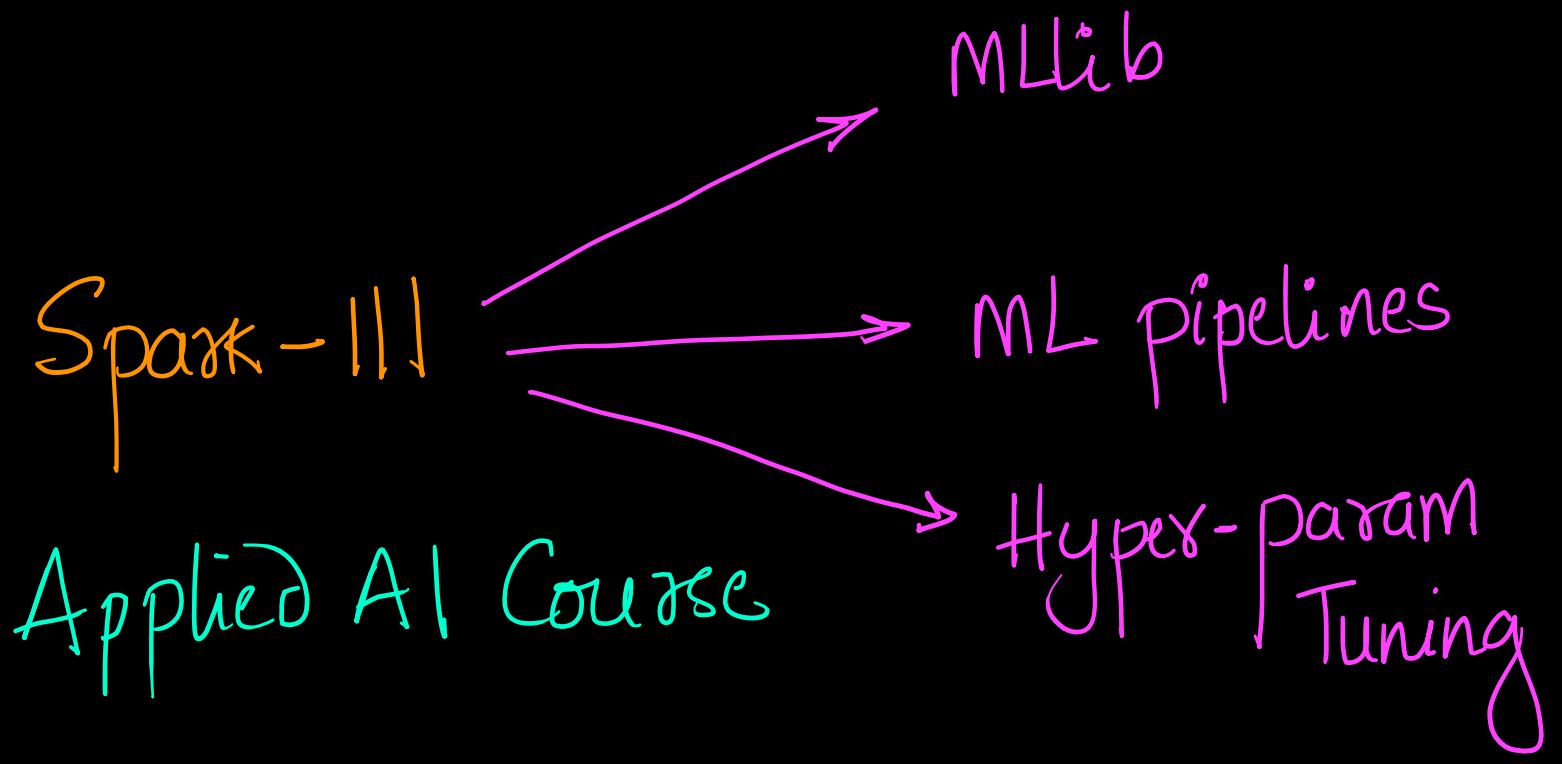

pySpark MLlib

https://spark.apache.org/docs/latest/api/python/pyspark.mllib.html

https://spark.apache.org/docs/2.1.0/mllib-data-types.html#labeled-point

Classification, Regression, Clustering, MF

↓

code-walkthroughs

ML - pipeline

https://spark.apache.org/docs/latest/ml-pipeline.html#code-examples

Spark-3-Session.ipnb

- **DataFrame**: This ML API uses DataFrame from Spark SQL as an ML dataset, which can hold a variety of data types. E.g., a DataFrame could have different columns storing text, feature vectors, true labels, and predictions.

- **Transformer**: A Transformer is an algorithm which can transform one DataFrame into another DataFrame. E.g., an ML model is a Transformer which transforms a DataFrame with features into a DataFrame with predictions.

- **Estimator**: An Estimator is an algorithm which can be fit on a DataFrame to produce a Transformer. E.g., a learning algorithm is an Estimator which trains on a DataFrame and produces a model.

- **Pipeline**: A Pipeline chains multiple Transformers and Estimators together to specify an ML workflow.

- **Parameter**: All Transformers and Estimators now share a common API for specifying parameters.

Hyperparam Tuning

https://spark.apache.org/docs/latest/ml-tuning.html

Spark-3-session.ipynb