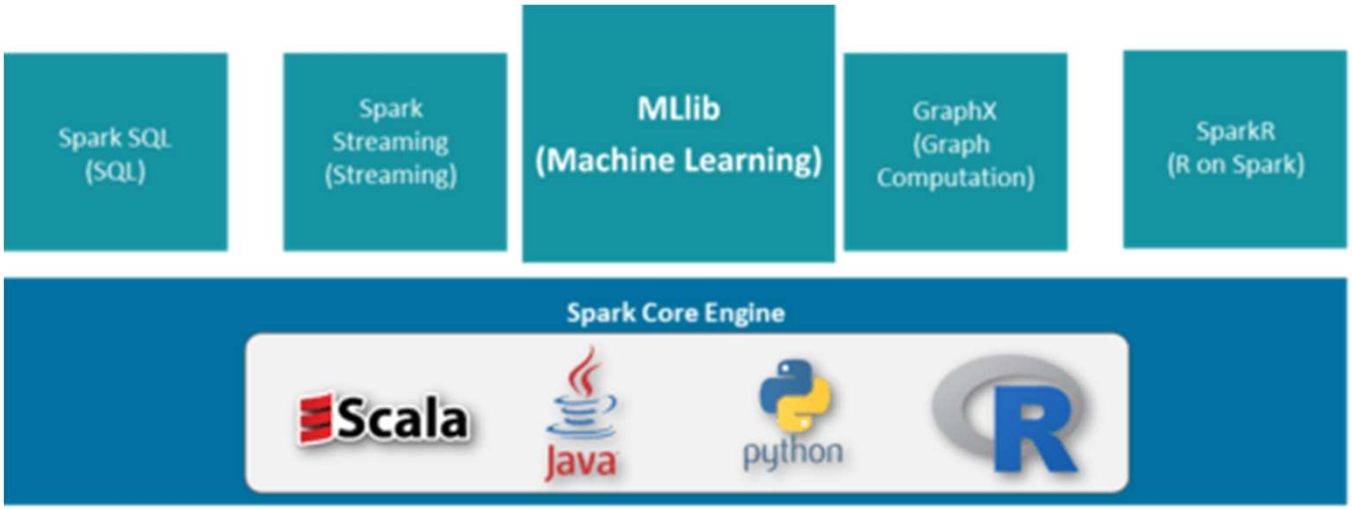
Spark MLLib

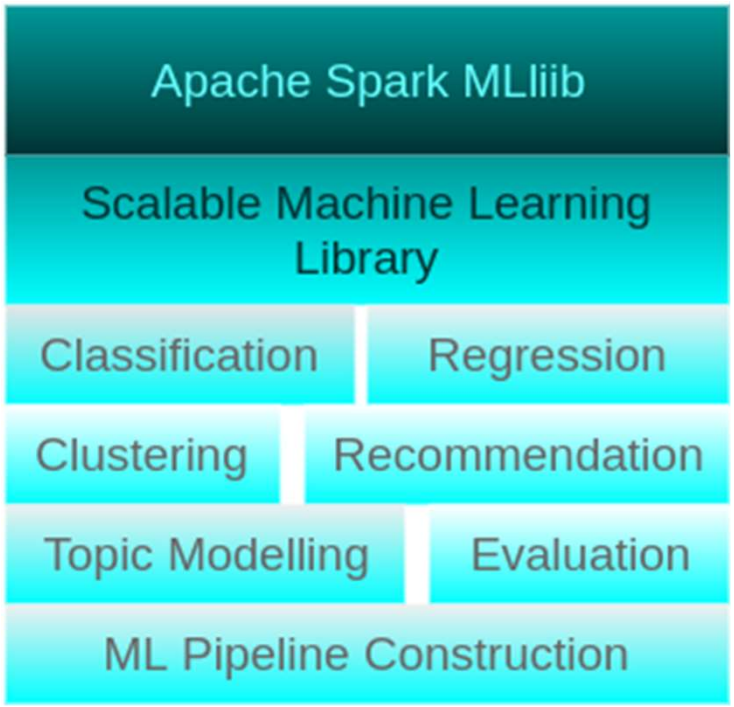
What is PySpark MLlib?

- A machine-learning library
- A wrapper over PySpark Core to do data analysis using machine-learning algorithms
- Has implementations of
  - Classification
  - Clustering
  - Linear regression



MLlib Features

- Data Source
  - HDFS
  - HBase
  - Local Files
- Excels at iterative computation



Algorithms

- Classification using Logistic Regression
- Classification using Naive Bayes
- Generalized Regression
- Survival Regression
- Decision Trees
- Random Forests
- Gradient Boosted Trees
- Recommendation using Alternating Least Squares (ALS)
- Clustering using KMeans
- Clustering using Gaussian Mixtures
- Topic Modelling using Latent Dirichlet Conditions
- Frequent Itemsets
- Association Rules
- Sequential Pattern Mining

MLlib Utilities

- Feature Transformation
- ML Pipeline construction
- Model Evaluation
- Hyper-parameter tuning
- Saving and loading of models and pipelines
- Distributed Linear Algebra
- Statistics

MLlib vs scikit-learn

- Scikit-Learn has fantastic performance if your data fits into RAM
- Spark's ML Lib is suitable when you're doing relatively simple ML on a large data set
- ML Lib is not computationally efficient for small data sets, and you're better off using scikit-learn for small and medium sized data sets (megabytes, up to a few gigabytes). For much larger data sets, use Spark ML.
- MLlib lacks in visualization
- Sklearn is far richer in terms of decent implementations of a large number of commonly used algorithms as compared to spark mllib
- Scikit-Learn integrates very well with Pandas dataframes and can work on Numpy and Pandas data structures. This is not the case with Spark - where you have to use Spark's native dataframes or RDDs to execute the computation.

Various ML algorithms supported by MLlib

- Classification
  - Logistic regression
    - Binomial logistic regression
    - Multinomial logistic regression
  - Decision tree classifier
  - Random forest classifier
  - Gradient-boosted tree classifier
  - Multilayer perceptron classifier
  - Linear Support Vector Machine
  - One-vs-Rest classifier (a.k.a. One-vs-All)
  - Naive Bayes

Various ML algorithms supported by MLlib

- Regression
  - Linear regression
  - Generalized linear regression
  - Decision tree regression
  - Random forest regression
  - Gradient-boosted tree regression
  - Survival regression
  - Isotonic regression
- Decision trees
- Tree Ensembles
  - Random Forests
  - Gradient-Boosted Trees (GBTs)

Various ML algorithms supported by MLlib

- Clustering
  - K-means
  - Bisecting k-means
  - Gaussian Mixture Model (GMM)
- Collaborative filtering

Thanks