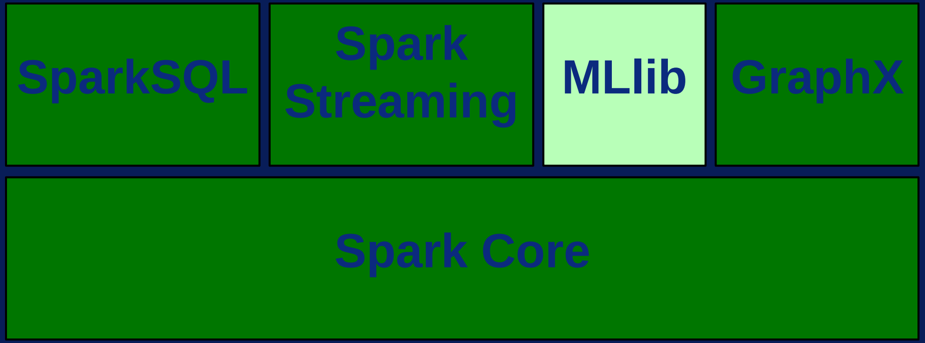
Spark MLlib

After this video you will be able to..

- Describe what MLlib is
- List main categories of techniques available in MLlib.
- Explain code segments containing MLlib algorithms.

Spark MLlib



SparkSQL

Spark Streaming

MLlib

GraphX

Spark Core

- Scalable machine learning library
- Provides distributed implementations of common machine learning algorithms and utilities
- Has APIs for Scala, Java, Python, and R

MLlib Algorithms & Techniques

- Machine Learning
  - Classification, regression, clustering, etc.
  - Evaluation metrics
- Statistics
  - Summary statistics, sampling, etc.
- Utilities
  - Dimensionality reduction, transformation, etc.

MlLib Example – Summary Statistics

- Compute column summary statistics

```
from pyspark.mllib.stat import Statistics        1

# Data as RDD of Vectors
dataMatrix = sc.parallelize([ [1, 2, 3], [4, 5, 6], [7, 8, 9], [10, 11, 12] ])   2


# Compute column summary statistics.
summary = Statistics.colStats(dataMatrix)        3
print(summary.mean())
print(summary.variance())                        4
print(summary.numNonzeros())
```

MLlib Example – Classification

- Build decision tree model for classification

```
from pyspark.mllib.tree import DecisionTree, DecisionTreeModel    1
from pyspark.mllib.util import MLUtils                            2

# Read and parse data
data = sc.textFile(“data.txt”)    3


# Decision tree for classification
model = DecisionTree.trainClassifier                  4
              (parsedData, numClasses=2)
print(model.toDebugString())                          5
model.save(sc, “decisionTreeModel”)                   6
```

MLlib Example – Clustering

- Build k-means model for clustering

```
from pyspark.mllib.clustering import KMeans, KMeansModel    1
from numpy import array
                                                            2
# Read and parse data
data = sc.textFile(“data.txt”)                              3
parsedData = data.map(lambda line:
        array([float(x) for x in line.split(‘ ‘)]))

# k-means model for clustering
clusters = Kmeans.train (parsedData, k=3)                   4

print(clusters.centers)                                     5
```

Main Take-Aways

- MLlib is Spark’s machine learning library.
  - Distributed implementations
- Main categories of algorithms and techniques:
  - Machine learning
  - Statistics
  - Utility for ML pipeline