Spark Ecosystem

Goals

- Scala Spark vs PySpark
- Spark Libraries and MLlib

Goals

- Scala Spark vs PySpark
- Spark Libraries and MLlib

Spark vs PySpark

- Spark is written in Scala
- PySpark is slower (Dynamically typed)
- More libraries in Spark (GraphX)
- Spark RDDs are statically typed

Static Typing

Spark RDD

Key	Value
1	'hey'
2	'go'
3	'yeah'

PySpark RDD

Key	Value
1	'hey'
1.0	'go'
[1, 2]	'yeah'

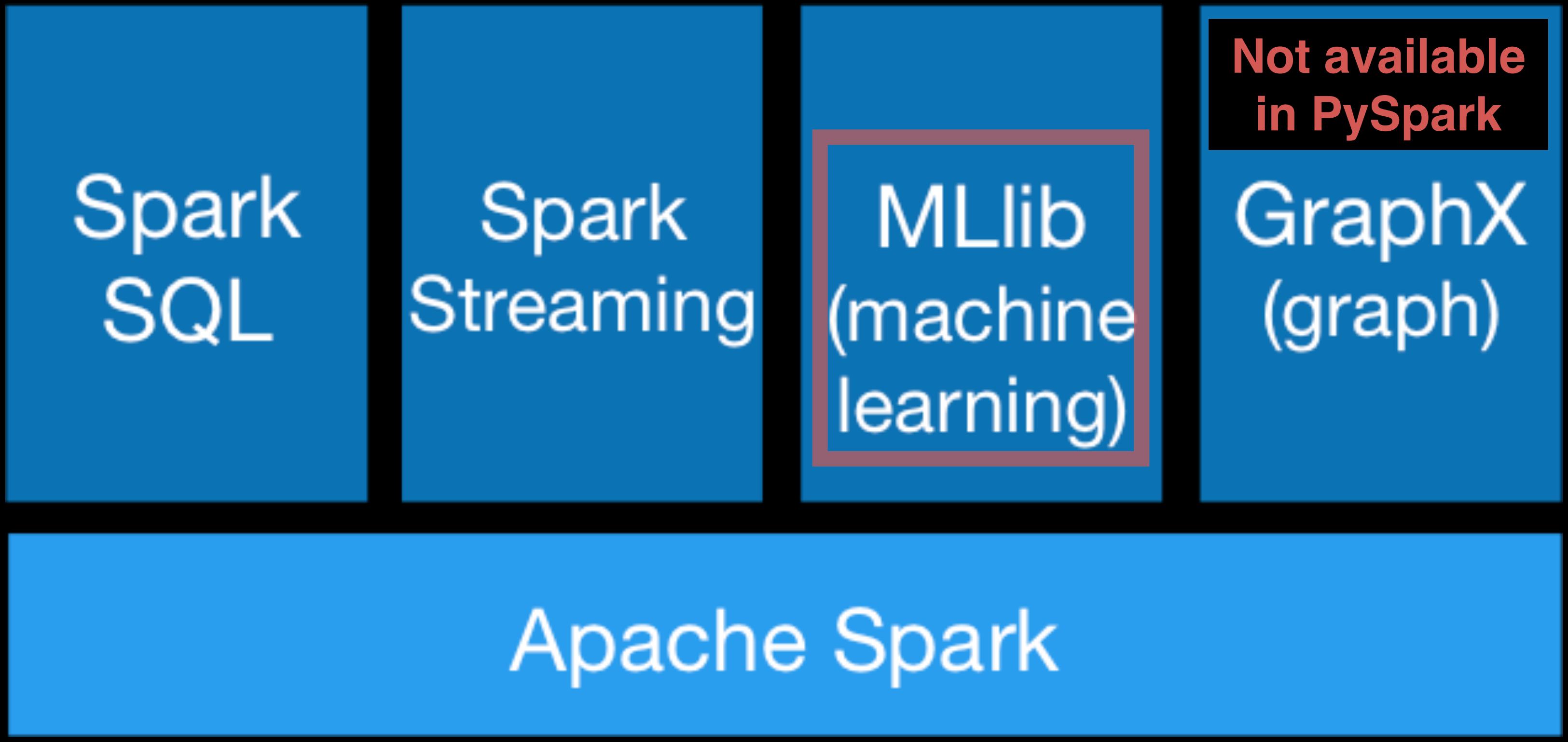
PySpark

- Calls functions in Scala Spark
- PySpark acts as a client that sends command to Scale Spark
- Scala Spark acts as a server receives commands and executes it

Goals

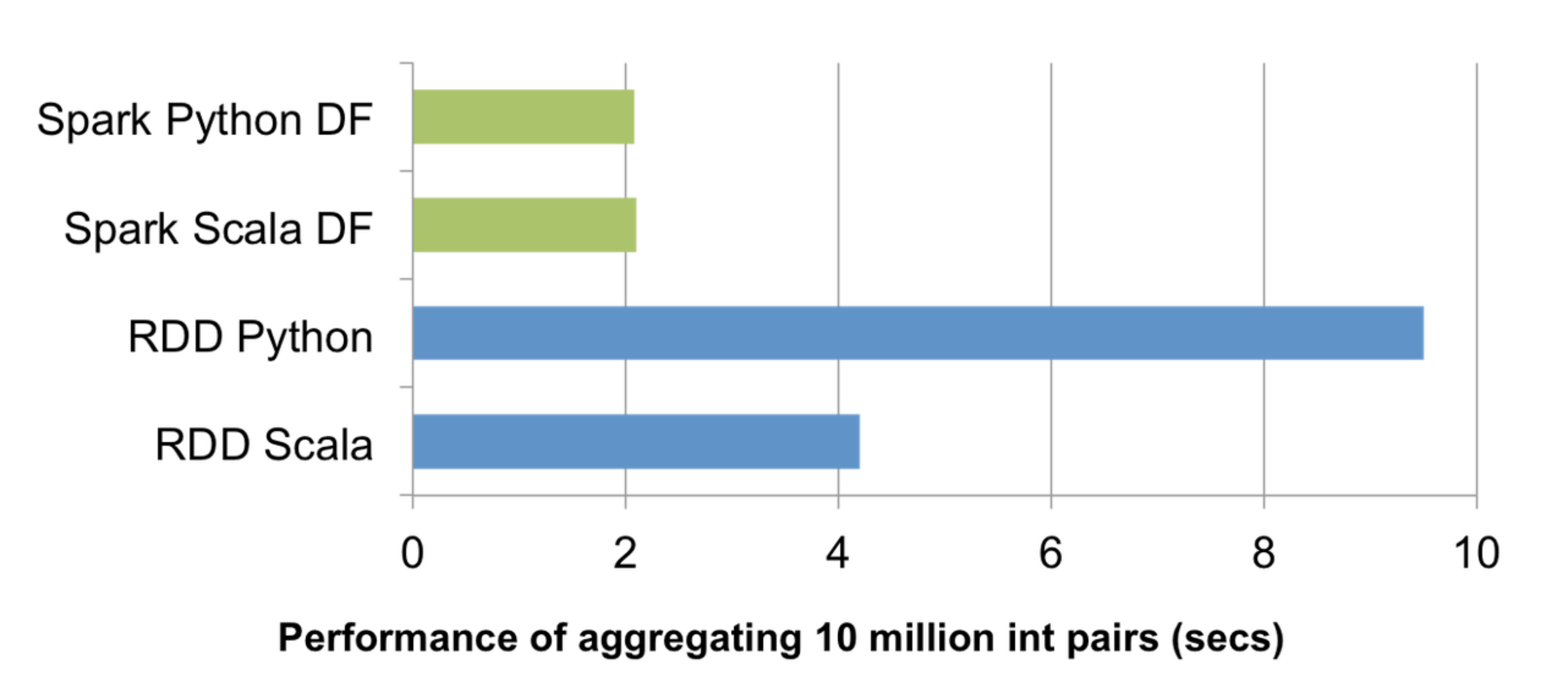
- Scala Spark vs PySpark
- Spark Libraries and MLlib

Spark Libraries



SQL / DataFrames

- Since v1.3, there is DataFrame support



ML-lib

- **Classification**
  - ⋆ Logistic Regression, SVM, Naive Bayes, GradientBoostedTrees
- **Regression**
  - ⋆ Generalized linear regression (GLM)
- **Recommender**
  - ⋆ NMF
- **Clustering:**
  - ⋆ k-means
- **Decomposition**
  - ⋆ SVD & PCA

MLlib Scala Source

https://github.com/apache/spark/tree/master/mllib/src/main/scala/org/apache/spark/mllib

Conventions

- For Supervised learning
  - ★ `LabeledPoints(target, feature)`
  - ★ `target` (numeric)
  - ★ `feature` (numeric vector)