What is MLlib?

What is MLlib?

MLlib is a Spark subproject providing machine learning primitives:

- initial contribution from AMPLab, UC Berkeley
- shipped with Spark since version 0.8
- 35 contributors

What is MLlib?

Algorithms:

- **classification:** logistic regression, linear support vector machine (SVM), naive Bayes, classification tree
- **regression:** generalized linear models (GLMs), regression tree
- **collaborative filtering:** alternating least squares (ALS)
- **clustering:** k-means
- **decomposition:** singular value decomposition (SVD), principal component analysis (PCA)

Why MLlib?

scikit-learn?

Algorithms:

- **classification:** SVM, nearest neighbors, random forest, …
- **regression:** support vector regression (SVR), ridge regression, Lasso, logistic regression, …
- **clustering:** k-means, spectral clustering, …
- **decomposition:** PCA, non-negative matrix factorization (NMF), independent component analysis (ICA), …

Mahout?

Algorithms:

- **classification:** logistic regression, naive Bayes, random forest, …
- **collaborative filtering:** ALS, …
- **clustering:** k-means, fuzzy k-means, …
- **decomposition:** SVD, randomized SVD, …

Mahout?

LIBLINEAR?

H_2O?

Vowpal Wabbit?

MATLAB?

R?

scikit-learn?

Weka?

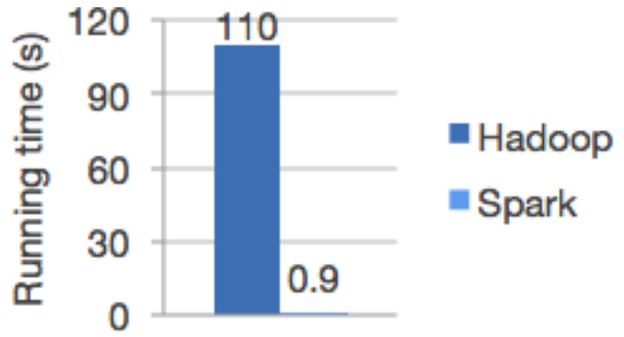
GraphLab?

Why MLlib?

Why MLlib?

- It is built on Apache Spark, a fast and general engine for large-scale data processing.

- Run programs up to 100x faster than Hadoop MapReduce in memory, or 10x faster on disk.



- Write applications quickly in Java, Scala, or Python.

Spark philosophy

Make life easy and productive for data scientists:

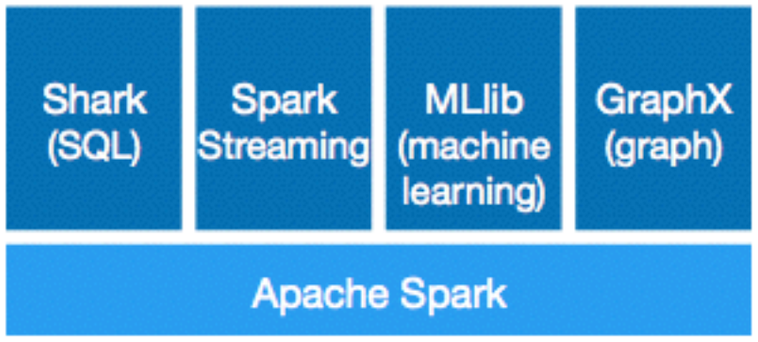
- Well documented, expressive API's
- Powerful domain specific libraries
- Easy integration with storage systems
- … and caching to avoid data movement

k-means (python)

```
# Load and parse the data
data = sc.textFile("kmeans_data.txt")
parsedData = data.map(lambda line:
                array([float(x) for x in line.split(' ')])).cache()

# Build the model (cluster the data)
clusters = KMeans.train(parsedData, 2, maxIterations = 10,
                runs = 1, initialization_mode = "kmeans||")

# Evaluate clustering by computing the sum of squared errors
def error(point):
    center = clusters.centers[clusters.predict(point)]
    return sqrt(sum([x**2 for x in (point - center)]))

cost = parsedData.map(lambda point: error(point))
          .reduce(lambda x, y: x + y)
print("Sum of squared error = " + str(cost))
```

Why MLlib?

- It ships with Spark as a standard component.



DATABRICKS

Why MLlib?

A special-purpose device may be better at one aspect than a general-purpose device. But the cost of context switching is high:

- different languages or APIs
- different data formats
- different tuning tricks

Spark SQL + MLlib

```
// Data can easily be extracted from existing sources,
// such as Apache Hive.
val trainingTable = sql("""
  SELECT e.action,
         u.age,
         u.latitude,
         u.longitude
  FROM Users u
  JOIN Events e
  ON u.userId = e.userId""")

// Since `sql` returns an RDD, the results of the above
// query can be easily used in MLlib.
val training = trainingTable.map { row =>
  val features = Vectors.dense(row(1), row(2), row(3))
  LabeledPoint(row(0), features)
}

val model = SVMWithSGD.train(training)
```

DATABRICKS

Why MLlib?

- Built on Spark's lightweight yet powerful APIs.
- Spark's open source community.
- Seamless integration with Spark's other components.
- Comparable to or even better than other libraries specialized in large-scale machine learning.

Why MLlib?

- Scalability
- Performance
- User-friendly APIs
- Integration with Spark and its other components

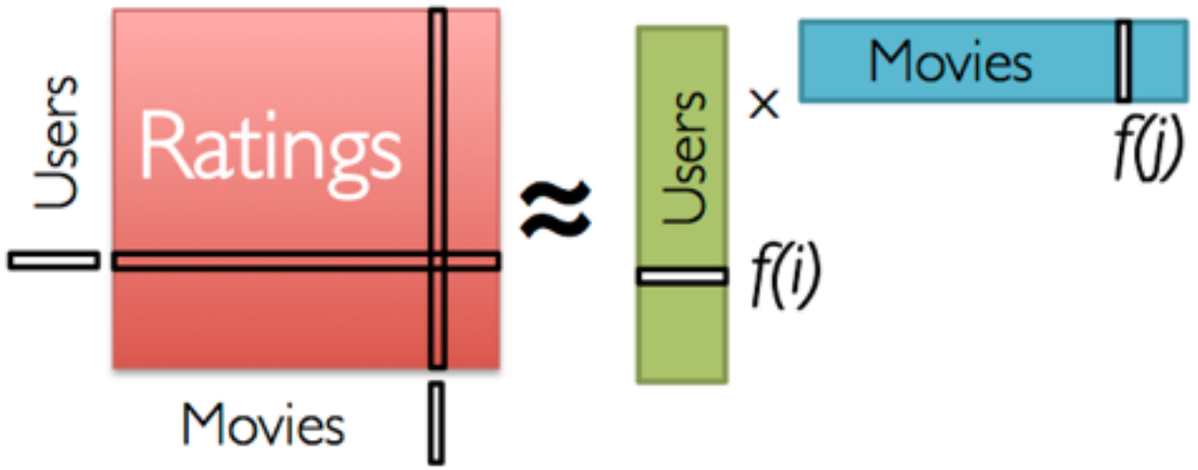
Collaborative filtering

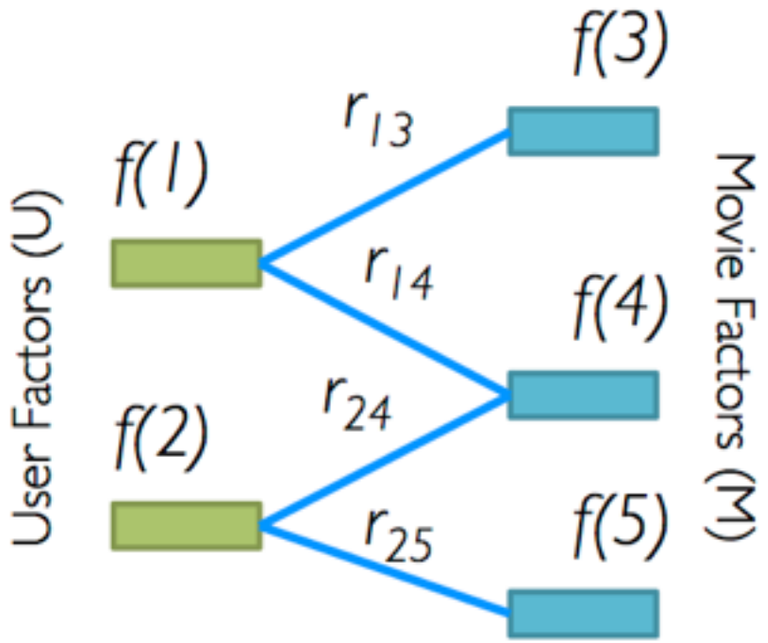
Collaborative filtering

★	★★★★	?
★	★★★	★★
★★★★	?	★
★	?	★★
?	★★★	★★
★★★★	★★	?

- Recover a rating matrix from a subset of its entries.

Alternating least squares (ALS)





Iterate:

$$f[i] = \arg\min_{w \in \mathbb{R}^d} \sum_{j \in \mathrm{Nbrs}(i)} \left(r_{ij} - w^T f[j]\right)^2 + \lambda ||w||_2^2$$

ALS - wall-clock time

System	Wall-clock time (seconds)
Matlab	15443
Mahout	4206
GraphLab	291
MLlib	481

- Dataset: scaled version of Netflix data (9X in size).
- Cluster: 9 machines.
- MLlib is an order of magnitude faster than Mahout.
- MLlib is within factor of 2 of GraphLab.