

A vertical line on the left side of the slide, with a small circle at its midpoint.

Hello!

I AM RAJA HASEEB

I am a second year student in Electrical Department
under supervision of professor Kim Jong-Hwan

My research interests include ML/DL, AI and computer
vision



MLlib: Scalable Machine Learning on Spark

Xiangrui Meng, Ameet Talwalkar, Evan Sparks, Virginia Smith, Xinghao Pan, Shivaram Venkataraman, Matei Zaharia, Rean Griffith, John Duchi, Joseph Gonzalez, Michael Franklin, Michael I. Jordan, Tim Kraska, etc.

● Contents

○ THIS PRESENTATION

What is Apache Spark

Component Stack

Why Spark?

Data processing vs machine learning
frameworks

MLlib

Some comparisons

Conclusion

MAIN IDEAS

About Apache Spark

MLlib of Spark

1

ABOUT APACHE SPARK?

What is Apache Spark?

Components

Why Spark?

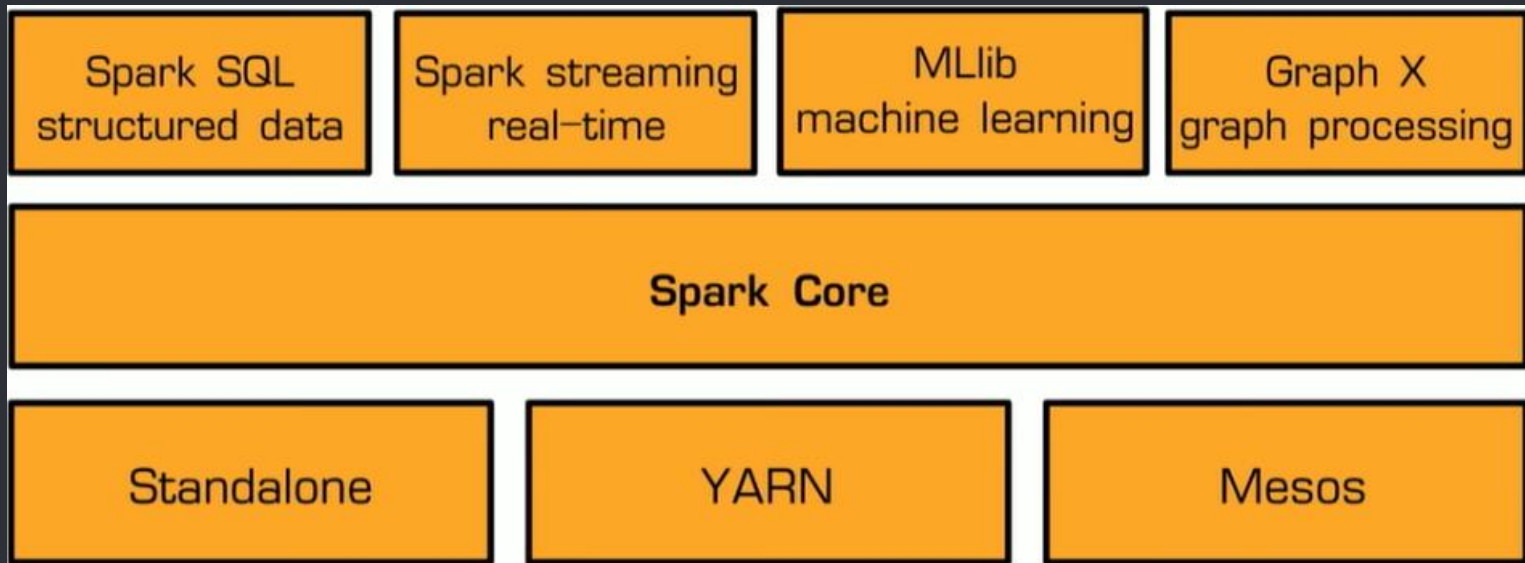
- What is Apache Spark?

- **Apache Spark** is a data processing framework that can quickly perform processing tasks on very large data sets, and can also distribute data processing tasks across multiple computers, either on its own or in tandem with other distributed computing tools
- Spark also takes some of the programming burdens of these tasks off the shoulders of developers with an easy-to-use API
- It was developed motivated by the limitations in the MapReduce/Hadoop paradigm

- What is Apache Spark?



- Component Stack



● Why Spark?

- Apache Spark is among one of the few competing big data frameworks for parallel computing that provides a combination of in-memory processing, fault-tolerance, scalability, speed and ease of programming.
- From its humble beginnings in the AMPLab at U.C. Berkeley in 2009, Apache Spark has become one of the key big data distributed processing frameworks in the world

● Why Spark?

- High speed data querying, analysis, and transformation with large data sets
- Compared to MapReduce, Spark offers much less reading and writing to and from the disk
- Great for iterative algorithms
- Easy-to-use API makes a big difference in terms of ease of development, readability, and maintenance
- Super fast, especially for interactive queries. (100x faster than classic Hadoop Hive queries)
- Supports multiple languages and integrations with other popular products.
- Helps make complex data pipelines coherent and easy.

2

Big Data vs ML frameworks

Spark vs Tensorflow

● Spark vs TensorFlow

- Spark, essentially a big data framework, has made it possible for a large number of companies generating huge amount of user data to process it efficient.
- TensorFlow, on the other hand, is a short library developed by Google that helps in improving the performance of numerical computation and neural networks and generating data flow as graphs



MLlib

Spark Machine Learning library

● 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
- 33 contributors

- What is MLlib?

- Algorithms

- **Classification:** logistic regression, linear support vector machine (SVM), naive Bayes
- **Regression:** generalized linear regression (GLM)
- **Collaborative filtering:** alternating least squares (ALS)
- **Clustering:** k-means
- **Decomposition:** singular value decomposition (SVD), principal component analysis (PCA)

- What is MLlib?

- ML workflow utilities include:

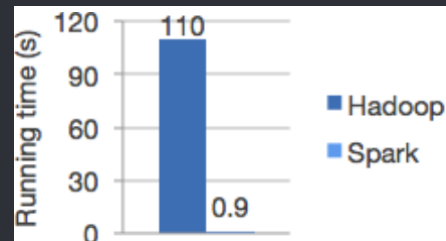
- Feature transformations: standardization, normalization, hashing, tokenization etc.
- ML Pipeline construction
- Model evaluation and hyper-parameter tuning
- ML persistence: saving and loading models and Pipelines

- Other utilities include:

- Linear Algebra and statistics

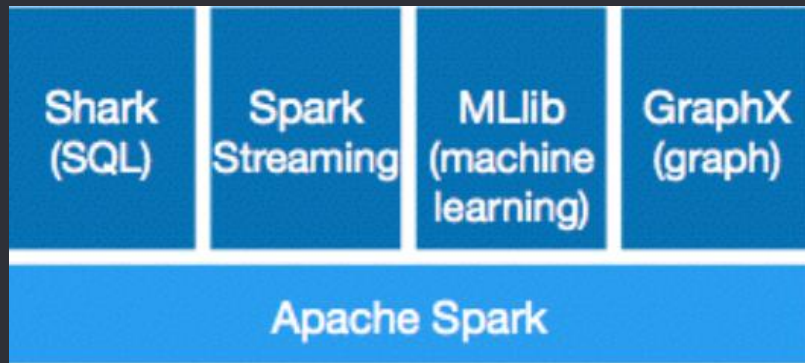
Why MLlib?

- Many other machine learning frameworks exist e.g. Mahout, scikit-learn, LABLINEAR, Matlab, Weka, Vowpal Wabbit etc.
- 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.



Why MLlib?

- It ships with Spark as a standard component.



- Scalability, performance, user-friendly APIs, integration with spark and its other components

● Why MLlib?

- Spark solves big data problem. By using cluster of machines instead of one, we are able to use more data. The more data we can throw in our models, the better they perform.
- MLlib benefits from its tight integration with various spark components. MLlib leverages high level libraries packaged with the Spark framework
- MLlib provides a package called `spark.ml` to simplify the development and performance tuning of multi-stage machine learning pipelines.

● Why MLlib?

- MLlib provides fast and distributed implementations of common machine learning algorithms along with a number of low-level primitives and various utilities for statistical analysis, feature extraction, convex optimizations, and distributed linear algebra.
- It also works really well with number of different pipelines like tensorflow, sk-learn etc.

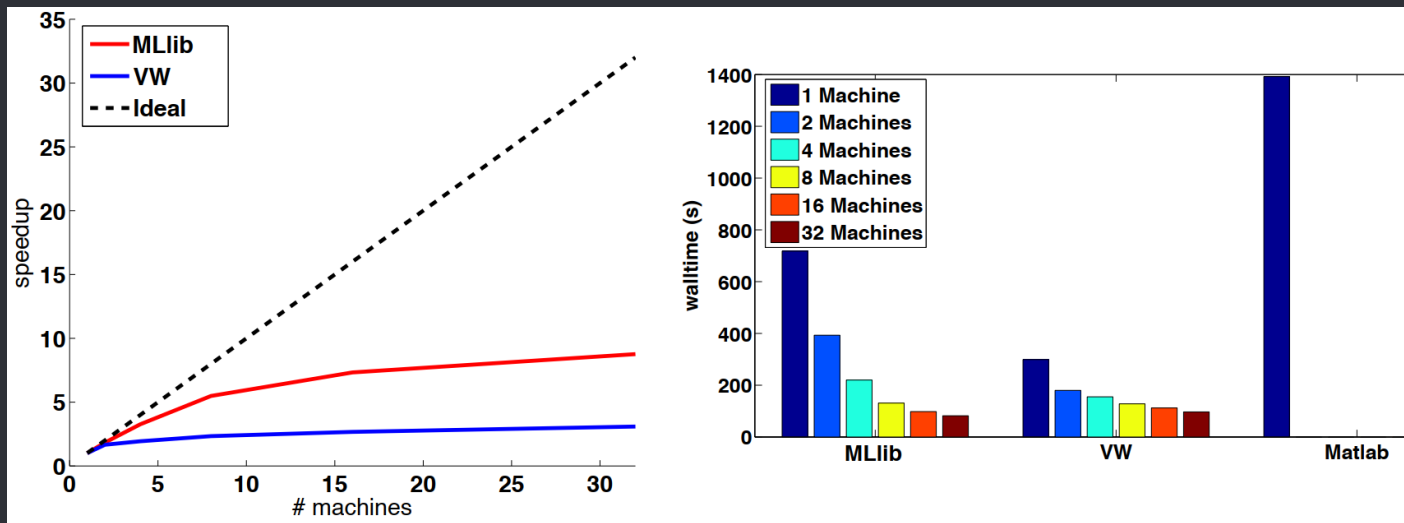
3

Comparison

Logistic regression
Alternating least squares

● Logistic regression - strong scaling

- Fixed Dataset: 50K images, 160K dense features.
- MLlib exhibits better scaling properties.
- MLlib is faster than VW with 16 and 32 machines.



● ALS wall-clock time

- 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

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

4

Spark MLlib Use Cases



● Common uses

- Operational Optimization
- Risk Assessment
- Fraud Detection
- Marketing optimization
- Advertising Optimization
- Security Monitoring
- Customer Segmentation
- Product Recommendations.

Other cases

- Spark MLlib is used for frequent pattern mining and is core to the analytics platform of Huawei's big data solution, Fusion Insight that is used by more than 100 customers across the world
- Toyota's Customer 360 Insights Platform leverages MLlib library for categorizing and prioritizing its customers social media interactions in real-time
- Spark MLlib is an integral part of Open Table's dining recommendations.
- ING 's machine learning pipeline uses Spark MLlib's K-Means Clustering and Decision Tree Ensembles for anomaly detection.
- Netflix and Spotify use Spark Streaming and Spark MLlib to make user recommendations that best fit in its customer tastes and buying histories.

4

Conclusion

Final remarks


● Final remarks

● Spark MLlib is in active development and all thanks to all the MLlib contributors for joining hands to speed up machine learning development

● Due to its high performance and scalability, there is a surge in adoption of Spark in Big data industries

● MLlib benefits from its tight integration with various spark components. MLlib leverages high level libraries packaged with the Spark framework. We can use more data which results in better models.

● Main Contributions



However, features are limited. Neural network support via only multi-layer perceptron classifier. Also no tensor support.

However, there are other deep learning pipelines available on Apache Spark

Thank you!

○ ANY QUESTIONS?