

Lecture – 11

Spark Built in Libraries

Refer slide time: (0:14)

Spark Built-in Libraries



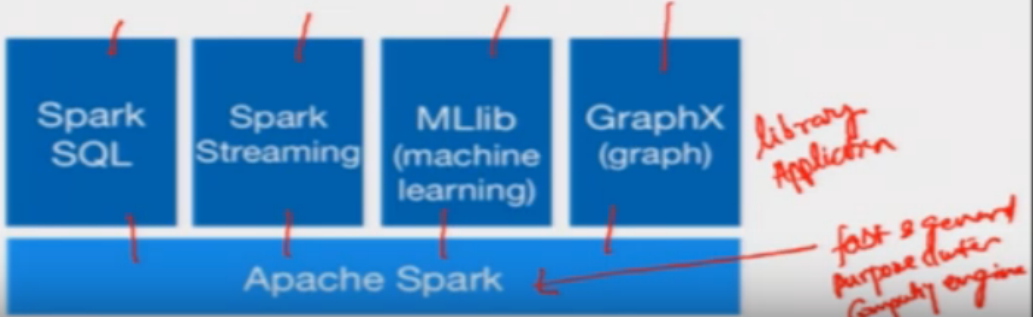
Dr. Rajiv Misra
Dept. of Computer Science & Engg.
Indian Institute of Technology Patna
rajivm@iitp.ac.in

The Spark Built-in Libraries.

Refer slide time: (0:16)

Introduction

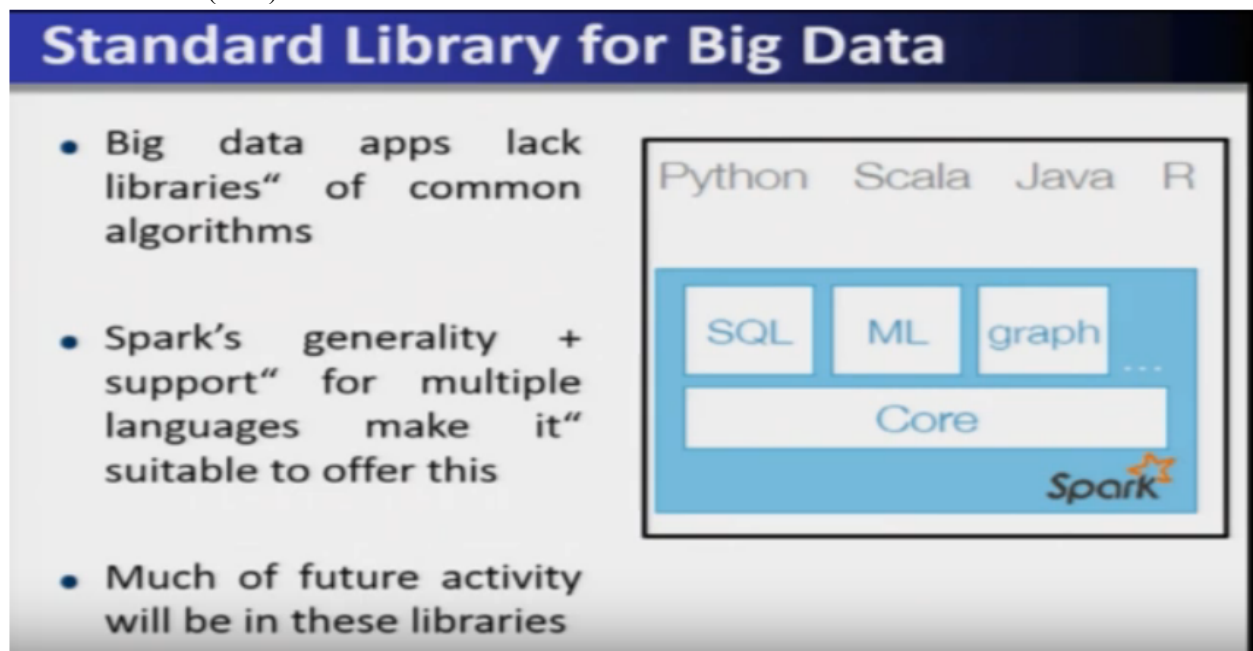
- Apache Spark is a fast and general-purpose cluster computing system for large scale data processing
- High-level APIs in Java, Scala, Python and R



Now Spark, Apache Spark is a fast and general-purpose, cluster computing, for large-scale, data processing. The Spark supports, high level APIs, in the form with, Java, Scala, Python and R. Now let us,

see the Spark core, which is the fast and a general purpose, cluster computing engine. Now this particular Spark is giving a fast and general-purpose computing, which provides or which has, enabled, various applications, to be run on, top of this Spark core or Spark engine. Such as the various components, on various libraries, which are supported, by the Spark, for different applications, are summarized in this particular diagram. So the first one is called, 'Spark Sequel.' So that means, Sequel like, commands are being provided and by that, the key value store and various programming can be supported, using, Sequel like commands. The another one is called as, 'Spark streaming', for real-time, streaming applications. Here the data will be in the form of, micro batches and the streams will be computed in, real-time. Another one is called, 'Spark MLlib', that is, machine learning libraries, which are provided, over the, Spark Core. Finally, the graph processing is done, over top of, the Spark, in the form of library, which is called, 'Graph X'.

Refer slide time: (2:12)



So these are the standard libraries, which are used for, various big data applications and also supports, various common algorithms, which are used for, big data analytics.

Refer slide time: (2:26)

Machine Learning Library (MLlib)

MLlib algorithms:

- (i) **Classification:** logistic regression, linear SVM, "naïve Bayes, classification tree
- (ii) **Regression:** generalized linear models (GLMs), regression tree
- (iii) **Collaborative filtering:** alternating least squares (ALS), non-negative matrix factorization (NMF)
- (iv) **Clustering:** k-means
- (v) **Decomposition:** SVD, PCA
- (vi) **Optimization:** stochastic gradient descent, L-BFGS

So let us see, these libraries, standard libraries, which are part of the Spark core, in more details.

Refer slide time: (2:31)

Machine Learning Library (MLlib)

MLlib algorithms:

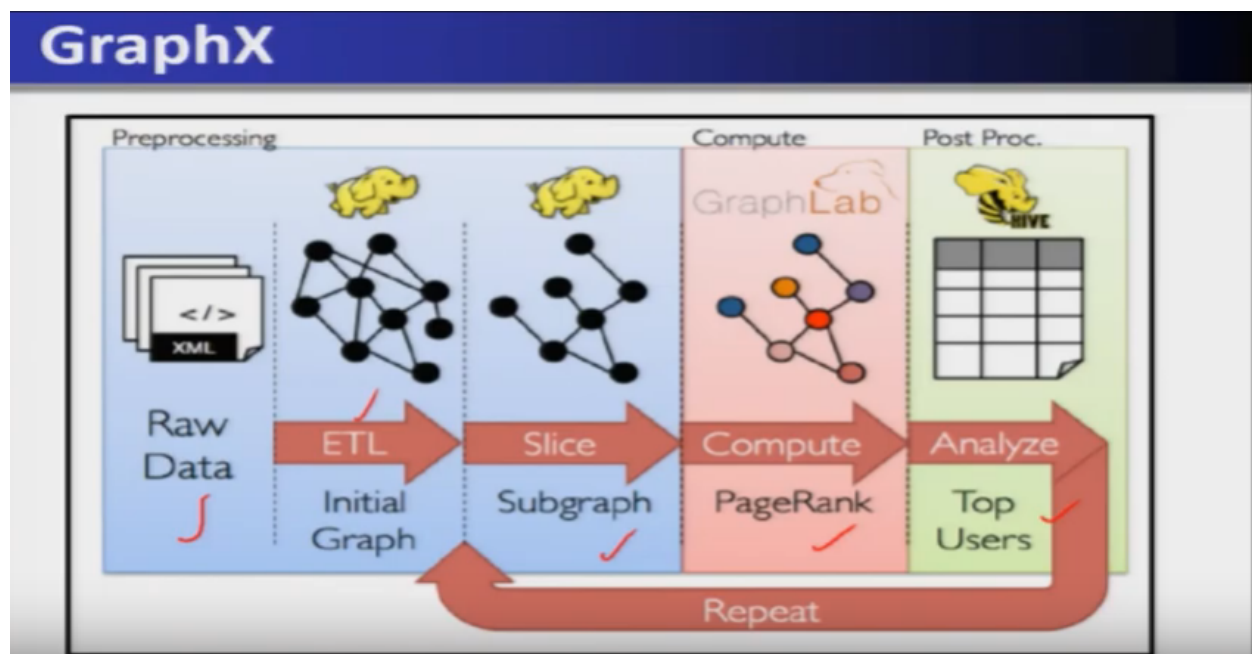
- (i) **Classification:** logistic regression, linear SVM, "naïve Bayes, classification tree
- (ii) **Regression:** generalized linear models (GLMs), regression tree
- (iii) **Collaborative filtering:** alternating least squares (ALS), non-negative matrix factorization (NMF)
- (iv) **Clustering:** k-means
- (v) **Decomposition:** SVD, PCA
- (vi) **Optimization:** stochastic gradient descent, L-BFGS

SPARK MLlib
Scalable Machine Learning
Algorithms for
big data
Analytics

For example, the Spark machine learning library, which is called, 'A Spark MLlib Library'. Now this particular m, Spark MLlib Libraries, provides a collection of, machine learning algorithms, which are, provided here, for doing the big data analytics, that is called, 'Scalable Machine Learning Algorithms', for big data analytics. Let us summarize what our, what are the different, machine learning, scalable

algorithms, are available, in the form of MLlib. For the classification application, the algorithms like, logistic regression, linear support vector machine, Naive Bayes and decision trees are available, as part of the MLlib library. For regression application, generalized linear model, GLM and regression tree is available. Similarly for, collaborative filtering, alternating least squares and non-negative matrix factorization, is available, as part of MLlib. For unsupervised clustering or cluster analysis, parallel K, means, algorithm, is available as part of MLlib. Similarly, for decomposition, support, SVD and principal component analysis, is available, for decomposition. And for the optimization, various libraries are available, will such as, Stochastic, Gradient descent and L-BFGS.

Refer slide time: (4:24)



Let us see the, another library, which is supporting the, graph applications, over the Spark, that is large-scale graph computation, over the Spark, which is supported, in the form of, the graphics. So here the graph, here the raw data, it takes and this figure shows that, that extract, transform and load this particular, ETL, it will through that, it will create the initial graph. It will perform various transformations, operations, on the graph, such as; it will create the sub graph, it will perform the graph algorithms. Which are, such as, page rank and it will do the different analysis. So all these operations are supported, in the form of graphic X.

Refer slide time: (5:15)

GraphX

- General graph processing library
- Build graph using RDDs of nodes and edges
- Large library of graph algorithms with composable steps

So graph X, is a general purpose, graph processing library and it builds, the graph using, RDD's of the nodes and edges. Large library of graph algorithms, are available, as part of the graph X.

Refer slide time: (5:31)

GraphX Algorithms

(i) Collaborative Filtering

Alternating Least Squares
Stochastic Gradient Descent
Tensor Factorization

(ii) Structured Prediction

Loopy Belief Propagation
Max-Product Linear Programs
Gibbs Sampling

(iii) Semi-supervised ML

Graph SSL
CoEM

(iv) Community Detection

Triangle-Counting
K-core Decomposition
K-Truss

(v) Graph Analytics

PageRank
Personalized PageRank
Shortest Path
Graph Coloring

(vi) Classification

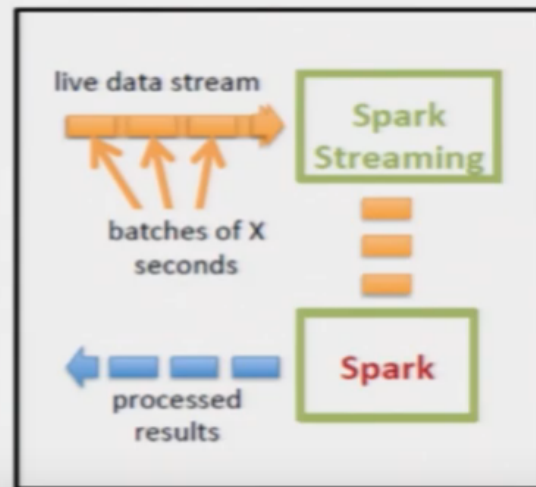
Neural Networks

Let us summarize some of the algorithms, which are available, as part of the graph X library and doing, and for the Graph processing. Collaborative filtering and then structured prediction, then, semi-supervised machine learning and then, community detection, graph analytics and classification, using neural networks. So all these graph analytics, are available, such as, page rank algorithm, on graphs, personalized page rank algorithm, shortest path graph coloring. All these algorithms, are available, as part of, the graphic X library.

Refer slide time: (6:23)

Spark Streaming

- Large scale streaming computation
- Ensure exactly one semantics
- Integrated with Spark → unifies batch, interactive, and streaming computations!



Similarly, for community Detection, triangle counting, K-core, Decomposition, K-Truss, all these algorithms are, also available, for doing the graph analytics. Similarly for, SPARK streaming, this particular Spark core, provides, library for, supporting the Spark streaming, Large-scale streaming applications are supported, in the form of, Spark streaming, Spark standard streaming library system. And it will unify, the integrate with, the Spark, to unify for batch interactive and streaming computations together.

Refer slide time: (6:57)

Spark SQL

Enables loading & querying structured data in Spark

From Hive:

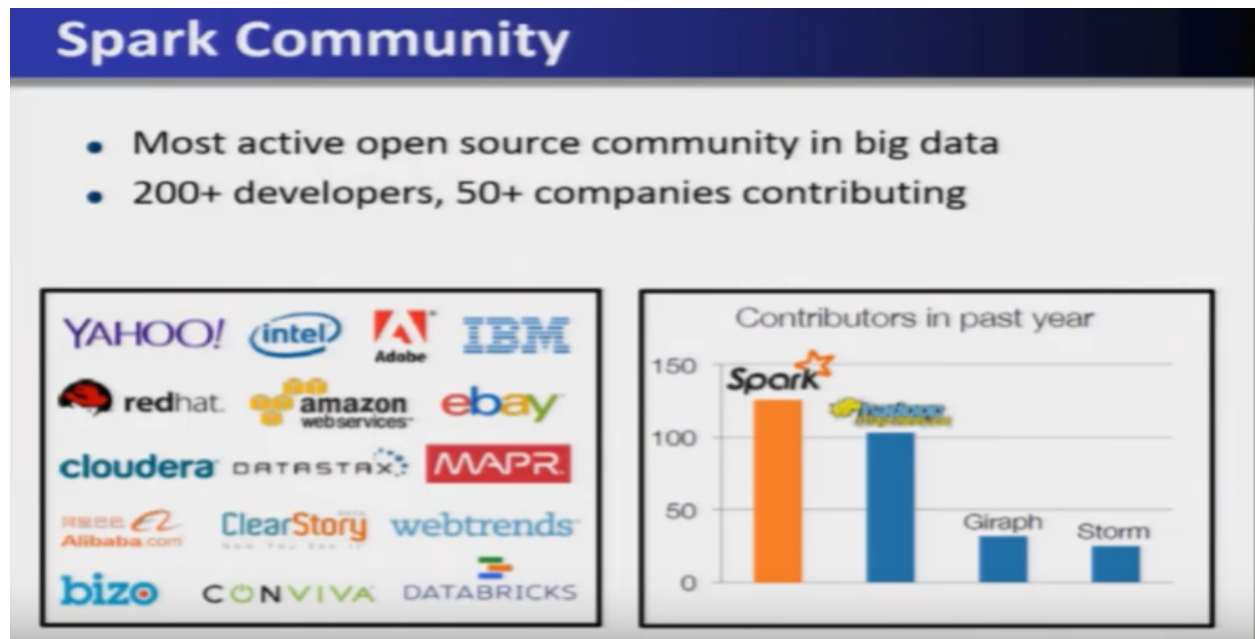
```
c = HiveContext(sc)
rows = c.sql("select text, year from hivetable")
rows.filter(lambda r: r.year > 2013).collect()
```

From JSON:

```
c.jsonFile("tweets.json").registerAsTable("tweets")
c.sql("select text, user.name from tweets")
```


Next library which is available, with the Spark core, is called, 'Spark Sequel'. Now it will enable loading and querying structured data, in the form of the Spark. It is also having the links or APIs with, the hive and with, JSON.

Refer slide time: (7:16)



So Spark community, now is, most open-source, most active open-source community, in the Big Data and 200-plus developers and 550 plus companies, are contributing. And these icons shows that, the complete presence of a Spark, in all the, the production clusters, at most of these companies, are shown over here and you can see that. The Spark, I surpassed, the Hadoop Map Reduce, in the number of contributions. Thank you.