(/) *Lightning-fast cluster computing*

Latest News

Spark 2.1.1 released (/news/spark-2-1-1-released.html) (May 02, 2017)

Spark Summit (June 5-7th, 2017, San Francisco) agenda posted (/news/spark-summit-june-2017-agenda-posted.html) (Mar 31, 2017)

Spark Summit East (Feb 7-9th, 2017, Boston) agenda posted (/news/spark-summit-east-2017-agenda-posted.html) (Jan 04, 2017)

Spark 2.1.0 released (/news/spark-2-1-0-released.html) (Dec 28, 2016)

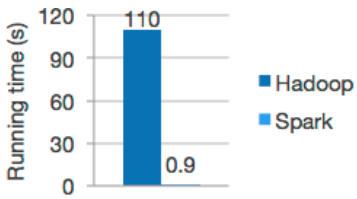
Archive (/news/index.html)

Apache Spark™ is a fast and general engine for large-scale data processing.

Speed

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

Apache Spark has an advanced DAG execution engine that supports acyclic data flow and in-memory computing.



Logistic regression in Hadoop and Spark

Ease of Use

Write applications quickly in Java, Scala, Python, R.

Spark offers over 80 high-level operators that make it easy to build parallel apps. And you can use it *interactively* from the Scala, Python and R shells.

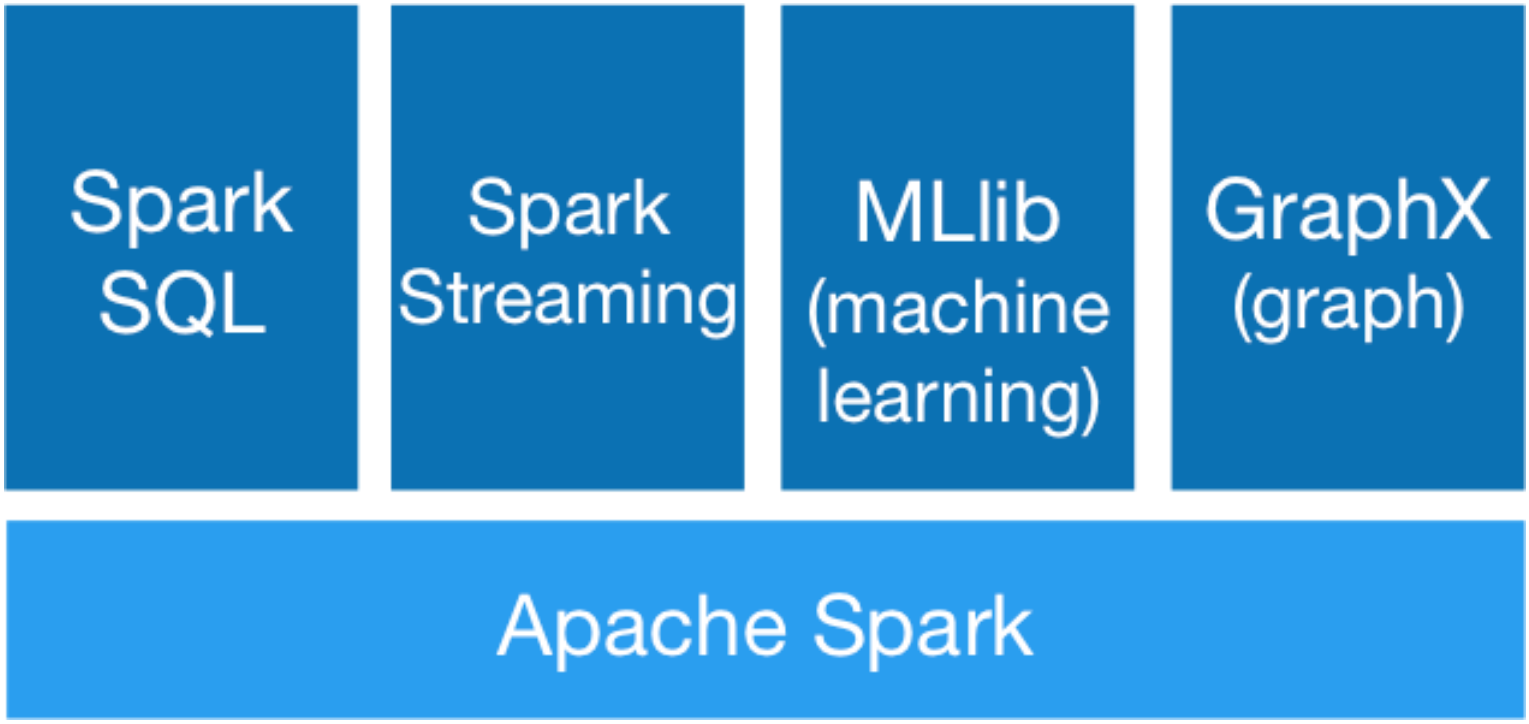
```
text_file = spark.textFile("hdfs://...")

text_file.flatMap(lambda line: line.split())
    .map(lambda word: (word, 1))
    .reduceByKey(lambda a, b: a+b)
```

Word count in Spark's Python API

Generality

Combine SQL, streaming, and complex analytics.

Spark powers a stack of libraries including SQL and DataFrames (/sql/), MLlib (/mllib/) for machine learning, GraphX (/graphx/), and Spark Streaming (/streaming/). You can combine these libraries seamlessly in the same application.



Runs Everywhere

Spark runs on Hadoop, Mesos, standalone, or in the cloud. It can access diverse data sources including HDFS, Cassandra, HBase, and S3.

You can run Spark using its standalone cluster mode (/docs/latest/spark-standalone.html), on EC2 (https://github.com/amplab/spark-ec2), on Hadoop YARN (https://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html), or on Apache Mesos (https://mesos.apache.org). Access data in HDFS (https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-hdfs/HdfsUserGuide.html), Cassandra (https://cassandra.apache.org), HBase (https://hbase.apache.org), Hive (https://hive.apache.org), Tachyon (http://tachyon-project.org), and any Hadoop data source.