(/)

Spark Research

Apache Spark started as a research project at UC Berkeley in the AMPLab (https://amplab.cs.berkeley.edu), which focuses on big data analytics.

Our goal was to design a programming model that supports a much wider class of applications than MapReduce, while maintaining its automatic fault tolerance. In particular, MapReduce is inefficient for *multi-pass* applications that require low-latency data sharing across multiple parallel operations. These applications are quite common in analytics, and include:

- *Iterative algorithms*, including many machine learning algorithms and graph algorithms like PageRank.
- *Interactive data mining*, where a user would like to load data into RAM across a cluster and query it repeatedly.
- *Streaming applications* that maintain aggregate state over time.

Traditional MapReduce and DAG engines are suboptimal for these applications because they are based on acyclic data flow: an application has to run as a series of distinct jobs, each of which reads data from stable storage (e.g. a distributed file system) and writes it back to stable storage. They incur significant cost loading the data on each step and writing it back to replicated storage.

Spark offers an abstraction called *resilient distributed datasets (RDDs)* (http://people.csail.mit.edu/matei/papers/2012/nsdi_spark.pdf) to support these applications efficiently. RDDs can be stored in memory between queries *without* requiring replication. Instead, they rebuild lost data on failure using *lineage*: each RDD remembers how it was built from other datasets (by transformations like `map`, `join` or `groupBy`) to rebuild itself. RDDs allow Spark to outperform existing models by up to 100x in multi-pass analytics. We showed that RDDs can support a wide variety of iterative algorithms, as well as interactive data mining and a highly efficient SQL engine (Shark (http://shark.cs.berkeley.edu)).

You can find more about the research behind Spark in the following papers:

- SparkR: Scaling R Programs with Spark (https://people.csail.mit.edu/matei/papers/2016/sigmod_sparkr.pdf), Shivaram Venkataraman, Zongheng Yang, Davies Liu, Eric Liang, Hossein Falaki, Xiangrui Meng, Reynold Xin, Ali Ghodsi, Michael Franklin, Ion Stoica, and Matei Zaharia. *SIGMOD 2016*. June 2016.
- MLlib: Machine Learning in Apache Spark (http://www.jmlr.org/papers/volume17/15-237/15-237.pdf), Xiangrui Meng, Joseph Bradley, Burak Yavuz, Evan Sparks, Shivaram Venkataraman, Davies Liu, Jeremy Freeman, DB Tsai, Manish Amde, Sean Owen, Doris Xin, Reynold Xin, Michael J. Franklin, Reza Zadeh, Matei Zaharia, and Ameet Talwalkar. *Journal of Machine Learning Research (JMLR)*. 2016.
- Spark SQL: Relational Data Processing in Spark (http://people.csail.mit.edu/matei/papers/2015/sigmod_spark_sql.pdf). Michael Armbrust, Reynold S. Xin, Cheng Lian, Yin Huai, Davies Liu, Joseph K. Bradley, Xiangrui Meng, Tomer Kaftan, Michael J. Franklin, Ali Ghodsi, Matei Zaharia. *SIGMOD 2015*. June 2015.
- GraphX: Unifying Data-Parallel and Graph-Parallel Analytics (https://amplab.cs.berkeley.edu/wp-content/uploads/2014/02/graphx.pdf). Reynold S. Xin, Daniel Crankshaw, Ankur Dave, Joseph E. Gonzalez, Michael J. Franklin, Ion Stoica. *OSDI 2014*. October 2014.
- Discretized Streams: Fault-Tolerant Streaming Computation at Scale (http://people.csail.mit.edu/matei/papers/2013/sosp_spark_streaming.pdf). Matei Zaharia, Tathagata Das, Haoyuan Li, Timothy Hunter, Scott Shenker, Ion Stoica. *SOSP 2013*. November 2013.
- Shark: SQL and Rich Analytics at Scale (http://people.csail.mit.edu/matei/papers/2013/sigmod_shark.pdf). Reynold S. Xin, Joshua Rosen, Matei Zaharia, Michael J. Franklin, Scott Shenker, Ion Stoica. *SIGMOD 2013*. June 2013.
- Discretized Streams: An Efficient and Fault-Tolerant Model for Stream Processing on Large Clusters (http://people.csail.mit.edu/matei/papers/2012/hotcloud_spark_streaming.pdf). Matei Zaharia, Tathagata Das, Haoyuan Li, Scott Shenker, Ion Stoica. *HotCloud 2012*. June 2012.
- Shark: Fast Data Analysis Using Coarse-grained Distributed Memory (http://people.csail.mit.edu/matei/papers/2012/sigmod_shark_demo.pdf) (demo). Cliff Engle, Antonio Lupher, Reynold S. Xin,

Matei Zaharia, Haoyuan Li, Scott Shenker, Ion Stoica. *SIGMOD 2012*. May 2012. **Best Demo Award**.

- Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing (http://people.csail.mit.edu/matei/papers/2012/nsdi_spark.pdf). Matei Zaharia, Mosharaf Chowdhury, Tathagata Das, Ankur Dave, Justin Ma, Murphy McCauley, Michael J. Franklin, Scott Shenker, Ion Stoica. *NSDI 2012*. April 2012. **Best Paper Award**.
- Spark: Cluster Computing with Working Sets (http://people.csail.mit.edu/matei/papers/2010/hotcloud_spark.pdf). Matei Zaharia, Mosharaf Chowdhury, Michael J. Franklin, Scott Shenker, Ion Stoica. *HotCloud 2010*. June 2010.

Latest News

Preview release of Spark 4.0 (/news/spark-4.0.0-preview1.html) (Jun 03, 2024)

Spark 3.4.3 released (/news/spark-3-4-3-released.html) (Apr 18, 2024)

Spark 3.5.1 released (/news/spark-3-5-1-released.html) (Feb 23, 2024)

Spark 3.3.4 released (/news/spark-3-3-4-released.html) (Dec 16, 2023)

Archive (/news/index.html)

COMMUNITY
THE ASF CONFERENCE
CODE

(https://www.apache.org/events/current-event.html)

DOWNLOAD SPARK (/DOWNLOADS.HTML)

Built-in Libraries:

SQL and DataFrames (/sql/)
Spark Streaming (/streaming/)
MLlib (machine learning) (/mllib/)
GraphX (graph) (/graphx/)

Third-Party Projects (/third-party-projects.html)