How Companies are Using Spark

And where the Edge in Big Data will be

Matei Zaharia

DATABRICKS

MIT

Spark
History

Decreasing storage costs have led to an explosion of big data. Commodity cluster software, like Hadoop, has made it 10-20x cheaper to store large datasets. Broadly available from multiple vendors.
Implication

Big data storage is becoming commoditized, so how will organizations get an edge?

What matters now is what you can do with the data.
Two Factors

**Speed**: how quickly can you go from data to decisions?

**Sophistication**: can you run the best algorithms on the data?

These factors have usually required separate, non-commodity tools.
Apache Spark

A compute engine for Hadoop data that is:

**Fast:** up to 100x faster than MapReduce
Apache Spark

A compute engine for Hadoop data that is:

**Fast:** up to 100x faster than MapReduce

**Sophisticated:** can run today’s most advanced algorithms
Apache Spark

A compute engine for Hadoop data that is:

**Fast:** up to 100x faster than MapReduce

**Sophisticated:** can run today’s most advanced algorithms

**Fully open source:** one of most active projects in big data

Contributors in past year:

- Spark
- Giraph
- Storm
- Tez
Apache Spark

A compute engine for Hadoop data that is:

**Fast:** up to 100x faster than MapReduce

**Sophisticated:** can run today’s most advanced algorithms

**Fully open source:** one of most active projects in big data

Spark brings top-end data analysis to commodity Hadoop clusters
Spark Use Cases
1. Yahoo! Personalization

Yahoo! properties are highly personalized to maximize relevance.

Reaction must be fast, as stories, etc change in time.

Best algorithms are highly sophisticated.
1. Yahoo! Personalization

Example challenge: relevance of news stories

Relevance models must be updated throughout the day
1. Yahoo! Personalization

Spark at Yahoo!
» Runs in Hadoop YARN to use existing data & clusters

Result: pilot for stream ads
» 120 lines in Scala, compared to 15K in C++
» 30 min to run on 100 million samples

Major contributor on YARN support, scalability, operations
2. Yahoo! Ad Analytics

Yahoo! Ads wanted interactive BI on terabytes of data

Chose Shark (Hive on Spark) to provide this through standard Hive server API + Tableau

Result: interactive-speed queries on terabytes from Tableau

Major contributor on columnar compression, statistics, JDBC
3. Conviva Real-Time Video Optimization

Conviva manages 4+ billion video streams per month

Dynamically selects sources to optimize quality

**Time is critical:** 1 second buffering = lost viewers
3. Conviva Real-Time Video Optimization

Using **Spark Streaming**, Conviva learns network conditions in real time.

Results fed directly to video players to optimize streams.

System running in production.
4. ClearStory Data: Multi-source, Fast-cycle Analysis

Same-day results from data updating at disparate sources

Dozens of disparate sources converged in seconds/minutes
4. ClearStory Data: Multi-source, Fast-cycle Analysis
Get Started

Download and resources: spark.incubator.apache.org

Free video tutorials: spark-summit.org/2013

Commercial support:

DATABRICKS + CLOUDERA®
Conclusion

Big data will be standard: everyone will have it.

Organizations will gain an edge through **speed** of action and **sophistication** of analysis.

Apache Spark brings these to Hadoop clusters.