Spark Streaming

Mateusz, Kopeć, Michał Okulewicz

Institute of Computer Science
Polish Academy of Sciences

Big Data
27 November 2014

Presentation Plan

1. Introduction
2. Spark architecture
3. Example stream processing task
4. How to run Spark?

What is Apache Spark™?

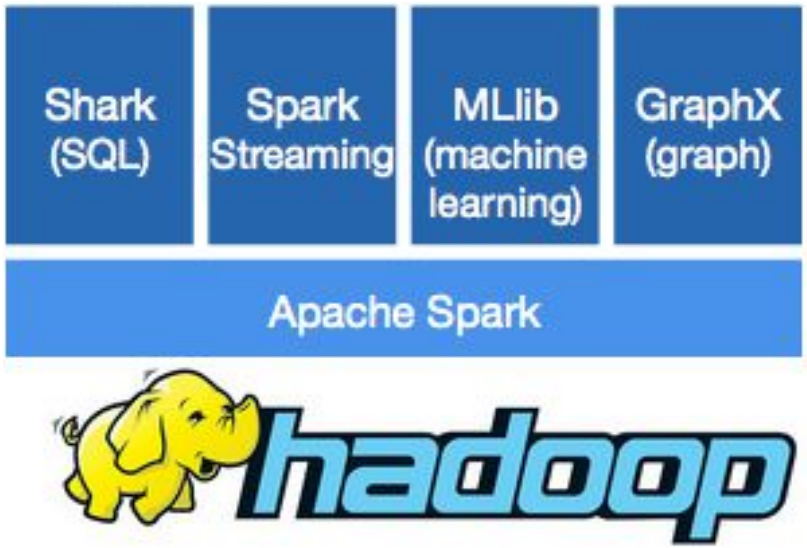
Apache Spark™

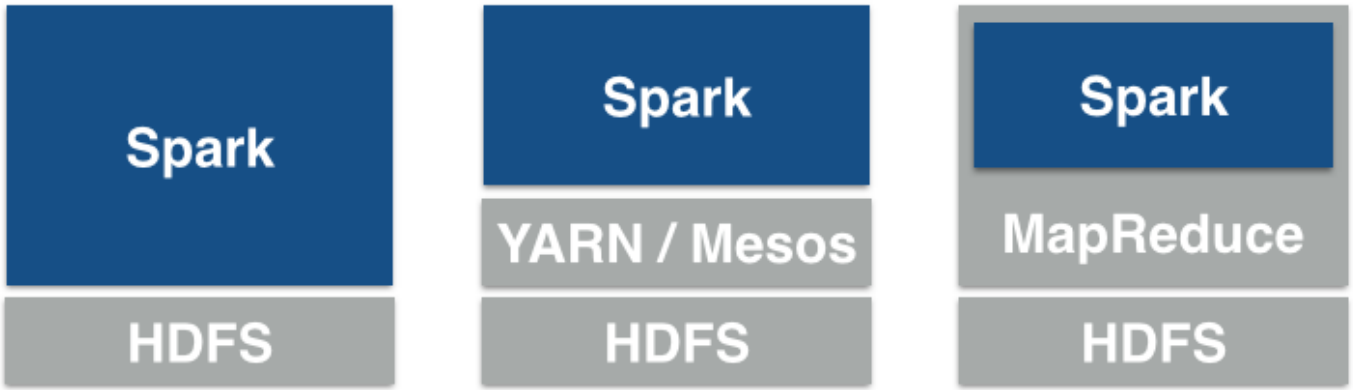
- distributed computations system
- not only MapReduce applications
- supports in-memory operations (Resilient Distributed Dataset)
- may use HDFS

APIs

- Scala
- Java
- python

Spark architecture

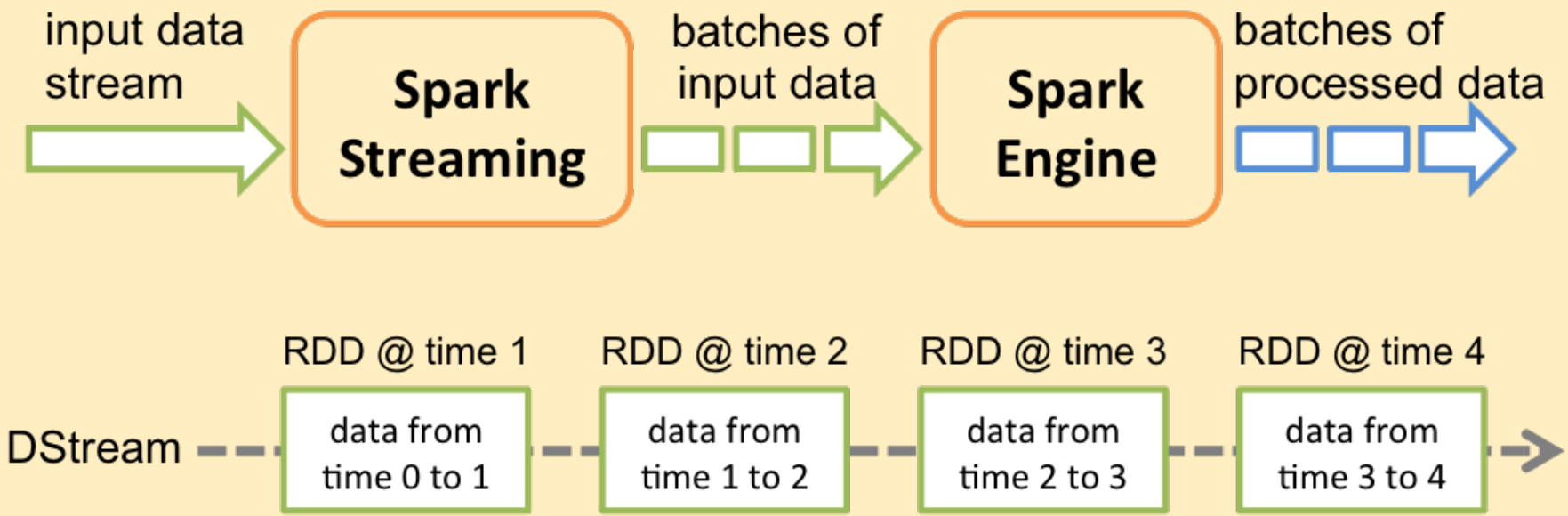




What is Spark Streaming?

Spark Streaming

- subproject of Apache Spark™
- allows for real-time distributed stream processing
- utilizes an idea called Discretized Stream (DStream)



Why do we want Spark Streaming?

- Fraud detection
- Financial market analysis
- On-line surveillance
- Early earthquakes detection
- ...

Example stream processing task

Data producer

Generate another number every 100ms

Data analyzer

Count all distinct numbers

How to run Spark? I

Running worker on Linux

- Get precompiled Spark 1.1.0 for Hadoop 1.x from `/home/2012/m.okulewicz/spark` and unpack it
- If necessary edit: `conf/spark-env.sh` and add location of `JAVA_HOME`
- Run `sbin/start-slave.sh 1 spark://phd03.phd.ipipan.waw.pl`
- Check in browser if `http://localhost:8081` is available and master points to `phd03.phd.ipipan.waw.pl`

How to run Spark? II

Running master and task on Linux

- Run `sbin/start-master.sh`
- Check in browser if `http://localhost:8080` is available
- TO BE CHANGED Run `bin/spark-submit`
  `--class org.apache.spark.examples.SparkPi`
  `--master spark://207.184.161.138:7077`
  `--executor-memory 20G`
  `--total-executor-cores 100`
  `/path/to/examples.jar 1000`

Bibliography I


https://spark.apache.org/docs/0.9.0/streaming-programming guide.html.
Streaming Programming Guide, 2014.


Matei Zaharia, Tathagata Das, Haoyuan Li, Timothy Hunter, Scott Shenker, and Ion Stoica.
Discretized Streams: Fault-tolerant Streaming Computation at Scale.
In *Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles*, SOSP '13, pages 423–438, New York, NY, USA, 2013. ACM.


Matei Zaharia, Tathagata Das, Haoyuan Li, Scott Shenker, and Ion Stoica.
Discretized Streams: An Efficient and Fault-tolerant Model for Stream Processing on Large Clusters.
In *Proceedings of the 4th USENIX Conference on Hot Topics in Cloud Ccomputing*, HotCloud'12, pages 10–10, Berkeley, CA, USA, 2012. USENIX Association.