Spark Streaming

Mateusz Kopeć, Michał Okulewicz

Institute of Computer Science
Polish Academy of Sciences

Big Data
27 November 2014

Presentation Plan

1. Spark
2. Spark Streaming
3. Example 1: Stream processing task
4. Example 2: Twitter

What is Apache Spark™?

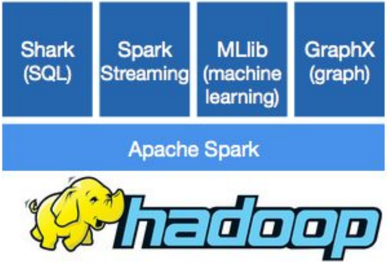
Apache Spark™

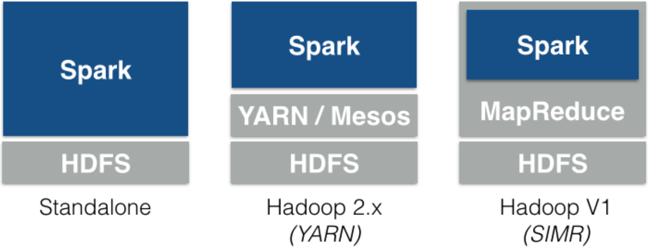
- distributed computations system
- not only MapReduce applications
- supports in-memory operations (Resilient Distributed Dataset)
- may use HDFS

APIs

- Scala
- Java
- python

Spark architecture





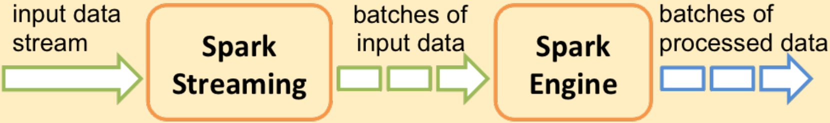
Why do we want Spark Streaming?

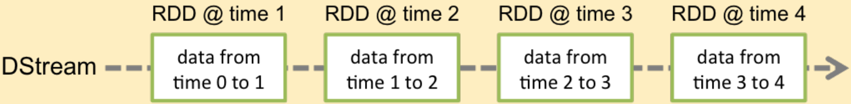
- Fraud detection
- Financial market analysis
- On-line surveillance
- Early earthquakes detection
- . . .

What is Spark Streaming?

Spark Streaming

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





Example stream processing task

Data producer

Generates next integer every 100ms

Data analyser

Counts all distinct numbers

How to run Spark? I

Running master on Linux

- Run `sbin/start-master.sh`
- Check in browser if `http://localhost:8080` is available

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://phd01.phd.ipipan.waw.pl`
- Check in browser if `http://localhost:8081` is available and master points to `phd01.phd.ipipan.waw.pl`

How to run Spark? II

Running task on Linux

- Run:

```
./bin/spark-submit
--class
pl.waw.ipipan.phd.mkopec.sparkReceiver.SocketReceiver
--master spark://phd01.phd.ipipan.waw.pl:7077
--executor-memory 20G
--total-executor-cores 100
/path/to/jar.jar localhost 9999 1000 1
```

Example stream processing task

Data producer

Retrieves stream of tweets containing specified keywords

Data analyser

Counts most frequent words in tweets

How to run Spark? I

Running task on Linux

- Run:

```
./bin/spark-submit
--class
pl.waw.ipipan.phd.mkopec.sparkReceiver.TwitterReceiver
--master spark://phd01.phd.ipipan.waw.pl:7077
--executor-memory 20G
--total-executor-cores 100
/path/to/jar.jar localhost 1000 1 polska,poland
```

Bibliography I