Introduction to Spark

Outline

(this segment)

- What is Big Data?
- What is Spark?
- Wasn’t Hadoop good enough?
- Architecture of Spark

What is Big Data?

Big

The explosion of data

Approximately 90% of the data that existed in the world (May, 2013) was created in the preceding two years

- http://www.sintef.no/home/corporate-news/Big-Data--for-better-or-worse/

Big and Bigger

The explosion of slow and fast-moving data

Digital universe is doubling in size every two years (April, 2014)

- http://www.emc.com/about/news/press/2014/20140409-01.htm

Big and Faster

The explosion of fast-moving data

Streaming sources of Data

- Twitter
- IoT
- Log Files

What is a Big Data Solution?

And why do we need one?

- Vertical vs Horizontal Scaling
- Batch vs Stream Processing
  - Lambda Architecture

Bandwidth and Latency

What is Spark

What is Apache Spark?

Why is it important?

Apache Spark is an open source big data processing framework built around speed, ease of use, and sophisticated analytics.

It was originally developed in 2009 in UC Berkeley's AMPLab, and open sourced in 2010 as an Apache project.

Speed

- In-memory
- DAG
- Distributed (Scalable)
- Stream Processing (micro-batches)

Ease of use

- Interactive Queries
- Well-integrated
  - Clients (Scala, Python, Java, ...)
  - Other big data tools (Hadoop, Cluster Managers, Data Sources)
- Comprehensive
  - Almost every requirement is addressed
  - Rich built-in libraries

Sophisticated Analysis

- Vectors and DataFrames
- More as we proceed

Wasn’t Hadoop Good Enough?

Spark vs Hadoop

- MapReduce
  - **If computation of a value depends on previously computed values, then MapReduce cannot be used (Fibonacci)**
  - **Complex Machine Learning Algorithms (SVM)**
  - **Iterations (kmeans)**
- DAG
- Stream Processing
- Machine Learning

Computation Architecture

DAG vs MapReduce

Let’s build a graph representing the computational model:

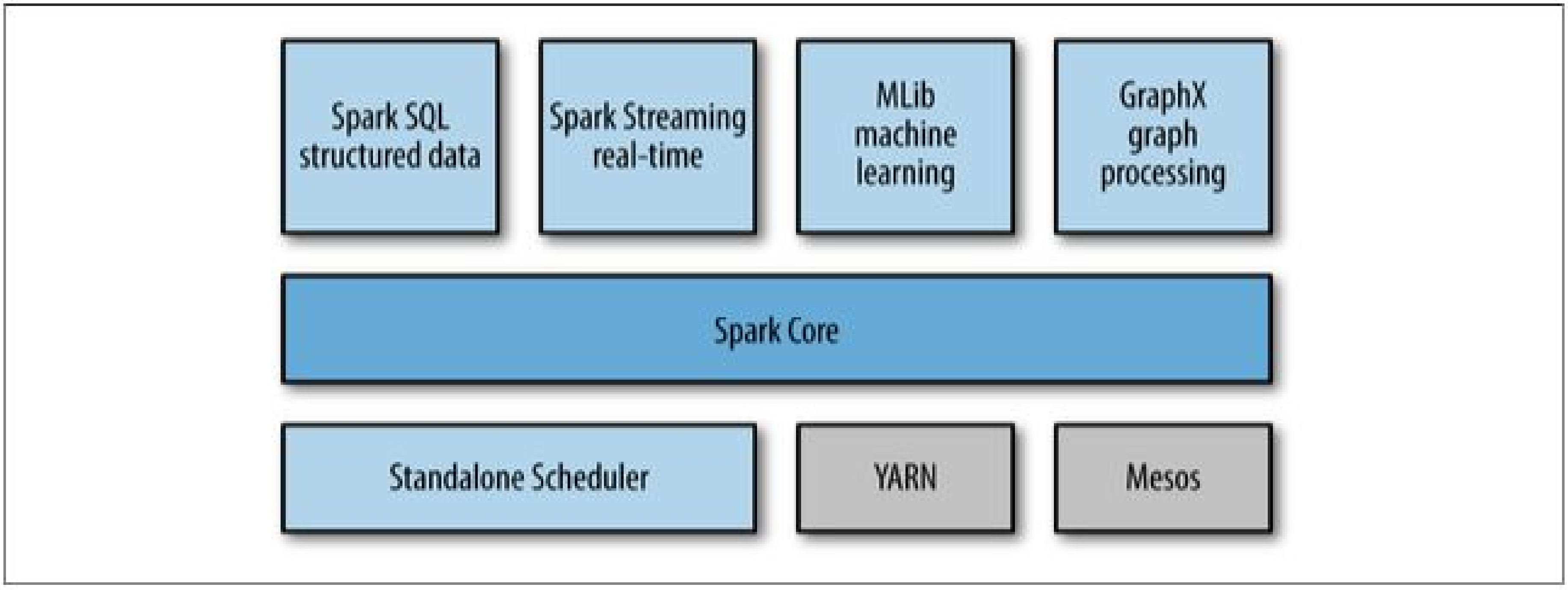
- Fibonacci

Architecture of Spark

Components of Spark

It’s a fairly complete Stack

- Spark Core
- Spark SQL
- Spark Streaming
- MLlib
- GraphX
- Cluster Manager


Spark SQL structured data
Spark Streaming real-time
MLib machine learning
GraphX graph processing
Spark Core
Standalone Scheduler
YARN
Mesos

Spark Core

- Basic functionalities
  - task scheduling
  - memory management
  - fault recovery
  - interacting with storage systems
  - ...
- RDD api

Spark SQL

- Query Data
  - SQL
  - HiveQL (Drop-in replacement for Hive and HiveQL)
- Programmatic Manipulation

Spark Streaming

- Processing of live streams of data in real time
- Extends RDD api
    - microbatches

MLlib

- Common machine learning algorithms implemented as Spark operations on RDDs
    - Classification
    - Regression
    - Clustering
    - Collaborative filtering
    - Model evaluation
- Lower-level constructs as well
    - generic gradient descent optimization

GraphX

- Manipulate graphs and perform parallel graph operations and computations
- Extends RDD api

Cluster Manager

Can run on many cluster managers

- Apache Mesos
- Hadoop YARN

Comes with one included

- Standalone Scheduler
  - We’ll get started with this

Application Scenarios

Why don’t you guys pitch in?

- ???
- ???
- ???

Application Scenarios

Spark@Amazon

- Generating Recommendations at Amazon Scale with Apache Spark and Amazon DSSTNE
    - https://aws.amazon.com/blogs/big-data/generating-recommendations-at-amazon-scale-with-apache-spark-and-amazon-dsstne/

Application Scenarios

Spark@Facebook

- Apache Spark @Scale: A 60 TB+ production use case
  - https://code.facebook.com/posts/1671373793181703/apache-spark-scale-a-60-tb-production-use-case/

Application Scenarios

More...

- https://cwiki.apache.org/confluence/display/SPARK/Powered+By+Spark