Introduction to Data Engineering and Spark Architecture

Agenda

- Peek into Data Engineering & Big Data
  - Data Engineering
  - Big Data Frameworks
- Introduction to Spark
  - Spark vs. Hadoop
  - Spark Architecture
- Setting up a Spark Cluster

Peek into Data Engineering & Big Data

Data Engineering

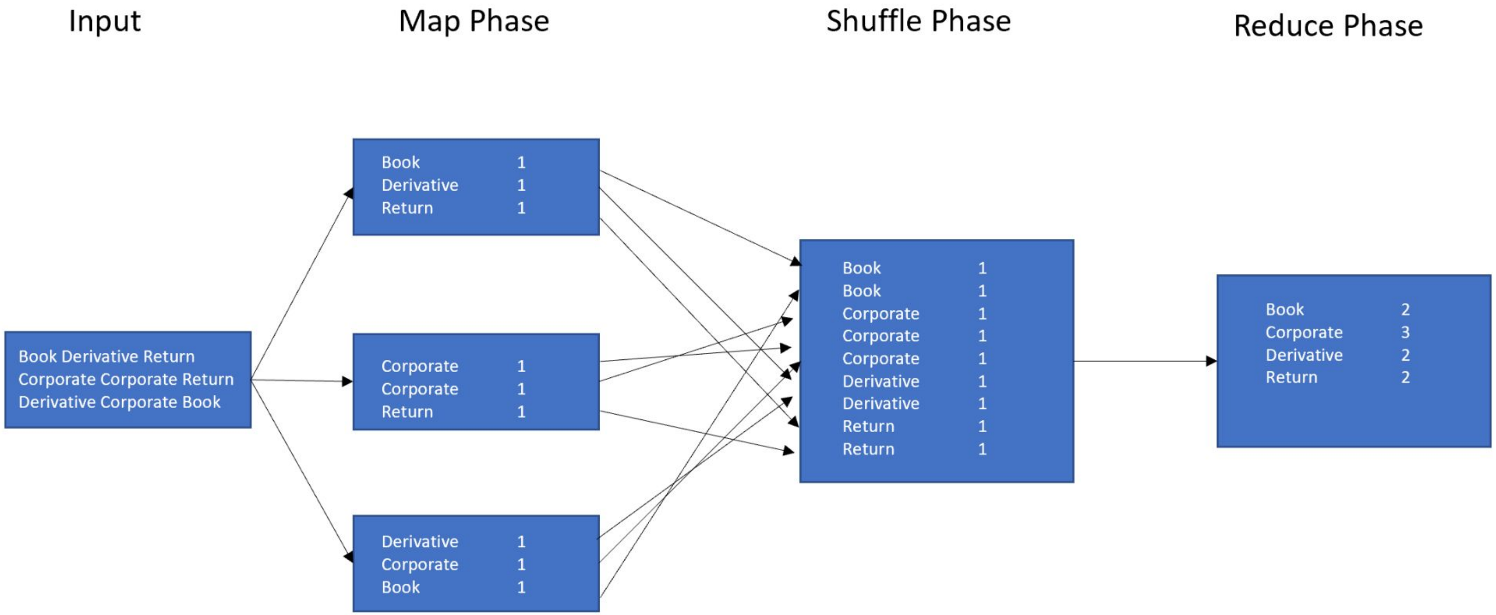
- Data Engineers support Data Scientists.
    - In charge of designing, creating, deploying, and supporting data pipelines.
    - Depending of the side of the company, the Data Scientist can be a “Full Stack Data Scientist”: in charge of its own Data Engineering.
- Skills needed
    - Computer Science
    - Business knowledge
    - Database Expertise (SQL and NoSQL)
    - Big Data Architectures
- Why Become a Data Engineer?
    - Explosion on roles over the last 5 years.
    - Unlike Data Scientist, there is no clear path.

Big Data Frameworks

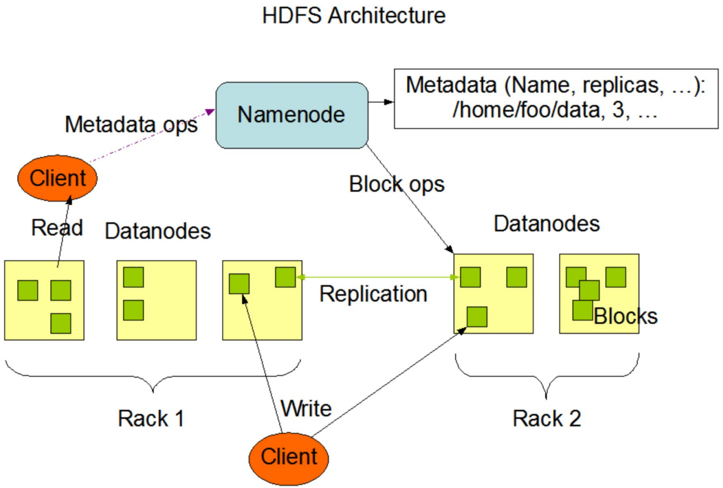
- Big Data
  - When is considered big? Loosely defined.
  - Depends on both external and internal factors.
  - Rule of Thumb: *“Too Big to fit in a Pandas Dataframe”.*
- Parallel computation
  - Divide and conquer - Map Reduce!
  - Map phase - Activities that can be done independently.
  - Reduce phase - Aggregation done at the end.
  - Data structure: Key-Value pairs

Big Data Frameworks

- MapReduce example



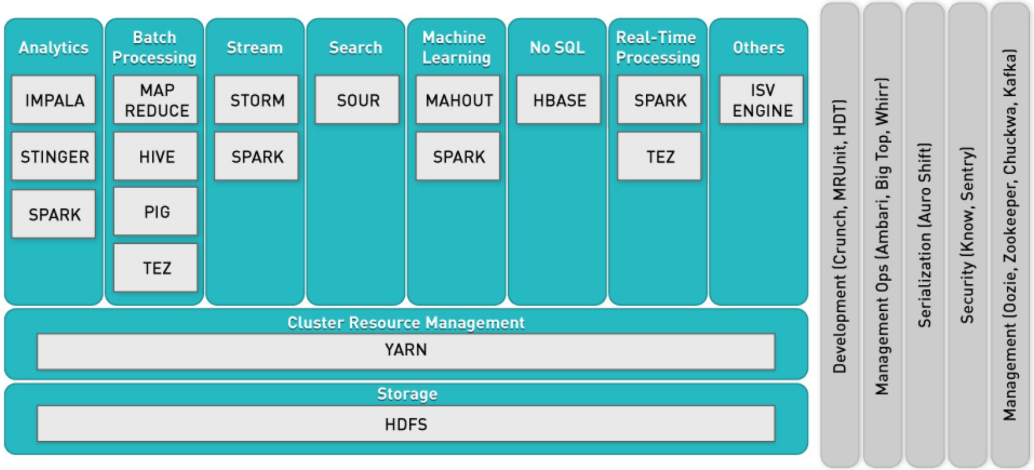
Big Data Frameworks



- Distributed computing
  - Instead of a big powerful machine - Several simpler ones
  - Code lives in Master - Work is done in Workers.

Big Data Frameworks

Apache Hadoop 2.0 Ecosystem



Analytics	Batch Processing	Stream	Search	Machine Learning	No SQL	Real-Time Processing	Others
IMPALA	MAP REDUCE	STORM	SOUR	MAHOUT	HBASE	SPARK	ISV ENGINE
STINGER	HIVE	SPARK		SPARK		TEZ	
SPARK	PIG						
	TEZ						

Cluster Resource Management

YARN

Storage

HDFS

Development (Crunch, MRUnit, HDT)

Management Ops (Ambari, Big Top, Whirr)

Serialization (Auro Shift)

Security (Know, Sentry)

Management (Oozie, Zookeeper, Chuckwa, Kafka)

http://incubator.apache.org/projects/

- Hadoop
  - Distributed File System (HDFS)
  - Manages both work distribution and fault tolerance.

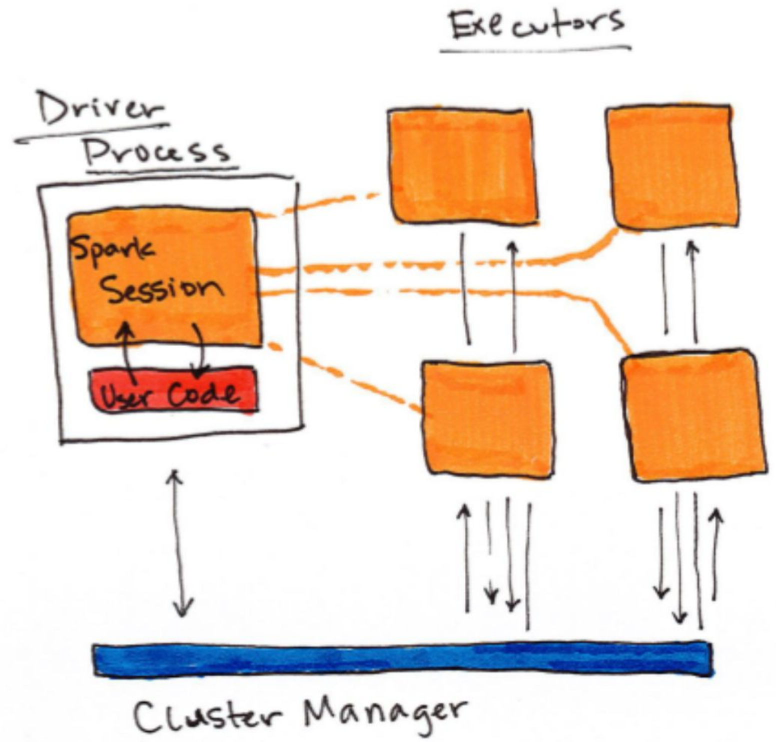
Introduction to Spark

Spark vs. Hadoop

- MapReduce has been the major framework for distributed computing
  - Hadoop’s limitations include programmability and performance.
  - Computational frameworks are becoming specialized.
- Spark: 100x faster than Hadoop
  - Spark is the Compute Engine - Hadoop still provides the environment.
- Specialized libraries for machine learning, graph processing, and database management.
- APIs include:
  - Java
  - Scala
  - Python
  - R

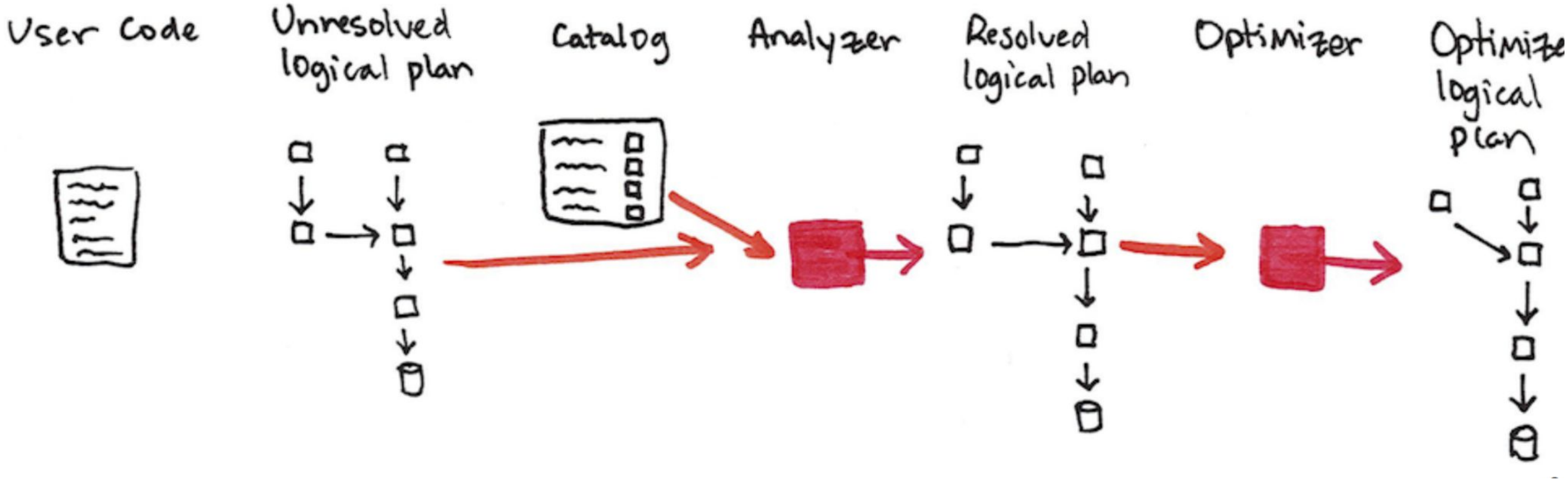
Spark Architecture (I)

- Driver and Executors



Spark Architecture (II)

- SparkContext and Lazy Evaluation



Great Learning

Happy Learning !