Introduction to Spark

Introduction to Spark

OBJECTIVES

- **Describe** the pros/cons of Spark compared to Hadoop MapReduce
- **Define** what an RDD is, by its properties and operations
- **Explain** the different between transformations and actions on an RDD
- **Implement** the different transformations through use cases
- **Explain** what persisting/caching an RDD means, and situations where this is useful

Why Spark?

Data science friendly parallel computing

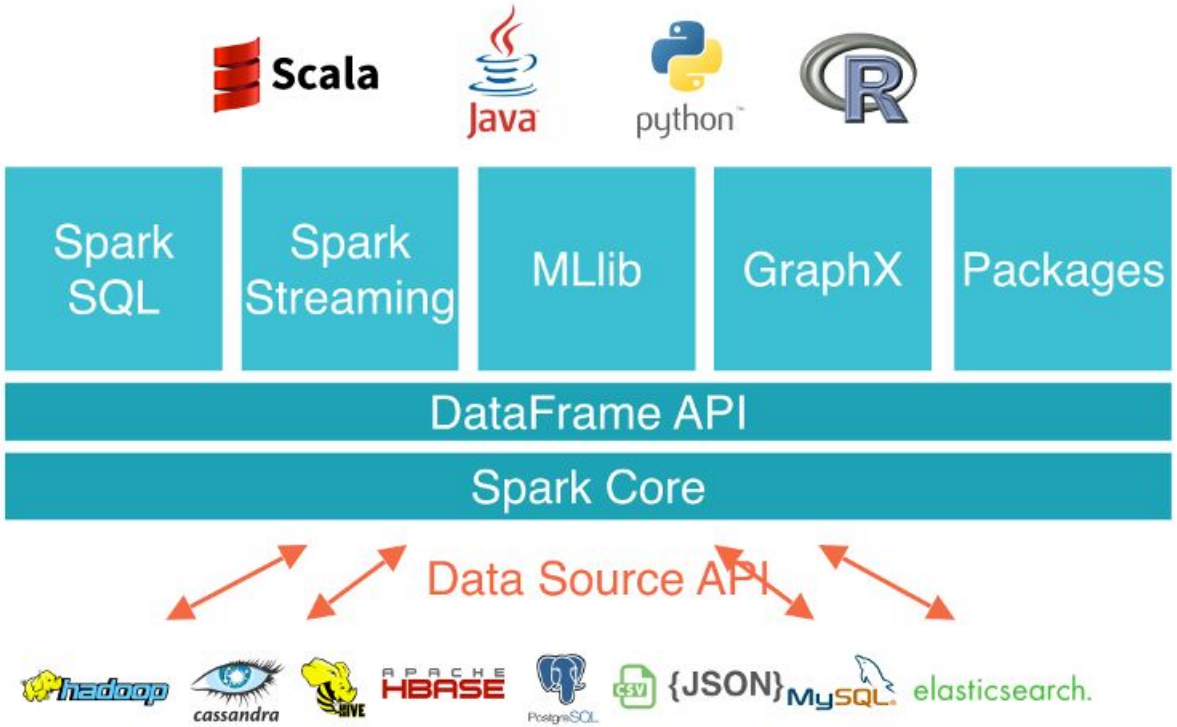
- Processing massive data sets
- Highly efficient distributed operations
- More use cases than just MapReduce
- Python and SQL supported natively

Apache Hadoop integration

- ~~Seamless~~ relatively easy integration into existing eco-systems (HDFS)
- Scalability, reliability, resilience

And… machine learning functions available!

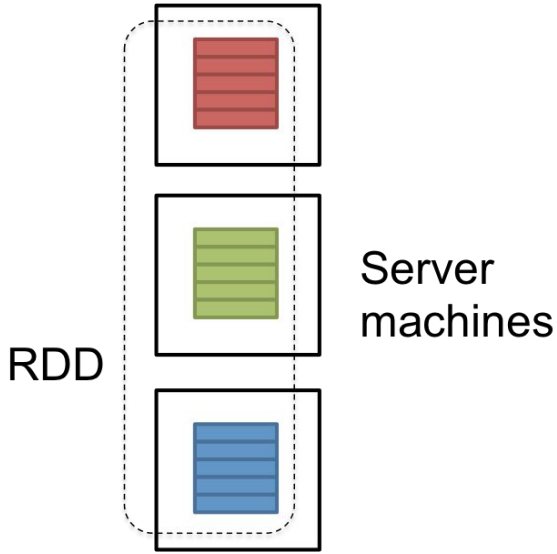
Spark Ecosystem



https://weidongzhou.wordpress.com/2015/09/08/hadoop-hdfs-mapreduce-and-spark-on-big-data/

Resilient Distributed Datasets

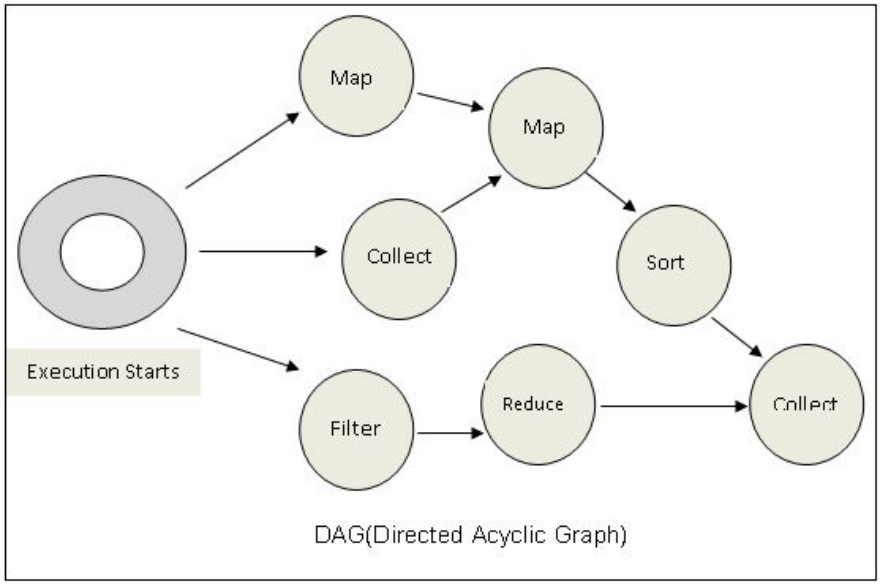
- created from HDFS, S3, HBase, JSON, text, local
- distributed across the cluster as partitions (atomic chunks of data)
- can recover from errors (node failure, slow process)
- traceability of each partition, can re-run the processing
- **immutable** : you *cannot* modify an RDD in place



[Image Source]

A “Functional” Programming paradigm

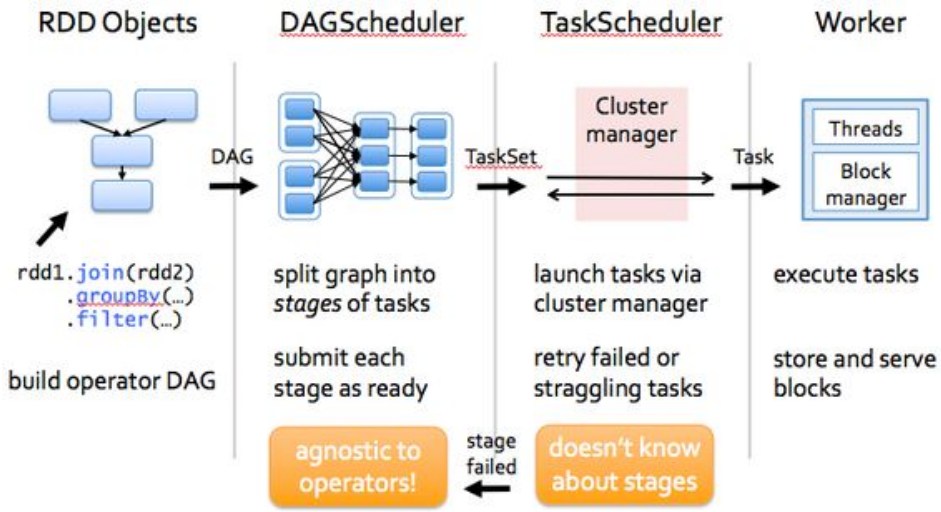
- RDDs are immutable !
  You can **only transform** an existing RDD into another one.
- Spark provides many **transformations functions**.
- Programming = construct a **Directed Acyclic Graph** (DAG).
- **Passed from the client to the master**, who then distributes them to workers, who apply them accross their partitions of the RDD.



[Image Source]

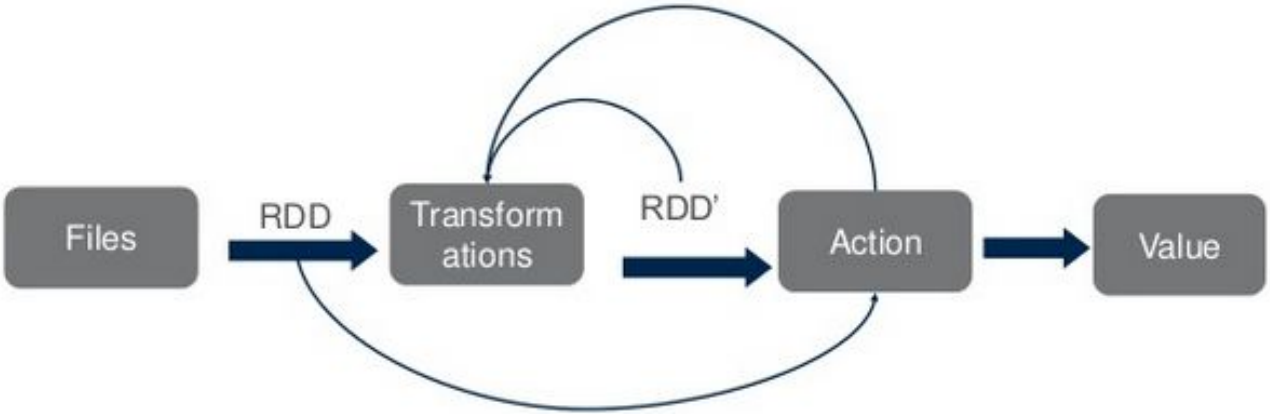
Directed-Acyclic-Graph

- You construct your sequence of transformations in python.
- Spark functional programming interface builds up a DAG.
- This DAG is sent by the driver for execution to the cluster manager.



[Image Source]

Operational Spark Workflow



Brainstorming: So, let's suppose you have this thing called an RDD, which is just basically a dataset made of rows and values.

What are all the operations you'd like to do to that RDD ?