Introduction to Apache Spark

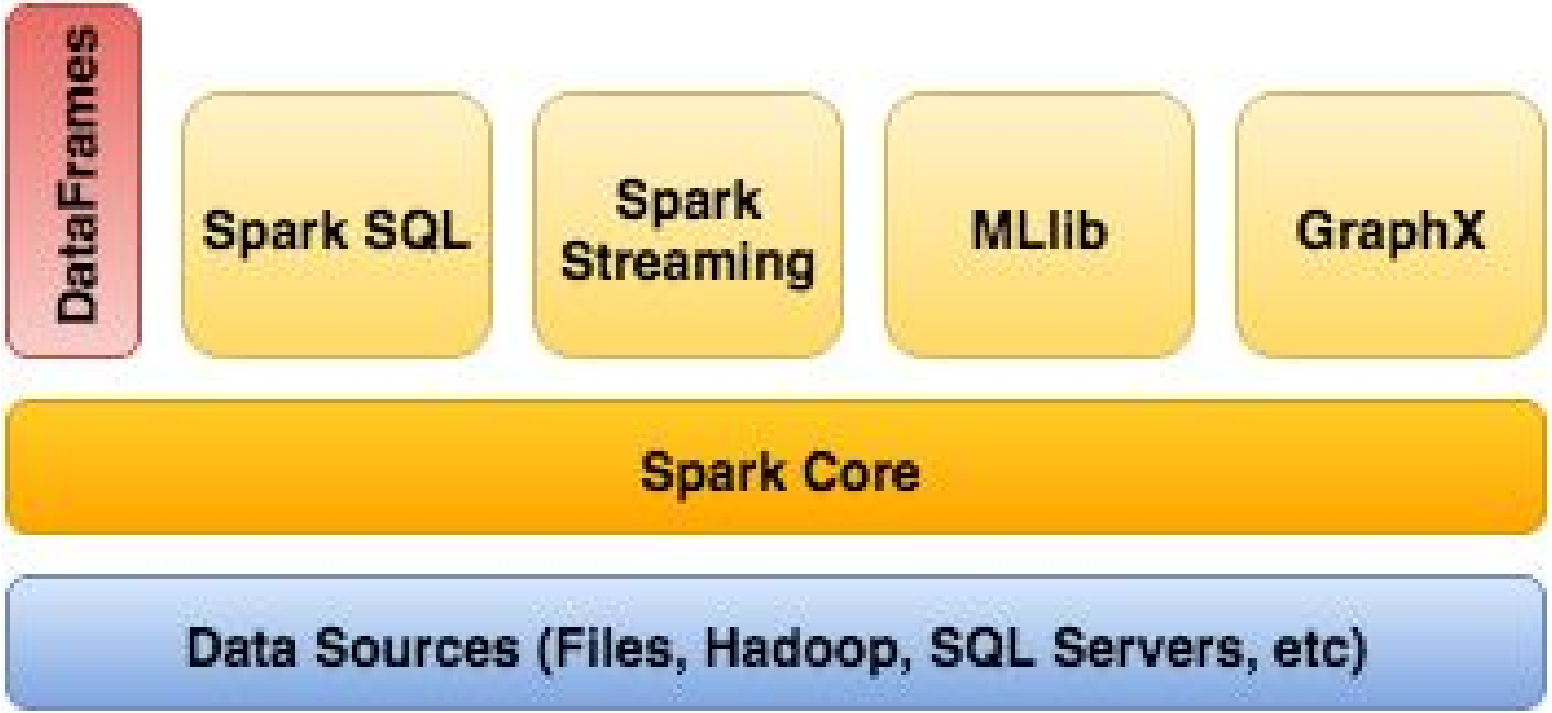
Goals

- Understand the general idea of how Apache Spark works
- Become aware of the tools available in Spark for Data Scientists
- Learn the benefit of Spark over MapReduce
- Understand what RDDs are and how it works

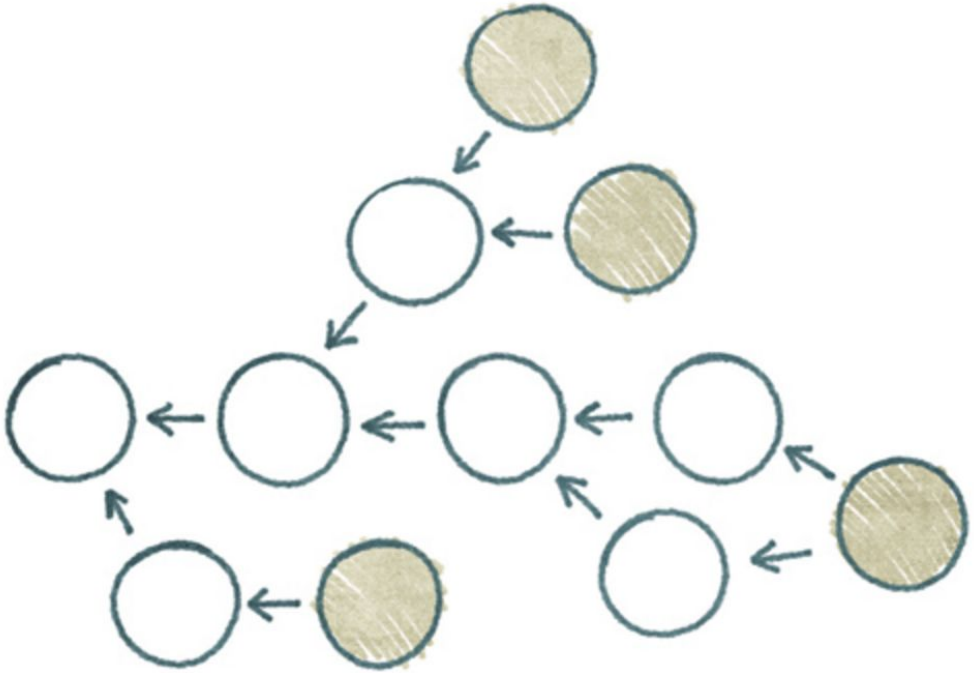
Spark Good to Knows

- Open source distributed data processing framework
- Is a MapReduce paradigm focused on memory usage
- Available in Python, Java, and Scala
- Is an unified system that includes SQL, ML, DataFrames, etc
- Up to 100 times faster than Hadoop MapReduce
- Great for tasks involving multiple iterations on the same data
- Fault Tolerance
- Lazy in operation

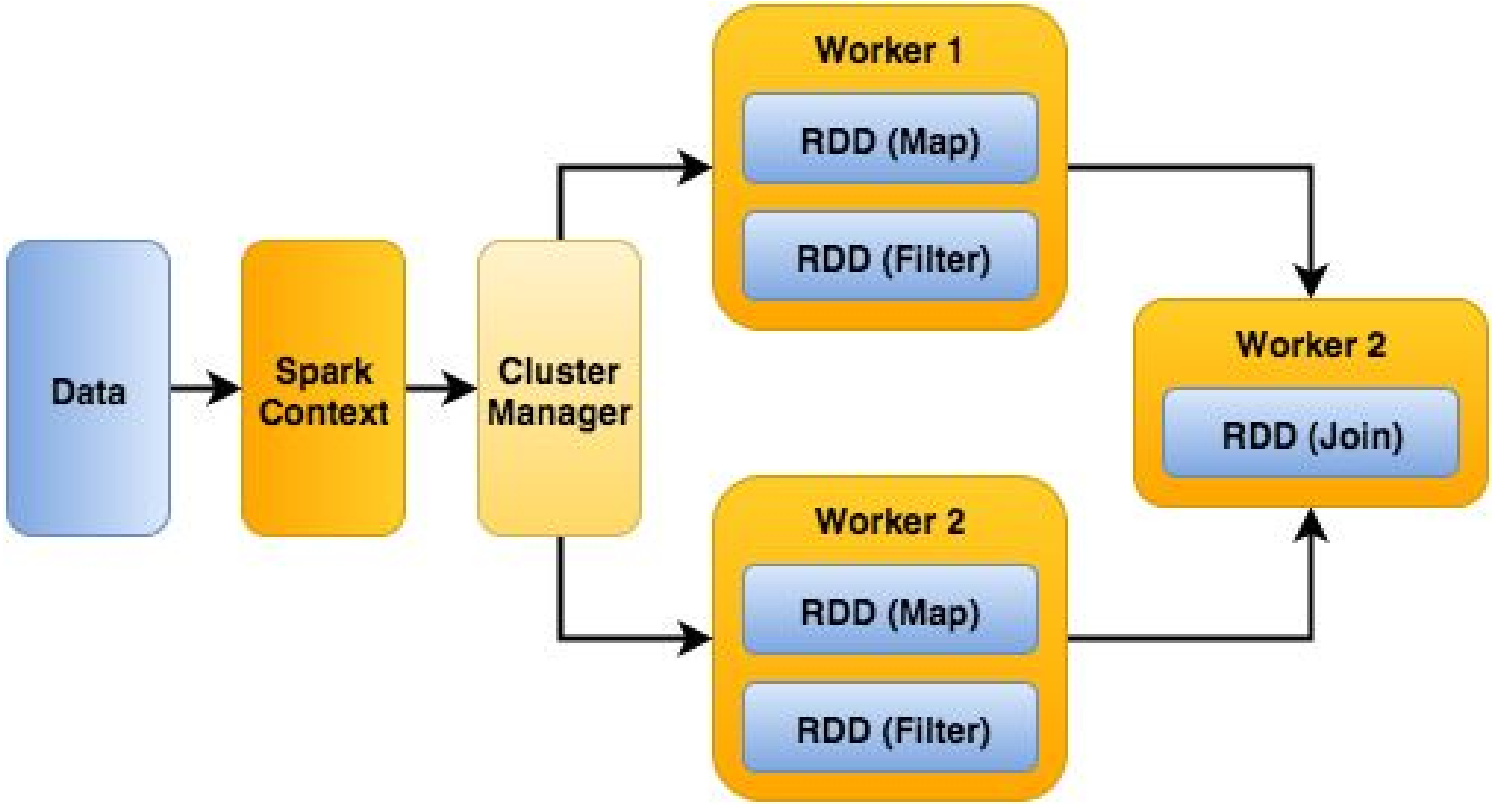
Apache Spark



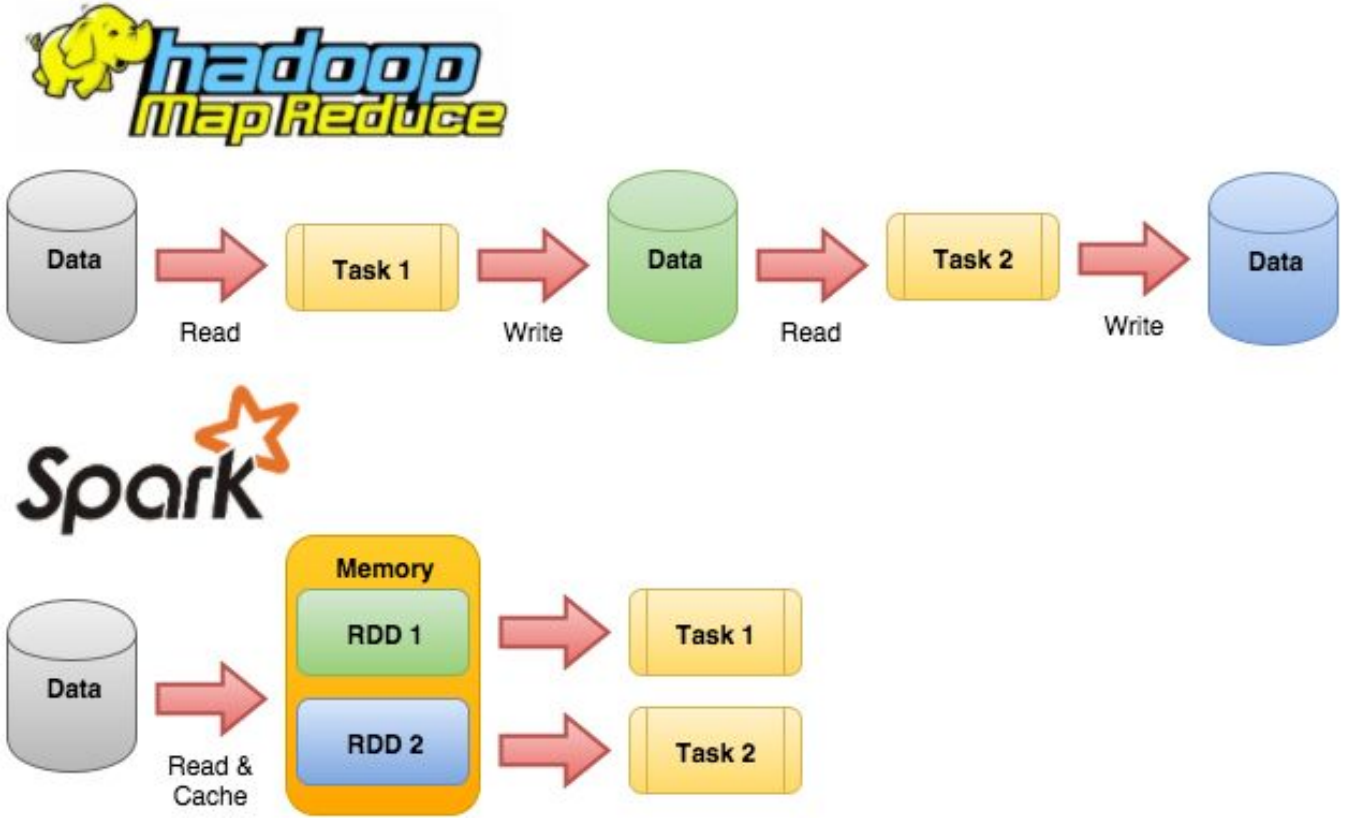
Concept of Spark - Directed Acyclic Graphs (DAG)



Apache Spark Architecture



Comparison of MapReduce and Spark



RDDs (Resilient Distributed Datasets)

They are:

- Partitioned collections of data
- Distributed across clusters
- Lazy (Does not execute until forced to)
- Can be persisted (cached)
- Functional Programming (Declarative and uses building blocks)

RDDs (continued)

- Constructors
  - parallelize, textFile
- Transformations
  - map, filter, union, distinct,
- Actions
  - count, sum, mean, collect

Until you execute an action RDD, transformation RDDs will not run. However, you can cache your RDDs to avoid the doing the same transformation or action again in applications where similar tasks may be needed repeatedly.

RDDs - Constructors

Constructors creates the RDDs that you will use.

- Some common construction RDDs are:
  - parallelize - Takes an iterable (list) and creates a RDD
  - textFile - reads text file from some file system and creates a RDD
  - hadoopFile - reads a Hadoop file and creates a RDD

RDDs - Transformations

Transformations mutate a given RDD and creates another RDD with the mutation.

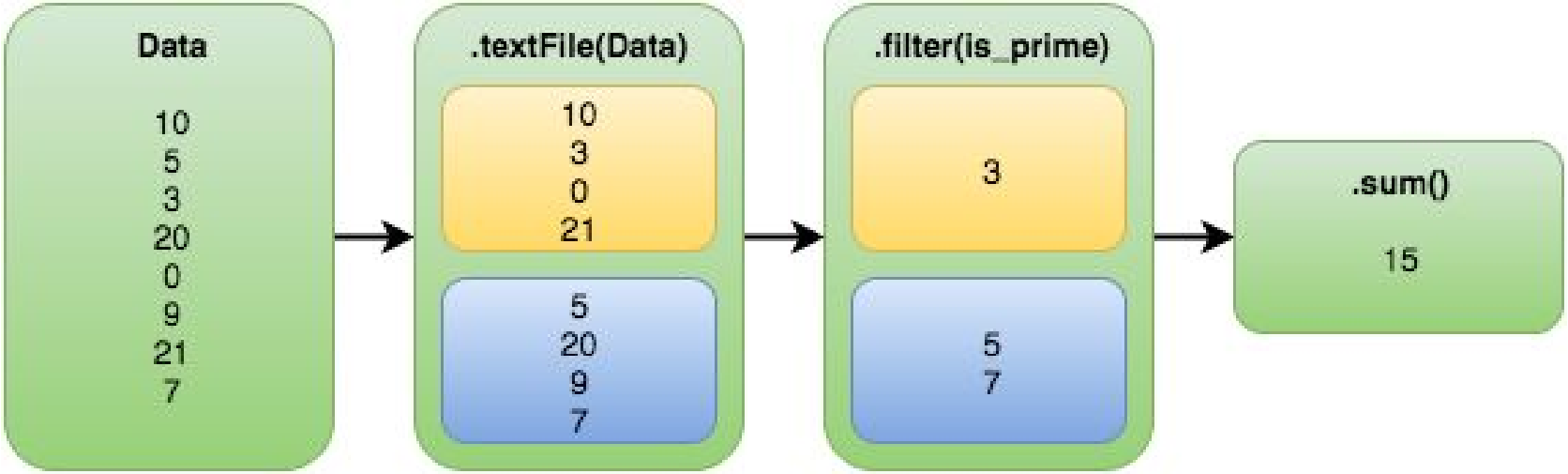
- Some common transformation RDDs are:
  - filter - filters RDD by some boolean condition
  - map - applies some function to RDD
  - flatMap - applies some function that returns an iterator and flatten
  - groupByKey - takes a key value RDD and collapse to unique keys and list of values

RDDs - Actions

Actions initializes the constructor and transformation RDDs, and returns some values as a result of the action.

- Some common action RDDs are:
  - sum
  - mean
  - stdev
  - collect - returns values in RDDs as a list

RDDs Visualized



Functional Programing

A declarative and building block style of programming to avoid mutable data.

IE: construct_rdd(from data).filter_data(some condition).take_sum()

Using the previous RDDs visualized example and actual Spark functions:

sc.textFile(Data).filter(is_prime).sum()

*sc is the SparkContext