Apache Spark

* Adds even more flexibility
* Has a rich ecosystem on top
  + ML, Data Mining, Graph Analysis, etc.
* Can use programming languages to write code for
  + Java, Python, or Scala
* Scalable
* Written in Scala
  + Python is slow in comparison
    - Python is better for prototyping
  + Similar syntax to Python(?)

Structure

1. Driver Program (Spark Context)
2. Cluster Manager
   1. Spark’s own cluster manager
   2. YARN
   3. MESOS
3. Executor
   1. Cache
      1. Spark is a “Memory based solution”
      2. RAM
      3. DAG
      4. Fast due to utilizing cache (memory)
   2. Tasks

Benefits

* DAG Engine to optimize workflow
* A lot faster than MapReduce

Resilient Distributed Dataset (RDD)

* Object that represents a dataset
* Various functions (API) that can transform the dataset object

Components

* Spark Core
* Spark Streaming
  + Real-time instead of batch
  + Webserver logs fed in real time
    - Analyzed in real time
* Spark SQL
  + SQL Interface to Spark
  + Use SQL like functions
  + This allows SQL optimizations
* MLLib
  + MapReduce tasks for ML-related
  + Simplifies complex MapReduce formations
* GraphX
  + Graph Theory
  + Analyzing Graphs
  + i.e. Social Network Graphs