Assignment 3

1. (25 pts) What is HBase? What’s its main differences with HDFS?

HBase is a NoSQL datastore on top of HDFS (used for storage) with a column family oriented database. With HBase, data can be retrieved quickly or batch processed with MapReduce

1. (25 pts) How many main components in YARN and what are the responsibilities of each component?

Resource Manager - Master of processing layer. Replace Job Tracker of Hadoop1

Scheduler - Takes care of allocating resources

Applications Manager - negotiates resources with node managers

Node managers - Slave of the processing layer. Replace Task Tracker of Hadoop1. Constantly gives resource manager updates about resources and application

App Master - Temporary Denim that takes care of execution of tasks in other containers

Containers - Resources. Using the containers, the app master runs processing tasks within the designated resources

1. (25 pts) List 5 main differences between MapReduce and Yarn. And give brief explanations.
   1. Hadoop 1 uses HDFS and MapReduce, but Hadoop 2 uses HSFA and YARN (usually called MapReduce version 2)
   2. YARY prevents single point failures because of Active name node and standby name nodes. -> When the active node stops the passive node takes over as the active to make sure execution continues
   3. MapReduce has a single master and multiple slave architecture. This means that in a single point of failure the entire slave stops working. In Hadoop2 with YARN, if one master goes down, there are multiple others to resume its process
   4. YARN is responsible for resource management, but MapReduce is a framework responsible for executing a certain job.
   5. Each node in MapReduce is run individually, but each node in YARN is run by a node manager
   6. YARN uses a resource manager for each cluster in addition to each data node running on a node manager -> one slave node monitors resources/tasks for each job. MapReduce keeps some nodes idle because it uses a job tracker to create/assign tasks to a tracker.

1. (25 pts) List 5 key benefits of YARN.
   1. Scalability - cluster size of 10,000+ nodes and run 1,000,000+ tasks at once
   2. Compatibility - Applications for Hadoop1 still run on YARN
   3. Resource Utilization - Dynamic allocation of resources to clusters
   4. Multitenancy - Open-source data access engines with real time analysis and ad-hoc query
   5. Redundancy - With a multiple master and slave architecture, YARN will resume its process and prevent single point failures