CS4830 Assignment 2

Shashank H S BE18B006

1. Files are in the zip folder.

2. a. Hadoop Distributed File System (**HDFS**) – it is the storage component of Hadoop. Large datasets and files need to be stored in a distributed manner across a cluster of machines since one disk will not have enough storage. HDFS is the distributed file system which manages the files across a cluster of machines. HDFS can detect errors and automate the recovery from hardware failures. It can also accommodate the scaling of the cluster and its data from a few nodes to hundreds / thousands of nodes in a cluster.

b. **Hive** - it is an open-source, distributed, fault-tolerant data warehouse framework for analysis and querying of data stored in HDFS. It is developed on top of Hadoop. It provides SQL-like declarative language, called HiveQL, to express queries. Using HiveQL, users can read, write and manage petabytes of data.

c. **Pig** - Pig is a high-level data flow platform for analysing large datasets and executing MapReduce programs of Hadoop. The structure of Pig programs is such that they are amenable to massive parallelization which enables them to handle very large data sets. Pig's infrastructure layer consists of a compiler that produces sequences of Map-Reduce programs, for which large-scale parallel implementations already exist (e.g. the Hadoop subproject). The language used for Pig’s language layer is Pig Latin. This language uses less lines of code for any operation and it is flexible to be used for unstructured and structured data. It also contains several useful in-built operators like sort and filter.

d. Yet Another Resource Navigator (**YARN**) – it is a resource management framework which works like an operating system for a cluster of machines and sets the execution environment for the machines. YARN divides resource management functionalities into global ResourceManager (RM) and per-application ApplicationMaster (AM). An application is the unit of scheduling on a YARN cluster, which could either be a single job or a directed acyclic graph of jobs. The ResourceManager and the NodeManager form the data-computation framework. The ResourceManager is the ultimate authority that allocates resources among all the applications in the system. The NodeManager is the agent who is responsible for containers, monitoring resource usage (CPU, memory, disk, network) for each machine and reporting the same to the ResourceManager/Scheduler

References:

1. https://www.ibm.com/topics/hdfs
2. https://aws.amazon.com/big-data/what-is-hive/
3. https://pig.apache.org/
4. https://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html