**Assignment 02**

Apache Hadoop and Distributions

**Name: Dean D’souza**

**H.U. ID: 168424**

# Apache Hadoop

Apache Hadoop is an open-source software library which aims at providing a framework for distributed file storage and distributed processing of large data sets residing on commodity grade hardware in the form of clusters of computers. This framework was developed as a result of two papers published by Google, the Google File System paper and the MapReduce: Simplified Data Processing on Large Clusters, and was mostly developed in Java programming language, with native code and command line utilities written in C and as shell scripts respectively.

In order to process large data sets,

The most important components of the Apache Hadoop framework are:

1. the Hadoop Distributed File System (HDFS), which is responsible for storage and high-throughput access of large data sets.
2. MapReduce, which is a YARN (Yet Another Resource Negotiator) based system for the distributed and parallel processing of such data sets.

The Apache Hadoop framework also has libraries and utilities required by the different modules stored in the Hadoop Common Module and a job scheduling framework that manages cluster resources, i.e., the computing resources in each cluster, and their utilization for user applications, known as Hadoop YARN.

# References:

[1] Apache Hadoop homepage, The Apache Software Foundation, (<http://hadoop.apache.org>)

[2] Apache Hadoop Wikipedia, (<https://en.wikipedia.org/wiki/Apache_Hadoop>)

[3] Cloudera vs Hortonworks vs MapR: Comparing Hadoop Distributions, Experfy Editor, (<https://www.experfy.com/blog/cloudera-vs-hortonworks-comparing-hadoop-distributions/>)