DBMS Models and implementation
Instructor: Sharma Chakravarthy
Project 3: Data Analysis using Map/Reduce

Made available on:	**10/24/2019**
Submit by:	**12/3/2019 (11:55 PM) No extension for this project**
Demo on:	**12/4/2019 (Choice of Slots will be given)**
Submit to:	**Canvas (1 zipped folder containing all the files/sub-folders)**
Weight:	**15% of total**
Total Points:	**100**

One of the advantages of cloud computing is its ability to deal with very large data sets and still have a reasonable response time. Typically, the map/reduce paradigm is used for these types of problems in contrast to the RDBMS approach for storing, managing, and manipulating this data. An immediate analysis of a large data set does not require designing a schema and loading the data set into an RDBMS. Hadoop is a widely used open source map/reduce platform. Hadoop Map/Reduce is a software framework for writing applications which process vast amounts of data in parallel on large clusters. In this project, you will use the IMDB (International Movies) dataset and develop programs to get interesting insights into the dataset using Hadoop map/reduce paradigm. Please use the following links for a better understanding of Hadoop and Map/Reduce (https://hadoop.apache.org/docs/stable/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html)

1. **Installation: There are two options**
   A. **RECOMMENDED - Hadoop Single Node Cluster Setup (Hadoop 2.9.1):** We advise you to use a Linux installation such as Ubuntu 16.04 on your system in order to breeze through the steps of Hadoop cluster set up. You can use a virtual machine for this purpose. The steps to install a *single node cluster in the pseudo distributed mode* are given here - https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-common/SingleCluster.html . Some important points that must be considered while setting up the cluster are:
      - Preferred Mirror site for download: http://mirror.metrocast.net/apache/hadoop/common/hadoop-3.2.1/
      - After completing the prerequisites and running the command ($ bin/hadoop), jump to https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-common/SingleCluster.html#Pseudo-Distributed_Operation, and follow the steps for Configuration, Setup passphraseless ssh, Execution (Only steps 1 – 4) and YARN on single node (Steps 1-3)
      - The cluster will start up by running the commands,
        $ sbin/start-dfs.sh
        $ sbin/start-yarn.sh
        which you must have done while following the above steps.
      - To check if the Namenode, Datanode, Secondary Namenode, ResourceManager, NodeManager are running as separate processes, run the command
        $ jps

- Once the cluster is up, run the most basic code, WordCount, that counts the number of occurrences of each word in the input files
  - Download the WordCount.java code from canvas into the updated Hadoop folder
  - Make a directory (example, inputFiles) on the HDFS (Hadoop Distributed File System) to store the input files
    $ **bin/hdfs dfs -mkdir** /inputFiles
  - Copy a file from the local file system to the HDFS using the command
    $ **bin/hdfs dfs -put** <path_of_file_on_local_system> /inputFiles
  - To execute the code, follow the steps given here, https://hadoop.apache.org/docs/stable/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html#Example:_WordCount_v1.0
- To stop the cluster, run the commands
  $ sbin/stop-dfs.sh
  $ sbin/stop-yarn.sh
- Some useful shell commands for the HDFS can be found here, https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/FileSystemShell.html

B. **Cloud Services (self-exploration required):** The second option is to use Amazon Elastic Map/Reduce. Amazon EMR (Elastic Map/Reduce) is a web service provided by Amazon that uses Hadoop and distributes large datasets and processes them into multiple EC2 instances. You have to sign up for AWS. Please make sure you read carefully your AWS agreements/contracts/free use. You may have sufficient free services to complete the projects, but you should monitor and understand your use. Don't leave things running when not necessary. Note that if you exceed your monthly quota, you will get charged. Also, a credit/debit card is necessary for signing up for this service. You may use other cloud environments like IBM cloud, Google cloud or Microsoft Azure. **For these, we will not be able to provide assistance.** There should be online help available for different cloud providers regarding how to set up a cluster and run the first basic map/reduce code.

Note that if you use Amazon/ECS, you can use multiple mappers and reducers and will be able to better understand and appreciate distributed aspect of map/reduce and how the response time changes (should decrease) with more resources. You may not be able to do this on your installation of Hadoop on your laptop or desktop.