Lab 01	Map Reduce Programming
IT413 No SQL Databases, Winter'2020, DAIICT, Gandhinagar; pm_jat	

In this lab we practice some Map Reduce programs. Do following-

1. Setup a single node cluster Hadoop on your laptop. May take help from instructuction at https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-common/SingleCluster.html

2. Download source code and data of maximum temperature problem discussed in the lecture, and run the program on your hadoop setup.

   Source code: https://github.com/tomwhite/hadoop-book/tree/master/ch02-mr-intro/src/main/java

   NCDC data:

   1901: https://github.com/tomwhite/hadoop-book/blob/master/input/ncdc/all/1901.gz

   1902: https://github.com/tomwhite/hadoop-book/blob/master/input/ncdc/all/1902.gz

   You can download both file, and have combined file as input!

3. You are given a file web access log file ("web_access_log") produced by a web server. You can understand about the content of web access log file from https://httpd.apache.org/docs/1.3/logs.html#common

   You are required to create a Map Reduce program called "ImageCounter" that counts the number of times GIF, JPG, and other image files that have been accessed by clients.

   Your MR job should output three figures: number of gif requests, number of jpeg requests, and number of other images!

4. Create another Map Reduce program that reads from same log file, and outputs following summaries on Monthly Basis –

   (a) Total number of requests.

   (b) Total download size (in Mega Bytes)

   It should output: <Year-Month, Number of Requests, Download Size> for every months like Dec-2016, Jan-2017, and so!

5. Create another Map Reduce program that lists Timestamp, URL for which http response status has been 404.

   You should able to do this by creating a map only MR job. Some idea you should get from following blog:
   https://data-flair.training/blogs/map-only-job-in-hadoop-mapreduce/

6. [Optional] Download Left Join source code from here; have some sample data, and attempt running the code.