CSE 3244: GCP Lab Assignment

The overall goal of this lab is to make you comfortable with:
- Using public cloud service such as AWS, Google Cloud Platform(GCP) and learn serverless functions.
- Handling big datasets and pre-processing them.
- Storing data in cloud relational databases to perform queries
- Writing queries and executing them using command line or any other tool
- Using Map-Reduce to extract structure from text
- Using Map-Reduce to create dimension tables for OLAP workloads
- End-to-end business intelligence using structured & unstructured data

You will work in groups of 4-5 students. Each group will be composed carefully to include people with Java and Python coding experience, access to GCP, and experience with data analysis. Everyone in the group will receive the same grade. If a group votes unanimously to remove a non-participating member and writes 2-page document explaining their reasoning and efforts to engage with the non-participating member, then that person will be removed from the team. That person will still have the opportunity to work alone to submit tasks.

Using GCP: Note, you have to pay to use GCP. This assignment can be completed for less than $40 USD. However, overruns can happen if you fail to terminate instances or employ costly resources. You are warned! Always, make sure to manage GCP costs. ***No extra time or points will be given for technical issues related to GCP costs.***

Grading: Each group will be given 9 tasks to complete by the last day of class. Everyone in the group will receive the same grade based on the number of tasks completed. Groups that complete fewer than 5 tasks will receive zero points. Groups that complete fewer than 7 but more than 5 tasks will receive 70 points. 8 tasks earns is 80 points. Groups that complete all tasks will be eligible for 100 points depending on the subjective assessment of the last task (worth up to 20 points).

Tasks Completed		Points
Lower range	Upper range	
0	5	50
5	7	70
8	8	80
9	9	80-100 (quality of last task)

Submission: Each task is confirmed via screen shot and/or report as detailed in the task description. Submit a single PDF on Carmen that includes all screenshots and reports.
Background information before starting:
(i) What is GCP? https://cloud.google.com/why-google-cloud/
(ii) What is Docker?https://docs.docker.com/v17.09/engine/userguide/storagedriver/imagesandcontainers/#data-volumes-and-the-storage-driver
(iii) What is VM? https://www.redhat.com/en/topics/virtualization/what-is-a-virtual-machine

TASK DESCRIPTIONS:

1. Use GCP to create a cloudera quickstart Docker container and create tennisDb using MySql terminal. Under Carmen Files CSE3244_GCPLab/, access GCPLab_Tasks1thru7.pdf and follow the tutorial. You will create GCP instances running the Cloudera quickstart Docker container.

2. You are given raw data collected about tennis players, matches and rankings in CSV format below. We will run OLAP workloads against the data. First, you must store in Star in Schema. In this setup, the Fact table will be world tennis rankings from the 1980's to current. The singular dimension table will describe players (wta_players.csv). Integrate the data and view using HUE or MYSQL. (More detailed instructions in GCPLab_Tasks1thru7.pdf). **Submit a screenshot of the HUE web interface and the GCP command line.** ***[DataSet Source: https://github.com/JeffSackmann/tennis_wta]***

3-7. Write queries and verify the output. (Refer to **GCPLab_Tasks1thru7.pdf**). **Submit the queries developed and screenshots of the output for each query.**

8. Access GCPLab_Tasks8an9.pdf. Fill in the blanks on the MapReduce program template to count the number of appearances for each player in qualifying matches since 2000 (wta_matches_qual_itf_20XX). Output the results in CSV format in an HDFS directory.
Submit a screenshot of (1) the MapReduce execution and (2) the files in the output HDFS directory

9. Write your own MapReduce program to extract structure from the qualifying matches. Link the resulting table to the tennis data set and write a 1-3 page report about the business intelligence learned from linking your new table to existing results. Feel free to produce graphs. **Submit only the report (the more analysis, the higher the grade).**

This lab was created by students in the ReROUT Lab. Specifically, I want to thank Naveen Babu (MS,2020) for creating the initial version of this project and iterating over 2 semesters. I also must thank Zichen Zhang (current 2021, PhD student).

-Chris