Data Analysis and Visualization Using Hive and Tableau
Project II

Instructor: Dr. Daisy Zhe Wang

TA: Kun Li *kli@cise.ufl.edu*

February 16, 2014

Department of Computer and Information Science and Engineering
University of Florida

1 Project description

1.1 Project Overview

In this project, you are required to analyze two datasets using Hive and visualize the results using Tableau. The two datasets are the Enron dataset and the Netflix dataset.

1. Enron Dataset: This dataset was collected and prepared by the CALO Project (A Cognitive Assistant that Learns and Organizes). It contains data from about 150 users, mostly senior management of Enron, organized into folders. The corpus contains a total of about 0.5M messages. This data was originally made public, and posted to the web, by the Federal Energy Regulatory Commission during its investigation. The dataset resides in Amazon S3: email dataset. The dataset consists of one database table: (primary key are highlighted)

   - Emails(*eid:varchar*, timestamp:varchar, from:varchar, to:varchar, cc:varchar, subject:varchar, context:varchar)

   The data columns are separated with tab in the dataset.
2. Netflix Dataset: Netflix provided a training data set of 100,480,507 ratings that 480,189 users gave to 17,770 movies. The user and movie fields are integer IDs, while grades are from 1 to 5 (integral) stars. The data set consists of two tables *movie_titles* and *movie_ratings*. The two data files reside in Amazon S3: movie titles. movie ratings. (primary key are highlighted)

   - movie_titles(*mid:integer*, yearOfRelease:integer, title:varchar)
   - movie_ratings(*mid:integer, custormer_id:integer*, date:varchar, rating: integer)

   The data columns are separated with comma in the dataset.

1.2 Project Report & Presentation

You are required to submit a maximum 6 pages PDF report. Name it report.pdf. At the very beginning of the report, you need to describe the division of labor for each group member in your group. Your report should at least answer the following questions for both datasets separately.

1. What is your data processing pipeline? (Graphs and words description)
2. What kind of analytics do you apply on the dataset? What are the Hive queries?
3. Which visualization do you use on the dataset using Tableau?
4. What are the programing lessons? And what are the good resources you found?
5. What is the runtime experience for queries and visualizations?
6. What difficulties you faced and what you learned from this project?

You are also required to submit a project presentation. You can make the presentation using MS PowerPoint, iWork Keynote and other tools. Finally it needs to be converted into a PDF format file named 'presentation.pdf'

1.3 Tutorials

You can have a look at the following tutorials that might be helpful in doing this project:

1. http://hive.apache.org/
2. https://cwiki.apache.org/confluence/display/Hive/Home
3. http://docs.aws.amazon.com/ElasticMapReduce/latest/DeveloperGuide/emr-hive.html
4. http://hortonworks.com/products/hortonworks-sandbox/#tutorial_gallery
5. http://aws.amazon.com/redshift/partners/tableau/
6. http://www.tableausoftware.com/learn
7. https://www.cs.cmu.edu/~enron/
8. http://en.wikipedia.org/wiki/Netflix_Prize

1.4 Grading Guidelines

In the project part I, you will primarily be graded based on the correctness of your implementation. For project part II, we will grade it based on the report and presentation only.

Components	Enron Dataset	Netflix Dataset
Grade(%)	50	50

Your submission will be graded based on the following criterion.

1. How well is the report/presentation written(Graphs/Tables and description of development&results).
2. Interesting and complexity of your hive query.
3. Interesting and complexity of your visualization.
4. A good description of the processing pipelines & experience with AWS Hive.
5. Interesting insights/discussion on data, development, results and visualization.
6. A good description of the lesson learned.

1.5 Project Submission

You project submission should include source code, a project report and a project presentation. Your project submission layout should be as follows:

```
Project2.tar
       report.pdf
       presentation.pdf
```

```
src/all the source code
```

You should tar all you source code files and the report in a single tar file using this command on Linux/SunOS:

tar cvf Project2.tar report.pdf presentation.pdf src/*

1.6 Questions and Answers

All the questions should be asked through Piazza. The TAs should answer your questions *ASAP*. Please let the TAs know *ASAP* if you find any ambiguity or any issue which may lead to non-unique outputs through Piazza. Bonus points are possible!