CIS 6930 Project 3 Report

Tae Seung Kang

In the project report you need to describe the implementation and performance results of the programs in both tasks.

A. Task 1 (OpenMP)

1. Parameters

The following is the illustration of data processing pipelining with table dependencies.

Damping factor = 0.85

2. What kind of analytics do you apply on the dataset? What are the Hive queries?

We analyze the data based on the following questions. For Hive queries, refer to the attachments (/src/enron folder). The HiveQLs for base tables are in /src/enron/ken.sql and /src/enron/steve.sql.

3. Which visualization do you use on the dataset using Tableau?

[Yifei] For visualization by Tableau, I used horizontal bars, pie chart, packed bubbles to make the plots.

4. What are the programming lessons? And what are the good resources you found?

- Whenever we found a bug, we needed to rerun the relevant queries.

- Aggregate functions such as collect_set, collect_list are very useful to deal with recipient and cc column which consist of an email list separated by comma.

5. What is the runtime experience for queries and visualizations?

- It was surprising that inserting the enron dataset which is almost 1GB took only a few minutes.

6. What difficulties you faced and what you learned from this project?

- The data given to us didn't fit into Hive format as the enron data file contains ^M character which is considered as tab. Hence, we had to remove the character before loading to Hive table.

B. Task 2 (MPI)

1. What is your data processing pipeline? (Graphs and words description)

The following is the illustration of data processing pipelining with table dependencies.

2. What kind of analytics do you apply on the dataset? What are the Hive queries?

We analyzed the dataset based on the following questions. For Hive queries, refer to the attachments (/src/netflix folder).

– Movies from which period do people watch/love/hate the most? (50's, 60's 70's, 80's, 90's, 2000's?):
/src/netflix/most_watched_periods.sql, /src/netflix/most_popular_periods.sql

3. Which visualization do you use on the dataset using Tableau?

[Yifei] For visualization by Tableau, I used horizontal bars, pie chart, packed bubbles to make the plots.

4. What are the programming lessons? And what are the good resources you found?

For large tables, we had to optimize the HiveQL queries to reduce the running time. For example, when joining two large tables, we needed to reduce the rows of the first table and then iterate over the second table. Also, we rearranged the query execution order by putting a WHERE clause inside of FROM clause from outside of the FROM clause.

5. What is the runtime experience for queries and visualizations?

- Inserting the netflix movie titles, movie ratings, and most queries data took no longer than 2 minutes.
- Nested query took 3 min which is longer than non-nested query.
- join query on triple tables to get the list of similar movie pairs took the longest time.

6. What difficulties you faced and what you learned from this project?

- join on large dataset like movie_ratings (100M x 100M records) is challenging because of runtime and intermediate disk space. We needed to come up with a smart way to deal with the issue. When we ran the query on the original datasets, we got the checksum error or were unable to rename the output file due to lack of disk space.