W205 summer 2017 section 2 – Lab03 - K Iwasaki

1. ***List the execution time of the weblog aggregation query for Hive, SparkSQL, and SparkSQL on Parquet.***

Execution times are 10.150 seconds, 10.037 seconds, and 3.055 seconds.

2. ***How many jobs does Hive launch? Does SparkSQL launch jobs?***

Hive had 2 jobs for the query. It seems that SparkSQL didn't launch jobs for the query.

3. ***Write a query which joins weblogs_parquet to user_info and counts the top 5 locations. List the locations.***

```
spark-sql>
         > select location, count(location) as location_count
         > from user_info2
         > join weblogs_parquet
         > on weblogs_parquet.user_id = user_info2.user_id
         > group by location
         > order by location_count desc
         > limit 5;
La Fayette      49
Leeds   47
Blountsville    46
Hayden  45
Hamilton        45
```

Above screenshot is the query and the results. Since "Top 5 locations" is clearly not defined, this question can be a tricky. I dig deeper a little by looking at a specific location and data associated with it. Below is a part of 49 user_id associated with La Fayette. It turns out there are lots of duplicated IDs. For example, there are five D22BD and two D1CA3.

```
spark-sql> select location, user_info2.user_id
         > from user_info2
         > join weblogs_parquet
         > on weblogs_parquet.user_id = user_info2.user_id
         > where location = 'La Fayette';
La Fayette      __RequestVerificationToken_Lw__=D22BD
La Fayette      __RequestVerificationToken_Lw__=D22BD
La Fayette      __RequestVerificationToken_Lw__=D22BD
La Fayette      __RequestVerificationToken_Lw__=D22BD
La Fayette      __RequestVerificationToken_Lw__=D22BD
La Fayette      __RequestVerificationToken_Lw__=D1CA3
La Fayette      __RequestVerificationToken_Lw__=D1CA3
La Fayette      __RequestVerificationToken_Lw__=CAB2C
La Fayette      __RequestVerificationToken_Lw__=CAB2C
La Fayette      __RequestVerificationToken_Lw__=CAB2C
```

I wanted to count locations without duplicated IDs so I created another query which is simply excluding join clause from the original query. Below is the query and the results, showing the different set of locations.

```
spark-sql>
         > select location, count(location) as location_count
         > from user_info2
         > group by location
         > order by location_count desc
         > limit 5;
Hamilton        19
Axis    18
Hazel Green     17
La Fayette      17
Headland        17
Time taken: 1.814 seconds, Fetched 5 row(s)
```