W205 summer 2017 section 2 – Lab03 - K Iwasaki

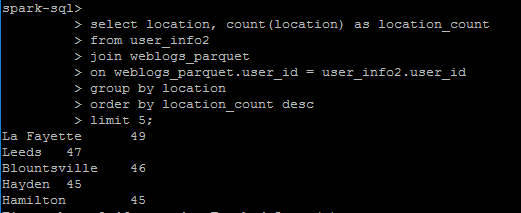
1. ***List the execution time of the weblog aggregation query for Hive, SparkSQL, and SparkSQL on Parquet.***

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

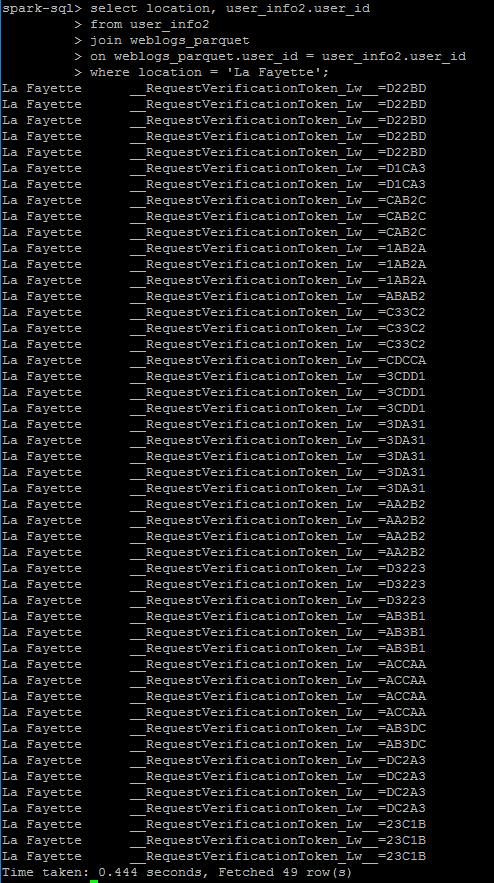
1. ***How many jobs does Hive launch? Does SparkSQL launch jobs?***

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

1. ***Write a query which joins weblogs\_parquet to user\_info and counts the top 5 locations. List the locations.***



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.



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.

