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

Hive

Time taken: 53.613 seconds, Fetched: 50 row(s)

SparkSQL

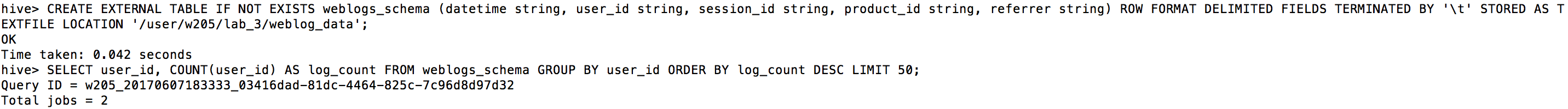
Time taken: 10.892 seconds, Fetched 50 row(s)

Parquet

Time taken: 3.224 seconds, Fetched 50 row(s)

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

Hive: 2 Jobs. SparkSQL: no job.



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

SELECT location, COUNT(location) AS loc\_count

FROM weblogs\_parquet

INNER JOIN user\_info

ON weblogs\_parquet.user\_id=user\_info.user\_id

GROUP BY location

ORDER BY loc\_count DESC

LIMIT 5;

