1. List the execution time of the weblog aggregation query for Hive, SparkSQL, and SparkSQL on Parquet.
   1. Hive: 115.82 seconds
   2. SparkSQL: 27.623 seconds
   3. SparkSQL on Parquet: 7.968 seconds
2. How many jobs does Hive launch? Does SparkSQL launch jobs?
   1. Hive – 2
   2. SparkSQL - No
3. Write a query which joins weblogs\_parquet to user\_info and counts the top 5 locations. List the locations.
   1. Locations
      1. La Fayette
      2. Leeds
      3. Blountsville
      4. Hamilton
      5. Hayden
   2. Code

|  |
| --- |
| * + 1. 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 10; |
| * + 1. ***Note:*** I printed the top ten rather than just the top five to reassure myself that my JOIN was working correctly. |