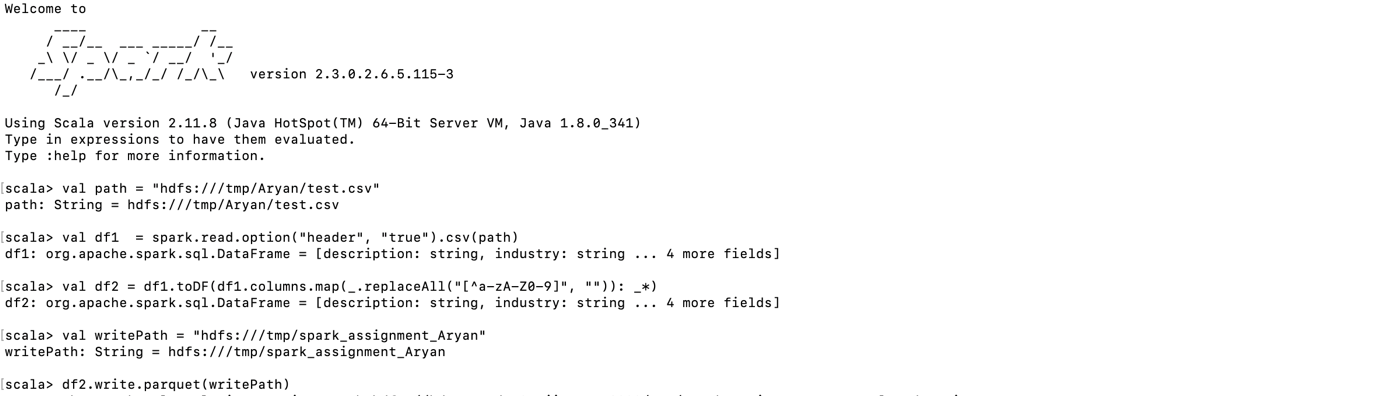
Spark Assignment 4

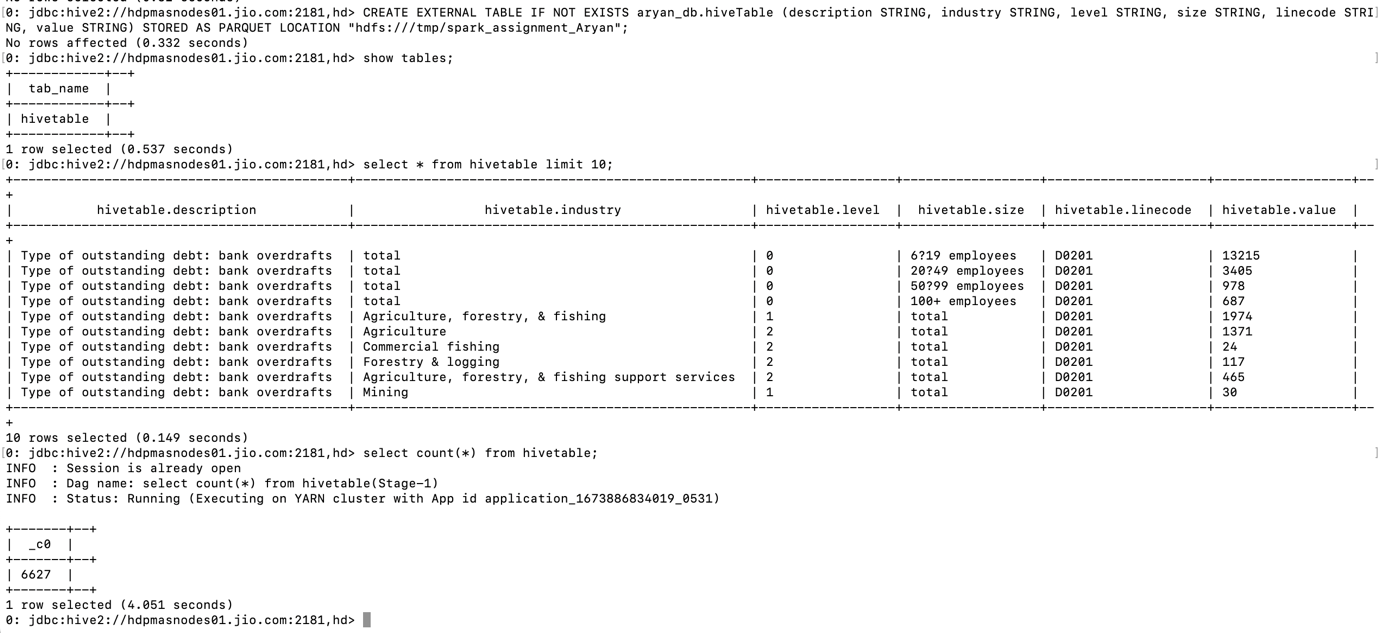
1-9:

Spark shell:





beeline:

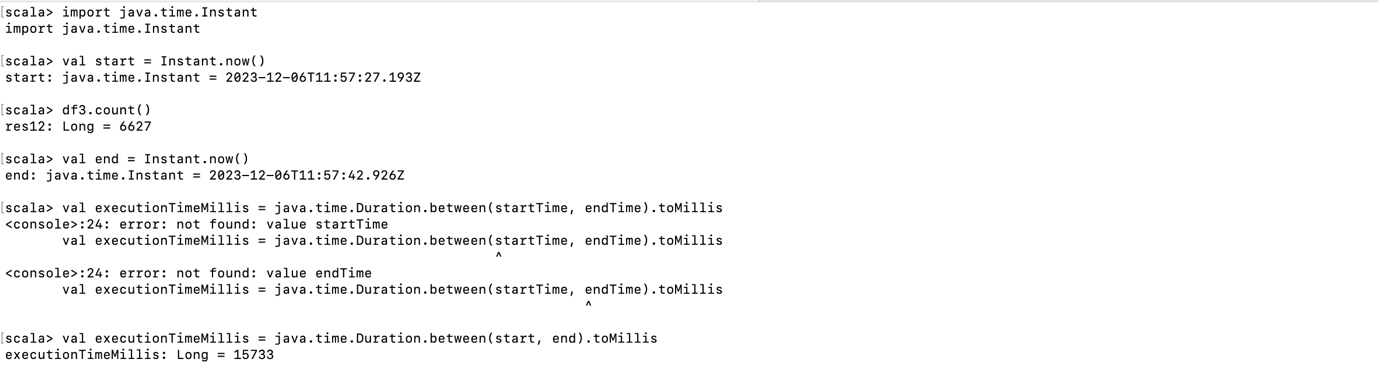


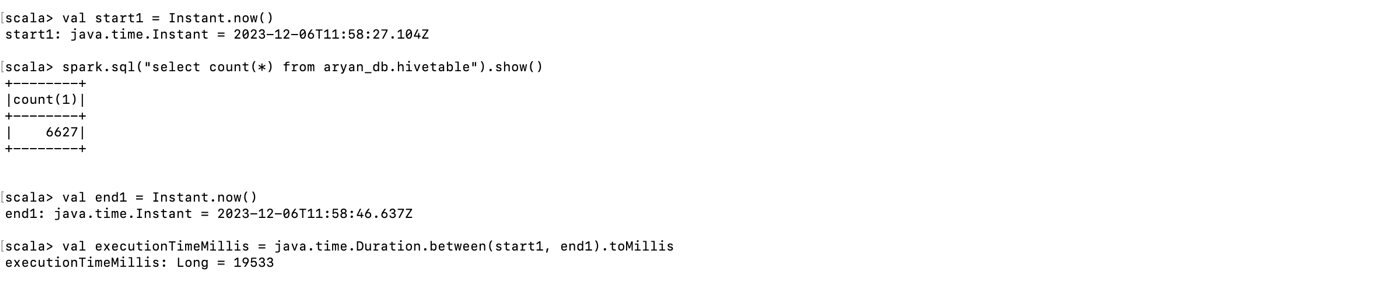
Spark shell:



We can see that all the counts are same. In Dataframe api, Spark SQL api and Hive.

Comparing runtimes:





Now comparing the execution strategies, we can see that Dataframe api is faster than Spark SQL, since in Dataframes the count is calculated programmatically whereas in Spark SQL it is queried which adds the overhead for parsing the SQL query. And the hive runs on Tez execution engine and hence will be slower than both Dataframe and spark SQL since these utilize spark optimizer in memory executions which are much faster than Map reduce.

Aryan.Sahoo