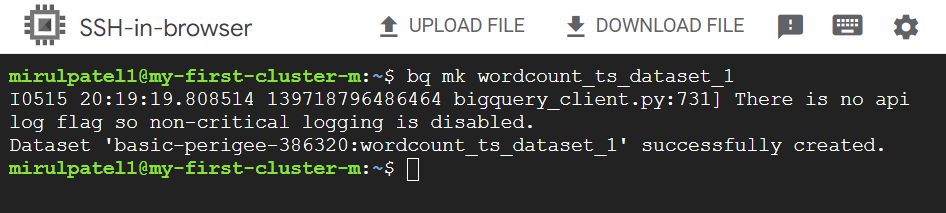
**Author: Mirul Patel**

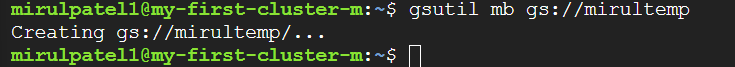
1. Created on Hadoop-spark managed cluster on compute engine in GCP Dataproc and it has started a VM instance. And accessed that instance through SSH in browser.

2. Created a dataset in Big Query by using this command **(bq mk wordcount\_ts\_dataset\_1)**



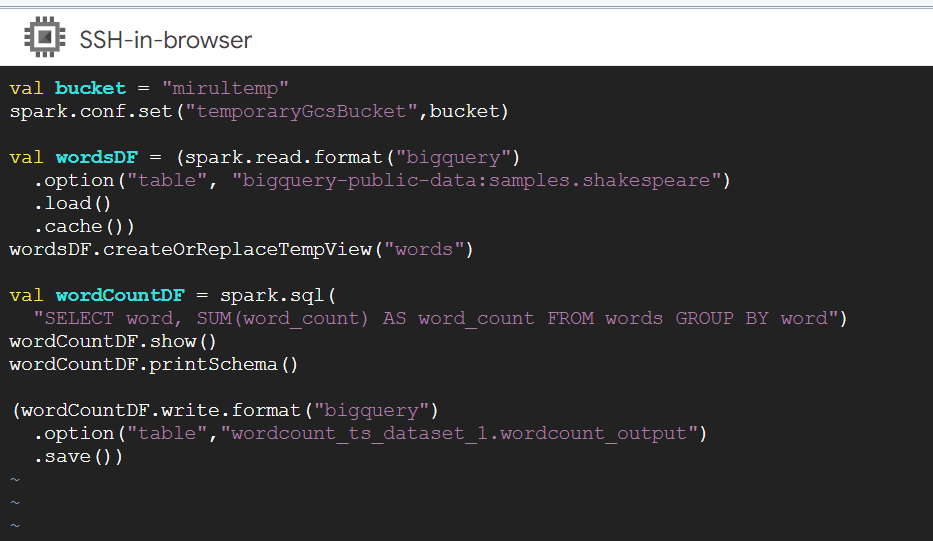
3. Have used gsutil command to create bucket to store the temp output. Used by Dataproc temporarily. Command **(gsutil mb gs://)**

Because Dataproc store the output first in Google storage/bucket before storing the O/P into Big Query.

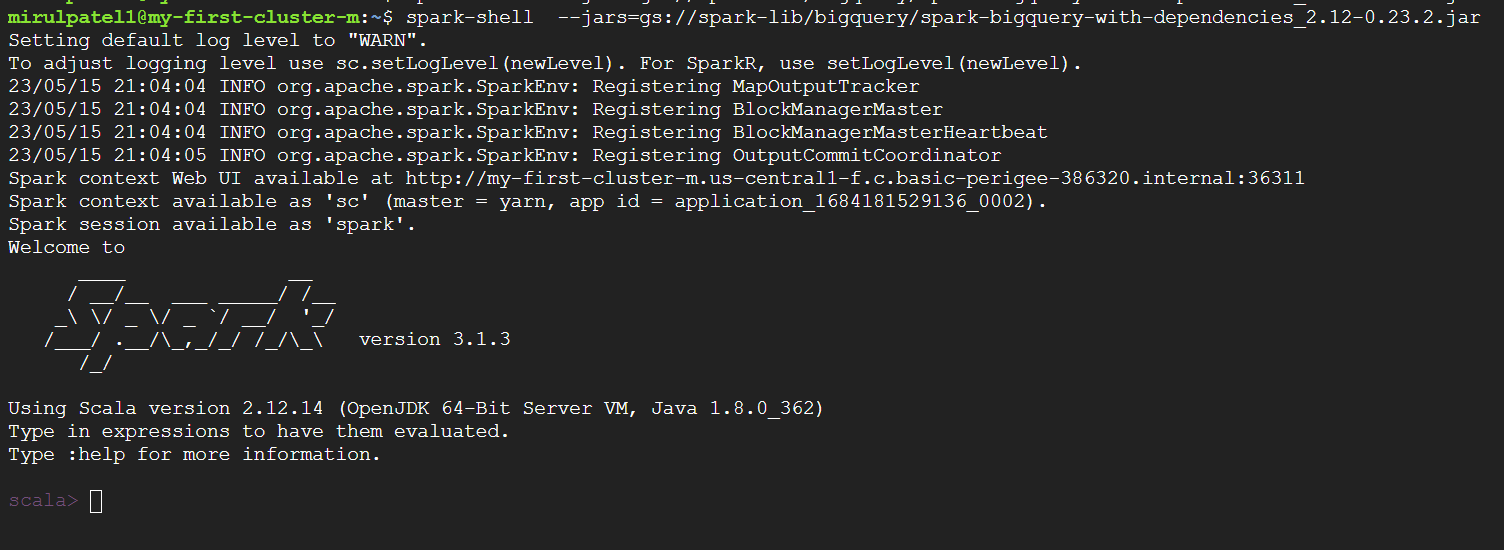


4. written a code to read data from Big Query’s public dataset (samples) consist a table called Shakespeare.

Created a temp view from that data frame and counted words by using Spark SQL query and then to written the output of that query to Big Query.



5. run a command to open spark-shell with this connector jar **(spark-shell --jars=gs://spark-lib/bigquery/spark-bigquery-with-dependencies\_2.12-0.23.2.jar)** as Dataproc job has to communicate with Big Query and Google Storage to run/perform this task.



6.Loaded the wordcount.scala file and run that file to perform a task/job.

