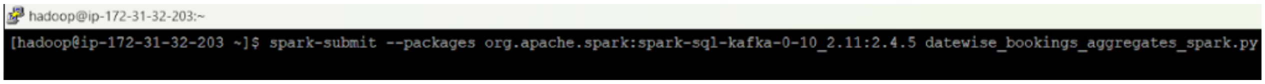
Load data from AWS RDS to Hadoop

Command to run the python file

Use WinScp to transfer the python file to the cluster.

Run the spark submit command:

>> spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.5 datewise_bookings_aggregates_spark.py

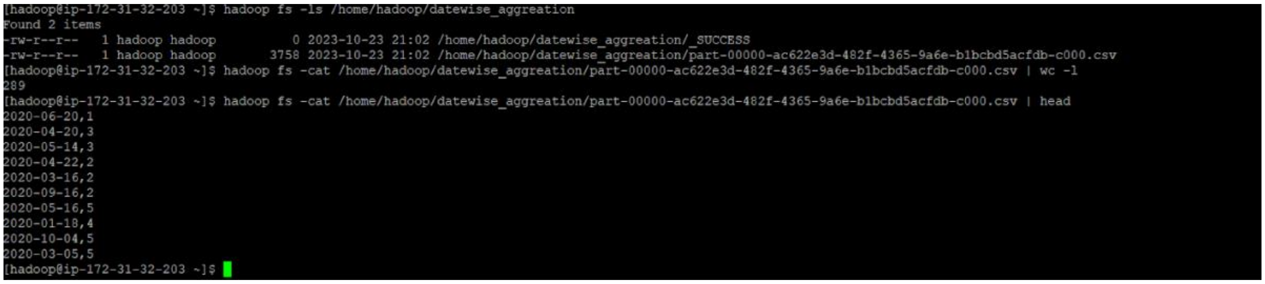


```
[hadoop@ip-172-31-32-203 ~]$ spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.5 datewise_bookings_aggregates_spark.py
```

Command to move the csv file to HDFS

The data is already loaded in Hdfs by the above pyhon file.

Screenshot of the file in HDFS



```
[hadoop@ip-172-31-32-203 ~]$ hadoop fs -ls /home/hadoop/datewise_aggreation
Found 2 items
-rw-r--r--   1 hadoop hadoop          0 2023-10-23 21:02 /home/hadoop/datewise_aggreation/_SUCCESS
-rw-r--r--   1 hadoop hadoop       3758 2023-10-23 21:02 /home/hadoop/datewise_aggreation/part-00000-ac622e3d-482f-4365-9a6e-b1bcbd5acfdb-c000.csv
[hadoop@ip-172-31-32-203 ~]$ hadoop fs -cat /home/hadoop/datewise_aggreation/part-00000-ac622e3d-482f-4365-9a6e-b1bcbd5acfdb-c000.csv | wc -l
289
[hadoop@ip-172-31-32-203 ~]$ hadoop fs -cat /home/hadoop/datewise_aggreation/part-00000-ac622e3d-482f-4365-9a6e-b1bcbd5acfdb-c000.csv | head
2020-06-20,1
2020-04-20,3
2020-05-14,3
2020-04-22,2
2020-03-16,2
2020-09-16,2
2020-05-16,5
2020-01-18,4
2020-10-04,5
2020-03-05,5
[hadoop@ip-172-31-32-203 ~]$ █
```