Load data from AWS RDS to Hadoop

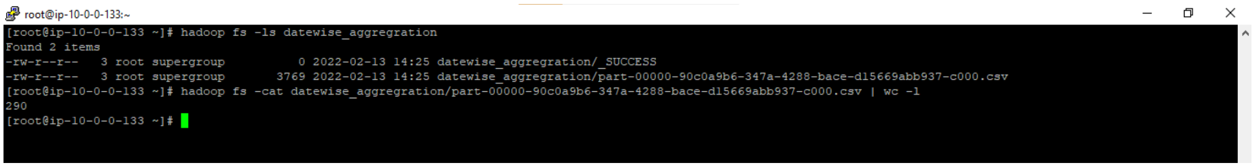
<Command to run the python file>

1. Connect to the ec2 instance
2. Switch the user from ec2 to root using. (**sudo –i**)
3. Create a python file with code that can ingest real-time clickstream data from Kafka Server and save it to the local directory. **(vi datewise_bookings_aggregates_spark.py)**
4. Submit the spark job using the python file with spark jar file **(spark2-submit --jars spark-sql-kafka-0-10_2.11-2.3.0.jar datewise_bookings_aggregates_spark.py**)

<Command to move the csv file to HDFS>

We’ve specified the path **(/user/root/datewise_aggregration)** for direct saving of the formatted csv and easy retrieval process.

<Screenshot of the file in HDFS>



```
[root@ip-10-0-0-133 ~]# hadoop fs -ls datewise_aggregration
Found 2 items
-rw-r--r--   3 root supergroup          0 2022-02-13 14:25 datewise_aggregration/_SUCCESS
-rw-r--r--   3 root supergroup       3769 2022-02-13 14:25 datewise_aggregration/part-00000-90c0a9b6-347a-4288-bace-d15669abb937-c000.csv
[root@ip-10-0-0-133 ~]# hadoop fs -cat datewise_aggregration/part-00000-90c0a9b6-347a-4288-bace-d15669abb937-c000.csv | wc -l
290
[root@ip-10-0-0-133 ~]# 
```