Load data from AWS RDS to Hadoop

1.Command to run the python file

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

2. Steps to create the datewise bookings aggregate

a. Import required libraries

```
from pyspark.sql import SparkSession
from pyspark.sql.functions import *
```

b. Create spark session

```
spark=SparkSession.builder.appName("datewise_bookings_aggregates_spark").master("local")
.getOrCreate()
```

c.Read data from HDFS

```
df=spark.read.csv("/user/root/cab_rides/part-m-00000")
```

d. Rename the columns

```
new_col =
["booking_id","customer_id","driver_id","customer_app_version","customer_phone_os_version",
"pickup_lat","pickup_lon","drop_lat",

"drop_lon","pickup_timestamp","drop_timestamp","trip_fare","tip_amount","currency_code","cab
_color","cab_registration_no","customer_rating_by_driver",
        "rating_by_customer","passenger_count"]


new_df = df.toDF(*new_col)
```

e. Convert pickup_timestamp to date by extracting date from pickup_timestamp for aggregation

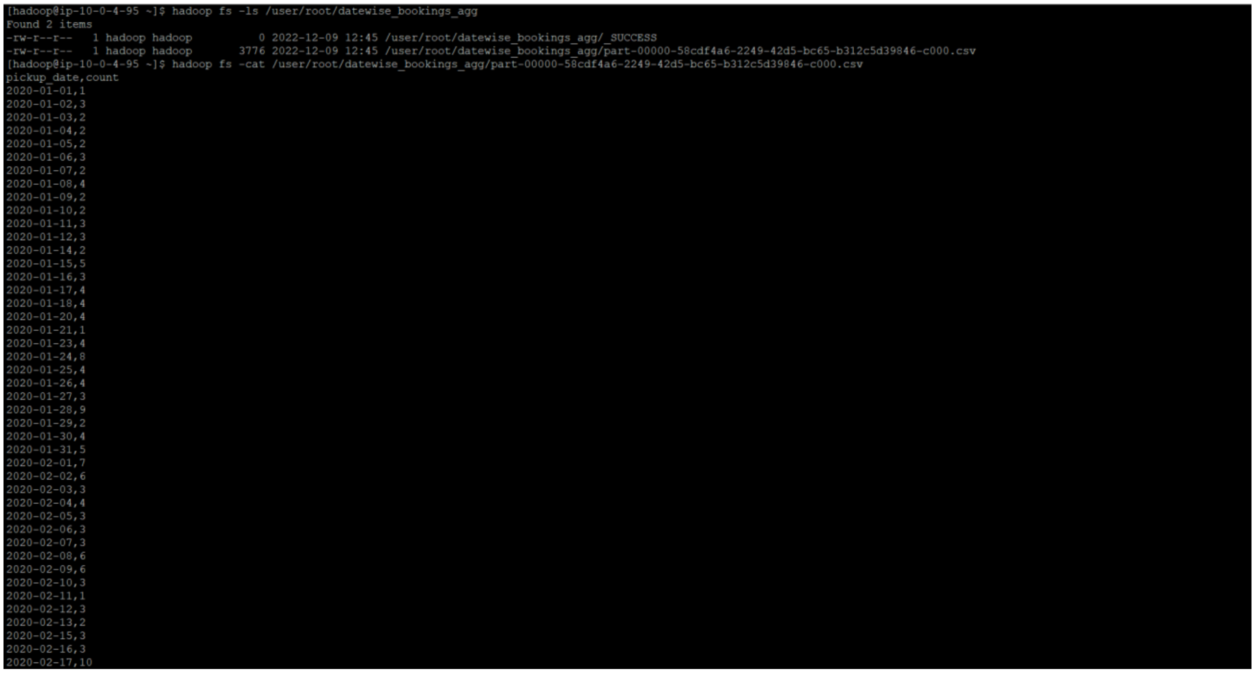
```
new_df=new_df.select("booking_id","customer_id","driver_id","customer_app_version","custom
er_phone_os_version","pickup_lat","pickup_lon","drop_lat",
```

"drop_lon",to_date(col('pickup_timestamp')).alias('pickup_date').cast("date"),"drop_timestamp","trip_fare","tip_amount","currency_code","cab_color","cab_registration_no","customer_rating_by_driver",
"rating_by_customer","passenger_count")

f. Aggregate data on pickup_date
agg_df=new_df.groupBy("pickup_date").count().orderBy("pickup_date")

3.Command to move the csv file to HDFS
agg_df.coalesce(1).write.format('csv').mode('overwrite').save('/user/root/datewise_bookings_agg',header='true')

Screenshot of the file in HDFS



```
[hadoop@ip-10-0-4-95 ~]$ hadoop fs -ls /user/root/datewise_bookings_agg
Found 2 items
-rw-r--r--   1 hadoop hadoop          0 2022-12-09 12:45 /user/root/datewise_bookings_agg/_SUCCESS
-rw-r--r--   1 hadoop hadoop       3776 2022-12-09 12:45 /user/root/datewise_bookings_agg/part-00000-58cdf4a6-2249-42d5-bc65-b312c5d39846-c000.csv
[hadoop@ip-10-0-4-95 ~]$ hadoop fs -cat /user/root/datewise_bookings_agg/part-00000-58cdf4a6-2249-42d5-bc65-b312c5d39846-c000.csv
pickup_date,count
2020-01-01,1
2020-01-02,3
2020-01-03,2
2020-01-04,2
2020-01-05,2
2020-01-06,3
2020-01-07,2
2020-01-08,4
2020-01-09,2
2020-01-10,2
2020-01-11,3
2020-01-12,3
2020-01-14,2
2020-01-15,5
2020-01-16,3
2020-01-17,4
2020-01-18,4
2020-01-20,4
2020-01-21,1
2020-01-23,4
2020-01-24,8
2020-01-25,4
2020-01-26,4
2020-01-27,3
2020-01-28,9
2020-01-29,2
2020-01-30,4
2020-01-31,5
2020-02-01,7
2020-02-02,6
2020-02-03,3
2020-02-04,4
2020-02-05,3
2020-02-06,3
2020-02-07,3
2020-02-08,6
2020-02-09,6
2020-02-10,3
2020-02-11,1
2020-02-12,3
2020-02-13,2
2020-02-15,3
2020-02-16,3
2020-02-17,10
```