Load data from AWS RDS to Hadoop

Steps to calculate date wise bookings aggregate data

1. **Establishing spark connection:**

```
spark = SparkSession \
        .builder \
        .appName('aggregateBatchData') \
        .master('yarn') \
        .getOrCreate()
```

2. **To read data from csv file extracted from AWS RDS and stored in HDFS:**

```
df=spark.read.csv('/user/hadoop/bookings-
data/',header=False,inferSchema = True)
```

3. **To add column headers according to given data:**

```
new_columns=["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_cus
tomer","passenger_count"]

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

4. **To create a new column with date extracted from pickup_timestamp column:**

```
new_df = new_df.withColumn("date", F.to_date(F.col("pickup_timestamp")))
```

5. **To get the datewise bookings aggregate:**

```
aggregate_df = new_df.groupby('date').count()
```

© Copyright 2020. upGrad Education Pvt. Ltd. All rights reserved

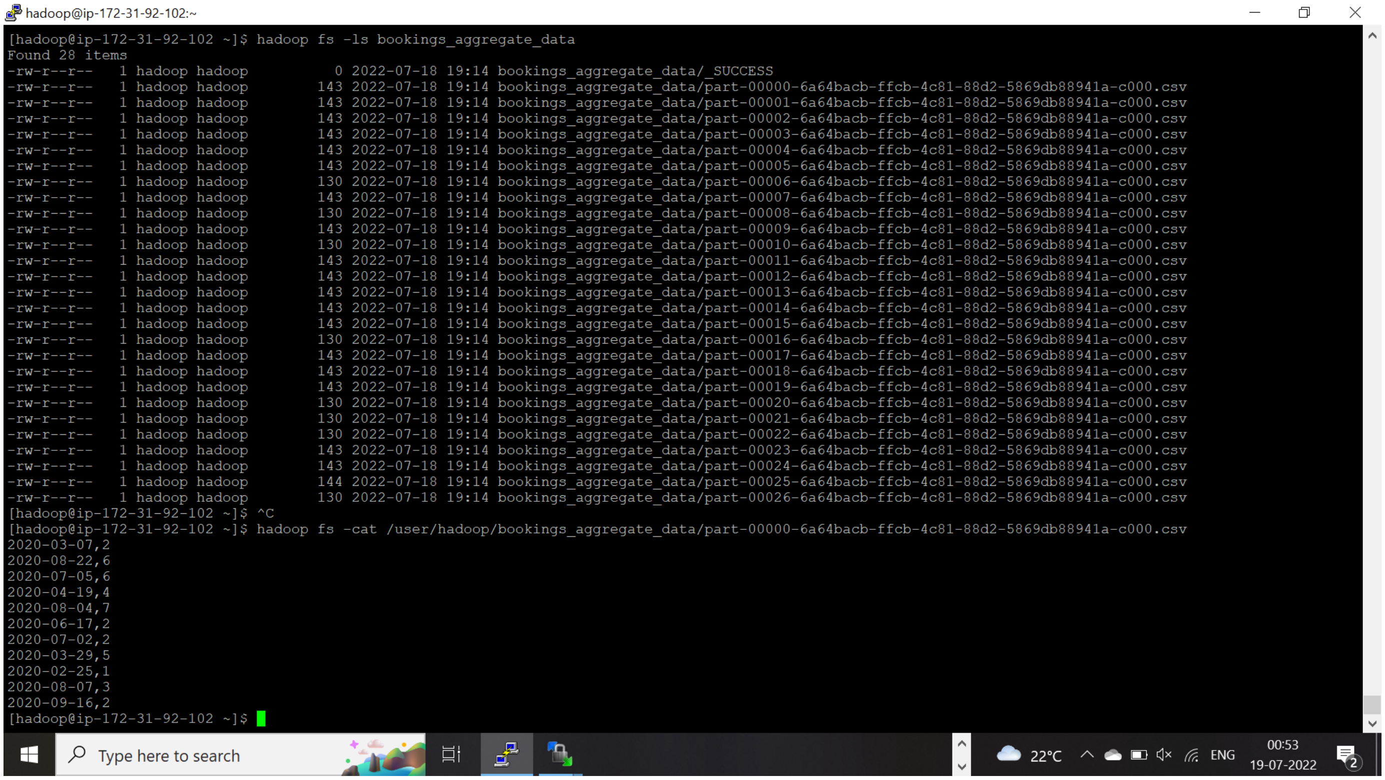
6. **To write the resultant dataframe in csv files in HDFS:**

```
aggregate_df.write.csv('/user/hadoop/bookings_aggregate_data/')
```

7. **Command to run datewise_bookings_aggregates_spark.py file:**

```
spark-submit datewise_bookings_aggregates_spark.py
```

8. **Screenshot of the csv files and data:**



```
[hadoop@ip-172-31-92-102 ~]$ hadoop fs -ls bookings_aggregate_data
Found 28 items
-rw-r--r--   1 hadoop hadoop          0 2022-07-18 19:14 bookings_aggregate_data/_SUCCESS
-rw-r--r--   1 hadoop hadoop        143 2022-07-18 19:14 bookings_aggregate_data/part-00000-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
-rw-r--r--   1 hadoop hadoop        143 2022-07-18 19:14 bookings_aggregate_data/part-00001-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
-rw-r--r--   1 hadoop hadoop        143 2022-07-18 19:14 bookings_aggregate_data/part-00002-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
-rw-r--r--   1 hadoop hadoop        143 2022-07-18 19:14 bookings_aggregate_data/part-00003-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
-rw-r--r--   1 hadoop hadoop        143 2022-07-18 19:14 bookings_aggregate_data/part-00004-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
-rw-r--r--   1 hadoop hadoop        143 2022-07-18 19:14 bookings_aggregate_data/part-00005-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
-rw-r--r--   1 hadoop hadoop        130 2022-07-18 19:14 bookings_aggregate_data/part-00006-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
-rw-r--r--   1 hadoop hadoop        143 2022-07-18 19:14 bookings_aggregate_data/part-00007-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
-rw-r--r--   1 hadoop hadoop        130 2022-07-18 19:14 bookings_aggregate_data/part-00008-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
-rw-r--r--   1 hadoop hadoop        143 2022-07-18 19:14 bookings_aggregate_data/part-00009-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
-rw-r--r--   1 hadoop hadoop        130 2022-07-18 19:14 bookings_aggregate_data/part-00010-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
-rw-r--r--   1 hadoop hadoop        143 2022-07-18 19:14 bookings_aggregate_data/part-00011-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
-rw-r--r--   1 hadoop hadoop        143 2022-07-18 19:14 bookings_aggregate_data/part-00012-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
-rw-r--r--   1 hadoop hadoop        143 2022-07-18 19:14 bookings_aggregate_data/part-00013-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
-rw-r--r--   1 hadoop hadoop        143 2022-07-18 19:14 bookings_aggregate_data/part-00014-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
-rw-r--r--   1 hadoop hadoop        143 2022-07-18 19:14 bookings_aggregate_data/part-00015-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
-rw-r--r--   1 hadoop hadoop        130 2022-07-18 19:14 bookings_aggregate_data/part-00016-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
-rw-r--r--   1 hadoop hadoop        143 2022-07-18 19:14 bookings_aggregate_data/part-00017-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
-rw-r--r--   1 hadoop hadoop        143 2022-07-18 19:14 bookings_aggregate_data/part-00018-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
-rw-r--r--   1 hadoop hadoop        143 2022-07-18 19:14 bookings_aggregate_data/part-00019-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
-rw-r--r--   1 hadoop hadoop        130 2022-07-18 19:14 bookings_aggregate_data/part-00020-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
-rw-r--r--   1 hadoop hadoop        130 2022-07-18 19:14 bookings_aggregate_data/part-00021-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
-rw-r--r--   1 hadoop hadoop        130 2022-07-18 19:14 bookings_aggregate_data/part-00022-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
-rw-r--r--   1 hadoop hadoop        143 2022-07-18 19:14 bookings_aggregate_data/part-00023-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
-rw-r--r--   1 hadoop hadoop        143 2022-07-18 19:14 bookings_aggregate_data/part-00024-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
-rw-r--r--   1 hadoop hadoop        144 2022-07-18 19:14 bookings_aggregate_data/part-00025-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
-rw-r--r--   1 hadoop hadoop        130 2022-07-18 19:14 bookings_aggregate_data/part-00026-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
[hadoop@ip-172-31-92-102 ~]$ ^C
[hadoop@ip-172-31-92-102 ~]$ hadoop fs -cat /user/hadoop/bookings_aggregate_data/part-00000-6a64bacb-ffcb-4c81-88d2-5869db88941a-c000.csv
2020-03-07,2
2020-08-22,6
2020-07-05,6
2020-04-19,4
2020-08-04,7
2020-06-17,2
2020-07-02,2
2020-03-29,5
2020-02-25,1
2020-08-07,3
2020-09-16,2
[hadoop@ip-172-31-92-102 ~]$
```

© Copyright 2020. upGrad Education Pvt. Ltd. All rights reserved