Load data from AWS RDS to Hadoop

<Command to run the python file>

1. **Download mysql.jdbc Driver and place under /lib/sqoop/lib/**

2. **Create python file datewise_bookings_aggregates_spark.py to create aggregates for finding date-wise total bookings**

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

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

#Reading data from HDFS
df=spark.read.csv("/user/hadoop/cab_rides/part-m-00000")

df.show(10)

df.printSchema()

#The count of the dataset
df.count()

#Renaming 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)

new_df.show(truncate=False)

#Converting 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","t
rip_fare","tip_amount","currency_code","cab_color","cab_registration_no","customer_rating_by_
driver",
        "rating_by_customer","passenger_count")

new_df.show()

#Aggregation on pickup_date
agg_df=new_df.groupBy("pickup_date").count().orderBy("pickup_date")

agg_df.show()

#The count of Bookings aggregates_table
agg_df.count()

agg_df.coalesce(1).write.format('csv').mode('overwrite').save('/user/hadoop/datewise_bookings_
agg',header='true')
```

3. **Run Spark Submit Job**

```
spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.6
datewise_bookings_aggregates_spark.py
```

<Command to move the csv file to HDFS>

```
agg_df.coalesce(1).write.format('csv').mode('overwrite').save('/user/hadoop/datewise_bookings_
agg',header='true')
```

<Screenshot of the file in HDFS>

```
[hadoop@ip-10-0-0-52 ~]$ hadoop fs -ls datewise_bookings_agg
Found 2 items
-rw-r--r--   1 hadoop hdfsadmingroup          0 2022-07-24 20:58 datewise_bookings_agg/_SUCCESS
-rw-r--r--   1 hadoop hdfsadmingroup       3776 2022-07-24 20:58 datewise_bookings_agg/part-00000-8bff7fc8-bf32-4cbd-bc83-5b82c09640cc-c000.csv
[hadoop@ip-10-0-0-52 ~]$
```

Copy file to local

hadoop fs -get /user/hadoop/datewise_bookings_agg/part-00000-8bff7fc8-bf32-4cbd-bc83-5b82c09640cc-c000.csv /home/hadoop/