

# Load data from AWS RDS to Hadoop

## <Command to run the python file>

1. Create a python file to consume data from kafka

```
vi datewise_bookings_aggregates_spark.py
```

2. Run spark submit command

```
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>

1. Make a directory using mkdir command

```
hadoop fs -mkdir datewise_aggregated_data
```

2. Loading the data from local file system to hadoop file system

```
hadoop fs-put ~/datewise_aggregated_data datewise_aggregated_data
```

3. Checking the data file in hadoop

```
hadoop fs -ls datewise_aggregated_data
```

```
hadoop fs -cat datewise_aggregated_data / part-00000-a085f9bc-ecd9-4ee7-b21b-  
3329d4fabfd3-c000.csv | wc -l
```

## <Screenshot of the file in HDFS>

```
[hadoop@ip-172-31-8-156 ~]$ hadoop fs -ls  
Found 7 items  
drwxr-xr-x - hadoop hadoop 0 2022-07-22 10:23 .sparkStaging  
drwxr-xr-x - hadoop hadoop 0 2022-07-22 10:23 booking_data_csv  
drwxr-xr-x - hadoop hadoop 0 2022-07-22 10:11 bookings_data  
drwxr-xr-x - hadoop hadoop 0 2022-07-22 07:53 clickstream_checkpoint  
drwxr-xr-x - hadoop hadoop 0 2022-07-22 07:53 clickstream_data  
drwxr-xr-x - hadoop hadoop 0 2022-07-22 09:02 clickstream_data_flatten  
drwxr-xr-x - hadoop hadoop 0 2022-07-22 10:23 datewise_aggregated_data  
[hadoop@ip-172-31-8-156 ~]$ hadoop fs -ls datewise_aggregated_data  
Found 2 items  
-rw-r--r-- 1 hadoop hadoop 0 2022-07-22 10:23 datewise_aggregated_data/_SUCCESS  
-rw-r--r-- 1 hadoop hadoop 3758 2022-07-22 10:23 datewise_aggregated_data/part-00000-a085f9bc-ecd9-4ee7-b21b-3329d4fabfd3-c000.csv  
[hadoop@ip-172-31-8-156 ~]$ hadoop fs -cat datewise_aggregated_data/part-00000-a085f9bc-ecd9-4ee7-b21b-3329d4fabfd3-c000.csv | wc -l  
289  
[hadoop@ip-172-31-8-156 ~]$
```

### <Screenshot of the booking\_data in csv format in HDFS>

```
[hadoop@ip-172-31-8-156 ~]$ hadoop fs -ls booking_data_csv
Found 2 items
-rw-r--r-- 1 hadoop hadoop 0 2022-07-22 10:23 booking_data_csv/_SUCCESS
-rw-r--r-- 1 hadoop hadoop 182968 2022-07-22 10:23 booking_data_csv/part-00000-ee83ab78-57ec-4982-9d50-e2ea2e04d9bf-c000.csv
[hadoop@ip-172-31-8-156 ~]$ hadoop fs -cat booking_data_csv/part-00000-ee83ab78-57ec-4982-9d50-e2ea2e04d9bf-c000.csv | wc -l
1001
[hadoop@ip-172-31-8-156 ~]$
```