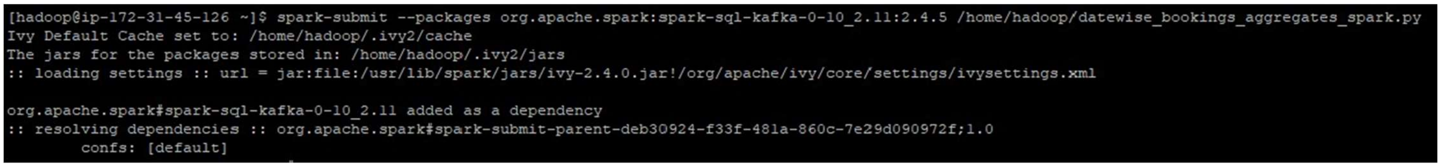
Loading Aggregated Bookings data into Hadoop

1. **Pyspark file “datewise_bookings_aggregates_spark.py” created to aggregate total number of bookings by pickup date.**

2. **Executing this file to save aggregated file as .csv format into HDFS location**

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



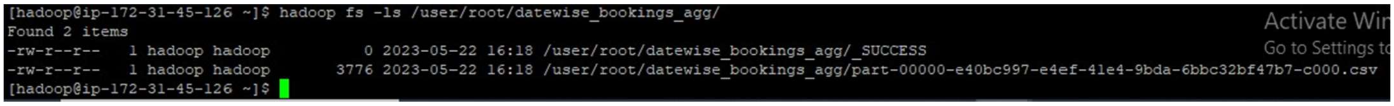
```
[hadoop@ip-172-31-45-126 ~]$ spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.5 /home/hadoop/datewise_bookings_aggregates_spark.py
Ivy Default Cache set to: /home/hadoop/.ivy2/cache
The jars for the packages stored in: /home/hadoop/.ivy2/jars
:: loading settings :: url = jar:file:/usr/lib/spark/jars/ivy-2.4.0.jar!/org/apache/ivy/core/settings/ivysettings.xml

org.apache.spark#spark-sql-kafka-0-10_2.11 added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-parent-deb30924-f33f-481a-860c-7e29d090972f;1.0
        confs: [default]
```

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')

4. **Screenshot of the .csv file created in HDFS**

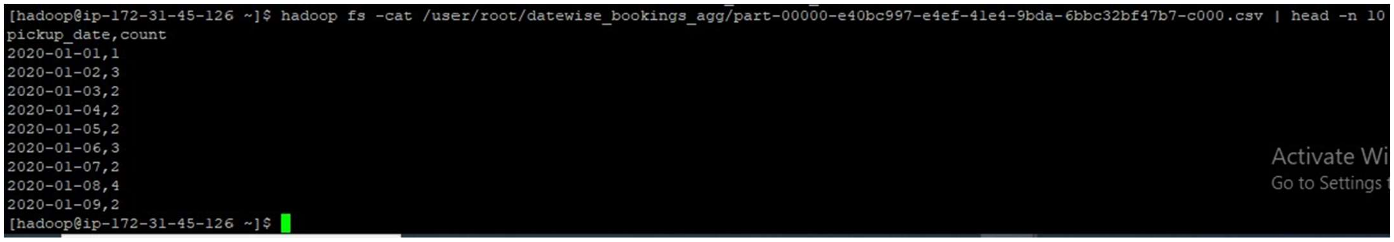
   hadoop fs -ls /user/root/datewise_bookings_agg/



```
[hadoop@ip-172-31-45-126 ~]$ hadoop fs -ls /user/root/datewise_bookings_agg/
Found 2 items
-rw-r--r--   1 hadoop hadoop          0 2023-05-22 16:18 /user/root/datewise_bookings_agg/_SUCCESS
-rw-r--r--   1 hadoop hadoop       3776 2023-05-22 16:18 /user/root/datewise_bookings_agg/part-00000-e40bc997-e4ef-41e4-9bda-6bbc32bf47b7-c000.csv
[hadoop@ip-172-31-45-126 ~]$
```

5. **Command and Screenshot of the aggregated file output in HDFS**

   hadoop fs -cat /user/root/datewise_bookings_agg/part-00000-e40bc997-e4ef-41e4-9bda-6bbc32bf47b7-c000.csv | head -n 10



```
[hadoop@ip-172-31-45-126 ~]$ hadoop fs -cat /user/root/datewise_bookings_agg/part-00000-e40bc997-e4ef-41e4-9bda-6bbc32bf47b7-c000.csv | head -n 10
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
[hadoop@ip-172-31-45-126 ~]$
```