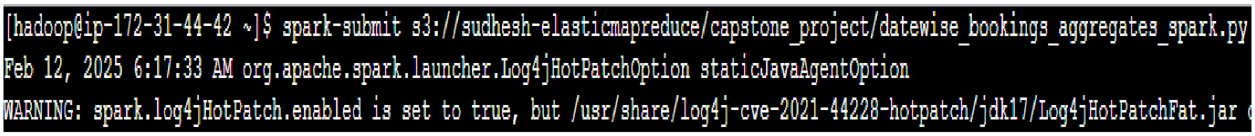
Load Aggregated Bookings data to HDFS

1. **Pyspark file “datewise_bookings_aggregates_spark.py number of bookings by pickup date. Python file is stored in s3 bucket.**

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

   spark-submit s3://sudhesh-elasticmapreduce/capstone_project/datewise_bookings_aggregates_spark.py

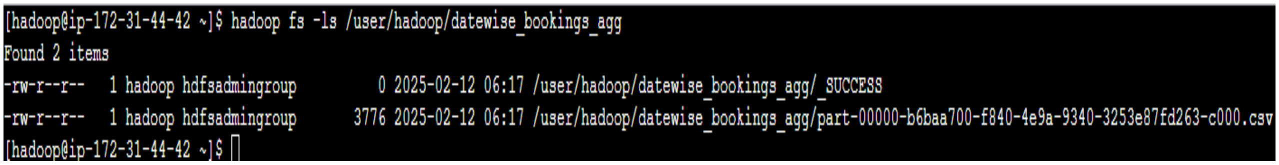


```
[hadoop@ip-172-31-44-42 ~]$ spark-submit s3://sudhesh-elasticmapreduce/capstone_project/datewise_bookings_aggregates_spark.py
Feb 12, 2025 6:17:33 AM org.apache.spark.launcher.Log4jHotPatchOption staticJavaAgentOption
WARNING: spark.log4jHotPatch.enabled is set to true, but /usr/share/log4j-cve-2021-44228-hotpatch/jdk17/Log4jHotPatchFat.jar
```

3. **Command to move aggregated csv file to HDFS.**

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

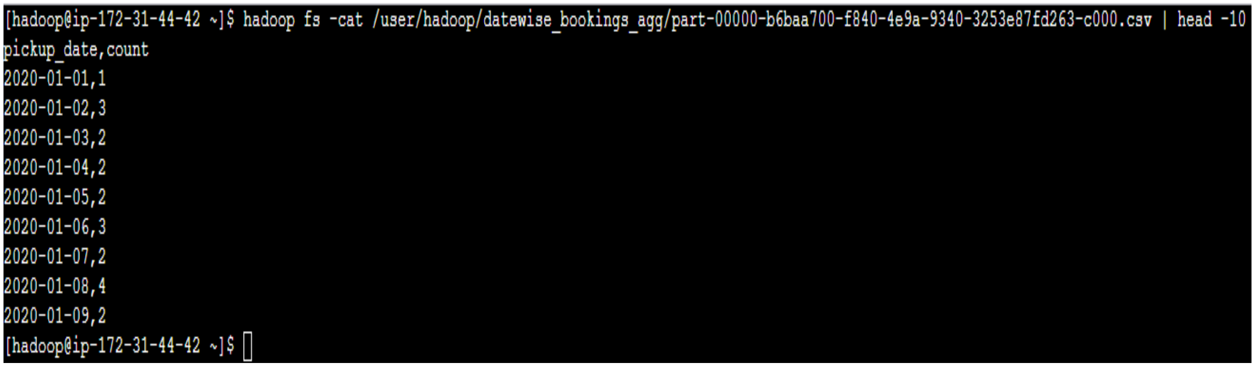
4. **Screenshot of the csv file in HDFS.**



```
[hadoop@ip-172-31-44-42 ~]$ hadoop fs -ls /user/hadoop/datewise_bookings_agg
Found 2 items
-rw-r--r--   1 hadoop hdfsadmingroup          0 2025-02-12 06:17 /user/hadoop/datewise_bookings_agg/_SUCCESS
-rw-r--r--   1 hadoop hdfsadmingroup       3776 2025-02-12 06:17 /user/hadoop/datewise_bookings_agg/part-00000-b6baa700-f840-4e9a-9340-3253e87fd263-c000.csv
[hadoop@ip-172-31-44-42 ~]$ 
```

5. **Screenshot of the aggregated csv file in HDFS.**

   hadoop fs -cat /user/hadoop/datewise_bookings_agg/part-00000-b6baa700-f840-4e9a-9340-3253e87fd263-c000.csv | head -10



```
[hadoop@ip-172-31-44-42 ~]$ hadoop fs -cat /user/hadoop/datewise_bookings_agg/part-00000-b6baa700-f840-4e9a-9340-3253e87fd263-c000.csv | head -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-172-31-44-42 ~]$ 
```