

Load aggregate bookings data to Hadoop

1. Creating aggregates for finding date-wise total bookings using the Spark script 'datewise_bookings_aggregates_spark.py'

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

```
hadoop@ip-172-31-32-102:~$
[hadoop@ip-172-31-32-102 ~]$
[hadoop@ip-172-31-32-102 ~]$ spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.5 datewise_bookings_aggregat
es_spark.py
hadoop@ip-172-31-32-102:~$
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-ac38adc0-af7d-493b-aelc-101dcd86d73a;1.0
  confs: [default]
  found org.apache.spark#spark-sql-kafka-0-10_2.11;2.4.5 in central
  found org.apache.kafka#kafka-clients;2.0.0 in central
  found org.lz4#lz4-java;1.4.0 in central
  found org.xerial.snappy#snappy-java;1.1.7.3 in central
  found org.slf4j#slf4j-api;1.7.16 in central
  found org.spark-project.spark#unused;1.0.0 in central
:: resolution report :: resolve 341ms :: artifacts dl 10ms
  :: modules in use:
  org.apache.kafka#kafka-clients;2.0.0 from central in [default]
  org.apache.spark#spark-sql-kafka-0-10_2.11;2.4.5 from central in [default]
  org.lz4#lz4-java;1.4.0 from central in [default]
  org.slf4j#slf4j-api;1.7.16 from central in [default]
  org.spark-project.spark#unused;1.0.0 from central in [default]
  org.xerial.snappy#snappy-java;1.1.7.3 from central in [default]
-----
|               | modules | artifacts |
|   conf       | number | search | dwnlded | evicted | number | dwnlded |
-----
|   default    | 6      | 0      | 0      | 0      | 6      | 0      |
:: retrieving :: org.apache.spark#spark-submit-parent-ac38adc0-af7d-493b-aelc-101dcd86d73a
  confs: [default]
  0 artifacts copied, 6 already retrieved (0kB/10ms)
```

2. Command to move the csv file to HDFS

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

3. Screenshot of the files containing the aggregated bookings data in HDFS

```
hadoop fs -ls /user/root/deduped_bookings_agg/
```

```
hadoop fs -cat /user/root/datetime_bookings_agg/part-00000-18cdabab-83e6-4637-8841-  
cf5ccb3fae86-c000.csv | head -n 5
```

```
hadoop@ip-172-31-32-102:~  
[hadoop@ip-172-31-32-102 ~]$  
[hadoop@ip-172-31-32-102 ~]$ hadoop fs -ls /user/root/  
Found 5 items  
drwxr-xr-x - hadoop hadoop 0 2024-07-26 11:25 /user/root/bookings  
drwxr-xr-x - hadoop hadoop 0 2024-07-26 11:16 /user/root/clickstream  
drwxr-xr-x - hadoop hadoop 0 2024-07-26 11:00 /user/root/clickstream_dump  
drwxr-xr-x - hadoop hadoop 0 2024-07-26 10:58 /user/root/clickstream_dump_checkpoint  
drwxr-xr-x - hadoop hadoop 0 2024-07-26 11:31 /user/root/datetime_bookings_agg  
[hadoop@ip-172-31-32-102 ~]$  
[hadoop@ip-172-31-32-102 ~]$ hadoop fs -ls /user/root/datetime_bookings_agg/  
Found 2 items  
-rw-r--r-- 1 hadoop hadoop 0 2024-07-26 11:31 /user/root/datetime_bookings_agg/_SUCCESS  
-rw-r--r-- 1 hadoop hadoop 3776 2024-07-26 11:31 /user/root/datetime_bookings_agg/part-00000-18cdabab-83e6-4637-8841-cf5  
ccb3fae86-c000.csv  
[hadoop@ip-172-31-32-102 ~]$  
[hadoop@ip-172-31-32-102 ~]$ hadoop fs -cat /user/root/datetime_bookings_agg/part-00000-18cdabab-83e6-4637-8841-cf5ccb3fae86-c00  
0.csv | head -n 5  
pickup_date,count  
2020-01-01,1  
2020-01-02,3  
2020-01-03,2  
2020-01-04,2  
[hadoop@ip-172-31-32-102 ~]$
```