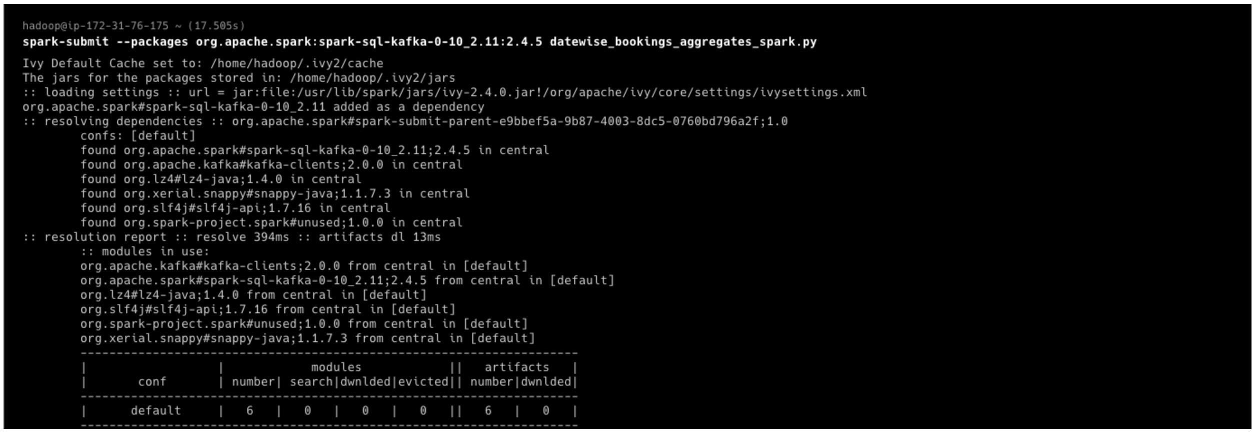
Load data from AWS RDS to Hadoop

Load Booking Data Aggregation into Hadoop

1. pyspark file "datewise_bookings_aggregates_spark.py" created to aggregate total number of booking by pickup date

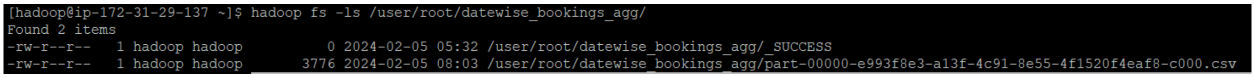
   ```
   spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.5
   datewise_bookings_aggregates_spark.py
   ```



```
hadoop@ip-172-31-76-175 ~ (17.505s)
spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.5 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-e9bbef5a-9b87-4003-8dc5-0760bd796a2f;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 394ms :: artifacts dl 13ms
        :: 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   |
        ---------------------------------------------------------------------
```

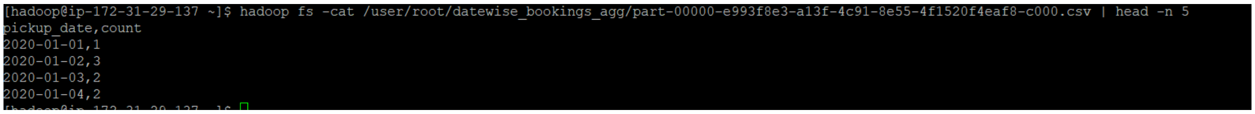
2. Verify the aggregated Data

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



```
[hadoop@ip-172-31-29-137 ~]$ hadoop fs -ls /user/root/datewise_bookings_agg/
Found 2 items
-rw-r--r--   1 hadoop hadoop          0 2024-02-05 05:32 /user/root/datewise_bookings_agg/_SUCCESS
-rw-r--r--   1 hadoop hadoop       3776 2024-02-05 08:03 /user/root/datewise_bookings_agg/part-00000-e993f8e3-a13f-4c91-8e55-4f1520f4eaf8-c000.csv
```

   ```
   hadoop fs -cat /user/root/datewise_bookings_agg/part-00000-e993f8e3-a13f-
   4c91-8e55-4f1520f4eaf8-c000.csv | head -n 5
   ```



```
[hadoop@ip-172-31-29-137 ~]$ hadoop fs -cat /user/root/datewise_bookings_agg/part-00000-e993f8e3-a13f-4c91-8e55-4f1520f4eaf8-c000.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-172-31-29-137 ~]$
```