Load data from Spark to HDFS

Commands used to clean and aggregate data

Run python file named `datewise_bookings_aggregates_spark.py` using the command below.

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

This python file will load all tables created directly to Hive database inside HDFS

To check if all the tables already exist inside Hive database, we use this command below. `hadoop fs -ls /user/hive/warehouse/`

```
[hadoop@ip-172-31-1-153 ~]$ hadoop fs -ls /user/hive/warehouse
Found 3 items
drwxrwxrwt   - hadoop hadoop          0 2022-06-15 14:09 /user/hive/warehouse/bookings
drwxrwxrwt   - hadoop hadoop          0 2022-06-15 14:09 /user/hive/warehouse/bookings_per_day
drwxrwxrwt   - hadoop hadoop          0 2022-06-15 14:09 /user/hive/warehouse/clickstream
[hadoop@ip-172-31-1-153 ~]$
```

Screenshots of command process

```
hadoop@ip-172-31-1-153:~
[hadoop@ip-172-31-1-153 ~]$ 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-096e9364-1a7f-4f62-93be-08bd9fleeffd;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 379ms :: artifacts dl 11ms
        :: 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-096e9364-1a7f-4f62-93be-08bd9fleeffd
```

Bookings table

```
[hadoop@ip-172-31-1-153 ~]$ hive
select * from bookings
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
limit 10;hive> select * from bookings limit 10;
OK
BK8968087150    51811359        15055660        2.2.14  Android -49     104     -59     146     2020-06-23 19:33:10     2020-06-06 09:02:10     534     83      INR     b
lack    054-38-4479     4       3       3
BK629851904     31663218        60872180        3.4.1   iOS     -84     176     86      128     2020-05-23 12:22:04     2020-08-09 19:02:56     126     67      INR     l
ime     796-39-6801     3       2       4
BK1797410350    86869399        94276051        4.1.36  iOS     -68     55      -51     -31     2020-05-19 14:14:32     2020-08-23 18:38:39     297     63      INR     o
live    748-73-1579     1       3       3
BK5788246325    58230837        45457227        2.4.27  Android 14      113     54      -18     2020-03-24 01:30:15     2020-05-19 11:16:45     932     32      INR     w
hite    558-80-6346     3       2       2
BK8342703255    84232510        86494681        4.1.34  Android -6      -115    23      70      2020-08-03 19:10:52     2020-03-24 08:25:40     260     7       INR     b
lue     068-72-1637     3       3       3
BK6015582453    11981042        35862658        2.4.39  iOS     -19     -70     -10     174     2020-07-17 05:33:48     2020-04-30 04:54:27     907     53      INR     p
urple   102-10-5639     3       2       3
BK4529355854    60071878        78022360        2.1.9   iOS     1       -56     35      104     2020-01-02 01:48:40     2020-02-16 04:28:55     547     17      INR     t
eal     866-83-4349     2       3       4
BK9720088219    14327312        94427067        3.1.2   Android -55     173     65      51      2020-04-10 15:11:07     2020-01-20 21:17:42     259     33      INR     m
aroon   572-73-6526     3       3       2
BK7157532607    46407210        43160003        1.3.4   Android 46      -17     8       -156    2020-06-09 05:56:31     2020-03-19 01:53:16     787     21      INR     o
live    667-23-5880     2       2       3
BK5014871433    65861573        64708618        1.3.28  iOS     -30     65      84      -50     2020-08-14 20:43:42     2020-06-03 09:39:59     586     5       INR     f
uchsia  255-52-5654     5       5       1
Time taken: 3.843 seconds, Fetched: 10 row(s)
```

Clickstream table

```
hive> select * from clickstream limit 10;
OK
Android 3.2.35  fcba68aa-1231-11eb-adc1-0242ac120002    26564820        No      Yes     Yes     No      16.4454865      99.902065       de545711-3914-4450-8c11-b17b8dab
b5e1    "2020-09-14 09:59:07"
iOS     2.4.7   a95dd57b-779f-49db-819d-b6960483e554    31906387        No      No      Yes     Yes     -64.813749      -133.527040     de545711-3914-4450-8c11-b17b8dab
b5e1    "2020-05-16 16:30:21"
Android 3.4.12  fcba68aa-1231-11eb-adc1-0242ac120002    25713677        No      No      No      Yes     89.943435       127.313415      b328829e-17ae-11eb-adc1-0242ac12
0002    "2020-02-09 00:52:13"
Android 3.1.8   ele99492-17ae-11eb-adc1-0242ac120002    83474293        Yes     No      No      Yes     -69.939070      -36.451670      e7bc5fb2-1231-11eb-adc1-0242ac12
0002    "2020-06-17 10:42:50"
iOS     2.2.9   fcba68aa-1231-11eb-adc1-0242ac120002    63727807        No      Yes     Yes     Yes     64.082108       -81.822078      e7bc5fb2-1231-11eb-adc1-0242ac12
0002    "2020-07-06 02:51:53"
Android 4.3.19  ele99492-17ae-11eb-adc1-0242ac120002    73737907        No      Yes     Yes     No      -18.850508      -116.358375     b328829e-17ae-11eb-adc1-0242ac12
0002    "2020-04-26 06:18:16"
iOS     3.2.26  a95dd57b-779f-49db-819d-b6960483e554    36927433        Yes     Yes     Yes     No      -84.6857245     -146.507678     de545711-3914-4450-8c11-b17b8dab
b5e1    "2020-02-06 10:21:18"
Android 3.3.11  ele99492-17ae-11eb-adc1-0242ac120002    12691783        Yes     Yes     No      No      54.3852925      -37.411814      de545711-3914-4450-8c11-b17b8dab
b5e1    "2020-08-08 04:23:56"
iOS     4.4.36  a95dd57b-779f-49db-819d-b6960483e554    22635021        No      No      No      No      -31.805500      150.655650      e7bc5fb2-1231-11eb-adc1-0242ac12
0002    "2020-08-02 00:33:50"
Android 1.2.16  ele99492-17ae-11eb-adc1-0242ac120002    23593546        Yes     No      No      Yes     8.8918475       -83.929878      de545711-3914-4450-8c11-b17b8dab
b5e1    "2020-07-23 23:59:19"
Time taken: 3.491 seconds, Fetched: 10 row(s)
hive>
```

Datewise aggregate booking table

```
hive> select * from bookings_per_day limit 10;
OK
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
2020-01-10      2
Time taken: 0.265 seconds, Fetched: 10 row(s)
hive>
```