Load data from Spark to HDFS

Commands used to clean and aggregate data

The python file will need some configurations. First, we will need to install **pyspark** using the command below.

```
conda install pyspark==2.3.0
```

We have to install this version of pyspark because we have Apache Spark 2.3.0 installed in our ec2-instance.

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

```
python spark_local_flatten_datewise_aggregates.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/
```

Screenshots of command process

```
[root@ip-10-0-0-174 capstone_project]# python spark_local_flatten_datewise_aggregates.py
WARNING: User-defined SPARK_HOME (/opt/cloudera/parcels/SPARK2-2.3.0.cloudera2-1.cdh5.13.3.p0.316101/lib/spark2) overrides detected (/opt/cloudera/parcels/S
PARK2-2.3.0.cloudera2-1.cdh5.13.3.p0.316101/lib/spark2/).
WARNING: Running spark-class from user-defined location.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
[Stage 8:==============>                                          (50 + 3) / 200]
```

```
[root@ip-10-0-0-174 capstone_project]# python spark_local_flatten_datewise_aggregates.py
WARNING: User-defined SPARK_HOME (/opt/cloudera/parcels/SPARK2-2.3.0.cloudera2-1.cdh5.13.3.p0.316101/lib/spark2) overrides detected (/opt/cloudera/parcels/S
PARK2-2.3.0.cloudera2-1.cdh5.13.3.p0.316101/lib/spark2/).
WARNING: Running spark-class from user-defined location.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
22/04/23 12:20:55 WARN util.Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.
Successfully loaded dataframes to Hive
[root@ip-10-0-0-174 capstone_project]#
```

Bookings table

```
hive> select * from bookings limit 10;
OK
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/opt/cloudera/parcels/CDH-5.15.1-1.cdh5.15.1.p0.4/jars/parquet-pig-bundle-1.5.0-cdh5.15.1.jar!/shaded/parquet/org/slf4j/im
pl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/opt/cloudera/parcels/CDH-5.15.1-1.cdh5.15.1.p0.4/jars/parquet-hadoop-bundle-1.5.0-cdh5.15.1.jar!/shaded/parquet/org/slf4j
/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/opt/cloudera/parcels/CDH-5.15.1-1.cdh5.15.1.p0.4/jars/parquet-format-2.1.0-cdh5.15.1.jar!/shaded/parquet/org/slf4j/impl/S
taticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/opt/cloudera/parcels/CDH-5.15.1-1.cdh5.15.1.p0.4/jars/hive-exec-1.1.0-cdh5.15.1.jar!/shaded/parquet/org/slf4j/impl/Static
LoggerBinder.class]
SLF4J: Found binding in [jar:file:/opt/cloudera/parcels/CDH-5.15.1-1.cdh5.15.1.p0.4/jars/hive-jdbc-1.1.0-cdh5.15.1-standalone.jar!/shaded/parquet/org/slf4j/
impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [shaded.parquet.org.slf4j.helpers.NOPLoggerFactory]
BK8968087150    51811359        15055660        2.2.14  Android -49.431965      103.917854      -58.804386      146.47737       2020-06-23 19:33:10     2020
-06-06 09:02:10 534     83      INR     black   054-38-4479     4       3       3
BK629851904     31663218        60872180        3.4.1   iOS     -83.54084       175.80086       86.20705        128.36723       2020-05-23 12:22:04     2020
-08-09 19:02:56 126     67      INR     lime    796-39-6801     3       2       4
BK1797410350    86869399        94276051        4.1.36  iOS     -67.89307       55.234127       -51.1079        -31.07475       2020-05-19 14:14:32     2020
-08-23 18:38:39 297     63      INR     olive   748-73-1579     1       3       3
BK5788246325    58230837        45457227        2.4.27  Android 13.707887       113.49995       54.38129        -18.437752      2020-03-24 01:30:15     2020
-05-19 11:16:45 932     32      INR     white   558-80-6346     3       2       2
BK8342703255    84232510        86494681        4.1.34  Android -6.091461       -114.64979      22.84495        70.137825       2020-08-03 19:10:52     2020
-03-24 08:25:40 260     7       INR     blue    068-72-1637     3       3       3
BK6015582453    11981042        35862658        2.4.39  iOS     -18.910034      -70.1931        -10.182921      173.87721       2020-07-17 05:33:48     2020
-04-30 04:54:27 907     53      INR     purple  102-10-5639     3       2       3
BK4529355854    60071878        78022360        2.1.9   iOS     1.215274        -56.014904      35.152878       104.324905      2020-01-02 01:48:40     2020
-02-16 04:28:55 547     17      INR     teal    866-83-4349     2       3       4
BK9720088219    14327312        94427067        3.1.2   Android -55.482224       173.36226       65.01212        51.39075        2020-04-10 15:11:07     2020
-01-20 21:17:42 259     33      INR     maroon  572-73-6526     3       3       2
BK7157532607    46407210        43160003        1.3.4   Android 46.005844       -16.826145      7.6126013       -156.42857      2020-06-09 05:56:31     2020
-03-19 01:53:16 787     21      INR     olive   667-23-5880     2       2       3
BK5014871433    65861573        64708618        1.3.28  iOS     -29.565327      64.84371        84.06811        -49.820835      2020-08-14 20:43:42     2020
-06-03 09:39:59 586     5       INR     fuchsia 255-52-5654     5       5       1
Time taken: 0.337 seconds, Fetched: 10 row(s)
hive>
```

Clicking_stream table

```
hive> select * from clicking_stream limit 10;
OK
26564820        3.2.35  Android 16.4454865      99.902065       de545711-3914-4450-8c11-b17b8dabb5e1    fcba68aa-1231-11eb-adc1-0242ac120002    No      YesN
o       Yes     2020-09-14 09:59:07
31906387        2.4.7   iOS     -64.813749      -133.52704      de545711-3914-4450-8c11-b17b8dabb5e1    a95dd57b-779f-49db-819d-b6960483e554    No      No Y
es      Yes     2020-05-16 16:30:21
25713677        3.4.12  Android 89.943435       127.313415      b328829e-17ae-11eb-adc1-0242ac120002    fcba68aa-1231-11eb-adc1-0242ac120002    No      No Y
es      No      2020-02-09 00:52:13
83474293        3.1.8   Android -69.93907       -36.45167       e7bc5fb2-1231-11eb-adc1-0242ac120002    e1e99492-17ae-11eb-adc1-0242ac120002    Yes     No Y
es      No      2020-06-17 10:42:50
63727807        2.2.9   iOS     64.082108       -81.822078      e7bc5fb2-1231-11eb-adc1-0242ac120002    fcba68aa-1231-11eb-adc1-0242ac120002    No      YesY
es      Yes     2020-07-06 02:51:53
73737907        4.3.19  Android -18.850508      -116.358375     b328829e-17ae-11eb-adc1-0242ac120002    e1e99492-17ae-11eb-adc1-0242ac120002    No      YesN
o       Yes     2020-04-26 06:18:16
36927433        3.2.26  iOS     -84.6857245     -146.507678     de545711-3914-4450-8c11-b17b8dabb5e1    a95dd57b-779f-49db-819d-b6960483e554    Yes     YesN
o       Yes     2020-02-06 10:21:18
12691783        3.3.11  Android 54.3852925      -37.411814      de545711-3914-4450-8c11-b17b8dabb5e1    e1e99492-17ae-11eb-adc1-0242ac120002    Yes     YesN
o       No      2020-08-08 04:23:56
22635021        4.4.36  iOS     -31.8055        150.65565       e7bc5fb2-1231-11eb-adc1-0242ac120002    a95dd57b-779f-49db-819d-b6960483e554    No      No N
o       No      2020-08-02 00:33:50
23593546        1.2.16  Android 8.8918475       -83.929878      de545711-3914-4450-8c11-b17b8dabb5e1    e1e99492-17ae-11eb-adc1-0242ac120002    Yes     No Y
es      No      2020-07-23 23:59:19
Time taken: 0.055 seconds, Fetched: 10 row(s)
hive>
```

Datewise aggregate booking table

```
hive> select * from datewise_bookings 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: 0.078 seconds, Fetched: 10 row(s)
hive>
```