Create Hive-Managed Tables

For Clickstream and RDS data, we have previously produced a hive table in steps one and two, respectively.

1. **Hive table for clickstream data:**

```
CREATE EXTERNAL TABLE clickstream_cleaned (
   customer_id STRING,
   app_version STRING,
   OS_version STRING,
   lat DOUBLE,
   lon DOUBLE,
   page_id STRING,
   button_id STRING,
   is_button_click STRING,
   is_page_view STRING,
   is_scroll_up STRING,
   is_scroll_down STRING,
   `timestamp` STRING
    )
   ROW FORMAT DELIMITED
   FIELDS TERMINATED BY ','
   STORED AS TEXTFILE
   LOCATION '/user/poushali/clickstream_flattened';
```

2. **Hive table for Booking data:**

```
CREATE EXTERNAL TABLE IF NOT EXISTS rds_bookings (
   booking_id STRING,
   customer_id BIGINT,
   driver_id BIGINT,
   customer_app_version STRING,
   customer_phone_os_version STRING,
   pickup_lat DOUBLE,
   pickup_lon DOUBLE,
   drop_lat DOUBLE,
   drop_lon DOUBLE,
   pickup_timestamp BIGINT,
   drop_timestamp BIGINT,
   trip_fare INT,
   tip_amount INT,
   currency_code STRING,
   cab_color STRING,
    cab_registration_no STRING,
    customer_rating_by_driver INT,
    rating_by_customer INT,
    passenger_count INT
)
STORED AS PARQUET
LOCATION '/user/poushali/rds_import/bookings';
```

3. Hive table for aggregated data:

```
CREATE TABLE datewise_booking_aggregates (
  booking_date DATE,
  total_bookings INT
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE;
```

- The datewise_booking_aggregates table is created by the supplied Hive query and is intended to hold daily booking summaries.
- Booking_date, which records the date of each booking activity, and total_bookings, which stores the total number of bookings for that particular date, make up its two columns.
- Using ROW FORMAT DELIMITED and fields separated by commas (FIELDS TERMINATED BY ','), the table is set up to handle data in a CSV format. \
- Since the table data is saved as a TEXTFILE, it will be in plain text format. Simple, legible datasets are usually stored in this configuration, which is also appropriate for loading outputs such as the CSV file produced by the PySpark aggregation.

For clarity and consistency, we renamed the previous tables:

```
ALTER TABLE clickstream_cleaned RENAME TO clickstream_data;
ALTER TABLE rds_bookings RENAME TO bookings_data;
```

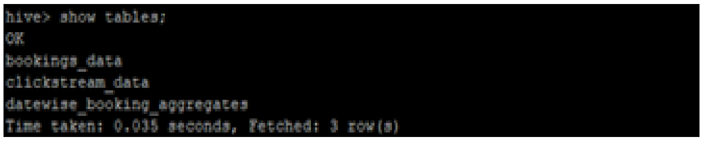
4. Command to load the data into Hive tables

Only datewise_booking_aggregates requires a data load from HDFS.

Command/s

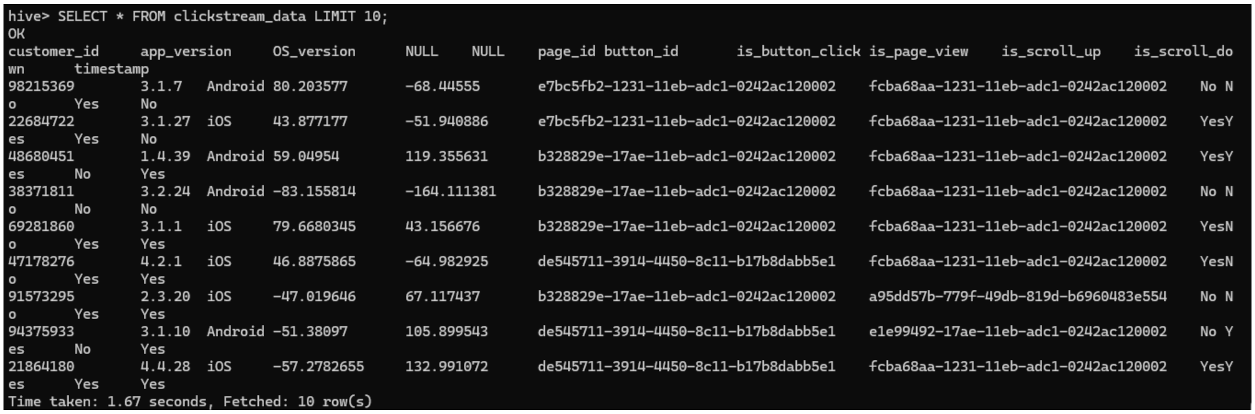
```
LOAD DATA INPATH '/user/poushali/datewise_bookings_output/part-00000-80d4e987-cbeb-
44b9-b304-703ae2d9adcb-c000.csv'
INTO TABLE datewise_booking_aggregates;
```

Checking data in all 3 tables:



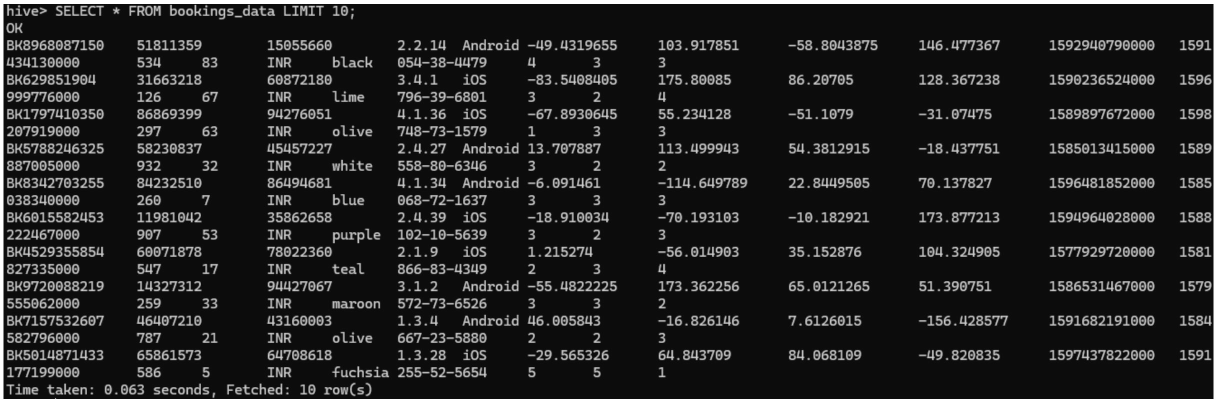
```
hive> show tables;
OK
bookings_data
clickstream_data
datewise_booking_aggregates
Time taken: 0.035 seconds, Fetched: 3 row(s)
```

SELECT * FROM clickstream_data LIMIT 10;



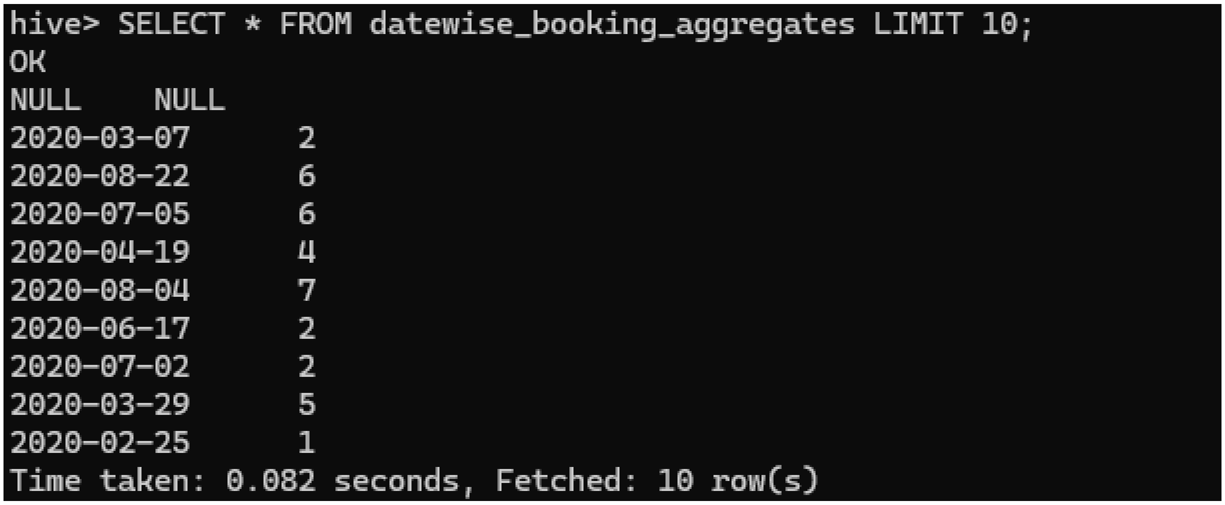
```
hive> SELECT * FROM clickstream_data LIMIT 10;
OK
customer_id     app_version     OS_version      NULL    NULL    page_id button_id       is_button_click is_page_view    is_scroll_up    is_scroll_do
wn      timestamp
98215369        3.1.7   Android 80.203577       -68.44555       e7bc5fb2-1231-11eb-adc1-0242ac120002    fcba68aa-1231-11eb-adc1-0242ac120002    No N
o       Yes     No
22684722        3.1.27  iOS     43.877177       -51.940886      e7bc5fb2-1231-11eb-adc1-0242ac120002    fcba68aa-1231-11eb-adc1-0242ac120002    YesY
es      Yes     No
48680451        1.4.39  Android 59.04954        119.355631      b328829e-17ae-11eb-adc1-0242ac120002    fcba68aa-1231-11eb-adc1-0242ac120002    YesY
es      No      Yes
38371811        3.2.24  Android -83.155814      -164.111381     b328829e-17ae-11eb-adc1-0242ac120002    fcba68aa-1231-11eb-adc1-0242ac120002    No N
o       No      No
69281860        3.1.1   iOS     79.6680345      43.156676       b328829e-17ae-11eb-adc1-0242ac120002    fcba68aa-1231-11eb-adc1-0242ac120002    YesN
o       Yes     Yes
47178276        4.2.1   iOS     46.8875865      -64.982925      de545711-3914-4450-8c11-b17b8dabb5e1    fcba68aa-1231-11eb-adc1-0242ac120002    YesN
o       Yes     Yes
91573295        2.3.20  iOS     -47.019646      67.117437       b328829e-17ae-11eb-adc1-0242ac120002    a95dd57b-779f-49db-819d-b6960483e554    No N
o       Yes     Yes
94375933        3.1.10  Android -51.38097       105.899543      de545711-3914-4450-8c11-b17b8dabb5e1    e1e99492-17ae-11eb-adc1-0242ac120002    No Y
es      No      Yes
21864180        4.4.28  iOS     -57.2782655     132.991072      de545711-3914-4450-8c11-b17b8dabb5e1    fcba68aa-1231-11eb-adc1-0242ac120002    YesY
es      Yes     Yes
Time taken: 1.67 seconds, Fetched: 10 row(s)
```

SELECT * FROM bookings_data LIMIT 10;



```
hive> SELECT * FROM bookings_data LIMIT 10;
OK
BK8968087150    51811359        15055660        2.2.14  Android -49.4319655     103.917851      -58.8043875     146.477367      1592940790000   1591
434130000       534     83      INR     black   054-38-4479     4       3       3
BK629851904     31663218        60872180        3.4.1   iOS     -83.5408405     175.80085       86.20705        128.367238      1590236524000   1596
999776000       126     67      INR     lime    796-39-6801     3       2       4
BK1797410350    86869399        94276051        4.1.36  iOS     -67.8930645     55.234128       -51.1079        -31.07475       1589897672000   1598
207919000       297     63      INR     olive   748-73-1579     1       3       3
BK5788246325    58230837        45457227        2.4.27  Android 13.707887       113.499943      54.3812915      -18.437751      1585013415000   1589
887005000       932     32      INR     white   558-80-6346     3       2       2
BK8342703255    84232510        86494681        4.1.34  Android -6.091461       -114.649789     22.8449505      70.137827       1596481852000   1585
038340000       260     7       INR     blue    068-72-1637     3       3       3
BK6015582453    11981042        35862658        2.4.39  iOS     -18.910034      -70.193103      -10.182921      173.877213      1594964028000   1588
222467000       907     53      INR     purple  102-10-5639     3       2       3
BK4529355854    60071878        78022360        2.1.9   iOS     1.215274        -56.014903      35.152876       104.324905      1577929720000   1581
827335000       547     17      INR     teal    866-83-4349     2       3       4
BK9720088219    14327312        94427067        3.1.2   Android -55.4822225     173.362256      65.0121265      51.390751       1586531467000   1579
555062000       259     33      INR     maroon  572-73-6526     3       3       2
BK7157532607    46407210        43160003        1.3.4   Android 46.005843       -16.826146      7.6126015       -156.428577     1591682191000   1584
582796000       787     21      INR     olive   667-23-5880     2       2       3
BK5014871433    65861573        64708618        1.3.28  iOS     -29.565326      64.843709       84.068109       -49.820835      1597437822000   1591
177199000       586     5       INR     fuchsia 255-52-5654     5       5       1
Time taken: 0.063 seconds, Fetched: 10 row(s)
```

SELECT * FROM datewise_booking_aggregates LIMIT 10;



```
hive> SELECT * FROM datewise_booking_aggregates LIMIT 10;
OK
NULL    NULL
2020-03-07      2
2020-08-22      6
2020-07-05      6
2020-04-19      4
2020-08-04      7
2020-06-17      2
2020-07-02      2
2020-03-29      5
2020-02-25      1
Time taken: 0.082 seconds, Fetched: 10 row(s)
```