Create Hive-Managed Tables

<Command to create the Hive tables>

Before creating the hive tables, Follow the following commands:

Type Hive to get into Hive CLI
Then Create a Database:

- ➔ Create Database CabRideProject;
- ➔ Use CabRideProject;

1. **Creating the Hive table to store the Booking data**

➔

```
CREATE TABLE IF NOT EXISTS Bookings
(
booking_id STRING,
customer_id INT,
driver_id INT,
customer_app_version STRING,
customer_phone_os_version STRING,
pickup_lat DOUBLE,
pickup_lon DOUBLE,
drop_lat DOUBLE,
drop_lon DOUBLE,
pickup_timestamp TIMESTAMP,
drop_timestamp TIMESTAMP,
trip_fare DECIMAL(10, 2),
tip_amount DECIMAL(10, 2),
 currency_code STRING,
cab_color STRING,
cab_registration_no STRING,
customer_rating_by_driver INT,
rating_by_customer INT,
passenger_count INT
);
```

Screen shot of the table creation

```
[hadoop@ip-172-31-36-71 ~]$ hive
Hive Session ID = 4ecd7cf1-bfa2-4955-af4f-869c43b9aee1

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: true
hive> create database CabrideProject;
OK
Time taken: 0.388 seconds
hive> CREATE TABLE IF NOT EXISTS Bookings
    > (
    > booking_id STRING,
    > customer_id INT,
    > driver_id INT,
    > customer_app_version STRING,
    > customer_phone_os_version STRING,
    > pickup_lat DOUBLE,
    > pickup_lon DOUBLE,
    > drop_lat DOUBLE,
    > drop_lon DOUBLE,
    > pickup_timestamp TIMESTAMP,
    > drop_timestamp TIMESTAMP,
    > trip_fare DECIMAL(10, 2),
    > tip_amount DECIMAL(10, 2),
    >  currency_code STRING,
    > cab_color STRING,
    > cab_registration_no STRING,
    > customer_rating_by_driver INT,
    > rating_by_customer INT,
    > passenger_count INT
    > );
OK
Time taken: 0.415 seconds
hive> show tables;
OK
bookings
Time taken: 0.092 seconds, Fetched: 1 row(s)
hive>
```

2. **Creating table for storing then aggregated booking values [daywise]**

➔

```
CREATE TABLE IF NOT EXISTS datewise_total_bookings
( pickup_date DATE,
total_bookings INT );
```

Screenshot for the same :

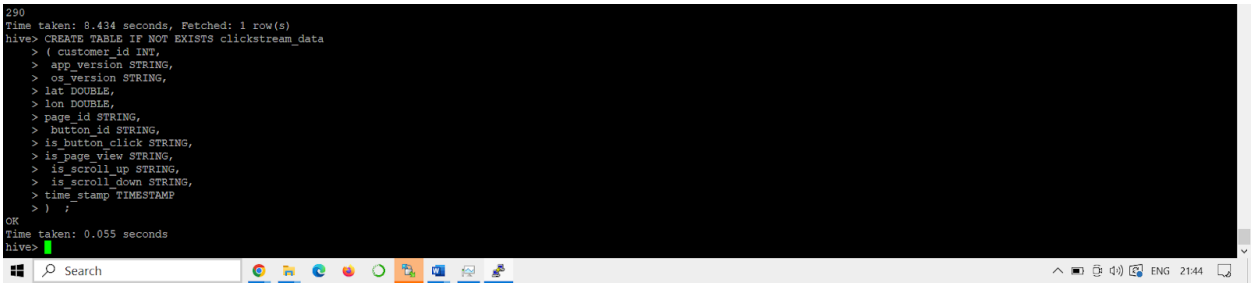
```
[hadoop@ip-172-31-36-71 ~]$ hive
Hive Session ID = fcee3ff1-0411-425a-8feb-08e1338c9820

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: true
hive> show tables;
OK
bookings
Time taken: 0.426 seconds, Fetched: 1 row(s)
hive> CREATE TABLE IF NOT EXISTS datewise_total_bookings
    > ( pickup_date DATE,
    > total_bookings INT );
OK
Time taken: 0.218 seconds
hive>
```

3. **Creating table for storing data for the click stream**

➔

```
CREATE TABLE IF NOT EXISTS clickstream_data
( customer_id INT,
 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,
time_stamp TIMESTAMP
)  ;
```

Screenshot for the table creation command execution:



<Command to load the data into Hive tables>

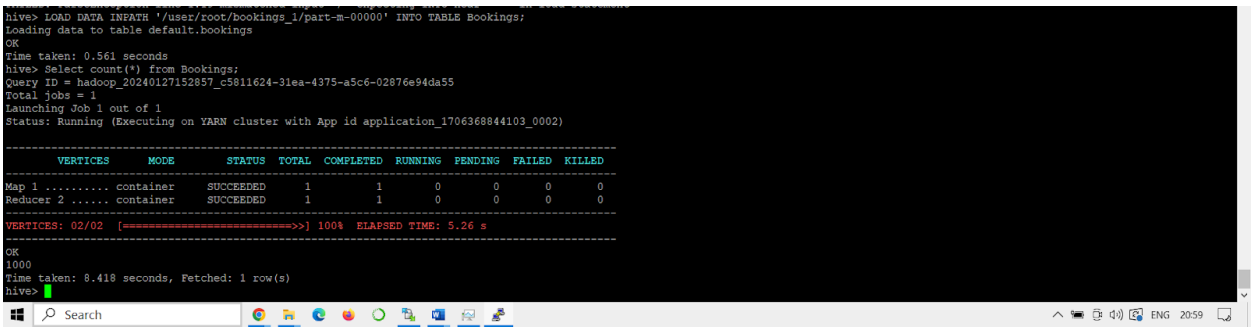
Loading the data from hadoop to the hive tables;

1. Uploading the data from hadoop to Hive table for storing the bookings data ingested from AWS RDS to hadoop using sqoop

   ➔

   LOAD DATA INPATH '/user/root/bookings_1/part-m-00000’ OVERWRITE INTO TABLE Bookings

   Screenshot for the running the same command and verifying the number of rows in the hive table



2. Uploading the data from hadoop to Hive table for the aggregated data created as a datewise aggregation of bookings

   ➔

   LOAD DATA INPATH '/user/root/datewise_bookings_agg' OVERWRITE INTO TABLE datewise_total_bookings;

The image also contains verification of rows present in the hive table

```
hive> LOAD DATA INPATH '/user/root/datewise_bookings_agg' OVERWRITE INTO TABLE datewise_total_bookings;
Loading data to table default.datewise_total_bookings
OK
Time taken: 0.362 seconds
hive> Select count(*) from datewise_total_bookings;
Query ID = hadoop_20240127161007_5866e85a-5ed6-41ef-a56f-c0bc7947ff01
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1706368844103_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 5.65 s
--------------------------------------------------------------------------------
OK
290
Time taken: 8.434 seconds, Fetched: 1 row(s)
hive>
```

3. Uploading the data form hadoop to hive -> This is for the data pulled form kafka and stored in the hadoop

   ➔

   LOAD DATA INPATH ‘/user/root/clickstream_flattened’ into table clickstream_data;

   Below image also holds the verification query to run the numbers of rows

```
hive> load data inpath '/user/root/clickstream_flattened' into table clickstream_data ;
Loading data to table default.clickstream_data
OK
Time taken: 1.004 seconds
hive> Select count(*) from clickstream_data
    > ;
Query ID = hadoop_20240127162154_8017e86a-c776-4bf0-bc55-7cc6c7fb067a
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1706368844103_0006)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 5.72 s
--------------------------------------------------------------------------------
OK
3004
Time taken: 8.511 seconds, Fetched: 1 row(s)
hive>
```