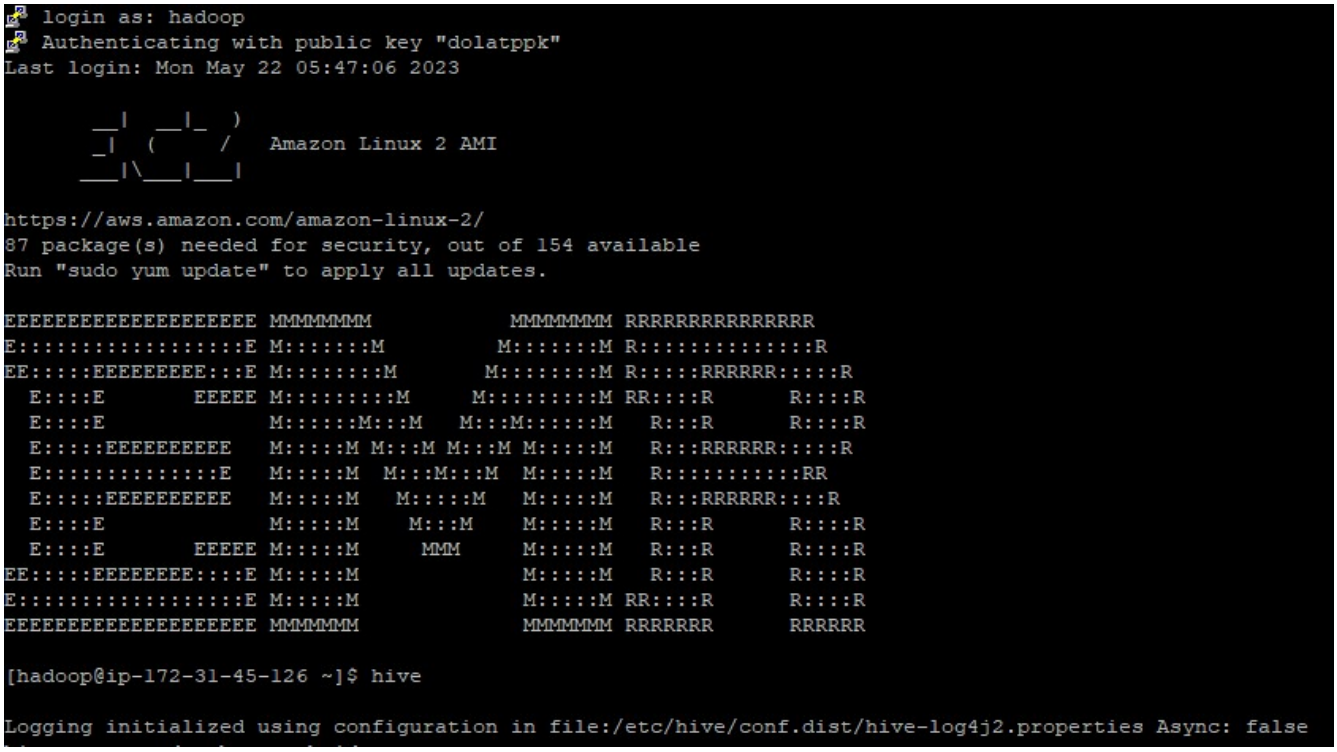
Create Hive-Managed Tables

Creating Hive database and tables:

1. **Connect to Hive instance:**



2. **Create database to create bookings, clickstream and aggregated tables in it:**

create database cabrides;
use cabrides;

3. **Creating bookings, clickstreaming and datewise_total_bookings (aggregated) tables:**

```
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 BOOLEAN,
 is_page_view BOOLEAN,
 is_scroll_up BOOLEAN,
 is_scroll_down BOOLEAN,
 time_stamp TIMESTAMP
)
COMMENT 'This table will store click streaming data red from kafka';
```

```
hive> 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 BOOLEAN,
    >   is_page_view BOOLEAN,
    >   is_scroll_up BOOLEAN,
    >   is_scroll_down BOOLEAN,
    >   time_stamp TIMESTAMP
    > )
    > COMMENT 'This table will store click streaming data red from kafka';
OK
Time taken: 0.067 seconds
hive>
```

```
CREATE TABLE IF NOT EXISTS bookings_detail (
 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
)
COMMENT 'This table will store Bookings data red from MySQL RDS';
```

```
hive> CREATE TABLE IF NOT EXISTS bookings_detail (
    >   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
    > )
    > COMMENT 'This table will store Bookings data red from MySQL RDS';
OK
Time taken: 0.069 seconds
hive>
```

```
CREATE TABLE IF NOT EXISTS datewise_total_bookings (
 pickup_date DATE,
 total_bookings INT
)
COMMENT 'This table will store aggregated count of booking by pickup date';
```

```
hive> CREATE TABLE IF NOT EXISTS datewise_total_bookings (
    >   pickup_date DATE,
    >   total_bookings INT
    > )
    > COMMENT 'This table will store aggregated count of booking by pickup date';
OK
Time taken: 0.075 seconds
hive>
```

Loading the data into Hive tables from HDFS files:

1. **Loading streaming file into clickstream_data table:**

LOAD DATA INPATH '/user/root/clickstream_data_dump/part-00000-e08eb778-7b21-4c2b-84cd-9e2c6467c5ea-c000.json' OVERWRITE INTO TABLE clickstream_data;

```
hive> LOAD DATA INPATH '/user/root/clickstream_data_dump/part-00000-e08eb778-7b21-4c2b-84cd-9e2c6467c5ea-c000.json' OVERWRITE INTO TABLE clickstream_data;
Loading data to table cabrides.clickstream_data
OK
Time taken: 1.073 seconds
```

Verifying count of records in clickstream_data table:

```
hive> select count(*) from clickstream_data;
Query ID = hadoop_20230523142148_d6bd8179-6ea4-4fa2-9b50-4b19bc4a9fab
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1684823754082_0014)

--------------------------------------------------------------------------------
        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: 5.20 s
--------------------------------------------------------------------------------
OK
3003
Time taken: 14.435 seconds, Fetched: 1 row(s)
hive>
```

Note: Here we could see that all around 3003 streaming events are captured and loaded into into bookings_detail Hive table.

2. **Loading bookings file into bookings_detail table:**

LOAD DATA INPATH '/user/root/bookings/part-m-00000' OVERWRITE INTO TABLE bookings_detail;

```
hive> LOAD DATA INPATH '/user/root/bookings/part-m-00000' OVERWRITE INTO TABLE bookings_detail;
Loading data to table cabrides.bookings_detail
OK
Time taken: 0.615 seconds
hive>
```

Verifying count of records in bookings_detail table:

select count(*) from bookings_detail;

```
hive> select count(*) from bookings_detail;
Query ID = hadoop_20230522070613_0a860f69-63e0-40d4-bda0-317ddf70a90a
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1684732171643_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: 5.63 s
--------------------------------------------------------------------------------
OK
1000
Time taken: 16.287 seconds, Fetched: 1 row(s)
hive>
```

Note: Here we could see that all 1000 records are inserted into bookings_detail Hive table.

3. **Loading aggregated bookings into datewise_total_bookings table:**

LOAD DATA INPATH '/user/root/datewise_bookings_agg/part-00000-e40bc997-e4ef-41e4-9bda-6bbc32bf47b7-c000.csv' OVERWRITE INTO TABLE datewise_total_bookings;

```
hive> LOAD DATA INPATH '/user/root/datewise_bookings_agg/part-00000-e40bc997-e4ef-41e4-9bda-6bbc32bf47b7-c000.csv' OVERWRITE INTO TABLE datewise_total_bookings;
Loading data to table default.datewise_total_bookings
OK
Time taken: 0.416 seconds
hive>
```


Verifying count of records in datewise_total_bookings table:

```
hive> select count(*) from datewise_total_bookings;
Query ID = hadoop_20230522164903_abc976d8-1466-446a-b9f4-fe5ace38dbb2
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1684732171643_0009)

--------------------------------------------------------------------------------
        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: 6.37 s
--------------------------------------------------------------------------------
OK
290
Time taken: 14.522 seconds, Fetched: 1 row(s)
hive>
```

Note: Here we could see that 290 unique pickup_date are identified and no of bookings are populated against each pickup_date.