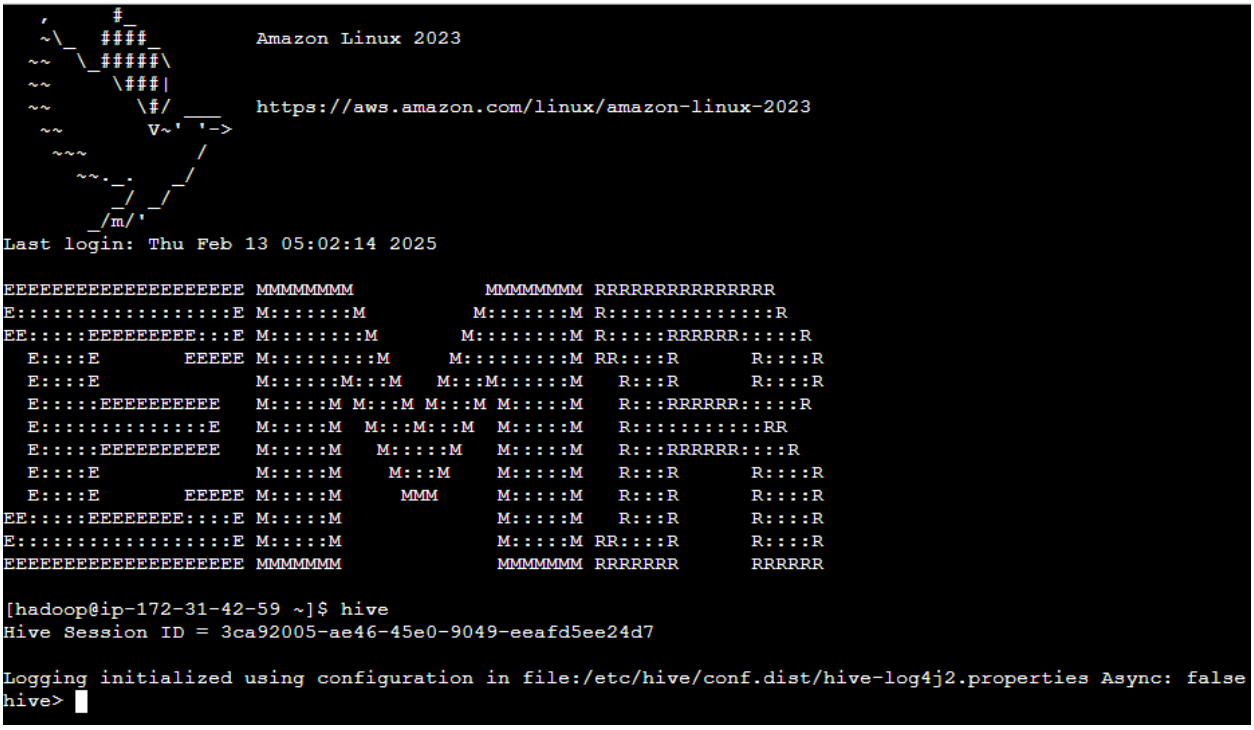
Create Hive-Managed Tables

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;
show databases;
use cabrides;
```

```
hive> create database cabrides;
OK
Time taken: 0.402 seconds
hive> show databases;
OK
cabrides
default
Time taken: 0.112 seconds, Fetched: 2 row(s)
hive> use cabrides;
OK
Time taken: 0.03 seconds
hive>
```

3. **Create clickstream, bookings and datewise_total_booking table;**

```
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)
COMMENT 'This table is for storing clickstream data 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 STRING,
    > is_page_view STRING,
    > is_scroll_up STRING,
    > is_scroll_down STRING,
    > time_stamp TIMESTAMP)
    > COMMENT 'This table is for storing clickstream data from Kafka';
OK
Time taken: 0.283 seconds
```

```
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 is for bookins detail data read from AWS 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 is for bookins detail data read from AWS RDS';
OK
Time taken: 0.068 seconds
```

```
CREATE TABLE IF NOT EXISTS datewise_total_bookings (
 pickup_date DATE,
 total_bookings INT)
 COMMENT 'This table is for datewise total bookings aggreagte data';
```

```
hive>  CREATE TABLE IF NOT EXISTS datewise_total_bookings (
    >  pickup_date DATE,
    >  total_bookings INT)
    >  COMMENT 'This table is for datewise total bookings aggreagte data';
OK
Time taken: 0.048 seconds
```

Loading the data into Hive tables from HDFS files:

1. **Loading clickstream data into clickstream_data table:**

```
LOAD DATA INPATH '/user/hadoop/kafka_stream/clickstream/part-00000-dfde3e6e-e227-4be7-9235-37d34707c461-c000.csv' OVERWRITE INTO TABLE clickstream_data;
```

```
hive> LOAD DATA INPATH '/user/hadoop/kafka_stream/clickstream/part-00000-dfde3e6e-e227-4be7-9235-37d34707c461-c000.csv' OVERWRITE INTO TABLE clickstream_data;
Loading data to table cabrides.clickstream_data
OK
Time taken: 0.365 seconds
```

Verifying count of records in clickstream_data table:

```
SELECT count(*) from clickstream_data;
```

```
hive> select count(*) from clickstream_data;
Query ID = hadoop_20250213055410_3d8fb584-3702-48de-b482-c124fa66c4cd
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1739422493977_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   container     SUCCEEDED     10         10        0        0       0       0
Reducer 2 ......   container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 20.07 s
--------------------------------------------------------------------------------
OK
4595792
Time taken: 26.31 seconds, Fetched: 1 row(s)
```

Total Number of records in clickstream_data: 4595792

2. **Loading data to bookings_detail table:**

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

```
hive> LOAD DATA INPATH '/user/hadoop/bookings/part-m-00000' OVERWRITE INTO TABLE bookings_detail;
Loading data to table cabrides.bookings_detail
OK
Time taken: 0.152 seconds
```

Verifying count of records in bookings_detail table:

SELECT count(*) from bookings_detail;

```
hive> select count(*) from bookings_detail;
Query ID = hadoop_20250213060441_835806e1-759c-4500-a7be-ea101d97c797
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1739422493977_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: 5.04 s
--------------------------------------------------------------------------------
OK
1000
Time taken: 11.709 seconds, Fetched: 1 row(s)
```

Total Number of records in bookings_detail table: 1000

3. **Loading data to datewise_total_bookings table:**

LOAD DATA INPATH '/user/hadoop/datewise_bookings_agg/part-00000-4c12c1f6-53f3-42d5-83e4-bec7b52f80bd-c000.csv' OVERWRITE INTO TABLE datewise_total_bookings;

```
hive> LOAD DATA INPATH '/user/hadoop/datewise_bookings_agg/part-00000-4c12c1f6-53f3-42d5-83e4-bec7b52f80bd-c000.csv' OVERWRITE INTO TABLE datewise_total_bookings;
Loading data to table cabrides.datewise_total_bookings
OK
Time taken: 0.204 seconds
```

Verifying count of records in datewise_total_bookigs table:

SELECT count(*) from datewise_total_bookings;

```
hive> select count(*) from datewise_total_bookings;
Query ID = hadoop_20250213061044_38b6a1e7-56a0-4202-af4d-3147007e9165
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1739422493977_0007)

--------------------------------------------------------------------------------
        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: 6.28 s
--------------------------------------------------------------------------------
OK
289
Time taken: 11.799 seconds, Fetched: 1 row(s)
```

Total Number of records in datewise_total_bookings table: 289