**Problem statement:**

**In this case study, we are giving a real world example of how to use HIVE on top of the HADOOP for different exploratory data analysis. In here, we have a predefined dataset (yellow\_tripdata\_2015-01-06.csv) having more than 15 columns.**

**Queries for Hive Case study**

**Tasks:**

**1.Create a table named taxidata . Required ddl script is given below.**

Create database challenge;

Use challenge;

CREATE TABLE IF NOT EXISTS taxidata(vendor\_id string, pickup\_datetime string,dropoff\_datetime string, passenger\_count int,

trip\_distance Float,pickup\_longitude Float, pickup\_latitude Float, rate\_code int,store\_and\_fwd\_flag string, dropoff\_longitude Float,

dropoff\_latitude Float,payment\_type string, fare\_amount Float, extra Float,mta\_tax Float, tip\_amount Float, tolls\_amount Float,

total\_amount Float, trip\_time\_in\_secs int )ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

TBLPROPERTIES ("skip.header.line.count"="1");

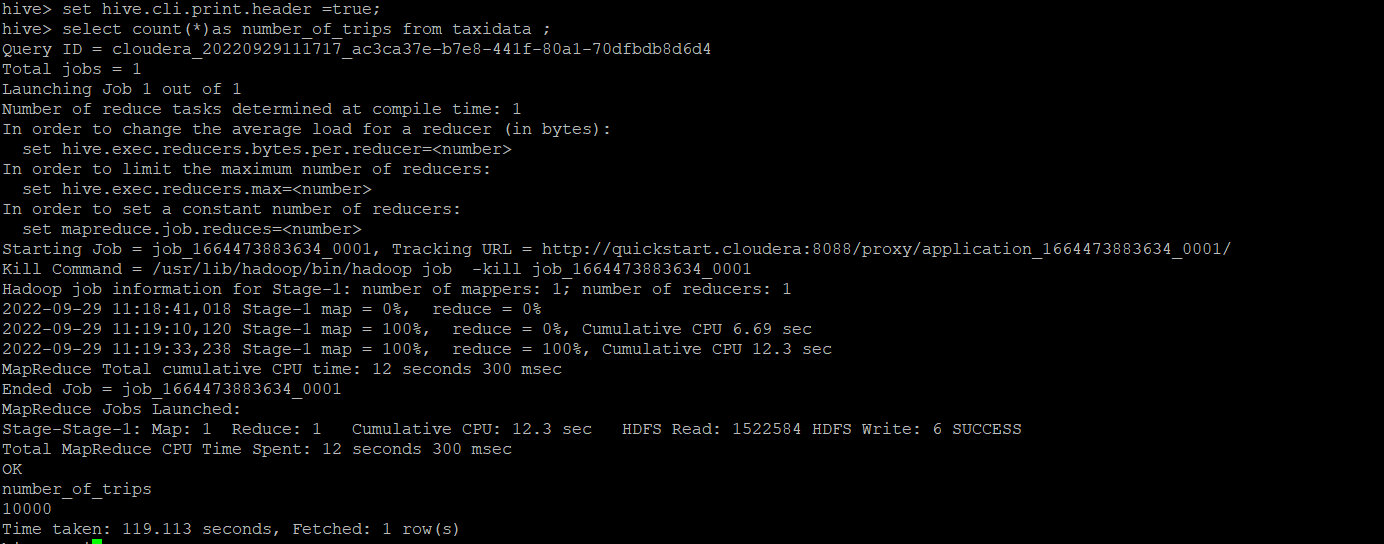
**2.Load data from the csv file - yellow\_tripdata\_2015-01-06.csv**

load data local inpath"/home/cloudera/sidd/Challenge/Mini project 3/yellow\_tripdata.csv" into TABLE taxidata;

**Perform taxi trip analysis by solving the questions below:**

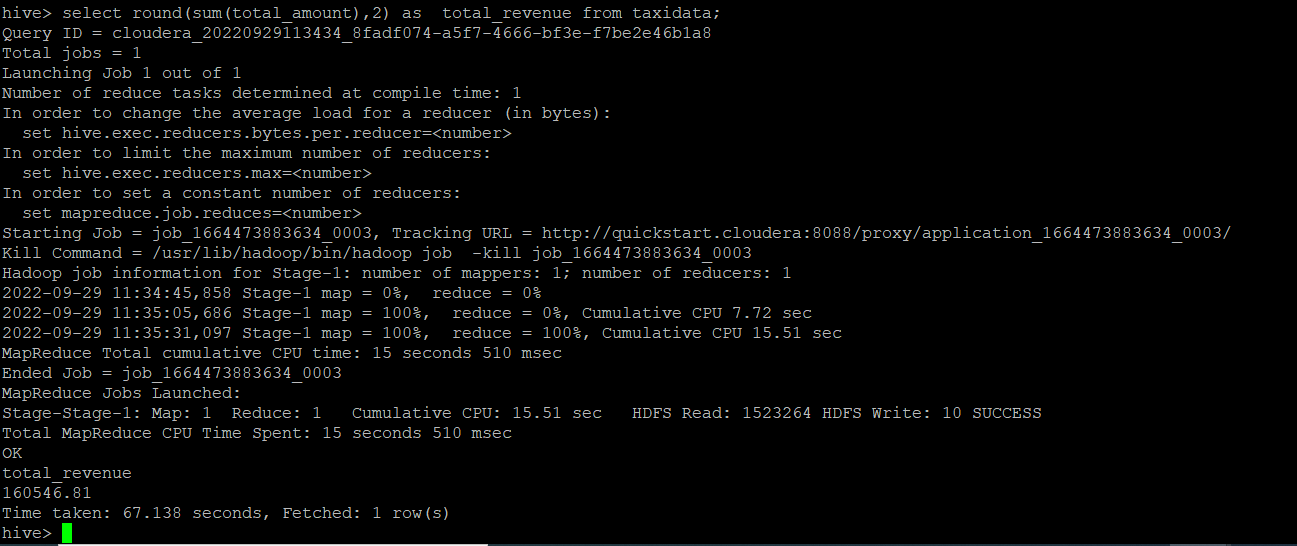
1. **What is the total number of trips ( equal to the number of rows)?**

**Hive> select count(\*) as number\_of\_trips from taxidata;**



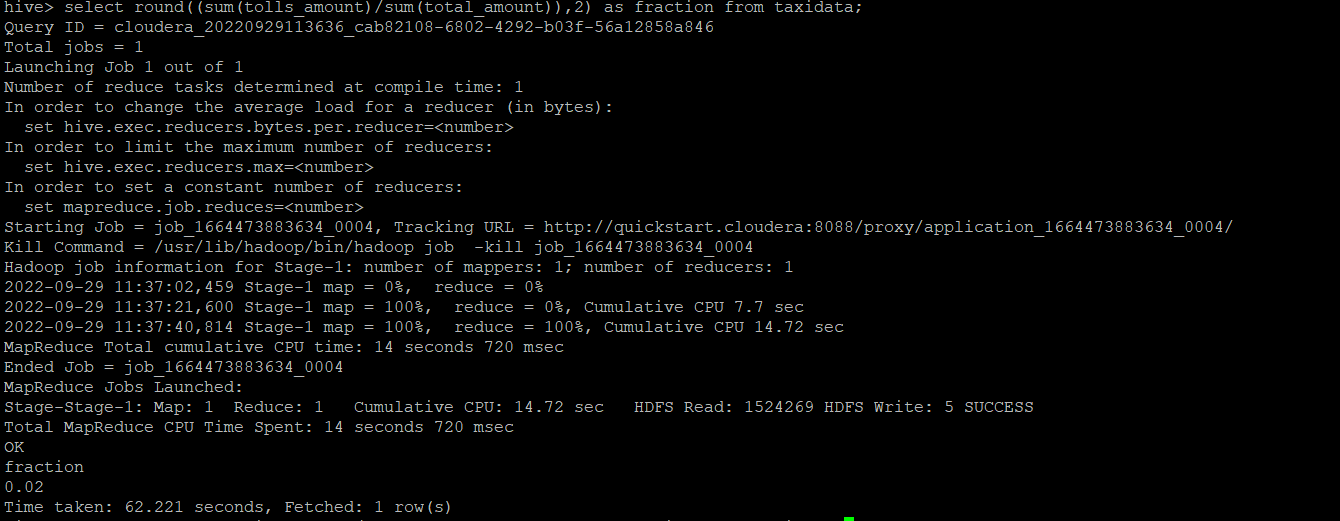
1. **What is the total revenue generated by all the trips? The fare is stored in the column total\_amount.**

**Hive> select round(sum(total\_amount),2) as total\_revenue from taxidata;**



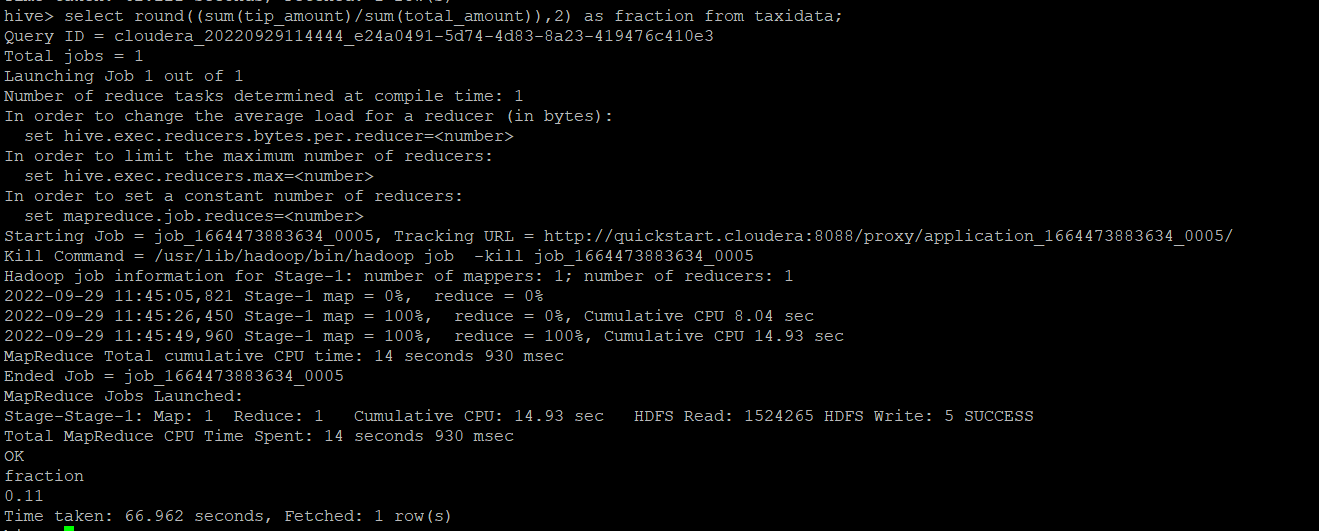
1. **What fraction of the total is paid for tolls? The toll is stored in tolls\_amount.**

**Hive> select round((sum(tolls\_amount)/sum(total\_amount)),2) as fraction from taxidata;**



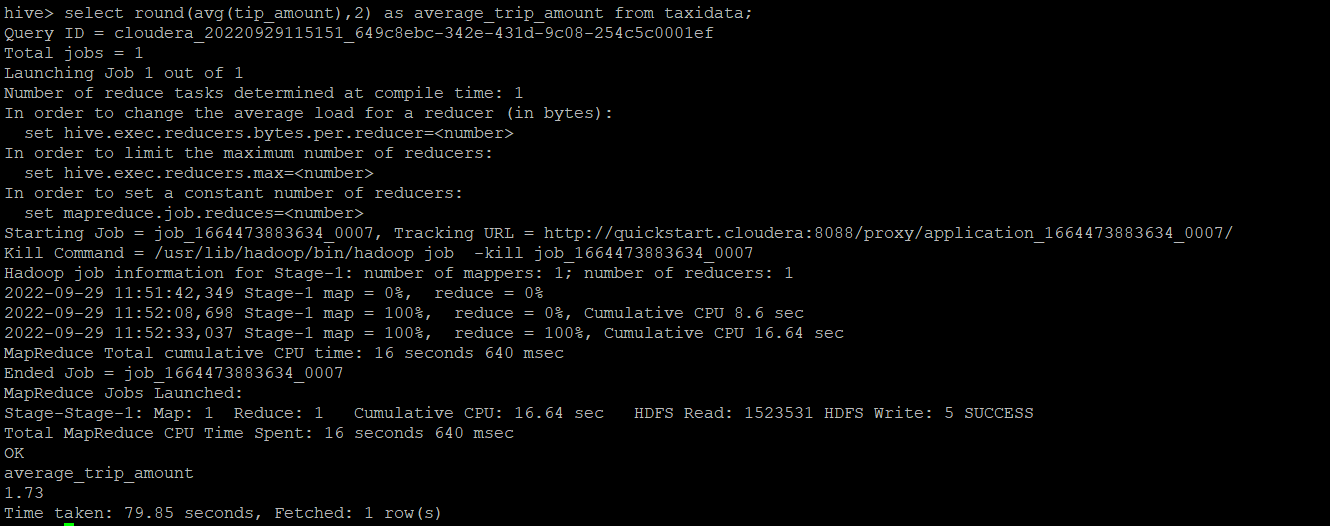
1. **What fraction of it is driver tips? The tip is stored in tip\_amount.**

**Hive> select round((sum(tip\_amount)/sum(total\_amount)),2) as fraction from taxidata;**



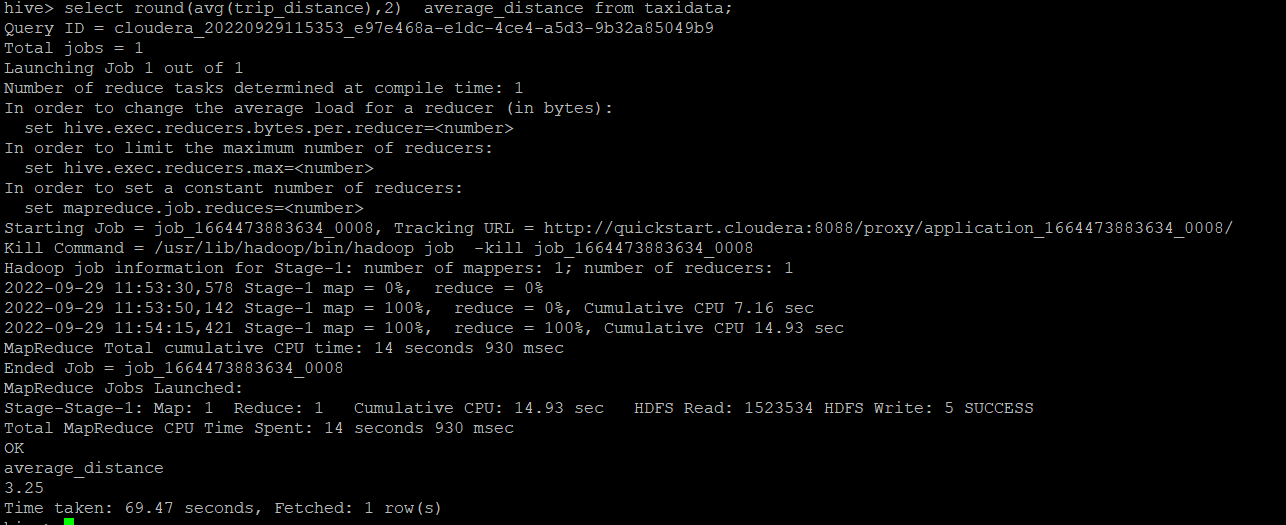
1. **What is the average trip amount?**

**Hive> select round(avg(tip\_amount),2) as average\_trip\_amount from taxidata;**



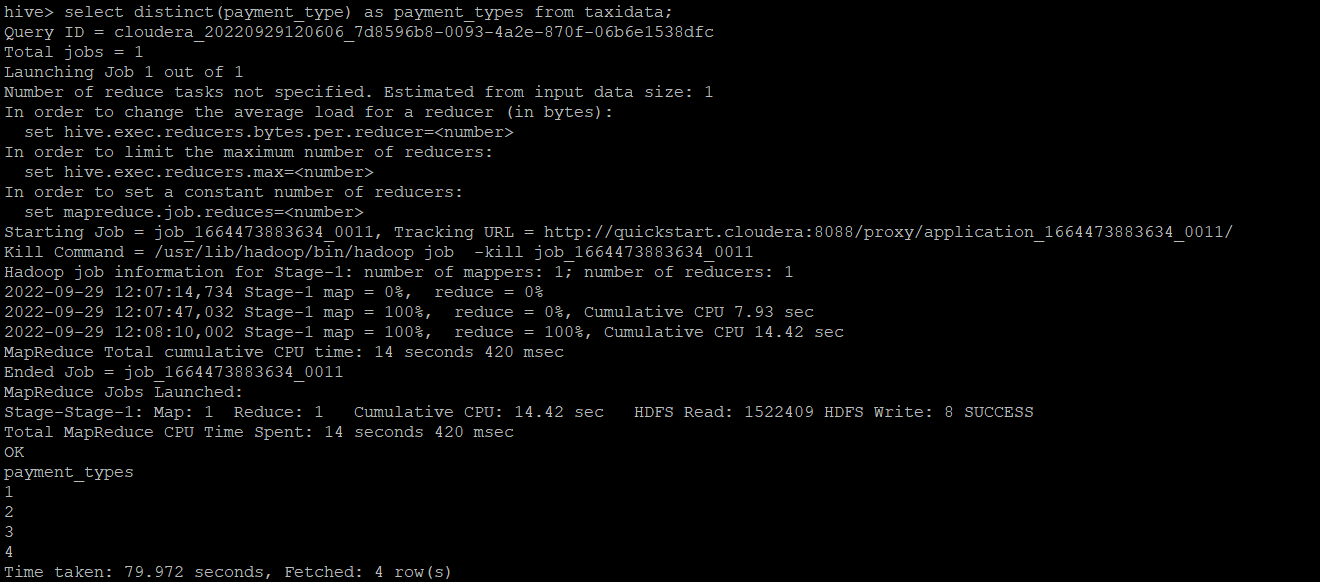
1. **What is the average distance of the trips? Distance is stored in the column trip\_distance.**

**Hive> select round(avg(trip\_distance),2) average\_distance from taxidata;**



1. **How many different payment types are used?**

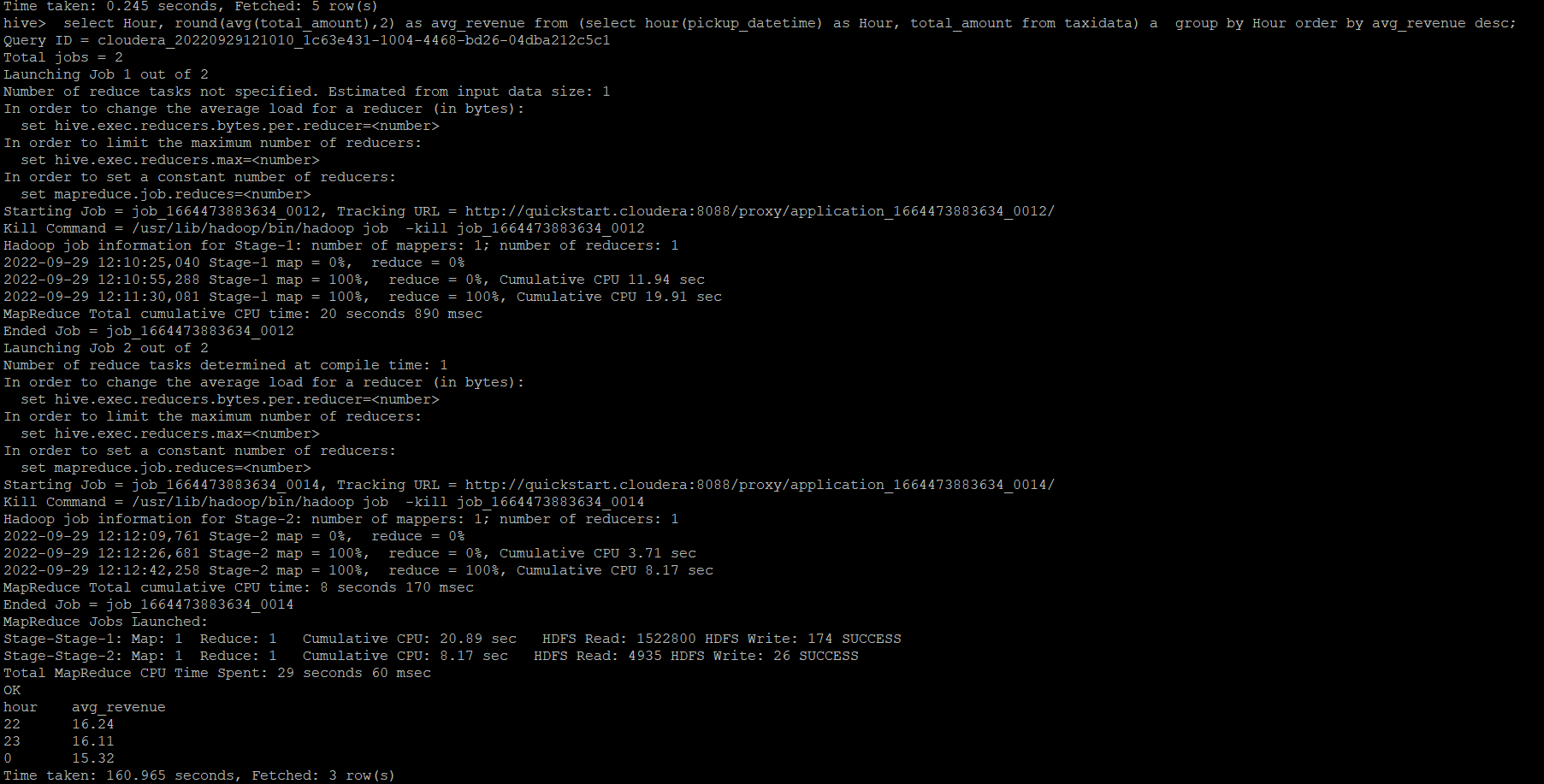
**Hive> select distinct(payment\_type) as payment\_types from taxidata;**



1. **For each payment type, display the following details:**

* **Average fare generated**
* **Average tip**
* **Average tax – tax is stored in column mta\_tax**

**Hive> select payment\_type, round(avg(fare\_amount),2) as average\_fare, round(avg(tip\_amount),2) as average\_tip, round(avg(mta\_tax),2) as average\_tax from taxidata group by payment\_type;**



1. **On average which hour of the day generates the highest revenue?**

**Hive> select Hour, round(avg(total\_amount),2) as avg\_revenue from (select hour(pickup\_datetime) as Hour, total\_amount from taxidata) a group by Hour order by avg\_revenue desc;**

