**Final Project**

**Dhrumil Rana**

**Project 3: Hadoop- HDFS-Sqoop-HIVE**

**Dataset: Yellow Taxicabs**

**Convert CSV file into MySQL database table then use Sqoop to Export database table into Hive table and then using Hive query to perform analysis**

**Create table in mysql**

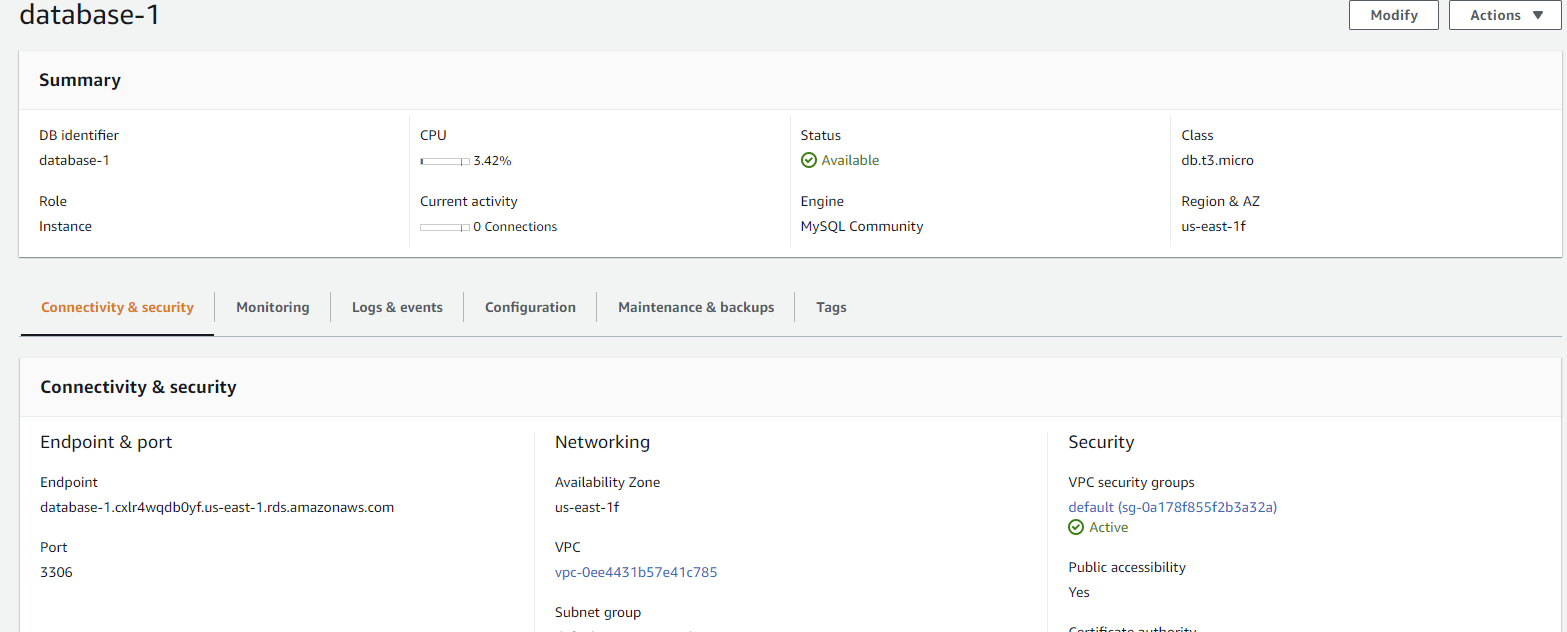
create table tripdata(VendorID int, tpep\_pickup\_datetime datetime, tpep\_dropoff\_datetime datetime, passenger\_count int,trip\_distance float,

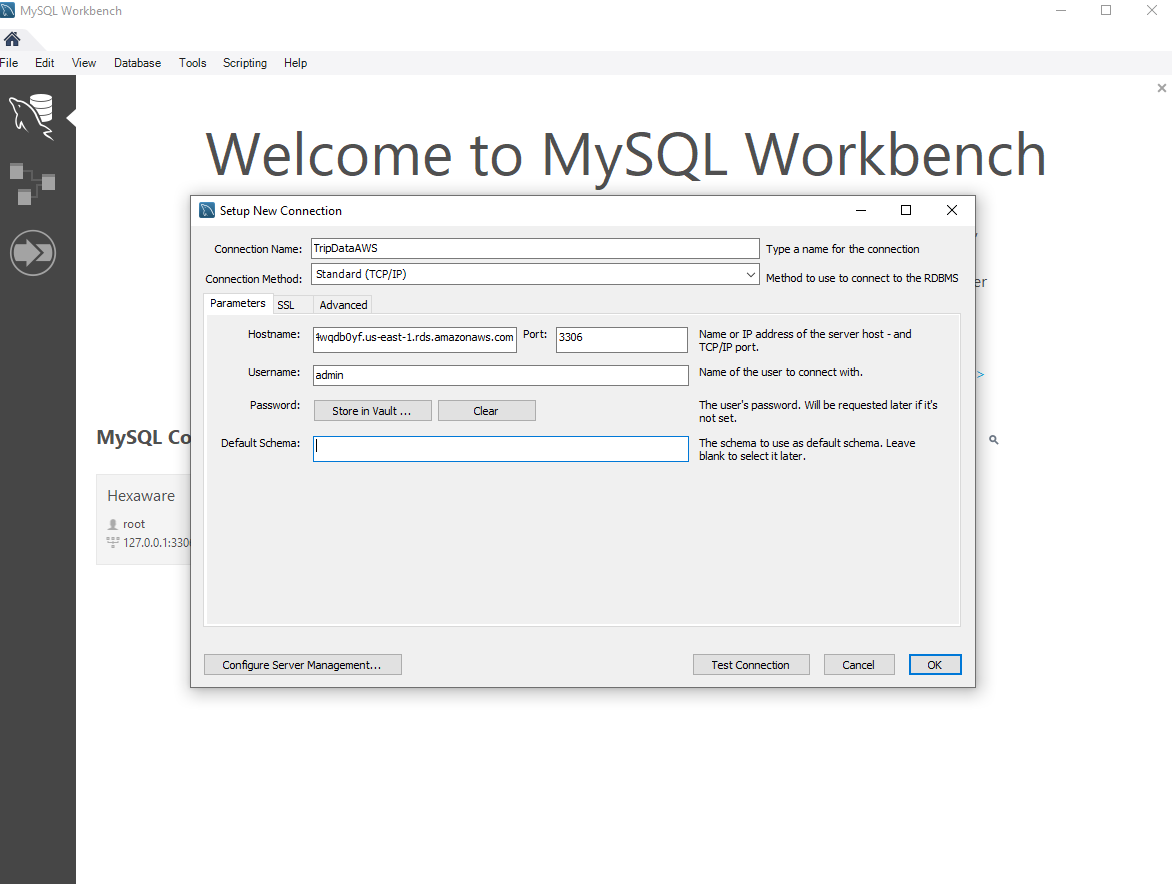
Pickup\_longitude float, pickup\_latitude float, RateCodeID int, store\_and\_fwd\_flag varchar(10), dropoff\_longitude float, dropoff\_latitude float, payment\_type int, fare\_amount float, extra float, mta\_tax float, tip\_amount float, tolls\_amount float, total\_amount float, trip\_time float)ROW FORMAT DELIMITED FIELDS TERMINATED BY "," LINES TERMINATED BY "\n";

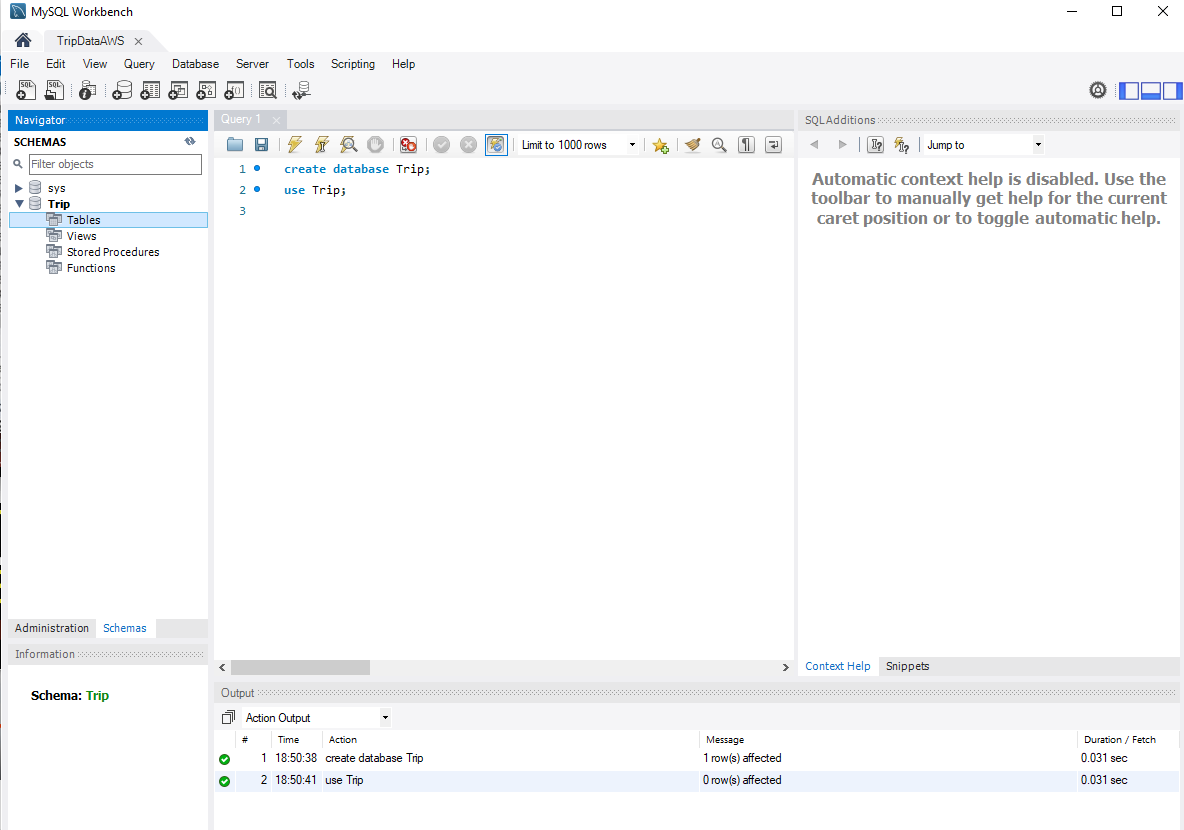
**Load data in mysql**

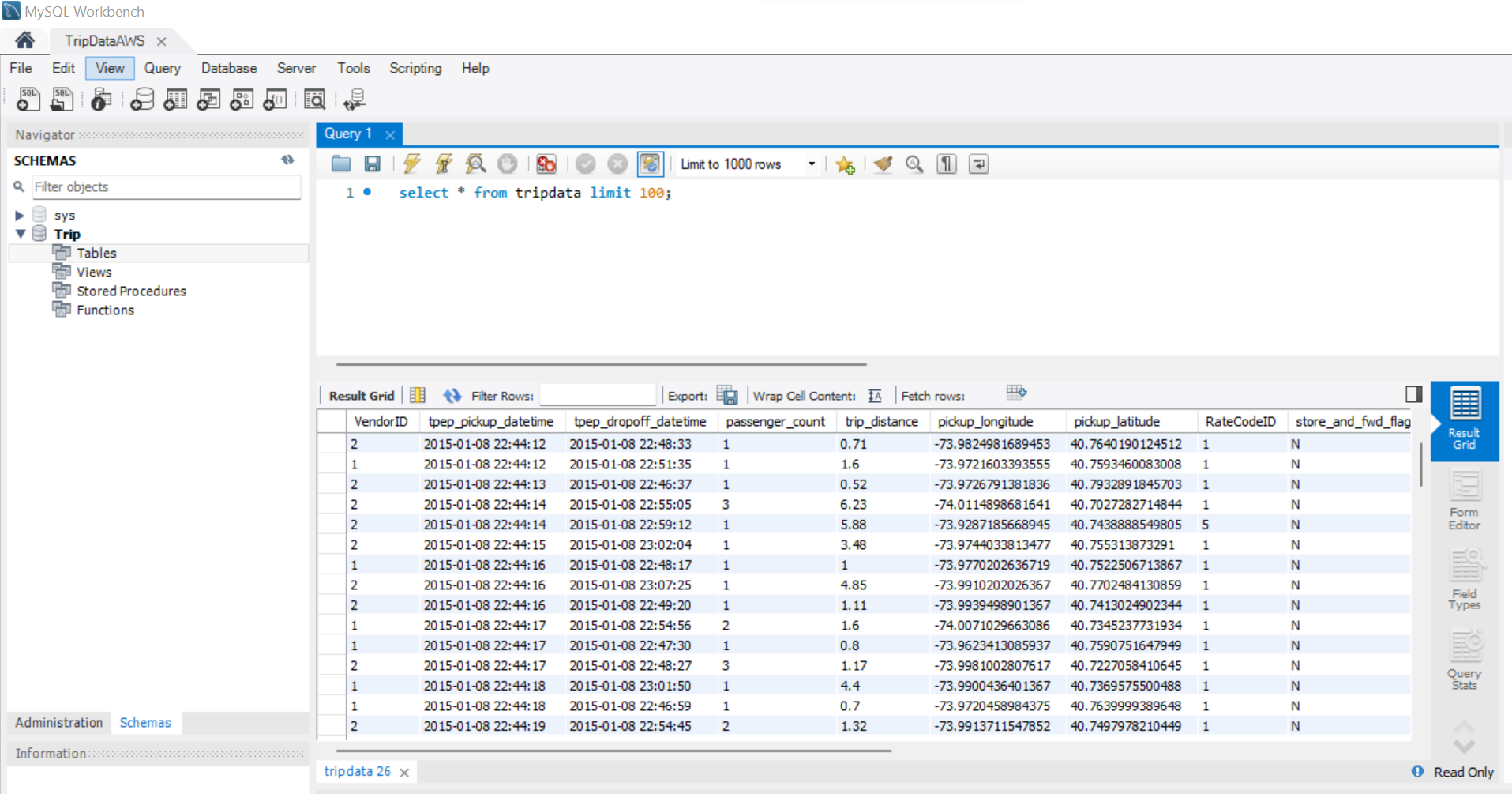
load data local infile '/root/tripdata.csv' into table tripdata FIELDS TERMINATED BY "," ENCLOSED BY "" LINES TERMINATED BY "\n" IGNORE 1 ROWS;

**RDS**



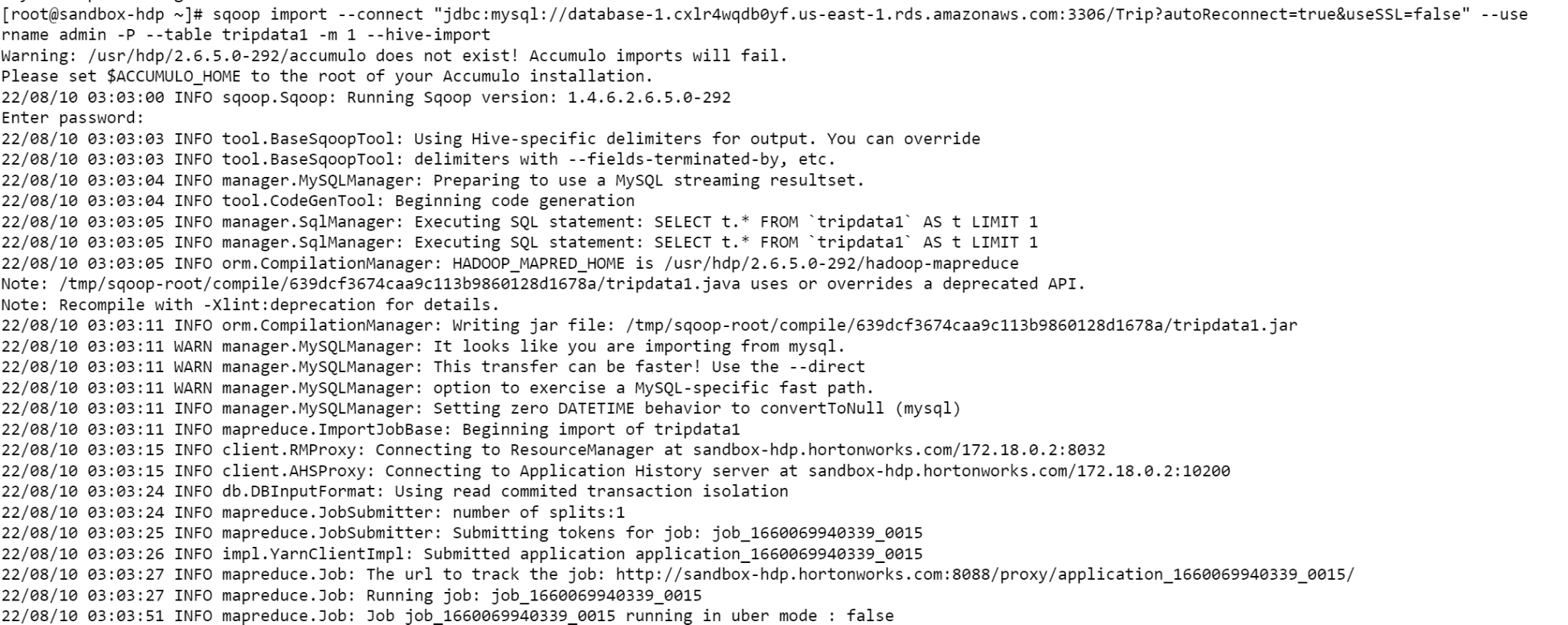






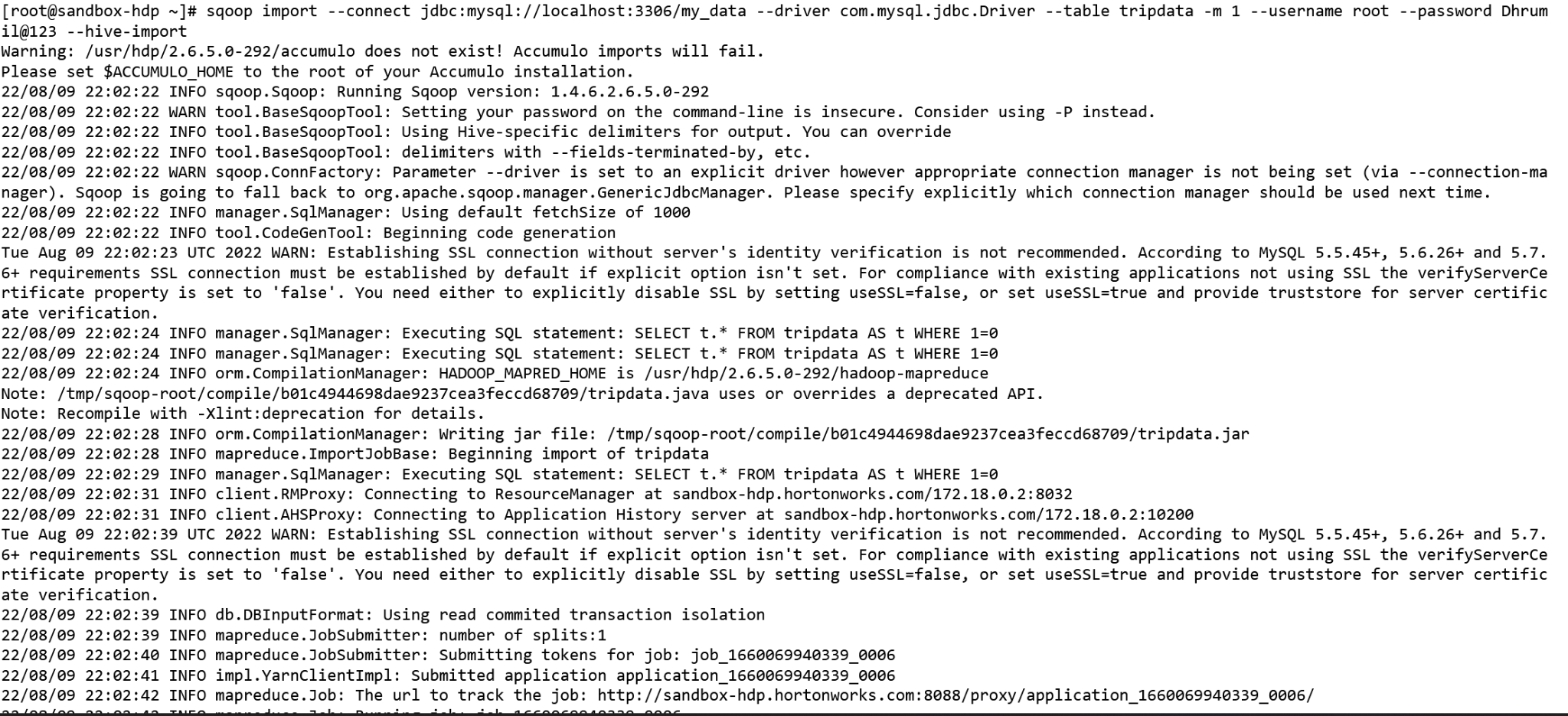
**Sqoop Command (for Aws)**

sqoop import --connect "jdbc:mysql://database-1.cxlr4wqdb0yf.us-east-1.rds.amazonaws.com:3306/Trip?autoReconnect=true&useSSL=false" --username admin -P --table tripdata1 -m 1 --hive-import

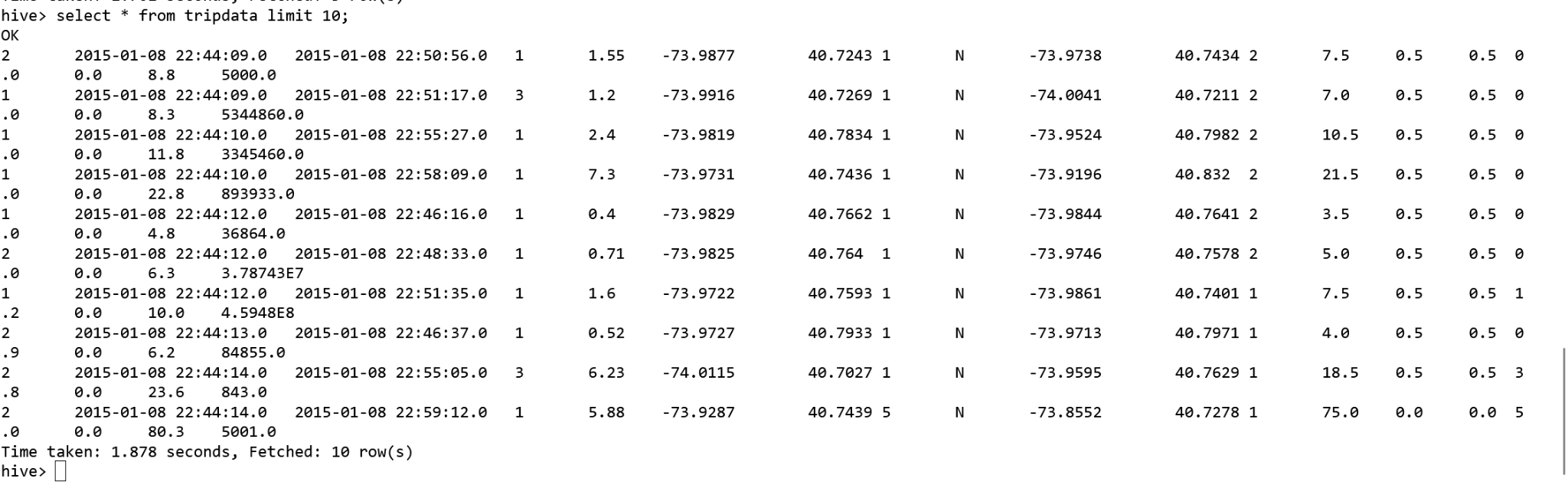


**Sqoop command (for localhost)**

sqoop import --connect jdbc:mysql://localhost:3306/my\_data --driver com.mysql.jdbc.Driver --table tripdata -m 1 --username root --password Dhrumil@123 --hive-import

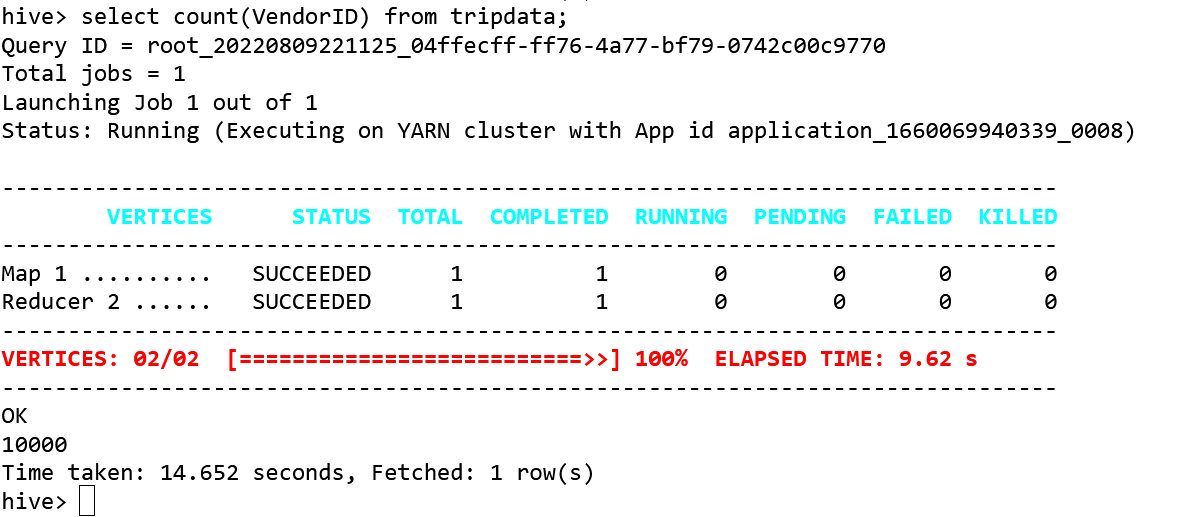


**Hive table view of the first 10 rows**



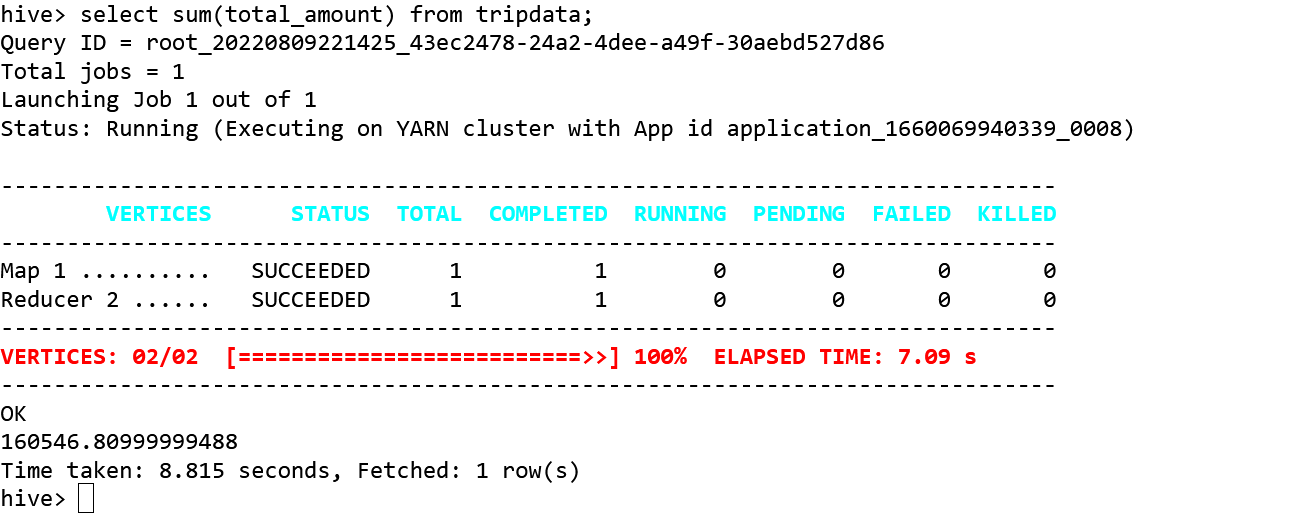
1. **What is the total number of trips (equal to the number of rows)?**

select count(VendorID) from tripdata;



**What is the total revenue generated by all the trips?**

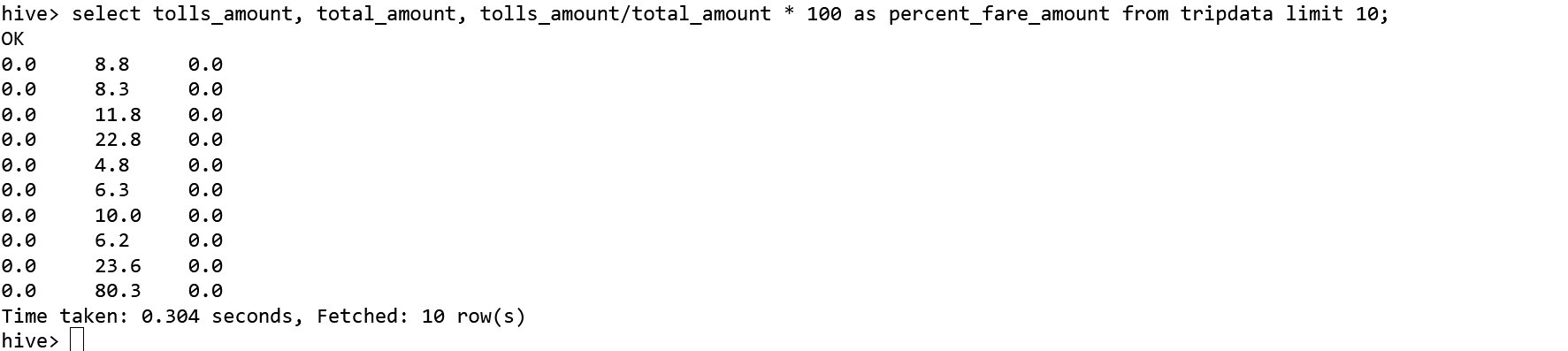
select sum(total\_amount) from tripdata;



**What fraction of the total is paid for tolls?**

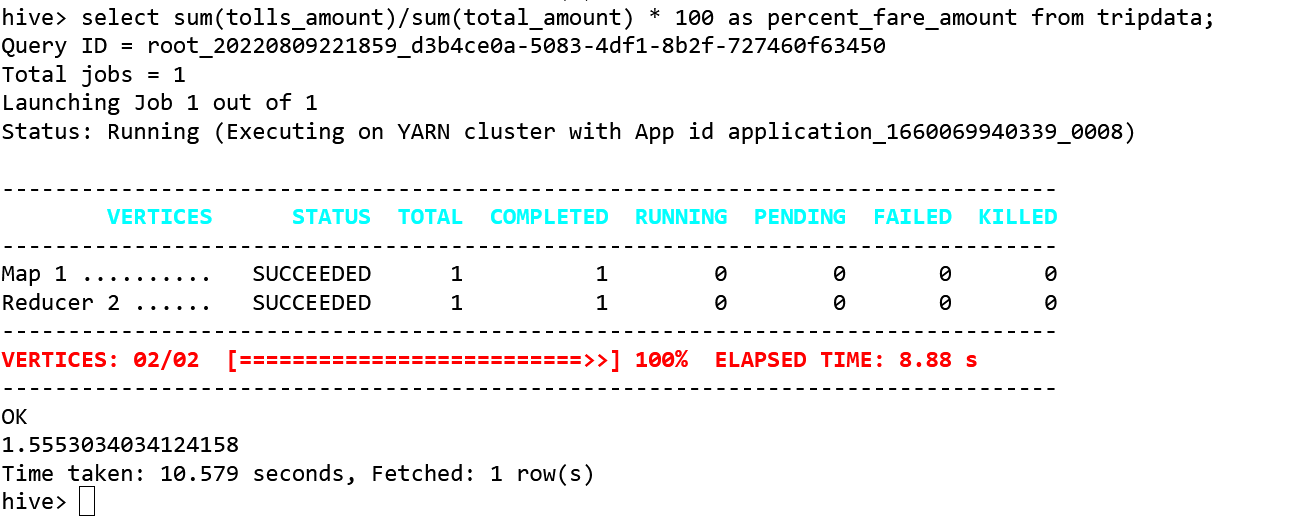
**Row-wise**

select tolls\_amount, total\_amount, tolls\_amount/total\_amount \* 100 as percent\_fare\_amount from tripdata;



**Total**

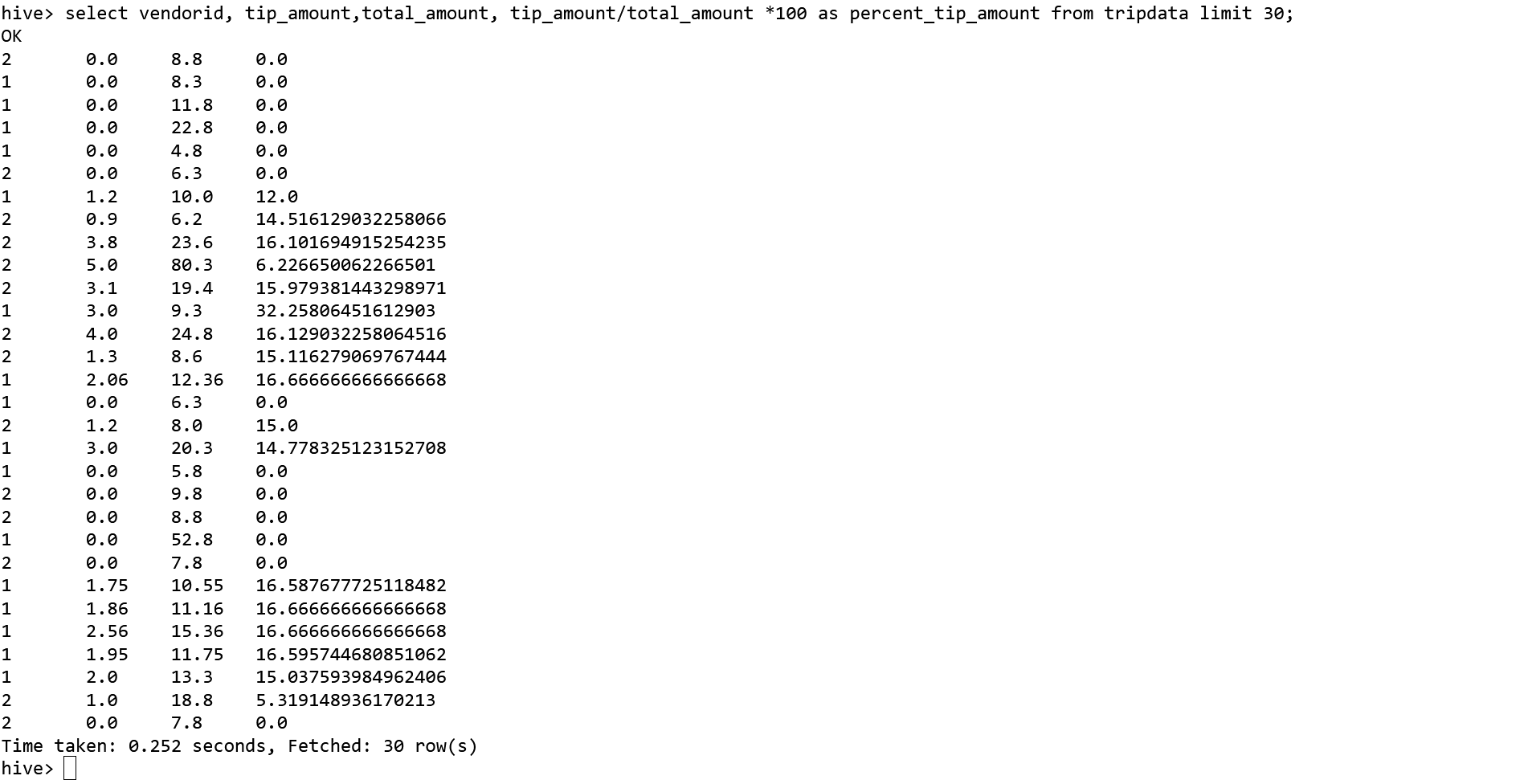
select sum(tolls\_amount)/sum(total\_amount) \* 100 as percent\_fare\_amount from tripdata;



**What fraction of it is driver tips?**

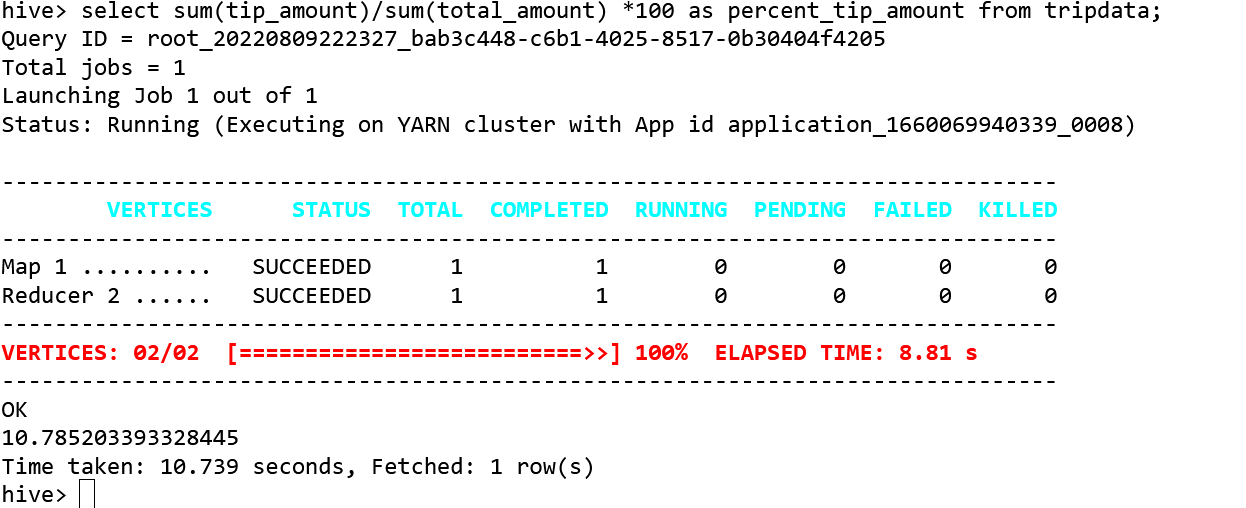
**Row-wise**

select vendorid,tip\_amount,total\_amount, tip\_amount/total\_amount \*100 as percent\_tip\_amount from tripdata;



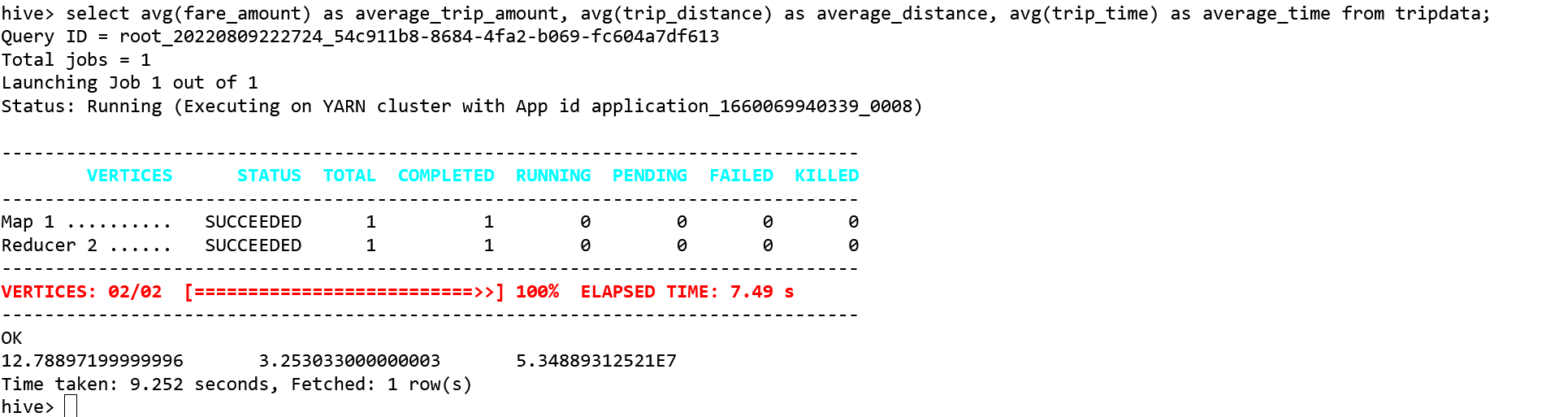
**total**

select sum(tip\_amount)/sum(total\_amount) \*100 as percent\_tip\_amount from tripdata;



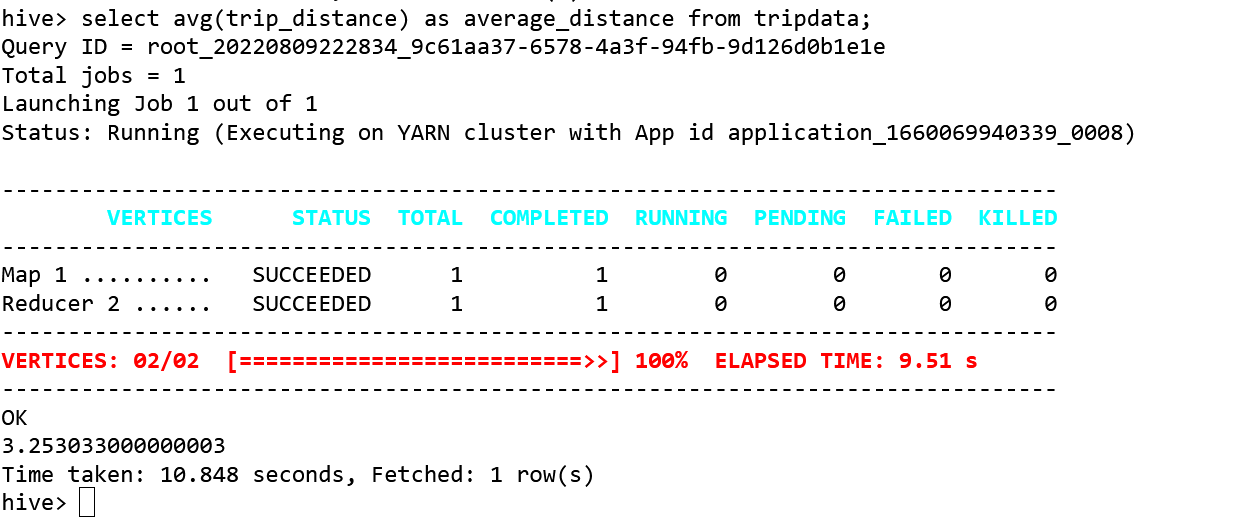
**What is the average trip amount?**

select avg(fare\_amount) as average\_trip\_amount, avg(trip\_distance) as average\_distance, avg(trip\_time) as average\_time from tripdata;



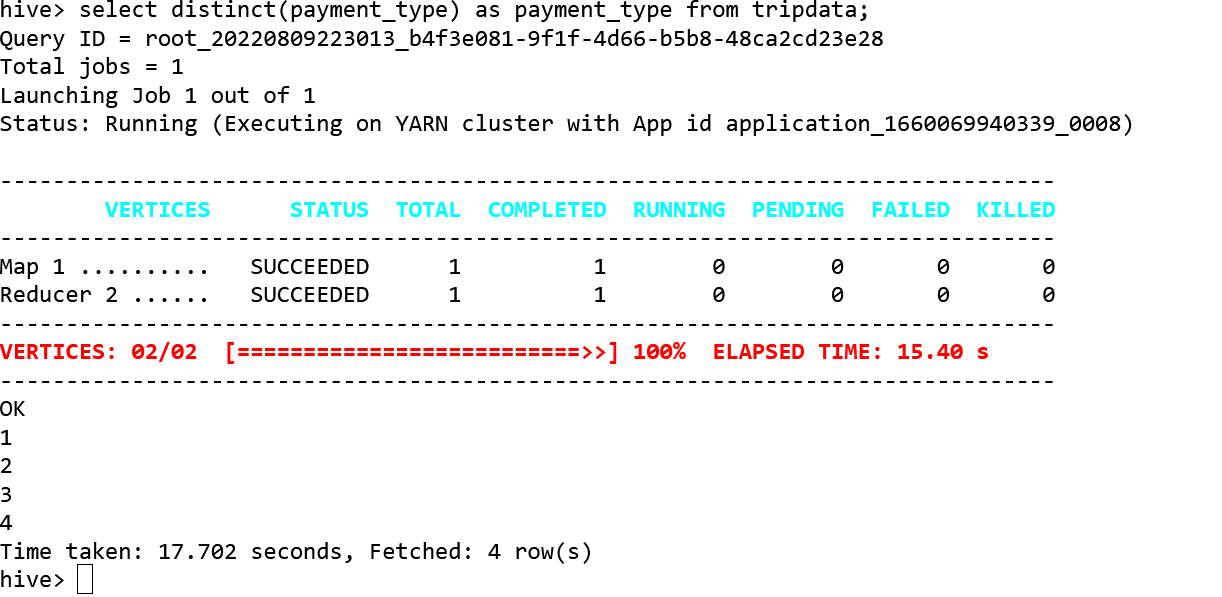
**What is the average distance of the trips?**

select avg(trip\_distance) as average\_distance from tripdata;



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

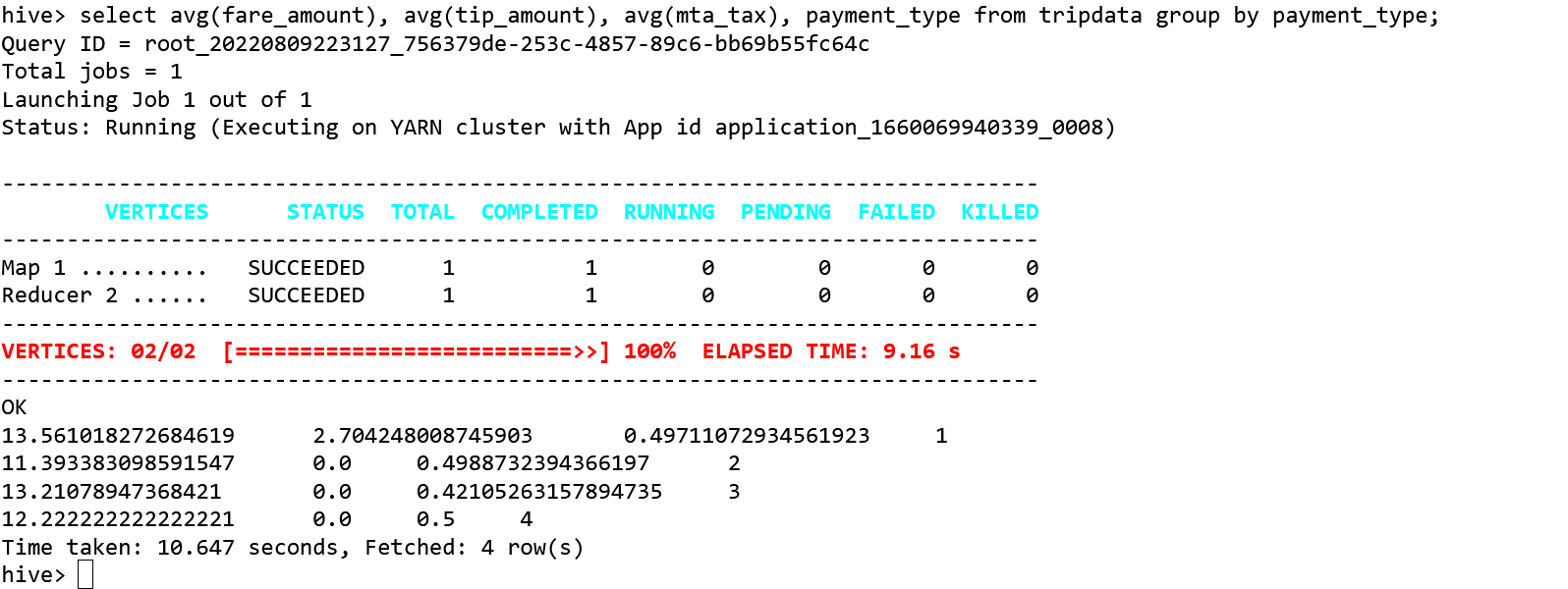
select distinct(payment\_type) as payment\_type from tripdata;



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

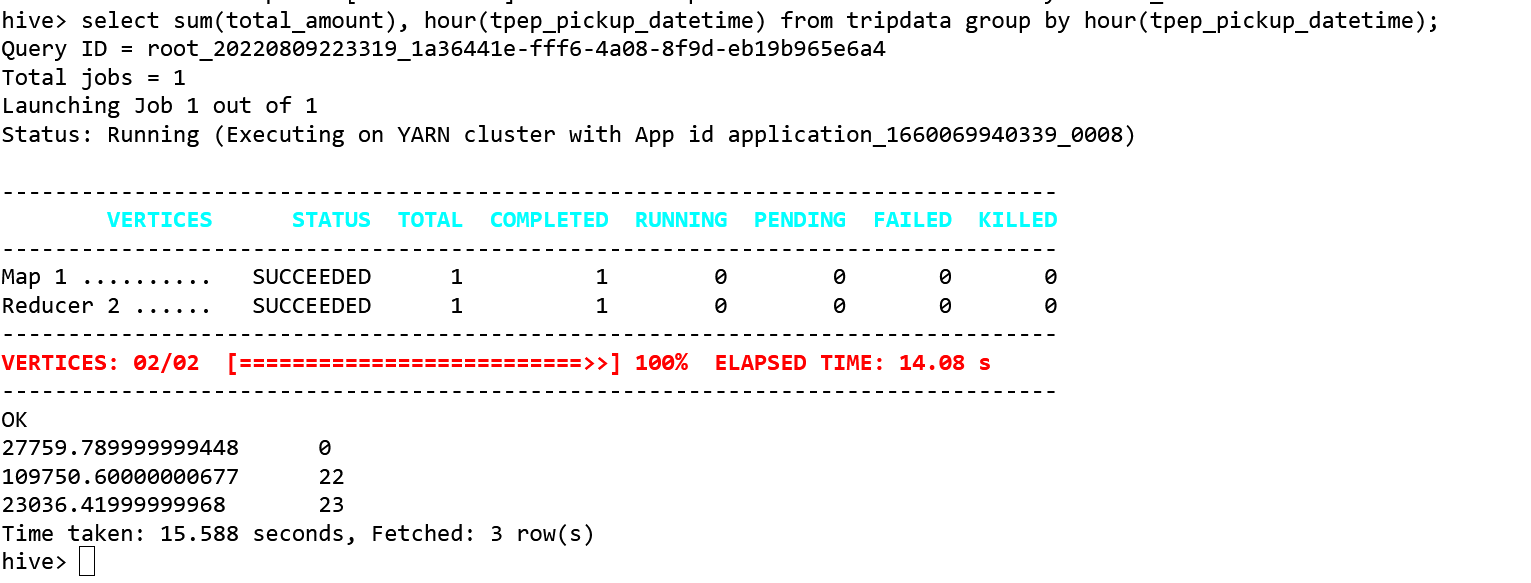
1. **Average fare generated**
2. **Average tip**
3. **Average tax**

select avg(fare\_amount), avg(tip\_amount), avg(mta\_tax), payment\_type from tripdata group by payment\_type;



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

select sum(total\_amount), hour(tpep\_pickup\_datetime) from tripdata group by hour(tpep\_pickup\_datetime);



ct avg(fare\_amount), avg(tip\_amount), avg(mta\_tax), payment\_type from tripdata group by payment\_type; the trips