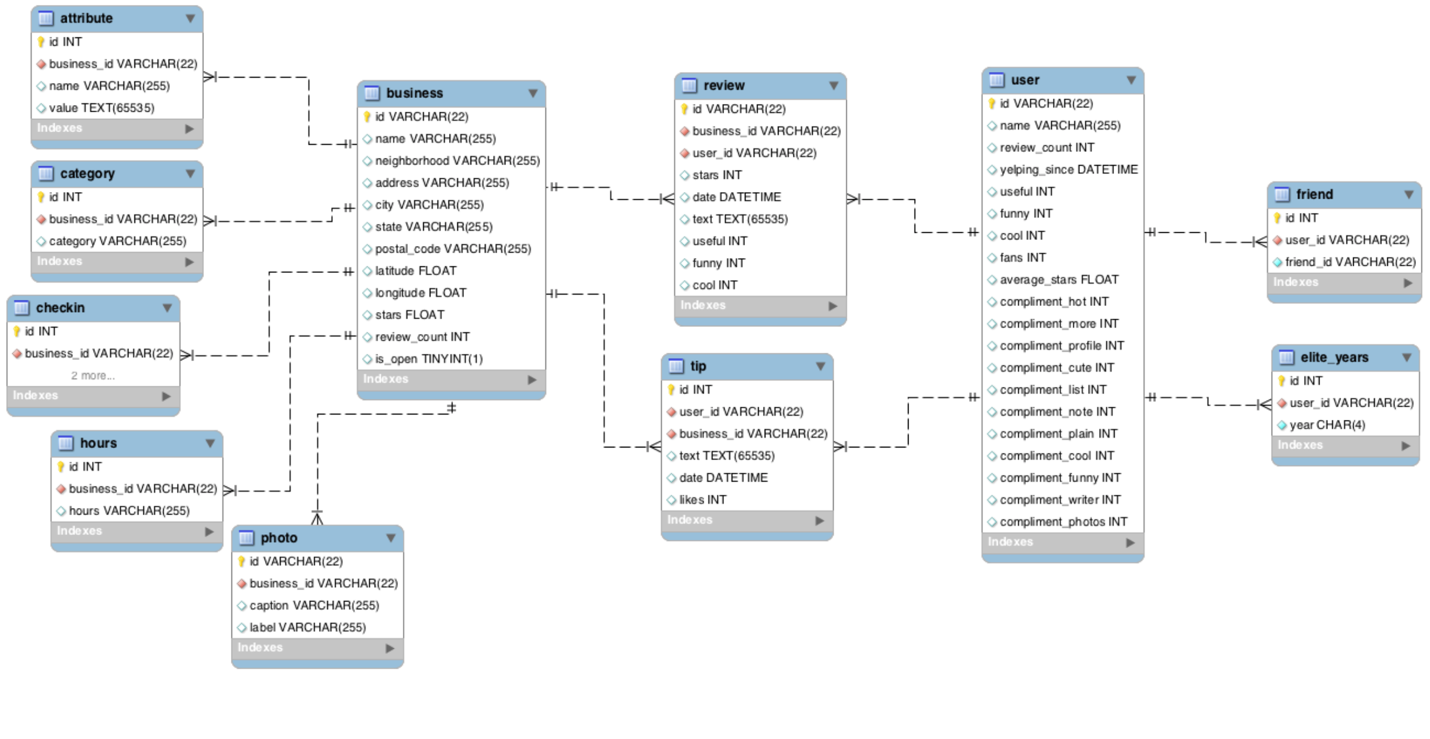
Task1: which day and which hour of the week a business will be busy

I have downloaded the dataset from yelp. It has schema as follows



It contains an yelp\_db with tables like:

1)attribute

2)business

3)category

4)checkin

5)elite years

6)friend

7)hours

8)photo

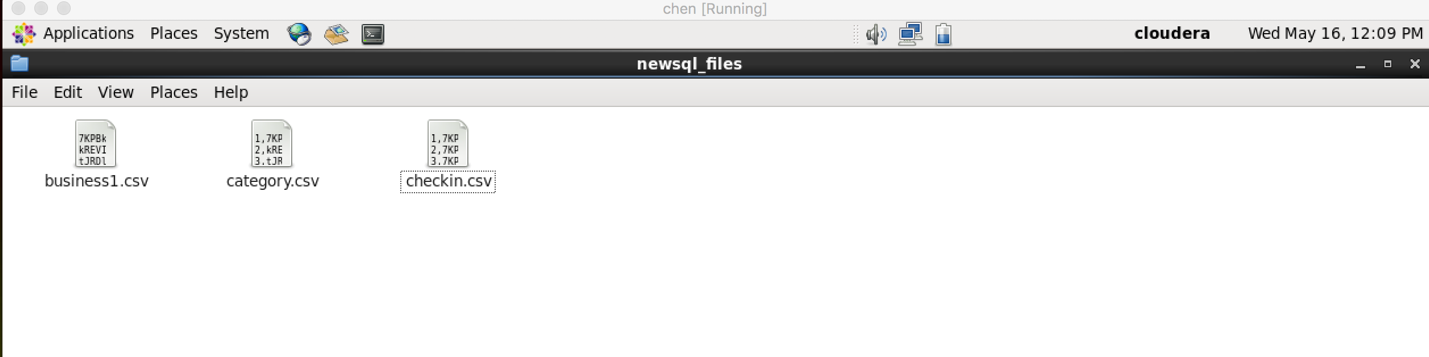
9)review

10)user

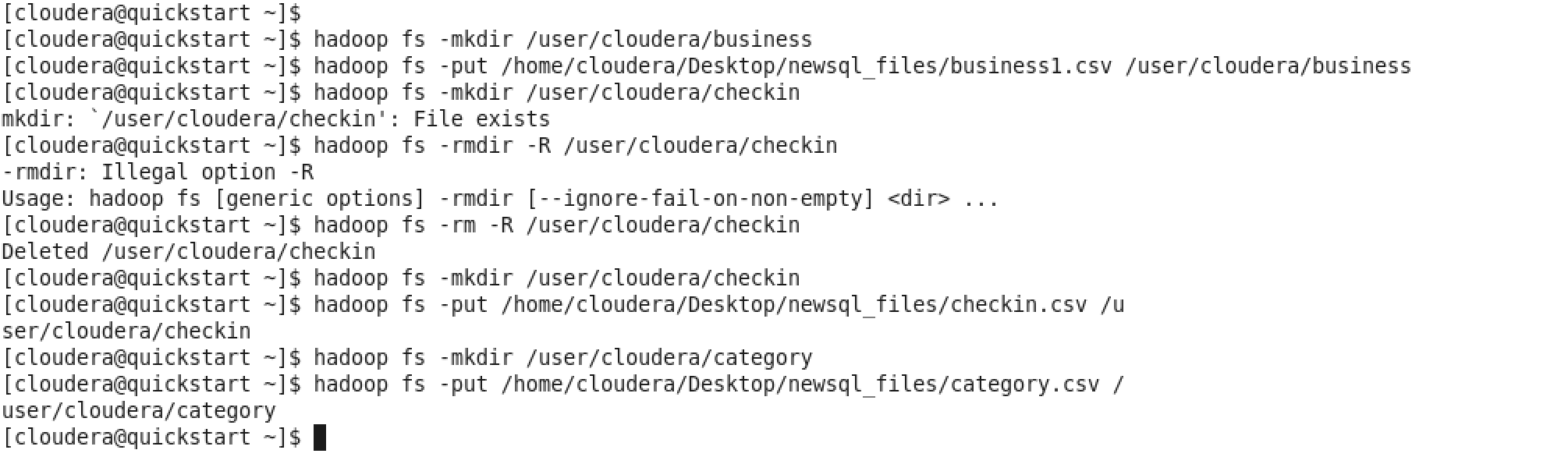
11)tip

LOADED the data into mysql database;

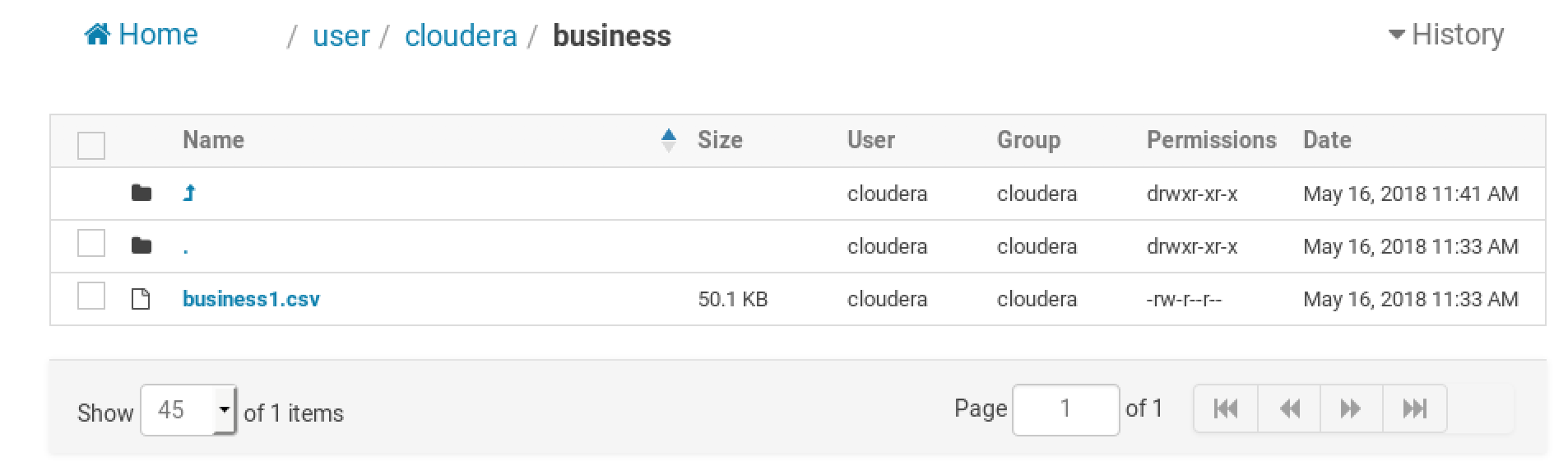
Generated the .csv files. – business.csv, checkin.csv and category.csv

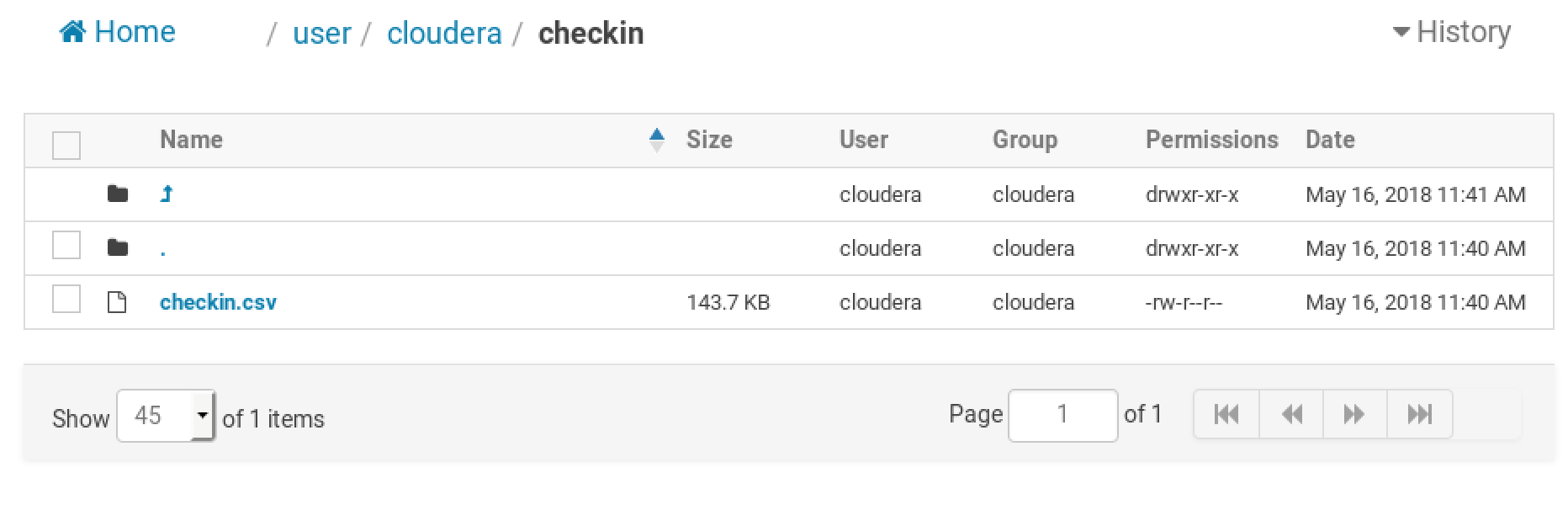


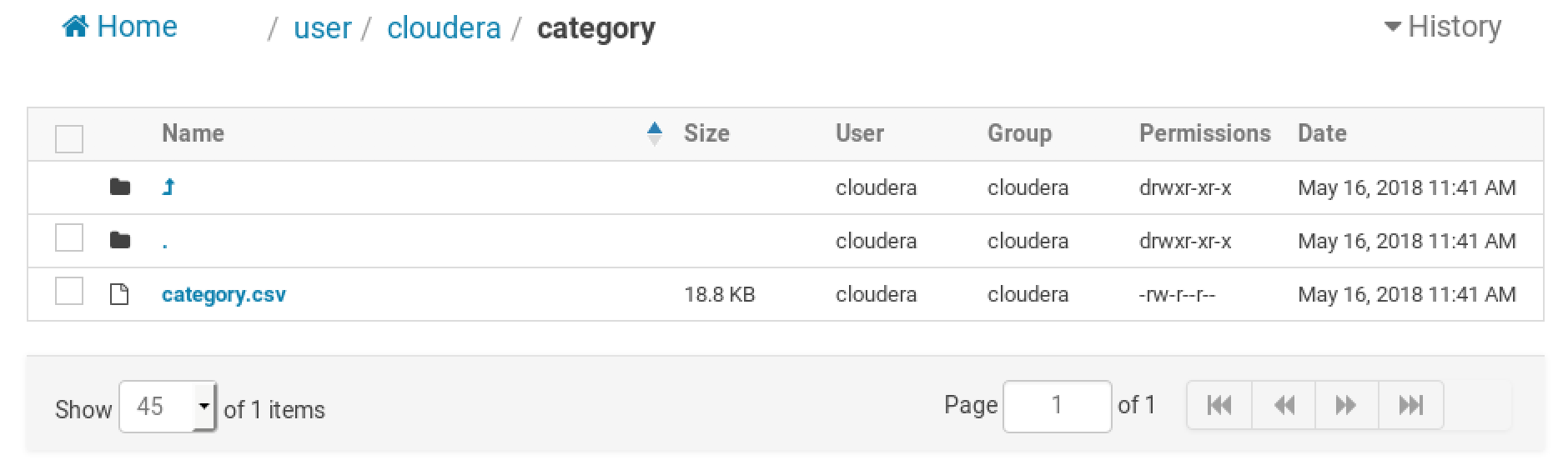
Loaded the 3 files into CLOUDERA virtual machine



Loaded the files into HDFS(Hadoop distributed file system)

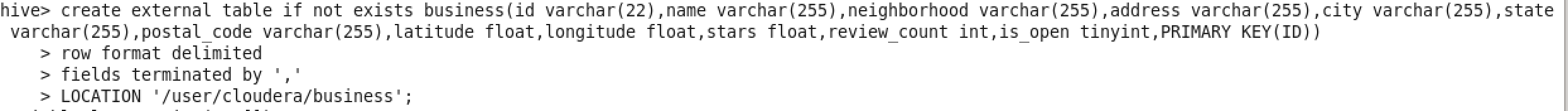




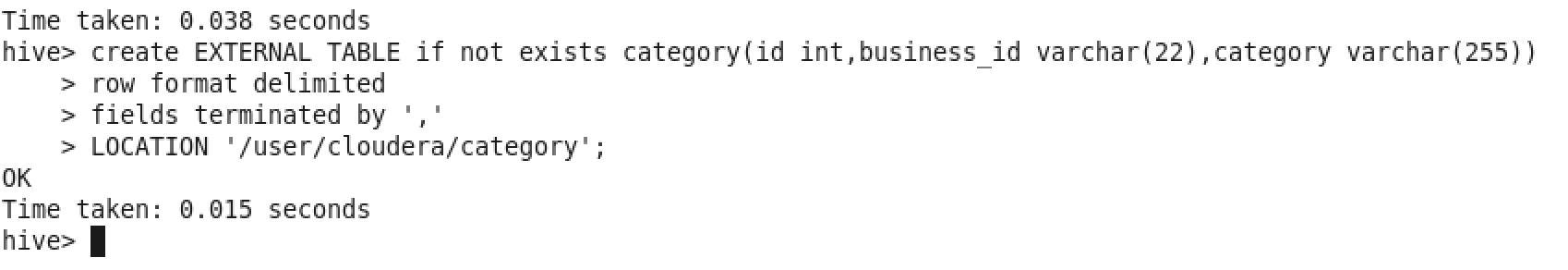


Created hive external tables

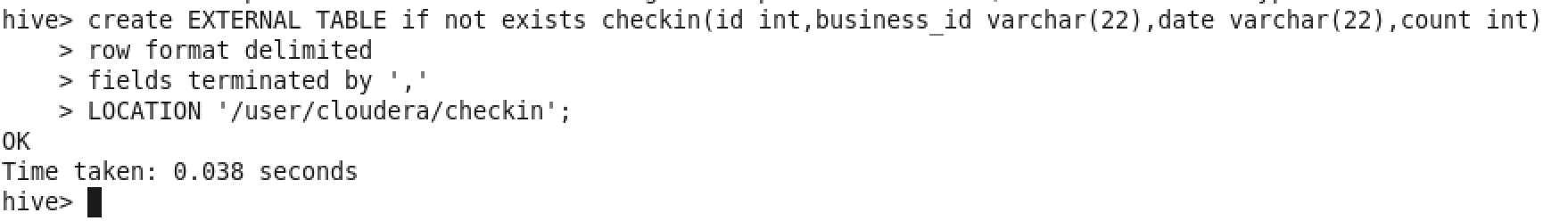
Business



category



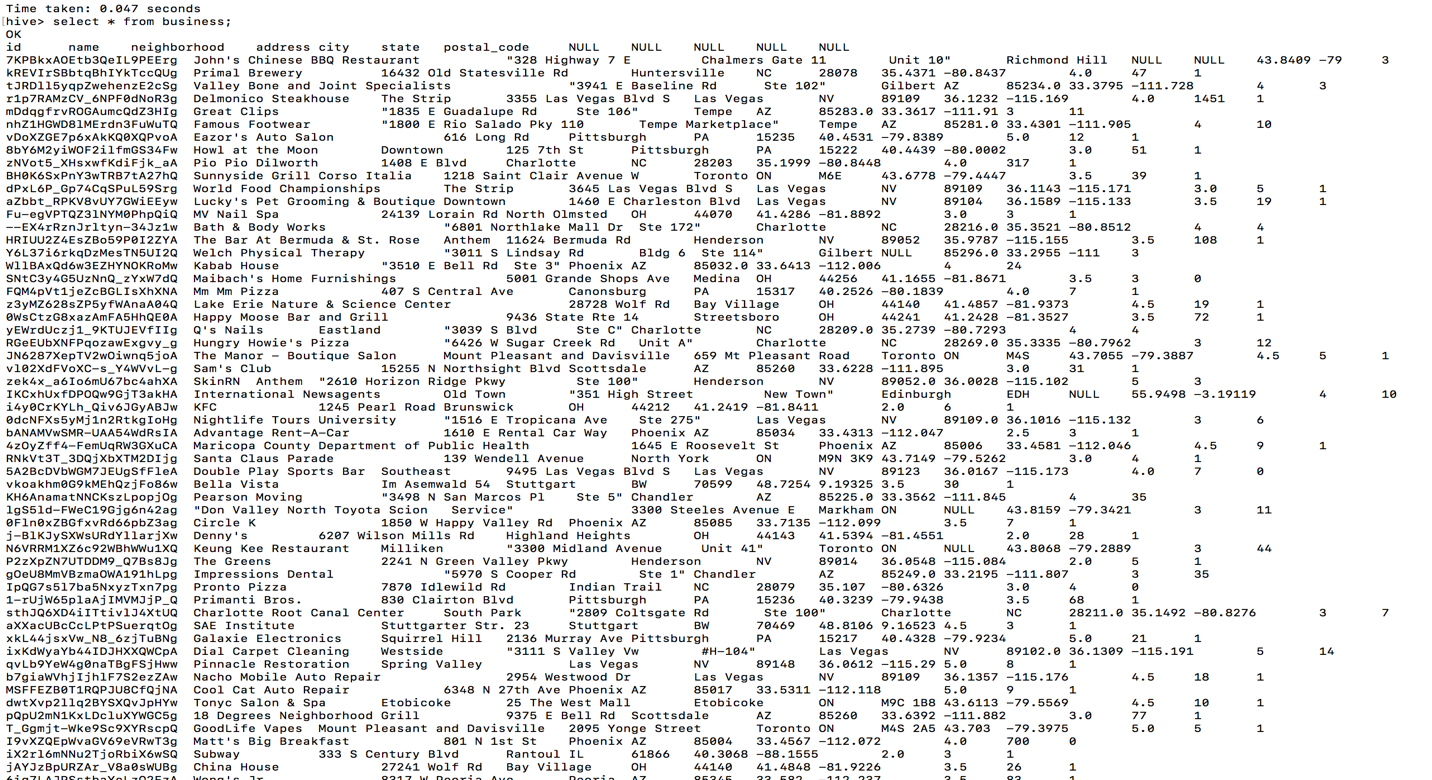
Checkin table



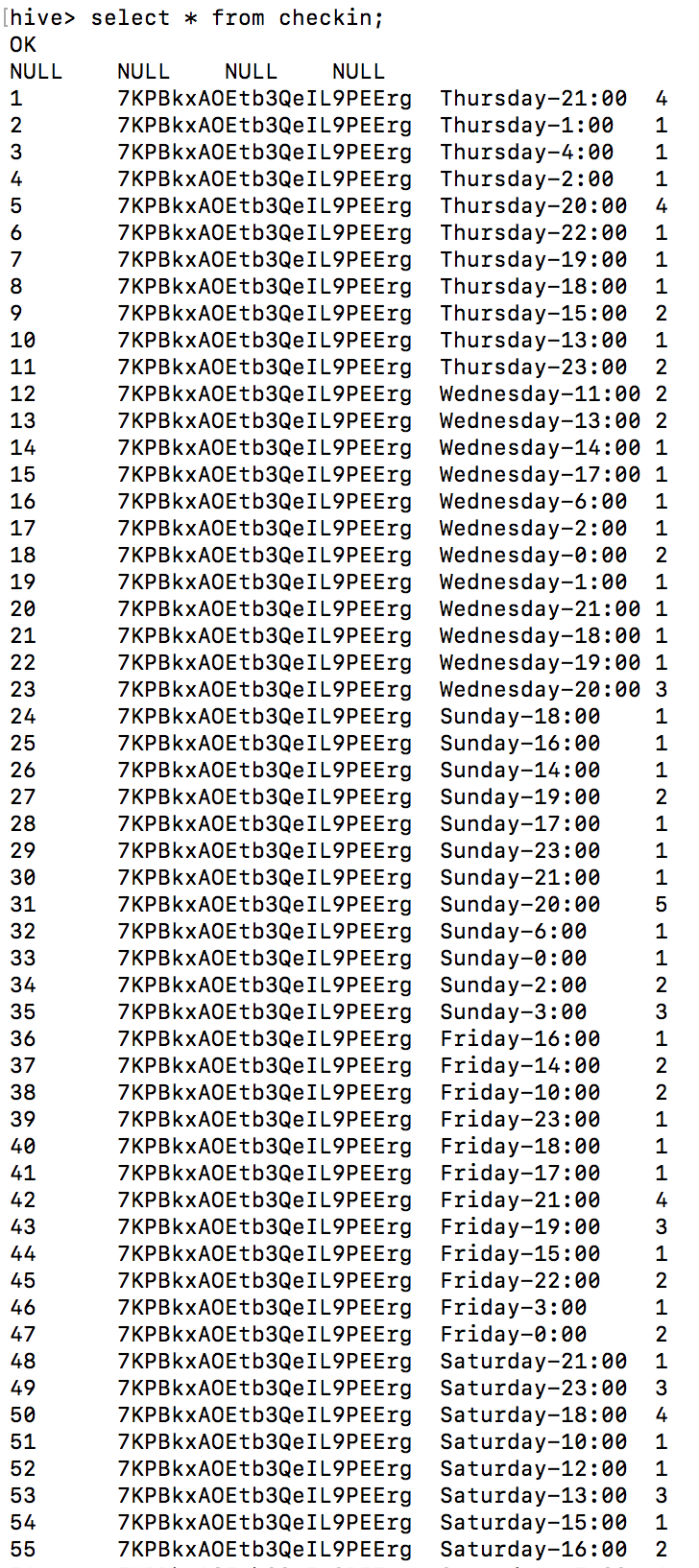
The data will be automatically loaded from hdfs to the hive tables as they are external tables

Now the data is displayed as follows

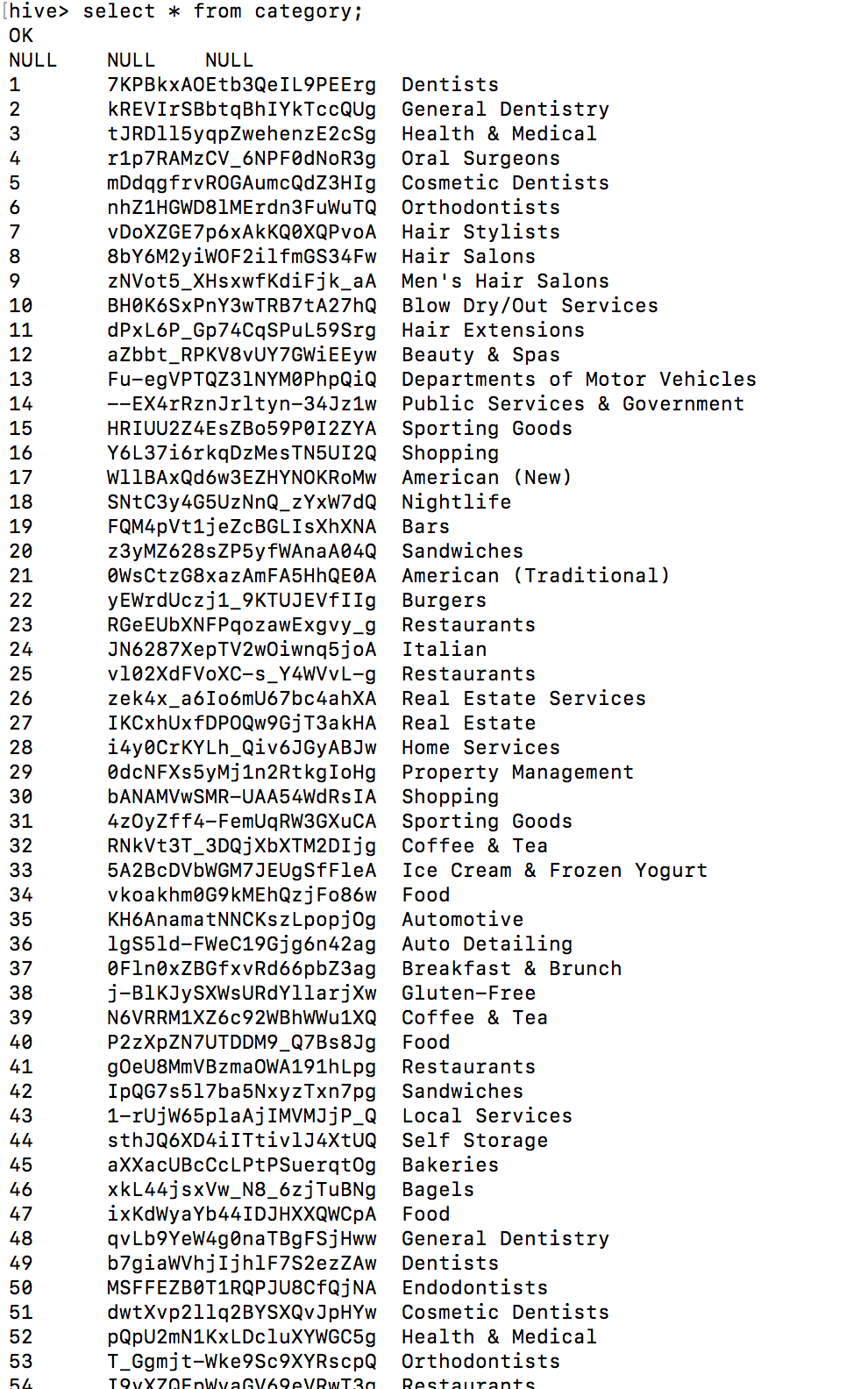
Business table:



Checkin table:



Category table:



The checkin part contains id, business\_id, date, count.

According to first task to know which day and which hours of a week a business is busy. I have written task1.script

Which contains the following:

1. Create view as select c.date,c.count,b.name,b.id from checkin c LEFT JOIN business b ON c.business\_id = b.id ORDER BY c.business\_id;
2. Select max(count),date,name,id from day group by id order by count desc

OUTPUT:

