Commands:

wh\_data = load 'comqa\_train.txt' using PigStorage('\n') as (text:chararray);

questions = filter wh\_data by (text matches '(.\*(who|what|where|when|how|why|how|whom|which|whose)+.\*)+.\*\\?.\*');

words = foreach questions generate flatten(TOKENIZE(text)) as word;

all\_wh\_words = filter words by word matches '(who|what|where|when|how|why|how|whom|which|whose).\*';

words\_without\_extras = foreach all\_wh\_words generate REPLACE(word ,'(\\?|\'s)','') as word;

pure\_wh\_words = filter words\_without\_extras by word matches '(who|what|where|when|how|why|how|whom|which|whose)';

wh\_words = group pure\_wh\_words by word;

finalout = foreach wh\_words generate group, COUNT(pure\_wh\_words);

dump finalout;

Output:

(how, 20)

(who, 1124)

(why, 1)

(what, 2433)

(when, 643)

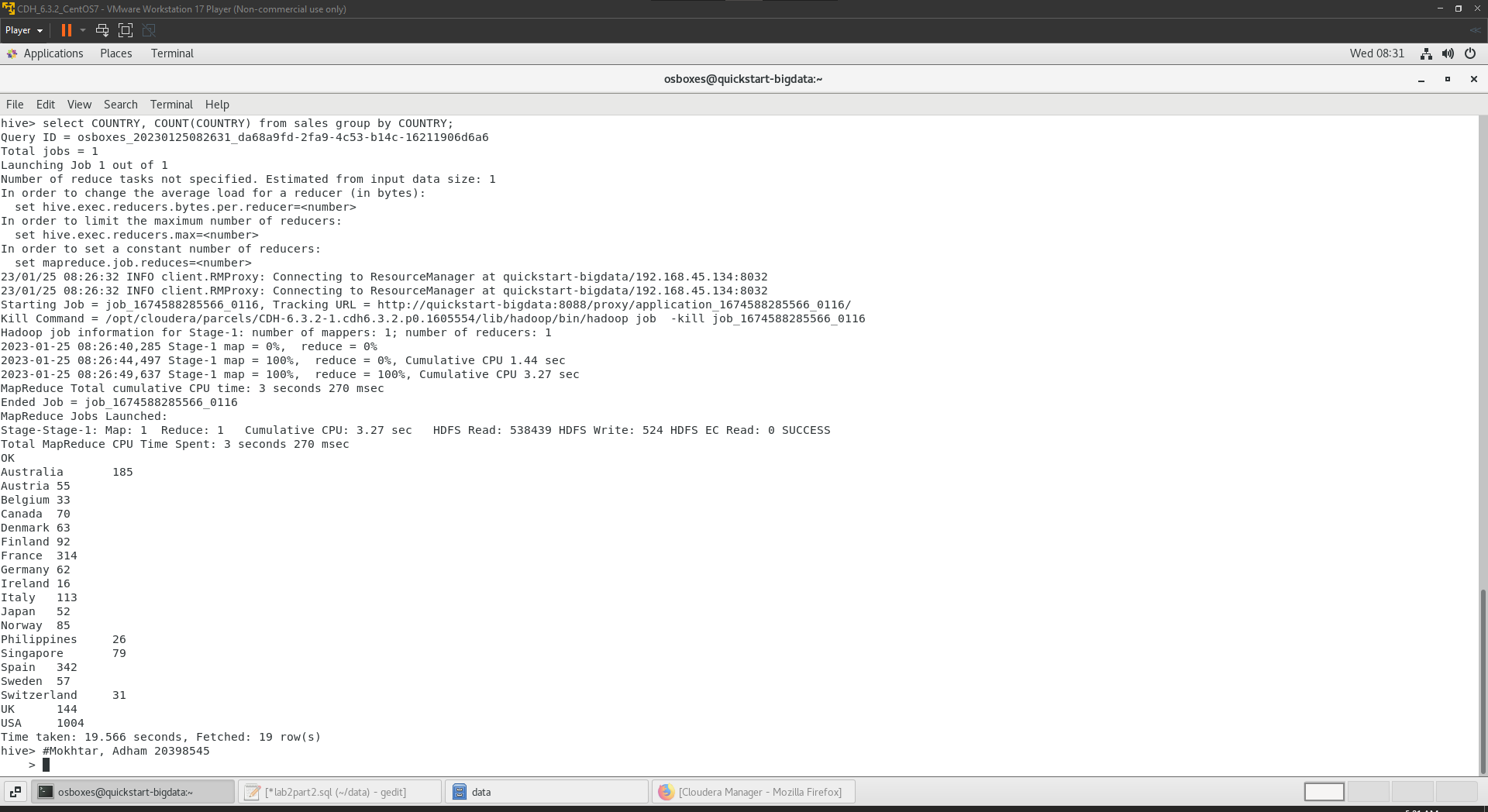
(whom, 5)

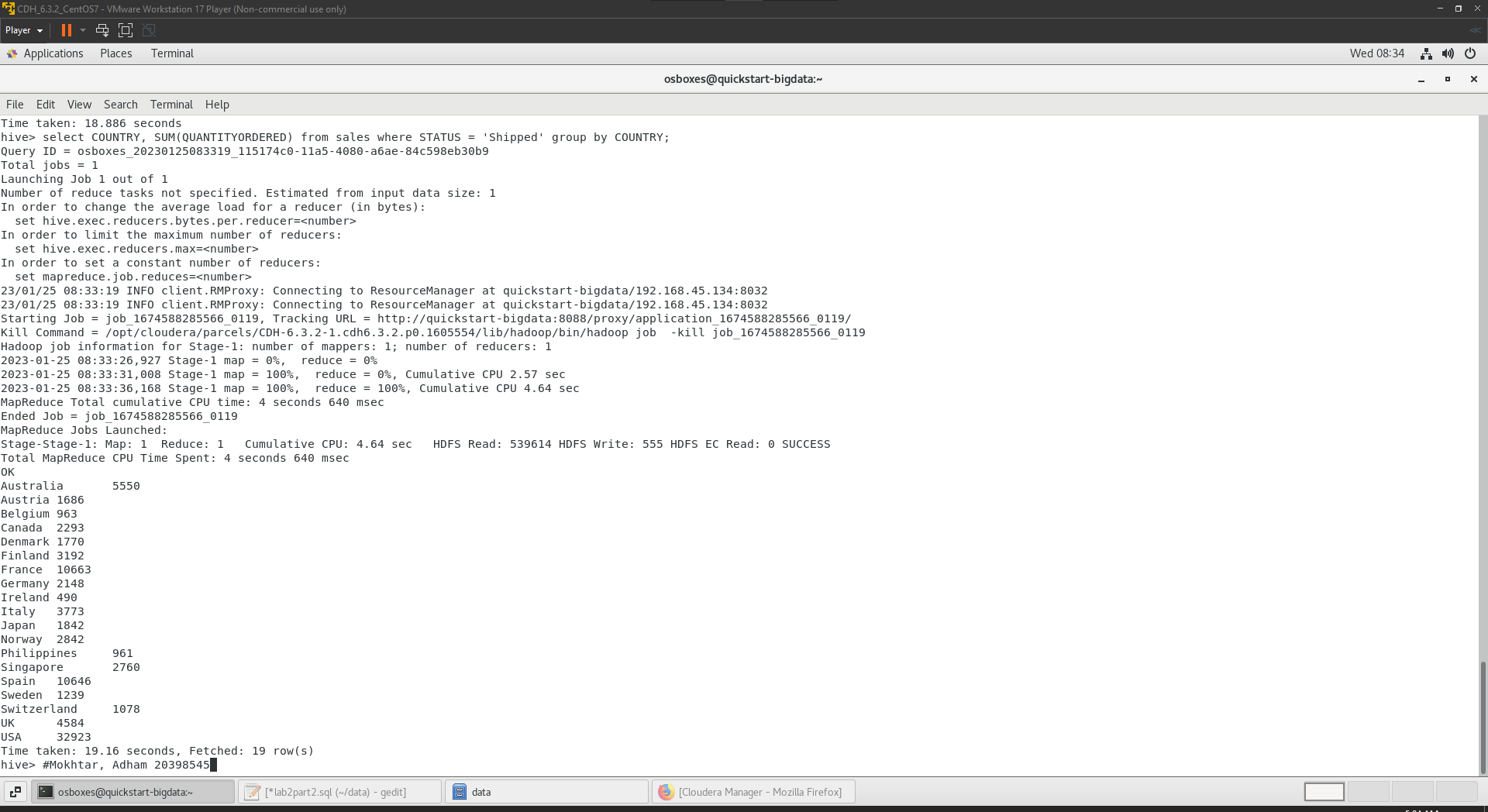
(where, 234)

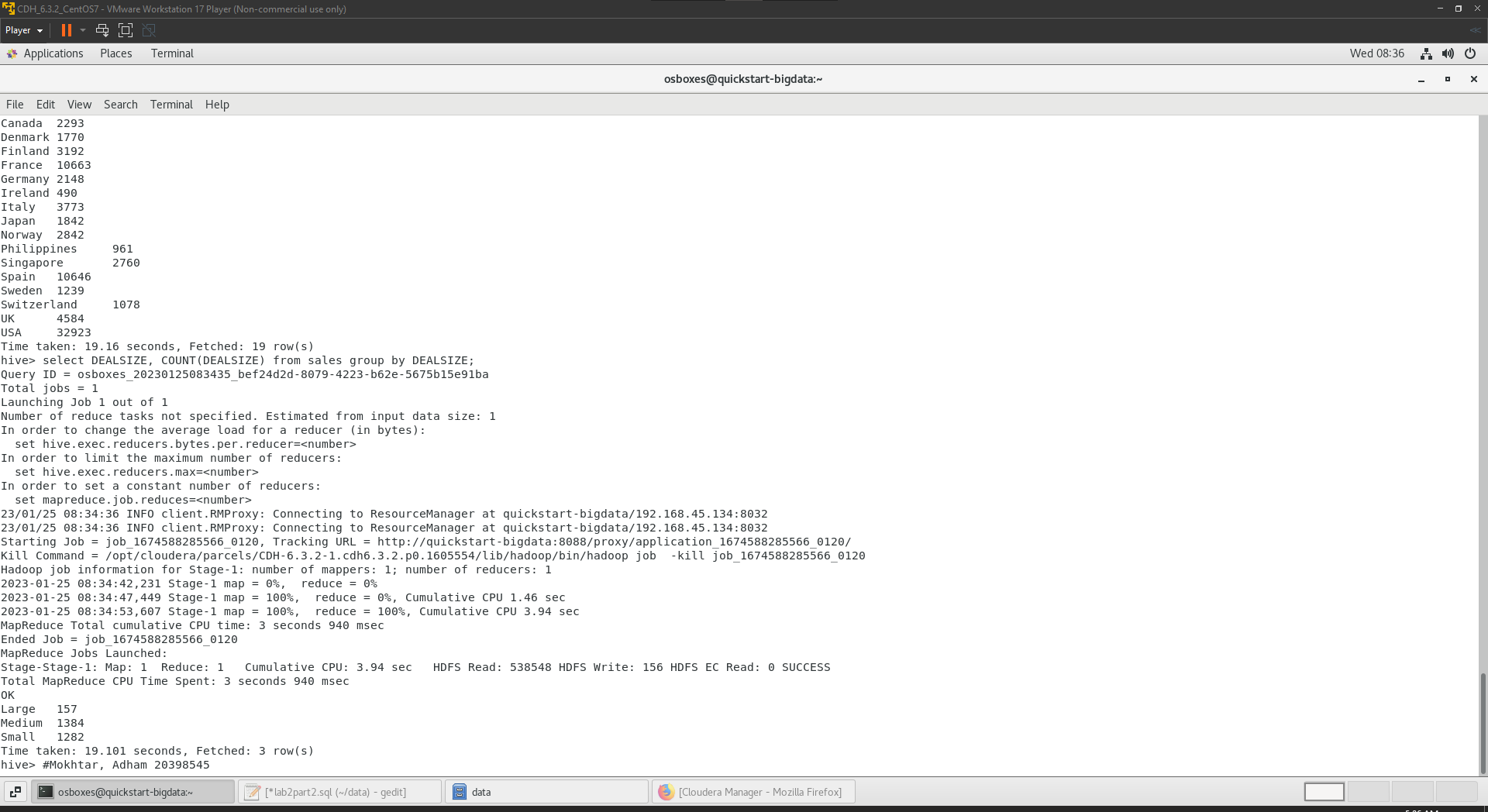
(which, 389)

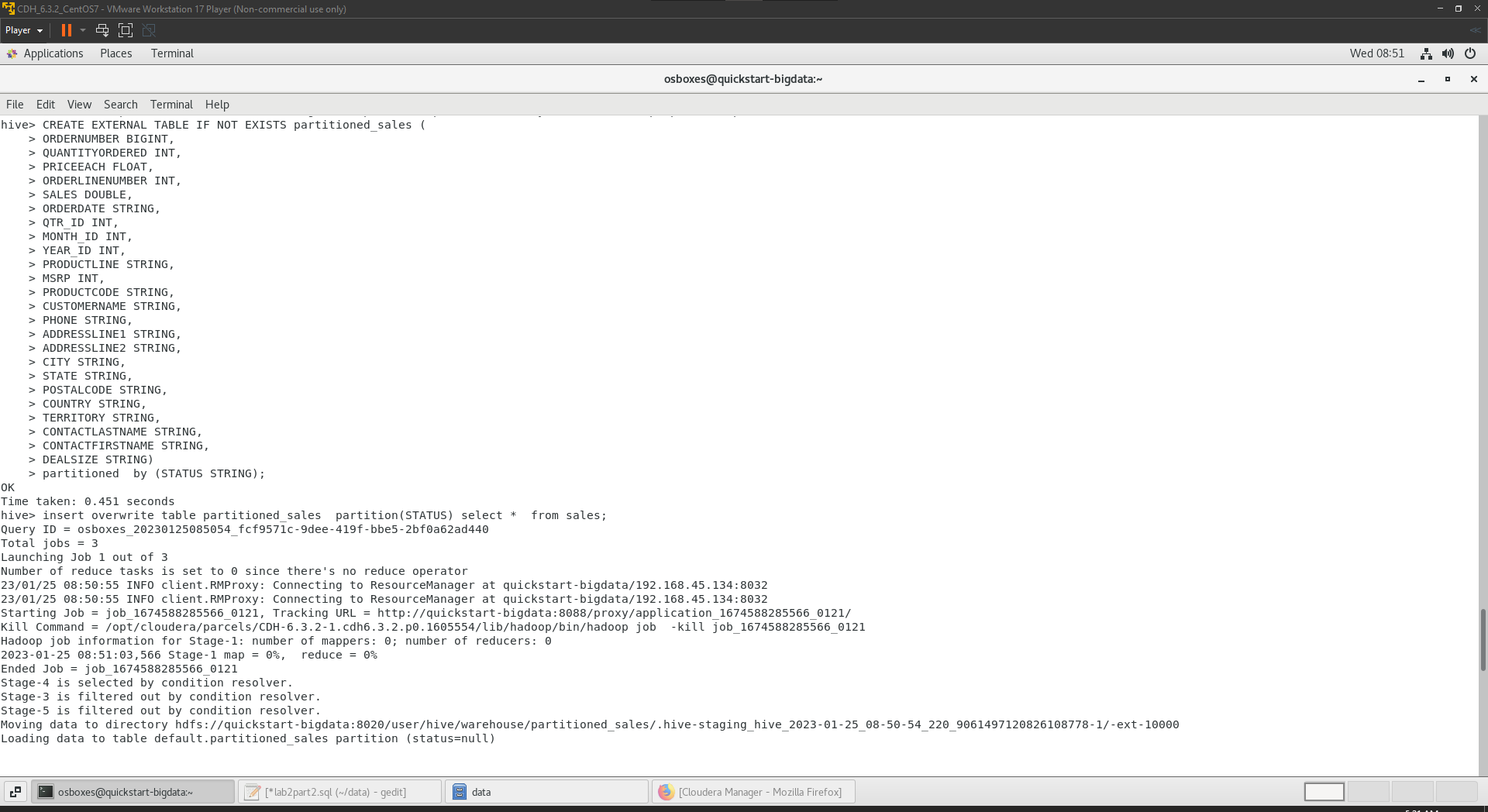
(whose, 6)

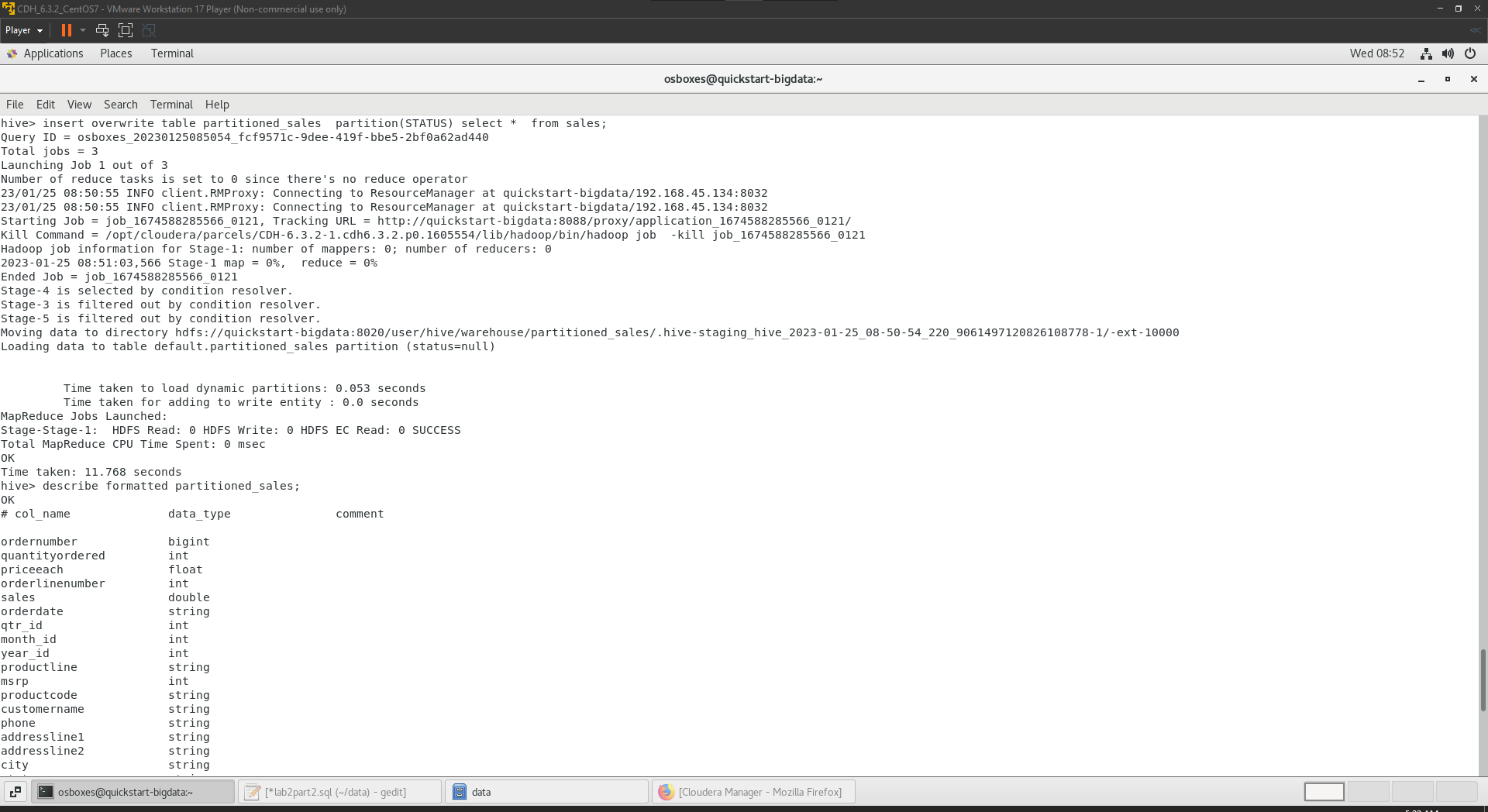
Part B:

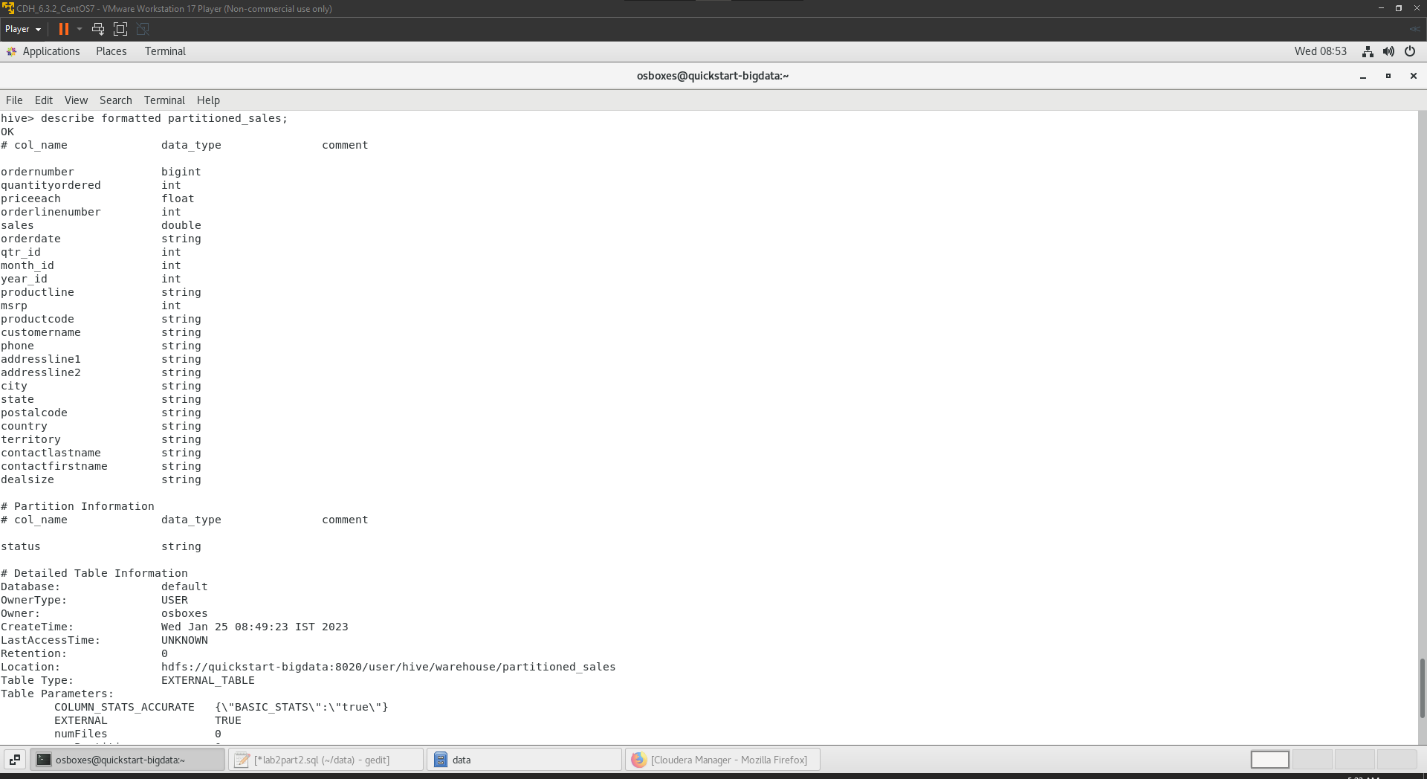




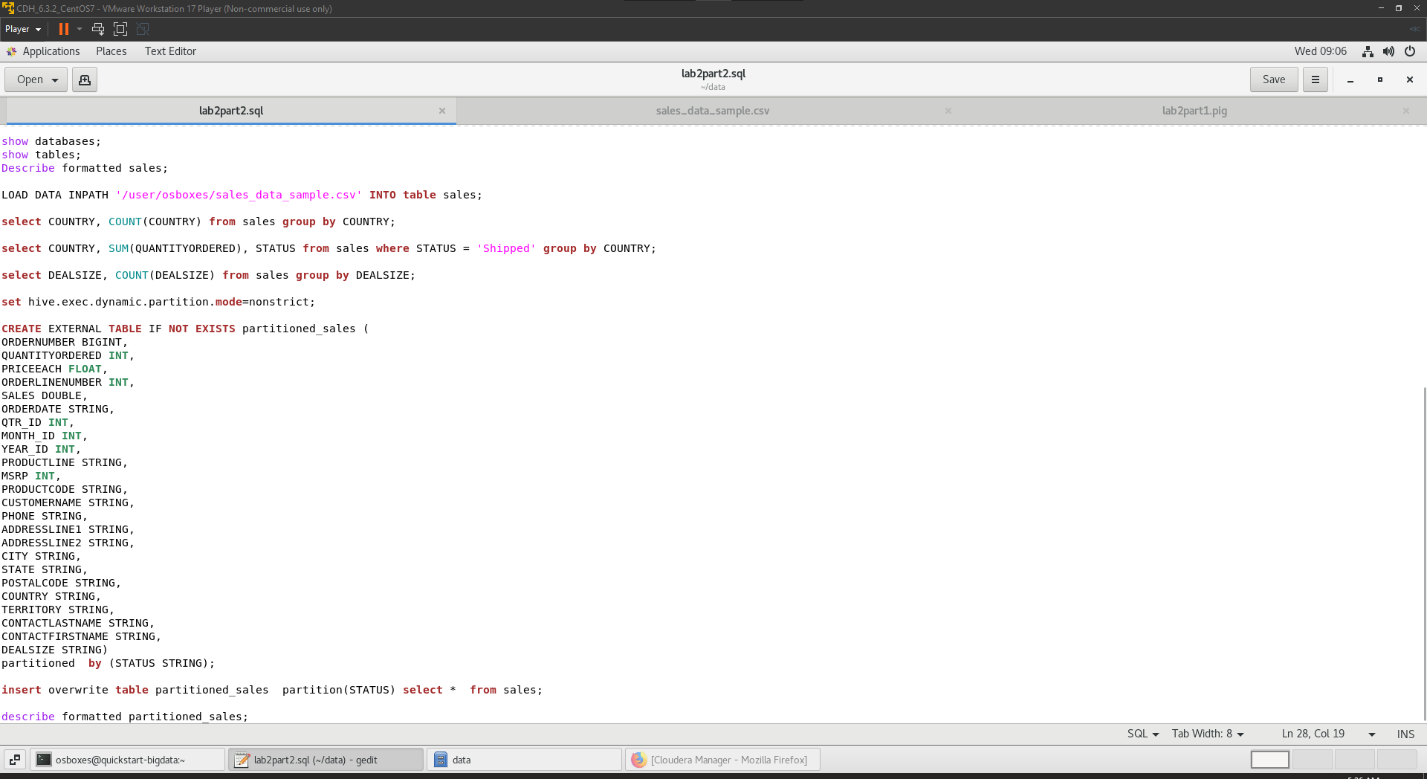


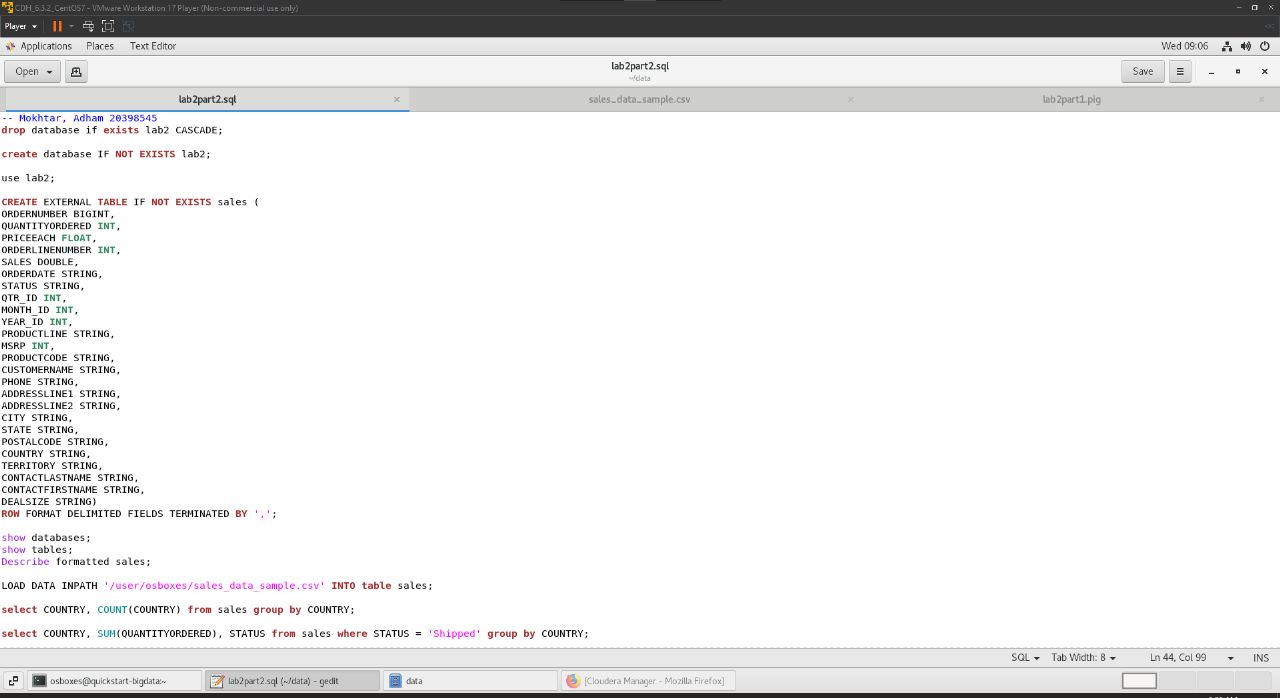






Code:





Commands:

drop database if exists lab2 CASCADE;

create database IF NOT EXISTS lab2;

use lab2;

CREATE EXTERNAL TABLE IF NOT EXISTS sales (

ORDERNUMBER BIGINT,

QUANTITYORDERED INT,

PRICEEACH FLOAT,

ORDERLINENUMBER INT,

SALES DOUBLE,

ORDERDATE STRING,

STATUS STRING,

QTR\_ID INT,

MONTH\_ID INT,

YEAR\_ID INT,

PRODUCTLINE STRING,

MSRP INT,

PRODUCTCODE STRING,

CUSTOMERNAME STRING,

PHONE STRING,

ADDRESSLINE1 STRING,

ADDRESSLINE2 STRING,

CITY STRING,

STATE STRING,

POSTALCODE STRING,

COUNTRY STRING,

TERRITORY STRING,

CONTACTLASTNAME STRING,

CONTACTFIRSTNAME STRING,

DEALSIZE STRING)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';

show databases;

show tables;

Describe formatted sales;

LOAD DATA INPATH '/user/osboxes/sales\_data\_sample.csv' INTO table sales;

select COUNTRY, COUNT(COUNTRY) from sales group by COUNTRY;

select COUNTRY, SUM(QUANTITYORDERED), STATUS from sales where STATUS = 'shipped' group by COUNTRY;

select DEALSIZE, COUNT(DEALSIZE) from sales group by DEALSIZE;

set hive.exec.dynamic.partition.mode=nonstrict;

CREATE EXTERNAL TABLE IF NOT EXISTS partitioned\_sales (

ORDERNUMBER BIGINT,

QUANTITYORDERED INT,

PRICEEACH FLOAT,

ORDERLINENUMBER INT,

SALES DOUBLE,

ORDERDATE STRING,

QTR\_ID INT,

MONTH\_ID INT,

YEAR\_ID INT,

PRODUCTLINE STRING,

MSRP INT,

PRODUCTCODE STRING,

CUSTOMERNAME STRING,

PHONE STRING,

ADDRESSLINE1 STRING,

ADDRESSLINE2 STRING,

CITY STRING,

STATE STRING,

POSTALCODE STRING,

COUNTRY STRING,

TERRITORY STRING,

CONTACTLASTNAME STRING,

CONTACTFIRSTNAME STRING,

DEALSIZE STRING)

partitioned by (STATUS STRING);

insert overwrite table partitioned\_sales partition(STATUS) select \* from sales;

describe formatted partitioned\_sales;