**Bash shell:**

hadoop fs -mkdir /user/cloudera/boston

hadoop fs -put boston.csv /user/cloudera/boston

#For hive table metastore access:

sudo su

cp /usr/lib/hive/conf/hive-site.xml /usr/lib/spark/conf

**Hive command prompt:**

hive

use default;

create table boston\_home(

crim double,

zn double,

indus double,

chas int,

nox double,

rm double,

age double,

dis double,

rad int,

tax int,

pt double,

b double,

lstat double,

mv double)

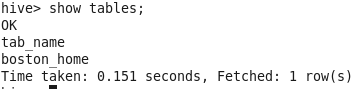
row format delimited

fields terminated by ','

stored as textfile

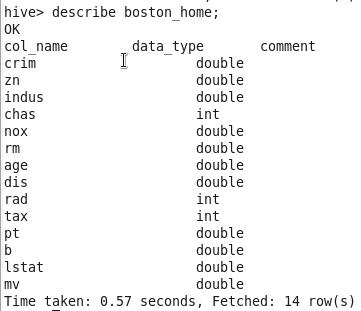
location '/user/cloudera/boston';

show tables;



set hive.cli.print.header = true; #to allow printing header

describe boston\_home;



select \* from boston\_home;

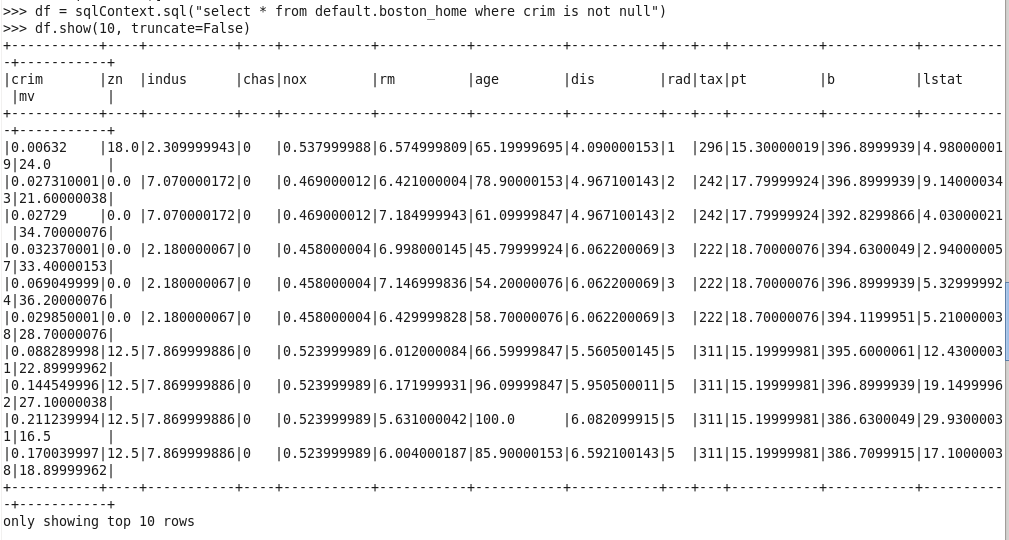
select \* from boston\_home limit 10;

**Pyspark bash:**

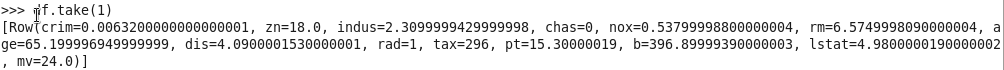
pyspark

df = sqlContext.sql("select \* from default.boston\_home where crim is not null")

df.show(10, truncate=False)



df.take(1)



df.cache()



df.printSchema()



for i in df.columns:

print("Correlation of ", i, "mv", df.stat.corr("mv", i))

