SPARK SQL COMMANDS

Launch Spark using python library: pyspark

```
from pyspark.sql import Row

 row1 = Row("Barack Obama", "President", "United States")
row1[0], row1[1]
row2 = Row(name="Alex", age=20)
row2
row2.name, row2.age


# DataFrame creation from RDD using ‘toDF’ function

rdd1 = sc.parallelize([Row(name='Alice', age=5, height=80),Row(name='Alice', age=5, height=80),Row(name='Alice', age=10, height=80)])

df = rdd1.toDF()
df.show()
df.printSchema()

# DataFrame creation from RDD using ‘createDataFrame’ function
rdd = sc.parallelize([('Alice', 1)])

sqlContext.createDataFrame(rdd).collect()
df = sqlContext.createDataFrame(rdd, ['name', 'age'])
df.collect()
df.show()
```

```
# Constructing Dataframe from a data source

df = spark.read.json("file:///home/hduser/data/people.json")

df.show()
df.printSchema()
df.select("name").show()
df.select("name", df.age + 1).show()
-------------------------------------------



# Practice SQL commands on ‘Yelp’ Dataset

biz = spark.read.json("file:///home/hduser/data/business.json")
biz.printSchema()

biz.registerTempTable("biz")
biz.cache()

sqlContext.sql("SELECT count(1) as businesses FROM biz").show()

sqlContext.sql("SELECT state, count(1) as businesses FROM biz GROUP BY
state").show(50)

sqlContext.sql("SELECT state, count(1) as businesses FROM biz GROUP BY state
ORDER BY businesses DESC").show(5)

sqlContext.sql("SELECT name, stars, review_count, city, state FROM biz WHERE
stars=5.0").show(5)

sqlContext.sql("SELECT name, stars, review_count, city, state FROM biz WHERE
state = 'NV' AND stars = 5.0").show(3)

sqlContext.sql("SELECT state, sum(review_count) as reviews FROM biz GROUP BY
state").show()
```

```
sqlContext.sql("SELECT stars, count(1) as businesses FROM biz GROUP BY
stars").show()

sqlContext.sql("SELECT state, AVG(review_count) as avg_reviews FROM biz
GROUP BY state").show()

sqlContext.sql("SELECT state, ROUND(AVG(review_count)) as avg_reviews FROM
biz GROUP BY state ORDER BY avg_reviews DESC LIMIT 5").show()

sqlContext.sql("SELECT name, stars, review_count FROM biz WHERE city = 'Las
Vegas' ORDER BY stars DESC, review_count DESC LIMIT 5 ").show()
```