Ritika Kumari(A20414073)

CSP554—Big Data Technologies

Assignment #7

Exercise 1)

Step A

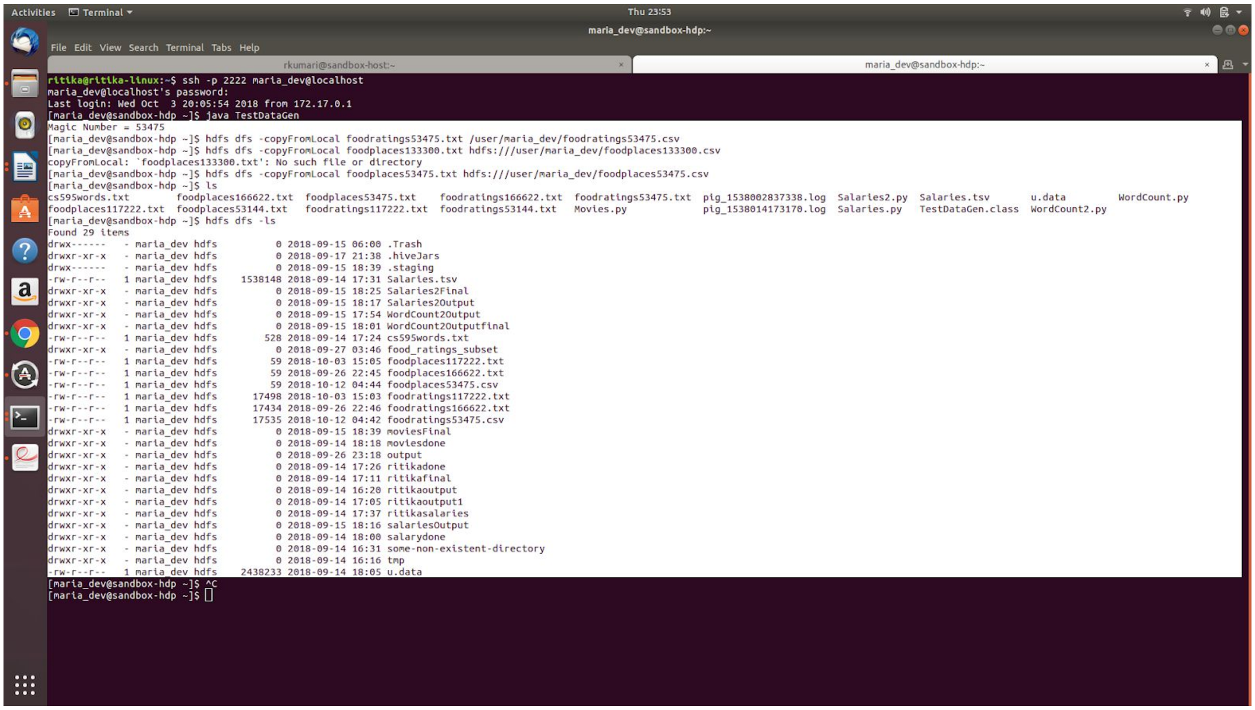
Use the TestDataGen program from previous assignments to generate new data files

Copy the files to HDFS.

Command Executed:

```
java TestDataGen

hdfs dfs -copyFromLocal foodratings53475.txt /user/maria_dev/foodratings53475.csv

hdfs dfs -copyFromLocal foodplaces53475.txt hdfs:///user/maria_dev/foodplaces53475.csv
```

Magic Number = 53475



```
ritika@ritika-linux:~$ ssh -p 2222 maria_dev@localhost
maria_dev@localhost's password:
Last login: Wed Oct  3 20:05:54 2018 from 172.17.0.1
[maria_dev@sandbox-hdp ~]$ java TestDataGen
Magic Number = 53475
[maria_dev@sandbox-hdp ~]$ hdfs dfs -copyFromLocal foodratings53475.txt /user/maria_dev/foodratings53475.csv
[maria_dev@sandbox-hdp ~]$ hdfs dfs -copyFromLocal foodplaces133300.txt hdfs:///user/maria_dev/foodplaces133300.csv
copyFromLocal: `foodplaces133300.txt': No such file or directory
[maria_dev@sandbox-hdp ~]$ hdfs dfs -copyFromLocal foodplaces53475.txt hdfs:///user/maria_dev/foodplaces53475.csv
[maria_dev@sandbox-hdp ~]$ ls
cs595words.txt        foodplaces166622.txt  foodplaces53475.txt    foodratings166622.txt  foodratings53475.txt  pig_1538002837338.log  Salaries2.py  Salaries.tsv       u.data         WordCount.py
foodplaces117222.txt  foodplaces53144.txt   foodratings117222.txt  foodratings53144.txt   Movies.py             pig_1538014173170.log  Salaries.py   TestDataGen.class  WordCount2.py
[maria_dev@sandbox-hdp ~]$ hdfs dfs -ls
Found 29 items
drwx------   - maria_dev hdfs          0 2018-09-15 06:00 .Trash
drwxr-xr-x   - maria_dev hdfs          0 2018-09-17 21:38 .hiveJars
drwx------   - maria_dev hdfs          0 2018-09-15 18:39 .staging
-rw-r--r--   1 maria_dev hdfs    1538148 2018-09-14 17:31 Salaries.tsv
drwxr-xr-x   - maria_dev hdfs          0 2018-09-15 18:25 Salaries2Final
drwxr-xr-x   - maria_dev hdfs          0 2018-09-15 18:17 Salaries2Output
drwxr-xr-x   - maria_dev hdfs          0 2018-09-15 17:54 WordCount2Output
drwxr-xr-x   - maria_dev hdfs          0 2018-09-15 18:01 WordCount2Outputfinal
-rw-r--r--   1 maria_dev hdfs        528 2018-09-14 17:24 cs595words.txt
drwxr-xr-x   - maria_dev hdfs          0 2018-09-27 03:46 food_ratings_subset
-rw-r--r--   1 maria_dev hdfs         59 2018-10-03 15:05 foodplaces117222.txt
-rw-r--r--   1 maria_dev hdfs         59 2018-09-26 22:45 foodplaces166622.txt
-rw-r--r--   1 maria_dev hdfs         59 2018-10-12 04:44 foodplaces53475.csv
-rw-r--r--   1 maria_dev hdfs      17498 2018-10-03 15:03 foodratings117222.txt
-rw-r--r--   1 maria_dev hdfs      17434 2018-09-26 22:46 foodratings166622.txt
-rw-r--r--   1 maria_dev hdfs      17535 2018-10-12 04:42 foodratings53475.csv
drwxr-xr-x   - maria_dev hdfs          0 2018-09-15 18:39 moviesFinal
drwxr-xr-x   - maria_dev hdfs          0 2018-09-14 18:18 moviesdone
drwxr-xr-x   - maria_dev hdfs          0 2018-09-26 23:18 output
drwxr-xr-x   - maria_dev hdfs          0 2018-09-14 17:26 ritikadone
drwxr-xr-x   - maria_dev hdfs          0 2018-09-14 17:11 ritikafinal
drwxr-xr-x   - maria_dev hdfs          0 2018-09-14 16:20 ritikaoutput
drwxr-xr-x   - maria_dev hdfs          0 2018-09-14 17:05 ritikaoutput1
drwxr-xr-x   - maria_dev hdfs          0 2018-09-14 17:37 ritikasalaries
drwxr-xr-x   - maria_dev hdfs          0 2018-09-15 18:16 salariesOutput
drwxr-xr-x   - maria_dev hdfs          0 2018-09-14 18:00 salarydone
drwxr-xr-x   - maria_dev hdfs          0 2018-09-14 16:31 some-non-existent-directory
drwxr-xr-x   - maria_dev hdfs          0 2018-09-14 16:16 tmp
-rw-r--r--   1 maria_dev hdfs    2438233 2018-09-14 18:05 u.data
[maria_dev@sandbox-hdp ~]$ ^C
[maria_dev@sandbox-hdp ~]$ 
```

Step B

Load the ‘foodratings’ file as a ‘csv’ file into a DataFrame called ex1_foodratings. When doing so specify a schema having fields of the following names and types:

Field Name	Field Type
name	String
food1	Integer
food1	Integer
food1	Integer
food1	Integer
placeid	Integer

As the results of this exercise provide the magic number, the code you execute and screen shots of the following commands:

foodratings.printSchema()

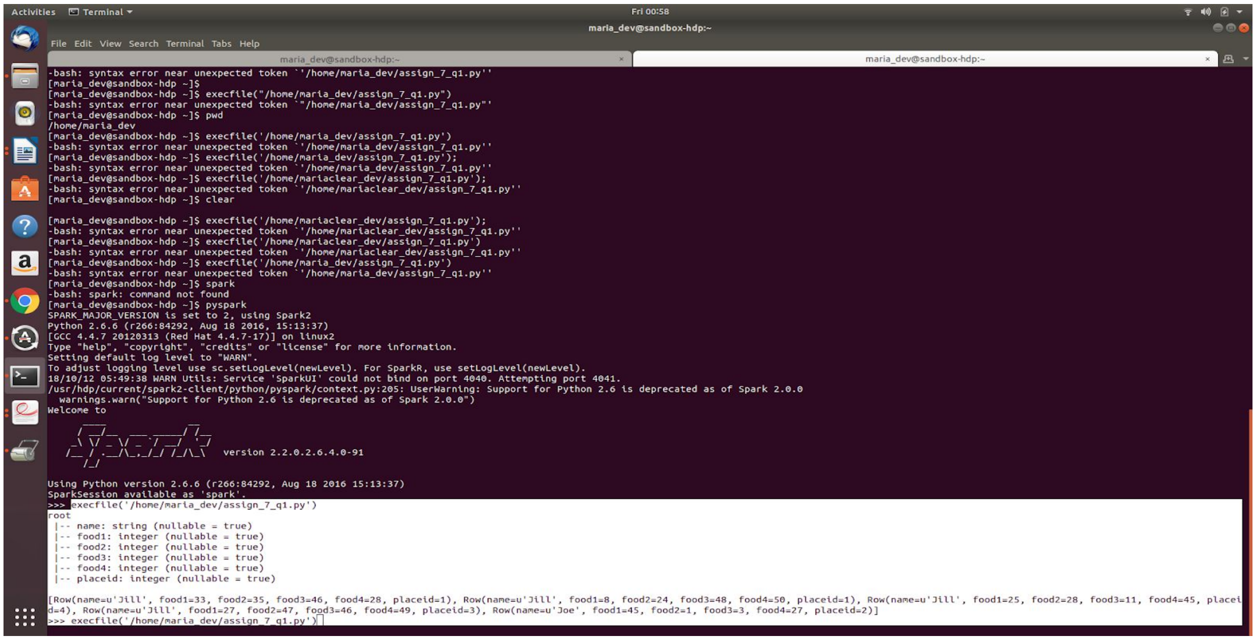
foodratings.head(5)

Command Executed:

vi assign_7_q1.py

```
from pyspark.sql.types import *

struct1 = StructType(

        [

                StructField("name", StringType(), True),

                StructField("food1",IntegerType(), True),

                StructField("food2",IntegerType(), True),

                StructField("food3",IntegerType(), True),

                StructField("food4",IntegerType(), True),

                StructField("placeid",IntegerType(), True)

        ]

)
```

execfile('/home/maria_dev/assign_7_q1.py')



```
-bash: syntax error near unexpected token `'/home/maria_dev/assign_7_q1.py''
[maria_dev@sandbox-hdp ~]$
[maria_dev@sandbox-hdp ~]$ execfile("/home/maria_dev/assign_7_q1.py")
-bash: syntax error near unexpected token `"/home/maria_dev/assign_7_q1.py"'
[maria_dev@sandbox-hdp ~]$ pwd
/home/maria_dev
[maria_dev@sandbox-hdp ~]$ execfile('/home/maria_dev/assign_7_q1.py')
-bash: syntax error near unexpected token `'/home/maria_dev/assign_7_q1.py''
[maria_dev@sandbox-hdp ~]$ execfile('/home/maria_dev/assign_7_q1.py');
-bash: syntax error near unexpected token `'/home/maria_dev/assign_7_q1.py''
[maria_dev@sandbox-hdp ~]$ execfile('/home/mariaclear_dev/assign_7_q1.py');
-bash: syntax error near unexpected token `'/home/mariaclear_dev/assign_7_q1.py''
[maria_dev@sandbox-hdp ~]$ clear

[maria_dev@sandbox-hdp ~]$ execfile('/home/mariaclear_dev/assign_7_q1.py');
-bash: syntax error near unexpected token `'/home/mariaclear_dev/assign_7_q1.py''
[maria_dev@sandbox-hdp ~]$ execfile('/home/mariaclear_dev/assign_7_q1.py')
-bash: syntax error near unexpected token `'/home/mariaclear_dev/assign_7_q1.py''
[maria_dev@sandbox-hdp ~]$ execfile('/home/maria_dev/assign_7_q1.py')
-bash: syntax error near unexpected token `'/home/maria_dev/assign_7_q1.py''
[maria_dev@sandbox-hdp ~]$ spark
-bash: spark: command not found
[maria_dev@sandbox-hdp ~]$ pyspark
SPARK_MAJOR_VERSION is set to 2, using Spark2
Python 2.6.6 (r266:84292, Aug 18 2016, 15:13:37)
[GCC 4.4.7 20120313 (Red Hat 4.4.7-17)] on linux2
Type "help", "copyright", "credits" or "license" for more information.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
18/10/12 05:49:38 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.
/usr/hdp/current/spark2-client/python/pyspark/context.py:205: UserWarning: Support for Python 2.6 is deprecated as of Spark 2.0.0
  warnings.warn("Support for Python 2.6 is deprecated as of Spark 2.0.0")
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 2.2.0.2.6.4.0-91
      /_/

Using Python version 2.6.6 (r266:84292, Aug 18 2016 15:13:37)
SparkSession available as 'spark'.
>>> execfile('/home/maria_dev/assign_7_q1.py')
root
 |-- name: string (nullable = true)
 |-- food1: integer (nullable = true)
 |-- food2: integer (nullable = true)
 |-- food3: integer (nullable = true)
 |-- food4: integer (nullable = true)
 |-- placeid: integer (nullable = true)

[Row(name=u'Jill', food1=33, food2=35, food3=46, food4=28, placeid=1), Row(name=u'Jill', food1=8, food2=24, food3=48, food4=50, placeid=1), Row(name=u'Jill', food1=25, food2=28, food3=11, food4=45, placei
d=4), Row(name=u'Jill', food1=27, food2=47, food3=46, food4=49, placeid=3), Row(name=u'Joe', food1=45, food2=1, food3=3, food4=27, placeid=2)]
>>> execfile('/home/maria_dev/assign_7_q1.py')
```

Exercise 2)

Load the ‘foodplaces’ file as a ‘csv’ file into a DataFrame called foodplaces. When doing so specify a schema having fields of the following names and types:

Field Nampee	Field Ty
placeid	integer
placename	string

As the results of this exercise provide the code you execute and screen shots of the following commands:

foodratings.printSchema()

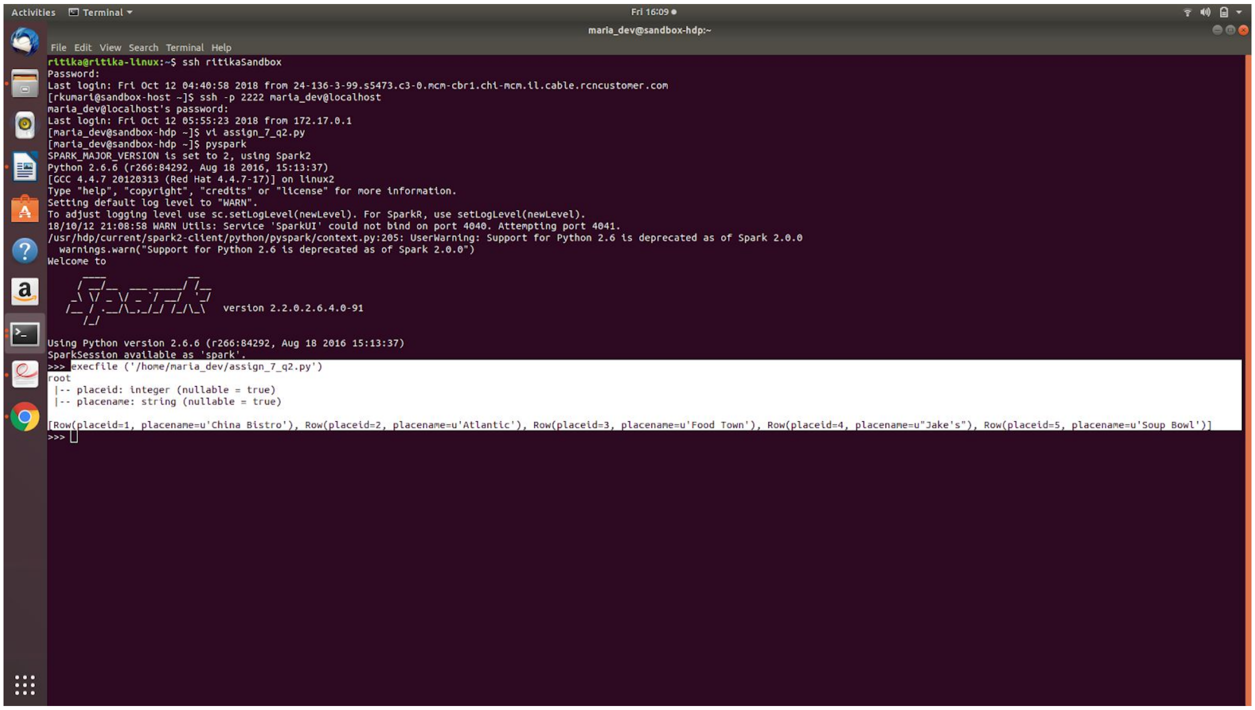
foodratings.head(5)

Command Executed:

```
vi assign_7_q2.py
```

```
from pyspark.sql.types import *

struct1 = StructType().add("placeid", IntegerType(), True).add("placename",StringType(), True)

foodplaces = spark.read.schema(struct1).csv('/user/maria_dev/foodplaces53475.csv')

foodplaces.printSchema()

print foodplaces.head(5)
```

```
execfile ('/home/maria_dev/assign_7_q2.py')
```



```
ritika@ritika-linux:~$ ssh ritikaSandbox
Password:
Last login: Fri Oct 12 04:40:58 2018 from 24-136-3-99.s5473.c3-0.mcm-cbr1.chi-mcm.il.cable.rcncustomer.com
[rkumari@sandbox-host ~]$ ssh -p 2222 maria_dev@localhost
maria_dev@localhost's password:
Last login: Fri Oct 12 05:55:23 2018 from 172.17.0.1
[maria_dev@sandbox-hdp ~]$ vi assign_7_q2.py
[maria_dev@sandbox-hdp ~]$ pyspark
SPARK_MAJOR_VERSION is set to 2, using Spark2
Python 2.6.6 (r266:84292, Aug 18 2016, 15:13:37)
[GCC 4.4.7 20120313 (Red Hat 4.4.7-17)] on linux2
Type "help", "copyright", "credits" or "license" for more information.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
18/10/12 21:08:58 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.
/usr/hdp/current/spark2-client/python/pyspark/context.py:205: UserWarning: Support for Python 2.6 is deprecated as of Spark 2.0.0
  warnings.warn("Support for Python 2.6 is deprecated as of Spark 2.0.0")
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 2.2.0.2.6.4.0-91
      /_/

Using Python version 2.6.6 (r266:84292, Aug 18 2016 15:13:37)
SparkSession available as 'spark'.
>>> execfile ('/home/maria_dev/assign_7_q2.py')
root
 |-- placeid: integer (nullable = true)
 |-- placename: string (nullable = true)

[Row(placeid=1, placename=u'China Bistro'), Row(placeid=2, placename=u'Atlantic'), Row(placeid=3, placename=u'Food Town'), Row(placeid=4, placename=u"Jake's"), Row(placeid=5, placename=u'Soup Bowl')]
>>> 
```

Exercise 3)

Step A

Register the DataFrames created in exercise 1 and 2 as tables called “foodratingsT” and “foodplacesT”

Step B

Use a SQL query on the table “foodratingsT” to create a new DataFrame called foodratings_ex3 holding records which meet the following condition: food2 < 25 and food4 > 40

As the results of this step provide the code you execute and screen shots of the following commands:

foodratings.printSchema()

foodratings.head(5)

Step C

Use a SQL query on the table “foodplacesT” to create a new DataFrame called foodplaces_ex3 holding records which meet the following condition: placeid > 3

As the results of this step provide the code you execute and screen shots of the following commands:

```
        foodratings.printSchema()

        foodratings.head(5)
```

Command Executed:

```
vi assign_7_q3.py

from pyspark.sql.types import *

structfr = StructType(

    [

        StructField("name", StringType(), True),

        StructField("food1",IntegerType(), True),

        StructField("food2",IntegerType(), True),

        StructField("food3",IntegerType(), True),

        StructField("food4",IntegerType(), True),

        StructField("placeid",IntegerType(), True)

    ]

)


structfp = StructType().add("placeid", IntegerType(), True).add("placename",StringType(), True)

foodratings = spark.read.schema(structfr).csv('/user/maria_dev/foodratings53475.csv')

foodplaces = spark.read.schema(structfp).csv('/user/maria_dev/foodplaces53475.csv')


foodratings.createOrReplaceTempView("foodratingsT")

foodplaces.createOrReplaceTempView("foodplacesT")


foodratings_ex3 = spark.sql("SELECT * FROM foodratingsT WHERE food2 < 25 AND food4 > 40")

foodratings_ex3.printSchema()
```

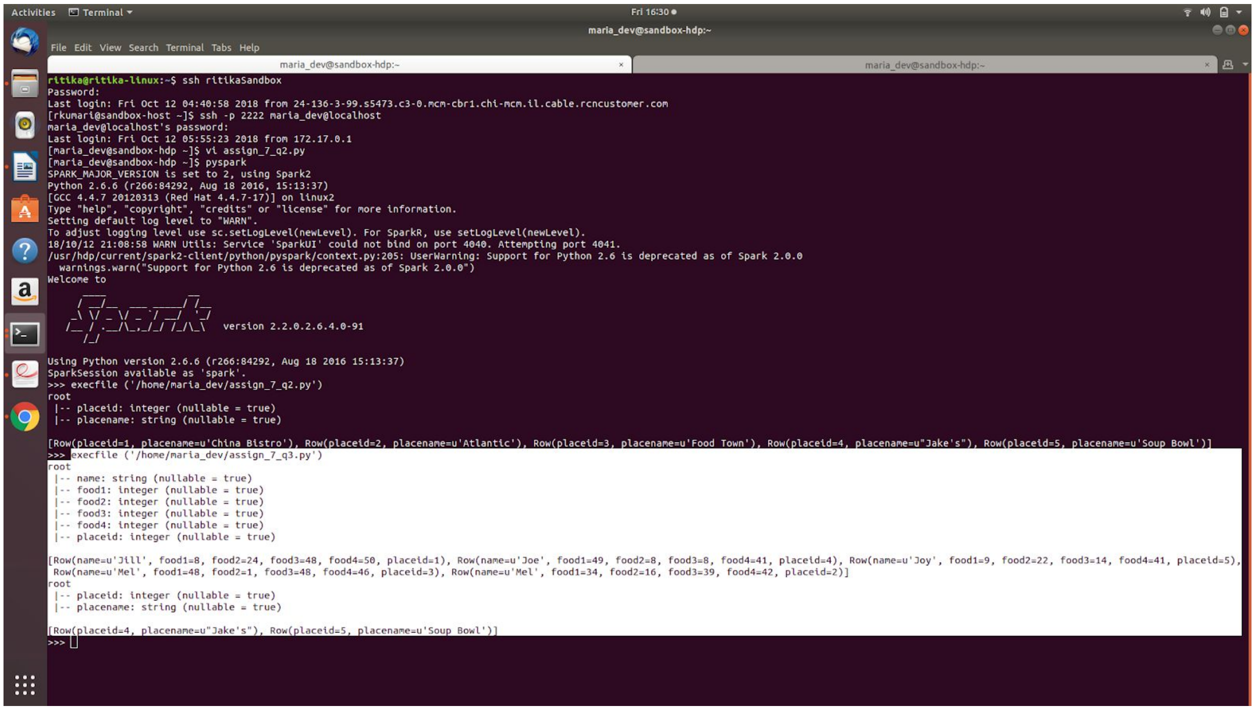
print foodratings_ex3.head(5)

foodplaces_ex3 = spark.sql("SELECT * FROM foodplacesT WHERE placeid > 3")

foodplaces_ex3.printSchema()

print foodplaces_ex3.head(5)

execfile ('/home/maria_dev/assign_7_q3.py')



```
ritika@ritika-linux:~$ ssh ritikaSandbox
Password:
Last login: Fri Oct 12 04:40:58 2018 from 24-136-3-99.s5473.c3-0.mcm-cbr1.chi-mcm.il.cable.rcncustomer.com
[rkumari@sandbox-host ~]$ ssh -p 2222 maria_dev@localhost
maria_dev@localhost's password:
Last login: Fri Oct 12 05:55:23 2018 from 172.17.0.1
[maria_dev@sandbox-hdp ~]$ vi assign_7_q2.py
[maria_dev@sandbox-hdp ~]$ pyspark
SPARK_MAJOR_VERSION is set to 2, using Spark2
Python 2.6.6 (r266:84292, Aug 18 2016, 15:13:37)
[GCC 4.4.7 20120313 (Red Hat 4.4.7-17)] on linux2
Type "help", "copyright", "credits" or "license" for more information.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
18/10/12 21:08:58 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.
/usr/hdp/current/spark2-client/python/pyspark/context.py:205: UserWarning: Support for Python 2.6 is deprecated as of Spark 2.0.0
  warnings.warn("Support for Python 2.6 is deprecated as of Spark 2.0.0")
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 2.2.0.2.6.4.0-91
      /_/

Using Python version 2.6.6 (r266:84292, Aug 18 2016 15:13:37)
SparkSession available as 'spark'.
>>> execfile ('/home/maria_dev/assign_7_q2.py')
root
 |-- placeid: integer (nullable = true)
 |-- placename: string (nullable = true)

[Row(placeid=1, placename=u'China Bistro'), Row(placeid=2, placename=u'Atlantic'), Row(placeid=3, placename=u'Food Town'), Row(placeid=4, placename=u"Jake's"), Row(placeid=5, placename=u'Soup Bowl')]
>>> execfile ('/home/maria_dev/assign_7_q3.py')
root
 |-- name: string (nullable = true)
 |-- food1: integer (nullable = true)
 |-- food2: integer (nullable = true)
 |-- food3: integer (nullable = true)
 |-- food4: integer (nullable = true)
 |-- placeid: integer (nullable = true)

[Row(name=u'Jill', food1=8, food2=24, food3=48, food4=50, placeid=1), Row(name=u'Joe', food1=49, food2=8, food3=8, food4=41, placeid=4), Row(name=u'Joy', food1=9, food2=22, food3=14, food4=41, placeid=5),
 Row(name=u'Mel', food1=48, food2=1, food3=48, food4=46, placeid=3), Row(name=u'Mel', food1=34, food2=16, food3=39, food4=42, placeid=2)]
root
 |-- placeid: integer (nullable = true)
 |-- placename: string (nullable = true)

[Row(placeid=4, placename=u"Jake's"), Row(placeid=5, placename=u'Soup Bowl')]
>>> 
```

Exercise 4)

Use an operation (not a SQL query) on the DataFrame ‘foodratings’ create in exercise 1 to create a new DataFrame called foodratings_ex4 that includes only those records (rows) where the ‘name’ field is “Mel” and food3 < 25.

As the results of this step provide the code you execute and screen shots of the following commands:

foodratings.printSchema()

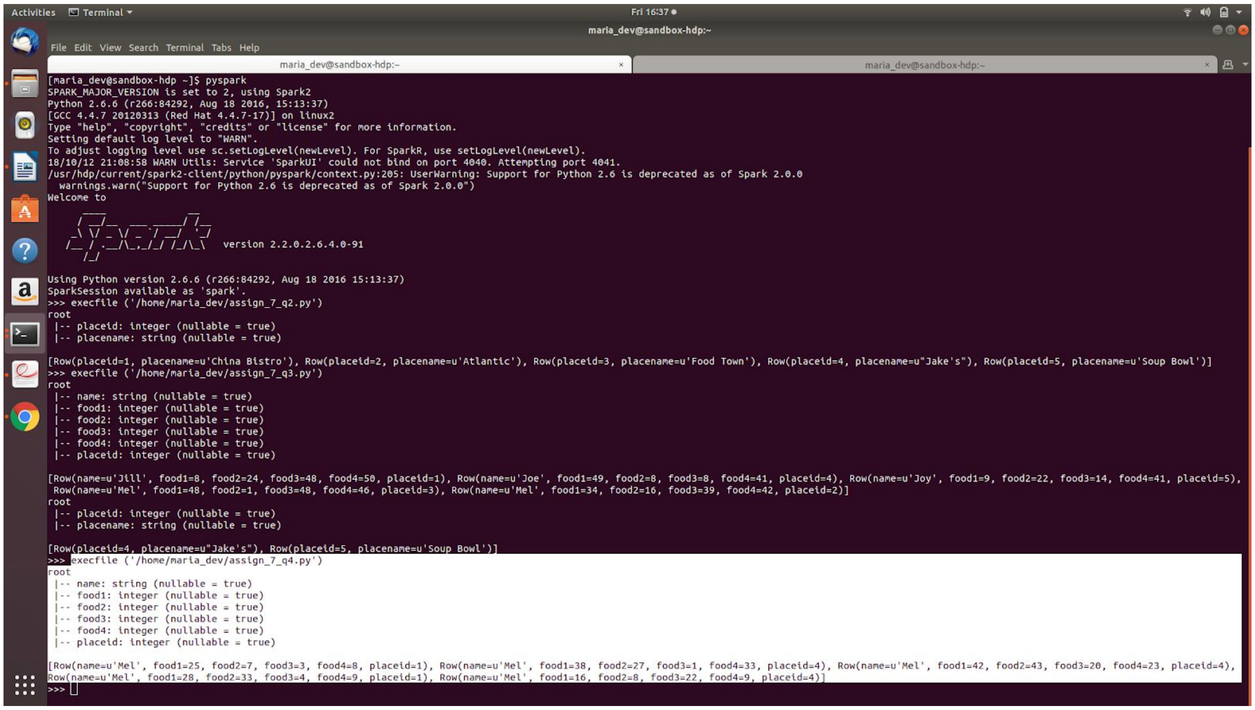
foodratings.head(5)

Command Executed:

```
vi assign_7_q4.py
```

```
from pyspark.sql.types import *
struct1 = StructType(
    [
        StructField("name", StringType(), True),
        StructField("food1",IntegerType(), True),
        StructField("food2",IntegerType(), True),
        StructField("food3",IntegerType(), True),
        StructField("food4",IntegerType(), True),
        StructField("placeid",IntegerType(), True)
    ]
)

foodratings = spark.read.schema(struct1).csv('/user/maria_dev/foodratings53475.csv')
foodratings_ex4 = foodratings.filter((foodratings['name'] == "Mel") & (foodratings['food3'] < 25))
foodratings_ex4.printSchema()
print foodratings_ex4.head(5)
```

```
execfile ('/home/maria_dev/assign_7_q4.py')
```



```
[maria_dev@sandbox-hdp ~]$ pyspark
SPARK_MAJOR_VERSION is set to 2, using Spark2
Python 2.6.6 (r266:84292, Aug 18 2016, 15:13:37)
[GCC 4.4.7 20120313 (Red Hat 4.4.7-17)] on linux2
Type "help", "copyright", "credits" or "license" for more information.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
18/10/12 21:08:58 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.
/usr/hdp/current/spark2-client/python/pyspark/context.py:205: UserWarning: Support for Python 2.6 is deprecated as of Spark 2.0.0
  warnings.warn("Support for Python 2.6 is deprecated as of Spark 2.0.0")
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 2.2.0.2.6.4.0-91
      /_/

Using Python version 2.6.6 (r266:84292, Aug 18 2016 15:13:37)
SparkSession available as 'spark'.
>>> execfile ('/home/maria_dev/assign_7_q2.py')
root
 |-- placeid: integer (nullable = true)
 |-- placename: string (nullable = true)

[Row(placeid=1, placename=u'China Bistro'), Row(placeid=2, placename=u'Atlantic'), Row(placeid=3, placename=u'Food Town'), Row(placeid=4, placename=u"Jake's"), Row(placeid=5, placename=u'Soup Bowl')]
>>> execfile ('/home/maria_dev/assign_7_q3.py')
root
 |-- name: string (nullable = true)
 |-- food1: integer (nullable = true)
 |-- food2: integer (nullable = true)
 |-- food3: integer (nullable = true)
 |-- food4: integer (nullable = true)
 |-- placeid: integer (nullable = true)

[Row(name=u'Jill', food1=8, food2=24, food3=48, food4=50, placeid=1), Row(name=u'Joe', food1=49, food2=8, food3=8, food4=41, placeid=4), Row(name=u'Joy', food1=9, food2=22, food3=14, food4=41, placeid=5),
 Row(name=u'Mel', food1=48, food2=1, food3=48, food4=46, placeid=3), Row(name=u'Mel', food1=34, food2=16, food3=39, food4=42, placeid=2)]
root
 |-- placeid: integer (nullable = true)
 |-- placename: string (nullable = true)

[Row(placeid=4, placename=u"Jake's"), Row(placeid=5, placename=u'Soup Bowl')]
>>> execfile ('/home/maria_dev/assign_7_q4.py')
root
 |-- name: string (nullable = true)
 |-- food1: integer (nullable = true)
 |-- food2: integer (nullable = true)
 |-- food3: integer (nullable = true)
 |-- food4: integer (nullable = true)
 |-- placeid: integer (nullable = true)

[Row(name=u'Mel', food1=25, food2=7, food3=3, food4=8, placeid=1), Row(name=u'Mel', food1=38, food2=27, food3=1, food4=33, placeid=4), Row(name=u'Mel', food1=42, food2=43, food3=20, food4=23, placeid=4),
Row(name=u'Mel', food1=28, food2=33, food3=4, food4=9, placeid=1), Row(name=u'Mel', food1=16, food2=8, food3=22, food4=9, placeid=4)]
>>> 
```

Exercise 5)

Use an operation (not a SQL query) on the DataFrame ‘foodratings’ create in exercise 1 to create a new DataFrame called foodratings_ex5 that includes only the columns (fields) ‘name’ and ‘placeid’

As the results of this step provide the code you execute and screen shots of the following commands:

foodratings.printSchema()

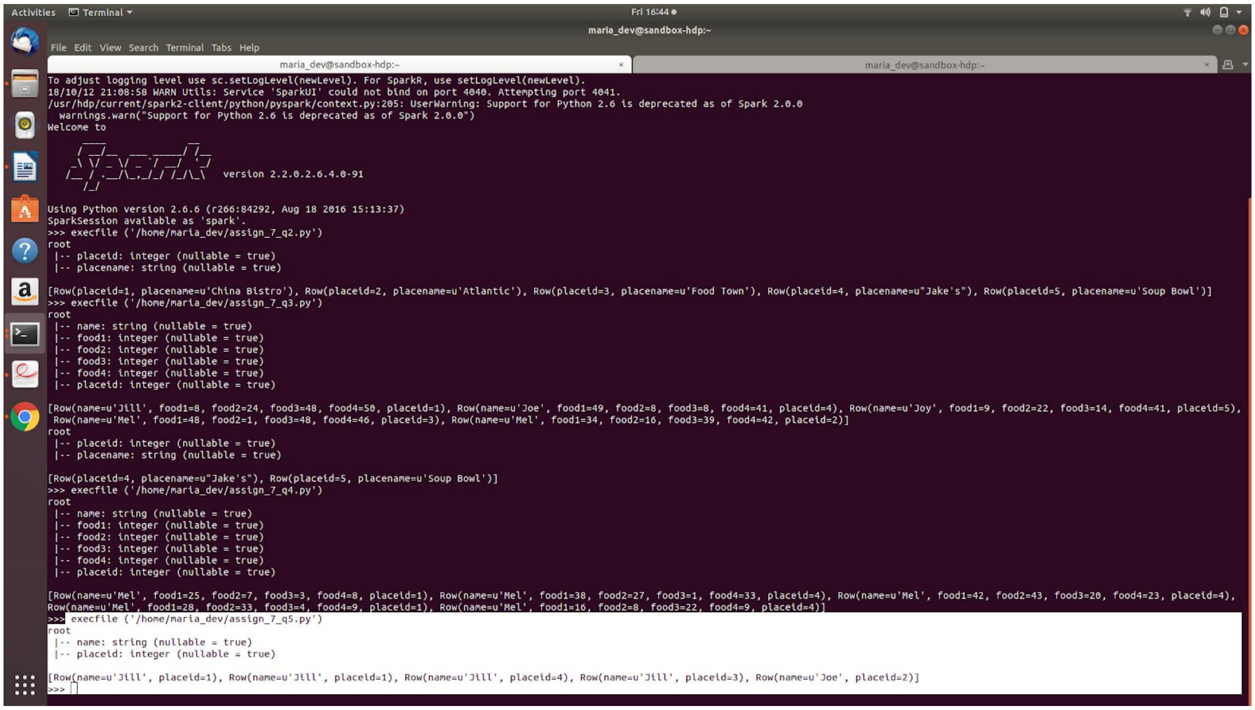
foodratings.head(5)

Command Executed:

vi assign_7_q5.py

```
from pyspark.sql.types import *

struct1 = StructType(

    [

        StructField("name", StringType(), True),

        StructField("food1",IntegerType(), True),
            StructField("food2",IntegerType(), True),

            StructField("food3",IntegerType(), True),

            StructField("food4",IntegerType(), True),

            StructField("placeid",IntegerType(), True)

    ]

)



foodratings = spark.read.schema(struct1).csv('/user/maria_dev/foodratings53475.csv')

foodratings_ex5 = foodratings.select(foodratings['name'],foodratings['placeid'])

foodratings_ex5.printSchema()

print foodratings_ex5.head(5)
```

```
execfile ('/home/maria_dev/assign_7_q5.py')
```



```
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
18/10/12 21:08:58 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.
/usr/hdp/current/spark2-client/python/pyspark/context.py:205: UserWarning: Support for Python 2.6 is deprecated as of Spark 2.0.0
  warnings.warn("Support for Python 2.6 is deprecated as of Spark 2.0.0")
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 2.2.0.2.6.4.0-91
      /_/

Using Python version 2.6.6 (r266:84292, Aug 18 2016 15:13:37)
SparkSession available as 'spark'.
>>> execfile ('/home/maria_dev/assign_7_q2.py')
root
 |-- placeid: integer (nullable = true)
 |-- placename: string (nullable = true)

[Row(placeid=1, placename=u'China Bistro'), Row(placeid=2, placename=u'Atlantic'), Row(placeid=3, placename=u'Food Town'), Row(placeid=4, placename=u"Jake's"), Row(placeid=5, placename=u'Soup Bowl')]
>>> execfile ('/home/maria_dev/assign_7_q3.py')
root
 |-- name: string (nullable = true)
 |-- food1: integer (nullable = true)
 |-- food2: integer (nullable = true)
 |-- food3: integer (nullable = true)
 |-- food4: integer (nullable = true)
 |-- placeid: integer (nullable = true)

[Row(name=u'Jill', food1=8, food2=24, food3=48, food4=50, placeid=1), Row(name=u'Joe', food1=49, food2=8, food3=8, food4=41, placeid=4), Row(name=u'Joy', food1=9, food2=22, food3=14, food4=41, placeid=5),
 Row(name=u'Mel', food1=48, food2=1, food3=48, food4=46, placeid=3), Row(name=u'Mel', food1=34, food2=16, food3=39, food4=42, placeid=2)]
root
 |-- placeid: integer (nullable = true)
 |-- placename: string (nullable = true)

[Row(placeid=4, placename=u"Jake's"), Row(placeid=5, placename=u'Soup Bowl')]
>>> execfile ('/home/maria_dev/assign_7_q4.py')
root
 |-- name: string (nullable = true)
 |-- food1: integer (nullable = true)
 |-- food2: integer (nullable = true)
 |-- food3: integer (nullable = true)
 |-- food4: integer (nullable = true)
 |-- placeid: integer (nullable = true)

[Row(name=u'Mel', food1=25, food2=7, food3=3, food4=8, placeid=1), Row(name=u'Mel', food1=38, food2=27, food3=1, food4=33, placeid=4), Row(name=u'Mel', food1=42, food2=43, food3=20, food4=23, placeid=4),
Row(name=u'Mel', food1=28, food2=33, food3=4, food4=9, placeid=1), Row(name=u'Mel', food1=16, food2=8, food3=22, food4=9, placeid=4)]
>>> execfile ('/home/maria_dev/assign_7_q5.py')
root
 |-- name: string (nullable = true)
 |-- placeid: integer (nullable = true)

[Row(name=u'Jill', placeid=1), Row(name=u'Jill', placeid=1), Row(name=u'Jill', placeid=4), Row(name=u'Jill', placeid=3), Row(name=u'Joe', placeid=2)]
>>> 
```

Exercise 6)

Use an operation on the DataFrame ‘to create a new DataFrame called ex6 which is the inner join, on placeid, of the DataFrames ‘foodratings; and ‘foodplaces’ created in exercises 1 and 2

As the results of this step provide the code you execute and screen shots of the following commands:

```
ex6.printSchema()

ex6.head(5)
```

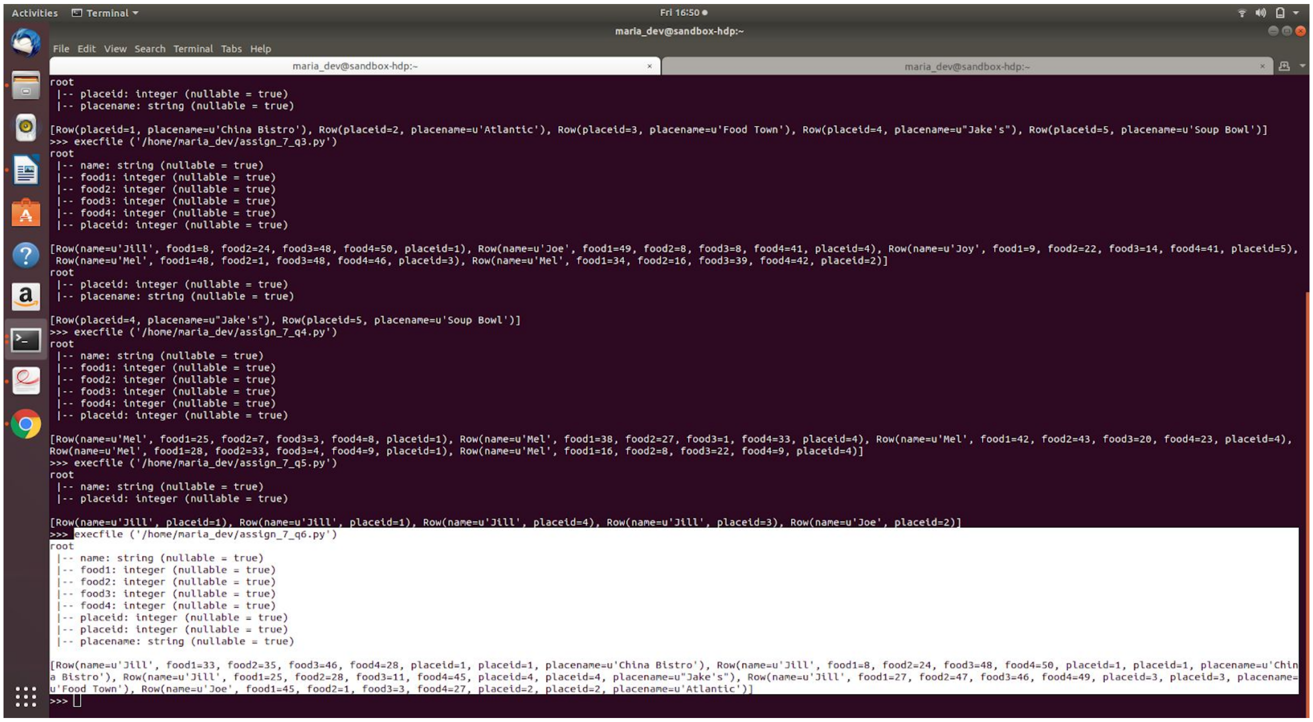
Command Executed:

```
vi assign_7_q6.py
```

```
from pyspark.sql.types import *

structfr = StructType(

    [

        StructField("name", StringType(), True),

        StructField("food1",IntegerType(), True),

        StructField("food2",IntegerType(), True),

        StructField("food3",IntegerType(), True),

        StructField("food4",IntegerType(), True),

        StructField("placeid",IntegerType(), True)

    ]

)


structfp = StructType().add("placeid", IntegerType(), True).add("placename",StringType(), True)


foodratings = spark.read.schema(structfr).csv('/user/maria_dev/foodratings53475.csv')

foodplaces = spark.read.schema(structfp).csv('/user/maria_dev/foodplaces53475.csv')
```

```
ex6 = foodratings.join(foodplaces, foodratings.placeid == foodplaces.placeid, 'inner')

ex6.printSchema()

print ex6.head(5)
```

```
execfile ('/home/maria_dev/assign_7_q6.py')
```



```
root
 |-- placeid: integer (nullable = true)
 |-- placename: string (nullable = true)

[Row(placeid=1, placename=u'China Bistro'), Row(placeid=2, placename=u'Atlantic'), Row(placeid=3, placename=u'Food Town'), Row(placeid=4, placename=u"Jake's"), Row(placeid=5, placename=u'Soup Bowl')]
>>> execfile ('/home/maria_dev/assign_7_q3.py')
root
 |-- name: string (nullable = true)
 |-- food1: integer (nullable = true)
 |-- food2: integer (nullable = true)
 |-- food3: integer (nullable = true)
 |-- food4: integer (nullable = true)
 |-- placeid: integer (nullable = true)

[Row(name=u'Jill', food1=8, food2=24, food3=48, food4=50, placeid=1), Row(name=u'Joe', food1=49, food2=8, food3=8, food4=41, placeid=4), Row(name=u'Joy', food1=9, food2=22, food3=14, food4=41, placeid=5),
 Row(name=u'Mel', food1=48, food2=1, food3=48, food4=46, placeid=3), Row(name=u'Mel', food1=34, food2=16, food3=39, food4=42, placeid=2)]
root
 |-- placeid: integer (nullable = true)
 |-- placename: string (nullable = true)

[Row(placeid=4, placename=u"Jake's"), Row(placeid=5, placename=u'Soup Bowl')]
>>> execfile ('/home/maria_dev/assign_7_q4.py')
root
 |-- name: string (nullable = true)
 |-- food1: integer (nullable = true)
 |-- food2: integer (nullable = true)
 |-- food3: integer (nullable = true)
 |-- food4: integer (nullable = true)
 |-- placeid: integer (nullable = true)

[Row(name=u'Mel', food1=25, food2=7, food3=3, food4=8, placeid=1), Row(name=u'Mel', food1=38, food2=27, food3=1, food4=33, placeid=4), Row(name=u'Mel', food1=42, food2=43, food3=20, food4=23, placeid=4),
Row(name=u'Mel', food1=28, food2=33, food3=4, food4=9, placeid=1), Row(name=u'Mel', food1=16, food2=8, food3=22, food4=9, placeid=4)]
>>> execfile ('/home/maria_dev/assign_7_q5.py')
root
 |-- name: string (nullable = true)
 |-- placeid: integer (nullable = true)

[Row(name=u'Jill', placeid=1), Row(name=u'Jill', placeid=1), Row(name=u'Jill', placeid=4), Row(name=u'Jill', placeid=3), Row(name=u'Joe', placeid=2)]
>>> execfile ('/home/maria_dev/assign_7_q6.py')
root
 |-- name: string (nullable = true)
 |-- food1: integer (nullable = true)
 |-- food2: integer (nullable = true)
 |-- food3: integer (nullable = true)
 |-- food4: integer (nullable = true)
 |-- placeid: integer (nullable = true)
 |-- placeid: integer (nullable = true)
 |-- placename: string (nullable = true)

[Row(name=u'Jill', food1=33, food2=35, food3=46, food4=28, placeid=1, placeid=1, placename=u'China Bistro'), Row(name=u'Jill', food1=8, food2=24, food3=48, food4=50, placeid=1, placeid=1, placename=u'Chin
a Bistro'), Row(name=u'Jill', food1=25, food2=28, food3=11, food4=45, placeid=4, placeid=4, placename=u"Jake's"), Row(name=u'Jill', food1=27, food2=47, food3=46, food4=49, placeid=3, placeid=3, placename=
u'Food Town'), Row(name=u'Joe', food1=45, food2=1, food3=3, food4=27, placeid=2, placeid=2, placename=u'Atlantic')]
>>> 
```