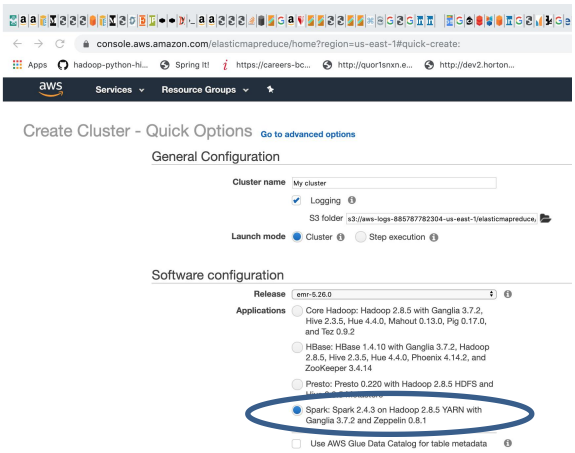
Assignment #7 A20512400

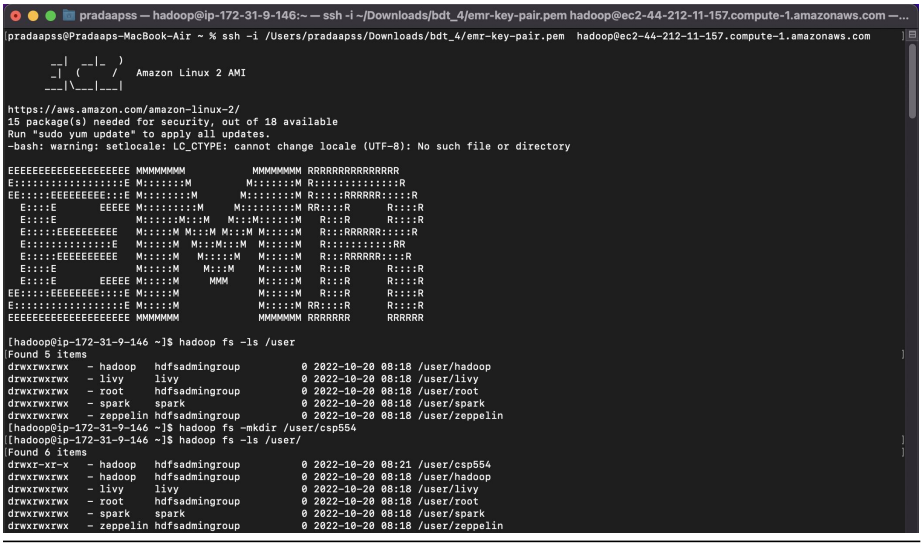
Exercise 1)

Step A

Start up a Hadoop cluster as previously, but instead of choosing the “Core Hadoop” configuration chose the “Spark” configuration (see below), otherwise proceed as before.



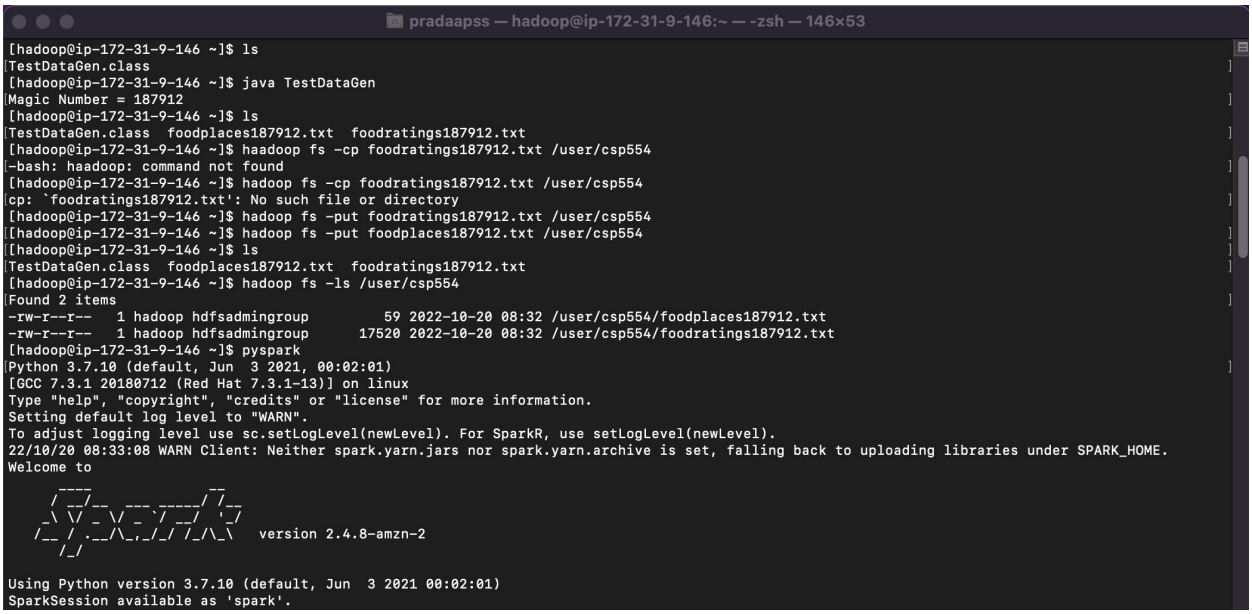
Step B



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

Copy both generated files to the HDFS directory “/user/hadoop”

Magic Number: 187912

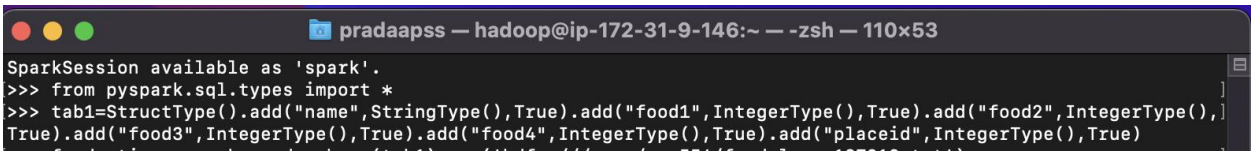


Step C

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

Field Name	Field Type
name	String
food1	Integer
food2	Integer
food3	Integer
food4	Integer
placeid	Integer

```
>>>from pyspark.sql.types import *
>>>tab1=StructType().add("name",StringType(),True).add("food1",Integer
Type(),True).add("food2",IntegerType(),True).add("food3",IntegerType(),
True).add("food4",IntegerType(),True).add("placeid",IntegerType(),True)
>>>foodratings=spark.read.schema(tab1).csv('hdfs:///user/csp554/foodra
tings187912.txt')
```



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

foodratings.printSchema()

foodratings.show(5)

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:

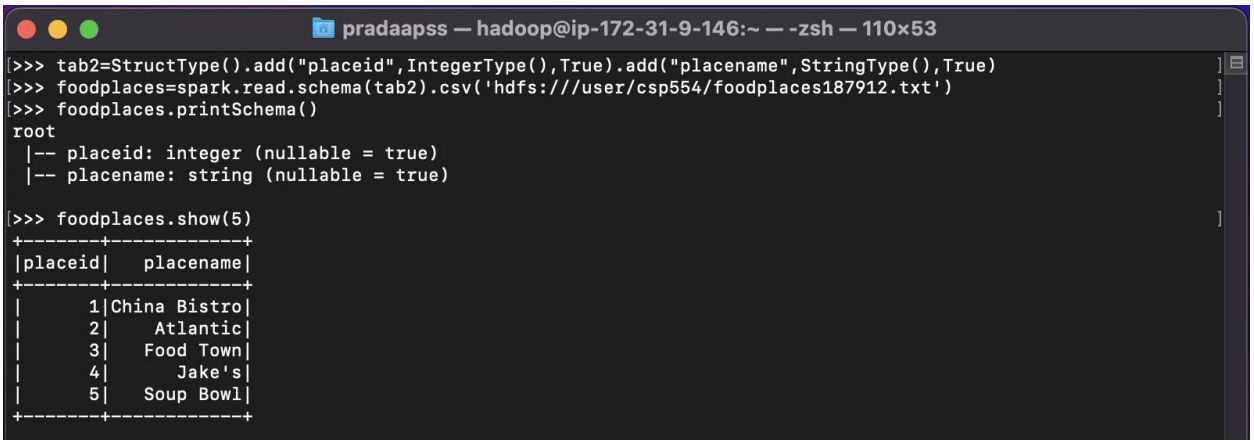
Field Nampee	Field Type
placeid	Integer
placename	String

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

```
>>>tab2=StructType().add("placeid",IntegerType(),True).add("placename",
StringType(),True)
>>>foodplaces=spark.read.schema(tab2).csv('hdfs:///user/csp554/foodpla
ces187912.txt')
```

foodplaces.printSchema()

foodplaces.show(5)

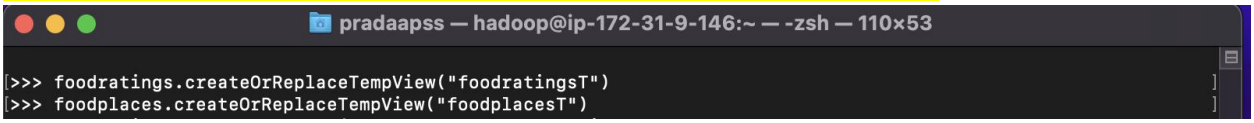


Exercise 3)

Step A

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

```
>>> foodratings.createOrReplaceTempView("foodratingsT")
>>> foodplaces.createOrReplaceTempView("foodplacesT")
```



Step B

Use a SQL query on the table “foodratingsT” to create a new DataFrame called foodratings_ex3a holding records which meet the following condition: food2 < 25 and food4 > 40. Remember, when defining conditions in your code use maximum parentheses.

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

```
>>> foodratings_ex3a=spark.sql("select * from foodratingsT where
food2<25 and food4>40")
```

foodratings_ex3a.printSchema()

foodratings_ex3a.show(5)

Step C

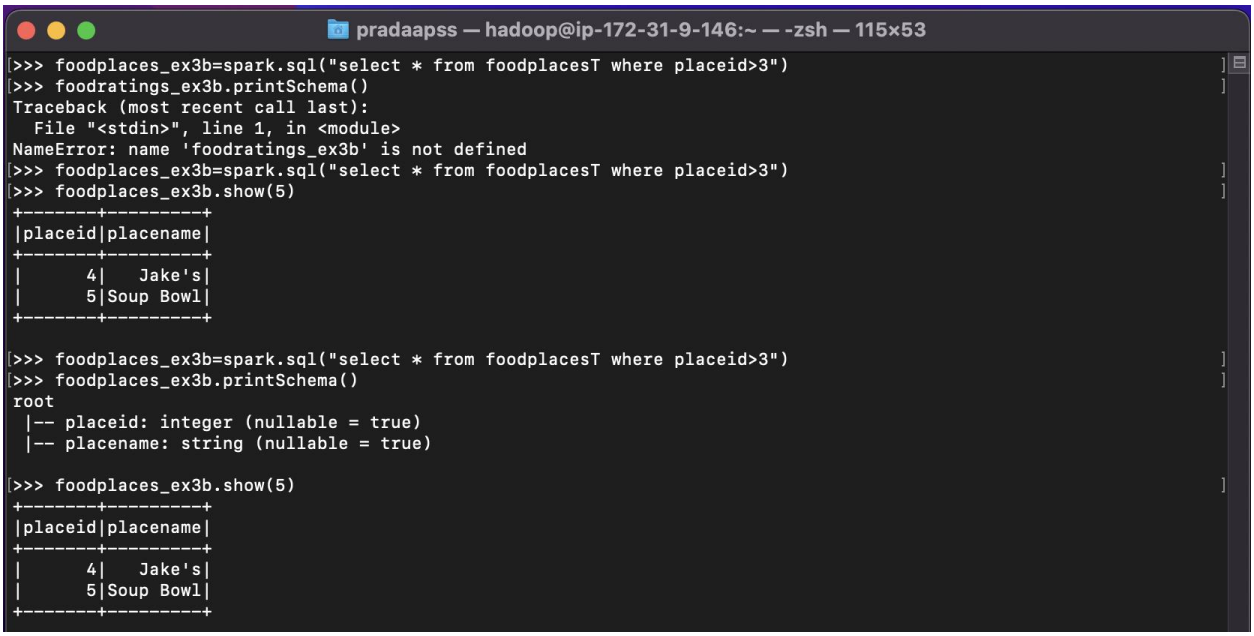
Use a SQL query on the table “foodplacesT” to create a new DataFrame called foodplaces_ex3b holding records which meet the following condition: placeid > 3

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

```
>>> foodplaces_ex3b=spark.sql("select * from foodplacesT where placeid> 3")
```

foodplaces_ex3b.printSchema()

foodplaces_ex3b.show(5)



Exercise 4)

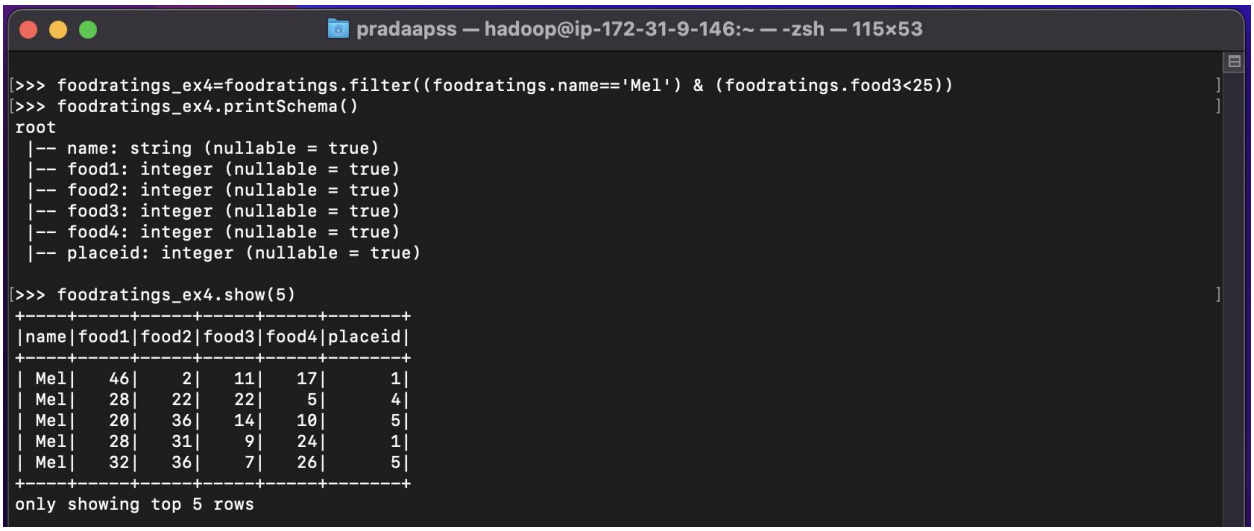
Use a transformation (not a SparkSQL query) on the DataFrame ‘foodratings’ created in exercise 1 to create a new DataFrame called foodratings_ex4 that includes only those records (rows) where the ‘name’ field is “Mel” and food3 < 25.

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

```
>>> foodratings_ex4=foodratings.filter((foodratings.name=='Mel') & (foodratings.food3<25))
```

foodratings_ex4.printSchema()

foodratings_ex4.show(5)



Exercise 5)

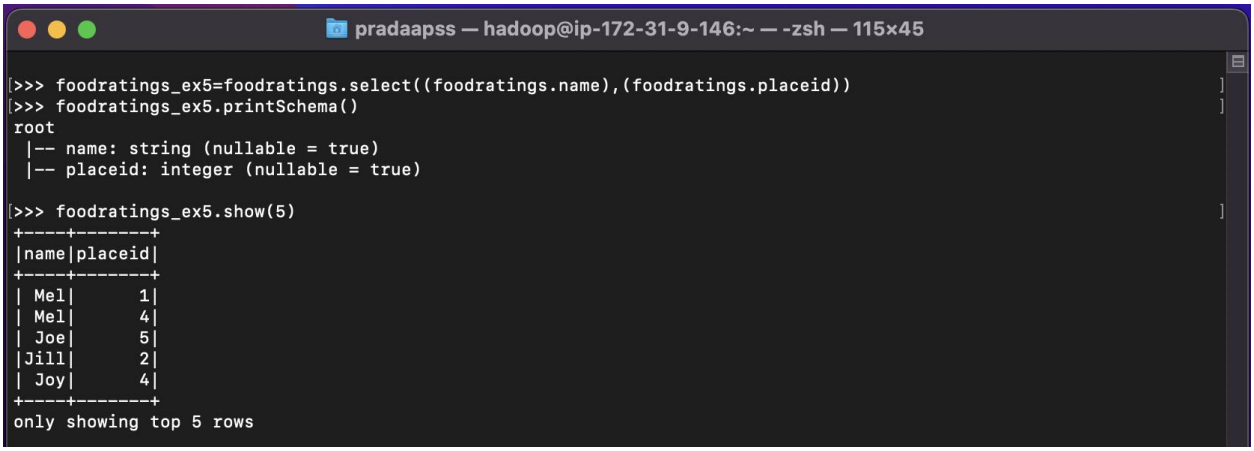
Use a transformation (not a SparkSQL query) on the DataFrame ‘foodratings’ created in exercise 1 to create a new DataFrame called foodratings_ex5 that includes only the columns (fields) ‘name’ and ‘placeid’

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

```
>>>foodratings_ex5=foodratings.select((foodratings.name),(foodratings.placeid))
```

foodratings_ex5.printSchema()

foodratings_ex5.show(5)



Exercise 6)

Use a transformation (not a SparkSQL query) 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=foodratings.join(foodplaces,foodratings.placeid==foodplaces.placeid,'inner').drop(foodplaces.placeid)
```

ex6.printSchema()

ex6.show(5)