Assignment-7
BIG DATA Technologies

EMR Connection:

```
Last login: Wed Sep 13 22:31:22 on ttys005
(base) sailavanyanarthu@Sailavanyas-MacBook-Air ~ % cd downloads
(base) sailavanyanarthu@Sailavanyas-MacBook-Air downloads % cd assignment7
(base) sailavanyanarthu@Sailavanyas-MacBook-Air assignment7 % chmod 400 keypair.pem
```

```
(base) sailavanyanarthu@Sailavanyas-MacBook-Air assignment7 % ssh -i keypair.pem hadoop@ec2-18-188-244-79.us-east-2.compute.amazonaws.com
The authenticity of host 'ec2-18-188-244-79.us-east-2.compute.amazonaws.com (18.188.244.79)' can't be established.
ED25519 key fingerprint is SHA256:ID5lAOJXYwWOQgTk/uTsOKBa0f9twIGyRcHyd88bgFg.
This key is not known by any other names
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'ec2-18-188-244-79.us-east-2.compute.amazonaws.com' (ED25519) to the list of known hosts.

       __|  __|_  )
       _|  (     /   Amazon Linux 2 AMI
      ___|\___|___|

https://aws.amazon.com/amazon-linux-2/

EEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::E M:::::::M         M:::::::M R::::::::::::::R
EE:::::EEEEEEEEE:::E M::::::::M       M::::::::M R:::::RRRRRR:::::R
  E::::E       EEEEE M:::::::::M     M:::::::::M RR::::R      R::::R
  E::::E             M::::::M:::M   M:::M::::::M   R:::R      R::::R
  E::::::EEEEEEEEEE   M:::::M M:::M M:::M M:::::M   R:::RRRRRR:::::R
  E::::::::::::::E   M:::::M  M:::M:::M  M:::::M   R:::::::::::RR
  E::::::EEEEEEEEEE   M:::::M   M:::::M   M:::::M   R:::RRRRRR::::R
  E::::E             M:::::M    M:::M    M:::::M   R:::R      R::::R
  E::::E       EEEEE M:::::M     MMM     M:::::M   R:::R      R::::R
EE:::::EEEEEEEE::::E M:::::M             M:::::M   R:::R      R::::R
E::::::::::::::::::E M:::::M             M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR
```

```
(base) sailavanyanarthu@Sailavanyas-MacBook-Air assignment7 % scp -i keypair.pem(base) sailavanyanarthu@Sailavanyas-MacBook-Air assignment7 % scp -i keypair.pem(base) sailavanyanarthu@Sailavanyas-MacBook-Air
assignment7 % scp -i keypair.pem TestDataGen.class hadoop@ec2-18-188-244-79.us-east-2.compute.amazonaws.com:/home/hadoop
TestDataGen.class                             100% 2189    54.2KB/s   00:00
(base) sailavanyanarthu@Sailavanyas-MacBook-Air assignment7 %
```

Exercise 1

Magic Number = 147006
hadoop fs -put /home/hadoop/foodplaces147006.txt /user/hadoop
hadoop fs -put /home/hadoop/foodratings147006.txt /user/hadoop

```
[hadoop@ip-172-31-11-194 ~]$ java TestDataGen
Magic Number = 147006
[hadoop@ip-172-31-11-194 ~]$ hadoop fs -put /home/hadoop/foodplaces147006.txt /user/hadoop
[hadoop@ip-172-31-11-194 ~]$ hadoop fs -put /home/hadoop/foodratings147006.txt /user/hadoop
[hadoop@ip-172-31-11-194 ~]$ ls
foodplaces147006.txt  foodratings147006.txt  TestDataGen.class
[hadoop@ip-172-31-11-194 ~]$
```

```
[hadoop@ip-172-31-11-194 ~]$ pyspark
Python 3.7.16 (default, Aug 30 2023, 20:37:53)
[GCC 7.3.1 20180712 (Red Hat 7.3.1-15)] on linux
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).
23/10/12 06:08:02 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries under SPARK_HOME.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 3.4.0-amzn-0
      /_/

Using Python version 3.7.16 (default, Aug 30 2023 20:37:53)
Spark context Web UI available at http://ip-172-31-11-194.us-east-2.compute.internal:4040
Spark context available as 'sc' (master = yarn, app id = application_1697089230538_0001).
SparkSession available as 'spark'.
>>>
>>>
```

from pyspark.sql.types import *
Table1 = StructType().add("name", StringType(), True).add("food1",IntegerType(), True).add("food2",IntegerType(), True).add("food3",IntegerType(), True).add("food4",IntegerType(), True).add("placeid",IntegerType(), True)

foodratings = spark.read.schema(Table1).csv(“/user/hadoop/foodratings147006.txt")

foodratings.printSchema()

foodratings.show(5)

```
>>>
>>> from pyspark.sql.types import*
>>> Table1 = StructType().add("name", StringType(), True).add("food1",IntegerType(), True).add("food2",IntegerType(), True).add("food3",IntegerType(), True).add("food4",IntegerType(), True).add("placeid",Inte
gerType(), True)
>>> foodratings= spark.read.schema(Table1).csv("/user/hadoop/foodratings147006.txt")
>>> foodratings.printSchema()
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)

>>> foodratings.show(5)
+----+-----+-----+-----+-----+-------+
|name|food1|food2|food3|food4|placeid|
+----+-----+-----+-----+-----+-------+
| Sam|    2|   50|   28|   16|      3|
| Joe|   22|   28|   40|   35|      2|
| Joy|   23|   27|   24|   43|      4|
| Sam|   40|   35|    9|   41|      1|
| Sam|   47|   43|   25|   34|      2|
+----+-----+-----+-----+-----+-------+
only showing top 5 rows

>>>
>>>
```

Exercise 2

Table2 = StructType().add("placeid", IntegerType(), True).add("placename", StringType(), True)

foodplaces = spark.read.schema(Table2).csv("/user/hadoop/foodplaces147006.txt")

foodplaces.printSchema()

foodplaces.show(5)

```
>>>
>>>
>>> Table2 = StructType().add("placeid", IntegerType(), True).add("placename", StringType(), True)
>>> foodplaces = spark.read.schema(Table2).csv("/user/hadoop/foodplaces147006.txt")
>>> foodplaces.printSchema()
root
 |-- placeid: integer (nullable = true)
 |-- placename: string (nullable = true)

>>> foodplaces.show(5)
+-------+------------+
|placeid|   placename|
+-------+------------+
|      1|China Bistro|
|      2|    Atlantic|
|      3|   Food Town|
|      4|      Jake's|
|      5|   Soup Bowl|
+-------+------------+
```

Exercise 3

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

foodratings_ex3a = spark.sql("SELECT * FROM foodratingsT WHERE food2 < 25 AND food4 > 40")
foodratings_ex3a.printSchema()
foodratings_ex3a.show(5)

```
>>>
>>>
>>> foodratings.createOrReplaceTempView("foodratingsT")
>>> foodplaces.createOrReplaceTempView("foodplacesT")
>>> foodratings_ex3a = spark.sql("SELECT * FROM foodratingsT WHERE food2 < 25 AND food4 > 40")
>>> foodratings_ex3a.printSchema()
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)

>>> foodratings_ex3a.show(5)
+----+-----+-----+-----+-----+-------+
|name|food1|food2|food3|food4|placeid|
+----+-----+-----+-----+-----+-------+
| Joy|   41|    2|   46|   47|      1|
|Jill|   27|   24|   43|   43|      4|
| Sam|   15|   17|   24|   42|      1|
| Joy|   30|   19|   38|   46|      4|
| Joy|   18|   21|   29|   50|      4|
+----+-----+-----+-----+-----+-------+
only showing top 5 rows
```

foodplaces_ex3b = spark.sql("SELECT * FROM foodplacesT WHERE placeid > 3")
foodplaces_ex3b.printSchema()
foodplaces_ex3b.show(5)

```
>>>
>>>
>>> foodplaces_ex3b = spark.sql("SELECT * FROM foodplacesT WHERE placeid > 3")
>>> foodplaces_ex3b.printSchema()
root
 |-- placeid: integer (nullable = true)
 |-- placename: string (nullable = true)

>>> foodplaces_ex3b.show(5)
+-------+---------+
|placeid|placename|
+-------+---------+
|      4|   Jake's|
|      5|Soup Bowl|
+-------+---------+
```

Exercise 4

foodratings_ex4 = foodratings.filter((foodratings["name"] == 'Mel') & (foodratings["food3"] < 25))
foodratings_ex4.printSchema()
foodratings_ex4.show(5)

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

>>> foodratings_ex4.show(5)
+----+-----+-----+-----+-----+-------+
|name|food1|food2|food3|food4|placeid|
+----+-----+-----+-----+-----+-------+
| Mel|   41|   43|   13|   17|      5|
| Mel|   14|    7|    8|   43|      4|
| Mel|   16|   23|   12|   33|      4|
| Mel|    3|   44|   17|   45|      2|
| Mel|   17|   26|   16|   24|      5|
+----+-----+-----+-----+-----+-------+
only showing top 5 rows

>>>
```

Exercise 5

foodratings_ex5 = foodratings.select(foodratings["name"], foodratings["placeid"])
foodratings_ex5.printSchema()
foodratings_ex5.show(5)

```
>>>
>>>
>>> foodratings_ex5 = foodratings.select(foodratings["name"], foodratings["placeid"])
>>> foodratings_ex5.printSchema()
root
 |-- name: string (nullable = true)
 |-- placeid: integer (nullable = true)

>>> foodratings_ex5.show(5)
+----+-------+
|name|placeid|
+----+-------+
| Sam|      3|
| Joe|      2|
| Joy|      4|
| Sam|      1|
| Sam|      2|
+----+-------+
only showing top 5 rows

>>>
```

Exercise 6

ex6 = foodratings.join(foodplaces, foodratings.placeid == foodplaces.placeid, "inner")
ex6.printSchema()
ex6.show(5)

```
>>>
>>>
>>> ex6 = foodratings.join(foodplaces, foodratings.placeid == foodplaces.placeid, "inner")
>>> ex6.printSchema()
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)

>>> ex6.show(5)
+----+-----+-----+-----+-----+-------+-------+------------+
|name|food1|food2|food3|food4|placeid|placeid|   placename|
+----+-----+-----+-----+-----+-------+-------+------------+
| Sam|    2|   50|   28|   16|      3|      3|   Food Town|
| Joe|   22|   28|   40|   35|      2|      2|    Atlantic|
| Joy|   23|   27|   24|   43|      4|      4|      Jake's|
| Sam|   40|   35|    9|   41|      1|      1|China Bistro|
| Sam|   47|   43|   25|   34|      2|      2|    Atlantic|
+----+-----+-----+-----+-----+-------+-------+------------+
only showing top 5 rows

>>>
```

Submitted by:
Sailavanya Narthu
A20516764
snarthu@hawk.iit.edu