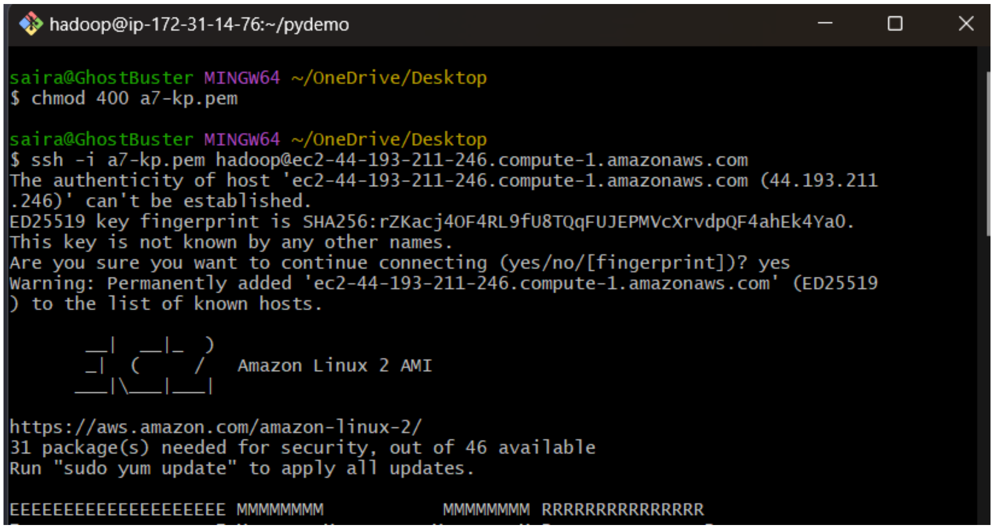
Assignment 7

A20522183

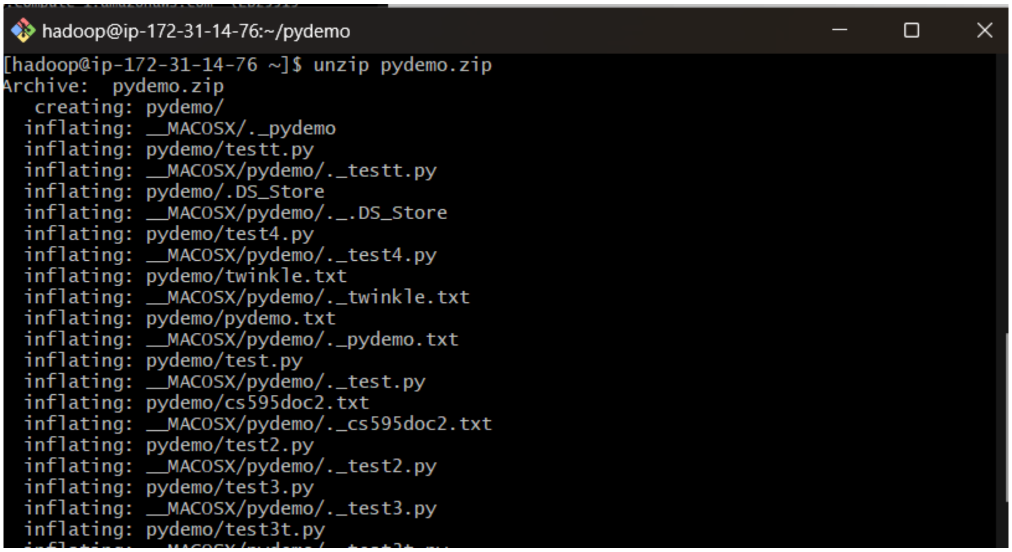
Oduri Sai Ram

Demo



```
saira@GhostBuster MINGW64 ~/OneDrive/Desktop
$ chmod 400 a7-kp.pem

saira@GhostBuster MINGW64 ~/OneDrive/Desktop
$ ssh -i a7-kp.pem hadoop@ec2-44-193-211-246.compute-1.amazonaws.com
The authenticity of host 'ec2-44-193-211-246.compute-1.amazonaws.com (44.193.211
.246)' can't be established.
ED25519 key fingerprint is SHA256:rZKacj4OF4RL9fU8TQqFUJEPMVcXrvdpQF4ahEk4Ya0.
This key is not known by any other names.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'ec2-44-193-211-246.compute-1.amazonaws.com' (ED25519
) to the list of known hosts.

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

https://aws.amazon.com/amazon-linux-2/
31 package(s) needed for security, out of 46 available
Run "sudo yum update" to apply all updates.

EEEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRR
```

Unzipping Pydemo

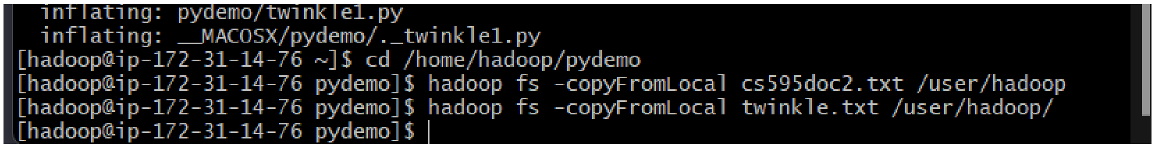


```
[hadoop@ip-172-31-14-76 ~]$ unzip pydemo.zip
Archive:  pydemo.zip
   creating: pydemo/
  inflating: __MACOSX/._pydemo
  inflating: pydemo/testt.py
  inflating: __MACOSX/pydemo/._testt.py
  inflating: pydemo/.DS_Store
  inflating: __MACOSX/pydemo/._.DS_Store
  inflating: pydemo/test4.py
  inflating: __MACOSX/pydemo/._test4.py
  inflating: pydemo/twinkle.txt
  inflating: __MACOSX/pydemo/._twinkle.txt
  inflating: pydemo/pydemo.txt
  inflating: __MACOSX/pydemo/._pydemo.txt
  inflating: pydemo/test.py
  inflating: __MACOSX/pydemo/._test.py
  inflating: pydemo/cs595doc2.txt
  inflating: __MACOSX/pydemo/._cs595doc2.txt
  inflating: pydemo/test2.py
  inflating: __MACOSX/pydemo/._test2.py
  inflating: pydemo/test3.py
  inflating: __MACOSX/pydemo/._test3.py
  inflating: pydemo/test3t.py
```

Demo instructions for Pydemo

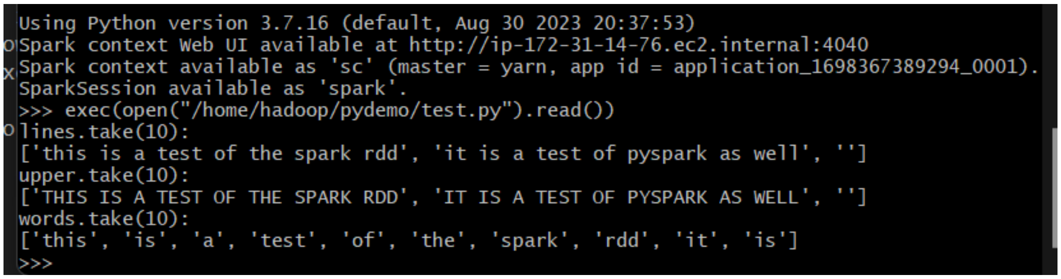
Hadoop fs -copyfromlocal cs595doc2.txt /user/hadoop

Hadoop fs -copyfromlocal twinkle.txt /user/hadoop

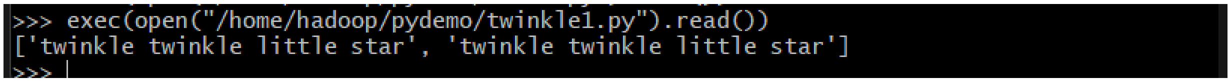


```
  inflating: pydemo/twinkle1.py
  inflating: __MACOSX/pydemo/._twinkle1.py
[hadoop@ip-172-31-14-76 ~]$ cd /home/hadoop/pydemo
[hadoop@ip-172-31-14-76 pydemo]$ hadoop fs -copyFromLocal cs595doc2.txt /user/hadoop
[hadoop@ip-172-31-14-76 pydemo]$ hadoop fs -copyFromLocal twinkle.txt /user/hadoop/
[hadoop@ip-172-31-14-76 pydemo]$ 
```

Examining the contents of the text.py file

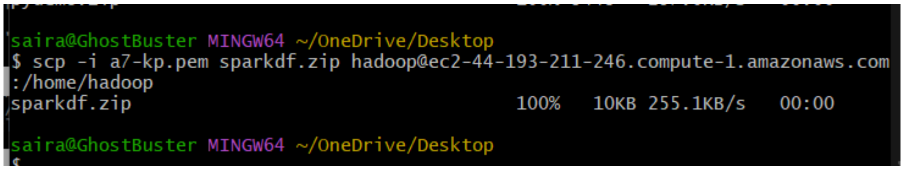


```
Using Python version 3.7.16 (default, Aug 30 2023 20:37:53)
Spark context Web UI available at http://ip-172-31-14-76.ec2.internal:4040
Spark context available as 'sc' (master = yarn, app id = application_1698367389294_0001).
SparkSession available as 'spark'.
>>> exec(open("/home/hadoop/pydemo/test.py").read())
lines.take(10):
['this is a test of the spark rdd', 'it is a test of pyspark as well', '']
upper.take(10):
['THIS IS A TEST OF THE SPARK RDD', 'IT IS A TEST OF PYSPARK AS WELL', '']
words.take(10):
['this', 'is', 'a', 'test', 'of', 'the', 'spark', 'rdd', 'it', 'is']
>>>
```

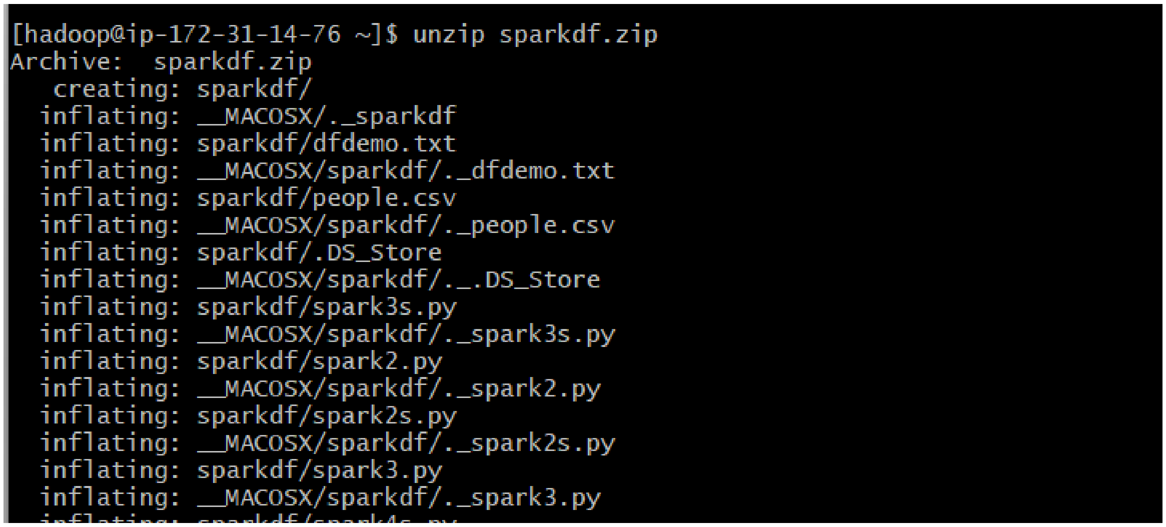


```
>>> exec(open("/home/hadoop/pydemo/twinkle1.py").read())
['twinkle twinkle little star', 'twinkle twinkle little star']
>>> 
```

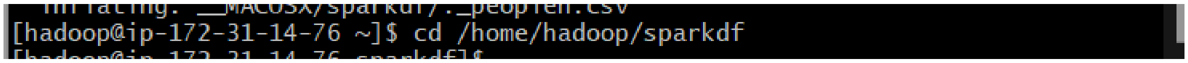
Uzipping Sparkdf



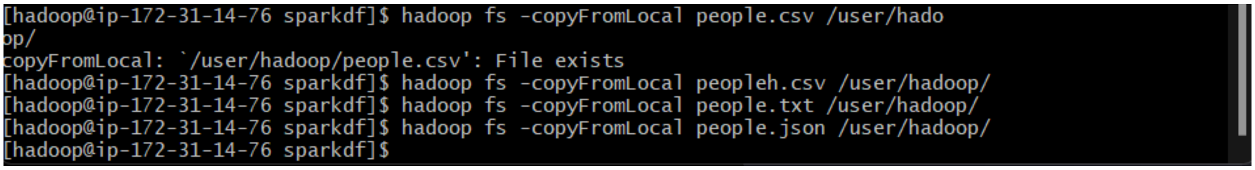
```
saira@GhostBuster MINGW64 ~/OneDrive/Desktop
$ scp -i a7-kp.pem sparkdf.zip hadoop@ec2-44-193-211-246.compute-1.amazonaws.com
:/home/hadoop
sparkdf.zip                                   100%   10KB 255.1KB/s   00:00

saira@GhostBuster MINGW64 ~/OneDrive/Desktop
$
```



```
[hadoop@ip-172-31-14-76 ~]$ unzip sparkdf.zip
Archive:  sparkdf.zip
   creating: sparkdf/
  inflating: __MACOSX/._sparkdf
  inflating: sparkdf/dfdemo.txt
  inflating: __MACOSX/sparkdf/._dfdemo.txt
  inflating: sparkdf/people.csv
  inflating: __MACOSX/sparkdf/._people.csv
  inflating: sparkdf/.DS_Store
  inflating: __MACOSX/sparkdf/._.DS_Store
  inflating: sparkdf/spark3s.py
  inflating: __MACOSX/sparkdf/._spark3s.py
  inflating: sparkdf/spark2.py
  inflating: __MACOSX/sparkdf/._spark2.py
  inflating: sparkdf/spark2s.py
  inflating: __MACOSX/sparkdf/._spark2s.py
  inflating: sparkdf/spark3.py
  inflating: __MACOSX/sparkdf/._spark3.py
  inflating: sparkdf/spark4s.py
```

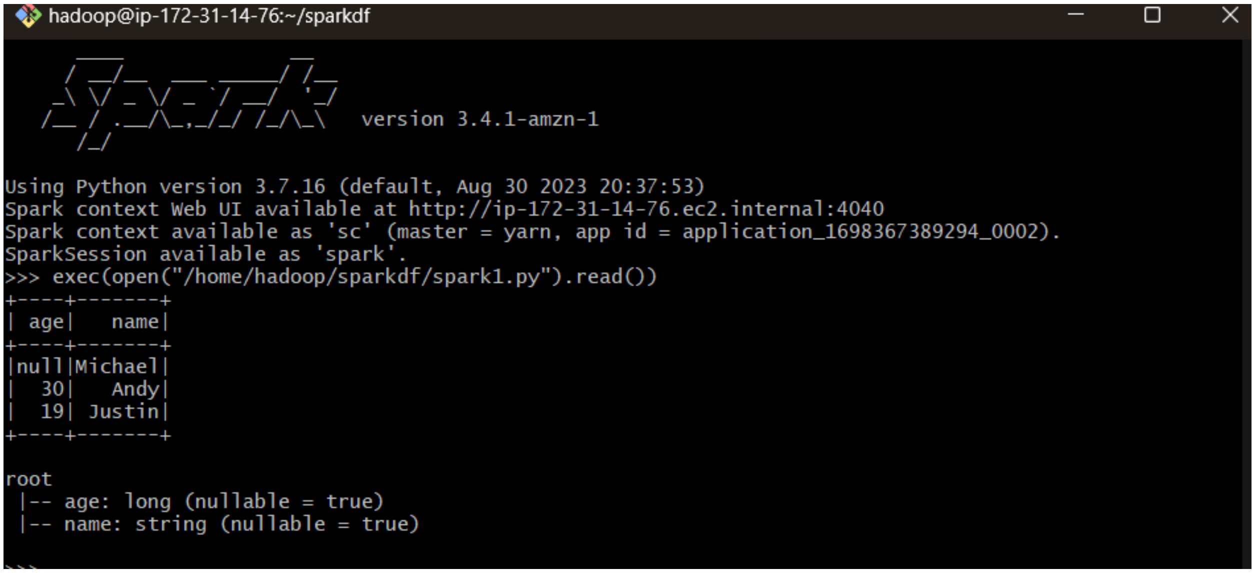


```
  inflating: __MACOSX/sparkdf/._people1.csv
[hadoop@ip-172-31-14-76 ~]$ cd /home/hadoop/sparkdf
[hadoop@ip-172-31-14-76 sparkdf]$
```



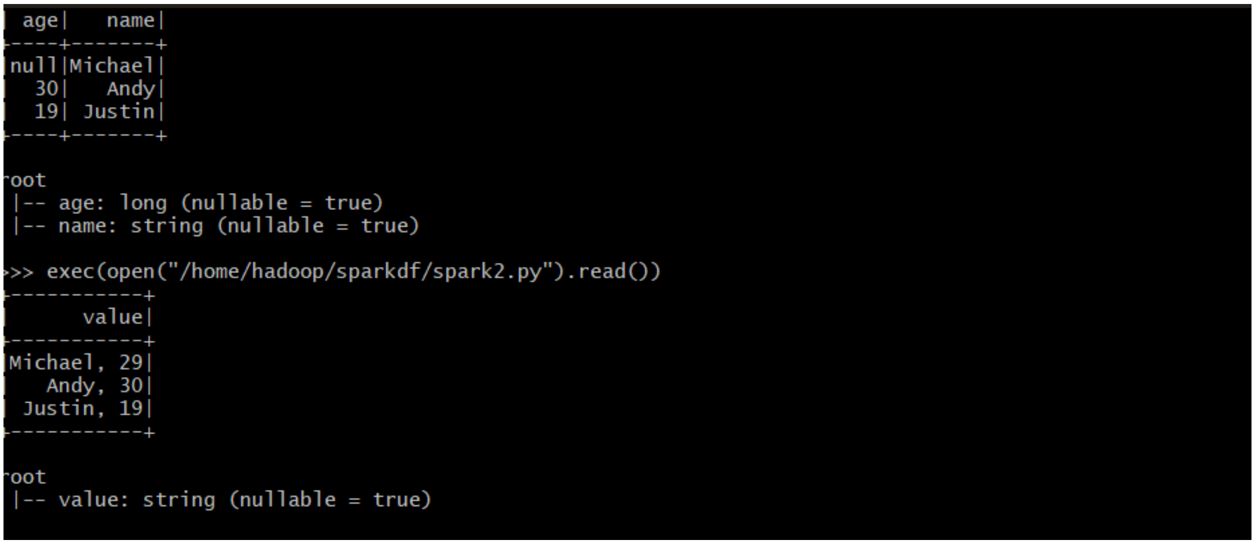
```
[hadoop@ip-172-31-14-76 sparkdf]$ hadoop fs -copyFromLocal people.csv /user/hado
op/
copyFromLocal: `/user/hadoop/people.csv': File exists
[hadoop@ip-172-31-14-76 sparkdf]$ hadoop fs -copyFromLocal peopleh.csv /user/hadoop/
[hadoop@ip-172-31-14-76 sparkdf]$ hadoop fs -copyFromLocal people.txt /user/hadoop/
[hadoop@ip-172-31-14-76 sparkdf]$ hadoop fs -copyFromLocal people.json /user/hadoop/
[hadoop@ip-172-31-14-76 sparkdf]$
```

Spark read datasets

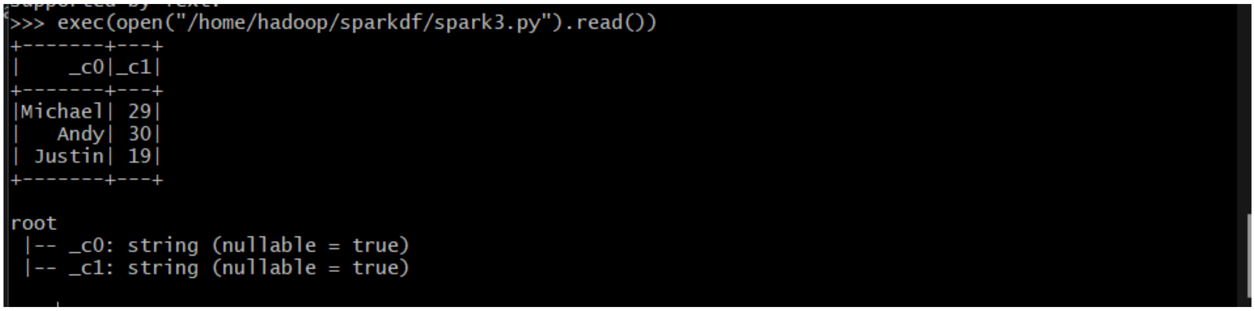


hadoop@ip-172-31-14-76:~/sparkdf

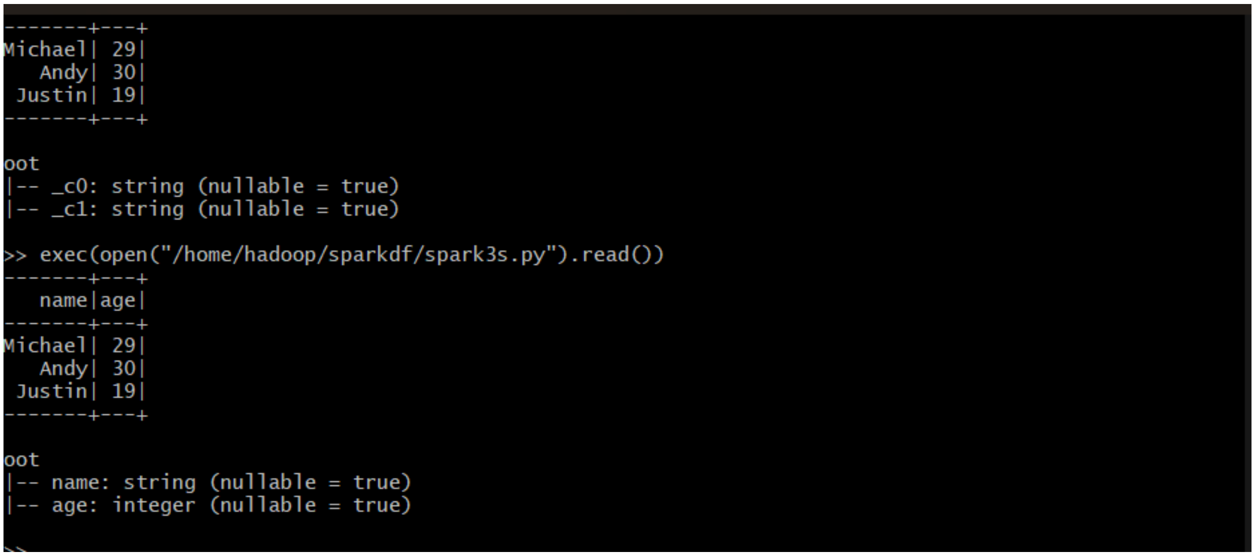
```
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 3.4.1-amzn-1
      /_/

Using Python version 3.7.16 (default, Aug 30 2023 20:37:53)
Spark context Web UI available at http://ip-172-31-14-76.ec2.internal:4040
Spark context available as 'sc' (master = yarn, app id = application_1698367389294_0002).
SparkSession available as 'spark'.
>>> exec(open("/home/hadoop/sparkdf/spark1.py").read())
+----+-------+
| age|   name|
+----+-------+
|null|Michael|
|  30|   Andy|
|  19| Justin|
+----+-------+

root
 |-- age: long (nullable = true)
 |-- name: string (nullable = true)

>>>
```



```
 age|   name|
----+-------+
null|Michael|
  30|   Andy|
  19| Justin|
----+-------+

oot
 |-- age: long (nullable = true)
 |-- name: string (nullable = true)

>> exec(open("/home/hadoop/sparkdf/spark2.py").read())
-----------+
      value|
-----------+
Michael, 29|
   Andy, 30|
 Justin, 19|
-----------+

oot
 |-- value: string (nullable = true)
```



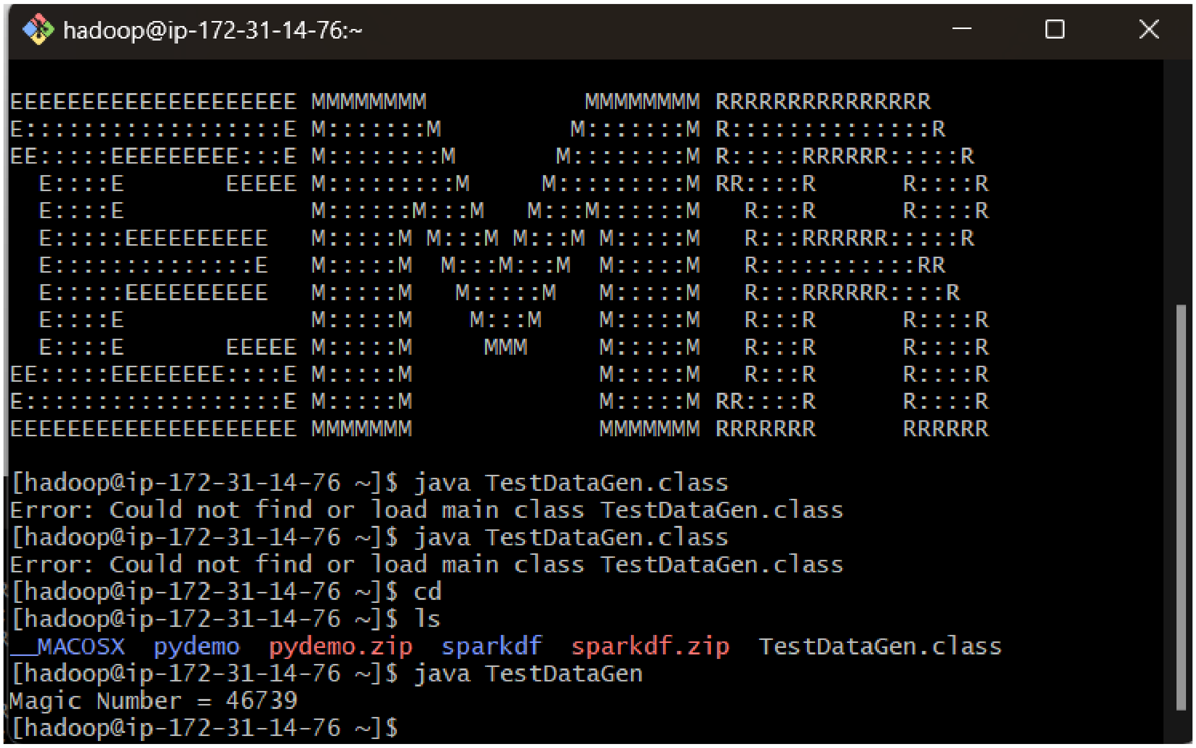
```
>>> exec(open("/home/hadoop/sparkdf/spark3.py").read())
+-------+---+
|    _c0|_c1|
+-------+---+
|Michael| 29|
|   Andy| 30|
| Justin| 19|
+-------+---+

root
 |-- _c0: string (nullable = true)
 |-- _c1: string (nullable = true)
```



```
-------+---+
Michael| 29|
   Andy| 30|
 Justin| 19|
-------+---+

oot
|-- _c0: string (nullable = true)
|-- _c1: string (nullable = true)

>> exec(open("/home/hadoop/sparkdf/spark3s.py").read())
-------+---+
   name|age|
-------+---+
Michael| 29|
   Andy| 30|
 Justin| 19|
-------+---+

oot
|-- name: string (nullable = true)
|-- age: integer (nullable = true)
```

Exercises

Reading Test data gen class



hadoop@ip-172-31-14-76:~

```
EEEEEEEEEEEEEEEEEEEE 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
EEEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-14-76 ~]$ java TestDataGen.class
Error: Could not find or load main class TestDataGen.class
[hadoop@ip-172-31-14-76 ~]$ java TestDataGen.class
Error: Could not find or load main class TestDataGen.class
[hadoop@ip-172-31-14-76 ~]$ cd
[hadoop@ip-172-31-14-76 ~]$ ls
__MACOSX  pydemo  pydemo.zip  sparkdf  sparkdf.zip  TestDataGen.class
[hadoop@ip-172-31-14-76 ~]$ java TestDataGen
Magic Number = 46739
[hadoop@ip-172-31-14-76 ~]$
```

1 Ans.

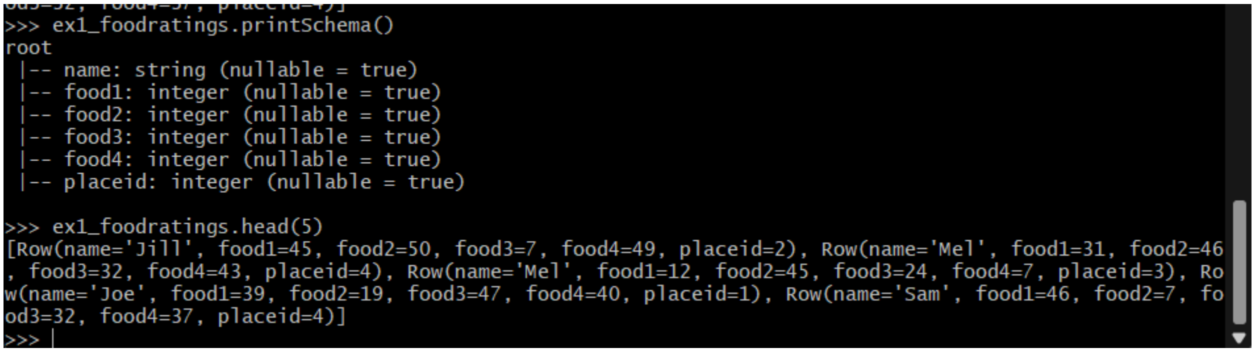
Code :

from pyspark.sql.types import *

st1 = StructType().add("placeid", IntegerType(), True).add("placename", StringType(), True) foodplaces = spark.read.schema(struct1).csv('/user/maria_dev/foodplaces41641.txt')

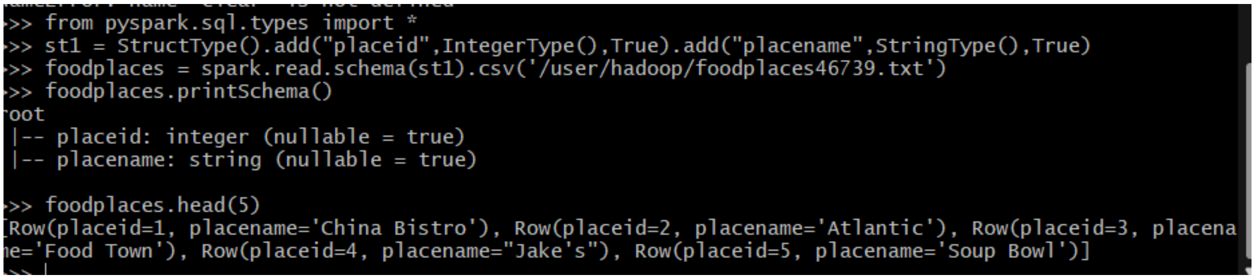
foodplaces.printSchema()

foodplaces.head(5)



```
od3=32, food4=37, placeid=4)]
>>> ex1_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)

>>> ex1_foodratings.head(5)
[Row(name='Jill', food1=45, food2=50, food3=7, food4=49, placeid=2), Row(name='Mel', food1=31, food2=46
, food3=32, food4=43, placeid=4), Row(name='Mel', food1=12, food2=45, food3=24, food4=7, placeid=3), Ro
w(name='Joe', food1=39, food2=19, food3=47, food4=40, placeid=1), Row(name='Sam', food1=46, food2=7, fo
od3=32, food4=37, placeid=4)]
>>> |
```

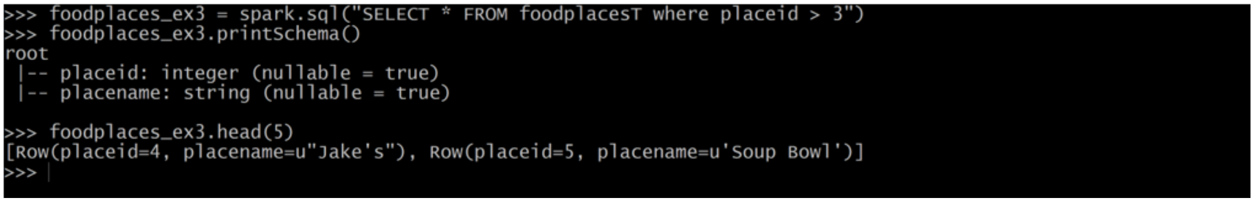
2Ans.



```
>> from pyspark.sql.types import *
>> st1 = StructType().add("placeid",IntegerType(),True).add("placename",StringType(),True)
>> foodplaces = spark.read.schema(st1).csv('/user/hadoop/foodplaces46739.txt')
>> foodplaces.printSchema()
oot
 |-- placeid: integer (nullable = true)
 |-- placename: string (nullable = true)

>> foodplaces.head(5)
Row(placeid=1, placename='China Bistro'), Row(placeid=2, placename='Atlantic'), Row(placeid=3, placena
e='Food Town'), Row(placeid=4, placename="Jake's"), Row(placeid=5, placename='Soup Bowl')]
```

3Ans.

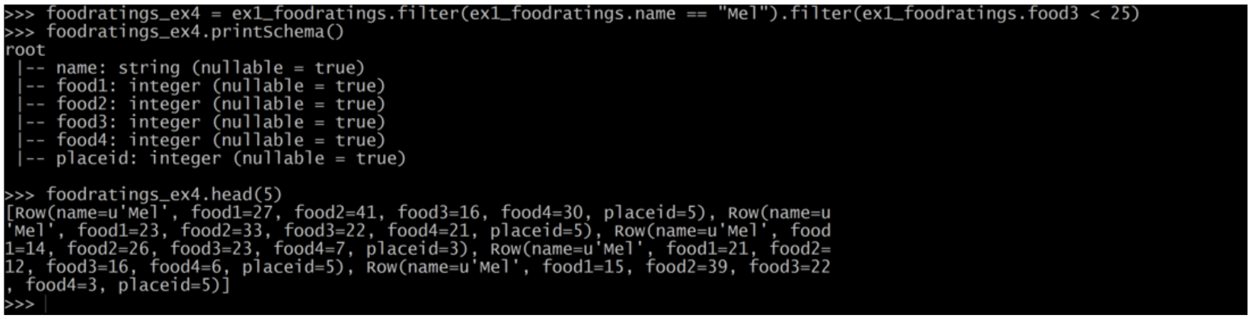
foodplaces_ex3 = spark.sql("SELECT * from foodplacesT where placeid > 3")



```
>>> foodplaces_ex3 = spark.sql("SELECT * FROM foodplacesT where placeid > 3")
>>> foodplaces_ex3.printSchema()
root
 |-- placeid: integer (nullable = true)
 |-- placename: string (nullable = true)

>>> foodplaces_ex3.head(5)
[Row(placeid=4, placename=u"Jake's"), Row(placeid=5, placename=u'Soup Bowl')]
>>>
```

4Ans.

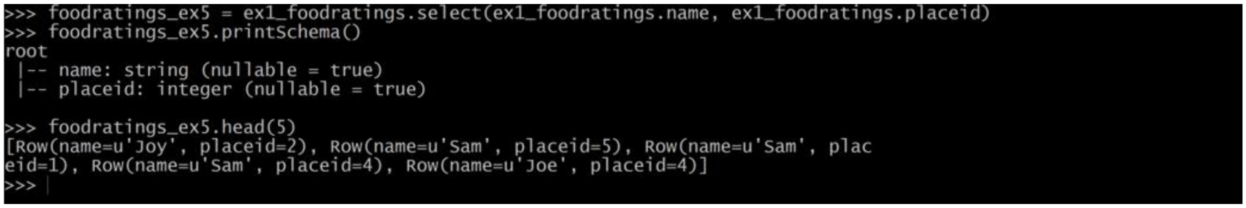
foodratings_ex4 = ex1_foodratings.filter(ex1_foodratings.name == "Mel").filter(ex1_foodratings.food3 < 25)



```
>>> foodratings_ex4 = ex1_foodratings.filter(ex1_foodratings.name == "Mel").filter(ex1_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.head(5)
[Row(name=u'Mel', food1=27, food2=41, food3=16, food4=30, placeid=5), Row(name=u
'Mel', food1=23, food2=33, food3=22, food4=21, placeid=5), Row(name=u'Mel', food
1=14, food2=26, food3=23, food4=7, placeid=3), Row(name=u'Mel', food1=21, food2=
12, food3=16, food4=6, placeid=5), Row(name=u'Mel', food1=15, food2=39, food3=22
, food4=3, placeid=5)]
>>>
```

5ans.

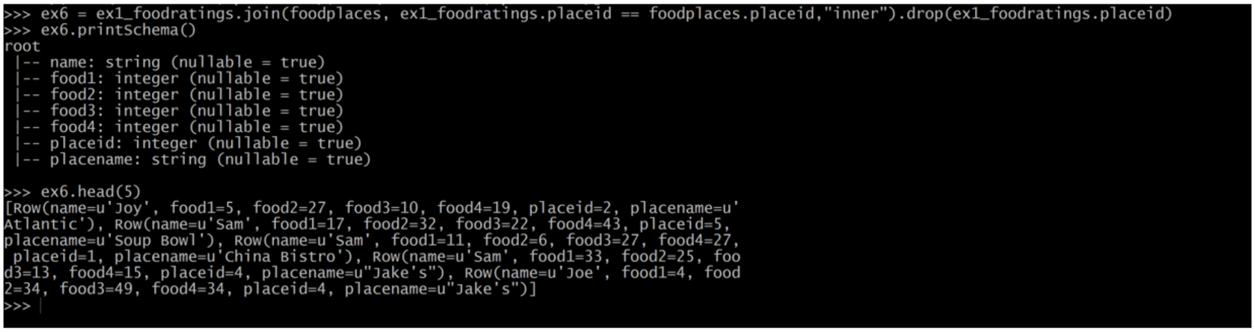
foodratings_ex5 = ex1_foodratings.select(ex1_foodratings.name, ex1_foodratings.placeid)



```
>>> foodratings_ex5 = ex1_foodratings.select(ex1_foodratings.name, ex1_foodratings.placeid)
>>> foodratings_ex5.printSchema()
root
 |-- name: string (nullable = true)
 |-- placeid: integer (nullable = true)

>>> foodratings_ex5.head(5)
[Row(name=u'Joy', placeid=2), Row(name=u'Sam', placeid=5), Row(name=u'Sam', plac
eid=1), Row(name=u'Sam', placeid=4), Row(name=u'Joe', placeid=4)]
>>> 
```

6Ans.

ex6 = ex1_foodratings.join(foodplaces, ex1_foodratings.placeid == foodplaces.placeid, "inner").drop(ex1_foodratings.placeid)



```
>>> ex6 = ex1_foodratings.join(foodplaces, ex1_foodratings.placeid == foodplaces.placeid,"inner").drop(ex1_foodratings.placeid)
>>> 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)
 |-- placename: string (nullable = true)

>>> ex6.head(5)
[Row(name=u'Joy', food1=5, food2=27, food3=10, food4=19, placeid=2, placename=u'
Atlantic'), Row(name=u'Sam', food1=17, food2=32, food3=22, food4=43, placeid=5,
placename=u'Soup Bowl'), Row(name=u'Sam', food1=11, food2=6, food3=27, food4=27,
 placeid=1, placename=u'China Bistro'), Row(name=u'Sam', food1=33, food2=25, foo
d3=13, food4=15, placeid=4, placename=u"Jake's"), Row(name=u'Joe', food1=4, food
2=34, food3=49, food4=34, placeid=4, placename=u"Jake's")]
>>> 
```