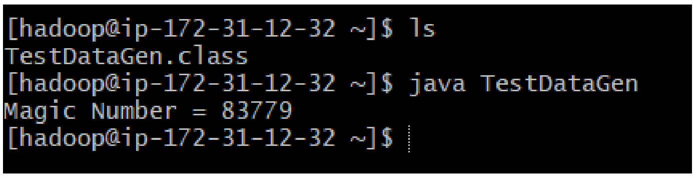
Exercise 1)

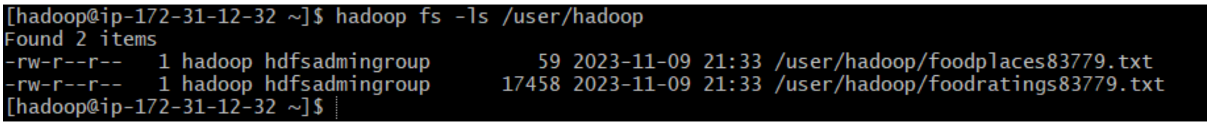
Use the TestDataGen program from previous assignments to generate a new foodratings.txt data file. Copy the file to HDFS, say into the /user/hadoop directory. Read in the text file into an RDD named ex1RDD. This RDD should now have records each consisting of a single string having 6 comma-separated parts. List the first five records of the RDD using the “take(5)” action and copy them and the “magic number to your assignment submission for this exercise.



```
[hadoop@ip-172-31-12-32 ~]$ ls
TestDataGen.class
[hadoop@ip-172-31-12-32 ~]$ java TestDataGen
Magic Number = 83779
[hadoop@ip-172-31-12-32 ~]$
```

Command:

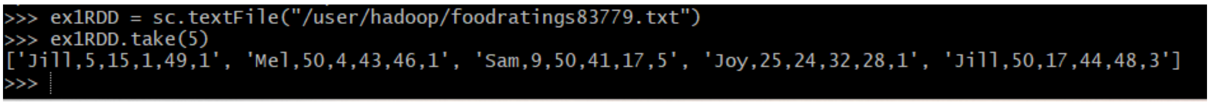
hadoop fs -ls /user/hadoop



```
[hadoop@ip-172-31-12-32 ~]$ hadoop fs -ls /user/hadoop
Found 2 items
-rw-r--r--   1 hadoop hdfsadmingroup         59 2023-11-09 21:33 /user/hadoop/foodplaces83779.txt
-rw-r--r--   1 hadoop hdfsadmingroup      17458 2023-11-09 21:33 /user/hadoop/foodratings83779.txt
[hadoop@ip-172-31-12-32 ~]$
```

Commands:

ex1RDD = sc.textfile(“/user/hadoop/foodratings83779.txt”)
ex1RDD.take(5)



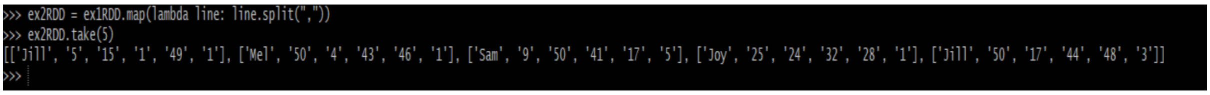
```
>>> ex1RDD = sc.textFile("/user/hadoop/foodratings83779.txt")
>>> ex1RDD.take(5)
['Jill,5,15,1,49,1', 'Mel,50,4,43,46,1', 'Sam,9,50,41,17,5', 'Joy,25,24,32,28,1', 'Jill,50,17,44,48,3']
>>>
```

Exercise 2)

Create another RDD called ex2RDD where each record of this new RDD has 6 fields, each a string, by splitting apart each record on “,” boundaries from the ex1RDD. List the first five records of this RDD using the “take(5)” action and copy them to your assignment submission for this exercise.

Command:

ex2RDD = ex1RDD.map(lambda line: line.split(“,”))
ex2RDD.take(5)



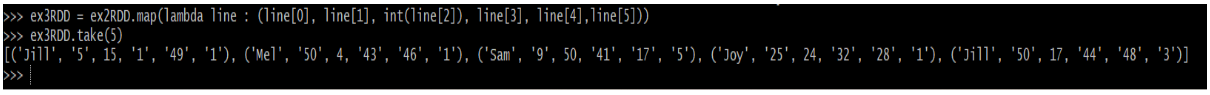
```
>>> ex2RDD = ex1RDD.map(lambda line: line.split(","))
>>> ex2RDD.take(5)
[['Jill', '5', '15', '1', '49', '1'], ['Mel', '50', '4', '43', '46', '1'], ['Sam', '9', '50', '41', '17', '5'], ['Joy', '25', '24', '32', '28', '1'], ['Jill', '50', '17', '44', '48', '3']]
>>>
```

Exercise 3)

Create another RDD called ex3RDD from ex2RDD where each record of this new RDD has its third column converted from a string to an integer.
Hint: Use a lambda function something like the following: lambda line : [line[0], line[1], int(line[2]), line[3], line[4], line[5]] List the first five records of this RDD using the “take(5)” action and copy them to your assignment submission for this exercise.

Commands:

```
ex3RDD = ex2RDD.map(lambda line : (line[0], line[1], line[2], line[3], line[4],line[5]))
ex3RDD.take(5)
```



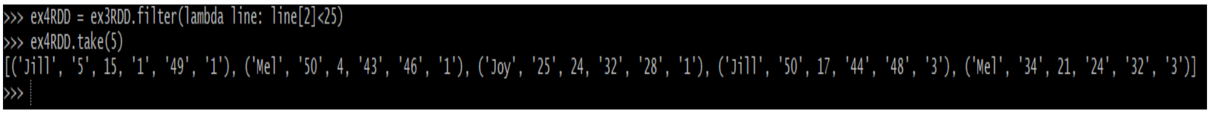
```
>>> ex3RDD = ex2RDD.map(lambda line : (line[0], line[1], int(line[2]), line[3], line[4],line[5]))
>>> ex3RDD.take(5)
[('Jill', '5', 15, '1', '49', '1'), ('Mel', '50', 4, '43', '46', '1'), ('Sam', '9', 50, '41', '17', '5'), ('Joy', '25', 24, '32', '28', '1'), ('Jill', '50', 17, '44', '48', '3')]
>>>
```

Exercise 4)

Create another RDD called ex4RDD from ex3RDD where each record of this new RDD is allowed to have a value for its third field that is less than 25 (<25). List the first five records of this RDD using the “take(5)” action and copy them to your assignment submission for this exercise.

Commands:

```
ex4RDD = ex3RDD.filter(lambda line: line[2]<25)
ex4RDD.take(5)
```



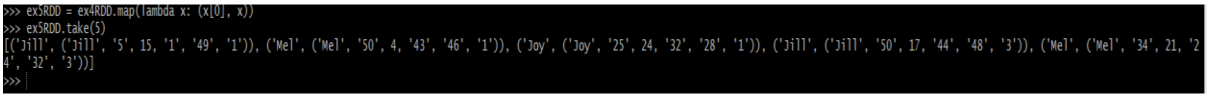
```
>>> ex4RDD = ex3RDD.filter(lambda line: line[2]<25)
>>> ex4RDD.take(5)
[('Jill', '5', 15, '1', '49', '1'), ('Mel', '50', 4, '43', '46', '1'), ('Joy', '25', 24, '32', '28', '1'), ('Jill', '50', 17, '44', '48', '3'), ('Mel', '34', 21, '24', '32', '3')]
>>>
```

Exercise 5)

Create another RDD called ex5RDD from ex4RDD where each record is a key value pair where the key is the first field of the record and the value is the entire record. List the first five records of this RDD using the “take(5)” action and copy them to your assignment submission for this exercise.

Commands:

```
ex5RDD = ex4RDD.map(lambda x: (x[0], x))
ex5RDD.take(5)
```



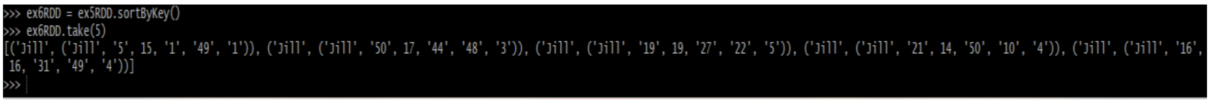
```
>>> ex5RDD = ex4RDD.map(lambda x: (x[0], x))
>>> ex5RDD.take(5)
[('Jill', ('Jill', '5', 15, '1', '49', '1')), ('Mel', ('Mel', '50', 4, '43', '46', '1')), ('Joy', ('Joy', '25', 24, '32', '28', '1')), ('Jill', ('Jill', '50', 17, '44', '48', '3')), ('Mel', ('Mel', '34', 21, '2
4', '32', '3'))]
>>>
```

Exercise 6)

Create another RDD called ex6RDD from ex5RDD where the records are organized in ascending order by key. List the first five records of this RDD using the “take(5)” action and copy them to your assignment submission for this exercise.

Commands:

ex6RDD = ex5RDD.sortByKey()
ex6RDD.take(5)



```
>>> ex6RDD = ex5RDD.sortByKey()
>>> ex6RDD.take(5)
[('Jill', ('Jill', '5', 15, '1', '49', '1')), ('Jill', ('Jill', '50', 17, '44', '48', '3')), ('Jill', ('Jill', '19', 19, '27', '22', '5')), ('Jill', ('Jill', '21', 14, '50', '10', '4')), ('Jill', ('Jill', '16',
 16, '31', '49', '4'))]
>>>
```