

Rakshith Churchagundi Amarnath

A20424771

## Assignment 7

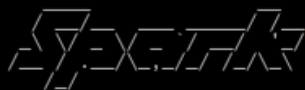
1)

Magic Number = 230700

```
[hadoop@ip-172-31-4-17 ~]$ java TestDataGen
Magic Number = 230700
[hadoop@ip-172-31-4-17 ~]$
```

```
[hadoop@ip-172-31-4-17 ~]$ hadoop fs -copyFromLocal foodratings230700.txt /user/hadoop
[hadoop@ip-172-31-4-17 ~]$ hadoop fs -copyFromLocal foodplaces230700.txt /user/hadoop
[hadoop@ip-172-31-4-17 ~]$ hadoop fs -ls /user/hadoop
Found 3 items
drwxr-xr-x - hadoop hadoop          0 2020-03-02 18:38 /user/hadoop/.sparkStaging
-rw-r--r-- 1 hadoop hadoop          59 2020-03-02 18:40 /user/hadoop/foodplaces230700.txt
-rw-r--r-- 1 hadoop hadoop    17449 2020-03-02 18:39 /user/hadoop/foodratings230700.txt
[hadoop@ip-172-31-4-17 ~]$
```

```
[hadoop@ip-172-31-4-17 ~]$ pyspark
Python 2.7.16 (default, Oct 14 2019, 21:26:56)
[GCC 4.8.5 20150623 (Red Hat 4.8.5-28)] 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).
20/03/02 18:34:13 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries under SPARK_HOME.
Welcome to
```

 version 2.4.4

```
Using Python version 2.7.16 (default, Oct 14 2019 21:26:56)
SparkSession available as 'spark'.
>>>
```

```
>>> ex1RDD=sc.textFile("./foodratings230700.txt")
>>> print ex1RDD.take(5)
```

```
>>> ex1RDD=sc.textFile("./foodratings230700.txt")
>>> print ex1RDD.take(5)
[u'Jill,32,50,24,29,4', u'Joe,29,42,7,6,4', u'Joy,19,43,39,12,1', u'Jill,21,18,45,15,1', u'Mel,16,20,1,26,5']
>>>
```

2)

```
>>> ex2RDD = ex1RDD.map(lambda line: line.split(","))
```

```
>>> print ex2RDD.take(5)
```

```
>>> ex2RDD = ex1RDD.map(lambda line: line.split(","))
>>> print ex2RDD.take(5)
[[u'Jill', u'32', u'50', u'24', u'29', u'4'], [u'Joe', u'29', u'42', u'7', u'6', u'4'], [u'Joy', u'19', u'43', u'39', u'12', u'1'], [u'Jill', u'21', u'18', u'45', u'15', u'1'], [u'Mel', u'16', u'20', u'1', u'26', u'5']]
>>>
```

3)

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

```
>>> print ex3RDD.take(5)
```

```
>>> ex3RDD = ex2RDD.map(lambda line : [line[0], line[1], int(line[2]), line[3], line[4], line[5]])
>>> print ex3RDD.take(5)
[[('Jill', '32', 50, '24', '29', '4'), ('Joe', '29', 42, '7', '6', '4'), ('Joy', '19', 43, '39', '12', '1'), ('Jill', '21', 18, '45', '15', '1'), ('Mel', '16', 20, '1', '26', '5')]
>>>
```

4)

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

```
>>> print ex4RDD.take(5)
```

```
>>> ex4RDD=ex3RDD.filter(lambda line: line[2]<25)
>>> print ex4RDD.take(5)
[[('Jill', '21', 18, '45', '15', '1'), ('Mel', '16', 20, '1', '26', '5'), ('Sam', '7', 9, '7', '40', '2'), ('Joe', '36', 14, '47', '11', '1'), ('Sam', '33', 15, '39', '11', '4')]
>>>
```

5) &gt;&gt;&gt; ex5RDD=ex4RDD.map(lambda x:(x[0], x))

```
>>> print ex5RDD.take(5)
```

```
>>> ex5RDD=ex4RDD.map(lambda x:(x[0], x))
...
>>> print ex5RDD.take(5)
[(('Jill', ['Jill', '21', 18, '45', '15', '1']), ('Mel', ['Mel', '16', 20, '1', '26', '5']), ('Sam', ['Sam', '7', 9, '7', '40', '2']), ('Joe', ['Joe', '36', 14, '47', '11', '1']), ('Sam', ['Sam', '33', 15, '39', '11', '4'])]
>>>
```

6)

```
>>> ex6RDD=ex5RDD.sortByKey()
```

```
>>> print ex5RDD.take(5)
```

```
>>> ex6RDD=ex5RDD.sortByKey()
>>> print ex5RDD.take(5)
[(('Jill', ['Jill', '21', 18, '45', '15', '1']), ('Mel', ['Mel', '16', 20, '1', '26', '5']), ('Sam', ['Sam', '7', 9, '7', '40', '2']), ('Joe', ['Joe', '36', 14, '47', '11', '1']), ('Sam', ['Sam', '33', 15, '39', '11', '4'])]
>>>
```