CWID: A20526972

CSP-554 Big Data Technologies

Bairi Rohith Reddy- Assignment #6

Java TestDataGen

Magic Number: 228345

hadoop fs -put foodratings228345.txt /user/hadoop

```
hadoop@ip-172-31-60-188:~
[hadoop@ip-172-31-60-188 ~]$
[hadoop@ip-172-31-60-188 ~]$ ls
foodplaces228345.txt  foodratings228345.txt  pydemo  pydemo.zip  TestDataGen.class
[hadoop@ip-172-31-60-188 ~]$ hadoop fs -put foodratings228345.txt /user/hadoop
put: `/user/hadoop/foodratings228345.txt': File exists
[hadoop@ip-172-31-60-188 ~]$ hadoop fs -ls /user/hadoop
Found 4 items
drwxr-xr-x   - hadoop hdfsadmingroup          0 2023-03-25 19:50 /user/hadoop/.sparkStaging
-rw-r--r--   1 hadoop hdfsadmingroup         65 2023-03-25 19:17 /user/hadoop/cs595doc2.txt
-rw-r--r--   1 hadoop hdfsadmingroup      17499 2023-03-25 19:22 /user/hadoop/foodratings228345.txt
-rw-r--r--   1 hadoop hdfsadmingroup        136 2023-03-25 19:17 /user/hadoop/twinkle.txt
[hadoop@ip-172-31-60-188 ~]$
```

Exercise 1)

ex1RDD = sc.textFile('/user/hadoop/foodratings228345.txt')

ex1RDD.take(5)

['Joy,23,36,41,43,3', 'Joy,29,34,29,6,3', 'Sam,23,34,18,42,1', 'Jill,13,40,6,39,5', 'Jill,33,23,32,50,4']

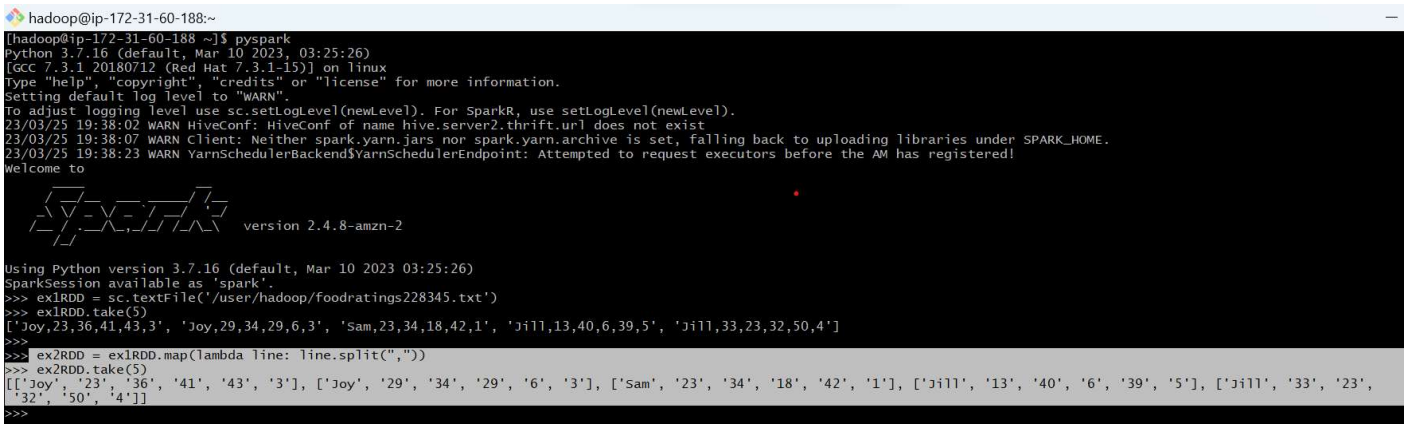
```
hadoop@ip-172-31-60-188:~
[hadoop@ip-172-31-60-188 ~]$ pyspark
Python 3.7.16 (default, Mar 10 2023, 03:25:26)
[GCC 7.3.1 20180712 (Red Hat 7.3.1-15)] on linux
Type "help", "copyright", "credits" or "license" for more information.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
23/03/25 19:38:02 WARN HiveConf: HiveConf of name hive.server2.thrift.url does not exist
23/03/25 19:38:07 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries under SPARK_HOME.
23/03/25 19:38:23 WARN YarnSchedulerBackend$YarnSchedulerEndpoint: Attempted to request executors before the AM has registered!
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 2.4.8-amzn-2
      /_/

Using Python version 3.7.16 (default, Mar 10 2023 03:25:26)
SparkSession available as 'spark'.
>>> ex1RDD = sc.textFile('/user/hadoop/foodratings228345.txt')
>>> ex1RDD.take(5)
['Joy,23,36,41,43,3', 'Joy,29,34,29,6,3', 'Sam,23,34,18,42,1', 'Jill,13,40,6,39,5', 'Jill,33,23,32,50,4']
>>>
```

Exercise 2)

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

ex2RDD.take(5)

[['Joy', '23', '36', '41', '43', '3'], ['Joy', '29', '34', '29', '6', '3'], ['Sam', '23', '34', '18', '42', '1'], ['Jill', '13', '40', '6', '39', '5'], ['Jill', '33', '23', '32', '50', '4']]

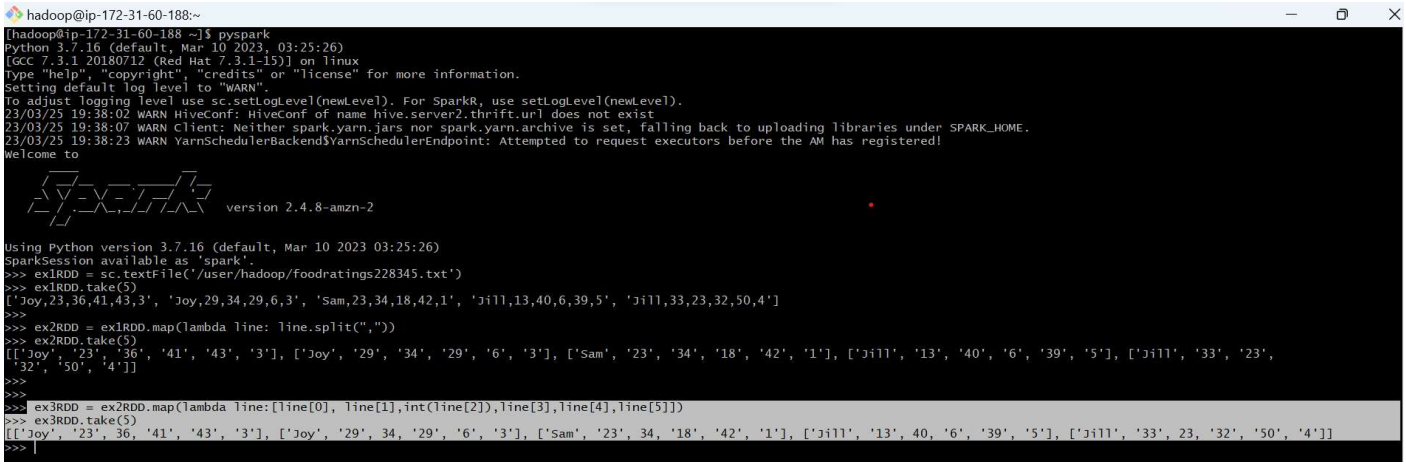


Exercise 3)

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

ex3RDD.take(5)

[['Joy', '23', 36, '41', '43', '3'], ['Joy', '29', 34, '29', '6', '3'], ['Sam', '23', 34, '18', '42', '1'], ['Jill', '13', 40, '6', '39', '5'], ['Jill', '33', 23, '32', '50', '4']]

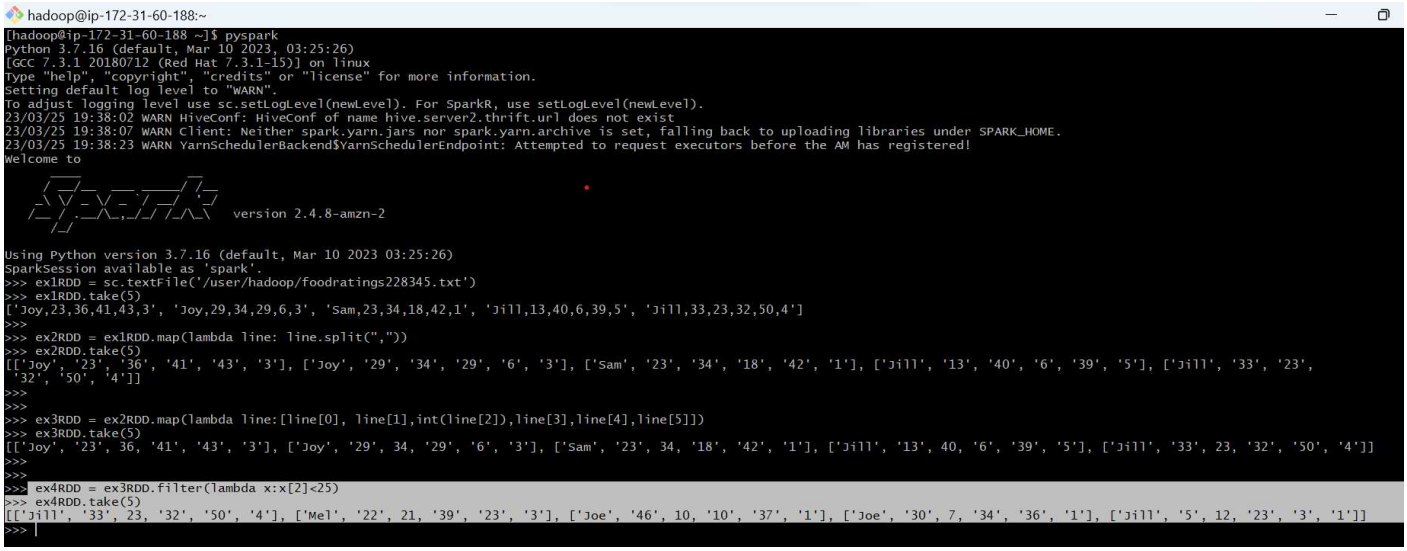


Exercise 4)

ex4RDD = ex3RDD.filter(lambda x:x[2]<25)

ex4RDD.take(5)

[['Jill', '33', 23, '32', '50', '4'], ['Mel', '22', 21, '39', '23', '3'], ['Joe', '46', 10, '10', '37', '1'], ['Joe', '30', 7, '34', '36', '1'], ['Jill', '5', 12, '23', '3', '1']]

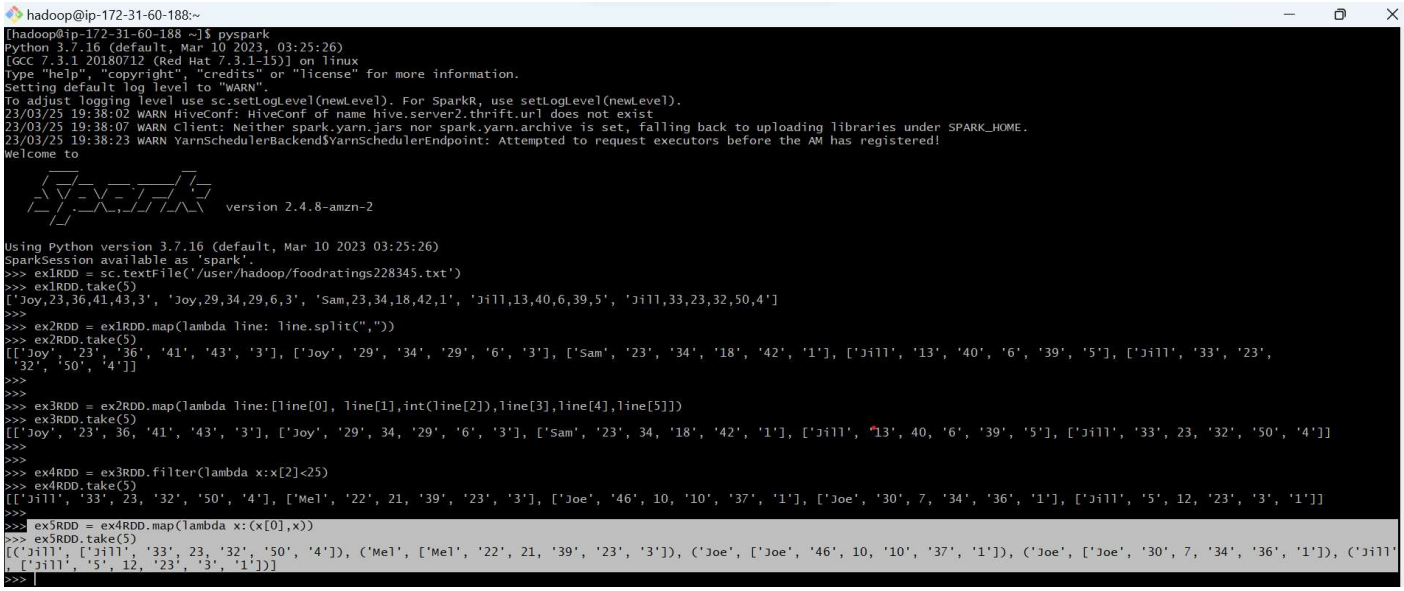


Exercise 5)

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

ex5RDD.take(5)

[('Jill', ['Jill', '33', 23, '32', '50', '4']), ('Mel', ['Mel', '22', 21, '39', '23', '3']), ('Joe', ['Joe', '46', 10, '10', '37', '1']), ('Joe', ['Joe', '30', 7, '34', '36', '1']), ('Jill', ['Jill', '5', 12, '23', '3', '1'])]



Exercise 6)

ex6RDD = ex5RDD.sortByKey()

ex6RDD.take(5)

[('Jill', ['Jill', '33', 23, '32', '50', '4']), ('Jill', ['Jill', '5', 12, '23', '3', '1']), ('Jill', ['Jill', '13', 17, '29', '42', '3']), ('Jill', ['Jill', '29', 5, '45', '20', '1']), ('Jill', ['Jill', '3', 1, '37', '6', '3'])]