TP_SPARK_CORE_3

Axel Jacquin

January 17, 2017

In this exercice, wi will work with same dataset that we used in TP_SPARK_CORE.

Tip: In this exercise you will be reducing and joining large datasets, which can take a lot of time. You may wish to perform the exercises below using a smaller dataset, consisting of only a few of the web log files, rather than all of them. Remember that you can specify a wildcard; textFile("/loudacre/weblogs/*6") would include only filenames ending with the digit 6.

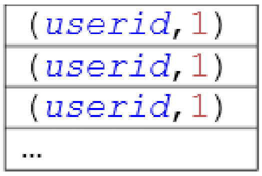
Tip 2 : To access attributes of a tuple which is an argument of a function, you could use this syntax: (p is the tuple)

```
def function (p:(Int,Int)): Int = {
        return p._1 + p._2
}
```

1 Exploring Key-Value Pair RDD functions

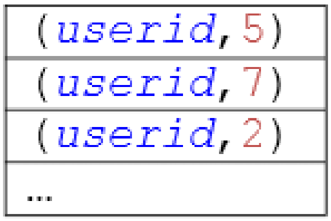
1. In this question we are going to count the number or requests from each user

1.1 Use *map* to create a Pair RDD with the user ID as the key, and the integer 1 as the value. (The user ID is the third field in each line.) Your data will look something like this:



```
(userid,1)
(userid,1)
(userid,1)
...
```

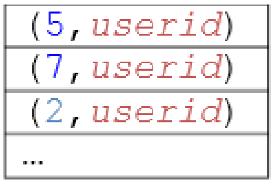
1.2 Use *reduceByKey* to sum the value for each user ID. Your RDD data will be similar to:



(userid,5)
(userid,7)
(userid,2)
...

2. In this question we are going to use countByKey to determine how many users visited the site for each frequency. That is, how many users visited once, twice, three times and so on

2.1 Use a map to reverse the key and the value, like this :



(5,userid)
(7,userid)
(2,userid)
...

2.2 Use the countByKey action to return a map of frequency:user-count

3. Create an RDD where the user id is the key, and the value is the list of all the IP addresses that user has connected from. (IP address is the first field in each request line.)
Hint: Map to (userid,ipaddress) and then use groupByKey.

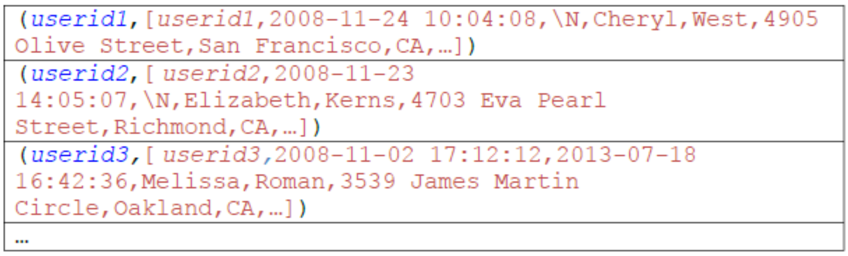
2 Join Web Log Data with Account Data

4. Copy the file /home/training/training_materials/sparkdev/data/accounts.csv into /loudacre directory on HDFS.

This data set consists of information about Loudacre's user accounts. The first field in each line is the user ID, which corresponds to the user ID in the web server logs. The other fields include account details such as creation date, first and last name and so on.

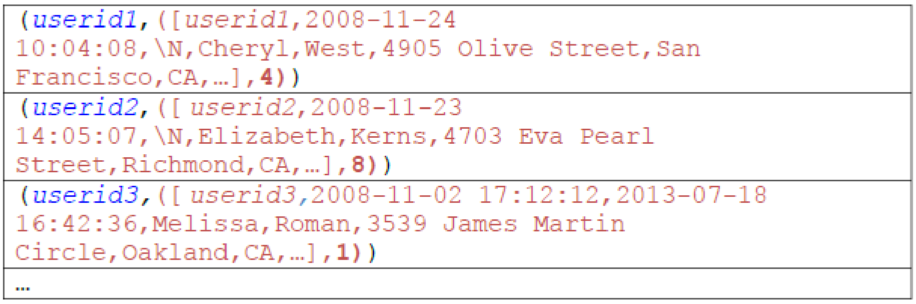
5. Join the accounts data with the weblog data to produce a dataset keyed by user ID which contains the user account information and the number of website hits for that user.

5.1 Map the accounts data to key/value RDD (userid, [values. . .]) like :



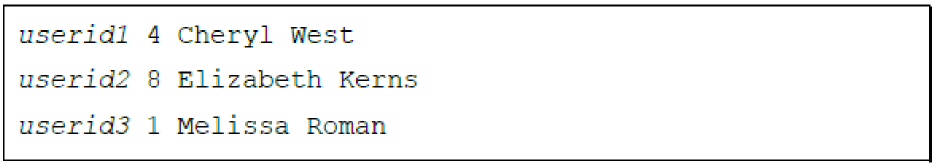
```
(userid1,[userid1,2008-11-24 10:04:08,\N,Cheryl,West,4905
Olive Street,San Francisco,CA,…])
(userid2,[ userid2,2008-11-23
14:05:07,\N,Elizabeth,Kerns,4703 Eva Pearl
Street,Richmond,CA,…])
(userid3,[ userid3,2008-11-02 17:12:12,2013-07-18
16:42:36,Melissa,Roman,3539 James Martin
Circle,Oakland,CA,…])
…
```

5.2 Join the Pair RDD with the set of userid/hit counts calculated in **2.2** like :



```
(userid1,([userid1,2008-11-24
10:04:08,\N,Cheryl,West,4905 Olive Street,San
Francisco,CA,…],4))
(userid2,([ userid2,2008-11-23
14:05:07,\N,Elizabeth,Kerns,4703 Eva Pearl
Street,Richmond,CA,…],8))
(userid3,([ userid3,2008-11-02 17:12:12,2013-07-18
16:42:36,Melissa,Roman,3539 James Martin
Circle,Oakland,CA,…],1))
…
```

5.3 What is the Type of each element of the previous RDD ?
Difficult: Display the user ID, hit count , and first name (3rd value) and last name (4th value) for the first 5 elements, e.g.:



```
userid1 4 Cheryl West
userid2 8 Elizabeth Kerns
userid3 1 Melissa Roman
```