CPSC 4330 Big Data Analytics

In-class Exercise 10 – Use Pair RDDs to Join Two Datasets

In this exercise, you will continue exploring the Loudacre web server log files (which were prepared in Exercise 9), as well as the Loudacre user account data, using key-value Pair RDDs.

Task 1: Explore Web Log Files

Data files that are used in this task are located on HDFS. The directory is:
/loudacre/weblogs

You may wish to perform the exercise below using a smaller dataset, consisting of the only a few of the web log files, rather than all of them (which can take a lot of time). Remember that you can specify a wildcard (e.g. /loudacre/weblogs/*6.log would include only log files which names end with the digit 6).

1. Using map-reduce, count the number of requests from each user.
    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 of the log file).
    2) Use reduce to sum the values for each user ID.

2. Determine how many users visited the site for each frequency. That is, how many users visited once, twice, three times and so on.
    1) Use map to reverse the key and value from the result of step 1.
    2) Use the countByKey action to return a Map (the Map here is a data structure) of *frequency:user-count* pairs.

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 line of the log file).
    1) Use map to create a Pair RDD with the user ID as the key, and the IP address as the value.
    2) Use the groupByKey to group the list of all the IP addresses that user has connected from for each user.
    3) You can use the following code to print out the first 5 user ids, and their IP list. "userips" the name of the rdd where the user id is the key, and the value is the list of all the IP addresses that user has connected from.

```
for (userid,ips) in userips.take(5):
    print(userid, ":")
    for ip in ips: print ("\t", ip)
```

Task 2: Join Web Log Data with Account Data

1. Download the file "accounts" from Canvas and copy it to Docker. Then copy the file to HDFS under the /loudacre directory.

2. Browse contents of the file "accounts" using the following command.

```
#hadoop fs -tail /loudacre/accounts
```

Each line is comma delimited. 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.

3. 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.

1) Create an RDD (named "accounts") based on the accounts data consisting of key/value-array pairs.

```
(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,…])
…
```

2) Join the Pair RDD with the set of user-id/hit-count pairs (e.g. named "userreqs") calculated in step 1 of task 1.

```
(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))
…
```

3) Display the userID, the hit count, the first name (3rd value), the last name (4th value) for the first 5 elements.

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

You can use the following code for this step (where “result” is the name of rdd)

```
>>> for (userid, (values, count)) in result.take(5):
        print(userid,  count, values[3], values[4])
```