Lab 4

Introduction

In this lab, you will use Spark to analyze a small dataset using Pair RDDs.

Objectives

- Practice Pair RDD operations.
- Practice MapReduce programming in Spark.

Preparation

In this lab, you will use the same 2014-03-15.log that you used in Lab 2. Refer to Lab 2 for its format. You should already have this file in HDFS at `loudacre/weblog/2014-03-15.log`.

Task 1: visit frequency

Step 1

Count the number of requests from each user and save the result to an RDD named `requestCountsRdd`.

You will need to use the **User ID** field, which is the third field in each line of the weblogs data.

Your RDD data will look like:

```
(userIdA, 5)
(userIdB, 7)
(userIdC, 5)
```

Lastly, run `requestCountsRdd.take(10)`. Take screenshots of all your commands and results.

Step 2

Determine how many users visited once, twice, three times, etc., and save the result to an RDD named `visitFrequencyRdd`.

In the previous example, two users visited five times, and one user visited seven times, so your RDD data will look like:

```
(5, 2)
(7, 1)
```

Lastly, run `visitFrequencyRdd.collect()`. Take screenshots of all your commands and results.

Task 2: user IP list

In this task, you need to analyze the weblogs data to figure out where the users are located. Specifically, for each user, you will generate a list of all the **unique IP addresses** that the user has connected from.

You should save your results to text file(s) in an HDFS directory named `loudacre/useriplist`. Here is the format of your output file(s):

- Your output should be sorted by User ID numerically (*i.e.*, 9 goes before 10), one user per line.
- Each line should begin with the User ID, followed by a colon and a space (`: `), followed by a space-delimited list of all the unique IP addresses that the user has connected from.
- Within each line, the IP addresses can be in any order – they don't have to be sorted.

Here is an example of the first few lines in your output file(s):

```
1: 5.173.74.120 18.180.122.7 89.35.59.219 210.32.249.72
2: 101.44.132.252 103.122.148.36 105.195.97.44 125.92.103.215 134.214.2
3: 74.118.240.165
```

Retrieve all output part files from HDFS and include them in your submission. Also, take screenshots of what you did.

Submission

To receive full credit, please compress all of the following items into a single file named `YourName_NetID_lab4.zip` and upload it to NYU Brightspace.

- For Task 1, put your screenshots in a folder named `task1`.
- For Task 2, put your screenshots and all output part files in a folder named `task2`.

Tips

Please use Discord if you experience any difficulties. The graders and I will help you get your environment working.

Don't procrastinate. The Hadoop cluster tends to get crowded near the due date.