cloudera®

Ask Bigger Questions

CAP - Developing with Spark and Hadoop:

Homework Assignment Guide for Students

Homework: Use Pair RDDs to Join Two Datasets

Files and Data Used in This Homework:

Exercise Directory: `$DEV1/exercises/spark-pairs`

Data files (HDFS):

`/loudacre/weblogs`

`/loudacre/accounts`

In this Exercise you will continue exploring the Loudacre web server log files, as well as the Loudacre user account data, using key-value Pair RDDs.

Explore Web Log Files

Continue working with the web log files, as in the previous exercise.

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.

1. Using map-reduce, count the number of requests from each user.

    a. 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)
…

b. Use `reduce` to sum the values for each user ID. Your RDD data will be similar to:

(*userid*,5)
(*userid*,7)
(*userid*,2)
…

2. 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.

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

(5,*userid*)
(7,*userid*)
(2,*userid*)
…

   b. Use the `countByKey` action to return a Map 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 request line.)

   - Hint: Map to `(userid,ipaddress)` and then use `groupByKey`.

(*userid*,20.1.34.55)
(*userid*,245.33.1.1)
(*userid*,*65.50.196.141*)
…

(*userid*,[20.1.34.55, 74.125.239.98])
(*userid*,[75.175.32.10, 245.33.1.1, 66.79.233.99])
(*userid*,[65.50.196.141])
…

Join Web Log Data with Account Data

In the Sqoop exercise you completed earlier, you imported data files containing Loudacre's customer account data from MySQL to HDFS. Review that data now (located in `/loudacre/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.

4. 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.

    a. Create an RDD based on the accounts data consisting of key/value-array pairs: `(userid,[values…])`

```
(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,…])
…
```

    b. Join the Pair RDD with the set of user-id/hit-count pairs calculated in the first step.

```
(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))
…
```

c. 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
…
```

Optional Homework

If you have more time, attempt the following challenges:

1. Challenge 1: Use `keyBy` to create an RDD of account data with the postal code (9th field in the CSV file) as the key.
    - Tip: Assign this new RDD to a variable for use in the next challenge
2. Challenge 2: Create a pair RDD with postal code as the key and a list of names (Last Name,First Name) in that postal code as the value.
    - Hint: First name and last name are the 4th and 5th fields respectively
    - Optional: Try using the `mapValues` operation
3. Challenge 3: Sort the data by postal code, then for the first five postal codes, display the code and list the names in that postal zone, e.g.

```
--- 85003
Jenkins,Thad
Rick,Edward
Lindsay,Ivy
…
--- 85004
Morris,Eric
Reiser,Hazel
Gregg,Alicia
Preston,Elizabeth
…
```

This is the end of the Homework