CS561: Advanced Topics In Database Systems
Spring-2016

Assignment 4
Team-based Coding Project

Total Points: **150**

Release Date: **03/14/2016**

Due Date: **03/31/2016 (11:59PM)**

Teams: **Project to be done in teams of two.**

Short Description

In this project, you will write java map-reduce jobs that implement advanced operations in Hadoop as well as learn more details about Hadoop's Input Formats.

Problem 1 (Spatial Join) [50 points]

Spatial join is a common type of joins in many applications that manage multi-dimensional data. A typical example of spatial join is to have two datasets: ***Dataset P*** (set of points in two dimensional space) as shown in Figure 1a, and ***Dataset R*** (set of rectangles in two dimensional space) as shown in Figure 1b. The spatial join operation is to join these two datasets and report any pair (rectangle *r*, point *p*) where *p* is contained in *r* (or even in the border of *r*).

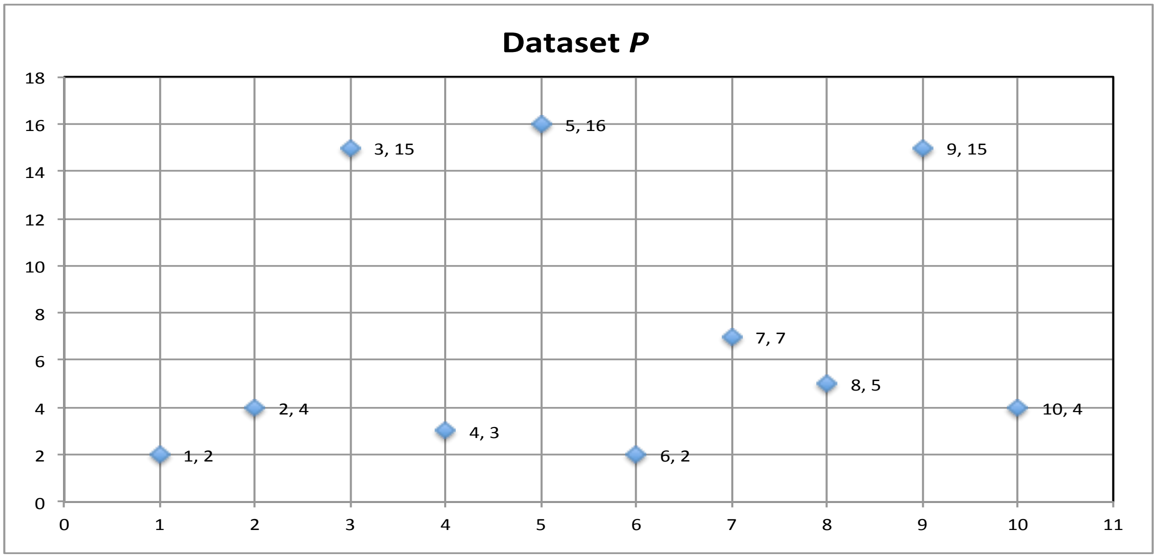

Figure 1a: Set of 2D Points

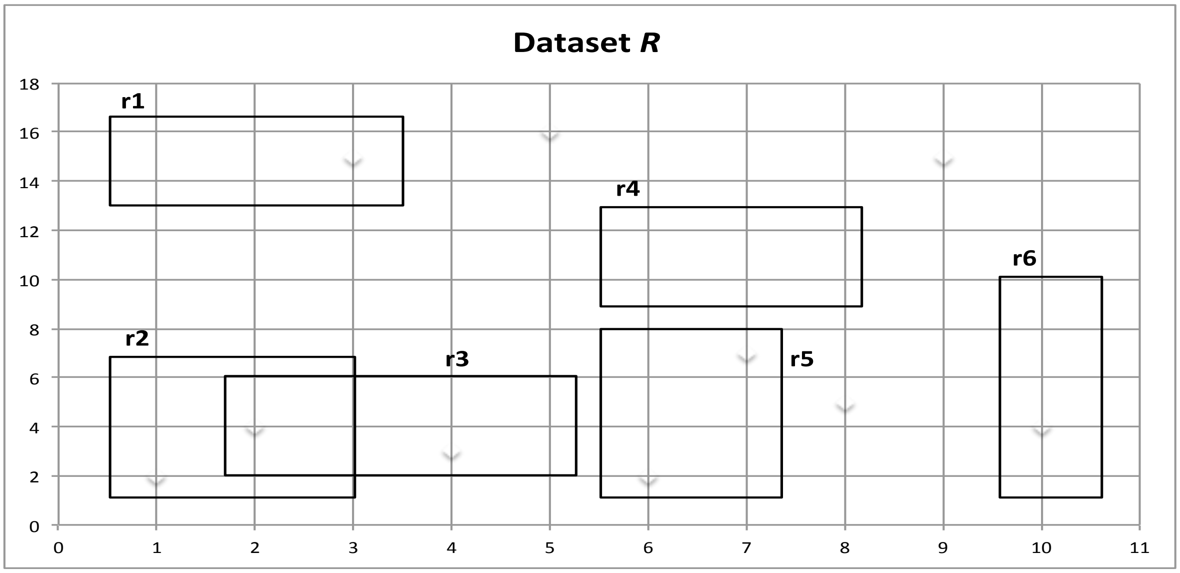

Figure 1b: Set of 2D Rectangles

For example, the join between the two datasets shown in Figure 1, will result in.

<r1, (3,15)>
<r2, (1,2)>
<r2, (2,4)>
<r3, (2,4)>
<r3, (4,3)>
<r5, (6,2)>
<r5, (7,7)>
<r6, (10,4)>

Step 1 (Create the Datasets)[10 Points]

- Your task in this step is to create the two datasets *P* (set of 2D points) and *R* (set of 2D rectangles). Assume the space extends from 1…10,000 in the both the X and Y axis. Each line will contain one object.
- Scale each dataset *P* or *R* to be at least 100MB.
- Choose the appropriate random function (of your choice) to create the points. For the rectangles, you will need to also select a point at random (say the top-left corner), and then select two random variables that define the *height* and *width* of the rectangle. For example, the height random variable can be uniform between [1,20] and the width is also uniform between [1,5].

Step 2 (MapReduce job for Spatial Join)[40 Points]

In this step, you will write a java map-reduce job that implements the spatial join operation between the two datasets *P* and *R* based on the following requirements:

- The program takes an optional input parameter W(x1, y1, x2, y2) that indicate a spatial window (rectangle) of interest within which we want to report the joined objects. If W is omitted, then the entire two sets should be joined.
    - Example, referring to Figure 1, if the window parameter is W(1, 3, 3, 20), then the reported joined objects should be:

      <r1, (3,15)>
      <r2, (2,4)>
      <r3, (2,4)>

- You should have a single map-reduce job to implement the spatial join operation.

Problem 2 (Custom Input Format) [50 points]

So far, all of the given assignments use text files as input, and hence you use 'TextInputFormat()' to read the files. In this problem, you will learn more about Hadoop input formats and you will write your custom one to read the input data.

Step 1 (Create the Datasets)[10 Points]

Assume we have a customer dataset, where each record is stored as follows:

```
{ Customer ID: <id>,
  Name: <name>,
  Address: <addr>,
  Salary: <salary>,
  Gender: <gender>
},
{ Customer ID: <id>,
  Name: <name>,
  Address: <addr>,
  Salary: <salary>,
  Gender: <gender>
},
….
```

- In this step, you are required to create a customer dataset with the above format.
- This format is called JSON format. Some of its characteristics are the following:
  - Every record starts with "{" and it ends with "}"
  - Each record has many fields separated by ",".
  - Each field in a record has a name followed by ":" followed by a value.
  - The value can be either one of the simple data types, e.g., string, date, number, or an array of values, e.g., array of numbers.
- Scale the dataset to be at least 100MBs
  - Customer Ids should be incremental unique numbers
  - Salary should be a random integer between 100 and 1000
  - Gender should be a random value of either "male" or "female"
  - Name and Address should be random strings each of length 100 character.
- Make sure within a single record, there is no "{" or "}". These brackets should identify the start and end of each record.

Step 2 (Map Job with a Custom Input Format)[40 Points]

- Now, to do any job on the above dataset using the standard "TextInputFormat()", the map function must be complex as it needs to collect many lines to form a single record. This complexity will repeat with each written job over the above dataset.
- A better way is to write a custom input format, call it "***CustomerInputFormat***" that reads many lines from the input file until it gets a complete record, and then coverts them to a list of comma separated values in a single line, and then pass it to the map function.
  - E.g., each input to the map function should be: *<id>, <name>, <addr>, <salary>, <gender>*

- Your task is to write this new "CustomerInputFormat", and use it in a map-reduce job that aggregates the records based on the Salary field, and for each salary value it should report the number of males and females having this salary.

- ***Hint:*** You need to understand first the "***FileInputFormat***", "***TextInputFormat***", and "***LineRecordReader***" classes.

Problem 3 (K-Means Clustering) [50 points]

K-Means clustering is a popular algorithm for clustering similar objects into *K* groups (clusters). It starts with an initial seed of K points (randomly chosen) as centers, and then the algorithm iteratively tries to enhance these centers. The algorithm terminates either when two consecutive iterations generate the same K centers, i.e., the centers did not change, or a maximum number of iterations is reached.

Hint: You may reference these links to get some ideas (in addition to the course slides):
http://en.wikipedia.org/wiki/K-means_clustering#Standard_algorithm
https://cwiki.apache.org/confluence/display/MAHOUT/K-Means+Clustering

Step 1 (Creation of Dataset) [10 points]:

- Create a dataset that consists of 2-dimenional points, i.e., each point has (x, y) values. X and Y values each range from 0 to 10,000. Each point is in a separate line.
- Scale the dataset such that its size is around 100MB.
- Create another file that will contain K initial seed points. ***Make the "K" value as a parameter to your program***, such that your program will generate these K seeds randomly, and then you upload the generated file to HDFS.

Step 2 (Clustering the Data) [40 points]:

Write map-reduce job(s) that implement the K-Means clustering algorithm as given in the course slides. The algorithm should terminates if either of these two conditions become true:

a) The K centers did not change over two consecutive iterations
b) The maximum number of iterations (make it six (6) iterations) has reached.

- Apply the tricks given in class and in the 2nd link above such as:
  - Use of a combiner
  - Use a single reducer
  - The reducer should indicate in its output file whether centers have changed or not.

Hint: Since the algorithm is iterative, then you need your program that generates the map-reduce jobs to control whether it should start another iteration or not.

What to Submit

You will submit a single zip file containing the java code needed to answer the queries above. Also include a .doc or .pdf report file containing any required documentation.

How to Submit

Use blackboard system to submit your files.