CSCI 585 - Database Systems
Spring 2014
Homework Assignment 3
Due: 04/25/2014 08:59 PM

Objective:

With previous assignments, you have designed and implemented a database for a hypothetical online retailer and developed an interface applications to execute queries on top of this database.

Now, assume you are expected to support a real online retailer with large number of users. At scale, running queries on a single machine is unacceptably slow and impractical.

As an alternative to address the aforementioned efficiency problem, with this assignment you will develop sample queries in *MapReduce* to be executed on top of a cloud computing system (namely, Amazon EC2), which allows parallel computation for efficient execution of queries at large scale (Part I). Once you develop these queries, you will also evaluate and report on their performance as the number of machines allocated to execute the queries grow (Part II).

Your AWS Account:

You need to sign up for a free AWS account if you do not have already one. You can Sign Up here. You can find a tutorial on this here.

After creating an account, you need to send an email from your USC Email to *ylu720@usc.edu* in the following format:

Subject: *AWS Account*

Body: Body of the email should contain the email address that has been used to create your AWS account.

Once we receive your email, we will provide you with information that allows you to obtain $40 worth of free resources from Amazon EC2. In particular, you will receive an email from *webservices@amazon.com* with the following subject: "Request to add account to Consolidated Bill". Once you received this email, you just need to click on the link included in the email body and sign in to your AWS account and accept the consolidated billing request to redeem the credit allocated to you for this assignment.

Important Notes About Your AWS Credit:

- Make sure you do NOT exceed your credit quota ($40)! You will not receive any more credits beyond your quota limit. Make sure to stop or terminate any machines and services once you are not actively working on your assignment; otherwise, your credit will diminish quickly before you get to finish your assignment! If you have any questions, in this regard feel free to consult with the TAs before starting to use your credit.
- When you want to create a MapReduce cluster in the "Elastic Map Reduce Job Flow" make sure to set the "Auto Terminate" option to "Yes".
- You should develop and test your applications locally and only run it on the Amazon clusters once verified locally
- You are only allowed to use EMR (Elastic Map Reduce), EC2 (Elastic Compute Cloud) and S3 (Simple Storage) services. DO NOT USE ANY OTHER AWS SERVICES!
- You can implement your Hadoop MapReduce applications using your desired language and IDE. However you need to follow the MapReduce paradigm.

.

Description

Input Data Files:

You need to use the provided data files as the input and you are not allowed to change them at all. The data is only used as input.
The Customers.txt contains the information of 40 million users in the following format:

```
CID, city, x coordinate of the customer's location, y coordination of the customer's location
```

Note that (x, y) indicates the Cartesian coordinates (not latitude and longitude) of customers. Given the two points (x_1, y_1) and (x_2, y_2), the distance (d) between these points is given by the formula:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

The LikeReviews.txt contains the information that which customers like which reviews in the following format:

```
CID, RID
```

Data files are uploaded on Blackboard along with this assignment description.

Part 1: Query Development (66 points total)

In this part of assignment, you are allowed to develop and run your application locally or on the Amazon platform. If you decide to run your program on the Amazon EMR, run it on a single small machine.

1- (**33 points**) Write a Hadoop MapReduce (MR) application to find how many reviews each customer likes (don't need to output the customers who do not like any reviews). For this query you need to submit the following:

- **1io.txt** which shows the input and output of your Map and Reduce functions. for example:
  Map: <key 1, value 1> ⟶ List <key 2, value 2>
  Reduce: List <key 2, value 2> ⟶ <key 3, value 3>
  You need to briefly describe what your Map() and Reduce() functions do.
- **1code.txt** which contains your Hadoop code. You need to include your Map, Reduce and Job (main) functions (or classes according to your code) in a text file. If you would like to use additional classes or functions, include those in this text file as well.
- **1out.txt** which contains the result of Question 1 in the following format:
  userid number_of_liked reviews
  userid number_of_liked reviews
  …

2- (**33 points**) Assume for each city, there is a fixed reference location (x*, y*). You can find the reference location for each city in the "city_location.txt" uploaded on Blackboard along with this assignment. Write a Hadoop MR application to find the number of all people who are from the same city and are within a 5 mile range from each city reference point. For this query, you need to submit the following:

- **2io.txt** (for description, see similar description for **1io.txt** above)
- **2code.txt** (for description, see similar description for **1code.txt** above)
- **2out.txt** which contains the result of question 2 in the following format:
  City number_of_customers
  City number_of_customers
  …

Part 2: Evaluation of Scale-out Effect (34 points total)

In this part of the assignment, you need to run the query No. 1 (from Part 1 above) using 1, 2 and 4 EC2 small machines, respectively. Accordingly, you must prepare and submit the following:

- **3.jpg** which contains a histogram in which the X axis shows the number of machines used to run the query, while the Y axis shows the time it takes to process the query (in seconds) in each case.

Submission Guidelines:

You need to submit your assignment through the Blackboard before the deadline. No deadline extensions will be awarded! Please compress all of your files into one .zip file named “Your_USC_Email_ID.zip”. For example if your USC Email ID is “john”, name the file as “john.zip”. The zip file should contain following items:

- **1io.txt**
- **1code.txt**
- **1out.txt**
- **2io.txt**
- **2code.txt**
- **2out.txt**
- **3.jpg**

Useful Links:

- Hadoop
- AWS SDK For eclipse shows how you can add AWS plugin to your eclipse.
- AWS MapReduce Training. This link contains some video tutorials for AWS MapReduce.
- You can find some good video tutorials regarding MapReduce basics and programming here.
- There is another programming MR programming example here.
- Learn how to install Hadoop on a Windows machine here and here.
- Learb how to install Hadoop on Mac OS here.

.

As always, please feel free to post your questions on the Blackboard Discussions.