CS 524 Lab Assignment #4: Performing Data Analytics in the Cloud

Due: April 25, 2017

This is lab assignment is for a total of **100** points, which involves setting up a **free** Amazon *RedShift* cluster and learning how data analytics can be performed on the Cloud. Although this is seemingly simple and straight-forward, there is **much to read and learn** here, so make sure you start working at once.

You need to read the following documents:

https://aws.amazon.com/redshift/ (Links to an external site.)
https://aws.amazon.com/redshift/getting-started/ (Links to an external site.)

Again, make sure that you understand what you need to do to keep this experiment free of charge. When in doubt, ask a CA!

Please proceed as follows:

Step 1:

Visit *https://aws.amazon.com/redshift/* and sign In to your account.

After that, set up *your Amazon RedShift Cluster*. Please configure the cluster in the US West (Oregon) region as we will be loading data for that region in the steps that follow.

Step 2:

Amazon Redshift allows you to query your database using a third party client software. The client software to be used for this lab is *JackDB*. Visit https://www.jackdb.com and start your free trial.

In *JackDB*, add *Amazon RedShift* as the data source that you will be making a connection to.

Step 3:

Configuration in AWS:

To grant this client access to your cluster, you need to modify the Security Group to allow inbound TCP traffic from the IP address used by the client.

Configuration in JackDb client:

In addition, you will need the Endpoint to connect the client to your database. This Endpoint can be found in the Configuration tab of the Amazon Redshift cluster.

Step 4:

Once you have successfully established a connection with JackDb, you need to load sample data into the Amazon Redshift Cluster.

On the JackDb client, please execute the following commands:

a) Command for Initial Table creation in Redshift:

```
create table users( userid integer not null distkey sortkey, username char(8), firstname varchar(30), lastname varchar(30), city varchar(30), state char(2), email varchar(100), phone char(14), likesports boolean, liketheatre boolean, likeconcerts boolean, likejazz boolean, likeclassical boolean, likeopera boolean, likerock boolean, likevegas boolean, likebroadway boolean, likemusicals boolean);
```

b) Command to load sample data from S3 into your database table:

```
copy users from 's3://awssampledbuswest2/tickit/allusers_pipe.txt'

CREDENTIALS 'aws_access_key_id=<Enter your access key>;aws_secret_access_key=<Enter your secret access key>'

delimiter '|';
```

NOTE: Make sure you load the sample S3 data as given in the command above and do not create your own S3 `bucket` or upload any data as that will result in your incurring a charge!

To obtain your access key id and secret access key, click on the 'My security credentials' option in your AWS account.

c) Commands to check if the data have been loaded into the database successfully:

```
SELECT * from users;
```

```
SELECT userid,firstname,lastname,city,likesports from users where
likesports = 'true' order by firstname;
```

Step 5:

After loading data into your database, now establish a connection from *Tableau* software (Desktop version) (www.tableau.com) to your Amazon Redshift cluster.

Make sure you add a security rule to grant this software access to your cluster.

Step 6:

Once you have connected Tableau to your cluster, analyze the data and create charts that explain the data pattern. You have the freedom to create charts as you find fit, but make sure that they convey meaningful information about the data.

Step 7: **Make sure you delete your cluster after completing this assignment.**

Submission:

Please submit one word document containing:

a) Explanation and screenshots of each step that has been executed,
b) Screenshot of the cluster in the available state and should also show your login in the AWS.
c) Screenshot of the cluster configuration. (Present under the configuration tab)
d) Screenshots of the connection settings made from an external client (*JackDb, Tableau*) to the Amazon Redshift cluster. (Data Source details in *JackDb* and in *Tableau* and connection successful screenshot)
e) Screenshots of the inbound rules added in Security Groups settings.
f) Screenshots of the commands being executed on the *JackDb* client along with their output. It is alright if the entire output is not visible in the screenshot.
g) Screenshots of the charts created in *Tableau*.