INFSCI 2750: Cloud Computing

Mini Project 3

Objective

The objective of this mini project is to get familiarized with Apache Cassandra. This project can be done in groups of 2 members and you will be working on the previously assigned VM's.

Part 1: Setting up Cassandra: (50 points)

The first task is to configure Cassandra distribution on the previous cluster. The entire Cassandra setup should be configured on top of a one or two node cluster.

Because Cassandra has a "master less" architecture, all of the Cassandra nodes can be configured as same as others. In addition, the best way to know an open source project is to read the official documents. Here are the official documents for Cassandra:

http://cassandra.apache.org/doc/latest/

You can check here to install Cassandra:

http://cassandra.apache.org/download/

On Ubuntu, you can install Cassandra easily with the debian packages.

Then, you need to configure Cassandra nodes to make them work together:

http://cassandra.apache.org/doc/latest/getting_started/configuring.html

Finally, you can start your Cassandra nodes on all the VMs (make sure the previous Hadoop and Spark services are all shutdown to empty the memory and use –R parameter if you run Cassandra with root user)

Part 2: Import Data into Cassandra (25 points)

As part of the project you will be working with the log data set which has been provided in access_log.zip in the Mini Project 1.

You need to use CQL (Cassandra Query Language: http://cassandra.apache.org/doc/latest/cql/index.html) or JAVA driver of Cassandra (https://github.com/datastax/java-driver) to import the access logs into Cassandra.

You need to create one keyspace and one table at least in Cassandra to store all the logs.

You can check https://docs.datastax.com/en/cql/3.3/cql/cql_reference/cqlshCopy.html for some helps of the COPY commands or use the bulk loader to import the data: https://docs.datastax.com/en/archived/cassandra/2.0/cassandra/tools/toolsBulkloader_t.html . In addition, for creating the table, you can check here: http://cassandra.apache.org/doc/latest/cql/ddl.html#create-table .

Part 2: Operate Data in Cassandra (25 points)

As part of the project you will be working with the log data set which has been stored in Cassandra.

You need to use CQL (Cassandra Query Language: http://cassandra.apache.org/doc/latest/cql/index.html) or JAVA driver of Cassandra (https://github.com/datastax/java-driver) to operate the access logs in Cassandra.

You need to get the result for the questions below:

Problems:

1. How many hits were made to the website item "/assets/img/release-schedule-logo.png"?
2. How many hits were made from the IP: 10.207.188.188
3. Which path in the website has been hit most? How many hits were made to the path?
4. Which IP accesses the website most? How many accesses were made by it?

The first two questions can be answered by SELECT in CQL. You can check http://cassandra.apache.org/doc/latest/cql/dml.html#select if you have any difficulties.

The last two questions need one more step to get: you can either use the java-driver to insert the counts of the items into a new table and use another CQL to get the answer or just use one user-defined function to get the answer of the group-max query, you can refer: http://christopher-batey.blogspot.com/2015/05/cassandra-aggregates-min-max-avg-group.html .

Project Submission: Submit a **single ZIP file** with your *Pitt email ID* as its filename via the CourseWeb system. The package should contain all your source files and a *readme* file that explains how to execute your program. Also include screenshots of your programs output and cql shell. The IP address of the master machine should be clearly visible in the screenshots. In addition for Part2, the *readme* file needs include the screenshots of showing parts of the importing dataset. For Part3, you need to show the results in the screenshots.