INFSCI 2750: Cloud Computing

Mini Project 2

Spring 2024

Objective

The objective of this mini project is to learn and develop applications using Apache Cassandra. You will using the previously assigned VMs.

Part 1: Setting up Cassandra: (50 points)

The first task is to configure a Cassandra distribution in your cluster of VMs. The entire Cassandra setup should be configured on top of a two-node or three-node cluster.

As Cassandra has a "master-less" architecture, all of the Cassandra nodes can be configured in the same way. Here are some 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 from 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 shut down 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 is similar to Mini Project 1. You may find the link to download the file in Canvas.

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 help of the COPY commands or use the bulk loader to import the data: https://github.com/datastax/dsbulk. In addition, for creating the table, you can check here: http://cassandra.apache.org/doc/latest/cql/ddl.html#create-table .

Part 3: 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 “/administrator/index.php”?
2. How many hits were made from the IP: 96.32.128.5
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?
5. How many accesses were made by Firefox(Mozilla)?
6. For all requests on 02/Apr/2022, what is the ratio of GET request?
7. How many requests are lower than or equal to 404 bytes?
8. List the IPs that have more than **ten** 404 requests. If no ip fulfills, print the ip that has most 404 requests and the number of requests.

You can check http://cassandra.apache.org/doc/latest/cql/dml.html#select if you have any difficulties.

For some questions, you may need more than one 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 and https://docs.datastax.com/en/cql-oss/3.3/cql/cql_using/useCreateUDA.html

Project Submission: Submit a **single ZIP file** with you and your teammates’ *Pitt email ID* as its filename. **For example: if studentA(abc@pitt.edu) and student(def@pitt.edu) are in a group, the filename should be abc_def.zip**.

The package should contain all your source files and a *readme* file that explains how to execute your program. Also include screenshots of your program’s output and CQL shell. The IP address of the master machine should be clearly visible in the screenshots. In addition, for Part 2, the *readme* file needs include the screenshots of showing parts of the importing dataset. For Part 3, you need to show the results in the screenshots.