Rushi Samirbhai Patel rs397441@dal.ca
B00886157

Assignment 4

Problem-1

Task 1

Cluster Setup - Apache Spark Framework on GCP

Explanation:

Initially I recalled the steps discussed in the lab where I learnt to install and configure Apache Spark on GCP. My GCP credits were out so after contacting TA, I started my free trial which gave me 350 dollars credit in my account.

- Firstly, I created a **new compute engine instance on a GCP virtual machine**. I configured my virtual machine with a machine type E2-medium and selected Ubuntu as BootDisk. I also allowed all traffic under the firewall (both HTTP and HTTPS).
- Secondly, I c**onnected to this configured virtual instance** with the help of **browser cloud shell.**
- Under the **VPC network**, I **allowed all protocols and ports** for both HTTP and HTTPS in the default firewall.
- Now, I **installed all the prerequisites** required to install Apache Spark - Scala, Java, and Git and checked the versions of these prerequisites in the VM.
- I then **installed Apache Spark in a new directory** and **configured the path** in the VM. I even configured the path permanently in the profile file by mentioning the Spark path as a variable in the path.
- After performing the above steps to check if Spark is installed in the VM, I tried **starting the master node** and **checked the logs** using the tail command. I then checked the spark dashboard in the browser. I did the same thing for **slave node.**

Task 2

Data Extraction and Preprocessing Engine

Tweet Extraction Algorithm:

Step 1: Create a Twitter developer account

Step 2: Generate API key, API secret key, Access Token key, Access Token secret key and Bearer token and save it for future use.

Step 3: Apply for elevated access
Step 4: Create a search query which includes all the search keywords (lang:en (mask OR cold OR flu OR snow OR immune OR vaccine)).
Step 5: Use Twitter4j to interact with Twitter and its properties to fetch tweet data.
Step 6: For authentication purpose, use ConfigurationBuilder of twitter4j and set the keys for OAuth that were saved in step 2. Build theconfigurationBuilder and get the instance of TwitterFactory and use this instance to fetch tweet data.
Step 7: The instance created in Step 6 should be used and the count of the search query (developed in step 4) must be set to 100. Since we require more than 3000 tweet data, iterate 32 times and search the query to get 3200 tweet data.
Step 8: The data is not in the proper JSON format since the data of all the 3200 tweets are not separated with comma so separate each tweet data with a comma and introduce a parent key to the entire file document named "data". This will make it easy to fetch this data in the tweet transformation algorithm.
Step 9: For every tweet data, create a text file to store the raw data fetched on searching the query.

Tweet Transformation Algorithm:

Step 1: Fetch all the text files created in the Tweet Extraction program.
Step 2: Convert the text file data to JSON and use JSONParser to parse the file data in the program.
Step 3: Store the data in the JSONObject and this data will be in the form of JSONArray since we had assigned a parent key of the entire file data as "data" in the Tweet Extraction program.
Step 4: Iterate over the JSONArray of step 3 and fetch the 'id', 'text', 'retweet count' and 'lang' parameters of the tweet data from the raw data stored in the text file.
Step 5: Transform the data by eliminating all the word characters and white spaces for the id data of the tweet.
Step 6: Transform the data by eliminating the url and whitespaces from the text data of the tweet.
Step 7: Establish a connection with MongoDB with the connection string.
Step 8: Create a document and insert the transformed data of the tweets into this document which will store it in the MongoDB.

Rushi Samirbhai Patel rs397441@dal.ca
B00886157

References:

[1] H. -->twitter4j.conf.ConfigurationBuilder, "twitter4j.conf.ConfigurationBuilder.setOAuthConsumerKey java code examples | Tabnine", *Tabnine.com*, 2022. [Online]. Available: https://www.tabnine.com/code/java/methods/twitter4j.conf.ConfigurationBuilder/setOAuthConsumerKey. [Accessed: 13- Mar- 2022].

[2] "Java Create and Write To Files", *W3schools.com*, 2022. [Online]. Available: https://www.w3schools.com/java/java_files_create.asp. [Accessed: 13- Mar- 2022].

[3] R. [duplicate], D. Agrawal, S. V and R. M.Tuman, "Regex for website or url validation", *Stack Overflow*, 2022. [Online]. Available: https://stackoverflow.com/questions/42618872/regex-for-website-or-url-validation. [Accessed: 13- Mar- 2022].

[4] H. -->com.mongodb.MongoClientSettings, "com.mongodb.MongoClientSettings.builder java code examples | Tabnine", *Tabnine.com*, 2022. [Online]. Available: https://www.tabnine.com/code/java/methods/com.mongodb.MongoClientSettings/builder. [Accessed: 13- Mar- 2022].