CSCI 5408 – Data Management, Warehousing and Analytics

Assignment-3

1. Cloud Setup Process

I have followed the following steps for setting up the cloud environment.

Java Installation

1. sudo apt install openjdk-8-jdk [1]
2. sudo apt-get update

Oracle Installation

1. sudo apt-get -y install openjdk-8-jdk-headless

Python Installation

1. sudo apt-get install python3

Apache Spark Installation

1. wget http://apache.forsale.plus/spark/spark-2.4.5/spark-2.4.5-bin-hadoop2.7.tgz
2. sudo tar -zxvf spark-2.4.5-bin-hadoop2.7.tgz

Setting up the environment variables

1. eport JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64/
2. export SPARK_HOME=~/server/spark-2.4.5-bin-hadoop2.7
3. export PYSPARK_PYTHON=python3

MongoDB Installation

1. sudo apt-key adv --keyserver hkp://keyserver.ubuntu.com:80 --recv 4B7C549A058F8B6B
2. echo "deb [arch=amd64] https://repo.mongodb.org/apt/ubuntu bionic/mongodb-org/4.2 multiverse" | sudo tee /etc/apt/sources.list.d/mongodb.list [2]
3. sudo apt install mongodb-org

2. API Setup Process

Twitter API

1. Created a twitter developer account [3]
2. Created an application in the name of CSCI 5408 Satya
3. Received Consumer API keys and access token and access token secret
4. Using these API keys accessed search API and streaming API.

News API

1. Launched URL: https://newsapi.org/ [4]
2. Created a developer account and received the API Key
3. Using https://newsapi.org/v2/everything, retrieved the news articles.

Movies API

1. Launched URL: http://www.omdbapi.com/ [5]
2. Created a developer account and received the API Key
3. Using http://www.omdbapi.com/?apikey= , retrieved the movies data and ratings.

3. Data Extraction Process

Twitter Data

By using the API keys and tokens, I have used Search API (1750 tweets) [6] by limiting using tweet cursor and Stream API (1750 tweets) by using stream listener to extract the data. I have stored them in a JSON file after performing cleaning of data (Please refer tweets.py file). Author, Date, Location, Text attributes are retrieved from the twitter data.

News Data

I have passed the required words to search and number of articles to retrieve as parameters and retrieved articles from the News API. Author, Content, Date, Title attributes are retrieved from the news article.

Movies Data

I have passed the required words to search to retrieve as parameters and retrieved movies data from the Movies API. Genre, Plot, Ratings, Title attributes are retrieved from the movies data.

4. Data Cleaning Process

I have used emoji patterns and regular expressions to replace all the emoticon, special characters, URL's with an empty string. I have also used the python regular expression package called 're'.

5. Word Count Process

Apache Spark Application is created, and I have written a map reduce function [7] using Python language for finding and counting the specified words by reading the text files and using the flatmap found the word count. I have also scripts for finding out the single words and double words a and stored in the respective text files such as tweets.txt and news.txt.

This created python script is deployed on the AWS EC2 instance and ran the following command:

1. pyspark - sudo ./spark-2.4.5-binhadoop2.7/bin/spark-submit --deploy-mode client wordcount.py

The specified wordcount.txt is the text file is the output generated by the spark slave.

6. Sample output format files

1. tweets.json, tweets.csv, tweets.txt - cleaned files for tweets data
2. news.json, news.txt, news.csv - cleaned files for news data
3. movies.json, movies.txt, movies.csv – cleaned files for movies data
4. wordcount.txt – word count of each word using Apache Spark
5. movieratings_mongodb.json – movie ratings json file that is generated from the python that retrieves data from mongodb.

7. References

[1]"Installing PySpark with JAVA 8 on ubuntu 18.04", *Medium*, 2020. [Online]. Available: https://towardsdatascience.com/installing-pyspark-with-java-8-on-ubuntu-18-04-6a9dea915b5b. [Accessed: 24- Mar- 2020].

[2]"How to Install MongoDB 4.2 on Ubuntu 18.04 & 16.04 - TecAdmin", *TecAdmin*, 2020. [Online]. Available: https://tecadmin.net/install-mongodb-on-ubuntu/. [Accessed: 24- Mar- 2020].

[3]*Developer.twitter.com*, 2020. [Online]. Available: https://developer.twitter.com/en/apps/17491175. [Accessed: 24- Mar- 2020].

[4]"Login - News API", *Newsapi.org*, 2020. [Online]. Available: https://newsapi.org/account. [Accessed: 24- Mar- 2020].

[5]"OMDb API - The Open Movie Database", *Omdbapi.com*, 2020. [Online]. Available: http://www.omdbapi.com/. [Accessed: 24- Mar- 2020].

[6] Tweepy's Documentation on 'API', 'Authentication', 'Cursor', 'Extended Tweets', Available: https://tweepy.readthedocs.io/en/latest/index.html [Accessed on 24- Mar- 2020]

[7] TutorialKart's , 'MongoDB Map Reduce', Available: https://www.tutorialkart.com/mongodb/mongodb-map-reduce/ [Accessed on 24- Mar- 2020]