Assignment Report

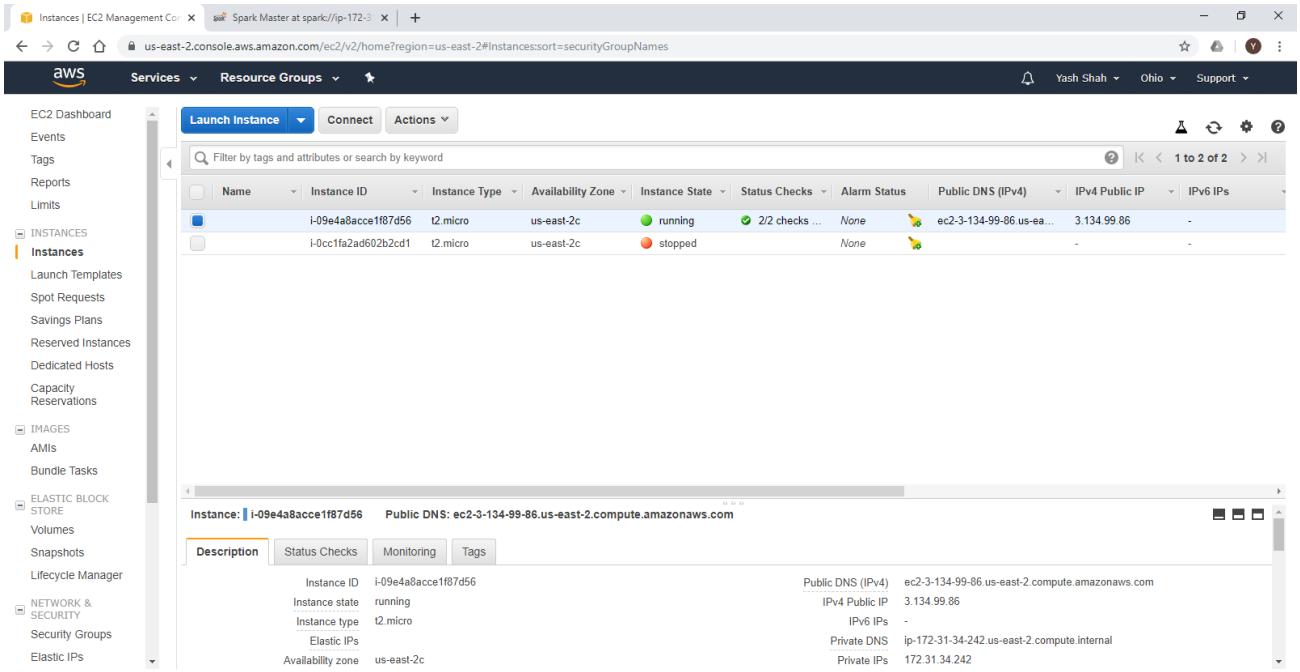
- Cloud Setup:

→ Starting from the basic; the first step in setting up the cloud storage was to create an account on Amazon Website Service (AWS).

→ After selecting the most suitable instance type 'Ubuntu Server 18.04LTS(HVM), SSD Volume Type', I configured it to make the most of this instance.

→ For better security, AWS uses asymmetric encryption concept and provides us with a key for secure login. Software named 'PuTTY' is used to create the public and private key of the AWS instance.

→ To allow interaction with the instance, I had enabled few ports (8080, 8081, 80, 22) in the inbound rules section under 'Security Groups' option.

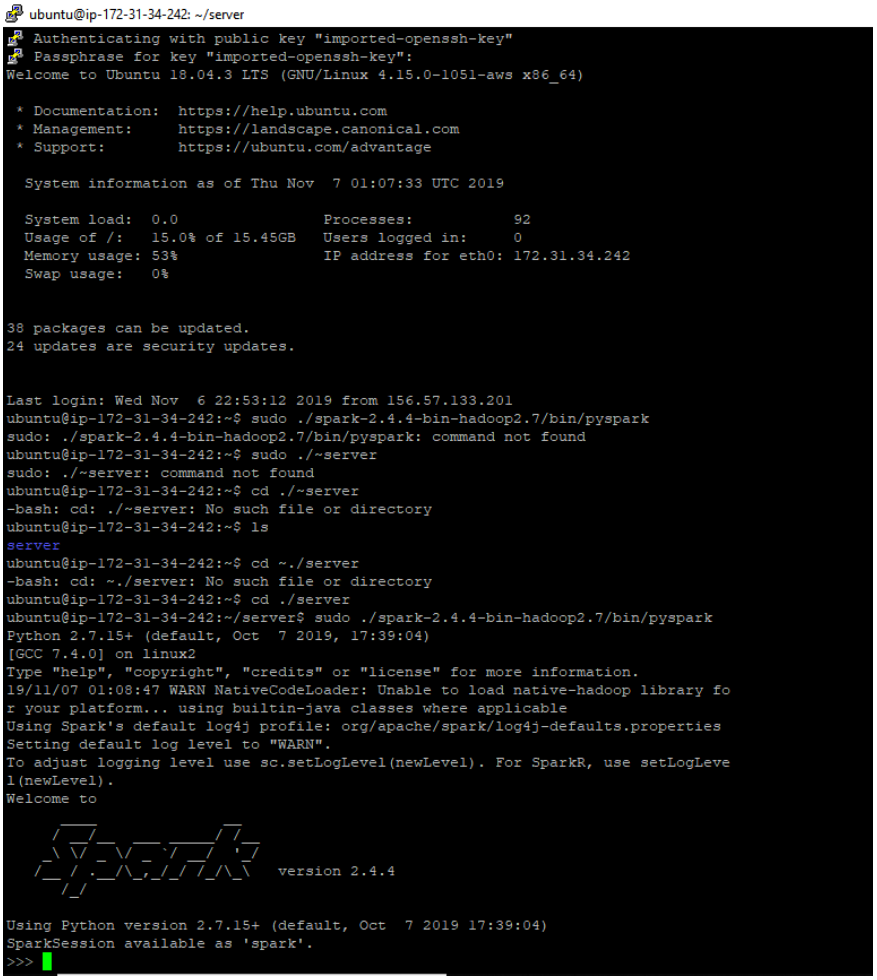


1. My Cloud Setup with a running instance

→ Post this, in order to run 'Apache Spark' on my instance, I installed Java and Python as the first step. Then, I created a directory 'server' and installed 'Apache Spark 2.4.4' using the following commands – 'wget http://apache.forsale.plus/spark/spark-2.4.4/spark-2.4.4-bin-hadoop2.7.tgz' and 'sudo tar zxvf spark-2.4.3-bin-hadoop2.7.tgz'.

→ After this, I made some changes in the file name 'profile' in order to use Java and Python with Spark.

→ Later, I started a Spark master node and slave node and then I started the Spark shell to run my MapReduce query.



2.Initialised Apache Spark 2.4.4

- Data Extraction and Transformation:

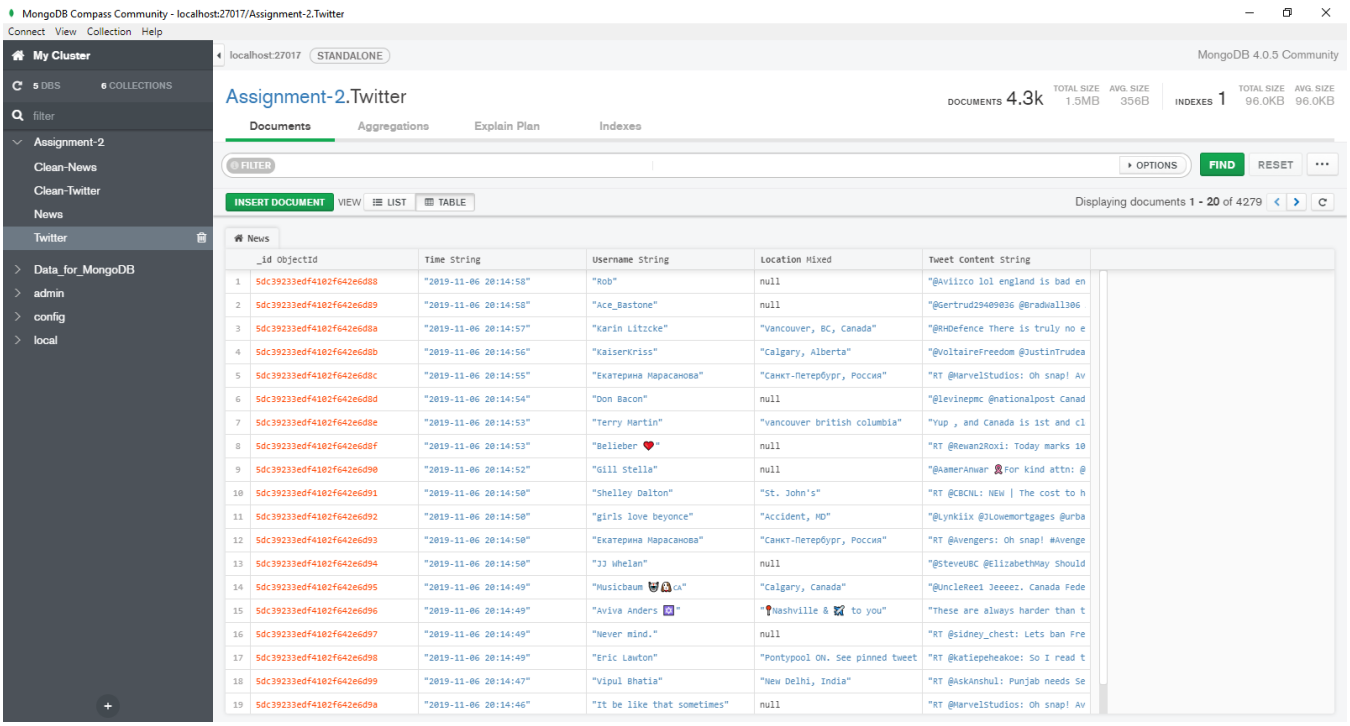
  → Content is extracted from 2 sources - Twitter and NewsAPI datastore.

  → A maximum of 1000 tweets/retweets (Tweet content) were extracted for each of the five keywords – Canada, University, Dalhousie University, Halifax, Canada Education. Along with the tweet/retweet, meta data like the name of the user (Username), time of creation (Time) and location of the user (Location) were also extracted and stored in a csv file

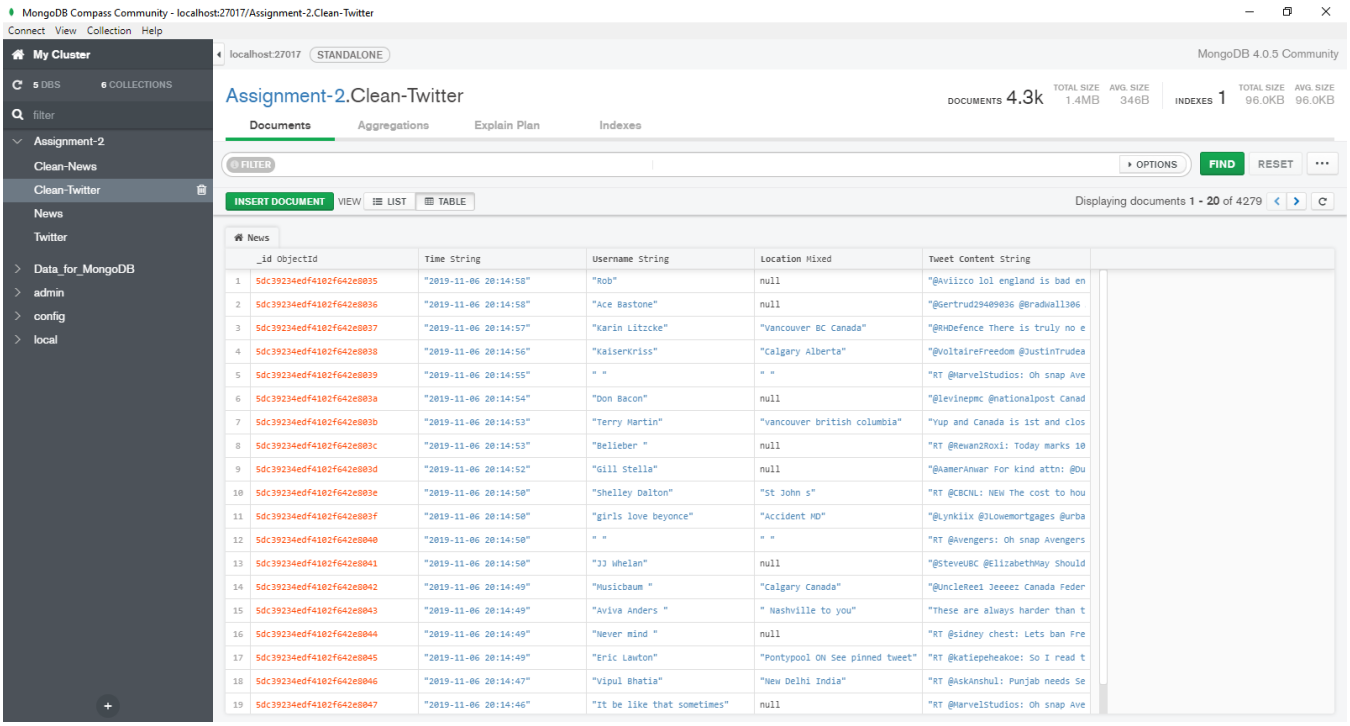
  → Similarly, a maximum of 100 articles (News content) were extracted for each of the five keywords - Canada, University, Dalhousie University, Halifax, Canada Education. Along with them, meta data like the author of the article (Author), the title of the news article (Title) and the time of creation (Time) were also extracted and stored in a csv file.

  → Both the datasets were checked for having special characters i.e. characters apart from our alphabets, numbers, '@' and ':'. These special characters plus any URLs present in the data were removed using Regex. The cleaned data were stored in another csv files.

  → All of these csv files were then uploaded to the local MongoDB instance and the transformation in the data can be seen in the images.



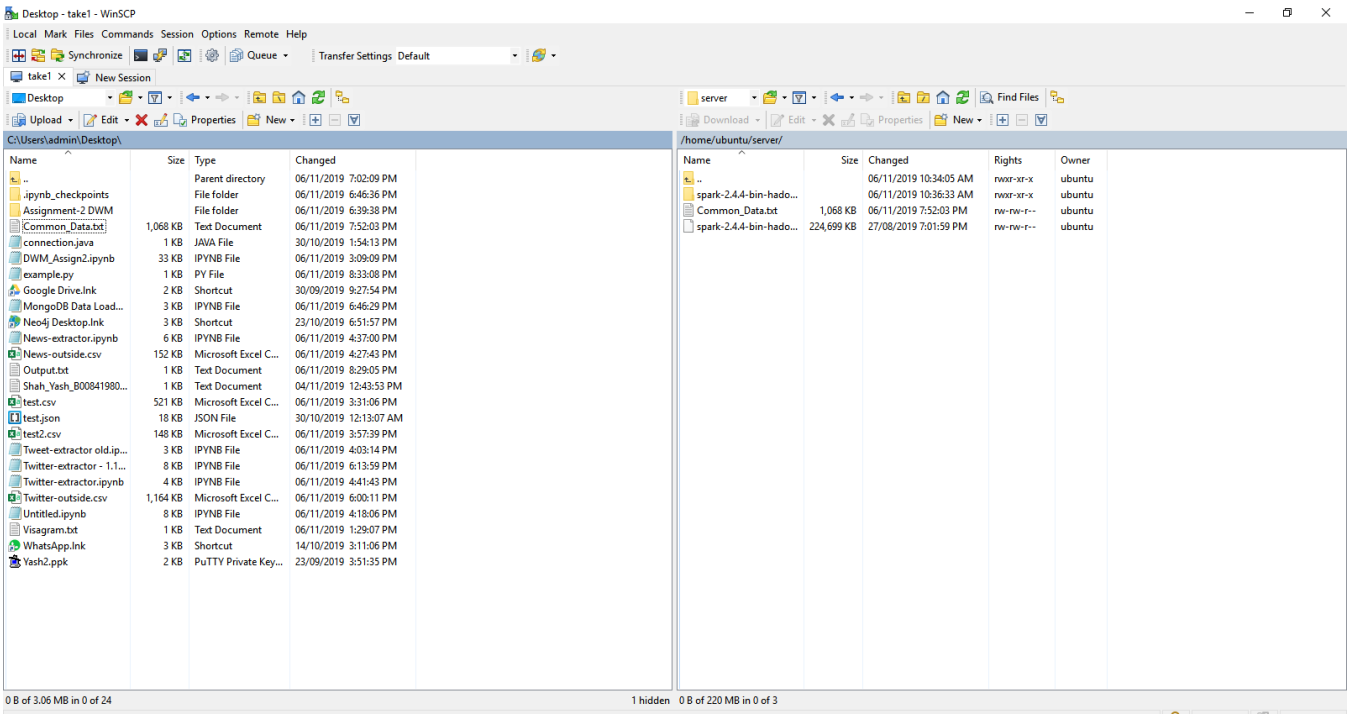
3.Unformatted Twitter Dataset



4.Cleaned Twitter Dataset

- Query Execution:

  → Only the tweets/retweets and news article content were transferred into a text file named ‘Common_Data.txt’.
  → This text file was uploaded to my cloud instance using a software ‘WinSCP’.
  → The MapReduce query was executed on this file and the output was stored in another text file named ‘Output.txt’.



5.Operating window of WinSCP

- Note:

  → File used for extraction of tweets is named ‘Twitter-extractor.ipynb’ and that used for news article is named ‘News-extractor.ipynb’.

  → File used for cleaning the csv files is named ‘Content-cleaner.ipynb’.

  → File used for loading data to MongoDB is named ‘MongoDB Data Loader.ipynb’.

  → The unformatted csv files are named ‘Twitter,csv’ and ‘News.csv’. The cleaned csv files are named ‘Clean-Twitter.csv’ and ‘Clean-News.csv’.

- References:

  → “API Reference for Tweepy”, Joshua Roesslein, 2019. [Online]. Available: http://docs.tweepy.org/en/v3.8.0/api.html#api-reference. [Accessed: 04 November, 2019].

  → “Extracting Twitter Data, Preprocessing and Sentiment Analysis using Python 3.0”, Dilan Jayasekara, 2019. [Online]. Available: https://towardsdatascience.com/extracting-twitter-data-pre-processing-and-sentiment-analysis-using-python-3-0-7192bd8b47cf. [Accessed: 04 November, 2019].