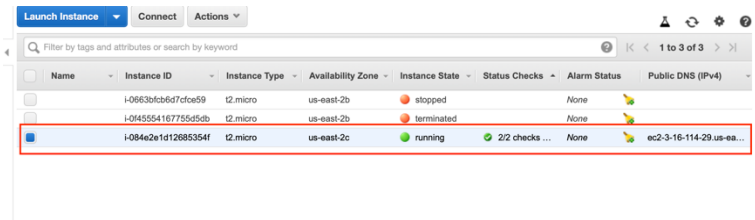
Assignment 2

CSCI5408 -Data Warehousing Management and Analytics

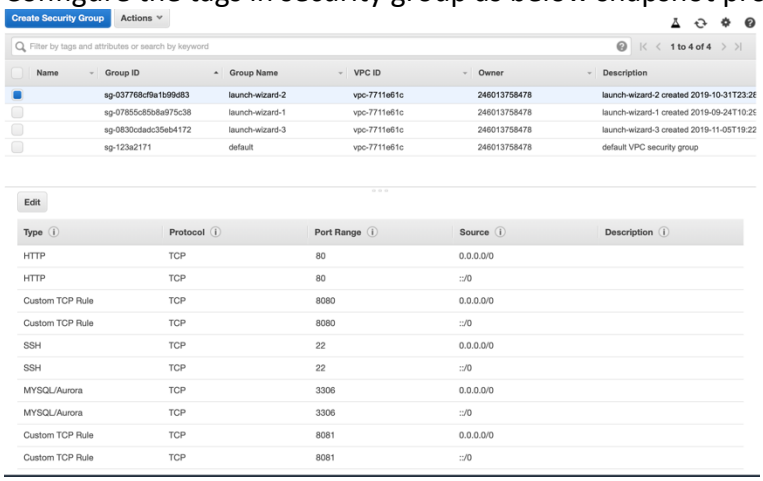
Sources:

tweet_dal.py – Extract data from twitter (imported tweepy and StreamListner libraries)
tweet_clean.py- Cleaning the extracted data(imported pandas and np libraries)
news_api.py – Extract data from newsapi.org(imported requests library)
news_clean.py- Clean the extracted news data(imported pandas and np libraries)
sparkWordCount.py – Perform word count on spark framework with collected and cleaned data (imported sparkContext and sparkConf libraries)

1. **Cloud setup steps [1]:**
   - Sign into AWS instance and click on Launch Instance
   - Select Free Tire and choose "Ubuntu Server 18.04 LTS(HVM), SSD Volume Type
   - Select Next: Configuration Instance Details
   - Click on Add Storage and increase the size to 16 GB
   - Create key value pair and download it
   - Launch the instance

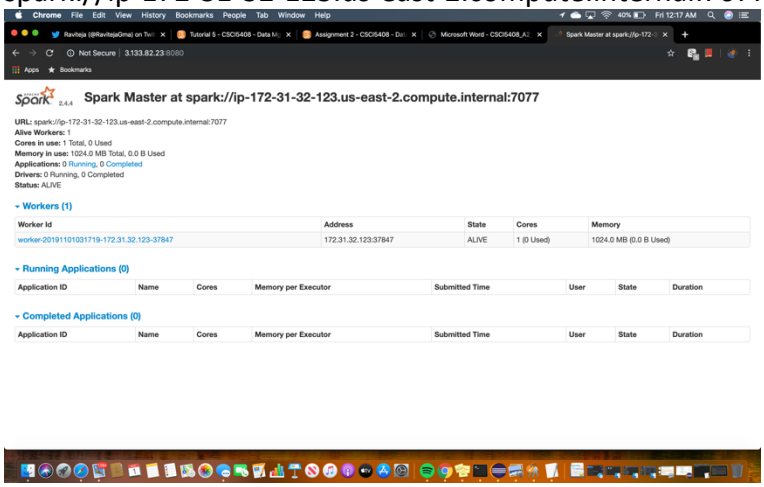


   - Configure the tags in security group as below snapshot provided



   - After setting up from terminal , ssh into AWS instance
   - Once the ubuntu starts running, download and install spark framework from the following commands
   - mkdir server (Create a server folder)
   - wget http://mirror.csclub.waterloo.ca/apache/spark/spark-2.4.4/spark-2.4.4-bin-hadoop2.7.tgz (download the zip file)
   - sudo tar zxvf spark-2.4.4-bin-haddop2.7.tgz (extract the zipped folder)
   - export JAVA_HOME=/usr/lib/jvm/jvm-8-openjdk-amd64/
   - export SPARK_HOME=/server/spark-2.4.4-bin-hadoop2.7

- export PYSPARK_PYTHON=python3
- start the master (sudo ./spark-2.4.4-bin-hadoop2.7/sbin/start-master.sh)
- start the slave(sudo ./spark-2.4.4-bin-hadoop2.7/sbin/start-master.sh spark://ip-172-31-32-123.us-east-2.compute.internal:7077)



2. **Data Extraction Process:**
   - Tweet data is extracted after creating developer account[2] and using keys and secrets generated after creating an app. (tweet_dal.py script).
   - Tweet data is extracted using both search API[4] (1750 tweets) and stream API [3] (1750 tweets) , with the list of keywords provided and written into the OutputStreaming.csv file.
     Search API:
     Tweets are limited to 1750 by setting limit on tweet cursor
     Stream API:
     Tweets are limited to 1750 , by initialising the num_tweets variable to zero and setting a limit to 1750 and incremented in on_status method by value 1 and written to data OutputStreaming.csv file.
   - Screen name, text, created at, location of the user, place from where it's been tweeted, Retweeted status, retweeted count are the fields collected for each tweet .
   - News articles are extracted by fetching the JSON responses[6] from news _url and with news_api key after creating a developer account [5] and appending the data to news.csv file .(news_api.py script)
   - Author, title, Description, Content , published at are the fields extracted for each article.
3. **Cleaning Process:**
   - OutputStreaming.csv is read and cleaned by removing urls , removing emojis and special characters (tweet_clean.py script)[3]
   - Cleaned data is placed in tweet_cleaned.csv for loading into mongoDB and tweet_cleaned.txt for processing in spark .
   - news.csv is read and cleaned by removing tags , removing emojis and special characters (news_clean.py script)[3]

- Cleaned data is placed in news_cleaned.csv for loading into mongoDB and news_cleaned.txt for processing in spark .
- Tweet_data in MongoDB:

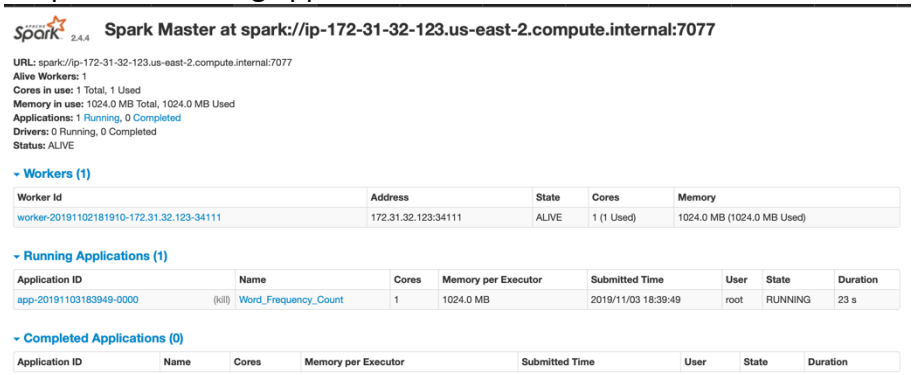
```
Desktop — ubuntu@ip-172-31-32-123: ~/csvFiles — ssh ◂ sudo — 80×24
> show collections
news_data
tweet_data
> db.tweet_data.find().pretty()
{
        "_id" : ObjectId("5dc1f2ed600344f1c45d92f9"),
        "Author" : "FrankMeloche",
        "Date" : "2019-11-05 21:47:18",
        "Text" : "RT nspector4 No one can begrudge Trudeau taking a break after
 a gruelling elxn43 campaign to recharge his batteries. Flying across Canad",
        "Location" : NaN,
        "TweetedFromLocation" : NaN,
        "RetweetStatus" : "False",
        "RetweetCount" : 0
}
{
        "_id" : ObjectId("5dc1f2ed600344f1c45d92fa"),
        "Author" : "hananelissian2",
        "Date" : "2019-11-05 21:47:19",
        "Text" : "RT AmrMoha94051490 Elissa Will Performing Live 22nd of Novembe
r at Theatre Rialto in Montreal Canada For Reservations 1 866 908",
        "Location" : NaN,
        "TweetedFromLocation" : NaN,
        "RetweetStatus" : "False",
```

- News_data in MongoDB:

```
Desktop — ubuntu@ip-172-31-32-123: ~/csvFiles — ssh ◂ sudo — 80×24
}
Type "it" for more
> db.news_data.find().pretty()
{
        "_id" : ObjectId("5dc1f321600344f1c45d9a38"),
        "author" : "Kirsten Korosec",
        "title" : "Uber Freight expands app to Canada",
        "description" : "Uber Freight, the Uber business unit that helps truck d
rivers connect with shipping companies, said Wednesday its launching the app in
Canada as part of its global expansion plan. The move into Canada will give Uber
 Freight access to the countrys 68 billio",
        "content" : "Uber Freight, the Uber business unit that helps truck drive
rs connect with shipping companies, said Wednesday its launching the app in Cana
da as part of its global expansion plan. The move into Canada will give Uber Fre
ight access to the countrys 68 billi 2181 chars ",
        "date" : "2019-10-30T11:37:16Z"
}
{
        "_id" : ObjectId("5dc1f321600344f1c45d9a39"),
        "author" : "Ian Austen and Dan Bilefsky",
        "title" : "Canada Votes Today. Heres What You Need to Know",
        "description" : "Prime Minister Justin Trudeau faces the political battl
e of his life in a tight race.",
        "content" : "Mr. Scheers authenticity was also challenged when it emerge
```

4. Sample CSV Files after cleaning:
   Files are placed in the zipped folder
   tweet_cleaned.csv-cleaned csv file for tweets data
   news_cleaned.csv- cleaned csv files for news data
   tweet_cleaned.txt- cleaned text file for tweets data
   news_cleaned.txt- cleaned txt file for news data
5. **Word Count in spark framework:**
   - Spark application is created and the master spark url is provided for creating app (sparkWordCount.py)[7]
   - Map reduce method is used to count the single words by reading the text files into RDD and then using flatMap and splitting the words and used map and reduce to get the count.[8]
   - The collected RDD is stored into Dictionary and count of words are accessed.
   - To count bi gram words , textfiles are read into RDD and using map words are split and joined for each pair.

- Such collected pairs are counted s=again using map and reduce and the obtained RDD is stored in Dictionary and the values are accessed for each bi gram word.
- Final count of the words is stored in Output.txt file .
- Such created python script is stored in AWS cluster and submitted to run through the command[7] to run in pyspark - sudo ./spark-2.4.4-bin-hadoop2.7/bin/spark-submit --deploy-mode client sparkWordCount.py
- Snapshot of running application on the cloud dashboard.



Spark 2.4.4 **Spark Master at spark://ip-172-31-32-123.us-east-2.compute.internal:7077**

URL: spark://ip-172-31-32-123.us-east-2.compute.internal:7077
Alive Workers: 1
Cores in use: 1 Total, 1 Used
Memory in use: 1024.0 MB Total, 1024.0 MB Used
Applications: 1 Running, 0 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers (1)

Worker Id	Address	State	Cores	Memory
worker-20191102181910-172.31.32.123-34111	172.31.32.123:34111	ALIVE	1 (1 Used)	1024.0 MB (1024.0 MB Used)

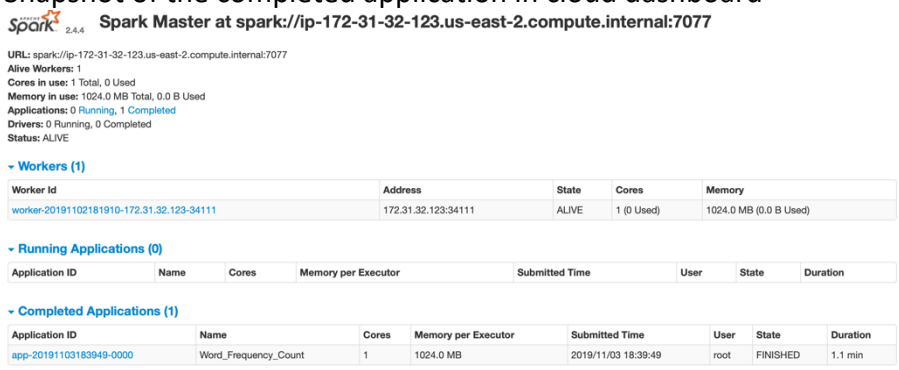
Running Applications (1)

Application ID	Name	Cores	Memory per Executor	Submitted Time	User	State	Duration
app-20191103183949-0000 (kill)	Word_Frequency_Count	1	1024.0 MB	2019/11/03 18:39:49	root	RUNNING	23 s

Completed Applications (0)

Application ID	Name	Cores	Memory per Executor	Submitted Time	User	State	Duration

- Snapshot of the completed application in cloud dashboard



Spark 2.4.4 **Spark Master at spark://ip-172-31-32-123.us-east-2.compute.internal:7077**

URL: spark://ip-172-31-32-123.us-east-2.compute.internal:7077
Alive Workers: 1
Cores in use: 1 Total, 0 Used
Memory in use: 1024.0 MB Total, 0.0 B Used
Applications: 0 Running, 1 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers (1)

Worker Id	Address	State	Cores	Memory
worker-20191102181910-172.31.32.123-34111	172.31.32.123:34111	ALIVE	1 (0 Used)	1024.0 MB (0.0 B Used)

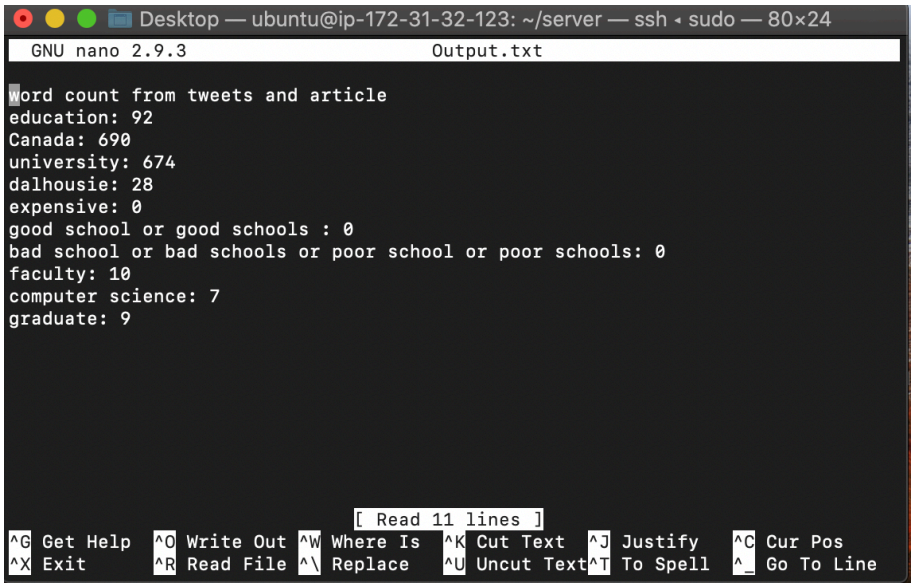
Running Applications (0)

Application ID	Name	Cores	Memory per Executor	Submitted Time	User	State	Duration

Completed Applications (1)

Application ID	Name	Cores	Memory per Executor	Submitted Time	User	State	Duration
app-20191103183949-0000	Word_Frequency_Count	1	1024.0 MB	2019/11/03 18:39:49	root	FINISHED	1.1 min

- Output.txt file from AWS instance



word count from tweets and article
education: 92
Canada: 690
university: 674
dalhousie: 28
expensive: 0
good school or good schools : 0
bad school or bad schools or poor school or poor schools: 0
faculty: 10
computer science: 7
graduate: 9

References

[1].Tutorial 5 [Online]: Lab Slides in *Brightspace.*
[2].Apply for access – Twitter Developers .[Online].
Available: https://developer.twitter.com/en/apply-for-access.html . [Accessed October 24th, 2019].
[3]. (Almost) Real-Time Twitter Sentiment Analysis with Tweep & Vader. [Online].
Available: https://towardsdatascience.com/almost-real-time-twitter-sentiment-analysis-with-tweep-vader-f88ed5b93b1c . [Accessed October 28th , 2019].
[4].Python Twitter Search API.[Online].
Available: https://twitterdev.github.io/search-tweets-python/ . [Accessed October 29th,2019]
[5].News API – A JSON API for live news and blog articles. [Online].
Available: https://newsapi.org/register .[Accessed October 25th , 2019].
[6].News API: Extracting News Headlines and Articles. [Online].
Available: https://python.gotrained.com/news-api/ .[Accessed October 28th ,2019].
[7]. How to Run an Application on Spark Standalone Cluster.[Online].

Available: https://medium.com/@cxu24/how-to-run-an-application-on-spark-standalone-cluster-3168ec12ba68 .[Accessed November 3rd, 2019].
[8].Examples |Apache Spark – The Apache Software Foundation !. [Online]. Available : https://spark.apache.org/examples.html .[Accessed November1st, 2019].