DALHOUSIE UNIVERSITY

CSCI 5408 – Data Management, Warehousing, Analytics

Assignment 4

Work done by,

Name: Guturu Rama Mohan Vishnu

Banner ID: B00871849

Email: rm286720@dal.ca

DECLARATION

I, Guturu Rama Mohan Vishnu, declare that in assignment 3 of CSCI 5408 course, I declare that all the work done was done by myself and I have not collaborated with anyone for the assignment.

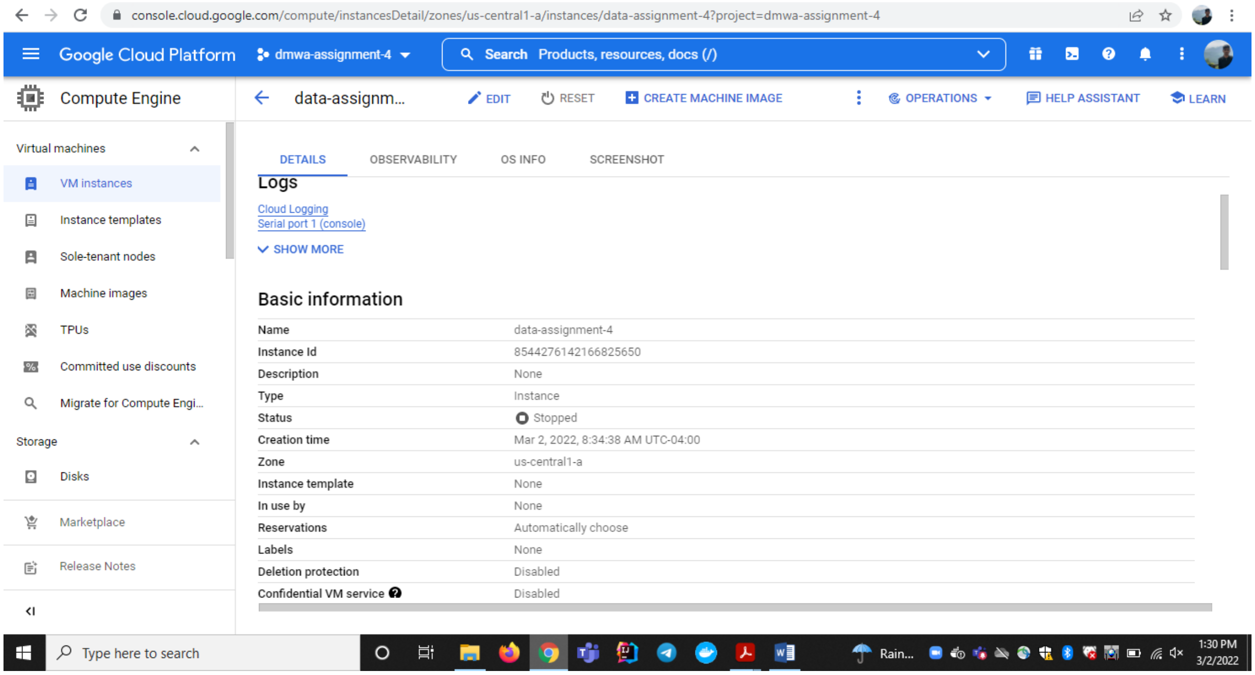
Problem #1

Task #1: Cluster Setup – Apache Spark Framework on GCP

In order to explain how I completed this task, I am creating a flowchart and adding it to a folder in my drive. I am sharing the link of that drive in this pdf and the reason I couldn't add the flowchart here is that the image is too long and it won't fit in a single page.

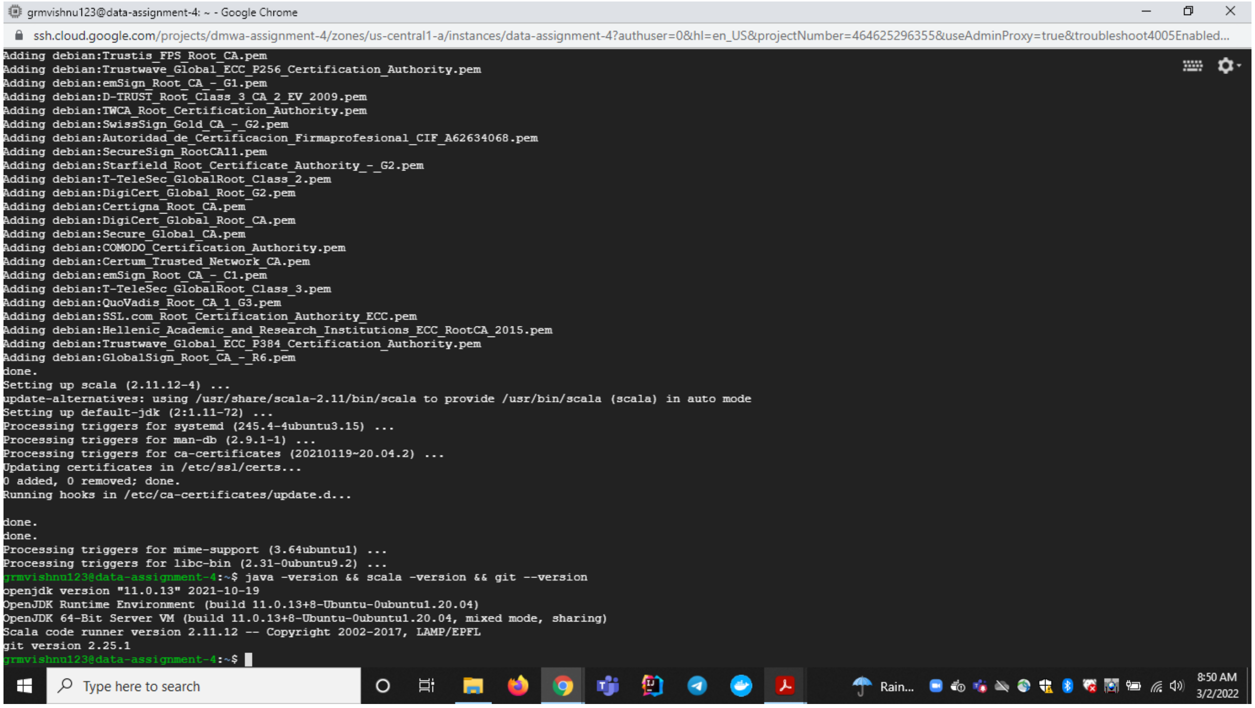
Drive link for flowchart: https://drive.google.com/drive/folders/1RnxczW2VeqWuyzAlfIV3kX0UX3o-mGIN?usp=sharing

Also, I am adding the screenshots of the work alongside the flowchart.

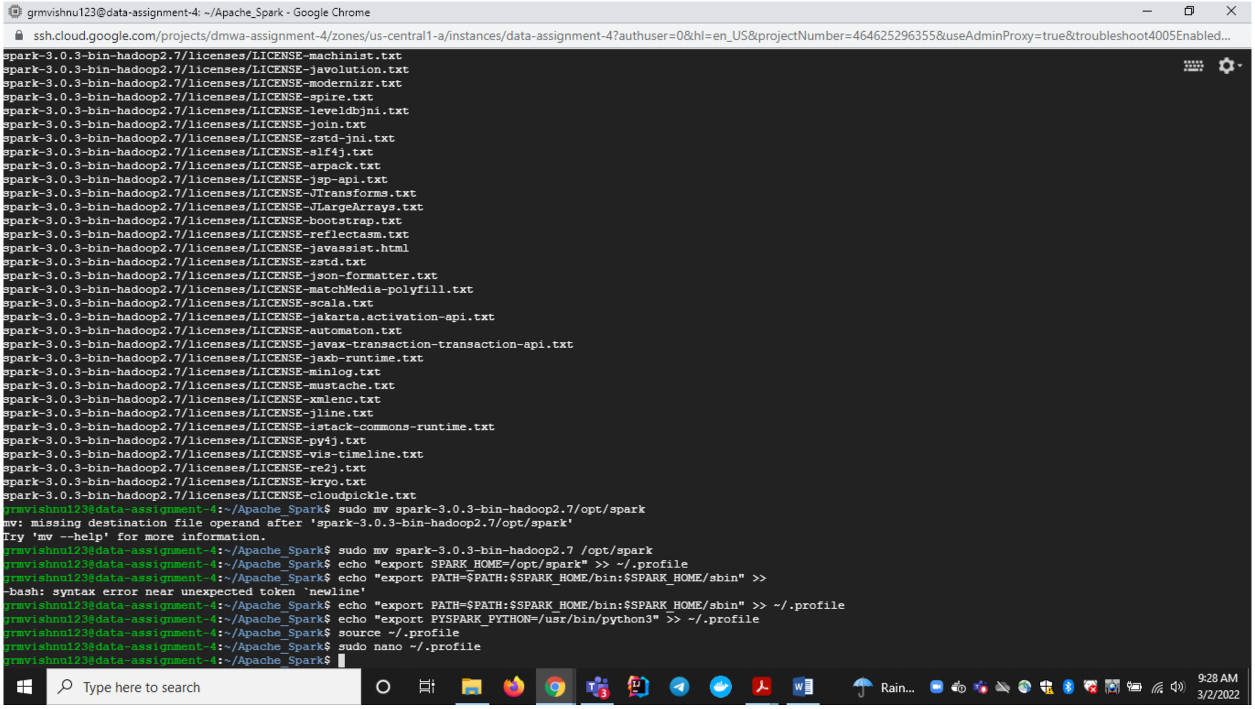


Basic information

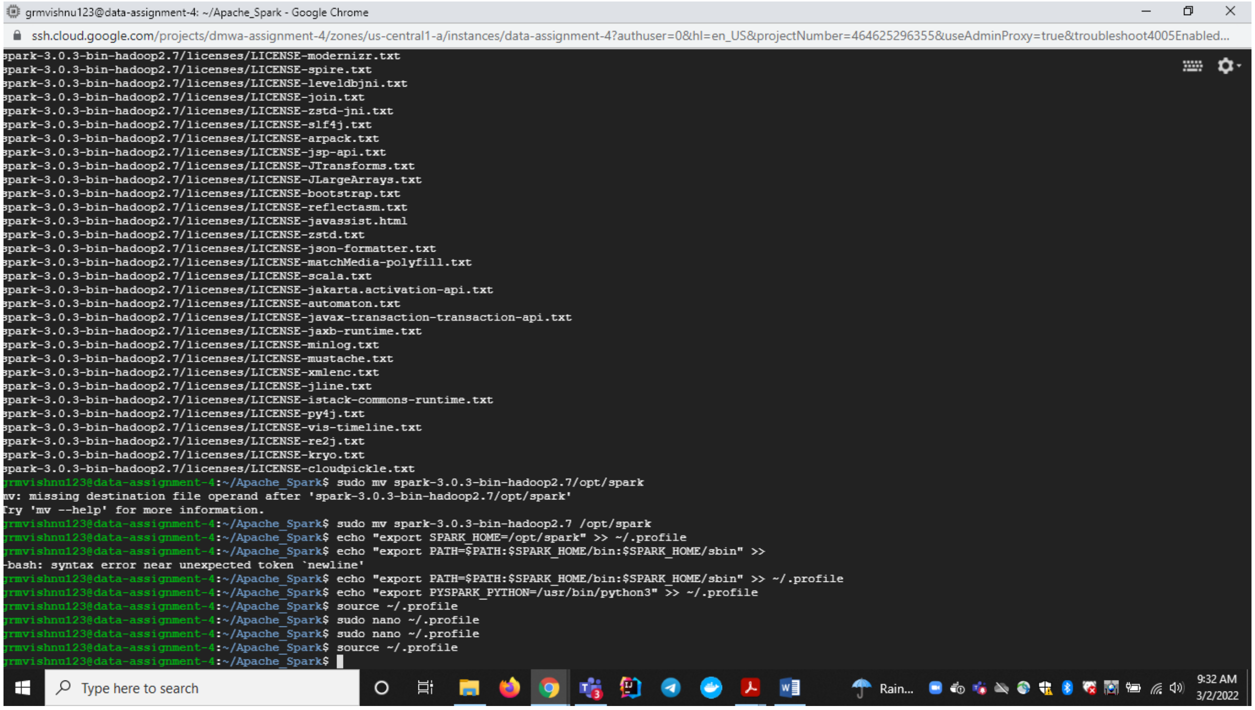
Name	data-assignment-4
Instance Id	8544276142166825650
Description	None
Type	Instance
Status	Stopped
Creation time	Mar 2, 2022, 8:34:38 AM UTC-04:00
Zone	us-central1-a
Instance template	None
In use by	None
Reservations	Automatically choose
Labels	None
Deletion protection	Disabled
Confidential VM service	Disabled


grmvishnu123@data-assignment-4: ~ - Google Chrome
ssh.cloud.google.com/projects/dmwa-assignment-4/zones/us-central1-a/instances/data-assignment-4?authuser=0&hl=en_US&projectNumber=464625296355&useAdminProxy=true&troubleshoot4005Enabled...
Type here to search
Rain...
8:50 AM
3/2/2022

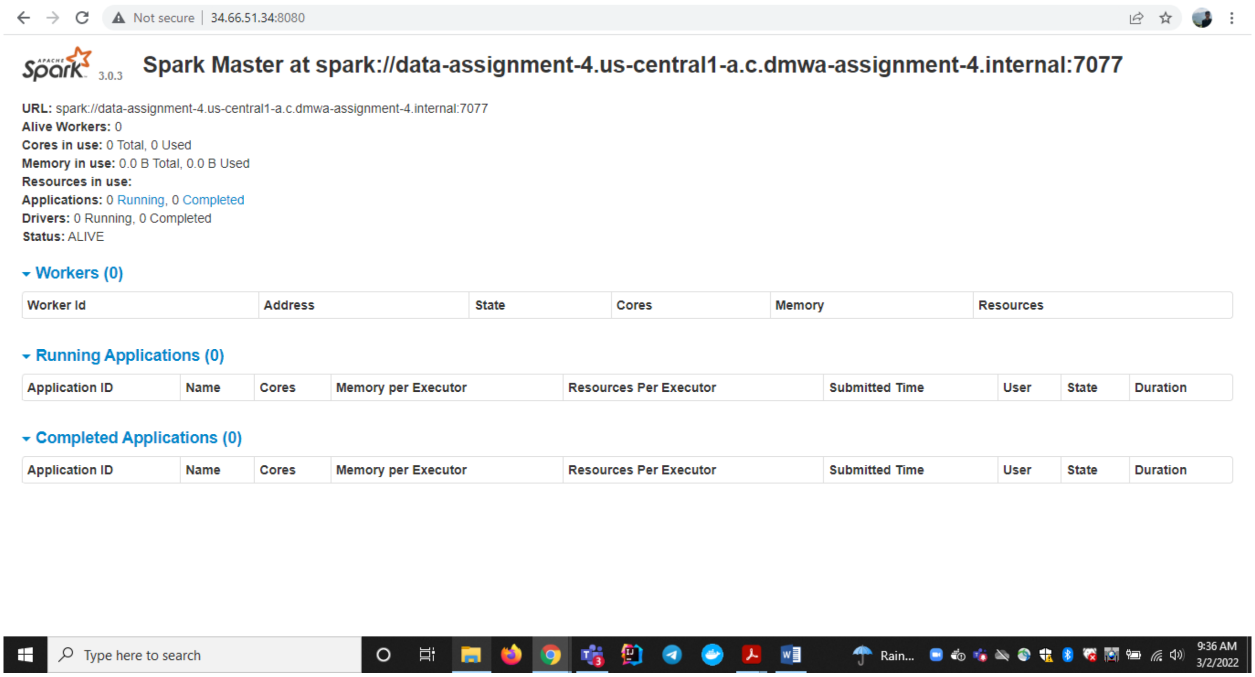
```
Adding debian:Trustis_FPS_Root_CA.pem
Adding debian:Trustwave_Global_ECC_P256_Certification_Authority.pem
Adding debian:emSign_Root_CA_-_G1.pem
Adding debian:D-TRUST_Root_Class_3_CA_2_EV_2009.pem
Adding debian:TWCA_Root_Certification_Authority.pem
Adding debian:SwissSign_Gold_CA_-_G2.pem
Adding debian:Autoridad_de_Certificacion_Firmaprofesional_CIF_A62634068.pem
Adding debian:SecureSign_RootCA11.pem
Adding debian:Starfield_Root_Certificate_Authority_-_G2.pem
Adding debian:T-TeleSec_GlobalRoot_Class_2.pem
Adding debian:DigiCert_Global_Root_G2.pem
Adding debian:Certigna_Root_CA.pem
Adding debian:DigiCert_Global_Root_CA.pem
Adding debian:Secure_Global_CA.pem
Adding debian:COMODO_Certification_Authority.pem
Adding debian:Certum_Trusted_Network_CA.pem
Adding debian:emSign_Root_CA_-_C1.pem
Adding debian:T-TeleSec_GlobalRoot_Class_3.pem
Adding debian:QuoVadis_Root_CA_1_G3.pem
Adding debian:SSL.com_Root_Certification_Authority_ECC.pem
Adding debian:Hellenic_Academic_and_Research_Institutions_ECC_RootCA_2015.pem
Adding debian:Trustwave_Global_ECC_P384_Certification_Authority.pem
Adding debian:GlobalSign_Root_CA_-_R6.pem
done.
Setting up scala (2.11.12-4) ...
update-alternatives: using /usr/share/scala-2.11/bin/scala to provide /usr/bin/scala (scala) in auto mode
Setting up default-jdk (2:1.11-72) ...
Processing triggers for systemd (245.4-4ubuntu3.15) ...
Processing triggers for man-db (2.9.1-1) ...
Processing triggers for ca-certificates (20210119~20.04.2) ...
Updating certificates in /etc/ssl/certs...
0 added, 0 removed; done.
Running hooks in /etc/ca-certificates/update.d...

done.
done.
Processing triggers for mime-support (3.64ubuntu1) ...
Processing triggers for libc-bin (2.31-0ubuntu9.2) ...
grmvishnu123@data-assignment-4:~$ java -version && scala -version && git --version
openjdk version "11.0.13" 2021-10-19
OpenJDK Runtime Environment (build 11.0.13+8-Ubuntu-0ubuntu1.20.04)
OpenJDK 64-Bit Server VM (build 11.0.13+8-Ubuntu-0ubuntu1.20.04, mixed mode, sharing)
Scala code runner version 2.11.12 -- Copyright 2002-2017, LAMP/EPFL
git version 2.25.1
grmvishnu123@data-assignment-4:~$ 
```


grmvishnu123@data-assignment-4: ~/Apache_Spark - Google Chrome
ssh.cloud.google.com/projects/dmwa-assignment-4/zones/us-central1-a/instances/data-assignment-4?authuser=0&hl=en_US&projectNumber=464625296355&useAdminProxy=true&troubleshoot4005Enabled...
Type here to search
Rain...
9:28 AM
3/2/2022

```
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-machinist.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-javolution.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-modernizr.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-spire.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-leveldbjni.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-join.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-zstd-jni.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-slf4j.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-arpack.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-jsp-api.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-JTransforms.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-JLargeArrays.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-bootstrap.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-reflectasm.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-javassist.html
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-zstd.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-json-formatter.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-matchMedia-polyfill.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-scala.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-jakarta.activation-api.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-automaton.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-javax-transaction-transaction-api.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-jaxb-runtime.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-minlog.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-mustache.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-xmlenc.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-jline.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-istack-commons-runtime.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-py4j.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-vis-timeline.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-re2j.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-kryo.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-cloudpickle.txt
grmvishnu123@data-assignment-4:~/Apache_Spark$ sudo mv spark-3.0.3-bin-hadoop2.7/opt/spark
mv: missing destination file operand after 'spark-3.0.3-bin-hadoop2.7/opt/spark'
Try 'mv --help' for more information.
grmvishnu123@data-assignment-4:~/Apache_Spark$ sudo mv spark-3.0.3-bin-hadoop2.7 /opt/spark
grmvishnu123@data-assignment-4:~/Apache_Spark$ echo "export SPARK_HOME=/opt/spark" >> ~/.profile
grmvishnu123@data-assignment-4:~/Apache_Spark$ echo "export PATH=$PATH:$SPARK_HOME/bin:$SPARK_HOME/sbin" >>
-bash: syntax error near unexpected token `newline'
grmvishnu123@data-assignment-4:~/Apache_Spark$ echo "export PATH=$PATH:$SPARK_HOME/bin:$SPARK_HOME/sbin" >> ~/.profile
grmvishnu123@data-assignment-4:~/Apache_Spark$ echo "export PYSPARK_PYTHON=/usr/bin/python3" >> ~/.profile
grmvishnu123@data-assignment-4:~/Apache_Spark$ source ~/.profile
grmvishnu123@data-assignment-4:~/Apache_Spark$ sudo nano ~/.profile
grmvishnu123@data-assignment-4:~/Apache_Spark$ 
```


grmvishnu123@data-assignment-4: ~/Apache_Spark - Google Chrome
ssh.cloud.google.com/projects/dmwa-assignment-4/zones/us-central1-a/instances/data-assignment-4?authuser=0&hl=en_US&projectNumber=464625296355&useAdminProxy=true&troubleshoot4005Enabled...
Type here to search
Rain...
9:32 AM
3/2/2022

```
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-modernizr.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-spire.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-leveldbjni.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-join.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-zstd-jni.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-slf4j.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-arpack.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-jsp-api.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-JTransforms.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-JLargeArrays.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-bootstrap.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-reflectasm.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-javassist.html
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-zstd.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-json-formatter.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-matchMedia-polyfill.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-scala.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-jakarta.activation-api.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-automaton.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-javax-transaction-transaction-api.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-jaxb-runtime.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-minlog.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-mustache.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-xmlenc.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-jline.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-istack-commons-runtime.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-py4j.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-vis-timeline.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-re2j.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-kryo.txt
spark-3.0.3-bin-hadoop2.7/licenses/LICENSE-cloudpickle.txt
grmvishnu123@data-assignment-4:~/Apache_Spark$ sudo mv spark-3.0.3-bin-hadoop2.7/opt/spark
mv: missing destination file operand after 'spark-3.0.3-bin-hadoop2.7/opt/spark'
Try 'mv --help' for more information.
grmvishnu123@data-assignment-4:~/Apache_Spark$ sudo mv spark-3.0.3-bin-hadoop2.7 /opt/spark
grmvishnu123@data-assignment-4:~/Apache_Spark$ echo "export SPARK_HOME=/opt/spark" >> ~/.profile
grmvishnu123@data-assignment-4:~/Apache_Spark$ echo "export PATH=$PATH:$SPARK_HOME/bin:$SPARK_HOME/sbin" >>
-bash: syntax error near unexpected token `newline'
grmvishnu123@data-assignment-4:~/Apache_Spark$ echo "export PATH=$PATH:$SPARK_HOME/bin:$SPARK_HOME/sbin" >> ~/.profile
grmvishnu123@data-assignment-4:~/Apache_Spark$ echo "export PYSPARK_PYTHON=/usr/bin/python3" >> ~/.profile
grmvishnu123@data-assignment-4:~/Apache_Spark$ source ~/.profile
grmvishnu123@data-assignment-4:~/Apache_Spark$ sudo nano ~/.profile
grmvishnu123@data-assignment-4:~/Apache_Spark$ sudo nano ~/.profile
grmvishnu123@data-assignment-4:~/Apache_Spark$ source ~/.profile
grmvishnu123@data-assignment-4:~/Apache_Spark$
```



Spark Master at spark://data-assignment-4.us-central1-a.c.dmwa-assignment-4.internal:7077

URL: spark://data-assignment-4.us-central1-a.c.dmwa-assignment-4.internal:7077
Alive Workers: 0
Cores in use: 0 Total, 0 Used
Memory in use: 0.0 B Total, 0.0 B Used
Resources in use:
Applications: 0 Running, 0 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers (0)

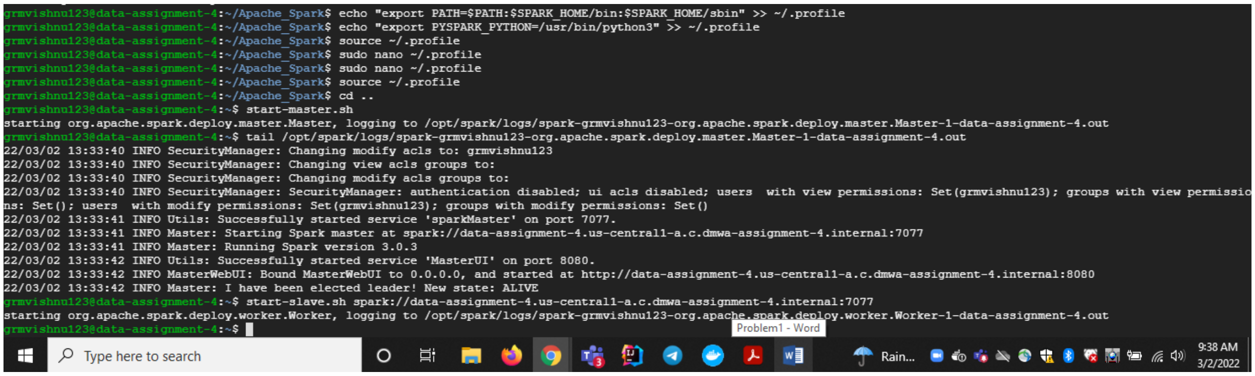
Worker Id	Address	State	Cores	Memory	Resources

Running Applications (0)

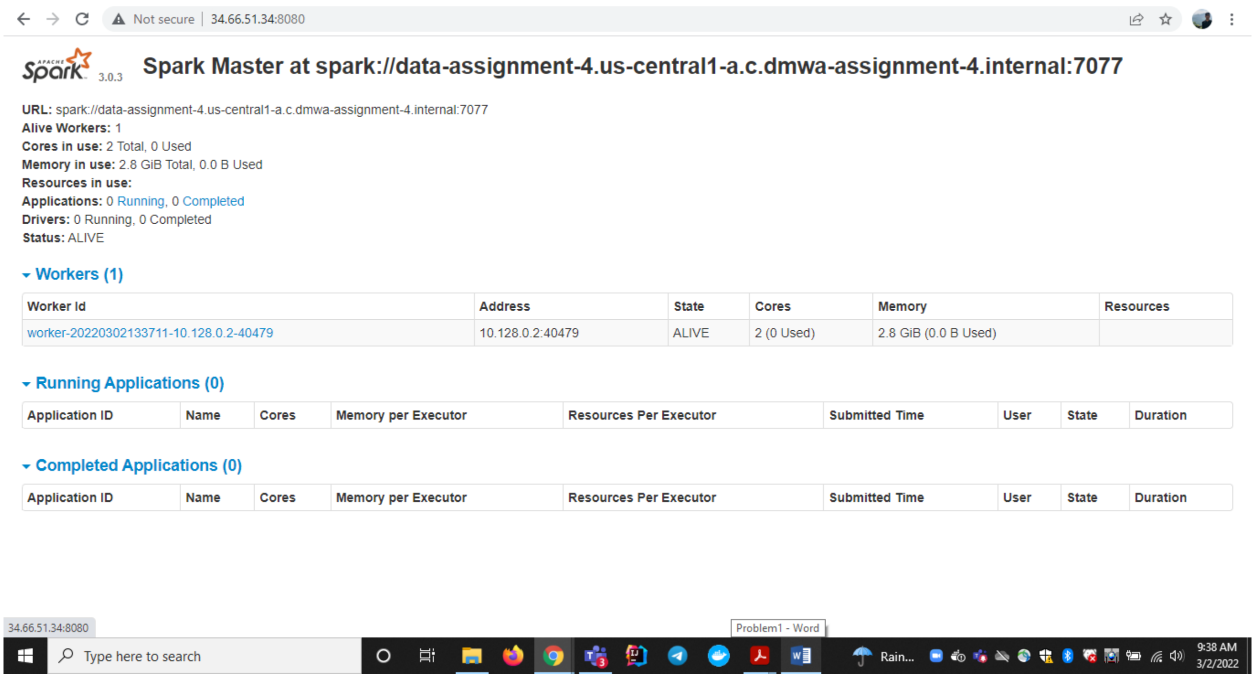
Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration

Completed Applications (0)

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



```
grmvishnu123@data-assignment-4:~/Apache_Spark$ echo "export PATH=$PATH:$SPARK_HOME/bin:$SPARK_HOME/sbin" >> ~/.profile
grmvishnu123@data-assignment-4:~/Apache_Spark$ echo "export PYSPARK_PYTHON=/usr/bin/python3" >> ~/.profile
grmvishnu123@data-assignment-4:~/Apache_Spark$ source ~/.profile
grmvishnu123@data-assignment-4:~/Apache_Spark$ sudo nano ~/.profile
grmvishnu123@data-assignment-4:~/Apache_Spark$ sudo nano ~/.profile
grmvishnu123@data-assignment-4:~/Apache_Spark$ source ~/.profile
grmvishnu123@data-assignment-4:~/Apache_Spark$ cd ..
grmvishnu123@data-assignment-4:~$ start-master.sh
starting org.apache.spark.deploy.master.Master, logging to /opt/spark/logs/spark-grmvishnu123-org.apache.spark.deploy.master.Master-1-data-assignment-4.out
grmvishnu123@data-assignment-4:~$ tail /opt/spark/logs/spark-grmvishnu123-org.apache.spark.deploy.master.Master-1-data-assignment-4.out
22/03/02 13:33:40 INFO SecurityManager: Changing modify acls to: grmvishnu123
22/03/02 13:33:40 INFO SecurityManager: Changing view acls groups to:
22/03/02 13:33:40 INFO SecurityManager: Changing modify acls groups to:
22/03/02 13:33:40 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  with view permissions: Set(grmvishnu123); groups with view permissio
ns: Set(); users  with modify permissions: Set(grmvishnu123); groups with modify permissions: Set()
22/03/02 13:33:41 INFO Utils: Successfully started service 'sparkMaster' on port 7077.
22/03/02 13:33:41 INFO Master: Starting Spark master at spark://data-assignment-4.us-central1-a.c.dmwa-assignment-4.internal:7077
22/03/02 13:33:41 INFO Master: Running Spark version 3.0.3
22/03/02 13:33:42 INFO Utils: Successfully started service 'MasterUI' on port 8080.
22/03/02 13:33:42 INFO MasterWebUI: Bound MasterWebUI to 0.0.0.0, and started at http://data-assignment-4.us-central1-a.c.dmwa-assignment-4.internal:8080
22/03/02 13:33:42 INFO Master: I have been elected leader! New state: ALIVE
grmvishnu123@data-assignment-4:~$ start-slave.sh spark://data-assignment-4.us-central1-a.c.dmwa-assignment-4.internal:7077
starting org.apache.spark.deploy.worker.Worker, logging to /opt/spark/logs/spark-grmvishnu123-org.apache.spark.deploy.worker.Worker-1-data-assignment-4.out
grmvishnu123@data-assignment-4:~$
```



Spark Master at spark://data-assignment-4.us-central1-a.c.dmwa-assignment-4.internal:7077

URL: spark://data-assignment-4.us-central1-a.c.dmwa-assignment-4.internal:7077
Alive Workers: 1
Cores in use: 2 Total, 0 Used
Memory in use: 2.8 GiB Total, 0.0 B Used
Resources in use:
Applications: 0 Running, 0 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers (1)

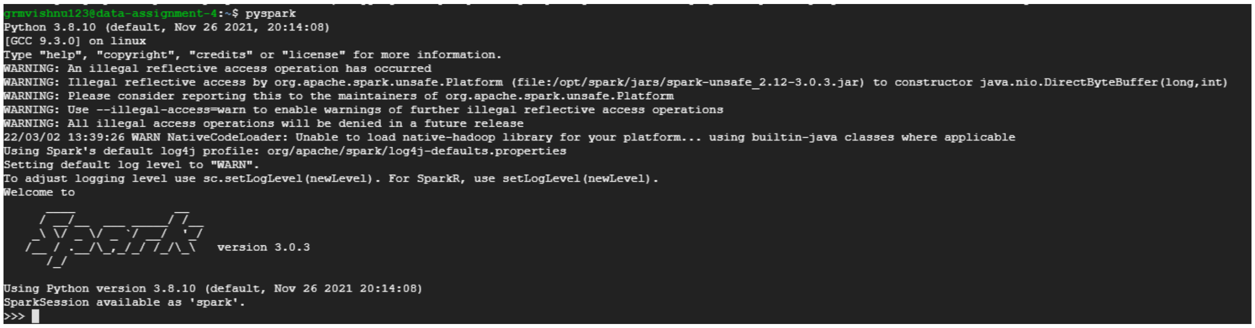
Worker Id	Address	State	Cores	Memory	Resources
worker-20220302133711-10.128.0.2-40479	10.128.0.2:40479	ALIVE	2 (0 Used)	2.8 GiB (0.0 B Used)	

Running Applications (0)

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

Completed Applications (0)

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

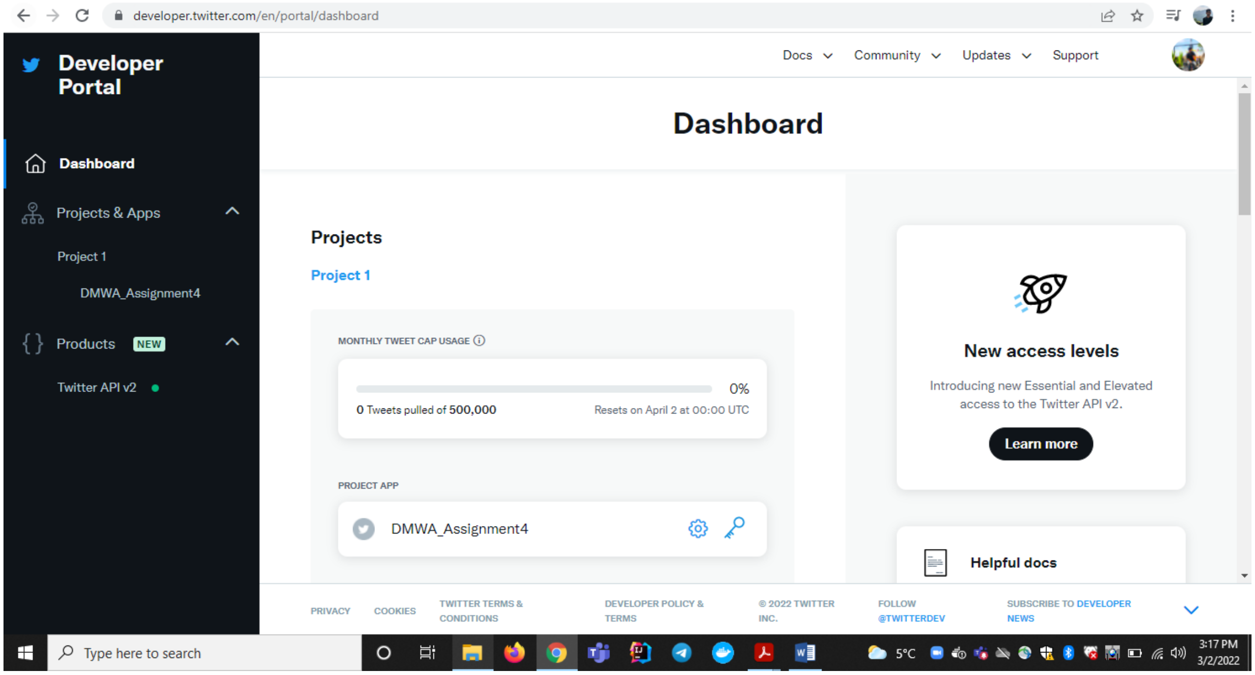


```
grmvishnu123@data-assignment-4:~$ pyspark
Python 3.8.10 (default, Nov 26 2021, 20:14:08)
[GCC 9.3.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.spark.unsafe.Platform (file:/opt/spark/jars/spark-unsafe_2.12-3.0.3.jar) to constructor java.nio.DirectByteBuffer(long,int)
WARNING: Please consider reporting this to the maintainers of org.apache.spark.unsafe.Platform
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
22/03/02 13:39:26 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 3.0.3
      /_/

Using Python version 3.8.10 (default, Nov 26 2021 20:14:08)
SparkSession available as 'spark'.
>>>
```

Task #2: Data Extraction and Preprocessing Engine

Step 1:

Created a developer account to access the tweets. And I created a project named "Project 1" and created an app in it called "DMWA_Assignment4" with the required keys.



0 Tweets pulled of 500,000 Resets on April 2 at 00:00 UTC

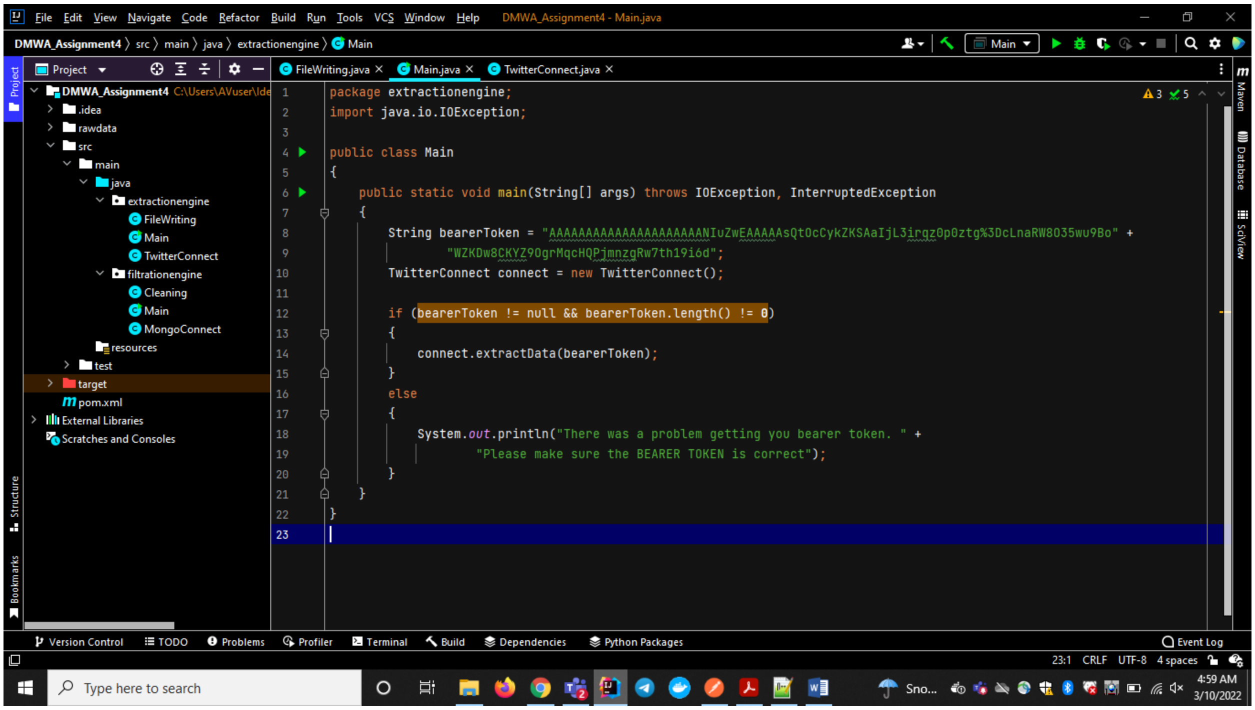
Introducing new Essential and Elevated access to the Twitter API v2.

Step 2:

Searching tweets is a very useful and important feature in developing applications related to tweet data and metadata. To help developers in searching tweets, twitter provides us with two endpoints: recent search and full-archive search. Operators can be combined into set of queries with the help of boolean logic and parentheses to help in refining the queries. These endpoints allow you to navigate the results by time and tweet ID ranges once you’ve set up the query and started getting tweets. Developers can use the filtered stream endpoint group to filter the real-time stream of public tweets. Multiple endpoints in this endpoint group allows to build and manage rules, as well as apply those rules to filter a stream of real-time tweets and return matching public tweets. The REST rules endpoint allows developers to add and remove rules from a persistent stream connection without having to disconnect. After adding a set of rules, we can create a streaming connection that will begin delivering tweet objects in JSON format over a persistent HTTP Streaming connection. There are 3 levels of access in filtering: Essential, Elevated and Academic Research.

Step 3:

Git URL for java code to extract data from twitter: https://git.cs.dal.ca/rguturu/csci-5408-w2022-b00871849-gutururamamohanvishnu/-/tree/main/assignment-4



```java
package extractionengine;
import java.io.IOException;

public class Main
{
    public static void main(String[] args) throws IOException, InterruptedException
    {
        String bearerToken = "AAAAAAAAAAAAAAAAAAAAANIuZwEAAAAAsQtOcCykZKSAaIjL3irqz0p0ztg%3DcLnaRW8035wu9Bo" +
                "WZKDw8CKYZ9OgrMqcHQPjmnzgRw7th19i6d";
        TwitterConnect connect = new TwitterConnect();

        if (bearerToken != null && bearerToken.length() != 0)
        {
            connect.extractData(bearerToken);
        }
        else
        {
            System.out.println("There was a problem getting you bearer token. " +
                    "Please make sure the BEARER TOKEN is correct");
        }
    }
}

```

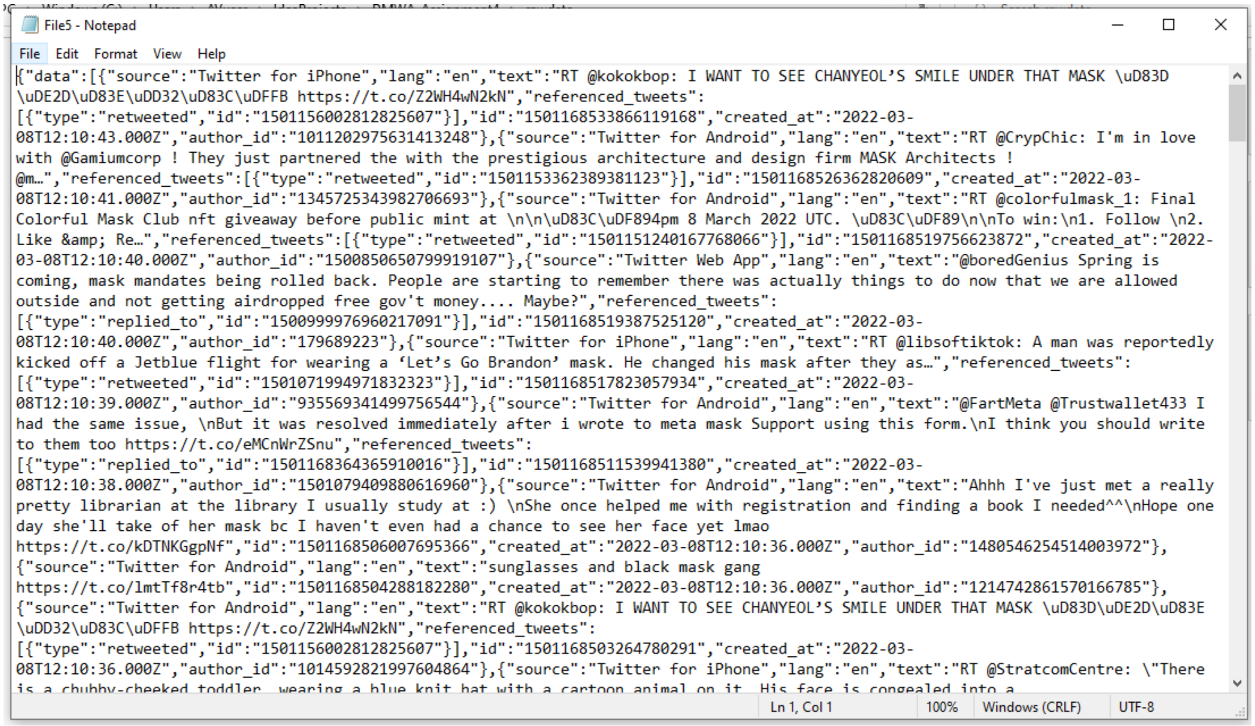
Step 4:

In order to explain how I wrote the program, I am creating a flowchart and adding it to a folder in my drive. I am sharing the link of that drive in this pdf and the reason I couldn’t add the flowchart here is that the image is long and it won’t fit in a single page.

Drive link for flowchart:
https://drive.google.com/drive/folders/1RnxczW2VeqWuyzAlfIV3kX0UX3o-mGIN?usp=sharing

Step 5:

I have extracted around 4200 tweets and metadata related to the given keywords by using the java program I wrote.



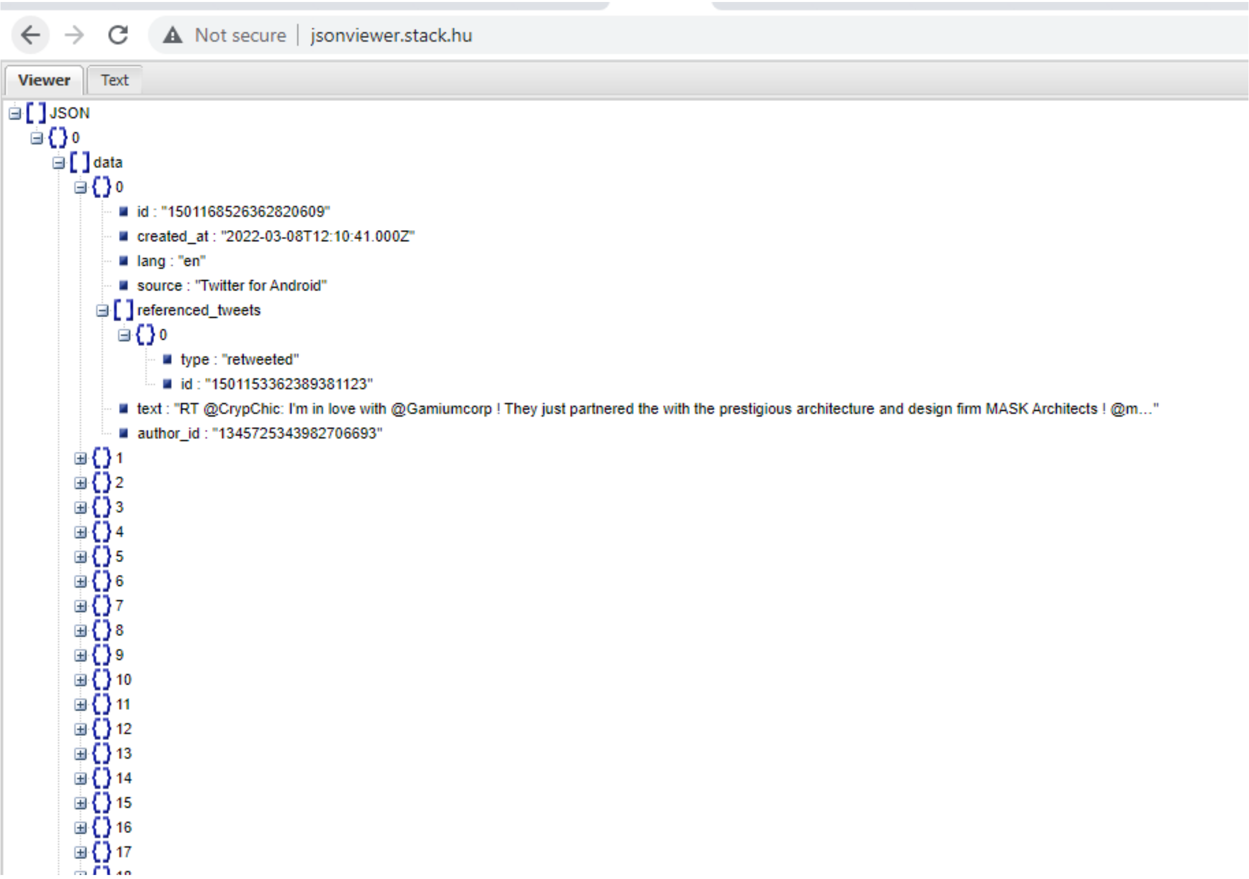
```
{"data":[{"source":"Twitter for iPhone","lang":"en","text":"RT @kokokbop: I WANT TO SEE CHANYEOL’S SMILE UNDER THAT MASK \uD83D
\uDE2D\uD83E\uDD32\uD83C\uDFFB https://t.co/Z2WH4wN2kN","referenced_tweets":
[{"type":"retweeted","id":"1501156002812825607"}],"id":"1501168533866119168","created_at":"2022-03-
08T12:10:43.000Z","author_id":"1011202975631413248"},{"source":"Twitter for Android","lang":"en","text":"RT @CrypChic: I'm in love
with @Gamiumcorp ! They just partnered the with the prestigious architecture and design firm MASK Architects !
@m…","referenced_tweets":[{"type":"retweeted","id":"1501153362389381123"}],"id":"1501168526362820609","created_at":"2022-03-
08T12:10:41.000Z","author_id":"1345725343982706693"},{"source":"Twitter for Android","lang":"en","text":"RT @colorfulmask_1: Final
Colorful Mask Club nft giveaway before public mint at \n\n\uD83C\uDF894pm 8 March 2022 UTC. \uD83C\uDF89\n\nTo win:\n1. Follow \n2.
Like &amp; Re…","referenced_tweets":[{"type":"retweeted","id":"1501151240167768066"}],"id":"1501168519756623872","created_at":"2022-
03-08T12:10:40.000Z","author_id":"1500850650799919107"},{"source":"Twitter Web App","lang":"en","text":"@boredGenius Spring is
coming, mask mandates being rolled back. People are starting to remember there was actually things to do now that we are allowed
outside and not getting airdropped free gov't money.... Maybe?","referenced_tweets":
[{"type":"replied_to","id":"1500999976960217091"}],"id":"1501168519387525120","created_at":"2022-03-
08T12:10:40.000Z","author_id":"179689223"},{"source":"Twitter for iPhone","lang":"en","text":"RT @libsoftiktok: A man was reportedly
kicked off a Jetblue flight for wearing a ‘Let’s Go Brandon’ mask. He changed his mask after they as…","referenced_tweets":
[{"type":"retweeted","id":"1501071994971832323"}],"id":"1501168517823057934","created_at":"2022-03-
08T12:10:39.000Z","author_id":"935569341499756544"},{"source":"Twitter for Android","lang":"en","text":"@FartMeta @Trustwallet433 I
had the same issue, \nBut it was resolved immediately after i wrote to meta mask Support using this form.\nI think you should write
to them too https://t.co/eMCnWrZSnu","referenced_tweets":
[{"type":"replied_to","id":"1501168364365910016"}],"id":"1501168511539941380","created_at":"2022-03-
08T12:10:38.000Z","author_id":"1501079409880616960"},{"source":"Twitter for Android","lang":"en","text":"Ahhh I've just met a really
pretty librarian at the library I usually study at :) \nShe once helped me with registration and finding a book I needed^^\nHope one
day she'll take of her mask bc I haven't even had a chance to see her face yet lmao
https://t.co/kDTNKGgpNf","id":"1501168506007695366","created_at":"2022-03-08T12:10:36.000Z","author_id":"1480546254514003972"},
{"source":"Twitter for Android","lang":"en","text":"sunglasses and black mask gang
https://t.co/lmtTf8r4tb","id":"1501168504288182280","created_at":"2022-03-08T12:10:36.000Z","author_id":"1214742861570166785"},
{"source":"Twitter for Android","lang":"en","text":"RT @kokokbop: I WANT TO SEE CHANYEOL’S SMILE UNDER THAT MASK \uD83D\uDE2D\uD83E
\uDD32\uD83C\uDFFB https://t.co/Z2WH4wN2kN","referenced_tweets":
[{"type":"retweeted","id":"1501156002812825607"}],"id":"1501168503264780291","created_at":"2022-03-
08T12:10:36.000Z","author_id":"1014592821997604864"},{"source":"Twitter for iPhone","lang":"en","text":"RT @StratcomCentre: \"There
is a chubby-cheeked toddler, wearing a blue knit hat with a cartoon animal on it. His face is congealed into a
```

Step 6:

I have enough data of 4200 tweets. So, there is no need for this step in my process.

Step 7:

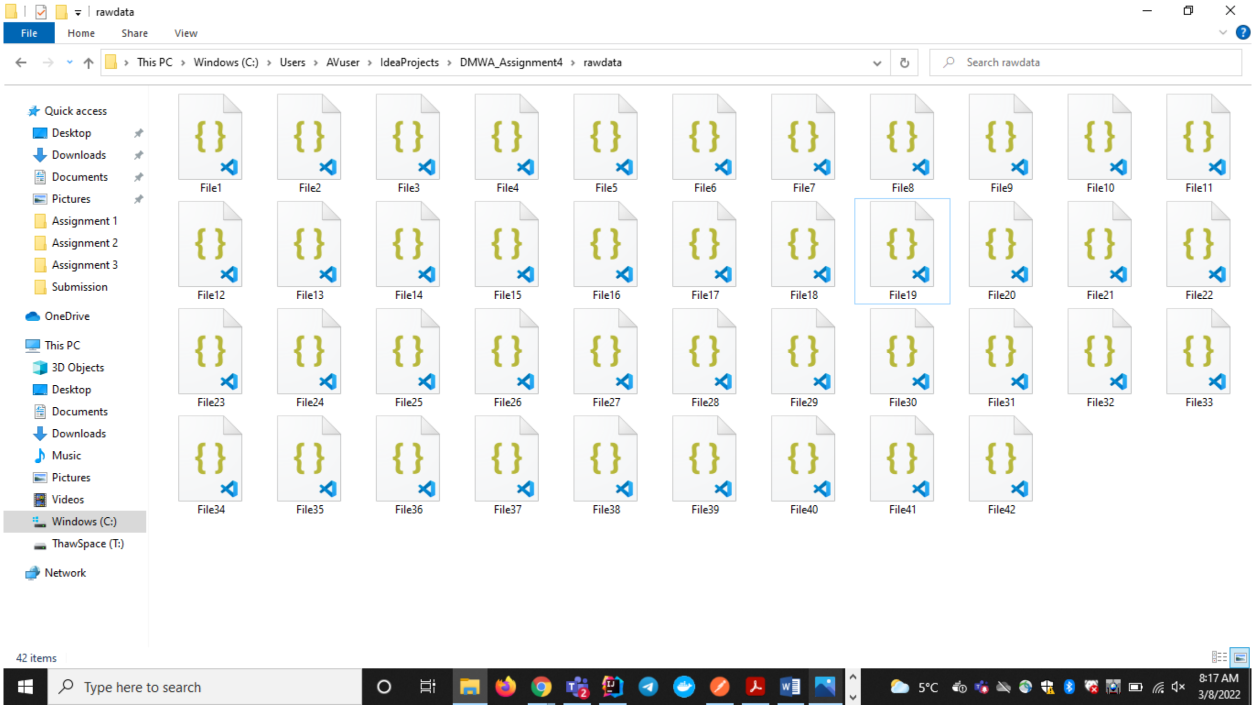
Example of the tweets and metadata extracted using the java program is given below in the screenshot.



```
[ ] JSON
  { } 0
    [ ] data
      { } 0
          id : "1501168526362820609"
          created_at : "2022-03-08T12:10:41.000Z"
          lang : "en"
          source : "Twitter for Android"
        [ ] referenced_tweets
          { } 0
              type : "retweeted"
              id : "1501153362389381123"
          text : "RT @CrypChic: I'm in love with @Gamiumcorp ! They just partnered the with the prestigious architecture and design firm MASK Architects ! @m..."
          author_id : "1345725343982706693"
      { } 1
      { } 2
      { } 3
      { } 4
      { } 5
      { } 6
      { } 7
      { } 8
      { } 9
      { } 10
      { } 11
      { } 12
      { } 13
      { } 14
      { } 15
      { } 16
      { } 17
```

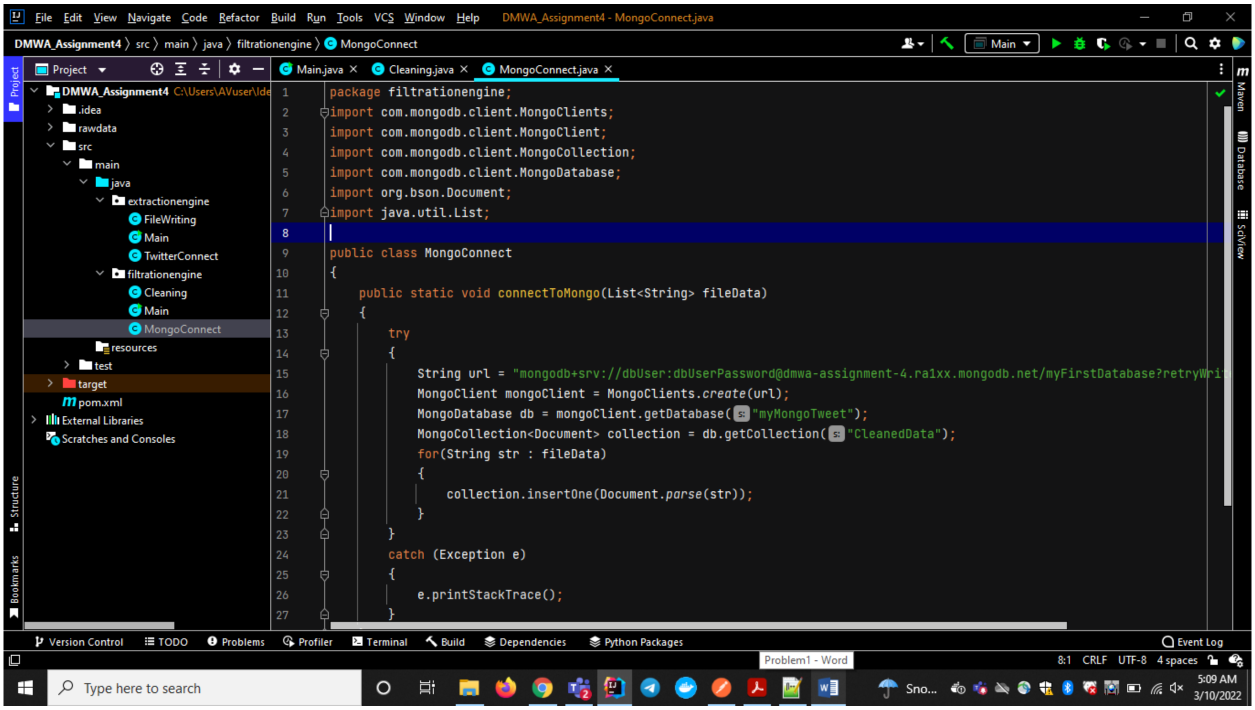
Step 8:

The captured raw data was programmatically kept in files and each file doesn't contain more than 100 tweets. So, there are a total of 42 files for 4200 tweets.

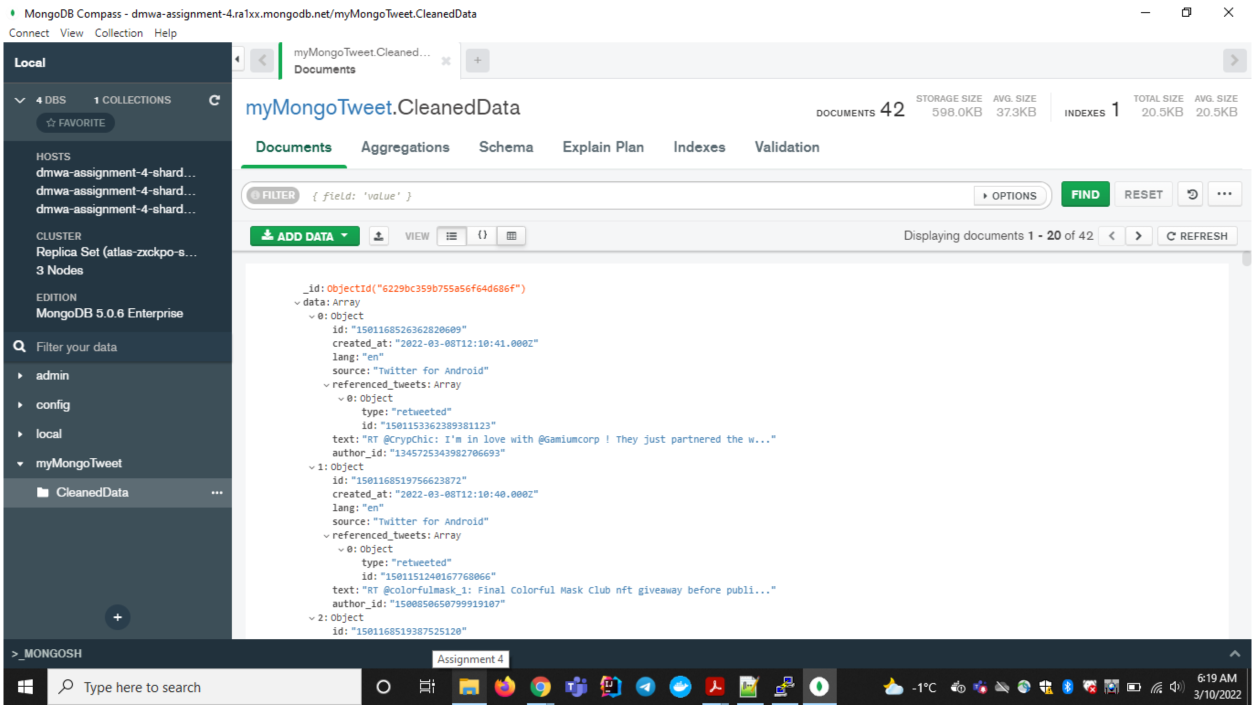


Step 9:

Git URL for java code to clean the data: https://git.cs.dal.ca/rguturu/csci-5408-w2022-b00871849-gutururamamohanvishnu/-/tree/main/assignment-4



```java
package filtrationengine;
import com.mongodb.client.MongoClients;
import com.mongodb.client.MongoClient;
import com.mongodb.client.MongoCollection;
import com.mongodb.client.MongoDatabase;
import org.bson.Document;
import java.util.List;

public class MongoConnect
{
    public static void connectToMongo(List<String> fileData)
    {
        try
        {
            String url = "mongodb+srv://dbUser:dbUserPassword@dmwa-assignment-4.ra1xx.mongodb.net/myFirstDatabase?retryWri
            MongoClient mongoClient = MongoClients.create(url);
            MongoDatabase db = mongoClient.getDatabase("myMongoTweet");
            MongoCollection<Document> collection = db.getCollection("CleanedData");
            for(String str : fileData)
            {
                collection.insertOne(Document.parse(str));
            }
        }
        catch (Exception e)
        {
            e.printStackTrace();
        }
```



```
_id: ObjectId("6229bc359b755a56f64d686f")
data: Array
  0: Object
    id: "1501168526362820609"
    created_at: "2022-03-08T12:10:41.000Z"
    lang: "en"
    source: "Twitter for Android"
    referenced_tweets: Array
      0: Object
        type: "retweeted"
        id: "1501153362389381123"
    text: "RT @CrypChic: I'm in love with @Gamiumcorp ! They just partnered the w..."
    author_id: "1345725343982706693"
  1: Object
    id: "1501168519756623872"
    created_at: "2022-03-08T12:10:40.000Z"
    lang: "en"
    source: "Twitter for Android"
    referenced_tweets: Array
      0: Object
        type: "retweeted"
        id: "1501151240167768066"
    text: "RT @colorfulmask_1: Final Colorful Mask Club nft giveaway before publi..."
    author_id: "1500850650799919107"
  2: Object
    id: "1501168519387525120"
```

Step 10:

In order to explain how I wrote the program, I am creating a flowchart and adding it to a folder in my drive. I am sharing the link of that drive in this pdf and the reason I couldn’t add the flowchart here is that the image is long and it won’t fit in a single page.

Drive link for flowchart:
https://drive.google.com/drive/folders/1RnxczW2VeqWuyzAlfIV3kX0UX3o-mGIN?usp=sharing

References:

[1] "*Search Tweets introduction*," Twitter [Online]. Available at: https://developer.twitter.com/en/docs/twitter-api/tweets/search/introduction [Accessed: March 2, 2022].

[2] "*Filtered stream introduction*," Twitter [Online]. Available at: https://developer.twitter.com/en/docs/twitter-api/tweets/filtered-stream/introduction [Accessed: March 2, 2022].

[3] "*Extracting tweets of a specific hashtag using twitter4j*," Stack Overflow [Online]. Available at: https://stackoverflow.com/questions/23341215/extracting-tweets-of-a-specific-hashtag-using-twitter4j [Accessed: March 3, 2022].

[4] "*Java create multiple new files*," Stack Overflow [Online]. Available at: https://stackoverflow.com/questions/54462609/java-create-multiple-new-files/54463212 [Accessed: March 8, 2022].

[5] "*Java Create and Write To Files*," W3schools.com [Online]. Available at: https://www.w3schools.com/java/java_files_create.asp [Accessed: March 8, 2022].

[6] "*Standard search API*," Developer.twitter.com [Online]. Available at: https://developer.twitter.com/en/docs/twitter-api/v1/tweets/search/api-reference/get-search-tweets [Accessed: March 9, 2022].

[7] "*GitHub - twitterdev/Twitter-API-v2-sample-code: Sample code for the Twitter API v2 endpoints*," GitHub [Online]. Available at:

https://github.com/twitterdev/Twitter-API-v2-sample-code [Accessed: March 10, 2022].

[8] "*Java HTTP Client - Examples and Recipes*," Openjdk.java.net [Online]. Available at: https://openjdk.java.net/groups/net/httpclient/recipes.html [Accessed: March 10, 2022].

[9] "*Getting Started with MongoDB and Java - CRUD Operations Tutorial*," Mongodb.com [Online]. Available at: https://www.mongodb.com/developer/quickstart/java-setup-crud-operations/?utm_campaign=javainsertingdocuments&utm_source=facebook&utm_medium=organic_social [Accessed: March 10, 2022].

[10] Vega, D., Youtube.com [Online]. Available at: https://www.youtube.com/watch?v=8QBJMxyXIqc&ab_channel=DanVega [Accessed: March 10, 2022].