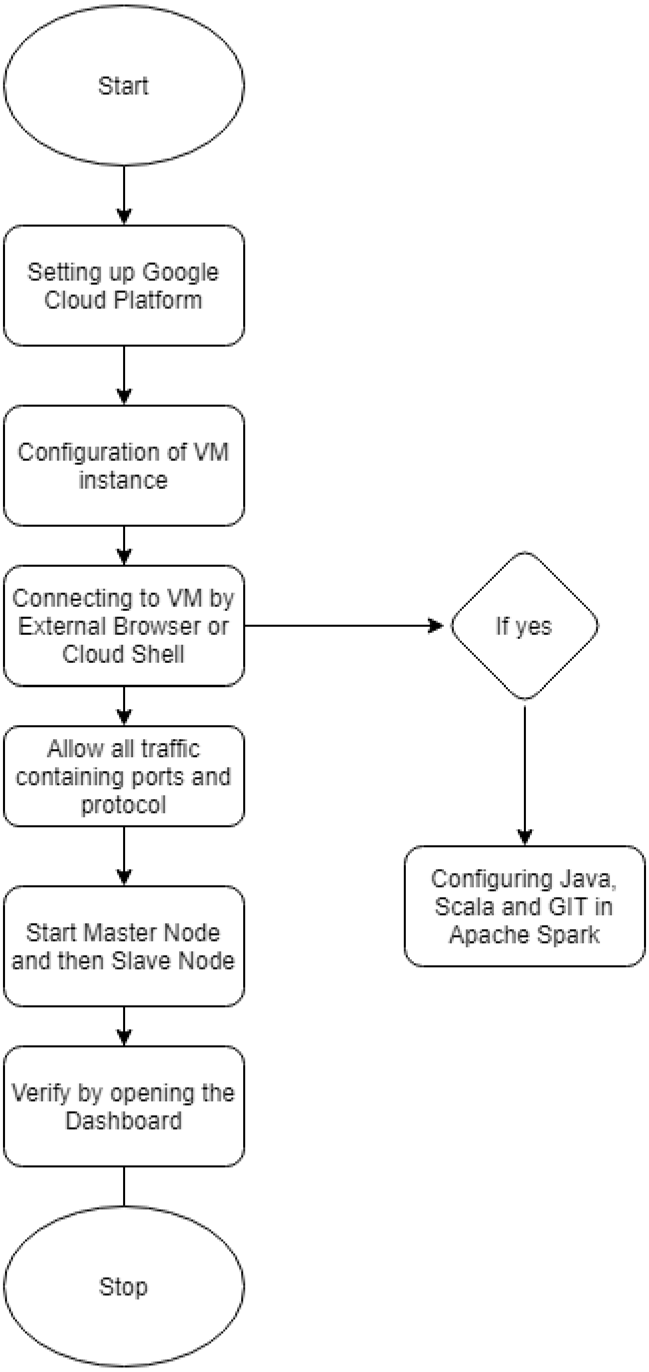
Report: Cluster Setup

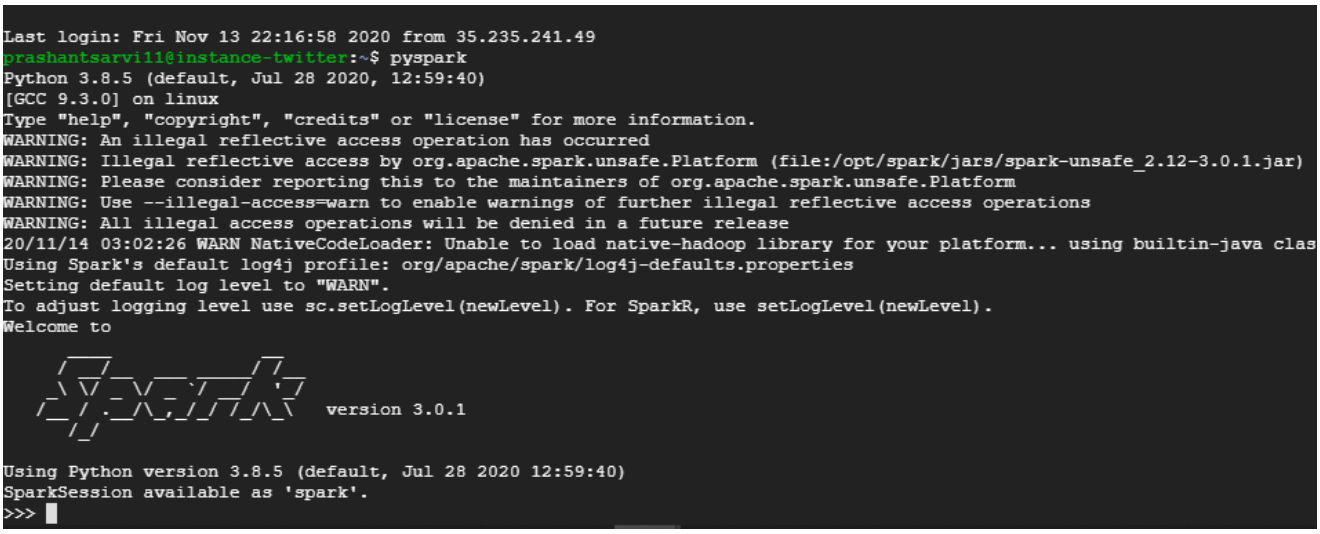
Flowchart:


Start
Setting up Google Cloud Platform
Configuration of VM instance
Connecting to VM by External Browser or Cloud Shell
If yes
Configuring Java, Scala and GIT in Apache Spark
Allow all traffic containing ports and protocol
Start Master Node and then Slave Node
Verify by opening the Dashboard
Stop

Source: https:// app.diagram.net

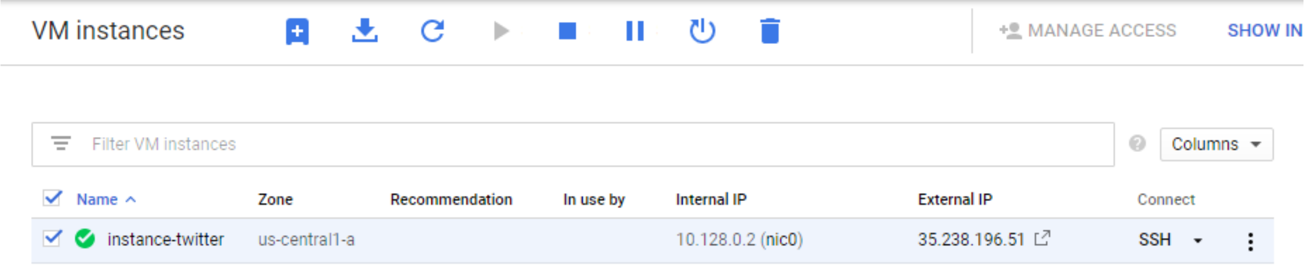
Screenshots:

Configured & Initialized Apache Spark Cluster:



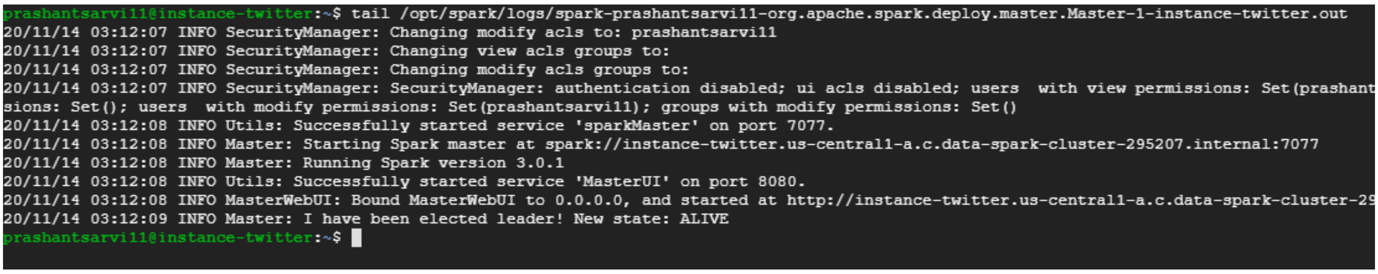
```
Last login: Fri Nov 13 22:16:58 2020 from 35.235.241.49
prashantsarvi11@instance-twitter:~$ pyspark
Python 3.8.5 (default, Jul 28 2020, 12:59:40)
[GCC 9.3.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.spark.unsafe.Platform (file:/opt/spark/jars/spark-unsafe_2.12-3.0.1.jar)
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
20/11/14 03:02:26 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java clas
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 3.0.1
      /_/

Using Python version 3.8.5 (default, Jul 28 2020 12:59:40)
SparkSession available as 'spark'.
>>> 
```



Name	Zone	Recommendation	In use by	Internal IP	External IP	Connect
instance-twitter	us-central1-a			10.128.0.2 (nic0)	35.238.196.51	SSH

Started Master Node:



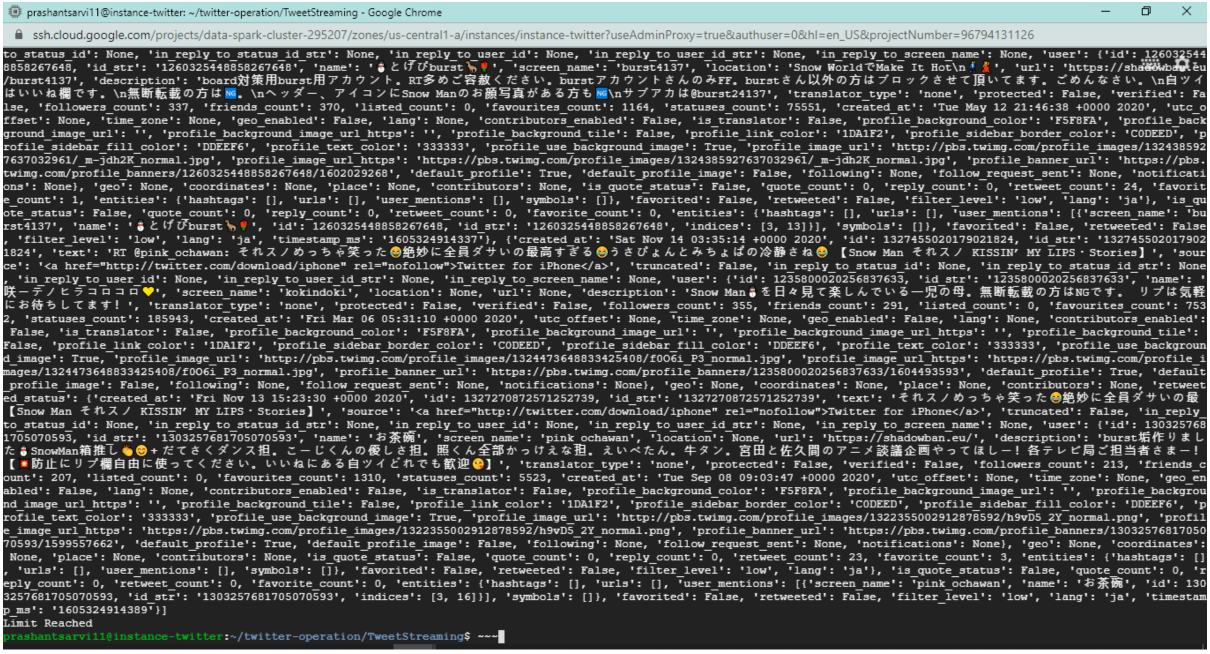
```
prashantsarvi11@instance-twitter:~$ tail /opt/spark/logs/spark-prashantsarvi11-org.apache.spark.deploy.master.Master-1-instance-twitter.out
20/11/14 03:12:07 INFO SecurityManager: Changing modify acls to: prashantsarvi11
20/11/14 03:12:07 INFO SecurityManager: Changing view acls groups to:
20/11/14 03:12:07 INFO SecurityManager: Changing modify acls groups to:
20/11/14 03:12:07 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  with view permissions: Set(prashant
sions: Set(); users  with modify permissions: Set(prashantsarvi11); groups with modify permissions: Set()
20/11/14 03:12:08 INFO Utils: Successfully started service 'sparkMaster' on port 7077.
20/11/14 03:12:08 INFO Master: Starting Spark master at spark://instance-twitter.us-central1-a.c.data-spark-cluster-295207.internal:7077
20/11/14 03:12:08 INFO Master: Running Spark version 3.0.1
20/11/14 03:12:08 INFO Utils: Successfully started service 'MasterUI' on port 8080.
20/11/14 03:12:08 INFO MasterWebUI: Bound MasterWebUI to 0.0.0.0, and started at http://instance-twitter.us-central1-a.c.data-spark-cluster-29
20/11/14 03:12:09 INFO Master: I have been elected leader! New state: ALIVE
prashantsarvi11@instance-twitter:~$ 
```

Report: Data Extraction

Algorithm:

1. Importing and installing pymongo, dnspython and tweepy.
2. Tweetcounter variable declared and initialised.
3. Loading the live tweets with the help of on_data method.
4. Appending all the incoming tweets inside a list and setting a condition after a certain number of tweets is reached.
5. On_error keeps track and handles the exception of the internal call.
6. In the main method connection to the MongoDB is set.
7. The database and the tables containing the tweets with particular attributes are inserted at the final step inside the DB.

Ouput when run on the VM instance:



prashantsarvi11@instance-twitter: ~/twitter-operation/TweetStreaming - Google Chrome
ssh.cloud.google.com/projects/data-spark-cluster-295207/zones/us-central1-a/instances/instance-twitter?useAdminProxy=true&authuser=0&hl=en_US&projectNumber=96794131126

```
to_status_id': None, 'in_reply_to_status_id_str': None, 'in_reply_to_user_id': None, 'in_reply_to_user_id_str': None, 'in_reply_to_screen_name': None, 'user': {'id': 126032544
8858267648, 'id_str': '1260325448858267648', 'name': '⛄とげびburst🦙🌹', 'screen_name': 'burst4137', 'location': 'Snow WorldでMake It Hot\n🕺💃', 'url': 'https://shadowban.eu
/burst4137', 'description': 'board対策用burst用アカウント。RT多めご容赦ください。burstアカウントさんのみFF。burstさん以外の方はブロックさせて頂いてます。ごめんなさい。\n自ツイ
はいいね欄です。\n無断転載の方は🆖。\nヘッダー、アイコンにSnow Manのお顔写真がある方も🆖\nサブアカは@burst24137', 'translator_type': 'none', 'protected': False, 'verified': Fa
lse, 'followers_count': 337, 'friends_count': 370, 'listed_count': 0, 'favourites_count': 1164, 'statuses_count': 75551, 'created_at': 'Tue May 12 21:46:38 +0000 2020', 'utc_o
ffset': None, 'time_zone': None, 'geo_enabled': False, 'lang': None, 'contributors_enabled': False, 'is_translator': False, 'profile_background_color': 'F5F8FA', 'profile_back
ground_image_url': '', 'profile_background_image_url_https': '', 'profile_background_tile': False, 'profile_link_color': '1DA1F2', 'profile_sidebar_border_color': 'C0DEED', 'p
rofile_sidebar_fill_color': 'DDEEF6', 'profile_text_color': '333333', 'profile_use_background_image': True, 'profile_image_url': 'http://pbs.twimg.com/profile_images/132438592
7637032961/_m-jdh2K_normal.jpg', 'profile_image_url_https': 'https://pbs.twimg.com/profile_images/1324385927637032961/_m-jdh2K_normal.jpg', 'profile_banner_url': 'https://pbs.
twimg.com/profile_banners/1260325448858267648/1602029268', 'default_profile': True, 'default_profile_image': False, 'following': None, 'follow_request_sent': None, 'notificati
ons': None}, 'geo': None, 'coordinates': None, 'place': None, 'contributors': None, 'is_quote_status': False, 'quote_count': 0, 'reply_count': 0, 'retweet_count': 24, 'favorit
e_count': 1, 'entities': {'hashtags': [], 'urls': [], 'user_mentions': [], 'symbols': []}, 'favorited': False, 'retweeted': False, 'filter_level': 'low', 'lang': 'ja'}, 'is_qu
ote_status': False, 'quote_count': 0, 'reply_count': 0, 'retweet_count': 0, 'favorite_count': 0, 'entities': {'hashtags': [], 'urls': [], 'user_mentions': [{'screen_name': 'bu
rst4137', 'name': '⛄とげびburst🦙🌹', 'id': 1260325448858267648, 'id_str': '1260325448858267648', 'indices': [3, 13]}], 'symbols': []}, 'favorited': False, 'retweeted': False
, 'filter_level': 'low', 'lang': 'ja', 'timestamp_ms': '1605324914337'}, {'created_at': 'Sat Nov 14 03:35:14 +0000 2020', 'id': 1327455020179021824, 'id_str': '132745502017902
1824', 'text': 'RT @pink_ochawan: それスノめっちゃ笑った😂絶妙に全員ダサいの最高すぎる😂うさぴょんとみちょぱの冷静さね😂 【Snow Man それスノ KISSIN' MY LIPS・Stories】', 'sour
ce': '<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>', 'truncated': False, 'in_reply_to_status_id': None, 'in_reply_to_status_id_str': None
, 'in_reply_to_user_id': None, 'in_reply_to_user_id_str': None, 'in_reply_to_screen_name': None, 'user': {'id': 1235800020256837633, 'id_str': '1235800020256837633', 'name': '
咲ーテノヒラコロコロ💛', 'screen_name': 'kokindoki', 'location': None, 'url': None, 'description': 'Snow Man⛄を日々見て楽しんでいる一児の母。無断転載の方はNGです。 リプは気軽
にお待ちしてます!', 'translator_type': 'none', 'protected': False, 'verified': False, 'followers_count': 355, 'friends_count': 291, 'listed_count': 0, 'favourites_count': 753
2, 'statuses_count': 185943, 'created_at': 'Fri Mar 06 05:31:10 +0000 2020', 'utc_offset': None, 'time_zone': None, 'geo_enabled': False, 'lang': None, 'contributors_enabled':
 False, 'is_translator': False, 'profile_background_color': 'F5F8FA', 'profile_background_image_url': '', 'profile_background_image_url_https': '', 'profile_background_tile':
False, 'profile_link_color': '1DA1F2', 'profile_sidebar_border_color': 'C0DEED', 'profile_sidebar_fill_color': 'DDEEF6', 'profile_text_color': '333333', 'profile_use_backgroun
d_image': True, 'profile_image_url': 'http://pbs.twimg.com/profile_images/1324473648833425408/f0O6i_P3_normal.jpg', 'profile_image_url_https': 'https://pbs.twimg.com/profile_i
mages/1324473648833425408/f0O6i_P3_normal.jpg', 'profile_banner_url': 'https://pbs.twimg.com/profile_banners/1235800020256837633/1604493593', 'default_profile': True, 'default
_profile_image': False, 'following': None, 'follow_request_sent': None, 'notifications': None}, 'geo': None, 'coordinates': None, 'place': None, 'contributors': None, 'retweet
ed_status': {'created_at': 'Fri Nov 13 15:23:30 +0000 2020', 'id': 1327270872571252739, 'id_str': '1327270872571252739', 'text': 'それスノめっちゃ笑った😂絶妙に全員ダサいの最
【Snow Man それスノ KISSIN' MY LIPS・Stories】', 'source': '<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>', 'truncated': False, 'in_reply_
to_status_id': None, 'in_reply_to_status_id_str': None, 'in_reply_to_user_id': None, 'in_reply_to_user_id_str': None, 'in_reply_to_screen_name': None, 'user': {'id': 130325768
1705070593, 'id_str': '1303257681705070593', 'name': 'お茶碗', 'screen_name': 'pink_ochawan', 'location': None, 'url': 'https://shadowban.eu/', 'description': 'burst垢作りまし
た⛄SnowMan箱推し👏😊+だてさくダンス担。こーじくんの優しさ担。照くん全部かっけえな担。えいぺたん。牛タン。宮田と佐久間のアニメ談議企画やってほしー!各テレビ局ご担当者さまー!
【💥防止にリプ欄自由に使ってください。いいねにある自ツイどれでも歓迎😘】', 'translator_type': 'none', 'protected': False, 'verified': False, 'followers_count': 213, 'friends_c
ount': 207, 'listed_count': 0, 'favourites_count': 1310, 'statuses_count': 5523, 'created_at': 'Tue Sep 08 09:03:47 +0000 2020', 'utc_offset': None, 'time_zone': None, 'geo_en
abled': False, 'lang': None, 'contributors_enabled': False, 'is_translator': False, 'profile_background_color': 'F5F8FA', 'profile_background_image_url': '', 'profile_backgrou
nd_image_url_https': '', 'profile_background_tile': False, 'profile_link_color': '1DA1F2', 'profile_sidebar_border_color': 'C0DEED', 'profile_sidebar_fill_color': 'DDEEF6', 'p
rofile_text_color': '333333', 'profile_use_background_image': True, 'profile_image_url': 'http://pbs.twimg.com/profile_images/1322355002912878592/h9vD5_2Y_normal.png', 'profil
e_image_url_https': 'https://pbs.twimg.com/profile_images/1322355002912878592/h9vD5_2Y_normal.png', 'profile_banner_url': 'https://pbs.twimg.com/profile_banners/13032576817050
70593/1599557662', 'default_profile': True, 'default_profile_image': False, 'following': None, 'follow_request_sent': None, 'notifications': None}, 'geo': None, 'coordinates':
 None, 'place': None, 'contributors': None, 'is_quote_status': False, 'quote_count': 0, 'reply_count': 0, 'retweet_count': 23, 'favorite_count': 3, 'entities': {'hashtags': []
, 'urls': [], 'user_mentions': [], 'symbols': []}, 'favorited': False, 'retweeted': False, 'filter_level': 'low', 'lang': 'ja'}, 'is_quote_status': False, 'quote_count': 0, 'r
eply_count': 0, 'retweet_count': 0, 'favorite_count': 0, 'entities': {'hashtags': [], 'urls': [], 'user_mentions': [{'screen_name': 'pink_ochawan', 'name': 'お茶碗', 'id': 130
3257681705070593, 'id_str': '1303257681705070593', 'indices': [3, 16]}], 'symbols': []}, 'favorited': False, 'retweeted': False, 'filter_level': 'low', 'lang': 'ja', 'timestam
p_ms': '1605324914389'}]
```

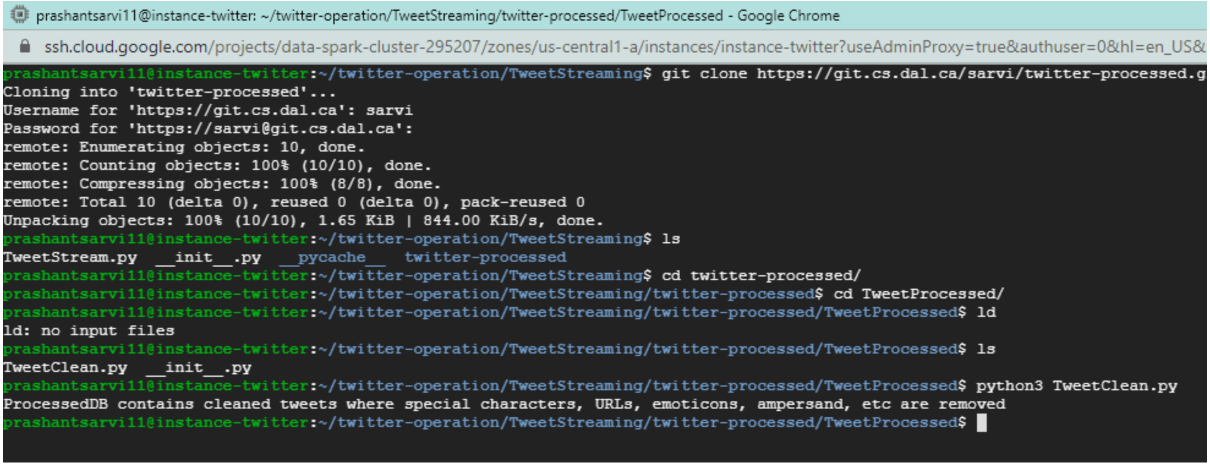
```
Limit Reached
prashantsarvi11@instance-twitter:~/twitter-operation/TweetStreaming$ ~~~
```

Report: Pre-processing Engine

Algorithm:

1. Establishing connection to the MongoDB client
2. Storing RawDB, its collections and processedDB and its collection inside variables
3. Iterating through the collection of tables from the RawDB and passing each document of table in the JSON format through RegEx or Regular expression.
4. The regular expression removes the URLs, special symbols and white spaces.
5. The cleaned JSONs are appended to a new list which is finally inserted in ProcessedDB

Ouput when run on the VM instance:



prashantsarvi11@instance-twitter: ~/twitter-operation/TweetStreaming/twitter-processed/TweetProcessed - Google Chrome
ssh.cloud.google.com/projects/data-spark-cluster-295207/zones/us-central1-a/instances/instance-twitter?useAdminProxy=true&authuser=0&hl=en_US&

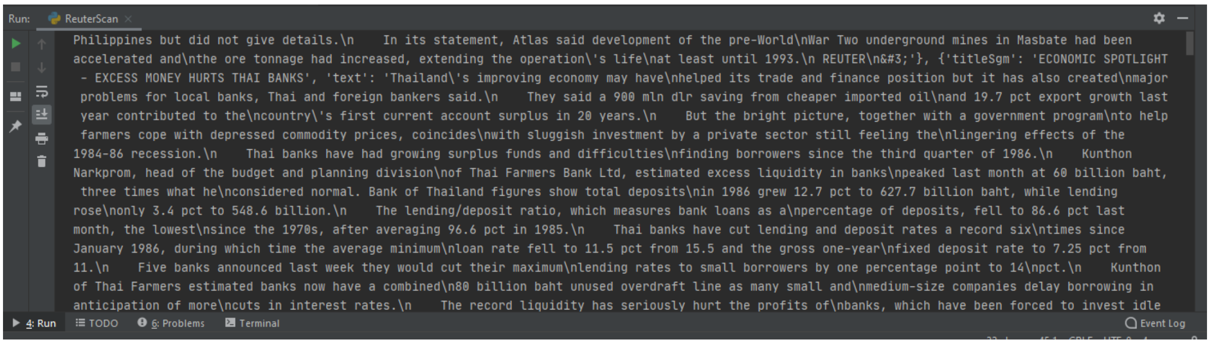
```
prashantsarvi11@instance-twitter:~/twitter-operation/TweetStreaming$ git clone https://git.cs.dal.ca/sarvi/twitter-processed.g
Cloning into 'twitter-processed'...
Username for 'https://git.cs.dal.ca': sarvi
Password for 'https://sarvi@git.cs.dal.ca':
remote: Enumerating objects: 10, done.
remote: Counting objects: 100% (10/10), done.
remote: Compressing objects: 100% (8/8), done.
remote: Total 10 (delta 0), reused 0 (delta 0), pack-reused 0
Unpacking objects: 100% (10/10), 1.65 KiB | 844.00 KiB/s, done.
prashantsarvi11@instance-twitter:~/twitter-operation/TweetStreaming$ ls
TweetStream.py  __init__.py  __pycache__  twitter-processed
prashantsarvi11@instance-twitter:~/twitter-operation/TweetStreaming$ cd twitter-processed/
prashantsarvi11@instance-twitter:~/twitter-operation/TweetStreaming/twitter-processed$ cd TweetProcessed/
prashantsarvi11@instance-twitter:~/twitter-operation/TweetStreaming/twitter-processed/TweetProcessed$ ld
ld: no input files
prashantsarvi11@instance-twitter:~/twitter-operation/TweetStreaming/twitter-processed/TweetProcessed$ ls
TweetClean.py  __init__.py
prashantsarvi11@instance-twitter:~/twitter-operation/TweetStreaming/twitter-processed/TweetProcessed$ python3 TweetClean.py
ProcessedDB contains cleaned tweets where special characters, URLs, emoticons, ampersand, etc are removed
prashantsarvi11@instance-twitter:~/twitter-operation/TweetStreaming/twitter-processed/TweetProcessed$ 
```

Report: Reuter News Articles

Algorithm:

1. Connection to be established to the MongoDb by creating ReuterDb.
2. The .sgm files are read and stored in different variables
3. The findall method is used to find the specific tags <REUTERS></REUTERS>, <TEXT></TEXT> and <TITLE></TITLE>.
4. The scanned texts between the tags are read appended in the list.
5. The final list is then inserted into the ReuterDb with the cleaned data.

ScreenShot:



```
Philippines but did not give details.\n    In its statement, Atlas said development of the pre-World\nWar Two underground mines in Masbate had been
accelerated and\nthe ore tonnage had increased, extending the operation\'s life\nat least until 1993.\n REUTER\n&#3;'}, {'titleSgm': 'ECONOMIC SPOTLIGHT
 - EXCESS MONEY HURTS THAI BANKS', 'text': 'Thailand\'s improving economy may have\nhelped its trade and finance position but it has also created\nmajor
 problems for local banks, Thai and foreign bankers said.\n    They said a 900 mln dlr saving from cheaper imported oil\nand 19.7 pct export growth last
 year contributed to the\ncountry\'s first current account surplus in 20 years.\n    But the bright picture, together with a government program\nto help
 farmers cope with depressed commodity prices, coincides\nwith sluggish investment by a private sector still feeling the\nlingering effects of the
1984-86 recession.\n    Thai banks have had growing surplus funds and difficulties\nfinding borrowers since the third quarter of 1986.\n    Kunthon
Narkprom, head of the budget and planning division\nof Thai Farmers Bank Ltd, estimated excess liquidity in banks\npeaked last month at 60 billion baht,
 three times what he\nconsidered normal. Bank of Thailand figures show total deposits\nin 1986 grew 12.7 pct to 627.7 billion baht, while lending
rose\nonly 3.4 pct to 548.6 billion.\n    The lending/deposit ratio, which measures bank loans as a\npercentage of deposits, fell to 86.6 pct last
month, the lowest\nsince the 1970s, after averaging 96.6 pct in 1985.\n    Thai banks have cut lending and deposit rates a record six\ntimes since
January 1986, during which time the average minimum\nloan rate fell to 11.5 pct from 15.5 and the gross one-year\nfixed deposit rate to 7.25 pct from
11.\n    Five banks announced last week they would cut their maximum\nlending rates to small borrowers by one percentage point to 14\npct.\n    Kunthon
of Thai Farmers estimated banks now have a combined\n80 billion baht unused overdraft line as many small and\nmedium-size companies delay borrowing in
anticipation of more\ncuts in interest rates.\n    The record liquidity has seriously hurt the profits of\nbanks, which have been forced to invest idle
```

Report: Data Visualization using Graph Database

Cypher:

```
CREATE (canada:country {title: 'Canada', continent: 'North America', climate:
'cold'}) - [:HAS_SEASON] -> (winter:weather {title: 'Winter', type: 'extreme',
duration: '5 months'})
```

```
MATCH (canada:country {title: 'Canada', continent: 'North America', climate:
'cold'})

CREATE (canada) - [:GETS_FREQUENTLY_HITBY] -> (storm:Calamity
{title: 'Storm', stormType: 'harsh', damage: 'extreme'})

RETURN canada, storm
```

```
MATCH (winter:weather {title: 'Winter', type: 'extreme', duration: '5 months'})

CREATE(flu:HealthCondition {title: 'Flu', type: 'pandemic', damage:
'medium'}) - [:MORE_CHANCES_TO_AFFECT_IN] -> (winter)

RETURN flu, winter
```

```
MATCH (winter:weather {title: 'Winter', type: 'extreme', duration: '5 months'})

CREATE (winter) - [:BREEZE_IS_VERY] -> (cold:Season {title: 'Cold', type:
'extreme', duration: '8 months'})

RETURN winter, cold
```

```
MATCH (cold:Season {title: 'Cold', type: 'extreme', duration: '8 months'})

CREATE (cold) - [:SO_RECOMMENDED_TO_STAY] ->
(indoor:SafetyMeasure {title: 'Indoor', type: 'Outdoor/Indoor'})
```

```
MATCH(flu:HealthCondition {title: 'Flu', type: 'pandemic', damage: 'medium'})

CREATE (flu) - [:TAKE_PRECAUTIONS] -> (safety:Precaution {title:
'Safety', count: 4, priority: 'high'})
```

```
MATCH (safety:Precaution {title: 'Safety', count: 4, priority: 'high'})
MATCH (indoor:SafetyMeasure {title: 'Indoor', type: 'Outdoor/Indoor'})
CREATE (indoor) - [:A_PRECAUTION_FOR] -> (safety)
```

```
MATCH (storm:Calamity {title: 'Storm', stormType: 'harsh', damage:
'extreme'})
MATCH (safety:Precaution {title: 'Safety', count: 4, priority: 'high'})
CREATE (storm) - [:HIGHEST_PRIORITY_TO_TAKE] -> (safety)
MATCH (n) RETURN n
```

```
MATCH (storm:Calamity {title: 'Storm', stormType: 'harsh', damage:
'extreme'})
CREATE (storm) - [:LEADS_TO_HEAVY] -> (rain:SeasonAffect {title: 'Rain',
type:'heavy', size:'508mm'})
```

```
MATCH (cold:Season {title: 'Cold', type: 'extreme', duration: '8 months'})
CREATE (cold) - [:MAKES_ROAD_TURN_INTO] -> (ice:consequence {title:
'Ice', damage:'medium', geography:'arctic', area: 'fourty-percent'})
```

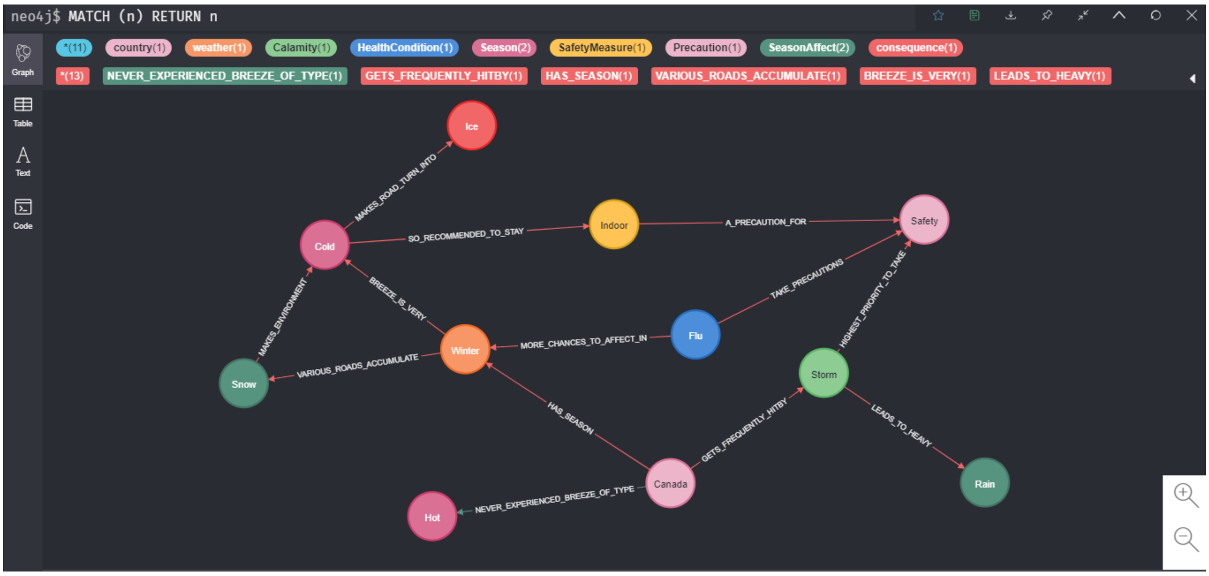
```
MATCH (winter:weather {title: 'Winter', type: 'extreme', duration: '5 months'})
CREATE (winter) - [:VARIOUS_ROADS_ACCUMULATE] ->
(snow:SeasonAffect {title: 'Snow', type:'heavy', size:'400mm'})
```

```
MATCH (snow:SeasonAffect {title: 'Snow', type:'heavy', size:'400mm'})
MATCH (cold:Season {title: 'Cold', type: 'extreme', duration: '8 months'})
CREATE (snow) - [:MAKES_ENVIRONMENT] -> (cold)
```

MATCH (canada:country {title: 'Canada', continent: 'North America', climate: 'cold'})

CREATE (canada) - [:NEVER_EXPERIENCED_BREEZE_OF_TYPE] -> (hot:Season {title: 'Hot', type: 'low', duration: '4 months'})

Graph:



```
neo4j$ MATCH (n) RETURN n
```