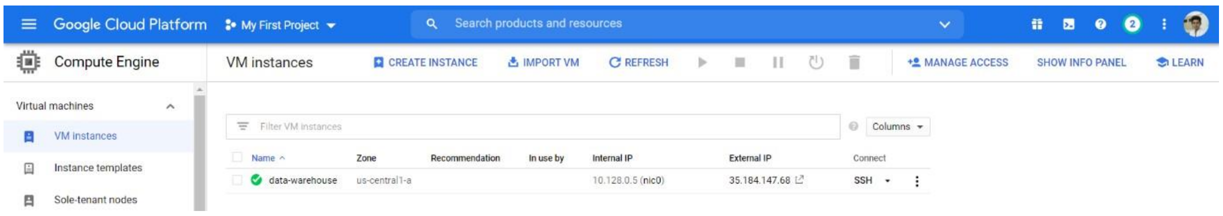
Juhilkumar P. Zalavadiya
B00872280

Problem #1:

Task 1: Cluster Setup - Apache Spark Framework on GCP

1. I opened my google cloud account [1].
2. Searched for VM instance and clicked on Create button.
3. Selected E2-medium and Ubuntu 20.04 LTS as per instructions.
4. Enabled HTTP and HTTPS traffic and permitted to allow all protocols and ports.
5. Opened terminal and installed JDK, scala, and git with the command: sudo apt install default-jdk scala git –y
6. Created a new folder Apache_Spark and installed apache-spark [2], and configured the path.
7. Started the master node with the command: start-master.sh
8. After that, I started the slave node using the command: start-slave.sh <url>
9. Opened <VM adress>:8080 to confirm the working of the cluster.
10. I installed pyspark for future use.

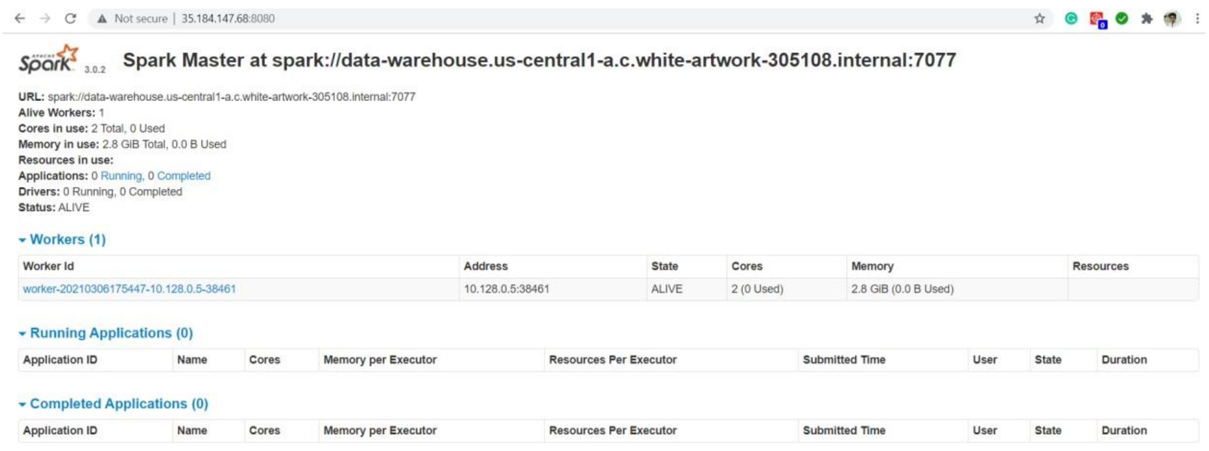
➢ Figure 1 shows the VM instance used for the apache spark cluster.



Name	Zone	Recommendation	In use by	Internal IP	External IP	Connect
data-warehouse	us-central1-a			10.128.0.5 (nic0)	35.184.147.68	SSH

Figure 1 VM instance

➢ Figure 2 shows the spark dashboard with master and slave



Spark Master at spark://data-warehouse.us-central1-a.c.white-artwork-305108.internal:7077

URL: spark://data-warehouse.us-central1-a.c.white-artwork-305108.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-20210306175447-10.128.0.5-38461	10.128.0.5:38461	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

Figure 2 spark dashboard

Juhilkumar P. Zalavadiya
B00872280

Task 2:Data Extraction from Twitter:

- At first, I have created a Twitter developer account
- To extract data from Twitter, I have used tweepy library [4].
- In tweepy, I have used search and stream API for searching previously generated data and live streaming data.

1) Stream API:

- Stream API is helpful in extracting live tweets.
- I have used the MongoClient library to connect with the MongoDB remote database [3].
- Also, I have authenticated the user based on Twitter-generated keys.
- I have used StreamListner class to search real-time tweets of the required keywords.
- In that class, I have given the path for the database and collection.
- I have defined a function on_status() to fetch tweet text with some metadata and inserted it into the database with the help of the insert() function provided by StreamListener class.
- To make 40 separate files for this data, I have used two counters: count and fileName. To change the file, I have used an if condition that checks for the count, and if the count is divisible by 100, it creates a new file and flushes 100 tweets in that file.
- I have used the fileName counter to give different file names in the database.
- One more if the condition will check for the counter. As mentioned in the assignment requirement, I have given an If the condition checks for the count value. If it is 4000, then the program will stop automatically.
- In the case of any error, I have handled it with the on_error() function.
- After running this program on GCP, files will automatically upload to the MongoDb remote database.

2) Search API:

- Search API is helpful in extracting previously generated tweets based on date. You have to specify a date, and the tweepy library will search it with the help of the Cursor() function and give you the tweets since that date.
- I have authenticated the user based on the keys given by Twitter.
- After that, I have connected MongoDB remote database with the help of the pymongo library.
- I have defined one list named “WORDS” to define keywords mentioned in the assignment requirements.
- I have defined a for loop to take each tweet with the [key, value] pair and extracted the text and metadata. After that, with the help of a predefined method insert(), I have inserted those fields into a MongoDB remote database collection.
- To store these data in separate files, I have used two counters: count and fileName. When the count is divisible by 100, this condition will store 100 tweets in one file in the MongoDb database named “myMongoTweet.”
- I have used the fileName counter to give different file names in the database.
- I have defined one more if condition to terminate the program when the count value reaches 4000.

Juhilkumar P. Zalavadiya
B00872280

Output screenshots for twitter data extraction:

1) Before generating the "myMongoTweet" database

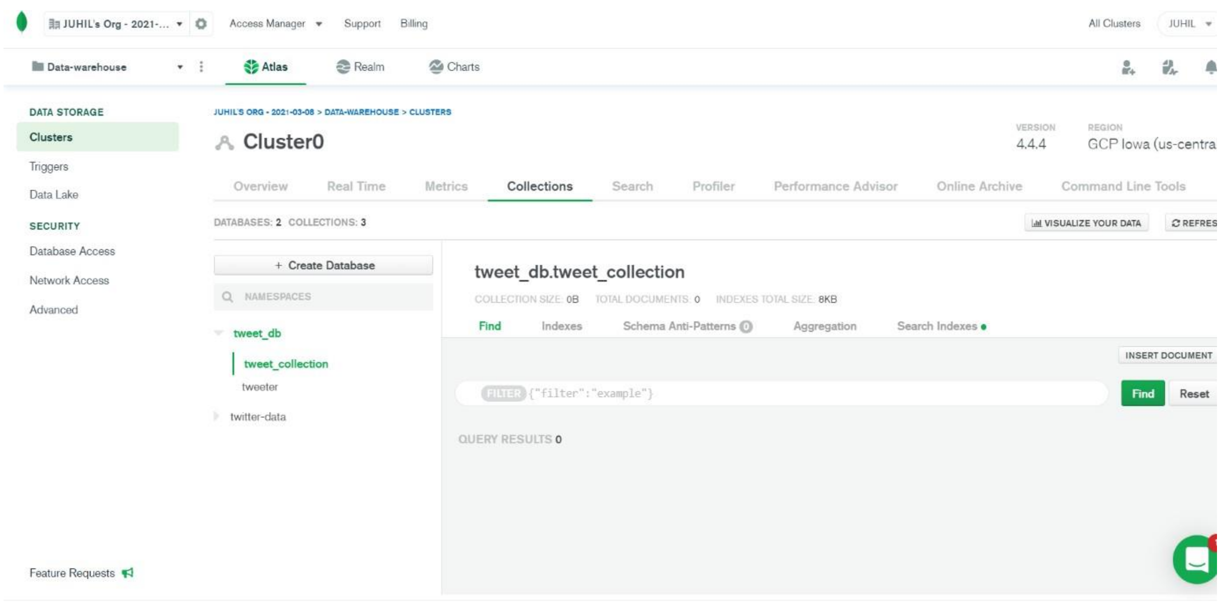
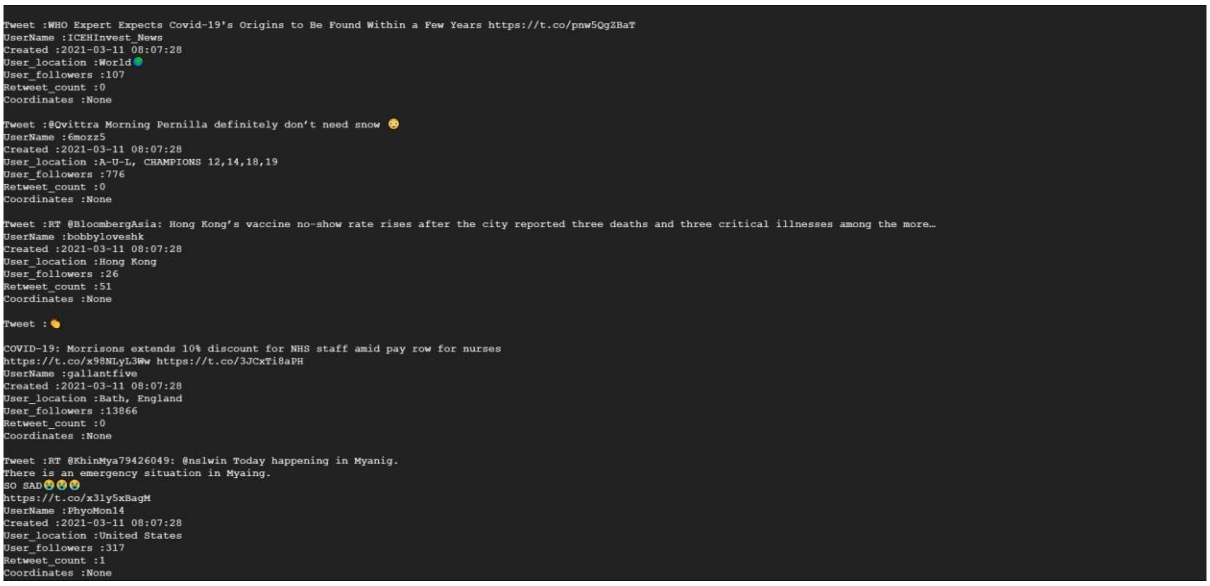


Figure 2. Before running the script

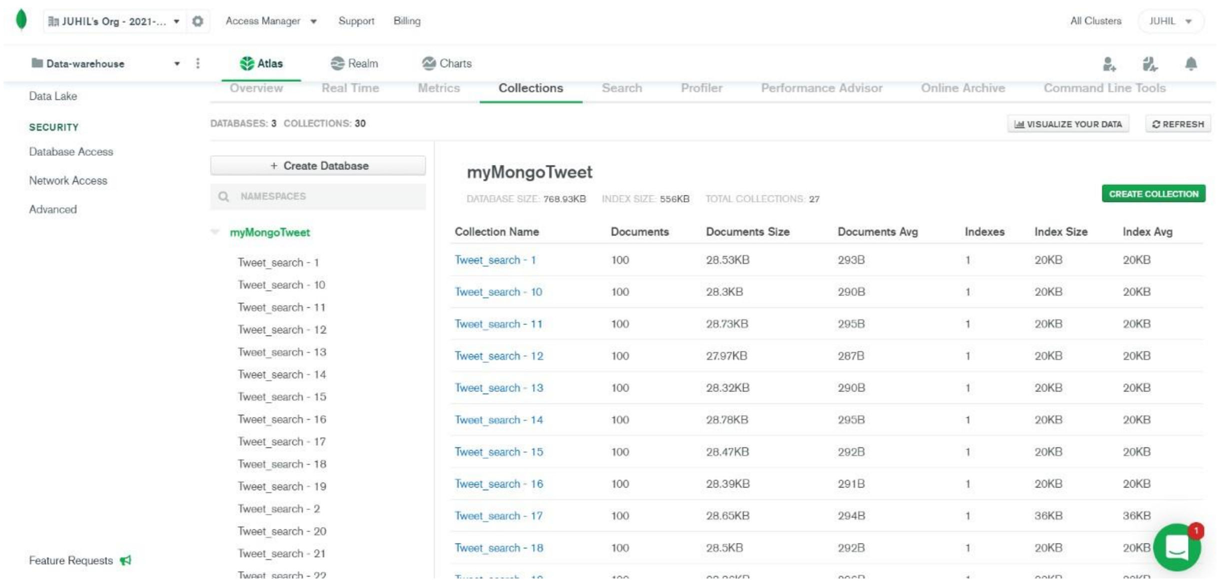
2) After running the search data extraction script



```
Tweet :WHO Expert Expects Covid-19's Origins to Be Found Within a Few Years https://t.co/pnw5QgZBaT
UserName :ICEHInvest_News
Created :2021-03-11 08:07:28
User_location :World🌏
User_followers :107
Retweet_count :0
Coordinates :None

Tweet :@Qvittra Morning Pernilla definitely don't need snow 😔
UserName :6mozz5
Created :2021-03-11 08:07:28
User_location :A-U-L, CHAMPIONS 12,14,18,19
User_followers :776
Retweet_count :0
Coordinates :None

Tweet :RT @BloombergAsia: Hong Kong's vaccine no-show rate rises after the city reported three deaths and three critical illnesses among the more…
UserName :bobbyloveshk
Created :2021-03-11 08:07:28
User_location :Hong Kong
User_followers :26
Retweet_count :51
Coordinates :None

Tweet :👏

COVID-19: Morrisons extends 10% discount for NHS staff amid pay row for nurses
https://t.co/x98NLyL3Ww https://t.co/3JCxTi8aPH
UserName :gallantfive
Created :2021-03-11 08:07:28
User_location :Bath, England
User_followers :13866
Retweet_count :0
Coordinates :None

Tweet :RT @KhinMya79426049: @nslwin Today happening in Myanig.
There is an emergency situation in Myaing.
SO SAD😭😭😭
https://t.co/x3ly5xBagM
UserName :PhyoMon14
Created :2021-03-11 08:07:28
User_location :United States
User_followers :317
Retweet_count :1
Coordinates :None
```

Figure 3. After running the search data extraction script

3) Output of the search data extraction script on MongoDB



DATABASES: 3 COLLECTIONS: 30

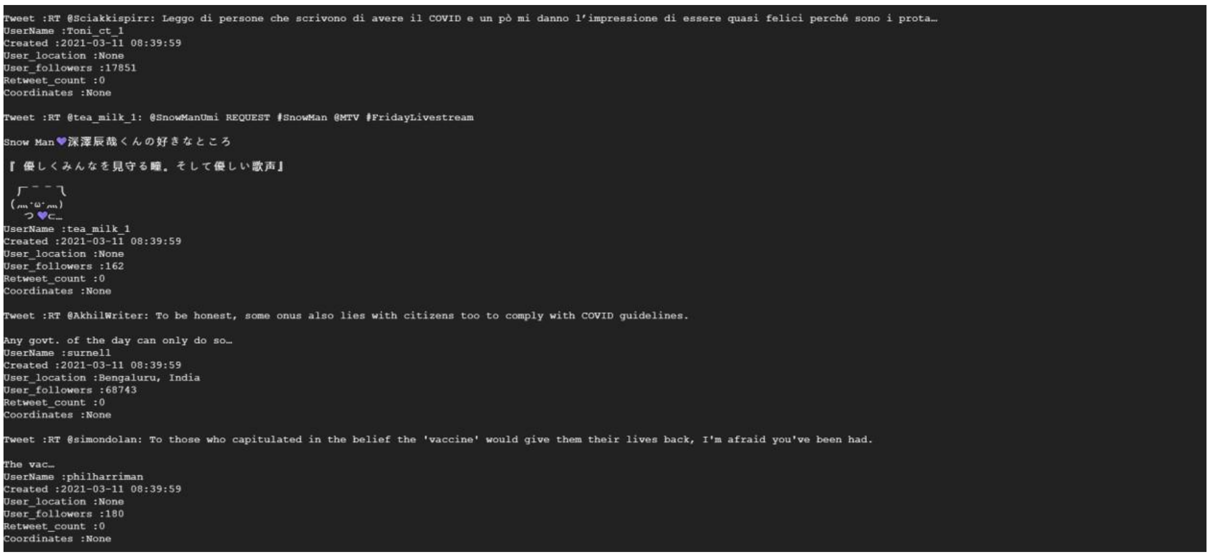
myMongoTweet

DATABASE SIZE: 768.93KB INDEX SIZE: 556KB TOTAL COLLECTIONS: 27

Collection Name	Documents	Documents Size	Documents Avg	Indexes	Index Size	Index Avg
Tweet_search - 1	100	28.53KB	293B	1	20KB	20KB
Tweet_search - 10	100	28.3KB	290B	1	20KB	20KB
Tweet_search - 11	100	28.73KB	295B	1	20KB	20KB
Tweet_search - 12	100	27.97KB	287B	1	20KB	20KB
Tweet_search - 13	100	28.32KB	290B	1	20KB	20KB
Tweet_search - 14	100	28.78KB	295B	1	20KB	20KB
Tweet_search - 15	100	28.47KB	292B	1	20KB	20KB
Tweet_search - 16	100	28.39KB	291B	1	20KB	20KB
Tweet_search - 17	100	28.65KB	294B	1	36KB	36KB
Tweet_search - 18	100	28.5KB	292B	1	20KB	20KB

Figure 4. The output of search data extraction on MongoDB

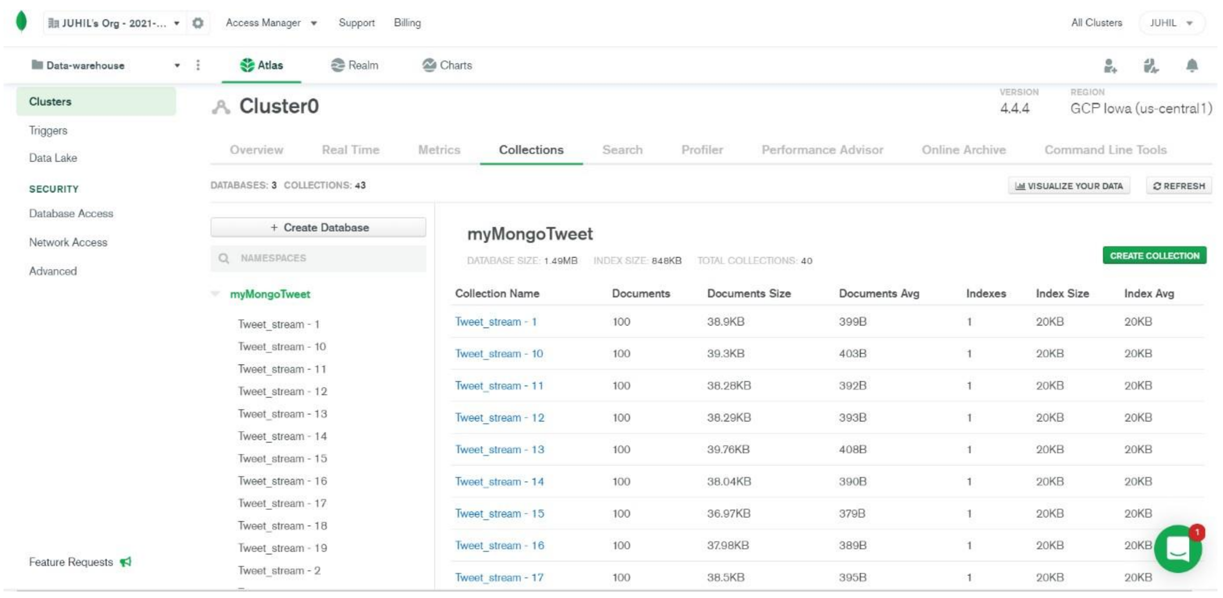
4) After running the stream data extraction script



```
Tweet :RT @Sciakkispirr: Leggo di persone che scrivono di avere il COVID e un pò mi danno l'impressione di essere quasi felici perché sono i prota…
UserName :Toni_ct_1
Created :2021-03-11 08:39:59
User_location :None
User_followers :17851
Retweet_count :0
Coordinates :None

Tweet :RT @tea_milk_1: @SnowManUmi REQUEST #SnowMan @MTV #FridayLivestream

Snow Man💜深澤辰哉くんの好きなところ

『 優しくみんなを見守る瞳。そして優しい歌声』

  ╭￣￣╮
 (ᵔ.˘ω˘.ᵔ)
   つ💜c…
UserName :tea_milk_1
Created :2021-03-11 08:39:59
User_location :None
User_followers :162
Retweet_count :0
Coordinates :None

Tweet :RT @AkhilWriter: To be honest, some onus also lies with citizens too to comply with COVID guidelines.

Any govt. of the day can only do so…
UserName :surnell
Created :2021-03-11 08:39:59
User_location :Bengaluru, India
User_followers :68743
Retweet_count :0
Coordinates :None

Tweet :RT @simondolan: To those who capitulated in the belief the 'vaccine' would give them their lives back, I'm afraid you've been had.

The vac…
UserName :philharriman
Created :2021-03-11 08:39:59
User_location :None
User_followers :180
Retweet_count :0
Coordinates :None
```

Figure 5. After running the stream data extraction script

5) Output of the stream data extraction script on MongoDB



Cluster0

VERSION 4.4.4 REGION GCP Iowa (us-central1)

DATABASES: 3 COLLECTIONS: 43

myMongoTweet
- Tweet_stream - 1
- Tweet_stream - 10
- Tweet_stream - 11
- Tweet_stream - 12
- Tweet_stream - 13
- Tweet_stream - 14
- Tweet_stream - 15
- Tweet_stream - 16
- Tweet_stream - 17
- Tweet_stream - 18
- Tweet_stream - 19
- Tweet_stream - 2

myMongoTweet

DATABASE SIZE: 1.49MB INDEX SIZE: 848KB TOTAL COLLECTIONS: 40

Collection Name	Documents	Documents Size	Documents Avg	Indexes	Index Size	Index Avg
Tweet_stream - 1	100	38.9KB	399B	1	20KB	20KB
Tweet_stream - 10	100	39.3KB	403B	1	20KB	20KB
Tweet_stream - 11	100	38.28KB	392B	1	20KB	20KB
Tweet_stream - 12	100	38.29KB	393B	1	20KB	20KB
Tweet_stream - 13	100	39.76KB	408B	1	20KB	20KB
Tweet_stream - 14	100	38.04KB	390B	1	20KB	20KB
Tweet_stream - 15	100	36.97KB	379B	1	20KB	20KB
Tweet_stream - 16	100	37.98KB	389B	1	20KB	20KB
Tweet_stream - 17	100	38.5KB	395B	1	20KB	20KB

Figure 6. The output of stream data extraction on MongoDB

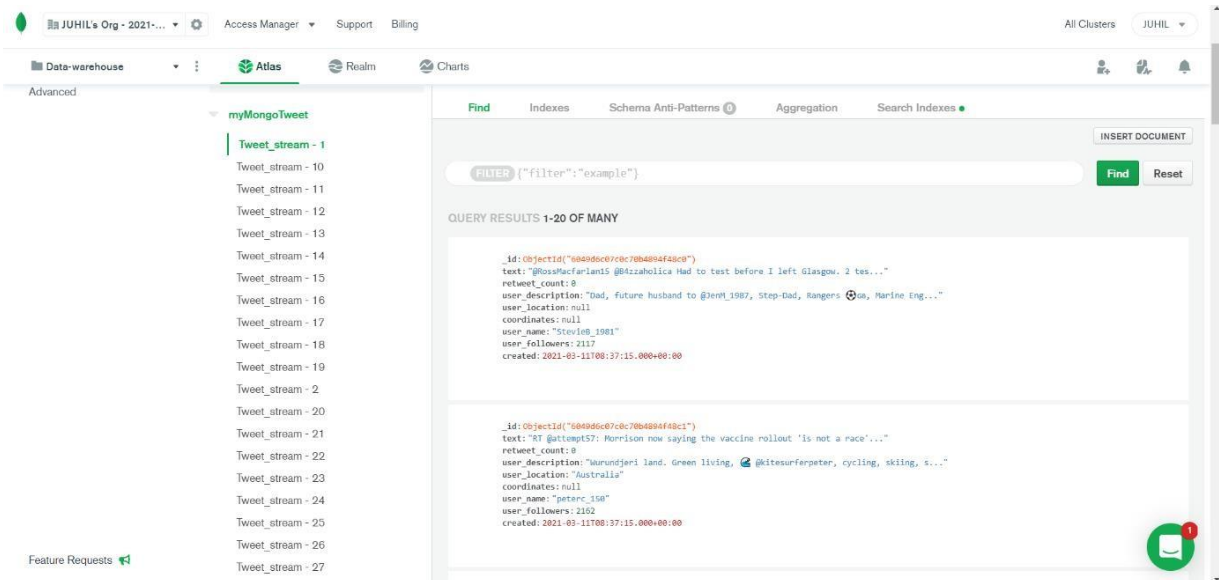
3) Cleaning process:

- At first, I have established a connection to store data in the MongoDB remote database.
- After that, I have defined paths for a database named “myMongoTweet” where all the streaming files are located. I have used previously generated streaming data for cleaning.
- I have used the find() method to find the tweets in the file.
- With the help of for loop, I have accessed each field of tweets like id, text, location and cleaned those fields with the clean_data() function. This function will take that field as input and checks it with the regular expression. As per the assignment requirement, I have defined regular expressions to remove the emoticons, special characters, and URLs.
- After the cleaning part, I have used the insert_one() function to insert all that data into a separate file.

Juhilkumar P. Zalavadiya
B00872280

Output screenshots for twitter data extraction:

1) Before applying the cleaning

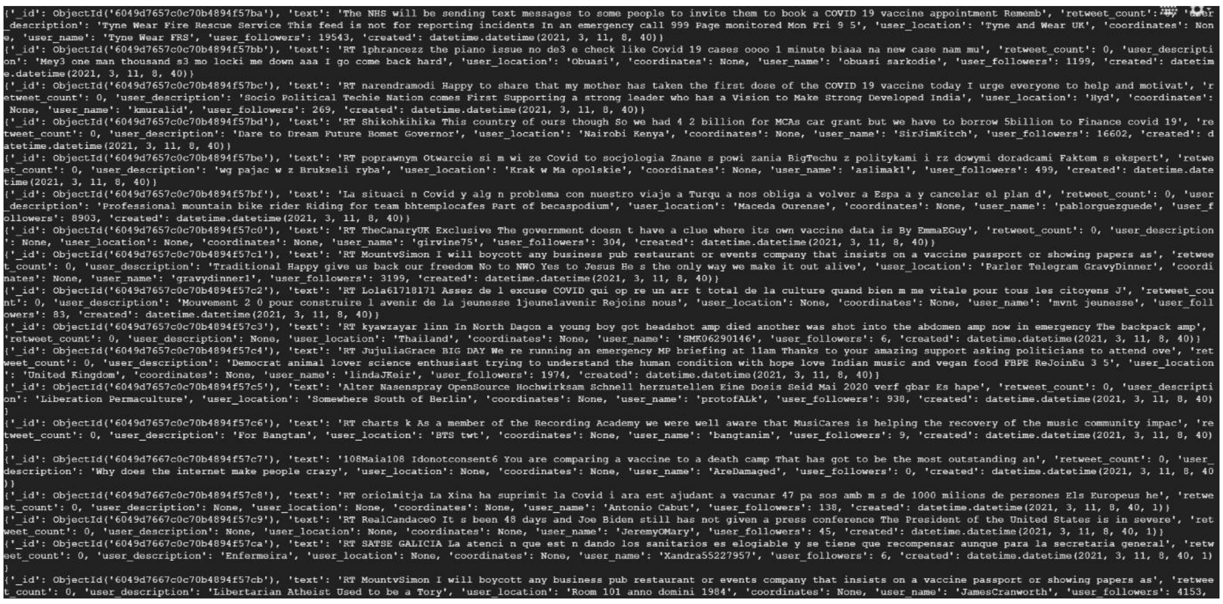


QUERY RESULTS 1-20 OF MANY

```
_id: ObjectId("6049d6c07c0c70b4894f48c0")
text: "@RossMacfarlan15 @B4zzaholica Had to test before I left Glasgow. 2 tes..."
retweet_count: 0
user_description: "Dad, future husband to @JenM_1987, Step-Dad, Rangers ⚽GB, Marine Eng..."
user_location: null
coordinates: null
user_name: "StevieB_1981"
user_followers: 2117
created: 2021-03-11T08:37:15.000+00:00

_id: ObjectId("6049d6c07c0c70b4894f48c1")
text: "RT @attempt57: Morrison now saying the vaccine rollout 'is not a race'..."
retweet_count: 0
user_description: "Wurundjeri land. Green living, 🌊 @kitesurferpeter, cycling, skiing, s..."
user_location: "Australia"
coordinates: null
user_name: "peterc_150"
user_followers: 2162
created: 2021-03-11T08:37:15.000+00:00
```

Figure 7. Before applying the cleaning

2) After running the cleaning script

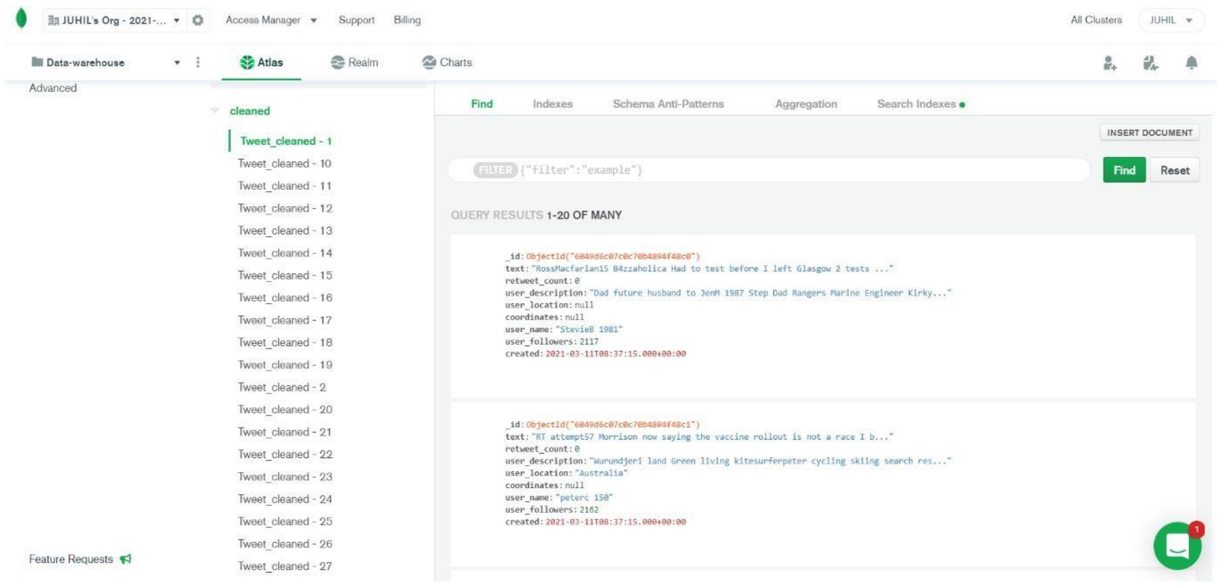


```
{'_id': ObjectId('6049d7657c0c70b4894f57ba'), 'text': 'The NHS will be sending text messages to some people to invite them to book a COVID 19 vaccine appointment Rememb', 'retweet_count': [illegible] 'user
_description': 'Tyne Wear Fire Rescue Service This feed is not for reporting incidents In an emergency call 999 Page monitored Mon Fri 9 5', 'user_location': 'Tyne and Wear UK', 'coordinates': Non
e, 'user_name': 'Tyne Wear FRS', 'user_followers': 19543, 'created': datetime.datetime(2021, 3, 11, 8, 40)}
{'_id': ObjectId('6049d7657c0c70b4894f57bb'), 'text': 'RT 1phrancezz the piano issue no de3 e check like Covid 19 cases oooo 1 minute biaaa na new case nam mu', 'retweet_count': 0, 'user_descripti
on': 'Mey3 one man thousand s3 mo locki me down aaa I go come back hard', 'user_location': 'Obuasi', 'coordinates': None, 'user_name': 'obuasi sarkodie', 'user_followers': 1199, 'created': datetim
e.datetime(2021, 3, 11, 8, 40)}
{'_id': ObjectId('6049d7657c0c70b4894f57bc'), 'text': 'RT narendramodi Happy to share that my mother has taken the first dose of the COVID 19 vaccine today I urge everyone to help and motivat', 'r
etweet_count': 0, 'user_description': 'Socio Political Techie Nation comes First Supporting a strong leader who has a Vision to Make Strong Developed India', 'user_location': 'Hyd', 'coordinates':
 None, 'user_name': 'kmuralid', 'user_followers': 269, 'created': datetime.datetime(2021, 3, 11, 8, 40)}
{'_id': ObjectId('6049d7657c0c70b4894f57bd'), 'text': 'RT Shikohkihika This country of ours though So we had 4 2 billion for MCAs car grant but we have to borrow 5billion to Finance covid 19', 're
tweet_count': 0, 'user_description': 'Dare to Dream Future Bomet Governor', 'user_location': 'Nairobi Kenya', 'coordinates': None, 'user_name': 'SirJimKitch', 'user_followers': 16602, 'created': d
atetime.datetime(2021, 3, 11, 8, 40)}
{'_id': ObjectId('6049d7657c0c70b4894f57be'), 'text': 'RT poprawnym Otwarcie si m wi ze Covid to socjologia Znane s powi zania BigTechu z politykami i rz dowymi doradcami Faktem s ekspert', 'retwe
et_count': 0, 'user_description': 'wg pajac w z Brukseli ryba', 'user_location': 'Krak w Ma opolskie', 'coordinates': None, 'user_name': 'aslimak1', 'user_followers': 499, 'created': datetime.date
time(2021, 3, 11, 8, 40)}
{'_id': ObjectId('6049d7657c0c70b4894f57bf'), 'text': 'La situaci n Covid y alg n problema con nuestro viaje a Turqu a nos obliga a volver a Espa a y cancelar el plan d', 'retweet_count': 0, 'user
_description': 'Professional mountain bike rider Riding for team bhtemplocafes Part of becaspodium', 'user_location': 'Maceda Ourense', 'coordinates': None, 'user_name': 'pablorguezquede', 'user_f
ollowers': 8903, 'created': datetime.datetime(2021, 3, 11, 8, 40)}
{'_id': ObjectId('6049d7657c0c70b4894f57c0'), 'text': 'RT TheCanaryUK Exclusive The government doesn t have a clue where its own vaccine data is By EmmaEGuy', 'retweet_count': 0, 'user_description
': None, 'user_location': None, 'coordinates': None, 'user_name': 'girvine75', 'user_followers': 304, 'created': datetime.datetime(2021, 3, 11, 8, 40)}
{'_id': ObjectId('6049d7657c0c70b4894f57c1'), 'text': 'RT MountvSimon I will boycott any business pub restaurant or events company that insists on a vaccine passport or showing papers as', 'retwee
t_count': 0, 'user_description': 'Traditional Happy give us back our freedom No to NWO Yes to Jesus He s the only way we make it out alive', 'user_location': 'Parler Telegram GravyDinner', 'coordi
nates': None, 'user_name': 'gravydinner1', 'user_followers': 3199, 'created': datetime.datetime(2021, 3, 11, 8, 40)}
{'_id': ObjectId('6049d7657c0c70b4894f57c2'), 'text': 'RT Lola61718171 Assez de l excuse COVID qui op re un arr t total de la culture quand bien m me vitale pour tous les citoyens J', 'retweet_cou
nt': 0, 'user_description': 'Mouvement 2 0 pour construire l avenir de la jeunesse 1jeune1avenir Rejoins nous', 'user_location': None, 'coordinates': None, 'user_name': 'mvnt jeunesse', 'user_foll
owers': 83, 'created': datetime.datetime(2021, 3, 11, 8, 40)}
{'_id': ObjectId('6049d7657c0c70b4894f57c3'), 'text': 'RT kyawzayar linn In North Dagon a young boy got headshot amp died another was shot into the abdomen amp now in emergency The backpack amp',
'retweet_count': 0, 'user_description': None, 'user_location': 'Thailand', 'coordinates': None, 'user_name': 'SMK06290146', 'user_followers': 6, 'created': datetime.datetime(2021, 3, 11, 8, 40)}
{'_id': ObjectId('6049d7657c0c70b4894f57c4'), 'text': 'RT JujuliaGrace BIG DAY We re running an emergency MP briefing at 11am Thanks to your amazing support asking politicians to attend ove', 'ret
weet_count': 0, 'user_description': 'Democrat animal lover science enthusiast trying to understand the human condition with hope love Indian music and vegan food FBPE ReJoinEu 3 5', 'user_location
': 'United Kingdom', 'coordinates': None, 'user_name': 'lindaJKeir', 'user_followers': 1974, 'created': datetime.datetime(2021, 3, 11, 8, 40)}
{'_id': ObjectId('6049d7657c0c70b4894f57c5'), 'text': 'Alter Nasenspray OpenSource Hochwirksam Schnell herzustellen Eine Dosis Seid Mai 2020 verf gbar Es hape', 'retweet_count': 0, 'user_descripti
on': 'Liberation Permaculture', 'user_location': 'Somewhere South of Berlin', 'coordinates': None, 'user_name': 'protofALk', 'user_followers': 938, 'created': datetime.datetime(2021, 3, 11, 8, 40)
}
{'_id': ObjectId('6049d7657c0c70b4894f57c6'), 'text': 'RT charts k As a member of the Recording Academy we were well aware that MusiCares is helping the recovery of the music community impac', 're
tweet_count': 0, 'user_description': 'For Bangtan', 'user_location': 'BTS twt', 'coordinates': None, 'user_name': 'bangtanim', 'user_followers': 9, 'created': datetime.datetime(2021, 3, 11, 8, 40)
}
{'_id': ObjectId('6049d7667c0c70b4894f57c7'), 'text': '108Maia108 Idonotconsent6 You are comparing a vaccine to a death camp That has got to be the most outstanding an', 'retweet_count': 0, 'user_
description': 'Why does the internet make people crazy', 'user_location': None, 'coordinates': None, 'user_name': 'AreDamaged', 'user_followers': 0, 'created': datetime.datetime(2021, 3, 11, 8, 40
)}
{'_id': ObjectId('6049d7667c0c70b4894f57c8'), 'text': 'RT oriolmitja La Xina ha suprimit la Covid i ara est ajudant a vacunar 47 pa sos amb m s de 1000 milions de persones Els Europeus he', 'retwe
et_count': 0, 'user_description': None, 'user_location': None, 'coordinates': None, 'user_name': 'Antonio Cabut', 'user_followers': 138, 'created': datetime.datetime(2021, 3, 11, 8, 40, 1)}
{'_id': ObjectId('6049d7667c0c70b4894f57c9'), 'text': 'RT RealCandaceO It s been 48 days and Joe Biden still has not given a press conference The President of the United States is in severe', 'ret
weet_count': 0, 'user_description': None, 'user_location': None, 'coordinates': None, 'user_name': 'JeremyOMary', 'user_followers': 45, 'created': datetime.datetime(2021, 3, 11, 8, 40, 1)}
{'_id': ObjectId('6049d7667c0c70b4894f57ca'), 'text': 'RT SATSE GALICIA La atenci n que est n dando los sanitarios es elogiable y se tiene que recompensar aunque para la secretaria general', 'retw
eet_count': 0, 'user_description': 'Enfermeira', 'user_location': None, 'coordinates': None, 'user_name': 'Xandra55227957', 'user_followers': 6, 'created': datetime.datetime(2021, 3, 11, 8, 40, 1)
}
{'_id': ObjectId('6049d7667c0c70b4894f57cb'), 'text': 'RT MountvSimon I will boycott any business pub restaurant or events company that insists on a vaccine passport or showing papers as', 'retwee
t_count': 0, 'user_description': 'Libertarian Atheist Used to be a Tory', 'user_location': 'Room 101 anno domini 1984', 'coordinates': None, 'user_name': 'JamesCranworth', 'user_followers': 4153,
```

Figure 8. After running the cleaning script

Juhilkumar P. Zalavadiya
B00872280

3) Output of the cleaned data on MongoDB



```
{"filter":"example"}

QUERY RESULTS 1-20 OF MANY

_id: ObjectId("6049d6c07c0c70b4894f48c0")
text: "RossMacfarlan15 B4zzaholica Had to test before I left Glasgow 2 tests ..."
retweet_count: 0
user_description: "Dad future husband to JenM 1987 Step Dad Rangers Marine Engineer Kirky..."
user_location: null
coordinates: null
user_name: "StevieB 1981"
user_followers: 2117
created: 2021-03-11T08:37:15.000+00:00

_id: ObjectId("6049d6c07c0c70b4894f48c1")
text: "RT attempt57 Morrison now saying the vaccine rollout is not a race I b..."
retweet_count: 0
user_description: "Wurundjeri land Green living kitesurferpeter cycling skiing search res..."
user_location: "Australia"
coordinates: null
user_name: "peterc 150"
user_followers: 2162
created: 2021-03-11T08:37:15.000+00:00
```

Figure 9. The output of the cleaned data on MongoDB

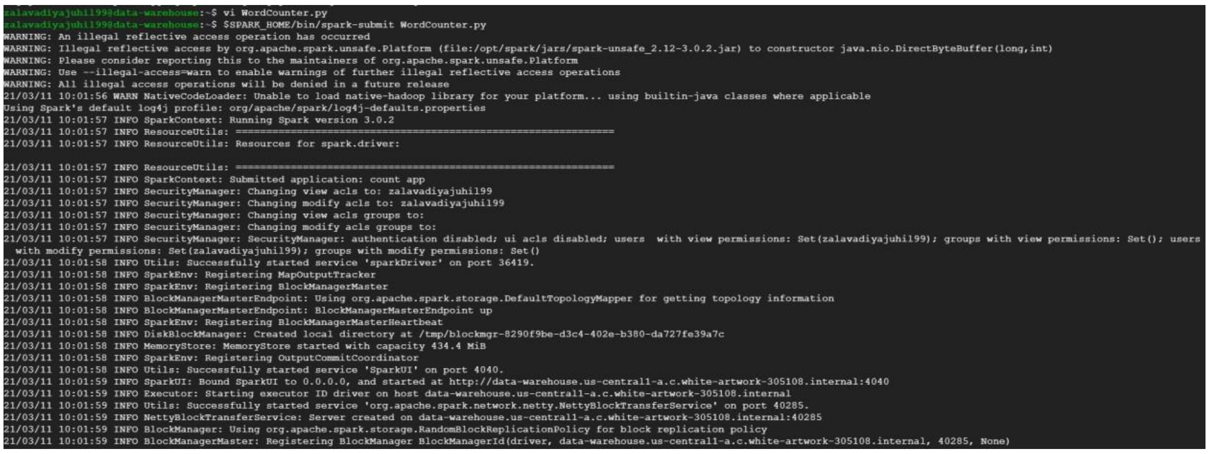
Juhilkumar P. Zalavadiya
B00872280

Problem #2

Task 1: Data Processing using Spark - MapReduce to perform count

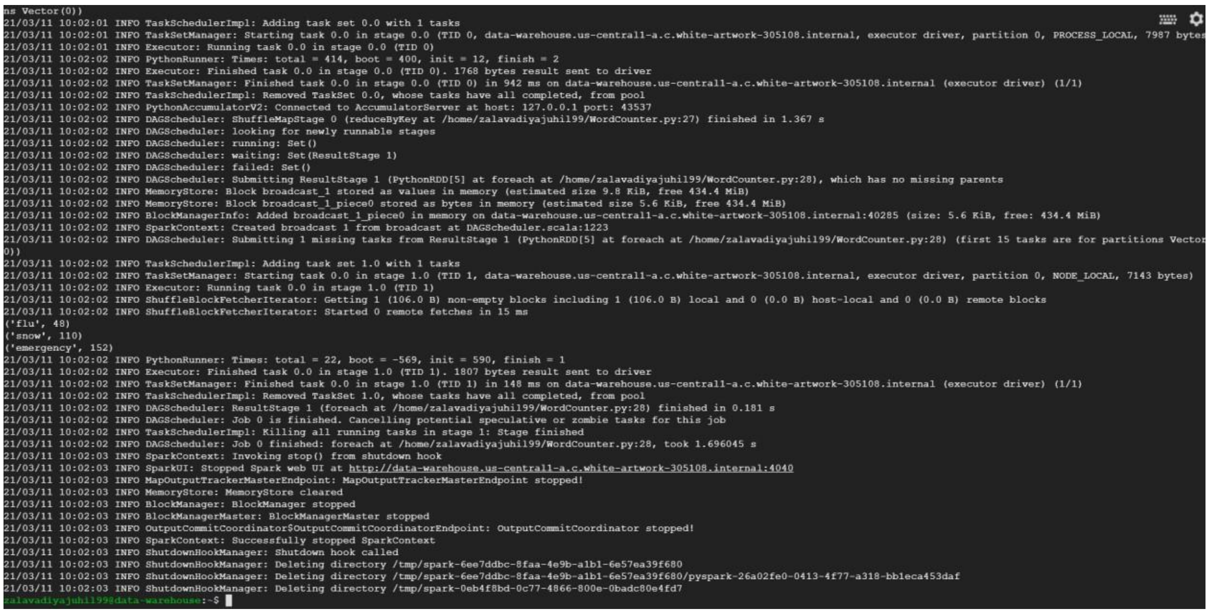
- MapReduce is used for the processing of extensive data set. The task is divided into many smaller parts to make the processing efficient and fast.
- To perform this task, I have used the SparkContext library for mapping and reduce tasks. I also used the MongoClient library to connect with the MongoDB remote database.
- After the connection part, I have given the path for the database where all the cleaned data is stored in different files.
- For storing the list of searched keywords mentioned in the assignment requirement, I have defined the initial_list.
- After that, I have used one for loop to access different files in the database.
- I have used another for loop to break the string into a list of words and compare it with the keywords I want to search.
- After all these steps, I will have a list of words required for the given task. I have stored all these keywords in the initial_list.
- I used parallelized() method to convert that initial_list into Resilient Distributed Dataset.
- After that, mapping is done with the help of the map() function. Its output will be given for the reducing part. SpartkContext library has one method named reduceByKey() for the reducing part. So, I have used it and stored the output in the reduced_map list.

The output of the MapReduce program:



```
zalavadiyajuhil99@data-warehouse:~$ vi WordCounter.py
zalavadiyajuhil99@data-warehouse:~$ $SPARK_HOME/bin/spark-submit WordCounter.py
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.spark.unsafe.Platform (file:/opt/spark/jars/spark-unsafe_2.12-3.0.2.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
21/03/11 10:01:56 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
21/03/11 10:01:57 INFO SparkContext: Running Spark version 3.0.2
21/03/11 10:01:57 INFO ResourceUtils: ==============================================================
21/03/11 10:01:57 INFO ResourceUtils: Resources for spark.driver:

21/03/11 10:01:57 INFO ResourceUtils: ==============================================================
21/03/11 10:01:57 INFO SparkContext: Submitted application: count app
21/03/11 10:01:57 INFO SecurityManager: Changing view acls to: zalavadiyajuhil99
21/03/11 10:01:57 INFO SecurityManager: Changing modify acls to: zalavadiyajuhil99
21/03/11 10:01:57 INFO SecurityManager: Changing view acls groups to:
21/03/11 10:01:57 INFO SecurityManager: Changing modify acls groups to:
21/03/11 10:01:57 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  with view permissions: Set(zalavadiyajuhil99); groups with view permissions: Set(); users
  with modify permissions: Set(zalavadiyajuhil99); groups with modify permissions: Set()
21/03/11 10:01:58 INFO Utils: Successfully started service 'sparkDriver' on port 36419.
21/03/11 10:01:58 INFO SparkEnv: Registering MapOutputTracker
21/03/11 10:01:58 INFO SparkEnv: Registering BlockManagerMaster
21/03/11 10:01:58 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
21/03/11 10:01:58 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
21/03/11 10:01:58 INFO SparkEnv: Registering BlockManagerMasterHeartbeat
21/03/11 10:01:58 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-8290f9be-d3c4-402e-b380-da727fe39a7c
21/03/11 10:01:58 INFO MemoryStore: MemoryStore started with capacity 434.4 MiB
21/03/11 10:01:58 INFO SparkEnv: Registering OutputCommitCoordinator
21/03/11 10:01:58 INFO Utils: Successfully started service 'SparkUI' on port 4040.
21/03/11 10:01:59 INFO SparkUI: Bound SparkUI to 0.0.0.0, and started at http://data-warehouse.us-central1-a.c.white-artwork-305108.internal:4040
21/03/11 10:01:59 INFO Executor: Starting executor ID driver on host data-warehouse.us-central1-a.c.white-artwork-305108.internal
21/03/11 10:01:59 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 40285.
21/03/11 10:01:59 INFO NettyBlockTransferService: Server created on data-warehouse.us-central1-a.c.white-artwork-305108.internal:40285
21/03/11 10:01:59 INFO BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy
21/03/11 10:01:59 INFO BlockManagerMaster: Registering BlockManager BlockManagerId(driver, data-warehouse.us-central1-a.c.white-artwork-305108.internal, 40285, None)
```

Figure 10. The output of the MapReduce - 1



```
ns Vector(0))
21/03/11 10:02:01 INFO TaskSchedulerImpl: Adding task set 0.0 with 1 tasks
21/03/11 10:02:01 INFO TaskSetManager: Starting task 0.0 in stage 0.0 (TID 0, data-warehouse.us-central1-a.c.white-artwork-305108.internal, executor driver, partition 0, PROCESS_LOCAL, 7987 bytes
21/03/11 10:02:01 INFO Executor: Running task 0.0 in stage 0.0 (TID 0)
21/03/11 10:02:02 INFO PythonRunner: Times: total = 414, boot = 400, init = 12, finish = 2
21/03/11 10:02:02 INFO Executor: Finished task 0.0 in stage 0.0 (TID 0). 1768 bytes result sent to driver
21/03/11 10:02:02 INFO TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0) in 942 ms on data-warehouse.us-central1-a.c.white-artwork-305108.internal (executor driver) (1/1)
21/03/11 10:02:02 INFO TaskSchedulerImpl: Removed TaskSet 0.0, whose tasks have all completed, from pool
21/03/11 10:02:02 INFO PythonAccumulatorV2: Connected to AccumulatorServer at host: 127.0.0.1 port: 43537
21/03/11 10:02:02 INFO DAGScheduler: ShuffleMapStage 0 (reduceByKey at /home/zalavadiyajuhil99/WordCounter.py:27) finished in 1.367 s
21/03/11 10:02:02 INFO DAGScheduler: looking for newly runnable stages
21/03/11 10:02:02 INFO DAGScheduler: running: Set()
21/03/11 10:02:02 INFO DAGScheduler: waiting: Set(ResultStage 1)
21/03/11 10:02:02 INFO DAGScheduler: failed: Set()
21/03/11 10:02:02 INFO DAGScheduler: Submitting ResultStage 1 (PythonRDD[5] at foreach at /home/zalavadiyajuhil99/WordCounter.py:28), which has no missing parents
21/03/11 10:02:02 INFO MemoryStore: Block broadcast_1 stored as values in memory (estimated size 9.8 KiB, free 434.4 MiB)
21/03/11 10:02:02 INFO MemoryStore: Block broadcast_1_piece0 stored as bytes in memory (estimated size 5.6 KiB, free 434.4 MiB)
21/03/11 10:02:02 INFO BlockManagerInfo: Added broadcast_1_piece0 in memory on data-warehouse.us-central1-a.c.white-artwork-305108.internal:40285 (size: 5.6 KiB, free: 434.4 MiB)
21/03/11 10:02:02 INFO SparkContext: Created broadcast 1 from broadcast at DAGScheduler.scala:1223
21/03/11 10:02:02 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 1 (PythonRDD[5] at foreach at /home/zalavadiyajuhil99/WordCounter.py:28) (first 15 tasks are for partitions Vector
0))
21/03/11 10:02:02 INFO TaskSchedulerImpl: Adding task set 1.0 with 1 tasks
21/03/11 10:02:02 INFO TaskSetManager: Starting task 0.0 in stage 1.0 (TID 1, data-warehouse.us-central1-a.c.white-artwork-305108.internal, executor driver, partition 0, NODE_LOCAL, 7143 bytes)
21/03/11 10:02:02 INFO Executor: Running task 0.0 in stage 1.0 (TID 1)
21/03/11 10:02:02 INFO ShuffleBlockFetcherIterator: Getting 1 (106.0 B) non-empty blocks including 1 (106.0 B) local and 0 (0.0 B) host-local and 0 (0.0 B) remote blocks
21/03/11 10:02:02 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 15 ms
('flu', 48)
('snow', 110)
('emergency', 152)
21/03/11 10:02:02 INFO PythonRunner: Times: total = 22, boot = -569, init = 590, finish = 1
21/03/11 10:02:02 INFO Executor: Finished task 0.0 in stage 1.0 (TID 1). 1807 bytes result sent to driver
21/03/11 10:02:02 INFO TaskSetManager: Finished task 0.0 in stage 1.0 (TID 1) in 148 ms on data-warehouse.us-central1-a.c.white-artwork-305108.internal (executor driver) (1/1)
21/03/11 10:02:02 INFO TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed, from pool
21/03/11 10:02:02 INFO DAGScheduler: ResultStage 1 (foreach at /home/zalavadiyajuhil99/WordCounter.py:28) finished in 0.181 s
21/03/11 10:02:02 INFO DAGScheduler: Job 0 is finished. Cancelling potential speculative or zombie tasks for this job
21/03/11 10:02:02 INFO TaskSchedulerImpl: Killing all running tasks in stage 1: Stage finished
21/03/11 10:02:02 INFO DAGScheduler: Job 0 finished: foreach at /home/zalavadiyajuhil99/WordCounter.py:28, took 1.696045 s
21/03/11 10:02:03 INFO SparkContext: Invoking stop() from shutdown hook
21/03/11 10:02:03 INFO SparkUI: Stopped Spark web UI at http://data-warehouse.us-central1-a.c.white-artwork-305108.internal:4040
21/03/11 10:02:03 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
21/03/11 10:02:03 INFO MemoryStore: MemoryStore cleared
21/03/11 10:02:03 INFO BlockManager: BlockManager stopped
21/03/11 10:02:03 INFO BlockManagerMaster: BlockManagerMaster stopped
21/03/11 10:02:03 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
21/03/11 10:02:03 INFO SparkContext: Successfully stopped SparkContext
21/03/11 10:02:03 INFO ShutdownHookManager: Shutdown hook called
21/03/11 10:02:03 INFO ShutdownHookManager: Deleting directory /tmp/spark-6ee7ddbc-8faa-4e9b-a1b1-6e57ea39f680
21/03/11 10:02:03 INFO ShutdownHookManager: Deleting directory /tmp/spark-6ee7ddbc-8faa-4e9b-a1b1-6e57ea39f680/pyspark-26a02fe0-0413-4f77-a318-bb1eca453daf
21/03/11 10:02:03 INFO ShutdownHookManager: Deleting directory /tmp/spark-0eb4f8bd-0c77-4866-800e-0badc80e4fd7
zalavadiyajuhil99@data-warehouse:~$
```

Figure 11. The output of the MapReduce - 2

Highest frequencies: emergency -> 152
Lowest frequencies: flu -> 48

Juhilkumar P. Zalavadiya
B00872280

References:

[1] "Google Cloud Services" , Google Cloud, 2021. [Online]. Available: https://cloud.google.com/ [Accessed: 5-March 2021]

[2] "Apache Spark – Python Programming Guide" [Online]. Available: https://spark.apache.org/docs/0.9.0/python-programming-guide.html [Accessed: 6-March 2021]

[3] "MongoDB Cloud", MongoDB,2021. [Online]. Available: https://www.mongodb.com/cloud [Accessed: 8-March 2021]

[4] "Tweepy – An easy-to-use Python library for accessing the Twitter API" [Online]. Available: https://www.tweepy.org/ [Accessed: 8-March 2021]