**CSEE5590/490: Big Data Programming**

**Project Increment 2**

**Due Date: March26(Fri) 2021, 11:59PM**

**Project Increment 02 Report**

**PROJECT TITLE : COVID 19 Sentiment Analysis**

***GROUP 5***

***VYOMA DESAI - 16314631***

***AFFAN CHAROLIA - 16305677***

***ALI ALYAMI- 16293096***

**INTRODUCTION**

COVID-19 has affected every single person around the globe and has had a huge impact on businesses everywhere. People are sounding their emotions on various social networking platforms and one of those platforms is twitter.

The hashtag “Covid19” is trending and people are voicing their sentiments on it.Twitter makes public Tweets and replies available to developers.

These endpoints can easily be used by people to identify, understand and counter misinformation around public health initiatives.

**MOTIVATION**

COVID-19 has affected every single person around the globe and has had a huge impact on businesses everywhere. People are sounding their emotions on various social networking platforms and one of those platforms is twitter. The hashtag “Covid19” is trending and people are voicing their sentiments on it.Twitter makes public Tweets and replies available to developers. Hence, it allows developers to post Tweets via API. Developers can access Tweets by searching for specific keywords,or requesting a sample of Tweets from specific accounts. These endpoints can easily be used by people to identify, understand and counter misinformation around public health initiatives.

**SIGNIFICANCE**

The pandemic of Covid-19 has taken everyone by surprise. Many people have lost their jobs and their businesses. It's a topic which has grappled everyone and is a point of discussion and debate. There have been cases where people have had breakdowns because of this\environment that has been created because of it. So it is significant to know how people are feeling about this and what are their sentiments towards it.

**OBJECTIVES**

1. To understand the sentiments of people.

2. To find out which people are affected the most.

3. Which locations are affected the most.

**FEATURES**

We collected tweets on the topic “Covid-19” and will be pre-processing the data to remove unnecessary data. Further we will be using Hadoop and Spark to analyze the data. We will be using various Hadoop tools like MapReduce, sqoop and hive and also some of the spark tools to help assist us in this project.

**DATASET :**

We are using COVID-19 tweets dataset from all over the world that we got from Kaggle website. Our aim is to perform sentiment analysis on the dataset so that we can conclude about the impact of COVID-19 around the world. Whatever they express on social media like twitter.

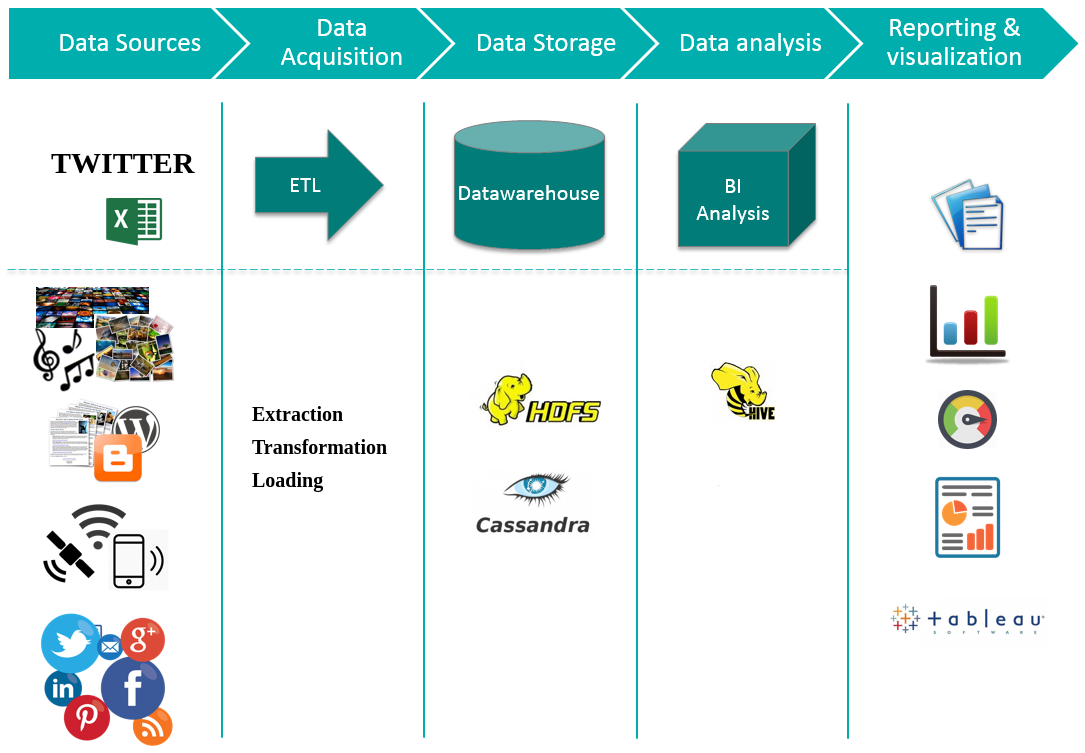
Dataset URL:<https://www.kaggle.com/gpreda/covid19-tweets>

Dataset Features:

Provided dataset contains the following features or tables. User\_name, user\_location, user\_description, user\_created, user\_followers, user\_friends, user\_favourites, user\_verified, date, text, hashtag, source, is\_retweet.

If we consider these fields then we can see that we need very few of them, such as we will be using the “text” field to get the text of tweet, whatever someone has written about the COVID-19. Other tables such as user location can be targeted to find out about the people of a certain area or geo position. We can also use hastags to find out how much the people are expressing the terms.

**PROJECT WORKFLOW:**



**Data Analysis: Fields Description:**

|  |  |
| --- | --- |
| Field Title | Description |
| User\_name | Describes the user name of person on twitter |
| User\_location | Describes the location of that person |
| User\_description | Describes the description of that user from twitter account |
| User\_created | Describes when the account of user was created |
| User\_followers | Describes how many followers that user have |
| User\_friends | Describes how many friends the user have |
| User\_favourites | Describes the favorites of the user |
| User\_verified | Describes if the user is verified or not |
| Date | Date when the tweet was made |
| Text | Describes the text that was written in the tweet |
| Hashtag | Describes the hashtags that were used in the tweet |
| Source | Describes the source from where tweet was made such as Phone/Desktop |
| Is\_retweet | Describes if it’s a retweet or not |

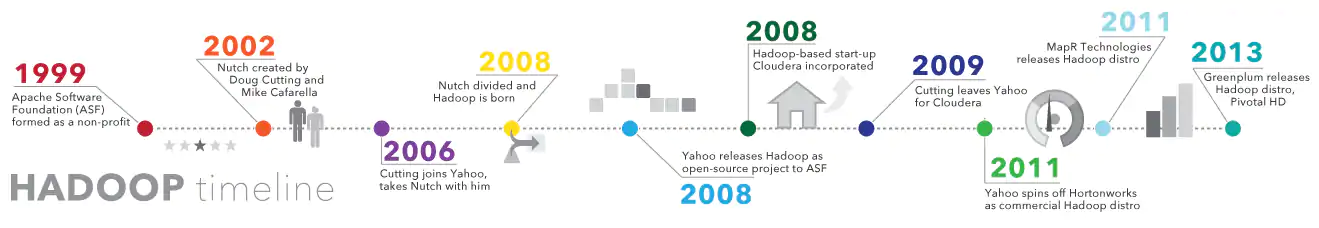
the fields we are looking for are text, date, and user\_location.

**IMPLEMENTATION :**

Hadoop:

Apache Hadoop is an open-source suite of software utilities that make it easy to use a multicomputer network to solve data and computing problems. It provides a software framework for the process of distributing and storing large data using the Mapreduce programming model.

As the World Wide Web grew in the late 1900s and late 2000s, search engines and indexes were developed to help find relevant information between text-pages. In the early years, people returned search results. But as the Internet grew from tens to millions of pages, automation was needed. Web crawlers were developed, most of which were research projects, and new search engines (Yahoo, AltaVista, etc.) began to emerge.



One such project was an open source search engine called Dutch Cutter and Mike Caffrela's Brain Builder. They wanted to quickly return web search results by spreading data and calculations across different computers so that they could perform more than one task at a time. Now, another search engine project called Google is being developed. It was based on the same concept - splitting, automatically storing and processing data so that relevant search results could be returned quickly.

In 2006, Cutting joined Yahoo! and, along with the Nich project, embraced the idea of ​​Google's initial work to automate distributed data storage and processing. The Natch Project split: part of the web crawler remained nuts, and part of the distributed computing and processing became (named after the cutting son's toy elephant). In 2008, Yahoo released Hadoop as an open-source project. Today, the Hadoop Framework and Technology Ecosystem is managed and maintained by the nonprofit Apache Software Foundation (ASF), the global community of software developers and partners.

**Why is Hadoop important?**

Ability to quickly store and process large amounts of any data: This is an important consideration as the amount of data is increasing in various ways, especially from social media and the Internet of Things (IoT).

**Computing power**: The HUDP grid computing model is faster in processing large data. The more compute nodes you use, the more computational power you have.

**Fault tolerance**: is protected from data and application processing hardware failures. If the node fails, the jobs are automatically redirected to other nodes to ensure that grid computing does not fail. Multiple copies of all data are automatically saved.

**Flexibility**: Unlike traditional relative databases, you do not need to pre-process the data before storing it. You can store as much as you want and decide how to use it later. This includes unstructured data such as text, images and videos.

**Low cost:** open-source platform is free and uses common hardware to store large amounts of data.

**Scalability**: You can easily extend your system to handle more data by adding nodes. Requires small administration.

**MapReduce:**

MapReduce is a processing method and programming model for Java-based distributed computing. MapReduce is based on java distributed computing programming model and processing model. The MapReduce algorithm has two main functions, namely mapping and reduction. The map takes one dataset and converts it into another dataset, where the individual elements are divided into topples (key / value pairs).

Steps to complete a MapReduce Task:

**Step 1**: A block is processed by a mapper at the same time. In mapper, the developer can define his proposed logic as needed. As such, the map runs across all cluster nodes and processes data blocks in parallel.

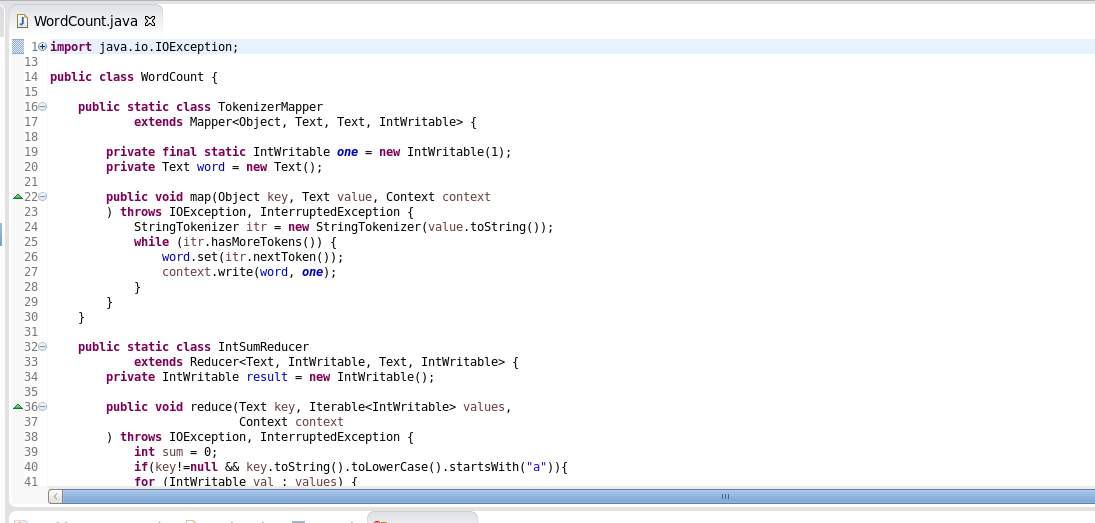
**Step 2**: Mapper output, also called staging output, is written to the local disk. Mapper output is not stored in HDFS because it is temporary data and writing to HDFS will create unnecessary copies.

**Step 3**: Output is converted to a reducer node (which is a common slave node, phase will continue here, hence it is called reducer node). Shuffling / copying is the physical movement of data across a network.

**Step 4:** After all the matches have been completed and their output is adjusted on the reducer nodes, this intermediate output is integrated and configured. Which is then provided as input to reduce the phase.

**Step 5:** Reduce is the second processing step where the user can articulate their business logic as needed. An input to a reducer is provided from all the mappers. The reducer output is the final output listed on the HDFS.

**Java Code (MapReduce WordCount):**

****

**hadoop fs -mkdir /user/cloudera/tweets**

It will create a directory named tweets under user/cloudera in hadoop file system

**hadoop fs -mkdir /user/cloudera/tweets/input**

We will create a sub directory named as input to save our input file here

**hadoop fs -put text.txt /user/cloudera/tweets/input**

It will save the text file in input directory

**hadoop jar WordFreqCount.jar WordCount /user/cloudera/tweets/input /user/cloudera/tweets/output**

It will run the jar file and count word in our input file and store them to ouput which we can either check through

HUE or we can run the command below

**hadoop fs -cat /user/cloudera/tweets/ouput/part-r-00000**

****

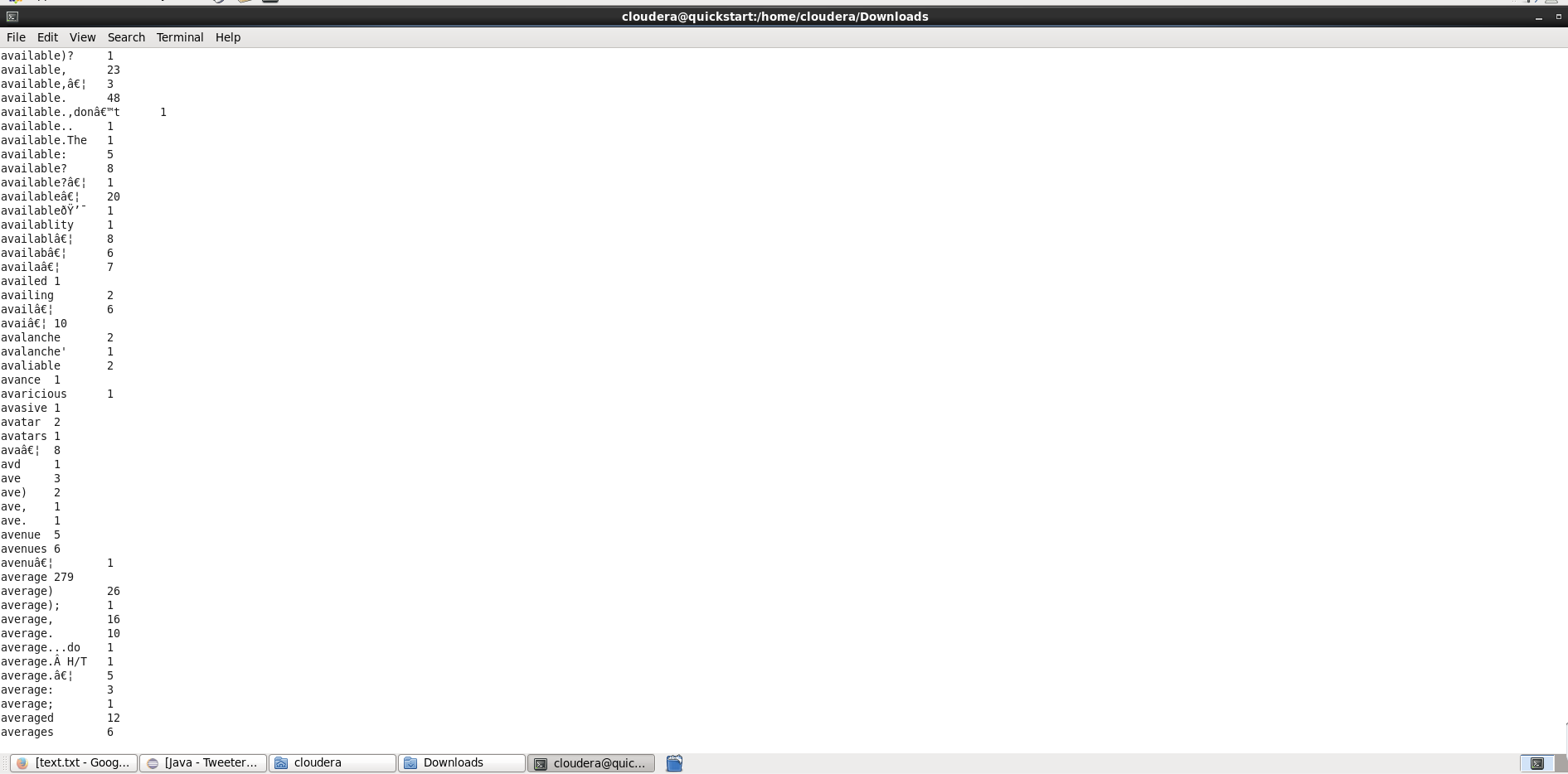
WordCount Output here is showing how many times a certain word was found in the text of tweets we provided from the COVID-19 dataset. From this WordCount we can also create a WordCloud.

**hadoop jar WordFreqaCount.jar WordCount /user/cloudera/tweets/input /user/cloudera/tweets/outputa**

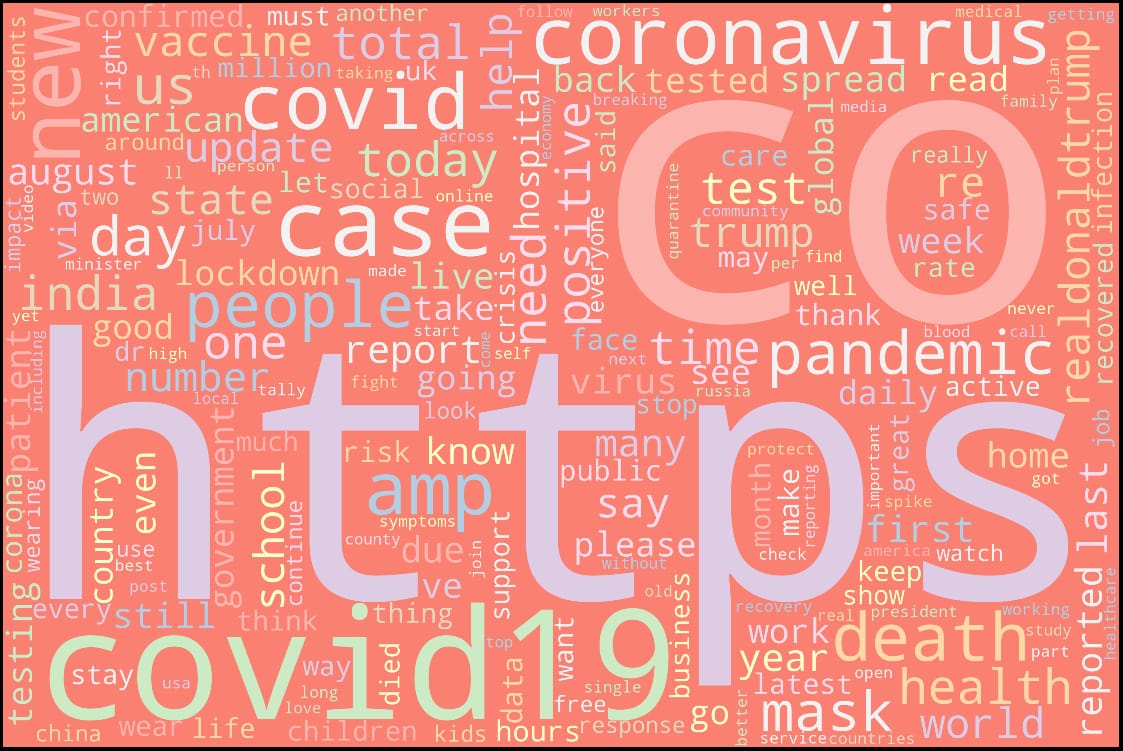
It will run the jar file and count word starting with a letter 'a' in our input file and store them to ouput which

we can either check through HUE or we can run the command below

**hadoop fs -cat /user/cloudera/tweets/ouputa/part-r-00000**



WordCount Output here is showing how many times a certain word that starts with a letter a was found in the text of tweets we provided from the COVID-19 dataset. From this WordCount we can also create a WordCloud.



We have generated WordCloud from our text dataset from which we can view that letter https occurs most of the time and then CO after that covid19 and coronavirus, list goes on.

**CASSANDRA IMPLEMENTATION**

* Covid-19 Twitter Data Implementation

1. We have used Covid-19 Twitter Dataset for our sentimental analysis.
2. Pre-processing of our dataset eliminated special characters.
3. Making use of Cassandra query we have implemented queries to represent data in the form of visualization.

* Dataset:

These tweets are collected using Twitter API and a Python script.

A query for this high-frequency hashtag (#covid19) is run on a daily basis for a certain time period, to collect a larger number of tweets samples.

Our data set mainly contains covid-19 tweets with hashtags #covid-19

* Kaggle Link:

Downloaded KAGGLE Link :

https://www.kaggle.com/gpreda/covid19-tweets

* Data Pre-Processing

1. This Twitter data contained many special characters which failed to upload the large data on cassandra database.
2. Removing special characters from the dataset using python
3. Add csv\_path and define a function clean\_text(string) which will return only those elements which we want.

def clean\_text(string):

return re.sub(r"[^a-zA-Z0-9:/.,@#&]+", ' ', string)

1. For each columns apply lambda function to eliminate all special characters.

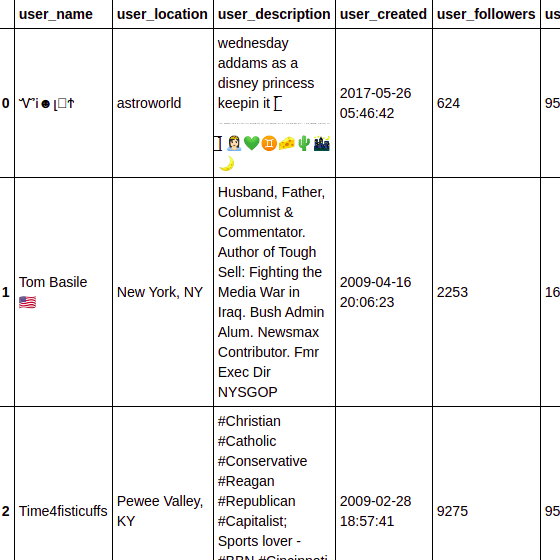
for column in columns:

df[column] = df[column].apply(lambda x: clean\_text(str(x)))

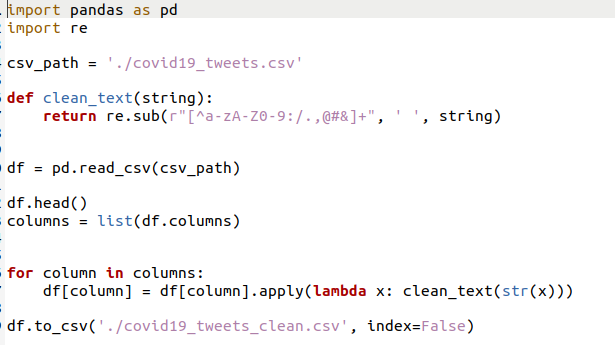
1. Store the new COVID-19 dataset inside covid19\_tweets\_clean.csv

df.to\_csv('./covid19\_tweets\_clean.csv', index=False)

1. We will now use new dataset to run queries for analysing the data.



*Fig : Twitter dataset with special characters.*



*Fig : Preprocessing Twitter Data Set using Python*

* Query Implementation using CQLSH

1. Create a keyspace named covid19\_tweets and create a table named as covid19tweetsdata for a collection column and other columns.

CREATE KEYSPACE covid19tweets

WITH replication = {'class':'SimpleStrategy', 'replication\_factor' : 3};

Describe Keyspace



Using keyspace

cqlsh> USE covid19tweets;

cqlsh:covid19tweets>

1. Create a table and describe table covid19tweetsdata.

Fields Description:

|  |  |
| --- | --- |
| **Field Title** | **Description** |
| User\_name | Describes the user name of person on twitter |
| User\_location | Describes the location of that person |
| User\_description | Describes the description of that user from twitter account |
| User\_created | Describes when the account of user was created |
| User\_followers | Describes how many followers that user have |
| User\_friends | Describes how many friends the user have |
| User\_favourites | Describes the favorites of the user |
| User\_verified | Describes if the user is verified or not |
| Date | Date when the tweet was made |
| Text | Describes the text that was written in the tweet |
| Hashtag | Describes the hashtags that were used in the tweet |
| Source | Describes the source from where tweet was made such as Phone/Desktop |
| Is\_retweet | Describes if it’s a retweet or not |



## Create Primary Key

The primary key is a column that is used to uniquely identify a row. Therefore,defining a primary key is mandatory while creating a table. A primary key is made of one or more columns of a table. You can define a primary key of a table as shown below.

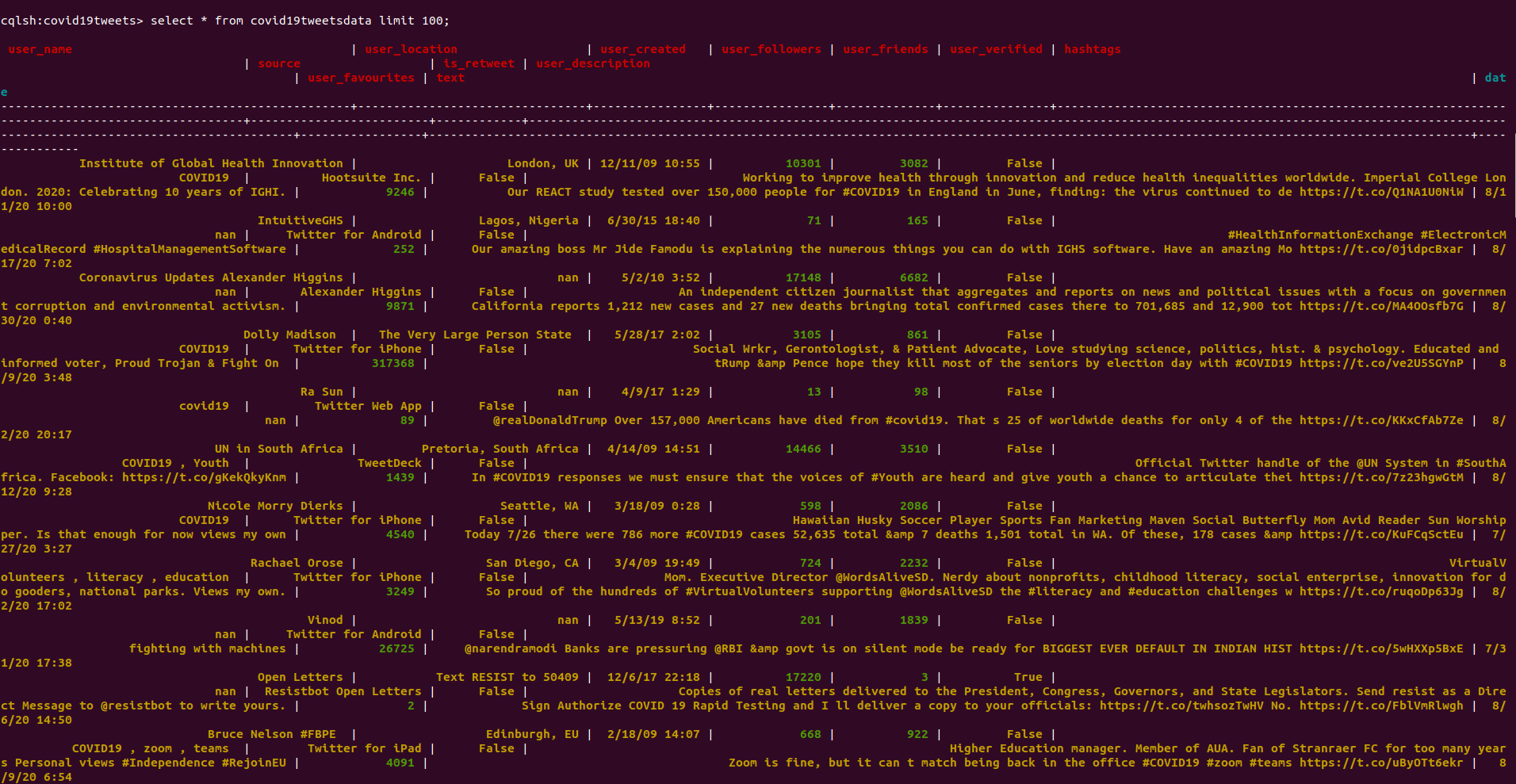
CREATE TABLE tablename(column1 name datatype PRIMARYKEY, column2 name data type, column3 name data type, PRIMARY KEY (column1))

1. Import the CSV values using COPY Command from where the data is stored in home directory covid19\_tweets\_clean,csv using delimiter = “,” and header =true;

COPY covid19tweetsdata (user\_name, user\_location,user\_description, user\_created, user\_followers ,user\_friends , user\_favourites , user\_verified , date , text , hashtags , source ,is\_retweet ) FROM '/home/vyoma/Desktop/BigDataSyllabus/Project Proposal/finalDataset/big\_data\_cassandra/big\_data\_cassandra/covid19\_tweets\_clean.csv' WITH DELIMITER=',' AND HEADER=TRUE;

1. Use a SELECT statement to retrieve a results set from a table as standard output or in JSON format.

Select \* from covid19tweetsdata limit 100;

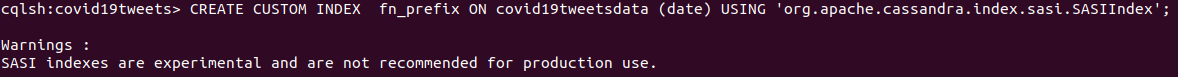


1. Define a new index on a single column of a table. If data already exists for the column, Cassandra indexes the data during the execution of this statement. After the index is created, Cassandra indexes new data for the column automatically when new data is inserted.

**CREATE** **INDEX** ***IF******NOT******EXISTS*** *index\_name*

**ON** *keyspace\_name.*table\_name (***KEYS*** ( column\_name ) )

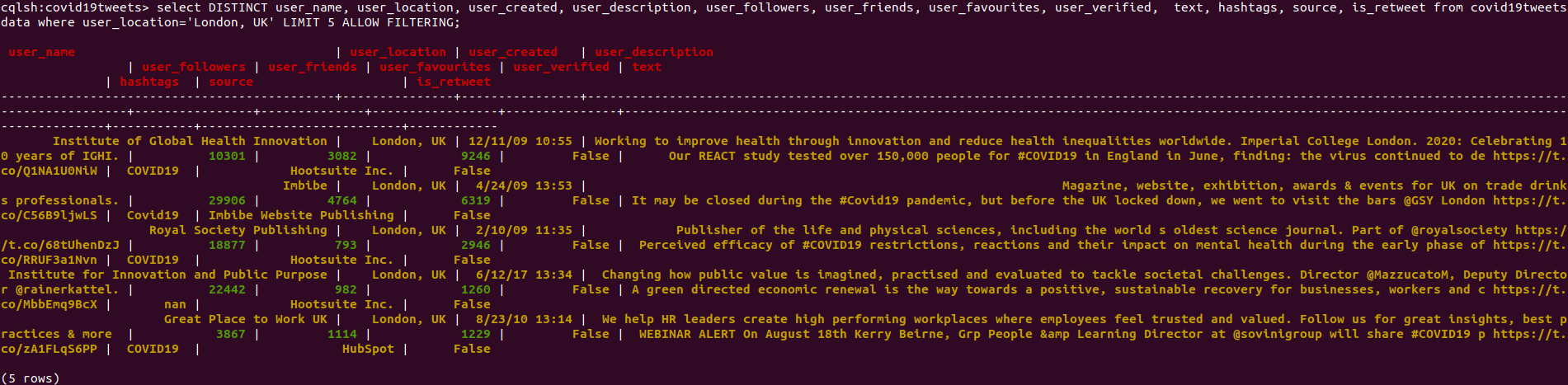
CREATE CUSTOM INDEX fn\_prefix ON covid19tweetsdata (date) using cassandra SASI INDEX



QUERIES ON COVID 19 TWITTER ANALYSIS

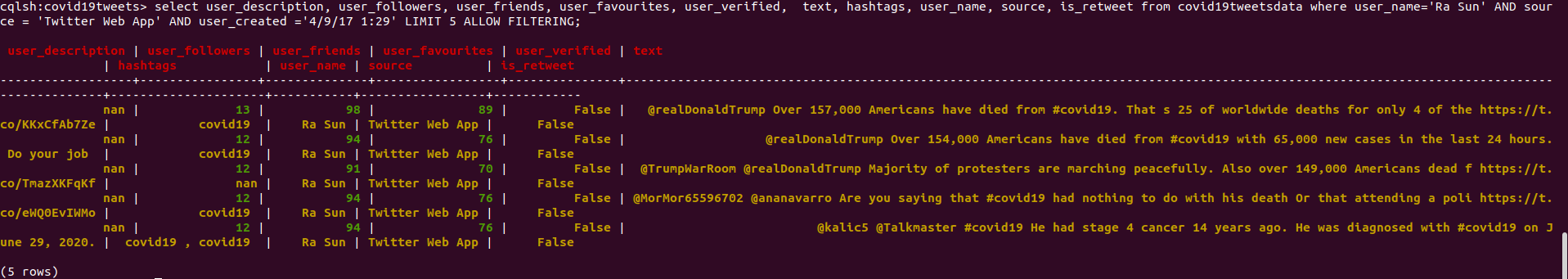
* Select DISTINCT fields form data where place belongs to london

select DISTINCT user\_name, user\_location, user\_created, user\_description, user\_followers, user\_friends, user\_favourites, user\_verified, text, hashtags, source, is\_retweet from covid19tweetsdata where user\_location='London, UK' ALLOW FILTERING;



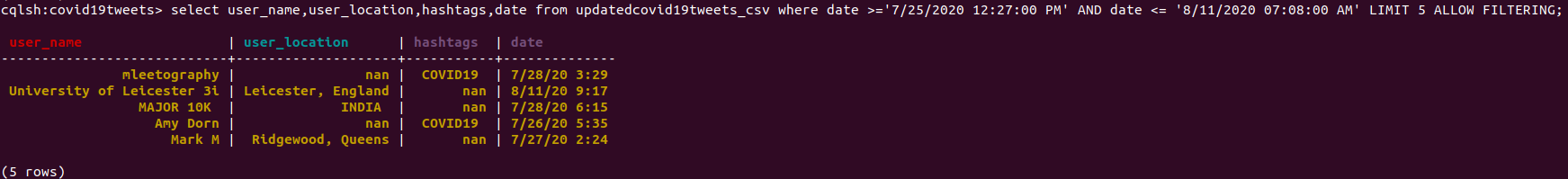
* With twitter web app , display tweets from Ra Sun dated on 4th September?

select user\_description, user\_followers, user\_friends, user\_favourites, user\_verified, text, hashtags, user\_name, source, is\_retweet from covid19tweetsdata where user\_name='Ra Sun' AND source = 'Twitter Web App' AND user\_created ='4/9/17 1:29' ALLOW FILTERING;



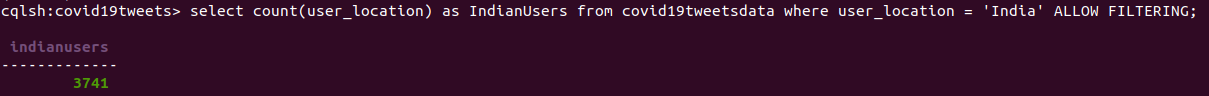
* Display tweets data from 25th July 2020 to 11th Nov 2020

select user\_name,user\_location,hashtags,date from updatedcovid19tweets\_csv where date >='7/25/2020 12:27:00 PM' AND date <= '8/11/2020 07:08:00 AM' ALLOW FILTERING;



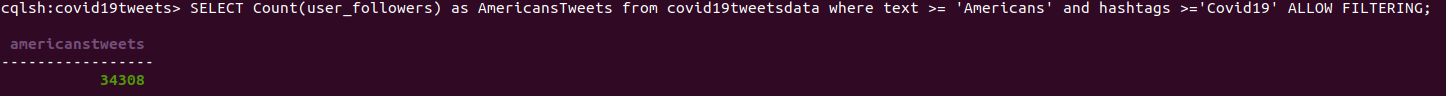
* Retrieve total number of Indians

select count(user\_location) as totalLoc from covid19tweetsdata where user\_location = 'India' ALLOW FILTERING;



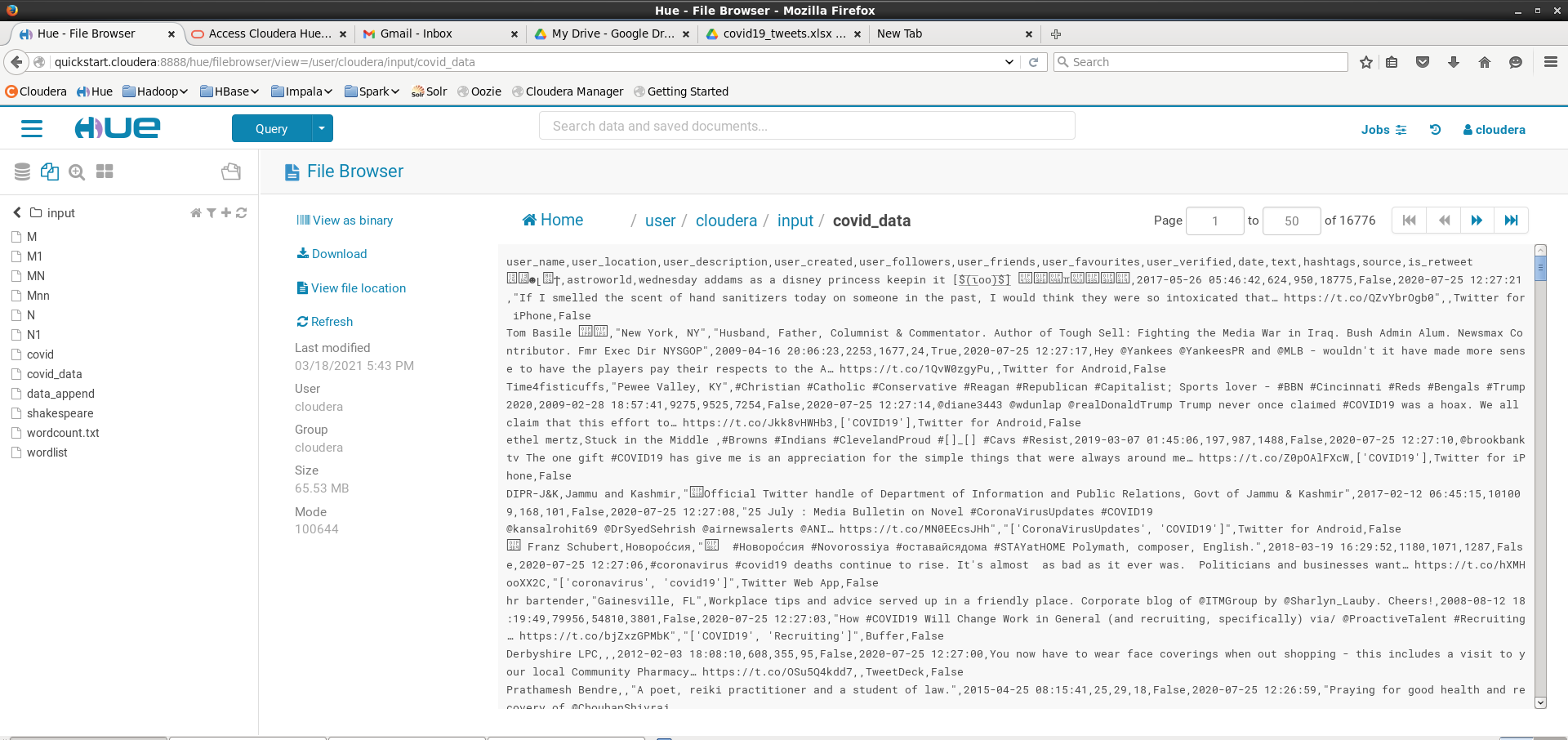
* Americans who twitted about coronavirus

CREATE CUSTOM INDEX comments ON covid19tweets.covid19tweetsdata (text) USING 'org.apache.cassandra.index.sasi.SASIIndex' WITH OPTIONS = {'mode': 'CONTAINS', 'analyzer\_class': 'org.apache.cassandra.index.sasi.analyzer.NonTokenizingAnalyzer', 'case\_sensitive': 'false'};



**Hue**

Apache hue is used for visualization and querying databases. It is an open source which helps look HDFS file system and manage databases like hive, etc. Where you can visualize your queries. We have added our database to the HDFS and visualize it using hue.

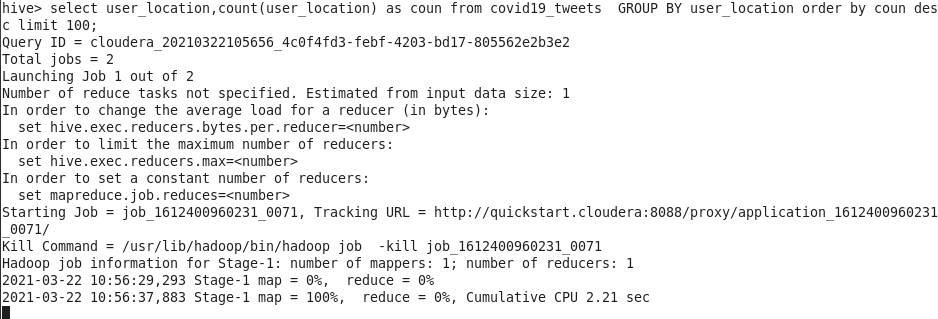


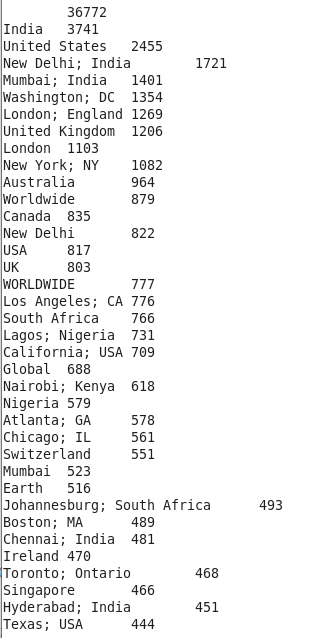
**Hive**

Hive is a data warehouse built on top of Hadoop for structured data. It has SQL-like query and hence is very popular and simple to operate. We have used it to analyze our dataset to find trends in it. We run these queries in hive and visualize it using tableau.

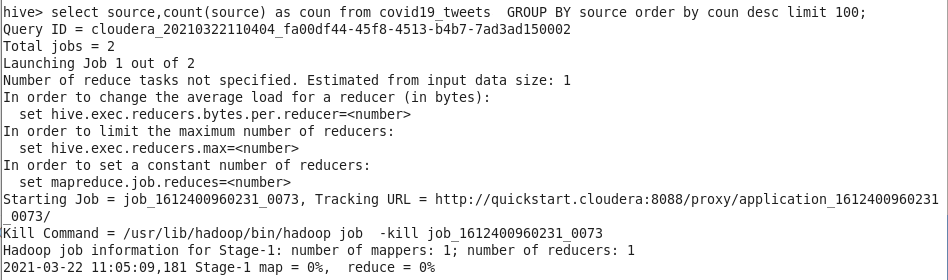
The queries operated on this dataset are:

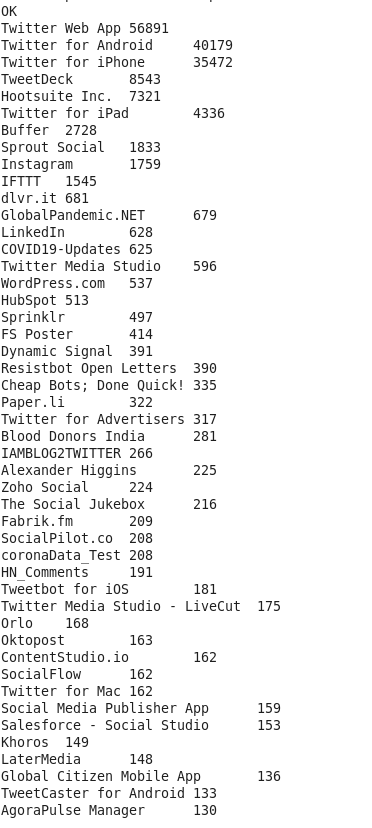
To see from where tweets are coming from majorly.



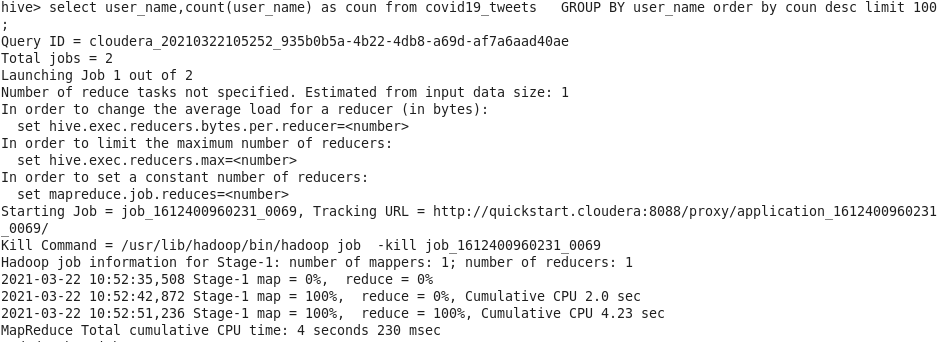


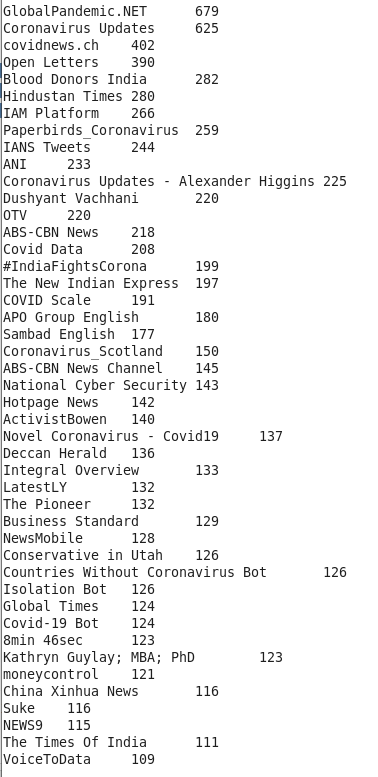
To see from which source are tweets majorly being generated from.



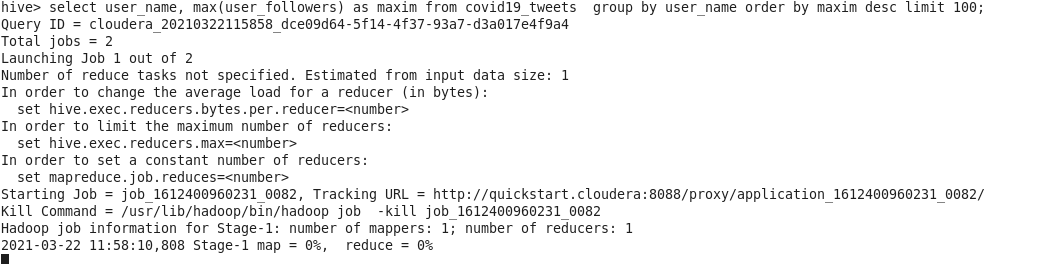


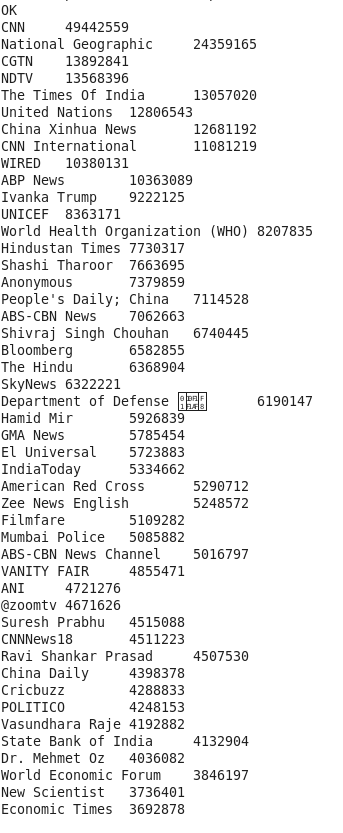
To see from which users have most number of tweets.



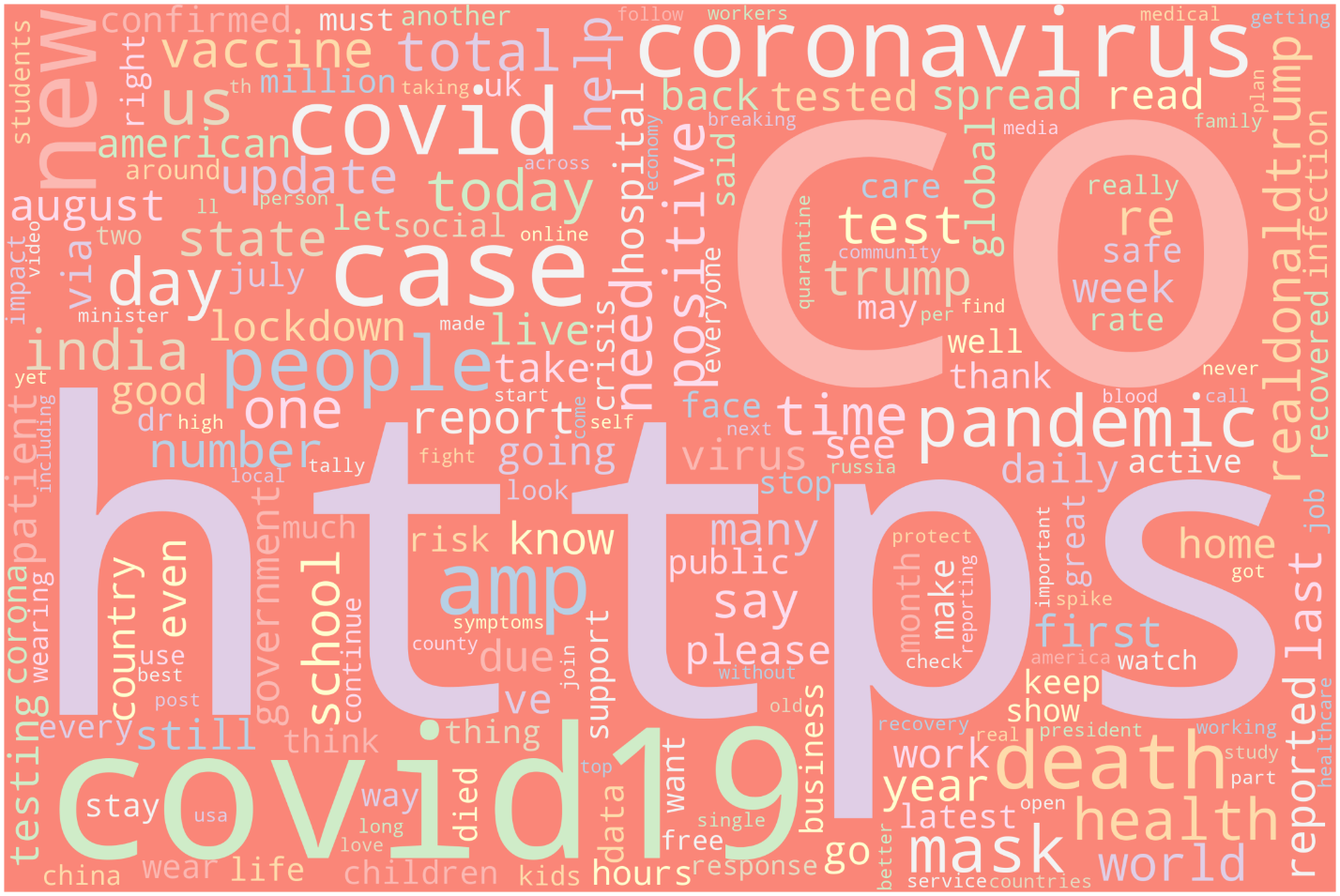


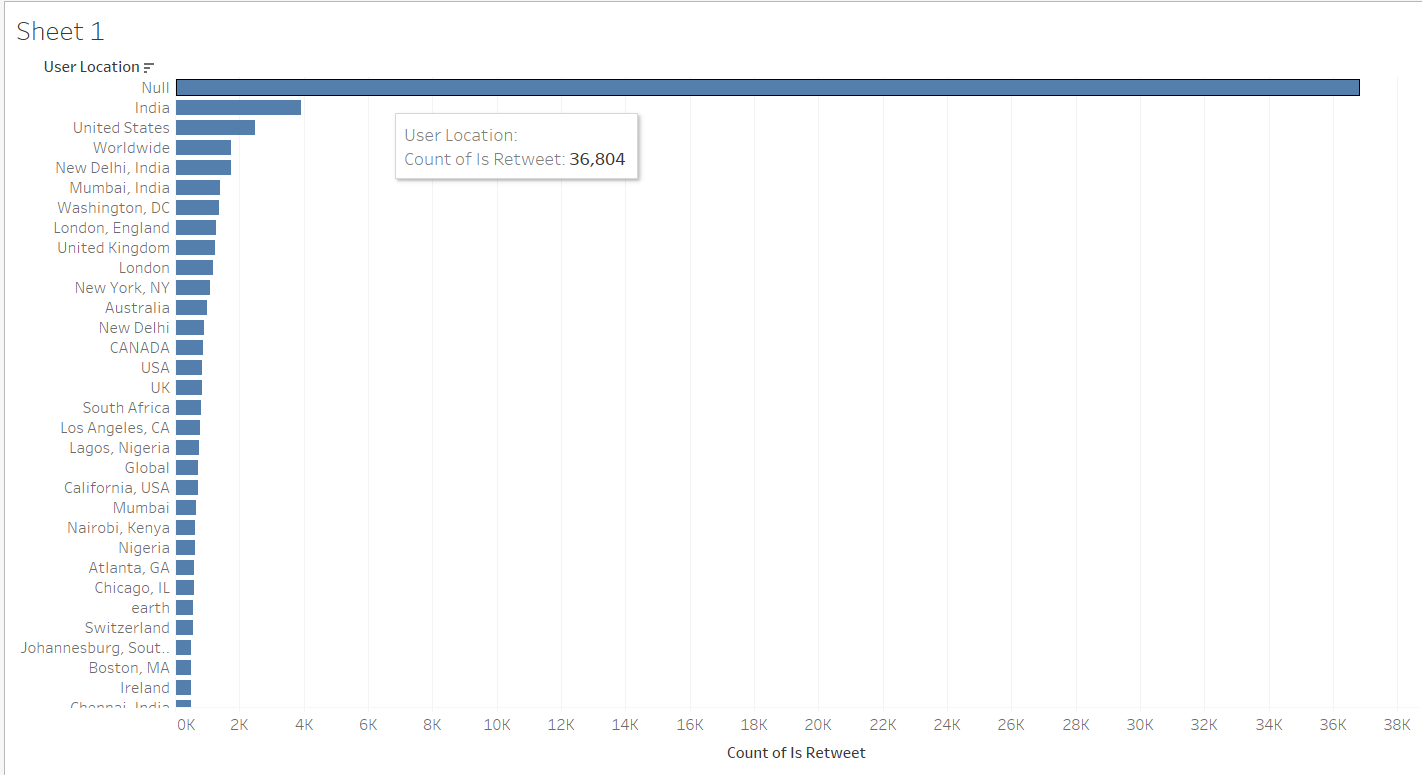
To see which users have most number of followers.

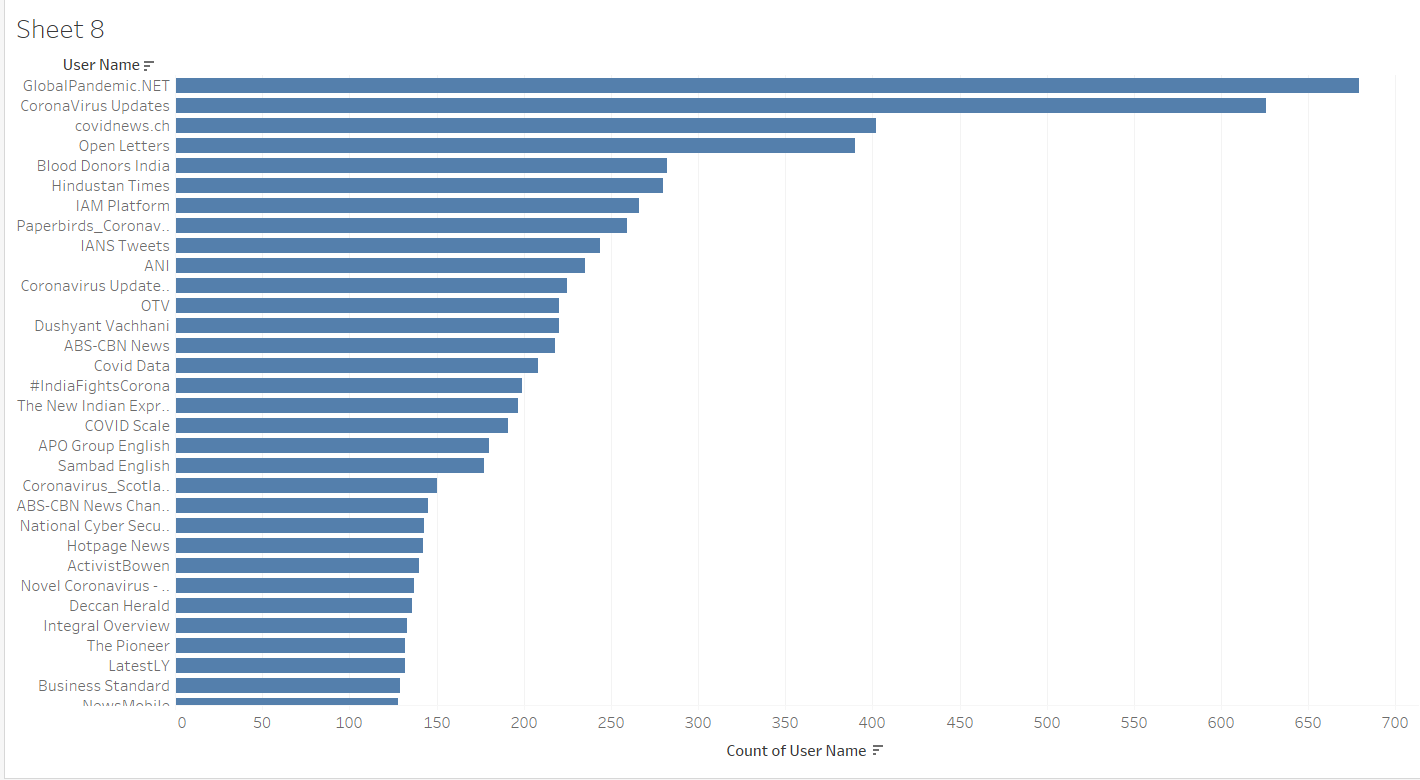
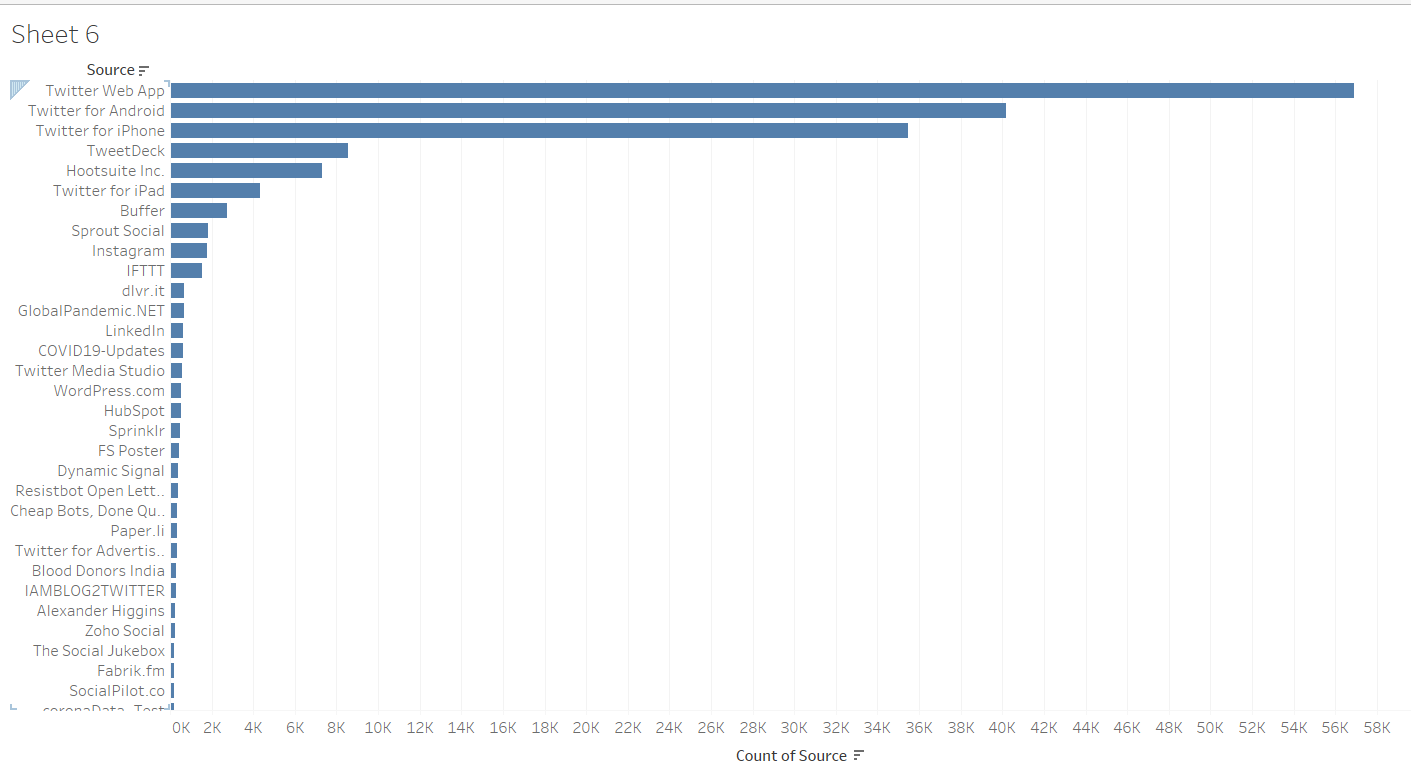


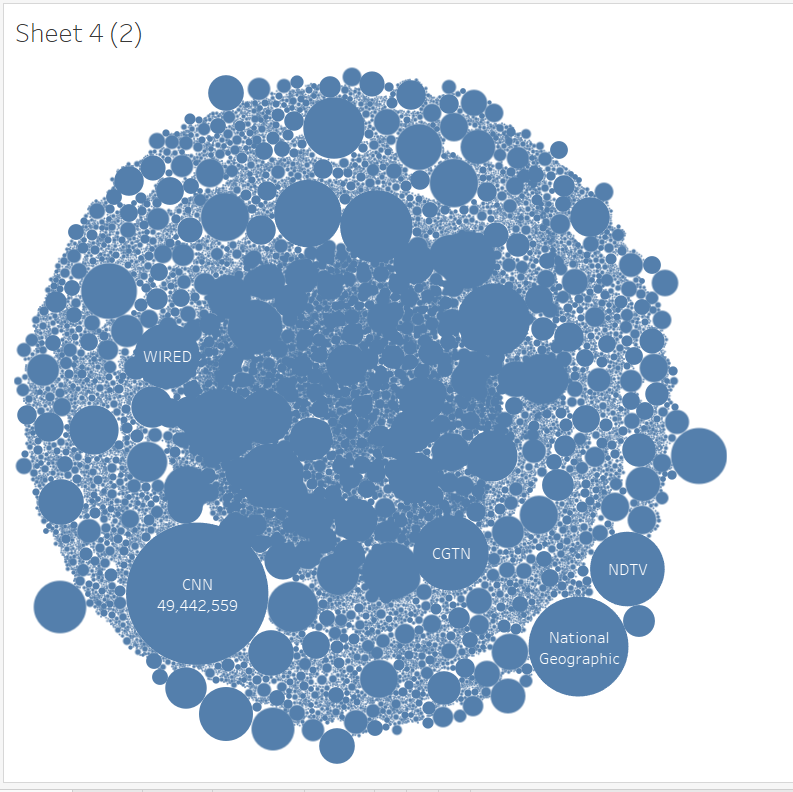


Preliminary Results:









**CHAPTER 1(LIFE)**

**WHO:**

Covid 19 pandemic has affected everyone around the world in one or the other way.

1. Cities, with their dense international setups were particularly affected which includes markets, business, travel, tourism. These were the common entry points for the virus.
2. Rural areas and regions with high numbers of elderly people were affected as well.
3. The pandemic has deeply impacted jobs and businesses. In the developing world,micro-level, small-level and medium-level enterprises are under intense pressure.

**WHAT :**

1. The pandemic has taken a lot of lives.
2. COVID-19 has rapidly affected businesses, world trade, our day to day life and movements.
3. Presently the impacts of COVID-19 in daily life are extensive and have put tremendous pressure on the healthcare sector and affected the social lives of the people.

**WHEN :**

The early traces of the virus were found in November 2019 but it spread rapidly across the world and was declared as a pandemic in March 2020 by the World HealthOrganization and is still going on.

**WHERE :**

200+ Countries and Territories around the world have reported a total of 113M+cases of the coronavirus COVID-19 that originated from Wuhan, China, and a death toll of2.5M+ (as of this writing)

**WHY :**

There is no definite answer to this question. The primary cause of this virus has been linked to the animal “bat” by the mainstream media.

**CHAPTER 2(DATA)**

**WHO:**

The dataset covers all of the twitter users who have used hashtag Covid19 or used any covid or related material to it in their text. It covers people from every corner across the globe.

**WHAT:**

The dataset records all the basic information of the users social media handle, which in this case is twitter and along with that it covers the twitter text that the user is sending which is one of the major analysis points in our dataset.

**WHEN:** The collection of dataset began from november 2019 when Covid 19 started spreading from china to everywhere across the globe. The dataset is collected from twitter.

**WHERE**: It was a local topic trending in China, but as it started spreading across the world, people started tweeting about it and that’s how the dataset was collected.

**WHY:** The dataset was collected to get an understanding of how people are reacting to this pandemic and how they are being affected by it.

**CONCLUSION:**

1. We have used various Hadoop and Cassandra modules to try and understand the twitter data-set and try to analyze it.
2. We have seen which of the words are frequently used in the twitter texts.
3. Also we tried to query the tweets based on various factors like location, etc.
4. Finally we used hive queries to visualize our data-set and understand more about it.

**FUTURE WORK:**

We have analyzed the data-set with various Hadoop modules and will see how spark performs when analyzing the data-set and how it helps us to further our work in pursuit for answers under this topic. We will be using the queries from hive and analyze that data in spark to further understand our dataset.

**CONTRIBUTION OF WORK:**

**Vyoma Desai ​**

Significance, features part, chapter life. Data research and implementing queries on CASSANDRA, d3.Js.

**Affan Asad Charolia**

​Motivation, objectives part. Implementing queries using Hue and Hive and representing graphs in Tableau.

**Ali Alyami**

Project dataset analysis, Conclusion. Implementing dataset wordcount using Hadoop Map Reduce