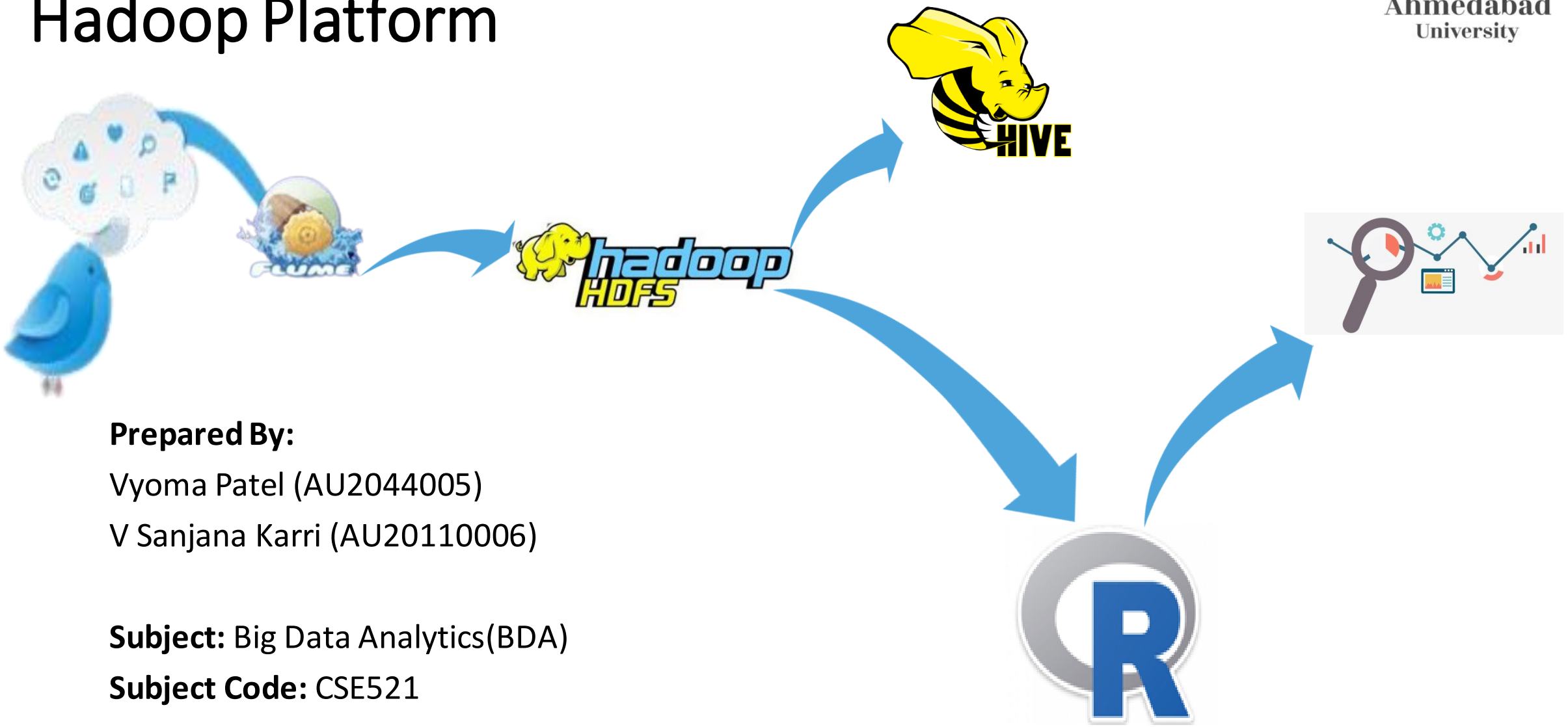
Twitter Data Analysis with R on top of Hadoop Platform

Ahmedabad University



Prepared By:

Vyoma Patel (AU2044005)

V Sanjana Karri (AU20110006)

Subject: Big Data Analytics(BDA)

Subject Code: CSE521

Proposed Methodology

01

Twitter Application

- Creating twitter developer account
- Generating API keys and tokens

02

Installation Part

- Hadoop-3.1.4
- Flume-1.9.0
- Hive-3.1.2

03

Twitter to HDFS

- Create twitter config file
- Start dfs and call twitter agent via flume

04

Hadoop Cluster

- Flume Data generation
- Data in AVRO format

05

Changing Data format

- Calling Avro tool, getting schema
- Converting Avro to JSON

06

Hive

- Starting metastore
- Table Creation and Retrieval

07

Connecting with R

- JSON file generated sent
- Data Preprocessing

08

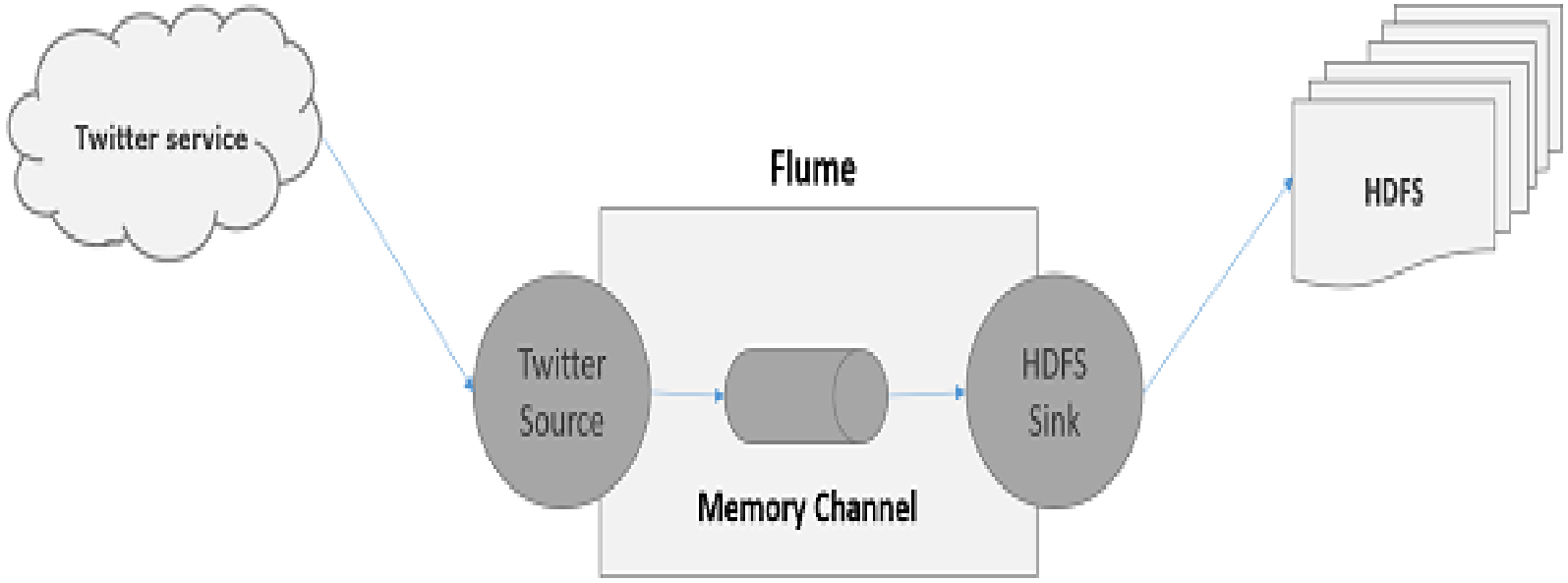
Visualizations

- Using appropriate libraries
- Analysing data with various plots

Basic Definitions

- **Apache Hadoop** - software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models.
- **Apache Flume** – It's a tool for data ingestion in HDFS. To capture streaming data from various web servers to HDFS.
- **Apache Hive** - data warehouse software facilitates reading, writing, and managing large datasets residing in distributed storage using SQL.
- **Apache Avro** - is a data serialization system. Provides compact, fast, binary data format on Hadoop.
- **JSON - J**ava**S**cript **O**bject **N**otation. Lightweight format for storing and transporting data. Widely supported and easily understandable.
- **R** - is a programming language and free software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing.

Process of Tweets stored on Hadoop Cluster



Twitter Config File

The flume agent has 3 components: source, sink, and channel.

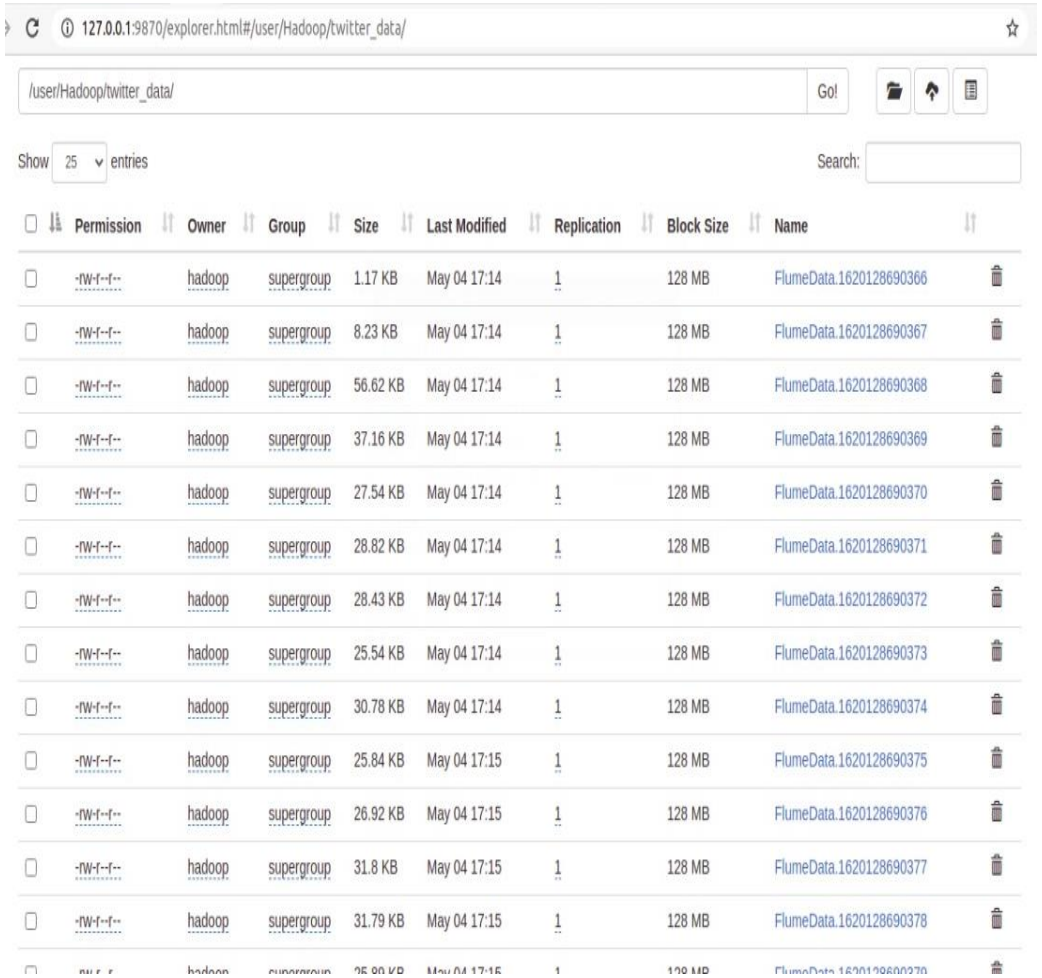
- **Source**: It accepts the data from the incoming streamline and stores the data in the channel.
- **Channel**: In general, the reading speed is faster than the writing speed. Channel acts as the local storage(buffer) or temporary storage between the source of data and persistent data in the HDFS.
- **Sink**: Collects the data from the channel and commits or writes the data in the HDFS permanently.

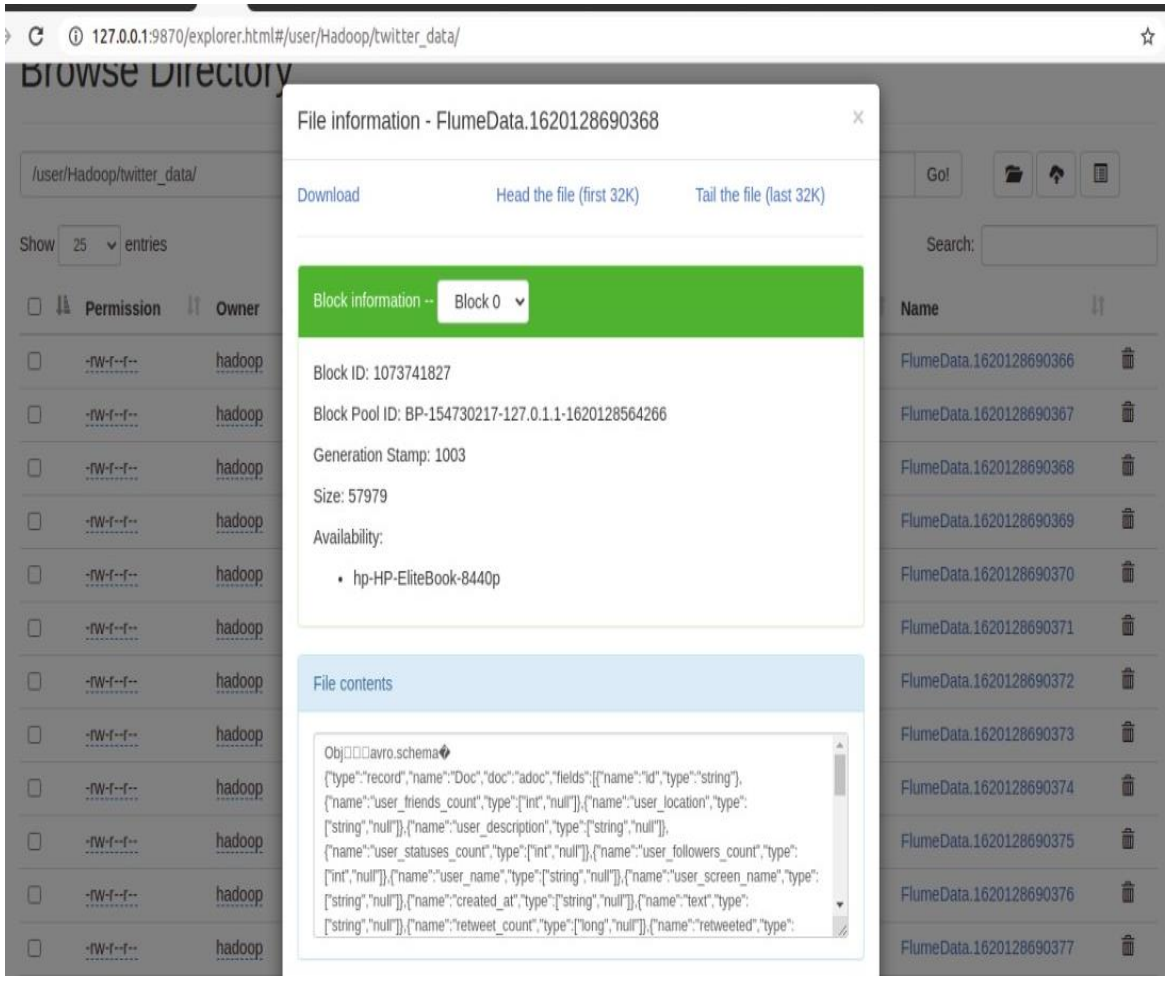
```
#####################################################
# Twitter agent for collecting Twitter data to HDFS.
#####################################################
TwitterAgent.sources = Twitter
TwitterAgent.channels = MemChannel
TwitterAgent.sinks = HDFS
#####################################################
# Describing and configuring the sources
#####################################################
TwitterAgent.sources.Twitter.type = org.apache.flume.source.twitter.TwitterSource
TwitterAgent.sources.Twitter.channels = MemChannel
TwitterAgent.sources.Twitter.consumerKey = pKyqnqrQSVcPIJFLBeyumm8OF
TwitterAgent.sources.Twitter.consumerSecret = UwSIdQnKtAle69rGwlfO0AfEWxzHgIcG4QDhsrzo2VyXRRxAXX
TwitterAgent.sources.Twitter.accessToken = 1320030990685777921-3U05qalVJ7qNQQpv3gvgEHDjH5fb5R
TwitterAgent.sources.Twitter.accessTokenSecret = 4g7NDwt4v5tRBspLt0Sm6QfADfPlxZOtPi86pox3eZLrS
TwitterAgent.sources.Twitter.keywords = covid-19,corona,aid,virus,India,COVID,help,sos,shortage,crisis,shutdown,lockdown,lack
#####################################################
# Twitter configuring  HDFS sink
#####################################################
TwitterAgent.sinks.HDFS.channel = MemChannel
TwitterAgent.sinks.HDFS.type = hdfs
TwitterAgent.sinks.HDFS.hdfs.path = hdfs://localhost:9000/user/Hadoop/twitter_data/
TwitterAgent.sinks.HDFS.hdfs.fileType = DataStream
# Test the hdfs.writeFormat to json
TwitterAgent.sinks.HDFS.
TwitterAgent.sinks.HDFS.hdfs.batchSize = 1000
#####################################################
# Twitter Channel
#####################################################
TwitterAgent.channels.MemChannel.type = memory
TwitterAgent.channels.MemChannel.capacity = 1000
TwitterAgent.channels.MemChannel.transactionCapacity = 1000
#####################################################
# Binding the Source and the Sink to the Channel
#####################################################

TwitterAgent.sources.Twitter.channels = MemChannel
TwitterAgent.sinks.HDFS.channels = MemChannel
```

Starting Hadoop and calling flume agent

```
2021-05-04 09:39:17,565 INFO util.GSet: VM type       = 64-bit
2021-05-04 09:39:17,565 INFO util.GSet: 0.25% max memory 828.5 MB = 2.1 MB
2021-05-04 09:39:17,565 INFO util.GSet: capacity      = 2^18 = 262144 entries
2021-05-04 09:39:17,581 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.window.num.buckets = 10
2021-05-04 09:39:17,581 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.num.users = 10
2021-05-04 09:39:17,582 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.windows.minutes = 1,5,25
2021-05-04 09:39:17,587 INFO namenode.FSNamesystem: Retry cache on namenode is enabled
2021-05-04 09:39:17,588 INFO namenode.FSNamesystem: Retry cache will use 0.03 of total heap and retry cache entry expiry time is 600000 millis
2021-05-04 09:39:17,592 INFO util.GSet: Computing capacity for map NameNodeRetryCache
2021-05-04 09:39:17,592 INFO util.GSet: VM type       = 64-bit
2021-05-04 09:39:17,592 INFO util.GSet: 0.029999999329447746% max memory 828.5 MB = 254.5 KB
2021-05-04 09:39:17,592 INFO util.GSet: capacity      = 2^15 = 32768 entries
2021-05-04 09:39:17,742 INFO namenode.FSImage: Allocated new BlockPoolId: BP-812740381-127.0.1.1-1620101357664
2021-05-04 09:39:17,914 INFO common.Storage: Storage directory /home/hadoop/hdfs/namenode has been successfully formatted.
2021-05-04 09:39:17,968 INFO namenode.FSImageFormatProtobuf: Saving image file /home/hadoop/hdfs/namenode/current/fsimage.ckpt_000000000000000
0000 using no compression
2021-05-04 09:39:18,196 INFO namenode.FSImageFormatProtobuf: Image file /home/hadoop/hdfs/namenode/current/fsimage.ckpt_0000000000000000000 of
 size 390 bytes saved in 0 seconds .
2021-05-04 09:39:18,268 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0
2021-05-04 09:39:18,279 INFO namenode.FSImage: FSImageSaver clean checkpoint: txid = 0 when meet shutdown.
2021-05-04 09:39:18,281 INFO namenode.NameNode: SHUTDOWN_MSG:
/************************************************************
SHUTDOWN_MSG: Shutting down NameNode at hp-HP-EliteBook-8440p/127.0.1.1
************************************************************/
hadoop@hp-HP-EliteBook-8440p:~/hadoop-3.1.4/bin$ cd $HADOOP_HOME/sbin/
hadoop@hp-HP-EliteBook-8440p:~/hadoop-3.1.4/sbin$ start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [hp-HP-EliteBook-8440p]
hadoop@hp-HP-EliteBook-8440p:~/hadoop-3.1.4/sbin$ jps
5186 NameNode
5382 DataNode
5753 Jps
5563 SecondaryNameNode
hadoop@hp-HP-EliteBook-8440p:~/hadoop-3.1.4/sbin$ cd
hadoop@hp-HP-EliteBook-8440p:~$ cd $FLUME_HOME
hadoop@hp-HP-EliteBook-8440p:/home/hp/apache-flume-1.9.0-bin$ bin/flume-ng agent --conf ./conf/ -f conf/new_twitter.conf -Dflume.root.logger=D
EBUG,console -n TwitterAgent
```

Tweets stored in Avro format using Flume on Hadoop Cluster





Storing Data generated using flume agent in a folder and Avro initialization

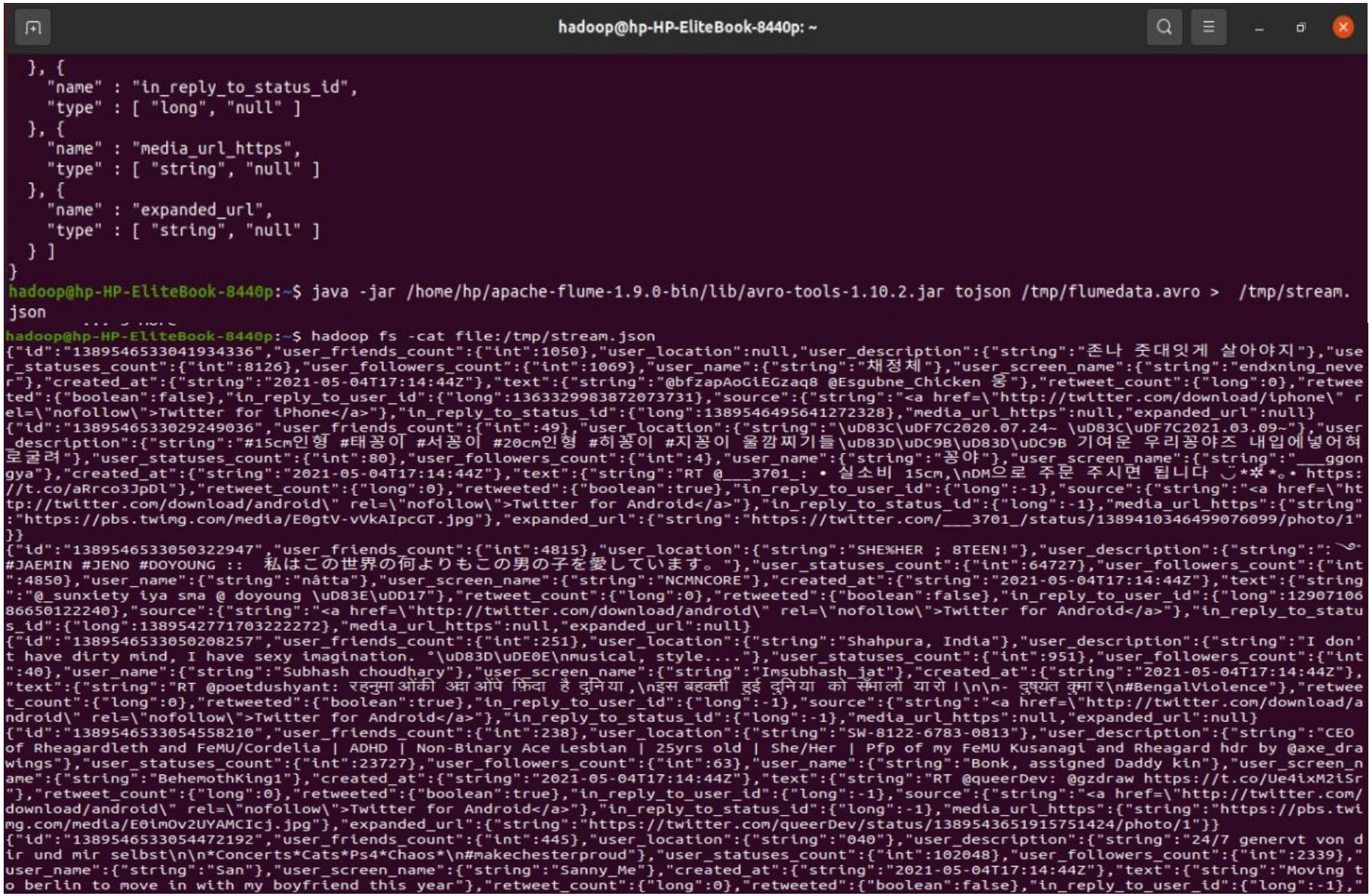
```
hadoop@hp-HP-EliteBook-8440p:~$ hdfs dfs -copyToLocal /user/Hadoop/twitter_data/FlumeData.1620128690368 /tmp
hadoop@hp-HP-EliteBook-8440p:~$ java -jar /home/hp/apache-flume-1.9.0-bin/lib/avro-tools-1.10.2.jar
Version 1.10.2 of Apache Avro
Copyright 2010-2015 The Apache Software Foundation

This product includes software developed at
The Apache Software Foundation (https://www.apache.org/).
---------------
Available tools:
    canonical  Converts an Avro Schema to its canonical form
          cat  Extracts samples from files
      compile  Generates Java code for the given schema.
       concat  Concatenates avro files without re-compressing.
        count  Counts the records in avro files or folders
  fingerprint  Returns the fingerprint for the schemas.
   fragtojson  Renders a binary-encoded Avro datum as JSON.
     fromjson  Reads JSON records and writes an Avro data file.
     fromtext  Imports a text file into an avro data file.
      getmeta  Prints out the metadata of an Avro data file.
    getschema  Prints out schema of an Avro data file.
          idl  Generates a JSON schema from an Avro IDL file
 idl2schemata  Extract JSON schemata of the types from an Avro IDL file
       induce  Induce schema/protocol from Java class/interface via reflection.
   jsontofrag  Renders a JSON-encoded Avro datum as binary.
       random  Creates a file with randomly generated instances of a schema.
      recodec  Alters the codec of a data file.
       repair  Recovers data from a corrupt Avro Data file
  rpcprotocol  Output the protocol of a RPC service
   rpcreceive  Opens an RPC Server and listens for one message.
      rpcsend  Sends a single RPC message.
       tether  Run a tethered mapreduce job.
       tojson  Dumps an Avro data file as JSON, record per line or pretty.
       totext  Converts an Avro data file to a text file.
     totrevni  Converts an Avro data file to a Trevni file.
  trevni_meta  Dumps a Trevni file's metadata as JSON.
trevni_random  Create a Trevni file filled with random instances of a schema.
trevni_tojson  Dumps a Trevni file as JSON.
hadoop@hp-HP-EliteBook-8440p:~$
```

Getting Schema of Data Generated in Avro Format

```
hadoop@hp-HP-EliteBook-8440p:~$ java -jar /home/hp/apache-flume-1.9.0-bin/lib/avro-tools-1.10.2.jar getschema /tmp/flumedata.avro
21/05/04 17:38:36 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where appli
cable
{
  "type" : "record",
  "name" : "Doc",
  "doc" : "adoc",
  "fields" : [ {
    "name" : "id",
    "type" : "string"
  }, {
    "name" : "user_friends_count",
    "type" : [ "int", "null" ]
  }, {
    "name" : "user_location",
    "type" : [ "string", "null" ]
  }, {
    "name" : "user_description",
    "type" : [ "string", "null" ]
  }, {
    "name" : "user_statuses_count",
    "type" : [ "int", "null" ]
  }, {
    "name" : "user_followers_count",
    "type" : [ "int", "null" ]
  }, {
    "name" : "user_name",
    "type" : [ "string", "null" ]
  }, {
    "name" : "user_screen_name",
    "type" : [ "string", "null" ]
  }, {
    "name" : "created_at",
    "type" : [ "string", "null" ]
  }, {
    "name" : "text",
    "type" : [ "string", "null" ]
  }, {
```

```
 "name" : "user_followers_count",
 "type" : [ "int", "null" ]
, {
 "name" : "user_name",
 "type" : [ "string", "null" ]
, {
 "name" : "user_screen_name",
 "type" : [ "string", "null" ]
, {
 "name" : "created_at",
 "type" : [ "string", "null" ]
, {
 "name" : "text",
 "type" : [ "string", "null" ]
, {
 "name" : "retweet_count",
 "type" : [ "long", "null" ]
, {
 "name" : "retweeted",
 "type" : [ "boolean", "null" ]
, {
 "name" : "in_reply_to_user_id",
 "type" : [ "long", "null" ]
, {
 "name" : "source",
 "type" : [ "string", "null" ]
, {
 "name" : "in_reply_to_status_id",
 "type" : [ "long", "null" ]
, {
 "name" : "media_url_https",
 "type" : [ "string", "null" ]
, {
 "name" : "expanded_url",
 "type" : [ "string", "null" ]
]
```

Converting Avro to JSON Data format



rHadoop

```
> install.packages("/home/hp/rHadoopClient_0.2.tar.gz")
Installing package into '/home/hp/R/x86_64-pc-linux-gnu-library/4.0'
(as 'lib' is unspecified)
inferring 'repos = NULL' from 'pkgs'
* installing *source* package 'rHadoopClient' ...
** package 'rHadoopClient' successfully unpacked and MD5 sums checked
** using staged installation
** R
** byte-compile and prepare package for lazy loading
** help
*** installing help indices
** building package indices
** testing if installed package can be loaded from temporary location
** testing if installed package can be loaded from final location
** testing if installed package keeps a record of temporary installation path
* DONE (rHadoopClient)
```

Avro data format read

```
> rHadoopClient::read.hdfs("/user/Hadoop/twitter_data/FlumeData.1620203595094")
                                                                        V1
1                                          Obj\001\002\026avro.schema\xe4
2                                                             {type:record
3                                        His face (especially eyes): 🧸🌸🥺✨🥰
4                                                  https://t.co/yYFhU7wnZx
5                                                           証明されちゃった
6                          鍵50付ける暇があったら勉強したらw https://t.co/9erkh36yRs
7
8                It is reported that 90 percent of Covid beds have been occupied....
9                                            EP.  5 https://t.co/7ttTtjf4S5
10                                                        いま行く　待っててね
11                                                                   India
12                              #최현석 #CHOIHYUNSUK https://t.co/KBRiJ8iGdF
13                                                                   cc bg
14 4/23より、ヘザー店舗とWEBストア[.st]にてヘザー商品1万円以上ご購入で、リップティントとポーチをプレゼント🎁
15                      リンクの診断テストでおすすめスタイル&リップをCHECKして♡…
16                                                          真ん中の家おるー！！
17                                                     원우 도겸 우지 양도 받아용
18                                                            항상 즉입가능
19                                                                    \002
20
21                                                 https://t.co/UsjzIhmtCu
22                              Islamic nigeriA Killed 3.5M+ Christian Biafrans
23                                                     #BiafraReferendum
24                                                  INDEPENDENCE FOR BIAFRA
25                                                   one nigeriA is a FRAUD
26                                                     #UnityIsNotByForce
27                           #Referendum  =! #War… https://t.co/IiMoJCva7I
28                                     On YouTube: https://t.co/A2tM3OKFCg
29                                #FridaysForFuture #Berlin👑 #ClimateCh…
30                                                        噛んでるのはスルーして
31                                                          全編はこちら〜〜〜♪
32                                                              #RuNaライブ
33                                                              #RuNaカット
34                       https://t.co/Cp8tmr2fSC https://t.co/xOCM9bpE4f
```

JSON Format data read

```
> rHadoopClient::read.hdfs("file:/tmp/stream.json")
                         V1                             V2
1  {id:1389860711581552640   user_friends_count:{int:150}
2  {id:1389860711594176514   user_friends_count:{int:143}
3  {id:1389860711589941253     user_friends_count:{int:1}
4  {id:1389860711602417667    user_friends_count:{int:57}
5  {id:1389860711602421761   user_friends_count:{int:134}
6  {id:1389860711581421573   user_friends_count:{int:392}
7  {id:1389860711577227265   user_friends_count:{int:171}
8  {id:1389860711577366529   user_friends_count:{int:175}
9  {id:1389860711589978116  user_friends_count:{int:1863}
10 {id:1389860711564595203   user_friends_count:{int:974}
11 {id:1389860711589826561   user_friends_count:{int:173}
12 {id:1389860715771531269    user_friends_count:{int:31}
13 {id:1389860715779936261   user_friends_count:{int:163}
14 {id:1389860715767373825    user_friends_count:{int:11}
15 {id:1389860715788333060   user_friends_count:{int:862}
16 {id:1389860715771559938   user_friends_count:{int:152}
17 {id:1389860715758981120   user_friends_count:{int:118}
18 {id:1389860715758964744   user_friends_count:{int:899}
19 {id:1389860715775680513  user_friends_count:{int:2814}
20 {id:1389860715792531457   user_friends_count:{int:377}
21 {id:1389860715796729856  user_friends_count:{int:2691}
22 {id:1389860715767414789   user_friends_count:{int:183}
23 {id:1389860715792670720  user_friends_count:{int:5001}
24 {id:1389860715796725760   user_friends_count:{int:189}
25 {id:1389860715775750146    user_friends_count:{int:25}
```

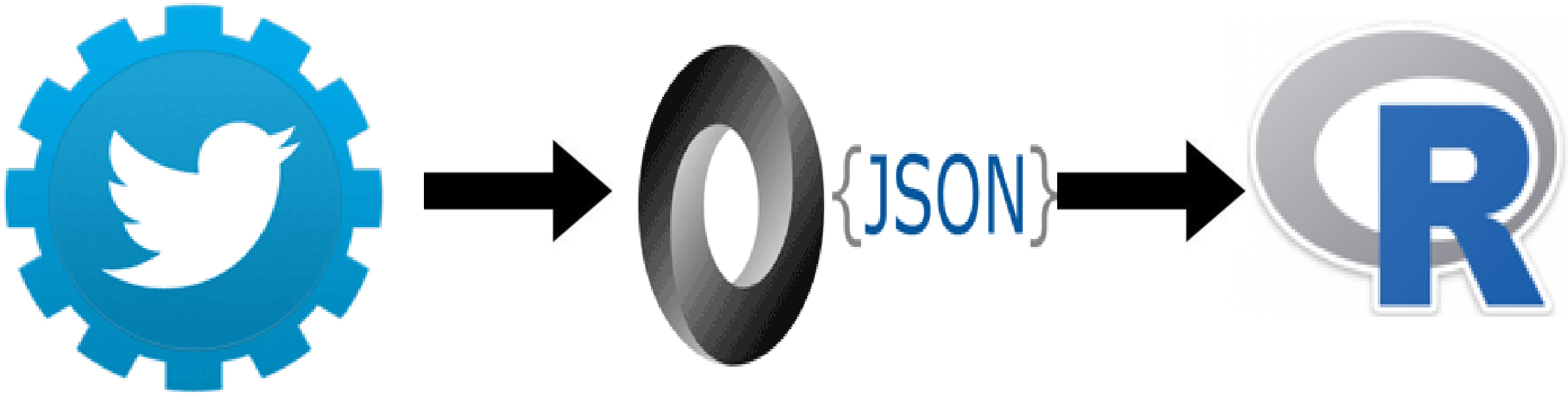
Starting Hive

```
hadoop@hp-HP-EliteBook-8440p:~$ mv metastore_db metastore_db.tmp
hadoop@hp-HP-EliteBook-8440p:~$ schematool -dbType derby -initSchema
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/hadoop/apache-hive-3.1.2-bin/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/hadoop/hadoop-3.1.4/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinde
r.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Metastore connection URL:        jdbc:derby:;databaseName=metastore_db;create=true
Metastore Connection Driver :    org.apache.derby.jdbc.EmbeddedDriver
Metastore connection User:       APP
Starting metastore schema initialization to 3.1.0
Initialization script hive-schema-3.1.0.derby.sql









Initialization script completed
schemaTool completed
hadoop@hp-HP-EliteBook-8440p:~$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/hadoop/apache-hive-3.1.2-bin/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/hadoop/hadoop-3.1.4/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinde
r.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Hive Session ID = 7f392ec7-ec03-420b-a521-7272e611e7a7

Logging initialized using configuration in jar:file:/home/hadoop/apache-hive-3.1.2-bin/lib/hive-common-3.1.2.jar!/hive-log4j2.properties Async
: true
Hive Session ID = f17f5181-056a-428f-abc9-534354122d65
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, t
ez) or using Hive 1.X releases.
hive> CREATE EXTERNAL TABLE TwitterData
    > ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.avro.AvroSerDe'
    > WITH SERDEPROPERTIES ('avro.schema.literal'='
    > {
    >
    >   "type" : "record",
    >
    >   "name" : "Doc",
    >
    >   "doc" : "adoc",
    >
    >   "fields" : [ {
    >
    >     "name" : "id",
```

Creating and Showing Table using Twitter(Avro) Data using Hive

```
    >   }, {
    >     "name" : "media_url_https",
    >     "type" : [ "string", "null" ]
    >   }, {
    >     "name" : "expanded_url",
    >     "type" : [ "string", "null" ]
    >   } ]
    > }
    >
    > ')
    > STORED AS
    > INPUTFORMAT 'org.apache.hadoop.hive.ql.io.avro.AvroContainerInputFormat'
    > OUTPUTFORMAT 'org.apache.hadoop.hive.ql.io.avro.AvroContainerOutputFormat'
    > LOCATION '/user/Hadoop/twitter_data'
    > ;
OK
Time taken: 2.363 seconds
hive> DESCRIBE TwitterData;
OK
id                      string
user_friends_count      int
user_location           string
user_description        string
user_statuses_count     int
user_followers_count    int
user_name               string
user_screen_name        string
created_at              string
text                    string
retweet_count           bigint
retweeted               boolean
in_reply_to_user_id     bigint
source                  string
in_reply_to_status_id   bigint
media_url_https         string
expanded_url            string
Time taken: 1.565 seconds, Fetched: 17 row(s)
hive>
```

Reconnecting with R

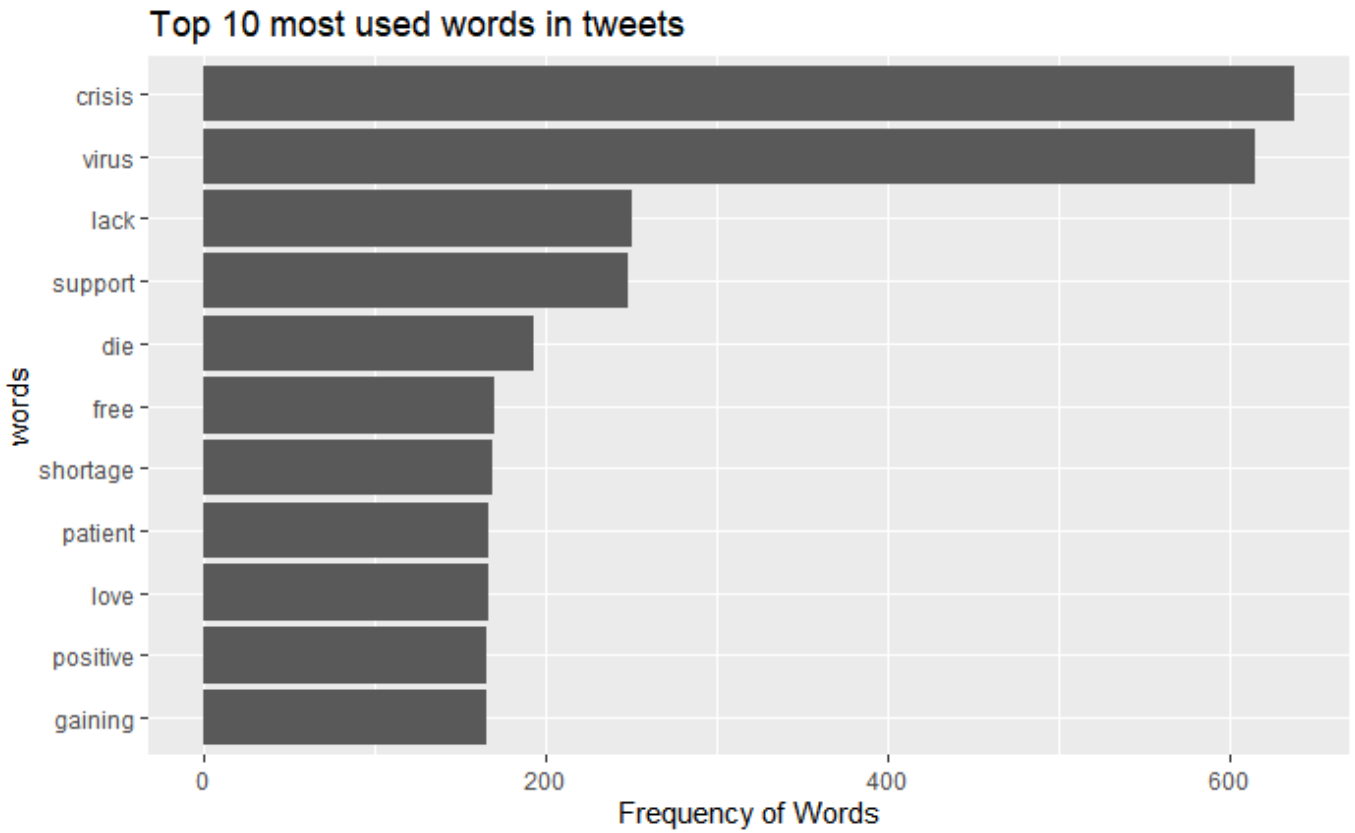


Cleaned JSON file with tweets in English Language

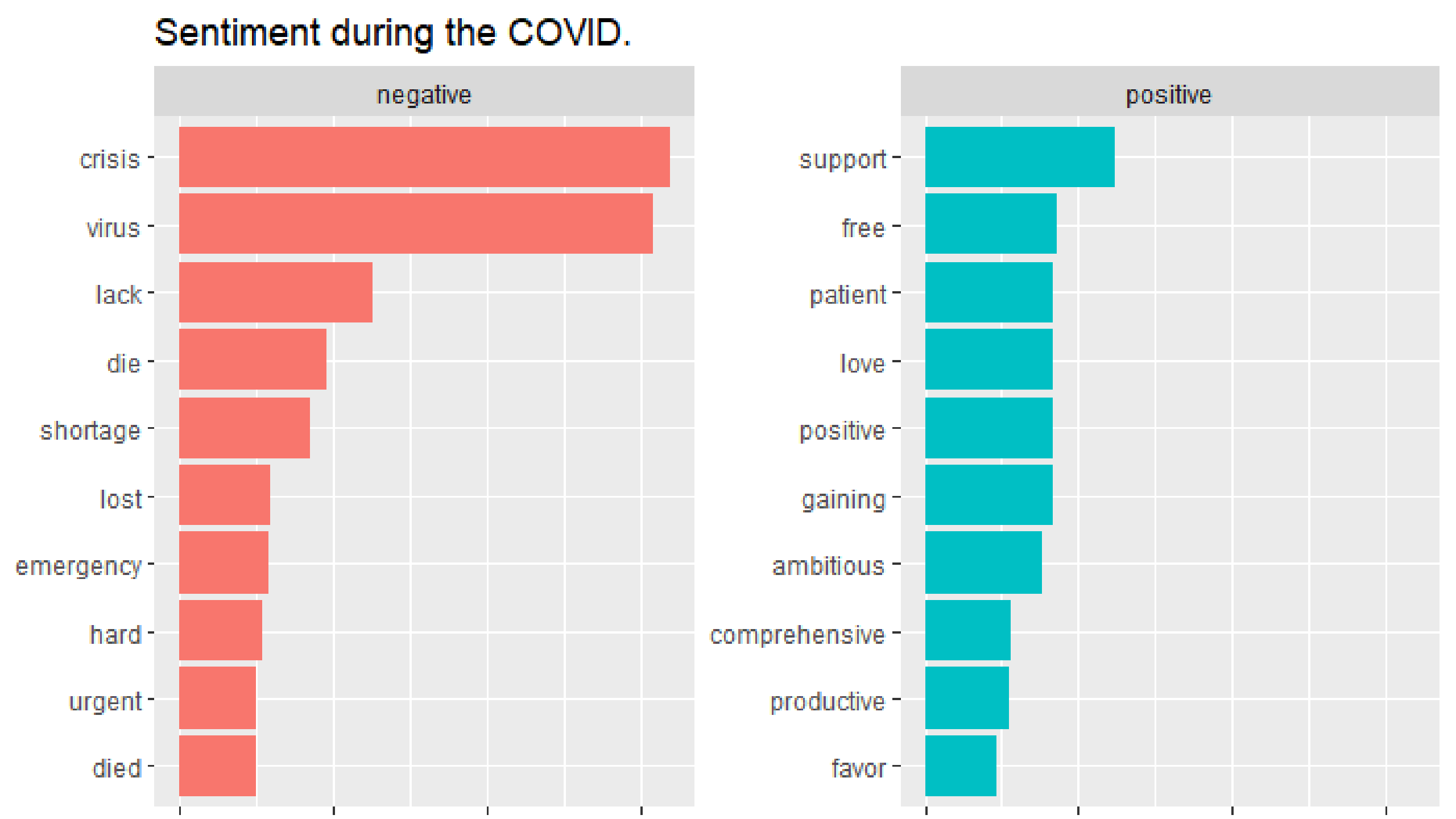

```
{"created_at":"Tue May 04 09:43:25 +0000 2021","id":1389516003801976836,"id_str":"1389516003801976836","text":"RT @Sriram37915703: Please provide compensatory attempt to all CSE last
{"created_at":"Tue May 04 09:43:25 +0000 2021","id":1389516004020166658,"id_str":"1389516004020166658","text":"RT @pinky_collectio: Name: Okigbo chidera Vera. \nAge: 15years \nLast s
{"created_at":"Tue May 04 09:43:25 +0000 2021","id":1389516004103950336,"id_str":"1389516004103950336","text":"RT @sangeetabasa: During times like pandemic, are the rules not same fo
{"created_at":"Tue May 04 09:43:25 +0000 2021","id":1389516003986509826,"id_str":"1389516003986509826","text":"#Indien - Fatale Massenveranstaltungen @msargentini @Corriere https:\/\
{"created_at":"Tue May 04 09:43:26 +0000 2021","id":1389516004187938816,"id_str":"1389516004187938816","text":"@manojpandey66 No. Although a protest against a protest for not followi
{"created_at":"Tue May 04 09:43:26 +0000 2021","id":1389516004313780224,"id_str":"1389516004313780224","text":"RT @YeoBeenmyQueen: THEY ARE SO CUTE \nHELP!!!!\n#VincenzoEp20 https:\/
{"created_at":"Tue May 04 09:43:26 +0000 2021","id":1389516004179517440,"id_str":"1389516004179517440","text":"RT @RICHA_LAKHERA: Engrave names of each &amp;amp; every one we lost on the
{"created_at":"Tue May 04 09:43:26 +0000 2021","id":1389516004380786689,"id_str":"1389516004380786689","text":"FUERZA COLOMBIA \ud83c\udde8\ud83c\uddf4 \u270a\u270a\u270a\n @ONU_es\n
{"created_at":"Tue May 04 09:43:26 +0000 2021","id":1389516004250771456,"id_str":"1389516004250771456","text":"Dear @MinPres sorry again. I hope you could translate it but it's about
{"created_at":"Tue May 04 09:43:26 +0000 2021","id":1389516004519288832,"id_str":"1389516004519288832","text":"RT @ReaganGomez: Why this picture\ud83d\ude02","source":"\u003ca href=\
{"created_at":"Tue May 04 09:43:26 +0000 2021","id":1389516004657713153,"id_str":"1389516004657713153","text":"RT @AnaCabrera: NEW: Los Angeles reports zero Covid deaths for second s
{"created_at":"Tue May 04 09:43:26 +0000 2021","id":1389516004502511617,"id_str":"1389516004502511617","text":"@chrislittlewoo8 @BeamLevels @wonderw12494002 India,Brazil and many cou
{"created_at":"Tue May 04 09:43:26 +0000 2021","id":1389516004661899265,"id_str":"1389516004661899265","text":"RT @Benarasiyaa: This @CNN report from UP's Meerut is glimpse of the ha
{"created_at":"Tue May 04 09:43:26 +0000 2021","id":1389516004842262531,"id_str":"1389516004842262531","text":"RT @Giyourdad: \u0e19\u0e35\u0e48\u0e01\u0e47\u0e16\u0e32\u0e21\u0e40\u
{"created_at":"Tue May 04 09:43:26 +0000 2021","id":1389516005014130694,"id_str":"1389516005014130694","text":"RT @ideal_granada: \ud83d\udea8 #\u00daltimaHora Largas colas en los ac
{"created_at":"Tue May 04 09:43:26 +0000 2021","id":1389516004552757248,"id_str":"1389516004552757248","text":"I'm getting tired of these Liverpool transfer rumours everytime yet we
{"created_at":"Tue May 04 09:43:26 +0000 2021","id":1389516005089677314,"id_str":"1389516005089677314","text":"RT @brajeshksingh: \u0915\u094b\u0932\u0915\u093e\u0924\u093e \u092e\u0
{"created at":"Tue May 04 09:43:26 +0000 2021","id":1389516004968042497,"id_str":"1389516004968042497","text":"RT @ThalaSudharshan: Rudhramadevi \ud83d\udd25\ud83d\udd25\n\nEarned 50
```

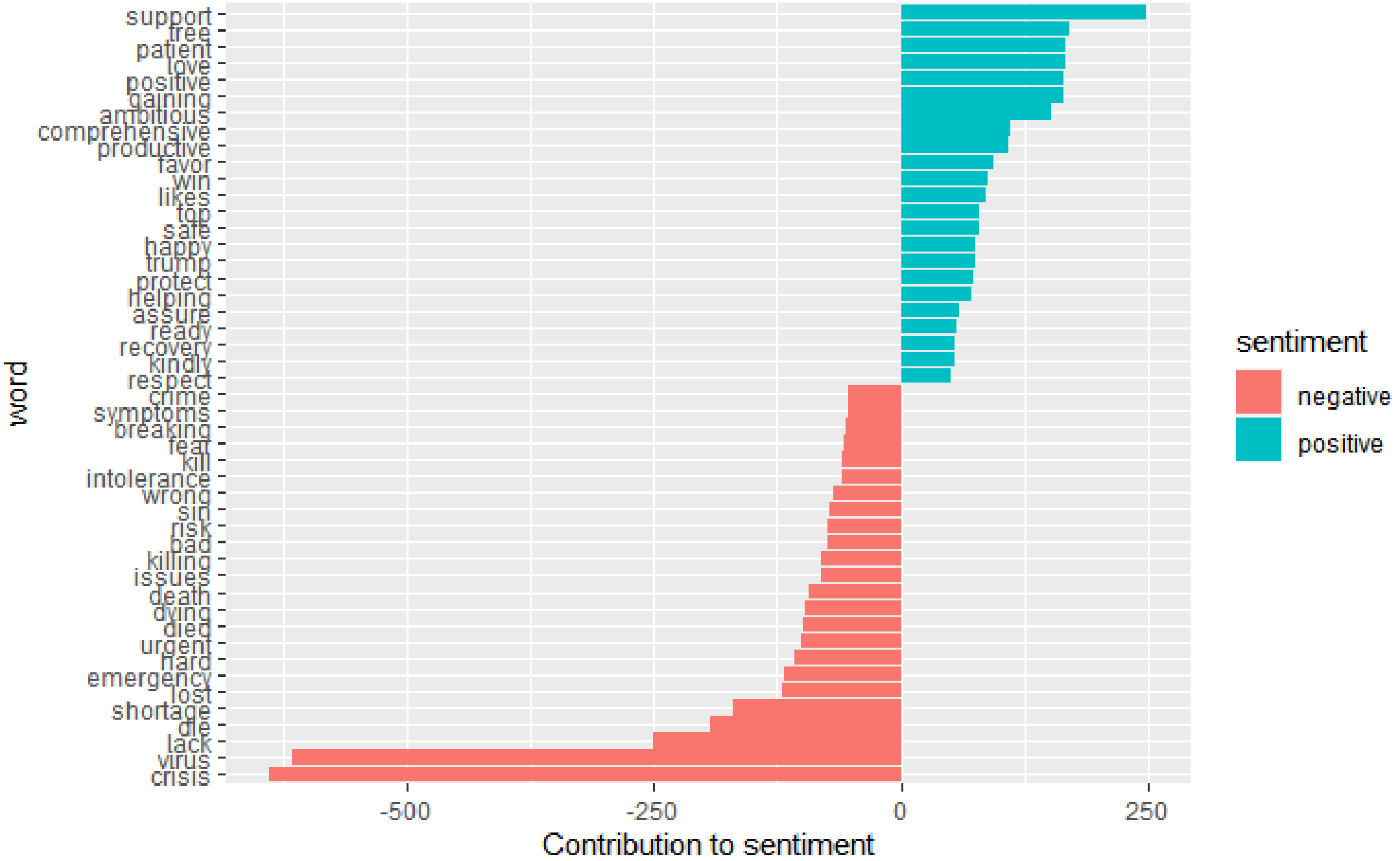
```r
library(deeplr)
library(translateR)
library(dplyr)
library(tidytext)
library(ggplot2)
library(corpus)
library(tm)
library(wordcloud)
library(igraph)
library(ggraph)
```

Libraries used for Data Analysis




```r
bing_word_counts %>%
  top_n(10) %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(x = word, y = n)) +
  geom_col() +
  xlab(NULL) +
  coord_flip() +
  labs(y = "Frequency of Words",
       x = "words",
       title = "Top 10 most used words in tweets")
```

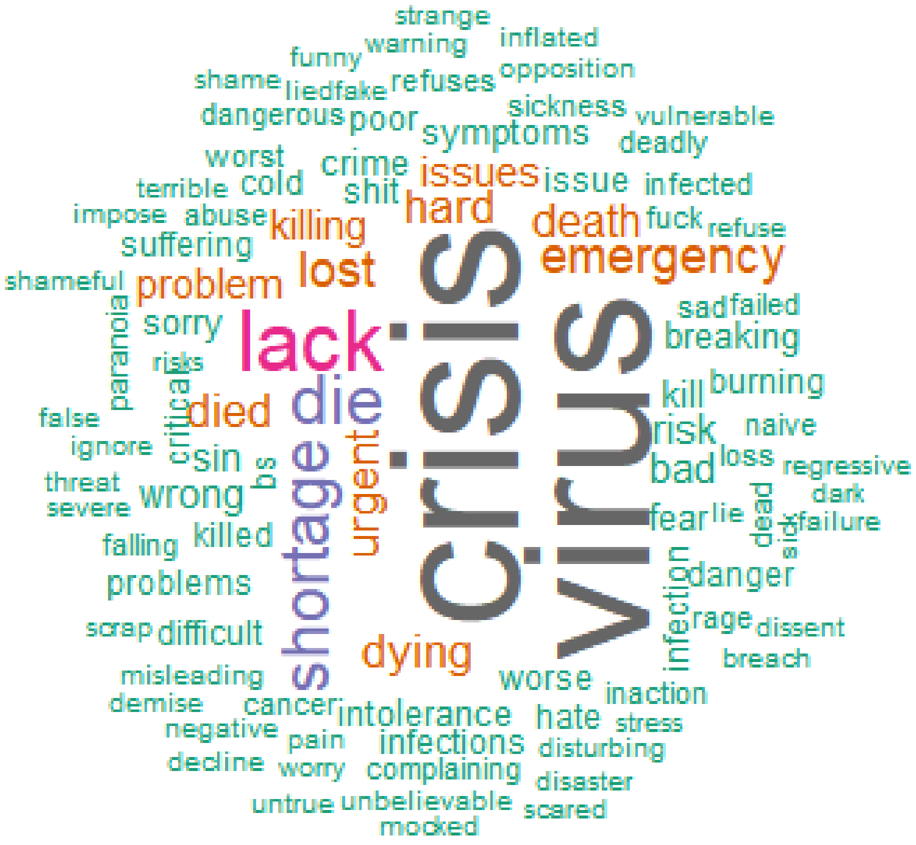

Sentiment during the COVID.
negative
crisis
virus
lack
die
shortage
lost
emergency
hard
urgent
died
positive
support
free
patient
love
positive
gaining
ambitious
comprehensive
productive
favor


support
free
patient
love
positive
gaining
ambitious
comprehensive
productive
favor
win
likes
top
safe
happy
trump
protect
helping
assure
ready
recovery
kindly
respect
crime
symptoms
breaking
fear
kill
intolerance
wrong
sin
risk
bad
killing
issues
death
dying
died
urgent
hard
emergency
lost
shortage
die
lack
virus
crisis
word
-500
-250
0
250
Contribution to sentiment
sentiment
negative
positive

```{r}
streamdata_clean %>%
  count(word) %>%
  inner_join(get_sentiments("bing"), by = "word") %>%
  with(wordcloud(word, n, max.words = 100,min.freq = 10,
          random.order = FALSE,colors = brewer.pal(8, "Dark2")))

```

Word Network: Tweets using the hashtag - Covid

Text mining twitter data


strange
warning
inflated
funny
opposition
shame
liedfake
refuses
dangerous
poor
sickness
vulnerable
symptoms
deadly
worst
crime
issues
issue
infected
terrible
cold
shit
hard
impose
abuse
killing
death
fuck
refuse
suffering
emergency
lost
shameful
problem
crisis
virus
sadfailed
paranoia
sorry
lack
breaking
risks
burning
critical
died
die
kill
false
risk
naive
ignore
sin
urgent
bad
loss
regressive
threat
bs
shortage
dark
severe
wrong
fear
lie
dead
sick
failure
falling
killed
infection
danger
problems
rage
dissent
scrap
difficult
dying
breach
misleading
worse
inaction
demise
cancer
intolerance
hate
stress
negative
pain
infections
disturbing
decline
worry
complaining
disaster
untrue
unbelievable
scared
mocked


encourage
confidence
meaningful
leading
commitment
appreciated
effective
smile
positive
trump
glad
helping
safe
patient
peace
loved
assure
favor
free
ready
win
respect
relief
won
leads
recovery
support
kindly
super
protect
enjoy
top
love
productive
amazing
tough
worth
lead
gaining
likes
happy
strong
helped
pretty
fans
ambitious
trust
proud
comprehensive
approval

References

1. https://flume.apache.org/FlumeUserGuide.html
2. https://www.tutorialspoint.com/apache_flume/fetching_twitter_data.htm
3. https://towardsdatascience.com/apache-flume-71ed475eee6d
4. https://www.youtube.com/watch?v=PdY31i25SL0
5. https://data-flair.training/blogs/apache-hive-installation/
6. https://stackoverflow.com/questions/11889261/datanode-process-not-running-in-hadoop
7. http://dbmentors.blogspot.com/2017/06/streaming-twitter-data-using-apache.html
8. https://medium.com/edureka/apache-flume-tutorial-6f7150210c76
9. https://www.h2kinfosys.com/blog/apache-flume-tutorial/
10. https://www.confluent.io/blog/avro-kafka-data/#:~:text=Avro%20has%20a%20JSON%20like,in%20a%20compact%20binary%20form.&text=It%20has%20a%20direct%20mapping,inefficient%20for%20high%2Dvolume%20usage.

11. https://towardsdatascience.com/twitter-sentiment-analysis-and-visualization-using-r-22e1f70f6967
12. https://towardsdatascience.com/twitter-data-visualization-fb4f45b63728
13. https://codeburst.io/sentiment-analysis-of-twitter-data-359fa9f86bd6
14. https://dataaspirant.com/twitter-sentiment-analysis-using-r/
15. https://www.r-bloggers.com/2014/04/twitter-sentiment-analysis-with-r/

Thank You