University of Windsor

Master of Applied Computing

Advanced Database Topics
COMP – 8157

Project on
Sentiment Analysis of Twitter Data

Guide and Instructor:
Prof. Kalyani Selvarajah

Group Members:

Bhavin Vaghasiya(110015301)

Karan Aggarwal (110015913)

Param Kewale (110021969)

Ravi Kanani(110013139)

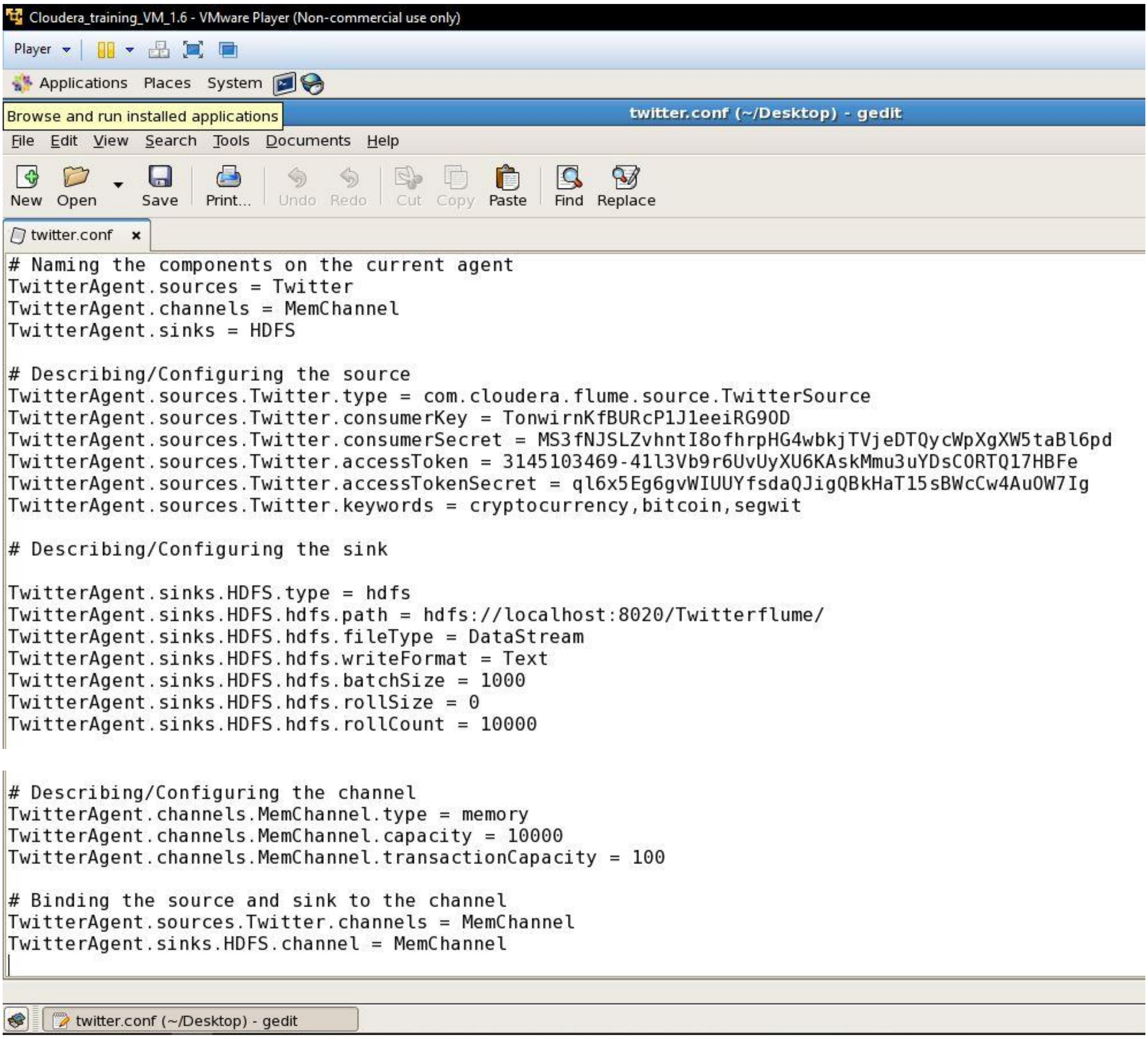
Yug Bhadreshkumar Rawal(110009769)

STEP BY STEP PROCESS

Step 1: Tweets are collected in real time from twitter using Apache Flume

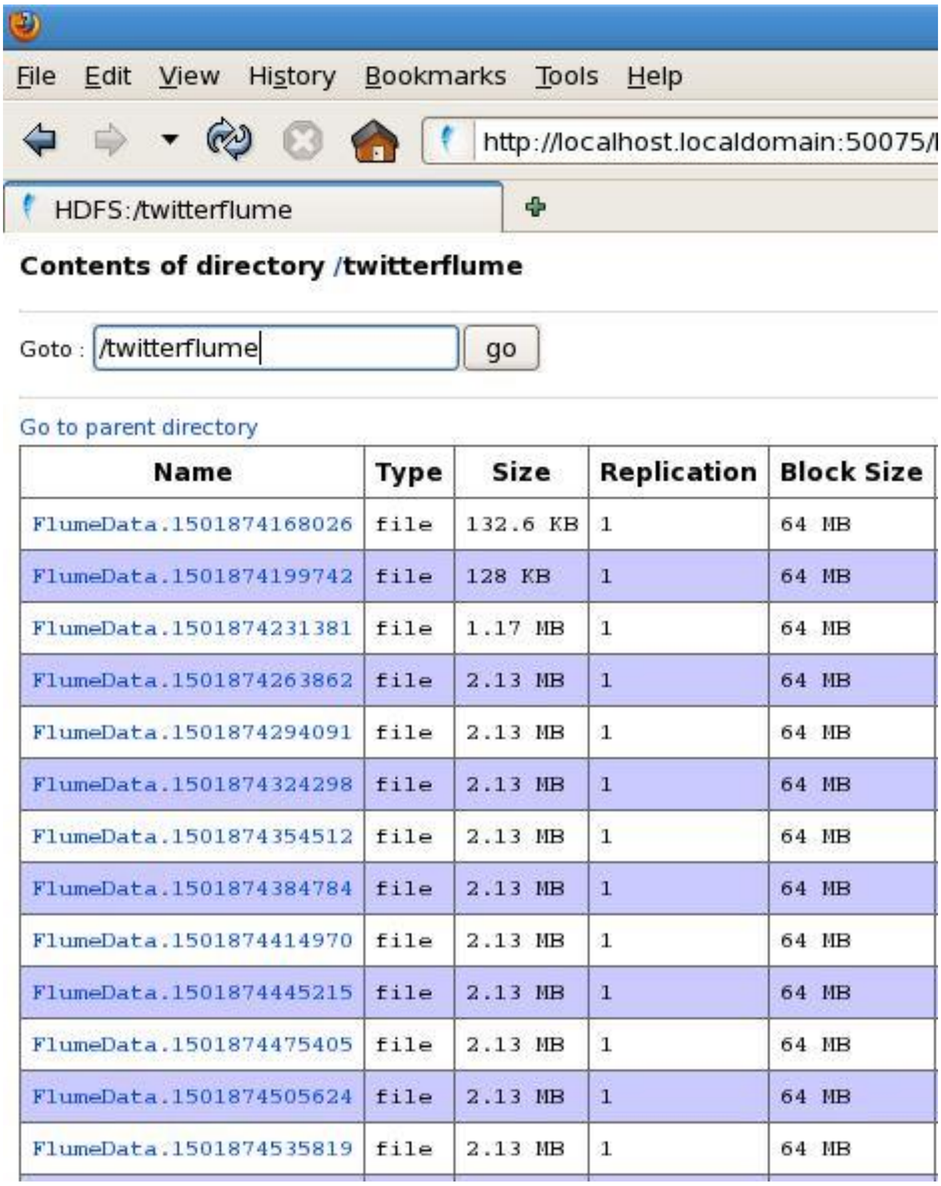
Flume uses three components Source, Channel and Sink in order to fetch and store the data in the centralized stores.

The Screenshot below shows how Source, Sink and Channel are configured and then how Source and Sink are connected to the Channel


Cloudera_training_VM_1.6 - VMware Player (Non-commercial use only)
Player
Applications Places System
Browse and run installed applications
twitter.conf (~/Desktop) - gedit
File Edit View Search Tools Documents Help
New Open Save Print... Undo Redo Cut Copy Paste Find Replace
twitter.conf
twitter.conf (~/Desktop) - gedit

```
# Naming the components on the current agent
TwitterAgent.sources = Twitter
TwitterAgent.channels = MemChannel
TwitterAgent.sinks = HDFS

# Describing/Configuring the source
TwitterAgent.sources.Twitter.type = com.cloudera.flume.source.TwitterSource
TwitterAgent.sources.Twitter.consumerKey = TonwirnKfBURcP1J1eeiRG90D
TwitterAgent.sources.Twitter.consumerSecret = MS3fNJSLZvhntI8ofhrpHG4wbkjTVjeDTQycWpXgXW5taBl6pd
TwitterAgent.sources.Twitter.accessToken = 3145103469-41l3Vb9r6UvUyXU6KAskMmu3uYDsC0RTQ17HBFe
TwitterAgent.sources.Twitter.accessTokenSecret = ql6x5Eg6gvWIUUYfsdaQJigQBkHaT15sBWcCw4Au0W7Ig
TwitterAgent.sources.Twitter.keywords = cryptocurrency,bitcoin,segwit

# Describing/Configuring the sink

TwitterAgent.sinks.HDFS.type = hdfs
TwitterAgent.sinks.HDFS.hdfs.path = hdfs://localhost:8020/Twitterflume/
TwitterAgent.sinks.HDFS.hdfs.fileType = DataStream
TwitterAgent.sinks.HDFS.hdfs.writeFormat = Text
TwitterAgent.sinks.HDFS.hdfs.batchSize = 1000
TwitterAgent.sinks.HDFS.hdfs.rollSize = 0
TwitterAgent.sinks.HDFS.hdfs.rollCount = 10000


# Describing/Configuring the channel
TwitterAgent.channels.MemChannel.type = memory
TwitterAgent.channels.MemChannel.capacity = 10000
TwitterAgent.channels.MemChannel.transactionCapacity = 100

# Binding the source and sink to the channel
TwitterAgent.sources.Twitter.channels = MemChannel
TwitterAgent.sinks.HDFS.channel = MemChannel
```

Step 2: The Screenshot Below shows the data obtained from Twitter using Flume in JSON format.

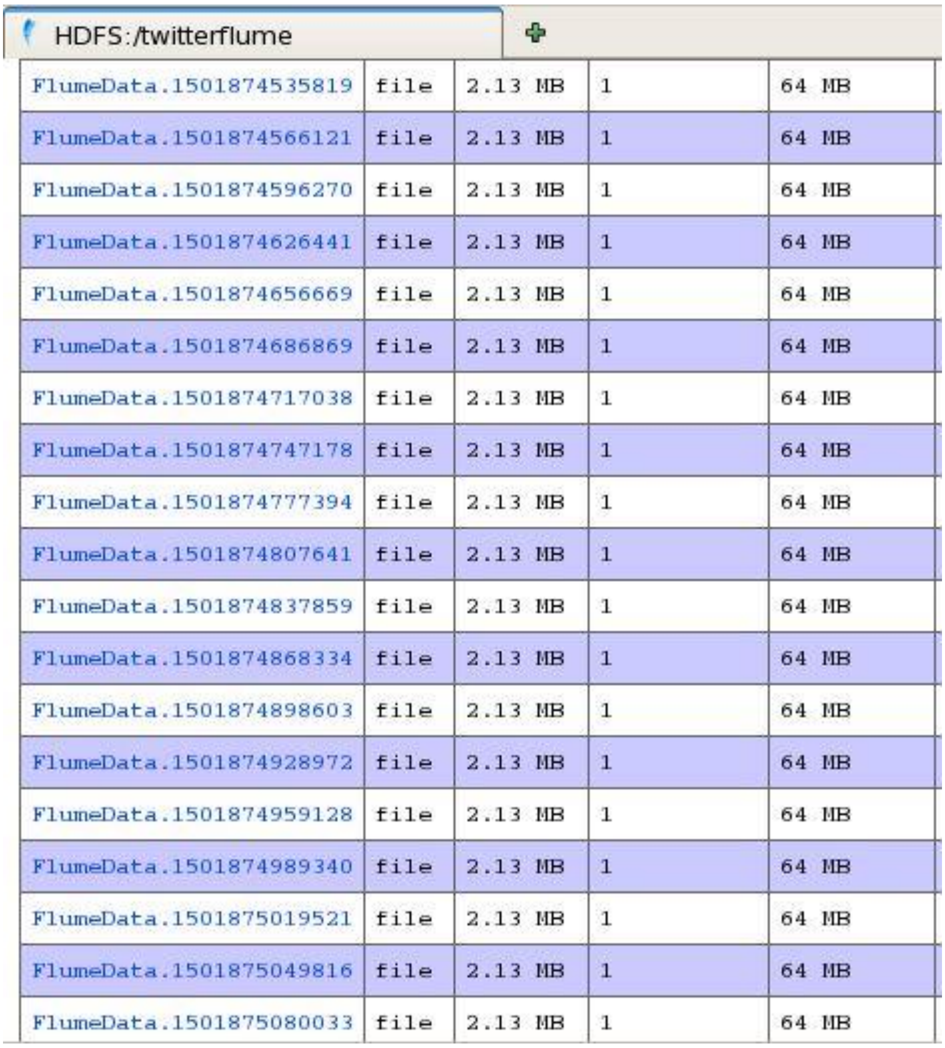


Contents of directory /twitterflume

Goto : /twitterflume

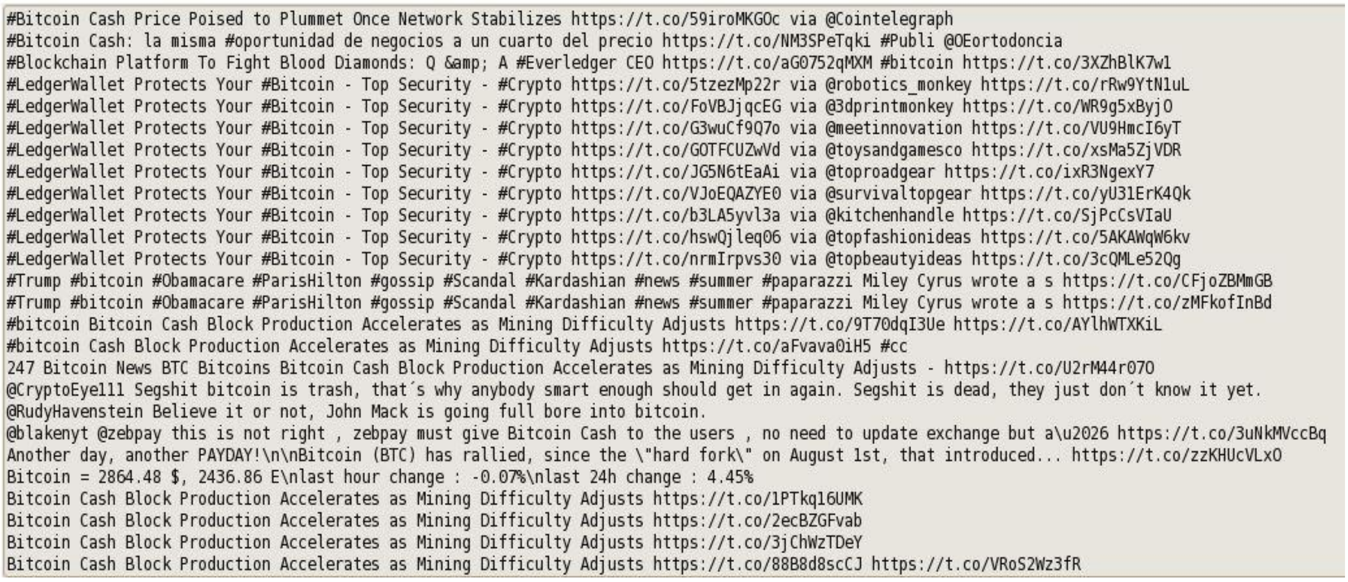
Go to parent directory

Name	Type	Size	Replication	Block Size
FlumeData.1501874168026	file	132.6 KB	1	64 MB
FlumeData.1501874199742	file	128 KB	1	64 MB
FlumeData.1501874231381	file	1.17 MB	1	64 MB
FlumeData.1501874263862	file	2.13 MB	1	64 MB
FlumeData.1501874294091	file	2.13 MB	1	64 MB
FlumeData.1501874324298	file	2.13 MB	1	64 MB
FlumeData.1501874354512	file	2.13 MB	1	64 MB
FlumeData.1501874384784	file	2.13 MB	1	64 MB
FlumeData.1501874414970	file	2.13 MB	1	64 MB
FlumeData.1501874445215	file	2.13 MB	1	64 MB
FlumeData.1501874475405	file	2.13 MB	1	64 MB
FlumeData.1501874505624	file	2.13 MB	1	64 MB
FlumeData.1501874535819	file	2.13 MB	1	64 MB

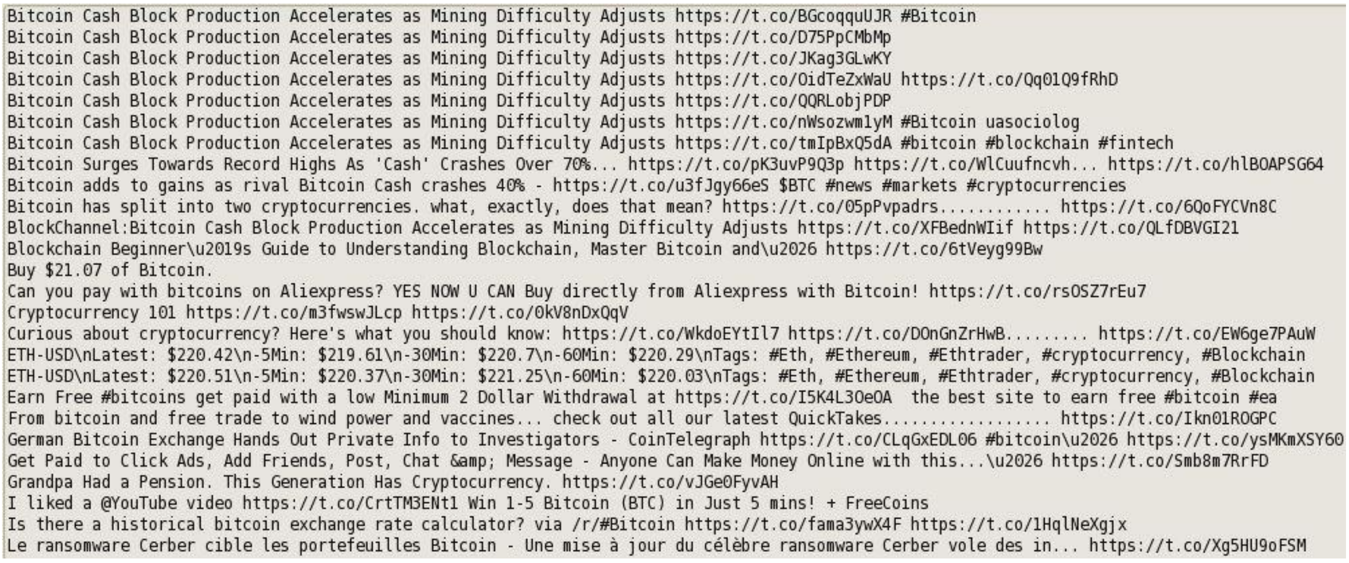


FlumeData.1501874535819	file	2.13 MB	1	64 MB
FlumeData.1501874566121	file	2.13 MB	1	64 MB
FlumeData.1501874596270	file	2.13 MB	1	64 MB
FlumeData.1501874626441	file	2.13 MB	1	64 MB
FlumeData.1501874656669	file	2.13 MB	1	64 MB
FlumeData.1501874686869	file	2.13 MB	1	64 MB
FlumeData.1501874717038	file	2.13 MB	1	64 MB
FlumeData.1501874747178	file	2.13 MB	1	64 MB
FlumeData.1501874777394	file	2.13 MB	1	64 MB
FlumeData.1501874807641	file	2.13 MB	1	64 MB
FlumeData.1501874837859	file	2.13 MB	1	64 MB
FlumeData.1501874868334	file	2.13 MB	1	64 MB
FlumeData.1501874898603	file	2.13 MB	1	64 MB
FlumeData.1501874928972	file	2.13 MB	1	64 MB
FlumeData.1501874959128	file	2.13 MB	1	64 MB
FlumeData.1501874989340	file	2.13 MB	1	64 MB
FlumeData.1501875019521	file	2.13 MB	1	64 MB
FlumeData.1501875049816	file	2.13 MB	1	64 MB
FlumeData.1501875080033	file	2.13 MB	1	64 MB

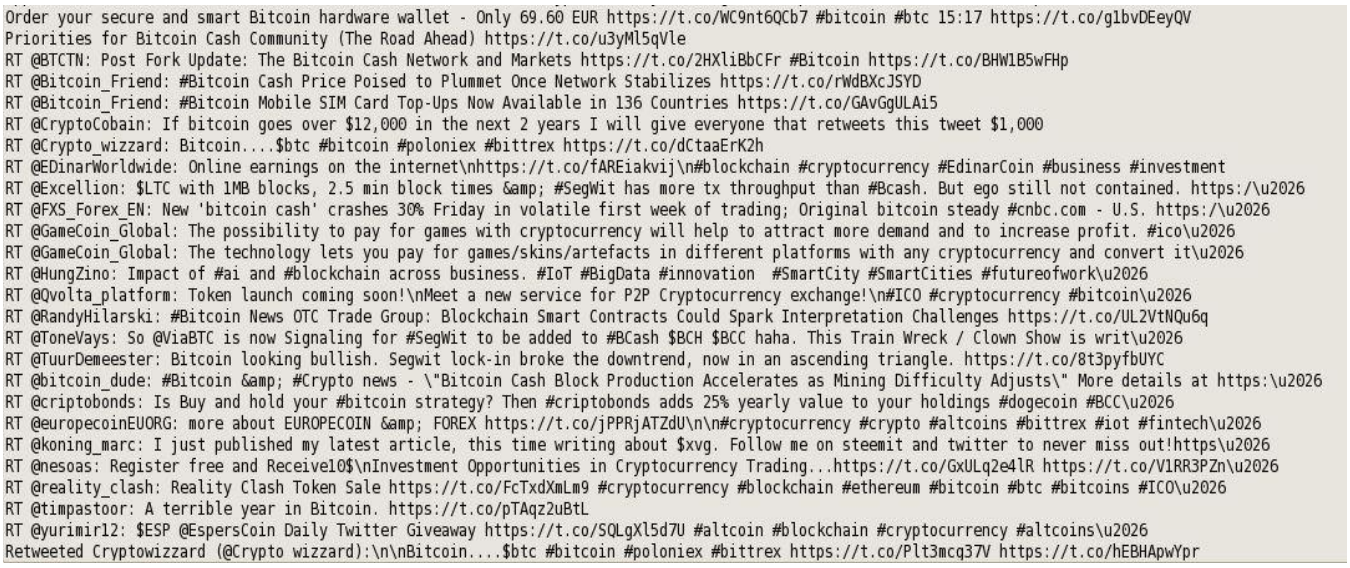
Step 3: Now this data is then transformed into text format. So that futher processing can be performed.The screenshots below shows the tweets related to Bitcoin in text format.



```
#Bitcoin Cash Price Poised to Plummet Once Network Stabilizes https://t.co/59iroMKGOc via @Cointelegraph
#Bitcoin Cash: la misma #oportunidad de negocios a un cuarto del precio https://t.co/NM3SPeTqki #Publi @OEortodoncia
#Blockchain Platform To Fight Blood Diamonds: Q &amp; A #Everledger CEO https://t.co/aG0752qMXM #bitcoin https://t.co/3XZhBlK7w1
#LedgerWallet Protects Your #Bitcoin - Top Security - #Crypto https://t.co/5tzezMp22r via @robotics_monkey https://t.co/rRw9YtN1uL
#LedgerWallet Protects Your #Bitcoin - Top Security - #Crypto https://t.co/FoVBJjqcEG via @3dprintmonkey https://t.co/WR9g5xByjO
#LedgerWallet Protects Your #Bitcoin - Top Security - #Crypto https://t.co/G3wuCf9Q7o via @meetinnovation https://t.co/VU9HmcI6yT
#LedgerWallet Protects Your #Bitcoin - Top Security - #Crypto https://t.co/GOTFCUZwVd via @toysandgamesco https://t.co/xsMa5ZjVDR
#LedgerWallet Protects Your #Bitcoin - Top Security - #Crypto https://t.co/JG5N6tEaAi via @toproadgear https://t.co/ixR3NgexY7
#LedgerWallet Protects Your #Bitcoin - Top Security - #Crypto https://t.co/VJoEQAZYE0 via @survivaltopgear https://t.co/yU31ErK4Qk
#LedgerWallet Protects Your #Bitcoin - Top Security - #Crypto https://t.co/b3LA5yvl3a via @kitchenhandle https://t.co/SjPcCsVIaU
#LedgerWallet Protects Your #Bitcoin - Top Security - #Crypto https://t.co/hswQjleq06 via @topfashionideas https://t.co/5AKAWqW6kv
#LedgerWallet Protects Your #Bitcoin - Top Security - #Crypto https://t.co/nrmIrpvs30 via @topbeautyideas https://t.co/3cQMLe52Qg
#Trump #bitcoin #Obamacare #ParisHilton #gossip #Scandal #Kardashian #news #summer #paparazzi Miley Cyrus wrote a s https://t.co/CFjoZBMmGB
#Trump #bitcoin #Obamacare #ParisHilton #gossip #Scandal #Kardashian #news #summer #paparazzi Miley Cyrus wrote a s https://t.co/zMFkofInBd
#bitcoin Bitcoin Cash Block Production Accelerates as Mining Difficulty Adjusts https://t.co/9T70dqI3Ue https://t.co/AYlhWTXKiL
#bitcoin Cash Block Production Accelerates as Mining Difficulty Adjusts https://t.co/aFvava0iH5 #cc
247 Bitcoin News BTC Bitcoins Bitcoin Cash Block Production Accelerates as Mining Difficulty Adjusts - https://t.co/U2rM44r07O
@CryptoEye111 Segshit bitcoin is trash, that´s why anybody smart enough should get in again. Segshit is dead, they just don´t know it yet.
@RudyHavenstein Believe it or not, John Mack is going full bore into bitcoin.
@blakenyt @zebpay this is not right , zebpay must give Bitcoin Cash to the users , no need to update exchange but a\u2026 https://t.co/3uNkMVccBq
Another day, another PAYDAY!\n\nBitcoin (BTC) has rallied, since the \"hard fork\" on August 1st, that introduced... https://t.co/zzKHUcVLxO
Bitcoin = 2864.48 $, 2436.86 E\nlast hour change : -0.07%\nlast 24h change : 4.45%
Bitcoin Cash Block Production Accelerates as Mining Difficulty Adjusts https://t.co/1PTkq16UMK
Bitcoin Cash Block Production Accelerates as Mining Difficulty Adjusts https://t.co/2ecBZGFvab
Bitcoin Cash Block Production Accelerates as Mining Difficulty Adjusts https://t.co/3jChWzTDeY
Bitcoin Cash Block Production Accelerates as Mining Difficulty Adjusts https://t.co/88B8d8scCJ https://t.co/VRoS2Wz3fR
```



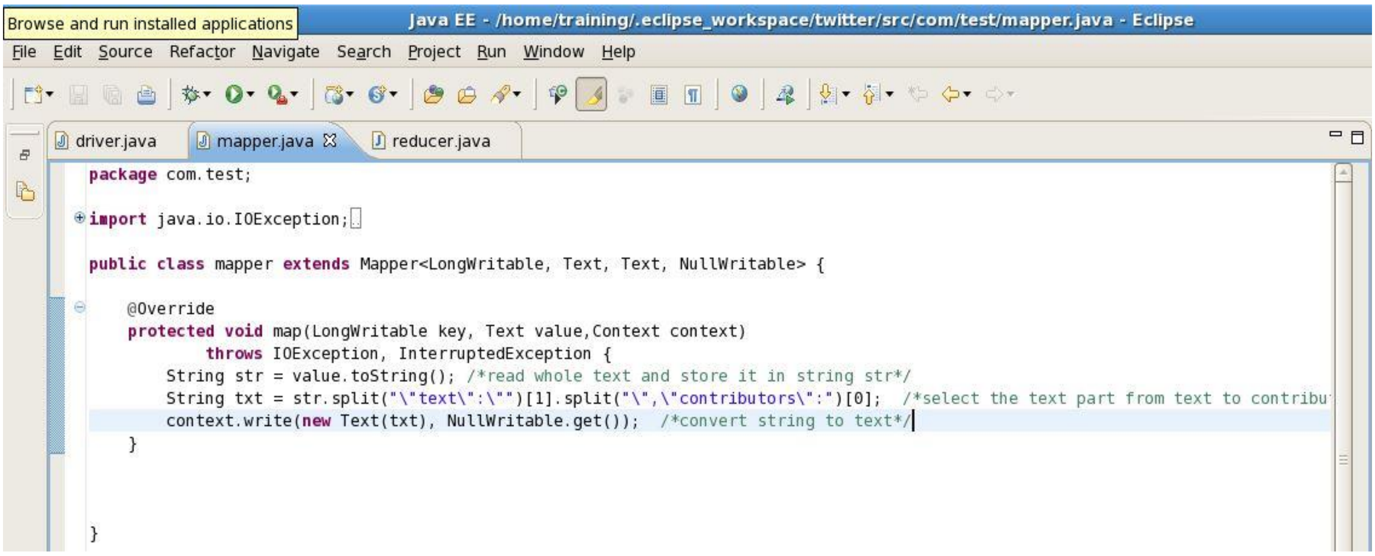
```
Bitcoin Cash Block Production Accelerates as Mining Difficulty Adjusts https://t.co/BGcoqquUJR #Bitcoin
Bitcoin Cash Block Production Accelerates as Mining Difficulty Adjusts https://t.co/D75PpCMbMp
Bitcoin Cash Block Production Accelerates as Mining Difficulty Adjusts https://t.co/JKag3GLwKY
Bitcoin Cash Block Production Accelerates as Mining Difficulty Adjusts https://t.co/OidTeZxWaU https://t.co/Qq01Q9fRhD
Bitcoin Cash Block Production Accelerates as Mining Difficulty Adjusts https://t.co/QQRLobjPDP
Bitcoin Cash Block Production Accelerates as Mining Difficulty Adjusts https://t.co/nWsozwm1yM #Bitcoin uasociolog
Bitcoin Cash Block Production Accelerates as Mining Difficulty Adjusts https://t.co/tmIpBxQ5dA #bitcoin #blockchain #fintech
Bitcoin Surges Towards Record Highs As 'Cash' Crashes Over 70%... https://t.co/pK3uvP9Q3p https://t.co/WlCuufncvh... https://t.co/hlBOAPSG64
Bitcoin adds to gains as rival Bitcoin Cash crashes 40% - https://t.co/u3fJgy66eS $BTC #news #markets #cryptocurrencies
Bitcoin has split into two cryptocurrencies. what, exactly, does that mean? https://t.co/05pPvpadrs............ https://t.co/6QoFYCVn8C
BlockChannel:Bitcoin Cash Block Production Accelerates as Mining Difficulty Adjusts https://t.co/XFBednWIif https://t.co/QLfDBVGI21
Blockchain Beginner\u2019s Guide to Understanding Blockchain, Master Bitcoin and\u2026 https://t.co/6tVeyg99Bw
Buy $21.07 of Bitcoin.
Can you pay with bitcoins on Aliexpress? YES NOW U CAN Buy directly from Aliexpress with Bitcoin! https://t.co/rsOSZ7rEu7
Cryptocurrency 101 https://t.co/m3fwswJLcp https://t.co/0kV8nDxQqV
Curious about cryptocurrency? Here's what you should know: https://t.co/WkdoEYtIl7 https://t.co/DOnGnZrHwB......... https://t.co/EW6ge7PAuW
ETH-USD\nLatest: $220.42\n-5Min: $219.61\n-30Min: $220.7\n-60Min: $220.29\nTags: #Eth, #Ethereum, #Ethtrader, #cryptocurrency, #Blockchain
ETH-USD\nLatest: $220.51\n-5Min: $220.37\n-30Min: $221.25\n-60Min: $220.03\nTags: #Eth, #Ethereum, #Ethtrader, #cryptocurrency, #Blockchain
Earn Free #bitcoins get paid with a low Minimum 2 Dollar Withdrawal at https://t.co/I5K4L3OeOA  the best site to earn free #bitcoin #ea
From bitcoin and free trade to wind power and vaccines... check out all our latest QuickTakes................. https://t.co/Ikn01ROGPC
German Bitcoin Exchange Hands Out Private Info to Investigators - CoinTelegraph https://t.co/CLqGxEDL06 #bitcoin\u2026 https://t.co/ysMKmXSY60
Get Paid to Click Ads, Add Friends, Post, Chat &amp; Message - Anyone Can Make Money Online with this...\u2026 https://t.co/Smb8m7RrFD
Grandpa Had a Pension. This Generation Has Cryptocurrency. https://t.co/vJGe0FyvAH
I liked a @YouTube video https://t.co/CrtTM3ENt1 Win 1-5 Bitcoin (BTC) in Just 5 mins! + FreeCoins
Is there a historical bitcoin exchange rate calculator? via /r/#Bitcoin https://t.co/fama3ywX4F https://t.co/1HqlNeXgjx
Le ransomware Cerber cible les portefeuilles Bitcoin - Une mise à jour du célèbre ransomware Cerber vole des in... https://t.co/Xg5HU9oFSM
```



```
Order your secure and smart Bitcoin hardware wallet - Only 69.60 EUR https://t.co/WC9nt6QCb7 #bitcoin #btc 15:17 https://t.co/g1bvDEeyQV
Priorities for Bitcoin Cash Community (The Road Ahead) https://t.co/u3yMl5qVle
RT @BTCTN: Post Fork Update: The Bitcoin Cash Network and Markets https://t.co/2HXliBbCFr #Bitcoin https://t.co/BHW1B5wFHp
RT @Bitcoin_Friend: #Bitcoin Cash Price Poised to Plummet Once Network Stabilizes https://t.co/rWdBXcJSYD
RT @Bitcoin_Friend: #Bitcoin Mobile SIM Card Top-Ups Now Available in 136 Countries https://t.co/GAvGgULAi5
RT @CryptoCobain: If bitcoin goes over $12,000 in the next 2 years I will give everyone that retweets this tweet $1,000
RT @Crypto_wizzard: Bitcoin....$btc #bitcoin #poloniex #bittrex https://t.co/dCtaaErK2h
RT @EDinarWorldwide: Online earnings on the internet\nhttps://t.co/fAREiakvij\n#blockchain #cryptocurrency #EdinarCoin #business #investment
RT @Excellion: $LTC with 1MB blocks, 2.5 min block times &amp; #SegWit has more tx throughput than #Bcash. But ego still not contained. https:/\u2026
RT @FXS_Forex_EN: New 'bitcoin cash' crashes 30% Friday in volatile first week of trading; Original bitcoin steady #cnbc.com - U.S. https:/\u2026
RT @GameCoin_Global: The possibility to pay for games with cryptocurrency will help to attract more demand and to increase profit. #ico\u2026
RT @GameCoin_Global: The technology lets you pay for games/skins/artefacts in different platforms with any cryptocurrency and convert it\u2026
RT @HungZino: Impact of #ai and #blockchain across business. #IoT #BigData #innovation  #SmartCity #SmartCities #futureofwork\u2026
RT @Qvolta_platform: Token launch coming soon!\nMeet a new service for P2P Cryptocurrency exchange!\n#ICO #cryptocurrency #bitcoin\u2026
RT @RandyHilarski: #Bitcoin News OTC Trade Group: Blockchain Smart Contracts Could Spark Interpretation Challenges https://t.co/UL2VtNQu6q
RT @ToneVays: So @ViaBTC is now Signaling for #SegWit to be added to #BCash $BCH $BCC haha. This Train Wreck / Clown Show is writ\u2026
RT @TuurDemeester: Bitcoin looking bullish. Segwit lock-in broke the downtrend, now in an ascending triangle. https://t.co/8t3pyfbUYC
RT @bitcoin_dude: #Bitcoin &amp; #Crypto news - \"Bitcoin Cash Block Production Accelerates as Mining Difficulty Adjusts\" More details at https:\u2026
RT @criptobonds: Is Buy and hold your #bitcoin strategy? Then #criptobonds adds 25% yearly value to your holdings #dogecoin #BCC\u2026
RT @europecoinEUORG: more about EUROPECOIN &amp; FOREX https://t.co/jPPRjATZdU\n\n#cryptocurrency #crypto #altcoins #bittrex #iot #fintech\u2026
RT @koning_marc: I just published my latest article, this time writing about $xvg. Follow me on steemit and twitter to never miss out!https\u2026
RT @nesoas: Register free and Receive10$\nInvestment Opportunities in Cryptocurrency Trading...https://t.co/GxULq2e4lR https://t.co/V1RR3PZn\u2026
RT @reality_clash: Reality Clash Token Sale https://t.co/FcTxdXmLm9 #cryptocurrency #blockchain #ethereum #bitcoin #btc #bitcoins #ICO\u2026
RT @timpastoor: A terrible year in Bitcoin. https://t.co/pTAqz2uBtL
RT @yurimir12: $ESP @EspersCoin Daily Twitter Giveaway https://t.co/SQLgXl5d7U #altcoin #blockchain #cryptocurrency #altcoins\u2026
Retweeted Cryptowizzard (@Crypto wizzard):\n\nBitcoin....$btc #bitcoin #poloniex #bittrex https://t.co/Plt3mcq37V https://t.co/hEBHApwYpr
```

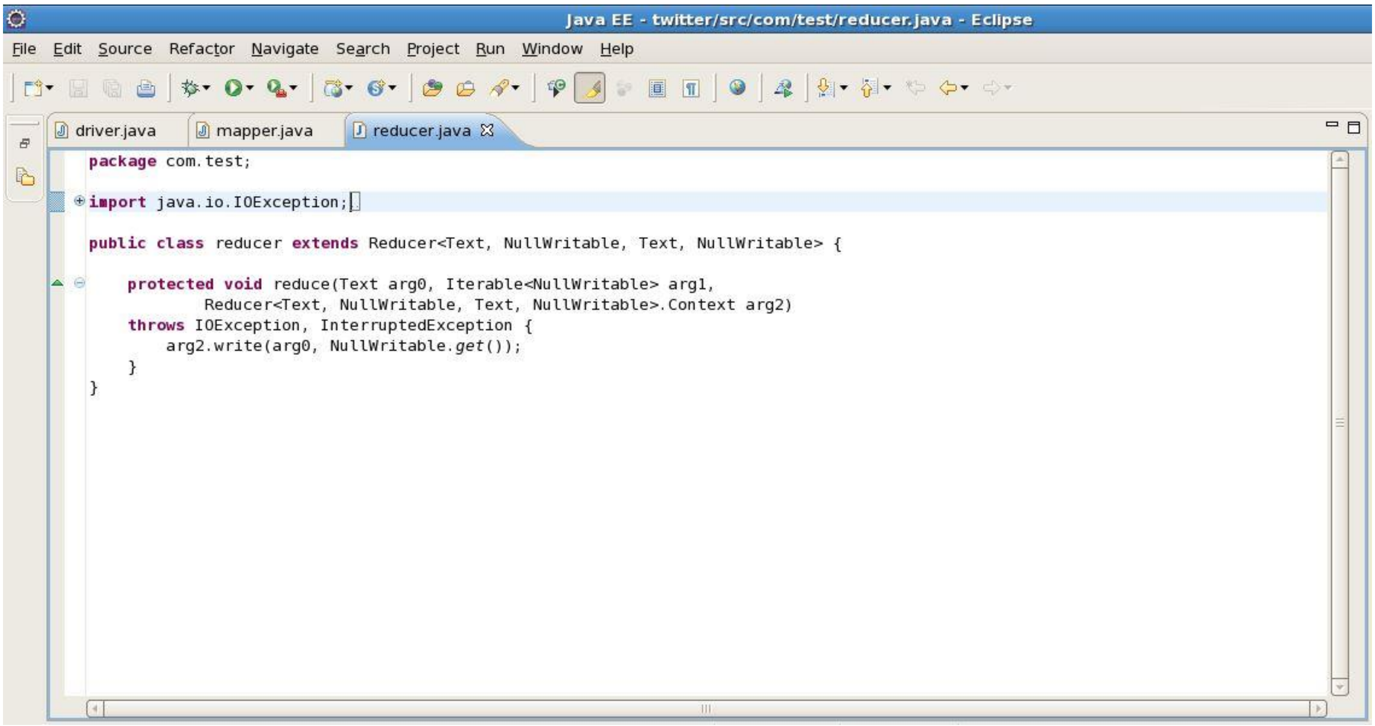
Step 4: Now the file is passed to the mapper which reads the file line by line. The data will be processed and smaller chunks of data will be created

Below is the screenshot of Mapper Class


Browse and run installed applications
Java EE - /home/training/.eclipse_workspace/twitter/src/com/test/mapper.java - Eclipse
File Edit Source Refactor Navigate Search Project Run Window Help
driver.java
mapper.java
reducer.java

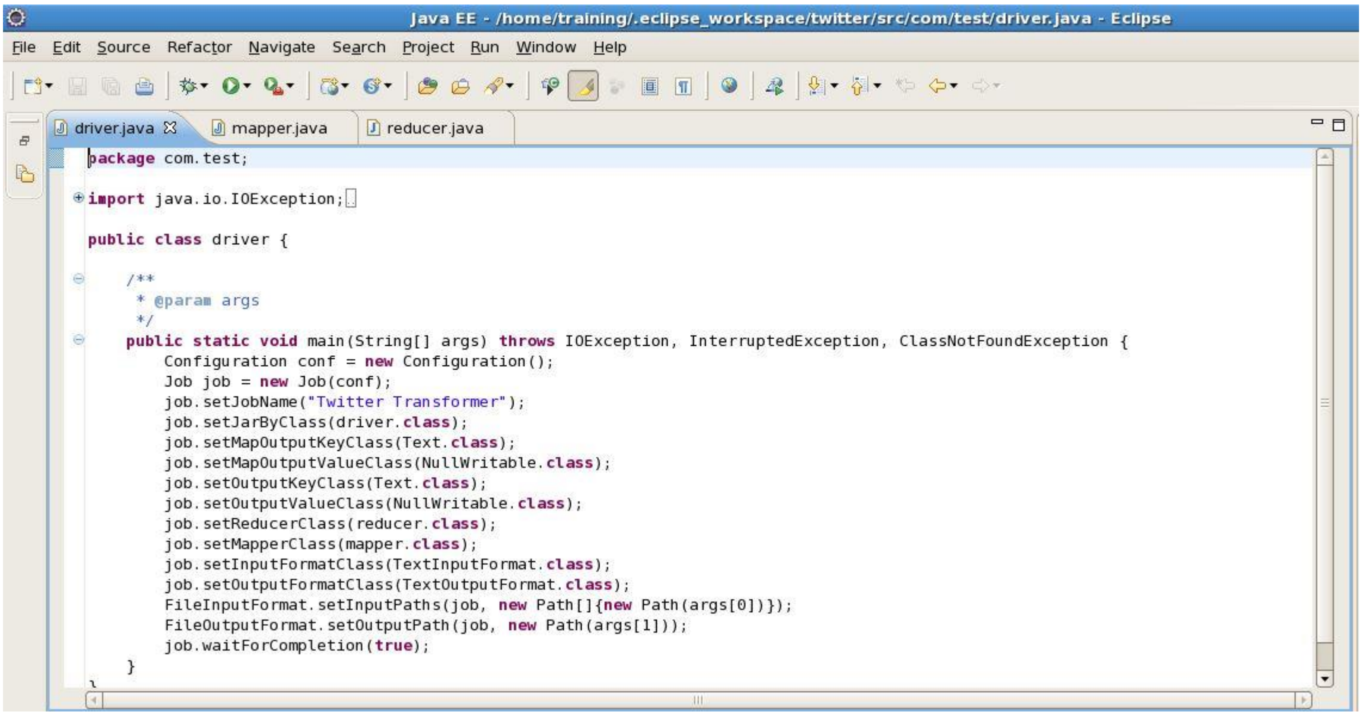
```
package com.test;

import java.io.IOException;

public class mapper extends Mapper<LongWritable, Text, Text, NullWritable> {

    @Override
    protected void map(LongWritable key, Text value,Context context)
            throws IOException, InterruptedException {
        String str = value.toString(); /*read whole text and store it in string str*/
        String txt = str.split("\"text\":\"")[1].split("\",\"contributors\":")[0];  /*select the text part from text to contribu
        context.write(new Text(txt), NullWritable.get());  /*convert string to text*/
    }


}
```

The output of the Mapper is given as input to the Reducer. After processing new set of output is produced. Below is the screenshot of reducer class.


Java EE - twitter/src/com/test/reducer.java - Eclipse
File Edit Source Refactor Navigate Search Project Run Window Help
driver.java
mapper.java
reducer.java

```
package com.test;

import java.io.IOException;

public class reducer extends Reducer<Text, NullWritable, Text, NullWritable> {

    protected void reduce(Text arg0, Iterable<NullWritable> arg1,
            Reducer<Text, NullWritable, Text, NullWritable>.Context arg2)
    throws IOException, InterruptedException {
        arg2.write(arg0, NullWritable.get());
    }
}
```

The mapper and the reducer class are being driven by the driver class. Below is the screenshot of the Driver class.



Java EE - /home/training/.eclipse_workspace/twitter/src/com/test/driver.java - Eclipse
File Edit Source Refactor Navigate Search Project Run Window Help
driver.java
mapper.java
reducer.java

```
package com.test;

import java.io.IOException;

public class driver {

    /**
     * @param args
     */
    public static void main(String[] args) throws IOException, InterruptedException, ClassNotFoundException {
        Configuration conf = new Configuration();
        Job job = new Job(conf);
        job.setJobName("Twitter Transformer");
        job.setJarByClass(driver.class);
        job.setMapOutputKeyClass(Text.class);
        job.setMapOutputValueClass(NullWritable.class);
        job.setOutputKeyClass(Text.class);
        job.setOutputValueClass(NullWritable.class);
        job.setReducerClass(reducer.class);
        job.setMapperClass(mapper.class);
        job.setInputFormatClass(TextInputFormat.class);
        job.setOutputFormatClass(TextOutputFormat.class);
        FileInputFormat.setInputPaths(job, new Path[]{new Path(args[0])});
        FileOutputFormat.setOutputPath(job, new Path(args[1]));
        job.waitForCompletion(true);
    }
}
```

In the upcoming phase will be linking the AFFIN Dictionary to our data in order to analyze the sentiments of the tweets. The words will be rated from +5 to -5 using that dictionary.