Project 1.1

Twitter Sentiment Analysis

1. Introduction You need to stream data from twitter and you need to perform sentiment analysis on the data.

**Flume.conf**

*TwitterAgent.sources = Twitter*

*TwitterAgent.channels = MemChannel*

*TwitterAgent.sinks = HDFS*

*TwitterAgent.sources.Twitter.type = com.cloudera.flume.source.TwitterSource*

*TwitterAgent.sources.Twitter.channels = MemChannel*

*TwitterAgent.sources.Twitter.consumerKey = O1B0dBZiazZuTtlcn3iD1zN8c*

*TwitterAgent.sources.Twitter.consumerSecret = dJ0gy05ZNCCOznICIC9RncVnIUOI6ajf5U4zC5tzXIrvoMvlss*

*TwitterAgent.sources.Twitter.accessToken = 2306605970-z7WpGnYkx7J8Y2BclnRpsqsCiI6EiUbq1QkHEET*

*TwitterAgent.sources.Twitter.accessTokenSecret = Lr31L0EwKhCAnfMP3ZMSHtkVmaFFPH0rLYYPbDik6XS7d*

*TwitterAgent.sources.Twitter.keywords = cricket*

*TwitterAgent.sinks.HDFS.channel = MemChannel*

*TwitterAgent.sinks.HDFS.type = hdfs*

*TwitterAgent.sinks.HDFS.hdfs.path = hdfs://localhost/user/ec2-user/flume/tweets*

*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*

*TwitterAgent.channels.MemChannel.type = memory*

*TwitterAgent.channels.MemChannel.capacity = 10000*

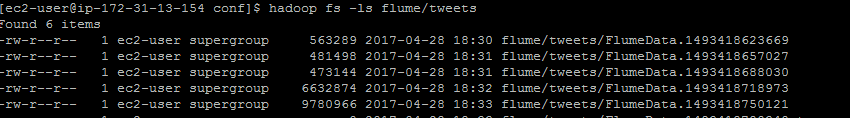
*TwitterAgent.channels.MemChannel.transactionCapacity = 100*

**Flume agent start**

flume-ng agent --conf-file /usr/lib/flume-ng/conf/flume.conf Dflume.root.logger=DEBUG,console -n TwitterAgent

**data streamed to hdfs**





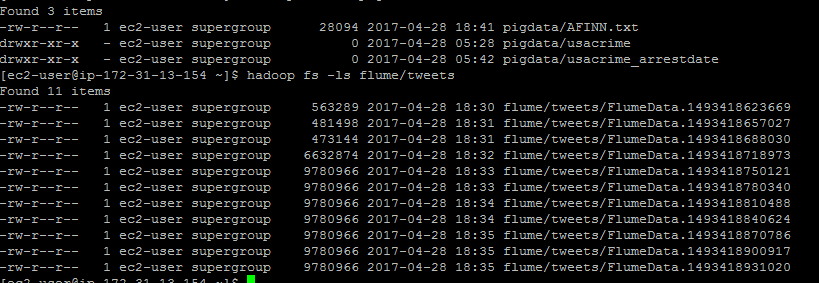
2. Objective

3. Prerequisites You need to have a Hadoop cluster installed in your system.

4. Associated Data Files

<https://acadgild.com/blog/sentiment-analysis-on-tweets-using-afinn-dictionary/>

tweets and afinn are loaded to hdfs



5. Problem Statement Follow the below blog and perform sentiment analysis using Mapreduce/pig

<https://acadgild.com/blog/sentiment-analysis-on-tweets-using-afinn-dictionary/>

Submit the screen shots of the final results with the Source code

load\_tweets = LOAD '/user/ec2-user/flume/tweets/' USING com.twitter.elephantbird.pig.load.JsonLoader('-nestedLoad') AS myMap;

extract\_details = FOREACH load\_tweets GENERATE myMap#'id' as id,myMap#'text' as text;

tokens = foreach extract\_details generate id,text, FLATTEN(TOKENIZE(text)) As word;

dictionary = load '/user/ec2-user/pigdata/AFINN.txt' using PigStorage('\t') AS(word:chararray,rating:int);

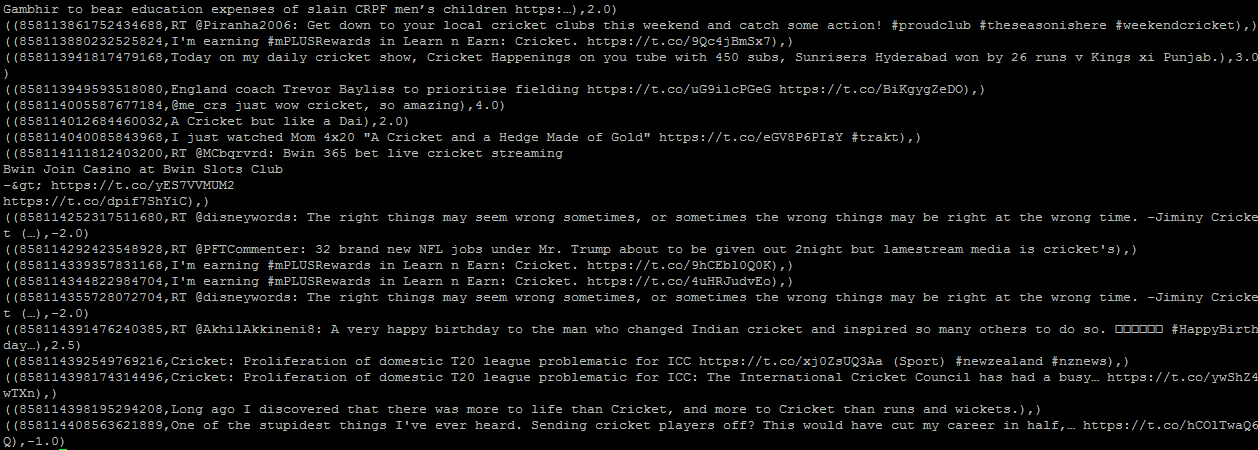
word\_rating = join tokens by word left outer, dictionary by word using 'replicated';

rating = foreach word\_rating generate tokens::id as id,tokens::text as text, dictionary::rating as rate;

word\_group = group rating by (id,text);

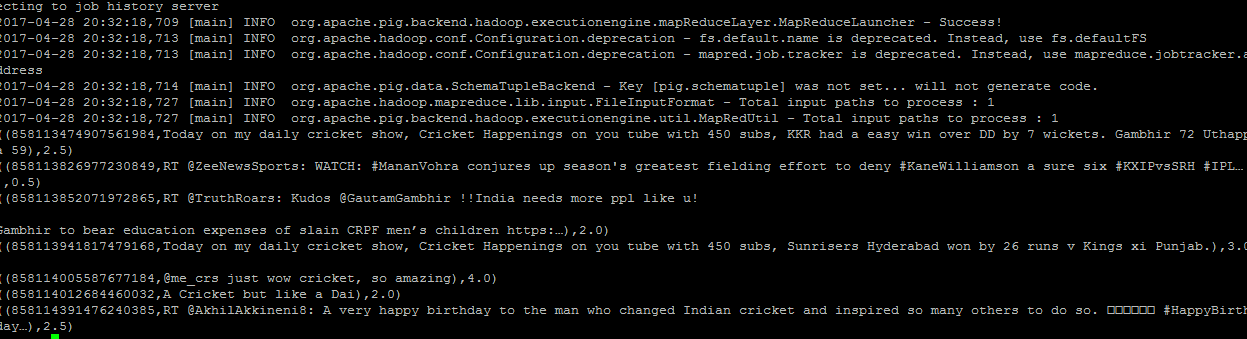
avg\_rate = foreach word\_group generate group, AVG(rating.rate) as tweet\_rating;

dump avg\_rate;



positive\_tweets = filter avg\_rate by tweet\_rating>=0;

dump positive\_tweets;



2017-04-28 20:32:18,727 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1

((858113474907561984,Today on my daily cricket show, Cricket Happenings on you tube with 450 subs, KKR had a easy win over DD by 7 wickets. Gambhir 72 Uthappa 59),2.5)

((858113826977230849,RT @ZeeNewsSports: WATCH: #MananVohra conjures up season's greatest fielding effort to deny #KaneWilliamson a sure six #KXIPvsSRH #IPL… ),0.5)

((858113852071972865,RT @TruthRoars: Kudos @GautamGambhir !!India needs more ppl like u!

Gambhir to bear education expenses of slain CRPF men’s children https:…),2.0)

((858113941817479168,Today on my daily cricket show, Cricket Happenings on you tube with 450 subs, Sunrisers Hyderabad won by 26 runs v Kings xi Punjab.),3.0)

((858114005587677184,@me\_crs just wow cricket, so amazing),4.0)

((858114012684460032,A Cricket but like a Dai),2.0)

((858114391476240385,RT @AkhilAkkineni8: A very happy birthday to the man who changed Indian cricket and inspired so many others to do so. 🙏🏻🙏🏻🙏🏻 #HappyBirthday…),2.5)