ANALYZING SOCIAL MEDIA AND CUSTOMER SENTIMENT WITH APACHE NIFI AND HDP SEARCH

Tools and Technologies:

1. Hortonworks
2. Apache NIFI
3. Apache SOLR
4. LucidWorks HDP Search
5. Apache Zeppelin
6. Twitter API/Social Media Data API

OUTLINE

1. Install Apache NiFi
2. Configure and Start Solr
3. Create a Twitter Application
4. Create a Data Flow with Nifi
5. (Optional) Generating Random Twitter Data
6. Analyze and Search Data with Solr
7. Analyze Tweet Data in Hive
8. Visualize Sentiment with Zeppelin

Install Apache NiFi:

Installation steps:

<http://hortonworks.com/hadoop-tutorial/learning-ropes-apache-nifi/#section_3>

Tutorial:

<http://hortonworks.com/hadoop-tutorial/learning-ropes-apache-nifi/#section_2>

Launch NIFI:

<http://sandbox.hortonworks.com:9090/nifi/>

OR

<http://hostname:9090/nifi/>

Configure and Start Solr:

Install using Ambari->add service\_>solr

SSH into HDP CLI

Type below commands:

1. Login as user solr

su solr

1. Create cofig file for twitter API in solr

cp -r /opt/lucidworks-hdpsearch/solr/server/solr/configsets/data\_driven\_schema\_configs /opt/lucidworks-hdpsearch/solr/server/solr/configsets/tweet\_configs

1. Edit solrconfig.xml file to recognize twitter API timestamp Data

vi /opt/lucidworks-hdpsearch/solr/server/solr/configsets/tweet\_configs/conf/solrconfig.xml

1. Add below line under processor class **solr.ParseDateFieldUpdateProcessorFactory**

<str>EEE MMM d HH:mm:ss Z yyyy</str>

<processor class="solr.ParseLongFieldUpdateProcessorFactory"/>

<processor class="solr.ParseDateFieldUpdateProcessorFactory">

<arr name="format">

<str>EEE MMM d HH:mm:ss Z yyyy</str>

<str>yyyy-MM-dd'T'HH:mm:ss.SSSZ</str>

<str>yyyy-MM-dd'T'HH:mm:ss,SSSZ</str>

<str>yyyy-MM-dd'T'HH:mm:ss.SSS</str>

<str>yyyy-MM-dd'T'HH:mm:ss,SSS</str>

<str>yyyy-MM-dd'T'HH:mm:ssZ</str>

</arr>

</processor>

</processor>

1. Upload default.json file to run BANANA dashboard for twitter data

cd /opt/lucidworks-hdpsearch/solr/server/solr-webapp/webapp/banana/app/dashboards/

mv default.json default.json.orig

wget <https://raw.githubusercontent.com/abajwa-hw/ambari-nifi-service/master/demofiles/default.json>

1. start solr cloud instance for the first time

/opt/lucidworks-hdpsearch/solr/bin/solr start -e cloud -noprompt -z localhost:2181

This will start solr on ports 8983 and 7574

Connect to solr <http://localhost:8983/solr/#/> cloud and confirm the cloud setup

Here is the link for steps for starting and stopping solr server

<https://support.lucidworks.com/hc/en-us/articles/206568297-SolrCloud-5-X-with-External-Zookeeper-3-4-6->

1. create a collection called “tweets”

/opt/lucidworks-hdpsearch/solr/bin/solr create -c tweets -d tweet\_configs -s 1 -rf 1 -p 8983

Note: Here -c indicates the name

-d is the config directory

-s is the number of shards

-rf is the replication factor

-p is the port at which Solr is running

1. exit solr

type command ‘exit’

now you are logged out of ‘solr’ user.

1. Launch SOLR UI

<http://sandbox.hortonworks.com:8983/solr/>

CREATE A TWITTER APPLICATION

Go to <https://apps.twitter.com/> twitter api website and create a new App

App name - YETI Twitter Demo

Update permission tab in the APP to ‘read only’.

Note down Keys and Access tokens either by clicking ‘testauth’ button on the app or by heading to ‘Keys and Access Tokens’ tab.

CREATE A DATA FLOW WITH NIFI

1. download NIFI template from below link and upload it template section on the right side.

<https://raw.githubusercontent.com/hortonworks/tutorials/hdp-2.3/assets/nifi-sentiment-analytics/assets/Twitter_Flow.xml>

1. Add Twitter\_Flow.xml template to the NIFI UI
2. Right click on Grab Garden Hose and click Configure to update KEYS AND ACCESS TOKENS of ‘YETI Twitter Demo’ twitter APP
3. Update Terms to filter on with keywords “YETI”.
4. Make sure all processors are showing STOP (red color). (if it shows YELLOW color on it then the configuration wasn’t set up properly)
5. Select all processors and click start button to start NIFI workflow.

Run Time ERRORS:

1. You may get errors such as collection not foung in ‘putsolrcontentstream’ processor.

You need to update the solr cloud port to 9983 instead of 2181.

OPTIONAL: GENERATING RANDOM TWEET DATA FOR HIVE AND SOLR

1. If you don’t want to create NIFI workflow, here is the sample twitter data
2. Follow below steps to generate random twitter data
   1. SSH into HDP CLI
   2. wget <https://raw.githubusercontent.com/hortonworks/tutorials/hdp/assets/nifi-sentiment-analytics/assets/twitter-gen.sh>
   3. bash twitter-gen.sh 2000 (generated 2000 tweets)
   4. Twitter data will be uploaded to /tmp/data/ folder

ANALYZE AND SEARCH DATA WITH SOLR

1. If you are using NIFI workflow, you can see the real-time data coming into solr cloud in banana dashboard.

<http://sandbox.hortonworks.com:8983/solr/banana/index.html>

1. Click on below link to heading to solr instance

<http://sandbox.hortonworks.com:8983/solr>

1. Select tweet shard under ‘core selector’ menu
   1. You can see the no.of documents coming into Solr cloud in ‘OVERVIEW” section.
   2. Head to ‘Query’ section and execute some queries to see the data flowing.
   3. Example:
      1. For q type language\_s:en
      2. For sort type screeName\_s asc
      3. For rows type 150
      4. For fl type screenName\_s, text\_t
      5. For wt choose csv

ANALYZE TWEET DATA IN HIVE

1. Stop NIFI workflow
2. SSH into HDP CLI and type below commands.

sudo -u hdfs hadoop fs -chown -R maria\_dev /tmp/tweets\_staging

sudo -u hdfs hadoop fs -chmod -R 777 /tmp/tweets\_staging

1. Login to AMBARI HIVE VIEW and type below scripts

CREATE EXTERNAL TABLE IF NOT EXISTS tweets\_text(

tweet\_id bigint,

created\_unixtime bigint,

created\_time string,

lang string,

displayname string,

time\_zone string,

msg string,

fulltext string)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY "|"

LOCATION "/tmp/tweets\_staging";

1. Download dictionary.tsv file from here

<https://raw.githubusercontent.com/hortonworks/tutorials/hdp-2.3/assets/nifi-sentiment-analytics/assets/dictionary.tsv>

1. Download time\_zone\_map.tsv file from here

<https://raw.githubusercontent.com/hortonworks/tutorials/hdp-2.3/assets/nifi-sentiment-analytics/assets/time_zone_map.tsv>

1. SSH into HDP CLI and grant access to below tables.

sudo -u hdfs hadoop fs -chmod -R 777 /tmp/data/tables

1. Run below scripts

CREATE EXTERNAL TABLE if not exists dictionary (

type string,

length int,

word string,

pos string,

stemmed string,

polarity string )

ROW FORMAT DELIMITED

FIELDS TERMINATED BY '\t'

STORED AS TEXTFILE

LOCATION '/tmp/data/tables/dictionary';

CREATE EXTERNAL TABLE if not exists time\_zone\_map (

time\_zone string,

country string,

notes string )

ROW FORMAT DELIMITED

FIELDS TERMINATED BY '\t'

STORED AS TEXTFILE

LOCATION '/tmp/data/tables/time\_zone\_map';

CREATE VIEW IF NOT EXISTS tweets\_simple AS

SELECT

tweet\_id,

cast ( from\_unixtime( unix\_timestamp(concat( '2015 ', substring(created\_time,5,15)), 'yyyy MMM dd hh:mm:ss')) as timestamp) ts,

msg,

time\_zone

FROM tweets\_text;

CREATE VIEW IF NOT EXISTS tweets\_clean AS

SELECT

t.tweet\_id,

t.ts,

t.msg,

m.country

FROM tweets\_simple t LEFT OUTER JOIN time\_zone\_map m ON t.time\_zone = m.time\_zone;

Compute sentiment for each tweet

-- Compute sentiment

create view IF NOT EXISTS l1 as select tweet\_id, words from tweets\_text lateral view explode(sentences(lower(msg))) dummy as words;

create view IF NOT EXISTS l2 as select tweet\_id, word from l1 lateral view explode( words ) dummy as word;

create view IF NOT EXISTS l3 as select

tweet\_id,

l2.word,

case d.polarity

when 'negative' then -1

when 'positive' then 1

else 0 end as polarity

from l2 left outer join dictionary d on l2.word = d.word;

create table IF NOT EXISTS tweets\_sentiment stored as orc as select

tweet\_id,

case

when sum( polarity ) > 0 then 'positive'

when sum( polarity ) < 0 then 'negative'

else 'neutral' end as sentiment

from l3 group by tweet\_id;

CREATE TABLE IF NOT EXISTS tweetsbi

STORED AS ORC

AS SELECT

t.\*,

case s.sentiment

when 'positive' then 2

when 'neutral' then 1

when 'negative' then 0

end as sentiment

FROM tweets\_clean t LEFT OUTER JOIN tweets\_sentiment s on t.tweet\_id = s.tweet\_id;

1. Run below command to see the sentiment of each tweet.

Select \* from tweetsbi LIMIT 100;

VISUALIZE SENTIMENT WITH ZEPPELIN

1. Launch Zeppelin server

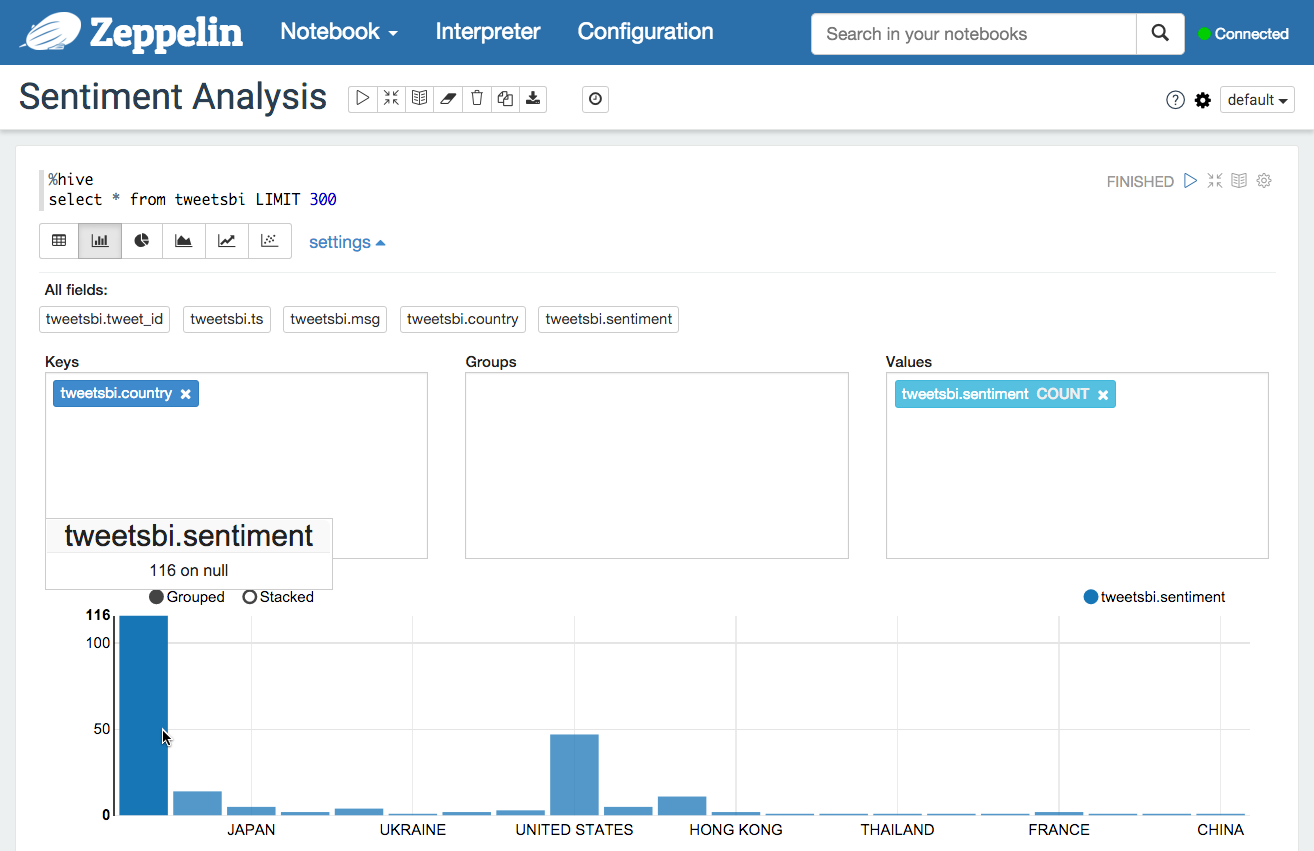
<http://sandbox.hortonworks.com:9995>

1. Create a new notebook
2. Bind hive interpreter to the notebook
3. Run below command

%hive

select \* from tweetsbi LIMIT 300

1. Select bar graph and go to settings and select country as KEY and sentiment as value to visualize



1. Try some other queries as per business requirements.

%hive

select \* from tweetsbi where country != "null" LIMIT 500

%hive

select lang, time\_zone from tweets\_text LIMIT 1000

%hive

select sentiment, count(country), country from tweetsbi group by sentiment, country having country != "null"

