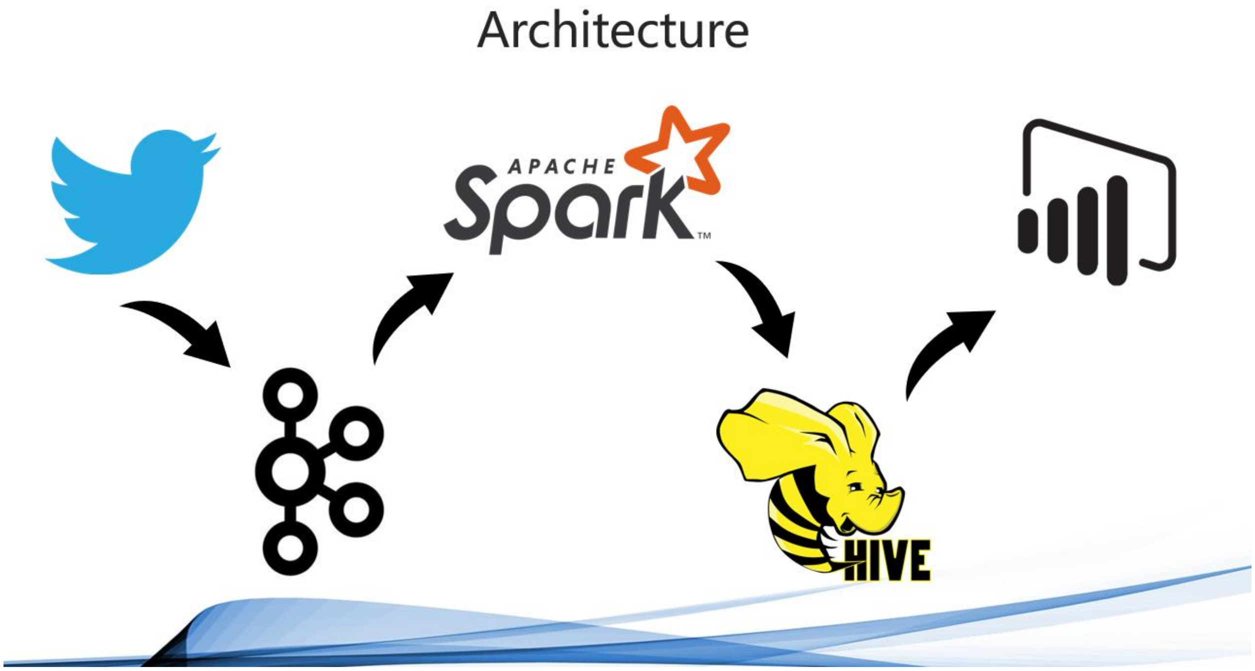
Introduction:

- The objective of this tool is to capture tweets and their data which were written with specific hashtag then doing sentiment analysis on it to figure out whether it is good, bad, or natural. Then reply to those Tweets and save data to use in building dashboards.
- In our case here let's assume that we have a movie realized four days ago we want to gather tweets with its name as hashtag and use it to answer business questions.

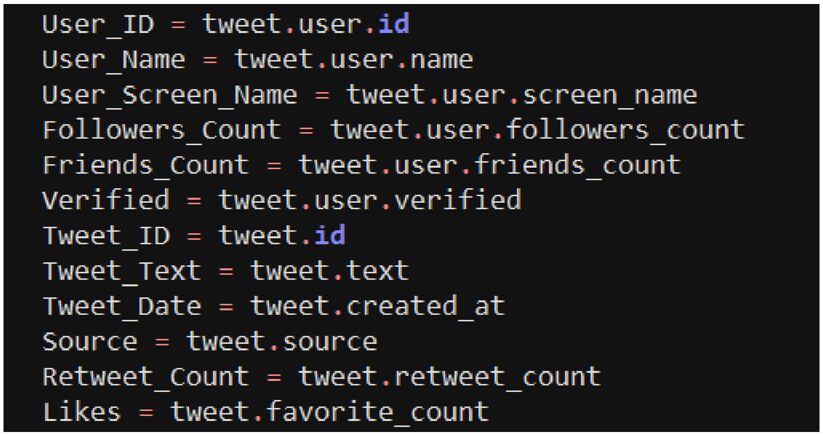
Architecture:

- The tool was built using the following architecture.
- Twitter API was used to capture data from twitter and write it to Kafka topic using Kafka Producer. Then Spark was used to consume the data using Kafka consumer and save it as parquet files to HDFS directory. Hive table was created on top of the HDFS directory to retrieve the data and to connect Power BI to that table to be able to make dashboards and answer questions.

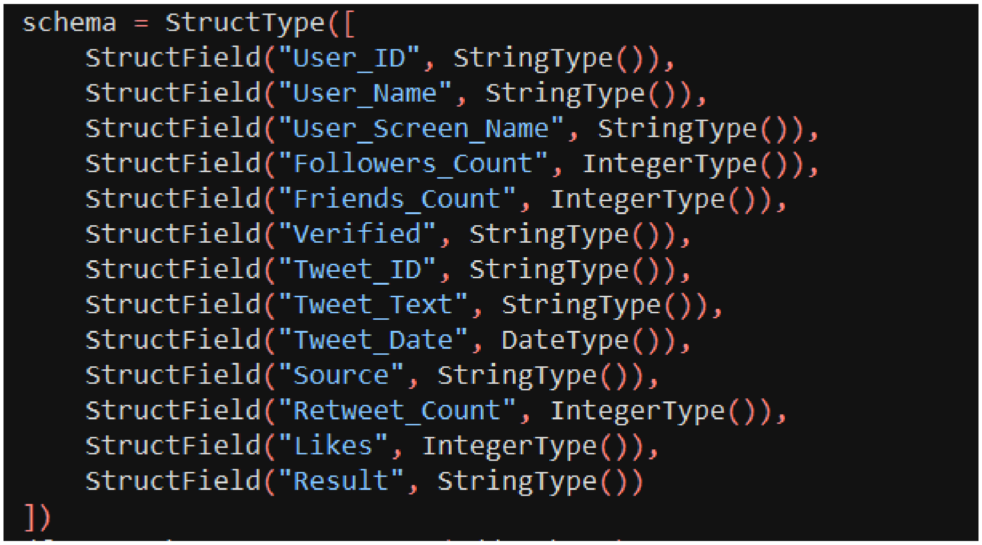


Data captured from Twitter:

The following data was captured from twitter and was used to answer business questions that will be mentioned later with dashboard section.



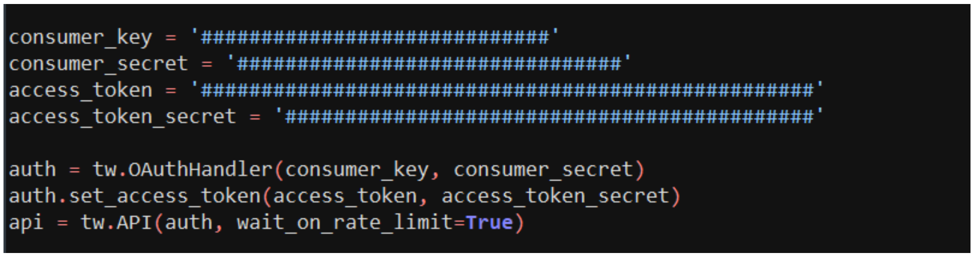
```
User_ID = tweet.user.id
User_Name = tweet.user.name
User_Screen_Name = tweet.user.screen_name
Followers_Count = tweet.user.followers_count
Friends_Count = tweet.user.friends_count
Verified = tweet.user.verified
Tweet_ID = tweet.id
Tweet_Text = tweet.text
Tweet_Date = tweet.created_at
Source = tweet.source
Retweet_Count = tweet.retweet_count
Likes = tweet.favorite_count
```



```
schema = StructType([
    StructField("User_ID", StringType()),
    StructField("User_Name", StringType()),
    StructField("User_Screen_Name", StringType()),
    StructField("Followers_Count", IntegerType()),
    StructField("Friends_Count", IntegerType()),
    StructField("Verified", StringType()),
    StructField("Tweet_ID", StringType()),
    StructField("Tweet_Text", StringType()),
    StructField("Tweet_Date", DateType()),
    StructField("Source", StringType()),
    StructField("Retweet_Count", IntegerType()),
    StructField("Likes", IntegerType()),
    StructField("Result", StringType())
])
```

Configurations:

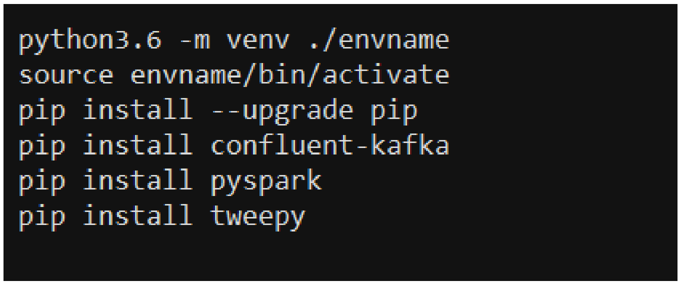
To be able to connect to twitter and capture data you must create Twitter Developer account and then create an app on your account so you can get your four keys to authorize your access. Then you will be able to connect to Twitter API with tweepy package as example and use its different functions with Twitter.



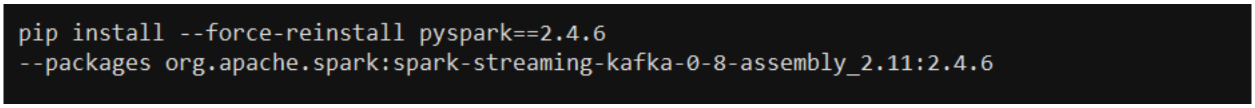
```
consumer_key = '#############################'
consumer_secret = '################################'
access_token = '###################################################'
access_token_secret = '############################################'

auth = tw.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)
api = tw.API(auth, wait_on_rate_limit=True)
```

Scripts was tested on Hortonworks Sandbox HDP 2.6.5. Virtual environment was created using the following steps to install the right version from Python and Spark. And to integrate Kafka and Spark the mentioned jar in the screen was given to –packages option while running Spark script.



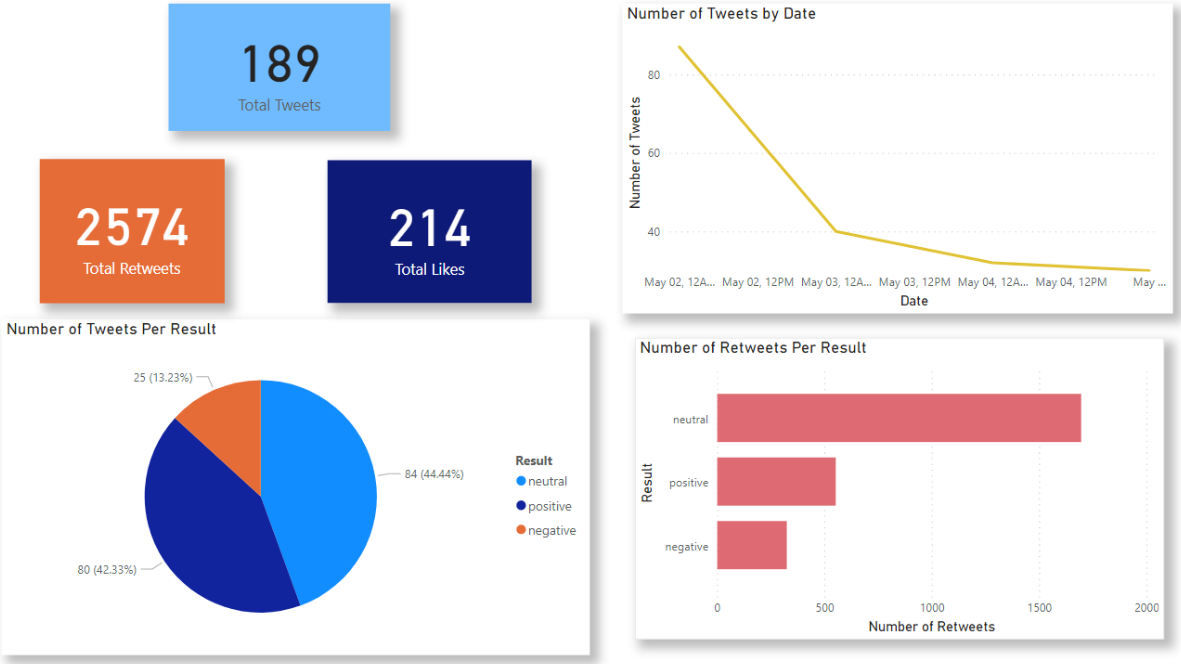
```
python3.6 -m venv ./envname
source envname/bin/activate
pip install --upgrade pip
pip install confluent-kafka
pip install pyspark
pip install tweepy
```



```
pip install --force-reinstall pyspark==2.4.6
--packages org.apache.spark:spark-streaming-kafka-0-8-assembly_2.11:2.4.6
```

Business questions and Dashboard:

1. What is total number of tweets about our movie?
2. What is total number of likes on tweets about our movie?
3. What is total number of retweets on tweets about our movie?
4. The number of tweets with each result of sentiment analysis whether tweets talking about the movie in good or bad way?
5. The number of retweets with each type (positive, negative, neutral) of tweets?
6. The number of tweets over the last four days?


Total Tweets
Total Retweets
Total Likes
Number of Tweets by Date
Number of Tweets
May 02, 12A...
May 02, 12PM
May 03, 12A...
May 03, 12PM
May 04, 12A...
May 04, 12PM
May ...
Date
Number of Tweets Per Result
25 (13.23%)
84 (44.44%)
80 (42.33%)
Result
● neutral
● positive
● negative
Number of Retweets Per Result
Result
neutral
positive
negative
Number of Retweets