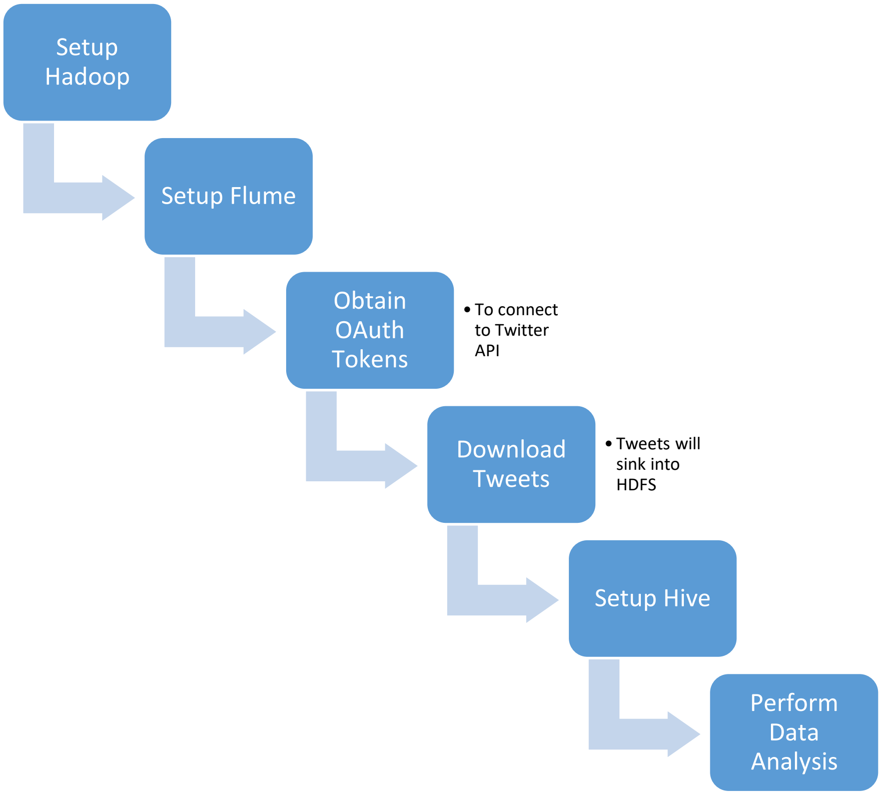
Twitter Data Analysis

Twitter, one of the largest social media site receives millions of tweets every day on variety of important issues. This huge amount of raw data can be used for industrial, Social, Economic, Government policies or business purpose by organizing according to our requirement and processing. Hadoop is one of the best tool options for twitter data analysis as it works for distributed Big Data, Streaming Data, Time Stamped Data, Text Data etc.

This project will discuss how to use FLUME and HIVE for twitter data analysis.

- FLUME is used to extract real time twitter data into HDFS
- Hive which is SQL like query language is used for some extraction and analysis

Steps to be followed to accomplish the project are as below:



Steps in Detail:

1. Hadoop should be already setup
2. Download and extract Flume to the HadoopInstallations directory
3. Generate the keys by creating a twitter application on https://apps.twitter.com/

4. Download the flume-sources-1.0-SNAPSHOT.jar from http://files.cloudera.com/samples/flumesources-1.0-SNAPSHOT.jar (Note: The jar contains the java classes to pull the Tweets and save them into HDFS. This jar file is available in /home/user1/Downloads/07_Packages directory)
5. Copy the ‘flume-env.sh’ and ‘twitter.conf’ files to /home/user1/HadoopInstallations/apache-flume-1.6.0-bin/conf
6. The conf/twitter.conf should have all the agents (source, channel and sink) as defined below:

   ```
   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 = <consumerKey>
   TwitterAgent.sources.Twitter.consumerSecret = <consumerSecret>
   TwitterAgent.sources.Twitter.accessToken = <accessToken>
   TwitterAgent.sources.Twitter.accessTokenSecret = <accessTokenSecret>
   TwitterAgent.sources.Twitter.keywords = hadoop, big data, analytics, bigdata, cloudera, data science, data scientist, business intelligence, mapreduce
   TwitterAgent.sinks.HDFS.channel = MemChannel
   TwitterAgent.sinks.HDFS.type = hdfs
   TwitterAgent.sinks.HDFS.hdfs.path = hdfs://localhost:9000/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
   ```

7. Start flume agent using the below command:

   ```
   flume-ng agent -n TwitterAgent -c conf -f /home/user1/HadoopInstallations/apache-flume-1.6.0-bin/conf/twitter.conf
   ```

   After a couple of minutes, Tweets should appear in HDFS

8. Download and extract Hive
9. Download hive-serdes-1.0-SNAPSHOT.jar from http://files.cloudera.com/samples/hive-serdes-1.0-SNAPSHOT.jar to the lib directory in Hive. (Note: Twitter returns Tweets in the JSON format and this library will help Hive understand the JSON format. This jar file is available in /home/user1/Downloads/.05_Programs/10_Project1/ directory)
10. Start the Hive shell using the hive command and register the hive-serdes-1.0-SNAPSHOT.jar file

    ```
    ADD JAR /home/user1/Downloads/.05_Programs/10_Project1/hive-serdes-1.0-SNAPSHOT.jar
    ```

11. Hive correction (More details: https://issues.apache.org/jira/browse/HIVE-10294)

    ```
    set hive.support.sql11.reserved.keywords=false
    ```

12. Create the tweets table in Hive

    ```
    CREATE EXTERNAL TABLE tweets (
    id BIGINT,
    created_at STRING,
    source STRING,
    favorited BOOLEAN,
    retweet_count INT,
    retweeted_status STRUCT<
    text:STRING,
    user:STRUCT<screen_name:STRING,name:STRING>>,
    entities STRUCT<
    urls:ARRAY<STRUCT<expanded_url:STRING>>,
    user_mentions:ARRAY<STRUCT<screen_name:STRING,name:STRING>>,
    hashtags:ARRAY<STRUCT<text:STRING>>>,
    text STRING,
    user STRUCT<
    screen_name:STRING,
    name:STRING,
    friends_count:INT,
    followers_count:INT,
    statuses_count:INT,
    verified:BOOLEAN,
    utc_offset:INT,
    time_zone:STRING>,
    in_reply_to_screen_name STRING
    )
    ROW FORMAT SERDE 'com.cloudera.hive.serde.JSONSerDe'
    LOCATION '/tweets';
    ```

13. To know which user has the most number of followers, the below query helps

```
select user.screen_name, user.followers_count c from tweets order by c
desc limit 10;
```

14. To know the most influential person, the below query helps

```
SELECT t.retweeted_screen_name, sum(retweets) AS total_retweets,
count(*) AS tweet_count FROM (SELECT retweeted_status.user.screen_name
as retweeted_screen_name, retweeted_status.text, max(retweet_count) as
retweets FROM tweets GROUP BY retweeted_status.user.screen_name,
retweeted_status.text) t GROUP BY t.retweeted_screen_name ORDER BY
total_retweets DESC LIMIT 10;
```

More information: https://github.com/cloudera/cdh-twitter-example