Kalyan Big Data Projects – Project 2
How To Stream Twitter Data Into Hadoop in JSON format Using Apache Flume

Pre-Requisites of Flume Project:

hadoop-2.6.0
flume-1.6.0
java-1.7

NOTE: Make sure that install all the above components

Flume Project Download Links:

`hadoop-2.6.0.tar.gz` ==> link
(https://archive.apache.org/dist/hadoop/core/hadoop-2.6.0/hadoop-2.6.0.tar.gz)

`apache-flume-1.6.0-bin.tar.gz` ==> link
(https://archive.apache.org/dist/flume/1.6.0/apache-flume-1.6.0-bin.tar.gz)

`kalyan-twitter-hdfs-agent.conf` ==> link
(https://github.com/kalyanhadooptraining/kalyan-bigdata-realtime-projects/blob/master/flume/project2-twitter-hadoop-json/kalyan-twitter-hdfs-agent.conf)

`kalyan-flume-project-0.1.jar` ==> link
(https://github.com/kalyanhadooptraining/kalyan-bigdata-realtime-projects/blob/master/kalyan/kalyan-flume-project-0.1.jar)

Learnings of this Project:

- We will learn Flume Configurations and Commands
- Flume Agent
  1. Source (Twitter Source)
  2. Channel (Memory Channel)
  3. Sink (Hdfs Sink)
- Major project in Real Time `Social Media (Twitter) Sentiment Analysis`
  1. We are extracting the data from twitter using twitter api credentials
  2. This data will be useful to do setiment analysis on twitter tweets
  3. JSON is the output format
- We can use hive / pig / mapreduce to analyze this data
  1. explore hive query to analysis
  2. explore pig scripts to analysis
  3. explore mapreduce to analysis

---

1. create "**kalyan-twitter-hdfs-agent.conf**" file with below content

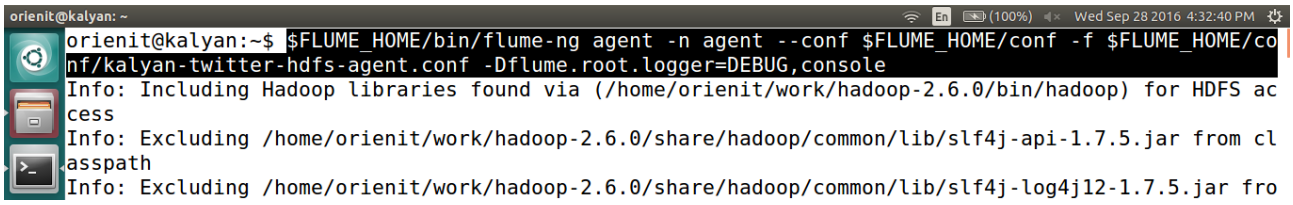
```
agent.sources = Twitter
agent.channels = MemChannel
agent.sinks = HDFS

agent.sources.Twitter.type = com.orienit.kalyan.flume.source.KalyanTwitterSource
agent.sources.Twitter.channels = MemChannel
agent.sources.Twitter.consumerKey = ********
agent.sources.Twitter.consumerSecret = ********
agent.sources.Twitter.accessToken = ********
agent.sources.Twitter.accessTokenSecret = ********
agent.sources.Twitter.keywords = hadoop, big data, analytics, bigdata, cloudera, data science, data scientiest, business intelligence, mapreduce, data warehouse, data warehousing, mahout, hbase, nosql, newsql, businessintelligence, cloudcomputing

agent.sinks.HDFS.type = hdfs
agent.sinks.HDFS.channel = MemChannel
agent.sinks.HDFS.hdfs.path = hdfs://localhost:8020/user/flume/tweets
agent.sinks.HDFS.hdfs.fileType = DataStream
agent.sinks.HDFS.hdfs.writeFormat = Text
agent.sinks.HDFS.hdfs.batchSize = 100
agent.sinks.HDFS.hdfs.rollSize = 0
agent.sinks.HDFS.hdfs.rollCount = 100
agent.sinks.HDFS.hdfs.useLocalTimeStamp = true

agent.channels.MemChannel.type = memory
agent.channels.MemChannel.capacity = 1000
agent.channels.MemChannel.transactionCapacity = 100
```

2. Copy "**kalyan-twitter-hdfs-agent.conf**" file into "**$FUME_HOME/conf**" folder

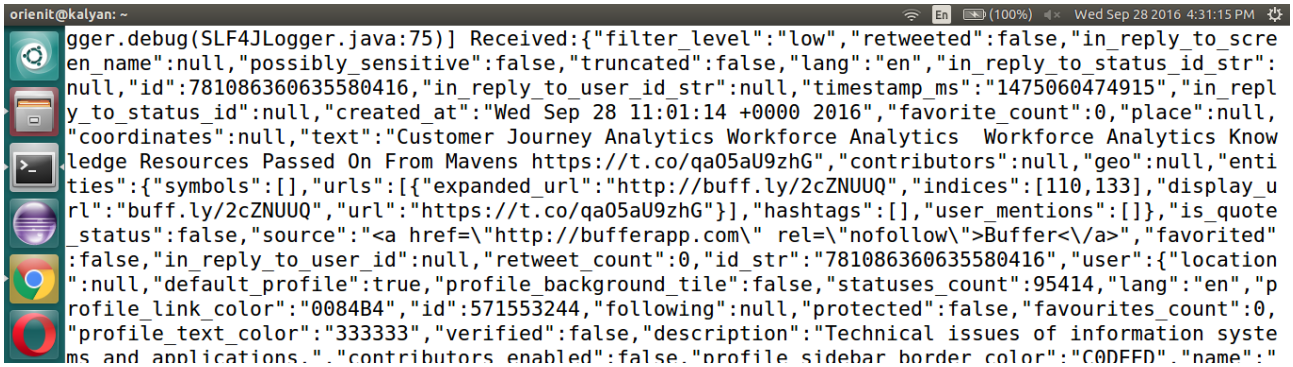
3. Copy "**kalyan-flume-project-0.1.jar**" file into "**$FLUME_HOME/lib**" folder

4. Execute the below command to `**Extract data from Twitter into Hadoop using Flume**`

```
$FLUME_HOME/bin/flume-ng agent -n agent --conf $FLUME_HOME/conf -f
$FLUME_HOME/conf/kalyan-twitter-hdfs-agent.conf -Dflume.root.logger=DEBUG,console
```



5. Verify the data in console



6. Verify the data in hdfs location is "**hdfs://localhost:8020/user/flume/tweets**"