Session 12:

Oozie and Flume

Create a flume agent that streams data from Twitter and stores in the HDFS.

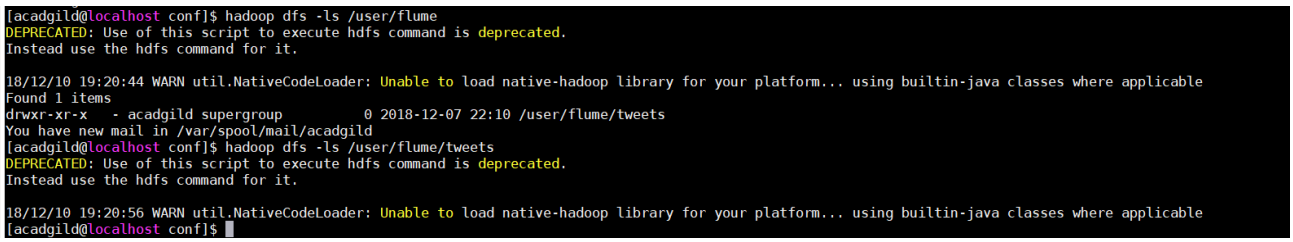
Twitter didn't provide access for developer account and I didn't get application token that will used for flume config file. I wrote the same thing to acadgild and they told me to share the procedure how to stream data and save into HDFS. Ticket number #17531

Steps / Procedure

1. For streaming the data we need to have a twitter account and Hadoop cluster
2. Log into a developer account and create a new aaplication
3. You will get below keys
   TwitterAgent.sources.Twitter.consumerKey
   TwitterAgent.sources.Twitter.consumerSecret
   TwitterAgent.sources.Twitter.accessToken-
   TwitterAgent.sources.Twitter.accessTokenSecret
4. You need install flume for us it is already present in VM
5. Create an empty file and make sure all the twitter jars are present in $FLUME_HOME/lib folder
6. The is config file attached at end and paste content into the empty file that you created in step 5
7. You need to define some key words like below which data need to be collected TwitterAgent.sources.Twitter.keywords=hadoop, bigdata, mapreduce, mahout, hbase, nosql

8. Open the terminal and start Hadoop using the start-all.sh, this will start [name mode, data node, secondary name node, Resource Manager, Node Manager]. Check by using jps commands whether all daemons are started

9. Create a new directory inside HDFS path, where the Twitter tweet data should be stored.
   Hadoop dfs –mkdir –p /user/flume/tweets



10. For fetching data from Twitter, Use the below command to fetch the twitter tweet data into the HDFS cluster path.
    flume-ng agent -n TwitterAgent -f /home/acadgild/install/flume/apache-flume-1.8.0-bin/conf/flume.conf
11. Once, the tweet data started streaming it into the given HDFS path we can use 'Ctrl+c' command to stop the streaming process.
12. To check the contents of the tweet data we can use the following command:
    hadoop dfs –ls /user/flume/tweets
13. We can use the 'cat' command to display the tweet data inside the /user/flume/tweets/FlumeData.145* path.
    hadoop dfs –cat /us er/flume/tweets/(file name generated)*

Config file

[acadgild@localhost conf]$ cat flume.conf

tterAgent.sources = Twitter

TwitterAgent.channels = MemChannel

TwitterAgent.sinks = HDFS

Describing/Configuring the source

TwitterAgent.sources.Twitter.type = org.apache.flume.source.twitter.TwitterSource

TwitterAgent.sources.Twitter.consumerKey

TwitterAgent.sources.Twitter.consumerSecret

TwitterAgent.sources.Twitter.accessToken

TwitterAgent.sources.Twitter.accessTokenSecretTwitterAgent.sources.Twitter.keywords=hadoop, bigdata, mapreduce, mahout, hbase, nosql

Describing/Configuring the sink

TwitterAgent.sources.Twitter.keywords= hadoop,election,sports, cricket,Big data

TwitterAgent.sinks.HDFS.channel=MemChannel

TwitterAgent.sinks.HDFS.type=hdfs

TwitterAgent.sinks.HDFS.hdfs.path=hdfs://localhost:9000/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.sinks.HDFS.hdfs.rollInterval=600

TwitterAgent.channels.MemChannel.type=memory
TwitterAgent.channels.MemChannel.capacity=10000
TwitterAgent.channels.MemChannel.transactionCapacity=1000

TwitterAgent.sources.Twitter.channels = MemChannel
TwitterAgent.sinks.HDFS.channel = MemChannel
```