CS-5433 – Big Data Management

Assignment-1

Part-1: Twitter Data Collection Using Flume

Note: Due Date for Assignment will be posted soon

Start with Part-1 of the Assignment below are the Pre-Requisites:

Follow steps to login to Hadoop and Collect Twitter data using Flume.

Procedure to create an account in CSX

If you are taking a CS prefix course, you already have an account.

To get an initial password:

1. Login to https://cs.okstate.edu/pwreset to set the password for CSX.

Download Putty:

Windows users use the below link to download PUTTY.

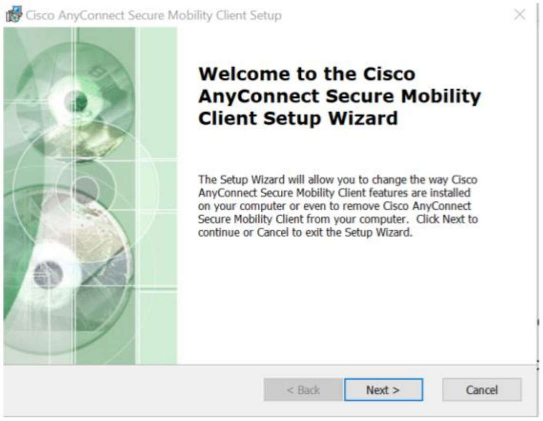
https://www.chiark.greenend.org.uk/~sgtatham/putty/latest.html

Connecting from outside the university:

1. Login to https://osuvpn.okstate.edu/+CSCOE+/logon.html with OKEY username and password

2. You will be redirected to the below screenshot go to download section and download AnyConnectVPN

3. Connect every time when you want to use the servers in CS Department.



Cisco AnyConnect Secure Mobility Client Setup

Welcome to the Cisco AnyConnect Secure Mobility Client Setup Wizard

The Setup Wizard will allow you to change the way Cisco AnyConnect Secure Mobility Client features are installed on your computer or even to remove Cisco AnyConnect Secure Mobility Client from your computer. Click Next to continue or Cancel to exit the Setup Wizard.

4. After Connecting open the putty and in Hostname give hadoop1.cs.okstate.edu for login cluster, set port as 22 and connection Type as SSH.

Reference: https://computerscience.okstate.edu/loggingon

Login to Hadoop1:

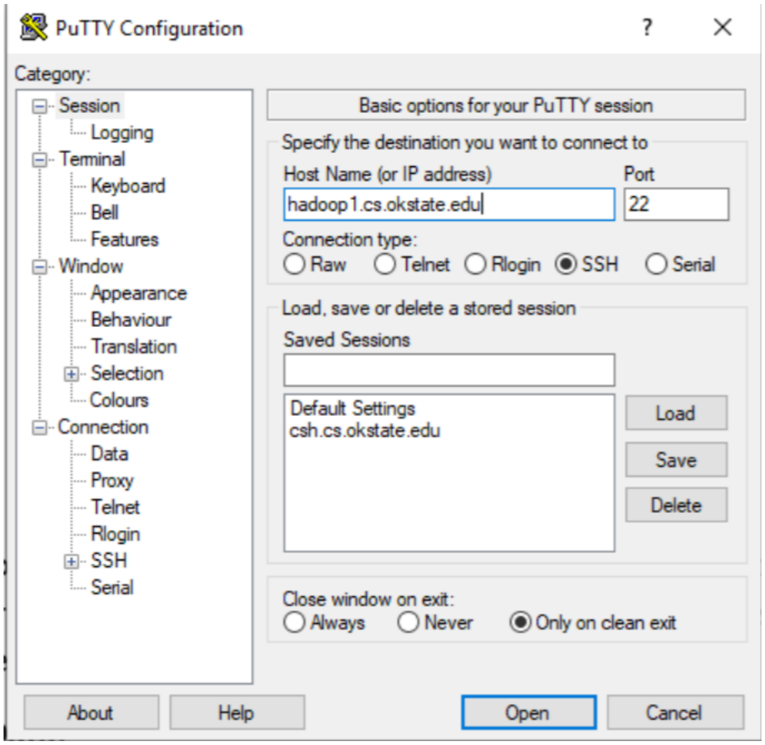
1. For windows users Login with your username (CSX username) to hadoop1.cs.okstate.edu from putty.

2. MAC/LINUX users type ssh user@hadoop1.cs.okstate.edu from your terminal.

3. Enter your CSX password when asked.

4. You will have to access only the folder /user_name in HDFS.

5. You can use Hadoop in the cluster.



PuTTY Configuration

Basic options for your PuTTY session

Specify the destination you want to connect to

Host Name (or IP address): hadoop1.cs.okstate.edu

Port: 22

Connection type: ○ Raw ○ Telnet ○ Rlogin ● SSH ○ Serial

Load, save or delete a stored session

Saved Sessions:

Default Settings
csh.cs.okstate.edu

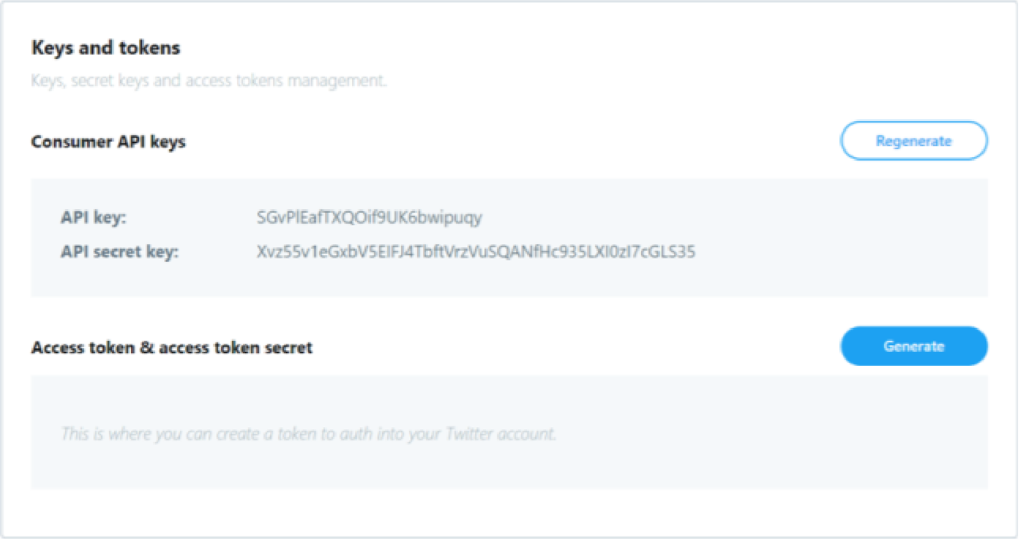
Close window on exit: ○ Always ○ Never ● Only on clean exit

Apache Flume

Apache Flume is an open-source software that helps to store the streaming data into HDFS. A flume agent should be created through which we can stream the data. The following steps describe the process to collect the twitter data.

Process to collect Twitter Data:

1. Create an account in twitter and login with the credentials.
2. Navigate to https://apps.twitter.com/ and create a new app.
3. Give the Name and the Description on what are you the data, and the website URL is https://twitter.com/
4. After the developer account is approved, create an app and get the consumer secret, consumer token, access token and access token secret for your application. To get the consumer secret, consumer token:
   (a)Go to Keys and Access Tokens get the consumer key and consumer secret, and then generate Your Access tokens and get the access tokens and access token secret.



Keys and tokens

Keys, secret keys and access tokens management.

Consumer API keys

API key: SGvPlEafTXQOif9UK6bwipuqy

API secret key: Xvz55v1eGxbV5ElFJ4TbftVrzVuSQANfHc935LXl0zI7cGLS35

Access token & access token secret

This is where you can create a token to auth into your Twitter account.

There are 3 components for a twitter agent namely source, sink and channel.

The flume source connects to Twitter API and receives data in JSON format which in turn stored into HDFS.

Now, create a configuration file for the flume agent by specifying the consumer key, consumer secret, access token, access token secret, keywords and HDFS path.

A sample configuration file with file extension .conf is shown below. It shows all the keys and keywords to be used to collect the twitter data.

Configuration File:

```
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 =
TwitterAgent.sources.Twitter.consumerSecret =
TwitterAgent.sources.Twitter.accessToken =
TwitterAgent.sources.Twitter.accessTokenSecret =
TwitterAgent.sources.Twitter.keywords = Mention keywords here
TwitterAgent.sinks.HDFS.channel = MemChannel
TwitterAgent.sinks.HDFS.type = hdfs
TwitterAgent.sinks.HDFS.hdfs.path = hdfs://hadoop1:9000/nmeka/Food_data/%Y/%m/%d/%H
TwitterAgent.sinks.HDFS.hdfs.fileType = DataStream
TwitterAgent.sinks.HDFS.hdfs.writeFormat = Text
TwitterAgent.sinks.HDFS.hdfs.batchSize = 100
TwitterAgent.sinks.HDFS.hdfs.rollSize = 0
TwtterAgent.sinks.HDFS.hdfs.rollCount = 0
TwitterAgent.channels.MemChannel.type = memory
TwitterAgent.channels.MemChannel.capacity = 10000
TwitterAgent.channels.MemChannel.transactionCapacity = 10000
```

Command to start the flume agent:

```
nohup $FLUME_HOME/bin/flume-ng agent -n TwitterAgent -f Configuration File Path &
```

Example: nohup $FLUME_HOME/bin/flume-ng agent -n TwitterAgent -f /autohome/nmeka/bitcoin.conf &

nohup will make sure the data collection process runs continuously at the backend. nohup.out file is the log file that will be created as we start the process. The data collected will be in JSON format.

Command to check the count of the files.
hdfs dfs –count /username/(folder name) there is a limit on the number of files. No more data is put into files when the limit is reached.

Reference:
https://acadgild.com/blog/streaming-twitter-data-using-flume/
https://www.tutorialspoint.com/apache_flume/fetching_twitter_data.htm

Part-2 will be posted soon