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

Apache Flume is a reliable service used for efficiently collecting, aggregating, and moving large amounts of log data.

Flume allows the user to do the following:

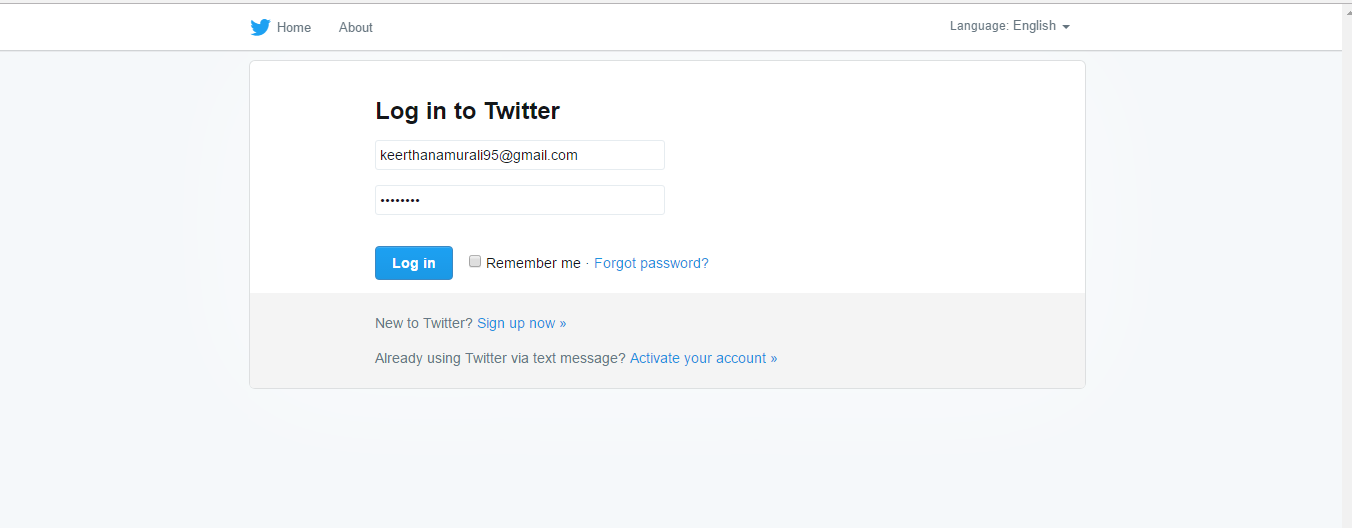
* Stream data into Hadoop from multiple sources for analysis.
* Collect high-volume web logs in real-time.
* It acts as a buffer when the rate of incoming data exceeds the rate at which the data can be written. Thereby preventing data loss.
* Guarantees data delivery.
* Scales horizontally (connects commodity system in parallel) to handle additional data volume.

There are 3 major components to get data from any live streaming source , namely: Source, Channel, and Sink, which are involved in ingesting data, moving data and storing data, respectively.

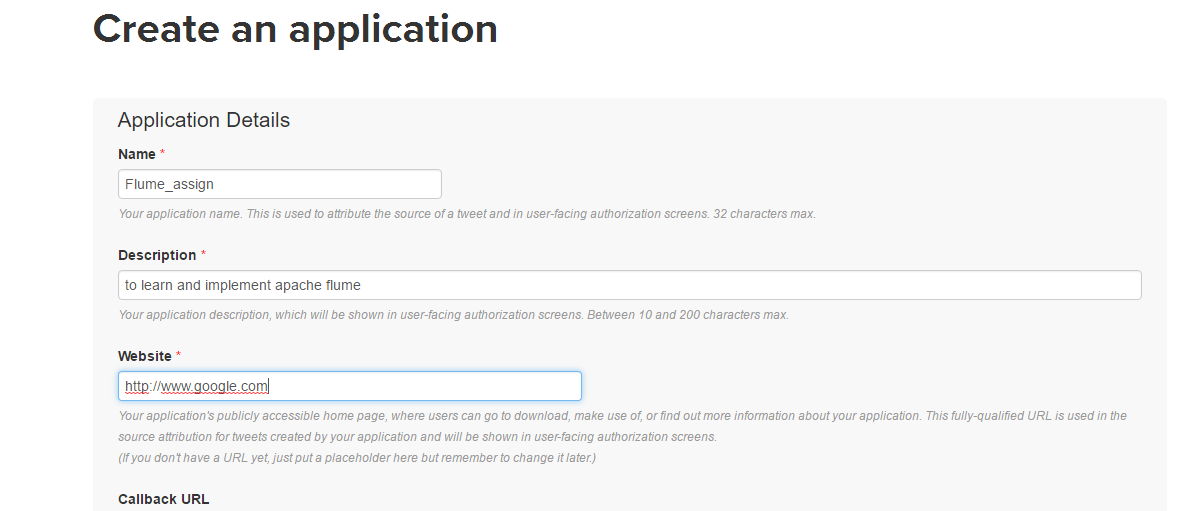
Subscribe, It’s Free!

### Step-by-step Tutorial: Data Streaming from Twitter to HDFS

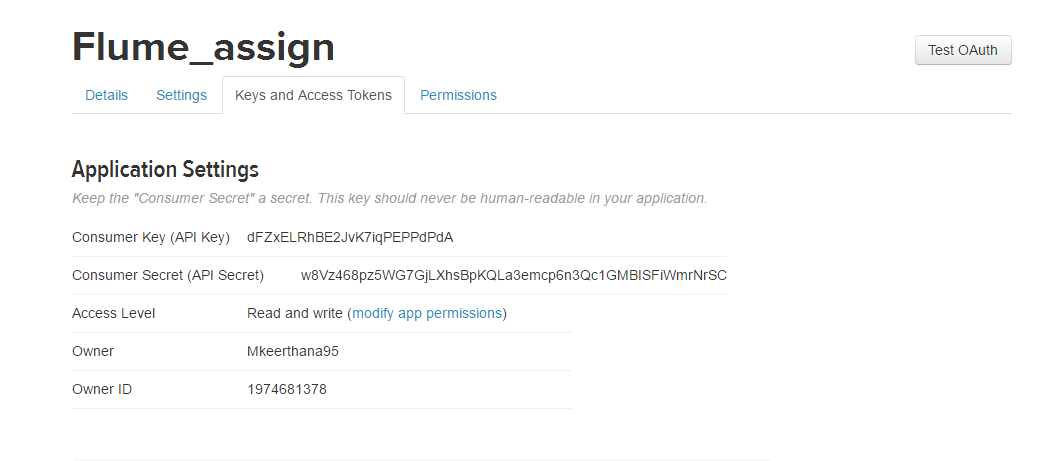
Open a Twitter account

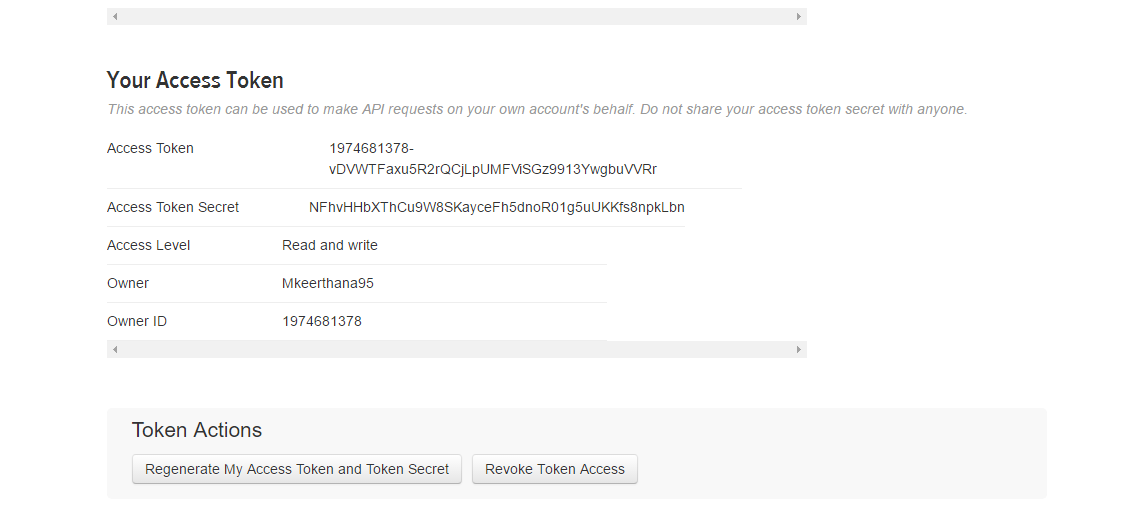


create an twitter app

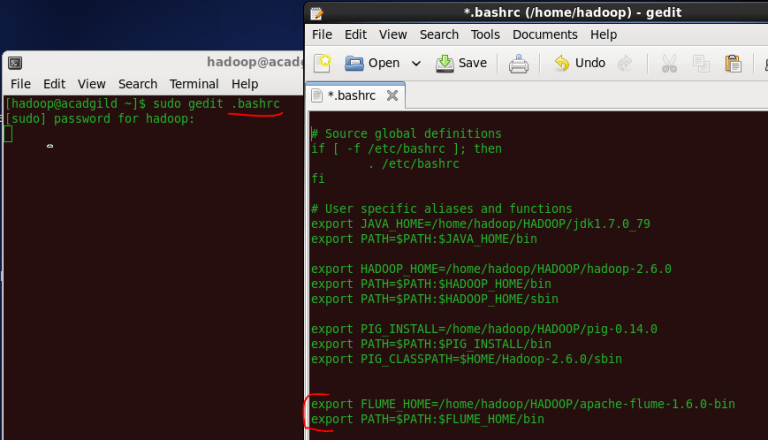


View

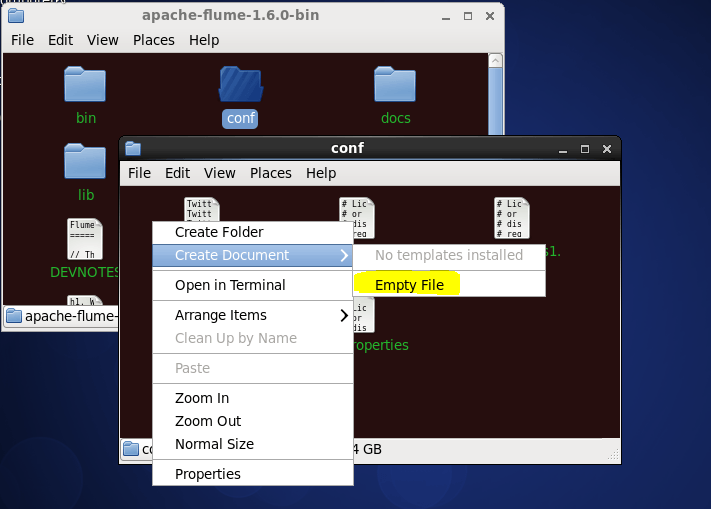


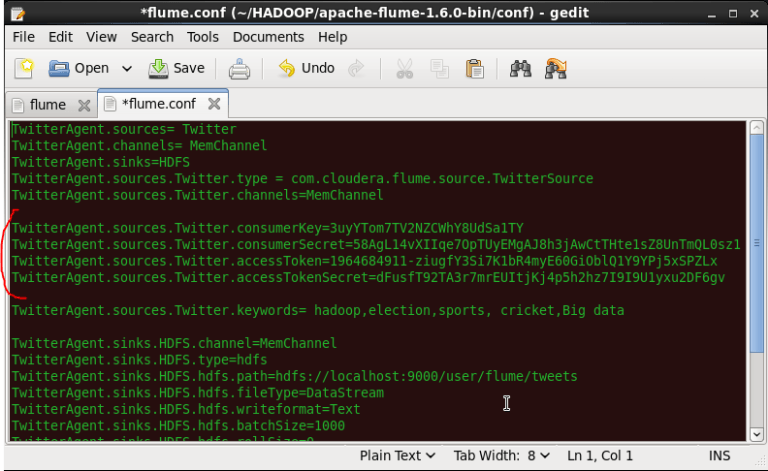


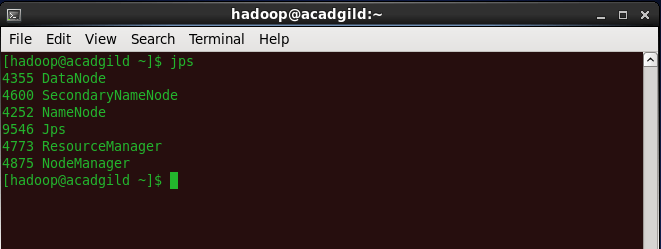
EXTRACT THE FILE AND UPDATE IT

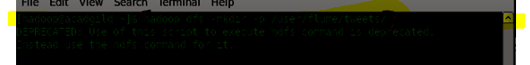


Create a new file inside the ‘conf’ directory inside the Flume-extracted directory.



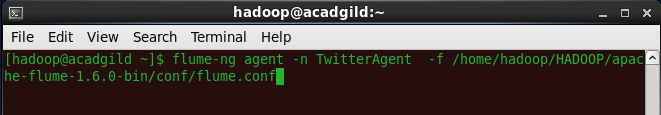


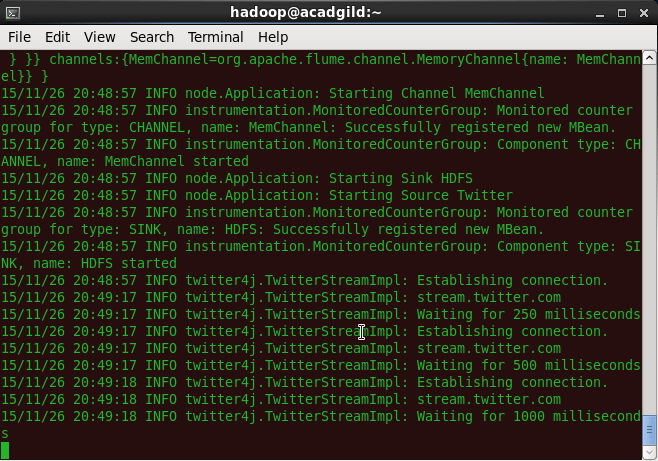




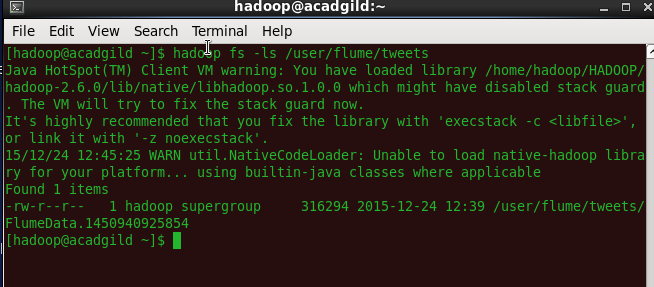
FETCHING DATA FROM TWITTER

Press Ctrl+C to stop fetching else it is an unending process





hadoop dfs –ls /user/flume/tweets



hadoop dfs –cat /us er/flume/tweets/<flumeData file name>

