All the required scripts, documentations and necessary JAR files have been uploaded to a GitHub repository.

URL - <https://github.com/hexabig/sentiment.git>

Username – hexabig

Repository – sentiment

Password – \* to be shared \*

**Contents in GitHub Repository –**

* ***Dictionary*** (folder) – create TwitterData directory in HDFS and put the folder in that directory.
* ***Time\_zone\_map*** (folder) - create TwitterData directory in HDFS and put the folder in that directory.
* ***Oozie job for adding partitions to Twitter data.docx*** – step by step documentation for executing Oozie workflow.
* ***Twitter Data Ingestion Using Flume.docx***­- step by step documentation for Data Ingestion using Flume.
* ***TwitterDDLV1*.*2.sql***– contains all the Hive DDL scripts that needs to be executed once the data is ingested.
* ***TwitterFlume-Source-T4J JAR*** *–* This needs to be copied into */usr/hdp/2.3.2.0-2950/flume/lib.*
* ***TwitterFlume.conf*** *–* Flume configuration file
* ***Json-serde-1.1.6-SNAPSHOT-jar-with-dependencies.jar*** *–* This needs to be copied into a /tmp folder in HDFS and then execute in HIVE.

**Steps required**

1. Data Ingestion - using Flume
2. Data Transformation – Hive DDL Scripts using HIVE
3. Hive DDL scripts automation – Oozie

Download the HDP 2.3.2 Sanbox using [HDP 2.3.2 Sandbox](http://hortonworks.com/hdp/downloads/) before you begin with Data Ingestion.

Once you have downloaded the sandbox, please make sure you have maven and java installed in your sandbox. To check that type:

* *mvn –verison*
* java –verison If they are not installed please follow the instructions in Twitter Data Ingestion Using Flume.docx.

You have to copy the ***TwitterFlume-Source-T4J JAR*** file into */usr/hdp/2.3.2.0-2950/flume/lib* by using the following command:

* *cp TwitterFlume-Source-T4J /usr/hdp/2.3.2.0-2950/flume/lib*

Also, it is necessary to update your system time according to the flume service so as to avoid any from Twitter in later steps (for Ex – during oozie workflow where you provided date as parameters).

To do this we use the NTP (Network Time Protocol), which is the time protocol that most servers in the world use to keep in sync.

Run the below steps in your terminal to install NTP.

* *yum install -y ntp*
* *service ntpd stop*
* *ntpdate pool.ntp.org*
* *service ntpd start*

Please follow the instructions on the Twitter Data Ingestion Using Flume.docx document for detailed instructions on Flume Ingestion. (**NOTE: The credentials for Twitter consumerkey, consumersecret, access and token are already included in TwitterFlume.conf. Do not alter the conf. file**).

Once you are done with Data Ingestion and executing the Hive scripts, we have to copy a few JAR files from your local to oozie share lib directory to make oozie work.

Copy Atlas jars from Atlas directory in your sandbox to HDFS.

* hdfs dfs -put /usr/hdp/2.3.0.0-2557/atlas/hook/hive/\* /user/oozie/share/lib/lib\_20150819123609/hive/
* hdfs dfs -put /usr/hdp/2.3.0.0-2557/atlas/bridge/hive/\* /user/oozie/share/lib/lib\_20150819123609/hive/

There should be 256 jars in total in the directory. (NOTE: ***2.3.0.0-2557*** is the version of your sandbox. Please check before running the above command.)

Edit hive-env.sh file from /etc/hive/2.3.0.0-2557/0/hive-env.sh and add the line “***export HIVE\_AUX\_JARS\_PATH=<atlas package>/hook/hive***”.

* To edit the hive-env.sh file type ***vi /etc/hive/2.3.0.0-2557/0/hive-env.sh***

Add the json serde (mentioned in GitHub) to the oozie share lib which adds it a total to 257 jars in the folder.

* ***hdfs dfs –put json-serde-1.1.6-SNAPSHOT-jar-with-dependencies.jar /user/oozie/share/lib/lib\_20150819123609/hive/***