**INSTALLATION:**

# **IMPORTANT** *- Place all the attached project files in your home folder (/home/<user>) or other preferred location*

**A. Elasticsearch**

*# download and install Elasticsearch*

1. sudo apt install curl
2. wget <https://artifacts.elastic.co/downloads/elasticsearch/elasticsearch-6.7.0.deb>
3. wget <https://artifacts.elastic.co/downloads/elasticsearch/elasticsearch-6.7.0.deb.sha512>
4. shasum -a 512 -c elasticsearch-6.7.0.deb.sha512
5. sudo dpkg -i elasticsearch-6.7.0.deb

**B. Confluent Hub**

*# You can download the confluent hub tar from the shared drive link*

*# look for - confluent-5.2.1-2.12.tar.gz*

1. <https://drive.google.com/drive/folders/1LZhmoBptbCW4LBcYEKwYv3EbuCGbKyiP?usp=sharing>

*# OR*

*# download Confluent Platform from the official website (requires email id)*

1. <https://www.confluent.io/download/>

# **IMPORTANT** *- Place confluent hub in the same folder as the project folder (usually home) and unzip it there using -*

1. tar -xvf confluent-5.2.1-2.12.tar.gz
2. git clone <https://github.com/jcustenborder/kafka-connect-twitter.git>

**C. Install maven and openjdk-8** *(this java version is required for mvn clean package)*

1. sudo apt install maven
2. sudo apt-get install openjdk-8-jdk

**D. Install jq for JSON parsing**

1. sudo apt-get install jq

**E. Required Python libraries**

1. sudo pip3 install pyspark
2. sudo pip3 install mmh3
3. sudo pip3 install bitarray
4. sudo pip3 install elasticsearch --upgrade

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**OPTIONAL** - *Install of kafka and spark is not required on vcl*

*(Since kafka and Spark are already installed on vcl and kafka comes with confluent as well)*

Optional 1. Spark Installation

*# if pip3 is not installed*

# sudo apt install python3-pip

*# downloading latest spark*

1. wget <https://www-us.apache.org/dist/spark/spark-2.4.2/spark-2.4.2-bin-hadoop2.7.tgz>

*# OR in case wget fails use our share drive link to download - spark-2.4.2-bin-hadoop2.7.tgz*

1. <https://drive.google.com/drive/folders/1LZhmoBptbCW4LBcYEKwYv3EbuCGbKyiP?usp=sharing>

# unpack spark

1. tar -zxvf spark-2.4.2-bin-hadoop2.7.tgz

*# edit environment variables to launch pyspark with python3*

1. echo "export SPARK\_HOME=~/spark-2.4.2-bin-hadoop2.7" >> ~/.bashrc
2. source ~/.bashrc
3. echo "export PATH=$SPARK\_HOME/bin:$PATH" >> ~/.bashrc
4. source ~/.bashrc
5. echo "export PYSPARK\_PYTHON=python3" >> ~/.bashrc
6. source ~/.bashrc

Optional 2. Kafka

*# download and install*

1. wget <https://www-us.apache.org/dist/kafka/2.2.0/kafka_2.12-2.2.0.tgz>

*# OR from the share drive link download - kafka\_2.12-2.2.0.tgz*

1. <https://drive.google.com/drive/folders/1LZhmoBptbCW4LBcYEKwYv3EbuCGbKyiP?usp=sharing>
2. sudo mkdir /opt/KAFKA
3. tar xzf kafka\_2.12-2.2.0.tgz
4. sudo mv kafka\_2.12-2.2.0 /opt/KAFKA

*# setup environment variables*

1. echo "export KAFKA\_HOME="/opt/KAFKA/kafka\_2.12-2.2.0"" >> ~/.bashrc
2. source ~/.bashrc

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**SETTING ENVIRONMENT VARIABLES:**

*# run these commands on terminal or set open .bashrc and add JAVA\_HOME and* *PYTHONPATH at the end of the file*

1. echo "export JAVA\_HOME="/usr/lib/jvm/java-1.8.0-openjdk-amd64/"" >> ~/.bashrc
2. echo "export PYTHONPATH=$SPARK\_HOME/python/:$PYTHONPATH" >> ~/.bashrc
3. source ~/.bashrc

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**SOME REQUIRED STEPS BEFORE ACTUAL RUN STARTS:**

1. cd kafka-connect-twitter
2. mvn clean package
3. cd target
4. tar -xvf kafka-connect-twitter-0.2-SNAPSHOT.tar.gz

*# Move back to your home folder location on the terminal or where you unzipped confluent hub*

1. cd confluent-5.2.1/etc/schema-registry

*# We need to edit - connect-avro-distributed.properties file*

*# Simply open connect-avro-distributed.properties file using a text editor and*

1. Find plugin.path value at the end of the file

*# edit its value to (replace <unityid> with your unityid or username)*

1. plugin.path=share/java,/home/<unityid>/kafka-connect-twitter/

*# save and close the file*

*# OR approach using vim*

1. vim connect-avro-distributed.properties

*# Add to it (edit plugin.path)*

1. plugin.path=share/java,/home/<unityid>/kafka-connect-twitter/

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**RUNNING INSTRUCTIONS:**

*#* **IMPORTANT** *- Make sure you are in the correct directory (/home/<user>) or where all the* *project files are placed*

*# start elasticsearch*

1. sudo systemctl start elasticsearch.service

*# Start all services using Confluent*

1. ./confluent-5.2.1/bin/confluent start

*# Load Sink*

*# (***IMPORTANT** *- Sometimes this will not work the first time, so wait for a minute and run the command again till you a prettified json format packet on terminal)*

1. ./confluent-5.2.1/bin/confluent load twitter-kafka-elastic-sink -d ./twitter-kafka-connect-elasticsearch-sink.json

*# Load Source*

1. ./confluent-5.2.1/bin/confluent load twitter\_source\_json -d ./twitter-source-json.json

*# Run the code using this instruction*

1. $SPARK\_HOME/bin/spark-submit --packages org.apache.spark:spark-streaming-kafka-0-8\_2.11:2.0.0 streamFromKafka.py

*# Run ElasticSearch.py*

1. python3 ElasticSearch.py

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**TO STOP RUNNING SERVICES:**

*# stop elasticsearch*

1. sudo systemctl stop elasticsearch.service

*# unload sink*

1. ./confluent-5.2.1/bin/confluent unload twitter-kafka-elastic-sink -d ./twitter-kafka-connect-elasticsearch-sink.json

*# unload source*

1. ./confluent-5.2.1/bin/confluent unload twitter\_source\_json -d ./twitter-source-json.json

*# stop confluent services*

1. ./confluent-5.2.1/bin/confluent stop

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**SOME TROUBLESHOOTING INSTRUCTIONS:**

A. If the first time run of the streamFromKafka.py file fails, try running the command again.

B. cURL check using (*check if it's working properly*)

1. curl localhost:9200

C. Restart confluent services:

1. ./confluent-5.2.1/bin/confluent stop connect
2. ./confluent-5.2.1/bin/confluent start connect

D. Unload and reload twitter-source-json.json and twitter-kafka-connect-elasticsearch-sink.json if facing any problem using steps mentioned in running and stop instructions

E. Consumer (If you want to check data is coming in Kafka through twitter)

1. ./confluent-5.2.1/bin/kafka-console-consumer --bootstrap-server localhost:9092 --topic twitterDataJson --from-beginning

F. Check Logs

1. ./confluent-5.2.1/bin/confluent log connect -f

*# It adds data to an index named twitterdatajson in elasticsearch*

***#*** *Check it with this command*

1. curl -XGET 'http://localhost:9200/twitterdatajson/\_search?pretty'

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**REFERENCES:**

[1] <https://www.confluent.io/blog/using-ksql-to-analyse-query-and-transform-data-in-kafka>

[2] <https://docs.confluent.io/current/connect/kafka-connect-elasticsearch/index.html>

[3]<https://www.confluent.io/blog/the-simplest-useful-kafka-connect-data-pipeline-in-the-world-or-thereabouts-part-2/>

[4] <https://www.youtube.com/watch?v=UPkqFvjN-yI>

[5] <https://www.youtube.com/watch?v=1EnvkPf7t6Y>

[6] <https://www.youtube.com/watch?v=ibxXO-b14j4>

[7] <https://www.youtube.com/watch?v=Bay3X9PAX5k>

[8]<https://www.rittmanmead.com/blog/2015/08/three-easy-ways-to-stream-twitter-data-into-elasticsearch/>

[9] <https://qbox.io/blog/building-an-elasticsearch-index-with-python>

[10] Project reference material provided by the professor.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_