**Data streaming using Apache Spark**

Overview of Docs:

The overall of this document deals about how the log file is transferred from Flume to Kafka and sending the real time data log file in apache spark streaming. Once the data gets streamed it makes the data in a structured format and insert the data in Cassandra DB.

**Flume**:

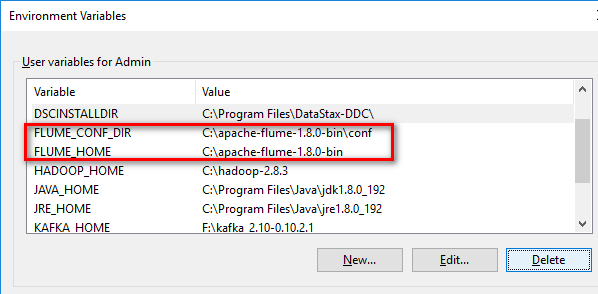
Flume is mainly used for getting the log file in a defined in local file path location. The configuration done in the properties file.

Location to download flume:

<http://www.apache.org/dyn/closer.lua/flume/1.8.0/apache-flume-1.8.0-src.tar.gz>

Steps:

1. Extract the tarball and place in separate directory.
2. Edit the environment variables in windows



1. Place the properties file under \apache-flume-1.8.0-bin\conf



1. Run the flume command

flume-ng agent -conf conf -conf-file conf\flume-conf.properties --name a1

**Kafka:**

In this application kafka mainly used for the purpose of consuming the messages from the flume. Once the flume reads the data immediately transfers to the kafka topic and sends to the consumer.

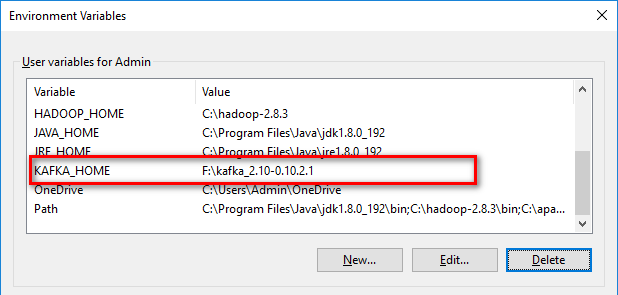
Zookeeper acts as a mediator to transfer the information from producer to consumer

Location to download kafka:

<https://www.apache.org/dyn/closer.cgi?path=/kafka/0.10.2.2/kafka_2.10-0.10.2.2.tgz>

Steps to install Kafka:

1. Extract the tarball in separate location.
2. Set environment variable



1. Update the config file under (kafka\_2.10-0.10.2.2\config) (server.properties and zookeeper.properties)
2. Update the log directory in both the file



1. Start the Zookeeper go to the folder location (kafka\_2.10-0.10.2.2\bin\windows)
2. From file location go to the windows folder run command(start zookeeper)

zookeeper-server-start.bat %KAFKA\_HOME%\config\zookeeper.properties

1. From file location go to the windows folder run command (start zookeeper)

kafka-server-start.bat %KAFKA\_HOME%\config\server.properties

1. From the kafka folder go to windows folder under (kafka\_2.10-0.10.2.2\bin\windows)
2. Run to create kafka topic

kafka-topics.bat --create --zookeeper localhost:2181 --replication-factor 1 --partitions 1 --topic sample\_topic

To check kafka working run the producer and consumer go to kafka\_2.10-0.10.2.2\bin\windows)

1. kafka-console-producer.bat --broker-list localhost:9092 --topic sample\_topic
2. kafka-console-consumer.bat --zookeeper localhost:2181 --topic sample\_topic -- from-beginning

Send some messages from producer to consumer

To run Apache spark streaming:

**Run the Cassandra DB:**

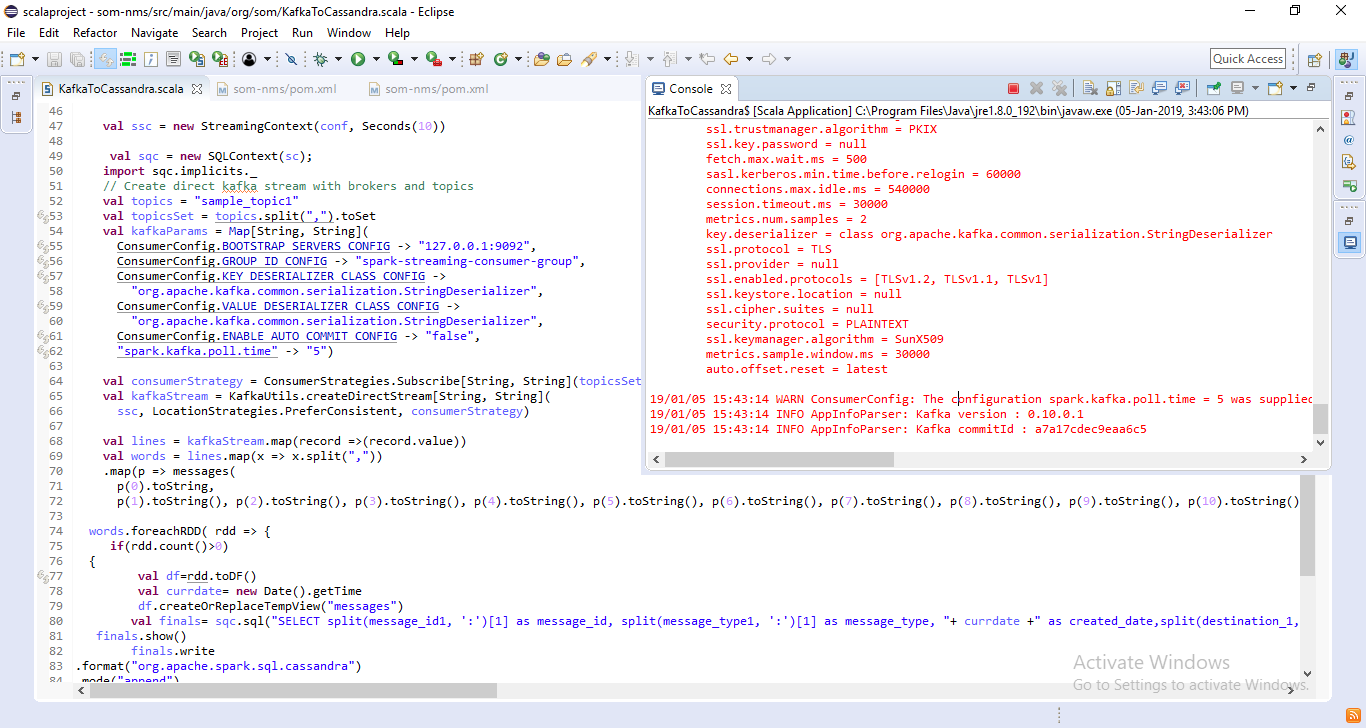
Keyspace:

CREATE KEYSPACE som WITH REPLICATION = { 'class' : 'org.apache.cassandra.locator.SimpleStrategy', 'replication\_factor': '1' } AND DURABLE\_WRITES = true;

**Create the table in db:**

CREATE TABLE som.messages (message\_id text,message\_type text,created\_date timestamp,destination text,message text,message\_status text,pin text,producer\_id text,send\_id text,template text,ticket\_id text,users text,PRIMARY KEY (message\_type,created\_date)) WITH CLUSTERING ORDER BY (created\_date ASC);

Run the Kafkatocassandra.scala program



The above program will receive the kafka consumer record and write in to the Cassandra database.

**To run with data add the text in the log file**

message\_id : ddddd , message\_type : EmailConfig, created\_date : 2018-01-05, destination : TamilDest, message : This is a test message from SOM Micro Services, message\_status : New , pin : tag.tamilselvan@gmail.com , producer\_id : e2dcdb8acf13c9f0b473ae70240ff216, send\_id :12345, template : Default, ticket\_id : 1028 , users :Tamil