Name: Khanh Nguyen

Date: 7/16/2022

Project 4.30: Data Ingestion into Big Data Systems and ETL

DESCRIPTION

Target is one of the biggest e-commerce giants in the US.  
They have a web portal where customers can purchase items, and the sellers can add products, remove the products, and do inventory-related operations.  
Recently, they started getting a lot of errors on the portal. They have collected these errors from all the applications and compiled them in a text file. Processing logs is a big task as an application can generate a lot of logs in a single day. They want to send all logs to HDFS so they can check which are the most frequent errors they are getting.

You have given an error log file containing the below details.

1. Dates
2. Server
3. Error message

You must read data from the text file and send it to Kafka and flume script.  
Also, you should be able to read data from Kafka and push it into HDFS.

Your task is to create:

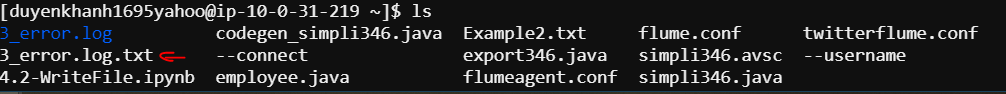
1. A Kafka producer which reads the CSV file data one by one and pushes it to Kafka
2. Write flume configuration where the source is Kafka and the sink is HDFS

**Objective**: To understand how to create a real-time pipeline using Kafka and Apache flume to send data to HDFS.

You can download the datasets from [here](https://github.com/Simplilearn-Edu/BDHS/tree/master/Lesson%20end%20project%20Datasets/Lesson-3)   
This is a non-gradable project

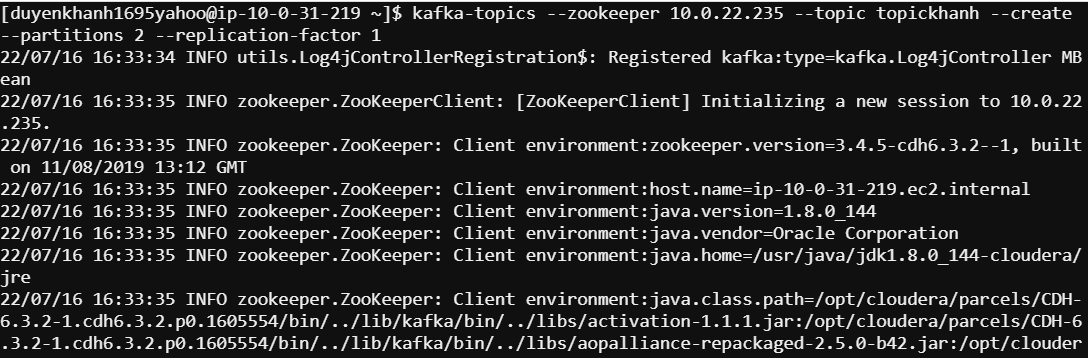
Work:

1. Upload “3\_error.log.txt” to FTP by drag and dropping from local machine.



**2. Create Kafka Topic:**

kafka-topics --zookeeper 10.0.22.235 --topic topickhanh –create --partitions 2 –replication-factor 1



**3. Create Flume Agent**

Write a flumeagent.conf file like this:

flumeagent.sources = kafka-source-1

flumeagent.channels = hdfs-channel-1

flumeagent.sinks = hdfs-sink-1

flumeagent.sources.kafka-source-1.type = org.apache.flume.source.kafka.KafkaSource

flumeagent.sources.kafka-source-1.zookeeperConnect = 10.0.22.235

flumeagent.sources.kafka-source-1.topic = topickhanh

flumeagent.sources.kafka-source-1.batchSize = 100

flumeagent.sinks.hdfs-sink-1.type = hdfs

flumeagent.sinks.hdfs-sink-1.hdfs.writeFormat = Text

flumeagent.sinks.hdfs-sink-1.hdfs.fileType = DataStream

flumeagent.sinks.hdfs-sink-1.hdfs.filePrefix = test-events

flumeagent.sinks.hdfs-sink-1.hdfs.useLocalTimeStamp = true

flumeagent.sinks.hdfs-sink-1.hdfs.path = /user/duyenkhanh1695yahoo/%{topic}/%y-%m-%d

#flumeagent.sinks.hdfs-sink-1.hdfs.path = /user/duyenkhanh1695yahoo/flumejob

flumeagent.sinks.hdfs-sink-1.hdfs.rollCount=100

flumeagent.sinks.hdfs-sink-1.hdfs.rollSize=0

flumeagent.channels.hdfs-channel-1.type = memory

flumeagent.channels.hdfs-channel-1.capacity = 10000

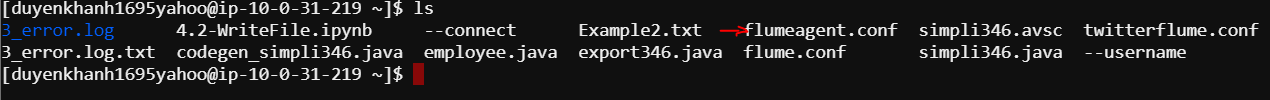
flumeagent.channels.hdfs-channel-1.transactionCapacity = 1000

#Integrations

flumeagent.sources.kafka-source-1.channels = hdfs-channel-1

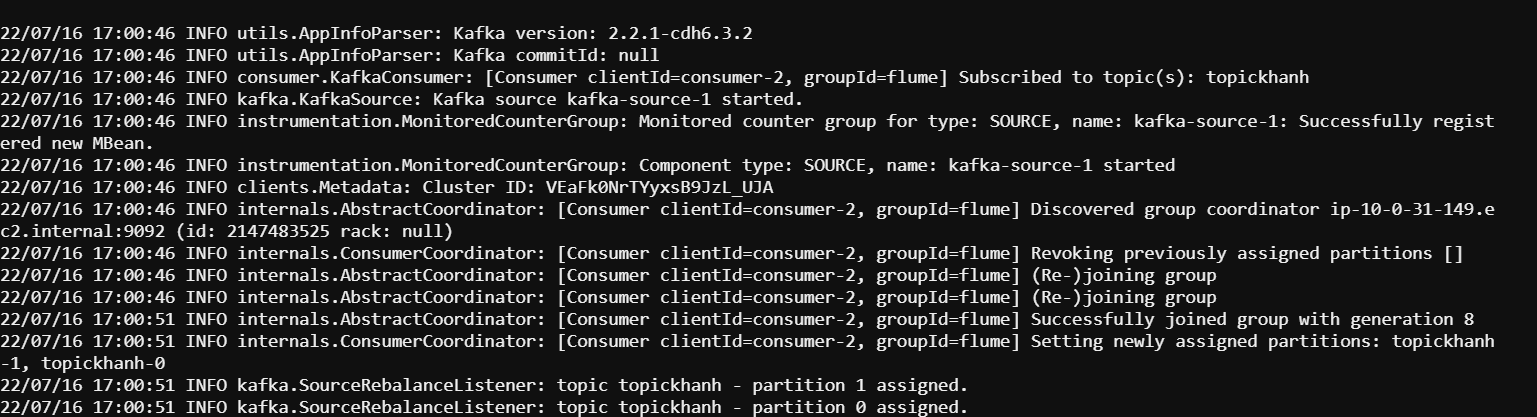
flumeagent.sinks.hdfs-sink-1.channel = hdfs-channel-1

4. Upload flumeagent.conf to FTP by drag and dropping the file from local machine.



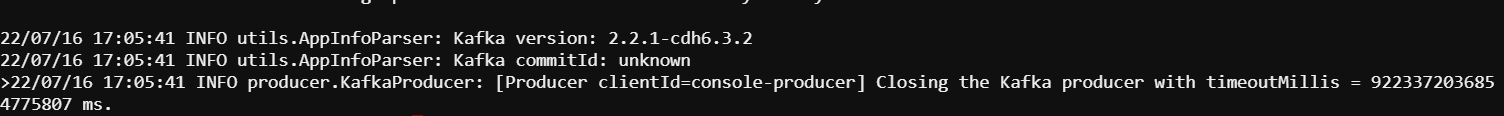
**5. Start Flume Agent**

Flume-ng agent -name flumeagent --conf-file flumeagent.conf



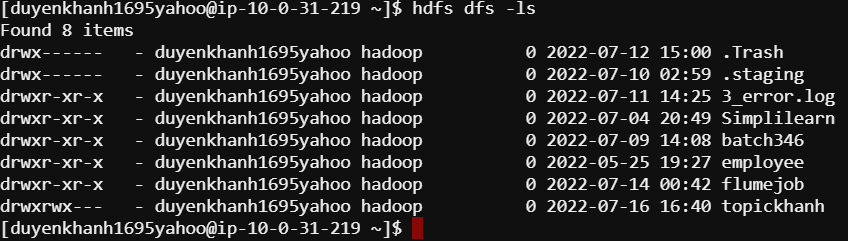
**6. Passing message from Kafka producer**

cat 3\_error.log.txt | kafka-console-producer –-broker-list 10.0.32.71:9092 –-topic topickhanh



7. Checking output in HDFS:

hdfs dfs -ls



hdfs dfs -ls topickhanh/22-07-16