Kalyan Big Data Projects – Project 5
How To Stream CSV Data Into Hadoop
Using Apache Flume – Kafka Sink

Pre-Requisites of Flume Project:

hadoop-2.6.0
flume-1.6.0
kafka-0.9.0
java-1.7

NOTE: Make sure that install all the above components

Flume Project Download Links:

`hadoop-2.6.0.tar.gz` ==> link
(https://archive.apache.org/dist/hadoop/core/hadoop-2.6.0/hadoop-2.6.0.tar.gz)

`apache-flume-1.6.0-bin.tar.gz` ==> link
(https://archive.apache.org/dist/flume/1.6.0/apache-flume-1.6.0-bin.tar.gz)

`kafka_2.11-0.9.0.0.tgz` ==> link
(https://archive.apache.org/dist/kafka/0.9.0.0/kafka_2.11-0.9.0.0.tgz)

`kalyan-kafka-sink-agent.conf` ==> link
(https://github.com/kalyanhadooptraining/kalyan-bigdata-realtime-
projects/blob/master/kafka/project5-flume-kafka-sink/kalyan-kafka-sink-agent.conf)

`kalyan-bigdata-examples.jar` ==> link
(https://github.com/kalyanhadooptraining/kalyan-bigdata-realtime-projects/blob/master/kalyan/kalyan-
bigdata-examples.jar)

ORIEN IT

Mr.Kalyan, Apache Contributor, Cloudera CCA175 Certified Consultant,
6+ years of Big Data exp, IIT Kharagpur, Gold Medalist

Learnings of this Project:

➢ We will learn Flume Configurations and Commands
➢ Flume Agent
   1. Source (Exec Source)
   2. Channel (Memory Channel)
   3. Sink (Kafka Sink)
➢ We will learn Kafka Configurations and Commands
➢ Kafka Information
   1. Kalyan Util (CSV data generator)
   2. Kafka Producer (Listen on CSV data)
   3. Kafka Consumer (Recieves the data from Kafka Producer)
   4. Flume Kafka Sink (Will Recieves the Kafka Sink data from Flume Channel)
➢ Major project in Real Time `Product Log Analysis`
   1. We are extracting the data from server logs
   2. This data will be useful to do analysis on product views
   3. CSV is the output format
➢ We can use hive / pig / mapreduce to analyze this data
   1. explore hive query to analysis
   2. explore pig scripts to analysis
   3. explore mapreduce to analysis

----

1. create "**kalyan-kafka-sink-agent.conf**" file with below content

```
agent.sources = EXEC
agent.channels = MemChannel
agent.sinks = KAFKA

agent.sources.EXEC.type = exec
agent.sources.EXEC.command = tail -F /tmp/users.csv
agent.sources.EXEC.channels = MemChannel

agent.sinks.KAFKA.type = org.apache.flume.sink.kafka.KafkaSink
agent.sinks.KAFKA.channel = MemChannel
agent.sinks.KAFKA.kafka.topic = flume-csv-topic
agent.sinks.KAFKA.kafka.bootstrap.servers = localhost:9092
agent.sinks.KAFKA.kafka.flumeBatchSize = 10

agent.channels.MemChannel.type = memory
agent.channels.MemChannel.capacity = 1000
agent.channels.MemChannel.transactionCapacity = 100
```

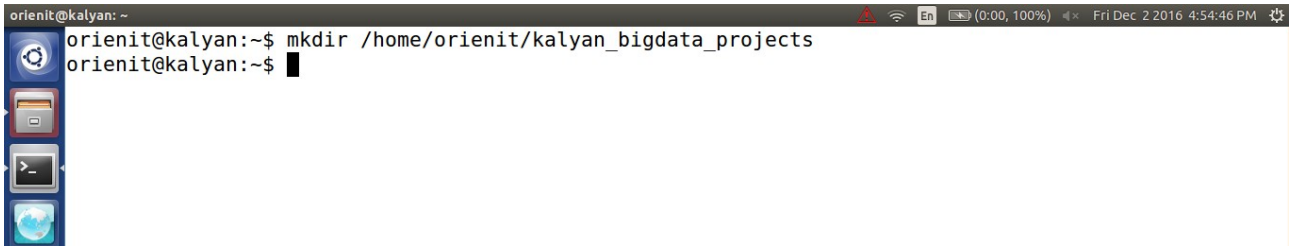2. Copy "**kalyan-kafka-sink-agent.conf**" file into "**$FUME_HOME/conf**" folder

3. Generate Large Amount of Sample CSV data follow this [article](http://kalyanbigdatatraining.blogspot.com/2016/12/how-to-generate-large-amount-of-sample.html).
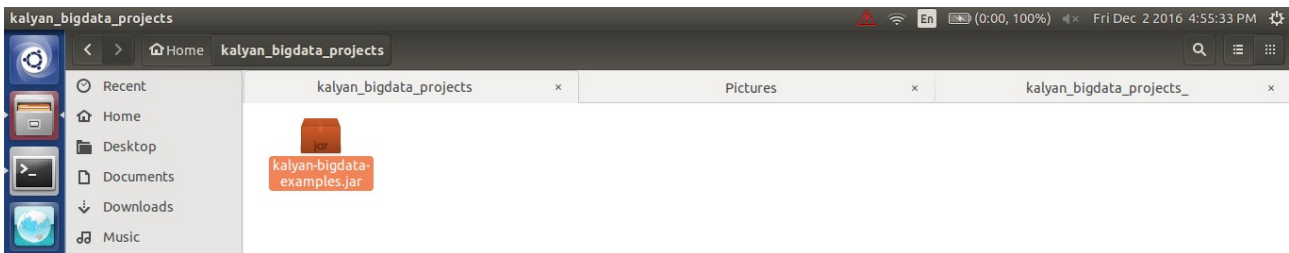
(http://kalyanbigdatatraining.blogspot.com/2016/12/how-to-generate-large-amount-of-sample.html)

4. Follow below steps...

i) Create '**kalyan_bigdata_projects**' folder in **user home** (i.e **/home/orienit**)

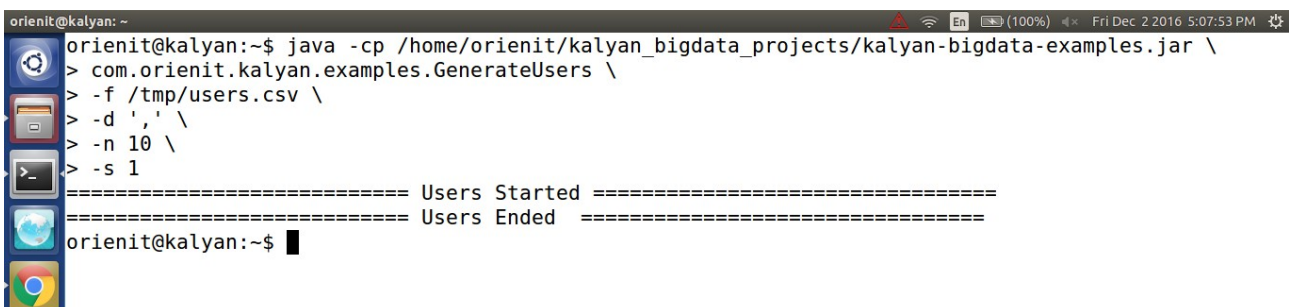**Command:** *mkdir /home/orienit/kalyan_bigdata_projects*

ii) Copy '**kalyan-bigdata-examples.jar**' jar file into '**/home/orienit/kalyan_bigdata_projects**' folder

iii) Execute Below Command to Generate Sample CSV data with 100 lines. Increase this number to get more data ...

java -cp /home/orienit/kalyan_bigdata_projects/kalyan-bigdata-examples.jar \
com.orienit.kalyan.examples.GenerateUsers \
-f /tmp/users.csv \
-d ',' \
-n 10 \
-s 1

ORIEN IT

Mr.Kalyan, Apache Contributor, Cloudera CCA175 Certified Consultant,
6+ years of Big Data exp, IIT Kharagpur, Gold Medalist

5. Verify the Sample CSV data in Console, using below command

cat /tmp/users.csv

```
orienit@kalyan:~$ cat /tmp/users.csv
1,user1,user1,user1@gmail.com,US,Washington,Seattle,2016-07-02 05:07:48
2,user2,user2,user2@gmail.com,US,Florida,Orlando,2016-07-02 05:07:48
3,user3,user3,user3@gmail.com,US,New York,Little Falls,2016-07-02 05:07:49
4,user4,user4,user4@gmail.com,India,Karnataka,Mangaluru,2016-07-02 05:07:49
5,user5,user5,user5@gmail.com,US,Hawaii,Hanapepe,2016-07-02 05:07:49
6,user6,user6,user6@gmail.com,India,Chennai,Kottur,2016-07-02 05:07:49
7,user7,user7,user7@gmail.com,India,Andhra Pradesh,Kakinada,2016-07-02 05:07:49
8,user8,user8,user8@gmail.com,US,Hawaii,East Honolulu,2016-07-02 05:07:49
9,user9,user9,user9@gmail.com,US,Florida,Hollywood,2016-07-02 05:07:49
10,user10,user10,user10@gmail.com,US,Washington,Bellevue,2016-07-02 05:07:49
orienit@kalyan:~$
```
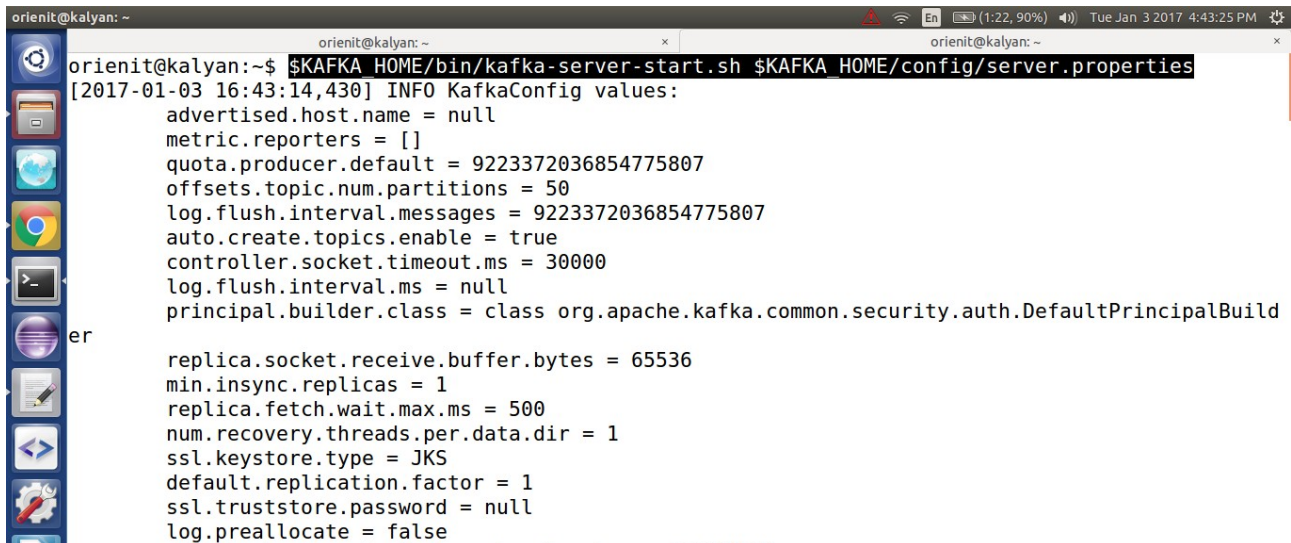
6. Start the `zookeeper` using below command (New Terminal)

$KAFKA_HOME/bin/zookeeper-server-start.sh $KAFKA_HOME/config/zookeeper.properties

```
orienit@kalyan:~$ $KAFKA_HOME/bin/zookeeper-server-start.sh $KAFKA_HOME/config/zookeeper.properties

[2017-01-03 16:41:44,986] INFO Reading configuration from: /home/orienit/work/kafka_2.11-0.9.0.0/config/zookeeper.properties (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2017-01-03 16:41:45,024] INFO autopurge.snapRetainCount set to 3 (org.apache.zookeeper.server.DatadirCleanupManager)
[2017-01-03 16:41:45,024] INFO autopurge.purgeInterval set to 0 (org.apache.zookeeper.server.DatadirCleanupManager)
[2017-01-03 16:41:45,024] INFO Purge task is not scheduled. (org.apache.zookeeper.server.DatadirCleanupManager)
[2017-01-03 16:41:45,025] WARN Either no config or no quorum defined in config, running  in standalone mode (org.apache.zookeeper.server.quorum.QuorumPeerMain)
[2017-01-03 16:41:45,086] INFO Reading configuration from: /home/orienit/work/kafka_2.11-0.9.0.0/config/zookeeper.properties (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2017-01-03 16:41:45,086] INFO Starting server (org.apache.zookeeper.server.ZooKeeperServerMain)
[2017-01-03 16:41:45,117] INFO Server environment:zookeeper.version=3.4.6-1569965, built on 02/20/2014 09:09 GMT (org.apache.zookeeper.server.ZooKeeperServer)
[2017-01-03 16:41:45,117] INFO Server environment:host.name=localhost (org.apache.zookeeper.server.ZooKeeperServer)
[2017-01-03 16:41:45,117] INFO Server environment:java.version=1.8.0_66-internal (org.apache.zookeeper.server.ZooKeeperServer)
[2017-01-03 16:41:45,117] INFO Server environment:java.vendor=Oracle Corporation (org.apache.zookee
```

ORIEN IT

Mr.Kalyan, Apache Contributor, Cloudera CCA175 Certified Consultant,
6+ years of Big Data exp, IIT Kharagpur, Gold Medalist

7. Start the `**kafka server**` using below command (New Terminal)

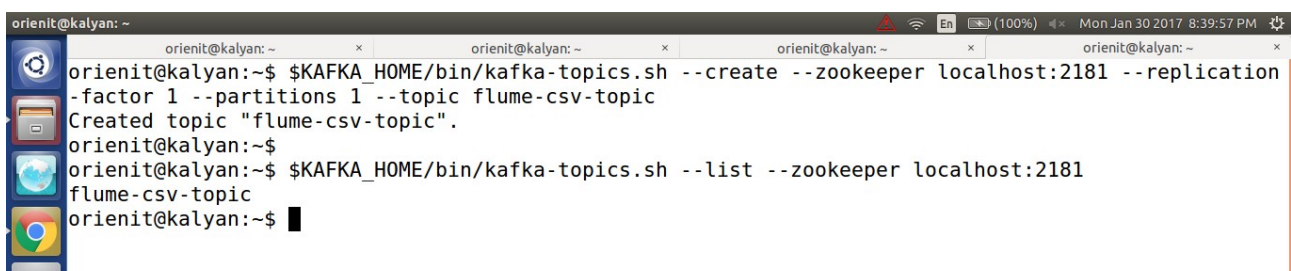$KAFKA_HOME/bin/kafka-server-start.sh $KAFKA_HOME/config/server.properties

8. Create a `**flume-csv-topic**` topic using below command (New Terminal)

$KAFKA_HOME/bin/kafka-topics.sh --create --zookeeper localhost:2181 --replication-factor 1
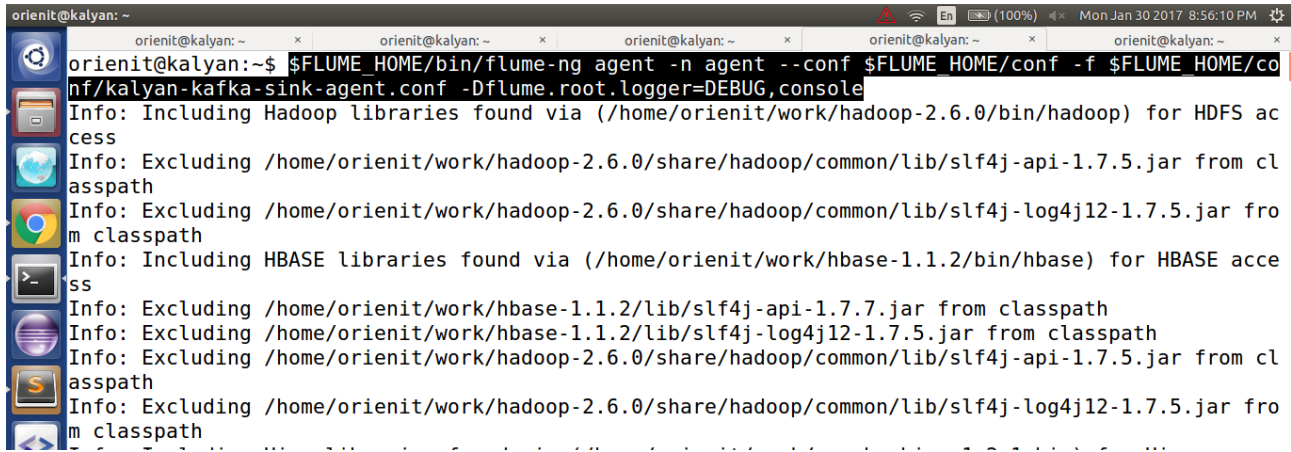--partitions 1 --topic flume-csv-topic

9. List out all the topics

$KAFKA_HOME/bin/kafka-topics.sh --list --zookeeper localhost:2181

10. Execute the below command to `**Extract data from CSV into KAFKA using Flume**`

$FLUME_HOME/bin/flume-ng agent -n agent --conf $FLUME_HOME/conf -f
$FLUME_HOME/conf/kalyan-kafka-sink-agent.conf -Dflume.root.logger=DEBUG,console

```
orienit@kalyan:~$ $FLUME_HOME/bin/flume-ng agent -n agent --conf $FLUME_HOME/conf -f $FLUME_HOME/co
nf/kalyan-kafka-sink-agent.conf -Dflume.root.logger=DEBUG,console
Info: Including Hadoop libraries found via (/home/orienit/work/hadoop-2.6.0/bin/hadoop) for HDFS ac
cess
Info: Excluding /home/orienit/work/hadoop-2.6.0/share/hadoop/common/lib/slf4j-api-1.7.5.jar from cl
asspath
Info: Excluding /home/orienit/work/hadoop-2.6.0/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar fro
m classpath
Info: Including HBASE libraries found via (/home/orienit/work/hbase-1.1.2/bin/hbase) for HBASE acce
ss
Info: Excluding /home/orienit/work/hbase-1.1.2/lib/slf4j-api-1.7.7.jar from classpath
Info: Excluding /home/orienit/work/hbase-1.1.2/lib/slf4j-log4j12-1.7.5.jar from classpath
Info: Excluding /home/orienit/work/hadoop-2.6.0/share/hadoop/common/lib/slf4j-api-1.7.5.jar from cl
asspath
Info: Excluding /home/orienit/work/hadoop-2.6.0/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar fro
m classpath
```
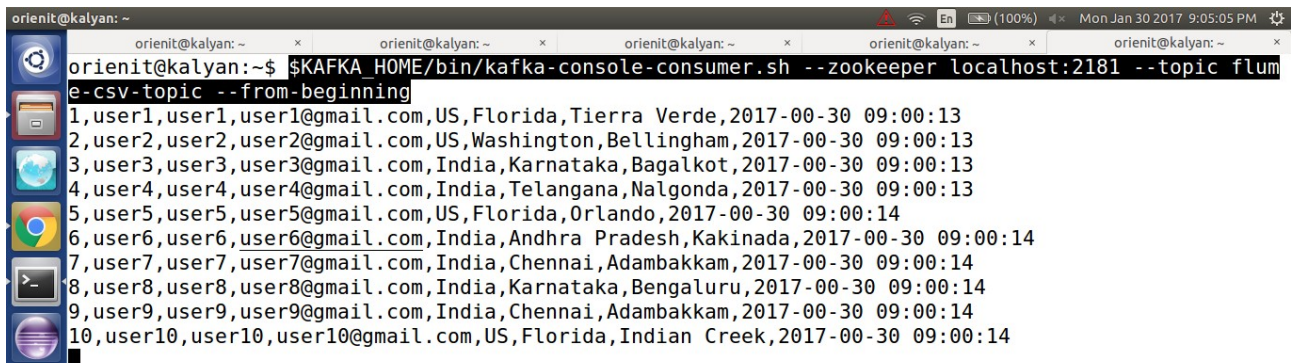
11. Start the `**kafka consumer**` using below command (New Terminal)

$KAFKA_HOME/bin/kafka-console-consumer.sh --zookeeper localhost:2181 --topic flume-csv-
topic --from-beginning

```
orienit@kalyan:~$ $KAFKA_HOME/bin/kafka-console-consumer.sh --zookeeper localhost:2181 --topic flum
e-csv-topic --from-beginning
1,user1,user1,user1@gmail.com,US,Florida,Tierra Verde,2017-00-30 09:00:13
2,user2,user2,user2@gmail.com,US,Washington,Bellingham,2017-00-30 09:00:13
3,user3,user3,user3@gmail.com,India,Karnataka,Bagalkot,2017-00-30 09:00:13
4,user4,user4,user4@gmail.com,India,Telangana,Nalgonda,2017-00-30 09:00:13
5,user5,user5,user5@gmail.com,US,Florida,Orlando,2017-00-30 09:00:14
6,user6,user6,user6@gmail.com,India,Andhra Pradesh,Kakinada,2017-00-30 09:00:14
7,user7,user7,user7@gmail.com,India,Chennai,Adambakkam,2017-00-30 09:00:14
8,user8,user8,user8@gmail.com,India,Karnataka,Bengaluru,2017-00-30 09:00:14
9,user9,user9,user9@gmail.com,India,Chennai,Adambakkam,2017-00-30 09:00:14
10,user10,user10,user10@gmail.com,US,Florida,Indian Creek,2017-00-30 09:00:14
```