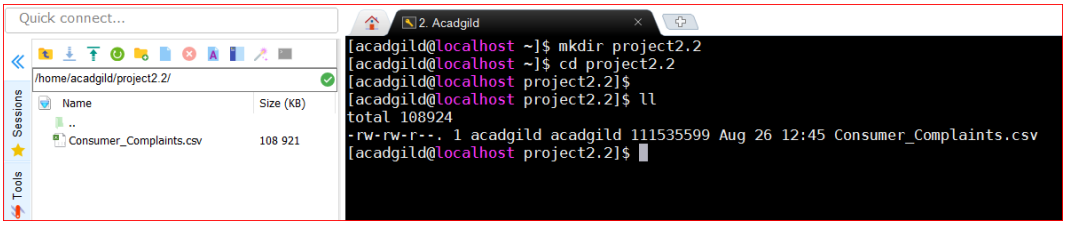
# ***Copying Dataset using Flume***

## Download the dataset using the below link:

https://drive.google.com/file/d/0Bxr27gVaXO5sUjd2RWFQS3hQQUE/view?usp=sharing




## Start Hadoop daemons:

*$ start-all.sh*

*$ hadoop fs -ls /*

```
[acadgild@localhost project2.2]$ start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
17/08/26 12:49:16 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Starting namenodes on [localhost]
localhost: starting namenode, logging to /usr/local/hadoop-2.6.0/logs/hadoop-acadgild-namenode-localhost.localdomain.out
localhost: starting datanode, logging to /usr/local/hadoop-2.6.0/logs/hadoop-acadgild-datanode-localhost.localdomain.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop-2.6.0/logs/hadoop-acadgild-secondarynamenode-localhost.localdomain.out
17/08/26 12:49:35 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop-2.6.0/logs/yarn-acadgild-resourcemanager-localhost.localdomain.out
localhost: starting nodemanager, logging to /usr/local/hadoop-2.6.0/logs/yarn-acadgild-nodemanager-localhost.localdomain.out
```

```
[acadgild@localhost project2.2]$ hadoop fs -ls /
17/08/26 12:59:06 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 8 items
drwxr-xr-x   - acadgild supergroup          0 2017-08-25 16:54 /flume_sink
drwxr-xr-x   - acadgild supergroup          0 2017-06-03 16:59 /hadoop
drwxr-xr-x   - acadgild supergroup          0 2017-08-19 13:53 /hbase_output_dir
drwxr-xr-x   - acadgild supergroup          0 2017-08-25 20:42 /hbasestorage
drwxr-xr-x   - acadgild supergroup          0 2017-08-24 01:16 /sqoop
drwxrwxr-x   - acadgild supergroup          0 2017-08-25 22:01 /tmp
drwxr-xr-x   - acadgild supergroup          0 2017-07-11 15:57 /user
drwxr-xr-x   - acadgild supergroup          0 2015-11-05 12:56 /zookeeper
[acadgild@localhost project2.2]$
```

## Copy dataset from local file system to HDFS using flume:

**Step 1: create source directory and move dataset file into it:**

*$ mkdir flume_sink*

*$ mv Consumer_Complaints.csv flume_sink/*

*$ ls flume_sink/*

```
[acadgild@localhost project2.2]$ mkdir flume_sink
[acadgild@localhost project2.2]$ mv Consumer_Complaints.csv flume_sink/
[acadgild@localhost project2.2]$ ls flume_sink/
Consumer_Complaints.csv
[acadgild@localhost project2.2]$
```

**Step 2: create flume config file which uses below source and target directory:**

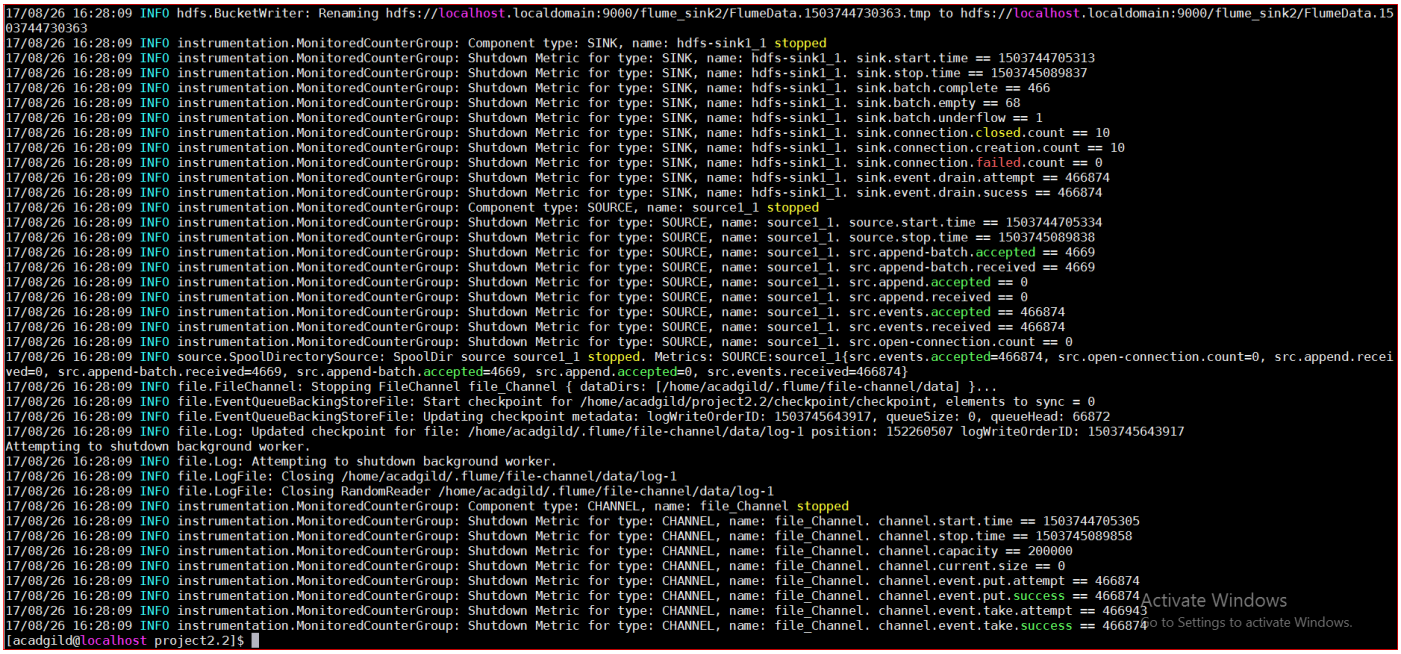
*Source Directory: /home/acadgild/project2.2/flume_sink*

*Target Directory: hdfs://localhost.localdomain:9000/flume_sink2*

```
#specify source, channel and sink
agent1.sinks =  hdfs-sink1_1
agent1.sources = source1_1
agent1.channels = file_Channel

#Flume Configuration Starts
# Define a file channel called fileChannel on agent1

agent1.channels.file_Channel.type = file
agent1.channels.file_Channel.checkpointDir = /home/acadgild/project2.2/checkpoint
agent1.channels.file_Channel.checkpointInterval = 30000

# on linux FS
agent1.channels.file_Channel.capacity = 200000
agent1.channels.file_Channel.transactionCapacity = 100000

# Define a source for agent1
agent1.sources.source1_1.type = spooldir

# on linux FS
#Spooldir in my case is /home/acadgild/project2.2/flume_sink

agent1.sources.source1_1.spoolDir = /home/acadgild/project2.2/flume_sink
agent1.sources.source1_1.fileHeader = false
agent1.sources.source1_1.fileSuffix = .COMPLETED
agent1.sources.source1_1.deserializer.maxLineLength=10000
agent1.sinks.hdfs-sink1_1.type = hdfs

#Sink is /flume_import under hdfs

agent1.sinks.hdfs-sink1_1.hdfs.path = hdfs://localhost.localdomain:9000/flume_sink2
agent1.sinks.hdfs-sink1_1.hdfs.batchSize = 1000
agent1.sinks.hdfs-sink1_1.hdfs.rollSize = 268435456
agent1.sinks.hdfs-sink1_1.hdfs.rollInterval = 0
agent1.sinks.hdfs-sink1_1.hdfs.rollCount = 50000
agent1.sinks.hdfs-sink1_1.hdfs.writeFormat=Text

agent1.sinks.hdfs-sink1_1.hdfs.fileType = DataStream
agent1.sources.source1_1.channels = file_Channel
agent1.sinks.hdfs-sink1_1.channel = file_Channel

~
```

**Step 3: run flume agent agent1 from spool_dir.config file:**

*$ flume-ng agent -n agent1 -f spool_dir.conf*




We have changed below parameter since event size was causing the issue.

**agent1.channels.file_Channel.checkpointDir = /home/acadgild/project2.2/checkpoint**

**agent1.channels.file_Channel.checkpointInterval = 30000**

Above statements are intended to create a checkpoint directory where file which is being moved is check-pointed in order to drive smooth transfer. Second statement is intended for interval of check pointing which is 30 sec in our case.

**agent1.sources.source1_1.deserializer.maxLineLength=10000**

The default maxLineLength for the LINE deserializer is 2048, which is very small than the each line length in our dataset. We can set it as per our needs to accommodate our large events.

**Step 4: verify if file has been transferred completely:**

Check if .COMPLETED have been added to the filename in the local file system under directory flume_sink. Also check the number of lines in the dataset file:

*$ ll flume_sink/**

*$ wc -l flume_sink/**

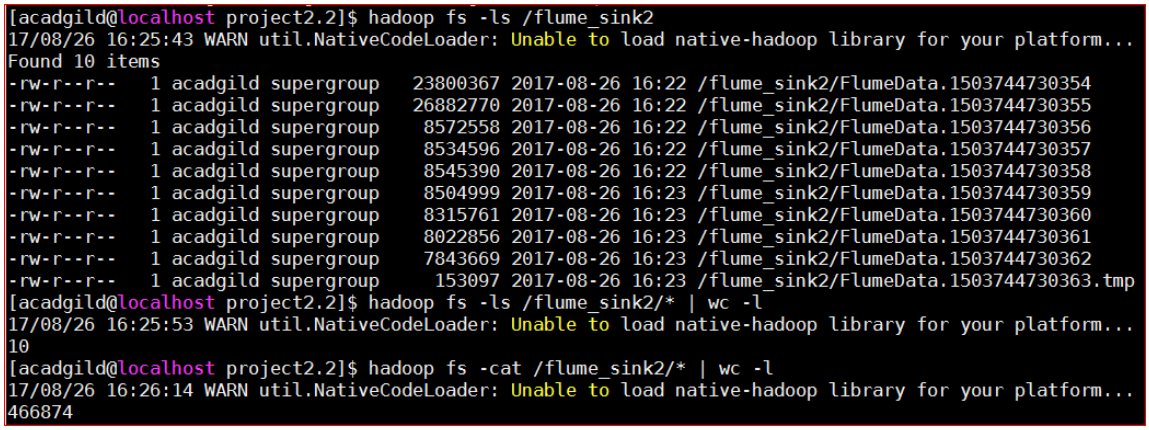
```
[acadgild@localhost project2.2]$ ll flume_sink/*
-rw-rw-r--. 1 acadgild acadgild 111535599 Aug 26 12:45 flume_sink/ConsumerComplaints.csv.COMPLETED
[acadgild@localhost project2.2]$ wc -l flume_sink/*
466874 flume_sink/ConsumerComplaints.csv.COMPLETED
[acadgild@localhost project2.2]$
```

Check if /flume_sink2 has same number of lines in the FlumeData files:

*$ hadoop fs -ls /flume_sink2*

*$ hadoop fs -ls /flume_sink2/* | wc -l*

*$ hadoop fs -cat /flume_sink2/* | wc -l*




Press cntrl+c for killing the flume agent since file have been copied as per the logs. As we can see that files have same number of lines concluding that whole file has been transferred without any loss of data.