**HADOOP: FLUME CASE STUDY**

**CASE STUDY-1: SOCIAL MEDIA**

1. Download Streaming data from Twitter (or any other social media website) into HDFS
2. Create Twitter Application and Get access Keys
3. Create configuration File (Agent)
4. Download the data for tweets related to specific key words

[Using latest twitter4j 4.0.6 jars]

TwitterAgent.sources = Twitter

TwitterAgent.channels = MemChannel

TwitterAgent.sinks = HDFS

# Describe & configure the source

# TwitterAgent.sources.Twitter.type = org.apache.flume.source.twitter.TwitterSource

TwitterAgent.sources.Twitter.type = com.cloudera.flume.source.TwitterSource

TwitterAgent.sources.Twitter.channels = MemChannel

TwitterAgent.sources.Twitter.consumerKey = XXXXXXXXXXXXXXXXXX

TwitterAgent.sources.Twitter.consumerSecret = XXXXXXXXXXXXXXXXX

TwitterAgent.sources.Twitter.accessToken = XXXXXXXXXXXXXXXXX

TwitterAgent.sources.Twitter.accessTokenSecret = XXXXXXXXXXXXXX

TwitterAgent.sources.Twitter.maxBatchSize =1000

TwitterAgent.sources.Twitter.keywords = big data,hadoop,data analytics,data science,machine learning,AI

# Describe & configure the sink

TwitterAgent.sinks.HDFS.channel = MemChannel

TwitterAgent.sinks.HDFS.type = hdfs

TwitterAgent.sinks.HDFS.hdfs.useLocalTimeStamp = true

TwitterAgent.sinks.HDFS.hdfs.path = /user/cloudera/tweets

TwitterAgent.sinks.HDFS.hdfs.fileType = DataStream

TwitterAgent.sinks.HDFS.hdfs.writeFormat = Text

TwitterAgent.sinks.HDFS.hdfs.batchSize = 10

TwitterAgent.sinks.HDFS.hdfs.rollSize = 0

TwitterAgent.sinks.HDFS.hdfs.rollCount = 10000

TwitterAgent.sinks.HDFS.hdfs.rollInterval = 600

# Use a channel which buffers events in memory

TwitterAgent.channels.MemChannel.type = memory

TwitterAgent.channels.MemChannel.capacity = 10000

TwitterAgent.channels.MemChannel.transactionCapacity = 1000

[cloudera@quickstart flume-demo]$ flume-ng agent --conf-file twitter.conf --name TwitterAgent --conf $FLUME\_HOME/conf -Dflume.root.logger=INFO,console -Dtwitter4j.streamBaseURL=https://stream.twitter.com/1.1/





**CASE STUDY-2: Transfer Hadoop logs to HDFS**

1. In a Hadoop node logs are created continuously
2. Design a solution to download these logs into HDFS or Hive
3. Create configuration File(Agent)
4. Download the data of all the logs to HDFS.

**Solution 1: Using spooldir type**

logs\_to\_hdfs.conf

agent1.sources = source1

agent1.sinks = sink1

agent1.channels = channel1

# Describe/configure the source

agent1.sources.source1.type = spooldir

agent1.sources.source1.spoolDir = /var/log/hadoop-hdfs

# Describe the sink

agent1.sinks.sink1.type = hdfs

agent1.sinks.sink1.fileType = DataStream

agent1.sinks.sink1.hdfs.path = /tmp/flume

# Use a channel which buffers events in memory

agent1.channels.channel1.type = memory

agent1.channels.channel1.capacity = 1000

agent1.channels.channel1.transactionCapacity = 100

# Bind the source and sink to the channel

agent1.sources.source1.channels = channel1

agent1.sinks.sink1.channel = channel1

[cloudera@quickstart flume-demo]$ sudo flume-ng agent --conf-file logs\_to\_hdfs.conf --name agent1 --conf $FLUME\_HOME/conf -Dflume.root.logger=INFO,console

**Solution 2: Using Taildir type**

logs\_to\_hdfs\_taildir.conf

agent1.sources = source1

agent1.sinks = sink1

agent1.channels = channel1

# Describe/configure the source

agent1.sources.source1.type = TAILDIR

agent1.sources.source1.positionFile = /var/log/flume/taildir\_position.json

agent1.sources.source1.filegroups = f1 f2

agent1.sources.source1.filegroups.f1 = /var/log/hadoop-hdfs/.\*log.\*

agent1.sources.source1.headers.f1.headerKey1 = value1

agent1.sources.source1.filegroups.f2 = /var/log/hadoop-mapreduce/.\*log.\*

agent1.sources.source1.headers.f2.headerKey1 = value2

agent1.sources.source1.fileHeader = true

# Describe the sink

agent1.sinks.sink1.type = hdfs

agent1.sinks.sink1.hdfs.writeFormat = Text

agent1.sinks.sink1.hdfs.fileType = DataStream

agent1.sinks.sink1.hdfs.path = /tmp/flume

# Use a channel which buffers events in memory

agent1.channels.channel1.type = memory

agent1.channels.channel1.capacity = 1000

agent1.channels.channel1.transactionCapacity = 100

# Bind the source and sink to the channel

agent1.sources.source1.channels = channel1

agent1.sinks.sink1.channel = channel1

[cloudera@quickstart flume-demo]$ sudo flume-ng agent --conf-file logs\_to\_hdfs\_taildir.conf --name agent1 --conf $FLUME\_HOME/conf -Dflume.root.logger=INFO,console

