YCBS-257 - Data at Scale

Hadoop Data ingestion Workshop - Keys

Apache Flume

```
$ mkdir /home/cloudera/flume

$ mkdir /home/cloudera/flume/spooldirsource
```

Exercise 1:

```
# Agent Components names
agtex1.sources =      source_ncat
agtex1.sinks =        log_sink
agtex1.channels =     mem_channel

# Describing the source
# Listen to port 41415 on the locale machine
agtex1.sources.source_ncat.type =     netcat
agtex1.sources.source_ncat.bind =     localhost
agtex1.sources.source_ncat.port =     41415
agtex1.sources.source_ncat.channels =  mem_channel

# Sink Description
# Logger -> only on screen
agtex1.sinks.log_sink.type =    logger
agtex1.sinks.log_sink.channel = mem_channel

# Describe the Channel
agtex1.channels.mem_channel.type =    memory
agtex1.channels.mem_channel.capacity =     10000
agtex1.channels.mem_channel.transactionCapacity = 100



# Agent Components names
agtex1.sources =      source_ncat
agtex1.sinks =        hdfs_sink
agtex1.channels =     mem_channel

# Describing the source
# Listen to port 41415 on the locale machine
agtex1.sources.source_ncat.type =     netcat
agtex1.sources.source_ncat.bind =     localhost
agtex1.sources.source_ncat.port =     41415
agtex1.sources.source_ncat.channels =  mem_channel

# Sink Description
# HDFS Sink description
agtex1.sinks.hdfs_sink.type =   hdfs
agtex1.sinks.hdfs-sink.hdfs.path = /tmp/flume-spooldir/
agtex1.sinks.hdfs-sink.hdfs.fileType = DataStream
agtex1.sinks.hdfs_sink.channel =     mem_channel

# Describe the Channel
agtex1.channels.mem_channel.type =   memory
agtex1.channels.mem_channel.capacity =     10000
agtex1.channels.mem_channel.transactionCapacity = 100
```

Exercise 2:

```
# Name the components of this agent
agtex2.sources = spooldir-source
agtex2.channels = mem-channel
agtex2.sinks = hdfs-sink

# Source
agtex2.sources.spooldir-source.type = spooldir
agtex2.sources.spooldir-source.spoolDir =
/home/cloudera/flume/spooldirsource
agtex2.sources.spooldir-source.fileHeader = false
agtex2.sources.spooldir-source.channels = mem-channel

# Channel
agtex2.channels.mem-channel.type = memory
agtex2.channels.mem-channel.capacity = 10000
agtex2.channels.mem-channel.transactionCapacity = 100

# Sink
agtex2.sinks.hdfs-sink.type = hdfs
agtex2.sinks.hdfs-sink.hdfs.path =
hdfs://localhost:8020/tmp/spooldir/
agtex2.sinks.hdfs-sink.hdfs.fileType = DataStream
agtex2.sinks.hdfs-sink.channel = mem-channel
```

Extra exercise: Routing Events based on criteria

```
agt1.sources = src1
agt1.sinks = HDFS_Hadoop HDFS_Spark
agt1.channels = MemChannel_Hadoop MemChannel_Spark

# Describe/configure the source
agt1.sources.src1.type = netcat
agt1.sources.src1.bind = localhost
agt1.sources.src1.port = 45454

agt1.sources.src1.interceptors = i1
agt1.sources.src1.interceptors.i1.type = regex_extractor

agt1.sources.src1.interceptors.i1.regex = (Hadoop|Spark)
agt1.sources.src1.interceptors.i1.serializers = s1
agt1.sources.src1.interceptors.i1.serializers.s1.name = BigData

agt1.sources.src1.selector.type = multiplexing
agt1.sources.src1.selector.header = BigData
agt1.sources.src1.selector.mapping.Hadoop = MemChannel_Hadoop
agt1.sources.src1.selector.mapping.Spark = MemChannel_Spark

# Bind the source and sink to the channel
agt1.sources.src1.channels = MemChannel_Hadoop MemChannel_Spark

# Use a channel which buffers events in memory
agt1.channels.MemChannel_Hadoop.type = memory
agt1.channels.MemChannel_Hadoop.capacity = 1000
agt1.channels.MemChannel_Hadoop.transactionCapacity = 100

agt1.channels.MemChannel_Spark.type = memory
agt1.channels.MemChannel_Spark.capacity = 1000
agt1.channels.MemChannel_Spark.transactionCapacity = 100

agt1.sinks.HDFS_Hadoop.channel = MemChannel_Hadoop
agt1.sinks.HDFS_Hadoop.type = hdfs
agt1.sinks.HDFS_Hadoop.hdfs.path = /user/flume/Hadoop
agt1.sinks.HDFS_Hadoop.hdfs.fileType = DataStream
agt1.sinks.HDFS_Hadoop.hdfs.writeFormat = Text
agt1.sinks.HDFS_Hadoop.hdfs.batchSize = 1000
agt1.sinks.HDFS_Hadoop.hdfs.rollSize = 0
agt1.sinks.HDFS_Hadoop.hdfs.rollCount = 100000

agt1.sinks.HDFS_Spark.channel = MemChannel_Spark
agt1.sinks.HDFS_Spark.type = hdfs
agt1.sinks.HDFS_Spark.hdfs.path = /user/flume/Spark
agt1.sinks.HDFS_Spark.hdfs.fileType = DataStream
agt1.sinks.HDFS_Spark.hdfs.writeFormat = Text
agt1.sinks.HDFS_Spark.hdfs.batchSize = 1000
agt1.sinks.HDFS_Spark.hdfs.rollSize = 0
agt1.sinks.HDFS_Spark.hdfs.rollCount = 100000
```