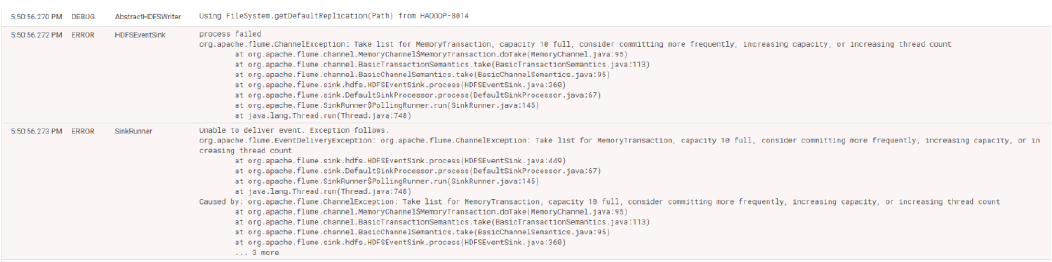
Для части по потоковой обработке (Flume)

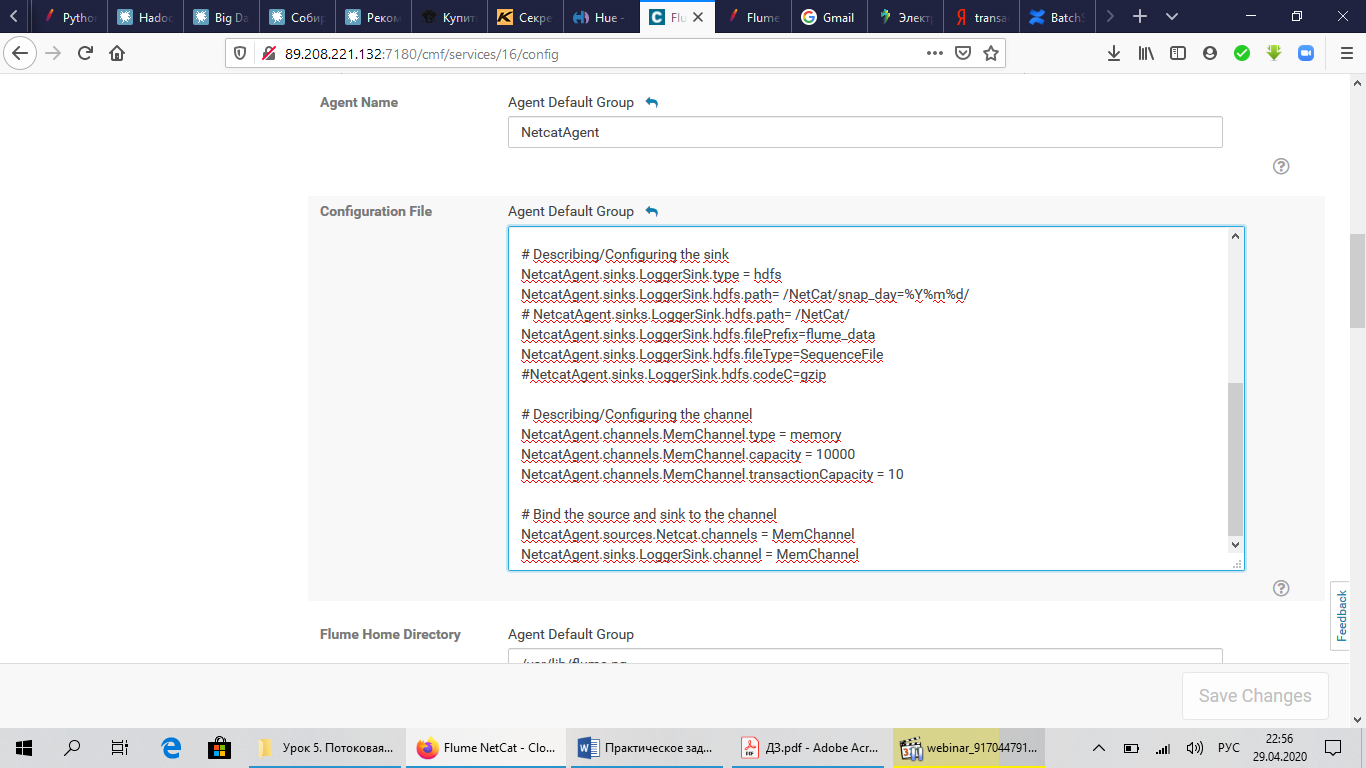
1. Посмотреть, что не так в конфигурации NetCat Flume agent которого я сделал. Описать и аргументировать.

Смотрим логи по NetCat Flume agent:

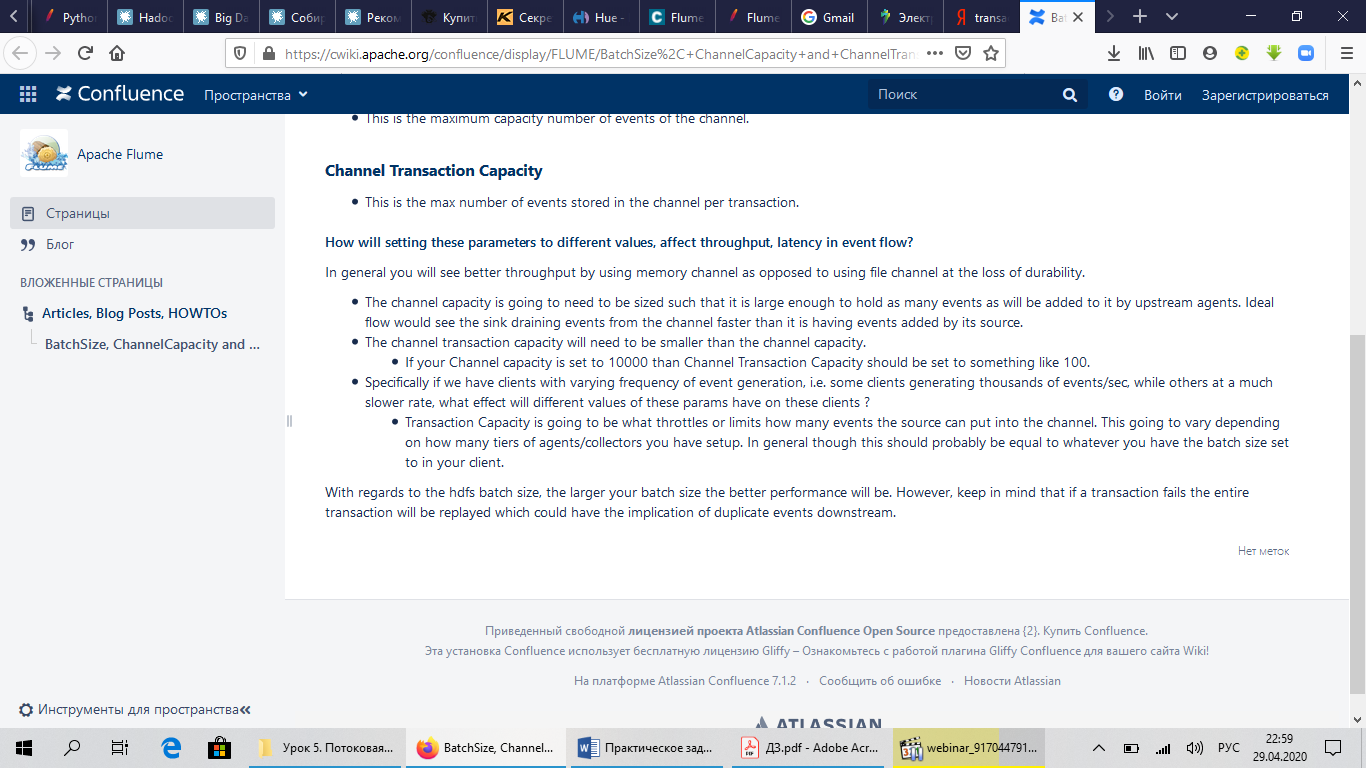


В логах есть ошибки, связанные с переполнением (capacity 10 full, increasing capacity or increasing thread count).

В конфигурации находим источник проблем:



В wiki было найдено решение (<https://cwiki.apache.org/confluence/display/FLUME/BatchSize%2C+ChannelCapacity+and+ChannelTransactionCapacity+Properties>):



Необходимо поменять параметр NetcatAgent.channels.MemChannel.transactionCapacity на 100.

2. Создать любой Flume поток используя Flume сервис соответствующего номера.

• Тип источника источник – exeс

• Тип канала – memory

• Тип слива – hdfs

Создал Flume поток. Конфигурация ниже:

# Naming the components on the current agent

Agent6.sources = Exec

Agent6.channels = MemChannel

Agent6.sinks = HDFSSink

# Describing/Configuring the source

Agent6.sources.Exec.type = exec

Agent6.sources.Exec.command = /usr/bin/tailf /tmp/myfile2

Agent6.sources.Exec.logStdErr = true

# Describing/Configuring the HDFS sink

Agent6.sinks.HDFSSink.type = hdfs

Agent6.sinks.HDFSSink.channel = MemChannel

Agent6.sinks.HDFSSink.hdfs.useLocalTimeStamp = true

Agent6.sinks.HDFSSink.hdfs.path = /user/student3\_6/flume/%y-%m-%d

Agent6.sinks.HDFSSink.hdfs.filePrefix = hdfs-st310-

Agent6.sinks.HDFSSink.hdfs.rollSize = 1000

Agent6.sinks.HDFSSink.hdfs.rollInterval = 0

Agent6.sinks.HDFSSink.hdfs.rollCount = 0

Agent6.sinks.HDFSSink.hdfs.fileType = SequenceFile

Agent6.sinks.HDFSSink.hdfs.codeC = gzip

# Describing/Configuring the channel

Agent6.channels.MemChannel.type = memory

Agent6.channels.MemChannel.capacity=1000

Agent6.channels.MemChannel.transactionCapacity=100

# Bind the source and sink to the channel

Agent6.sources.Exec.channels = MemChannel

Agent6.sinks.HDFSSink.channel = MemChannel

3. [Продвинутый вариант] Сделать то-же самое используя несколько сливов в разные места, например, в HDFS и в HIve одновременно

4. [Продвинутый вариант] Повторить стандартный пример с выборкой сообщений из Twitter. Перед этим связаться со мной :)