BIG DATA STRATEGIES

APACHE FLUME

Presented by
Megha Navin
Akhila Saladi
Manpreet Kaur
Nirmala Shaymala

WHAT IS APACHE FLUME?

OUR MAIN TOPIC TODAY

Apache Flume is a data ingestion tool in the Hadoop World. It collects, aggregates and moves large amount of streaming data from various data sources to a centralized data store like HDFS

FEATURES OF APACHE FLUME

RELIABILITY	SCALABILITY	EXTENSIBILITY	MANAGEABILITY
Agents installed on many machines Scalability	Add more machines to transfer more events	Durable storage, failover and/or replication	Easy to install configure, reconfigure and run

FLUME VS SQOOP

FLUME

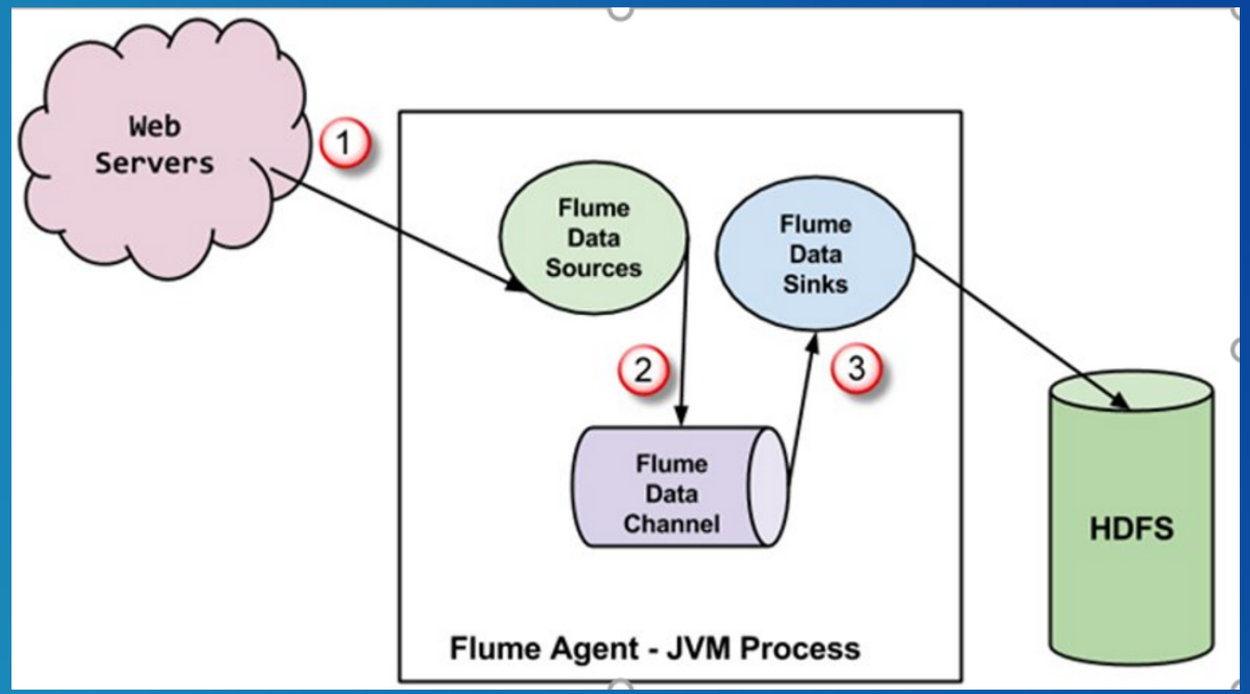
Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data. Flume helps to collect data from a variety of sources, like logs, jms, Directory etc. Multiple flume agents can be configured to collect high volume of data.It scales horizontally.

SQOOP

Apache Sqoop(TM) is a tool designed for efficiently transferring bulk data between Apache Hadoop and structured datastores such as relational databases. Sqoop helps to move data between hadoop and other databases and it can transfer data in parallel for performance.

FLUME ARCHITECTURE

- FLUME SOURCE
- FLUME CHANNEL
- FLUME SINK



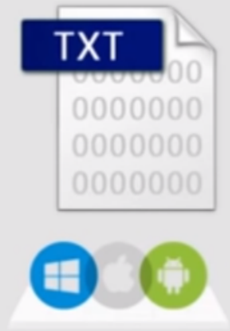
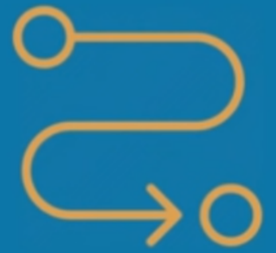
SO…HOW DOES IT WORK???

The events generated by external source (WebServer) are consumed by Flume Data Source. The external source sends events to Flume source in a format that is recognized by the target source.

Source receives an event and stores it into one or more channels. The channel acts as a store which keeps the event until it is consumed by the flume sink. This channel may use a local file system in order to store these events.

Flume sink removes the event from a channel and stores it into an external repository like e.g., HDFS. There could be multiple flume agents, in which case flume sink forwards the event to the flume source of next flume agent in the flow.

ADVANTAGES OF APACHE FLUME

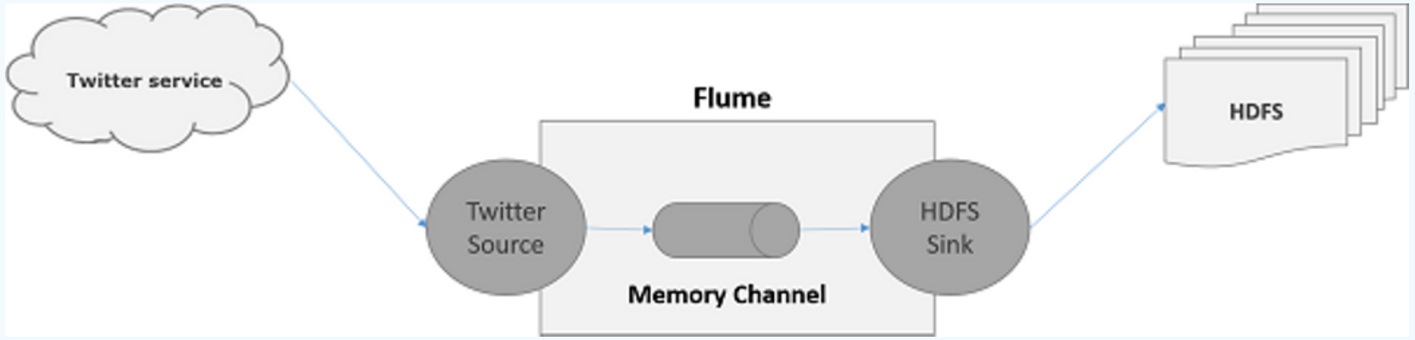
	Scalable		Reliable		Fault Tolerant
Centralized Data Storage		Data Ingestion		Large Data Sets	
	Multiple hop flows		Real-time data streaming		Source & Destination

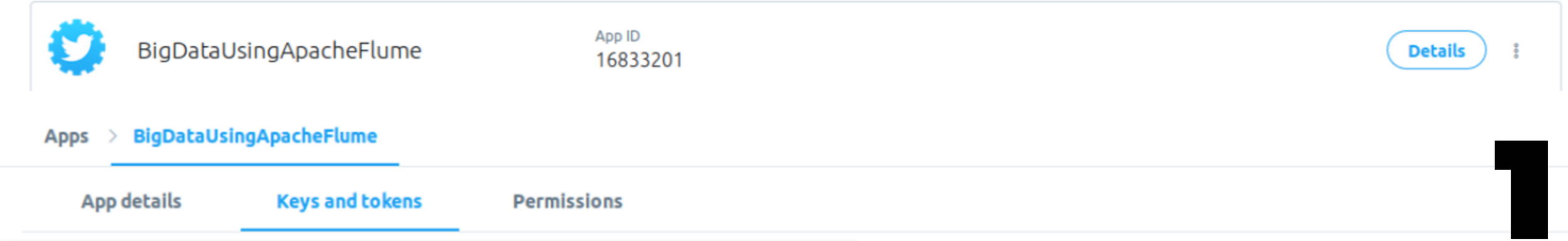
TWITTER EXAMPLE

we have created an application and to get the tweets from it using the experimental twitter source provided by Apache Flume. We will use the memory channel to buffer these tweets and HDFS sink to push these tweets into the HDFS.

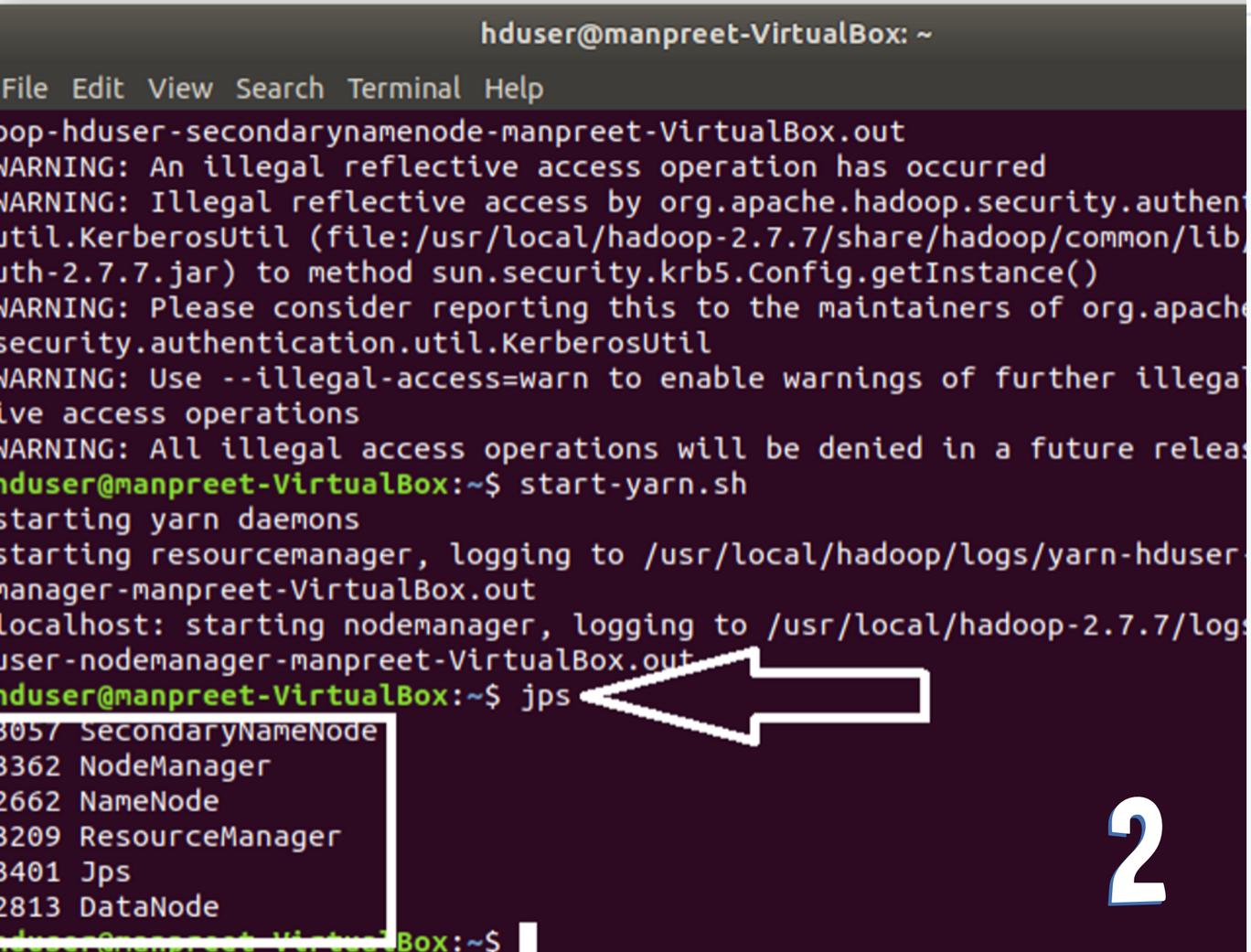
Steps Involved:

- Create a twitter Application
- Install / Start HDFS
- Configure Flume

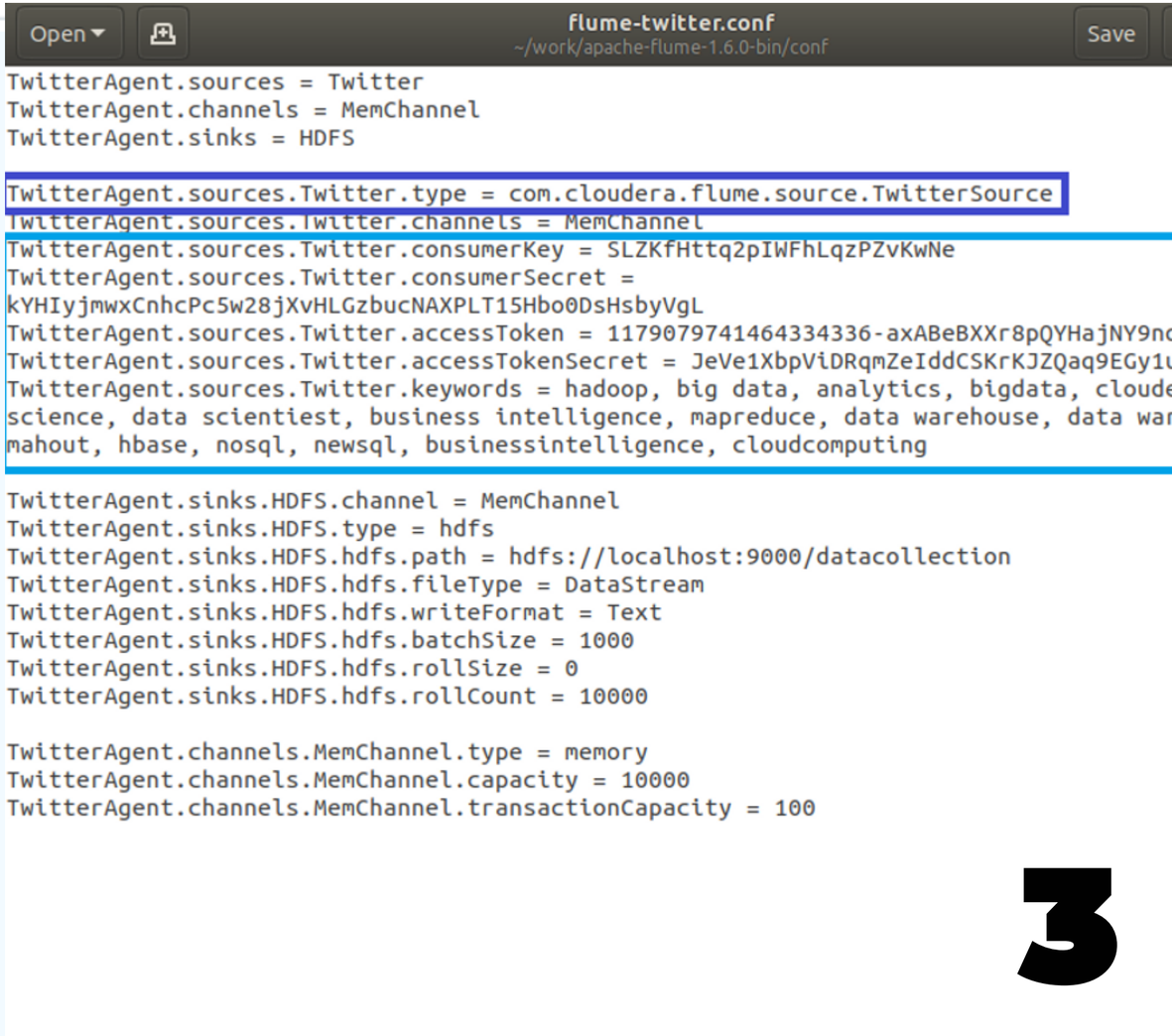



BigDataUsingApacheFlume
App ID
16833201
Details
Apps
BigDataUsingApacheFlume
App details
Keys and tokens
Permissions

1


hduser@manpreet-VirtualBox: ~
File Edit View Search Terminal Help
oop-hduser-secondarynamenode-manpreet-VirtualBox.out
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authen
util.KerberosUtil (file:/usr/local/hadoop-2.7.7/share/hadoop/common/lib/
uth-2.7.7.jar) to method sun.security.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache
security.authentication.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal
ive access operations
WARNING: All illegal access operations will be denied in a future releas
hduser@manpreet-VirtualBox:~$ start-yarn.sh
starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop/logs/yarn-hduser-
manager-manpreet-VirtualBox.out
localhost: starting nodemanager, logging to /usr/local/hadoop-2.7.7/logs
user-nodemanager-manpreet-VirtualBox.out
hduser@manpreet-VirtualBox:~$ jps
3057 SecondaryNameNode
3362 NodeManager
2662 NameNode
3209 ResourceManager
3401 Jps
2813 DataNode

2


Open
flume-twitter.conf
~/work/apache-flume-1.6.0-bin/conf
Save
TwitterAgent.sources = Twitter
TwitterAgent.channels = MemChannel
TwitterAgent.sinks = HDFS
TwitterAgent.sources.Twitter.type = com.cloudera.flume.source.TwitterSource
TwitterAgent.sources.Twitter.channels = MemChannel
TwitterAgent.sources.Twitter.consumerKey = SLZKfHttq2pIWFhLqzPZvKwNe
TwitterAgent.sources.Twitter.consumerSecret =
kYHIyjmwxCnhcPc5w28jXvHLGzbucNAXPLT15Hbo0DsHsbyVgL
TwitterAgent.sources.Twitter.accessToken = 1179079741464334336-axABeBXXr8pQYHajNY9nc
TwitterAgent.sources.Twitter.accessTokenSecret = JeVe1XbpViDRqmZeIddCSKrKJZQaq9EGy1u
TwitterAgent.sources.Twitter.keywords = hadoop, big data, analytics, bigdata, cloude
science, data scientiest, business intelligence, mapreduce, data warehouse, data war
mahout, hbase, nosql, newsql, businessintelligence, cloudcomputing
TwitterAgent.sinks.HDFS.channel = MemChannel
TwitterAgent.sinks.HDFS.type = hdfs
TwitterAgent.sinks.HDFS.hdfs.path = hdfs://localhost:9000/datacollection
TwitterAgent.sinks.HDFS.hdfs.fileType = DataStream
TwitterAgent.sinks.HDFS.hdfs.writeFormat = Text
TwitterAgent.sinks.HDFS.hdfs.batchSize = 1000
TwitterAgent.sinks.HDFS.hdfs.rollSize = 0
TwitterAgent.sinks.HDFS.hdfs.rollCount = 10000
TwitterAgent.channels.MemChannel.type = memory
TwitterAgent.channels.MemChannel.capacity = 10000
TwitterAgent.channels.MemChannel.transactionCapacity = 100

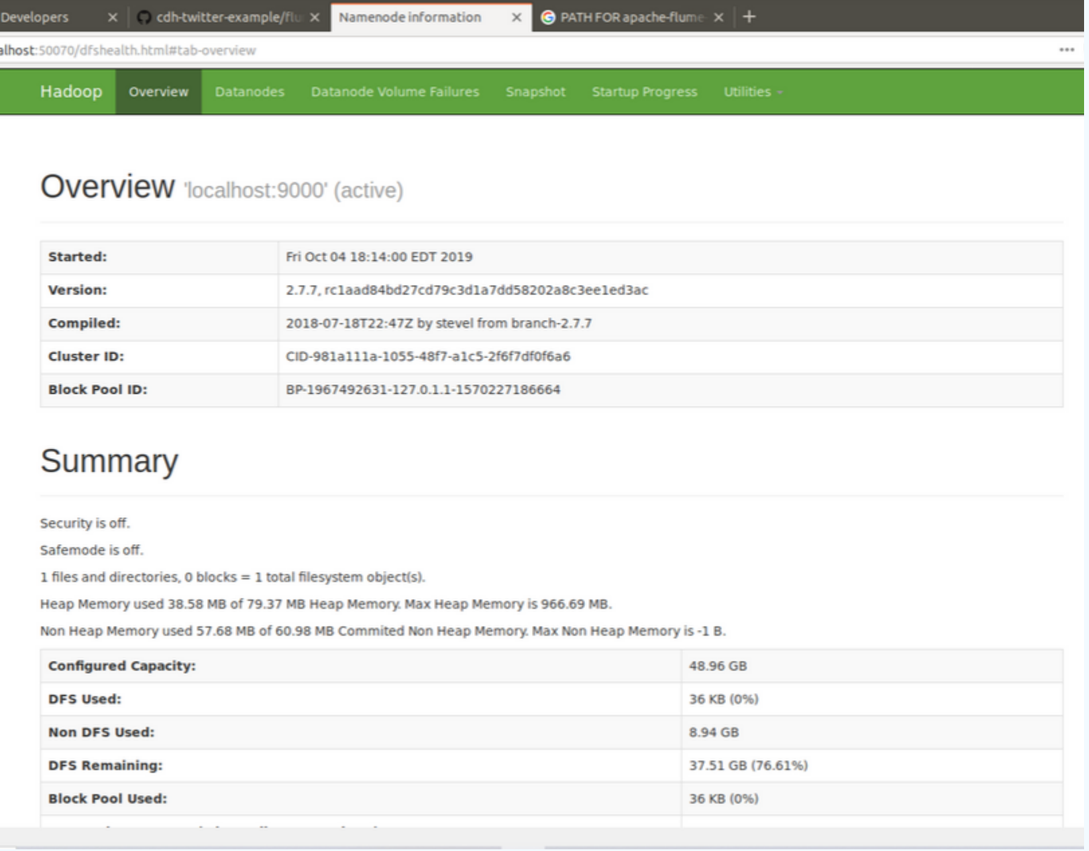
3

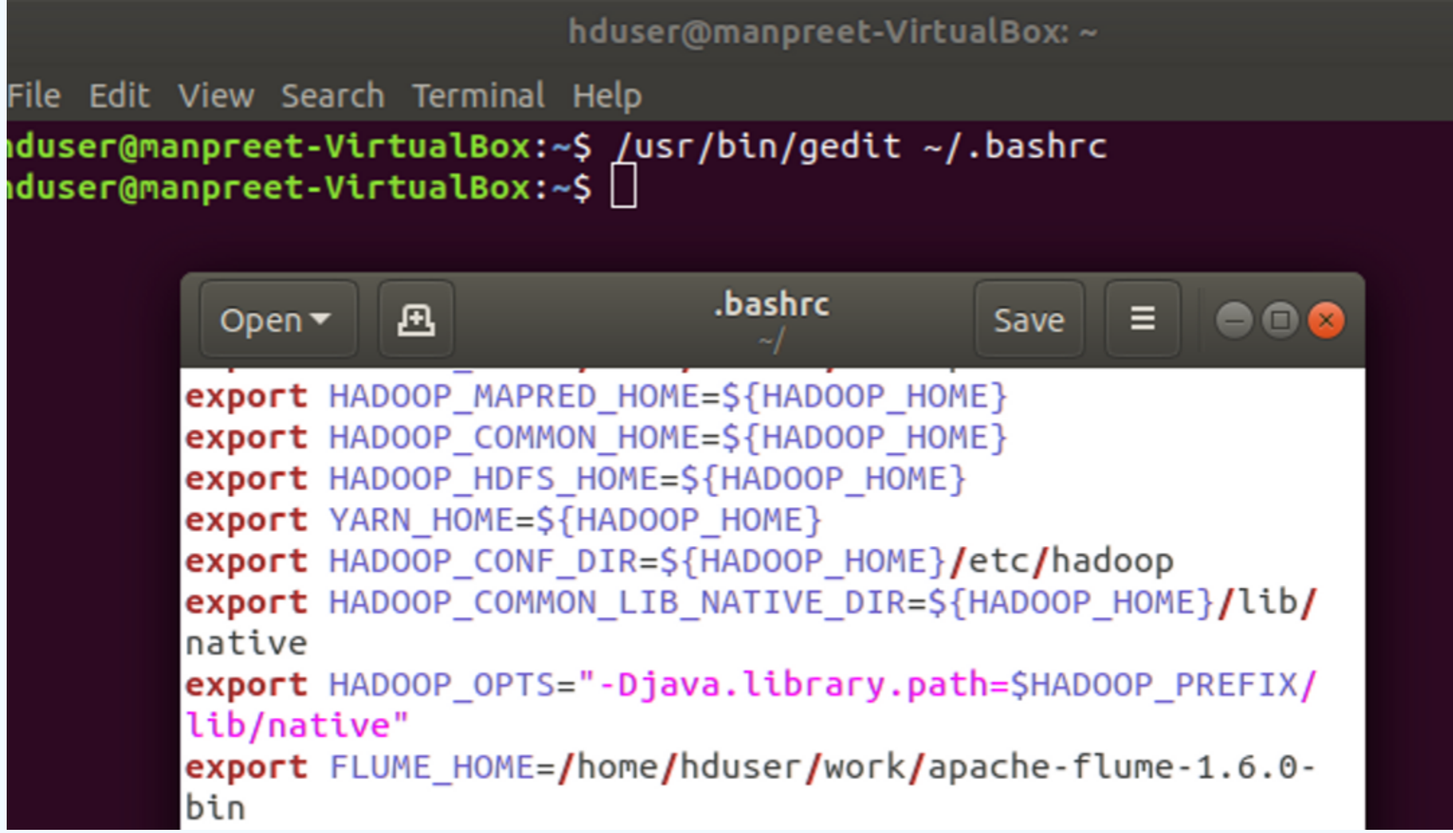
FLUME COMMANDS

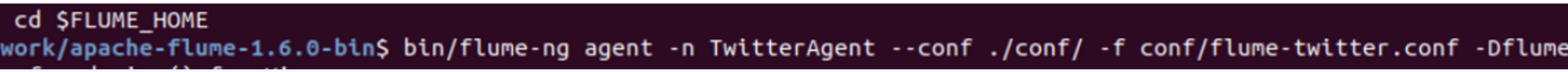
STARTING A FLUME AGENT

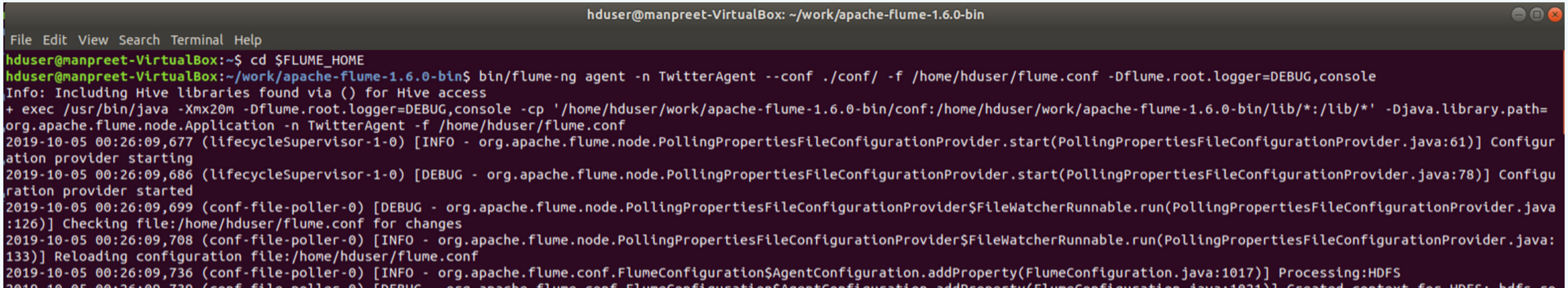
flume-ng agent -n agentName -f conf/flumexample.com -Dflume.root.logger=Info,CONSOLE

- agent - Command to start the Flume agent
- --name, -n <name> - Name of the twitter agent
- --conf, -c<conf> - Use configuration file in conf directory
- -f<file> - Specifies a config file path
- -D property = value - Sets a java system property value


Developers
cdh-twitter-example/flu
Namenode information
PATH FOR apache-flume
alhost:50070/dfshealth.html#tab-overview
Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Utilities
Overview 'localhost:9000' (active)
Started: Fri Oct 04 18:14:00 EDT 2019
Version: 2.7.7, rc1aad84bd27cd79c3d1a7dd58202a8c3ee1ed3ac
Compiled: 2018-07-18T22:47Z by stevel from branch-2.7.7
Cluster ID: CID-981a111a-1055-48f7-a1c5-2f6f7df0f6a6
Block Pool ID: BP-1967492631-127.0.1.1-1570227186664
Summary
Security is off.
Safemode is off.
1 files and directories, 0 blocks = 1 total filesystem object(s).
Heap Memory used 38.58 MB of 79.37 MB Heap Memory. Max Heap Memory is 966.69 MB.
Non Heap Memory used 57.68 MB of 60.98 MB Commited Non Heap Memory. Max Non Heap Memory is -1 B.
Configured Capacity: 48.96 GB
DFS Used: 36 KB (0%)
Non DFS Used: 8.94 GB
DFS Remaining: 37.51 GB (76.61%)
Block Pool Used: 36 KB (0%)


hduser@manpreet-VirtualBox: ~
File Edit View Search Terminal Help
hduser@manpreet-VirtualBox:~$ /usr/bin/gedit ~/.bashrc
hduser@manpreet-VirtualBox:~$
Open
.bashrc
~/
Save
export HADOOP_MAPRED_HOME=${HADOOP_HOME}
export HADOOP_COMMON_HOME=${HADOOP_HOME}
export HADOOP_HDFS_HOME=${HADOOP_HOME}
export YARN_HOME=${HADOOP_HOME}
export HADOOP_CONF_DIR=${HADOOP_HOME}/etc/hadoop
export HADOOP_COMMON_LIB_NATIVE_DIR=${HADOOP_HOME}/lib/native
export HADOOP_OPTS="-Djava.library.path=$HADOOP_PREFIX/lib/native"
export FLUME_HOME=/home/hduser/work/apache-flume-1.6.0-bin


cd $FLUME_HOME
work/apache-flume-1.6.0-bin$ bin/flume-ng agent -n TwitterAgent --conf ./conf/ -f conf/flume-twitter.conf -Dflume


hduser@manpreet-VirtualBox: ~/work/apache-flume-1.6.0-bin
File Edit View Search Terminal Help
hduser@manpreet-VirtualBox:~$ cd $FLUME_HOME
hduser@manpreet-VirtualBox:~/work/apache-flume-1.6.0-bin$ bin/flume-ng agent -n TwitterAgent --conf ./conf/ -f /home/hduser/flume.conf -Dflume.root.logger=DEBUG,console
Info: Including Hive libraries found via () for Hive access
+ exec /usr/bin/java -Xmx20m -Dflume.root.logger=DEBUG,console -cp '/home/hduser/work/apache-flume-1.6.0-bin/conf:/home/hduser/work/apache-flume-1.6.0-bin/lib/*:/lib/*' -Djava.library.path=
org.apache.flume.node.Application -n TwitterAgent -f /home/hduser/flume.conf
2019-10-05 00:26:09,677 (lifecycleSupervisor-1-0) [INFO - org.apache.flume.node.PollingPropertiesFileConfigurationProvider.start(PollingPropertiesFileConfigurationProvider.java:61)] Configur
ation provider starting
2019-10-05 00:26:09,686 (lifecycleSupervisor-1-0) [DEBUG - org.apache.flume.node.PollingPropertiesFileConfigurationProvider.start(PollingPropertiesFileConfigurationProvider.java:78)] Configu
ration provider started
2019-10-05 00:26:09,699 (conf-file-poller-0) [DEBUG - org.apache.flume.node.PollingPropertiesFileConfigurationProvider$FileWatcherRunnable.run(PollingPropertiesFileConfigurationProvider.java
:126)] Checking file:/home/hduser/flume.conf for changes
2019-10-05 00:26:09,708 (conf-file-poller-0) [INFO - org.apache.flume.node.PollingPropertiesFileConfigurationProvider$FileWatcherRunnable.run(PollingPropertiesFileConfigurationProvider.java:
133)] Reloading configuration file:/home/hduser/flume.conf
2019-10-05 00:26:09,736 (conf-file-poller-0) [INFO - org.apache.flume.conf.FlumeConfiguration$AgentConfiguration.addProperty(FlumeConfiguration.java:1017)] Processing:HDFS

REFERENCES

Hadoop configuration

https://www.youtube.com/watch?v=Y6oit3rCsZo
IonT's Mindhouse Channel

Apache Flume Configuration

https://www.youtube.com/watch?v=Vb49rhlXFYA&t=107s
Hadoop Home Channel

Other resources

Webpage :
https://www.tutorialspoint.com/apache_flume/fetching_twitter_data.htm
Coso IT (youtube)
https://www.youtube.com/watch?v=l6KP2TRIhYY

THANKS FOR YOUR TIME!