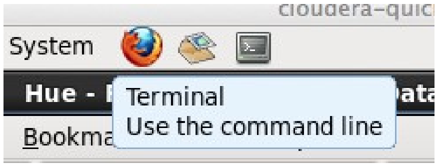
FLUME LAB

Overview: Flume brings data into Hadoop from log streams such as Avro, Thrift, Syslog, and **Netcat**. Cloudera Quickstart VM comes with Flume already installed. We will start the Flume service which, by default, is configured to listen to network traffic on port 9999 using a Netcat service. We will alter the configuration to add a sink that will write the input to Netcat to a file in HDFS.

1) In Cloudera Manager, start the Flume service.

2) Open a Terminal session. The icon to do this is at the top of your VM window.



3) In the Terminal, start a netcat client with “nc 127.0.0.1 9999”. Now type in a few lines, hitting Enter after each one. You should receive an “OK” message each time.

4) We want to verify that Flume's netcat is listening. Open another Terminal window and “tail -f “ the log file in your /var/log/flume-ng directory. The inputs you entered in the previous step should be seen here.

5) Now we will change the Flume configuration in Cloudera Manager to make it write to HDFS. We will do this by configuring a second sink.
   Cloudera Manager > Flume > Configuration > Agent > Configuration File

a) change the "tier1.sinks" to add the value "sink2" after sink1 separated by a space.

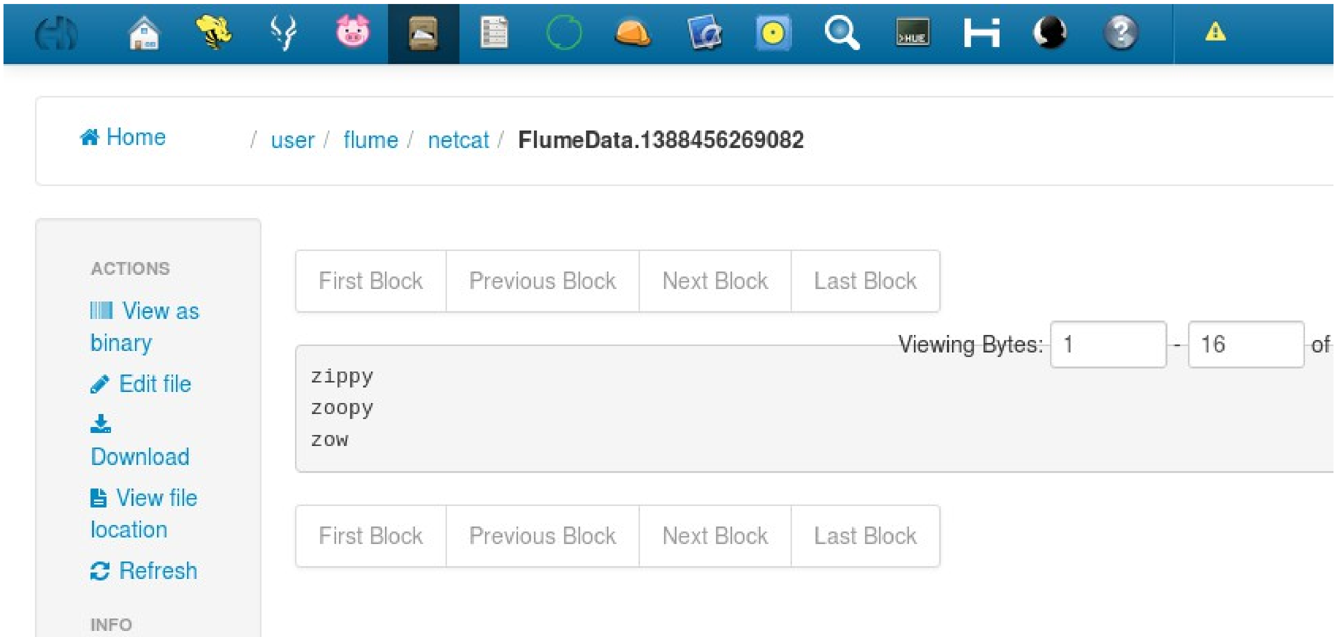
b) Add the following properties to define the new sink:

```
tier1.sink.sink2.channel    = channel1
tier1.sinks.sink2.type        = hdfs
tier1.sinks.sink2.hdfs.path = /user/flume/netcat
tier1.sinks.sink2.hdfs.fileType = DataStream
```

6) Restart the Flume Service in Cloudera Manager

7) Add a few more inputs to your netcat client Terminal.

8) In Hue > File Manager you should be able to navigate to /user/flume/netcat and see a file there that contains the inputs from your latest netcat session. If it's not there try refreshing your browser and checking your tailed log for errors.



Home / user / flume / netcat / FlumeData.1388456269082

Viewing Bytes: 1 - 16 of

```
zippy
zoopy
zow
```

Seeing something like this? Great job!

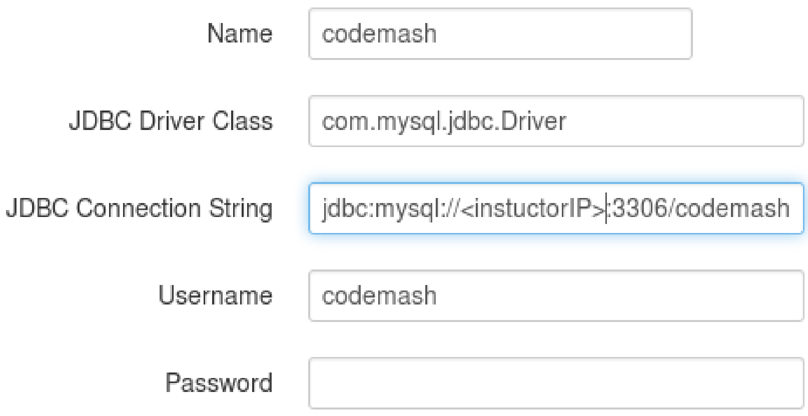
SQOOP LAB

Overview: Sqoop is used to transfer data between HDFS and outside structured databases. We will use Sqoop to connect to a MySQL instance running on the instructor's machine and pull data from a table there into your VM's HDFS.

1) In Cloudera Manager, start the Sqoop service if it is not already running.

2) In Hue > Sqoop, select Add New Job.

3) Name your job, leave the Job type as “Import”, and add a new connection with the following values. Note that you should replace <instructorIP> with the IP address the instructor gives you and Password should be left blank.



Name	codemash
JDBC Driver Class	com.mysql.jdbc.Driver
JDBC Connection String	jdbc:mysql://<instuctorIP>:3306/codemash
Username	codemash
Password	

(Job creation continued on next page)

4) (cont.) The next step of Job Creation involves setting up information about the MySQL table being read.
   Schema name: codemash
   Table name: names
   Table column names: id,first,last
   Partition column name: id

5) (cont.) The final step of Job Creation defines how to store the input data in HDFS.
   Storage type: HDFS
   Output format: SEQUENCE_FILE
   Output directory: /user/sqoop2/names

6) Save and Run the Job

7) There should be at least a single popup message to say that the job is running. You can “tail -f” the log file with a large name in the /var/log/sqoop2 directory in a Terminal window to look for any problems. It is possible that Hue could end up not giving you any success or error messages and simply fail if, for instance, you forgot to set the “Partition column name” during Job Creation.

8) Look for your new HDFS file(s) under /user/sqoop2

Bonus: Most of the commands that can be run in Hue are simply prettier UI versions of what you could do in a command line. In the case of Sqoop, you may find it significantly easier to just use the command line. Try it:

```
sqoop import --connect jdbc:mysql://<instructorIP>/codemash --username codemash --table fullnames
```

Your Terminal session runs as the cloudera user, so the above command will place your command line sqoop import HDFS data into the /user/**cloudera** directory.

Now, inspect your home directory with “ls ~”. There are now magical java files for your perusal from the Sqoop jobs you ran in this lab! Check them out!