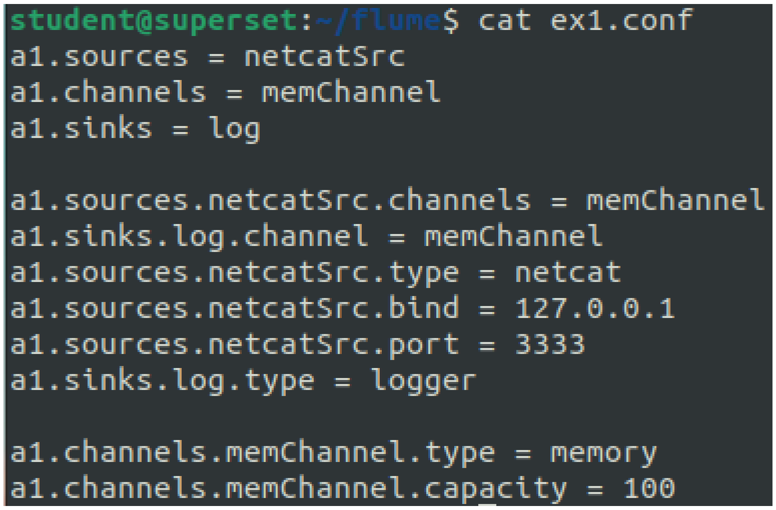
Practice №1. Workshop on Apache Flume.

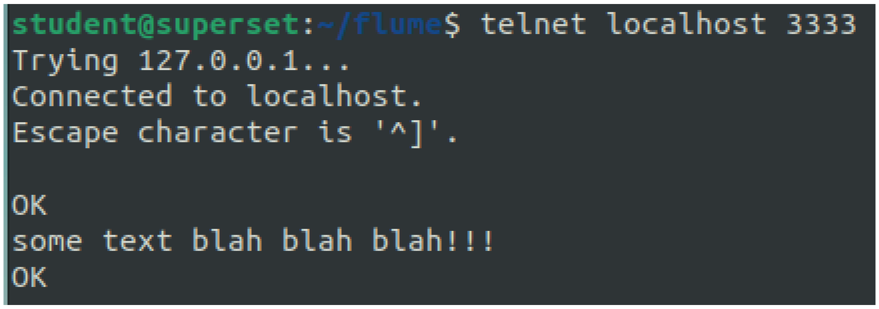
Ex.1 Creating a logger.

For this I used the following configuration file:



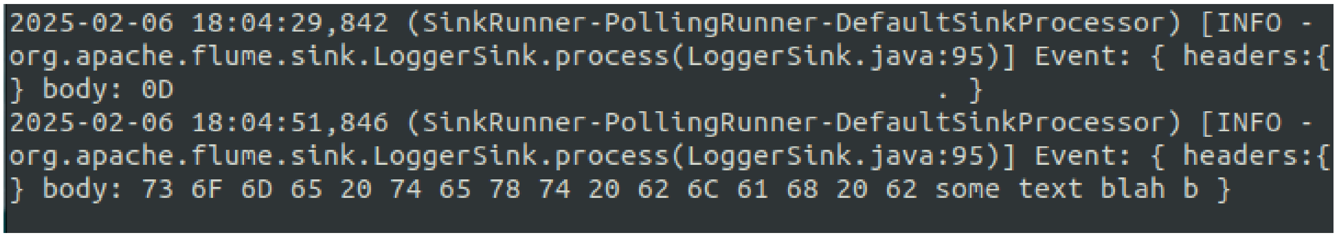
```
student@superset:~/flume$ cat ex1.conf
a1.sources = netcatSrc
a1.channels = memChannel
a1.sinks = log

a1.sources.netcatSrc.channels = memChannel
a1.sinks.log.channel = memChannel
a1.sources.netcatSrc.type = netcat
a1.sources.netcatSrc.bind = 127.0.0.1
a1.sources.netcatSrc.port = 3333
a1.sinks.log.type = logger

a1.channels.memChannel.type = memory
a1.channels.memChannel.capacity = 100
```

and following input



```
student@superset:~/flume$ telnet localhost 3333
Trying 127.0.0.1...
Connected to localhost.
Escape character is '^]'.

OK
some text blah blah blah!!!
OK
```

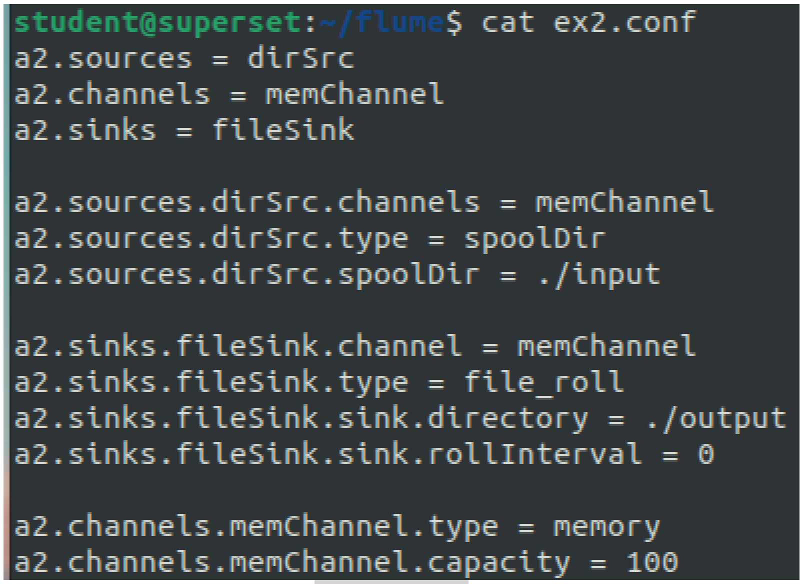
This is how logs look like after running the agent:



```
2025-02-06 18:04:29,842 (SinkRunner-PollingRunner-DefaultSinkProcessor) [INFO -
org.apache.flume.sink.LoggerSink.process(LoggerSink.java:95)] Event: { headers:{
} body: 0D                                              . }
2025-02-06 18:04:51,846 (SinkRunner-PollingRunner-DefaultSinkProcessor) [INFO -
org.apache.flume.sink.LoggerSink.process(LoggerSink.java:95)] Event: { headers:{
} body: 73 6F 6D 65 20 74 65 78 74 20 62 6C 61 68 20 62 some text blah b }
```

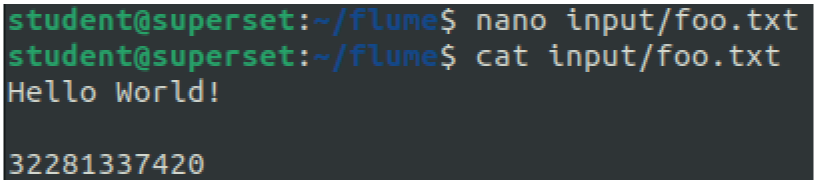
Ex.2 Collecting log files in a specified directory.

For this, I used the following configuration file:



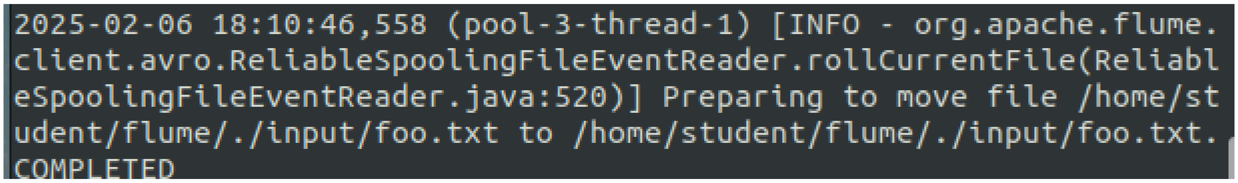
```
student@superset:~/flume$ cat ex2.conf
a2.sources = dirSrc
a2.channels = memChannel
a2.sinks = fileSink

a2.sources.dirSrc.channels = memChannel
a2.sources.dirSrc.type = spoolDir
a2.sources.dirSrc.spoolDir = ./input

a2.sinks.fileSink.channel = memChannel
a2.sinks.fileSink.type = file_roll
a2.sinks.fileSink.sink.directory = ./output
a2.sinks.fileSink.sink.rollInterval = 0

a2.channels.memChannel.type = memory
a2.channels.memChannel.capacity = 100
```

Here, the sink directory is `./output` and the source directory is `./input`, which were manually made beforehand and `./input` in particular was populated with text data:

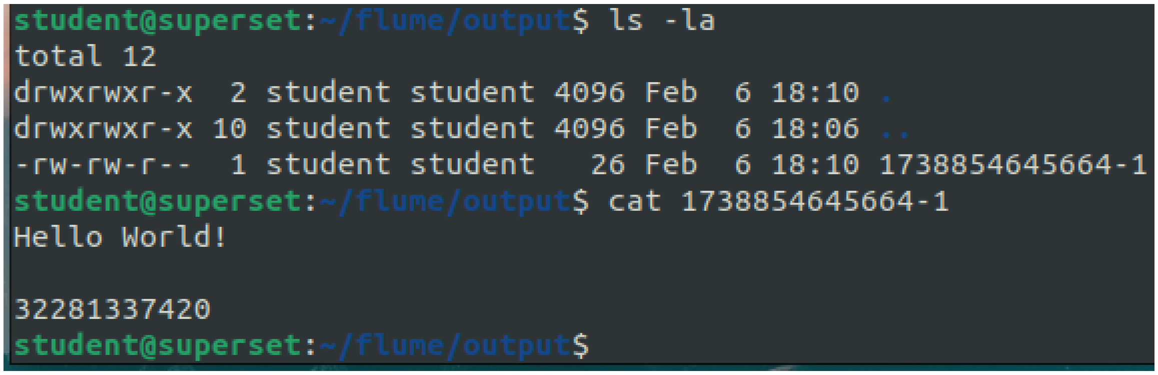


```
student@superset:~/flume$ nano input/foo.txt
student@superset:~/flume$ cat input/foo.txt
Hello World!

32281337420
```

After running the agent we see that it registers the files and successfully transfers them:



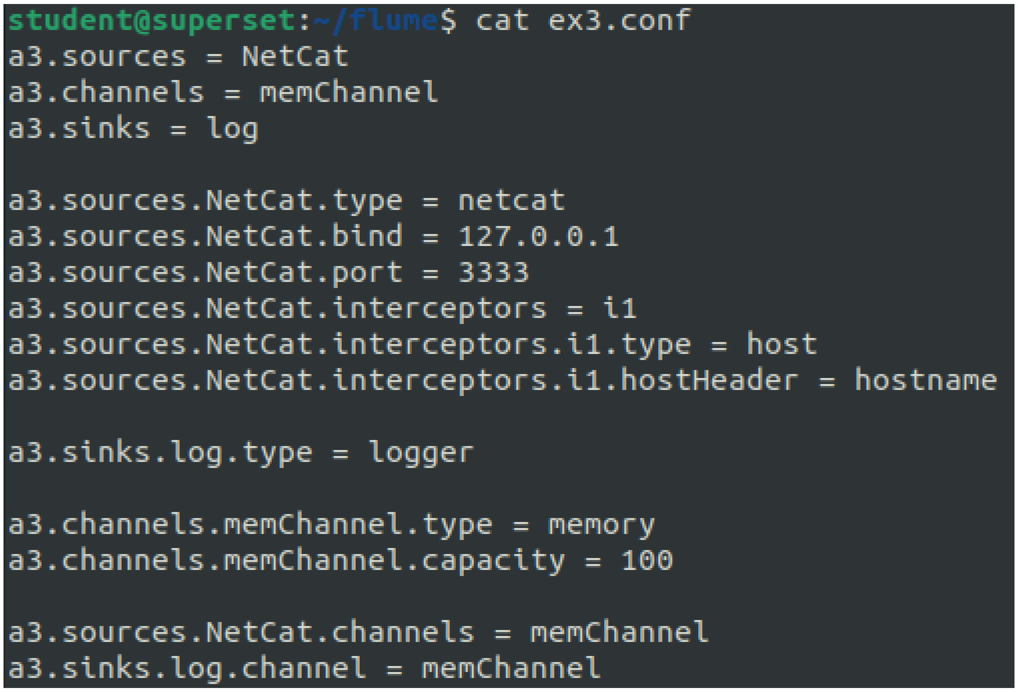
```
2025-02-06 18:10:46,558 (pool-3-thread-1) [INFO - org.apache.flume.
client.avro.ReliableSpoolingFileEventReader.rollCurrentFile(Reliabl
eSpoolingFileEventReader.java:520)] Preparing to move file /home/st
udent/flume/./input/foo.txt to /home/student/flume/./input/foo.txt.
COMPLETED
```



```
student@superset:~/flume/output$ ls -la
total 12
drwxrwxr-x  2 student student 4096 Feb  6 18:10 .
drwxrwxr-x 10 student student 4096 Feb  6 18:06 ..
-rw-rw-r--  1 student student   26 Feb  6 18:10 1738854645664-1
student@superset:~/flume/output$ cat 1738854645664-1
Hello World!

32281337420
student@superset:~/flume/output$
```

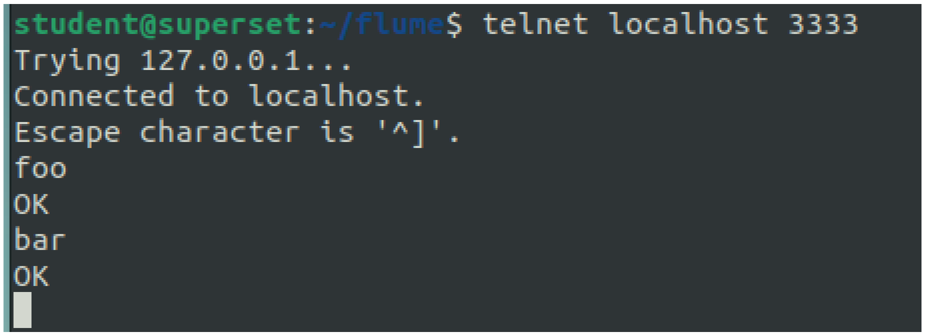
Ex.3 Using an interceptor to process incoming events.

For this I used the following configuration file(it adds hostname at the beginning):



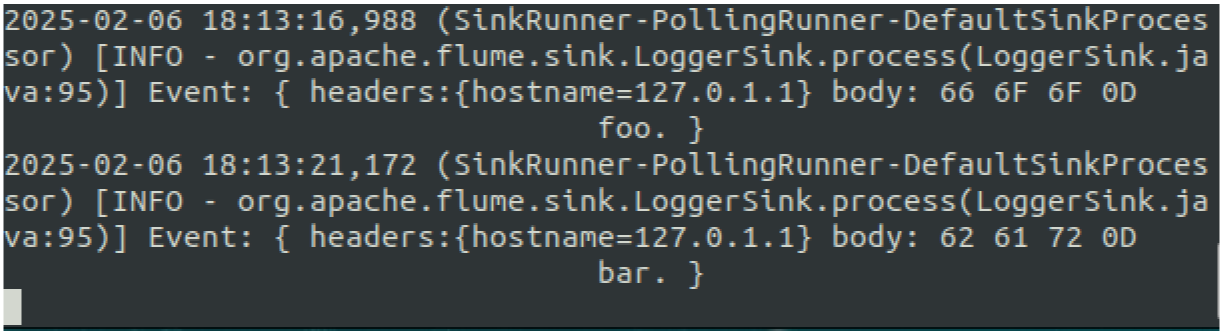
```
student@superset:~/flume$ cat ex3.conf
a3.sources = NetCat
a3.channels = memChannel
a3.sinks = log

a3.sources.NetCat.type = netcat
a3.sources.NetCat.bind = 127.0.0.1
a3.sources.NetCat.port = 3333
a3.sources.NetCat.interceptors = i1
a3.sources.NetCat.interceptors.i1.type = host
a3.sources.NetCat.interceptors.i1.hostHeader = hostname

a3.sinks.log.type = logger

a3.channels.memChannel.type = memory
a3.channels.memChannel.capacity = 100

a3.sources.NetCat.channels = memChannel
a3.sinks.log.channel = memChannel
```

With following input:



```
student@superset:~/flume$ telnet localhost 3333
Trying 127.0.0.1...
Connected to localhost.
Escape character is '^]'.
foo
OK
bar
OK
```

There results were as follows:

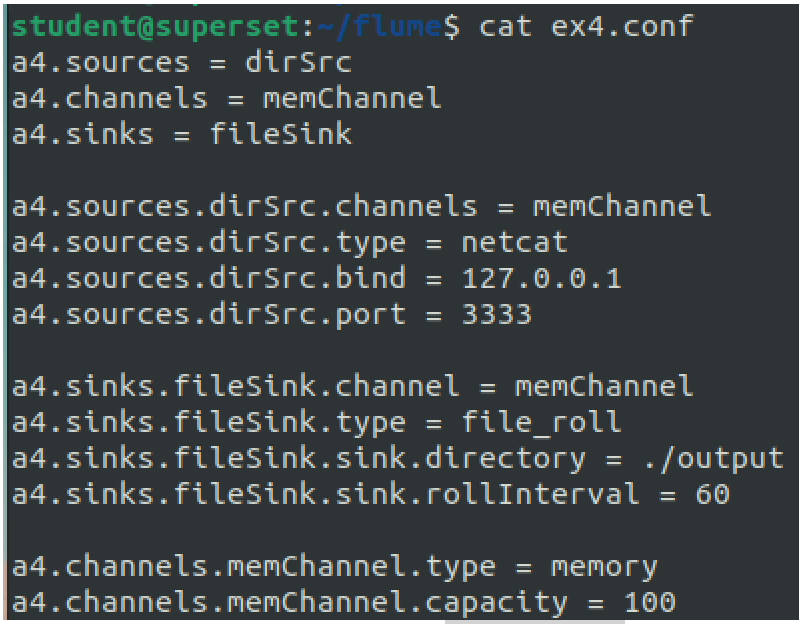


```
2025-02-06 18:13:16,988 (SinkRunner-PollingRunner-DefaultSinkProces
sor) [INFO - org.apache.flume.sink.LoggerSink.process(LoggerSink.ja
va:95)] Event: { headers:{hostname=127.0.1.1} body: 66 6F 6F 0D
                                  foo. }
2025-02-06 18:13:21,172 (SinkRunner-PollingRunner-DefaultSinkProces
sor) [INFO - org.apache.flume.sink.LoggerSink.process(LoggerSink.ja
va:95)] Event: { headers:{hostname=127.0.1.1} body: 62 61 72 0D
                                  bar. }
```

One can easily see that the input was indeed inctercepted and hostname was added as its header.

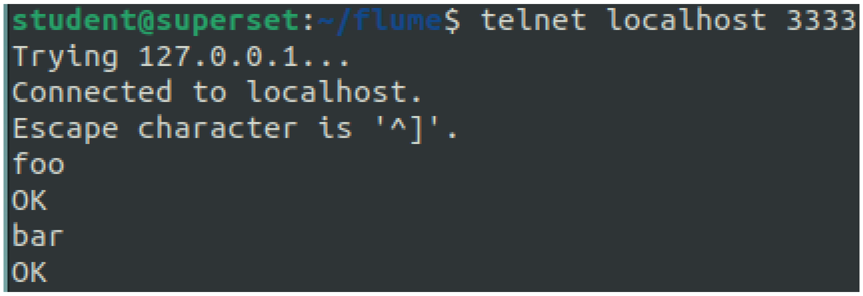
Ex.4 Using File roll to save events to files.

For this task I wrote the following configuration file:



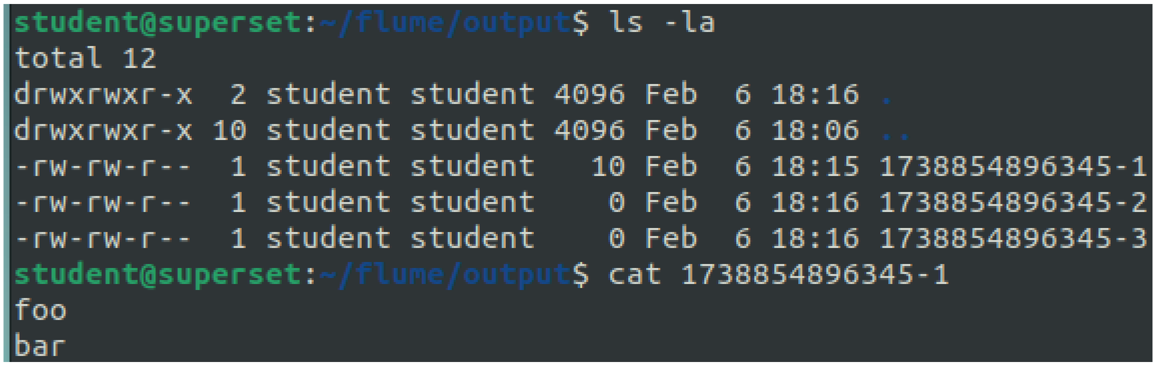
```
student@superset:~/flume$ cat ex4.conf
a4.sources = dirSrc
a4.channels = memChannel
a4.sinks = fileSink

a4.sources.dirSrc.channels = memChannel
a4.sources.dirSrc.type = netcat
a4.sources.dirSrc.bind = 127.0.0.1
a4.sources.dirSrc.port = 3333

a4.sinks.fileSink.channel = memChannel
a4.sinks.fileSink.type = file_roll
a4.sinks.fileSink.sink.directory = ./output
a4.sinks.fileSink.sink.rollInterval = 60

a4.channels.memChannel.type = memory
a4.channels.memChannel.capacity = 100
```

It should accumulate events in log files in `./output`. With the following input:



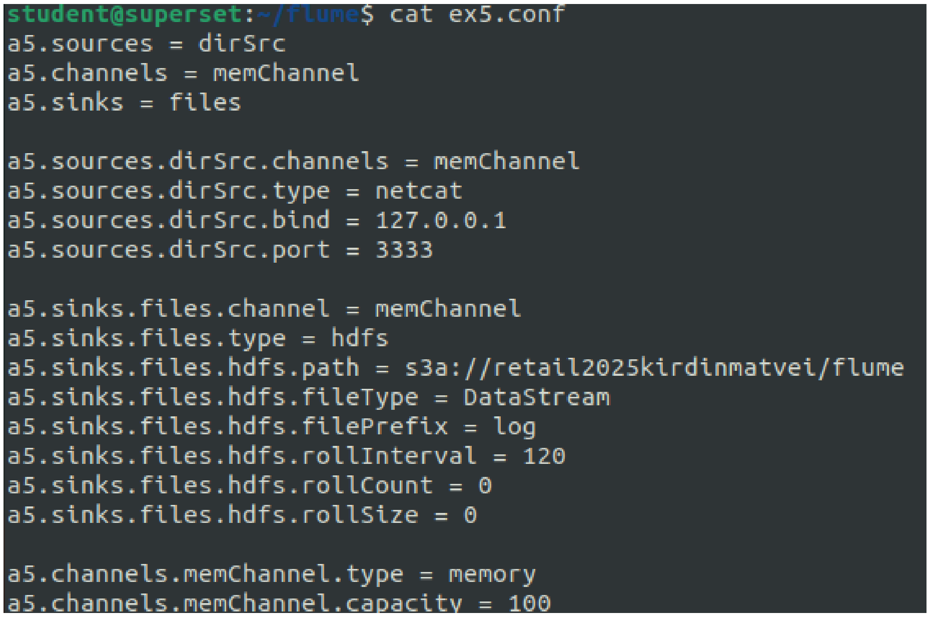
```
student@superset:~/flume$ telnet localhost 3333
Trying 127.0.0.1...
Connected to localhost.
Escape character is '^]'.
foo
OK
bar
OK
```

We get that

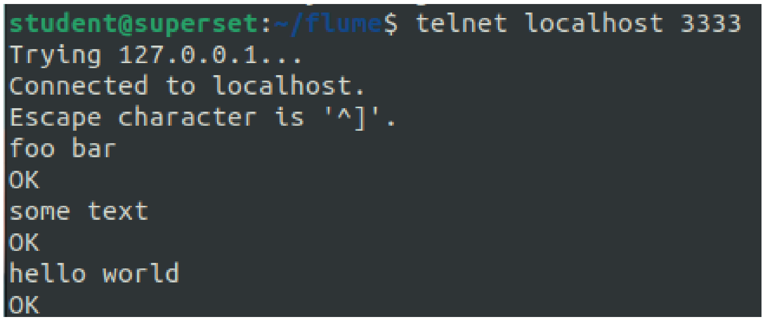


```
student@superset:~/flume/output$ ls -la
total 12
drwxrwxr-x  2 student student 4096 Feb  6 18:16 .
drwxrwxr-x 10 student student 4096 Feb  6 18:06 ..
-rw-rw-r--  1 student student   10 Feb  6 18:15 1738854896345-1
-rw-rw-r--  1 student student    0 Feb  6 18:16 1738854896345-2
-rw-rw-r--  1 student student    0 Feb  6 18:16 1738854896345-3
student@superset:~/flume/output$ cat 1738854896345-1
foo
bar
```

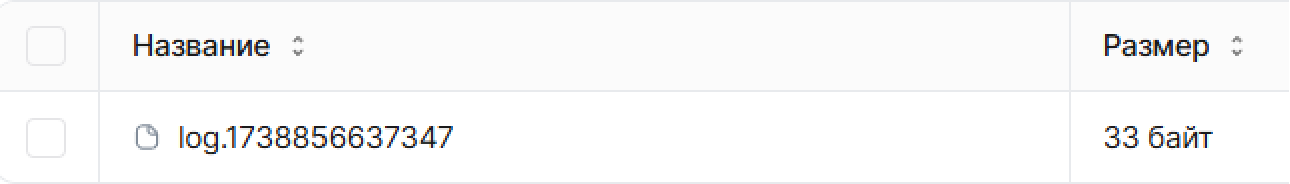
Ex.5 Uploading data to an S3 bucket.



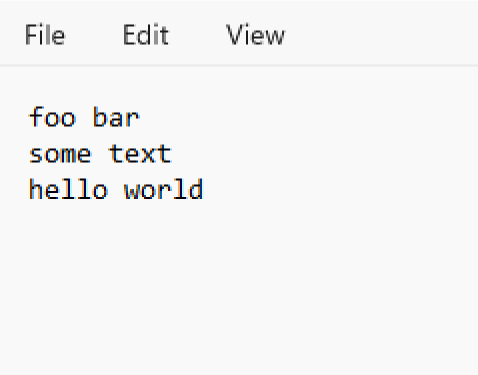
```
student@superset:~/flume$ cat ex5.conf
a5.sources = dirSrc
a5.channels = memChannel
a5.sinks = files

a5.sources.dirSrc.channels = memChannel
a5.sources.dirSrc.type = netcat
a5.sources.dirSrc.bind = 127.0.0.1
a5.sources.dirSrc.port = 3333

a5.sinks.files.channel = memChannel
a5.sinks.files.type = hdfs
a5.sinks.files.hdfs.path = s3a://retail2025kirdinmatvei/flume
a5.sinks.files.hdfs.fileType = DataStream
a5.sinks.files.hdfs.filePrefix = log
a5.sinks.files.hdfs.rollInterval = 120
a5.sinks.files.hdfs.rollCount = 0
a5.sinks.files.hdfs.rollSize = 0

a5.channels.memChannel.type = memory
a5.channels.memChannel.capacity = 100
```



```
student@superset:~/flume$ telnet localhost 3333
Trying 127.0.0.1...
Connected to localhost.
Escape character is '^]'.
foo bar
OK
some text
OK
hello world
OK
```



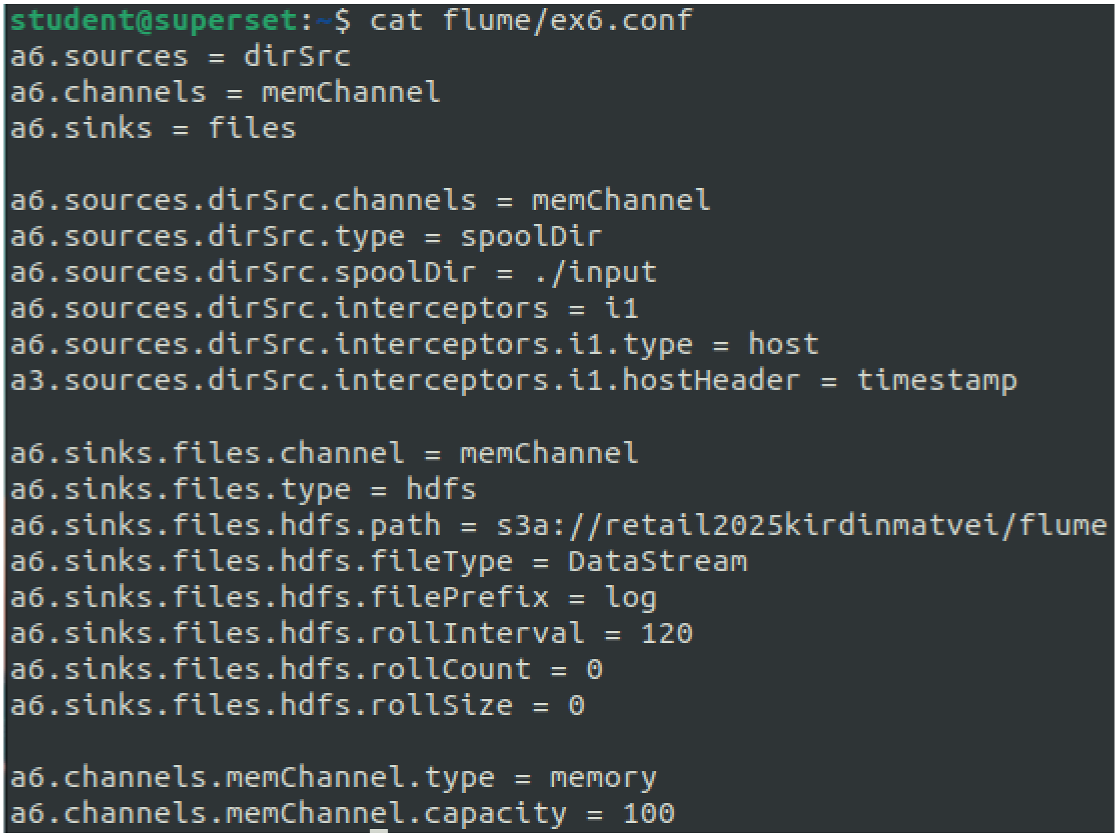
□	Название	Размер
□	log.1738856637347	33 байт



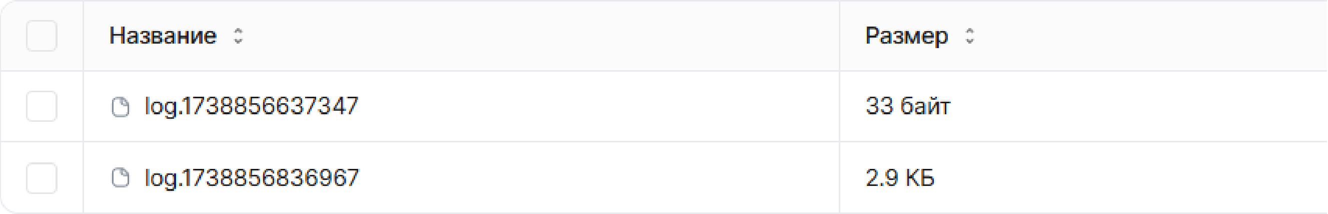
```
foo bar
some text
hello world
```

Ex. 6 Uploading data to an S3 bucket.

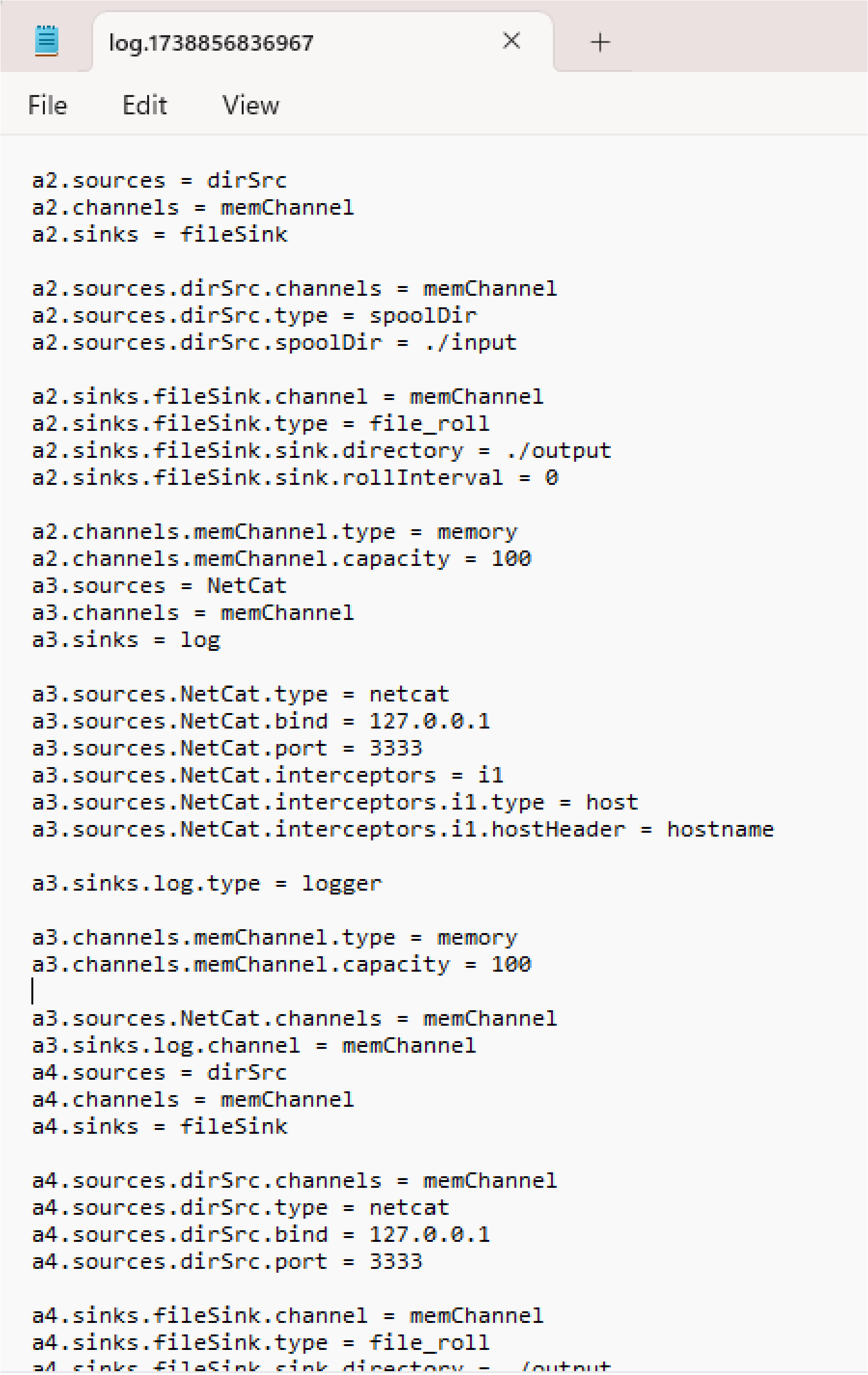
Here is the .conf file. I decided to copy conf files to S3. Let us see how it went:



```
student@superset:~$ cat flume/ex6.conf
a6.sources = dirSrc
a6.channels = memChannel
a6.sinks = files

a6.sources.dirSrc.channels = memChannel
a6.sources.dirSrc.type = spoolDir
a6.sources.dirSrc.spoolDir = ./input
a6.sources.dirSrc.interceptors = i1
a6.sources.dirSrc.interceptors.i1.type = host
a3.sources.dirSrc.interceptors.i1.hostHeader = timestamp

a6.sinks.files.channel = memChannel
a6.sinks.files.type = hdfs
a6.sinks.files.hdfs.path = s3a://retail2025kirdinmatvei/flume
a6.sinks.files.hdfs.fileType = DataStream
a6.sinks.files.hdfs.filePrefix = log
a6.sinks.files.hdfs.rollInterval = 120
a6.sinks.files.hdfs.rollCount = 0
a6.sinks.files.hdfs.rollSize = 0

a6.channels.memChannel.type = memory
a6.channels.memChannel.capacity = 100
```



☐	Название	Размер
☐	log.1738856637347	33 байт
☐	log.1738856836967	2.9 КБ



```
a2.sources = dirSrc
a2.channels = memChannel
a2.sinks = fileSink

a2.sources.dirSrc.channels = memChannel
a2.sources.dirSrc.type = spoolDir
a2.sources.dirSrc.spoolDir = ./input

a2.sinks.fileSink.channel = memChannel
a2.sinks.fileSink.type = file_roll
a2.sinks.fileSink.sink.directory = ./output
a2.sinks.fileSink.sink.rollInterval = 0

a2.channels.memChannel.type = memory
a2.channels.memChannel.capacity = 100
a3.sources = NetCat
a3.channels = memChannel
a3.sinks = log

a3.sources.NetCat.type = netcat
a3.sources.NetCat.bind = 127.0.0.1
a3.sources.NetCat.port = 3333
a3.sources.NetCat.interceptors = i1
a3.sources.NetCat.interceptors.i1.type = host
a3.sources.NetCat.interceptors.i1.hostHeader = hostname

a3.sinks.log.type = logger

a3.channels.memChannel.type = memory
a3.channels.memChannel.capacity = 100

a3.sources.NetCat.channels = memChannel
a3.sinks.log.channel = memChannel
a4.sources = dirSrc
a4.channels = memChannel
a4.sinks = fileSink

a4.sources.dirSrc.channels = memChannel
a4.sources.dirSrc.type = netcat
a4.sources.dirSrc.bind = 127.0.0.1
a4.sources.dirSrc.port = 3333

a4.sinks.fileSink.channel = memChannel
a4.sinks.fileSink.type = file_roll
a4.sinks.fileSink.sink.directory = ./output
```