FLUME PEC2 – Tania Batista

1. Crear un Pipeline donde la fuente sean los ficheros que sean colgados en un directorio del sistema local en la ruta /home/bigdata/flume/source (NOTA: Si la ruta no existe crearla), de tal forma que una vez que sea insertado un fichero en dicho directorio se desencadene el proceso (spooldir). Los datos ha de almacenarse en HDFS en la ruta /home/bigdata/pec2/ex1 y en formato de texto.

* Source directory: (files get stored here 🡪 a1.sources.r1.spoolDir):

**$ mkdir /home/bigdata/flume/source**

* HDFS route / directory for agent (where messages will be visible in console):

**$ mkdir /home/bigdata/pec2/ex1**

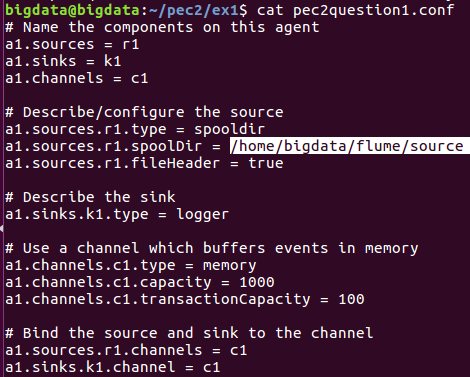
* Agent configuration:

**$ cd /home/bigdata/pec2/ex1/**

**$ touch pec2question1.conf**

**$ nano pec2question1.conf**

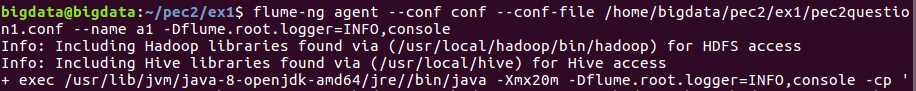
* + - Copy text from: example2.conf
    - a1.sources.r1.spoolDir = /home/bigdata/test



* Initiate the agent:
  + - Directory: /home/bigdata/pec2/ex1/pec2question1.conf

**$  flume-ng agent --conf conf --conf-file /home/bigdata/pec2/ex1/pec2question1.conf --name a1 -Dflume.root.logger=INFO,console**

* + SUCCESS!



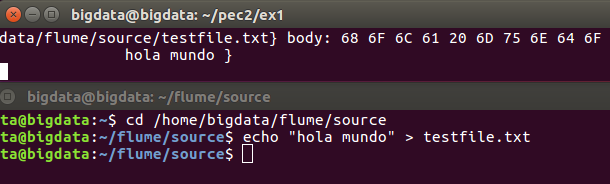
* Type messages

> new window

**$ cd /home/bigdata/flume/source**

**$ echo "hola mundo" > testfile.txt**

* + SUCCESS!



2. Crear un Pipeline donde la fuente sean los ficheros que sean colgados en un directorio del sistema local en la ruta /home/bigdata/flume/source, de tal forma que una vez que sea insertado un fichero en dicho directorio se desencadene el proceso (fuente spooldir). Los datos ha de almacenarse en HDFS en la ruta /home/bigdata/pec2/ex2 en formato de texto en una estructura de carpetas de acuerdo a la fecha y hora a la que han llegado por ejemplo /home/bigdata/pec2/ex2/18-04-15/1020 (significa que el fichero fue colocado en la carpeta el 15/04/2018 a las 10:20 aproximadamente).

* HDFS route / directory for agent (where messages will be visible in console):

**$ mkdir /home/bigdata/pec2/ex2**

* Agent configuration:

**$ cd /home/bigdata/pec2/ex2**

**$ touch /home/bigdata/pec2/ex2/pec2question2.conf**

**$ nano pec2question2.conf**

* + - * Copy text from: example4.conf

# Name the components on this agent

a1.sources = r1

a1.sinks = k1

a1.channels = c1

a1.sources.r1.interceptors = i1

a1.sources.r1.interceptors.i1.type = timestamp

# Describe/configure the source

a1.sources.r1.type = netcat

a1.sources.r1.bind = localhost

a1.sources.r1.port = 44444

# Describe the sink

a1.sinks.k1.type = hdfs

a1.sinks.k1.hdfs.path = /flume/events/%y-%m-%d/%H%M

a1.sinks.k1.hdfs.filePrefix = events-

a1.sinks.k1.hdfs.round = true

a1.sinks.k1.hdfs.roundValue = 10

a1.sinks.k1.hdfs.roundUnit = minute

# Use a channel which buffers events in memory

a1.channels.c1.type = memory

a1.channels.c1.capacity = 1000

a1.channels.c1.transactionCapacity = 100

# Bind the source and sink to the channel

a1.sources.r1.channels = c1

a1.sinks.k1.channel = c1

* Initiate agent
  + Directory 🡪/home/bigdata/pec2/ex2/pec2question2.conf

**$ flume-ng agent --conf conf --conf-file /home/bigdata/pec2/ex2/pec2question2.conf --name a1 -Dflume.root.logger=INFO,console**

* Start demons and connect to host

> new window

**$ start-dfs.sh**

**$ start-yarn.sh**

**$ jps**

**$ telnet localhost 44444**

* Send some test messages:

1, José Antonio, Ingeniero

2, María Antonia, Medico

3, María Fernanda, Ingeniero

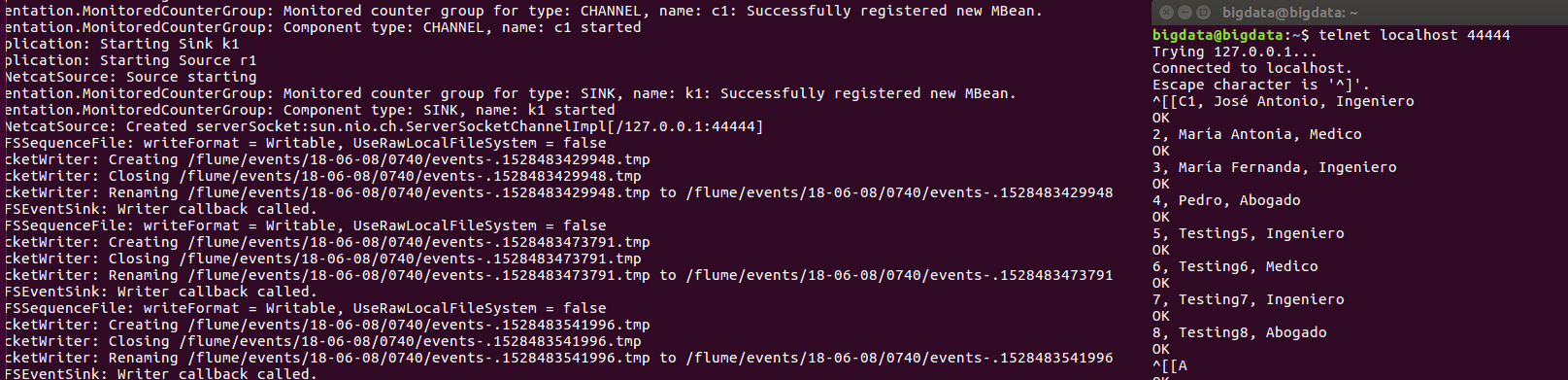
4, Pedro, Abogado

5, Testing5, Ingeniero

6, Testing6, Medico

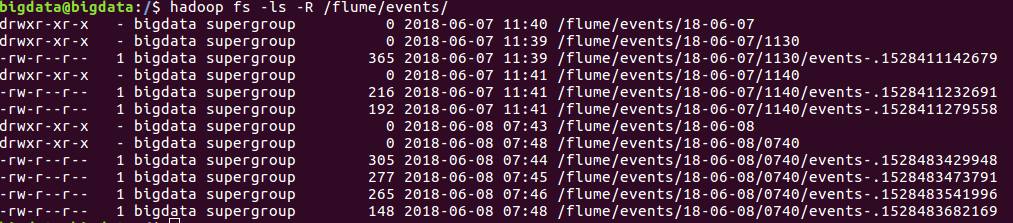
7, Testing7, Ingeniero

8, Testing8, Abogado



* HDFS File directory with timestamps:

**$ hadoop fs -ls -R /flume/events/**



**$ hadoop fs -ls -R /flume/events*/18-06-08/0740/events-.1528483429948***



**$ hadoop fs -cat /flume/events*/18-06-08/0740/events-.1528483429948***



3. Crear un pipeline donde la fuente sean los ficheros que sean colgados en un directorio del sistema local en la ruta /home/bigdata/flume/source, de tal forma que una vez que sea insertado un fichero en dicho directorio se desencadene el proceso (spooldir). Los ficheros que serán volcados al sistema local tendrán el siguiente formato:

1, José Antonio, Ingeniero

2, María Antonia, Medico

3, María Fernanda, Ingeniero

1, Pedro, Abogado

Los datos ha de almacenarse en HDFS en 3 rutas distintas de acuerdo al valor entero con el que inicia cada línea dentro de los ficheros. el cual no es mas que el código del departamento al que pertenece. Los datos ha de guardarse en HDFS en formato de texto.

* HDFS route / directory for agent (where messages will be visible in console):

**$ mkdir /user/bigdata/pec2/ex3/**

* Agent configuration:

**$ touch /home/bigdata/pec2/ex3/pec2question3.conf**

**$ nano pec2question3.conf**

* + - Copied from example4.conf and multiplexing.conf

# Describe source, channels and sinks

a1.sources = r1

a1.sinks = k1 k2 **k3**

a1.channels = c1 c2 **c3**

# 1 SOURCE SPOOL DIRECTORY

**a1.sources.r1.type = spooldir**

**a1.sources.r1.spoolDir =** /home/bigdata/flume/source

**a1.sources.r1.fileHeader = true**

# CHANNEL C1

a1.channels.c1.type = memory

a1.channels.c1.capacity = 100

**a1.channels.c1.transactionCapacity = 100**

# CHANNEL C2

a1.channels.c2.type = memory

a1.channels.c2.capacity = 100

**a1.channels.c2.transactionCapacity = 100**

**# CHANNEL C3 (newly added)**

a1.channels.c3.type = memory

a1.channels.c3.capacity = 100

**a1.channels.c3.transactionCapacity = 100**

# Describe regex\_extractor to extract different patterns (tipo1, tipo2, tipo3)

# Capture event attribute and add it as event header with attribute name as "department"

a1.sources.r1.interceptors = i1

a1.sources.r1.interceptors.i1.type = regex\_extractor

a1.sources.r1.interceptors.i1.regex = ^(\\d)

a1.sources.r1.interceptors.i1.serializers = t

a1.sources.r1.interceptors.i1.serializers.t.name = department

# HDFS sink k1

a1.sinks.k1.type = hdfs

a1.sinks.k1.channel = c1

**a1.sinks.k1.hdfs.path = user/bigdata/pec2/ex3/tipo1**

a1.sinks.k1.hdfs.fileType=DataStream

a1.sinks.k1.hdfs.writeFormat=Text

# HDFS sink k2

a1.sinks.k2.channel = c2

a1.sinks.k2.type = hdfs

**a1.sinks.k2.hdfs.path = user/bigdata/pec2/ex3/tipo2**

a1.sinks.k2.hdfs.fileType=DataStream

a1.sinks.k2.hdfs.writeFormat=Text

# HDFS sink k3

a1.sinks.k3.channel = c3

a1.sinks.k3.type = hdfs

**a1.sinks.k3.hdfs.path = user/bigdata/pec2/ex3/tipo3**

a1.sinks.k3.hdfs.fileType=DataStream

a1.sinks.k3.hdfs.writeFormat=Text

# Define channel selector and define mapping

a1.sources.r1.selector.type = multiplexing

a1.sources.r1.selector.header = department

a1.sources.r1.selector.mapping.1 = c1

a1.sources.r1.selector.mapping.2 = c2

a1.sources.r1.selector.mapping.3 = c3

# Bind the source and sink to the channel

a1.sources.r1.channels = c1 c2 **c3**

* Initiate agent
  + Directory 🡪/home/bigdata/pec2/ex3/pec2question3.conf

**$ flume-ng agent --conf conf --conf-file /home/bigdata/pec2/ex3/pec2question3.conf --name a1 -Dflume.root.logger=INFO,console**

* Source directory: (files get stored here 🡪 a1.sources.r1.spoolDir):

**$ cd /home/bigdata/flume/source**

**$ nano simpsonsDepartment.txt**

1, Maggie Simpson, Ingeniero

2, Lisa Simpson, Medico

3, Bart Simpson, Ingeniero

1, Homer Simpson, Abogado

2, Marge Simpson, Medico

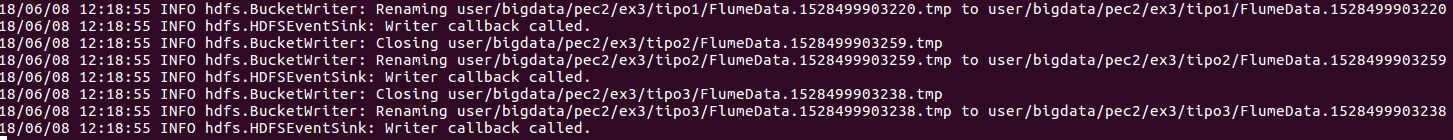
3, Ned Flanders, Ingeniero

* HDFS File directories

user/bigdata/pec2/ex3/tipo1/FlumeData.1528497594686

user/bigdata/pec2/ex3/tipo2/FlumeData.1528497627824

user/bigdata/pec2/ex3/tipo3/FlumeData.1528497492118

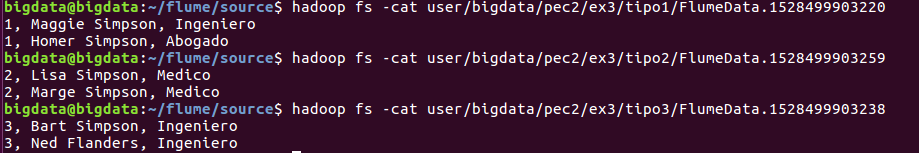


* Print HDFS file directories data: SUCCESS! Divided by type #!

**$ hadoop fs -cat user/bigdata/pec2/ex3/tipo1/FlumeData.1528499903220**

**$ hadoop fs -cat user/bigdata/pec2/ex3/tipo2/FlumeData.1528499903259**

**$ hadoop fs -cat user/bigdata/pec2/ex3/tipo3/FlumeData.1528499903238**



Estos comandos los tuve que encontrar en internet, porque los equivalentes de la aula virtual no funcionaban:

$  flume-ng agent --conf conf --conf-file ruta/fichero.conf --name a1 -Dflume.root.logger=INFO,console

$ hadoop fs -cat user/

$ hadoop fs -ls -R