Memoria de prácticas Flume.

Programa de experto en Big Data

Francisco L. López López

Curso 2013-2014

Tabla de contenido

[1. Planteamiento de la práctica. 3](#_Toc404114242)

[1.1. Primera parte. 3](#_Toc404114243)

[1.1.1. Código AGENTE 1. 4](#_Toc404114244)

[1.1.2. Código AGENTE 2. 5](#_Toc404114245)

[1.2. Segunda parte. 6](#_Toc404114246)

[1.2.1. Código agente. 7](#_Toc404114247)

[1.3. Tercera parte. 8](#_Toc404114248)

[1.3.1. Código agente colector. 9](#_Toc404114249)

[1.3.2. Código agente cliente1 en FR-CENTRAL1 10](#_Toc404114250)

[1.3.3. Código agente cliente2 en FR-CENTRAL2 10](#_Toc404114251)

# 1. Planteamiento de la práctica.

La solución a la práctica se puede resolver dividiendo en tres partes el planteamiento que se ha proporcionado en el texto.

1) Primera parte, definiendo el agente que comunicara entre los n webservices y un datacenter. Esto es un planteamiento que se tiene que extrapolar a los 50 datacenters.

2) Segunda parte, un agente entre un datacenter y las dos máquinas 'FR-CENTRAL1' y 'FR-CENTRAL2'. Esto es, que los 50 datacenters estarán conectados a las dos máquinas.

3) Tercera parte, un agente o conjunto de agentes entre las dos máquinas centrales y el sistema de almacenamiento HDFS.

## 1.1. Primera parte.

N webservices se conectan a un datanode.

DATACENTER

servidor web 1

servidor web 2

servidor web 3

servidor webn

En el DATACENTER existe un agente flume que recoge los eventos de los n servidores web

Agente en servidor web.

Agente 1

Agente en Data Center

Agente 2

....

Agente en servidor web.

Agente n

### 1.1.1. Código AGENTE 1.

Pasa del log generado por el webservice a un directorio interno, donde va grabando lo que va saliendo del log.

#Defino agente source channels sinks

agente.sources=s1

agente.channels=c1

agente.sinks=k1

#defino source

agente.sources.s1.type = syslogtcp

agente.sources.s1.channels=c1

# Un puerto diferente por cada DATANODE

agente.sources.s1.port=10000

agente.sources.s1.eventSize=300 #300 bytes por evento

agente.sources.s1.batchSize=60 # Número máximo de eventos (media 60)

#Defino channels

agente.channels.c1.type = memory

agente.channels.c1.capacity = 100

# Grabación del log

agente.sinks.k1.type=file\_roll

agente.sinks.k1.channel=c1

agente.sinks.k1.sink.directory=/var/log/operaciones.log

### 1.1.2. Código AGENTE 2.

Rota diariamente el diario de una carpeta interna a otra del DATACENTER.

#Defino agente source channels sinks

agente2.sources=s1

agente2.channels=c1

agente2.sinks=k1

#defino source

agente2.sources.s1.type = spooldir

agente2.sources.s1.channels=c1

agente2.sources.s1.spoolDir = /var/log

agente2.sources.s1.bufferMaxLines= 86400 #60 eventos /min\*1440 minutos dia = 86400 lineas

agente2.sources.s1.interceptors = i1

agente2.sources.s1.interceptors.i1.type = org.apache.flume.interceptor.TimestampInterceptor$Builder

#defino channel

agente2.channels.c1.type = memory

agente2.channels.c1.capacity = 10000

#defino sink

agente2.sinks.k1.type=file\_roll

agente2.sinks.k1.channel=c1

agente2.sinks.k1.sink.directory=/var/operaciones

agente2.sinks.k1.sink.rollInterval=86400 #Segundos que tiene 1 día

agente2.sinks.k1.sink.filePrefix=operaciones-%Y-%m-%d.log

## 1.2. Segunda parte.

Se vuelcan los datos de los 50 datacenters a las máquinas FR-CENTRAL1 y FR-CENTRAL2. Con sistema FAILOVER + regex\_filter para VENTAS y PEDIDOS.

DATACENTER1

FR-CENTRAL 1

DATACENTER 2

FR-CENTRAL 2

DATACENTER n

Descripción.

Por cada uno de los datacenters tendremos un agente con el código del apartado siguiente.

Se asume que cada DATACENTER tendrá:

* + 1. ip\_data\_center. IP asignada a ese datacenter

1.1 Filtraremos aquellos que tengan VENTA o PEDIDOS

* + 1. Puerto de entrada = 10000
    2. Tiene 2 sinks de salida, uno para frcentral1 y otro para frcentral2

Para el datacenter 1 tendra puerto 10001 para frcentral1 y 10051 para frcentral2

Para el datacenter 2 tendrá puerto 10002 para frcentral1 y 10052 para frcentral2

Para el datacenter 50 tendrá puerto 10050 para frcentral1 y 10100 para frcentral2

### 1.2.1. Código agente.

# Agente que guarda en salida los eventos tipo VENTA y PEDIDO

#Defino agente source channels sinks

agente.sources=s1

agente.channels=c1

agente.sinks= frcentra1\_sink frcentral2\_sink

agente.sources.s1.type = spooldir

agente.sources.s1.channels=c1

agente.sources.s1.spoolDir =/var/operaciones

agente.sources.s1.interceptors = i1

agente.sources.s1.interceptors.i1.type = regex\_filter

agente.sources.s1.interceptors.i1.regex = (\\d+)\\s(VENTA|PEDIDO)\\s(\\w+)

#channel

agente.channels.c1.type = memory

agente.channels.c1.capacity = 1000

# sinks

agente.sinks.frcentral1\_sink.type = avro

agente.sinks.frcentral1\_sink.hostname = ip\_frcentral1

agente.sinks.frcentral1\_sink.port = (10000 + n) // n es el valor del numero de datacenter

agente.sinks.frcentral1\_sink.channel = c1

agente.sinks.frcentral2\_sink.type = avro

agente.sinks.frcentral2\_sink.hostname = ip\_frcentral2

agente.sinks.frcentral2\_sink.port = (10050 + n) // n es el valor del numero de datacenter

agente.sinks.frcentral2\_sink.channel = c1

# sink groups

agente.sinkgroups = failover\_group

agente.sinkgroups.failover\_group.sinks = frcentral1 frcentral2

agente.sinkgroups.failover\_group.processor.type = failover

agente.sinkgroups.failover\_group.processor.priority.frcentral1 = 10

agente.sinkgroups.failover\_group.processor.priority.frcentral2 = 20

## 1.3. Tercera parte.

Descripción.

Tenemos 2 máquinas FR-CENTRAL 1 y 2 que para llegar al HDFS usamos un recolector usando un par de avro-source/avro sink.

Por lo tanto, declaramos 3 agentes:

* + 1. Agente de FR-CENTRAL 1. Cliente 1.
    2. Agente de FR-CENTRAL 2 .Cliente 2.
    3. Agente colector. Con un reg\_extractor, para clasificar por VENTA o PEDIDO y escribir en el directorio TIPO PEDIDO/FECHA con destino URI: hdfs://fr-hdfs:9000

Con tiempo de keep-alive = 3 horas.

Con channel en formato fichero en el directorio /.flume/file-channel/data

para que se guarde como máximo ese tiempo.



### 1.3.1. Código agente colector.

#Definicion del colector

colector.sources=s1

colector.channels=c1

colector.sinks=k1

#Sources

colector.sources.s1.type=avro

colector.sources.s1.bind=10.240.240.90

colector.sources.s1.port=12345

colector.sources.s1.channels=c1

colector.sources.s1.interceptors = i1 i2

colector.sources.s1.interceptors.i1.type = regex\_extractor

colector.sources.s1.interceptors.i1.regex = (\\d+) (\\w\*) (\\w+)

colector.sources.s1.interceptors.i1.serializers = t1 t2 t3

colector.sources.s1.interceptors.i1.serializers.t1.name = numero

colector.sources.s1.interceptors.i1.serializers.t2.name = tipo

colector.sources.s1.interceptors.i1.serializers.t3.name = resto

colector.sources.s1.interceptors.i2.type = org.apache.flume.interceptor.TimestampInterceptor$Builder

#Channels con 3 horas de backup para poder esperar a grabar

colector.channels.c1.type=file

colector.channels.c1.keep-alive = 10800

colector.channels.c1.dataDirs=/.flume/file-channel/data

#Sinks

#Solo interesa el valor tipo.

colector.sinks.k1.type=hdfs

colector.sinks.k1.channel=c1

colector.sinks.k1.hdfs.path =hdfs://fr-hdfs:9000/%{tipo}/%Y%m%d

colector.sinks.k1.hdfs.filePrefix=%{tipo}

colector.sinks.k1.hdfs.fileType=DataStream

colector.sinks.k1.hdfs.writeFormat=Text

### 1.3.2. Código agente cliente1 en FR-CENTRAL1

#Definicion del cliente1

cliente1.sources=s1

cliente1.channels=c1

cliente1.sinks=k1

#Sources

cliente1.sources.s1.type=avro

cliente1.sources.s1.bind= ip\_frcentral1

cliente1.sources.s1.port=12345

#Channels

cliente1.channels.c1.type=memory

#Sinks

cliente1.sinks.k1.type=avro

cliente1.sinks.k1.channel=c1

cliente1.sinks.k1.hostname=10.240.240.90

cliente1.sinks.k1.port=12345

### 1.3.3. Código agente cliente2 en FR-CENTRAL2

#Definicion del cliente2

cliente2.sources=s1

cliente2.channels=c1

cliente2.sinks=k1

#Sources

cliente2.sources.s1.type=avro

cliente2.sources.s1.bind= ip\_frcentral2

cliente2.sources.s1.port=12345

#Channels

cliente2.channels.c1.type=memory

#Sinks

cliente2.sinks.k1.type=avro

cliente2.sinks.k1.channel=c1

cliente2.sinks.k1.hostname=10.240.240.90

cliente2.sinks.k1.port=12345