Tarea 1 UD 3

Juan Manuel García Moyano
IABD
Informática y comunicaciones

Índice

Caso práctico 1

La empresa Jardines Online S.L. (para la cual trabajas) se dedica a la venta online de productos de jardín a través de su tienda online.

Sus directivos llevan varios meses evaluando la posibilidad de comenzar a emplear la gran cantidad de datos de los que se dispone con la intención de obtener valor de ellos. En concreto quieren analizar los datos transaccionales producidos por el departamento de ventas.

Su intención es realizar la analítica básica con personal propio y subcontratar la generación la generación de elaboración de perfiles y modelos predictivos a una empresa especializada.

Los datos de los que quieren obtener valor son los siguientes:

- ✔ Las transacciones realizadas por el departamento de ventas (las ventas online de productos).
- ✔ Información relativa a la cuenta que cada cliente ha creado en la tienda online, en la cual se incluyen datos personales.

Gracias a que cada cliente necesita iniciar sesión con su cuenta para poder hacer compras, la empresa sabe qué ha comprado cada uno. Por lo tanto, a partir de la información que tiene el departamento de ventas (las transacciones de ventas online), la empresa desea ser capaz de realizar un perfilado de sus clientes para ser capaz de mostrarles ofertas personalizadas en función de sus gastos.

La reunión más interesante de los últimos años acaba de terminar y te informan de que se te ha encomendado la misión de generar un documento de análisis de situación que servirá para comenzar con el proceso.

En ese documento te piden que hagas determinadas indicaciones acerca de lo que habrá que hacer para estar conformes al RGPD y de actividades de Gobierno de Datos para asegurar que la gestión que se haga de los datos sea correcta.

¿Qué te pedimos que hagas?

Utiliza la información que encontrarás en los contenidos de esta unidad para crear un documento PDF en el que irás resolviendo los siguientes apartados.

Apartado 1

En lo referente a la conformidad con el RGPD:

1. Indica quién es el interesado.→ Los clientes de Jardines Online S.L., cuyos datos personales y transaccionales serán procesados para la elaboración de perfiles y modelos predictivos.

2. Indica quién es el responsable del tratamiento. → Jardines Online S.L., como entidad que decide los fines y medios del tratamiento de los datos personales.

3. Indica quién es el encargado del tratamiento. → La empresa especializada subcontratada para la generación de perfiles y modelos predictivos.

4. Indica quién es el destinatario. → Los datos pueden ser accedidos por el departamento de ventas y marketing de Jardines Online S.L., con el objetivo de generar ofertas personalizadas.

5. Indica quién es el tercero. → La empresa especializada subcontratada que procesará los datos para la elaboración de perfiles y modelos predictivos.

6. Indica si será necesario obtener consentimiento de los interesados. → Sí, se debe obtener el consentimiento explícito de los clientes para el tratamiento de sus datos con fines de perfilado y marketing personalizado.

7. Indica si será necesario realizar seudonimización de los datos. →Sí, es recomendable para reducir riesgos y cumplir con el principio de minimización de datos.

8. Indica si los usuarios tendrán derecho de acceso sobre sus datos. → Sí, los clientes tienen derecho a acceder a sus datos personales, solicitar su modificación o eliminación, según lo estipulado en el RGPD.

9. Indica si habrá que hacer algún tipo de comprobación respecto de la empresa subcontratada. → Sí, se debe verificar que la empresa cumple con el RGPD y tiene medidas de seguridad adecuadas para el tratamiento de los datos.

10. Indica si será necesario firmar algún tipo de contrato escrito con la empresa subcontratada. → Sí, se debe formalizar un contrato de encargo de tratamiento donde se estipulen las responsabilidades y medidas de protección de datos.

Apartado 2

En lo referente a actividades de Gobierno de Datos para asegurar que la gestión de los datos sea la correcta:

1. Indica si los datos podrán ser considerados un activo de la empresa.

Sí, los datos transaccionales y de clientes representan un activo estratégico para la empresa y deben ser gestionados adecuadamente.

2. Indica si habrá que realizar una gestión general de riesgos respecto a posibles incidentes.

Sí, es necesario implementar una gestión de riesgos para mitigar posibles incidentes como accesos no autorizados, fugas de datos o usos indebidos.

3. Indica si se considera necesario crear una guía de calidad de los datos.

Sí, se recomienda la elaboración de una guía de calidad de datos que establezca estándares para la integridad, exactitud y disponibilidad de la información.

4. Indica si debería de haber un *Chief Data Officer* (CDO) y cuales serían sus atribuciones.

   Sí, se recomienda designar un CDO cuya función sea supervisar la estrategia de datos, garantizar su seguridad y cumplimiento normativo, y optimizar su uso en la empresa.

5. Indica qué persona sería el *Data Owner* (propietario del dato).

   El departamento de ventas y marketing de Jardines Online S.L., como responsables de la recolección y uso de los datos para la generación de ofertas personalizad

Caso práctico 2.

Ana, en su tiempo libre, está haciendo un MOOC sobre la plataforma de Hadoop. Aunque está trabajando gracias a su titulación de Técnico Superior en Desarrollo de Aplicaciones Multiplataforma, aspira a cambiar de trabajo y está mirando ofertas como esta Oferta Big Data Engineer -Technical Support – Google Cloud in Barcelona.

En el módulo de hoy le han explicado someramente que Apache Flume es un software distribuido para recopilar, agregar y mover, de manera eficiente, grandes cantidades de datos de muchas fuentes diferentes a un almacén de datos centralizado. La verdad es que ha entendido el texto pero cree que se aclararía más si pudiera probarlo. Ha preguntado en el foro del MOOC y una compañera le ha sugerido que cree un cuaderno en Colab y se dispone a hacerlo.

¿Qué te pedimos que hagas?

Práctica 1

En esta práctica tienes que instalar Hadoop y Flume en una máquina virtual, y configurarás el entorno para que ambas herramientas estén operativas.

1. Máquina virtual instalada con Hadoop (como vimos en la UD2). Puedes apoyarte en videotutoriales en Internet o con este enlace (ten en cuenta las versiones de Hadoop y el sistema operativo para evitar errores en la instalción).
2. Instalación de Flume: descarga e instala el binario de Flume desde aquí. Para ayudarte con Flume utiliza los enlaces a Flume que hemos subido en la plataforma.
3. Realiza los siguientes ejercicios:
    a) Crear un agente Flume, cuyo source será de tipo spooldir, comprobará un determinado directorio, donde debería encontrar un fichero .csv y procesarlo para colocar su contenido en un directorio final de HDFS. Esta acción podrá generar en el destino uno o varios ficheros nuevos. A continuación ejecutarás el agente y comprobarás que su funcionamiento es

correcto. Para el desarrollo de esta práctica necesitarás un fichero csv con un dataset. Puedes descargar alguno que te interese desde la página web: https://www.kaggle.com/datasets.

b) Crear un agente Flume, cuyo source será de tipo NetCat donde se especificará un puerto de escucha. Mediante esa fuente Flume se quedará en escucha en dicho puerto. El canal será de tipo fichero para que el evento sea perdurable y los datos se ingesten en un directorio final de HDFS. Para su comprobación podemos hacer uso en otra terminar (mientras se lanza el agente Flume) de "curl telnet://localhost:<puerto>", introduciremos datos y podremos ver la generación de archivos con esa información en la interfaz gráfica de HDFS.

a) La máquina virtual ya tiene instalado Hadoop. Compruebo la versión.

```
hadoop@dhcp2:~$ hadoop version
Hadoop 3.4.1
Source code repository https://github.com/apache/hadoop.git -r 4d782530934895633
6b8f06a08322b78422849b1
Compiled by mthakur on 2024-10-09T14:57Z
Compiled on platform linux-x86_64
Compiled with protoc 3.23.4
From source with checksum 7292fe9dba5e2e44e3a9f763fce3e680
This command was run using /usr/local/hadoop/share/hadoop/common/hadoop-common-3
.4.1.jar
```

b) Descargo Flume.

```
hadoop@dhcp2:~$ wget https://downloads.apache.org/flume/1.11.0/apache-flume-1.11
.0-bin.tar.gz
--2025-02-04 17:26:31--  https://downloads.apache.org/flume/1.11.0/apache-flume-
1.11.0-bin.tar.gz
Resolviendo downloads.apache.org (downloads.apache.org)... 88.99.208.237, 135.18
1.214.104, 2a01:4f8:10a:39da::2, ...
Conectando con downloads.apache.org (downloads.apache.org)[88.99.208.237]:443...
 conectado.
Petición HTTP enviada, esperando respuesta... 200 OK
Longitud: 87380462 (83M) [application/x-gzip]
Guardando como: “apache-flume-1.11.0-bin.tar.gz”

apache-flume-1.11.0 100%[===================>]  83,33M  6,28MB/s    en 15s

2025-02-04 17:26:50 (5,61 MB/s) - “apache-flume-1.11.0-bin.tar.gz” guardado [873
80462/87380462]
```

c) Extraigo los datos del .zip.

```
hadoop@dhcp2:~$ tar -xvzf apache-flume-1.11.0-bin.tar.gz
apache-flume-1.11.0-bin/LICENSE
apache-flume-1.11.0-bin/NOTICE
apache-flume-1.11.0-bin/bin/
apache-flume-1.11.0-bin/conf/
apache-flume-1.11.0-bin/DEVNOTES
apache-flume-1.11.0-bin/bin/flume-ng.cmd
```

d) Muevo Flume al directorio /usr/local/flume

```
hadoop@dhcp2:~$ sudo mv apache-flume-1.11.0-bin /usr/local/flume
[sudo] contraseña para hadoop:
hadoop@dhcp2:~$
```

e) Modifico el fichero ~/.bashrc

```
hadoop@dhcp2:~$ nano ~/.bashrc
hadoop@dhcp2:~$
```

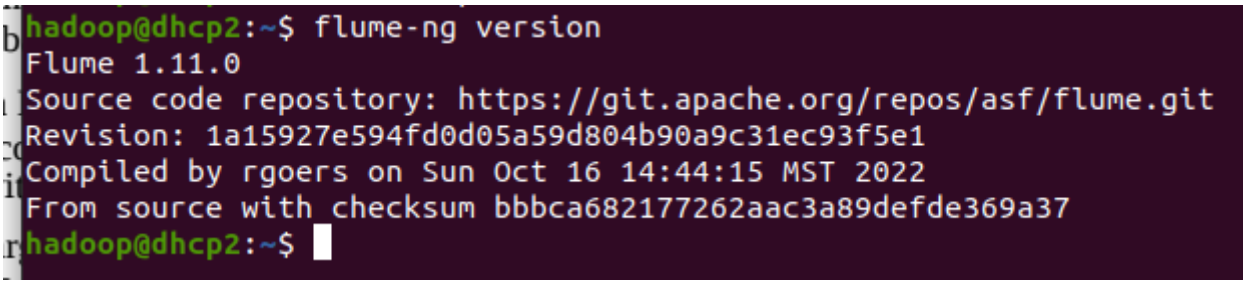
f) Al final del fichero exporto las siguientes rutas:

```
 GNU nano 4.8                    /home/hadoop/.bashrc

# Hadoop environment variables
export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64
export HADOOP_HOME=/usr/local/hadoop
export HADOOP_INSTALL=$HADOOP_HOME
export HADOOP_MAPRED_HOME=$HADOOP_HOME
export HADOOP_COMMON_HOME=$HADOOP_HOME
export HADOOP_HDFS_HOME=$HADOOP_HOME
export HADOOP_YARN_HOME=$HADOOP_HOME
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
export PATH=$PATH:$HADOOP_HOME/sbin:$HADOOP_HOME/bin
export HADOOP_OPTS="-Djava.library.path=$HADOOP_HOME/lib/native"

export FLUME_HOME=/usr/local/flume
export PATH=$PATH:$FLUME_HOME/bin
```

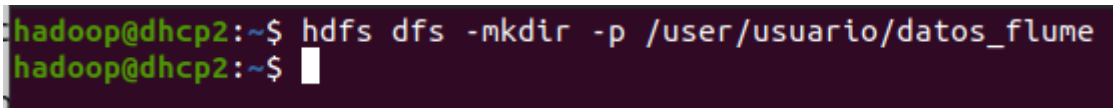
g) Ejecuto el fichero que acabo de modificar para que pille los cambios.

```
hadoop@dhcp2:~$ source ~/.bashrc
hadoop@dhcp2:~$
```

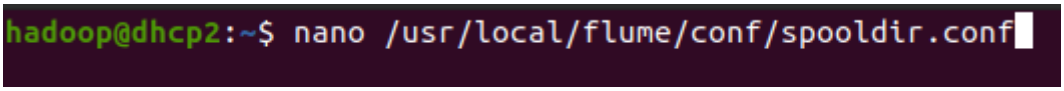
h) Compruebo la versión de Flume para ver si se ha instalado correctamente



i) Creo el directorio con hdfs /user/usuario/datos_flume



j) Modifico el fichero o creo spooldir.conf



k) Pongo las siguientes filas

```
  GNU nano 4.8          /usr/local/flume/conf/spooldir.conf          Modificado
agent.sources = spool-source
agent.channels = memory-channel
agent.sinks = hdfs-sink

agent.sources.spool-source.type = spooldir
agent.sources.spool-source.spoolDir = /home/hadoop/flume/spooldir
agent.sources.spool-source.fileHeader = true

agent.channels.memory-channel.type = memory
agent.channels.memory-channel.capacity = 1000

agent.sinks.hdfs-sink.type = hdfs
agent.sinks.hdfs-sink.hdfs.path = hdfs://localhost:9000/user/hadoop/datos_flume
agent.sinks.hdfs-sink.hdfs.fileType = DataStream
agent.sinks.hdfs-sink.hdfs.writeFormat = Text
agent.sinks.hdfs-sink.hdfs.batchSize = 1000

agent.sources.spool-source.channels = memory-channel
agent.sinks.hdfs-sink.channel = memory-channel
```

l) Me descargo un dataset csv del siguiente enlace https://www.kaggle.com/datasets/arshmankhalid/shopify-streaming-history-dataset

m) Muevo el fichero spotify_history.csv

```
hadoop@dhcp2:~$ mv ~/Descargas/archive/spotify_history.csv /home/hadoop/flume/sp
ooldir/
hadoop@dhcp2:~$
```

n) Inicio el agente de Flume

```
hadoop@dhcp2:~$ flume-ng agent --name agent --conf /usr/local/flume/conf --conf-
file /usr/local/flume/conf/spooldir.conf -Dflume.root.logger=INFO,console
Info: Including Hadoop libraries found via (/usr/local/hadoop/bin/hadoop) for HD
FS access
Info: Including Hive libraries found via () for Hive access
+ exec /usr/lib/jvm/java-11-openjdk-amd64/bin/java -Xmx20m -Dflume.root.logger=I
NFO,console -cp '/usr/local/flume/conf:/usr/local/flume/lib/*:/usr/local/hadoop/
etc/hadoop:/usr/local/hadoop/share/hadoop/common/lib/*:/usr/local/hadoop/share/h
adoop/common/*:/usr/local/hadoop/share/hadoop/hdfs:/usr/local/hadoop/share/hadoo
p/hdfs/lib/*:/usr/local/hadoop/share/hadoop/hdfs/*:/usr/local/hadoop/share/hadoo
p/mapreduce/*:/usr/local/hadoop/share/hadoop/yarn:/usr/local/hadoop/share/hadoop
/yarn/lib/*:/usr/local/hadoop/share/hadoop/yarn/*:/lib/*' -Djava.library.path=:/
usr/local/hadoop/lib/native org.apache.flume.node.Application --name agent --con
f-file /usr/local/flume/conf/spooldir.conf

SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/flume/lib/log4j-slf4j-impl-2.18.0.j
ar!/org/slf4j/impl/StaticLoggerBinder.class]
```

o) Continuamos con netcat. Creamos un fichero de configuración

```
hadoop@dhcp2:~$ nano /usr/local/flume/conf/netcat.conf
```

p) Al fichero anterior le añadimos las siguientes filas:

```
  GNU nano 4.8          /usr/local/flume/conf/netcat.conf          Modificado
agent.sources = netcat-source
agent.channels = file-channel
agent.sinks = hdfs-sink

agent.sources.netcat-source.type = netcat
agent.sources.netcat-source.bind = localhost
agent.sources.netcat-source.port = 44444

agent.channels.file-channel.type = file
agent.channels.file-channel.checkpointDir = /tmp/flume-checkpoint
agent.channels.file-channel.dataDirs = /tmp/flume-data

agent.sinks.hdfs-sink.type = hdfs
agent.sinks.hdfs-sink.hdfs.path = hdfs://localhost:9000/user/usuario/netcat_data
agent.sinks.hdfs-sink.hdfs.fileType = DataStream
agent.sinks.hdfs-sink.hdfs.batchSize = 1000

agent.sources.netcat-source.channels = file-channel
agent.sinks.hdfs-sink.channel = file-channel
```

q) Ejecutamos el agente

```
hadoop@dhcp2:~$ flume-ng agent --name agent --conf /usr/local/flume/conf --conf-
file /usr/local/flume/conf/spooldir.conf -Dflume.root.logger=INFO,console
Info: Including Hadoop libraries found via (/usr/local/hadoop/bin/hadoop) for HD
FS access
Info: Including Hive libraries found via () for Hive access
+ exec /usr/lib/jvm/java-11-openjdk-amd64/bin/java -Xmx20m -Dflume.root.logger=I
NFO,console -cp '/usr/local/flume/conf:/usr/local/flume/lib/*:/usr/local/hadoop/
etc/hadoop:/usr/local/hadoop/share/hadoop/common/lib/*:/usr/local/hadoop/share/h
adoop/common/*:/usr/local/hadoop/share/hadoop/hdfs:/usr/local/hadoop/share/hadoo
p/hdfs/lib/*:/usr/local/hadoop/share/hadoop/hdfs/*:/usr/local/hadoop/share/hadoo
p/mapreduce/*:/usr/local/hadoop/share/hadoop/yarn:/usr/local/hadoop/share/hadoop
/yarn/lib/*:/usr/local/hadoop/share/hadoop/yarn/*:/lib/*' -Djava.library.path=:/
usr/local/hadoop/lib/native org.apache.flume.node.Application --name agent --con
f-file /usr/local/flume/conf/spooldir.conf

SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/flume/lib/log4j-slf4j-impl-2.18.0.j
ar!/org/slf4j/impl/StaticLoggerBinder.class]
```