Caso de Uso: Ingesta de Archivos de Texto con Flume

Objetivo:

Automatizar la transferencia de archivos de texto desde un directorio local al sistema HDFS utilizando Flume y así analizar los datos con MapReduce (por ejemplo, el `wordcount` del ejercicio anterior).

---

1. Configuración del Entorno de Flume

1. **Verifica que Apache Flume esté instalado y configurado.** En Cloudera, puedes instalar Flume desde Cloudera Manager si aún no está disponible.

Inicia el Servicio de Flume:

sudo service flume-ng start

---

2. Crear una Configuración para Flume

Archivo de Configuración (`agenteflumeexec.conf`):

```
# Nombre del agente
agent.sources = exec-source
agent.sinks = hdfs-sink
agent.channels = memory-channel

agent.sources.exec-source.type = exec
agent.sources.exec-source.command = curl -o /tmp/quijote.txt https://babel.upm.es/~angel/teaching/pps/quijote.txt
agent.sources.exec-source.command = cat /tmp/quijote.txt
agent.sources.exec-source.channels = memory-channel

# Configuración del canal (MemoryChannel)
agent.channels.memory-channel.type = memory
agent.channels.memory-channel.capacity = 1000
agent.channels.memory-channel.transactionCapacity = 100

# Configuración del sink (HDFSSink) agent.sinks.hdfs-sink.type = hdfs
agent.sinks.hdfs-sink.hdfs.path = /user/cloudera/flume-output
agent.sinks.hdfs-sink.hdfs.fileType = DataStream
agent.sinks.hdfs-sink.hdfs.writeFormat = Text
agent.sinks.hdfs-sink.hdfs.rollSize = 0
```

agent.sinks.hdfs-sink.hdfs.rollCount = 10
agent.sinks.hdfs-sink.hdfs.batchSize = 100
agent.sinks.hdfs-sink.channel = memory-channel

---

3. Preparar el Directorio Local de Entrada

Crea el Directorio de Entrada Local:

```
mkdir -p tmp
```

---

4. Ejecutar Apache Flume

Inicia el Agente Flume: Ejecuta Flume con el archivo de configuración:

```
flume-ng agent --conf /path/to/flume/conf --conf-file /path/to/agenteflumeexec.conf --name agent -Dflume.root.logger=INFO,console
```

Verifica que los Datos Llegaron a HDFS: Lista los archivos en la ruta configurada en el sink (`/user/cloudera/flume-output`):

```
hdfs dfs -ls /user/cloudera/flume-output
```

Muestra el Contenido Ingerido:

```
hdfs dfs -cat /user/cloudera/flume-output/*
```

Entrega un documento con las capturas de pantalla de todo el proceso.