1. Implement flume script for spooling a directory from local file system to HDFS.

To implement a Flume script for spooling a directory from the local file system to HDFS, follow these steps:

**Step 1: Install and Configure Apache Flume**

1. **Download Apache Flume**:
   * Download the Flume binary from the [Apache Flume downloads page](https://flume.apache.org/download.html).
   * Extract the downloaded archive.
2. **Set Up Environment Variables**:
   * Add the Flume bin directory to your PATH.
   * Example (in ~/.bashrc or ~/.bash\_profile

export FLUME\_HOME=/path/to/flume

export PATH=$PATH:$FLUME\_HOME/bin

1. **Create Directories**:
   * Create a directory for the files to be spooled, e.g., /path/to/spooldir.
   * Ensure the HDFS directory you want to write to exists, e.g., /user/flume/spooldir.

**Step 2: Create the Flume Configuration File**

Create a Flume configuration file, e.g., flume-conf.properties, with the following content:

# Name the components on this agent

agent1.sources = spool-source

agent1.sinks = hdfs-sink

agent1.channels = memory-channel

# Describe/configure the source

agent1.sources.spool-source.type = spooldir

agent1.sources.spool-source.spoolDir = /path/to/spooldir

agent1.sources.spool-source.fileHeader = true

# Describe the sink

agent1.sinks.hdfs-sink.type = hdfs

agent1.sinks.hdfs-sink.hdfs.path = hdfs://namenode:8020/user/flume/spooldir

agent1.sinks.hdfs-sink.hdfs.fileType = DataStream

agent1.sinks.hdfs-sink.hdfs.writeFormat = Text

agent1.sinks.hdfs-sink.hdfs.batchSize = 1000

agent1.sinks.hdfs-sink.hdfs.rollSize = 0

agent1.sinks.hdfs-sink.hdfs.rollCount = 10000

agent1.sinks.hdfs-sink.hdfs.rollInterval = 600

# Use a memory channel

agent1.channels.memory-channel.type = memory

agent1.channels.memory-channel.capacity = 10000

agent1.channels.memory-channel.transactionCapacity = 1000

# Bind the source and sink to the channel

agent1.sources.spool-source.channels = memory-channel

agent1.sinks.hdfs-sink.channel = memory-channel

### Step 3: Start the Flume Agent

Run the Flume agent with the configuration file:

flume-ng agent --conf $FLUME\_HOME/conf --conf-file /path/to/flume-conf.properties --name agent1 -Dflume.root.logger=INFO,console

Replace /path/to/flume-conf.properties with the path to your configuration file.

**Step 4: Verify Data in HDFS**

1. **Check HDFS Directory**: Ensure the data is being written to the HDFS directory specified in the configuration file:

hdfs dfs -ls /user/flume/spooldir

**Explanation of Configuration**

* **Source**:
  + spool-source.type = spooldir: Specifies that the source is a spooling directory.
  + spool-source.spoolDir = /path/to/spooldir: The local directory to be monitored for new files.
  + spool-source.fileHeader = true: Adds a file header to each event indicating the source file.
* **Sink**:
  + hdfs-sink.type = hdfs: Specifies that the sink is HDFS.
  + hdfs-sink.hdfs.path = hdfs://namenode:8020/user/flume/spooldir: The HDFS directory where data will be written.
  + hdfs-sink.hdfs.fileType = DataStream: Uses a DataStream to write data to HDFS.
  + hdfs-sink.hdfs.writeFormat = Text: Writes data in text format.
  + hdfs-sink.hdfs.batchSize = 1000: Number of events written to HDFS per transaction.
  + hdfs-sink.hdfs.rollSize = 0, hdfs-sink.hdfs.rollCount = 10000, hdfs-sink.hdfs.rollInterval = 600: Conditions for rolling the HDFS file.
* **Channel**:
  + memory-channel.type = memory: Specifies a memory channel.
  + memory-channel.capacity = 10000: Maximum number of events the channel can hold.
  + memory-channel.transactionCapacity = 1000: Maximum number of events per transaction.
* **Binding**:
  + agent1.sources.spool-source.channels = memory-channel: Binds the source to the channel.
  + agent1.sinks.hdfs-sink.channel = memory-channel: Binds the sink to the channel.

This configuration sets up a Flume agent to monitor a local directory for new files and write the contents of these files to HDFS. The memory channel serves as an intermediary buffer. Adjust the paths and parameters as needed for your environment.