1. 准备工作

确保以下条件已满足：

- **安装和配置 Flume**：确保 Flume 已经在你的系统中安装和配置好。
- **Hadoop 和 HDFS 配置**：确保 Hadoop 和 HDFS 在你的环境中运行，并且有足够的权限将数据写入 HDFS。
- **目标 HDFS 目录**：确定将数据写入 HDFS 的目标路径。

2. 编写 Flume 配置文件

Flume 使用配置文件来定义数据流和传输的管道。以下是一个基本的 Flume 配置文件示例，用于监控本地文件并将其内容传输到 HDFS：

```
# flume.conf

# 定义 agent 名和其使用的组件
agent.sources = local_source
agent.sinks = hdfs_sink
agent.channels = memory_channel

# 配置 Source：监控本地文件
agent.sources.local_source.type = spooldir
agent.sources.local_source.spoolDir = /path/to/your/local/directory
agent.sources.local_source.fileHeader = true
agent.sources.local_source.channels = memory_channel

# 配置 Sink：将数据写入 HDFS
agent.sinks.hdfs_sink.type = hdfs
agent.sinks.hdfs_sink.hdfs.path = hdfs://namenode_host:8020/path/to/your/hdfs/directory
agent.sinks.hdfs_sink.hdfs.fileType = DataStream
agent.sinks.hdfs_sink.hdfs.filePrefix = events-
agent.sinks.hdfs_sink.hdfs.fileSuffix = .log
agent.sinks.hdfs_sink.hdfs.rollSize = 0
agent.sinks.hdfs_sink.hdfs.rollCount = 10000
agent.sinks.hdfs_sink.hdfs.batchSize = 1000
agent.sinks.hdfs_sink.hdfs.useLocalTimeStamp = true
agent.sinks.hdfs_sink.channel = memory_channel

# 配置 Channel：内存通道，用于在 Source 和 Sink 之间传输数据
agent.channels.memory_channel.type = memory
agent.channels.memory_channel.capacity = 10000
agent.channels.memory_channel.transactionCapacity = 1000

# 将数据从 Source 传输到 Sink 的绑定
agent.sources.local_source.channels = memory_channel
agent.sinks.hdfs_sink.channel = memory_channel
```

解释配置文件内容：

- **Source (`local_source`)**：使用 `spooldir` 类型，监控指定的本地目录（`spoolDir`）。这里假设词频数据文件是作为文本文件存储在本地目录中。
- **Sink (`hdfs_sink`)**：使用 `hdfs` 类型，将数据写入到 HDFS 中的指定路径（`hdfs.path`）。可以配置文件名前缀、后缀以及数据滚动（`rollSize`、`rollCount` 等）。

- **Channel (`memory_channel`)**: 使用 `memory` 类型的通道，用于在 Source 和 Sink 之间传输数据。可以配置通道的容量（`capacity`）和事务容量（`transactionCapacity`）。

3. 启动 Flume Agent

将上述配置保存为 `flume.conf` 文件，并通过以下命令启动 Flume Agent：

```bash
flume-ng agent -n agent -c conf -f /path/to/your/flume.conf -Dflume.root.logger=INFO,console
```

注意事项：

- 确保本地目录（`spoolDir`）中包含你想要监控的词频数据文件，并且 Flume 有权限读取这些文件。
- 替换配置中的路径和主机名为你的实际路径和主机名。
- 根据实际需求调整 `hdfs` sink 的配置，比如文件的滚动策略和批处理大小。

通过这些步骤，Flume 将会持续监控指定的本地目录中的文件，并将其内容实时传输到指定的 HDFS 路径中，实现数据的持续流动和存储。