课堂实验

一、实验信息

文档编号		文档版本	1.0
实验名称	Flume 安装与使用		
所属课程	第 6 讲 分布式数据采集工具 Flume	认证等级	数据工程师（初级）
授课形式	上机实验	实验批次	第 6 次 / 共 11 次

二、实验目的

- 掌握 Flume 的安装和使用方法。
- 理解 Flume Agent 及其组件的基本功能。

三、实验准备

- 实验目录与命名规划
  （1）本地目录
    ① 工作目录：/mnt/disk1/{student_name}
    ② Flume 监听目录：/mnt/disk1/{student_name}/flumeSpool
  （2）HDFS 目录
    工作目录：/tmp/{student_name}
    Flume 数据采集目录：/tmp/{student_name}/flume_data
  （3）{student_name}为变量，代表学员姓名全拼
- 文件服务器
  （1）IP：172.16.140.111
  （2）目录：/mnt/disk1/de_training

四、实验内容

1、安装 Flume

- 任务：将 Flume 安装包复制到集群服务器中，并解压完成安装。
- 步骤

Linux：

```
// 登录文件服务器，将 Flume 安装包复制到集群的第一个节点
1. cd /mnt/disk1/de_training
2. scp apache-flume-1.7.0-bin.tar.gz 172.16.140.85:/mnt/disk1/{student_name}
// 登录集群节点，解压 Flume 安装包
3. cd /mnt/disk1/{student_name}
4. tar -xzvf apache-flume-1.7.0-bin.tar.gz
```

2、配置 Flume

- 任务：新建 Flume 配置文件，设置 source 类型为 spooldir、sink 类型为 hdfs、channel 类型为 memory。
- 步骤

Linux：

```
// 新建 flume.conf，文件内容如下所示
1. cd /mnt/disk1/{student_name}/apache-flume-1.7.0-bin/conf
2. vim flume.conf
```

flume.conf 文件内容：

```
# 定义 Agent 组件名
a1.sources = r1
a1.sinks = k1
a1.channels = c1
# 配置 Souce 组件
a1.sources.r1.type = spooldir
a1.sources.r1.spoolDir = /mnt/disk1/{student_name}/flumeSpool
a1.sources.r1.fileHeader = true
# 配置 Sink 组件
a1.sinks.k1.type = hdfs
a1.sinks.k1.channel = c1
a1.sinks.k1.hdfs.path = /tmp/{student_name}/flume_data
a1.sinks.k1.hdfs.rollSize = 0
a1.sinks.k1.hdfs.rollCount = 0
# 配置 Channel 组件
a1.channels.c1.type = memory
a1.channels.c1.capacity = 5000
a1.channels.c1.transactionCapacity = 5000
# 设置 Source、Sink 和 Channel 的关系
a1.sources.r1.channels = c1
a1.sinks.k1.channel = c1
```

3、启动 TDH Client

- 任务：执行 TDH Client 的 init.sh 脚本，启动 TDH Client。

- 步骤

```
Linux：
  // 执行 TDH Client 的 init.sh 脚本
  1. source {TDH_Client_install_dir}/init.sh
```

4、创建 Flume 监听和存储目录

- 任务：在本地创建 Flume 监听目录并复制日志文件，在 HDFS 中创建 Flume 存储目录。
- 步骤

```
Linux：
  // 在本地创建 Flume 监听目录
  1. mkdir -p /mnt/disk1/{student_name}/flumeSpool
  // 将 transwarp-manager.log 复制到本地监听目录中
  2. cp /var/log/transwarp-manager/master/transwarp-manager.log
     /mnt/disk1/{student_name}/flumeSpool
  // 将 Hadoop 当前用户设置为 hdfs，进行访问授权
  3. export HADOOP_USER_NAME=hdfs
  // 在 HDFS 上创建 Flume 存储目录，并设置权限为 777
  4. hadoop fs -mkdir -p /tmp/{student_name}/flume_data
  5. hadoop fs -chmod -R 777 /tmp/{student_name}/flume_data
```

5、运行 Flume

- 任务：运行 Flume Agent，对本地目录进行监听，如有新增文件，则将文件内容采集到 HDFS 存储目录中。
- 步骤

```
Linux：
  /* 运行 Flume Agent，其中 a1 为 Agent 名称，conf 为 Flume 配置文件所在目录，flume.conf
     为 Flume 配置文件 */
  1. cd /mnt/disk1/{student_name}/apache-flume-1.7.0-bin
  2. bin/flume-ng agent -n a1 -c conf -f conf/flume.conf
  // 查看 Flume 是否采集到新文件
  3. hadoop fs -ls /tmp/{student_name}/flume_data
```