《数据采集与预处理》

教材官网： http://dblab.xmu.edu.cn/post/data-collection/

温馨提示：编辑幻灯片母版，可以修改每页PPT的厦大校徽和底部文字

第5章 日志采集系统Flume

（PPT版本号：2022年1月版本）

林子雨 副教授

厦门大学计算机科学与技术系

E-mail: ziyulin@xmu.edu.cn

主页：**http://dblab.xmu.edu.cn/linziyu**

提纲

5.1 Flume简介
5.2Flume的安装和使用
5.3 Flume和Kafka的组合使用
5.4 采集日志文件到HDFS
5.5采集MySQL数据到HDFS

本PPT是以下教材的配套讲义
林子雨编著《数据采集与预处理》
人民邮电出版社

教材官网：
http://dblab.xmu.edu.cn/post/data-collection

5.1 Flume简介

Flume 运行的核心是Agent。Flume以Agent为最小的独立运行单位，一个Agent就是一个JVM（Java Virtual Machine），它是一个完整的数据采集工具，包含三个核心组件，分别是数据源（Source）、数据通道（Channel）和数据槽（Sink）。通过这些组件，“事件”可以从一个地方流向另一个地方。每个组件的具体功能如下（如图5-1所示）：

- （1）数据源是数据的收集端，负责将数据捕获后进行特殊的格式化，将数据封装到事件（Event）里，然后将事件推入数据通道中。常用的数据源的类型包括Avro、Thrift、Exec、JMS、Spooling Directory、Taildir、Kafka、NetCat、Syslog、HTTP等。
- （2）数据通道是连接数据源和数据槽的组件，可以将它看作一个数据的缓冲区（数据队列），它可以将事件暂存到内存中，也可以持久化到本地磁盘上，直到数据槽处理完该事件。常用的数据通道类型包括Memory、JDBC、Kafka、File、Custom等。
- （3）数据槽取出数据通道中的数据，存储到文件系统和数据库，或者提交到远程服务器。常用的数据槽包括HDFS、Hive、Logger、Avro、Thrift、IRC、File Roll、HBase、ElasticSearch、Kafka、HTTP等。

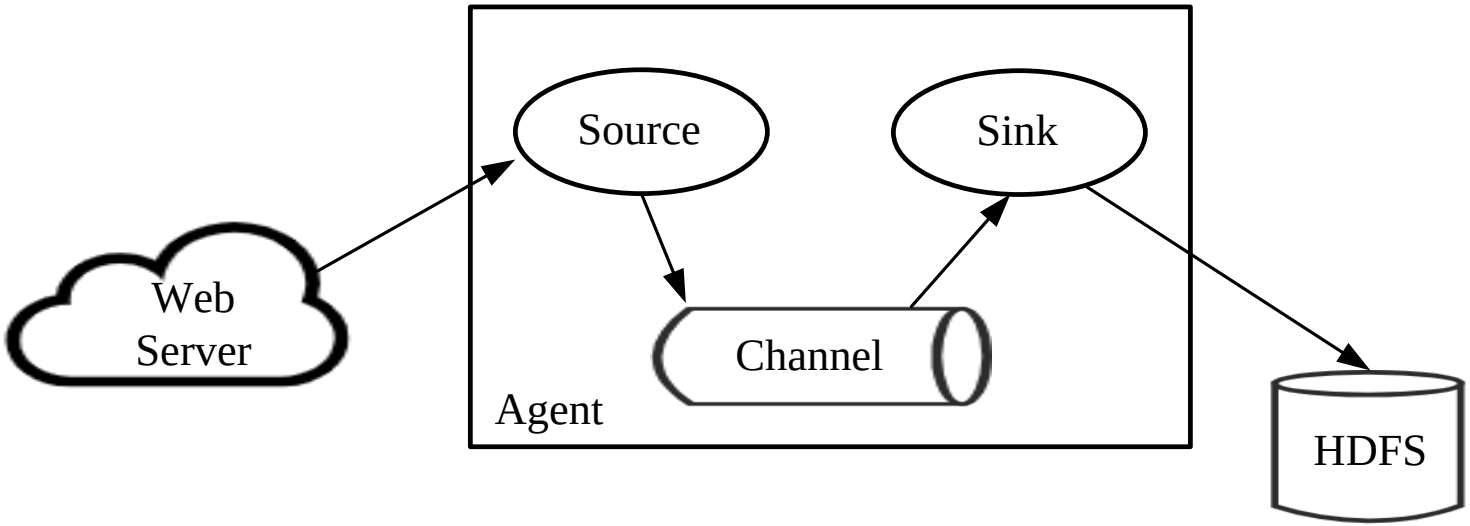


图5-1 Flume的技术架构

Flume提供了大量内置的数据源、数据通道和数据槽类型。不同类型的数据源、数据通道和数据槽可以自由组合。组合方式基于用户设置的配置文件，非常灵活。比如，数据通道可以把事件暂存在内存里，也可以持久化到本地硬盘上；数据槽可以把日志写入HDFS、HBase甚至是另外一个数据源等等。

5.2Flume的安装和使用

5.2.1 Flume的安装
5.2.2 Flume的使用

5.2.1 Flume的安装

Flume的运行需要Java环境的支持，因此，需要在Windows系统中安装JDK。请参照第2章内容完成JDK的安装。

访问Flume官网（http://flume.apache.org/download.html），下载Flume安装文件apache-flume-1.9.0-bin.tar.gz。把安装文件解压缩到Windows系统的“C:\”目录下，然后，执行如下命令测试是否安装成功：

```
> cd c:\apache-flume-1.9.0-bin\bin
> flume-ng version
```

如果能够返回类似如下的信息，则表示启动成功：

```
Flume 1.9.0
Source code repository: https://git-wip-us.apache.org/repos/asf/flume.git
Revision: d4fcab4f501d41597bc616921329a4339f73585e
Compiled by fszabo on Mon Dec 17 20:45:25 CET 2018
From source with checksum 35db629a3bda49d23e9b3690c80737f9
```

5.2.2 Flume的使用

1.实例1：采集NetCat数据显示到控制台

这里给出一个简单的实例，假设Source为NetCat类型，使用Telnet连接Source写入数据，产生日志数据输出到控制台（屏幕）。下面首先介绍操作系统Windows7中的操作方法，然后再介绍Windows10中的操作方法。

为了顺利完成后面的测试，首先开启Windows7的telnet服务。具体方法是，打开“控制面板”，进入“默认程序”，在出现的界面的左侧底部点击“程序和功能”，再在出现的界面的左侧顶部点击“打开或关闭Windows功能”，会出现如图5-2所示的界面，把“Telnet服务器”和“Telnet客户端”都选中，然后点击“确定”按钮。

5.2.2 Flume的使用

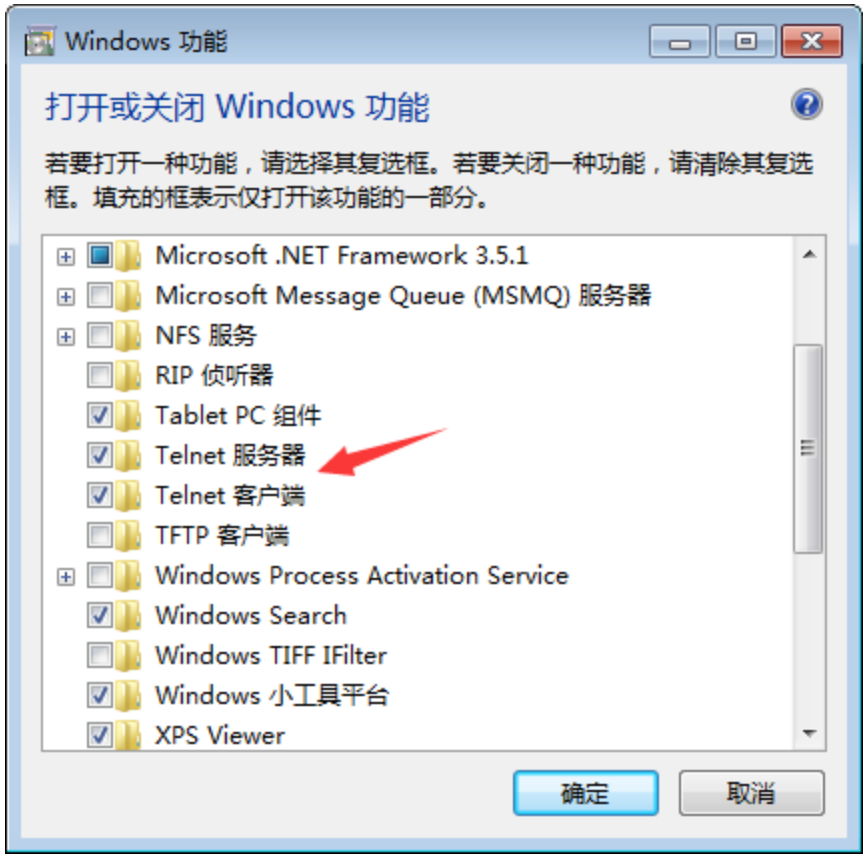


图5-2 打开或关闭Windows功能

5.2.2 Flume的使用

在Flume安装目录的conf子目录下，新建一个名称为example.conf的配置文件，该文件的内容如下：

```
# 设置Agent上的各个组件名称
a1.sources = r1
a1.sinks = k1
a1.channels = c1

# 配置Source
a1.sources.r1.type = netcat
a1.sources.r1.bind = localhost
a1.sources.r1.port = 44444

# 配置Sink
a1.sinks.k1.type = logger
```

5.2.2 Flume的使用

```
# 配置Channel
a1.channels.c1.type = memory
a1.channels.c1.capacity = 1000
a1.channels.c1.transactionCapacity = 100

# 把Source和Sink绑定到Channel上
a1.sources.r1.channels = c1
a1.sinks.k1.channel = c1
```

5.2.2 Flume的使用

在这个配置文件中，设置了Source类型为netcat，Channel类型为memory，Sink的类型为logger。

然后，新建一个cmd窗口（称为“Flume窗口”），并执行如下命令：

> cd c:\apache-flume-1.9.0-bin

> .\bin\flume-ng agent --conf .\conf --conf-file .\conf\example.conf --name a1 -property flume.root.logger=INFO,console

再新建一个cmd窗口，并执行如下命令：

> telnet localhost 44444

这时就可以从键盘输入一些英文单词，比如“Hello World”，切换到Flume窗口，就可以看到屏幕上显示了“Hello World”（如图5-3所示），说明Flume成功地接收到了信息。

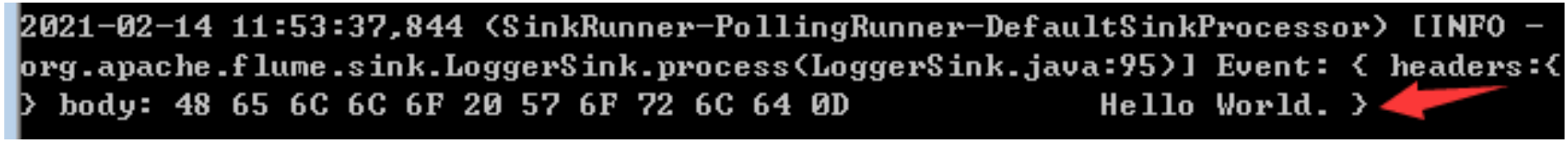


图5-3 Flume窗口中显示接收到的信息

5.2.2 Flume的使用

上面介绍了Windows7中的操作方法，现在介绍Windows10中的操作方法。在Windows10中，运行Flume的操作和Windows7一样，不同是的Telnet操作。由于Telnet服务端的安全性问题，Windows10中移除了Telnet服务端组件，也就是说，在Windows10中无法找到Telnet服务端组件，也就无法执行“telnet localhost 44444”命令，因此，操作方法不同于Windows7。为了能够执行“telnet localhost 44444”命令，这里使用子系统的方法通过Linux的telnet命令进行操作，操作步骤如下：

5.2.2 Flume的使用

（1）进入Windows10自带的“软件商店”（Mircrosoft Store），在软件商店中搜索“Ubuntu”，选择第一个进行下载（如图5-4所示）。下载结束后，如图5-5所示，点击“安装”按钮，完成Ubuntu系统的安装。

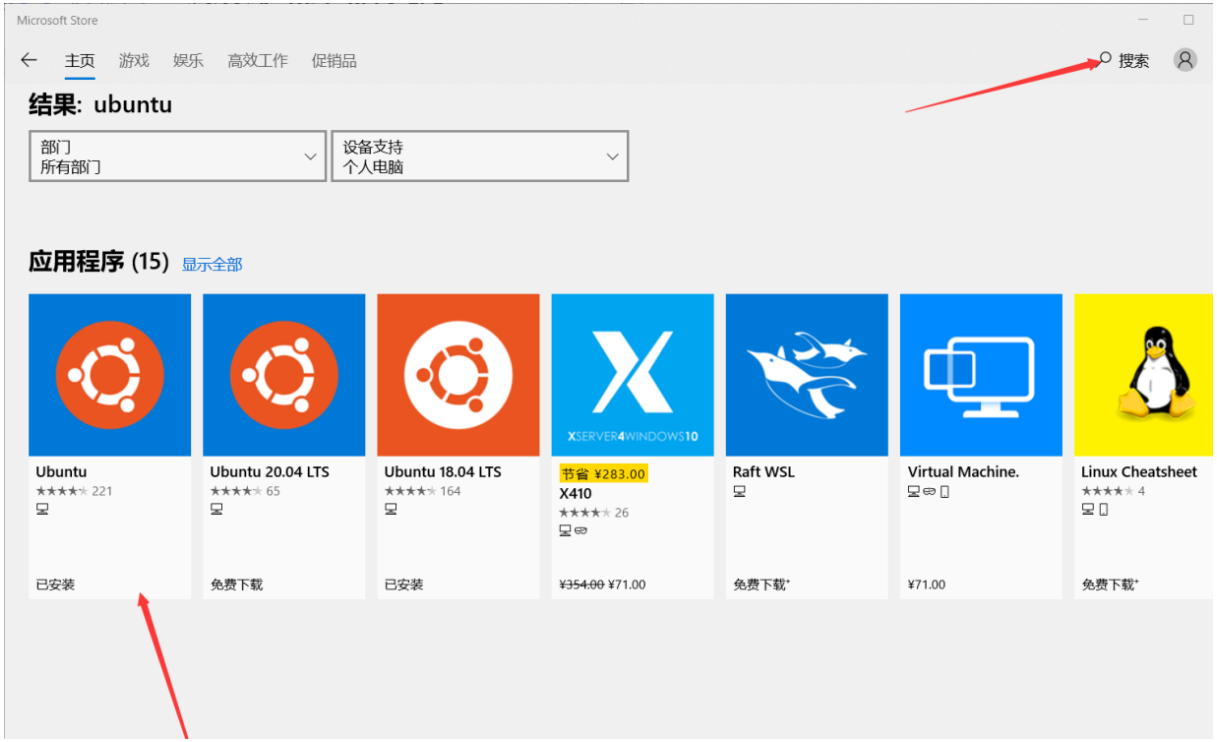


图5-4 在“软件商店”中下载Ubuntu

5.2.2 Flume的使用

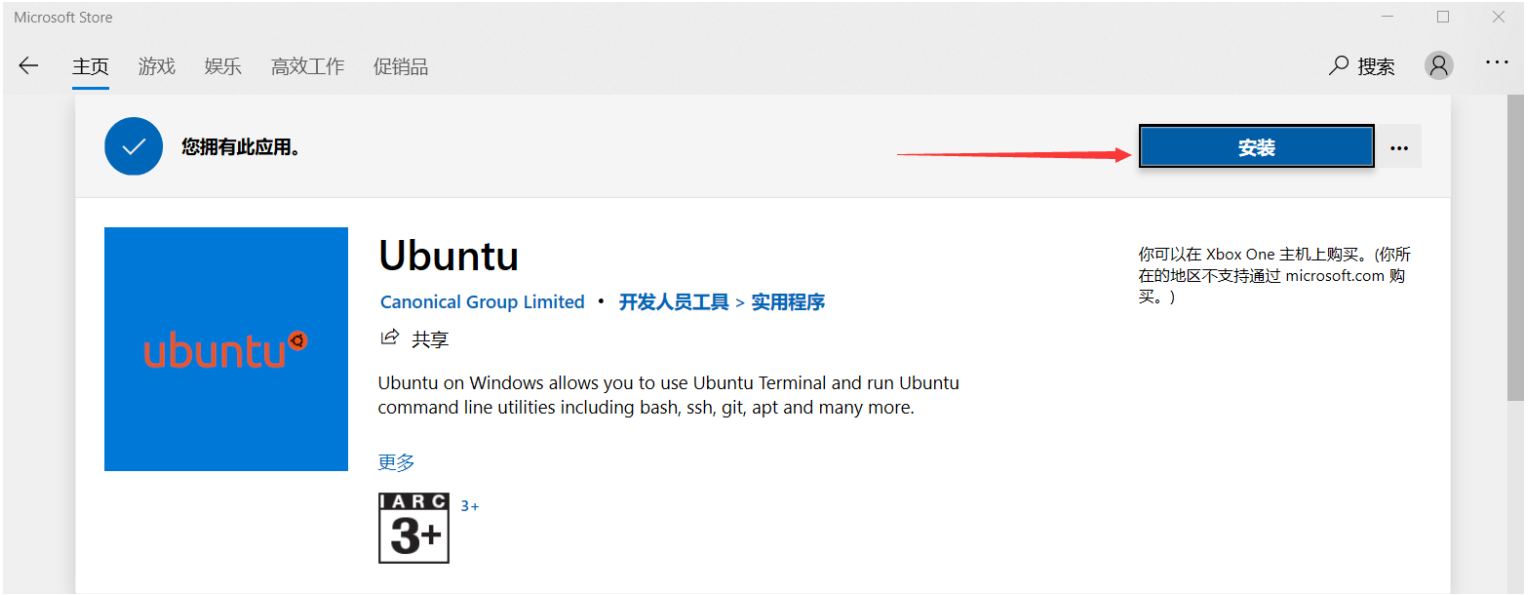


图5-5 安装Ubuntu

5.2.2 Flume的使用

（2）安装完成以后，可以从“开始”菜单中启动Ubuntu（如图5-6所示）。初次启动时会要求设置用户名和密码，设置以后就可以进入Ubuntu的命令界面。

图5-6 从“开始”菜单启动Ubuntu

5.2.2 Flume的使用

（3）在命令提示符后面输入“telnet localhost 44444”命令即可（如图5-7所示），Ubuntu子系统和原系统Windows10的端口信息可以互通，效果等同于Windows7中的telnet命令。这时就可以从键盘输入一些英文单词，比如“Hello World”，切换到Windows10中的Flume窗口，就可以看到屏幕上显示了“Hello World”，说明Flume成功地接收到了信息。

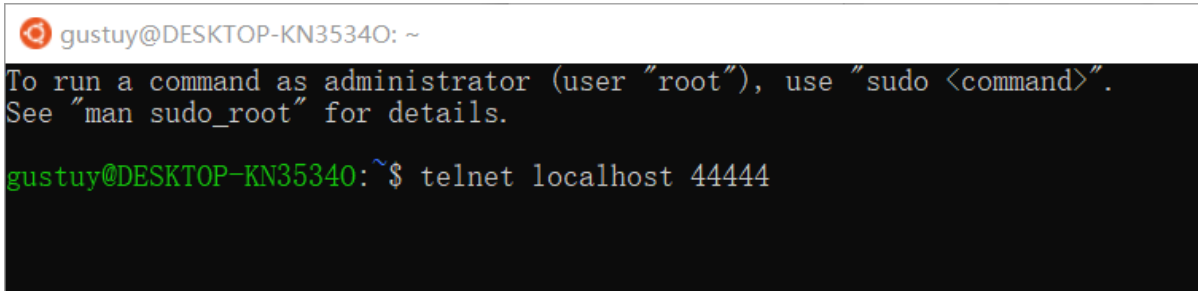


图5-7 执行telnet命令

5.2.2 Flume的使用

2.采集目录下的数据显示到控制台

假设Windows系统中有一个目录“C:\mylogs”，这个目录下不断有新的文件生成，使用Flume采集这个目录下的文件，并把文件内容显示到控制台（屏幕）。

在Flume安装目录的conf子目录下，新建一个名称为example1.conf的配置文件，该文件的内容如下：

5.2.2 Flume的使用

```
#定义三大组件名称
a1.sources = r1
a1.channels = c1
a1.sinks = k1

#定义Source
a1.sources.r1.type = spooldir
a1.sources.r1.spoolDir = C:/mylogs/

#定义Channel
a1.channels.c1.type = memory
a1.channels.c1.capacity = 10000
a1.channels.c1.transactionCapacity = 100

#定义Sink
a1.sinks.k1.type = logger

#组装Source、Channel、Sink
a1.sources.r1.channels = c1
a1.sinks.k1.channel = c1
```

5.2.2 Flume的使用

清空“C:\mylogs”目录（即删除该目录下的所有内容），然后新建一个cmd窗口（称为“Flume窗口”），并执行如下命令：

> cd c:\apache-flume-1.9.0-bin

➢.\bin\flume-ng agent --conf .\conf --conf-file .\conf\example1.conf --name a1 -property flume.root.logger=INFO,console

然后，在“C:\”目录下新建一个文件mylog.txt，里面输入一些内容，比如“I love Flume”，保存该文件，并把该文件复制到“C:\mylogs”目录下，可以看到，mylog.txt很快会变成mylog.txt.COMPLETED，这时，在Flume窗口中就可以看到mylog.txt中的内容，比如“I love Flume”。

5.3 Flume和Kafka的组合使用

在Windows系统中打开第1个cmd窗口，执行如下命令启动Zookeeper服务：

```
> cd c:\kafka_2.12-2.4.0
> .\bin\windows\zookeeper-server-start.bat .\config\zookeeper.Properties
```

打开第2个cmd窗口，然后执行下面命令启动Kafka服务：

```
> cd c:\kafka_2.12-2.4.0
> .\bin\windows\kafka-server-start.bat .\config\server.properties
```

打开第3个cmd窗口，执行如下命令创建一个名为test的Topic：

```
> cd c:\kafka_2.12-2.4.0
> .\bin\windows\kafka-topics.bat --create --zookeeper localhost:2181 --replication-factor 1 --partitions 1 --topic test
```

5.3 Flume和Kafka的组合使用

在Flume的安装目录的conf子目录下创建一个配置文件kafka.conf，内容如下：

```
# Name the components on this agent
a1.sources = r1
a1.sinks = k1
a1.channels = c1

# source
a1.sources.r1.type = netcat
a1.sources.r1.bind = localhost
a1.sources.r1.port = 44444
```

5.3 Flume和Kafka的组合使用

```
# sink
a1.sinks.k1.type = org.apache.flume.sink.kafka.KafkaSink
a1.sinks.k1.kafka.topic = test
a1.sinks.k1.kafka.bootstrap.servers = localhost:9092
a1.sinks.k1.kafka.flumeBatchSize = 20
a1.sinks.k1.kafka.producer.acks = 1
a1.sinks.k1.kafka.producer.linger.ms = 1
a1.sinks.k1.kafka.producer.compression.type = snappy

# channel
a1.channels.c1.type = memory
a1.channels.c1.capacity = 1000
a1.channels.c1.transactionCapacity = 100

# Bind the source and sink to the channel
a1.sources.r1.channels = c1
a1.sinks.k1.channel = c1
```

5.3 Flume和Kafka的组合使用

打开第4个cmd窗口，执行如下命令启动Flume：
> cd c:\apache-flume-1.9.0-bin
> .\bin\flume-ng.cmd agent --conf ./conf --conf-file ./conf/kafka.conf --name a1 -property flume.root.logger=INFO,console
打开第5个cmd窗口，执行如下命令：
> telnet localhost 44444
执行上面命令以后，在该窗口内用键盘输入一些单词，比如“hadoop”。这个单词会发送给Flume，然后，Flume发送给Kafka。
打开第6个cmd窗口，执行如下命令：
> cd c:\kafka_2.12-2.4.0
> .\bin\windows\kafka-console-consumer.bat --bootstrap-server localhost:9092 --topic test --from-beginning
上面命令执行以后，就可以在屏幕上看到“hadoop”，说明Kafka成功接收到了数据。

5.4 采集日志文件到HDFS

5.4.1 采集目录到HDFS
5.4.2 采集文件到HDFS

5.4.1 采集目录到HDFS

采集需求是某服务器的某特定目录下（比如“C:\mylog”），会不断产生新的文件，每当有新文件出现，就需要把文件采集到HDFS中去。

根据需求，首先定义以下3大要素：

- Source：因为要监控文件目录，所以Souce的类型是spooldir；
- Sink：因为要把文件采集到HDFS中，所以，Sink的类型是hdfs；
- Channel：Channel的类型可以设置为memory。

5.4.1 采集目录到HDFS

在Flume安装目录的conf子目录下，编写一个配置文件spooldir_hdfs.conf，其内容如下：

```
#定义三大组件的名称
agent1.sources = source1
agent1.sinks = sink1
agent1.channels = channel1

# 配置Source组件
agent1.sources.source1.type = spooldir
agent1.sources.source1.spoolDir = C:/mylogs/
agent1.sources.source1.fileHeader = false
```

5.4.1 采集目录到HDFS

```
# 配置Sink组件
agent1.sinks.sink1.type = hdfs
agent1.sinks.sink1.hdfs.path =hdfs://localhost:9000/weblog/%y-%m-%d/%H-%M
agent1.sinks.sink1.hdfs.filePrefix = access_log
agent1.sinks.sink1.hdfs.maxOpenFiles = 5000
agent1.sinks.sink1.hdfs.batchSize= 100
agent1.sinks.sink1.hdfs.fileType = DataStream
agent1.sinks.sink1.hdfs.writeFormat =Text
agent1.sinks.sink1.hdfs.rollSize = 102400
agent1.sinks.sink1.hdfs.rollCount = 1000000
agent1.sinks.sink1.hdfs.rollInterval = 60
#agent1.sinks.sink1.hdfs.round = true
#agent1.sinks.sink1.hdfs.roundValue = 10
#agent1.sinks.sink1.hdfs.roundUnit = minute
agent1.sinks.sink1.hdfs.useLocalTimeStamp = true
```

5.4.1 采集目录到HDFS

```
# 设置Channel
agent1.channels.channel1.type = memory
agent1.channels.channel1.keep-alive = 120
agent1.channels.channel1.capacity = 500000
agent1.channels.channel1.transactionCapacity = 600

# 把Source和Sink绑定到Channel上
agent1.sources.source1.channels = channel1
agent1.sinks.sink1.channel = channel1
```

5.4.1 采集目录到HDFS

为了让Flume能够顺利访问HDFS，需要把Flume安装目录下的lib子目录下的guava-11.0.2.jar文件删除，然后，把Hadoop安装目录下的“share\hadoop\common\lib”目录下的guava-27.0-jre.jar文件复制到Flume安装目录下的lib子目录下。

在Windows系统中，新建一个cmd窗口，使用如下命令启动Hadoop的HDFS：

```
> cd c:\hadoop-3.1.3\sbin
> start-dfs.cmd
```

执行JDK自带的命令jps查看Hadoop已经启动的进程：

```
> jps
```

需要注意的是，这里在使用jps命令的时候，没有带上绝对路径，是因为已经把JDK添加到了Path环境变量中。

执行jps命令以后，如果能够看到“DataNode”和“NameNode”这两个进程，就说明Hadoop启动成功。

5.4.1 采集目录到HDFS

再新建一个cmd窗口，执行如下命令启动Flume：

```
> cd c:\apache-flume-1.9.0-bin
> .\bin\flume-ng agent --conf .\conf --conf-file .\conf\spooldir_hdfs.conf --name agent1 -property flume.root.logger=INFO,console
```

执行上述命令以后，Flume就被启动了，开始实时监控“C:/mylogs/”目录，只要这个目录下有新的文件生成，就会被Flume捕捉到，并把文件内容保存到HDFS中。在C盘根目录下新建一个文本文件mylog1.txt，里面写入一些句子，比如“This is mylog1”，然后，把mylog1.txt文件复制到“C:\mylog”目录下，过一会儿，就会看到mylog1.txt文件名被修改成了mylog4.txt.COMPLETED，说明该文件已经成功被Flume捕捉到。可以在HDFS的WEB管理页面中（http://localhost:9870）查看生成的文件及其内容。

5.4.2 采集文件到HDFS

采集需求是某服务器的某特定目录下的文件（比如“C:\mylog\log1.txt”），会不断发生更新，每当文件被更新时，就需要把更新的数据采集到HDFS中去。

根据需求，首先定义以下3大要素：

- Source：因为要监控文件内容，所以Souce的类型是exec；
- Sink：因为要把文件采集到HDFS中，所以，Sink的类型是hdfs；
- Channel：Channel的类型可以设置为memory。

5.4.2 采集文件到HDFS

在Flume安装目录的conf子目录下，编写一个配置文件exec_hdfs.conf，其内容如下：

```
#定义三大组件的名称
agent1.sources = source1
agent1.sinks = sink1
agent1.channels = channel1

# 配置Source组件
agent1.sources.source1.type = exec
agent1.sources.source1.command = tail -F C:/mylogs/log1.txt
agent1.sources.source1.channels = channel1
```

5.4.2 采集文件到HDFS

```
# 配置Sink组件
agent1.sinks.sink1.type = hdfs
agent1.sinks.sink1.hdfs.path =hdfs://localhost:9000/weblog/%y-%m-%d/%H-%M
agent1.sinks.sink1.hdfs.filePrefix = access_log
agent1.sinks.sink1.hdfs.maxOpenFiles = 5000
agent1.sinks.sink1.hdfs.batchSize= 100
agent1.sinks.sink1.hdfs.fileType = DataStream
agent1.sinks.sink1.hdfs.writeFormat =Text
agent1.sinks.sink1.hdfs.rollSize = 102400
agent1.sinks.sink1.hdfs.rollCount = 1000000
agent1.sinks.sink1.hdfs.rollInterval = 60
#agent1.sinks.sink1.hdfs.round = true
#agent1.sinks.sink1.hdfs.roundValue = 10
#agent1.sinks.sink1.hdfs.roundUnit = minute
agent1.sinks.sink1.hdfs.useLocalTimeStamp = true
```

5.4.2 采集文件到HDFS

```
# 配置Channel组件
agent1.channels.channel1.type = memory
agent1.channels.channel1.keep-alive = 120
agent1.channels.channel1.capacity = 500000
agent1.channels.channel1.transactionCapacity = 600

# 把Source和Sink绑定到Channel
agent1.sources.source1.channels = channel1
agent1.sinks.sink1.channel = channel1
```

5.4.2 采集文件到HDFS

在上面的配置文件中，有一行内容如下：

agent1.sources.source1.command = tail -F C:/mylogs/log1.txt

在这个配置信息中，使用了tail命令，Windows系统没有自带tail命令，因此，需要单独安装。可以到网络查找tail.exe文件，或者直接到教材官网的“下载专区”的“软件”目录中下载tail.zip文件，解压缩生成tail.exe文件，再把tail.exe文件复制到“C:\Windows\System32”目录下，然后，可以测试一下该命令的效果。首先新建一个文件“C:\mylog\log1.txt”，文件内容可以为空，然后，打开一个cmd窗口（这里称为“tail窗口”），输入如下命令：

> tail -f c:\mylogs\log1.txt

然后，用记事本打开log1.txt，向里面输入一些内容（比如“I love Flume”）并且保存文件，这时，tail窗口内就会显示刚刚输入到log1.txt中的这些内容。

5.4.2 采集文件到HDFS

再新建一个cmd窗口，启动HDFS，然后执行如下命令启动Flume：

```
> cd c:\apache-flume-1.9.0-bin
> .\bin\flume-ng agent --conf .\conf --conf-file .\conf\exec_hdfs.conf --name agent1 -property flume.root.logger=INFO,console
```

执行上述命令以后，Flume就被启动了，开始实时监控“C:/mylogs/log1.txt”文件，只要这个文件发生了内容更新，就会被Flume捕捉到，并把更新内容保存到HDFS中。作为测试，可以在log1.txt文件中输入一些内容，然后到HDFS的WEB管理页面中（http://localhost:9870）查看生成的文件及其内容。

5.5采集MySQL数据到HDFS

5.5.1 准备工作
5.5.2 创建MySQL数据库
5.5.3 配置和启动Flume

5.5.1 准备工作

在采集MySQL数据库中的数据到HDFS时，需要用到一个第三方JAR包，即flume-ng-sql-source-1.5.2.jar。这个JAR包可以直接从网络上下载，或者也可以到教材官网的“下载专区”的“软件”目录中下载。但是，直接下载的JAR包一般都不是最新的版本，或者可能与已经安装的Flume不兼容，因此，这里介绍自己下载源代码进行编译得到JAR包的方法。

为了对源代码进行编译，这里需要用到Maven工具，可以到Maven官网（https://maven.apache.org/download.cgi）下载安装包apache-maven-3.6.3-bin.zip，然后，解压缩到Windows系统的“C:\”目录下。

访问github网站（https://github.com/keedio/flume-ng-sql-source/tree/release/1.5.2），在出现的页面（如图5-8所示）中，点击右上角的“Code”按钮，在弹出的菜单中点击“Download ZIP”，就可以把压缩文件flume-ng-sql-source-release-1.5.2.zip下载到本地。然后，把文件解压缩到Windows系统的“C:\”目录下。

5.5.1 准备工作

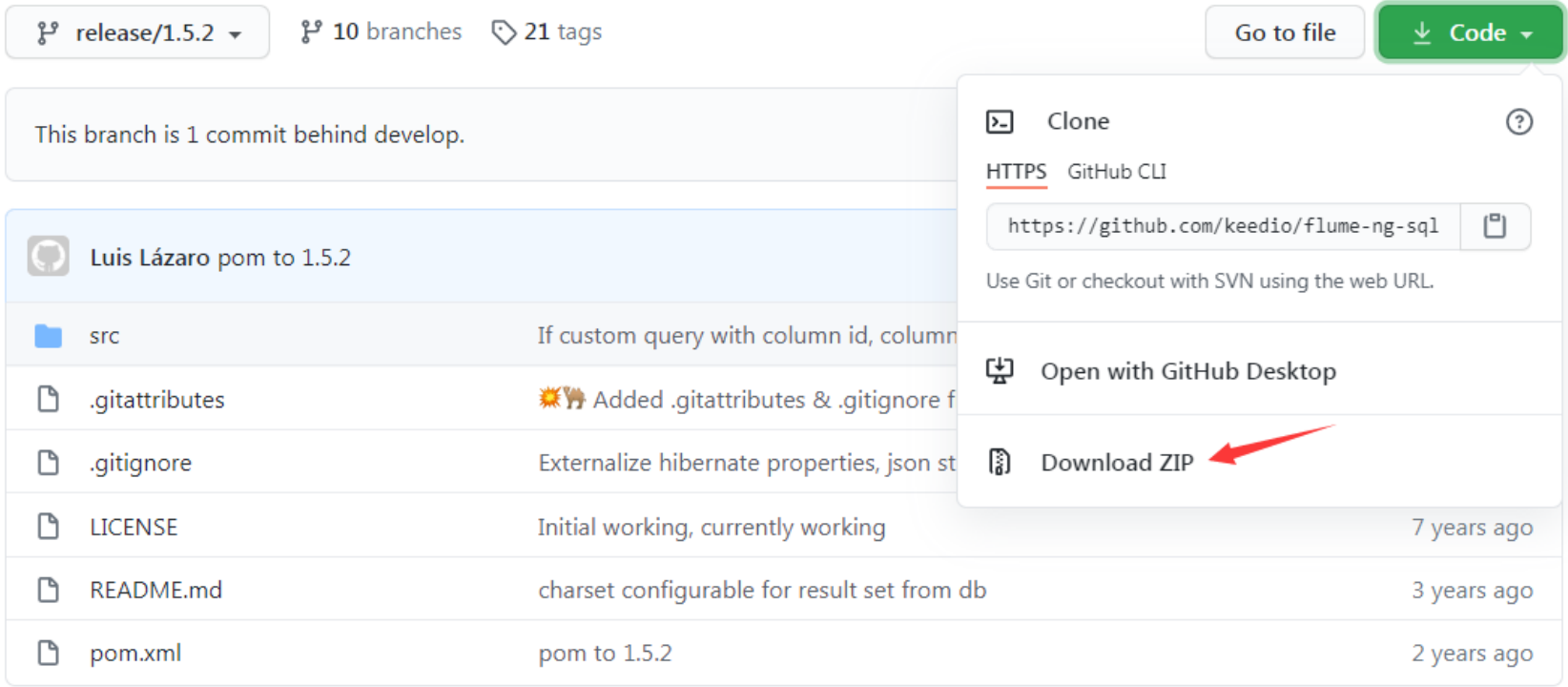


图5-8 github网站页面

5.5.1 准备工作

打开一个cmd窗口，执行如下命令执行编译打包：

```
> cd C:\flume-ng-sql-source-release-1.5.2
> C:\apache-maven-3.6.3\bin\mvn package
```

编译打包过程会持续一段时间，最终，如果编译打包成功，会返回类似如下的信息：

```
[INFO] Building jar: c:\flume-ng-sql-source-release-1.5.2\target\flume-ng-sql-source-1.5.2-javadoc.jar
[INFO] ------------------------------------------------------------------------
[INFO] BUILD SUCCESS
[INFO] ------------------------------------------------------------------------
[INFO] Total time:  04:27 min
[INFO] Finished at: 2021-02-17T09:36:13+08:00
[INFO] ------------------------------------------------------------------------
```

5.5.1 准备工作

这时，在“C:\flume-ng-sql-source-release-1.5.2\target”目录下，可以看到一个JAR包文件flume-ng-sql-source-1.5.2.jar，把这个文件复制到Flume安装目录的lib子目录下（比如“C:\apache-flume-1.9.0-bin\lib”）。

此外，为了让Flume能够顺利连接MySQL数据库，还需要用到一个连接驱动程序JAR包。可以访问MySQL官网（https://dev.mysql.com/downloads/connector/j/?os=26）下载驱动程序压缩文件mysql-connector-java-8.0.23.tar.gz（也可以到教材官网下载），然后，对该压缩文件进行解压缩，在解压后的目录中，找到文件mysql-connector-java-8.0.23.jar，把这个文件复制到Flume安装目录的lib子目录下（比如“C:\apache-flume-1.9.0-bin\lib”）。

5.5.2 创建MySQL数据库

参照第2章中关于MySQL数据库的内容，完成MySQL的安装，并学习其基本使用方法，这里假设读者已经成功安装了MySQL数据库并掌握了基本的使用方法。在Windows系统中，启动MySQL服务进程，然后，打开MySQL的命令行客户端，执行如下SQL语句创建数据库和表：

```
mysql>CREATE DATABASE school;
mysql> USE school;
mysql> CREATE TABLE student1(
    -> id INT NOT NULL,
    -> name VARCHAR(40),
    -> age INT,
    -> grade INT,
    -> PRIMARY KEY (id));
```

需要注意的是，在创建表的时候，一定要设置一个主键（比如，这里id是主键），否则后面Flume会捕捉数据失败。
创建好MySQL数据库以后，再执行如下命令启动Hadoop的HDFS：
> cd c:\hadoop-3.1.3\sbin
> start-dfs.cmd
执行JDK自带的命令jps查看Hadoop已经启动的进程：
> jps
执行jps命令以后，如果能够看到“DataNode”和“NameNode”这两个进程，就说明Hadoop启动成功。

5.5.3 配置和启动Flume

根据需求，首先定义以下3大要素：

•Source：因为要监控MySQL数据库，所以Souce的类型是org.keedio.flume.source.SQLSource；

•Sink：因为要把文件采集到HDFS中，所以，Sink的类型是hdfs；

•Channel：Channel的类型可以设置为memory。

5.5.3 配置和启动Flume

在Flume安装目录的conf子目录下，编写一个配置文件mysql_hdfs.conf，其内容如下：

```
#设置三大组件
agent1.channels = ch1
agent1.sinks = HDFS
agent1.sources = sql-source

#设置Source组件
agent1.sources.sql-source.type = org.keedio.flume.source.SQLSource
agent1.sources.sql-source.hibernate.connection.url = jdbc:mysql://localhost:3306/school
agent1.sources.sql-source.hibernate.connection.user = root  #数据库用户名
agent1.sources.sql-source.hibernate.connection.password = 123456  #数据库密码
agent1.sources.sql-source.hibernate.connection.autocommit = true
agent1.sources.sql-source.table = student    #数据库中的表名称
agent1.sources.sql-source.run.query.delay=5000
agent1.sources.sql-source.status.file.path = C:/apache-flume-1.9.0-bin/
agent1.sources.sql-source.status.file.name = sql-source.status
```

5.5.3 配置和启动Flume

```
#设置Sink组件
agent1.sinks.HDFS.type = hdfs
agent1.sinks.HDFS.hdfs.path = hdfs://localhost:9000/flume/mysql
agent1.sinks.HDFS.hdfs.fileType = DataStream
agent1.sinks.HDFS.hdfs.writeFormat = Text
agent1.sinks.HDFS.hdfs.rollSize = 268435456
agent1.sinks.HDFS.hdfs.rollInterval = 0
agent1.sinks.HDFS.hdfs.rollCount = 0

#设置Channel
agent1.channels.ch1.type = memory

#把Source和Sink绑定到Channel
agent1.sinks.HDFS.channel = ch1
agent1.sources.sql-source.channels = ch1
```

5.5.3 配置和启动Flume

在配置文件mysql_hdfs.conf中，有如下两行：
agent1.sources.sql-source.status.file.path = C:/apache-flume-1.9.0-bin/
agent1.sources.sql-source.status.file.name = sql-source.status
这两行设置了Flume状态信息的保存位置，即保存在“C:/apache-flume-1.9.0-bin/”目录下的“sql-source.status”这个文件中。需要重点强调的是，sql-source.status这个文件一定不要自己创建（如果自己创建，启动Flume时会报错），Flume在启动过程中会自动创建这个文件。如果已经存在sql-source.status这个文件，也要删除。

5.5.3 配置和启动Flume

在配置文件mysql_hdfs.conf中，还有如下一行：
agent1.sinks.HDFS.hdfs.path = hdfs://localhost:9000/flume/mysql
这行配置信息设置了数据在HDFS中的保存目录。需要注意的是，这个目录不需要自己创建，Flume会自动在HDFS中创建该目录。
下面执行如下命令启动Flume：
> cd c:\apache-flume-1.9.0-bin
> .\bin\flume-ng agent --conf .\conf --conf-file .\conf\mysql_hdfs.conf --name agent1 -property flume.root.logger=INFO,console
执行上述命令以后，Flume就被启动了，一定要注意启动过程中的返回信息，看看是否有返回错误信息，当返回的信息中没有包含任何错误信息时，就表示启动成功了。

5.5.3 配置和启动Flume

然后，在MySQL命令行客户端中执行如下语句向MySQL数据库中插入数据：

```
mysql> insert into student(id,name,age,grade) values(1,'Xiaoming',23,98)
mysql> insert into student(id,name,age,grade) values(2,'Zhangsan',24,96);
mysql> insert into student(id,name,age,grade) values(3,'Lisi',24,93);
```

到“C:/apache-flume-1.9.0-bin/”目录下查看“sql-source.status”这个文件，这个文件会包含类似如下的信息：

{"SourceName":"sql-source","URL":"jdbc:mysql:\/\/localhost:3306\/school","LastIndex":"3","ColumnsToSelect":"*","Table":"student"}

5.5.3 配置和启动Flume

在浏览器中输入网址“http://localhost:9870”打开Hadoop的WEB管理界面（如图5-9所示），就可以看到新生成的文件。

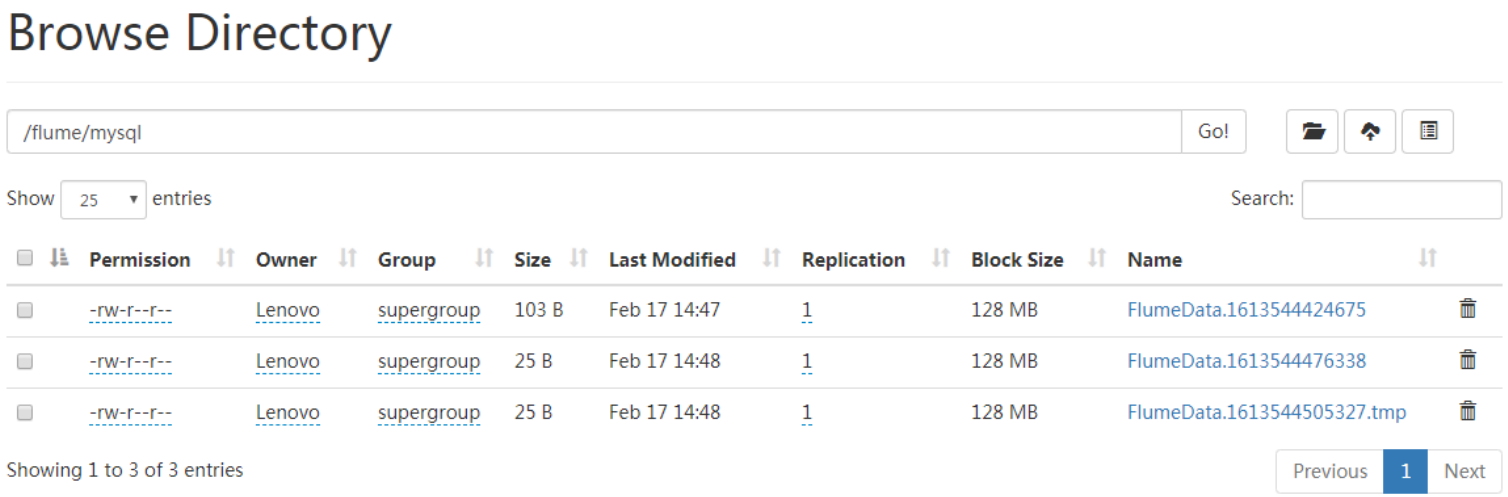


图5-9 HDFS的文件浏览页面

5.5.3 配置和启动Flume

打开其中一个文件（如图5-10所示），在出现的页面中点击“Tail the file(last 32K)”，就会显示文件的内容（如图5-11所示）。

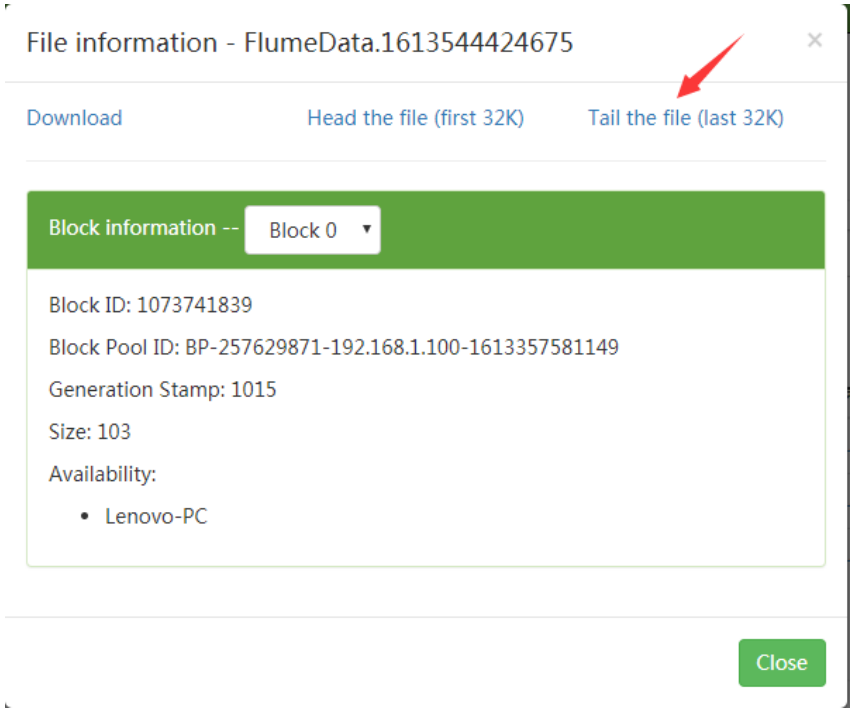


图5-10 HDFS的文件信息页面

5.5.3 配置和启动Flume

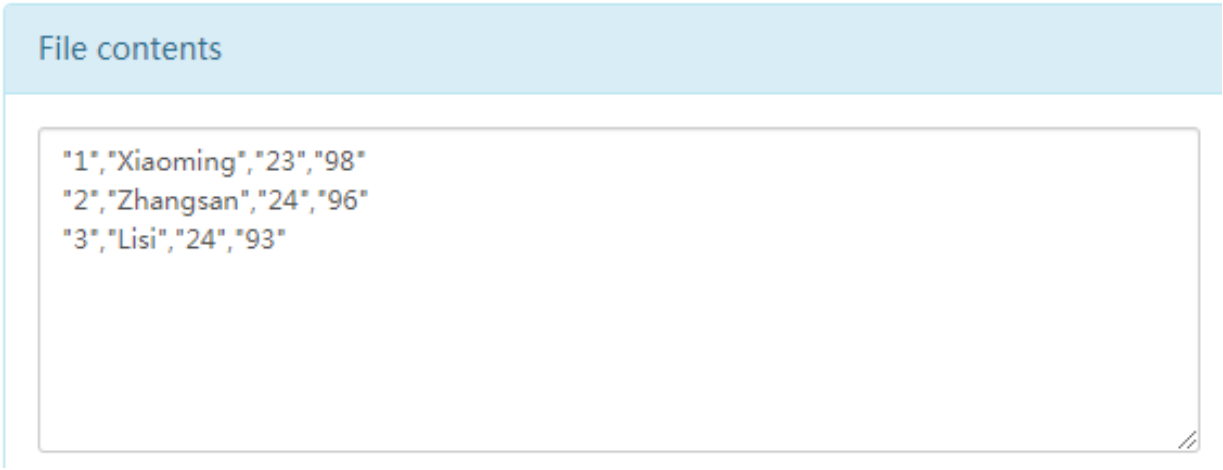


图5-11 HDFS的文件内容页面

5.6 本章小结

Flume最早是Cloudera提供的日志收集系统，是Apache下的一个孵化项目，Flume支持在日志系统中定制各类数据发送方，用于收集数据。Flume提供对数据进行简单处理，并写到各种数据接受方（可定制）的能力。Flume提供了从console（控制台）、RPC（Thrift-RPC）、text（文件）、tail（UNIX tail）、syslog（syslog日志系统）、exec（命令执行）等数据源上收集数据的能力。本章内容介绍了Flume的技术架构，并给出了Flume的安装和使用方法。本章介绍的Flume使用方法较为基础，如果要了解更加高级的使用方法，读者可以参考相关书籍或者网络资料。

附录A：主讲教师林子雨简介

主讲教师：林子雨

单位：厦门大学计算机科学与技术系
E-mail: ziyulin@xmu.edu.cn
个人网页：http://dblab.xmu.edu.cn/post/linziyu
数据库实验室网站：http://dblab.xmu.edu.cn

扫一扫访问个人主页

林子雨，男，1978年出生，博士（毕业于北京大学），全国高校知名大数据教师，现为厦门大学计算机科学系副教授，厦门大学信息学院实验教学中心主任，曾任厦门大学信息科学与技术学院院长助理、晋江市发展和改革局副局长。中国计算机学会数据库专业委员会委员，中国计算机学会信息系统专业委员会委员。国内高校首个“数字教师”提出者和建设者，厦门大学数据库实验室负责人，厦门大学云计算与大数据研究中心主要建设者和骨干成员，2013年度、2017年度和2020年度厦门大学教学类奖教金获得者，荣获2019年福建省精品在线开放课程、2018年厦门大学高等教育成果特等奖、2018年福建省高等教育教学成果二等奖、2018年国家精品在线开放课程。主要研究方向为数据库、数据仓库、数据挖掘、大数据、云计算和物联网，并以第一作者身份在《软件学报》《计算机学报》和《计算机研究与发展》等国家重点期刊以及国际学术会议上发表多篇学术论文。作为项目负责人主持的科研项目包括1项国家自然科学青年基金项目(No.61303004)、1项福建省自然科学青年基金项目(No.2013J05099)和1项中央高校基本科研业务费项目(No.2011121049)，主持的教改课题包括1项2016年福建省教改课题和1项2016年教育部产学协作育人项目，同时，作为课题负责人完成了国家发改委城市信息化重大课题、国家物联网重大应用示范工程区域试点泉州市工作方案、2015泉州市互联网经济调研等课题。中国高校首个“数字教师”提出者和建设者，2009年至今，“数字教师”大平台累计向网络免费发布超过1000万字高价值的研究和教学资料，累计网络访问量超过1000万次。打造了中国高校大数据教学知名品牌，编著出版了中国高校第一本系统介绍大数据知识的专业教材《大数据技术原理与应用》，并成为京东、当当网等网店畅销书籍；建设了国内高校首个大数据课程公共服务平台，为教师教学和学生学习大数据课程提供全方位、一站式服务，年访问量超过400万次，累计访问量超过1500万次。

附录B：大数据学习路线图

大数据学习路线图访问地址：http://dblab.xmu.edu.cn/post/10164/

附录C：林子雨大数据系列教材

了解全部教材信息：http://dblab.xmu.edu.cn/post/bigdatabook/

附录D：《大数据导论（通识课版）》教材

开设全校公共选修课的优质教材

本课程旨在实现以下几个培养目标：

- 引导学生步入大数据时代，积极投身大数据的变革浪潮之中
- 了解大数据概念，培养大数据思维，养成数据安全意识
- 认识大数据伦理，努力使自己的行为符合大数据伦理规范要求
- 熟悉大数据应用，探寻大数据与自己专业的应用结合点
- 激发学生基于大数据的创新创业热情

高等教育出版社 ISBN:978-7-04-053577-8 定价：32元 版次：2020年2月第1版
教材官网：http://dblab.xmu.edu.cn/post/bigdataintroduction/

附录E：《大数据导论》教材

- 林子雨 编著《大数据导论》
- 人民邮电出版社，2020年9月第1版
- ISBN:978-7-115-54446-9 定价：49.80元

教材官网：http://dblab.xmu.edu.cn/post/bigdata-introduction/

开设大数据专业导论课的优质教材

扫一扫访问教材官网

附录F：《大数据技术原理与应用（第3版）》教材

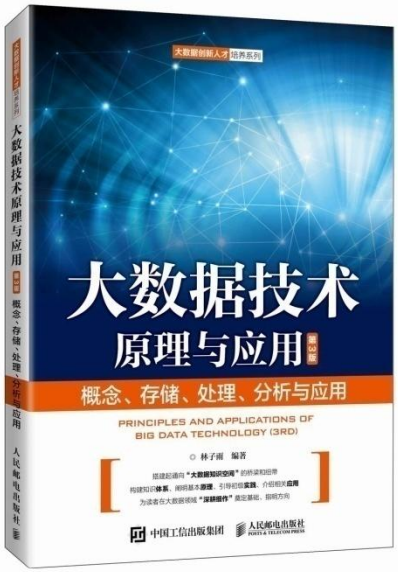
《大数据技术原理与应用——概念、存储、处理、分析与应用（第3版）》，由厦门大学计算机科学系林子雨博士编著，是国内高校第一本系统介绍大数据知识的专业教材。人民邮电出版社　ISBN:978-7-115-54405-6　定价：59.80元

全书共有17章，系统地论述了大数据的基本概念、大数据处理架构Hadoop、分布式文件系统HDFS、分布式数据 库HBase、NoSQL数据库、云数据库、分布式并行编程模型MapReduce、Spark、流计算、Flink、图计算、数据可视化以及大数据在互联网、生物医学和物流等各个领域的应用。在Hadoop、HDFS、HBase、MapReduce、Spark和Flink等重要章节，安排了入门级的实践操作，让读者更好地学习和掌握大数据关键技术。

本书可以作为高等院校计算机专业、信息管理等相关专业的大数据课程教材，也可供相关技术人员参考、学习、培训之用。

欢迎访问《大数据技术原理与应用——概念、存储、处理、分析与应用》教材官方网站：
http://dblab.xmu.edu.cn/post/bigdata3

扫一扫访问教材官网



附录G：《大数据基础编程、实验和案例教程（第2版）》

本书是与《大数据技术原理与应用（第3版）》教材配套的唯一指定实验指导书

大数据教材

1+1黄金组合

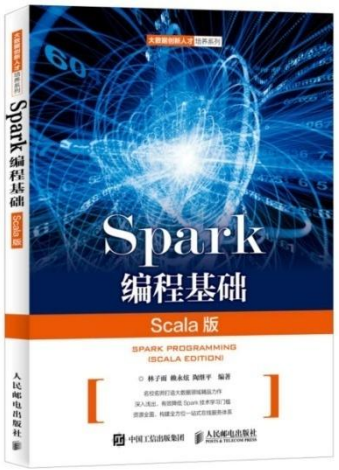
厦门大学林子雨编著

配套实验指导书

- 步步引导，循序渐进，详尽的安装指南为顺利搭建大数据实验环境铺平道路
- 深入浅出，去粗取精，丰富的代码实例帮助快速掌握大数据基础编程方法
- 精心设计，巧妙融合，八套大数据实验题目促进理论与编程知识的消化和吸收
- 结合理论，联系实际，大数据课程综合实验案例精彩呈现大数据分析全流程

林子雨编著《大数据基础编程、实验和案例教程（第2版）》
清华大学出版社 ISBN:978-7-302-55977-1 定价：69元 2020年10月第2版

附录H：《Spark编程基础（Scala版）》



《Spark编程基础（Scala版）》

厦门大学 林子雨，赖永炫，陶继平 编著

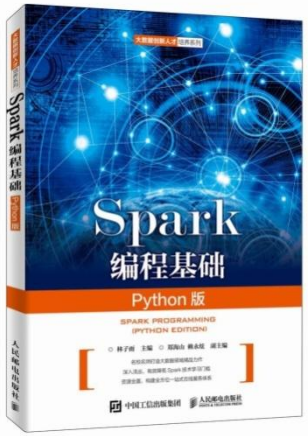
披荆斩棘，在大数据丛林中开辟学习捷径
填沟削坎，为快速学习Spark技术铺平道路
深入浅出，有效降低Spark技术学习门槛
资源全面，构建全方位一站式在线服务体系

人民邮电出版社出版发行，ISBN:978-7-115-48816-9

教材官网：http://dblab.xmu.edu.cn/post/spark/

本书以Scala作为开发Spark应用程序的编程语言，系统介绍了Spark编程的基础知识。全书共8章，内容包括大数据技术概述、Scala语言基础、Spark的设计与运行原理、Spark环境搭建和使用方法、RDD编程、Spark SQL、Spark Streaming、Spark MLlib等。本书每个章节都安排了入门级的编程实践操作，以便读者更好地学习和掌握Spark编程方法。本书官网免费提供了全套的在线教学资源，包括讲义PPT、习题、源代码、软件、数据集、授课视频、上机实验指南等。

附录I：《Spark编程基础（Python版）》



《Spark编程基础（Python版）》

厦门大学 林子雨，郑海山，赖永炫 编著

披荆斩棘，在大数据丛林中开辟学习捷径
填沟削坎，为快速学习Spark技术铺平道路
深入浅出，有效降低Spark技术学习门槛
资源全面，构建全方位一站式在线服务体系

人民邮电出版社出版发行，ISBN:978-7-115-52439-3

教材官网： http://dblab.xmu.edu.cn/post/spark-python/

本书以Python作为开发Spark应用程序的编程语言，系统介绍了Spark编程的基础知识。全书共8章，内容包括大数据技术概述、Spark的设计与运行原理、Spark环境搭建和使用方法、RDD编程、Spark SQL、Spark Streaming、Structured Streaming、Spark MLlib等。本书每个章节都安排了入门级的编程实践操作，以便读者更好地学习和掌握Spark编程方法。本书官网免费提供了全套的在线教学资源，包括讲义PPT、习题、源代码、软件、数据集、上机实验指南等。

附录J：高校大数据课程公共服务平台

高校大数据课程

公 共 服 务 平 台

http://dblab.xmu.edu.cn/post/bigdata-teaching-platform/

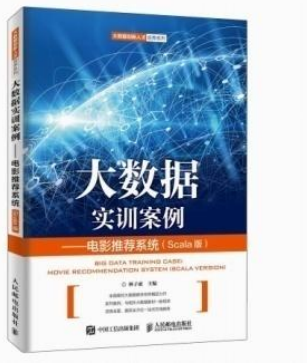
扫一扫访问平台主页

扫一扫观看3分钟FLASH动画宣传片

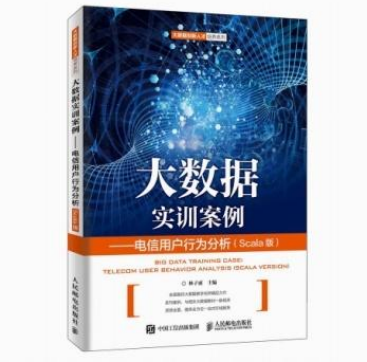
附录K：高校大数据实训课程系列案例教材

为了更好满足高校开设大数据实训课程的教材需求，厦门大学数据库实验室林子雨老师团队联合企业共同开发了《高校大数据实训课程系列案例》，目前已经完成开发的系列案例包括：

《电影推荐系统》（已经于2019年5月出版）
《电信用户行为分析》（已经于2019年5月出版）
《实时日志流处理分析》
《微博用户情感分析》
《互联网广告预测分析》
《网站日志处理分析》



系列案例教材将于2019年陆续出版发行，教材相关信息，敬请关注网页后续更新！
http://dblab.xmu.edu.cn/post/shixunkecheng/



扫一扫访问大数据实训课程系列案例教材主页


Thank You!