第3章 数据采集

学习目标

- 了解数据采集概述，能够说出数据采集的方法；
- 熟悉 Flume 的部署，能够完成 Flume 的安装和配置；
- 掌握采集招聘信息的操作，能够根据需求制定采集方案。

本项目的需求是对大数据职位的招聘信息进行分析。这些信息需要通过系统化的数据采集方式获取，因此，数据采集成为推动本项目实施的重要一环。本章将详细介绍如何通过数据采集来获取招聘信息。

3.1 数据采集概述

在当今信息时代，数据已经成为企业决策、科学研究、市场分析等方面的重要资源。为了有效地获取并充分利用这一宝贵资源，数据采集成为一项至关重要的任务。数据采集是一个从多种来源搜集信息的过程，包括传感器、网络、数据库、日志等。其目标是收集有关特定主题的信息，以支持决策制定和研究分析等活动。

数据采集方法可以分为手动采集和自动采集，具体介绍如下。

1. 手动采集

手动采集是最传统的数据采集方式，通常由专业人员通过键盘输入、扫描等方式将数据录入计算机。这种方法的优点是准确性高，但效率较低，不适合大规模的数据采集。在实际操作中，手动采集通常用于收集一些结构化的、需要人工审核的数据，如调查问卷、财务报表。

2. 自动采集

自动采集通过技术手段自动从各种来源获取数据。常见的自动采集方式包括网络爬虫、传感器采集、调用 API、数据库查询和日志采集，具体如下。

（1）网络爬虫。网络爬虫是一种自动采集网页信息的程序。通过网络爬虫可以在互联网上抓取大量文本、图片等信息。实际应用中，网络爬虫可用于采集新闻、社交媒体和电商网站等多种类型的数据。例如，新闻机构可使用网络爬虫自动抓取互联网上的新闻文章，以便快速更新新闻库；电商公司可使用网络爬虫自动抓取竞争对手的产品信息和价格，以便进行市场分析和定价策略制定。

（2）传感器采集。传感器是一种能感知被测量信息并将其转换成可用信号的设备。通过传感器技术，我们能实时采集环境参数和设备状态等信息。在实际应用中，传感器采集可用于环境监测、工业生产和智能家居等领域。例如，环境监测站可使用传感器采集空气质量、水质等环境参数，以便实时监测环境变化；工厂可使用传感器采集生产线上的设备状态和生产数据，以便优化生产过程和提高生产效率。

（3）调用 API。API（应用程序接口）是一种允许在不同软件之间进行交互的接口。通过调用 API，我们能方便地从其他系统中获取数据。实际应用中，调用 API 可用于采集企业内部数据、第三方服务数据等。例如，企业可使用调用 API 自动获取员工信息、项目进度等内部数据，以便进行人力资源管理和工作协同；开发者可使用调用 API 自动获取天气信息、地图数据等外部数据，以便开发更丰富的应用功能。

（4）数据库查询。数据库查询是一种直接从数据库中获取数据的方式。通过 SQL 或其他查询语言，我们能快速从数据库中获取所需的数据。例如，财务人员可使用数据库查询快速查找和分析财务数据，以便进行财务报告和决策支持；销售人员可使用数据库查询快速查找和分析客户数据，以便进行客户关系管理和销售策略制定。

（5）日志采集。日志采集是一种从系统的日志文件中获取数据的方式。通过解析和采集这些日志文件，可以获得关于系统性能、错误和用户行为等方面的有用信息。例如，通过日志采集实时监控服务器和网络设备的运行状态，以便及时发现和解决系统故障；通过日志采集分析用户行为数据，以便优化产品功能和提高用户体验。

相比手动采集，自动采集更适用于大规模数据采集，但实现自动采集通常需要借助数据采集工具。这些工具能够帮助我们更高效、更准确地从各种数据源中获取数据。例如，网络爬虫工具 Scrapy 可从网站上抓取文本、图片等信息。传感器采集工具 Arduino 可实时采集环境参数、设备状态等信息。在本项目中，我们主要使用 Hadoop 生态系统中的日志采集工具 Flume 从本地文件系统采集招聘信息。

3.2　部署 Flume

Flume 是一个高可用的、高可靠的日志采集工具，它能够将不同数据源的海量日志进行高效采集、汇总和移动，最终将这些日志存储到指定的存储系统，如 HDFS、HBase 等。Flume 在实际应用过程中，不仅仅用于日志的采集，由于 Flume 采集的数据源是可定制的（所谓数据源可定制是指用户可以根据实际应用场景指定 Flume 采集的数据），所以 Flume 还可以用于传输大量的网络流量数据、社交媒体生成的数据和电子邮件等。

Flume 的核心是通过数据采集器（Source）采集数据源（如 Web Server）的数据，再把采集到的数据通过缓冲通道（Channel）汇总到指定接收器（Sink），接收器将数据存储到指定的存储设备，如 HDFS。接下来，通过图 3-1 展示 Flume 的基本结构。

在图 3-1 中，Agent（代理）实际是一个 JVM 进程，该进程运行着 Flume 的三个核心组件：Source、Channel 和 Sink，具体介绍如下。

- Source：用于采集数据源的数据，并将数据写入 Channel。一个 Source 可以连接一个或多个 Channel。

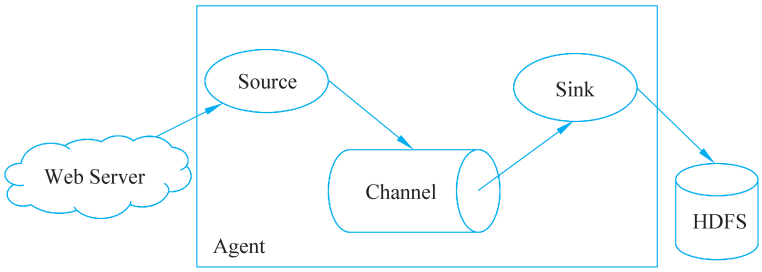

图 3-1 Flume 的基本结构

- Channel：用于缓存 Source 写入的数据，并将数据写入 Sink，待 Sink 将数据写入存储设备或者下一个 Source 之后，Flume 会删除 Channel 中缓存的数据。
- Sink：用于接收 Channel 写入的数据，并将数据写入存储设备。

在整个数据传输过程中，Flume 将流动的数据封装到一个事件(Event)中，它是 Flume 内部传输数据的基本单元。一个完整的事件包含 headers 和 body 两部分，其中 headers 用于存放数据的属性，属性以键值对的形式存在。body 用于存放数据信息，也就是实际采集到的数据。

接下来，将演示如何通过虚拟机 Hadoop3 部署 Flume，具体操作步骤如下。

1. 上传 Flume 安装包

将 Flume 安装包 apache-flume-1.9.0-bin.tar.gz 上传至虚拟机的/export/software 目录中。

2. 安装 Flume

采用解压缩方式将 Flume 安装至/export/servers 目录。在虚拟机上执行如下命令。

```
$tar -zxvf /export/software/apache-flume-1.9.0-bin.tar.gz \
-C /export/servers/
```

上述命令执行完成后，在虚拟机的/export/servers 目录会看到一个名称为 apache-flume-1.9.0-bin 的文件夹。为了便于后续使用 Flume，这里将 Flume 的安装目录重命名为 flume-1.9.0，在虚拟机的/export/servers 目录中执行如下命令。

```
$mv apache-flume-1.9.0-bin flume-1.9.0
```

3. 同步依赖文件

Flume 和 Hadoop 都依赖于 Guava 库来实现一些功能。然而，在 Flume 1.9.0 和 Hadoop 3.3.0 中，这二者使用的 Guava 库版本不一致，即 Flume 1.9.0 使用的 Guava 库版本为 11.0.2，而 Hadoop 3.3.0 使用的 Guava 库版本为 27.0。为了确保 Flume 能够将采集的数据成功存储到 HDFS，我们需要统一 Flume 和 Hadoop 使用 Guava 库的版本。鉴于向下兼容性的考虑，这里用 Hadoop 3.3.0 所使用的 Guava 库来替换 Flume 1.9.0 中的 Guava 库，具体操作步骤如下。

首先，进入 Hadoop 存放 Guava 库的目录，具体命令如下。

```
$cd /export/servers/hadoop-3.3.0/share/hadoop/common/lib
```

然后，将 Hadoop 使用的 Guava 库 guava-27.0-jre.jar 复制到 Flume 安装目录的 lib 目录中，具体命令如下。

```
$cp guava-27.0-jre.jar /export/servers/flume-1.9.0/lib
```

最后，删除 Flume 自带的 Guava 库 guava-11.0.2.jar，具体命令如下。

```
$rm -fr /export/servers/flume-1.9.0/lib/guava-11.0.2.jar
```

4. 创建配置文件 flume-env.sh

Flume 运行时使用配置文件 flume-env.sh 来定义环境变量。然而，Flume 默认没有提供给用户可编辑的配置文件 flume-env.sh，而是提供了一个供用户参考的模板文件 flume-env.sh.template。为了创建配置文件 flume-env.sh，用户可以复制该模板文件并将其重命名为 flume-env.sh。

在虚拟机中，进入存放 Flume 配置文件的目录/export/servers/flume-1.9.0/conf，复制该目录中的模板文件 flume-env.sh.template 并将其重命名为 flume-env.sh，具体命令如下。

```
$cp flume-env.sh.template flume-env.sh
```

上述命令执行完成后，将会在/export/servers/flume-1.9.0/conf 目录中生成配置文件 flume-env.sh。

5. 修改配置文件 flume-env.sh

为了确保 Flume 在运行时能够准确地识别 JDK，我们需要使用 vi 编辑器来编辑配置文件 flume-env.sh，并在该文件中配置 JDK 的安装目录。在配置文件 flume-env.sh 的末尾添加以下内容。

```
export JAVA_HOME=/export/servers/jdk1.8.0_333
```

在配置文件 flume-env.sh 中添加上述内容后，保存并退出编辑。

6. 配置 Flume 系统环境变量

为了方便后续使用 Flume，我们需要在虚拟机中配置 Flume 的系统环境变量。在虚拟机的/etc 目录下，使用 vi 编辑器编辑系统环境变量文件 profile，并在文件的末尾添加如下内容。

```
export FLUME_HOME=/export/servers/flume-1.9.0
export PATH=$FLUME_HOME/bin:$PATH
```

在系统环境变量文件 profile 中添加上述内容后，保存并退出编辑。然后，初始化虚拟机的系统环境变量，使系统环境变量文件 profile 中修改的内容生效。

7. 验证 Flume 是否部署成功

通过 Flume 提供的命令行工具 flume-ng 查看虚拟机中 Flume 的版本号,具体命令如下。

```
$flume-ng version
```

上述命令执行完成的效果如图 3-2 所示。

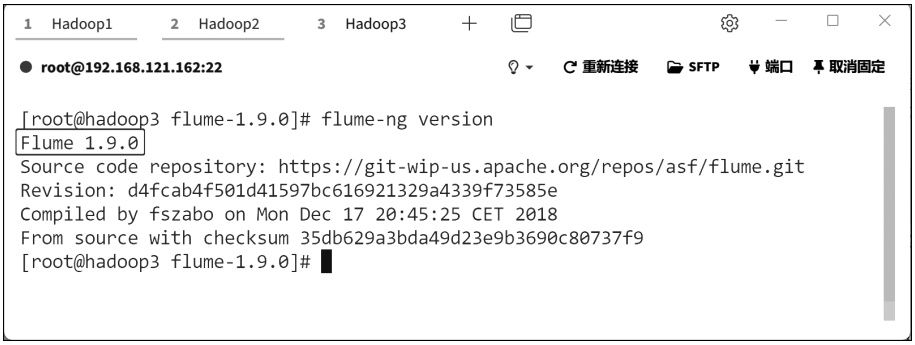

图 3-2 查看虚拟机中 Flume 的版本号

在图 3-2 中显示了 Flume 的版本号为 1.9.0,说明在虚拟机中成功部署了 Flume。

3.3 采集招聘信息

本项目所采集的招聘信息存储在虚拟机 Hadoop3 的/export/data/offer_job/20231208 目录中,这些招聘信息通过网络爬虫程序获取,涵盖了某招聘网站在 2023 年 12 月 8 日发布的全国范围内与大数据相关的职位。我们需要使用 Flume 从/export/data/offer_job/20231208 目录采集数据,并将其写入 HDFS 的/job/origin/2023-12-08 目录,具体操作步骤如下。

1. 指定采集方案

使用 Flume 采集招聘信息之前,需要先创建一个配置文件来指定采集方案。针对本项目中采集招聘信息的需求,在配置文件中进行如下配置。

- 指定 Source 通过虚拟机 Hadoop3 的/export/data/offer_job/20231208 目录采集数据。
- 指定 Channel 通过内存缓存事件。
- 指定 Sink 将事件写入 HDFS 的/job/origin/2023-12-08 目录。

在虚拟机 Hadoop3 创建/export/data/flume_conf 目录,用于存放采集方案的配置文件,具体命令如下。

```
$mkdir /export/data/flume_conf
```

上述命令执行完成后,进入虚拟机 Hadoop3 的/export/data/flume_conf 目录。使用

vi 编辑器编辑配置文件 job-hdfs.conf，在该文件中添加如下内容。

```
#指定 Source 的唯一标识 r1
a1.sources =r1
#指定 Channel 的唯一标识 c1
a1.channels =c1
#指定 Sink 的唯一标识 k1
a1.sinks =k1
#指定 Source 的类型为 Taildir Source，表示通过监控指定目录中文件的变化采集数据
a1.sources.r1.type =TAILDIR
#指定文件分组为 f1
a1.sources.r1.filegroups =f1
#指定监控/export/data/offer_job/20231208 目录中文件的变化
a1.sources.r1.filegroups.f1 =/export/data/offer_job/20231208/.*
#指定在/export/data/flume 目录的 job_position.json 文件中记录文件的变化
a1.sources.r1.positionFile =/export/data/flume/job_position.json
#指定拦截器的唯一标识 i1
a1.sources.r1.interceptors =i1
#指定拦截器的类型为 Timestamp Interceptor，表示在事件的 headers 中添加时间戳
a1.sources.r1.interceptors.i1.type =timestamp
#指定 Channel 的类型为 Memory Channel，表示通过内存缓存事件
a1.channels.c1.type =memory
#指定 Channel 缓存事件的最大容量为 1000
a1.channels.c1.capacity =1000
#指定 Channel 中每个事务能够处理事件数量的上限为 100
a1.channels.c1.transactionCapacity =100
#指定 Sink 的类型为 HDFS Sink，表示将事件写入 HDFS 的文件中
a1.sinks.k1.type =hdfs
#指定 HDFS 的目录为/job/origin/2023-12-08
a1.sinks.k1.hdfs.path =/job/origin/2023-12-08
#指定文件名称的前缀为 job-
a1.sinks.k1.hdfs.filePrefix =job-
#指定启用轮询模式将事件批量写入 HDFS
a1.sinks.k1.hdfs.round =true
#指定生成新文件的时间间隔为 10 秒
a1.sinks.k1.hdfs.rollInterval =10
#指定当累积的事件大小达到 128MB 时生成新文件
a1.sinks.k1.hdfs.rollSize =134217728
#指定新文件的生成与事件数量无关
a1.sinks.k1.hdfs.rollCount =0
#指定文件类型为 DataStream
a1.sinks.k1.hdfs.fileType =DataStream
#将 Channel 与 Source 关联
```

```
42  a1.sources.r1.channels =c1
43  #将 Channel 与 Sink 关联
44  a1.sinks.k1.channel=c1
```

上述代码指定 Agent 的唯一标识为 a1,其中第 34 行和第 36 行代码用于定义生成新文件的条件,只要满足其中任意条件,HDFS Sink 便会在 HDFS 的/job/origin/2023-12-08 目录下生成一个以 job-为前缀的新文件。

在配置文件 job-hdfs.conf 中添加完上述内容后,保存并退出编辑。

2. 启动 Agent

在虚拟机 Hadoop3 中,通过 Flume 提供的命令行工具 flume-ng 启动 Agent,具体命令如下。

```
$flume-ng agent --name a1 --conf conf/ --conf-file \
/export/data/flume_conf/job-hdfs.conf -Dflume.root.logger=INFO,console
```

上述命令中,参数--name 用于指定 Agent 的名称为 a1,该名称需要与配置文件 job-hdfs.conf 中 Agent 的唯一标识一致;参数--conf 用于指定 Flume 配置文件的路径,由于在虚拟机 Hadoop3 中配置了 Flume 的系统环境变量,所以这里默认会从 Flume 安装目录的 conf 目录中找到 Flume 配置文件;参数--conf-file 用于指定配置文件 job-hdfs.conf 的路径;参数-Dflume.root.logger 用于指定 Agent 的日志级别为 INFO,并将日志输出到控制台(console)。

上述命令执行完成的效果如图 3-3 所示。

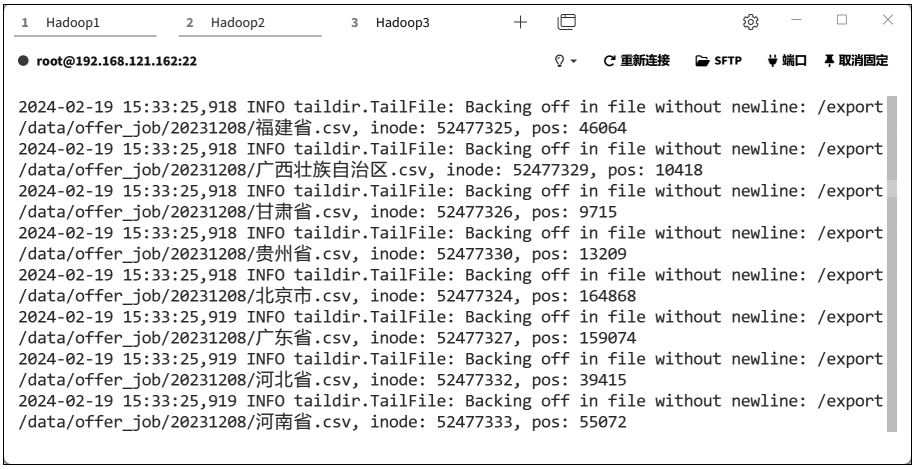

图 3-3 启动 Agent

从图 3-3 显示的日志信息可以看出,Agent 的 Source 开始从/export/data/offer_job/20231208 目录中的文件采集数据。

需要说明的是,启动 Agent 的窗口会一直保持占用状态。若关闭此窗口或按下组合键 Ctrl+C,则 Agent 将停止运行。

3. 查看采集的招聘信息

首先，通过检查 HDFS 的/job/origin/2023-12-08 目录，查看 Flume 将采集的招聘信息写入 HDFS 的文件。在虚拟机 Hadoop2 上执行如下命令。

```
$hdfs dfs -ls /job/origin/2023-12-08
```

上述命令执行完成的效果如图 3-4 所示。

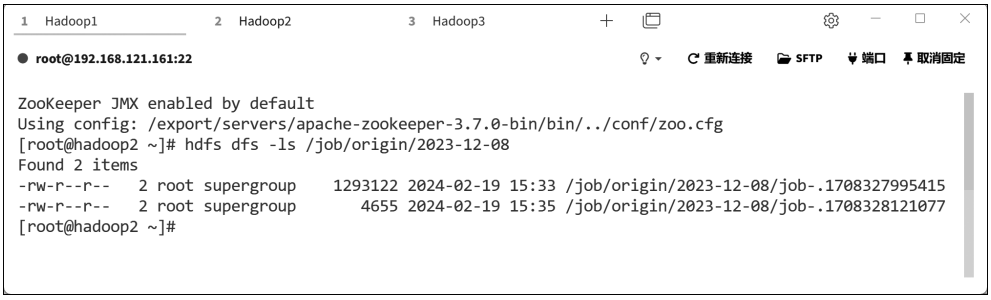

图 3-4　检查 HDFS 的/job/origin/2023-12-08 目录

从图 3-4 中可以看出，Flume 将采集的招聘信息写入 HDFS 的/job/origin/2023-12-08 目录下的文件 job-.1708327995415 和 job-.1708328121077 中。

然后，通过读取文件 job-.1708327995415 的前 10 行，查看招聘信息的内容。在虚拟机 Hadoop2 上执行如下命令。

```
$hdfs dfs -cat /job/origin/2023-12-08/job-.1708327995415 | head -n 10
```

上述命令执行完成的效果如图 3-5 所示。

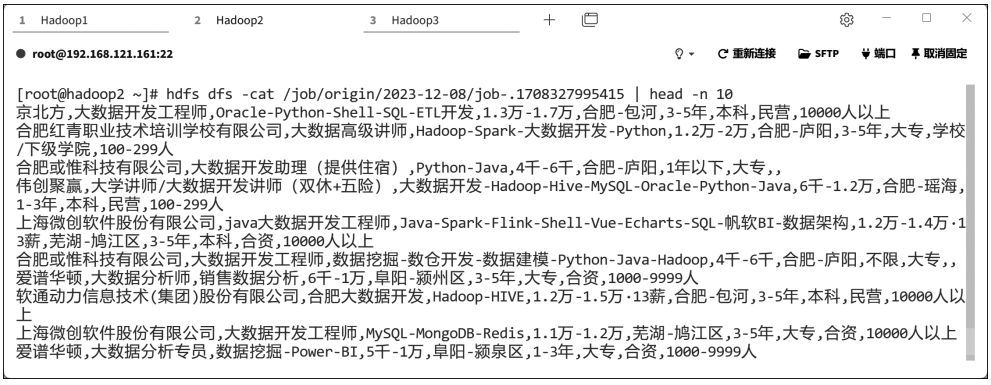

图 3-5　查看招聘信息的内容

从图 3-5 中可以看出，招聘信息包含职位名称、公司名称、公司地址、职位薪资等信息。

3.4 本章小结

本章主要讲解了数据采集的相关内容。首先,介绍了数据采集的概念。然后,介绍了部署 Flume 的相关内容。最后,介绍了采集招聘信息的相关内容。通过本章的学习,读者可以熟悉如何使用 Flume 采集数据。