Data Ingestion Tools 实验

211250109 赵政杰

2023 年 9 月 27 日

本次实验在实验二所创建的云实例上进行。

1 任务 1：使用 Apache Kafka 进行数据流

1.1 安装 Kafka

在实验 2 中已经在服务器上完成安装 Java.

使用`wget https://mirrors.aliyun.com/apache/kafka/3.5.1/kafka_2.12-3.5.1.tgz`获取压缩包并解压。

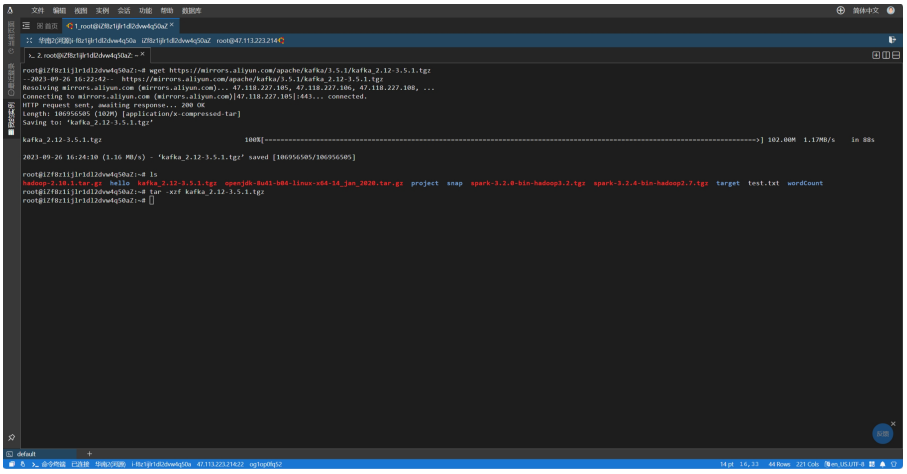

图 1: 安装 Kafka

1.2 启动 Kafka 服务

- 在 Kafka 目录中,启动 Zookeeper 服务器。命令为`bin/zookeeper-server-start.sh config/zookeeper.properties`

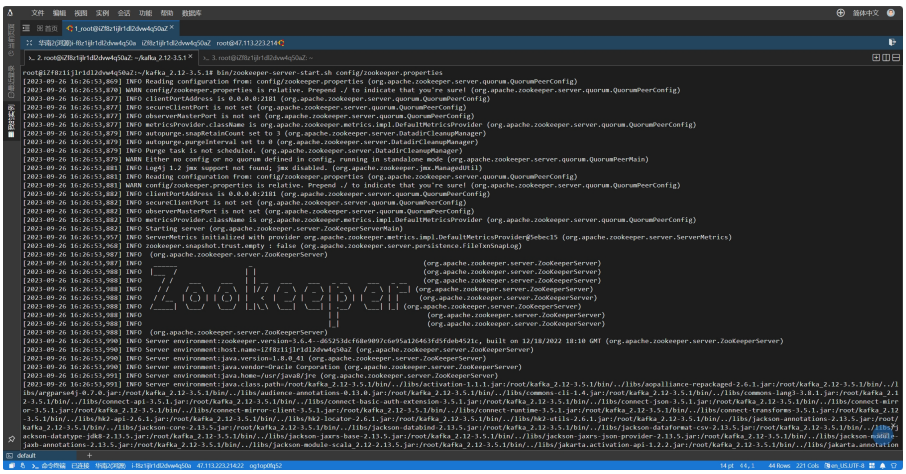

图 2: 启动 Zookeeper

- 在新的终端窗口中，启动 Kafka 服务器。命令为`bin/kafka-server-start.sh config/server.properties`

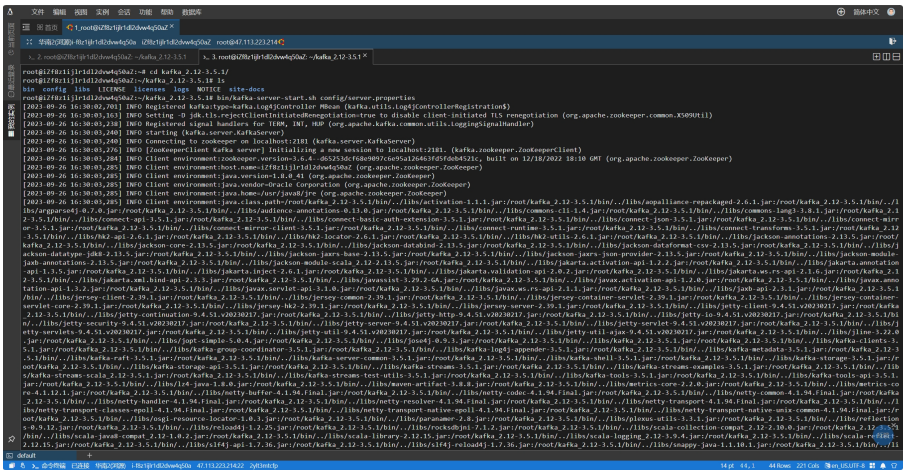

图 3: 启动 Kafka

1.3 创建 Kafka 主题

使用`bin/kafka-topics.sh --create --topic my-topic --bootstrap-server localhost:9092 --partitions 1 --replication-factor 1`

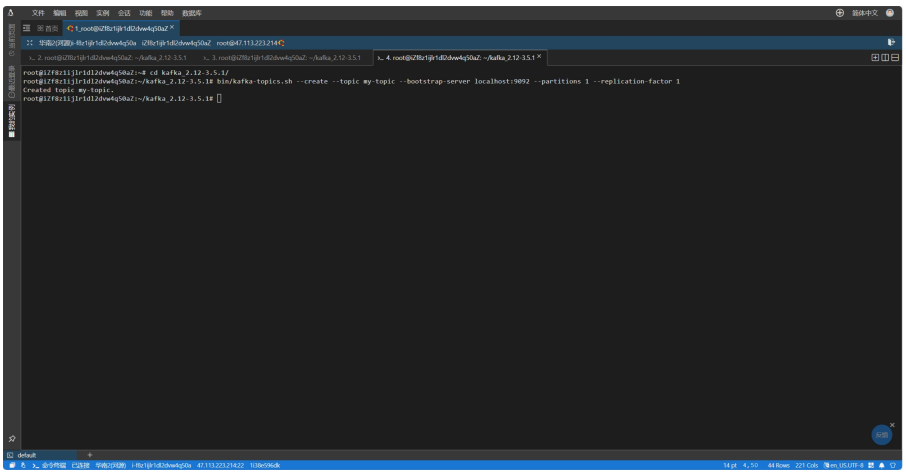

图 4: 创建 my-topic 主题

1.4 生产和消费消息

- 使用`bin/kafka-console-producer.sh --topic my-topic --bootstrap-server localhost:9092`创建生产者。

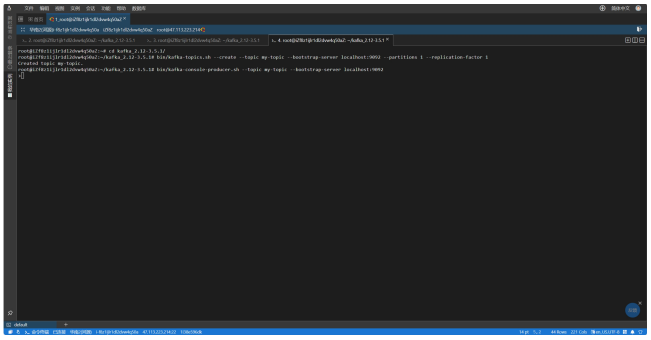

图 5: 创建生产者

- 使用`bin/kafka-console-consumer.sh --topic my-topic --bootstrap-server localhost:9092 --from-beginning`创建消费者。

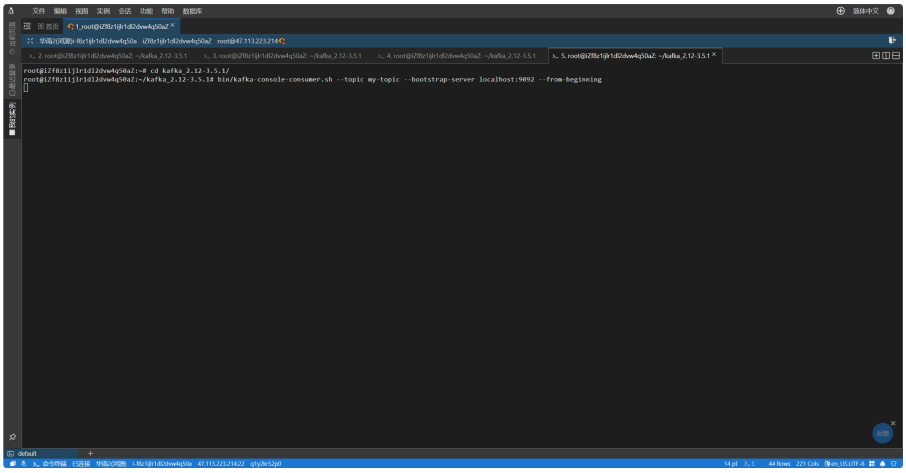

图 6: 创建消费者

1.5 确认 Kafka 主题中的消息

生产者生产消息，消费者消费消息

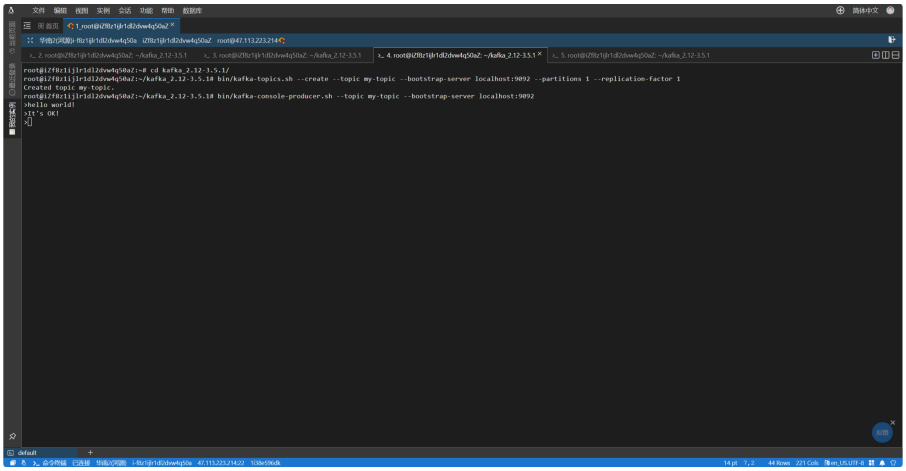

图 7: 生产消息

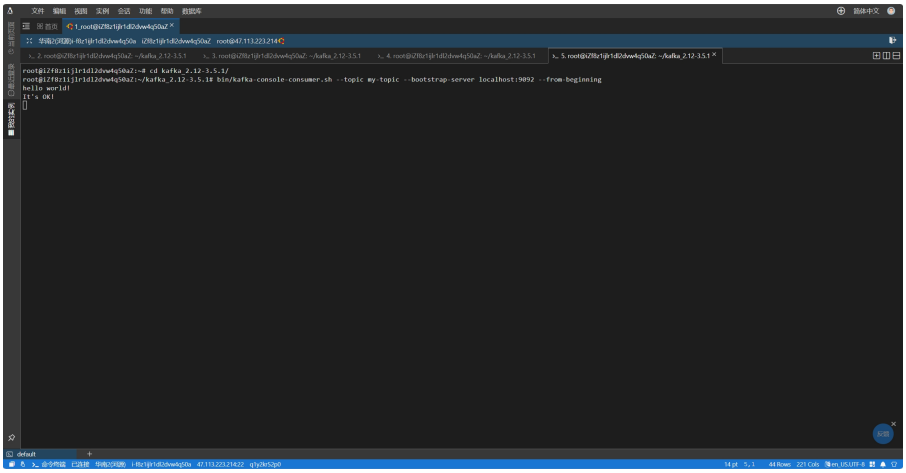

图 8: 消费消息

2 任务 4：使用 Flume 收集日志

1.1 安装 Flume

通过`wget https://mirrors.aliyun.com/apache/flume/1.11.0/apache-flume-1.11.0-bin.tar.gz`获取 Flume 并解压至 flume-1.11.0.

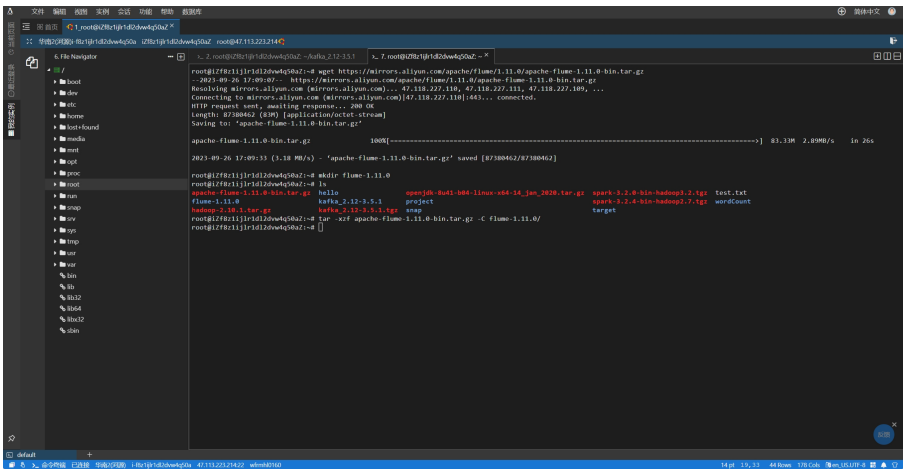

图 9: 安装 Flume

通过`bin/flume-ng version`查看是否安装成功。

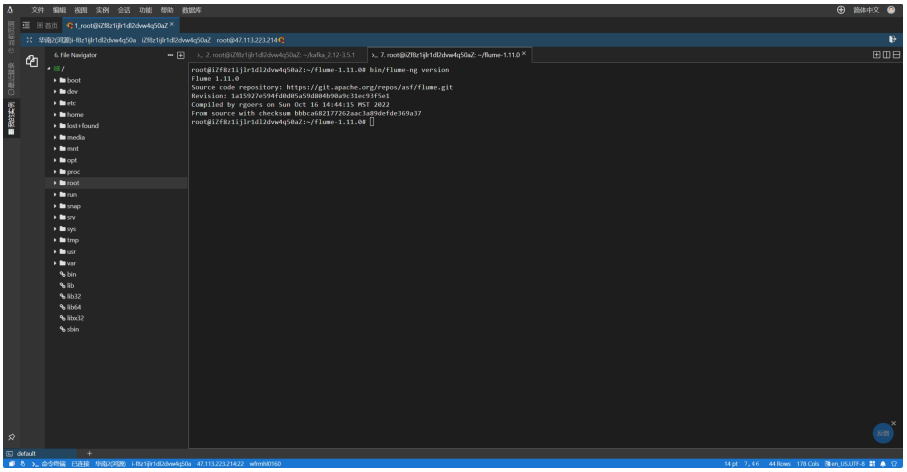

图 10: 查看 Flume 版本

1.2 启动 Hadoop

Hadoop 环境已经在实验 2 中配置完成。

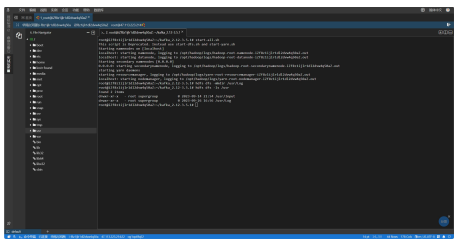

图 11: 启动 Hadoop

1.3 配置 Flume

配置文件的内容如下：

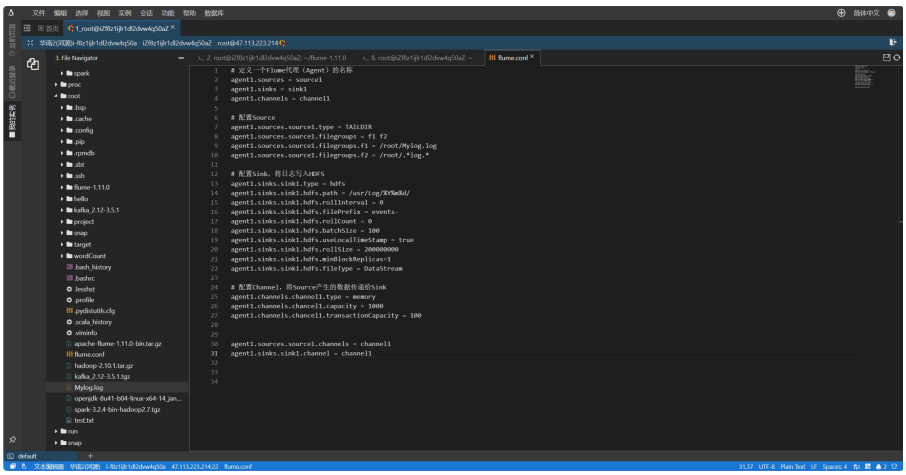

图 12: flume.conf

1.4 启动 Flume

完成配置后，可以启动 Flume 代理来开始日志收集和传输。使用以下命令启动 Flume：

```
bin/flume-ng agent --conf conf --conf-file /root/flume.conf -name agent1 -Dflume.root.logger=INFO,console
```

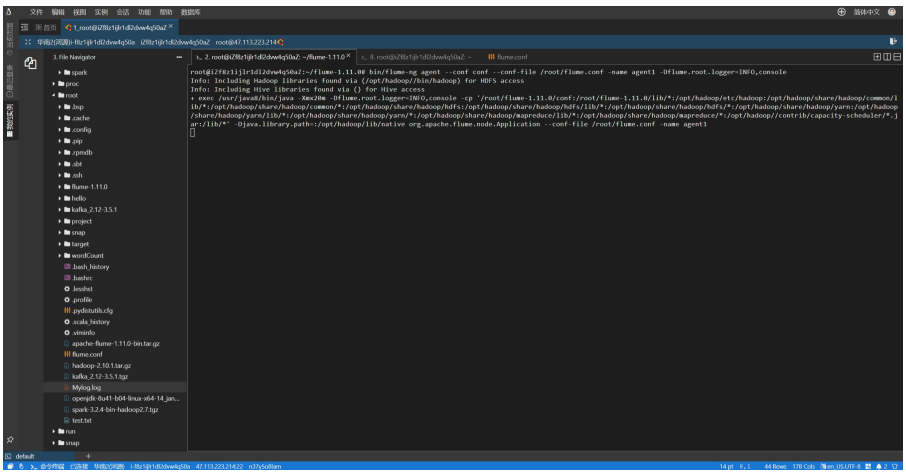

图 13: 启动 Flume

1.5 验证日志存储在 HDFS 中

- 在 root 目录下新建 Mylog.log 文件。

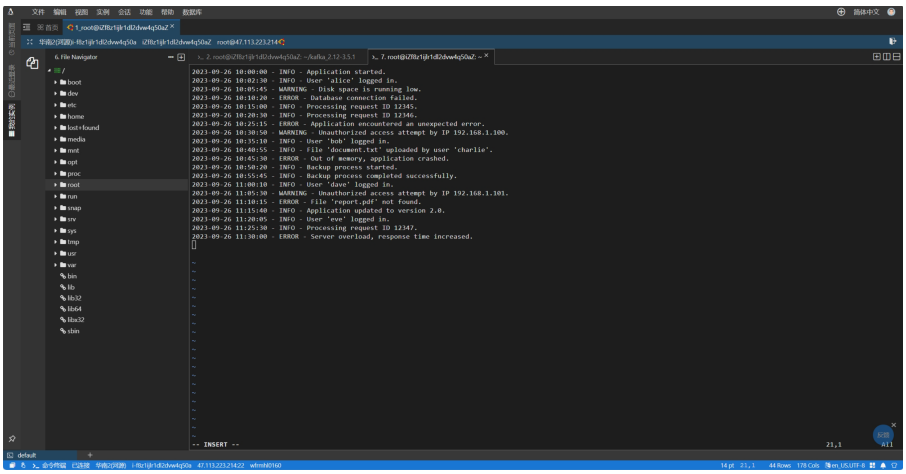

图 14: Mylog.log

- 在 Hadoop 的 web 界面查看，可以看到存储的日志。

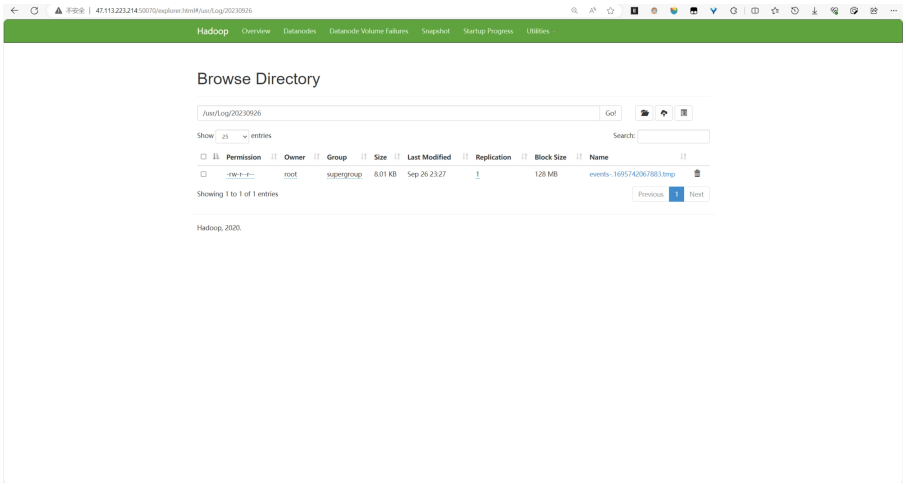

图 15: Hadoop

3 困难及克服方法

在做任务 4 时发现 Flume 一直无法把日志文件写入到 HDFS 里，经过不断查找发现不应该在 Flume 服务启动前就将.log 文件写入目标目录，这样会导致 Flume 没有监听到。

在启动 Flume 服务后再在目标目录中新建.log 文件并写入内容，就可以监听到了。