

Data Ingestion Tools 实验

211250109 赵政杰

2023 年 9 月 27 日

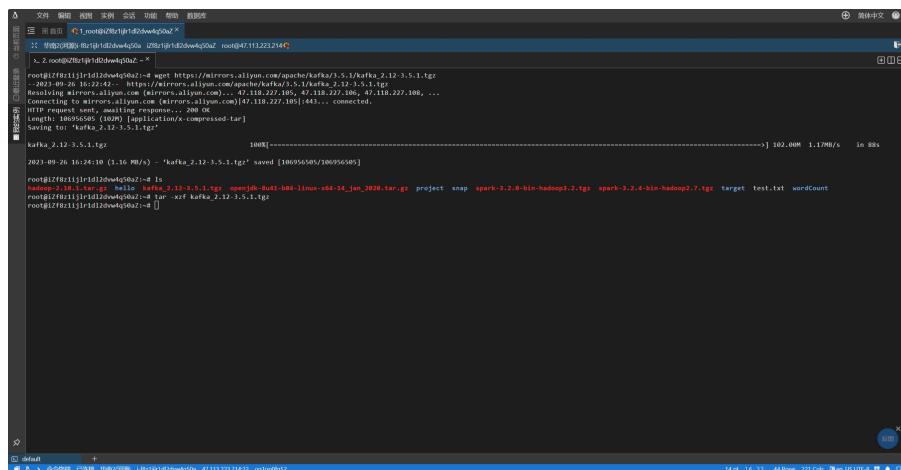
本次实验在实验二所创建的云实例上进行。

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。



```
root@iZb131j1r1d2wq6o2z:~# wget https://mirrors.aliyun.com/apache/kafka/3.5.1/kafka_2.12-3.5.1.tgz
--2023-09-26 16:24:10 (1.16 MB/s) - `kafka_2.12-3.5.1.tgz' saved [106956505/106956505]
root@iZb131j1r1d2wq6o2z:~# ls
kafka_2.12-3.5.1.tgz
root@iZb131j1r1d2wq6o2z:~# tar -xzf kafka_2.12-3.5.1.tgz
root@iZb131j1r1d2wq6o2z:~# cd kafka_2.12-3.5.1
root@iZb131j1r1d2wq6o2z:~/kafka_2.12-3.5.1# ./bin/zookeeper-server-start.sh config/zookeeper.properties
root@iZb131j1r1d2wq6o2z:~/kafka_2.12-3.5.1#
```

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

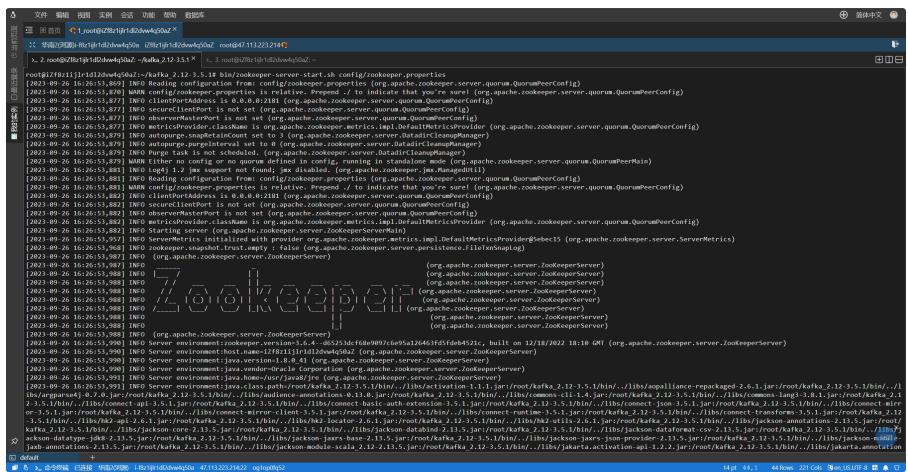


图 2: 启动 Zookeeper

- 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`

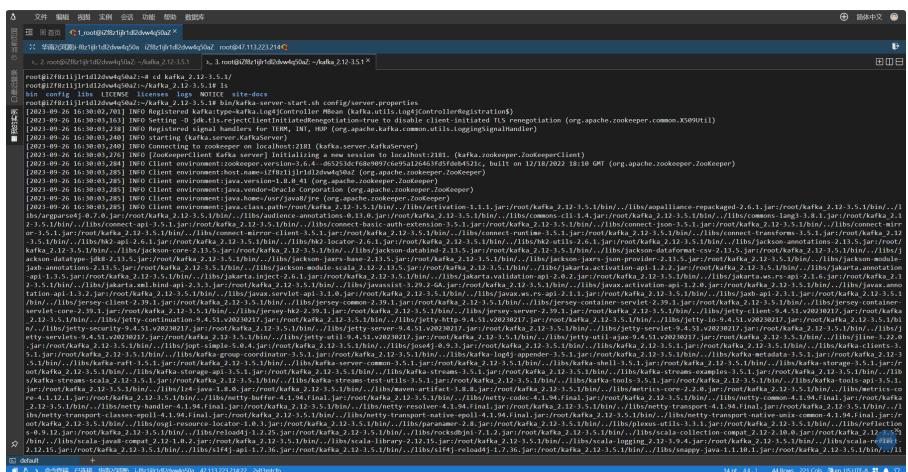


图 3: 启动 Kafka

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
```

```

root@iZbm11j1rid12dwq50z:~# bin/kafka-topic.sh --create --topic my-topic --bootstrap-server localhost:9092 --partitions 1 --replication-factor 1
Created topic: my-topic
root@iZbm11j1rid12dwq50z:~/kafka_2.12-3.5.1

```

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。

```

root@iZbm11j1rid12dwq50z:~# bin/kafka-console-producer.sh --topic my-topic --bootstrap-server localhost:9092

```

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。

```

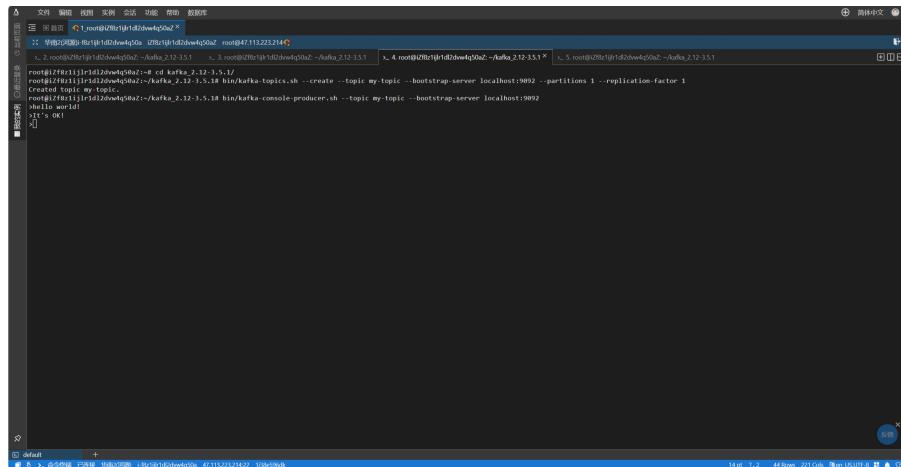
root@iZbm11j1rid12dwq50z:~# bin/kafka-console-consumer.sh --topic my-topic --bootstrap-server localhost:9092 --from-beginning

```

图 6: 创建消费者

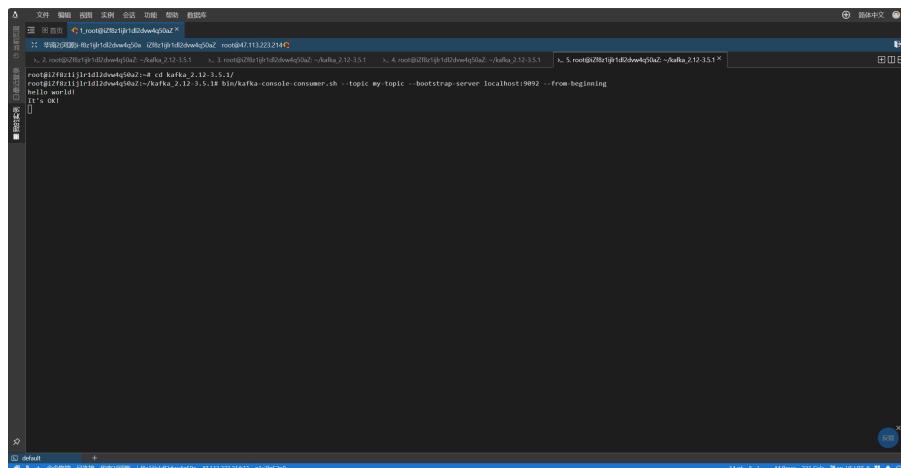
1.5 确认 Kafka 主题中的消息

生产者生产消息，消费者消费消息



```
root@iZ2fr1j1rld2dw4q5oZ:~# cd kafka_2.12-3.5.1/
root@iZ2fr1j1rld2dw4q5oZ:~/kafka_2.12-3.5.1> bin/kafka-topics.sh --create --topic my_topic --bootstrap-server localhost:9092 --partitions 1 --replication-factor 1
root@iZ2fr1j1rld2dw4q5oZ:~/kafka_2.12-3.5.1> bin/kafka-console-producer.sh --topic my_topic --bootstrap-server localhost:9092
Hello World!
It's OK!
```

图 7: 生产消息



```
root@iZ2fr1j1rld2dw4q5oZ:~# cd kafka_2.12-3.5.1/
root@iZ2fr1j1rld2dw4q5oZ:~/kafka_2.12-3.5.1> bin/kafka-console-consumer.sh --topic my_topic --bootstrap-server localhost:9092 --from-beginning
Hello World!
It's OK!
```

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.

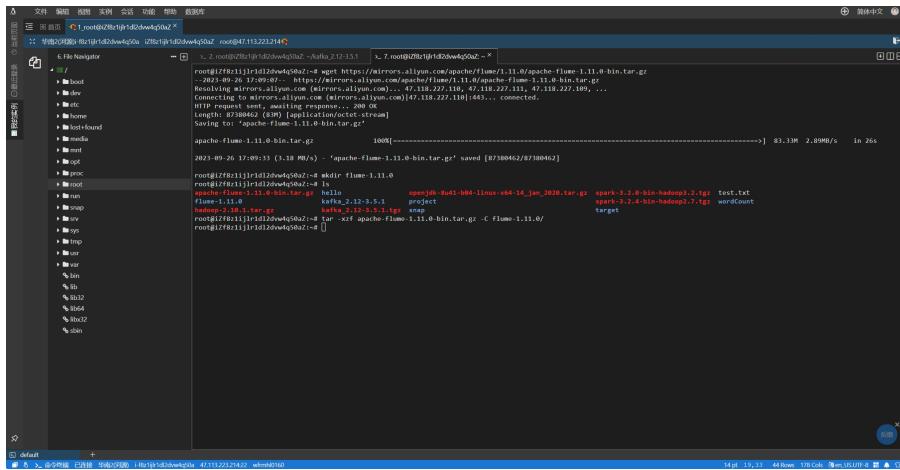


图 9: 安装 Flume

通过bin/flume-ng version查看是否安装成功。

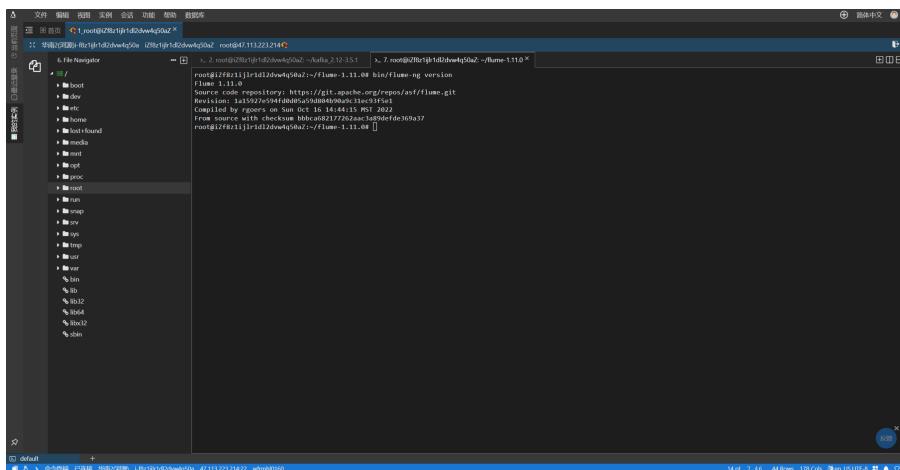


图 10: 查看 Flume 版本

1.2 启动 Hadoop

Hadoop 环境已经在实验 2 中配置完成。

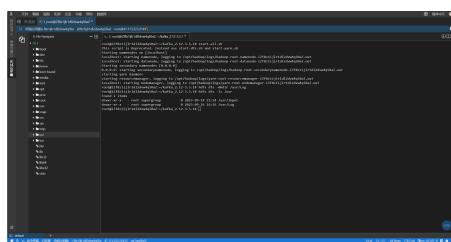


图 11: 启动 Hadoop

1.3 配置 Flume

配置文件的内容如下：

```
文件 编辑 调试 视频 窗口 功能 帮助 数据流
```

Locally(Zulu8j1jdzbwqfj1j) d:\dzw\log\flume-1.11.0 -> root@Zulu8j1jdzbwqfj1j:~%

flume-1.11.0

flume.conf

```
file://Navigator
  +-- spark
  +-- proc
  +-- agent
  +-- Zookeeper
  +-- Zookeeper
  +-- cache
  +-- config
  +-- app
  +-- memfd
  +-- log
  +-- sink
  +-- file
  +-- flume-1.11.0
  +-- hello
  +-- kafka-2.12-3.5.1
  +-- log4j
  +-- sleep
  +-- target
  +-- wordCount
  +-- JobHistory
  +-- hadoop
  +-- local
  +-- profile
  +-- pyflinklib.cfg
  +-- stale历史
  +-- apache-flume-1.11.0-bin.gz
  +-- flumeconf
  +-- hadoop-2.10.1.tar.gz
  +-- kafka-2.12-3.5.1.tgz
  +-- MySQL
  +-- openjdk-8u171-b11-linux-x64-14_jar-
  +-- spark-3.2.4-bin-hadoop3.7.tgz
  +-- test.txt
  +-- run
  +-- config
  +-- default
```

定义一个叫 Agent1 的名称
1 agent1.sources = source1
2 agent1.sinks = sink1
3 agent1.channels = channel1
4
5 # 配置Source
6 agent1.sources.source1.type =尾端
7 agent1.sources.source1.interceptors = fi f2
8 agent1.sources.source1.interceptors.f1 = /root/flog.log
9 agent1.sources.source1.interceptors.f2 = /root/*.log
10 agent1.sources.source1.interceptors.f3 = /root/*.log.
11
12 # 配置Sink, 将日志写入MongoDB
13 agent1.sinks.sink1.type = MongoDB
14 agent1.sinks.sink1.hostname = 192.168.1.100
15 agent1.sinks.sink1.mongodb.collection = flume
16 agent1.sinks.sink1.mongodb.rollInterval = 0
17 agent1.sinks.sink1.mongodb.batchSize = 100
18 agent1.sinks.sink1.mongodb.batchSize = 100
19 agent1.sinks.sink1.mongodb.useElasticSearchTimestamp = true
20 agent1.sinks.sink1.mongodb.timestampFormat = %Y-%m-%d %H:%M:%S
21 agent1.sinks.sink1.mongodb.minIngestReplicas = 1
22 agent1.sinks.sink1.mongodb.fileType = DataStream
23
24 # 配置Channel, 将Source产生的数据存到channel
25 agent1.channels.channel1.type = memory
26 agent1.channels.channel1.capacity = 1000
27 agent1.channels.channel1.transactionCapacity = 100
28
29
30 agent1.sources.source1.channels = channel1
31 agent1.sinks.sink1.channel = channel1
32
33
34

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
```

A screenshot of a Linux terminal window titled 'flume'. The command entered is 'flink run -c flume -m local[1] /opt/flume/conf/flume-conf.xml'. The output shows the job starting up, with logs indicating the configuration file path and the start of the Flink cluster. The terminal also displays the Java classpath and the main application class 'org.apache.flume.Application'. The background shows a file browser with various Hadoop and Flink configuration files.

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。

```

x. 2 root@192.168.1.100:~#ls -l /var/log
total 0
x. 2 root@192.168.1.100:~#tail -f /var/log/mylog.log
2023-09-20 10:00:00 INFO Application started.
2023-09-20 10:05:00 INFO User "root" logged in.
2023-09-20 10:05:45 WARNING Disk space is running low.
2023-09-20 10:10:20 ERROR Database connection failed.
2023-09-20 10:15:30 INFO Application updated to version 2.0.
2023-09-20 10:20:30 INFO Processing request ID 12345.
2023-09-20 10:25:10 INFO Application updated to version 2.1.
2023-09-20 10:30:50 WARNING Unauthorized access attempt by IP 192.168.1.100.
2023-09-20 10:35:10 INFO User "bob" logged in.
2023-09-20 10:40:30 INFO Application updated to version 2.2.
2023-09-20 10:45:30 ERROR Out of memory, application crashed.
2023-09-20 10:50:20 INFO Backup process started.
2023-09-20 10:55:10 INFO Backup process completed successfully.
2023-09-20 11:00:10 INFO User "dave" logged in.
2023-09-20 11:05:10 INFO Application updated to version 2.3.
2023-09-20 11:10:15 - ERROR - File "report.pdf" not found.
2023-09-20 11:15:40 INFO Application updated to version 2.4.
2023-09-20 11:20:10 INFO Application updated to version 2.5.
2023-09-20 11:25:30 INFO Processing request ID 12345.
2023-09-20 11:30:00 ERROR - Server overload, response time increased.
[]

x. 2 root@192.168.1.100:~#

```

图 14: Mylog.log

- 在 Hadoop 的 web 界面查看，可以看到存储的日志。

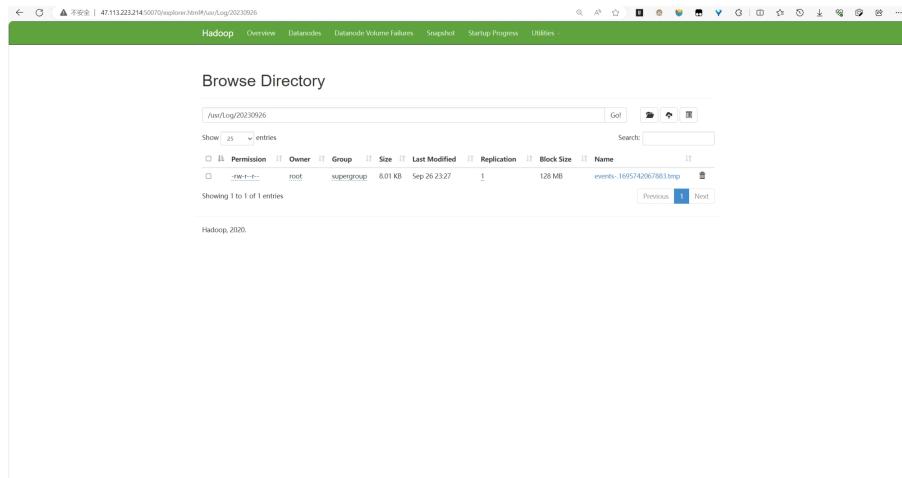


图 15: Hadoop

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。