实验报告：数据流工具

程智镝

2023 年 10 月 1 日

涵盖工具

本实验涵盖以下数据流工具：

1. Apache Kafka
2. AWS Kinesis
3. Apache NiFi
4. Flume

实验任务

1. 任务一：使用 Apache Kafka 进行数据流
    - **要求**：安装 Apache Kafka。
    - **任务**：生产和消费一个简单的消息。
    - **验证**：确认 Kafka 主题中的消息。
2. 任务二：使用 AWS Kinesis 进行实时数据摄取
    - **要求**：在 AWS 控制台中创建一个 Kinesis 流。
    - **任务**：使用 AWS SDK 发送一批消息。
    - **验证**：在 Kinesis 控制台中监控传入数据。
3. 任务三：使用 Apache NiFi 进行数据流管理

- **要求**：安装 Apache NiFi。
- **任务**：创建一个简单的数据流，将数据从平面文件移动到数据库。
- **验证**：确认数据库中的记录。

4. 任务四：使用 Flume 收集日志
   - **要求**：安装 Flume。
   - **任务**：配置 Flume 以收集日志并将其存储在 HDFS 中。
   - **验证**：确认 HDFS 中存储的日志。

实验难点

- 未使用过该配置
- 构造数据
- 我使用的是阿里云而不是 AWS，配置上可能会有差距

任务一

任务流程参照:https://blog.csdn.net/sun_hong_likeIT/article/details/123502688

下载 Apache Kafka:`sudo wget https://mirrors.tuna.tsinghua.edu.cn/apache/kafka/3.5.1/k`

这里使用的是清华镜像网站，比教程给的网站更好用。

使用 `bin/zookeeper-server-start.sh -daemon config/zookeeper.properties` 以守护进程启动)

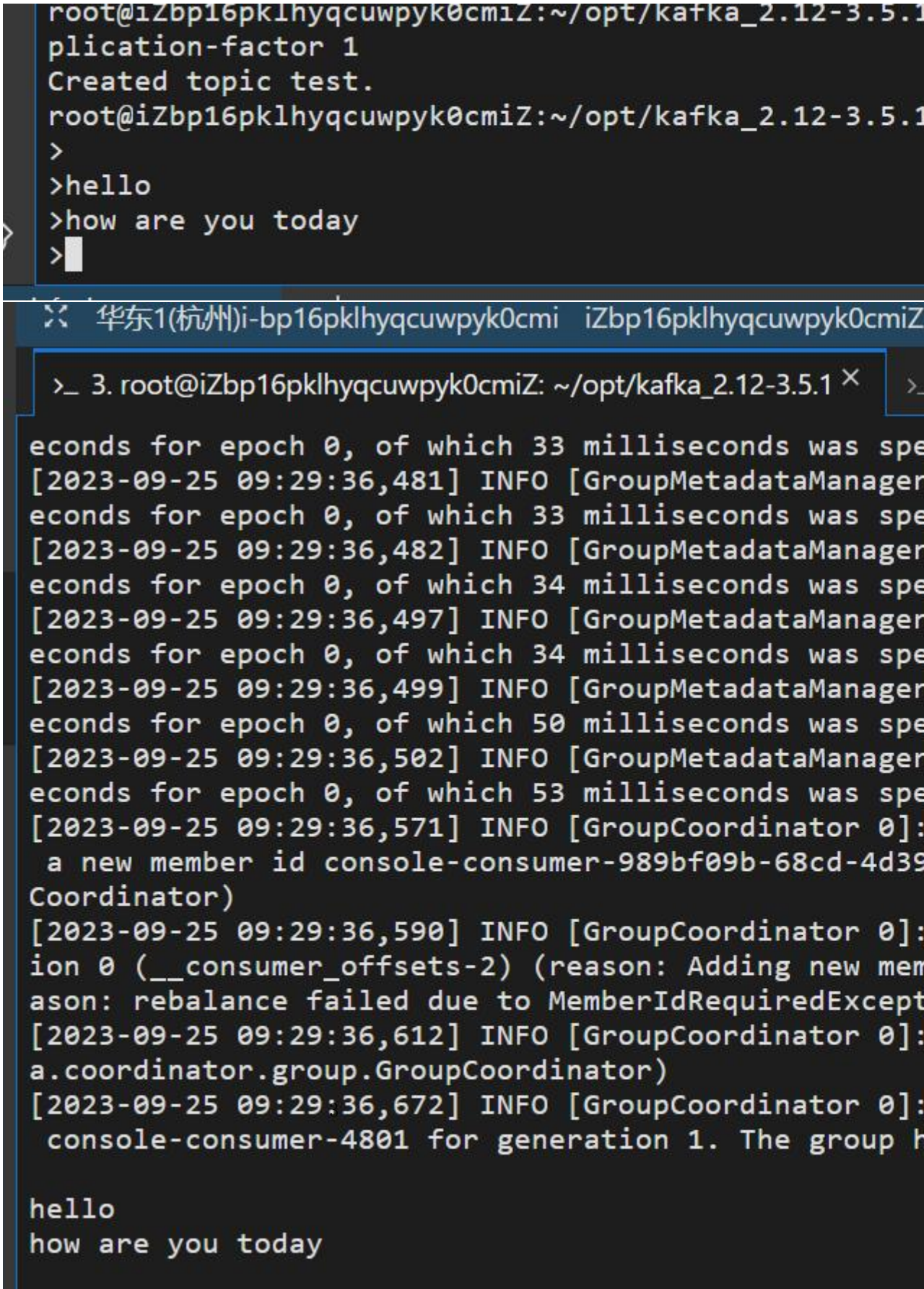
启动 Kafka:`bin/kafka-server-start.sh config/server.properties &`

新建终端创建主题:`bin/kafka-topics.sh --bootstrap-server localhost:9092 --create --topic test --partitions 2 --replication-factor 1`

在主题终端发送消息:`bin/kafka-console-producer.sh --broker-list localhost:9092 --topic test`

服务终端接收到消息:`bin/kafka-console-consumer.sh --bootstrap-server localhost:9092 --topic test --from-beginning`

消息生产和消费过程:


root@iZbp16pklhyqcuwpyk0cmiZ:~/opt/kafka_2.12-3.5.1
plication-factor 1
Created topic test.
root@iZbp16pklhyqcuwpyk0cmiZ:~/opt/kafka_2.12-3.5.1
>
>hello
>how are you today
>
华东1(杭州)i-bp16pklhyqcuwpyk0cmi iZbp16pklhyqcuwpyk0cmiZ
>_ 3. root@iZbp16pklhyqcuwpyk0cmiZ: ~/opt/kafka_2.12-3.5.1
econds for epoch 0, of which 33 milliseconds was spe
[2023-09-25 09:29:36,481] INFO [GroupMetadataManager
econds for epoch 0, of which 33 milliseconds was spe
[2023-09-25 09:29:36,482] INFO [GroupMetadataManager
econds for epoch 0, of which 34 milliseconds was spe
[2023-09-25 09:29:36,497] INFO [GroupMetadataManager
econds for epoch 0, of which 34 milliseconds was spe
[2023-09-25 09:29:36,499] INFO [GroupMetadataManager
econds for epoch 0, of which 50 milliseconds was spe
[2023-09-25 09:29:36,502] INFO [GroupMetadataManager
econds for epoch 0, of which 53 milliseconds was spe
[2023-09-25 09:29:36,571] INFO [GroupCoordinator 0]:
a new member id console-consumer-989bf09b-68cd-4d39
Coordinator)
[2023-09-25 09:29:36,590] INFO [GroupCoordinator 0]:
ion 0 (__consumer_offsets-2) (reason: Adding new mem
ason: rebalance failed due to MemberIdRequiredExcept
[2023-09-25 09:29:36,612] INFO [GroupCoordinator 0]:
a.coordinator.group.GroupCoordinator)
[2023-09-25 09:29:36,672] INFO [GroupCoordinator 0]:
console-consumer-4801 for generation 1. The group h
hello
how are you today

任务二

由于我使用的不是 AWS，我在阿里云上使用了数据总线来进行平替：https://www.aliyun.com/product/bigdata/datahub

任务三

步骤一：

NiFi 安装:wget https://mirrors.tuna.tsinghua.edu.cn/apache/nifi/1.23.2/nifi-1.23.2-bin.z

解压：unzip nifi-1.23.2.zip

提示：不知道为什么，网上大多数下载方法（使用镜像）我都无法成功下载，只能用官网的来下，真的巨慢……最后还是自己找了清华镜像自己 wget，别看网上教程了

用官网会很慢，建议使用国内镜像

步骤二：

配置 JAVA_HOME，sudo update-alternatives --config java 这个指令可以快速帮你找到 java 路径。export JAVA_HOME= 你的 Java 所在的路径，比如我的就是 /usr/lib/jvm/java-11-openjdk-amd64

NiFi，启动！./bin/nifi.sh start

任务四

参考:https://www.cnblogs.com/j-y-s/p/16018612.html 安装 Flume:wget https://mirrors.tuna.tsinghua.edu.cn/apache/flume/1.11.0/apache-flume-1.11.0-bin.tar.gz 解压 tar -zvxf 包名中途遇到了云实例不允许 root 使用密码登录，解决办法：VNC 关防火墙：(sudo) ufw disable 云实例没办法使用 jps 指令：修复 dpkg：sudo dpkg --configure -a 安装：sudo apt install openjdk-11-jdk-headless 途中遇到了云实例连接超时的问题，网络上所有的关于配置 ip 白名单的都试过了，都没用，到最后还是因为防火墙没关，登录 vnc，sudo ufw disable 即可