数据流工具实验报告

程智镝

2023 年 9 月 26 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

nifi 安装:wget https://dlcdn.apache.org/nifi/1.23.2/nifi-1.23.2-source-release.zip
解压：unzip nifi-1.23.2-source-release.zip 用官网会很慢，建议使用国内镜像

任务四

安装 Flume:https://downloads.apache.org/flume/1.11.0/apache-flume-1.11.0-bin.tar.gz