南京大學

NANJING UNIVERSITY

课程作业报告

大数据分析 03

学生姓名：左皓升

学生学号：211250074

2023 年 9 月

Data Ingestion Tools

使用华为云 ECS 服务器，实现本次作业中的任务一和任务四

1． 任务一：使用 Apache Kafka 进行数据流

1.1 下载最新版本的 kafka 并使用 scp 命令上传到云服务器

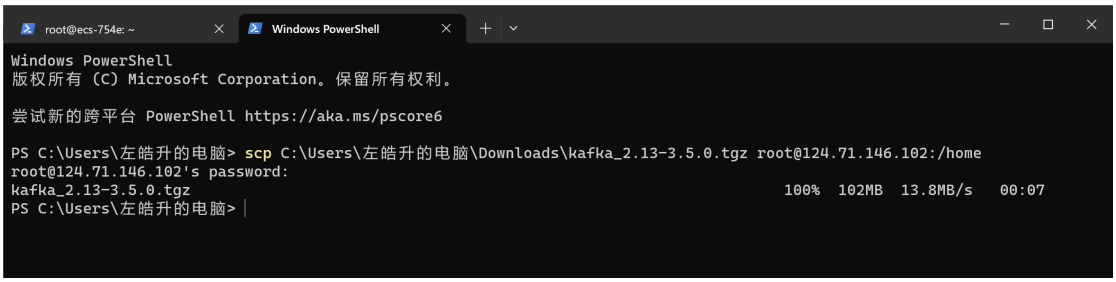


图 1: 下载上传

1.2 SSH 连接服务器，解压上传的 tgz 并进入解压后的文件夹

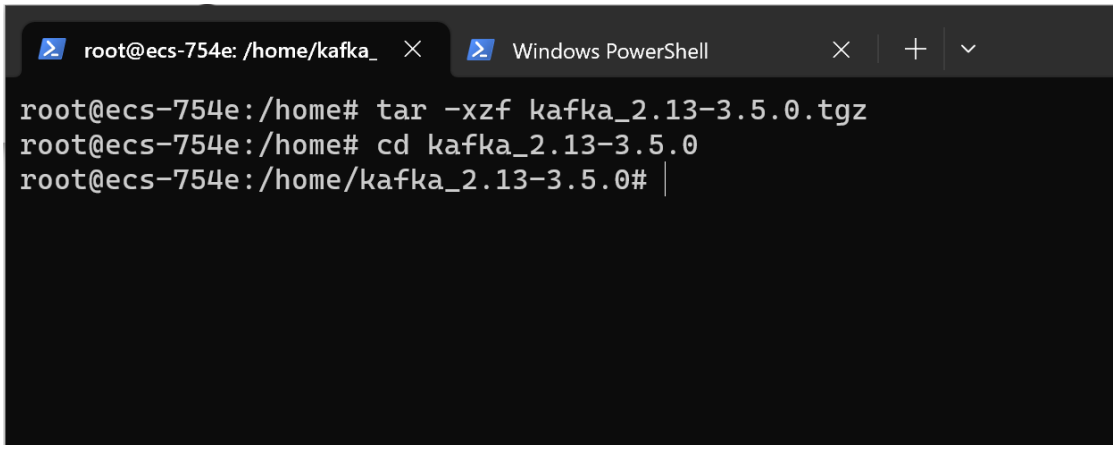
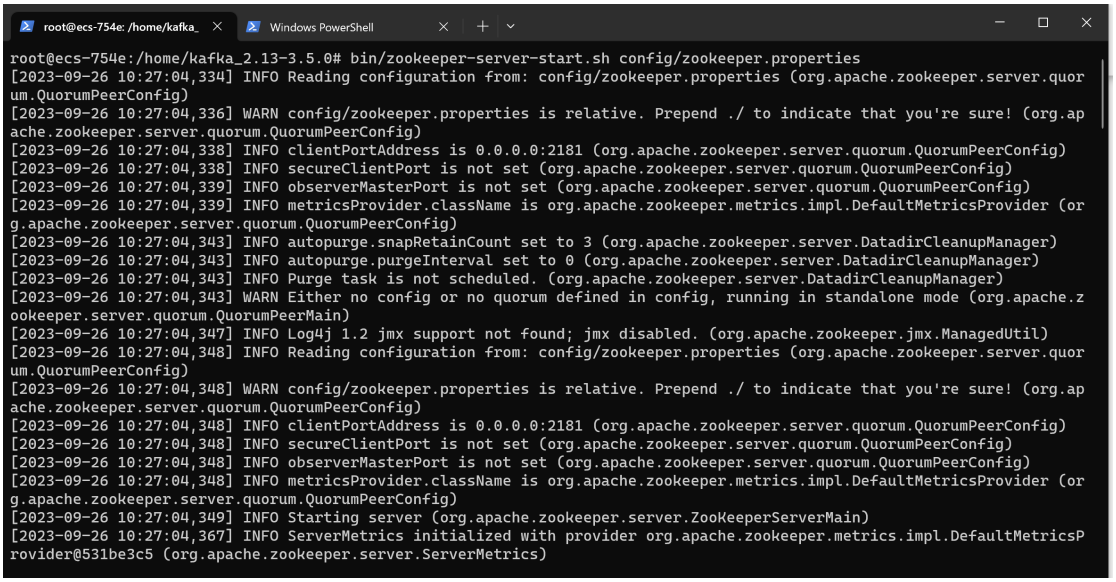


图 2: 解压

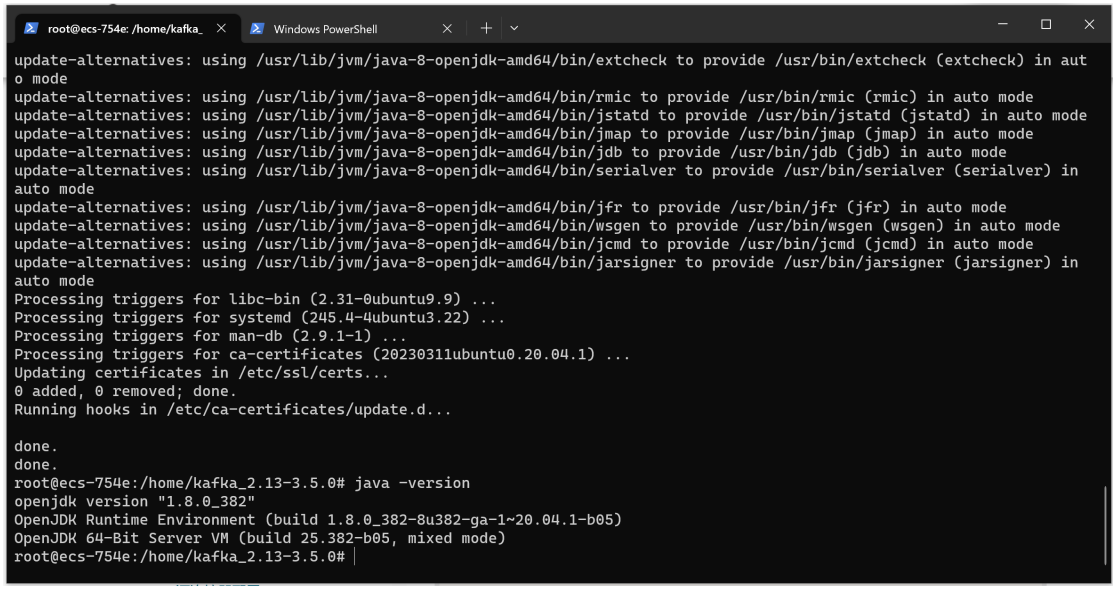
1.3 启动 zookeeper

bin/zookeeper-server-start.sh config/zookeeper.properties



```
root@ecs-754e: /home/kafka_   Windows PowerShell
root@ecs-754e:/home/kafka_2.13-3.5.0# bin/zookeeper-server-start.sh config/zookeeper.properties
[2023-09-26 10:27:04,334] INFO Reading configuration from: config/zookeeper.properties (org.apache.zookeeper.server.quor
um.QuorumPeerConfig)
[2023-09-26 10:27:04,336] WARN config/zookeeper.properties is relative. Prepend ./ to indicate that you're sure! (org.ap
ache.zookeeper.server.quorum.QuorumPeerConfig)
[2023-09-26 10:27:04,338] INFO clientPortAddress is 0.0.0.0:2181 (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2023-09-26 10:27:04,338] INFO secureClientPort is not set (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2023-09-26 10:27:04,339] INFO observerMasterPort is not set (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2023-09-26 10:27:04,339] INFO metricsProvider.className is org.apache.zookeeper.metrics.impl.DefaultMetricsProvider (or
g.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2023-09-26 10:27:04,343] INFO autopurge.snapRetainCount set to 3 (org.apache.zookeeper.server.DatadirCleanupManager)
[2023-09-26 10:27:04,343] INFO autopurge.purgeInterval set to 0 (org.apache.zookeeper.server.DatadirCleanupManager)
[2023-09-26 10:27:04,343] INFO Purge task is not scheduled. (org.apache.zookeeper.server.DatadirCleanupManager)
[2023-09-26 10:27:04,343] WARN Either no config or no quorum defined in config, running in standalone mode (org.apache.z
ookeeper.server.quorum.QuorumPeerMain)
[2023-09-26 10:27:04,347] INFO Log4j 1.2 jmx support not found; jmx disabled. (org.apache.zookeeper.jmx.ManagedUtil)
[2023-09-26 10:27:04,348] INFO Reading configuration from: config/zookeeper.properties (org.apache.zookeeper.server.quor
um.QuorumPeerConfig)
[2023-09-26 10:27:04,348] WARN config/zookeeper.properties is relative. Prepend ./ to indicate that you're sure! (org.ap
ache.zookeeper.server.quorum.QuorumPeerConfig)
[2023-09-26 10:27:04,348] INFO clientPortAddress is 0.0.0.0:2181 (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2023-09-26 10:27:04,348] INFO secureClientPort is not set (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2023-09-26 10:27:04,348] INFO observerMasterPort is not set (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2023-09-26 10:27:04,348] INFO metricsProvider.className is org.apache.zookeeper.metrics.impl.DefaultMetricsProvider (or
g.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2023-09-26 10:27:04,349] INFO Starting server (org.apache.zookeeper.server.ZooKeeperServerMain)
[2023-09-26 10:27:04,367] INFO ServerMetrics initialized with provider org.apache.zookeeper.metrics.impl.DefaultMetricsP
rovider@531be3c5 (org.apache.zookeeper.server.ServerMetrics)
```

图 3: 运行 1



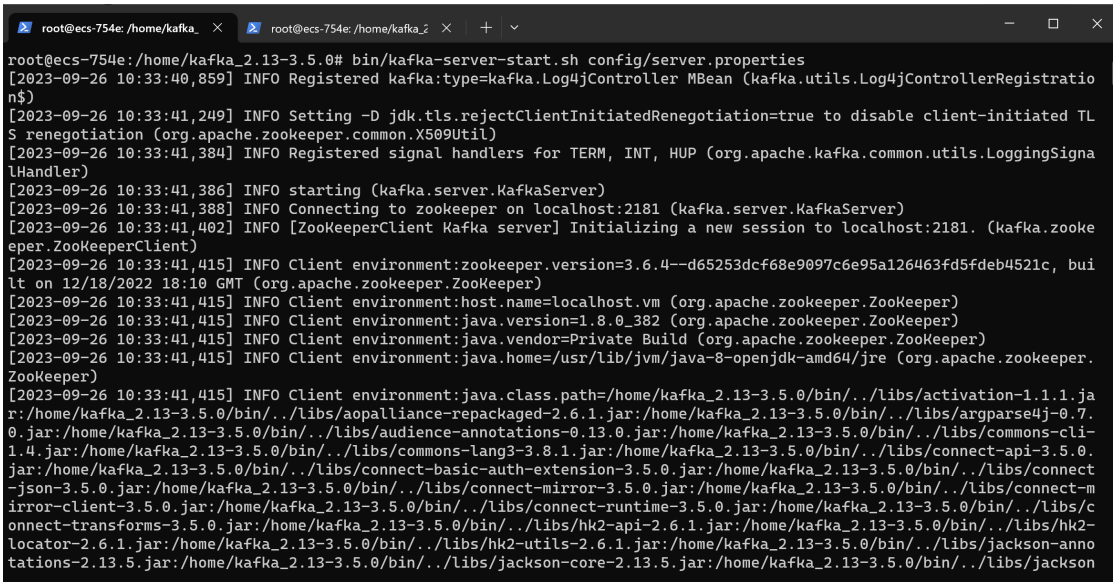
```
root@ecs-754e: /home/kafka_   Windows PowerShell
update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/bin/extcheck to provide /usr/bin/extcheck (extcheck) in aut
o mode
update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/bin/rmic to provide /usr/bin/rmic (rmic) in auto mode
update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/bin/jstatd to provide /usr/bin/jstatd (jstatd) in auto mode
update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/bin/jmap to provide /usr/bin/jmap (jmap) in auto mode
update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/bin/jdb to provide /usr/bin/jdb (jdb) in auto mode
update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/bin/serialver to provide /usr/bin/serialver (serialver) in
auto mode
update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/bin/jfr to provide /usr/bin/jfr (jfr) in auto mode
update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/bin/wsgen to provide /usr/bin/wsgen (wsgen) in auto mode
update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/bin/jcmd to provide /usr/bin/jcmd (jcmd) in auto mode
update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/bin/jarsigner to provide /usr/bin/jarsigner (jarsigner) in
auto mode
Processing triggers for libc-bin (2.31-0ubuntu9.9) ...
Processing triggers for systemd (245.4-4ubuntu3.22) ...
Processing triggers for man-db (2.9.1-1) ...
Processing triggers for ca-certificates (20230311ubuntu0.20.04.1) ...
Updating certificates in /etc/ssl/certs...
0 added, 0 removed; done.
Running hooks in /etc/ca-certificates/update.d...

done.
done.
root@ecs-754e:/home/kafka_2.13-3.5.0# java -version
openjdk version "1.8.0_382"
OpenJDK Runtime Environment (build 1.8.0_382-8u382-ga-1~20.04.1-b05)
OpenJDK 64-Bit Server VM (build 25.382-b05, mixed mode)
root@ecs-754e:/home/kafka_2.13-3.5.0#
```

图 4: 运行 2

1.4 启动 server

bin/kafka-server-start.sh config/server.properties



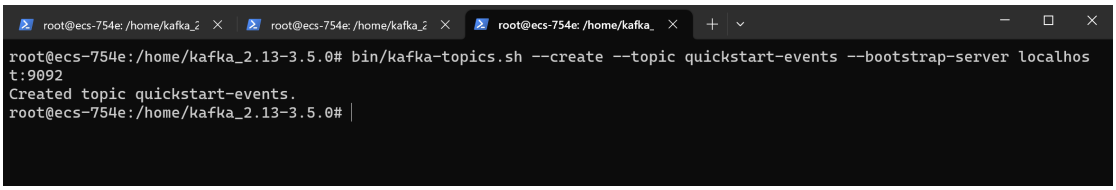
```
root@ecs-754e:/home/kafka_2.13-3.5.0# bin/kafka-server-start.sh config/server.properties
[2023-09-26 10:33:40,859] INFO Registered kafka:type=kafka.Log4jController MBean (kafka.utils.Log4jControllerRegistratio
n$)
[2023-09-26 10:33:41,249] INFO Setting -D jdk.tls.rejectClientInitiatedRenegotiation=true to disable client-initiated TL
S renegotiation (org.apache.zookeeper.common.X509Util)
[2023-09-26 10:33:41,384] INFO Registered signal handlers for TERM, INT, HUP (org.apache.kafka.common.utils.LoggingSigna
lHandler)
[2023-09-26 10:33:41,386] INFO starting (kafka.server.KafkaServer)
[2023-09-26 10:33:41,388] INFO Connecting to zookeeper on localhost:2181 (kafka.server.KafkaServer)
[2023-09-26 10:33:41,402] INFO [ZooKeeperClient Kafka server] Initializing a new session to localhost:2181. (kafka.zooke
eper.ZooKeeperClient)
[2023-09-26 10:33:41,415] INFO Client environment:zookeeper.version=3.6.4--d65253dcf68e9097c6e95a126463fd5fdeb4521c, bui
lt on 12/18/2022 18:10 GMT (org.apache.zookeeper.ZooKeeper)
[2023-09-26 10:33:41,415] INFO Client environment:host.name=localhost.vm (org.apache.zookeeper.ZooKeeper)
[2023-09-26 10:33:41,415] INFO Client environment:java.version=1.8.0_382 (org.apache.zookeeper.ZooKeeper)
[2023-09-26 10:33:41,415] INFO Client environment:java.vendor=Private Build (org.apache.zookeeper.ZooKeeper)
[2023-09-26 10:33:41,415] INFO Client environment:java.home=/usr/lib/jvm/java-8-openjdk-amd64/jre (org.apache.zookeeper.
ZooKeeper)
[2023-09-26 10:33:41,415] INFO Client environment:java.class.path=/home/kafka_2.13-3.5.0/bin/../libs/activation-1.1.1.ja
r:/home/kafka_2.13-3.5.0/bin/../libs/aopalliance-repackaged-2.6.1.jar:/home/kafka_2.13-3.5.0/bin/../libs/argparse4j-0.7.
0.jar:/home/kafka_2.13-3.5.0/bin/../libs/audience-annotations-0.13.0.jar:/home/kafka_2.13-3.5.0/bin/../libs/commons-cli-
1.4.jar:/home/kafka_2.13-3.5.0/bin/../libs/commons-lang3-3.8.1.jar:/home/kafka_2.13-3.5.0/bin/../libs/connect-api-3.5.0.
jar:/home/kafka_2.13-3.5.0/bin/../libs/connect-basic-auth-extension-3.5.0.jar:/home/kafka_2.13-3.5.0/bin/../libs/connect
-json-3.5.0.jar:/home/kafka_2.13-3.5.0/bin/../libs/connect-mirror-3.5.0.jar:/home/kafka_2.13-3.5.0/bin/../libs/connect-m
irror-client-3.5.0.jar:/home/kafka_2.13-3.5.0/bin/../libs/connect-runtime-3.5.0.jar:/home/kafka_2.13-3.5.0/bin/../libs/c
onnect-transforms-3.5.0.jar:/home/kafka_2.13-3.5.0/bin/../libs/hk2-api-2.6.1.jar:/home/kafka_2.13-3.5.0/bin/../libs/hk2-
locator-2.6.1.jar:/home/kafka_2.13-3.5.0/bin/../libs/hk2-utils-2.6.1.jar:/home/kafka_2.13-3.5.0/bin/../libs/jackson-anno
tations-2.13.5.jar:/home/kafka_2.13-3.5.0/bin/../libs/jackson-core-2.13.5.jar:/home/kafka_2.13-3.5.0/bin/../libs/jackson
```

图 5: 运行 3

1.5 创建一个主题

打开一个新终端

bin/kafka-topics.sh –create –topic quickstart-events –bootstrap-server localhost:9092



```
root@ecs-754e:/home/kafka_2.13-3.5.0# bin/kafka-topics.sh --create --topic quickstart-events --bootstrap-server localhos
t:9092
Created topic quickstart-events.
root@ecs-754e:/home/kafka_2.13-3.5.0#
```

图 6: 创建主题

1.6 打开生产者，将事件写入主题

bin/kafka-console-producer.sh –topic quickstart-events –bootstrap-server localhost:9092

```
root@ecs-754e:/home/kafka_2.13-3.5.0# bin/kafka-console-producer.sh --topic quickstart-events --bootstrap-server localho
st:9092
>Hello World!
>This is a messgae
>^Croot@ecs-754e:/home/kafka_2.13-3.5.0#
```

图 7: 创建主题

1.7 打开消费者，读取事件

bin/kafka-console-consumer.sh –topic quickstart-events –from-beginning –bootstrap-server localhost:9092

成功读取到了生产者的信息

图 8: 阅读事件

确认了 kafka 主题中的信息，实验成功

2． 任务四：使用 Flume 收集日志

2.1 安装 hadoop 环境

在上一次作业的某次尝试中，已经在我的 ECS 中安装好了 java 环境和 hadoop，没有留下截图。参考了阿里云手册，过程较为简单

2.2 将本地下载的 Flume 上传到云服务器

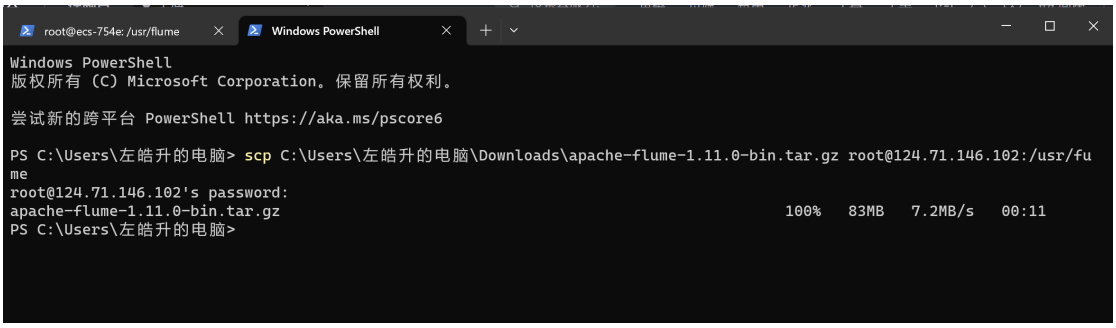


图 9: 上传

2.3 SSH 连接服务器解压并进入 conf 文件夹

```
Last login: Sat Sep 30 16:12:04 2023 from 49.70.116.43
root@ecs-ea1f:~# cd /usr/flume
root@ecs-ea1f:/usr/flume# tar -zxf apache-flume-1.11.0-bin.tar.gz
root@ecs-ea1f:/usr/flume# cd apache-flume-1.11.0-bin/
root@ecs-ea1f:/usr/flume/apache-flume-1.11.0-bin# cd conf
root@ecs-ea1f:/usr/flume/apache-flume-1.11.0-bin/conf#
```

图 10: 解压

2.4 新建.conf 文件并写入配置文件

```
root@ecs-ea1f:/usr/flume/apache-flume-1.11.0-bin/conf# vim flume.conf
```

图 11: 新建.conf

```
# Name the components on this agent
a1.sources = r1
a1.sinks = k1
a1.channels = c1

a1.sources.r1.type = TAILDIR
a1.sources.r1.positionFile = /export/servers/flume/taildir_position.json
a1.sources.r1.filegroups = f1 f2
a1.sources.r1.filegroups.f1 = /export/data/test1/example.log
a1.sources.r1.filegroups.f2 = /export/data/test2/.*log.*

# Describe the sink
#指定hdfs sink
a1.sinks.k1.type = hdfs
#hdfs目录，带有时间信息
a1.sinks.k1.hdfs.path = /flume/tailout/%Y-%m-%d/
#生成的hdfs文件名的前缀
a1.sinks.k1.hdfs.filePrefix = events-
#指定滚动时间，默认是30秒，设置为0表示禁用该策略
a1.sinks.k1.hdfs.rollInterval = 0
#指定滚动大小，设置为0表示禁用该策略
a1.sinks.k1.hdfs.rollSize = 200000000
#指定滚动条数
a1.sinks.k1.hdfs.rollCount = 0
a1.sinks.k1.hdfs.batchSize = 100
a1.sinks.k1.hdfs.useLocalTimeStamp = true
#副本策略
a1.sinks.k1.hdfs.minBlockReplicas=1
#生成的文件类型，默认是Sequencefile，可用DataStream，则为普通文本
a1.sinks.k1.hdfs.fileType = DataStream

# Use a channel which buffers events in memory
a1.channels.c1.type = memory
a1.channels.c1.capacity = 1000
a1.channels.c1.transactionCapacity = 100

# Bind the source and sink to the channel
a1.sources.r1.channels = c1
a1.sinks.k1.channel = c1

"flume.conf" 40L, 1297C                                40,0-1        All
```

图 12: 上传

2.5 新建文件夹和.log 文件

```
root@ecs-ea1f:/usr/flume/apache-flume-1.11.0-bin/conf# mkdir -p /export/
data/test1/
root@ecs-ea1f:/usr/flume/apache-flume-1.11.0-bin/conf# mkdir -p /export/
data/test2/
root@ecs-ea1f:/usr/flume/apache-flume-1.11.0-bin/conf#
```

图 13: 新建文件夹

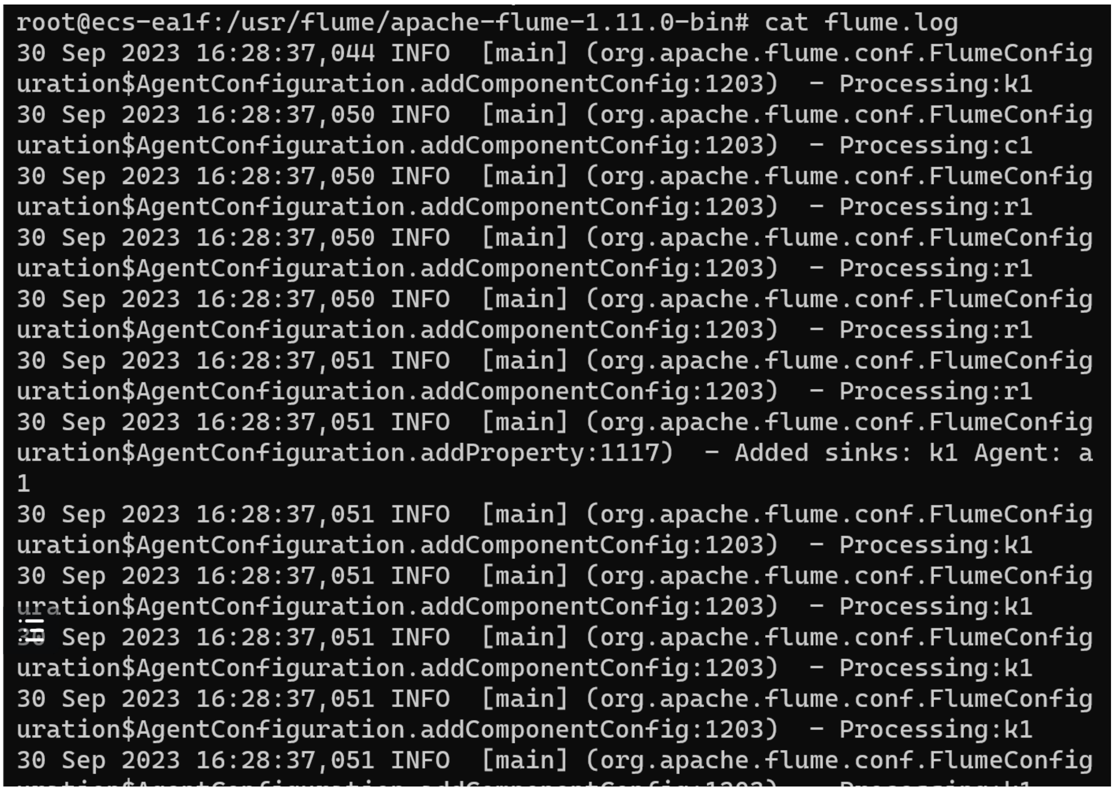
```
root@ecs-ea1f:/usr/flume/apache-flume-1.11.0-bin/conf# vim /export/data/
test1/example.log
```

图 14: 新建.log

2.6 启动 flume

```
root@ecs-ea1f:/usr/flume/apache-flume-1.11.0-bin# bin/flume-ng agent -na
me a1 -c conf/ -f /usr/flume/apache-flume-1.11.0-bin/conf/flume.conf
Info: Including Hadoop libraries found via (/opt/hadoop//bin/hadoop) for
 HDFS access
Info: Including Hive libraries found via () for Hive access
+ exec /usr/java8/bin/java -Xmx20m -cp '/usr/flume/apache-flume-1.11.0-b
in/conf:/usr/flume/apache-flume-1.11.0-bin/lib/*:/opt/hadoop/etc/hadoop:
/opt/hadoop/share/hadoop/common/lib/*:/opt/hadoop/share/hadoop/common/*:
/opt/hadoop/share/hadoop/hdfs:/opt/hadoop/share/hadoop/hdfs/lib/*:/opt/h
adoop/share/hadoop/hdfs/*:/opt/hadoop/share/hadoop/yarn:/opt/hadoop/shar
e/hadoop/yarn/lib/*:/opt/hadoop/share/hadoop/yarn/*:/opt/hadoop/share/ha
doop/mapreduce/lib/*:/opt/hadoop/share/hadoop/mapreduce/*:/opt/hadoop//c
ontrib/capacity-scheduler/*.jar:/lib/*' -Djava.library.path=:/opt/hadoop
/lib/native org.apache.flume.node.Application -name a1 -f /usr/flume/apa
che-flume-1.11.0-bin/conf/flume.conf
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/flume/apache-flume-1.11.0-bin/lib
/log4j-slf4j-impl-2.18.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/opt/hadoop/share/hadoop/common/lib/sl
f4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an expl
anation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFa
ctory]
```

图 15: 启动 flume

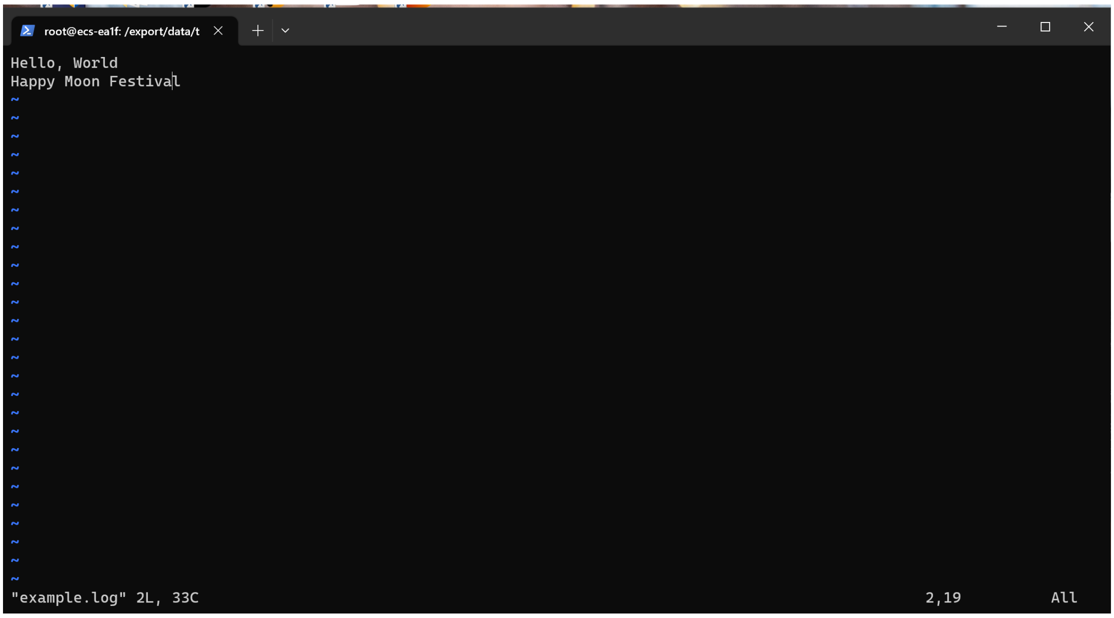


```
root@ecs-ea1f:/usr/flume/apache-flume-1.11.0-bin# cat flume.log
30 Sep 2023 16:28:37,044 INFO  [main] (org.apache.flume.conf.FlumeConfig
uration$AgentConfiguration.addComponentConfig:1203)  - Processing:k1
30 Sep 2023 16:28:37,050 INFO  [main] (org.apache.flume.conf.FlumeConfig
uration$AgentConfiguration.addComponentConfig:1203)  - Processing:c1
30 Sep 2023 16:28:37,050 INFO  [main] (org.apache.flume.conf.FlumeConfig
uration$AgentConfiguration.addComponentConfig:1203)  - Processing:r1
30 Sep 2023 16:28:37,050 INFO  [main] (org.apache.flume.conf.FlumeConfig
uration$AgentConfiguration.addComponentConfig:1203)  - Processing:r1
30 Sep 2023 16:28:37,050 INFO  [main] (org.apache.flume.conf.FlumeConfig
uration$AgentConfiguration.addComponentConfig:1203)  - Processing:r1
30 Sep 2023 16:28:37,051 INFO  [main] (org.apache.flume.conf.FlumeConfig
uration$AgentConfiguration.addComponentConfig:1203)  - Processing:r1
30 Sep 2023 16:28:37,051 INFO  [main] (org.apache.flume.conf.FlumeConfig
uration$AgentConfiguration.addProperty:1117)  - Added sinks: k1 Agent: a
1
30 Sep 2023 16:28:37,051 INFO  [main] (org.apache.flume.conf.FlumeConfig
uration$AgentConfiguration.addComponentConfig:1203)  - Processing:k1
30 Sep 2023 16:28:37,051 INFO  [main] (org.apache.flume.conf.FlumeConfig
uration$AgentConfiguration.addComponentConfig:1203)  - Processing:k1
30 Sep 2023 16:28:37,051 INFO  [main] (org.apache.flume.conf.FlumeConfig
uration$AgentConfiguration.addComponentConfig:1203)  - Processing:k1
30 Sep 2023 16:28:37,051 INFO  [main] (org.apache.flume.conf.FlumeConfig
uration$AgentConfiguration.addComponentConfig:1203)  - Processing:k1
30 Sep 2023 16:28:37,051 INFO  [main] (org.apache.flume.conf.FlumeConfig
```

图 16: 启动成功

2.7 写入 example.log 文件

向 example.log 文件中写入数据



```
Hello, World
Happy Moon Festival
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
"example.log" 2L, 33C                                          2,19          All
```

图 17: 50070

2.8 查看 50070 端口

通过服务器的公网 IP 访问 50070 端口，可以看到.tmp 文件

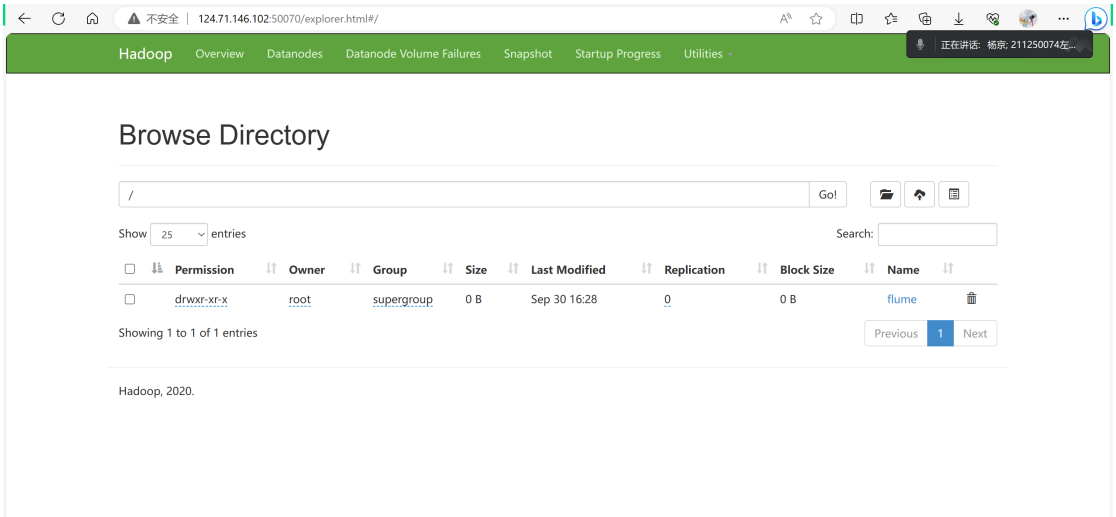


图 18: 50070

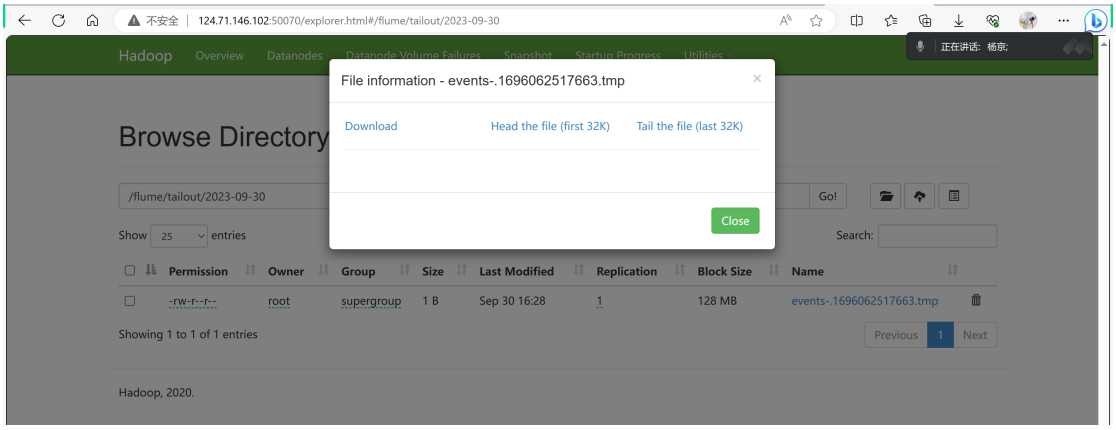


图 19: 50070

2.9 查看.tmp 文件

通过命令行查看.tmp 文件，成功看到 example.log 中我们写入的信息

```
root@ecs-ea1f:/opt/hadoop# hadoop dfs -cat /flume/tailout/2023-09-30/eve
nts-.1696062517663.tmp
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.


Hello, World
Happy Moon Festival
root@ecs-ea1f:/opt/hadoop#
```

图 20: 成功查看到

成功实现了使用 Flume 收集日志，实验成功

3． 实验中遇到的困难

3.1 任务一

任务一较为简单，并且再 kafka 的官网提供了较为详近的手册，在实验的过程中比较顺利，未遇到较大的困难。

3.2 任务四

任务四涉及到 hadoop 和 flume，遇到的最大困难是 flume 中的配置文件如何使用。一开始使用了几个网络上的配置文件，均无法成功实验。后来通过去学习了解配置文件中的每一行的意思，找到了合适的配置文件。

另外，在 50070 端口的图形界面中查看 event 文件会有延迟，让我一度怀疑自己没有实验成功。后来通过命令行打开.tmp 文件，成功查看到了信息。