综合实验：实现推荐系统

班级：2018211310

学号：2018211527

姓名：陈泊任

时间： 2021年6月25日

(单人完成，无分工)

综合实验：实现推荐系统

一、实验目的

掌握 Lambda 架构

二、实验平台

操作系统：Ubuntu 18.04 LTS

三、实验环境

Java：java-1.8.0

Python：python-3.8.5

Scala：scala-2.11.8

Flume：apach-flume-1.9.0

Kafka：kafka_2.11-2.4.1

Hadoop：hadoop-2.10.1

HBase：hbase-2.3.5

Spark：spark-2.4.7

Redis：redis-6.0.6

其他环境都是之前的四次实验配置好的

安装redis

从官网下载，编译 src 即可。

遇到的问题：

如果编译src报错，则需按照下面的步骤

```
make distclean  && make
```

Redis的测试在实验**步骤 5.5**

四、实验内容

3.1 基于 Lambda 架构的大数据推荐系统架构概述

3.2 使用 Flume、kafka 实现数据的手机

3.3 使用 HBase 实现数据的存储

3.4 使用 Redis 作为缓存数据实现特征/模型参数的存储

3.5 使用 MapReduce 或 Spark 实现定时启动的批示计算任务

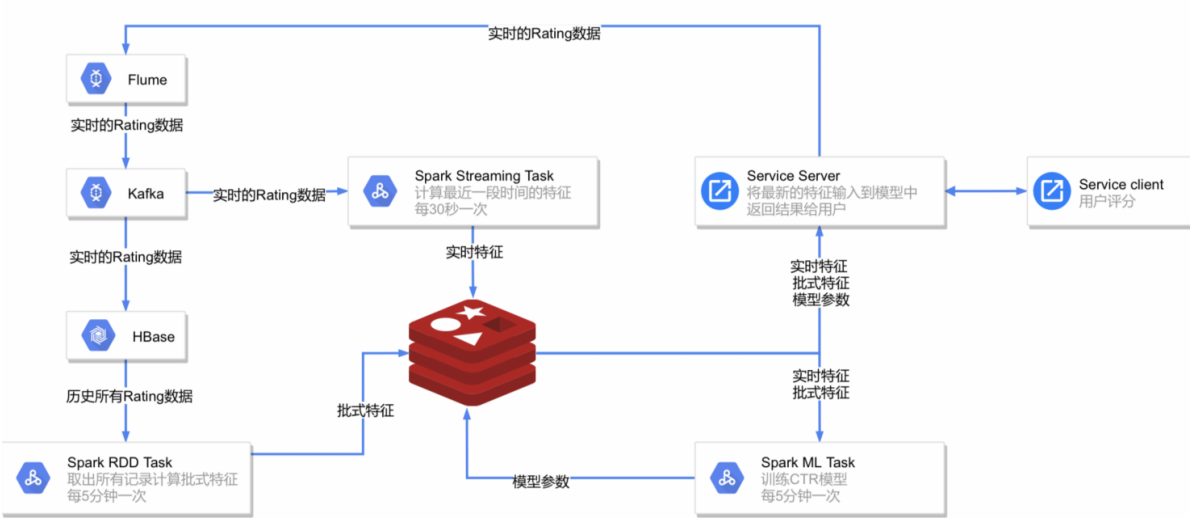
3.6 使用 Spark Streaming 或 Strom 实现流式计算任务

3.7 实现 服务应答

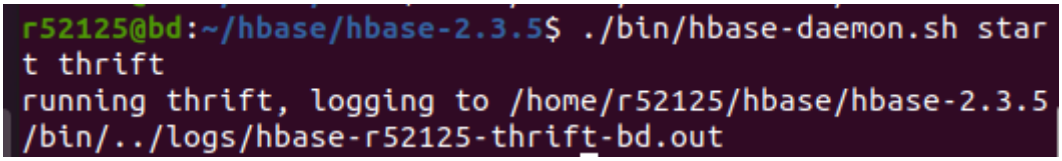
五、实验步骤

5.0 了解整个模型的流程

根据学长的实验指导，了解整个模型的流动。画出一个简易的流程图

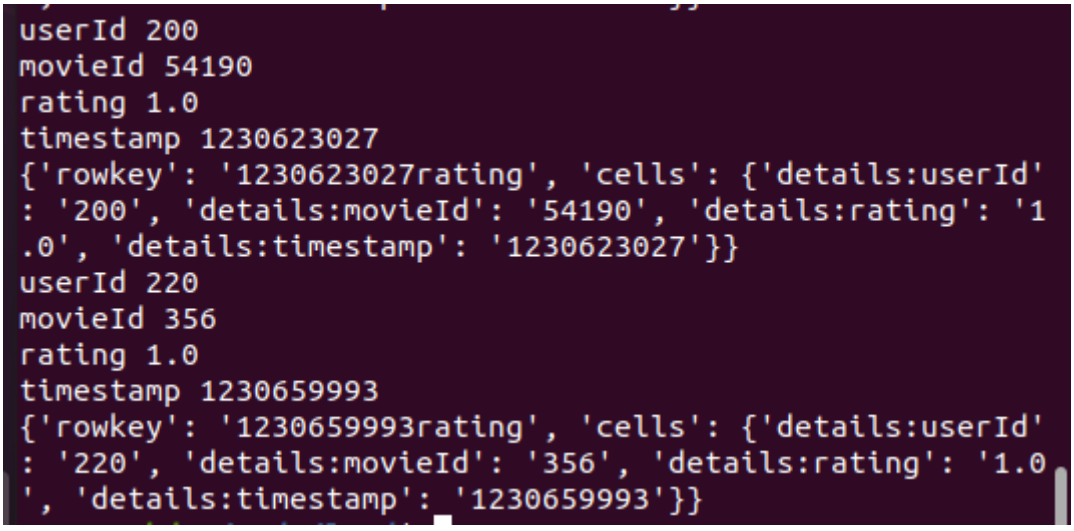


5.1 开启HBase thrift



5.2 启动 load_train_ratings_hbase.py

```
python3 load_train_ratings_hbase.py bd 9090 movie_records
~/data/json_train_ratings.json
```



查看'movie_records'表的内容

```
scan 'movie_records'
```

```
999978543ratin column=details:timestamp, timestamp=2021-0
g              6-15T22:19:03.257, value=999978543
999978543ratin column=details:userId, timestamp=2021-06-1
g              5T22:19:03.257, value=452
999995350ratin column=details:movieId, timestamp=2021-06-
g              15T22:19:03.258, value=4617
999995350ratin column=details:rating, timestamp=2021-06-1
g              5T22:19:03.258, value=1.0
999995350ratin column=details:timestamp, timestamp=2021-0
g              6-15T22:19:03.258, value=999995350
999995350ratin column=details:userId, timestamp=2021-06-1
g              5T22:19:03.258, value=414
10000 row(s)
Took 14.5294 seconds
```

5.3 启动 load_movie_redis.py

```
python3 load_movie_redis.py bd 6379 ~/data/movies.csv
```

```
193579 Jon Stewart Has Left the Building (2015)

193581 Black Butler: Book of the Atlantic (2017)

193583 No Game No Life: Zero (2017)

193585 Flint (2017)

193587 Bungo Stray Dogs: Dead Apple (2018)

193609 Andrew Dice Clay: Dice Rules (1991)

Load /home/r52125/data/movies.csv success!
```

5.4 开启 Flume

与实验二相同

5.5 启动 Kafka 并创建 Kafka Topic

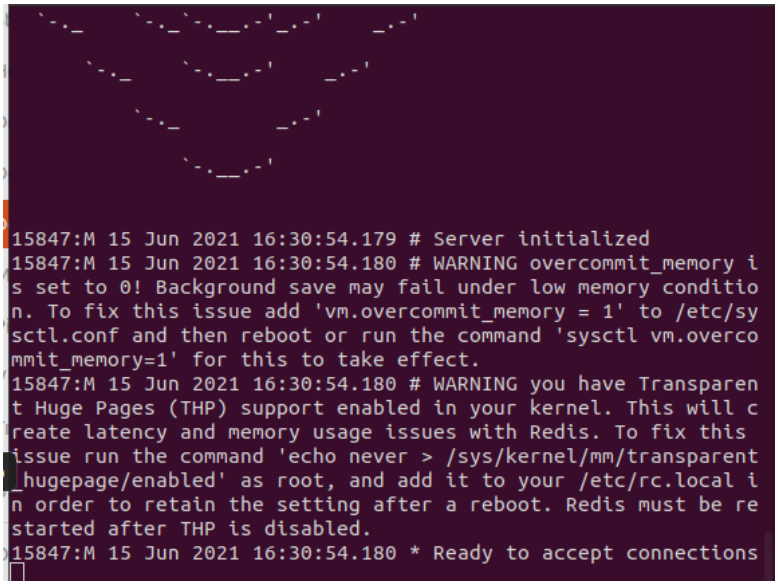
Kafka Topic："movie_rating_records"

与实验二相同

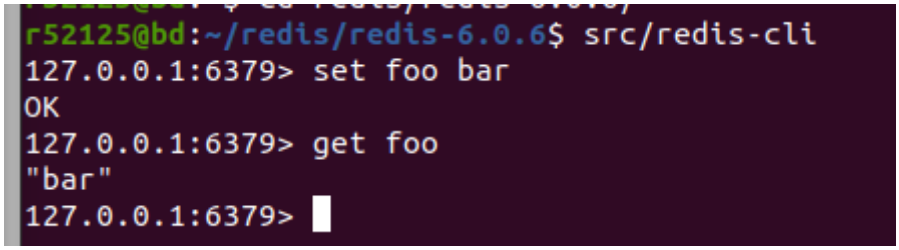
5.6 启动 Redis

测试

```
src/redis-server
```




```
src/redis-cli
```

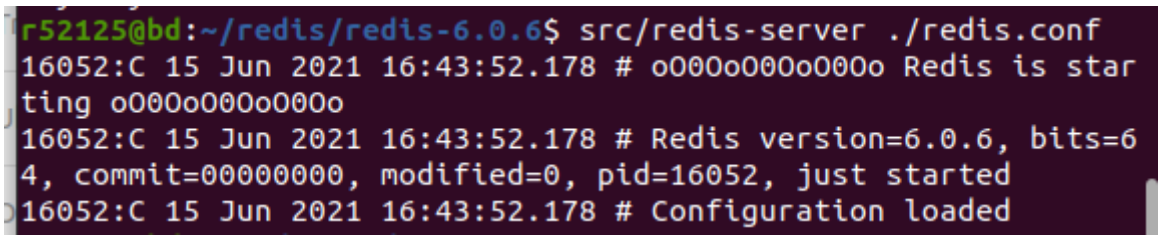


后台启动

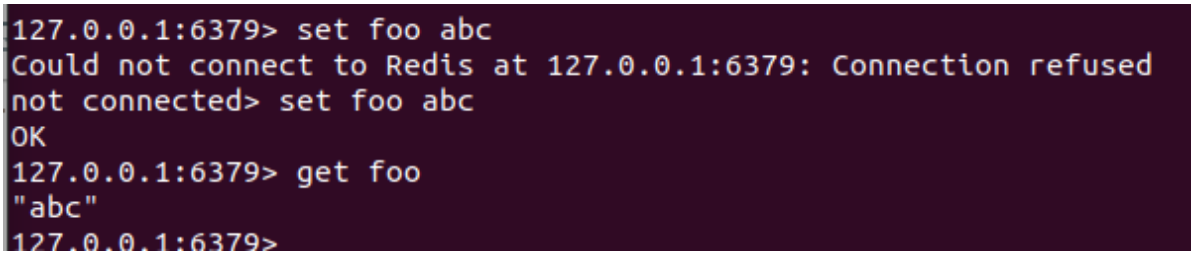
更改 redis.conf 的 daemonize

```
221 # Note that Redis will write a pid file in /var/run/redis.pid when daemonized.
222 daemonize yes
```

启动



再次测试



成功。

5.7 启动 generatorRecord.py

```
python3 generatorRecord.py -h bd -p 44444 -f ~/data/json_test_ratings.json
```

5.8 启动 hbase2spark.scala、kafkaStreaming.scala、recommend.scala

5.8.1 hbase2spark.scala

```
./bin/spark-submit ./hbase2spark9_jar/hbase2spark.jar
```

```
r52125@bd:~/spark/spark-2.4.7-bin-without-hadoop$ ./bin/spark-submit ./hbase2sp
ark9_jar/hbase2spark.jar
```

```
r52125@bd:~/spark/spark-2.4.7-bin-without-hadoop$ ./bin/spark-submit ./hbase2sp
ark9_jar/hbase2spark.jar
21/06/17 17:48:09 WARN util.Utils: Your hostname, bd resolves to a loopback add
ress: 127.0.1.1; using 192.168.223.161 instead (on interface ens33)
21/06/17 17:48:09 WARN util.Utils: Set SPARK_LOCAL_IP if you need to bind to an
other address
21/06/17 17:48:14 WARN util.NativeCodeLoader: Unable to load native-hadoop libr
ary for your platform... using builtin-java classes where applicable
2021-06-17 17:48:14 [INFO] Begin to calculate batch features
21/06/17 17:48:23 INFO util.log: Logging initialized @26560ms
21/06/17 17:48:23 INFO server.Server: jetty-9.3.z-SNAPSHOT, build timestamp: un
known, git hash: unknown
21/06/17 17:48:23 INFO server.Server: Started @26787ms
21/06/17 17:48:23 INFO server.AbstractConnector: Started ServerConnector@2d140a
7{HTTP/1.1,[http/1.1]}{0.0.0.0:4040}
21/06/17 17:48:23 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHa
ndler@2fa3be26{/jobs,null,AVAILABLE,@Spark}
21/06/17 17:48:23 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHa
ndler@3956b302{/jobs/json,null,AVAILABLE,@Spark}
21/06/17 17:48:23 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHa
ndler@1500e009{/jobs/job,null,AVAILABLE,@Spark}
```

五分钟计算一次批示特征

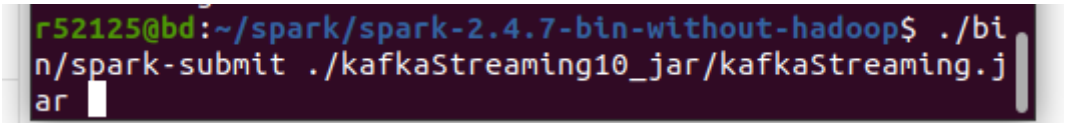
空行是区分两次计算，由时间可以看出，每隔五分钟计算一次，成功

```
21/06/17 17:48:34 INFO mapreduce.TableInputFormatBase: Input split length: 2 M
bytes.
21/06/17 17:48:45 INFO zookeeper.ZooKeeper: Session: 0x1000002997a0073 closed
21/06/17 17:48:45 INFO zookeeper.ClientCnxn: EventThread shut down for session:
 0x1000002997a0073
21/06/17 17:48:55 INFO server.AbstractConnector: Stopped Spark@2d140a7{HTTP/1.1
,[http/1.1]}{0.0.0.0:4040}
2021-06-17 17:48:56 [INFO] Success!

2021-06-17 17:53:56 [INFO] Begin to calculate batch features
21/06/17 17:53:56 INFO server.Server: jetty-9.3.z-SNAPSHOT, build timestamp: un
known, git hash: unknown
21/06/17 17:53:56 INFO server.Server: Started @359914ms
21/06/17 17:53:56 INFO server.AbstractConnector: Started ServerConnector@4130a6
48{HTTP/1.1,[http/1.1]}{0.0.0.0:4040}
21/06/17 17:53:56 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHa
ndler@61ff6a49{/jobs,null,AVAILABLE,@Spark}
21/06/17 17:53:56 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHa
ndler@18dd5ed3{/jobs/json,null,AVAILABLE,@Spark}
```

5.8.2 kafkaStreaming.scala

```
./bin/spark-submit ./kafkaStreaming10_jar/kafkaStreaming.jar
```




```
r52125@bd:~/spark/spark-2.4.7-bin-without-hadoop$ ./bin/spark-submit ./kafkaStre
aming10_jar/kafkaStreaming.jar
21/06/17 16:09:12 WARN util.Utils: Your hostname, bd resolves to a loopback addr
ess: 127.0.1.1; using 192.168.223.161 instead (on interface ens33)
21/06/17 16:09:12 WARN util.Utils: Set SPARK_LOCAL_IP if you need to bind to ano
ther address
21/06/17 16:09:14 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
21/06/17 16:09:34 INFO util.log: Logging initialized @30573ms
21/06/17 16:09:34 INFO server.Server: jetty-9.3.z-SNAPSHOT, build timestamp: unk
nown, git hash: unknown
21/06/17 16:09:34 INFO server.Server: Started @30841ms
21/06/17 16:09:34 INFO server.AbstractConnector: Started ServerConnector@3d4d1ec
1{HTTP/1.1,[http/1.1]}{0.0.0.0:4040}
21/06/17 16:09:35 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHan
dler@40620d8e{/jobs,null,AVAILABLE,@Spark}
21/06/17 16:09:35 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHan
dler@712ca57b{/jobs/json,null,AVAILABLE,@Spark}
21/06/17 16:09:35 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHan
dler@4564e94b{/jobs/job,null,AVAILABLE,@Spark}
21/06/17 16:09:35 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHan
dler@45673f68{/jobs/job/json,null,AVAILABLE,@Spark}
21/06/17 16:09:35 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHan
```

每过30秒计算一次，并将结果写入redis，中间的空行是我测试时间是否正确时输入回车隔开的，可以开出来，下面部分和上面部分除了时间相差30秒以外，其余都相同，说明确实实现了间隔30s将数据写入redis

```
21/06/17 23:05:37 INFO consumer.KafkaConsumer: [Consumer clientId=consumer-spa
rk-executor-kafkaStreaming-2, groupId=spark-executor-kafkaStreaming] Seeking t
o offset 717 for partition movie_rating_records-0
21/06/17 23:05:38 INFO consumer.KafkaConsumer: [Consumer clientId=consumer-spa
rk-executor-kafkaStreaming-2, groupId=spark-executor-kafkaStreaming] Seeking t
o offset 717 for partition movie_rating_records-0

21/06/17 23:06:00 INFO internals.SubscriptionState: [Consumer clientId=consume
r-kafkaStreaming-1, groupId=kafkaStreaming] Seeking to LATEST offset of partit
ion movie_rating_records-0
21/06/17 23:06:00 INFO internals.SubscriptionState: [Consumer clientId=consume
r-kafkaStreaming-1, groupId=kafkaStreaming] Resetting offset for partition mov
ie_rating_records-0 to offset 754.
21/06/17 23:06:00 INFO zookeeper.ZooKeeper: Initiating client connection, conn
ectString=bd:2181 sessionTimeout=90000 watcher=org.apache.hadoop.hbase.zookeep
er.ReadOnlyZKClient$$Lambda$201/907412982@57dbb36a
21/06/17 23:06:00 INFO zookeeper.ClientCnxn: Opening socket connection to serv
er bd/127.0.1.1:2181. Will not attempt to authenticate using SASL (unknown err
or)
```

```
21/06/17 17:26:30 INFO internals.SubscriptionState: [Consumer clientId=consumer-kafkaStreaming-1, gro
upId=kafkaStreaming] Seeking to LATEST offset of partition test-flume-pykafka-0
21/06/17 17:26:30 INFO internals.SubscriptionState: [Consumer clientId=consumer-kafkaStreaming-1, gro
upId=kafkaStreaming] Resetting offset for partition test-flume-pykafka-0 to offset 0.
21/06/17 17:26:30 INFO zookeeper.ZooKeeper: Initiating client connection, connectString=bd:2181 sessi
onTimeout=90000 watcher=org.apache.hadoop.hbase.zookeeper.ReadOnlyZKClient$$Lambda$202/17538726 4@53fa
52f5
21/06/17 17:26:30 INFO zookeeper.ClientCnxn: Opening socket connection to server bd/127.0.1.1:2181. W
ill not attempt to authenticate using SASL (unknown error)
21/06/17 17:26:30 INFO zookeeper.ClientCnxn: Socket connection established to bd/127.0.1.1:2181, init
iating session
21/06/17 17:26:30 INFO zookeeper.ClientCnxn: Session establishment complete on server bd/127.0.1.1:21
81, sessionid = 0x1000002997a0058, negotiated timeout = 90000
21/06/17 17:26:34 INFO zookeeper.ZooKeeper: Session: 0x1000002997a0058 closed
21/06/17 17:26:34 INFO zookeeper.ClientCnxn: EventThread shut down for session: 0x1000002997a0058
21/06/17 17:26:34 INFO zookeeper.ZooKeeper: Session: 0x1000002997a0057 closed
21/06/17 17:26:34 INFO zookeeper.ClientCnxn: EventThread shut down for session: 0x1000002997a0057

21/06/17 17:27:00 INFO internals.SubscriptionState: [Consumer clientId=consumer-kafkaStreaming-1, gro
upId=kafkaStreaming] Seeking to LATEST offset of partition test-flume-pykafka-0
21/06/17 17:27:00 INFO internals.SubscriptionState: [Consumer clientId=consumer-kafkaStreaming-1, gro
upId=kafkaStreaming] Resetting offset for partition test-flume-pykafka-0 to offset 0.
21/06/17 17:27:00 INFO zookeeper.ZooKeeper: Initiating client connection, connectString=bd:2181 sessi
onTimeout=90000 watcher=org.apache.hadoop.hbase.zookeeper.ReadOnlyZKClient$$Lambda$202/17538726 4@53fa
52f5
21/06/17 17:27:00 INFO zookeeper.ClientCnxn: Opening socket connection to server bd/127.0.1.1:2181. W
ill not attempt to authenticate using SASL (unknown error)
21/06/17 17:27:00 INFO zookeeper.ClientCnxn: Socket connection established to bd/127.0.1.1:2181, init
iating session
21/06/17 17:27:00 INFO zookeeper.ClientCnxn: Session establishment complete on server bd/127.0.1.1:21
81, sessionid = 0x1000002997a0059, negotiated timeout = 90000
S21/06/17 17:27:30 INFO internals.SubscriptionState: [Consumer clientId=consumer-kafkaStreaming-1, gr
oupId=kafkaStreaming] Seeking to LATEST offset of partition test-flume-pykafka-0
```

5.8.3 recommend.scala

```
./bin/spark-submit ./recommend1_jar/recommend.jar
```

```
r52125@bd:~/spark/spark-2.4.7-bin-without-hadoop$ ./bin/spark-submit ./recomme
nd1_jar/recommend.jar
21/06/17 23:15:07 WARN util.Utils: Your hostname, bd resolves to a loopback ad
dress: 127.0.1.1; using 192.168.223.161 instead (on interface ens33)
21/06/17 23:15:07 WARN util.Utils: Set SPARK_LOCAL_IP if you need to bind to a
nother address
21/06/17 23:15:10 WARN util.NativeCodeLoader: Unable to load native-hadoop lib
rary for your platform... using builtin-java classes where applicable
2021-06-17 23:15:11 [INFO] Begin to train lr model21/06/17 23:15:13 INFO util.
log: Logging initialized @11604ms
21/06/17 23:15:14 INFO server.Server: jetty-9.3.z-SNAPSHOT, build timestamp: u
nknown, git hash: unknown
21/06/17 23:15:14 INFO server.Server: Started @12075ms
21/06/17 23:15:14 INFO server.AbstractConnector: Started ServerConnector@4604b
900{HTTP/1.1,[http/1.1]}{0.0.0.0:4040}
```

五分钟一次

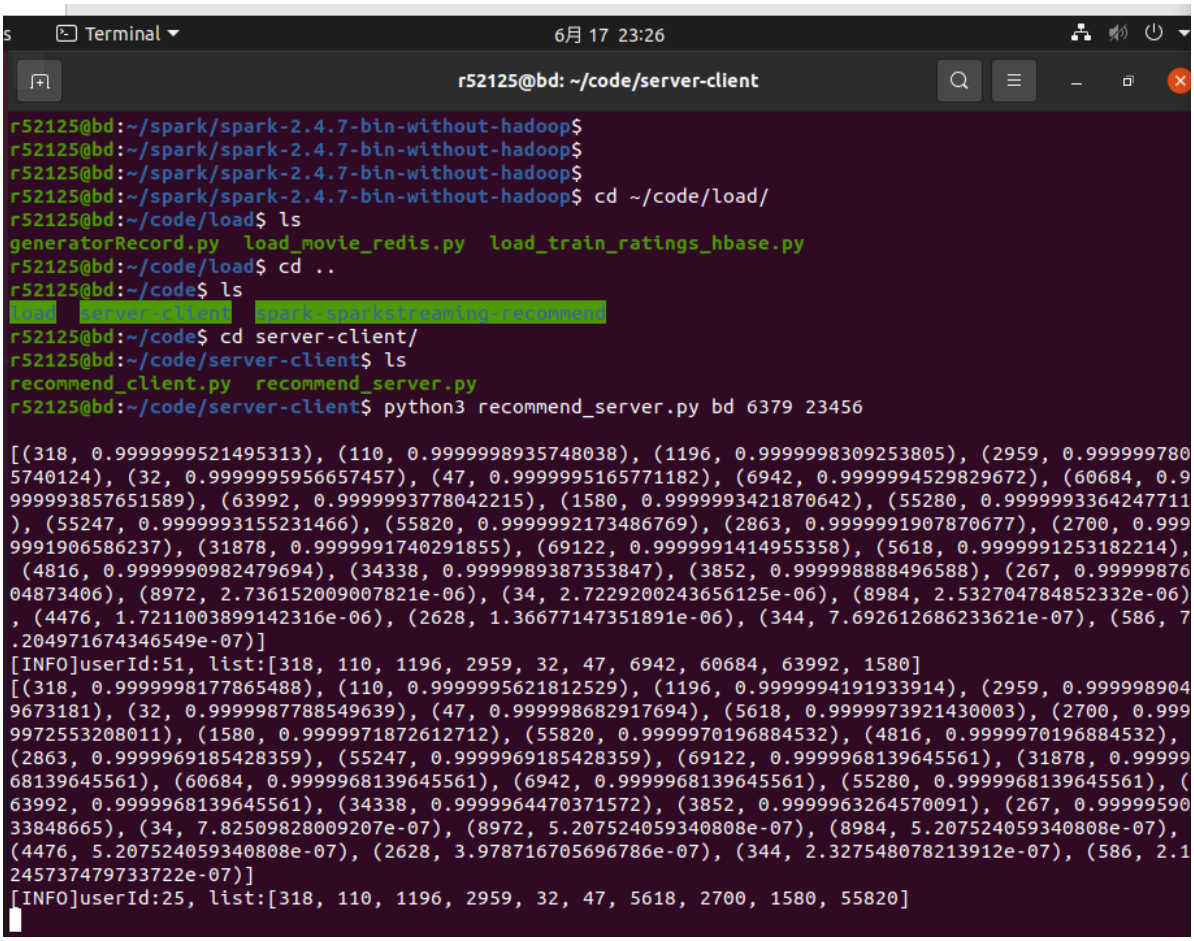
```
21/06/17 23:16:05 INFO optimize.LBFGS: Val and Grad Norm: 3.53372e-07 (rel: 0.
471) 6.94381e-07
21/06/17 23:16:05 INFO optimize.LBFGS: Converged because max iterations reache
d
Array(0.009709879938833745, 0.01744755078064542, 0.03337485056393552, -0.14237
974017074986, -0.0973645967389315, 0.31061732555098326, -0.19062967831155253,
10.680847626806608, -16.44387302471306, 0.8571366890927801)
21/06/17 23:16:44 INFO server.AbstractConnector: Stopped Spark@4604b900{HTTP/1
.1,[http/1.1]}{0.0.0.0:4040}
2021-06-17 23:16:46 [INFO] Success!



2021-06-17 23:21:46 [INFO] Begin to train lr model21/06/17 23:21:47 INFO serve
r.Server: jetty-9.3.z-SNAPSHOT, build timestamp: unknown, git hash: unknown
21/06/17 23:21:47 INFO server.Server: Started @404956ms
21/06/17 23:21:47 INFO server.AbstractConnector: Started ServerConnector@23cb8
fcb{HTTP/1.1,[http/1.1]}{0.0.0.0:4040}
21/06/17 23:21:47 INFO handler.ContextHandler: Started o.s.j.s.ServletContextH
andler@25a2f432{/jobs,null,AVAILABLE,@Spark}
21/06/17 23:21:47 INFO handler.ContextHandler: Started o.s.j.s.ServletContextH
andler@7efba9b9{/jobs/json,null,AVAILABLE,@Spark}
```

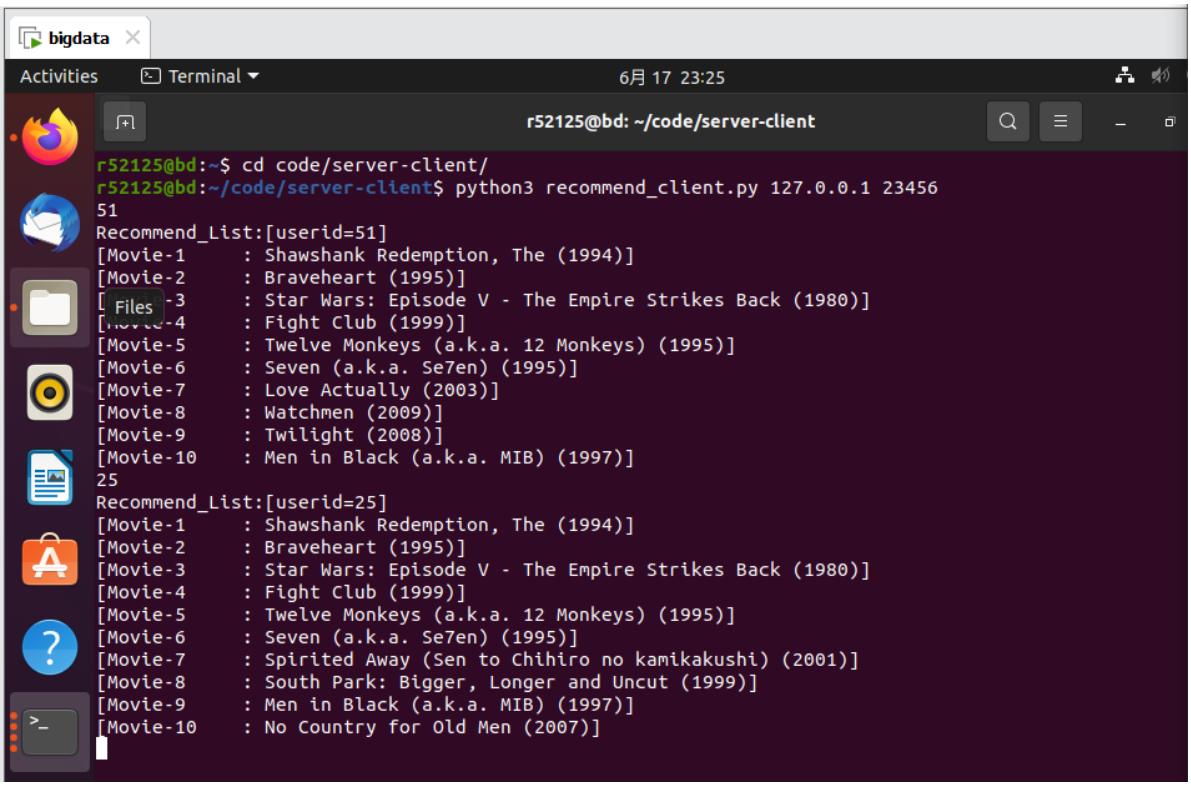
5.9 启动 recommend_server.py

```
python3 recommend_server.py bd 6379 23456
```



5.10 启动 recommend_client.py

```
python3 recommend_client.py 127.0.0.1 23456
```



六、遇到的问题

6.1 HMaster无法正常打开

解决方法：不知道什么原因，重新配置hbase

6.2 运行代码导致虚拟机死机

产生的问题：hbase的 zookeeper 的时间和 HMaster 的时间不同，导致 HMaster 异常退出

解决方法：回退到配置没有问题的地方（使用快照）