福建師範大學
FUJIAN NORMAL UNIVERSITY

计算机与网络空间安全学院学生实验报告

实验课程名称：大数据导论 教师：林鑫泓

实验名称	大数据软件引用与技术实践			实验成绩	
学生姓名	叶建行	学 号	116052020005	年级专业班级	2020 级软件工程一班
小组成员	无			实验日期	2022 年 12 月 17

一、实验要求

1. 结合 Netflix Dataset 或 ml-25m，按照实验步骤完成实验。

2. Netflix 数据集包含了 1999.12.31 至 2005.12.31 期间由网站用户提供的超过一亿条电影评价。Netflix Dataset.7z 压缩文件包含电影信息、training set（训练集）、probe set（探测集）和 qualifying set(评估集)。压缩文件的详细信息如下图 1 所示：

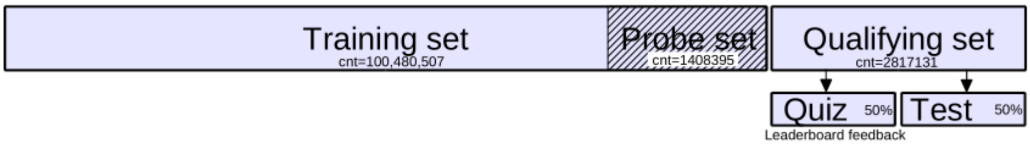


图 1 Netflix Dataset 说明

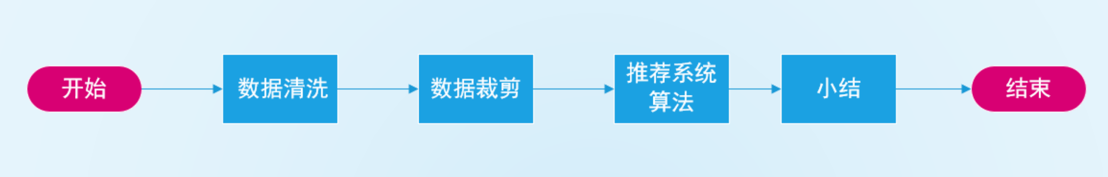
3. 实验成绩按照各步骤完成情况给分。

4. 大数据软件需包括教材当中有的，可以是介绍过的也可以是没介绍过的，尽可能地使用本学期学习并实验过的大数据软件和方法。

二、实验步骤

1. 独立设计一套切实可行的数据挖掘任务 task，并简要介绍该任务的各个步骤 step，使用的关键技术与软件应用。（10 分）

任务流程图：



数据清洗：通过 hive 实现，清除掉看的人少的，差评多的电影，全给好评或差评的用户。

数据裁剪：通过 hivesql 实现，随机抽样。

推荐系统算法：通过使用 mahout 推荐算法实现。

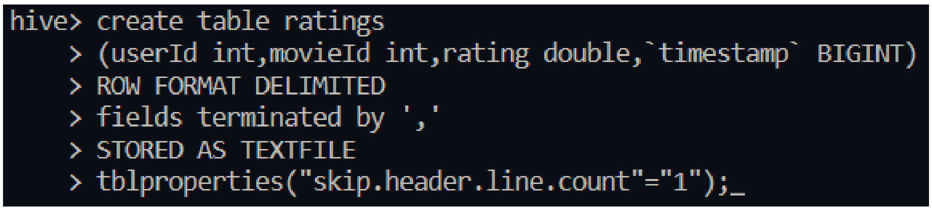
小结：对得到的结果进行分析和总结。

2. 详细介绍数据准备过程，例如：如何对数据进行预处理，如何做数据持久化存储。（10 分）

（1） 导入数据（hive 导表过程过程相似，只展示 ratings 表的导表过程）

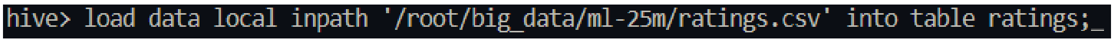
（由于之前没有保存结果截图，因此只是语句的截图，希望老师不要在意）

- **建表：**



```
hive> create table ratings
    > (userId int,movieId int,rating double,`timestamp` BIGINT)
    > ROW FORMAT DELIMITED
    > fields terminated by ','
    > STORED AS TEXTFILE
    > tblproperties("skip.header.line.count"="1");_
```

- **导入本地数据：**



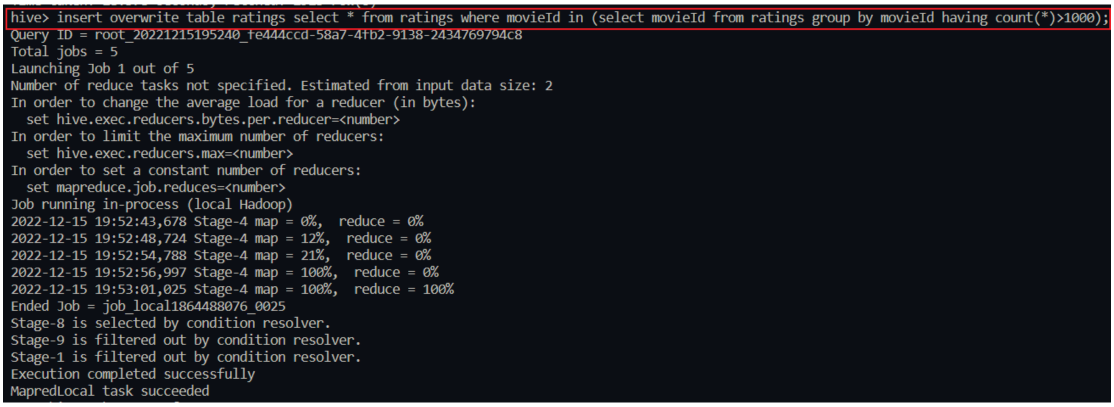
```
hive> load data local inpath '/root/big_data/ml-25m/ratings.csv' into table ratings;_
```

（2） ratings 表筛选了评分人数<=1000 的电影

执行语句：

```
insert overwrite table ratings select * from ratings where movieId in (select movieId from ratings group by movieId having count(*)>1000);
```

执行结果：



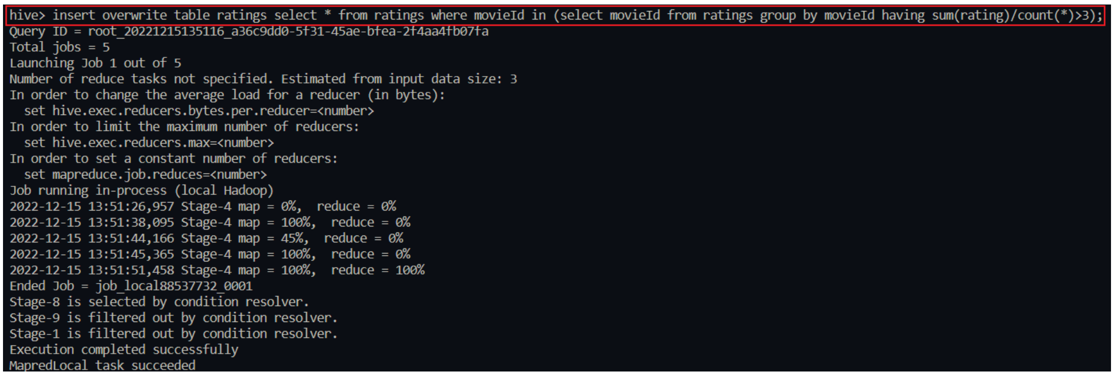
```
hive> insert overwrite table ratings select * from ratings where movieId in (select movieId from ratings group by movieId having count(*)>1000);
Query ID = root_20221215195240_fe444ccd-58a7-4fb2-9138-2434769794c8
Total jobs = 5
Launching Job 1 out of 5
Number of reduce tasks not specified. Estimated from input data size: 2
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Job running in-process (local Hadoop)
2022-12-15 19:52:43,678 Stage-4 map = 0%,  reduce = 0%
2022-12-15 19:52:48,724 Stage-4 map = 12%,  reduce = 0%
2022-12-15 19:52:54,788 Stage-4 map = 21%,  reduce = 0%
2022-12-15 19:52:56,997 Stage-4 map = 100%,  reduce = 0%
2022-12-15 19:53:01,025 Stage-4 map = 100%,  reduce = 100%
Ended Job = job_local1864488076_0025
Stage-8 is selected by condition resolver.
Stage-9 is filtered out by condition resolver.
Stage-1 is filtered out by condition resolver.
Execution completed successfully
MapredLocal task succeeded
```

（3）ratings 表筛选了电影评分的平均值<=3 的电影

执行语句：

```
insert overwrite table ratings select * from ratings where movieId in (select movieId from ratings group by movieId having sum(rating)/count(*)>3);
```

执行结果：



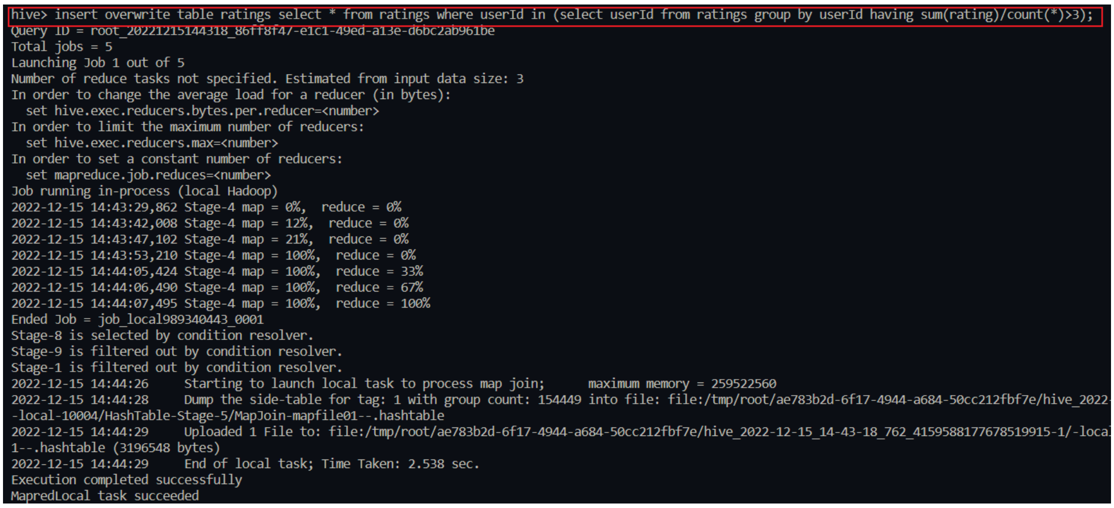
```
hive> insert overwrite table ratings select * from ratings where movieId in (select movieId from ratings group by movieId having sum(rating)/count(*)>3);
Query ID = root_20221215135116_a36c9dd0-5f31-45ae-bfea-2f4aa4fb07fa
Total jobs = 5
Launching Job 1 out of 5
Number of reduce tasks not specified. Estimated from input data size: 3
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Job running in-process (local Hadoop)
2022-12-15 13:51:26,957 Stage-4 map = 0%,  reduce = 0%
2022-12-15 13:51:38,095 Stage-4 map = 100%,  reduce = 0%
2022-12-15 13:51:44,166 Stage-4 map = 45%,  reduce = 0%
2022-12-15 13:51:45,365 Stage-4 map = 100%,  reduce = 0%
2022-12-15 13:51:51,458 Stage-4 map = 100%,  reduce = 100%
Ended Job = job_local88537732_0001
Stage-8 is selected by condition resolver.
Stage-9 is filtered out by condition resolver.
Stage-1 is filtered out by condition resolver.
Execution completed successfully
MapredLocal task succeeded
```

（4）ratings 表筛选了给差评多的用户的评价（评价平均分<=3）

执行语句：

```
insert overwrite table ratings select * from ratings where userId in (select userId from ratings group by userId having sum(rating)/count(*)>3);
```

执行结果：



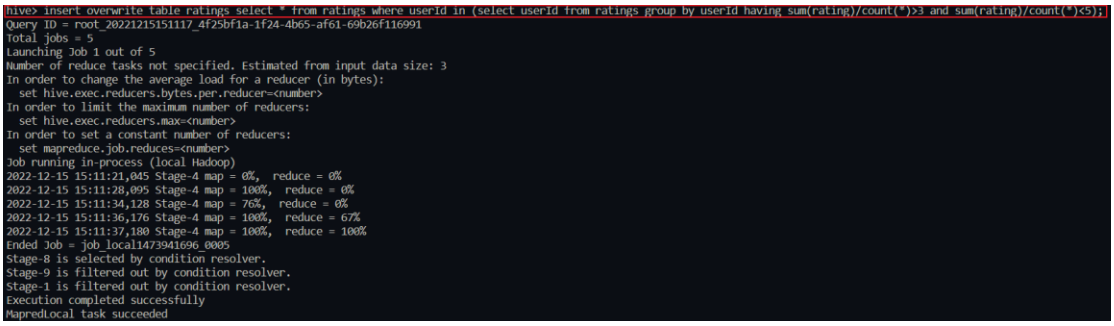
```
hive> insert overwrite table ratings select * from ratings where userId in (select userId from ratings group by userId having sum(rating)/count(*)>3);
Query ID = root_20221215144318_86ff8f47-e1c1-49ed-a13e-d6bc2ab961be
Total jobs = 5
Launching Job 1 out of 5
Number of reduce tasks not specified. Estimated from input data size: 3
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Job running in-process (local Hadoop)
2022-12-15 14:43:29,862 Stage-4 map = 0%,  reduce = 0%
2022-12-15 14:43:42,008 Stage-4 map = 12%,  reduce = 0%
2022-12-15 14:43:47,102 Stage-4 map = 21%,  reduce = 0%
2022-12-15 14:43:53,210 Stage-4 map = 100%,  reduce = 0%
2022-12-15 14:44:05,424 Stage-4 map = 100%,  reduce = 33%
2022-12-15 14:44:06,490 Stage-4 map = 100%,  reduce = 67%
2022-12-15 14:44:07,495 Stage-4 map = 100%,  reduce = 100%
Ended Job = job_local989340443_0001
Stage-8 is selected by condition resolver.
Stage-9 is filtered out by condition resolver.
Stage-1 is filtered out by condition resolver.
2022-12-15 14:44:26     Starting to launch local task to process map join;      maximum memory = 259522560
2022-12-15 14:44:28     Dump the side-table for tag: 1 with group count: 154449 into file: file:/tmp/root/ae783b2d-6f17-4944-a684-50cc212fbf7e/hive_2022
-local-10004/HashTable-Stage-5/MapJoin-mapfile01--.hashtable
2022-12-15 14:44:29     Uploaded 1 File to: file:/tmp/root/ae783b2d-6f17-4944-a684-50cc212fbf7e/hive_2022-12-15_14-43-18_762_4159588177678519915-1/-local
1--.hashtable (3196548 bytes)
2022-12-15 14:44:29     End of local task; Time Taken: 2.538 sec.
Execution completed successfully
MapredLocal task succeeded
```

（5）**ratings 表筛选了全给好评的用户的评价（评价平均分==5）**

执行语句：

```
insert overwrite table ratings select * from ratings where userId in (select userId from ratings group by userId having sum(rating)/count(*)>3 and sum(rating)/count(*)<5);
```

执行结果：



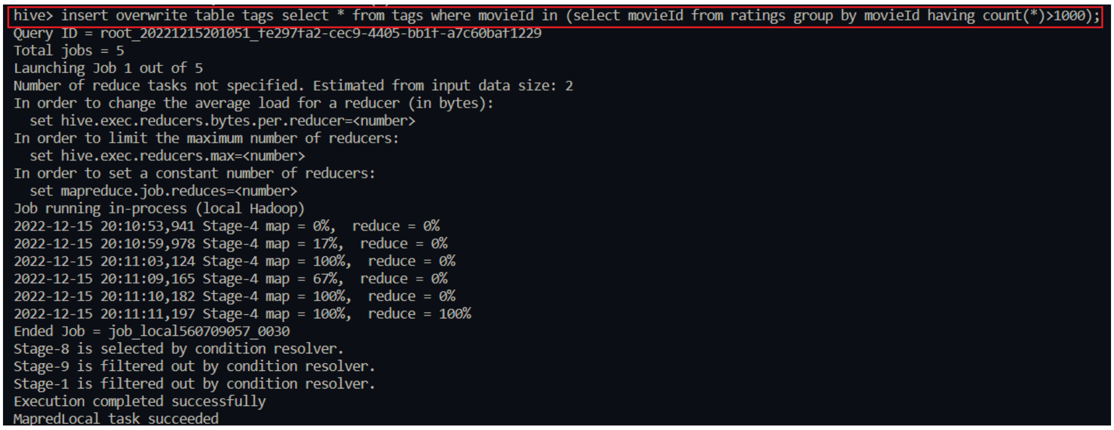
```
hive> insert overwrite table ratings select * from ratings where userId in (select userId from ratings group by userId having sum(rating)/count(*)>3 and sum(rating)/count(*)<5);
Query ID = root_20221215151117_4f25bf1a-1f24-4b65-af61-69b26f116991
Total jobs = 5
Launching Job 1 out of 5
Number of reduce tasks not specified. Estimated from input data size: 3
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Job running in-process (local Hadoop)
2022-12-15 15:11:21,045 Stage-4 map = 0%,  reduce = 0%
2022-12-15 15:11:28,095 Stage-4 map = 100%,  reduce = 0%
2022-12-15 15:11:34,128 Stage-4 map = 76%,  reduce = 0%
2022-12-15 15:11:36,176 Stage-4 map = 100%,  reduce = 67%
2022-12-15 15:11:37,180 Stage-4 map = 100%,  reduce = 100%
Ended Job = job_local1473941696_0005
Stage-8 is selected by condition resolver.
Stage-9 is filtered out by condition resolver.
Stage-1 is filtered out by condition resolver.
Execution completed successfully
MapredLocal task succeeded
```

（6）**tags 表筛选了评分人数<=1000 的电影**

执行语句：

```
insert overwrite table tags select * from tags where movieId in (select movieId from ratings group by movieId having count(*)>1000);
```

执行结果：



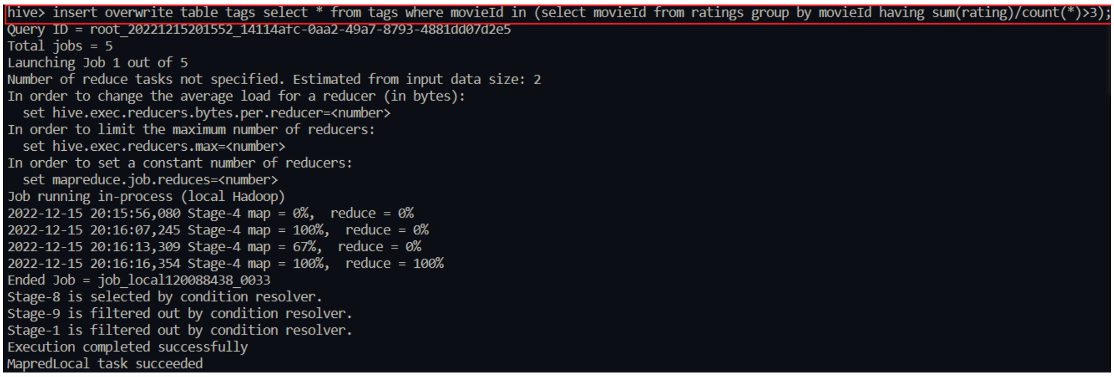
```
hive> insert overwrite table tags select * from tags where movieId in (select movieId from ratings group by movieId having count(*)>1000);
Query ID = root_20221215201051_fe297fa2-cec9-4405-bb1f-a7c60baf1229
Total jobs = 5
Launching Job 1 out of 5
Number of reduce tasks not specified. Estimated from input data size: 2
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Job running in-process (local Hadoop)
2022-12-15 20:10:53,941 Stage-4 map = 0%,  reduce = 0%
2022-12-15 20:10:59,978 Stage-4 map = 17%,  reduce = 0%
2022-12-15 20:11:03,124 Stage-4 map = 100%,  reduce = 0%
2022-12-15 20:11:09,165 Stage-4 map = 67%,  reduce = 0%
2022-12-15 20:11:10,182 Stage-4 map = 100%,  reduce = 0%
2022-12-15 20:11:11,197 Stage-4 map = 100%,  reduce = 100%
Ended Job = job_local560709057_0030
Stage-8 is selected by condition resolver.
Stage-9 is filtered out by condition resolver.
Stage-1 is filtered out by condition resolver.
Execution completed successfully
MapredLocal task succeeded
```

（7）tags 表筛选了电影评分的平均值<=3 的电影

执行语句：

```
insert overwrite table tags select * from tags where movieId in (select movieId from ratings group by movieId having sum(rating)/count(*)>3);
```

执行结果：



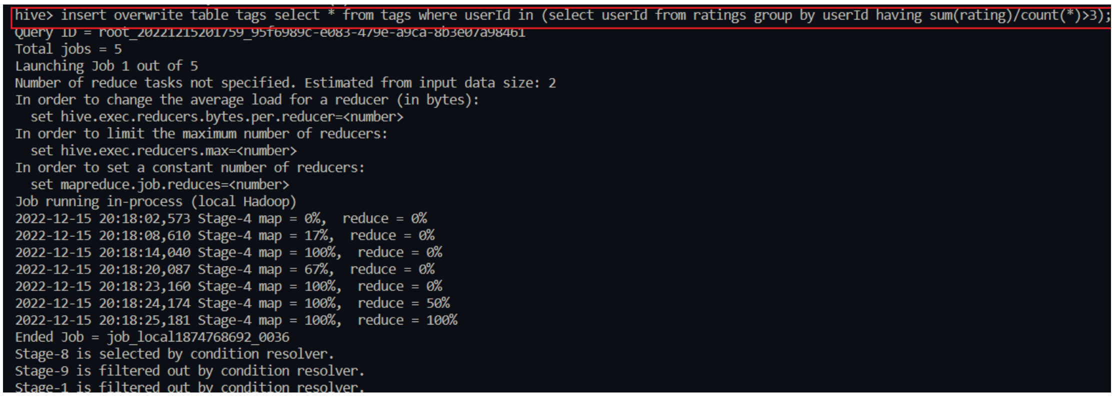
```
hive> insert overwrite table tags select * from tags where movieId in (select movieId from ratings group by movieId having sum(rating)/count(*)>3);
Query ID = root_20221215201552_14114afc-0aa2-49a7-8793-4881dd07d2e5
Total jobs = 5
Launching Job 1 out of 5
Number of reduce tasks not specified. Estimated from input data size: 2
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Job running in-process (local Hadoop)
2022-12-15 20:15:56,080 Stage-4 map = 0%,  reduce = 0%
2022-12-15 20:16:07,245 Stage-4 map = 100%,  reduce = 0%
2022-12-15 20:16:13,309 Stage-4 map = 67%,  reduce = 0%
2022-12-15 20:16:16,354 Stage-4 map = 100%,  reduce = 100%
Ended Job = job_local120088438_0033
Stage-8 is selected by condition resolver.
Stage-9 is filtered out by condition resolver.
Stage-1 is filtered out by condition resolver.
Execution completed successfully
MapredLocal task succeeded
```

（8）tags 表筛选了给差评多的用户的评价（评价平均分<=3）

执行语句：

```
insert overwrite table tags select * from tags where userId in (select userId from ratings group by userId having sum(rating)/count(*)>3);
```

执行结果：



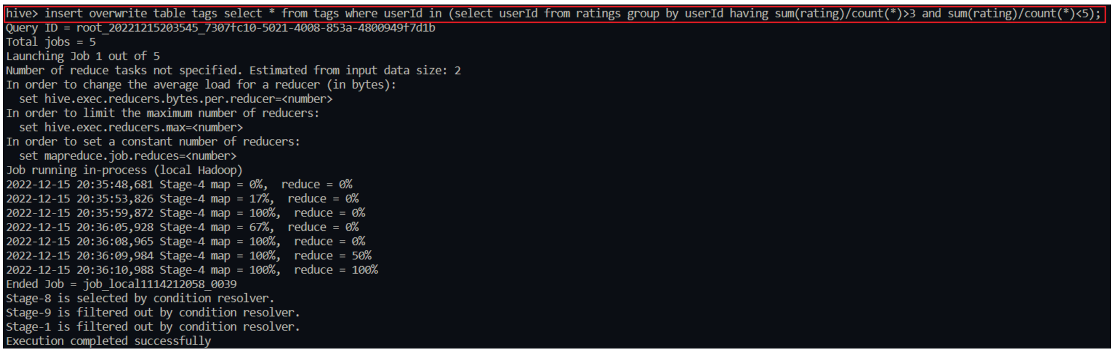
```
hive> insert overwrite table tags select * from tags where userId in (select userId from ratings group by userId having sum(rating)/count(*)>3);
Query ID = root_20221215201759_95f6989c-e083-479e-a9ca-8b3e07a98461
Total jobs = 5
Launching Job 1 out of 5
Number of reduce tasks not specified. Estimated from input data size: 2
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Job running in-process (local Hadoop)
2022-12-15 20:18:02,573 Stage-4 map = 0%,  reduce = 0%
2022-12-15 20:18:08,610 Stage-4 map = 17%,  reduce = 0%
2022-12-15 20:18:14,040 Stage-4 map = 100%,  reduce = 0%
2022-12-15 20:18:20,087 Stage-4 map = 67%,  reduce = 0%
2022-12-15 20:18:23,160 Stage-4 map = 100%,  reduce = 0%
2022-12-15 20:18:24,174 Stage-4 map = 100%,  reduce = 50%
2022-12-15 20:18:25,181 Stage-4 map = 100%,  reduce = 100%
Ended Job = job_local1874768692_0036
Stage-8 is selected by condition resolver.
Stage-9 is filtered out by condition resolver.
Stage-1 is filtered out by condition resolver.
```

（9）tags 表筛选了全给好评的用户的评价（评价平均分==5）

执行语句：

```
insert overwrite table tags select * from tags where userId in (select userId from ratings group by userId having sum(rating)/count(*)>3 and sum(rating)/count(*)<5);
```

执行结果：

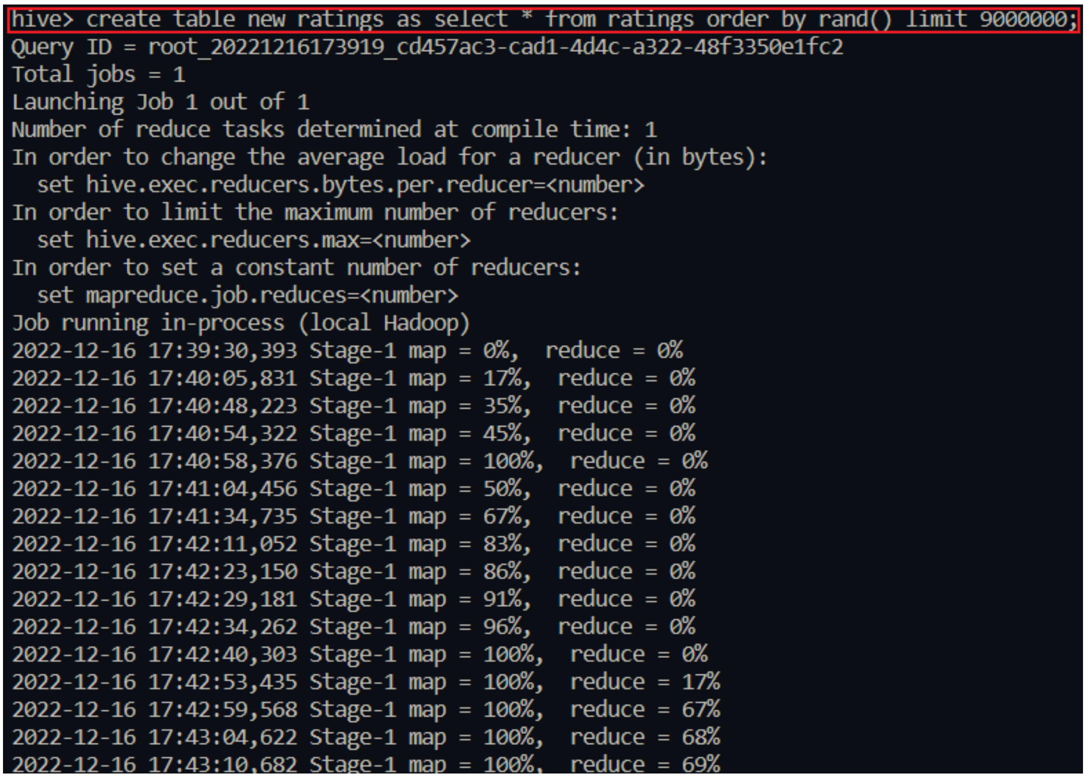


```
hive> insert overwrite table tags select * from tags where userId in (select userId from ratings group by userId having sum(rating)/count(*)>3 and sum(rating)/count(*)<5);
Query ID = root_20221215203545_7307fc10-5021-4008-853a-4800949f7d1b
Total jobs = 5
Launching Job 1 out of 5
Number of reduce tasks not specified. Estimated from input data size: 2
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Job running in-process (local Hadoop)
2022-12-15 20:35:48,681 Stage-4 map = 0%,  reduce = 0%
2022-12-15 20:35:53,826 Stage-4 map = 17%,  reduce = 0%
2022-12-15 20:35:59,872 Stage-4 map = 100%,  reduce = 0%
2022-12-15 20:36:05,928 Stage-4 map = 67%,  reduce = 0%
2022-12-15 20:36:08,965 Stage-4 map = 100%,  reduce = 0%
2022-12-15 20:36:09,984 Stage-4 map = 100%,  reduce = 50%
2022-12-15 20:36:10,988 Stage-4 map = 100%,  reduce = 100%
Ended Job = job_local1114212058_0039
Stage-8 is selected by condition resolver.
Stage-9 is filtered out by condition resolver.
Stage-1 is filtered out by condition resolver.
Execution completed successfully
```

（10）数据裁剪（数据随机排序，然后截取数据）

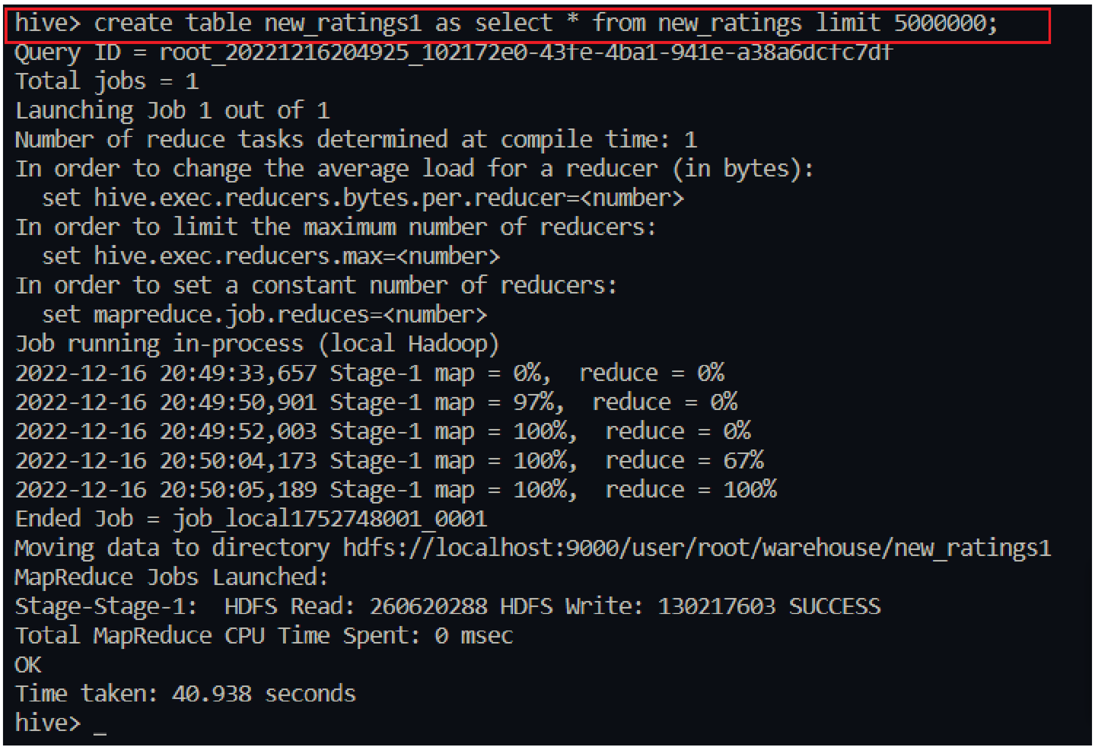
（共有 18000000 条数据，没截取算法跑了很多次，内存调到 15G 都崩了）

（截取一半的数据）



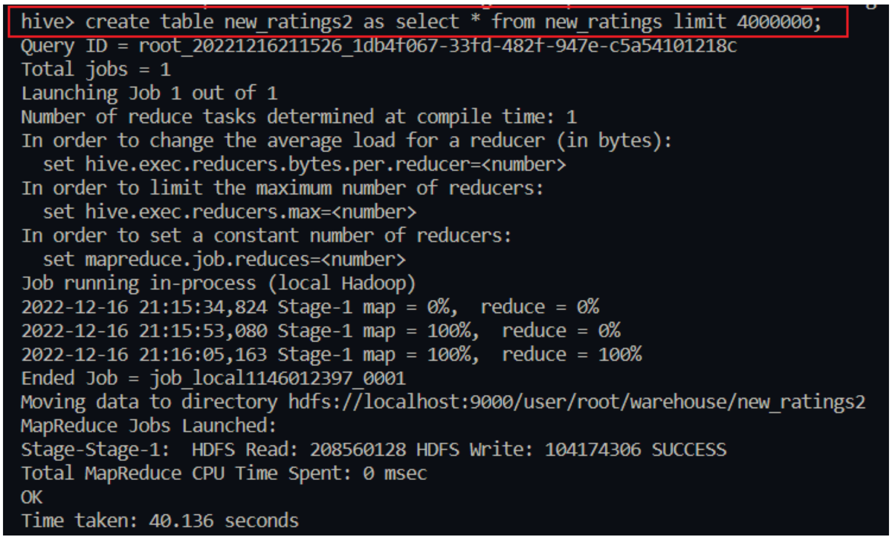
```
hive> create table new ratings as select * from ratings order by rand() limit 9000000;
Query ID = root_20221216173919_cd457ac3-cad1-4d4c-a322-48f3350e1fc2
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Job running in-process (local Hadoop)
2022-12-16 17:39:30,393 Stage-1 map = 0%,  reduce = 0%
2022-12-16 17:40:05,831 Stage-1 map = 17%,  reduce = 0%
2022-12-16 17:40:48,223 Stage-1 map = 35%,  reduce = 0%
2022-12-16 17:40:54,322 Stage-1 map = 45%,  reduce = 0%
2022-12-16 17:40:58,376 Stage-1 map = 100%,  reduce = 0%
2022-12-16 17:41:04,456 Stage-1 map = 50%,  reduce = 0%
2022-12-16 17:41:34,735 Stage-1 map = 67%,  reduce = 0%
2022-12-16 17:42:11,052 Stage-1 map = 83%,  reduce = 0%
2022-12-16 17:42:23,150 Stage-1 map = 86%,  reduce = 0%
2022-12-16 17:42:29,181 Stage-1 map = 91%,  reduce = 0%
2022-12-16 17:42:34,262 Stage-1 map = 96%,  reduce = 0%
2022-12-16 17:42:40,303 Stage-1 map = 100%,  reduce = 0%
2022-12-16 17:42:53,435 Stage-1 map = 100%,  reduce = 17%
2022-12-16 17:42:59,568 Stage-1 map = 100%,  reduce = 67%
2022-12-16 17:43:04,622 Stage-1 map = 100%,  reduce = 68%
2022-12-16 17:43:10,682 Stage-1 map = 100%,  reduce = 69%
```

(还是崩溃，降到了 5000000 条数据)



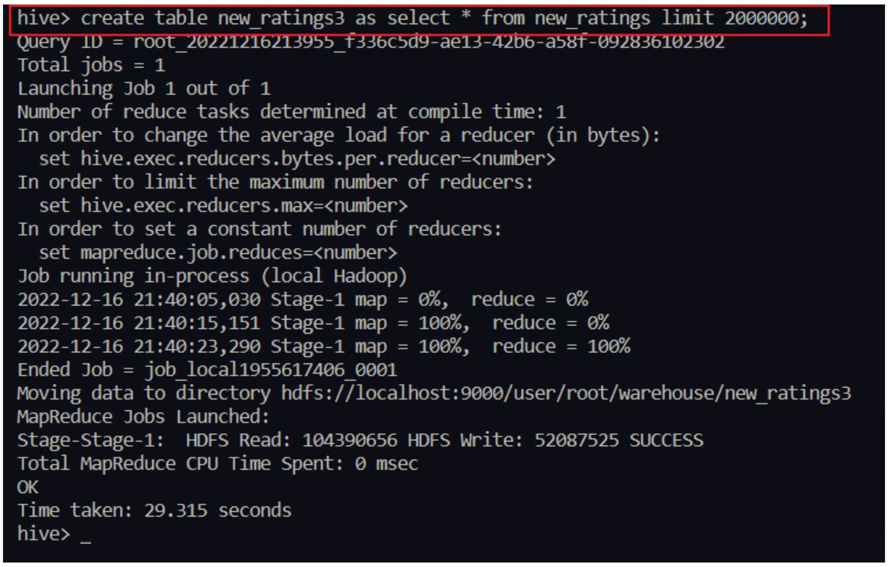
```
hive> create table new_ratings1 as select * from new_ratings limit 5000000;
Query ID = root_20221216204925_102172e0-43fe-4ba1-941e-a38a6dcfc7df
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Job running in-process (local Hadoop)
2022-12-16 20:49:33,657 Stage-1 map = 0%,  reduce = 0%
2022-12-16 20:49:50,901 Stage-1 map = 97%,  reduce = 0%
2022-12-16 20:49:52,003 Stage-1 map = 100%,  reduce = 0%
2022-12-16 20:50:04,173 Stage-1 map = 100%,  reduce = 67%
2022-12-16 20:50:05,189 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local1752748001_0001
Moving data to directory hdfs://localhost:9000/user/root/warehouse/new_ratings1
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 260620288 HDFS Write: 130217603 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
Time taken: 40.938 seconds
hive> _
```

(还是崩溃，降到了 4000000 条数据)



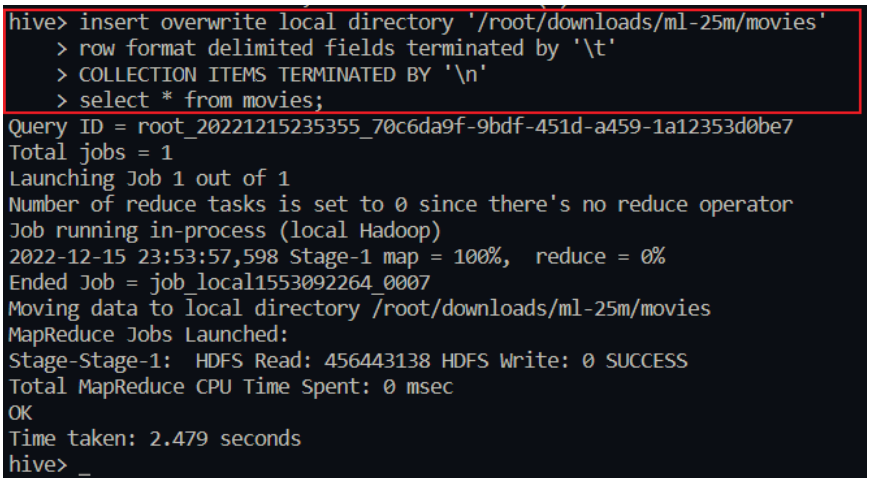
```
hive> create table new_ratings2 as select * from new_ratings limit 4000000;
Query ID = root_20221216211526_1db4f067-33fd-482f-947e-c5a54101218c
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Job running in-process (local Hadoop)
2022-12-16 21:15:34,824 Stage-1 map = 0%,  reduce = 0%
2022-12-16 21:15:53,080 Stage-1 map = 100%,  reduce = 0%
2022-12-16 21:16:05,163 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local1146012397_0001
Moving data to directory hdfs://localhost:9000/user/root/warehouse/new_ratings2
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 208560128 HDFS Write: 104174306 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
Time taken: 40.136 seconds
```

(还是崩溃，降到了 2000000 条数据)



```
hive> create table new_ratings3 as select * from new_ratings limit 2000000;
Query ID = root_20221216213955_f336c5d9-ae13-42b6-a58f-092836102302
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Job running in-process (local Hadoop)
2022-12-16 21:40:05,030 Stage-1 map = 0%,  reduce = 0%
2022-12-16 21:40:15,151 Stage-1 map = 100%,  reduce = 0%
2022-12-16 21:40:23,290 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local1955617406_0001
Moving data to directory hdfs://localhost:9000/user/root/warehouse/new_ratings3
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 104390656 HDFS Write: 52087525 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
Time taken: 29.315 seconds
hive> _
```

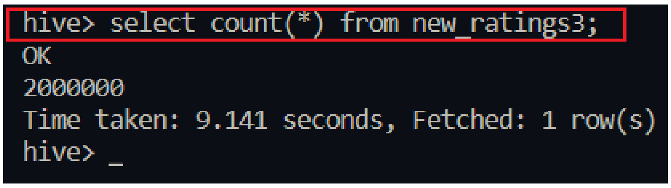
（10）数据持久化存储（将所有的表导出存在本地,导出过程相似，只展示导出 movies 表的过程）



```
hive> insert overwrite local directory '/root/downloads/ml-25m/movies'
    > row format delimited fields terminated by '\t'
    > COLLECTION ITEMS TERMINATED BY '\n'
    > select * from movies;
Query ID = root_20221215235355_70c6da9f-9bdf-451d-a459-1a12353d0be7
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Job running in-process (local Hadoop)
2022-12-15 23:53:57,598 Stage-1 map = 100%,  reduce = 0%
Ended Job = job_local1553092264_0007
Moving data to local directory /root/downloads/ml-25m/movies
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 456443138 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
Time taken: 2.479 seconds
hive> _
```

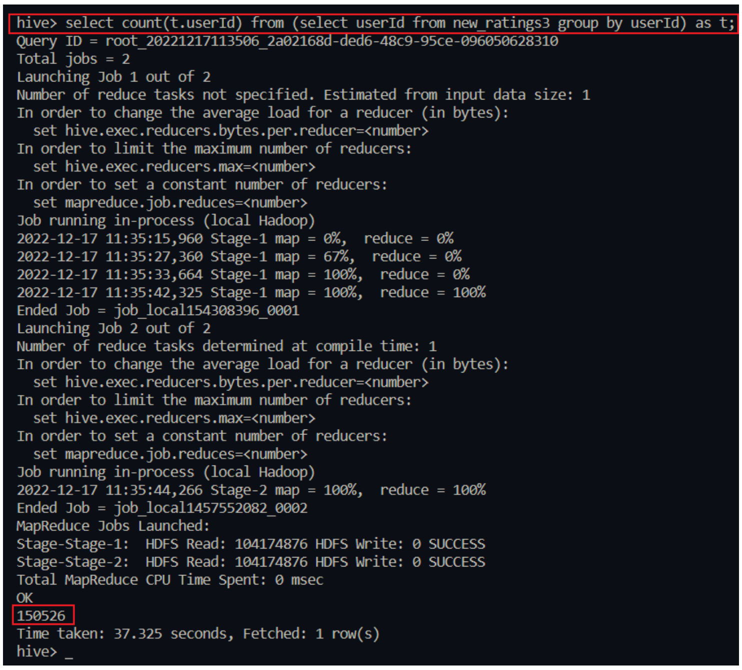
3. 详细介绍数据统计分析的过程，例如：统计、分析电影评价的数量、用户的数量、电影的数量等，统计、分析电影评价的平均值、方差等。（10 分）

（1）电影评价的数量：



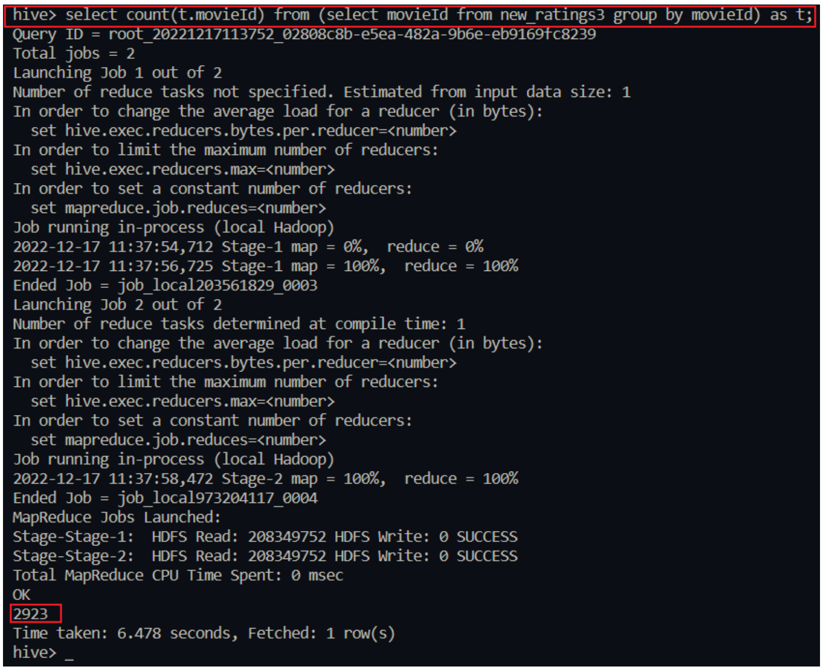
```
hive> select count(*) from new_ratings3;
OK
2000000
Time taken: 9.141 seconds, Fetched: 1 row(s)
hive> _
```

（2）用户的数量：



```
hive> select count(t.userId) from (select userId from new_ratings3 group by userId) as t;
Query ID = root_20221217113506_2a02168d-ded6-48c9-95ce-096050628310
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Job running in-process (local Hadoop)
2022-12-17 11:35:15,960 Stage-1 map = 0%,  reduce = 0%
2022-12-17 11:35:27,360 Stage-1 map = 67%,  reduce = 0%
2022-12-17 11:35:33,664 Stage-1 map = 100%,  reduce = 0%
2022-12-17 11:35:42,325 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local154308396_0001
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Job running in-process (local Hadoop)
2022-12-17 11:35:44,266 Stage-2 map = 100%,  reduce = 100%
Ended Job = job_local1457552082_0002
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 104174876 HDFS Write: 0 SUCCESS
Stage-Stage-2:  HDFS Read: 104174876 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
150526
Time taken: 37.325 seconds, Fetched: 1 row(s)
hive> _
```

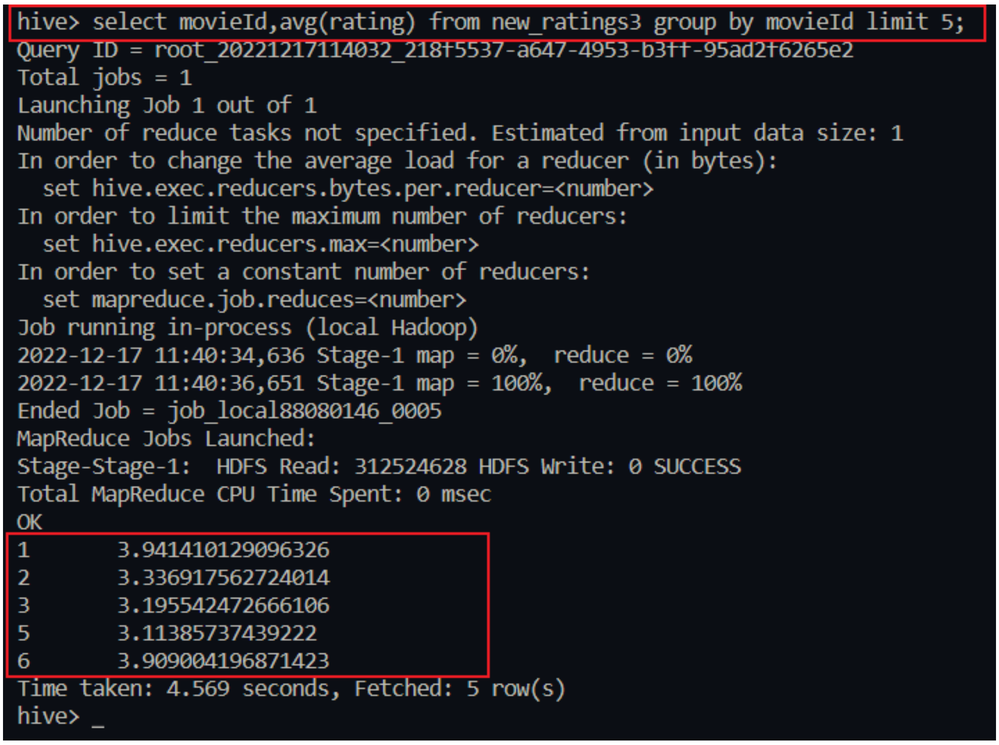
（3）电影的数量：



```
hive> select count(t.movieId) from (select movieId from new_ratings3 group by movieId) as t;
Query ID = root_20221217113752_02808c8b-e5ea-482a-9b6e-eb9169fc8239
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Job running in-process (local Hadoop)
2022-12-17 11:37:54,712 Stage-1 map = 0%,  reduce = 0%
2022-12-17 11:37:56,725 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local203561829_0003
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Job running in-process (local Hadoop)
2022-12-17 11:37:58,472 Stage-2 map = 100%,  reduce = 100%
Ended Job = job_local973204117_0004
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 208349752 HDFS Write: 0 SUCCESS
Stage-Stage-2:  HDFS Read: 208349752 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
2923
Time taken: 6.478 seconds, Fetched: 1 row(s)
hive> _
```

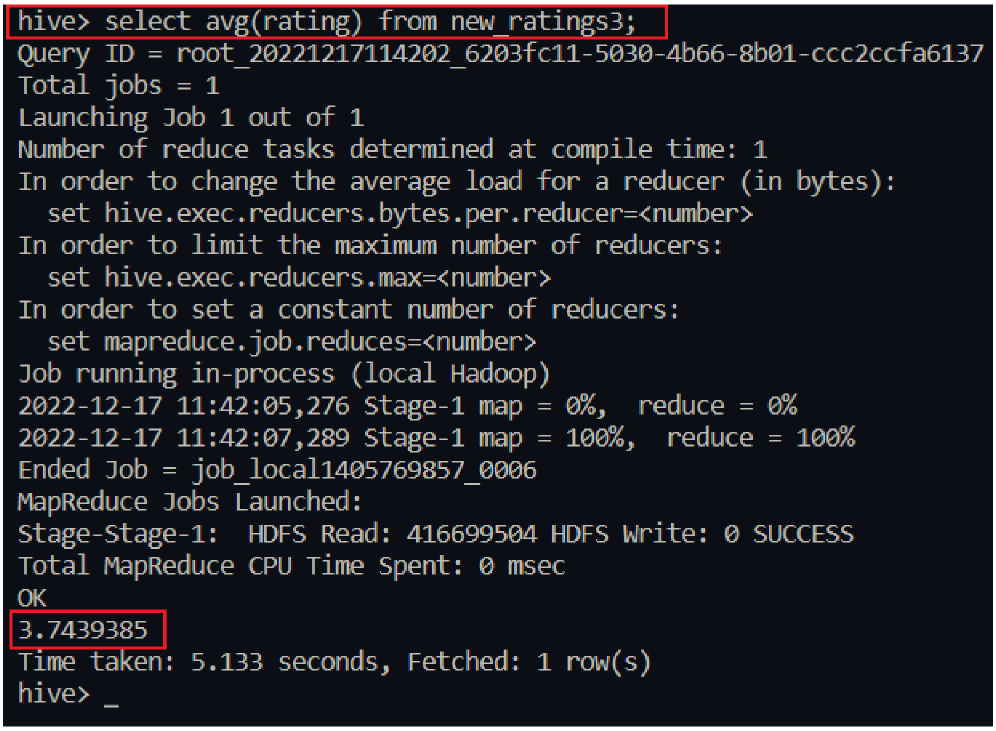
（4）电影评价的平均值

- 电影评价的每部电影的评分平均值（前 5 个）：



```
hive> select movieId,avg(rating) from new_ratings3 group by movieId limit 5;
Query ID = root_20221217114032_218f5537-a647-4953-b3ff-95ad2f6265e2
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Job running in-process (local Hadoop)
2022-12-17 11:40:34,636 Stage-1 map = 0%,  reduce = 0%
2022-12-17 11:40:36,651 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local88080146_0005
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 312524628 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
1       3.941410129096326
2       3.336917562724014
3       3.195542472666106
5       3.11385737439222
6       3.909004196871423
Time taken: 4.569 seconds, Fetched: 5 row(s)
hive> _
```

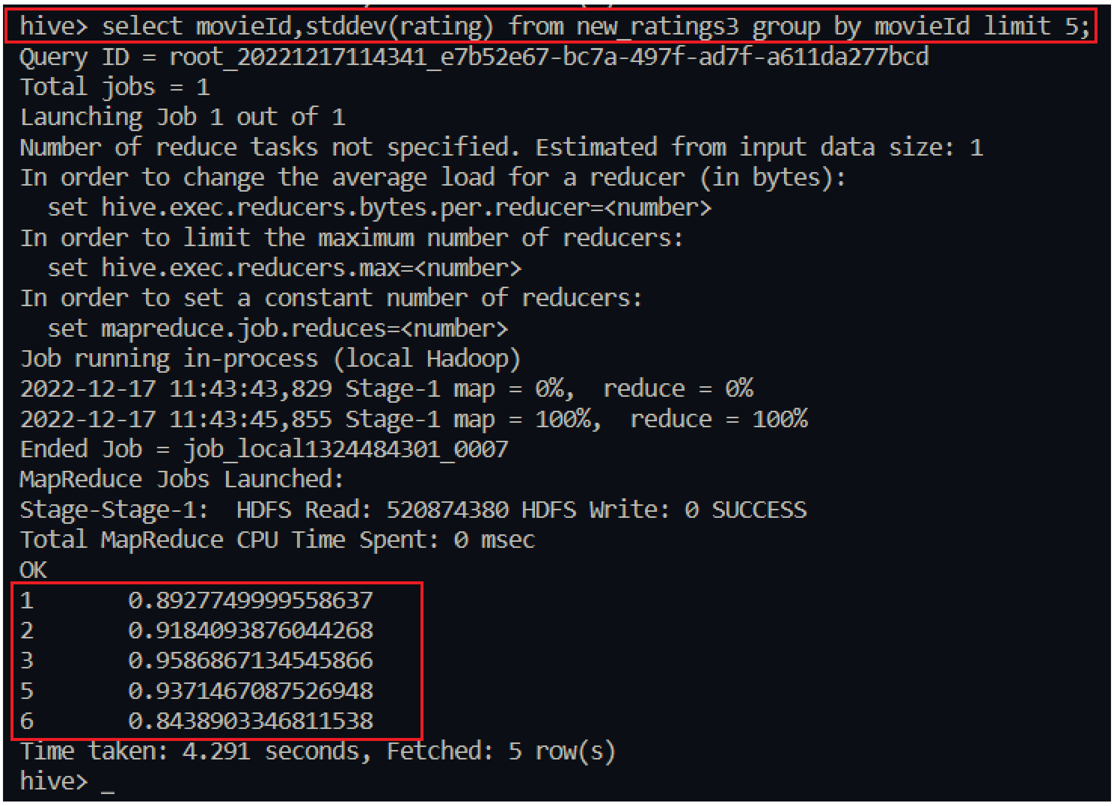
- 电影评价的所有电影评分的评价值



```
hive> select avg(rating) from new_ratings3;
Query ID = root_20221217114202_6203fc11-5030-4b66-8b01-ccc2ccfa6137
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Job running in-process (local Hadoop)
2022-12-17 11:42:05,276 Stage-1 map = 0%,  reduce = 0%
2022-12-17 11:42:07,289 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local1405769857_0006
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 416699504 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
3.7439385
Time taken: 5.133 seconds, Fetched: 1 row(s)
hive> _
```

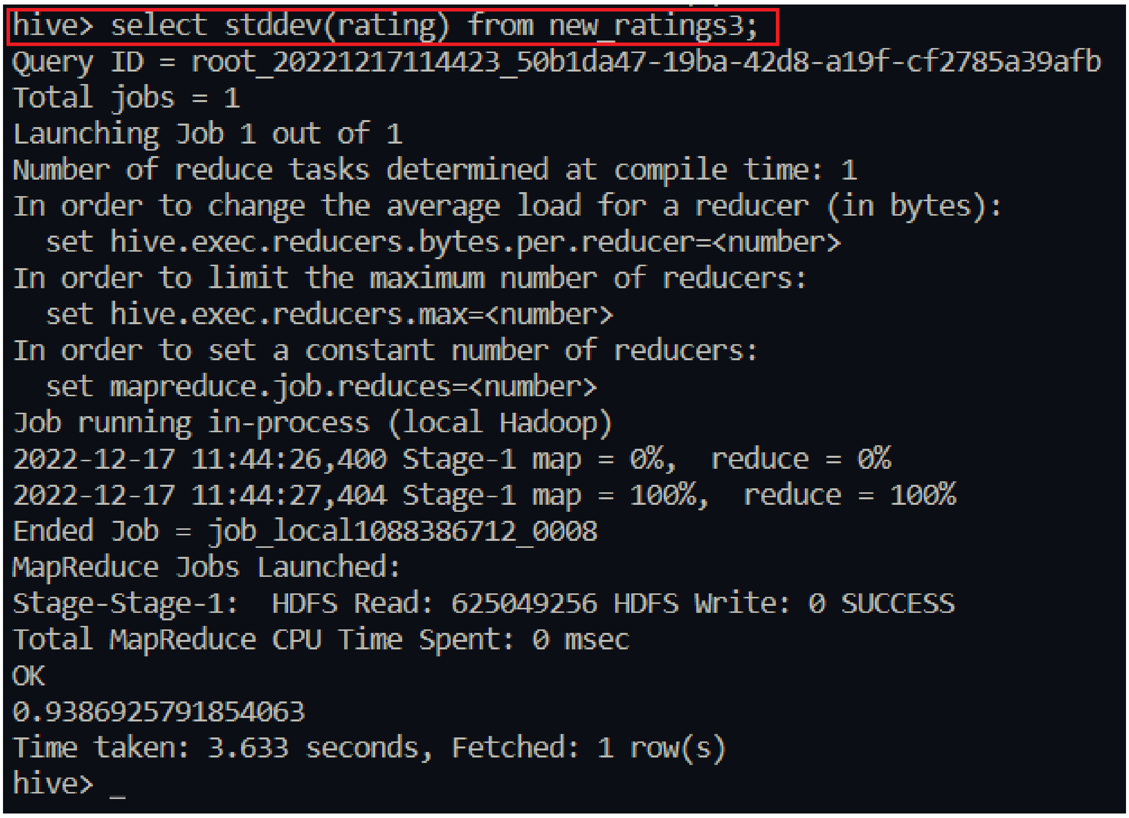
（5）电影评价的方差

- 电影评价的每部电影的评分的方差（前 5 个）：



```
hive> select movieId,stddev(rating) from new_ratings3 group by movieId limit 5;
Query ID = root_20221217114341_e7b52e67-bc7a-497f-ad7f-a611da277bcd
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Job running in-process (local Hadoop)
2022-12-17 11:43:43,829 Stage-1 map = 0%,  reduce = 0%
2022-12-17 11:43:45,855 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local1324484301_0007
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 520874380 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
1       0.8927749999558637
2       0.9184093876044268
3       0.9586867134545866
5       0.9371467087526948
6       0.8438903346811538
Time taken: 4.291 seconds, Fetched: 5 row(s)
hive> _
```

- 电影评价的所有电影评分的方差



```
hive> select stddev(rating) from new_ratings3;
Query ID = root_20221217114423_50b1da47-19ba-42d8-a19f-cf2785a39afb
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Job running in-process (local Hadoop)
2022-12-17 11:44:26,400 Stage-1 map = 0%,  reduce = 0%
2022-12-17 11:44:27,404 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local1088386712_0008
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 625049256 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
0.9386925791854063
Time taken: 3.633 seconds, Fetched: 1 row(s)
hive> _
```

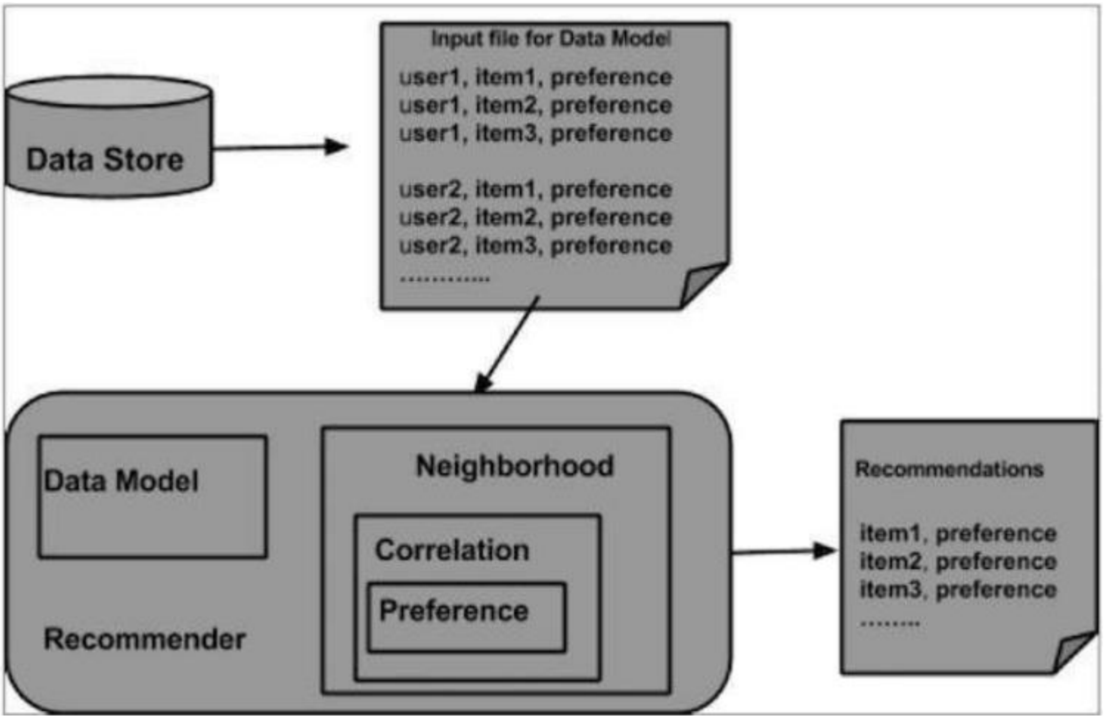
4. 详细介绍数据挖掘的算法、过程和结果。例如：使用了某个数据挖掘的算法，或编辑了某段代码，接着对多少用户和多少电影进行了协同过滤，得到了什么结果或什么成果，有什么结论。（10 分）

（1）mahout 的推荐算法：

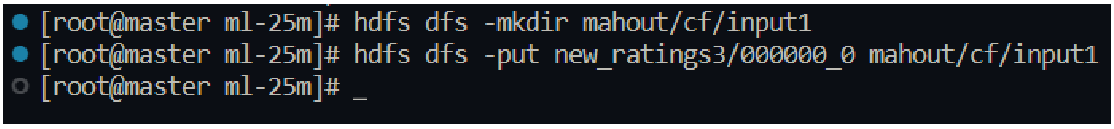
mahout 的推荐算法是基于 userbase 的,通过输入传递具有用户对项目首选项的文本文档，输出特定用户对其他项目的估计偏好。

从数据存储中，准备数据模型，并将其作为输入传递到推荐引擎。推荐器引擎为特定用户生成推荐。

推荐引擎的架构：



（2）将处理过后导出到本地的 ratings 表上传到 HDFS



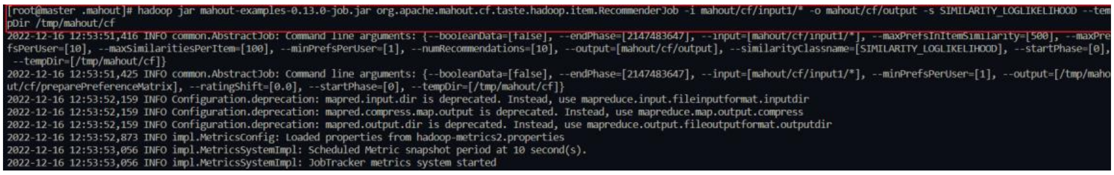
```
[root@master ml-25m]# hdfs dfs -mkdir mahout/cf/input1
[root@master ml-25m]# hdfs dfs -put new_ratings3/000000_0 mahout/cf/input1
[root@master ml-25m]# _
```

（3）使用 mahout 的推荐算法处理上述步骤中上传的文件

实验代码：

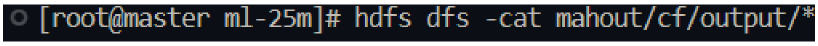
```
hadoop jar mahout-examples-0.13.0-job.jar

org.apache.mahout.cf.taste.hadoop.item.RecommenderJob -i mahout/cf/input1/* -o

mahout/cf/output -s SIMILARITY_LOGLIKELIHOOD --tempDir /tmp/mahout/cf
```

实验执行过程：

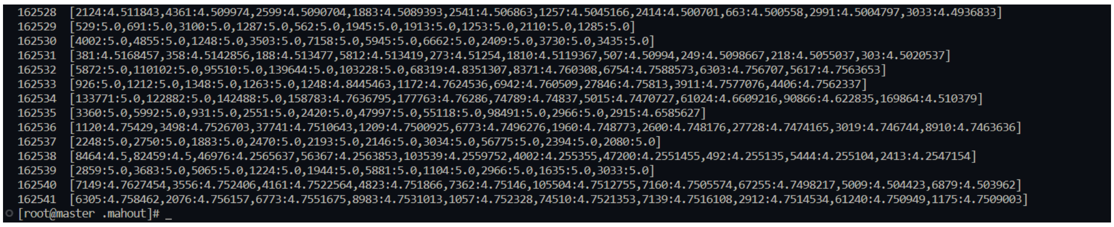


```
[root@master .mahout]# hadoop jar mahout-examples-0.13.0-job.jar org.apache.mahout.cf.taste.hadoop.item.RecommenderJob -i mahout/cf/input1/* -o mahout/cf/output -s SIMILARITY_LOGLIKELIHOOD --tem
pDir /tmp/mahout/cf
2022-12-16 12:53:51,416 INFO common.AbstractJob: Command line arguments: {--booleanData=[false], --endPhase=[2147483647], --input=[mahout/cf/input1/*], --maxPrefsInItemSimilarity=[500], --maxPre
fsPerUser=[10], --maxSimilaritiesPerItem=[100], --minPrefsPerUser=[1], --numRecommendations=[10], --output=[mahout/cf/output], --similarityClassname=[SIMILARITY_LOGLIKELIHOOD], --startPhase=[0],
 --tempDir=[/tmp/mahout/cf]}
2022-12-16 12:53:51,425 INFO common.AbstractJob: Command line arguments: {--booleanData=[false], --endPhase=[2147483647], --input=[mahout/cf/input1/*], --minPrefsPerUser=[1], --output=[/tmp/maho
ut/cf/preparePreferenceMatrix], --ratingShift=[0.0], --startPhase=[0], --tempDir=[/tmp/mahout/cf]}
2022-12-16 12:53:52,159 INFO Configuration.deprecation: mapred.input.dir is deprecated. Instead, use mapreduce.input.fileinputformat.inputdir
2022-12-16 12:53:52,159 INFO Configuration.deprecation: mapred.compress.map.output is deprecated. Instead, use mapreduce.map.output.compress
2022-12-16 12:53:52,159 INFO Configuration.deprecation: mapred.output.dir is deprecated. Instead, use mapreduce.output.fileoutputformat.outputdir
2022-12-16 12:53:52,873 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2022-12-16 12:53:53,056 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2022-12-16 12:53:53,056 INFO impl.MetricsSystemImpl: JobTracker metrics system started
```

实验结果：



```
[root@master ml-25m]# hdfs dfs -cat mahout/cf/output/*
```



```
  162528   [2124:4.511843,4361:4.509974,2599:4.5090704,1883:4.5089393,2541:4.506863,1257:4.5045166,2414:4.500701,663:4.500558,2991:4.5004797,3033:4.4936833]
  162529   [529:5.0,691:5.0,3100:5.0,1287:5.0,562:5.0,1945:5.0,1913:5.0,1253:5.0,2110:5.0,1285:5.0]
  162530   [4002:5.0,4855:5.0,1248:5.0,3503:5.0,7158:5.0,5945:5.0,6662:5.0,2409:5.0,3730:5.0,3435:5.0]
  162531   [381:4.5168457,358:4.5142856,188:4.513477,5812:4.513419,273:4.51254,1810:4.5119367,507:4.50994,249:4.5098667,218:4.5055037,303:4.5020537]
  162532   [5872:5.0,110102:5.0,95510:5.0,139644:5.0,103228:5.0,68319:4.8351307,8371:4.760308,6754:4.7588573,6303:4.756707,5617:4.7563653]
  162533   [926:5.0,1212:5.0,1348:5.0,1263:5.0,1248:4.8445463,1172:4.7624536,6942:4.760509,27846:4.75813,3911:4.7577076,4406:4.7562337]
  162534   [133771:5.0,122882:5.0,142488:5.0,158783:4.7636795,177763:4.76286,74789:4.74837,5015:4.7470727,61024:4.6609216,90866:4.622835,169864:4.510379]
  162535   [3360:5.0,5992:5.0,931:5.0,2551:5.0,2420:5.0,47997:5.0,55118:5.0,98491:5.0,2966:5.0,2915:4.6585627]
  162536   [1120:4.75429,3498:4.7526703,37741:4.7510643,1209:4.7500925,6773:4.7496276,1960:4.748773,2600:4.748176,27728:4.7474165,3019:4.746744,8910:4.7463636]
  162537   [2248:5.0,2750:5.0,1883:5.0,2470:5.0,2193:5.0,2146:5.0,3034:5.0,56775:5.0,2394:5.0,2080:5.0]
  162538   [8464:4.5,82459:4.5,46976:4.2565637,56367:4.2563853,103539:4.2559752,4002:4.255355,47200:4.2551455,492:4.255135,5444:4.255104,2413:4.2547154]
  162539   [2859:5.0,3683:5.0,5065:5.0,1224:5.0,1944:5.0,5881:5.0,1104:5.0,2966:5.0,1635:5.0,3033:5.0]
  162540   [7149:4.7627454,3556:4.752406,4161:4.7522564,4823:4.751866,7362:4.75146,105504:4.7512755,7160:4.7505574,67255:4.7498217,5009:4.504423,6879:4.503962]
  162541   [6305:4.758462,2076:4.756157,6773:4.7551675,8983:4.7531013,1057:4.752328,74510:4.7521353,7139:4.7516108,2912:4.7514534,61240:4.750949,1175:4.7509003]
[root@master .mahout]# _
```

（4）结论：

根据上面‘3’的数据统计分析可知，推荐算法对 150526 名用户和 2923 部电影进行了协同过滤，得到了为 150526 名用户推荐的不同类型和不同个数的电影列表。