实验3： MapReduce和Spark编程实验

1. 实验任务

1. MapReduce：
    - 统计各省的双十一前十热门关注产品（“点击+添加购物车+购买+关注”总量最多前10的产品）
    - 统计各省的双十一前十热门销售产品（购买最多前10的产品）
2. Hive
    - 把精简数据集导入到数据仓库Hive中，并对数据仓库Hive中的数据进行查询分析
    - 查询双11那天有多少人购买了商品
    - 查询双11那天男女买家购买商品的比例
    - 查询双11那天浏览次数前十的品牌
3. Spark：
    - 统计各省销售最好的产品类别前十（销售最多前10的产品类别）
    - 统计各省的双十一前十热门销售产品（购买最多前10的产品）-- 和MapReduce作业对比结果
    - 查询双11那天浏览次数前十的品牌 -- 和Hive作业对比结果
4. 数据挖掘：
    - 针对预处理后的训练集和测试集，基于MapReduce或Spark MLlib编写程序预测回头客
    - 评估预测准确率

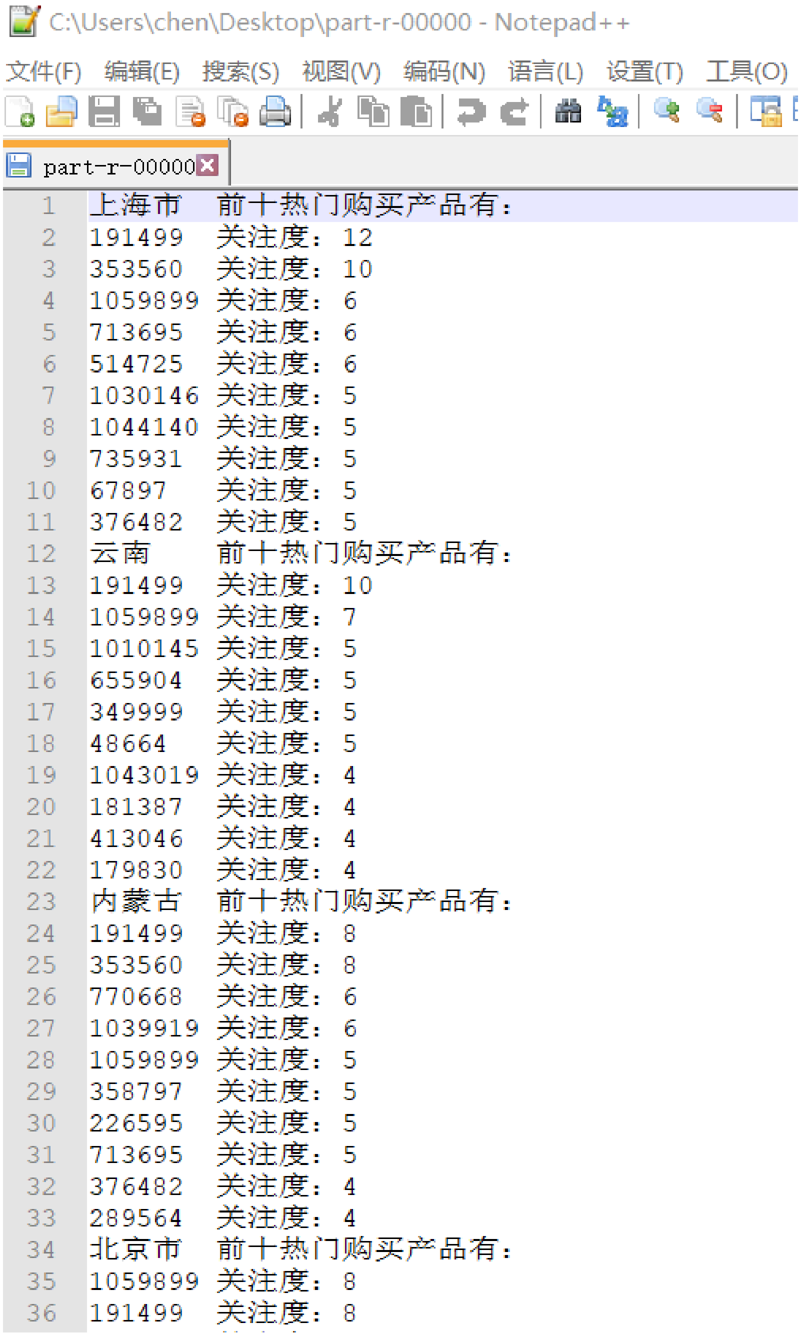
2. 实验环境

- Java 1.8+
- Hadoop 3.2.x
- Spark 2.4.x
- Hive 2.3.x

3. 实验过程

3.1 MapReduce（代码分别见Attention.java和Sell.java）

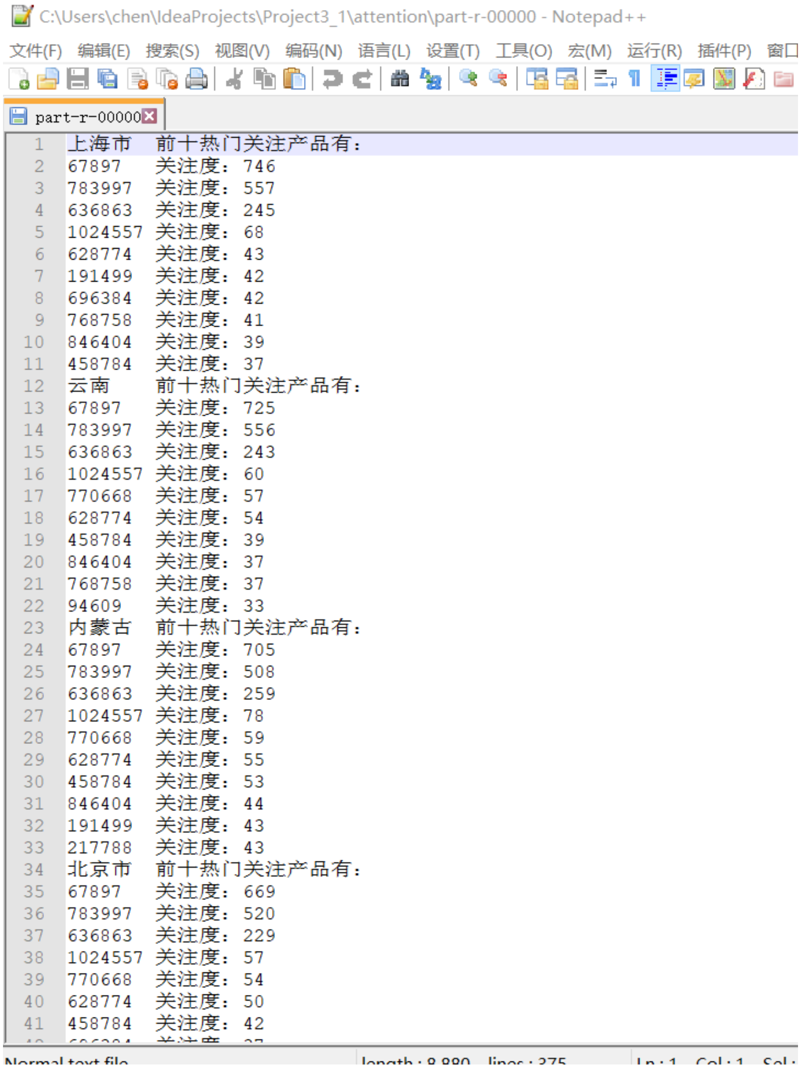
- 统计各省的双十一前十热门关注产品


C:\Users\chen\Desktop\part-r-00000 - Notepad++
文件(F) 编辑(E) 搜索(S) 视图(V) 编码(N) 语言(L) 设置(T) 工具(O)
part-r-00000

```
上海市    前十热门购买产品有：
191499   关注度：12
353560   关注度：10
1059899  关注度：6
713695   关注度：6
514725   关注度：6
1030146  关注度：5
1044140  关注度：5
735931   关注度：5
67897    关注度：5
376482   关注度：5
云南      前十热门购买产品有：
191499   关注度：10
1059899  关注度：7
1010145  关注度：5
655904   关注度：5
349999   关注度：5
48664    关注度：5
1043019  关注度：4
181387   关注度：4
413046   关注度：4
179830   关注度：4
内蒙古    前十热门购买产品有：
191499   关注度：8
353560   关注度：8
770668   关注度：6
1039919  关注度：6
1059899  关注度：5
358797   关注度：5
226595   关注度：5
713695   关注度：5
376482   关注度：4
289564   关注度：4
北京市    前十热门购买产品有：
1059899  关注度：8
191499   关注度：8
```

命令行参数：<输入文件路径>

- 统计各省的双十一前十热门销售产品

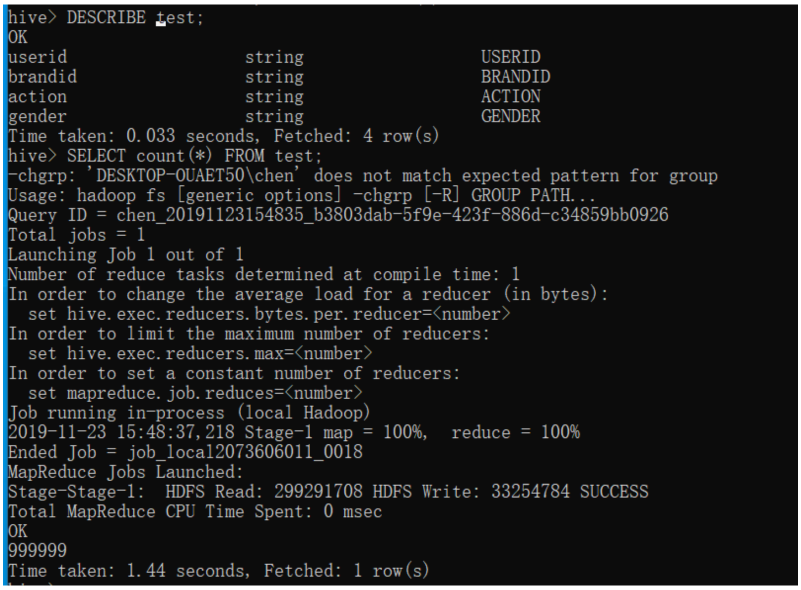

C:\Users\chen\IdeaProjects\Project3_1\attention\part-r-00000 - Notepad++
文件(F) 编辑(E) 搜索(S) 视图(V) 编码(N) 语言(L) 设置(T) 工具(O) 宏(M) 运行(R) 插件(P) 窗口
part-r-00000

```
上海市   前十热门关注产品有：
67897    关注度：746
783997   关注度：557
636863   关注度：245
1024557  关注度：68
628774   关注度：43
191499   关注度：42
696384   关注度：42
768758   关注度：41
846404   关注度：39
458784   关注度：37
云南     前十热门关注产品有：
67897    关注度：725
783997   关注度：556
636863   关注度：243
1024557  关注度：60
770668   关注度：57
628774   关注度：54
458784   关注度：39
846404   关注度：37
768758   关注度：37
94609    关注度：33
内蒙古   前十热门关注产品有：
67897    关注度：705
783997   关注度：508
636863   关注度：259
1024557  关注度：78
770668   关注度：59
628774   关注度：55
458784   关注度：53
846404   关注度：44
191499   关注度：43
217788   关注度：43
北京市   前十热门关注产品有：
67897    关注度：669
783997   关注度：520
636863   关注度：229
1024557  关注度：57
770668   关注度：54
628774   关注度：50
458784   关注度：42
```

```
命令行参数：<输入文件路径>
```

3.2 Hive：

- 把精简数据集导入到数据仓库Hive中

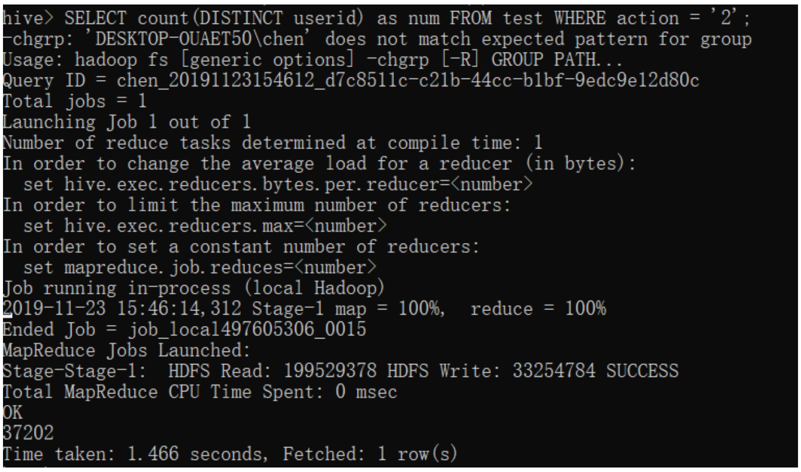


```
hive> DESCRIBE test;
OK
userid                  string                  USERID
brandid                 string                  BRANDID
action                  string                  ACTION
gender                  string                  GENDER
Time taken: 0.033 seconds, Fetched: 4 row(s)
hive> SELECT count(*) FROM test;
-chgrp: 'DESKTOP-OUAET50\chen' does not match expected pattern for group
Usage: hadoop fs [generic options] -chgrp [-R] GROUP PATH...
Query ID = chen_20191123154835_b3803dab-5f9e-423f-886d-c34859bb0926
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)
2019-11-23 15:48:37,218 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local2073606011_0018
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 299291708 HDFS Write: 33254784 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
999999
Time taken: 1.44 seconds, Fetched: 1 row(s)
```

- 查询双11那天有多少人购买了商品

```
SELECT count(DISTINCT userid) as num FROM test
WHERE action = '2';
```

输出37202

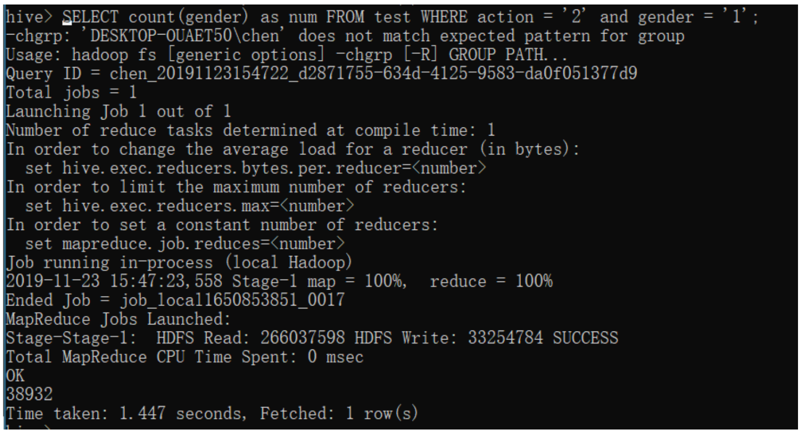


```
hive> SELECT count(DISTINCT userid) as num FROM test WHERE action = '2';
-chgrp: 'DESKTOP-OUAET50\chen' does not match expected pattern for group
Usage: hadoop fs [generic options] -chgrp [-R] GROUP PATH...
Query ID = chen_20191123154612_d7c8511c-c21b-44cc-b1bf-9edc9e12d80c
Total jobs = 1
Launching 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)
2019-11-23 15:46:14,312 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local497605306_0015
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 199529378 HDFS Write: 33254784 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
37202
Time taken: 1.466 seconds, Fetched: 1 row(s)
```

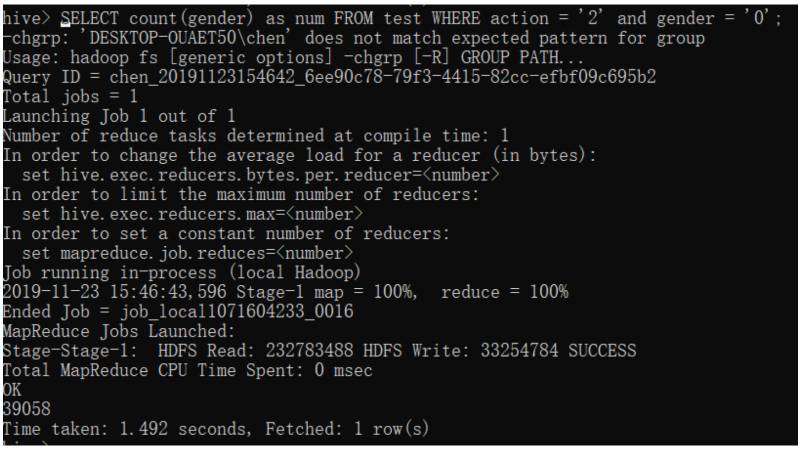
- 查询双11那天男女买家购买商品的比例

```
SELECT count(gender) as num FROM test WHERE
action = '2' and gender = '0';
SELECT count(gender) as num FROM test WHERE
action = '2' and gender = '1';
```

分别输出38932和39058



```
hive> SELECT count(gender) as num FROM test WHERE action = '2' and gender = '1';
-chgrp: 'DESKTOP-OUAET50\chen' does not match expected pattern for group
Usage: hadoop fs [generic options] -chgrp [-R] GROUP PATH...
Query ID = chen_20191123154722_d2871755-634d-4125-9583-da0f051377d9
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)
2019-11-23 15:47:23,558 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local1650853851_0017
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 266037598 HDFS Write: 33254784 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
38932
Time taken: 1.447 seconds, Fetched: 1 row(s)
```



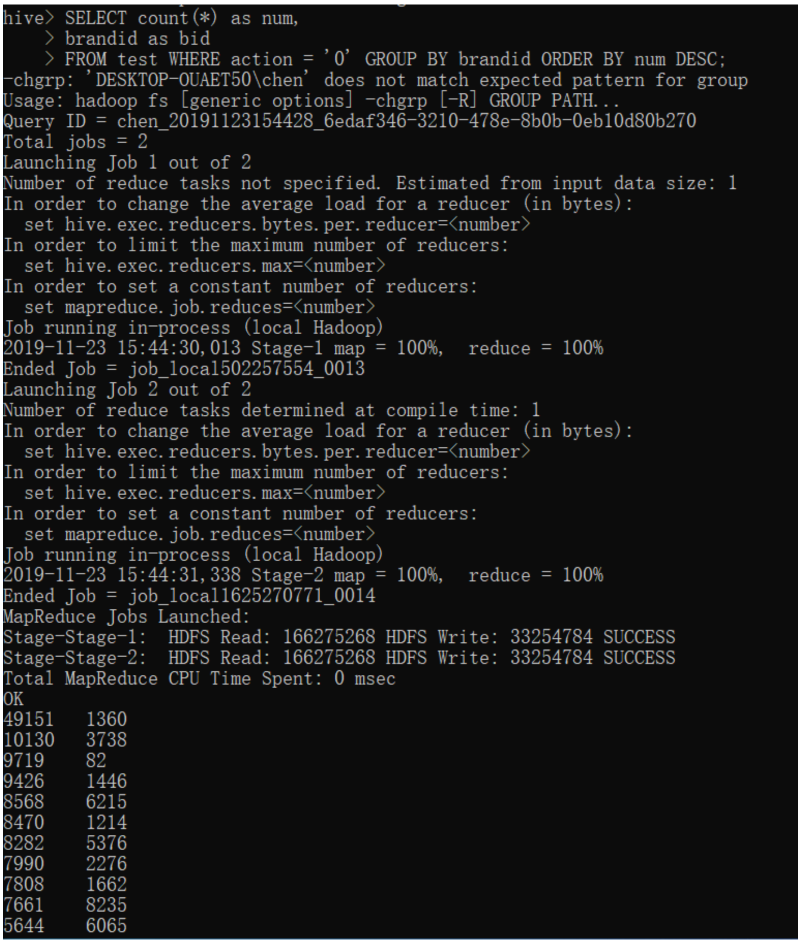
```
hive> SELECT count(gender) as num FROM test WHERE action = '2' and gender = '0';
-chgrp: 'DESKTOP-OUAET50\chen' does not match expected pattern for group
Usage: hadoop fs [generic options] -chgrp [-R] GROUP PATH...
Query ID = chen_20191123154642_6ee90c78-79f3-4415-82cc-efbf09c695b2
Total jobs = 1
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)
2019-11-23 15:46:43,596 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local1071604233_0016
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 232783488 HDFS Write: 33254784 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
39058
Time taken: 1.492 seconds, Fetched: 1 row(s)
```

- 查询双11那天浏览次数前十的品牌

```
SELECT count(*) as num brandid as bid FROM test
WHERE action = '0' GROUP BY brandid ORDER BY num
DESC;
```

数量	品牌

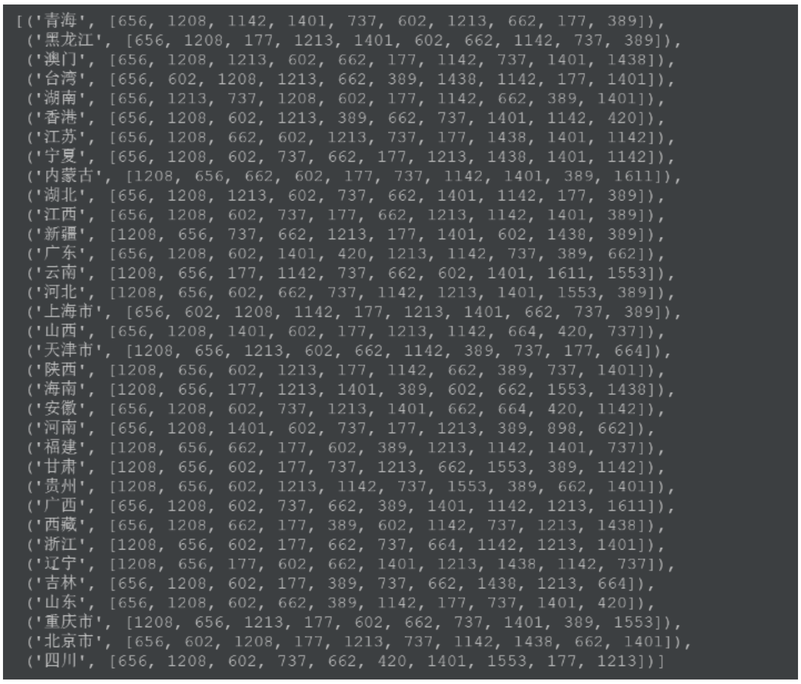
数量	品牌
49151	1360
10130	3738
9719	82
9426	1446
8568	6215
8470	1214
8282	5376
7990	2276
7808	1662
7661	8235



```
hive> SELECT count(*) as num,
    > brandid as bid
    > FROM test WHERE action = '0' GROUP BY brandid ORDER BY num DESC;
-chgrp: 'DESKTOP-OUAET50\chen' does not match expected pattern for group
Usage: hadoop fs [generic options] -chgrp [-R] GROUP PATH...
Query ID = chen_20191123154428_6edaf346-3210-478e-8b0b-0eb10d80b270
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)
2019-11-23 15:44:30,013 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local502257554_0013
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)
2019-11-23 15:44:31,338 Stage-2 map = 100%,  reduce = 100%
Ended Job = job_local1625270771_0014
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 166275268 HDFS Write: 33254784 SUCCESS
Stage-Stage-2:  HDFS Read: 166275268 HDFS Write: 33254784 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
49151   1360
10130   3738
9719    82
9426    1446
8568    6215
8470    1214
8282    5376
7990    2276
7808    1662
7661    8235
5644    6065
```

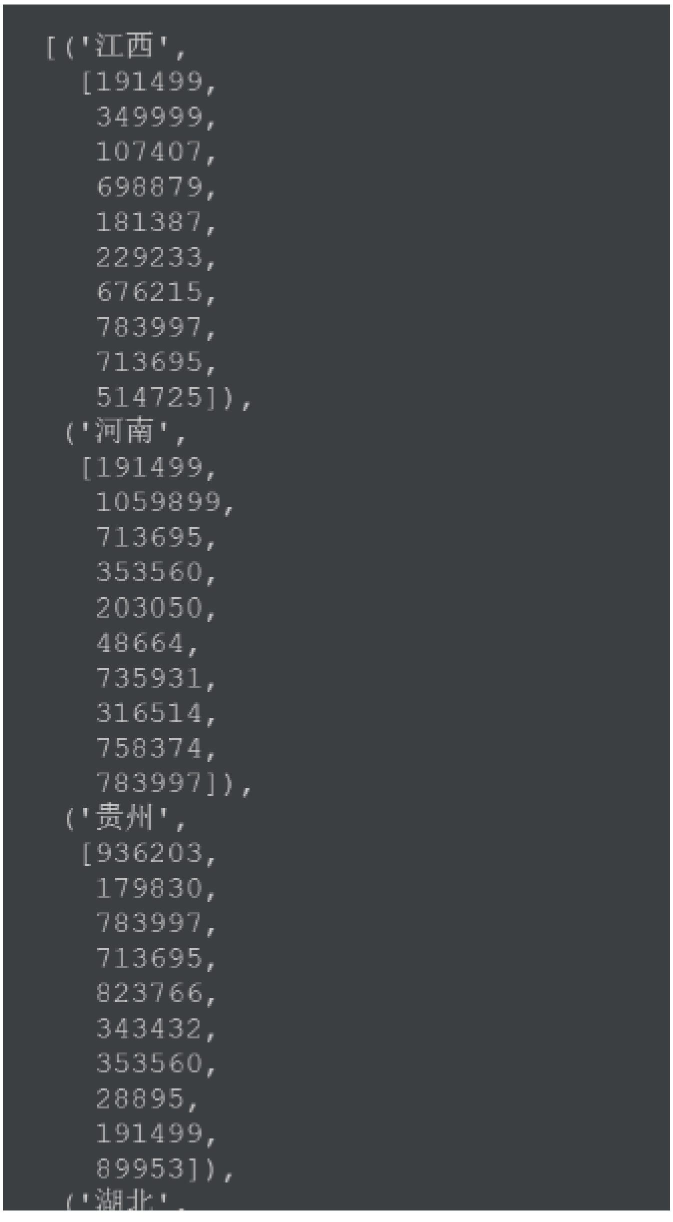
3.3 Spark（代码见task3.py）

- 统计各省销售最好的产品类别前十（销售最多前10的产品类别）



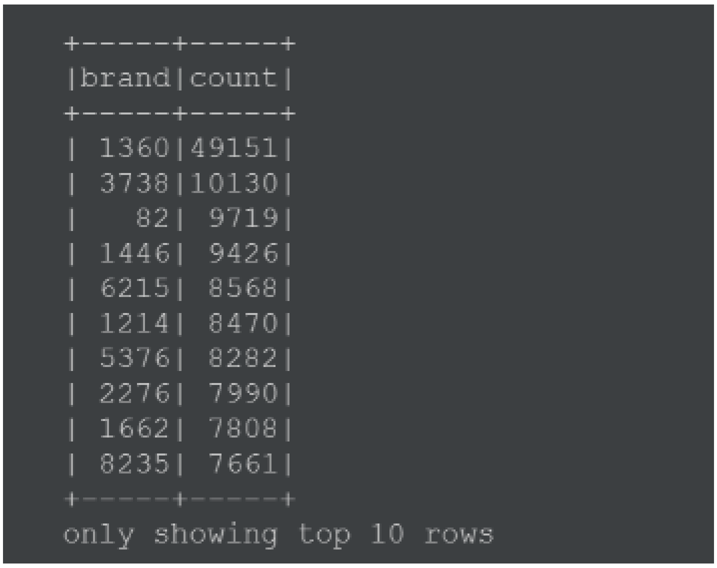
```
[('青海', [656, 1208, 1142, 1401, 737, 602, 1213, 662, 177, 389]),
 ('黑龙江', [656, 1208, 177, 1213, 1401, 602, 662, 1142, 737, 389]),
 ('澳门', [656, 1208, 1213, 602, 662, 177, 1142, 737, 1401, 1438]),
 ('台湾', [656, 602, 1208, 1213, 662, 389, 1438, 1142, 177, 1401]),
 ('湖南', [656, 1213, 737, 1208, 602, 177, 1142, 662, 389, 1401]),
 ('香港', [656, 1208, 602, 1213, 389, 662, 737, 1401, 1142, 420]),
 ('江苏', [656, 1208, 662, 602, 1213, 737, 177, 1438, 1401, 1142]),
 ('宁夏', [656, 1208, 602, 737, 662, 177, 1213, 1438, 1401, 1142]),
 ('内蒙古', [1208, 656, 662, 602, 177, 737, 1142, 1401, 389, 1611]),
 ('湖北', [656, 1208, 1213, 602, 737, 662, 1401, 1142, 177, 389]),
 ('江西', [656, 1208, 602, 737, 177, 662, 1213, 1142, 1401, 389]),
 ('新疆', [1208, 656, 737, 662, 1213, 177, 1401, 602, 1438, 389]),
 ('广东', [656, 1208, 602, 1401, 420, 1213, 1142, 737, 389, 662]),
 ('云南', [1208, 656, 177, 1142, 737, 662, 602, 1401, 1611, 1553]),
 ('河北', [1208, 656, 602, 662, 737, 1142, 1213, 1401, 1553, 389]),
 ('上海市', [656, 602, 1208, 1142, 177, 1213, 1401, 662, 737, 389]),
 ('山西', [656, 1208, 1401, 602, 177, 1213, 1142, 664, 420, 737]),
 ('天津市', [1208, 656, 1213, 602, 662, 1142, 389, 737, 177, 664]),
 ('陕西', [1208, 656, 602, 1213, 177, 1142, 662, 389, 737, 1401]),
 ('海南', [1208, 656, 177, 1213, 1401, 389, 602, 662, 1553, 1438]),
 ('安徽', [656, 1208, 602, 737, 1213, 1401, 662, 664, 420, 1142]),
 ('河南', [656, 1208, 1401, 602, 737, 177, 1213, 389, 898, 662]),
 ('福建', [1208, 656, 662, 177, 602, 389, 1213, 1142, 1401, 737]),
 ('甘肃', [1208, 656, 602, 177, 737, 1213, 662, 1553, 389, 1142]),
 ('贵州', [1208, 656, 602, 1213, 1142, 737, 1553, 389, 662, 1401]),
 ('广西', [656, 1208, 602, 737, 662, 389, 1401, 1142, 1213, 1611]),
 ('西藏', [656, 1208, 662, 177, 389, 602, 1142, 737, 1213, 1438]),
 ('浙江', [1208, 656, 602, 177, 662, 737, 664, 1142, 1213, 1401]),
 ('辽宁', [1208, 656, 177, 602, 662, 1401, 1213, 1438, 1142, 737]),
 ('吉林', [656, 1208, 602, 177, 389, 737, 662, 1438, 1213, 664]),
 ('山东', [656, 1208, 602, 662, 389, 1142, 177, 737, 1401, 420]),
 ('重庆市', [1208, 656, 1213, 177, 602, 662, 737, 1401, 389, 1553]),
 ('北京市', [656, 602, 1208, 177, 1213, 737, 1142, 1438, 662, 1401]),
 ('四川', [656, 1208, 602, 737, 662, 420, 1401, 1553, 177, 1213])]
```

- 统计各省的双十一前十热门销售产品（购买最多前10的产品）-- 和MapReduce作业对比结果



```
[('江西',
  [191499,
   349999,
   107407,
   698879,
   181387,
   229233,
   676215,
   783997,
   713695,
   514725]),
 ('河南',
  [191499,
   1059899,
   713695,
   353560,
   203050,
   48664,
   735931,
   316514,
   758374,
   783997]),
 ('贵州',
  [936203,
   179830,
   783997,
   713695,
   823766,
   343432,
   353560,
   28895,
   191499,
   89953]),
 ('湖北',
```

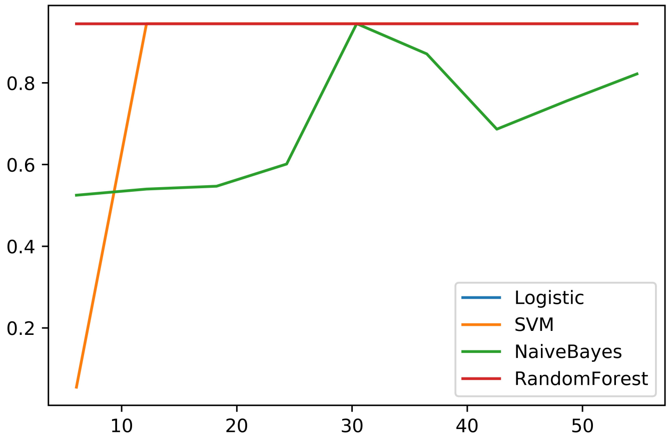
- 查询双11那天浏览次数前十的品牌 -- 和Hive作业对比结果



```
+-----+-----+
|brand|count|
+-----+-----+
| 1360|49151|
| 3738|10130|
|   82| 9719|
| 1446| 9426|
| 6215| 8568|
| 1214| 8470|
| 5376| 8282|
| 2276| 7990|
| 1662| 7808|
| 8235| 7661|
+-----+-----+
only showing top 10 rows
```

3.4 数据挖掘（代码见task4.py）：

- 使用MLlib中Logistic、SVM、NaiveBayes和RandomForest编写程序
- 使用 用户年龄段、性别和卖家id 进行预测
- 将train_after按照70%:30%划分成训练集和测试集
- 使用accuracy_score对预测的准确率进行评估
- 通过**改变训练集中正反例的比例**，每个算法训练十个模型，绘出训练集中正反例比例与预测的准确率的图像



- data.txt是对test_after的预测