

实验报告——tmdb数据集统计

朱政烨 1700017760

薛飞跃 1700017831

2020 年 4 月 11 日

1 实验目标

计算出每个公司的高分电影（得分6.5以上）总收入

2 实验步骤

2.1 清洗数据（只需要收入、评分、公司这三列）

	revenue	vote_average	production_companies
0	2787965087	7.2	[Ingenious Film Partners, Twentieth Century Fo...]
1	961000000	6.9	[Walt Disney Pictures, Jerry Bruckheimer Films...]
2	880674609	6.3	[Columbia Pictures, Danjaq, B24]
3	1084939099	7.6	[Legendary Pictures, Warner Bros., DC Entertai...]
4	284139100	6.1	[Walt Disney Pictures]
...
4798	2040920	6.6	[Columbia Pictures]
4799	0	5.9	[]
4800	0	7.0	[Front Street Pictures, Muse Entertainment Ent...]
4801	0	5.7	[]
4802	0	6.3	[rusty bear entertainment, lucky crow films]

4803 rows × 3 columns

图 1: 清洗后格式

2.2 编写代码

关键部分在于map。类似于词频统计，对于每行取出收入、得分、公司，如果满足高分，就写入(公司,收入)的键值对。

```
1 package tmdb;
2
3 import java.io.IOException;
4
5 import org.apache.hadoop.conf.Configuration;
6 import org.apache.hadoop.fs.Path;
7 import org.apache.hadoop.io.LongWritable;
8 import org.apache.hadoop.io.Text;
9 import org.apache.hadoop.mapreduce.Job;
10 import org.apache.hadoop.mapreduce.Mapper;
11 import org.apache.hadoop.mapreduce.Reducer;
12 import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
13 import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
14
15 public class tmdb {
16
17     public static class TokenizerMapper extends Mapper<Object, Text, Text,
18         LongWritable> {
19
20         public void map(Object key, Text value, Context context) throws
21             IOException, InterruptedException {
22             String line = value.toString();
23             String[] words = line.split(" ", 3);
24
25             Long revenue = Long.valueOf(words[0]);
26             Double votes = Double.valueOf(words[1]);
27
28             String[] companys = words[2].split(", ");
29             if (votes >= 6.5 & revenue > 0) {
30                 for (String c : companys) {
31                     context.write(new Text(c.replaceAll("\\pP", "")),
32                         new LongWritable(revenue));
33                 }
34             }
35         }
36     }
37 }
```

```
33     }
34
35     public static class LongSumReducer extends Reducer<Text, LongWritable, Text,
36         LongWritable> {
37         private LongWritable result = new LongWritable();
38
39         public void reduce(Text key, Iterable<LongWritable> values, Context
40             context)
41             throws IOException, InterruptedException {
42             long sum = 0;
43             for (LongWritable val : values) {
44                 sum += val.get();
45             }
46             result.set(sum);
47             context.write(key, result);
48         }
49     }
50
51     public static void main(String[] args) throws Exception {
52         Configuration conf = new Configuration();
53         Job job = Job.getInstance(conf, "tmdb count");
54         job.setJarByClass(tmdb.class);
55         job.setMapperClass(TokenizerMapper.class);
56         job.setCombinerClass(LongSumReducer.class);
57         job.setReducerClass(LongSumReducer.class);
58         job.setOutputKeyClass(Text.class);
59         job.setOutputValueClass(LongWritable.class);
60         FileInputFormat.addInputPath(job, new Path(args[0]));
61         FileOutputFormat.setOutputPath(job, new Path(args[1]));
62         System.exit(job.waitForCompletion(true) ? 0 : 1);
63     }
64 }
```

2.3 导出为jar

2.4 启动Hadoop并执行程序

```

sz@master:~$ hadoop jar tndb.jar input output
20/04/11 11:41:34 INFO client.RMProxy: Connecting to ResourceManager at Master/192.168.8.100:8032
20/04/11 11:41:34 WARN mapreduce.JobResourceLoader: Hadoop command-line option parsing not performed. Implement the Tool Interface and execute your application with ToolRunner to remedy this.
20/04/11 11:41:35 INFO InputFileInputFormat: Total input paths to process : 1
20/04/11 11:41:35 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
    at java.lang.Object.wait(Native Method)
    at java.lang.Thread.join(Thread.java:1252)
    at java.lang.Thread.join(Thread.java:1326)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:716)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:476)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:652)
20/04/11 11:41:35 INFO mapreduce.JobSubmitter: number of splits:1
20/04/11 11:41:35 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1580499079194_0009
20/04/11 11:41:36 INFO Unl.VarClientImpl: Submitted application application_1580499079194_0009
20/04/11 11:41:36 INFO mapreduce.Job: The url to track the job: http://Master:8080/Proxy/application_1580499079194_0009/
20/04/11 11:41:36 INFO mapreduce.Job: Running job: job_1580499079194_0009
20/04/11 11:41:47 INFO mapreduce.Job: Job job_1580499079194_0009 running in uber mode : false
20/04/11 11:41:47 INFO mapreduce.Job: map 0% reduce 0%
20/04/11 11:41:57 INFO mapreduce.Job: map 100% reduce 0%
20/04/11 11:42:06 INFO mapreduce.Job: map 100% reduce 100%
20/04/11 11:42:06 INFO mapreduce.Job: Job job_1580499079194_0009 completed successfully
20/04/11 11:42:06 INFO mapreduce.Job: Counters: 49
  File System Counters
    FILE: Number of bytes read=07834
    FILE: Number of bytes written=379077
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=38520
    HDFS: Number of bytes written=5123
    HDFS: Number of read operations=0
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=2
  Job Counters
    Launched map tasks=1
    Launched reduce tasks=1
    Data-local map tasks=1
    Total time spent by all maps in occupied slots (ms)=6177
    Total time spent by all reduces in occupied slots (ms)=6252
    Total time spent by all map tasks (ms)=6177
    Total time spent by all reduce tasks (ms)=6252
    Total vcore-milliseconds taken by all map tasks=6177
    Total vcore-milliseconds taken by all reduce tasks=6252
    Total megabyte-milliseconds taken by all map tasks=6325248
    Total megabyte-milliseconds taken by all reduce tasks=6482048
  Map-Reduce Framework
    Map input records=4803
    Map output records=4886
    Map output bytes=137044
    Map output materialized bytes=07834
    Input split bytes=107
    Combine input records=4886
    Combine output records=2238
    Reduce input groups=2238
    Reduce shuffle bytes=07834
    Reduce input records=2238
    Reduce output records=2238
    Spilled Records=4400
    Shuffled Map =1
    Failed Shuffles=0
    Merged Map outputs=1
    GC time elapsed (ms)=191
    CPU time spent (ms)=2080
    Physical memory (bytes) snapshot=312340480
    Virtual memory (bytes) snapshot=3391562880
    Total committed heap usage (bytes)=170604480
  Shuffle Errors
    MAP_ID=0
    CONNECTION=0
    IO_ERROR=0
    WRONG_LENGTH=0
    WRONG_MAP=0
    WRONG_REDUCE=0
  File Input Format Counters
    Bytes Read=385713
  File Output Format Counters
    Bytes Written=5123

```

图 2: 输出信息

3 结果分析

3.1 结果文件

Open	part-r-00000
100 Bares	33965843
10th Hole Productions	34705850
120 Films	792
1200B Films	71904
1492 Pictures	2892611330
2 Entertain	37311672
21 Laps Entertainment	347419439
22 Indiana Pictures	542307423
2929 Productions	180385023
2DUX2	531865000
3 Arts Entertainment	603395563
33andOut Productions	188441614
3D Entertainment	7518876
3Foot7	956019788
3Mark Entertainment	4720371
3ality Digital Entertainment	22730842
40 Acres A Mule Filmworks	372503304
4DH Films	2260712
5150 Action	156909231
838 Productions	122126687
98 MPH Productions	75597042
A Band Apart	896304889
A Plus Image	17511906
A24	94485910
AB Producoes	109423648
ABC Pictures	814666
AE Television Networks	351040419
AIE	93820758
AJOZ Films	60965854
Aline Pictures	49084830
ANA Media	3665069
ARD Degeto Film	146182575
AVCO Embassy Pictures	50244700
Aardman Animations	417287396
Abu Dhabi Film Commission	1506249360
Access Films	46118097
Achte Babelsberg Film	200276000
Across the River Productions	44862187
Act III 52287414	
Act III Communications	83145228
Ad Hominem Enterprises	177243185
Ada Films	3200000
Adan Fields Productions	1270522
Adan Schroeder Productions	145000000
Aegis Film Fund	433771023
Affirm Films	115067528

图 3: 公司-收入

3.2 排序

前十的都是著名公司：华纳、环球、派拉蒙、福克斯...

Warner Bros:	49155747874
Universal Pictures:	42588465942
Paramount Pictures:	40878523165
Twentieth Century Fox Film Corporation:	39357151309
Walt Disney Pictures:	28683256048
Columbia Pictures:	28599634775
New Line Cinema:	19444865804
Amblin Entertainment:	16090835147
DreamWorks SKG:	14384533626
Dune Entertainment:	13797504190

图 4: 收入TOP10

4 实验心得

1. 通过本次实习，掌握了MapReduce原理，学会了简单的MapReduce编程
2. MapReduce易于编程，实现简单，并不需要在意底层的存储、通信等方面
3. 由于数据集不大，未能体现出并行计算的威力