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，如果满足高评分，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,
               LongWritable> {
18
19                 public void map(Object key, Text value, Context context) throws
                       IOException, InterruptedException {
20                         String line = value.toString();
21                         String[] words = line.split(" ", 3);
22
23                         Long revenue = Long.valueOf(words[0]);
24                         Double votes = Double.valueOf(words[1]);
25
26                         String[] companys = words[2].split(", ");
27                         if (votes >= 6.5 & revenue > 0) {
28                                 for (String c : companys) {
29                                         context.write(new Text(c.replaceAll("\\pP", "")),
                                               new LongWritable(revenue));
30                                 }
31                         }
32                 }
```

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

2.3 导出为jar

2.4 启动Hadoop并执行程序

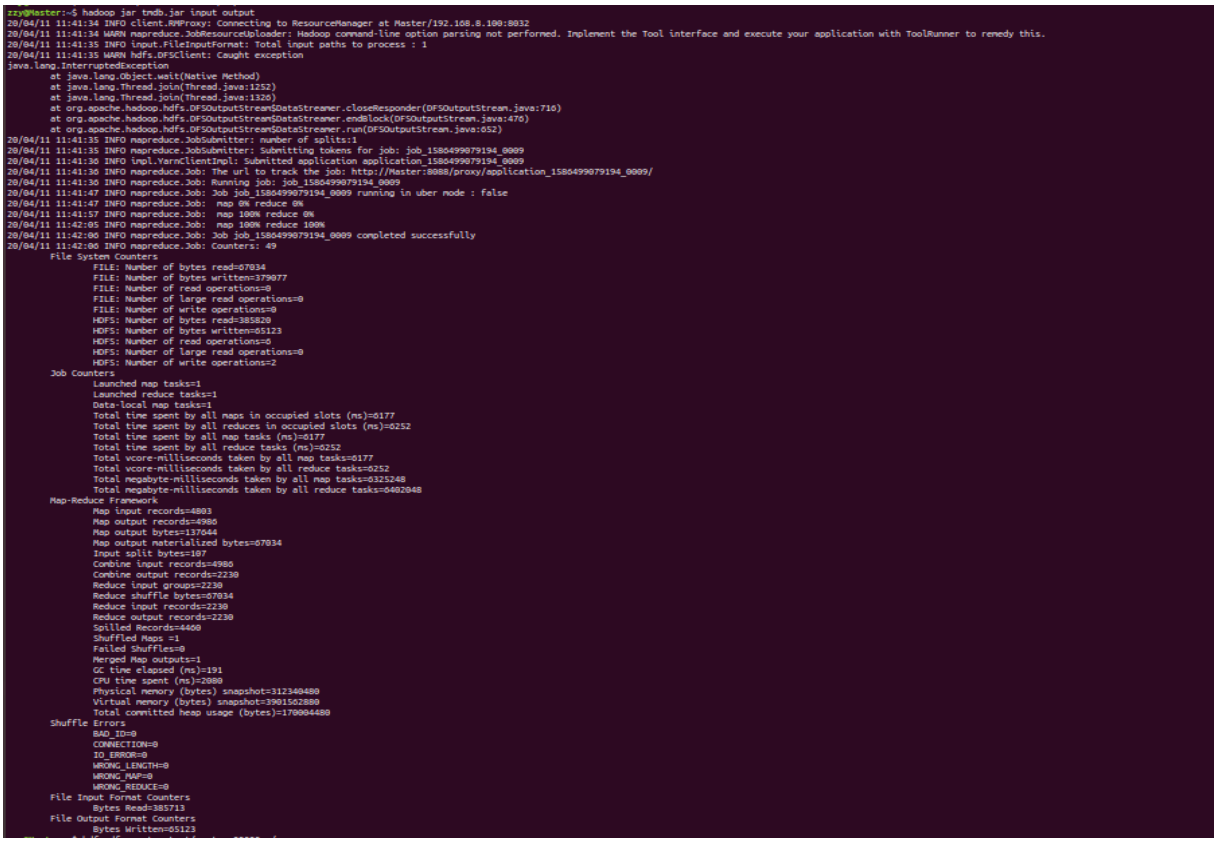

图 2: 输出信息

3 结果分析

3.1 结果文件

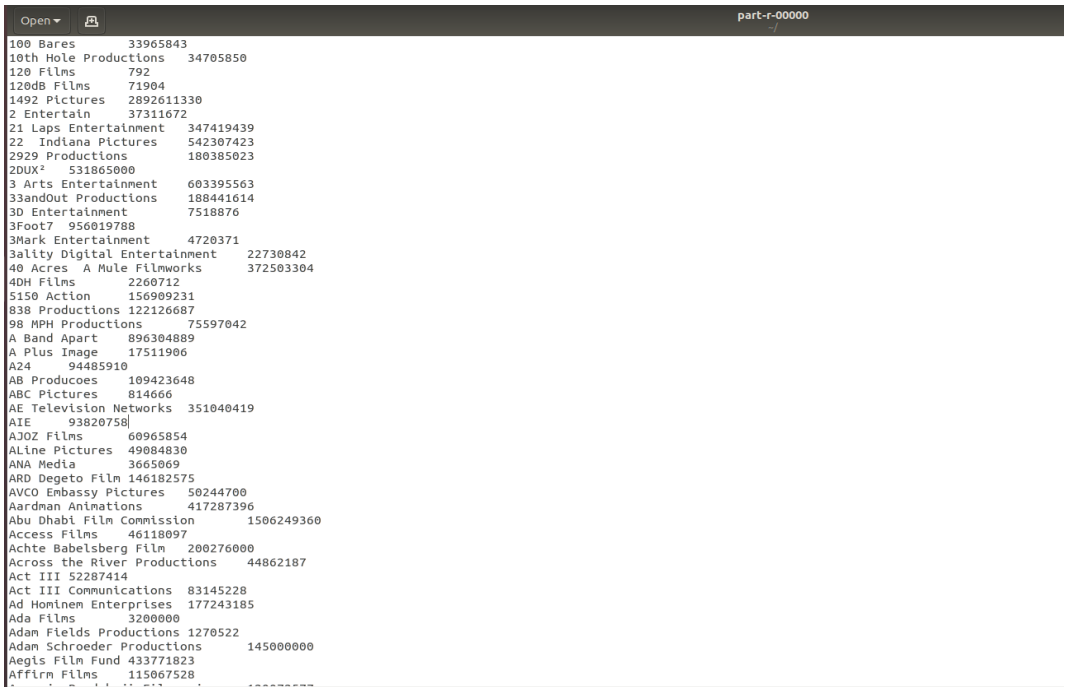

图 3: 公司-收入

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