

# 实验3 Hive表的Join操作

小组成员：张涵之@191220154 林芳麒@191220057

日期：2022.5.7

## 实验3 Hive表的Join操作

### 一、实验内容

#### 任务一 UniversityCountry

- (一) Map 和 Reduce 的设计思路
- (二) MapReduce 中 Map 和 Reduce 代码
- (三) hive输出结果文件的部分截图
- (四) WebUI 执行报告

#### 任务二 CourseUniversity

- (一) Map设计思路
- (二) Drive和MapReduce 中 Map 代码
- (三) hive输出结果文件的部分截图
- (四) WebUI 执行报告

### 二、遇到的问题

## 一、实验内容

### 任务一 UniversityCountry

执行命令：`hadoop jar exp3_UniversityCountry.jar ReduceJoin /data/2022s/hive_join /user/2021sg07/exp3/UniversityCountry`

#### (一) Map 和 Reduce 的设计思路

首先设计一个 `TableBean` 对象，为最终表 `UniversityCountry` 的一条记录。

```
private String n_name;//国家名
private int u_key;//学校序号
private String u_name;//学校名
private String u_webpage;//学校主页
private String flag;//标记是哪个表 university country
```

#### Map:

输出 `Key` 类型为 `String`，`Value` 类型为 `TableBean` (自定义类)

对来自学校信息表 (`university.tbl`) 和国家信息表 (`country.tbl`) 的信息进行切分，用join连接字段国家2字母编码 (`n_alpha-2-code`) 和学校所属国家2字母编码 (`u_alpha-2-code`) 作为输出 **Key**，其余信息学校信息表的 (`u_key`)、(`u_name`)、(`u_webpage`) 字段和国家信息表的 (`n_name`) 字段作为输出 **value**。

并且为了在Reduce端区分 `value` 来自哪个表，需要在map阶段对每条信息打标签 `flag` (即文件名)。具体操作为：在 `setup()` 中获取即将被切分的文件名，在每个表即文件切分之前记录下其表名。

## Reducer:

输入 Key 类型为 String (字段 alpha-2-code), 输入 Value 类型为 Iterable<TableBean> (以 key 为分组的 TableBean 集合)。

输出 Key 类型为 TableBean (自定义类), Value 类型为 NullWritable。

根据在map阶段打的标签 flag 分开values, 每一组中有多个学校信息记录和一个国家信息记录, 将学校信息放入 ArrayList 集合, 遍历该集合, 合并国家信息, 然后输出一个新 TableBean 对象。

## (二) MapReduce 中 Map 和 Reduce 代码

```
/*
TableMapper.java
*/
public class TableMapper extends Mapper<LongWritable, Text, Text, TableBean> {

    private String fileName;
    private Text outK = new Text();
    private TableBean outV = new TableBean();
    /*
    获取每个文件文件名, 放在setup中只需要获取一次, 在Map中则需每获取一行数据重新获取一次文件名
    */
    @Override
    protected void setup(Context context) throws IOException, InterruptedException {
        //初始化 country university
        FileSplit split = (FileSplit) context.getInputSplit();
        fileName = split.getPath().getName();
    }

    @Override
    protected void map(LongWritable key, Text value, Context context) throws
    IOException, InterruptedException {
        //1. 获取一行数据
        String line = value.toString();
        //2. 判断是哪个文件, 存储不同信息
        if (fileName.contains("university")){
            String[] split = line.split("\\|");

            outK.set(split[2]); //国家编码2
            outV.setU_key(Integer.parseInt(split[0])); //学校序号
            outV.setU_name(split[1]); //学校名
            outV.setU_webpage(split[4]); //学校主页
            outV.setFlag("university");

        } else if (fileName.contains("country")){
            String[] split = line.split("\\|");

            outK.set(split[0]); //国家编码2
            outV.setN_name(split[1]); //国家名
```

```

        outV.setFlag("country");
    }
    //3.输出
    context.write(outK,outV);
}
}

```

```

/*
TableReduce.java
*/
public class TableReduce extends Reducer<Text, TableBean, TableBean, NullWritable> {

    /*
    以join连接字段(alpha-2-code)作为key的分组输入，再合并
    */
    @Override
    protected void reduce(Text key, Iterable<TableBean> values, Context context) throws
IOException, InterruptedException {
        ArrayList<TableBean> uniBeans = new ArrayList<>();//学校集合
        TableBean cntryBean = new TableBean();//国家信息

        //1.遍历values，区分两表信息
        for (TableBean value : values) {
            if("university".equals(value.getFlag())){
                TableBean tmpBean = new TableBean();
                uniBeans.add(tmpBean);
            }else {
                BeanUtils.copyProperties(cntryBean,value);
            }
        }

        //2.遍历学校集合，合并国家信息
        for (TableBean uniBean : uniBeans) {
            uniBean.setN_name(cntryBean.getN_name());
            //3.输出
            context.write(uniBean,NullWritable.get());
        }
    }
}

```

### (三) hive输出结果文件的部分截图

Universitycountry.u Key	Universitycountry.u Name	Universitycountry.u Webpage	Universitycountry.n Name
1339	University of Andorra	http://www.uda.ad/	Andorra
8380	Khalifa University of Science, Technology and Research	http://www.ku.ac.ae/	United Arab Emirates
8379	Al Khawarizmi International College	http://www.khawarizmi.com/	United Arab Emirates
8378	Jumeira University	http://www.ju.ac.ae/	United Arab Emirates
8377	Ittihad University	http://www.ittihad.ac.ae/	United Arab Emirates
8376	Higher Colleges of Technology	http://www.hct.ac.ae/	United Arab Emirates
8375	Hamdan Bin Mohammed e-University	http://www.hbmeu.ac.ae/	United Arab Emirates
8374	Gulf Medical University	http://www.gmu.ac.ae/	United Arab Emirates
8373	The Emirates Academy of Hotel Managment	http://www.emiratesacademy.edu/	United Arab Emirates
8372	Etisalat University College	http://www.ece.ac.ae/	United Arab Emirates
8371	Rochester Institute of Technology, Dubai	http://dubai.rit.edu/	United Arab Emirates
8370	Dubai Pharmacy College	http://www.dpc.edu/	United Arab Emirates

« < 1 2 3 > » 共25条

## (四) WebUI 执行报告



### Application application\_1626070675586\_10723

Logged in as: dr.who

Cluster

- About
- Nodes
- Node Labels
- Applications
  - NEW
  - NEW\_SAVING
  - SUBMITTED
  - ACCEPTED
  - RUNNING
  - FINISHED
  - FAILED
  - KILLED
- Scheduler

Tools

Kill Application

Application Overview

User: 2021sg07  
Name: join table  
Application Type: MAPREDUCE  
Application Tags:  
YarnApplicationState: FINISHED  
Queue: root.2021s  
FinalStatus Reported by AM: SUCCEEDED  
Started: Sat May 07 19:27:44 +0800 2022  
Elapsed: 28sec  
Tracking URL: History  
Diagnostics:

Application Metrics

Total Resource Preempted: <memory:0, vCores:0>  
Total Number of Non-AM Containers Preempted: 0  
Total Number of AM Containers Preempted: 0  
Resource Preempted from Current Attempt: <memory:0, vCores:0>  
Number of Non-AM Containers Preempted from Current Attempt: 0  
Aggregate Resource Allocation: 194732 MB-seconds, 48 vcore-seconds

Show 20 entries

Search:

Attempt ID	Started	Node	Logs	Blacklisted Nodes
appattempt_1626070675586_10723_000001	Sat May 7 19:27:44 +0800 2022	http://slave003:8042	Logs	N/A

Showing 1 to 1 of 1 entries

First Previous 1 Next Last

## 任务二 CourseUniversity

执行命令: `hadoop jar exp3_CourseUniversity.jar MapJoin /data/2022s/hive_join /user/2021sg07/exp3/CourseUniversity`

### (一) Map设计思路

首先设计一个 `TableBean` 对象, 为最终表 `CourseUniversity` 的一条记录。

```

private int c_key;
private String c_name;
private String c_subject;
private String c_hours;
private int u_key;
private String u_name;
private String u_webpage;
private String flag;//标记是哪个表 university course

```

若在 Reduce 端进行合并，由于 course 表十分庞大，运行时长长，且需要消耗大量传输宽带，因此我们考虑在 Map 端直接进行两表合并。

把较小的数据源文件 university.tbl 复制到每个 Map 节点,然后在 Map 阶段完成Join操作。总之，用文件复制方法实现Map端Join。

1.首先将 university.tbl 放置到 distributed cache file 中

2.Map阶段：

- (1) 获取缓存文件 university.tbl 内容到 HashMap 集合 uniMap
- (2) 获取输入数据 course.tbl ,切分，将每条记录与 uniMap 进行 join
- (3) 封装为Bean对象输出

## (二) Drive和MapReduce 中 Map 代码

```

/*
MapJoinDrive.java
*/
.....

//加载缓存数据
job.addCacheFile(new URI(args[0]+"/university.tbl"));
//跳过Reduce阶段
job.setNumReduceTasks(0);

.....

```

```

/*
TableMapper.java
*/
public class TableMapper extends Mapper<LongWritable, Text, TableBean, NullWritable> {

    private HashMap<String,String[]> uniMap = new HashMap<>();//
    private TableBean outK = new TableBean();//封装输出对象

    @Override
    protected void setup(Context context) throws IOException, InterruptedException {

```

```

//1. 获取缓存文件地址 university.tbl
URI[] cacheFiles = context.getCacheFiles();
//2. 获取文件内容
FileSystem fs= FileSystem.get(context.getConfiguration());
FSDataInputStream fis = fs.open(new Path(cacheFiles[0]));
BufferedReader reader = new BufferedReader(new InputStreamReader(fis,"UTF-8"));

String line;
//2.1 一行一行处理内容
while (StringUtils.isNotEmpty(line = reader.readLine())) {
    //2.1.1切割
    String[] fields = line.split("\\|");
    String[] uniInfo = new String[2];
    uniInfo[0] = fields[1];//u_name
    uniInfo[1] = fields[4];//u_webpage
    //2.1.2 缓存数据到HashMap集合uniMap中
    uniMap.put(fields[0],uniInfo);
}
//3. 关流
IOUtils.closeStream(reader);
}

@Override
protected void map(LongWritable key, Text value, Context context) throws
IOException, InterruptedException {
    //1.获取一行数据 course.tbl
    String line = value.toString();
    String[] split = line.split("\\|");

    //2.赋值结果Bean对象 course.tbl
    outK.setC_key(Integer.parseInt(split[0]));
    outK.setC_name(split[1]);
    outK.setC_subject(split[2]);
    outK.setC_hours(split[3]);

    //3.遍历uniMap集合进行连接join
    for (HashMap.Entry<String,String[]> uni : uniMap.entrySet()) {
        int u_key = Integer.parseInt(uni.getKey());
        String u_name = uni.getValue()[0];
        String u_web = uni.getValue()[1];

        //4.赋值结果Bean对象 university.tbl
        outK.setU_key(u_key);
        outK.setU_name(u_name);
        outK.setU_webpage(u_web);
        //5.输出
        context.write(outK,NullWritable.get());
    }
}
}

```

```
}
```

(三) hive输出结果文件的部分截图

1 select \* from courseuniversity limit 25;

执行的历史记录	保存的语句 US Citizenship Comparatively	AAS	gendered	4303	Namimatsu University	http://www.namimatsu-u.ac.jp/
0	US Citizenship Comparatively	AAS	gendered	2305	Beijing Foreign Studies University	http://www.bfsu.edu.cn/
0	US Citizenship Comparatively	AAS	gendered	3639	Evangelische Fachhochschule für Sozialpädagogik der "Diakonenanstalt des Rauhen Hauses" Hamburg	http://www.rauheshaus.de/fachhochschule/
0	US Citizenship Comparatively	AAS	gendered	2304	Beijing University of Agriculture	http://www.bac.edu.cn/
0	US Citizenship Comparatively	AAS	gendered	2303	Anhui University of Finance and Economics	http://www.aufe.edu.cn/
0	US Citizenship Comparatively	AAS	gendered	2302	Anhui University	http://www.ahu.edu.cn/
0	US Citizenship Comparatively	AAS	gendered	2301	Anhui University of Traditional Chinese Medicine	http://www.ahtcm.edu.cn/

« < 1 2 3 > »

共25条

(四) WebUI 执行报告

Kill Application

Application Overview

User: 2021sg07

Name: MapJoin

Application Type: MAPREDUCE

Application Tags:

YarnApplicationState: FINISHED

Queue: root.2021s

FinalStatus Reported by AM: SUCCEEDED

Started: Sat May 07 21:07:29 +0800 2022

Elapsed: 1mins, 49sec

Tracking URL: History

Diagnostics:

Application Metrics

Total Resource Preempted: <memory:0, vCores:0>

Total Number of Non-AM Containers Preempted: 0

Total Number of AM Containers Preempted: 0

Resource Preempted from Current Attempt: <memory:0, vCores:0>

Number of Non-AM Containers Preempted from Current Attempt: 0

Aggregate Resource Allocation: 1074706 MB-seconds, 217 vcore-seconds

Show 20 entries

Search:

Attempt ID	Started	Node	Logs	Blacklisted Nodes
appattempt_1626070675586_10727_000001	Sat May 7 21:07:29 +0800 2022	http://slave002:8042	Logs	N/A

Showing 1 to 1 of 1 entries

First Previous 1 Next Last

二、遇到的问题

任务一Reduce端遍历输入Values时，不可直接将迭代器value加入 `ArrayList<TableBean> uniBeans` 中。

原因：hadoop中迭代器 `value` 优化为一个地址，如果直接 `add(value)` 永远是指向一个对象。

解决：每次遍历新建一个 `TableBean` 对象，由 `value` 克隆而来，再加入集合 `uniBeans` 中。