第4章MapReduce编程入门实训

(源自:https://biglab.site)

(版本:Ver1.4-20231024)

实训1获取成绩表的最高分记录

训练要点

1. 掌握MapReduce的编程的基本操作
2. 掌握MapReduce编程的一些常用的数据处理方法，如最大值，去重等
3. 掌握对MapReduce的执行流程
4. 掌握MapReduce程序的输入输出格式

需求说明

样例文件Subject_score成绩表,每行包含两个字段 科目和分数，要求获取成绩列表中每个科目成绩的最高记录，输出到最高成绩表中

实现思路及步骤

1. 准备测试数据，从服务器上下载测试文件subject_score.txt到linux系统; 然后从linux系统上传测试文件到hdfs
2. 编写脚本，实现成绩表最高分统计的MR脚 ，包括FindMaxScore驱动类，FindMaxMapper类和FindMaxReducer类
3. 编译和导出hadoop.jar并上传到集群的master服务器上
4. 使用hadoop jar hadoop.jar命令运行MR程序，观察运行过程
5. 观察hdfs上输出文件内容是否达到预期，内容中应该包含每科的最高分。

作业要求

1. 提供集群网络拓朴图，图中需说明master,slave的名称、IP、JPS进程名。
2. 在http://master:9870上截取本小组集群中本成员目录下/user/myname中上传的文件,需包含subject_score.txt文件的截图
3. 在eclipse中，分别截图 map类，reduce类，main方法的源码图
4. 使用hadoop jar hadoop.jar命令运行MR程 序，运行过程截图
5. 在http://master:9870的文件系统中，打开运行输出结果:/user/myname/output_maxscore/下的文件内容，截图

实现参考

准备测试数据

进入服务器/root/hadoop目录:

```
cd /root/hadoop
```

获取实验数据

```
wget https://biglab.site/b37066/file/subject_score.txt
```

上传数据到hdfs

```
hdfs dfs -put ./subject_score.txt /user/myname/
```

验证数据是否已上传

```
hdfs dfs -ls /user/myname/sub*
```

编写脚本

FindMaxScore类

```
package chap4_score;

import java.io.IOException;

import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.FileSystem;
```

```java
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
import org.apache.hadoop.util.GenericOptionsParser;

import utils.ConfUtil;

public class FindMaxScore {

    public static void main(String[] args) throws IOException, ClassNotFoundException, InterruptedException {
        Configuration conf = ConfUtil.GetConf(FindMaxScore.class);
        String[] otherArgs = new GenericOptionsParser(conf, args).getRemainingArgs();
        if (otherArgs.length < 2) {
            otherArgs = new String[] { "/user/limm/subject_score.txt", "/user/limm/output_subject_score" };
        }

        if (otherArgs.length < 2) {
            System.err.println("Usage: FindMaxScore <in> [<in>...] <out>");
            System.exit(2);
        }
        Job job = Job.getInstance(conf, "FindMaxScore");
        job.setJarByClass(FindMaxScore.class);
        job.setMapperClass(FindMaxMapper.class);
//          job.setCombinerClass(FindMaxReducer.class);
        job.setReducerClass(FindMaxReducer.class);

        job.setMapOutputKeyClass(Text.class);
        job.setMapOutputValueClass(IntWritable.class);

        job.setOutputKeyClass(Text.class);
        job.setOutputValueClass(IntWritable.class);
        job.setNumReduceTasks(1);
        for (int i = 0; i < otherArgs.length - 1; ++i) {
            FileInputFormat.addInputPath(job, new Path(otherArgs[i]));
        }
        FileSystem.get(conf).delete(new Path(otherArgs[otherArgs.length - 1]), true);
        FileOutputFormat.setOutputPath(job, new Path(otherArgs[otherArgs.length - 1]));
        System.exit(job.waitForCompletion(true) ? 0 : 1);

    }

}

```

FindMaxMapper类

```
package chap4_score;

import java.io.IOException;

import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Mapper;

public class FindMaxMapper extends
Mapper<LongWritable,Text,Text,IntWritable>
{
    Text cource = new Text();
    IntWritable score = new IntWritable();
    @Override
    protected void map(LongWritable key, Text value, Context context)
            throws IOException, InterruptedException {
        String line = new
String(value.getBytes(),0,value.getLength(),"GBK"); //为了处理乱码
        String[] values= line.toString().trim().split(" ");
        cource.set(values[0]);
        score.set(Integer.parseInt(values[1]));
        context.write(cource, score); //这里可能出现乱码,如果不进行 GBK转码的
    }



}

```

FindMaxReducer类

```
package chap4_score;

import java.io.IOException;

import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Reducer;

public class FindMaxReducer extends
Reducer<Text,IntWritable,Text,IntWritable>{

    @Override
    protected void reduce(Text key, Iterable<IntWritable> values,
            Context context)
                    throws IOException, InterruptedException {
        int maxScore=-1;
        Text course = new Text();
        for (IntWritable score:values) {
            if (score.get()>maxScore) {
                maxScore= score.get();
                course=key;
```

```
            }
        }
        context.write(course,new IntWritable(maxScore));
    }


}

```

编译与导出jar

编译生成jar包

菜单-》Build -》 Build Artifacts -》hadoop-》Build， 参考4.3节编译生成jar包图例

上传jar包到master

1. 在Hadoop项目，左侧树图中-》out-》artifacts -》hadoop -》 hadoop.jar ，右击hadoop.jar，菜单中选择复制
2. 打开xftp，进入master-》root-》hadoop目录，右击选择粘贴

运行MR程序

(以下myname需换成本人姓名的拼音，全拼或简拼)

使用crt或xshell进入master

```
cd /root/hadoop
hadoop  jar hadoop.jar chap4_score.FindMaxScore \
/user/myname/subject_score.txt \
/user/myname/output_subject_score
```

运行结果如：

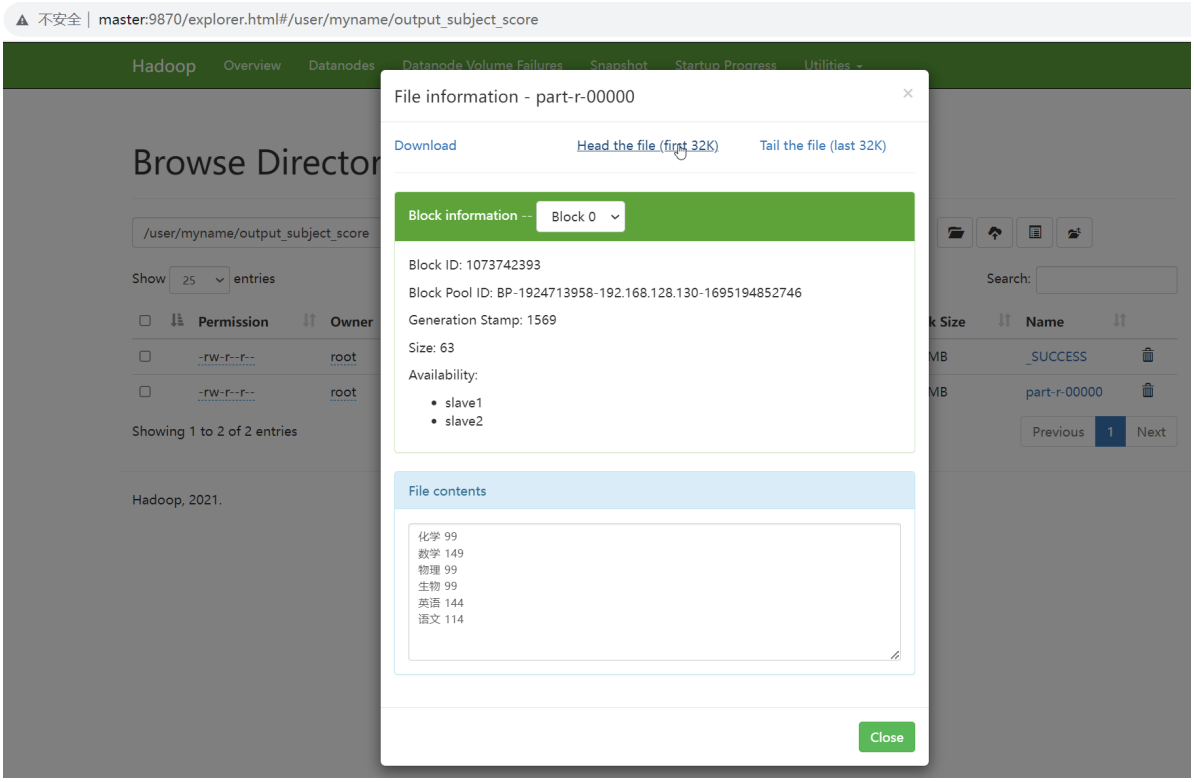
```
[root@master ~]# cd /root/hadoop
[root@master hadoop]# hadoop  jar hadoop.jar chap4_score.FindMaxScore \
> /user/myname/subject_score.txt \
> /user/myname/output_subject_score
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hadoop-3.1.4/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop-3.1.4/share/hadoop/common/slf4j-log4j12-1.7.30.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
class name:chap4_score.FindMaxScore
2023-10-24 14:56:12,755 INFO client.RMProxy: Connecting to ResourceManager at master/192.168.128.130:8032
2023-10-24 14:56:13,930 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1698130548927_0001
2023-10-24 14:56:15,365 INFO input.FileInputFormat: Total input files to process : 1
```

```
2023-10-24 14:56:15,560 INFO mapreduce.JobSubmitter: number of splits:1
2023-10-24 14:56:15,906 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1698130548927_0001
2023-10-24 14:56:15,909 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-10-24 14:56:16,274 INFO conf.Configuration: resource-types.xml not found
2023-10-24 14:56:16,274 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2023-10-24 14:56:16,886 INFO impl.YarnClientImpl: Submitted application application_1698130548927_0001
2023-10-24 14:56:16,956 INFO mapreduce.Job: The url to track the job: http://master:8088/proxy/application_1698130548927_0001/
2023-10-24 14:56:16,957 INFO mapreduce.Job: Running job: job_1698130548927_0001
2023-10-24 14:56:31,544 INFO mapreduce.Job: Job job_1698130548927_0001 running in uber mode : false
2023-10-24 14:56:31,545 INFO mapreduce.Job:  map 0% reduce 0%
2023-10-24 14:56:41,704 INFO mapreduce.Job:  map 100% reduce 0%
2023-10-24 14:56:48,790 INFO mapreduce.Job:  map 100% reduce 100%
2023-10-24 14:56:48,809 INFO mapreduce.Job: Job job_1698130548927_0001 completed successfully
2023-10-24 14:56:48,982 INFO mapreduce.Job: Counters: 53
        File System Counters
                FILE: Number of bytes read=780006
                FILE: Number of bytes written=2004973
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=482714
                HDFS: Number of bytes written=63
                HDFS: Number of read operations=8
                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)=30800
                Total time spent by all reduces in occupied slots (ms)=17756
                Total time spent by all map tasks (ms)=7700
                Total time spent by all reduce tasks (ms)=4439
                Total vcore-milliseconds taken by all map tasks=7700
                Total vcore-milliseconds taken by all reduce tasks=4439
                Total megabyte-milliseconds taken by all map tasks=15769600
                Total megabyte-milliseconds taken by all reduce tasks=9091072
        Map-Reduce Framework
                Map input records=60000
                Map output records=60000
                Map output bytes=660000
                Map output materialized bytes=780006
                Input split bytes=113
                Combine input records=0
                Combine output records=0
```

```
                Reduce input groups=6
                Reduce shuffle bytes=780006
                Reduce input records=60000
                Reduce output records=6
                Spilled Records=120000
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=292
                CPU time spent (ms)=4430
                Physical memory (bytes) snapshot=326193152
                Virtual memory (bytes) snapshot=7199862784
                Total committed heap usage (bytes)=141520896
                Peak Map Physical memory (bytes)=210395136
                Peak Map Virtual memory (bytes)=3597443072
                Peak Reduce Physical memory (bytes)=115798016
                Peak Reduce Virtual memory (bytes)=3602419712
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=482601
        File Output Format Counters
                Bytes Written=63
[root@master hadoop]#
```

查看HDFS结果输出

查看hdfs:/user/myname/output_maxscore目录下结果如：



####