

暨南大学本科实验报告专用纸

课程名称 云计算实验 成绩评定
实验项目名称 使用 MapReduce 完成本学期网课考勤统计
指导教师 魏林锋
实验项目编号 实验项目类型 设计 实验地点 N116
学生姓名 陈宇 学号 2020101642
学院 信息科学技术 系 计算机 专业 软件工程
实验时间 2022 年 11 月 9 日 上 午 ~ 11 月 9 日 上 午

7.1 实验目的

- 1) 理解分布式离线计算框架 MapRecue 的工作原理。
- 2) 通过实验掌握分布式离线计算框架 MapRecue 的编程。
- 3) 使用 MapReuce 完成本学期网课考勤数据的统计工作。

7.2 实验内容

使用 MapReuce 完成本学期网课考勤数据的统计工作。

7.3 实验环境

已经配置完成的 Hadoop 伪分布式或完全分布式环境。环境配置如下：

Hadoop01: 192.168.145.91
Hadoop02: 192.168.145.92
Hadoop03: 192.168.145.93
管理员用户: root / admin@1
Hadoop 用户: hadoop / hadoop

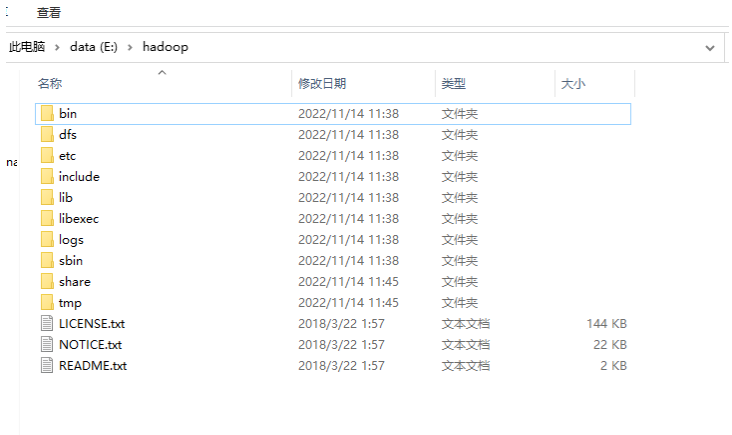
7.4 实验步骤

- 1、安装 Java 环境，下载在 windows 系统中下载 jdk1.8.0_131，图形化界面安装。

2、配置环境变量，右键点击“此电脑”，选择“属性”，点击“高级系统设置”，点击“环境变量”，在“系统变量”中按以下要求新增或修改环境变量。环境变量配置如下：

变量名	变量值	备注
JAVA_HOME	C:\Program Files\Java\jdk1.8.0_131	新增，填写的值为 windows 系统下 Java 环境的目录
HADOOP_HOME	D:\hadoop	新增，填写的值为 windows 系统下 hadoop 环境的目录
HADOOP_USER_HOME	jiahui	新增，填写的值为当前登录的 windows 用户名（不能使用中文）
CLASSPATH	.;%JAVA_HOME%\lib;%JAVA_HOME%\lib\dt.jar;%JAVA_HOME%\lib\tools.jar	新增
PATH	%JAVA_HOME%\bin; %JAVA_HOME%\jre\bin;	在原有 PATH 中添加此两项（切记不是覆盖）。

3、使用 XFTP，登录到 192.168.145.91，下载/usr/hadoop 目录到 E 盘中。



4、将 hadoop-eclipse-plugin-2.7.0.jar 拷贝到 “E:\Program Files (x86)\sts-bundle\sts-3.9.1.RELEASE\plugins” 目录中。

电脑 > data (E:) > Program Files (x86) > sts-bundle > sts-3.7.0.RELEASE > plugins

名称	修改日期	类型	大小
org.springframework.transaction_4.1....	2015/6/28 23:57	Executable Jar File	269 KB
org.springframework.web.servlet_4.1....	2015/6/28 23:57	Executable Jar File	809 KB
org.springframework.web_4.1.4.2015...	2015/6/28 23:57	Executable Jar File	747 KB
org.springframework.ide.eclipse.commo...	2015/6/28 23:57	Executable Jar File	36 KB
org.springframework.ide.eclipse.commo...	2015/6/28 23:57	Executable Jar File	110 KB
org.springframework.ide.eclipse.commo...	2015/6/28 23:57	Executable Jar File	36 KB
org.springframework.ide.eclipse.commo...	2015/6/28 23:57	Executable Jar File	97 KB
org.springframework.ide.eclipse.commo...	2015/6/28 23:57	Executable Jar File	85 KB
org.springframework.ide.eclipse.commo...	2015/6/28 23:57	Executable Jar File	221 KB
org.springframework.ide.eclipse.commo...	2015/6/28 23:57	Executable Jar File	104 KB
org.springframework.ide.eclipse.commo...	2015/6/28 23:57	Executable Jar File	89 KB
org.springframework.ide.eclipse.commo...	2015/6/28 23:57	Executable Jar File	170 KB
org.springframework.ide.eclipse.dashbo...	2015/6/28 23:57	Executable Jar File	240 KB
org.springframework.sts_3.7.0.201506290...	2015/6/28 23:56	Executable Jar File	192 KB
org.tukaani.xz_1.3.0.v201308270617.jar	2015/6/28 23:57	Executable Jar File	108 KB
org.uddi4j_2.0.5.v200805270300.jar	2015/6/28 23:57	Executable Jar File	192 KB
org.w3c.css.sac_1.3.1.v200903091627....	2015/6/28 23:57	Executable Jar File	33 KB
org.w3c.dom.events_3.0.0.draft20060...	2015/6/28 23:57	Executable Jar File	13 KB
org.w3c.dom.smil_1.0.1.v2009030916...	2015/6/28 23:57	Executable Jar File	19 KB
org.w3c.dom.svg_1.1.0.v20101104143...	2015/6/28 23:57	Executable Jar File	86 KB
org.xmlpull_1.1.3.4_v201201052148.jar	2015/6/28 23:57	Executable Jar File	32 KB
org.yaml.snakeyaml_1.14.0.v2015050...	2015/6/28 23:57	Executable Jar File	294 KB
hadoop-eclipse-plugin-2.7.0.jar	2022/11/14 11:22	Executable Jar File	32,867 KB

5、将 winutils.exe 拷贝到 E:\hadoop\bin 目录中。

6、将 hadoop.dll 拷贝到 C:\Windows\System32 目录中。

7、点击 Window->Preferences->Hadoop Map/Reduce, 设置 Hadoop installation directory 为: E:\hadoop, 点击 Apply and Close。

8、打开 STS.exe, 创建项目, 选择 File->New->Project->Map/Reduce Project, 点击“NEXT”, 输入项目名称: MapReduce, 点击“NEXT”, 点击“Finish”。

9、在下方窗口找到 Map/Reduce Locations, 右键点击“New Hadoop Location”。设置如下: Location name: 192.168.123.62; Host: 192.168.123.62; 左边 Port: 9001; 右边 Port: 9000; 勾选“Use M/R Master host”; User name: jiahui, 点击 Finish。配置完成后, 可间左方窗口的 DFS Locations 下有 192.168.123.62 的信息。

General **Advanced parameters**

Location name: 192.168.123.62

Map/Reduce(V2) Master

Host: 192.168.123.62

Port: 9001

DFS Master

☒ Use M/R Master hos

Host: 192.168.123.62

Port: 9000

User name: 85340

10、右键点击左边窗口的 192.168.123.62, 选择 Refresh, 即可将 HDFS 的目录刷新出

来。

11、使用 xshell 连接到 192.168.123.62，使用管理员用户登录。

12、创建用户 jiahui，命令如下：

```
[root@master ~]# useradd jiahui
```

13、切换到 hadoop 用户，命令如下：

```
[root@master ~]# su hadoop
```

14、打开 hadoop 集群，命令如下：

```
[hadoop@master ~]$ start-all.sh
```

15、在 hdfs 上创建 jiahui 文件夹，并将该文件夹的权限赋予给 jiahui 用户，命令如下：

```
[hadoop@master ~]$ hdfs dfs -mkdir /user/jiahui  
[hadoop@master ~]$ hdfs dfs -chown -R jiahui:jiahui /user/jiahui
```

16、右键点击左边窗口的 192.168.24.91，选择 Refresh，即可将 HDFS 的 /user/jiahui 目录刷新出来。

17、在 src 目录右键点击 New->Class，新建 WordMain 类，在 name 中输入：WordMain，点击“Finish”。在 src 目录右键点击 New->Class，新建 WordMapper 类，在 name 中输入：WordMapper，点击“Finish”。在 src 目录右键点击 New->Class，新建 WordReducer 类，在 name 中输入：WordReducer，点击“Finish”。

18、编写 WordMain.java 程序，程序如下：

```
import org.apache.hadoop.conf.Configuration;  
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;  
  
public class WordMain  
{  
    public static void main(String[] args) throws Exception  
    {  
        // Configuration 类：读取 Hadoop 的配置文件，如 site-core.xml...;  
        // 也可用 set 方法重新设置（会覆盖）：conf.set("fs.default.name", "hdfs://xxxx:9000")  
        Configuration conf = new Configuration();  
  
        // 将命令行中参数自动设置到变量 conf 中
```

```

String[] otherArgs = new GenericOptionsParser(conf, args).getRemainingArgs();

/**
 * 这里必须有输入输出
 */

if (otherArgs.length != 2)
{
    System.err.println("Usage: wordcount <in> <out>");
    System.exit(2);
}

Job job = new Job(conf, "word count"); // 新建一个 job, 传入配置信息
job.setJarByClass(WordMain.class); // 设置 job 的主类
job.setMapperClass(WordMapper.class); // 设置 job 的 Mapper 类
job.setCombinerClass(WordReducer.class); // 设置 job 的作业合成类
job.setReducerClass(WordReducer.class); // 设置 job 的 Reducer 类
job.setOutputKeyClass(Text.class); // 设置 job 输出数据的关键类
job.setOutputValueClass(IntWritable.class); // 设置 job 输出值类
FileInputFormat.addInputPath(job, new Path(otherArgs[0])); // 文件输入
FileOutputFormat.setOutputPath(job, new Path(otherArgs[1])); // 文件输出
// System.out.println(job);
System.out.println(job.waitForCompletion(true) ); // 等待完成退出
}
}

```

19、编写 WordMapper.java 程序，程序如下：

```

import java.io.IOException;
import java.util.StringTokenizer;

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

// 创建一个 WordMapper 类 继承于 Mapper 抽象类
public class WordMapper extends Mapper<Object, Text, Text, IntWritable>
{
    private final static IntWritable one = new IntWritable(1);
    private Text word = new Text();

    // Mapper 抽象类的核心方法，三个参数
    public void map(Object key, // 首字符偏移量
        Text value, // 文件的一行内容

```

```

        Context context) // Mapper 端的上下文，与 OutputCollector 和 Reporter 的功能
类似
        throws IOException, InterruptedException
    {
        StringTokenizer itr = new StringTokenizer(value.toString());
        while (itr.hasMoreTokens())
        {
            word.set(itr.nextToken());
            context.write(word, one);
        }
    }
}

```

20、编写 WordReducer.java 程序，程序如下：

```

import java.io.IOException;

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

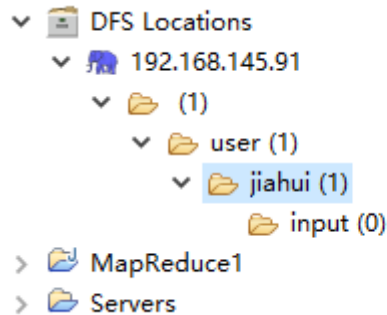
// 创建一个 WordReducer 类 继承于 Reducer 抽象类
public class WordReducer extends Reducer<Text, IntWritable, Text, IntWritable>
{
    private IntWritable result = new IntWritable(); // 用于记录 key 的最终词频数

    // Reducer 抽象类的核心方法，三个参数
    public void reduce(Text key, // Map 端 输出的 key 值
        Iterable<IntWritable> values, // Map 端 输出的 Value 集合（相同 key 的集合）
        Context context) // Reduce 端的上下文，与 OutputCollector 和 Reporter 的功能
类似
        throws IOException, InterruptedException
    {
        int sum = 0;
        for (IntWritable val : values) // 遍历 values 集合，并把值相加
        {
            sum += val.get();
        }
        result.set(sum); // 得到最终词频数
        context.write(key, result); // 写入结果
    }
}

```

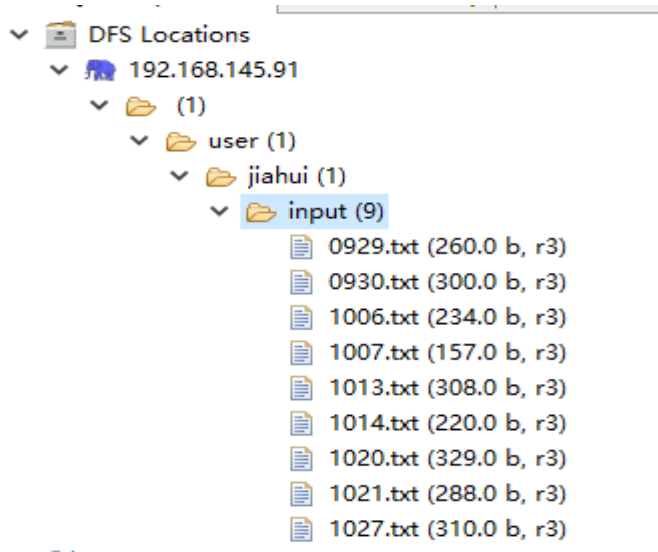
21、右键点击左边窗口的 192.168.24.91, 选择/user/jiahui 文件夹, 右键点击“Create new directory”, 输入新建目录名为: input, 点击 OK 按钮。


22、右击左边窗口的 192.168.24.91 下的/user/jiahui 文件夹, 点击 Refresh, 可见 input 文件夹被创建。



23、右击左边窗口的 192.168.24.91 下的/user/jiahui/input 文件夹, 选择 upload file to hdfs, 批量选择要处理的考勤文件进行上传。

24、右击左边窗口的 192.168.24.91 下的/user/jiahui/input 文件夹, 点击 Refresh, 可见文件已上传到 hdfs 上。

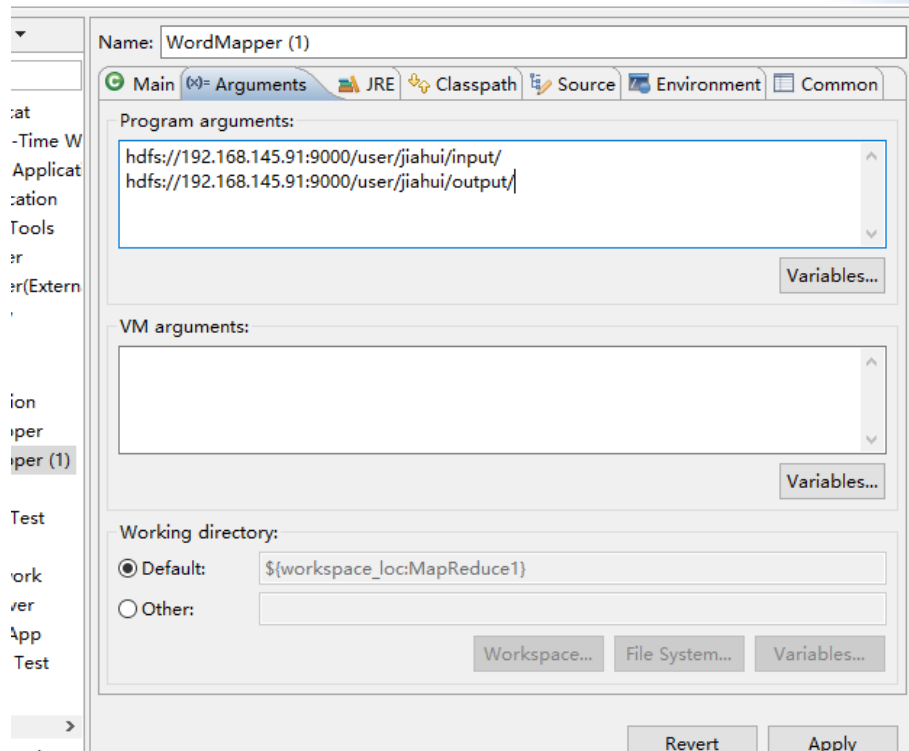


25、在菜单栏中找到  按钮, 点击小三角形, 案选择 “Run Configurations”, 选中 “WordMain”, 点击 “Arguments”, 输入 Program argument, 内容为:

hdfs://192.168.24.91:9000/user/jiahui/input/
hdfs://192.168.24.91:9000/user/jiahui/output/

and run configurations

on



26、点击“Run”运行程序，运行成功后，右击左边窗口的 192.168.145.91 下的 /user/jiahui 文件夹，点击 Refresh，可见 output 文件夹被创建，点击 output 文件夹，可见生成两个文件，选中 part-r-00000 文件，右键点击“Download from DFS”，可将统计信息下载到本地电脑。

27、在本地电脑中使用记事本打开“part-r-00000”，即可查看本学期网课考勤情况统计。

part-r-00000 - 记事本

文件(F) 编辑(E) 格式(O) 查看(V)

DD	1	
blue	1	
p_pmqyang(杨梦清)	6	
zhi.	7	
。	2	
一只猪	1	
丹丹	3	
久别	4	
之琳	1	
代敏	7	
何以解忧	1	
凌顺祥	9	
刘东红	3	
刘建军	2	
刘政正	5	
刘树宁	8	
卢潘周	9	
叶展浩	9	
向勇杰	1	
吴环杰	8	
咋噜秋	1	
哇唧唧哇	3	
唐志鏢	6	
小新卖蜡笔		2
庞冬雪	9	
彭岳富	4	
敖丹丹	1	
文信	5	
李云彬	4	
李剑	7	
李哈哈	1	
杨东玲	7	
杨立文	2	
林晓嘉	1	
梁红梅	2	
牛亭亭	5	
王腊梅	9	
田占文	9	
田婷	8	
罗任昌	4	
罗志锋	7	
谭坚铨	8	
邹苏军	3	
郑宗誉	4	
陈书田	5	
陈川恒	5	

