Hadoop客户端开发环境与WordCount本地测试

**1 说明**

以小组为单位完成此次作业，该部分为大数据导论实验课实验一的准备工作之一，请认真完成，否则会影响实验课程的进行。本作业共包含以下两部分内容：

1. 完成hadoop客户端开发环境的搭建。
2. 编写WordCount代码并在本地进行测试。

**此次作业已提供WordCount代码模板，包含已经实现的词频统计功能，鼓励大家修改Mapper和Reducer的代码，实现不同需求的WordCount功能，作为加分项，例如：统计特定词；统计总词数；筛除英文和数字；加权统计；词长度分布统计（统计长度为1，2，3…的各有多少个）等等。在词频统计基础上实现其他功能，如词频统计排序；词频均值、中位数或标准差计算等等同样适用。**

**本地测试请使用小文件进行测试，比如从爬虫文件中截取一小部分。**

**此次作业需要提交报告，如实现了额外功能需提交代码并在报告中进行阐述。**

希望同学们通过此次作业对真实的大数据开发环境有更深入的了解，同时巩固MapReduce的相关知识。

**2 hadoop客户端安装**

**2.1 基础环境**

需已安装java环境（1.8及更高，建议1.8），已经配置好java环境变量（JAVA\_HOME,Path,CLASS\_PATH）。**Java安装路径不能有空格（C:\Program Files\这种）和中文。**

**2.2 windows客户端安装**

**1 解压**

以【**管理员身份】**解压hadoop-3.2.2.tar.gz，假设解压后的目录为D:\software\hadoop-3.2.2，**将winutil.exe放到bin目录下，（仅windows需要）。如无法解压，请安装Bindzip等软件解压。切记以管理员身份运行。**

**2 添加hadoop环境变量**

**【系统环境变量】里新增**【HADOOP\_HOME】系统变量，值为hadoop解压位置。

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

找到系统变量中的【Path】，在最后追加【%HADOOP\_HOME%\bin】

图形用户界面, 文本, 应用程序

描述已自动生成

添加用户变量或系统变量【HADOOP\_USER\_NAME】，值为【hitsz】,这是实验一所需的环境变量。

图形用户界面, 文本, 应用程序

描述已自动生成

**3 检查**

双击hadoop安装目录bin文件夹下的winutils.exe，检查是否有缺少dll的错误。一闪而过表示没有问题。

打开cmd，输入 hadoop version检查安装是否成功。

**2.3macos和linux客户端安装**

**1 解压**

使用tar命令或其他方式解压hadoop-3.2.2.tar.gz，假设解压目录为/opt/software/ hadoop-3.2.2

**2 编辑全局环境变量文件，添加hadoop环境变量**

sudo vim /etc/profile.d/my\_env.sh

添加以下内容

#HADOOP\_HOME

export HADOOP\_HOME=/opt/module/hadoop-3.2.2

export PATH=$PATH:$HADOOP\_HOME/bin

export PATH=$PATH:$HADOOP\_HOME/sbin

export HADOOP\_USER\_NAME=hitsz

**3使环境变量修改生效**

source /etc/profile

**3 WordCount本地测试（基于IDEA）**

（1）在IDEA中创建一个Maven工程WordCountDemo，导入相应的依赖坐标+日志添加

<dependencies>

<dependency>

<groupId>org.apache.hadoop</groupId>

<artifactId>hadoop-client</artifactId>

<version>3.2.2</version>

</dependency>

<dependency>

<groupId>junit</groupId>

<artifactId>junit</artifactId>

<version>4.12</version>

</dependency>

<dependency>

<groupId>org.slf4j</groupId>

<artifactId>slf4j-log4j12</artifactId>

<version>1.7.30</version>

</dependency>

</dependencies>

（2） 在项目的src/main/resources目录下，新建一个文件，命名为“log4j.properties”，在文件中填入

log4j.rootLogger=INFO, stdout

log4j.appender.stdout=org.apache.log4j.ConsoleAppender

log4j.appender.stdout.layout=org.apache.log4j.PatternLayout

log4j.appender.stdout.layout.ConversionPattern=%d %p [%c] - %m%n

log4j.appender.logfile=org.apache.log4j.FileAppender

log4j.appender.logfile.File=target/spring.log

log4j.appender.logfile.layout=org.apache.log4j.PatternLayout

log4j.appender.logfile.layout.ConversionPattern=%d %p [%c] - %m%n

（3）在src\main\java目录下创建包名 com.hitsz.mapreduce

（4）编写WordCountMapper类，WordCountReducer类，WordCountDriver类

（5）在Project Strucure中添加Artifacts实现自动打包，点击Artifacts -> JAR -> From modules with dependencies

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

（6）选择Main Class，设置META-INF的存储位置，保存

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

（7）右上角新增运行配置 Edit Configuration -> + -> Application ，Main class 选择主程序Program arguments 添加参数 输入路径和输出路径。例如

D:\input D:\output

（8）输入路径为你统计词频的文本文件所在目录，输出路径不能已经存在。

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

（9）点击运行查看结果。