**厦门大学林子雨编著《大数据技术原理与应用》教材配套实验**

五个实验已经收录到林子雨编著《大数据基础编程、实验和案例教程》

教材官网：

http://dblab.xmu.edu.cn/post/bigdatapractice/

http://dblab.xmu.edu.cn/post/bigdata/

## 实验一：熟悉常用的Linux操作和Hadoop操作

### 一、实验目的

Hadoop运行在Linux系统上，因此，需要学习实践一些常用的Linux命令。本实验旨在熟悉常用的Linux操作和Hadoop操作，为顺利开展后续其他实验奠定基础。

### 二、实验平台

* 操作系统：Linux（建议Ubuntu16.04）；
* Hadoop版本：2.7.1。

### 三、实验步骤

**（一）熟悉常用的Linux操作**

* cd命令：切换目录

1. 切换到目录“/usr/local”
2. 切换到当前目录的上一级目录
3. 切换到当前登录Linux系统的用户的自己的主文件夹

* ls命令：查看文件与目录

（4）查看目录“/usr”下的所有文件和目录

* mkdir命令：新建目录

（5）进入“/tmp”目录，创建一个名为“a”的目录，并查看“/tmp”目录下已经存在哪些目录

（6）进入“/tmp”目录，创建目录“a1/a2/a3/a4”

* rmdir命令：删除空的目录

（7）将上面创建的目录a（在“/tmp”目录下面）删除

（8）删除上面创建的目录“a1/a2/a3/a4” （在“/tmp”目录下面），然后查看“/tmp”目录下面存在哪些目录

* cp命令：复制文件或目录

（9）将当前用户的主文件夹下的文件.bashrc复制到目录“/usr”下，并重命名为bashrc1

（10）在目录“/tmp”下新建目录test，再把这个目录复制到“/usr”目录下

* mv命令：移动文件与目录，或更名

（11）将“/usr”目录下的文件bashrc1移动到“/usr/test”目录下

（12）将“/usr”目录下的test目录重命名为test2

* rm命令：移除文件或目录

（13）将“/usr/test2”目录下的bashrc1文件删除

（14）将“/usr”目录下的test2目录删除

* cat命令：查看文件内容

（15）查看当前用户主文件夹下的.bashrc文件内容

* tac命令：反向查看文件内容

（16）反向查看当前用户主文件夹下的.bashrc文件的内容

* more命令：一页一页翻动查看

（17）翻页查看当前用户主文件夹下的.bashrc文件的内容

* head命令：取出前面几行

（18）查看当前用户主文件夹下.bashrc文件内容前20行

（19）查看当前用户主文件夹下.bashrc文件内容，后面50行不显示，只显示前面几行

* tail命令：取出后面几行

（20）查看当前用户主文件夹下.bashrc文件内容最后20行

（21）查看当前用户主文件夹下.bashrc文件内容，并且只列出50行以后的数据

* touch命令：修改文件时间或创建新文件

（22）在“/tmp”目录下创建一个空文件hello，并查看文件时间

（23）修改hello文件，将文件时间整为5天前

* chown命令：修改文件所有者权限

（24）将hello文件所有者改为root帐号，并查看属性

* find命令：文件查找

（25）找出主文件夹下文件名为.bashrc的文件

* tar命令：压缩命令

（26）在根目录“/”下新建文件夹test，然后在根目录“/”下打包成test.tar.gz

（27）把上面的test.tar.gz压缩包，解压缩到“/tmp”目录

* grep命令：查找字符串

（28）从“～/.bashrc”文件中查找字符串'examples'

* 配置环境变量

（29）请在“～/.bashrc”中设置，配置Java环境变量

（30）查看JAVA\_HOME变量的值

**（二）熟悉常用的Hadoop操作**

（31）使用hadoop用户登录Linux系统，启动Hadoop（Hadoop的安装目录为“/usr/local/hadoop”），为hadoop用户在HDFS中创建用户目录“/user/hadoop”

（32）接着在HDFS的目录“/user/hadoop”下，创建test文件夹，并查看文件列表

（33）将Linux系统本地的“～/.bashrc”文件上传到HDFS的test文件夹中，并查看test

（34）将HDFS文件夹test复制到Linux系统本地文件系统的“/usr/local/hadoop”目录下

### 四、实验报告

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 实验报告 | | | | |
| 题目： |  | 姓名 |  | 日期 |
| 实验环境： | | | | |
| 实验内容与完成情况： | | | | |
| 出现的问题： | | | | |
| 解决方案（列出遇到的问题和解决办法，列出没有解决的问题）： | | | | |

（备注：实验答案请见附录A）

## 实验二：熟悉常用的HDFS操作

### 一、实验目的

* 理解HDFS在Hadoop体系结构中的角色；
* 熟练使用HDFS操作常用的Shell命令；
* 熟悉HDFS操作常用的Java API。

### 二、实验平台

* 操作系统：Linux（建议Ubuntu16.04）；
* Hadoop版本：2.7.1；
* JDK版本：1.7或以上版本；
* Java IDE：Eclipse。

### 三、实验步骤

（一）编程实现以下功能，并利用Hadoop提供的Shell命令完成相同任务：

1. 向HDFS中上传任意文本文件，如果指定的文件在HDFS中已经存在，则由用户来指定是追加到原有文件末尾还是覆盖原有的文件；
2. 从HDFS中下载指定文件，如果本地文件与要下载的文件名称相同，则自动对下载的文件重命名；
3. 将HDFS中指定文件的内容输出到终端中；
4. 显示HDFS中指定的文件的读写权限、大小、创建时间、路径等信息；
5. 给定HDFS中某一个目录，输出该目录下的所有文件的读写权限、大小、创建时间、路径等信息，如果该文件是目录，则递归输出该目录下所有文件相关信息；
6. 提供一个HDFS内的文件的路径，对该文件进行创建和删除操作。如果文件所在目录不存在，则自动创建目录；
7. 提供一个HDFS的目录的路径，对该目录进行创建和删除操作。创建目录时，如果目录文件所在目录不存在，则自动创建相应目录；删除目录时，由用户指定当该目录不为空时是否还删除该目录；
8. 向HDFS中指定的文件追加内容，由用户指定内容追加到原有文件的开头或结尾；
9. 删除HDFS中指定的文件；
10. 在HDFS中，将文件从源路径移动到目的路径。

（二）编程实现一个类“MyFSDataInputStream”，该类继承“org.apache.hadoop.fs.FSDataInputStream”，要求如下：实现按行读取HDFS中指定文件的方法“readLine()”，如果读到文件末尾，则返回空，否则返回文件一行的文本。

（三）查看Java帮助手册或其它资料，用“java.net.URL”和“org.apache.hadoop.fs.FsURLStreamHandlerFactory”编程完成输出HDFS中指定文件的文本到终端中。

### 实验报告

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 实验报告 | | | | |
| 题目： |  | 姓名 |  | 日期 |
| 实验环境： | | | | |
| 实验内容与完成情况： | | | | |
| 出现的问题： | | | | |
| 解决方案（列出遇到的问题和解决办法，列出没有解决的问题）： | | | | |

（备注：实验答案请见附录A）

## 实验三：熟悉常用的HBase操作

### 一、实验目的

* + - * 理解HBase在Hadoop体系结构中的角色；
      * 熟练使用HBase操作常用的Shell命令；
      * 熟悉HBase操作常用的Java API。

### 二、实验平台

* 操作系统：Linux（建议Ubuntu16.04）；
* Hadoop版本：2.7.1；
* HBase版本：1.1.5；
* JDK版本：1.7或以上版本；
* Java IDE：Eclipse。

### 三、实验步骤

（一）编程实现以下指定功能，并用Hadoop提供的HBase Shell命令完成相同任务：

1. 列出HBase所有的表的相关信息，例如表名；
2. 在终端打印出指定的表的所有记录数据；
3. 向已经创建好的表添加和删除指定的列族或列；
4. 清空指定的表的所有记录数据；
5. 统计表的行数。

（二）HBase数据库操作

1. 现有以下关系型数据库中的表和数据，要求将其转换为适合于HBase存储的表并插入数据：

学生表（Student）

|  |  |  |  |
| --- | --- | --- | --- |
| 学号（S\_No） | 姓名（S\_Name） | 性别（S\_Sex） | 年龄（S\_Age） |
| 2015001 | Zhangsan | male | 23 |
| 2015003 | Mary | female | 22 |
| 2015003 | Lisi | male | 24 |

课程表（Course）

|  |  |  |
| --- | --- | --- |
| 课程号（C\_No） | 课程名（C\_Name） | 学分（C\_Credit） |
| 123001 | Math | 2.0 |
| 123002 | Computer Science | 5.0 |
| 123003 | English | 3.0 |

选课表（SC）

|  |  |  |
| --- | --- | --- |
| 学号（SC\_Sno） | 课程号（SC\_Cno） | 成绩（SC\_Score） |
| 2015001 | 123001 | 86 |
| 2015001 | 123003 | 69 |
| 2015002 | 123002 | 77 |
| 2015002 | 123003 | 99 |
| 2015003 | 123001 | 98 |
| 2015003 | 123002 | 95 |

2. 请编程实现以下功能：

（1）createTable(String tableName, String[] fields)

创建表，参数tableName为表的名称，字符串数组fields为存储记录各个字段名称的数组。要求当HBase已经存在名为tableName的表的时候，先删除原有的表，然后再创建新的表。

（2）addRecord(String tableName, String row, String[] fields, String[] values)

向表tableName、行row（用S\_Name表示）和字符串数组fields指定的单元格中添加对应的数据values。其中，fields中每个元素如果对应的列族下还有相应的列限定符的话，用“columnFamily:column”表示。例如，同时向“Math”、“Computer Science”、“English”三列添加成绩时，字符串数组fields为{“Score:Math”, ”Score:Computer Science”, ”Score:English”}，数组values存储这三门课的成绩。

（3）scanColumn(String tableName, String column)

浏览表tableName某一列的数据，如果某一行记录中该列数据不存在，则返回null。要求当参数column为某一列族名称时，如果底下有若干个列限定符，则要列出每个列限定符代表的列的数据；当参数column为某一列具体名称（例如“Score:Math”）时，只需要列出该列的数据。

（4）modifyData(String tableName, String row, String column)

修改表tableName，行row（可以用学生姓名S\_Name表示），列column指定的单元格的数据。

（5）deleteRow(String tableName, String row)

删除表tableName中row指定的行的记录。

### 四、实验报告

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 实验报告 | | | | |
| 题目： |  | 姓名 |  | 日期 |
| 实验环境： | | | | |
| 实验内容与完成情况： | | | | |
| 出现的问题： | | | | |
| 解决方案（列出遇到的问题和解决办法，列出没有解决的问题）： | | | | |

（备注：实验答案请见附录A）

## 实验四：NoSQL和关系数据库的操作比较

### 一、实验目的

* 理解四种数据库(MySQL、HBase、Redis和MongoDB)的概念以及不同点；
* 熟练使用四种数据库操作常用的Shell命令；
* 熟悉四种数据库操作常用的Java API。

### 二、实验平台

* 操作系统：Linux（建议Ubuntu16.04）；
* Hadoop版本：2.7.1；
* MySQL版本：5.6；
* HBase版本：1.1.2；
* Redis版本：3.0.6；
* MongoDB版本：3.2.6；
* JDK版本：1.7或以上版本；
* Java IDE：Eclipse；

### 三、实验步骤

**（一） MySQL数据库操作**

学生表Student

|  |  |  |  |
| --- | --- | --- | --- |
| Name | English | Math | Computer |
| zhangsan | 69 | 86 | 77 |
| lisi | 55 | 100 | 88 |

1. 根据上面给出的Student表，在MySQL数据库中完成如下操作：

（1）在MySQL中创建Student表，并录入数据；

（2）用SQL语句输出Student表中的所有记录；

（3）查询zhangsan的Computer成绩；

（4）修改lisi的Math成绩，改为95。

1. 根据上面已经设计出的Student表，使用MySQL的JAVA客户端编程实现以下操作：

（1）向Student表中添加如下所示的一条记录：

|  |  |  |  |
| --- | --- | --- | --- |
| scofield | 45 | 89 | 100 |

（2）获取scofield的English成绩信息

**（二）HBase数据库操作**

学生表Student

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| name | score | | | |
| English | Math | Computer | |
| zhangsan | 69 | 86 | | 77 |
| lisi | 55 | 100 | | 88 |

1. 根据上面给出的学生表Student的信息，执行如下操作：

（1）用Hbase Shell命令创建学生表Student；

（2）用scan命令浏览Student表的相关信息；

（3）查询zhangsan的Computer成绩；

（4）修改lisi的Math成绩，改为95。

2.根据上面已经设计出的Student表，用HBase API编程实现以下操作：

（1）添加数据：English:45 Math:89 Computer:100

|  |  |  |  |
| --- | --- | --- | --- |
| scofield | 45 | 89 | 100 |

（2）获取scofield的English成绩信息。

**（三）Redis数据库操作**

Student键值对如下：

|  |
| --- |
| zhangsan:｛  English: 69  Math: 86  Computer: 77  ｝  lisi:｛  English: 55  Math: 100  Computer: 88  ｝ |

1. 根据上面给出的键值对，完成如下操作：

（1）用Redis的哈希结构设计出学生表Student（键值可以用student.zhangsan和student.lisi来表示两个键值属于同一个表）；

（2）用hgetall命令分别输出zhangsan和lisi的成绩信息；

（3）用hget命令查询zhangsan的Computer成绩；

（4）修改lisi的Math成绩，改为95。

2.根据上面已经设计出的学生表Student，用Redis的JAVA客户端编程(jedis)，实现如下操作：

（1）添加数据：English:45 Math:89 Computer:100

该数据对应的键值对形式如下：

|  |
| --- |
| scofield:｛  English: 45  Math: 89  Computer: 100  ｝ |

（2）获取scofield的English成绩信息

**（四）MongoDB数据库操作**

Student文档如下:

|  |
| --- |
| {  “name”: “zhangsan”,  “score”: {  “English”: 69,  “Math”: 86,  “Computer”: 77  }  }  {  “name”: “lisi”,  “score”: {  “English”: 55,  “Math”: 100,  “Computer”: 88  }  } |

1.根据上面给出的文档，完成如下操作：

（1）用MongoDB Shell设计出student集合；

（2）用find()方法输出两个学生的信息；

（3）用find()方法查询zhangsan的所有成绩(只显示score列)；

（4）修改lisi的Math成绩，改为95。

2.根据上面已经设计出的Student集合，用MongoDB的Java客户端编程，实现如下操作：

（1）添加数据：English:45 Math:89 Computer:100

与上述数据对应的文档形式如下：

|  |
| --- |
| {  “name”: “scofield”,  “score”: {  “English”: 45,  “Math”: 89,  “Computer”: 100  }  } |

（2）获取scofield的所有成绩成绩信息(只显示score列)

### 四、实验报告

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 实验报告 | | | | |
| 题目： |  | 姓名 |  | 日期 |
| 实验环境： | | | | |
| 实验内容与完成情况： | | | | |
| 出现的问题： | | | | |
| 解决方案（列出遇到的问题和解决办法，列出没有解决的问题）： | | | | |

（备注：实验答案请见附录A）

## 实验五：MapReduce初级编程实践

### 一、实验目的

* 通过实验掌握基本的MapReduce编程方法；
* 掌握用MapReduce解决一些常见的数据处理问题，包括数据去重、数据排序和数据挖掘等。

### 二、实验平台

* 操作系统：Linux（建议Ubuntu16.04）
* Hadoop版本：2.7.1

### 三、实验步骤

**（一）编程实现文件合并和去重操作**

对于两个输入文件，即文件A和文件B，请编写MapReduce程序，对两个文件进行合并，并剔除其中重复的内容，得到一个新的输出文件C。下面是输入文件和输出文件的一个样例供参考。

输入文件A的样例如下：

|  |
| --- |
| 20170101 x  20170102 y  20170103 x  20170104 y  20170105 z  20170106 x |

输入文件B的样例如下：

|  |
| --- |
| 20170101 y  20170102 y  20170103 x  20170104 z  20170105 y |

根据输入文件A和B合并得到的输出文件C的样例如下：

|  |
| --- |
| 20170101 x  20170101 y  20170102 y  20170103 x  20170104 y  20170104 z  20170105 y  20170105 z  20170106 x |

**（二）编写程序实现对输入文件的排序**

现在有多个输入文件，每个文件中的每行内容均为一个整数。要求读取所有文件中的整数，进行升序排序后，输出到一个新的文件中，输出的数据格式为每行两个整数，第一个数字为第二个整数的排序位次，第二个整数为原待排列的整数。下面是输入文件和输出文件的一个样例供参考。

输入文件1的样例如下：

|  |
| --- |
| 33  37  12  40 |

输入文件2的样例如下：

|  |
| --- |
| 4  16  39  5 |

输入文件3的样例如下：

|  |
| --- |
| 1  45  25 |

根据输入文件1、2和3得到的输出文件如下：

|  |
| --- |
| 1 1  2 4  3 5  4 12  5 16  6 25  7 33  8 37  9 39  10 40  11 45 |

**（三）对给定的表格进行信息挖掘**

下面给出一个child-parent的表格，要求挖掘其中的父子辈关系，给出祖孙辈关系的表格。

输入文件内容如下：

|  |
| --- |
| child parent  Steven Lucy  Steven Jack  Jone Lucy  Jone Jack  Lucy Mary  Lucy Frank  Jack Alice  Jack Jesse  David Alice  David Jesse  Philip David  Philip Alma  Mark David  Mark Alma |

输出文件内容如下：

|  |
| --- |
| grandchild grandparent  Steven Alice  Steven Jesse  Jone Alice  Jone Jesse  Steven Mary  Steven Frank  Jone Mary  Jone Frank  Philip Alice  Philip Jesse  Mark Alice  Mark Jesse |

### 四、实验报告

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 实验报告 | | | | |
| 题目： |  | 姓名 |  | 日期 |
| 实验环境： | | | | |
| 解决问题的思路： | | | | |
| 实验内容与完成情况： | | | | |
| 出现的问题： | | | | |
| 解决方案（列出遇到的问题和解决办法，列出没有解决的问题）： | | | | |

（备注：实验答案请见附录A）