计算机科学与技术学院 大数据管理与分析 课程实验报告

|  |  |  |
| --- | --- | --- |
| 实验题目：大数据系统基本实验 | | 学号：201605130116 |
| 日期：2019.3.10 | 班级：2016级泰山学堂 | 姓名：杜洪超 |
| Email：[1503345074@qq.com](mailto:1503345074@qq.com) | | |
| 实验目的：  1.熟悉常用的Linux操作和Hadoop操作；  2.熟悉常用的HDFS操作；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；  3.熟悉常用的HBase操作；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；  4.通过NoSQL数据库和常用数据库的比较，理解4种数据库(MySQL、HBase、Redis 和 MongDB)的概念及不同点，熟练使用 4 种数据库操作常用的 shell 命令，熟悉 4 种数据库操作常用的 Java API；  5. 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。 | | |
| 实验软件和硬件环境：  软件环境：  系统：Ubuntu16.04 LTS 64位  软件：openjdk-8-jre,openjdk-8-jdk,java1.8.0\_191  Hadoop 2.9.2，HBase 1.2.11  MySQL 5.7.25，Redis 3.0.6，MongoDB 2.6.10  Eclipse,ssh  硬件环境：  CPU: Intel® Core™ i5-6260U CPU @ 1.80GHz × 4  磁盘：121.8 GB  内存：7.7 GiB | | |
| 实验原理和方法：   1. 熟悉常用的Linux操作和hadoop操作；通过命令行实践Linux命令和hadoop HDFS操作 2. 熟悉常用的HDFS操作；练习HDFS命令行操作和并用JAVA API实现相同功能； 3. 熟悉常用的HBase操作；练习HBase命令行操作和并用JAVA API实现相同功能； 4. 比较NoSQL数据库和常用数据库；练习使用四种数据库的命令行操作并用相应的JAVA API实现相同功能； 5. 通过实验掌握基本的 MapReduce 编程方法； | | |
| 实验步骤：（不要求罗列完整源代码）   1. 熟悉常用的Linux操作和hadoop操作；   常用的Linux命令举例如下：  cp -r src dst 递归复制文件夹  head/tail -n -count 不显示最后count行和只显示最后count行  touch -t time file 修改文件时间  find path -name file\_name 查找文件  tar -czf \*.tar.gz file\_list 压缩文件  tar -xzf \*.tar.gz -C path 解压缩文件到目录  grep string file 查找指定字符串   1. 熟悉常用的HDFS操作；   常用的HDFS命令行如下：    使用HDFA java API编写的函数如下，实现了相同的用shell命令完成的十项任务：     1. 熟悉常用的HBase操作；   HBase shell常用命令如下：    使用HBase java API编写的函数如下，实现了相同的用shell命令完成的几项任务：     1. 比较NoSQL数据库和常用数据库 2. MySQL常用命令：   mysql -u root -p 使用密码登陆MySQL root用户  create database d\_name; 创建数据  use d\_name; 使用某个数据库  create table test(field type); 创建表格  insert into t\_name values(field,value); 插入数据  select \* from t\_name 查看数据  update t\_name set field=value 更新数据   1. HBase常用命令   见实验第三部分   1. Redis常用命令：   redis\_cli 启动命令行客户端  hset table.row field value 插入key-value数据  hget table.row filed 获取指定数据  hgetall table.row 获取某一范围所有key-value数据   1. MongoDB常用命令：   mongo 启动命令行客户端  use t\_name 创建并使用某个数据库  var stus=[{key:value}] 创建文档  db.t\_name.insert(stus) 插入数据  db.t\_name.find().pretty() 输出表格信息  db.t\_name.find({field},{field:0}) 只输出指定数据  db.t\_name.update({field},{$set:{}}) 更新数据  除HBase外其它三种数据库使用Java API编写的类如下：     1. 通过实验掌握基本的 MapReduce 编程方法 2. 文件的合并和去重   对多个文件的行进行合并去重操作；  可以在Map函数中设置key为行，value为空，Reduce函数中对每个key-values只输出key，函数如下：     1. 数据排序   对不同文件的数据进行排序，输出名次和数据；  利用reduce前的sort过程自动排序，Map中把数据作为key,value可以设为任意值，就能在Reduce中获得排好序的数据，只需再处理名次的输出即可，函数如下：     1. 亲属关系挖掘   通过原始数据中的父母和子女关系，挖掘爷爷奶奶和孙子孙女关系；  对给定的一个人，他的父母和和他的子女肯定满足爷孙类关系，同样的某个爷孙类关系肯定存在与某一个人的父母和子女关系中；因此对原始数据的每一行数据，Map函数生成两组key-value，对孩子：key为孩子的姓名，value为parent+父母姓名；对父母，key为父母姓名，value为value+孩子姓名；在Reduce函数中，对key为某个人的values，包括父母关系和子女关系两部分，这两部分两两配对均为爷孙类关系；函数如下： | | |
| 结论分析与体会：  通过本次实验，熟悉了常用的Linux操作和Hadoop操作；了解了常用的数据库，如MySQL,Redis，MongoDB，HBase的基本操作和Java API;并通过几个Map\_Reduce程序的编写和实现，掌握了MapReduce编程的基本方法，加深了对大数据的理解 | | |
| 就实验过程中遇到和出现的问题，你是如何解决和处理的，自拟1－3道问答题：   1. HDFS编程中实现在文件首部添加内容；   为了实现添加多行，在本地新建了临时文件用于存储用户输入，如果在HDFS文件系统中新建临时文件会出现用户不匹配问题，因此把HDFS中的目标文件移动到本地文件系统中，在根据是在首部添加还是尾部添加，把两个文件的内容按顺序写到第三个文件中，最后把第三个文件移回HDFS，删除两个临时文件；   1. HBase无法正常启动；   HBase版本要和HDFS统一，官网中有版本对应关系，更换版本后解决 | | |