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：安装单机Hadoop系统与WordCount程序实验 | | 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：  安装和熟悉单机Hadoop系统，并运行简单的WordCount程序，加深对Hadoop MapReduce程序开发的理解，熟悉实验环境，为后续实验打下基础。 | | |
| 实验软件和硬件环境：  软件环境：  系统：Ubuntu16.04 LTS 64位  软件：openjdk-7-jre,openjdk-7-jdk,java1.7.0\_95  Hadoop 2.9.2  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. 安装和配置环境；配置java和Hadoop环境，安装ssh和eclipse用于远程登陆和编写程序； 2. 运行Hadoop,编写WordCount程序导出为jar包，或利用Hadoop自带程序，收集测试数据，运行并观察结果。 | | |
| 实验步骤：（不要求罗列完整源代码）   1. 安装与配置环境   安装好java和Hadoop环境，测试如下图：    Web界面如下：     1. 测试数据   测试数据使用了山东大学官网及相关链接的网页内容文件，一共爬取了154个网页文件，共2.73M。   1. 编写程序，导出jar包；将数据和jar包复制到HDFS后，利用jar命令下发作业，执行情况如下： | | |
| 结论分析与体会：  实验最终没有实现预期效果，出现把大段文字识别为单词的现象；但原因是java的字符串分解过于简单，不能做到有效分词，不过这与本次实验目的无关，可以通过使用其它更成熟的分词工具解决。  通过本次实验，我熟悉了Hadoop系统和HDFS，对MapReduce编程有了更深的理解，为以后的实验打下了良好的基础。 | | |
| 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. Hadoop安装时出现不能识别java类的错误；   解决办法：更换低版本的Hadoop或使用高版本的java   1. 使用jar命令提交作业时不能识别java类；   解决办法：最终找出原因在于导出jar包时没有指定默认入口类，因此需要在命令中显示加上类的全称或提前指定；   1. 课件或书中提供的代码存在已被舍弃的用法警告；可使用其它合法语法 2. Datanode节点出现无法启动现象；   把tmp目录中data和name下的VERSION文件中的集群ID统一即可。   1. Web界面找不到jobs执行情况；安装yarn可解决。 | | |