

深入理解大数据-大数据处理与编程实践

课程实验与课程设计 内容与要求

鸣谢：本课程得到Google（北京）与Intel公司
中国大学合作部精品课程计划资助

南京大学计算机科学与技术系

主讲人：黄宜华，顾荣

2018年秋季学期

实验1：大数据基本开发工具平台安装使用

实验内容与要求

1. 在本机上安装Maven、Git；
2. 创建自己的Github帐号；
3. 将Alluxio/alluxio (<https://github.com/Alluxio/alluxio>) fork到自己的Github仓库，clone到本地，建立新的分支完成给定任务（任务待分配），并在本地使用maven进行编译测试（需在linux环境下）；
4. 编译测试成功后提交commit，并push到自己Github帐号远程仓库相应分支，最后创建并提交pull request至Alluxio/alluxio (<https://github.com/Alluxio/alluxio>)（参考<http://www.alluxio.org/docs/master/cn/Contributing-to-Alluxio.html>）；
5. 及时处理PR页面中他人提出的修改意见，本地修改后push到自己Github帐号远程仓库即可（不需要重新创建PR），并等待最终merge。

实验完成时间：30天，预计分两批完成；

本课程课件请从<ftp://pasa-bigdata.nju.edu.cn>的2018课程目录下载。

实验2：单机Hadoop系统安装与WordCount实验

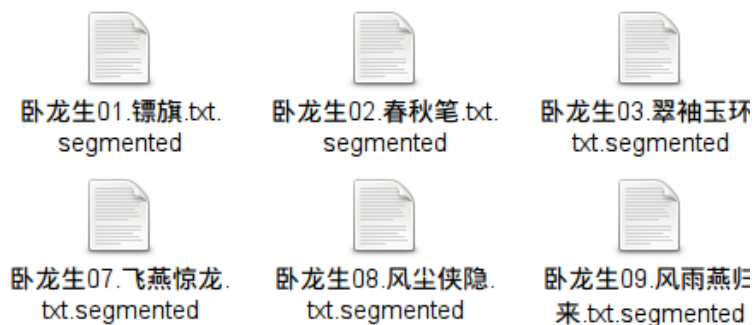
实验内容与要求

1. 每人在自己本地电脑上正确安装和运行伪分布式Hadoop系统
本课程课件请从<ftp://pasa-bigdata.nju.edu.cn>的2018课程目录下载。
2. 安装完成后,自己寻找一组英文网页数据,在本机上运行Hadoop系统自带的WordCount可执行程序文件,并产生输出结果
3. 实验结果提交：要求书写一个实验报告，其中包括：
 1. 系统安装运行的情况
 2. 实验数据说明（下载的什么网页数据，多少个HTML或text文件）
 3. 程序运行后在Hadoop Web作业状态查看界面上的作业运行状态屏幕拷贝
 4. 实验输出结果开头部分的屏幕拷贝
 5. 实验体会
 6. 实验报告文件命名规则：MPExp2-学号-姓名.doc
 7. 实验报告提交至：<ftp://pasa-bigdata.nju.edu.cn>
 8. 实验完成时间：7天

实验3：倒排索引

实验内容与要求

1. 请实现课堂上介绍的“带词频属性的文档倒排算法”。
2. 在统计词语的倒排索引时，除了要输出带词频属性的倒排索引，还请计算每个词语的“平均提及次数”并输出。
$$\text{平均提及次数} = \frac{\text{词语在全部文档中出现的频数总和}}{\text{包含该词语的文档数}}$$
3. 两个计算任务请在同一个MapReduce Job中完成，输出时两个内容可以混杂在一起。
4. 输入输出文件的格式和其他具体要求请见FTP上“实验要求”文件夹下对应的PDF文档。



江湖	古龙48.飘香剑雨.txt.segmented:100,李凉0
江湖	15.867
解药	梁羽生25.牧野流星.TXT.segmented:8,金庸1
解药	6.975

实验3：倒排索引

实验内容与要求

5.实验结果提交：要求书写一个实验报告，其中包括：

1. 实验设计说明，包括主要设计思路、算法设计、程序和各个类的设计说明
2. 程序运行和实验结果说明和分析
3. 性能、扩展性等方面存在的不足和可能的改进之处
4. 源程序，执行程序
5. 运行结果文件
6. 实验报告文件命名规则： MPExp3-组号-组长姓名.doc
7. 实验报告提交至： <ftp://pasa-bigdata.nju.edu.cn>
- 8.实验完成时间： 10天

实验4：社交网络图三角形计数

实验内容与要求

1. 本次实验的输入数据是一张社交网络的关系图
2. 首先请将输入的有向图转换为无向图，然后在无向图上统计图中出现的所有三角形的个数，这是一个典型的图分析问题。
3. 本次实验需要通过多个MapReduce Job完成，请编写一个Driver程序将这些MapReduce Job组织在一起执行。
4. 实验结果提交：要求书写一个实验报告，其中包括：
 1. 实验设计说明，包括主要设计思路、算法设计、程序和各个类的设计说明
 2. 程序运行和实验结果说明和分析,包括最终统计出的三角形个数
 3. 性能、扩展性等方面存在的不足和可能的改进之处
 4. 源程序，执行程序
 5. 更多细节见“实验要求”文件夹下的具体内容
 6. 实验报告文件命名规则： MPExp4-组号-组长姓名.doc
 7. 实验报告提交至： <ftp://pasa-bigdata.nju.edu.cn>
 8. 实验完成时间： 10天

实验5：基于Spark的数据分析

实验内容与要求

待定（设计中）

实验结果提交：要求书写一个实验报告，其中包括：

1. 实验输出结果的屏幕拷贝、相关操作步骤的屏幕拷贝
2. 实验体会
3. 实验报告文件命名规则：MPExp5-学号-姓名.doc
4. 实验报告提交至：FTP：ftp://pasa-bigdata.nju.edu.cn
7. 实验完成时间：10天

课程设计

开题报告

- **目的**：为了评估课程设计选题的内容和难度是否达到一定要求，需要提交开题报告
- **主要内容**
 1. 小组信息（人员，学号，联系信息）
 2. 课题分工：各个成员初步的课题分工计划
 3. 研究题目
 4. 研究问题背景
 5. 主要技术难点和拟解决的问题，尤其要解释说明哪些地方、为什么需要采用MapReduce
 6. 基本解决方法和设计思路、可行性分析，尤其要解释说明如何采用MapReduce并行化算法解决问题
 7. 参考文献
- **提交时间**：12 月 20 日
 - 开题报告文件命名规则：开题报告-组号-组长姓名.doc
 - 课题报告提交至：<ftp://pasa-bigdata.nju.edu.cn>
- **审阅意见返回**：12 月 27 日

课程设计(研究生)

最终课题完成与提交

■ 课程设计结果提交（以下内容打包提交）

● 课程设计报告，内容包括

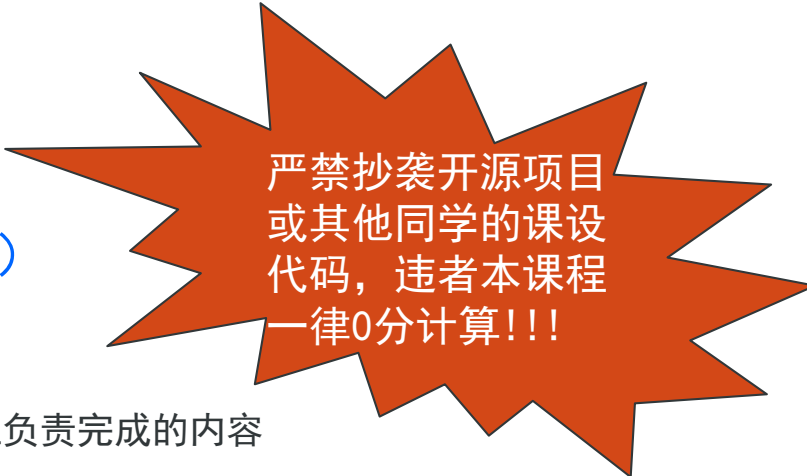
1. 小组信息（人员，学号，联系信息，导师及研究领域）
2. 课题小组分工：需要明确说明各成员在整个课题中分工负责完成的内容
3. 课程设计题目
4. 摘要
5. 研究问题背景
6. 主要技术难点和拟解决的问题，尤其要解释说明哪些地方、为什么需要采用MapReduce
7. 主要解决方法和设计思路，尤其要解释说明如何采用MapReduce并行化算法解决问题
8. 详细设计说明，包括详细算法设计、程序框架、功能模块、主要类的设计说明，包括主要类、函数的输入输出参数、**尤其是map和reduce函数的输入输出键值对详细数据格式和含义**，主要功能和算法代码中加清晰的注释说明
9. **输入文件数据和详细输入数据格式**，输出结果文件数据片段和详细输出数据格式（**必须清晰描述**）
10. 程序运行实验结果说明和分析
11. 总结：特点总结，功能、性能、扩展性等方面存在的不足和可能的改进之处
12. 参考文献

- 带注释的源程序 (**必须提交源程序以备检查实现情况，无源程序的以未完成课程设计处理**)
- 输入数据文件和运行结果文件（**必须提交输入输出文件数据，数据量太大可取部分数据**）
- 执行程序

■ 课题报告文件命名规则：课程设计报告-组号-组长姓名.doc

课题报告提交至：<ftp://pasa-bigdata.nju.edu.cn>

提交截止时间：寒假内完成，具体截止日期待定



严禁抄袭开源项目
或其他同学的课设
代码，违者本课程
一律0分计算!!!