大数据处理综合实验

2024-2025学年 第2学期

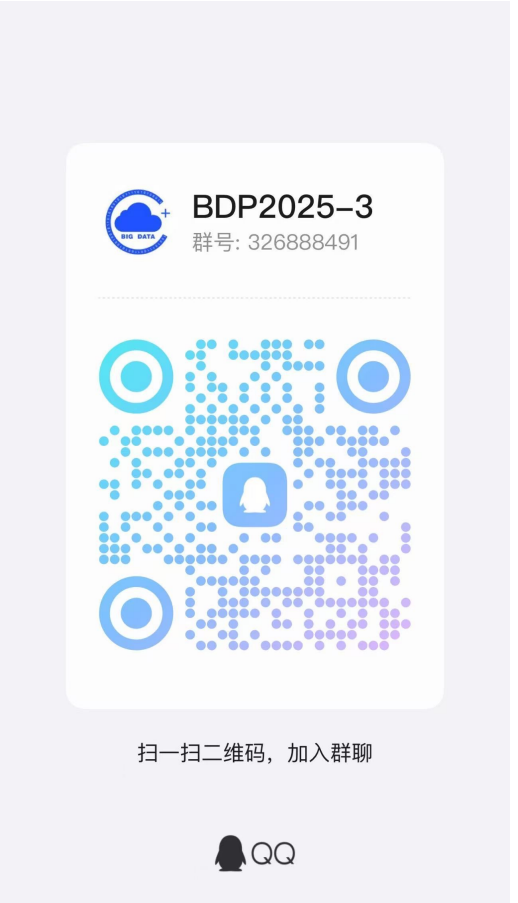
相关信息

- 教师：余萍
  - Email: yuping@nju.edu.cn
  - Office: 仙林计算机系大楼818
- 助教：路知行

- 课程网页(2024-2025第二学期大数据处理综合实验3班):

http://cslabcms.nju.edu.cn/course/view.php?id=1264

- 班级QQ群：326888491



课程简介

- 教学内容简介
  - 系统介绍大数据并行处理技术和编程方法。课程首先介绍并行计算技术的基本概念、原理、方法和技术，在此基础上，介绍基于集群的大数据并行处理技术原理和方法，重点介绍Hadoop MapReduce并行计算集群的构架、用于大数据存储和计算的分布式文件系统、以及基于MapReduce集群的大数据并行处理技术和编程方法，MapReduce并行化算法设计技术、并行化算法应用研究案例。

- 选课要求
  - 具有Java程序设计能力、Linux系统操作使用能力、以及机器学习与数据挖掘算法基础知识。

课程性质

- 计算机专业平台课程（5学分）

- 综合实验课程（理论+实践）：16周
  - 理论：每周2课时（白天）
    - 周二 3-4节 (10:10-12:00) 仙II -104
  - 实践：每周3课时（晚上）
    - 周二 9-11节 (18:30-21:20) 基础实验楼乙125

注意：根据实验安排动态调整白天和晚上授课内容

“具备本领域分析问题解决问题的优秀潜质，具备驾驭本领域复杂计算机系统的良好能力，具备良好的实践技能和新技术应用能力”

“设计/开发解决方案能力：能够独立或者带领一个团队设计复杂问题的计算解决方案，并能够有效开展该计算系统软硬件设计和实现，并能够开展该系统的性能和效率分析。”

课程性质

- 不是又一门语言/编程课
  - 虽然可能需要自学Java、Python、Scala等语言
- 不是又一门分布并行计算系统课
  - 但要求会操作典型的分布并行计算系统
- 不是又一门数据挖掘课程
  - 但会讲授和学习使用一些重要的算法和相关工具

教学目标

- 深入理解大数据处理技术的基本概念、并行计算技术思想、并行计算系统基本架构。
- 理解和掌握Hadoop、HBase、Hive、Spark等大数据处理系统的基本组成和工作原理。
- 理解和掌握MapReduce并行编程模型和方法。
- 通过课程实验，熟悉Hadoop、HBase、Hive 等大数据处理系统的安装、操作、管理和使用。
- 通过课程设计，将大数据处理技术应用到领域应用中。

选修本课程的重要性

- 并行处理成为计算技术的重大发展趋势
- IT行业和应用已进入“大数据（Big Data）”和“数据为王”的时代
- 计算机专业人员面临挑战，市场迫切需要相应的专业技术人才

课程开课情况

- 2016年首次为本科生开设大数据综合实验课程，120人选修，课程情况良好，绝大多数同学一次通过。
- 安排4-5次由易到难的课程实验，从基本工具平台使用到分布式编程实验。
- 分小组完成一个具有一定难度的课程项目设计，每年都会出现一批相当出色的课程设计项目。
- 三个平行班，教学内容和教学进度一致，实验和评分要求一致。

ZQ

我想学习大数据处理技术，应该从何入手

学习大数据处理技术的最佳入门方法取决于您的背景和具体需求。但是，通常从下面的步骤入手是一个不错的选择：

1. 了解大数据基础知识：了解大数据概念、分布式系统、数据存储、数据处理框架等。
2. 学习NoSQL数据库：如MongoDB、Cassandra、HBase等。
3. 学习Hadoop生态系统：如Hadoop Distributed File System（HDFS）、MapReduce、YARN等。
4. 学习Spark：了解Spark的架构、RDD、Spark SQL、Spark Streaming等。
5. 学习数据处理工具：如Apache Pig、Apache Hive等。
6. 学习机器学习：了解机器学习的概念、算法、工具等。
7. 实践：通过项目实践练习，熟练掌握大数据处理技术。

以上是学习大数据处理技术的基本步骤，根据您的需求，可以选择关注的内容。希望以上信息对您有所帮助。

ZQ 还有没有必要学习Hadoop MapReduce

学习Hadoop MapReduce是否有必要取决于您的具体情况。如果您正在寻求大数据处理方案，并且希望使用开源软件，则学习Hadoop MapReduce是有益的。

Hadoop MapReduce是一种分布式数据处理框架，可以处理大量数据。它是Hadoop生态系统的一部分，提供了一种通过在大量计算节点上并行执行任务来处理大量数据的方法。Hadoop MapReduce具有很强的容错能力，可以轻松处理PB级别的数据。

如果您想要在大数据处理领域工作，则学习Hadoop MapReduce是很有必要的。如果您只是想深入了解分布式数据处理，则可以选择学习Hadoop MapReduce或其他类似的技术，如Apache Spark。

总的来说，学习Hadoop MapReduce是否有必要取决于您的具体需求和目标。但是，在今天的数据处理领域，学习Hadoop MapReduce是一个有益的技能。

课程内容

- Ch.1 并行技术和大数据处理技术简介
  - 简要介绍并行计算技术的概况，基本分类，主要技术问题，MPI并行程序设计，大规模并行数据处理技术。
- Ch.2 MapReduce简介
  - 简要介绍MapReduce技术的由来，基本构思，编程模型，主要设计思想和技术特征，基本应用。
- Ch.3 Google 和Hadoop MapReduce的基本架构
  - 介绍Google MapReduce并行计算框架的基本结构、工作原理，Google分布式文件系统GFS的基本构架与工作原理，Google结构化数据管理系统BigTable的基本结构与工作原理。介绍开源MapReduce系统Hadoop 的基本结构、工作原理，Hadoop分布式文件系统HDFS的基本构架与工作原理，并介绍HDFS的基本编程。

课程内容

- Ch.4 Hadoop系统安装运行与程序开发
  - 介绍单机和集群Hadoop系统安装方法和步骤，以及程序开发环境与开发过程。
- Ch.5 MapReduce算法设计
  - 介绍排序算法、文档倒排索引、文档共现算法、专利文献数据分析应用。
- Ch.6 Hadoop HBase与Hive基本原理与程序设计
  - 介绍Hadoop分布式数据管理系统HBase基本工作原理、基本操作和编程方法示例。
  - 介绍Hadoop数据仓库Hive基本原理、基本操作和程序设计。
- Ch.7 高级MapReduce编程技术
  - 介绍复杂I/O数据表示、、用复合键值对完成特殊处理、自定义的I/O格式、Partitioner、Combiner，基于迭代的MapReduce求解方法、数据相关MapReduce任务计算、链式MapReduce计算、多数据源连接、访问关系数据库等高级技术。

课程内容

- Ch.8 基于MapReduce的搜索引擎算法
  - 介绍图算法中广泛使用的网页排名算法PageRank，及其基于MapReduce模型的并行化算法设计与实现。
- Ch.9 MapReduce数据挖掘基础算法
  - 介绍聚类算法，分类算法，频繁项集挖掘等算法的MapReduce并行化方法。
- Ch.10 Spark系统和编程技术
  - 介绍基于内存计算的Spark系统及其基本编程技术。

课程内容

- 实验
  - 五个实验，覆盖Hadoop安装、运行，MapReduce编程，HBase/Hive编程等，独立完成。
- 课程设计
  - 由老师指定具有一定难度和工作量的题目（一般是三选一），分组完成一个综合性大数据课程设计（要求每组人数最多不超过3人，建议2人一组，不跨班）

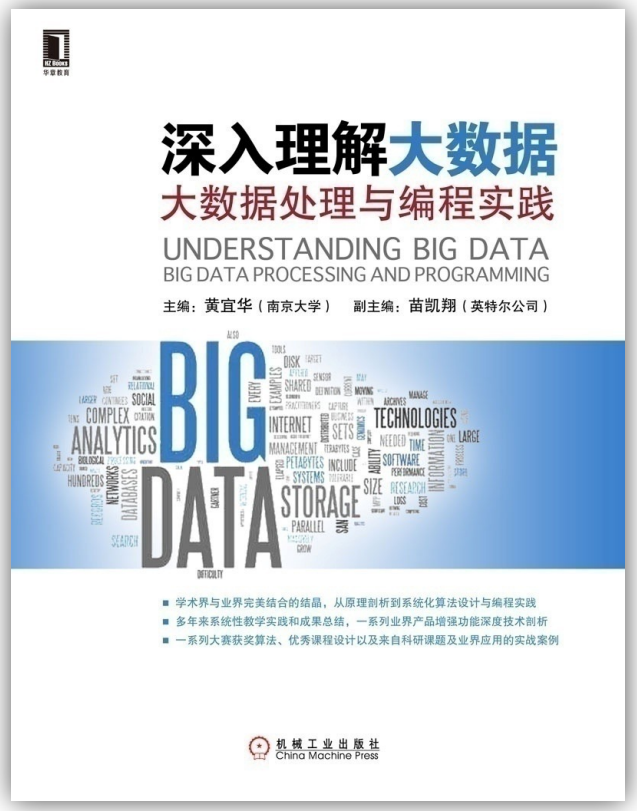
教学周历

周次	教学内容
第1周(02月17-02月23)	白天：大数据和并行计算技术简介（1）；晚上：并行计算技术简介（2）
第2周(02月24-03月02)	白天：MapReduce简介；晚上：Google和Hadoop MapReduce的基本构架 (1)
第3周(03月03-03月09)	白天：Google和Hadoop MapReduce的基本构架 (2)；晚上：Hadoop系统安装运行与程序开发，布置实验1
第4周(03月10-03月16)	白天：MapReduce算法设计 (1)；晚上：大数据技术前沿报告-1，实验1
第5周(03月17-03月23)	白天：MapReduce算法设计 (2)；晚上：布置实验2
第6周(03月24-03月30)	白天：Hadoop HBase/Hive原理与编程技术；晚上：实验2，布置实验3
第7周(03月31-04月06)	白天：高级MapReduce编程技术 (1)；晚上：实验3，布置实验4
第8周(04月07-04月13)	白天：高级MapReduce编程技术 (2)；晚上：大数据技术前沿报告-2，实验4
第9周(04月14-04月20)	白天：基于MapReduce的搜索引擎算法；晚上：布置实验5
第10周(04月21-04月27)	白天：基于MapReduce的数据挖掘基础算法 (1)；晚上：大数据技术前沿报告-3，实验5
第11周(04月28-05月04)	白天：基于MapReduce的数据挖掘基础算法 (2)；晚上：大数据技术前沿报告-4，实验5
第12周(05月05-05月11)	白天：基于MapReduce的数据挖掘基础算法 (3)；晚上：布置课程设计大作业
第13周(05月12-05月18)	白天：Spark系统及其编程技术简介（1）；晚上：大数据技术前沿报告-5，课程设计大作业
第14周(05月19-05月25)	白天：Spark系统及其编程技术简介（2）；晚上：课程设计大作业
第15周(05月26-06月01)	白天：Spark系统及其编程技术简介（3）；晚上：课程设计大作业
第16周(06月02-06月08)	白天：Spark系统及其编程技术简介（4）；晚上：课程设计大作业完成与验收

进度和安排可能动态调整

教材与参考资料

- 《深入理解大数据——大数据处理与编程实践》，黄宜华，2016，机械工业出版社
- 《Hadoop权威指南：大数据的存储与分析》，Tom White，2017，清华大学出版社
- 《数据算法 Hadoop/Spark大数据处理技巧》，Mahmoud Parsian，2016，中国电力出版社
- 《Spark快速大数据分析》，Holden Karau等，2015，人民邮电出版社
- 《Spark高级数据分析》，Sandy Ryza等，2018，人民邮电出版社



考核方式

- 课程实验：5次，共占30%
- 课程设计：1项，占20%
- 期末考试：笔试，闭卷，占50%

大数据竞赛

- 天池大数据竞赛
- CCF BDCI大数据与计算智能大赛
- 中国大学生计算机设计大赛
- Kaggle

THANK YOU

南京大学
NANJING UNIVERSITY

南京大学计算机软件研究所
Institute of Computer Software, Nanjing University