深入理解大数据
-大数据处理与编程实践
课 程 简 介

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

助教：李崇杰，胡求

课程简介

教学内容简介

本课程将系统介绍目前业界和学术界最新的并行计算和大规模海量数据并行处理技术和方法。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，介绍基于集群的大规模海量数据并行处理技术原理和方法，重点介绍MapReduce并行计算集群的构架、用于海量数据存储和计算的分布式文件系统、以及基于MapReduce集群的大规模海量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。

教学目标

课程的主要目标是通过介绍多处理器并行处理技术、以及基于集群的大规模海量数据并行处理技术和MapReduce并行编程模型和方法，要求学生理解和掌握并行处理技术的基本概念、原理和构架、以及基于集群的大规模海量数据并行处理与编程技术方法，并能够用MapReduce实际设计和编写具体的大数据处理应用问题的算法和程序。

选课要求

具有Java程序设计能力，除课堂听课外需要完成编程实验；研究生还要求在学期结束时自选课题完成一个课程设计

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

- 单处理器性能提升达到极限，多核/多处理并行计算成为计算技术发展必然趋势
- 并行计算技术将渗透到每个计算应用领域
- 并行计算技术将影响传统计算技术的各个层面，与传统计算技术相互结合产生很多新的研究热点和课题

IT行业和应用已进入“大数据(Big Data)”和“数据为王”的时代

IT行业应用规模急剧扩大，出现越来越多的超大规模数据处理应用需求，传统系统难以提供足够的存储和计算资源进行处理

- 2008年国际著名的《Nature》杂志出版一期专刊专门讨论未来大数据(Big Data)处理相关的技术问题和挑战
- 世界权威的IT数据分析公司IDC：全世界的数据量2009年为800EB，到2020年将增长44倍，达到35ZB(35,000EB)
- 未来的IT行业中，价值在于数据，未来是“数据为王”的时代

选修本课程的重要性

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

- 并行计算技术将从硬件到软件全面影响传统计算技术的各个层面，新的技术发展挑战和需求迫使我们软件开发和程序设计人员必须尽快掌握并行计算技术
- 20-30年前程序设计技术最大的革命是面向对象技术，而下一个程序设计技术的革命将是并行程序设计技术
- 今天绝大多数程序员不懂并行设计技术，就像20年前绝大多数程序员不懂面向对象技术一样
- 目前国内外的知名IT企业迫切需要大量掌握大规模数据并行处理技术的人才

课程内容

Ch.1 并行计算技术简介

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的基本构架

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

介绍开源大数据处理系统Hadoop 的基本组成结构和工作原理，MapReduce基本框架和工作原理，HDFS基本组成及工作原理，并介绍HDFS的基本编程

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

课程内容

实验1：Hadoop系统安装与WordCount词频统计编程实验

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

实验2：搜索引擎文档倒排索引编程实验

Ch.6 Hadoop HBase与Hive原理与编程技术

介绍Hadoop 分布式数据管理系统HBase工作原理及其编程技术；介绍Hadoop数据仓库 Hive基本结构、工作原理及其编程技术

实验3：Hadoop HBase和Hive编程实验

课程内容

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、程序员定制的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

Ch.8 基于MapReduce的搜索引擎算法

介绍网页排名算法PageRank，搜索引擎文档倒排索引算法，以及全文检索系统的设计实现

实验4：社交网络图三角形计数

Ch.9 基于MapReduce的数据挖掘基础算法

介绍机器学习和数据挖掘中的聚类算法、分类算法、频繁项集挖掘等算法的MapReduce并行化设计技术方法

课程内容

Ch.10 Spark系统和编程技术介绍

介绍基于内存计算的Spark系统及其基本编程技术

Ch.11 云计算技术简介

介绍云计算技术基本概念、发展现状、关键技术与云计算应用

课程设计大作业

自选具有一定难度和工作量的题目，鼓励结合导师的研究工作自选课程设计题目，完成课程设计

课时安排

2017学期

9月18日-12月8日：课堂讲授，课程实验，期末考试
每周2课时，共计12次课堂讲授(24课时)

12月14日：期末考试

12月16日-18年2月28日：研究生分组完成课程设计

考核方法

期末考试

笔试，占50%

课程实验

实验，占25%

课程设计

研究生：自主选题或结合导师研究课题选题，占25%

课件与参考书目、文献

课件参考资料下载：pasa-bigdata.nju.edu.cn，
用户名:hadoopcourse，密码:course2017

参考书目：

1.《深入理解大数据—大数据编程技术与实践》，黄宜华，苗凯翔主编，机械工业出版社，2014
2.《MapReduce大数据并行处理》课程PPT，南京大学，黄宜华
3.《Hadoop in Action》，Chuck Lam，2010，Manning Publications
4.《Data-Intensive Text Processing with MapReduce》，Jimmy Lin and Chris Dyer，2010，University of Maryland, College Park
5.《Mining of Massive Datasets》, 2010, Anand Rajaraman(Kosmix, Inc), Jeffrey D. Ullman (Stanford Univ.)
6.《云计算的关键技术与应用实例》王鹏著，2010，人民邮电出版社

阅读文献：

1. Jeffrey Dean and Sanjay Ghemawat，**MapReduce: Simplied Data Processing on Large Clusters**，OSDI '04
2. Sanjay Ghemawat, et.al, **The Google File System**, SOSP'03
3. Sergey Brin and Lawrence Page, **The Anatomy of a Large-Scale Hypertextual Web Search Engine**
4. Andrew McCallumzy, *et.al,* **Efficient Clustering of HighDimensional Data Sets with Application to Reference Matching**
5. ……

课程开课情况

- 2011-2016学年分别开设了6个学期，每周2学时，研究生选修人数720多人
- 安排4次又简到难的课程实验，从Hadoop安装到编程实验
- 要求结合导师研究课题或自行选题完成一个大的课程设计
- 课程结束后，要求研究生结合导师课题选题或自主选题，分小组完成一个具有一定难度的课程项目设计。每年有30多个小组提交了开题报告，出现一批相当出色的课程设计项目。

课程开课情况

课程项目设计（开题报告和评审意见）

分组		成员	题目	难度（	工作量	可行性与开题报告质量	开题报告评审意见
st52	MF1033037	殷昆燕	基于Hadoop的影片推荐系统	4.5	4	可行，4.5	选题有较大的技术难度（4.5）和设计实现工作量（4），技术方案可行，达到课程设计要求，同意按开题报告进行，请按开题报告的设计目标认真完成课题。
	MF1033011	金国平					
	MG1033057	余宗桥					
st53	MF0933002	陈光鹏	基于Mapreduce的频繁闭项集挖掘算法研究及其实现	4	3	可行，4	选题"基于Mapreduce的频繁闭项集挖掘算法研究及其实现"有一定的难度(4)和设计实现工作量(3)，技术方案可行，达到课程设计要求，同意按开题报告进行，请按开题报告的设计目标认真完成课题。但开题报告中主要技术难点挖掘和分析不足，请在最终设计报告中加强这部分的讨论。
	MF0933009	黄刚					
st54	MG0933035	王团团	基于Map-Reduce框架的SQL语句解析及执行系统	5+	5+	偏难，4	选题类似于Hadoop的子项目Hive，难度（5+）和工作量（5+）很高，具有很大的技术挑战性，难度和工作量都达到并超过课程设计的要求，同意按开题报告进行。但选题目标过大，可能无法按期完成，且开题报告中对基本解决方法和设计实现思路缺少讨论，可行性分析不足，请在这方面进行一定的讨论和可行性分析，在此基础上确定一个适当难度和工作量、可以如期完成的设计目标，最终按设计目标认真完成课题。
	MG1033080	江凯					
	MG1033088	陆瑶					
	MG1033075	顾小东					
st55	MG1033060	张航	基于MapReduce的本体匹配技术	4	3.5	可行,4	选题具有较好的应用问题背景，有一定的技术难度和工作量，技术方案可行，达到课程设计的要求，同意按开题报告进行，请按开题报告的设计目标认真完成课题。但开题报告对具体的本体匹配技术的并行化处理的难点分析讨论不足，对研究问题以及引入MapReduce并行化处理的必要性缺少清晰的描述和讨论，也缺少参考文献。请在最终设计报告中补充这些方面的内容。
	MG1033052	杨琬琪					
	MF1033023	陶承恺					
st56	MG1033015	李文凯	汽车推荐系统	4	4	基本可行,3.5	选题新颖有趣，具有较好的应用前景。选题具有较大的难度（4）和工作量（4）,达到课程设计的要求，同意按开题报告进行，请按开题报告的设计目标认真完成课题。但开题报告内容较为单薄，报告中对课题的主要技术难点及其并行化处理的难点和必要性缺少足够的分析讨论，尽管原型系统可能数据量不会太大，但作为课程设计，要体现出对大数据量并行处理的特点。在基本技术方案中对具体的汽车评估模型及其并行化处理也缺少基本的讨论和描述。请在最终的设计报告中就以上问题提供足够的内容叙述。
	MF1033014	李若冰					
	MF1033041	赵靓					
st57	mf1033015	刘敏	NBA球员数据分析工具	4	4	基本可行，3	选题新颖有趣，具有一定的潜在应用价值。选题具有较大的难度（4）和工作量（4），达到课程设计的要求，同意按开题报告进行，请按开题报告的设计目标认真完成课题。但开题报告过于简单，报告中对主要技术问题及其难点、课题的可行性等缺少足够的分析讨论；基本技术方案中对数据爬取的MapRedcuce并行化处理设计可行性不足，对基本的球员评估模型缺少基本的讨论和描述。请详细研究以上问题，确定课题的可行性，并在最终的设计报告中就以上问题提供足够的内容叙述。
	mf1033019	鲁林					
	mf1033017	刘振兴					
st59	MF1033001	鲍慧慧	Netflix电影推荐	4.5	3.5	可行,3.5	选题有较大的技术难度（4.5）和一定的设计实现工作量（3.5），达到课程设计要求，同意按开题报告进行，请按开题报告的设计目标认真完成课题。但主要评估算法的有效性和可行性研究和分析不足，请详细研究并确定所选算法的有效性和可行性，如果所选算法不是足够有效，需要考虑更为复杂和有效的算法；另外开题报告内容较为单薄，最终设计报告中请注意保证有足够和完整的内容。
	MF1033009	蒋慧					
	Mf033035	杨丽					
	MF1033033	薛艳					
st60	MF1033002	蔡希辉	？？推荐系统			重新提交	选题没有题目，主要研究内容是什么？问题关键点在哪里？与MapReduce有什么关系？都没有交待。如果是做推荐系统，需要说明具体是哪个应用领域、什么样的具体推荐系统问题。开题报告简单粗糙，内容不清，不符合要求，请重新提交开题报告。
	MF1033036	杨阳					
	MF1033030	徐佳一					
	MG1033102	朱正文					

优秀课程项目设计示例

陈虎，笪庆小组：基于内容的图像搜索引擎EagleEye
—MapReduce海量数据并行处理项目

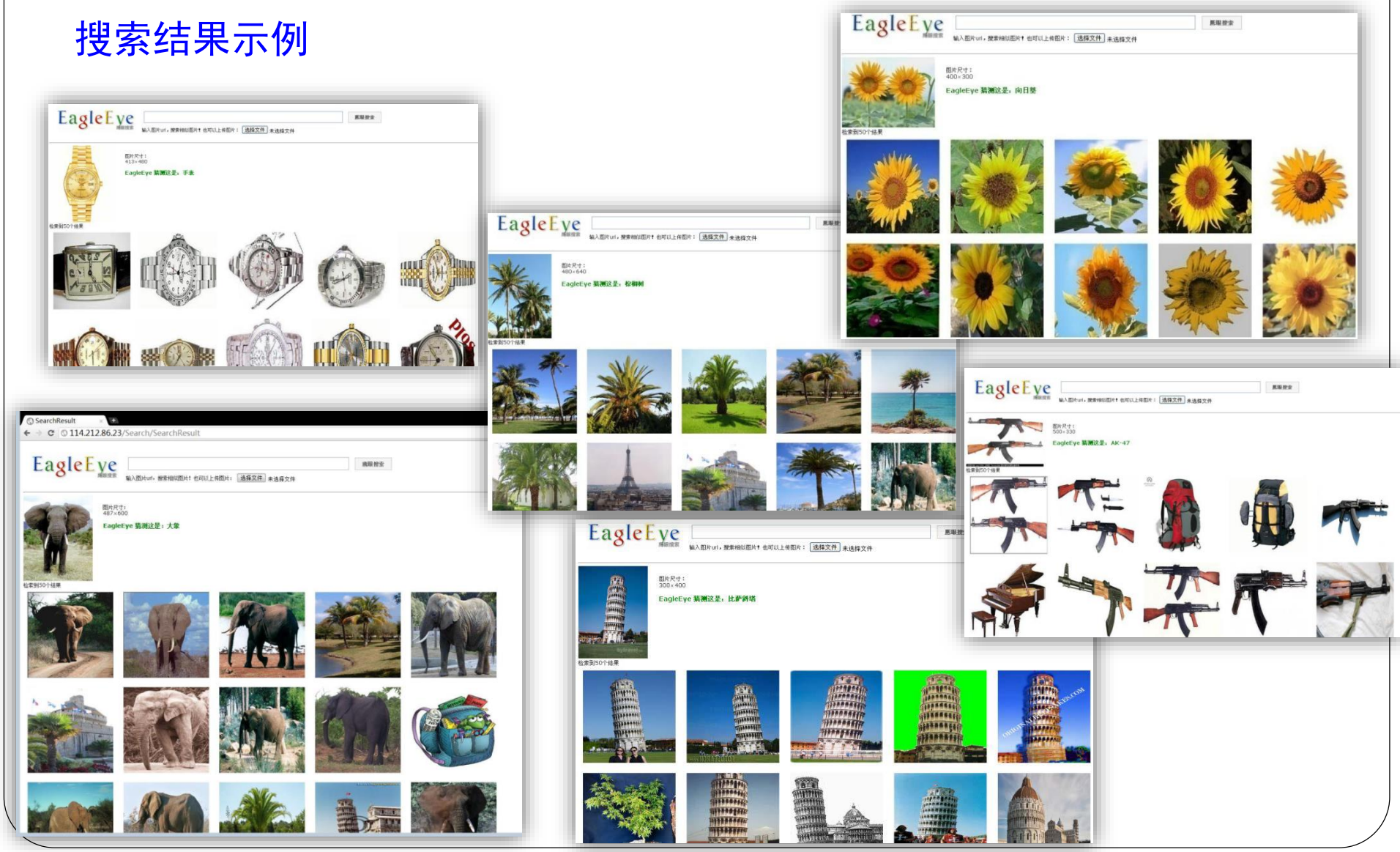
主要研究内容：

1、研究解决了有效的图像特征表示和快速提取方法：表示和提取图像的特征使其在基于内容的图像检索中能够更准确地表征不同图像之间的相似程度。
2、研究解决了基于MapReduce的海量图像特征索引和图像搜索算法
3、完成了一个基于内容的图像搜索EagleEye原型系统的设计实现

优秀课程项目设计示例

陈虎，笪庆小组：基于内容的图像搜索引擎EagleEye

搜索结果示例

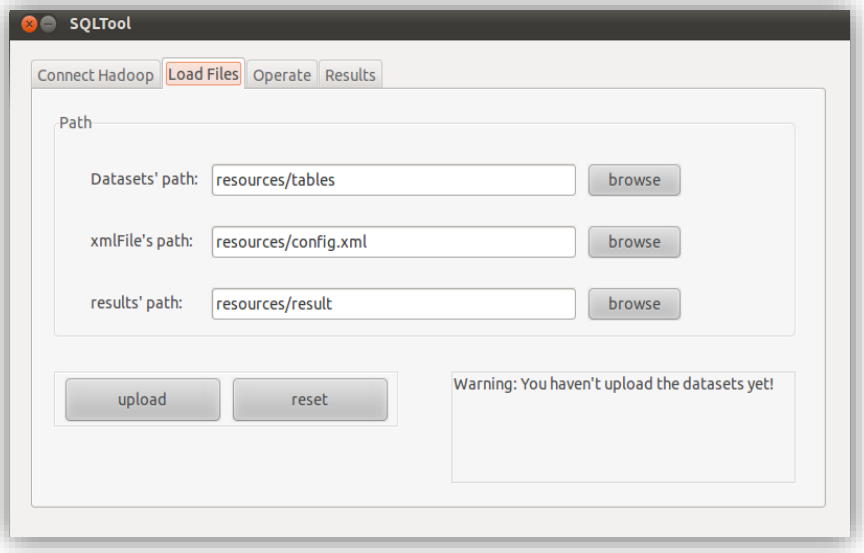


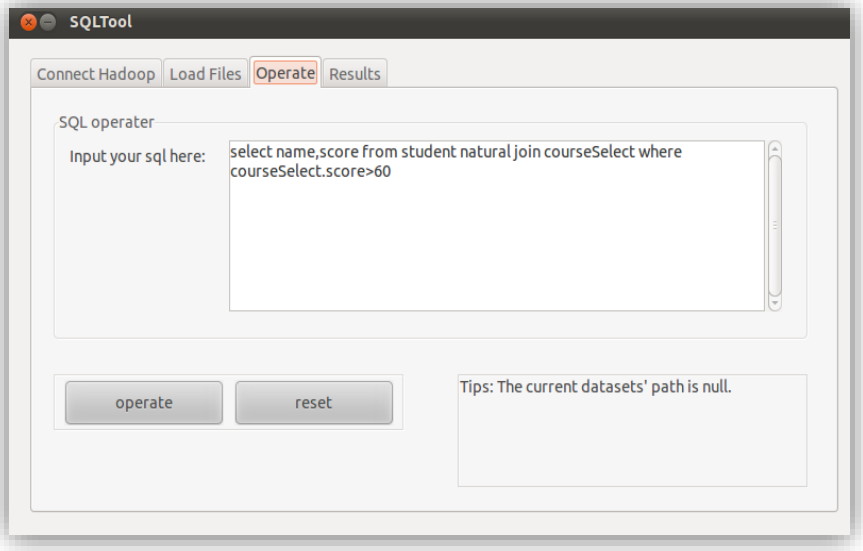
优秀课程项目设计示例

江凯, 顾小东, 陆瑶, 王团团小组：**基于Hadoop的SQL查询工具**

主要研究了在Hadoop分布式文件系统环境下设计和模拟一个管理和查询结构化数据的原型数据库系统，主要技术内容包括：

- 设计了基于XML的数据库Schema的描述和处理方法
- 设计了基本的SQL查询语言
- 完成SQL语句的解析处理
- 完成SQL到关系代数的转换处理
- 基于MapReduce并行计算框架完成关系代数的并行化处理，提高计算效率
- 设计实现了一个原型的查询工具





优秀课程项目设计示例

梁亚澜, 李杰, 钮鑫涛：Hadoop平台下覆盖表生成遗传算法参数配置启发式演化工具

主要研究内容：

1. 采用启发式演化方法对遗传算法的种群规模、进化机制、交叉概率、变异概率及其变种算法5个因素进行取值组合演化，系统地探索各个因素对遗传算法覆盖表生成效果的影响程度和性质，并以覆盖表规模和消耗时间为依据寻找出最佳配置
2. 遗传算法生成覆盖表的计算量极大，设种群规模为100，进化代数为1000，则完整的进化过程需运行遗传算法100*1000 = 100,000次，以一次生成覆盖表的时间为1分钟为例，采用串行计算共需100000分钟，约71天。课题研究实现了基于Hadoop MapReduce的并行化遗传算法生成覆盖表算法, 大大缩短了计算时间

表 3：各待测实例的最终最优配置和覆盖表生成结果

	Algorithm	m	T	Pc	Pm	Size	Time		Algorithm	m	T	Pc	Pm	Size	Time
4^{10}	GAr climb	100	100	0.2	0.2	28	0.234	6^{30}	GA climb	100	1100	0.2	0.2	87	52.6
3^{13}	GAr climb	100	1100	0.8	0.2	17	2.28	10^{11}	GA climb	100	1100	0.8	0.2	154	19.8
6^{10}	GA climb	6100	1100	0.2	0.2	58	402	$7^6 6^7 5^6$	GAr climb	100	1100	0.8	0.2	82	23.5
4^{20}	GAr climb	100	1100	0.8	0.2	35	10.1	$8^2 7^2 6^2 5^2$	GA- climb	2100	600	0.8	0.6	70	277
8^{10}	GA climb	2100	600	0.6	0.2	98	604	$6^1 5^1 4^6 3^8 2^3$	GAr climb	4100	1100	0.8	0.4	36	568.1
3^{20}	GA- climb	100	600	0.2	0.2	21	3.31	6^4	GAr climb	100	100	0.6	0.2	41	0.03
6^{20}	GA climb	100	1100	0.8	0.2	74	22.9	$5^1 3^8 2^2$	GAr climb	100	100	0.8	0.2	20	0.43
4^{30}	GAr climb	100	600	0.2	0.2	40	12.4								

基于本课程设计项目的研究成果作者和导师发表了两篇学术论文

1. 梁亚澜，聂长海，覆盖表生成的遗传算法配置参数优化 2011年6月，计算机学报已录用.
2. Liang Yalan, Changhai Nie, Jonathan M. Kau_man, Gregory M. Kapfhammer, and Hareton Leung. **Empirically identifying the best genetic algorithm for covering array generation**. In Proceedings of the 3rd International Symposium on Search Based Software Engineering, Szeged, Hungary, September 2011

课程项目设计

梁亚澜, 李杰, 钮鑫涛:Hadoop平台下覆盖表生成遗传算法参数配置启发式演化工具
李袁奎, 刘文杰, 王姜: 使用Mapreduce框架进行软件代码分析

软件工程

黄刚, 陈光鹏: 一种基于MapReduce的频繁闭项集挖掘算法研究及其实现
王苏琦, 金龑, 罗爱宝, 王灵江: 基于模型的协同过滤并行化算法
胡昊然，冯子陵，窦文科，刘晶晶：面向新浪微博的关注推荐系统
段轶：Netflix电影数据聚类分析

机器学习
数据挖掘

孙道平：基
刘敏, 刘
刘正, 朱小
王尧, 苏宗

- 选题覆盖了我系大多数研究方向
- 随着研究问题数据规模越来越大，越来越多的研究领域都需要使用并行计算技术提供新的计算方法
- 本课程的开设对推动我系各方向的研究将起到积极的作用

社会网络
分析

化的研究
的分析实验

金惠益, 刘
基于短

机器翻译

式设计

张旭，何良朋：P2P流媒体中的结点分簇与最短路径构造

网络通信

陈虎，笪庆小组：基于内容的图像搜索引擎EagleEye

多媒体检索

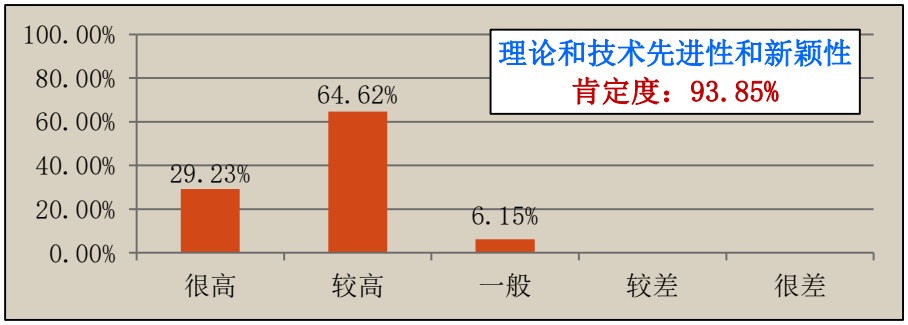
张航, 杨琬琪, 陶承恺: 基于MapReduce的本体匹配技术

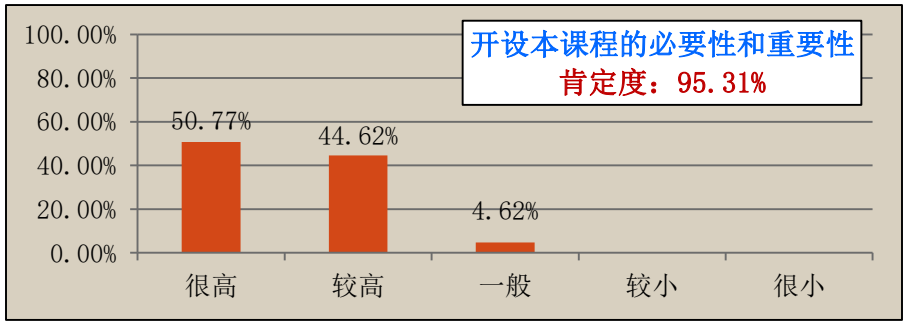
Web本体

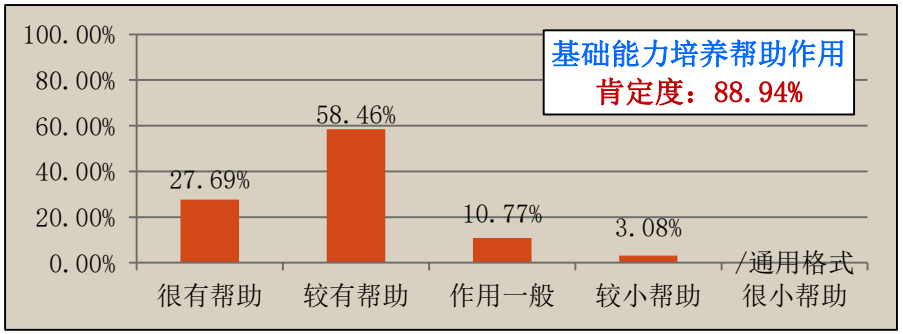
江凯, 顾小东, 陆瑶, 王团团小组：基于Hadoop的SQL查询工具

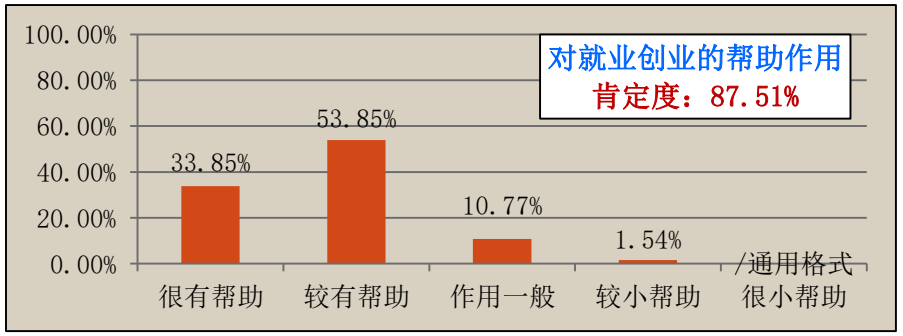
数据库

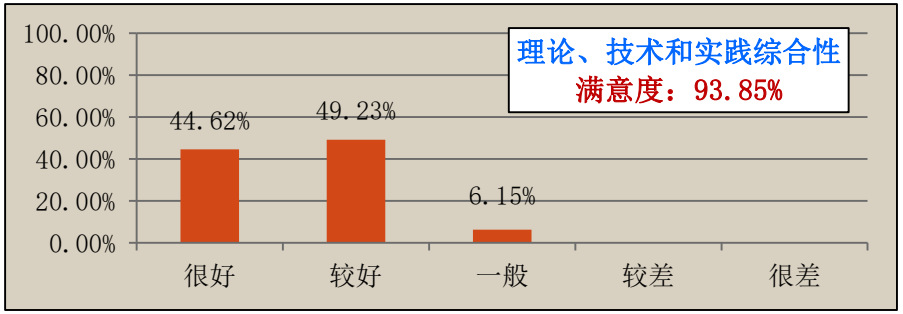
课程教学评估

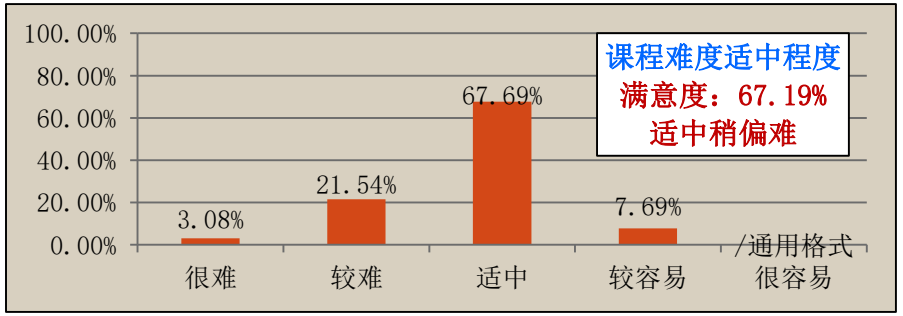

理论和技术先进性和新颖性
肯定度：93.85%
100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
29.23%
64.62%
6.15%
很高
较高
一般
较差
很差


开设本课程的必要性和重要性
肯定度：95.31%
100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
50.77%
44.62%
4.62%
很高
较高
一般
较小
很小

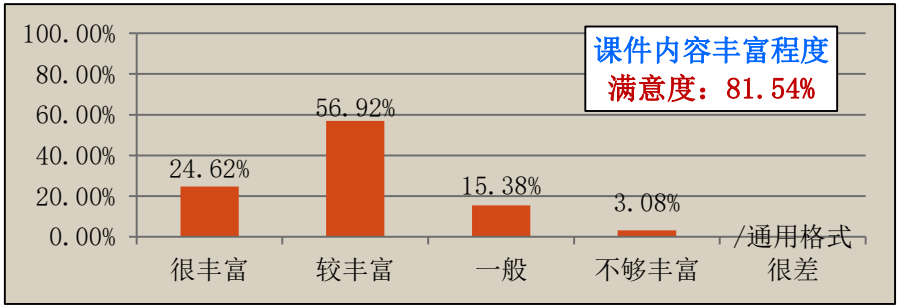

基础能力培养帮助作用
肯定度：88.94%
100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
27.69%
58.46%
10.77%
3.08%
很有帮助
较有帮助
作用一般
较小帮助
很小帮助
/通用格式

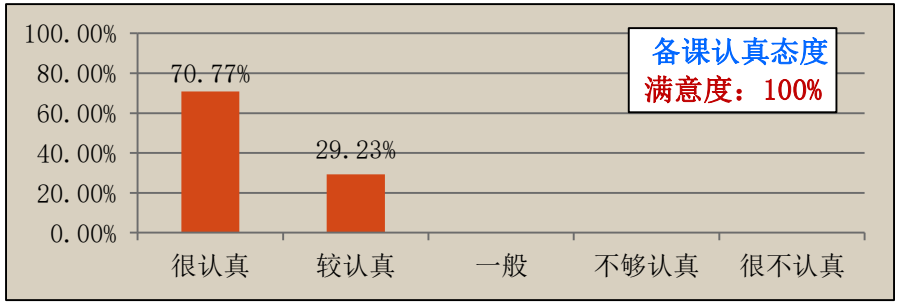

对就业创业的帮助作用
肯定度：87.51%
100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
33.85%
53.85%
10.77%
1.54%
很有帮助
较有帮助
作用一般
较小帮助
很小帮助
/通用格式

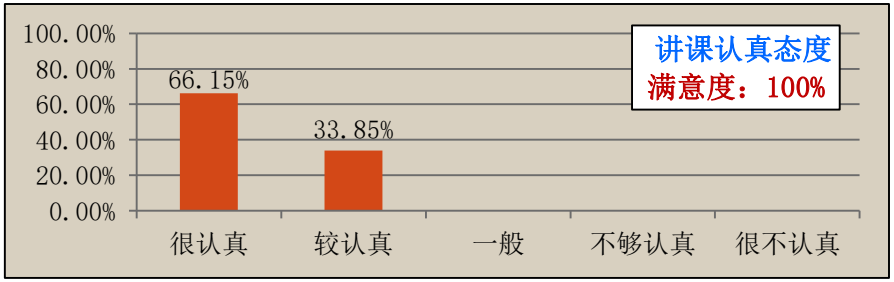

理论、技术和实践综合性
满意度：93.85%
100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
44.62%
49.23%
6.15%
很好
较好
一般
较差
很差

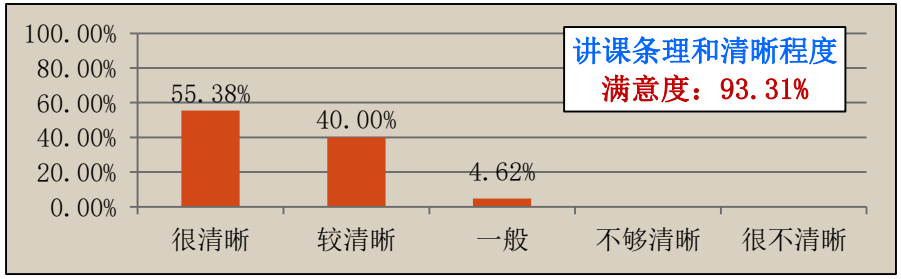

课程难度适中程度
满意度：67.19%
适中稍偏难
100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
3.08%
21.54%
67.69%
7.69%
很难
较难
适中
较容易
很容易
/通用格式

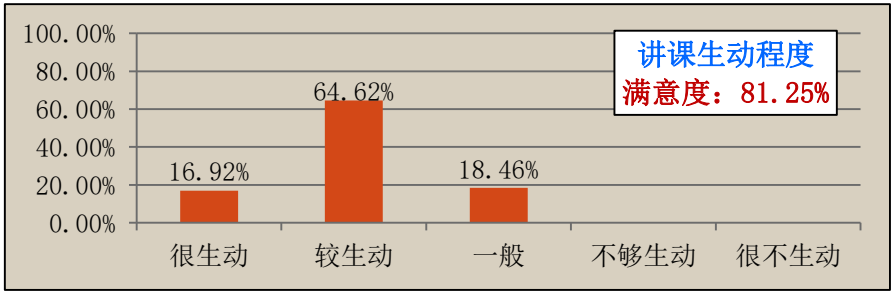
课程教学评估

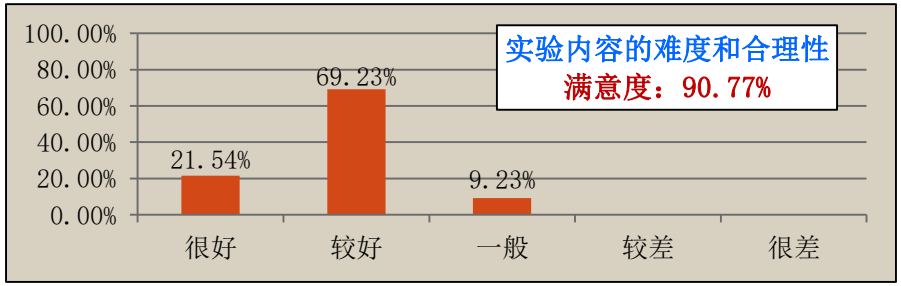

课件内容丰富程度
满意度：81.54%
100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
24.62%
56.92%
15.38%
3.08%
/通用格式
很丰富
较丰富
一般
不够丰富
很差

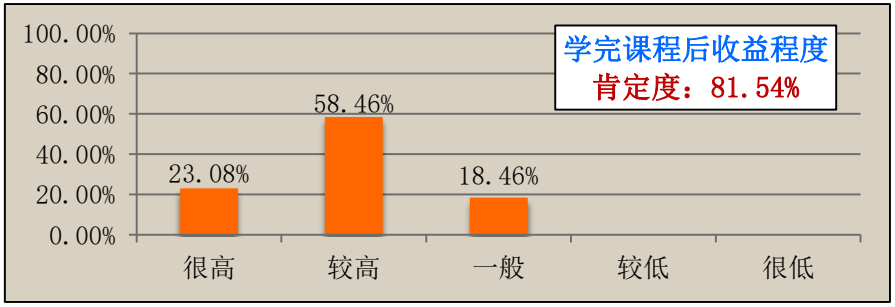

备课认真态度
满意度：100%
100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
70.77%
29.23%
很认真
较认真
一般
不够认真
很不认真

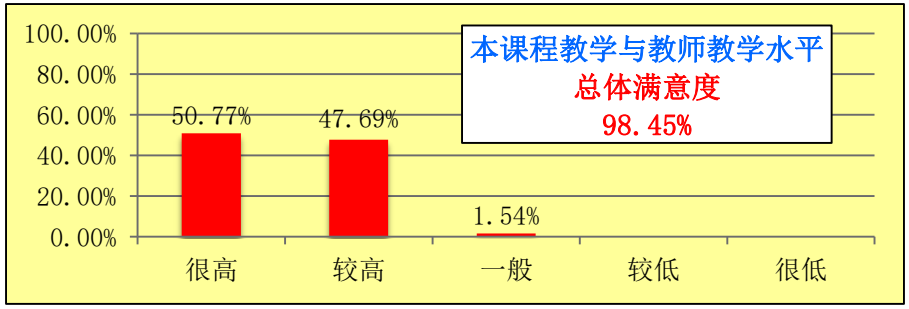

讲课认真态度
满意度：100%
100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
66.15%
33.85%
很认真
较认真
一般
不够认真
很不认真


讲课条理和清晰程度
满意度：93.31%
100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
55.38%
40.00%
4.62%
很清晰
较清晰
一般
不够清晰
很不清晰


讲课生动程度
满意度：81.25%
100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
16.92%
64.62%
18.46%
很生动
较生动
一般
不够生动
很不生动


实验内容的难度和合理性
满意度：90.77%
100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
21.54%
69.23%
9.23%
很好
较好
一般
较差
很差


学完课程后收益程度
肯定度：81.54%
100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
23.08%
58.46%
18.46%
很高
较高
一般
较低
很低


本课程教学与教师教学水平
总体满意度
98.45%
100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
50.77%
47.69%
1.54%
很高
较高
一般
较低
很低

教学效果

第一届“中国云/移动互联网创新大奖赛”

本课程开设后，课程同学组织了4支研究生代表队在“中国云产业联盟”组织的首届“中国云·移动互联网创新大奖赛”中参赛并荣获9项优胜奖（一等奖2项，二等奖4项，三等奖3项）和4项优秀领队奖，并获得大赛奖金20万元！占据大赛全部30个奖项中的9项，4道大数据赛题全部17个奖项中的8项！

教学效果

第一届“中国云/移动互联网创新大奖赛”

技术类赛题 1：调色板搜图—在百万图片中搜索与指定调色板相近的图片

技术类赛题 2：多快好省的速递员 — 动态路况环境下的物流规划

技术类赛题 3：你不知道我知道 — 互联网问答系统用户行为分析

技术类赛题 4：难舍难分 — 大规模搜索关键字（短文本）分类

技术类赛题 5：麻雀级云数据中心 — 规定时间内在小规模硬件环境上部署大量虚拟机

创意类竞赛说明： 创意类赛题没有具体的问题约束。

教学效果

第一届“中国云/移动互联网创新大奖赛”

本课程4支研究生代表队荣获9项优胜奖和4项优秀领队奖，获得奖金20万

教学效果

“全国高校云计算创新应用大赛”大数据技能赛

- 2015-2017年，荣获教育部主办的“全国高校云计算创新应用大赛”大数据技能赛冠军，实现该项赛事全国三连冠！

教学效果

荣获2012年度Google奖教金

由于在本课程建设教学工作中取得了显著的成绩，Google公司授予2012年度“Google奖教金”

课程联系群与课程邮箱

2017研究生大数据课程群

该二维码7天内(9月27日前)有效，重新进入将更新

2017研究生大数据课程

扫一扫二维码，加入该群。

课程邮箱： mapreduce_2017@163.com