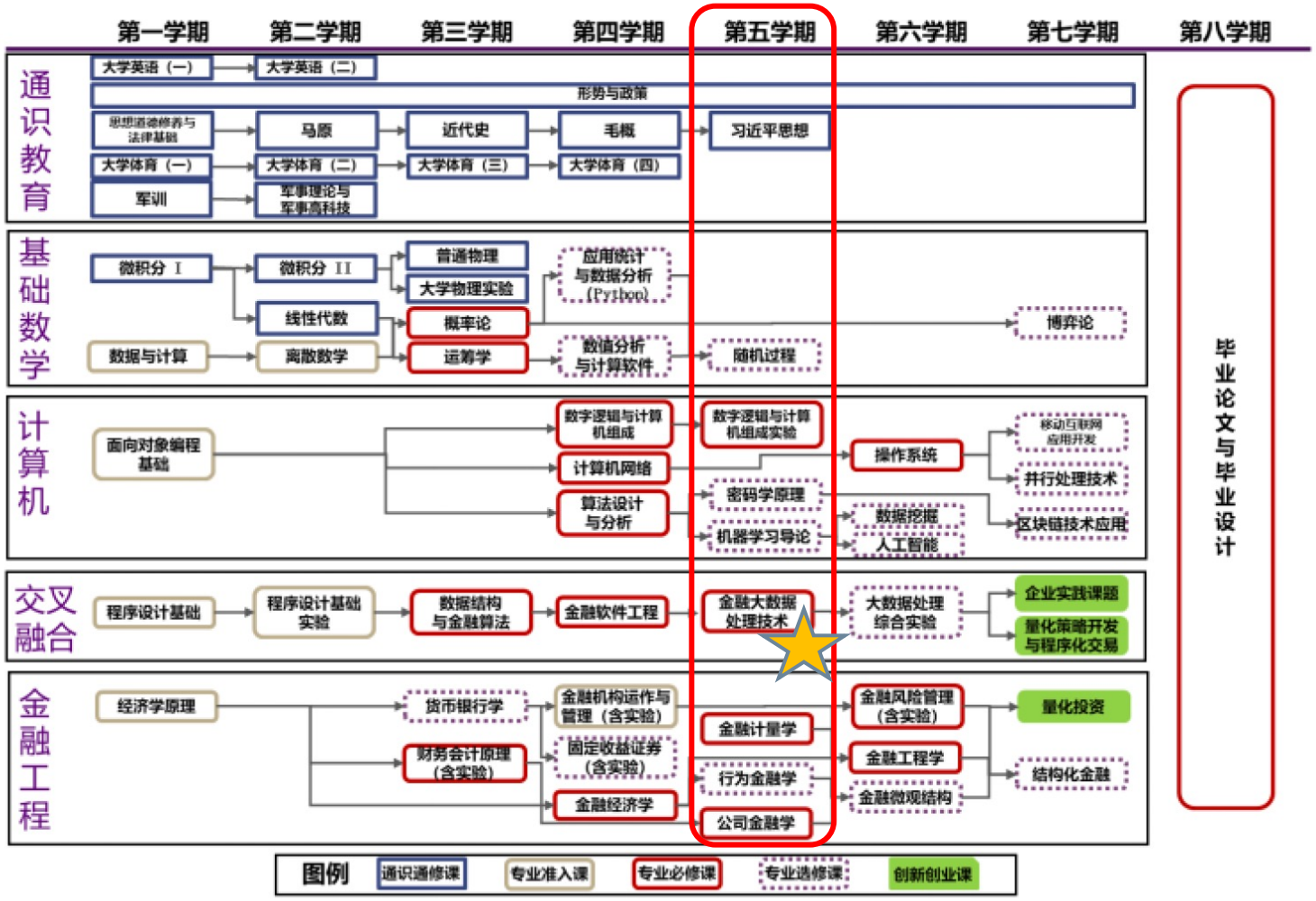
金融大数据处理技术

2023-2024 学年第一学期（秋季）

Review

计金班课程结构拓扑



大学生创新训练计划

\+

创新创业大赛

暑期社会实践

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。
- 学习Hadoop、Spark等大数据处理系统的基本组成和工作原理。
- 学习MapReduce和Spark并行程序设计和基础算法。
- 通过课程实验，熟悉Hadoop、Spark、HBase、Hive等大数据处理系统的安装、操作管理和使用。
- 通过课程实践，将大数据处理技术应用到金融领域的应用中。

教学目标

- 更深入地掌握大数据处理的基本原理
- 更广泛地了解大数据领域的新兴技术
- 更自信地面对金融科技领域的技术需求

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
- 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

教学安排

- 第1周：课程介绍&大数据技术简介
  - 概述课程情况、要求；
  - 简要介绍大数据技术背景和概念
- 第2周：并行计算技术简介
  - 简要介绍并行计算的主要技术问题，MPI并行程序设计，大规模并行数据处理技术
- 第3周：Google 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，GFS的基本构架与工作原理，BigTable的基本结构与工作原理
- 第4周： Hadoop MapReduce的基本构架
  - 介绍Hadoop MapReduce基本框架和工作原理，HDFS基本组成及工作原理
- 第5周： 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，以及程序开发流程

=>实验1

教学安排

- 第6周：MapReduce基础编程（I）
  - 介绍MapReduce所能处理的算法问题
- 第7周：MapReduce基础编程（II）
  - 介绍文档倒排索引、单词同现算法、专利文献数据分析应用
- 第8周：MapReduce高级编程
  - 介绍基于迭代的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

=>实验2

- 第9周：MapReduce数据挖掘基础算法
  - 介绍聚类、分类算法、频繁项集挖掘等数据挖掘经典算法的MapReduce设计和实现
- 第10周：NoSQL数据库
  - 介绍NoSQL数据库理论和典型的NoSQL数据库

教学安排

- 第11周：HBase基本原理与程序设计
  - 介绍HBase基本原理、基本操作和编程方法
- 第12周：Hive简介&大数据技术的最新进展 =>实验3
  - 介绍Hive基本原理和操作方法；
  - 介绍Alluxio、Fluid、数据湖等最新的大数据技术
- 第13周：Spark简介
  - 介绍Spark发展起源、设计理念和基本思想
- 第14周：Spark基础编程 =>实验4
  - 介绍Spark核心功能和编程方法
- 第15周：Spark高级编程
  - 介绍Spark上层组件，包括Spark SQL、Spark ML、Spark Steaming、GraphX
- 第16周：云计算&复习
  - 云计算技术简介；
  - 对本学期课程内容进行一次完整的复习整理讲解

实验

- Ex.1 Hadoop安装与运行
- Ex.2 MapReduce编程（银行贷款违约）
  - MapReduce基础编程
  - MapReduce数据挖掘
- Ex.3 HBase安装与运行
- Ex.4 Spark编程（银行贷款违约）
  - Spark基础编程
  - Hive操作或Spark SQL编程
  - 数据挖掘应用

课程内容

- Ch.1 大数据简介
  - 大数据背景
    - Scale up vs. Scale out
  - 什么是数据？什么是大数据？
  - 大数据的5V特征：Volume，Variety，Velocity，Veracity，Value
  - 大数据的类型
    - 结构特征；获取和处理方式；关联特征
  - 大数据涉及的关键技术
    - 存储，实时处理，高速传输，搜索，数据分析等
    - 新平台，新服务，新传输方案

课程内容

- Ch.2 并行计算和MPI基础编程
  - 提高计算机硬件性能的主要手段
  - 为什么需要并行计算?
  - 并行计算的分类
    - 按数据和指令处理结构；按并行类型；按存储访问构架；按系统类型；按计算特征；按并行程序设计模型/方法
  - MPI并行程序设计的特点
  - MPI通信机制
    - 点对点通信
    - 节点集合通信
    - 用户自定义的复合数据类型传输
  - MPI的不足

课程内容

- Ch.3 MapReduce简介
  - MapReduce的基本模型和处理思想
    - 对付大数据处理：分而治之
    - 构建抽象模型：Map和Reduce
    - 上升到构架：自动并行化并隐藏低层细节

课程内容

- Ch.4 Google MapReduce的基本架构
  - Google MapReduce
    - 基本工作原理
    - 失效处理，带宽优化，计算优化
  - GFS
    - 基本设计原则
    - 基本工作原理
  - BigTable基本工作原理
    - 设计目标
    - Data Model
    - 基本构架

课程内容

- Ch.5 Hadoop MapReduce基本架构
  - Hadoop平台的基本组成和生态系统
  - HDFS
    - 基本特征
    - 基本构架
    - 数据分布设计及设计要点
  - Hadoop MapReduce
    - 基本构架
    - 主要组件
    - MapReduce v1.0 vs. YARN (v2.0)
    - 容错及优化

课程内容

- Ch.6/7 MapReduce基础编程
  - MapReduce流水线
  - WordCount
  - 矩阵乘法
  - 关系代数运算
  - 排序算法
  - 二级排序
  - 单词同现
  - 倒排索引
  - 专利文献数据分析

课程内容

- Ch.8 MapReduce高级编程
  - 复合键值对的使用
  - 用户自定义数据类型
  - 用户自定义输入输出格式
  - 用户自定义Partitioner和Combiner
  - 迭代完成MapReduce计算
  - 链式MapReduce任务
  - 全局参数/数据文件的传递
  - 其它处理技术

课程内容

- Ch.9 基于MapReduce的搜索引擎算法
  - 图表示：邻接矩阵，邻接表
  - PageRank设计思想和设计原则
    - rank leak；rank sink
    - 随机浏览模型

课程内容

- Ch.10/11/12 MapReduce数据挖掘基础算法
  - K-Means聚类算法
  - KNN最邻近分类算法
  - 朴素贝叶斯分类算法
  - 决策树分类算法
  - 频繁项集挖掘算法

课程内容

- Ch.13 NoSQL数据库
  - NoSQL简介
  - NoSQL与RDBMS
  - NoSQL的四大类型
    - 键值数据库、列族数据库、文档数据库和图形数据库
  - NoSQL的三大基石
    - CAP、BASE、最终一致性
  - 从NoSQL到NewSQL

课程内容

- Ch.14 HBase基础原理与程序设计
  - HBase基本工作原理
    - HBase vs. RDBMS
    - 数据模型
  - 数据存储管理方法
    - 三级索引结构
  - 基本操作

课程内容

- Ch.15 Hive简介
  - RDBMS vs. Hive
  - HBase vs. Hive
  - Hive的体系结构
  - Hive的数据模型
  - Hive QL
    - DDL，DML，QUERY

课程内容

- Ch.16 Spark简介
  - Spark特点
  - Spark vs. Hadoop
  - Spark生态圈
  - Spark的基本构架和组件
  - Spark的技术特点

课程内容

- Ch.17/18 Spark基础编程
  - Spark安装与运行
  - Spark编程模型
    - RDD的操作、容错、依赖和持久化
    - 键值对操作
    - 共享变量
  - Spark编程实例
    - WordCount
    - K-Means

课程内容

- Ch.19/20 Spark高级编程
  - Spark SQL
  - Spark Mllib
  - Spark Streaming & Spark Structured Streaming
  - GraphX

课程内容

- Ch.21 云计算简介
  - 什么是云计算？云计算解决什么主要问题?
  - 云计算的主要特点
  - 云计算的分类
    - 按服务层面的分类：IaaS，PaaS，SaaS
    - 按系统类型的分类：公用云，私有云，社区云，混合云
  - 云计算的关键技术
  - 容器云
  - 云原生：属性；四要素
  - 数据湖 vs. 数据仓库

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
- 《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
- 《数据算法 Hadoop/Spark大数据处理技巧》，Mahmoud Parsian，2016，中国电力出版社
- 《Hadoop金融大数据分析》，Rajiv Tiwari，2017，电子工业出版社
- 《云计算》，刘鹏，2010，电子工业出版社

考核方式

- 平时10%
- 实验30%
- 期末笔试60%

考试题型

- 填空题（20分）：概念
- 简答题（20分）：概念与原理
- 论述题（60分）：分析与设计

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