金融大数据处理技术

2020-2021 秋季学期

相关信息

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

- 课程网页：
  http://cslabcms.nju.edu.cn/course/view.php?id=716

课程简介

- 课程简介
  - - “选不选这门课？”

- 背景概述
  - - “所学内容在专业知识结构中的位置？”

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

课程内容

- Ch.1 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 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。
- Ch.4 Hadoop的基本架构
  - 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，Hadoop数据管理系统的基本结构与工作原理。

课程内容

- Ch.5 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.6 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.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 复杂问题的MapReduce编程
  - 介绍图算法（如宽度优先搜索），PageRank（Web网页排序）等。

课程内容

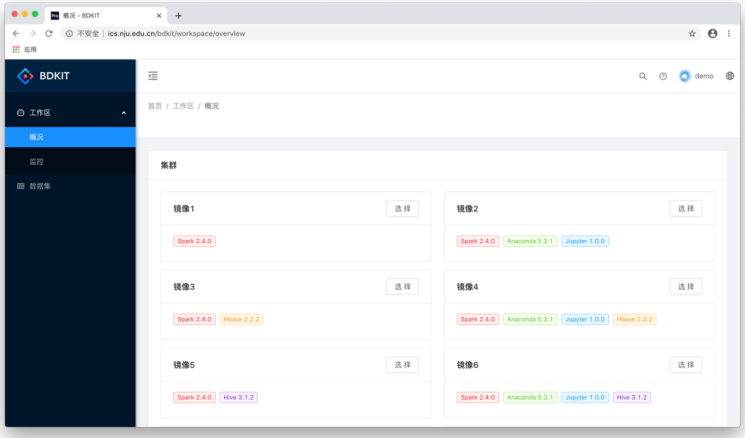
- Ch.9 MapReduce数据挖掘基础算法
  - 介绍聚类算法，分类算法，频繁项集挖掘等算法及应用。
- Ch.10 HBase基本原理与程序设计
  - 介绍HBase基本工作原理、基本操作和编程方法示例。
- Ch. 11 Hive简介
  - 介绍Hive基本原理、基本操作和程序设计。

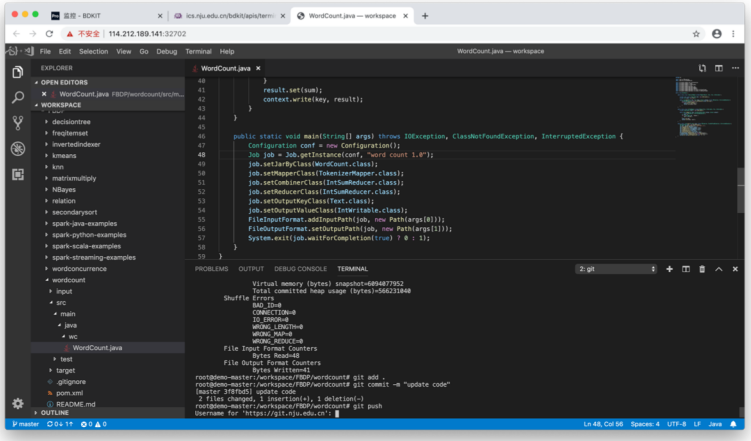
课程内容

- Ch.12 Spark简介
  - 介绍Spark的基本概念和原理，以及Spark的安装和运行。
- Ch.13 Spark基础编程
  - 介绍Spark基本编程模型，键值对操作，共享变量等编程技术。
- Ch.14 Spark高级编程
  - 介绍Spark Streaming，Spark SQL，Spark ML等技术。
- Ch.15 云计算技术简介
  - 介绍云计算的基本概念，大数据与云计算的关系，云资源的调度与管理，虚拟机与容器技术，以及云原生技术。
- Ch.16 金融大数据应用编程案例
  - 通过金融大数据应用案例分析，深入理解大数据处理技术。

在线实验平台

- BDKit
  - **BDKit是一个大数据应用开发和运行支撑平台，支持Hadoop集群的创建、MapReduce/Spark/Hive等大数据应用的在线开发、运行、监控等功能。**
  - 基础功能：根据前端发送的创建请求和具体设置在Kubernetes集群中生成一个带有并行计算环境的用户集群，其中包括一个主节点和多个子节点，并自动监控、维护用户集群的稳定运行。





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
- 《未来架构：从服务化到云原生》
- 《持续演进的Cloud Native：云原生架构下微服务最佳实践》

考核方式

- 平时10%
- 实验30%
- 期末笔试60%

FBDP2020

扫一扫二维码，加入群聊。

QQ