【大数据应用开发】

【Big Data Application Development】

一、基本信息

课程代码：【2055018】

课程学分：【2】

面向专业：【软件工程】

课程性质：【选修】

开课院系：信息技术学院 软件工程系

使用教材：

教材【《大数据技术基础与实践》薛志东 主编. 人民邮电出版社.2021.9】

参考书目【《大数据技术原理与应用-概念、存储、处理、分析与应用》（第 2 版）.林子雨编著.人民邮电出版社.2017.1】

【《深入理解大数据-大数据处理与编程实践》.黄宜华 主编.机械工业出版社.2014.8】

【《Hadoop 大数据开发实战》. 杨力 主编. 人民邮电出版社 2019.2 】

课程网站网址：

先修课程：【面向对象程序设计 2050218（3）】

【数据库原理 2050217（3）】

二、课程简介

大数据技术基础与实践是计算机专业学生的一门重要专业拓展选修课。本课程中内容覆盖全面、讲解详细，其中第 1 章大数据概述 首先介绍大数据的概念和特性、大数据的处理流程，接着介绍了 Hadoop 大数据技术，最后介绍了配置基本实践环境；第 2 章主要讲解 Linux 基础和集群搭建；第 3 章 Hadoop 集群配置；第 4～第 5 章讲解 HDFS 分布式文件系统、MapReduce 分布式编程；第 6 章～第 9 章主要讲解 Hadoop 生态圈中的相关辅助系统，包括 Hive、HBase 分布式存储系统、Flume、Spark；第 10 章利用大数据平台处理图像。

三、选课建议

本课程是软件工程专业、计算机科学与技术专业的选修课，建议在第五学期开设。

四、课程与专业毕业要求

五、课程目标/课程预期学习成果

序号	课程预期学习成果	课程目标（细化的预期学习成果）	教与学方式	评价方式
1	L024 在充分理解专业知识的基础上，能够运用所学专业知识并借助文献研究，获得解决问题的总体思路和方案	按照学习目标，课后可通过讨论的方式查找文献和资料，设计完成学习目标的学习计划	团队讨论、自主学习	平时作业
2	L032 能针对需求分析独立进行算法设计和程序实现，并能测试验证算法与程序的正确性	能够掌握 Hadoop IDE 开发环境及集群的搭建； 能够掌握 Hive 的工作原理和服务； 能够编写小模块	讲授、练习	课堂实验
3	L041 能够基于科学原理，结合软件行业，通过文献研究等相关方法，调研和分析软件系统	能够设计开发系统，并能分析所开发系统的优缺点	自主学习、团队学习	大作业

	设计问题			
4	L071 了解与本专业相关的职业和行业的生产、设计、研究与开发、环境保护和可持续发展等方面的方针、政策、法律和法规	能够将行业的开发方案应用于自己的实作中，能够体现实作系统的可持续性	课后阅读、自主学习、团队讨论	大作业

六、课程内容（根据实际更改）

第 1 单元 大数据技术概述

通过本单元学习，使学生知道

1）大数据的概念与基本特性

2）大数据处理流程

3）Hadoop 大数据技术

4）实践环境准备

本单元的重点和难点是大数据处理流程、Hadoop 大数据技术、实践环境准备　本单元 2 学时

第 2 单元 Linux 基础与集群搭建

通过本单元学习，使学生知道

1）Linux 常用命令

2）网络配置

3）Linux 集群配置

4）快速配置 Linux 集群

本单元的重点和难点是如何进行网络配置、 Linux 集群配置、快速配置 Linux 集群　本单元理论课时 2 课时，实践课时数 2 学时。

第 3 单元 Hadoop 集群配置

通过本单元学习，使学生知道

1）Hadoop 集群安装

2）Hadoop 集群初始化和日志查看

本单元的重点和难点是 HDFS 存储架构和数据读写流程。本单元理论课 2 课时，实践课时数 4 学时

第 4 单元 HDFS

通过本单元学习，使学生知道

1）HDFS 简介

2）HDFS 基本命令

3）HDFS 数据平衡优化

4）HDFS API 的使用方法

本单元的重点是掌握 HDFS 基本命令、HDFS 数据平衡优化、HDFS API 的使用方法。 本单元理论课时 2 课时，实践课时数 2 学时。

第 5 单元 MapReduce

通过本单元学习，使学生知道

1）认识 MapReduce

2）MapReduce 编程组件

3）MapReduce 作业解析

4）MapReduce 工作原理

5）Shuffle 阶段

6）优化—数据倾斜

7）MapReduce 典型案例—排序

8）MapReduce 典型案例—倒排索引

9）MapReduce 典型案例—连接

10）MapReduce 典型案例—平均分以及百分比

MapReduce 典型案例—过滤敏感词汇本单元的重点和难点是掌握理解 MapReduce 经典案例 WorldCount 的实现原理、掌握 MapReduce 运行流程、掌握 MapReduce 程序设计方法。本单元理论课时 2 课时，实践课时数 4 学时。

第 6 单元 Hive 大数据库仓库

通过本单元学习，使学生知道

1）Hive 简介

2）Hive 安装与配置

3）从创建数据库到创建表

4）数据查询及自定义函数运算

5）Hive 自定义函数编程

本单元的重点是掌握 Hive 安装与配置、从创建数据库到创建表、数据查询及自定义函数运算、Hive 自定义函数编程。 本单元理论课时 2 课时，实践课时数 4 学时。

第 7 单元　HBase 数据库部署与操作

通过本单元学习，使学生知道

1）认识 HBase

2）HBase 表设计

3）HBase 安装

4）HBase Shell 常用操作

5）HBase 编程

6）HBase 过滤器和比较器

7）HBase 与 Hive 结合

8）HBase 性能优化

本单元的重点是掌握 HBase 架构及其原理、掌握 HBase 的存储流程、HBase 的安装和使用理解 HBase 与 Hive 之间的关.系 本单元理论课时 2 课时，实践课时数 4 学时

第 8 数据获取与 Flume 应用

通过本单元学习，使学生知道

1）公开数据资源获取

2）使用网络爬虫获取数据

3）使用 Flume 获取数据

4）综合案例

本单元的重点是掌握使用网络爬虫获取数据使用 Flume 获取数据。本单元理论课时 2 课时，实践课时数 2 学时。

第 9 单元　基于 Spark 的内存计算

通过本单元学习，使学生知道

1）Spark 简介

2）Spark 快速部署

3）Spark 程序

4）Spark RDD 编程

5）Spark 生态系统

6）Spark 应用案例

本单元的重点是掌握 Spark 快速部署、Spark 程序、Spark RDD 编程.本单元理论课时 2 课时，实践课时数 4 学时

第 10 单元 利用大数据平台处理图像

通过本单元学习，使学生知道

1) 图像的基本概念

2) Hadoop 处理图像的问题与对策

3) HIPI 安装与部署

4) 使用 HIPI 进行图像处理

5) HIPI 工具 hibDownload

本单元的重点是掌握HIPI安装与部署、使用HIPI进行图像处理、HIPI工具hibDownload .本单元理论课时 2 课时，实践课时数 2 学时

七、课内实验名称及基本要求

列出课程实验的名称、学时数、实验类型（演示型、验证型、设计型、综合型）及每个实验的内容简述。

序号	实验名称	主要内容	实验时数	实验类型	备注
1	Hadoop 集群搭建	虚拟机安装、克隆、网络配置、SSH 服务配置、Hadoop 集群搭建、Zookeeper 安装、Hadoop 的 HA 模式	4	验证型	1 台 PC 机/1 人；
2	案例编程	Java HDFS 、MapReduce、Hive、Hbase 编程各 2-3 个案例做成模块（1 个项目）	8	设计型	同上
3	综合案例	完成一个综合性的大数据应用程序的设计和编程	20	综合型	同上

八、评价方式与成绩

总评构成（3X）	评价方式	占比
X1	大作业	60%
X2	实验报告	25%
X2	课堂实验	15%

撰写人：宋建虎　　系主任审核签名：朱丽娟

审核时间：2024 年 8 月 31 日