【大数据基础概论】

【Introdution to Big Data Foundation】

一、基本信息

课程代码：【2050545】

课程学分：【3】

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》.宋旭东 主编.清华大学出版社.2020.6】

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】

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

本课程是软件工程专业的一门专业选修课程。本课程的任务使学生了解大数据基本概念、大数据框架体系、大数据采集与预处理技术、数据存储和管理技术、数据分析与挖掘技术、数据可视化等技术，以及大数据并行计算框架Hadoop平台及其核心组件，为其未来的研究和应用打下坚实基础。

三、选课建议

本课程是软件工程专业的选修课，建议在第六或第七学期开设。

四、课程与专业毕业要求的关联性

软件工程专业毕业要求	关联
L02：问题分析：能够应用数学、自然科学和工程科学的基本原理，识别、表达、并通过文献研究分析复杂工程问题，以获得有效结论	●
L03：设计/开发解决方案：能够设计针对复杂工程问题的解决方案，设计满足特定需求的系统、单元（部件）或工艺流程，并能够在设计环节中体现创新意识，考虑社会、健康、安全、法律、文化以及环境等因素	●
L04：研究：能够基于科学原理并采用科学方法对复杂工程问题进行研究，包括设计实验、分析与解释数据、并通过信息综合得到合理有效的结论	●
L07：环境和可持续发展：能够理解和评价针对复杂工程问题的专业工程实践对环境、社会可持续发展的影响	●

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	实验报告
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设计问题	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
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

第 1 单元 大数据概论

通过本单元学习，使学生知道

1）大数据的定义、特征、框架体系、关键技术

2）大数据平台 Hadoop 概述、原理及其组件

3）熟悉 Hadoop 安装，配置及使用

本单元的重点和难点是如何正确搭建 Hadoop 的开发环境。本单元理论课时 4 课时，实践课时数 2 学时。

第 2 单元 大数据存储技术

通过本单元学习，使学生知道

1)大数据的数据类型，数据管理技术的发展，NoSQL 数据库，大数据存储与管理技术

2)大数据分布式文件系统 HDFS 概述、工作原理

3)HDFS 文件操作命令

4)HDFS 工作流程及编程接口

5)大数据分布式数据库系统 HBase 数据模型，工作原理

6)HBase 安装及 HBase 操作命令及编程接口

7)大数据分布式数据仓库系统 Hive 数据模型、HiveSQL 查询语法

本单元的重点和难点是如何灵活且正确地使用 HDFS、Hbase 和 Hive。本单元理论课时 6 课时，实践课时数 2 学时。

第 3 单元 大数据采集与预处理

通过本单元学习，使学生知道

1)大数据采集与预处理技术、大数据采集工具

2)Sqoop 采集工具的使用

3)大数据计算模式，包括批处理、查询分析、流计算、迭代计算、图计算等

本单元的重点和难点是掌握大数据的计算模式。本单元理论课时 6 课时，实践课时数 2 学时

第 4 单元 MapReduce

通过本单元学习，使学生知道

1）大数据 MapReduce 模型框架、数据处理过程

2）熟悉 WordCount MapReduce 程序结构及程序运行测试

3）大数据 MapReduce 程序执行过程及编程接口

本单元的重点和难点是掌握 MapReduce 的工作原理和编程技术。本单元理论课时 6 课时，实践课时数 2 学时。

第 5 单元 Spark

通过本单元学习，使学生知道

1）大数据 Spark 工作原理及访问接口

2）Spark 环境下 WordCount 编程实现

3）Scala 编程语言

本单元的重点是掌握 Scala 编程技术和 Spark 体系的工作原理。本单元理论课时 10 课时，实践课时数 4 学时。

第 6 单元 Hadoop 大数据平台实践

本单将要求学生独立完成一个基于大数据技术的小型应用程序的开发，通过本单元，学生应将之前学习的各种技术进行综合实践。本单元理论课时 2 课时，实践课时数 4 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 平台的安装和配置	在 windows 环境下安装和配置 Hadoop 平台及集群搭建	2	验证型	1 台 PC 机/1 人；
2	Hive 安装配置及服务使用	Hive 相关服务的使用	4	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	10	综合型	同上

八、评价方式与成绩

总评构成（1+X）	评价方式	占比
1	大作业	60%
X2	实验报告	25%
X3	课堂实验	15%

撰写人：朱茂坤　　系主任审核签名：朱丽娟

审核时间：2024 年 1 月 3 日