【大数据基础概论】

【Introduction 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）】

【Java 程序设计(双语)2050010（3）】

【数据库原理 2050217（3）】

二、课程简介

大数据基础。着重介绍大数据基本概念，大数据的4V特征及在其应用，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；大数据并行计算框架Hadoop平台及其核心组件。

大数据存储与管理。着重介绍大数据存储与管理的基本概念和技术，大数据数据类型，大数分布式系统基础理论，NoSQL数据库，分布式存储技术、虚拟化技术和云存储技术；大数据分布式文件系统HDFS,包括HDFS的设计特点,体系结构和工作组件;大数据分布式数据库系统HBase，包括HBase列式数据库的逻辑模型和物理模型，HBase体系结构及其工作原理；大数据分布式数据仓库系统Hive，包括Hive的工作原理和执行流程、Hive的数据类型与数据模型，以及Hive主要访问接口等。

大数据采集与预处理。着重介绍大数据采集与预处理相关技术，包括数据抽取、转换和加载技术，数据爬虫技术、数据清理、数据集成、数据变换和数据归约的方法和技术；大数据采集工具，包括Sqoop关系型大数据采集工具，Flume日志大数据采集工具和分布式大数据Nutch爬虫系统。

大数据分析与挖掘。着重介绍大数据计算模式，包括大数据批处理、大数据查询分析计算、

大数据流计算、大数据迭代计算、大数据图计算；大数据 MapReduce 计算模型、模型框架和数据处理过程，以及 MapReduce 主要编程接口；大数据 Spark 计算模型，包括 Spark 的工作流程与运行模式；大数据 MapReduce 基础算法和挖掘算法（这部分内容可选讲）。

大数据平台 Hadoop 实践与应用案例。着重介绍大数据 Hadoop 平台的实践操作；给出大数据技术在开敞式码头系泊缆力预测中的应用以及中科曙光 XData 大数据平台架构、关键技术及其智能交通应用案例（这部分内容可选讲）。

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的关联性

软件工程专业毕业要求	关联
L011：要求能领会用户诉求，正确表达自己的观点，具有专业文档的撰写能力	
L021：能根据环境需要确定自己的学习目标，并主动的通过搜集信息、分析信息、讨论、实践、质疑、创造等方法来实现学习目标。	●
L031：工程素养：掌握数学、自然科学知识，具有工程意识，能结合计算机、计算机网络相关专业知识解决复杂工程问题	
L032：软件开发：应用主流开发技术和程序设计思维对各类应用软件进行开发和实现的能力	●
L033：系统设计：应用软硬件基础理论知识及软件工程知识对软件系统进行分析设计、模块划分及整合能力	●
L034：软件测试：应用专业知识能够编写软件测试计划和测试报告能力，具备白盒测试、黑盒测试、自动化测试能力及测试管理能力	
L035：系统运维：应用软硬件和网络知识能够搭建软件应用环境、具备软件系统安全管理和维护能力	
L036：移动应用：应用主流移动平台开发工具实现移动应用软件开发能力、移动网络数据应用能力和新技术应用创新能力	
L041：遵守纪律、守信守责；具有耐挫折、抗压力的能力	●
L051：能与团队保持良好关系，积极参与其中，保持对信息技术发展的好奇心和探索精神，具有创新性解决问题的能力	●
L061：能发掘信息的价值，综合运用计算机相关的专业知识和技能，解决实际问题	●
L071：愿意服务他人、服务企业、服务社会；为人热忱，富于爱心	
L081：具有基本外语表达沟通能力，积极关注发达国家和地区信息技术发展新动向	

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	L0211 能根据需要确定学习目标，并设计学习计划。	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	学习报告
2	L032 软件开发：应用主流开发技术和程序设计思维对各类应用软件进行开发和实现的能力。	1. 能够掌握 Hadoop IDE 开发环境的搭建	讲授、练习	课堂展示
		2. 能够掌握 MapReduce 的工作原理和编程技术	讲授、练习	课堂展示
		3. 掌握 Scala 语言的核心语法和编程技术	讲授、练习	课堂展示
		4. 掌握 Spark 平台的安装和配置	讲授、练习	课堂展示
3	L0412 诚实守信：为人诚实，信守承诺，尽职尽责。	能够以团队的形式帮助团队中其他学习有困难的同学，帮助他们战胜学习上的困难，培养他们学习兴趣和开发能力	自主学习、团队学习	自我评估 同辈评估
4	L0514 了解行业前沿知识技术。	能够利用课后的扩展阅读，了解行业的前沿知识技术，并能通过团队的力量进行协作学习、共同探究了解到的前沿知识技术，并能在软件或软件的某一模块中运用	课后阅读、自主学习、团队讨论、协作开发	实作评估

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 4 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 4 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 4 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 12 课时，实践课时数 6 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	2	设计型	1 台 PC 机/1 人；
2	大数据存储技术	在 HDFS 和 Hbase 中存储和操作数据	2	设计型	同上
3	MapReduce 编程	使用 MapReduce 算法实现 WordCount 功能	2	设计型	同上
4	Saprk 平台搭建和配置	在 Windows 上搭建和配置 Spark 平台	2	设计型	同上
5	Scala 编程技术	使用 Scala 语言编写基于 Spark 平台的应用程序	4	设计型	同上

6	综合应用	完成一个综合性的大数据应用程序的设计和编程	4	设计型	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	平时作业	25%
X3	课堂表现	15%

“1”一般为总结性评价，“X”为过程性评价，“X”的次数一般不少于 3 次，无论是“1”、还是“X”，都可以是纸笔测试，也可以是表现性评价。与能力本位相适应的课程评价方式，较少采用纸笔测试，较多采用表现性评价。

常用的评价方式有：课堂展示、口头报告、论文、日志、反思、调查报告、个人项目报告、小组项目报告、实验报告、读书报告、作品（选集）、口试、课堂小测验、期终闭卷考、期终开卷考、工作现场评估、自我评估、同辈评估等等。**一般课外扩展阅读的检查评价应该成为“X”中的一部分。**

同一门课程由多个教师共同授课的，由课程组共同讨论决定 X 的内容、次数及比例。

撰写人：刘俊　　系主任审核签名：朱丽娟

审核时间：2022 年 9 月 10 日