《大数据计算技术》课程教学大纲

课程英文名称：Big Data Computing Technology

课程代码：　　学 时 数 ：32　　学 分 数：2

课程类型： 专业选修课

适用学科专业： 软件工程专业各培养方向

先修课程：数据结构与算法、数据库原理与应用

执 笔 者：林迪　　**编写日期：**2018.08　**审 核 人：** 汤羽

一、课程简介

本课程是针对软件工程和信息技术相关专业开设的数据分析计算课程。通过本课程的学习，学生将对大数据分析的价值、意义和基本原理建立清晰和比较全面的认识，掌握有关数据发掘、处理、建模的基本原理和方法，学习大数据处理的计算架构、计算模型及开发技术，了解和熟悉大数据分析和计算技术在医疗卫生、金融、电子商务、公共管理等领域的实际应用案例。

This course is a basic course of data analysis for information technology related programs. In this course, students will establish a clear and comprehensive understanding of the value, meaning and basic principles of large-scaled data analysis, master the basic principles and methods of data mining, processing and modeling, study the computing architecture, computing model and development technologies in big data processing, acquire the methods of data analysis in health care, Finance, E-commerce, public management and the other application fields.

二、课程目标

课程目标 (CO)	CO1：建立大数据的基本概念，了解大数据分析的要素和设计原理 Establish the fundamental concepts of big data and understand the key attributes and design principles CO2：理解数据分析的特点，并掌握基于 Python 和 Java 语言的数据分析方法 Understand the characters of various data analysis and master the analysis methods based on the programming language of Python and Java. CO3：理解分布式数据计算的设计原理及实现方法 Understand the design of distributed data computing and its implementation. CO4：通过课程学习和完成课程项目，初步具备按照业务需求进行数据分析算法设计和实现的能力 Develop the preliminary capability of data analysis and its implementation through the lectures and a course

	project on the requirements of data analysis.
课程模块 (CM)	**CM1：大数据的基本概念与模型** 大数据的基本概念，大数据分析的基本要素和设计原理。 **CM2：大数据计算体系与模式** 各种大数据计算的模式特点及设计原理。 **CM3：数据采集方法** 数据采集的来源、方法和步骤内容。 **CM4：数据清洗和分析方法** 数据清洗、数据规约、数据分析方法。 **CM5：数据处理及分析技术** 文本读写技术、数据处理技术、数据分析技术、可视化技术、大数据计算系统架构。 **CM6：海量数据处理平台** Hadoop 生态集群、MapReduce 方法、图并行计算框架、交互式计算框架、流计算系统、内存计算模式、Spark 计算平台。 **CM7：课程作业(Project)** 业务需求分析、分布式数据存储设计、数据清洗与规约方法设计、数据分析算法设计、数据分析报告。

培养目标	毕业要求	指标点	课程目标	课程模块
PO3	GR3	GR3.3	CO2	CM2, CM3, CM4
		GR3.4	CO3	CM3, CM5, CM7
PO5	GR6	GR6.1	CO4	CM6，CM7
PO9	GR3	GR3.4	CO2	CM3，CM6，CM7

课程达成度评价

课程目标	考核方式					考核标准	权重系数	考核模块
	考试	考查	作业	实验	其他			
CO1	√					百分制	0.10	CM1
CO2	√					百分制	0.20	CM2, CM3，CM4
CO3	√					百分制	0.30	CM5, CM6
CO4				√		百分制	0.40	CM6，CM7

达成度评价方式	CO1 达成度	(期末考试达成度) * 1.0
	CO2 达成度	(期末考试达成度) * 1.0
	CO3 达成度	(期末考试达成度) * 1.0
	CO4 达成度	(课程实验达成度) * 1.0
课程达成度	(CO1 达成度)*0.1 + (CO2 达成度)*0.2 + (CO3 达成度)*0.3 + (CO4 达成度)*0.4	

指标点达成度评价

指标点	*权重系数	考核方式					考核模块
		考试	考查	作业	实验	其他	
GR3.3	0.20	√			√		CM2, CM3, CM7
GR3.4	0.20	√					CM3, CM4，CM5
GR6.1	0.50				√		CM6, CM7
达成度评价方式	GR3.3 达成度	(考试达到值/考试预期值) * 0.4 + (实验达到值/实验预期值) * 0.6 =					
	GR3.4 达成度	(考试达到值/考试预期值) * 1.0					
	GR6.1 达成度	(实验达到值/实验预期值) * 1.0					

*此权重系数指本课程对某项指标点达成度（一个指标点的达成度通常由多门课程支撑）的贡献度，由表 5.1.7（毕业要求与高关联课程的支撑关系）定义。在此处此权重系数仅表示本课程对支撑某项指标点达成度的重要性，而非本课程分配该该指标点的权重比例。

三、教学计划

（一）教学内容、要求及教学方法

第 1 讲　大数据概念及计算技术简介

课程模块：CM1

学时分配：2 学时

教学方法：课堂面授

教学要求：对本课程的教学目标、内容、方式做一个全面概要介绍

教学内容：本章主要让学生了解数据科学的发展背景和数据科学所要解决的问题，介绍大数据的概念，以及大数据在现代服务行业的应用情况。本章重点为大数据的概念和数据科学的发展历史。

第 2 讲　大数据计算体系与模式

课程模块：CM2

学时分配：2 学时

教学方法：课堂面授

教学要求：介绍大数据存储系统和数据处理平台

教学内容：让学生了解主要的大数据存储系统，包括数据的清洗、建模、分布式文件存储、NoSQL数据库、数据访问接口。向学生介绍目前数据工程界采用的主要数据处理平台，通过实例介绍各类数据分析算法的特点和功能，使学生初步了解计算处理模型和计算平台引擎。本章重点为大数据存储系统和数据处理平台。

第 3 讲　数据采集方法

课程模块：CM3

学时分配：2 学时

教学方法：课堂面授

教学要求：讲授数据的采集方法和数据接口

教学内容：讲授内容包括日志数据的采集、互联网数据的采集等，让学生掌握网络爬虫技术。本章重点为互联网数据采集。

第 4 讲　数据清洗与规约方法

课程模块：CM4

学时分配：2 学时

教学方法：课堂面授

教学要求：介绍数据预处理技术、数据清洗技术、数据规约技术的基本原理和方法

教学内容：讲授内容包括数据预处理技术、数据清洗技术、数据规约技术的基本原理和方法等，掌握各类数据质量问题以及相应的数据清洗及规约方法。本章重点为数据清洗和规约方法。

第 5 讲　数据分析算法

课程模块：CM4

学时分配：2 学时

教学方法：课堂面授

教学要求：讲授常用的数据分析算法的原理

教学内容：讲授内容包括常用的数据分析算法的原理，并比较不同数据分析算法之间的区别，让学生掌握各种数据分析方法的原理，并能够选择适当的方法解决数据科学中的问题。本章重点为常用的数据分析算法的原理。

第 6 讲　文本读写技术

课程模块：CM5

学时分配：2 学时

教学方法：课堂面授

教学要求：讲授文本读写技术的工作原理及方法

教学内容：让学生掌握文本读写技术的组成特点，了解常见的文本读写技术的特点，掌握读取文件、写入文件、连接数据库的方法等。本章重点为文本读写技术的工作原理。

第 7 讲　数据处理技术

课程模块：CM5

学时分配：2 学时

教学方法：课堂面授

教学要求：介绍数据处理技术的基本原理及主要方法

教学内容：让学生了解数据处理技术的概念和特点，了解其基本原理、主要功能特点等，让学生对数据处理技术有一个初步理解。本章重点为数据处理技术的基本原理。

第 8 讲　数据分析技术

课程模块：CM5

学时分配：2 学时

教学方法：课堂面授

教学要求：介绍数据分析技术的概念、算法及应用场景

教学内容：让学生了解数据分析技术的概念和特点，了解其原理、算法、应用场景等，让学生对数据分析算法体系有一个初步理解。本章重点为数据分析算法的基本原理。

第 9 讲　数据可视化技术

课程模块：CM5

学时分配：2 学时

教学方法：课堂面授

教学要求： 讲解数据可视化技术的基本原理和主要功能

教学内容：讲授数据可视化技术的基本原理和主要功能,介绍数据可视化技术的应用场景。本章重点为数据可视化技术的基本原理。

第 10 讲　Hadoop 生态系统

课程模块：CM6

学时分配：2 学时

教学方法：课堂面授

教学要求：介绍Hadoop生态系统架构、HDFS分布式文件系统、分布式存储架构

教学内容：讲解 Hadoop 生态系统架构的基本框架、重点介绍 HDFS 分布式文件系统、分布式存储架构、HBase 索引与检索、资源管理与作业调度。

第 11 讲　MapReduce 计算模型

课程模块：CM6

学时分配：2 学时

教学方法：课堂面授

教学要求：讲解分布式并行计算系统、MapReduce计算架构、以及键值对的映射

教学内容：讲授分布式并行计算系统、MapReduce 计算架构、键值对与输入格式、映射与化简、应用编程接口。

第 12 讲　图并行计算框架

课程模块：CM6

学时分配：2 学时

教学方法：课堂面授

教学要求：讲解图的基本概念、图并行计算框架

教学内容：讲授图的基本概念、BSP 模型、Pregel 图计算引擎、Hama 开源框架、应用编程接口。

第 13 讲 交互式计算模式

课程模块：CM6

学时分配：2 学时

教学方法：课堂面授

教学要求：介绍数据模型、存储结构、并行查询算法实现

教学内容：讲授交互式计算中数据模型、存储结构的设计原理和功能特点，让学生掌握数据并行查询算法的实现方法，并介绍几种主要的开源框架的编程实现。

第 14 讲 流计算系统

课程模块：CM6

学时分配：2 学时

教学方法：课堂面授

教学要求：介绍流计算模型、Storm计算架构及其工作机制

教学内容：讲授流计算模型、Storm 计算架构及其工作机制、Storm 编程接口等。

第 15 讲 内存计算模式

课程模块：CM6

学时分配：2 学时

教学方法：课堂面授

教学要求：介绍内存计算模式中的分布式缓存体系、内存数据库等

教学内容：讲授内存计算模式中的分布式缓存体系、内存数据库、内存云 MemCloud、Spark 内存计算。

第 16 讲 Spark 计算平台

课程模块：CM6

学时分配：2 学时

教学方法：课堂面授

教学要求：讲授Spark平台内存计算模型及计算架构

教学内容：让学生了解大数据标准体系的内容及原理，掌握数据计算模式的基本要点及设计的基本方法。本章重点为数据计算模式的基本要点。

（二）自学内容和要求

- Linux 编程基础
- Python 编程基础

（三）实践性教学环节和要求

本课程有 4 个学时的课外实验，主要内容为采用 Hadoop 平台及 python 开发环境动手完成一个大数据分析的课程设计。

实验 1　Hadoop 单机安装配置

课程模块：CM7

学时分配：2 学时

教学方法：课外实验

教学要求：介绍hadoop单机模式安装

教学内容：默认情况下运行为一个单独机器上的独立Java进程，主要用于调试环境

实验 2　Mapreduce 实现 Wordcount 实例

课程模块：CM7

学时分配：2 学时

教学方法：课外实验

教学要求：基于MapReduce思想，编写WordCount程序

教学内容：基于MapReduce平台，实现Wordcount实例测试及源代码透析。

实验 3　Spark 安装部署

课程模块：CM7

学时分配：2 学时

教学方法：课外实验

教学要求：介绍Spark的安装与部署

教学内容：Spark的安装、部署、运行

实验 4　Mapreduce 和 Spark 的计算性能比对案例

课程模块：CM7

学时分配：2 学时

教学方法：课外实验

教学要求：基于MapReduce和Spark思想，编写WordCount程序

教学内容：使用Hadoop MapReduce、Spark框架分别运行wordcount分析程序，来对MapReduce和Spark的性能进行对比。

四、考核方式

本课程的考核方式为平时考核 (课程实验) (40%)，期末考核 (60%)。 期末考试为开卷考试。

五、建议教材及参考资料

（一）教材：

《大数据分析与计算》 汤羽、林迪、范爱华、吴薇薇 编著，清华大学出版社，第 1 版，2017

（二）参考资料：

1.《利用 Python 进行数据分析》 麦金尼 编著，机械工业出版社， 第 1 版，2014

2. 《大数据时代的算法: 机器学习、人工智能及其典型实例》 刘凡平 编著，电子工业出版社，第 1 版，2017