

《基于 Hadoop 的应用开发技术》

课程教学大纲

一、课程基本信息

课程类型	总学时为学时数	☒ 理论课（含上机、实验学时）			
	总学时为周数	☐ 实习 ☐ 课程设计 ☐ 毕业设计			
课程编码	7294511	总学时	32	学分	2
课程名称	基于 Hadoop 的应用开发技术				
课程英文名称	Hadoop-based application development technology				
适用专业	数据科学与大数据技术				
先修课程	（7030701）高等数学、（7101201）线性代数、（7029501）概率论与数理统计				
开课部门	信息学院人工智能系（大数据）				

二、课程性质与目标

本课程为数据科学与大数据技术专业大学本科生选修的一门课程，目的是让学生了解并掌握四个领域（即大数据系统的起源及系统特征、大数据系统的架构设计及功能目标设计、大数据系统程序开发、企业大数据案例分析）的内容，同时利用上机实验环节来提升学生对大数据开发的实践能力；

课程目标 1：学生应掌握 HDFS 使用操作、MapReduce 开发、HBase 数据库的开发、Hive 数据仓库开发、大数据案例分析等内容。

课程目标 2：学生应能从大数据分析技术中构建解决实际问题的方法，锻炼灵活运用经典算法的能力；同时从经典算法中得到启发，培养发现问题、解决问题的能力及思维创新能力。

课程思政目标：作为大数据开发的核心课程，将从科技应用方向和敬业精神方面开展，培养学生的科学精神、创新精神，坚定学生理想信念，厚植爱国主义情怀，弘扬社会主义精神。

三、课程教学基本内容及要求

第一章 大数据概述

（一）基本要求：

- 1、了解大数据概念、特征、数据计量单位以及大数据的类型；
- 2、了解大数据系统的设计背景、以及当前大数据系统存在的不足；

- 3、理解大数据系统的设计思想、设计目标和设计原则；
- 4、理解大数据系统的整体逻辑架构设计及运行逻辑，当前大数据系统的主流架构；

第二章 大数据应用开发思路和环境配置

（一）基本要求：

- 1、掌握大数据系统应用读写操作的开发流程；
- 2、掌握分析大数据开发技术及思路；
- 3、掌握大数据 Hadoop 环境配置；

第三章 HDFS 分布式文件系统

（一）基本要求：

- 1、理解 HDFS 设计目标、基本概念；
- 2、掌握 HDFS 文件系统的命令操作；
- 3、掌握对 HDFS 的程序开发操作，包含目录管路、文件列表、读取、导入导出、文件压缩等开发；
- 4、上机实操训练（实验环节 1）；

第四章 MapReduce 分布式编程

（一）基本要求：

- 1、了解 MapReduce 的设计思想、基本概念；
- 2、理解 MapReduce 的系统架构、作业运行机制和关键技术；
- 3、掌握 MapReduce 的数据类型的自定义以及数据类型的使用；
- 4、掌握 MapReduce 开发，定制输入输出的数据格式；
- 5、掌握将 HDFS 文件系统中整个文件作为输入数据的开发；
- 6、掌握利用 MapReduce 完成小文件聚合成一个大文件的开发；
- 7、掌握压缩数据处理程序开发；
- 8、掌握多数据源连接的开发，包含 Map 端开发以及 Reduce 端开发；
- 9、掌握 Hadoop 全局参数的使用，全局文件的使用；
- 10、上机实操训练（实验环节 2）

第五章 Hbase 分布式数据库

（一）基本要求：

- 1、了解 Hbase 分布式数据库的设计目标、基本概念；
- 2、理解 Hbase 逻辑架构以及物理架构
- 3、掌握 Hbase 分布式数据库 Shell 命令操作；
- 4、掌握 Hbase 数据库系统的开发，包含创建表、删除表、查询所有表操作；

- 5、掌握 Hbase 数据库系统的开发，包含插入记录、查询数据、组合查询、修改删除记录等开发；
- 6、上机实操训练（实验环节 3）；

第六章 Hive 数据仓库开发

（一）基本要求：

- 1、了解 Hive 数据仓库的工作原理及特点；
- 2、了解 Hive 架构设计，包含数据类型、数据存储方式以及查询方式；
- 3、掌握 Hive 数据仓库系统的 HQL 语言语法；
- 4、掌握 HQL 的创建表、查看表及修改表、删除表；
- 5、掌握利用 HQL 语句将 HDFS 的文件导入数据仓库；
- 6、掌握分区表、桶表、外部表的使用；
- 7、掌握 HQL 语句的联合查询、子查询、创建视图等操作；
- 8、掌握连接 Hive 数据仓库进行数据查询；
- 9、上机实操训练（实验环节 4）；

第七章 Mahout 算法

（一）基本要求：

- 1、了解 Mahout 算法的原理及特点；
- 2、掌握 Mahout 的配置及使用；
- 3、掌握运行一个 Mahout 案例；

四、课程学时分配

教学内容	讲授	实验	上机	课内学时小计	课外学时
1. 大数据概述	2	0	0	2	0
2. 大数据应用开发思路和环境配置	2	0	0	2	0
3. HDFS 分布式文件系统	4	2	0	6	0
4. MapReduce 分布式编程	4	2	0	6	0
5. Hbase 分布式数据库	4	2	0	6	0
6. Hive 数据仓库开发	4	2	0	6	0
7. Mahout 算法	4	0	0	4	0
合 计	24	8	0	32	0

五、 教学设计与教学组织

1、本课程以课堂讲授为主，辅以一定的讨论环节，充分调动学生的积极性，加深学生对知识点的理解。

2、使用 MS PowerPoint 幻灯片作为主要教学辅助工具，主要方法用软件演示其效果。

六、 教材与参考资料

1. 教材

《大数据技术丛书:Hadoop 应用开发技术详解》，刘刚著，机械工业出版社 2014 年第 1 版，9787111452447

七、 课程考核方式与成绩评定标准

本课程成绩由平时成绩及期末考试成绩二部分组成。课程成绩以百分制计算，平时成绩占 50%（其中出勤成绩占 10%，作业成绩占 40%），期末考试成绩占 50%。

八、 大纲制(修)订说明

无。

大纲执笔人：田青

大纲审核人：王彦平

开课系主任：王彦平

开课学院教学副院长：宋威

制（修）订日期：2022 年 2 月