【大数据开发技术】

【Big Data Development Technology】

一、基本信息

课程代码：【2055036】

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【《Hadoop 大数据开发实战》 慕课版. 千峰教育 主编. 人民邮电出版社.2021.8.6】

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】

【《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

Hadoop 大数据开发实战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 章首先让读者对大数据及 Hadoop 有一个总体的认识；第 2 章主要讲解如何搭建 Hadoop 集群；第 3～第 5 章讲解 HDFS 分布式文件系统、MapReduce 分布式计算框架以及 Zookeeper 分布式协调服务；第 6 章讲解 Hadoop 2.0 的新特性；第 7～第 10 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、Sqoop；第 11 章讲解了综合项目——电商精准营销，该项目涵盖从前期设计到最终实施的整个过程的内容，对全书知识点进行串联和巩固，使读者融会贯通，加深对 Hadoop 技术的理解。

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

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	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 初识 Hadoop

通过本单元学习，使学生知道

1) 大数据简介

2) 大数据技术的核心需求

3）Hadoop 简介

离线数据分析流程介绍

大数据学习流程

本单元的重点和难点是 Hadoop 简介、离线数据分析流程介绍、大数据学习流程　本单元 2 学时

第 2 单元 搭建 Hadoop 集群

通过本单元学习，使学生知道

1）安装准备

2）Linux 基本命令

3）Hadoop 集群搭建

4）Hadoop 集群测试

本单元的重点和难点是如何进行 Hadoop 集群搭建、Hadoop 集群测试、使用 Hadoop 集群 本单元实践课时 4 学时

第 3 单元 HDFS 分布式文件系统

通过本单元学习，使学生知道

1）HDFS 简介

2）HDFS 存储架构和数据读写流程

3）HDFS 的 Shell 命令

4）Java 程序操作 HDFS

5）Hadoop 序列化

6）Hadoop 小文件处理

7）通信机制 RPC

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 2 学时

第 4 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 2 学时。

第 5 单元 Zookeeper 分布式协调服务

通过本单元学习，使学生知道

1）认识 Zookeeper

2）Zookeeper 安装和常用命令

3）Zookeeper 客户端编程

4）Zookeeper 典型应用场景

本单元的重点是掌握　Zookeeper 安装和常用命令、Zookeeper 客户端编程、Zookeeper 典型应用场景。 本单元理论课时 1 课时，实践课时数 1 学时。

第 6 单元 Hadoop2.0 新特性

通过本单元学习，使学生知道

1） Hadoop2.0 的改进

2） YARN 资源管理框架

3） Hadoop 的 HA 模式

本单将要求学生熟悉 Hadoop2.0 的改进与提升、理解 YARN 架构的理由、理解 Hadoop 的 HA 模 本单元理论课时 1 课时，实践课时数 1 学时

第 7 单元 Zookeeper 分布式协调服务

通过本单元学习，使学生知道

1）数据仓库简介

2）认识 Hive

3）Hive 安装

4）Hive 数据类型

5）Hive 数据库操作

6）Hive 表

7）Hive 表的查询

8）Hive 函数

9）Hive 性能优化

10） Hive 案例分析

本单元的重点是掌握 Hive 的安装、Hive 架构及其原理、Hive 的数据库和表的操作方法、Hive 函数的使用、Hive 的性能调优。 本单元理论课时 2 课时，实践课时数 2 学时。

第 8 单元　HBase 分布式存储系统

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 2 学时

第 9 单元 Flume

通过本单元学习，使学生知道

1）认识 Flume

2）Flume 基本组件

3）Flume 安装

4）Flume 数据流模型

5）Flume 的可靠性保证

6）Flume 拦截器

本单元的重点是掌掌握 Flume 架构及其原理、掌握 Flume 的安装和使用.本单 元理论课时 1 课时，实践课时数 1 学时

第 10 单元 Sqoop

通过本单元学习，使学生知道

1) 认识 Sqoop

2) Sqoop 安装

3) Sqoop 命令

4) Sqoop 数据导入

5) Sqoop 数据导出

6) Sqoop job

本单元的重点是掌掌握 Sqoop 的安装及其安装、Sqoop 的框架、掌握 Sqoop 的 import、export、job 命令的用法.本单 元理论课时 1 课时，实践课时数 1 学时

第 11 单元 综合项目

通过本单元学习，使学生知道

1) 项目概述

2) 项目详细介绍

3) 项目模块分析

4) 数据采集

5) 数据清洗

6)使用数据仓库进行数据分析

7)可视化

本单元的重点是掌握项目背景及需求、项目中的架构设计、数据来源、数据清洗流程、数据仓库操作流程、应用测试方法 元理论课时 1 课时，实践课时数 1 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（1+X）	评价方式	占比
1	大作业	60%
X2	平时作业	25%

X3	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：2023 年 9 月 3 日