《Hadoop 大数据技术原理与应用》课程教学大纲

（课程英文名称）

课程编号：201800522062

学　　分：5 学分

学　　时：63 学时　（其中：讲课学时 51　上机学时：12）

先修课程：

后续课程：Spark

适用专业：大数据应用技术

开课部门：

一、课程的性质与目标

《Hadoop 大数据技术原理与应用》是互联网+创业教育学院软件工程（大数据、人工智能）专业的一门校定必修专业课。通过学习课程使得学生掌握大数据分析的主要思想和基本步骤，并通过编程练习和典型应用实例加深了解；同时对 Hadoop 平台应用与开发的一般理论有所了解，如分布式数据收集、分布式数据存储、分布式数据计算、分布式数据展示。

开设本学科的目的是让学生掌握如何使用大数据分析技术解决特定业务领域的问题。完成本课程学习后能够熟练的应用大数据技术解决企业中的实际生产问题。

二、教学条件要求

操作系统：Center OS

Hadoop 版本：Hadoop2.7.4

开发工具：Eclipse

三、课程的主要内容及基本要求

第 1 章　初识 Hadoop

章名	初识 Hadoop	学时	4
学习目标	1.　了解什么是大数据及其特征 2.　熟悉大数据的典型应用		

	3. 了解 Hadoop 的发展历史及其版本 4. 掌握 Hadoop 的生态体系			
知识点	了解	掌握	重点	难点
什么是大数据	√			
大数据的特征	√			
研究大数据的意义	√			
大数据的应用场景		√		
Hadoop 的发展历史	√			
Hadoop 的优势		√	√	
Hadoop 的生态体系		√	√	√
Hadoop 的版本	√			

第 2 章 构建 Hadoop 集群

章名	Hadoop 集群构建	学时	5	
学习目标	1. 了解虚拟机的安装和克隆 2. 熟悉 Linux 系统的网络配置和 SSH 配置 3. 掌握 Hadoop 集群的搭建和配置 4. 掌握 Hadoop 集群测试 5. 熟悉 Hadoop 集群初体验的操作			
知识点	了解	掌握	重点	难点
虚拟机安装	√			
虚拟机克隆	√			
Linux 系统网络配置		√		
SSH 服务配置		√		
Hadoop 集群部署模式		√		
JDK 安装		√		
Hadoop 安装		√		
Hadoop 集群配置		√	√	√
格式化文件系统		√		
启动和关闭 Hadoop 集群		√	√	
通过 UI 界面查看 Hadoop 运行状态		√	√	
Hadoop 集群初体验		√	√	

第 3 章 HDFS 分布式文件系统

章名	HDFS 分布式文件系统	学时	5
学习目标	1. 了解 HDFS 演变 2. 掌握 HDFS 特点 3. 掌握 HDFS 的架构和原理 4. 掌握 HDFS 的 Shell 和 Java Api 操作		

知识点	了解	掌握	重点	难点
HDFS 的演变	√			
HDFS 的基本概念		√		
HDFS 的特点		√	√	
HDFS 架构和原理		√	√	√
HDFS 的 Shell 操作		√		
HDFS 的 Java API 操作		√		

第 4 章 MapReduce 分布式计算系统

章名	MapReduce分布式计算框架	学时	8
学习目标	1. 理解 MapReduce 的核心思想 2. 掌握 MapReduce 的编程模型 3. 掌握 MapReduce 的工作原理 4. 掌握 MapReduce 常见编程组件的使用		

知识点	了解	掌握	重点	难点
MapReduce 核心思想		√		
MapReduce 编程模型		√	√	
MapReduce 编程实例——词频统计		√		
MapReduce 工作过程		√	√	√
MapTask 工作原理		√	√	
ReduceTask 工作原理		√	√	
Shuffle 工作原理		√	√	
MapReduce 编程组件		√	√	√
MapReduce 运行模式		√		
MapReduce 性能优化策略	√			
MapReduce 经典案例——倒排索引		√		
MapReduce 经典案例——数据去重		√		
MapReduce 经典案例——TopN		√		

第 5 章 Zookeeper 分布式协调服务

章名	Hadoop 进阶	学时	10
学习目标	1. 了解 Zookeeper 的概念和特性 2. 理解 Zookeeper 数据模型 3. 掌握 Zookeeper 的 Watch 机制和选举机制 4. 掌握 Zookeeper 的集群部署 5. 掌握 Zookeeper 的 Shell 操作和 Java API 操作 6. 熟悉 Zookeeper 的应用场景		

知识点	了解	掌握	重点	难点
Zookeeper 的简介	√			
Zookeeper 的特性	√			

Zookeeper 集群角色		√		
Zookeeper 的数据模型		√		
Zookeeper 的 Watch 机制		√	√	
Zookeeper 的选举机制		√	√	
Zookeeper 分布式集群部署		√	√	√
Zookeeper Shell 操作		√		
Zookeeper Java API 操作		√		
Zookeeper 典型应用场景		√		

第 6 章 Hadoop2.0 新特性

章名	Hadoop2.0 新特性	学时	3	
学习目标	1. 掌握 YARN 的体系结构和工作流程 2. 掌握 HDFS 的高可用架构 3. 会搭建 Hadoop 高可用集群			
知识点	了解	掌握	重点	难点
Hadoop2.0 改进与提升		√		
YARN 体系结构		√	√	
YARN 工作流程		√		√
HDFS HA 的搭建方式		√		
Hadoop 的高可用架构		√		
启动 Hadoop HA 方式		√		

第 7 章 Hive 数据仓库

章名	Hive 数据仓库	学时	7	
学习目标	1. 了解 Hive 的相关功能和特点 2. 熟悉 Hive 的简单安装和配置 3. 掌握 HiveQL 的相关操作			
知识点	了解	掌握	重点	难点
数据仓库简介	√			
数据仓库的结构	√	√		
数据仓库数据模型	√			√
Hive 简介		√		
Hive 系统架构		√	√	
Hive 工作原理		√	√	√
Hive 数据模型		√	√	
Hive 安装模式		√		
Hive 的管理方式		√		
Hive 内置数据类型		√		√
Hive 的操作方式		√		

第 8 章 Flume 日志采集系统

章名	Flume 日志采集系统		学时	6
学习目标	1. 了解 Flume 的作用 2. 熟悉 Flume 的运行机制 3. 掌握 Flume 的安装部署 4. 熟悉 Flume 的可靠性保证 5. 熟悉案例——日志采集的编写			
知识点	了解	掌握	重点	难点
Flume 简介	√			
Flume 运行机制		√	√	√
Flume 日志采集系统结构		√		
Flume 基本使用		√		
Flume 安装配置		√		
Flume Sources		√		
Flume Channels		√		
Flume Sinks		√		
Flume 负载均衡		√		
Flume 故障转移		√		
Flume 拦截器	√			

第 9 章 Azkaban 工作流管理器

章名	Azkaban 工作流管理器		学时	5
学习目标	1. 了解 Azkaban 的结构 2. 掌握 Azkaban 的部署 3. 熟悉 Azkaban 的基本使用			
知识点	了解	掌握	重点	难点
工作流管理器简介	√			
Azkaban 特点	√			
Azkaban 组织结构	√			
Azkaban 部署模式		√		
Azkaban 安装配置		√		
Azkaban 启动方式		√		
Azkaban Job		√		
Azkaban 工作流		√		
Azkaban 嵌入流		√		
依赖任务调度管理		√		

MapReduce 任务调度管理		√		
Hive 脚本任务调度管理		√		

第 10 章 Sqoop 数据迁移

章名	Sqoop 数据迁移	学时	3	
学习目标	1. 了解 Sqoop 基本概念 2. 掌握 Sqoop 安装配置 3. 熟悉 Sqoop 常用的相关指令 4. 掌握使用 Sqoop 进行导入导出			
知识点	了解	掌握	重点	难点
Sqoop 简介	√			
Sqoop 导入导出工作原理		√	√	
Sqoop 安装配置		√		
Sqoop 指令介绍		√		
MySQL 表数据导入 HDFS		√		
增量导入		√		
MySQL 表数据导入 Hive		√		
MySQL 表数据子集导入		√		
Sqoop 数据导出		√		

第 11 章 综合项目——网站流量日志数据分析系统

章名	综合项目——网站流量日志数据分析系统	学时	7	
学习目标	1. 熟悉日志分析系统的架构 2. 熟悉系统环境搭建的步骤 3. 掌握日志分析系统业务流程 4. 掌握人均浏览页面模块的实现方法			
知识点	了解	掌握	重点	难点
系统背景介绍	√			
系统架构设计		√	√	
模块开发-数据预处理		√		
模块开发-数据仓库开发		√		√
模块开发-数据分析		√		
模块开发-数据导出		√		√
模块开发-日志分析系统报表展示		√		

四、学时分配

章目	讲课	上机	合计
第 1 章　初识 Hadoop	4 学时	0 学时	4 学时
第 2 章　构建 Hadoop 集群	4 学时	1 学时	5 学时
第 3 章　HDFS 分布式文件系统	4 学时	1 学时	5 学时
第 4 章　MapReduce 分布式计算系统	7 学时	1 学时	8 学时
第 5 章　Zookeeper 分布式协调服务	7 学时	3 学时	10 学时
第 6 章　Hadoop2.0 新特性	2 学时	1 学时	3 学时
第 7 章　Hive 数据仓库	6 学时	1 学时	7 学时
第 8 章　Flume 日志采集系统	5 学时	1 学时	6 学时
第 9 章　Azkaban 工作流管理器	4 学时	1 学时	5 学时
第 10 章　Sqoop 数据迁移	2 学时	1 学时	3 学时
第 11 章 综合项目——网站流量日志数据分析系统	6 学时	1 学时	7 学时
合计	51 学时	12 学时	63 学时

五、考核模式与成绩评定办法

本课程为考试课程，期末考试采用百分制的闭卷考试模式。学生的考试成绩由平时成绩（30%）和期末考试（70%）组成，其中，平时成绩包括出勤（5%）、作业（5%）、上机成绩（20%）。

六、选用教材和主要参考书

本大纲是根据教材《Hadoop 大数据技术原理与应用》所设计的。

七、大纲说明

本课程的授课模式为：课堂授课+上机，其中，课堂主要采用多媒体的方式进行授课，并且会通过测试题阶段测试学生的掌握程度；上机主要是编写程序，要求学生动手完成指定的程序设计或验证。

撰写人：　　审定人：

批准人：　　执行时间：