《大数据开发技术（Hadoop）》教学大纲

课程编号：071333B

课程类型：□通识教育必修课 □通识教育选修课

□学科基础课 □专业核心课

☑专业提升课 □专业拓展课

总学时：48 讲课学时：16 实验（上机）学时：32

学　　分：3

考试类型：□考试 ☑考查

适用对象：计算机科学与技术专业、大数据专业，大三本科生

☑是 □否 适合作为其他专业学生的个性化选修课

先修课程：人工智能、面向对象程序设计、数据库原理、高性能计算

一、教学目标

目标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，同时利用真机实验环节以及大数据实训一体机来提升学生对大数据开发的实践能力；同时，我国在党中央的领导下，近年来在大数据领域取得了一系列的世界瞩目的成就，通过本门课的学习来树立学生对我国在大数据技术领域的认同感和民族自豪感。

本课程重点让学生掌握四个方面的内容：

目标1：HDFS使用操作

目标2：Map/Reduce开发

目标3：HBase数据库的开发

目标4：Hive数据仓库开发

二、教学内容及其与毕业要求的对应关系

**1：Hadoop**大数据处理平台安装与部署

目标及要求：让学生掌握Hadoop平台的安装，熟悉大数据处理的软硬件环境。

**2**：**HDFS**操作命令操作

目标及要求：让学生掌握分布式文件系统HDFS的文件操作，包含导入导出文件、列表、查找、删除文件等。

**3**：**MapReduce**开发

目标及要求：让学生掌握在HDFS文件系统内的Map及Reduce的Java开发，实现对指定文本文件的单词计数，将统计结果输出至控制台。

**4**：**HBase**数据库命令操作

目标及要求：让学生掌握分布式文件数据库系统HBase的数据库操作，包含创建表、删除表、增加列、导入记录、删除记录等。

**5**：**Hive**数据仓库基础使用

目标及要求：让学生掌握在Hive数据仓库的基本命令的操作，包含创建数据仓库、创建表、删除表、导入及导出数据，统计查询等在CLI模式下的使用操作。

三、各教学环节学时分配

**教学课时分配**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 序号 | 章节内容 | 讲课 | 实验 | 其他 | 合计 |
| 1 | 大数据概述 | 1 | 0 |  |  |
| 2 | 大数据应用开发思路和开发环境配置 | 1 | 6 |  |  |
| 3 | HDFS 分布式文件系统 | 2 | 6 |  |  |
| 4 | MapReduce 分布式编程 | 4 | 8 |  |  |
| 5 | HBase 分布式数据库 | 4 | 6 |  |  |
| 6 | Hive数据仓库开发 | 4 | 6 |  |  |
| **合计** | **48** | **16** | **32** |  |  |

四、教学内容

第一章 大数据概述

授课学时：1

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. 了解大数据系统的设计思想、设计目标和设计原则；

4. 了解大数据系统的整体逻辑架构设计及运行逻辑，了解当前大数据系统的主流架构；

5. 介绍我国近年来在党的领导下在超级计算机领域取得的成就。

第二章 大数据应用开发思路和开发环境配置

授课学时：1

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Java开发的环境配置、Plugin插件的安装，Hadoop环境配置；

4. 真机实操训练（实验环节1）；

第三章 **HDFS** 分布式文件系统

授课学时：2

基本要求：

1. 了解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Java对HDFS的程序开发操作，包含目录管理、文件列表、读取、导入导出、文件压缩等开发；

4. 真机实操训练（实验环节2）；

5. 介绍我国近年来在党的领导下在大数据技术领域取得的成就。

第四章 **MapReduce** 分布式编程

授课学时：4

基本要求：

1. 了解MapReduce的设计思想、基本概念；

2. 了解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. 掌握任务组合过程，掌握迭代组合、并行组合及串行组合；

9. 掌握任务的前后链式组合；

10. 掌握多数据源连接的开发，包含Map端开发以及Reduce端开发；

11. 掌握Hadoop全局参数的使用，全局文件的使用；

12. 掌握与关系型数据库的访问连接；

13. 真机实操训练（实验环节3）；

第五章 **HBase** 分布式数据库

授课学时：4

基本要求：

1. 了解HBase分布式数据库的设计目标、基本概念；

2. 了解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Java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Java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. 真机实操训练（实验环节4）；

7. 介绍我国近年来在党的领导下在国际互联网领域取得的成就。

第六章 **Hive**数据仓库开发

授课学时：4

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、查看表及查询有结构，修改表以及删除表；

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. 掌握利用Java开发UDF自定义函数，以及自定义函数的使用；

9. 掌握Java连接Hive数据仓库进行数据查询；

10. 真机实操训练（实验环节5）；

五、考核方式、成绩评定

考核方式：考查，采用 平时成绩+期末课程设计。

成绩评定：采用 平时成绩 占40%，期末课程设计 占60%。

六、主要参考书及其他内容

[1] 陆嘉恒. 分布式系统及云计算概论. 北京：清华大学出版社. 2010年.

[2] 刘军. Hadoop大数据处理. 北京：人民邮电出版社. 2013年.

[3] Tom White. Hadoop权威指南. 北京：清华大学出版社. 2015年.

执笔人：王汀 教研室主任：高静　　系教学主任审核签名：