**《大数据技术》教学大纲**

课程代码：05228913

课程名称：大数据技术

总 学 时：48学时（含实践教学16学时）

总 学 分：3学分

课程性质：专业必修课

适用专业：数据科学与大数据技术

**一、教学目的及任务**

**（一）教学目的**

本课程的教学目的旨在通过对大数据的相关知识的学习，使学生掌握大数据的概念和原理，熟悉大数据的理论与算法，了解大数据未来发展趋势，能够利用所学知识，进行大数据应用实现和算法设计，培养学生运用大数据技术解决大数据行业应用问题。

**（二）教学任务**

本课程的教学任务旨在学生通过本课程的学习，学生利用大数据组件能具备数据采集与预处理、数据挖掘算法与工具、R语言、深度学习以及大数据可视化等基本技能，并掌握大数据在互联网、商业和典型行业的应用。学生对大数据处理技术有比较深入的理解，能够从具体问题或实例入手，利用所学的大数据知识实现数据分析和数据挖掘等应用。

**二、教学内容和教学要求**

**第一章 大数据概述**

**（一）教学目标**

通过本章的学习，要求学生能够建立对大数据知识体系的轮廓性认识，了解大数据发展历程、基本概念、主要影响、应用领域、关键技术、计算模式和产业发展，并了解云计算、物联网的概念及其与大数据之间的紧密关系。

**（二）教学要求**

**1．教学重点：**

大数据的发展历程，大数据的基本概念和大数据的计算模式；云计算和物联网的概念。

**2．教学难点：**

大数据、云计算与物联网之间的关系。

**3．教学方法：**

讲授法、多媒体技术教学法和课堂讨论法。

**（三）教学内容**

1．大数据时代

（1）第三次信息化浪潮

（2）信息科技为大数据时代提供技术支撑

（3）数据产生方式的变革促成大数据时代的来临

（4）大数据的发展历程

2．大数据的概念

（1）数据量大

（2）数据类型繁多

（3）处理速度快

（4）价值密度低

3．大数据的影响

4．大数据的应用

5. 大数据关键技术

6. 大数据计算模式

7. 大数据与云计算、物联网

（1）云计算

（2）物联网

（3）大数据与云计算、物联网的关系

**第二章 大数据处理架构Hadoop**

**（一）教学目标**

通过本章的学习，要求学生能够了解Hadoop的发展历史、重要特性和应用现状，Hadoop项目结构及其各个组件，掌握Hadoop原理与运行机制；并熟练掌握Hadoop平台的安装和使用方法。了解当前的百度、阿里巴巴、腾讯、华为和中国移动的Hadoop平台。

**（二）教学要求**

**1．教学重点：**

Hadoop原理与运行机制；Hadoop生态系统；Hadoop平台的安装。

**2．教学难点：**

Hadoop平台的安装。

**3．教学方法：**

多媒体理论教学与操作演示。

**（三）教学内容**

1．概述

（1）Hadoop简介

（2）Hadoop发展简史

（3）Hadoop的特性

（4）Hadoop的应用现状

（5）Hadoop的版本

2．Hadoop生态系统

（1）HDFS

（2）Hbase

（3）Map Reduce

（4）Hive

（5）Pig

（6）Zookeeper

（7）Flume

（8）Sqoop

3．Hadoop的安装与使用

（1）创建Hadoop用户

（2）Java安装

（3）SSH登录权限设置

（4）安装单机Hadoop

（5）Hadoop伪分布式安装

**第三章 分布式文件系统HDFS**

**（一）教学目标**

通过本章的学习，要求学生了解分布式文件系统的基本概念、结构和设计需求，掌握Hadoop分布式文件系统HDFS的重要概念、存储原理和读写过程，并熟练掌握分布式文件系统HDFS的使用方法。

**（二）教学要求**

**1．教学重点：**

HDFS的概念、体系结构、存储原理和读写过程。

**2．教学难点：**

分布式文件系统HDFS的使用方法。

**3．教学方法：**

多媒体理论教学。

**（三）教学内容**

1．分布式文件系统

（1）计算机集群结构

（2）分布式文件系统的结构

（3）分布式文件系统的设计需求

2．HDFS的相关概念

（1）块

（2）名称节点和数据节点

（3）第二名称节点

3．HDFS体系结构

（1）概述

（2）HDFS命名空间管理

（3）通信协议

（4）客户端

（5）HDFS体系结构的局限性

4．HDFS的存储原理

（1）数据的冗余存储

（2）数据存取策略

（3）数据错误与恢复

5. HDFS的读写过程

（1）读数据的过程

（2）写数据的过程

6. HDFS编程实践

**第四章 资源管理调度框架YARN**

**（一）教学目标**

通过本章的学习，要求学生了解Hadoop的局限和不足及什么是YARN框架；掌握YARN的体系结构和工作流程；掌握搭建YARN分布式集群的方法，并能够使用YARN集群提交一些简单的任务；理解YARN作为Hadoop生态中的资源管理器的意义。

**（二）教学要求**

**1．教学重点：**

YARN体系结构和工作流程，搭建YARN分布式集群。

**2．教学难点：**

使用YARN集群提交简单任务。

**3．教学方法：**

多媒体理论教学与操作演示。

**（三）教学内容**

1．Hadoop的优化与发展

（1）Hadoop的局限与不足

（2）针对Hadoop的改进与提升

2．HDFS2.0的新特性

（1）HDFS HA

（2）HDFS联邦

3. 新一代资源管理调度框架YARN

（1）YARN设计思路

（2）YARN体系结构

（3）YARN工作流程

（4）YARN框架与MapReduce1.0框架的对比分析

（5）YARN的发展目标

**第五章 NoSQL数据库**

**（一）教学目标**

通过本章的学习，要求学生了解NoSQL数据库与传统的关系数据库的差异、NoSQL数据库的四大类型以及NoSQL数据库的三大基石；基本掌握Redis、MongoDB等NoSQL数据库的使用方法。

**（二）教学要求**

**1．教学重点：**

NoSQL数据库的概念和基本原理。

**2．教学难点：**

Redis、MongoDB的使用方法。

**3．教学方法：**

多媒体理论教学与操作演示。

**（三）教学内容**

1．NoSQL简介

2.NoSQL与关系数据库的比较

3.NoSQL的四大类型

（1）键值数据库

（2）列族数据库

（3）文档数据库

（4）图数据库

4．NoSQL的三大基石

（1）CAP

（2）BASE

（3）最终一致性

5．从NoSQL到NewSQL数据库

6.云数据库

**第六章 MapReduce计算模式**

**（一）教学目标**

通过本章的学习，要求学生理解MapReduce计算模型，MapReduce作业的运行流程，MapReduce作业调度，MapReduce异常处理；掌握MapReduce应用开发流程，通过Web界面分析MapReduce应用，多个MapReduce过程的组合模式；掌握MapReduce编程思想。

**（二）教学要求**

**1．教学重点：**

MapReduce计算模型、MapReduce编程思想。

**2．教学难点：**

编写MapReduce算法。

**3．教学方法：**

多媒体理论教学与操作演示。

**（三）教学内容**

1．概述

（1）分布式并行编程

（2）MapReduce模型简介

（3）Map和Reduce函数

2．MapReduce的工作流程

（1）工作流程概述

（2）MapReduce的各个执行阶段

（3）Shuffle过程详解

3．实例分析：WordCount

（1）WordCount的程序任务

（2）WordCount的设计思路

（3）WordCount的具体执行过程

4．MapReduce的具体应用

（1）MapReduce在关系代数运算中的应用

（2）分组与聚合运算

5．MapReduce编程实践

**第七章 基于Hadoop的数据仓库Hive**

**（一）教学目标**

通过本章的学习，要求学生了解数据仓库的概念；掌握Hive的系统架构及组件，Hive数据结构和数据存储格式；掌握Hive支持的数据类型和使用HiveQL访问数据；熟练使用Hive进行简单编程。

**（二）教学要求**

**1．教学重点：**

Hive的系统架构及组件、Hive数据结构和数据存储格式、Hive支持的数据类型以及如何使用HiveQL访问数据。

**2．教学难点：**

Hive的应用。

**3．教学方法：**

多媒体理论教学与操作演示。

**（三）教学内容**

1．大数据的分析处理概述

2．Hive

（1）系统架构及组件

（2）Hive数据结构

（3）数据存储格式

（4）Hive支持的数据类型

（5）使用HiveQL访问数据

3．Hive的具体应用及Hive HA原理

4. Hive编程相对于MapReduce编程的优势

**第八章 分布式数据库HBase**

**（一）教学目标**

通过本章的学习，要求学生了解大数据环境下的数据库HBase的访问接口和数据模型；理解HBase架构与原理，物理部署与读写流程；熟练掌握HBase的使用方法。

**（二）教学要求**

**1．教学重点：**

HBase的数据模型与物理存储，物理部署与读写流程，Shell管理HBase中的数据。

**2．教学难点：**

从RDBMS到HBase数据库概念的转化，包括：行到列与主键到行关键字、联合查询（Join）。

**3．教学方法：**

多媒体理论教学与操作演示。

**（三）教学内容**

1．概述

（1）Hbase简介

（2）HBase与传统关系数据库的对比分析

2．HBase访问接口

3. Hbase数据模型

（1）数据模型的相关概念

（2）数据坐标

（3）概念视图

（4）物理视图

（5）面向列的存储

3．HBase的实现原理

（1）HBase的功能组件

（2）表和Region

（3）Region的定位

4．HBase运行机制

（1）HBase系统架构

（2）Region服务器的工作原理

5．HBase编程实践

（1）HBase常用的Shell命令

（2）HBase常用的JavaAPI及应用实例

**第九章 Spark**

**（一）教学目标**

通过本章的学习，要求学生了解Spark与Scala编程语言，了解 Spark与Hadoop的区别，认识Hadoop MapReduce计算模型的缺陷与Spark的优势；掌握Spark的生态系统和架构设计，并掌握Spark SQL以及Spark的部署和应用方式；掌握Spark的安装与基本的编程实践。

**（二）教学要求**

**1．教学重点：**

Spark体系架构和Spark计算模型。

**2．教学难点：**

部署Spark集群。

**3．教学方法：**

多媒体理论教学与操作演示。

**（三）教学内容**

1．概述

（1）Spark简介

（2）Scala简介

（3）Spark与Hadoop的对比

2．Spark生态系统

3. Spark运行架构

（1）基本概念

（2）架构设计

（3）Spark运行基本流程

（4）RDD的设计与运行原理

4．Spark的部署与应用方式

（1）Spark三种部署方式

（2）从“Hadoop+Storm”架构转向Spark架构

（3）Hadoop和Spark的统一部署

5. Spark编程实践

**第十章 ZooKeeper**

**（一）教学目标**

通过本章的学习，要求学生了解什么是ZooKeeper；掌握ZooKeeper的相关概念及工作流程；掌握Java代码连接ZooKeeper集群及通过代码读写ZooKeeper集群的目录下数据的方法。

**（二）教学要求**

**1．教学重点：**

使用ZooKeeper对Hadoop的集群进行配置管理和协调，ZooKeeper架构、ZooKeeper的数据模型。

**2．教学难点：**

ZooKeeper的配置。

**3．教学方法：**

多媒体理论教学与操作演示。

**（三）教学内容**

1．云计算平台的管理体系

2．ZooKeeper——集群中的配置管理与协调者

（1）集群环境下的配置管理

（2）ZooKeeper架构

（3）ZooKeeper的数据模型

3．ZooKeeper的相关概念

（1）节点

（2）角色

4. ZooKeeper的读写机制

5. ZooKeeper的安装与配置

**第十一章 数据采集**

**（一）教学目标**

通过本章的学习，要求学生了解数据采集架构，数据来源，掌握日志采集工具Flume和Kafka、数据迁移工具Sqoop的使用方法 。

**（二）教学要求**

**1．教学重点：**

Flume的概念和结构。

**2．教学难点：**

数据采集工具的使用。

**3．教学方法：**

多媒体理论教学与操作演示。

**（三）教学内容**

1．大数据采集架构

（1）概述

（2）常用大数据采集工具

（3）Apache Kafka数据采集

2．数据预处理

（1）数据清洗

（2）数据集成

（3）数据变换

3．数据仓库与ETL工具

（1）数据仓库与ETL工具

（2）常用ETL工具

**第十二章 大数据在互联网领域的应用**

**（一）教学目标**

通过本章的学习，要求学生了解推荐系统的概念，了解长尾理论与推荐系统的内在联系；掌握不同的推荐方法以及推荐系统的应用；然后，掌握协同过滤算法的基本思想。

**（二）教学要求**

**1．教学重点：**

推荐系统模型及推荐系统的应用。

**2．教学难点：**

协同过滤算法。

**3．教学方法：**

多媒体理论教学。

**（三）教学内容**

1．推荐系统概述

（1）什么是推荐系统

（2）长尾理论

（3）推荐系统模型

（4）推荐系统的应用

2．协同过滤

（1）基于用户的协同过滤

（2）基于物品的协同过滤

（3）UserCF算法和ItemCF算法的对比

3．协同过滤实践

**三、主要教学环节和学时安排**

本课程计划 48学时，其中讲授 32 学时，实验（或实训）16 学时。课程主要内容和学时安排见课程教学环节和学时安排表。

**课程教学环节和学时安排表**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 教学环节  课程内容 | 讲授 | 实验 | 实训 | 其它 | 小计 |
| 第一章：大数据概述 | 2 |  |  |  | 2 |
| 第二章：大数据处理架构Hadoop | 2 | 2 |  |  | 4 |
| 第三章：分布式文件系统HDFS | 4 | 2 |  |  | 6 |
| 第四章：资源管理调度框架YARN | 2 | 2 |  |  | 4 |
| 第五章：NoSQL数据库 | 4 |  |  |  | 4 |
| 第六章：MapReduce计算模式 | 4 | 2 |  |  | 6 |
| 第七章：基于Hadoop的数据仓库Hive | 2 | 2 |  |  | 4 |
| 第八章：分布式数据库HBase | 2 | 2 |  |  | 4 |
| 第九章：Spark | 4 | 2 |  |  | 6 |
| 第十章：ZooKeeper | 2 | 2 |  |  | 4 |
| 第十一章：数据采集 | 2 |  |  |  | 2 |
| 第十二章：大数据在互联网领域的应用 | 2 |  |  |  | 2 |
| 总计 | 32 | 16 |  |  | 48 |

**四、实验教学的基本要求**

1.课内实验项目一览表

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | 实验项目名称 | 实验类型（演示性、验证性、综合性、设计研究、其他） | 学时 |
| 1 | HDFS实验：部署HDFS | 验证性 | 2 |
| 2 | HDFS实验：读写HDFS文件 | 验证性 | 2 |
| 3 | 部署YARN集群 | 验证性 | 2 |
| 4 | MapReduce实验：单词计数 | 验证性 | 2 |
| 5 | Hive实验：新建Hive表 | 验证性 | 2 |
| 6 | 部署HBase | 验证性 | 2 |
| 7 | 部署Spark集群 | 验证性 | 2 |
| 8 | 部署ZooKeeper | 验证性 | 2 |
| 总计 | | | 16 |

2.实验教学要求

教学目的：通过本课程实验加深学生对课堂教学内容的理解，使学生熟练掌握系统地掌握大数据处理概论和基于Hadoop的大数据处理框架，具备利用大数据组件工具进行数据分析的能力，提高动手实践能力和综合应用能力。

教学要求：具备Hadoop集群的管理与维护能力；基于MapReduce的数据挖掘实践能力；具有使用HDFS存储大数据和HBase大数据库的能力，利用Hadoop环境进行大数据的分析处理和数据整合能力。

**五、课程考核**

**（一）考核类型：**考查。

**（二）考核方式与成绩评定**

**1．考核方式：**本课程采用上机实际操作方式进行考核。

**2．课程考核的成绩评定：**

以百分制计分，最终成绩的构成为平时成绩占40%、期末考试成绩占50%、实验（上机）操作考核成绩占10%。

**（三）命题要求：**

考试题型可有选择题、判断题、简答题、论述题和计算题等。命题依据教学大纲要求，重点考核学生对基本概念、基本理论、重点知识的学习和掌握情况，考查学生对上述内容的综合应用能力。考试内容包括各章基本概念、基本原理和重点知识，试题内容覆盖教学大纲各章，A、B两套试题重复率不超过20%，与往年试题的重复率不超过10%。

**六、建议使用教材及主要教学参考资料**

**（一）教材：**

《Hadoop大数据原理及应用》，徐鲁辉，西安电子科技大学出版社，2020年3月。

**（二）主要教学参考资料：**

1．《Hadoop基础教程》，（英）特金顿 著，张治起 译编，人民邮电出版社，2014年。

2．《Hadoop权威指南(第3版)》，（美）Tom White著，华东师范大学数据科学与工程学院 译，清华大学出版社，2015年。

3．《Hadoop大数据原理与应用实验教程》，徐鲁辉主编，西安电子科技大学出版社，2020年3月。

4．《实战Hadoop 2．0（第二版）——从云计算到大数据》，叶晓江主编，电子工业出版社，2016年。

5.《大数据原理及应用》，林子雨，人民邮电出版社，2017年1月，第2版。

**七、大纲说明**

本课程教学大纲依据2019版人才培养方案制定。本课程具有理论和实践紧密结合的特点，既要求学生掌握大数据的基础知识，又要求学生能够运用所学解决实际问题，在人才培养中具有比较重要的地位。先修课程为Linux操作系统、面向对象程序设计等。

**执笔人：刘敏娟 审核人：白鑫**