《基于Hadoop的大数据概论》教学大纲

1. 课程基本信息

|  |  |  |  |
| --- | --- | --- | --- |
| 课程代码 |  | 课程名称（中/英） | 基于Hadoop的大数据概论 |
| 开课学院 | 智能制造与信息工程学院 | 适用专业 | 大数据技术 |
| 学 分 | 3.5分  | 总学时 | 56学时  | 理论学时 | 44 |
| 实践学时 | 12 |
| 课程性质 | 必修 | 课程类别 |  | 考核方式 | 考查 |
| 先修课程 | 大学信息技术、数据库技术及应用（MySQL)、Linux操作系统基础 |
| 选用教材 | 1.《Hadoop大数据开发基础》张军 张良均 主编 人民邮电出版社 2021.11,ISBN 978-7-115-57562-3 |
| 网络课程 | 超星创建课程 |

1. 课程简介

本课程是大数据技术专业的必修课程。随着时代的发展，大数据已经成为一个耳熟能详的词汇。与此同时，针对大数据处理的新技术也在不断的开发和运用中，逐渐成为数据处理挖掘行业广泛使用的主流技术之一。Hadoop分布式集群系统架构，具有高可用性、高容错性和高扩展性等优点，由于它提供了一个开放式的平台，用户可以在不了解底层实现细节的情形下，开发适合自身应用的分布式程序。经过多年的发展，目前Hadoop已经成长为一个全栈式的大数据技术生态圈，包括了Hive、HBase、Spark等一系列组件，成为应用最广泛、最具有代表性的大数据技术之一。因此，学习Hadoop技术是从事大数据行业工作所必不可少的一步。为了满足企业的大数据人才需求，帮助学者掌握相关技术知识解决实际的业务需求，特开设Hadoop大数据开发基础课程。通过本课程的学习，使学生掌握Hadoop、Hive和HBase集群的安装配置，能够根据具体需求编写MapReduce程序解决实际业务问题，使用Hive、HBase进行数据存储、查询、分析，最后详细拆解并学习电影网站用户影评分析案例，将理论与实践相结合，为将来从事数据分析挖掘研究、工作奠定基础。

1. 课程目标

课程目标1：课程思政目标——用审辩思维（critical thinking）理性看待事物，明辨是非；引导学生形成积极向上的世界观、价值观和人生观。

课程目标2：依据阶段练习与维护典型工作任务的能力要求，确定各教学单元的能力、知识和素质目标，并以此进一步进行教学单元设计，使学生掌握Hadoop基本原理与架构、集群安装及配置，MapReduce编程等，能够理论结合实践，运用相关技术解决实际的业务需求。全书按照解决实际任务的工作流程路线，逐步展开介绍相关的理论知识点，推导生成可行的解决方案，最后落实任务实现环节。培养学生分析问题和解决问题的能力。

课程目标3：掌握Hadoop的相关知识，包括Hadoop集群的搭建及配置、基础操作、MapReduce入门编程、MapReduce进阶编程、Hive数据仓库和HBase分布式数据库等等，能够独立完成电影网站用户影评分析。

1. 课程内容与教学要求

**第一章 Hadoop介绍**

**（**一）课程内容

1. 通过直接引入的方式导入新课、补充线上学习的相关习题和背景知识。在大数据时代，Hadoop作为处理大数据的分布式存储和计算框架，在国内外大、中、小型企业中已得到广泛应用，掌握Hadoop技术是从事大数据行业工作必不可少的一步。本章主要介绍Hadoop框架的理论知识，包括Hadoop框架及其发展历史、特点；然后重点讲解Hadoop的三大核心组件HDFS、MapReduce和YARN；接着对Hadoop生态系统中常用的组件进行简单的介绍，包括组件的特点和应用；最后简要介绍Hadoop的应用场景。

2.通过讲解知识点，然后提出问题，让学生分析问题并独立解决问题，然后将自己理解的答案分享给同学们。

3.示例操作部分，通过教师先演示，学生接着完成的方式，学生遇到问题及时解决，可以找老师帮忙，也可以同学之间相互帮忙。

4.课程思政：锻炼学生工匠精神，一个标点符号都会影响整个代码的输出。还要锻炼团队合作精神，同学间相互帮助，完成代码的效果会更高。

**（二）教学要求**

1.了解Hadoop框架及其发展历史、特点；

2.了解Hadoop核心组件；

3.了解Hadoop生态系统；

4.了解Hadoop应用场景；

5.掌握Hadoop的三大核心组件HDFS、MapReduce和YARN；

重点：

1.掌握Hadoop生态系统；

2.掌握Hadoop的三大核心组件HDFS、MapReduce和YARN。

难点：

1.掌握Hadoop的三大核心组件HDFS、MapReduce和YARN。

**第二章 数据Vue.js开发基础**

**（一）课程内容**

1.复习巩固。教师通过上节课作业的完成情况，对学生吸收不好的知识点进行再次巩固讲解。

2.教学导入。

为了更好的学习Hadoop，我们要先搭建好Hadoop集群环境。本章将详细讲解Hadoop集群的搭建及配置。

3.通过讲解知识点，然后提出问题，让学生分析问题并独立解决问题，然后将自己理解的答案分享给同学们。

4.示例操作部分，通过教师先演示，学生接着完成的方式，学生遇到问题及时解决，可以找老师帮忙，也可以同学之间相互帮忙。

5.前面的示例是按照源代码练习，最后的代码只给部分代码学生，自己在运行过程中补齐补全代码。

6.课程思政：锻炼学生工匠精神，一个标点符号都会影响整个代码的输出。还要锻炼团队合作精神，同学间相互帮助，完成代码的效果会更高。

**（二）教学要求**

1.掌握虚拟机的安装及配置方法

2.掌握在Linux下JDK的安装方法

3.掌握Hadoop完全分布式集群环境的搭建过程

4.掌握Hadoop集群的监控方法

重点：Hadoop集群的搭建及配置成功。

难点：Hadoop集群的搭建及配置成功。

**第三章 Hadoop基础操作**

**（一）课程内容**

1.复习巩固。教师通过上节课作业的完成情况，对学生吸收不好的知识点进行再次巩固讲解。

2.教学导入。

在学习完第2章的基础知识后，Hadoop集群的环境搭配好了，Hadoop框架的核心设计为HDFS和MapReduce，HDFS负责数据的存储，MapReduce则负责数据的计算。对海量数据进行计算前，数据的存储是一步必要的操作，因此，需要先掌握HDFS的基本操作方法。

3.通过讲解知识点，然后提出问题，让学生分析问题并独立解决问题，然后将自己理解的答案分享给同学们。

4.示例操作部分，通过教师先演示，学生接着完成的方式，学生遇到问题及时解决，可以找老师帮忙，也可以同学之间相互帮忙。

5.前面的示例是按照源代码练习，最后的代码只给部分代码学生，自己在运行过程中补齐补全代码。

6.课程思政：锻炼学生工匠精神，一个标点符号都会影响整个代码的输出。还要锻炼团队合作精神，同学间相互帮助，完成代码的效果会更高。

**（二）教学要求**

1.了解Hadoop安全模式

2.掌握查看、解除与开启Hadoop安全模式的操作

3.掌握查看Hadoop集群存储系统和计算资源信息的方法

4.了解HDFS分布式文件系统

5.掌握HDFS的基本操作

6.掌握提交MapReduce任务的基本操作

7.掌握多个MapReduce任务的管理方法

重点：掌握如何查看Hadoop集群存储系统和计算资源信息的方法；掌握提交MapReduce任务的基本操作

难点：掌握多个MapReduce任务的管理方法

**第四章 MapReduce入门编程**

**（一）课程内容**

1.复习巩固。教师通过上节课作业的完成情况，对学生吸收不好的知识点进行再次巩固讲解。

2.教学导入。

随着互联网的发展，加入互联网的用户越来越多，互联网的用户规模已不容小觑。互联网市场潜力巨大，各大网站的运营商都在积极采取措施，分析用户的特征，根据不同的用户群体向其提供差异化的服务，进而达到精准营销的目的。接下来，本章编写基础的MapReduce程序处理简单任务的方法。

3.通过讲解知识点，然后提出问题，让学生分析问题并独立解决问题，然后将自己理解的答案分享给同学们。

4.示例操作部分，通过教师先演示，学生接着完成的方式，学生遇到问题及时解决，可以找老师帮忙，也可以同学之间相互帮忙。

5.前面的示例是按照源代码练习，最后的代码只给部分代码学生，自己在运行过程中补齐补全代码。

6.课程思政：锻炼学生工匠精神，一个标点符号都会影响整个代码的输出。还要锻炼团队合作精神，同学间相互帮助，完成代码的效果会更高。

**（二）教学要求**

1.掌握在Windows下安装Javat和IntelliJ IDEA的方法

2.掌握IntelliJ IDEA中创建MapReduce工程和配置MapReduce环境的方法

3.熟练掌握MapReduce的工作原理及执行流程

4.了解Hadoop官方示例中的WordCount源码

5.了解MapReduce编程的基本思路

6.掌握map()方法与reduce()方法的处理逻辑

7.掌握编写基础的MapReduce程序处理简单任务的方法

重点：掌握编写基础的MapReduce程序处理简单任务的方法

难点：掌握MapReduce编程的基本思路

**第五章 MapReduce进阶编程**

**（一）课程内容**

1.复习巩固。教师通过上节课作业的完成情况，对学生吸收不好的知识点进行再次巩固讲解。

2.教学导入。

通过对MapReduce入门编程的学习，我们实现了对某竞赛网站每日访问次数的统计任务。本章将继续学习MapReduce编程，通过一些高级的编程技巧，使得整个编程过程更加高效。

3.通过讲解知识点，然后提出问题，让学生分析问题并独立解决问题，然后将自己理解的答案分享给同学们。

4.示例操作部分，通过教师先演示，学生接着完成的方式，学生遇到问题及时解决，可以找老师帮忙，也可以同学之间相互帮忙。

5.前面的示例是按照源代码练习，最后的代码只给部分代码学生，自己在运行过程中补齐补全代码。

6.课程思政：锻炼学生工匠精神，一个标点符号都会影响整个代码的输出。还要锻炼团队合作精神，同学间相互帮助，完成代码的效果会更高。

**（二）教学要求**

1.掌握MapReduce输入和输出格式的设置方法

2.掌握Hadoop Java API的使用方法

3.掌握自定义键值类型的方法

4.了解Combiner的工作原理

5.掌握Combiner、Partitioner和自定义计数器的使用方法

6.掌握MapReduce参数传递流程

7.掌握使用IntelliJ IDEA自动打包并提交MapReduce任务的方法

重点：掌握MapReduce输入和输出格式的设置方法；掌握Hadoop Java API的使用方法；掌握自定义键值类型的方法；掌握Combiner、Partitioner和自定义计数器的使用方法。

难点：掌握MapReduce参数传递流程；掌握使用IntelliJ IDEA自动打包并提交MapReduce任务的方法。

**第六章 Hive数据仓库**

**（一）课程内容**

1.复习巩固。教师通过上节课作业的完成情况，对学生吸收不好的知识点进行再次巩固讲解。

2.教学导入。

随着信息技术的普及和企业信息化建设步伐的加快，企业逐步认识到建立企业范围内的统一数据存储的重要性，越来越多的企业已经建立或正着手建立企业数据仓库。为了保证较高的处理效率与灵活性，选用Hive数据仓库对基站数据进行存储与处理分析。本章将详细讲解如何通过Hive编程解决实际问题，首先介绍Hive数据仓库的系统架构、数据模型和执行流程；其次讲解访问Hive的3种试及其配置过程；接着介绍Hive数据仓库中表的创建与修改的基本语法，并结合官方的示例介绍Hive表数据的增删查改；最后通过编写Hive语句实现基站掉话率的统计分析。

3.通过讲解知识点，然后提出问题，让学生分析问题并独立解决问题，然后将自己理解的答案分享给同学们。

4.示例操作部分，通过教师先演示，学生接着完成的方式，学生遇到问题及时解决，可以找老师帮忙，也可以同学之间相互帮忙。

5.前面的示例是按照源代码练习，最后的代码只给部分代码学生，自己在运行过程中补齐补全代码。

6.课程思政：锻炼学生工匠精神，一个标点符号都会影响整个代码的输出。还要锻炼团队合作精神，同学间相互帮助，完成代码的效果会更高。

**（二）教学要求**

1.了解Hive的概念及Hive与传统数据库的对比

2.了解Hive系统架构、数据模型和执行流程

3.熟悉3种访问Hive的方式及配置过程

4.掌握Hive中数据库与表的创建、修改操作方法

5.掌握Hive表数据增删查改的操作方法

重点：Hive系统架构、数据模型和执行流程；掌握Hive中数据库与表的创建、修改操作方法；掌握Hive表数据增删查改的操作方法。

难点：熟悉3种访问Hive的方式及配置过程。

**第七章 HBase分布式数据库**

**（一）课程内容**

1.复习巩固。教师通过上节课作业的完成情况，对学生吸收不好的知识点进行再次巩固讲解。

2.教学导入。

对于推动企业发展新动能而言，大数据的价值表现在掌握业务战略决策，提高服务水平，提高商品购买率，研发创新产品等，不断塑造发展新动能新优势。为了保证较高的处理效率与灵活性，一般选择使用HBase分布式数据库存储通话记录，并采用HBase Java API实现通话记录数据的查询分析。本章首先详细讲解HBase分布式数据库的系统架构、数据模型、读/写流程；其次介绍HBaser的安装及配置过程和HBase常用的Shell命令；接着重点介绍采用HBase Java API实现表创建、表数据导入；最后通过编写HBase Java API实现通话记录数据表的创建与查询分析。

3.通过讲解知识点，然后提出问题，让学生分析问题并独立解决问题，然后将自己理解的答案分享给同学们。

4.示例操作部分，通过教师先演示，学生接着完成的方式，学生遇到问题及时解决，可以找老师帮忙，也可以同学之间相互帮忙。

5.前面的示例是按照源代码练习，最后的代码只给部分代码学生，自己在运行过程中补齐补全代码。

6.课程思政：锻炼学生工匠精神，一个标点符号都会影响整个代码的输出。还要锻炼团队合作精神，同学间相互帮助，完成代码的效果会更高。

**（二）教学要求**

1.了解分布式数据库HBase

2.熟悉HBase的系统架构、数据模型和读/写流程

3.了解ZooKeeper的概念、集群角色及选举机选

4.了解ZooKeeper集群的安装部署方法

5.掌握HBase的安装部署方法

6.掌握HBase常用的Shell命令

7.掌握HBase Java API的使用方法

重点：HBase常用的Shell命令；掌握HBase Java API的使用方法。

难点：掌握HBase Java API的使用方法。

**第八章 项目案例——电影网站用户影评分析**

**（一）课程内容**

1.复习巩固。教师通过上节课作业的完成情况，对学生吸收不好的知识点进行再次巩固讲解。

2.教学导入。

常规的数据分析工具在大数据场景下，处理数据的效率低，显然不选用于大数据处理分析。分析式计算框架的出现，为分析处理大数据提供了很好的解决方案。本章将使用Hadoop框架并结合电影评分数据，编写MapReduce程序实现对用户影评的分析，从多维度分析用户的观影兴趣偏好。

3.通过讲解知识点，然后提出问题，让学生分析问题并独立解决问题，然后将自己理解的答案分享给同学们。

4.示例操作部分，通过教师先演示，学生接着完成的方式，学生遇到问题及时解决，可以找老师帮忙，也可以同学之间相互帮忙。

5.前面的示例是按照源代码练习，最后的代码只给部分代码学生，自己在运行过程中补齐补全代码。

6.课程思政：锻炼学生工匠精神，一个标点符号都会影响整个代码的输出。还要锻炼团队合作精神，同学间相互帮助，完成代码的效果会更高。

**（二）教学要求**

1.掌握根据业务场景设计map()方法和reduce()方法的计算逻辑

2.掌握编写MapReduce程序解决常见的数据处理问题

3.掌握编写MapReduce程序实现电影网站用户影评分析的方法

重点：掌握根据业务场景设计map()方法和reduce()方法的计算逻辑；掌握编写MapReduce程序解决常见的数据处理问题；掌握编写MapReduce程序实现电影网站用户影评分析的方法。

难点：掌握根据业务场景设计map()方法和reduce()方法的计算逻辑；掌握编写MapReduce程序解决常见的数据处理问题；掌握编写MapReduce程序实现电影网站用户影评分析的方法。

五、学时分配、教学方法及支撑课程目标

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **序号** | **教学单元名称** | **学时** | **主要教学方法** | **支撑的课程目标** |
| 1 | 第一章 Hadoop介绍 | 4 | 讲授法、演示法、讨论法、练习课 | 课程目标1，2，3 |
| 2 | 第二章Hadoop集群的搭建及配置 | 12 | 讲授法、演示法、讨论法、练习课 | 课程目标1，2，3 |
| 3 | 第三章 Hadoop基础操作 | 12 | 讲授法、演示法、讨论法、练习课 | 课程目标1，2，3 |
| 4 | 第四章 MapReduce入门编程 | 12 | 讲授法、演示法、讨论法、练习课 | 课程目标1，2，3 |
| 5 | 第五章 MapReduce进阶编程 | 12 | 讲授法、演示法、讨论法、练习课 | 课程目标1，2，3 |
| 6 | 期末考试复习 | 2 |  |  |
| 7 | 期末考试 | 2 |  |  |
| 合计 | 56 |  |  |  |
|  |  |  |  |

六、课程考核

**（一）考核要求**

1. 本课程为考查科目，成绩评定采用百分制。试卷命题依据教学大纲要求，侧重教材上所学的知识点，并适当采用部分课外资源。命题符合教学大纲中规定的教学内容和教学要求。

2. 重点考核范围：课内部分考核内容包含教材里的教学单元内容，重点考核所学的知识点和课外习题。

3. 考核目标：本课程注重学生平时的学习投入、综合技能训练和积累，涉及学生的理论基础知识能力、操作技能和综合知识应用能力。

4. 成绩评定方法：总评成绩由平时表现（占40%）和考试成绩（占60%）两部分构成。平时成绩包括课堂出勤（占20%），课堂表现及随堂练习（占20%）；考试成绩包括期末考试成绩（占60%）。

**（二）成绩评定**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **序号** | **考核形式** | **考核方法** | **考核权重** | **考核对应的课程目标** | **备注** |
| 1 | 出勤 | 出勤次数 | 20% | 课程目标1，3 |  |
| 2 | 课堂表现及随堂练习 | 课堂作业及平时表现 | 20% | 课程目标1，2，3 |  |
| 3 | 期末考试 | 闭卷考试 | 60% | 课程目标2，3 |  |
| 总评成绩 | 各项考核按权重相加 | 100% | 课程目标1，2，3 |  |
| **说明**：学生不提交或被认定为抄袭者，以0分计算。 |

七、参考书目及学习资料

1. 《Hadoop大数据技术原理与应用》黑马程序员 主编 清华大学出版社

八、大纲说明

本课程的授课模式为：课堂授课+演示+上机，其中，课堂主要采用多媒体和超星进行授课，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。