**江西开放大学**

**《大数据技术综合实训》课程实施方案**

**一、课程基本信息**

|  |  |
| --- | --- |
| **课程名称** | 《大数据技术综合实训》 |
| **课程代码** | 05192 |
| **课程性质** | 专业必修课 / 集中实践环节 |
| **学分** | 4学分 |
| **学时** | 72学时  （其中：理论学习20学时，实践操作52学时） |
| **先修课程** | 《Python程序设计》、《Linux操作系统》、  《数据库原理与应用》、《大数据技术概论》 |
| **开设学期** | 第4或第5学期 |
| **适用专业** | 大数据技术 |

**二、课程定位与目标**

**1、课程定位**

本课程是大数据技术专业及相关专业人才培养方案中的综合性、实践性核心环节。旨在将前序课程的分散知识点进行有机整合，通过完整的、贴近企业真实场景的项目实践，培养学生解决复杂大数据问题的综合能力和职业素养。

**2、课程目标**

**（1）知识目标**

* 整合并深化理解大数据生命周期（数据采集→存储→处理→分析→可视化）的全流程知识。
* 掌握一种主流的大数据技术栈（如Hadoop+Spark）的核心组件及其协同工作原理。

**（2）能力目标**

* 技术实践能力：能够独立或在团队中完成从数据获取到最终可视化展示的完整大数据项目。
* 问题解决能力：能够针对特定业务需求，设计合理的技术方案，并解决实施过程中的技术难题。
* 工具使用能力：熟练使用至少一种主流大数据组件及相关开发工具。
* 文档撰写能力：能够规范地撰写项目方案、设计文档、实验报告和总结报告。

**（3）素质目标**

* 培养严谨的科学态度、规范的操作习惯和创新精神。
* 提升团队协作、沟通表达和项目管理能力。
* 建立数据安全与隐私保护的意识。

**三、课程的学习内容和基本要求**

**1、教学内容**

按照教学大纲和教材的体系结构，本课程的教学内容主要包括以下十个章节：

第一章 大数据技术综合实训概况及实战模板

第二章 Linux基础与集群搭建

第三章 Hadoop集群配置

第四章 HDFS

第五章 MapReduce分布式编程

第六章 Hive大数据仓库

第七章 HBase数据库部署与操作

第八章 数据获取与Flume应用

第九章 基于Spark的内存计算

第十章 利用大数据平台处理图像

**2、教学目的和要求**

**教学目的：**

通过本课程的学习和实践，使学生系统地掌握大数据技术平台构建与数据处理的全流程技能，能够独立完成从环境搭建、数据采集到分析可视化的完整项目，培养符合行业需求的复合型技术人才。

**教学要求：**

**第1章 大数据技术综合实训概况及实战模板**

【教学内容】

1.1 大数据的概念与基本特性

1.2 大数据处理流程

1.3 Hadoop大数据技术

1.4 实践环境准备

【教学要求】

了解：大数据的概念与基本特性；

掌握：大数据处理流程及实践环境的搭建。

**第2章 Linux基础与集群搭建**

【教学内容】

2.1 Linux常用命令

2.2 网络配置

2.3 Linux集群配置

【教学要求】

掌握：Linux 常用命令、网络配置及集群配置。

**第3章 Hadoop集群配置**

【教学内容】

3.1 Hadoop集群安装

3.2 Hadoop集群初始化和日志查看

【教学要求】

掌握：Hadoop 集群的安装及其安装流程，熟悉 Hadoop 集群的初始化和日志查看。

**第4章 HDFS**

【教学内容】

4.1 HDFS简介

4.2 HDFS基本命令

4.3 HDFS数据平衡优化

4.4 HDFS API的使用方法

【教学要求】

了解：HDFS 的基本知识；

掌握：HDFS 的基本命令、数据平衡优化及 API 的使用。

**第5章 MapReduce分布式编程**

【教学内容】

5.1 MapReduce简介

5.2 词频统计编程实例

5.3 MapReduce Shuffle 过程开发

5.4 MapReduce 的性能优化

5.5 YARN 数据处理框架

【教学要求】

了解：MapReduce 的基本知识；

掌握：MapReduce Shuffle 过程开发及性能优化，熟悉 YARN 数据处理框架。

**第6章 Hive大数据仓库**

【教学内容】

6.1 Hive简介

6.2 Hive安装及配置

6.3 从创建数据库到创建表

6.4 数据查询及自定义函数运算

6.5 Hive 自定义函数编程

6.6 Hive 实战

【教学要求】

了解：Hive 大数据仓库的基本概念；

掌握：Hive 大数据仓库的安装配置及使用，熟悉 Hive 自定义操作。

**第7章 HBase数据库部署与操作**

【教学内容】

7.1 HBase简介

7.2 HBase的安装

7.3 HBase Shell操作

7.4 HBase客户端API

7.5 HBase客户端选择及配置优化

7.6 HBase与MapReduce集成

7.7 HBase集群监控

【教学要求】

了解：HBase 数据库的基本概念；

掌握：HBase 数据仓库的安装配置及使用，熟悉 HBase 客户端 API 的基本使用及集群监控。

**第8章 数据获取与Flume应用**

【教学内容】

8.1 公开数据资源获取

8.2 使用网络爬虫获取数据

8.3 使用Flume获取数据

【教学要求】

了解：获取数据的三种方法；

熟悉：网络爬虫的工作原理；

掌握：Flume 获取数据。

**第9章 基于Spark的内存计算**

【教学内容】

9.1 Spark简介

9.2 Spark快速部署

9.3 Spark程序

9.4 RDD 编程

9.5 Spark 生态系统

【教学要求】

了解：Spark 的基本概念；

掌握：Spark 的部署及应用，掌握 RDD 编程，熟悉 Spark 的生态系统。

**第10章 利用大数据平台处理图像**

【教学内容】

10.1 图像的基本概念

10.2 Hadoop处理图像的问题及对策

10.3 HIPI 安装与部署

10.4 使用 HIPI 进行图像处理

10.5 HIPI 工具 hibDownload

【教学要求】

了解：图像的基本概念；

熟悉：Hadoop 处理图像的问题及对策；

掌握：HIPI 的安装及运用 HIPI 处理图像。

**四、学时进度安排**

|  |  |  |  |
| --- | --- | --- | --- |
| **章节** | **教学内容** | **理论学时** | **实验学时** |
| 1 | 大数据技术综合实训概况及实战模板 | 2 | 2 |
| 2 | Linux基础与集群搭建 | 2 | 6 |
| 3 | Hadoop集群配置 | 2 | 6 |
| 4 | HDFS | 2 | 6 |
| 5 | MapReduce分布式编程 | 2 | 6 |
| 6 | Hive大数据仓库 | 2 | 4 |
| 7 | HBase数据库部署与操作 | 2 | 6 |
| 8 | 数据获取与Flume应用 | 2 | 6 |
| 9 | 基于Spark的内存计算 | 2 | 4 |
| 10 | 利用大数据平台处理图像 | 2 | 6 |
| **合计** | | **20** | **52** |

**五、考核要求**

本课程考主要采用形成性考核方式，重点考察学生的实际操作能力与项目完成质量。考核总成绩由实训项目成绩和学习过程表现两部分组成。

**1、考核方式与比例**

|  |  |  |
| --- | --- | --- |
| 考核维度 | 总权重 | 评价标准 |
| 实训项目成绩 | 80% | 基于实训项目的完成质量（如实训报告完整性、操作准确性、问题解决能力） |
| 学习过程表现 | 20% | 包括出勤率、作业完成情况、课堂互动、网上学习参与度（如讨论区发言、视频观看时长） |
| 合计 | 100% |  |

**2、形成性考核说明**

形成性考核成绩由指导教师根据学生在实训项目中的表现和学习过程表现综合评定。考核重点在于学生的实际操作能力、问题解决能力和项目完成质量。

**3、实训项目说明**

由 4 个选做实训项目构成，学生根据自身兴趣和条件，选择其中之一作为自己的实训项目。并最将最终成果以“实训报告”形式提交（见附录）。项目说明如下：

|  |  |  |
| --- | --- | --- |
| **项目编号** | **实训项目名称** | **对应章节**  **总权重** |
| 项目一 | 天猫电商平台用户行为分析 | 基于阿里云天池的“天猫用户行为数据集”，分析用户点击和购买行为，生成简单的用户行为报告。 |
| 项目二 | 豆瓣电影评分数据分析 | 基于教师提供的模拟豆瓣电影评分数据集（CSV，约10MB），分析电影评分分布和热门电影排名。 |
| 项目三 | 微博评论活跃度统计 | 基于教师提供的模拟微博评论数据集（CSV，约50MB，包含评论ID、用户ID、评论内容、时间戳），统计每日评论活跃度。 |
| 项目四 | 学生成绩分析 | 基于教师提供的模拟学生成绩数据集（CSV，约20MB，包含学生ID、课程ID、成绩、考试时间），分析课程的成绩分布。 |

**实训报告要求：**包含“封面”、“目的”、“内容”、“过程”、“总结”等五部分（参见附录），以 Word 文档上传。

**4、实训项目评价标准**

|  |  |  |
| --- | --- | --- |
| 评价指标 | 权重 | 评价标准 |
| 实训报告完整性 | 50% | 含“封面”、“目的”、“内容”、“过程”、“总结”，结构完整，逻辑清晰。 |
| 操作准确性 | 20% | 操作步骤完整无误，技术参数配置正确，程序代码无语法/逻辑错误，最终运行结果符合项目预期要求。 |
| 问题解决能力 | 20% | 能独立解决操作中的常见问题。 |
| 实训体会深度 | 10% | 能总结操作经验、反思不足。 |

**5、学习过程表现评价标准**

|  |  |  |
| --- | --- | --- |
| 考核维度 | 权重 | 评价标准 |
| 出勤率 | 50% | 面授或网上学习时长达标  （如出勤率≥80%） |
| 作业完成情况 | 40% | 按时提交作业，答案正确、逻辑清晰 |
| 课堂互动 | 10% | 积极参与讨论区发言、回答问题  （如至少 1 次有效发言） |

**六、实训内容与项目设置**

**综合实训项目（选其中之一）**

**1、项目一（天猫电商平台用户行为分析）**  
**（1）实训内容**

基于阿里云天池的“天猫用户行为数据集”，分析用户点击和购买行为，生成简单的用户行为报告。   
**（2）实训要求**

* 数据采集：从天池下载“天猫用户行为数据集”（约100MB，包含用户ID、商品ID、行为类型、时间戳，使用Pandas清洗缺失值和重复记录（参考模块一）。
* 数据存储：将清洗后的数据上传至HDFS，创建Hive表tianmao\_behavior（字段：user\_id、item\_id、behavior\_type、timestamp），导入数据（参考模块二）。
* 数据分析：使用Spark SQL统计每天的点击量和购买量，输出CSV文件（参考模块三）。
* 数据可视化：使用Matplotlib绘制折线图，展示7天的点击量和购买量趋势（参考模块四）。
* 提交清洗代码、HDFS/Hive操作记录、Spark SQL代码、可视化图表和项目文档。  
  **（3）考核要求**

按照要求完成项目，提交一份不小于500字的实验报告，包含项目背景、数据处理流程、分析结果和可视化图表解读。

**2、项目二（豆瓣电影评分数据分析）**  
**（1）实训内容**

基于教师提供的模拟豆瓣电影评分数据集（CSV，约10MB），分析电影评分分布和热门电影排名。

**（2）实训要求**

* 数据采集：从教师提供的模拟豆瓣评分数据集（包含用户ID、电影ID、评分、时间戳，字段格式：user\_id、movie\_id、rating、timestamp），使用Pandas清洗无效评分（如0分或空值）（参考模块一）。
* 数据存储：将清洗数据上传至HDFS，创建Hive表douban\_ratings（字段：user\_id、movie\_id、rating、 timestamp），导入数据（参考模块二）。
* 数据分析：使用Spark SQL统计每部电影的平均评分，筛选评分次数大于20的电影，输出前10部高分电影到CSV（参考模块三）。
* 数据可视化：使用Matplotlib绘制柱状图，展示前10部电影的平均评分（参考模块四）。
* 提交清洗代码、HDFS/Hive操作记录、Spark SQL代码、可视化图表和项目文档。  
  **（3）考核要求**

按照要求完成项目，提交一份不小于500字的实验报告，包含项目背景、数据处理流程、分析结果和可视化图表解读。

**3、项目三（微博评论活跃度统计）**  
**（1）实训内容**

基于教师提供的模拟微博评论数据集（CSV，约50MB，包含评论ID、用户ID、评论内容、时间戳），统计每日评论活跃度。

**（2）实训要求**

* 数据采集：使用Pandas读取教师提供的微博评论CSV，清洗空值评论和无效时间戳（参考模块一）。
* 数据存储：将清洗数据上传至HDFS，创建Hive表weibo\_comments（字段：comment\_id、user\_id、content、timestamp），导入数据（参考模块二）。
* 数据分析：使用Spark SQL统计每日评论数量，输出CSV文件（参考模块三）。
* 数据可视化：使用Matplotlib绘制折线图，展示一周的评论数量趋势（参考模块四）。
* 提交清洗代码、HDFS/Hive操作记录、Spark SQL代码、可视化图表和项目文档。  
  **（3）考核要求**

按照要求完成项目，提交一份不小于500字的实验报告，包含项目背景、数据处理流程、分析结果和可视化图表解读。

**4、项目四（学生成绩分析）**  
**（1）实训内容**

基于教师提供的模拟学生成绩数据集（CSV，约20MB，包含学生ID、课程ID、成绩、考试时间），分析课程的成绩分布。

**（2）实训要求**

* 数据采集：使用Pandas读取成绩CSV，清洗负分或空值成绩（参考模块一）。
* 数据存储：将清洗数据上传至HDFS，创建Hive表student\_scores（字段：student\_id、course\_id、score、exam\_time），导入数据（参考模块二）。
* 数据分析：使用Spark SQL统计每门课程的平均成绩和及格率，输出CSV文件（参考模块三）。
* 数据可视化：使用Matplotlib绘制柱状图，展示前5门课程的平均成绩（参考模块四）。
* 提交清洗代码、HDFS/Hive操作记录、Spark SQL代码、可视化图表和项目文档。  
  **（3）考核要求**

按照要求完成项目，提交一份不小于500字的实验报告，包含项目背景、数据处理流程、分析结果和可视化图表解读。

**七、教学组织与实施方式**

采用"线上为主、线下为辅、混合式教学"的组织形式。

**1、线上教学**

* 学习平台：省开大在线学习平台（资源发布、任务布置、在线答疑）

**2、实践环境**

* 首选方案：云端大数据实验平台（基于Docker容器化技术）
* 备选方案：虚拟机镜像或Docker-Compose脚本（本地部署）

**3、教学环节**

**第一阶段：开题与组队（第1周）**

* 发布项目指南，学生自由组队（3-4人/组）或独立完成
* 选定项目主题，提交初步项目计划书

**第二阶段：核心技能训练（第2-3周）**

* 教师直播讲解四个核心模块的关键技术
* 学生完成基础实验

**第三阶段：项目实施与迭代（第4-7周）**

* 学生按计划开展项目
* 中期线上汇报（展示进度和问题）

**第四阶段：验收与总结（第8周）**

* 提交全部成果物
* 提交个人/小组实训报告

**八、课程资源**

**1、教材与参考书**

* 《Spark权威指南》，Bill，Chambers，Matei，Zaharia，中国电力出版社，2020年4月第1版
* 《Hadoop权威指南》，[美] 汤姆，怀特（Tom White） 著，王海，华东，刘喻，吕粤海 译，清华大学出版社，2017年7月第4版
* 《大数据技术概论》，阮敬，国家开放大学出版社，2022年12月第1版
* 《大数据技术原理与应用 ——概念、存储、处理、 分析与应》，林子雨，人民邮电出版社，2024年8月第4版
* 《数据挖掘实用案例分析（第2版）（大数据技术与应用专业系列教材）》，赵卫东、董亮，清华大学出版社，2024年3月第2版
* 《大数据导论技术实训(大数据应用人才培养系列教材)》，刘鹏，清华大学出版社，2024年1月第1版
* 《Spark项目实训（Python版）（大数据技术与应用丛书）》，黑马程序员，清华大学出版社，2025年3月第1版

**2、工具与平台**

* 大数据组件：Hadoop、Hive、Spark, Kafka、Flume
* 开发语言：Python、Scala
* 开发工具：Jupyter Notebook、PyCharm、VS Code
* 可视化工具：ECharts、Tableau Public

**3、实验报告**

见附录。

**《大数据技术综合实训》**

**实验报告**

**实验项目：**

**专 业：**

**班 级：**

**姓 名：**

**学 号：**

**日 期：**

**指导教师：**

# **实验目的**

|  |
| --- |
| （仿宋，四号字体，单倍间距） |

# **实验内容**

|  |
| --- |
| （仿宋，四号字体，单倍间距） |

# **实验过程**

（仿宋，四号字体，单倍间距）

# **实验总结**

（仿宋，四号字体，单倍间距）