**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：理论指导8学时，实践操作64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、大数据与会计、计算机应用技术、  电子商务等相关专业 |

**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、实训内容与项目设置**

**（一）核心技能模块（必选基础任务）**

|  |  |  |
| --- | --- | --- |
| 模块 | 内容 | 关键技术 |
| 模块一 | 数据采集与接入 | Flume/Kafka/Sqoop/API调用 |
| 模块二 | 数据存储与管理 | HDFS/Hive/Spark SQL |
| 模块三 | 数据处理与分析 | Spark Core/Spark SQL/Spark MLlib |
| 模块四 | 数据可视化与呈现 | ECharts/Tableau/Matplotlib/Seaborn |

**1、模块一：数据采集与简单预处理  
（1）实训内容**

* 使用Python脚本从阿里云天池的公开CSV数据集（如“天猫用户行为数据集”）采集数据。
* 使用Python的Pandas库进行简单数据清洗，处理缺失值和格式标准化。

**（2）实训要求**

* 从阿里云天池下载“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读取CSV，移除缺失值（空值填充为0或均值），统一时间戳格式（如YYYY-MM-DD）。
* 输出清洗后的CSV文件（约1000条记录）。
* 编写不超过50行的Python脚本，注释清晰。
* 提交数据集下载链接、清洗代码和清洗后数据文件。  
  **（3）考核要求**

按照要求完成数据采集与清洗，提交一份不小于500字的实验报告，包含数据集描述、清洗步骤和结果截图。

**2、模块二：数据存储与查询**  
**（1）实训内容**

* 使用HDFS存储清洗后的CSV数据，学习基本文件操作。
* 使用Hive创建简单表，导入数据并执行基础SQL查询。  
  **（2）实训要求**
* 在HDFS上创建目录（如/user/student/data），上传清洗后的CSV文件。
* 在Hive中创建一张表（如user\_behavior），包含字段：用户ID、商品ID、行为类型、时间戳。
* 导入CSV数据到Hive表，执行简单查询（如统计每天的行为次数）。
* 提交HDFS上传命令记录和Hive查询结果（截图）。
* 确保数据量不超过1GB，查询返回结果少于100行。  
  **（3）考核要求**

按照要求完成数据存储与查询，提交一份不小于500字的实验报告，包含HDFS操作步骤、Hive表结构和查询结果。

**3、模块三：简单数据分析**  
**（1）实训内容**

* 使用Spark SQL对Hive中的数据进行简单统计分析。
* 实现一个分析任务（如按商品类别统计点击量或按时间统计评分）。  
  **（2）实训要求**
* 使用Spark SQL读取Hive表，编写查询统计商品类别（或评分等级）的总数。
* 输出结果到CSV文件（包含类别和计数两列）。
* 编写不超过30行的Spark SQL代码，使用Python或Scala。
* 验证结果正确性（如与Hive查询结果对比）。
* 提交Spark代码和分析结果文件。  
  **（3）考核要求**

按照要求完成分析任务，提交一份不小于500字的实验报告，包含分析目标、代码说明和结果解读。

**4、模块四：基础数据可视化**  
**（1）实训内容**

* 使用Python的Matplotlib库对分析结果绘制简单图表。
* 设计柱状图或折线图展示分析结果（如商品点击量趋势或评分分布）。  
  **（2）实训要求**
* 从模块三的CSV结果文件中提取数据，生成柱状图（显示前5个商品类别/评分的计数）。
* 图表需包含标题、轴标签和图例，保存为PNG格式。
* 编写不超过20行的Matplotlib代码。
* 提交图表PNG文件和代码。
* 提供图表结果的简要文字说明（约100字）。  
  **（3）考核要求**

按照要求完成可视化，提交一份不小于500字的实验报告，包含图表设计过程、代码说明和结果解读。

**（二）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1000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1000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1000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1000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。

**四、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）
* 代码管理：Git/GitLab（版本管理和协作）
* 直播互动：腾讯会议（每周线上指导和阶段性汇报）

**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

**五、考核与评价方式**

采用过程性评价与终结性评价相结合的方式。

|  |  |  |  |
| --- | --- | --- | --- |
| 考核项目 | | 权重 | 考核内容与标准 |
| 过程性考核 | 学习过程参与度 | 10% | 在线讨论、出勤、提问与互动的积极性 |
| 核心模块实验报告 | 40% | 四个基础模块的实验完成度、代码质量与报告规范性 |
| 终结性考核 | 最终项目成果 | 50% | **代码与数据(20%)：**功能完整性、代码规范性与可读性  **可视化作品(10%)：**美观度、交互性、洞察力  **实训报告(20%)：**结构完整性、内容详实度、总结反思深度 |
| 总计 | | 100% |  |

**六、课程资源**

**1、教材与参考书**

* 《Spark权威指南》
* 《Hadoop权威指南》
* 其他相关技术文档

**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**

见附件。

**《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）