《Spark大数据计算技术》教学大纲

**课程名称：**Spark大数据计算技术

**课程类别：**必修

**适用专业：**大数据技术类相关专业

**总学时：**56学时（其中理论40学时，实验16学时）

**总学分：**3.5学分

# 课程的性质

近年来，大数据、云计算、人工智能等数字技术与各行业加速融合，数字经济快速发展，数字经济深化发展的核心引擎是数据要素。企业急需具有数据分析技术能力的人才，以便在数字经济浪潮中保持竞争优势。然而，数据源多变、数据量巨大、处理速度缓慢和计算能力不足等问题，使得企业难以用传统的数据分析方法有效分析和利用海量数据。Spark作为一种快速、通用的大数据分析框架，具有兼容多种数据源、支持内存计算、支持分布式计算和可扩展性等优点，得到了广泛的认可和应用。为了推动我国大数据，云计算，人工智能行业的发展，满足日益增长的数据分析人才需求，特开设Spark大数据分析实务课程。

# 课程的任务

通过本课程的学习，使学生学会使用Spark进行数据读取、数据处理、分析与建模，并详细拆解学习广告流量检测违规识别企业案例，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。

# 课程学时分配

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 序号 | 教学内容 | 理论学时 | 实验学时 | 其它 |
| 1 | 项目1 广告流量检测违规识别需求分析 | 2 |  |  |
| 2 | 项目2 Spark大数据环境安装搭建 | 2 | 4 |  |
| 3 | 项目3 基于Hive实现广告流量检测数据存储 | 4 | 4 |  |
| 4 | 项目4 基于Spark SQL实现广告流量检测数据探索分析 | 4 | 4 |  |
| 5 | 项目5 基于Spark SQL实现广告流量检测数据预处理 | 6 | 6 |  |
| 6 | 项目6 基于Spark MLlib实现广告流量检测违规识别模型构建与评估 | 6 | 6 |  |
| 7 | 项目7 基于Spark开发环境实现广告流量检测违规识别 | 6 | 6 |  |
| 8 | 项目8 基于TipDM大数据挖掘建模平台实现广告流量检测违规识别 | 2 | 2 |  |
|  | 总计 | 32 | 32 |  |

# 教学内容及学时安排

### 理论教学

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **序号** | **章节名称** | **主要内容** | **教学目标** | **学时** |
| 1 | 广告流量检测违规识别需求分析 | 1. 大数据的概念 2. 大数据的特征 3. 大数据分析的概念 4. 大数据分析的流程 5. 大数据分析的工具 6. 了解正常的广告投放流程 7. 了解广告流量违规现状 8. 了解广告流量常见违规方式 9. 了解识别虚假流量的常见维度 10. 广告流量检测违规识别流程分析 | 1. 了解大数据的概念和特征。 2. 了解常见的广告流量违规方式。 3. 了解广告流量检测违规识别项目的流程分析。 4. 能够根据项目目标进行流程分析。 5. 能够根据业务需求设计项目总体流程。 | 2 |
| 2 | Spark大数据环境安装搭建 | 1. Hadoop的发展历程 2. Hadoop的特点 3. Hadoop生态系统 4. Hadoop的应用场景 5. Hive简介 6. Hive的特点 7. Hive的应用场景 8. Hive与关系数据库的区别 9. Spark简介 10. Spark的发展历程 11. Spark的特点 12. Spark生态系统 13. Spark的应用场景 14. 安装搭建Hadoop集群 15. 安装搭建Hive 16. 安装搭建Spark集群 | 1. 了解Hadoop框架的历程、特点、生态系统、应用场景。 2. 了解Hive的特点、应用场景以及Hive与关系数据库的区别。 3. 了解Spark的发展历程、特点、生态系统、应用场景。 4. 熟悉Hadoop框架和Spark的架构组成。 5. 能够理解Hadoop、Hive、Spark组件的作用。 6. 能够完成Hadoop集群、Hive和Spark集群的搭建与配置。 | 2 |
| 3 | 基于Hive实现广告流量检测数据存储 | 1. 了解Hive中的数据类型 2. 创建数据库语法格式 3. 删除数据库语法格式 4. 更改数据库语法格式 5. 使用数据库语法格式 6. 创建内部表 7. 创建外部表 8. 创建分区表 9. 创建桶表 10. 修改表 11. 将文件系统中的数据导入Hive表 12. 通过查询语句向表中插入数据 13. 将Hive表中的数据导出至本地文件系统 14. 将Hive表中的数据导出至HDFS 15. 创建数据表 16. 导入数据至Hive表 | 1. 了解Hive中的数据类型。 2. 掌握创建和管理数据库的操作方法。 3. 掌握创建和修改表的操作方法。 4. 掌握Hive数据导入与导出的操作方法。 5. 能够认识Hive中的数据类型。 6. 能够完成数据库的创建与管理。 7. 能够根据要求创建表并导入相应的数据。 | 4 |
| 4 | 基于Spark SQL实现广告流量检测数据探索分析 | 1. Spark SQL简介 2. 配置Spark SQL CLI 3. Spark SQL与Shell交互 4. 通过结构化数据文件创建DataFrame 5. 通过外部数据库创建DataFrame 6. 通过RDD创建DataFrame 7. 通过Hive表创建DataFrame 8. printSchema()：输出数据模式 9. show()：查看数据 10. first(head(take(takeAsList()：获取若干行记录 11. 条件查询 12. 查询指定字段的数据信息 13. 查询指定行数的数据 14. 排序查询 15. 分组查询 16. 读取数据创建DataFrame对象 17. 简单查询DataFrame数据 18. 探索分析日流量特征 19. 探索分析IP地址的访问次数特征 20. 探索分析虚假流量数据特征 | 1. 了解Spark SQL框架的功能及运行过程。 2. 了解Spark SQL与Shell交互。 3. 掌握Spark SQL的可编程数据模型DataFrame的创建、查询等操作方法。 4. 能够配置Spark SQL CLI，提供Spark SQL与Shell交互环境。 5. 能够通过不同数据源创建DataFrame。 6. 能够实现DataFrame数据及行列表的查询操作。 | 4 |
| 5 | 基于Spark SQL实现广告流量检测数据预处理 | 1. 新增数据列 2. 删除数据列 3. 创建与使用用户自定义函数 4. 5种join()方法 5. 根据特定字段进行表联合操作 6. 指定类型进行表联合操作 7. 使用Column类型进行表联合操作 8. 保存为文件 9. 保存为持久化的表 10. 使用drop语句删除数据 11. 自定义函数构建关键特征 12. 保存DataFrame数据至Hive表中 | 1. 掌握DataFrame行列表的增、删操作方法。 2. 掌握用户自定义函数的创建与使用方法。 3. 掌握多种DataFrame表联合操作的方法。 4. 掌握DataFrame保存数据的多种方式。 5. 能够处理DataFrame中的缺失值。 6. 能够实现DataFrame表联合操作。 7. 能够创建和使用用户自定义函数。 8. 能够按照不同需求采用不同方式保存DataFrame数据。 | 6 |
| 6 | 基于Spark MLlib实现广告流量检测违规识别模型构建与评估 | 1. 了解机器学习算法 2. Spark MLlib简介 3. 数据类型 4. 基本统计 5. 管道 6. 特征提取 7. 特征处理 8. 回归 9. 分类 10. 聚类 11. 关联规则 12. 智能推荐 13. 掌握Spark MLlib的评估器与模型评估 14. 使用spark.ml.classification模块构建分类模型 15. 使用评估器实现模型评估 | 1. 了解Spark MLlib算法库。 2. 熟悉Spark MLlib中的算法与算法包。 3. 掌握Spark MLlib的评估器与模型评估的使用方法。 4. 能够掌握Spark MLlib特征提取的方法。 5. 能够使用Spark MLlib回归与分类相关算法包构建模型。 6. 能够使用Spark MLlib评估器对模型进行评估。 | 6 |
| 7 | 基于Spark开发环境实现广告流量检测违规识别 | 1. 安装JDK 8 2. 设置环境变量 3. 下载与安装IntelliJ IDEA 4. Scala插件安装与使用 5. 在IntelliJ IDEA中配置Spark运行环境 6. 运行Spark程序 7. 开发环境下实现流量数据违规识别 8. 模型应用 | 1. 掌握JDK 8安装和环境配置。 2. 掌握IntelliJ IDEA中创建Scala工程。 3. 掌握Spark运行环境配置。 4. 能够在本地环境下配置JDK环境，实现Java环境搭建。 5. 能够在IntelliJ IDEA中配置相关插件和开发依赖包，实现Scala工程创建。 6. 能够在IntelliJ IDEA中配置Spark运行环境，实现Spark程序运行。 | 6 |
| 8 | 基于TipDM大数据挖掘建模平台实现广告流量检测违规识别 | 1. 共享库 2. 数据连接 3. 数据集 4. 我的工程 5. 个人组件 6. 快速构建广告流量检测违规识别工程 | 1. 了解TipDM大数据挖掘建模平台的相关概念和特点。 2. 能够使用TipDM大数据挖掘建模平台完成广告流量检测违规识别项目总体流程的设配置。 3. 能够使用TipDM大数据挖掘建模平台完成项目流程的实现。 | 2 |
| 学 时 合 计 | | | | 32 |

### 实验教学

|  |  |  |  |
| --- | --- | --- | --- |
| **序号** | **实验项目名称** | **实验要求** | **学时** |
| 1 | Spark大数据环境安装搭建 | 1. 创建Linux虚拟机 2. 设置固定IP地址 3. 远程连接虚拟机 4. 配置本地yum源及安装常用软件 5. 在Linux系统下安装Java 6. 修改配置文件 7. 克隆虚拟机 8. 配置SSH免密登录 9. 配置时间同步服务 10. 添加地址映射 11. 启动关闭集群 12. 安装MySQL 13. 下载和安装Hive 14. 修改Hive配置文件 15. 设置环境变量 16. 初始化元数据库与启动Hive 17. 解压并配置Spark集群 18. 启动Spark集群 19. 查看Spark监控服务 | 4 |
| 2 | 基于Hive实现广告流量检测数据存储 | 1. 创建数据库语法格式 2. 删除数据库语法格式 3. 更改数据库语法格式 4. 使用数据库语法格式 5. 创建内部表 6. 创建外部表 7. 创建分区表 8. 创建桶表 9. 修改表 10. 将文件系统中的数据导入Hive表 11. 通过查询语句向表中插入数据 12. 将Hive表中的数据导出至本地文件系统 13. 将Hive表中的数据导出至HDFS 14. 创建数据表 15. 导入数据至Hive表 | 4 |
| 3 | 基于Spark SQL实现广告流量检测数据探索分析 | 1. 配置Spark SQL CLI 2. Spark SQL与Shell交互 3. 通过结构化数据文件创建DataFrame 4. 通过外部数据库创建DataFrame 5. 通过RDD创建DataFrame 6. 通过Hive表创建DataFrame 7. 查看DataFrame数据 8. printSchema()：输出数据模式 9. show()：查看数据 10. first(head(take(takeAsList()：获取若干行记录 11. 条件查询 12. 查询指定字段的数据信息 13. 查询指定行数的数据 14. 排序查询 15. 分组查询 16. 读取数据创建DataFrame对象 17. 查询数据记录数 18. 查询数据缺失值 19. 探索分析日流量特征 20. 探索分析IP地址的访问次数特征 21. 探索分析虚假流量数据特征 | 4 |
| 4 | 基于Spark SQL实现广告流量检测数据预处理 | 1. 新增数据列 2. 删除数据列 3. 创建与使用用户自定义函数 4. 根据特定字段进行表联合操作 5. 指定类型进行表联合操作 6. 使用Column类型进行表联合操作 7. 保存为文件 8. 保存为持久化的表 9. 使用drop语句删除数据 10. 划分时间区间 11. 构建关键特征并保存至Hive表中 12. 保存DataFrame数据至Hive表中 | 6 |
| 5 | 基于Spark MLlib实现广告流量检测违规识别模型构建与评估 | 1. 基本统计 2. 管道 3. 特征提取 4. 特征处理 5. 回归 6. 分类 7. 聚类 8. 关联规则 9. 智能推荐 10. 数据归一化 11. 构建建模样本 12. 使用逻辑回归算法实现广告流量检测违规识别 13. 使用随机森林算法实现广告流量检测违规识别 14. 使用评估器实现模型评估 | 6 |
| 6 | 基于Spark开发环境实现广告流量检测违规识别 | 1. 安装JDK 8 2. 设置环境变量 3. 下载与安装IntelliJ IDEA 4. Scala插件安装与使用 5. 在IntelliJ IDEA中配置Spark运行环境 6. 运行Spark程序 7. 集群连接参数设置 8. 封装代码 9. 运行Spark程序 | 6 |
| 7 | 基于TipDM大数据挖掘建模平台实现广告流量检测违规识别 | 1. 数据源配置 2. 数据处理 3. 模型构建与评估 | 2 |
| 学 时 合 计 | | | 32 |

# 考核方式

突出学生解决实际问题的能力，加强过程性考核。课程考核的成绩构成 = 平时作业（10%）+ 课堂参与（20%）+ 期末考核（70%），期末考试建议采用开卷形式，试题应包括基本概念、分组聚合、数据合并、数据清洗、数据变换、模型构建等部分，题型可采用判断题、选择、简答、应用题等方式。

# 教材与参考资料

### 教材

郑浩森，张荣．Spark大数据技术分析[M]．北京：人民邮电出版社．2024．

### 参考资料

[1] 肖芳，张良均．Spark大数据技术与应用（第2版）（微课版）[M]．北京：人民邮电出版社．2022．

[2] 王哲，张良均．Hadoop与大数据挖掘（第2版）[M]．北京：机械工业出版社．2022．

[3] 张军，张良均．Hadoop大数据开发基础（第2版）（微课版）[M]．北京：人民邮电出版社．2021．