**《Spark大数据技术综合实训》实训课程教学大纲**

**字体、字号请参考范例**

**注意：**

**首字母大写**

**植物拉丁学名斜体**

**一、课程简介**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **课程中文名** | Spark大数据技术综合实训 | | | | | | | | |
| **课程英文名** | Comprehensive Training of Spark Big Data Technology | | | | | | **双语授课** | | □是 ☑否 |
| **课程代码** | 10114091 | **课程学分** | | 2 | **周（学时）** | | | 2周 | |
| **课程类别** | □专业认知实习  □专业见习  ☑工程实训  □毕业实习  □其他 | **课程性质** | | ☑必修  □选修  □其他 | **课程形态** | | | □线上  ☑线下  □线上线下混合式  □社会实践  □虚拟仿真实验教学 | |
| **考核方式** | □闭卷 □开卷 □课程论文 □课程作品 □汇报展示 ☑报告  □课堂表现 □阶段性测试 □平时作业 □其他（可多选） | | | | | | | | |
| **开课学院** | 大数据与智能工程学院 | | **开课**  **系(教研室)** | | | 数据科学与大数据技术 | | | |
| **面向专业** | 数据科学与大数据技术 | | **开课学期** | | | 第5学期 | | | |
| **课程负责人** | 黄金龙 | | **审核人** | | | 张素兰 | | | |
| **先修课程** | 数据采集与预处理、Hadoop大数据技术 | | | | | | | | |
| **后续课程** |  | | | | | | | | |
| **选用教材** | 1. 林子雨. Spark编程基础（Scala版）. 人民邮电出版社, 2022年. | | | | | | | | |
| **参考书目** | 1. [肖芳](https://book.jd.com/writer/%E8%82%96%E8%8A%B3_1.html)，[张良均](https://book.jd.com/writer/%E5%BC%A0%E8%89%AF%E5%9D%87_1.html). Spark大数据技术与应用. 人民邮电出版社, 2018年.  2. [刘彬斌](https://book.jd.com/writer/%E5%88%98%E5%BD%AC%E6%96%8C_1.html)，[李柏章](https://book.jd.com/writer/%E6%9D%8E%E6%9F%8F%E7%AB%A0_1.html)等. Hadoop+Spark大数据技术. [清华大学出版社](https://book.jd.com/publish/%E6%B8%85%E5%8D%8E%E5%A4%A7%E5%AD%A6%E5%87%BA%E7%89%88%E7%A4%BE_1.html), 2018年. | | | | | | | | |
| **课程资源** | 1.大数据spark精选全集视频教程. 网易云课堂. | | | | | | | | |
| **课程简介** | Spark大数据技术综合实训是数据科学雨大数据技术专业必修的综合实践课程。该课程内容主要包括Spark大数据框架的搭建及其工作流程，RDD高级特性，SparkSession，DataFrameg概念及其持久化操作，Spark Streaming，Spark监控与优化等。使学生能够选择不同的Spark组件和技术方法，解决国内主流的大数据应用场景中存在的问题。 | | | | | | | | |

**二、课程目标**

**表2-1 课程目标**

|  |  |
| --- | --- |
| **序号** | **具体课程目标** |
| **课程目标1** | 能够应用计算机和软件工程、操作系统等领域知识，分析和评价大数据相关的工程实践和复杂工程问题的解决方案对社会、健康、安全、法律以及文化的潜在影响，以及这些因素对项目实施的影响，并理解应承担的责。 |
| **课程目标2** | 能够应用Spark的理论与方法、技能与工具，设计出针对Spark框架问题的解决方案，解决Spark框架搭建过程中的问题，并具有健全的人格和健康的身心，具备一定的人际交往能力，能够胜任团队成员的角，与团队其他成员有效合作，并能承担相应责任。 |
| **课程目标3** | 能够针对大数据挖掘，大数据系统开发过程中涉及的复杂工程问题，选择和使用恰当的大数据处理技术和工具，进行信息处理、表达、建模、设计、模拟、验证，并能够在实践中了解国内主流的Spark框架在各个应用中的优势和局限性；树立使用国产软件、保护版权的意识，激发民族自豪感；坚持自主创新，为建设世界科技强国而奋斗。 |

**表2-2 课程目标与毕业要求对应关系**

| **毕业要求** | **指标点** | **课程目标** |
| --- | --- | --- |
| **毕业要求6：**基于大数据应用领域相关背景知识，能够合理分析和评价大数据应用实践和复杂工程问题解决方案对社会、健康、安全、法律以及文化的影响，并理解应承担的责任。【H】 | 6.2能够分析和评价大数据相关的工程实践和复杂工程问题的解决方案对社会、健康、安全、法律以及文化的潜在影响，以及这些因素对项目实施的影响，并理解应承担的责。 | 1 |
| **毕业要求9：**具有良好的综合素质，能够在多学科背景下的团队中理解与承担个体、团队成员以及负责人角色，并发挥相应的作用。【M】 | 9.1具有健全的人格和健康的身心，具备一定的人际交往能力，能够胜任团队成员的角，与团队其他成员有效合作，并能承担相应责任。 | 2 |
| **毕业要求11：**能够理解并掌握工程管理与经济决策方法，熟悉大数据应用领域工程项目管理的基本方法和技术，并能在多学科环境中应用。【M】 | 11.2在多学科背景下，将大数据工程项目方案设计中涉及的时间及成本管理、质量及风险管理、人力资源管理等问题进行最优求。 | 3 |

**三、教学内容及要求**

**（一）学习内容**

通过本次课程设计，进一步掌握和应用Spark分布式计算框架及其相关组件的知识，需掌握的内容包括：

1. Spark高级应用：
2. RDD的内部表现形式，RDD持久化操作，CheckPoint检查点机制 原理。
3. RDD内部特性，包括：RDD分区案例、RDD依赖关系案例、RDD首选位置案例、RDD分区函数案例。
4. RDD持久化操作案例、RDD CheckPoint操作案例，查看其依赖图。
5. 案例：PageRank网页排名。
6. Spark作业的执行原理和Shuffle详解，并介绍共享变量的概念。
7. Spark累加器、广播变量
8. 案例：求共同好友
9. Spark扩展应用：
10. Spark SQL概念和特点，简述DataFrame和DataSet用法。
11. 构建IDEA工程，创建SparkSession对象，并查看其相关内容，三种 方式构建DataFrame 对象。
12. DataFrame 常见SQL 相关操作和几种Spark数据源格式，DataFrame持久化 参数 SaveMode。
13. 案例：SparkSQL 分析 欧洲足球运行员基本信息。
14. Spark Streaming基本概念，Spark Streaming编程模型。
15. 访问文件系统构建DStream流，使用updateStateByKey来更新状态。

**（二）时间安排**

|  |  |  |
| --- | --- | --- |
|  | 上午 | 下午 |
| 周一 | 教师讲解 | RDD持久化操作及检查点机制 |
| 周二 | 教师讲解 | RDD内部特性 |
| 周三 | RDD持久化操作案例、及其依赖图 | 案例：PageRank网页排名 |
| 周四 | 教师讲解 | Spark作业的执行原理 |
| 周五 | Spark累加器、广播变量 | 案例：求共同好友 |
| 周一 | 教师讲解 | Spark SQL概念和特点 |
| 周二 | 创建SparkSession对象的三种方式 | DataFrame 常见SQL 相关操作 |
| 周三 | 教师讲解 | Spark数据源格式，DataFrame持久化 参数 SaveMode |
| 周四 | SparkSQL 分析 欧洲足球运行员基本信息 | Spark Streaming基本概念，Spark Streaming编程模型。 |
| 周五 | 教师讲解 | 访问文件系统构建DStream流，使用updateStateByKey来更新状态 |

**（三）工作流程**

1)技术讲解：教师应用讲授法、演示法、实验法、任务驱动法、讨论法、自主学习法，讲解综合实训应用的知识点,及各个软件的使用要点。

2)问题分析：通过对项目需要解决的问题进行分析，确定各个步骤的技术方案，书写对问题分析的文档。

3）技术方案设计：对项目所需要的总体技术方案进行设计。

4)撰写总结材料：总结实验过程，并形成实验报告，分析收获和不足。

5)评分：根据评分标准对每个同学的项目的设计、完成度、项目演示、文档撰写等结果进行综合评分。

**（四）业务指导**

在综合课程设计期间，由校内老师1名和企业老师1名，每天深入机房，随时指导解答问题。

## 四、课程考核

**（一）考核内容与考核方式**

**表4-1 课程目标、考核内容与考核方式对应关系**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **课程目标** | **考核内容** | **所属环节** | **考核**  **占比** | **考核方式** |
| 课程目标 1 | 1. RDD的内部表现形式，RDD持久化操作 | Spark高级应用 | 46% | 实验报告 |
| 2. RDD内部特性 |
| 3.Spark SQL概念和特点 | Spark扩展应用 |
| 4. 创建SparkSession对象，并查看其相关内容 |
| 课程目标 2 | 1. RDD持久化操作案例、RDD CheckPoint操作案例，查看其依赖图 | Spark高级应用 | 35% | 实验报告 |
| 2. Spark作业的执行原理和Shuffle详解 |
| 3. Spark累加器、广播变量 |
| 4. DataFrame 常见SQL 相关操作和几种Spark数据源格式 | Spark扩展应用 |
| 课程目标 3 | 1. PageRank网页排名 | Spark高级应用 | 19% | 实验报告 |
| 2. 求共同好友 |
| 3. SparkSQL 分析 欧洲足球运行员基本信息 | Spark扩展应用 |

**表4-2 课程目标与考核方式矩阵关系**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 课程  目标 | 考核方式 | | | | | | | | 考核占比 | |
| 平时考核（40%） | | | | | | 期末考核  （60%） | |
| 课堂交互（10%） | | 实验报告  （20%） | | 课后练习  （10%） | |
| 课程目标1 | 60% | 60% | | 40% | | 40% | | 46%=10%\*60%+20%\*60%+10%\*40%+60%\*40% | |
| 课程目标2 | 30% | 20% | | 40% | | 40% | | 35%=10%\*30%+20%\*20%+10%\*40%+60%\*40% | |
| 课程目标3 | 10% | 20% | | 20% | | 20% | | 19%=10%\*10%+20%\*20%+10%\*20%+60%\*20% | |

**（二）成绩评定**

**1.平时成绩评定**

1）课堂交互：通过学生在课堂上的表现情况、提问与回答、与教师互动情况，评价学生相关的能力。（包括理想信念、家国情怀、社会责任、学习态度、学习兴趣、团队合作与终身学习意识等）

2）实验报告：评价学生收集课程知识相关资料、撰写文本的能力；研究并提出问题，进而解决问题和合作研究的能力。

3）团队协作：学生在过程中协调、沟通等的情况。

平时成绩（100%）=课堂交互（25%）+实验报告（50%）+团队协作（25%）

**2.期末成绩评定**

综合课程设计过程中，根据学生的设计报告、实验的参与和完成情况，以及其他能够评价学生设计学习情况的因素评定成绩。期末成绩的评定为作品设计成绩，即系统代码实现。

**3.总成绩评定**

总成绩（100%）=平时成绩（40%）+期末成绩（60%）

**（三）评分标准**

评分依据：

1)优秀（90-100分）

按设计任务书要求圆满完成规定任务；综合运用知识能力和实践动手能力强，软件设计方案合理，实验效果好；设计态度认真，独立工作能力强，并具有良好的团队协作精神。设计报告条理清晰、论述充分、图表规范、符合设计报告文本格式要求。 答辩过程中，思路清晰、论点正确、对设计方案理解深入，问题回答正确。

2)良好（80-89分）

按设计任务书要求完成规定设计任务；综合运用知识能力和实践动手能力较强，软件设计方案较合理，实验效果较好；设计成果质量较高；设计态度认真，有一定的独立工作能力，并具有较好的团队协作精神。设计报告条理清晰、论述正确、图表较为规范、符合设计报告文本格式要求。 答辩过程中，思路清晰、论点基本正确、对设计方案理解较深入，主要问题回答基本正确。

3)中等（70-79分）

按设计任务书要求完成规定设计任务；能够一定程度的综合运用所学知识，软件设计基本合理，有一定的实践动手能力，设计成果质量一般；设计态度较为认真，设计报告条理基本清晰、论述基本正确、文字通顺、图表基本规范、符合设计报告文本格式要求，但独立工作能力较差；答辩过程中，思路比较清晰、论点有个别错误，分析不够深入。

4)及格（60-69分）

在指导教师及同学的帮助下，能按期完成规定设计任务；综合运用所学知识能力及实践动手能力较差，设计方案基本合理，设计成果质量一般；独立工作能力差；或设计报告条理不够清晰、论述不够充分但没有原则性错误、文字基本通顺、图表不够规范、符合设计报告文本格式要求；或答辩过程中，主要问题经启发能回答，但分析较为肤浅。

5)不及格（60分以下）

未能按期完成规定设计任务。不能综合运用所学知识，实践动手能力差，设计方案存在原则性错误，计算、分析错误较多；或设计报告条理不清、论述有原则性错误、图表不规范、质量很差；或答辩过程中，主要问题阐述不清，对设计内容缺乏了解，概念模糊，问题基本回答不出。

## 五、其他说明

本课程大纲依据2023版数据科学与大数据技术专业人才培养方案，由大数据与智能工程院（部）数据科学与大数据技术教学系（教研室）讨论制定，大数据与智能工程院（部）教学工作委员会审定，教务处审核批准，自2023级开始执行。

**字体、字号请参考范例**

**注意：**

**首字母大写**

**植物拉丁学名斜体**

**字体、字号请参考范例**

**注意：**

**首字母大写**

**植物拉丁学名斜体**

**字体、字号请参考范例**

**注意：**

**首字母大写**

**植物拉丁学名斜体**