《Spark技术与应用》教学大纲

**一、课程基本信息**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **课程类别** | 专业课程 | | **课程性质** | 选修 | **课程属性** | 理论 | |
| **课程名称** | Spark技术与应用 | | | **课程英文名称** | Spark Technology and Application | | |
| **课程编码** | H35X029D | | | **适用专业** | 数据科学与大数据 | | |
| **考核方式** | 考查 | | | **先修课程** | python程序设计、Hadoop应用开发 | | |
| **总学时** | 32 | **学分** | | **2** | **理论学时** | | 16 |
| **实验学时/实训学时/ 实践学时/上机学时** | | | | 上机学时：16 | | | |

**二、课程简介**

《Spark技术与应用》是本科高等学校计算机类大数据方向一门重要的专业应用课程。本课程以编程语言（python语言）为平台，通过对Spark分布式框架、编程、组件等内容的讲授，培养学生对海量数据的处理、分析以及应用开发能力。Spark作为一款分布式计算框架款，除了继承Hadoop MapReduce所具备的优点外，进一步支持输出结果数据存储在内存中，从而大幅度减少了HDFS的读写次数，达到实时计算的需求。目前，Spark技术已广泛应用于企业之中，学习spark框架已成为进入大数据开发行业不可或缺的知识内容。

**三、课程教学目标**

|  |  |  |  |
| --- | --- | --- | --- |
| **课程教学目标** | | **支撑人才培养规格指标点** | **支撑人才培养规格** |
| **知**  **识**  **目**  **标** | **目标1：**  学生需要掌握spark框架基本概念；RDD的相关编程；spark sql、streaming、mllib的使用; | 1-1：具备大数据专业理论知识和专业知识 | 1.工程知识 |
| **能**  **力**  **目**  **标** | **目标2：**  在实践中掌握spark编程语法，逐步培养在数据处理、开发整体流程的应用能力，能够根据实际问题对spark程序进行调优处理。 | 2-1：具备大数据知识运用能力，能将大数据和云计算基本知识用于分析和解决复杂工程问题；  3-1：具备大数据应用程序开发实践能力，能根据软件需要，设计简单的解决方案。 | 2.问题分析  3.设计/开发解决方案 |
| **素**  **质**  **目**  **标** | **目标3：**  通过本课程的学习，培养作为一个工程技术人员必须具备的坚持不懈的学习精神，严谨治学的科学态度和积极向上的价值观，为未来的学习、工作和生活奠定良好的基础。 | 12-3：能针对个人或职业发展的需求，采用合适的方法自主学习，以适应社会和行业发展。 | 12. 终身学习 |

**四、课程主要教学内容、学时安排及教学策略**

**（一）理论教学**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **教学模块** | **学时** | **主要教学内容与策略** | **学习任务安排** | **支撑课程目标** |
| Spark概述 | 2 | **重点：**Spark基本概念；Spark框架组成与原理；Spark发展与应用；Spark安装与配置  **难点：**Spark框架组成与原理。  **教学方法与策略：**线下教学。对于概念性原理在课堂上予以讲授，对于Spark安装与配置应安排上机实践。课堂运用主要运用讲授法和案例法开展教学，从而使学生对spark相关概念有基本的理解。 | 课前：查阅资料了解spark框架的应用场景  课堂：掌握Spark框架的组成以及相关概念  课后：  巩固复习，通过做题加深印象 | 目标1  目标3 |
| Spark RDD编程 | 2 | **重点：**RDD的定义以及相关操作；transform算子与action算子；文件读取输出；  **难点：** RDD的相关编程操作；transform算子与action算子编程  **教学方法与策略：**线下教学。对于涉及RDD相关编程的内容需要在课堂上进行演示，同时增加对照组辅助学生理解每一个操作的应用效果以及使用场合 | 课前：查阅资料了解RDD的概念  课堂：掌握了解RDD的两种算子  课后：  对两种算子包含的操作进行实际运用 | 目标1  目标3 |
| Spark进阶编程 | **2** | **重点：**key-value算子以及相关操作；依赖关系；共享变量；RDD持久化；  **难点：**键值对概念的引入以及在拥有key的情况下对RDD可进行的相关操作；对Spark依赖关系的解释；  **教学方法与策略：**线下教学。对于涉及到键值对的操作应有相关动图辅助学生理解数据划分的过程。同时需在课堂上演示不同模块的代码，并穿插有关依赖关系的知识进行讲授。 | 课前：查阅资料了解键值对的概念  课堂：掌握了解键值对RDD的相关操作，并从中理解宽窄依赖的判断方式。同时掌握变量共享的方法，并  课后：  巩固复习，通过实际编程加深印象 | 目标1  目标3 |
| Spark SQL | **4** | **重点：**Spark SQL的创建，以及Dataframe常见函数的使用；利用相关语句结合matplot、echarts进行数据可视化；结合MongoDB实现数据存储；  **难点：**spark sql的相关语句使用；常见数据可视化图的实现；MongoDB相关内容介绍  **教学方法与策略：**线下教学，对于dataframe操作的常见函数功能、数据可视化图表、Mongo数据库相关内容进行实际演示。同时应讲授如何利用spark sql官方文档查询所需功能，培养学生自主探索能力。 | 课前：查阅资料了解非关系型数据库的概念  课堂：掌握了解spark sql常见语句，数据可视化操作，从MongoDB中读写数据  课后：  巩固复习，通过实际编程加深印象。同时学习查阅官方文档 | 目标1  目标2  目标3 |
| Spark streaming | **4** | **重点：**Kafka相关功能介绍；spark streaming运行的原理；Dstream数据流的转换；  **难点：**Kafka数据发送与接收实现；spark streaming运行原理；Dstream数据流的转换操作  **教学方法与策略：**线下教学，应对kafka的工作原理、相关角色进行简要讲授，辅助学生理解kafka数据发送、接收、订阅等过程。streaming的运行原理解析需有图解辅助学生理解，同时应演示Dstream数据流的相关转换操作 | 课前：查阅资料了解KafKa的作用  课堂：掌握了解spark streaming的处理原理以及应用  课后：  巩固复习，通过实际编程加深印象。 | 目标1  目标2  目标3 |
| Spark MLlib | **2** | **重点：**机器学习库的介绍，数据预处理过程；常见机器学习算法的应用；  **难点：** 数据预处理过程；常见机器学习算法的介绍与使用  **教学方法与策略：**线下教学，可利用案例进行解析数据预处理带来的影响。同时可演示一些简单案例进行ml算法的讲授，辅助学生理解不同算法的使用以及应用场景 | 课前：查阅资料了解机器学习算法的应用场景  课堂：掌握了解数据预处理方法，对于预处理完毕的数据可以使用ML算法进行应用  课后：  巩固复习，通过实际编程加深印象 ，同时学习查阅官方文档，学习调参方法 | 目标1  目标2  目标3 |

**（二）实践教学**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **实践类型** | **项目名称** | **学时** | **主要教学内容** | **项目**  **类型** | **项目**  **要求** | **支撑课程目标** |
| 上机 | RDD编程 | 2 | **重点：**rdd的创建；文件的输入输出；action与transform算子的应用；  **难点：**action与transform算子的实践  **思政元素：**要求学生处理数据必须坚持实事求实、严谨的科学态度。 | 验证 | 实验单人完成，须完成实验报告。实验报告须有详细的实验记录 | 目标1  目标3 |
| 上机 | RDD进阶编程 | 4 | 重点：创建键值对RDD； 键值对RDD的相关算子应用；共享变量的实践；RDD持久化实践；  难点： 键值对RDD相关算子的应用  思政元素：要求学生处理数据必须坚持实事求实、严谨的科学态度。 | 验证 | 实验单人完成，须完成实验报告。实验报告须有详细的实验记录 | 目标1  目标3 |
| 实验 | Spark SQL | 4 | 重点：Spark SQL常见语句的应用；读写入mongo数据操作；数据可视化实现；  难点： 从mongo中读取数据，利用Spark SQL进行相应处理，最终利用数据可视化进行分析。  思政元素：要求学生在应用所学语句之外，需自行利用网络资源如spark sql官方文档。培养学生主动探索、不断发展的能力 | 综合 | 实验单人完成，须完成实验报告。实验报告须有详细的实验记录 | 目标1  目标2  目标3 |
| 上机 | Spark streaming | 4 | 重点：利用kafka模拟实时数据的发送与接收；topic订阅；Dstream数据流的相关应用  难点： kafka相关模拟实时数据的相关操作；对实时数据的处理应用 | 验证 | 实验单人完成，须完成实验报告。实验报告须有详细的实验记录 | 目标1  目标2  目标3 |
| 上机 | Spark MLlib | 2 | 重点：数据预处理方法；常见机器学习算法的应用  难点： 数据预处理方法；机器学习算法的应用以及参数调整  思政元素：要求学生在应用所学算法时，需结合网络资源如spark mllib官方文档进行调整。培养学生主动探索、不断发展的能力 | 综合 | 实验单人完成，须完成实验报告。实验报告须有详细的实验记录 | 目标1目标2  目标3 |
|  | 备注： 项目类型填写验证、综合、设计、训练等。 | | | | | |

**五、学生学习成效评估方式及标准**

考核与评价是对课程教学目标中的知识目标、能力目标和素质目标等进行综合评价。在本课程中，学生的最终成绩是由平时成绩、上机报告成绩、期末考试成绩三个部分组成。

1.平时成绩（占总成绩的40%）：采用百分制。平时成绩分作业（占10%）、上机报告成绩（占20%）和考勤（占10%）三个部分。评分标准如下表：

|  |  |
| --- | --- |
| **分数** | **评 分 标 准** |
| **1.作业；2.上机报告；3.考勤** |
| 90～100分 | 1.作业书写工整、书面整洁、流程描述逻辑清晰；90％以上的习题解答正确或实验习题结果准确无误。  2.未出现迟到、缺勤现象 |
| 80～89分 | 1.作业书写工整、书面整洁、流程描述逻辑较为清晰；；80％以上的习题解答正确或实验习题结果准确无误。  2.未出现迟到、缺勤现象 |
| 70～79分 | 1.作业书写较工整、书面较整洁、流程描述逻辑较为清晰；70％以上的习题解答正确或实验习题结果准确无误。  2. 有不超过3次的迟到或缺勤记录 |
| 60～69分 | 1.作业书写一般、书面整洁度一般、流程描述逻辑较为混乱；60％以上的习题解答正确或实验习题结果准确无误。  2. 有不超过3次的迟到或缺勤记录 |
| 60分以下 | 1.字迹模糊、卷面书写零乱、流程描述逻辑混乱；超过40％的习题解答不正确或实验习题结果错误。  2.有不超过3次的迟到或缺勤记录 |

2.期末考试（占总成绩的60%）：采用百分制。期末考试的考核内容、题型和分值分配情况请见下表：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **考核**  **模块** | **考核内容** | **主要**  **题型** | **支撑目标** | **分值** |
| Spark基本概念 | spark框架的基本概念、原理。 | 简答题 | 目标3 | 6 |
| spark框架的组成 | 填空题 | 目标3 | 4 |
| RDD编程 | 读取文件、rdd的相关算子应用 | 编程题 | 目标3 | 6 |
| Spark core、RDD相关概念 | 填空题 | 目标4 | 4 |
| RDD进阶编程 | 键值对RDD的相关算子应用 | 编程题 | 目标3 | 6 |
| RDD持久化 | 选择题 | 目标3 | 2 |
| 共享变量的作用与应用 | 选择题 | 目标3 | 4 |
| 宽窄依赖关系判断 | 填空题 | 目标3 | 2 |
| shuffle与DAG相关概念 | 选择题 | 目标3 | 6 |
| Spark SQL | mongoDB的相关概念 | 简答题 | 目标3 | 5 |
| spark sql命令应用 | 编程题 | 目标4 | 10 |
| 数据可视化处理与分析 | 编程题 | 目标4 | 5 |
| Spark Streaming | Kafka的相关概念与作用 | 简答题 | 目标3 | 5 |
| Dstream处理应用 | 编程题 | 目标4 | 10 |
| Spark MLlib | 数据预处理过程描述 | 简答题 | 目标3 | 5 |
| 数据预处理实际应用 | 编程题 | 目标4 | 8 |
| 机器学习算法的应用 | 编程题 | 目标4 | 12 |

**六、教学安排及要求**

|  |  |  |
| --- | --- | --- |
| **序号** | **教学安排事项** | **要 求** |
| 1 | 授课教师 | 职称：讲师或以上 学历（位）：硕士或以上  其他： |
| 2 | 课程时间 | 周次：1次/周  节次：2节/次 |
| 3 | 授课地点 | ☑教室 ☑实验室 □室外场地  □其他： |
| 4 | 学生辅导 | 线上方式及时间安排：企业微信，周五14:30~16:30  线下地点及时间安排：办公室3A111，周四下午14:30~16:30 |

**七、选用教材**

[1] 卡劳,肯维尼斯科,温德尔,扎哈利亚. Spark快速大数据分析 [M].北京:人民邮电出版社,2021-05-01.

[2] 赵红艳,许桂秋.Spark大数据技术与应用 [M].北京:人民邮电出版社,2019-04-01.

**八、参考资料**

[1] 卡劳,肯维尼斯科,温德尔,扎哈利亚. Spark快速大数据分析 [M].北京:人民邮电出版社,2021-05-01.

[2] 赵红艳,许桂秋.Spark大数据技术与应用 [M].北京:人民邮电出版社,2019-04-01.

[3] 肖芳,张良均.Spark大数据技术与应用 [M].北京:人民邮电出版社,2018-02-01.

**网络资料**

[1] PySpark官方文档,https://spark.apache.org/docs/latest/api/python/  
index.html

执笔人： 李红艳

参与人：邓美玲、杨来

系（教研室）主任：陈强

学院（部）审核人：郭松