**广州城建职业学院**

**课 程 标 准**

**课 程 名 称： Spark技术应用**

**教 研 室： 软件技术**

**教 学 单 位： 信息工程学院**

**执 笔 人： 蔡启明**

**审 核 人： 张 宇**

**制 订 时 间： 2024年2月**

**教务处制**

**2024年2月**

**《Spark技术应用》课程标准**

**一、课程基本信息**

|  |  |  |  |
| --- | --- | --- | --- |
| 课程代码 | 专业限选课4 | 课程名称 | Spark技术应用 |
| 课程学分 | 6 | 课程学时 | 96 |
| 课程类别 | 专业限选课 | 考核方式 | 机试 |
| 开设学期 | 大二第2个学期 | 适用专业 | 大数据技术 |
| 先修课程 | 大数据概论、Linux操作系统、Java程序语言设计、Hadoop大数据技术 | | |
| 后续课程 | 大数据项目开发实战 | | |

**二、课程定位**

**（一）课程性质**

本课程是大数据技术专业选修的一门专业核心技能课程，是在学习大数据概论、Linux操作系统、Hadoop大数据技术等课程、具备了大数据基础知识和大数据平台部署能力的基础上，开设的一门理论+实践课程，其功能是对接专业人才培养目标，面向大数据平台开发工程师和大数据分析工程师工作岗位，培养构建Spark实时处理平台能力和基于Spark进行大数据分析和处理的能力，为后续在Spark技术应用领域深耕细作奠定基础。

**（二）课程功能定位**

表1 课程功能定位分析表

|  |  |  |
| --- | --- | --- |
| 类别 | 对接的工作岗位/内容 | 对接培养的职业岗位能力 |
| 岗位 | 大数据平台开发工程师 | 1．熟练掌握Spark工作原理，能解决分布式计算实施过程中的问题。 |
| 2．能构建Spark实时处理平台。 |
| 大数据分析工程师 | 1．能基于Spark进行大数据分析和处理。 |
| 2．能使用Scala在Spark环境中进行编程。 |
| X证书 | 大数据应用技术（hadoop/spark） 初、中、高级证书 | Spark平台环境下大数据分析和数据处理能力。 |
| 竞赛 | 全国职业院校学生专业技术大赛“大数据技术与应用”省赛 | 基于 Spark平台环境下，充分利用 Spark Core、Spark SQL等技术的特点，综合软件开发相关技术，解决实际问题的能力。 |

**三、课程目标**

**（一）课程总目标**

通过本课程的学习，让学生掌握大数据技术概述、Scala语言基础、Spark的设计与运行原理、Spark环境搭建和使用方法、RDD编程、Spark SQL、Spark Streaming、Spark MLlib等知识，为学生在Spark技术应用领域深耕细作奠定基础、指明方向。

1. **课程具体目标**
2. 知识目标

（1）掌握大数据关键技术和大数据相关软件的用途。

（2）掌握Spark的技术优势和在大数据技术中的地位和作用。

（3）掌握Scala编程基础，为学习基于Scala语言的Spark编程奠定基础。

（4）掌握Spark的设计与运行原理。

（5）掌握Spark环境的搭建和使用方法，为开展Spark编程实践铺平道路。

（6）掌握Spark RDD编程，包括RDD的创建、操作API、持久化、分区以及键值对RDD。

（7）掌握结构化数据处理组件Spark SQL，包括DataFrame数据模型、创建方法和常用操作。

（8）掌握Spark流计算框架Spark Streaming，掌握流数据实时计算需求。

（9）掌握Spark机器学习库MLlib，包括MLlib的基本原理、算法、模型选择和超参数调整方法等。

2. 能力目标

（1）能够建立对大数据知识体系的轮廓性认识，了解大数据发展历程、基本概念、主要影响、应用领域、关键技术、计算模式和产业发展，并了解云计算、物联网的概念及其与大数据之间的紧密关系；

（2）能够画出Spark架构图并模拟Spark生态系统结构图，理解Spark的技术优势。

（3）能够理解Scala基础语法，读懂Scala程序代码，使用Scala语言进行简单编程。

（4）能够掌握Spark运行架构、RDD的设计与运行原理和Spark的部署模式。

（5）能够完成Spark的安装和使用并编写WordCount程序。

（6）能够完成开发工具Intellij idea和Eclipse的搭建，并使用这2种工具完成Spark程序的编写、打包和运行。

（7）能够进行Spark基础编程，包括：RDD编程、键值对RDD、共享变量、数据读写、文件数据读写和读写HBase数据。

（8）能够进行Spark SQL编程，包括：Spark SQL简介、DataFrame与RDD的区别、DataFrame的创建、从RDD转换得到DataFrame、读写Parquet(DataFrame)、通过JDBC连接数据库(DataFrame)和连接Hive读写数据。

（9）能够进行Spark Streaming编程，包括：DStream操作，基本输入源文件流、 套接字流和RDD队列流操作，高级数据源Apache Kafka、Apache Flume操作，转换操作和输出操作。

（10）能够进行Spark机器学习库MLlib编程，包括：构建机器学习工作流、特征抽取、转化和选择、聚类算法、推荐算法和机器学习参数调优。

3. 素质目标（含课程思政目标）

（1）养成善于思考、深入研究的良好自主学习的习惯和创新精神；

（2）培养结构化程序设计思想和良好的编码规范；

（3）培养细致缜密的工作态度、团结协作的良好品质、沟通交流和学生的书面表达能力；

（4）养成爱岗敬业、遵守职业道德规范、诚实、守信的高尚品质。

**四、课程内容与教学设计**

**（一）内容模块**

表2 课程内容模块及学时分配

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 序号 | 项目（模块） | 教学内容 | 学时 | | |
| 理论 | 实践 | 小计 |
| 1 | 大数据技术概述 | 大数据的基本概念、关键技术和代表性软件 | 1 | 1 | 2 |
| 2 | Scala语言基础 | 介绍Scala语言基础语法 | 12 | 12 | 24 |
| 3 | Spark运行架构与原理与集群搭建 | Spark简介、运行架构、核心数据集RDD、Spark伪分布式搭建、Spark On Yarn模式搭建 | 3 | 3 | 6 |
| 4 | Spark编程基础 | RDD编程、键值对RDD、数据读写 | 8 | 8 | 16 |
| 5 | Spark SQL | Spark SQL简介、DataFrame、读写Parquet、通过JDBC连接数据库、连接Hive读写数据 | 8 | 8 | 16 |
| 6 | Spark Streaming | 流计算简介、Spark Streaming简介、DStream操作 | 8 | 8 | 16 |
| 7 | Spark MLlib | Spark MLlib简介、spark.mllib库、spark.ml库 | 8 | 8 | 16 |
| 合计 | | | 48 | 48 | 96 |

**（二）教学设计**

表3 课程教学设计

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 序号 | 项目（模块） | 教学内容 | 任务名称 | 教学方法与手段 | 学时安排 | 考核方式 |
| 1 | 单元01：大数据技术概述 | 大数据的关键技术；Spark生态图和Spark架构图。 | 认识Spark生态图 | 教学做一体化 | 2 | 实训项目考核 |
| 2 | 单元02：Scala基础 | Scala声明值和变量、基本数据类型和操作、从控制台读取数据、打印信息到控制台。 | Scala基础语法 | 教学做一体化 | 4 | 实训项目考核 |
| 3 | 单元03：Scala控制结构 | Scala三种控制结构：if条件表达式、while循环和for循环； | Scala数组和列表 | 教学做一体化 | 4 | 实训项目考核 |
| 4 | 单元04：Scala集合的运用 | Scala数据结构：数组、列表、元组、集合和映射。 | Scala数组、列表、元组、集合和映射的使用 | 教学做一体化 | 4 | 实训项目考核 |
| 5 | 单元05：Scala类、对象、继承、特质和模式匹配 | Scala面向对象技术：类、对象、继承、特质和模式匹配。 | Scala面向对象技术 | 教学做一体化 | 4 | 实训项目考核 |
| 6 | 单元06：Scala函数式编程开发技巧 | Scala函数式编程，包括Scala函数定义和高阶函数；Scala集合操作：map操作、flatMap操作。 | Scala函数式编程 | 教学做一体化 | 4 | 实训项目考核 |
| 7 | 单元07： 函数组合器介绍 | Scala组合器操作：filter操作、reduce操作、fold操作；  函数式编程实例WordCount。 | Scala集合操作 | 教学做一体化 | 4 | 实训项目考核 |
| 8 | 单元08：Spark运行架构与原理 | Spark架构、作业运行流程、核心数据集RDD核心原理。 | Spark的运行架构 | 教学做一体化 | 2 | 实训项目考核 |
| 9 | 单元09：Spark集群搭建 | Spark伪分布式集群搭建；Spark On Yarn模式搭建 | Spark的部署与安装 | 教学做一体化 | 4 | 实训项目考核 |
| 10 | 单元10：RDD编程 | RDD的含义；RDD的创建步骤；RDD的转换操作和行动操作。 | RDD的创建和转换操作 | 教学做一体化 | 4 | 实训项目考核 |
| 11 | 单元11：键值对RDD | 键值对RDD的创建方式；  常用的键值对转换操作。 | RDD常用键值对转换 | 教学做一体化 | 4 | 实训项目考核 |
| 12 | 单元12：共享变量 | 掌握广播变量和累计器2种共享变量； | 共享变量 | 教学做一体化 | 2 | 实训项目考核 |
| 13 | 单元13：数据读写—文件数据读写 | 数据读写操作，包括从文件数据中读写和从HDFS中进行数据读写操作。 | 文件数据读写 | 教学做一体化 | 4 | 实训项目考核 |
| 14 | 单元14：RDD持久化与数据分区 | 数据缓存、数据检查点机制、分区partition函数的使用 | 数据持久化与分区 | 教学做一体化 | 2 | 实训项目考核 |
| 15 | 单元15：Spark SQL的基本原理 | Spark SQL的基本原理；DataFrame与RDD的区别。 | Spark SQL的基本原理 | 教学做一体化 | 2 | 实训项目考核 |
| 16 | 单元16：DataFrame的创建 | DataFrame的创建过程；从RDD转换得到DataFrame的过程。 | DataFrame的创建过程 | 教学做一体化 | 4 | 实训项目考核 |
| 17 | 单元17：读取和保存数据 | Spark SQL读取和保存数据：读写csv、JSON。 | Spark SQL读取和保存数据 | 教学做一体化 | 4 | 实训项目考核 |
| 18 | 单元18：读取和保存数据—通过JDBC连接数据库(DataFrame) | Spark SQL读取和保存数据：通过JDBC连接数据库(DataFrame)。 | 通过JDBC连接数据库 | 教学做一体化 | 2 | 实训项目考核 |
| 19 | 单元19：读取和保存数据—连接Hive读写数据 | Spark SQL读取和保存数据：连接Hive读写数据。 | 连接Hive读写数据 | 教学做一体化 | 4 | 实训项目考核 |
| 20 | 单元20：Spark Streaming基本原理 | 流计算简介和Spark Streaming基本原理。 | Spark Streaming基本原理 | 教学做一体化 | 2 | 实训项目考核 |
| 21 | 单元21： DStream基本输入源—文件流、套接字流 | DStream各种操作过程，包括基本输入源、高级数据源、转换操作和输出操作。 | DStream操作 | 教学做一体化 | 4 | 实训项目考核 |
| 22 | 单元22：Dstream转换操作 | DStream基本转换操作于transform转换操作 | DStream转换操作 | 教学做一体化 | 4 | 实训项目考核 |
| 23 | 单元23：DStream高级输入源—Apache Kafka | DStream高级输入源操作——KaFka。 | DStream高级输入源操作 | 教学做一体化 | 4 | 实训项目考核 |
| 24 | 单元24： DStream窗口计算 | DStream窗口计算操作 | DStream窗口计算 | 教学做一体化 | 2 | 实训项目考核 |
| 25 | 单元25：Spark MLlib机器学习工作流 | Spark MLlib的基本原理；掌握一个机器学习工作流的构建。 | Spark MLlib的基本原理 | 教学做一体化 | 4 | 实训项目考核 |
| 26 | 单元26：特征抽取、转化和选择操作 | 特征抽取（TF-IDF、Word2Vec和CountVectorizer）、特征变化（标签和索引的转化）和特征选择（卡方选择器）操作。 | 特征抽取（TF-IDF、Word2Ve | 教学做一体化 | 4 | 实训项目考核 |
| 27 | 单元27：分类与回归算法 | 分类与回归算法（逻辑斯蒂回归分类器和决策树分类器）。 | 分类与回归算法（逻辑斯蒂回归分类器和决策树分类器）。 | 教学做一体化 | 4 | 实训项目考核 |
| 28 | 单元28：聚类算法 | 聚类算法和推荐算法；机器学习参数调优。 | 聚类算法和推荐算法 | 教学做一体化 | 4 | 实训项目考核 |

**（三）实践项目（任务）设计**

表4 课程实践项目（任务）设计

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | 项目（任务）名称 | 学生实践结果（可展示） | 学时安排 |
| 1 | 单元01：大数据技术概述 | 项目实训程序代码 | 1 |
| 2 | 单元02：Scala基础 | 项目实训程序代码 | 2 |
| 3 | 单元03：Scala控制结构 | 项目实训程序代码 | 2 |
| 4 | 单元04：Scala集合的运用 | 项目实训程序代码 | 2 |
| 5 | 单元05：Scala类、对象、继承、特质和模式匹配 | 项目实训程序代码 | 2 |
| 6 | 单元06：Scala函数式编程开发技巧 | 项目实训程序代码 | 2 |
| 7 | 单元07：函数组合器介绍 | 项目实训程序代码 | 2 |
| 8 | 单元08：Spark简介和运行架构 | 项目实训程序代码 | 1 |
| 9 | 单元09：Spark集群搭建 | 项目实训程序代码 | 2 |
| 10 | 单元10：RDD编程 | 项目实训程序代码 | 2 |
| 11 | 单元11：键值对RDD | 项目实训程序代码 | 2 |
| 12 | 单元12：共享变量 | 项目实训程序代码 | 1 |
| 13 | 单元13：数据读写—文件数据读写 | 项目实训程序代码 | 2 |
| 14 | 单元14：RDD持久化与数据分区 | 项目实训程序代码 | 1 |
| 15 | 单元15：Spark SQL的基本原理 | 项目实训程序代码 | 1 |
| 16 | 单元16：DataFrame的创建 | 项目实训程序代码 | 2 |
| 17 | 单元17：读取和保存数据 | 项目实训程序代码 | 2 |
| 18 | 单元18：读取和保存数据—通过JDBC连接数据库(DataFrame) | 项目实训程序代码 | 1 |
| 19 | 单元19：读取和保存数据—连接Hive读写数据 | 项目实训程序代码 | 2 |
| 20 | 单元20：Spark Streaming基本原理 | 项目实训程序代码 | 1 |
| 21 | 单元21： DStream基本输入源—文件流、套接字流 | 项目实训程序代码 | 2 |
| 22 | 单元22： Dstream转换操作 | 项目实训程序代码 | 2 |
| 23 | 单元23：DStream高级输入源—Apache Kafka | 项目实训程序代码 | 2 |
| 24 | 单元24： DStream窗口计算 | 项目实训程序代码 | 1 |
| 25 | 单元25：Spark MLlib机器学习工作流 | 项目实训程序代码 | 2 |
| 26 | 单元26：特征抽取、转化和选择操作 | 项目实训程序代码 | 2 |
| 27 | 单元27：分类与回归算法 | 项目实训程序代码 | 2 |
| 28 | 单元28：聚类算法 | 项目实训程序代码 | 2 |

**五、课程考核**

**（一）成绩构成**

本课程考核采用形成性考核方式，总评成绩由形成性考核的各项成绩组成，全面考核学生的动手能力、基础理论和平时学习状况。分值比例如下：

综合素质评价（平时表现+通用能力考核） 20%

过程考核（课堂实践考核+课后实践考核+项目开发实践） 50%

终结考核(期终考核) 30%

**（二）评价指标**

1、综合素质评价标准

|  |  |  |
| --- | --- | --- |
| 考核内容 | 评价标准 | 考核分值 |
| 出勤情况 | 迟到、早退1 次扣0.5 分，旷课1 次扣2 分，扣完5 分为止，旷课3 次及以上不得参加终结性考核。 | 5% |
| 学习态度 | 不认真听课、不按时完成课堂或课外实践1 次扣0.5 分，扣完5 分为止。 | 5% |
| 学习表现 | 主动提出问题、回答问题1 次加0.5 分；课堂或课外实践完成质量高（正确率80%以上）1 次加0.5 分，加满5 分为止。 | 5% |
| 通用能力 | 提出问题、回答问题时语言表达准确、逻辑性强1 次加0.5分，加满5 分为止。 | 5% |

2、过程考核标准

|  |  |  |
| --- | --- | --- |
| 考核内容 | 评价标准 | 考核分值 |
| 课堂实践 | 课堂实践完成质量高（正确率80%以上）1 次加0.5 分，加  满15 分为止。 | 15% |
| 课外实践 | 课外实践完成质量高（正确率80%以上）1 次加0.5 分，加  满5 分为止。 | 5% |
| 项目开发实践 | 需求分析0 至3 分、整体设计0 至10 分、界面设计0 至2  分、程序编码0 至10 分、程序测试0 至5 分。 | 30% |

3、终结考核标准

采用笔试闭卷方式，题型包括：选择题、知识填空题、程序填空题、程序阅读题、程序设计题。分值比例：选择题占30%，知识填空题占10%，程序填空题占10%，程序阅读题占20%，程序设计题占30%。

**六、教学实施建议**

**（一）授课教师基本要求**

（1）熟练Spark技术；

（2）熟悉Linux操作系统、大数据平台工具的部署；

（3）具有在Spark平台下Scala语言开发的技能；

（4）课内主讲教师必须具备现场实际工作经历1年以上或实践指导教学2年以上；

（5）具备设计基于行动导向的教学法的设计应用能力。

**（二）实践教学条件基本要求**

表6　课程校内外实践教学条件

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | 实践教学场地名称 | 校内/校外 | 主要实践设备（含软件） |
| 1 | 曙光大数据应用创新中心 | 校内二教109 | Linux开发平台、Spark软件、Scala软件等 |
| 2 | 软件实习工场（一） | 校内二教417 | Linux开发平台、Spark软件、Scala软件等 |
| 3 | 软件实习工场（二） | 校内二教418 | Linux开发平台、Spark软件、Scala软件等 |
| 4 | 广州市扬海科技有限公司大学生校外实践教学基地 | 校外（省级） | Linux开发平台、Spark软件、Scala软件等 |
| 5 | 广州市清源信息科技有限公司大学生校外实践教学基地 | 校外 | Linux开发平台、Spark软件、Scala软件等 |
| 6 | 广州市乐商软件科技有限公司大学生校外实践教学基地 | 校外 | Linux开发平台、Spark软件、Scala软件等 |
| 7 | 广州市天懋信息股份有限公司大学生校外实践教学基地 | 校外 | Linux开发平台、Spark软件、Scala软件等 |
| 8 | 广州慧图信息科技有限公司大学生校外实践教学基地 | 校外 | Linux开发平台、Spark软件、Scala软件等 |
| 9 | 广州凌动数码科技有限公司大学生校外实践教学基地 | 校外 | Linux开发平台、Spark软件、Scala软件等 |

**（三）教材选用与编写**

本课程所用的教材，在内容和结构上必须体现大数据技术概述、Scala语言基础、Spark的设计与运行原理、Spark环境搭建和使用方法、RDD编程、Spark SQL、Spark Streaming、Spark MLlib等知识，并能适应大数据技术专业学生的教学要求。

表7　课程教材选用表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 序号 | 教材名称 | 教材类型 | 出版社 | 主编 | 出版日期 |
| 1 | Spark大数据技术与应用 | 大数据创新人才培养教材 | 人民邮电出版社 | 肖芳 张良均 | 2022.09 |

表8　课程教学参考书选用表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 序号 | 教材名称 | 教材类型 | 出版社 | 主编 | 出版日期 |
| 1 | Spark SQL入门与数据分析实践 | 系列规划教材 | 人民邮电出版社 | 杨虹 | 2021.9 |
| 2 | Spark大数据技术与应用 | 系列规划教材 | 人民邮电出版社 | 赵红艳 | 2019.3 |

**（四）课程数字化教学资源**

表9　课程数字化资源表

|  |  |  |
| --- | --- | --- |
| 序号 | 数字化资源名称 | 资源网址 |
| 1 | Spark课程博客 | http://dblab.xmu.edu.cn/blog/spark/ |
| 2 | Spark教材配套资源 | http://dblab.xmu.edu.cn/post/spark |

附件：

**授课计划表**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 周次 | 教学内容（章节名称、主要知识点） | 课时数 | | | 备注 |
| 理论 | 实操 | 小计 |
| 1 | 单元01：大数据技术概述 | 1 | 1 | 2 |  |
| 1 | 单元02：Scala基础 | 2 | 2 | 4 |  |
| 2 | 单元03：Scala控制结构 | 2 | 2 | 4 |  |
| 2 | 单元04：Scala集合的运用 | 1 | 1 | 2 |  |
| 3 | 单元04：Scala集合的运用 | 1 | 1 | 2 |  |
| 3 | 单元05：Scala类、对象、继承、特质和模式匹配 | 2 | 2 | 4 |  |
| 4 | 单元06：Scala函数式编程开发技巧 | 2 | 2 | 4 |  |
| 4 | 单元07：函数组合器介绍 | 1 | 1 | 2 |  |
| 5 | 单元07：函数组合器介绍 | 1 | 1 | 2 |  |
| 5 | 单元08：Spark运行架构与原理 | 1 | 1 | 2 |  |
| 5 | 单元09：Spark集群搭建 | 1 | 1 | 2 |  |
| 6 | 单元09：Spark集群搭建 | 1 | 1 | 2 |  |
| 6 | 单元10：RDD编程 | 2 | 2 | 4 |  |
| 7 | 单元11：键值对RDD | 2 | 2 | 4 |  |
| 7 | 单元12：共享变量 | 1 | 1 | 2 |  |
| 8 | 单元13：数据读写—文件数据读写 | 2 | 2 | 4 |  |
| 8 | 单元14：RDD持久化与数据分区 | 1 | 1 | 2 |  |
| 9 | 单元15：Spark SQL的基本原理 | 1 | 1 | 2 |  |
| 9 | 单元16：DataFrame的创建 | 2 | 2 | 4 |  |
| 10 | 单元17：读取和保存数据 | 2 | 2 | 4 |  |
| 10 | 单元18：读取和保存数据—通过JDBC连接数据库(DataFrame) | 1 | 1 | 2 |  |
| 11 | 单元19：读取和保存数据—连接Hive读写数据 | 2 | 2 | 4 |  |
| 11 | 单元20：Spark Streaming基本原理 | 1 | 1 | 2 |  |
| 12 | 单元21： DStream基本输入源—文件流、套接字流 | 2 | 2 | 4 |  |
| 12 | 单元22： DStream基本输入源—RDD队列流 | 1 | 1 | 2 |  |
| 13 | 单元22： DStream基本输入源—RDD队列流 | 1 | 1 | 2 |  |
| 13 | 单元23：DStream高级输入源—Apache Kafka、Apache Flume | 1 | 1 | 2 |  |
| 13 | 单元23：DStream高级输入源—Apache Kafka | 1 | 1 | 2 |  |
| 14 | 单元24： DStream窗口计算 | 1 | 1 | 2 |  |
| 14 | 单元25：Spark MLlib机器学习工作流 | 2 | 2 | 4 |  |
| 15 | 单元26：特征抽取、转化和选择操作 | 2 | 2 | 4 |  |
| 15 | 单元27：分类与回归算法 | 1 | 1 | 2 |  |
| 16 | 单元27：分类与回归算法 | 1 | 1 | 2 |  |
| 16 | 单元28：聚类算法 | 2 | 2 | 4 |  |
| 合计 | | 48 | 48 | 96 |  |