《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** | Flink原理与应用 | | | **课程英文名称** | Flink Principles and Applications | | |
| **课程编码** | H35X017D | | | **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、面向对象程序设计 | | |
| **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**

《Flink原理与应用》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。Flink是一种针对数据流的大数据计算引擎。比起已有的大数据计算引擎Hadoop和Spark，Flink集流处理和批量处理于一体，具有高吞吐，低延迟，支持状态和故障恢复等优势，尤其是其流处理技术为业界领先水平，目前被广泛应用在阿里巴巴、腾讯等大公司的生产环境上。通过本课程的学习，学生掌握Flink流数据处理相关技术，具体包括：流处理的基本概念、Flink分布式架构、使用DataStream API进行数据流的转换操作、在时间窗口上处理数据流、使用状态并进行备份以便用于故障恢复、Flink与外部系统进行连接以及Flink部署和配置等问题。课程按照从基础到应用，从基本功能到高级功能的逻辑循序渐进进行讲授，要求学生通过动手实践来掌握Flink数据流开发技能。

**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：**了解大数据开发相关生态和工作流程；掌握流处理的基础概念；掌握编程所必备的继承、多态、泛型和函数式编程等基础知识；掌握Flink数据流图和分布式架构等核心概念；掌握使用不同的方式完成用户自定义函数；掌握Flink的序列化和反序列化机制；掌握DataStream上常用的转换操作；掌握Flink应用的部署和简单配置。 | 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：**学生掌握大数据开发工作流程；学生基于Flink进行流数据处理以及发现问题、分析问题和解决问题的能力。 | 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，培养学生良好的文化修养、职业道德、服务意识和敬业精神培养团队合作和协调沟通能力。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** |
| 大数据技术概述 | 2 | **重点：**了解大数据生态圈常见组件和相应功能；分析哪些应用适合使用流处理的模式；构建一个基于Kafka的实时数据流。  **难点：**构建一个基于Kafka的实时数据流。  **思政元素：**介绍大数据技术的发展历程，大数据技术生态圈的典型事例，激发学生对大数据技术的学习热情。  **教学方法与策略：**线下教学，其次通过课外延伸阅读使学生对基本概念有更深层次对了解。课堂运用主要运用讲授法、课外延伸阅读和实践方法、案例分析法，辅以启发式提问拓宽学生学习思路。 | 课前：了解数据流的构建  课堂：认真听课，做好笔记  课后：及时完成并按时提交作业 | 目标1  目标2  目标3 |
| Flink的设计与运行原理 | 2 | **重点：**Flink的数据流图概念；通过Flink作业提交的过程掌握Flink的核心组件；算子子任务的划分、Slot计算资源的划分；使用Web UI查看并行的算子子任务和Slot信息。  **难点：**Flink的数据流图概念；通过Flink作业提交的过程掌握Flink的核心组件；算子子任务的划分、Slot计算资源的划分。  **教学方法与策略：**线下教学。通过课外延伸阅读使学生对基本概念有更深层次对了解。课堂运用主要运用讲授法、课外延伸阅读、案例分析法，辅以启发式提问拓宽学生学习思路。必要的时候课堂演示操作案例。 | 课前：了解Flink的设计与运行原理  课堂：认真听课，做好笔记  课后：及时完成并按时提交作业 | 目标1  目标2  目标3 |
| DataStream API的使用 | **4** | **重点：**一个Flink程序的组成；使用map()、flatMap()或filter()对数据流进行转换处理；对数据流使用keyBy()进行分组；Flink算子在并行环境下的算子切分和数据流动机制。  **难点：**使用map()、flatMap()或filter()对数据流进行转换处理；对数据流使用keyBy()进行分组；Flink算子在并行环境下的算子切分和数据流动机制。  **教学方法与策略：**线下教学。对于涉及到DataStreamAPI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，并穿插有关数据流知识的讲授。 | 课前：了解DataStream API的使用  课堂：认真听课，做好笔记  课后：及时完成并按时提交作业 | 目标1  目标2  目标3 |
| Flink序列化和反序列化与用户自定义函数 | **2** | **重点：**根据业务场景设计合适的数据结构，且数据结构可被快速序列化；设计Java POJO或Scala case class数据结构；Flink TypeInformation及相关类型推断和序列化机制；在Flink中如何使用Avro、Kryo、Thrift、Protobuf等序列化框架；分别使用接口、Lambda表达式和Rich函数类实现用户自定义函数。  **难点：**Flink TypeInformation及相关类型推断和序列化机制；在Flink中如何使用Avro、Kryo、Thrift、Protobuf等序列化框架。  **教学方法与策略：**线下教学。对于涉及到Flink序列化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，并穿插有关序列化知识的讲授。 | 课前：了解Flink序列化的知识  课堂：认真听课，做好笔记  课后：及时完成并按时提交作业 | 目标1  目标2  目标3 |
| 时间和ProcessFunction函数 | **2** | **重点：**Flink三种时间语义的区别，根据业务场景和需求选择合适的时间语义；Event Time下设置时间戳并生成Watermark的方法；ProcessFunction中Timer的使用方法，能够根据业务场景注册Timer；两个数据流上使用ProcessFunction的方法。  **难点：**Event Time下设置时间戳并生成Watermark的方法；ProcessFunction中Timer的使用方法，能够根据业务场景注册Timer。  **教学方法与策略：**线下教学。对于涉及到时间Timer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Timer的讲授。 | 课前：了解ProcessFunction函数的知识  课堂：认真听课，做好笔记  课后：及时完成并按时提交作业 | 目标1  目标2  目标3 |
| 窗口算子、双流关联和迟到数据的处理 | **2** | **重点：**窗口程序的骨架结构;根据业务场景选择合适的窗口划分方法;根据业务场景选择合适的窗口处理函数;能够使用Join对两个数据流进行操作;掌握迟到数据处理方法。  **难点：**根据业务场景选择合适的窗口划分方法;根据业务场景选择合适的窗口处理函数;能够使用Join对两个数据流进行操作。  **教学方法与策略：**线下教学。对于涉及到窗口算子、双流关联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，并穿插有关窗口算子、双流关联的讲授。 | 课前：了解窗口算子、双流关联的知识  课堂：认真听课，做好笔记  课后：及时完成并按时提交作业 | 目标1  目标2  目标3 |
| 状态和检查点 | **2** | **重点：**流处理中状态的概念以及状态在作业重启和横向扩展时将会遇到的问题；Keyed State和Operator State的使用方法，使用合适的API进行有状态的计算；Flink分布式快照原理；State Backend和Checkpoint的配置方法。  **难点：**Keyed State和Operator State的使用方法，使用合适的API进行有状态的计算；State Backend和Checkpoint的配置方法。  **教学方法与策略：**线下教学。对于涉及到Keyed State和Operator State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Flink分布式快照原理的讲授。 | 课前：了解状态和检查点的知识  课堂：认真听课，做好笔记  课后：及时完成并按时提交作业 | 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 | Flink环境搭建；DataStream API的使用。 | 2 | **重点：**map()、flatMap()或filter()对数据流进行转换处理；对数据流使用keyBy()进行分组。  **难点：**对数据流使用keyBy()进行分组的实践  **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 | 时间和ProcessFunction函数 | 4 | 重点：Event Time下设置时间戳并生成Watermark的方法；ProcessFunction中Timer的使用方法，根据业务场景注册Timer。  难点：Event Time下设置时间戳并生成Watermark  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 | 窗口算子、双流关联和迟到数据的处理 | 4 | 重点：窗口的划分;窗口处理函数的计算过程;Window Join和Interval Join两种双流关联的窗口计算过程。  难点：Window Join和Interval Join两种双流关联的窗口计算过程。  思政元素：要求学生在应用所学操作之外，需自行搜索网络资源如Window Join和Interval Join的知识。培养学生主动探索、不断发展的能力 | 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 | 状态和检查点 | 2 | 重点：Keyed State的使用方法；Operator State的使用方法；Checkpoint相关配置；Savepoint的使用方法  难点： Keyed State的使用方法；Operator State的使用方法。 | 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 | Flink连接器 | 4 | 重点：Source接口的使用方法；Sink接口的使用方法；内置I/O Connector的使用方法；Kafka Connector的使用方法。  难点：内置I/O Connector的使用方法；Kafka Connector的使用方法。 | 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** |
| 大数据技术概述 | 大数据生态圈常见组件和相应功能 | 简答题  填空题  选择题 | 目标1  目标2  目标3 | 5 |
| 分析哪些应用适合使用流处理的模式 |
| Flink的设计与运行原理 | Flink的数据流图概念、Flink的核心组件 | 简答题  填空题  选择题  编程题 | 目标1  目标2  目标3 | 10 |
| 算子子任务的划分、Slot计算资源的划分 |
| DataStream API的使用 | 一个Flink程序的组成 | 简答题  填空题  选择题  编程题 | 目标1  目标2  目标3 | 20 |
| 使用map()、flatMap()或filter()对数据流进行转换处理 |
| 对数据流使用keyBy()进行分组 |
| Flink算子在并行环境下的算子切分和数据流动机制 |
| Flink序列化和反序列化与用户自定义函数 | 设计Java POJO或Scala case class数据结构 | 简答题  填空题  选择题  编程题 | 目标1  目标2  目标3 | 20 |
| Flink TypeInformation及相关类型推断和序列化机制 |
| 分别使用接口、Lambda表达式和Rich函数类实现用户自定义函数。 |
| 时间和ProcessFunction函数 | Flink三种时间语义及Event Time下设置时间戳并生成Watermark的方法 | 简答题  填空题  选择题  编程题 | 目标1  目标2  目标3 | 15 |
| ProcessFunction中Timer的使用方法及两个数据流上使用ProcessFunction的方法 |
| 窗口算子、双流关联和迟到数据的处理 | 根据业务场景选择合适的窗口划分方法根据业务场景选择合适的窗口处理函数 | 简答题  填空题  选择题  编程题 | 目标1  目标2  目标3 | 15 |
| 使用Join对两个数据流进行操作 |
| 迟到数据的处理方法。 |
| 状态和检查点 | 流处理中状态的概念 | 简答题  填空题  选择题  编程题 | 目标1  目标2  目标3 | 15 |
| Keyed State和Operator State的使用方法 |
| Flink分布式快照原理 |

**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] 鲁蔚征. Flink原理与实践[M].北京:人民邮电出版社,2021-03-01.

[2] 罗江宇. Flink设计与实现[M].北京:机械工业出版社,2022-01-01.

**八、参考资料**

[1] 张伟洋. Flink大数据分析实战 [M].北京:清华大学出版社,2022-02-01.

[2] 辛立伟. Flink原理深入与编程实战 [M].北京:清华大学出版社,2023-02-01.

**网络资料**

[1]https://www.icourse163.org/course/XMU-1002335004?from=searchPage&outVendor=zw\_mooc\_pcssjg\_

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