|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 课程编号 | 编制人 | 制定日期 | 修订日期 | 审定组(人) | 审定日期 |
|  |  |  | 2021.1 |  |  |

**《****Flink原理与实践》课程教学大纲**

一、课程基本信息

开课单位：

课程名称：《Flink原理与实践》

课程编号：

课程类型：专业课

学 分： 4

学 时： 48

开设专业： 专业

前导课程：

二、课程任务和目标

**（一）课程任务**

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。

本课程是计算机软件专业的专业拓展课程。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。

**（二）课程目标**

理论上，要求学生掌握Flink数据流开发的基础知识，理解数据流、数据流图、算子、时间、窗口、检查点和数据一致性等概念。技能上，能够使用常见的Flink Java/Scala API进行编程开发，尤其是分组和窗口等相关API；能够使用连接器读取和写入外部系统；能够进行必要的配置并将一个Flink应用部署到集群上。

1、能力目标

（1）培养学生掌握大数据开发工作流程

（2）培养学生基于Flink进行流数据处理的能力

（3）培养学生发现问题、分析问题和解决问题的能力

（4）培养良好的文化修养、职业道德、服务意识和敬业精神

（5）培养团队合作和协调沟通能力

2、知识目标

（1）了解大数据开发相关生态和工作流程

（2）掌握流处理的基础概念

（3）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

（4）掌握Flink数据流图和分布式架构等核心概念

（5）掌握使用不同的方式完成用户自定义函数

（6）掌握Flink的序列化和反序列化机制

（7）掌握DataStream上常用的转换操作

（8）掌握时间以及时间语义

（9）掌握窗口算子的使用方法

（10）掌握状态的使用方法

（11）掌握检查点的原理和常用设置

（12）掌握Flink应用的部署和简单配置

三、教学条件

技术网络机房，在较高配置的类Unix操作系统上组建实验环境，确保系统能够访问互联网。最后一个单元本课程对实验环境要求更高，需要搭建3个主机节点的集群。

建议在虚拟机上进行实验操作，可以充分利用虚拟机的快照功能来备份和恢复实验环境，利用其克隆功能来快速安装系统。

四、教学内容及学时安排

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **序号** | **单元** | **教学内容** | | **教学目标** | **学时** |
| **训练项目** | **必备知识** |
| 1 | 大数据技术概述 | 1．了解大数据生态圈常见组件和相应功能  2．分析哪些应用适合使用流处理的模式  3．构建一个基于Kafka的实时数据流 | 1．大数据分而治之的编程思想  2．批处理和流处理的区别  3．代表性大数据技术  4．流处理基础概念  5．Flink所支持的编程语言 | 了解大数据分而治之的编程思想，弄清批处理与流处理的区别，了解流处理面临的挑战，了解代表性大数据技术，了解Lambda和Kappa大数据处理架构，理解延迟和吞吐、窗口和时间、状态与检查点等流处理基础概念，能够使用Kafka构建数据流 | 4 |
| 2 | 大数据必备编程知识 | 1．掌握继承和多态的区别  2．掌握泛型的应用场景  3．掌握Lambda表达式的使用方法  4．搭建Flink开发环境，能够运行Flink样例程序 | 1．继承和多态  2．泛型  3．函数式编程思想 | 掌握大数据开发所必须的继承、多态、泛型和函数式编程知识，搭建Flink开发环境 | 4 |
| 3 | Flink的设计与运行原理 | 1．掌握Flink的数据流图概念  2．通过Flink作业提交的过程掌握Flink的核心组件  3．熟悉算子子任务的划分、Slot计算资源的划分  4．能够使用Web UI查看并行的算子子任务和Slot信息 | 1．数据流图的逻辑视图和物理执行图  2．Flink分布式架构与核心组件  3．资源划分机制、算子子任务与算子链  4．计算资源中Slot的划分机制 | 通过WordCount样例程序，掌握Flink的数据流图，了解Flink分布式架构与核心组件，了解Flink任务执行与资源划分机制 | 4 |
| 4 | DataStream API的使用 | 1．一个Flink程序包含哪几大部分  2．使用map()、flatMap()或filter()对数据流进行转换处理  3．对数据流使用keyBy()进行分组  4．掌握Flink算子在并行环境下的算子切分和数据流动机制 | 1．Flink骨架结构所包含的几大重要部分  2．单数据流基本转换map()、flatMap()、filter()  3．分组转换keyBy()  4．多数据流转换union()和connect()  5．并行度与数据重分布 | 重点掌握Flink骨架结构所包含的设置执行环境、Source、Transformation、Sink和执行五大部分，掌握map()、flatMap()、filter()、keyBy()的功能和使用方法，掌握并行环境下算子切分和数据流动机制 | 4 |
| 5 | Flink序列化和反序列化与用户自定义函数 | 1．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  2．能够设计Java POJO或Scala case class数据结构  3．了解Flink TypeInformation及相关类型推断和序列化机制  4．了解在Flink中如何使用Avro、Kryo、Thrift、Protobuf等序列化框架  5．分别使用接口、Lambda表达式和Rich函数类实现用户自定义函数 | 1．Flink支持的数据类型，重点掌握元组、Java POJO  2．Flink基于TypeInformation的类型推断和序列化机制  3．Avro、Kryo等序列化框架在Flink中的使用方法  4．三种常用的用户自定义函数方式：接口、Lambda表达式和Rich函数类 | 掌握Flink所支持的数据类型、序列化和反序列化机制，能够设计可被快速序列化的数据结构，掌握常用的用户自定义函数的方法 | 4 |
| 6 | 时间和ProcessFunction函数 | 1．掌握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  2．掌握Event Time下设置时间戳并生成Watermark的方法  3．掌握ProcessFunction中Timer的使用方法，能够根据业务场景注册Timer  4．了解两个数据流上使用ProcessFunction的方法 | 1．Flink的三种时间语义  2．Event Time和Watermark的关系，如何设置时间戳并生成Watermark  3．ProcessFunction系列函数中Timer的使用方法  4. 两个数据流上使用ProcessFunction的方法 | 掌握Flink提供的三种时间语义的区别和应用场景，掌握Event Time下如何设置时间戳并生成Watermark，掌握ProcessFunction系列函数中Timer的使用方法，了解两个数据流上使用ProcessFunction的方法 | 4 |
| 7 | 窗口算子、双流关联和迟到数据的处理 | 1．掌握窗口程序的骨架结构  2．根据业务场景选择合适的窗口划分方法  3．根据业务场景选择合适的窗口处理函数  4．能够使用Join对两个数据流进行操作  5．掌握迟到数据处理方法 | 1．窗口程序骨架结构  2．窗口的划分  3．窗口处理函数的计算过程  4．Window Join和Interval Join两种双流关联的窗口计算过程  5．迟到数据的处理方式 | 掌握常见窗口程序的骨架结构、窗口划分方法、窗口处理函数的计算过程，掌握Window Join和Interval Join两种双流关联方式掌握常见的迟到数据处理方式 | 4 |
| 8 | 状态和检查点 | 1．掌握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  2．掌握Keyed State和Operator State的使用方法，能够使用合适的API进行有状态的计算  3．掌握Flink分布式快照原理  4．掌握State Backend和Checkpoint的配置方法 | 1．状态的概念  2．状态的横向扩展问题  3．Keyed State的使用方法  4．Operator State的使用方法  5．Flink Checkpoint机制的流程和分布式快照原理  6．三种State Backend  7．Checkpoint相关配置  8．Savepoint和Checkpoint的区别，Savepoint的使用方法 | 掌握状态的概念，掌握Keyed State和Operator State的使用方法，掌握Checkpoint的机制的原理、State Backend和Checkpoint相关配置，了解Savepoint与Checkpoint的区别 | 4 |
| 9 | Flink连接器 | 1．掌握端到端Exactly-Once所需要解决的问题和挑战  2．掌握幂等写和事务写的原理  3．掌握Source接口的使用方法，能够结合状态实现数据重发  4．掌握Sink接口的使用方法  5．能够查阅文档，使用合适的Connector，读写到外部系统 | 1．Flink端到端Exactly-Once解决方案：数据重发、幂等写和事务写  2．Source接口使用方法  3．Sink接口使用方法  4．内置I/O Connector使用方法  5．Kafka Connector使用方法 | 掌握数据一致性保障面临的挑战以及Flink端到端Exactly-Once的解决方案，掌握Source和Sink接口，了解常用Connector的使用方法 | 4 |
| 10 | Table API & SQL | 1．掌握Table API & SQL的骨架结构和编程方式  2．掌握基于表的流处理背后的基本原理  3．掌握Flink SQL中时间和窗口计算  4．掌握Flink SQL Join，能够根据业务场景选择合适的Join  5．能够使用SQL函数 | 1．Table API & SQL骨架结构  2．动态表和持续查询  3．Flink SQL中进行时间和窗口计算  4．Flink SQL中的Join  5．流处理SQL DDL  6．SQL函数 | 了解Table API & SQL程序的骨架结构及其与Java/Scala API的区别和联系，了解动态表和持续查询的概念，了解Flink SQL上如何进行时间和窗口计算，了解Flink SQL上如何进行Join，了解Flink SQL中的DDL语句 | 4 |
| 11 | Flink的部署和配置 | 1．能够使用多节点搭建Standalone集群  2．能够在搭建好的YARN集群上提交Flink作业  3．掌握并行度、槽位共享、算子链的设置  4．能够使用命令行工具提交和管理作业 | 1．Standalone集群部署方法  2．YARN集群部署和作业提交流程  3．并行度和槽位划分和算子链的配置  4．内存设置  5．使用命令行工具提交和管理作业  6．Flink与Hadoop集成所进行的设置 | 了解Flink常见的三种集群部署模式，了解配置文件中所涉及的并行度、槽位划分和内存设置，掌握命令行工具的基本使用方法，了解如何将Flink于Hadoop集成 | 4 |
| 12 | 复习与习题解答 |  |  |  | 4 |
| 学 时 合 计 | | | | | 48 |

六、考核方式及评分办法

1、出勤、平时作业：20%

2、理论考试： 40%

3、课程设计：40%

五、教法说明

本课程将采用理论与实践相结合的教学方法。原则上先讲基础和原理知识，再让学生上机操作。课堂上阐述重要的原理和示意图，演示一些重要API的使用方式。课后学生基于课堂上所学的原理，自己动手实践。认真完成书中演示的案例有助于学生掌握Flink编程的。

本课程的重点：DataStream API使用方法、时间和窗口、状态的使用方法、外部连接器Connector。

本课程的难点：多种Flink用户自定义函数的使用方法、分组转换、多数据流转换、窗口上进行Timer编程、状态的使用方法、Checkpoint原理、端到端数据一致性保障、并行度和槽位划分。

七、教材和参考书

教 材：

《Flink原理与实践》 鲁蔚征编著，人民邮电出版社