《分布式计算系统》课程教学大纲

课程代码	DATA0031131017.02	课程性质	专业必修
课程名称：	分布式计算系统		
英文名称	Distributed Computing System		
学时/学分	72	其中实验/实践学时	36
开课单位	数据科学与工程学院	适用专业：	数据科学与大数据技术
先修课程	计算机系统、操作系统、计算机网络原理、当代数据管理系统		
大纲撰写人	徐辰	大纲审核人	钱卫宁
课程网址	https://dasebigdata.github.io	授课语言	中文

注：课程性质选择下列类别之一：学科基础、大类平台、专业必修、专业选修、教师教育

一、课程说明

这门课程是为数据科学与大数据技术专业本科生开设的核心专业必修课。本课程将理论与实践为一体。一方面，本课程选取了典型的分布式计算系统，包括MapReduce、Spark、Flink，统一从设计思想、体系架构、工作原理、容错机制和编程使用五个维度进行剖析，以原理探究为核心，培养学生的思辨和分析能力。另一方面，本课程同步配套实践环节，以实践探索为核心，着力培养学生的动手实践能力。

二、课程目标

本课程的目标是使已经具备编程能力和计算机系统相关知识的学习者，获得关于分布式计算系统的知识，具备运用分布式计算系统处理大规模数据的能力，以适应学习和工作的需要。对学习者的具体目标有：

目标 1：了解分布式计算系统的基本原理与技术。（支撑毕业要求 1、3）

目标 2：理解各类分布式计算系统架构联系与区别，能够阐述系统发展演变的内在逻辑。（支撑毕业要求 3、5）

目标 3：掌握分布式计算系统设计的核心原理，能够结合实践分析比较各类系统的优缺点。（支撑毕业要求 3、4、5）

目标 4：运用分布式计算系统进行编程，能够根据理论知识合理地优化程序设计。（支撑毕业要求 2、4、5）

二、课程目标与毕业要求的对应关系

毕业要求	指标点	课程目标
1.理想信念坚定	具有正确的价值观和道德观，爱国、诚信、守法；	课程目标1
	具有高度的社会责任感和良好的协作精神；	
	具备工科学生所需要的科学精神和人文社会科学素养。	
2.专业技能扎实	掌握工科学生所必须的数学知识；	课程目标4
	掌握数据科学与工程的基础知识，包括相关的计算机、统计与应用数学、信息系统的基础知识；	
	掌握数据分析和机器学习的基本模型和算法。	
3.学科理念先进	深刻理解数据的获取、建模、管理、利用的全生命周期，深刻理解数据科学与工程相关技术发展与社会经济发展的关系；	课程目标1、2、3
	深刻理解数据对于社会经济发展的赋能作用，了解金融、物流、零售、制造等领域的典型应用的技术问题并掌握主要解决方法。	
4.工程能力全面	掌握主要的数据管理和处理工具以及系统平台的使用，熟知它们的特点、系统架构，具备基本的数据系统的设计和开发能力；	课程目标3、4
	了解大数据应用中需求分析、数据和应用建模、系统选型、应用设计、开发和实施的过程，具备合作进行系统和应用研发能力；	
	掌握开源软件的设计和开发方法，掌握云计算平台的使用技术，掌握基于云计算的应用设计、开发、实施、运维方法与技术；	
	具备参与数据系统或数据应用设计、开发、运维工程所需的沟通交流与协作能力，掌握基本的工程管理知识与能力。	

5.研究能力突出	了解“数据科学与工程”学科领域，以及相关应用领域的技术发展前沿； 具有初步的从事数据科学与工程研究工作的科学训练，具有从事相关学科科学研究、教学或工程开发的技术工作的能力。	课程目标2、3、4

三、教学内容与学时安排

第一章 绪论（支撑课程目标 1）

学时：理论 3、实验 4

1. 从数据管理角度看分布式系统

2. 分布式计算系统的内涵与外延

要求学生：初步了解分布式计算系统有关基本概念

实验：熟悉 Linux 基本操作和集群的访问；熟悉上机操作环境；完成编程环境的设置，掌握编写 Java 程序的基本流程。

第二章 Hadoop 文件系统（支撑课程目标 1、2）

学时：理论 3、实验 2

1. HDFS 的设计思想和体系架构

2. HDFS 工作原理（文件读写与访问模型）

3. HDFS 容错机制和文件读写编程

要求学生：理解文件系统与分布式文件系统的联系与区别，了解 HDFS 的架构和工作流程，理解文件访问模型；掌握单机伪分布式部署 Hadoop 的方法，掌握基本的 HDFS 文件访问 API。

实验：学习 Hadoop 的部署，熟悉常用的命令，并且编写程序利用 HDFS 的 API 进行文件读写

第三章 批处理系统 MapReduce ◈（支撑课程目标 4）

学时：理论 8、实验 8

1. MapReduce 设计思想和体系架构

2. MapReduce 工作原理（Map、Shuffle、Reduce 的工作过程）

3. MapReduce 的容错机制

4. MapReduce 基础编程

5. 编程作业评析

要求学生：理解 MapReduce 的体系架构和程序运行过程，了解容错机制；掌握 MapReduce 基本程序设计，并熟练运用。

实验：掌握 Hadoop 2.0 的部署方法；学习使用 Java 语言编写 MapReduce 程序

第四章 批处理系统 Spark ◈（支撑课程目标 3、4）

学时：理论 8、实验 8

1. Spark 设计思想（RDD 数据模型、DAG 计算模型）

2. Spark 体系架构与工作原理

3. Spark 容错机制

4. RDD 基础编程

5. 编程作业评析

要求学生：掌握 Spark 的体系架构，理解 Spark 与 MapReduce 之间的差异，了解 Spark 内部工作流程与容错机制；掌握 Spark 的部署与编程方式，理解各类 API 之间的区别。

实验：Spark 部署、使用 Spark Shell 编程，以及使用 Java 语言编写的程序

第五章 资源管理系统 Yarn（支撑课程目标 1、2）

学时：理论 2、实验 2

1. Yarn 设计思想与工作原理

2. Yarn 应用案例（第二代 MapReduce、基于 Yarn 的 Spark 部署）

要求学生：理解资源管理系统的作用以及平台和框架的概念，了解 Yarn 工作原理。

实验：基于 Yarn 部署 MapReduce 和 Spark，并运用 Yarn 的 UI 进行作业监控

第六章 流计算系统 Flink（支撑课程目标 3、4）

学时：理论 8、实验 10

1. Flink 设计思想（DataStream 数据模型、DAG 计算模型、迭代模型）

2. Flink 体系架构与数据传输方式

3. Flink 状态管理与容错机制

4. DataStream 基础编程

5. 编程作业评析

要求学生：理解 Flink 的设计思想，流计算系统与批处理系统的区别；掌握 Flink 的 DataStream 编程，并且熟练运用。

实验：部署 Flink；并利用 Java 语言进行 DataStream 编程；基于 Yarn 运行 Flink 程序

第七章 总结（支撑课程目标 1、2、3、4）

学时：理论 4、实验 2

1. 课程总结

2. 课程答疑

要求学生：回顾课程内容，从更高层面理解内容之间的逻辑关系

实验：回顾课程实验，总结编程部署的一般方法

四、教学方法

主要教学手段为大班讲授及个别辅导，理论教学方式结合多媒体课件，采用线下教学形式，依托课程网站、水杉在线、钉钉群等进行课程资源发布、作业收集等，课前采用“讨论式”“启发式”检查预习效果和复习巩固。实验教学中的关键的实验操作等提前进行演示录制，随堂对学生操作等进行个别指导，强化学生实验操作技能和规范实验习惯。此外，理论教学过程中，结合实验教学内容，引导学生针对实验现象验证理论内容并进行分析，提升科学思维能力。

五、考核方式

课程考核由平时成绩和期末成绩组成，分别占 50%、50%，详见下表。

	考核方式	占该项比例	占总评比例
平时成绩	考勤	10%	5%
	课堂表现	15%	7.5%
	MapReduce 编程作业	15%	7.5%
	Spark 编程作业	15%	7.5%

	Hadoop 实验	15%	7.5%
	Spark 实验	15%	7.5%
	Flink 实验	15%	7.5%
期末成绩	笔试	100%	50%

课程目标与考核方式对应关系如下：

考核方式 / 课程目标	**考勤**	**课堂表现**	**MapReduce/Spark编程作业**	**Hadoop/Spark/Flink实验**	**笔试**
课程目标1	√	√		√	√
课程目标2		√	√		√
课程目标3		√		√	√
课程目标4		√	√	√	

六、推荐教材和参考资料

（一）推荐教材：

1.《分布式计算系统》，徐辰编著，高等教育出版社，2022 年。

（二）参考资料：

1.《设计数据密集型应用》（影印版），Martin Kleppmann 著，东南大学出版社，2017 年。

2.《大数据处理框架 Apache Spark 设计与实现》，许利杰、方亚芬著，电子工业出版社，2020 年。

3.《大数据计算系统：原理、技术与应用》，王宏志、刘海龙、张立臣、石胜飞编著，机械工业出版社，2023 年。

七、评分标准

课程目标	评分标准				
	90-100	80-89	70-79	60-69	0-59
目标 1：了解分布式计算系统的基本原理与技术。	能够准确、全面地了解分布式计算系统的基本原理与技术	能够准确地了解分布式计算系统的基本原理与技术	能够大致了解分布式计算系统的基本原理与技术	了解一些分布式计算系统的基本原理与技术	完全不了解分布式计算系统的基本原理与技术
目标 2：理解各类分布式计算系统架构联系与区别，能够阐述系统发展演变的内在逻辑。	深刻理解各类分布式计算系统架构联系与区别，熟练阐述、分析系统发展演变的内在逻辑	理解各类分布式计算系统架构联系与区别，较为熟练阐述、分析系统发展演变的内在逻辑	基本能够理解各类分布式计算系统架构联系与区别，可以阐述、分析一些系统发展演变的内在逻辑	能够勉强地理解各类分布式计算系统架构联系与区别，但对系统发展演变的内在逻辑缺乏认识	不能够理解各类分布式计算系统架构联系与区别，对系统发展演变的内在逻辑缺乏认识
目标 3：掌握分布式计算系统设计的核心原理的基础，能够结合实践分析比较各类系统的优缺点。	在完全理解、掌握分布式计算系统设计的核心原理的基础上，熟练结合实践分析比较各类系统的优缺点	在理解、掌握分布式计算系统设计的核心原理的基础上，能够结合实践分析比较各类系统的优缺点	在大致理解、掌握分布式计算系统设计的核心原理，可以结合实践分析比较各类系统的优缺点	大致理解、掌握分布式计算系统设计的核心原理，但无法结合实践分析比较各类系统的优缺点	不能理解分布式计算系统设计的核心原理的基础上，无法结合实践分析比较各类系统的优缺点

目标 4：运用分布式计算系统进行编程，能够根据理论知识合理地优化程序设计。	能熟练地运用分布式计算系统进行编程，能够根据理论知识合理地优化程序设计	能较为熟练地运用分布式计算系统进行编程，能够根据理论知识优化程序设计	能大致运用分布式计算系统进行编程，能够进行一些程序设计的优化	大致能够运用分布式计算系统进行编程，但无法对程序进行优化	不能够运用分布式计算系统进行编程，也无法对程序进行优化