南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

分布式存储与并行计算（STA321）课程大纲

课程详述

COURSE SPECIFICATION

以下课程信息可能根据实际授课需要或在课程检讨之后产生变动。如对课程有任何疑问，请联系授课教师。

The course information as follows may be subject to change, either during the session because of unforeseen circumstances, or following review of the course at the end of the session. Queries about the course should be directed to the course instructor.

1.	**课程名称 Course Title**	分布式存储与并行计算 Distributed storage and parallel computing
2.	**授课院系 Originating Department**	统计与数据科学系 Department of Statistics and Data Science
3.	**课程编号 Course Code**	STA321
4.	**课程学分 Credit Value**	3
5.	**课程类别 Course Type**	专业核心课 Major Core Courses
6.	**授课学期 Semester**	秋季 Spring 2023 年秋季学期开始
7.	**授课语言 Teaching Language**	中英双语 Chinese and English
8.	**授课教师、所属学系、联系方式（如属团队授课，请列明其他授课教师） Instructor(s), Affiliation& Contact （For team teaching, please list all instructors）**	杨鹏，统计与数据科学系. yangp@sustech.edu.cn
9.	**实验员/助教、所属学系、联系方式 Tutor/TA(s), Contact**	待公布 To be announced
10.	**选课人数限额(可不填) Maximum Enrolment （Optional）**	

11. 授课方式 Delivery Method	讲授 Lectures	习题/辅导/讨论 Tutorials	实验/实习 Lab/Practical	其它(请具体注明) Other（Please specify）	总学时 Total
学时数 Credit Hours	32		32		64

12. 先修课程、其它学习要求 Pre-requisites or Other Academic Requirements	计算机程序设计基础 Introduction to Computer Programming （CS109） 数据结构与算法分析 Data Structures and Algorithm Analysis（ CS203）/ 数据结构与算法分析 B Data Structures and Algorithm Analysis（ CS203B）
13. 后续课程、其它学习规划 Courses for which this course is a pre-requisite	
14. 其它要求修读本课程的学系 Cross-listing Dept.	

教学大纲及教学日历 SYLLABUS

15. 教学目标 Course Objectives

本课程在计算机程序设计、数据结构等课程的基础上，围绕大数据处理，让学生了解掌握目前分布式存储和并行计算的模式与框架，初步掌握分布式编程的实践能力。

This course focuses on big data processing on the basis of computer programming, data structure and other courses to make students understand and master the current distributed storage and parallel computing mode and framework, and preliminarily master the practical ability of distributed programming.

16. 预达学习成果 Learning Outcomes

通过本课程的学习，学生预期可达到：

- 了解大数据技术的硬件和软件
- 了解系统体系结构
- 了解大数据整体分析框架及关键实现技术
- 掌握新的编程范式
- 掌握创建高性能集群和分布式编程实践能力
- 了解并行分布式计算技术最新研究进展。

On successful completion of the course, students should be able to：

- Be familiar with the hardware and software of big data management,
- Be familiar with system architecture,
- Be familiar with the overall framework of big data and key implementation technologies,
- Be equipped with the new programming paradigm,
- Be equipped with the ability to create high-performance clusters and distributed programming practices,
- Be aware of the latest research progress of parallel distributed computing technology.

17. 课程内容及教学日历 （如授课语言以英文为主，则课程内容介绍可以用英文；如团队教学或模块教学，教学日历须注明主讲人）
Course Contents (in Parts/Chapters/Sections/Weeks. Please notify name of instructor for course section(s), if this is a team teaching or module course.)

理论课教学大纲

第 1 章 大数据对分布式存储与并行计算的需求（2 学时）

大数据基本概念，基础架构、现状及发展、应用前景

Chapter 1 Introduction to Big Data Processing (2 hours)

Basic concepts of big data, infrastructure, current situation and development, application prospects

第 2 章 基于 Hadoop 的分布式处理框架 （4 学时）

Hadoop 基本概念，基础架构，以及相关的技术和应用、发展现状

Chapter 2 Big data processing framework based on Hadoop (4 hours)

Hadoop basic concepts, infrastructure, related technologies, applications, and development status

第 3 章 MapReduce 并行计算模式 （4 学时）

MapReduce 工作机制，MapReduce 的负载均衡和容错机制，基于 MapReduce 的并行算法设计

Chapter 3 MapReduce computing patterns (4 hours)

MapReduce work mechanism, MapReduce load balancing and fault tolerance mechanism, parallel algorithm design based on MapReduce

第 4 章 使用 HDFS 分布式存储大数据 （4 学时）

HDFS 架构和流程，HDFS 的访问与控制机制

Chapter 4 Use HDFS to store big data (4 hours)

HDFS architecture and process, HDFS access and control mechanism

第 5 章 HBase 分布式数据库（4 学时）

HBase 架构与原理，HBase 性能优化

Chapter 5 HBase database (4 hours)

HBase architecture and principle, HBase performance optimization

第 6 章 大数据并行分析处理（4 学时）

Hive 和 Pig 编程机制和原理

Chapter 6 Big data analysis and processing (4 hours)

Hive and Pig programming mechanisms and principles

第 7 章 基于 MapReduce 的分布式数据挖掘（4 学时）

基于 MapReduce 的数据挖掘算法编程

Chapter 7 Data mining based on MapReduce (4 hours)

Data mining algorithm programming based on MapReduce

第 8 章 Hadoop 分布式集群的管理与维护机制（4 学时）

ZooKeeper 管理机制和基于 Kerberos 的 Hadoop 管理机制

Chapter 8 Hadoop cluster management and maintenance (4 hours)

ZooKeeper management mechanism and Kerberos-based Hadoop management mechanism

第 9 章 代表性大数据并行计算框架简介（2 学时）

Spark、Tensorflow、Ray

Chapter 9 Typical Big Data Parallel Computing Framework (2 hours)

Spark、Tensorflow、Ray

实验课教学大纲

实验 1 （2 学时）

- Hadoop 系统安装及环境配置

Lab 1 (2 hours)

- Installing and Configuration of Hadoop system

实验 2 （4 学时）

- 熟悉 Hadoop 系统下的典型命令行操作

Lab 2 (4 hours)

- Be familiar with typical commands in Hadoop system

实验 3 (4 学时)

- 基于 MapReduce 基本 API 的并行计算编程练习

Lab 3 (4 hours)

- Parallel programming with basic APIs of MapReduce

实验 4（4 学时）

- HDFS 分布式存储基本访问与控制指令练习

Lab 4 (4 hours)

- Basic commands of accessing and controlling HDFS

实验 5 （4 学时）

- HBase 安装及环境配置
- HBase 分布式数据库查询语句编程练习

Lab 6 (4 hours)

- Installing and Configuration of HBase
- Learning Querying language of HBase

实验 6（4 学时）

- Hive 安装及环境配置
- Hive 编程 API 学习与大数据并行计算练习

Lab 6 (4 hours)

- Installing and Configuration of Hive
- Learning programing language of HBase for large-scale data

实验 7（4 学时）

- 基于 MapReduce 机器学习 API 接口的智能数据分析编程练习

Lab 7 (4 hours)

- Learning machine learning APIs in MapReduce for intelligent data analysis

实验 8（2 学时）

- Zookeeper 安装与环境配置
- Zookeeper 基本操作指令练习

Lab 8 (2 hours)

- Installing and Configuration of Zookeeper
- Learning basic commands of Zookeeper

实验 9 （4 学时）

- 课程项目指导
- 课程项目答辩展示

Lab 9 (4 hours)

- Tutorials for the course project
- Project presentation

18. 教材及其它参考资料 Textbook and Supplementary Readings

Textbook:
《Hadoop: The Definitive Guide》3rd Edition，Tom White，O'Reilly Media，2012
《Hadoop 大数据处理》，刘军编著，人民邮电出版社，2013 年

Supplementary Readings:
《Hadoop 权威指南》第三版，Tom White 编著，清华大学出版社，2015 年
Distributed Systems http://code.google.com/edu/parallel/index.html

课程评估 ASSESSMENT

19.

评估形式 Type of Assessment	评估时间 Time	占考试总成绩百分比 % of final score	违纪处罚 Penalty	备注 Notes
出勤 Attendance		10		
课堂表现 Class				

Performance				
小测验 **Quiz**				
课程项目 **Projects**				
平时作业 **Assignments**		20		
期中考试 **Mid-Term Test**		30		
期末考试 **Final Exam**				
期末报告 **Final Presentation**		40		
其它（可根据需要改写以上评估方式） **Others (The above may be modified as necessary)**				

20. 记分方式 **GRADING SYSTEM**

☑ A. 十三级等级制 **Letter Grading**
☐ B. 二级记分制（通过/不通过） **Pass/Fail Grading**

课程审批 REVIEW AND APPROVAL

21. 本课程设置已经过以下责任人/委员会审议通过
This Course has been approved by the following person or committee of authority

课程详述

COURSE SPECIFICATION

以下课程信息可能根据实际授课需要或在课程检讨之后产生变动。如对课程有任何疑问，请联系授课教师。

The course information as follows may be subject to change, either during the session because of unforeseen circumstances, or following review of the course at the end of the session. Queries about the course should be directed to the course instructor.

1.	**课程名称 Course Title**	分布式存储与并行计算 Distributed storage and parallel computing
2.	**授课院系 Originating Department**	统计与数据科学系 Department of Statistics and Data Science
3.	**课程编号 Course Code**	STA321
4.	**课程学分 Credit Value**	3
5.	**课程类别 Course Type**	专业核心课 Major Core Courses
6.	**授课学期 Semester**	秋季 Spring
7.	**授课语言 Teaching Language**	中英双语 English & Chinese
8.	**授课教师、所属学系、联系方式（如属团队授课，请列明其他授课教师） Instructor(s), Affiliation& Contact （For team teaching, please list all instructors）**	胡延庆, 统计与数据科学系. huyq@sustech.edu.cn
9.	**实验员/助教、所属学系、联系方式 Tutor/TA(s), Contact**	待公布 To be announced
10.	**选课人数限额(可不填) Maximum Enrolment （Optional）**	

11. 授课方式 Delivery Method	讲授 Lectures	习题/辅导/讨论 Tutorials	实验/实习 Lab/Practical	其它(请具体注明) Other（Please specify）	总学时 Total
学时数 Credit Hours	48				48

12. 先修课程、其它学习要求 Pre-requisites or Other Academic Requirements	计算机程序设计基础 Introduction to Computer Programming （CS102） 数据结构与算法分析 Data Structures and Algorithm Analysis（ CS203）
13. 后续课程、其它学习规划 Courses for which this course is a pre-requisite	
14. 其它要求修读本课程的学系 Cross-listing Dept.	

教学大纲及教学日历 SYLLABUS

15. 教学目标 Course Objectives

本课程在计算机程序设计、数据结构等课程的基础上，围绕大数据处理，让学生了解掌握目前分布式存储和并行计算的模式与框架，初步掌握分布式编程的实践能力。

This course focuses on big data processing on the basis of computer programming, data structure and other courses to make students understand and master the current distributed storage and parallel computing mode and framework, and preliminarily master the practical ability of distributed programming.

16. 预达学习成果 Learning Outcomes

通过本课程的学习，学生预期可达到：

- 了解大数据管理的硬件和软件、系统体系结构、新的编程范式，以及并行分布式计算技术最新研究进展。
- 了解云计算的整体框架及关键实现技术、业务模式，掌握创建高性能集群和分布式编程实践能力。

On successful completion of the course, students should be able to:

- Be familiar with the hardware and software of big data management, system architecture, new programming paradigms, and the latest research progress of parallel distributed computing technology.

Understand the overall framework of cloud computing, key implementation technologies and business models, and master the ability to create high-performance clusters and distributed programming practices.

17. 课程内容及教学日历 （如授课语言以英文为主，则课程内容介绍可以用英文；如团队教学或模块教学，教学日历须注明主讲人）

Course Contents (in Parts/Chapters/Sections/Weeks. Please notify name of instructor for course section(s), if this is a team teaching or module course.)

First Part: Introduction to distributed storage and parallel computing [4 hours]

> Basic concepts, motivations, current situation and development, application prospects [2 hours]

> Examples of parallel computing using Linux and Python [2 hours]

Second Part: Methods for distributed storage and parallel computing [18 hours]

> Infrastructures of distributed systems [2 hours]

> Map and Reduce for parallel computing [4 hours]

> Workload balance and scheduling [2 hours]

> Communication and synchronisation in parallel computing [4 hours]

> Transactions and locks [2 hours]

> Fault-tolerance, Byzantine fault, and Paxos/RAFT protocols [2 hours]

> Distributed file system for distributed storage (e.g. HDFS) [2 hours]

Third Part: Parallel computing in practice [14 hours]

> Data processing in multi-threads/multi-processes [4 hours]

Data crawling, cleaning, preprocessing [2 hours]

Experiments [2 hours]

> Hadoop and PySpark [4 hours]

Basic concepts and usages of Hadoop and PySpark [2 hours]

Experiments [2 hours]

> Machine learning with multiple GPUs [6 hours]

Clustering, regression, classification, collaborative filter [4 hours]

Experiments [2 hours]

Fourth Part: Distributed storage and parallel computing in the future [6 hours]

> Training LARGE neural networks : data parallelism, model parallelism and beyond [2 hours]

> Blockchain -- decentralized distributed system [4 hours]

18. 教材及其它参考资料 **Textbook and Supplementary Readings**

Textbook:
- Tomasz Drabas, Denny Lee. Learning PySpark: Build data-intensive applications locally and deploy at scale using the combined powers of Python and Spark 2.0. Packt Publishing. 2017. Available in https://k-state.instructure.com/files/7013786/download?download_frd=1
- Zaccone, Giancarlo. Python parallel programming cookbook. Packt Publishing Ltd, 2015. https://docs.google.com/viewer?a=v&pid=sites&srcid=b2JqZWN0bWFnZS5jb218cHJpdmF0ZS10cmFpbmluZ3xneDoyZjU2M2U4NGJiN2M0NWU2

Supplementary Readings:
- PySpark tutorial: https://sparkbyexamples.com/pyspark-tutorial/
- Parallel computing. Stanford CS128, Fall 2021. https://gfxcourses.stanford.edu/cs149/fall21

- Wenqiang Feng. Learning Apache Spark with Python. https://runawayhorse001.github.io/LearningApacheSpark/pyspark.pdf or https://runawayhorse001.github.io/LearningApacheSpark/

课程评估 ASSESSMENT

19.

评估形式 Type of Assessment	评估时间 Time	占考试总成绩百分比 % of final score	违纪处罚 Penalty	备注 Notes
出勤 Attendance				
课堂表现 Class Performance				
小测验 Quiz				
课程项目 Projects				
平时作业 Assignments		25		
期中考试 Mid-Term Test				
期末考试 Final Exam		50		
期末报告 Final Presentation		25		
其它（可根据需要改写以上评估方式） Others (The above may be modified as necessary)				

20. 记分方式 **GRADING SYSTEM**

☑ A. 十三级等级制 **Letter Grading**

课程审批 REVIEW AND APPROVAL

21. **本课程设置已经过以下责任人/委员会审议通过**
This Course has been approved by the following person or committee of authority
