北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY

课程介绍

Course Overview

大数据处理技术

计算机学院

课程基本信息

- **课程编号：** 100081062
- **课程名称：** 大数据处理技术
- **课程总学时：** 32，其中包括实验/上机8学时
- **课程学分：** 2
- **上课时间/地点：** 9-16周
  - 星期二　0809节　　文萃楼F102
  - 星期五　0607节　　文萃楼F102
- **授课教师：** 张美慧
  - 办公地点：中关村校区 中心教学楼 1211
  - 邮箱：meihui_zhang@bit.edu.cn

授课教师

张美慧，**教授，博士生导师**。毕业于**新加坡国立大学**，入选2018年第十四批**“国家海外高层次人才引进计划”青年项目**。凭借在大数据关联关系挖掘和融合分析方面做出的突出贡献，以及在医疗领域的成功应用，获**2019 年度CCF-IEEE CS青年科学家奖、2020年度VLDB Early Career Research Contribution Award（亚洲仅两位）及2025年度CCFF科技成果奖自然科学二等奖**。课题组致力于发表**国际高水平研究成果**，主要研究方向为**大数据、人工智能、区块链**等，具体包括大数据管理与分析、新型数据库系统、区块链数据管理、知识图谱、医疗+人工智能等方面的研究。课题组**项目经费充足**，目前承担国家自然科学基金联合基金项目、面上项目、原创项目、科技部重点研发课题、CCF-蚂蚁科研基金、CCF-华为胡杨林基金等。与国内多家知名医院合作联合开展医疗+大数据+人工智能方向合作研究，同时与蚂蚁金服、OceanBase、华为等头部企业开展区块链、智能数据管理等方向的合作研究。

课程定位

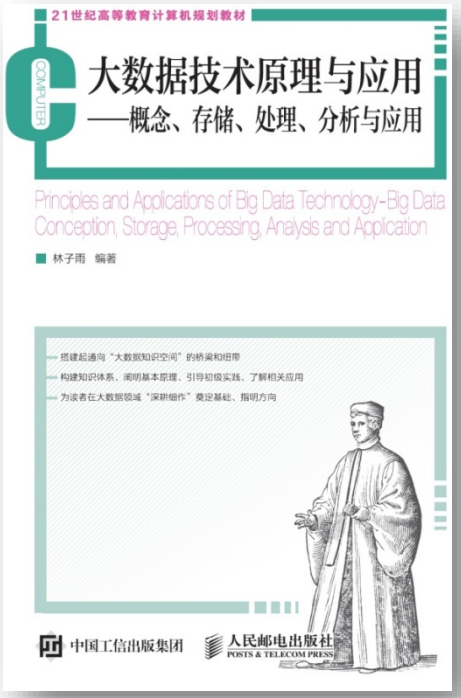

BIG
DATA
STORAGE
PROCESSING
VOLUME
VARIETY
VELOCITY
INFORMATION
SHARING
DECISION
SOFTWARE
BUSINESS
SPEED
OPTIMIZATION
RELATIONAL
PETABYTES
AMOUNT
SOURCES
SYSTEMS
DATABASE
DIFFICULT
CHALLENGES
SEARCH
ANALYSIS
CAPTURE
COLLECTION
TOOLS
TRANSFER
DATA
STREAM
ENABLE
TRIGGER
ENHANCED
LOGS
TYPES
DENSITY
AREAS
MAKING
PROCESS
CURATE
STATISTICS
MANAGE
RANGE
EXPLORE
APPLICATIONS
SIMULATION
大数据之门

教材介绍

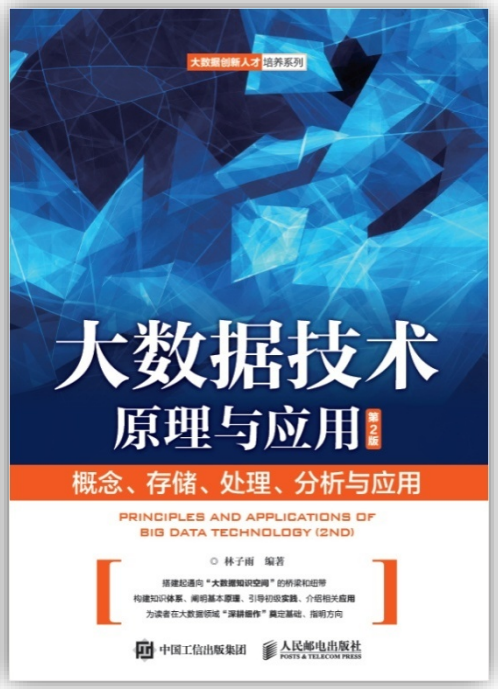
《大数据技术原理与应用——大数据概念、存储、处理、分析与应用》

厦门大学 林子雨编著　人民邮电出版社　2017年2月第2版
ISBN:978-7-115-44330-4

- 国内高校第一本系统介绍大数据知识专业教材
- 京东、当当等各大网店畅销书籍
- 致力于打造成为大数据入门教材精品
- 工信部“全国云计算与大数据应用技术人才考试、认证项目”目前唯一指定大数据教材



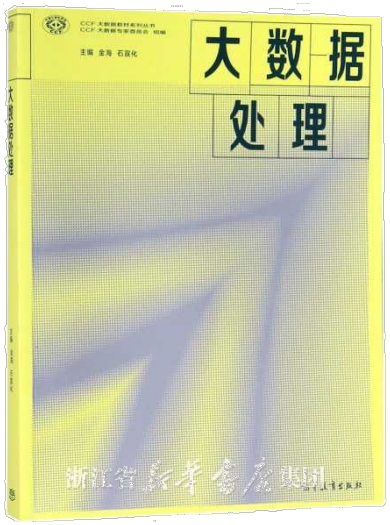
2015年8月第1版



2017年2月第2版

课外选读

- 大数据导论. 梅宏, CCF大数据教材系列丛书. 高等教育出版社. 2018
- 大数据处理. 金海、石宣化, CCF大数据教材系列丛书. 高等教育出版社. 2018



课外选读

http://www.mmds.org/

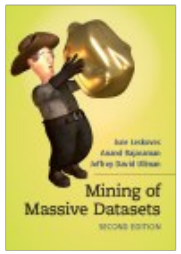
Mining of Massive Datasets

Jure Leskovec, Anand Rajaraman, Jeff Ullman

- Home
- Book & Slides
- MOOC
- Stanford Courses
- Supporting Materials

Big-data is transforming the world. Here you will learn data mining and machine learning techniques to process large datasets and extract valuable knowledge from them.

The book



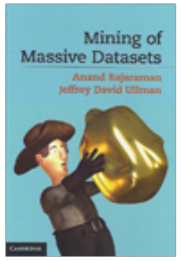
The book is based on Stanford Computer Science course CS246: Mining Massive Datasets (and CS345A: Data Mining).

The book, like the course, is designed at the undergraduate computer science level with no formal prerequisites. To support deeper explorations, most of the chapters are supplemented with further reading references.

The Mining of Massive Datasets book has been published by Cambridge University Press. You can get a 20% discount by applying the code MMDS20 at checkout.

By agreement with the publisher, you can download the book for free from this page. Cambridge University Press does, however, retain copyright on the work, and we expect that you will obtain their permission and acknowledge our authorship if you republish parts or all of it.

We welcome your feedback on the manuscript.



The MOOC (Massive Open Online Course)

We are running the third edition of an online course based on the Mining Massive Datases book:

Mining Massive Datasets MOOC

The course starts September 12 2015 and will run for 9 weeks with 7 weeks of lectures. Additional information and registration.

教学内容

- 本课程系统介绍了大数据存储+处理相关知识
- 大数据的基本概念、大数据处理架构Hadoop、分布式文件系统HDFS、分布式数据库HBase、分布式并行编程模型MapReduce、NoSQL数据库、基于内存的大数据处理架构Spark及流计算和图计算系统
- 在Hadoop、HDFS、HBase、MapReduce等重要章节，安排了实践操作课程（实验/作业），以更好地学习和掌握大数据关键技术

进度安排

第12周		
11.11 星期二	第1讲	大数据概述
11.14 星期五	第2讲	Hadoop
第13周		
11.18 星期二	**实验1**	**Hadoop安装与使用**
11.21 星期五	第3讲	HDFS
第14周		
11.25 星期二	**实验2**	**HDFS操作**
11.28 星期五	第4讲	HBase
第15周		
12.02 星期二	**实验3**	**HBase操作**
12.05 星期五	第5讲	MapReduce (1)
第16周		
12.09 星期二	第6讲	MapReduce (2) + Hadoop优化与发展
12.12 星期五	**实验4**	**MapReduce编程实践**
第17周		
12.16 星期二	第7讲	NoSQL
12.19 星期五	**大作业中期检查（报告）**	
第18周		
12.23 星期二	第8讲	Spark
12.16 星期五	第9讲	流计算
第19周		
12.30 星期二	第10讲	图计算
01.02 星期五	**大作业终期检查（汇报+系统演示）**	

课程设置

- 课件下载及作业及实验提交：
  - 乐学 https://lexue.bit.edu.cn/
  - 课程链接：https://lexue.bit.edu.cn/course/view.php?id=17619
  - 课程名称：大数据处理技术（2025秋）
  - 选课密码：bigdata2025

考核与成绩评定

- 考核方式：考试
- 成绩构成：
  - **课堂表现　　5%**
  - **实验（4次）　　20%**
  - **课程设计　　35%**
    - 汇报演示　　15%
    - 报告　　10%
    - 代码　　10%
  - **期末考试　　40%**

课程设计

- 组队：1-2人/组
- 基本要求：
  - 使用大数据处理框架（不局限于课上内容）
  - 不限开发语言
  - 不限应用（推荐系统、预测分析 ...）
- 关于数据
  - Kaggle https://www.kaggle.com/datasets
  - DF竞赛平台 https://www.datafountain.cn/dataSets

关于数据

- Amazon Web Services (AWS) datasets
  https://registry.opendata.aws/
  - Common Crawl: data from a crawl of over 5 billion web pages
  - COVID-19 Open Research Dataset (CORD-19): full-text and metadata dataset of COVID-19 and coronavirus-related research articles
- Google Public Data sets
  https://cloud.google.com/bigquery/public-data
  - USA Names: contains all Social Security name applications in the US, from 1879 to 2015
  - Github Activity: contains all public activity on over 2.8 million public Github repositories
  - Historical Weather: data from 9000 NOAA weather stations from 1929 to 2016

Google Cloud Platform

课程设计

报告内容：

- 课程设计题目、背景意义
- 组号、小组成员、分工
- 设计内容、主要功能
- 开发工具
- 平台/架构
- 具体实现
- 系统演示、测试结果等
- 实现难点分析等