大数据技术原理与应用

0. 大数据课程介绍

陈建文
电子信息与通信学院
chenjw@hust.edu.cn

0. 课程介绍

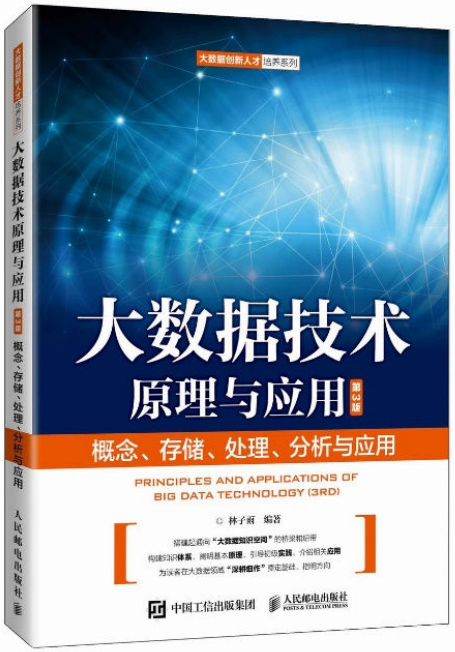
0.1 课程特色

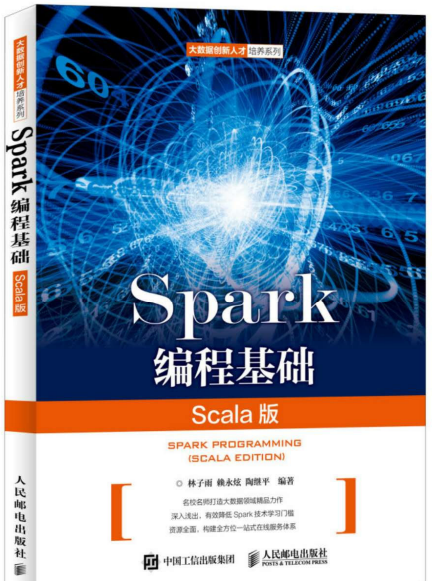
0.2 课程内容

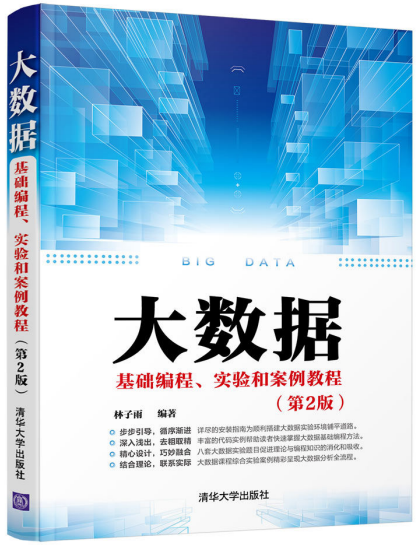
0.3 教材介绍

0.4 电子资源

0.5 考核方式







0.1 课程特色

搭建起通向“大数据知识空间”的桥梁和纽带

构建知识体系、阐明基本原理
引导初级实践、了解相关应用

为学生在大数据领域“深耕细作”奠定基础、指明方向

0.2 课程内容

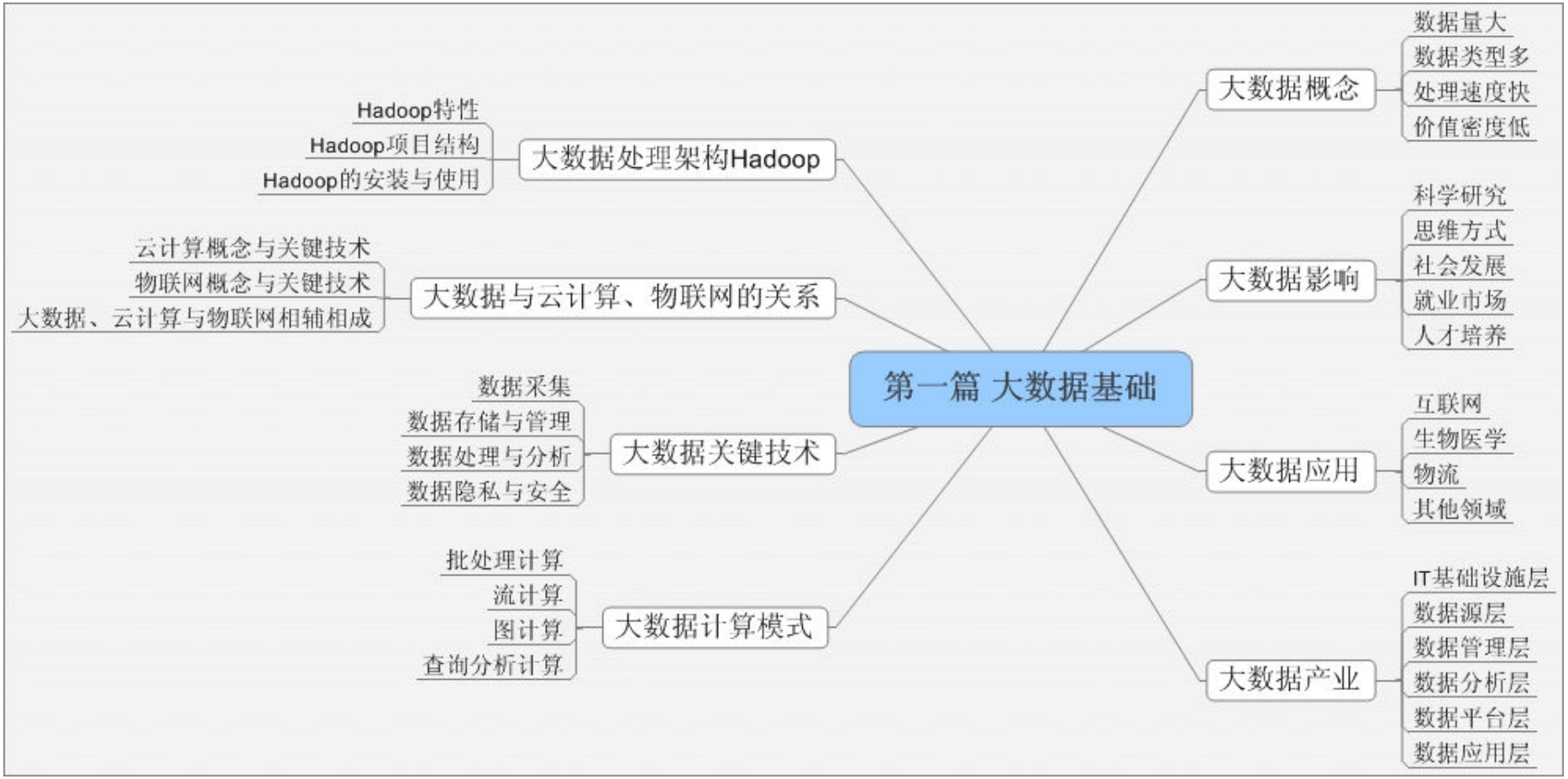
第一篇：大数据基础
第二篇：大数据存储
第三篇：大数据处理与分析
第四篇：大数据应用

- 本课程系统介绍了大数据相关知识；
- 本课程系统地描述了大数据的基本概念、大数据处理架构Hadoop、分布式文件系统 HDFS、分布式数据库 HBase、NoSQL数据库、云数据库、分布式并行编程模型 MapReduce、数据仓库Hive、基于内存的分布式计算框架Spark、流计算、流处理框架Flink、图计算、数据可视化以及大数据在互联网、生物医学和物流等各个领域的应用。
- 在Hadoop、HDFS、HBase、MapReduce、Hive、Spark和Flink等重要章节，安排了入门级的实践操作，更好地学习和掌握大数据关键技术。

第一篇：大数据基础

本篇内容介绍大数据（Big Data）的基本概念、影响和应用领域，并阐述大数据、云计算和物联网的相互关系，同时还将介绍大数据处理架构Hadoop。由于Hadoop已经成为应用最为广泛的大数据技术，因此，本书的大数据相关技术主要围绕Hadoop展开，包括Hadoop MapReduce、HDFS和HBase。本篇内容是理解后续其他篇章内容的基础。

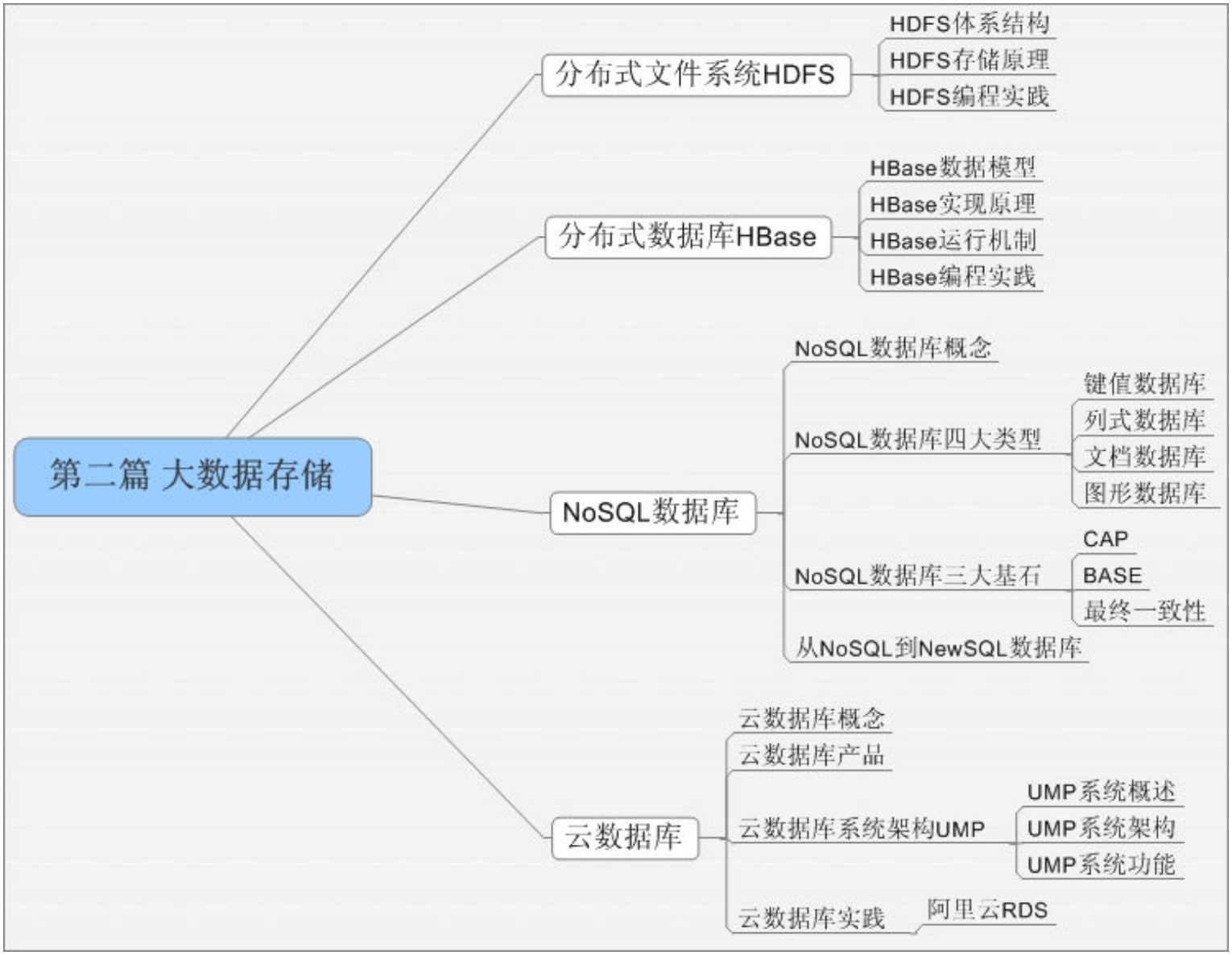
本篇包括2章。第一章介绍大数据的概念和应用，分析了大数据、云计算和物联网的相互关系；第二章介绍大数据处理架构Hadoop。


第一篇 大数据基础
大数据概念
数据量大
数据类型多
处理速度快
价值密度低
大数据影响
科学研究
思维方式
社会发展
就业市场
人才培养
大数据应用
互联网
生物医学
物流
其他领域
大数据产业
IT基础设施层
数据源层
数据管理层
数据分析层
数据平台层
数据应用层
大数据处理架构Hadoop
Hadoop特性
Hadoop项目结构
Hadoop的安装与使用
大数据与云计算、物联网的关系
云计算概念与关键技术
物联网概念与关键技术
大数据、云计算与物联网相辅相成
大数据关键技术
数据采集
数据存储与管理
数据处理与分析
数据隐私与安全
大数据计算模式
批处理计算
流计算
图计算
查询分析计算

第二篇：大数据存储

本篇介绍大数据存储相关技术的概念与原理，包括分布式文件系统HDFS、分布式数据库HBase、NoSQL数据库和云数据库。HDFS提供了在廉价服务器集群中进行大规模分布式文件存储的能力。HBase是一个高可靠、高性能、面向列、可伸缩的分布式数据库，主要用来存储非结构化和半结构化的松散数据。NoSQL数据库可以支持超大规模数据存储，灵活的数据模型可以很好地支持Web2.0应用，具有强大的横向扩展能力，可以有效弥补传统关系型数据库的不足。云数据库是部署和虚拟化在云计算环境中的数据库，可以将用户从繁琐的数据库硬件定制中解放出来，同时让用户拥有强大的数据库扩展能力，满足各种不同类型用户的数据存储需求。

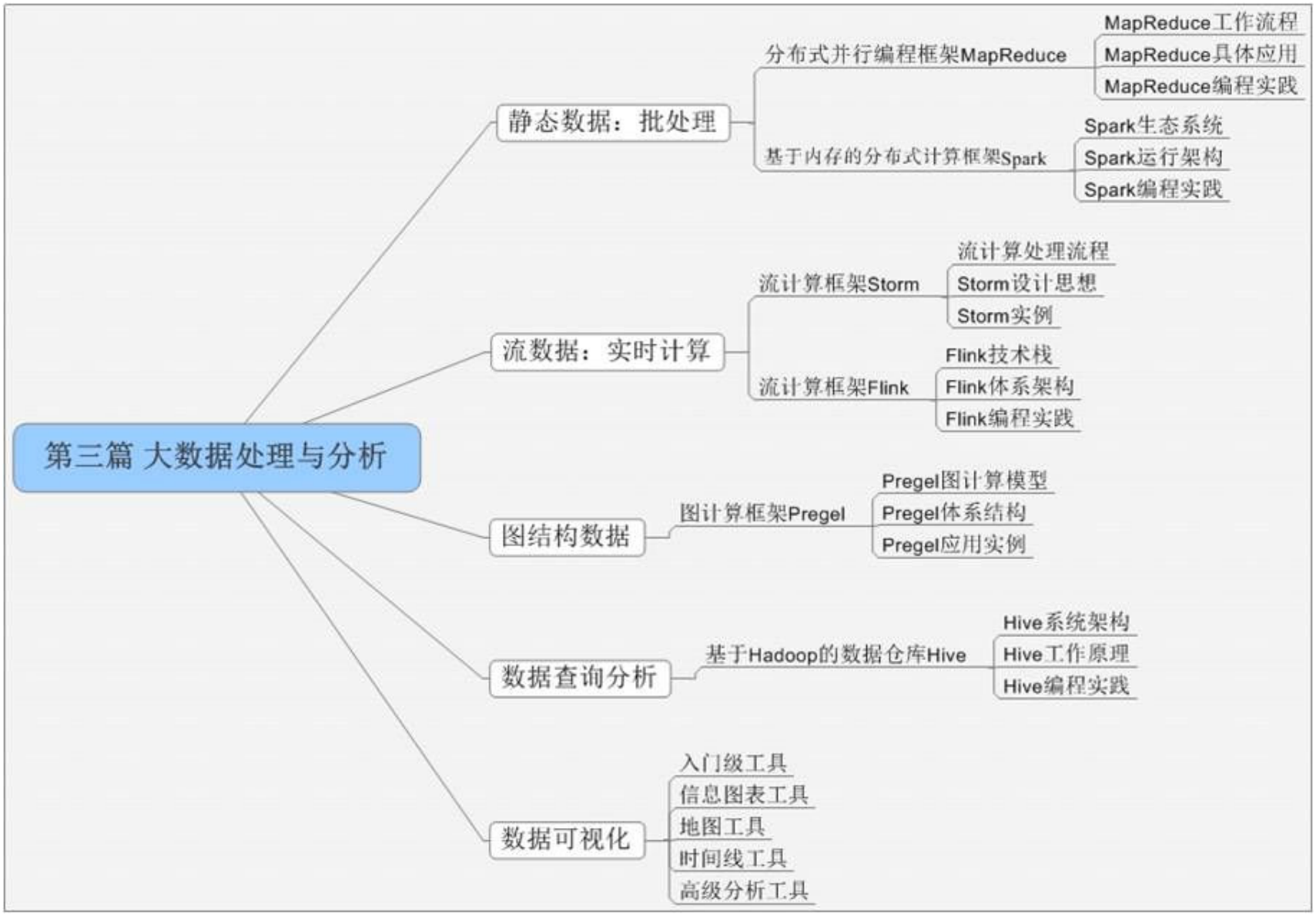
本篇包括4章。第三章介绍分布式文件系统HDFS；第四章介绍分布式数据库HBase；第五章介绍NoSQL数据库；第六章介绍云数据库。


第二篇 大数据存储
分布式文件系统HDFS
HDFS体系结构
HDFS存储原理
HDFS编程实践
分布式数据库HBase
HBase数据模型
HBase实现原理
HBase运行机制
HBase编程实践
NoSQL数据库
NoSQL数据库概念
NoSQL数据库四大类型
键值数据库
列式数据库
文档数据库
图形数据库
NoSQL数据库三大基石
CAP
BASE
最终一致性
从NoSQL到NewSQL数据库
云数据库
云数据库概念
云数据库产品
云数据库系统架构UMP
UMP系统概述
UMP系统架构
UMP系统功能
云数据库实践
阿里云RDS

第三篇：大数据处理与分析

本篇介绍大数据处理与分析的相关技术。大数据包括静态数据和动态数据（流数据），静态数据适合采用批处理方式，动态数据需要进行实时计算。分布式并行编程框架MapReduce实现高效的批量数据处理。Hive是一个基于Hadoop的数据仓库工具，用户通过编写类似SQL的HiveQL语句就可以运MapReduce任务，不必编写复杂的MapReduce应用程序。基于内存的分布式计算框架Spark，是一个可应用于大规模数据处理的快速、通用引擎，成为当今大数据领域最热门的大数据计算平台。流计算框架Storm是一个低延迟、可扩展、高可靠的处理引擎，可以有效解决流数据的实时计算问题。Flink是一种具有代表性的开源流处理架构，具有十分强大的功能，同时支持批处理和流处理。大数据中包括很多图结构数据，Pregel就是其中一种具有代表性的产品。本篇还简要介绍了数据可视化的概念和相关工具。

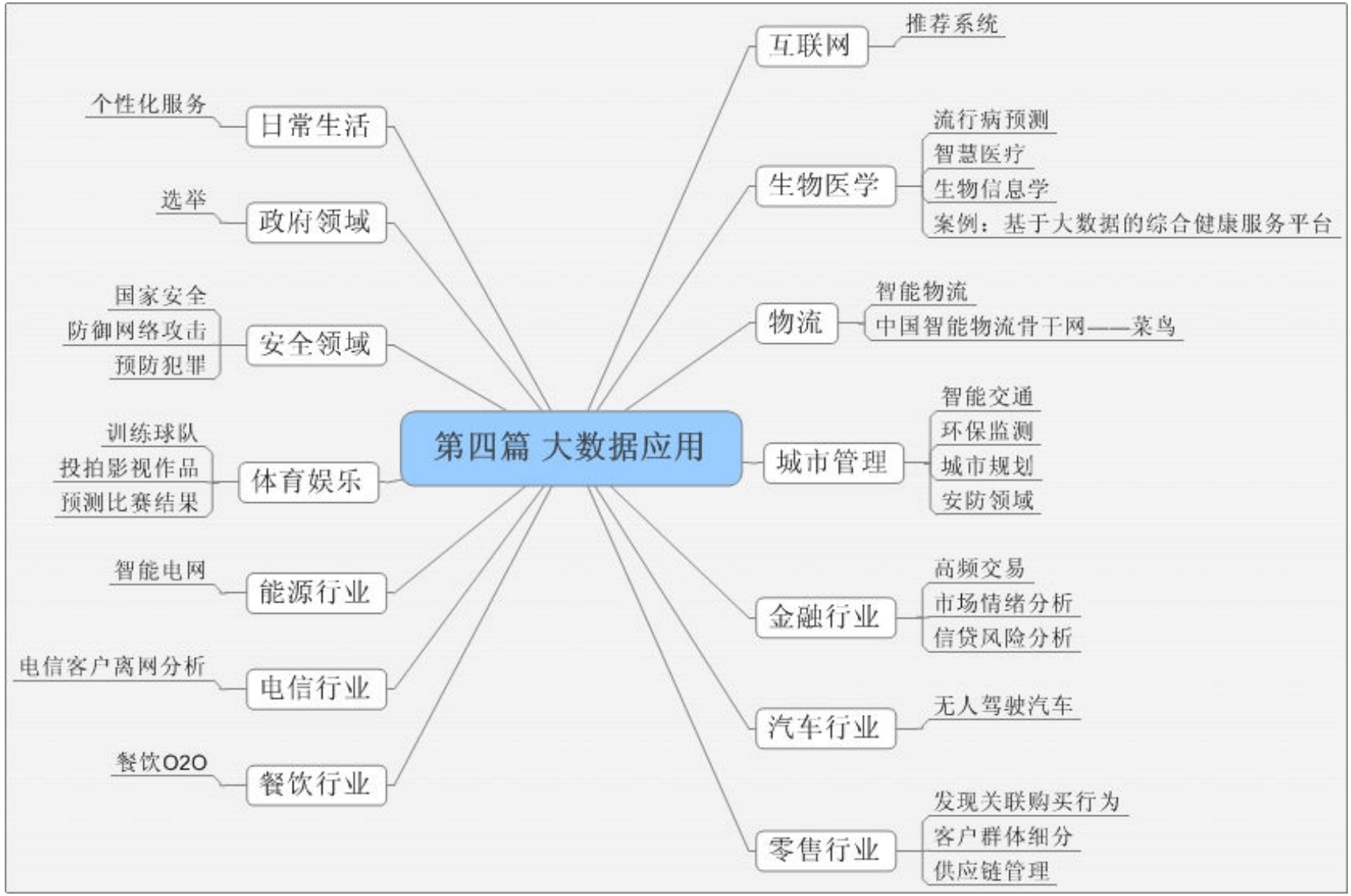
本篇包括八章。第7章介绍分布式并行编程框架MapReduce；第8章对Hadoop进行了再探讨；第9章介绍基于Hadoop的数据仓库Hive；第10章介绍基于内存的分布式计算框架Spark；第11章介绍流计算和开源流计算框架Storm；第12章介绍开源流处理框架Flink；第13章介绍图计算框架Pregel；第14章简要介绍数据可视化的概念和相关工具。


第三篇 大数据处理与分析
静态数据：批处理
分布式并行编程框架MapReduce
MapReduce工作流程
MapReduce具体应用
MapReduce编程实践
基于内存的分布式计算框架Spark
Spark生态系统
Spark运行架构
Spark编程实践
流数据：实时计算
流计算框架Storm
流计算处理流程
Storm设计思想
Storm实例
流计算框架Flink
Flink技术栈
Flink体系架构
Flink编程实践
图结构数据
图计算框架Pregel
Pregel图计算模型
Pregel体系结构
Pregel应用实例
数据查询分析
基于Hadoop的数据仓库Hive
Hive系统架构
Hive工作原理
Hive编程实践
数据可视化
入门级工具
信息图表工具
地图工具
时间线工具
高级分析工具

第四篇：大数据应用

大数据已经在社会生产和日常生活中得到了广泛的应用，对人类社会的发展进步起着重要的推动作用。本篇介绍大数据在互联网、生物医学、物流、城市管理、金融、汽车、零售、餐饮、电信、能源、体育娱乐、安全、政府、日常生活等方面的应用，从中我们可以深刻地感受到大数据对社会的影响及其重要价值。

本篇包括3章。第15章以推荐系统为核心介绍大数据在互联网领域的应用；第16章介绍大数据在生物医学领域的应用；第17章介绍大数据在其他领域的应用。其中，第15章需要重点理解，其他章节可以作为开拓视野的拓展性阅读材料。


第四篇 大数据应用
互联网
推荐系统
生物医学
流行病预测
智慧医疗
生物信息学
案例：基于大数据的综合健康服务平台
物流
智能物流
中国智能物流骨干网——菜鸟
城市管理
智能交通
环保监测
城市规划
安防领域
金融行业
高频交易
市场情绪分析
信贷风险分析
汽车行业
无人驾驶汽车
零售行业
发现关联购买行为
客户群体细分
供应链管理
餐饮行业
餐饮O2O
电信行业
电信客户离网分析
能源行业
智能电网
体育娱乐
训练球队
投拍影视作品
预测比赛结果
安全领域
国家安全
防御网络攻击
预防犯罪
政府领域
选举
日常生活
个性化服务

0.3 教材介绍

《大数据技术原理与应用——概念、存储、处理、分析与应用》第3版
林子雨编著，人民邮电出版社，2021年1月

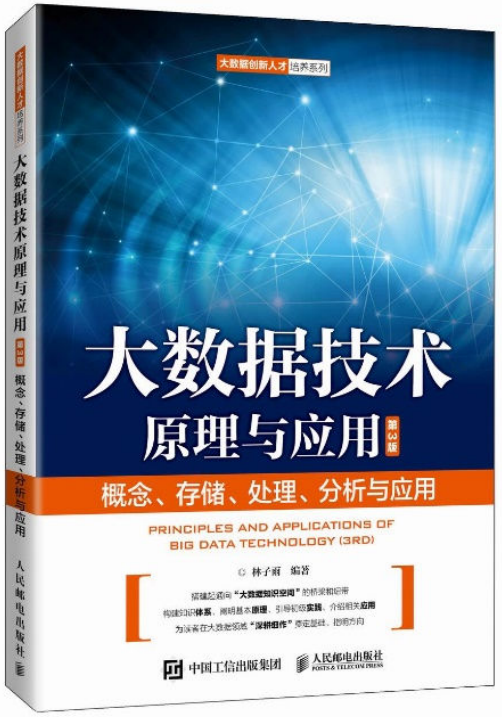
内容简介：

（1） **大数据基础篇**：介绍当前紧密关联的最新IT领域技术云计算、大数据和物联网。

（2） **大数据存储篇**：介绍分布式数据存储的概念、原理和技术，包括HDFS、HBase、NoSQL数据库、云数据库等。

（3） **大数据处理与分析篇**：介绍MapReduce、Hive、Spark、流计算、流处理框架Flink、图计算等。

（4） **大数据应用篇**：在互联网、生物医学和物流等各个领域的应用。

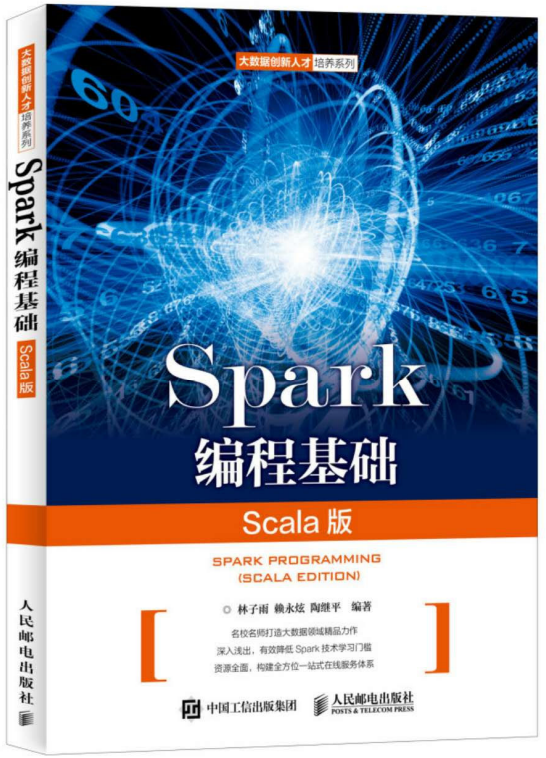


第1章 大数据概述
第2章 大数据处理架构Hadoop
第3章 分布式文件系统HDFS
第4章 分布式数据库HBase
第5章 NoSQL数据库
第6章 云数据库
第7章 MapReduce
第8章 Hadoop架构再探讨
第9章 数据仓库Hive
第10章 Spark
第11章 流计算
第12章 Flink
第13章 图计算
第14章 数据可视化
第15-17章 大数据在不同领域的应用

《Spark编程基础（Python版）》

林子雨，郑海山，赖永炫 编著　人民邮电出版社，2020年7月第1版

本书以Python作为开发Spark应用程序的编程语言，系统介绍了Spark编程的基础知识。全书共8章，内容包括大数据技术概述、Spark的设计与运行原理、Spark环境搭建和使用方法、RDD编程、Spark SQL、Spark Streaming、Structured Streaming、 Spark MLlib等。本书每个章节都安排了入门级的编程实践操作，以便读者更好地学习和掌握Spark编程方法。



第1章 大数据技术概述

第2章 Spark的设计与运行原理

第3章 Spark环境搭建和使用方法

第4章 RDD编程

第5章 Spark SQL

第6章 Spark Streaming

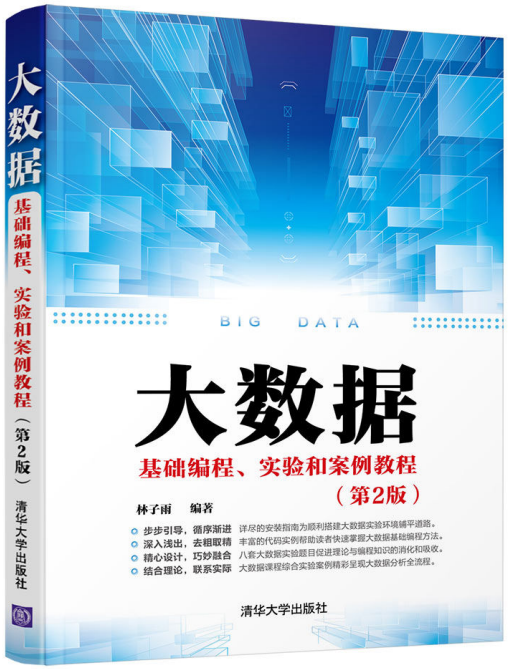
第7章 Structured Streaming

第8章 Spark MLlib

《大数据基础编程、实验和案例教程》第2版
林子雨编著，清华大学出版社，2020年10月

内容简介：

本教程是林子雨编著《大数据技术原理与应用》的配套实验指导书。《大数据技术原理与应用》侧重于大数据知识框架和理论介绍，而本教程侧重于介绍大数据软件的安装、使用和基础编程方法，并提供了大量实验和案例。由于大数据软件都是开源软件，安装过程一般比较复杂，也很耗费时间。为了尽量减少读者搭建大数据实验环境时的障碍，笔者在本教程中详细写出了各种大数据软件的详细安装过程，可以确保读者顺利完成大数据实验环境搭建。



第1章 大数据技术概述
第2章 Linux系统的安装和使用
第3章 Hadoop的安装和使用
第4章 HDFS操作方法和基础编程
第5章 HBase的安装和基础编程
第6章 典型NoSQL数据库的安装和使用
第7章 MapReduce基础编程
第8章 数据仓库Hive的安装和使用
第9章 Spark的安装和基础编程
第10章 典型可视化工具的使用方法
第11章 数据采集工具的安装和使用
第12章 大数据课程综合实验案例

0.4 电子资源

中国高校大数据课程

公 共 服 务 平 台

http://dblab.xmu.edu.cn/post/bigdata-teaching-platform/

扫一扫访问平台主页

扫一扫观看动画宣传片

访问地址：**http://dblab.xmu.edu.cn/post/7553/**

1.《大数据技术原理与应用》教材
官网：http://dblab.xmu.edu.cn/post/bigdata3/
2.大数据软件安装和编程实践指南
官网http://dblab.xmu.edu.cn/post/13741/
3.备课指南
官网：http://dblab.xmu.edu.cn/post/5637/
4.授课视频
官网：http://dblab.xmu.edu.cn/post/bigdata-online-course/
5.实验指南
官网：http://dblab.xmu.edu.cn/post/6131/
6.电子书籍
官网：http://dblab.xmu.edu.cn/post/4782/
7.Spark入门教程
官网：http://dblab.xmu.edu.cn/blog/spark/
8.大数据课程实验案例《网站用户购物行为分析》
官网：http://dblab.xmu.edu.cn/post/7499/

大数据技术原理与应用/爱课程（中国大学MOOC）

大数据技术原理与应用（厦门大学）

林子雨

- 课程名称:大数据技术原理与应用
- 主要建设单位：厦门大学
- 课程负责人：林子雨
- 主要开课平台：爱课程（中国大学MOOC）
- 认定年份：2018
- 课程官网：http://www.icourse163.org/course/XMU-1002335004

国家精品在线开放课程 /数据科学与大数据技术专业领域

前言

2019年1月22日，教育部办公厅公布了第二批国家精品在线开放课程认定结果。本文为您盘点数据科学与大数据专业领域的国家精品在线开放课程，包括2017年第一批认定课程（注：排名不分先后）。

1 大数据算法（哈尔滨工业大学）

王宏志

- 课程名称:大数据算法
- 主要建设单位：哈尔滨工业大学
- 课程负责人：王宏志
- 主要开课平台：爱课程（中国大学MOOC）
- 认定年份：2017
- 课程官网：http://www.icourse163.org/course/HIT-10001

2 大数据平台核心技术（清华大学）

武永卫

- 课程名称:大数据平台核心技术
- 主要建设单位：清华大学
- 课程负责人：武永卫
- 主要开课平台：学堂在线
- 认定年份：2017
- 课程官网：http://www.xuetangx.com/courses/course-v1:TsinghuaX+60240202X+sp/about

5 大数据技术原理与应用（厦门大学）

林子雨

- 课程名称:大数据技术原理与应用
- 主要建设单位：厦门大学
- 课程负责人：林子雨
- 主要开课平台：爱课程（中国大学MOOC）
- 认定年份：2018
- 课程官网：http://www.icourse163.org/course/XMU-1002335004

6 生物大数据（福建农林大学）

何华勤

- 课程名称:生物大数据
- 主要建设单位：福建农林大学
- 课程负责人：何华勤
- 主要开课平台：爱课程（中国大学MOOC）
- 认定年份：2018
- 课程官网：http://www.icourse163.org/course/FAFU-1001766004

3 数据科学导论（中国人民大学）

朝乐门

- 课程名称:数据科学导论
- 主要建设单位：中国人民大学
- 课程负责人：朝乐门
- 主要开课平台：北京高校优质课程研究会
- 认定年份：2018
- 课程官网：http://www.livedu.com.cn/ispace4.0/moocxjkc/toKcView.do?kcid=238

4 大数据系统基础（清华大学）

王建民

- 课程名称:大数据系统基础
- 主要建设单位：清华大学
- 课程负责人：王建民
- 主要开课平台：学堂在线
- 认定年份：2018
- 课程官网：http://www.xuetangx.com/courses/course-v1:TsinghuaX+64100033X+2018_T2/about

0.5 考核方式

- **上课时间**
  - 星期二：(1-2节) 1-12周，东九楼 D101
  - 星期五：(7-8节) 1-12周，东九楼 D101
- **实验时间（待确定）**
  - 时间：14周，周五晚上9-12 节（2023-05-19）

    15周，周五晚上9-12 节（2023-05-26）
  - 地点：东17楼F206（主校区）

教学方式

- 线上教学：讲授大数据技术基础知识
- 实验教学：Hadoop 和 Spark 安装与使用
- 线上讨论：大数据应用案例大讨论

考核方式（考查形式：作业+实验报告+设计报告）

- 小组课程报告题目（题目自定）
  - **A1.** 超市零售大数据分析报告
  - **A2** **Netflix**电影大数据分析报告
  - **A3.** 社交资源共享站点用户行为大数据分析报告
  - **A4.** 新浪微博消息大数据分析报告
  - **A5.** 带有感情标签的微博大数据分析报告
  - **A6.** 网络安全日志大数据分析报告
  - **A7.** 出租车**GPS**位置大数据分析报告
  - ……

美好课堂，携手共建！