概述

- 大数据的特征：数据量大、数据类型繁多、处理速度快、价值密度低
- 大数据技术包含 P15
- 大数据的核心：（猜测）云计算、物联网 P18
- 大数据的来源
  - 大数据的发展历程 P6
  - 数据来源：交易数据、移动通信数据、人为数据、机器和传感数据、互联网上的开放数据

Hadoop

- 安装模式：单机模式、伪分布模式、全分布模式 P33
- 核心是 HDFS 和 MapReduce P31
- 生态系统 P31
- 版本 P30
- 局限性 P155

HDFS

- 结构模型 P43、P47（NameNode、DataNode）
- 特点：设计需求P44、实现目标P45
- SecondaryNameNode 的作用 P47
- 通信协议 P49
- 冗余因子 P50
- 数据存储策略 P51
- 读写特征 P53
- 常用命令 P55

MapReduce

- 体系结构：（百度）Client、JobTracker、TaskTracker以及Task
- 工作流程：P134
- shuffle过程：P136
- Reduce端的shuffle过程：领取数据、归并数据、把数据输入给Reduce任务 P138

HBase

- NoSQL的列族数据库：P100
- 底层数据存在：HDFS P64
- 表的索引：行键、列祖、列限定符、时间戳 P66
- 三层结构：P73
- 强大的计算能力：MapReduce P64
- 系统架构：P74
- 启动后进程：NameNode、SecondaryNameNode、DataNode、HRegionServer、Jps、HQuorumPeer、HMaster
- 基本shell命令 P78

Spark

- 特点：P173
- 生态系统包含组件：Spark Core、Spark SQL、Spark Streaming、MLib、GraphX P176
- 运行框架 P177

RDD

- 特点：P181
- 依赖关系分类：P182
- 操作分类： P187
- 转换操作和行动操作：P188