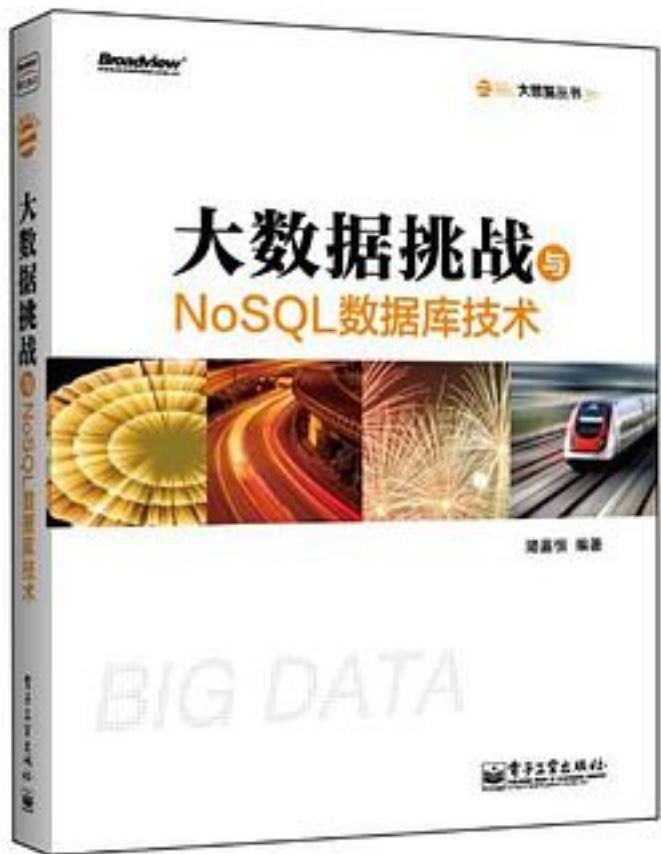


# 大数据挑战与NoSQL数据库技术



[大数据挑战与NoSQL数据库技术\\_下载链接1](#)

著者:陆嘉恒

出版者:电子工业出版社

出版时间:2013-4-15

装帧:平装

isbn:9787121196607

本书共分为三部分。理论篇重点介绍大数据时代下数据处理的基本理论及相关处理技术，并引入NoSQL数据库；系统篇主要介绍了各种类型NoSQL数据库的基本知识；应用篇对国内外几家知名公司在利用NoSQL数据库处理海量数据方面的实践做了阐述。

本书对大数据时代面临的挑战，以及NoSQL数据库的基本知识做了清晰的阐述，有助

于读者整理思路，了解需求，并更有针对性、有选择地深入学习相关知识。

## 作者介绍：

陆嘉恒，中国人民大学教授，博士生导师。2006年毕业于新加坡国立大学计算机科学系，获博士学位；2006-2008年在美国加利福尼亚大学尔湾分校(University of California, Irvine)进行博士后研究；2008年加入中国人民大学，2012年破格晋升为教授。主要研究领域包括数据库技术和云计算技术。先后在SIGMOD、VLDB、ICDE、WWW等国际重要会议和期刊上发表数据库方向的论文40多篇，主编多本云计算和大数据的教材和著作。

## 目录: 第1章 概论 1

- 1.1 引子 2
- 1.2 大数据挑战 3
- 1.3 大数据的存储和管理 5
  - 1.3.1 并行数据库 5
  - 1.3.2 NoSQL数据管理系统 6
  - 1.3.3 NewSQL数据管理系统 8
  - 1.3.4 云数据管理 11
- 1.4 大数据的处理和分析 11
- 1.5 小结 13

## 参考文献 13

### 理论篇

- 第2章 数据一致性理论 16
  - 2.1 CAP理论 17
  - 2.2 数据一致性模型 21
  - 2.3 ACID与BASE 22
  - 2.4 数据一致性实现技术 23
    - 2.4.1 Quorum系统NRW策略 23
    - 2.4.2 两阶段提交协议 24
    - 2.4.3 时间戳策略 27
    - 2.4.4 Paxos 30
    - 2.4.5 向量时钟 38
  - 2.5 小结 43

## 参考文献 43

- 第3章 数据存储模型 45
  - 3.1 总论 46
  - 3.2 键值存储 48
    - 3.2.1 Redis 49
    - 3.2.2 Dynamo 49
  - 3.3 列式存储 50
    - 3.3.1 Bigtable 51
    - 3.3.2 Cassandra与HBase 51
  - 3.4 文档存储 52
    - 3.4.1 MongoDB 53
    - 3.4.2 CouchDB 53
  - 3.5 图形存储 54
    - 3.5.1 Neo4j 55
    - 3.5.2 GraphDB 55
  - 3.6 本章小结 56

## 参考文献 56

- 第4章 数据分区与放置策略 58
  - 4.1 分区的意义 59

4.1.1 为什么要分区	59
4.1.2 分区的优点	60
4.2 范围分区	61
4.3 列表分区	62
4.4 哈希分区	63
4.5 三种分区的比较	64
4.6 放置策略	64
4.6.1 一致性哈希算法	65
4.6.2 容错性与可扩展性分析	66
4.6.3 虚拟节点	68
4.7 小结	69
参考文献	69
第5章 海量数据处理方法	70
5.1 MapReduce简介	71
5.2 MapReduce数据流	72
5.3 MapReduce数据处理	75
5.3.1 提交作业	76
5.3.2 初始化作业	78
5.3.3 分配任务	78
5.3.4 执行任务	79
5.3.5 更新任务执行进度和状态	80
5.3.6 完成作业	81
5.4 Dryad简介	81
5.4.1 DFS Cosmos介绍	82
5.4.2 Dryad执行引擎	84
5.4.3 DryadLINQ解释引擎	86
5.4.4 DryadLINQ编程	88
5.5 Dryad数据处理步骤	90
5.6 MapReduce vs Dryad	92
5.7 小结	94
参考文献	95
第6章 数据复制与容错技术	96
6.1 海量数据复制的作用和代价	97
6.2 海量数据复制的策略	97
6.2.1 Dynamo的数据库复制策略	97
6.2.2 CouchDB的复制策略	99
6.2.3 PNUTS的复制策略	99
6.3 海量数据的故障发现与处理	101
6.3.1 Dynamo的数据库的故障发现与处理	101
6.3.2 CouchDB的故障发现与处理	103
6.3.3 PNUTS的故障发现与处理	103
6.4 小结	104
参考文献	104
第7章 数据压缩技术	105
7.1 数据压缩原理	106
7.1.1 数据压缩的定义	106
7.1.2 数据为什么可以压缩	107
7.1.3 数据压缩分类	107
7.2 传统压缩技术[1]	108
7.2.1 霍夫曼编码	108
7.2.2 LZ77算法	109
7.3 海量数据带来的3V挑战	112
7.4 Oracle混合列压缩	113
7.4.1 仓库压缩	114

7.4.2 存档压缩	114
7.5 Google数据压缩技术	115
7.5.1 寻找长的重复串	115
7.5.2 压缩算法	116
7.6 Hadoop压缩技术	118
7.6.1 LZO简介	118
7.6.2 LZO原理[5]	119
7.7 小结	121
参考文献	121
第8章 缓存技术	122
8.1 分布式缓存简介	123
8.1.1 分布式缓存的产生	123
8.1.2 分布式缓存的应用	123
8.1.3 分布式缓存的性能	124
8.1.4 衡量可用性的标准	125
8.2 分布式缓存的内部机制	125
8.2.1 生命期机制	126
8.2.2 一致性机制	126
8.2.3 直读与直写机制	129
8.2.4 查询机制	130
8.2.5 事件触发机制	130
8.3 分布式缓存的拓扑结构	130
8.3.1 复制式拓扑	131
8.3.2 分割式拓扑	131
8.3.3 客户端缓存拓扑	131
8.4 小结	132
参考文献	132
系统篇	
第9章 key-value数据库	134
9.1 key-value模型综述	134
9.2 Redis	135
9.2.1 Redis概述	135
9.2.2 Redis下载与安装	135
9.2.3 Redis入门操作	136
9.2.4 Redis在业内的应用	143
9.3 Voldemort	143
9.3.1 Voldemort概述	143
9.3.2 Voldemort下载与安装	144
9.3.3 Voldemort配置	145
9.3.4 Voldemort开发介绍[3]	147
9.4 小结	149
参考文献	149
第10章 Column-Oriented数据库	150
10.1 Column-Oriented数据库简介	151
10.2 Bigtable数据库	151
10.2.1 Bigtable数据库简介	151
10.2.2 Bigtable数据模型	152
10.2.3 Bigtable基础架构	154
10.3 Hypertable数据库	157
10.3.1 Hypertable简介	157
10.3.2 Hypertable安装	157
10.3.3 Hypertable架构	163
10.3.4 基本概念和原理	164
10.3.5 Hypertable的查询	168

10.4 Cassandra数据库 175  
10.4.1 Cassandra简介 175  
10.4.2 Cassandra配置 175  
10.4.3 Cassandra数据库的连接 177  
10.4.4 Cassandra集群机制 180  
10.4.5 Cassandra的读/写机制 182  
10.5 小结 183

参考文献 183

第11章 文档数据库 185  
11.1 文档数据库简介 186  
11.2 CouchDB数据库 186  
11.2.1 CouchDB简介 186  
11.2.2 CouchDB安装 188  
11.2.3 CouchDB入门 189  
11.2.4 CouchDB查询 200  
11.2.5 CouchDB的存储结构 207

11.2.6 SQL和CouchDB 209  
11.2.7 分布式环境中的CouchDB 210  
11.3 MongoDB数据库 211  
11.3.1 MongoDB简介 211  
11.3.2 MongoDB的安装 212  
11.3.3 MongoDB入门 215  
11.3.4 MongoDB索引 224  
11.3.5 SQL与MongoDB 226  
11.3.6 MapReduce与MongoDB 229  
11.3.7 MongoDB与CouchDB对比 234  
11.4 小结 236

参考文献 237

第12章 图存数据库 238  
12.1 图存数据库的由来及基本概念 239  
12.1.1 图存数据库的由来 239  
12.1.2 图存数据库的基本概念 239  
12.2 Neo4j图存数据库 240  
12.2.1 Neo4j简介 240  
12.2.2 Neo4j使用教程 241  
12.2.3 分布式Neo4j——Neo4j HA 251  
12.2.4 Neo4j工作机制及优缺点浅析 256  
12.3 GraphDB 258  
12.3.1 GraphDB简介 258  
12.3.2 GraphDB的整体架构 260  
12.3.3 GraphDB的数据模型 264  
12.3.4 GraphDB的安装 266  
12.3.5 GraphDB的使用 268  
12.4 OrientDB 276  
12.4.1 背景 276  
12.4.2 OrientDB是什么 276  
12.4.3 OrientDB的原理及相关技术 277  
12.4.4 Windows下OrientDB的安装与使用 282  
12.4.5 相关Web应用 286  
12.5 三种图存数据库的比较 288  
12.5.1 特征矩阵 288  
12.5.2 分布式模式及应用比较 289  
12.6 小结 289  
参考文献 290

## 第13章 基于Hadoop的数据管理系统 291

13.1 Hadoop简介 292

13.2 HBase 293

13.2.1 HBase体系结构 293

13.2.2 HBase数据模型 297

13.2.3 HBase的安装和使用 298

13.2.4 HBase与RDBMS 303

13.3 Pig 304

13.3.1 Pig的安装和使用 304

13.3.2 Pig Latin语言 306

13.3.3 Pig实例 311

13.4 Hive 315

13.4.1 Hive的数据存储 316

13.4.2 Hive的元数据存储 316

13.4.3 安装Hive 317

13.4.4 HiveQL简介 318

13.4.5 Hive的网络接口 (WebUI) 328

13.4.6 Hive的JDBC接口 328

13.5 小结 330

参考文献 331

## 第14章 NewSQL数据库 332

14.1 NewSQL数据库简介 333

14.2 MySQL Cluster 333

14.2.1 概述 334

14.2.2 MySQL Cluster的层次结构 336

14.2.3 MySQL Cluster的优势和应用 337

14.2.4 海量数据处理中的sharding技术 339

14.2.5 单机环境下MySQL Cluster的安装 343

14.2.6 MySQL Cluster的分布式安装与配置指导 348

14.3 VoltDB 350

14.3.1 传统关系数据库与VoltDB 351

14.3.2 VoltDB的安装与配置 351

14.3.3 VoltDB组件 354

14.3.4 Hello World 355

14.3.5 使用Generate脚本 361

14.3.6 Eclipse集成开发 362

14.4 小结 365

参考文献 365

## 第15章 分布式缓存系统 366

15.1 Memcached缓存技术 367

15.1.1 背景介绍 367

15.1.2 Memcached缓存技术的特点 368

15.1.3 Memcached安装[3] 374

15.1.4 Memcached中的数据操作 375

15.1.5 Memcached的使用 376

15.2 Microsoft Velocity分布式缓存系统 378

15.2.1 Microsoft Velocity简介 378

15.2.2 数据分类 379

15.2.3 Velocity核心概念 380

15.2.4 Velocity安装 382

15.2.5 一个简单的Velocity客户端应用 385

15.2.6 扩展型和可用性 387

15.3 小结 388

参考文献 388

## 应用篇

第16章 企业应用 392

16.1 Instagram 393

16.1.1 Instagram如何应对数据的急剧增长 395

16.1.2 Instagram的数据分片策略 398

16.2 Facebook对Hadoop以及HBase的应用 400

16.2.1 工作负载类型 401

16.2.2 为什么采用Apache Hadoop和HBase 403

16.2.3 实时HDFS 405

16.2.4 Hadoop HBase的实现 409

16.3 淘宝大数据解决之道 411

16.3.1 淘宝数据分析 412

16.3.2 淘宝大数据挑战 413

16.3.3 淘宝OceanBase数据库 414

16.3.4 淘宝将来的工作 422

16.4 小结 423

参考文献 423

• • • • • (收起)

[大数据挑战与NoSQL数据库技术\\_下载链接1](#)

## 标签

大数据

nosql

数据库

计算机科学

数据挖掘

计算机

数据分析

云计算

## 评论

: TP274/7549.2

好水啊……不过对我还是有用的。NoSQL是Not only SQL

学院派 作品， 非学生可以从中了解 cpa , base 理论

前1/3还行， 后面就太水了

综述，但没有参考文献...太简单了

虎头蛇尾

笔记：<http://artech.farbox.com/post/note-book/-da-shu-ju-tiao-zhan-yu-nosqlshu-ju-ku-ji-zhu-du-shu-bi-ji>

果然是“编著” 啊

可以学到不少基础知识。

综述，目测有四五个学生一起写的。综述的文章、书如果写好也很有价值，但这本...就当个引子看吧

开拓视野系列图书。

走马观花看看还行，开头几章理论部份还有点用

题目惊悚。看得晕乎。基本上是数据库软件下载安装手册。浮光般的知识值得你掠影般过。

入门大数据和NoSQL的一本书，还可以。感觉整本书应该不是一个人完成的，估计分章节找不同的人写的。

数据库面面观，当百度百科看了

浅显易懂，对于了解NoSQL数据库技术的全貌有帮助

读了1,2章，讲的不错～

基本上都是些基本的东西，非常宽泛，还不如去看官方文档。

基础上，概念书

[大数据挑战与NoSQL数据库技术\\_下载链接1](#)

## 书评

我读过《NoSQL数据库入门》和《NoSQL

Distilled》，感觉还是这本最适合初学者;cap、base、2pc、paxos理论深入浅出，对各种分布式数据库又介绍到位，操作讲的不是太多，理论又基本覆盖，假如希望研究还可以参考作者列举的论文。强力推荐！

---

[大数据挑战与NoSQL数据库技术\\_下载链接1](#)