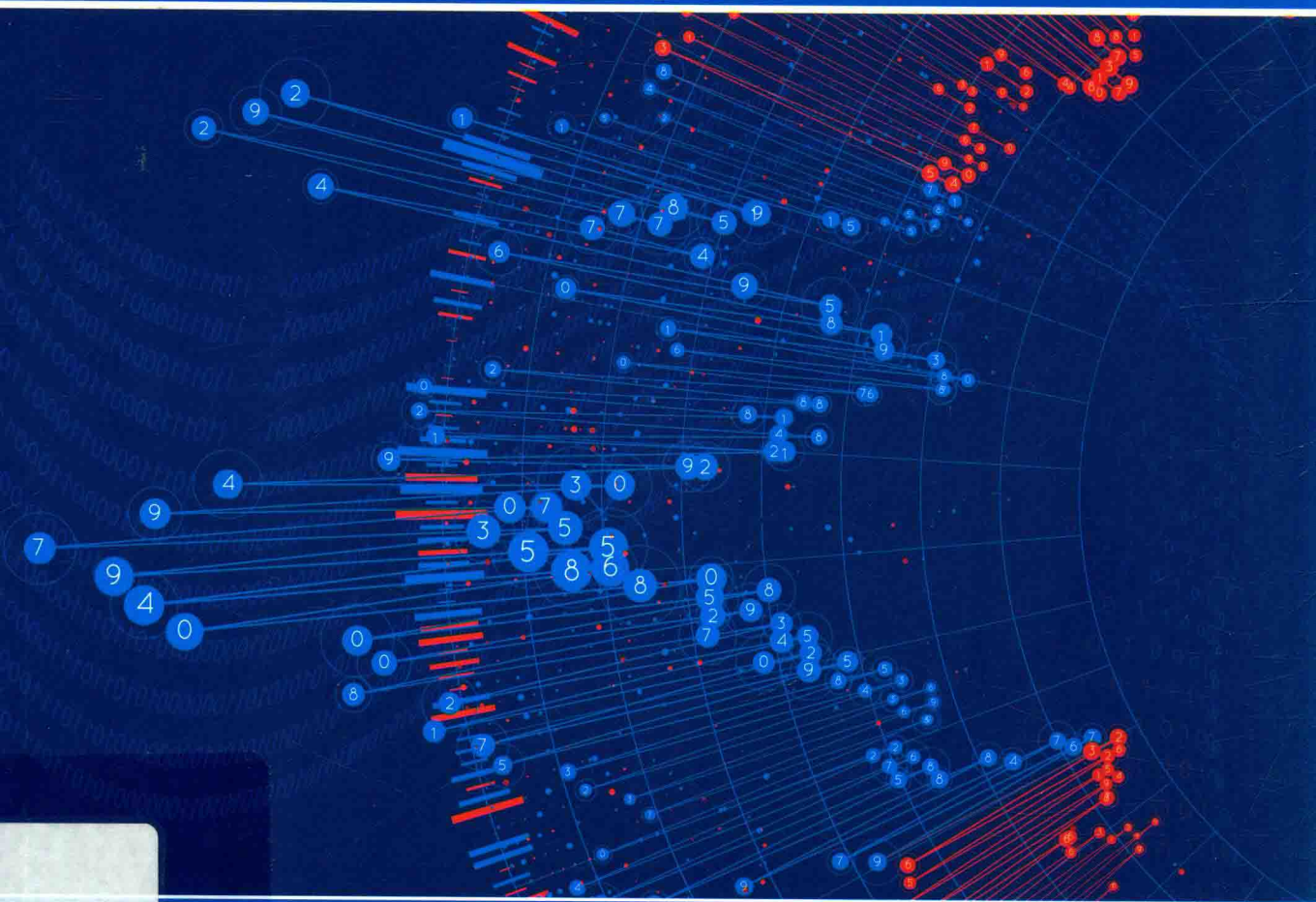
大数据处理与存储技术

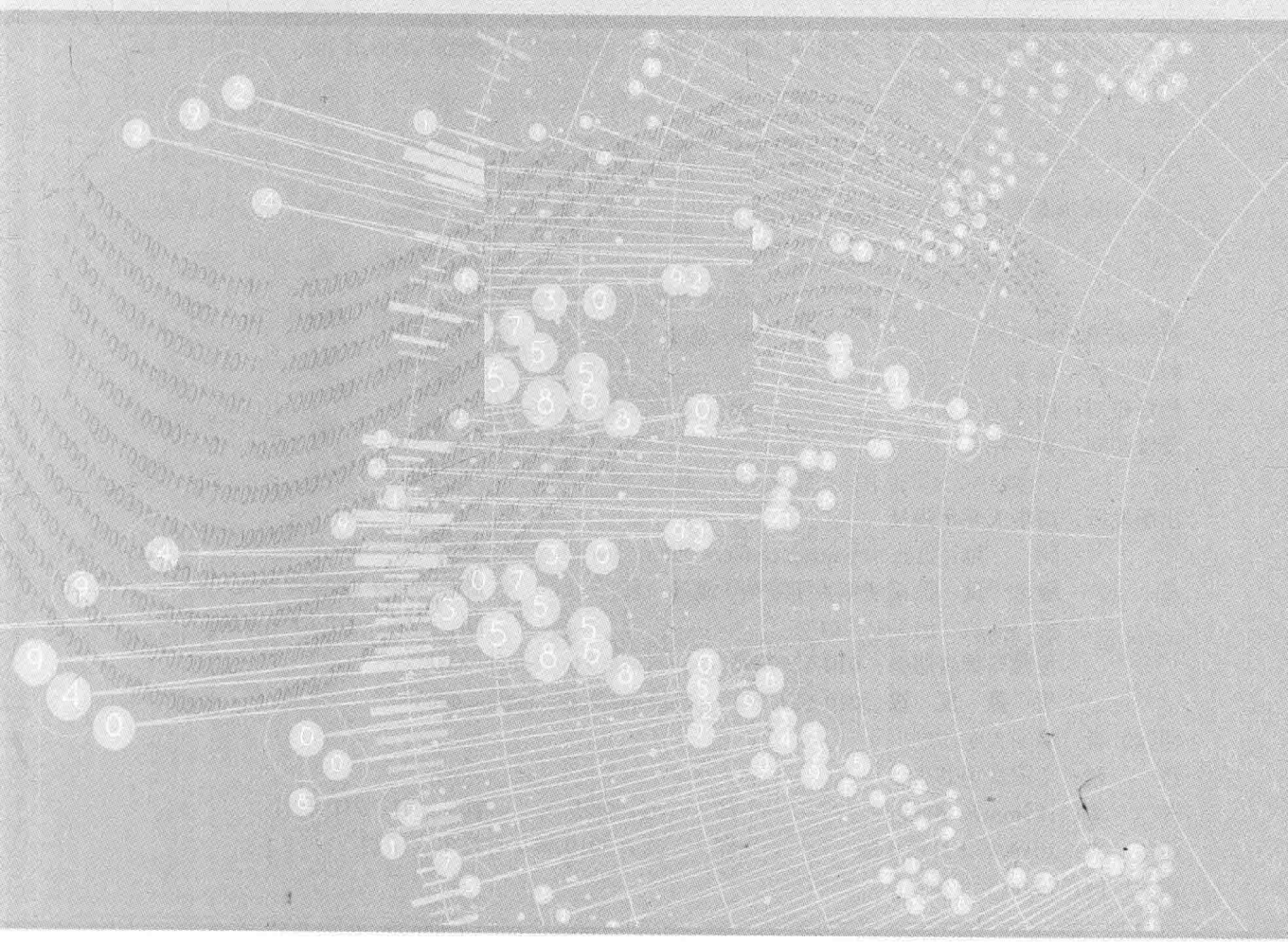
葛维春 主编



清華大學出版社

大数据处理与存储技术

葛维春 主编



清華大學出版社
北京

内 容 简 介

本书归纳和总结了主流数据库软件和常用数据处理工具的常见问题与应用技巧，为大数据技术与传统数据存储和转换技术相结合提供了技术参考，为促进大数据技术的发展，为数据库和ETL开发人员、运维人员提供了技术支撑。

本书分为3篇，共5章，主要内容包括Oracle数据库应用、MySQL数据库应用、Informatica PowerCenter工具应用、Kettle工具应用、数据库调优与ETL工具应用技巧。本书分别从数据存储软件、数据抽取与清洗软件等方面，向读者展示了Oracle、MySQL、Informatica和Kettle的常见问题、优化与提升的技巧。

本书所涉及的内容均为生产实践中必要的过程和阶段，讲解由浅入深、通俗易懂，适合从事数据库开发、维护、管理、优化任务和高可用设计的工程技术人员及从事ETL开发、优化的工程技术人员使用或参考。

图书在版编目（CIP）数据

大数据处理与存储技术 / 葛维春主编. — 北京：清华大学出版社，2019
ISBN 978-7-302-51720-7

Ⅰ. ①大… Ⅱ. ①葛… Ⅲ. ①数据处理 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2018）第 266967 号

责任编辑：杨如林
封面设计：杨玉兰
责任校对：胡伟民
责任印制：杨 艳

出版发行：清华大学出版社
网 址：http://www.tup.com.cn，http://www.wqbook.com
地 址：北京清华大学学研大厦 A 座 邮 编：100084
社 总 机：010-62770175 邮 购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：清华大学印刷厂
经 销：全国新华书店
开 本：185mm×260mm 印 张：25.5 字 数：711 千字
版 次：2019 年 2 月第 1 版 印 次：2019 年 2 月第 1 次印刷
定 价：79.00 元

产品编号：079262-01

编委会名单

主　编：葛维春

副主编（排名不分先后）

王　磊　程志华　范鹏展　柏俊峰　郭昆亚　郭永贵　杨　光　蔡立群　阎青春
苏安龙　马显智　顾洪群　申　扬　刘树吉　李　伟　田小蕾　黄文思

编写组成员（排名不分先后）

朱洪斌　梅文明　刘　虎　张朝阳　邢蒙蒙　纪永满　王显波　田海韬　左　壮
刘忠威　齐　悦　葛延峰　崔万里　雷振江　刘　颖　冉　冉　周大鹏　王丹妮
刘雪松　胡　畔　刘鹏宇　李云鹏　夏　雨　李占军　刘　洋　王　刚　于　海
刘　岩　李　岩　刘少军　曹　凯　金　兰　魏　霞　王　昊　王小溪　王　宁
历　丽　蒯继鹏　杨海峰　周晓明　王丽霞　赵　军　郑永健　王　勇　宋　季
丛海洋　于泓维　张幼明　李凤强　王鹏宇　龙　云　郭　勇　赵明江　郭志彤
曹丽娜　高　潇　吕旭明　毛洪涛　黄笑伯　周武明　郭长彪　周英男　刘　双
陈　蓉　金　妍　刘中彦　巴明强　陈娇茵　李耀宗　田庆阳　孙峰烈　王　阳
茹满辉　金福国　李学斌　赵　军　王浩淼　王天博　陈　硕　王　磊　杨　超
陈　龙　刘　瑞　焦　勇　李　斌　胡　楠　罗义旺　李金湖　罗顺辉　林　燊
余仰淇　刘　青　张毅琦　陈　坤　黄鑫烨　吴胜竹　林海玉　邢聪辉　陈彦达
韩天阳　朱继阳　白雨佳　陈永强　谢宏宇　曹国强

前　言

关于本书

“大数据”是当前最热门的话题之一，虽然其实现技术多种多样，但是其应用和实践依然是基于大量实际业务数据的。本书从数据存储和数据清洗与转换的角度出发，针对当前主流数据库软件（Oracle、MySQL）和主流ETL工具（Informatica PowerCenter、Kettle）的常见问题、应用技巧进行归纳和总结，为大数据技术与传统数据存储技术和数据转换技术相结合提供技术参考，促进大数据技术的发展，为数据库和ETL开发人员、运维人员提供技术支撑。

本书内容主要来源于项目实践，如有不当之处，恳请读者批评指正，部分支撑材料来源于网络，如涉及版权问题请相关作者及时联系沟通，联系方式：838743142@qq.com。

本书的读者群体

本书的目标读者是从事数据库开发、维护、管理、优化任务和高可用设计的工程技术人员和从事ETL开发、优化的工程技术人员。

从技术角度看，本书涉及的内容均为生产实践中必要的过程和阶段，因此本书的内容精炼、通俗易懂，尤其适合作为开发人员的基础工具书。本书读者无须拥有非常深厚的专业技术基础。

本书内容安排

本书共5章，分为数据库软件篇、ETL工具篇及高级调优篇，各章的内容分述如下。

数据库软件篇	第1章：Oracle数据库应用 该章从Oracle简介、安装配置、数据库函数、常用查询命令、常见问题参考等方面，介绍了Oracle的产品特点、常见问题及解决技巧
	第2章：MySQL数据库应用 该章从MySQL简介、安装配置、数据库函数、常见问题参考等方面，介绍了MySQL的产品特点、常见问题及解决技巧

（续表）

ETL工具篇	第3章：Informatica PowerCenter工具应用 该章从Informatica简介、安装配置及常见问题参考，介绍了Informatica的产品特点、常见问题及解决办法
	第4章：Kettle工具应用 该章从Kettle简介、安装配置及常见问题，介绍了Kettle的产品特点、常见问题及解决办法
高级调优篇	第5章：数据库调优与ETL工具应用技巧 该章对Oracle和MySQL数据库的性能调优和Informatica PowerCenter工具的应用技巧进行介绍

致谢

首先感谢国网辽宁省电力有限公司全业务统一数据中心分析域项目组全体技术人员对本书的技术支撑，他们从项目实践的角度为本书提供数据库和ETL技术的实践支撑，他们在工作和技术领域中的不断探索促进了本书内容的不断完善。

还要特别感谢本书的编辑，感谢他们审查书稿，并提出他们的观点和建议，他们的宝贵意见为本书的成功出版提供了方向指引。

编　者

目　录

第一篇　数据库软件篇

第二篇 ETL工具篇

第三篇 高级调优篇

第一篇
数据库软件篇

第1章
Oracle数据库应用

本章从Oracle简介、安装配置、数据库函数、常用查询命令、常见问题参考等方面，介绍Oracle的产品特点、常见问题及解决技巧。

- Oracle简介
- 安装配置
- 数据库函数
- 常用查询命令
- 常见问题参考

1.1 Oracle简介

Oracle Database又名Oracle RDBMS（简称Oracle），是甲骨文公司的一款关系数据库管理系统。它在数据库领域是一直处于领先地位的产品，是目前世界上流行的关系数据库管理系统。Oracle的系统具有可移植性好、使用方便、功能强大等特点，适用于各类大、中、小型机及微机环境，它也是一种高效可靠的适应高吞吐量的数据库解决方案。

Oracle数据库系统是以分布式数据库为核心的一组软件产品，是目前最流行的浏览器/服务器（Browser/Server，B/S）体系结构的数据库之一。比如，SilverStream就是基于数据库的一种中间件。作为一个通用的数据库系统，Oracle具有完整的数据管理功能；作为一个关系数据库，Oracle是一个具有完备关系的产品；作为分布式数据库，Oracle实现了分布式处理功能。针对Oracle数据库操作和应用的知识与技能，只要在一种机型上学习了Oracle，便能在各种机型上使用。

Oracle数据库的最新版本为Oracle Database12c。该版本引入了一个新的多承租方架构，使用该架构可以轻松部署和管理数据库云。此外，一些创新特性可最大限度地提高资源利用率和灵活性。例如，Oracle Multitenant可以快速整合多个数据库，而Automatic Data Optimization和HeatMap能以更高的密度压缩数据和对数据分层。这些独一无二的技术再结合其在可用性、安全性和大数据支持方面的增强，使得该版成为私有云和公有云部署的理想平台。

1.1.1 产品历史

1979年夏季，RSI发布了Oracle第2版。这个数据库产品整合了比较完整的SQL实现，其中包括子查询、连接及其他特性。

1983年3月，RSI发布了Oracle第3版。从该版本起Oracle产品有了一个关键的特性，即可移植性。

1984年10月，Oracle发布了第4版产品。该版增加了读一致性这个重要特性。

1985年，Oracle发布了5.0版。这个版本算得上是Oracle数据库的稳定版本，这也是首批可以在客户/服务器（Client/Server，C/S）模式下运行的RDBMS产品。

1986年，Oracle发布了5.1版。该版本支持分布式查询，允许通过一次性查询访问存储

在多个位置的数据。

1988年，Oracle发布了第6版。该版本引入了行级锁这个重要的特性，同时引入了联机热备份功能。

1992年6月，Oracle发布了第7版。该版本增加了许多新的性能特性，包括分布式事务处理功能、增强的管理功能、用于应用程序开发的新工具以及安全性方法。

1997年6月，Oracle第8版发布。Oracle 8支持面向对象的开发及新的多媒体应用，这个版本也为支持Internet、网络计算等奠定了基础。

1998年9月，Oracle公司正式发布Oracle 8i。这一版本添加了大量为支持Internet而设计的特性，同时这一版本为数据库用户提供了全方位的Java支持。

2001年6月，Oracle发布了Oracle 9i。在Oracle 9i的诸多新特性中，最重要的就是Real Application Clusters（RAC）了。

2003年9月，Oracle发布了Oracle 10g。该版本的最大的特性就是加入了网格计算的功能。

2007年7月，Oracle发布了Oracle 11g。Oracle 11g是甲骨文公司30年来发布的最重要的数据库版本，根据用户的需求实现了信息生命周期管理(Information Lifecycle Management)等多项创新。

2013年6月，Oracle 发布了Oracle 12c。该版本之前的Oracle 10g和Oracle 11g中的g代表grid，而Oracle 12c中的c代表cloud，代表云计算。

1.1.2 支撑的平台

在2001年发布的Oracle 9i之前，甲骨文公司把他们的数据库产品广泛地移植到了不同的平台上。截至2015年1月，甲骨文公司的Oracle 10g/11g/12c支持以下操作系统和硬件。

- Apple Mac OS X Server:PowerPC
- HP-UX:PA-RISC,Itanium
- HP Tru64 UNIX:Alpha
- HP OpenVMS: Alpha,Itanium
- IBM AIX 5L:IBM POWER
- IBM z/OS:zSeries
- Linux:x86,x86-64, PowerPC,zSeries,Itanium
- Microsoft Windows:x86,x86-64,Itanium
- Sun Solaris:SPARC,x86,x86-64

1.1.3 数据库特点

（1）完整的数据管理功能。

- 数据的大量性；

- 数据保存的持久性；
- 数据的共享性；
- 数据的可靠性。

（2）完备关系的产品。

- 信息准则——关系型DBMS的所有信息都应在逻辑上用一种方法，即表中的值显式表示；
- 保证访问的准则；
- 视图更新准则——只要形成视图的表中的数据变化了，相应视图中的数据也同时变化；
- 数据物理性和逻辑性独立准则。

（3）分布式处理功能。

Oracle数据库自第5版起就具有了分布式处理能力，到第7版就有比较完善的分布式数据库功能了。一个Oracle分布式数据库由oraclerdbms、SQL*Net、SQL*connect和其他非Oracle的关系型产品构成。

（4）用Oracle能轻松实现数据仓库的操作。

1.2 安装配置

1.2.1 安装环境

Oracle数据库所需安装的软件、安装环境和操作系统如表1-1所示。

表1-1 Oracle安装环境

序号	服务	软件环境
1	操作系统	RedHat 6.5 64
2	数据库软件	Oracle11g-64

1.2.2 系统配置

操作系统须在Root用户下进行如下配置。

（1）关闭SELinux、防火墙（后续要打开防火墙，须开放1521端口并允许ip通过），命令如下。

```
service  iptables stop
chkconfig iptables off
vi /etc/selinux/config
```

把SELINUX=enforcing改为SELINUX=disabled，重启计算机或者用命令使之立刻生效。

```
# setenforce 0
```

（2）检查hosts文件。

```
vim  /etc/hosts
127.0.0.1      localhost.localdomain  localhost
172.0.0.214    localhost.localdomain  localhost
```

（3）修改Linux内核，修改/etc/sysctl.conf文件，输入命令:vim /etc/sysctl.conf，按i键进入编辑模式，修改或添加下列内容，编辑完成后按Esc键，输入“:wq”保存并退出，使用命令：sysctl –p使之立刻生效。

```
fs.suid_dumpable = 1
fs.aio-max-nr = 1048576
fs.file-max = 6815744
kernel.shmall = 2097152
kernel.shmmax = 536870912
kernel.shmmni = 4096
kernel.sem = 250 32000 100 128
net.ipv4.ip_local_port_range = 9000 65500
net.core.rmem_default=4194304
net.core.rmem_max=4194304
net.core.wmem_default=262144
net.core.wmem_max=1048586
```

（4）修改用户的SHELL限制，输入命令：vim /etc/security/limits.conf，按i键进入编辑模式，添加下列内容，编辑完成后按Esc键，输入“:wq”保存并退出。

```
oracle              soft    nproc   2047
oracle              hard    nproc   16384
oracle              soft    nofile  4096
oracle              hard    nofile  65536
oracle              soft    stack   10240
```

（5）修改/etc/pam.d/login文件，输入命令：vim /etc/pam.d/login，按i键进入编辑模式，添加下列内容，编辑完成后按Esc键，输入“:wq”保存并退出。

```
session    required     /lib/security/pam_limits.so
session    required     pam_limits.so
```

（6）编辑/etc/profile，输入命令：vim /etc/profile，添加下列内容，编辑完成后按Esc键，输入“:wq”存盘并退出。

```
if [ $USER = "oracle" ]; then
```

```
if [ $SHELL = "/bin/ksh" ]; then
     ulimit -p 16384
     ulimit -n 65536
else
     ulimit -u 16384 -n 65536
fi
fi
```

（7）检查所需的包，是否缺少如下安装包。

```
binutils-2.20.51.0.2-5.11.el6 (x86_64)
compat-libcap1-1.10-1 (x86_64)
compat-libstdc++-33-3.2.3-69.el6 (x86_64)
compat-libstdc++-33-3.2.3-69.el6.i686
gcc-4.4.4-13.el6 (x86_64)
gcc-c++-4.4.4-13.el6 (x86_64)
glibc-2.12-1.7.el6 (i686)
glibc-2.12-1.7.el6 (x86_64)
glibc-devel-2.12-1.7.el6 (x86_64)
glibc-devel-2.12-1.7.el6.i686
ksh
libgcc-4.4.4-13.el6 (i686)
libgcc-4.4.4-13.el6 (x86_64)
libstdc++-4.4.4-13.el6 (x86_64)
libstdc++-4.4.4-13.el6.i686
libstdc++-devel-4.4.4-13.el6 (x86_64)
libstdc++-devel-4.4.4-13.el6.i686
libaio-0.3.107-10.el6 (x86_64)
libaio-0.3.107-10.el6.i686
libaio-devel-0.3.107-10.el6 (x86_64)
libaio-devel-0.3.107-10.el6.i686
make-3.81-19.el6
sysstat-9.0.4-11.el6 (x86_64)
unixODBC-2.2.14-12.el6_3.i686.rpm
unixODBC-2.2.14-12.el6_3.x86_64.rpm
unixODBC-devel-2.2.14-12.el6_3.i686.rpm
unixODBC-devel-2.2.14-12.el6_3.x86_64.rpm
libXp-1.0.0-15.1.el6.i686.rpm
libXp-devel-1.0.0-15.1.el6.i686.rpm
libXp-1.0.0-15.1.el6.x86_64.rpm
libXp-devel-1.0.0-15.1.el6.x86_64.rpm
elfutils-libelf-devel-0.152-1.el6.x86_64.rpm
```

（8）创建Oracle用户和组。

①创建组，使用如下命令。

```
groupadd  oinstall
```