

The Internals of PostgreSQL
for Database Administrators and System Developers



PostgreSQL 指南

内幕探索

[日]铃木启修◎著
冯若航 刘阳明 张文升◎译



中国工信出版集团



电子工业出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
<http://www.phei.com.cn>

The Internals of PostgreSQL
for Database Administrators and System Developers

PostgreSQL指南

内幕探索

[日]铃木启修◎著
冯若航 刘阳明 张文升◎译

电子工业出版社
Publishing House of Electronics Industry
北京•BEIJING

内 容 简 介

本书介绍了 PostgreSQL 内部的工作原理，包括数据库对象的逻辑组织与物理实现、进程与内存的架构，并依次剖析了几个重要的子系统：查询处理、外部数据包装器、并发控制、清理过程、缓冲区管理、WAL、备份及流复制。本书为 DBA 与系统开发者提供了一幅全景概念地图，有助于读者形成对数据库实现的整体认识，亦可作为深入学习 PostgreSQL 源代码的导读手册，对于理解数据库原理与 PostgreSQL 内部实现大有裨益。

本书适合数据库开发人员及相关领域的研究人员、数据库 DBA 及高等院校相关专业的学生阅读。

The Internals of PostgreSQL for Database Administrators and System Developers, Copyright © Hironobu Suzuki. Chinese translation Copyright © 2019 by Publishing House of Electronics Industry.

本书中文简体版专有版权由 Hironobu Suzuki 授予电子工业出版社，未经许可，不得以任何方式复制或者抄袭本书的任何部分。

图书在版编目（CIP）数据

PostgreSQL 指南：内幕探索 /（日）铃木启修（Hironobu Suzuki）著；冯若航，刘阳明，张文升译。
北京：电子工业出版社，2019.6

书名原文：The Internals of PostgreSQL for Database Administrators and System Developers
ISBN 978-7-121-35709-1

I. ①P… II. ①铃… ②冯… ③刘… ④张… III. ①关系数据库系统 IV. ①TP311.138

中国版本图书馆 CIP 数据核字（2018）第 281117 号

策划编辑：符隆美

责任编辑：张春雨

印 刷：山东华立印务有限公司

装 订：山东华立印务有限公司

出版发行：电子工业出版社

北京市海淀区万寿路 173 信箱 邮编：100036

开 本：787×980 1/16 印张：15.25 字数：350 千字

版 次：2019 年 6 月第 1 版

印 次：2019 年 6 月第 1 次印刷

定 价：79.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，
联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：010-51260888-819，faq@phei.com.cn。

· 作者简介 ·

Hironobu Suzuki (铃木启修)

毕业于北海道大学信息工程研究生院，获得信息工程硕士学位，曾在多家公司担任软件开发人员和技术经理/技术主管。在数据库和系统集成领域出版了7本书。

2010年—2016年担任日本PostgreSQL用户组的主任，连续7年组织了日本PostgreSQL技术研讨会，并担任日本2013年PostgreSQL大会的委员会主席。

推荐序

PostgreSQL 已获得 DB-Engines 排行榜 2017 年和 2018 年的“年度数据库”称号，发展速度迅猛，PostgreSQL 被广泛应用的主要原因在于 PostgreSQL 和 Oracle 一样，特别适合于非常复杂的企业应用场景，代替 Oracle 的位置。同时 PostgreSQL 的开源协议允许用户非常友好地应用在自用业务、对外商用业务中，没有法律风险，也不要求用户开源等。

很多大型企业应用 PostgreSQL 也有数年，例如阿里巴巴、邮储、平安、中兴、苏宁、亚信、探探等，长期的 PostgreSQL 数据库应用使得公司内部的数据库团队在 PostgreSQL 的管理、开发、内核等各方面的经验、人才都得到了大量的积累。

文升是我多年的好友，同时也是 PostgreSQL 核心组成员，为 PostgreSQL 中文技术社区做出了巨大的贡献。

2019 年，张文升及其技术团队又要出书了，这次带来的是《The Internals of PostgreSQL for database Administrators and System Developers》一书的中文翻译书籍，作者为日本 PostgreSQL 数据库专家 Suzuki，对数据库的集群、架构、SQL 处理、外部表接口、并发控制、垃圾回收、HOT、堆表、索引存储结构、WAL 日志、时间点恢复、流复制等原理进行了深入浅出的讲解，对期望了解 PostgreSQL 内部原理的开发者、管理员、架构师来说，无疑是一本非常好的入门书籍。

感谢译者对 PostgreSQL 社区的付出，期待这本图片的出版。
我从小就对中国世界和未知的事物充满好奇心，这些好奇心也引领我走入软件工程师的道路。现在，我已经在中国出版了第一本书，面向读者——一个关于在日本家庭与中国之间传递爱的德哥
我很高心终于可以向中国读者介绍一本。这本书只是一份很小的礼物，2019 年 3 月
诚挚地将它献给您。

2017 年，我有幸访问中国，亲眼目睹了中国令人惊叹的发展。对我而言，这一次非常珍

译者序

PostgreSQL 内部的工作原理，包括数据库对象的逻辑组织与物理实现、锁机制、缓冲区管理、WAL、备份及流复制。本书为 DBA 与系统开发者提供了一幅全景概念地图，有助于读者形成对数据库原理的整体认识，亦可作为深入学习 PostgreSQL 源代码的参考读物，对于理解数据库原理与 PostgreSQL 内部实现大有裨益。

一本适合数据库开发人员及相关领域的研究人士、数据库 DBA 及高等院校相关专业的学生阅读。

The Internals of PostgreSQL for Database Administrators and System Developers. Copyright © Hisanobu Suzuki. Chinese translation Copyright © 2019 by Publishing House of Electronics Industry.

本书由日本作者 Hisanobu Suzuki 编著，由电子工业出版社·机械工业出版社出版。

相信选择这本书的读者，大多已经对 PostgreSQL 有所了解。本书从 PostgreSQL 的整体架构展开，依次介绍了各个功能模块的来龙去脉，方便 DBA（数据库管理员）与数据库系统开发人员了解数据库内部原理、阅读学习 PostgreSQL 源码。

数据库是信息系统的核心组件，关系型数据库则是数据库皇冠上的明珠，而 PostgreSQL 的头衔是“世界上最先进的开源关系型数据库”。PostgreSQL 在各行各业的各种场景下都有着广泛应用。但是会用只是“知其然”，知道背后的原理才能“知其所以然”。理解数据库原理及其具体实现，能让架构师以最小复杂度的代价实现所需的功能，让程序员以最小复杂度的代价写出更加高效可靠的代码，让 DBA 在遇到“疑难杂症”时拥有精准的直觉与深刻的洞察。

数据库是一个博大精深的领域，存储、I/O、计算，无所不包。PostgreSQL 可以视作关系型数据库实现的典范，用 100 万行不到的 C 代码实现了功能如此丰富的软件系统，非常凝练。它的每一个功能模块都值得用一本甚至几本书的篇幅去介绍。本书虽限于篇幅而无法一一深入所有细节，但它为读者进一步深入理解 PostgreSQL 提供了一幅全局的概念地图。读者完全可以顺着各个章节的线索，以点破面，深入挖掘源码背后的设计思路。

我们偶然发现了本书的英文版本，读完之后感觉受益匪浅。看到这么好的书没有中文译本，实在是遗憾，遂萌生了翻译的念头。译者不才，愿为 PostgreSQL 在中国的发展贡献一份力量，但鉴于水平有限，翻译如有疏漏，还望读者海涵。

更多关于电子工业出版社图书有缺损问题，请你将图片发给我。姓名：吴昊，地址：北京市海淀区学院路 30 号，邮编：100083，电话：(010) 88258833，88258888。

凡此章 010-88258833，邮箱：wuhao@zjgchina.com.cn，微信号：wuhao_010-88258833。

本书咨询联系方式：010-51260885-619，E-mail：wuhao@zjgchina.com.cn。

作者序

数算木餐

2018 年 11 月

中国的 PostgreSQL 用户们：

你们好！

本书详细解释了 PostgreSQL 的内部工作细节，目标读者为 DBA 与系统开发人员。理解数据库内部机制很有挑战，愿本书能在你们精通 PostgreSQL 的道路上有所助益。

本书能出中文版，我真的感到非常高兴，我这样认为是有原因的。

首先，这是我写的书中第一本被翻译的。当自己的书出版时，心情愉悦自不必说，而自己的作品能被翻译出版，更是一件非常令人激动的事情。

其次，我收到了来自世界各地的电子邮件，请求将这本书翻译成各种语言。实际上，至少有一半的邮件来自中国。许多邮件提及“这本书对中国 PostgreSQL 用户而言很有帮助”，因此，我很高兴终于能对他们的要求做出回应。

而最重要的原因与我的家族史有关。先父曾在中国的哈尔滨市住过几年，此后他回到日本，我出生了。他患心脏病很长时间，因此在我的印象中，他的心情总是不好，除了提到一件事时。

当我还是孩子的时候，每晚睡觉前父亲都会给我和妹妹讲他在中国的经历。在讲这些经历时，他总是显得非常高兴，经常说中国人民帮助了他。

他的故事给我留下了深刻的印象，当然也影响了我的人生。因为这些故事，我从小就对这个世界和未知的事物有强烈的好奇心，而我的好奇心也引领我走入软件工程领域。现在，我已经在中国出版了第一本书。简而言之，这是一个关于在日本家庭与中国之间传递爱的故事。

我很高兴终于可以回报中国人民了。当然，这本书只是一份很小的礼物，然而我想充满感激之情地将它献给您。

2017 年，我有幸访问中国，亲眼目睹了中国令人惊叹的发展。对我而言，这是一次非常难

忘的经历。我希望能再去中国，如果有机会，我还想在中国工作。

无论如何，让我们享受 PostgreSQL 吧！



铃木启修

2018 年 11 月

相信选择这本书的读者，大多已经对 PostgreSQL 有所了解。本书从 PostgreSQL 的各方面入手展开，依次介绍了各个功能模块的来龙去脉，方便 DBA、数据库管理员与数据版系统开发人员快速上手。同时，书中还提供了大量的 PostgreSQL 相关知识，帮助读者更好地理解 PostgreSQL。希望本书能为读者带来帮助，同时也希望读者能够通过本书学习到更多的 PostgreSQL 知识。

致谢

本书的出版离不开电子工业出版社工作人员与译者们共同的努力与付出，对此我深表感激。

读者服务

轻松注册成为博文视点社区用户（www.broadview.com.cn），扫码直达本书页面。

- **提交勘误：**您对书中内容的修改意见可在 提交勘误 处提交，若被采纳，将获赠博文视点社区积分（在您购买电子书时，积分可用来抵扣相应金额）。
- **交流互动：**在页面下方 读者评论 处留下您的疑问或观点，与我们和其他读者一同学习交流。

页面入口：<http://www.broadview.com.cn/35709>



· 译者简介 ·

冯若航

PostgreSQL DBA，全栈工程师。
PostgreSQL与Golang布道者，活跃于
PostgreSQL技术社区，致力于企业
PostgreSQL培训与技术支持。

刘阳明

毕业于中国人民大学信息学院，
对数据库内核有浓厚兴趣，有丰富的
PostgreSQL管理与使用经验，坚信
PostgreSQL是一个还未被大众了解的优
秀数据库产品，坚持推动PostgreSQL的
发展。

张文升

参与了《PostgreSQL实战》一
书的写作，中国开源软件推进联盟
PostgreSQL分会核心成员之一。常年活
跃于PostgreSQL、MySQL、Redis等
开源技术社区，坚持推动PostgreSQL在
中国的发展，多次参与组织PostgreSQL
全国用户大会。近年来致力于企业
PostgreSQL培训与技术支持。

目录

第 1 章 数据库集簇、数据库和数据表	1
1.1 数据库集簇的逻辑结构	1
1.2 数据库集簇的物理结构	2
1.2.1 数据库集簇的布局	3
1.2.2 数据库布局	4
1.2.3 表和索引相关文件的布局	5
1.2.4 PostgreSQL 中表空间的布局	7
1.3 堆表文件的内部布局	8
1.4 读写元组的方式	11
1.4.1 写入堆元组	11
1.4.2 读取堆元组	12
第 2 章 进程和内存架构	14
2.1 进程架构	14
2.1.1 Postgres 服务器进程	15
2.1.2 后端进程	15
2.1.3 后台进程	16
2.2 内存架构	17
2.2.1 本地内存区域	17
2.2.2 共享内存区域	18
第 3 章 SQL 语句处理	19
3.1 SQL 语句的执行流程	19
3.1.1 提交 SQL 语句	20
3.1.2 处理 SQL 语句	20
3.1.3 执行 SQL 语句	21
3.1.4 游标处理	22
3.1.5 错误处理	23
3.1.6 语句缓存	24
3.1.7 语句重用	25
3.1.8 语句计划	26
3.1.9 语句解释	27
3.1.10 语句优化	28
3.1.11 语句重写	29
3.1.12 语句替换	30
3.1.13 语句转换	31
3.1.14 语句翻译	32
3.1.15 语句解析	33
3.1.16 语句分析	34
3.1.17 语句生成	35
3.1.18 语句生成	36
3.1.19 语句生成	37
3.1.20 语句生成	38
3.1.21 语句生成	39
3.1.22 语句生成	40
3.1.23 语句生成	41
3.1.24 语句生成	42
3.1.25 语句生成	43
3.1.26 语句生成	44
3.1.27 语句生成	45
3.1.28 语句生成	46
3.1.29 语句生成	47
3.1.30 语句生成	48
3.1.31 语句生成	49
3.1.32 语句生成	50
3.1.33 语句生成	51
3.1.34 语句生成	52
3.1.35 语句生成	53
3.1.36 语句生成	54
3.1.37 语句生成	55
3.1.38 语句生成	56
3.1.39 语句生成	57
3.1.40 语句生成	58
3.1.41 语句生成	59
3.1.42 语句生成	60
3.1.43 语句生成	61
3.1.44 语句生成	62
3.1.45 语句生成	63
3.1.46 语句生成	64
3.1.47 语句生成	65
3.1.48 语句生成	66
3.1.49 语句生成	67
3.1.50 语句生成	68
3.1.51 语句生成	69
3.1.52 语句生成	70
3.1.53 语句生成	71
3.1.54 语句生成	72
3.1.55 语句生成	73
3.1.56 语句生成	74
3.1.57 语句生成	75
3.1.58 语句生成	76
3.1.59 语句生成	77
3.1.60 语句生成	78
3.1.61 语句生成	79
3.1.62 语句生成	80
3.1.63 语句生成	81
3.1.64 语句生成	82
3.1.65 语句生成	83
3.1.66 语句生成	84
3.1.67 语句生成	85
3.1.68 语句生成	86
3.1.69 语句生成	87
3.1.70 语句生成	88
3.1.71 语句生成	89
3.1.72 语句生成	90
3.1.73 语句生成	91
3.1.74 语句生成	92
3.1.75 语句生成	93
3.1.76 语句生成	94
3.1.77 语句生成	95
3.1.78 语句生成	96
3.1.79 语句生成	97
3.1.80 语句生成	98
3.1.81 语句生成	99
3.1.82 语句生成	100
3.1.83 语句生成	101
3.1.84 语句生成	102
3.1.85 语句生成	103
3.1.86 语句生成	104
3.1.87 语句生成	105
3.1.88 语句生成	106
3.1.89 语句生成	107
3.1.90 语句生成	108
3.1.91 语句生成	109
3.1.92 语句生成	110
3.1.93 语句生成	111
3.1.94 语句生成	112
3.1.95 语句生成	113
3.1.96 语句生成	114
3.1.97 语句生成	115
3.1.98 语句生成	116
3.1.99 语句生成	117
3.1.100 语句生成	118
3.1.101 语句生成	119
3.1.102 语句生成	120
3.1.103 语句生成	121
3.1.104 语句生成	122
3.1.105 语句生成	123
3.1.106 语句生成	124
3.1.107 语句生成	125
3.1.108 语句生成	126
3.1.109 语句生成	127
3.1.110 语句生成	128
3.1.111 语句生成	129
3.1.112 语句生成	130
3.1.113 语句生成	131
3.1.114 语句生成	132
3.1.115 语句生成	133
3.1.116 语句生成	134
3.1.117 语句生成	135
3.1.118 语句生成	136
3.1.119 语句生成	137
3.1.120 语句生成	138
3.1.121 语句生成	139
3.1.122 语句生成	140
3.1.123 语句生成	141
3.1.124 语句生成	142
3.1.125 语句生成	143
3.1.126 语句生成	144
3.1.127 语句生成	145
3.1.128 语句生成	146
3.1.129 语句生成	147
3.1.130 语句生成	148
3.1.131 语句生成	149
3.1.132 语句生成	150
3.1.133 语句生成	151
3.1.134 语句生成	152
3.1.135 语句生成	153
3.1.136 语句生成	154
3.1.137 语句生成	155
3.1.138 语句生成	156
3.1.139 语句生成	157
3.1.140 语句生成	158
3.1.141 语句生成	159
3.1.142 语句生成	160
3.1.143 语句生成	161
3.1.144 语句生成	162
3.1.145 语句生成	163
3.1.146 语句生成	164
3.1.147 语句生成	165
3.1.148 语句生成	166
3.1.149 语句生成	167
3.1.150 语句生成	168
3.1.151 语句生成	169
3.1.152 语句生成	170
3.1.153 语句生成	171
3.1.154 语句生成	172
3.1.155 语句生成	173
3.1.156 语句生成	174
3.1.157 语句生成	175
3.1.158 语句生成	176
3.1.159 语句生成	177
3.1.160 语句生成	178
3.1.161 语句生成	179
3.1.162 语句生成	180
3.1.163 语句生成	181
3.1.164 语句生成	182
3.1.165 语句生成	183
3.1.166 语句生成	184
3.1.167 语句生成	185
3.1.168 语句生成	186
3.1.169 语句生成	187
3.1.170 语句生成	188
3.1.171 语句生成	189
3.1.172 语句生成	190
3.1.173 语句生成	191
3.1.174 语句生成	192
3.1.175 语句生成	193
3.1.176 语句生成	194
3.1.177 语句生成	195
3.1.178 语句生成	196
3.1.179 语句生成	197
3.1.180 语句生成	198
3.1.181 语句生成	199
3.1.182 语句生成	200
3.1.183 语句生成	201
3.1.184 语句生成	202
3.1.185 语句生成	203
3.1.186 语句生成	204
3.1.187 语句生成	205
3.1.188 语句生成	206
3.1.189 语句生成	207
3.1.190 语句生成	208
3.1.191 语句生成	209
3.1.192 语句生成	210
3.1.193 语句生成	211
3.1.194 语句生成	212
3.1.195 语句生成	213
3.1.196 语句生成	214
3.1.197 语句生成	215
3.1.198 语句生成	216
3.1.199 语句生成	217
3.1.200 语句生成	218
3.1.201 语句生成	219
3.1.202 语句生成	220
3.1.203 语句生成	221
3.1.204 语句生成	222
3.1.205 语句生成	223
3.1.206 语句生成	224
3.1.207 语句生成	225
3.1.208 语句生成	226
3.1.209 语句生成	227
3.1.210 语句生成	228
3.1.211 语句生成	229
3.1.212 语句生成	230
3.1.213 语句生成	231
3.1.214 语句生成	232
3.1.215 语句生成	233
3.1.216 语句生成	234
3.1.217 语句生成	235
3.1.218 语句生成	236
3.1.219 语句生成	237
3.1.220 语句生成	238
3.1.221 语句生成	239
3.1.222 语句生成	240
3.1.223 语句生成	241
3.1.224 语句生成	242
3.1.225 语句生成	243
3.1.226 语句生成	244
3.1.227 语句生成	245
3.1.228 语句生成	246
3.1.229 语句生成	247
3.1.230 语句生成	248
3.1.231 语句生成	249
3.1.232 语句生成	250
3.1.233 语句生成	251
3.1.234 语句生成	252
3.1.235 语句生成	253
3.1.236 语句生成	254
3.1.237 语句生成	255
3.1.238 语句生成	256
3.1.239 语句生成	257
3.1.240 语句生成	258
3.1.241 语句生成	259
3.1.242 语句生成	260
3.1.243 语句生成	261
3.1.244 语句生成	262
3.1.245 语句生成	263
3.1.246 语句生成	264
3.1.247 语句生成	265
3.1.248 语句生成	266
3.1.249 语句生成	267
3.1.250 语句生成	268
3.1.251 语句生成	269
3.1.252 语句生成	270
3.1.253 语句生成	271
3.1.254 语句生成	272
3.1.255 语句生成	273
3.1.256 语句生成	274
3.1.257 语句生成	275
3.1.258 语句生成	276
3.1.259 语句生成	277
3.1.260 语句生成	278
3.1.261 语句生成	279
3.1.262 语句生成	280
3.1.263 语句生成	281
3.1.264 语句生成	282
3.1.265 语句生成	283
3.1.266 语句生成	284
3.1.267 语句生成	285
3.1.268 语句生成	286
3.1.269 语句生成	287
3.1.270 语句生成	288
3.1.271 语句生成	289
3.1.272 语句生成	290
3.1.273 语句生成	291
3.1.274 语句生成	292
3.1.275 语句生成	293
3.1.276 语句生成	294
3.1.277 语句生成	295
3.1.278 语句生成	296
3.1.279 语句生成	297
3.1.280 语句生成	298
3.1.281 语句生成	299
3.1.282 语句生成	300
3.1.283 语句生成	301
3.1.284 语句生成	302
3.1.285 语句生成	303
3.1.286 语句生成	304
3.1.287 语句生成	305
3.1.288 语句生成	306
3.1.289 语句生成	307
3.1.290 语句生成	308
3.1.291 语句生成	309
3.1.292 语句生成	310
3.1.293 语句生成	311
3.1.294 语句生成	312
3.1.295 语句生成	313
3.1.296 语句生成	314
3.1.297 语句生成	315
3.1.298 语句生成	316
3.1.299 语句生成	317
3.1.300 语句生成	318
3.1.301 语句生成	319
3.1.302 语句生成	320
3.1.303 语句生成	321
3.1.304 语句生成	322
3.1.305 语句生成	323
3.1.306 语句生成	324
3.1.307 语句生成	325
3.1.308 语句生成	326
3.1.309 语句生成	327
3.1.310 语句生成	328
3.1.311 语句生成	329
3.1.312 语句生成	330
3.1.313 语句生成	331
3.1.314 语句生成	332
3.1.315 语句生成	333
3.1.316 语句生成	334
3.1.317 语句生成	335
3.1.318 语句生成	336
3.1.319 语句生成	337
3.1.320 语句生成	338
3.1.321 语句生成	339
3.1.322 语句生成	340
3.1.323 语句生成	341
3.1.324 语句生成	342
3.1.325 语句生成	343
3.1.326 语句生成	344
3.1.327 语句生成	345
3.1.328 语句生成	346
3.1.329 语句生成	347
3.1.330 语句生成	348
3.1.331 语句生成	349
3.1.332 语句生成	350
3.1.333 语句生成	351
3.1.334 语句生成	352
3.1.335 语句生成	353
3.1.336 语句生成	354
3.1.337 语句生成	355
3.1.338 语句生成	356
3.1.339 语句生成	357
3.1.340 语句生成	358
3.1.341 语句生成	359
3.1.342 语句生成	360
3.1.343 语句生成	361
3.1.344 语句生成	362
3.1.345 语句生成	363
3.1.346 语句生成	364
3.1.347 语句生成	365
3.1.348 语句生成	366
3.1.349 语句生成	367
3.1.350 语句生成	368
3.1.351 语句生成	369
3.1.352 语句生成	370
3.1.353 语句生成	371
3.1.354 语句生成	372
3.1.355 语句生成	373
3.1.356 语句生成	374
3.1.357 语句生成	375
3.1.358 语句生成	376
3.1.359 语句生成	377
3.1.360 语句生成	378
3.1.361 语句生成	379
3.1.362 语句生成	380
3.1.363 语句生成	381
3.1.364 语句生成	382
3.1.365 语句生成	383
3.1.366 语句生成	384
3.1.367 语句生成	385
3.1.368 语句生成	386
3.1.369 语句生成	387
3.1.370 语句生成	388
3.1.371 语句生成	389
3.1.372 语句生成	390
3.1.373 语句生成	391
3.1.374 语句生成	392
3.1.375 语句生成	393
3.1.376 语句生成	394
3.1.377 语句生成	395
3.1.378 语句生成	396
3.1.379 语句生成	397
3.1.380 语句生成	398
3.1.381 语句生成	399
3.1.382 语句生成	400
3.1.383	

第 3 章 查询处理	19
3.1 概览	20
3.1.1 解析器	20
3.1.2 分析器	22
3.1.3 重写器	24
3.1.4 计划器与执行器	25
3.2 单表查询的代价估计	27
3.2.1 顺序扫描	28
3.2.2 索引扫描	29
3.2.3 排序	36
3.3 创建单表查询的计划树	38
3.3.1 预处理	41
3.3.2 找出代价最小的访问路径	42
3.3.3 创建计划树	51
3.4 执行器如何工作	55
3.5 连接	57
3.5.1 嵌套循环连接	57
3.5.2 归并连接	63
3.5.3 散列连接	67
3.5.4 连接访问路径与连接节点	73
3.6 创建多表查询计划树	76
3.6.1 预处理	76
3.6.2 获取代价最小的路径	77
3.6.3 获取三表查询代价最小的路径	81
参考文献	83
第 4 章 外部数据包装器	84
4.1 概述	85
4.1.1 创建一棵查询树	86
4.1.2 连接至远程服务器	86
4.1.3 使用 EXPLAIN 命令创建计划树（可选）	87
4.1.4 逆解析	87

4.1.5	发送 SQL 命令并接收结果	88
4.2	postgres_fdw 的工作原理	90
4.2.1	多表查询	91
4.2.2	排序操作	97
4.2.3	聚合函数	98
第 5 章	并发控制	101
5.1	事务标识	103
5.2	元组结构	104
5.3	元组的增、删、改	106
5.3.1	插入	106
5.3.2	删除	107
5.3.3	更新	108
5.3.4	空闲空间映射	109
5.4	提交日志	110
5.4.1	事务状态	110
5.4.2	提交日志如何工作	110
5.4.3	提交日志的维护	111
5.5	事务快照	111
5.6	可见性检查规则	114
5.6.1	t_xmin 的状态为 ABORTED	115
5.6.2	t_xmin 的状态为 IN_PROGRESS	115
5.6.3	t_xmin 的状态为 COMMITTED	116
5.7	可见性检查	118
5.7.1	可见性检查的过程	118
5.7.2	PostgreSQL 可重复读等级中的幻读	122
5.8	防止丢失更新	122
5.8.1	并发 UPDATE 命令的行为	123
5.8.2	例子	125
5.9	可串行化快照隔离	127
5.9.1	SSI 实现的基本策略	127

第 5 章 PostgreSQL 的 SSI 实现	128
5.9.2 PostgreSQL 的 SSI 实现	128
5.9.3 SSI 的原理	129
5.9.4 假阳性的串行化异常	132
5.10 需要的维护进程	134
参考文献	136
第 6 章 清理过程	137
6.1 并发清理概述	138
6.1.1 第一部分	139
6.1.2 第二部分	140
6.1.3 第三部分	140
6.1.4 后续处理	141
6.2 可见性映射	141
6.3 冻结过程	142
6.3.1 惰性模式	142
6.3.2 迫切模式	143
6.3.3 改进迫切模式中的冻结过程	146
6.4 移除不必要的 CLOG 文件	147
6.5 自动清理守护进程	148
6.6 完整清理	148
第 7 章 堆内元组和仅索引扫描	153
7.1 堆内元组	153
7.1.1 没有 HOT 时的行更新	153
7.1.2 HOT 如何工作	154
7.2 仅索引扫描	157
第 8 章 缓冲区管理器	160
8.1 概览	161
8.2 缓冲区管理器的结构	163
8.2.1 缓冲表	164
8.2.2 缓冲区描述符	165

8.2.3 缓冲区描述符层	167
8.2.4 缓冲池	169
8.3 缓冲区管理器锁	169
8.3.1 缓冲表锁	170
8.3.2 缓冲区描述符相关的锁	170
8.4 缓冲区管理器的工作原理	174
8.4.1 访问存储在缓冲池中的页面	174
8.4.2 将页面从存储加载到空槽	175
8.4.3 将页面从存储加载到受害者缓冲池槽	176
8.4.4 页面替换算法：时钟扫描	178
8.5 环形缓冲区	180
8.6 脏页刷盘	181
第 9 章 WAL	182
9.1 概述	183
9.1.1 没有 WAL 的插入操作	183
9.1.2 插入操作与数据库恢复	184
9.1.3 整页写入	186
9.2 事务日志与 WAL 段文件	188
9.3 WAL 段文件的内部布局	190
9.4 WAL 记录的内部布局	191
9.4.1 WAL 记录首部部分	191
9.4.2 XLOG 记录的数据部分（9.4 及更低版本）	193
9.4.3 XLOG 记录的数据部分（9.5 及更高版本）	196
9.5 WAL 记录的写入	200
9.6 WAL 写入进程	203
9.7 PostgreSQL 中的检查点进程	203
9.7.1 检查点进程概述	204
9.7.2 pg_crond 文件	205
9.8 PostgreSQL 中的数据库恢复	206
9.9 WAL 段文件管理	209

9.9.1 WAL 段切换	209
9.9.2 WAL 段管理（9.5 及更高版本）	209
9.9.3 WAL 段管理（9.4 及更低版本）	211
9.10 持续归档与归档日志	212
第 10 章 基础备份与时间点恢复	214
10.1 基础备份	215
10.1.1 pg_start_backup	215
10.1.2 pg_stop_backup	217
10.2 时间点恢复（PITR）的工作原理	217
10.3 时间线与时间线历史文件	220
10.3.1 时间线标识	220
10.3.2 时间线历史文件	221
10.4 时间点恢复与时间线历史文件	222
第 11 章 流复制	224
11.1 流复制的启动	225
11.2 如何实施流复制	227
11.2.1 主从间的通信	227
11.2.2 发生故障时的行为	229
11.3 管理多个备库	229
11.3.1 同步优先级与同步状态	229
11.3.2 主库如何管理多个备库	230
11.3.3 发生故障时的行为	231
11.4 备库的故障检测	232
第 12 章 异常处理	257
12.1 错误	259
12.1.1 错误的种类	259
12.1.2 错误的处理	260
12.2 错误消息的生成	263
12.2.1 错误消息	263
12.2.2 错误消息体	264
第 13 章 PostgreSQL 8.9	163
13.1 新特性	163
13.1.1 PostgreSQL 8.9	163
13.2 语义	164
13.2.1 语义	164
13.2.2 语义	165

第 1 章

数据库集群、数据库和数据表

第 1 章和第 2 章会简单介绍一些 PostgreSQL 的基础知识，以帮助读者理解后续章节的内容。本章包括以下几个主题：

- 数据库集群的逻辑结构
- 数据库集群的物理结构
- 堆表文件的内部布局
- 从表中读写数据的方式

如果你已经熟悉这些内容，可以跳过本章。

译者注：本书中的 database cluster 与 PostgreSQL 中文文档统一，译为数据库集群。与高可用数据库集群不同，这里的集群表示多个逻辑的数据库在同一个数据库实例中。

1.1 数据库集群的逻辑结构

数据库集群（database cluster）是一组数据库（database）的集合，由一个 PostgreSQL 服务器管理。读者第一次听到这个定义也许会产生疑惑，其实 PostgreSQL 中的术语“数据库集群”，并非意味着“一组数据库服务器”。一个 PostgreSQL 服务器只会在单机上运行并管理单个数据库集群。

图 1.1 展示了一个数据库集群的逻辑结构。数据库是数据库对象（database object）的集合。在关系型数据库理论中，数据库对象用于存储或引用数据的数据结构。（堆）表就是一个典型的

例子，还有更多对象，例如索引、序列、视图、函数等。在 PostgreSQL 中，数据库本身也是数据库对象，并在逻辑上彼此分离。所有其他的数据库对象（例如表、索引等）都归属于各自相应的数据库。

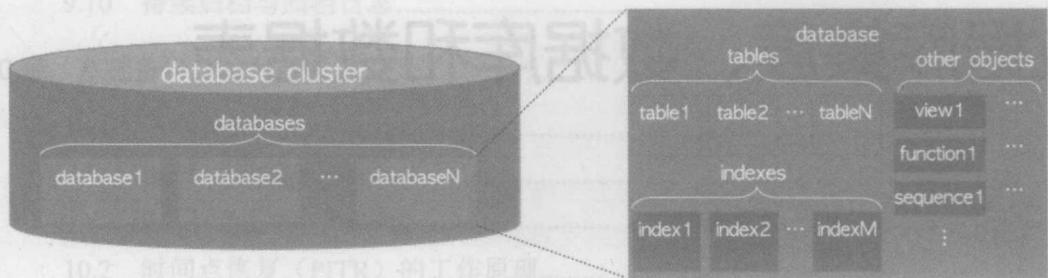


图 1.1 数据库集簇的逻辑结构

在 PostgreSQL 内部，所有的数据库对象都通过相应的对象标识符（object identifier, oid）进行管理，这些标识符是无符号的 4 字节整型。数据库对象与相应 oid 之间的关系存储在对应的系统目录中，依具体的对象类型而异。例如数据库和堆表对象的 oid 分别存储在 pg_database 和 pg_class 中，因此，当你希望找出 oid 时，可以执行以下查询：

```
sampledb=# SELECT datname, oid FROM pg_database WHERE datname = 'sampledb';
datname | oid
-----+-----
sampledb | 16384
(1 row)

sampledb=# SELECT relname, oid FROM pg_class WHERE relname = 'sampltbl';
relname | oid
-----+-----
sampltbl | 18740
(1 row)
```

1.2 数据库集簇的物理结构

数据库集簇在本质上就是一个文件目录，即基础目录，包含着一系列子目录与文件。执行 initdb 命令会在指定目录下创建基础目录，从而初始化一个新的数据库集簇。通常基础目录的路径会被配置到环境变量 PGDATA 中，但这并不是必要的。

图 1.2 展示了一个 PostgreSQL 数据库集簇的例子。base 子目录中的每一个子目录都对应一