

УДК 004.65  
ББК 32.972.134  
Р59

**Рогов Е. В.**

Р59 PostgreSQL изнутри. — М.: ДМК Пресс, 2022. — 660 с.  
ISBN 978-5-93700-122-1

В книге рассматривается внутреннее устройство СУБД PostgreSQL: детали реализации многоверсионности и изоляции на основе снимков данных, включая процедуру очистки неактуальных версий строк; буферный кеш и журнал предзаписи; использование блокировок различных уровней; планирование и выполнение SQL-запросов; принципы расширяемости и особенности имеющихся индексных методов доступа. Большое внимание уделяется возможностям, предоставляемым для самостоятельного изучения механизмов функционирования PostgreSQL.

Сайт книги: <https://postgrespro.ru/education/books/internals>.

Для администраторов и программистов.

УДК 004.65  
ББК 32.972.134

ISBN 978-5-6041193-9-6  
ISBN 978-5-93700-122-1

© Текст, оформление, ООО «ППГ», 2022  
© Издание, ДМК Пресс, 2022

# Содержание

<b>О книге</b>	<b>17</b>
<b>Глава 1. Введение</b>	<b>23</b>
1.1. Организация данных . . . . .	23
Базы данных . . . . .	23
Системный каталог . . . . .	24
Схемы . . . . .	25
Табличные пространства . . . . .	26
Отношения . . . . .	27
Слои и файлы . . . . .	28
Страницы . . . . .	33
TOAST . . . . .	33
1.2. Процессы и память . . . . .	39
1.3. Клиенты и клиент-серверный протокол . . . . .	41
 <b>Часть I. Изоляция и многоверсионность</b>	 <b>45</b>
<b>Глава 2. Изоляция</b>	<b>47</b>
2.1. Согласованность . . . . .	47
2.2. Уровни изоляции и аномалии в стандарте SQL . . . . .	49
Потерянное обновление . . . . .	50
Грязное чтение и Read Uncommitted . . . . .	50
Неповторяющееся чтение и Read Committed . . . . .	51
Фантомное чтение и Repeatable Read . . . . .	51
Отсутствие аномалий и Serializable . . . . .	52
Почему именно эти аномалии? . . . . .	52
2.3. Уровни изоляции в PostgreSQL . . . . .	54
Read Committed . . . . .	55
Repeatable Read . . . . .	64
Serializable . . . . .	70
2.4. Какой уровень изоляции использовать? . . . . .	73

<b>Глава 3. Страницы и версии строк</b>	<b>75</b>
3.1. Структура страниц . . . . .	75
Заголовок страницы . . . . .	75
Специальная область . . . . .	76
Версии строк . . . . .	76
Указатели на версии строк . . . . .	77
Свободное место . . . . .	78
3.2. Структура версий строк . . . . .	78
3.3. Выполнение операций над версиями строк . . . . .	80
Вставка . . . . .	81
Фиксация . . . . .	85
Удаление . . . . .	87
Отмена . . . . .	88
Обновление . . . . .	88
Индексы . . . . .	89
3.4. TOAST . . . . .	90
3.5. Виртуальные транзакции . . . . .	91
3.6. Вложенные транзакции . . . . .	92
Точки сохранения . . . . .	92
Ошибки и атомарность операций . . . . .	94
<b>Глава 4. Снимки данных</b>	<b>97</b>
4.1. Что такое снимок данных . . . . .	97
4.2. Видимость версий строк в снимке . . . . .	98
4.3. Из чего состоит снимок . . . . .	99
4.4. Видимость собственных изменений . . . . .	104
4.5. Горизонт транзакции . . . . .	105
4.6. Снимок данных для системного каталога . . . . .	108
4.7. Экспорт снимка данных . . . . .	109
<b>Глава 5. Внутривстраничная очистка и hot-обновления</b>	<b>111</b>
5.1. Внутривстраничная очистка . . . . .	111
5.2. Hot-обновления . . . . .	115
5.3. Внутривстраничная очистка при hot-обновлениях . . . . .	118
5.4. Разрыв hot-цепочки . . . . .	120
5.5. Внутривстраничная очистка индексов . . . . .	122

<b>Глава 6. Очистка и автоочистка</b>	<b>124</b>
6.1. Очистка вручную . . . . .	124
6.2. Еще раз о горизонте базы данных . . . . .	127
6.3. Этапы выполнения очистки . . . . .	130
Сканирование таблицы . . . . .	130
Очистка индексов . . . . .	130
Очистка таблицы . . . . .	131
Усечение таблицы . . . . .	132
6.4. Анализ . . . . .	133
6.5. Автоматическая очистка и анализ . . . . .	133
Устройство автоочистки . . . . .	134
Какие таблицы требуют очистки . . . . .	135
Какие таблицы требуют анализа . . . . .	137
Автоочистка в действии . . . . .	138
6.6. Регулирование нагрузки . . . . .	142
Управление интенсивностью обычной очистки . . . . .	143
Управление интенсивностью автоочистки . . . . .	143
6.7. Мониторинг очистки . . . . .	144
Отслеживание выполнения ручной очистки . . . . .	145
Отслеживание выполнения автоочистки . . . . .	148
<b>Глава 7. Заморозка</b>	<b>149</b>
7.1. Переполнение счетчика транзакций . . . . .	149
7.2. Заморозка версий и правила видимости . . . . .	150
7.3. Управление заморозкой . . . . .	153
Минимальный возраст для заморозки . . . . .	154
Возраст для «агрессивной» заморозки . . . . .	156
Возраст для аварийного срабатывания автоочистки . . . . .	158
Возраст для приоритетного режима заморозки . . . . .	160
7.4. Заморозка вручную . . . . .	160
Очистка с заморозкой . . . . .	160
Заморозка при загрузке . . . . .	161
<b>Глава 8. Перестроение таблиц и индексов</b>	<b>163</b>
8.1. Полная очистка . . . . .	163
Необходимость . . . . .	163
Оценка плотности информации . . . . .	164

## Содержание

Заморозка . . . . .	168
8.2. Другие способы перестроения . . . . .	169
Аналоги полной очистки . . . . .	169
Перестроение без долгих блокировок . . . . .	170
8.3. Профилактика . . . . .	171
Читающие запросы . . . . .	171
Обновление данных . . . . .	172
<b>Часть II. Буферный кеш и журнал</b>	<b>175</b>
<b>Глава 9. Буферный кеш</b>	<b>177</b>
9.1. Кеширование . . . . .	177
9.2. Устройство буферного кеша . . . . .	178
9.3. Попадание в кеш . . . . .	180
9.4. Промах кеша . . . . .	184
Поиск буфера и вытеснение . . . . .	186
9.5. Массовое вытеснение . . . . .	188
9.6. Настройка размера . . . . .	191
9.7. Прогрев кеша . . . . .	194
9.8. Локальный кеш . . . . .	196
<b>Глава 10. Журнал предзаписи</b>	<b>198</b>
10.1. Журналирование . . . . .	198
10.2. Устройство журнала . . . . .	200
Логическая структура . . . . .	200
Физическая структура . . . . .	203
10.3. Контрольная точка . . . . .	205
10.4. Восстановление . . . . .	209
10.5. Фоновая запись . . . . .	213
10.6. Настройка . . . . .	213
Настройка контрольной точки . . . . .	213
Настройка фоновой записи . . . . .	216
Мониторинг . . . . .	217
<b>Глава 11. Режимы журнала</b>	<b>220</b>
11.1. Производительность . . . . .	220

11.2. Надежность . . . . .	224
Кеширование . . . . .	225
Повреждение данных . . . . .	226
Неатомарность записи . . . . .	228
11.3. Уровни журнала . . . . .	232
Minimal . . . . .	232
Replica . . . . .	234
Logical . . . . .	237
 <b>Часть III. Блокировки</b>	 <b>239</b>
 <b>Глава 12. Блокировки отношений</b>	 <b>241</b>
12.1. Общие сведения о блокировках . . . . .	241
12.2. Тяжелые блокировки . . . . .	244
12.3. Блокировки номеров транзакций . . . . .	246
12.4. Блокировки отношений . . . . .	247
12.5. Очередь ожидания . . . . .	250
 <b>Глава 13. Блокировки строк</b>	 <b>254</b>
13.1. Устройство . . . . .	254
13.2. Режимы блокировки строки . . . . .	255
Исключительные режимы . . . . .	255
Разделяемые режимы . . . . .	257
13.3. Мультитранзакции . . . . .	258
13.4. Очередь ожидания . . . . .	260
Исключительные режимы . . . . .	260
Разделяемые режимы . . . . .	267
13.5. Блокировка без ожидания . . . . .	270
13.6. Взаимоблокировки . . . . .	272
Взаимоблокировка при обновлении строк . . . . .	274
Взаимоблокировка двух команд UPDATE . . . . .	275
 <b>Глава 14. Блокировки разных объектов</b>	 <b>279</b>
14.1. Блокировки не-отношений . . . . .	279
14.2. Блокировки расширения отношения . . . . .	281
14.3. Блокировки страниц . . . . .	282

## Содержание

14.4. Рекомендательные блокировки . . . . .	282
14.5. Предикатные блокировки . . . . .	284
<b>Глава 15. Блокировки в памяти</b>	<b>291</b>
15.1. Спин-блокировки . . . . .	291
15.2. Легкие блокировки . . . . .	292
15.3. Примеры . . . . .	292
Буферный кеш . . . . .	292
Буферы журнала предзаписи . . . . .	294
15.4. Мониторинг ожиданий . . . . .	295
15.5. Семплирование . . . . .	297
<b>Часть IV. Выполнение запросов</b>	<b>301</b>
<b>Глава 16. Этапы выполнения запросов</b>	<b>303</b>
16.1. Демонстрационная база данных . . . . .	303
16.2. Протокол простых запросов . . . . .	306
Разбор . . . . .	306
Трансформация . . . . .	308
Планирование . . . . .	310
Исполнение . . . . .	319
16.3. Протокол расширенных запросов . . . . .	321
Подготовка . . . . .	321
Привязка параметров . . . . .	322
Планирование и исполнение . . . . .	323
Получение результатов . . . . .	326
<b>Глава 17. Статистика</b>	<b>327</b>
17.1. Базовая статистика . . . . .	327
17.2. Неопределенные значения . . . . .	331
17.3. Уникальные значения . . . . .	332
17.4. Наиболее частые значения . . . . .	334
17.5. Гистограмма . . . . .	337
17.6. Статистика для нескалярных типов данных . . . . .	341
17.7. Средний размер поля . . . . .	342
17.8. Корреляция . . . . .	342

17.9. Статистика по выражению . . . . .	343
Расширенная статистика по выражению . . . . .	344
Статистика для индекса по выражению . . . . .	345
17.10. Многовариантная статистика . . . . .	346
Функциональные зависимости между столбцами . . . . .	346
Многовариантное число различных значений . . . . .	348
Многовариантные списки частых значений . . . . .	350
<b>Глава 18. Табличные методы доступа</b>	<b>352</b>
18.1. Подключаемые движки хранения . . . . .	352
18.2. Последовательное сканирование . . . . .	354
Оценка стоимости . . . . .	355
18.3. Параллельные планы выполнения . . . . .	359
18.4. Параллельное последовательное сканирование . . . . .	360
Оценка стоимости . . . . .	361
18.5. Ограничения параллельного выполнения . . . . .	365
Количество рабочих процессов . . . . .	365
Нераспараллеливаемые запросы . . . . .	369
Ограниченно распараллеливаемые запросы . . . . .	370
<b>Глава 19. Индексные методы доступа</b>	<b>375</b>
19.1. Индексы и расширяемость . . . . .	375
19.2. Классы и семейства операторов . . . . .	378
Класс операторов . . . . .	378
Семейство операторов . . . . .	383
19.3. Интерфейс механизма индексирования . . . . .	385
Свойства метода доступа . . . . .	386
Свойства индекса . . . . .	390
Свойства столбцов . . . . .	391
<b>Глава 20. Индексное сканирование</b>	<b>395</b>
20.1. Простое индексное сканирование . . . . .	395
Оценка стоимости . . . . .	396
Хороший случай: высокая корреляция . . . . .	397
Плохой случай: низкая корреляция . . . . .	400
20.2. Сканирование только индекса . . . . .	403
Include-индексы . . . . .	406



## Содержание

20.3. Сканирование по битовой карте . . . . .	408
Точность карты . . . . .	409
Действия с битовыми картами . . . . .	411
Оценка стоимости . . . . .	412
20.4. Параллельные версии индексного сканирования . . . . .	416
20.5. Сравнение методов доступа . . . . .	418
<b>Глава 21. Вложенный цикл</b>	<b>420</b>
21.1. Виды и способы соединений . . . . .	420
21.2. Соединение вложенным циклом . . . . .	422
Декартово произведение . . . . .	422
Параметризованное соединение . . . . .	426
Кеширование (мемоизация) строк . . . . .	431
Внешние соединения . . . . .	434
Анти- и полусоединения . . . . .	436
Не эквисоединения . . . . .	438
Параллельный режим . . . . .	439
<b>Глава 22. Хеширование</b>	<b>441</b>
22.1. Соединение хешированием . . . . .	441
Однопроходное соединение хешированием . . . . .	441
Двухпроходное соединение хешированием . . . . .	447
Динамические корректировки плана . . . . .	450
Соединение хешированием в параллельных планах . . . . .	454
Параллельное однопроходное хеш-соединение . . . . .	455
Параллельное двухпроходное хеш-соединение . . . . .	457
Модификации . . . . .	460
22.2. Группировка и уникальные значения . . . . .	463
<b>Глава 23. Сортировка и слияние</b>	<b>466</b>
23.1. Соединение слиянием . . . . .	466
Слияние отсортированных наборов . . . . .	466
Параллельный режим . . . . .	470
Модификации . . . . .	471
23.2. Сортировка . . . . .	472
Быстрая сортировка . . . . .	474
Частичная пирамидальная сортировка . . . . .	475

Внешняя сортировка . . . . .	477
Инкрементальная сортировка . . . . .	481
Параллельный режим . . . . .	483
23.3. Группировка и уникальные значения . . . . .	486
23.4. Сравнение способов соединения . . . . .	488
<b>Часть V. Типы индексов</b>	<b>491</b>
<b>Глава 24. Хеш-индекс</b>	<b>493</b>
24.1. Общий принцип . . . . .	493
24.2. Страничная организация . . . . .	494
24.3. Класс операторов . . . . .	501
24.4. Свойства . . . . .	502
Свойства метода доступа . . . . .	502
Свойства индекса . . . . .	503
Свойства столбцов . . . . .	504
<b>Глава 25. В-дерево</b>	<b>505</b>
25.1. Общий принцип . . . . .	505
25.2. Поиск и вставка . . . . .	506
Поиск по равенству . . . . .	506
Поиск по неравенству . . . . .	508
Поиск по диапазону . . . . .	509
Вставка . . . . .	509
25.3. Страничная организация . . . . .	511
Компактное хранение дубликатов . . . . .	515
Компактное хранение внутренних индексных записей . . . . .	517
25.4. Класс операторов . . . . .	518
Семантика сравнения . . . . .	518
Сортировка и составные индексы . . . . .	524
25.5. Свойства . . . . .	529
Свойства метода доступа . . . . .	529
Свойства индекса . . . . .	530
Свойства столбцов . . . . .	531

<b>Глава 26. Индекс GiST</b>	<b>532</b>
26.1. Общий принцип . . . . .	532
26.2. R-дерево для точек . . . . .	534
Страничная организация . . . . .	537
Класс операторов . . . . .	537
Поиск вхождения в область . . . . .	539
Поиск ближайших соседей . . . . .	542
Вставка . . . . .	547
Ограничение исключения . . . . .	548
Свойства . . . . .	551
26.3. RD-дерево для полнотекстового поиска . . . . .	554
Про полнотекстовый поиск . . . . .	554
Индексация tsvector . . . . .	555
Свойства . . . . .	563
26.4. Другие типы данных . . . . .	563
<b>Глава 27. Индекс SP-GiST</b>	<b>566</b>
27.1. Общий принцип . . . . .	566
27.2. Дерево квадрантов для точек . . . . .	568
Класс операторов . . . . .	569
Страничная организация . . . . .	573
Поиск . . . . .	574
Вставка . . . . .	575
Свойства . . . . .	578
27.3. К-мерные деревья для точек . . . . .	580
27.4. Префиксное дерево для строк . . . . .	582
Класс операторов . . . . .	583
Поиск . . . . .	584
Вставка . . . . .	586
Свойства . . . . .	587
27.5. Другие типы данных . . . . .	588
<b>Глава 28. Индекс GIN</b>	<b>590</b>
28.1. Общий принцип . . . . .	590
28.2. Индекс для полнотекстового поиска . . . . .	591
Страничная организация . . . . .	593
Класс операторов . . . . .	595

Поиск . . . . .	597
Частые и редкие лексемы . . . . .	598
Вставка . . . . .	602
Ограничение выборки . . . . .	604
Свойства . . . . .	605
Ограничения GIN и RUM-индекс . . . . .	607
28.3. Индекс для триграмм . . . . .	608
28.4. Индекс для массивов . . . . .	610
28.5. Индекс для JSON . . . . .	614
Класс операторов jsonb_ops . . . . .	614
Класс операторов jsonb_path_ops . . . . .	617
28.6. Другие типы данных . . . . .	619
<b>Глава 29. Индекс BRIN</b>	<b>620</b>
29.1. Общий принцип . . . . .	620
29.2. Пример . . . . .	621
29.3. Страничная организация . . . . .	623
29.4. Поиск . . . . .	625
29.5. Обновление сводной информации . . . . .	626
Вставка значений . . . . .	626
Обобщение зоны . . . . .	627
29.6. Диапазоны значений (minmax) . . . . .	628
Выбор столбцов для индексирования . . . . .	629
Размер зоны и эффективность поиска . . . . .	630
Свойства . . . . .	634
29.7. Мультидиапазоны значений (minmax-multi) . . . . .	637
29.8. Охватывающие значения (inclusion) . . . . .	640
29.9. Фильтры Блума (bloom) . . . . .	643
<b>Заключение</b>	<b>648</b>
<b>Предметный указатель</b>	<b>649</b>