

Оглавление

О книге	17
Глава 1. Введение	23
Часть I. Изоляция и многоверсионность	45
Глава 2. Изоляция	47
Глава 3. Страницы и версии строк	75
Глава 4. Снимки данных	97
Глава 5. Внутространичная очистка и hot-обновления	111
Глава 6. Очистка и автоочистка	124
Глава 7. Заморозка	149
Глава 8. Перестроение таблиц и индексов	163
Часть II. Буферный кеш и журнал	175
Глава 9. Буферный кеш	177
Глава 10. Журнал предзаписи	198
Глава 11. Режимы журнала	220
Часть III. Блокировки	239
Глава 12. Блокировки отношений	241
Глава 13. Блокировки строк	254
Глава 14. Блокировки разных объектов	279
Глава 15. Блокировки в памяти	291
Часть IV. Выполнение запросов	301
Глава 16. Этапы выполнения запросов	303
Глава 17. Статистика	327
Глава 18. Табличные методы доступа	352
Глава 19. Индексные методы доступа	375
Глава 20. Индексное сканирование	395
Глава 21. Вложенный цикл	420

Оглавление

Глава 22. Хеширование	441
Глава 23. Сортировка и слияние	466
Часть V. Типы индексов	491
Глава 24. Хеш-индекс	493
Глава 25. B-дерево	505
Глава 26. Индекс GiST	532
Глава 27. Индекс SP-GiST	566
Глава 28. Индекс GIN	590
Глава 29. Индекс BRIN	620
Заключение	648
Предметный указатель	649

Содержание

О книге	17
Глава 1. Введение	23
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
 Часть I. Изоляция и многоверсионность	 45
Глава 2. Изоляция	47
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

Глава 3. Страницы и версии строк	75
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
Глава 4. Снимки данных	97
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
Глава 5. Внутривстраничная очистка и hot-обновления	111
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

Глава 6. Очистка и автоочистка	124
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
Глава 7. Заморозка	149
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
Глава 8. Перестроение таблиц и индексов	163
8.1. Полная очистка	163
Необходимость	163
Оценка плотности информации	164

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
Часть II. Буферный кеш и журнал	175
Глава 9. Буферный кеш	177
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
Глава 10. Журнал предзаписи	198
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
Глава 11. Режимы журнала	220
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
Часть III. Блокировки	239
Глава 12. Блокировки отношений	241
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
Глава 13. Блокировки строк	254
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
Глава 14. Блокировки разных объектов	279
14.1. Блокировки не-отношений	279
14.2. Блокировки расширения отношения	281
14.3. Блокировки страниц	282

14.4. Рекомендательные блокировки	282
14.5. Предикатные блокировки	284
Глава 15. Блокировки в памяти	291
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
Часть IV. Выполнение запросов	301
Глава 16. Этапы выполнения запросов	303
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
Глава 17. Статистика	327
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
Глава 18. Табличные методы доступа	352
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
Глава 19. Индексные методы доступа	375
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
Глава 20. Индексное сканирование	395
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

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
Глава 21. Вложенный цикл	420
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
Глава 22. Хеширование	441
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
Глава 23. Сортировка и слияние	466
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
Часть V. Типы индексов	491
Глава 24. Хеш-индекс	493
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
Глава 25. В-дерево	505
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

Глава 26. Индекс GiST	532
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
Глава 27. Индекс SP-GiST	566
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
Глава 28. Индекс GIN	590
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
Глава 29. Индекс BRIN	620
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
Заключение	648
Предметный указатель	649

О книге

— До чего же это все просто! — воскликнул Шпунтик. — А я где-то читал, что писателю нужен какой-то вымысел, замысел...

— Э, замысел! — нетерпеливо перебил его Смекайло. — Это только в книгах так пишется, что нужен замысел, а попробуй задумай что-нибудь, когда все уже и без тебя задумано! Что ни возьми — все уже было.

Николай Носов, *Приключения Незнайки и его друзей*

Для кого эта книга

Эта книга для тех, кого не устраивает работать с базой данных как с черным ящиком. Если вы любознательны, не довольствуетесь авторитетными советами и хотите во всем разобраться сами — нам по пути.

Я ориентируюсь на читателей, имеющих определенный опыт использования PostgreSQL и хотя бы в общих чертах представляющих себе, что к чему. Для совсем новичков текст будет тяжеловат. Например, я ни слова не скажу о том, как устанавливать сервер, вводить команды в `psql` или изменять конфигурационные параметры.

Надеюсь, что книга будет полезной и тем, кто хорошо знаком с устройством другой СУБД, но переходит на PostgreSQL и хочет разобраться в отличиях. Несколько лет назад такая книга сэкономила бы мне много времени. Именно поэтому я ее в конце концов и написал.

Чего нет в книге

Эта книга — не сборник рецептов. На все случаи жизни готовых решений не напасешься, а понимание внутренней механики сложной системы дает

возможность критически переосмысливать чужой опыт и делать свои собственные выводы. Поэтому я и объясняю такие подробности устройства, знание которых на первый взгляд не имеет практического смысла.

Но эта книга и не учебник. Она углубляется в одни области (более интересные мне самому) и обходит стороной другие. Если вы изучаете SQL, обратите внимание на учебник Евгения Моргунова *PostgreSQL. Основы языка SQL*¹, а необходимый теоретический фундамент даст книга Бориса Новикова *Основы технологий баз данных*².

Называться справочником эта книга тоже не претендует. Я старался быть точным, но у меня не было цели заменить книгой документацию, поэтому я легко опускал непринципиальные на мой взгляд подробности. В любой непонятной ситуации читайте документацию.

Еще эта книга не учит разрабатывать ядро PostgreSQL. Я не предполагаю у читателя знания языка C и ориентируюсь на администраторов и прикладных разработчиков. Хотя и ссылаюсь постоянно на исходный код, из которого можно узнать столько подробностей, сколько душе угодно, и даже больше.

Что в книге есть

Во вводной главе без особых деталей я даю основные понятия, на которые опирается все дальнейшее повествование. Я предполагаю, что вы не почерпнете из этой главы практически ничего нового, но все-таки включаю ее для полноты картины. К тому же она может пригодиться тем, кто переходит с других СУБД.

Первая часть книги посвящена вопросам согласованности и изоляции, которые я сперва рассматриваю с позиции пользователя (какие уровни изоляции существуют и чем это грозит), а затем с точки зрения внутреннего устройства. Для этого мне приходится погрузиться в детали реализации многоверсионности и изоляции на основе снимков данных. Особенно много внимания требует процедура очистки неактуальных версий строк.

¹ postgrespro.ru/education/books/sqlprimer.

² postgrespro.ru/education/books/dbtech.

Во второй части я рассматриваю буферный кеш и механизм, позволяющий восстанавливать согласованность после сбоя, — журнал предзаписи.

В третьей части детально разбирается устройство и использование блокировок разных уровней: легких блокировок для оперативной памяти, тяжелых блокировок для отношений, блокировок табличных строк.

Четвертая часть объясняет, как сервер планирует и выполняет SQL-запросы. Я рассказываю, какие есть способы доступа к данным, какие применяются методы соединения и как используется статистическая информация.

В пятой части обсуждение индексов, сводившееся ранее к B-деревьям, добирается и до остальных методов доступа. Сначала я рассматриваю общие принципы расширяемости, устанавливающие границы между ядром системы индексирования, индексными методами доступа и типами данных (что требует введения понятия классов операторов), а затем подробно останавливаюсь на особенностях каждого из имеющихся методов.

В состав PostgreSQL входит масса «интроспективных» расширений, которые не нужны для обычной работы, но дают возможность заглянуть во внутреннюю жизнь сервера. В книге используются многие из них. Кроме того, что эти расширения позволяют лучше изучить устройство сервера, они могут облегчить диагностику в сложных случаях.

Обозначения

Я пытался писать книгу так, чтобы ее можно было читать последовательно, страница за страницей. Но всю правду не получается раскрыть сразу, и к одной и той же теме приходится возвращаться несколько раз. Если бы я каждый раз писал «это будет рассмотрено позже», книга сильно увеличилась бы в размере, поэтому в таких случаях я ставлю на полях номер страницы, на которой тема развивается дальше. Такой же номер, ведущий назад, отсылает к месту в книге, где уже что-то говорилось о предмете обсуждения. с. 19

Текст книги и все примеры актуальны для PostgreSQL 14. Некоторые абзацы имеют на полях отметку о номере версии. Это означает, что сказанное справедливо для версий PostgreSQL, начиная с указанной, а более ранние v. 14

версии либо вовсе не имели описанной возможности, либо были устроены как-то иначе. Такие пометки могут оказаться полезными для тех, кто еще не обновил систему до последнего выпуска.

Также на полях указываются значения по умолчанию для обсуждаемых параметров. Сами параметры (как обычные, так и параметры хранения) выделены курсивом: *work_mem*.

В сносках я постоянно ссылаюсь на первоисточники. Их несколько, и на первом месте стоит кладезь полезной информации — документация¹. Являясь органичной частью проекта, она всегда поддерживается в актуальном состоянии самими разработчиками. Но главный первоисточник — безусловно, исходный код². Удивительно, на какое количество вопросов можно найти ответы просто в комментариях и файлах README, даже не владея языком С. Реже я ссылаюсь на записи коммитфеста³: в переписках `pgsql-hackers` всегда можно проследить историю изменений и понять логику принятых разработчиками решений, но ценой чтения огромного массива обсуждений.

Лирические отступления и замечания, которые могут увести в сторону от основной мысли, но которые я не удержался и вставил в книгу, выделены так, чтобы их можно было пропустить.

Конечно, в книге много фрагментов кода, в основном на языке SQL. Код показан с приглашением `=>`; если необходимо, то следом за ним приведен и ответ сервера:

```
=> SELECT now();
               now
-----
2022-01-04 17:45:15.108324+03
(1 row)
```

Если аккуратно повторять все приведенные команды в PostgreSQL 14, должен получиться такой же результат (конечно, с точностью до номеров транзакций и прочих несущественных деталей). Во всяком случае, весь код в книге — результат выполнения скрипта, содержащего ровно эти команды.

¹ postgrespro.ru/docs/postgresql/14/index.

² git.postgresql.org/gitweb/?p=postgresql.git;a=summary.

³ commitfest.postgresql.org.

Когда требуется показать одновременную работу нескольких транзакций, код, выполняющийся в другом сеансе, выделен отступом и отчеркиванием:

```
=> SHOW server_version;  
    server_version  
-----  
    14.1  
(1 row)
```

Чтобы повторить такие команды (а это полезно для самообразования, как и любые эксперименты), удобно открыть два терминала с `psql`.

Отдельные команды и названия различных объектов базы данных (таких как таблицы и столбцы, функции, расширения) выделены в тексте моноширинным шрифтом: `UPDATE`, `pg_class`.

Вызовы утилит из операционной системы показаны с приглашением, оканчивающимся на `$`:

```
postgres$ whoami  
postgres
```

Я использую Linux, но без какой-либо специфики; достаточно будет самого базового понимания.

Благодарности

Книгу невозможно написать в одиночку, и это отличный повод сказать спасибо хорошим людям.

Я благодарен Павлу Лузанову, который в нужный момент предложил мне заняться чем-то действительно стоящим.

Я признателен компании Postgres Professional за возможность работать над этой книгой не только в свободное время. Но компания — это люди, и я хочу сказать отдельное спасибо Олегу Бартунову за неиссякаемую энергию и идеи и Ивану Панченко за всестороннюю поддержку и \LaTeX .

Спасибо моим товарищам по образовательному отделу за творческую атмосферу и дискуссии, в ходе которых формировался материал учебных курсов и способ его подачи, что нашло свое отражение и в книге. Спасибо Павлу Толмачеву за внимательную вычитку черновиков.

Многие главы книги впервые были опубликованы в виде статей на Хабре¹, и я благодарен читателям за замечания и отклики. Они показали необходимость этой работы, позволили разглядеть белые пятна в моих знаниях и сделать текст лучше.

Спасибо Людмиле Мантровой, проделавшей огромную работу над языком книги. Если вы не спотыкаетесь на каждом втором предложении, это ее заслуга.

В книге я не называю имен, но за каждой функцией и возможностью, про которые я пишу, стоит многолетний труд вполне конкретных людей. Я восхищаюсь разработчиками PostgreSQL, и мне особенно приятно, что многих из них я имею честь называть коллегами.

¹ habr.com/ru/company/postgrespro/blog.

Конец ознакомительного фрагмента.

Приобрести книгу можно

в интернет-магазине

«Электронный универс»

e-Univers.ru