

УДК 004.65
ББК 32.972.134
P59

Рогов Е. В.

P59 PostgreSQL 15 изнутри. — М.: ДМК Пресс, 2023. — 662 с.
ISBN 978-5-93700-178-8

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.

В настоящем издании учтены замечания читателей и исправлены опечатки, а также отражены изменения, произошедшие в версии PostgreSQL 15.

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

Для администраторов и программистов.

УДК 004.65
ББК 32.972.134

На обложке использован фрагмент гравюры из книги *The History of Four-footed Beasts and Serpents* Эдварда Топсела, изданной в 1658 году в Лондоне.

ISBN 978-5-6045970-2-6
ISBN 978-5-93700-178-8

© Текст, оформление, ООО «ППГ», 2023
© Издание, ДМК Пресс, 2023

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	63
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
3.4. Индексы	89
3.5. TOAST	90
3.6. Виртуальные транзакции	91
3.7. 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	119
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	132
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	147
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	161
Заморозка при загрузке	161
Глава 8. Перестроение таблиц и индексов	163
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
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	185
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	210
10.5. Фоновая запись	213
10.6. Настройка	214
Настройка контрольной точки	214
Настройка фоновой записи	217
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	233
Replica	235
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 неотношений	279
14.2. Блокировки расширения отношения	281
14.3. Блокировки страниц	282

Содержание

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 не скалярных 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

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
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	430
Внешние соединения	434
Анти- и полусоединения	436
Неэквисоединения	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	485
23.4. Сравнение способов соединения	488
Часть V. Типы индексов	493
Глава 24. Хеш-индекс	495
24.1. Общий принцип	495
24.2. Страничная организация	496
24.3. Класс операторов	503
24.4. Свойства	504
Свойства метода доступа	504
Свойства индекса	505
Свойства столбцов	506
Глава 25. В-дерево	507
25.1. Общий принцип	507
25.2. Поиск и вставка	508
Поиск по равенству	508
Поиск по неравенству	510
Поиск по диапазону	511
Вставка	511
25.3. Страничная организация	513
Компактное хранение дубликатов	517
Компактное хранение внутренних индексных записей	519
25.4. Класс операторов	520
Семантика сравнения	520
Сортировка и составные индексы	526
25.5. Свойства	531
Свойства метода доступа	531
Свойства индекса	532
Свойства столбцов	532

Глава 26. Индекс GiST	534
26.1. Общий принцип	534
26.2. R-дерево для точек	536
Страничная организация	539
Класс операторов	540
Поиск вхождения в область	542
Поиск ближайших соседей	544
Вставка	549
Ограничение исключения	550
Свойства	553
26.3. RD-дерево для полнотекстового поиска	556
Про полнотекстовый поиск	556
Индексация tsvector	557
Свойства	565
26.4. Другие типы данных	565
Глава 27. Индекс SP-GiST	568
27.1. Общий принцип	568
27.2. Дерево квадрантов для точек	570
Класс операторов	571
Страничная организация	575
Поиск	576
Вставка	577
Свойства	580
27.3. K-мерные деревья для точек	582
27.4. Префиксное дерево для строк	584
Класс операторов	585
Поиск	586
Вставка	587
Свойства	589
27.5. Другие типы данных	590
Глава 28. Индекс GIN	592
28.1. Общий принцип	592
28.2. Индекс для полнотекстового поиска	593
Страничная организация	595
Класс операторов	597

Поиск	599
Частые и редкие лексемы	600
Вставка	604
Ограничение выборки	606
Свойства	607
Ограничения GIN и RUM-индекс	609
28.3. Индекс для триграмм	610
28.4. Индекс для массивов	612
28.5. Индекс для JSON	616
Класс операторов jsonb_ops	616
Класс операторов jsonb_path_ops	619
28.6. Другие типы данных	621
Глава 29. Индекс BRIN	622
29.1. Общий принцип	622
29.2. Пример	623
29.3. Страничная организация	625
29.4. Поиск	627
29.5. Обновление сводной информации	628
Вставка значений	628
Обобщение зоны	629
29.6. Диапазоны значений (minmax)	630
Выбор столбцов для индексирования	631
Размер зоны и эффективность поиска	632
Свойства	636
29.7. Мультидиапазоны значений (minmax-multi)	639
29.8. Охватывающие значения (inclusion)	642
29.9. Фильтры Блума (bloom)	645
Заключение	650
Предметный указатель	651