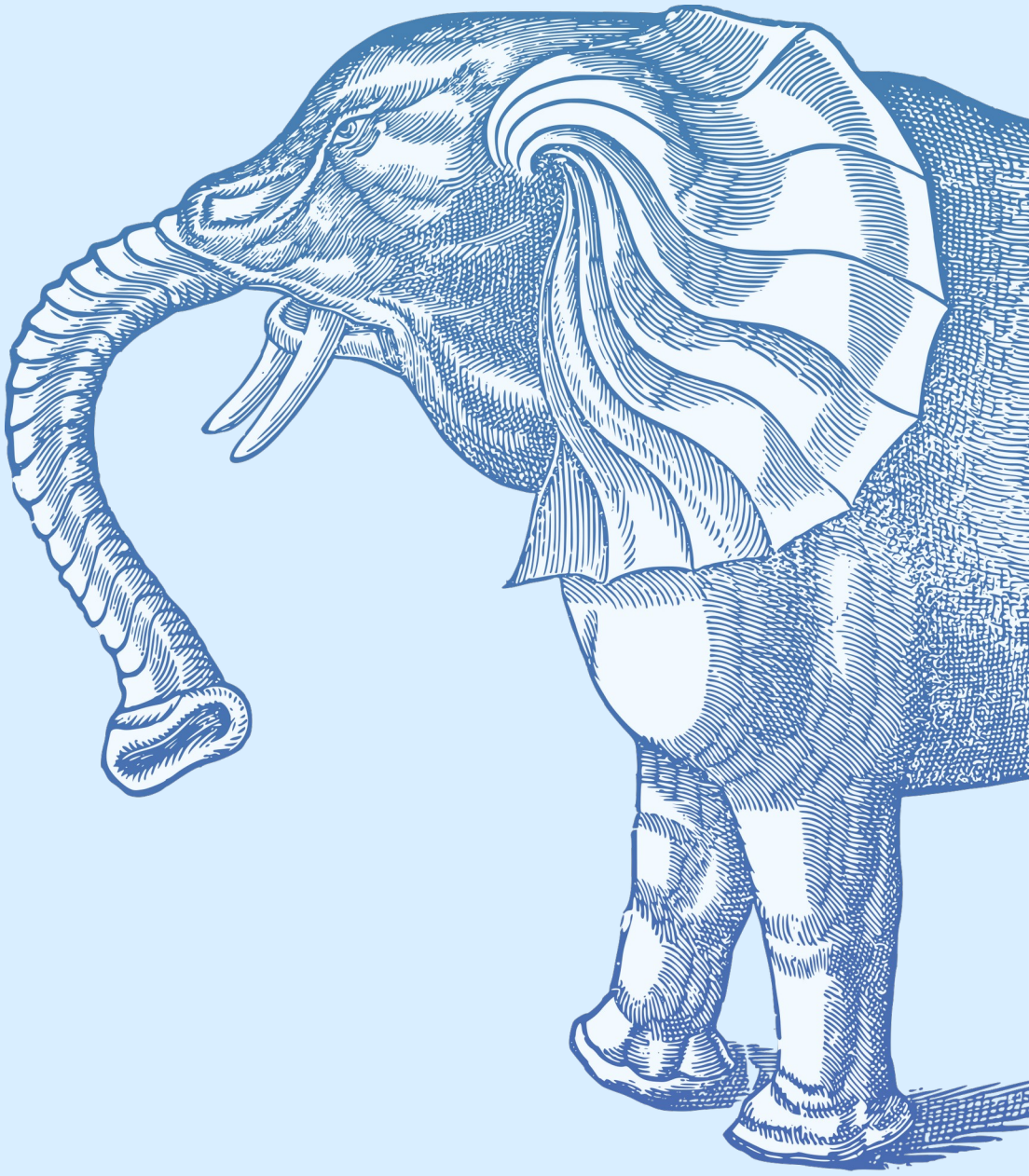
Егор Рогов
Postgres Pro

Изнанка «PostgreSQL 14 изнутри»



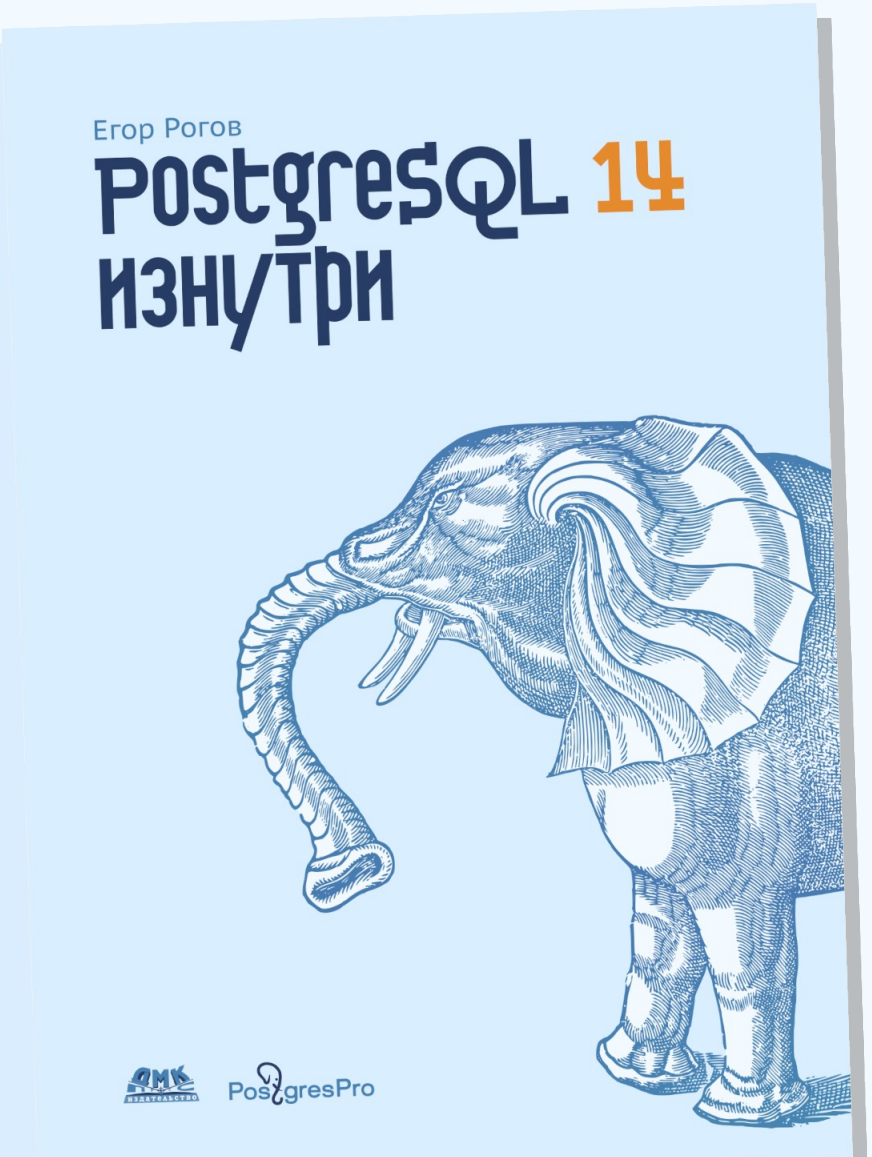
О себе

Егор Рогов

Отдел образовательных программ Postgres Professional

- edu@postgrespro.ru
- все в свободном доступе на postgrespro.ru/education

PostgreSQL изнутри



postgrespro.ru/education/books/internals

В основе — учебные курсы, в основном DBA1 и DBA2, и статьи на Хабре

Превращение в книгу заняло один год

О чем книга и для кого она

Всем, кому интересно заглядывать под капот черного ящика

- изоляция и многоверсионность
- буферный кеш и журнал
- блокировки
- выполнение запросов
- типы индексов

Не учебник, не сборник рецептов, не замена документации,
не руководство по разработке ядра

Как написать книгу

Папа обиженно встал и стал собирать свои листки.

— Если вам не нравится моя пьеса, пожалуйста, пишите сами новую, — сказал он.

— Голубчик, — утешала его мама, — мы находим пьесу замечательной. Не правда ли?

— Конечно, —подтвердили все в один голос.

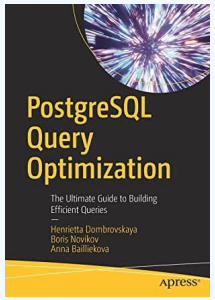
— Вот видишь, она всем нравится, —сказала мама. Только чуть подправь ее, измени содержание и стиль. Я позабочусь, чтобы никто не мешал тебе. И пока ты работаешь, возле тебя будет стоять вазочка с карамельками.

— *Туве Янссон,* Опасное лето

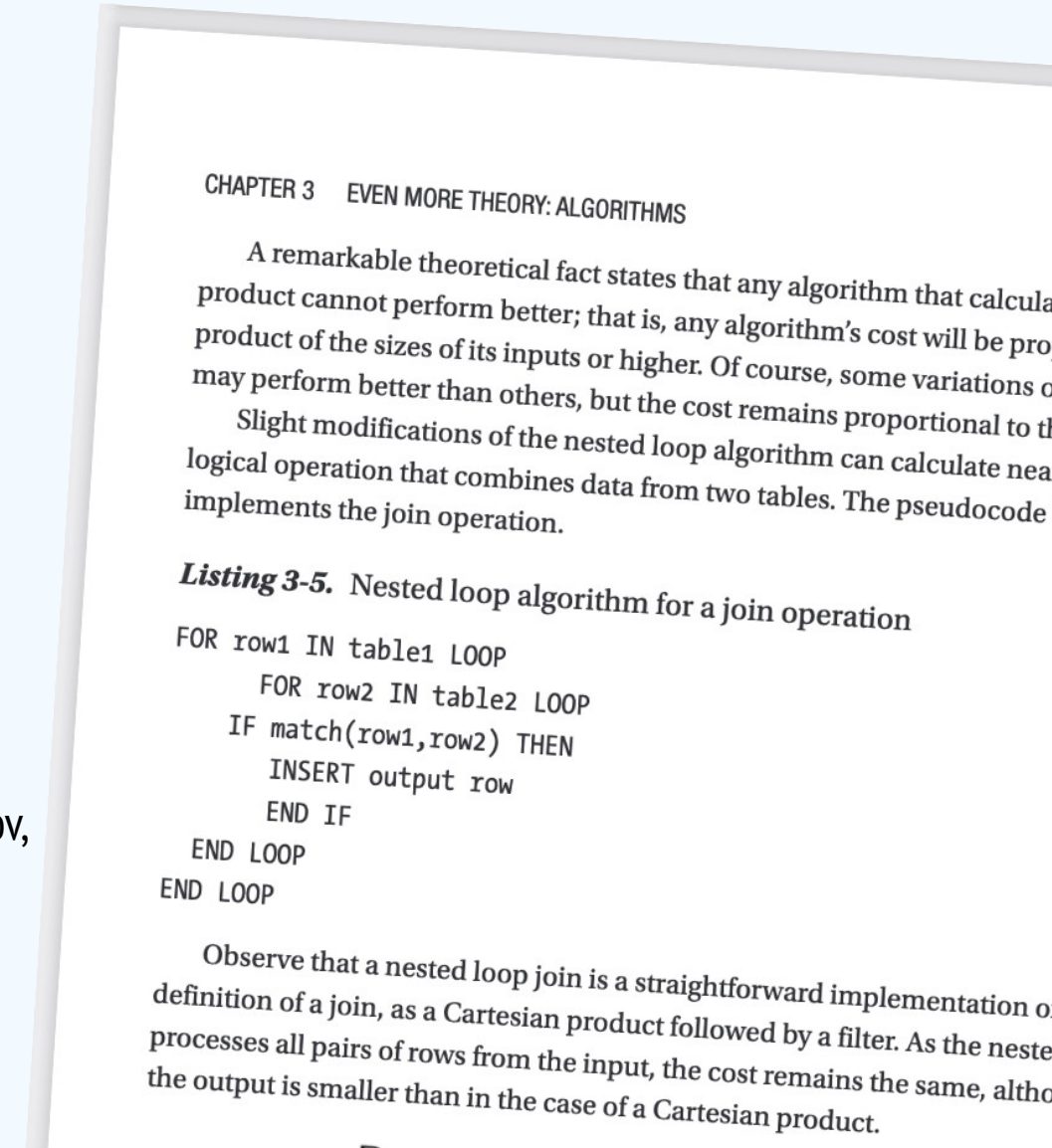
Хочешь сделать хорошо...

For programming books, it was vital that the programs were typeset directly from the source code, so we could be sure that what was printed was correct, that it hadn't been inadvertently changed by human intervention.

— *Brian Kernighan,*
Unix: a History and a Memoir



Henrietta Dombrovskaya, Boris Novikov, Anna Bailliekova.
PostgreSQL Query Optimization:
The Ultimate Guide to Building Efficient Queries



CHAPTER 3 EVEN MORE THEORY: ALGORITHMS

A remarkable theoretical fact states that any algorithm that calcula
product cannot perform better; that is, any algorithm's cost will be pro
product of the sizes of its inputs or higher. Of course, some variations o
may perform better than others, but the cost remains proportional to th
Slight modifications of the nested loop algorithm can calculate nea
logical operation that combines data from two tables. The pseudocode
implements the join operation.

Listing 3-5. Nested loop algorithm for a join operation

```
FOR row1 IN table1 LOOP
    FOR row2 IN table2 LOOP
  IF match(row1,row2) THEN
    INSERT output row
    END IF
  END LOOP
END LOOP
```

Observe that a nested loop join is a straightforward implementation o
definition of a join, as a Cartesian product followed by a filter. As the neste
processes all pairs of rows from the input, the cost remains the same, altho
the output is smaller than in the case of a Cartesian product.

…сделай сам!

He loved maps, as I have told you before;
and he also liked runes and letters and
cunning handwriting, though when he wrote
himself it was a bit thin and spidery.

— *J. R. R. Tolkien*, The Hobbit

TeX и LaTeX

After having spent ten years developing the TeX and METAFONT systems for computer typesetting, I am now able to fulfill the dream that I had when I began that work, by applying those systems to The Art of Computer Programming.

— *Donald E. Knuth,*
The Art of Computer Programming,
3rd ed.

Дональд Кнут

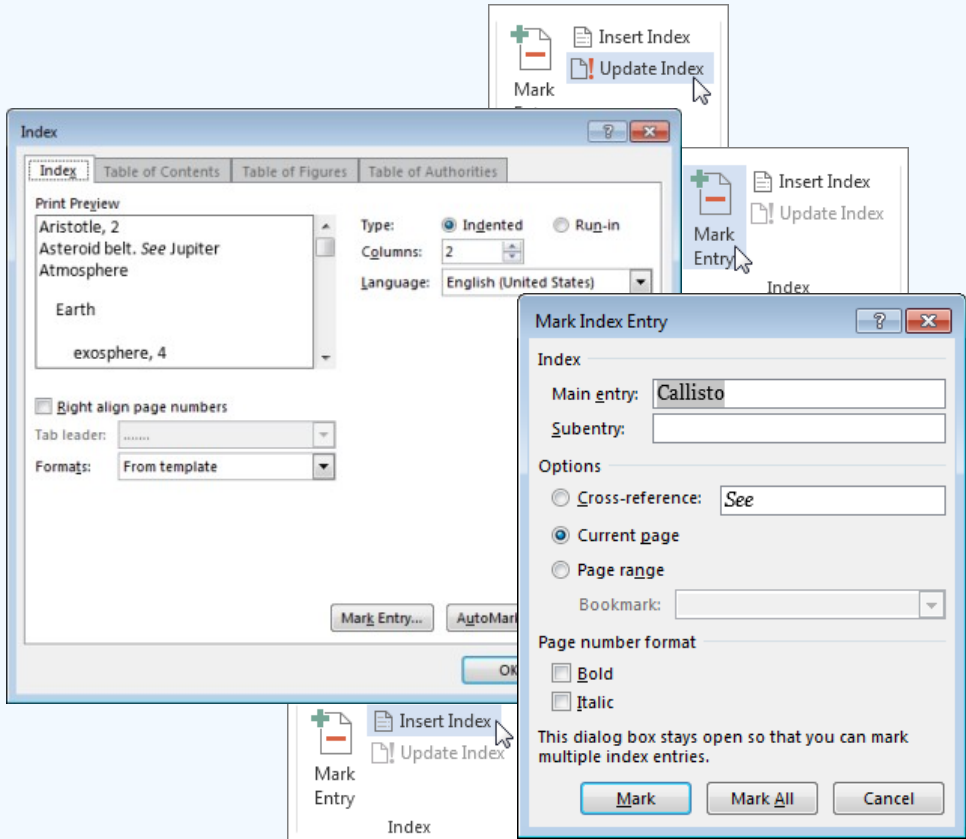
Лесли Лэмпорт

Книга как исходный код

LaTeX — не просто язык разметки,
но и полный по Тьюрингу язык макроподстановок

Обычный текстовый файл

- любимый текстовый редактор
- git
- diff
- make (latexmk)



Пример: моноширинные пробелы

ние, второй отдельно дочитывает пропущенное начало таблицы.

Стратегия массовой записи применяется для операций COPY FROM, CREATE TABLE AS SELECT, CREATE MATERIALIZED VIEW и тех вариантов ALTER TABLE, которые вызывают перезапись таблицы. Выделяется довольно большое кольцо размером 16 Мбайт (2048 стандартных страниц), но не больше $\frac{1}{8}$ от размера всего буферного кеша.

Стратегия массовой записи применяется для операций COPY FROM, CREATE TABLE AS SELECT, CREATE MATERIALIZED VIEW и тех вариантов ALTER TABLE, которые вызывают перезапись таблицы. Выделяется довольно большое кольцо размером 16 Мбайт (2048 стандартных страниц), но не больше $\frac{1}{8}$ от размера всего буферного кеша.

Стратегия очистки применяется

book — vim _styles.latex — 120×40

```
% код в тексте (моноширинные пробелы смотрятся ужасно, заменяем их на обычные)
\usepackage{xstring}
\newcommand{\codett}[1]{%
  \noexpandarg%
  \StrSubstitute{#1}{ }{\textnormal{ }}[\codewords]%
  \texttt{\codewords}%
}
```

Кстати, microtype

ычислительных системах по-
ратном уровне. У одного толь-
ровня кеша; свой кеш бывает
у самих дисков.

мпенсировать разную произ-
оторых быстрее, но дороже и
дешевле и больше. В быстрой
медленной памяти. Но в боль-
овременно только с *небольшой*
амяти под *кеш* для хранения
омить на обращениях к мед-

отношений, сглаживая разни-
амяти (наносекунды) и к дис-

ется и у операционной систе-
УБД стараются избегать двой-
к диску напрямую, минуя кеш
ные читаются и записываются

слительных системах по-
ом уровне. У одного толь-
я кеша; свой кеш бывает
мих дисков.

нсировать разную произ-
ых быстрее, но дороже и
евле и больше. В быстрой
ленной памяти. Но в боль-
менно только с *небольшой*
ти под *кеш* для хранения
ть на обращениях к мед-

ошений, сглаживая разни-
ги (наносекунды) и к дис-

и у операционной систе-
стараются избегать двой-
ску напрямую, минуя кеш
читаются и записываются

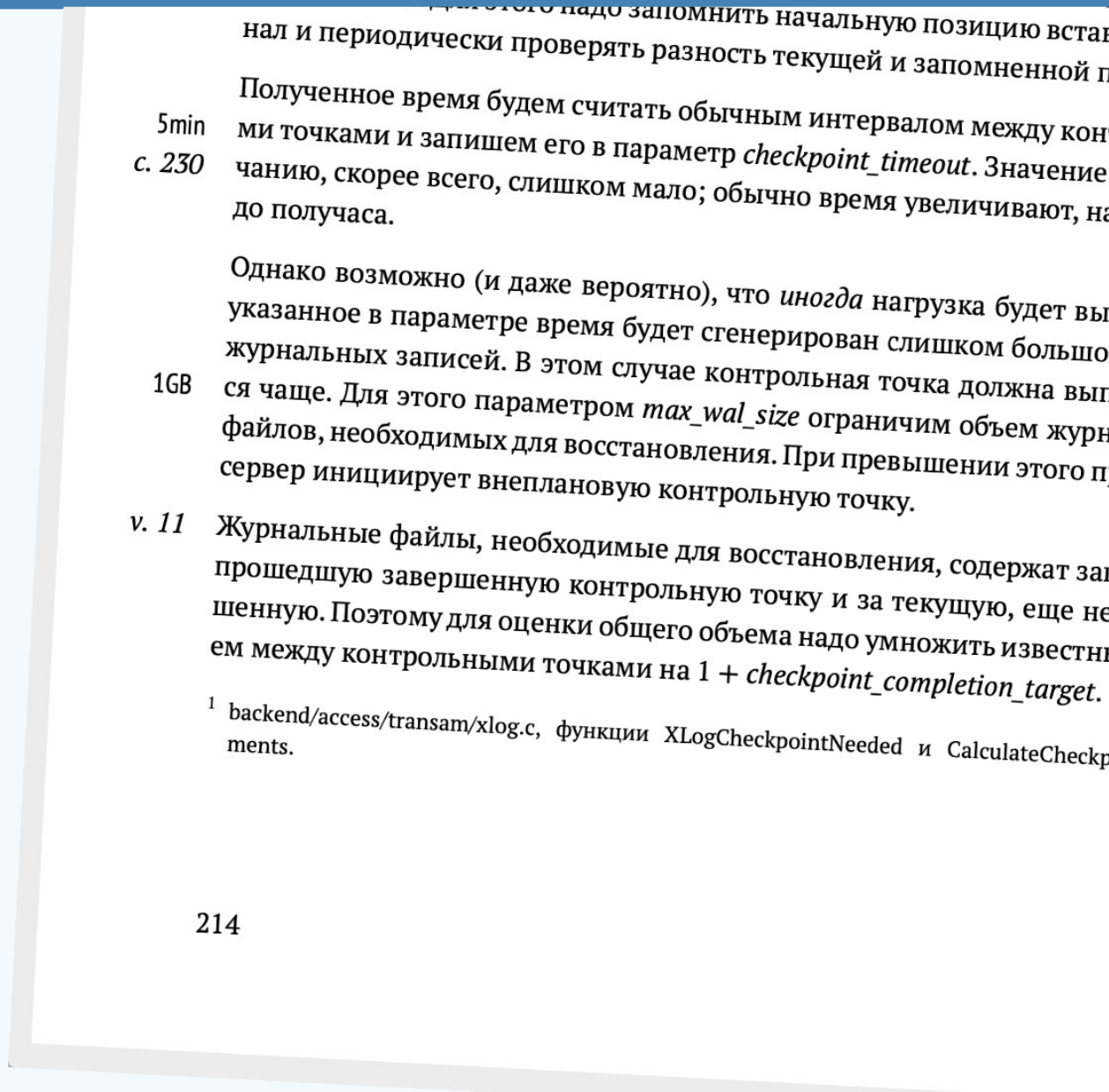
Micro-typography is the art of enhancing the appearance and readability of a document while exhibiting a minimum degree of visual obtrusion.

— *R Schlich,*
The microtype package:
Subliminal refinements towards typographical perfection

Пример: макет

Многие книги используются — или могут в принципе быть использованы — одновременно для нескольких, в том числе и не предусмотренных автором целей.

— *Елена Герчук*, Архитектура книги



нал и периодически проверять разность текущей и запомненной п

Полученное время будем считать обычным интервалом между кон
5min ми точками и запишем его в параметр *checkpoint_timeout*. Значение
c. 230 чанию, скорее всего, слишком мало; обычно время увеличивают, на
до получаса.

Однако возможно (и даже вероятно), что *иногда* нагрузка будет вы
указанное в параметре время будет сгенерирован слишком большо
журнальных записей. В этом случае контрольная точка должна вып
1GB ся чаще. Для этого параметром *max_wal_size* ограничим объем журн
файлов, необходимых для восстановления. При превышении этого п
сервер инициирует внеплановую контрольную точку.

v. 11 Журнальные файлы, необходимые для восстановления, содержат за
прошедшую завершенную контрольную точку и за текущую, еще не
шенную. Поэтому для оценки общего объема надо умножить известн
ем между контрольными точками на 1 + *checkpoint_completion_target*.

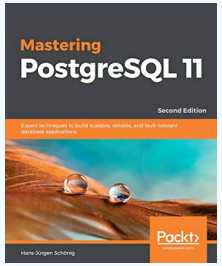
[1] backend/access/transam/xlog.c, функции XLogCheckpointNeeded и CalculateCheckp
ments.

214

Запросы и copy-paste

Голова удатна,
да лень перекатна.

— *народная пословица*



Hans-Jürgen Schönig.
Mastering PostgreSQL 11,
second edition

Creating partitions

First, we will take a closer look at the outdated method of partitioning data. Keep in mind that understanding this technique is important to gain a deeper overview of what PostgreSQL really does behind the scenes.

Before digging deeper into the advantages of partitioning, I want to show how partitions can be created. The entire thing starts with a parent table that we can create by using the following command:

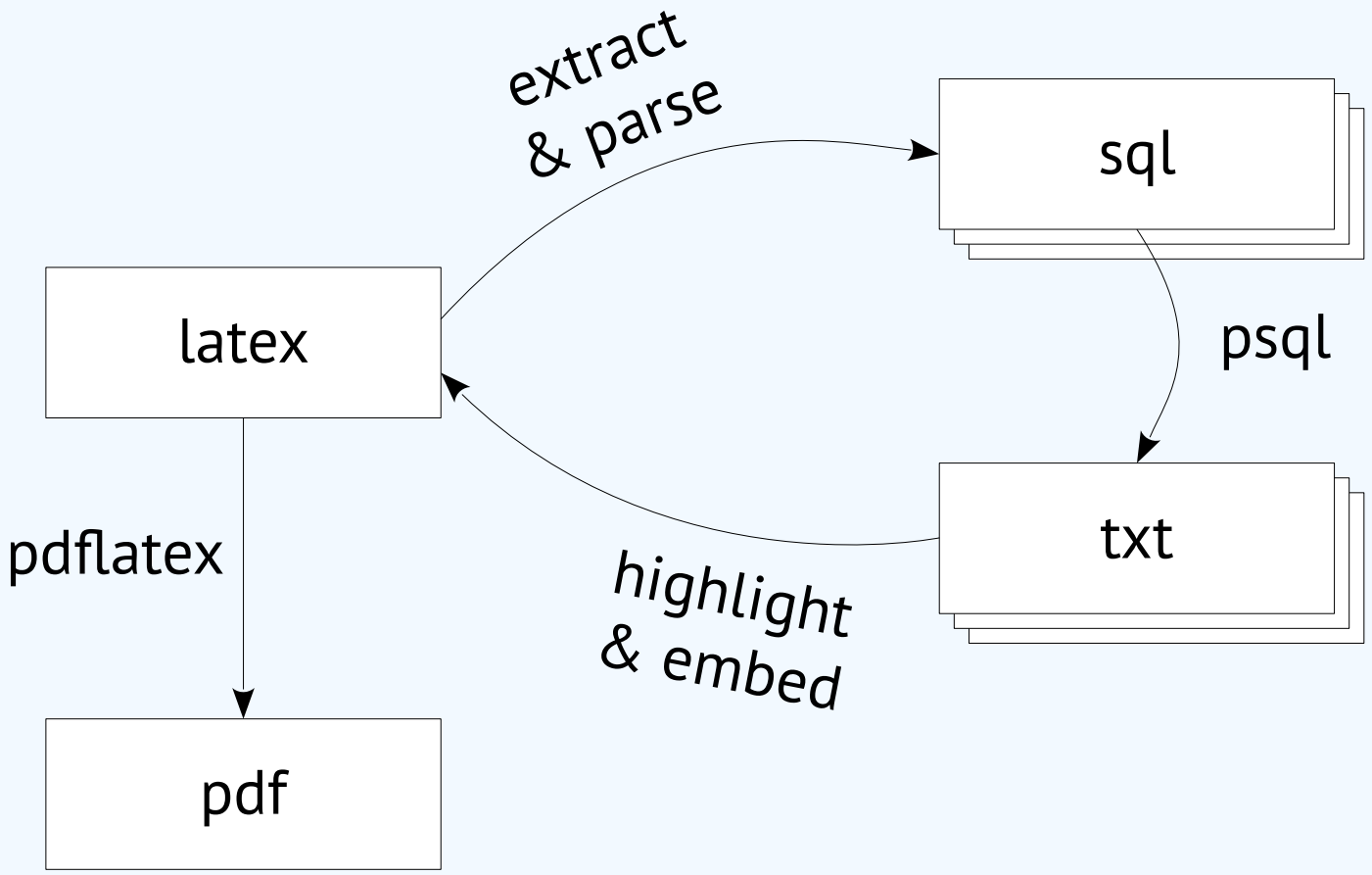
```
test=# CREATE TABLE t_data (id serial, t date, payload text);
CREATE TABLE
```

In this example, the parent table has three columns. The date column will be used for partitioning, but we'll cover more on that a bit later.

Now that the parent table is in place, the child tables can be created. This is how it works:

```
test=# CREATE TABLE t_data_2016 () INHERITS (t_data);
CREATE TABLE
test=# \d t_data_2016
                    Table "public.t_data_2016"
 Column  | Type    | Modifiers
---------+---------+-----------------------------------------------------
 id      | integer | not null default
                     nextval('t_data_id_seq'::regclass)
 t       | date    |
 payload | text    |
```

Запросы как часть книги



Запросы как часть книги

```
book -
Конечно, в книге много фрагментов кода, в осн
>}; если необходимо, то следом за ним приведе

%% start_chunk test-1
%% s 1 "SELECT now();"
\begin{nobar}
\begin{code}
=> \textbf{SELECT} now();
\end{code}
\begin{result}
                now
\mbox{-}\mbox{-}\mbox{-}\mbox{-}\mbox{-}\mbox{
{-}\mbox{-}\mbox{-}\mbox{-}\mbox{-}\mbox{-}\mb
box{-}\mbox{-}\mbox{-}\mbox{-}\mbox{-}\mbox{-}
 2022\mbox{-}01\mbox{-}04 17:45:15.108324+03
(1 row)
\end{result}
\end{nobar}
%% end_chunk
```

мысли, но которые я не удержал
было пропустить.

Конечно, в книге много фрагментов кода, в основном на язык
казан с приглашением =>; если необходимо, то следом за ни
ответ сервера:

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

Если аккуратно повторять все приведенные команды в Post
жен получиться такой же результат (конечно, с точностью до
закций и прочих несущественных деталей). Во всяком случае
ге — результат выполнения скрипта, содержащего ровно эти

1 postgrespro.ru/docs/postgresql/14/index.
2 git.postgresql.org/gitweb/?p=postgresql.git;a=summary.
3 commitfest.postgresql.org.

EXPLAIN

internals-14 — vim explain.patch — 101×36

```
            if (has_temp)
            {
-               appendStringInfoString(es->str, " temp");
+               appendStringInfoString(tmpstr, " temp");
                if (usage->temp_blks_read > 0)
-                   appendStringInfo(es->str, " read=%lld",
+                   appendStringInfo(tmpstr, " read=%lld",
                                     (long long) usage->temp_blks_read
                if (usage->temp_blks_written > 0)
-                   appendStringInfo(es->str, " written=%lld",
+                   appendStringInfo(tmpstr, " written=%lld",
                                     (long long) usage->temp_blks_writ
            }
-           appendStringInfoChar(es->str, '\n');
+           appendStringInfoChar(tmpstr, '\n');
+
+            if (Wrap_explain == true && strlen(tmpstr->data) > Limit
+            {
+                int pos = find_wrap_pos(tmpstr->data, Limit_explain
+                (tmpstr->data)[pos] = '\0';
+                appendStringInfo(es->str, "%s\n", tmpstr->data);
+                ExplainIndentText(es);
+                /* assuming one wrap is enough... */
+                appendStringInfo(es->str, "%s", tmpstr->data+pos+1);
+            }
+           else
+                appendStringInfo(es->str, "%s", tmpstr->data);
        }

        /* As above, show only positive counter values. */
@@ -4372,14 +4496,28 @@ static void
 ExplainProperty(const char *qlabel, const char *unit, const char *v
                 bool numeric, ExplainState *es)
```

21.2. Соединение вложенным циклом

```
                                QUERY PLAN
-----------------------------------------------------------------------
 Nested Loop Anti Join  (cost=0.28..4.65 rows=1 width=40)
   -> Seq Scan on aircrafts_data ml  (cost=0.00..1.09 rows=9 widt...
   -> Index Only Scan using seats_pkey on seats s
        (cost=0.28..5.55 rows=149 width=4)
        Index Cond: (aircraft_code = ml.aircraft_code)
(5 rows)
```

Тот же план будет построен и для эквивалентного запроса без NOT EXISTS:

```
=> EXPLAIN SELECT a.*
FROM aircrafts a
  LEFT JOIN seats s ON a.aircraft_code = s.aircraft_code
WHERE s.aircraft_code IS NULL;
                                QUERY PLAN
-----------------------------------------------------------------------
 Nested Loop Anti Join  (cost=0.28..4.65 rows=1 width=40)
   -> Seq Scan on aircrafts_data ml  (cost=0.00..1.09 rows=9 widt...
   -> Index Only Scan using seats_pkey on seats s
        (cost=0.28..5.55 rows=149 width=4)
        Index Cond: (aircraft_code = ml.aircraft_code)
(5 rows)
```

Полусоединение возвращает те строки

TikZ: картинка как код

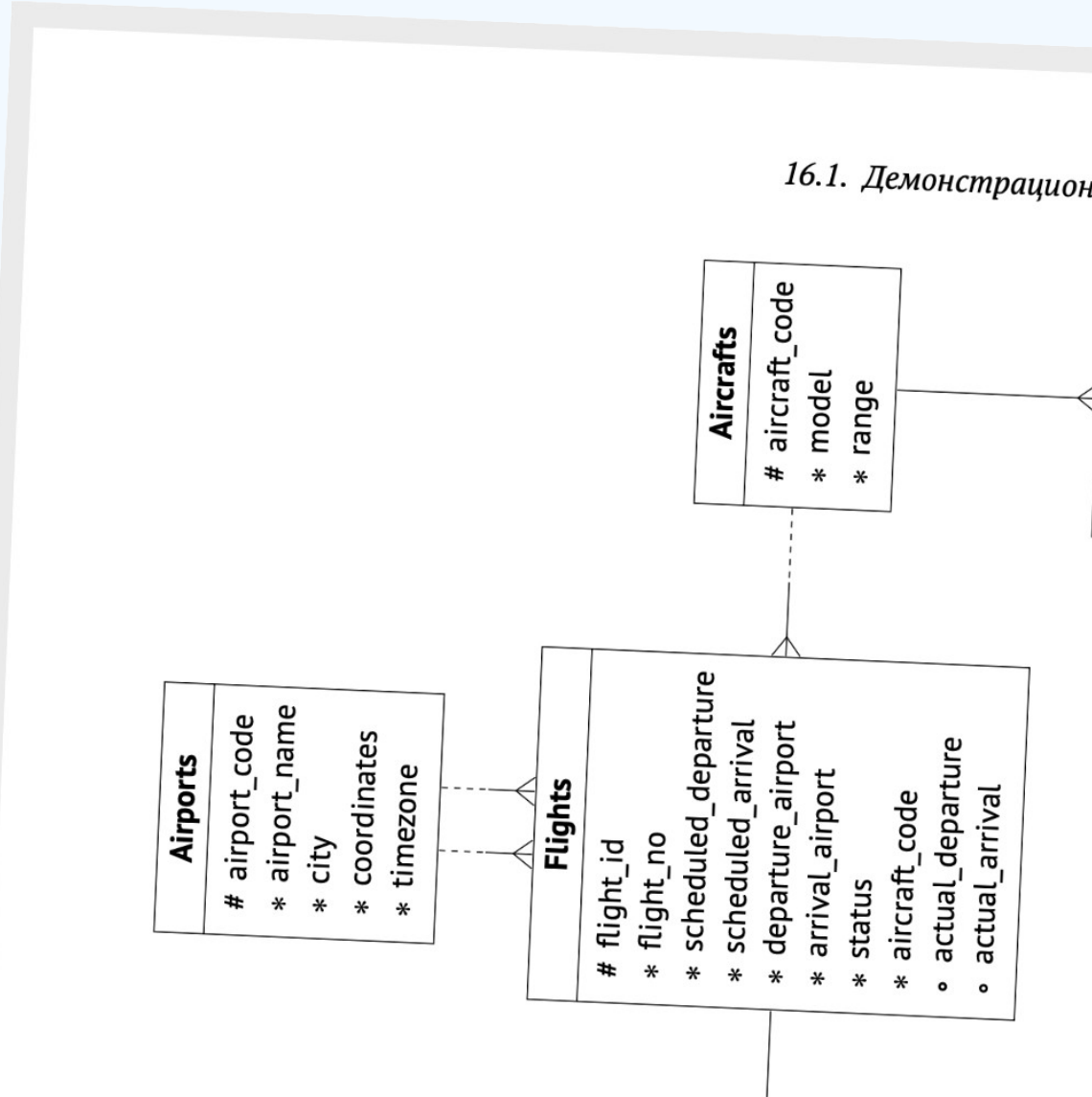
In a sense, when you use TikZ you “program” your graphics, just as you “program” your document when you use TeX.

— *Till Tantau*, The TikZ and PGF Packages

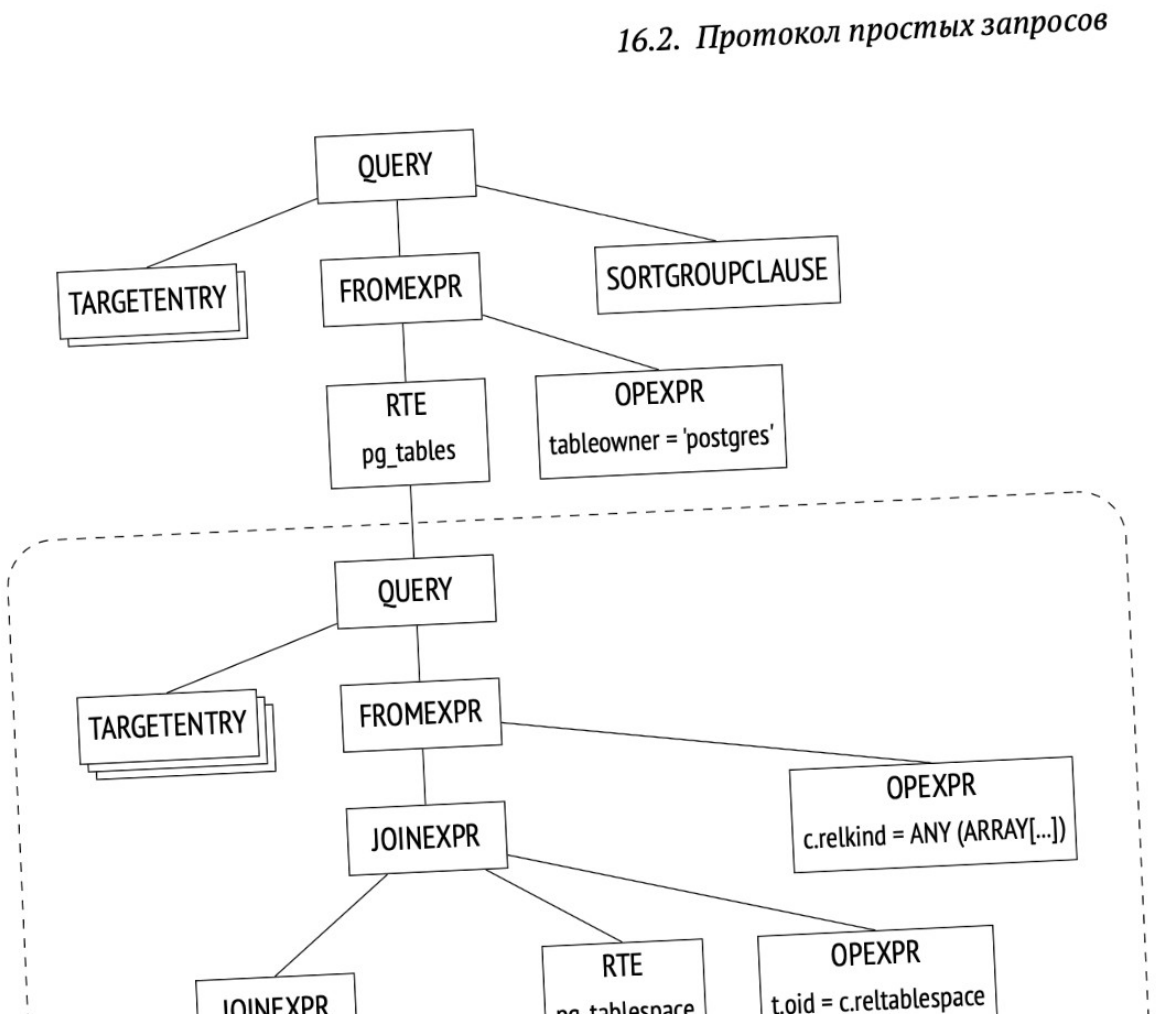
book — vim _query.latex

```
    \draw [one to many] (bookings) -- (tickets);
    \draw [one to many] (tickets) -- (ticket-flights);
    \draw [one to many] (flights) -- (ticket-flights);
    \draw [optional one to one] (ticket-flights) -- (
    \draw [optional one to many] (aircrafts) -- (fligl
    \draw [one to many] (aircrafts) -- (seats);
    \draw [optional one to many] ([xshift=-2ex]airpor
    \draw [optional one to many] ([xshift=+2ex]airpor
  \end{tikzpicture}

\end{document}
```



TikZ: картинка как код



book — vim _query.latex — 101×41

```
\node (query) {QUERY}
  child {
    node [fill=white, copy shadow={shadow xshift=0.5ex,shadow yshi
  }
  child {
    node {FROMEXPR}
    child [missing] {
    }
    child {
      node {RTE\\ \small pg\_tables}
      child {
        node (top) {QUERY}
        child {
          node (left) [fill=white, double copy shadow={shadow xshi
            {TARGETENTRY}
        }
        child {
          node {FROMEXPR}
          child [missing] {
          }
          child {
            node {JOINEXPR}
            child [missing] {
            }
            child [sibling distance=8em, xshift=-1em] {
              node {JOINEXPR}
              child [missing] {
              }
              child [xshift=1em] {
                node (bottom) {RTE\\ \small pg\_class}
```

TikZ: дополненная реальность

Вот пример такого плана:

```
=> SET work_mem = '64MB';
=> EXPLAIN (analyze, costs off, timing off, summary off)
SELECT count(*)
FROM bookings b
  JOIN tickets t ON t.book_ref = b.book_ref;
                              QUERY PLAN
-----------------------------------------------------------------------
 Finalize Aggregate (actual rows=1 loops=1)
   -> Gather (actual rows=3 loops=1)
        Workers Planned: 2
        Workers Launched: 2
        -> Partial Aggregate (actual rows=1 loops=3)
             -> Parallel Hash Join (actual rows=983286 loops=3)
                  Hash Cond: (t.book_ref = b.book_ref)
                  -> Parallel Index Only Scan using tickets_book_ref...
                       Heap Fetches: 0
                  -> Parallel Hash (actual rows=703703 loops=3)
                       Buckets: 4194304  Batches: 1  Memory Usage:
                       115392kB
                       -> Parallel Seq Scan on bookings b (actual row...
(13 rows)
=> RESET work_mem;
```

Это тот же запрос, что я показывал в предыдущем разделе, но там парал-

на эти версии гарантированно не было
строки была записана на освободившееся

лько раз:

```
ot',0);
  | xmax  | hhu | hot | t_ctid
--+-------+-----+-----+-------
  |       |     |     |
c | 817 c | t   | t   | (0,3)
c | 818   | t   | t   | (0,5)
c | 816 c | t   | t   | (0,2)
  | 0 a   |     | t   | (0,5)
```

вызывает внутристраничную очистку:

Шрифты

Спасибо Паратайпу за PT Sans, PT Serif и PT Mono

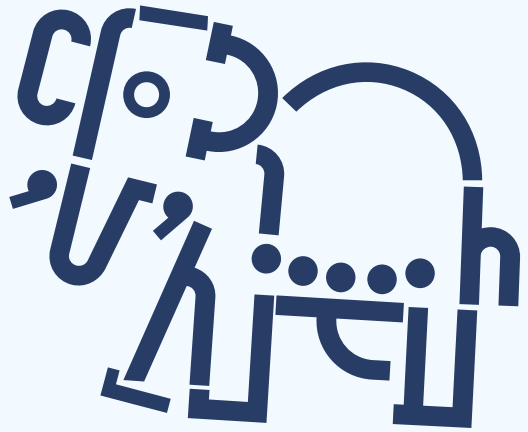
Общедоступные шрифты с открытой лицензией,
«универсальные по назначению, современные
по дизайну и соответствующие представлениям
о хорошей грамотной кириллице, которая не вызывает
раздражения у отечественных пользователей»

Шрифты

Спасибо Юг за две круглых

Шрифт построен нарочито грубо, из штрихов одинаковой толщины. Никаких оптических компенсаторов, никаких хитростей на стыках и сгибах. Фактически — трубосварка и ацетиленовая резка. Все отсылки шрифта — к конструктивизму, хотя тогда так не делали.

— *Юрий Гордон,* Книга про мои буквы



О слоне


Of the Elephant. 151

Here is no creature among all the Beasts of the world, which hath so great and ample demonstration of the power and wisdom of Almighty God as the Elephant.

— *Edward Topsell,*
The History of Four-footed Beats and Serpent
London, 1658

Скоро и на английском

Тот перевод хорош, который по возможности
наиболее точно соответствует оригиналу,
но вместе с тем производит впечатление,
будто написан он на языке перевода.

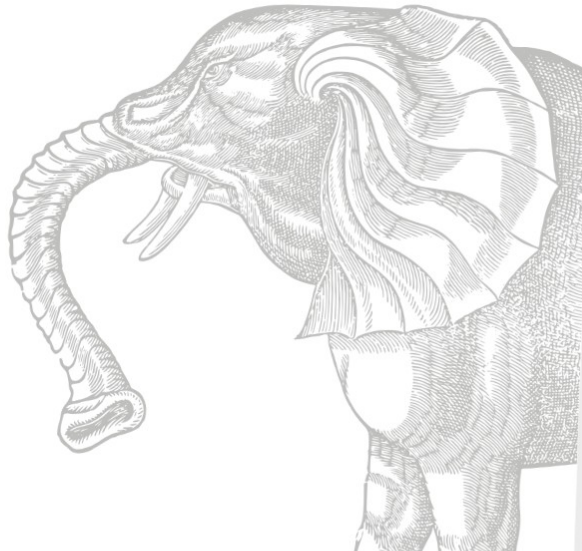
— *Като Ломб,* Как я изучаю языки

book_en — vim po/_about

```
#. type: Plain text
#: ../po/_about.latex:48
msgid ""
"Я ориентируюсь на читателей, имеющих определенный опыт использов
"PostgreSQL и хотя бы в общих чертах представляющих себе, что к ч
"совсем новичков текст будет тяжеловат. Например, я ни слова не с
"как устанавливать сервер, вводить команды в \\prog{psql} или изм
"конфигурационные параметры."
msgstr ""
"I assume that the reader has already tried using PostgreSQL and
"some general understanding of how it works. For entry-level use
```

Egor Rogov

PostgreSQL 14 Internals



This is Part I of the book.
The other parts will follow soon:

II. Buffer Cache and WAL

III. Locks

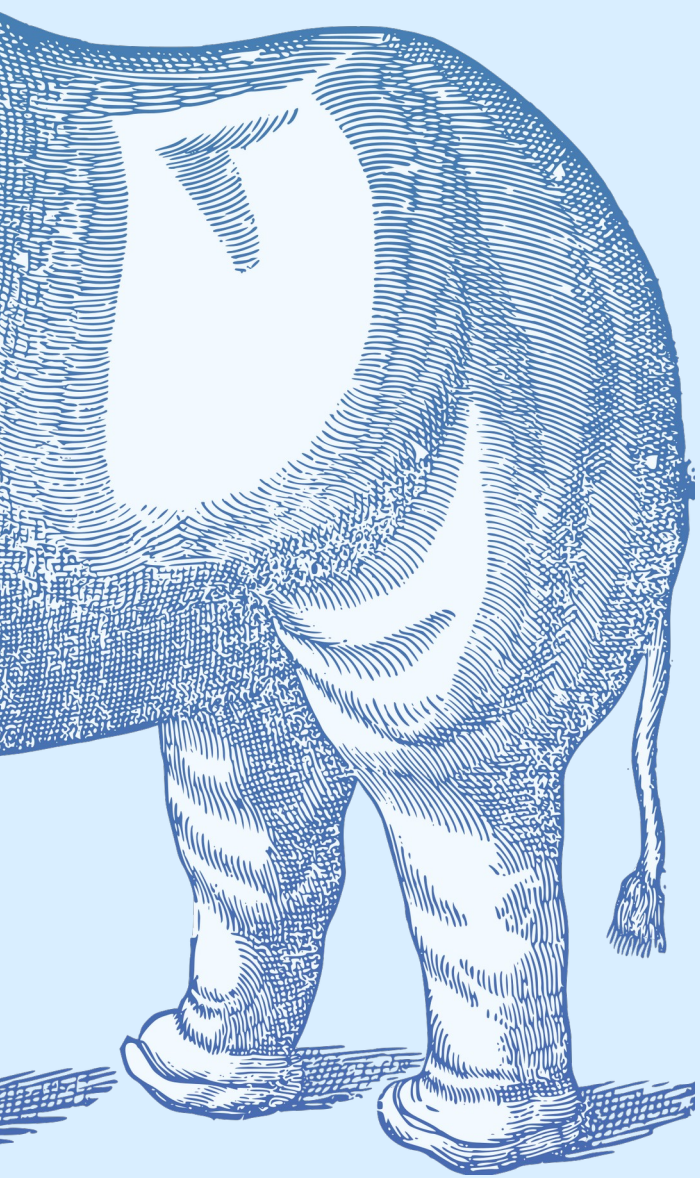
IV. Query Execution

V. Index Types

Кто хочет написать свою книгу?

Чем мы можем помочь:

- обратная связь
- редактура
- подготовка к печати
- издание



Спасибо за внимание

Остались вопросы?

edu@postgrespro.ru