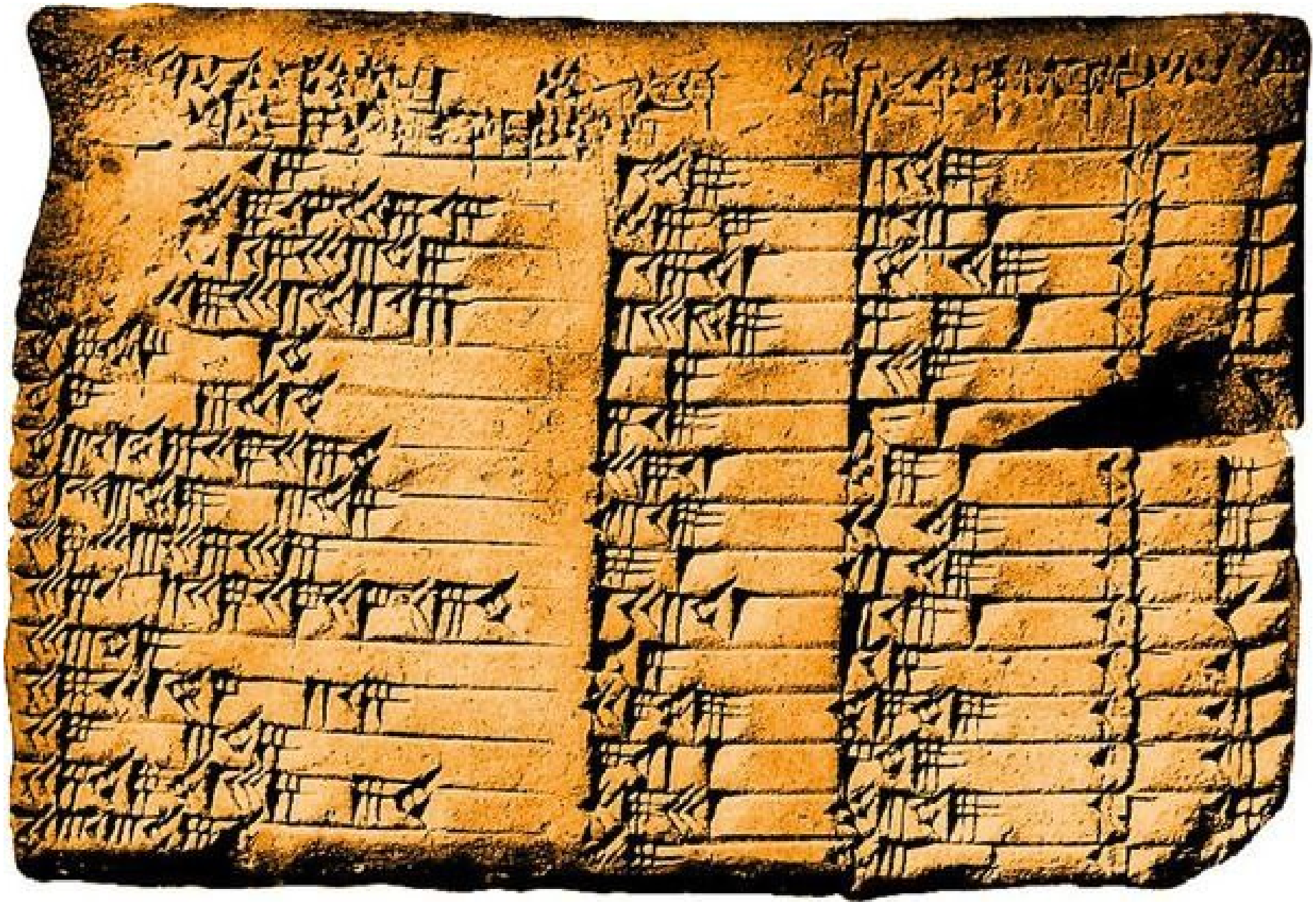
Postgres от начала веков до наших дней

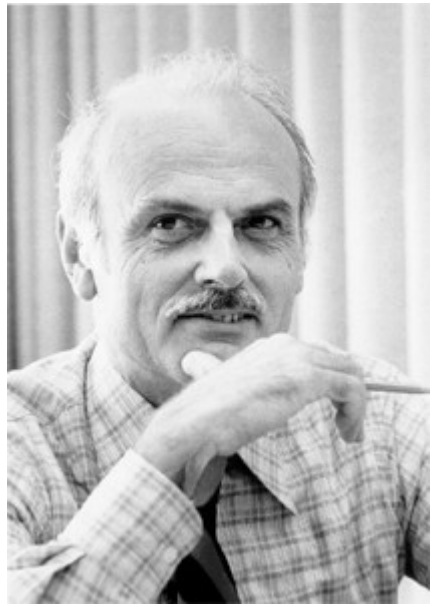
Олег Бартунов
Иван Панченко



Были ли СУБД до СУБД ?

- Можно ли считать вавилонские таблички СУБД
- Хм, а что вообще мы понимаем под СУБД?
    - ACID
    - Структурированность данных
    - Оптимизация для поиска
- Определённые идеи, конечно, были
    - Таблицы :)
    - Картотеки :)
    - Двойная запись в бухгалтерии

У основ СУБД ?

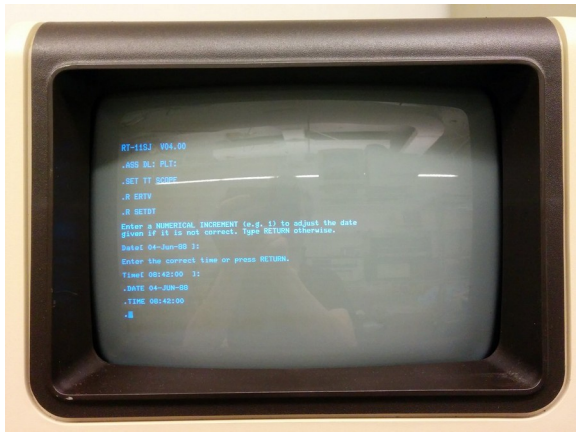


- Статья Эдгара Кодда:
- A relational model of data for large shared data banks. Communications of the ACMVolume **13,** 6 June 1970, pp 377–387
- https://dl.acm.org/doi/10.1145/362384.362685

Ingres

- Чтобы разработать СУБД, Майкл Стоунбрейкер приобрел на грант пару «ПК» PDP-11.

- Когда подвезли терминал и поставили UNIX…
- Ему так понравилась интерактивная работа в shell, что он назвал СУБД
- INteractive Graphic REtrieval System



PostgresPro

Что было хорошего в Ingres

- Язык QUEL — почти SQL.
- Cost-based planner
- Data dictionary
- Access methods
- Locks
- Wei Hong набрал его код руками с распечатки в 1985 г.
- См https://doi.org/10.1145/320473.320476
  https://doi.org/10.1145/320141.320158

Языки реляционных баз

- QUEL – QUEryLanguage
- PostQUEL

- SQUARE – Specifying Queries As Relational Expressions (1975) Raymond Boyce,

  Donald Chamberlin и др.

- SEQUEL - 1974
- SQL

SEQUEL: A STRUCTURED ENGLISH QUERY LANGUAGE

by

Donald D. Chamberlin
Raymond F. Boyce

IBM Research Laboratory
San Jose, California

ABSTRACT: In this paper we present the data manipulation facility for a structured English query language (SEQUEL) which can be used for accessing data in an integrated relational data base. Without resorting to the concepts of bound variables and quantifiers SEQUEL identifies a set of simple operations on tabular structures, which can be shown to be of equivalent power to the first order predicate calculus. A SEQUEL user is presented with a consistent set of keyword English templates which reflect how people use tables to obtain information. Moreover, the SEQUEL user is able to compose these basic templates in a structured manner in order to form more complex queries. SEQUEL is intended as a data base sublanguage for both the professional programmer and the more infrequent data base user.

Всё делали команды

Стоунбрейкер

- Лучшие ребята из Беркли
- QUEL — ошибка Стоунбрейкера.
- Она и убила Ingres (тогда)
- Академический и эксцентричный подход

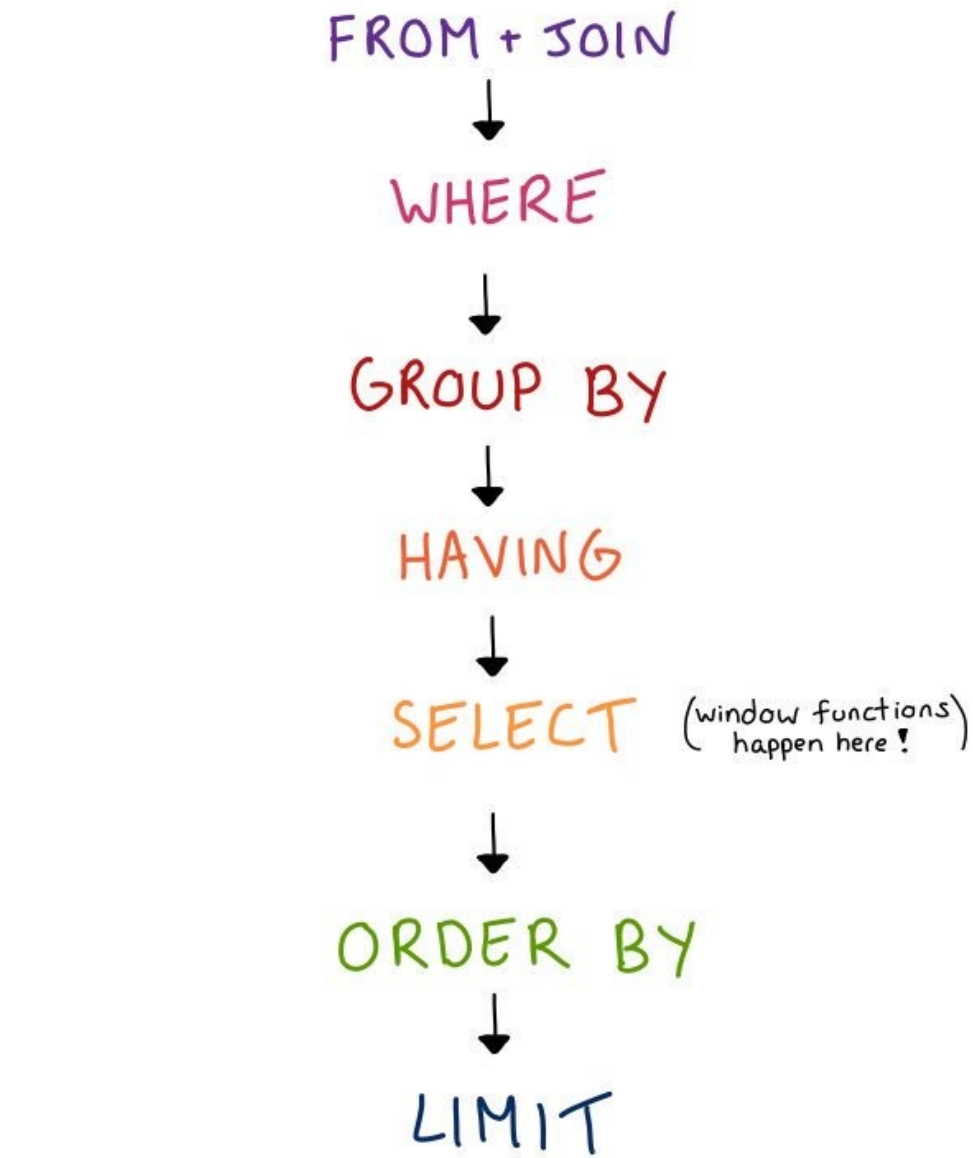
Ларри Элинсон

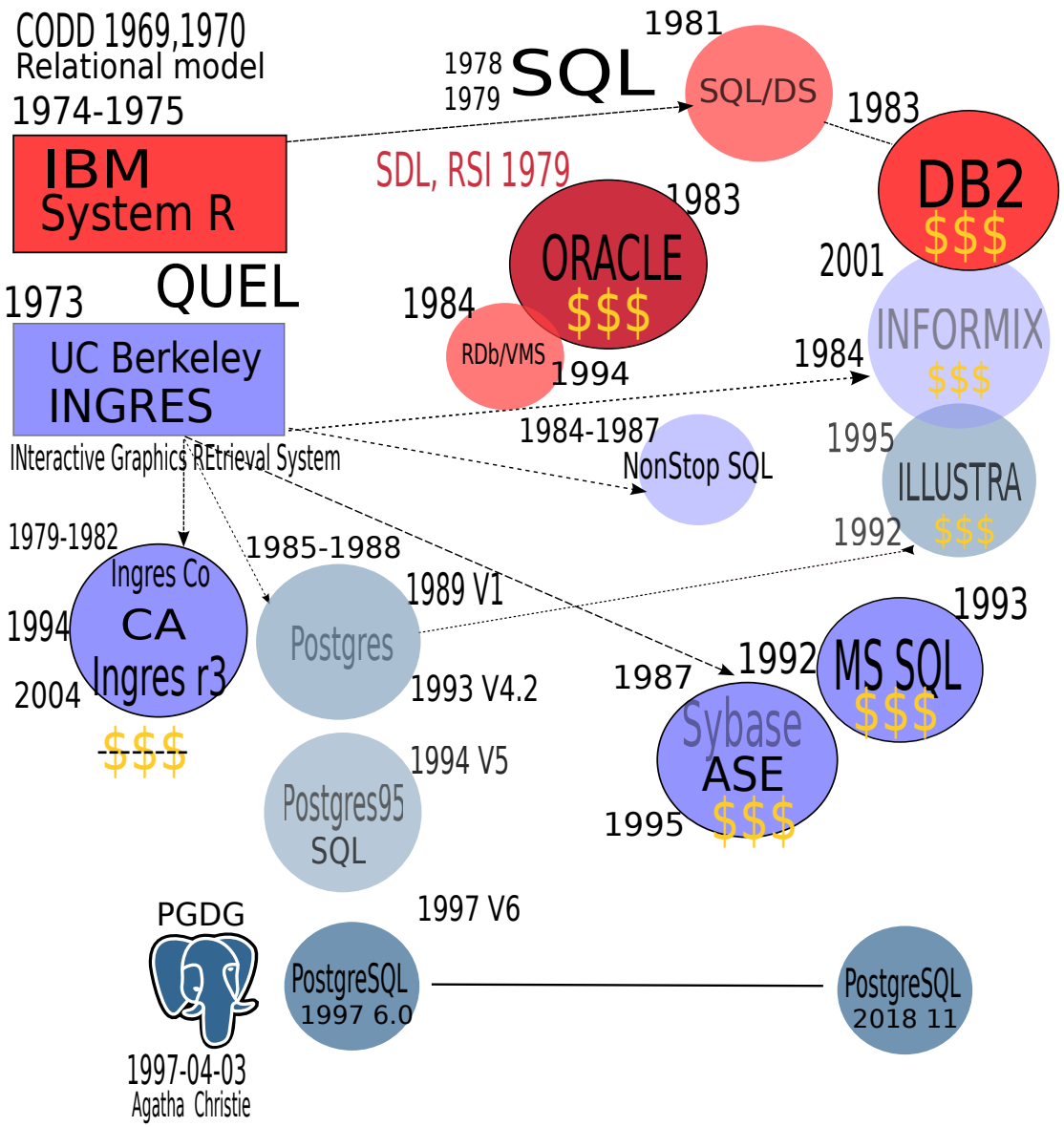
- Лучшие ребята из Стэнфорда
- Ставка на SQL
- Который и стал стандартом

- Бизнес.

Julia Evans
@b0rk

SQL queries run in this order




CODD 1969,1970
Relational model
1974-1975
IBM
System R
1978
1979
SQL
1981
SQL/DS
1983
DB2
$$$
SDL, RSI 1979
1983
ORACLE
$$$
1984
RDb/VMS
1994
2001
INFORMIX
$$$
1984
1973
QUEL
UC Berkeley
INGRES
INteractive Graphics REtrieval System
1984-1987
NonStop SQL
1995
ILLUSTRA
$$$
1992
1979-1982
Ingres Co
CA
Ingres r3
1994
2004
$$$
1985-1988
Postgres
1989 V1
1993 V4.2
1994 V5
Postgres95
SQL
1993
1992
MS SQL
$$$
1987
Sybase
ASE
$$$
1995
PGDG
1997 V6
PostgreSQL
1997 6.0
1997-04-03
Agatha Christie
PostgreSQL
2018 11

Postgres

- Stonebraker & Rowe. The design of Postgres (1986)
  https://doi.org/10.1145/16856.16888
- Stonebraker, Rowe & Hirohama. The implementation of Postgres (1990)
  https://doi.org/10.1109/69.50912
- Joseph M. Hellerstein. Looking Back at Postgres (2019)
  https://arxiv.org/abs/1901.01973

THE DESIGN OF POSTGRES

Michael Stonebraker and Lawrence A. Rowe

Department of Electrical Engineering
and Computer Sciences
University of California
Berkeley, CA 94720

Abstract

This paper presents the preliminary design of a new database management system, called POSTGRES, that is the successor to the INGRES relational database system. The main design goals of the new system are to:

1) provide better support for complex objects,

2) provide user extendibility for data types, operators and access methods,

3) provide facilities for active databases (i.e., alerters and triggers) and inferencing including forward- and backward-chaining,

4) simplify the DBMS code for crash recovery,

5) produce a design that can take advantage of optical disks, workstations composed of multiple tightly-coupled processors, and custom designed VLSI chips, and

6) make as few changes as possible (preferably none) to the relational model.

Чего хотел Стоунбрейкер

- Сложные объекты (неатомарные типы данных)
- Расширяемость типов данных, методов доступа, операторов.
- Элементы «Активной БД» (триггеры, правила, *alerters*)
- Развитие языка (PostQUEL)
- Рекурсивные запросы
- Компиляция запросов (использование планов)
- Process per user

Чего ещё хотел Стоунбрейкер

- VACUUM – перенос старых данных на оптический диск
- 64-битный transaction id (!)
- Исторические данные (версионность, снапшот)

Tuples will have all non-null fields stored adjacently in a physical record. Moreover, there will be a tuple prefix containing the following extra fields:

IID	: immutable id of this tuple
tmin	: the timestamp at which this tuple becomes valid
BXID	: the transaction identifier that assigned tmin
tmax	: the timestamp at which this tuple ceases to be valid
EXID	: the transaction identifier that assigned tmax
v-IID	: the immutable id of a tuple in this or some other version
descriptor	: descriptor on the front of a tuple

The descriptor contains the offset at which each non-null field starts, and is similar to the data structure attached to System R tuples [ASTR76]. The first transaction identifier and timestamp

И чего ещё хотел Стоунбрейкер

- Восстановление после сбоев дисков
- Лог транзакций с циклическим «ускорителем» в памяти

1. Supporting ADTs in a Database System
   a. Complex Objects (i.e., nested or non-first-normal form data)*
   b. User-Defined Abstract Data Types and Functions*
   c. Extensible Access Methods for New Data Types*
   d. Optimizer Handling of Queries with Expensive UDFs
2. Active Databases and Rules Systems (Triggers, Alerts)*
   a. Rules implemented as query rewrites†
   b. Rules implemented as record-level triggers†
3. Log-centric Storage and Recovery
   a. Reduced-complexity recovery code by treating the log as data,* using non-volatile memory for commit status†
   b. No-overwrite storage and time travel queries†
4. Support for querying data on new deep storage technologies, notably optical disks*
5. Support for multiprocessors or custom processors*
6. Support for a variety of language models
   a. Minimal changes to the relational model and support for declarative queries*
   b. Exposure of "fast path" access to internal APIs, bypassing the query language†
   c. Multi-lingual support†

Figure 1: Postgres features first mentioned in the 1986 paper* and the 1991 paper†.

От Postgres к PostgreSQL

- Postgres пошел в релиз в 1989 г.
- Andrew Yu, Jolly Chen, 1994-95г: Postgres95 (поддержка SQL)
- Тогда же: Postgres стал СПО.
- Вадим Михеев начал работать над MVCC
- 1996: PostgreSQL (6я версия вышла в начале 1997)

PostgresPro

Олег выходит на тропу

```
From megera@sai.msu.su Mon Sep 11 12:06:07 1995
Date: Mon, 11 Sep 1995 12:06:06 +0400 (MSK DST)
From: "O.Bartunov" <megera@sai.msu.su>
X-Sender: megera@ra
To: postgres <postgres95@nobozo.cs.berkeley.edu>
Subject: Q: Friendly interface for Postgres95
Message-ID: <pine.sv4.3.91.950911113736.16173a-100000@ra>
MIME-Version: 1.0
Content-Type: TEXT/PLAIN; charset=US-ASCII
Status: O
X-Status:

Hi,
now when postgres95 v.1.0 installed and passed all tests
(without regex :-()
```

NUMBER OF SHARES 1400
CERTIFICATE NUMBER A-10
CLASS A
DATE OF ISSUE SEP 30, 2000

CORPORATION POSTGRESQL, INC
REGISTERED HOLDER OLEG BARTUNOV
TRANSFER (OR ALLOTMENT) FROM Treasury

CERTIFICATE RECEIVED
DATE 20
Signature

A-10	A	1400
Share Certificate Number	Class of Shares	Number of shares

THIS CERTIFIES THAT OLEG BARTUNOV

is the registered holder of the above described fully paid shares in the capital of

POSTGRESQL, INC.

Incorporated under the "Canada Business Corporations Act"

IN WITNESS WHEREOF the Corporation has caused this Certificate to be signed by its duly authorized officer(s) this day of , 20 .

Внутри Postgres95

```
megera@ra:~/TT_DB$ ls -l
total 36
-rw-r--r-- 1 megera megera 1024 июн 16  1996 access_table.ind
-rw-r--r-- 1 megera megera 3072 июн 16  1996 access_table.rec
-rw-r--r-- 1 megera megera 2048 июн 16  1996 file_object_map.ind
-rw-r--r-- 1 megera megera 2048 июн 16  1996 file_object_map.rec
-rw-r--r-- 1 megera megera 2048 июн 16  1996 file_table.ind
-rw-r--r-- 1 megera megera 3072 июн 16  1996 file_table.rec
-rw-r--r-- 1 megera megera    0 июн 16  1996 file_table.var
-rw-r--r-- 1 megera megera 1024 окт  7  1996 property_table.ind
-rw-r--r-- 1 megera megera 3072 окт  7  1996 property_table.rec
-rw-r--r-- 1 megera megera 1024 окт  7  1996 property_table.var
megera@ra:~/TT_DB$
```

Первый патч Олега

```
commit 5b1311acfbd6b84dbb849752409142l4c51fcb48
Author: Marc G. Fournier <scrappy@hub.org>
Date:   Wed Apr 2 18:13:47 1997 +0000

    From: Oleg Bartunov <oleg@sai.msu.su>
    Subject: [HACKERS] locale patches !

    Hi there,

    here are little patches to get Postgres 6.1 works with locale stuff.
    This is a patch against 970402.tar.gz, there are no problem to apply them
    by hand to 6.0 release. Collate stuff tested about 1-2 months in real
    working database but I'm sure there must be no problem. US hackers
    could vote against locale implementation ( locale for sure will affect to
    speed of postgres ), so I introduce variable USE_LOCALE which
    controls locale stuff. Non-US users now could use ~* operator
    for searching and <order by> for strings with nation alphabet.
    Please, don't forget, as I did first time, to set environment variable
    LC_CTYPE and LC_COLLATE because backend get locale information from them.
    I start postmaster from a little script, assuming that shell is Bash shell
    it looks like:
```

Рождение PostgreSQL в письмах

From: "Vadim B. Mikheev" <vadim@sable.krasnoyarsk.su>
Date: Wed, 18 Sep 1996 15:04:22 +0800
Subject: Re: [PG95-DEV] postgres2.0

Marc G. Fournier wrote:

[...]

> How about PostgresV6?

And it's good name. But Ingres & Postgres are different in language (not only), and Postgres & Postgres'95 too, so why not to change prefix now again ?

> Please...though...**just not Postgres/SQL**...it just **doesn't slide off the tongue nicely** :(

Vote against too.

Получение имени

commit 9b41da6ce48e3bed6730faa6347a5461175cff83

Author: Bruce Momjian <bruce@momjian.us>

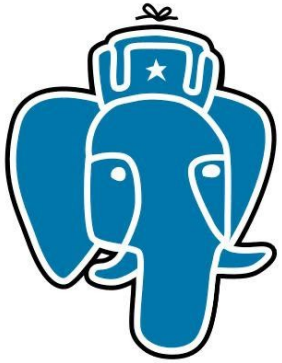
Date: Wed Dec 11 00:28:15 1996 +0000

Rename postgres95 to PostgreSQL.
Add comment for SELECT NULL

Русский след...

Slonik.gif,
Даниил Лундин,
Дмитрий Самерсов

https://obartunov.livejournal.com/186860.html



INTERNET ARCHIVE WayBackMachine http://www.ca.postgresql.org/images/ Go

1 captures
17 Dec 02 - 17 Dec 02

[IMG]	s3_off.gif	21-Sep-2001 10:28	1k
[IMG]	s3_on.gif	21-Sep-2001 10:28	1k
[IMG]	s4_off.gif	21-Sep-2001 10:28	1k
[IMG]	s4_on.gif	21-Sep-2001 10:28	1k
[IMG]	s5_off.gif	21-Sep-2001 10:28	1k
[IMG]	s5_on.gif	21-Sep-2001 10:28	1k
[IMG]	s6_off.gif	21-Sep-2001 10:28	1k
[IMG]	s6_on.gif	21-Sep-2001 10:28	1k
[IMG]	s7_off.gif	21-Sep-2001 10:28	1k
[IMG]	s7_on.gif	21-Sep-2001 10:28	1k
[IMG]	s8_off.gif	21-Sep-2001 10:28	1k
[IMG]	s8_on.gif	21-Sep-2001 10:28	1k
[IMG]	s9_off.gif	21-Sep-2001 10:28	1k
[IMG]	s9_on.gif	21-Sep-2001 10:28	1k
[IMG]	slonik.gif	21-Sep-2001 10:28	6k
[IMG]	so1.gif	21-Sep-2001 10:28	1k
[IMG]	so2.gif	21-Sep-2001 10:28	1k
[IMG]	so3.gif	21-Sep-2001 10:28	1k
[IMG]	so4.gif	21-Sep-2001 10:28	1k
[IMG]	so5.gif	21-Sep-2001 10:28	1k
[IMG]	spacer.gif	21-Sep-2001 10:28	1k
[IMG]	sqlephant.gif	21-Sep-2001 10:28	16k
[IMG]	ss.gif	21-Sep-2001 10:28	1k
[IMG]	tbkg.gif	21-Sep-2001 10:28	1k
[IMG]	title.gif	21-Sep-2001 10:28	2k
[IMG]	top.gif	21-Sep-2001 10:28	10k
[IMG]	verh1.gif	21-Sep-2001 10:28	1k
[IMG]	verh2.gif	21-Sep-2001 10:28	5k
[IMG]	verh3.gif	21-Sep-2001 10:28	1k
[IMG]	zaglushka.gif	21-Sep-2001 10:28	1k

Этапы развития

1996 — Стабилизация работы
1997 (6.1) — Интернационализация +World
2005 (8)
2010 (9) — Встроенная репликация +Enterprise users
2014 (9.4) — jsonb +NoSQL users
2016 (9.6) — Parallel Query +OLAP users
2017 (10) — Logical Replication, Declarative Partitioning
2018 (11) — JIT
2019 (12) — Pluggable storage, Jsonpath (SQL/JSON)
2022 (15) — JSON_TABLE,...(SQL/JSON), MERGE
2023 (16) — Pluggable TOAST, Fast Jsonb

MVCC

- Выпилить историческое хранилище Стоунбрейкера
- Уровни изоляции транзакций
- Сделал Вадим Михеев. (См).

```
Date: Mon, 09 Aug 1999 16:08:19 +0800
From: Vadim Mikheev
To: Oleg Bartunov
Subject: Re: indices grow !


----skipped ------
А что его понимать-то! -:)
Основное что надо помнить:

Запрос (те Query - то что читает записи из базы используя Seq/Index
scans и отбирает их в соответствии с условиями в WHERE используя
joins, subselects etc) видит (те возвращает) только те записи,
который были живы в момент старта
запроса(READ COMMITTED)/транзакции(SERIALIZABLE).

Всё остальное лишь производное -:)
```

WAL

- Сделал Вадим Михеев.
- Основывается на ARIES (1992), IBM (C.Mohan).
  - Запись вначале в WAL, на надёжный сторадж
  - Проигрывание истории после крэша
  - Rollback тоже, конечно же, пишется в WAL
  - Позже это пригодилось для репликации.

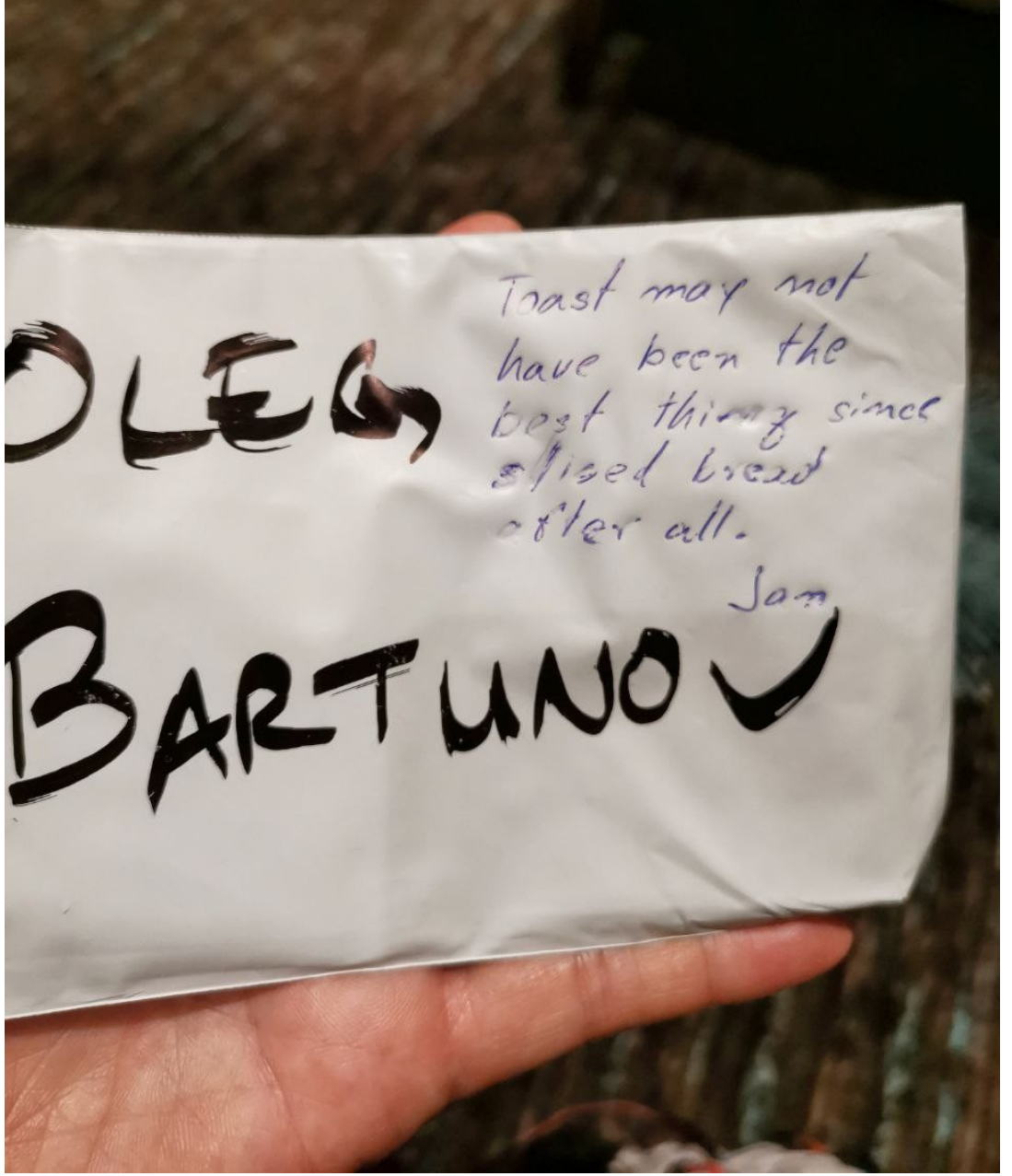
```
commit 47937403676d913c0e740eec6b85113865c6c8ab
Author: Vadim B. Mikheev <vadim4o@yahoo.com>
Date:   Wed Oct 6 21:58:18 1999 +0000

    XLOG (also known as WAL -:)) Bootstrap/Startup/Shutdown.
    First step in cleaning up backend initialization code.
    Fix for FATAL: now FATAL is ERROR + exit.
```

Дальнейшее развитие MVCC

- Проблема Write Amplification
  - HOT
  - Восходящее удаление
- Проблема определения видимости версий
  - hint bits
  - ProcArray vs CSN

История TOAST

- Был написан Jan Wieck, Dec 1999
- С тех пор не менялся, разве что:
  - Частичное чтение (Partial TOAST decompression)
  - Идёт обсуждение Pluggable TOASTers.



История GiST

- Разработан в Berkeley для PostgreSQL командой проф Joe Hellerstein http://gist.cs.berkeley.edu/
- Затем полностью переписан О.Бартуновым и Ф.Сигаевым в начале XXI в.
- GiST — обобщенная структура, поверх реализуются разные типы деревьев. RD-дерево для поиска в множествах.

Chapter 36. GiST Indices

Author: This extraction from an e-mail sent by Eugene Selkov Jr. contains good information on GiST. Hopefully we will learn more in the future and update this information. - thomas 1998-03-01

Well, I can't say I quite understand what's going on, but at least I (almost) succeeded in porting GiST examples to linux. The GiST access method is already in the postgres tree (`src/backend/access/gist`).

Examples at Berkeley come with an overview of the methods and demonstrate spatial index mechanisms for 2D boxes, polygons, integer intervals and text (see also GiST at Berkeley). In the box example, we are supposed to see a performance gain when using the GiST index; it did work for me but I do not have a reasonably large collection of boxes to check that. Other examples also worked, except polygons: I got an error doing

```
test=> create index pix on polytmp
test-> using gist (p:box gist_poly_ops) with (islossy);
ERROR:  cannot open pix
```

Sven Helmer. Index Structures for Databases Containing Data Items with Set-valued Attributes (1997)

Hellerstein о том, как Олег и Федор переписали GiST

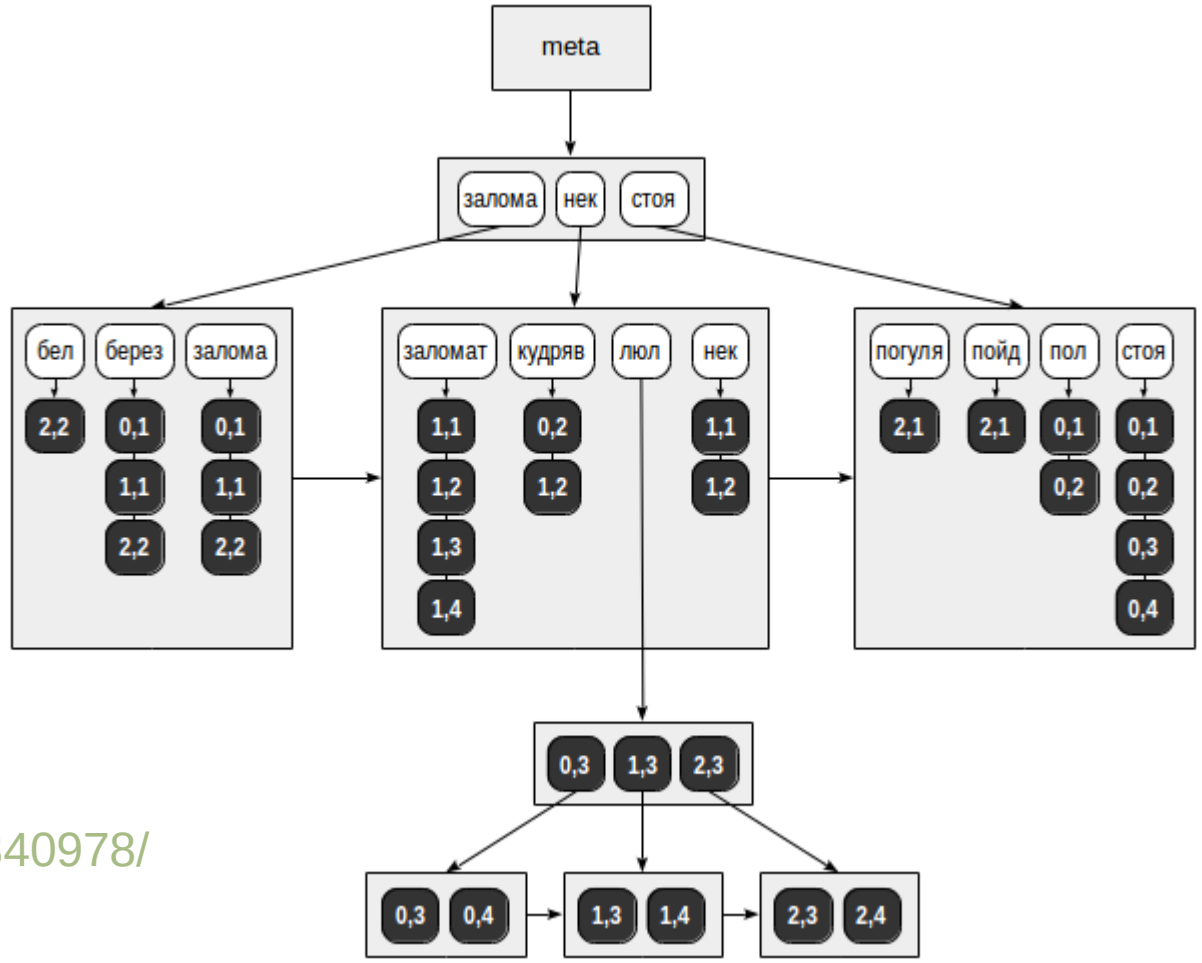
nity made significant modifications and improvements to the core of the system as well, from the optimizer to the access methods and the core transaction and storage system. Since the mid-1990s, very few of the PostgreSQL internals ca[illegible]out of the academic group at Berkeley—the last contribution may have been my GiST implementation in the latter half of the 1990s—but even that was rewritten and cleaned up substantially by open-source volunteers (from Russia, in that case). The open source community around PostgreSQL deserves enormo[illegible]credit for running a disciplined process that has soldiered on over decades to produce a remarkably high-impact and long-running project.

Довольные авторы GiST

- Проблема GiST (точнее, проблема superimposed сигнатур) — переполнение этих сигнатур.

GIN-индексы

- Обобщенный обратный индекс. Сигаев, Бартунов (8.4, 2008)
- Ключ => B-дерево из xid



https://habr.com/ru/company/postgrespro/blog/340978/

История репликации

- Transaction log shipping (Mammoth Replicator, Slony)
- Proxy-based replication (PgPool)

  Replication is not a single solution for a single problem; it is several solutions for a wide array of different problems.No one replication tool will ever be the "default" replication for PostgreSQL

  Josh Berkus, 2004

- WAL-based replication: Hans-Jürgen Schönig (idea, 2005). Релиз 9.0 (2010)
- Logical replication (pglogical, 9.4, 2015)

История репликации (2)

- Postgres XC
- Postgres XL
- BDR
- Postgres Pro Multimaster
-

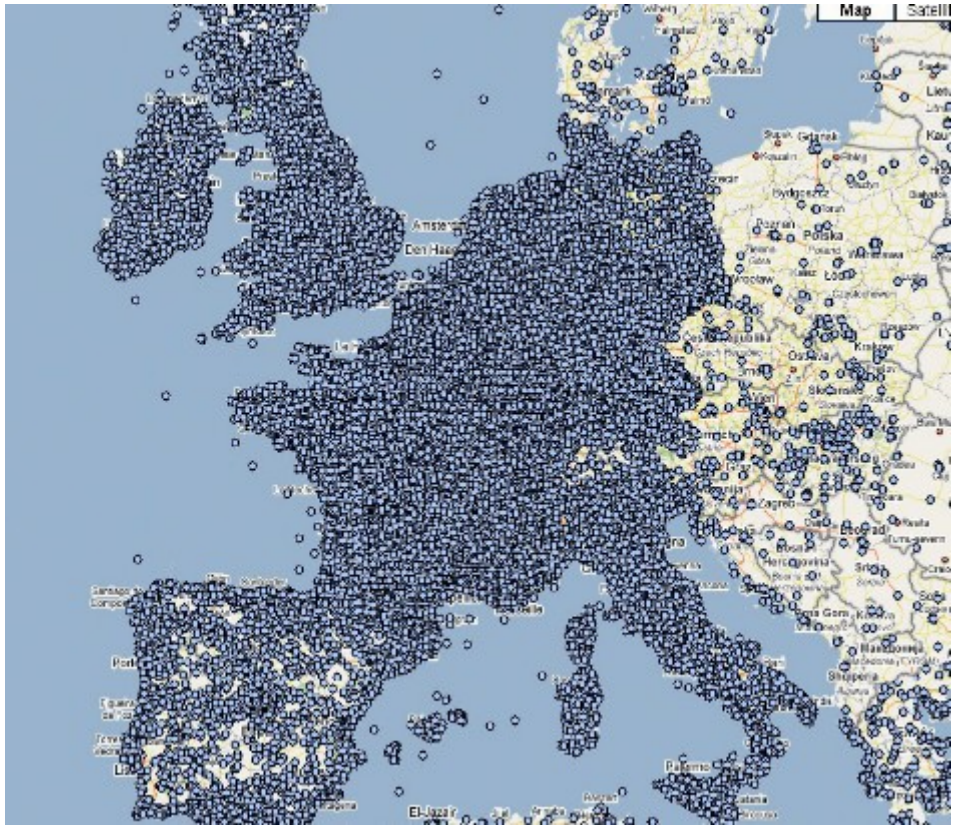
История JSON(,B!)

- Intarray, GiST (2000, Бартунов, Сигаев, Rambler)
- HStore с 2003 (с 2006 в PostgreSQL, Бартунов, Сигаев)
- JSON как текст 9.2
- JSONB (9.4) и GIN-индексы (Бартунов, Коротков, Сигаев)
- SQL/JSON (12… ? Бартунов, Коротков, Глухов)
- GSON ?

KNN

- 2005--2010. GiST, PostGIS
- 2020. SP-GiST (А. Лебедев)
- …… B-Tree

Бартунов, Сигаев, Глухов



PostgresPro

История полнотекстового поиска

- SQL-based
- OpenFTS (2000..2009, 7.4) - поиск на базе GiST (Сигаев, Бартунов)
- Tsearch
- Tsearch2
- Встроенный поиск
- GIN-индекс

История резервного копирования

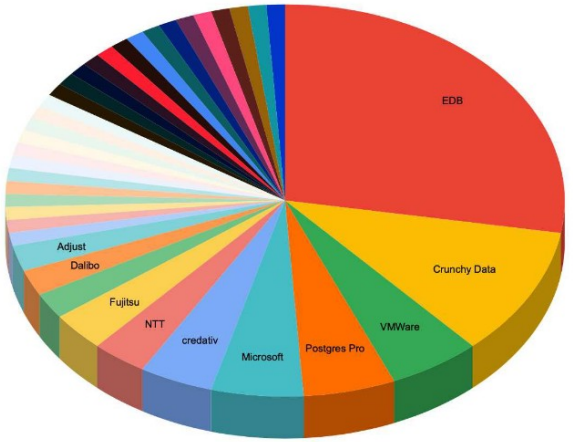
- pg_dump
- pg_basebackup + WAL archiving
- PITR
- pg_probackup + PTrack : инкрементальный Backup

Расширяемость

- Типы данных (tsvector, json, xml, pgsphere, PostGIS...)
- Индексы (GiST, SP-GiST, GIN, BRIN)
- Языки программирования (PL/Perl,Python,V8,Java).
- Табличные методы доступа (Heap...)
- TOAST (в процессе...)
-

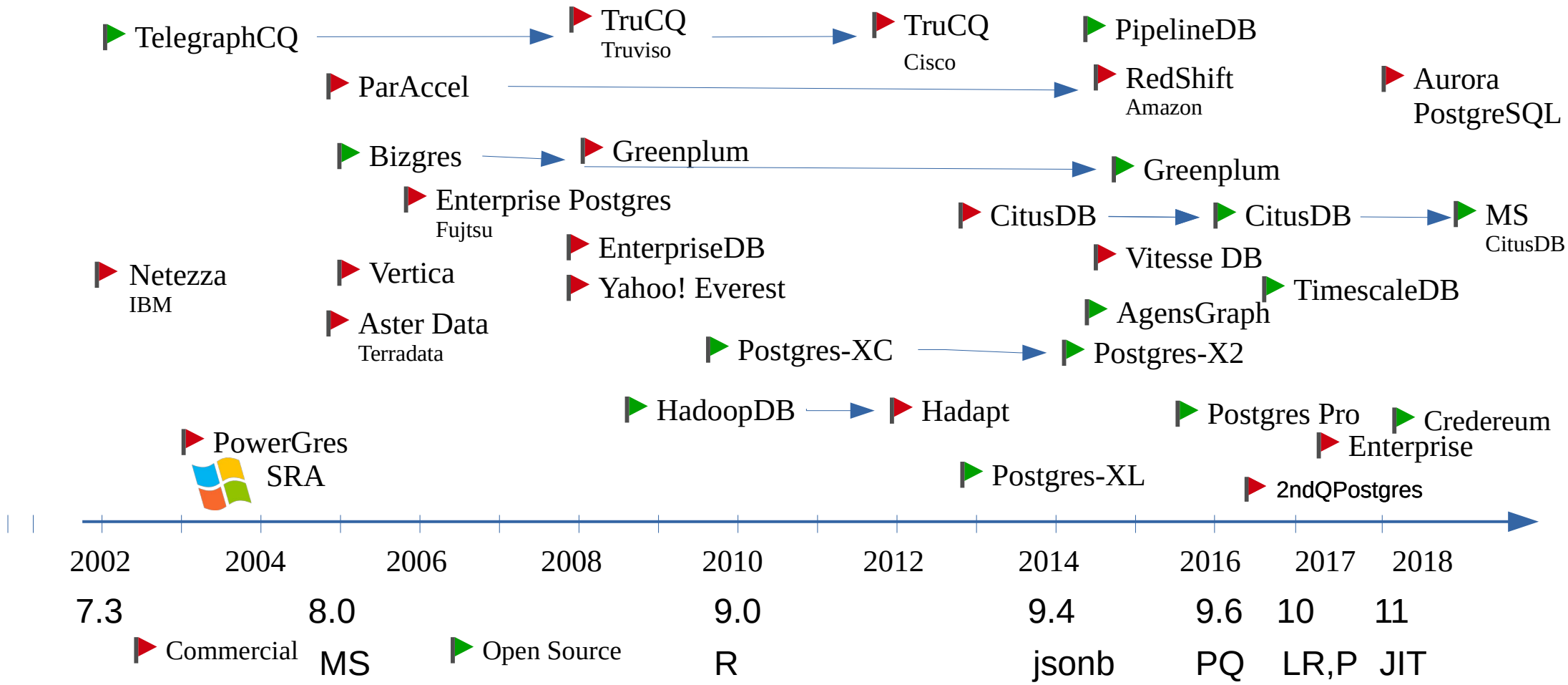
Сообщество

- Пермиссивная лицензия
  - Форки
  - Бэкпорт
- Эволюция
  - от гиков до корпораций
- Баланс между developer-driven и customer-driven



EDB Crunchy Data VMWare Postgres Pro Microsoft credativ NTT Fujitsu
PeopleDoc Dalibo Adjust Yandex University of Cambridge True Software Swarm64
SRA OSS Rakuten Postgres Experts Playtech Paragon
Open Standards Promotion Center, Japan Ongres MongoDB Minsait iovation ILande
Huawei HeteroDB Google Facebook Deriv.com Cybertec
conova communications GmbH Brandwatch Basefarm AS Appen Amazon Afilias

Форки


TelegraphCQ
TruCQ
Truviso
TruCQ
Cisco
PipelineDB
ParAccel
RedShift
Amazon
Aurora
PostgreSQL
Bizgres
Greenplum
Greenplum
Enterprise Postgres
Fujtsu
CitusDB
CitusDB
MS
CitusDB
EnterpriseDB
Vitesse DB
Netezza
IBM
Vertica
Yahoo! Everest
TimescaleDB
AgensGraph
Aster Data
Terradata
Postgres-XC
Postgres-X2
HadoopDB
Hadapt
Postgres Pro
Credereum
Enterprise
PowerGres
SRA
Postgres-XL
2ndQPostgres
2002
2004
2006
2008
2010
2012
2014
2016
2017
2018
7.3
8.0
9.0
9.4
9.6
10
11
Commercial
MS
Open Source
R
jsonb
PQ
LR,P
JIT

Торонто 2006

PostgresPro

Спрашивайте
Postgres Pro
и приходите на PgConf.RU !

PostgresPro