How we made WarehousePG Open Source (again)

Or: What happened to Greenplum?

Andreas Scherbaum

PostgreSQL Berlin Meetup

Berlin, 2026-03-05



Who am I

Andreas Scherbaum

- Works with databases since ~1997, with PostgreSQL since ~1998
- Founding member of PGEU
- Board of Directors: PGEU, Orga team for pgconf.[eu|de], FOSDEM
- Ran my own company around PostgreSQL for 7+ years
- Joined EMC in 2011
  - then Pivotal, then EMC, then Pivotal
  - working on PostgreSQL and Greenplum projects
- Joined Adjust GmbH in Berlin in 2020
  - Head of (around 1000) Databases, and a great team
- Now at EDB
  - Principal Engineer, WarehousePG

Get Social ...

- Mastodon: https://mastodon.social/@ascherbaum
- LinkedIn: https://www.linkedin.com/in/andreas-scherbaum-87b6663/
- ~~Twitter~~, Instagram: @ascherbaum
- GitHub: andreasscherbaum
- Blog: https://andreas.scherbaum.la/



Blog



Mastodon

Agenda

- What is WarehousePG?
- What problems does it solve?
- What about this Open Source story?
- Future outlook

What is WarehousePG?

What is WarehousePG

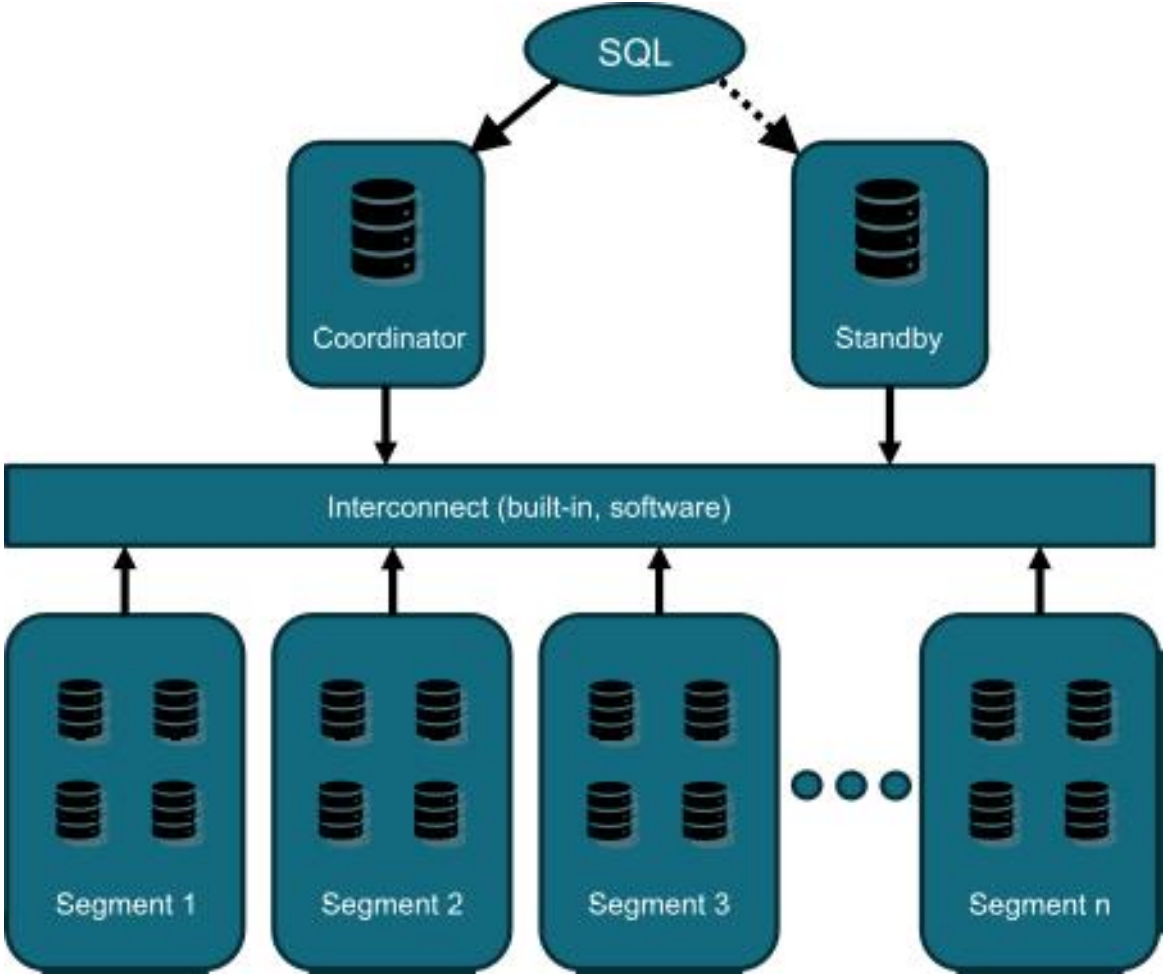
- A
  - Massive Parallel Processing (MPP)
  - Shared Nothing
  - Open Source
  - Hardware Agnostic
  - SQL
- Database

Massive Parallel Processing (MPP)

Massively parallel is the term for using a large number of computer processors (or separate computers) to simultaneously perform a set of coordinated computations in parallel.

-- Wikipedia (Massively parallel)

Shared Nothing



Segment Hosts :

- Have their own CPU, memory and disk (shared nothing)

Open Source (again)

- WarehousePG is Open Source since May 2025
  - Apache License

Hardware Agnostic

- Runs on commodity hardware
- Benefits from faster computing, storage and network

SQL

- PostgreSQL compatible
  - Mostly
- Almost fully compliant with SQL:1992
- Most of the features from SQL:1999
- Some features from SQL:2003

Other Shared-Nothing Databases

- **Greenplum**: WarehousePG is forked from Greenplum, functionality is similar
- **Greengage Database**: fork by Arenadata, out of Russia
- **Cloudberry**: fork by HashData, out of China
  - now an Apache Incubator project: Apache Cloudberry
- **open-gpdb**: Yandex Cloud

Other Shared-Nothing Databases

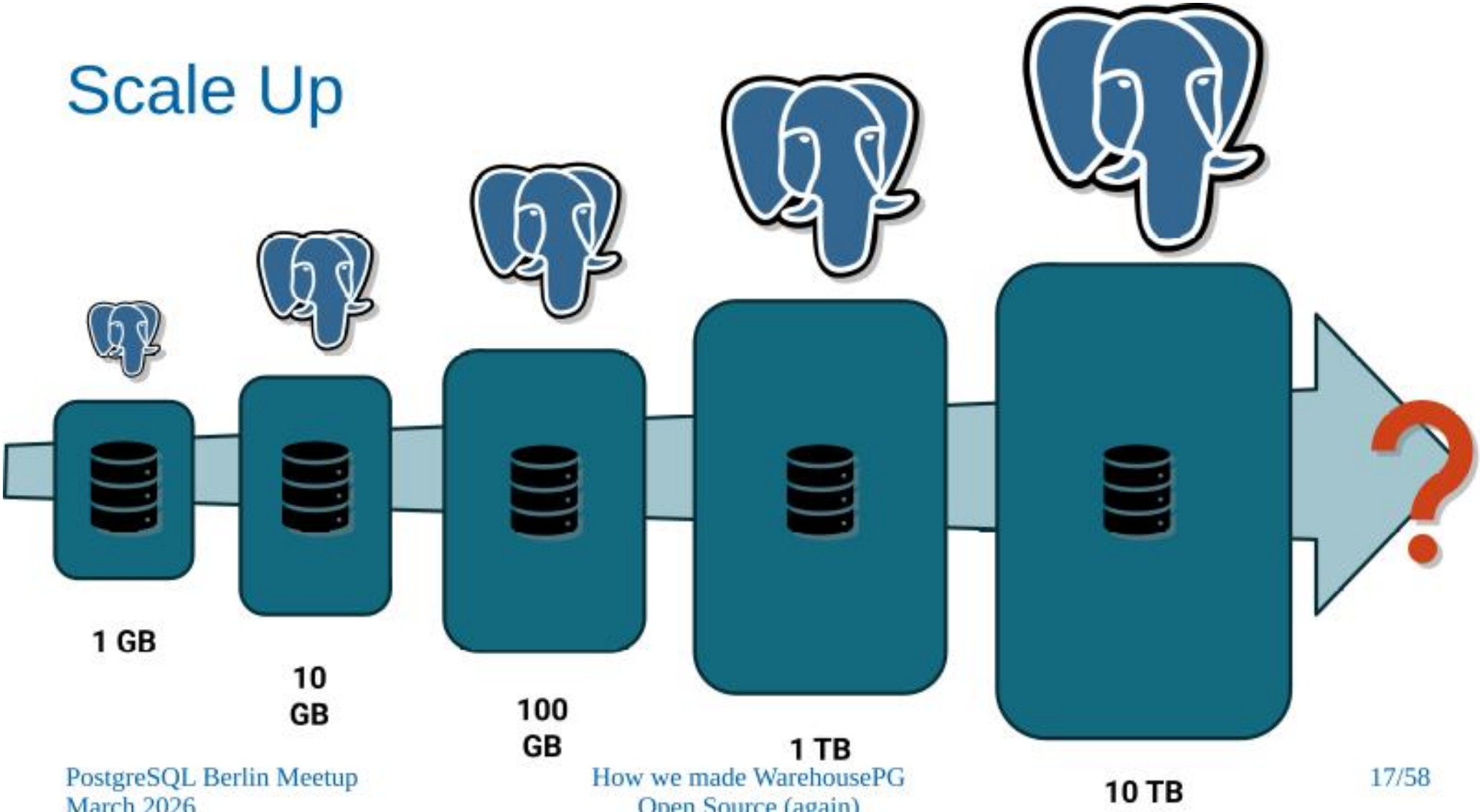
- Teradata: One of the earliest MPP database systems, designed for data warehousing and analytics.
- Netezza (now IBM): Early PostgreSQL fork (7.x), long-time appliance-based data warehousing platform.
- Vertica (now OpenText): columnar storage, MPP analytic database
- Exasol: An in-memory, MPP database management system
- ScyllaDB: A NoSQL database compatible with Apache Cassandra
- Amazon Redshift: A fully managed, petabyte-scale data warehouse service in the
- Snowflake: A cloud-based data warehousing platform that uses a hybrid architecture
- Google BigQuery: A fully managed, serverless data warehouse
- CockroachDB: A distributed SQL database
- TiDB: An open-source, distributed SQL database
- YugabyteDB: A distributed SQL database with PostgreSQL and Cassandra compability

What problems does it solve?

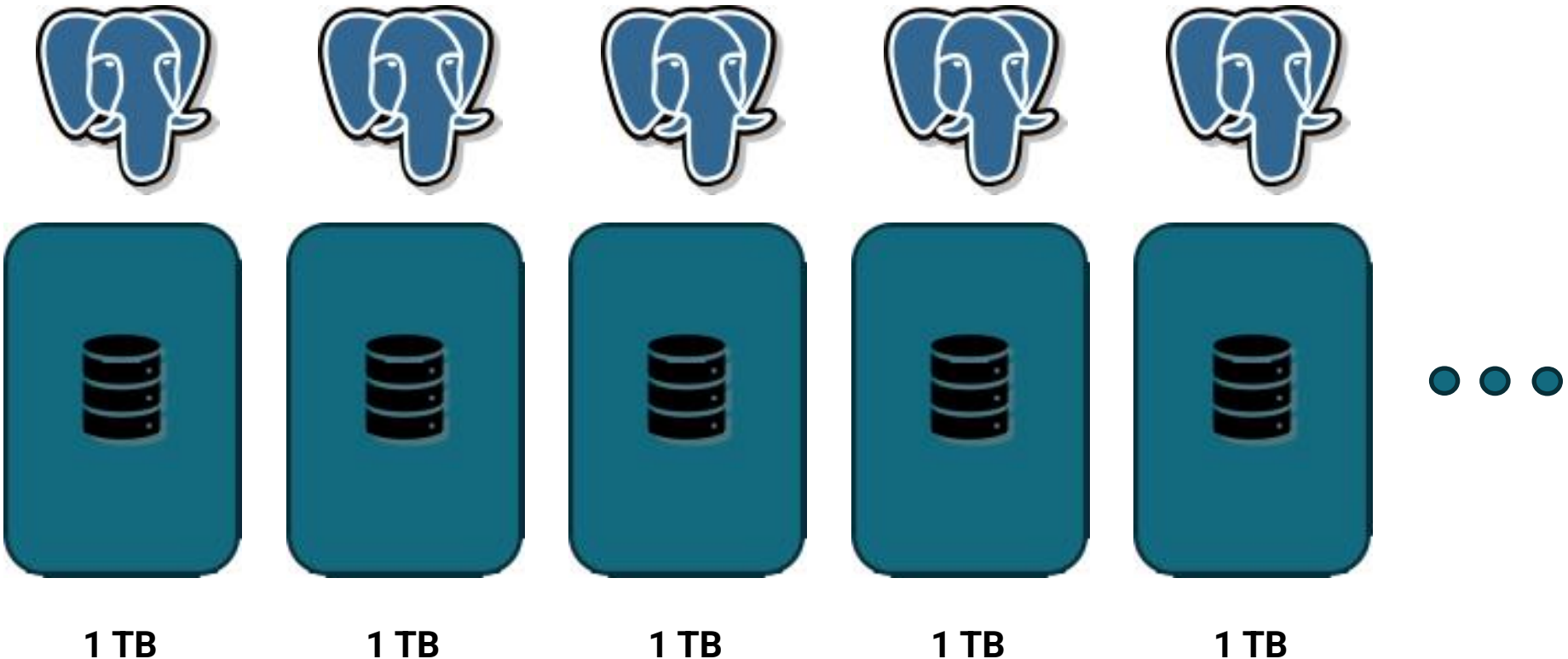
Poll ...

- Largest PostgreSQL Database you know of?

Scale Up



Scale Out



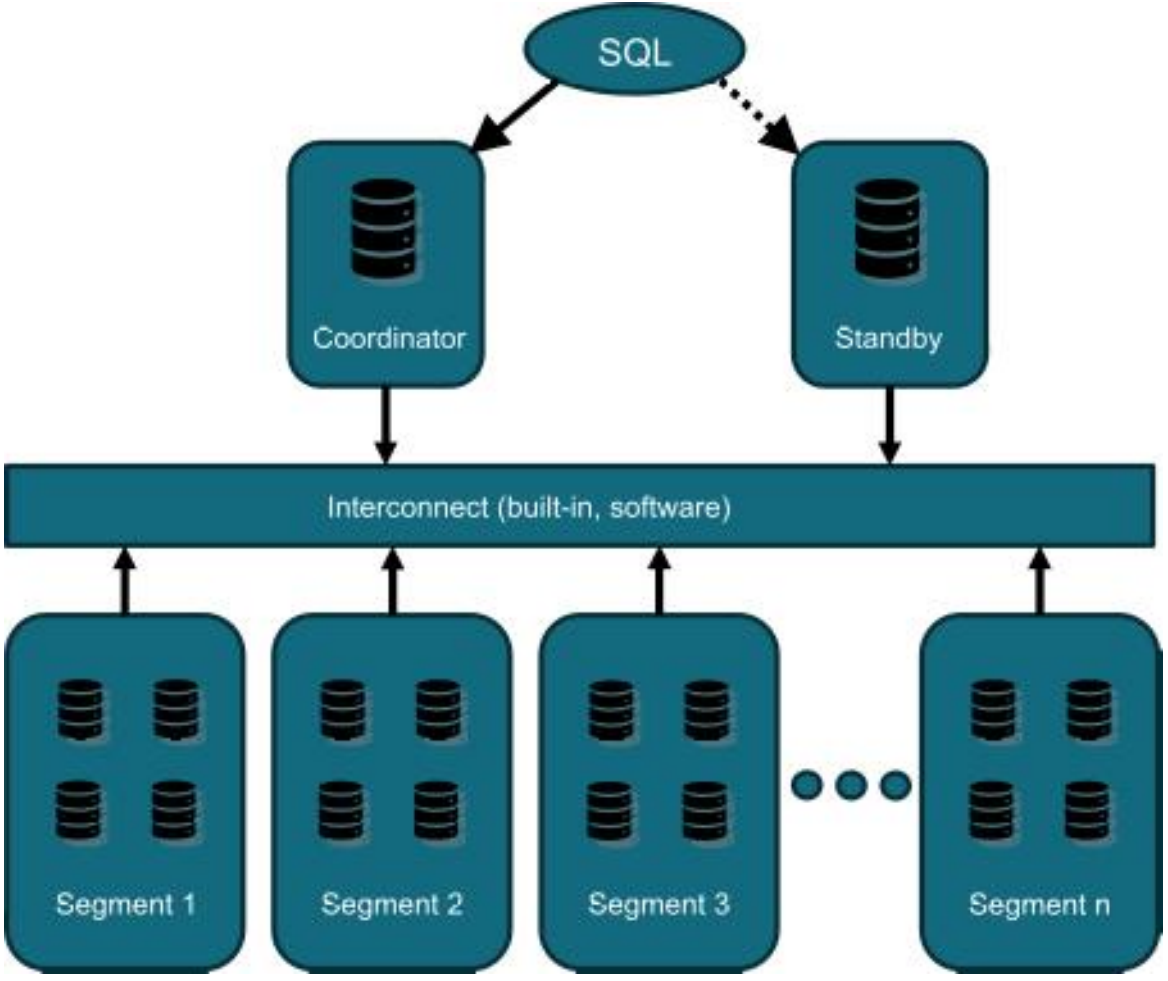
Scale Out

Clients connect to Coordinator Host which coordinates work with Segment Hosts

Built-in Interconnect software for continuous pipelining of data processing: connects all segments, Coordinator + Standby

Segment Hosts :

- Have their own CPU, memory and disk (shared nothing)
- Manage data and process queries
- Provide HA through data mirroring
- Can scale up/out



What problems does it solve

- Large scale analytics
- Compatibility with other applications
- Data Warehouse
- Data Silo
- Hot/warm/cold data storage
- Lower Total Cost of Ownership (TCO)
- In-database Machine Learning

Large scale analytics

- Database sizes in TB and PB
- Faster queries by scaling out

```
pbtest=# \dt+ data_table
                              List of relations
 Schema |    Name    | Type  |  Owner  | Storage |  Size   | Description
--------+------------+-------+---------+---------+---------+-------------
 public | data_table | table | gpadmin | ao_row  | 1019 TB |
(1 row)
```

Compatibility with other applications

- OLTP: PostgreSQL
- OLAP: WarehousePG
- Same drivers, same applications, same SQL queries
  - Different performance

Data Warehouse

- Central “Source of Truth” for a company
- Load all company data
  - And run all reports on one data set

Data Silo

- Department Data Warehouse
- Regular sync with company data warehouse

Hot/warm/cold data storage

- External Tables: allow reading/querying/writing external data sources
- Looks like a table for applications

Lower Total Cost of Ownership (TCO)

- It is Open Source
  - Build your own support model



In-database Machine Learning

- pgvector
- PGAI
- Apache MADlib

Open Source? Again?

2015-02-17: Announcement

- February 17th: Pivotal announces that it will Open Source most of it's products
  - Under Paul Maritz (CEO)







2015-10-29: Greenplum Open Source



2015-10-28: Greenplum Open Source

- October 28th: Greenplum is Open Source
  - During PostgreSQL Conference Europe 2015 (Vienna)



https://news.ycombinator.com/item?id=10462339

What happened between February 2015 and October 2015?

What is Greenplum Database?

- One of these PostgreSQL forks
  - Diverged away a long time ago (2007)
- And now wants to merge again (2015)





Open Source Preparations

- Check source for customer names
- Check patents
- Polish build system
- Find governance model
- Find release date
- Marketing …

2015 - 2023

- Update kernel version from PostgreSQL 8.2 to 12
  - Greenplum 5 (2017): based on PostgreSQL 8.3
  - Greenplum 6 (2019): based on PostgreSQL 9.4
  - Greenplum 7 (2023): based on PostgreSQL 12

2022-05-26: Broadcom to Acquire VMware

- Broadcom intends to acquire VMware
  - Transaction closed 2023-11-22
  - End of perpetually licenses 2023-12-13
  - subscription-based offerings


Broadcom's $61B VMware buy ranks as 3rd-largest tech deal

May 2024: Close Source

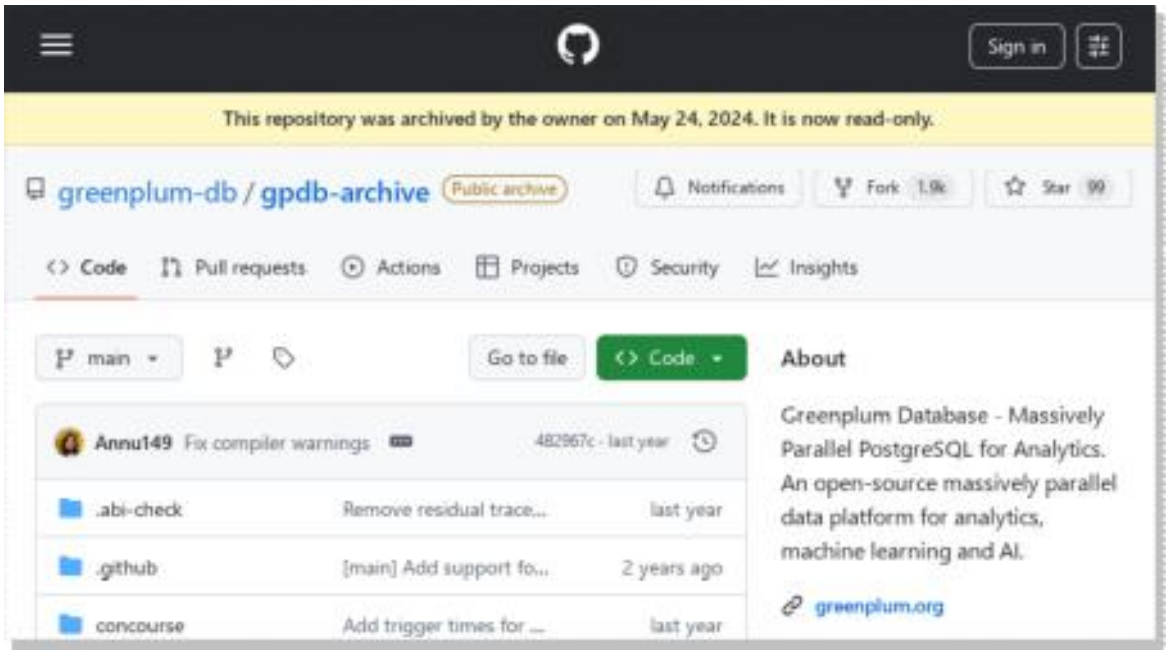
- Broadcom did:
  - Archive almost all GitHub repositories
  - Delete the Slack space for users
  - Mailing lists went silent
  - Deactivate CI/CD-pipelines (ConcourseCI)



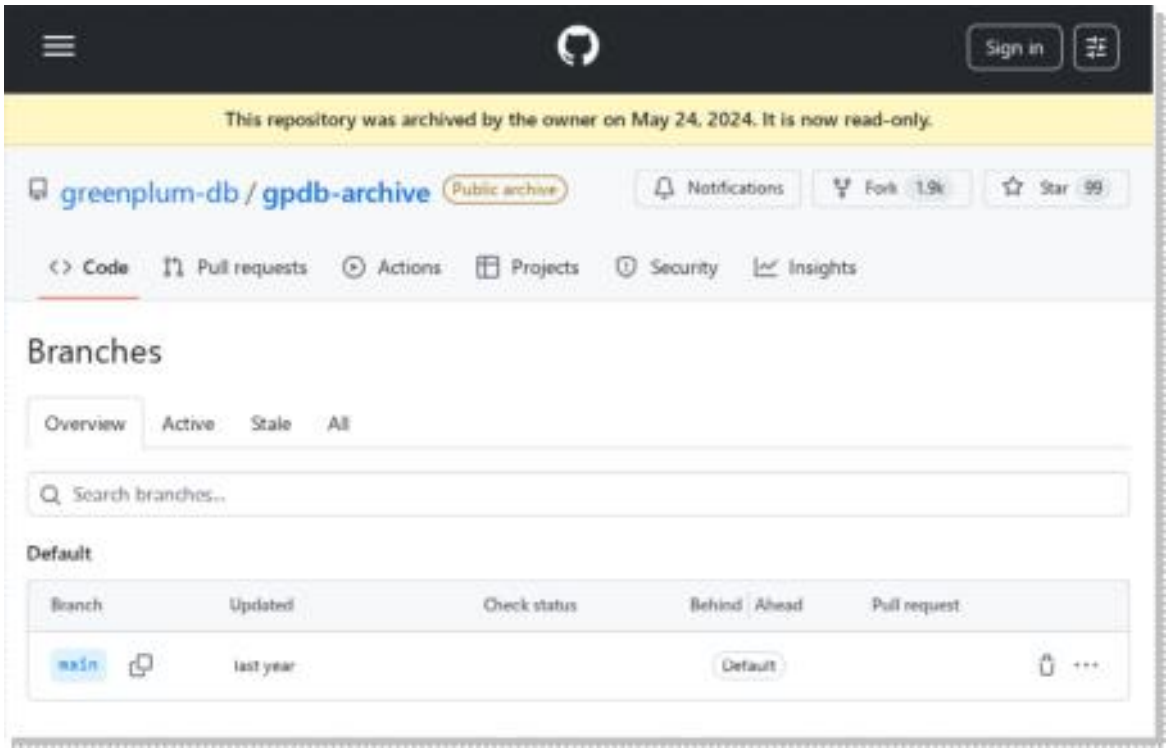
Broadcom Has Stored The Greenplum DBMS Repository In Their Archives

June 2, 2024 by Lee Michaelis

May 2024: Close Source



May 2024: Delete all branches and tags



May 2024: Delete all branches

- Only a single git commit timeline
- No release information
- No tag information
- Most likely: a new repository just with the main branch
- Mostly useless

Is it all lost?

Cloudberry

- Another, earlier Greenplum fork
- Upgraded kernel to PG 14 already
  - Not easy to support GP 6.x and 7.x on this
- Not a lot of activity in the repository
- Now an Apache Incubator project

Couple other repositories

- Questionable sources, unknown GitHub accounts
- Legal says “No”

gp-releng repositories

- Pivotal Release Engineering
- Probably long lost and forgotten
- Has almost up to date tags and releases

Lawyers (1)

- You can’t use that repository as source
  - But you can import the tags from it

Engineering (1)

```
git remote add origin <release engineering repo>

git fetch origin --prune --tags

git remote rm origin

git tag --list
```

Engineering (1)

- Turns out:
  - Importing tags also imports code for the tags
    - At least code which is not already in `main`
  - About 20k additional commits imported

Lawyers (2)

- Proof to us that this is not malicious
  - We checked checksums of files and commits against all kind of repositories

`https://github.com/warehouse-pg/warehouse-pg/blob/main/PROVENANCE.md`

Lawyers (2)

- Finally:
  - Approval to go ahead

Status of Source Code

- Released as Open Source in May 2025
- gpdb-archive as base
  - Tags imported from Pivotal releng
- No Contributor License Agreement (CLA)
  - We can’t change the license

Conclusion

Conclusion

- Broadcom made it Closed Source (license allows this)
- License allowed us to make it Open Source again
  - Also allowed us to import missing code (also open source)
- Project is alive



Outlook

Additionally

- Provide Red Hat RPMs
  - No Debian/Ubuntu packages yet, no requests so far
- Blackduck scans: identify risks in source code
  - Outdated libs, known CVEs
- Sonarcube scans: identify problems through code analysis

Working on

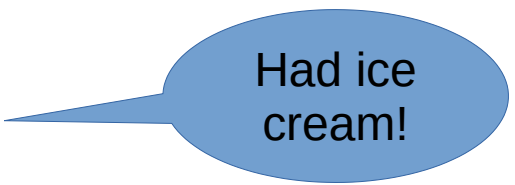
- CI pipelines ✅
- Backup Tools
- Observability (WEM)
- High Availability (WHPG-DR)
- DB kernel upgrades
- Tools and drivers compatibility

Questions?



Questions (already answered)

- Can I get the slides?
  - Yes, on my website (blog).
- Can I use WarehousePG?
  - Sure, please clone the repo
- Can I contribute?
  - Yes, it is Open Source
- ~~No ice cream today?~~
  - ~~Unfortunately not~~



WarehousePG

Blog

The End