{* **DATABASES** *}

DuckDB, database wrangler used by Google, Facebook, and Airbnb, hits 0.5.0

System look well suited to fill gaps in traditional OLAP market

Lindsay Clark   Fri 9 Sep 2022 // 09:45 UTC

11 🗨

DuckDB – the in-process analytical database management system used by Google, Facebook, and Airbnb – has released its 0.5.0 iteration.

The brainchild of academics at Amsterdam's Centrum Wiskunde & Informatica mathematical and theoretical computing research center, DuckDB is embedded within a host process. There is no DBMS server software to install, update or maintain.

For example, the DuckDB Python package can run queries directly on data in Python software library Pandas without importing or copying data. Written in C++, DuckDB is free and open source under the MIT License.

Consultancy and support are provided by DuckDB Labs. Co-founder and CEO Hannes Mühleisen, who also co-authored the code and maintains the project, told *The Register* it was inspired by SQLite, the serverless OLTP database engine, where he saw the opportunity for a similar approach, but for analytics.

"We were working a lot with practitioners in data science and they all had these problems that weren't theoretical problems anymore in the computing research – they were solved ages ago – but somehow the software just wasn't there for them. With the commercial software vendors, the technology was in some of these packages, but not accessible or hidden behind many, many layers of enterprise bullshit," he said.

Mühleisen and his co-founder began to realize that a rethink of the database architecture might be necessary for OLAP. "We took that idea of in process data management systems where the entire database manager runs within the process that you're in – for example, Python or even Excel – and we redesigned a system to be the first in class for OLAP using this approach," said Mühleisen, who is still a senior researcher at his academic institution.

DuckDB is also often used as a part of a broader analytics or data management stack. For example, if someone builds a custom application that collects data, and then wants to build a SQL interface, in the past they might have to copy the data and move it into another system, which might cause synchronization issues, he said. But DuckDB can query third-party datasets as if it was its own data. "You can engineer that on top of an existing application or dataset. And people do," he said.

The system's popularity among data tool builders has even prompted its own meme.

The first release was in 2019 and has since been steadily gaining in popularity, with users including Google, Facebook, and Airbnb.

This week the project released its 0.5.0 iteration.

- Open source databases: What are they and why do they matter?
- Teradata takes on cloud-native rivals with data lakes, MLOps
- Cloudera launches SaaS platform for the lakehouse crowd
- Ant Group's in-house DB set for global release, including Raspberry Pi edition

Highlights among new features include "out of core," which aims to solve the problems that can occur when in-flight data is bigger than memory by offering intermediate results. The project has also added join order optimization, a perennial problem in analytical databases. Hyoun Park, CEO and chief analyst at Amalgam Insights, said DuckDB's differentiation comes from being a small application that works within code-based processes to analyze large data stores quickly.

"This is increasingly important as workloads are distributed, performance is needed across a variety of analytic use cases, and as analytic data continues to double year over year in large organizations," Park said. "As an open source database that is easily embeddable within specific analytic jobs, DuckDB is well suited to fill in the gaps where traditional monolithic OLAP databases are more rigid, more expensive, or require transfer and duplication efforts to support analytic variety.

"DuckDB can often run queries directly on data without intermediate processing, which improves processing. From a pure technology perspective, it is somewhat similar to Actian Vector, which also takes a columnar vectorized OLAP query approach, though Actian is designed to bring in data rather than work within a specific process or workload."

But there are clear limits on when and where the system should and should not be used. Although in some ways it offers a cheap alternative to a data warehouse, and could offer each data scientist a system on their laptop, it does not necessarily replace enterprise data warehouse systems from companies such as Teradata, Oracle, and IBM. The home page clearly states that it should not be used for "large client/server installations for centralized enterprise data warehousing."

"It's a question of the priority for your organization or the data problem. Is it really dependent on everybody working on the same data? If so, then maybe this is not the best solution," Mühleisen said.

This being open source databases, the project arrives with an unusual name. While CockroachDB was named after its supposed unkillable nature, and MongoDB was a contraction of "humongous," DuckDB was of course named after Mühleisen's pet Wilbur, who has, incidentally, appeared in *The Guardian* newspaper.

The project is working towards its 1.0 release, after which backwards breaking changes will not feature. "I think we're getting there with a lot of work. We always say by the end of the year, but I fear is not going to happen this year," Mühleisen said. ®

Sponsored: Just how do you define software defined storage?

🔗 SHARE   11 Comments

Similar topics

Database

Broader topics

Oracle   Software

Narrower topics

MongoDB   MySQL   NoSQL   PostgreSQL   SQL Server

Corrections   Send us news

Open source databases: What are they and why do they matter?

FEATURE We trawled through the licensing terms and spoke to the vendors so you don't have to

DATABASES   8 days | 47 🗩

Want to know the future of FOSS? You can look it up in a database

OPINION If you make it here, you can make it anywhere

DATABASES   1 day | 8 🗩

Microsoft offers SQL Server 2022 release candidate to Linux world

Redmond loves open source so much, it keeps most of its stuff closed

DATABASES   5 days | 16 🗩

Managed document-based data stores come of age

Why enterprises are choosing Amazon DocumentDB to power their mission critical applications

SPONSORED FEATURE

Ant Group's in-house DB set for global release, including Raspberry Pi edition

OceanBase 4.0 is derived from tech behind Alipay and is MySQL
DATABASES compatible

28 days | 7 💬

Couldn't connect to West Europe SQL Databases last week? Blame operator error

Clouds form over PICNIC (Problem In Chair Not In Computer)

PAAS + IAAS   2 months | 5 💬

Cyber-mercenaries for hire represent shifting criminal business model

Emerging threat group offers a broad range of attack services

SECURITY   2 months | 9 💬

Oracle, Microsoft agree to shared custody of your workloads in the cloud

Larry gets your database, while Satya will take your apps

PAAS + IAAS   2 months | 15 💬

Oracle 12.1 users warned as mainstream support ends in July

12.1 friends without benefits as database pros question LTR with Oracle 12.2

DATABASES   2 months | 3 💬

Digital burglary at recruitment agency Morgan Hunt confirmed

Third-party software developer blamed for 'improperly storing credentials to our database'

SECURITY   2 months | 19 💬

Rows, columns, and the search for a database that can do everything

ANALYSIS Snowflake last week promised analytics and transactions in the same system. For some it was déjà vu all over again

DATABASES   3 months | 6 💬

Oracle, IBM losing ground to local databases in China, says IDC

One market down on China's march toward tech self-sufficiency, a bajillion more to go.

DATABASES   2 months | 5 💬