MotherDuck

How to Bootstrap a Data Warehouse with DuckDB

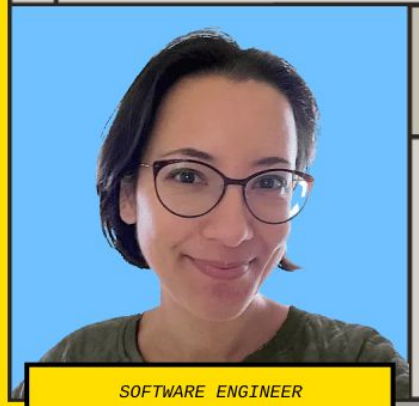
Guen Prawiroatmodjo, Alex Monahan, Elena Felder,
Nicholas Ursa, Mehdi Ouazza, Till Döhmen

SciPy conference 2024

▶ https://bit.ly/cookiecutter-data

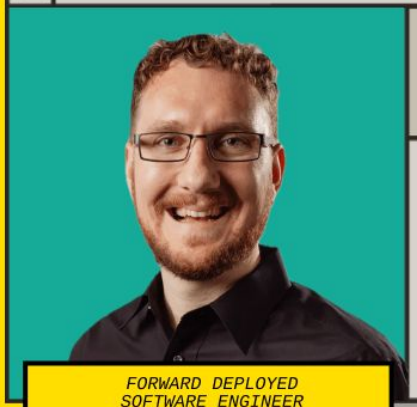
ELENA FELDER

SOFTWARE ENGINEER
MOTHERDUCK



GUEN PRAWIROATMODJO

SOFTWARE ENGINEER
MOTHERDUCK



ALEX MONAHAN

FORWARD DEPLOYED SOFTWARE ENGINEER
MOTHERDUCK

▶ https://bit.ly/quacky-sql

SciPy 2024 TACOMA, WA

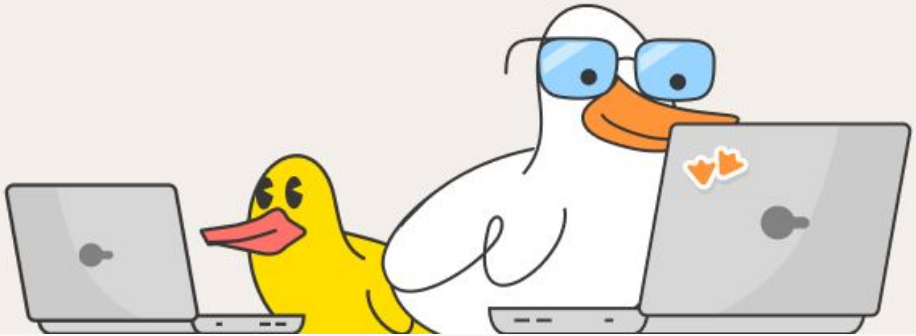
Tuesday, July 9th All the SQL a Pythonista needs to know

Friday, July 12th How to Bootstrap a Data Warehouse with DuckDB

DUCK TUB

DUCK-SHAPED TUB WITH DUCKIES
MOTHERDUCK

A Data Warehouse is a place used to collect, store and collaborate on your data



BIG DATA IS DEAD

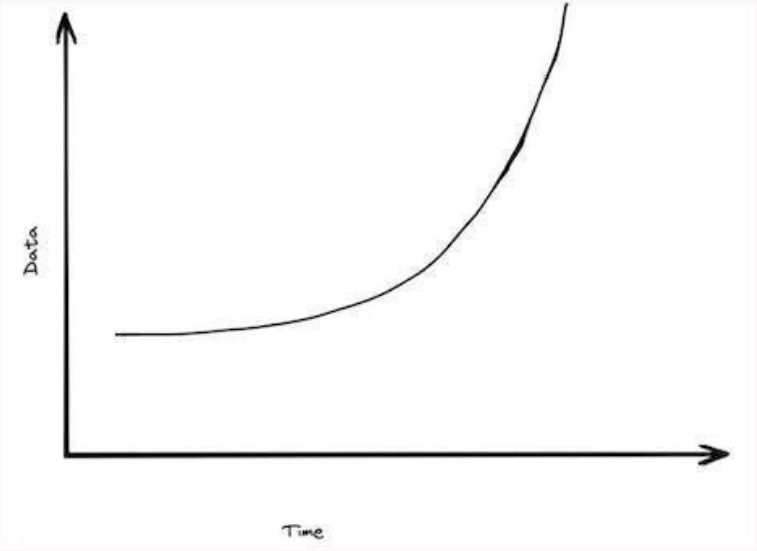
2023/02/07

BY JORDAN TIGANI

https://bit.ly/big-data-is-dead

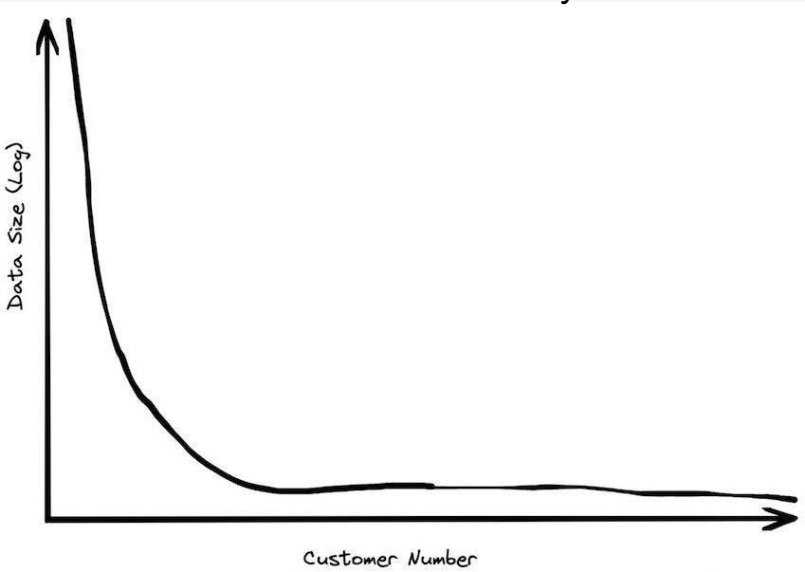
Expectation (~ 10 yrs ago)

By year {CURRENT_YEAR() + ε} there will be {unfathomably large amount} of data generated



Reality

“The vast majority of enterprises have data warehouses smaller than a terabyte.”

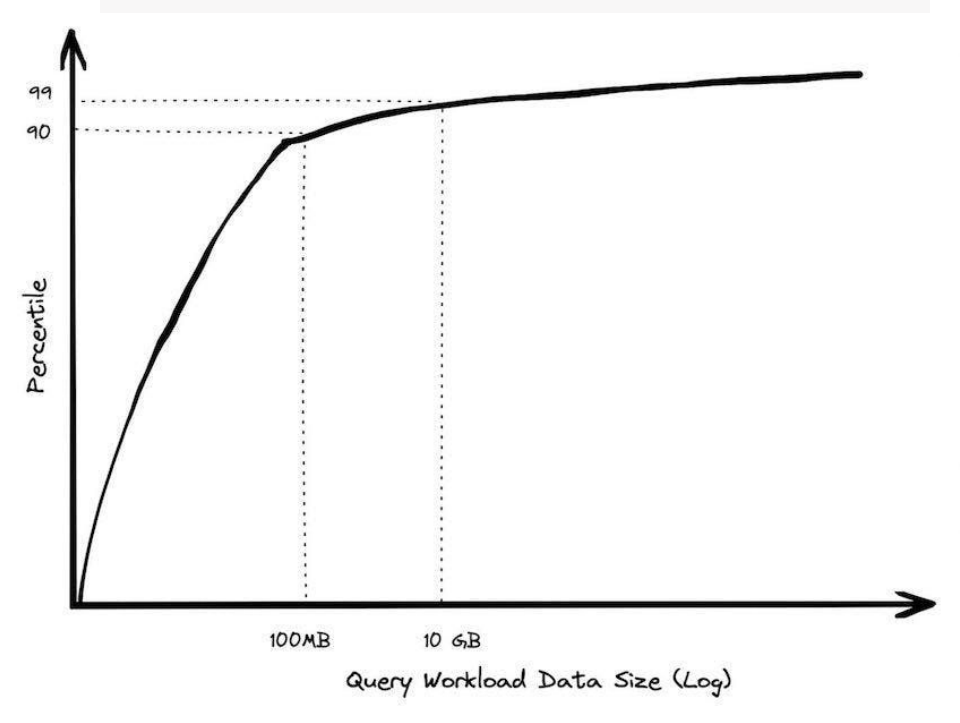


BIG DATA IS DEAD

2023/02/07

BY JORDAN TIGANI

Customers with giant data sizes almost never query huge amounts of data

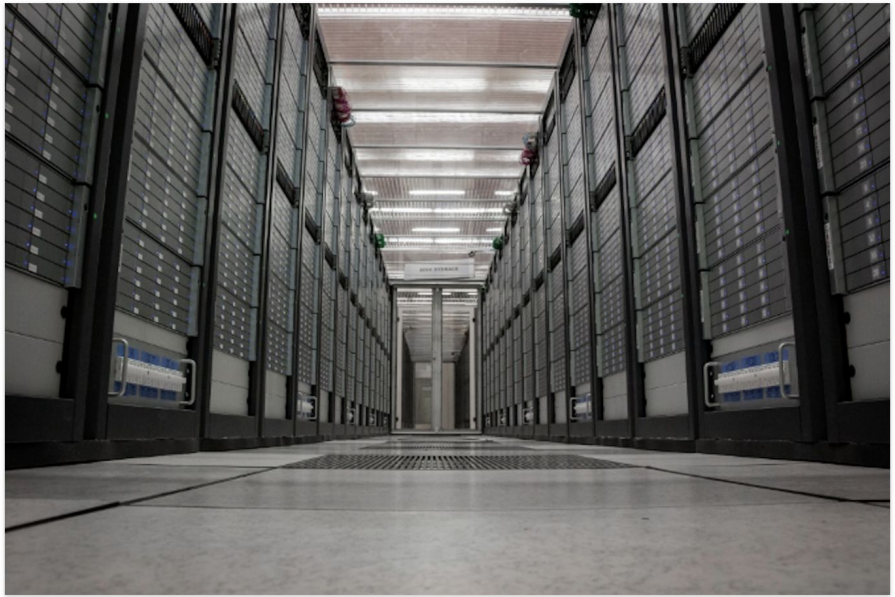


Big data in science?

An exabyte of disk storage at CERN

CERN disk storage capacity passes the threshold of one million terabytes of disk space

29 SEPTEMBER, 2023 | By Tim Smith



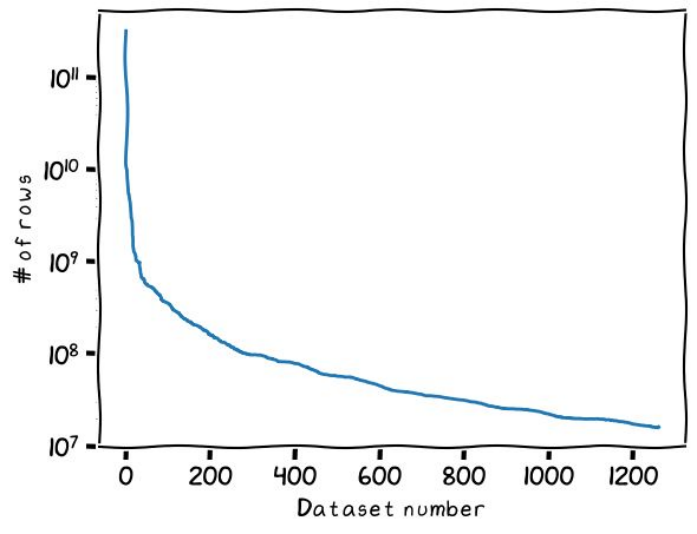
A fraction of the 111 000 devices that form CERN's data storage capacity. (Image: CERN)

CornellLab eBird

World All species, locations, and dates	**39+ GB** .tar
Sampling event data Effort data only	**5.5 GB** .tar

huggingface.co/datasets

>99% of datasets are <5GB



Why use a database for small to medium data workloads?

- Fly faster than Pandas
- Handle larger than memory data
- Describe the output you want, not how to get there (SQL is declarative)
- Package up all your tables into 1 DuckDB file for sharing
- Manage your data pipelines with modern data engineering tools

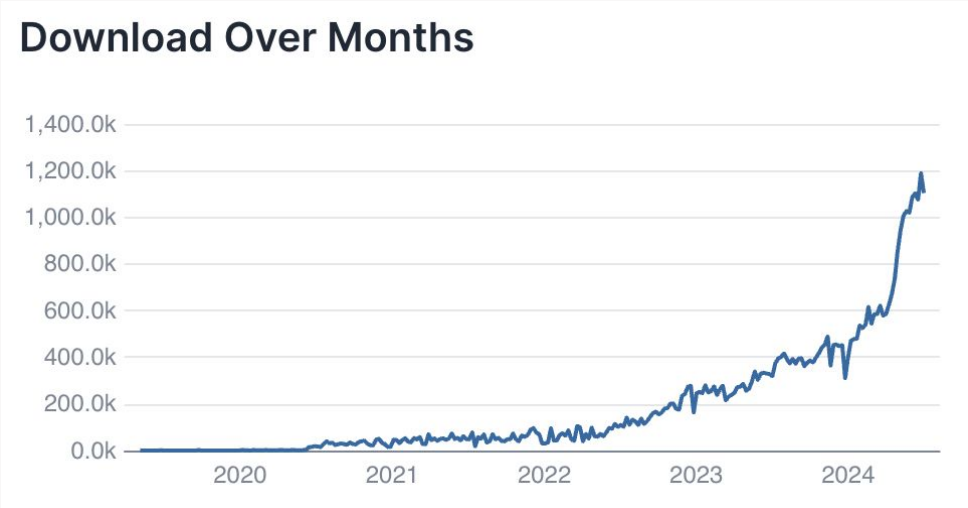
What is DuckDB?

A lightweight, in-process SQL Analytics Engine that is taking the data world by storm.

DuckDB Labs

DuckDB

- An analytical SQL database
- MIT licensed
- Clients in 15+ languages
- 1M+ downloads / week



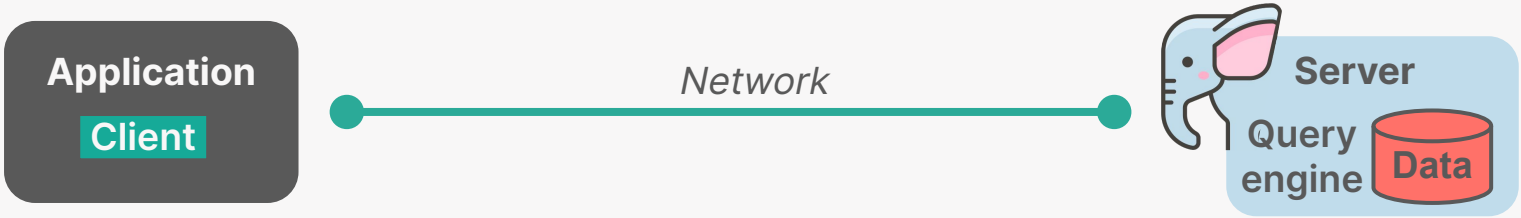
.duckdb CSV JSON Parquet Iceberg

APACHE ARROW pandas SQLite NumPy Polars

https://duckdbstats.com/

Typical Client-Server Architecture

eg: Postgres

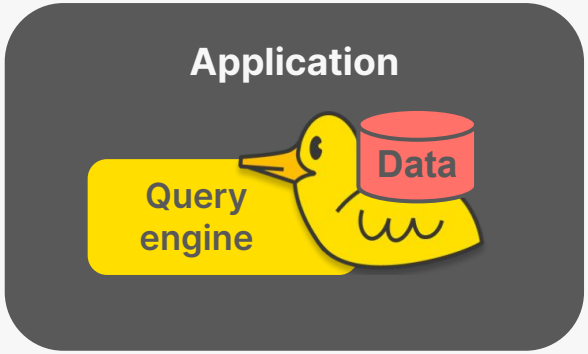


- DB hosts data and query engine
- Application logic split between app and server
- Network is slow and untrusted

```
psql -h <HOST> -p <PORT> -U <USER> <DB_NAME>
```

In-Process Database Architecture

eg: SQLite



- In-process data, engine and logic
- No trust boundaries to traverse
- High bandwidth

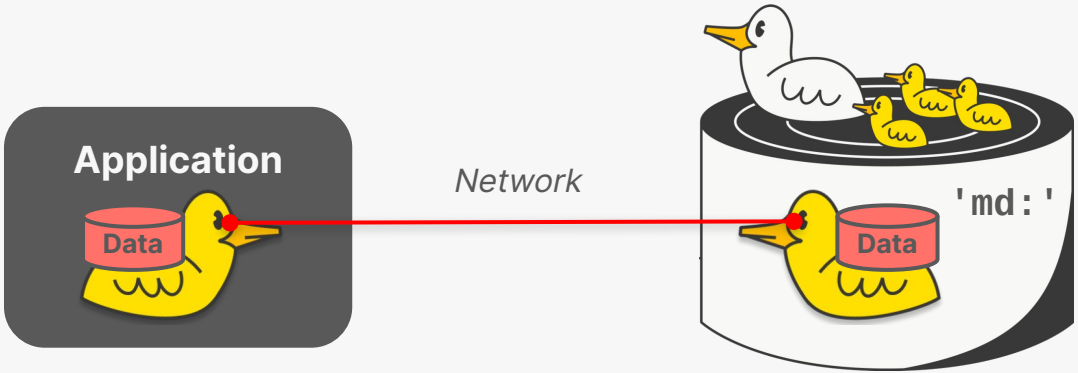
```
duckdb 'my_db.duckdb'
```

What is MotherDuck?

A serverless DuckDB platform for low-cost, low-friction analytics that will scale to support complex apps and data

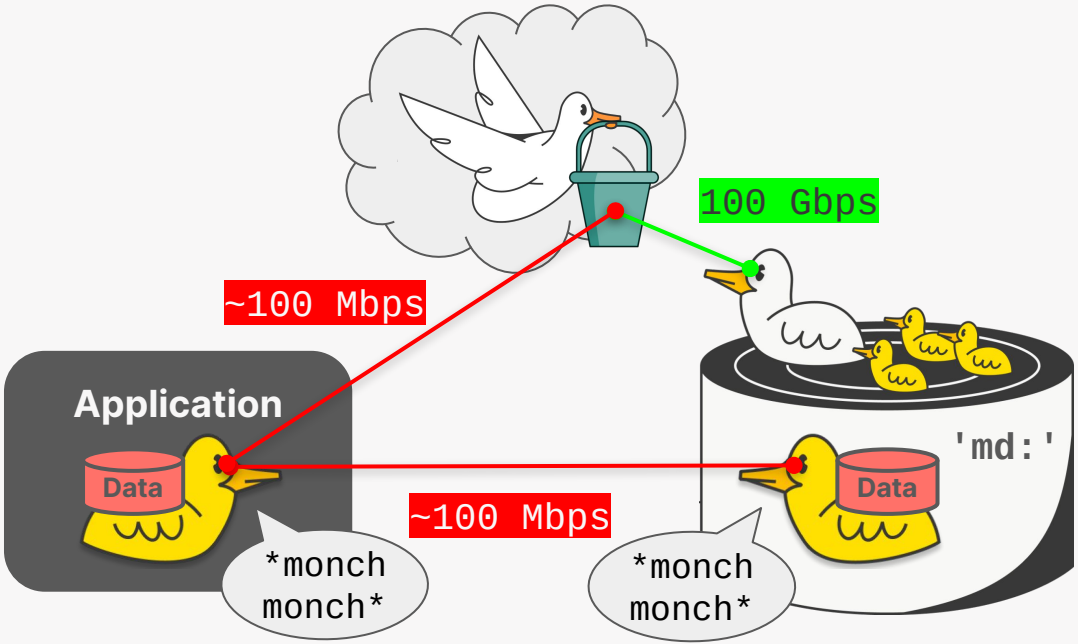
Dual Execution Architecture

MotherDuck



- Data and query engine both local & remote
- Better utilization of cache
- Serverless

```
duckdb 'md:my_remote_db'
```



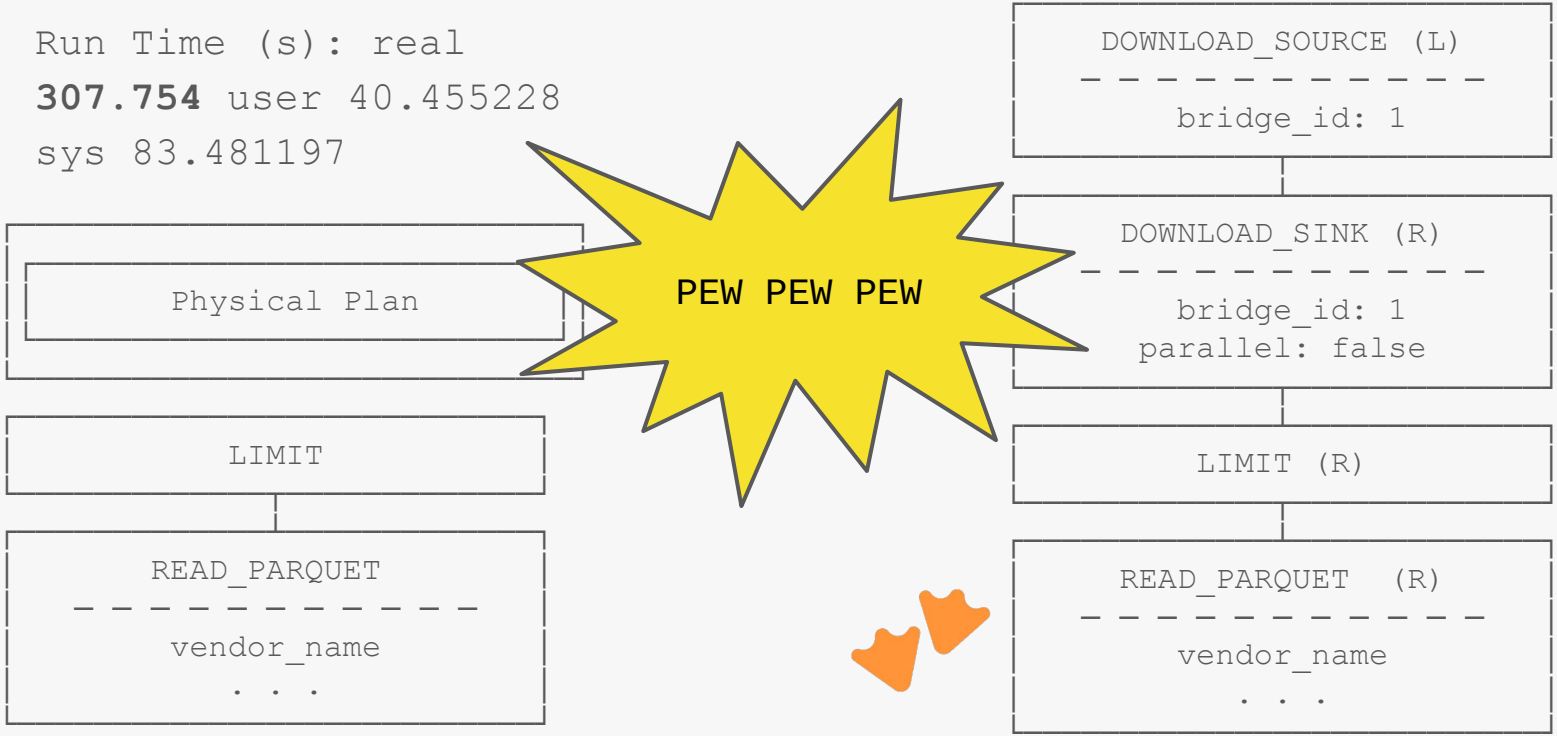
- Fast ingestion of cloud-based data
- Dual query engine abstracts away local-remote query planning

```
SELECT * FROM read_parquet('s3://xyz/*.parquet') LIMIT 100;
```

```
EXPLAIN SELECT * FROM
read_parquet('s3://xyz/*.parquet')
LIMIT 100;
```

Run Time (s): real
43.024 user 17.930957
sys 0.058618

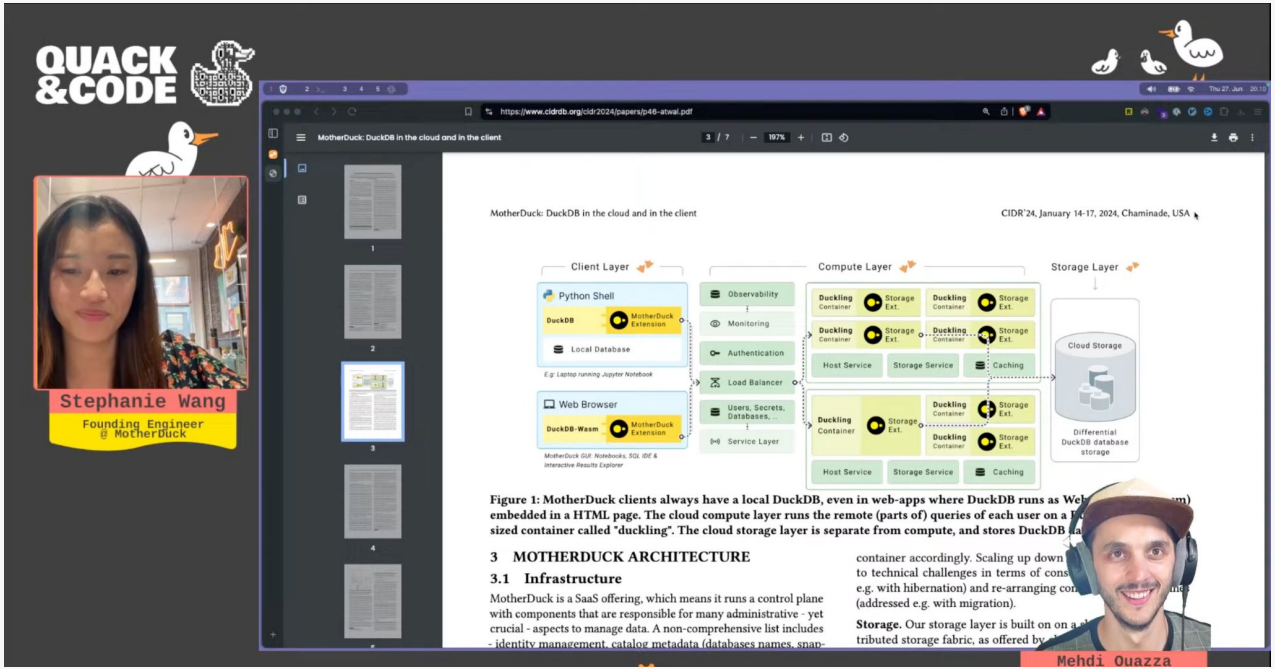
Run Time (s): real
307.754 user 40.455228
sys 83.481197



MotherDuck: DuckDB in the cloud and in the client

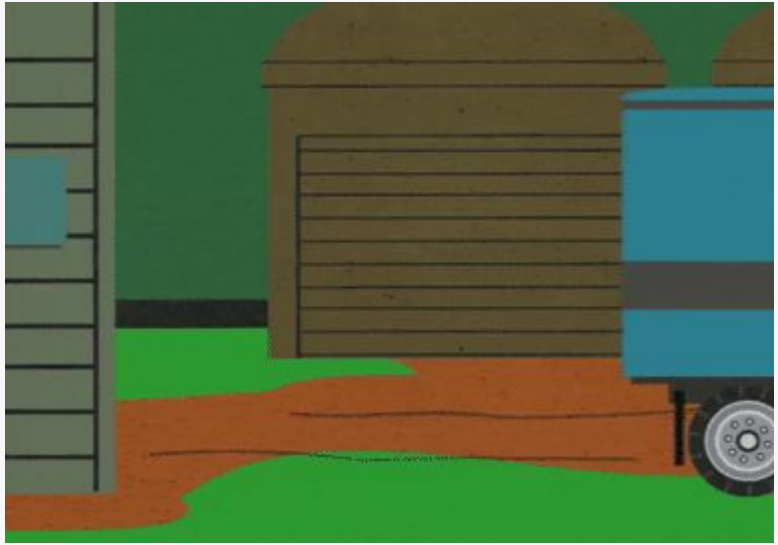
RJ Atwal, Peter Boncz, Ryan Boyd, Antony Courtney, Till Döhmen, Florian Gerlinghoff, Jeff Huang, Joseph Hwang, Raphael Hyde, Elena Felder, Jacob Lacouture, Yves LeMaout, Boaz Leskes, Yao Liu, Alex Monahan, Dan Perkins, Tino Tereshko, Jordan Tigani, Nick Ursa, Stephanie Wang, Yannick Welsch
firstname@motherduck.com

https://bit.ly/motherduck-paper

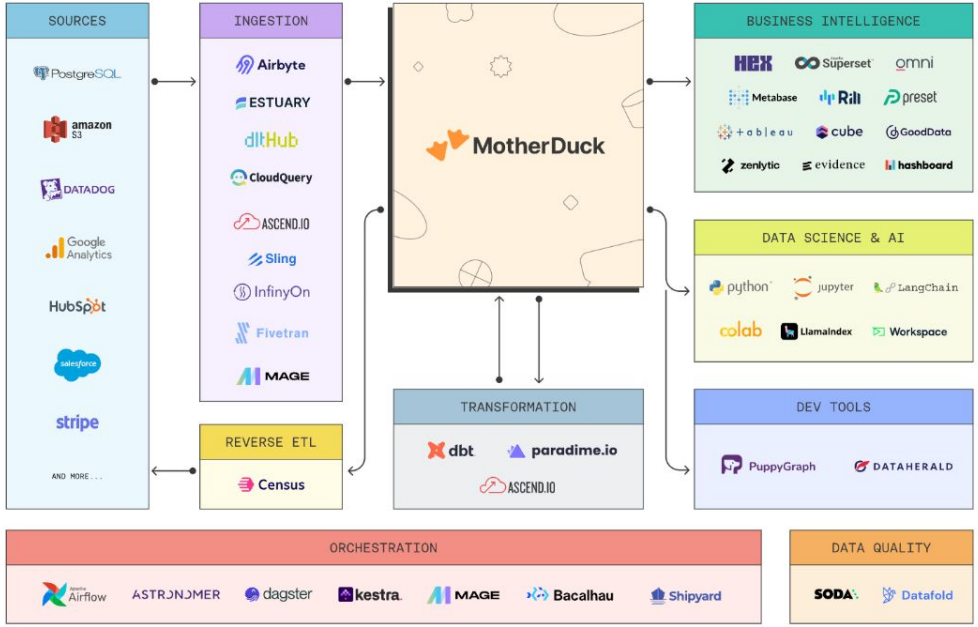


Quack & Code episode on Dual Execution: https://bit.ly/quack-and-code-dual-execution

Extract, Load & Transform (ELT)

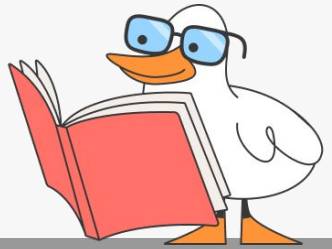


ECOSYSTEM Modern Duck Stack



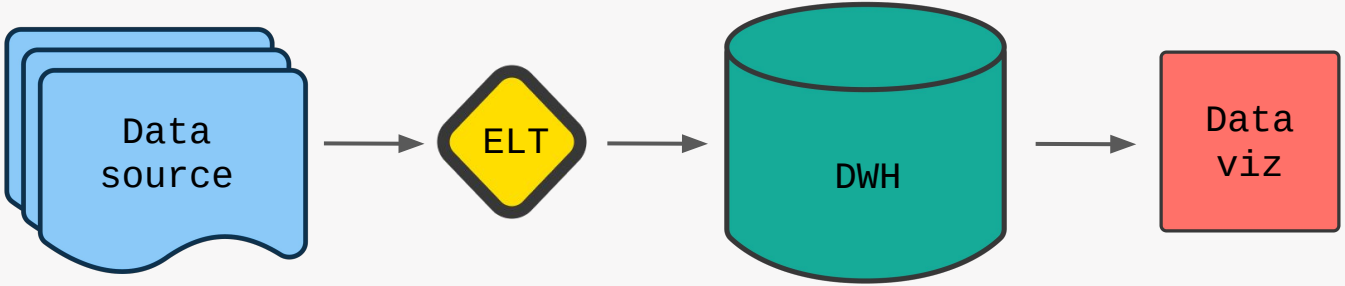
Create your own Data Warehouse with DuckDB

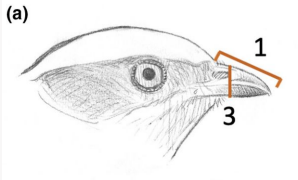
1. Build
2. Share
3. Collaborate



```
> pipx install cookiecutter
> pipx run cookiecutter gh:guenp/cookiecutter-data-warehouse
```

What tools are we using?





AVONET[1]

jupyter

DuckDB

MotherDuck

plotly

jupyter{book}

GitHub

[1] AVONET: morphological, ecological and geographical data for all birds. Ecology Letters, 25(3):581–597, 2 2022. URL: https://doi.org/10.1111/ele.13898, doi:10.1111/ele.13898.

Demo!

bit.ly/cookiecutter-data

THINK SMALL

SAVE THE DATE

SOMETHING SMALL IS COMING SOON

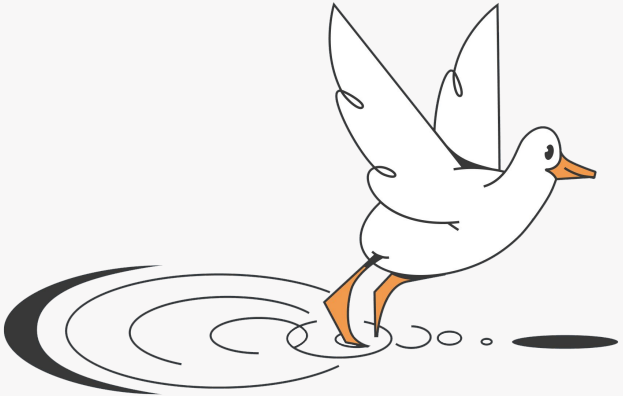
JULY 15 2024

Thank you

- https://bit.ly/cookiecutter-data
- https://slack.motherduck.com/

MotherDuck

Questions?



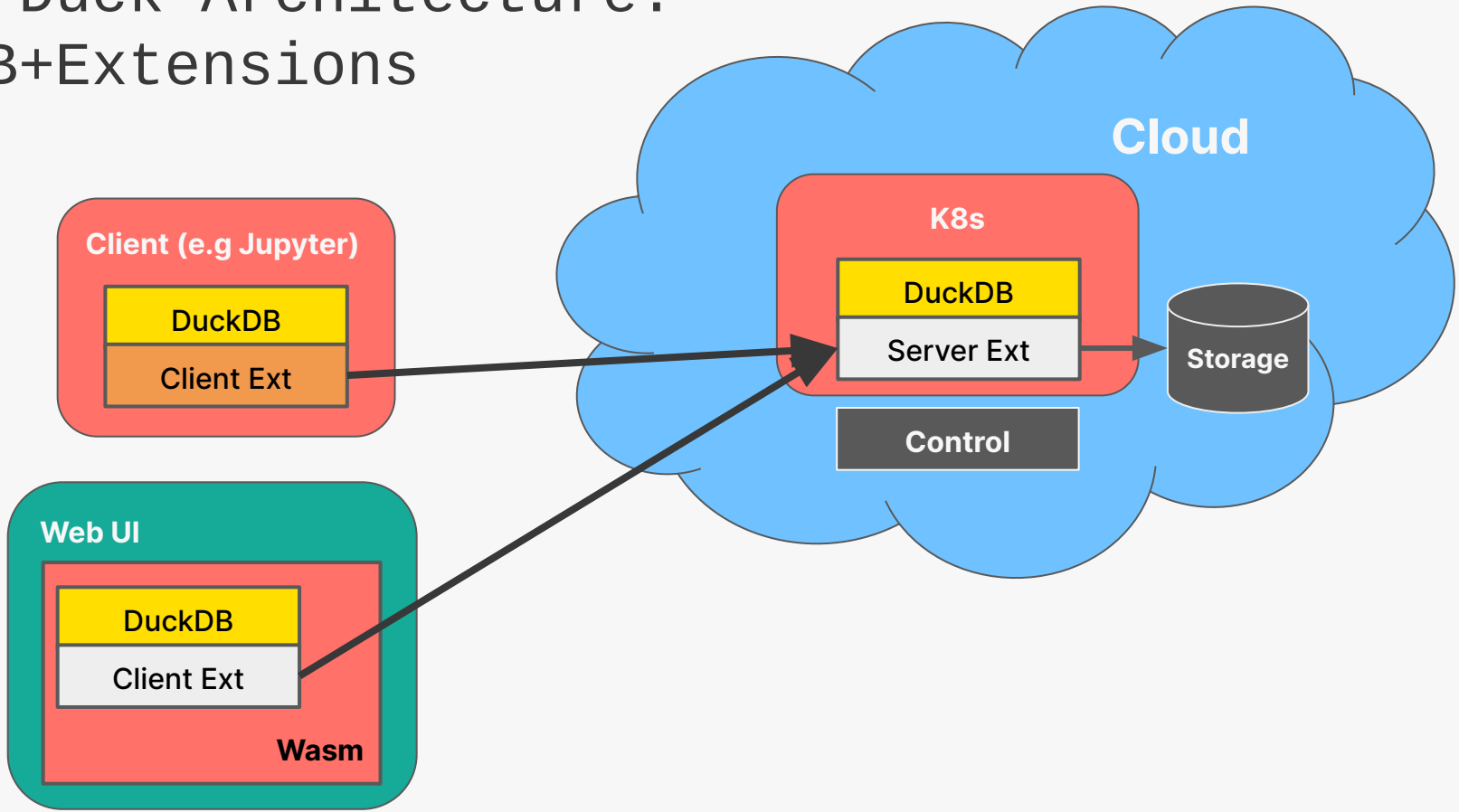
DuckDB Performance Tips

- for speed and correct types, prefer binary formats like `.parquet` to csv
- duckdb can work with larger than memory queries with spill to disk if you open a file (`duckdb my_db.db` or `SET temp_directory=temp.tmp`)
- Use `EXPLAIN` to see the query plan to optimize your query
  - It's efficient to apply filters early and apply sorts late.
- Use `.timer on` to time your queries in the CLI
- Split large queries into steps using CTEs (`WITH … AS` or `TEMP TABLE`) to make debugging each stage easier
- Blocking operators (`JOIN`, `ORDER BY`, `GROUP BY`) are usually where time and memory are consumed
- `GROUP BY ALL` is a convenience to avoid having to specify group by keys. It will group by any non-aggregated columns (DuckDB-specific!)

▶ https://bit.ly/quacky-sql

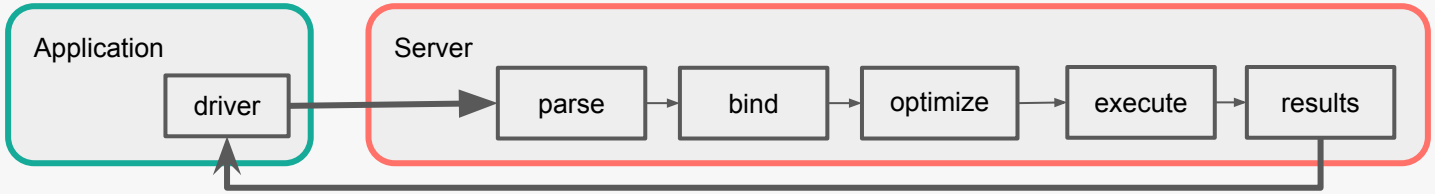
MotherDuck Architecture: DuckDB+Extensions

- Client APIs
  - Overview
  - Python
  - R
  - Java
  - Julia
  - C
  - C++
  - Node.js
  - WASM
  - ODBC
  - CLI



Hybrid cloud database

Typical Database Execution



MotherDuck Execution