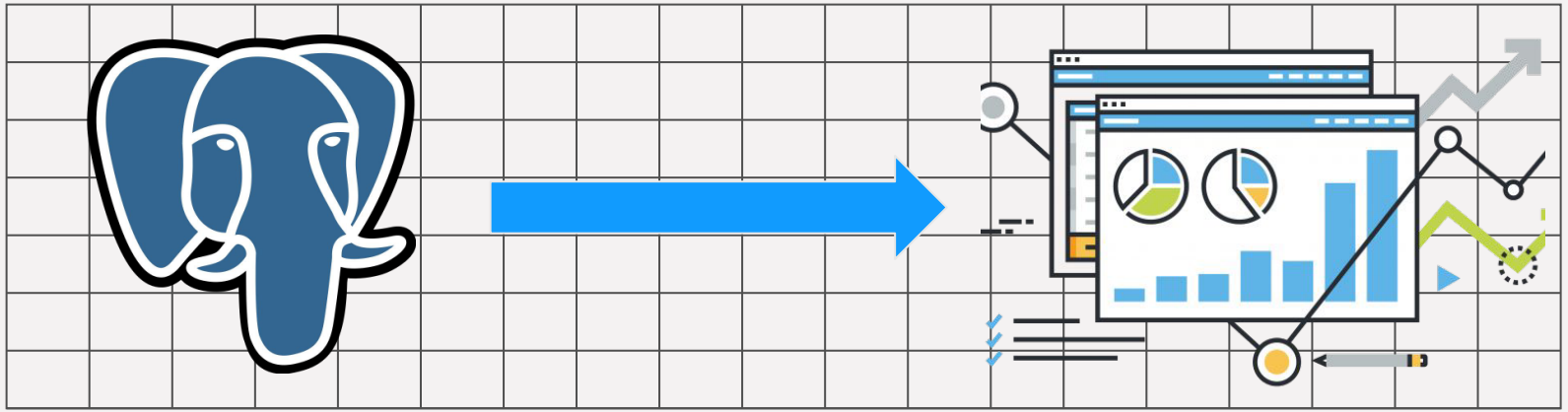
PostgreSQL & Analytics

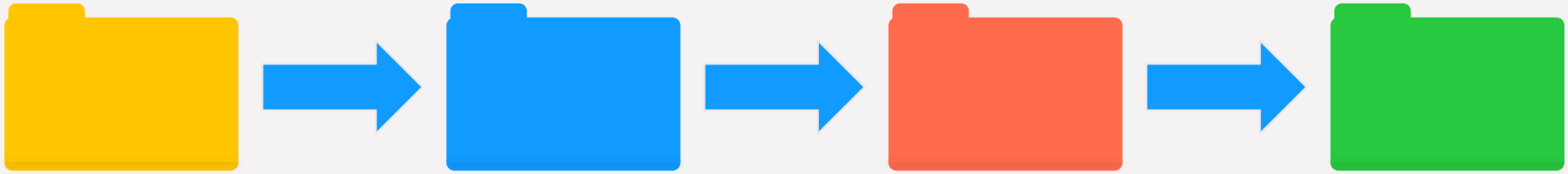
Exploring how extensions are making it possible



About Me

Akshat Jaimini

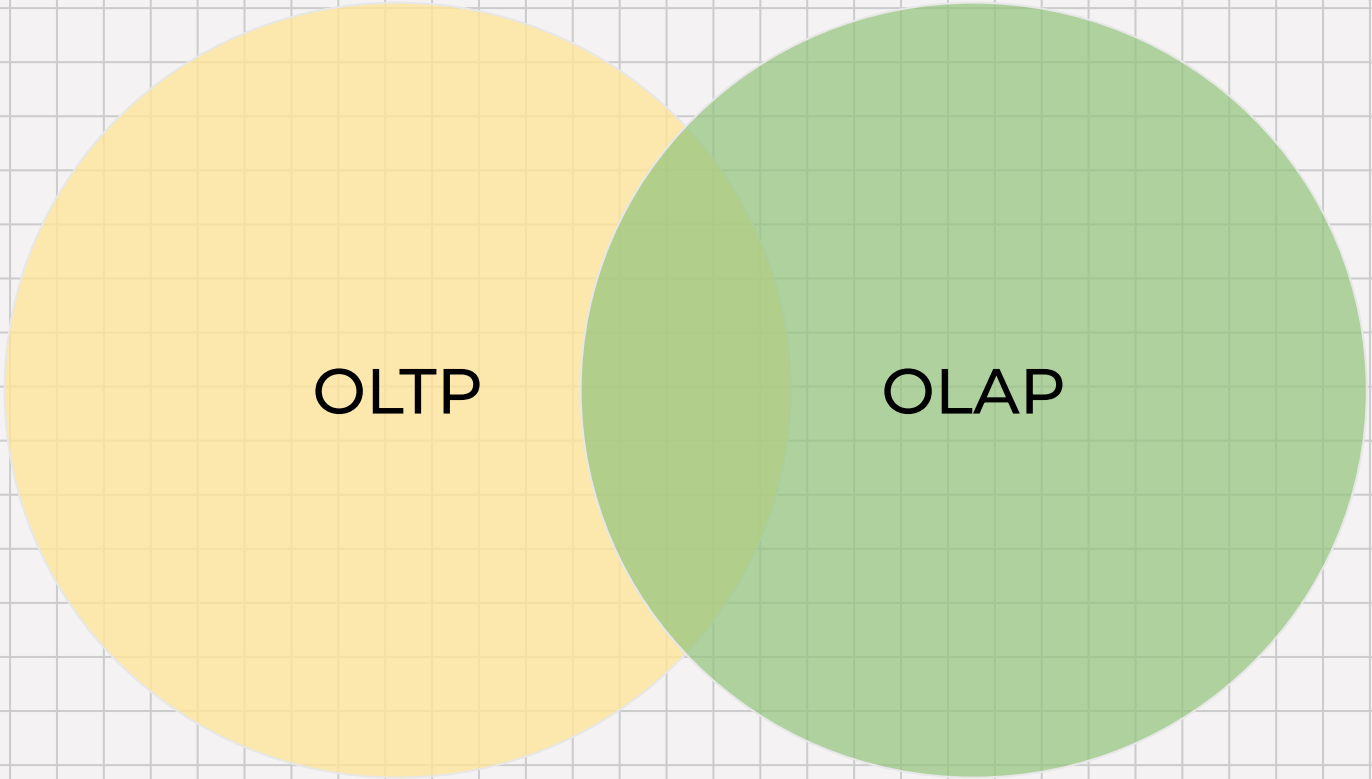
- Final Year Undergrad, TIET Patiala
- SDE Intern at ION Group, India
- PostgreSQL Community member since 2023
- Joined via GSoC 2023
- Maintainer pgweb-testing-harness
- Database Observability Tools
- Budding Postgres Hacker


Why OLAP and PostgreSQL?
Solutions with PostgreSQL
How extensions make it possible?
Demo Time!

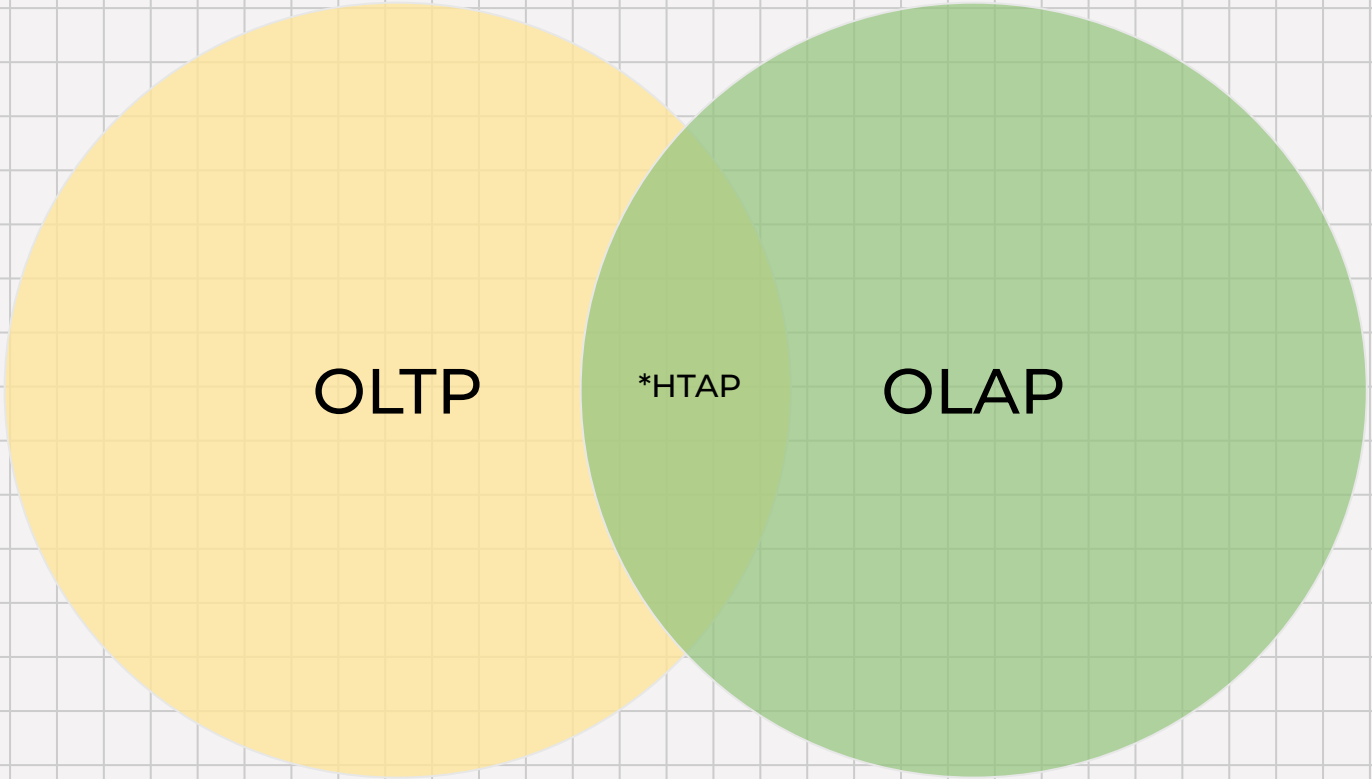
Why OLAP?

And PostgreSQL?

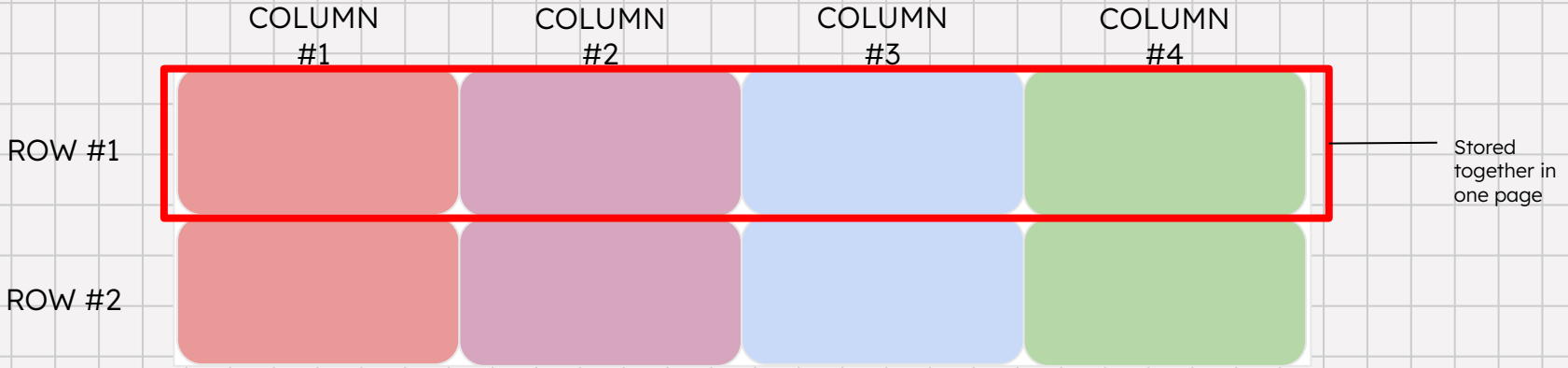
DIVISION ON TYPE OF WORKLOAD



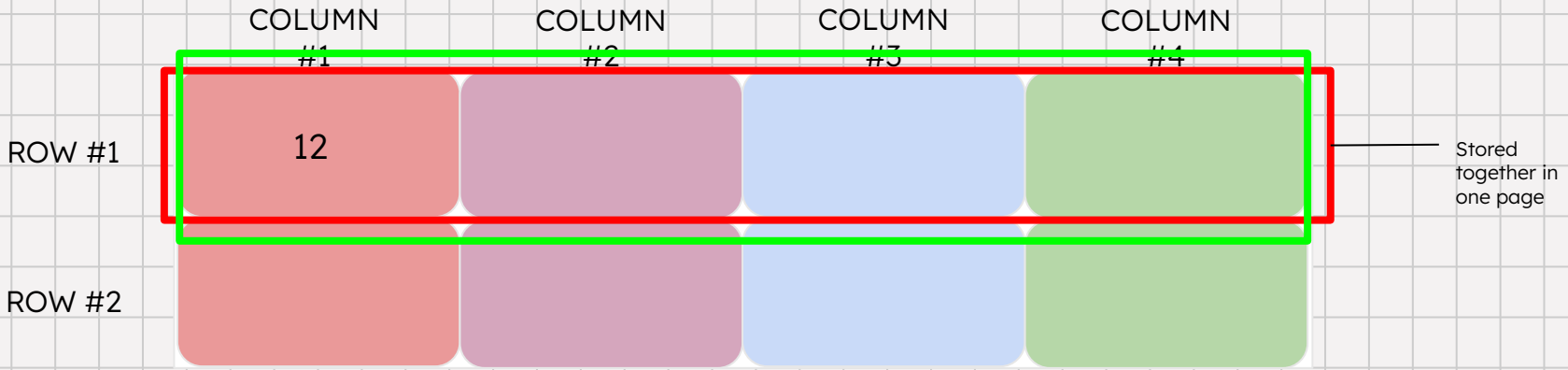
DIVISION ON TYPE OF WORKLOAD



OLTP

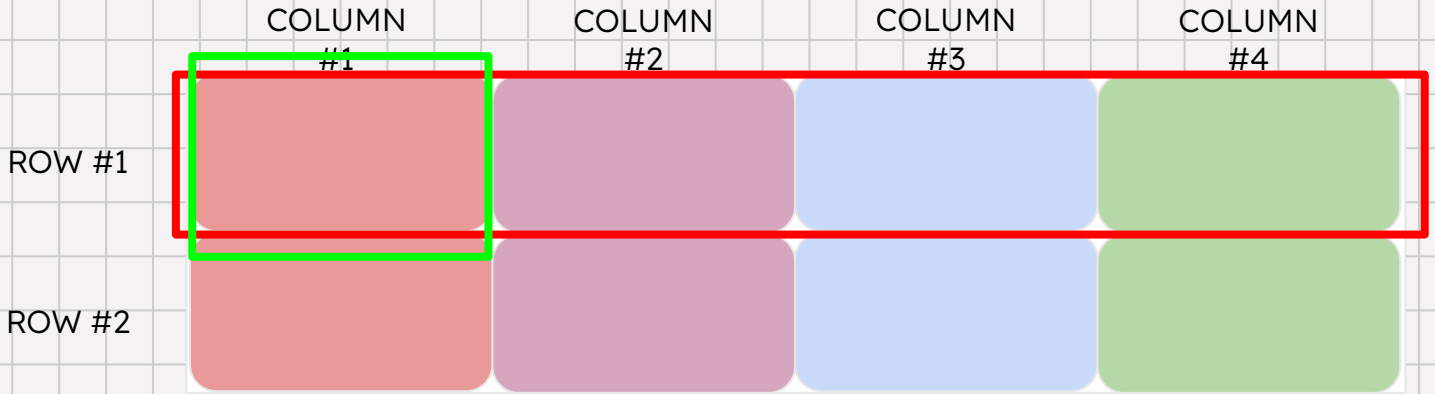

COLUMN #1
COLUMN #2
COLUMN #3
COLUMN #4
ROW #1
ROW #2
Stored together in one page

OLTP



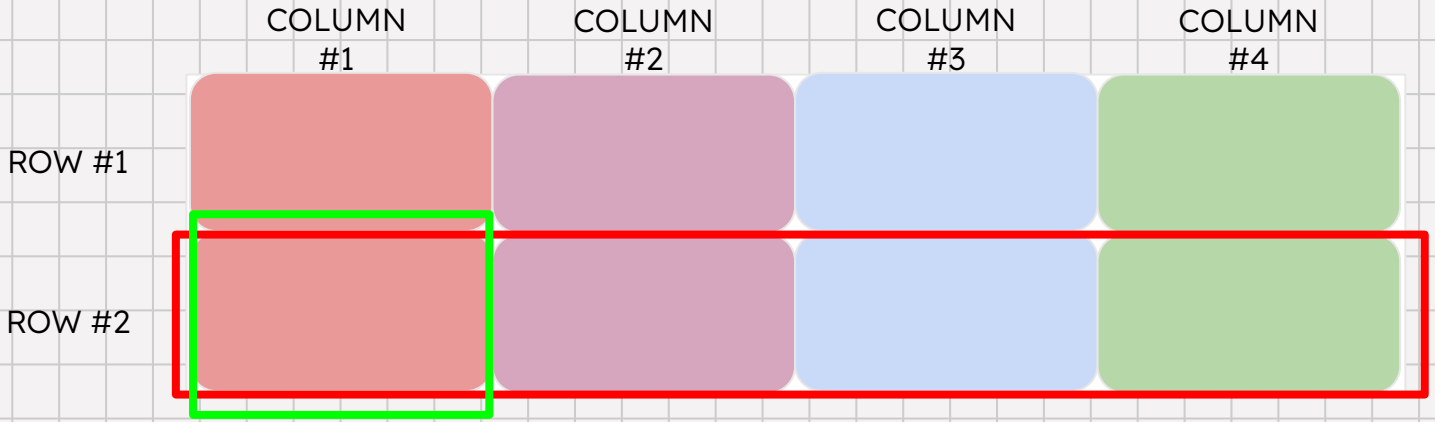

```
SELECT * FROM OLTP_DB WHERE COLUMN1=12;
```

OLTP



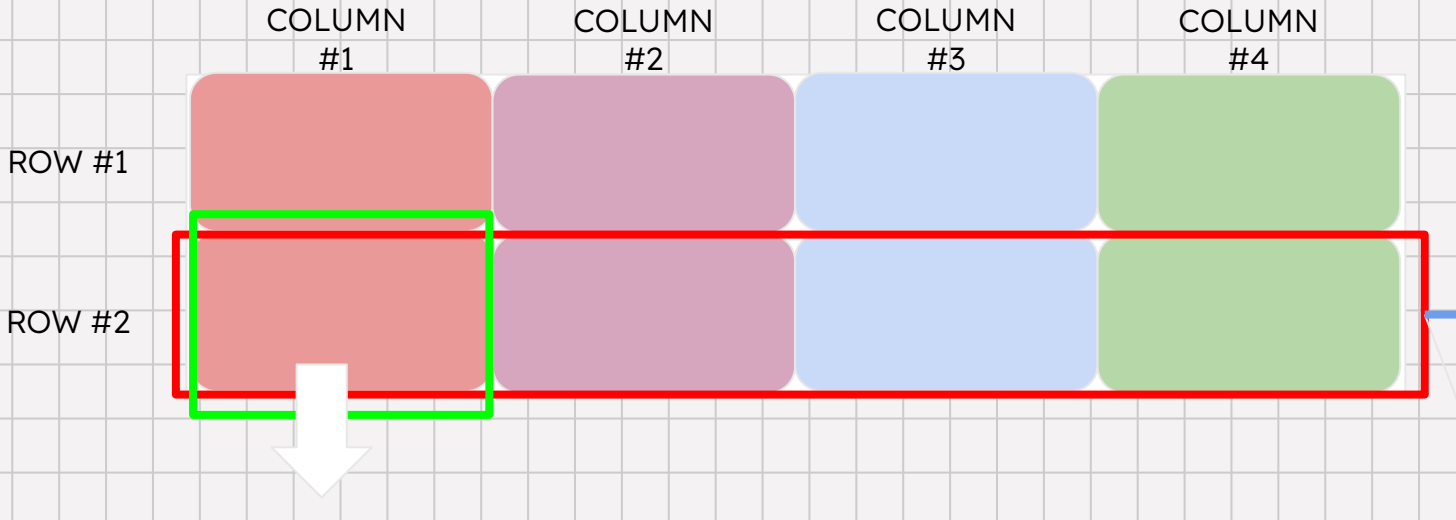

```
SELECT SUM(COLUMN#1) FROM OLTP_DB;
```

OLTP




```
SELECT SUM(COLUMN#1) FROM OLTP_DB;
```

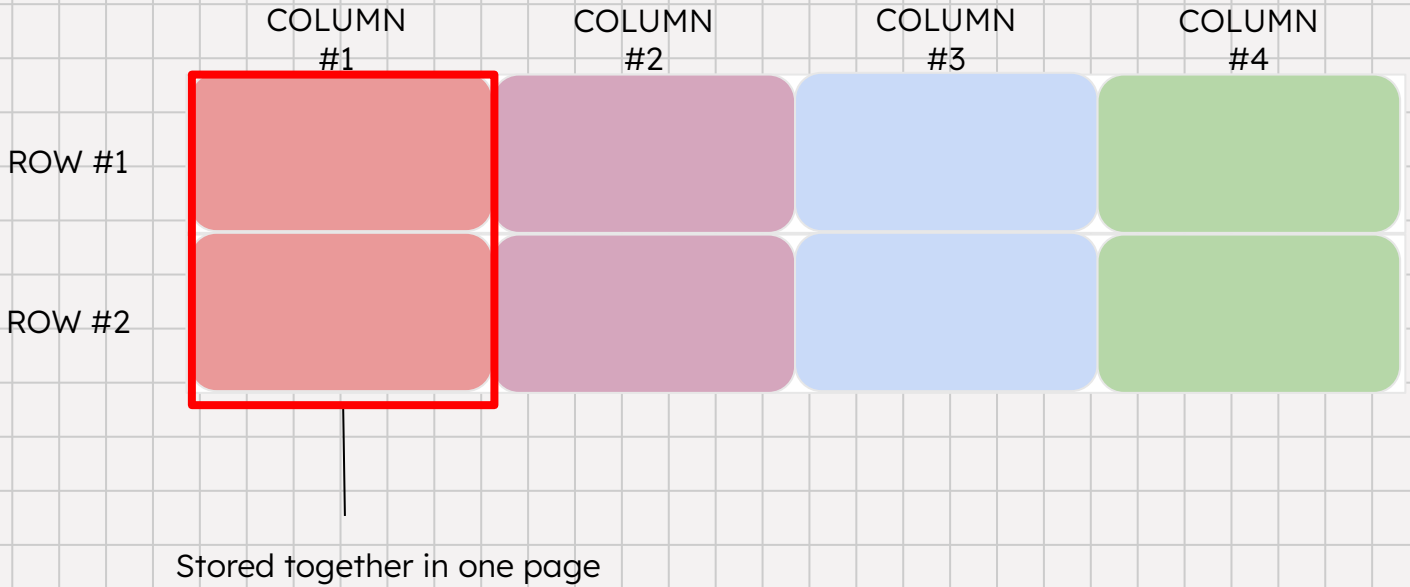
OLTP



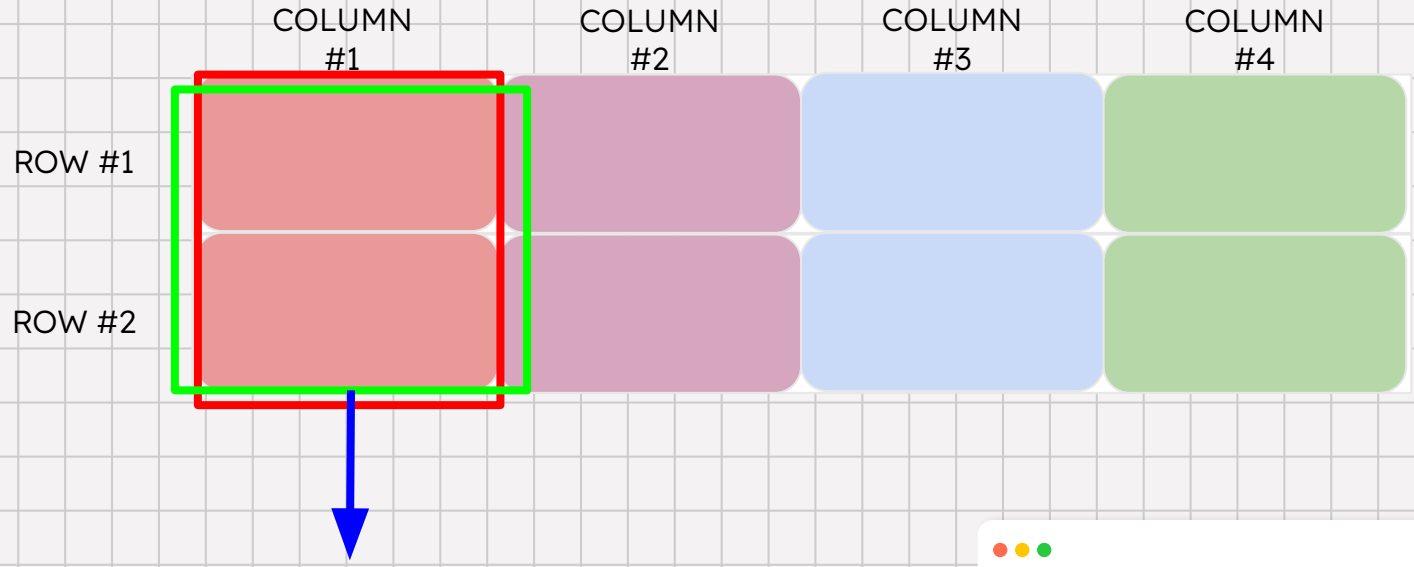

```
SELECT SUM(COLUMN#1) FROM OLTP_DB;
```

MIGHT BE IN A DIFFERENT PAGE!

OLAP - Column Stores



OLAP - Column Stores



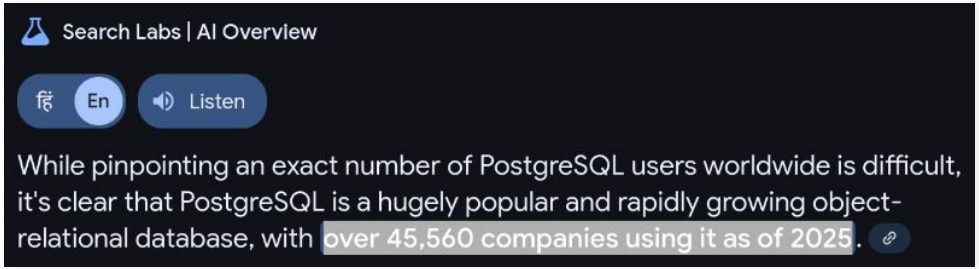

```
SELECT SUM(COLUMN#1) FROM OLTP_DB;
```

Why PostgreSQL

In the OLAP Space?

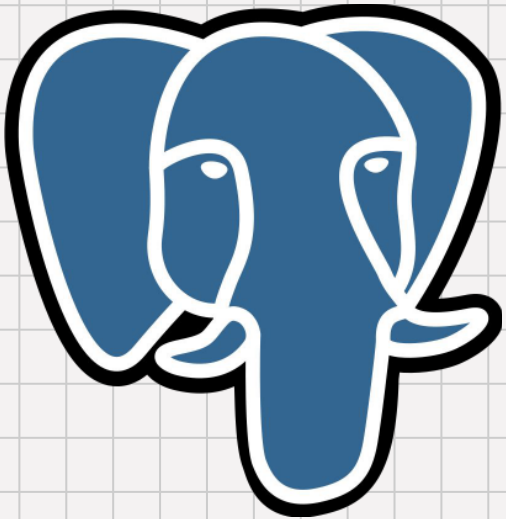
Why?

- PostgreSQL is one of the most used OSS databases



- Indie Hackers to Large MnC's rely on PostgreSQL
- Replicating Data from one Data System to another is a huge commitment
- PostgreSQL behaviour and interaction standards are easy to pick up on and have a large user base

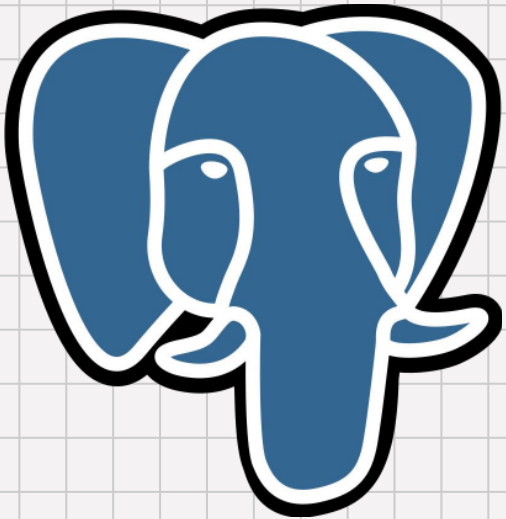
History of Pg - OLAP



Hacks to make OLAP work on PostgreSQL

- Tweaking shared_buffers and cache params - (Example)
- Tweaking WAL Configs and playing with replication strategies: (Tuning WAL to make Pg OLAP performant)

History of Pg - OLAP



Direct Commits to PostgreSQL

- COAST :

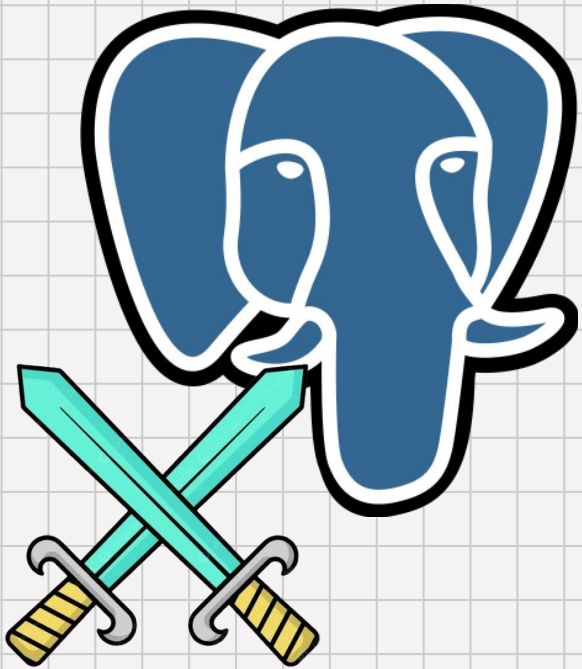
```
ALTER TABLE foo
ALTER COLUMN foocol
SET ORIENTATION COLUMN;
```

We might also want to store columns in groups together, like so

```
ALTER TABLE foo
ALTER COLUMN foocol1 SET COLUMN GROUP 2,
ALTER COLUMN foocol2 SET COLUMN GROUP 2;
```

- Vertical Columnar Index
- Using JSONB

Extensible PostgreSQL



- PostgreSQL extensions API
- Allows you to directly interact with the internal code
- Original API is in C
- PGRX: to write extensions in Rust

Where Extensions

position PostgreSQL?

- As an Analytics Generator System

Where Extensions position PostgreSQL?

- As an Analytics Generator System

Timescale

ParadeDB

Extensions

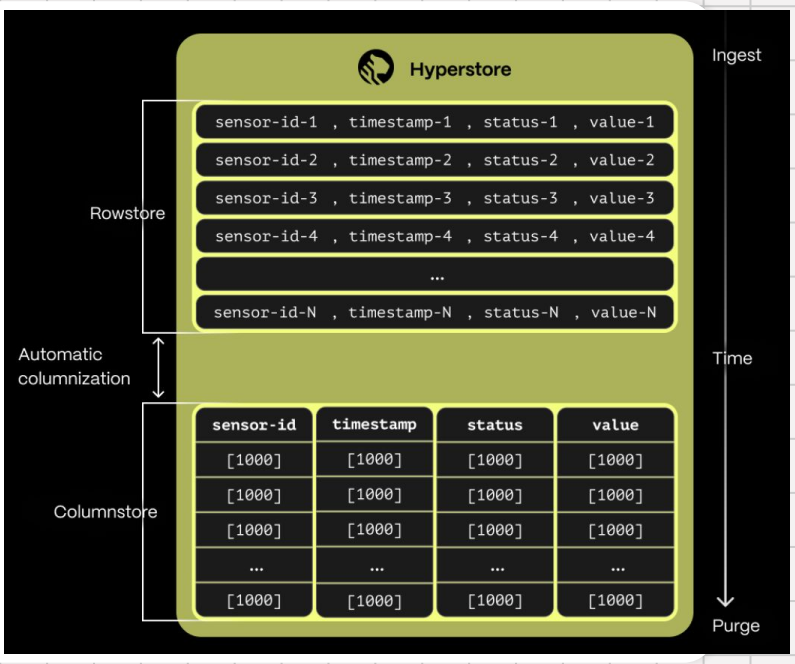
Timescale

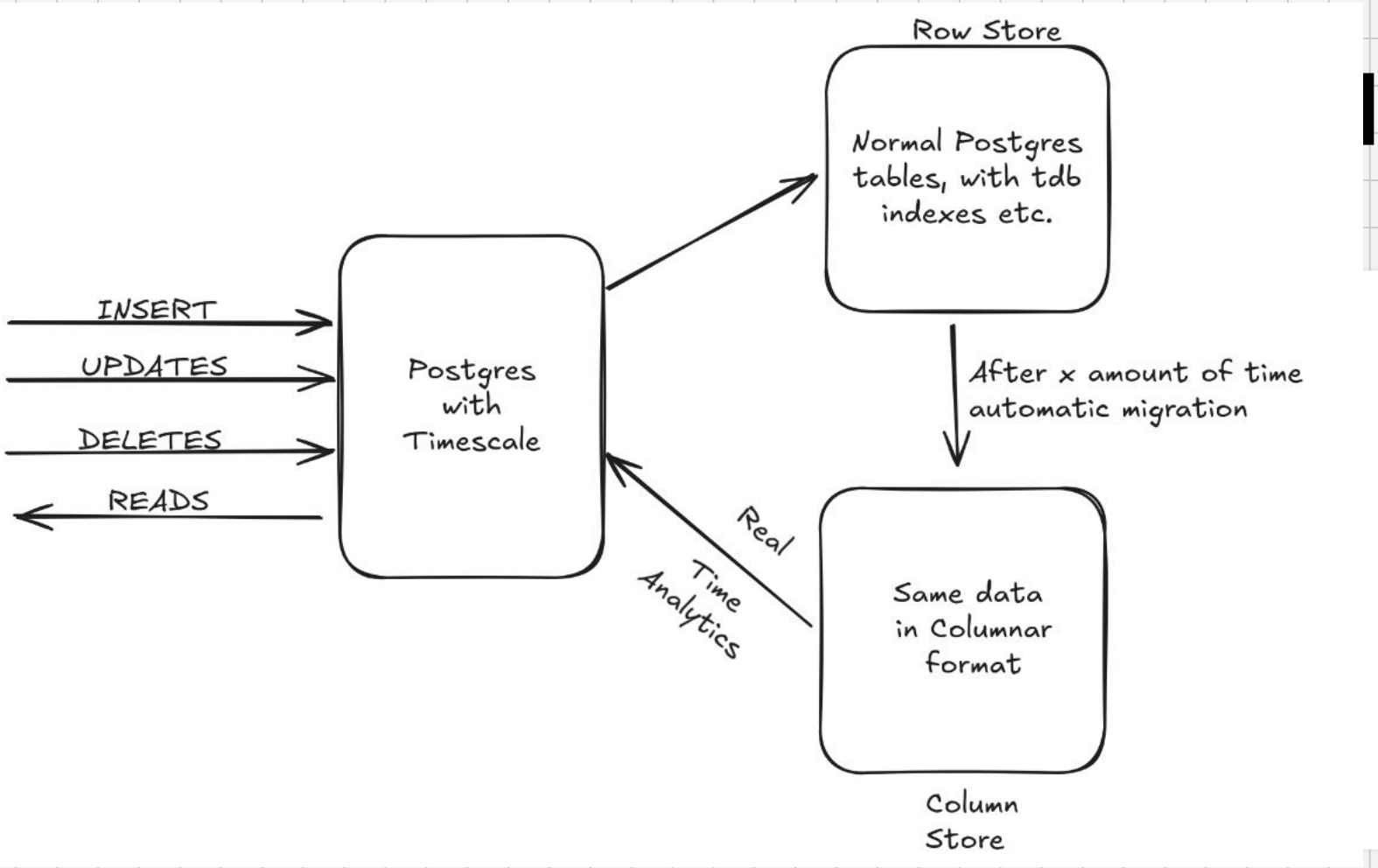
- Designed for handling time series
- Support for Real-Time Analytics
- Offer both row and column store options via *HyperCore*

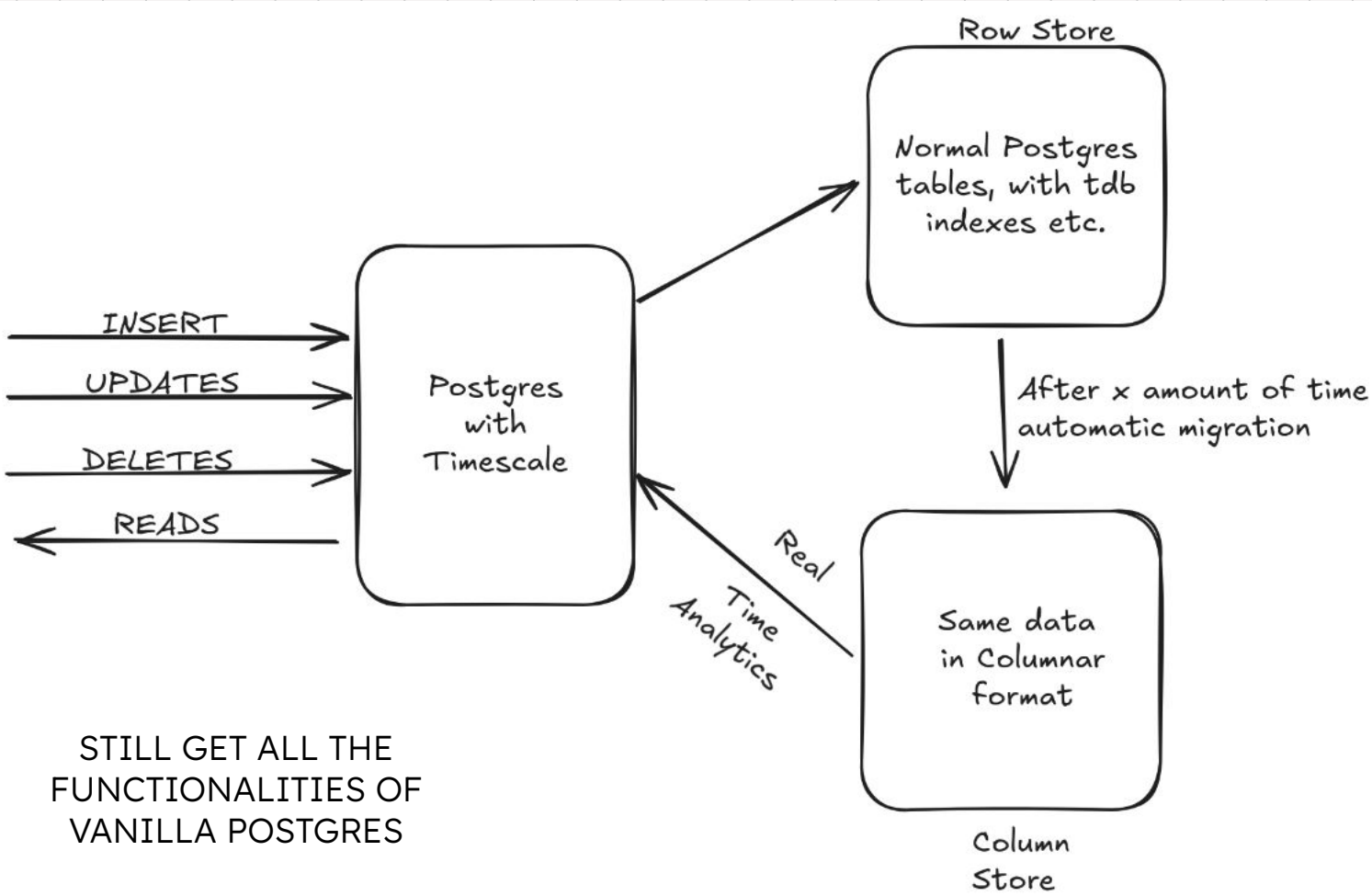
Extensions

Timescale

- Designed for handling time series
- Support for Real-Time Analytics
- Offer both row and column store options via *HyperCore*




Row Store
Normal Postgres tables, with tdb indexes etc.
INSERT
UPDATES
DELETES
READS
Postgres with Timescale
After x amount of time automatic migration
Real Time Analytics
Same data in Columnar format
Column Store



STILL GET ALL THE FUNCTIONALITIES OF VANILLA POSTGRES

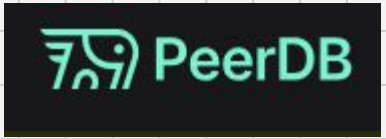
Extensions

Timescale

- Designed for handling time series
- Support for Real-Time Analytics
- Offer both row and column store options via *HyperCore*
- Enhanced Indexes for time series data
- *Hypertable* for enhanced partitioning and query optimization

Where Extensions position PostgreSQL?

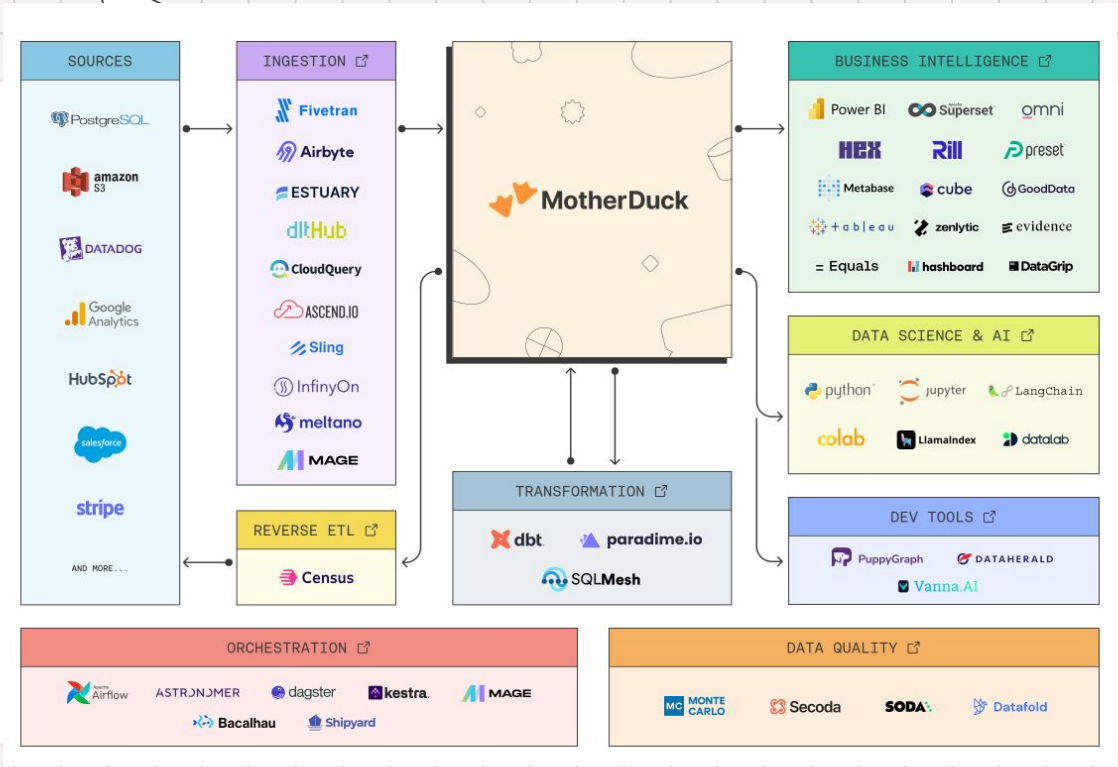
- As an Analytics Generator System
- As a Platform



Extensions

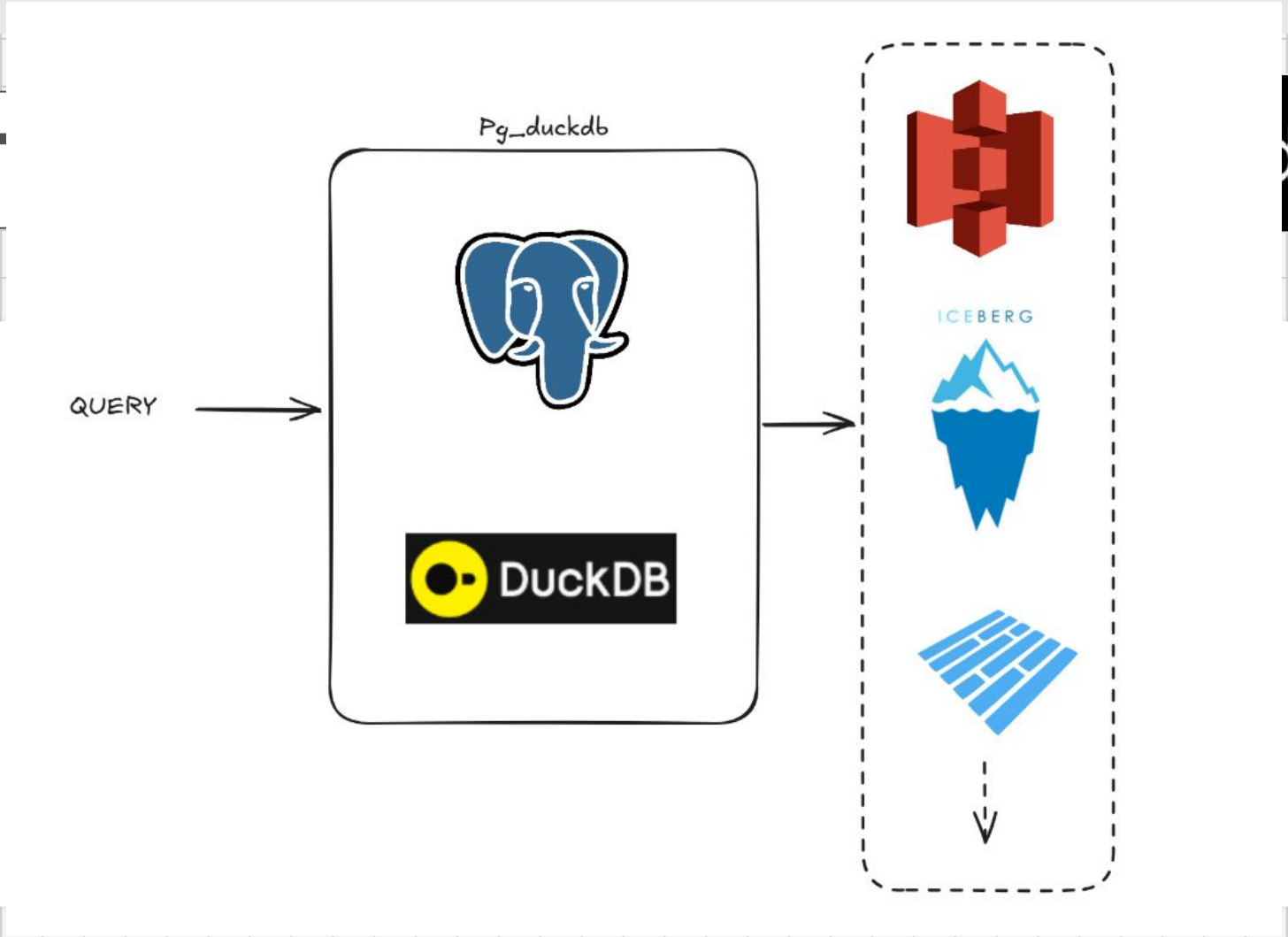
pg_duckdb

- Part of MotherDuck Ecosystem to ingest data from PostgreSQL into duckdb.
- With pg_duckdb we can treat PostgreSQL as an OLAP system alongwith its OLTP capabilities, using duckdb analytics engine as the base

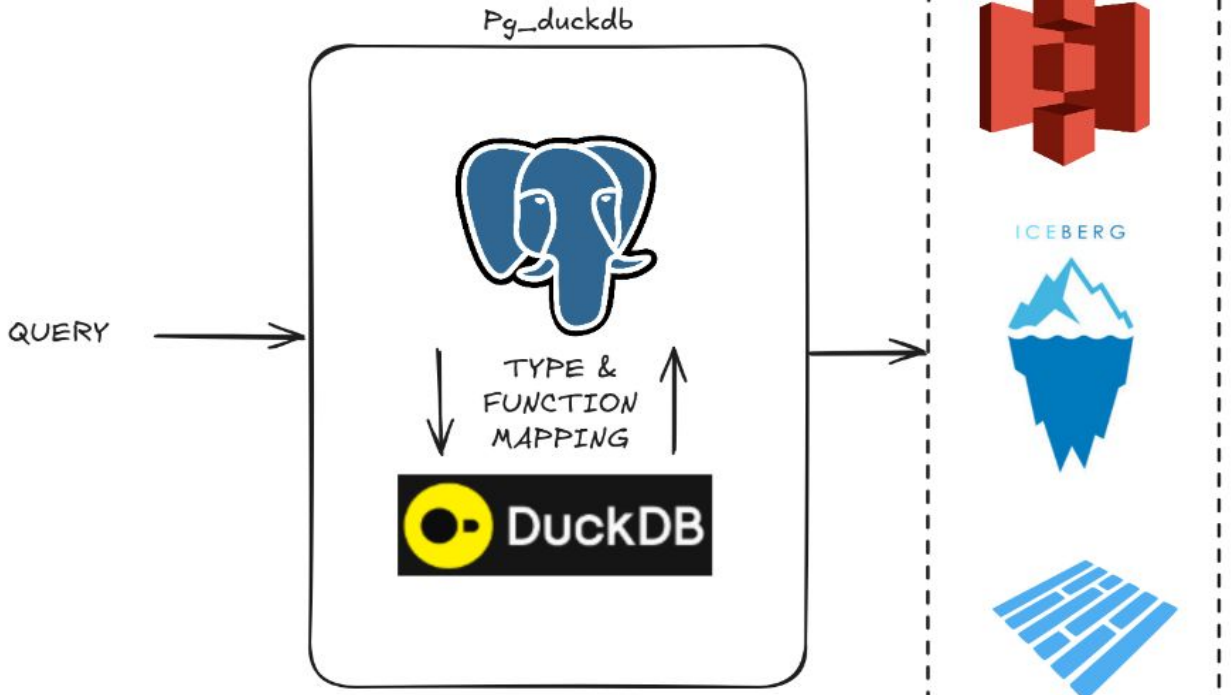


Ext

db



Ext



Extensions

- Uses DuckDB for query execution
- If DuckDB cannot execute the query, we fall back to PostgreSQL
- Build extensible OLAP pipelines using OpenTable formats like IceBerg
- Ability to integrate with:
    - Iceberg Tables
    - Parquet Files
    - S3 etc..

Extensions

pg_duckdb

- DataLake Usecases

LINK to DEMO Content

https://github.com/destrex271/talks

Extensions

pg_duckdb

- Data Lake Use cases

- Real Time Analytics

Thanks!