

Data Wrangling

DuckDB

[DuckDB](#) is a lightweight OLAP type database system written in C++ and designed to be used for EDA style activities:

When to use DuckDB



- Processing and storing tabular datasets, e.g. from CSV or Parquet files
- Interactive data analysis, e.g. Joining & aggregate multiple large tables
- Concurrent large changes, to multiple large tables, e.g. appending rows, adding/removing/updating columns
- Large result set transfer to client

When to not use DuckDB

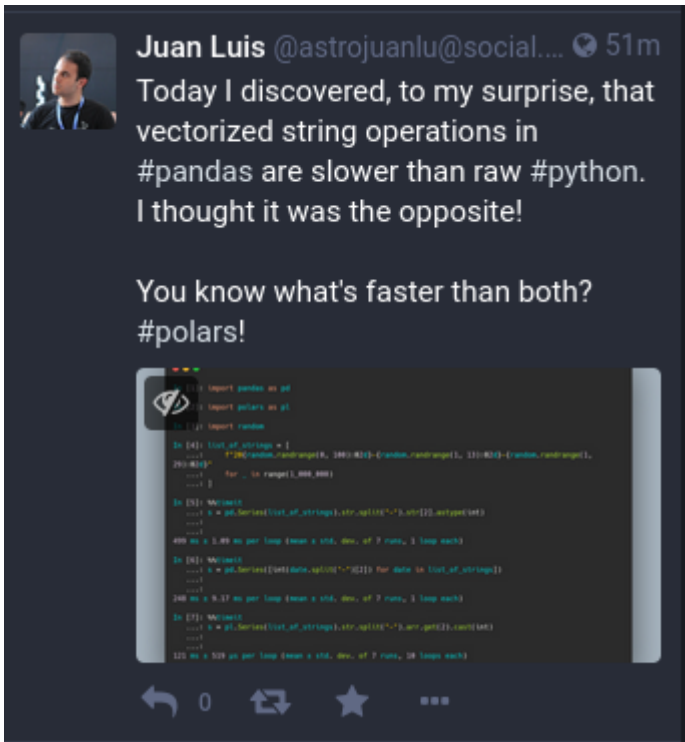


- High-volume transactional use cases (e.g. tracking orders in a webshop)
- Large client/server installations for centralized enterprise data warehousing
- Writing to a single database from multiple concurrent processes

From their website: advice on when to use and not to use DuckDB

Polars

[Polars](#) is a rust-based data frames library with Python bindings



[Here is a talk](#) that Juan Luis gave about the library

Revision #6

Created 4 November 2022 11:58:18 by James

Updated 8 February 2024 21:20:15 by James