

DuckDBはなぜ流行りはじめているのか？

LT@stapy#111

12/19, 2024

辻真吾 (www.tsjshg.info)

自己紹介

自己紹介は苦手です

1月
25

scipydata2025



SciPyData

ハッシュタグ : #scipydata2025

募集内容	現地参加 無料	先着順 11/100人
	Zoom登壇 無料	先着順 1/4人
出席登録	(イベント開始時間の2時間前から終了時間まで、参加者のみに公開されます)	

グループ

メンバーです

SciPyData



イベント数 1回
メンバー数 20人

開催前

2025/01/25(土)
13:00 ~ 18:30

 Googleカレンダー  icsファイル

✔ このイベントに参加できます

 受付票を見る

※受付や入場方法は主催者の案内に従ってください。

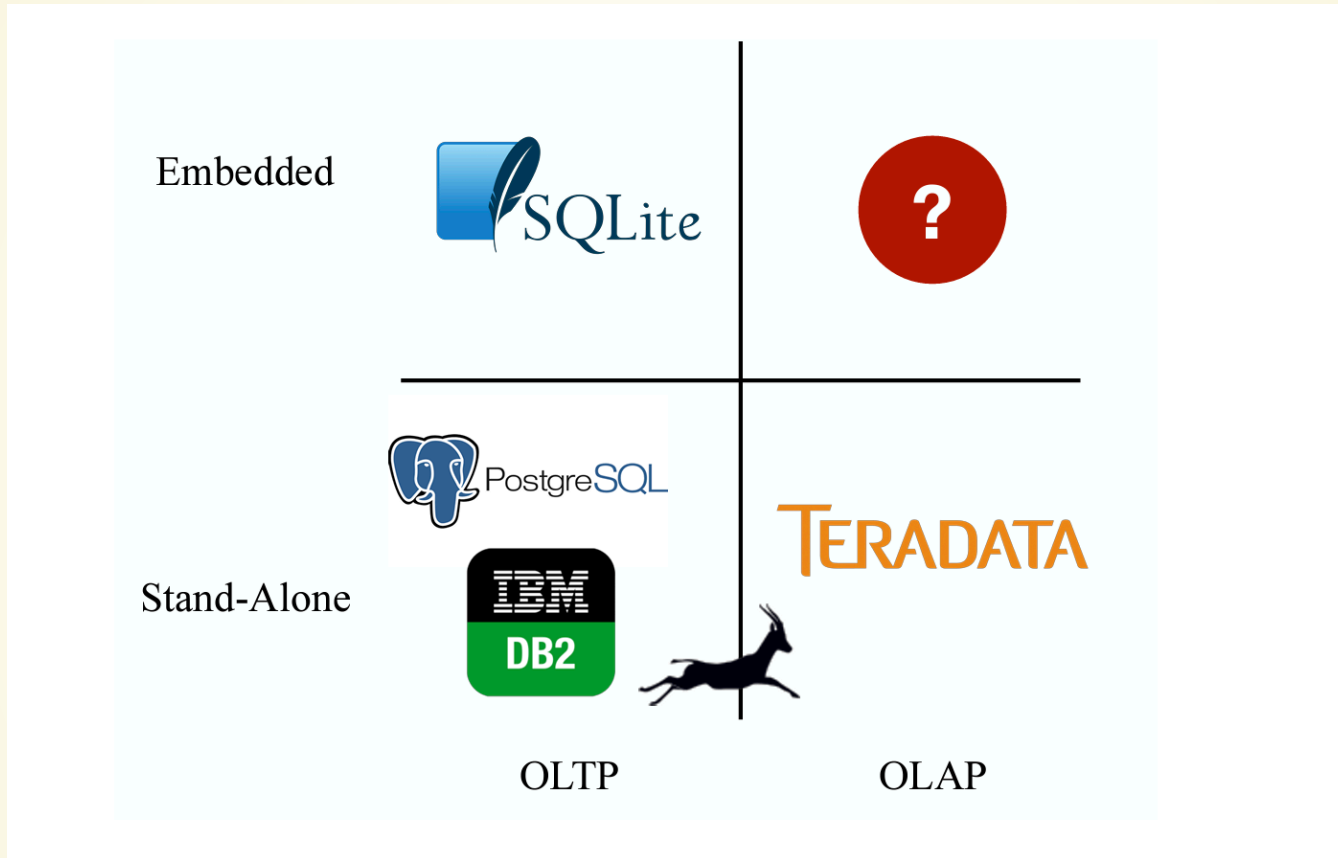
2025年1月25日（土）に茅場町で開催

<https://scipydata.connpass.com/event/337318/>

今日の内容

にわかに流行の兆しを見せつつあるDuckDBの紹介

DuckDBとは？



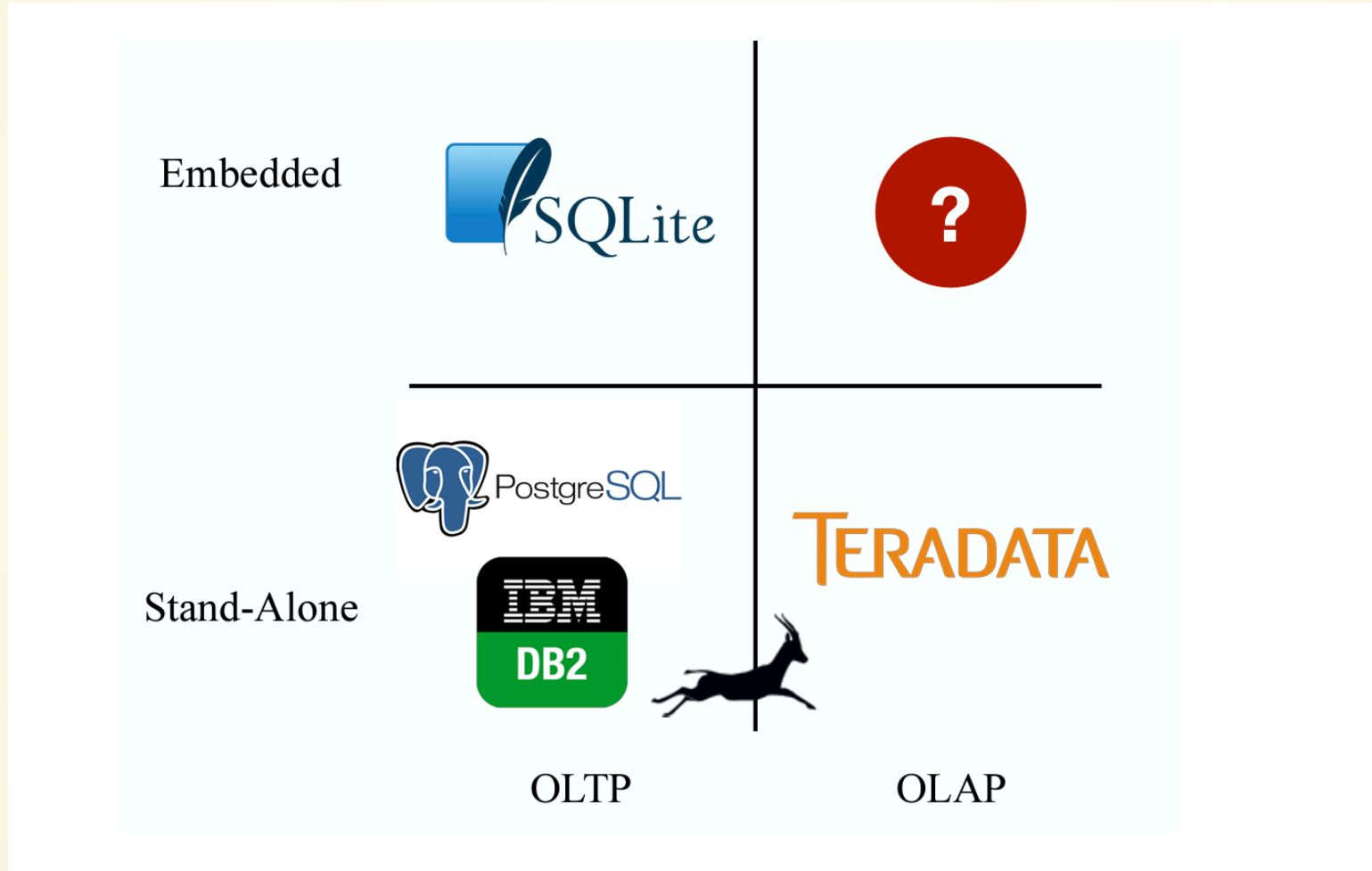
2019年の論文 **DuckDB: an Embeddable Analytical Database** から引用

DBの種類

- OLTP: Online Transaction Processing
 - オンラインショッピングでのお買い物を支えるデータベース
- OLAP: Online Analytical Processing
 - 集まったデータを分析するときに利用されるデータベース

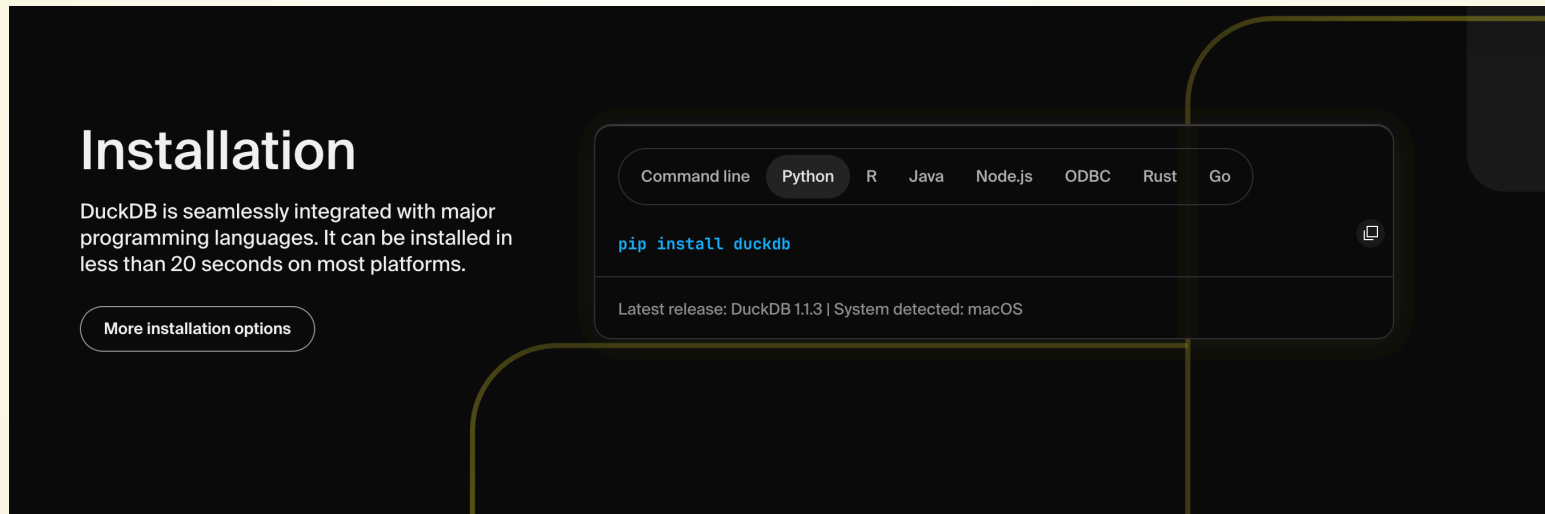
DuckDBの位置付け

SQLiteのように手軽に使える分析用データベース



DuckDBのインストール

各種プログラミング言語用のバンドルが用意されている



duckdb コマンド

コマンドラインからも起動できる（しかも1ファイル）

DuckDB Installation

This page contains installation options for DuckDB. For production use, we recommend the stable release.

Binaries are available for major programming languages and [platforms](#). If there are no pre-packaged binaries available, consider [building DuckDB from source](#).

Version

Stable release (v1.1.3)

Nightly build (bleeding edge)

Environment

Command line

Python

R

Java

Node.js

Rust

Go

C/C++

ODBC

Platform

Windows

macOS

Linux

Download method

Package manager

Direct download

Architecture

x86_64

arm64

Link

https://github.com/duckdb/duckdb/releases/download/v1.1.3/duckdb_cli-windows-arm64.zip

Note: Each DuckDB client is installed without relying on any other DuckDB clients. For example, the Python library can use a different version than the CLI client. Therefore, they need to be updated separately.

Usage example

`./duckdb`



サンプルデータの紹介

よく使われるワインの質を計測したデータ

フィルの拡張子はcsvなのに、;で区切られている・・・。

```
"fixed acidity";"volatile acidity";"citric acid";"residual sugar";"chlo  
7.4;0.7;0;1.9;0.076;11;34;0.9978;3.51;0.56;9.4;5  
7.8;0.88;0;2.6;0.098;25;67;0.9968;3.2;0.68;9.8;5  
7.8;0.76;0.04;2.3;0.092;15;54;0.997;3.26;0.65;9.8;5  
11.2;0.28;0.56;1.9;0.075;17;60;0.998;3.16;0.58;9.8;6  
7.4;0.7;0;1.9;0.076;11;34;0.9978;3.51;0.56;9.4;5  
7.4;0.66;0;1.8;0.075;13;40;0.9978;3.51;0.56;9.4;5  
7.9;0.6;0.06;1.6;0.069;15;59;0.9964;3.3;0.46;9.4;5  
7.3;0.65;0;1.2;0.065;15;21;0.9946;3.39;0.47;10;7  
7.8;0.58;0.02;2;0.073;9;18;0.9968;3.36;0.57;9.5;7  
7.5;0.5;0.36;6.1;0.071;17;102;0.9978;3.35;0.8;10.5;5  
6.7;0.58;0.08;1.8;0.097;15;65;0.9959;3.28;0.54;9.2;5  
7.5;0.5;0.36;6.1;0.071;17;102;0.9978;3.35;0.8;10.5;5  
5.6;0.615;0;1.6;0.089;16;59;0.9943;3.58;0.52;9.9;5  
7.8;0.61;0.29;1.6;0.114;9;29;0.9974;3.26;1.56;9.1;5  
8.9;0.62;0.18;3.8;0.176;52;145;0.9986;3.16;0.88;9.2;5  
8.9;0.62;0.19;3.9;0.17;51;148;0.9986;3.17;0.93;9.2;5  
8.5;0.28;0.56;1.8;0.092;35;103;0.9969;3.3;0.75;10.5;7
```

<https://archive.ics.uci.edu/dataset/186/wine+quality>

DuckDBの使い方

CSVファイルからテーブルを作ってデータベースを構築する

```
→ 20241219_DuckDB_LT duckdb wine.db
```

```
v1.1.3 19864453f7
```

```
Enter ".help" for usage hints.
```

```
D CREATE TABLE wine AS
```

```
    SELECT * FROM 'winequality-red.csv';
```

```
D select * from wine limit 5;
```

fixed acidity double	volatile acidity double	citric acid double	...	sulphates double	alcohol double	quality int64
7.4	0.7	0.0	...	0.56	9.4	5
7.8	0.88	0.0	...	0.68	9.8	5
7.8	0.76	0.04	...	0.65	9.8	5
11.2	0.28	0.56	...	0.58	9.8	6
7.4	0.7	0.0	...	0.56	9.4	5
5 rows						12 columns (6 shown)

```
D .exit
```

```
→ 20241219_DuckDB_LT
```

```
→ 20241219_DuckDB_LT duckdb wine.db
```

```
v1.1.3 19864453f7
```

```
Enter ".help" for usage hints.
```

```
D select count(*) from wine;
```

count_star() int64
1599

地味に好きな機能 (.columns)

```
D .columns
D select * from wine limit 5;
```

Column	Type	Row 1	Row 2	Row 3	Row 4	Row 5
fixed acidity	double	7.4	7.8	7.8	11.2	7.4
volatile acidity	double	0.7	0.88	0.76	0.28	0.7
citric acid	double	0.0	0.0	0.04	0.56	0.0
residual sugar	double	1.9	2.6	2.3	1.9	1.9
chlorides	double	0.076	0.098	0.092	0.075	0.076
free sulfur dioxide	double	11.0	25.0	15.0	17.0	11.0
total sulfur dioxide	double	34.0	67.0	54.0	60.0	34.0
density	double	0.9978	0.9968	0.997	0.998	0.9978
pH	double	3.51	3.2	3.26	3.16	3.51
sulphates	double	0.56	0.68	0.65	0.58	0.56
alcohol	double	9.4	9.8	9.8	9.8	9.4
quality	int64	5	5	5	6	5

全ページと同じ結果がちゃんと画面に入る

DuckDBのここがスゴイ

巨大なテーブルを扱える

```
select count(*) from page;
```

count_star() int64
60947455

英語版Wikipediaの全ページ（6094万7455）

ファイルサイズ113GB（マシンはメモリサイズ24GB）

DuckDBは速い

```
D select * from page where page_title like '%Random%forest%';
```

100%

page_id int32	page_namespace int32	page_title varchar	...	page_len int32	page_content_model varchar	page_lang varchar
1363880	0	Random_forest	...	47582	wikitext	NULL
2339973	1	Random_forest	...	24884	wikitext	NULL
6716209	0	Random_forests	...	27	wikitext	NULL
3 rows						12 columns (6 shown)

Run Time (s): real 6.060 user 3.561452 sys 0.796797

6000万行以上からlike文でそれっぽいものを探してくるのにたったの6秒（10コアのM4 Mac mini）

DuckDBは速い

```
D select * from page where page_title = 'Random_forests';
```

page_id int32	page_namespace int32	page_title varchar	...	page_len int32	page_content_model varchar	page_lang varchar
6716209	0	Random_forests	...	27	wikitext	NULL
1 rows						12 columns (6 shown)

```
Run Time (s): real 0.258 user 2.441850 sys 0.006320
```

=で指定するなら0.3秒

その他のすごいところ

- リリースから3年パフォーマンスが改善され続けている
 - <https://duckdb.org/2024/06/26/benchmarks-over-time.html>
- Pythonから使うとselectの結果をpandas.DataFrameにできる
- Wasm(WebAssembly)版がある
 - つまりWebブラウザの中だけで動く
- などなど

もちろん万能ではない

データサイズが500GBを超えてくるとパフォーマンスがそれほど良くないという報告（2023なのでちょっと古い）

How fast is DuckDB really?

まとめ

- 日常業務で扱うデータが肥大化する昨今、手軽でかつ高性能なDuckDBに注目が集まるのは必然か
- マシンのメモリサイズを超える数10GB～500GBくらいのデータにはDuckDBおすすめ
 - pandasではメモリ乗らない、乗っても遅いレンジのデータ
- 今日の結論
 - SQLを書こう！