FireDucks

Compiler Accelerated DataFrame Library with pandas API

2024/06/20

石坂一久

みんなのPython勉強会#105

自己紹介

石坂 一久

（NEC セキュアシステムプラットフォーム研究所所属）

＜これまでの関わってきた主な領域＞

自動並列化コンパイラ

並列処理・ベクトル処理

ソフトウェアの高速化が生業

Intel Xeon Phi
（メニコア）

NEC SX-Aurora TSUBASA
（スパコン）

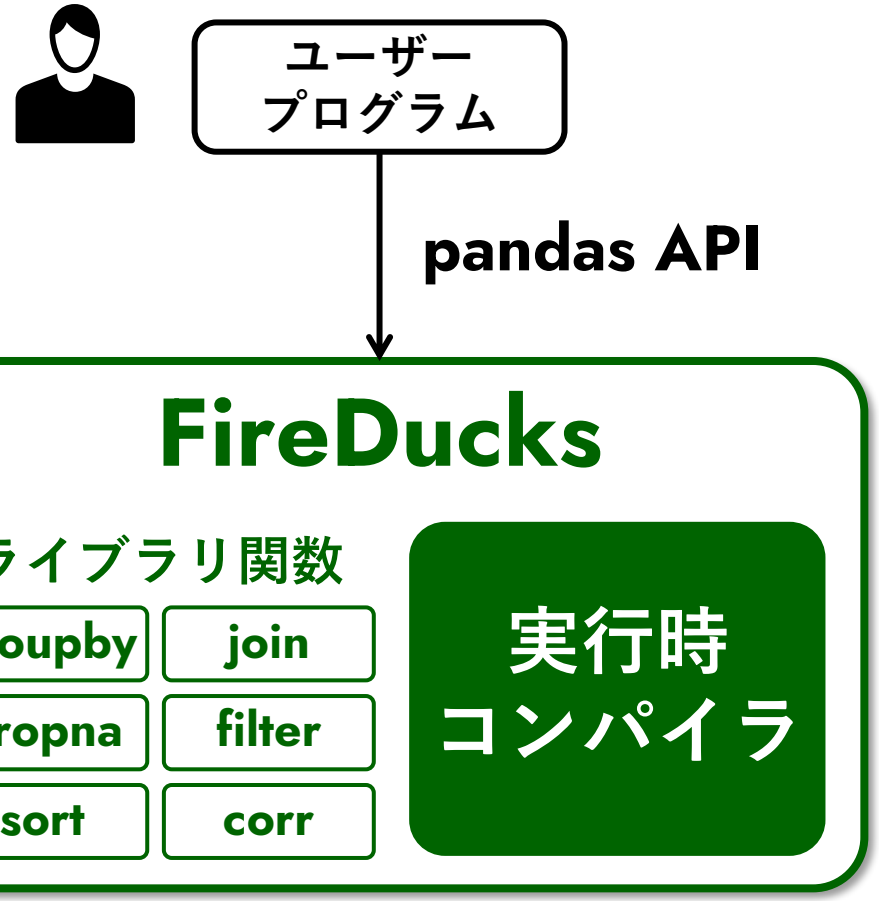
https://pc.watch.impress.co.jp/docs/news/yajiuma/1238340.html
https://jpn.nec.com/hpc/sxauroratsubasa/specification/index.html

FireDucks: pandas APIの高速実装

コンパイラ技術
で何か作りたい

Pythonを
高速化したい

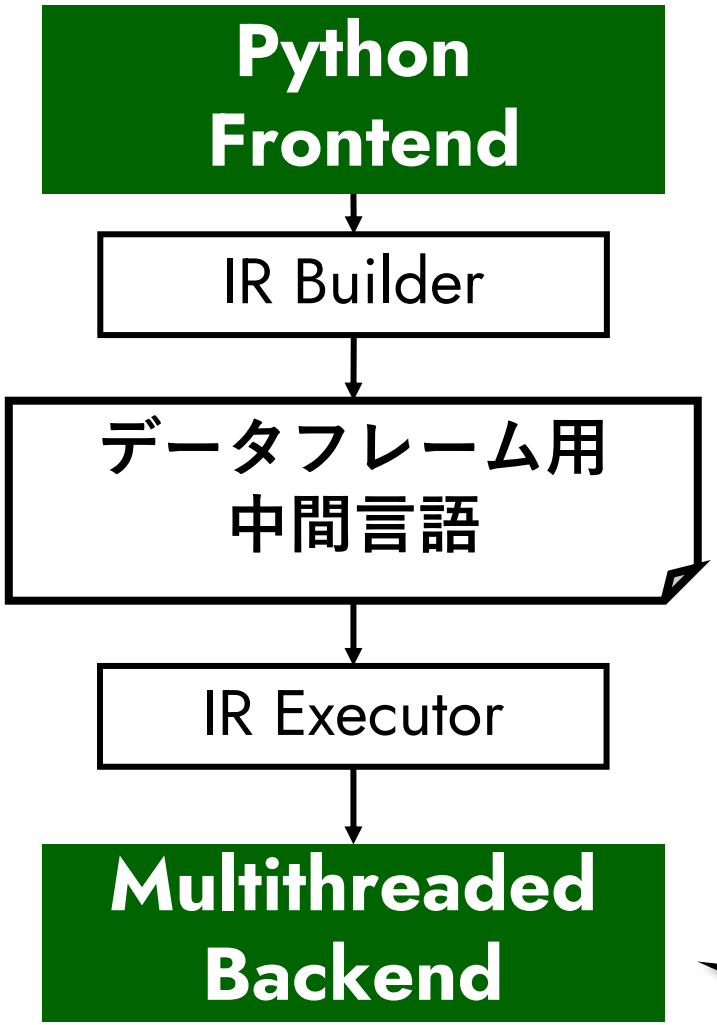
pandasが
遅くて困ってる
人がいる



pandasの高速化版を作ろう！

（DataFrame Compilerを作ろう）

FireDucksの構造と高速化の仕組み



FireDucksの構造

高速化1

pandasプログラムの
高速化ノウハウを自動化

自動最適化

MLIR

コンパイラ
フレームワーク

高速化2

マルチスレッド化された
バックエンドで中間言語を実行

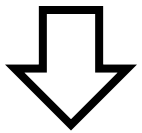
自動最適化の例

ユーザーが書いたプログラム

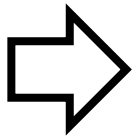
```
df = pd.read_csv("sample.csv")
sorted = df.sort_values("b")
result = sorted[["a"]]
```

実際に実行される処理

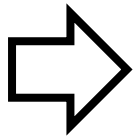
```
df = pd.read_csv("sample.csv")
df2 = df[[“a”, “b”]]
sorted = df2.sort_values("b")
result = sorted[["a"]]
```



Python Frontend



データフレーム用 中間言語

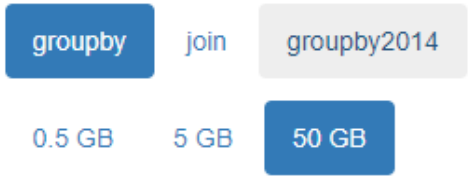


自動最適化

```
%v1 = "fireducks.read_csv"("sample.csv")
%v2 = "fireducks.sort_values"(%v1,"b")
%v3 = "fireducks.project"(%v2,["a"])
```

FireDucksの性能 (要素処理 groupby, join)

Database-like ops benchmark (https://duckdblabs.github.io/db-benchmark)



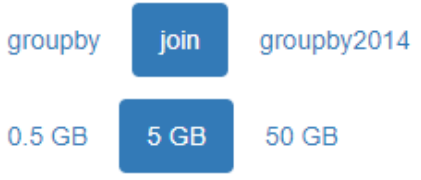
basic questions

Input table: 1,000,000,000 rows x 9 columns (50 GB)

1位

FireDucks	0.12.1	2024-06-17	14s
duckdb-latest	0.9.1.1	2023-10-26	24s
DuckDB	0.8.1.3	2023-10-26	25s
ClickHouse	23.10.4.25	2023-11-30	29s
Polars	0.19.8	2023-10-17	32s
DataFrames.jl	1.6.1	2023-10-17	84s
data.table	1.14.9	2023-10-17	89s
Datafusion	31.0.0	2023-10-24	133s
InMemoryDataSets.jl	0.7.13	2023-10-17	218s
collapse	2.0.3	2023-10-26	233s
spark	3.5.0	2023-10-24	297s
R-arrow	13.0.0.1	2023-10-17	511s
dask	2023.10.0	2023-11-29	544s
pandas	2.1.1	2023-10-17	773s
(py)datatable	1.1.0a0	2023-10-17	993s
dplyr	1.1.3	2023-10-17	1022s
Modin		see README	pending

Groupby



basic questions

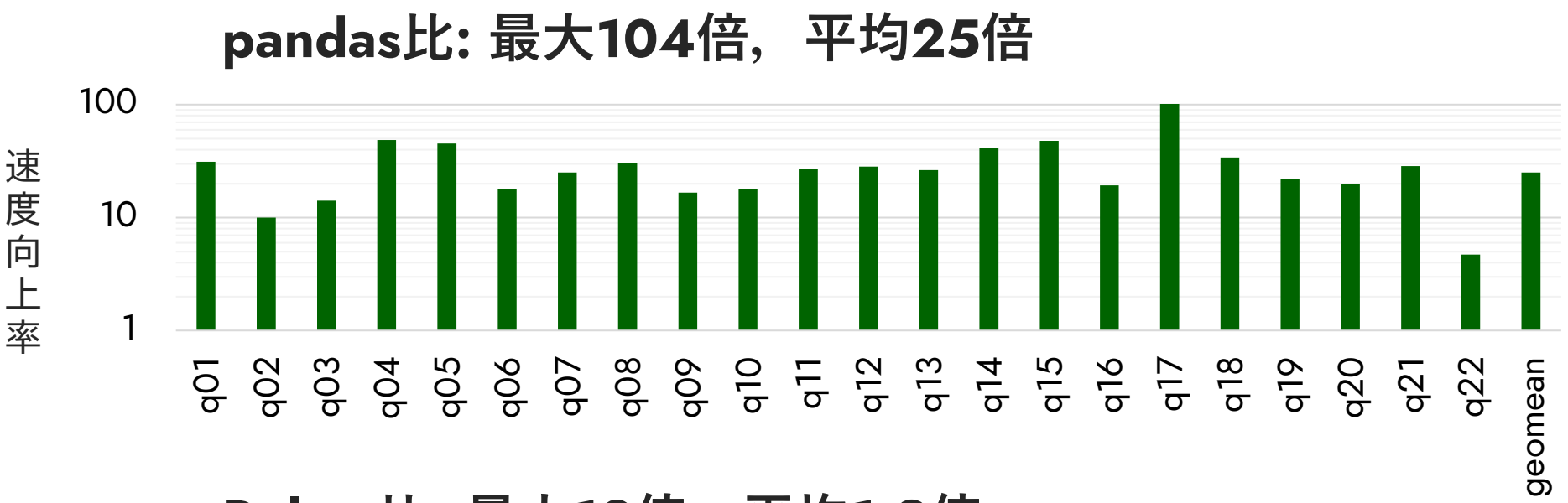
Input table: 100,000,000 rows x 7 columns (5 GB)

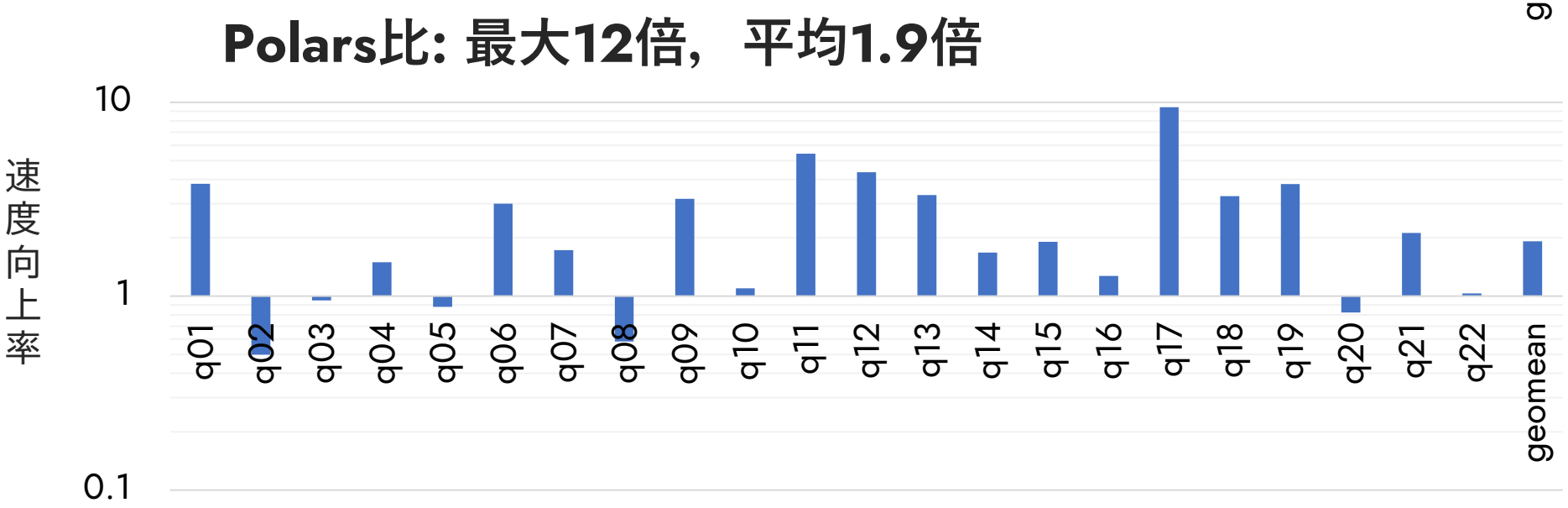
3位

DuckDB	0.8.1.3	2023-10-20	8s
duckdb-latest	0.9.1.1	2023-10-25	8s
FireDucks	0.12.1	2024-06-17	8s
Polars	0.19.8	2023-10-20	14s
Datafusion	31.0.0	2023-10-25	22s
InMemoryDataSets.jl	0.7.13	2023-10-20	25s
ClickHouse	23.10.4.25	2023-11-30	42s
data.table	1.14.9	2023-10-20	55s
collapse	2.0.3	2023-10-26	65s
DataFrames.jl	1.6.1	2023-10-20	71s
spark	3.5.0	2023-10-24	129s
dask	2023.10.0	2023-11-29	179s
dplyr	1.1.3	2023-10-20	225s
pandas	2.1.1	2023-10-20	265s
(py)datatable	1.1.0a0	2023-10-20	5699s
R-arrow	13.0.0.1	2023-10-20	out of memory
Modin		see README	pending

Join

FireDucksの性能 (TPC-Hベンチマーク Scale Factor=10)





評価環境

インテル® Xeon® Gold 5317 プロセッサー
(12コア x 2ソケット)

メモリ: 256GB
OS: Linux

pandas 2.2.0
polars 0.20.7
FireDucks 0.10.1

ベンチマークコード

https://github.com/fireducks-dev/polars-tpch/tree/fireducks

ぜひFireDucksをご利用ください

pipコマンドでインストール可能（BSDライセンス）

```
$ pip install fireducks
```

pandas互換のため既存プログラムの修正や新たな学習は不要

1) import pandasの自動置き換え（python起動オプション）

```
$ python3 -m fireducks.pandas program.py
```

jupyter notebookではマジックコマンド

```
%load_ext fireducks.pandas
```

2) もしくは, import文の書き換え

```
import fireducks.pandas as pd
```

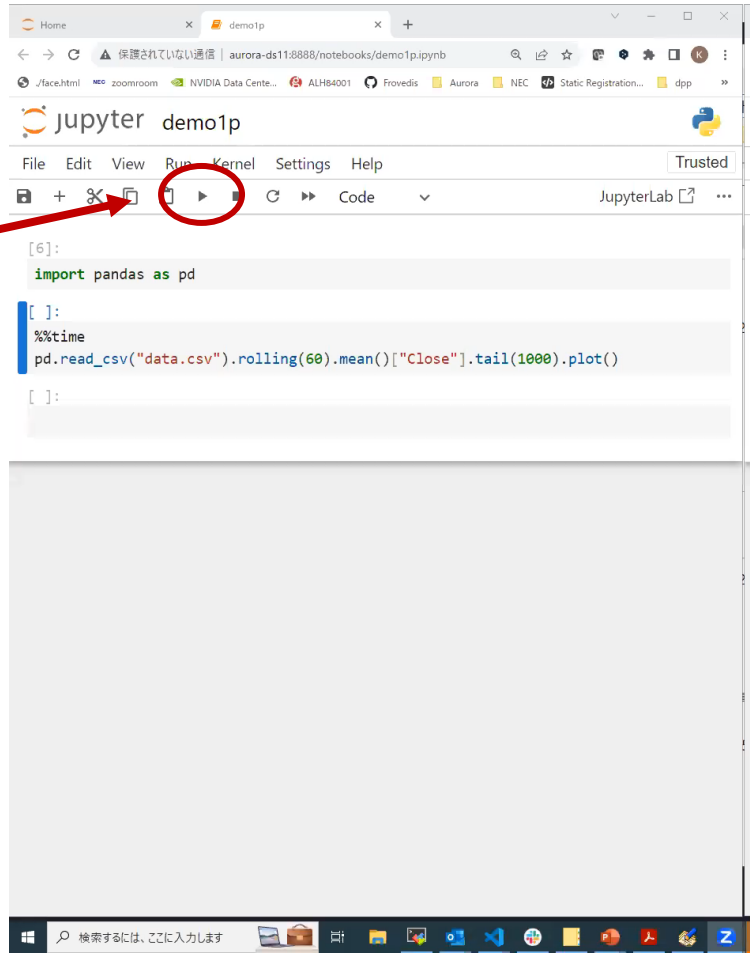
Demo

```
pd.read_csv(“data.csv”).rolling(60).mean()[“Close”].tail(1000).plot()
```

実行開始
ボタン

移動平均を
計算する
プログラム

pandas



FireDucks