Why you should move from Pandas to Polars

Alon Samuel

Where I took info from

Using polars for 4ish months

Jetbrains post

Medium post

ChatGPT

Polars

Introduction - what is Polars

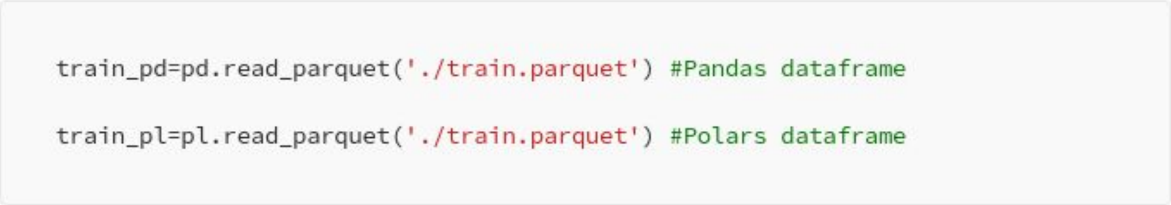
- Designed for Single-Machine Use
- Similar to pandas
- Written in Rust: Almost as Fast as C and C++
    - Safe Concurrency for Parallelism
- Based on Arrow: Language-Independent Memory Format

Why Choose Polars Over pandas?

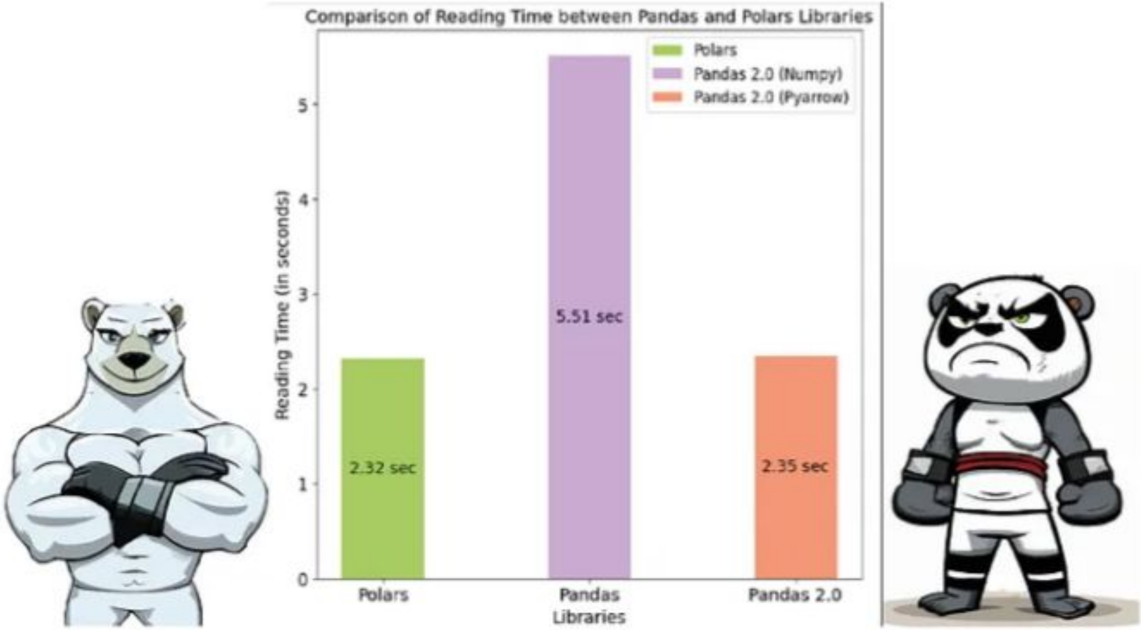
- Key Factor: **Performance**
- Speed: 5–10 Times Faster Than pandas
- Memory Requirement: 2–4 Times Smaller Than pandas

Examples

Reading

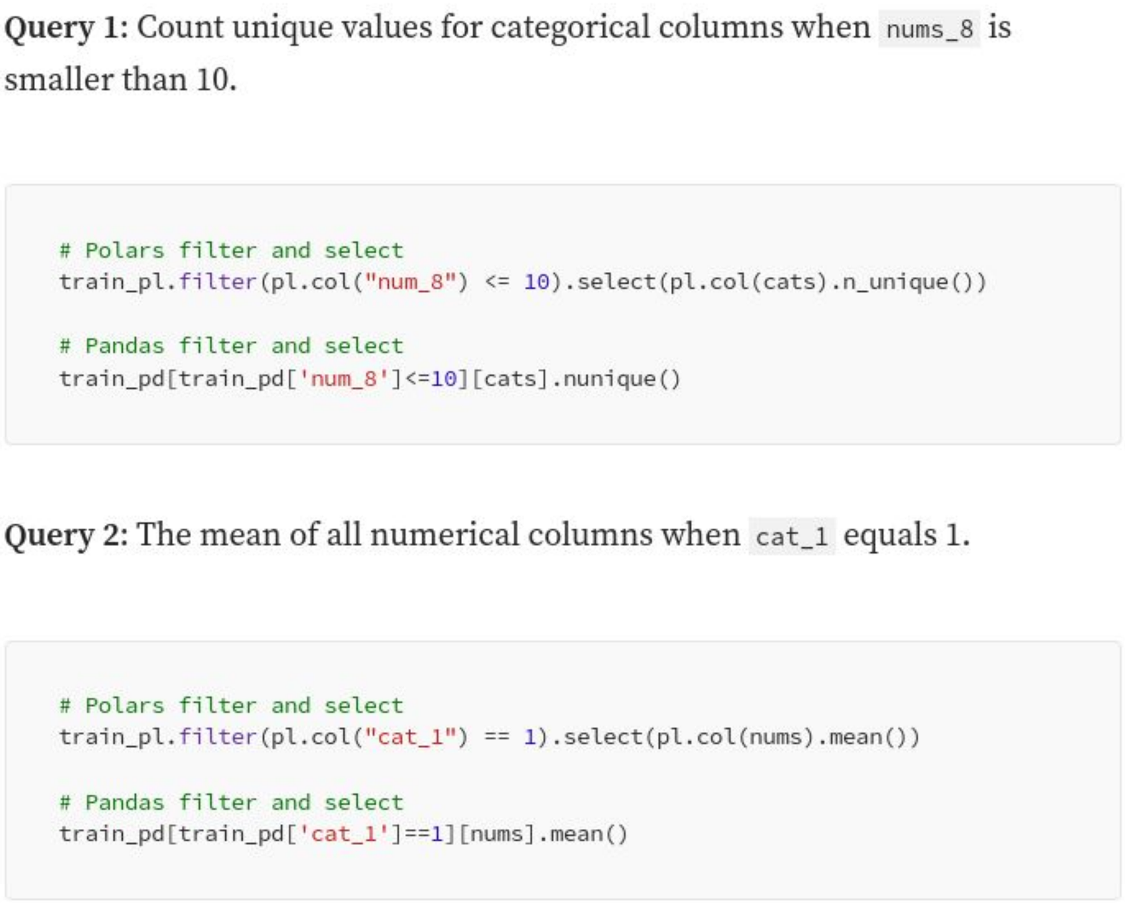


```
train_pd=pd.read_parquet('./train.parquet') #Pandas dataframe

train_pl=pl.read_parquet('./train.parquet') #Polars dataframe
```



Reading time comparison. Image by author

Filtering



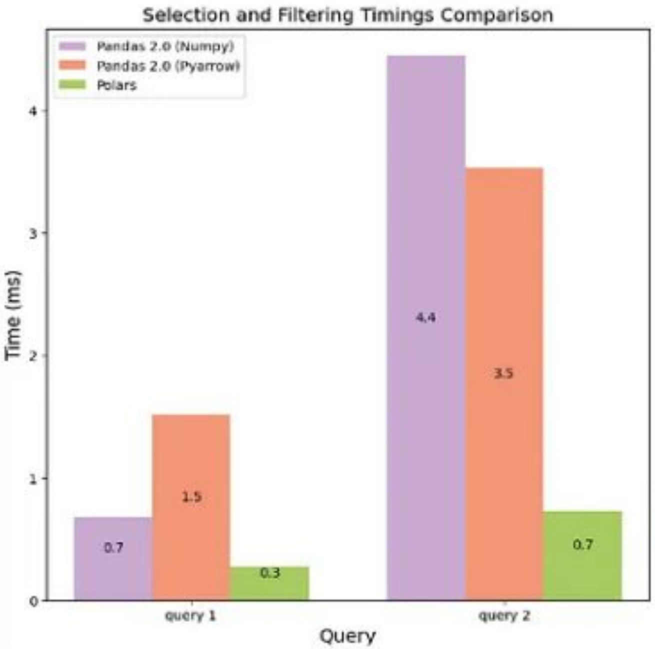
Query 1: Count unique values for categorical columns when `nums_8` is smaller than 10.

```
# Polars filter and select
train_pl.filter(pl.col("num_8") <= 10).select(pl.col(cats).n_unique())

# Pandas filter and select
train_pd[train_pd['num_8']<=10][cats].nunique()
```

Query 2: The mean of all numerical columns when `cat_1` equals 1.

```
# Polars filter and select
train_pl.filter(pl.col("cat_1") == 1).select(pl.col(nums).mean())

# Pandas filter and select
train_pd[train_pd['cat_1']==1][nums].mean()
```

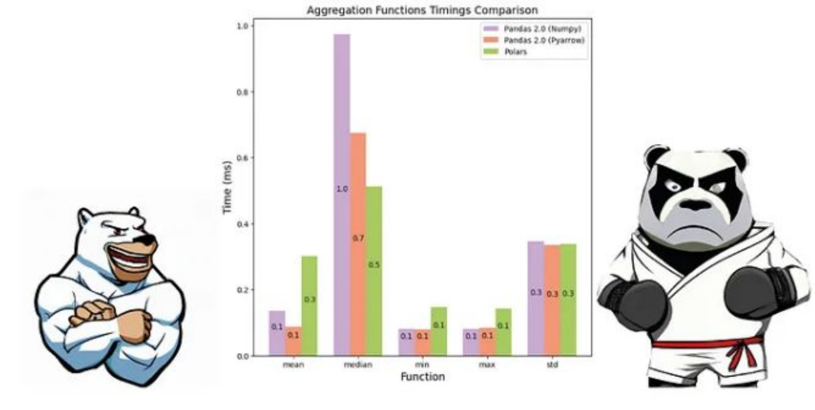
Filtering



Selection and filtering time comparison. Image by author

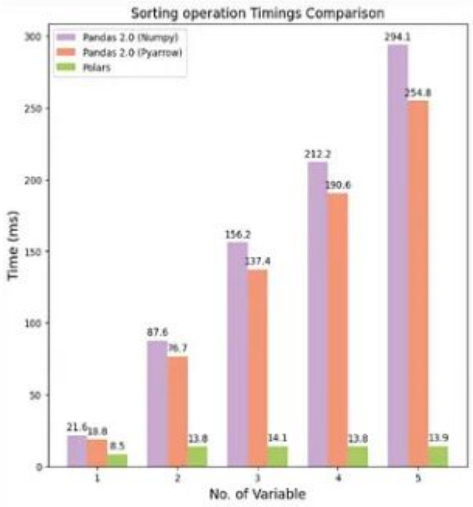
Aggregations

```
# pandas query
train_pd[nums].agg(['min','max','mean','median','std')
train[cats].agg(['nunique'])

# Polars query
train_pl.with_columns([
    pl.col(nums).min().suffix('_min'),
    pl.col(nums).max().suffix('_max'),
    pl.col(nums).mean().suffix('_mean'),
    pl.col(nums).median().suffix('_median'),
    pl.col(nums).std().suffix('_std'),
    pl.col(cats).nunique().suffix('_unique'),
])
```

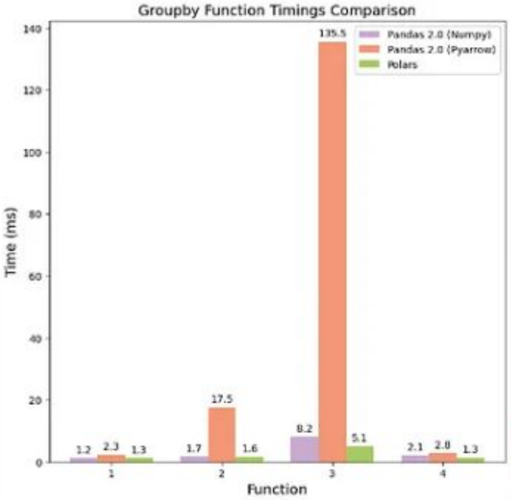


Aggregation function time comparison. Image by author

More examples



Sorting time comparison. Image by author



Comparison of groupby functions. Image by author

How to start

- Do not use ChatGPT!
  - Use StackOverflow!
- Only if needed - big DFs, Memory issues, multiple cores machines.

When to Stick with pandas?

- Polars for
  - Efficient transformations
  - Machine learning pipelines
- Pandas Continued Strengths:
  - Data Exploration
  - ‘Understanding’ data types
  - Excel sheets
  - Visualisations

Thank you!

Alon Samuel

Linkedin

alon1samuel@gmail.com