Polars: fast DataFrames in Python

Raoul Schram

January 25, 2023

Utrecht University

Python DataFrame libraries

- Pandas
  - Most widely used DataFrame library for python
  - Based on NumPy
  - Getting old/slow/quirks
- Dask
  - Based on Pandas
  - Distributed memory
- Polars
  - Based on Apache Arrow, written in Rust
  - Lower user base
  - Fast

Utrecht University

Pandas versus Polars

- No index
- Supports automatic parallel computing
- Lazy (optional)
- More consistent data typing
- More consistency with missing data
- Polars can convert between Polars and Pandas DataFrames

Utrecht University

Lazy evaluation

Pandas:

```
import pandas as pd
df = pd.read_csv(csvFile, use_columns=['id1', 'v1'])
grouped_df = df.loc[:,['id1','v1']].groupby('id1').sum('v1')
```

Polars:

```
import polars as pl
df = pl.scan_csv(csvFile)
grouped_df = df.groupby('id1').agg(pl.col('v1').sum()).collect()
```

- Saves memory
- Allows optimizations/parallelization

Utrecht University

Conclusion

- You **should** try/use Polars if:
  - You like shiny new things
  - You dislike / have trouble with Pandas API
  - Pandas is too slow for you
- You **should not** try/use Polars if:
  - You need distributed memory (dask)
  - You panic if you can't find your solution on Stack Overflow
  - A lack of extensive documentation is a deal breaker

Utrecht University