

Polars and RapidsAI



Peter Belonovskiy



Short Intro



Тренды 2023 года

Библиотеки и фреймворки

1. langchain
2. python-polars
3. faiss
4. sentence-transformers
5. huggingface-datasets
6. jax
7. stable-baselines
8. onnxruntime
9. huggingface-transformers
10. pytorch-geometric

Data Processing in Python



RAPIDS



Достоинства:

1. Простой синтаксис (все привыкли)
2. Большая экосистема (интеграция с другими библиотеки)
3. Фактический стандарт обработки данных в питоне



Недостатки (Скорость и оптимизация):

1. Single thread режим
2. ОЗУ переполняется
3. Не оптимизированные операции



Task

groupby

join


















0.5 GB

5 GB

50 GB

basic questions

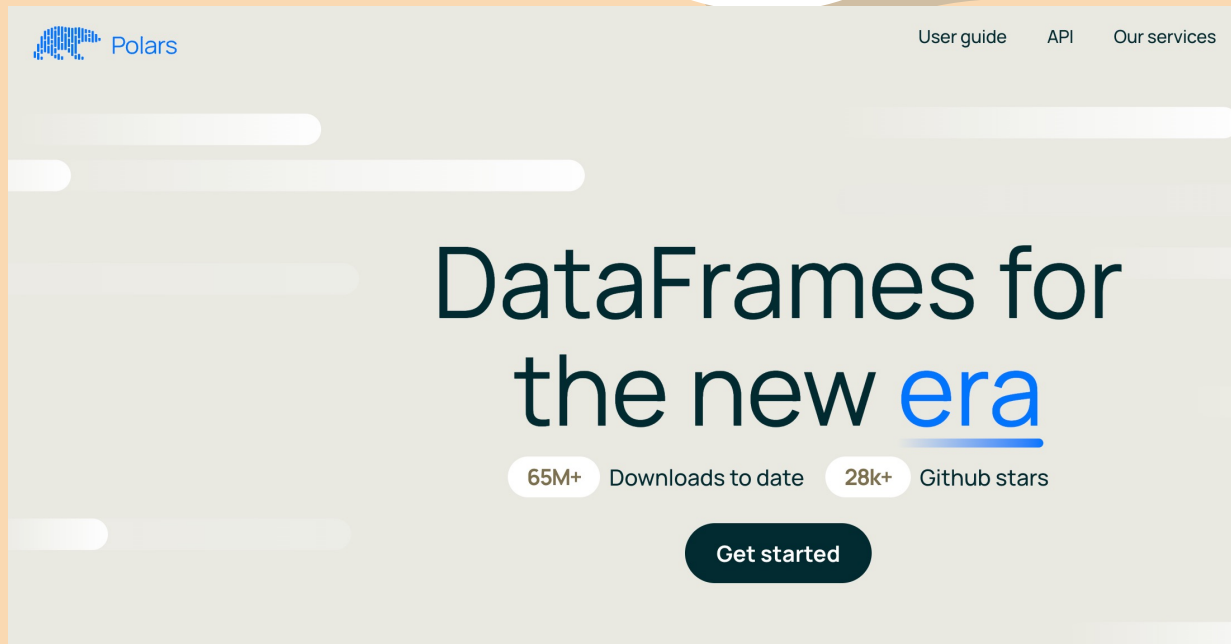
Input table: 100,000,000 rows x 7 columns (5 GB)

	DuckDB	1.0.0	2024-07-04	9s	
	Polars	1.1.0	2024-07-08	9s	
	Datafusion	38.0.1	2024-06-07	15s	
	InMemoryDataset.jl	0.7.18	2023-10-20	25s	
	ClickHouse	24.5.1.1763	2024-06-07	43s	
	data.table	1.15.99	2024-06-07	62s	
	collapse	2.0.14	2024-06-07	69s	
	DataFrames.jl	1.6.1	2024-06-07	77s	
	spark	3.5.1	2024-06-07	128s	
	dplyr	1.1.4	2024-06-07	214s	
	pandas	2.2.2	2024-06-07	244s	
	dask	2024.5.2	2024-06-07	635s	
	(py)datatable	1.2.0a0	2024-06-07	undefined exception	
	R-arrow	16.1.0	2024-06-07	out of memory	
	Modin		see README	pending	

<https://duckdblabs.github.io/db-benchmark/>



Polars

A mockup of the Polars website. The header features the Polars logo (a blue elephant) and the word "Polars" on the left, and navigation links "User guide", "API", and "Our services" on the right. The main content area has a large heading "DataFrames for the new era" where "era" is underlined in blue. Below the heading, there are two statistics: "65M+ Downloads to date" and "28k+ Github stars". At the bottom of the main area is a dark blue button with the text "Get started".

Polars

User guide API Our services

DataFrames for the new era

65M+ Downloads to date 28k+ Github stars

Get started



Rapids AI (CUDF)

RAPIDS

GPU Accelerated Data Science

QUICK START



NEW: CUDF NOW PRE-INSTALLED IN GOOGLE COLAB



NEW: CUVS VECTOR SEARCH



CATCH UP WITH GTC'24 ON-DEMAND



RAPIDS 24.08 RELEASED





Polars – логичный шаг в развитии инструментов обработки данных в Питоне. Быстрый, оптимизированный, параллельный, синтаксис легко выучить, предоставляет новые возможности обработки больших данных. Можем попробовать использовать в наших проектах.

RapidsAI(CuDF) – интересный инструмент, сильно ускоряет обработку данных, но не очень практичный. Наиболее подходит для исследовательской работы. Можно попробовать на ДатаЛабе.

Test

Сколько было показано мартышек в презентации? (Включая эту)

1. -1
2. 8
3. 1000

