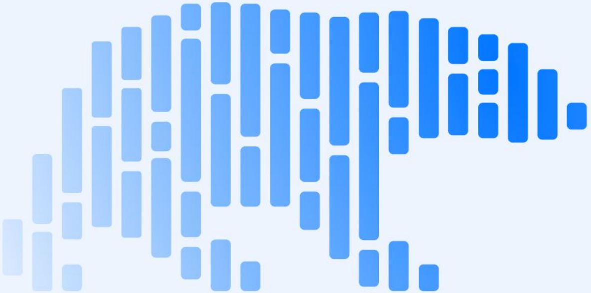


Polars

for Data Analysis in Python

dKd
Dr Kim Data

Kimberly Fessel, PhD

Kimberly Fessel, PhD

★ Data scientist, ~10 years

★ Data educator, ~10 years

★ ex-Director of Data Science Bootcamp

★ Founder of Dr Kim Data

dKd
DR KIM DATA

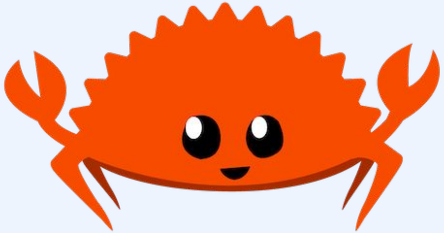
What is Polars?

Polars is…

A high-performance **DataFrame** library.

Objective

Efficient data handling and analysis



Built on Rust

Leverages speed of Rust programming language, with a Python API

Open Source

Free to download and modify

Development and Adoption

2020

Released in 2020
by Ritchie Vink

65M+

Over 65 million
downloads

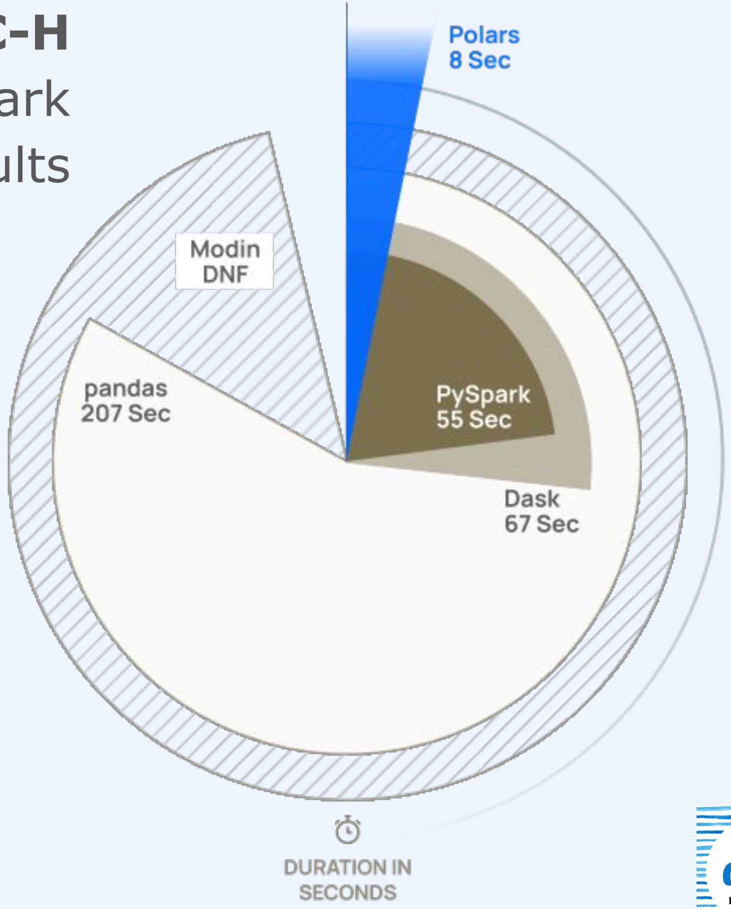
Key Features

Speed

30x

Up to 30x faster than pandas

TPC-H Benchmark Results



Parallel Computing and Memory Efficiency

★ Automatic multi-threading

★ Vectorized operations

★ Aggregations in parallel

★ Arrow memory format

Lazy Evaluation (Optional)

Computational strategy where expressions are **not evaluated until** their values are **needed**

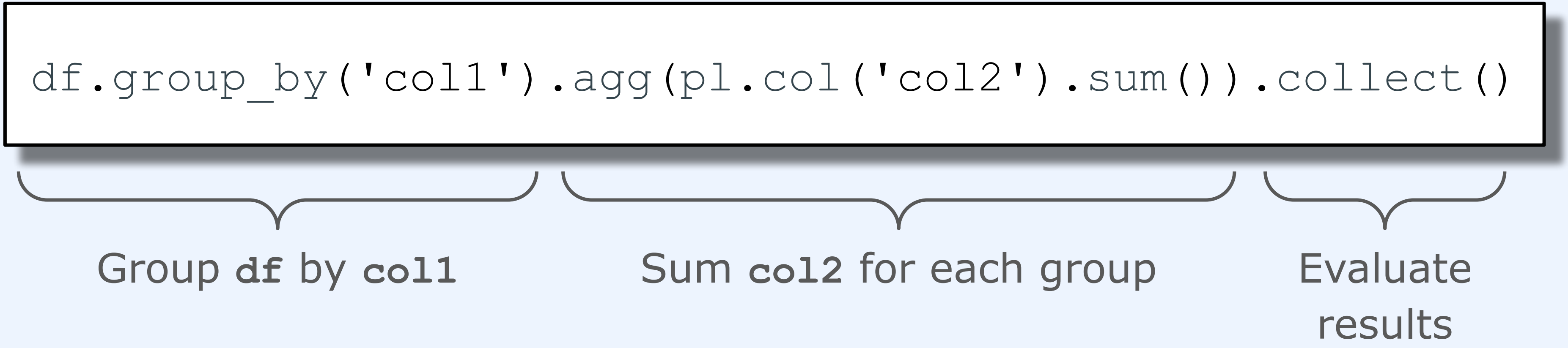
✓ Optimized execution plan

✓ Reduced resource usage

dKd
DR KIM DATA

Polars vs. Others

Polars Syntax



```
df.group_by('col1').agg(pl.col('col2').sum()).collect()
```

Group **df** by **col1**

Sum **col2** for each group

Evaluate results

Python Polars looks like a cross between **pandas** and **PySpark**.

Comparison with Pandas

Similarities

- ★ Data handling and analysis
- ★ DataFrame and Series objects
- ★ Many of the same operations
- ★ Familiar syntax

Differences

- ★ No index names
- ★ More parallelism
- ★ Optional lazy evaluation
- ★ Better syntax?

> *"Came for the **speed**, stayed for the **syntax**."*
>
> *-Polars Enthusiasts*

dask

★ Out-of-core processing

★ Parallelism, lazy evaluation

★ Python vs. Rust

★ Cluster vs. single machine

★ Pandas vs. unique syntax

APACHE Spark

★ Performant for big data

★ Parallelism, lazy evaluation

★ Java/Scala vs. Rust

★ Cluster vs. single machine

★ Dataset size

Getting Started

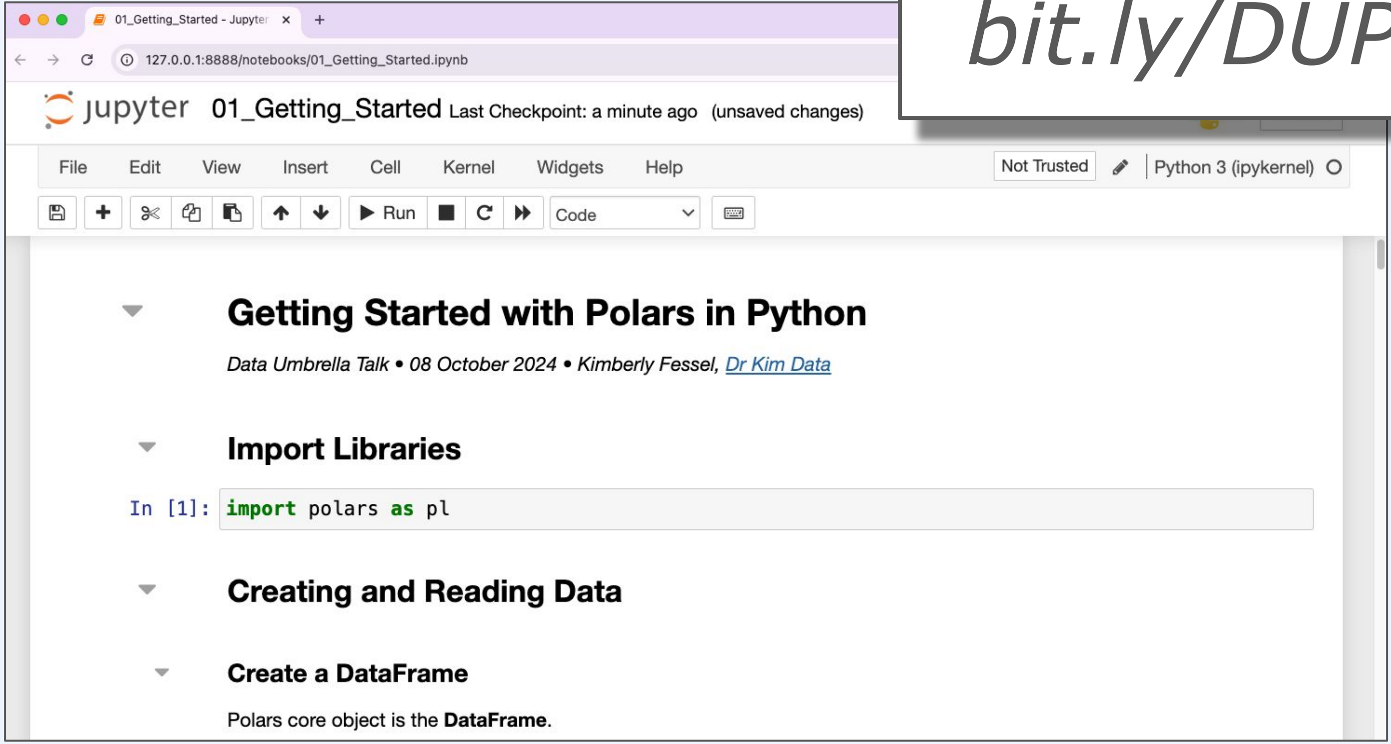
Installing Polars

```
pip install 'polars[plot]'
```

Adds plotting capabilities

Getting Started with Polars in Python

bit.ly/DUPolars


01_Getting_Started - Jupyter
127.0.0.1:8888/notebooks/01_Getting_Started.ipynb
jupyter
01_Getting_Started Last Checkpoint: a minute ago (unsaved changes)
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Not Trusted
Python 3 (ipykernel)
Run
Code

Getting Started with Polars in Python

Data Umbrella Talk • 08 October 2024 • Kimberly Fessel, [Dr Kim Data]

Import Libraries

```
In [1]: import polars as pl
```

Creating and Reading Data

Create a DataFrame

Polars core object is the **DataFrame**.

Should You Switch to Polars?

Additional Data Source Options

Excel

Parquet

JSON

Database

Working with Large Files

Lazy Evaluation

Allow Polars to create and execute optimized performance plan

Streaming

Work with **larger-than-RAM** data by passing a file path and utilizing the **streaming** option

dKd Dr Kim Data

Advanced Operations

- ★ Many more dataframe operations (pivot, transpose, to_dummies, …)
- ★ Many more **expressions** (abs, arg_max, floor, is_in, ...)
- ★ **String** manipulation (contains, split, replace, explode, …)
- ★ **Time series** commands (rolling_mean, group_by_dynamic, ...)
- ★ **Set** operations (union, intersection, compliment, …)
- ★ **Styling** resulting output

When should I use Polars?

✓

Large dataset on a **single machine**

✗

Small-scale exploratory analysis

Really large **cluster** computing

dKd

DR KIM DATA

Conclusion

Next Steps

Source: Yuki Kakagawa, LinkedIn

- ★ Try Polars on your data
- ★ Convert pandas to Polars
- ★ Determine best platform
- ★ Visit Polars homepage

Resources

- Polars Homepage: **pola.rs**
    - User Guide > Getting Started
    - API > Python
- YouTube
    - This Data Umbrella talk
    - Polars series on my channel: Kimberly Fessel

Stay in Touch

Kimberly Fessel

Kimberly Fessel

kimfetti

www.drkimdata.com

Kimberly Fessel, PhD

dKd
DR KIM DATA