Accelerating data analytics with RAPIDS cuDF

Nastaran Shahparian
SHARCNET | Compute Ontario| Digital Research Alliance of Canada

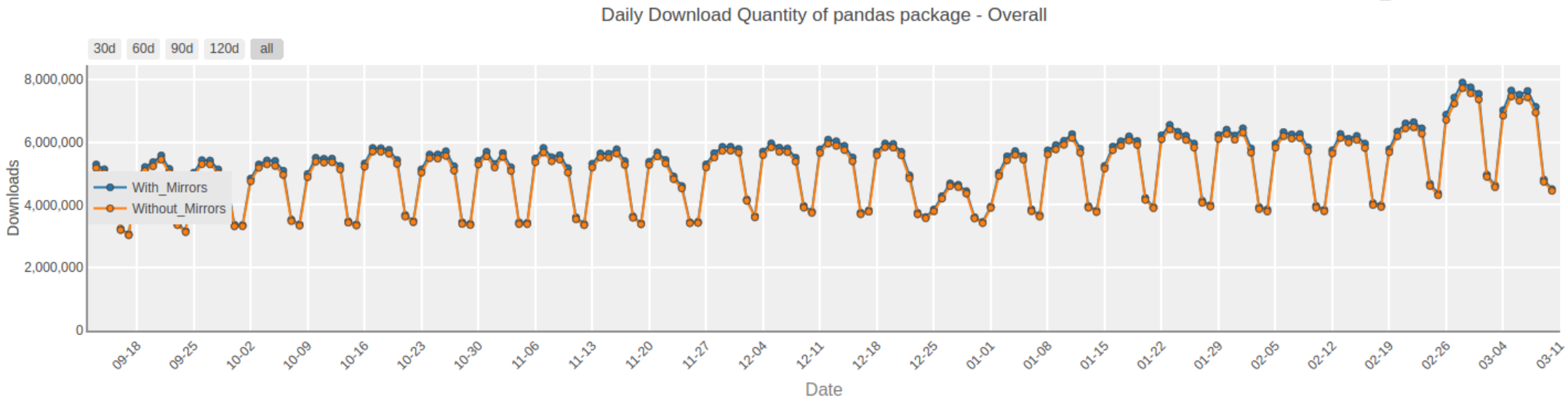
SHARCNET™

NumPy SciPy pandas

matplotlib scikit learn

Pandas: Everywhere You Look!

+176 M
monthly Downloads



https://pypistats.org/packages/pandas

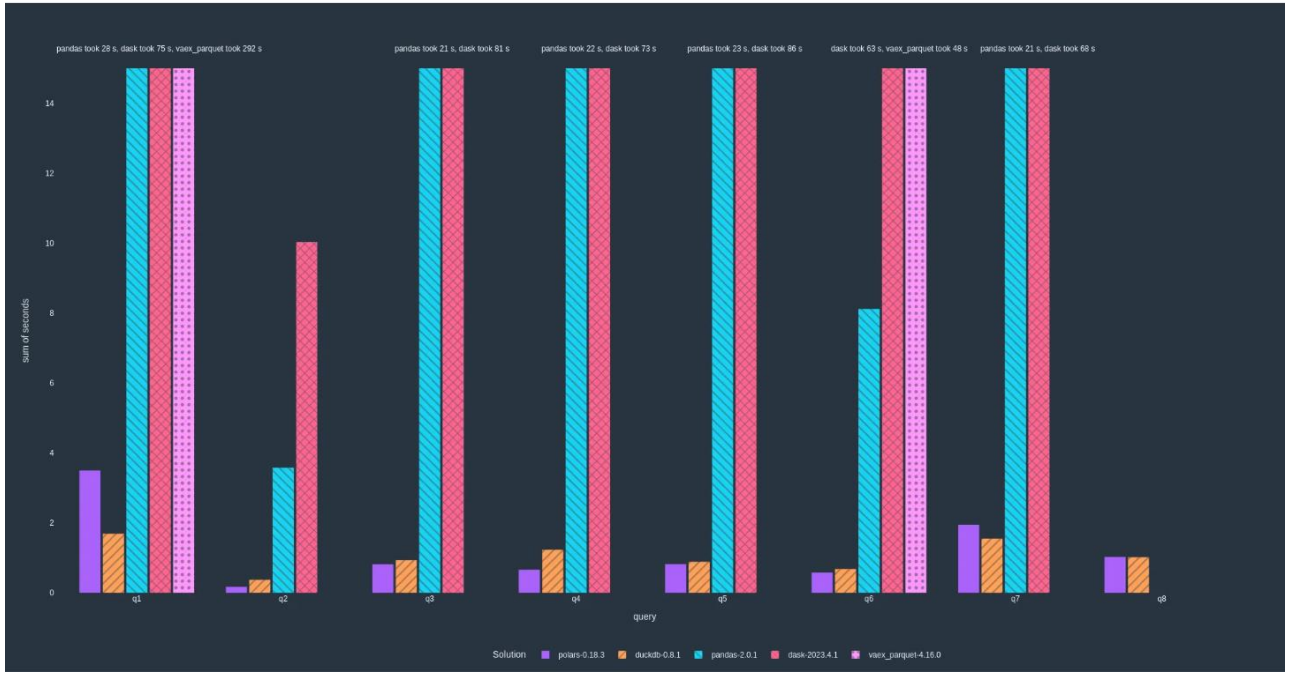
Pandas is slow for large datasets

- Single Threaded
  - Doing just one calculation at a time for a dataset
- Not a Query Language like SQL

Input table: 100,000,000 rows x 9 columns (5 GB)

DataFrames.jl	1.1.1	2021-05-15	9s
Polars	0.8.8	2021-06-30	11s
DuckDB	0.2.7	2021-06-15	14s
data.table	1.14.1	2021-06-30	15s
cuDF*	0.19.2	2021-05-31	17s
ClickHouse	21.3.2.5	2021-05-12	18s
spark	3.1.2	2021-05-31	34s
pandas	1.2.5	2021-06-30	70s
(py)datatable	1.0.0a0	2021-06-30	75s
dask	2021.04.1	2021-05-09	170s
dplyr	1.0.7	2021-06-20	175s
Arrow	4.0.1	2021-05-31	212s
Modin		see README	pending

https://h2oai.github.io/db-benchmark/



https://pola.rs/posts/benchmarks/

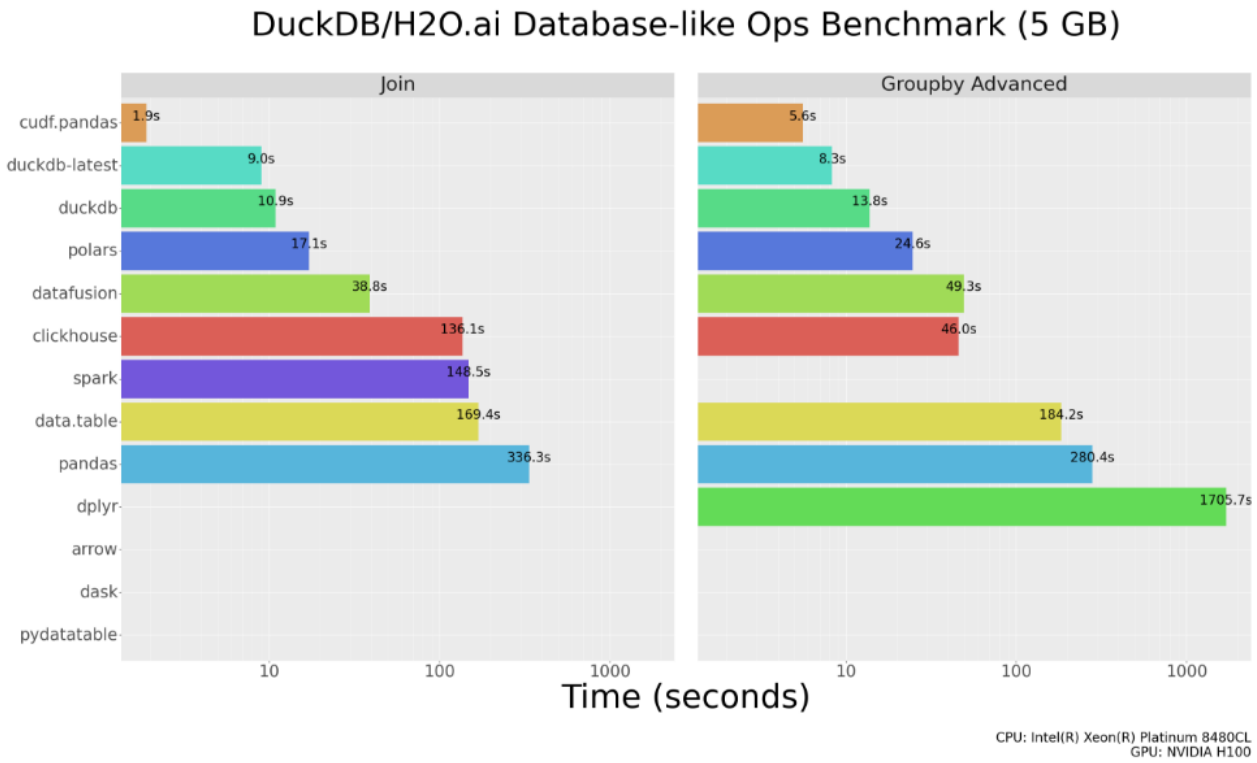
Alternatives to the Pandas

RAPIDS

- An open-source suite of GPU-accelerated AI libraries
- Developed by NVIDIA
- Provided with familiar Python Apl's
  - CuDF, dask-cuDF, cuML, cuGraph
- Helpful videos
  - https://www.youtube.com/watch?v=4xIdxwwbbic

cuDF vs. cudf.pandas

- Not all of pandas is supported
  - Only 60%
- A GPU was required for development and testing
- Required processor swapping



https://docs.rapids.ai/api/cudf/stable/cudf_pandas/benchmarks/

Using RAPIDS on the cluster

- RAPIDS Installation
  - https://docs.alliancecan.ca/wiki/RAPIDS
  - Getting the apptainer image
    - https://catalog.ngc.nvidia.com/orgs/nvidia/teams/rapidsai/containers/notebooks
    - Tag: 24.02-cuda12.0-py3.10
    - apptainer build rapids.sif docker:nvcr.io/nvidia/rapidsai/notebooks:24.02-cuda12.0-py3.10

Working interactively on a GPU node

- Request an interactive session on a compute node
  - salloc --ntasks=1 --cpus-per-task=2 --mem=16G --gres=gpu:t4:1 --time=1:0:0 --account=def-nast
- Starting the RAPIDS shell on the GPU node
  - [nast@gra1160 nast]$ module load apptainer
  - [nast@gra1160 nast]$ apptainer shell --nv -B /home -B /project -B /scratch rapids.sif
  - Apptainer> source /opt/conda/etc/profile.d/conda.sh

Working interactively on a GPU node

- launch the Jupyter Notebook serve
  - https://docs.alliancecan.ca/wiki/Advanced_Jupyter_configuration#Connecting_to_JupyterLab
  - jupyter-lab --ip $(hostname -f) --no-browser
    - http://node_name.int.cluster.computecanada.ca:8888/lab?token=101c368829...2728fad4eb
    - hostname:port token
  - Setup a SSH tunnel on a second terminal from your local computer
    - ssh -L 9999:<hostname:port> <username>@<cluster>.computecanada.ca
  - Paste this URL on a local web browser
    - http://localhost:9999/?token=<token>

Comparison between pandas and cudf.pandas

	Pandas	Cudf.pandas	Speed up
Common state violations	6.05 s	856ms	7.067
Total plates violations	1.81 s	90 ms	20.1
Total state fines	2.24 s	81.6 ms	27.4
Plates Popular violations	3.94 s	630 ms	6.25

Cudf.pandas drawbacks

- cudf.pandas is not designed for distributed or out-of-core computing workflows.

- Can’t currently interface smoothly with functions that interact with objects using a C API (such as the Python or NumPy C API)

Reference

- ✓Cudf.pandas documentation https://docs.rapids.ai/api/cudf/stable/cudf_pandas/
- ✓Cudf.pandas notebook https://colab.research.google.com/drive/12tCzP94zFG2BRduACucn5Q_OcX1TUKY3
- ✓Datacamp Blog NVIDIA Announces cuDF pandas Accelerator Mode, Nov 2023 by Richie Cotton
- ✓RAPIDS Docker Containers: https://catalog.ngc.nvidia.com/orgs/nvidia/teams/rapidsai/containers/notebooks
- ✓NVIDIA seminar Bringing Zero-Code Change Acceleration to PyData with RAPIDS cuDF and cuML