# RAPIDS

By Chaz Merritt, Murad Ali & Patrick Ihejirika

# What is RAPIDS

- A suite of open-source libraries which accelerate data science workflows using GPUs
- Enables …
  1. Significant improvements in data preparation, machine learning and graph analytics
  2. Interactive data exploration and model development
  3. Reduced costs for data processing & model training


RAPIDS

# What is RAPIDS

- Three primary libraries built on top of NVIDIA CUDA and Apache Arrow:
  - Library 1: cuDF (GPU-accelerated DataFrame library)
  - Library 2: cuML (GPU-accelerated machine learning library)
  - Library 3: cuGraph (GPU-acceleration graph analytics library)
  - And more!


RAPIDS

# RAPIDS and cuDF (Pandas but faster)

cuDF ⇒ GPU-accelerated DataFrame library

- Enables fast data manipulation & analysis on GPUs
- Provides a drop-in replacement for pandas
- **cuDF vs DASK cuDF**
  - cuDF ⇒ Used when data/workflow can be completed/saved fast enough on one GPU
  - DASK ⇒ parallel computing (Using many GPUs at once)

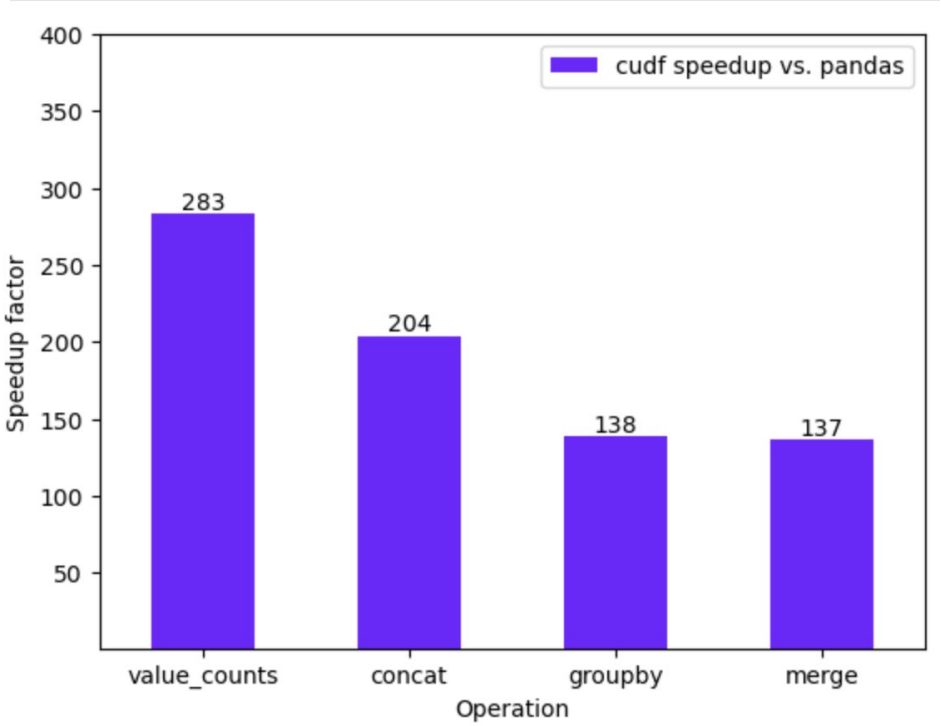
# RAPIDS' cuDF and data types

- cuDF supports the following data types: numeric, datetime, timedelta, categorical and string data types.
    - These data types are also supported by pandas and NumPY (allows for inter-function comparisons)

| Kind of data | Data type(s) |
|---|---|
| Signed integer | `'int8'`, `'int16'`, `'int32'`, `'int64'` |
| Unsigned integer | `'uint32'`, `'uint64'` |
| Floating-point | `'float32'`, `'float64'` |
| Datetime | `'datetime64[s]'`, `'datetime64[ms]'`, `'datetime64['us']`, `'datetime64[ns]'` |
| Timedelta (duration) | `'timedelta[s]'`, `'timedelta[ms]'`, `'timedelta['us']`, `'timedelta[ns]'` |
| Category | `CategoricalDtype` |
| String | `'object'` or `'string'` |
| Decimal | `Decimal32Dtype`, `Decimal64Dtype`, `Decimal128Dtype` |
| List | `ListDtype` |
| Struct | `StructDtype` |

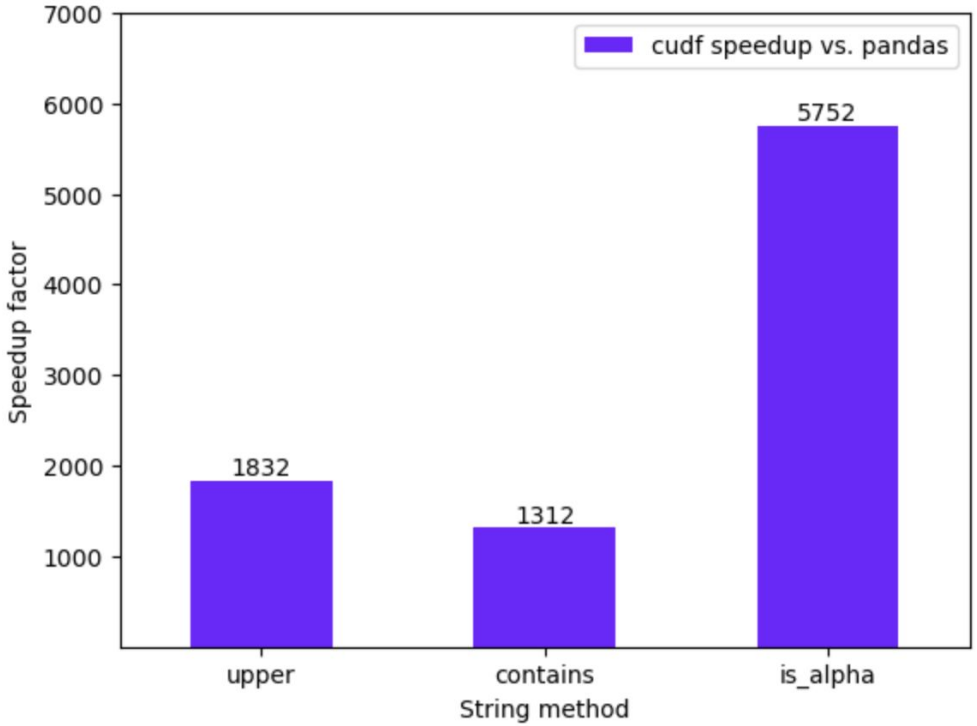
# RAPIDS' cuDF vs pandas

- Comparison in speedup factor between cuDF and pandas for dataframe operations



# RAPIDS' cuDF vs pandas

- Comparison in speedup factor between cuDF and pandas for string  operations



# RAPIDS and cuML (SciKit-Learn but Faster)

cuML ⇒ GPU-accelerated machine learning library

- Accelerates machine learning algorithms using GPU to load data and compute tasks
- Built upon Scikit Learn:
    - Includes a wide range of machine learning algorithms, including linear regression, logistic regression, dimensionality reduction and decision trees

# RAPIDS and cuML

- RAPIDS + cuML application (demonstration)

https://github.com/rapidsai/cuml/blob/branch-23.12/notebooks/kmeans_demo.ipynb

# RAPIDS and cuGraph

cuGraph ⇒ GPU-accelerated graph analytics library

- Collection of accelerated graph manipulation and analysis
    - Fast graph manipulation and analysis on GPU

# RAPIDS and cuGraph

- RAPIDS + cuGraph application (demonstration)

https://github.com/rapidsai/cugraph/blob/branch-23.08/notebooks/demo/mg_pagerank.ipynb

# RAPIDS v23.10

This new version of RAPIDS brings accelerated computing to pandas workflows with no code changes using pandas accelerator mode (which is currently in beta version)

By adding just one line of code, your existing pandas code can incorporate the speedups from RAPIDS.

https://colab.research.google.com/drive/12tCzP94zFG2BRduACucn5Q_OcX1TUKY3?usp=sharing

# Importance of RAPIDS v23.10:

Adopting cuDF previously required workarounds in three main areas:

1. Working around pandas functionality that was not yet implemented in cuDF
2. Designing separate code paths for CPU and GPU execution
3. Manually switching between cuDF and pandas when interacting with other PyData libraries

RAPIDS v23.10 resolves this by **enabling a unified GPU/CPU in pandas accelerator** which executes some operations on GPU where possible and on the CPU otherwise and **enabling zero code change acceleration:** To enable GPU acceleration for pandas operations, you can either install the cuDF Jupyter Notebook extension or import the cuDF Python module

# RAPIDS use cases

**Entreprise:**

Bumble

AT&T

Amazon

WalMart

**Research:**

KGMON

# RAPIDS use cases (Enterprise)

**Bumble:**

Buzzwords is Bumble's open-source GPU-powered topic modelling tool, it was developed in house and builds upon the work found in BERTopic and Top2Vec. Buzzwords uses cuML's algorithms

**AT&T:**

AT&T uses RAPIDS ETL to allow for faster analysis by moving data science workloads to GPU

# RAPIDS use cases (Enterprise)

**Amazon:**

RAPIDS is enabling work on the very cutting edge by enabling Graph Neural Networks (GNN) for applications including drug discovery, recommender systems, fraud detection, and cybersecurity

**Walmart:**

Walmart implemented cuML's XGBoost to increase forecast accuracy and potentially save billions of dollars

# RAPIDS use cases (Research)

**KGMON:**

Kaggle Grandmasters of NVIDIA (KGMoN) used RAPIDS to build winning recommender systems, predict degradation rates in RNA molecules, identify melanoma in medical imaging, and more

# References

rapids.ai