UDACITY Google spark

Palestine Launchpad Program

Data Analysis NanoDegree

Our agenda for today

1. Congratulation Students who passed
2. Recap about the last session
3. Project (Three) Overview
4. Case Study
5. New Concepts (Fundamentals of accelerating Data)
6. Knowledge Sharing (Text-Processing)
7. Q\A

Palestine Launchpad
منصة فلسطين التقنية
Google

Recap

What did we talk about last session

UDACITY

Last Session we covered the following:

Here's a breakdown of the key areas we covered:

📊 **Project 2 Statistics:** We discussed the students' progress and results regarding Project 2.

📱 **Exploratory Vs Explanatory Data Analysis:** differentiate between the two the terms.

📘 **Introduction to Project 3:** We provided an overview and objectives of Project 3.

📝 **Text Processing Pipeline:** We walked through the steps involved in processing and analyzing text data.

🔍 **Case Study:** We analyzed a dataset from a firewall and applied exploratory analysis using univariate analysis techniques.

Palestine Launchpad
منصة فلسطين التقنية
Google

Project Three : Communicate Data Findings

What is needed to make a successful submit

UDACITY

Project: Communicate Data Findings

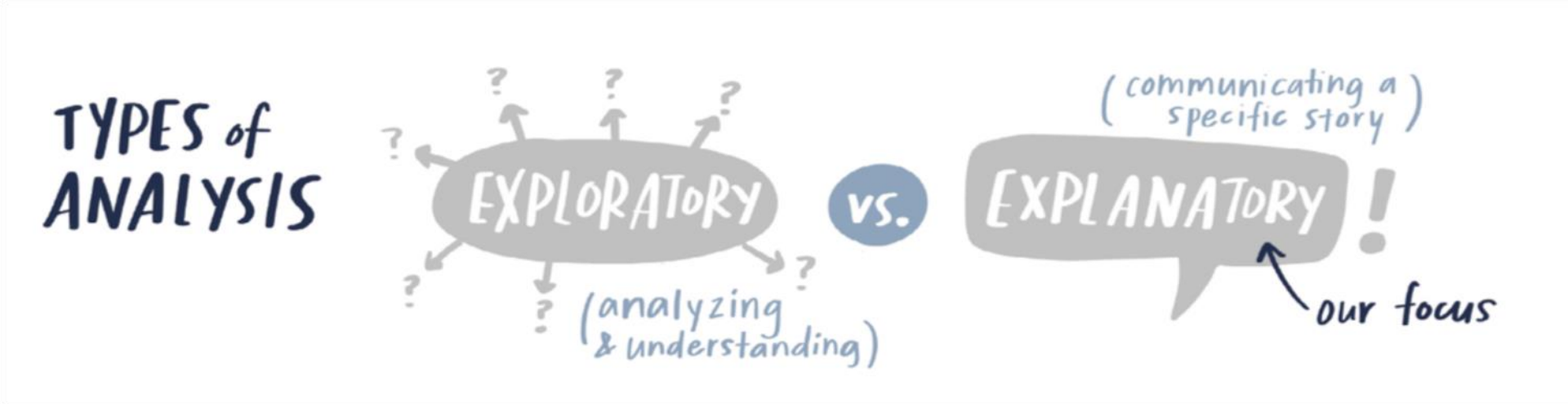
Project Overview
This project has two parts that demonstrate the importance and value of data visualization techniques in the data analysis process.

- ✓ **In Part I, Exploratory data visualization**, you will use Python visualization libraries to systematically explore a selected dataset, starting from plots of single variables and building up to plots of multiple variables.

- ✓ **In Part II, Explanatory data visualization**, you will produce a short presentation that illustrates interesting properties, trends, and relationships that you discovered in your selected dataset. The primary method of conveying your findings will be through transforming your exploratory visualizations from the first part into polished, explanatory visualizations.

- ✓ **Project Due Date: Aug 27, 2024**

Project: Communicate Data Findings



Choose Your Dataset

Below is a compiled list of the datasets you can choose from. However, you can explore other datasets that interest you:

- ❑ **Ford GoBike System Data(opens in a new tab)** (38 MB, CSV File)
- ❑ **Flights**
- ❑ **Loan Data from Prosper(opens in a new tab)** (82.5 MB, CSV File)
- ❑ **PISA Data - 2018**

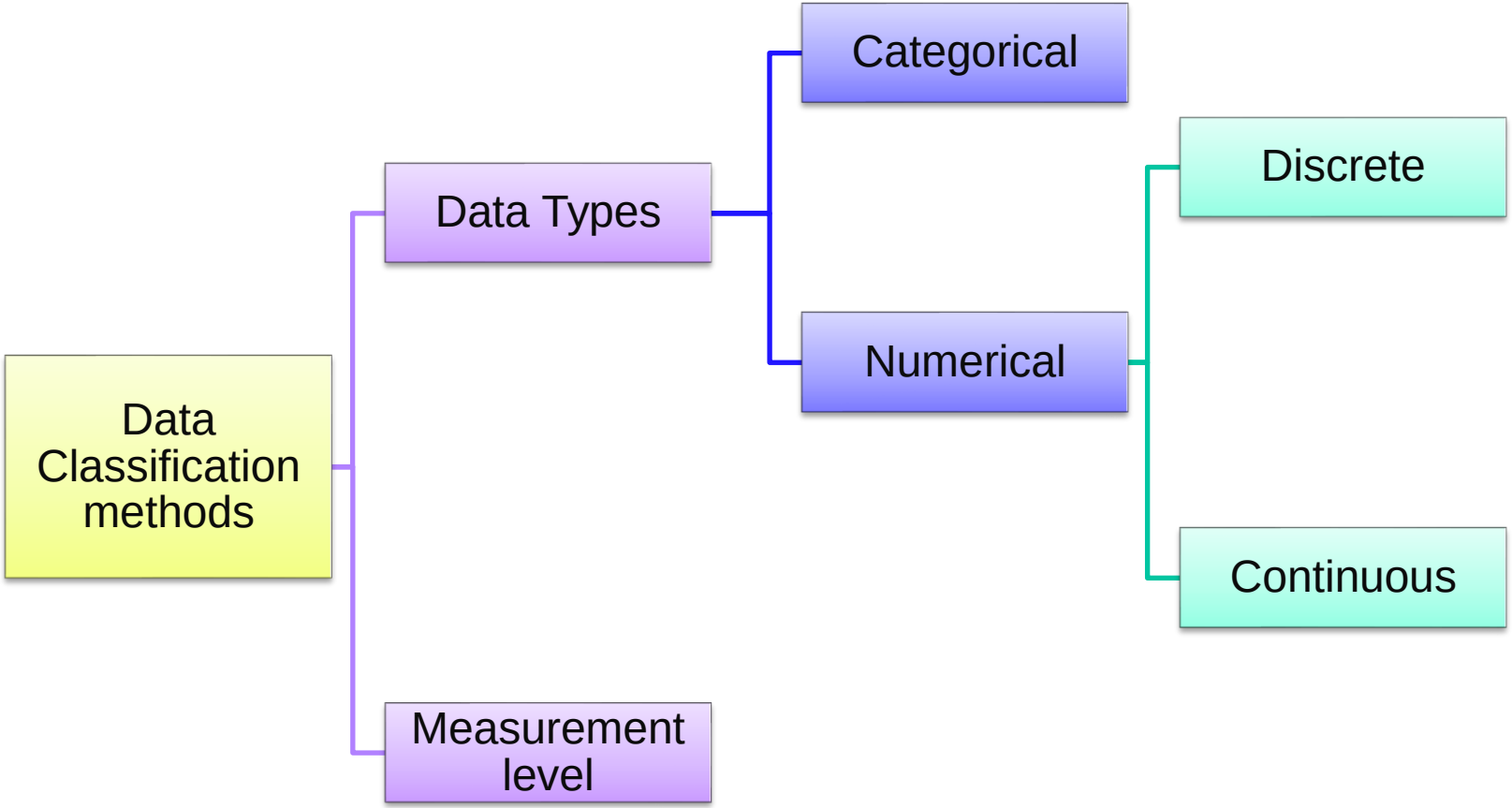
Pareto Charts & 80-20 Rule - Clinical Excellence Commission (nsw.gov.au)

Palestine Launchpad
منصة فلسطين التقنية
Google

Case Study

Viulize a log database

UDACITY


Data Classification methods
Data Types
Categorical
Numerical
Discrete
Continuous
Measurement level

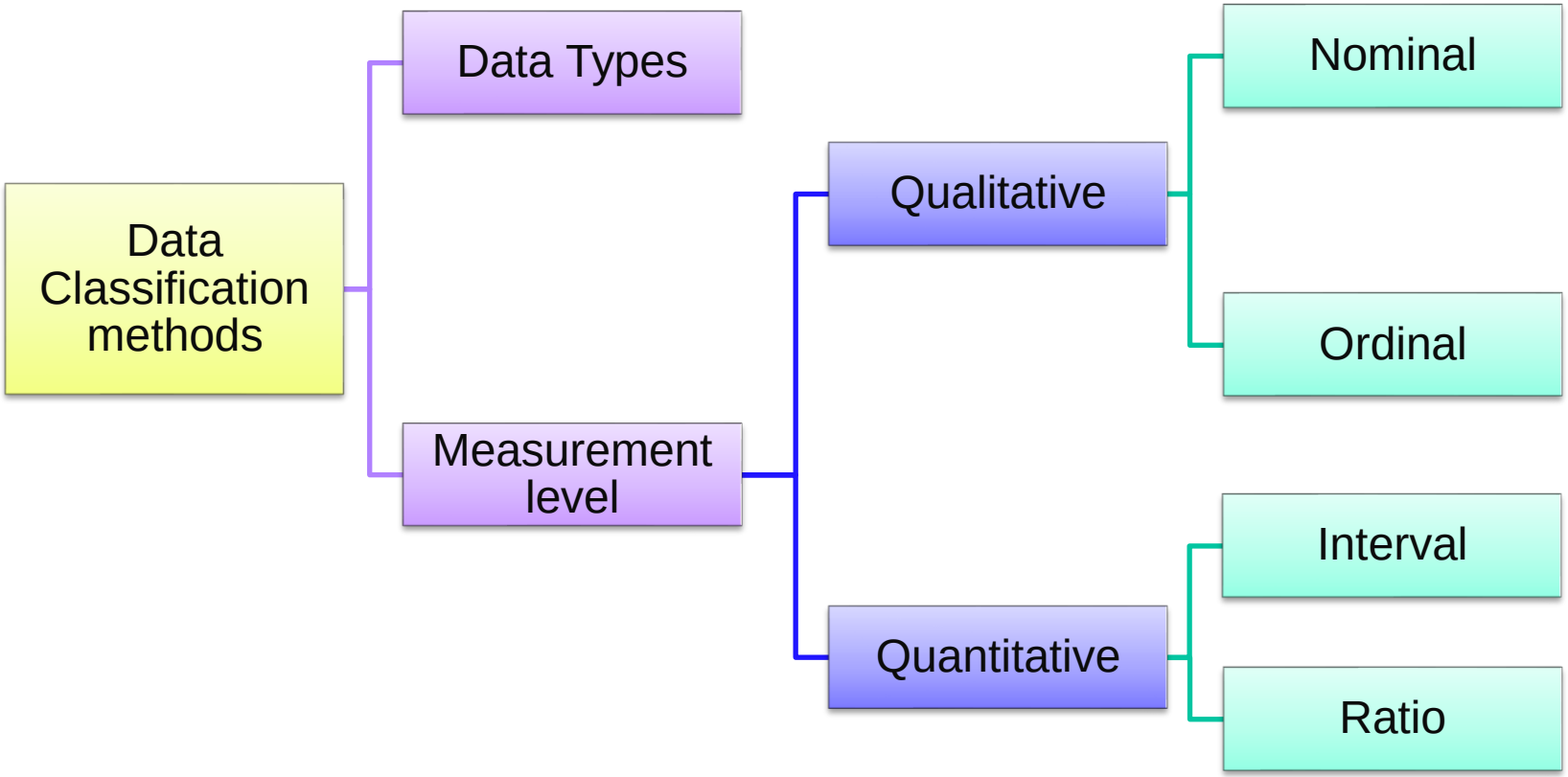
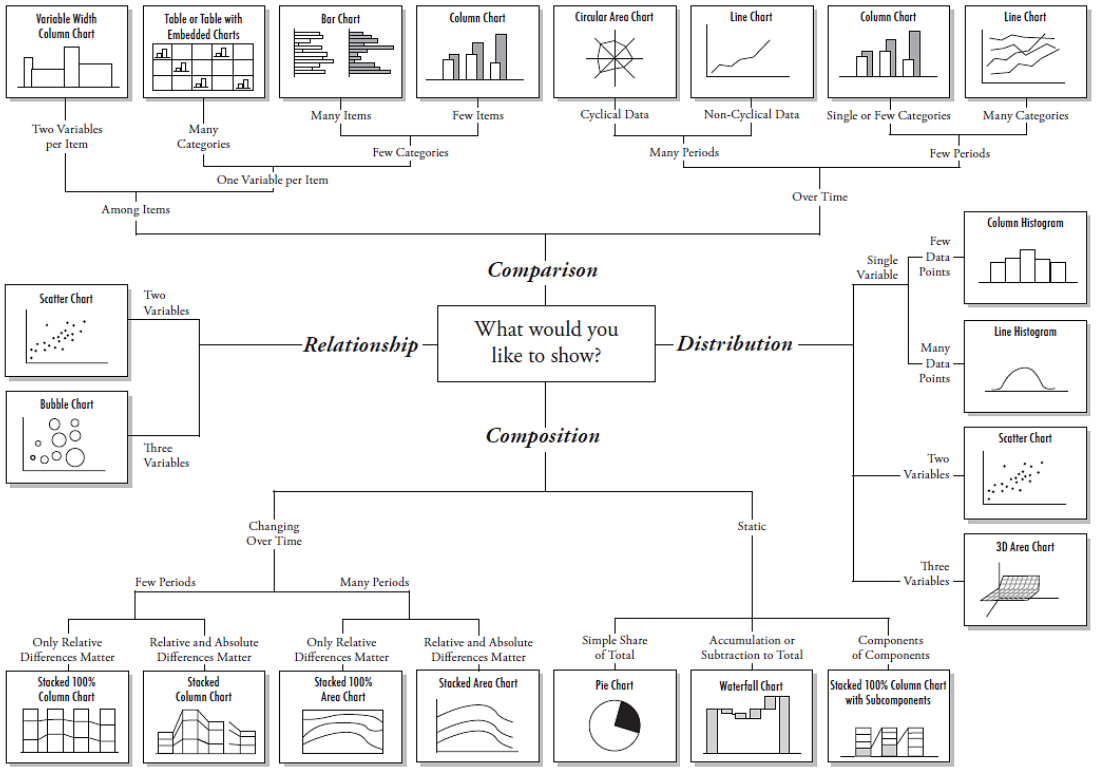
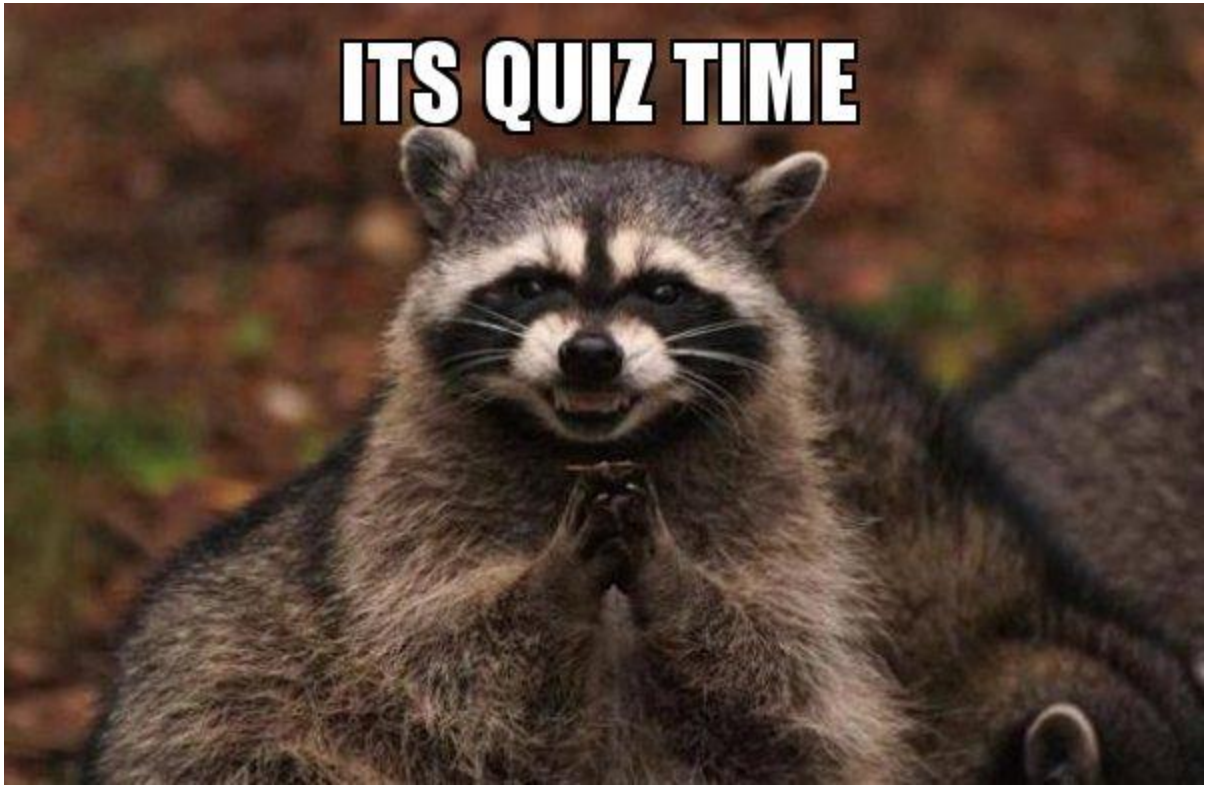

Data Classification methods
Data Types
Measurement level
Qualitative
Quantitative
Nominal
Ordinal
Interval
Ratio

Chart Suggestions—A Thought-Starter





- Data Visualization | Quizizz

Visualizing Data (Categorial Data Types) :

1. Frequency distribution table (Table with frequencies)
2. Bar Charts
3. Pie Charts (Percentage (Share of the total))
4. Pareto Diagram (Special Bar Chart)

https://colab.research.google.com/drive/1rNcwrSXRCnnEuRfVjzzd5HrL1L1izboc?usp=sharing

Palestine Launchpad
منصة فلسطين التقنية
Google

New Concepts

Udacity, Students and Session Leads

UDACITY

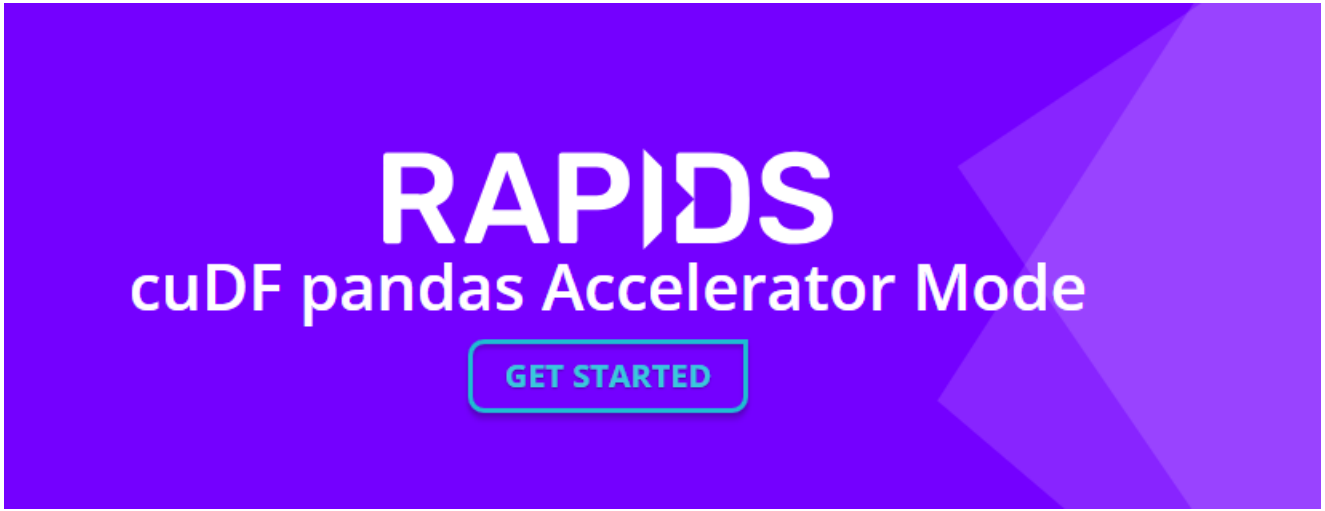
WHAT IS RAPIDS

AN ECOSYSTEM OF HARDWARE, SOFTWARE, AND DEVELOPERS

RAPIDS is a collection of open-source software libraries and APIs that gives you the ability to execute end-to-end data science and analytics pipelines entirely on NVIDIA GPUs using familiar PyData APIs

WHAT IS RAPIDS

- Dataframe processing with **cuDF** (similar API to pandas)
- Machine learning with **cuML** (similar API to scikit-learn)
- Graph processing with **cuGraph** (similar API to networkX)
- Spatial analytics with **cuSpatial** (similar API to geoPandas)
- Image processing with **cuCIM** (similar API to scikit-image)
- Seamless cross-filtered dashboards with **cuxfilter**
- Low level compute primitives with **RAFT**
- Apache Spark acceleration with **Spark RAPIDS**



https://colab.research.google.com/drive/1O7NYEbABJ_IVoFwWfvvRADeM6Kxf9f-J?usp=sharing

RAPIDS
cuDF pandas Accelerator Mode
GET STARTED

ACCELERATING PANDAS WITH ZERO CODE CHANGE

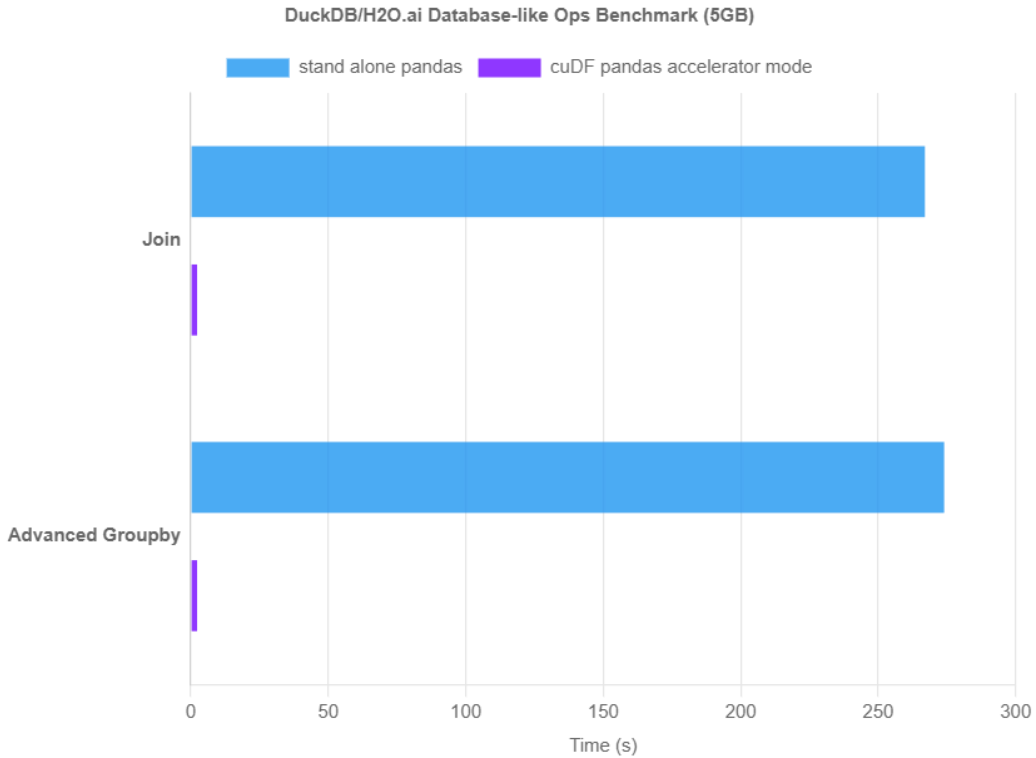
150X FASTER, ZERO CODE CHANGE

https://youtu.be/kmj1QOY71Ps

RAPIDS
cuDF pandas Accelerator Mode
GET STARTED

ACCELERATING PANDAS WITH ZERO CODE CHANGE

150X FASTER, ZERO CODE CHANGE



Standard DuckDB Data Benchmark (5GB)
GPU: NVIDIA Grace Hopper, CPU: Intel® Xeon® Platinum 8480C
pandas v2.2, RAPIDS cuDF v23.10

BRINGING THE SPEED OF CUDF TO EVERY PANDAS USER

HOW TO USE IT?

1. To accelerate IPython or Jupyter Notebooks, use the magic

```
%load_ext cudf.pandas
import pandas as pd
...
```

2. To accelerate a Python script, use the Python module flag on the command line:

```
python -m cudf.pandas script.py
```

BRINGING THE SPEED OF CUDF TO EVERY PANDAS USER

HOW TO USE IT?

Or, explicitly enable cudf.pandas via import if you can't use command line flags:

```
import cudf.pandas
cudf.pandas.install()

import pandas as pd
...
```

DEMO

https://colab.research.google.com/drive/1O7NYEbABJ_IVoFwWfvvRADeM6Kxf9f-J?usp=sharing

cuDF Pandas | RAPIDS | GPU Accelerated Data Science

Firewall Log dataset

https://colab.research.google.com/drive/1m_Q12L-CDGunKSsHI9Nrrh_m7Jx3igDR?usp=sharing

Palestine Launchpad
منصة فلسطين التقنية
Google

Knowledge Sahring

Udacity, Students and Session Leads

UDACITY

Text processing pipeline by Sondos

Palestine Launchpad
منصة فلسطين التقنية
Google

Next Steps

Let's get things started...

UDACITY

Second Project July 2, 2024

May | June | July | August | September

Strat working on Project three

1. Choose the dataset
2. Start Exploring your dataset

Today
July 6, 2024

Third Project Due Date

August 27, 2024
- Communicate Data findings

Program Period
May 9, 2024 - September 14, 2024

Palestine Launchpad
منصة فلسطين التقنية
Google

Thank You!

And Good Luck!

UDACITY