Introduction to RAPIDS
DATA SCIENCE WORKFLOW WITH RAPIDS
Open Source, GPU-accelerated ML Built On CUDA

Data preparation / wrangling

cuDF

ML model training

cuML

Dataset exploration

VISUALIZE

PREDICTIONS
WHAT IS RAPIDS?
The New GPU Data Science Pipeline

rapids.ai
Suit of open-source, end-to-end data science tools
Built on CUDA
Pandas-like API for data cleaning and transformation
Scikit-learn-like API
A unifying framework for GPU data science
“CLASSIC” MACHINE LEARNING
The daily work of most data scientists

Comprehensible to average data scientists and analysts

Higher level of interpretability

Solutions for unlabeled data

Techniques such as regression and decision trees, clustering

Scikit-learn
RAPIDS ROADMAP

DATA ANALYTICS

- **IO**
  - Data Formats (CSV, ORC, PARQUET, JSON)
  - Data Sources (Cloud, HDFS)
  - Data Types (INT64, FP64, STRINGS)

- **Operators**
  - joins
  - GroupBys
  - Windowing
  - Strings
  - UDFs

- **CuML Library**
  - dimension reduction
  - Clustering
  - Time Series
  - Preprocessing

MACHINE LEARNING

- **Classification**
  - GBDT
  - Logistic
  - Linear
  - Random Forest

- **Regression**
  - GBDT
  - Ridge
  - Lasso

- **Dimension Reduction**
  - UMAP
  - PCA
  - T-SNE

- **Clustering**
  - K-Means
  - DBSCAN
  - Kalman Filtering

- **Time Series**
  - Holt-Winters
  - ARIMA

- **Preprocessing**
  - Similarity
  - Weighted Jaccard
  - Triangle Counting

- **CuDF Library**
  - Up to 5-15X speedup
  - Up to 10-20X speedup
  - Up to 100-500X speedup

GRAPH ANALYSIS

- **CuGRAPH Library**
  - Centrality
  - Page Rank
  - Single Shortest Path
  - Louvain Clustering
  - Subgraph Extraction
  - Triangle Counting
  - Weighted Jaccard
  - Jaccard Similarity

- **Path Finding**
  - Breadth-First Search
  - Depth First Search
  - Spectral Clustering

- **Similarity**
  - KNN
  - K-Means
  - DBSCAN

- **Community Detection**
  - Spectral Clustering
  - Louvain Clustering

- **Page Rank**
  - Single Shortest Path
  - Breadth-First Search
  - Depth First Search

- **Similarity**
  - Weighted Jaccard
  - Jaccard Similarity
RAPIDS PREREQUISITES
See more at rapids.ai

- NVIDIA Pascal™ GPU architecture or better
- CUDA 9.2 or 10.0 compatible NVIDIA driver
- Ubuntu 16.04 or 18.04
- Docker CE v18+
- nvidia-docker v2+
Get RAPIDS

RAPIDS is available as conda or pip packages, docker images, and from source builds. Use the tool below to select your preferred method, packages, and environment to install RAPIDS. Certain combinations may not be possible and are dimmed automatically. Be sure to review the prerequisites section for more details about requirements to use RAPIDS.

<table>
<thead>
<tr>
<th>Method</th>
<th>Preferred</th>
<th>Beta</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conda</td>
<td>Docker</td>
<td>Pip</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Packages</td>
<td>cuDF</td>
<td>cuML</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linux</td>
<td>Ubuntu 16.04</td>
<td>Ubuntu 18.04</td>
</tr>
<tr>
<td>Python</td>
<td>Python 3.6</td>
<td></td>
</tr>
<tr>
<td>CUDA</td>
<td>CUDA 9.2</td>
<td></td>
</tr>
<tr>
<td>Command</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
|            | `conda install -c nvidia -c rapidsai -c pytorch -c numba -c conda-forge \`  
            |            |        | `-c defaults cudf=0.5 cuml=0.5 python=3.6` |

COPY COMMAND  

CONDA DETAILS
GETTING STARTED RESOURCES

Rapids.ai


Github: https://github.com/RAPIDSai

Twitter: @rapidsai
RMM Memory Pool Allocation

[https://github.com/rapidsai/rmm](https://github.com/rapidsai/rmm)

Use large `cudaMalloc` allocation as **memory pool**

Custom memory management in **pool**

Streams enable asynchronous `malloc/free`

RMM currently uses `CNMem` as it’s Sub-allocator

[https://github.com/NVIDIA/cnmem](https://github.com/NVIDIA/cnmem)

RMM is standalone and free to use in your own projects!
cuML architecture

- Python
- Cython
- cuML Algorithms
- cuML Prims
- CUDA Libraries
- CUDA
Let’s Dive into the Tutorial!
Getting GCP Set Up

Get GCP IP address

ssh pydata@[IP]
Password: gtc2019

conda activate rapids

Get the data: `wget -v -O black_friday.zip -L https://goo.gl/3EYV8r` (if you don’t have `wget`, you can install it on mac via homebrew)


jupyter-notebook --allow-root --ip=0.0.0.0 --port 8888 --no-browser --NotebookApp.token="
THANK YOU TO GOOGLE CLOUD PLATFORM
Kubeflow also has a RAPIDS container!

Google kindly donated the instances for this tutorial at GTC SJ 2019!