TechU

IBM

Explore GPU accelerated Data Preparation and Machine Learning with Nvidia RAPIDS

Dustin VanStee - Data Scientist
Loic Fura - Power System Specialist

IBM Worldwide Client Experience Centers

2019 IBM Systems Technical University
Oct 7-11 | Las Vegas, NV

IBM Systems Worldwide Client Experience Centers

IBM Systems Worldwide Client Experience Centers *maximize IBM Systems competitive advantage in the Cloud and Cognitive era by providing access to world class technical experts and infrastructure services to assist Clients with the transformation of their IT implementations..*

9 Worldwide Locations (* also Infrastructure Hubs):

Austin TX , *Poughkeepsie NY, Rochester MN, Tucson AZ, *Beijing CHINA, Boeblingen GERMANY, Guadalajara MEXICO,*Montpellier FRANCE, Tokyo JAPAN

Client Experience	Architecture & Design	Infrastructure Solutions	Content
Tailored, in-depth technology Innovation Exchange Events Relationship building Demonstrations Meetups Solution workshops Remote options	Advise clients, "Art of the Possible" Discovery & Design Workshops, Consulting, Showcases, Reference Architectures, Co-Creation of assets	Benchmarks, MVP & Proof of Technology "Test Drives" Demonstrations Infrastructure Services Certify ISV solutions Hosting Cloud Environment	Content Development IBM Redbooks Training Courses Video courses "Test Drives" Demonstrations
(Inbound & Outbound)	(Inbound & Outbound)	(Inbound to Centers)	

NEW: *Co-Creation Lab; CEC Cloud; RedHat Center of Competency*

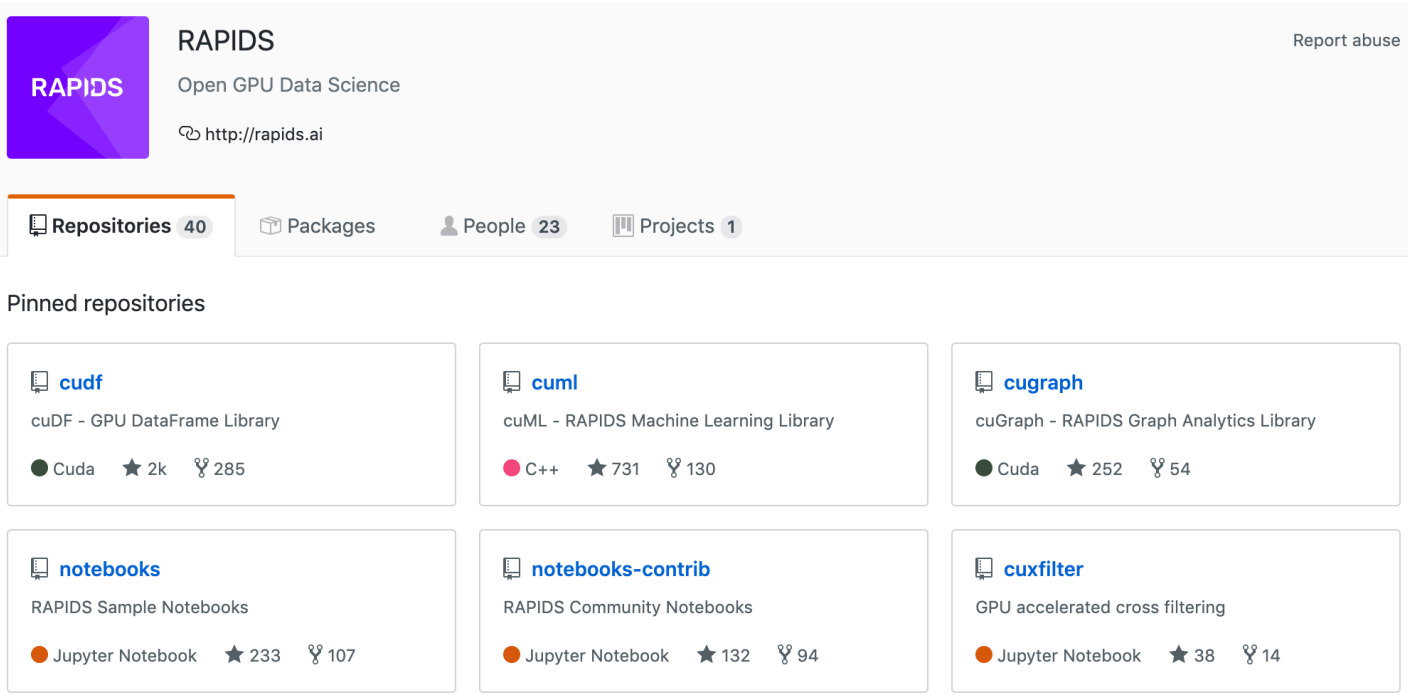
For further information, please contact the Centers via email at:
ccenter@us.ibm.com

Session Objectives

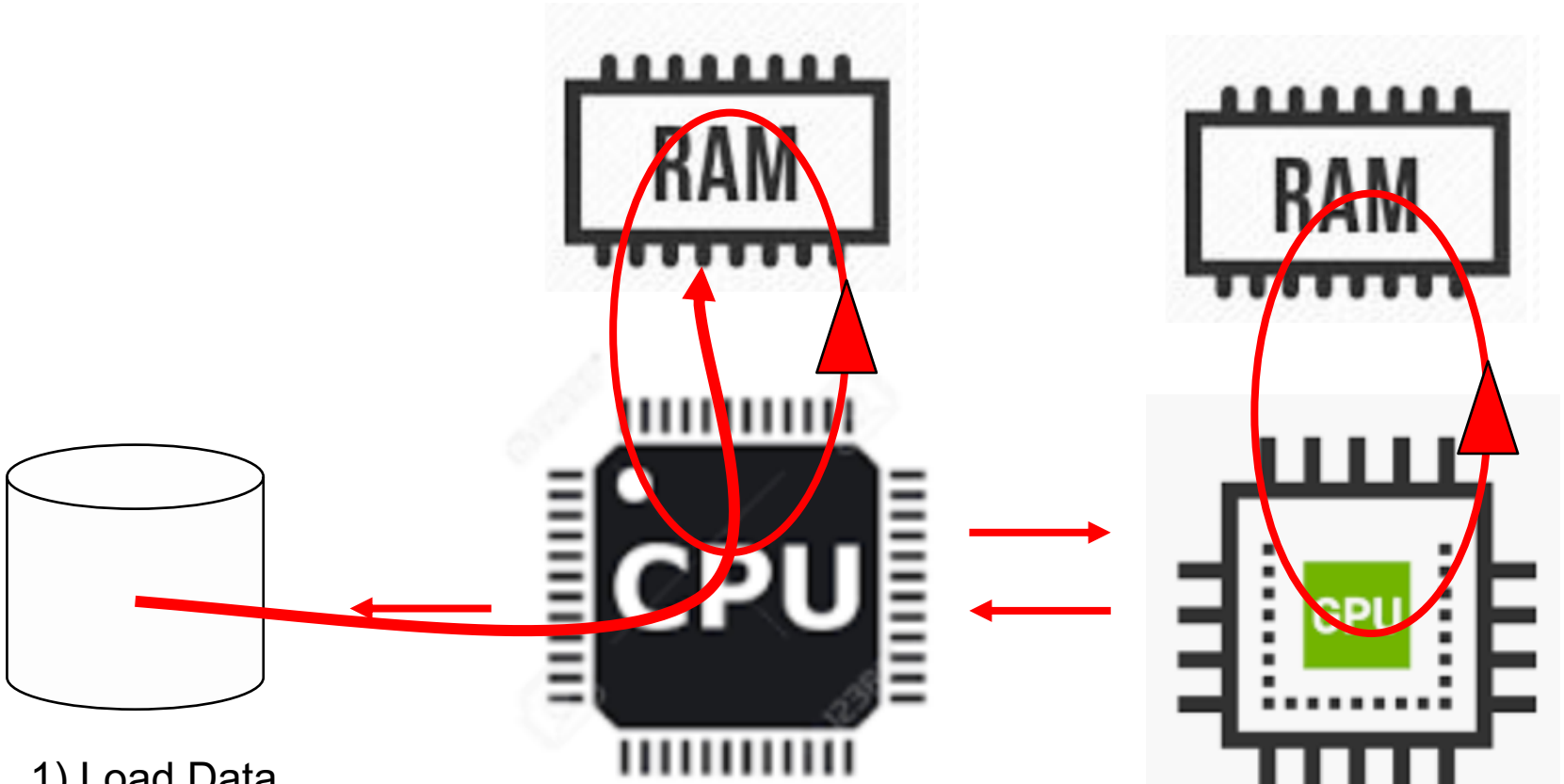
- What is RAPIDS
- Motivation for RAPIDS
- RAPIDS architecture
- How can I get RAPIDS on AC922 ?
- Quick Lab Demo to get started

What is RAPIDS

- RAPIDS is a software package from Nvidia
- Accelerates **Data Preparation** and **Machine Learning**
- Eliminates a lot of CPU to GPU transfers
- Meant to be used in place of Pandas and Sklearn python libraries
- Provides 10x-50x performance boost for many tasks
- Scales to multiple GPUs
- Designed to integrate with DL frameworks too



Typical CPU Based Workflow

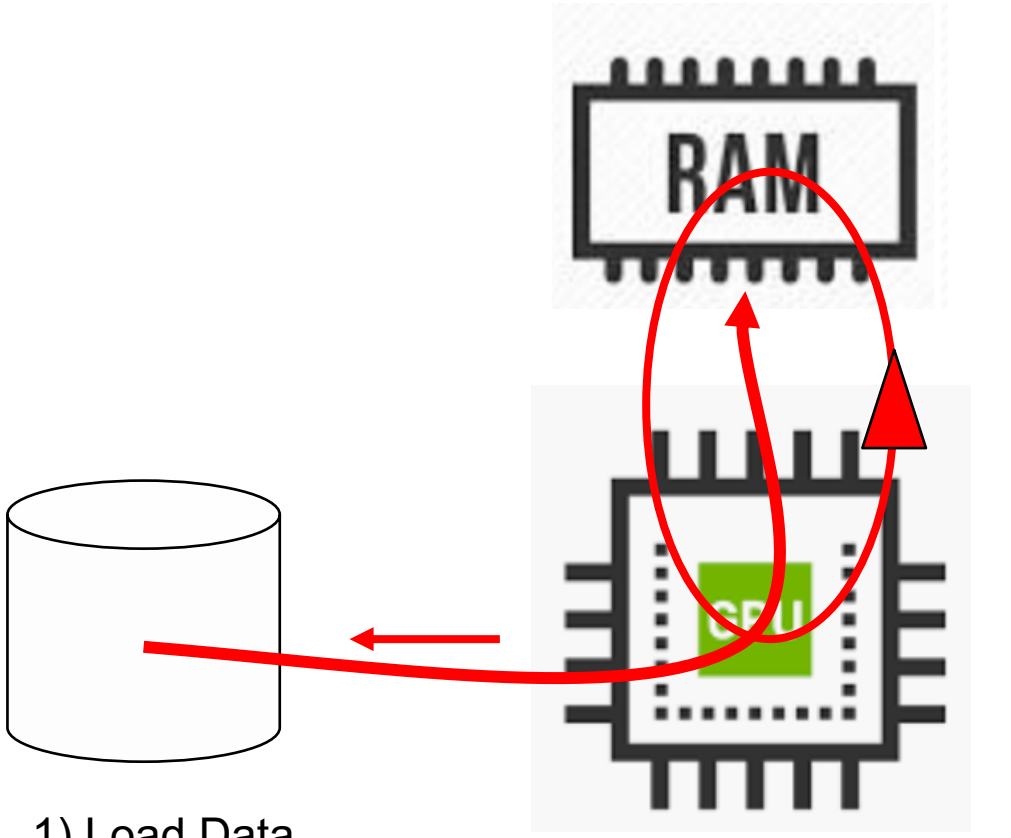


1) Load Data from Disk

2) Data Preparation Pandas

3) Machine Learning SkLearn

RAPIDS Based Workflow



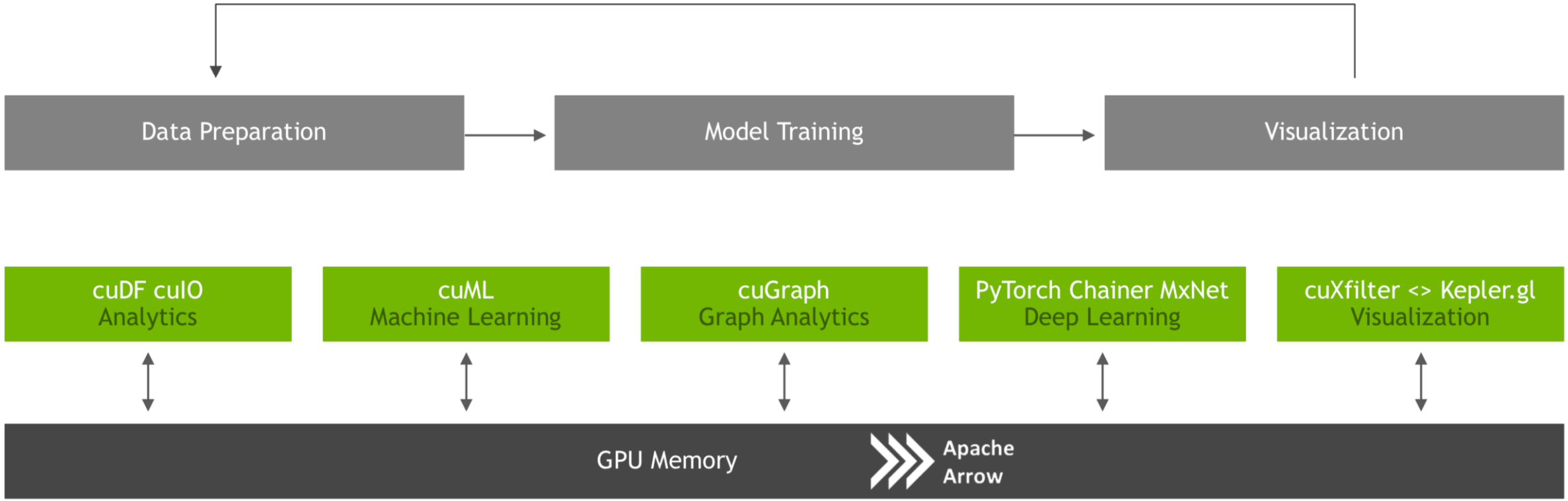
1) Load Data from Disk

2) Data Preparation Pandas

3) Machine Learning SkLearn

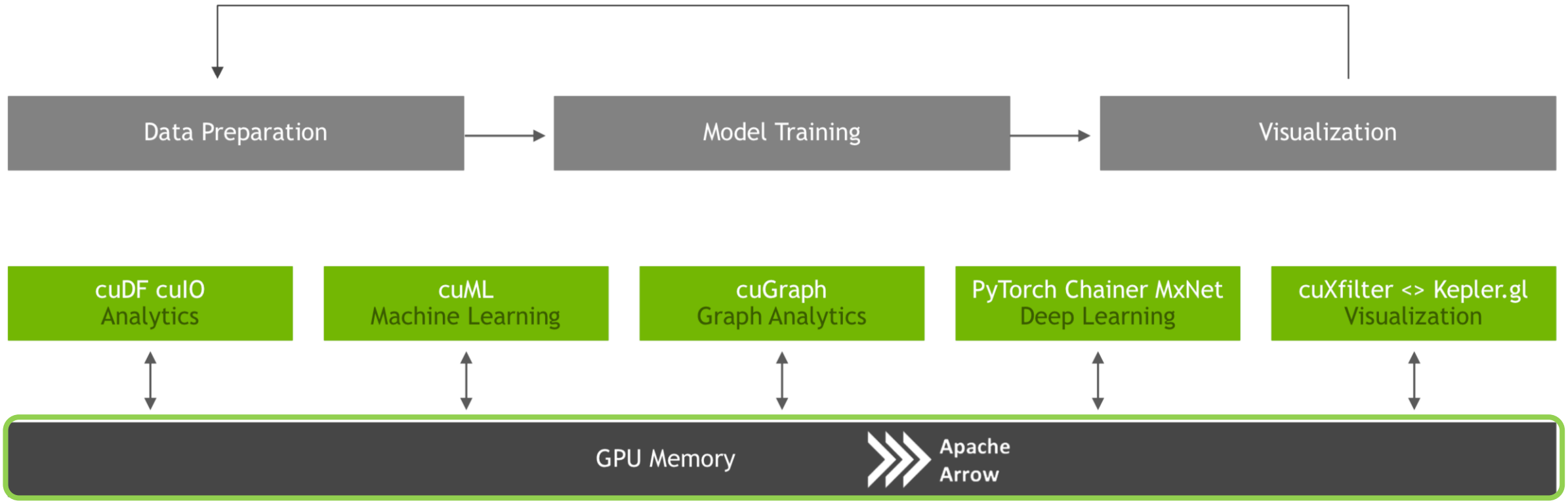
RAPIDS

End to End Accelerate GPU Data Science

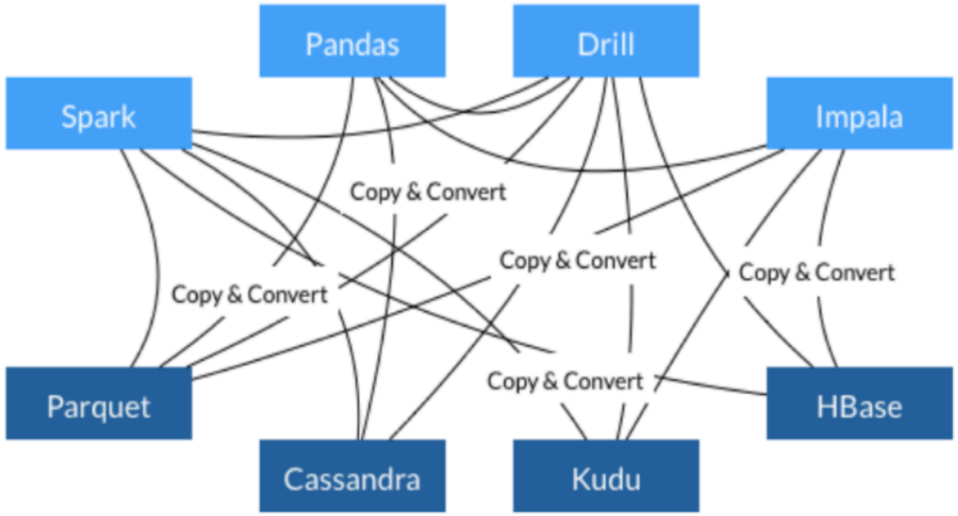


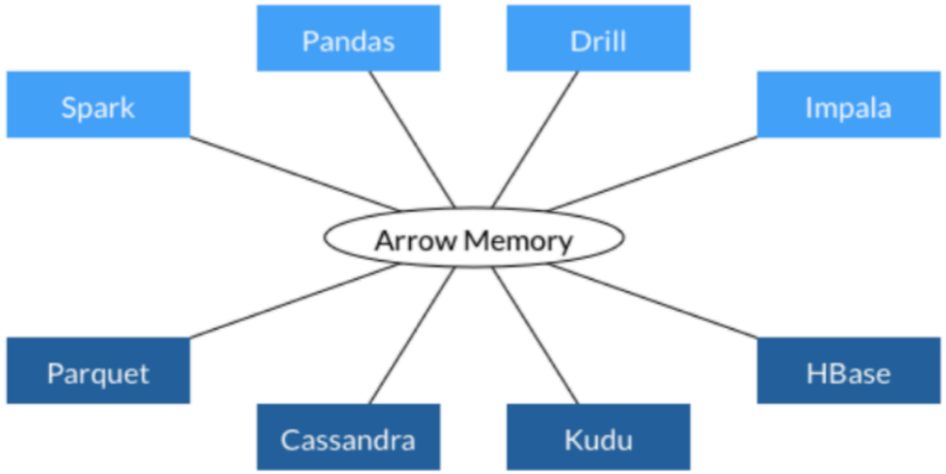
RAPIDS

End to End Accelerate GPU Data Science



LEARNING FROM APACHE ARROW





- Each system has its own internal memory format
- 70-80% computation wasted on serialization and deserialization
- Similar functionality implemented in multiple projects

- All systems utilize the same memory format
- No overhead for cross-system communication
- Projects can share functionality (eg, Parquet-to-Arrow reader)

From Apache Arrow Home Page - https://arrow.apache.org/

Data Preparation – Speeding up the 80%

cuDF is comprised of

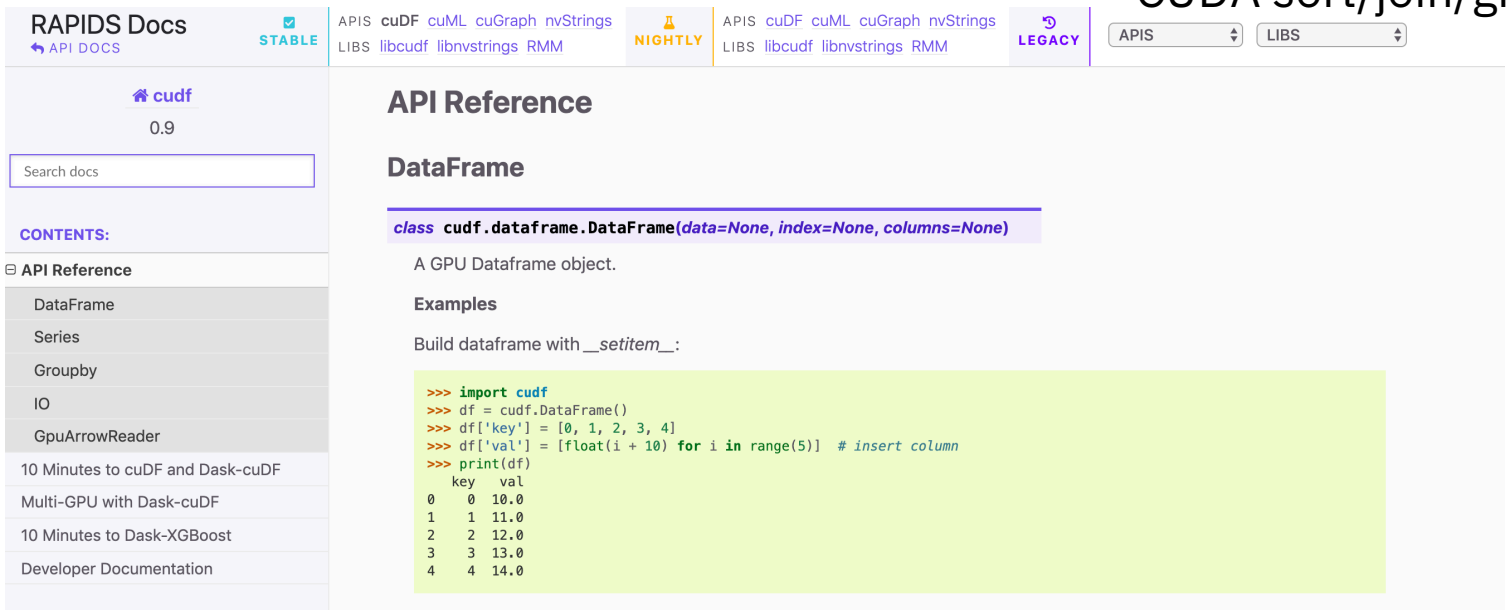
Python

- Python API for data prep
- Mirror Pandas API
- Creates Apache Arrow data frames
- JIT completion for UDFs via Numba

+

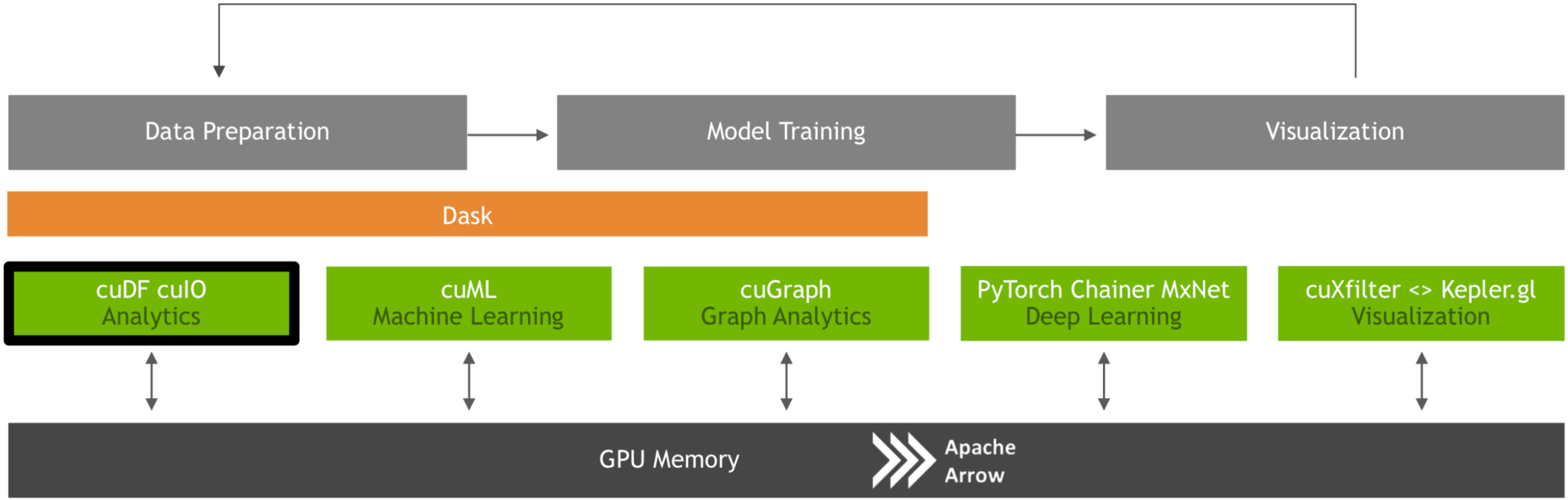
CUDA

- Low level implementations for data prep on GPU
- Implements functions for data movement using Apache Arrow
- CUDA kernels to perform element wise opps
- CUDA sort/join/groupby/reduce



ETL - THE BACKBONE OF DATA SCIENCE

cuDF is not the end of the story



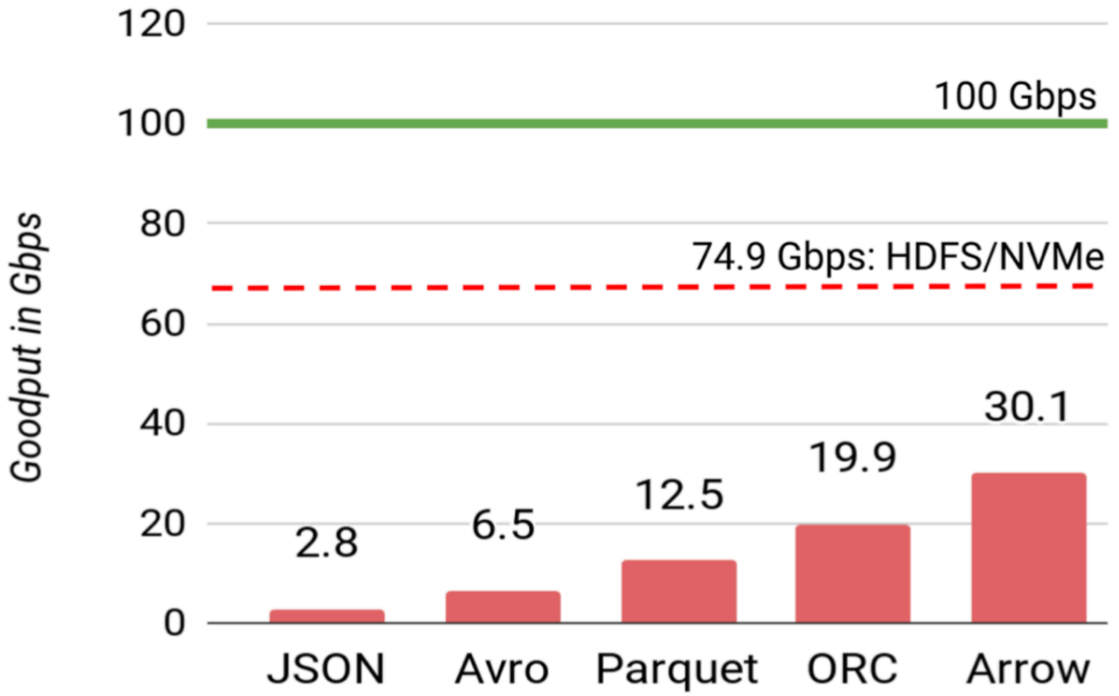
Scale out RAPIDS with DASK!

Scale out RAPIDS with DASK!

- **PyData Native**
  - Built on top of NumPy, Pandas Scikit-Learn, … (easy to migrate)
  - With the same APIs (easy to train)
  - With the same developer community (well trusted)
- **Scales**
  - Easy to install and use on a laptop
  - Scales out to thousand-node clusters
- **Popular**
  - Most common parallelism framework today at PyData and SciPy conferences
- **Deployable**
  - HPC: SLURM, PBS, LSF, SGE
  - Cloud: Kubernetes
  - Hadoop/Spark: Yarn

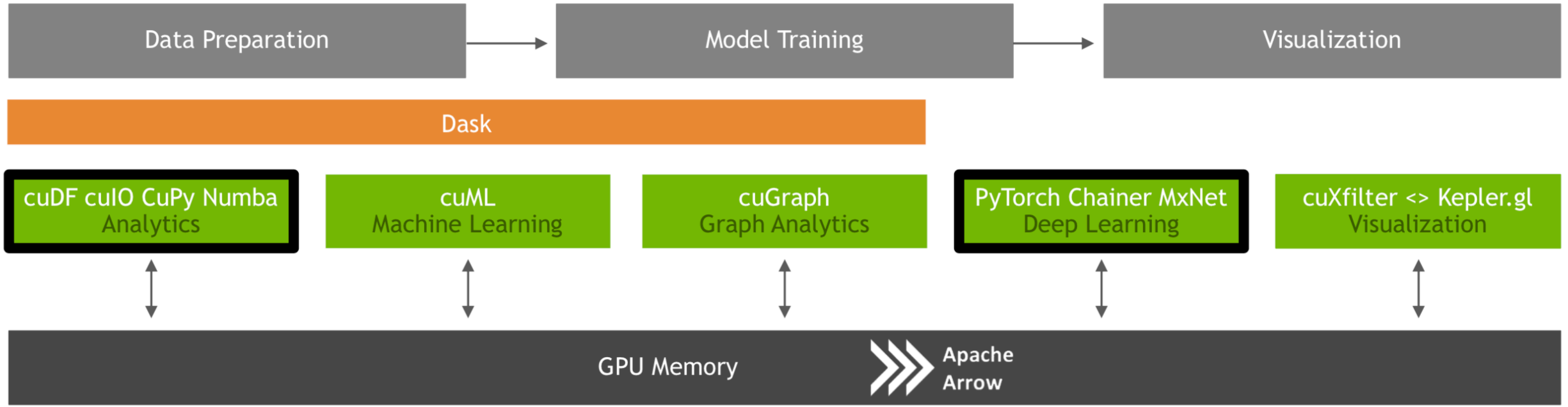
Data Loading is FAST

- CSV Reader
  - Follows API of pandas.read_csv
  - Current implementation is >10x speed improvement over pandas
- Parquet Reader – v0.7
  - Work in progress: Will follow API of pandas.read_parquet
- ORC Reader – v0.7
  - Work in progress: Will have similar API of Parquet reader
- Additionally looking towards GPU-accelerating decompression for common compression schemes



Source: Apache Crail blog: SQL Performance: Part 1 - Input File Formats

RAPIDS and Deep Learning Frameworks



RAPIDS PyTorch

Using RAPIDS with PyTorch

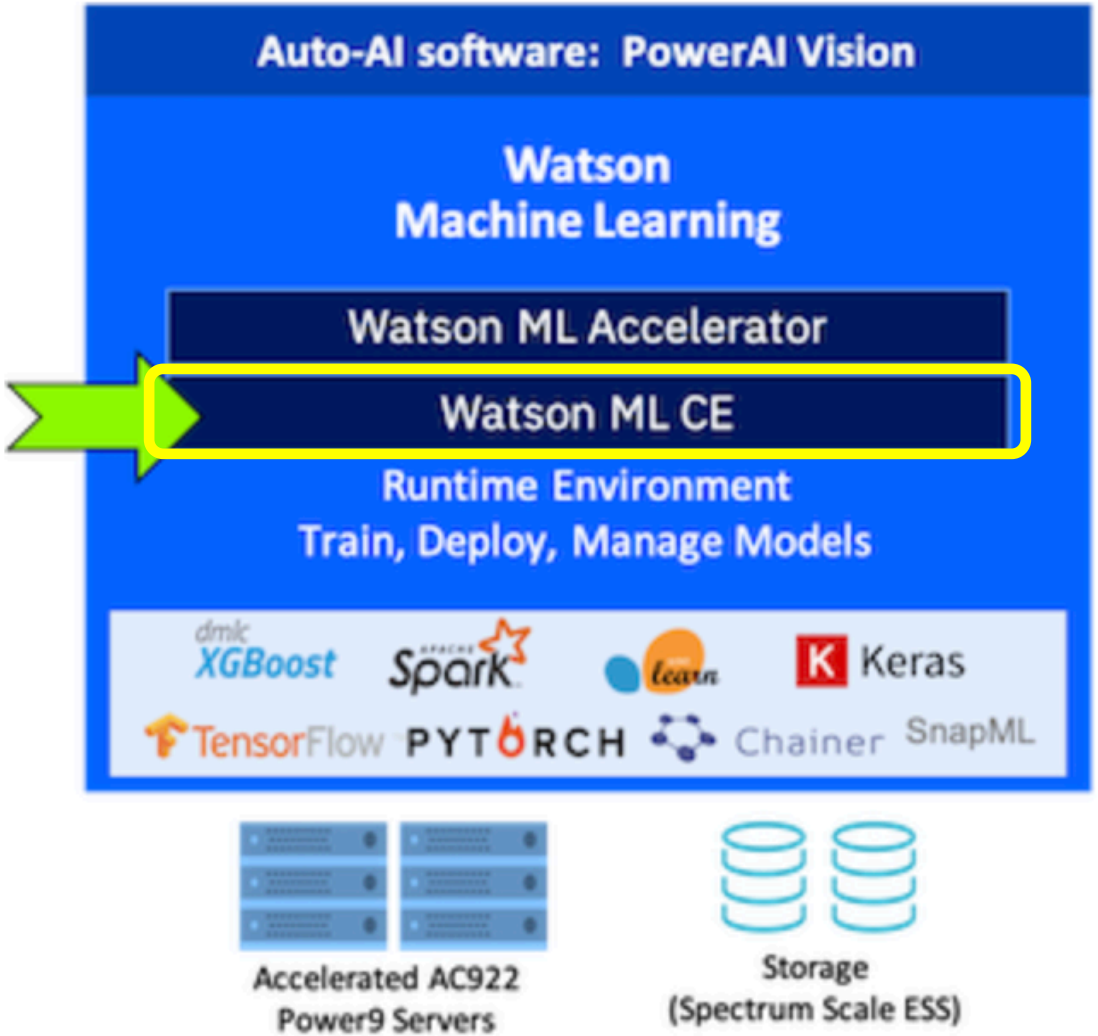
Using the GPU for ETL and preprocessing of deep learning workflows

Even Oldridge Follow

May 21 · 8 min read

https://medium.com/rapids-ai/using-rapids-with-pytorch-e602da018285

RAPIDS on Power – How can I use it ?



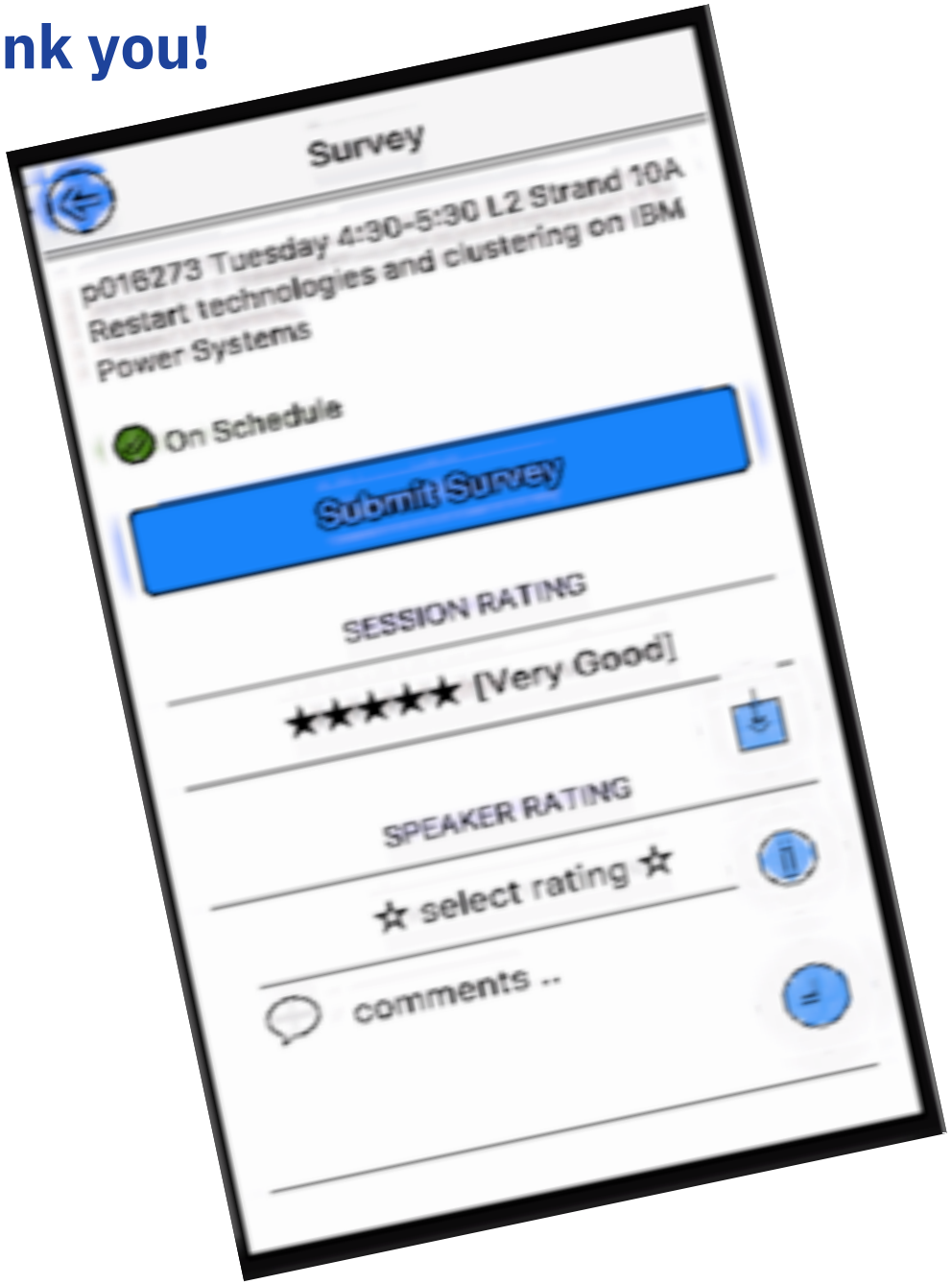

```
## Install PowerAI ! AKA WML-CE

conda config --prepend channels \
https://public.dhe.ibm.com/ibmdl/export/pub/sof
tware/server/ibm-ai/conda/

conda create --name wmlce_env python=3.6
conda activate wmlce_env

conda install powerai
```

```
## Python Program

import cudf
import cuml

…
…
```

https://www.ibm.com/support/knowledgecenter/SS5SF7_1.6.1/navigation/wmlce_install.htm

Jupyter Lab with Rapids

Thank you!



Dustin VanStee

Data Scientist

vanstee@us.ibm.com

Please complete the Session Evaluation!

IBM AI IMMERSION EXPERIENCE

2 Day Workshop

IMMERSIVE HANDS-ON SESSIONS USING ML/DL ON IBM GPU POWER SERVERS FOR DATA ENGINEERS, DATA SCIENTISTS AND PRACTITIONERS

Sample Agenda

Day1

- State of AI and General Use Cases
- Machine Learning Overview
- Lab: hands-on with ML using H2O DAI
- Auto AI Demo
- Deep Learning – Applications in Computer Vision
- Lab: Object Detection with PowerAIVision
- Building an AI Platform (Infrastructure)

Day 2

- Deep Learning with NLP Overview
- Lab: Aspect Based Sentiment using FastAI & PyTorch
- Deep Learning with Computer Vision
- Convolutional Neural Networks
- GAN Overview
- Lab: MNIST Data Generation Using GAN

Attendees will gain an understanding of the most recent breakthroughs in deep learning and how to apply these techniques using open source deep learning frameworks and IBM AI software.

Industry Focused

Financial
Comms/CSI
Retail
Healthcare
Distribution
Industrial
Research
Utilities

Customize the topics & agenda.

For a full menu of topics go to https://ibm.box.com/v/IBMAIBootcampTopics

SIGN UP BY CONTACTING
COGNITIVE SYSTEMS SOLUTIONS CENTER
CSSC@US.IBM.COM