RAPIDS

Accelerating Network Analysis
(and all your data science pipelines)

Brad Rees and Corey Nolet

Brad Rees

Brad Rees is a Manager in the AI Infrastructure group at NVIDIA and lead of the RAPIDS cuGraph team. Brad has been designing, implementing, and supporting a variety of advanced software and hardware systems for over 30 years. Brad specializes in complex analytic systems, primarily using graph analytic techniques for social and cyber network analysis. His technical interests are in HPC, machine learning, deep learning, and graph. Brad has a Ph.D. in Computer Science from the Florida Institute of Technology.

Corey Nolet

Corey Nolet is a Data Scientist & Senior Engineer on the RAPIDS cuML team at NVIDIA, where he focuses on building and scaling machine learning algorithms to support extreme data processing at light speed. Prior to NVIDIA, Corey spent over a decade building massive-scale exploratory data science & real-time analytics platforms for HPC environments in the defense industry. Corey holds BS. & MS. degrees in Computer Science and is currently working towards his PhD with a focus on scaling unsupervised machine learning algorithms. Corey has a passion for using data to make better sense of the world.

Agenda

- 15:30 – 16:00 Why RAPIDS
- *Break*
- 16:30 – 17:15 Components of RAPIDS
- 17:15 – 18:15 Analytic Workflow Example
- 18:15 - 18:30 Q&A

Why RAPIDS

Or why you should stay for the second half

Speed, UX, and Iteration

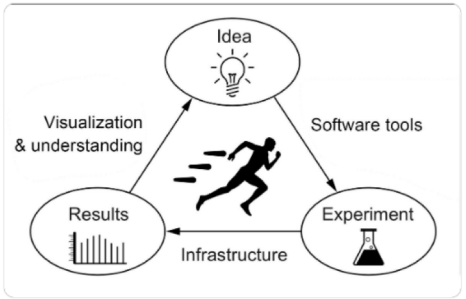
The Way to Win at Data Science

François Chollet ✓
@fchollet

Following

Winners are those who went through *more iterations* of the "loop of progress" -- going from an idea, to its implementation, to actionable results. So the winning teams are simply those able to run through this loop *faster*.

And this is were Keras gives you an edge.



12:31 PM - 3 Apr 2019

50 Retweets 158 Likes

5 50 158

François Chollet ✓ @fchollet · Apr 3
We often talk about how following UX best practices for API design makes Keras more accessible and easier to use, and how this helps beginners.

But those who stand to benefit most from good UX *aren't* the beginners. It's actually the very best practitioners in the world.

1 7 50

François Chollet ✓ @fchollet · Apr 3
Because good UX reduces the overhead (development overhead & cognitive overhead) to setting up new experiments. It means you will be able to iterate faster. You will be able to try more ideas.

And ultimately, that's how you win competitions or get papers published.

2 11 78

François Chollet ✓ @fchollet · Apr 3
So I don't think it's mere personal preference if Kaggle champions are overwhelmingly using Keras.

Using Keras means you're more likely to win, and inversely, those who practice the sort of fast experimentation strategy that sets them up to win are more likely to prefer Keras.

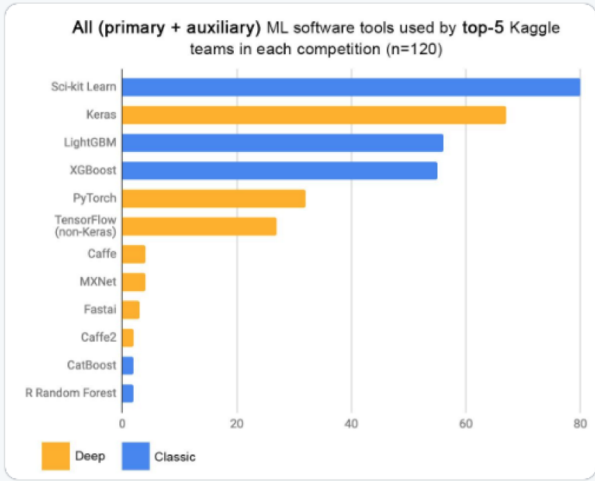
8 8 74

Joshua Patterson @datametrician · Apr 3
Replying to @fchollet
This is the fundamental believe that drives @RAPIDSai. @nvidia #GPU infrastructure is fast, people need to iterate quickly, people want a known #python interface. Combine them and you're off to the races!

2 11

François Chollet ✓ @fchollet · Apr 3
The second question asked about secondary frameworks -- usually teams win with an ensemble that involves many different ML frameworks. Here are *all* frameworks used.

Sklearn tops that ranking: everyone uses sklearn (although often as an auxiliary, for preprocessing or scoring).

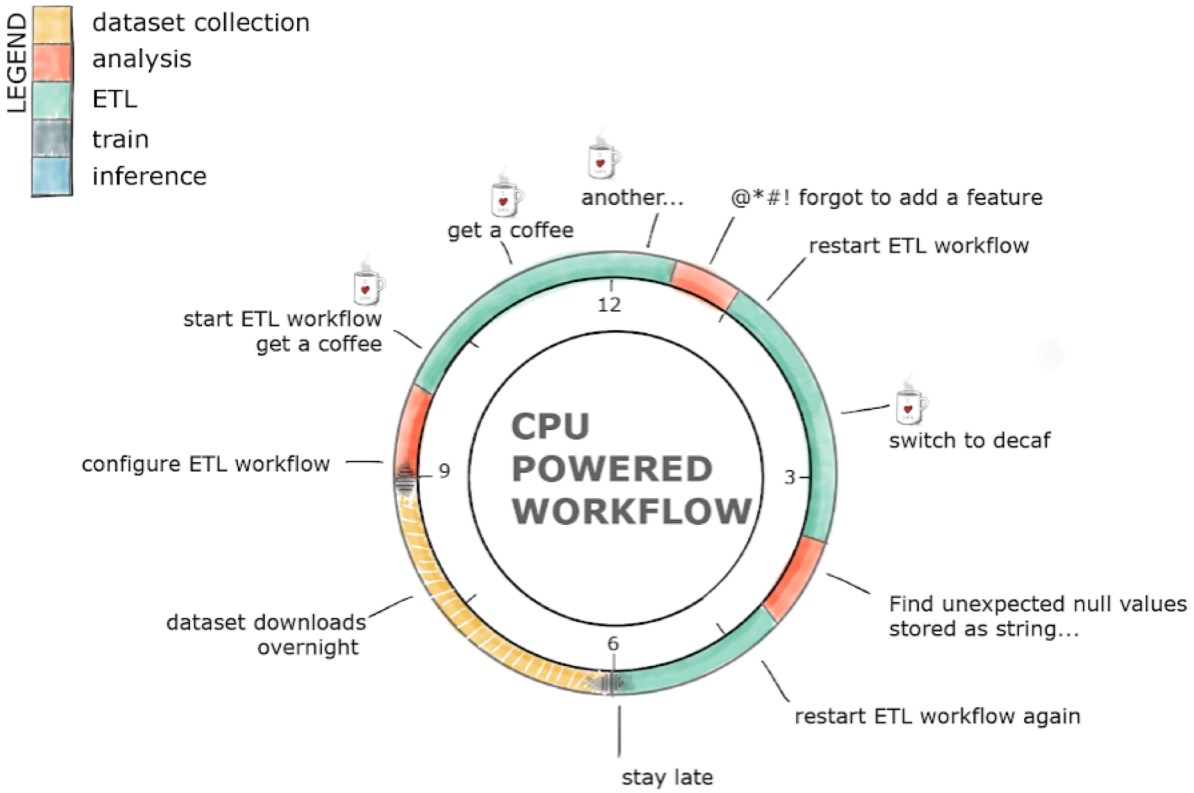


4 44 129

kaggle

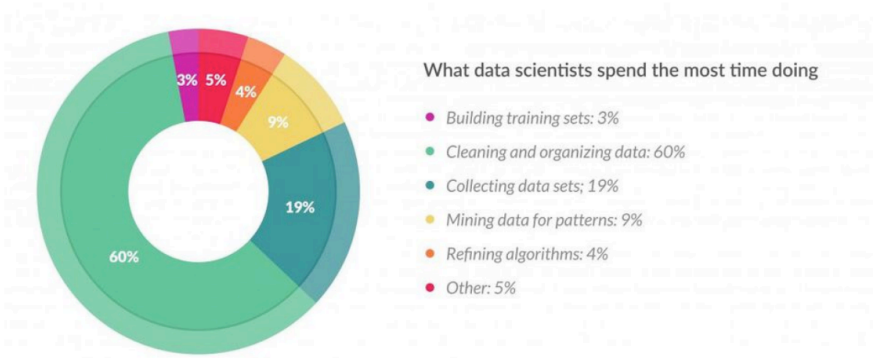
ITERATIONS TAKE TIME

We need more speed



Day in the life of a Data Scientist

Data preparation *accounts for about 80% of the work of data scientists*

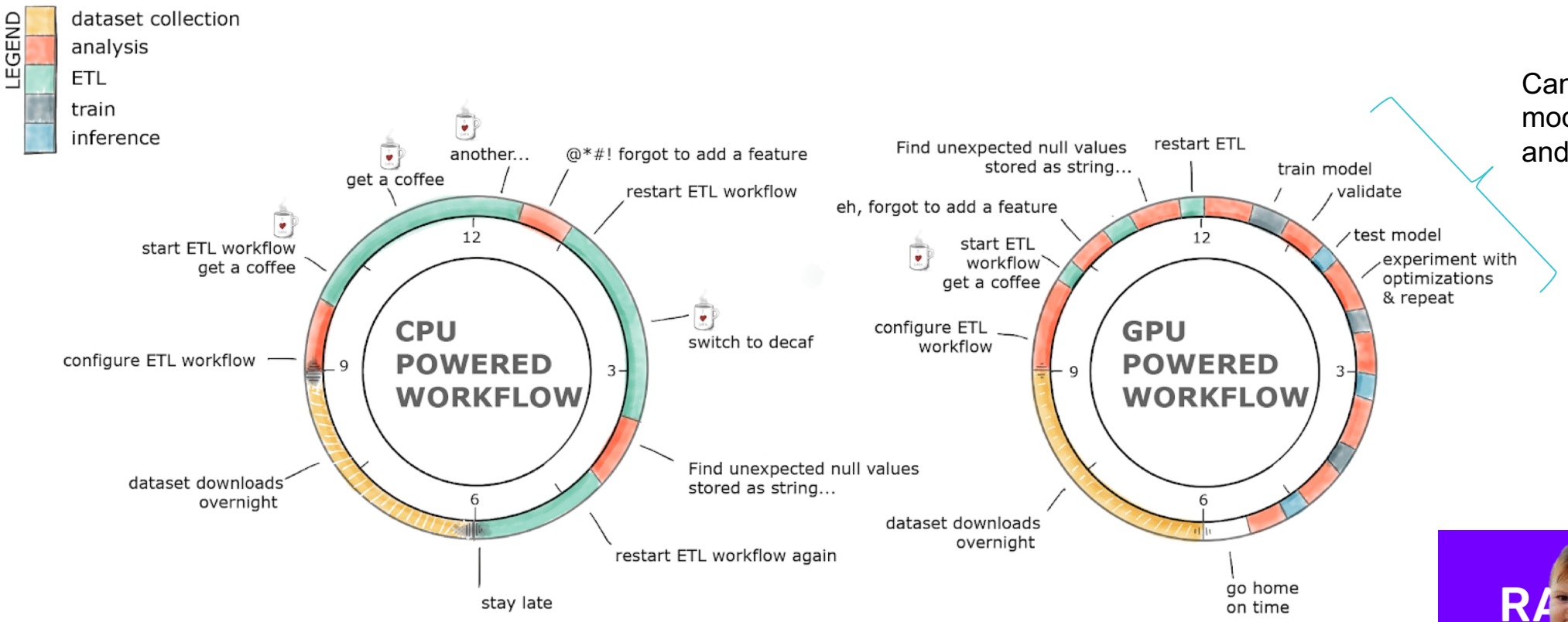


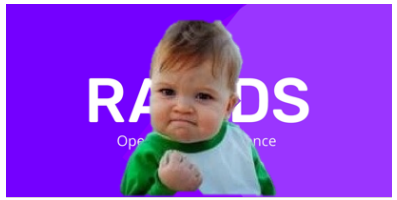
Forbes March 23, 2016

https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#15ff02c66f63

GPU-Accelerated Data Science

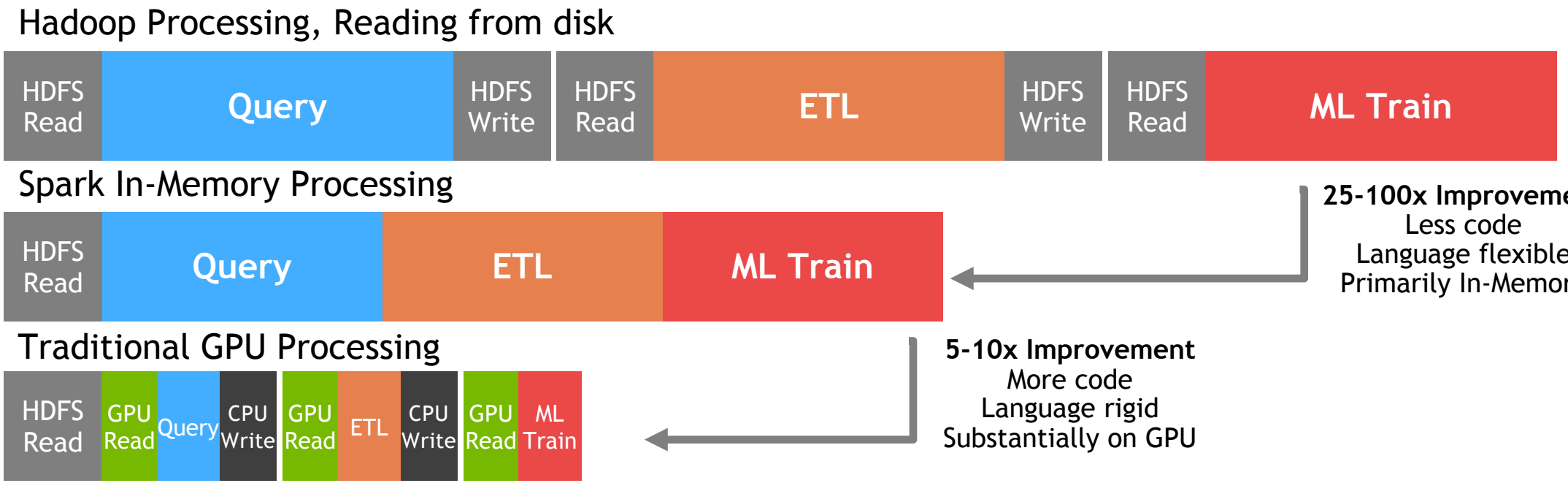
Shifting the focus from coffee to science





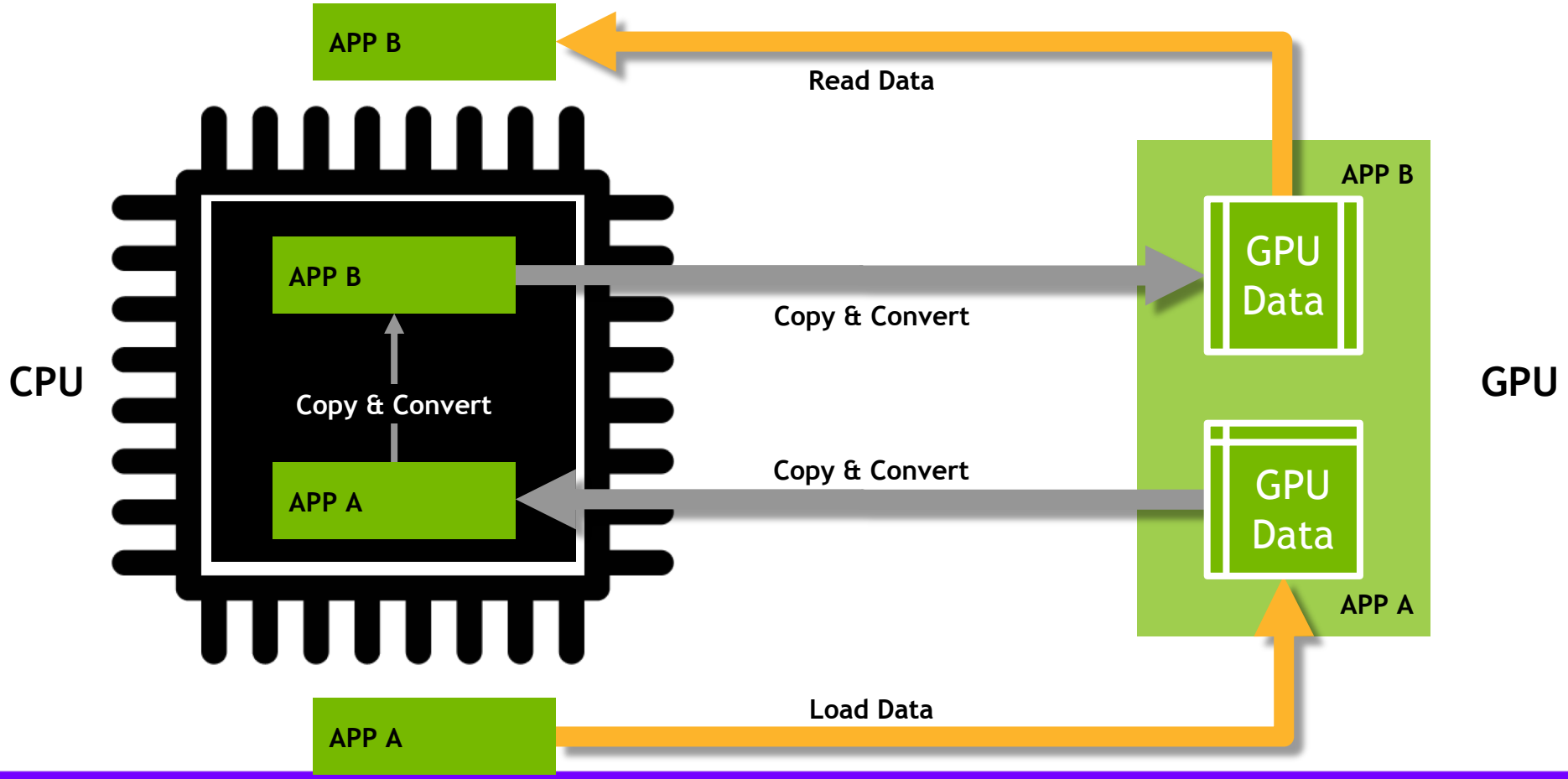
Data Processing Evolution

Faster data access, less data movement



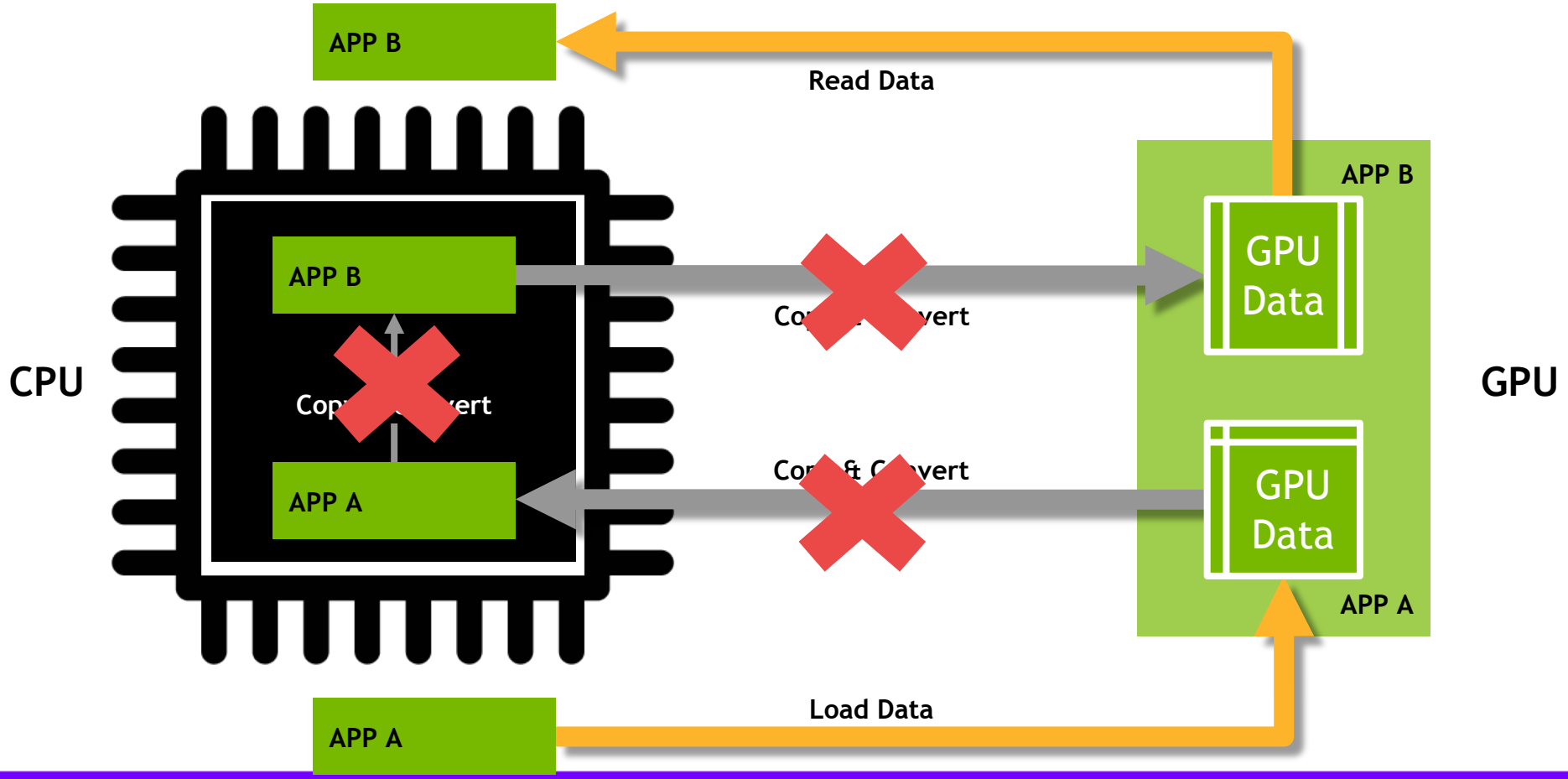
Data Movement and Transformation

The bane of productivity and performance

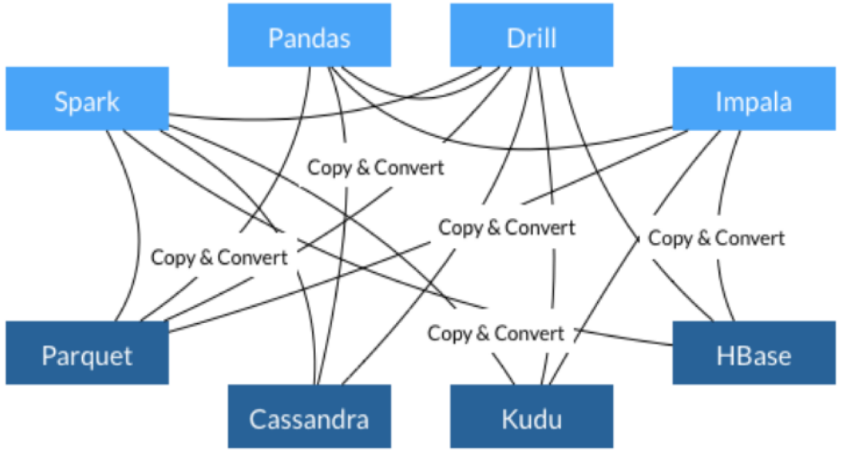


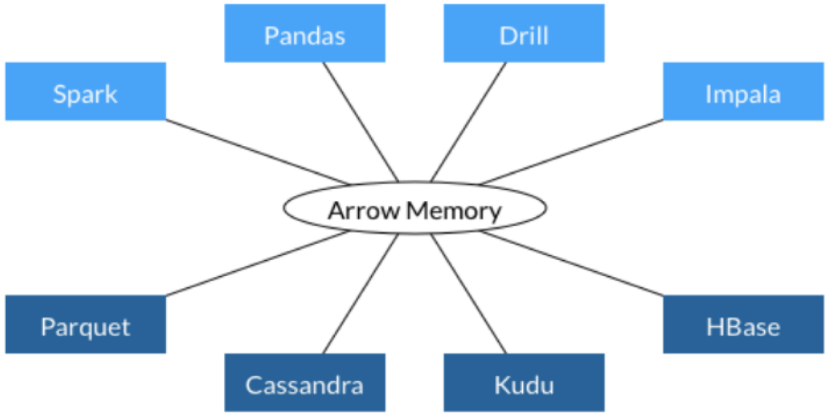
Data Movement and Transformation

What if we could keep data on the GPU?



Learning from Apache Arrow





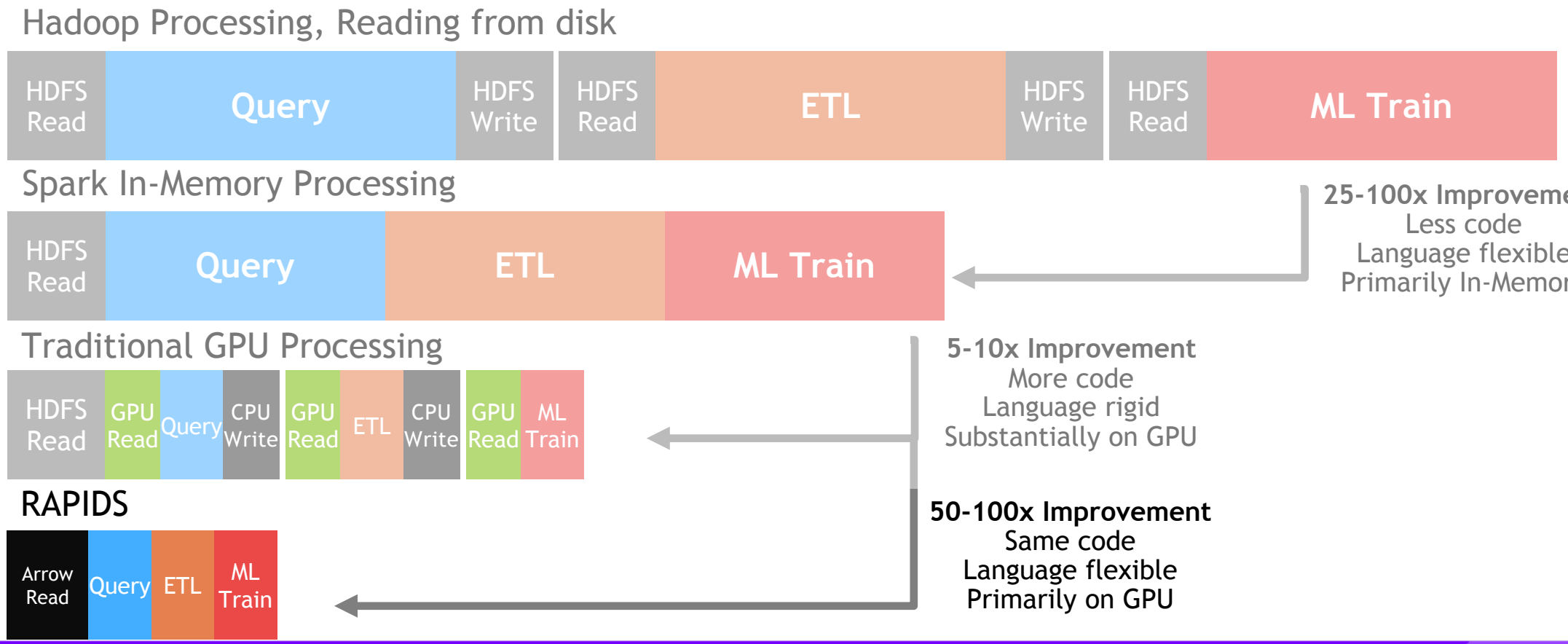
- Each system has its own internal memory format
- 70-80% computation wasted on serialization and deserialization
- Similar functionality implemented in multiple projects

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

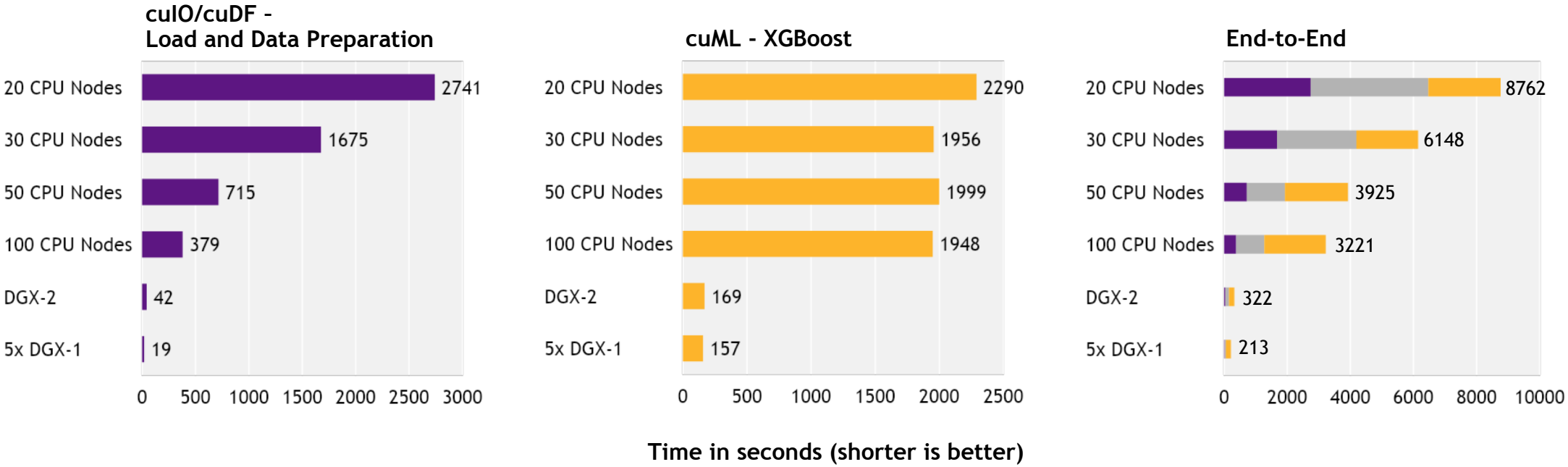
From Apache Arrow Home Page - https://arrow.apache.org/

Data Processing Evolution

Faster data access, less data movement



Faster Speeds, Real-World Benefits



Benchmark

200GB CSV dataset; Data prep includes joins, variable transformations

CPU Cluster Configuration

CPU nodes (61 GiB memory, 8 vCPUs, 64-bit platform), Apache Spark

DGX Cluster Configuration

5x DGX-1 on InfiniBand network

Customers

IMPROVING DEMAND FORECASTS

With over 100,000 different products in its 4,700 stores in the U.S., the Walmart Labs data science team must predict demand for 500 million items by store combinations every week.

By performing forecasting with the open-source RAPIDS data processing and machine learning (ML) libraries built on CUDA-X AI on NVIDIA GPUs, Walmart does ML feature engineering 100x faster and trains ML algorithms 20x faster. The company now reacts to shopper trends in real-time, speeds up product delivery, and realizes inventory cost savings at scale.

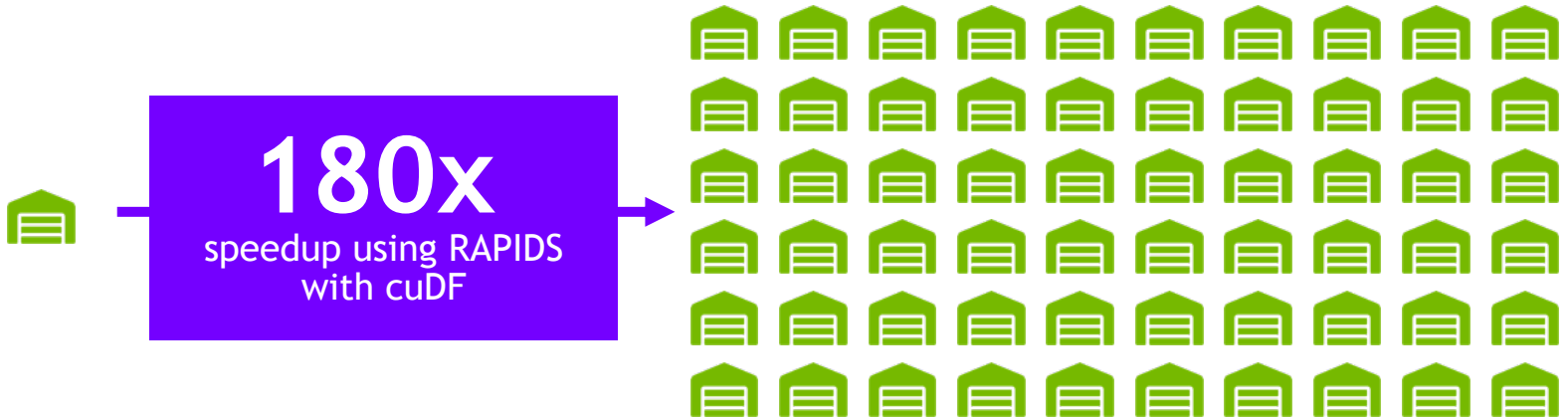
NVIDIA Walmart

“RAPIDS software has immensely improved how we use data — enabling the most complex models to run at scale and deliver even more accurate forecasting.”

— Jeremy King, EVP & CTO

Forecasting with accelerated data science

Optimizing Predictions to Maximize Savings



10 stores
1 million rows

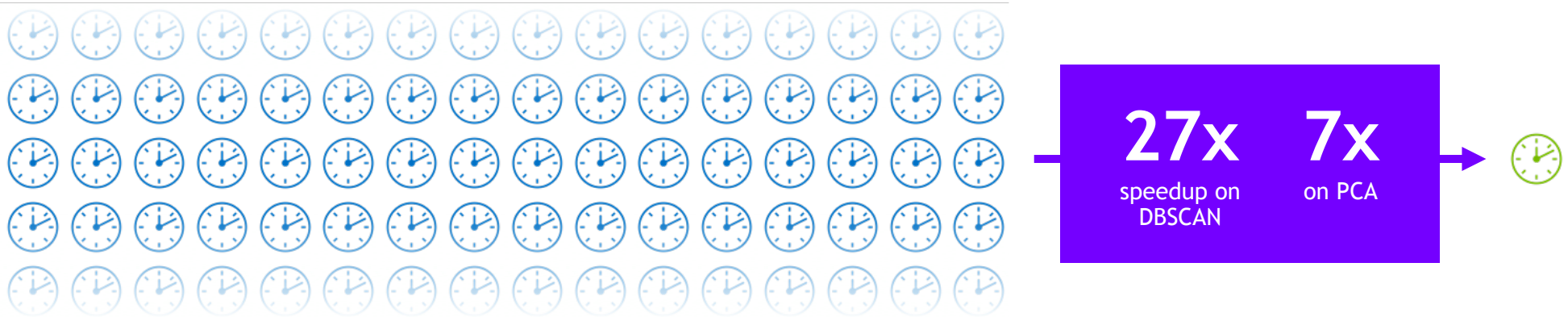
600 stores
60 million rows

"My previous bottleneck was I/O. …15 seconds to pull in data for 10 stores (about 1 Million rows). With RAPIDS, we can pull in data for about 600 stores (60 Million rows) in less than 5 seconds. … plain awesome."

— A mid-market specialty retailer with 6000 stores

Fraud detection with accelerated data science

Using the Power of AI to Reduce Loss



Days on CPUs

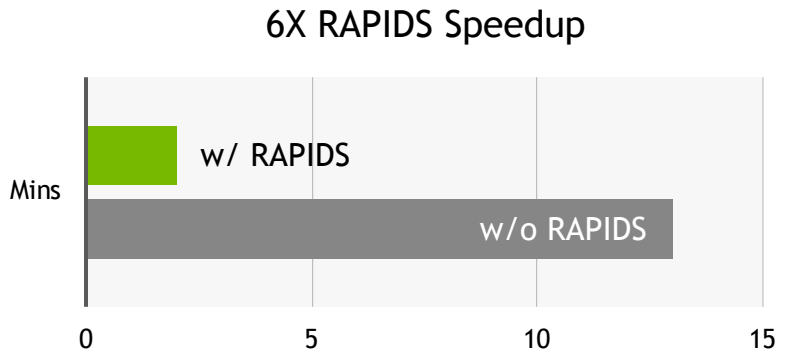
Hours on GPU

"RAPIDS reduced our model training time from days to hours. Now we can detect fraud with higher efficiency."

— Dr. Jian Zong Wang, Vice Chief Engineer and Senior AI Director
(from the Largest Insurance and Internet Finance Company in China)

RECOMMENDER SYSTEMS with accelerated data science

Personalized Consumer Health



Users of cuDF, PCA, XGBOOST, Pandas, SVD

"Using RAPIDS, we saw 6x speedup out of the gate! We can now apply machine learning to correlate microbiome features and Type-2 diabetes and deliver personalized health services such as diet suggestions or treatment planning."

— Hancheng Zheng, Director Of iCarbonX AI Lab

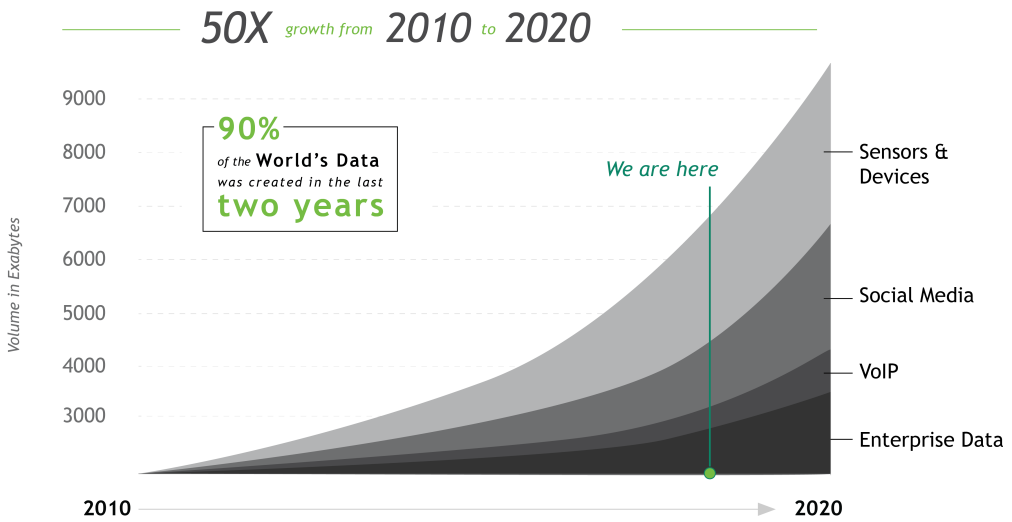
iCarbonX 碳云智能

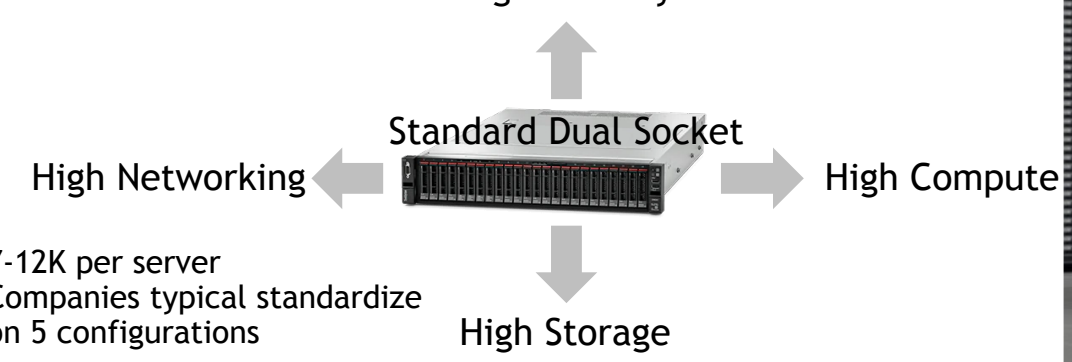
RAPIDS

NVIDIA Accelerated data science platforms

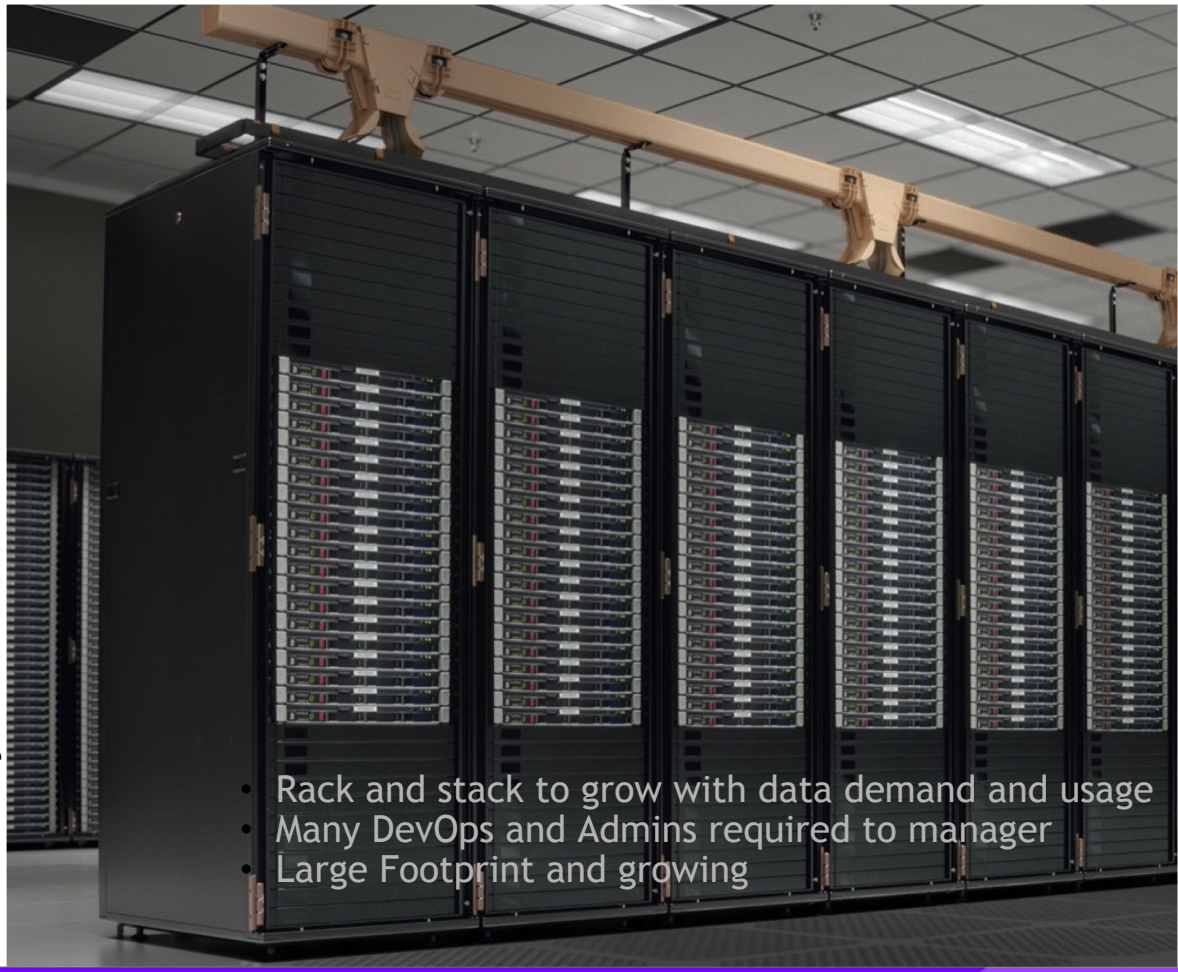
Data Science is Growing

In People, Footprint, Cost, Power, and more…





- 7-12K per server
- Companies typical standardize on 5 configurations



- Rack and stack to grow with data demand and usage
- Many DevOps and Admins required to manager
- Large Footprint and growing

Traditional data science CLUSTER

Workload Profile:

Fannie Mae Mortgage Data:

- 192GB data set
- 16 years, 68 quarters
- 34.7 Million single family mortgage loans
- 1.85 Billion performance records
- XGBoost training set: 50 features

300 Servers | $3M | 180 kW

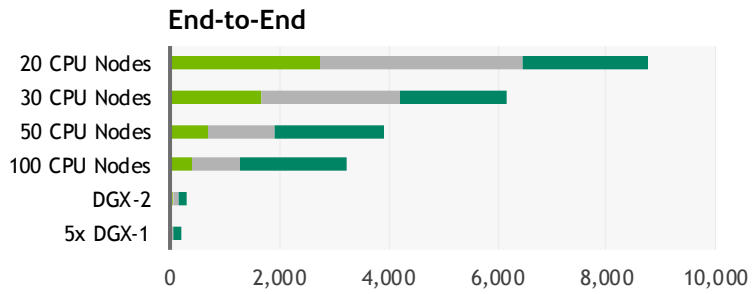
GPU-ACCELERATED DATA SCIENCE CLUSTER

GPU-accelerated XGBoost with DGX-2

1 DGX-2 | 10 kW

1/8 the Cost | 1/15 the Space

1/18 the Power



GPU-ACCELERATED DATA SCIENCE PLATFORMS

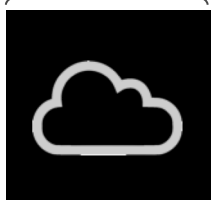
Unparalleled Performance and Productivity

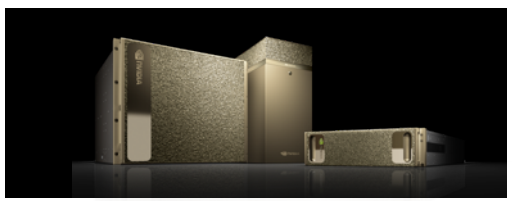
ML in the Cloud
All the top CSPs

ML Enthusiast
High-end PCs

Enterprise Desktop
Individual Workstations

Enterprise Data Center
Shared Infrastructure for Data Science Teams





Max Flexibility | **Max Performance**

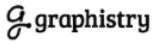
	NVIDIA GPUs in the Cloud	GeForce	TITAN RTX	NVIDIA-Powered Data Science Workstations	T4 Enterprise Servers	DGX Station, DGX-1 / HGX-1	DGX-2 / HGX-2
Benefit	Ease of getting started, low/no barrier to entry, elasticity of resources	Enthusiast PC solution, easy to acquire, low cost, great performance	The ultimate PC GPU for data scientists. Easy to acquire, deploy and get started experimenting.	Enterprise workstation for experienced data scientists	Standard GPU-accelerated data center infrastructures with the world's leading servers	Enterprise server, proven 4 or 8-way configuration, modular approach for scale-up, fastest multi-GPU & multi-node training	Largest compute and memory capacity in a single node, fastest training solution
Typical GPU Memory (system dependent)	varies depending on offering	22GB	48GB	96GB	16GB	128GB-256GB	512GB
GPU Fabric	varies depending on offering	2-way NVLink	2-way NVLink	2-way NVLink	Server dependent	4- and 8-way NVLink	16-way NVSwitch

Community

Ecosystem Partners

CONTRIBUTORS

ADOPTERS



OPEN SOURCE

Building on top of RAPIDS

A bigger, better, stronger ecosystem for all

High-Performance Serverless event and data processing that utilizes RAPIDS for GPU Acceleration

GPU accelerated SQL engine built on top of RAPIDS

Streamz

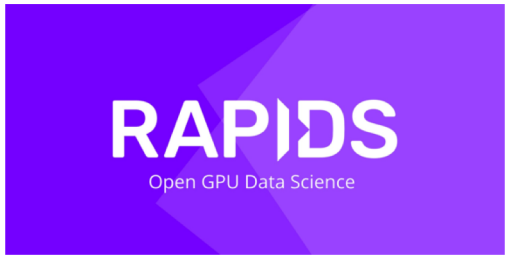
Distributed stream processing using RAPIDS and Dask

Deploy RAPIDS Everywhere

Focused on robust functionality, deployment, and user experience

Cloud Dataproc

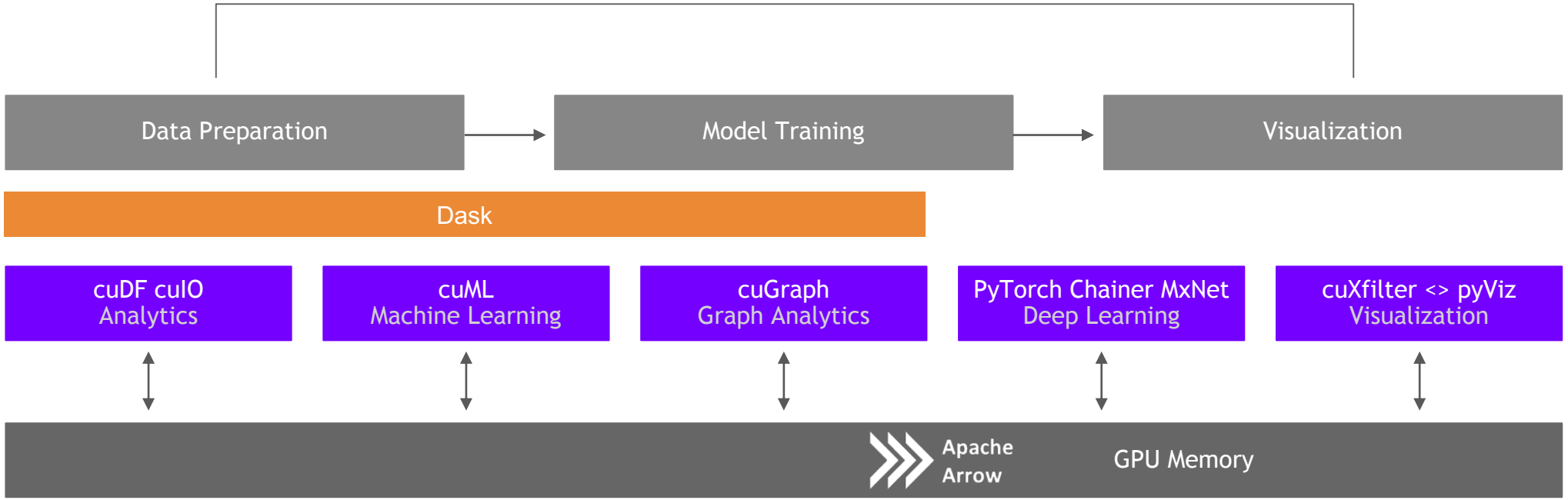
Azure Machine Learning



Integration with major cloud providers
Both containers and cloud specific machine instances
Support for Enterprise and HPC Orchestration Layers

RAPIDS

End-to-End Accelerated GPU Data Science



Coffee Break

When we come back:

- Short overview of RAPIDS libraries
- RAPIDS in active

    Sample Jupyter notebooks showing how to use RAPIDS for various analytic problems

RAPIDS