ACCELERATING TIME TO VALUE WITH XGBOOST ON NVIDIA GPUS

Chris Kawalek, NVIDIA

Paul Hendricks, NVIDIA

Chris Kawalek
Sr. Product Marketing Manager, AI and Data Science Platform Solutions

Paul Henricks
Solutions Architect, Deep Learning and AI

AGENDA

How Data Science is Transforming Business

NVIDIA Accelerated Data Science and XGBoost

GPU-Accelerated XGBoost Technical Overview

Distributed XGBoost with Apache Spark and Dask

How to Get Started with GPU-Accelerated XGBoost

Resources

Q&A and Wrap Up

DATA SCIENCE IS THE KEY TO MODERN BUSINESS

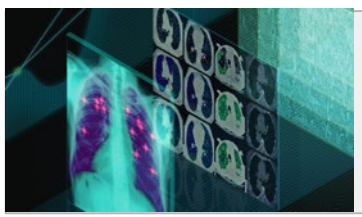
Forecasting, Fraud Detection, Recommendations, and More

RETAIL

Supply Chain & Inventory Management
Price Management / Markdown Optimization
Promotion Prioritization And Ad Targeting

TELECOM

Detect Network/Security Anomalies
Forecasting Network Performance
Network Resource Optimization (SON)



HEALTHCARE

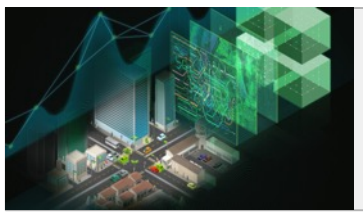
Improve Clinical Care
Drive Operational Efficiency
Speed Up Drug Discovery

MANUFACTURING

Remaining Useful Life Estimation
Failure Prediction
Demand Forecasting

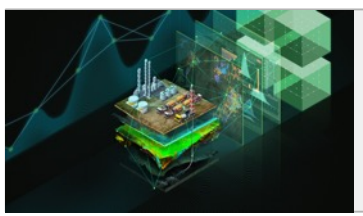
FINANCIAL SERVICES

Claim Fraud
Customer Service Chatbots/Routing
Risk Evaluation



CONSUMER INTERNET

Ad Personalization
Click Through Rate Optimization
Churn Reduction



OIL & GAS

Sensor Data Tag Mapping
Anomaly Detection
Robust Fault Prediction

AUTOMOTIVE

Personalization & Intelligent Customer Interactions
Connected Vehicle Predictive Maintenance
Forecasting, Demand, & Capacity Planning

CHALLENGES AFFECTING DATA SCIENCE TODAY

What Needs to be Solved to Empower Data Scientists

INCREASING DATA ONSLAUGHT

Data sets are continuing to dramatically increase in size

Multitude of sources

Different formats, varying quality

SLOW CPU PROCESSING

End of Moore's law, CPUs aren't getting faster

Many popular data science tools have been CPU-only

Can only throw so many CPUs at a job

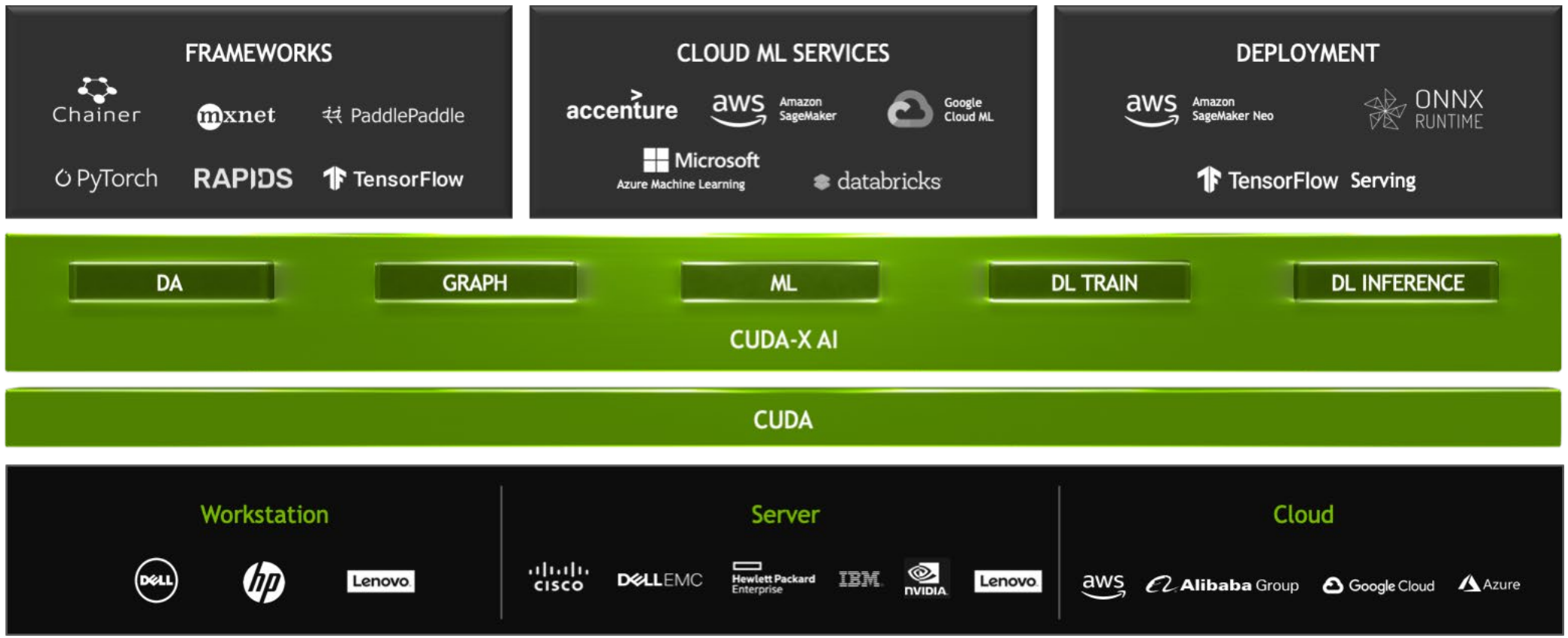
WORKING AT SCALE

Expertise required to scale beyond a single GPU

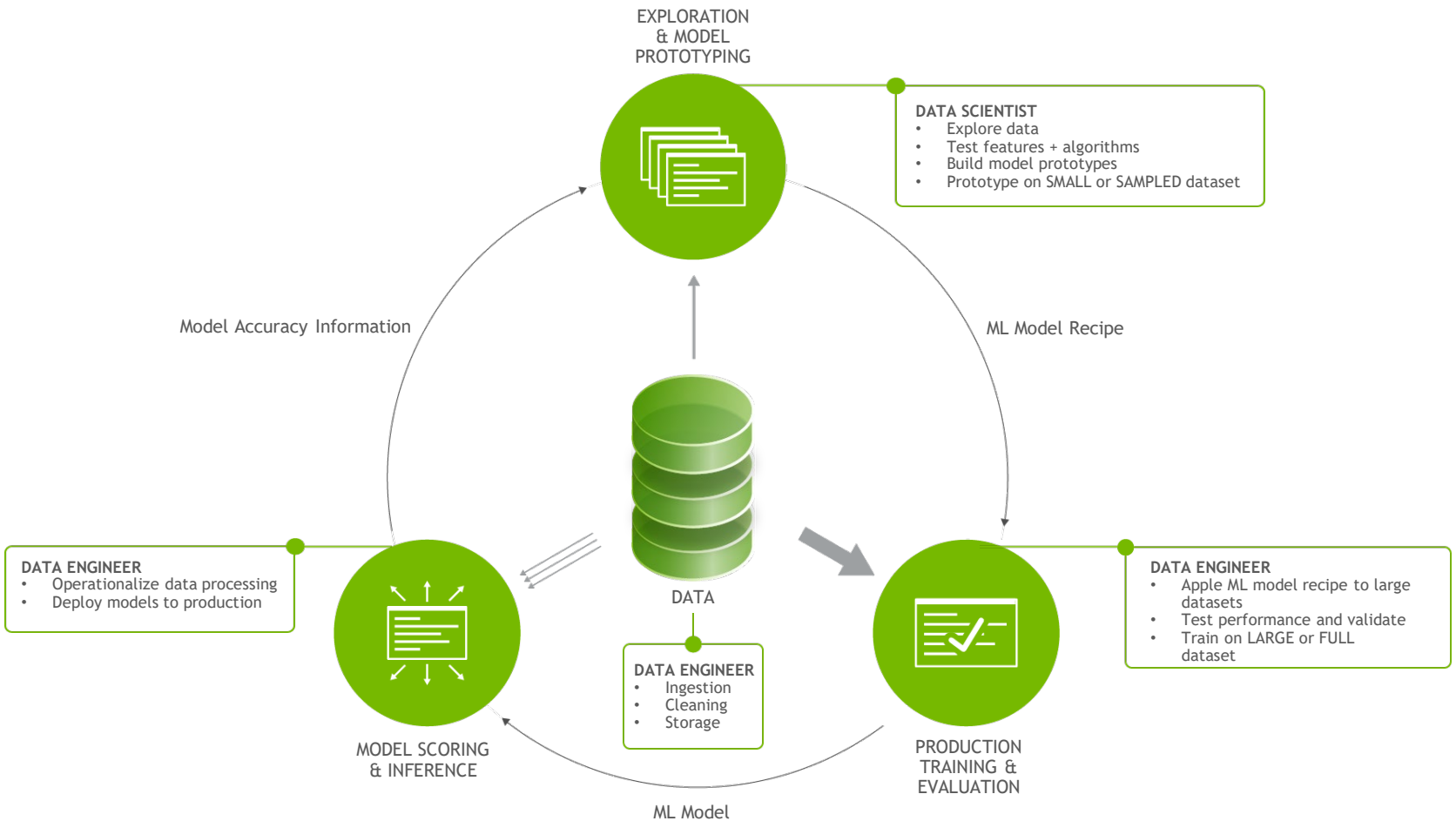
Challenging to scale to multiple nodes

CUDA-X AI TRANSFORMS DATA SCIENCE

From Data Science to NVIDIA **Accelerated** Data Science with CUDA-X AI



DATA SCIENCE DEVELOPMENT LIFECYCLE



XGBOOST: THE WORLD'S MOST POPULAR MACHINE LEARNING ALGORITHM

Versatile and High Performance

The leading algorithm for tabular data

Outperforms most ML algorithms on regression, classification and ranking

Winner of many data science Kaggle competitions

InfoWorld Technology of the Year Award, 2019

Well known in data science community and widely used for **forecasting**, **fraud detection**, **recommender engines**, and much more

dmlc XGBoost

HOW CAN XGBOOST BE IMPROVED?

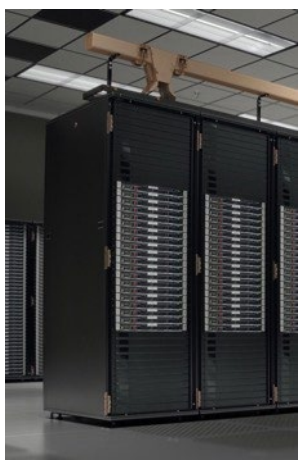
XGBoost Performance is Constrained by CPU Limitations

CPU processing is slow, creating issues for large data sets or when timeliness is crucial (e.g. intraday requirements for financial services)

Hyperparameter search is very slow, making search not feasible

Prediction speed limits the depth and number of trees in time sensitive applications

dmlc XGBoost



GPU-ACCELERATED XGBOOST

Unleashing the Power of NVIDIA GPUs for Users of XGBoost

Faster Time To Insight
XGBoost training on GPUs is significantly faster than CPUs, completely transforming the timescales of machine learning workflows.

Better Predictions, Sooner
Work with larger datasets and perform more model iterations without spending valuable time waiting.

Lower Costs
Reduce infrastructure investment and save money with improved business forecasting.

Easy to Use
Works seamlessly with the RAPIDS open source data processing and machine learning libraries and ecosystem for end-to-end GPU-accelerated workflows.

EASY TO USE, MINIMAL CODE CHANGES

GPU-Acceleration with the same XGBoost Usage

BEFORE

```
import xgboost as xgb

params = {'max_depth': 3,
        'learning_rate': 0.1}


dtrain = xgb.DMatrix(X, y)
bst = xgb.train(params, dtrain)
```

AFTER

```
import xgboost as xgb

params = {‘tree_method’: ‘gpu_hist’,
        'max_depth': 3,
        'learning_rate': 0.1}

dtrain = xgb.DMatrix(X, y)
bst = xgb.train(params, dtrain)
```

DISTRIBUTED XGBOOST

GPU-Accelerated XGBoost for Large Scale Workloads

GPU-acceleration for **XGBoost** with **Apache Spark** and **Dask**

Multiple nodes and multiple GPUs per node

Explore and prototype models on a PC, workstation, server, or cloud instance and scale to two or more nodes for production training

An ideal solution for GPU-accelerated clusters and enterprise scale workloads

Try out Dask support immediately using Google Cloud Dataproc

Download for on-prem and cloud deployments

GPU-ACCELERATED DATA SCIENCE PLATFORMS

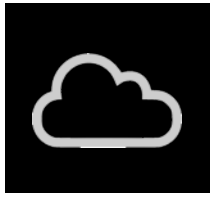
Unparalleled Performance and Productivity

ML in the Cloud
All the top CSPs

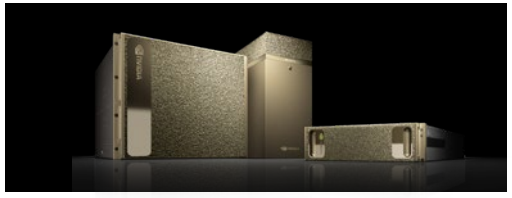
ML Enthusiast
High-end PCs

Enterprise Desktop
Individual Workstations

Enterprise Data Center
Shared Infrastructure for Data Science Teams





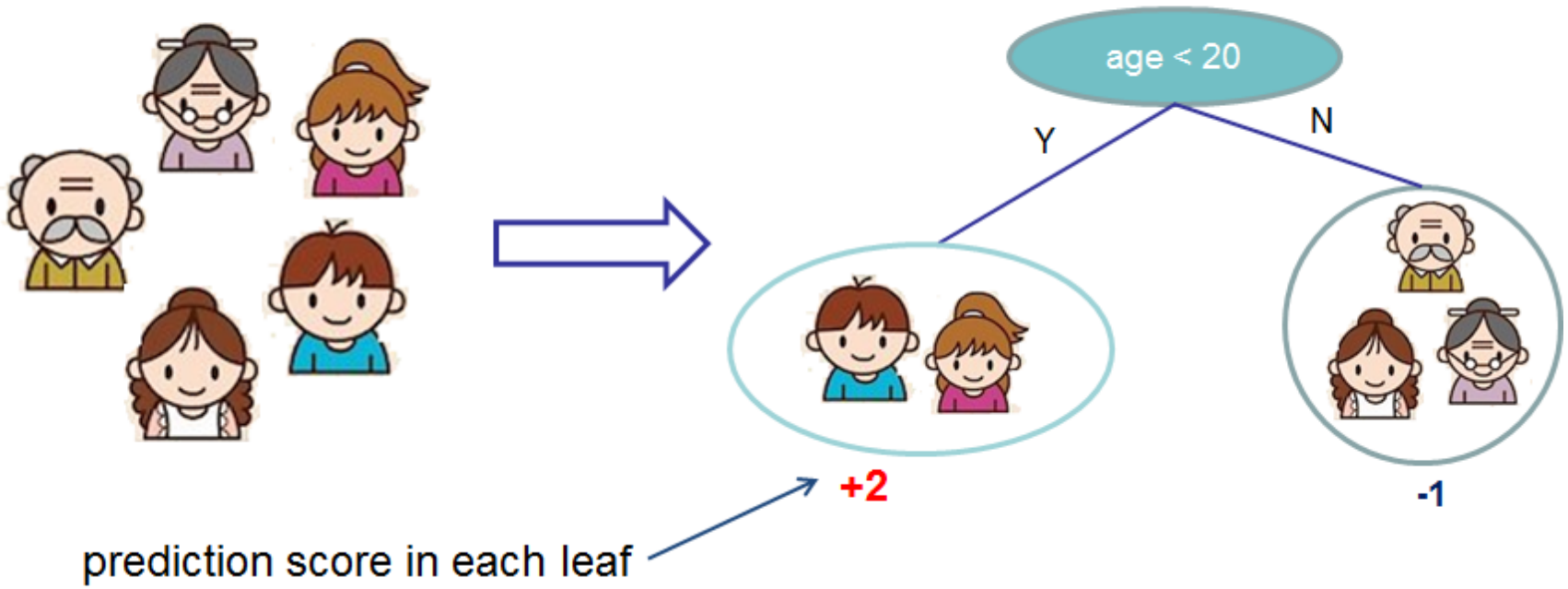
					Max Flexibility	Max Performance	
	NVIDIA GPUs in the Cloud	GeForce	TITAN RTX	NVIDIA-Powered Data Science Workstations	T4 Enterprise Servers	DGX Station, DGX-1 / HGX-1	DGX-2 / HGX-2
Benefit	Ease of getting started, low/no barrier to entry, elasticity of resources	Enthusiast PC solution, easy to acquire, 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	64 GB (4 x 16 GB)	128GB-256GB	512GB
GPU Fabric	varies depending on offering	2-way NVLink	2-way NVLink	2-way NVLink	PCIe 3.0	4- and 8-way NVLink	16-way NVSwitch

GPU-ACCELERATED XGBOOST TECHNICAL OVERVIEW

Example of a Decision Tree

Input: age, gender, occupation, …

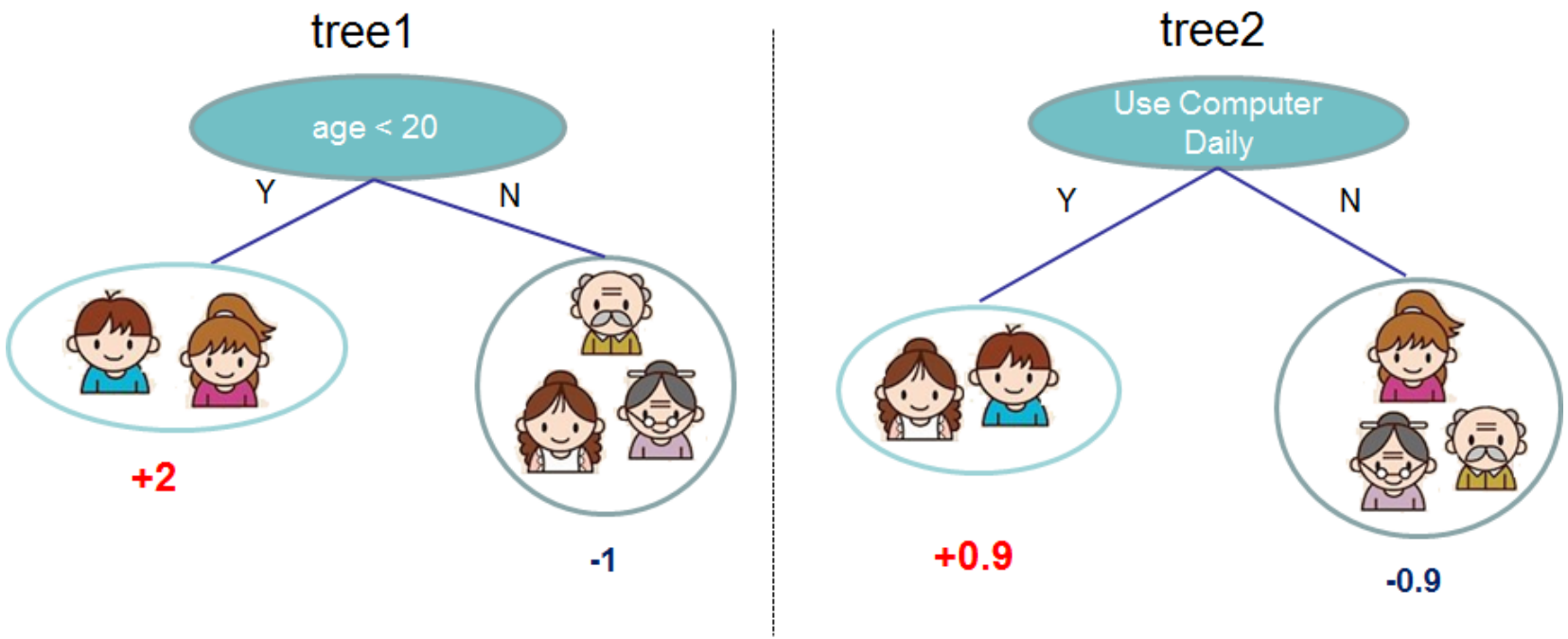
Like the computer game X



Source: https://xgboost.ai/2016/12/14/GPU-accelerated-xgboost.html

COMBINE TREES FOR STRONGER PREDICTION

If (age < 15) AND (use computer daily) = enjoy computer games



f() = 2 + 0.9= 2.9 f()= -1 - 0.9= -1.9

Source: https://xgboost.ai/2016/12/14/GPU-accelerated-xgboost.html

XGBOOST IMPROVEMENTS

Scaling Using Apache Spark and Dask

Apache Spark

Easy to use through Scala/Java API using an existing Spark cluster – Python API coming soon

YARN for managing resources

Dask

Simple to use Python API that is infrastructure agnostic

Kubernetes for managing resources

Both Apache Spark and Dask support loading CSV, Parquet, ORC, and other file formats from local disk, HDFS, S3, GS, Azure Storage

Near identical performance for both Apache Spark and Dask - ~9X speedup on NVIDIA T4

SCALING USING DASK

GPU-Accelerated ETL and Model Training

Dask - A Python library for distributed computing

Compose multiple Directed Acyclic Graphs (DAGs) and execute

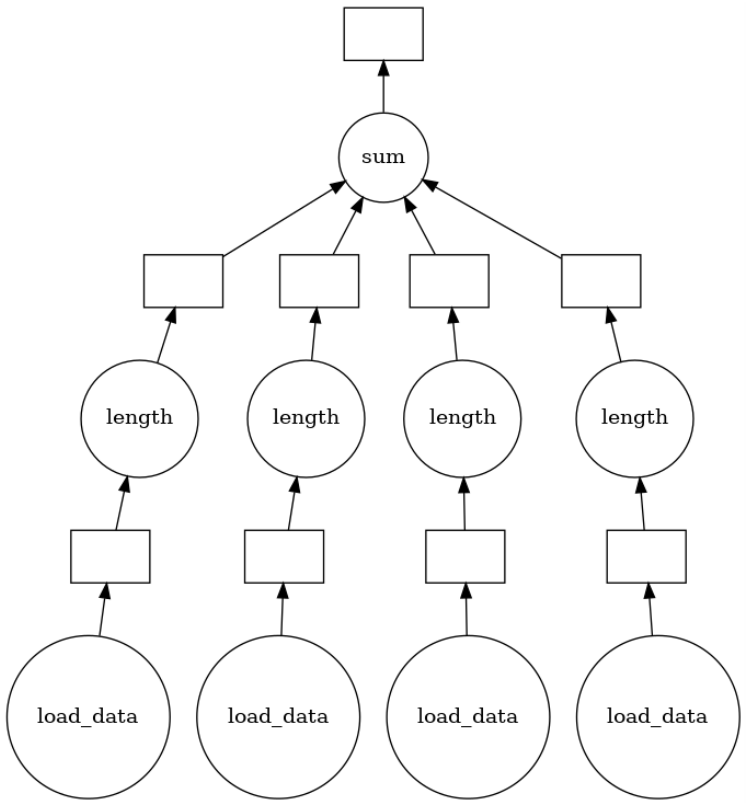
Create clusters of multiple nodes and interact with using a Client

In [3]:
```
from dask.distributed import Client
from dask_cuda import LocalCUDACluster


# create a local CUDA cluster
cluster = LocalCUDACluster()
client = Client(cluster)
client
```

Out[3]:

Client	Cluster
• **Scheduler:** tcp://127.0.0.1:40615 • **Dashboard:** http://127.0.0.1:8787/status	• **Workers:** 8 • **Cores:** 8 • **Memory:** 540.95 GB



EASY TO USE, MINIMAL CODE CHANGES

GPU-Acceleration with the Same XGBoost Usage

BEFORE

```
import xgboost as xgb




params = {‘tree_method’: ‘gpu_hist’,
        ‘n_gpus’: 1,
        'max_depth': 3,
        'learning_rate': 0.1,
        ‘num_boost_round’: 100}

dtrain = xgb.DMatrix(X, y)




bst = xgb.train(params, dtrain)
```

AFTER

```
import dask
import dask_cudf
import dask_xgboost as dask_xgb

client = Client()

params = {‘tree_method’: ‘gpu_hist’,
        ‘n_gpus’: 1.
        'max_depth': 3,
        'learning_rate': 0.1}

X_dask_df = dask.dataframe.from_array(X)
X_dask_cudf = dask_cudf.from_dask_dataframe(X_dask_df)

y_dask_df =  dask.dataframe.from_array(y)
y_dask_cudf = dask_cudf.from_dask_dataframe(y_dask_df)

bst = dask_xgb.train(client, params,
                X_dask_cudf,
                y_dask_cudf,
                num_boost_round=100)
```

XGBOOST WITH GOOGLE DATAPROC

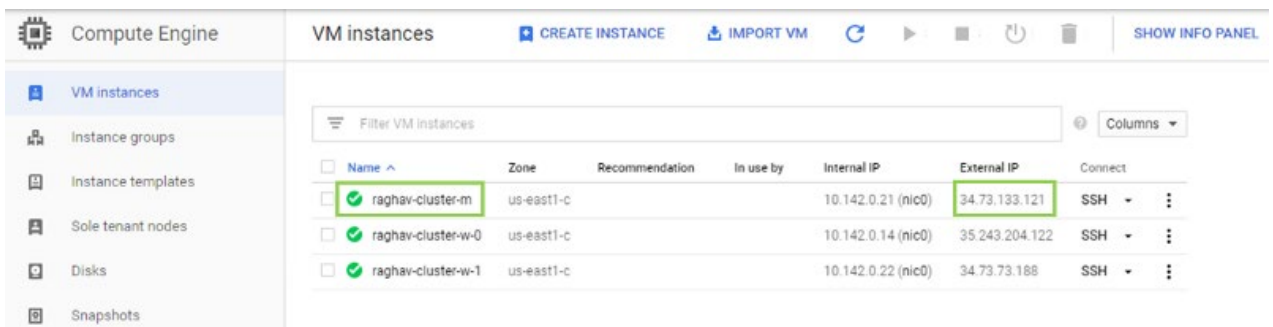
Easily Create, Interact With, and Monitor Dask Clusters

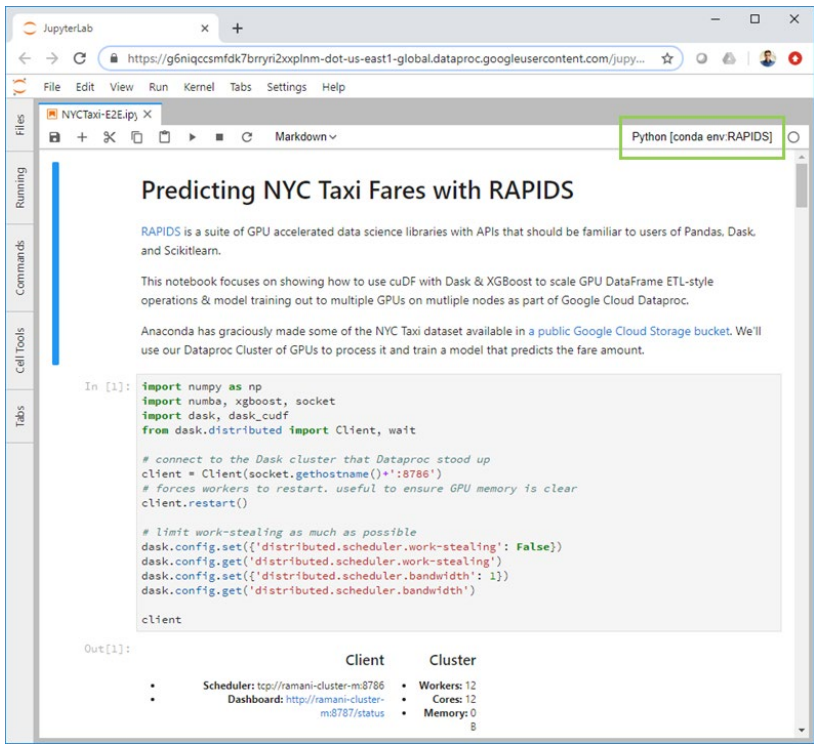
Create clusters of T4s and V100s on-demand

Load data from Google Cloud Storage Buckets

Interact with cluster through Jupyter Notebook

Monitor scheduled jobs through Dask Dashboard

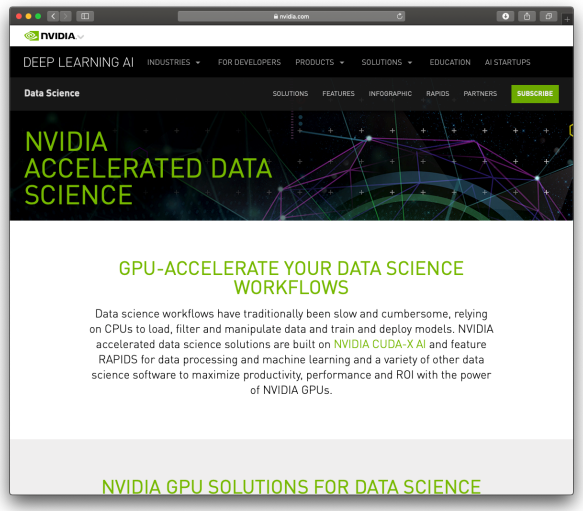




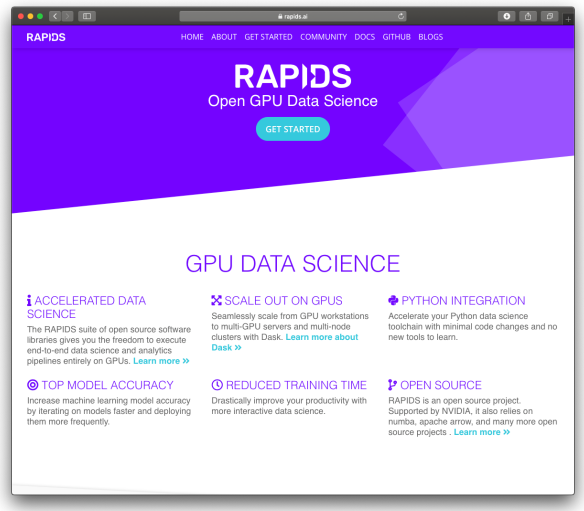
Source: https://medium.com/rapids-ai/scale-out-rapids-on-google-cloud-dataproc-8a873233258f?ncid=so-twi-n2-92572

DEMO VIDEO (NOTEBOOK WALKTHROUGHS)

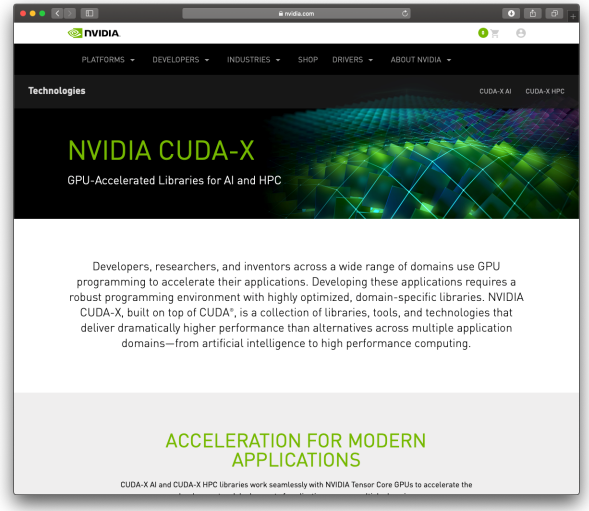
FOR MORE INFORMATION



nvidia.com/datascience



rapids.ai



nvidia.com/en-us/technologies/cuda-x/

THANK YOU!


NVIDIA