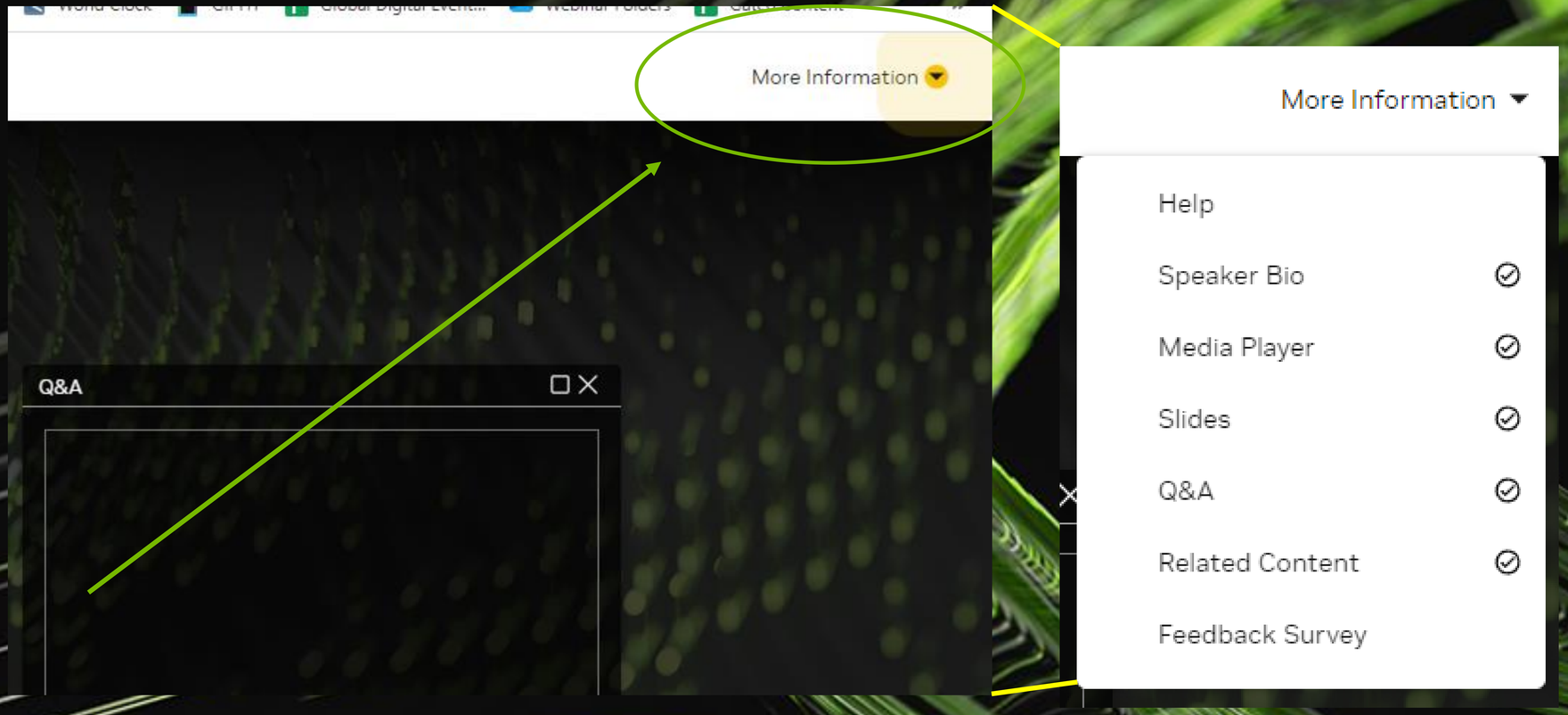
How To Use the Console



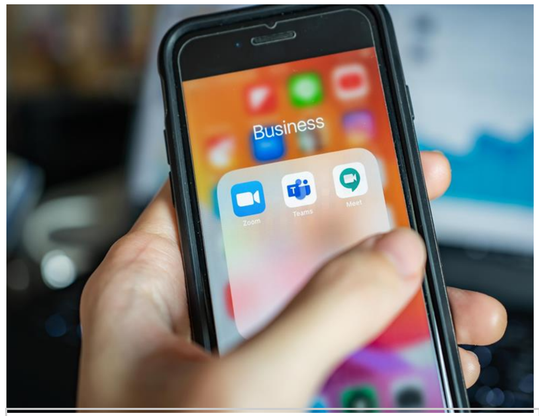
Improve Retail Business Outcomes With GPU-Accelerated Apache Spark

Charu Chaubal

Ward Eldred

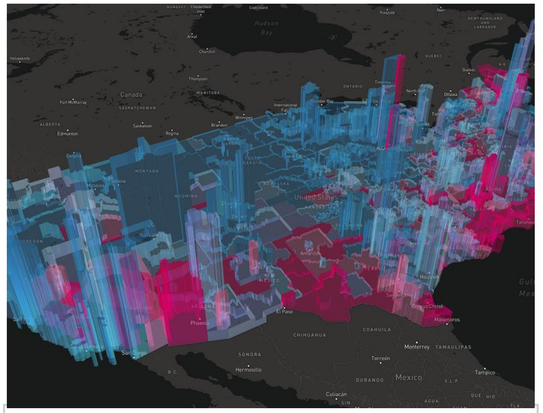
Data Analytics in Retail

Used for many core business functions

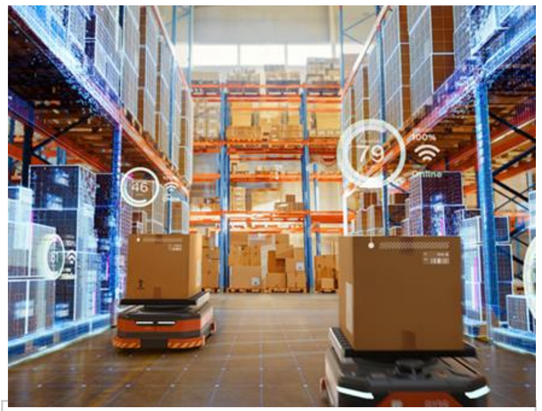


Recommender Systems
Advertising Analytics

Inventory Forecasting
Supply Chain Analysis



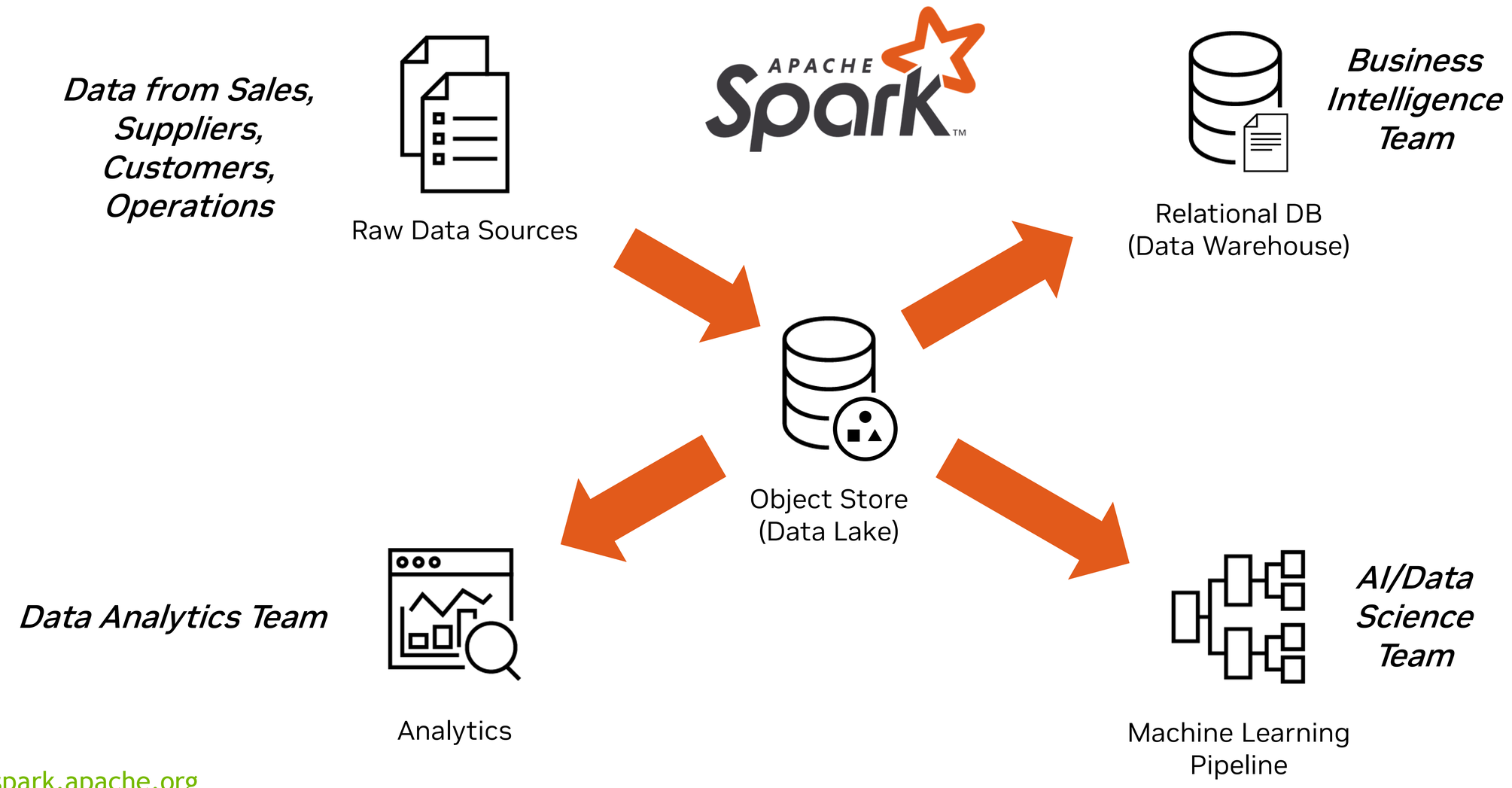
Audience Segmentation
Demand Forecasting



Price Optimization
Product Analysis

Apache Spark is widespread in the Modern Enterprise

60% of the Fortune 500 use Apache Spark 3.x[1]



[1] https://spark.apache.org

Data Processing Challenges

Common problems that enterprises are facing today

Time is Precious

Data processing workflows are constrained by slow compute.

Datasets are Growing

CPU-based infrastructure is no longer effective, resulting in higher cost and larger carbon footprint.

Reliability is Crucial

Reliance on community support for key business functions is risky.

NVIDIA RAPIDS Accelerator for Apache Spark

Improve Your Existing Data Processing Workflows

5x

Faster Execution Time

- Move data in and out of data lakes more quickly.
- Take advantage of faster analytics
- Accelerate AI pipelines

4x

Lower Costs

- Save on cloud usage costs
- Reduce power consumption and carbon footprint

Full Enterprise Support

NVIDIA AI Enterprise 3.1 offers

- Mission critical support
- Bug fixes
- Professional services

Amazon EMR

Google Cloud Dataproc

databricks

How it Works

Key technologies for GPU acceleration

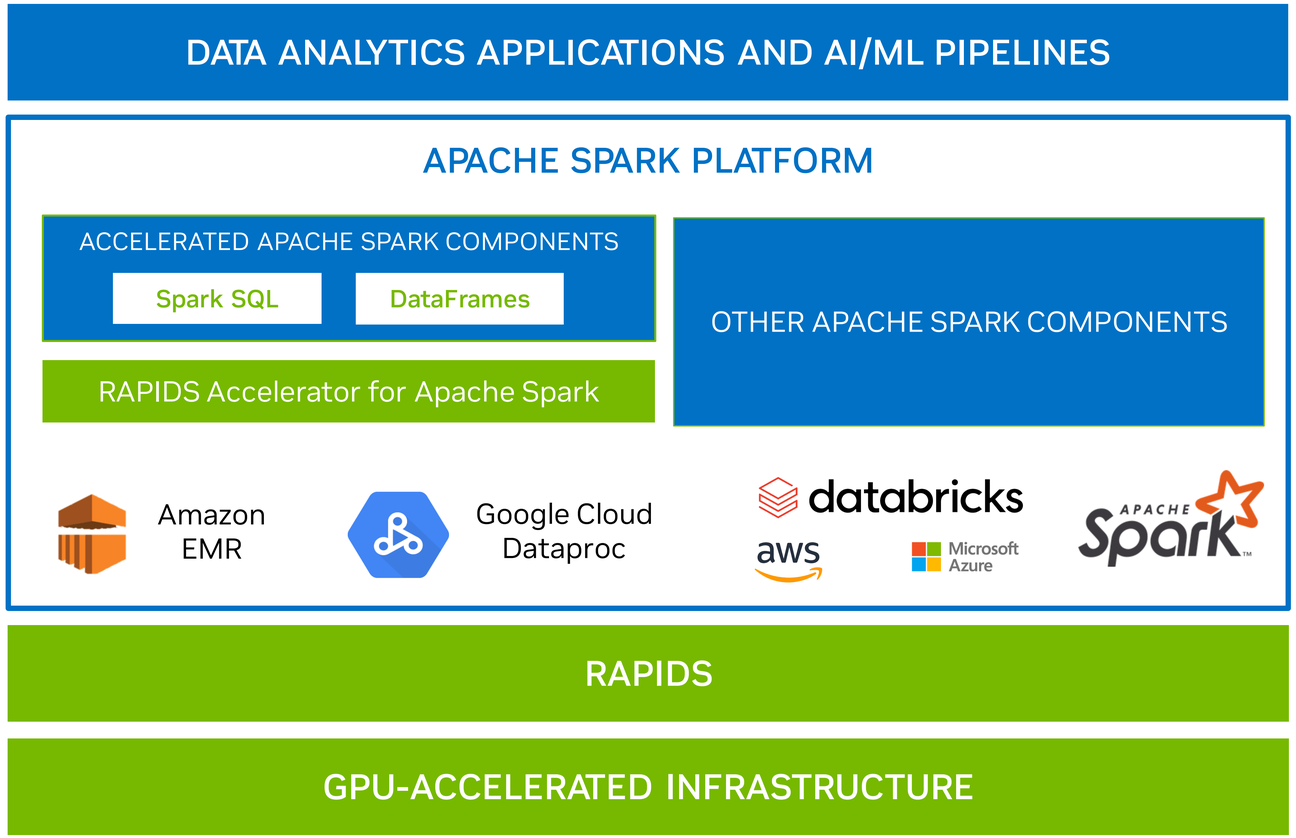
NVIDIA RAPIDS Accelerator

- Operates as a software plugin to popular Apache Spark platforms
  - Automatically accelerates supported operations
  - Requires no code changes
- Operations currently accelerated
  - Spark SQL
  - DataFrame
- Works with Spark standalone, YARN clusters, Kubernetes clusters

Key Spark 3 innovations

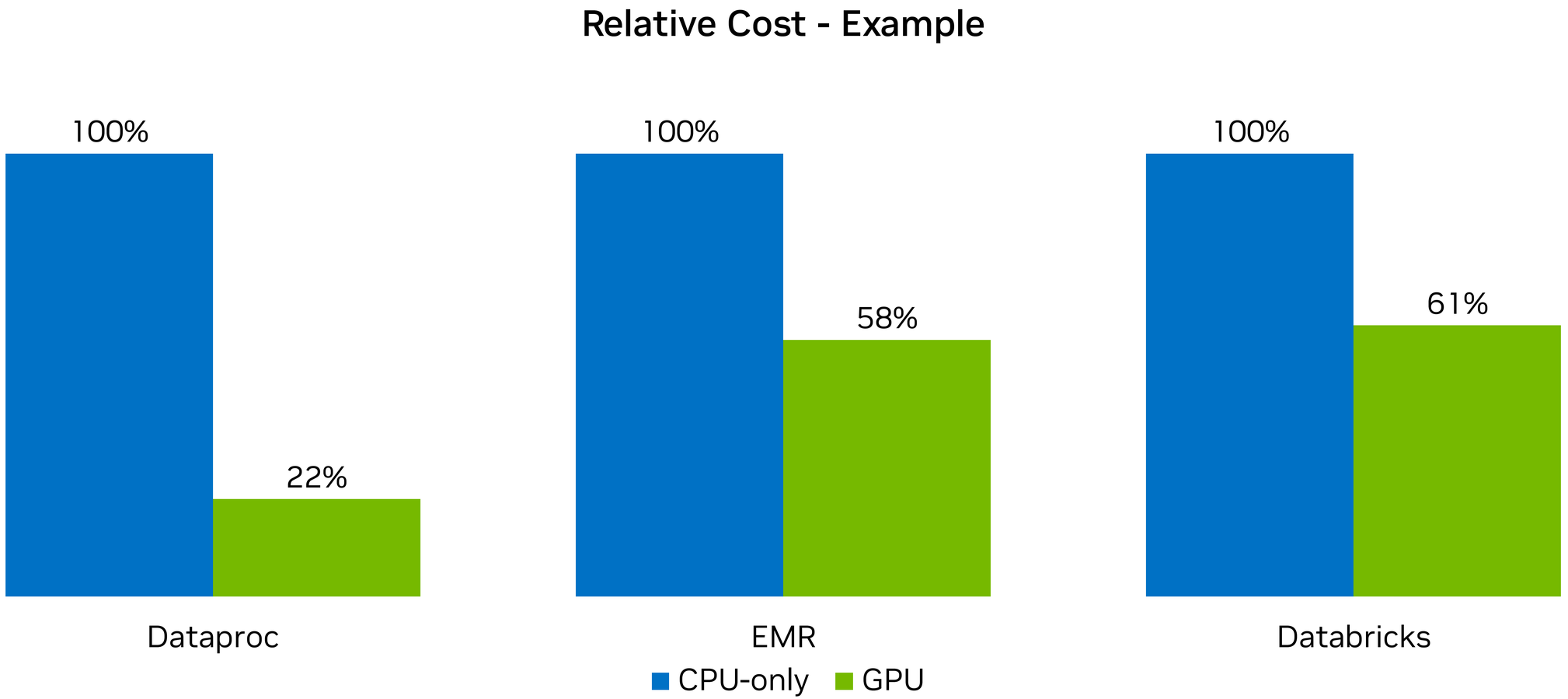
Columnar processing support in the Catalyst query optimizer – allows efficient GPU acceleration

GPU-aware scheduling of executors with a specified number of GPUs and how many GPUs for each task



Substantial Cost Savings

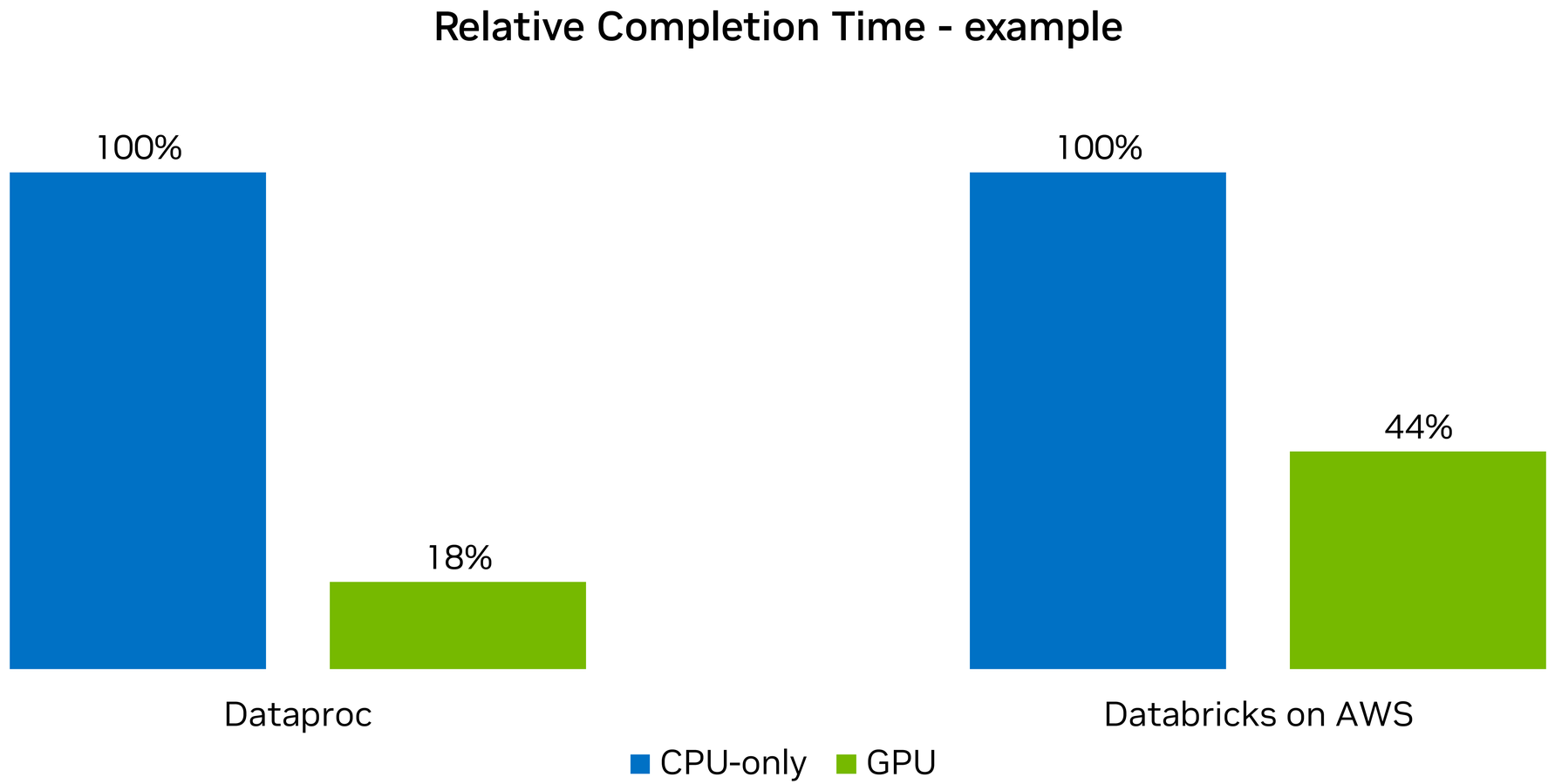
Equivalent work with less spend



Based on NDS benchmark
Does not include NVIDIA AI Enterprise license cost
Databricks on AWS

Faster Results

Allows more processing in the same time window



Based on NDS benchmark
Dataproc: 4x N1-highmem-32 vs. 4x N1-highmem-32 with 2 GPUs per instance
Databricks/AWS: 8x m6gd.2xlarge vs. 4x g5.8xlarge with 2 GPUs per instance

See Potential Savings on your own Apache Spark workloads

Quantify predicted time and cost savings and see results on jobs

Accelerated Spark Analysis Tool

- Analyze logs of existing Spark 2 or Spark 3 workloads to see time and cost savings estimates
- Also provides recommended configurations, and further optimizations based on initial run with GPUs

```
+--------------------+----------------------+-----------------+-----------------+---------------+-----------------+
| App Name           | Recommendation       |   Estimated GPU |   Estimated GPU |           App |   Estimated GPU |
|                    |                      |         Speedup |     Duration(s) |   Duration(s) |      Savings(%) |
+--------------------+----------------------+-----------------+-----------------+---------------+-----------------|
| Customer App #1    | Strongly Recommended |             3.7 |             651 |       2384.32 |              64 |
| Sales App #1       | Strongly Recommended |             3.1 |              89 |        281.62 |              58 |
| Sales App #2       |          Recommended |             2.1 |             447 |        939.21 |              58 |
| Customer App #2    |      Not Recommended |             1.6 |            1115 |       1783.65 |              38 |
+--------------------+----------------------+-----------------+-----------------+---------------+-----------------+

Report Summary:
----------------------------  ------
Total applications                 4
RAPIDS candidates                  3
Overall estimated speedup        3.0
Overall estimated cost savings   58%
```

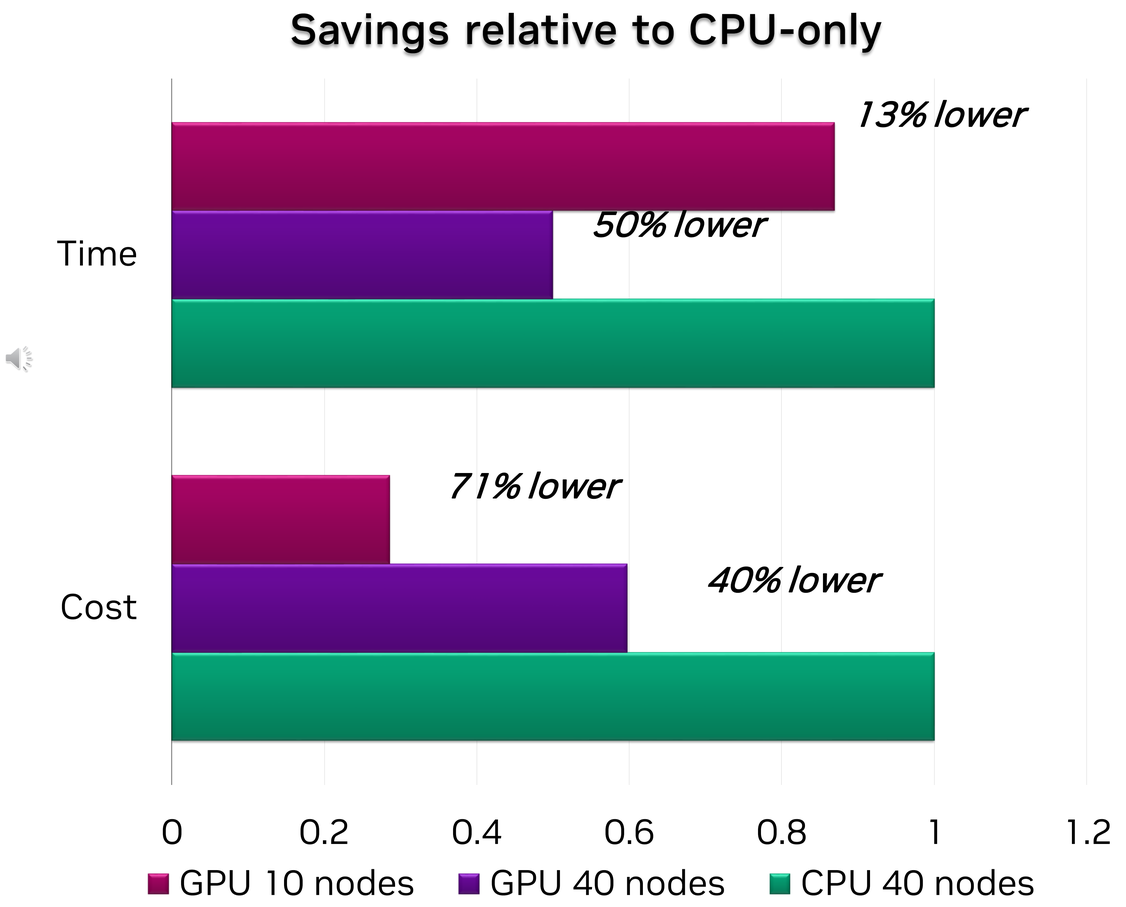
Customer Success Stories

Saving Time and Money in E-Commerce

Case Study

Large Retailer

- **Challenges**: increase sales in an increasingly online market
  - Internal tool rearranges online shelves based on price, popularity and other constraints, using a multiple stage ML pipeline that starts with ETL
  - Tool generated more than $300M in incremental revenue once implemented on Google Dataproc, but single run requires several hours
- **Solution**: RAPIDS Accelerator reduced job time to below one hour, while saving 70% in infrastructure costs
- **Outcome**:
  - Greater than $150K/year savings for this tool alone
  - Many other applications have similar potential

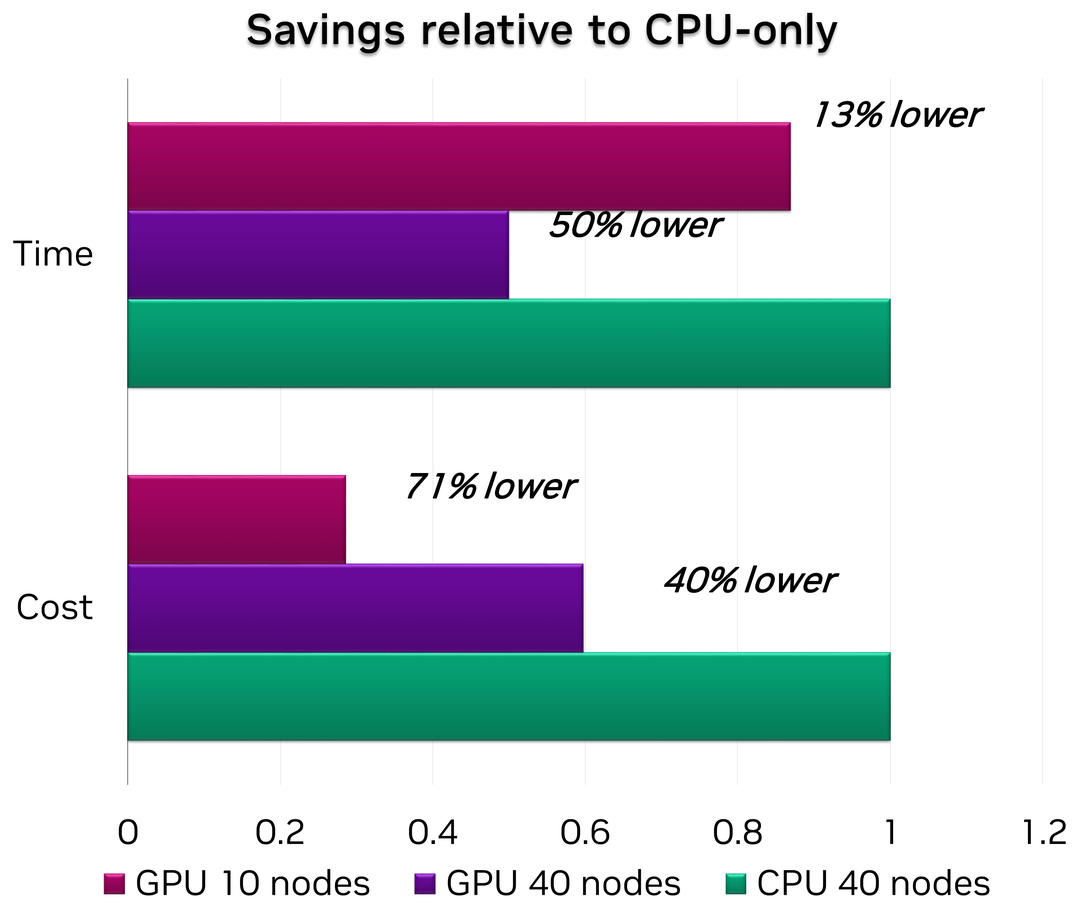


Saving Time and Money in E-Commerce

Case Study

Large Retailer

- **Challenges**: increase sales in an increasingly online market
  - Internal tool rearranges online shelves based on price, popularity and other constraints, using a multiple stage ML pipeline that starts with ETL
  - Tool generated more than $300M in incremental revenue once implemented on Google Dataproc, but single run requires several hours
- **Solution**: RAPIDS Accelerator reduced job time to below one hour, while saving 70% in infrastructure costs
- **Outcome**:
  - Greater than $150K/year savings for this tool alone
  - Many other applications have similar potential



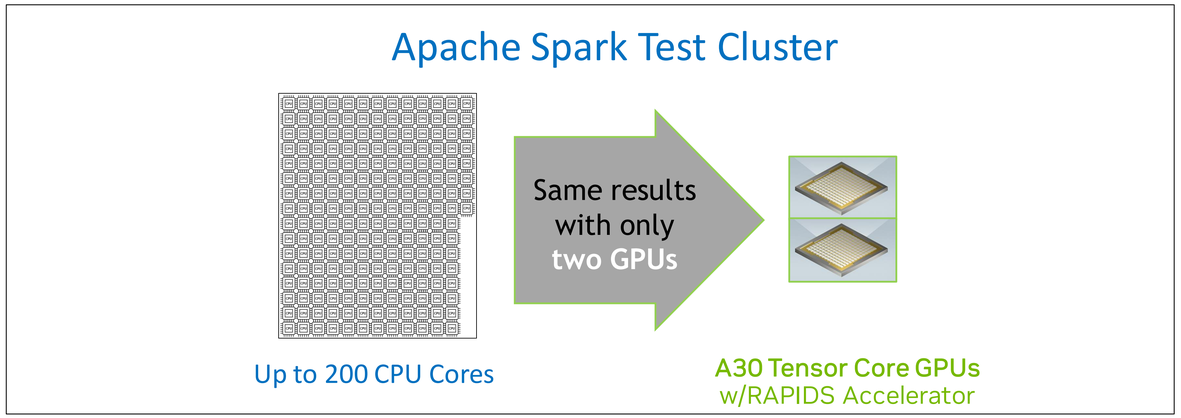
Taboola Optimizes Data Center Capacity and Cost

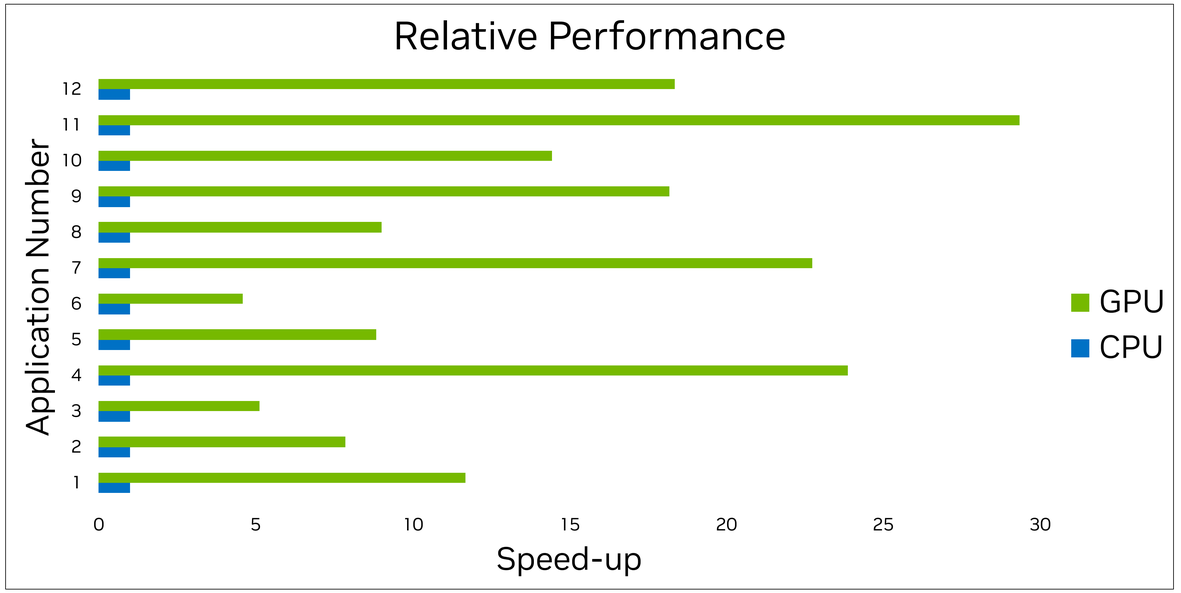
Case Study

- Most context-relevant webpage advertisements are served by Taboola's complex and compute-hungry data pipeline
- **Challenges**: Scaling capacity and minimizing cost for Apache Spark data pipelines
  - Frequent need to scale Apache Spark CPU cluster capacity to address constantly growing compute and storage requirements
  - Scaling CPU clusters was expensive
- **Solution**: RAPIDS Accelerator and A30 Tensor Core GPUs to accelerate data pipelines more cost-effectively than CPUs
- **Outcome**:
  - Greater scalability at lower cost
  - For some workloads, two A30 GPUs sustained the same production load as 200 CPU cores, and with greater energy efficiency

20X GPU Speed-up

Average measured across multiple workloads on Intel CPUs





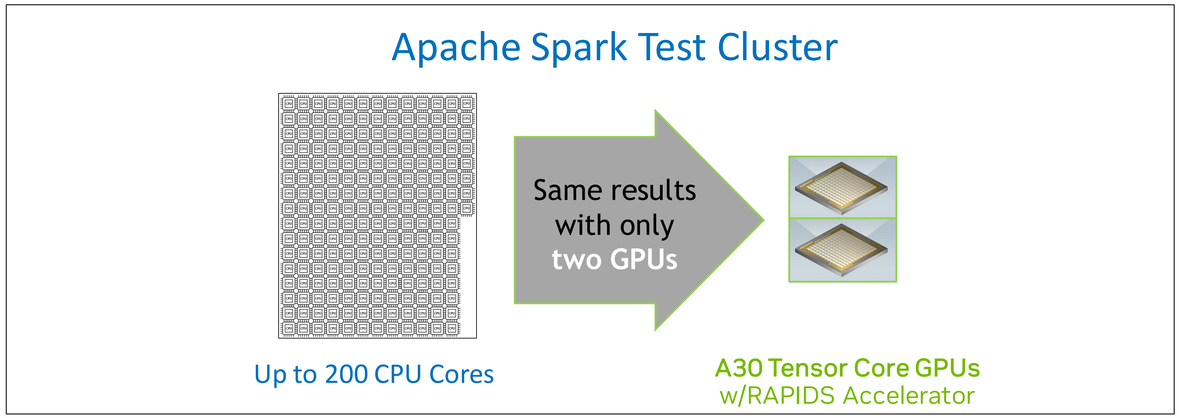
Taboola Optimizes Data Center Capacity and Cost

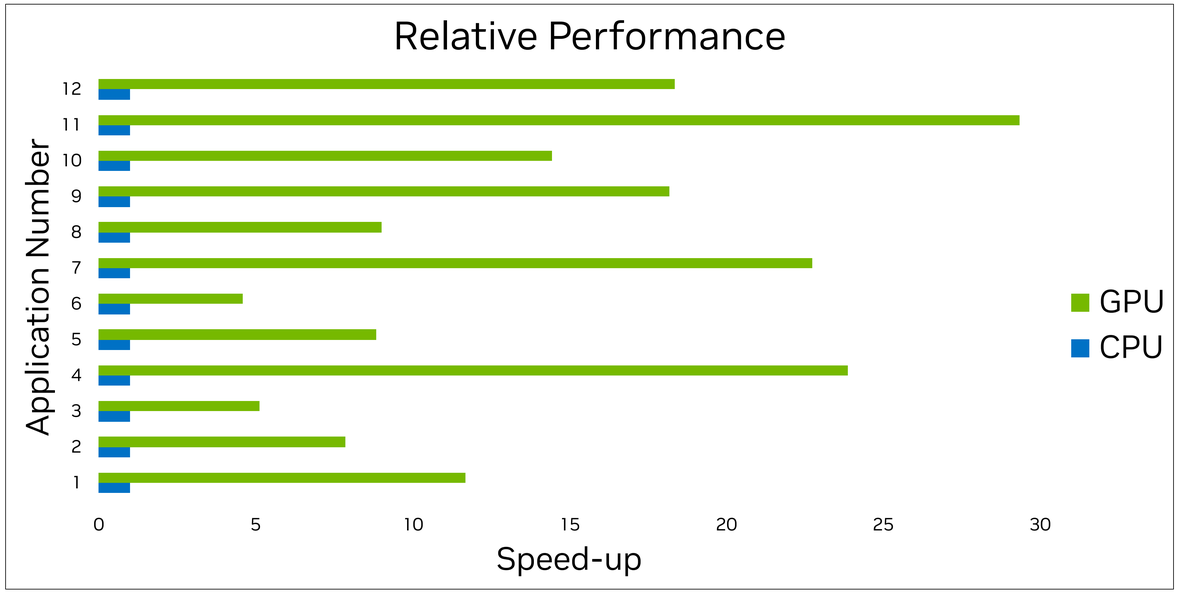
Case Study

- Most context-relevant webpage advertisements are served by Taboola's complex and compute-hungry data pipeline
- **Challenges**: Scaling capacity and minimizing cost for Apache Spark data pipelines
  - Frequent need to scale Apache Spark CPU cluster capacity to address constantly growing compute and storage requirements
  - Scaling CPU clusters was expensive
- **Solution**: RAPIDS Accelerator and A30 Tensor Core GPUs to accelerate data pipelines more cost-effectively than CPUs
- **Outcome**:
  - Greater scalability at lower cost
  - For some workloads, two A30 GPUs sustained the same production load as 200 CPU cores, and with greater energy efficiency

20X GPU Speed-up

Average measured across multiple workloads on Intel CPUs

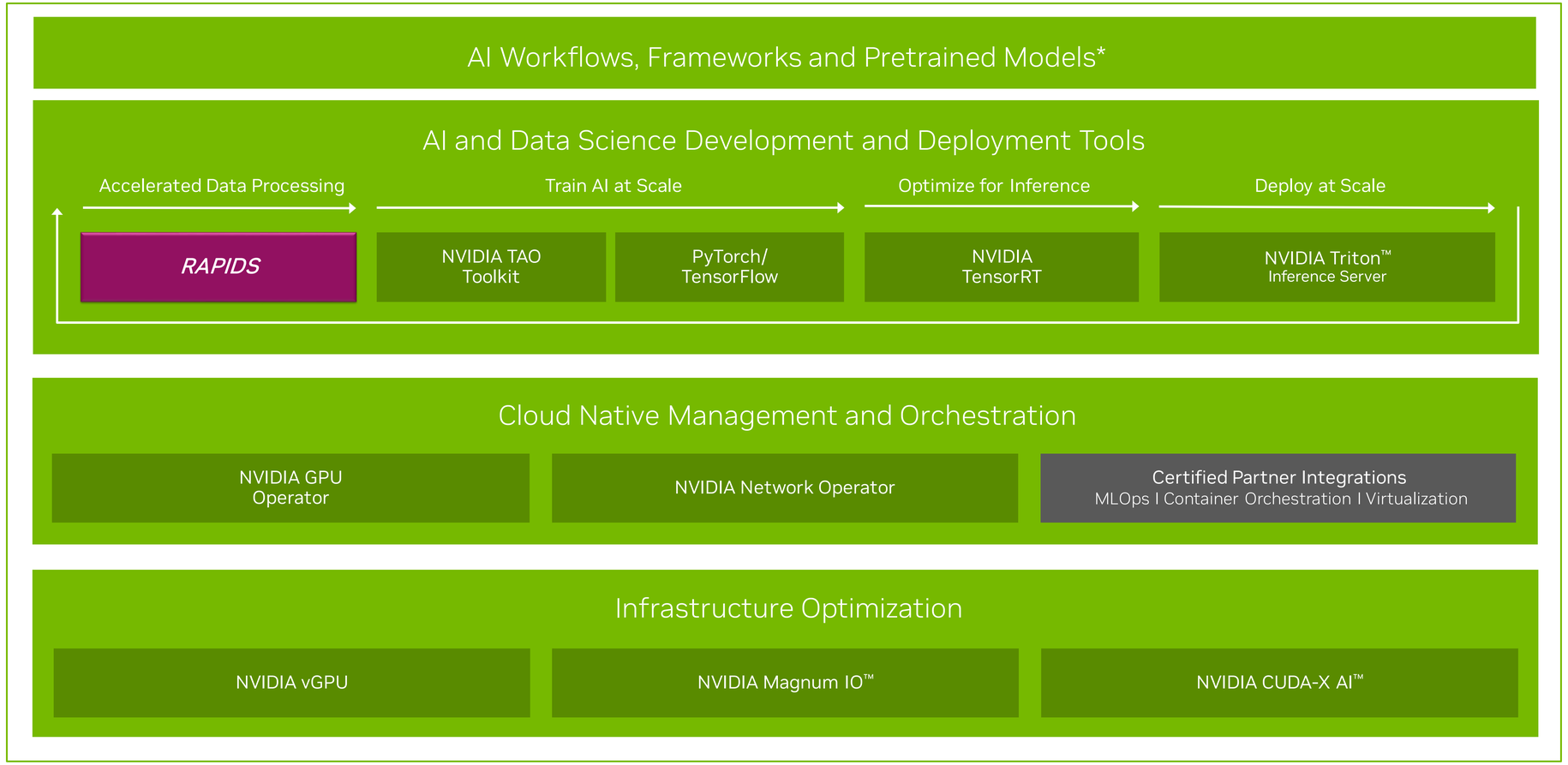




Enterprise Support and Services

NVIDIA AI Enterprise Software Suite

Includes Essential NVIDIA Software for Streamlined Development and Deployment



*NVIDIA NGC public catalog provides a complete listing of over 50 supported frameworks and pretrained models.

NVIDIA Enterprise Services

Delivering Customer Success

Support Services

Included Support

Business Standard Support
- ✓ Standard SLA
- ✓ Access to NVIDIA AI Experts
- ✓ Long Term Branch Support (LTS)
- ✓ Priority Notifications

Value Add Support

Business Critical Support
- ✓ 24/7 Support
- ✓ 1 hour response SLA

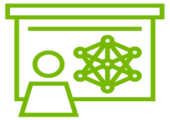
Technical Account Manager (TAM)
- ✓ Customer champion
- ✓ Designated support
- ✓ Performance reviews and support plans

Professional Services

Workload Onboarding
- ✓ Guidance and support to onboard workloads
- ✓ Specific use case instruction and best practices for running workloads

Custom Engagement
- ✓ Expert Guidance

Education Services



- ✓ NVIDIA AI Enterprise Administration Public Bootcamp
- ✓ Introduction to AI in the Data Center
- ✓ Accelerated Computing, Data Science, Deep Learning, and Graphics (Self-paced & Instructure-led)

Business Continuity

Speed to Value

Return on Investment

How to Purchase NVIDIA RAPIDS Accelerator

Sold as part of NVIDIA AI Enterprise

Bring Your Own License (BYOL)

- Apply existing entitlement to RAPIDS Accelerator
- Cloud BYOL allows one GPU per license
- EULA-based enforcement

Private Offer

Pay with committed cloud spend agreements

- Custom quote
- Long-term commitment

Customer can spend down existing cloud credits

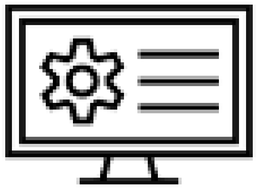
Getting Started

Perform Proof of Concept

Work with NVIDIA or partner Solution Architect

Qualification

Identify and estimate the cost savings and acceleration potential of your Spark workloads based on an analysis of the log files.



Bootstrap

Get recommended configuration parameters for GPU acceleration on your Spark cluster. Use them for initial run.

Tuning

See optimized configuration per application based on the initial (bootstrap) job run.

Sign up at nvidia.com/spark-tool


NVIDIA