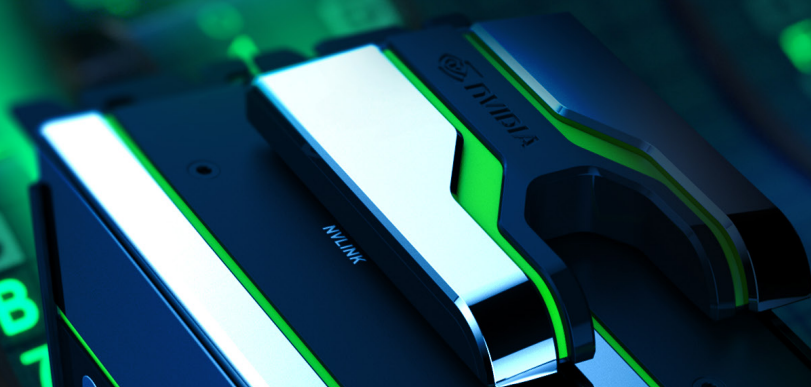




NVIDIA QUADRO[™]
AUTHORIZED PARTNER

PNY[®]

IMPROVING
INSIGHTS & OUTCOMES

NVIDIA Quadro RTX for Data Science and Big Data Analytics

2,500,000,000,000,000,000



With 2.5 quintillion (10¹⁸) bytes of data being generated every day, a trend that will only continue to exponentiate as the Internet of Things (IoT) undergoes explosive growth, information is everywhere while insights remain elusive. 90% of the world's data was created in the past two years alone. Sifting through this supersized data tsunami to find actionable information is the domain of Big Data Analytics. GPU acceleration, provided by the NVIDIA® Quadro® RTX™ line of professional graphics boards, is opening new vistas, possibilities, and capabilities for Big Data Analytics across academia, business, education, government, and numerous scientific disciplines.

Data scientists working on big data domain problems now have access to new technology that increases productivity while reducing time to insight. A portfolio of options ranging from big clusters with disk-based databases, small clusters with in-memory data, single systems with in-CPU memory data, and workstations with in-GPU memory data meet various Data Analytics use cases. This paper spans all options, but emphasizes GPU-based and accelerated solutions for single systems optimized for Data Science use and servers certified for data center deployment.

90%

of the world's data
was created in the
past two years.

010011100010111110110111010011100010111110110111010
11110110111101001111001111101101111010011110011111
1010100111000000000001100101010011100000000001100101
0001010000011011010000000000101000001101101000000000

Traditional Data Science workflows centered on an SQL server and other "Big Data" tools, usually running on clusters of CPU-only machines, and focused on manipulating data that was primarily stored on hard drives. Data scientists used to focus on data query, data manipulation, data access control, and data definition. Big-cluster Data Science is oriented towards efficient and well-defined data access and manipulation.

As computing power increased, and with the advent of high-DRAM single machines, it became increasingly feasible to run other more complex tasks in real-time on big datasets: exploratory analysis, data wrangling, building machine-learning models, etc. The flexibility and speed that in-memory Data Science tasks afford enabled data scientists to probe datasets for many more interesting questions, and experience a quick turnaround in terms of the time it takes to get the answers.



Today, state-of-the-art databases use in-GPU-memory data manipulation to get the basic database-like tasks accomplished in a matter of seconds on even moderately large datasets (several hundreds of GB). Just a generation ago these tasks would have taken days, if not weeks, to perform.

Approaches based on NVIDIA Quadro RTX, which features multi-precision Tensor Cores ideally suited for accelerating AI and Big Data Analytics, and unprecedented GPU memory capacity, are surrounded by an unrivaled ecosystem of NVIDIA and ISV development tools and applications to realize the promise of Big Data Analytics.

NVIDIA's RTX (Turing architecture) products can utilize a large collection of libraries that enable Data Scientists to easily scale their workflow from single small machine setups to an arbitrary combination of CPUs, GPUs, and a number of machines. DASK can facilitate most common Python libraries for work on multiple machines, including clusters of GPU-equipped ones; H2O Sparkling Water scales the H2O library API to a Spark cluster; and most modern Deep Learning libraries are capable of handling big data by default.

NVIDIA Quadro RTX Tensor Cores are used primarily for speeding up general matrix-to-matrix multiplications, which are an essential part of many deep learning algorithms. These algorithms are part of the NVIDIA cuDNN C++ library that is utilized by the popular high-level frameworks such as TensorFlow and PyTorch, and also fully support Python.

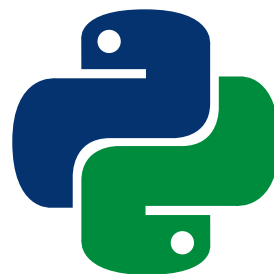
Most of the Python Data Science and Deep Learning libraries and frameworks have gone through several major iterations, and have become very mature and stable across different systems. Installing and running TensorFlow, PyTorch, Chainer, RAPIDS, etc. is just a matter of running a single install command. There is also increasing availability of NVIDIA Quadro RTX workstations

and laptops that have been specifically built with Data Science work in mind. Most of these machines come with preinstalled Data Science software stacks, like the NVIDIA-Powered Data Science Workstation, which provides practicing data scientists a turnkey solution that lets them start working on their projects as soon as they turn their machines on!

GPU memory capacity also plays an essential role in terms of batch size and training time for deep learning. The biggest advantage of having more GPU memory for deep learning is that it enables you to train ever larger, more accurate, models. The increasing value of bigger DL models is contingent on the size and quality of the dataset: bigger datasets can help bigger models learn more intricate representation of data without overfitting.

Large batch size is generally beneficial, especially if you use a lot of varied data augmentation. However, in many instances, a relatively smaller batch size can act as an additional regularization for model training.

As a rule of thumb, larger batch sizes also lead to faster training times, as the training process can go through the dataset in a fewer number of steps. Even so, the relationship between the number of steps in the training process and the training time is not linear.



As the power of Data Science solutions, frameworks, and applications has increased, so has the complexity of managing and maintaining them. All data scientists have horror stories about how they had to wear DevOps and SysAdmin hats, while still trying to be productive with their core job duties. A completely turnkey system that works right out of the box, and is easy to upgrade and maintain, enables data scientists to work smoothly and with minimum core workflow interruption. This has led NVIDIA and PNY, in conjunction with OEM partners, to deliver a completely turnkey GPU-accelerated solution based on NVIDIA Quadro RTX 8000 and RTX 6000 products - the NVIDIA-Powered Data Science Workstation - that lets Data Science workers focus on advancing the frontiers of research, not spending time spinning wheels because a software compatibility issue in a single library disrupted an entire workflow.

The NVIDIA-Powered Data Science workstation is a system specifically configured with hardware components that deliver a balanced systems architecture and software stack to handle the complex tasks that large datasets in Data Science require. Built to an NVIDIA defined reference platform by select PNY certified OEM partners, a base system utilizes an Intel Xeon Silver 4114 CPU, 192 GB to 384 GB of DRAM, 1 TB or 2 TB NVMe SSD storage in addition to NVIDIA Quadro RTX GPUs. GPU selection is based on the dataset sizes a data scientist will be working with. A single Quadro RTX 6000 provides 24 GB or 48 GB in a dual GPU configuration. NVIDIA's Quadro RTX 8000 with 48 GB of GPU memory, 96 GB in a dual GPU configuration, can handle even larger datasets. Dual GPUs are linked with NVIDIA NVLink, a 100 GB/sec bandwidth bi-directional bus between pairs of GPUs for the highest NN training performance. If you need double precision floating point training dataset support, you should utilize the NVIDIA Quadro GV100 with 32 GB of GPU memory, or 64 GB when configured in pairs with NVLink for high-performance NN Training.

NVIDIA CUDA-X AI is software foundation for the NVIDIA-Powered Data Science Workstation that opens up Data Analytics, machine learning and deep learning to massive speed boosts when utilizing NVIDIA Quadro RTX 6000, RTX 8000, or GV100 GPUs. It accelerates Data Science from ingest of data, to ETL, to model training, to deployment. Machine learning algorithms for regression, classification and clustering are also provided. Every major deep learning training framework has been optimized for NVIDIA Tensor Core equipped GPUs. Inference is also GPU accelerated and large-scale Kubernetes cloud deployment provides ultimate scalability. NVIDIA CUDA-X AI accelerates data analysis with cudDF, deep learning primitives with NVIDIA cuDNN, machine learning algorithms with NVIDIA cuML, and data processing with DALI, among others. The RAPIDS suite of software libraries, built on NVIDIA CUDA-X AI, gives you the freedom to execute end-to-end Data Science and analytics pipelines entirely on GPUs. Together these NVIDIA CUDA primitives and libraries accelerate every step in a typical AI workflow, whether it involves using deep learning to train speech and image recognition systems or Data Analytics to assess the risk profile of a mortgage portfolio. With NVIDIA CUDA-X AI and the NVIDIA-Powered Data Science Workstation the possibilities are boundless.



In addition to Tier 1 OEMs, specialized system builders like AMAX, COLFAX, BOXX, EXXACT, Images & Technologie, Microway, RAVE Computer, OSS and THINKMATE have been authorized to deliver turnkey NVIDIA-Powered Data Science workstations with all of the hardware and software capabilities we've discussed. Additionally, NVIDIA offers support agreements to ensure that data scientists spend their time doing research, not recompiling or troubleshooting.

Use cases that deliver significant business advantages through Data Science and Big Data Analytics abound. A major telecommunication carrier had to ensure network reliability by monitoring 10 billion records per week, yet their existing analytics solution couldn't handle the volume or velocity of the data. This delayed reporting of degradations and outages, and even forced the team to manually look through these records to spot trends.

OmniSci, a leader in Big Data Analytics software, built on NVIDIA technology and powered by Quadro RTX GPUs, enables this telco to process billions of rows of data in real-time, and to monitor network performance to maintain the network's high standards of reliability. The ease of use and installation facilitated adoption throughout the company and encouraged analysts to further explore their data.

The results are impressive; reports that once took over 20 minutes are produced in seconds. Engineers spot problems immediately, instead of 45 minutes later. This telco is now expanding the use of OmniSci to ensure quality of service during time-sensitive events, as well as data forensics to spot rogue users and devices.

Another example involves a pioneer in the audience-based, advanced TV advertising industry. Its goal is to improve the effectiveness and efficiency of TV advertising in an increasingly fragmented media landscape.

To achieve this, they process over 300 million viewing events per day, aggregated over 20 different sources. The company's existing spreadsheet-based analytics applications couldn't handle the influx of data. This delayed its time to insight and constrained the ability of national advertisers to see the effectiveness of their ad campaigns and spend. Again, OmniSci was chosen for its ability to handle data at extreme speed and scale, and because it allowed them to visualize and interact with their data to uncover hidden insights. Advertising reports, which once took over 40 minutes to produce, are now returned in milliseconds. The company now approaches prospective clients with a better understanding of national advertising, and identifies new opportunities for its clients to optimize their ad spend.

Another company, Graphistry, brings a human interface to big and complex data, by automatically transforming data into interactive, visual investigation maps built for the needs of analysts. It quickly surfaces relationships between events and entities without writing queries or wrangling data, harnesses all of your data without worrying about scale, and pivots on the fly to follow anywhere your investigation leads. Ideal for everything from security, fraud, and IT investigations to 3600 views of customers and supply chains, Graphistry turns the potential of your data into human insight and value, powered and accelerated by NVIDIA Quadro RTX.

A particularly relevant, but by no means the only, area where Graphistry excels is giving investigators the tools they need to drive fast, conclusive investigations into fraud, abuse, money laundering, and other financial crimes. Graphistry brings transactions, traffic logs, risk score APIs, account databases, and even CSV files into a unified investigative content. The platform remains fully interactive even at large scale, allowing investigators to see clusters, outliers, and links between entities, while always retaining the freedom to drill deeper and pivot to new data sources.

Battling retail fraud accounts of 7.6% of revenue for online retailers and existing detection tools simply aren't enough. With Graphistry, powered by NVIDIA Quadro RTX, analysts can quickly connect the dots across data sources to identify both account creation and transactional fraud. With the intelligence to see the traits that distinguish valid accounts and transactions from fraudulent ones. From classifying incidents to surfacing underlying fraud rings to tuning rules, GPU-accelerated Graphistry brings visibility and productivity across analyst tiers.

While still hospitalized patients face a number of potential threats to their recovery, including the development of sepsis, the acquisition of an antibiotic-resistant infection, or a sudden downturn due to pre-existing medical conditions (e.g. COVID-19 and BMI or diabetes). Big Data Analytics can help providers react as quickly as possible to changes in a patient's status, and can even identify an upcoming deterioration before symptoms clearly manifest themselves, even to experienced diagnosticians.

This is a small subset of the capabilities Big Data Analytics, powered by Quadro RTX GPUs, brings to Data Science and analytics professions. If you'd like to learn more, email us at gopny@pny.com and we will answer any questions you have. We also provide additional information on NVIDIA-Powered Data Science Workstations at www.pny.com/datascienceworkstations.