Interview: Global Technology Leader PNY

June 8, 2020 *by Editorial Team*

We recently caught up with our friends over at PNY to discuss a variety of topics affecting data scientists conducting work on big data problem domains including how "Big Data" is becoming increasingly accessible with big clusters with disk-based databases, small clusters with in-memory data, single systems with in-CPU-memory data, and single systems with in-GPU-memory data. Answering our inquiries were: Bojan Tunguz, Senior System Software Engineer, NVIDIA and Carl Flygare, NVIDIA Quadro Product Marketing Manager, PNY.

insideBIGDATA: Data scientists conducting work on big data problem domains have seen massive projects become more accessible with the availability of a range of specialized data science hardware and software platforms. Can you provide us with an overview of these different architectures, e.g. big clusters vs. single machines, CPUs vs. GPUs, various memory configurations?

PNY: 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.

insideBIGDATA: Please describe the typical software stack (NVIDIA CUDA-X AI, RAPIDS, etc.) a data scientist might utilize for big data applications? What would be the primary subsystems, for example, that would be optimized to utilize NVIDIA® Quadro® RTX™ Tensor Cores?

PNY: There is 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.

insideBIGDATA: How easy would it be for the above software stack to interact with the major frameworks like TensorFlow, Pytorch, Chainer, etc.?

PNY: 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 an increasing availability of 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 enables practicing data scientists to start working on their projects as soon as they turn their machines on!

insideBIGDATA: What role does GPU memory capacity play in terms of batch size and training time for deep learning?

PNY: 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.

insideBIGDATA: What value do completely turnkey systems offer to data scientists that want to unbox a system and go straight to useful data science work?

PNY: 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, 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.

insideBIGDATA: You have mentioned the NVIDIA-Powered Data Science workstation. What is it?

PNY: It's 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 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 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 Quadro GV100 with 32 GB of GPU memory, or 64 GB when configured in pairs with NVLink for high performance NN Training.

insideBIGDATA: What software does the NVIDIA-Powered Data Science Workstation utilize?

PNY: NVIDIA CUDA-X AI is software 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 A accelerates data analysis with cdDF, 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 GPIs. 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.

insideBIGDATA: Who offers NVIDIA-Powered Data Science Workstations and what about support?

PNY: 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 solutions with all of the hardware and software capabilities we've discussed. NVIDIA offers support agreements to ensure that data scientists spend their time doing research, not recompiling or troubleshooting.

insideBIGDATA: I'm interested and would like to learn more. How can I do so?

PNY: 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

NVIDIA QUADRO AUTHORIZED PARTNER PNY