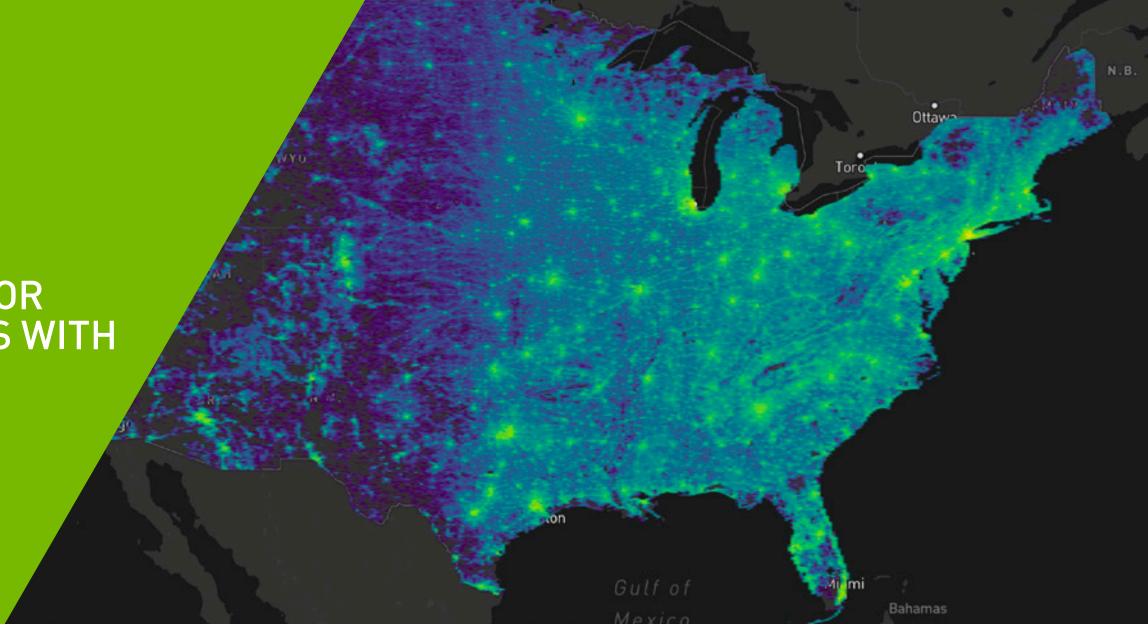
SERVER-SCALE EXPERIMENTATION FOR DATA SCIENCE TEAMS WITH

NVIDIA DGX STATION A100 AND RAPIDS



The modern enterprise is data-driven. With the rise of data science, enterprises across all industries are evolving to leverage large-scale datasets and better understand customer and market behaviors. This data-driven evolution has allowed businesses to tailor products, services, and decisions to meet market needs faster than ever.

While data science is invaluable for unlocking business potential and creating long-term value, generating business insights is a significant undertaking with substantial upfront costs and overhead. These pain points are felt early on in the modeling process when data science teams experiment to determine the proper path forward. Due to limited compute resources, complex software environments, and unfamiliar big data tools, data science teams often begin the arduous journey to valuable insights on the wrong foot.

Data science teams now have a new tool to address these challenges, offering server-scale computing that fits in an office setting without the need of a data center. NVIDIA DGX Station™ A100 with RAPIDS™ offers a purpose-built platform that gives data science teams the power to unleash insights, without being dependent on massive IT infrastructure. Now, data science teams can leverage familiar and intuitive tools to power massive-scale, collaborative experimentation. With a blend of accessibility and performance, DGX Station A100 is the supercharged data science workhorse for teams who need to shorten the journey to insights.

Innovation Stifled by Cumbersome Experimentation

As enterprises continue a data-driven evolution, practices and technologies must evolve as well. That evolution begins with enabling data science teams early in the modeling process to build innovative solutions.

Even in the most forward-thinking organizations, experimentation is a cumbersome process, especially when collaborating with a team. Data practitioners primarily rely on local laptops, cloud instances, or highly in-demand data center infrastructure to experiment with datasets. While all these avenues have their benefits, they are often accompanied by hindrances that limit innovation and progress:

- **Laptops** offer a comfortable experience for practitioners but make it incredibly difficult to share software environments with collaborators and leverage large datasets to build accurate models.
- **Cloud instances** provide flexibility and power but come at a substantial cost that decreases the return on investment of data-driven operations.
- **Data center infrastructure** can power high-performance experimentation but is usually a shared resource under significant demand and can present challenges when managing development environments.

These problems also present substantial challenges when data science teams are ready to deploy a production-ready model. Due to complex software environments and disparate experimentation and production infrastructure, deploying models regularly leads to comprehensive refactors, further lengthening the time to insight generation.

Collaborative Experimentation on an AI Supercomputer

NVIDIA DGX Station A100 brings AI supercomputing to data science teams, offering data center performance without a data center or additional IT investment. Designed for multiple, simultaneous users, DGX Station A100 leverages server-grade components in an easy-to-manage workstation form factor. It's the only system with four fully interconnected NVIDIA A100 Tensor Core GPUs—with up to 320 gigabytes (GB) of GPU memory and support for Multi-Instance GPU (MIG)—that can plug into a standard power outlet in the office or at home, resulting in a powerful AI appliance that can be placed anywhere.

DGX Station A100 provides a powerful solution to the problems faced by enterprise data science teams, especially during experimentation. Not only is it powerful, flexible, and easy to set up, DGX Station A100 offers an on-ramp for enterprises starting their journey in high-performance data science to evolve into a sophisticated AI-driven enterprise.

High-Performance Data Science Using RAPIDS

RAPIDS is a collection of open-source tools that accelerates highly popular data science tools and allows users to execute end-to-end data science operations completely on GPUs. While it relies on NVIDIA CUDA® primitives for low-level compute optimization, RAPIDS exposes GPU parallelism and high memory bandwidth through user-friendly Python and Java/Scala interfaces. By following and extending open-source standards, RAPIDS can be used as a "drop-in" replacement for pre-existing code for immediate performance gains with additional performance gained from purpose-built code.

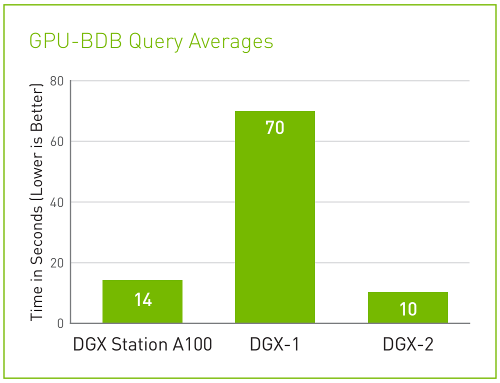
When pairing RAPIDS and DGX Station A100, data science teams can quickly take advantage of NVIDIA accelerated computation to conduct terabyte (TB)-scale experiments and produce innovative insights with ease.

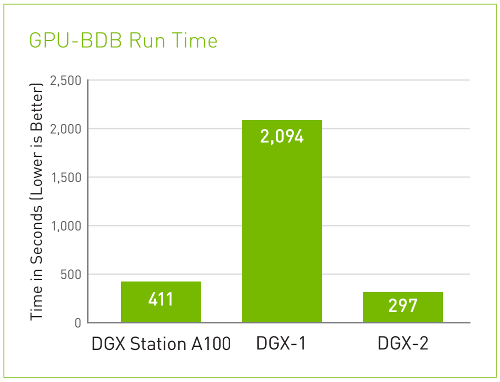
Near Real-Time Experimentation with DGX Station A100 and RAPIDS

To test our products, NVIDIA has built a standard benchmark, the **GPU Big Data Benchmark (GPU-BDB)**, to mimic real operations at a typical large retail or finance company. It consists of 30 queries requiring large-scale extract, transform, and load (ETL) operations, natural language processing, and machine learning using a mixture of structured and unstructured data at 1TB or 10TB scale. The benchmark is evaluated "end-to-end," meaning data starts and ends on disk, and everything in between is measured.

Performing the 1TB benchmark on DGX Station A100 with RAPIDS, each query was completed in an average of **13.7 seconds** with a total runtime of **5 minutes**. Running the same workflow on NVIDIA DGX-1® server, the average query-completion time was 68.7 seconds and total runtime was 35 minutes. When compared with DGX-1 server, DGX Station A100 is **5X more performant**.

Beyond delivering server-grade performance in a compact, office floor-friendly package, DGX Station A100 offers near real-time interactivity for data science teams solving challenging business problems. This level of interactivity makes it easy for teams to get more done in less time and helps enterprises get the most value out of their data.





Interested in supercharged experimentation with DGX Station A100 and RAPIDS?

Learn more about DGX Station A100: **www.nvidia.com/DGXStationA100**

Learn more about RAPIDS: **www.rapids.ai**

Contact an NVIDIA DGX sales representative: **www.nvidia.com/en-us/contact/sales/**