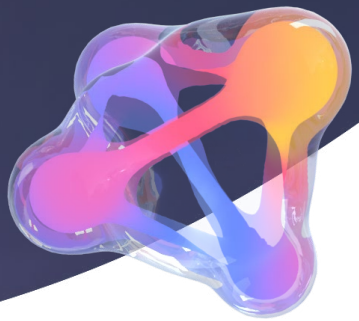
Iguazio's GPU-as-a-Service Powered by its Data Science Platform on NVIDIA DGX Systems

Enabling greater GPU efficiency, addressing larger scale workloads, and providing end to end ML automation



OVERVIEW

GPUs play an important role in improving performance and scalability of data science projects. However, GPUs are often under-utilized due to inefficient resource allocation, data bottlenecks, and complicated DevOps. In a cloud-native era, data scientists need a way to manage and control GPU resources as part of both training and inferencing layers - a solution that results in cost savings and reduced time to deployment.

The NVIDIA DGX - Ready Software program delivers proven enterprise-grade solutions that increase data science productivity, accelerate AI workflow, and improve accessibility and utilization of AI infrastructure. Iguazio's GPU-as-a-Service solution helps enterprises make more efficient use of their data science infrastructure, reduce complexity, and accelerate time to impact of AI projects. By being certified as part of the NVIDIA DGX-Ready Software program, the Iguazio Data Science Platform is helping democratize access to AI infrastructure and unleash the full potential of AI for every enterprise.

INDUSTRY CHALLENGES

Inefficient Resource Sharing

When multiple training jobs and data science exploration are running in parallel, data scientists are competing over GPU resources. Without proper admission control, jobs will be waiting in the queue because the GPU resources are not free, while other GPUs sit idle. Often the overall GPU usage is not optimized and not aligned with business needs.

Limited Data Access

Data intensive applications may require large amounts of data to process in the GPU. Since not all data can be uploaded to the GPU at once, there is a need for fast data access and smooth integration between the data layer and the GPU.

Inability to Run Training Jobs at Scale

Many data scientists write their code in Python, However, Python is a single-threaded language and therefore the typical process written in Python is not designed to run at scale. This poses a problem, especially when scale and performance are required, for example in real-time usecases.

The Need for an End-to-End Platform

Before enterprises can get to the point where they are running training and inference on GPUs, they would need to build an environment for data scientists and data engineers to collaborate on a full data science lifecycle: collect, explore, train and deploy models. This will require integration with various frameworks (e.g. Horovod, RAPIDS) and involves heavy lifting by the devops team, requiring extensive time and resources to build and maintain.

A COMPREHENSIVE SOLUTION FOR MORE EFFICIENT ML WORKLOADS & OPERATIONS

NVIDIA DGX POD provides an enterprise-grade AI infrastructure solution that eliminates design complexity, accelerates deployments and simplifies on-going operations.

As a validated ISV solution for NVIDIA DGX POD, the Iguazio Data Science Platform enables data scientists and engineers to work in one end-to-end data science platform, which includes a comprehensive solution for GPU sharing.

The Iguazio Data Science Platform takes full advantage of NVIDIA DGX Systems, allowing customers to utilize GPU-as-a-Service and NGC containers, making more efficient use of computational resources, saving costs and reducing infrastructure complexities. The solution automates pipelines across machine learning, deep learning and data analytics. The end result is a scalable way of processing data and computation.

The solution frees up resources when jobs are done. It also includes the ability to monitor GPU usage – not just on the infrastructure level but also on the application level, down to the individual user and specific GPU. With built in frameworks like Horovod for running deep learning (DL) on GPUs, and a built-in integration with cuDF, the solution ensures minimal devops, no matter how complex the computation.

IGUAZIO'S DATA SCIENCE PLATFORM FEATURING GPU-AS-A-SERVICE

THE FIRST SERVERLESS PLATFORM SUPPORTING GPUs

Serverless and Kubernetes target key challenges in data science: they simplify operationalization, eliminate manual devops processes and cut time to market. Iguazio's open source project, Nuclio, is the only serverless framework that enables enterprises to run any model on GPUs as a serverless function, accelerating performance and scalability of AI applications.

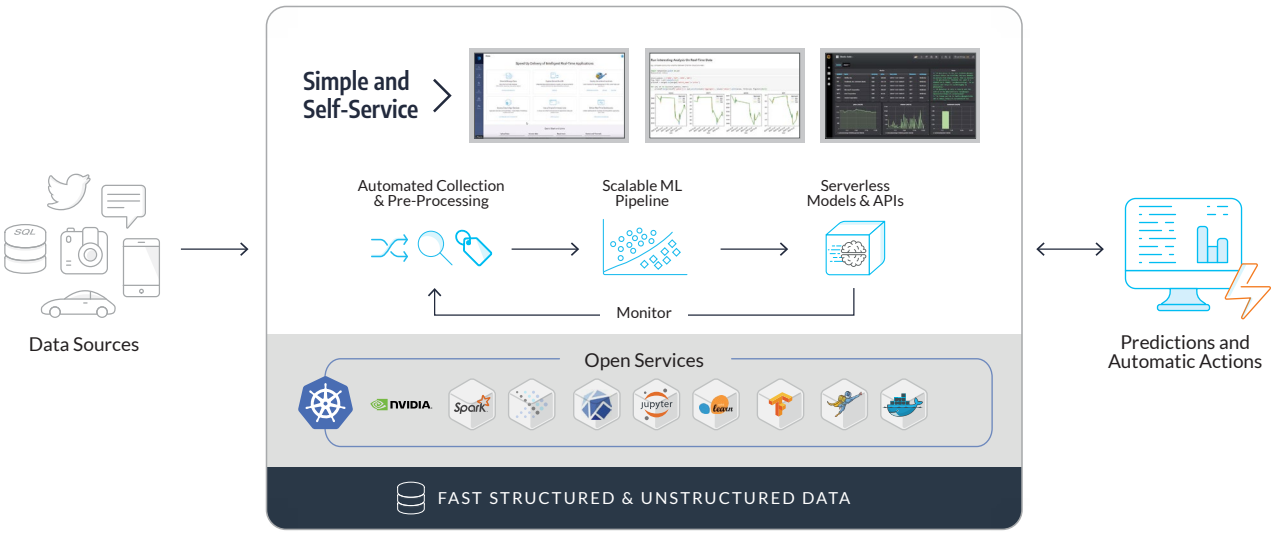
NO LIMIT SCALE OUT

Iguazio is integrated with NVIDIA's RAPIDS open-source machine learning libraries, for faster and scalable data processing. It enables:

- Direct writes/reads into/from the GPU's memory using RAPIDS data frames
- Streaming data in chunks directly into GPUs
- Full parallelism - multiple nodes read data and specific shards are allocated to specific GPU processors
- Predicate push down - offloading queries and analytics to Iguazio

HIGH SPEED DATA FABRIC FOR GPU OPTIMIZATION

The solution provides a microservice-based architecture for better CPU and GPU optimization running over a high-speed data fabric. This enables enterprises to scale GPUs in and out with serverless for maximum resource efficiency. In this way, data scientists can easily share GPUs in a collaborative, monitored environment and use them in an optimized way.



iguazio | NVIDIA

INDUSTRY INSIGHTS

Now, customers can accelerate machine learning, deep learning and data processing using a single GPU as a Service solution, for faster and scalable AI-based applications. This allows them to create business impact faster with their AI applications, with reduced costs and a simplified process.

FINANCIAL SERVICES

Proactively detect threats before they occur, prevent fraud, money laundering, cyber threats and other risks to your business by understanding patterns of illicit behavior and acting to proactively prevent risks using a distributed and simplified process.

GAMING AND AD-TECH

Get a better understanding of your users by leveraging AI and real-time data. Anticipate their next move – on a granular basis – to reduce churn, increase engagement and provide the best user experience possible.

RETAIL

Harness historic and real-time data to produce the most accurate, personalized recommendations delivered at the right place and time. Build and deploy models to create accurate sales forecasts, and plan your inventory accordingly to increase revenue. Increase loyalty and create differentiation strategies that make a difference to your bottom line.

Find Out More:

The NVIDIA DGX-Ready Software program delivers proven enterprise-grade solutions that increase data science productivity, accelerate AI workflow, and improve accessibility and utilization of AI infrastructure. In combination with NVIDIA DGX POD implementations, this full stack solution provides an enterprise-grade AI infrastructure that enables enterprise IT to support the full lifecycle of AI development.

Website: www.nvidia.com/dgx-pod

Twitter: @NVIDIAAI

Blog: blogs.nvidia.com

The Iguazio Data Science Platform enables enterprises to develop, deploy and manage AI applications at scale. With Iguazio, enterprises can run AI models in real time, deploy them anywhere (multi-cloud, on-prem or edge), and bring to life their most ambitious AI-driven strategies. Enterprises spanning a wide range of verticals, including financial services, manufacturing, smart mobility and telecoms use Iguazio to create business impact through a multitude of real-time use cases such as fraud prevention, self-healing networks and location-based recommendations. Iguazio brings data science to life.

Website: www.iguazio.com

Contact: info@iguazio.com

Twitter: @iguazio

Blog: iguazio.com/blog