Cloud-based Data Science at the speed of thought using RAPIDS – the Open GPU Data Science ecosystem on Azur

Josh Patterson
GM of Data Science
NVIDIA

Keith Kraus
RAPIDS cuDF Lead
NVIDIA

Tom Drabas, PhD
Senior Data Scientist
Microsoft

Brad Rees, PhD
RAPIDS cuGraph Lead
NVIDIA

Bartley Richardson, PhD
RAPIDS Data Science Lead
NVIDIA

Juan-Arturo Herrera, PhD
Senior Data Scientist
Microsoft

Corey Nolet
RAIDS cuML Sr Developer
NVIDIA

Data science is the exploration of vast amounts of data to discover actionable knowledge. It is about finding answers to hard questions and involves trial and error over multiple iterations. The data scientist's job is not easy; after all, if you know how to find the answer then you do not need a data scientist. To increase productive, the data scientist uses tools within the Python ecosystem, such as Pandas, Numpy, and Scikit-learn. While Python is the defacto standard for data science, it suffers from poor performance. This make the data scientist's job hard since the amount of time they spend waiting for results interrupts their train of thought. We must strive to make the data scientist environment more productive. As larger and larger amount of data is required to be explored, the amount of time waiting for results also increases.

The RAPIDS suite of open source software libraries gives the data scientist the freedom to execute end-to-end data science and analytics pipelines on GPUs. RAPIDS is incubated by NVIDIA based on years of accelerated analytics experience. RAPIDS relies on NVIDIA CUDA primitives for low-level compute optimization and exposes GPU parallelism and high-bandwidth memory speed through user-friendly Python interfaces. Through a familiar DataFrame API that integrates with a variety of machine learning algorithms, RAPIDS facilitates common data preparations tasks while removing typical serialization costs. RAPIDS includes support for multi-GPU deployments, enabling vastly accelerated processing and training on large dataset sizes.

Join NVIDIA's engineers as they walk through a collection of data science problems that introduce components and features of RAPIDS, including: feature engineering, data manipulation, statistical tasks, machine learning, and graph analysis. This tutorial focuses on accelerating a large data science workflow in Python on a multiple GPU.