

NVIDIA Accelerates Apache Spark, World's Leading Data Analytics Platform

Open Source Community Accelerates Spark 3.0 with Native NVIDIA GPU Support; Lightning-Fast ETL and SQL Processing on Hundreds of Terabytes of Data; Adobe Achieves 7x Speedup in Model Training with Spark 3.0 on Databricks

GTC 2020 -- NVIDIA today announced that it is collaborating with the open-source community to bring end-to-end GPU acceleration to Apache Spark 3.0, an analytics engine for big data processing used by more than 500,000 data scientists worldwide.

With the anticipated late spring release of Spark 3.0, data scientists and machine learning engineers will for the first time be able to apply revolutionary GPU acceleration to the ETL (extract, transform and load) data processing workloads widely conducted using SQL database operations.

In another first, AI model training will be able to be processed on the same Spark cluster, instead of running the workloads as separate processes on separate infrastructure. This enables high-performance data analytics across the entire data science pipeline, accelerating tens to thousands of terabytes of data from data lake to model training, without changes to existing code used for Spark applications running on premises and in the cloud.

"Data analytics is the greatest high performance computing challenge facing today's enterprises and researchers," said Manuvir Das, head of Enterprise Computing at NVIDIA. "Native GPU acceleration for the entire Spark 3.0 pipeline -- from ETL to training to inference -- delivers the performance and scale needed to finally connect the potential of big data with the power of AI."

Building on its [strategic AI partnership with NVIDIA](#), Adobe is one of the first companies working with a preview release of Spark 3.0 running on Databricks. It has achieved a 7x performance improvement and 90 percent cost savings in an initial test, using GPU-accelerated data analytics for product development in Adobe Experience Cloud and supporting features that power digital businesses.

The performance gains in Spark 3.0 enhance model accuracy by enabling scientists to train models with larger datasets and retrain models more frequently. This makes it possible to process terabytes of new data every day, which is critical for data scientists supporting online recommender systems or analyzing new research data. In addition, faster processing means that fewer hardware resources are needed to deliver results, providing significant cost savings.

"We're seeing significantly faster performance with NVIDIA-accelerated Spark 3.0 compared to running Spark on CPUs," said William Yan, senior director of Machine Learning at Adobe. "With these game-changing GPU performance gains, entirely new possibilities open up for enhancing AI-driven features in our full suite of Adobe Experience Cloud apps."

Databricks and NVIDIA Bring More Speed to Spark

Apache Spark was originally created by the founders of Databricks, whose cloud-based Unified Data Analytics Platform runs on over [1 million virtual machines every day](#). NVIDIA and Databricks have [collaborated to optimize Spark with the RAPIDS™](#) software suite for Databricks, bringing GPU acceleration to data science and machine learning workloads running on Databricks across healthcare, finance, retail and many other industries.

"Our continued work with NVIDIA improves performance with RAPIDS optimizations for Apache Spark 3.0 and Databricks to benefit our joint customers like Adobe," said Matei Zaharia, original creator of Apache Spark and chief technologist at Databricks. "These contributions lead to faster data pipelines, model training and scoring, that directly translate to more breakthroughs and insights for our community of data engineers and data scientists."

Faster ETL and Data Transfers in Spark with NVIDIA GPUs

NVIDIA is contributing a new open source [RAPIDS Accelerator for Apache Spark](#) to help data scientists increase the performance of their pipelines from end to end. The accelerator intercepts functions previously operated on by CPUs and instead uses GPUs to:

- Accelerate ETL pipelines in Spark by dramatically improving the performance of Spark SQL and DataFrame operations without requiring any code changes.
- Accelerate data preparation and model training on the same set of infrastructure, where a separate cluster is not required for machine learning and deep learning.
- Accelerate data transfer performance across nodes in a Spark distributed cluster. These libraries leverage the open source Unified Communication X (UCX) framework of the UCF Consortium and minimize latency by enabling data to move directly between GPU memory.

[A preview release of Spark 3.0](#) is now available from the Apache Software Foundation, with general availability expected in the coming months. More information is available at www.nvidia.com/spark.

About NVIDIA

NVIDIA's (NASDAQ: NVDA) invention of the GPU in 1999 sparked the growth of the PC gaming market, redefined modern computer graphics and revolutionized parallel computing. More recently, GPU deep learning ignited modern AI — the next era of computing — with the GPU acting as the brain of computers, robots and self-driving cars that can perceive and understand the world. More information at <https://nvidianews.nvidia.com/>.

Certain statements in this press release including, but not limited to, statements as to: NVIDIA and the open source community collaborating and accelerating Apache Spark; the anticipated release of Spark 3.0 and it enabling the GPU acceleration to ETL data processing workloads using SQL database operations; AI model training being able to be processed on the Spark cluster and enabling high-performance data analytics; the benefits, performance and abilities of our products and technologies, including GPU acceleration for Spark 3.0; data analytics being the greatest high performance computing challenge and native GPU acceleration for the Spark 3.0 pipeline being able to deliver the performance and scale needed to connect big data with AI; the performance and benefits from Adobe working with Spark 3.0 running on Databricks, including with RAPIDS; Spark 3.0 enabling scientists to train models with larger datasets, retrain models more frequently, process terabytes of data, requiring fewer hardware resources to deliver results and cost savings; the new possibilities opening up for AI-driven features in the Adobe Experience Cloud apps based on NVIDIA-accelerated Spark 3.0; the benefits and performance from NVIDIA's contributions to open source RAPIDS Accelerator for Apache Spark; and the availability of Spark 3.0 are forward-looking statements that are subject to risks and uncertainties that could cause results to be materially different than expectations. Important factors that could cause actual results to differ materially include: global economic conditions; our

reliance on third parties to manufacture, assemble, package and test our products; the impact of technological development and competition; development of new products and technologies or enhancements to our existing product and technologies; market acceptance of our products or our partners' products; design, manufacturing or software defects; changes in consumer preferences or demands; changes in industry standards and interfaces; unexpected loss of performance of our products or technologies when integrated into systems; as well as other factors detailed from time to time in the most recent reports NVIDIA files with the Securities and Exchange Commission, or SEC, including, but not limited to, its annual report on Form 10-K and quarterly reports on Form 10-Q. Copies of reports filed with the SEC are posted on the company's website and are available from NVIDIA without charge. These forward-looking statements are not guarantees of future performance and speak only as of the date hereof, and, except as required by law, NVIDIA disclaims any obligation to update these forward-looking statements to reflect future events or circumstances.

© 2020 NVIDIA Corporation. All rights reserved. NVIDIA, the NVIDIA logo and RAPIDS are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated. Features, pricing, availability and specifications are subject to change without notice.

Shannon McPhee

+1-310-920-9642

smcphee@nvidia.com