NVIDIA

GPU-ACCELERATED APACHE SPARK 3.0

GPU-accelerate your Apache Spark 3.0 data science pipelines—without code changes—and speed up data processing and model training while substantially lowering infrastructure costs.

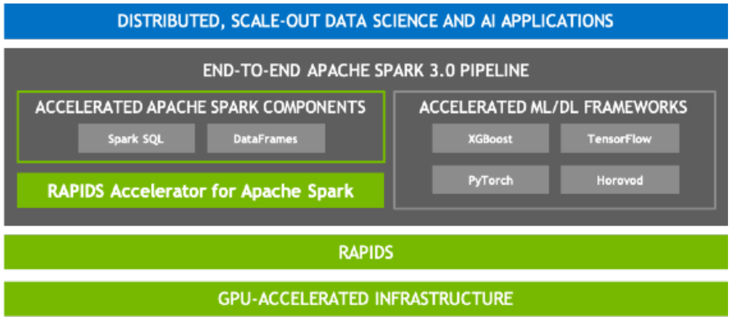
WHAT IS APACHE SPARK?

Apache Spark has become the de facto standard framework for distributed scale-out data processing. With Spark, organizations are able to process large amounts of data, in a short amount of time, using a farm of servers—either to curate and transform data or to analyze data and generate business insights.

- 100s of 1000s of data scientists and over 16,000 enterprises use Spark
- Spark is 100x faster at processing data than Hadoop
- 1000+ contributors across 250+ companies
- Databricks platform alone spins up 1 million virtual machines per day

KEY VALUE PROPS FOR GPU-ACCELERATED SPARK

- **Faster Execution Time** - Accelerate data preparation and quickly move to next stages of the pipeline
- **Streamline Analytics to AI** - Orchestrate end-to-end pipelines, from ETL to model training to visualization; Use same infrastructure for Spark and ML/DL frameworks
- **Reduced Infrastructure Costs** - Complete jobs faster with less hardware to save on-premises and in the cloud; do more with less


DISTRIBUTED, SCALE-OUT DATA SCIENCE AND AI APPLICATIONS

END-TO-END APACHE SPARK 3.0 PIPELINE

ACCELERATED APACHE SPARK COMPONENTS

Spark SQL

DataFrames

RAPIDS Accelerator for Apache Spark

ACCELERATED ML/DL FRAMEWORKS

XGBoost

TensorFlow

PyTorch

Horovod

RAPIDS

GPU-ACCELERATED INFRASTRUCTURE

SPARK 3.0 INNOVATIONS

- **RAPIDS Accelerator for Spark 3.0** - Intercepts and accelerates SQL and DataFrame operations, dramatically improving ETL performance
- **Modifications to Spark Components** - Columnar processing support in the Catalyst query optimizer; Spark shuffle implementation that optimizes the data transfer between Spark processes
- **GPU-Aware Scheduling in Spark** - Spark 3.0 places GPU-accelerated workloads directly onto servers containing the necessary GPU resources; Spark standalone, YARN, and Kubernetes clusters

LEADING SPARK PLATFORMS ARE ACCELERATED

Databricks

Google Cloud Dataproc

On-Premises / Open Source

IDENTIFY OPPORTUNITIES

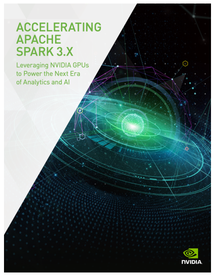
- Are you using Databricks, Google Cloud Dataproc, or Spark on-premises for data analytics?
- If you are using Spark, when do you plan to move to Spark 3?
- Is it taking to long to ingest and prepare data for your ML/DL workloads? It is common this consumes 70% of the time.
- Is data growing faster than your infrastructure can handle?
- Are you using Spark to prepare data for machine learning or deep learning?
- Do you use XGBoost for machine learning?
- Are you currently using GPUs for HPC or AI workloads?
- Is your budget shrinking amidst escalating infrastructure demand?

EDUCATE SPARK USERS

Spark 3.0 is new and it will take time to migrate from Spark 2.x. Also, this is the first version of Spark that is GPU accelerated, making it an ideal educational opportunity.

Share the free Spark eBook by going to:

www.nvidia.com/spark-book


ACCELERATING APACHE SPARK 3.X
Leveraging NVIDIA GPUs to Power the Next Era of Analytics and AI

> *“These contributions lead to faster data pipelines, model training and scoring for more breakthroughs and insights with Apache Spark 3.0 and Databricks.”*
>
> — Matei Zaharia, creator of Apache Spark and chief technologist at Databricks

> *“We’re seeing significantly faster performance with NVIDIA-accelerated Spark 3.0 compared to running Spark on CPUs. With these game-changing GPU performance gains, new possibilities open up for enhancing AI-driven features in our full suite of Adobe Experience Cloud apps.”*
>
> — William Yan, Senior Director of Machine Learning at Adobe

SPARK 3.X TIMELINE

Spark 3.0 Preview Release - November 2019

Spark 3.0 - June 2020

Databricks GPU Support - June 2020

Google Cloud Dataproc - June 2020

Spark 3.1 - Late 2020

Note: It is expected that significant customer traction will occur around the Spark 3.1 release timeframe.

RESOURCES:

www.nvidia.com/spark

www.nvidia.com/spark-book

www.rapids.ai

CHEAT SHEET | GPU-ACCELERATED APACHE SPARK 3.0 | MAY 2020