

Data Engineering is the foundation of data science and lays the groundwork for analysis and modeling. In order for organizations to extract knowledge and insights from structured and unstructured data, fast access to accurate and complete datasets is critical. Working with massive amounts of data from disparate sources requires complex infrastructure and expertise. Minor inefficiencies can result in major costs, both in terms of time and money, when scaled across millions to trillions of data points.

In this workshop, we'll explore how GPUs can improve data pipelines and how using advanced data engineering tools and techniques can result in significant performance acceleration. Faster pipelines produce fresher dashboards and machine learning (ML) models, so users can have the most current information at their fingertips.

## Learning Objectives

In this workshop, you will learn:

- > How data moves within a computer. How to build the right balance between CPU, DRAM, Disk Memory, and GPUs. How different file formats can be read and manipulated by hardware.
- > How to scale an ETL pipeline with multiple GPUs using NVTabular.
- > How to build an interactive Plotly dashboard where users can filter on millions of data points in less than a second.

## Overview

|                                          |                                                                                                                                                                                                                                                |
|------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Duration:</b>                         | 8 hours                                                                                                                                                                                                                                        |
| <b>Price:</b>                            | <a href="#">Contact us for pricing</a>                                                                                                                                                                                                         |
| <b>Prerequisites:</b>                    | <ul style="list-style-type: none"><li>&gt; Intermediate knowledge of <b>Python</b> (List comprehension, objects)</li><li>&gt; Familiarity with <b>pandas</b> a plus</li><li>&gt; <b>Introductory statistics</b> (mean, median, mode)</li></ul> |
| <b>Tools, libraries, and frameworks:</b> | Pandas, cuDF, Dask, NVTabular, Plotly                                                                                                                                                                                                          |
| <b>Assessment type:</b>                  | Skills-based coding assessments evaluate your ability to efficiently filter through millions of data points in the context of an interactive dashboard.                                                                                        |
| <b>Certificate:</b>                      | Upon successful completion of the assessment, you'll receive an NVIDIA DLI certificate to recognize your subject matter competency and support your professional career growth.                                                                |
| <b>Hardware/software requirements:</b>   | You'll need a desktop or laptop computer capable of running the latest version of Chrome or Firefox. You'll be provided with dedicated access to a fully configured, GPU-accelerated workstation in the cloud.                                 |
| <b>Language:</b>                         | English                                                                                                                                                                                                                                        |

## Sample Workshop Outline

|                                                   |                                                                                                                                                                                                                                                                                                                                                                   |
|---------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Introduction</b> (15 mins)                     | <ul style="list-style-type: none"><li>&gt; Meet the instructor.</li><li>&gt; Create an account at <a href="https://courses.nvidia.com/join">courses.nvidia.com/join</a></li></ul>                                                                                                                                                                                 |
| <b>Data on the Hardware Level</b><br>(60 mins)    | <p>Explore the strengths and weaknesses of different hardware approaches to data and the frameworks that support them:</p> <ul style="list-style-type: none"><li>&gt; Pandas</li><li>&gt; CuDF</li><li>&gt; Dask</li></ul>                                                                                                                                        |
| <b>Break</b> (15 mins)                            |                                                                                                                                                                                                                                                                                                                                                                   |
| <b>ETL with NVTabulars</b><br>(120 mins)          | <p>Learn how to scale an ETL pipeline from 1 GPU to many with NVTabular through the perspective of a big data recommender system.</p> <ul style="list-style-type: none"><li>&gt; Transform raw json into analysis-ready parquet files</li><li>&gt; Learn how to quickly add features to a dataset, such as Categorify and Lambda operators</li></ul>              |
| <b>Break</b> (60 mins)                            |                                                                                                                                                                                                                                                                                                                                                                   |
| <b>Data Visualization</b><br>(120 mins)           | <p>Step into the shoes of a meteorologist and learn how to plot precipitation data on a map.</p> <ul style="list-style-type: none"><li>&gt; Learn how to use descriptive statistics and plots like histograms in order to assess data quality</li><li>&gt; Learn effective memory usage, so users can quickly filter data through a graphical interface</li></ul> |
| <b>Final Project: Data Detective</b><br>(60 mins) | <p>Users are complaining that the dashboard is too slow. Apply the techniques learned in class to find and eliminate efficiencies in the backend code.</p>                                                                                                                                                                                                        |
| <b>Final Review</b> (15 mins)                     | <ul style="list-style-type: none"><li>&gt; Review key learnings and answer questions</li><li>&gt; Complete the assessment and earn your certificate</li><li>&gt; Complete the workshop survey</li><li>&gt; Learn how to set up your own AI application development environment</li></ul>                                                                          |
| <b>Next Steps</b>                                 | <p>Continue learning with these DLI trainings:</p> <ul style="list-style-type: none"><li>&gt; <a href="#">Fundamentals of Accelerated Computing with CUDA Python</a></li><li>&gt; <a href="#">Fundamentals of Accelerated Data Science</a></li><li>&gt; <a href="#">High-Performance Computing with Containers</a></li></ul>                                      |

## Why Choose NVIDIA Deep Learning Institute for Hands-On Training?

- > Access workshops from anywhere with just your desktop/laptop and an internet connection. You'll have access to a fully configured, GPU-accelerated server in the cloud.
- > Obtain hands-on experience with the most widely used, industry-standard software, tools, and frameworks.
- > Learn to build deep learning and accelerated computing applications for industries, such as healthcare, robotics, manufacturing, accelerated computing, and more.
- > Gain real-world experience through content designed in collaboration with industry leaders, such as the Children's Hospital of Los Angeles, Mayo Clinic, and PwC.
- > Earn an NVIDIA DLI certificate to demonstrate your subject matter competency and support your career growth. 

For the latest DLI workshops and trainings, visit [www.nvidia.com/dli](https://www.nvidia.com/dli)

For questions, contact us at [nvdli@nvidia.com](mailto:nvdli@nvidia.com)

© 2021 NVIDIA Corporation. All rights reserved. NVIDIA, the NVIDIA logo, DGX, Nsight, and Triton are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. All other trademarks and copyrights are the property of their respective owners. APR21

