NVIDIA AI Enterprise Test Drive

Application Note
# Document History

SWE-NVAIE-001-APNT

<table>
<thead>
<tr>
<th>Version</th>
<th>Date</th>
<th>Authors</th>
<th>Description of Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>30 July 2021</td>
<td>VNK, JC, JL, CW</td>
<td>Initial release</td>
</tr>
</tbody>
</table>
Table of Contents

Chapter 1. Introduction.............................................................................................................. 5
Chapter 2. Getting Started with NVIDIA AI Enterprise Test Drive .............................. 8
Chapter 3. Testing the NVIDIA and VMware Test Drive ................................................. 9
  3.1 Data Prep [RAPIDS]........................................................................................................ 9
  3.2 Training [TensorFlow] ................................................................................................... 9
  3.2.1 Launching NVIDIA AI Enterprise Test Drive ......................................................... 9
Chapter 4. Running the NVIDIA AI Enterprise Test Drive Demos................................. 11
  4.1 Training [TensorFlow] ................................................................................................ 11
  4.2 Data Prep [RAPIDS]..................................................................................................... 13
List of Figures

Figure 1-1. NVIDIA AI Enterprise Software Suite .......................................................... 6
Figure 1-2. Example of Data Science Workflow .............................................................. 6
Figure 3-1. NVIDIA AI Enterprise Test Drive – Ubuntu Desktop ................................ 10
Figure 4-1. BERT Question/Answer ............................................................................. 11
Figure 4-2. BERT Model Example Paragraph .............................................................. 12
Figure 4-3. BERT Demo Provided Questions ................................................................. 12
Figure 4-4. BERT Demo Custom Inputs .................................................................... 13
Figure 4-5. Taxi Fare Demo Overview .......................................................................... 13
Figure 4-6. Example of Taxi Fare Data ....................................................................... 14
Figure 4-7. Training and Wall Time ........................................................................... 14
Figure 4-8. Comparison of Taxi Fare Prediction and Actual Data ................................. 15
Chapter 1. Introduction

Artificial intelligence (AI) is transforming every industry, whether it is by improving customer relationships in financial services, streamlining manufacturer supply chains, or helping doctors deliver better outcomes for patients. While most organizations know they need to invest in AI to secure their future, they struggle with finding the strategy and platform that can enable success.

Unlike traditional enterprise applications, AI applications are a relatively recent development for many IT departments. They are anchored in rapidly evolving, open-source, bleeding-edge code and lack proven approaches that meet the rigors of scaled production settings in enterprises.

VMware and NVIDIA have partnered to unlock the power of AI for every business by delivering an end-to-end enterprise platform optimized for AI workloads. This integrated platform delivers best-in-class AI software, the NVIDIA AI Enterprise Suite, optimized and exclusively certified for the industry’s leading virtualization platform, VMware vSphere®. Running on NVIDIA-Certified Systems™, industry-leading accelerated servers, this platform accelerates the speed at which developers can build AI and high-performance data analytics, enables organizations to scale modern workloads on the same VMware vSphere infrastructure they have already invested in, and delivers enterprise-class manageability, security, and availability.
NVIDIA AI Enterprise is an end-to-end AI software suite that includes AI frameworks and tools that provide performance-optimized deep learning, machine learning, and data science tools that simplify building, sharing, and deploying AI software, so enterprises can gather insights faster and deliver business value sooner.

**Figure 1-1. NVIDIA AI Enterprise Software Suite**

**Figure 1-2. Example of Data Science Workflow**

**Data Prep** – Data science software such as the RAPIDS suite of software libraries, built on CUDA-X AI, enable data scientists to execute end-to-end data science and analytics pipelines entirely on GPUs.

**AI Training** – Deep learning frameworks for training and machine learning, such as PyTorch and TensorFlow, are optimized for GPU acceleration. With the NVIDIA Transfer Learning Toolkit, enterprises can take these pre-trained models and refine them for their own usage.
domain, greatly reducing development time. Leveraging these tools and pre-trained models, can reduce development time by up to 10x, from 80 weeks to 8 weeks, for example.

**AI Inference** – The newly trained models can be optimized using deep learning inference SDKs and libraries such as NVIDIA TensorRT. This is done by fusing layers and eliminating unneeded steps.

**Scale** – Triton Inference Server simplifies and optimizes the deployment of AI models at scale in production. It integrates with Kubernetes for orchestration and auto-scaling and allows front-end client applications to submit inference requests from an AI inference cluster and can service models from an AI model repository.

The NVIDIA AI Enterprise Test Drive is not a commercially available product. It is a technology demonstration developed by NVIDIA and VMware to showcase the benefits of adding NVIDIA AI Enterprise software suite to boost AI performance, deploy with confidence, and scale without compromise.

Take 48 hours to test drive a high-performance enterprise-ready AI experience.
Chapter 2. Getting Started with NVIDIA AI Enterprise Test Drive

Your personal virtual machine (VM) has been preconfigured with demos that showcase the benefits of the NVIDIA AI Enterprise Software Suite. These applications are categorized by type and found on the Home Tabs of the Google Chrome Browser.

To run any of these demos, simply click on any of the Tabs after launching Google Chrome.

The demos are divided into the following categories:

- Inference
- Training

The NVIDIA AI Enterprise Software Suite includes a curated set of AI and data science frameworks and tools, NVIDIA operators for cloud native deployment, and infrastructure optimization software.
Chapter 3. Testing the NVIDIA and VMware Test Drive

3.1 Data Prep (RAPIDS)

RAPIDS is a suite of GPU-accelerated data science libraries with APIs that should be familiar to users of Pandas, Dask, and Scikitlearn. This demo focuses on showing how to use cuDF with Dask and XGBoost to scale GPU DataFrame ETL-style operations. Anaconda has graciously made some of the NYC Taxi dataset available in a public Google Cloud Storage bucket. We use our vGPU-enabled VM to process it and train a model that predicts the fare amount.

3.2 Training (TensorFlow)

Bidirectional Embedding Representations from Transformers (BERT) is a method of pre-training language representations that obtains state-of-the-art results on a wide array of natural language processing (NLP) tasks. The original paper can be found here: https://arxiv.org/abs/1810.04805.

NVIDIA’s BERT is an optimized version of Google’s official implementation, leveraging mixed precision arithmetic target accuracy. This demo will demonstrate inference on question/answering tasks and the use of mixed precision models.

3.2.1 Launching NVIDIA AI Enterprise Test Drive

After login, click NVIDIA AI Enterprise Test Drive and launch the Ubuntu Desktop.
Login credentials are as follows:

**Username**: nvidia

**Password**: nvidiaAI!

Once the Ubuntu desktop is launched, open **Google Chrome** from the favorites **left side bar** for access to the demos.
Chapter 4. Running the NVIDIA AI Enterprise Test Drive Demos

This section describes how to execute the pre-installed demos in the NVIDIA AI Enterprise Test Drive.

- Training (TensorFlow)
- Bert Q&A System Data Prep [RAPIDS] – The section goes through the listed rich web experience demos.
- Data Prep [RAPIDS]
  - NYC Taxi Fair

4.1 Training (TensorFlow)

Figure 4-1. BERT Question/Answer

To interact with this demo, click on each step then press Shift + Enter to execute each Jupyter notebook text or code block.

During this demo, we demonstrate inference on the paragraph below:
Figure 4-2. BERT Model Example Paragraph

The demo answers the questions provided.

Figure 4-3. BERT Demo Provided Questions

After running inference, you are able to display and compare the Question/Answer results. You can also customize the number of questions, content, and questions. At the end, the answers are displayed based on the content and questions provided.
4.2 Data Prep (RAPIDS)

To interact with this demo, click on each step then press Shift + Enter to execute each Jupyter notebook text or code block.

First, you install tools needed for the demo. This should take about 1-2 minutes. This demo will go through the steps to inspect, clean up, and analyze the taxi fare data. You can also display the data based on distance, fare, and passenger count.
You can perform training based on the specified training set. Notice that the wall time is under 12 seconds, significantly faster than a CPU-only VM.

You can then compare the prediction to the actual fare.
Figure 4-8. Comparison of Taxi Fare Prediction and Actual Data

In [30]: prediction.head()
Out[30]:
   0   13.348164
   1   7.968383
   2   8.166205
   3   8.239672
   4  14.626629
Name: prediction, dtype: float32

In [31]: actual.head()
Out[31]:
   0   13.0
   1    7.5
   2    8.0
   3    8.0
   4   14.5
Name: fare_amount, dtype: float32