AI Made Easy: Unleash the Potential of Dell Enterprise Hub on Hugging Face

Seamlessly streamlining model training, fine-tuning, and deployment

May 2024

H20033

Technical White Paper

Abstract

This document describes the advantages and features of using the Dell Enterprise Portal on Hugging Face to select and deploy optimized self-contained containers on validated Dell infrastructure.

Dell Technologies AI Solutions

Dell
Reference Design

DELL Technologies

Copyright

The information in this publication is provided as is. Dell Inc. makes no representations or warranties of any kind with respect to the information in this publication, and specifically disclaims implied warranties of merchantability or fitness for a particular purpose.

Dell Inc. believes the information in this document is accurate as of its publication date. The information is subject to change without notice.

Contents

Executive summary

Overview

In our fast-paced digital landscape, harnessing the power of artificial intelligence (AI) is more critical than ever. However, the complexity of AI technologies can be a significant barrier, especially as numerous solutions claim to streamline AI deployment. Recognizing this challenge, Dell Technologies and Hugging Face have partnered to introduce a groundbreaking portal to bridge this gap. Introducing the Dell Enterprise Hub, a platform focused on simplifying and demystifying AI to make AI more accessible and manageable for all.

This partnership combines Dell Technologies' robust hardware and user-friendly platform with Hugging Face's cutting-edge AI models, libraries, and tools to provide a streamlined, easy-to-navigate solution that simplifies the AI deployment processes. This white paper covers Dell Enterprise Hub's innovative features and provides step-by-step instructions on leveraging Hugging Face's sophisticated models to optimize AI workflow and enhance business operations, leading users through the nuances of model training, fine-tuning, and deployment.

This guide empowers IT professionals, data scientists, and business decision-makers to implement and scale AI solutions effectively by reducing the complexity traditionally associated with AI projects. Unlike many existing platforms that offer piecemeal or overly complex solutions, our integrated approach ensures a smoother transition, better scalability, and enhanced operational efficiency, thereby driving innovation and building competitive advantage in a variety of fields.

This document provides comprehensive guidance on using Dell Enterprise Hub to access and deploy AI models optimized for enhanced performance, scalability, and security. Developed with Hugging Face, these models are specially tailored with containerization for robust on-premises execution.

This guide offers detailed, step-by-step instructions for installing, configuring, and operating the system, alongside practical examples of use cases and best practices. By detailing both the technical specifications and the operational guidance, this document ensures that users are equipped with the knowledge to maximize the capabilities of the Dell Enterprise Hub, optimize their AI workflows, and achieve substantial improvements in processing speed, data handling, and operational efficiency.

Audience

This document is intended for a technical audience, including IT professionals, data scientists, AI researchers, and technical teams involved in integrating and deploying AI solutions within their organizations. Whether you want to integrate these advanced models into your existing infrastructure, optimize AI performance at scale, or streamline your AI deployment processes, this white paper will provide the essential information needed to achieve these goals efficiently and effectively.

Revisions

Date	Part number/ revision	Description
May 2024	H20033	Initial release

We value your feedback

Dell Technologies and the authors of this document welcome your feedback on this document. Contact the Dell Technologies team by email.

Author: Tiffany Fahmy, Gary Pannell

Note: For links to other documentation for this topic, see the Artificial Intelligence Info Hub.

Solution overview

While it is feasible for companies to download AI models directly from Hugging Face, the solution covered here streamlines the entire lifecycle of AI deployment from selection and integration to scaling and management. Dell Enterprise Hub is a centralized platform where users can access pre-optimized AI models from Hugging Face. This integration reduces the technical overhead typically associated with configuring and maintaining AI models, offering a user-friendly interface that allows users to manage their AI resources easily. This simplification is crucial for businesses looking to adapt AI solutions quickly without extensive in-house expertise.

Dell Technologies' hardware is specifically designed to maximize the performance of AI models, providing the computational power needed to handle complex algorithms and large datasets efficiently. The solution ensures low-latency processing by running Hugging Face's AI models on Dell Technologies' optimized hardware, an essential for applications requiring real-time decision-making. The hardware-software integration provided by this partnership allows businesses to leverage AI capabilities to their full potential, ensuring that performance bottlenecks do not hinder operational efficiency.

The solution's architecture is designed for current needs, scalability, and future expansion. As businesses grow and their data processing requirements evolve, the solution can scale accordingly without requiring a complete overhaul of the existing system. This future proofing is vital for enterprises investing heavily in AI, ensuring that their initial investments yield returns even as their operational complexities increase.

For industries where data privacy is paramount, such as healthcare and finance, deploying AI solutions on-premises is often necessary to comply with regulatory standards. The combined solution from Dell Technologies and Hugging Face addresses this by ensuring that all data processing occurs within the controlled environment of the company's infrastructure, significantly reducing the risks associated with data breaches and unauthorized access. Furthermore, the solution incorporates advanced security protocols and compliance checks, updated continuously to keep pace with evolving regulations.

The section provides an overview of the design and solution implemented by Dell Technologies and Hugging Face, setting the context by highlighting the critical business challenges and technical environments these solutions are tailored to address. By integrating Hugging Face's advanced AI models with Dell Technologies' powerful hardware, this solution caters to the growing need for on-premises AI capabilities that comply with stringent data privacy regulations and offer low-latency processing.

Business challenges

As an emerging technology, it is vital to partner with trusted leaders with the expertise and dedication to further the success of your organization. Dell Technologies focuses on the end-to-end AI strategy for enterprises to ensure your AI approach can succeed in the modern data-centric digital era. In the first quarter of 2024, the Forrester Wave™: AI Infrastructure Solutions ranked Dell Technologies as a leader in the space.[1] The report also highlighted Dell Technologies' market presence, current offering, and strategy, demonstrating that Dell Technologies is a coveted partner for enterprises looking to evolve their AI infrastructure. Dell Technologies is also fiercely competitive and innovative in ensuring the best performance for our customers. Dell Technologies continues to expand its AI catalog by developing solutions, adding to its portfolio of services, and partnering with other proven leaders in AI.

Hugging Face is the apex community for AI models, built on a culture of collaboration and sharing, supporting multiple frameworks, and promoting an open-sourced approach to AI and data. The Hugging Face library is immense, with over 600,000 models, over 100,000 data sets, and hundreds of thousands of user-submitted applications (known as 'Spaces' on the Hugging Face). The Transformers library supplies flexible framework interoperability through APIs and tools combined with the provided pre-trained models, allowing developers to save time and resources. Dell Enterprise Hub subscriptions add many features, including single sign-on (SSO), granular access controls, comprehensive logs, and priority support from the Hugging Face Team.

Today's competitive landscape demands scalable AI solutions to innovate and lead the market. This is not without its challenges, however, leaving businesses to contend with:

1. **Supply Chain Security:** With AI models being integrated into various aspects of supply chain management, ensuring the security of these systems has become paramount. Breaches can disrupt operations, resulting in significant financial and reputational damage.
2. **Traceability:** Companies are increasingly required to trace data and decision flows, especially when using AI models. This need for traceability is crucial for accountability, compliance, and improving the transparency of AI-driven decisions.
3. **Indemnification:** The legal landscape surrounding AI models is still evolving. Therefore, organizations using third-party AI models must protect themselves from potential liabilities arising from misuse of models and/or unintended consequences.
4. **Scalability and Performance:** As AI adoption grows, businesses need help to scale their AI models while maintaining performance. The increased volume of data and increasing complexity of AI tasks require robust infrastructure and optimization.
5. **Model Validation and Optimization:** Organizations often waste valuable resources testing and tuning generic models. Finding a suitable model optimized for specific hardware and business needs can take time and effort, leading to inefficiencies in the workflow.

[1] https://www.dell.com/en-us/blog/dells-ai-infrastructure-makes-waves-in-forrester-report/

Use cases

Eliminate wasted time and resources when testing and tuning generic models

Data researchers and scientists can spend anywhere from days to weeks implementing and optimizing a suitable AI model. This can include time sunk into researching whether a model will perform at the expected level given the hardware environment and software dependencies. Even if the model is able to run, there is no guarantee that the parameters and variables are set correctly to maximize the model, leaving potential cycles lost in forums and support cases along the way.

Dell Technologies and Hugging Face engineers have tested, validated, and optimized select popular models to remove the burden of uncertainty from your workflow and increase your speed to solution delivery. A curated set of models, available in easy-to-download and launch containers, are listed on the Dell Enterprise Hub on Hugging Face to ensure you have *the* optimized model for the given server and GPU combination. There is peace of mind in knowing that a solution has been validated and tested by authorities and proven to work optimally so that the developers can focus on the AI workload outcomes.

Working together to keep up with the pace of change

"We are facing a step change in what's possible for individual people to do, and at a previously unthinkable pace."[2]

Advances in AI and AI models are ushering in an unprecedented rate of technology change and the way people interact with it. New models and improvements are being added to the landscape rapidly, rivaling or surpassing human capabilities. Organizations are challenged to stay abreast of the latest advancements in AI models and the hardware that drives them.

Partnering with Dell Technologies and Hugging Face helps enterprises stay current and efficient in AI. As new GPUs from various vendors and innovative improvements in Dell PowerEdge servers emerge into the marketplace, Dell Technologies and Hugging Face will research, test, and optimize the highest-demand models. New models will be regularly validated and added to the inventory for selection in the Dell Enterprise Hub on Hugging Face. Subscribers to this service can take advantage of Dell Technologies and Hugging Face's work to stay current on models and capabilities.

Solution approach

Dell Technologies announced the partnership with Hugging Face as organizations seek an easier way to bring optimized AI models to their on-premises environment. These organizations benefit from the time-to-value and user experience this partnership brings. As a result of this collaboration, Dell Technologies is the first infrastructure provider to partner with Hugging Face to offer optimized on-premises deployment of generative AI models.

The portal features a tiered architecture where Hugging Face's models are pre-packaged into containers optimized for Dell Technologies' hardware performance. These models can be easily accessed and managed through a user-friendly portal developed by Dell Technologies, allowing for effortless downloading, deployment, and scaling. Dell Technologies and Hugging Face will continue to collaborate to validate and optimize

[2] https://time.com/6310115/ai-revolution-reshape-the-world/

popular, in-demand models as they enter the market for existing Dell PowerEdge server platforms, GPU combinations, and new AI-focused hardware.

As subscribers to the Dell Enterprise Hub, organizations benefit from the enterprise-grade features of Hugging Face. Organizations can deploy single sign-on and receive increased security and audit capabilities, access to advanced computing options, a more comprehensive data set, and priority support from the Hugging Face team.

Solution design

Initial hardware platforms

The Dell Enterprise Hub on Hugging Face is an evolving utility that speeds up the deployment of optimized containers to customer environments. As new Dell hardware, GPUs, and AI models enter the market, Dell Technologies and Hugging Face will collaborate to curate self-contained containers for customers to rapidly deploy with confidence. The initial selection of Dell Server Platforms and GPU models reflects the most in-demand and versatile platforms currently on the market. New entrants into AI-focused Dell PowerEdge servers will be evaluated and selected for the Dell Enterprise Hub to expand the ease of deploying containers from Hugging Face.

Servers

Dell servers are engineered with high-performance computing in mind and designed to support intense AI and machine learning workloads. These servers have advanced CPUs and GPUs that offer the necessary power to handle multiple AI processes simultaneously, ensuring efficient data processing and model training.

The Dell PowerEdge XE9680 rack server is a purpose-built platform to maximize high-performance machine learning by leveraging up to 8 fully interconnected GPUs. Customers experience notable performance gains when leveraging this platform to train and fine-tune LLMs.

The Dell PowerEdge R760xa rack server is engineered for scalability and high performance and can support up to 4 enterprise-class GPUs per server. This server presents flexible options for AI acceleration in a compact 2U chassis while delivering the performance required to scale up large language models.

GPUs

Recent advances in AI technology have shown that high-performing GPUs can significantly improve task times for crucial types of workloads and are a required component of many Generative AI models. Their scalability, parallel processing, and extensive software tools make GPUs essential for maximizing AI workloads.

The NVIDIA H100 Tensor Core GPU is specifically designed for the most demanding Enterprise AI experience and is the current market leader in LLM training.

The NVIDIA L40S GPU is a popular option due to its performance and efficiency, paired with the scalable Dell PowerEdge R760xa platform for growth and flexibility.

Initial AI models

Dell Technologies and Hugging Face have worked together to establish the initial set of popular models for the Dell Enterprise Hub and will continue to add, test, and optimize models as the platform evolves. This joint venture is designed to be iterative and flexible, continually expanding organically to meet the demands of customers running AI workloads on Dell servers.

The Hugging Face community is a collaborative environment for data researchers and data scientists to work together and contribute to open-source projects for AI. The community rigorously tests and evaluates models, resulting in high engagement and feedback. Popular models boast their benchmark scores on various tests and competitive claims against their peers to validate their popularity and download counts.

Hugging Face and Dell Technologies have collaborated to optimize the selected models for inclusion in the initial group, aiming to represent the industry's leading LLMs to run on Dell PowerEdge Server and GPU combinations.

The models included can be used for both training and inferencing. The training process is more resource-intensive given that the model learns from the data set to capture patterns and generalization. Once the model has been trained, the inference operation works toward fast, efficient, and accurate predictions of new data while constituting less of a burden on the available resources.

The software suite from Hugging Face includes advanced AI solutions and workloads, as well as specialized libraries and tools that enhance the performance of AI models, streamlining their deployment and operation.

Models

- The Llama 3 70B model and Llama 3 8B model are two sizes of the latest pre-trained and instruction-tuned generative text models developed by Meta. These models outperform other open-source LLMs and are optimized for dialogue use cases.[3] The Llama 3 8B model offers strong performance on a smaller, less resource-intensive model, while the Llama 3 70B model requires substantially more resources to achieve its performance heights.
- The Zephyr 7B Beta model is a language model developed by Hugging Face based on a fine-tuned version of the Mistral 7B v.01 model. This model was trained on publicly available datasets and adjusted to be less filtered when producing responses.
- The Gemma 7B model is a lightweight, state-of-the-art open language model developed by Google that is designed for text generation tasks. One of the features of the Gemma 7B model is its small size, which allows it to run efficiently in containerized environments with less demand for GPU resources. This model also benefits from Google's focus on quality, safety, and ethical data preprocessing.
- The Mistral 8x22B model and the Mistral 7B model focus on being the most efficient and performance-centric models when compared to their contemporaries and are easy to fine-tune for specific tasks. While the Mistral 7B model uses grouped-query attention (GQA) to increase inference speed, the Mistral 8x22B model utilizes sparse Mixture of Experts (MoE) layers to allow the model to significantly scale up or improve the data size without greatly impacting resources.

[3] https://techcrunch.com/2024/04/18/meta-releases-llama-3-claims-its-among-the-best-open-models-available/

- The Dell Enterprise Hub also supports the Bring Your Own model (BYOM), which enables customers to deploy their own trained and fine-tuned model from the platform.

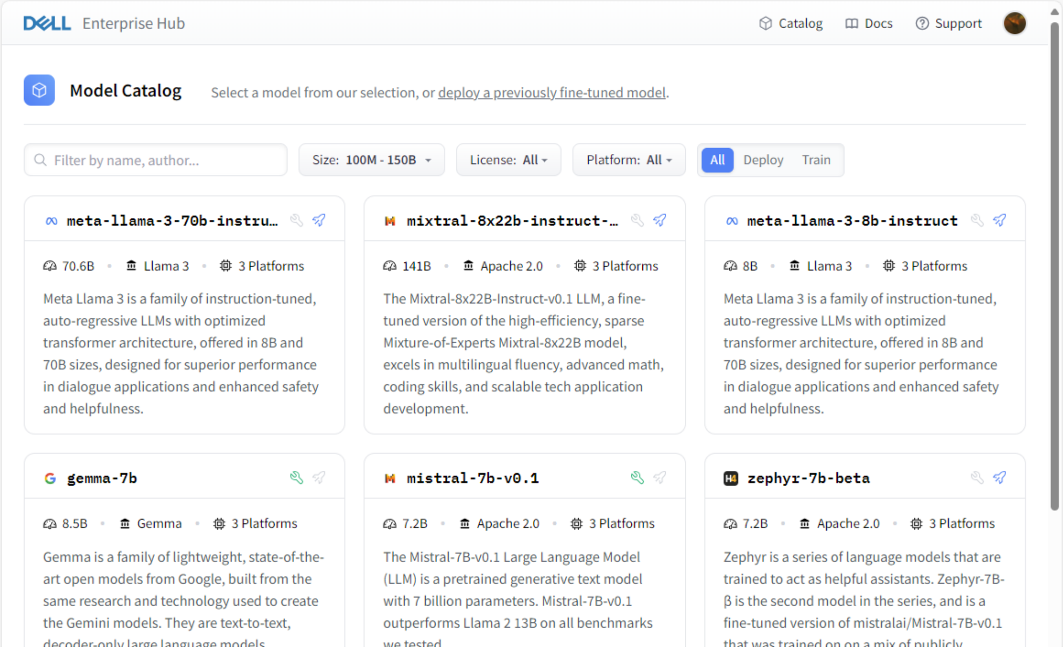
Implementation guidance from model selection to deployment

The Dell Enterprise Hub on Hugging Face is designed to make selecting, configuring, and deploying AI models simpler and more reliable by filtering the optimized model choices and delivering the models in self-contained containers to deploy on Dell PowerEdge Servers.

> *"…[That's the] magic of the experience, with minimal config from the enterprise hub through a simple copy and paste command, you are able to accomplish in minutes something that takes companies weeks through trial and error…"*
>
> *– Jeff Boudier, Head of Product at Hugging Face*

Dell Enterprise Hub on Hugging Face is available to all users and organizations. Subscribers to the Dell Enterprise Hub will use their Hugging Face credentials to log into https://dell.huggingface.co securely. The initial login process will redirect the user to authorize Dell Enterprise Hub to access the content that exists in your organization's Enterprise Hub subscription on Hugging Face. This is an OAuth connection to grant read access to the user content and models. Terms of service on models the user has previously accepted will carry through to models curated in the Dell Enterprise Hub.

Model Catalog



Model Catalog

Select a model from our selection, or deploy a previously fine-tuned model.

meta-llama-3-70b-instru…

70.6B • Llama 3 • 3 Platforms

Meta Llama 3 is a family of instruction-tuned, auto-regressive LLMs with optimized transformer architecture, offered in 8B and 70B sizes, designed for superior performance in dialogue applications and enhanced safety and helpfulness.

mixtral-8x22b-instruct-…

141B • Apache 2.0 • 3 Platforms

The Mixtral-8x22B-Instruct-v0.1 LLM, a fine-tuned version of the high-efficiency, sparse Mixture-of-Experts Mixtral-8x22B model, excels in multilingual fluency, advanced math, coding skills, and scalable tech application development.

meta-llama-3-8b-instruct

8B • Llama 3 • 3 Platforms

Meta Llama 3 is a family of instruction-tuned, auto-regressive LLMs with optimized transformer architecture, offered in 8B and 70B sizes, designed for superior performance in dialogue applications and enhanced safety and helpfulness.

gemma-7b

8.5B • Gemma • 3 Platforms

Gemma is a family of lightweight, state-of-the-art open models from Google, built from the same research and technology used to create the Gemini models. They are text-to-text, decoder-only large language models

mistral-7b-v0.1

7.2B • Apache 2.0 • 3 Platforms

The Mistral-7B-v0.1 Large Language Model (LLM) is a pretrained generative text model with 7 billion parameters. Mistral-7B-v0.1 outperforms Llama 2 13B on all benchmarks we tested

zephyr-7b-beta

7.2B • Apache 2.0 • 3 Platforms

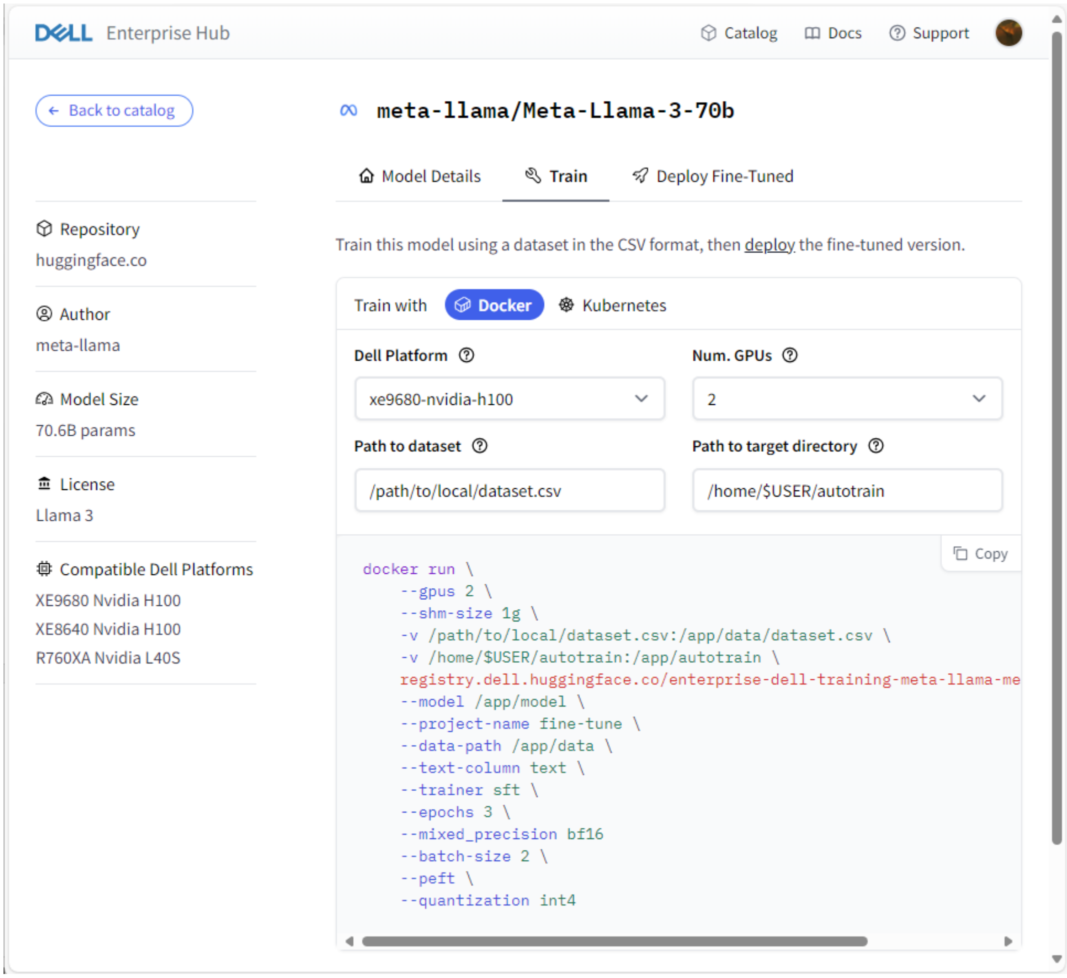
Zephyr is a series of language models that are trained to act as helpful assistants. Zephyr-7B-β is the second model in the series, and is a fine-tuned version of mistralai/Mistral-7B-v0.1 that was trained on on a mix of publicly

Figure 1. Sample Model Catalog, Dell Enterprise Hub on Hugging Face

While the filter functions include valid selectors such as model name, size of the model, and license type, the main filter is the Platform dropdown. This allows the Dell Enterprise Hub subscriber to select specific Dell PowerEdge server models with the available GPU options. This retrieves the exact model containers tuned specifically to Dell Server and GPU combinations to create optimized containers.

There are two types of Model Cards in the Model Catalog: **Deploy** and **Train**. The **Deploy** option, indicated by a small blue rocket icon next to the model title, enables developers to deploy the pretrained models to on-premises Dell platforms. The **Train** option, indicated by the small green wrench icon next to the model title, empowers users to leverage the model to train custom datasets using the on-premises Dell platforms and then deploy the fine-tuned model leveraging the trained data. The user clicks the model card to configure a small number of defined parameters and initiate the deployment process.

Model Card Train tab



meta-llama/Meta-Llama-3-70b

Repository
huggingface.co

Author
meta-llama

Model Size
70.6B params

License
Llama 3

Compatible Dell Platforms
XE9680 Nvidia H100
XE8640 Nvidia H100
R760XA Nvidia L40S

Train this model using a dataset in the CSV format, then deploy the fine-tuned version.

Train with Docker Kubernetes

Dell Platform: xe9680-nvidia-h100

Num. GPUs: 2

Path to dataset: /path/to/local/dataset.csv

Path to target directory: /home/$USER/autotrain

```
docker run \
    --gpus 2 \
    --shm-size 1g \
    -v /path/to/local/dataset.csv:/app/data/dataset.csv \
    -v /home/$USER/autotrain:/app/autotrain \
    registry.dell.huggingface.co/enterprise-dell-training-meta-llama-me
    --model /app/model \
    --project-name fine-tune \
    --data-path /app/data \
    --text-column text \
    --trainer sft \
    --epochs 3 \
    --mixed_precision bf16
    --batch-size 2 \
    --peft \
    --quantization int4
```

Figure 2. Sample Model Card Train tab, Dell Enterprise Hub on Hugging Face

Notice the details on the left, including Author, Model Size, License, and—specific to the Dell Enterprise Hub—Compatible Dell Platforms. The model card for the Train model allows the user to train then deploy the fine-tuned model with Docker or Kubernetes, with additional deployment options planned for future releases. To utilize the containerized feature of the Dell Enterprise Hub, the system must have one of these platforms fully deployed and configured in their on-premises Dell platform.

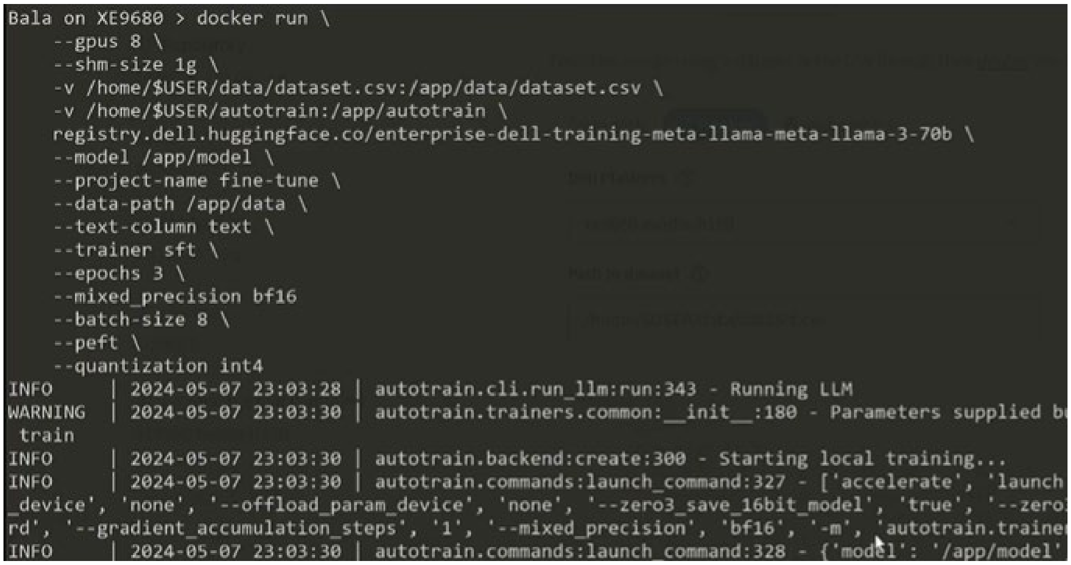
The **Dell Platform** drop-down allows users to select the appropriate hardware and GPU combination to which they wish to deploy the model. This ensures that the generated code and model information is the optimized model container for that exact hardware and GPU combination. Many Dell PowerEdge server and GPU combinations have more than

1 GPU, however the user may only want to allocate a fraction of the total to a particular model or project. The **Num.GPUs** field allows the user to adjust this quantity for the model.

The **Path to dataset** field identifies the user dataset to be trained. This can be either a mounted location on the host server or an accessible unc path from the host server. The **Path to target directory** defaults to the current user's home directory and will host artifacts such as weights from the model.

The default settings in the containers are set for PEFT (Parameter-Efficient Fine-Tuning) with QLoRA (Quantization and Low-Rank Adapters). They are already included in the container automation, however users can modify this setting if another method is preferred. The user can make additional edits to the generated code snippet section, such as epochs or batch size. The containers are self-contained and prepackaged with model weights.

The user will then copy the code snippet from the Model Card and paste this code into the command line on the host server (single node) or a control node for the cluster (multi-node). Running the code will pull the Dell-optimized training container from the Hugging Face container registry. If the model has previously been pulled in this environment, the code snippet will attempt to use the local cached container registry.

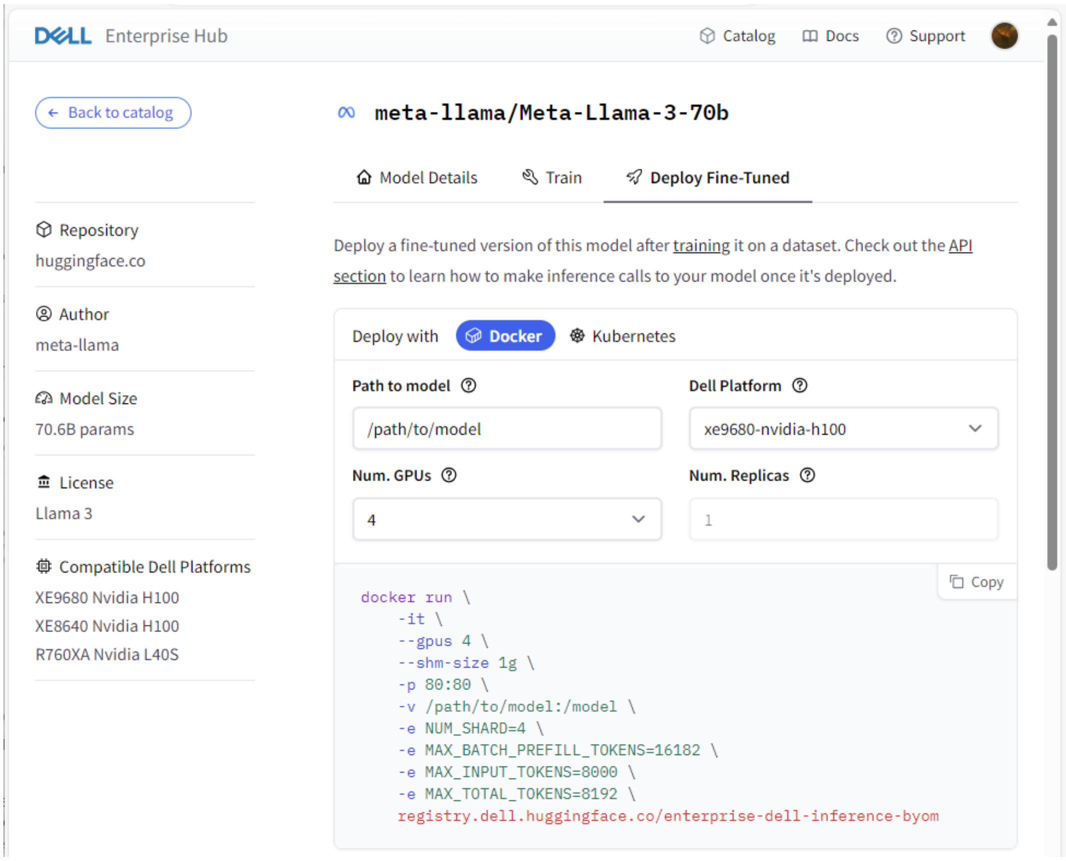


```
Bala on XE9680 > docker run \
    --gpus 8 \
    --shm-size 1g \
    -v /home/$USER/data/dataset.csv:/app/data/dataset.csv \
    -v /home/$USER/autotrain:/app/autotrain \
    registry.dell.huggingface.co/enterprise-dell-training-meta-llama-meta-llama-3-70b \
    --model /app/model \
    --project-name fine-tune \
    --data-path /app/data \
    --text-column text \
    --trainer sft \
    --epochs 3 \
    --mixed_precision bf16
    --batch-size 8 \
    --peft \
    --quantization int4
INFO      | 2024-05-07 23:03:28 | autotrain.cli.run_llm:run:343 - Running LLM
WARNING   | 2024-05-07 23:03:30 | autotrain.trainers.common:__init__:180 - Parameters supplied b
 train
INFO      | 2024-05-07 23:03:30 | autotrain.backend:create:300 - Starting local training...
INFO      | 2024-05-07 23:03:30 | autotrain.commands:launch_command:327 - ['accelerate', 'launch
_device', 'none', '--offload_param_device', 'none', '--zero3_save_16bit_model', 'true', '--zero
rd', '--gradient_accumulation_steps', '1', '--mixed_precision', 'bf16', '-m', 'autotrain.traine
INFO      | 2024-05-07 23:03:30 | autotrain.commands:launch_command:328 - {'model': '/app/model'
```

Figure 3. Sample code snippet command line – model training using docker

Hugging Face's autotrain, a powerful tool that simplifies the model training process, is an integrated component in the container. The model will immediately spin up the required resources and begin training the model using the dataset indicated in the model card parameter. As the model training progresses, informational messages in the command line relay data about the deployment status. Once the training is completed, return to the Dell Enterprise Hub Model Card to update the **Path to model** field and copy the code snippet.

Model Card Deploy Fine-Tuned tab



meta-llama/Meta-Llama-3-70b

Repository
huggingface.co

Author
meta-llama

Model Size
70.6B params

License
Llama 3

Compatible Dell Platforms
XE9680 Nvidia H100
XE8640 Nvidia H100
R760XA Nvidia L40S

Deploy a fine-tuned version of this model after training it on a dataset. Check out the API section to learn how to make inference calls to your model once it's deployed.

Path to model: /path/to/model

Dell Platform: xe9680-nvidia-h100

Num. GPUs: 4

Num. Replicas: 1

```
docker run \
    -it \
    --gpus 4 \
    --shm-size 1g \
    -p 80:80 \
    -v /path/to/model:/model \
    -e NUM_SHARD=4 \
    -e MAX_BATCH_PREFILL_TOKENS=16182 \
    -e MAX_INPUT_TOKENS=8000 \
    -e MAX_TOTAL_TOKENS=8192 \
    registry.dell.huggingface.co/enterprise-dell-inference-byom
```

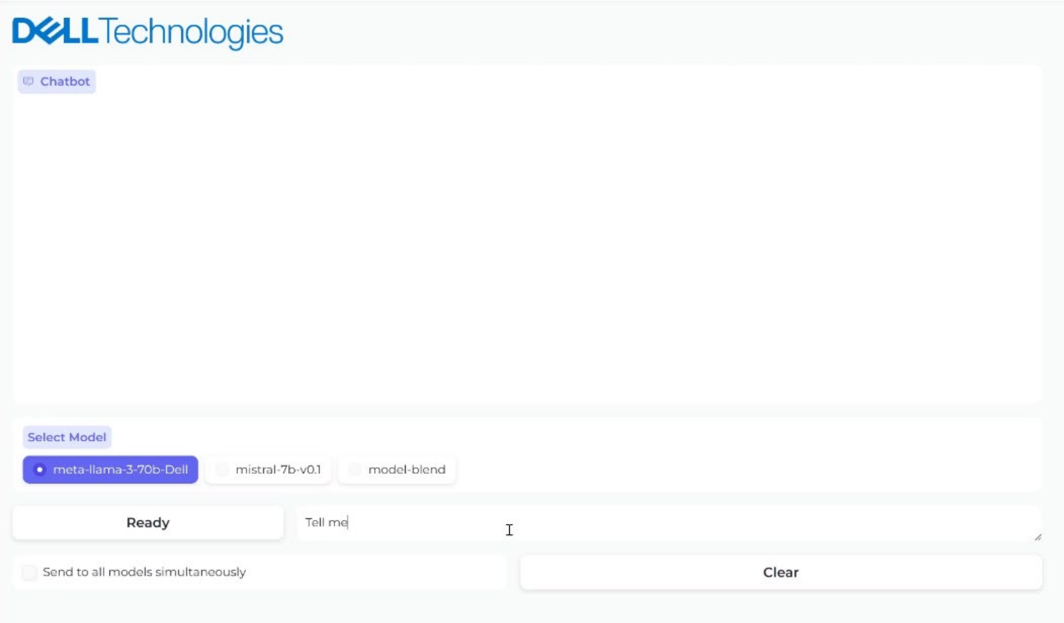
Figure 4. Sample Model Card Deploy Fine-Tuned tab, Dell Enterprise Hub on Hugging Face

During the previous training process, the model path defaults to the parent directory of the **Path to target** directory setting. If the default setting was used during the last process, enter the value `/home/$USER/model` in the Path to model field. Like the Train process, the number of GPUs can be designated. The listening port of the web service can also be set in the code snippet. Once the options are set, copy the code snipped and run it from the on-premises Dell platform to deploy the fine-tuned model.

The fine-tuned model deployment leverages the trained dataset. Once the deployment completes, the model is ready to be in the included web UI provided by the web service. A simple test with a curl command can verify the service is running and will respond with generated AI content:

```
curl 127.0.0.1:80/generate    \
  -X POST    \
  -d '{"inputs":"What is Deep
Learning?","parameters":{"max_new_tokens":50}}'    \
  -H 'Content-Type: application/json'
```

Figure 5 represents the web interface for the deployed model. Users can reach the UI using any network interface on the host using the port specified in the deployment code snippet or default port 80.



Select Model

● meta-llama-3-70b-Dell ○ mistral-7b-v0.1 ○ model-blend

Tell me

□ Send to all models simultaneously

Figure 5. Sample web interface for a deployed model from Dell Enterprise Hub

The containers also support API calls using the OpenAI-compatible Messages API. Developers can create dynamic applications that call the API endpoints to create custom interactions with the deployed model.

Conclusion

The Dell Enterprise Hub is the first ever portal designed by Hugging Face for on-premises deployment of generative AI models. This initiative is intended to meticulously tackle prevalent business challenges that industries face today, including supply chain security, data traceability, scalability, and stringent regulatory compliance.

The impetus behind the Dell Enterprise Hub stems from the need to simplify the integration and management of AI technologies in a business environment. Dell Technologies recognized the necessity of providing a solution that enhances the operational efficiency of AI deployments and makes these advanced technologies accessible to all sectors without requiring extensive in-house technical expertise.

By harmonizing Dell Technologies' robust hardware capabilities with Hugging Face's cutting-edge AI models, the Dell Enterprise Hub offers a seamless, user-friendly platform that significantly eases the AI deployment process. The portal's tiered architecture facilitates easy access, management, and scaling of AI operations, ensuring that AI models run optimally with minimal latency on specialized hardware. This integration is supported by sophisticated backend tools that manage resources efficiently, removing the manual overhead and simplifying the user experience.

Dell Enterprise Hub Key Features and Advantages:

- **Optimized Containers:** The AI models are delivered in containers that are rigorously tested and validated by engineers from both Hugging Face and Dell

Technologies. These containers are finely tuned to leverage the full potential of Dell hardware, ensuring superior performance and reliability.

- **Dynamic Updates and Scalability**: The platform is continually updated with the latest AI models and technological enhancements. This dynamic approach ensures businesses can access the most advanced tools without constant manual updates or upgrades.
- **Simplified Deployment and Management:** With single sign-on, detailed access controls, and comprehensive logs, the portal ensures that deploying and managing AI models is straightforward and secure. The added benefit of priority support from the Hugging Face Team enhances the overall user experience and troubleshooting capabilities.

The Dell Enterprise Hub redefines the landscape of on-premises AI model deployment. For businesses, the takeaway is clear: deploying state-of-the-art AI solutions has never been more accessible, secure, and optimized. This initiative addresses specific operational challenges and empowers organizations to leverage generative AI effectively, transforming their processes and securing a competitive edge in a rapidly evolving digital era.

For more information about Dell Enterprise Hub on Hugging Face, please check out these additional resources:

Model Selection Made Easy by Dell Enterprise Hub

Model Plug and Play Made Easy by Dell Enterprise Hub

Model Merging Made Easy by Dell Enterprise Hub

Open-Source RAG Made Easy by Dell Enterprise Hub

Code Assistant Made Easy by Dell Enterprise Hub

AI Agents Made Easy by Dell Enterprise Hub

References

Dell Technologies documentation

The following Dell Technologies documentation provides other information related to this document. Access to these documents depends on your login credentials. If you do not have access to a document, contact your Dell Technologies representative.

- Artificial Intelligence – Dell Technologies Info Hub
- End-to-End AI is Within Reach — Are You Ready?
- Fast Track AI With Dell PowerEdge XE9680
- Dell Scalable Architecture for Retrieval-Augmented Generation (RAG) with NVIDIA Microservices

Hugging Face documentation

The following Hugging Face documentation provides other information related to this document. Access to these documents depends on your login credentials. If you do not have access to a document, contact Hugging Face.

- Dell Enterprise Hub on Hugging Face
- Dell Enterprise Hub – Frequently Asked Questions
- Hugging Face AutoTrain