Optimize Machine Learning Through MLOps with Dell Technologies and cnvrg.io

A Dell Technologies Overview

December 2022

H19116.1

White Paper

Abstract

This white paper defines MLOps, describes its purpose and increasing relevance for practitioners of machine learning and artificial intelligence, and details how organizations can use MLOps to optimize their machine learning deployments and bring prototypes into production faster. It also summarizes the Dell Technologies Validated Design for AI, designed in collaboration with and to support cnvrg.io implementations.

Dell Technologies Solutions

Dell Validated Design

intel cnvrg.io

Dell Technologies

Copyright

Contents

Introduction

Executive summary

Intelligence—the ability to comprehend and react—is demonstrated both as natural intelligence in humans and animals and as artificial intelligence (AI) in computers and systems. AI can deliver significant benefits to organizations by better enabling the categorization and analysis of massive amounts of data collected through mechanisms like sensors, networks, and always-connected devices.

Organizations today use AI to support progress towards their business intelligence goals and to solve problems. Machine learning (ML) is the most critical enabler of AI, employing algorithms and learning models to parse large datasets. Image and speech recognition are two examples of applications where employing ML can improve AI far faster than human analysis can. But data scientists working with ML spend an inordinate amount of time designing, configuring, and testing ML platforms, causing these highly trained specialists to spend too little time working with the data and models, which is their primary specialty. By using a process called MLOps (Machine Learning combined with Operations)—inspired by the popular DevOps model for application development—organizations investing in ML can more easily automate the continuous training and deployment of ML models at scale. MLOps adapts DevOps principles and practices to the ML workflow.

Combining development and operations in DevOps creates a more standardized and accelerated methodology for application development and deployment. Similarly, by automating the complexity and variability of the ML process, MLOPs helps lead to far more reproducible, testable, and evolvable ML environments.

Standardized, predictable, and manageable ML deployments can drive the launch of new capabilities, discoveries, and services, enabling an organization to derive more insights and value from the data it collects. These insights and values are the goal of MLOps.

Document purpose

The information provided in this white paper enables data scientists and others working with ML to understand how they can better use their data skills by using MLOPs to minimize platform complexity and associated administrative work. By taking advantage of an MLOps environment with cnvrg.io software, data scientists and IT professionals can move ML models out of the lab and into production faster and more easily, bringing a better return on investment (ROI) for the organization's ML investments.

Note: The contents of this document are valid for the described software and hardware version. For information about updated configurations for newer software and hardware versions, contact your Dell Technologies sales representative.

Audience

This white paper is intended for data scientists, solution architects, system administrators, and others developing ML and AI applications.

Revision history

Date	Version	Change summary
April 2022	1.0	Initial release
December 2022	1.1	Added support for Symworld Cloud Native Platform (formerly known as Robin Cloud Native Platform)

The challenges of implementing machine learning

Market environment

We live in an accelerated world. Organizations are pressed to move quickly to meet the demands of markets and stakeholders, gain resources, maintain competitiveness, or maintain service levels. The long planning cycles and steady growth of lumbering, giant organizations and agencies have been challenged by nimble competitors that are either smaller but empowered to move faster or by large organizations that have continuously invested in quicker processes to drive new capabilities. Our ability to collect data is greater today than it has ever been, enabling organizations to amass staggeringly large amounts of information. Without powerful tools to turn that data into actionable insights quickly, the marginal value of the data is severely limited.

ML is a powerful tool that is now available to organizations of all kinds and sizes, enabling them to identify trends and patterns more easily from massive quantities of diverse data without requiring human interaction. Including so-called “deep learning (DL)”—which uses a multitude of highly adaptive analysis layers for the greatest possible sophistication in pattern recognition and description—ML is a key new strategy to reduce the time and increase the consistency and efficiency of day-to-day analysis and decision-making.

ML uses algorithms that allow computer systems to “learn” and, ultimately, support informed decisions based on that learning—AI, in other words. DL applies to highly interconnected datasets, stringing together multiple algorithms across the adaptive analysis layers using artificial neural networks for the greatest flexibility in observation and analysis. It also allows for the most impressive machine capabilities such as speech and image recognition, biochemical design, and climate science.

Increasingly, organizations are turning to ML and AI to attack their new data challenges, working to shorten the lead time between problem and answer with the automation and efficiency of ML. In this new area of development, however, there is a challenge that arises from mismatches in skills within teams, leading to suboptimal execution of ML strategies.

Data scientists have the domain expertise on data and extrapolation, and it is in the interest of the organization that they concentrate most of their time on these areas. However, they are spending significant time outside of ML modeling and training, dealing instead with the design, configuration, and testing of the infrastructure as performance and reliability are so critical to project success. As organizations built up their data science capabilities, not as much attention was paid to the other key aspects of ML life cycle, resulting in gaps when it came to model testing, deployment, monitoring, and optimization. As infrastructure is not the domain of data scientists and not where their expertise and training can deliver the greatest value, IT needed to be part of the equation. There needed to be a way to bring IT and data science together.

Comparison of MLOps with DevOps

For the past two decades, IT has steadily adopted DevOps for application development and delivery. DevOps features tight integration of software and infrastructure teams, allowing responsive creation, customization, and management of increasingly sophisticated application stacks. It also provides shorter times to value and more reliable scaling through capabilities such as continuous integration/continuous deployment (CI/CD). When large numbers of systems needing to be deployed combine with always-limited IT resources, the fastest way to deploy tested and proven stable systems was to avoid the need for numerous configuration iterations. Fusing developers with IT operations into unified DevOps teams and working on much smaller and faster integration exercises enabled faster response, innovation, and deployments.

Initial adoption of ML and AI proceeded without a similar model. The drawbacks of leaving the platform tasks to a separate IT team are amplified in ML and AI projects, which are driven to move and evolve more quickly than user-focused software. For example, continuous testing is easier to implement in standard web and enterprise applications because the underlying elements are relatively stable and consistent, but ML must use more volatile elements, requiring more continuous training and validation. No single standard yields the best performance in ML because conditional environments and changing pipelines require someone with domain expertise—that is the data scientist—to ensure optimization.

ML challenges

ML and AI are highly influenced by both complexity and tuning. By default, every ML and AI pipeline contains multiple component steps, some of which are highly repeatable and some of which might be highly specific to a particular problem. Even if a particular pipeline can be run for multiple workloads, there is a significant benefit to tuning and optimizing for the required dataset. ML development is a full cycle process with multiple phases. As an organization goes from modeling to testing to deployment, management, and monitoring, there are a host of challenges at each step in the process.

For instance, in the modeling phase, data scientists might have to vary parameters, adjust settings, or change variables and then evaluate the outputs to see how closely the initial small amount of test data aligns to their goals. The multiplicity of versions and model hosts (containers) that arises can be complex enough to make organizing and management of the models too time-consuming, slowing development, and delaying release.

When a model appears to be valid, testing does not end, as larger amounts of data are fed into the model and a new cycle of testing occurs. When deployed, all the pipeline process steps must be recorded and archived as the data output is highly dependent on the exact pipeline configuration and steps. For repeatability and audit reasons, it is not enough to deploy; there must be extensive recording of the deployment parameters.

Constant monitoring of the solution is also required. Just as workers need to constantly monitor factory robots that have automated physical production, organizations cannot simply deploy an ML solution and move on, because changes in data over time may have a fundamental impact on outputs. A continuous cycle of modeling, testing, deploying, managing, and monitoring is important for the best quality control, just as you would expect from an assembly line.

The result is that far too much time can be spent setting up the proper ML and AI environment and not enough time on the development of working AI/ML systems. Clearly, optimizing the deployment process can drive increased results and provide more value faster; but almost 85 percent of AI/ML projects fail to deliver on their goals and a staggering 47 percent never even make it out of the lab.[1] The disconnect between effort and result is a direct consequence of the lack of process standardization in ML and AI.

It is not a case, however, of simply applying the DevOps methodologies directly to ML. The organization must recognize that the practice of data science—where application development is the subservient, not primary function—needs to change, bringing it closer to IT as the two forces begin to work together to effect real change through MLOps. MLOps is a powerful path forward, allowing data scientists to get out from behind the system configuration console and back to driving ever more sophisticated and useful results.

Implementing MLOps

MLOps is a defined process and life cycle for ML data, models, and coding. The MLOps life cycle begins with **data extraction and preparation** as the dataset is massaged into a structure that can effectively feed the model. **Experimentation** on the data then occurs as data scientists conduct short experiments to initially gauge the usefulness of the outcomes. **Model training** follows as the algorithm is given data to process and "learn." Next, the outputs are **evaluated and validated** against the business objectives and expected outcomes. When the models are adjusted sufficiently, they can be **deployed** to start processing the data. Constant **monitoring** occurs along the way to ensure that the process is running smoothly. **Automatic retraining** can be implemented to help adjust the deployed process and tighten up the results with each iteration. (See the following figure.)

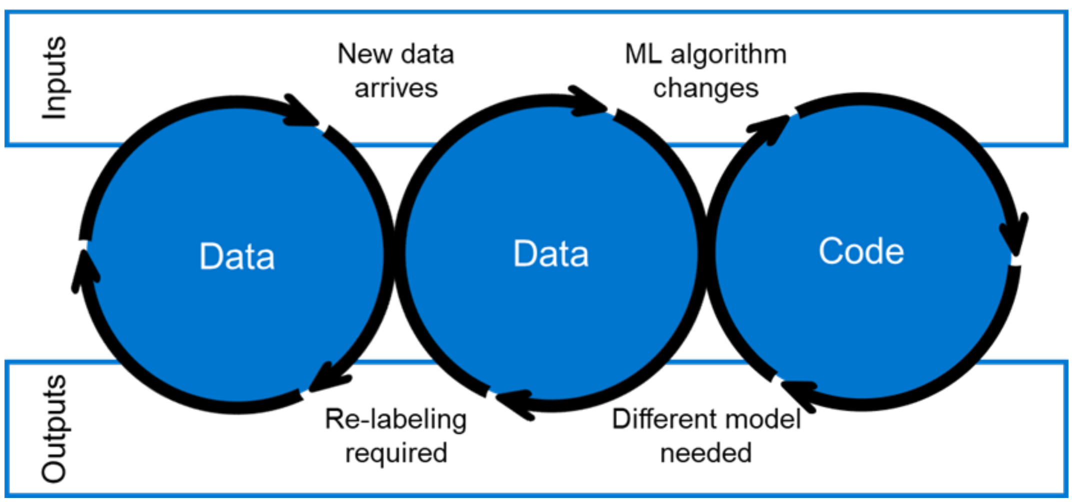

Figure 1. The MLOps pipeline

[1] Source: https://www.infoworld.com/article/3639028/why-ai-investments-fail-to-deliver.html

The primary goals of MLOps are:[2]

- Combining the ML platform, data, and software into a single, manageable release cycle
- Automating the testing of data, models and integration
- Introduction of Agile Principles into ML
- Treating ML models and datasets as first-class citizens in a CI/CD world
- Reducing technical variances across ML models
- Creating an agnostic environment concerning language, framework, platform, and infrastructure

A key difference between ML and traditional IT workloads is that ML tends to be more hardware- and platform-driven. ML, especially in DL use cases, has a wider range of underlying hardware types that can include more discrete compute elements. These compute elements include graphics processing units (GPUs) used in this case for general acceleration of highly parallelized analytics), tensor processing units (TPUs) used to support faster neural networks, and other specialized accelerator components. Because of this difference, MLOps needs to comprehend not only the classic data and compute models but must also comprehend more of the many compute elements that can underly each stage of the ML pipeline.

Another difference is that MLOps must go beyond the infrastructure and application focus of DevOps. A proper MLOps structure needs to address all three domains of machine learning: data, ML models, and code.

cnvrg.io for MLOps

A platform for managing the MLOps life cycle

Available either as self-hosted software or as the Metacloud SaaS offering, cnvrg.io delivers a full stack MLOps platform that helps simplify continuous training and deployment of AI and ML models. With cnvrg.io, organizations can automate end-to-end ML pipelines at scale in all environments, whether on premises or across clouds. cnvrg.io makes it easy to place training or inferencing workloads on CPUs, GPUs, TPUs, and other specialized accelerators depending on the wanted cost and performance trade-offs. Developers get a cloud-like self-service experience for choosing compute and storage resources from market leaders like Dell Technologies.

Regardless of the components in the pipeline, the result is a single end-to-end flow designed to maximize workload performance, optimizing with the right hardware and processing elements beneath each stage in the flow. ML jobs can be launched on demand, regardless of the underlying compute and storage elements.

Whether from the command line, SDK, or an intuitive visual interface, cnvrg.io provides access to all models, code, and datasets that can be run across an organization's compute and storage resources. Utilization and efficiency are boosted as data scientists can aggregate and use the best components, optimized for the job and then orchestrate those flows, all from a unified graphical interface. Through native Kubernetes pod and

[2] Source: ml-ops.org

cluster orchestration and cnvrg.io's internal job scheduler, both cloud and on-premises resources can be easily scaled to meet the computational and storage needs of an organization's ML workloads.

All these capabilities help remove friction and latency from the data science process, getting models out of the lab and into production quicker and reducing time to value for of the data. By removing much of the underlying complexity from the model, data scientists can spend more time delivering insight and spend less time dealing with configuration and testing. With cnvrg.io, enterprises can apply MLOps for continuous training and deployment of ML in the way that DevOps principles enable CI/CD for traditional IT workloads.

ML workflows with cnvrg.io

A typical workflow for an ML project using cnvrg.io has multiple steps, as shown in the following figure:

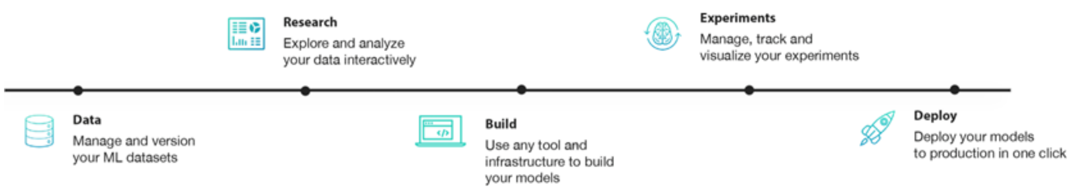

Figure 1. The cnvrg.io workflow (source: cnvrg.io)

The workflow begins with the data. Through cnvrg.io, the data scientist can manage and version multiple ML datasets from a single console without the demands for close administration or management of the underlying infrastructure.

When the data is defined, the next step is research, where data scientists can interactively model the data to analyze and visualize the different potential outcomes.

Coding starts as the data scientists begin building the solution. Because cnvrg.io enables repositories and categorization of pipeline elements, there is a library of existing elements from which to choose. Or new, custom constructs can be created if there is no existing analog. At this level, data scientists are not only working with the software elements but also matching workloads to the underlying compute elements, including processing and storage.

Experimentation then follows, with the ability to manage, track, and visualize the different scenarios and configurations, choosing the best combination of models and coding to meet your organization's needs.

When the optimal configuration is achieved, a single mouse click takes that complete pipeline from the lab to production where the data, models, and code can begin yielding results immediately. As this new solution moves into production, all the key elements are tracked in cnvrg.io for potential reusability.

cnvrg.io is an advanced modeling tool. The different data sources, models, and underlying elements can be added to an ML project easily through a visual workflow canvas, as shown in the following figure:

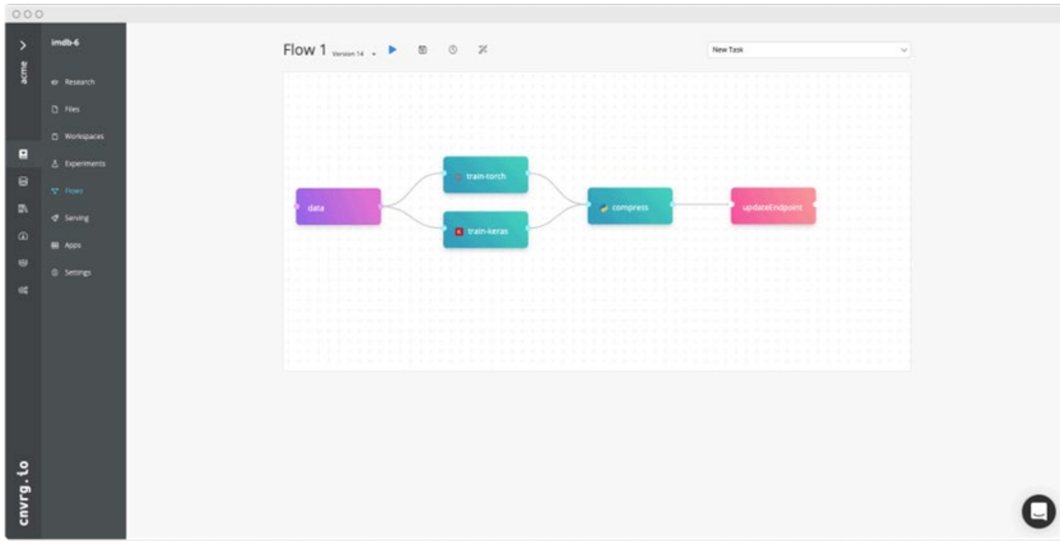

Figure 2. Creating an ML workflow visually with cnvrg.io (source: cnvrg.io)

Key features

The cnvrg.io platform enables the following elements of ML deployment and management:

- **Creation of ML pipelines**—Pipelines are key to the ML processes as raw data is ingested, cleansed, and transformed for use in modeling. cnvrg.io provides a drag-and-drop graphical interface for constructing and managing end-to-end AI and ML pipelines. These pipelines span from training to production and include dashboards for monitoring the state of jobs and resources, code reusability, and traceability.
- **AI library management**—The cnvrg.io AI Library is a native package manager for model code that promotes both code reusability and collaboration providing all the necessary container management for software from various open-source and OEM-provided registries. Data scientists can choose from open-source models or their own models through GitHub and Bitbucket integration. Changes can be pushed back to the remote repository when experimentation is complete.
- **Heterogeneous compute**—cnvrg.io abstracts compute and storage hardware into a cloud-like utility, making it easy to deploy jobs to various compute and storage resources with seamless management across environments. This capability means that locality of data or even locality of ML processing is not important, giving researchers more flexibility to handle and work with data or ML models in a location-agnostic manner.
- **Centralized dataset control and centralized version control**—Enables data scientists to connect and share any data from any source, creating datasets in a storage-agnostic manner. This centralized version-controlled system tracks every stage and all actions are committed, benefiting version control, explainability, and regulatory compliance.
- **Orchestrator and workload scheduler**—Uses any Kubernetes distribution such as VMware Tanzu, Symworld Cloud Native Platform, RedHat OpenShift, SUSE

Rancher, or one that is compliant with Cloud Native Computing Foundation (CNCF) as an orchestration, scheduling, and scaling layer. It makes jobs portable across environments and scales pods and clusters up and down on demand. cnvrg.io also uses Kubernetes' own native mechanisms, such as taints and tolerations, to place workloads only on appropriate nodes.

- **Machine learning tracking**—Automatically tracks and stores model code, statistics, and artifacts, bringing better control over the process with easy reproducibility and comprehensive monitoring. Real-time monitoring and interaction enable greater accuracy, performance, and reproducibility of models.
- **Machine learning model deployment**—Reliable deployment with automatic monitoring for zero down time. Deploy any ML model with a single click using various deployment models or interfaces.
- **Scalable inferencing endpoints**—Deploy large-scale and real-time machine learning to production in one click. cnvrg.io enables developers to serve models in batches, with RESTful APIs, or with streaming endpoints for real-time use cases.

cnvrg.io architecture

cnvrg.io has created an architecture with an ecosystem that enables an organization to build and deploy ML and AI on many infrastructures, including Dell optimized infrastructure and Dell Validated Designs for AI, as shown in the following figure:

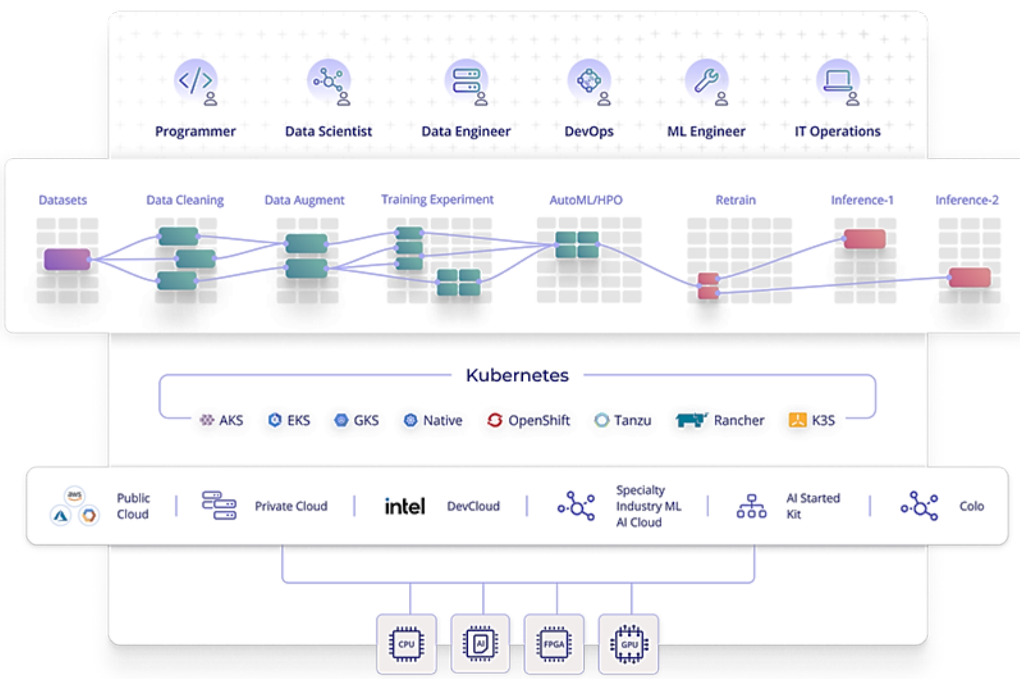

Figure 3. The cnvrg.io architecture (source: cnvrg.io)

This ecosystem is enabled by the cnvrg.io Operating System for AI that consists of the following components:

- **Control Plane**—The Control Plane is the management layer designed to operate with any CNCF-compliant Kubernetes distribution. The control plane provides a "single pane of glass" to manage all the elements in the ML stack, including datasets, model code, jobs, model performance, cluster, and resource statistics.

From the control plane, a data scientist has a consolidated view of all the elements open for manipulation in the project.

- **AI Library**—The AI Library is a package manager for the algorithms and data components. cnvrg.io integrates with GitHub, allowing data scientists to add their own custom repositories as well.
- **Pipelines**—The pipelines are where the all-important specialized work of the data scientist occurs. The drag-and-drop console enables a data scientist to build a complete end-to-end ML process that begins with data and ends with model serving and monitoring.
- **Orchestration and Scheduling**—Through the Kubernetes-based metascheduler, cnvrg.io enables all the tools for orchestrating pods, containers, jobs, and scaling resources across clusters.
- **Compute and Storage**—The compute and storage layers enable assignment of the underlying platform elements to pipeline stages, optimizing for the best type of compute, whether it be CPU, GPU, or any other specialized compute element.

Benefits

Adopting cnvrg.io delivers significant MLOps value to an organization as it accelerates the time to result for particularly data-driven problems:

- **Faster experimentation and model development**—Model optimization requires many tests with different model features and weights in parallel (also known as hyperparameter optimization). By streamlining and optimizing this process, MLOps shortens the cycle by empowering the data science team to train and test many models at once.
- **Faster deployment of updated models into production**—As noted above, only a small majority of ML experiments are deployed into production, and even then, the process takes an average of nine months (depending on the business requirements, models, model update schedule, and so on). MLOps automation and reusability principles help to quicken training and deployment.
- **Better quality assurance**—cnvrg.io gives data scientists and ML engineers full visibility into model performance, drift and erratic behaviors, as well as impacts on cluster heath and resource consumption. Users can either completely automate the retraining of models as new data arrives or put humans in the decision loop when expert judgment is needed. Improving the quality of model performance improves the quality of the answers coming out of the data.

The Dell Validated Designs for AI: MLOps

Optimizing ML through MLOps with cnvrg.io

Dell Technologies has worked closely with cnvrg.io to deliver MLOps for AI/ML adopters through a jointly engineered and tested solution to help organizations capitalize on the benefits of MLOps for ML and AI workloads, as shown in the following figure:

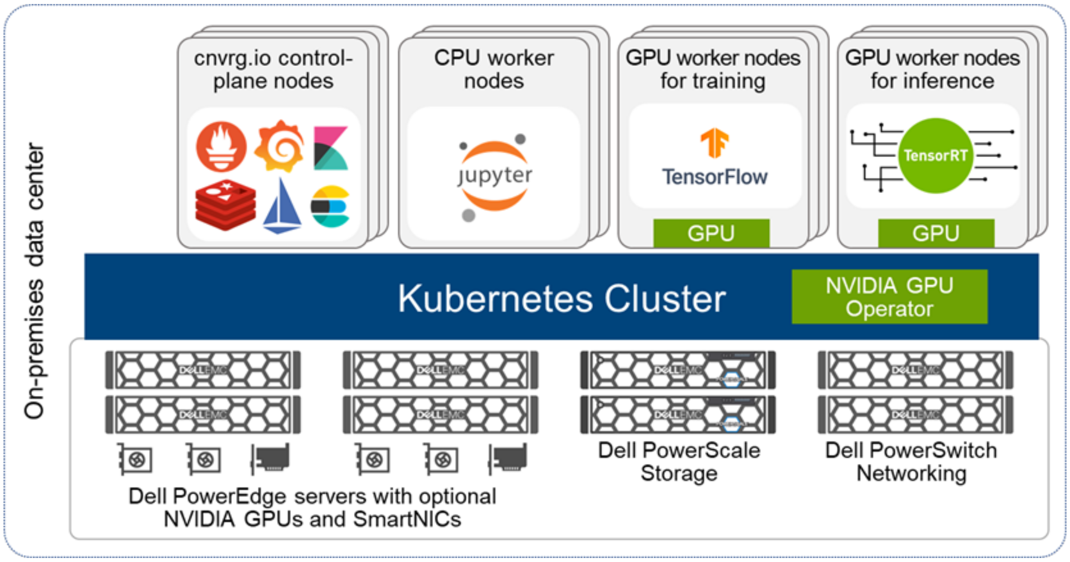

Figure 4. Overview of components of the Dell Validated Designs for AI: MLOps

The *Optimize Machine Learning Through MLOps with Dell Technologies and cnvrg.io Design Guide* provides guidance for architecting, deploying, and operating MLOps in the data center.

The design guide validates the cnvrg.io MLOps platform on two Kubernetes orchestration platforms:

- Dell Validated Design for AI provided by *Virtualizing GPUs for AI with VMware and NVIDIA*, which uses VMware vSphere with Tanzu for the Kubernetes layer and NVIDIA AI Enterprise software for additional applications, frameworks, and tools that researchers, data scientists, and developers can use to build ML models and analyze data
- *Dell Validated Design for Analytics—Data Lakehouse*, which uses Symworld Cloud Native Platform and open-source NVIDIA GPU operator

Powered with Dell PowerEdge servers for compute (with optional NVIDIA GPUs) and coupled with Dell PowerScale storage, the solution provides the analytics performance and concurrency at scale critical to consistently feeding the most data-hungry ML and AI algorithms.

Note: Additional Dell Validated Designs for container platforms can be used with similar results, but they have not been validated by Dell Technologies as of the publication of this white paper.

Deployment and support

Dell Technologies is also ready to support the joint solution by linking people, processes, and technology to accelerate innovation and enable optimal business outcomes:

- Consulting Services help you create a competitive advantage for your business. Our expert consultants work with companies at all stages of data analytics to help you plan, implement, and optimize solutions that enable you to unlock your data capital and support advanced techniques, such as AI, ML, and DL.
- Deployment Services help you streamline complexity and bring new IT investments online as quickly as possible. Leverage our 30 plus years of experience for efficient and reliable solution deployment to accelerate adoption and ROI while freeing IT staff for more strategic work.
- Support Services driven by AI and DL will change the way you think about support with smart, groundbreaking technology backed by experts to help you maximize productivity, uptime, and convenience. Experience more than fast problem resolution—our AI engine proactively detects and prevents issues before they impact performance. Select ProSupport Plus for a single point of contact for software and hardware support.
- Payment Solutions from Dell Financial Services help you maximize your IT budget and acquire the technology that you need today. Our portfolio includes traditional leasing and financing options, as well as advanced flexible consumption products.
- Managed Services can help reduce the cost, complexity, and risk of managing IT so that you can focus your resources on digital innovation and transformation while our experts help optimize your IT operations and investment.
- Residency Services provide the expertise needed to drive effective IT transformation and keep IT infrastructure running at its peak. Resident experts work tirelessly to address challenges and requirements, with the ability to adjust as priorities shift.

The Dell Technologies Customer Solutions Center

Adopting cnvrg.io delivers significant MLOps value to an organization as it accelerates the time to result for particularly data-driven problems:

- **Proof of Concept**—Validate that your preferred solution meets your needs with a custom Proof of Concept. Dell Technologies solution architects enable practical, hands-on implementation based on your test cases.
- **Design Session**—Collaborate with Dell Technologies experts to design a solution framework. Brainstorm with our experts to explore your current IT environment, your future objectives, and potential solutions.
- **Technical Deep Dive**—Dive into the technical solution details that you are considering for your organization. Learn from live product demonstrations and solution-focused discussions with Dell Technologies subject matter experts.

Dell Validated Designs for container platforms can be used with similar results, but they have not been validated by Dell as of the publication of this white paper.

Conclusion

Optimizing machine learning through MLOps

Machine learning is changing how organizations work, enabling them to use massive datasets that help drive decision making and automate processes. However, the deployment, configuration, testing, and operation of ML are critical to ML success on an ongoing basis, requiring data scientists to attend to operational and administrative infrastructure details and sapping the time that they would otherwise spend applying their expertise. DevOps processes have helped optimize the delivery of traditional IT workloads and services, and a similar set of methodologies, tailored for the life cycle and resource needs of ML, are needed to get the most out of the data.

This new set of methodologies—MLOps—is the approach to the deployment and management of ML solutions that can drive better ROI from ML and AI by allowing data scientists to focus more on model development, experimentation, and training and less on operational concerns. This focus empowers the organization to get ML and AI solutions out of the lab and into production much faster, driving more insights, better decision-making, and quicker, more assured achievement of goals. While roughly half of the ML projects never leave the lab, Dell Technologies and its partners are targeting this gap by delivering the cnvrg.io MLOps deployment and management solution that enables organizations to truly embrace ML and deliver far greater value.

We value your feedback

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

Terminology

The following table provides definitions for some of the terms that are used in this document.

Table 1. Terminology

Term	Definition
ML	Machine learning—A methodology that applies algorithms that can be automatically improved through experience and training with data
AI	Artificial intelligence—A system that perceives its environment and acts based on the machine intelligence (system-based) rather than natural intelligence (human- or animal-based)
DL	Deep learning—The stringing together of multiple algorithms across layers to create an artificial neural network that is designed to learn and make intelligent decisions on its own
MLOps	Machine learning + [IT] operations—A methodology that applies DevOps principles to ML environments
DevOps	[Application] development + [IT] operations—A set of practices in which application development and IT operations are combined into cohesive unified processes to provide continuous development and delivery pipelines that can iterate and improve systems more quickly and reliably

References

Dell Technologies documentation

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

- *Design Guide—Optimizing Machine Learning Through MLOps with Dell and cnvrg.io*
- *Design Guide—Virtualizing GPUs for AI with VMware and NVIDIA*
- *Implementation Guide - Virtualizing GPUs for AI with VMware and NVIDIA*
- *Design Guide—Dell Validated Design for Analytics—Data Lakehouse*
- *Design Guide—Red Hat OpenShift Container Platform 4.6 on Dell EMC Infrastructure*
- *Implementation Guide—Red Hat OpenShift Container Platform 4.6 on Dell EMC Infrastructure*

cnvrg.io documentation

The following cnvrg.io resources provide further additional and relevant information related to the MLOps platform discussed in this document.

- *cnvrg.io home page*
- *cnvrg.io Documentation*
- *Machine Learning Workbench: A Unified Code-First Data-Science Stack*
- *cnvrg.io for MLOps in the Dell Autonomous Driving Ecosystem*