DKube: An End-to-End MLOps Platform for Cost-Effective On-Premises Deep Learning

Dell EMC and One Convergence combine to offer a deep learning solution that executes on a unified, scalable, high-performance cluster & data analytics platform, coordinated using the Bright Cluster Manager

MLOps Where Your Data Is

Currently, the majority of deep learning users are doing their work on one of the public cloud platforms. There are a wide variety of easy-to-use applications that enable data scientists to prepare, train, and deploy their models.

Most of the cloud-based applications are tuned to the cloud, and some of them will only run on a specific cloud platform. Most of them address a piece of the overall deep learning puzzle.

The public cloud, though convenient, has significant drawbacks. Cloud-based solutions can be expensive for large, complex models and datasets, and the latency necessary to move those datasets back and forth can be limiting.

Some users cannot use the cloud for reasons of security, privacy, or corporate governance. This is especially true in markets such as **pharmaceutical/chemical research, medical imaging, defense**, and **manufacturing**.

For users who need secure access to their data, an on-premises approach is the only viable option. However, creating a working, production-ready on-premises deep learning cluster is often not an option. It is too time-consuming and complicated to pull together all of the software components and make them run reliably.

And once the initial deployment has been painstakingly implemented, supporting it and keeping it up-to-date makes the resulting solution hard to maintain.

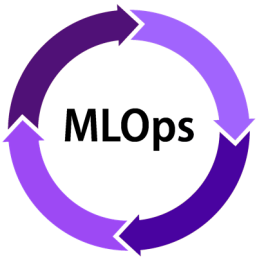
Most users also require the ability to move back and forth between their on-premises system and the cloud, and wish to maintain the same look, feel, and database for this migration.

In order to address this reality, One Convergence has created the **DKube deep learning application**. It provides an end-to-end MLOps platform that runs on-premises, on the cloud, or in a hybrid system consisting of both.

Resilient, On-Premises MLOps that Scales with You

The availability of affordable high performance computing components has moved complex problem-solving into the mainstream. Highly capable CPUs, AI-focused GPUs, high bandwidth storage subsystems, and powerful software frameworks are the foundation of these systems, but it is challenging to make them run together in a reliable manner.

In order to marry the advantages of on-premises computation, described in the sidebar, with the simplicity of cloud-like use, **One Convergence** is offering their flagship **DKube** application on **Dell EMC** platforms.



One Convergence has created DKube to offer an end-to-end **MLOps** capability for deep learning. Engineers can develop proof-of-concept models, optimize them, train them on large datasets, deploy them to production, and monitor them to ensure strong inference results.

The solution is truly **end-to-end**. Data scientists can develop their models on a Dell EMC workstation, and migrate their work seamlessly to train on an HPC cluster – taking advantage of high performance GPUs and as much scalability as they require. They can then deploy them for inference on yet another platform.

The Dell EMC HPC cluster allows the DKube deep learning application to share the same system with other numerically-intensive workloads such as simulation and data analysis.

The DKube application and the underlying hardware platform are both designed for resiliency. Mutliple nodes can be installed on the cluster, and the system will continue to operate even when some of the components become degraded or are inoperable.

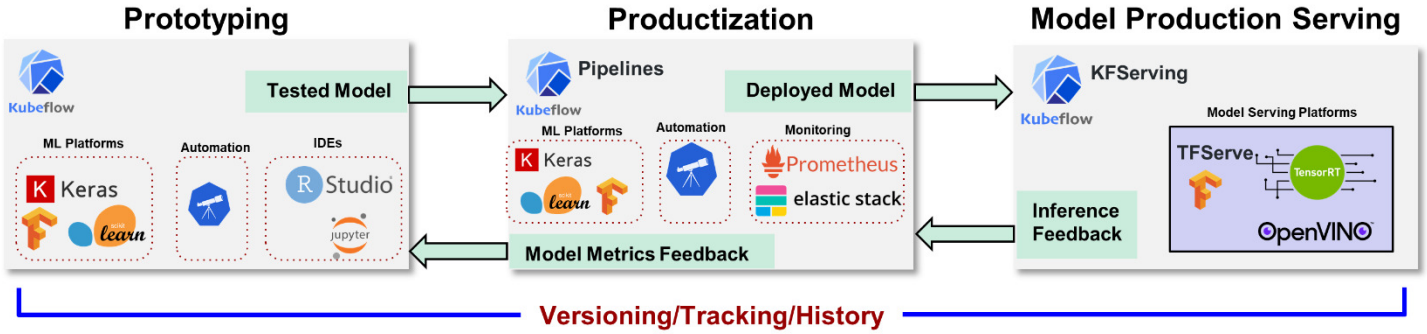
The Bright Cluster Manager (BCM) brings all of these capabilities together, managed by a single pane of glass through BrightView. Bright integrates directly with Dell EMC's iDRAC management system, allowing changes to BIOS settings and system firmware from the same console.

The Dell EMC HPC platform is described in more detail at https://www.dellemc.com/resources/en-us/asset/technical-guides-support-information/solutions/h18161-hpc-ra-for-ai-da-ra.pdf

DKube: An End-to-End MLOps Platform

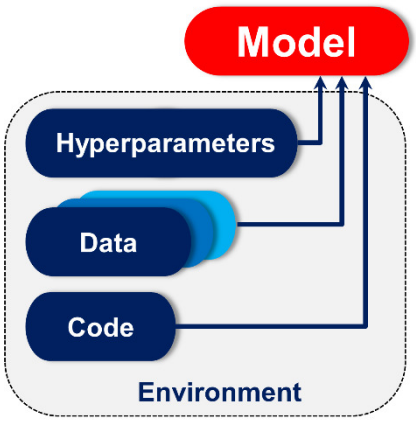
Deep learning consists of more than simply running model code through a notebook and identifying a few hyperparameters to optimize your output.

Deep learning development must incorporate a **full MLOps flow** to be useful. The eventual goal is to deploy and maintain an accurate model on a production server.



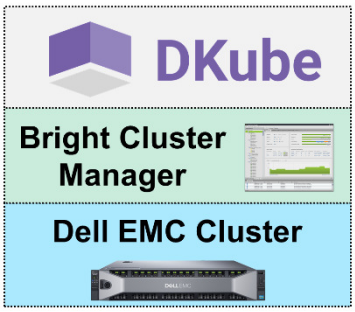
Secure, Flexible Workflow for the Enterprise

- The process must accommodate a small team, a large organization that spans many people – or anything in between.
- For large organizations the handoff between functions must be **clear** and **formal**.
- Workflow **governance** is critical. Privacy and security are important from both an ethical and regulatory point of view, and mandatory in some markets. The data must be **tracked** as it makes its way through the process, and be easily **reproducible** for later **audit**.
- It is important that any model provide complete insight into its **provenance**. The code, datasets, hyperparameters, and platform hardware/software details must be easily determined, and the **metadata** must follow the model through its life cycle.
- Since model development is iterative, **versioning** capability must integrate with the overall solution.



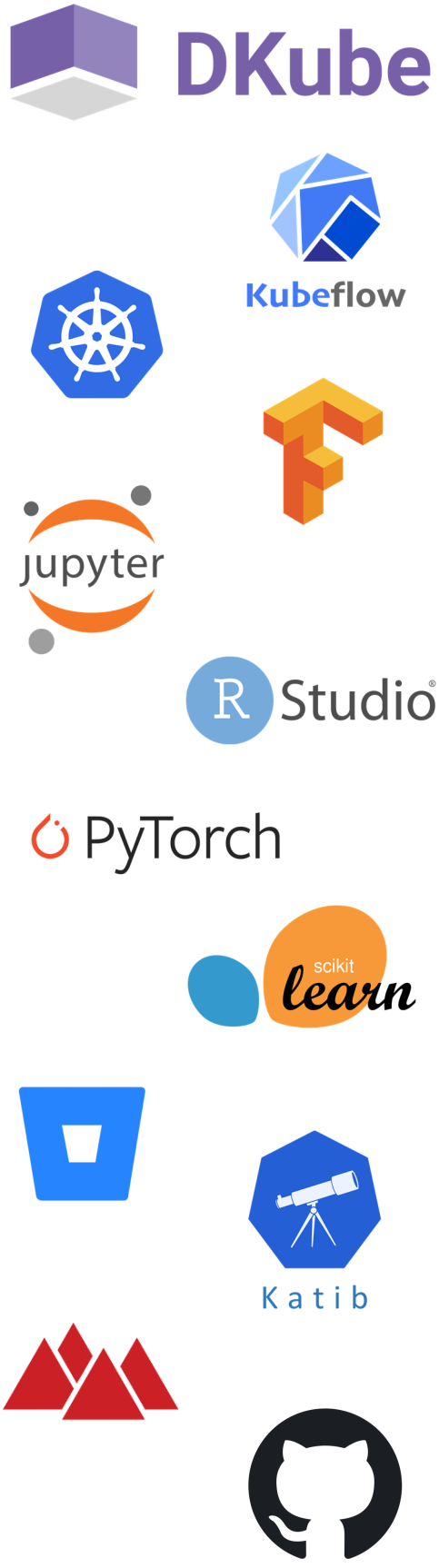
Flexible Deployment

- Since there is no perfect production server, the model must be deployable on a variety of frameworks. Different requirements for performance, cost, load balancing, and proximity mean that the serving capability must be flexible.



Focus on Deep Learning, Not Infrastructure

- The application must be **easily installed** on the target platform, and the workflow must be intuitive.
- The resources on the hardware platform must be **automatically** identified and configured for use. Allocation of resources must be flexible, such that they are **securely shared** to maximize their utilization.



Deep Learning where Your Data Is

End-to-End MLOps Platform

DKube provides a complete environment for deep learning. Data Engineers, Data Scientists, ML Engineers, and Production Engineers coordinate to seamlessly deliver powerful models.

Bare Metal to MLOps in Hours

Installation is simple and quick. One Convergence handles the difficult task of integrating the required standards, enhancing the flow and capability with added value software components, and ensuring that they work together in a GUI-based production environment.

On-Prem, Cloud, & Hybrid Platforms

Although an on-prem platform is most appropriate for large datasets and compute-intensive workloads, there may be situations when a cloud environment makes sense. DKube allows seamless migration between on-premises use and the cloud.

Versioning and Lineage

Complex models and datasets require substantial experimentation and iteration to get right. In order to keep all that activity in some manageable form, DKube provides full versioning. In addition, a particular run or model can be traced back to show all of the inputs used to create it, and a dataset can be traced forward to understand where it has been used.

Multi-Tenancy & Collaboration

Resources and manpower are both expensive. Resources on the system can be allocated to groups of users, and shared among those users. The resources can be easily moved between the groups based on utilization. This maximizes resource usage, and minimizes engineer queueing. Users share models and data, offering secure collaboration.

Support for Popular Standards

The system supports the most popular experimentation and training frameworks, including **Jupyter** and **RStudio** for model development, and **TensorFlow**, **Keras**, **PyTorch**, and **Scikit Learn** for training. Authentication is supported for **GitHub** and **LDAP**, and code management is provided by **GitHub** and **Bitbucket.** Hyperparameter optimization is offered using **Katib**, and DKube is compatible with **Kubeflow** Pipelines, providing full automation.

Scale with Your Workload

Different workloads require different resource envelopes. Model development and experimentation might need only CPUs on a workstation, while training will make use of powerful GPUs on scalable clusters. You can also scale-out by adding more resources, or scale-up by adding more nodes. DKube automatically identifies resources and makes them available for use.

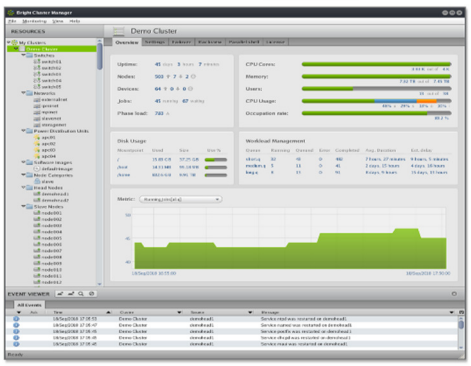
Resiliency

DKube provides multi-node resiliency to address critical, high availability environments.

High-Performance Dell EMC Platform

slurm workload manager

kubernetes



Bright Computing

The Scale of HPC. The Flexibility of Containers. All in One System.

The DKube deep learning MLOps application is tuned to run on a powerful, flexible Dell EMC cluster. The platform is tailored for a wide variety of high-performance computing (HPC), artificial intelligence (AI), and data analytics (HPDA) workloads. The resource pool is dynamically assigned and managed by an HPC resource manager, or to containerized AI/DA workloads orchestrated by the open-source Kubernetes container orchestration system. This allows IT managers to change the balance of the system over time to meet the evolving computational needs of users.

Built Using Dell EMC's Accelerated Compute Platforms

The **Dell EMC PowerEdge R740** can be configured as a compute-only platform, or as an accelerated platform which supports Nvidia Volta V100 GPUs, Nvidia Tesla T4 GPUs, or Intel® Performance Accelerator Cards (PACs) with Intel® Arria 10GX Field Programmable Gate Arrays (FPGAs).

The **Dell EMC PowerEdge C4140** is a purpose-built dense accelerated compute platform which can be configured with four (4) Nvidia Volta V100 GPUs with NVLink.

The **Dell EMC DSS8440** is a 4U purpose-built dense accelerated compute platform which can support up to ten (10) Nvidia Volta V100 GPUs for maximum performance, or with Nvidia Tesla T4 GPUs for power/performance optimization.

Connected with High-bandwidth Network Fabric

The Dell EMC HPC RA for AI/DA can have one or more network fabrics, ranging from **25 Gigabit Ethernet to 100 Gigabit InfiniBand** low-latency fabric from Mellanox Technologies, which is tailored specifically for HPC environments.

Avoid Data Motion with Scalable Shared File Storage

Avoid unnecessary data motion with shared storage from Dell EMC. Independently scale storage requirements using Dell EMC Isilon, Dell EMC Ready Solution for HPC BeeGFS Storage, or Dell EMC Ready Solutions for HPC PixStor Storage.

Single Pane of Glass Management

Using Bright Cluster Manager (BCM), the Dell EMC systems bring all these capabilities together, managed by a single pane of glass through BrightView. Bright integrates directly with Dell EMC's iDRAC management system, allowing changes to BIOS settings and system firmware from the same console.

Balance and Scale Your Resources

Your workload balance changes. Avoid idle resources and excessive wait times by moving compute resources from HPC scheduling to Kubernetes, and back again as needed. Independently scale your shared storage to create the right balance of compute and storage for your needs.