DELL Technologies

Solution Brief

Dell EMC PowerScale and the cnvrg.io Data Science Platform

Accelerating ML/AI workloads by a dataset caching tier

Dell EMC PowerScale benefits

- High performance AFA storage that scales from TB to PB
- Industry's first ML cache tier for Data Science pipelines
- High speed connectivity for ML compute clusters
- Scale from a single node (entry level) to multi-node clusters
- Kubernetes aware and compatible with Red Hat OpenShift

cnvrg.io benefits

- Full stack data science platform
- Provides dataset hub tightly integrated with PowerScale
- Container based, supporting multiple Kubernetes clusters – in the cloud or on-prem
- Code first platform, supporting Jupyter, VScode and R-Studio workspaces
- Certified as Red Hat OpenShift Operator
- Certified by NVIDIA NGC – integration with pre-packaged ML models
- Meta-scheduler technology - ability to schedule across disparate Kubernetes clusters, and other schedulers

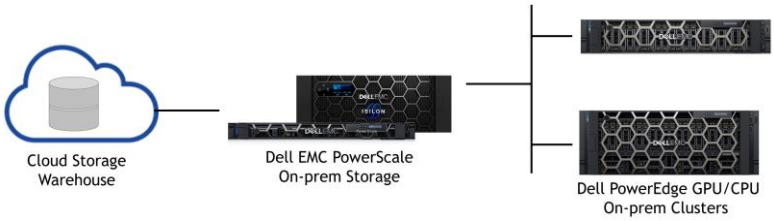
Machine Learning (ML) offers unprecedented promise to many applications, including automotive safety advancement, where the upcoming wave of fully autonomous vehicles must be able to accurately identify other vehicles, pedestrians, signage – really anything -- that might stray into their environment in real-time. In spite of all the interest, and obvious benefits, the promises remain largely unrealized, as data scientists are forced to spending 65% or more of their time on non-data science work. This is because most ML development efforts rely on a pipeline of ad-hoc tools, plug-ins, scripts and other siloed tools that are impeding organizations from streamlining ML development.

The most demanding element of the data science pipeline is the data used to train algorithms. Training requires vast datasets that need to feed GPU-accelerated compute servers for longer (and longer) runs. In real life, multiple datasets co-exist, each with many versions, stored in complex wide and deep directory hierarchies. This data feeds potentially different clusters of accelerated compute that are often spread across different physical locations, introducing variability in distance and latency that can dramatically impact GPU utilization, cost and ultimately, the quality of ML results.

Partnering with Dell Technologies and NVIDIA, a new ML architecture imperative was borne – solving the data proximity challenge by effectively bringing the data closer to the compute – by creating a dataset cache tier for ML models. With this new ML architectural paradigm, companies large and small can finally free their data scientists to focus on the science, not on managing the data.

Dell EMC PowerScale – dataset cache tier for ML pipelines

A key component of the cnvrg.io Enterprise Data Science Platform is the dataset hub, which was developed to provide a single, centralized repository for all the datasets to be managed. Hub management includes many capabilities including securing access, maintaining lineage of the datasets, creating different datasets versions, offering query meninism and more. With the cnvrg.io data science platform, when an ML pipeline is initially fired-up, the corresponding dataset, regardless where it is physically located, is copied once to a PowerScale cluster that is located nearest to the ML compute (e.g. one hop of 100GbE RDMA) creating a cached copy of the dataset. The data science platform then tracks its versions, usage and other metrics. Additional ML training runs will then access the PowerScale cache and pull the dataset across high speed interconnect. Other users no longer need to attach to remote storage for this same dataset, as they can rely on the cached copy. Creating different versions, queries and sub-sets will operate from the cache, without the performance lag caused by long latencies to/from a remote data lake or data warehouse. All datasets versions will operate from the local storage, implemented by the Dell EMC PowerScale cluster.

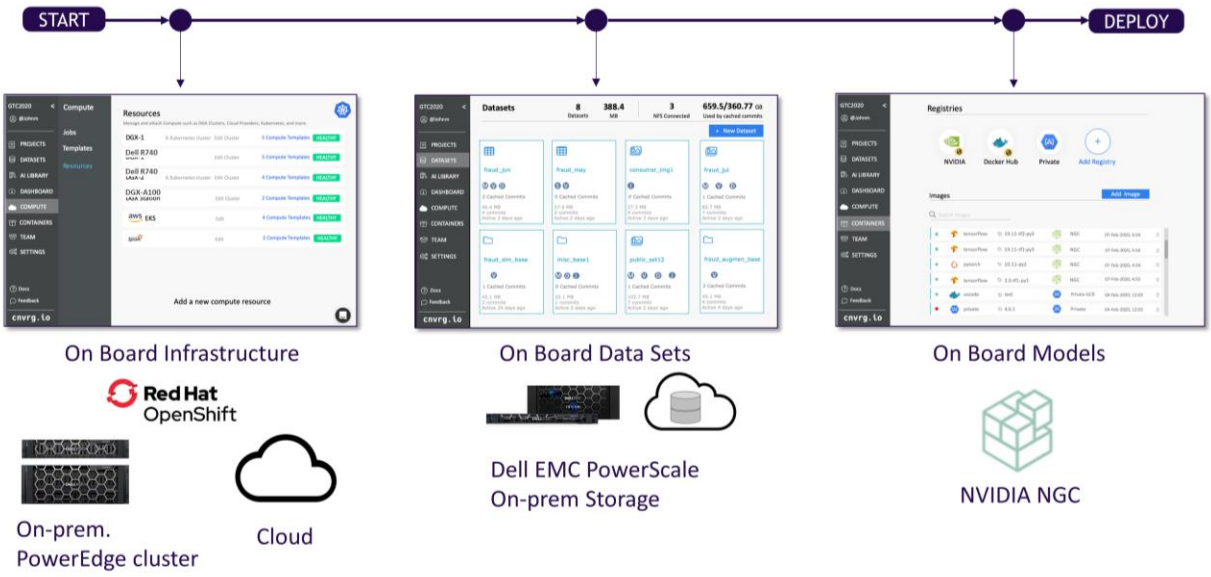


Accelerating ML/AI workloads with cnvrg.io

Solution details – Kubernetes-based dataset caching tier

The Dell and cnvrg.io solution incorporates the best in class Kubernetes managed clusters, cached datasets for extreme performance and the one-click attachments of models to datasets, with NVIDIA NGC integration. Now users can cache the needed datasets (and/or their versions) knowing they are located at the nearest PowerScale nodes attached to the GPU cluster or CPU cluster exercising the training. Once the needed datasets are cached they can be used multiple times by different team members. This solution provides all the elements of a complete ML infrastructure: storage, high throughput and low latency networking, GPU servers and CPU servers. Essentially Dell provides the complete physical ML infrastructure with cnvrg.io serving as the control plane for the data science platform. This solution delivers many business advantages including:

- **Productivity gains**: Datasets are ready to be used in seconds instead of hours
- **Sharing and collaboration**: the cached datasets can be authorized and be used by multiple teams, in the same compute cluster connected to the PowerScale cached data
- **Cost saving:** models are pulling the datasets from the cache, and not repeatedly from remote storage, which may require egress payments for each and every download
- **Cloud enablement**: The ML cache can be used as an on-prem storage mirror for the data-lake residing in the cloud
- **Tightly coupled integration with NVIDIA MIG** : cnvrg.io is the first data science platform to implement Multi-Instance GPUs, bring virtualization and GPU slicing to data science pipelines in one-click
- **ML Infrastructure Dashboard** – monitoring and insights dashboard for the stakeholders to measure utilization, jobs, capacity and create cost-center reports of the ML infrastructure usage across the users



The criticality of ML cache tier for hybrid cloud

It is no secret that hybrid cloud is a major trend for modern workloads, including ML. It is customary to have experiments and light developments in the cloud. But when there is a more structured business need, many of the ML models need to train on-prem. Dell solution for ML cache tier, provides the needed bridge between public and on-prem development, so there are no interruptions and unneeded delays when migrating the project from the public cloud to on-prem. cnvrg.io pipeline can run in the cloud and on-prem, and particularly the cache feature is hybrid cloud aware.

ML data, model and infrastructure fusion

cnvrg.io offers onboarding and caching of frequently used datasets creating a tightly integrated data science pipeline. Architected to be a container based platform, it is well integrated with the leading managed Kubernetes offerings including Red Hat OpenShift. cnvrg.io also attaches models directly from NVIDIA NGC catalogue, with a one click selection, minimizing data scientists' efforts to kickstart model development. Using cnvrg.io Data Science Platform combined with Dell Technologies PowerScale, user's can assign datasets with one click to the pipeline, while assuring data proximity to the compute for high performance.

About cnvrg.io

As the leading data science platform for MLOps and model management, cnvrg.io is a pioneer in building cutting-edge machine learning development solutions so you can build high-impact machine learning models in half the time. cnvrg.io was built by data scientists, for data scientists to streamline the machine learning process, so they can focus less on grunt work and more on the real magic – algorithms. Learn more about cnvrg.io.

About Dell EMC PowerScale

Dell EMC PowerScale provides an enterprise-grade, scale-out NAS platform that scales from terabytes to more than 10s of PB of capacity in a single file system. Industry-leading data protection guards against hardware failures and intentional or unintentional data corruption. PowerScale stays simple to manage, regardless of how large your automotive environment grows – reducing costs and allowing you to manage design development – not storage. Learn more about PowerScale.

Learn more about Dell Technologies automotive data storage solutions

Contact a Dell Technologies Expert

Join the conversation with #DellEMCStorage

DELL Technologies