E-BOOK

Building Your MLOps Pipeline

NetApp

netapp.com/ai
The Promise of AI

Artificial intelligence (AI) and machine learning (ML) are quickly becoming foundational to business decision making. But AI is demanding and building an architecture to handle the data challenges of AI can be complex. This is especially true for organizations looking to integrate their DevOps pipeline with their AI pipelines.

AI applications are just another piece of code that can be developed, tested, and deployed using key DevOps capabilities, such as containers, automation, and cloud, to speed innovation. While containers and Kubernetes offer powerful tools for accelerating AI and ML, home-grown solutions built on opensource tools can prove unreliable and difficult to manage—particularly at scale.

As AI moves from pilot to production, organizations need an enterprise-grade architecture to bring the latest cloud-native DevOps and AI technologies together in an MLOps pipeline. This eBook explores the challenges of using containers to deploy portable cloud applications for AI and ML and how to overcome them.

We will show you how to:
- Rapidly clone a data namespace
- Instantly create AI data and model baselines
- Provision Jupyter workspaces with access to full data sets
- Unify compute and data silos across sites and regions
The Rise of Containers in AI

Containers are quickly becoming an essential tool for enabling AI applications, offering developers a way to build and deploy lightweight, portable cloud applications for AI and ML. With an MLOps pipeline, data scientists can optimize their use of DevOps methodologies and cloud-native AI technologies to save time and money and get the resources they need to be successful.

Containerized applications can get complicated, however. You might require hundreds or thousands of separate containers in production. This is where container runtime environments, such as Docker, benefit from the use of other tools to orchestrate or manage all the containers in operation.

One of the most popular tools for this purpose is Kubernetes, a container orchestrator that automates deployment, management, and scaling functions for containerized applications.

Kubeflow, originally developed by Google, is an open-source AI and ML toolkit for Kubernetes that makes deployments of AI/ML workflows on Kubernetes simple, portable, and scalable. Kubeflow abstracts away the intricacies of Kubernetes, allowing data scientists to focus on what they know best—data science. As enterprise IT departments standardize on Kubernetes, Kubeflow has been gaining significant traction.

Tool Integrations
- Docker: Packages software into standardized units for development, shipment, and deployment.
- Kubernetes: An open-source, distributed, container orchestration platform originally developed by Google.
- Kubeflow: An open-source AI and ML toolkit for Kubernetes that was originally developed by Google.
The Promise of AI

The Rise of Containers in AI

Challenges to Implementing DevOps for AI

Integrate Your DevOps Pipeline with Your AI Pipeline

4 Ways to Prevent Unnecessary Toil

Next Steps

Challenges to Implementing DevOps for AI

As organizations work to optimize their infrastructures for MLOps, they face three major challenges:

- Workload scalability
- Data availability
- Deployment difficulty

There are many AI and DevOps frameworks in the marketplace that attempt to tackle workload scalability and deployment difficulty hurdles. Most of these frameworks, however, fail to address the problems of data availability and data mobility. Many feature proprietary data platforms that lack proven enterprise-class reliability and don't scale across different sites and regions.

Because many of the most powerful AI and DevOps tools, including Kubernetes and Kubeflow, come from the startup and webscale world, many of them are open-source or bespoke, which can be difficult for enterprises to implement and manage at scale. With data scientists having to manage their own home-grown infrastructures, they can waste valuable time and effort performing tedious routine tasks and troubleshooting issues. As enterprises move their AI initiatives from pilot to production, they need a solution that can improve reliability, break down data silos, and accelerate innovation.

As more enterprises develop AI pilots into production, IT leaders will need to deliver infrastructure stacks for DataOps, ModelOps and, most importantly, platforms that support the integration of streaming-data analytics into the enterprise architecture.1

---

Integrate Your DevOps Pipeline with Your AI Pipeline

The NetApp AI Control Plane is NetApp’s full-stack AI data and experiment management solution. It provides extreme scalability, streamlined deployment, and nonstop data availability—when and where you need it.

The NetApp AI Control Plane integrates Kubernetes and Kubeflow with a data fabric enabled by NetApp, which provides uncompromising data availability and portability—from edge, to core, to cloud.

Using NetApp Trident, NetApp’s open-source, persistent storage provisioner for Kubernetes, data volumes stored within your data fabric can be presented to Kubernetes workloads in an enterprise-class, Kubernetes- and cloud-native format.

With Trident, your developers and data scientists can create, manage, and interact with persistent storage volumes in the standard Kubernetes format that they are already familiar with. At the same time, they can take advantage of NetApp advanced data management capabilities and a data fabric that is powered by NetApp technology. Perhaps best of all, NetApp AI Control Plane is built completely on open-source components, so you can implement the solution today on any NetApp storage in your environment.

Figure 1) NetApp AI Control Plane architecture.
4 Ways to Prevent Unnecessary Toil

What Can You Do with NetApp AI Control Plane?

- Rapidly clone a data namespace—just as you would a Git repository
- Instantly create AI data and model baselines for traceability and versioning
- Swiftly provision Jupyter Notebook workspaces with access to massive datasets
- Aggregate and unify AI compute and data silos across sites and regions

1) Rapidly Clone a Data Namespace

For most AI developers, data scientists, and engineers, having access to quality data is critical for training, testing, and deployment. It can often take hours to copy production data sets for testing, wasting valuable time and resources.

With NetApp AI Control Plane, data scientists can rapidly clone a data namespace just as easily as they would a Git repository. Our validated architecture integrates NetApp replication technologies with Kubeflow to enable data scientists to create an exact copy of the production dataset with the press of a button.

“Data scientists and data engineers are struggling to manage AI implementation details today. The types of capabilities offered by the NetApp AI Control Plane can help simplify AI operations. Demand for these capabilities is growing as enterprises engage in more AI initiatives. Vendors that offer an ecosystem of solutions can help end users simplify the AI environment and avoid piecemeal solutions.”

- Ritu Jyoti, Program Vice President, Artificial Intelligence Research and Global AI Research Lead, IDC
4 Ways to Prevent Unnecessary Toil

2) Instantly Create AI Data and Model Baselines for Traceability and Versioning

AI traceability isn’t impossible, but it can be hard. If you have an ever-changing source dataset and you’re not saving versions of it each time you train a new model, you don’t have traceability. This can make it difficult to go back in time to troubleshoot. And, in certain industries, such as financial services, it could be a problem for compliance.

The NetApp AI Control Plane enables you to save and tag a baseline version of your source dataset and the model you’re training for traceability and versioning, making it fast and easy to locate and access a specific data set tied to a specific model. This feature also makes it easy to compare models with A/B testing. Quickly test models against a copy or subset of your data for best results in production.

3) Swiftly Provision Jupyter Workspaces with Access to Full Data Sets

Jupyter is quickly becoming an industry standard for data scientists to document and share projects. However, Jupyter workspaces can be difficult to provision, often requiring data scientists to submit tickets to IT operations. Depending on the length of the queue, it could take hours, if not days, just to get a Jupyter workspace set up with the access data scientists need to do their work. In the meantime, they are stuck waiting.

NetApp AI Control Plane enables IT ops or even the data scientists themselves to spin up a workspace in seconds, so your data scientists can get straight to work. This enables faster time to market and allows you to focus valuable resources on more strategic tasks.
4 Ways to Prevent Unnecessary Toil

4) Aggregate and Unify AI Compute and Data Silos Across Sites and Regions

As major cloud providers release new features and tools for AI you want to be in the best position possible to take advantage of them. Data silos and vendor lock-in will only slow you down.

With a data fabric enabled by NetApp, you can seamlessly move data between clouds, from on-premises storage to the cloud, and back from the cloud to your premises. Even if your core work is on-premises, your data scientists can go out and experiment in the cloud, using new features as they become available. If your on-premises infrastructure is over-burdened and you need additional capacity, you can go out and work in the cloud.

NetApp gives your data scientists the freedom to go anywhere they need to go and use any feature or compute resources they need to get their work done. The data fabric also enables you to seamlessly move your data from where you’re collecting it at the edge to the core and to the cloud—wherever you’re using it to train models or perform analysis.

Figure 2) Enable seamless data movement across edge, core, and cloud with a data fabric powered by NetApp.
Next Steps


Learn more online at www.netapp.com/ai

Refer to the Interoperability Matrix Tool (IMT) on the NetApp Support site to validate that the exact product and feature versions described in this document are supported for your specific environment. The NetApp IMT defines the product components and versions that can be used to construct configurations that are supported by NetApp. Specific results depend on each customer’s installation in accordance with published specifications.

Software derived from copyrighted NetApp material is subject to the following license and disclaimer:

This software is provided by NetApp "AS IS" and without any express or implied warranties, including, but not limited to, the implied warranties of merchantability and fitness for a particular purpose, which are hereby disclaimed. In no event shall NetApp be liable for any direct, indirect, incidental, special, exemplary, or consequential damages (including, but not limited to, procurement of substitute goods or services; loss of use, data, or profits; or business interruption) however caused and on any theory of liability, whether in contract, strict liability, or tort (including negligence or otherwise) arising in any way out of the use of this software, even if advised of the possibility of such damage.

NetApp reserves the right to change any products described herein at any time, and without notice. NetApp assumes no responsibility or liability arising from the use of products described herein, except as expressly agreed to in writing by NetApp. The use or purchase of this product does not convey a license under any patent rights, trademark rights, or any other intellectual property rights of NetApp.

The product described in this manual may be protected by one or more U.S. patents, foreign patents, or pending applications.

Data contained herein pertains to a commercial item (as defined in FAR 2.101) and is proprietary to NetApp, Inc. The U.S. Government has a non-exclusive, non-transferable, non-sublicensable, worldwide, limited revocable license to use the Data only in connection with and in support of the U.S. Government contract under which the Data was delivered. Except as provided herein, the Data may not be used, disclosed, reproduced, modified, performed, or displayed without the prior written approval of NetApp, Inc. United States Government license rights for the Department of Defense are limited to those rights identified in DFARS clause 252.227-7015(b).

Trademark Information

NetApp, the NetApp logo, and the marks listed at http://www.netapp.com/TM are trademarks of NetApp, Inc. Other company and product names may be trademarks of their respective owners. NA-430-0520

© 2020 NetApp, Inc. All rights reserved.