White Paper

MLOps powered by NetApp and Modzy

Scaled AI inferencing with Modzy and StorageGRID

Sathish Thyagarajan, NetApp

Seth Clark and Kirsten Lloyd, Modzy

November 2022 | WP-7358

In partnership with

Abstract

NetApp® and Modzy have partnered together to deliver a new way of applying artificial intelligence (AI) and machine learning (ML) at scale to any type of data, including imagery, audio, text and tables. This white paper offers guidelines for customers to take AI/ML projects from planning to deploying and scaling production workloads. The NetApp StorageGRID®-Modzy integration allows organizations to bring AI/ML analysis, powered by Modzy, to data securely stored in StorageGRID®. The solution enables customers to accelerate the secure analysis and processing of data, at-scale, with faster speeds and lower latency and costs, ultimately producing high value AI-generated insights.

TABLE OF CONTENTS

Executive summary 3

Solution overview 3

NetApp StorageGRID 3

Modzy MLOps platform 4

MLOps powered by NetApp and Modzy 4

Solution technology 7

Hardware requirements 7

Software requirements 8

Real-world applications 8

Document scanning 8

Satellite imagery processing 8

Conclusion 8

Acknowledgments 9

Version history 9

LIST OF TABLES

Table 1) Hardware requirements. 8

LIST OF FIGURES

Figure 1: NetApp StorageGRID for globally distributed AI/ML applications. 3

Figure 2: Modzy platform for production AI. 4

Figure 3: Modzy MLOps with NetApp StorageGRID for AI at scale 5

Figure 4: NetApp StorageGRID-Modzy integration. 5

Figure 5 Explainability for AI model predictions. 6

Figure 6: Data-centric AI model retraining. 7

Executive summary

Extracting value from data requires the use of analytics, ML and AI to distill insights from raw information. However, turning that data and those insights into a competitive differentiator that can drive business value means that AI and ML must operate at massive scale with explainable models that can also be trusted by humans. NetApp and Modzy have partnered together to deliver a new way of applying AI and ML at scale to any type of data, including imagery, audio, text and tables. The NetApp StorageGRID® and Modzy integration delivers secure AI-enabled analysis at speed and scale on-premises, in the cloud, and everywhere in between. By bringing AI-enabled processing directly to data stored in NetApp StorageGRID's secure data storage solution in a hybrid cloud configuration, organizations can accelerate AI-enabled processing of large volumes of data at faster speeds and lower costs, generating high value insights with low latency, and reduce the need for unnecessary data transfer.

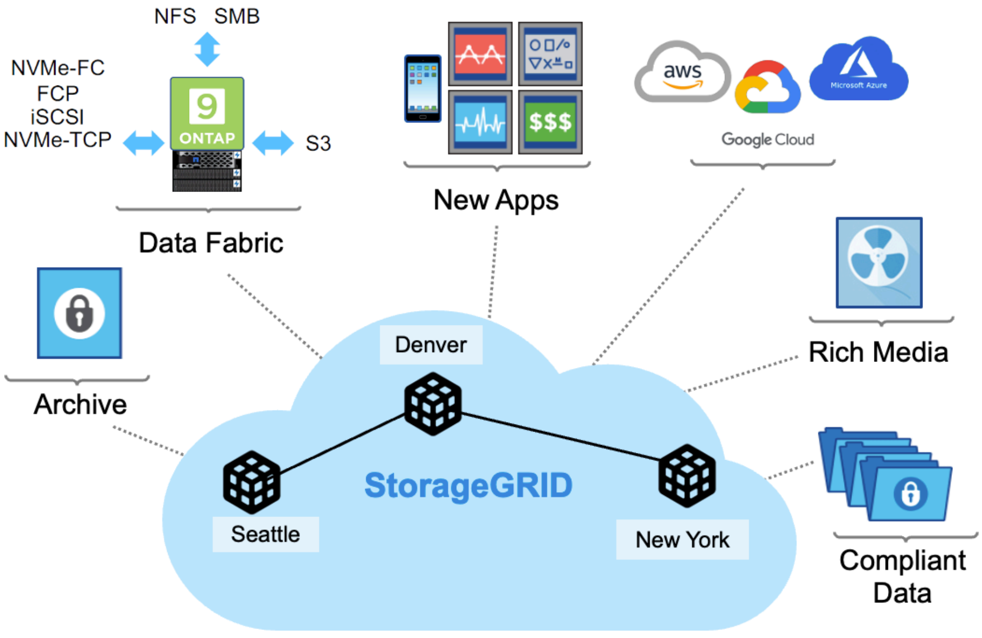
Solution overview

NetApp StorageGRID

NetApp StorageGRID is an enterprise grade software-defined object storage that natively talks S3 API. StorageGRID is a software defined solution that can be deployed on NetApp's purpose-built appliance, VMware based environments as well as bare metal environments. Within a single namespace StorageGRID can span across 16 sites within a single GRID. StorageGRID's intelligent policy-driven data management delivers industry-leading innovations such as automated lifecycle management to store, secure, and preserve data cost effectively over long periods, seamlessly moving data between on-premises and public cloud storage to optimize its availability, protection, performance, and cost. StorageGRID supports a wide range of use cases across public, private, hybrid or multi-cloud environments.

StorageGRID is composed of globally distributed, redundant, heterogeneous nodes, which can be integrated with both existing and next-generation AI/ML applications (Figure 1).

Figure 1: NetApp StorageGRID for globally distributed AI/ML applications.



Some of the advantages of the StorageGRID system include the following:

- Deploy multiple StorageGRID sites to access data from any location between data centers and the cloud through a single namespace that easily scales to hundreds of petabytes.
- Provide flexibility to deploy and centrally manage across infrastructures NetApp's purpose-built appliance, VMWare based environments and Bare Metal platforms.
- Provide unmatched durability with fifteen-nines of durability leveraging layered Erasure Coding (EC).
- Enable more hybrid multi-cloud capabilities with validated integrations into Amazon S3 Glacier and Azure Blob.
- Meet regulatory obligations and facilitate compliance through tamper-proof data retention, without proprietary APIs or vendor lock-in.

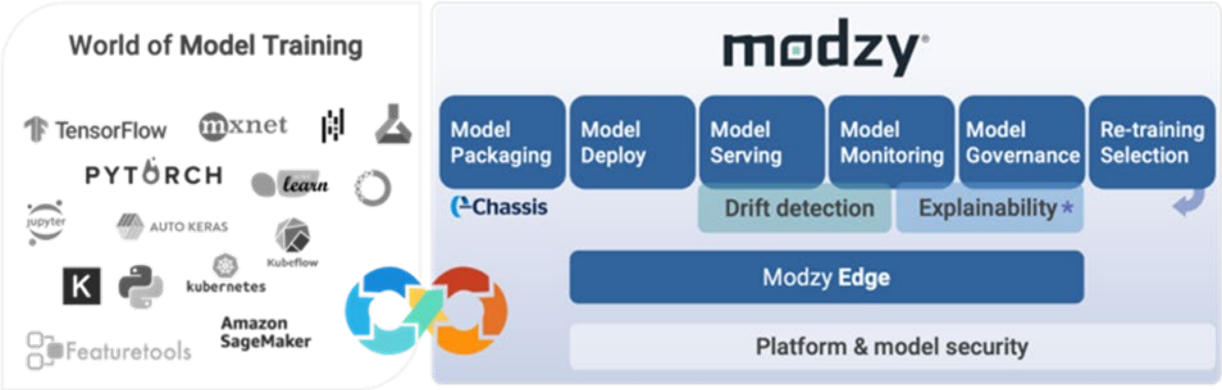
For more details you may refer to StorageGRID documentation and NetApp AI Solutions resources.

Modzy MLOps platform

Modzy is an MLOps platform for deploying, integrating, running, and monitoring AI models at scale, providing the missing link between the lab and production AI. With API connections for popular model training tools, data management solutions, CI/CD pipelines, and enterprise applications, Modzy provides an open and flexible solution that fits within an organization's existing tech stack. Modzy can be deployed to any combination of infrastructure – in the cloud, on-premises, hybrid, at the edge, meaning that AI models can be run and managed anywhere.

Modzy turns ML/AI models into API endpoints that can be integrated anywhere, with automated monitoring and governance to manage the full lifecycle of the models (Figure 2).

Figure 2: Modzy platform for production AI.

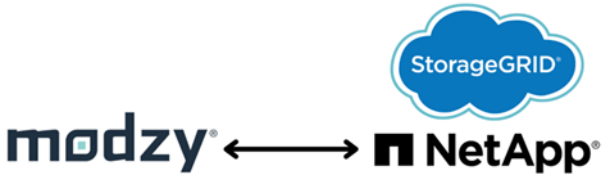


MLOps powered by NetApp and Modzy

The NetApp StorageGRID®-Modzy integration powers a central, secure solution to unify, govern, and analyze data for accelerated AI workloads. With flexible deployment options for cloud, on-premises, or edge deployments, organizations can quickly process and analyze data where it's collected, yielding fast results with low latency. With fast processing and features for model monitoring, explainability and drift detection, the integrated solution powers fast, AI-enabled analysis of data at scale, with results users can trust. By centralizing data management and AI-analysis, the integration also enables a data-centric retraining loop to update and keep models accurate, while ensuring data and models remain secure. Ultimately, the integration provides organizations with the ability to process AI workloads where data is stored, generating insights quickly, securely, with low latency at lower cost and without the need for unnecessary data transfer.

Modzy uses the information provided in the request to authenticate and request the specified file or files, which are then copied into Modzy from StorageGRID. Modzy processes the inputs from StorageGRID and saves the results (Figure 3).

Figure 3: Modzy MLOps with NetApp StorageGRID for AI at scale

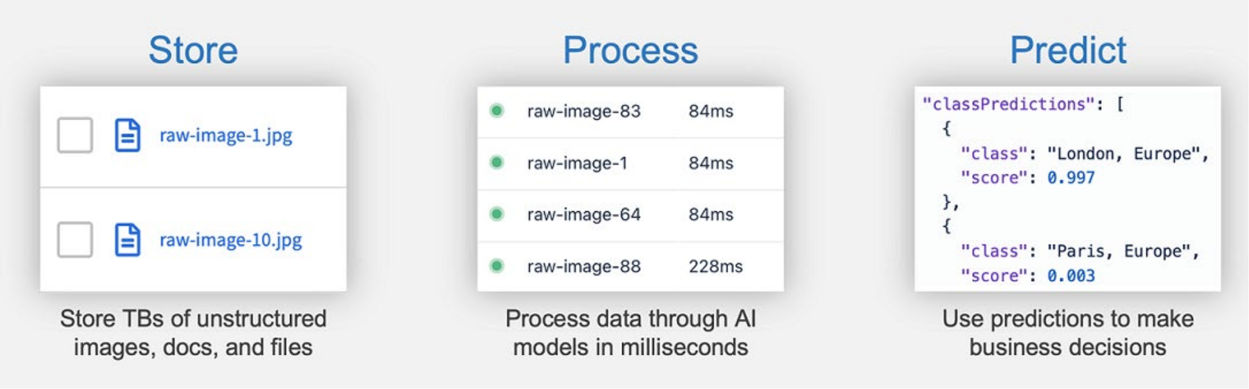


Large-scale AI processing

StorageGRID and Modzy enable large-scale batch processing of data through AI models, with high-quality results generated at lower cost. Modzy powers AI-enabled processing for any type of data stored in StorageGRID, and the integrated solution provides the most value in processing unstructured assets such as images, videos, audio files, documents, and other binary data formats.

The NetApp StorageGRID-Modzy integration unlocks parallelized inference at a massive scale by supporting batch processing, GPU acceleration, and dynamic queues. Data stored in StorageGRID buckets on-premises can be processed through ML models in 10s of milliseconds. This equates to processing 10s of thousands of images per minute on a single NVIDIA A100 GPU, through advanced AI models such as image classification, object detection, image segmentation, and object tracking (Figure 4).

Figure 4: NetApp StorageGRID-Modzy integration.



□ raw-image-1.jpg

□ raw-image-10.jpg

Store TBs of unstructured images, docs, and files

●	raw-image-83	84ms
●	raw-image-1	84ms
●	raw-image-64	84ms
●	raw-image-88	228ms

Process data through AI models in milliseconds

```
"classPredictions": [
  {
    "class": "London, Europe",
    "score": 0.997
  },
  {
    "class": "Paris, Europe",
    "score": 0.003
```

Use predictions to make business decisions

Hybrid deployment options

Accelerated AI analysis with StorageGRID and Modzy can be run nearly anywhere. Deploying StorageGRID and Modzy together generates significant costs savings and speeds up inference for AI processing by reducing unnecessary data transfer. StorageGRID supports on-premises and hybrid cloud object storage, ensuring that data is securely stored wherever it is needed. Modzy also offers cloud and on-premises deployment options, making it possible to co-locate AI processing with the large volumes of unstructured data stored in StorageGRID. Modzy Edge can also be used to support hybrid cloud deployments. Modzy can be centrally installed in AWS, Azure, or on-premises adjacent to a StorageGRID installation, thus enabling customers to run edge nodes co-located with other buckets in another public cloud or remote data centers. This offering ensures that AI inference happens as close to the data as possible, reducing the need for unnecessary data transfer.

Model management and monitoring

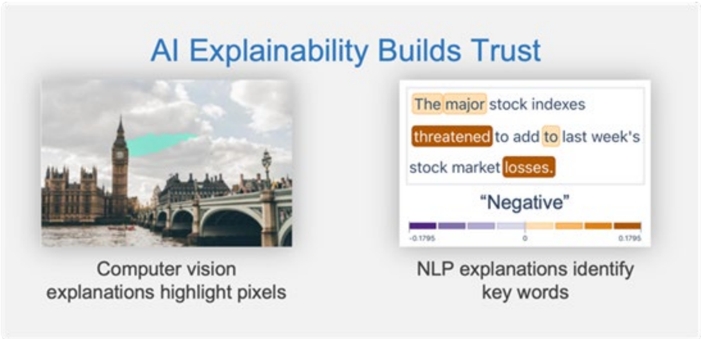
A major component of any MLOps solution is streamlined deployment, integration, and management of ML/AI models. Modzy provides organizations with a central location for all models to be deployed, run, and monitored, with explainability and drift detection, ensuring model performance remains robust. Organizations can deploy and execute their own custom-built models, open-source or commercial models, or pre-trained models from Modzy's marketplace to analyze data stored in StorageGRID. Model background information, like model architecture, training data, expected performance, and version history is documented, enabling model sharing and reuse amongst teams. From there, all model performance data is tracked, with the ability to see all job results, history and model performance details, like explainability for model results and drift detection. This enables full transparency and auditability for all AI-enabled predictions generated.

Explainability

Explainable AI solutions attempt to show how AI models make decisions in terms that humans understand, translating the process AI uses to transform data into real insights and value; this enables transparency and insight into the "why" behind model predictions.

Modzy makes it easy to add explainability to pipelines by adding the AI explainability code directly into a model container and flagging in the API request to turn on explainability; this includes support for Modzy's patent-pending explainability solution, or for popular open-source solutions like LIME or SHAP. Today, ML engineers rarely incorporate AI explainability solutions into their workflows because of integration challenges, and because they can significantly slow down the AI pipeline. Simplifying the process to integrate explainability into AI pipelines improves transparency and increases confidence and trust in AI-enabled decisions (Figure 5). For more information you may also refer to the following tech talk: Beyond LIME and SHAP: The fastest approach to AI Explainability.

Figure 5 Explainability for AI model predictions.


AI Explainability Builds Trust
The major stock indexes
threatened to add to last week's
stock market losses.
"Negative"
-0.1795
0
0.1795
Computer vision explanations highlight pixels
NLP explanations identify key words

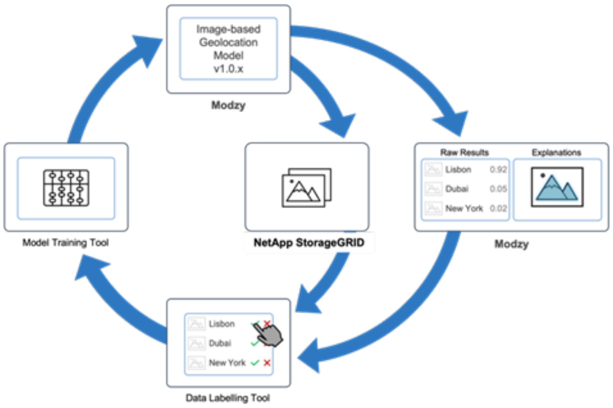
Drift detection

Machine learning models require careful monitoring and constant tuning to maintain peak performance. These measurements are often referred to as "model drift" as models drift away from accurate predictions over time. Modzy offers drift detection for computer vision and natural language processing, allowing data scientists to evaluate model performance against production data in real time. Data scientists can also examine the data associated with the greatest drift and find those data assets in StorageGRID for future analysis and review. Monitoring model drift is a central element to ensuring model performance evolves alongside changes in production data. For more information, please visit Data Drift Detection.

Data-centric model retraining

The NetApp StorageGRID®-Modzy integration can be used to create a data-centric model retraining workflow to update and improve models over time, reducing data labelling efforts and data transfer costs. StorageGRID provides an ideal data store for data-centric applications that use AI for advanced features such as computer vision or natural language processing. The architecture of a model retraining workflow starts with simultaneously sending production input files to both StorageGRID and Modzy. StorageGRID permanently saves these data assets for reuse in the future, while Modzy generates AI-powered predictions. Data inputs and AI outputs are named using a common reference so they can be mapped together in the future. When a model needs to be retrained, such as when it appears to be drifting, model predictions are imported from Modzy to a data labelling tool while a reference to the raw input is imported from StorageGRID. After re-labelling the incorrect predictions, a new model can be trained directly from the production data stored in StorageGRID. After retraining, the new and improved model can be deployed to production in just a few minutes, restarting the cycle. This approach reduces data labelling efforts, reduces data transfer costs, and accelerates the development of new models (Figure 6).

Figure 6: Data-centric AI model retraining.



AI security

The NetApp StorageGRID®-Modzy integration ensures that organizations meet the most stringent AI security requirements, while maintaining the flexibility to build and evolve their AI tech stacks over time. With a DISA-compliant open-source AI/ML model container specification, organizations can securely package models trained in any training tool or framework for production deployment. Modzy meets FISMA-moderate security controls and offers optional adversarial defense features that ensure models remain robust. For more information on Modzy's security features, please refer to the Modzy security overview white paper.

Security is particularly important for NetApp StorageGRID® because many types of rich content data that are well suited for object storage are also sensitive in nature and subject to regulations and compliance. StorageGRID® provides various features like S3 object lock, WORM compliance, encryption, Ransomware protection for cloud backups, and various other features within the product to help customers comply with regulations like SEC Rule 17a-4(f) or FINRA Rule 4511(c). For more information please refer to security features in StorageGRID and Ransomware protection for StorageGRID.

Solution technology

Hardware requirements

A minimal on-premises Modzy deployment requires three servers with the following configuration (high-availability setups might require more):

Table 1) Hardware requirements.

Operating system	VCPUS	RAM (GIB)	COUNT
UBUNTU 20.04 LTS	8	32	3 (or more)

Cloud formation templates are available for both AWS and Azure deployments.

For minimum installation requirements, please refer to Modzy minimum requirements documentation.

Software requirements

Modzy requires Kubernetes, S3-compliant object storage, and a PostgreSQL database. Modzy used StorageGRID 11.6 for this integration, testing and validation. However, the software components for StorageGRID® used in any implementation of a solution might vary based on customer requirements.

For platform specific software requirements please refer to StorageGRID documentation.

Real-world applications

Document scanning

The NetApp StorageGRID®-Modzy integration powers a simple and fast solution for processing documents, text files, and scanned PDFs. By storing these assets in StorageGRID, they can be processed through one or more natural language processing models such as optical character recognition (to extract text from images), language translation, and entity extraction. These results can be used to quickly find information embedded in documents that would otherwise take hours or days for humans to scan.

Satellite imagery processing

StorageGRID® and Modzy provide an excellent solution for satellite imagery applications such as environmental monitoring, economic activity detection, and global security. While StorageGRID provides an ideal environment for storing and hosting raw satellite imagery in the cloud or on-premises, Modzy has dozens of pre-trained models designed specifically for processing geospatial data such as overhead object tracking, geospatial image-registration, and infrastructure health monitoring. Modzy's integration with StorageGRID makes it possible to process huge volumes of satellite data either before or after storage.

Conclusion

By leveraging the NetApp StorageGRID®-Modzy integration, validated by the NetApp AI solutions and the Modzy team, organizations can accelerate the secure analysis of their data at-scale using AI and ML. By co-locating AI and ML processing with secure data storage, organizations reduce security risks from unnecessary data transfers, have lower costs and latency because of faster processing, and unlock insights hidden in data from AI-enabled analysis of data at-scale. By adopting an MLOps approach to AI management, users gain features for model monitoring, explainability and drift detection, all in service of generating results they can trust for various use-cases ranging from general object detection in imagery to geospatial analysis for ArcGIS and more. This also enables a data-centric model retraining loop to update and keep models accurate, while ensuring data and models remain secure. With flexible deployment options the NetApp StorageGRID®-Modzy integration can be run nearly anywhere, still while keeping security at the forefront for our customers.

To get started with this solution or if you have any questions, please reach out to the NetApp AI team at ng-ai-inquiry@netapp.com or the Modzy team at info@modzy.com.

Acknowledgments

- Joseph Kandatilparambil, Solution Architect, NetApp
- Raymond Vargas, Software Architect, Modzy

Version history

Version	Date	Document version history
1	November 2022	Initial release

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.

Copyright information

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.

LIMITED RIGHTS LEGEND: Use, duplication, or disclosure by the government is subject to restrictions as set forth in subparagraph (b)(3) of the Rights in Technical Data -Noncommercial Items at DFARS 252.227-7013 (FEB 2014) and FAR 52.227-19 (DEC 2007).

Data contained herein pertains to a commercial product and/or commercial service (as defined in FAR 2.101) and is proprietary to NetApp, Inc. All NetApp technical data and computer software provided under this Agreement is commercial in nature and developed solely at private expense. The U.S. Government has a non-exclusive, non-transferrable, nonsublicensable, worldwide, limited irrevocable 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) (FEB 2014).

Trademark information

WP-7358-1122