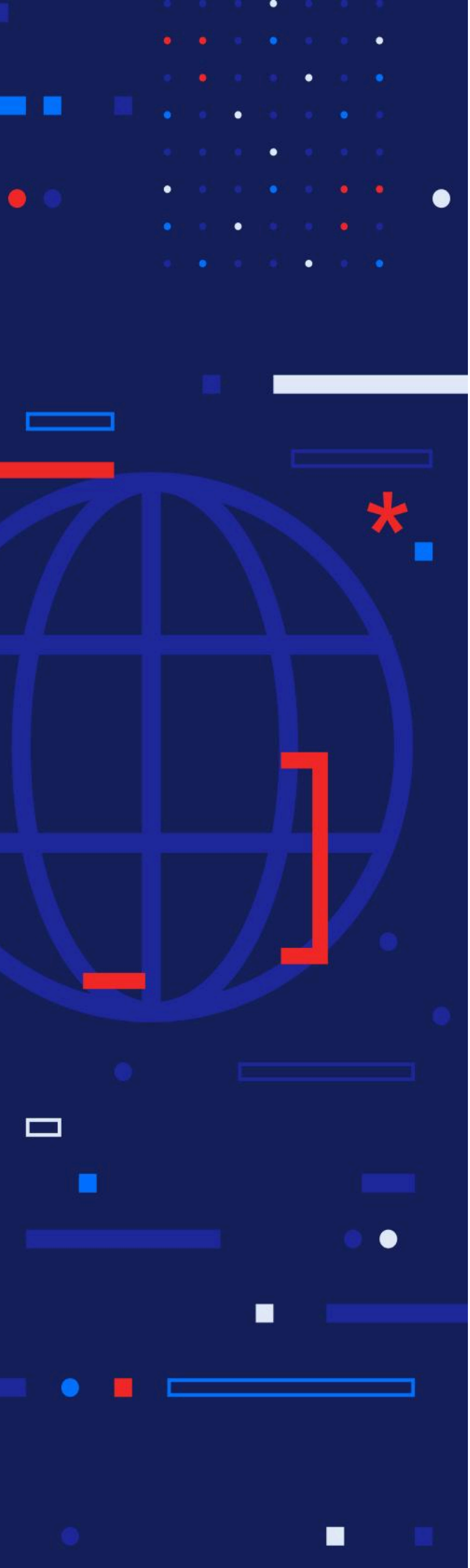


# Reference Architecture/ POC Whitepaper

DataNeuron LLM Studio
powered by NetApp


NetApp

**Balbeer Bhurjee**
balbeer.bhurjhee@netapp.com

**Shinil Vaish**
shinil.vaish@netapp.com

DataNeuron

**Bharath Rao**
bharath@dataneuron.ai

**Prakash Baskaran**
prakash@dataneuron.ai

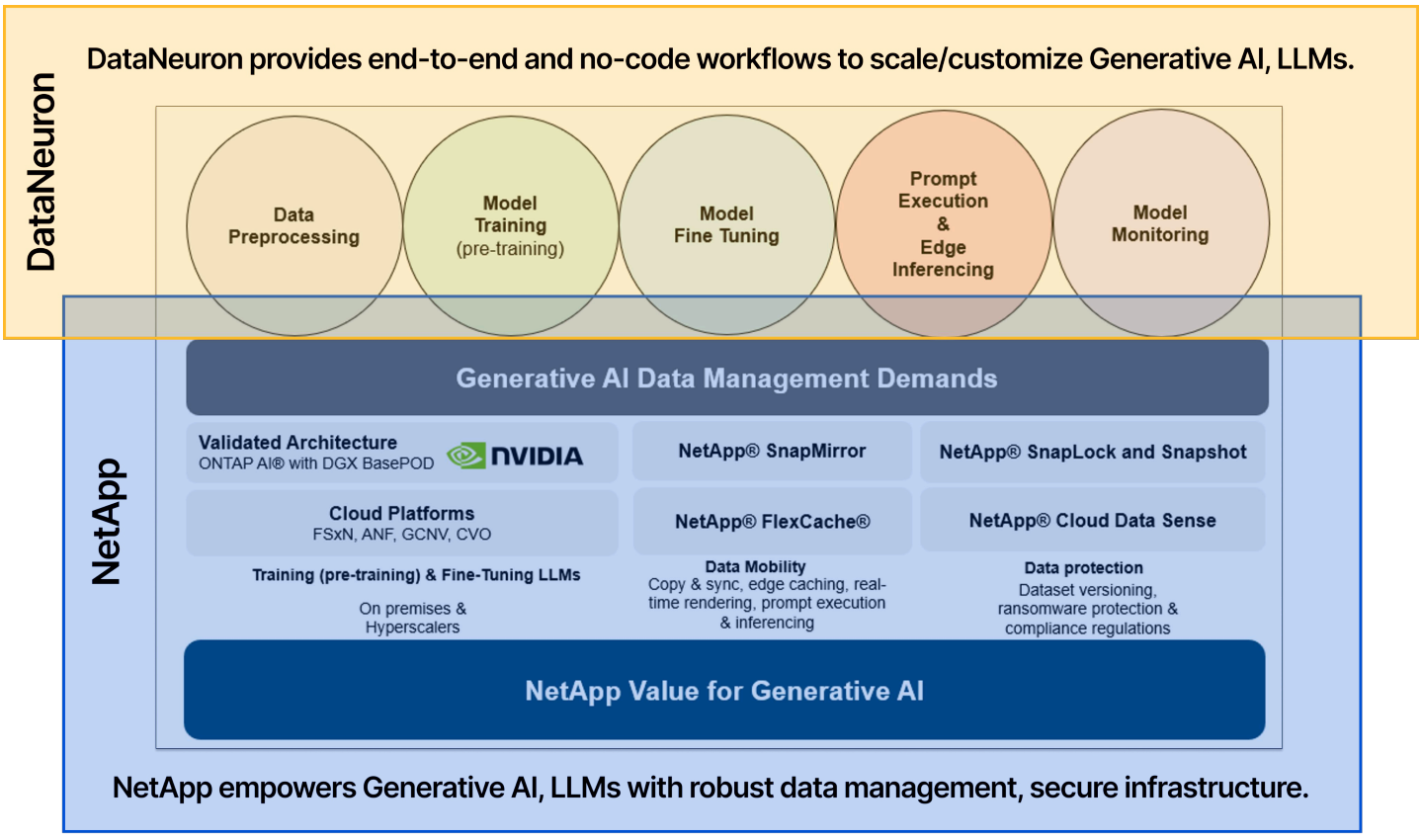
# Reference Architecture/ POC Whitepaper

## DataNeuron x NetApp

The rise of Large Language Models (LLMs) is reshaping industries, yet not every business has the resources or expertise to build their own foundational models. Fine-tuning and RAG (Retrieval-Augmented Generation) technologies overcomes this limitation and are increasingly popular among businesses looking to leverage LLMs. These technologies provide flexibility by allowing customers to refine existing LLM models with domain-specific data or augment pre-trained models with proprietary information, thereby enhancing accuracy and reliability. However, concerns surrounding data governance, compliance, and privacy present significant barriers for enterprises seeking to adopt these AI techniques.

**DataNeuron, a platform specializing in customized Large Language Model (LLM) solutions, has joined forces with NetApp, renowned for its Intelligent Data Infrastructure.**

This partnership aims to revolutionize the deployment and scalability of LLMs in enterprises, and address key concerns surrounding LLM integration, including data security, privacy, customization, and scalability.



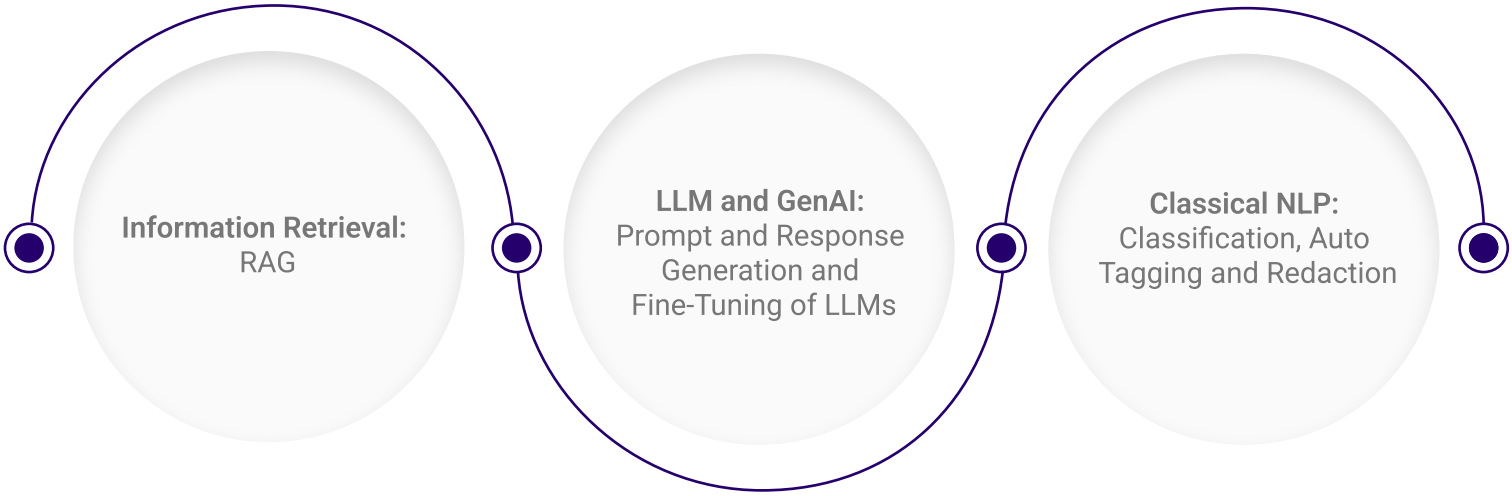

Source of the image: NetApp Generative AI Documentation

# Overview

- **Data Curation:**
  Fine-tuning LLMs requires precise data. This collaboration tackles the challenges associated with data curation, enabling organizations to optimize LLM performance effectively.
- **Model Lifespan and Selection:**
  Given the short lifespan of AI models, the ability to choose the most suitable model for a specific task or capability is crucial. The integration enhances this capability, enabling enterprises to select the best-fit model for their needs.
- **No-Code Pipelines:**
  Recognizing the scarcity of technical expertise in many organizations, the integration offers no-code pipelines, simplifying the deployment and management of LLMs without extensive technical knowledge.
- **Data Security and Privacy:**
  The integrated system ensures robust data security and privacy measures, mitigating the risk of data exposure to third-party systems. Enterprises are deeply concerned about data readiness and regulatory compliance in AI initiatives. The integrated system ensures that data management practices adhere to regulatory requirements, facilitating smooth AI deployments while minimizing compliance risks.
- **Intelligent Data Infrastructure:**
  Leveraging NetApp's Intelligent Data Infrastructure, the integrated system provides a solid foundation for LLM deployment, ensuring efficient data management and processing capabilities.
- **Robust and Responsible Data Management**
  Without robust and responsible data management practices, AI initiatives are prone to failure in delivering value. The collaboration between DataNeuron and NetApp prioritizes responsible data management, ensuring that data is handled ethically and efficiently throughout the AI lifecycle.

# POC

The Proof of Concept (POC) conducted showcases the seamless integration of DataNeuron's platform with NetApp's Intelligent Data Infrastructure, providing organizations with a robust solution for LLM implementation. We experimented on different workflows of the DataNeuron platform deployed on the NetApp Intelligent Data Infrastructure and NVIDIA GPUs :



## DataNeuron Workflows:

DataNeuron platform supports three no-code and automated workflows:

| | **LLM and GenAI:**<br>Prompt/Response Generation, Validation and Fine-Tuning. | **Classical NLP:**<br>Multi-label and Multi-class classification and NER. | **Information Retrieval:**<br>RAG and Playground/Q&A Interface |
|---|---|---|---|
| **Data Curation** | • Automated Prompt and Response Generation<br>• Prompt Annotation (Select, Ranking and Validation workflows) | • Automated Data Labeling for Classification (95% automation)<br>• Auto Tagging and Redaction | • Embeddings for better scaling, efficiency, and re-using same data for multiple use cases. |
| **Model Customization** | • Leading open-source LLMs available for customisation and fine-tuning.<br>• Deployment, Inferencing and Model Management | • Model Comparison<br>• Hyperparameter Selection<br>• Model Training<br>• Deployment, Inferencing & Model Management | • RAG<br>• Prompt and Response Generation<br>• LLM Playground and Q&A system |

## POC environment:

Our proof of concept (PoC) capitalizes on NetApp's robust data infrastructure, deployed on Google Cloud Platform (GCP), seamlessly operating in serverless cloud environments. To increase the performance of our language model (LLM) pipeline, we have deployed on NVIDIA Tensor Core A100 GPU.

To optimize resource utilization and streamline data access, we incorporated load balancers into our setup. These balancers intelligently distribute incoming traffic across kubernetes clusters, minimizing latency and maximizing compute efficiency.

For efficient data management and storage, we rely on NetApp ONTAP Storage Volumes (Extreme) via GCP. These fully managed file storage solutions provide reliability and scalability for our extensive datasets and knowledge base.

To seamlessly integrate components, we utilize the NFSv3 protocol to mount NetApp volumes onto our NVIDIA A100 GPU instances. This configuration ensures smooth data accessibility and operation throughout our pipeline, enhancing the overall efficiency of our PoC

### Minimum Compute Requirements/Operating System:

- **Operating System:** Linux Ubuntu (22.04 LTS)
- **GPU:** NVIDIA A100 80GB
- **CUDA Version:** 12.2

DataNeuron LLM Studio powered by NetApp

## Workflow

LLM and GenAI | Classical NLP | Information Retrieval

### Data Curation

- Prompt Generation
- Prompt Annotation
- Auto Tagging and Redaction
- Automated Labeling for Classification

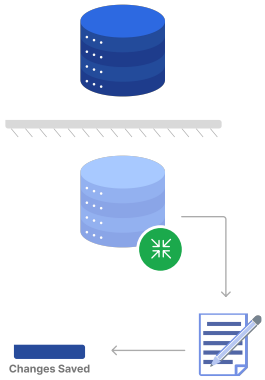
DataNeuron LLM Studio powered by NetApp

### Model Customization

- Fine-Tune / Train
- Deploy / Inference
- RAG
- Playground / Q&A Interface

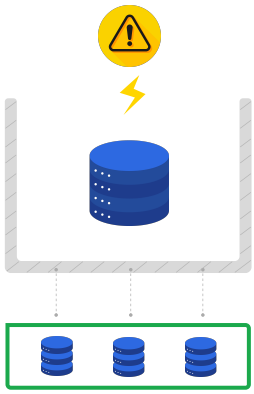
Data Engineering/DevOps using NetApp AI solutions

### FlexClone



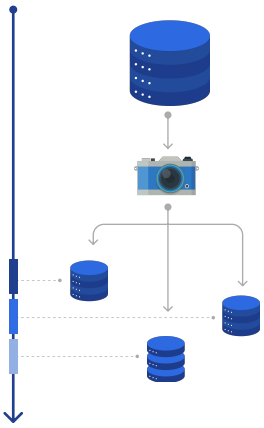

- Data/Model Versioning
- Iterative Learning

### Snapmirror



- Disaster Recovery
- Data Availability

### Snapshot



- Model Benchmarking
- Data Safeguarding

Cloud/Infra Layer

### NetApp Storage

ONTAP Storage

### NVIDIA GPUs

A100 / H100

On premises & Hyperscalers

# NetApp Data Engineering Solutions for DataNeuron platform:

- **Snapshot:**
  By employing snapshot technology, DataNeuron enhances workflow agility and efficiency by enabling users to easily revert to previous versions of the project's VectorDB volume for review or restoration. Enables Model benchmarking workflows. Within the DataNeuron platform, Snapshot integration spans both the frontend and backend, facilitating workflow versioning. This feature empowers users to leverage storage capacity more efficiently while benchmarking and experimenting with Generative AI and LLMs.

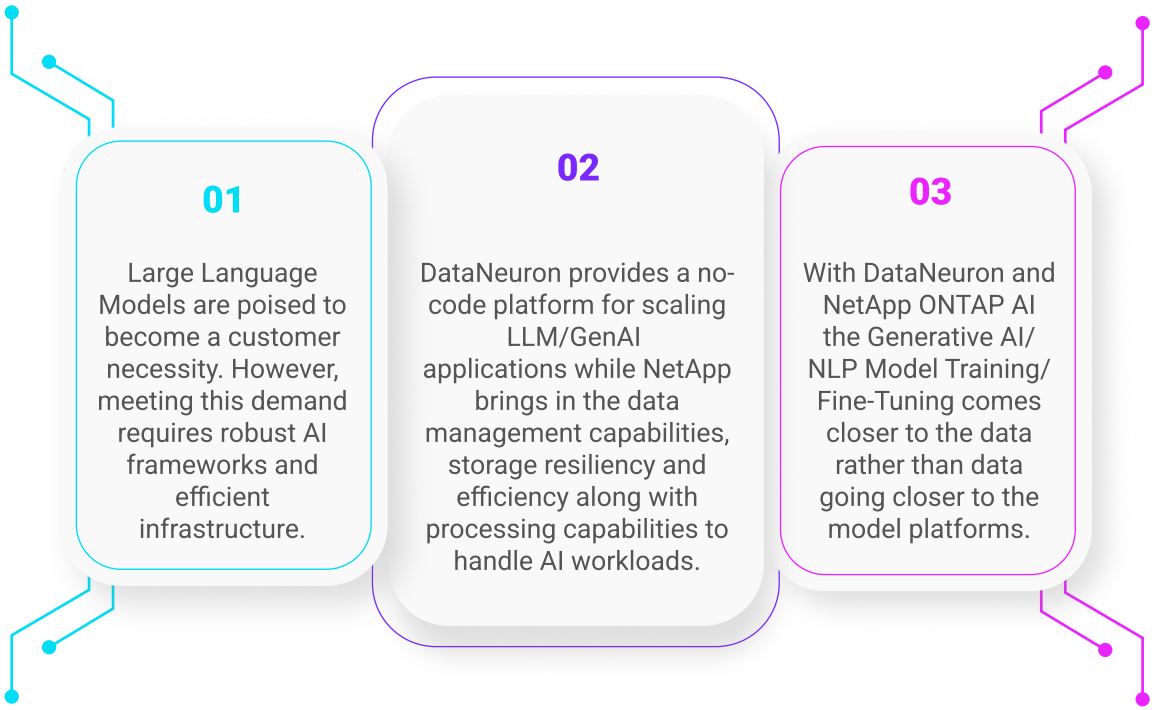
- **FlexClone:**
  Leveraging NetApp FlexClone, DataNeuron facilitates seamless experimentation by allowing users to create multiple workflows within each project, test data with different configurations, and generate clones for each workflow, enabling easy selection of the best RAG model.

- **SnapMirror:**
  DataNeuron can utilize SnapMirror to ensure data security and implement robust disaster recovery mechanisms, safeguarding all user-uploaded data and data generated by the RAG flow, including VectorDB, against unforeseen events. SnapMirror was not tested as part of the POC.

These features are available through the NetApp DataOps Toolkit, a python library that makes it easy for developers, data scientists, and data engineers to perform numerous data management tasks & streamline AI workflows. These features bring value to the deployment of real-time Generative AI models and help address data challenges from the edge to the data center to the cloud.
***We have used the GCP python library to enable these features in this POC.**

# Conclusion : DataNeuron + NetApp:

We successfully integrated and tested all the workflows of DataNeuron platform on the NetApp and Nvidia platform.



## About DataNeuron:

DataNeuron (DN) is a trailblazing venture-backed startup revolutionizing LLM and NLP workflows with cutting-edge SaaS solutions. With a distinguished team, comprising highly skilled data scientists, seasoned product experts, and visionary leaders, with recognition from Forbes under 30. Supported by a network of experienced board members, advisors, and investors, DN is poised to redefine the landscape of LLMs and NLP.

DN represents the next frontier in LLM and Generative AI solutions. With a focus on innovation, quality, and efficiency, DN is positioned to disrupt the market and set new standards for scaling LLMs.

For more information, please visit dataneuron.ai

## About NetApp:

NetApp is the intelligent data infrastructure company, combining unified data storage, integrated data services, and CloudOps solutions to turn a world of disruption into opportunity for every customer. NetApp creates silo-free infrastructure, harnessing observability and AI to enable the industry's best data management. As the only enterprise-grade storage service natively embedded in the world's biggest clouds, our data storage delivers seamless flexibility. In addition, our data services create a data advantage through superior cyber resilience, governance, and application agility. Our CloudOps solutions provide continuous optimization of performance and efficiency through observability and AI. No matter the data type, workload, or environment, with NetApp you can transform your data infrastructure to realize your business possibilities.

For more information on NetApp AI solutions, please visit netapp.com

References:

**Generative AI and NetApp Value:**
https://docs.netapp.com/us-en/netapp-solutions/ai/wp-genai.html#netapp-capabilities
**Private RAG:**
https://www.netapp.com/blog/private-rag-unlocking-generative-ai-for-enterprise/
**Ready to build your Generative AI:**
https://www.netapp.com/blog/ready-to-build-your-generative-ai/