

# Content Aware Storage

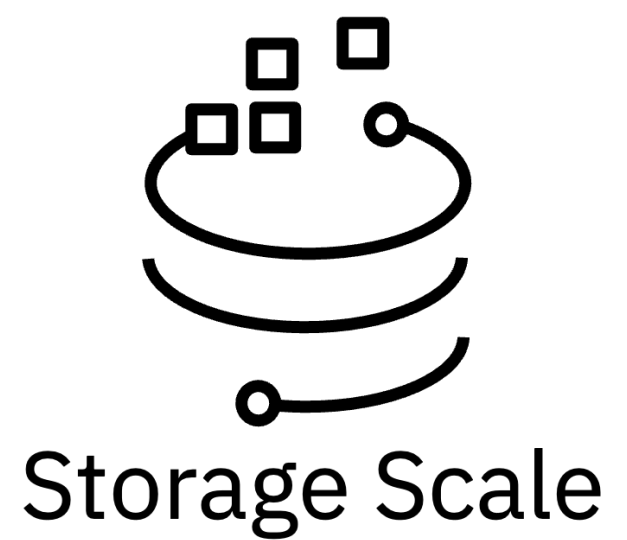
**IBM Storage Scale Days 2025 DE**

March 19th – 20th, 2025 | Heidelberg, Germany

Ulf Troppens

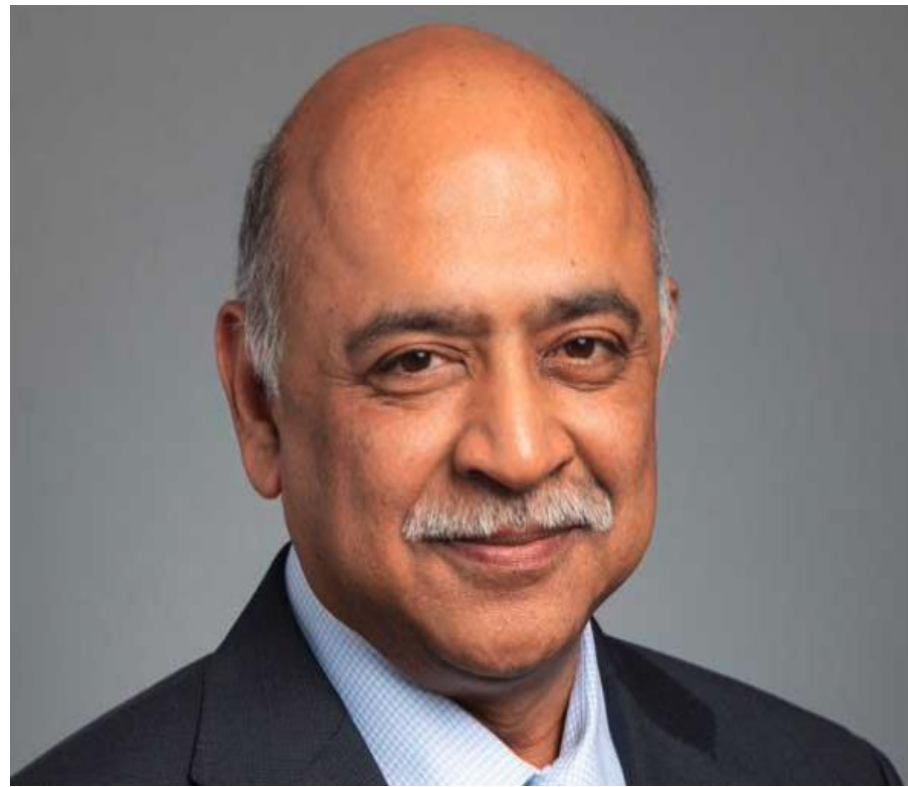


# Disclaimer



- IBM's statements regarding its plans, directions, and intent are subject to change or withdrawal without notice at IBM's sole discretion. Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision. The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code, or functionality. The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.
- IBM reserves the right to change product specifications and offerings at any time without notice. This publication could include technical inaccuracies or typographical errors. References herein to IBM products and services do not imply that IBM intends to make them available in all countries.

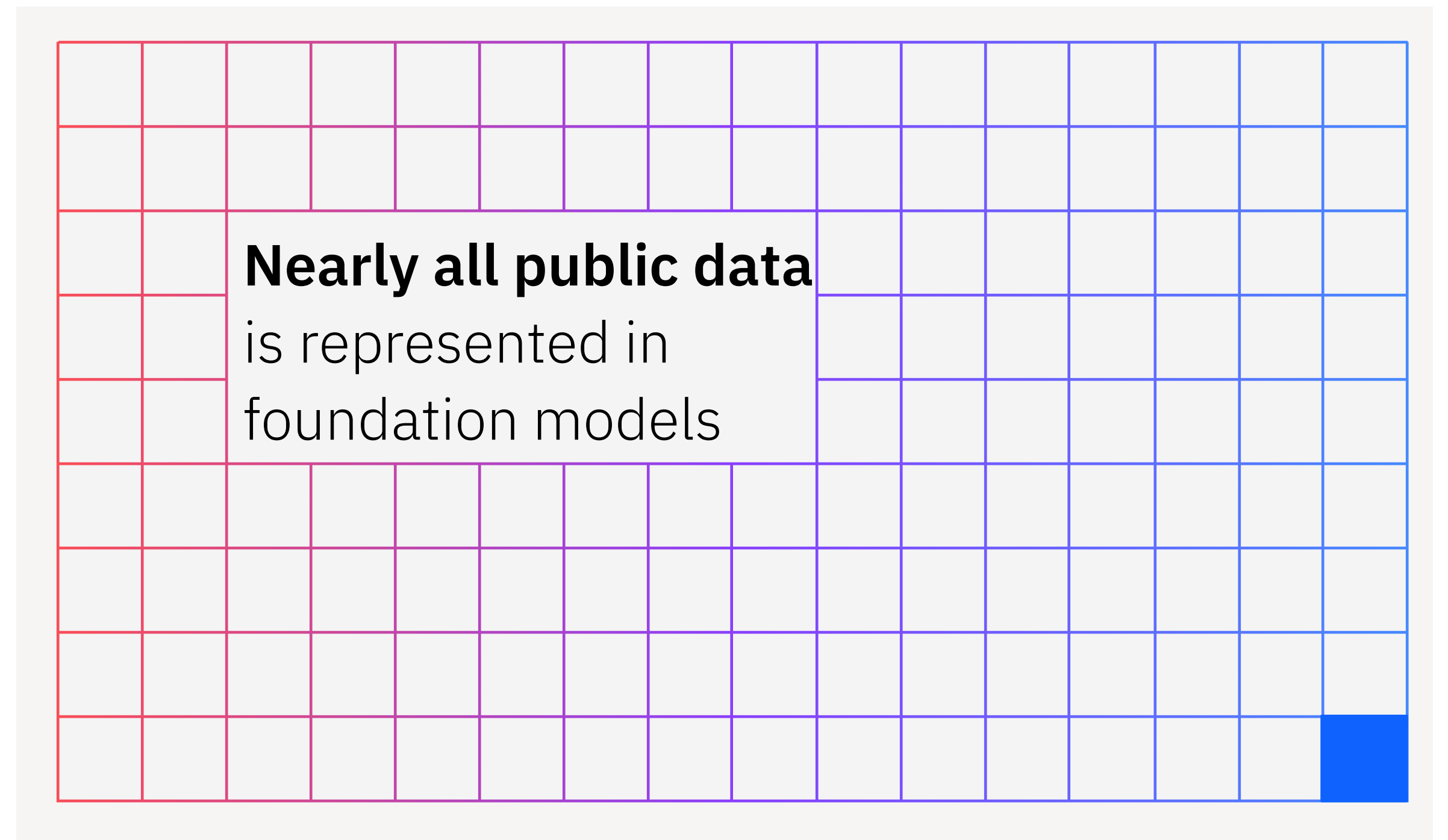
# Organizations Need to Unlock Value From Data... *Wherever it Resides*



“We fundamentally believe that core to the competitiveness of every company going forward will be their ability to use AI to **unlock real-time value from their data wherever the data resides.**”

Arvind Krishna  
IBM Chairman and CEO

Data is the fuel for an effective AI strategy  
Only small fraction of enterprise data is used in Gen AI



**But only 1% of  
enterprise data!**

A question  
answered with  
public data ...

You 16:30

How can S3 workloads be accelerated?



watsonx 16:30

1. Utilize Amazon S3 Transfer Acceleration: This feature uses optimized network paths to improve data transfer speeds between your client and an S3 bucket. It's particularly useful for applications that frequently upload or download large amounts of data from S3.
2. Employ Amazon S3 Intelligent-Tiering: This storage class automatically optimizes costs by moving objects between access tiers based on their access patterns. By using this tier, you can potentially reduce latency and increase performance without additional effort.
3. Leverage Amazon S3 Select: This is an optional feature that allows you to retrieve only specific subsets of data from an object instead of downloading entire objects. This reduces the amount of data transferred over the network, thereby improving performance.
4. Use AWS CloudFront with S3: Combining these services enables fast delivery of static website content, streaming media, and APIs directly to end users via edge locations around the world.



Type something



This productivity workflow leverages RAG to answer questions. Toggle the chevron icon to view the source text used from the document(s).

AskIBM leverages w3 and Publisher content, coupled with IBM's Granite™ family of AI models, to put generative AI in the hands of IBMers. As a user contributing data, you must adhere to this notice and abide by the [policies](#) and any contractual obligations applying to customer data. The files you upload to AskIBM will be deleted within 3 hrs

AskIBM output is created using generative AI, which may produce inaccuracies and is not intended to give advice. As such, please use discretion before using and relying on the output.

[Learn more](#)

Owner  
Ulf Troppens

Model  
granite-3-8b-instruct

Date created  
March 13, 2025

Feedback

The same question  
answered with  
enterprise data ...





# Chat with documents



watsonx 16:34

Welcome to the Chat with documents productivity workflow. To get started, upload the document(s) you'd like to chat with using the 'Upload' button in the input bar.

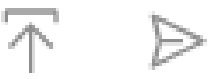
You 16:34

DOC under embargo.docx ✓



watsonx 16:34

☰ Type something



## Chat with documents



This productivity workflow leverages RAG to answer questions. Toggle the chevron icon to view the source text used from the document(s).

AskIBM leverages w3 and Publisher content, coupled with IBM's Granite™ family of AI models, to put generative AI in the hands of IBMers. As a user contributing data, you must adhere to this notice and abide by the [policies](#) and any contractual obligations applying to customer data. The files you upload to AskIBM will be deleted within 3 hrs

AskIBM output is created using generative AI, which may produce inaccuracies and is not intended to give advice. As such, please use discretion before using and relying on the output.

[Learn more](#)

Owner  
Ulf Troppens

Model  
granite-3-8b-instruct

Date created  
March 13, 2025

Feedback



# Chat with documents

## Prompt menu

5

What is S3 over RDMA?



Who developed S3 over RDMA?



What are the benefits of using S3 over RDMA for AI/ML applications?



What hardware and software requirements are needed for S3 over RDMA?



How does S3 over RDMA improve performance compared to traditional S3 over TCP?



≡ Type something



## Chat with documents



This productivity workflow leverages RAG to answer questions. Toggle the chevron icon to view the source text used from the document(s).

AskIBM leverages w3 and Publisher content, coupled with IBM's Granite™ family of AI models, to put generative AI in the hands of IBMers. As a user contributing data, you must adhere to this notice and abide by the [policies](#) and any contractual obligations applying to customer data. The files you upload to AskIBM will be deleted within 3 hrs

AskIBM output is created using generative AI, which may produce inaccuracies and is not intended to give advice. As such, please use discretion before using and relying on the output.

[Learn more](#)

Owner  
Ulf Troppens

Model  
granite-3-8b-instruct

Date created  
March 13, 2025

Feedback



You 16:40

How can S3 workloads be accelerated?



watsonx 16:40

S3 workloads can be accelerated by implementing Remote Direct Memory Access (RDMA) over InfiniBand or RoCE networks. This technology enables direct communication between the client's memory and the object storage system, bypassing the OS, TCP stack, and CPU overhead. It supports S3 operations like GetObject, PutObject, and PutPart, improving throughput and reducing latency significantly. To utilize this, you'll need high-performance RDMA networks, compatible compute nodes with optional GPU devices, and NVIDIA CUDA toolkit. Integration with various clients such as Python applications with boto3, PyTorch AI/ML applications with s3-connector-for-pytorch, and RAPIDS data science libraries using kvikio is planned. Future developments aim to optimize for IBM Storage Scale and conduct experiments with real-world storage systems and GPU-based AI/ML benchmarks.



## Chat with documents



This productivity workflow leverages RAG to answer questions. Toggle the chevron icon to view the source text used from the document(s).

AskIBM leverages w3 and Publisher content, coupled with IBM's Granite™ family of AI models, to put generative AI in the hands of IBMers. As a user contributing data, you must adhere to this notice and abide by the [policies](#) and any contractual obligations applying to customer data. The files you upload to AskIBM will be deleted within 3 hrs

AskIBM output is created using generative AI, which may produce inaccuracies and is not intended to give advice. As such, please use discretion before using and relying on the output.

[Learn more](#)

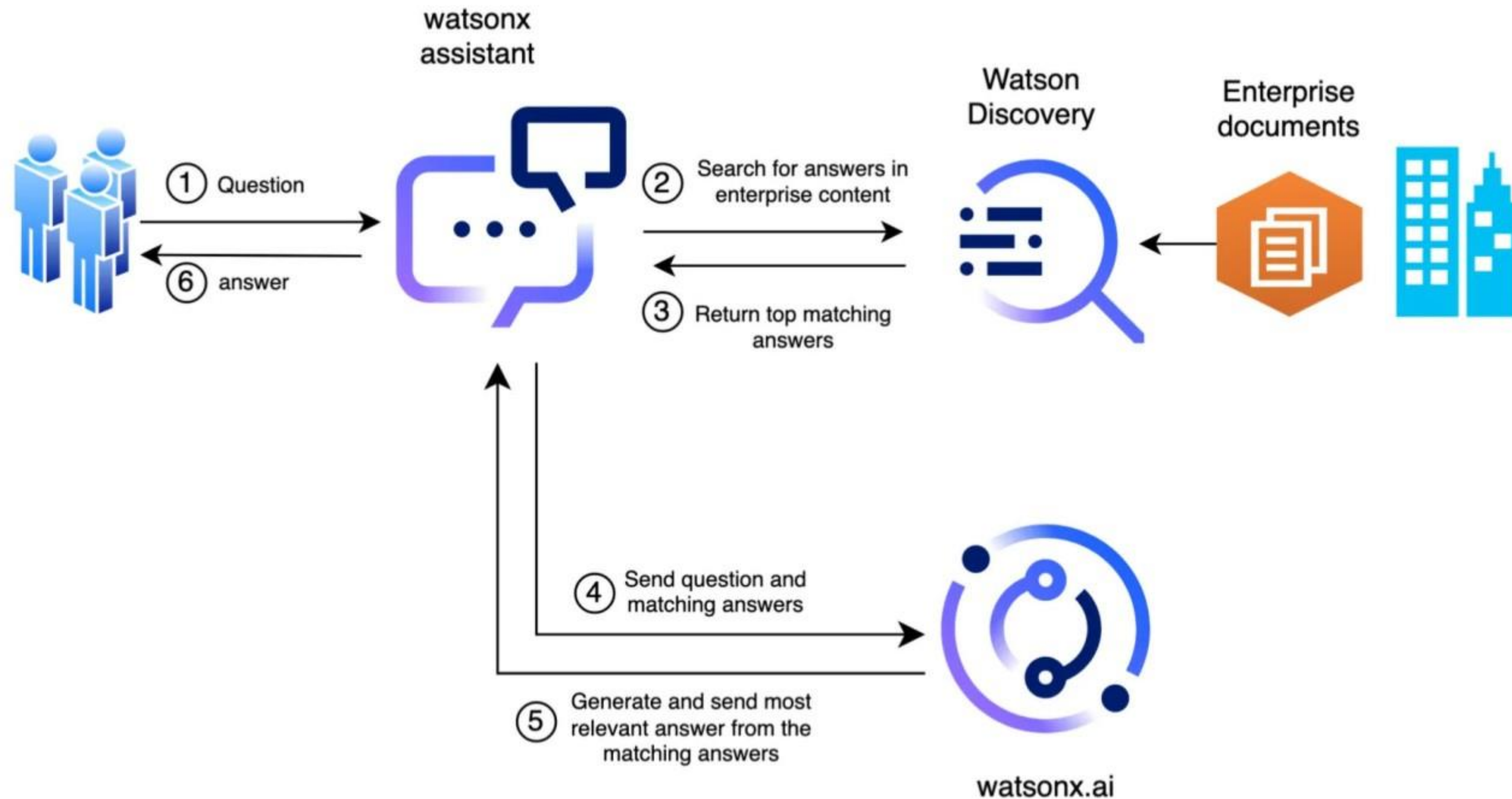
Owner  
Ulf Troppens

Model  
granite-3-8b-instruct

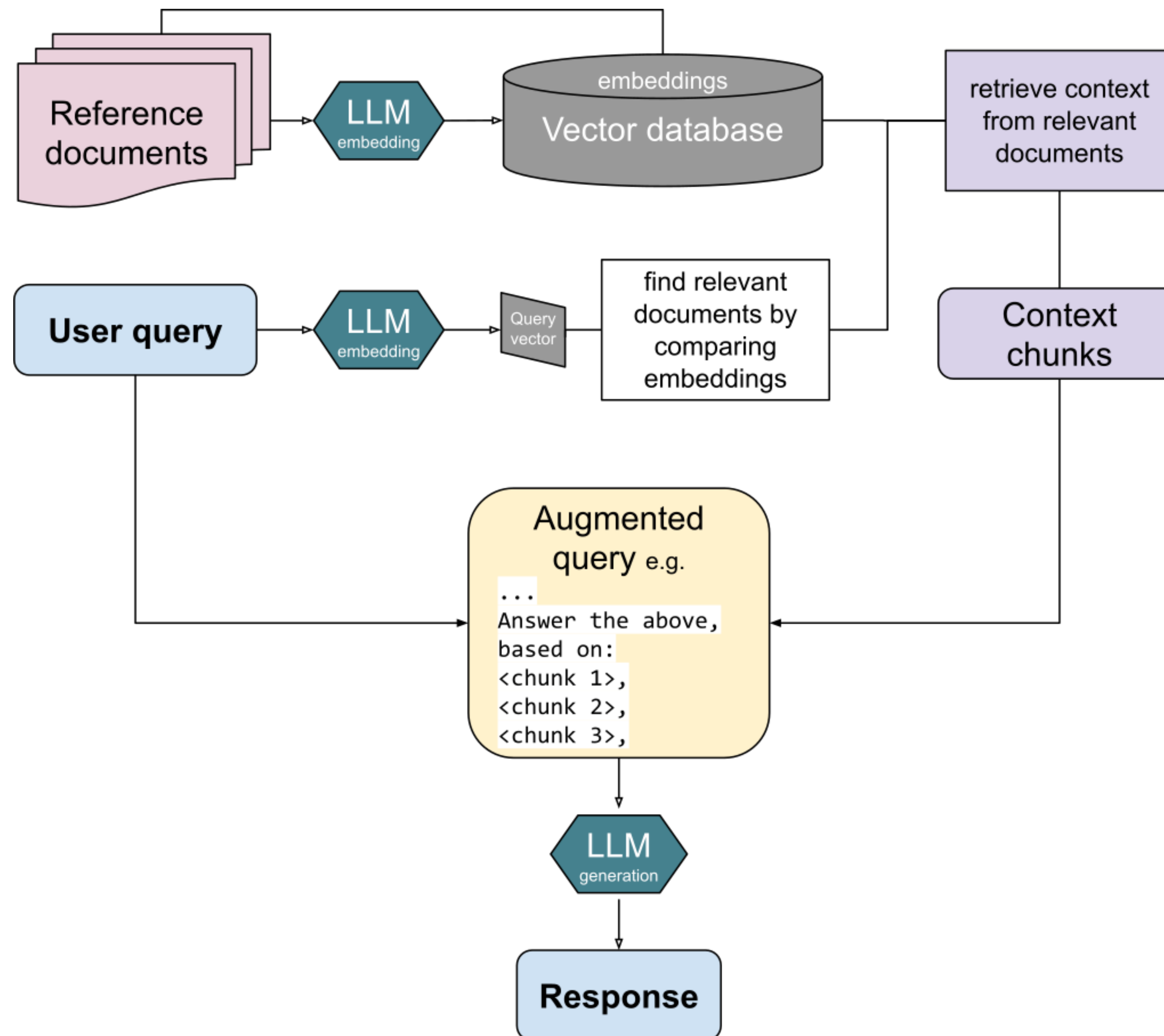
Date created  
March 13, 2025

Feedback

# Retrieval Augmented Generation (RAG) with IBM watsonx



# Retrieval Augmented Generation (RAG) under the hood



Content Aware Storage (CAS)  
is RAG integrated in storage.



# S71937: Enabling Intelligent Storage to Process Data for AI Applications

IBM Content Aware Storage  
*Accelerate enterprise Gen AI with NVIDIA NIMs + IBM Storage*

—  
Vincent Hsu  
IBM Fellow, VP & CTO of IBM Storage

Rob Davis  
VP, NVIDIA VP of Storage technology





# Years of Collaboration between NVIDIA and IBM

*2018: US DOE Summit & Sierra - Summit: 27,648 Tesla GPUs*

*2019: NVIDIA DGX-2H SuperPOD - 96 NVIDIA DGX2 nodes*

*2020: Red Hat OpenShift on DGX*

*2021: Magnum IO, GPUDirect Storage (GDS)*

*2021: DGX A100 SuperPOD RA*

*2022: 1st DGX SuperPOD BCM installation*

*2023: DGX SuperPOD H100 RA*

*2024: IBM Storage Scale 6000 Certification – 320 GBs  
for H200 and B200 DGX SuperPOD and NCP*

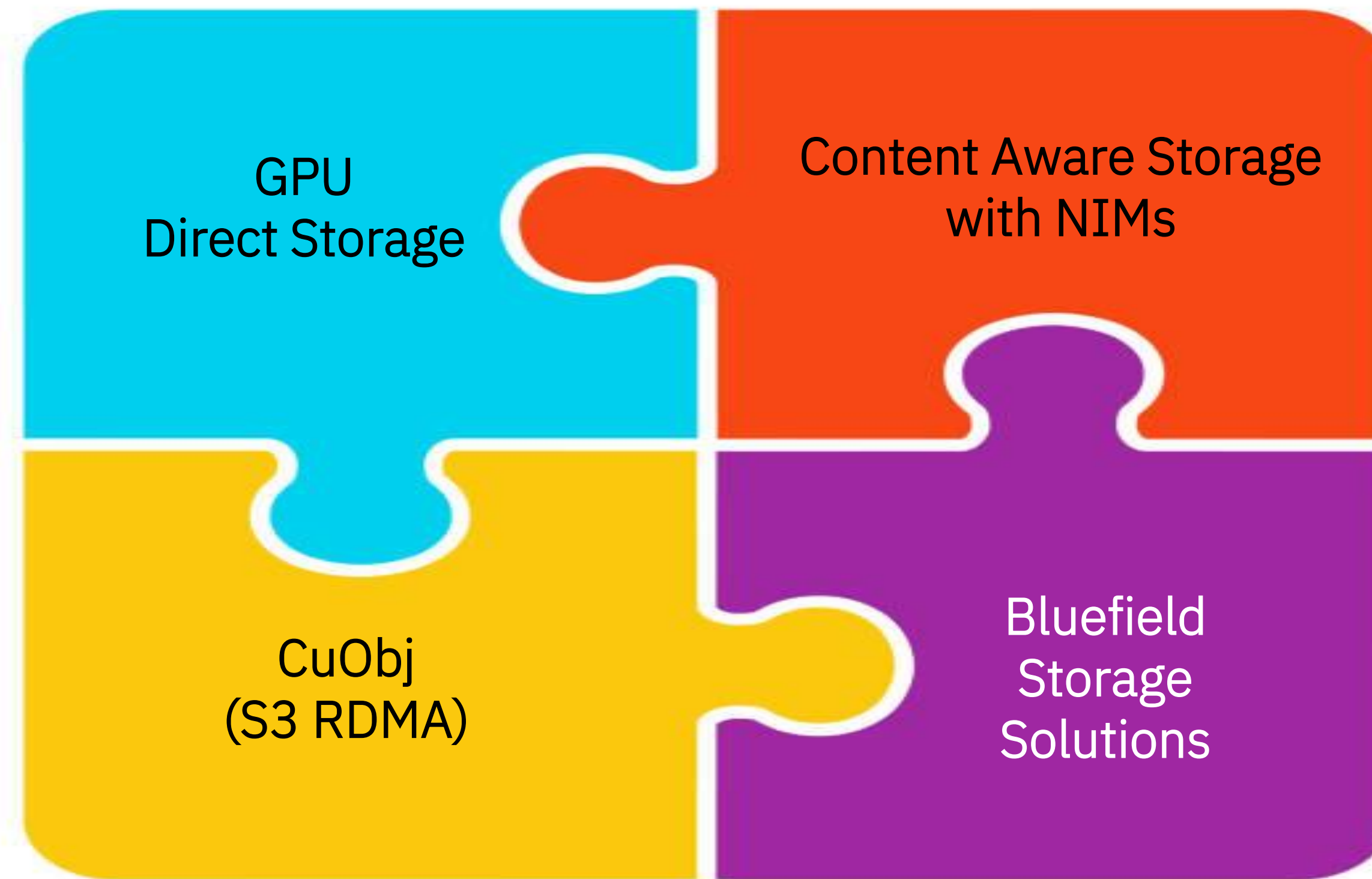
*2024: Jülich – Exascale Storage NVIDIA GH200*

*2024: IBM Model Factory – BlueVela with 6000 GPUs*

*2025: IBM Model Factory with CoreWeave and GB200NVL72  
including the IBM Storage Scale 6000*

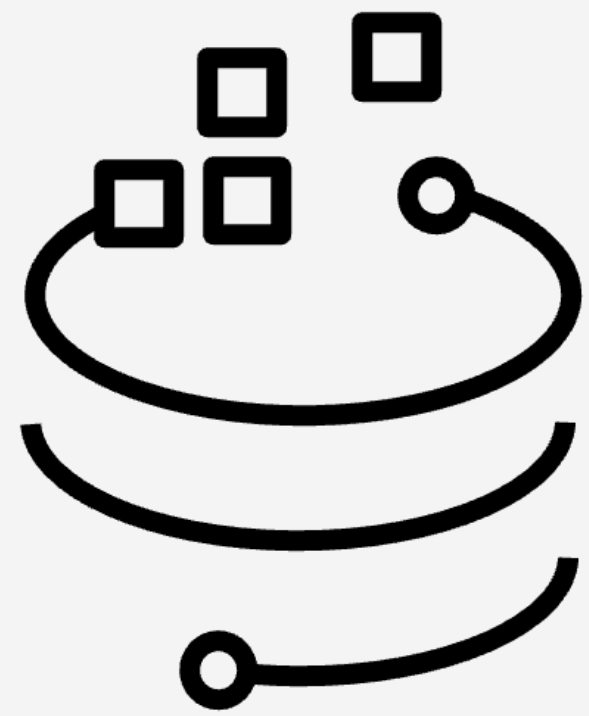
*2025: IBM Content Aware Storage (CAS) with NVIDIA NIMs*

# Strategic technical collaboration between IBM storage and NVIDIA

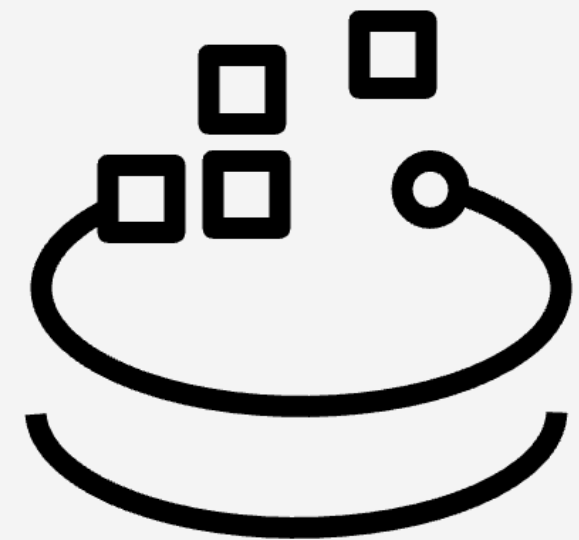




# Thank you for using



Storage Scale



Storage Scale  
System