



# ADVANCED TECHNOLOGY GROUP (ATG)

---



## Accelerate with ATG Webinar: IBM Fusion and Content Aware Storage (CAS)

**Shu Mookerjee**

ATG Senior Storage Technical Specialist

[Shu.Mookerjee@ibm.com](mailto:Shu.Mookerjee@ibm.com)



## Accelerate with ATG Technical Webinar Series

---

*Advanced Technology Group* experts cover a variety of technical topics.

**Audience:** Clients who have or are considering acquiring IBM Storage solutions. Business Partners and IBMers are also welcome.

To automatically receive announcements of upcoming Accelerate with ATG webinars - Clients, Business Partners and IBMers are welcome to send an email request to [accelerate-join@hursley.ibm.com](mailto:accelerate-join@hursley.ibm.com).

### 2025 Upcoming Webinars – Register Here!

[Forging Ahead - IBM Storage Virtualize 9.1.0 Technical Update](#) - August 12th, 2025



### *Important Links to Bookmark:*

**Accelerate with ATG** - Click here to access the Accelerate with ATG webinar schedule for 2025, view presentation materials, and watch past replays dating back two years. <https://ibm.biz/BdSUFN>

**ATG MediaCenter Channel** - This channel offers a wealth of additional videos covering a wide range of storage topics, including IBM Flash, DS8, Tape, Ceph, Fusion, Cyber Resiliency, Cloud Object Storage, and more. <https://ibm.biz/BdfEgQ>

## Offerings

---

### Client Technical Workshops

- **IBM Fusion & Ceph: August 6-7 (Coppell, TX)**
- **IBM Storage Scale & Storage Scale Functions: August 20-21 (San Jose, CA)**
- **IBM DS8000 G10 Advanced Functions: August 26-27 (Chicago, IL)**
- **IBM FlashSystem Deep Dive & Advanced Functions: September 10-11 (Durham, NC)**
- **Cyber Resilience with IBM Storage Defender**

### TechZone Test Drive / Demo's

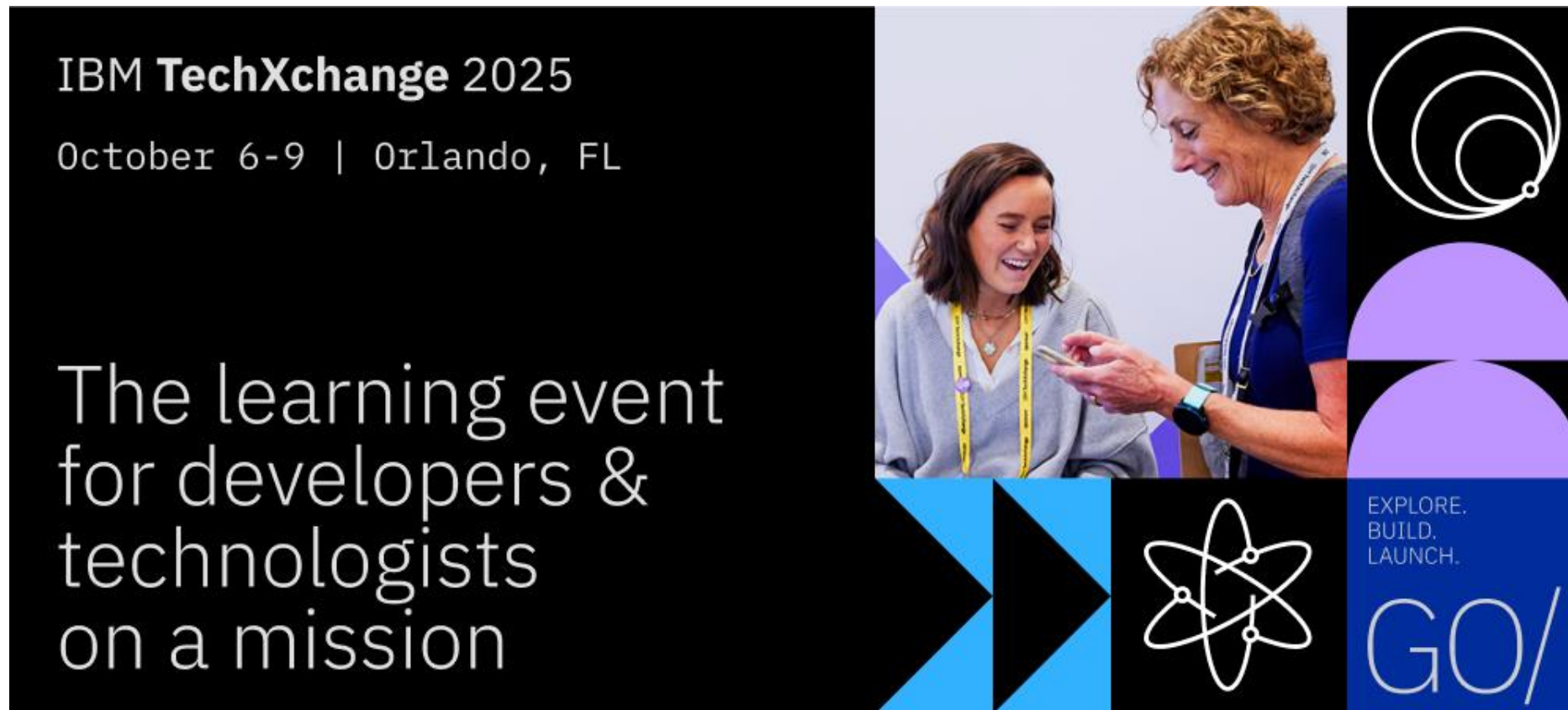
- IBM Cloud Object Storage Test Drive - (VMware based)
- IBM DS8900F Safeguarded Copy (SGC) Test Drive
- IBM DS8900F Storage Management Test Drive
- IBM Storage Scale and Storage Scale System GUI
- IBM Storage Virtualize Test Drive
- IBM Storage Ceph Test Drive
- IBM Storage Ceph Test Drive - (VMware based)
- IBM Storage Protect Live Test Drive
- Managing Copy Services on the DS8000 Using IBM Copy Services Manager Test Drive

Please reach out to your IBM Representative or Business Partner for more information.

**\*IMPORTANT\* The ATG team serves clients and Business Partners in the Americas, concentrating on North America.**

## Announcing the 2025 IBM TechXchange Conference

Our theme this year is simple but powerful: **GO / Explore. Build. Launch.**



IBM TechXchange 2025  
October 6-9 | Orlando, FL

The learning event  
for developers &  
technologists  
on a mission

EXPLORE.  
BUILD.  
LAUNCH.

GO/

The graphic is a collage. On the left, a black rectangle contains the event title and dates in white text, and below it, the tagline 'The learning event for developers & technologists on a mission' in a larger white font. To the right of this is a photograph of two women, one showing a smartphone to the other. Further right is a vertical strip with a black background containing a white spiral logo, two purple semi-circles, and a blue section with the text 'EXPLORE. BUILD. LAUNCH.' and 'GO/'. At the bottom center, there is a blue and black geometric arrow pointing right, and to its right, a white atomic symbol logo on a black background.

For more information, please visit - <https://www.ibm.com/community/ibm-techxchange-conference/>

## Accelerate with ATG Survey

---

Please take a moment to share your feedback with our team!

You can access this 6-question survey via [Menti.com](#) with code 51510447 or

Direct link <https://www.menti.com/alhsf3bgvxu6>

Or

QR Code







# ADVANCED TECHNOLOGY GROUP (ATG)

---



## Accelerate with ATG Webinar: IBM Fusion and Content Aware Storage (CAS)

**Shu Mookerjee**

ATG Senior Storage Technical Specialist

[Shu.Mookerjee@ibm.com](mailto:Shu.Mookerjee@ibm.com)



## Meet the Team

---

### Speaker



**Shu Mookerjee** is a Level 2 Certified Technical Specialist with nearly twenty-five years at IBM, working in a variety of roles including sales, management and technology. For the last twelve years, he has focused exclusively on storage and has authored three (3) Redbooks. Currently, Shu is part of the Advanced Technology Group where he provides education, technical guidance, Proofs of Concept and Proofs of Technology to IBMers, business partners and clients.

### Panelists



**Andrew Rice** is an Infrastructure/Storage Engineer with over 17 years of experience implementing cloud infrastructure design, storage solutions and virtualization. Andrew's expertise extends across IBM's storage portfolio primarily in IBM Storage Scale, Fusion, Storage Protect, IBM FlashSystems and encompasses technical proficiencies in VMWare and Red Hat OpenShift

## Meet the Team

---

### Host



**John Shubeck** is an information technology professional with over 42 years of industry experience spanning both the customer and technology provider experience. John is currently serving as a Senior Storage Technical Specialist on IBM Object Storage platforms across all market segments in the Americas.

### Special Thanks!



Chris Maestas



Joe Dain



## Agenda

---

- Goals and Objectives
- Introduction to Content Aware Storage
- Components
- CAS Architecture
  - Data Source Management
  - RAG Pipeline
  - NeMo Retrievers
  - Database
  - Querying

## Goals and Objectives

---

### **Objective:**

Provide a mid-level technical overview of Content Aware Storage

### **We WILL:**

- Cover the main objective and design point of the solution
- Review the architecture and components
- Walk through the pieces and processes

### **We WILL Not:**

- Cover installation and configuration in depth
- Review licensing and pricing
- Try to make you computer scientists!

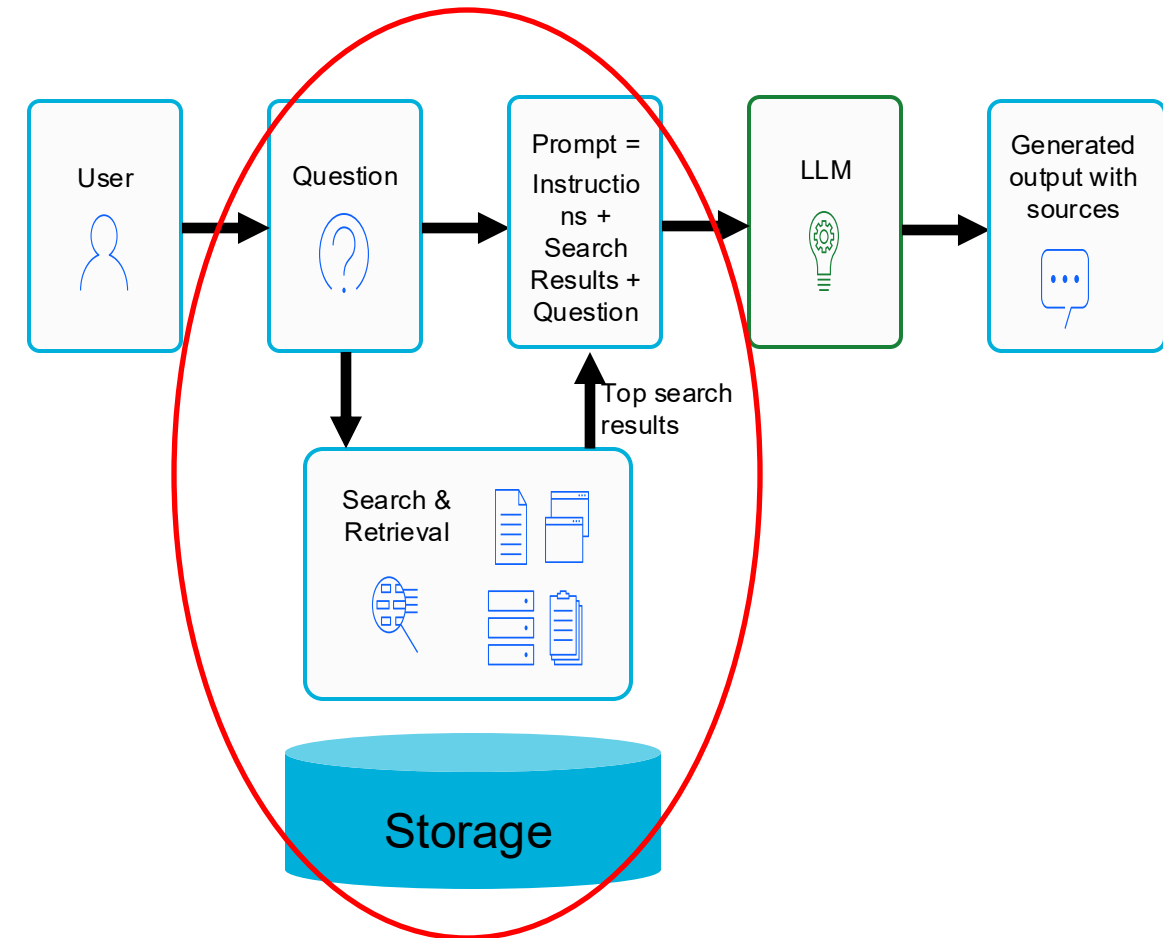
## Introduction to CAS – What is Content Aware Storage?

- Turn-key solution that provides the ability to process unstructured data for use in multi-modal AI/ML applications
- Leverages a variety of microservices including NVIDIA inferencing Models (NIMs) and NVIDIA NeMo Retriever Extraction applications
- Components include:
  - NVIDIA L40S or H100 GPUs for process acceleration
  - IBM Storage Scale as a file/object cache for enterprise data
  - IBM Fusion as the main CAS runtime platform
  - IBM Fusion HCI as the turnkey infrastructure
- Design to optimize the Retrieval Augmented Generation (RAG) process



## Introduction to CAS – What is Content Aware Storage? - RAG

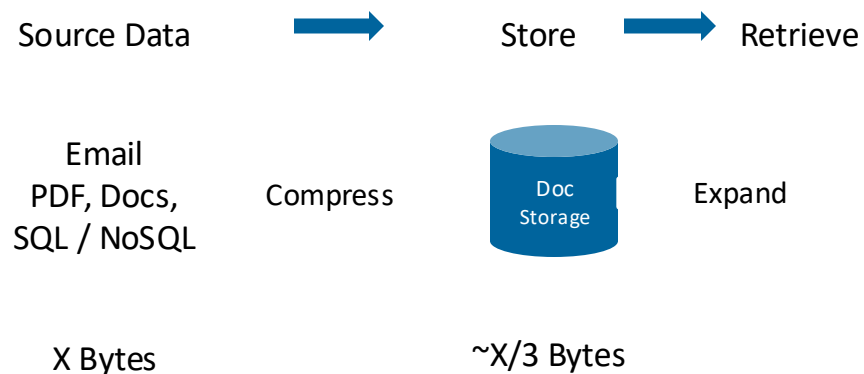
- Large Language Models (LLM) leverage an AI framework called Retrieval Augmented Generation (RAG) to analyze data context and return meaningful content
- The RAG Process consists of three steps:
  1. Search for relevant content in the knowledge base
  2. Pull the most relevant content into the model as context
  3. Send the combined prompt text to the model to generate output
- However, bringing the data to the RAG Framework can be:
  - Complex: too many technologies to integrate
  - Costly: too many copies of data
  - Security Risk: too many replicas of data with inconsistent Access Control Lists between raw data and embedding
  - Stale: long latency between data change and updated vectors
  - Limited scalability



## Introduction to CAS – What is Content Aware Storage? - RAG

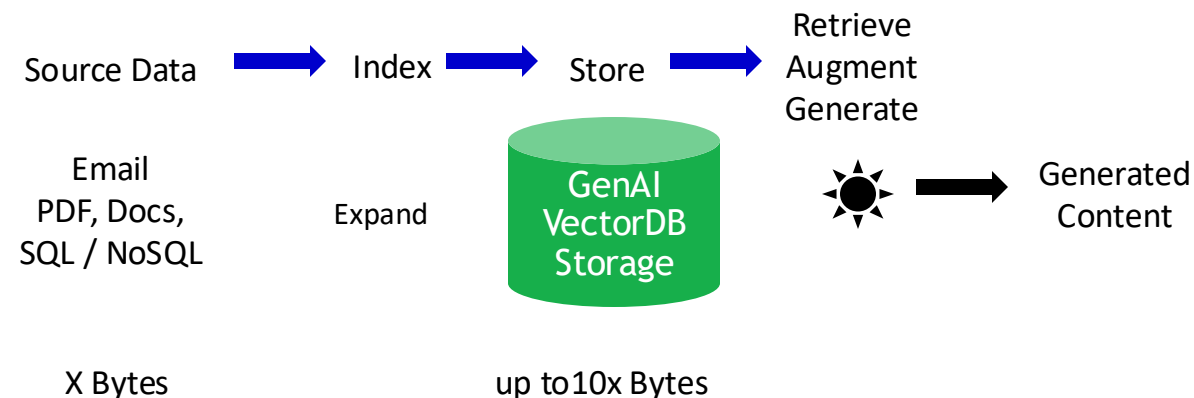
Up to 10X Data Growth Motivates Repatriation from high-cost Cloud to On-Prem Storage

### Traditional Enterprise Application Data Pipeline



- Traditional enterprise apps are retrieval based
- Data is compressed
- Retrieved and uncompressed as needed
- Enterprise content is static, largely text, and expanded on retrieve
- Access control is simple: Open(file) -> {allow,deny}

### Generative AI Enterprise Application Data Pipeline

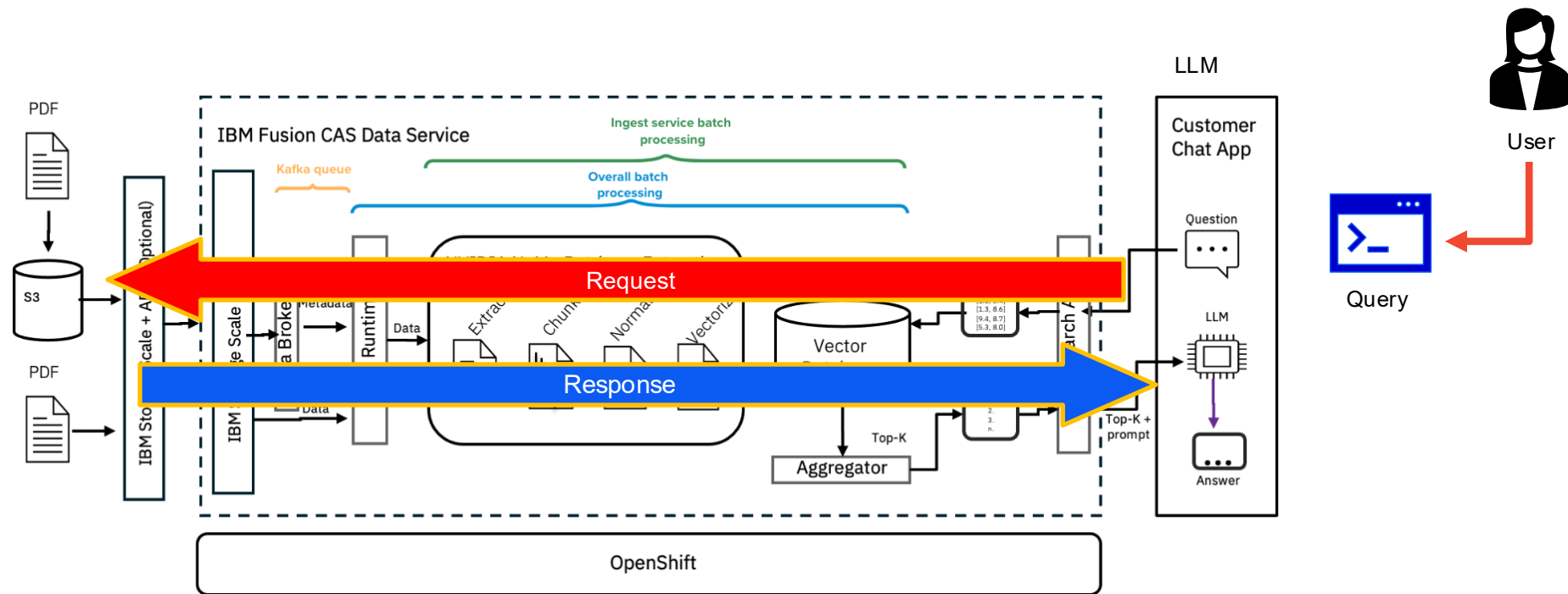


- GenAI & RAG fundamentally change workflow
- Data expands (up to 10X) when embedded and not compressible!
- Model & VectorDB memory requirement may limit RAG GenAI inference efficiency
- Access control is pan-document



## Introduction to CAS – What is Content Aware Storage?

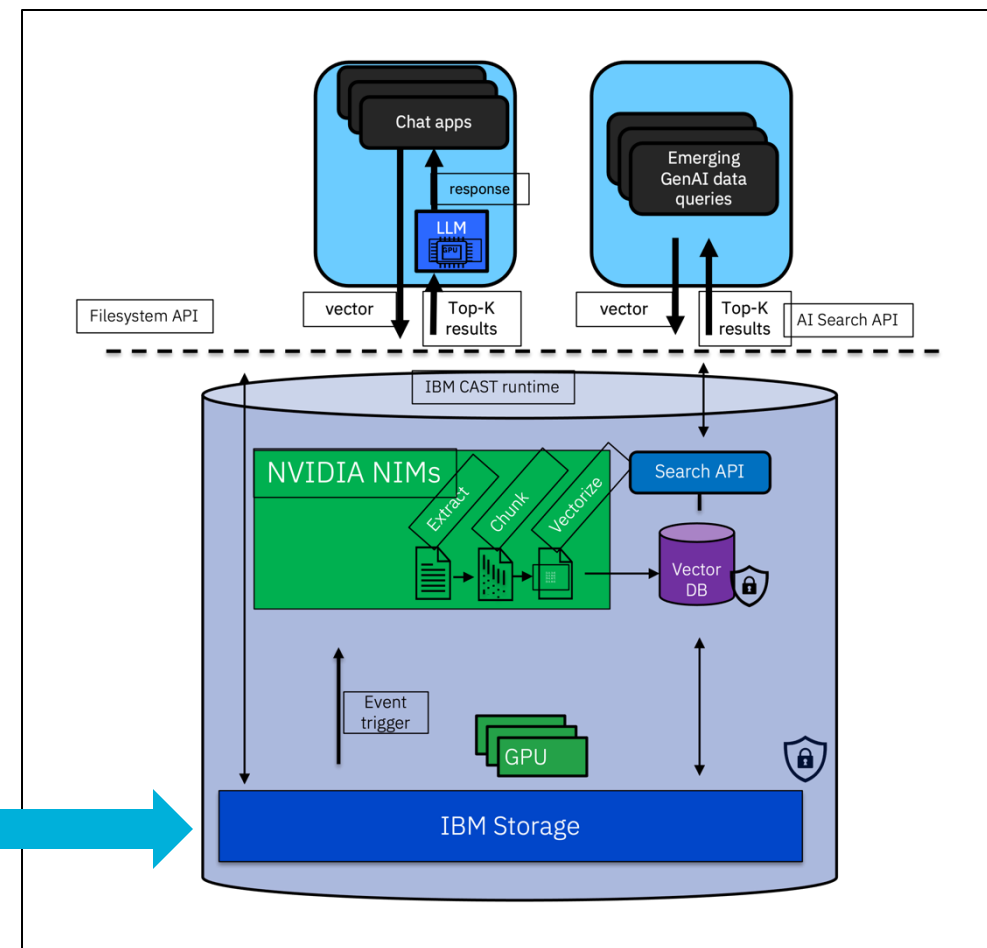
The Content-Aware Storage (CAS) architecture is designed to enhance the RAG Process by facilitating a smooth interaction between Large Language Models and extensive quantities of unstructured data while amplifying insight, derivation and suggestion functionalities.



## Components – Solution Overview

CAS is a turn-key solution that leverages:

- AI optimized storage – IBM Storage Scale
- Includes a vector database that scales to 1B vectors and beyond while preserving data source ACL permissions
- Supports RAG vectorization for enterprise data, with incremental data processing.
- NVIDIA Components:
  - NIM (NVIDIA Inferencing Microservices)
  - L40S GPU for hardware acceleration
  - NeMo Retriever with support for text extraction from charts and images.



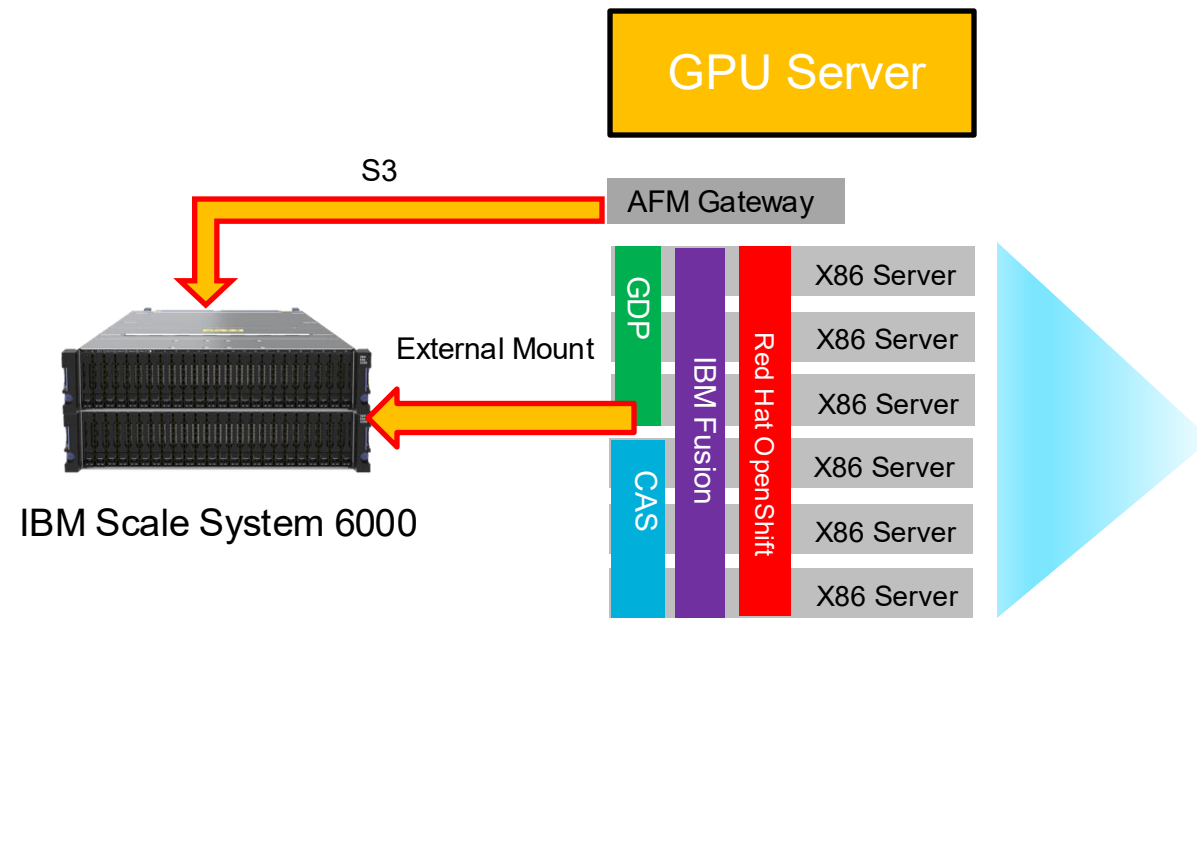
*Content-aware storage leverages AI storage and data processing pipelines*

## Components - Hardware Architecture

CAS can be run both on Fusion HCI or as part of a Fusion SDS Deployment

HCI Example:

- Six-node server cluster
- Deploy OpenShift
- Install Fusion Operator
- Install GDP (Scale CNSA Service)
- Integrate GPU node
- External mount to Scale (Scale System 6000 shown)
  - Note: If using S3 data, an AFM gateway for AFM S3 is required



# Components – Hardware Architecture – SDS Resource Requirements

Starter configuration <12TB CAS Managed Data:

- 1 GPU worker node
- 2 non-GPU worker nodes
- 160 vCPU (SMT=2)
- Min 768 GB RAM

>12TB CAS Managed Data with High Availability:

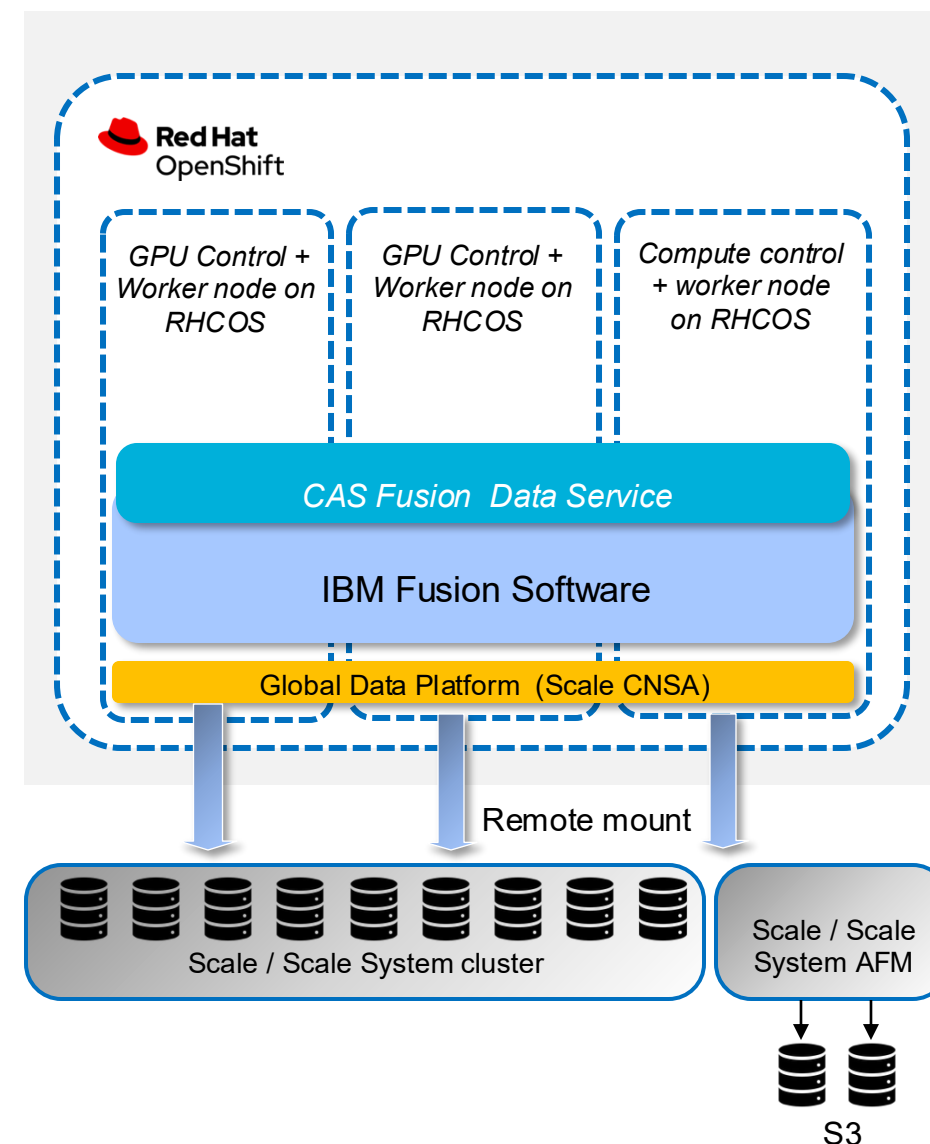
- 2 GPU worker nodes
- 1 non-GPU worker node
- 320 vGPU
- 2560GB RAM

Production with Multi-instance GPU (MIG) enabled	Production without Multi-instance GPU (MIG) enabled	Non-production with NVIDIA time slicing enabled
2 MIG Capable NVIDIA GPUs (A100, H100, RTX PRO 6000)	6 NVIDIA GPUs (L40S A100, H100,H200, RTX PRO 6000)	2 NVIDIA L40S, A10G GPUs

Production with Multi-instance GPU (MIG) enabled	Production without Multi-instance GPU (MIG) enabled	Non-production with NVIDIA time slicing enabled
4 MIG Capable NVIDIA GPUs (A100, H100, RTX PRO 6000)	12 NVIDIA GPUs (L40S A100, H100,H200, RTX PRO 6000)	4 NVIDIA L40S, A10G GPUs

## Components – Software Stack

<b>Software Packaging</b>	<ul style="list-style-type: none"> <li>CAS provided as Fusion Data Service offered as part of IBM Fusion SDS 2.10 and Fusion HCI 2.10</li> </ul>
<b>OpenShift Support</b>	<ul style="list-style-type: none"> <li>Requires OpenShift 4.17 w/CoreOS for worker nodes</li> <li>Customer provided OpenShift for Fusion SDS</li> <li>Combined control and worker node architecture (shown in figure)</li> <li>Separate Control + worker node configuration</li> </ul>
<b>Storage Support</b>	<ul style="list-style-type: none"> <li>Uses Storage Scale CNSA to mount new or existing Storage Scale software or Storage Scale System (hardware)</li> </ul>
<b>RAG Pipeline</b>	<ul style="list-style-type: none"> <li>NVIDIA NeMo Retriever Extraction Microservices</li> <li>Supports text, tables, and charts</li> </ul>
<b>Data Source Support</b>	<ul style="list-style-type: none"> <li>Storage Scale Data</li> <li>S3 sources with S3</li> </ul>

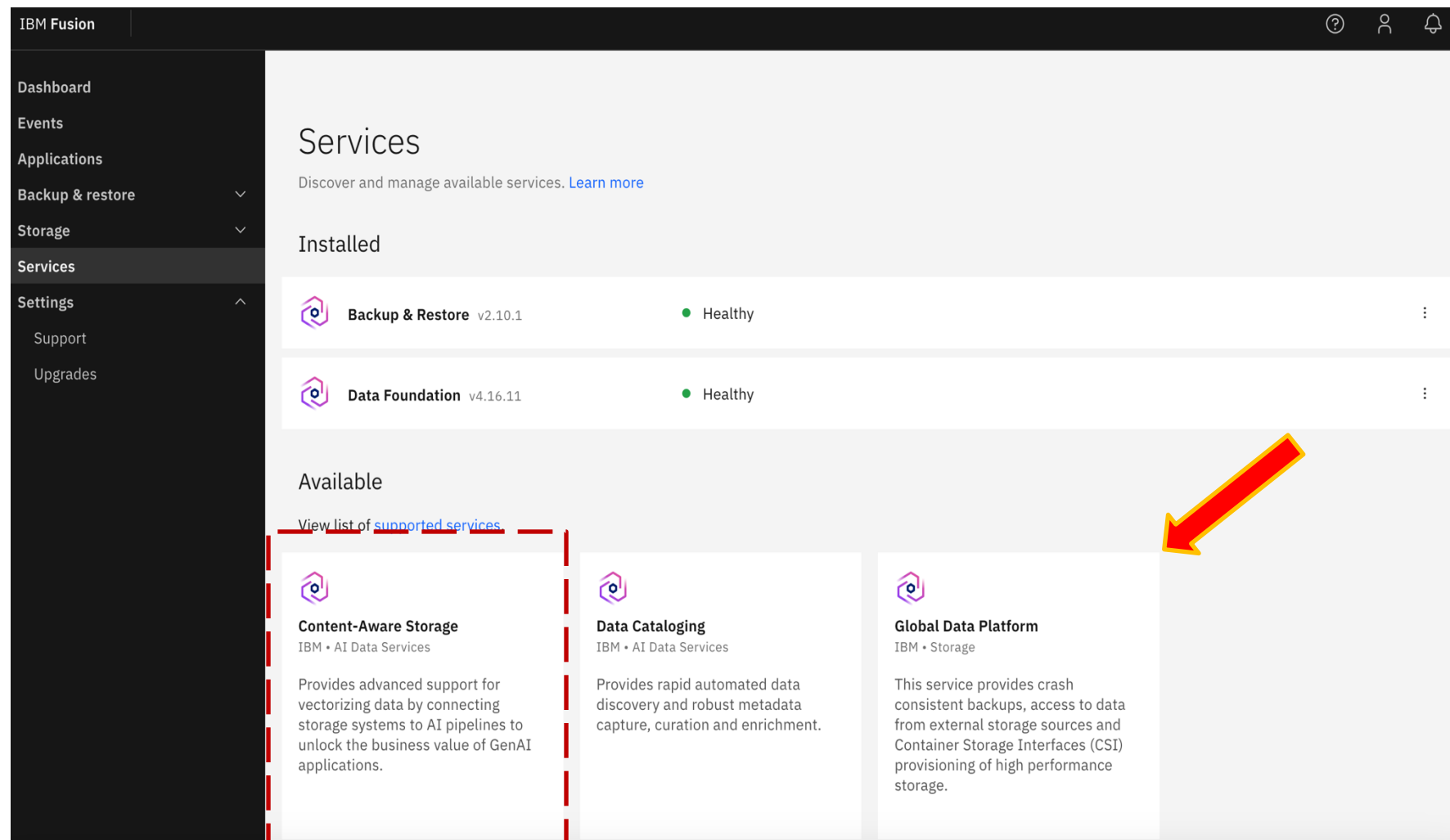




## Components – Deploying Through Fusion

CAS is deployed through Fusion

- Must be 2.10, OCP 4.17 or higher
- Automatically installs as a service in the Fusion UI
- Requires Global Data Platform external mount



The screenshot displays the IBM Fusion web interface. On the left is a dark sidebar with navigation links: Dashboard, Events, Applications, Backup & restore, Storage, Services (highlighted), Settings, Support, and Upgrades. The main content area is titled 'Services' with the subtitle 'Discover and manage available services. [Learn more](#)'. Below this, there are two sections: 'Installed' and 'Available'.

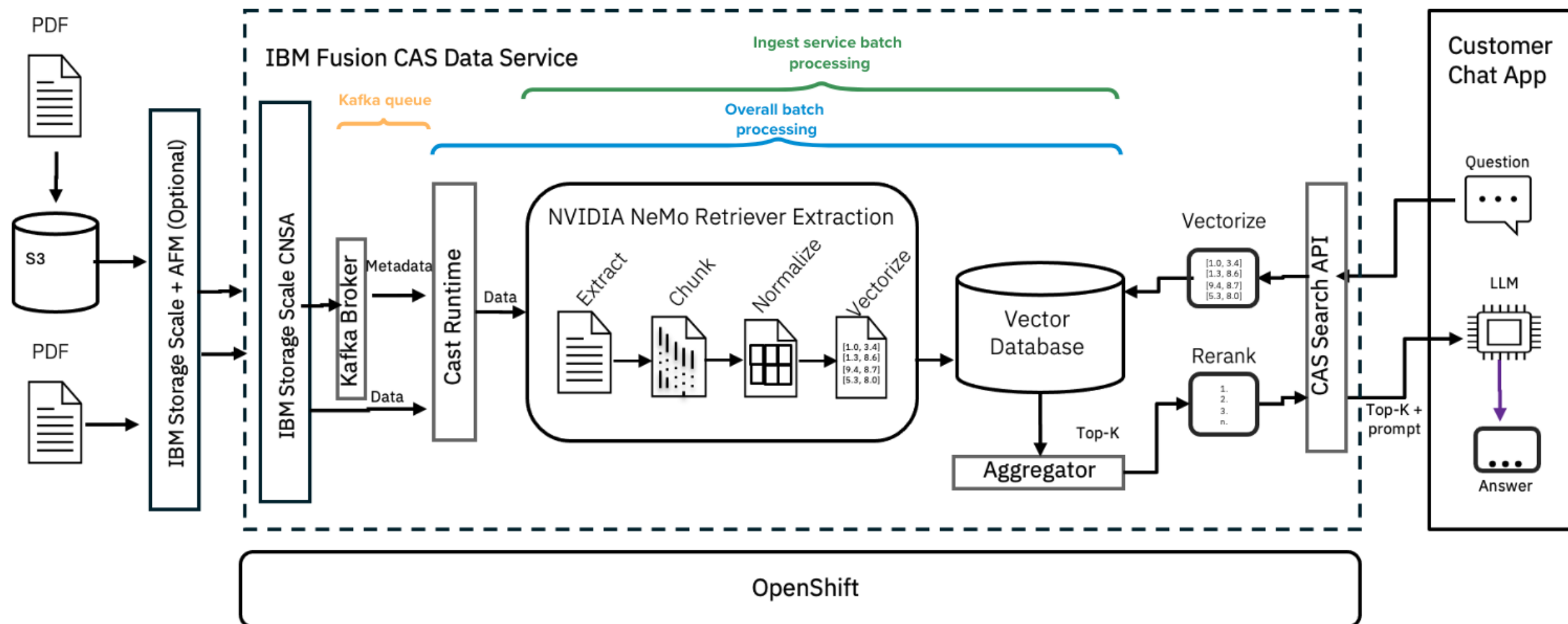
The 'Installed' section lists two services:

- Backup & Restore** v2.10.1, status: Healthy
- Data Foundation** v4.16.11, status: Healthy

The 'Available' section has a link 'View list of [supported services](#)'. It contains three service cards:

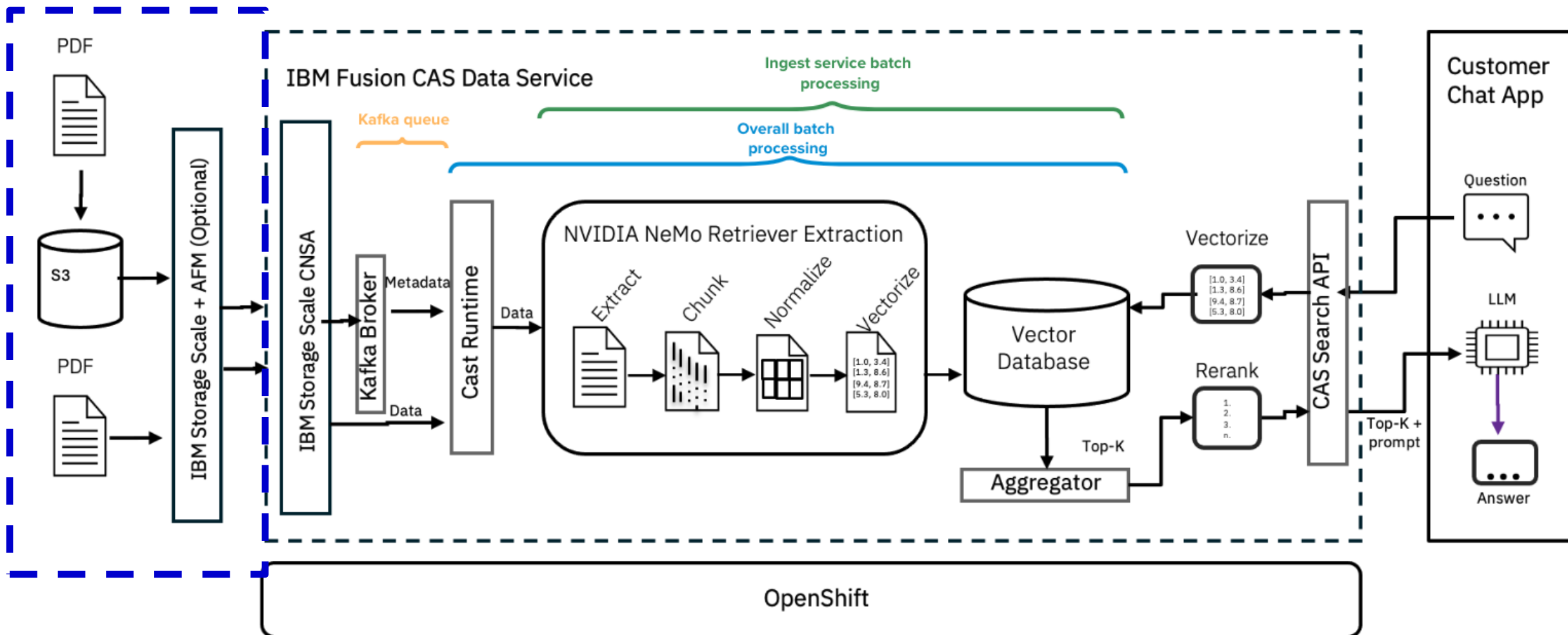
- Content-Aware Storage** (IBM • AI Data Services): Provides advanced support for vectorizing data by connecting storage systems to AI pipelines to unlock the business value of GenAI applications. This card is highlighted with a red dashed border.
- Data Cataloging** (IBM • AI Data Services): Provides rapid automated data discovery and robust metadata capture, curation and enrichment.
- Global Data Platform** (IBM • Storage): This service provides crash consistent backups, access to data from external storage sources and Container Storage Interfaces (CSI) provisioning of high performance storage. A large red arrow points to this card.

## CAS Service



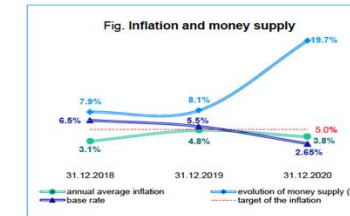
## CAS Service – Data Source Management

### Data Source Management

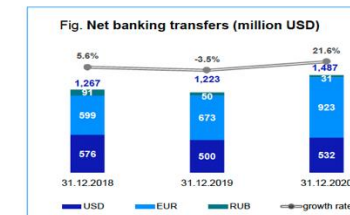


## CAS Service – Data Source Management

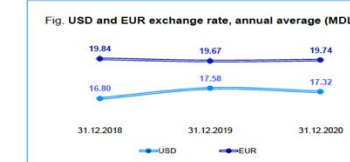
- Where does CAS Data come from?
- Current solution supports external Storage Scale
- Supported data types are multimodal unstructured data such as:
  - Text, tables, charts, and infographics
  - PDF, docx, pptx, txt documents
- However, expect more data types in the future



In 2020, the volume of remittances from abroad to individuals, on a net basis, increased by 21.6% compared to 2019 and amounted to 1,487 million US dollars, recording the maximum value from 2015 to present.



At the end of 2020, the exchange rate of the national currency was 17.21 MDL for 1 US dollar, similar to the beginning of the year. Against the Euro, the national currency depreciated by 9.7%. The average exchange rate of the national currency against the US dollar appreciated by 1.5% and against the Euro it depreciated by 0.4%. The stock of Official Reserve Assets of the NBM on 31.12.2020 reached a historical maximum of 3,783.5 million US dollars, increasing by 23.7% compared to the level recorded at the end of 2019.



## CAS Service – Data Source Management

CAS requires attachment to a remote IBM Storage Scale filesystem

- Can be configured via IBM Fusion UI
- Requires Storage Scale 5.2.2 or 5.2.3 software on remote Scale filesystem
- Supports SDS (e.g. IBM Storage Scale software on commodity hardware) and IBM Storage Scale Server (e.g. IBM Storage Scale Server 6000)

IBM Fusion

Dashboard  
Events  
Applications  
Storage  
Local storage  
**Remote file systems**  
Content-aware Storage  
Domains  
Data sources  
Services  
Settings

### Remote file systems

IBM Fusion provides OpenShift applications access to IBM Storage Scale file systems for accessing data and utilizing storage for backup jobs. [Learn more.](#)

#### Health

✓ Global Data Platform (service)

#### Encryption

No encryption server connected yet.

[Connect](#)

Search ⚙️ Add +

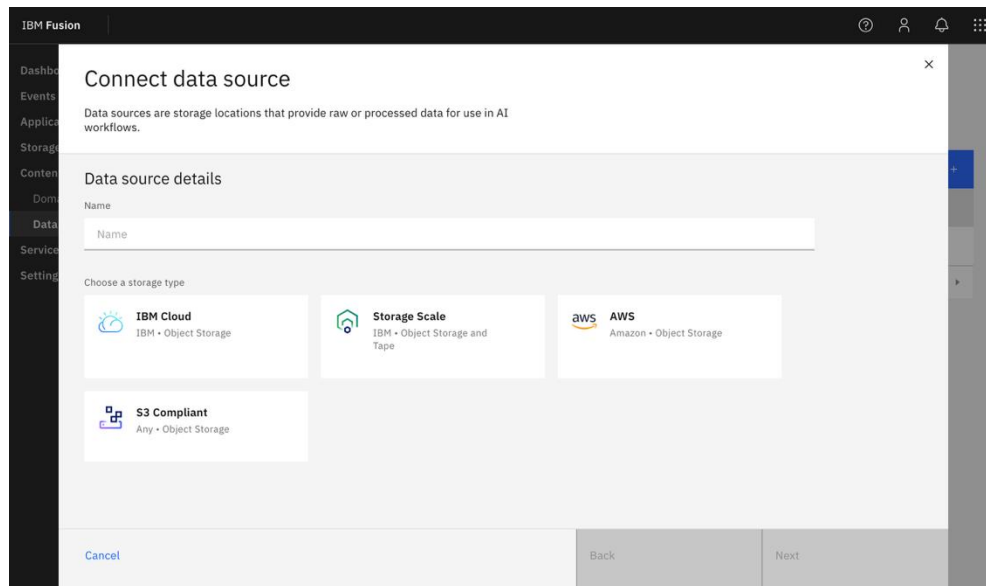
File system	FS status	Cluster	Cluster status	Used	Capacity	Storage class
<a href="#">gpfs2</a>	✓ Connected	tc11scale1.rtp.raleigh.ibm.com	✓ Connected	45.93 GiB	14,307.91 GiB	ibm-spectrum-fusion-gc3

Items per page: 25 1–1 of 1 item 1 1 of 1 page

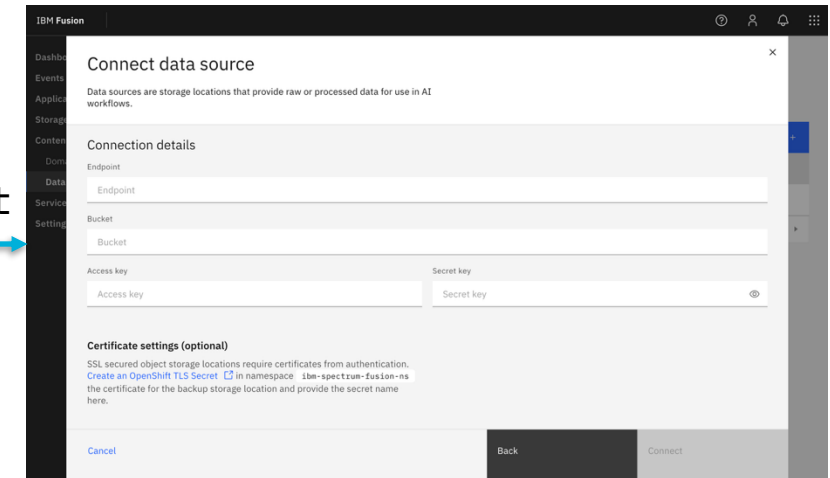


## CAS Service – Data Source Management

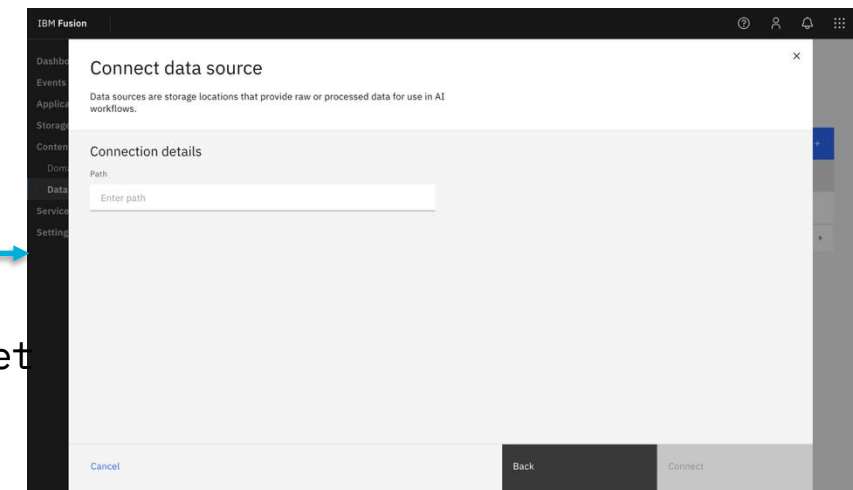
- Select the data source from the available selections
- IBM Storage Scale filesets on remote mounted IBM Storage Scale filesystem
- External S3 object storage buckets (IBM Cloud, Generic S3, AWS S3)
  - External S3 buckets require Storage Scale AFM on remote mounted IBM Storage Scale filesystem
- Storage Scale watch folders configured on remote mounted scale to enable incremental change processing
  - Remote mount Scale emits watch folder events to Kafka broker residing in CAS namespace



S3 Object



Scale Fileset



## CAS Service – Data Source Management - Configuration

### CAS Datasource CRD

Manages data sources for CAS

### Supported Data Source Types

- Supports external S3 buckets with AFM
- Supports internal Scale filesets

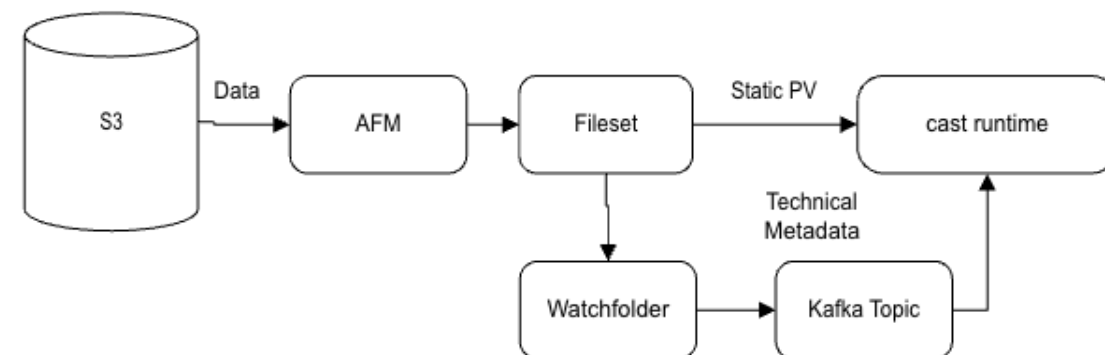
### Configuration Rules

- Max 25 CAS datasources<sup>1</sup>
- 1 to 1 mapping S3 bucket to AFM Fileset
- 1 to 1 mapping Scale fileset to WatchFolder
- 1 to 1 mapping Watch folder to Kafka topic
- 1 static PV per fileset using CSI volumeHandle
- Datasources map to one or more cas runtimes

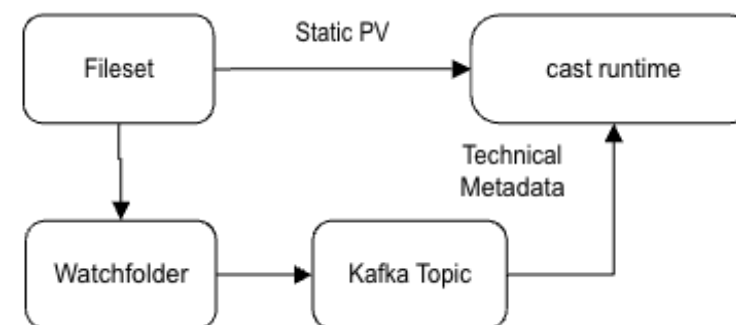
### Kafka / CAS Runtime Rules

- Each cast runtime is a separate Kafka consumer group
- Offset updated only after successful commit to vector DB (at least once semantics)
- File updates processed as delete and reinsert
- File deletes remove records from DB
- IN\_CLOSE\_WRITE, IN\_DELETE events tracked
- Delete datasource blocked until no pipelines using datasource

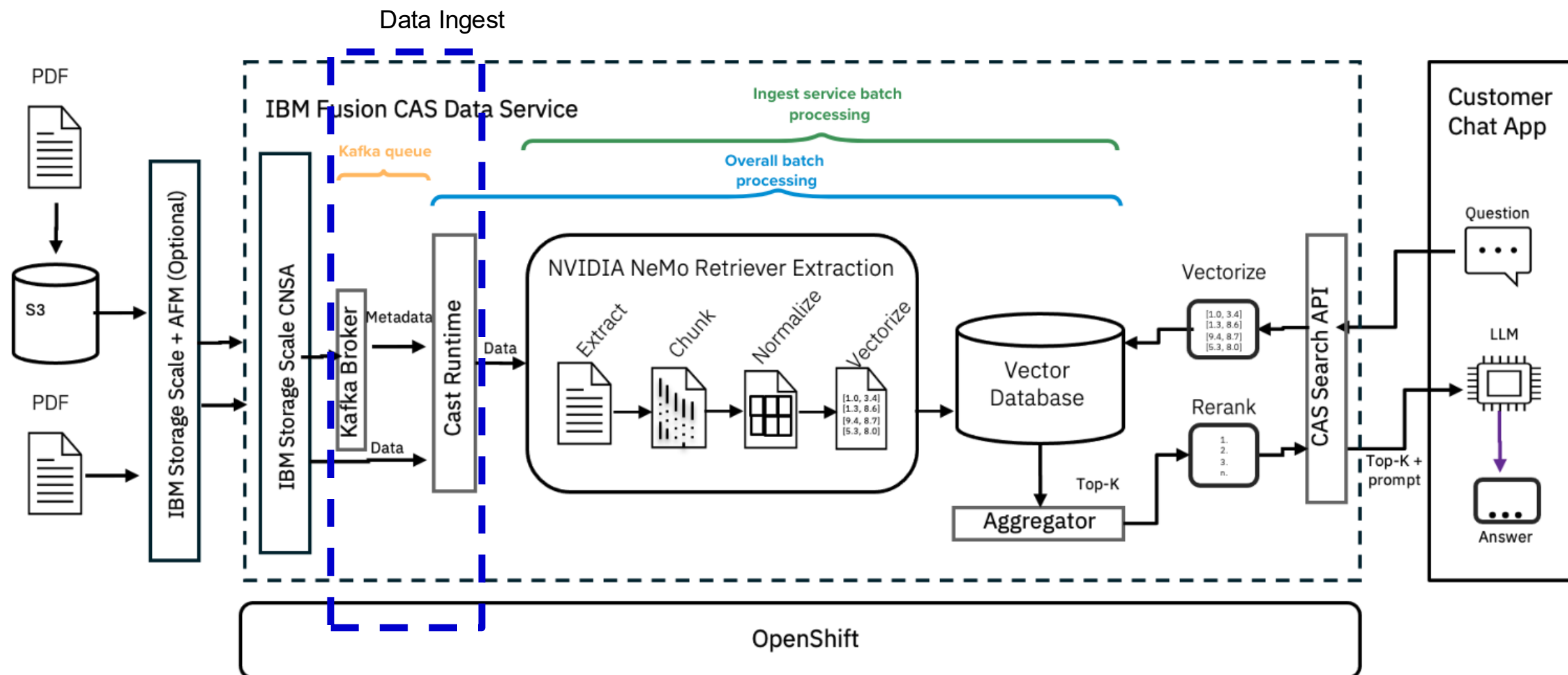
#### 1. External S3 Data Source



#### 2. Scale Fileset Data Source



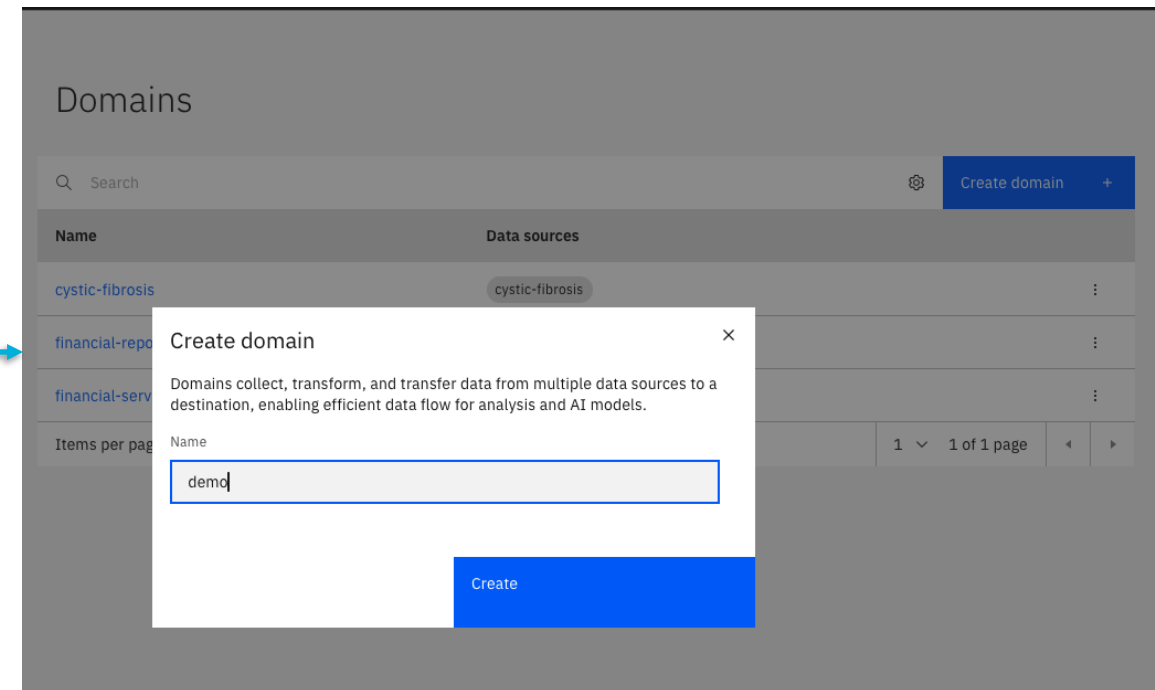
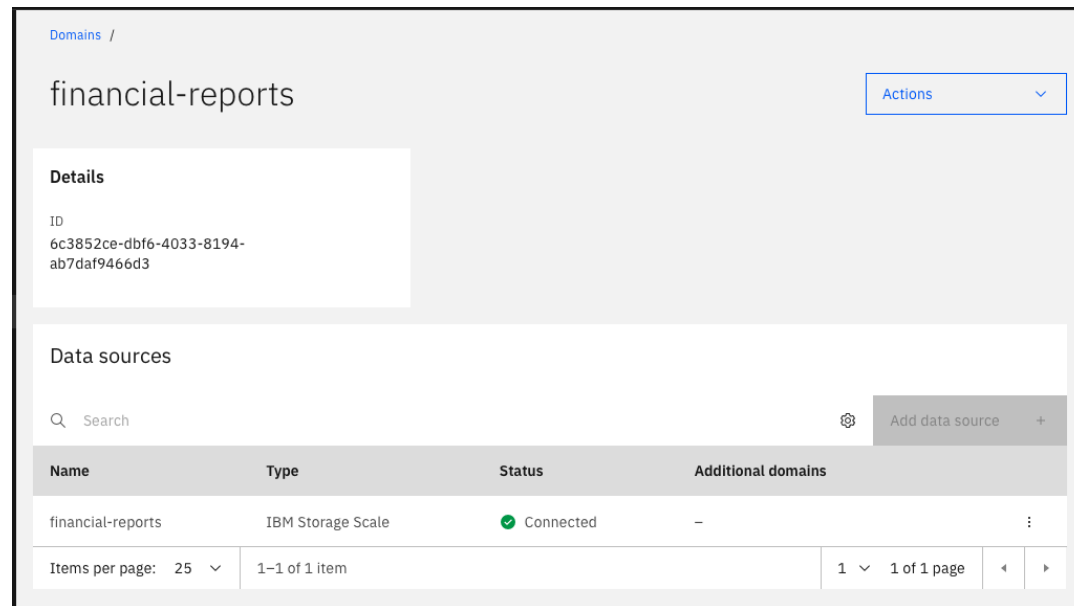
## CAS Service – Data Ingest



## CAS Service – Data Ingest

To ingest documents into CAS, create a domain

- The Domain CRD enables the users to configure which datasources can be group together
- It triggers the ability on CAS to process documents from this domain and store their vectors/embeddings in a common database table called collection.



# CAS Service – Data Ingest

Add datasource to domain. Choose between existing datasources or create a new one

Domains /

demo

Actions

Details

ID  
b7a80182-52b0-4019-b165-fb5087a2471e

Data sources

Search

Add data source

Name	Type	Status	Additional domains
<div><div></div><div>You haven't connected any data sources.</div></div>			

New Data Source

Connect data source

Close

Data sources are storage locations that provide raw or processed data for use in AI workflows.

Would you like to add a new or existing data source?

New

Existing

Data source details

Name

Choose a storage type

IBM Cloud

IBM • Object Storage

Storage Scale

IBM • Object Storage and Tape

AWS

Amazon • Object Storage

S3 Compliant

Any • Object Storage

Existing Data Source

Connect data source

Close

Data sources are storage locations that provide raw or processed data for use in AI workflows.

Would you like to add a new or existing data source?

New

Existing

Search

Name	Type	Status	Domains
<div></div> cystic-fibrosis	IBM Storage Scale	<div>Connected</div>	<div>cystic-fibrosis</div>
<div></div> financial-reports	IBM Storage Scale	<div>Connected</div>	<div>financial-reports</div>
<div></div> financial-services	IBM Storage Scale	<div>Unreachable</div>	<div>financial-services</div>

Items per page: 25

1-3 of 3 items

1

1 of 1 page

© Copyright IBM Corporation 2025

28



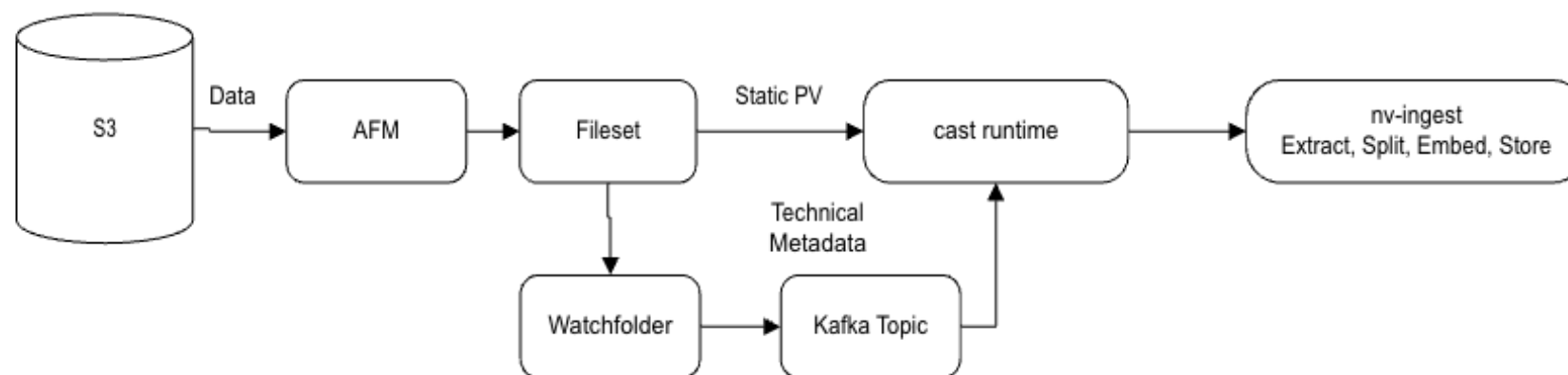
## CAS Service – Data Ingest – Domain Configuration Rules

### CAS Domain CRD

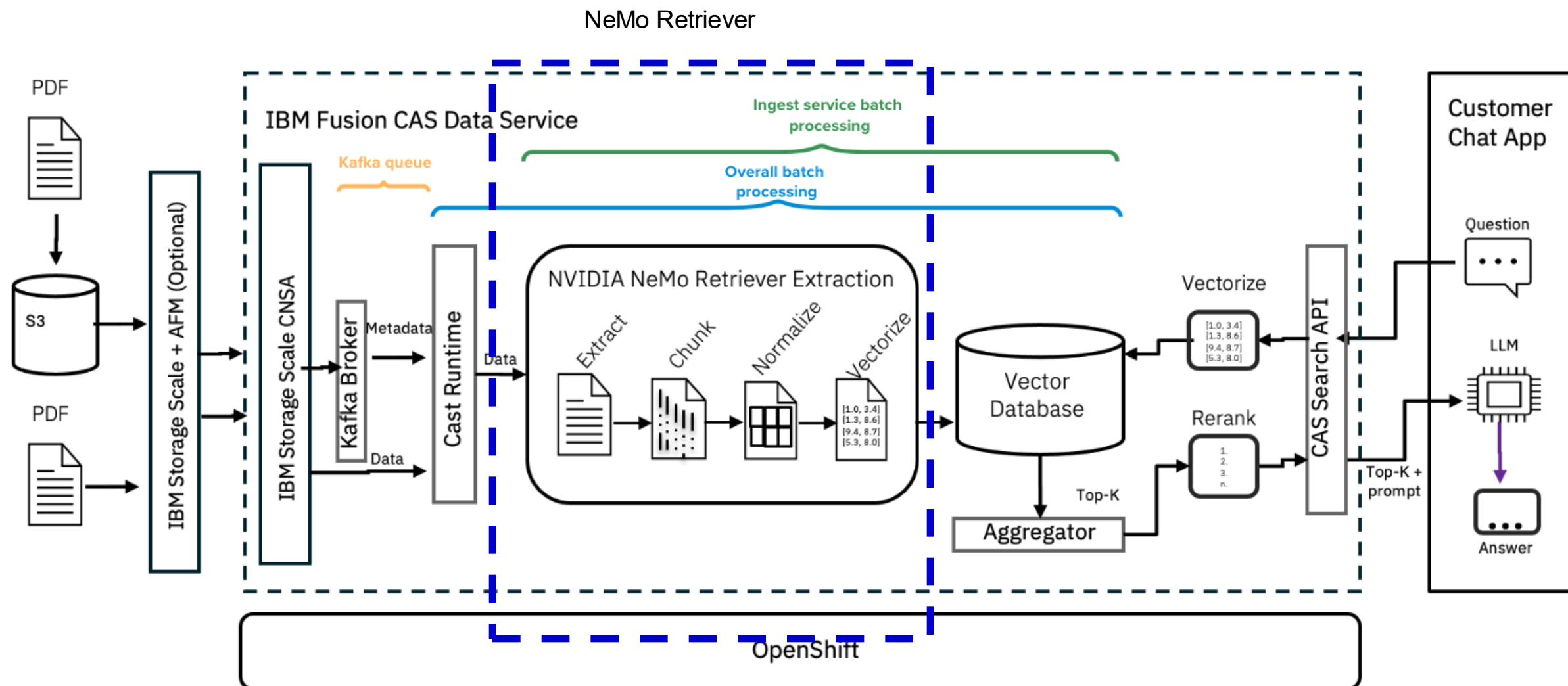
- Instantiates CAS runtime
- Manages ingest of data into RAG pipeline from one or more CAS data sources
- Triggers mapping of datasources into CAS runtime
- Subscribes to appropriate Kafka topics in cas runtime
- Performs ingest of data into RAG pipeline based on provided options (extract, split, embed, vector DB)

### Configuration Rules

- 1 to 1 mapping between pipeline and ParadeDB table
- Supports many to one mapping of datasources to pipeline
- Deleting pipeline removes database records

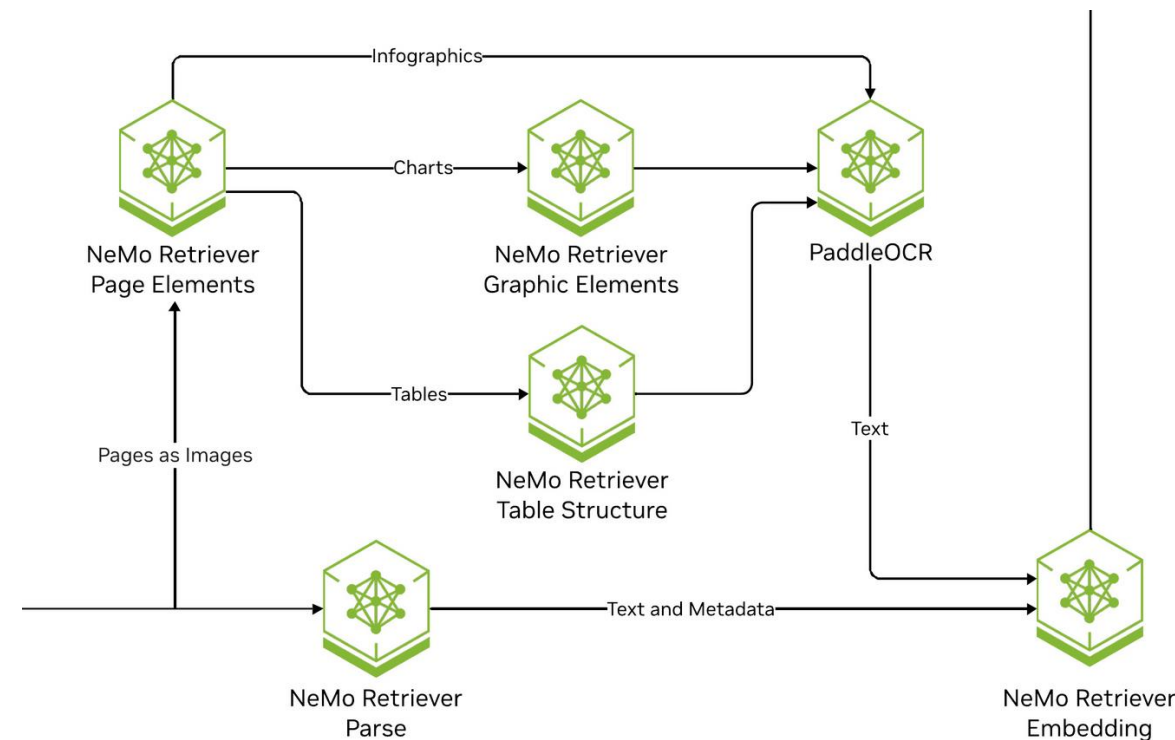


## CAS Service – NeMo Retriever

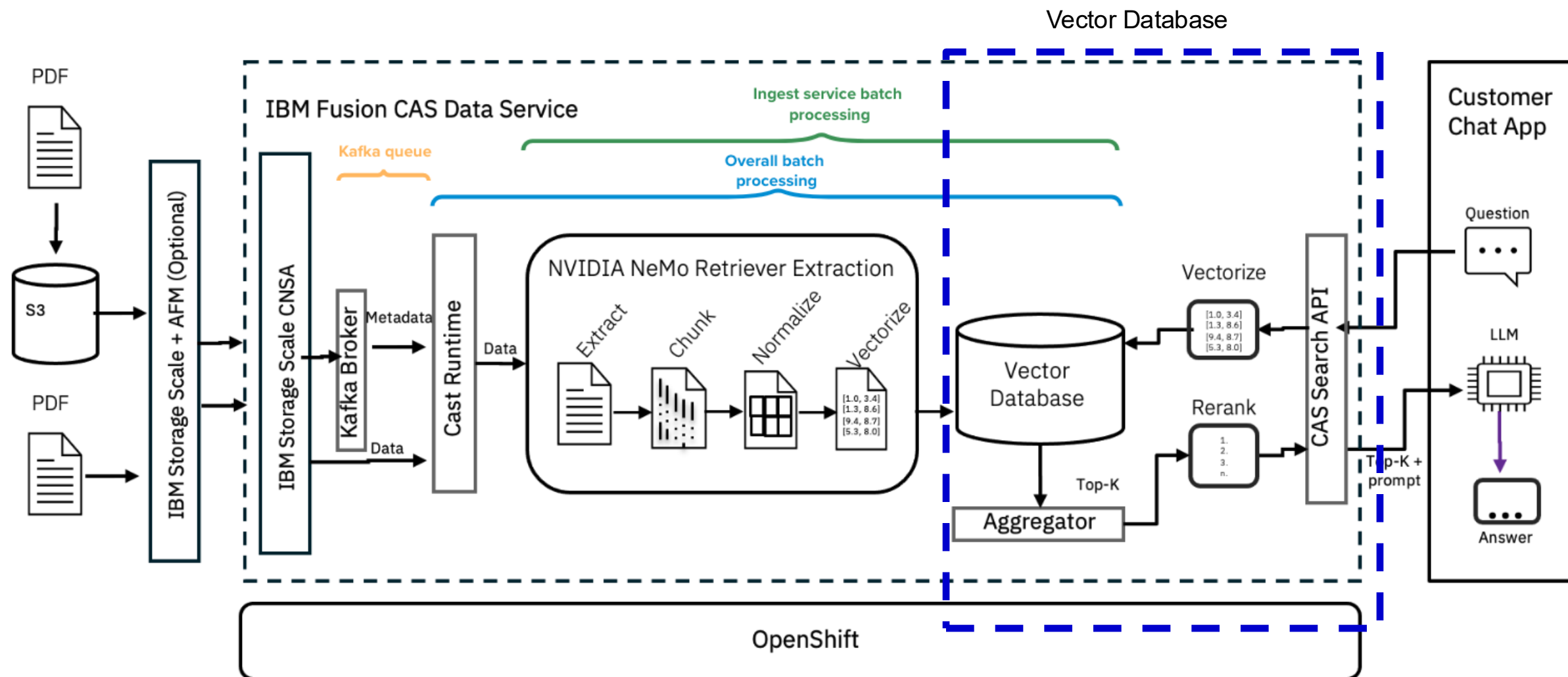


## CAS Service – NeMo Retriever

- The majority of the AI CAS processes occur within the NeMo Retriever
- NVIDIA NeMo™ Retriever is a collection of microservices for building multimodal extraction, reranking, and embedding pipelines with high accuracy and maximum data privacy
- Uses specialized NVIDIA Inferencing Microservices (NIM) to find, contextualize, and extract text, tables, charts and images that you can use in downstream generative applications.
- The NeMo Retriever extraction blueprint must be installed in the same OpenShift cluster running CAS
- Refer to the following links for additional details pertaining to the NeMo Retriever Extraction pipeline
  - <https://build.nvidia.com/nvidia/build-an-enterprise-rag-pipeline>
  - <https://github.com/NVIDIA/nv-ingest>
- Bonus! “NeMo” is short for “Neural Modules”

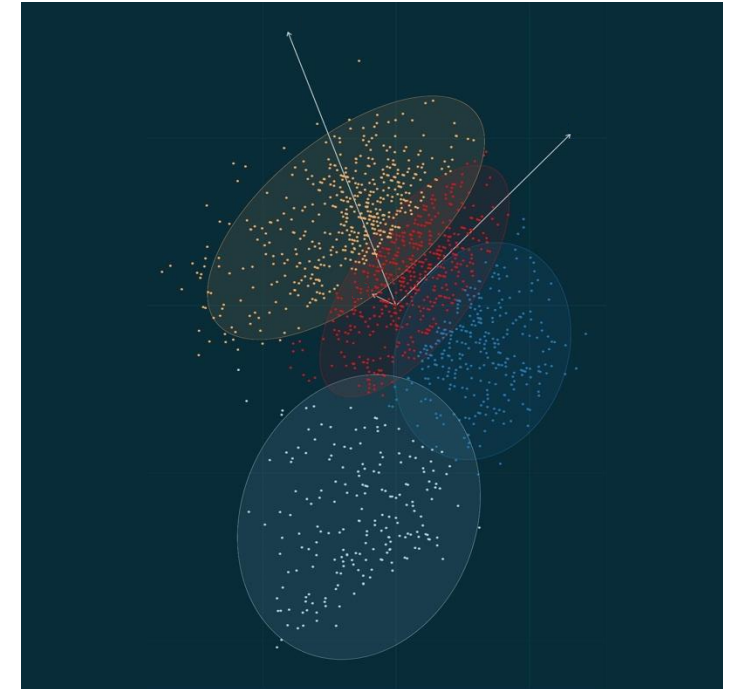


## CAS Service – Vector Database



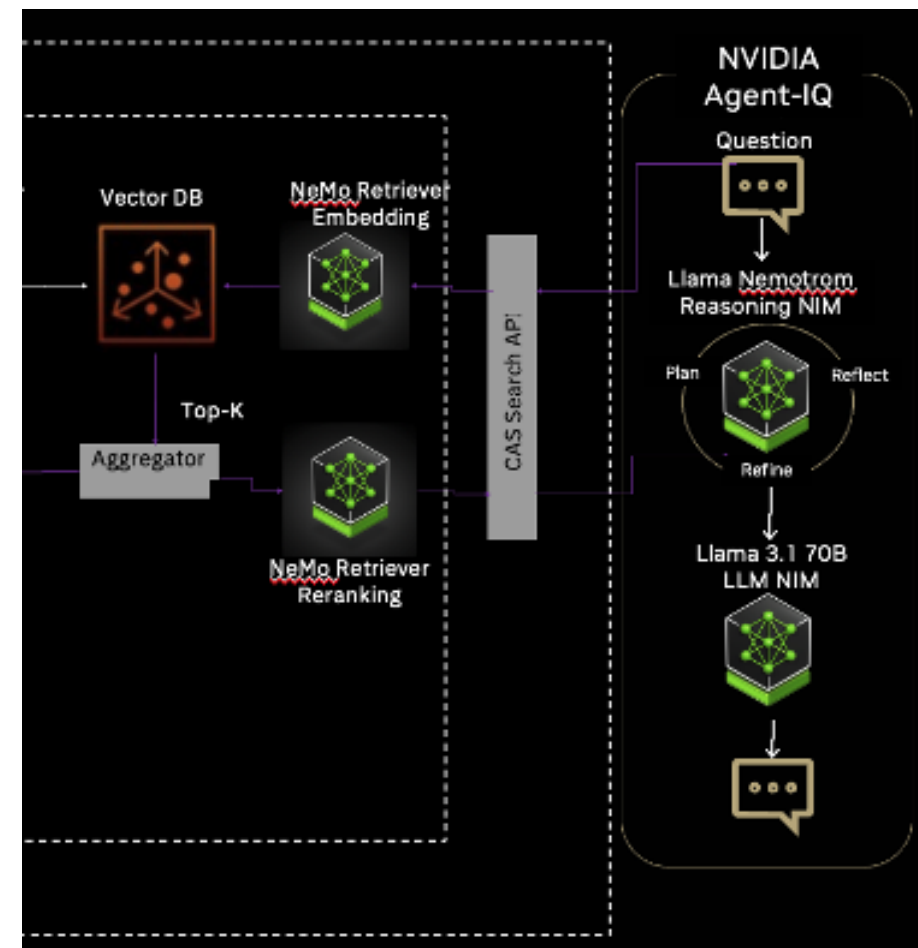
## CAS Service – Vector Database

- Vectors are a key component of the inferencing process.
- They are mathematical representations of data in a high-dimensional space.
  - Each dimension corresponds to a feature of the data
  - Number of dimensions ranging from a few hundred to tens of thousands, depending on the complexity of the data being represented
  - A vector's position represents its characteristics
  - Words, phrases, or entire documents, as well as images, audio, and other types of data, can all be vectorized
- Vectors are stored in the database as vector embeddings



## CAS Service – Vector Database

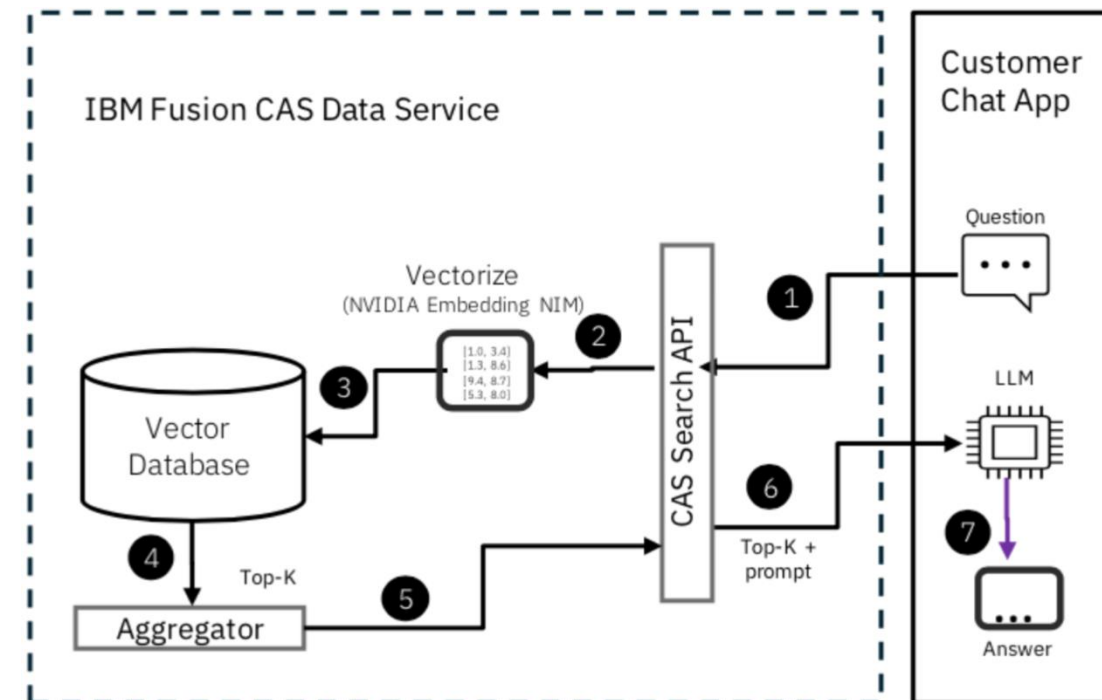
- The query workflow is divided into two parts:
  - Retrieval from CAS
  - Retrieval from a Large Language Model (LLM).
- The vector database exists within the CAS namespace to store data ingested from various files.
- The namespace also includes the query service pod and a FAST API (REST API) to plug into the querying application.
- For more information, see here: [Content-Aware Storage \(CAS\) APIs](#).



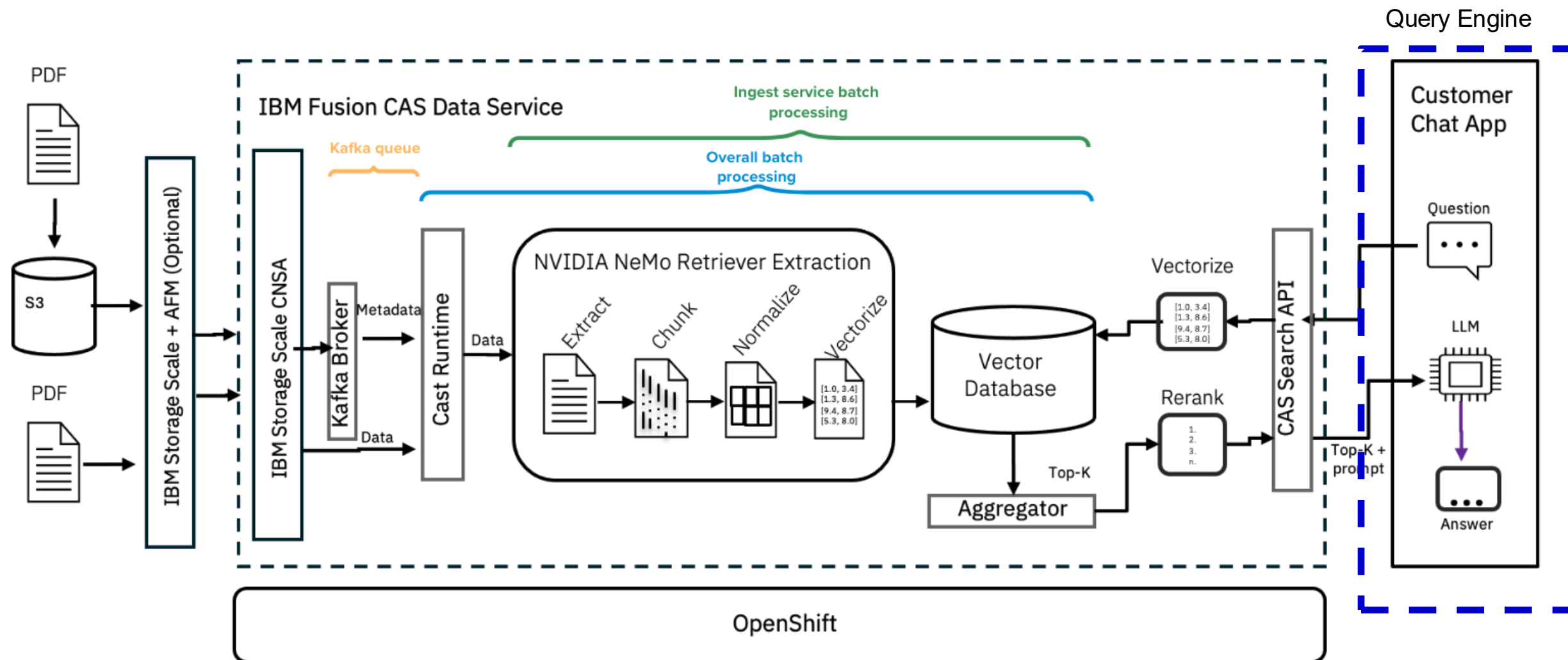


## CAS Service – Vector Database

1. A client application raises a question from its front end, sending a query REST Call to the query service pod through the FAST API.
2. The query service in turn goes with the query to the NVIDIA NIM Embedding Service.
3. This service generates embeddings for the data, enabling the query service to understand and process the request of the user effectively.
4. After the results are fetched, an SQL query is sent with the embedding to the database for top "K" number of Vector results.
5. When the user runs query from the front end, they can specify the value of this "K".
  - For example: "Select, column from table order by, limit by vector, with embedding."
  - The top "K" results are returned to the client application which in turn sends it to the LLM.
6. While embeddings make the retrieval process efficient, LLMs add a layer of contextual understanding, transforming raw data into meaningful information. Finally, the end results are published on the front end of the client application.

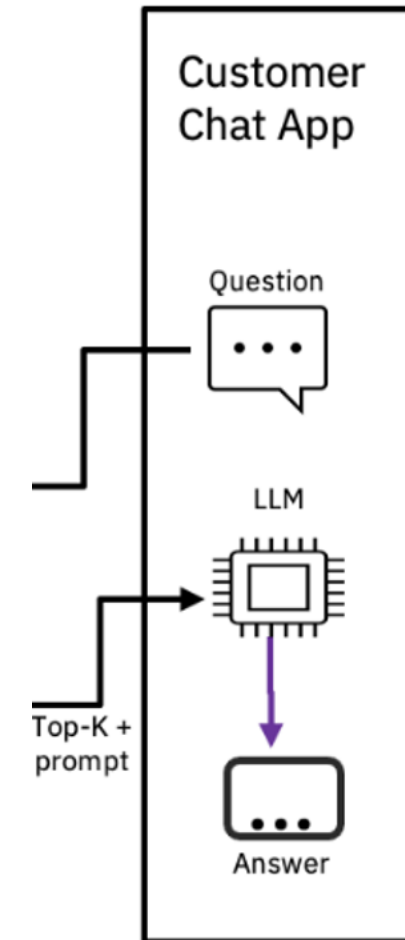


## CAS Service – Query Engine



## CAS Service – Query Engine

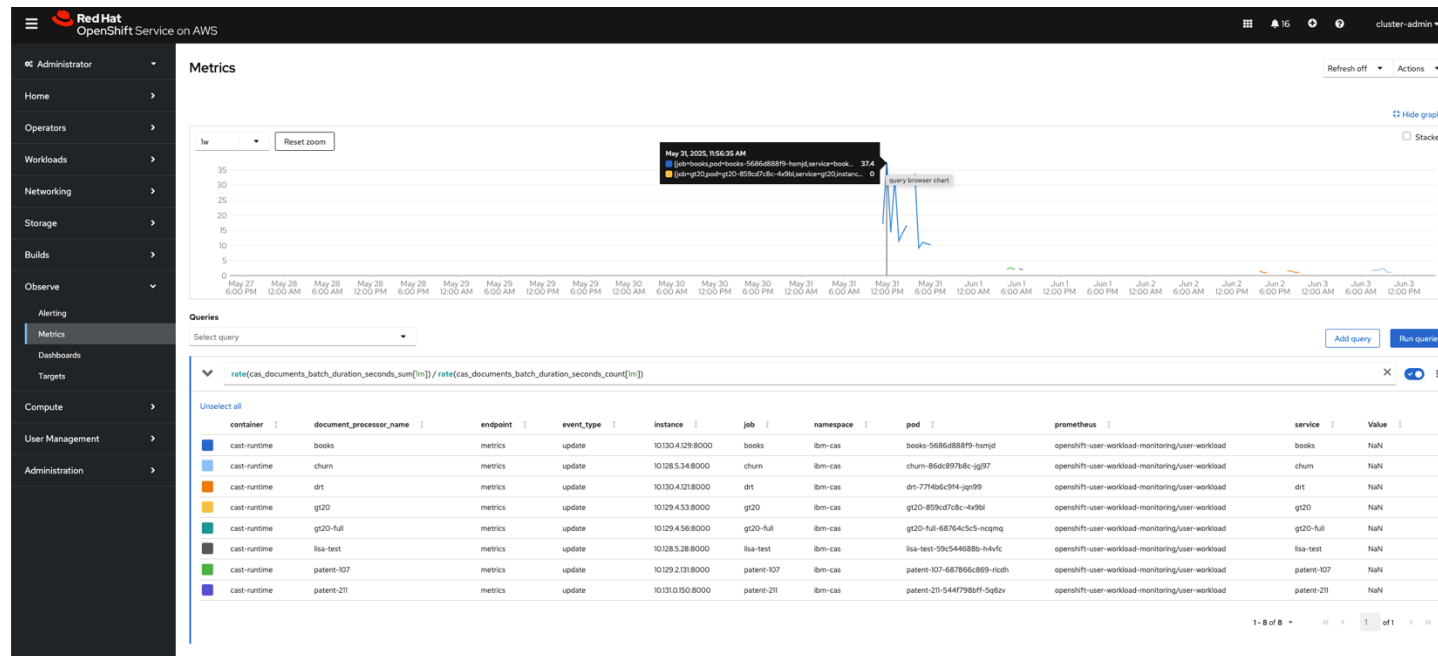
- NOT a part of the CAS Solution
- BUT...and integral part of the CAS Process
- Leverages natural learning models for query ingest
- CAS service provides a search API that supports Chat Applications,
  - BM25 keyword search
  - Semantic Search
  - Hybrid search (semantic + BM25 keyword).



## CAS Service – Bonus! Monitoring and Performance

CAS provides a set of Prometheus metrics to give insights on processing time and document sizes of the continuous data ingestion for your CAS domain

- Batch processing time
- Total documents in batch
- Individual document size
- Kafka queue, document size, and processing time



## CAS Service – Bonus! Monitoring and Performance

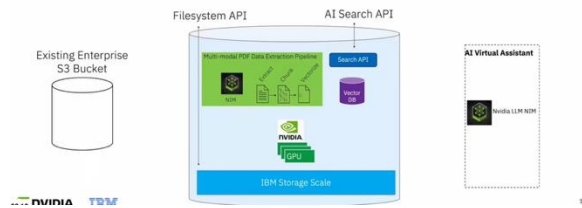
Insights on Kafka queue

- Document size
- Processing time



## Useful Links

Demo : IBM Content Aware Storage Scale



[Content Aware Storage Demo Presentation](#)



[CAS Blog by Vincent Hsu](#)



[CAS Accelerate by Chris Maestas](#)



Please take a moment to share your feedback with our team!

You can access this 6-question survey via [Menti.com](https://www.menti.com) with code 5151 0447 or

Direct link <https://www.menti.com/alhsf3bgvxu6>

Or

QR Code



# Thank you!