

# AI Fundamentals and 101



Matthew Klos  
Senior Solutions Architect  
Americas SWAT Team



# AI 101 Agenda

- What is AI?
- AI Ethics
- What is an AI model?
- Training vs. Inference?
- The 1% Problem
- What is Retrieval Augmented Generation (RAG)?
- What is Agentic AI?
- AI 101 Recap

# AI 101 - What is AI?

## Definition:

*Artificial intelligence (AI) is technology that enables computers and machines to simulate human learning, comprehension, problem solving, decision making, creativity and autonomy.*

LLMs

Deep  
Learning

Machine  
Learning

Generative  
AI

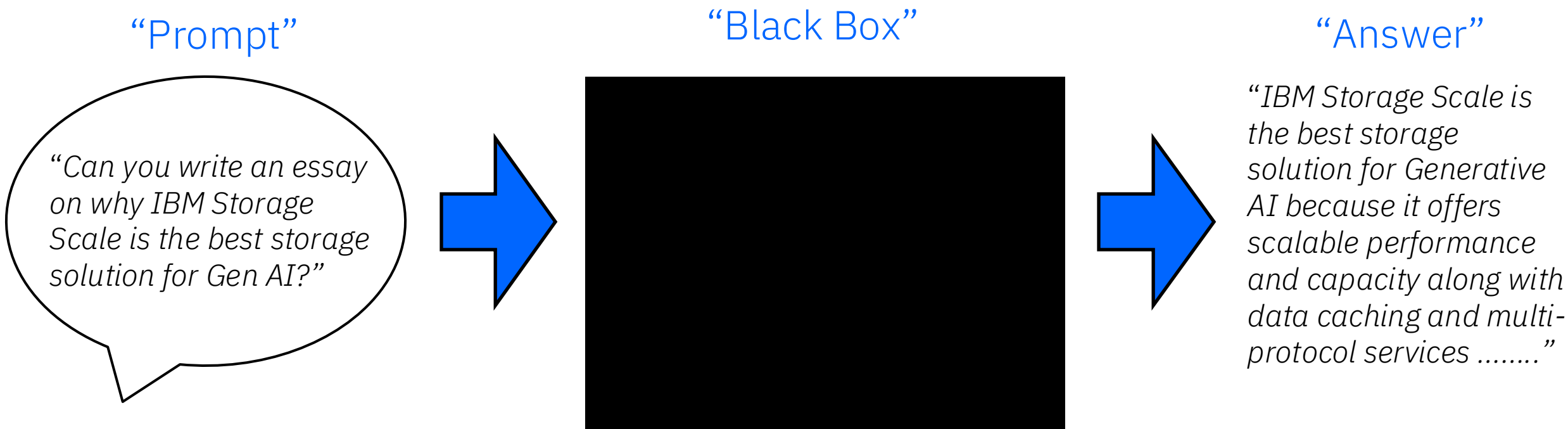
Agentic  
AI

And more....

# AI 101 - What is Generative AI?

## Definition:

Generative AI, sometimes called "gen AI", refers to deep learning models that can *create complex original content* such as long-form text, high-quality images, realistic video or audio and more in response to a user's prompt or request.



# AI 101 – AI Ethics

You'll hear a fair amount of discussion on the **governance** of AI to make sure that your AI has characteristics like:

- Fairness
- Transparency
- Explainability
- Regulatory Compliance



Every conference has sessions on Governance and ethics. It's an important part of what customers think about when embarking on AI initiatives

New technology can be **challenging** to adopt for enterprise business needs

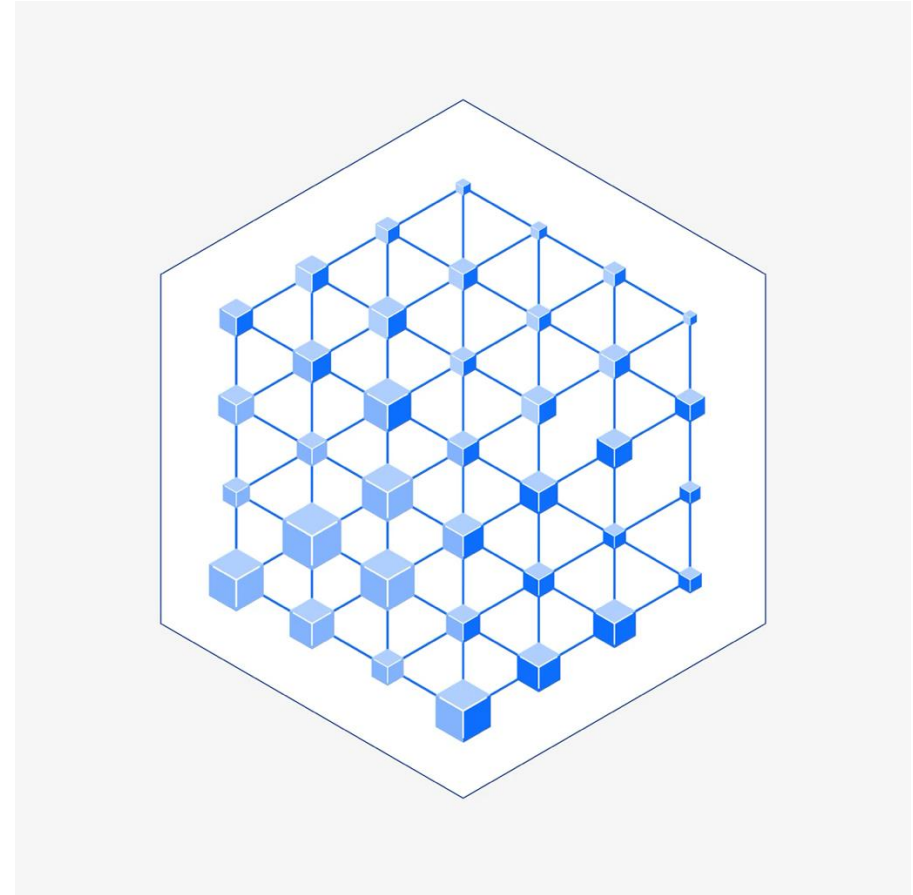
- LLMs may provide harmful, biased, or inaccurate results
- Vendors and users are sued for copyright infringement, not providing attribution or misuse of personal information
- Companies need more than just a general LLM
- Over 75% of consumers are concerned about misinformation from AI



# AI 101 – What is an AI model

## Definition:

*The AI model is a computational representation of real-world phenomenon or system learned from data. It is essentially a mathematical structure that learns patterns and relationships from data to make predictions or decisions about new, unseen data*



## AI Models Are Big !

Hundreds of billions of ‘parameters’

10 trillion parameters

Billions of ‘parameters’

Billions of ‘parameters’

Meta Llama at 405 Billion parameters

Alibaba M6 model

IBM Granite: 4B to 34B parameters

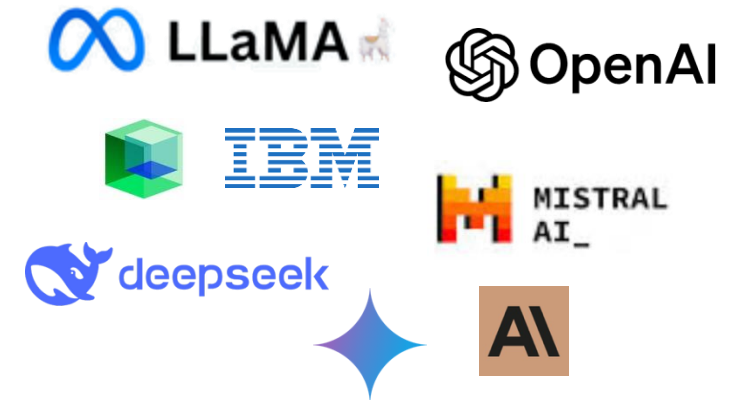
DeepSeek: 1.5B to 671B parameters



# AI 101 – Things to Know about Models

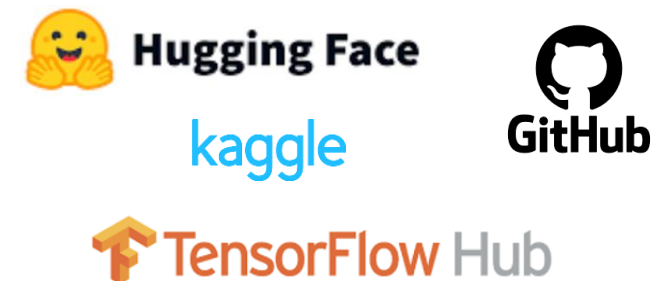
## 1) Customers may use one or many 3<sup>rd</sup> party models

- LLaMa Models from Meta
- GPT Models from OpenAI
- Granite Models from IBM
- Mistral from Mistral AI
- DeepSeek Models
- Anthropic Models
- Google Models



## 2) Models are acquired from model sites: (something like the Apple App store)

- Hugging Face
- Github Model Catalog
- TensorFlow Hub/PyTorch Hub
- Kaggle



# AI 101 – Training vs. Inference

## AI Training

AI training is the process of **teaching an artificial intelligence (AI) model** to perform tasks by exposing it to data. The goal is to enable the AI to learn and make accurate predictions.

---

## AI Inference

AI inference is the process of **using a trained AI model** to make predictions or conclusions from new data. It's the phase of the AI model lifecycle that comes after training.

# A little oversimplified but...

AI Training

Costs Money

AI Inference

Makes Money

or

Hopefully makes money

# AI 101 – AI isn't easy!

Puppy or Bagel ?



Mop or Sheepdog ?







- Day time vs nighttime
- The sun different colors at different times
- Shade
- Partial coverage from a tree branch
- Mist
- Hail
- Rain
- Snow
- Humidity
- Evening
- Nighttime

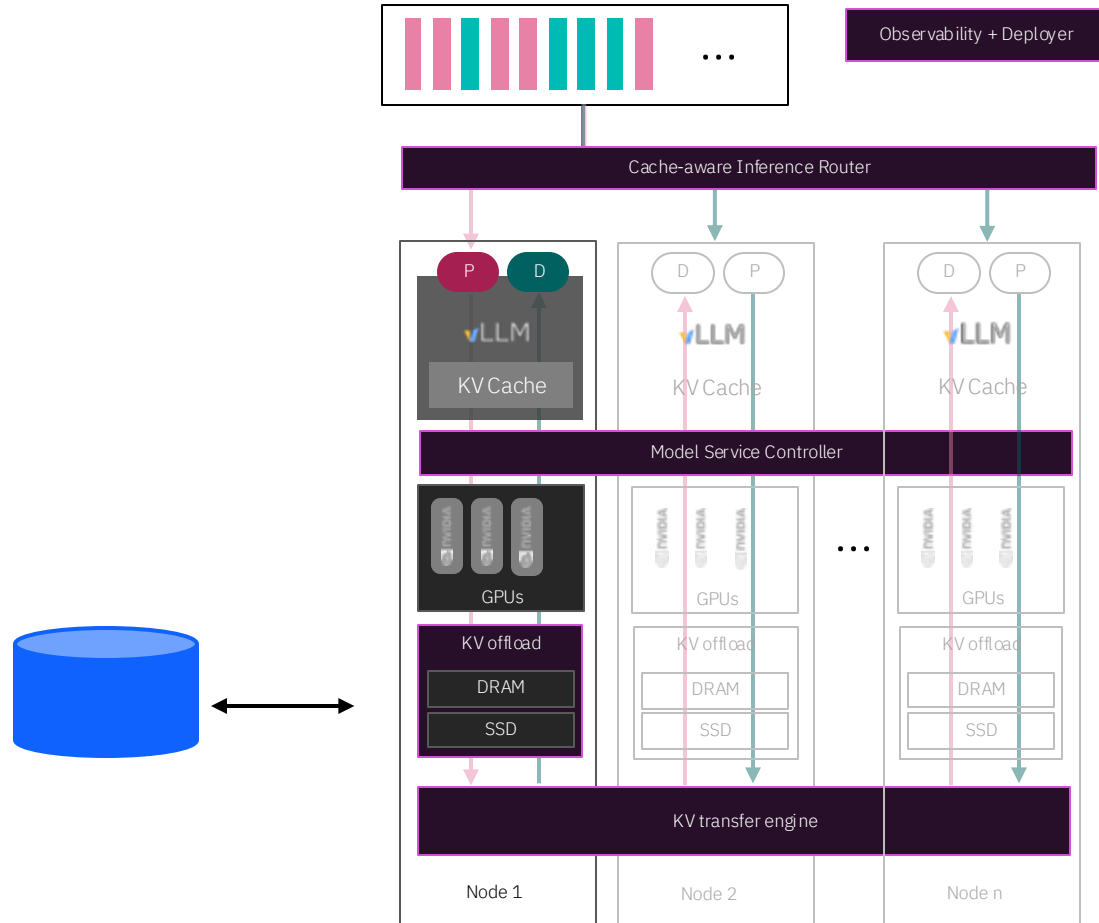
A good model requires a lot of training data.....

# AI 101 – AI Optimized Storage

## Fast storage accelerates and cost-optimizes AI

Storing and retrieving KV Cache from external storage offers many advantages:

- **Performance:** retrieving KV data from high-I/O storage dramatically reduces time-to-first-token
- **Scalability:** Persistence and reuse of larger volumes of KV data, beyond what fits in local compute resources
- **Cost efficiency:** Replacing GPU cycles with storage capacity can reduce overall costs
- **Persistence:** Ability to restore history and context across sessions, including GPU context-switching

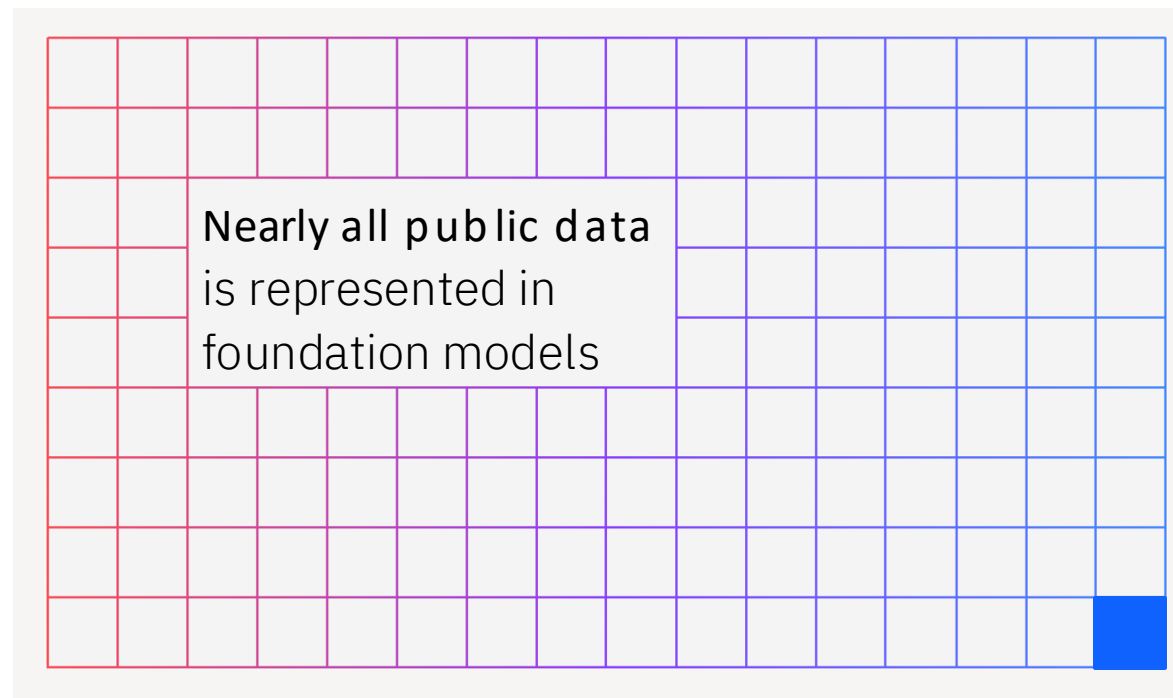


# AI 101 – The 1% Problem

- Organizations are swamped with unstructured data.
- But less than **1%** of all enterprise data was used to train major large language models.
- Retrieval Augmented Generation (RAG) improves inferencing by incorporating near real-time data.

Data is the fuel for an effective AI strategy

Only a small fraction of enterprise data is used in Gen AI



But only 1% of enterprise data!

# AI 101 – RAG

If the 3rd-party model only has 1% of my enterprise data that leaves a gap of 99% !!!!

What about AI for ..

- Customer service?
- Making my HR policies searchable?
- Using AI to help with IT support?
- Sales enablement?
- Predictive maintenance?
- Fraud detection?
- Writing SW code for me?

Something needs to fill this gap !





# AI 101 – Retrieval Augmented Generation (RAG)

An AI framework for improving the quality of LLM-generated responses

Grounds the model on additional sources of knowledge to supplement its internal representation of information

Examples of contextually relevant info:

- IBM Fusion or Nuclear Fusion?
- Flame (fire) or Flame (lover or crush)?
- Check in a bank or a body check in hockey?

RAG involves three basic steps:

- 1 Search for relevant content in your knowledge base
- 2 Pull the most relevant content into your model prompt as context
- 3 Send the combined prompt text to the model to generate output

Significantly elevates level of trust:

- Ensures that the model has access to the most current and reliable facts
- System becomes "business-aware"
- Sources are known, ensuring output can be checked for accuracy
- Less likely to make-up a factually inaccurate responses, with ability to say, "I don't know."

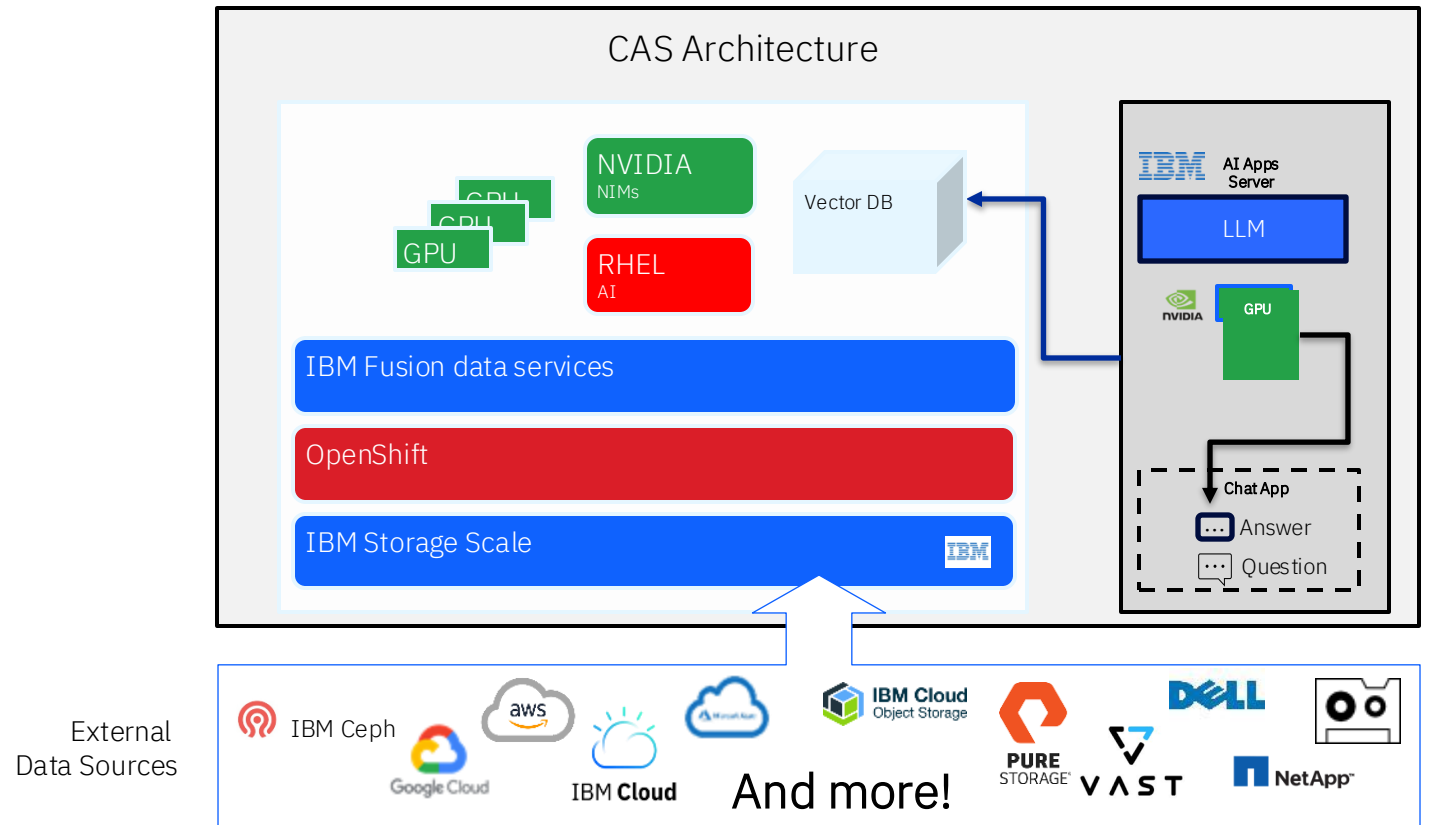


# AI 101 – 2025 Gartner Storage Trends

1. Enterprise storage platforms can serve generative AI (GenAI) workloads such as retrieval-augmented generation (RAG) inference and small-scale fine-tuning of language models, eliminating the need for a separate data lake.
2. Nearline SSD flash storage provides the opportunity to replace hybrid hard-disk drive (HDD) arrays with cost-efficient quad-level cell (QLC) flash solutions that improve overall performance, space efficiency and carbon emissions.
3. Cyberstorage solutions can be an additional layer of active defense at the storage layer to augment the traditional practice of deploying security at the network or application layers.
4. Integrated data intelligence enables unstructured data to become queryable data repositories to allow retrieval of data at a subobject level to support data analytics and AI and GenAI workflows.
5. Hybrid cloud storage presents the opportunity to optimize costs and enhance flexibility by seamlessly integrating on-premises and cloud resources. This approach supports global data growth, simplified management and the facilitation of hybrid cloud data workflows such as AI and GenAI.

# AI 101 – Content Aware Storage (CAS)

- Provides flexibility to integrate data pipelines with IBM Fusion data services
- Integrates Storage Scale for data access and storage optimization
- Integrates advanced vector database and enables hardware acceleration
- Near Real-time vector database updates



# AI 101 – Making your Data Searchable

## RAG (Retrieval Augmented Generation)

Retrieval Augmented Generation enhances the capabilities of large language models (LLMs) by integrating an information retrieval system with organizational data

- Simplified
- Topology
- Example



3<sup>rd</sup> party foundation  
model with public  
information



Vector database with  
organizational data

AI with:

More data  
Better answers  
Higher accuracy

- Plain Language Examples of what RAG helps solve:

IBM Fusion vs Fusion Energy or Nuclear Fusion  
Red Hat is a company, not just a hat  
Maximo is an IBM software offering, not a typo for Maximus



# AI 101 – RAG Use Cases

Assistants and Agents are two common AI systems used by many. It will be helpful to have a basic understanding of the differences between the two

## Assistants

Information retrieval  
Prescriptive tasks  
Single-Step processes

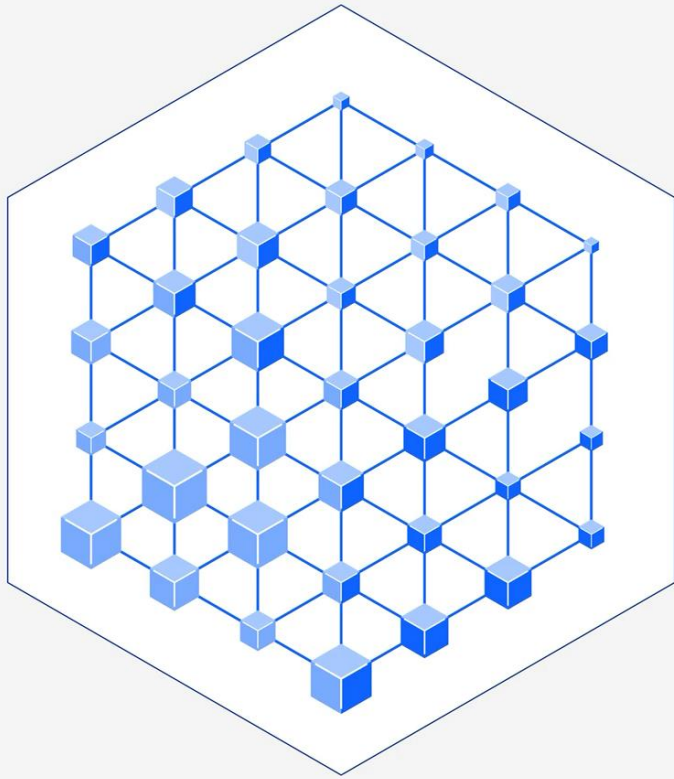


## Agents

Multi-step processes  
Autonomous action-taking  
Self-Correcting



# AI 101 -What is agentic AI?



Agentic AI is a **framework for accomplishing goals** with limited supervision that **consists of AI agents**.

In multiagent systems, **each agent performs a specific subtask** that's required to reach the goal.

# AI 101 – AI Agents, Agentic AI, MCP

## AI Agents

An AI Agent is an autonomous AI program, it can perform tasks and accomplish goals on behalf of a user or another system without human intervention, by designing its own workflow and using available tools (other applications or services).

---

## Agentic AI

Agentic AI is a system of multiple AI agents, the efforts of which are coordinated, or orchestrated, to accomplish a more complex task or a greater goal than any single agent in the system could accomplish.

---

## Model Context Protocol (MCP)

MCP is an open protocol that standardizes how applications provide context to large language models (LLMs)

# AI 101 – Recap

- AI automates tasks which require human intelligence
- AI Ethics and Governance is an important part of an AI project
- Models are big big big
- Training models is expensive and time consuming
- Training costs money and inference makes money
- Most organizations use 3<sup>rd</sup> party models
- 3<sup>rd</sup> party models lack organizational data which gets added with RAG
- IBM's Content Aware Storage helps customers to simplify, accelerate, and save money in AI RAG
- Agentic AI operationalizes Generative AI





# AI goes well beyond image recognition

Recommendation Engines

Predictive maintenance

Fraud detection

Virtual assistants

Code generation

Personalized Marketing

Drug discovery

Medical diagnosis

Sales enablement

Threat detection

Design

Analysis

Report creation

Image generation



# Suppliers of Training and Inference Chips

## AI Training

NVIDIA GPUs are the most commonly used accelerators



---

## AI Inference

There are many players who make inferencing accelerators



# Challenges with RAG

---

## Challenges:

- Enterprise data needs to be collected and curated
- Only a fraction of enterprise data is indexed today
- The data in the vector database keeps changing
- Updates to the DB are done daily or weekly as a batch process. This is time consuming and requires a lot of GPU resources
- Access to data needs to be handled properly
- Security policies for data access keeps changing

## Bottom Line:

- Difficult to implement
- Limited ability to scale

# IBM Content Aware Storage

Pssst....

IBM has Content Aware Storage to fill in some of these gaps

- Handles just the changes to the vector DB data in real time
- Handles just the changes to security access in real time

And guess what....

- This optimizes GPU usage to drive efficiency and sustainability

But that's a separate discussion....



# Limitations with RAG

---

## Limitations:

- RAG is not the only way to tune AI to improve results
- RAG may not work for tuning a VLM (Visual Language Model)
- One approach or one model may not work for many AI use cases
- Example: Try getting AI to generate a picture of a left-handed guitarist playing a guitar for lefties

## Bottom Line:

- AI is still difficult
- We are still early in the journey