HCLSoftware

openntf

HCL Domino IQ
Using RAG In Your Apps

Brian Arnold

Technical Advisor

HCL Domino+ Family

brian.arnold@hcl-software.com

DISCLAIMER

HCL’s statements regarding its plans, directions, and intent are subject to change or withdrawal without notice and at HCL’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.

Information regarding potential future products is intended to outline HCL's general product direction and should not be relied upon for any other purpose. Information relating to potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality, all products are made available for sale solely at HCL's discretion and shall be subject to relevant terms and conditions. Development, release, and timing of any future features or functionality described for our products remains at HCL's sole discretion.

Performance is based on measurements and projections using standard HCL benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon many factors, including considerations such as the amount of multiprogramming in the user’s job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve results similar to those stated here.

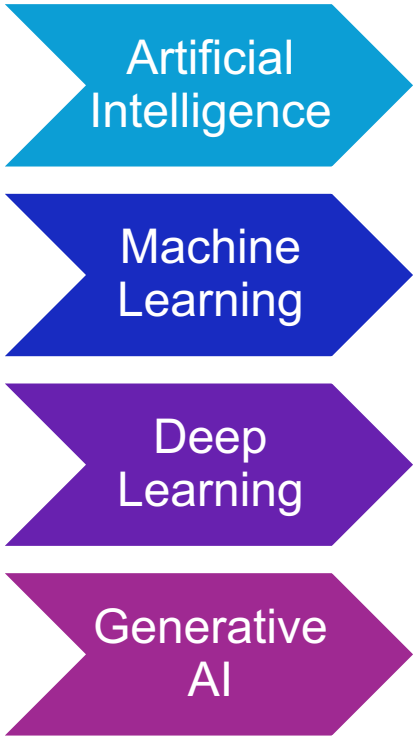
HCLSoftware

What is AI ?

Some Useful Information

HCLSoftware

Terminology



Artificial Intelligence (AI)
Overall name of the field of interest to have a machine to perform tasks typically associated with human intelligence, such as learning, reasoning, problem-solving, perception, and decision-making

Machine Learning (ML)
statistical algorithms that can learn from data and generalize to unseen data, and thus perform tasks without explicit instructions.
= Predicting patterns.

Deep Learning (DL)
Multi Layer Neural Networks (try to) simulate human brain.

Generative Artificial Intelligence (GenAI)
Create new content. A remix (not a repeat) of what it has been trained on.
Language, Model it, predict what next sentence, paragraph, …

Generative AI

What is it

- A (complex) system that creates new content based on patterns identified from training data.
- Uses Large Language Models (LLMs) to predict what comes next
- Output can be text, images, code, audio, video.

Watch out

- Generative AI does not “think” nor “understand” like a human.
- It's a pattern generator with strengths and limits
- Output quality depends on training data, prompts, and how well it’s integrated into real world scenarios.
- Hallucinations are a systematic problem of the industry.

What is an LLM?

Definition

- LLM = Large Language Model
- A neural network that learns language patterns and generates human-like text

What can it be used for?

- Write text (emails, poems, blog posts, code, etc.)
- Answer Questions
- Translate
- Summarize Information
- …and more

What is the size of an LLM ?

- Depends…
- From 2 GB to >150 GB+

Training Data

- Natural Language & Grammar of that language
- Facts
- Relationship between words

Training an LLM

- Requires large amounts of data
  - – Typically, Tera- or Peta-Bytes!
- E.g., Meta LLaMA-3 was trained on 24,000 GPU's!!

What is an LLM?

Definition

- LLM = Large Language Model
- A neural network that learns language patterns and generates human-like text

What can it be used for?

- Write text (emails, poems, blog posts, code, etc.)
- Answer Questions
- Translate
- Summarize Information
- …and more

Problem

- Can not provide answers that were not part of training set
- Tends to make things up (“Hallucination”)
- Output not limited to intended use-case

Guard Model

What is a Guard Model?

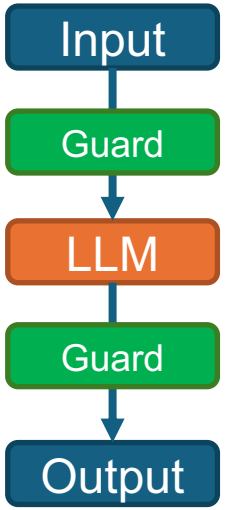
- An AI Model in front of your main LLM model
- Performs security checks against input / output given by / provided to user

What is it good for?

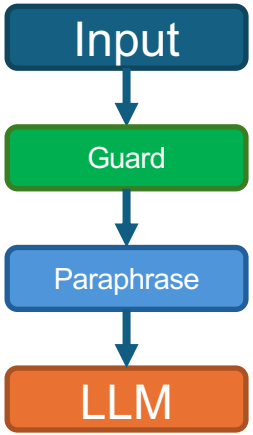
- Secure Input & Output of Domino IQ
- Detect prompt injection and jailbreak attempts
- Block data exfiltration (“Summarize these internal secrets…”)

Note:

- Guard Models are not perfect
- There is a performance overhead



Paraphrasing



- Rephrasing the input provided
- An additional layer of ‘security’
- Before the main model is queried, there is an additional query to “paraphrase” the user’s original query
  - Consumes time, tokens, and electricity
  - May alter the meaning of the query
  - (often) Translates the query into English
- Not all LLMs can do this
- Enabled by default, can turn it off with this Notes.ini
  - DOMIQ_DISABLE_PROMPT_PARAPHRASE=1

https://help.hcl-software.com/domino/14.5.1/admin/conf_security_considerations_for_iq.html

https://blog.nashcom.de/nashcomblog.nsf/dx/domino-iq-paraphrasing-explained.htm

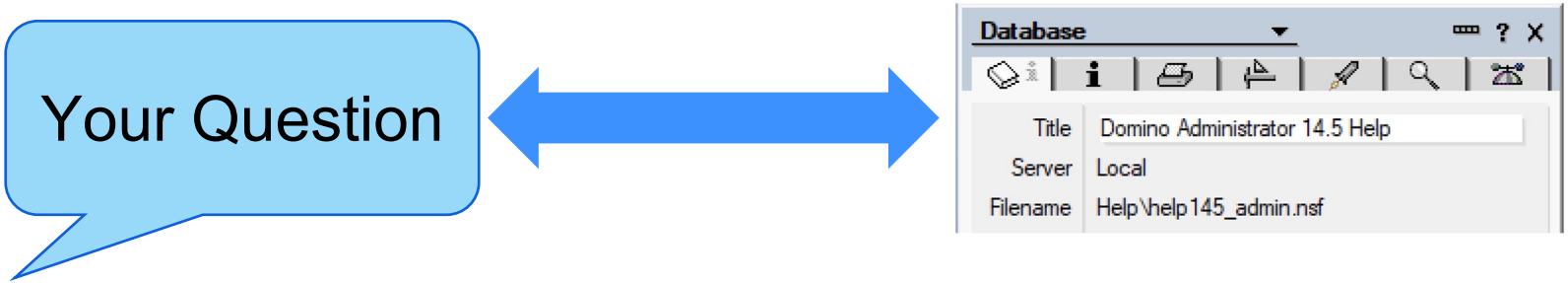
Retrieval Augmented GenAI (RAG)

Users expect:

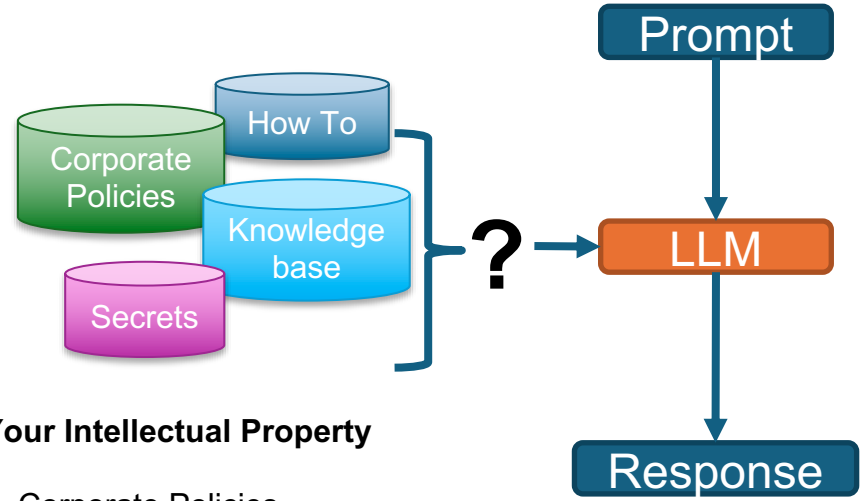
- Using the information stored in a (Domino) Data source (NSF) to correctly respond to an AI request.

Problem:

- Content of the data source (Domino Database) keeps changing
- Training an AI on this data would take too long, is too resource intensive, and is too expensive ($200k+)



What is RAG? (in General)



Your Intellectual Property

- Corporate Policies
- Knowledge Base
- Company Secretes
- Files, Wiki, Blogs, …

RAG = Retrieval Augmented GenAI

- Infuse additional information into the ‘Knowledge’ that the AI backend will have available to compose an answer from.
- Sovereign RAG : Leverage this information without exposing it to external entities.

Benefits:

- RAG allows new content to be provided to an AI (much) faster than training or updating an LLM
- Better quality of answers due to more current context
- Reducing “hallucinations”

Vectors – Context Matters (Example)

Tokyo Hotel

Music Band

Tokyo Hotel

In Chicago

Tokyo Hotel

Hotels in Tokyo

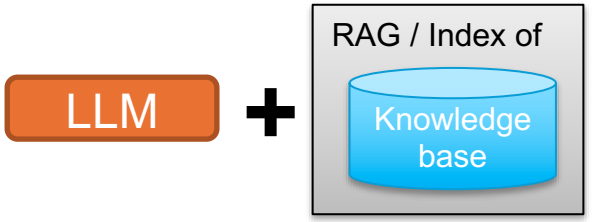
GenAI vs. "RAG" = Retrieval Augmented Generation AI

GenAI

- Answers are based on what it learned during training.
- Training data is **months** or even **years** old
- Can't see your company docs, PDFs, databases, etc. unless part of the actual (single) request

RAG

- Leverage current / more up to date data sources
- RAG 'index' usually **hours**, **days** or **weeks** old
- RAG technically is a matrix full of numbers that provide additional information to an LLM to provide better answers.



AI Systems: A Human Analogy

A simple way to understand how AI systems evolve from intelligence to enterprise action.

Domino 2026

LLM

LLM

- Core intelligence and reasoning
- Generates language and ideas
- The foundation model layer

Brain

RAG

RAG

- LLM + trusted external knowledge
- Grounded in documents and data
- Improves relevance and accuracy

Brain + Knowledge

Domino IQ with Agents

ACT

AI Agent

- Takes actions toward a goal
- Uses tools, memory, and workflows
- Moves from answers to execution

Brain + Hands

Domino Next

MCP

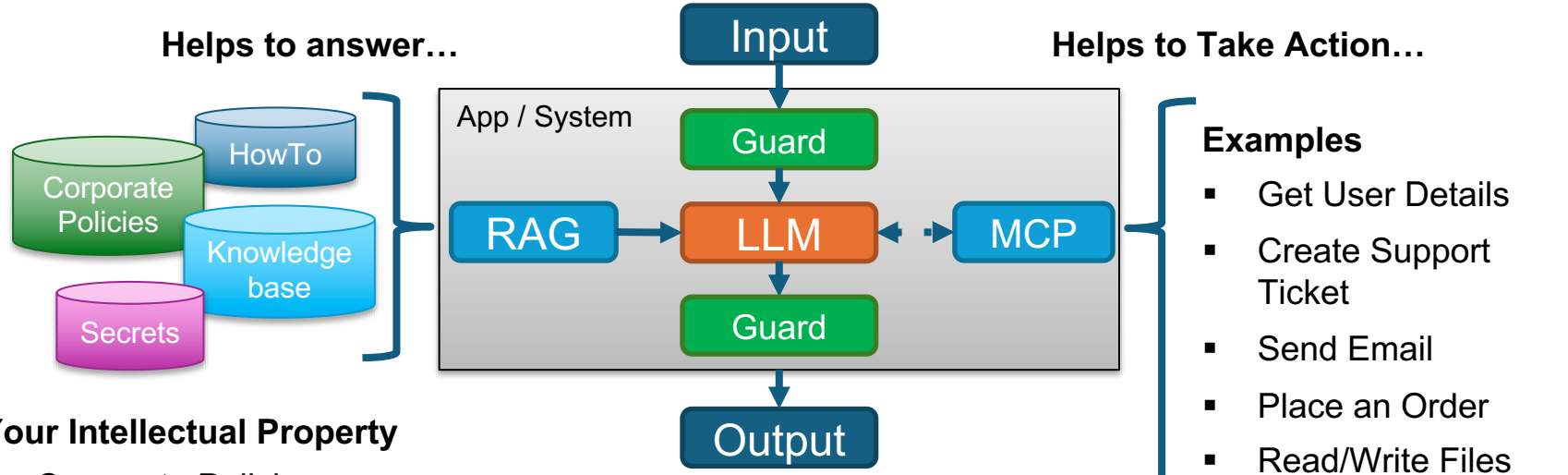
MCP

- Connects models, tools, and context
- Enables coordinated system communication
- The integration and orchestration layer

Nervous System

HCLSoftware

AI in Domino – Putting Together The Pieces



Your Intellectual Property

- Corporate Policies
- Knowledge Base
- Company Secretes
- Files, Wiki, Blogs, …

What you get with Domino IQ in 2026

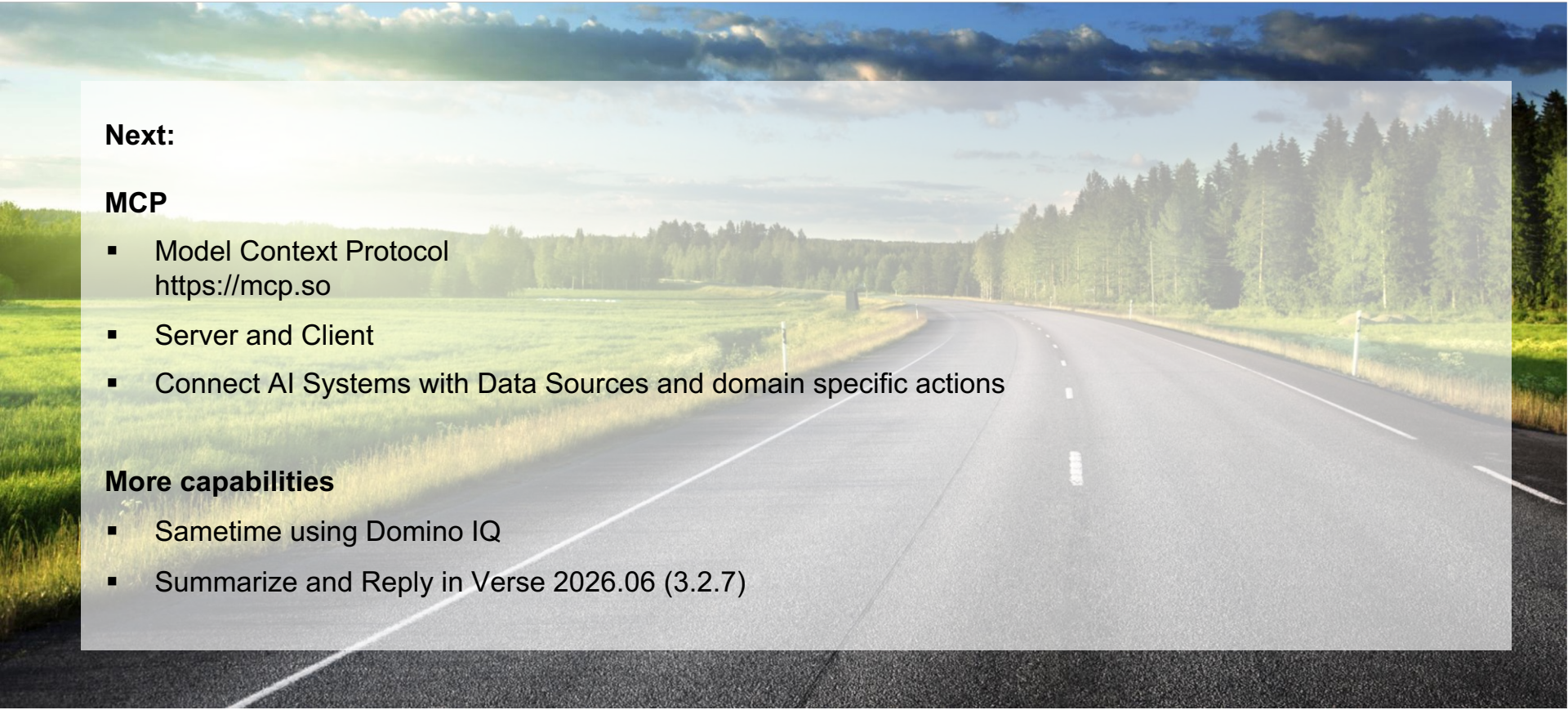
Available with HCL Domino 2026

- All local AI processing
  Fully integrated in Domino, but optional to use.
- Run a standard Large Language Models (LLM) of your choice
  - Optimized to your local language, geography or industry
  - Free from foreign interference
- New LotusScript Classes
- Retrieval-Augmented Generation AI
- Use domain specific knowledge from (configurable) Domino databases to generated answer
- **Vector database** integrated in Domino
- File Parser optimization for PDF, Office, etc.

Out of Scope (for now):

- Work with Images or Videos
- Support for IBMi
- Non-Nvidia GPUs
- ARM Processor based servers/PCs

What's NEXT with Domino 2026.x (v14.5.1FP?) and beyond



Next:

MCP

- Model Context Protocol
  https://mcp.so
- Server and Client
- Connect AI Systems with Data Sources and domain specific actions

More capabilities

- Sametime using Domino IQ
- Summarize and Reply in Verse 2026.06 (3.2.7)

Domino IQ Architecture

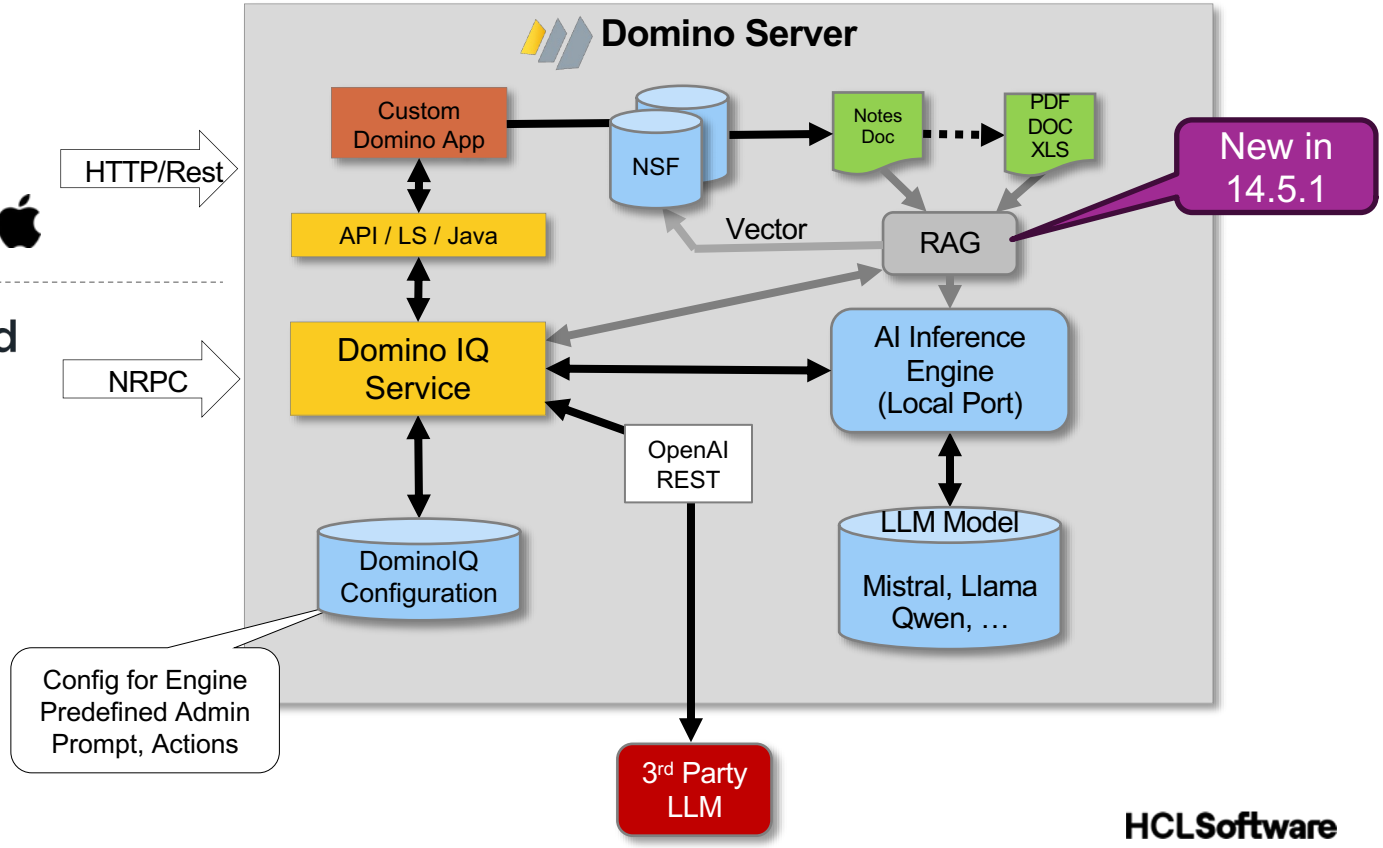
Client

Examples, but not limited to…

HCL Verse

HCL Nomad

HCL Notes



Hardware Requirements

Minimum hardware requirements for a Domino IQ server:

- 4 CPU Cores
- 8 GB of RAM
- NVIDIA GPU is mandatory
  - minimum of 8 GB of memory
  - compute capability (CC) of 5.2 or higher
  - Recommendation 8.0 and higher

Nvidia:

Cuda Cores = Number Parallel processors
Language Model size should fit into GPU Memory

Type	Nvidia Model	Est. Cost.
Entry	GTX 1660 Ti, RTX 3060	$120 - $475
Mid Range	RTX 3080, A4000	$575 - $1750
High End	RTX 6000, A5000	$1750 - $3500
Enterprise	A100, H100, V100	$12k - $35k
Insane	DGX A100 (=System)	"Contact us"

https://help.hcl-software.com/domino/14.5.1/admin/domino_iq_server.html#domino_iq_server__section_wgl_dwc_q2c

https://developer.nvidia.com/cuda/gpus

Our Hardware

Configuration:

- AWS Zone: eu-north-1b
- AWS Instance: g6e.2xlarge
- OS: Windows Server 2022 DATACENTER (OS Build 20348.3328, Ver 21H2)
- RAM: 64GB
- Processor: AMD EPYC 7R13 Processors - 2.65 GHz x 8
- Hard drives: 199GB & 419GB
- GPU: NVIDIA L40S - 48GB RAM (CUDA Version 12.8)
- Monthly Cost: ~$1,677 USD

Let’s See It In Action!

HCLSoftware