



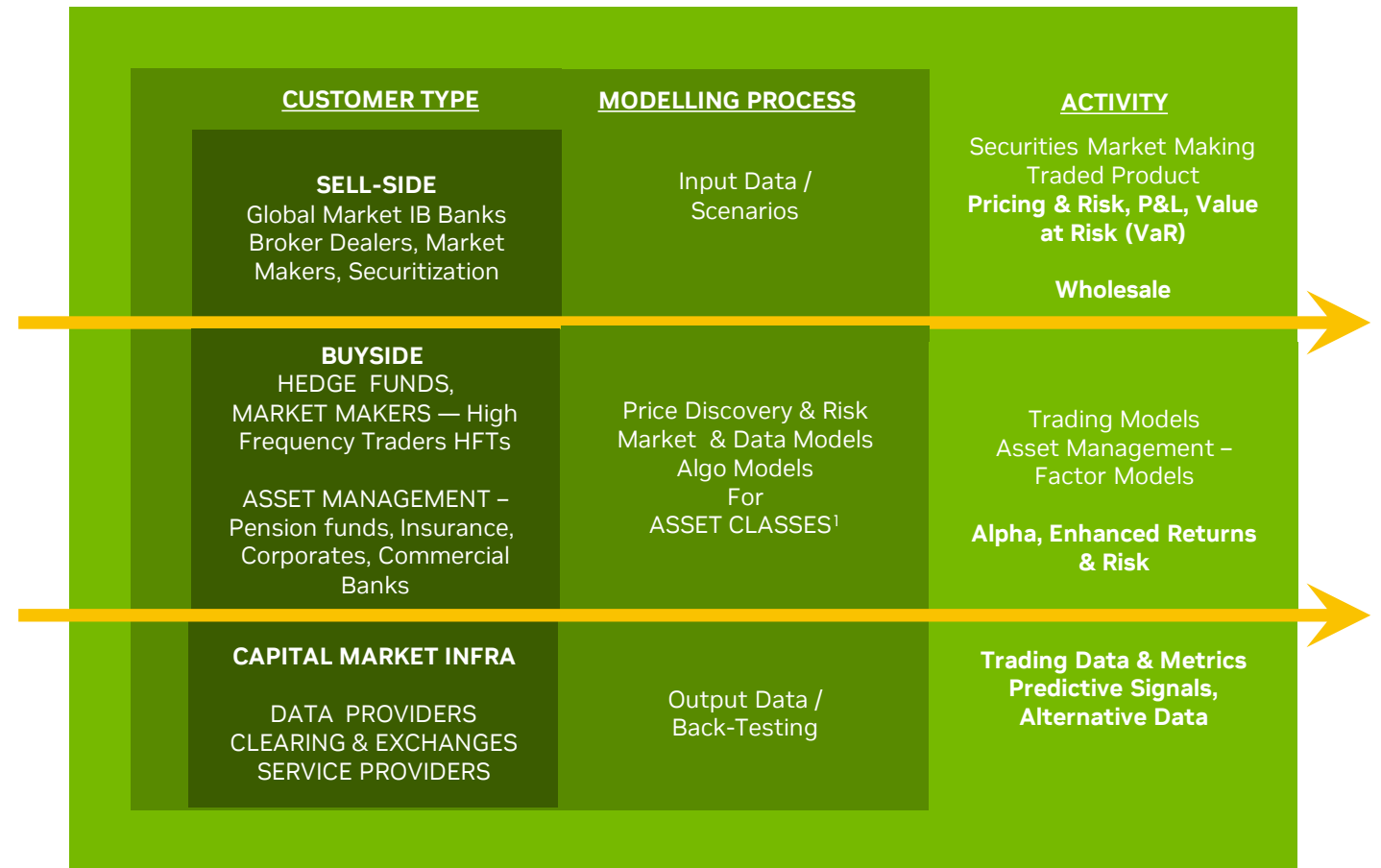
# Two birds, one (special) stone, Unifying FinHPC and AI

Accelerated & Converged AI +HPC solutions to unlock  
value in Trading, Banking and Financial Institutions

# Trading Participants in Capital Markets – HPC +AI needs

- Hedge Funds, High Frequency Trader HFTs, Market Makers
- Global Market Risk Investment Banks, Wholesale/Commercial & Retail Banking
- Asset managers – Pension funds, Insurance
- Capital Market Infra firms such as Data Providers and Exchanges in Front Office, Middle Office and Back Office areas

all of whom have extensive calculations and operations so can help your individual firm



## GPU's Used In All Areas in Front Office, Middle Office and Back Office

Price & Signal Discovery — using Scenario Gen, Simulation & Sampling  
 Accelerated Calcs for Trading & Risk — Pricing, VaR, Future Exposure, Intraday / T+1 calcs  
 Alternative & Unstructured Data for Alpha — NLP

Note<sup>1</sup> ASSET CLASSES - EQ, IR, FX, COMM, CR, Securitization products

# Financial Services Needs - Revenue, Operations & CX

AI Solutions - Seeking Higher AI ROI from Top to Bottom

## Financial Revenue



### Quant Finance

Market & Counterparty Risk  
Algo Trading & Backtesting



### Underwriting & Analytics

Underwriting with Alt Data  
Modelling & Back testing



### Alt Data

News & Sentiment NLP Data  
Satellite & Video Imagery  
Earnings transcriptions  
Synthetic data

## OPS



### KYC / AML / Fraud

Customer Onboarding  
ID Verification  
Cyber & Compliance



### Margining & Settlement

Collateral & Margin  
Exposure Analytics — Default,  
and Traded Underwriting



### Intelligent Automation

Intelligent Document  
Extraction - Contracts  
Backoffice & Claims

## Customer Experience CX/Internal Employee EX



### Conversational AI

ConvAI — ASR / TTS



### NLP Chatbot

Chatbot / Avatars,  
Transformer Models,  
Generative AI Q&A



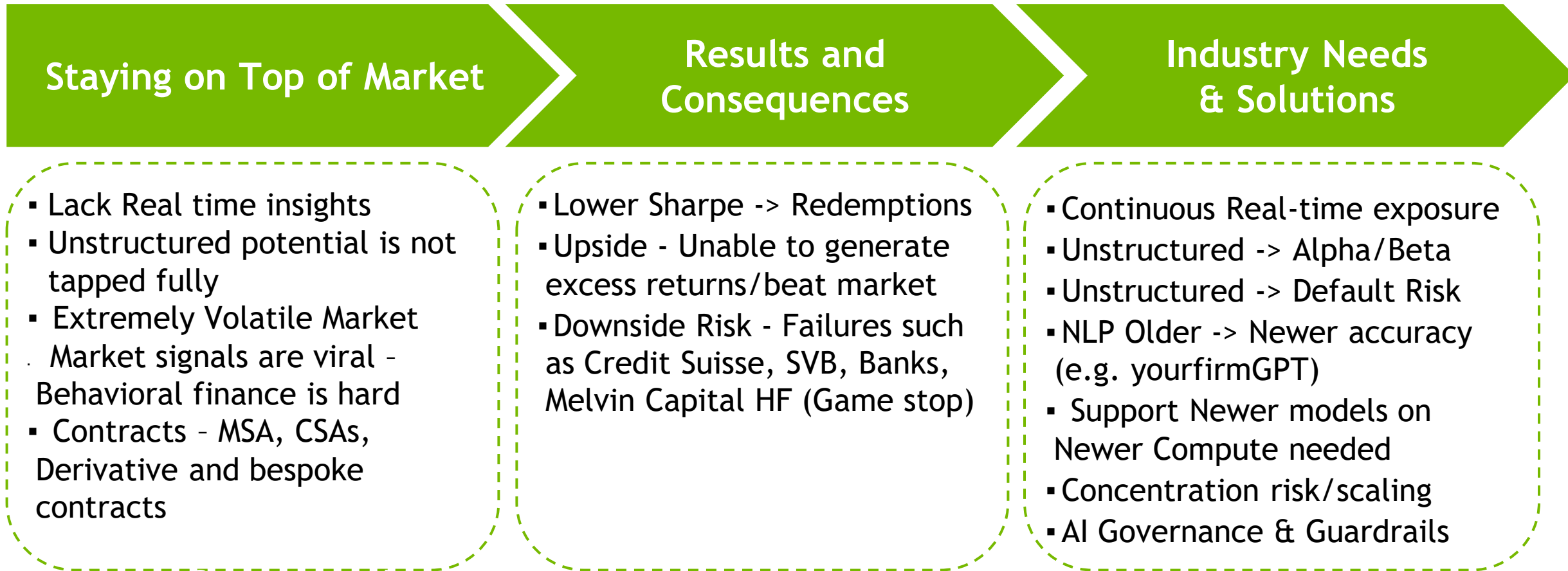
### Personalization

RoboAdvisory,  
Recommendation

# WHY NLP GENERATIVE AI/LLMs in Trading & Banking

Need for Real time calcs and AI Models

McKinsey reports - [Advanced analytics in asset management beyond the buzz](#), [Generative AI is here to change biz](#)



HPC+AI - Please see latest NVIDIA Accelerated compute in [STAC A2 Quant Finance results](#) and [supporting blog with quant finance+AI NLP convergence](#)

# NLP/Generative AI from Research to Production issues

Key Technical Challenges: Untapped Data in organizations and Modelling Challenges

HPC + Generative AI is revolutionizing Front/Middle/Back office where today's financial firm challenges are developing and backtesting trading AI strategies, calibrating models, and looking r causal risk scenario calculations

**Data Modelling challenges:** Organizations not relying and tapping on unstructured data where 70% or more of organizational information is unstructured alternative data, as in an IDC report.

**Generative AI Modelling challenges:** *Lack accuracy, full of hallucinations, high latency (not low) and costs high (in order of priority)*

**FSI Vertical domain needs:** Need **high accuracy and reducing misleading predictions** ("hallucinations"). Techniques like Retrieval Augmented Generation (RAG) with vector embeddings, fine-tuned hybrid language models for domain, and temporal time-series semantic search in low latency maximizing tokens

**Solution:** Empower financial firms to optimize quant research, accelerate alpha returns, refine trading strategies, and streamline risk management by building a cost-effective "AI factory".

- Facilitating new world trading, causal time based LLM analytics and exposure analytics
- End audience Financial Institutions for Traders, Lines of Business and Board of Institutions providing **the high accuracy, reduced hallucinations, reduced latency, scaling, reduced AI factory cost via hybrid RAG** for use cases.

# Satisfy GenAI/DL, Quant Finance, HPC and ETL/ML

Not only LLMs use with 1. Gen AI/DL Neural Nets 2. Quant Finance/HPC and 3. ETL/ML

## AI (Neural nets) - LLM/GenAI

- AI Unstructured Data using NLP with LLMs, Other Systematic Trading Algos
- Framework - PyTorch/TensorFlow, [NVIDIA NIM](#) e.g. RAG based retrieval, NVIDIA AI Enterprise

## Quant Finance / HPC

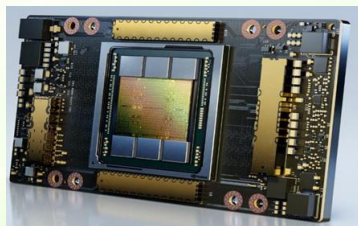
- Pricing, Risk (MC Sim, Margin, FRTB, CVA, SIMM, XVA) & Back testing
- Framework - CUDA C/C++, Parallel Algorithms C++, NVIDIA Accelerated Python - RAPIDS, Open ACC

## Data Processing - ETL/ML

- Feature Engineering, Data Prep, & Data Science (e.g., XGBOOST)
- Framework - NVIDIA Accelerated Python on Steroids - RAPIDS, Spark on GPU

## Full Stack plus Accelerated Software

NVIDIA A100  
Tensor Core  
GPU for  
LLMs and  
inference



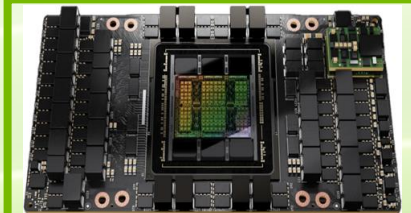
20X Volta Performance

NVIDIA L40S  
Tensor Core  
GPU for AI  
Inferencing



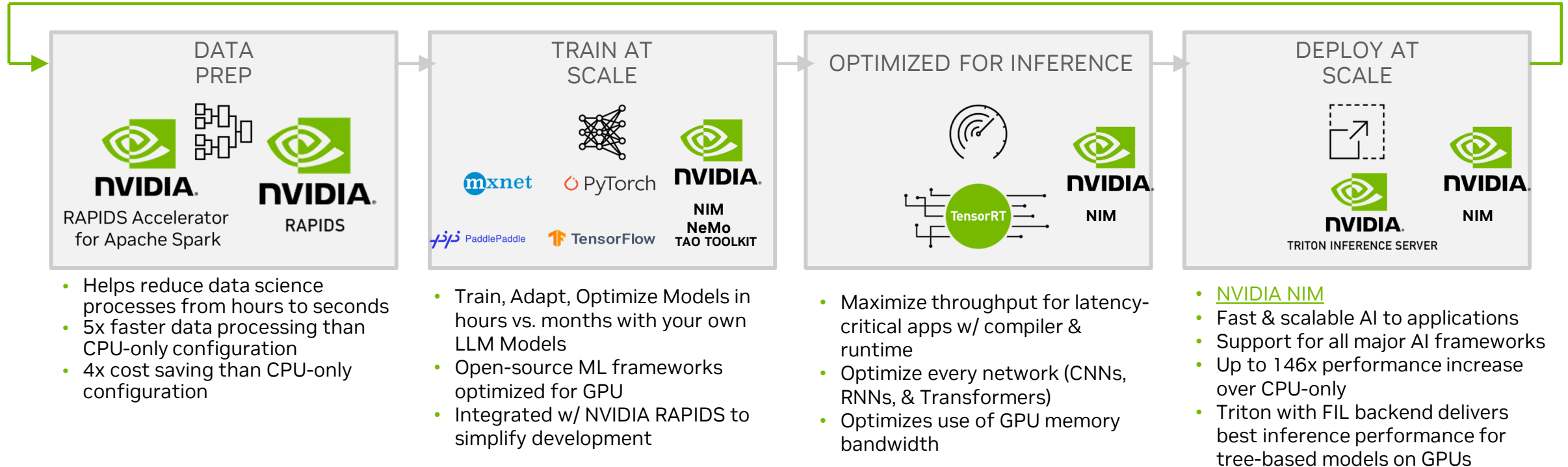
1.5X Inference Performance  
than the HGX A100

NVIDIA H100  
Tensor Core  
GPU for  
LLM Training/  
Deployment



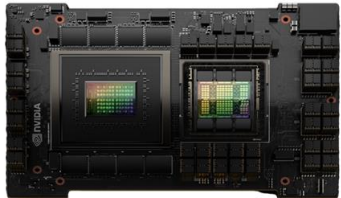
10X Faster Inference  
than the previous Gen A100s

# Converged HPC + AI: Needs Heterogenous Compute with Software (Software is key in Enterprises)




Full Stack plus Accelerated Software – NVIDIA AI Enterprise and NVIDIA NIM are also optimized for latest architectures

NVIDIA Grace Hopper GH100 CPU+ GPU for LLMs and inference



7X Faster data transfers and queries than PCIe Gen 5

NVIDIA Grace Hopper GH200 CPU+ GPU for LLMs and inference



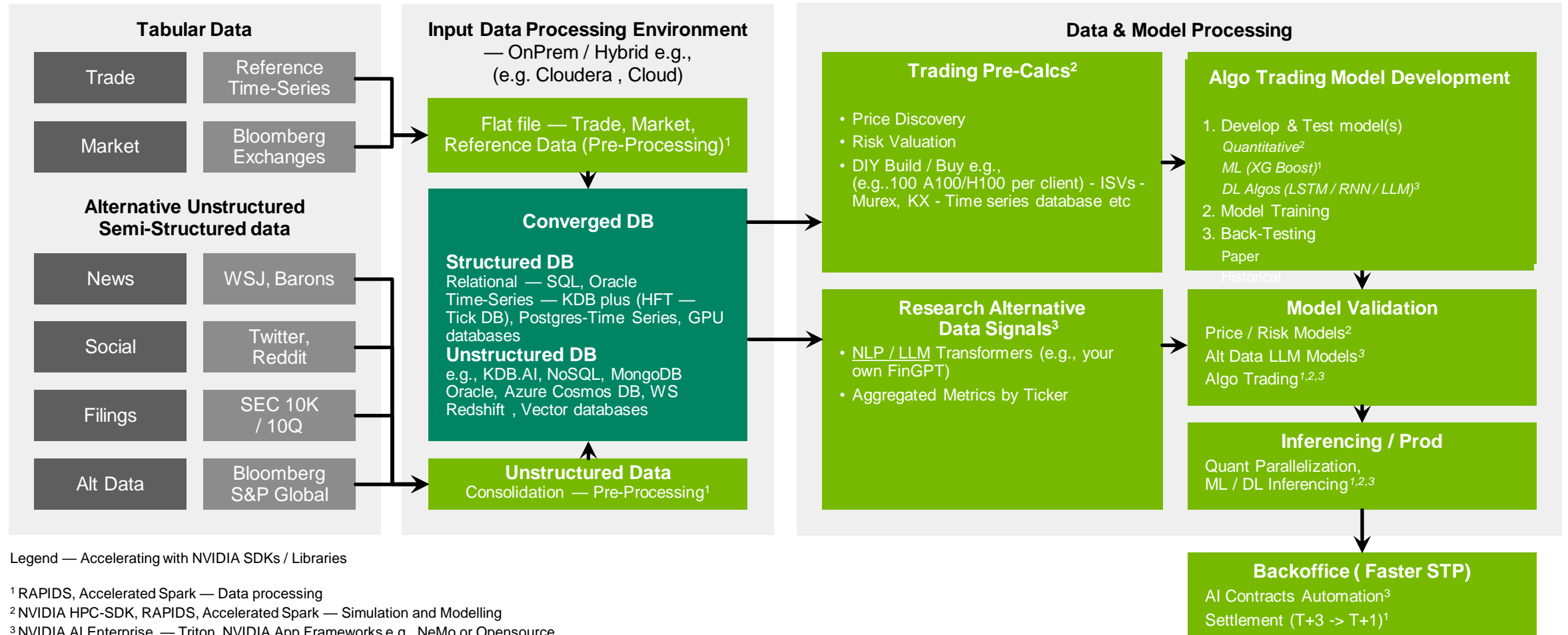
3.5x more memory and 3x more bandwidth than the current generation

- Latest from NVIDIA GTC 2024**
- NVIDIA Blackwell Architecture - <https://nvidianews.nvidia.com/news/nvidia-blackwell-platform-arrives-to-power-a-new-era-of-computing>
- [NVIDIA GB200 Grace Blackwell Superchip](#)

# Functional workflow Unifying HPC + AI

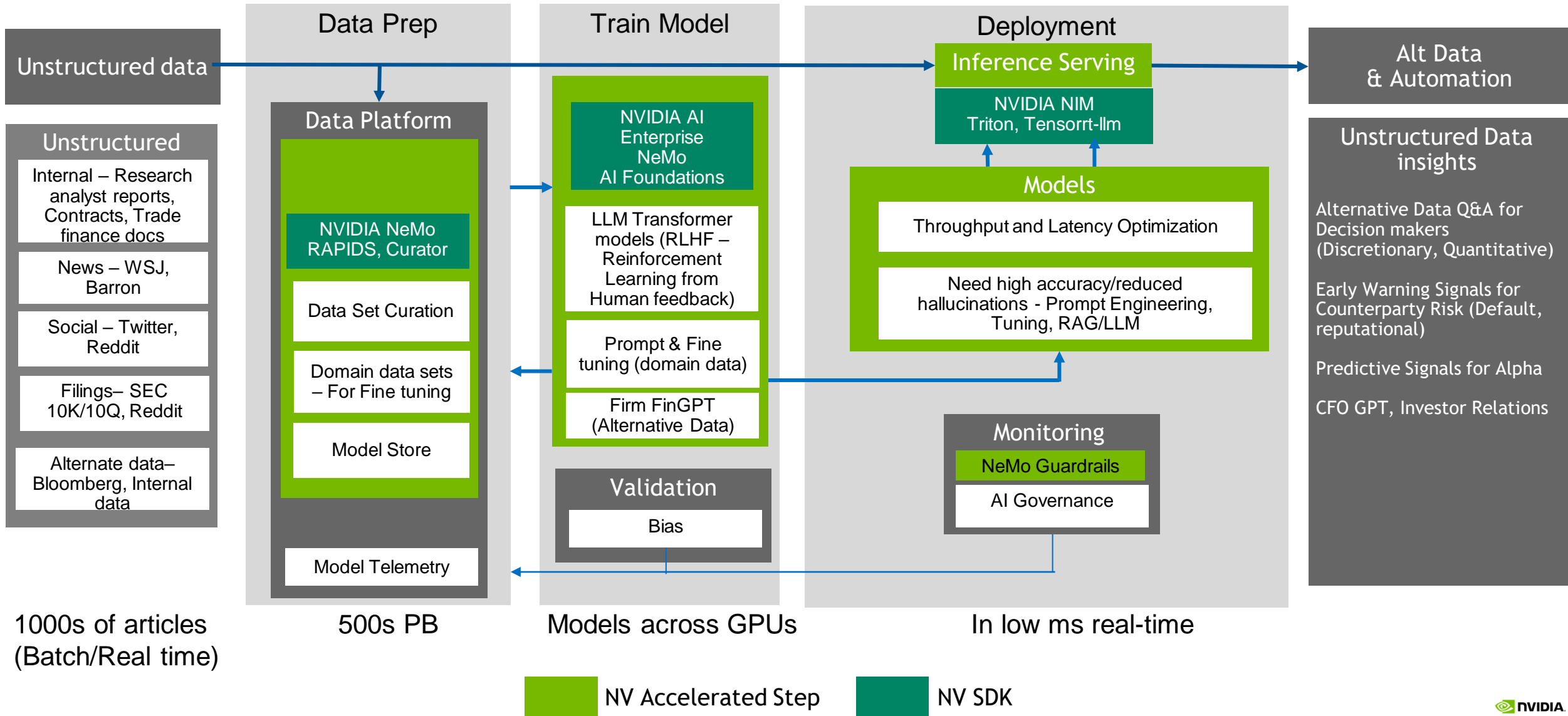
## End-to-end Data science pipeline

**Building Adaptive Trading Strategies — Real-time information analysis (e.g. information retrieval on contracts) with Early Warning Indicators, Alpha, Beat Market Returns – Enhanced Beta**



# End-End AI Data workflow example for Trading & Banking

Alpha Generation using Alternative Data (Gen AI LLM) and CFO GPT, Investor Relations Alpha Research





# Converged HPC+ AI use cases in action

KX/Murex

# Trade Ideation, Trade Execution and Risk Management

## Data Preparation

### Data Sources

#### Structured

- Market Data
- Order/Execution Data
- Security Master Data
- Corporate Action Data

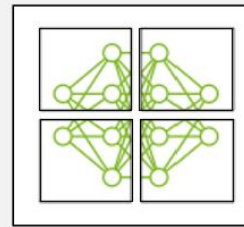
#### Unstructured

- SEC Filings
- Analyst Reports
- Company Transcripts
- News

## Model Training

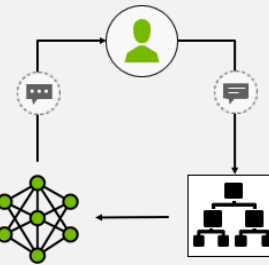


Hybrid RAG  
Fine tuned  
Model



NeMo Customizer

Vector  
Database



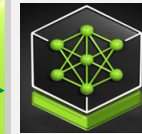
NeMo Retriever

Timeseries  
Database

## Inferencing Deployment



Trade Ideation & Research NIM



Trade Execution NIM



Risk Management NIM

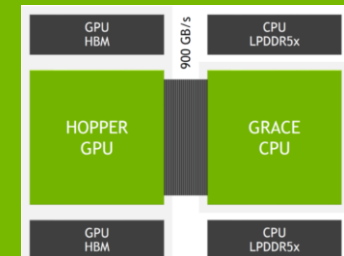
NVIDIA AI Enterprise

NVIDIA

RAPIDS NeMo



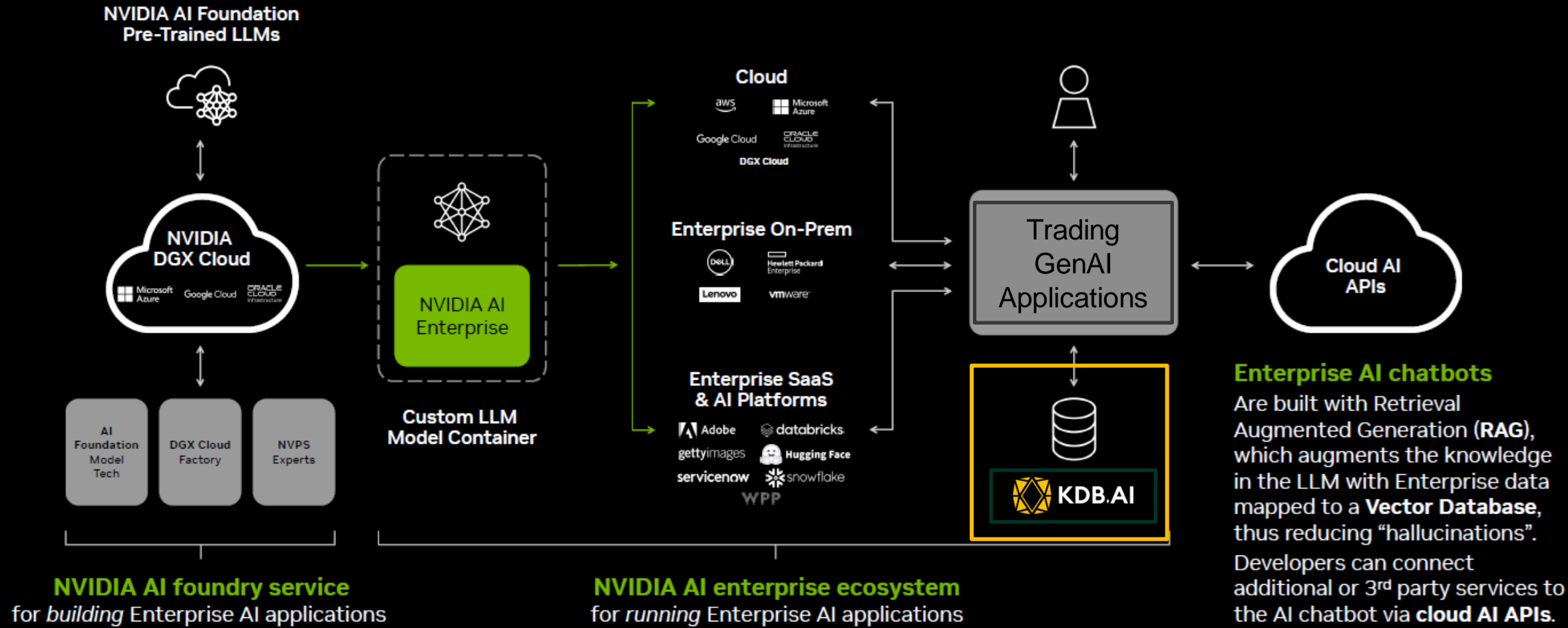
KDB.AI





# Powering the AI Industrial Revolution

## Building and Running Enterprise Gen AI Applications



**Enterprise AI chatbots** Are built with Retrieval Augmented Generation (RAG), which augments the knowledge in the LLM with Enterprise data mapped to a **Vector Database**, thus reducing “hallucinations”. Developers can connect additional or 3<sup>rd</sup> party services to the AI chatbot via **cloud AI APIs**.

# HPC Quant Finance - XVA Counterparty Credit Risk on Grace Hopper

RESULTS OF NVIDIA WORKING WITH MUREX R&D WITH THE NEW NVIDIA GRACE HOPPER SUPER CHIP

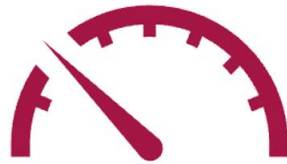


~5x in XVA Speedup and power

PERFORMANCE



POWER CONSUMPTION



HARDWARE COST



XVA Evaluation Comparison

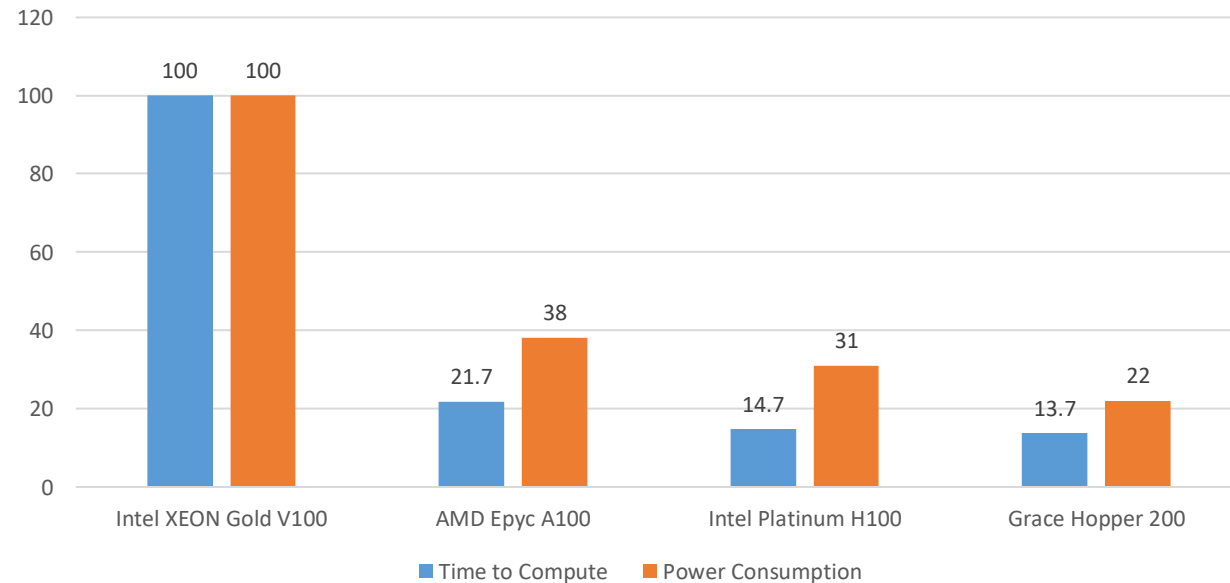


Figure 1- XVA Counterparty Credit Risk Benefits on Heterogenous Grace Hopper (CPU/GPU)

# TCO, ROI and Productivity in HPC+AI use cases

Resources and NVIDIA GTC Customer stories

## Capital One Example



**Before GPUs:**

26-Hour  
Run Time



**100x**  
Training Time  
Improvement

**98%**  
Decrease in  
total costs

**RAPIDS** | **dask**



**After GPUs:**

15-Minute  
Run Time

## Generic Savings Calculator Per Year

No. of Data Science models	50	400	5000
No. of AI models (e.g., Small, Medium, Large LLMs)	5	40	500
<b>Estimated Cost Savings</b>	<b>\$0.5 M</b>	<b>\$4-5 M</b>	<b>\$65M</b>

### Reduce TCO by 71%

- Reduce No of Servers (e.g., CVA Credit valuation) — 100 to 4 servers and 7X speeds up on Grace Hopper
- Baseline cost — Before 100 Servers @15,000 = 1,500,000
- After - 4 Servers @ 443,929
- Reduced TCO saving 71% **(1,056,071)** for one use case
- Avoid costly Maintenance, Reduced Admin server costs

### Increase ROI - Do More with the Hardware

- Do Generative AI + Data Science + Quant Finance
- Multiple accelerated workloads > 20x the Speeds for Quant Finance and Data Science
- **Leveraging [Spark 3 and NVIDIA's GPUs to Reduce Cloud Cost](#) by up to 70% for [Big Data Pipelines](#) - Paypal**

### Productivity

#### Choice - Language, Framework

- NVIDIA HPC SDK, C/C++
- CUDA, ISO C++ parallelism,
- RAPIDS in Python, NVIDIA SPARK on GPU, RAPIDS / DASK with single line code changes
- NVIDIA AI Enterprise (DL Model and Framework)

#### Software ISV Partners

- [Example ISVs -KX, Murex](#)
- [Algo Trading, CFO GPT, KYC/AML/Fraud, NVIDIA AI](#)

#### Other Resources

- STAC [A2 blog](#), STAC [A3 blog](#)
- [Book limit order prediction blog](#)
- JPMC [Risk Calculations](#)

#### NVIDIA GTC Sessions:

- [Capital One](#)
- [Cohen & Steers](#)
- [Wells Fargo](#)
- [Citibank NN For Exotic CBOE](#)
- [Bank of America](#)

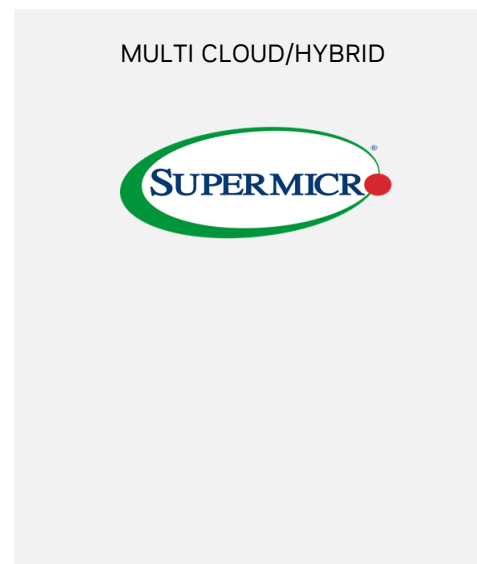
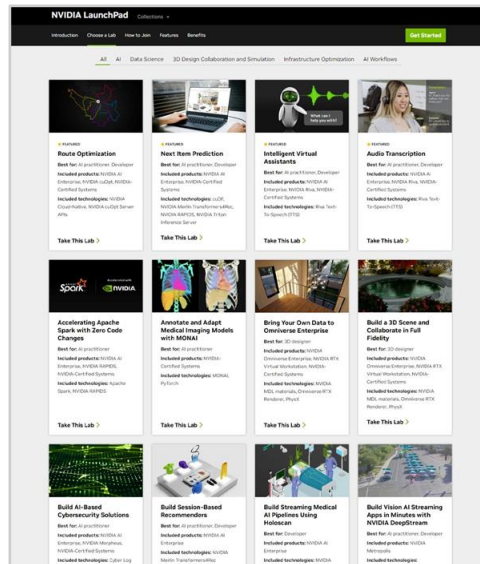
# Experience HPC +AI Convergence benefits

Accelerated & converged AI/HPC in Capital Markets on Heterogenous Compute with Super Micro

Test on NVIDIA Launchpad

OR

Test on Your Own Cloud



## What would a journey look like?

1. Run a workshop with the relevant team and validate the problem statement / use case with NVIDIA financial services team and a preferred partner (ISV or GSI consulting partner) on an agreed use case tied to workload (4-8-hour workshop)
2. Run Agreed Use case with workload type — Quant Finance or Generative AI workloads on accelerated compute on Super Micro for your use case (2-4 weeks)
3. Collaborate on shipping into production (2-3 weeks)
4. Document findings and present to key stakeholders (1-2 weeks)

Visit [nvidia.com/launchpad](https://nvidia.com/launchpad)