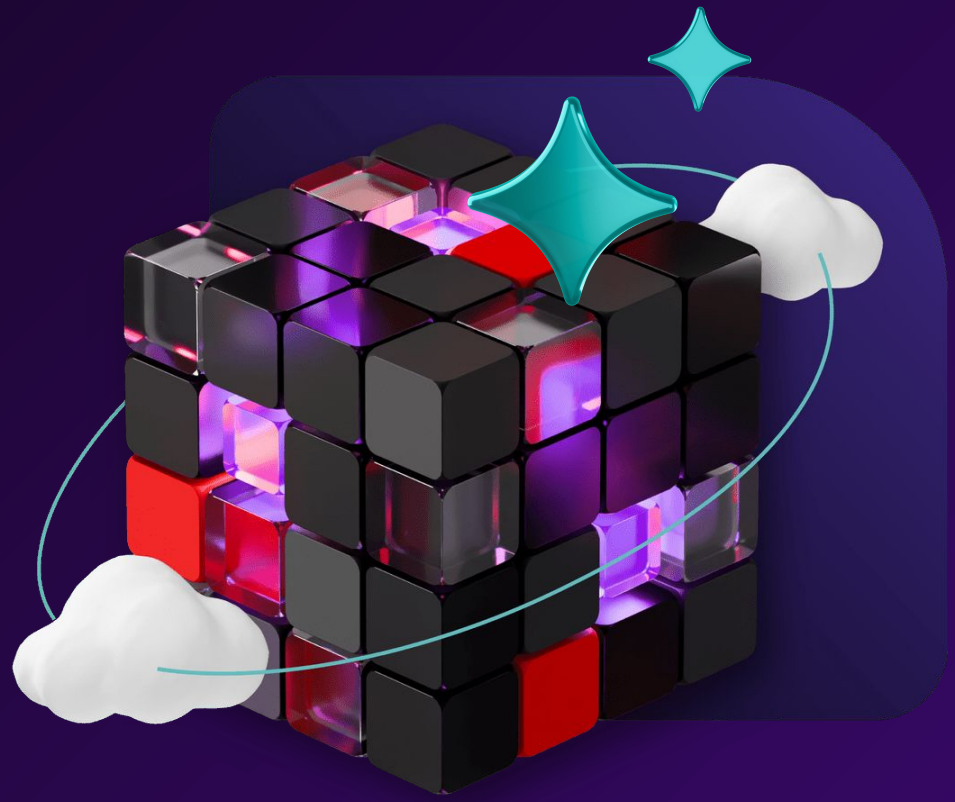
Last update: Oct, 2026

Red Hat AI Factory with NVIDIA

Why and why now?



Andreas Bergqvist Red Hat
Martin Westblom NVIDIA



Red Hat | NVIDIA

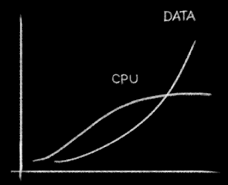

ABB
abbvie
accenture
Adobe
AETHERFLUX
AHEAD
Ai2
AiVRES
Akamai
AT&T
Alibaba Cloud
ALTIMETER
aws
AMP COALITION
ANTHROPIC
ASRock Rack
ASUS
ATREIDES
AXIOM SPACE
baseten
biofy
Black Forest Labs
BLUE ORIGIN
BNY
BOSCH
BYD
ByteDance
cadence
Caltech
Capital One
Carnegie Mellon University
CATERPILLAR
CISCO
CLOUDIAN
cohere
Cognition
COMCAST
COMPAL
CONVICTION
CoolIT systems
CoreWeave
coupang
CROWDSTRIKE
Crusoe
CURSOR
DASSAULT SYSTEMES
databricks
ddn
deepinfra
DELL Technologies
Deloitte.
DELTA
Disney Research
EATON
Edison
EQUINIX
ETH zürich
Everpure
EY
Fireworks AI
firmus
Ford
FOXCONN
GE Healthcare
GE VERNOVA
GEELY
GEICO
gm
Georgia Institute of Technology
GIGABYTE
GMI
Google Cloud
HITACHI
Honeywell
hp
HPE
HYUNDAI
hyve solutions
IBM
Ingrasys
Instacart
Inventec
IQVIA
IREN
Jacobs
JLR
JOHNS HOPKINS UNIVERSITY
Johnson&Johnson
KAIST
KLA
L'OREAL
Lambda
LangChain
Lenovo
Lilly
LOCKHEED MARTIN
Los Alamos NATIONAL LABORATORY
MEDBOT
Medtronic
Meta
micron
Microsoft
Mistral AI
MIT
MiTAC Computing
Morgan Stanley
msi
Nasdaq
NATIONAL TSING HUA UNIVERSITY
NEBIUS
NetApp
NISSAN MOTOR CORPORATION
NOKIA
NSCALE
NUTANIX
OpenAI
OpenClaw
OpenEvidence
ORACLE CLOUD Infrastructure
Palantir
paloalto
PayPal
PEGATRON
PEPSICO
perplexity
PHYSICS X
Pinterest
planet.
PNY
PRIME Intellect
PRINCETON UNIVERSITY
ptc
QCT
RADIANT
RAFAY
Red Hat
REDWIRE
Reflection
runway
salesforce
SAMSUNG
SJSU
SAP
Schneider Electric
서울대학교 SEOUL NATIONAL UNIVERSITY
SEQUOIA
servicenow
Shopee
shopify
SUPERMICRO
SIEMENS
SK hynix
SKILD AI
SONY
Stanford University
Starcloud
stryker
switch
SYNOPSYS
T-Mobile
TCS
TESLA
TACC
大阪大学 OSAKA UNIVERSITY
ThermoFisher SCIENTIFIC
THINKING MACHINES
together.ai
Tripadvisor
tsmc
Uber
UC Berkeley
UCLA
UNIVERSITY OF CAMBRIDGE
UNIVERSITY of WASHINGTON
VAST
VERTIV.
VOLKSWAGEN
VULTR
waabi
Walmart
WAYMO
wiwynn x wistron
WEKA
WATER DARSO
World Labs
World Wide Technology
xAI
XIAOMI
YOTTA
Yum!
Red Hat
NVIDIA

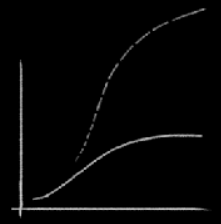
1964
IBM S/360





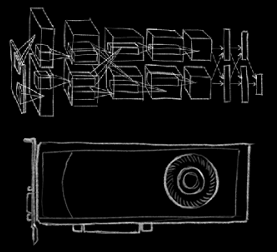
1995
WINDOWS 95
PENTIUM



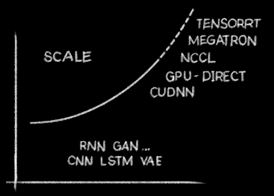


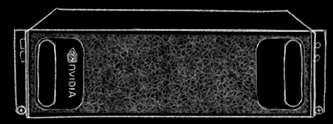
ACCELERATED COMPUTING
1993 UDA → 2003 CG → 2006 CUDA

2012
ALEXNET
"FIRST CONTACT"

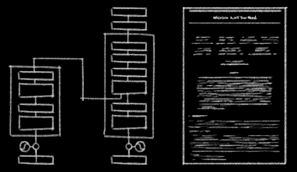


DOUBLING
EVERY 6 MONTHS





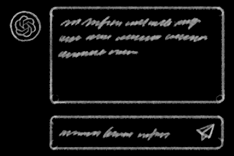
2017
TRANSFORMER



LEARN EVERYTHING

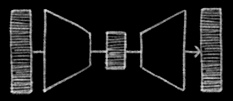
PROTEIN
LANGUAGE
SOUND
PHYSICS
VIDEO
3D
MANIPULATION
IMAGES
GESTURE

2022
OPEN AI CHATGPT

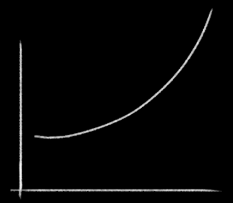


FINE TUNING
GUARDRAILING
ALIGNMENT
PROMPT ENGINEERING
RAG
VECTOR DB
MULTI-MODAL
COT & TOT
AGENTS

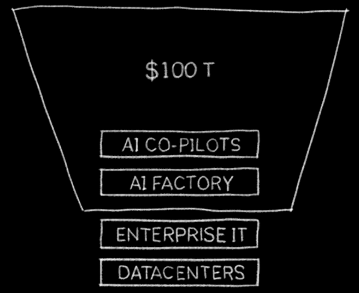
GENERATIVE AI



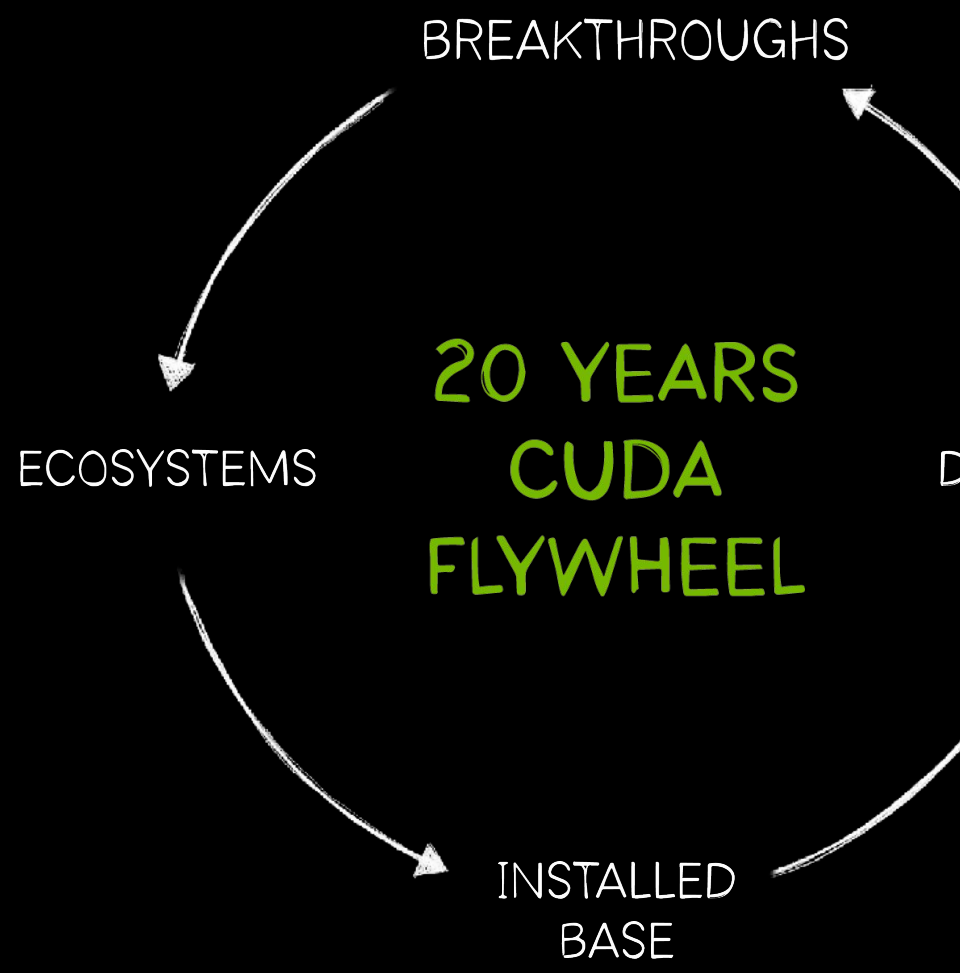
STARTUPS



A NEW
INDUSTRIAL
REVOLUTION



Red Hat | NVIDIA



Red Hat | NVIDIA

CUDA-X FOR EVERY INDUSTRY



WARP
PHYSICS

cuDF
cuML
DATA SCIENCE
AND PROCESSING

cuDSS
cuSPARSE
cuFFT
AMGX
COMPUTER AIDED
ENGINEERING

TRT-LLM
MEGATRON
NCCL
cuDNN
CUTLASS
cuBLAS
DEEP
LEARNING

cuEQUIVARIANCE
cuTENSOR
QUANTUM
CHEMISTRY

cuQUANTUM
CUDA-Q
QUANTUM
COMPUTING

EARTH-2
WEATHER
ANALYTICS

MONAI
MEDICAL
IMAGING

PARABRICKS
GENE
SEQUENCING

cuOPT
DECISION
OPTIMIZATION

AERIAL
SIONNA
5G/6G
SIGNAL
PROCESSING

cuLITHO
COMPUTATIONAL
LITHOGRAPHY

cuPYNUMERIC
NUMERICAL
COMPUTING

Red Hat | NVIDIA

AI is a Five Layer Cake

APPLICATIONS

MODELS

INFRASTRUCTURE

GPUS

ENERGY

CHATBOTS

DIGITAL BIOLOGY

ROBOTAXI

ENTERPRISE AI AGENTS

SCIENCE

ROBOTICS

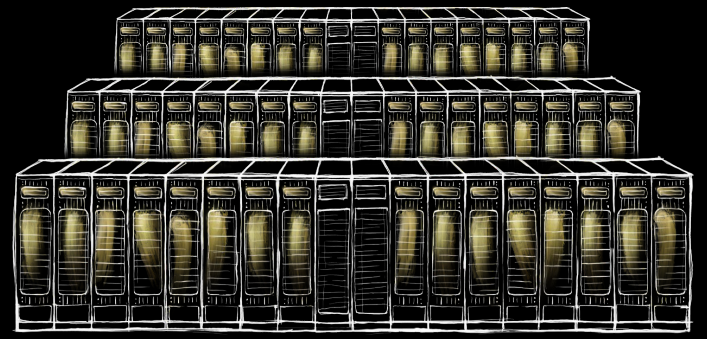
MANUFACTURING

AI CODER

LLM | VLM | VLA | MMLLM | GPT

DM | GNN | MOE | SSM | LBM

AI FACTORIES



Red Hat | NVIDIA





Red Hat

Storage

Server

Red Hat AI Factory

Llama3

Nemo Retriever

Maxine Eye Contact

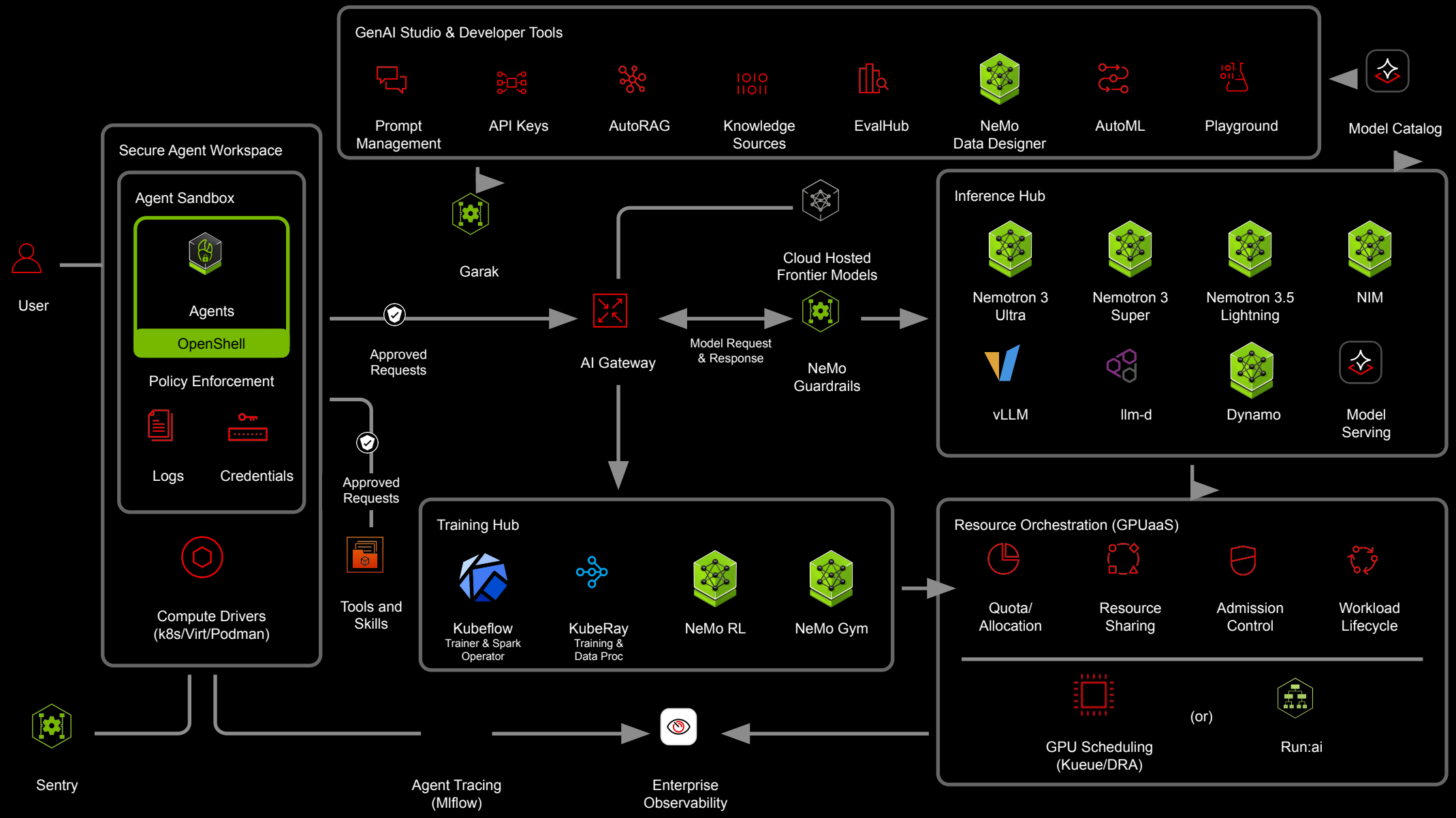
RIVA ASR

Audio2-Gesture

Parakeet

Red Hat | NVIDIA

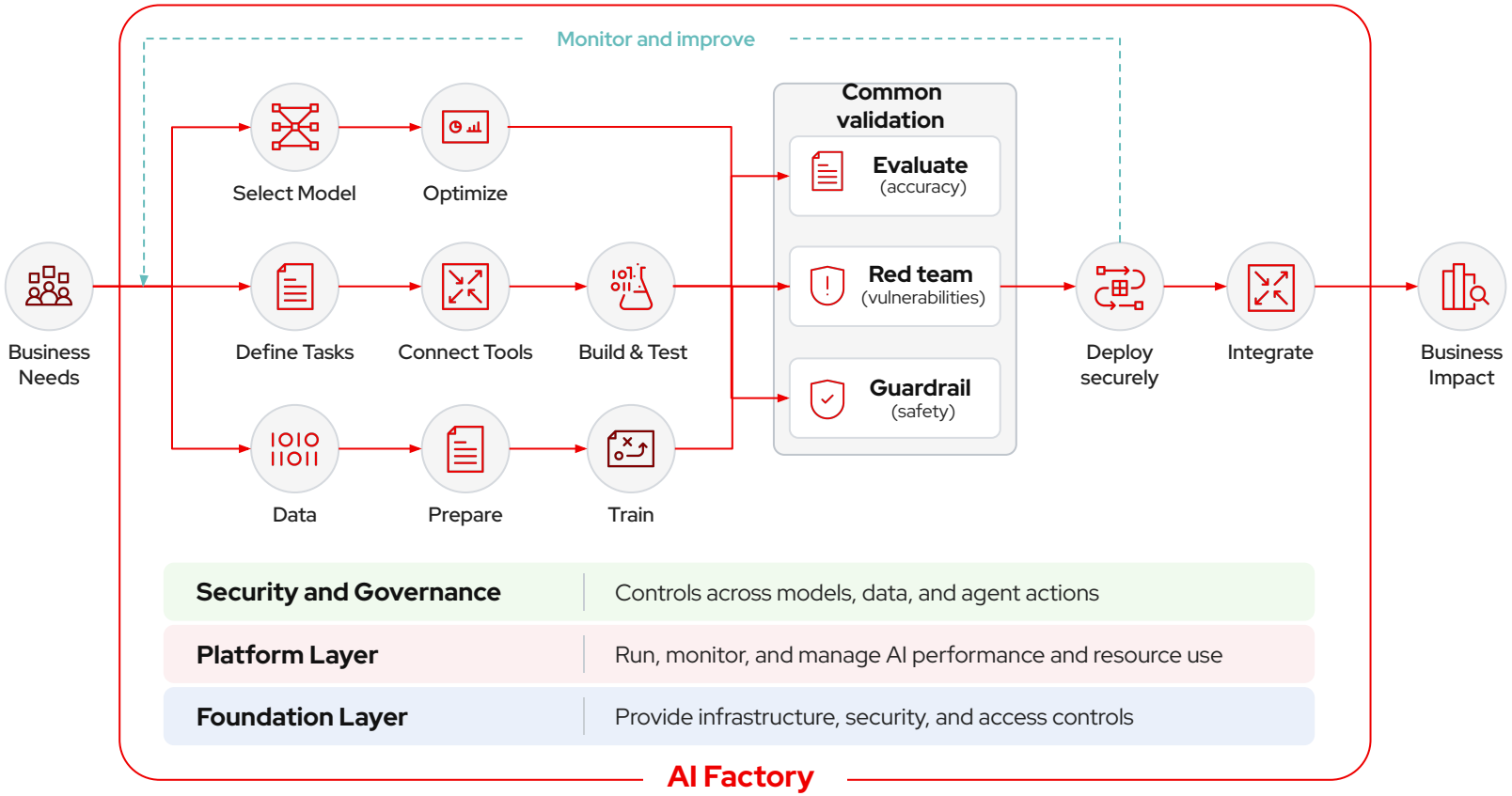
Technical architecture



NVIDIA

But why does AI need a factory approach?

Because AI is happening at scale!



Integrate AI into your applications with **full control** over data, models, and deployment

Standardize how teams build and operate AI through a common platform to **reduce complexity**

Operate AI across environments on a consistent foundation without re-architecting to **optimize cost**

Unlock innovation at scale using Red Hat AI Factory with NVIDIA

Red Hat | NVIDIA

Better Together: Red Hat AI Factory with NVIDIA

NVIDIA drives AI innovation and Red Hat delivers it reliably at enterprise scale

Together, we deliver:

Co-engineered for Enterprise AI

- **Jointly engineered** platform built through **ongoing collaboration**
- **Aligned to evolving AI** requirements
- **Delivers a curated, certified stack,** enabling customers to assemble with confidence and scale diverse AI workloads

Production-Ready out-of-the-box

- **Validated Blueprints and Quickstarts** provide a tested foundation for production
- **Eliminate the need to design architectures** from scratch
- **DAY 0 support**

Consistent AI Operations

- **Single platform** across data center, cloud, and edge
- **Consistent security, governance, and control** across environments

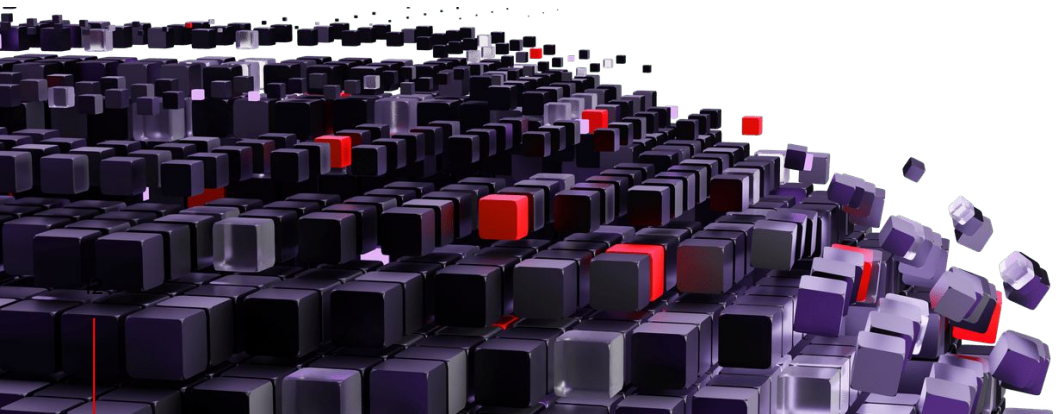
Predictable, Scalable Economics

- Bundled offering with **node-based pricing**
- **More cost-efficient** than traditional GPU-based pricing models
- **Predictable spend** with a **consistent way** to buy and scale AI

Red Hat | NVIDIA

Turn AI investment into business value

A unified foundation for production AI across your hybrid cloud.



Accelerate time to value

Bring AI use cases into production faster with simplified workflows, validated blueprints, and optimized models.

Improve cost efficiency

Make better use of AI infrastructure with optimized inference and shared GPU resources, while tracking usage and spending.

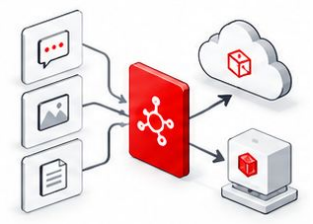
Reduce risks as AI scales

Help protect sensitive data and govern AI use with access controls, runtime safeguards, and monitoring across hybrid environments.

Red Hat | NVIDIA

AI Use Cases: From Platform to Production

Explore how AI can support your business



Inference Hub/MaaS

- Unified gateway for cloud or self-hosted AI models
- Control access, set usage limits, and track spending
- Share GPU resources and monitor model services

Learn more about MaaS



Post-Training

- Customize AI models through training
- Prepare or generate training data; track experiments
- Control access and resource usage across training workloads



Secure Agent Workspace

- Run AI agents in isolated user workspaces
- Control user access and agents' network access
- Protect credentials and enterprise access

Learn more about secure agent workspace



Enterprise Knowledge

- Find answers across company and external sources
- Get cited answers and research reports
- Reduce information gathering and review time

Learn more about enterprise research

Learn more about enterprise RAG



Video Search & Summarization

- Find events in recorded or live video
- Ask questions and get automatic summaries
- Reduce manual video review

Learn more about video search and summarization

Red Hat | NVIDIA

Customer Success

Red Hat AI Factory with NVIDIA in production

telenor

Challenge

Telenor needed to grow AI services for sensitive data while meeting rules on where that data is processed. Customers also needed freedom to choose their models and tools.

Solution

Telenor combined Red Hat OpenShift AI and OpenShift Platform Plus on NVIDIA infrastructure. Red Hat automation and consulting help teams deploy and operate the service consistently.

Results

- **Local data control:** Sensitive enterprise data is processed in Norway, supporting requirements on where data is processed.
- **Choice and flexibility:** Customers can choose models and tools, and run AI alongside existing applications.
- **Services in production:** Two sites in Norway serve internal and external customers with sensitive and critical workloads.



0.lat

Challenge

Industrial applications need fast AI responses. Running AI far from where data is created adds delays and makes time-sensitive tasks harder to support.

Solution

Zero Latency adopted Red Hat AI Factory with NVIDIA to run AI near customers on NVIDIA Blackwell GPUs. Red Hat management tools provide one consistent way to manage GPUs across locations.

Results

- **On-demand GPU access:** Customers can use AI compute without building private GPU infrastructure.
- **Simpler operations:** One workflow manages GPU resources across locations, reducing the complexity of managing each site separately.
- **Live services:** Initial datacenters are operational. Expansion to hundreds of locations is planned.

How to get Red Hat AI Factory with NVIDIA

Delivered as a validated system through a trusted partner ecosystem

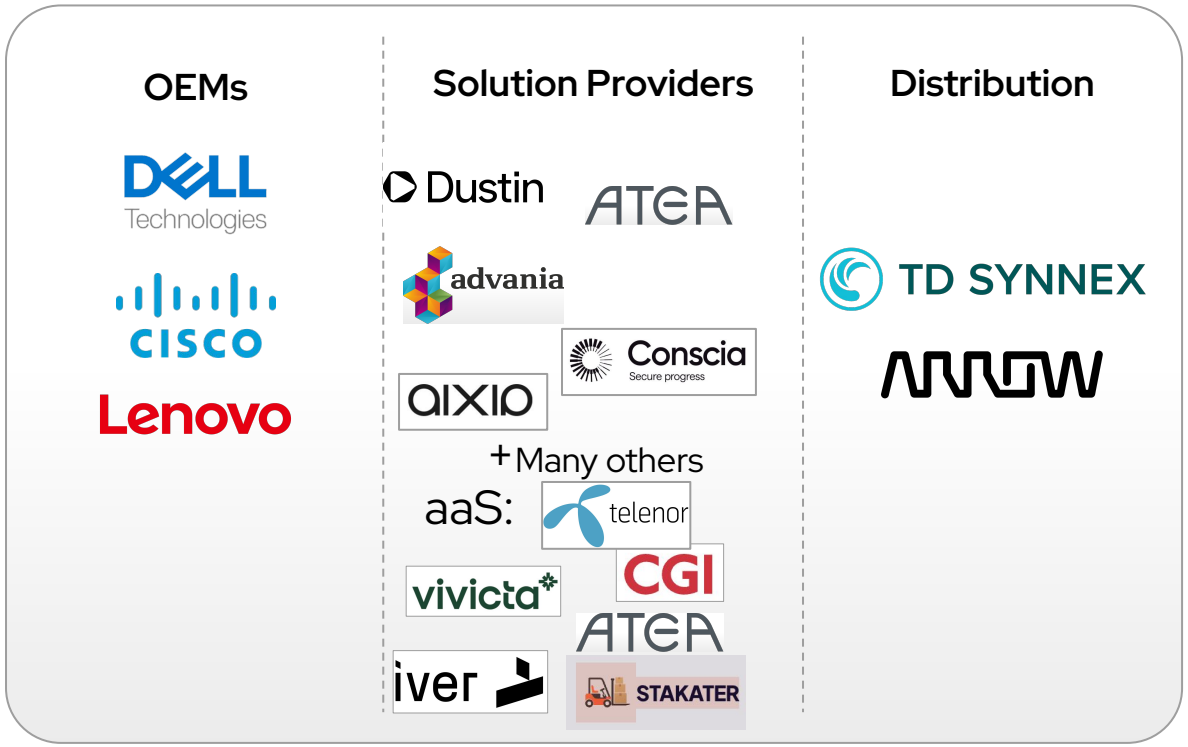
Delivered Through Partners

- Purchased and delivered via certified partners
- Partners handle quoting, and procurement
- Jointly supported by Red Hat and NVIDIA
- Partner services for deployment and development

✓ Available now on **NVIDIA Blackwell** systems

→ Day 0 support for upcoming NVIDIA platforms

Value Chain



***Note**: To deliver this joint offer seamlessly, RedHat and NVIDIA may share your order details with one another to coordinate service and support.*

Red Hat | NVIDIA

Bring your next AI use case to production

Next Steps

- Choose a priority use case and define the outcome you want to achieve.
- Engage your preferred OEM, VAR, or Distributor and Red Hat Services to assess requirements and shape the solution.
- Define deployment milestones, success measures, and support for ongoing operations.

Let’s schedule a focused discussion on your priority use case.

You’ll find us in the Expert Zone in the Exihibtion area

Thank you!

