Red Hat Summit

Connect

Build an AI Private/Sovereign Cloud with Kyndryl and Red Hat OpenShift AI

Pietro Iannucci | Principal Architect, Director Private Cloud architectures, Kyndryl

kyndryl | Red Hat

Agenda

- Market Insight, needs, challenges and opportunities
- How Red Hat AI can help
- How Kyndryl can help

kyndryl

AI/GenAI market perspective

Transition to strategic AI investments impact

- Applications
- Platforms
- Data
- Infrastructure

AI is the new strategic workload

Increasing infrastructure complexity

App ecosystem shift to Agentic AI

$749B

anticipated spend on AI technology by 2028[1]

$304B

Spend of GenAI technology by 2028[1]

kyndryl

Drivers

Efficiency and productivity

- Automate repetitive tasks and streamline workflows

Customer experience

- Automate and personalize customer interactions

Innovation

- Accelerate product development and focus on innovation

Cost

- Optimize processes, automate tasks and provide actionable insights

Decision-making

- Provide data-driven insights that enhance decision-making processes

[1] IDC: WW GenAI forecast

Major AI/GenAI use cases customers are implementing

Finance

- Fraud detection
- Risk analysis (*)
- Know your customer
- Anti-money laundering
- Personalized Banking (*)
- Investments insight

Healthcare

- Medical image analysis
- Drug discovery (*)
- Next-generation DNA/RNA sequencing
- Molecule simulation
- Clinical trial data analysis (*)

Retail

- Self-checkout
- Loss prevention
- Video surveillance (*)
- Personalized shopping
- Automated catalogs creation
- Automated price optimization (*)

Telecom

- Virtual assistants
- Network performance tuning
- Remote support (*)

Media and Entertainment

- Character development (*)
- Video editing and image creation
- Style augmentation
- Artistic content generation

Manufacturing

- Factory simulation
- Product design (*)
- Predictive maintenance
- Manufacturing safety
- Visual inspection for quality control
- Delivery robots
- Digital twins (*)
- Self-driving vehicles

Public Sector

- Document summarization (*)
- Audit compliance (*)
- Virtual assistants

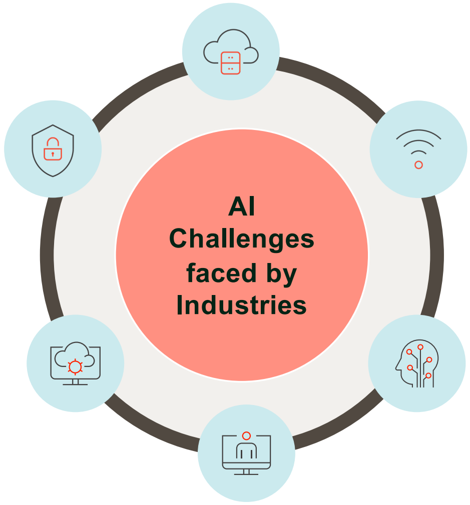
Energy

- Knowledge base QA (*)
- Predictive maintenance
- Customer service

(*) Since these use cases may use proprietary and potentially critical client data during the model's training phase or to populate a RAG database, they are among the best candidates to be implemented in an AI infrastructure on a Private Cloud

Industry challenges and opportunities for innovation

Challenges adopting AI



Business

Positive business outcomes
Proof of ROI

Data

Siloed / Legacy
Security & Privacy
Geo-political concerns

Gen AI Deployment

Model Selection
Business Process Integration
Lack of skills

Manage

Model management
Orchestration

Governance & Optimization

Lack of automation and cost transparency
Need Responsible AI Governance

Opportunities

Provide **Ready to implement Industry Use Cases**, leveraging Edge & AI capabilities

AI as a Service, with GPUs at Central DC or at Edge, AI Platform, complete with Security and Governance Framework

Data Preparation, Data management and Data Governance **as a Service**

Model Selection, development and training. **Skills for development** and human feedback.

AI Orchestration, Observability (AIOps) and Governance

Data & AI **Security**, Regulatory Compliance and **Sovereignty**

Major challenges in implementing AI/GenAI applications

Top concerns from customers ……

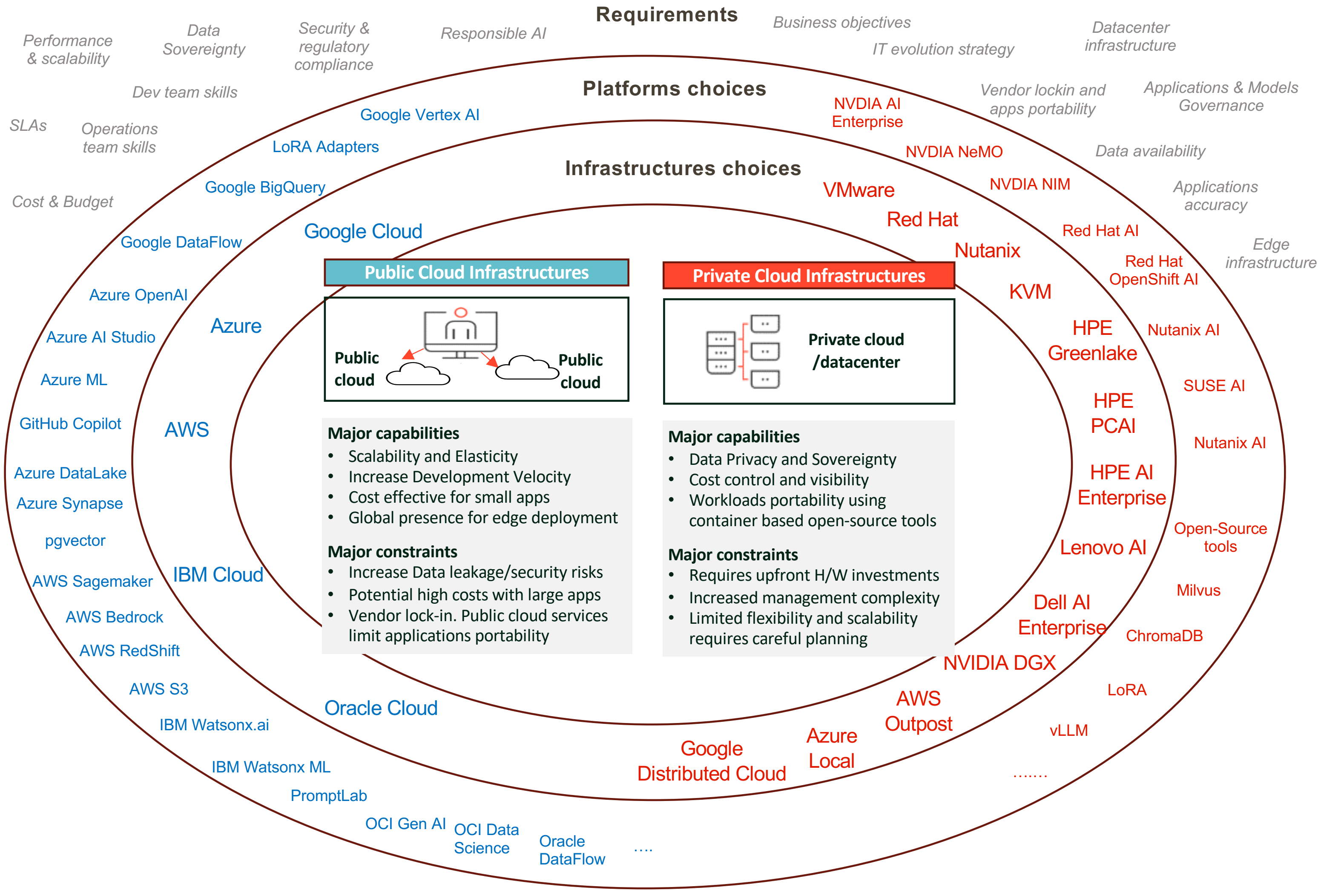
- Which **Public** or **Private cloud infrastructures** best address my needs in terms of
  - Data security, privacy and compliance, Sovereignty
  - Scalability, Flexibility and SLAs
  - Infrastructure costs and TCO
- Which **AI/GenAI platform** or **services** allows me to
  - Minimize the skills needed to develop and run my AI/GenAI applications
  - Expedite the development process and reduce the costs of running applications in production
  - Avoid vendor lock-in and allow to move my applications to any public or private cloud infrastructure without any or minimal code changes

Selecting the right infrastructure/platform for AI/GenAI workloads may be challenging

- No solution fits all

- All AI/GenAI infrastructures and platforms provide
  - Strong security
  - Support for the most common AI/GenAI tools and frameworks
  - Robustness and resiliency

- Right choice depends on
  - Applications architecture
  - Function and non-functional requirements
  - Customer IT strategy
  - Customer skills
  - …. and many other factors



Requirements

Performance & scalability, *Data Sovereignty*, *Security & regulatory compliance*, *Responsible AI*, *Business objectives*, *IT evolution strategy*, *Datacenter infrastructure*, *Dev team skills*, *SLAs*, *Operations team skills*, *Cost & Budget*, *Vendor lockin and apps portability*, *Applications & Models Governance*, *Data availability*, *Applications accuracy*, *Edge infrastructure*

Platforms choices

Google Vertex AI, LoRA Adapters, Google BigQuery, Google DataFlow, Azure OpenAI, Azure AI Studio, Azure ML, GitHub Copilot, Azure DataLake, Azure Synapse, pgvector, AWS Sagemaker, AWS Bedrock, AWS RedShift, AWS S3, IBM Watsonx.ai, IBM Watsonx ML, PromptLab, OCI Gen AI, OCI Data Science, Oracle DataFlow, ….

NVDIA AI Enterprise, NVDIA NeMO, NVDIA NIM, Red Hat AI, Red Hat OpenShift AI, Nutanix AI, SUSE AI, Nutanix AI, Open-Source tools, Milvus, ChromaDB, LoRA, vLLM, …….

Infrastructures choices

Google Cloud, Azure, AWS, IBM Cloud, Oracle Cloud

VMware, Red Hat, Nutanix, KVM, HPE Greenlake, HPE PCAI, HPE AI Enterprise, Lenovo AI, Dell AI Enterprise, NVIDIA DGX, AWS Outpost, Azure Local, Google Distributed Cloud

Public Cloud Infrastructures

Public cloud — Public cloud

Major capabilities
- Scalability and Elasticity
- Increase Development Velocity
- Cost effective for small apps
- Global presence for edge deployment

Major constraints
- Increase Data leakage/security risks
- Potential high costs with large apps
- Vendor lock-in. Public cloud services limit applications portability

Private Cloud Infrastructures

Private cloud /datacenter

Major capabilities
- Data Privacy and Sovereignty
- Cost control and visibility
- Workloads portability using container based open-source tools

Major constraints
- Requires upfront H/W investments
- Increased management complexity
- Limited flexibility and scalability requires careful planning

Examples of AI/GenAI applications mapping to different private/public infrastructures and platforms

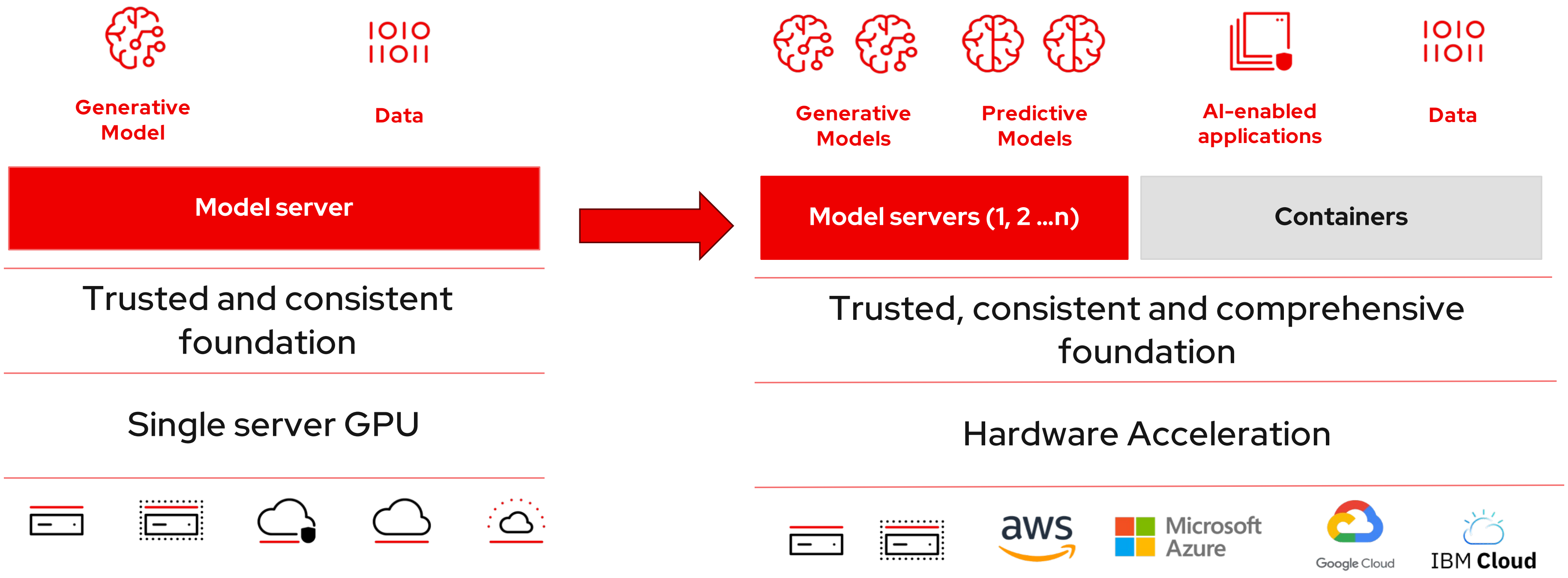
Use case	Training	Fine-tuning	Inference
Call-center Chatbot	**Platform**: OpenAI, or Azure OpenAI **Models/services**: No need for training, use pre-trained models like GPT-3, o-mini, etc. **Why**: Existing LLMs already provide NLP and context aware conversational capabilities	**Platform**: OpenAI, or Azure OpenAI **Models/services**: OpenAI fine-tuning APIs **Why** • require little customization for the specific domain • Does not use sensitive data for fine-tuning	**Platform**: OpenAI, or Azure OpenAI **Models/services**: OpenAI inference APIs **Why** • Does not use sensitive data • Can easily scale up/down based on demand
Legal documents summarizer	**Platform**: On-premises OpenShift with GPUs **Models/services**: No need for training use pre-trained models like LLama3, etc. **Why**: Existing LLMs already provide NLP and documents summarization capabilities	**Platform**: On-premises OpenShift with GPUs **Models/services**: Parameter-Efficient Fine-Tuning (PEFT) with LoRA, adapters, or prompt tuning **Why**: Model tuning may require use of sensitive data and data privacy protection	**Platform**: On-premises OpenShift with GPUs **Models/services**: NVIDIA NIM for Lllama3 **Why**: Inference phase may handle sensitive data that might be ingested to the system
Marketing campaign generation	**Platform**: AWS, Azure, GCP **Models/Services**: No need for training Llama 3, Stable Diffusion **Why**: Existing Llama and Stable-diffusion models already provide NLP, text and images generation capabilities	**Platform**: AWS, Azure, GCP **Models/**: Llama 3, Stable Diffusion **Services**: AWS Sagemaker training, Azure Machine Learning Training, Vertex AI Training **Why** • It may require significant resources to fine-tune the model to the specific domain • Marketing data are in general not very sensitive	**Platform**: AWS, Azure, GCP **Models/**: Llama 3, Stable Diffusion **Services**: AWS Sagemaker endpoint, Azure Machine Learning endpoint, Vertex AI Endpoint **Why** • Does not use sensitive data • Can easily scale up/down based on demand
Predictive maintenance	**Platform**: On-premises OpenShift with GPUs **Models:** Random Forest, SVM, Logistic regression, RNN, XGBoost **Services**: Pythorch/Tensorflow training, Scikit-learn **Why**: Models use sensor data or image analysis, that might not require too many GPU resources for training. Training data might be sensitive and contain industrial secrets that should be protected		**Platform**: OpenShift single node on bare-metal with GPUs **Models:** Random Forest, SVM, Logistic regression, RNN, **Services**: Single node OpenShift, K3s, KServe **Why**: Models need to run in edge sites, gathering sensitive data that need to be processed immediately close to the source

Agenda

- Market Insight, needs, challenges and opportunities
- How Red Hat AI can help
- How Kyndryl can help

Red Hat AI overview

From single server deployments to highly scaled-out platform architectures



Generative Model

Data

Model server

Trusted and consistent foundation

Single server GPU

Generative Models

Predictive Models

AI-enabled applications

Data

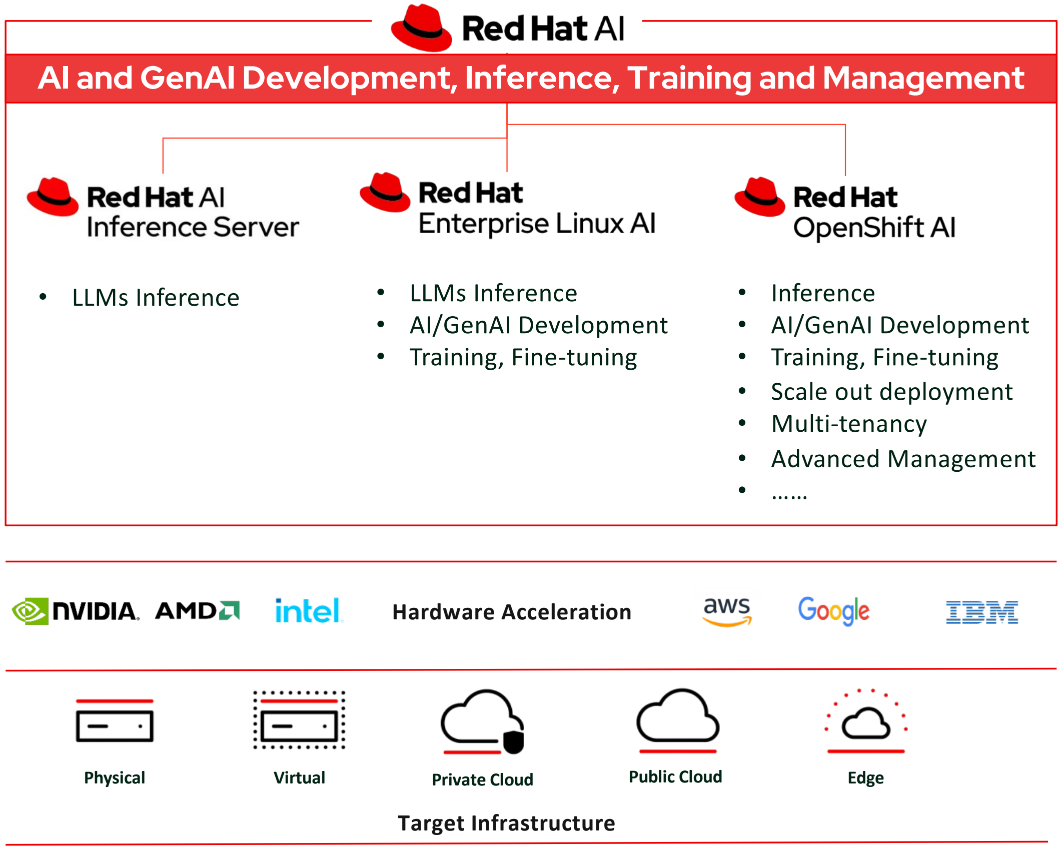
Model servers (1, 2 …n)

Containers

Trusted, consistent and comprehensive foundation

Hardware Acceleration

Red Hat AI overview



Red Hat AI

AI and GenAI Development, Inference, Training and Management

Red Hat AI Inference Server

- LLMs Inference

Red Hat Enterprise Linux AI

- LLMs Inference
- AI/GenAI Development
- Training, Fine-tuning

Red Hat OpenShift AI

- Inference
- AI/GenAI Development
- Training, Fine-tuning
- Scale out deployment
- Multi-tenancy
- Advanced Management
- ……

NVIDIA AMD intel Hardware Acceleration aws Google IBM

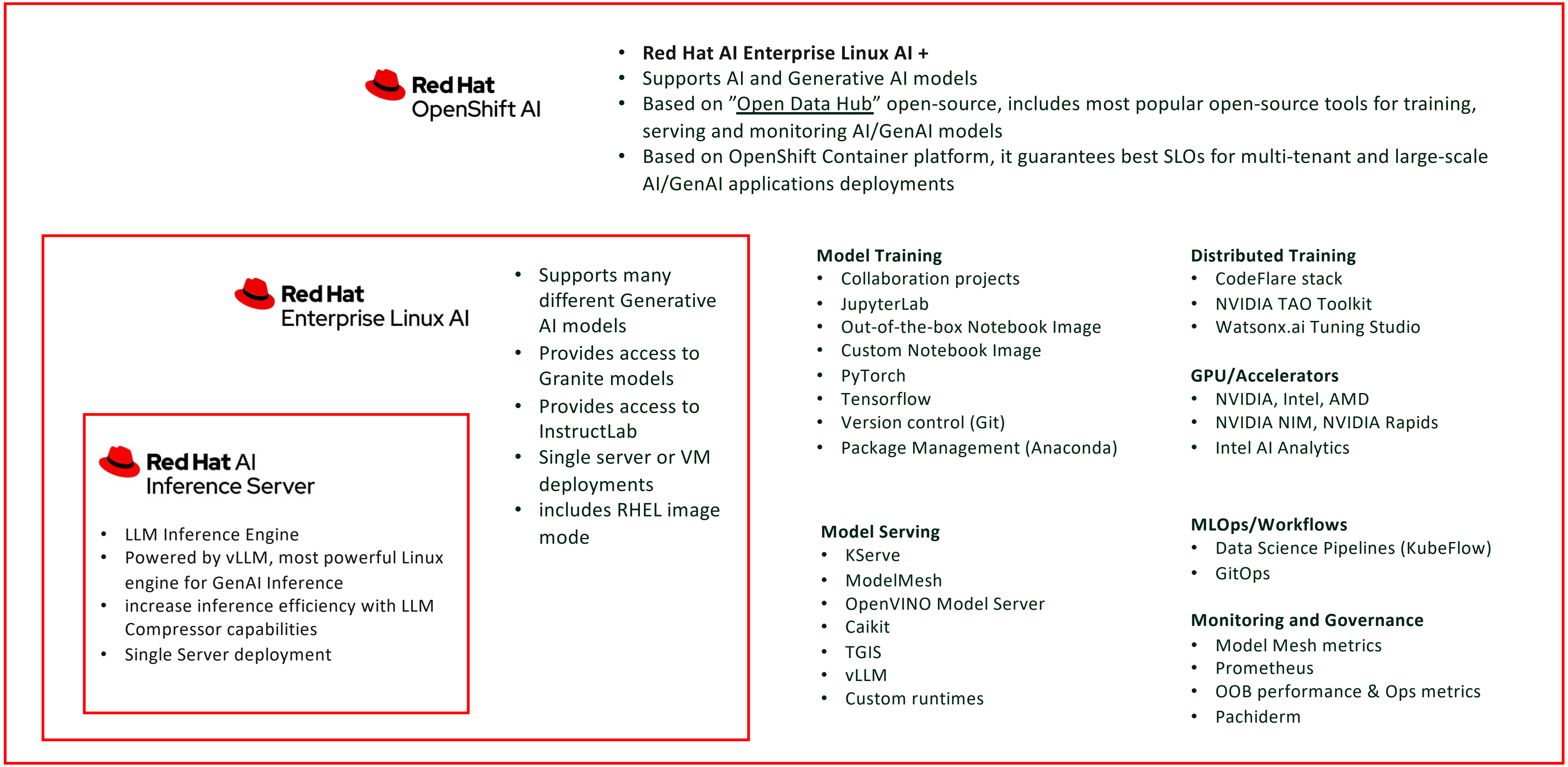
Physical Virtual Private Cloud Public Cloud Edge

Target Infrastructure

A suite of products to build and run AI and Gen AI solutions from the PoC phase to the full production deployment

- **Red Hat AI Inference Server** provides an optimized Inference Engine based on vLLM to run AI/GenAI apps
- **Red Hat Enterprise Linux AI** adds model development, training and tuning capabilities to "*Red Hat AI Inference Server*" ones
- **Red Hat OpenShift AI** includes capabilities of" *Red Hat Enterprise Linux AI*" and "*Red Hat AI Inference Server*".
  - Built on OpenShift Container Platform
  - provides the most advanced capabilities for Developing, training, fine-tuning, run and manage large scale AI/GenAI applications in an enterprise-grade production environment

Red Hat AI major features



Red Hat OpenShift AI

- **Red Hat AI Enterprise Linux AI +**
- Supports AI and Generative AI models
- Based on "Open Data Hub" open-source, includes most popular open-source tools for training, serving and monitoring AI/GenAI models
- Based on OpenShift Container platform, it guarantees best SLOs for multi-tenant and large-scale AI/GenAI applications deployments

Red Hat Enterprise Linux AI

- Supports many different Generative AI models
- Provides access to Granite models
- Provides access to InstructLab
- Single server or VM deployments
- includes RHEL image mode

Red Hat AI Inference Server

- LLM Inference Engine
- Powered by vLLM, most powerful Linux engine for GenAI Inference
- increase inference efficiency with LLM Compressor capabilities
- Single Server deployment

Model Training
- Collaboration projects
- JupyterLab
- Out-of-the-box Notebook Image
- Custom Notebook Image
- PyTorch
- Tensorflow
- Version control (Git)
- Package Management (Anaconda)

Model Serving
- KServe
- ModelMesh
- OpenVINO Model Server
- Caikit
- TGIS
- vLLM
- Custom runtimes

Distributed Training
- CodeFlare stack
- NVIDIA TAO Toolkit
- Watsonx.ai Tuning Studio

GPU/Accelerators
- NVIDIA, Intel, AMD
- NVIDIA NIM, NVIDIA Rapids
- Intel AI Analytics

MLOps/Workflows
- Data Science Pipelines (KubeFlow)
- GitOps

Monitoring and Governance
- Model Mesh metrics
- Prometheus
- OOB performance & Ops metrics
- Pachiderm

RedHat OpenShift AI new/enhanced features

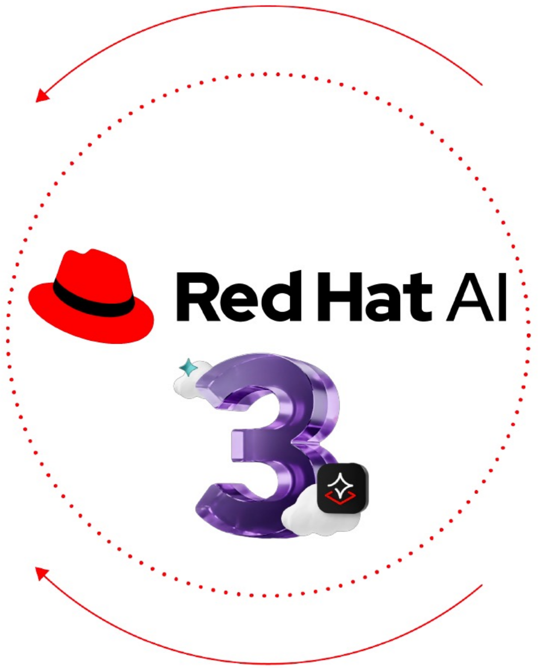
Announcements

Flexible and Efficient Inference

- GA distributed inference (llm-d)
- New validated and optimized models
- vLLM enhancements
- LLM Compressor GA

Connecting Models to Data

- Modular and extensible approach for: data ingestion, synthetic data generation, tuning, evaluations.
- RAG enhancements & partner integrations
- Continual Post Training Algorithm
- Feature Store GA



Agentic AI

- AI experiences: AI hub and gen AI studio
- Model Context Protocol support & MCP Server access in gen AI studio
- Llama Stack API integration

AI Platform

- Model catalog and registry GA
- Model as a Service provider enhancements and API Mgt integratior
- GPU as a Service enhancements

Single platform to run any model, on any accelerator, on any cloud

RedHat OpenShift AI high level architecture

Based on the open-source Open Data Hub project provides

AI/ML modeling and visualization tools

- JupyterLab UI with prebuilt notebook images and Python libraries
- TensorFlow, PyTorch, CUDA, Kubeflow, Anaconda (optional)
- Parallelized and distributed workloads (KubeRay, CodeFlare)
- Data drift and Bias detection
- Efficient fine-tuning with low-rank adapters (LoRA)

Model serving and monitoring

- Model serving (KServe with user interface), vLLM serving runtime
- Model monitoring,
- OpenShift Source-to-Image (S2I),
- Red Hat OpenShift API Management (optional add-on),
- Intel Distribution of OpenVINO toolkit (*)
- Support for NVIDIA NIM

Data engineering

Starburst (optional), Pachyderm (optional)

Data ingestion and storage

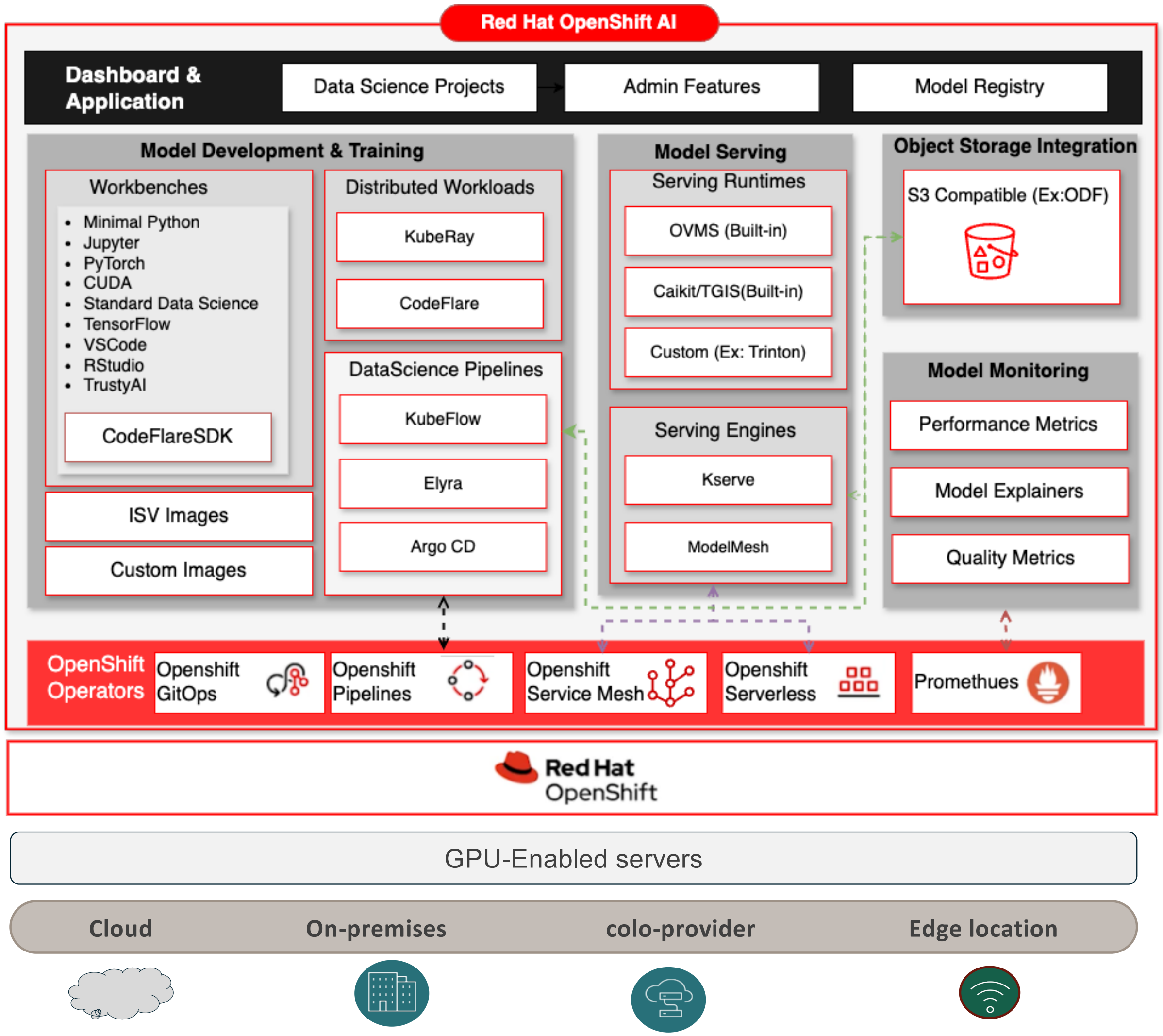
- Model registry, Red Hat AMQ (optional add-on);
- Amazon Simple Storage Service (S3)

Data science pipelines

Kubeflow Pipelines to chain together processes like data preparation, build models, and serve models

GPU support

- NVIDIA GPUs
- AMD GPUs
- Intel GPUs (including Xeon, Gaudi, and Intel Data Center GPU Flex Series))



Red Hat OpenShift AI

Dashboard & Application: Data Science Projects, Admin Features, Model Registry

Model Development & Training

Workbenches
- Minimal Python
- Jupyter
- PyTorch
- CUDA
- Standard Data Science
- TensorFlow
- VSCode
- RStudio
- TrustyAI

CodeFlareSDK

ISV Images

Custom Images

Distributed Workloads: KubeRay, CodeFlare

DataScience Pipelines: KubeFlow, Elyra, Argo CD

Model Serving

Serving Runtimes: OVMS (Built-in), Caikit/TGIS(Built-in), Custom (Ex: Trinton)

Serving Engines: Kserve, ModelMesh

Object Storage Integration: S3 Compatible (Ex:ODF)

Model Monitoring: Performance Metrics, Model Explainers, Quality Metrics

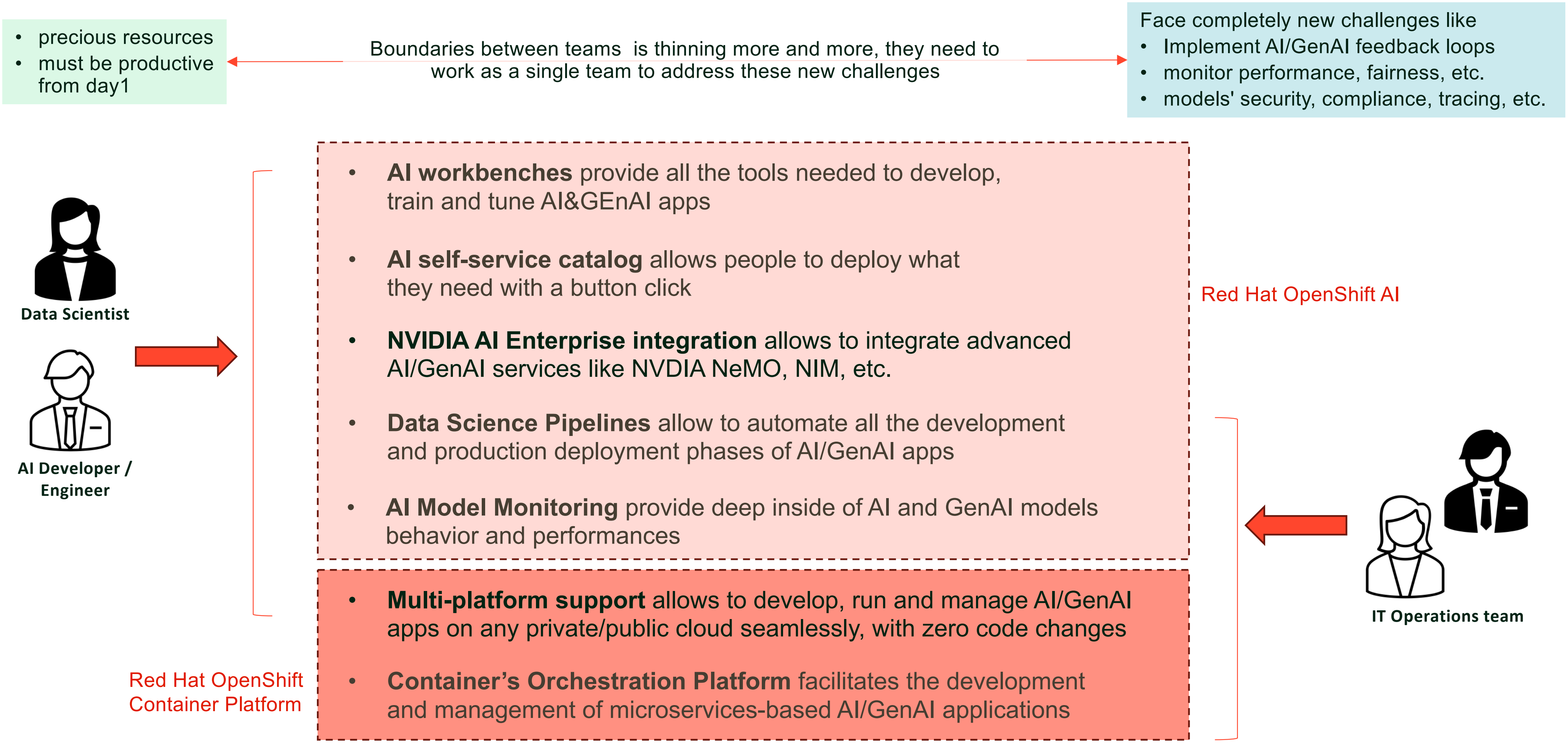
OpenShift Operators: Openshift GitOps, Openshift Pipelines, Openshift Service Mesh, Openshift Serverless, Promethues

Red Hat OpenShift

GPU-Enabled servers

Cloud, On-premises, colo-provider, Edge location

How Red Hat OpenShift AI address the needs of multiple actors in AI/GenAI space



- precious resources
- must be productive from day1

Boundaries between teams is thinning more and more, they need to work as a single team to address these new challenges

Face completely new challenges like

- Implement AI/GenAI feedback loops
- monitor performance, fairness, etc.
- models' security, compliance, tracing, etc.

Data Scientist

AI Developer / Engineer

- **AI workbenches** provide all the tools needed to develop, train and tune AI&GEnAI apps
- **AI self-service catalog** allows people to deploy what they need with a button click
- **NVIDIA AI Enterprise integration** allows to integrate advanced AI/GenAI services like NVDIA NeMO, NIM, etc.
- **Data Science Pipelines** allow to automate all the development and production deployment phases of AI/GenAI apps
- **AI Model Monitoring** provide deep inside of AI and GenAI models behavior and performances

Red Hat OpenShift AI

- **Multi-platform support** allows to develop, run and manage AI/GenAI apps on any private/public cloud seamlessly, with zero code changes
- **Container's Orchestration Platform** facilitates the development and management of microservices-based AI/GenAI applications

Red Hat OpenShift Container Platform

IT Operations team

Red Hat AI key benefits

Accelerate time to market:
Microservices architecture significantly decrease the time to develop and deploy scalable, resilient, and adaptable applications

Improve developers' productivity

- Leverage catalog of ready-to-use tools for AI/GenAI training, tuning, inference
- Leverage MLOPs built-in capabilities to automate the models tuning, inference and validation loop

Improve IT Operations productivity

- Leverage MLOPs built-in capabilities to automate the setup of dev/staging or production environments
- Leverage built-in models monitoring to get deep inside of AI and GenAI models behavior and performances

Enforce data compliance and sovereignty

- Keep mission-critical data secured in house and under your control
- reduce the risk of data leakages
- Operational resiliency & sovereignty
- Trust and transparency

Avoid Vendor Lock-in

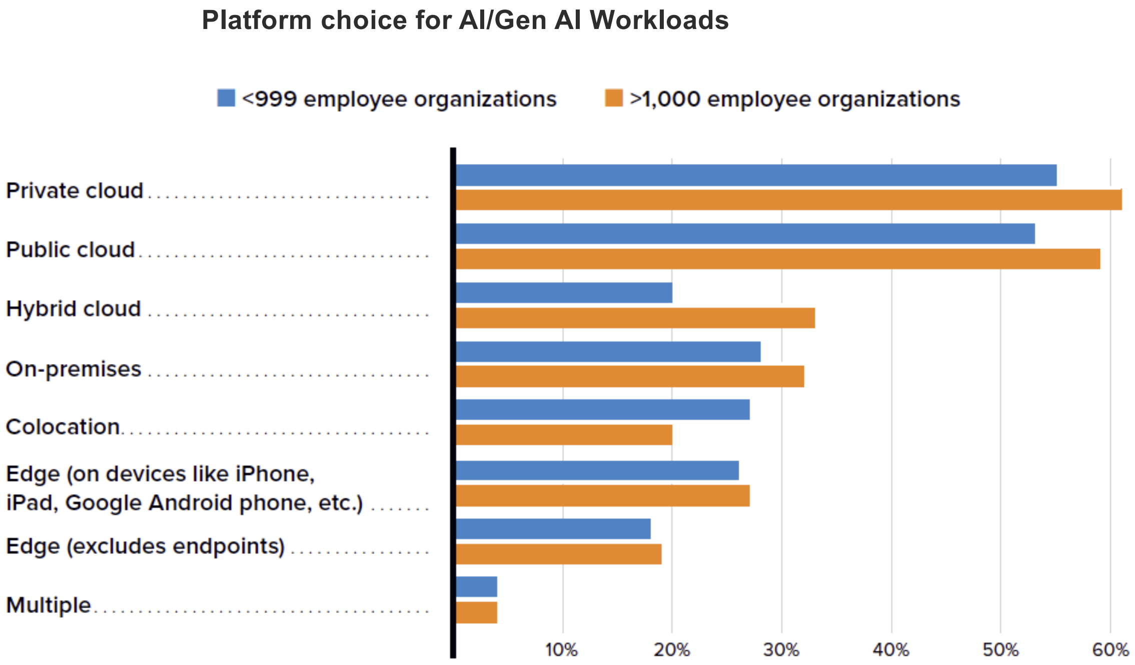
- Leverage Open-Source AI/GenAI frameworks and tools
- Build and Run AI/GenAI applications everywhere, on private or public clouds, with zero code changes

Meet most demanding SLOs:
OpenShift Container platform guarantees best SLOs for multi-tenant and large-scale AI/GenAI applications deployments

Optimize Infrastructure costs

- Leverage the support of multiple H/W vendors and different cloud providers to select the most convenient infrastructure solution
- Leverage built-in support for HCI storage as an alternative to more complex and costly SAN/NAS storage
- Choose between managed and self-managed solutions and Capex, Opex and pay-per-use models

Red Hat AI on Private Cloud Key benefits



Source: IDC's *AI StrategiesView 2022*

For AI/GenAI workloads, **Private Cloud is the first choice** for customers that need

- Complete control of the infrastructure
- Data security, sovereignty and compliance
- Cost control
- Strong SLAs requirements
- Avoid vendor lock-in
- Flexibility in the choice of AI/GenAI tools
- Run applications on central datacenter and edge locations
- Maintain some footprint on-premises

The combination of ReHat AI on a Private Cloud, can help to address many customer challenges

Agenda

- Market Insight, needs, challenges and opportunities
- How Red Hat AI can help
- How Kyndryl can help

Kyndryl helps clients to implement AI/GenAI solutions across the entire stack

Area	Description	Examples
Application Layer	Design, build and manage Business-facing solutions that apply AI/GenAI to solve specific use cases, improve productivity, or generate value	Agentic AI
		AI chatbots & virtual agents, Recommendation Engines, Fraud detection, etc…
		Predictive maintenance, Anomaly detection, etc…
Platform Layer	Design, build and manage the core virtualization, orchestration and AI/GenAI platforms to develop and run AI/GenAI models, workflows, and data pipelines	VMware, Red Hat, Nutanix, Suse, Azure local, AWS Outpost, Google Distributed Cloud
		Kubernetes, Tanzu, OpenShift, Rancher, AKS, EKS, GKE
		OpenShift AI, NVIDIA AI Enterprise, Open-source AI/GenAI frameworks
		PyTorch, TensorFlow, Triton, NVIDIA NIM, Milvus, KubeFlow,, Ray, TrustyAI, LangChain, Haystack , CrewAI, AutoGen, etc.
Infrastructure Layer (HW & SW)	Design, build and manage the underlying accelerated compute, storage, and network infrastructure that enable scalable and efficient AI operations	**Hardware**: Dell, HPE, Lenovo, NVIDIA, etc.
		H/W Acceleration: NVIDIA L40s/H100/H200, Bluefield, AMD Instinct, Alveo, Pensando, Intel Habana, Arc, etc…
		Software: NVIDIA CUDA X, CUDA X-AI, Magnum-IO, vGPU, GPU Operator, Network operator, etc..

Strategic Benefits

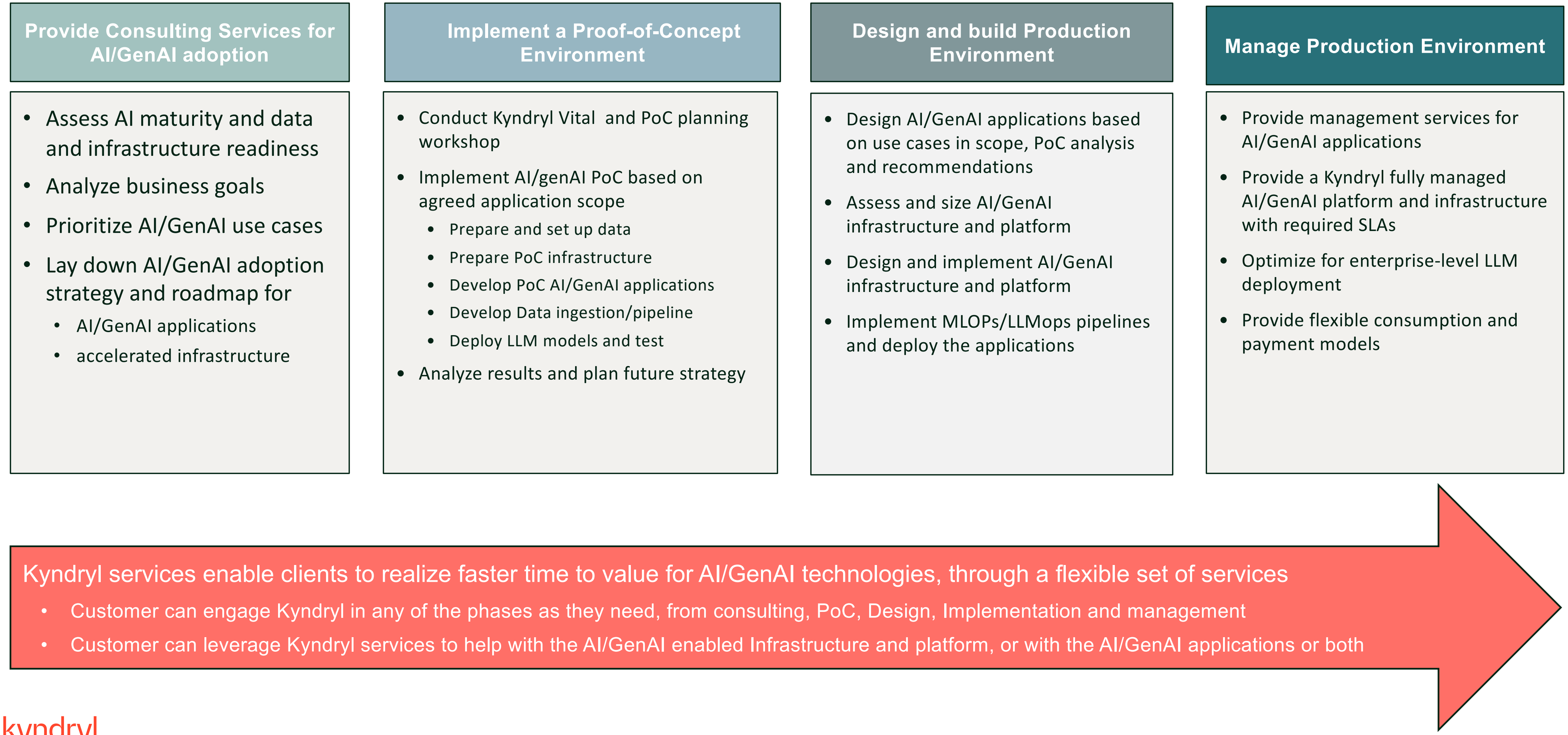
Data Privacy and Security: Deployment in strictly controlled environments guarantees data sovereignty and compliance.

Cost Control: Keep the direct control of H/W cost and S/W licenses usage.

Customization: Open-source frameworks + industry standard tools enable tailored AI deployments.

Avoid vendor lockin: Containerized AI/GenAI apps built on open-source frameworks can run on any private, hybrid or public cloud platform

Kyndryl approach from Consulting to PoC and production



Provide Consulting Services for AI/GenAI adoption

- Assess AI maturity and data and infrastructure readiness
- Analyze business goals
- Prioritize AI/GenAI use cases
- Lay down AI/GenAI adoption strategy and roadmap for
  - AI/GenAI applications
  - accelerated infrastructure

Implement a Proof-of-Concept Environment

- Conduct Kyndryl Vital and PoC planning workshop
- Implement AI/genAI PoC based on agreed application scope
  - Prepare and set up data
  - Prepare PoC infrastructure
  - Develop PoC AI/GenAI applications
  - Develop Data ingestion/pipeline
  - Deploy LLM models and test
- Analyze results and plan future strategy

Design and build Production Environment

- Design AI/GenAI applications based on use cases in scope, PoC analysis and recommendations
- Assess and size AI/GenAI infrastructure and platform
- Design and implement AI/GenAI infrastructure and platform
- Implement MLOPs/LLMops pipelines and deploy the applications

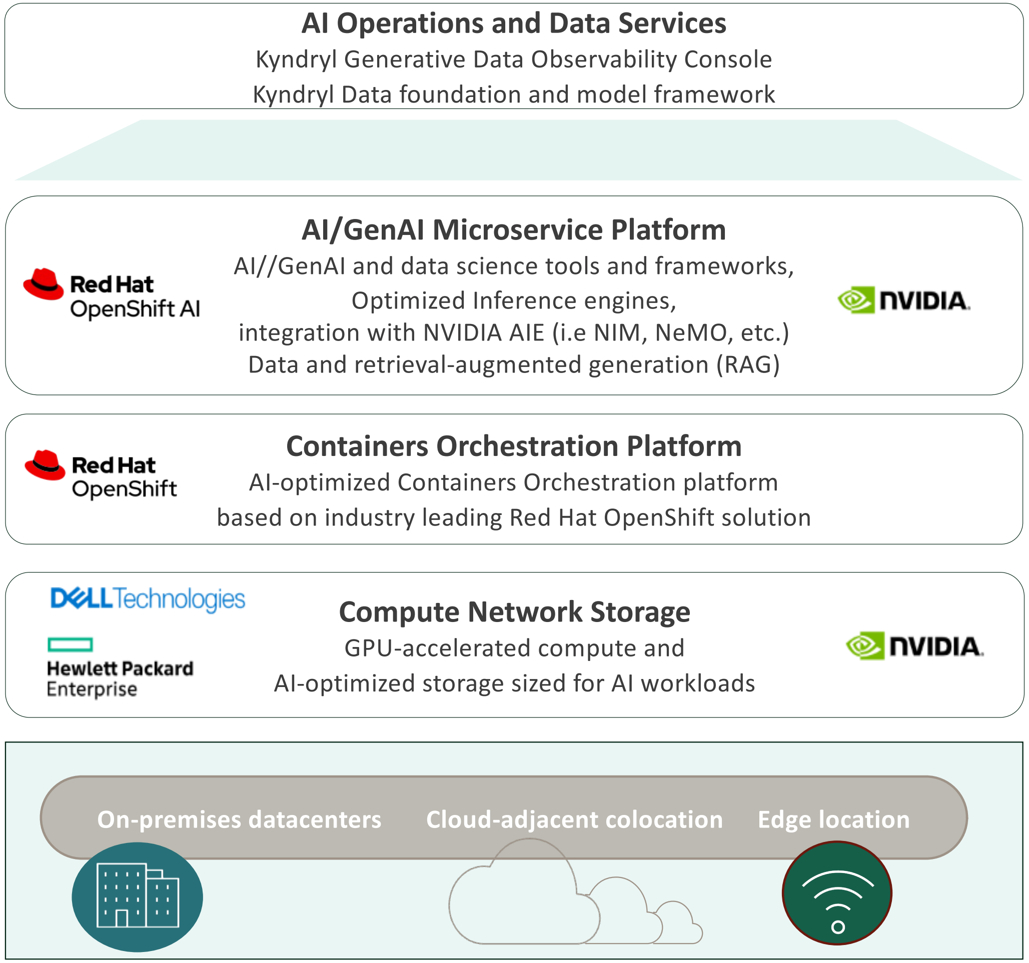
Manage Production Environment

- Provide management services for AI/GenAI applications
- Provide a Kyndryl fully managed AI/GenAI platform and infrastructure with required SLAs
- Optimize for enterprise-level LLM deployment
- Provide flexible consumption and payment models

Kyndryl services enable clients to realize faster time to value for AI/GenAI technologies, through a flexible set of services

- Customer can engage Kyndryl in any of the phases as they need, from consulting, PoC, Design, Implementation and management
- Customer can leverage Kyndryl services to help with the AI/GenAI enabled Infrastructure and platform, or with the AI/GenAI applications or both

Kyndryl services for Data and AI/GenAI Infrastructures with Red Hat OpenShift AI



AI Operations and Data Services
Kyndryl Generative Data Observability Console
Kyndryl Data foundation and model framework

AI/GenAI Microservice Platform
AI//GenAI and data science tools and frameworks,
Optimized Inference engines,
integration with NVIDIA AIE (i.e NIM, NeMO, etc.)
Data and retrieval-augmented generation (RAG)

Containers Orchestration Platform
AI-optimized Containers Orchestration platform
based on industry leading Red Hat OpenShift solution

Compute Network Storage
GPU-accelerated compute and
AI-optimized storage sized for AI workloads

On-premises datacenters
Cloud-adjacent colocation
Edge location

Data and AI operations services

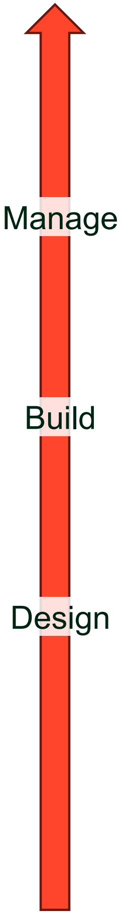
- Efficient LLM operations
- Comprehensive data foundation and governance services
- Complete Data Observability Services and Console

AI/GenAI platform services

- Container orchestration based on OpenShift
- Optional HCI storage based on OpenShift Data Foundation
- Model training, inferencing, fine-tuning using OpenShift AI
- MLOps/LLMOps based on OpenShift AI data pipelines
- Optional infrastructure for retrieval-augmented-generation
- Advisory, design and build services
- Managed by client, or Managed hosting by Kyndryl in as-a-service model

AI/GenAI enabled Infrastructure services

- GPU accelerated servers based on HPE or Dell
- Infrastructure sized according to application requirements
- Flexible hosting in client, colo, edge or Kyndryl datacenter
- Flexible cloud OPEX model
- Advisory, design and build services
- Managed by client, or Managed hosting by Kyndryl in as-a-service model



Manage
Build
Design

Kyndryl services for Data and AI/GenAI Infrastructures with Red Hat OpenShift AI

Secured, Dedicated, single-tenant, on-premises Platform deployment

- GPU-enabled servers, with storage and network designed to meet required LLM response time and throughput, based on Dell or HPE H/W with NVIDIA GPUs
- GPUs-enabled Red Hat OpenShift for microservice and application deployment
- Security rich, air gapped platform ideal for regulated or private workloads

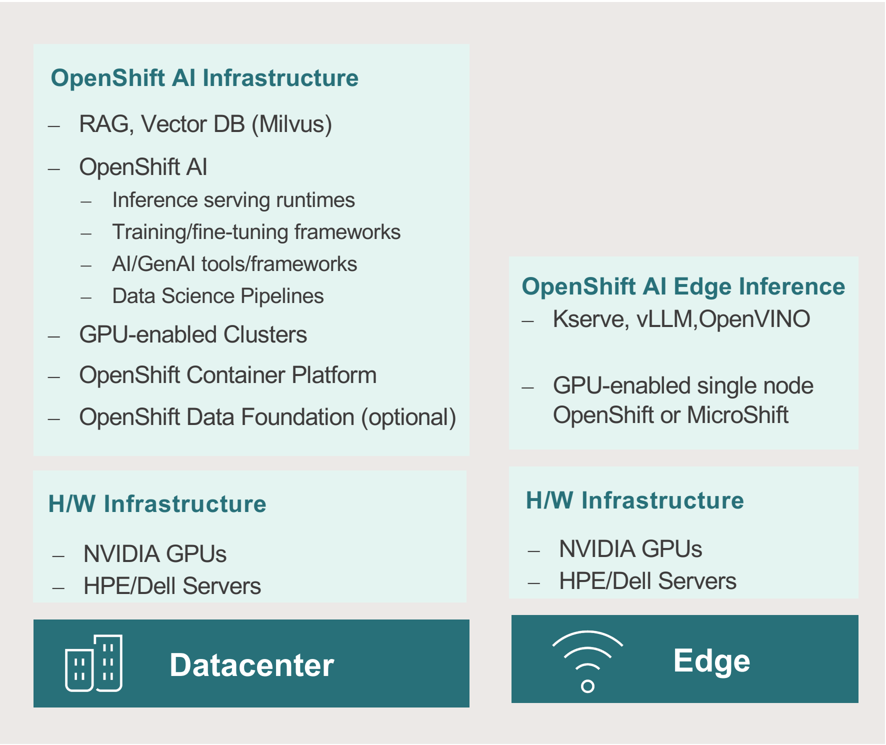
Optimized for AI – Inferencing with RAG

- Red Hat OpenShift AI used for AI/GenAI and data science tools and frameworks
- Optionally integrate NVDIA AI Enterprise services (e.g. Nemo, NIM, etc.)
- Data Science Pipelines based on embedded KubeFlow
- RAG infrastructure based on Milvus

Deployed by Kyndryl, can be managed by client or delivered as-a-service by Kyndryl

- Kyndryl Design and Build the platform in customer premise or colo provider location
- Client can manage the platform by itself or have it managed by Kyndryl (24x7x365 based on defined SLOs and KPIs for GenAI workloads)
- If managed by Kyndryl, the platform can be delivered as-a-service
- Platform components can be delivered independently if required by customer

Different OpenShift AI deployment models



OpenShift AI Infrastructure

- RAG, Vector DB (Milvus)
- OpenShift AI
  - Inference serving runtimes
  - Training/fine-tuning frameworks
  - AI/GenAI tools/frameworks
  - Data Science Pipelines
- GPU-enabled Clusters
- OpenShift Container Platform
- OpenShift Data Foundation (optional)

H/W Infrastructure

- NVIDIA GPUs
- HPE/Dell Servers

Datacenter

OpenShift AI Edge Inference

- Kserve, vLLM,OpenVINO
- GPU-enabled single node OpenShift or MicroShift

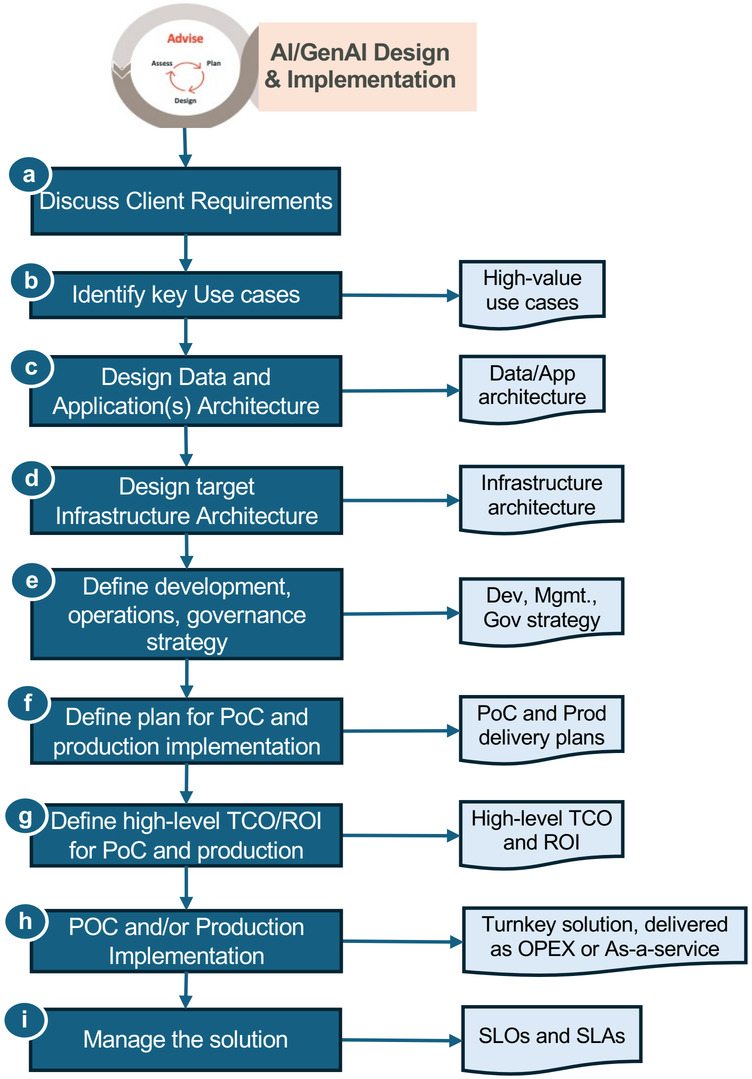
H/W Infrastructure

- NVIDIA GPUs
- HPE/Dell Servers

Edge

The Kyndryl approach

Kyndryl can help clients in the realization of a Private Cloud solution for AI/GenAI using an holistic approach that analyzes all the different aspects and phases of the analysis, design, implementation and management of such solution



AI/GenAI Design & Implementation

a Discuss Client Requirements

b Identify key Use cases → High-value use cases

c Design Data and Application(s) Architecture → Data/App architecture

d Design target Infrastructure Architecture → Infrastructure architecture

e Define development, operations, governance strategy → Dev, Mgmt., Gov strategy

f Define plan for PoC and production implementation → PoC and Prod delivery plans

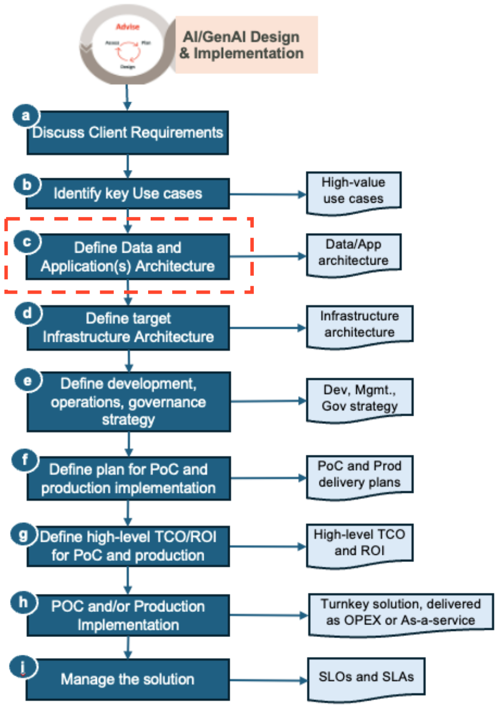
g Define high-level TCO/ROI for PoC and production → High-level TCO and ROI

h POC and/or Production Implementation → Turnkey solution, delivered as OPEX or As-a-service

i Manage the solution → SLOs and SLAs

kyndryl

An example – Design Data and Application(s) architecture



AI/GenAI Design & Implementation

a Discuss Client Requirements

b Identify key Use cases → High-value use cases

c Define Data and Application(s) Architecture → Data/App architecture

d Define target Infrastructure Architecture → Infrastructure architecture

e Define development, operations, governance strategy → Dev, Mgmt., Gov strategy

f Define plan for PoC and production implementation → PoC and Prod delivery plans

g Define high-level TCO/ROI for PoC and production → High-level TCO and ROI

h POC and/or Production Implementation → Turnkey solution, delivered as OPEX or As-a-service

i Manage the solution → SLOs and SLAs

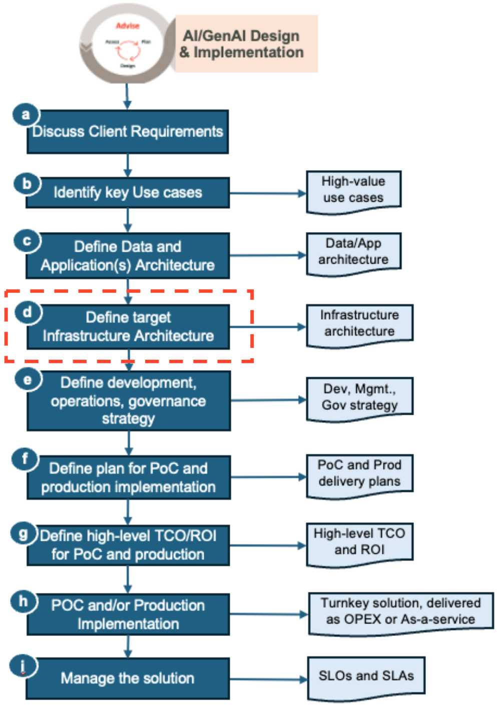
Define high level data and application(s) architectural design

- ➔ Understand which data are used, where they originally reside, how they should be prepared and where they should be stored for the AI/GenAI applications use
- ➔ Understand application scenario, models used, required latency/throughput, etc.
- ➔ Pay attention to application requirements in terms of training, fine-tune, RAG, inference
- ➔ Look for most popular applications in the same industry and search for applications blueprints, or for most common vendors/solutions in that industry
- ➔ Create the high-level design of the application(s), including the major data flows and the MLOps/LLMOps processes

Kyndryl Reusable assets and supporting material

- Use Data & Applications Architecture Patterns folder where you can find AI/GenAI Applications Blueprints from Google, Azure, NVIDIA and Red Hat

An example – Design Infrastructure Architecture



AI/GenAI Design & Implementation

a Discuss Client Requirements

b Identify key Use cases → High-value use cases

c Define Data and Application(s) Architecture → Data/App architecture

d Define target Infrastructure Architecture → Infrastructure architecture

e Define development, operations, governance strategy → Dev, Mgmt., Gov strategy

f Define plan for PoC and production implementation → PoC and Prod delivery plans

g Define high-level TCO/ROI for PoC and production → High-level TCO and ROI

h POC and/or Production Implementation → Turnkey solution, delivered as OPEX or As-a-service

i Manage the solution → SLOs and SLAs

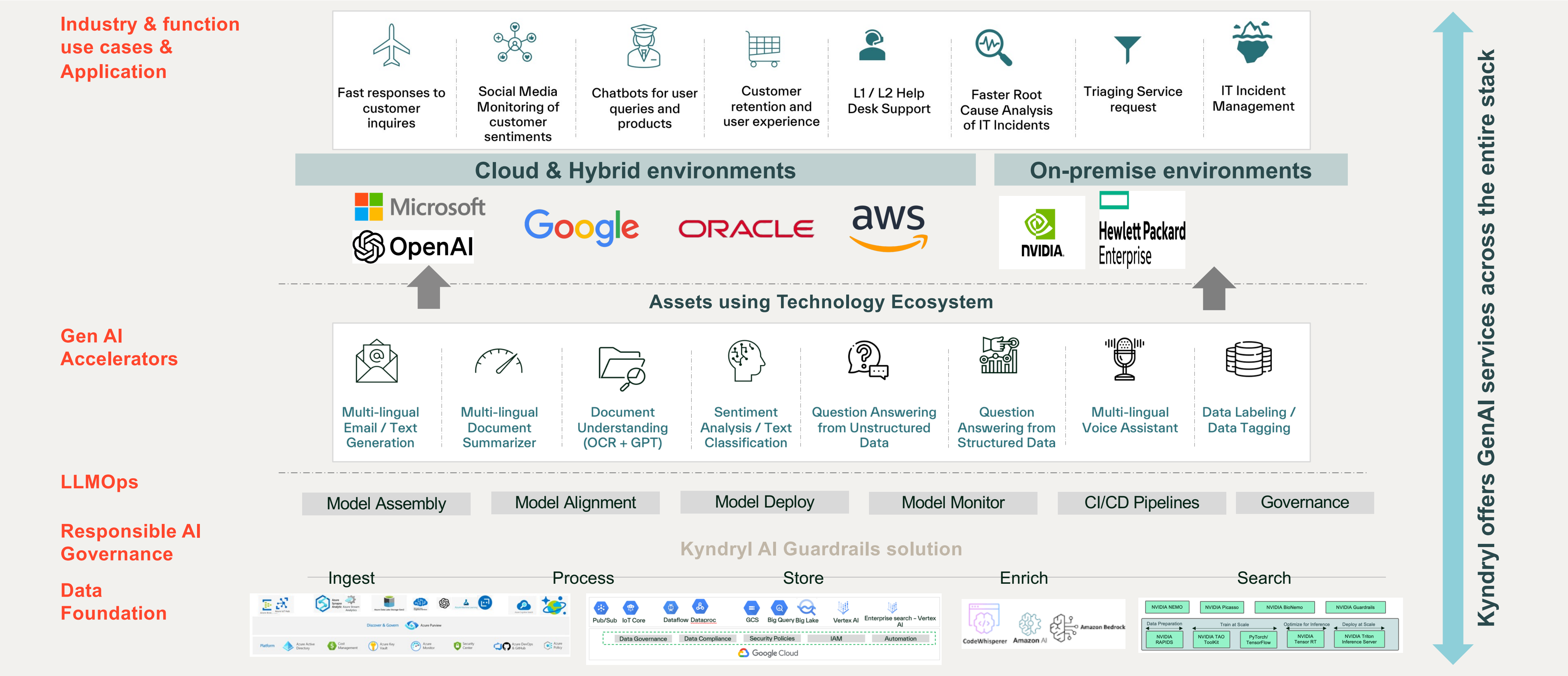
Design an AI/GenAI Infrastructure architecture requires deep investigation

- Right size the infrastructure based on application requirements
  - **Type of AI Models** – (Computer Vision / NLP/ Reinforcement learning /Deep Learning)
  - **Model size** (e.g. Llama-70B, Mistral-7B, Llama-13B, etc.)
  - **Application requirements and architecture –** Training, fine-tuning, RAG, inference only ?
  - **Data size** – Number of I/O tokens, system prompts) Image resolution, Audio sample rate, Text length etc.
  - **Performance objectives** – Time-to-first-token (TTFT), Inter-Token-Latency (ITL), E2E Latency
  - **Scalability objectives** – Number of concurrent requests per sec/min, Output Tokens generated per sec/min
  - **Model accuracy** – Precision used to train and evaluate the models
- GPUs – how many GPUs and GPUs types,
- GPUs topology – single, NVLink or GPUs Fabric
- Network and storage infrastructure – Infiniband, Ethernet, 25/100/200/400GB, RoCE, GPUDirect, etc..
- Servers' configuration – PCIe switches, Accelerated NICs/DPUs, CPU/RAM/DISK
- Datacenter power & cooling – wattage, heat dissipation, cooling type (air/liquid), cooling capacity, etc.
- Edge Deployment – Security, network, storage, etc

Kyndryl Reusable assets and supporting material

- AI-GenAI Infrastructure Architecture patterns
- AI-GenAI Sample BOMs
- GPU Sizing guidance & tool
- Red Hat OpenShit AI Reference architecture, etc

Kyndryl AI capabilites across different Technologies



Industry & function use cases & Application

Fast responses to customer inquires

Social Media Monitoring of customer sentiments

Chatbots for user queries and products

Customer retention and user experience

L1 / L2 Help Desk Support

Faster Root Cause Analysis of IT Incidents

Triaging Service request

IT Incident Management

Cloud & Hybrid environments

Microsoft · OpenAI · Google · ORACLE · aws

On-premise environments

NVIDIA · Hewlett Packard Enterprise

Assets using Technology Ecosystem

Gen AI Accelerators

Multi-lingual Email / Text Generation

Multi-lingual Document Summarizer

Document Understanding (OCR + GPT)

Sentiment Analysis / Text Classification

Question Answering from Unstructured Data

Question Answering from Structured Data

Multi-lingual Voice Assistant

Data Labeling / Data Tagging

LLMOps

Model Assembly · Model Alignment · Model Deploy · Model Monitor · CI/CD Pipelines · Governance

Responsible AI Governance

Kyndryl AI Guardrails solution

Data Foundation

Ingest · Process · Store · Enrich · Search

Azure IoT Hub · Azure Synapse Analytic · Azure Stream Analytics · Azure Data Lake Storage Gen2 · Azure Machine Learning · Azure Cognitive Search

Discover & Govern · Azure Purview

Platform · Azure Active Directory · Cost Management · Azure Key Vault · Azure Monitor · Security Center · Azure DevOps & GitHub · Azure Policy

Pub/Sub · IoT Core · Dataflow · Dataproc · GCS · Big Query · Big Lake · Vertex AI · Enterprise search – Vertex AI

Data Governance · Data Compliance · Security Policies · IAM · Automation

Google Cloud

CodeWhisperer · Amazon AI · Amazon Bedrock

NVIDIA NEMO · NVIDIA Picasso · NVIDIA BioNemo · NVIDIA Guardrails

Data Preparation: NVIDIA RAPIDS

Train at Scale: NVIDIA TAO ToolKit · PyTorch/ TensorFlow

Optimize for Inference: NVIDIA Tensor RT

Deploy at Scale: NVIDIA Triton Inference Server

Kyndryl offers GenAI services across the entire stack

Leverage Kyndryl expertise in the field of AI/Gen AI

Skills & Expertise

- **Dedicated team** of GenAI SMEs, consultants, architects and engineers certified across hyperscalers and on-prem with proven experience in GenAI
- **Deep domain expertise** after handling multiple customer scenarios
- Continued **investments** in **hiring and upskilling**
- **350+ Data and AI patents**

Investments

- Building a **CoE with Microsoft** across GenAI, Data Foundations (Fabric, Purview), & Apps Modernization and Responsible AI Partner (one of 11)
- Multi-year strategic agreement - **AWS Joint Innovation Factory** to build industry solutions in GenAI/ML
- **Joint partnership with Google** focused on GenAI solutions, AI Governance & Data, SAP data & AI modernization
- **Collaboration with Nvidia and Dell-** for private AI/ GenAI solutions

Kyndryl as Customer 0

- Embedding **GenAI in our delivery capabilities**: Kyndryl Bridge – Natural Language ChatOps for querying IT / ticketing data
- **Several GenAI solutions** for CFO/ Solutioning organizations, such as Kyndryl IR Advisor bot to support Q&A, meeting prep, information retrieval and summary from investor reports, calls, market news

Assets

- **LLMOps** (Large Language Model Ops) console & Platform to scale responsibly GenAI solutions & models
- **GenAI assets** such as: Kyndryl SRE Assist to speed up end2end DevOps automation, code modernization to translate code between languages
- **Industry solutions**, such as: Automated Quality Inspection leveraging Vision AI, Worker Safety, etc.
- Robust **consulting framework, and methodologies** (GenAI risk assessment, data foundations assessment framework, etc.)

Pilots & solutions in production

- **Social media** – sentiment analysis, personalised responses and responses to information seeking question
- **Natural language** queries to fetch data from database
- **Issue resolution for contact centre** leveraging FAQs and customer data in WhatsApp
- **Q&A** from large contractual documents for government entity
- …

kyndryl

Red Hat AI

Questions ?

Nov 2025

Red Hat Summit

Connect

Grazie

in linkedin.com/company/red-hat

f facebook.com/redhatinc

youtube.com/user/RedHatVideos

twitter.com/RedHat