Where to Run AI?

Factors to Consider

Derrick Monahan

Principal Solutions Architect
AI & High-Performance Architecture

Jason Campagna

Senior Director - AI Practice Strategy & GTM

World Wide Technology

Our Collective Point of View is Based on Real AI Work

For Customers

Since 2014, our AI & Data Science teams have delivered **50+ AI programs**

SELECT EXAMPLES

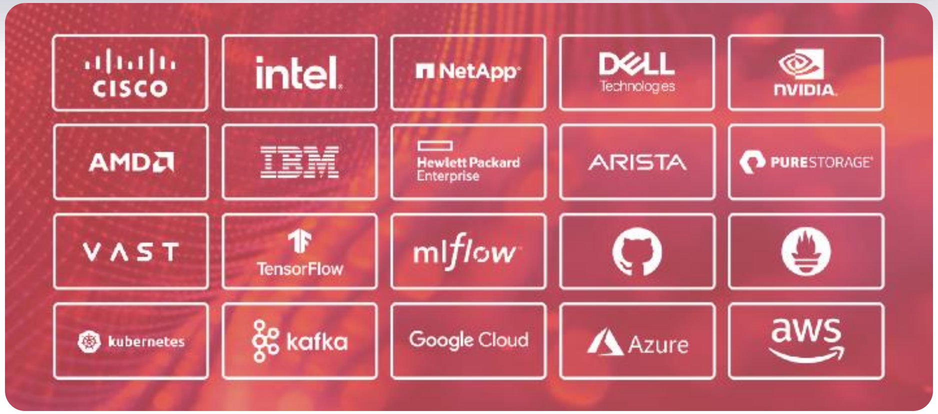
- Global Financials
- Life Sciences
- Manufacturing
- Healthcare
- Retail/QSR
- Academia
- Utilities & Mining
- Telecom & Media
- Government & Public Sector

With Partners

Cross-OEM AI testing ground, with access to the latest AI technologies

WWT AI Proving Ground

- Foundational data capabilities
- Generative AI and deep learning
- Edge compute and AI inference



For Employees

We are generating **significant productivity enhancements internally**

EARLY FOCUS

WWT AI Hub
Internally Focused AI Platform

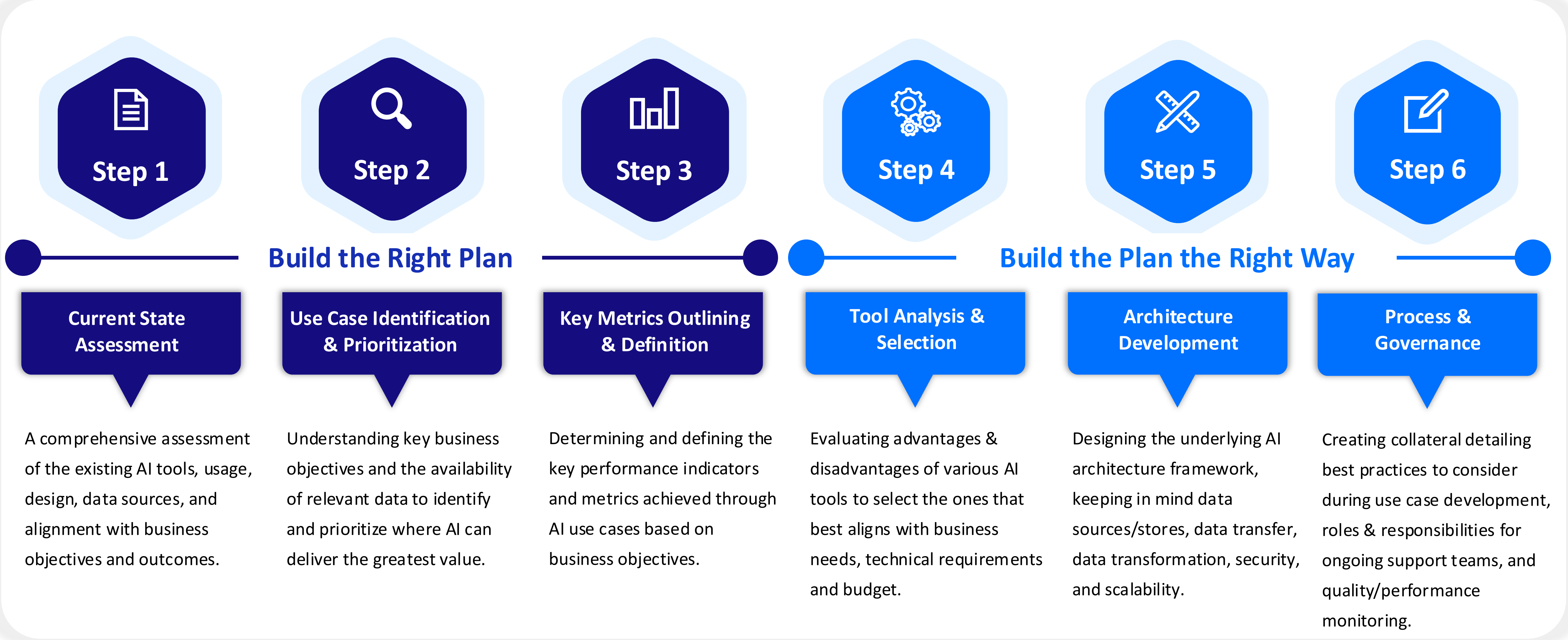
- **Scripting Copilot**
- **Proposal Assistant**
- **Atom (WWT GPT)**

BACKLOG: 50+ Product Ideas

200+ Customer AI Engagements

More than a technology challenge, it’s a business opportunity

Process to delivering an effective AI Strategy



Build the Right Plan

Step 1

Current State Assessment

A comprehensive assessment of the existing AI tools, usage, design, data sources, and alignment with business objectives and outcomes.

Step 2

Use Case Identification & Prioritization

Understanding key business objectives and the availability of relevant data to identify and prioritize where AI can deliver the greatest value.

Step 3

Key Metrics Outlining & Definition

Determining and defining the key performance indicators and metrics achieved through AI use cases based on business objectives.

Build the Plan the Right Way

Step 4

Tool Analysis & Selection

Evaluating advantages & disadvantages of various AI tools to select the ones that best aligns with business needs, technical requirements and budget.

Step 5

Architecture Development

Designing the underlying AI architecture framework, keeping in mind data sources/stores, data transfer, data transformation, security, and scalability.

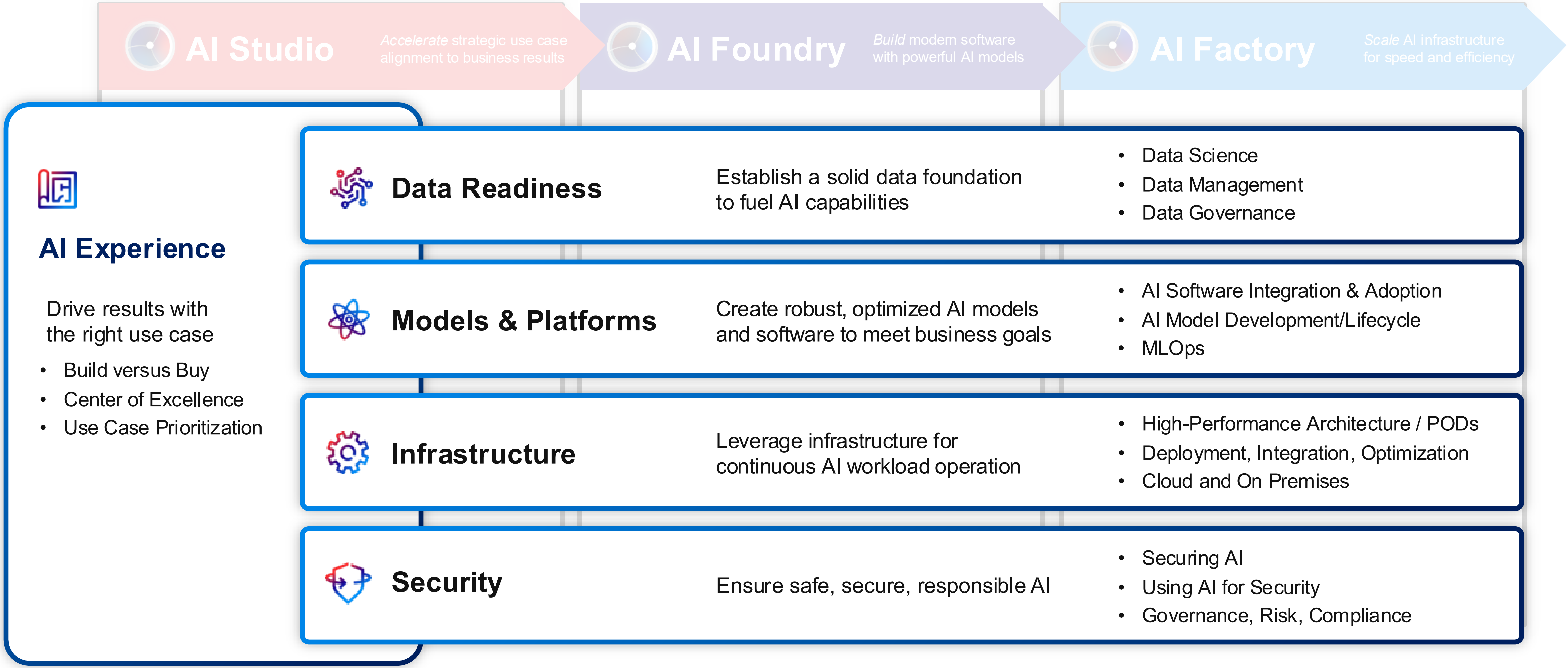
Step 6

Process & Governance

Creating collateral detailing best practices to consider during use case development, roles & responsibilities for ongoing support teams, and quality/performance monitoring.

Technology Stack Layers For AI

Your AI Journey Simplified: Accelerate, Build and Scale with Purpose-Driven Impact



AI Studio — *Accelerate* strategic use case alignment to business results

AI Foundry — *Build* modern software with powerful AI models

AI Factory — *Scale* AI infrastructure for speed and efficiency

AI Experience

Drive results with the right use case

- Build versus Buy
- Center of Excellence
- Use Case Prioritization

Data Readiness

Establish a solid data foundation to fuel AI capabilities

- Data Science
- Data Management
- Data Governance

Models & Platforms

Create robust, optimized AI models and software to meet business goals

- AI Software Integration & Adoption
- AI Model Development/Lifecycle
- MLOps

Infrastructure

Leverage infrastructure for continuous AI workload operation

- High-Performance Architecture / PODs
- Deployment, Integration, Optimization
- Cloud and On Premises

Security

Ensure safe, secure, responsible AI

- Securing AI
- Using AI for Security
- Governance, Risk, Compliance

Customer Challenges: Building AI Solutions

It could take organizations anywhere between 7 - 12 months to operationalize AI/ML from concept to deployment.

Optionality

Need choices, NVIDIA, AMD, Intel Accelerators, etc. GPUs (Graphics Processing Units), TPUs (Tensor Processing Units), ASICs (Application-Specific Integrated Circuits), CPUs

Vendor Lock-In

Need Flexibility, Vendor lock-in is a major concern for businesses adopting AI, as it can limit flexibility and increase costs in the long run.

Scalability

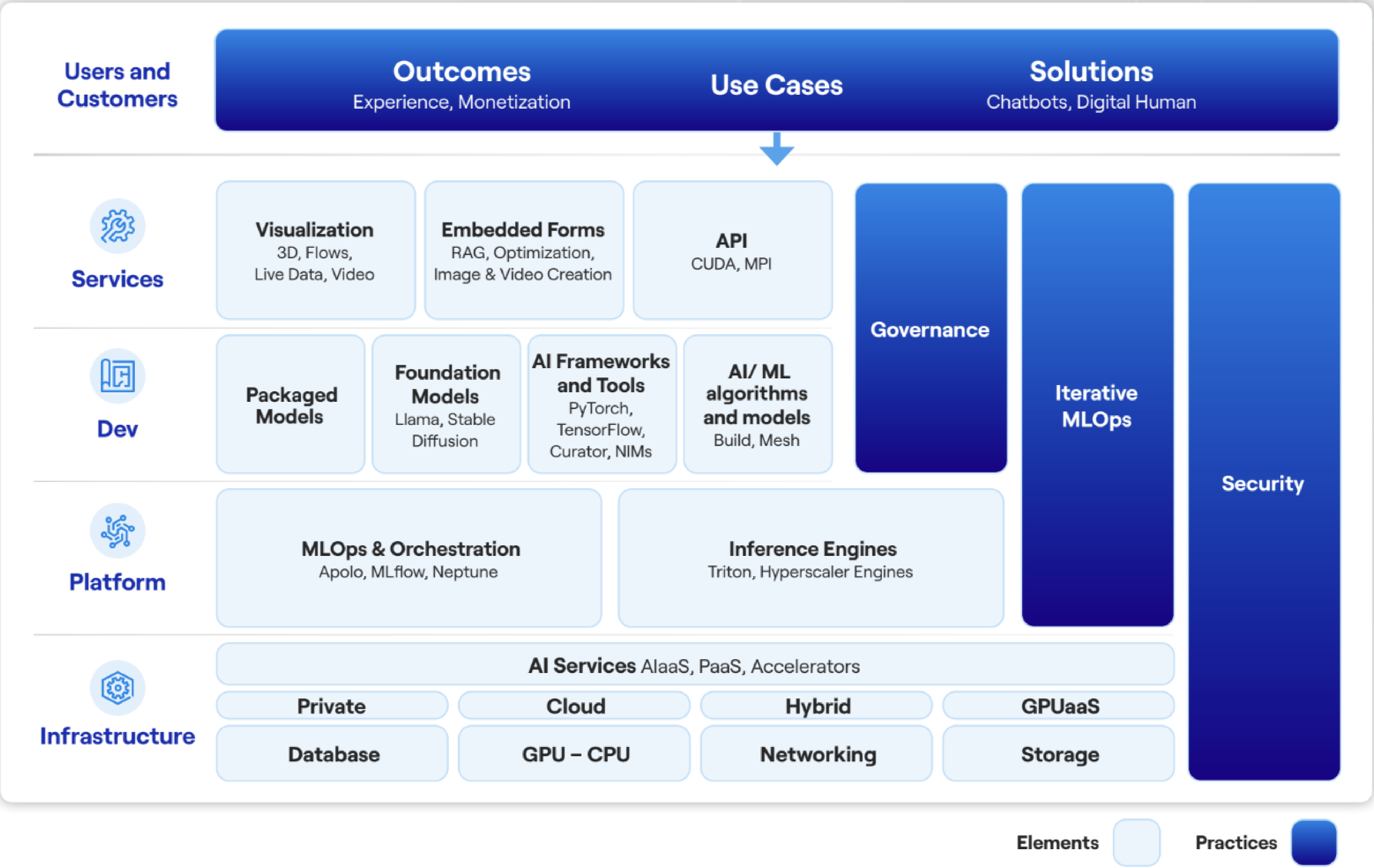
Need Scale, AI workloads can be highly demanding. Businesses need compute solutions that can scale seamlessly to handle increasing data volumes and model complexity.

High Cost

Need formalized TCO, Containing costs of infrastructure is imperative. The high cost of AI compute resources, including specialized hardware and software, can be a barrier to entry for many organizations.

Operational Complexity

Need Operationally Consistent Capabilities, Deploying and managing AI infrastructure can be complex, requiring specialized expertise and tools



Users and Customers

Outcomes Experience, Monetization

Use Cases

Solutions Chatbots, Digital Human

Services

- **Visualization** 3D, Flows, Live Data, Video
- **Embedded Forms** RAG, Optimization, Image & Video Creation
- **API** CUDA, MPI

Dev

- **Packaged Models**
- **Foundation Models** Llama, Stable Diffusion
- **AI Frameworks and Tools** PyTorch, TensorFlow, Curator, NIMs
- **AI/ ML algorithms and models** Build, Mesh

Platform

- **MLOps & Orchestration** Apolo, MLflow, Neptune
- **Inference Engines** Triton, Hyperscaler Engines

Infrastructure

AI Services AIaaS, PaaS, Accelerators

Private	Cloud	Hybrid	GPUaaS
Database	GPU – CPU	Networking	Storage

Governance

Iterative MLOps

Security

Elements □ Practices ■

AI Adoption: Matching Your Deployment Strategy to Your Needs

Data Sensitivity & Control

If the AI model processes highly sensitive or regulated data, on-premises deployment offers a high level of control and security. Compliance requirements also a factor

Data Gravity & Integration

AI models are data-hungry. The sheer volume of data can make it impractical or expensive to move. AI workloads are often deployed close to where the data resides. 'Data Gravity' is a powerful force in deployment decisions

Facilities Infrastructure

AI infrastructure requires significant power and cooling to support high-performance computing, with space considerations crucial for accommodating specialized hw and efficient airflow.

Time to Market

Buying can significantly accelerate the deployment of AI capabilities, as pre-built solutions are readily available. Building can take longer due to development and customization.

Performance & Latency

On-premises AI avoids network latency issues, critical for real-time applications (e.g., industrial automation, high-frequency trading).

Cost

Technical Expertise

Visualizing the Cost of GenAI

Application implementation and maintenance cost

Number of systems enhanced/built

- Custom app build
- Mod existing systems

+

Cost of models

Number and size of model (parameters), if fine-tuned, custom

- AI models
- Model tuning

+

Cost of running AI tech and inference infrastructure

Cloud, HW, Ops required

- AI infrastructure
- Cloud infrastructure

+

Cost of data

Amount and source of data

- Your data
- Data engineering
- Vendor/ public data
- Data management infrastructure

+

User cost

Number of users

- Training cost
- Per user license

×

Consumption cost/ inference RAG

Size and iterations of prompts/responses/ number of API calls

- Prompt input size
- Answer output size

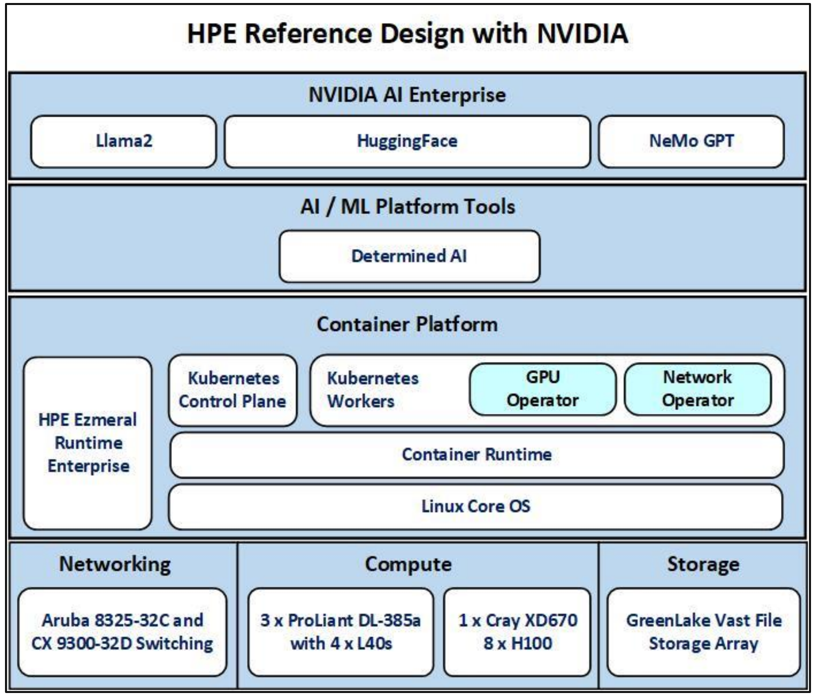
×

- Number of iterations to answer
- Average price per token

Hybrid AI

Driven by AI Workload Placement

	On Prem - HPA	Private Cloud (Co-Lo)	GPUaaS	Public Cloud
Testing Workloads	● ● ●			●
Inferencing Workloads	● ● ●			●



HPE Reference Design with NVIDIA

NVIDIA AI Enterprise

Llama2 | HuggingFace | NeMo GPT

AI / ML Platform Tools

Determined AI

Container Platform

HPE Ezmeral Runtime Enterprise

Kubernetes Control Plane

Kubernetes Workers | GPU Operator | Network Operator

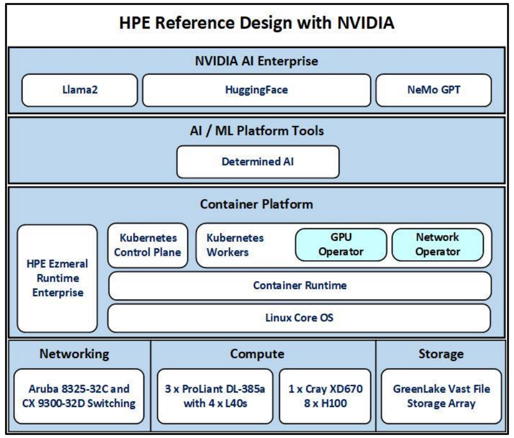
Container Runtime

Linux Core OS

Networking: Aruba 8325-32C and CX 9300-32D Switching

Compute: 3 x ProLiant DL-385a with 4 x L40s | 1 x Cray XD670 8 x H100

Storage: GreenLake Vast File Storage Array



HPE Reference Design with NVIDIA

NVIDIA AI Enterprise

Llama2 | HuggingFace | NeMo GPT

AI / ML Platform Tools

Determined AI

Container Platform

HPE Ezmeral Runtime Enterprise

Kubernetes Control Plane

Kubernetes Workers | GPU Operator | Network Operator

Container Runtime

Linux Core OS

Networking: Aruba 8325-32C and CX 9300-32D Switching

Compute: 3 x ProLiant DL-385a with 4 x L40s | 1 x Cray XD670 8 x H100

Storage: GreenLake Vast File Storage Array

EQUINIX

DIGITAL REALTY

SCOTTDATA center

Managed Services

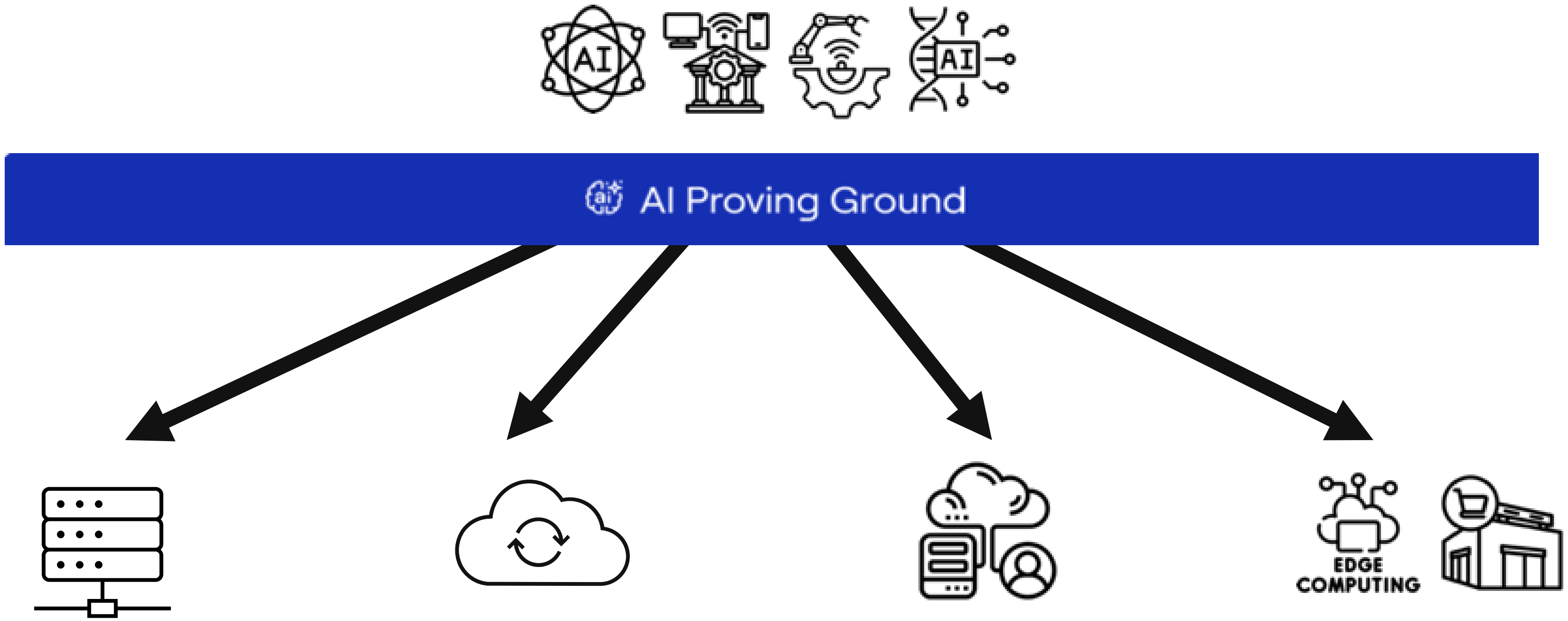
Google Cloud

Azure

aws

Architectural Evaluation Process

Ability to Develop and Deploy AI anywhere



AI Proving Ground

On-Premises

Cloud

Hybrid AI Cloud

Edge

Data Locality

Data Gravity challenges are further compounded by AI. Location of data varies as data processing can occur at the edge, model training, or being used for production applications

Data Sovereignty

Adhering to regulatory requirements governing data storage and processing.

Hybrid AI Strategies

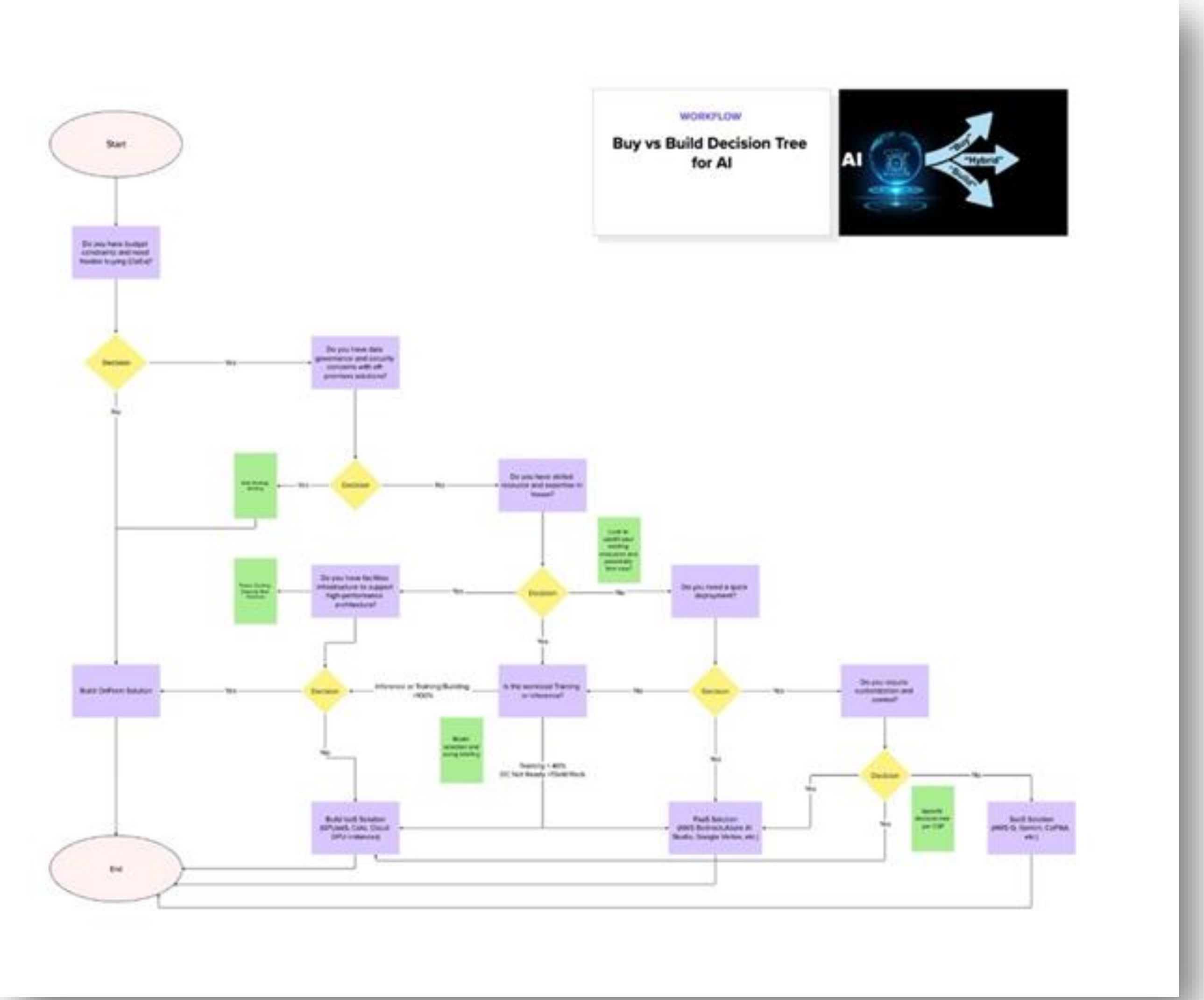
Combined approaches optimizing cost, performance, and security to leverage best-of-breed AI solutions,

Real-Time Performance

Supporting real-time analytics and insights that rely on sensor-generated data or that must react in real-time

Decision Tree

WWT Example - Where to Run AI


WORKFLOW
Buy vs Build Decision Tree for AI
AI
"Buy"
"Hybrid"
"Build"
Start
Do you have budget constraints and need flexible buying (OpEx)?
Decision
Yes
No
Do you have data governance and security concerns with off premises solutions?
Do you have skilled resources and expertise in house?
Do you have facilities infrastructure to support high-performance architecture?
Do you need a quick deployment?
Build OnPrem Solution
Inference or Training/Building >100%
Is the workload Training or Inference?
Do you require customization and control?
Model selection and sizing
Build IaaS Solution (GPUaaS, Colo, Cloud GPU instances)
PaaS Solution (AWS Bedrock, Azure AI Studio, Google Vertex, etc.)
SaaS Solution (AWS Q, Gemini, CoPilot, etc.)
End

Simplifies a complex landscape:

- AI can be overwhelming
- Decision trees make it easier to navigate the options

Exploring AI Approaches (HPA Workshop)

- The decision tree captures critical input using a weighted scoring model to explore best AI architecture approaches

Providing Recommendations:

- **Tailored solutions:** By following the decision tree, the customer arrives at a specific AI solution or a narrowed-down set of options that best fit their needs

Four key building blocks for High Performance Architecture (HPA)

Compute

HPC / supercomputing

Accelerated computing

Heterogenous computing

Emergent computing

Quantum computing

Storage

Parallel file system storage

Streaming storage

Data Platforms

Synthetic data

Computational storage

Emergent storage

Network

Connects users and infrastructure

Secure, smart, fast fabrics

SmartNICs and DPUs

Computational networking

Photonics (SOC, switches, backplanes)

Orchestration & AI Workflow

AI and Data Science Tools and Frameworks

Cloud-Native Management and Orchestration

Infrastructure Optimization

Cluster Management

Platform and MLOps

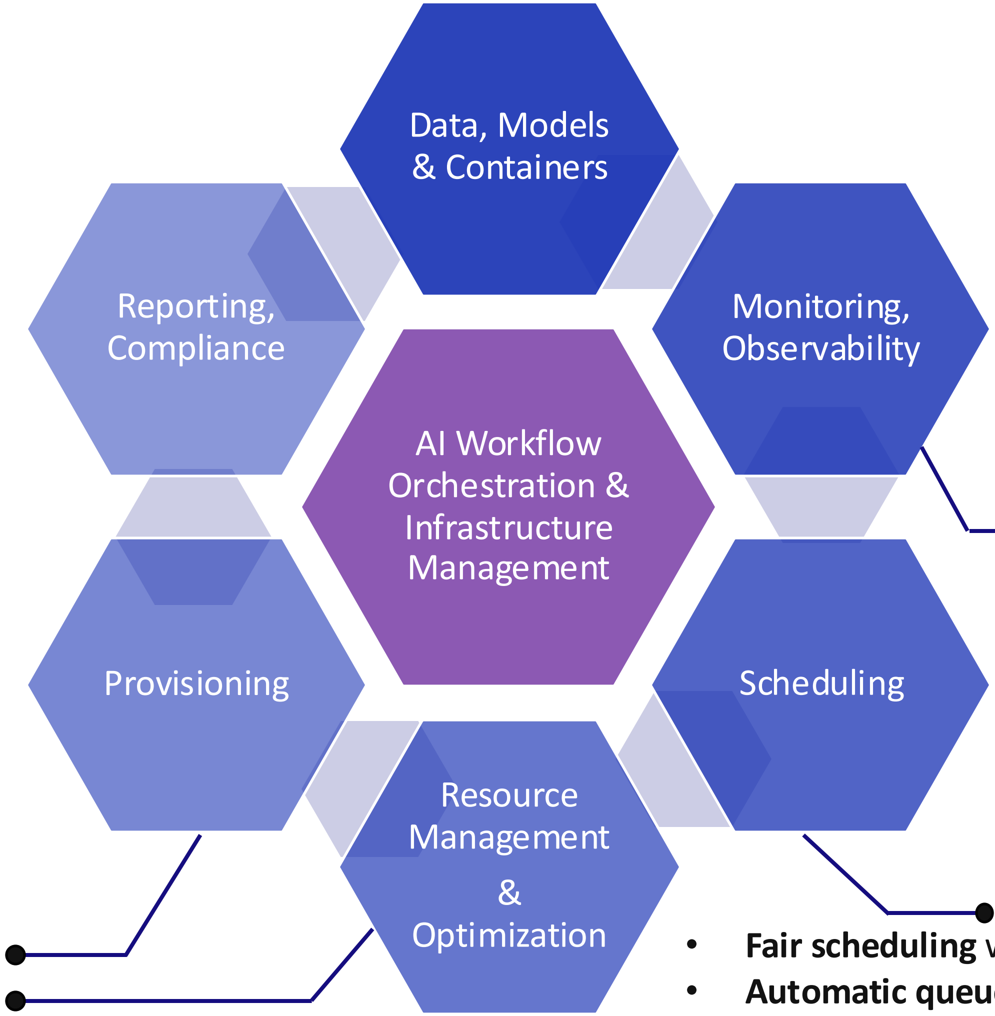
AI Workflow Orchestration & Infrastructure Management

Intersects with multiple AI disciplines

Low GPU Utilization

IT Management Complexity

AI NOT Getting Into Production


Data, Models & Containers
Reporting, Compliance
Monitoring, Observability
AI Workflow Orchestration & Infrastructure Management
Provisioning
Scheduling
Resource Management & Optimization

Fabric Visibility and Control

- Real-time network telemetry information
- Automated network discovery and validation
- Automated Anomaly Detection
- Automated Log Analysis
- Congestion tracking to identify traffic bottlenecks
- System Configuration and Validation
- Network performance tests
- Application workload usage

- **Isolated GPU Fractions**
- Increases utilization of GPU compute
  - (OnPrem, Cloud & Edge workloads)
- **Dynamic MIG Partitioning**
- **GPU Memory Overprovisioning**
- IT gains visibility and control to maximize GPU utilization

- **Fair scheduling** with management of multiple scheduling queues
- **Automatic queueing or dequeuing** of workloads – policy based
- **Gang scheduling -** efficient mgmt of multi-node distributed workloads
- Administrators gain control - align resources with business goals

Infrastructure & AI Workflow

AI and Data Science Tools and Frameworks

Cloud-Native Management and Orchestration

Infrastructure Optimization

Cluster Management

Foundational Components

Build Approach

Architecture Choice

- Requires upfront investment in GPUs, CPUs, memory, storage, and networking.
- Well-established Reference Architectures
- Tailor the hardware configuration to the specific requirements of your AI workloads

Customized Software Stack

- Operating system, data management platforms, machine learning frameworks, and MLOps tools
- Provide greater flexibility and control over your AI development and deployment process

Expert AI engineers & Data scientists

- Strong programming and DevOps skills
- Deep understanding of hardware and software infrastructure.
- Optimize and Monitor infrastructure health and troubleshoot issues.

Key Components of a Hybrid Approach

Optimal results obtained with a balance of speed, cost, and control

Pre-Built AI Platforms

- Accelerate with pre-trained models
- APIs for connectivity across several systems
- **Examples**:
  - Sagemaker
  - Vertex AI
  - Azure ML
  - Watson Studio
  - PyTorch
  - TensorFlow

Customize and Extend

- **Customize:**
  - Model selection
  - Data preparation
  - Integrated with existing systems
- **Extend:**
  - Add features such as NLP, computer vision, etc.
  - Scale and burst workloads
  - Deploy models near data

Leverage Enterprise Data

- Maintain governance, access controls, and security of your data
- Control bias and ethics within AI models
- **Example Solutions:**
  - Personalized recommendations
  - Drug discovery
  - Risk assessment
  - Predictive maintenance
  - Customer service chatbots

Foundational Concepts

Hybrid Approach

Mix of Infrastructure

- Workload placement and dynamic resource allocation
- Well established platforms in the public cloud
- Private cloud a must with connectivity, security, and access controls established

Systems of Systems

- APIs and cloud-native container platforms integration
- Modular architecture with standard interfaces/protocols
- Execution of multiple AI models to optimize performance + accuracy

Data Governance and Integration

- Data sovereignty opportunities with an established governance framework
- Must maintain a framework to collect, process, and manage data
- Continuous monitoring and evaluation

Key Takeaways of a Hybrid Approach

- Flexibility and scalability
- Cost optimization
- Data residency and compliance
- Operational Efficiency and Success
- Innovation and agility

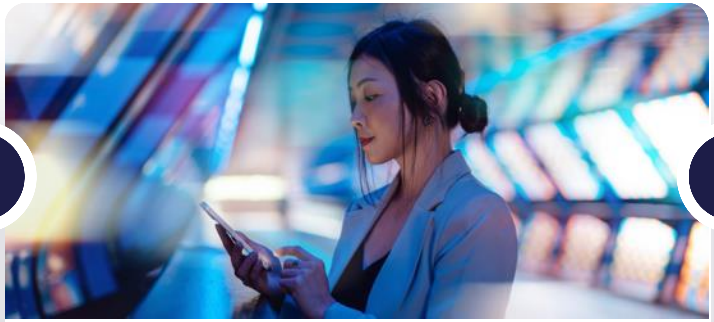
Accelerate AI Outcomes: Our Practical Approach

Simplifying the AI journey: Accelerate, build and scale with purpose-driven impact

AI Studio

Accelerate strategic alignment and business value

+



AI Foundry

Build AI applications rapidly with powerful models

+

AI Factory

Scale AI infrastructure for speed and efficiency

Rapidly achieve business impact with the right AI experiences

Business ROI Validation

Use Case Validation

Strategy & Roadmap

Center of Excellence

Workload Sizing

Build vs. Buy

Rapid Prototyping

Data Readiness

AI Security

High-Performance Architecture

Agentic Platforms

SaaS Solutions

Automation

Optimized Deployment

AI Operations

Thank You!

Questions

World Wide Technology