TD SYNNEX

RHEL AI und die Plattform Ihrer Wahl

Red Hat

Authorized Distributor

Sensitivity: Public

Date	Author	Version
07.05.2025	Sejid Canoski	1.0

Todays Agenda

RHEL AI and your platform of choice

- Current state of AI …
- Consideration: IBM Power platform
- Client-Use Cases | Examples

Foundation models are bringing an inflection point in AI...

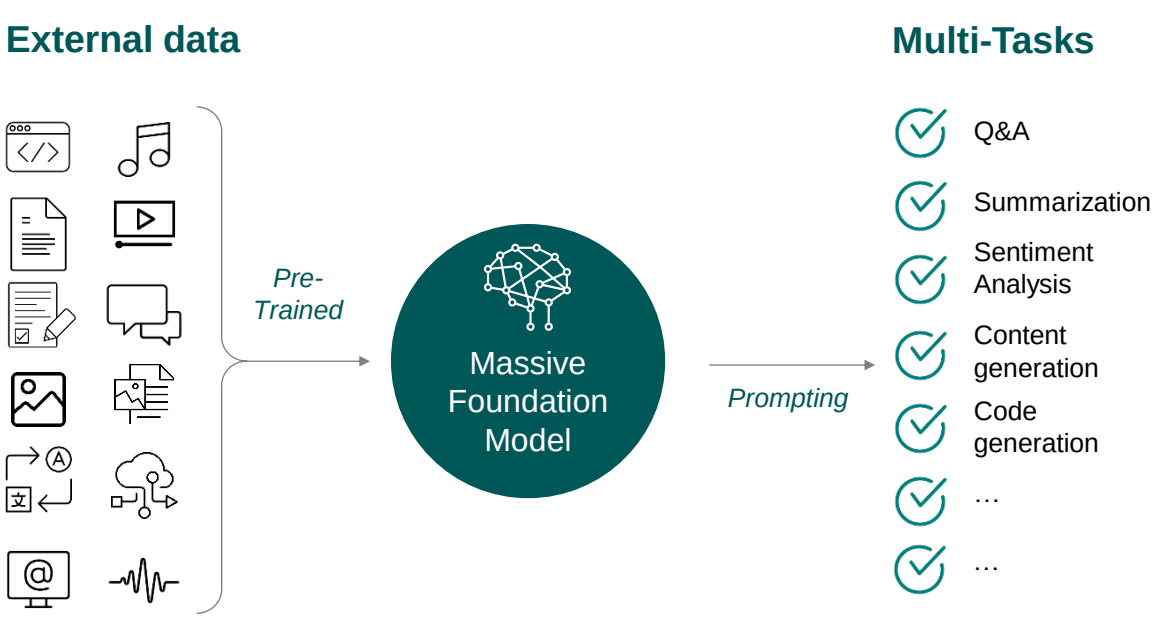
64%

of executives believe that generative AI is closing the gap between IT and the business*

...but how enterprises adopt and execute will define whether they unlock value at scale.

*IBM Institute for Business Value (IBV): The CEOs Guide to Generative AI

Foundation models …



Key advantages:

- Lower upfront costs through less labeling
- Faster deployment through fine tuning; inferencing
- Equal or better accuracy for multiple use cases
- Incremental revenue through better performance

70%

reduction in time for NLP tasks*

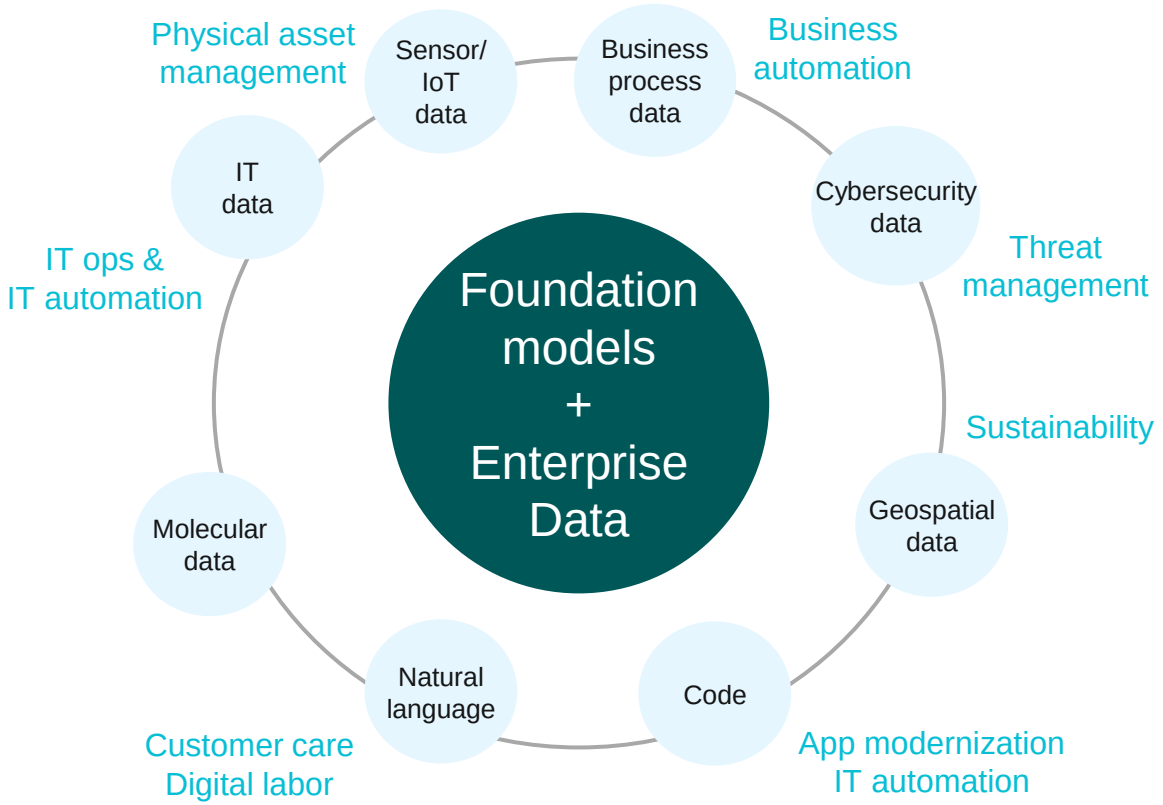
64%

of executives believe that generative AI is closing the gap between IT and the business**

*Client engagements with IBM Consulting: IBM blog
**IBM Institute for Business Value (IBV): The CEOs Guide to Generative AI

…fueled by enterprise data

"Enterprises that have large stores of high-quality data, monetize data effectively, and say their data is trusted by internal and external stakeholders **realize almost double the ROI from their AI capabilities (9% versus 4.8% for all others)"***



*IBM Institute for Business Value (IBV): The CEOs Guide to Generative AI

Enterprise considerations: Five truths of generative AI

Truth #1
Multi-model

Two thirds of 150+ enterprises surveyed report pursuing a multi-model strategy

- 60% + of enterprises pursuing multi-model are experimental with commercial & open-source models
- Commercial & open-source innovation
- Quickly prioritize use cases that will outlive the model
- Multi-modal (text, image, audio, etc.)
- One model will not rule them all

Truth #2
Multi | hybrid cloud

Gartner reports that most enterprises will deploy generative AI across hybrid/multi-cloud environments

- Run where the workflows, apps and data live
- Infer where business runs to drive performance, cost, and simplicity
- Data location to drive security benefits
- Regulatory compliance to influence location selection

Truth #3
Governance

Surveyed companies report governance as a top requirement, impact of generative AI makes governance more difficult

- Businesses must control bias and monitor drift
- Organizations must actively monitor hallucinations and ensure model explainability
- Leaders must seek practices and tools to ensure model and data provenance

Truth #4
Scale for value

Critical to pick the right use cases and deployment for generative AI ROI

- Different work tasks have strongly positive or negative ROI impact
- Time savings for a meaningful product innovation +40%; business problem solving - 23% time needed
- 60+ points difference in value for work tasks
- 25x difference in cost per inference, depending on model and deployment

Truth #5
Data matters

Two thirds of 150+ enterprises surveyed report pursuing a multi-model strategy

- **Short run**: model innovation creates value
- **Long run**: data quality will decide which enterprises win with generative AI

Enterprise considerations:

Traditional AI is still critical, Gen AI unlocks new possibilities

Generative AI

Foundation models majorly trained with ‘unlabeled’ data

Trained on very big data sets (100s of TBs)

Trained for no specific tasks

Works well for general tasks and can improve for specific tasks with relatively less training

Traditional AI

Traditional AI (ML/DL) models majorly trained with ‘labeled’ data

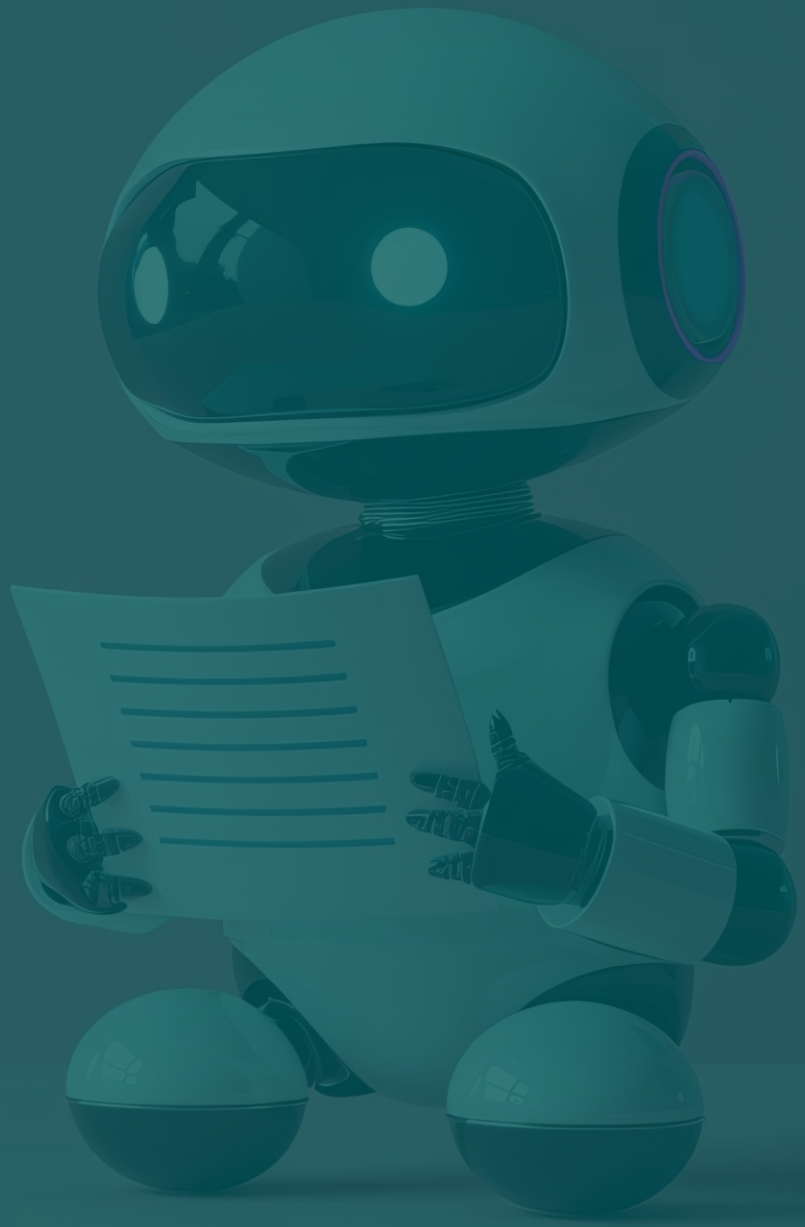
Trained on proper, large data sets (100s of GBs)

Trained for a specific task

A tuned model can be very efficient for the specific task it was designed for

TD SYNNEX

Where can “the artificial intelligence” process things?



TD SYNNEX

The *seemingly* complex world of processing

CPU

Central Processing Unit

The CPU, often called the "brain" of a computer, is responsible for performing general-purpose tasks. It executes instructions from applications, manages hardware, and handles data processing.

- Sequential Processing
- Versatility
- Cores and Threads

GPU

Graphics Processing Unit

The GPU is a specialized processor designed to handle parallel processing tasks, making it ideal for rendering graphics and processing large datasets simultaneously.

- Parallelism
- High Performance for Graphics
- Compute Acceleration

- CUDA and OpenCL

NPU

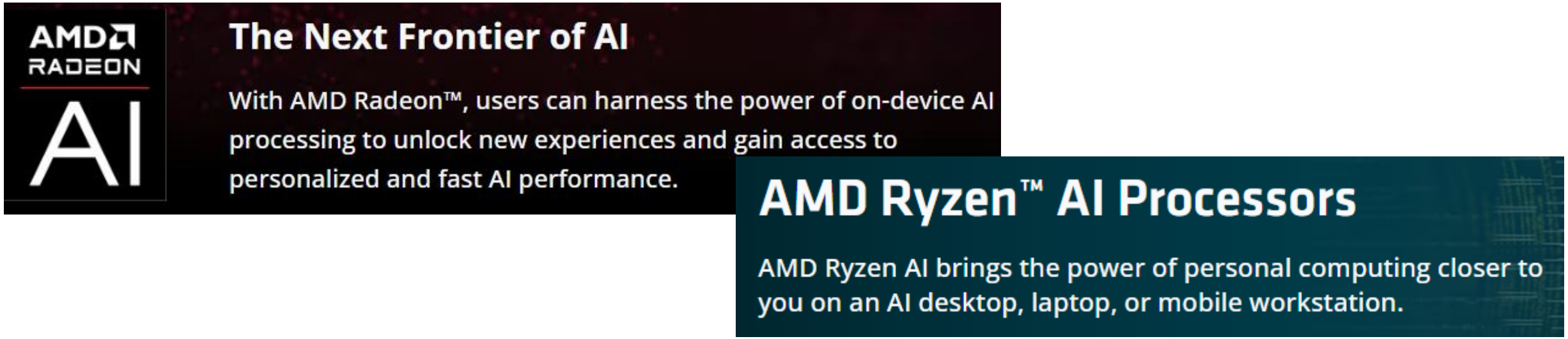
Neural Processing Unit

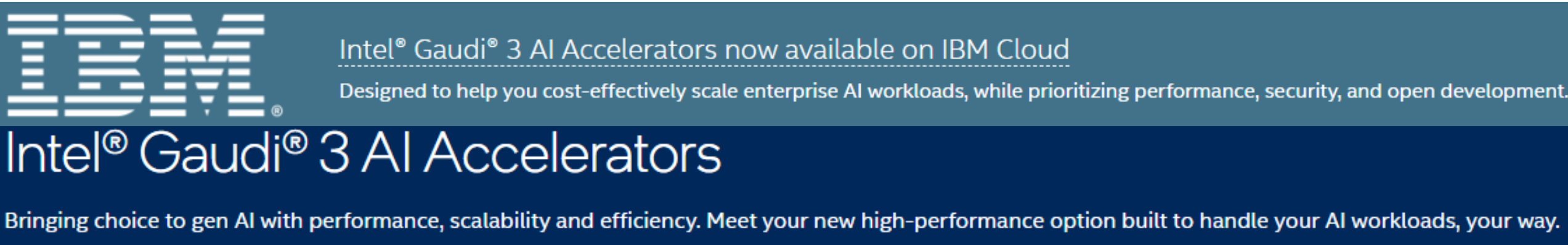
An NPU is a processor specifically designed to accelerate artificial intelligence (AI) and machine learning (ML) tasks.

- AI Optimization
- Energy Efficiency
- Matrix Multiplications

TD SYNNEX

Vendors providing the compute to fuel artificial intelligence





TD SYNNEX

ARM technology to fuel AI workloads

The Arm CPU already provides the foundation for accelerated AI globally. The combination of Arm's pervasiveness, performance at scale, energy-efficiency, security, and developer flexibility, means the number of AI-enabled chips shipped each quarter is increasing exponentially.

Generative AI at Scale

The arrival of sophisticated generative AI models has sparked a new wave of innovation, offering an opportunity for technology step change. Arm's compute platform offers the efficient and highest-performing capabilities, which enable GenAI to run on phones and PCs, and in datacenters.

AI Inference on CPU

With the growing number of AI applications, there comes an exponential increase in the need for AI inference capabilities. Arm CPUs provide the technology foundations for inference to run on all compute touchpoints, bringing AI into the hands of billions of people around the world.

TD SYNNEX

IBM’s – Red Hat POV for AI-ready IT Infrastructure.

Reliable performance

Scale AI inferencing for complex tasks like generative AI.

Hybrid flexibility

Create AI workflows based on where your data and applications reside.

Secured insights

Offer security and data protection to promote trust and support compliance demands.

IBM Power is built for AI

Accelerate AI efficiently

Leverage AI-optimized hard- & software

- Run AI models on a highly performant, sustainable platform
- Simplify solution architectures & save costs
- Scale AI solutions with a specialized ecosystem

Orchestrate AI flexibly

Create & run AI where and how needed

- Use hybrid cloud infrastructure seamlessly
- Consume resources elastically
- Combine enterprise & open-source AI software

Safeguard AI and Data

Ensure full-stack security for AI and data

- Minimize exposure & risks by converging AI with data
- Secure AI workloads across all stack layers
- Protect data through accelerated encryption

TD SYNNEX

42%

more inferencing throughput vs. compared x86 servers when using Large Language Models[1]

< 1s

inferencing latency on IBM Power10 when using a Q&A Large Language Model[2]

51%

lower 3-year TCO running highly parallel inferencing with IBM Cloud Pak for Data on IBM Power S1022 vs. compared x86 servers[3]

39%

more inferencing per watt with IBM Cloud Pak for Data on IBM Power S1022 vs compared x86 servers[3]

1 Based on IBM internal testing of question and answer inferencing using PrimeQA model (based on Dr. Decr and ColBERT models). Results valid as of Aug 22, 2023, and conducted under laboratory conditions, individual results can vary based on workload size, use of storage subsystems and other conditions. Comparison is based on total throughput in score (inferences) per second on IBM Power S1022 (1x20-core/512GB) versus Intel Xeon Platinum 8468V-based (1x48-core/512GB) systems. Test was run with Python and Anaconda environments including packages of Python 3.9 and PyTorch 2.0. The Python libraries used are platform-optimized for both Power and Intel. Configuration: OMP-NUM-THREADS = 4; batch size = 60. OMP_NUM_THREADS optimized across a variety of load levels.
IBM S1022 Power system: https://www.redbooks.ibm.com/abstracts/redp5675.html
Compared x86 system:
Supermicro SYS-221H-TNR system with x86 AME/AMX AI accelerators: https://www.supermicro.com/en/products/system/hyper/2u/sys-221h-tnr
PrimeQA models: https://github.com/primeqa
Models fine-tuned by IBM on a corpus of IBM-internal data

2 Based on IBM internal testing of question and answer inferencing using PrimeQA models (based on Dr. Decr and ColBERT models). Results valid as of Aug 22, 2023, and conducted under laboratory conditions, individual results can vary based on workload size, use of storage subsystems and other conditions. Estimated result for an IBM Power S1024 (2x12-core 3.4-4GHz/512GB) is based on a measured inferences time of 1.008 seconds on IBM Power S1022 (2x12-core 2.9-4GHz/512GB). Test was run with Python and Anaconda environments including packages of Python 3.9 and PyTorch 2.0. The Python libraries used are platform-optimized libraries for Power. Configuration: OMP-NUM-THREADS = 32; batch size = 1.
IBM S1024 Power system: https://www.redbooks.ibm.com/abstracts/redp5675.html
PrimeQA models: https://github.com/primeqa
Models fine-tuned by IBM on a corpus of IBM-internal data

3. 1.Based on IBM internal testing of data science components, (WML, WSL, Analytic Engine) of Cloud Pak for Data version 4.8 in OpenShift 4.12. Results valid as of 11/17/2023 and conducted under laboratory condition. Individual results can vary based on workload size, use of storage subsystems & other conditions.
2.The workload mimics a real-time fraud detection logic flow. JMeter is used to submit credit card transactions for different user id and card number combinations. The inferencing application running as microservices in Cloud Pak for Data deployment space extracts the user id and credit card number and uses them to look up 6 previous transactions of the same user and card combination from the Db2 database which is also running within the Cloud Pak for Data cluster. The data retrieved from the database is then combined with the new entry and pass to the LSTM model to determine whether the latest transaction is fraud or not. The score (value between 0 to 1) is returned to the JMeter client as an indicator of whether that transaction is likely a fraud or not.
3.The measurement used for both Power and Intel systems is the throughput result (score/second) reported by JMeter, when running 192 current threads (1 thread representing 1 user) against 96 inferencing end points.
4.Power10 S1022 has a total of 40 physical cores and 2 TB RAM (machine type 9105-22A). There are 7 LPAR on this system including 3 master nodes of 2 cores and 32 GB RAM each, 3 worker nodes of 10 cores and 490 GB RAM each, and a bastion node of 4 cores 128 GB RAM. A local 800 GB NVME drives are used as boot drives for each node, and one 1.6 TB NVMe used for NFS server storage running on the bastion node. There is one 100G Ethernet adapters virtualized through SRIOV, with each LPAR taken 10% of network bandwidth. Each LPAR ran with CPU frequency range 3.20GHz to 4.0GHz. All 3 worker nodes ran in SMT 4 mode, while master and bastion nodes ran in SMT 8 mode.
5.The Intel system is Xeon Platinum 8468V with 96 physical cores and 2 TB RAM. The KVM host takes 2 core and 32 GB RAM, which supports 7 KVM guests on this system, including 3 master nodes of 4 cores and 32 GB RAM each, 3 worker nodes of 24 cores and 490 GB RAM each, and a bastion node of 4 cores 128 GB RAM. Local 1.6 GB NVME drives are used as boot drives for these nodes, and one 1.6TB NVMe used for NFS storage on the bastion node. There is one 100G Ethernet adapters virtualized through SRIOV. Each KVM guest ran with CPU frequency range from 2.40GHz to 3.8GHz. All nodes are RHEL CoreOS KVM guests running on the server with hyperthreading enabled.
Pricing is based on: Power S1022 (see page 4). Typical industry standard Intel x86 (example on page 5) pricing https://www.synnexcorp.com/us/gpo/gds/pricing/ and IBM software pricing available at https://www.ibm.com/downloads/cas/EL8QWBEP.
Assumes energy usage for the Supermicro server is similar to a similarly configured Lenova server (a ThinkSystem SR650 V3) (https://www.supermicro.com/en/products/system/hyper/2u/sys-221h-tnr) , the relative energy usage between Power and x86 system using IDC QPI (https://www.idc.com/about/qpi), is similar for the batch queries workload, and energy usage scales based on the number of batch queries.

TD SYNNEX

Accelerate AI Efficiently with technical innovation

On-Chip AI acceleration

Large Memory & Cache

Highly Parallel

Enterprise Server Design

Per Core: 4x MMA* & 8x SIMD**

- Accelerate Matrix & Vector Math operations in the processor without GPUs
- Minimize data movement from processor to GPU and vice-versa
- Lower quantization for improved performance

TBs Main memory per core

- Store large AI models (like LLMs), large datasets (higher batch sizes), multiple AI models (parallel inferencing) in single memory (1TB-16TB)
- Large cache (4x) to speed up execution

4x more Threads per core

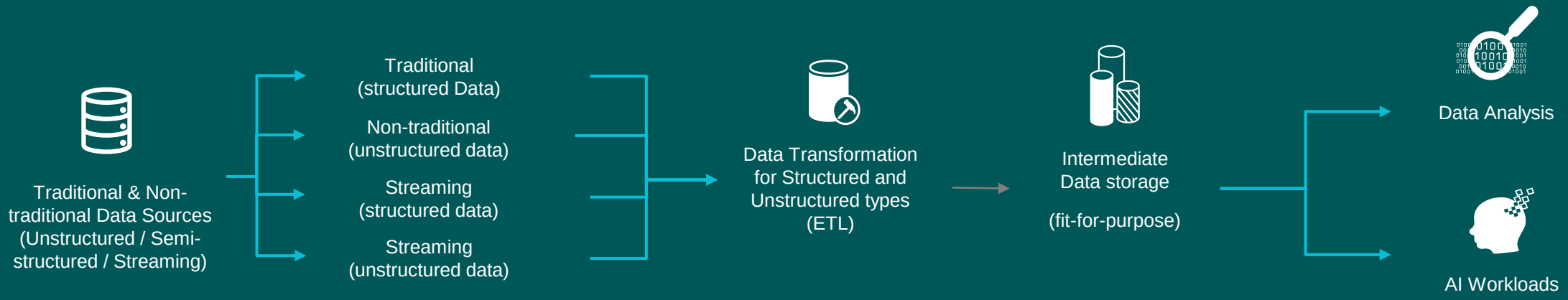
- Provides up to 8 threads per core (4x vs. Intel)
- Run parallel jobs for AI inferencing
- Improve throughput

Lower CO2 footprint

- Same work with lower energy usage
- Consolidate more workloads due to higher guaranteed utilization

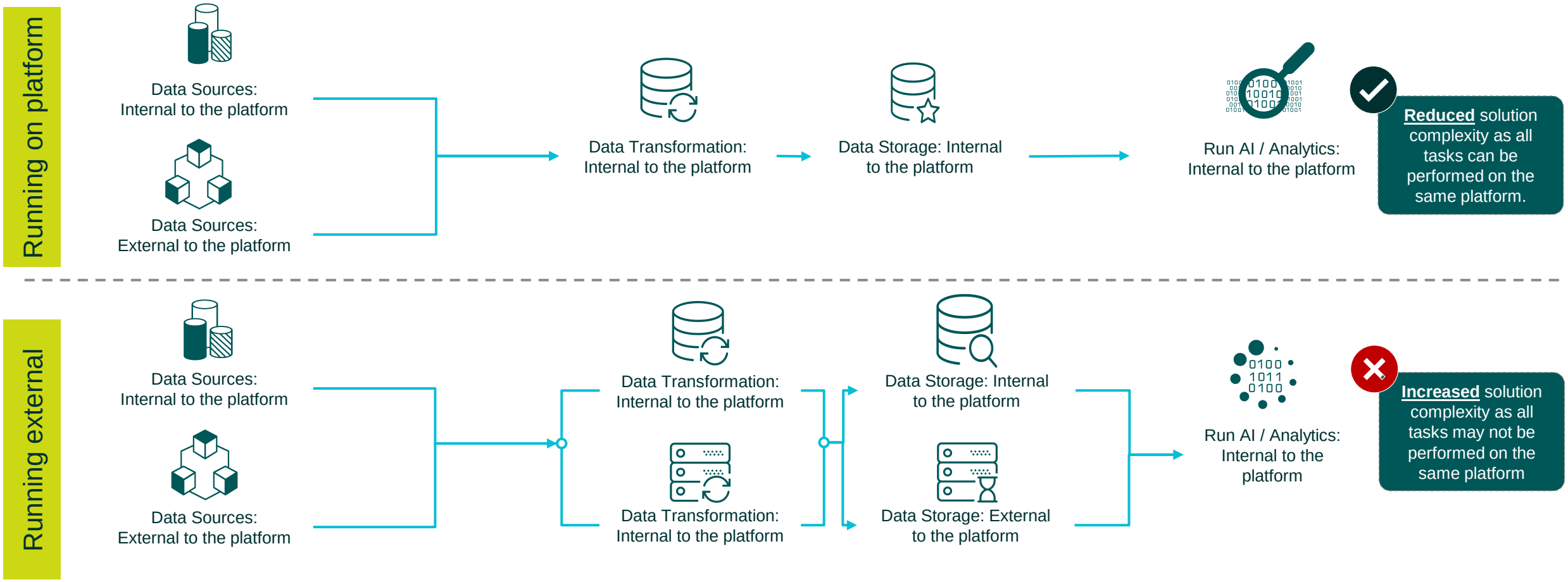
TD SYNNEX

Accelerate AI Efficiently by simplifying solution architecture



Typical AI Workflow

Reduce complexity: run on workloads on the same platform hardware accelerated



Orchestrate AI Flexibly with frictionless hybrid infrastructure

Frictionless Hybrid Experience

AI Solutions from Ecosystem Partners and/or Clients

IBM Enterprise AI (like Cloud Pak for Data or watsonx**)
Rocket AI Hub (Open-source Data and AI platform with support)

Red Hat OpenShift or RHEL

IBM PowerVM

On-Prem IaaS: Private Cloud with Dynamic Capacity

Red Hat OpenShift or RHEL

IBM PowerVM

Off-Prem IaaS: IBM Power Virtual Server in IBM Cloud

IBM Power Infrastructure (On-premise or in the Cloud)

AI Workload
Run on-premise or in the cloud seamlessly

AI Platform Software
Leverage the software across environments

Frictionless Hybrid Infrastructure for AI

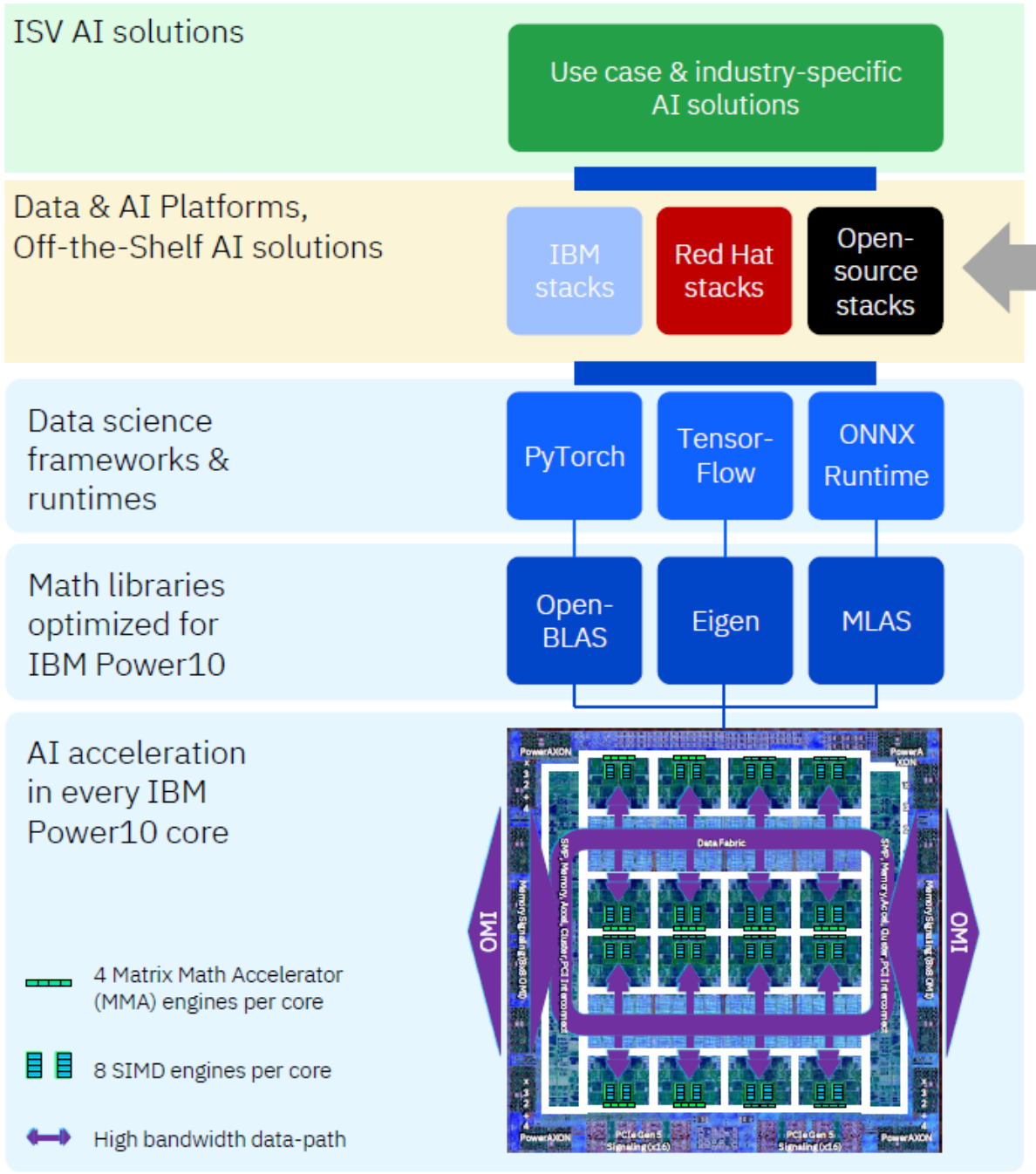
Use the same OS (RHEL or OpenShift) – on-premises or in the cloud

Use the same virtualization technology – on-premises or in the cloud

Use the same hardware platform – on-premises or in the cloud

Full stack optimization

Users get out-of-the-box acceleration without code changes, easing deployments of AI models such as large language models.



TD SYNNEX

AI assistants

Empower individuals to do work without expert knowledge across a variety of business processes and applications.

- **IBM Cloud Pak for AIOps**
- **IBM Cloud Pak for Business Automation**

SDKs & APIs

Embed watsonx platform in third party assistants and applications using programmatic interfaces.

watsonx.ai SDK
→ use foundation models in IBM Power applications (Python, ABAP).

Ecosystem integrations
- Equitus
- Trovares

AI & data platform

Leverage generative AI and machine learning — tuned with your data — with responsibility, transparency, and explainability.

Cloud Pak for Data
- IBM Watson Studio
- IBM Watson Machine Learning
- IBM Data Refinery
- IBM Watson Studio Runtimes

Foundation models

Granite***	*IBM*
Llama 2	*Meta AI*
Geospatial	*IBM + NASA*
…	*Deci, Elinar, Hugging Face, …*

Rocket AI Hub
- Kubeflow*
- KServe*
- RocketCE*
- Jupyter Hub & Jupyter Lab*

Vector Databases
- Milvus**
- Chroma**
- PGVector**
- Weaviate**
- OpenSearch**
- Faiss**

Data services

Define, organize, manage, and deliver trusted data to train and tune AI models with data fabric services.

Data fabric services
- IBM Analytics Engine
- IBM Event Streams
- IBM App Connect Enterprise
- IBM Db2 Warehouse
- IBM MQ
- RH Fuse
- RH AMQ Streams for Apache Kafka
- RH Data Grid
- Apache Spark**
- Apache Kafka**
- trino**

Hybrid cloud AI tools

Build on a consistent, scalable foundation based on open-source technology.

Red Hat OpenShift

Native LPAR deployment options

AI Infrastructure

Accelerated, converge, and safeguard AI efficiently with your data & workflows.

IBM **Power10**

Consulting

Generative AI strategy, experience, technology, operations

Ecosystem

System Integrators, Software and SaaS partners, Public Cloud providers

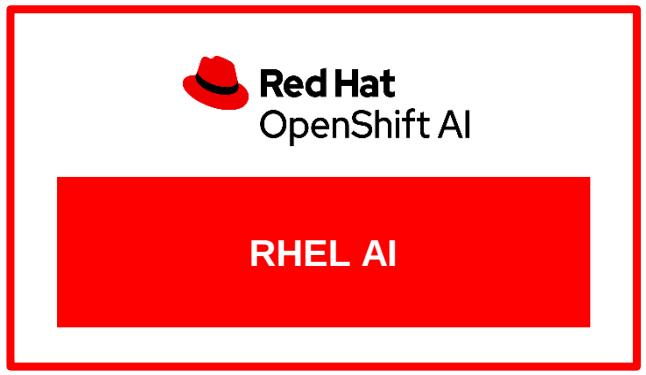
External AI services useful for IBM Power

Ansible Lightspeed
→ Generate playbooks for IBM i & AIX.

SAP
→ Use watsonx-embedded SAP applications with IBM Power.

*: Free-of-charge open-source software; enterprise support options available.
**: Free-of-charge open-source software; currently community-supported.
***: work-in-progress.
*** TBD

The power of community that Red Hat OpenShift AI and RHEL AI offer



Red Hat OpenShift AI

Model serving & monitoring

Deploy models anywhere with consistent cloud-to-edge production deployment, and monitoring capabilities

Resource optimization and management

Create model pipelines for tuning and alignment, validation and delivery; Share and optimize resources across teams

RHEL AI

Foundation model runtime engine

Develop, deploy, and test open Granite language/code models, supported & indemnified from Red Hat and IBM

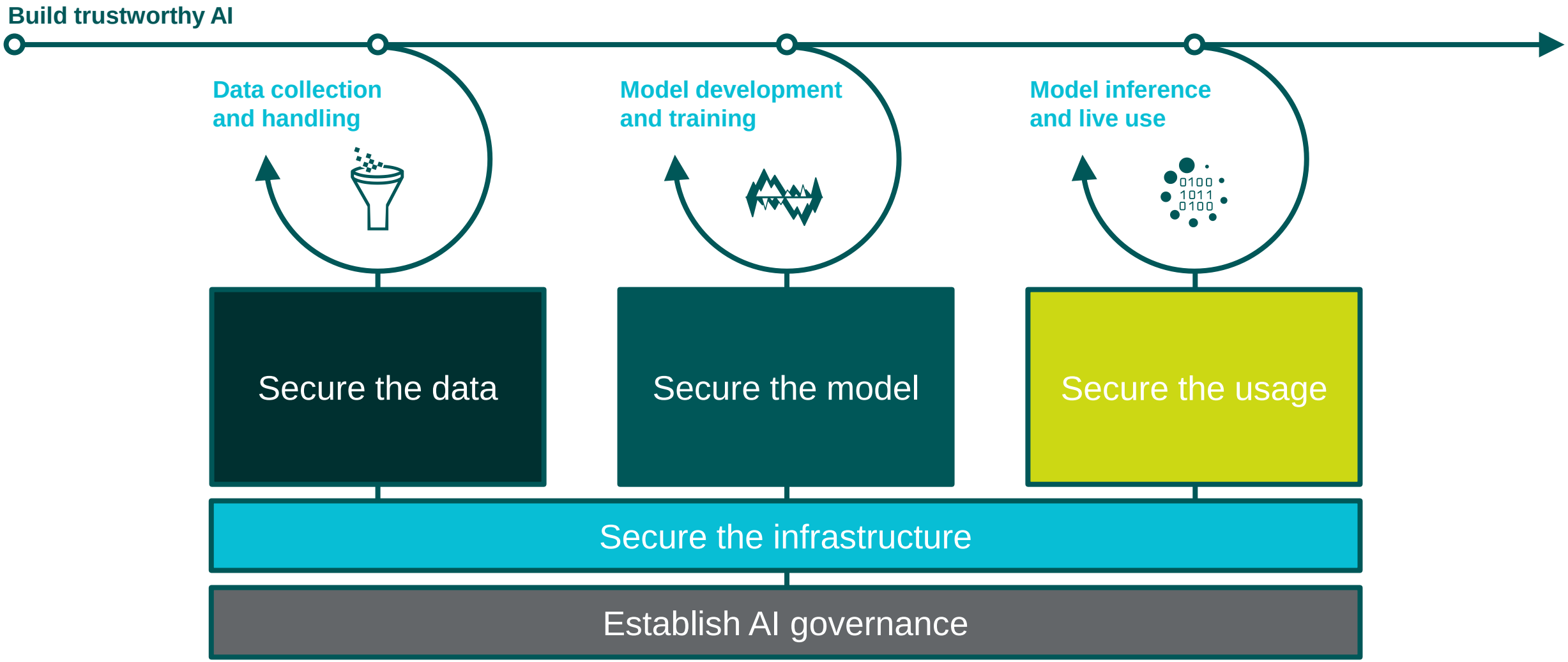
Developer-friendly model toolkit

Bring rapid LLM experimentation to developers via supported InstructLab CLI

Runs on any infrastructure

TD SYNNEX

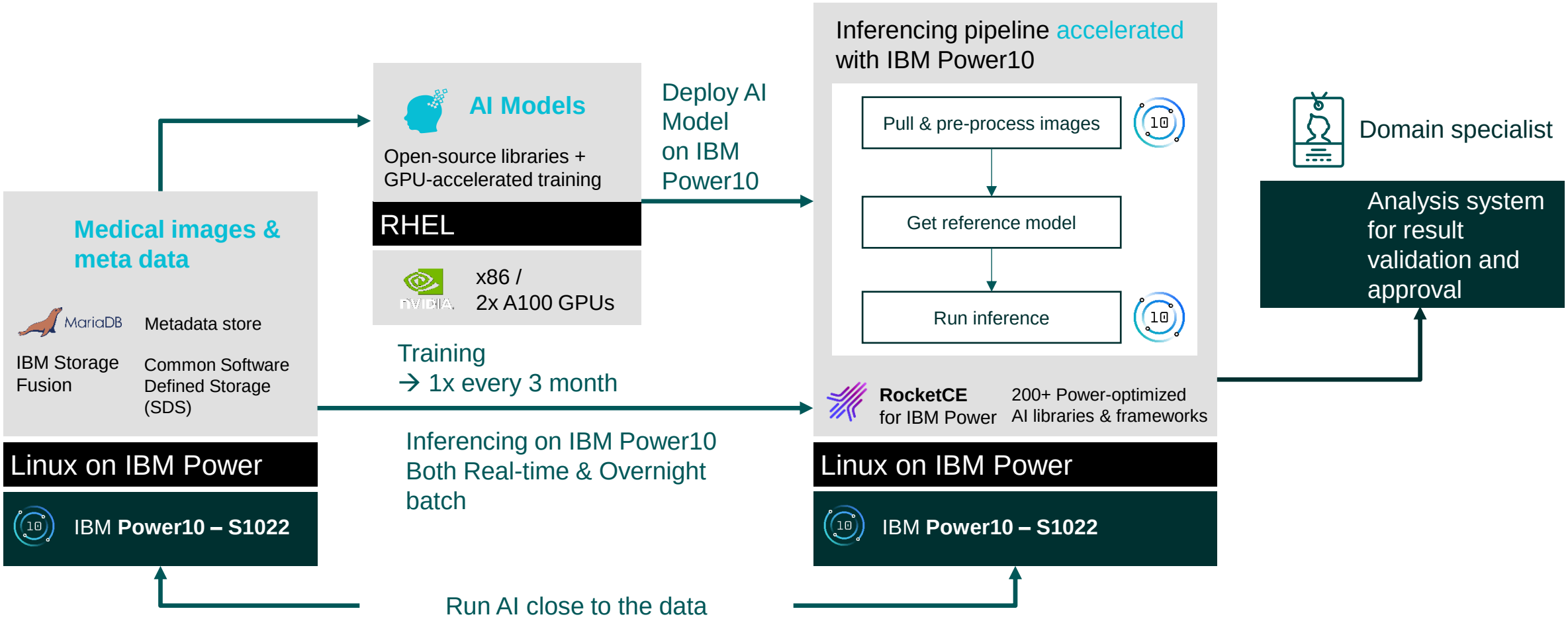
Safeguard AI Workflows - a holistic framework



TD SYNNEX

Client Use Cases – Run Traditional AI & GenAI Models on IBM Power10

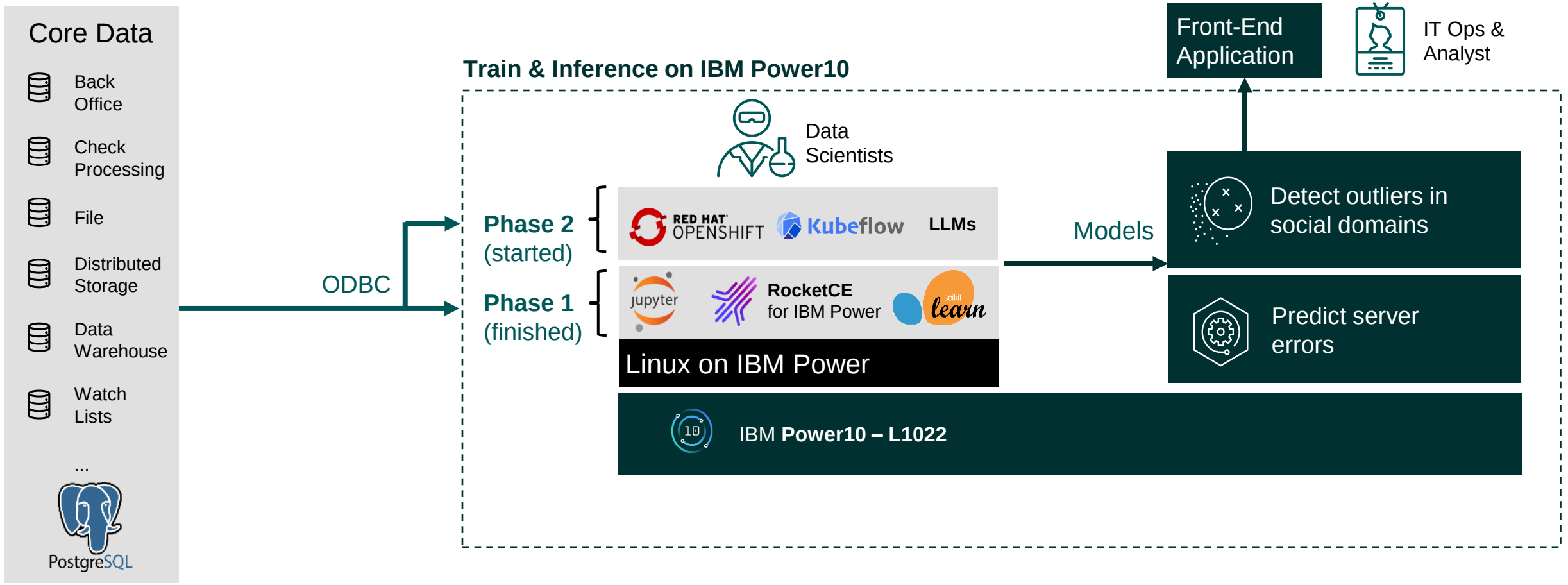
Multi-architecture solution with training deployed on x86 Inferencing on IBM Power10 for both Traditional AI and Gen AI models.



TD SYNNEX

Client Use Cases – Train and Run AI Models on IBM Power10

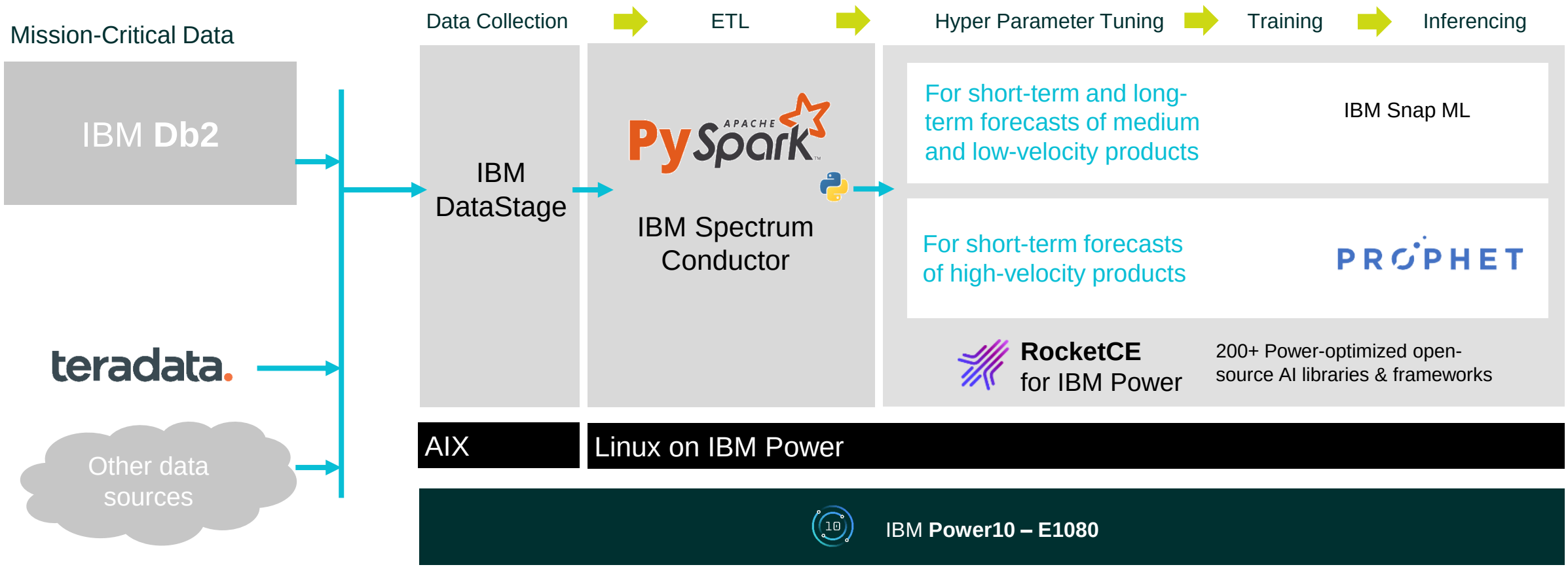
Accessing data from multiple external data sources Training and Inference (multiple ML models) on IBM Power10



TD SYNNEX

Client Use Cases – Prepare Data and Drive Inference on IBM Power10

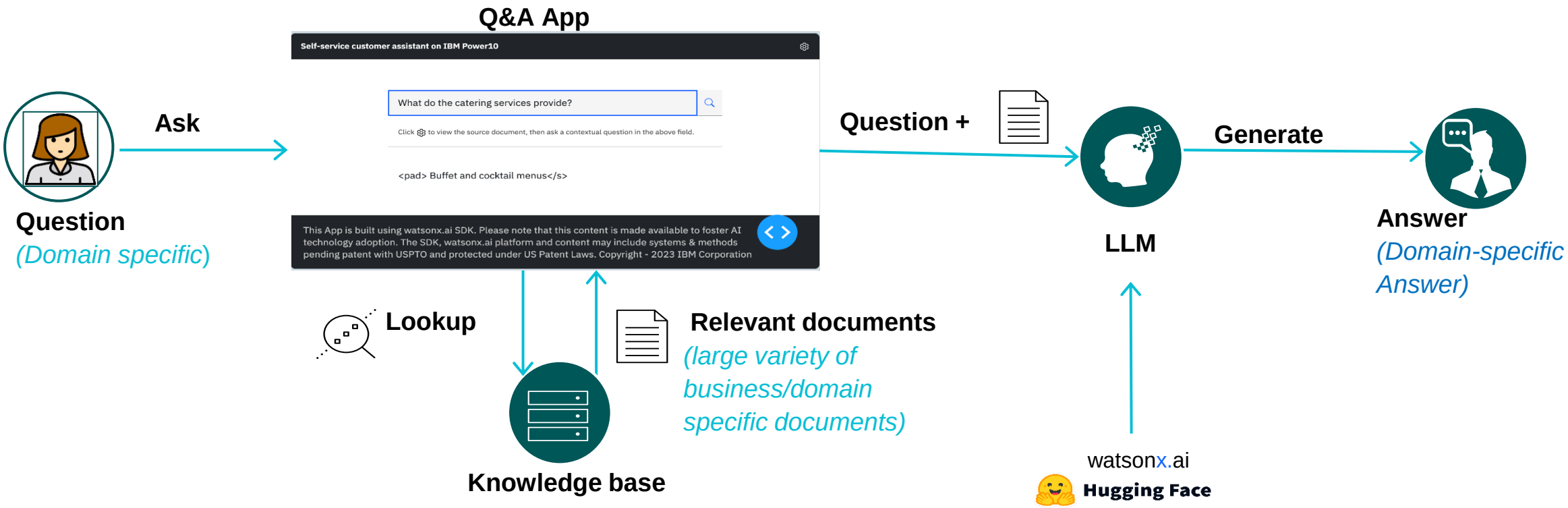
Accessing data residing on multiple data sources – Transformation of data and inferencing with Enterprise and Open-source SW on IBM Power10



TD SYNNEX

Client Use Cases – Digital Assistant with GenAI on IBM Power10

Use a Large Language Model (LLM) adapted to a specific domain on IBM Power10 with MMA – Digital Labor use case



IBM **Power10**

TD SYNNEX

IBM Power modernizes infrastructure and accelerates innovation with AI in the year ahead

Source: IBM Newsroom

TD SYNNEX

IBM Spyre Accelerator and IBM Power

Drive business growth, improve productivity, and remain in compliance with regulatory requirements

Jointly designed by IBM Research and IBM Infrastructure, Spyre's architecture is designed for more efficient AI computation. Notably, the chip will send data directly from one compute engine to the next, leading to an efficient use of energy. This family of processors also uses a range of lower precision numeric formats (such as int4 and int8), to make running an AI model more energy efficient and far less memory intensive.

Introducing IBM Power11 processor

Next generation IBM Power11 system will have innovations in the processor, system, and stack levels to help enterprises propel digital transformation initiatives for their mission-critical infrastructure. IBM Power also continues to support emerging enterprise AI use cases with the MMA (Matrix-Math Assist) architecture.

The Power11 processor is designed to deliver higher clock speeds and can add up to 25% more cores per processor chip than comparable IBM Power10 systems. The Power11 processor builds upon the key capabilities we delivered with Power10 including stronger reliability, availability, and serviceability (RAS) characteristics, better energy efficiency and energy management, and improved quantum-safe security.

Integrating IBM Spyre Accelerator and IBM Power

AI workloads are creating their own demands on client's infrastructure. We are committed to helping clients right size their workload placement and build use cases for even more advanced AI capabilities on IBM Power. I'm excited to reveal that IBM intends to incorporate the IBM Spyre Accelerator in future Power offerings to provide additional AI compute capabilities. Working together, IBM Power11 processors and the IBM Spyre Accelerator are designed to enable the next generation infrastructure to scale demanding AI workloads for businesses.

The IBM Spyre Accelerator is a purpose-built enterprise-grade accelerator offering scalable capabilities for complex AI models and generative AI use cases. The new accelerator features 32 individual accelerator cores onboard, and each Spyre is mounted on a PCIe card.


Questions &
Answers

Sejid Canoski

Technical Presales Consultant / Architect

Sejid Canoski is a solutions architect at TD SYNNEX. He designs and validates partner solutions based on IBM Security, IBM Power and IBM Storage products.
He also teaches IBM Security and IBM Power at the TD SYNNEX Academy.
When he is not busy with his main job he loves to tinker with IBM Cyber Resiliency solutions and watsonx.ai / watsonx.governance.

TD SYNNEX Germany GmbH & Co. OHG
Kistlerhofstraße 75
81379 Munich

E-Mail: Sejid.canoski@tdsynnex.com
Tel.: +49 (0) 89 4700 – 1812
Mobil: +49 (0) 176 200 83730

LinkedIn

Slack

IBM TechXchange Community

IBM Champion 2025
Awarded through the program
IBM Champion

TD SYNNEX

Red Hat

Authorized Distributor

The Power of us

Connect. Empower. Transform.