



TD SYNnex



Distributor

Let's Create

an IT Infrastructure engineered

for **SAP** workloads

Bob Friske

Vendor Business Executive

IBM Brand Manager

robert.friske@tdsynnex.com

Building on 50+ years of partnership...

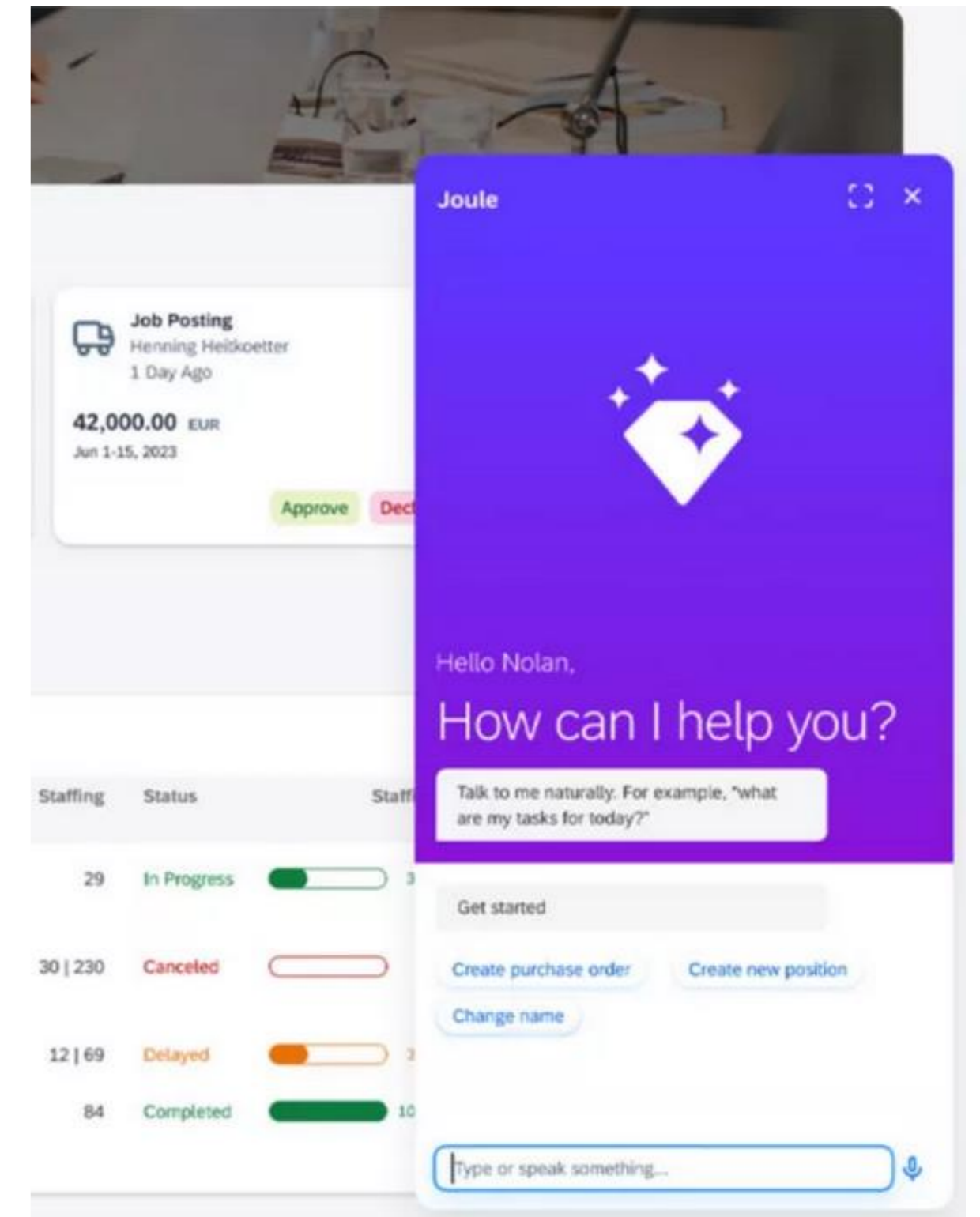
“IBM and SAP are joining forces to give businesses new and exciting ways to harness the transformative power of AI and use it as a source of competitive advantage.

... we are infusing IBM Watson’s powerful, enterprise-grade AI capabilities into SAP’s leading ERP platform to help businesses reimagine customer experiences, boost productivity and fuel growth.”

Arvind Krishna
Chairman and CEO, IBM

SAP Joule – SAP Assistant Framework
(based on watsonx – Nov 2023)

watsonx code assistant



```
java > com > ibm > wca > J CustomerRequest.java > ...
1 package com.ibm.wca;
2
3 class CustomerRequest extends Request {
4     Customer customer;
5
6     public CustomerRequest() {
7         customer = new Customer();
8     }
9
10    public Customer getCustomer() {
11        return customer;
12    }
13
14    public void setCustomer(Customer valCustomer) {
15        customer = valCustomer;
16    }
17
18    @Override
19    public String toString() {
20        return "customer = " + customer;
21    }
22
23    void insertCustomer(String lgacNcs, int db2CustomerNumInt){}
24
25    void obtainCustomerNumber(String lgacNcs, int db2CustomerNumInt){}
26 }
27
28 |
```

SAP clients host mission critical transactions, data, and [AI workflows](#) on IBM Power.

“Transitioning to IBM Power10 servers delivers performance gains of up to 75% while cutting energy consumption by 20% compared to POWER9. IBM is a true partner in our journey to run business-critical systems cost efficiently and sustainably, while delivering outstanding performance.”

- Christian Dummler, SAP Infrastructure, Bosch Group

“By deploying an AI inference solution for both speech-to-text and image analysis on IBM Power10, the pathology unit was able to increase sensitivity in detecting lesions and prioritize high probability cancer cases, leading to better clinical outcomes, a faster time to treatment for patients, and an efficient reduction in pathologist workloads.”

– Head of Pathology, Hospital Chain in APAC

↗ 4,800 SAP clients



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

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

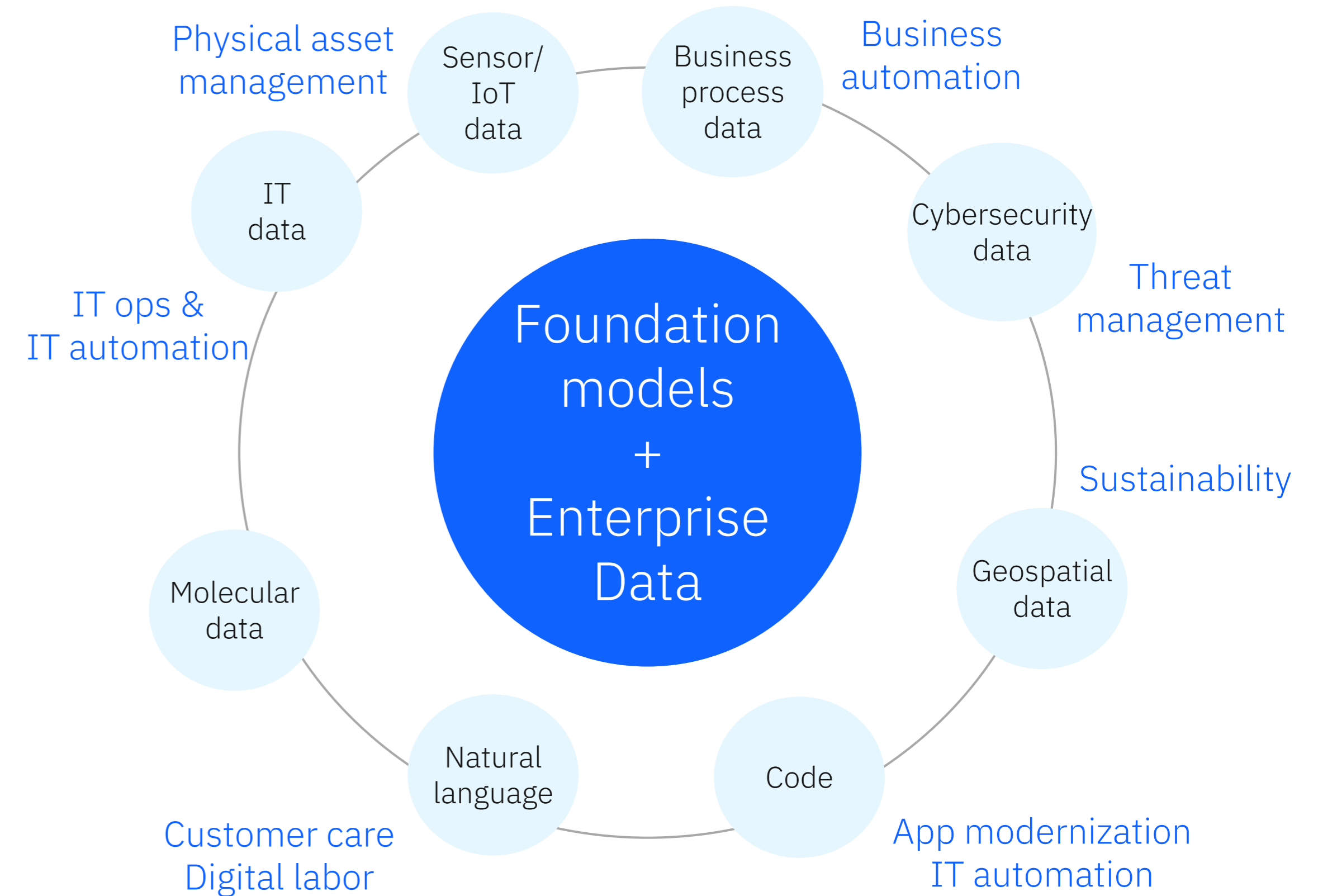
64%

of executives believe that generative AI is closing the gap between IT and the business¹

¹ IBM Institute for Business Value (IBV): The CEOs Guide to Generative AI

Data is the new oil: Proprietary data will be the source of competitive advantage

“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).”¹



¹ IBM Institute for Business Value (IBV): The CEOs Guide to Generative AI

Enterprise considerations:

Five truths we're hearing about generative AI

Truth 1 Multi-model

Two thirds of 150+ enterprises surveyed report pursuing a **multi-model strategy**¹

- 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 / multicloud environments**²

- 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 difficult³

- 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**⁴

- 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

Generative AI pilots have not made it to production due to **challenges with data quality, access, and security**

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

Our 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 Power10 is built for AI.

Inferencing vs.
compared x86 servers:

↑ 42%

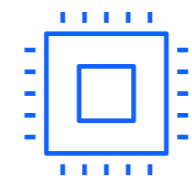
more throughput for large
language models¹

↑ 39%

more requests
per watt²

↓ 51%

lower TCO
over 3-years²



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







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.supernmicro.com/en/products/system/hyper/2u/sys-221h-trn>
PrimeQA models: <https://github.com/primeqa>
Models fine-tuned by IBM on a corpus of IBM-internal data

² 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 1.6TB 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 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. Pricing is based on: Power S1022 (see page 4). Typical industry standard Intel x86 (example on page 5) pricing <https://www.synnexcorp.com/us/govsolvr/pricing/> and IBM software pricing available at <https://www.ibm.com/downloads/cas/DLBCQWBPK>
Assumes energy usage for the Supermicro server is similar to a similarly configured Lenovo server (a ThinkSystem SR650 V3) <https://www.ibm.com/en/products/system/hyper/2u/sys-221h-trn>; 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.

Reference architecture for IBM's enterprise AI technology on IBM Power.

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

AI assistants 	Empower individuals to do work without expert knowledge across a variety of business processes and applications.	<ul style="list-style-type: none"> - 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.	<table border="0"> <tr> <td data-bbox="1459 489 1865 714"> watsonx.ai SDK → use foundation models in IBM Power applications (Python, SAP ABAP). </td> <td data-bbox="1885 489 2442 714"> Ecosystem integrations - Equitus - Trovares </td> </tr> </table>	watsonx.ai SDK → use foundation models in IBM Power applications (Python , SAP ABAP).	Ecosystem integrations - Equitus - Trovares														
watsonx.ai SDK → use foundation models in IBM Power applications (Python , SAP ABAP).	Ecosystem integrations - Equitus - Trovares																	
AI & data platform 	Leverage generative AI and machine learning — tuned with your data — with responsibility, transparency, and explainability.	<table border="0"> <tr> <td data-bbox="1459 744 1865 1119"> Cloud Pak for Data - IBM Watson Studio - IBM Watson Machine Learning - IBM Data Refinery - IBM Watson Studio Runtimes Rocket AI Hub - Kubeflow* - KServe* - RocketCE* - Jupyter Hub & Jupyter Lab* </td> <td data-bbox="1885 744 2442 1119"> Foundation models <table border="0"> <tr> <td>Granite***</td> <td> IBM</td> </tr> <tr> <td>Llama 2</td> <td> Meta AI</td> </tr> <tr> <td>Geospatial</td> <td> IBM + NASA</td> </tr> <tr> <td>...</td> <td> Deci, Elinar, Hugging Face, ...</td> </tr> </table> Vector Databases <table border="0"> <tr> <td>- Milvus**</td> <td>- Weaviate**</td> </tr> <tr> <td>- Chroma**</td> <td>- OpenSearch**</td> </tr> <tr> <td>- PGVector**</td> <td>- Faiss**</td> </tr> </table> </td> </tr> </table>	Cloud Pak for Data - IBM Watson Studio - IBM Watson Machine Learning - IBM Data Refinery - IBM Watson Studio Runtimes Rocket AI Hub - Kubeflow* - KServe* - RocketCE* - Jupyter Hub & Jupyter Lab*	Foundation models <table border="0"> <tr> <td>Granite***</td> <td> IBM</td> </tr> <tr> <td>Llama 2</td> <td> Meta AI</td> </tr> <tr> <td>Geospatial</td> <td> IBM + NASA</td> </tr> <tr> <td>...</td> <td> Deci, Elinar, Hugging Face, ...</td> </tr> </table> Vector Databases <table border="0"> <tr> <td>- Milvus**</td> <td>- Weaviate**</td> </tr> <tr> <td>- Chroma**</td> <td>- OpenSearch**</td> </tr> <tr> <td>- PGVector**</td> <td>- Faiss**</td> </tr> </table>	Granite***	IBM	Llama 2	Meta AI	Geospatial	IBM + NASA	...	Deci, Elinar, Hugging Face, ...	- Milvus**	- Weaviate**	- Chroma**	- OpenSearch**	- PGVector**	- Faiss**
Cloud Pak for Data - IBM Watson Studio - IBM Watson Machine Learning - IBM Data Refinery - IBM Watson Studio Runtimes Rocket AI Hub - Kubeflow* - KServe* - RocketCE* - Jupyter Hub & Jupyter Lab*	Foundation models <table border="0"> <tr> <td>Granite***</td> <td> IBM</td> </tr> <tr> <td>Llama 2</td> <td> Meta AI</td> </tr> <tr> <td>Geospatial</td> <td> IBM + NASA</td> </tr> <tr> <td>...</td> <td> Deci, Elinar, Hugging Face, ...</td> </tr> </table> Vector Databases <table border="0"> <tr> <td>- Milvus**</td> <td>- Weaviate**</td> </tr> <tr> <td>- Chroma**</td> <td>- OpenSearch**</td> </tr> <tr> <td>- PGVector**</td> <td>- Faiss**</td> </tr> </table>	Granite***	IBM	Llama 2	Meta AI	Geospatial	IBM + NASA	...	Deci, Elinar, Hugging Face, ...	- Milvus**	- Weaviate**	- Chroma**	- OpenSearch**	- PGVector**	- Faiss**			
Granite***	IBM																	
Llama 2	Meta AI																	
Geospatial	IBM + NASA																	
...	Deci, Elinar, Hugging Face, ...																	
- Milvus**	- Weaviate**																	
- Chroma**	- OpenSearch**																	
- PGVector**	- Faiss**																	
Data services 	Define, organize, manage, and deliver trusted data to train and tune AI models with data fabric services.	<table border="0"> <tr> <td data-bbox="1459 1149 1865 1375"> Data fabric services - IBM Analytics Engine - IBM Event Streams - IBM App Connect Enterprise - IBM Db2 Warehouse - IBM MQ </td> <td data-bbox="1885 1149 2442 1375"> <table border="0"> <tr> <td>- RH Fuse</td> <td>- Apache Spark**</td> </tr> <tr> <td>- RH AMQ Streams for Apache Kafka</td> <td>- Apache Kafka**</td> </tr> <tr> <td>- RH Data Grid</td> <td>- trino**</td> </tr> </table> </td> </tr> </table>	Data fabric services - IBM Analytics Engine - IBM Event Streams - IBM App Connect Enterprise - IBM Db2 Warehouse - IBM MQ	<table border="0"> <tr> <td>- RH Fuse</td> <td>- Apache Spark**</td> </tr> <tr> <td>- RH AMQ Streams for Apache Kafka</td> <td>- Apache Kafka**</td> </tr> <tr> <td>- RH Data Grid</td> <td>- trino**</td> </tr> </table>	- RH Fuse	- Apache Spark**	- RH AMQ Streams for Apache Kafka	- Apache Kafka**	- RH Data Grid	- trino**								
Data fabric services - IBM Analytics Engine - IBM Event Streams - IBM App Connect Enterprise - IBM Db2 Warehouse - IBM MQ	<table border="0"> <tr> <td>- RH Fuse</td> <td>- Apache Spark**</td> </tr> <tr> <td>- RH AMQ Streams for Apache Kafka</td> <td>- Apache Kafka**</td> </tr> <tr> <td>- RH Data Grid</td> <td>- trino**</td> </tr> </table>	- RH Fuse	- Apache Spark**	- RH AMQ Streams for Apache Kafka	- Apache Kafka**	- RH Data Grid	- trino**											
- RH Fuse	- Apache Spark**																	
- RH AMQ Streams for Apache Kafka	- Apache Kafka**																	
- RH Data Grid	- trino**																	
Hybrid cloud AI tools 	Build on a consistent, scalable foundation based on open-source technology.	<table border="0"> <tr> <td data-bbox="1459 1405 1865 1600"> Red Hat OpenShift </td> <td data-bbox="1885 1405 2442 1600"> Native LPAR deployment options </td> </tr> </table>	Red Hat OpenShift	Native LPAR deployment options														
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

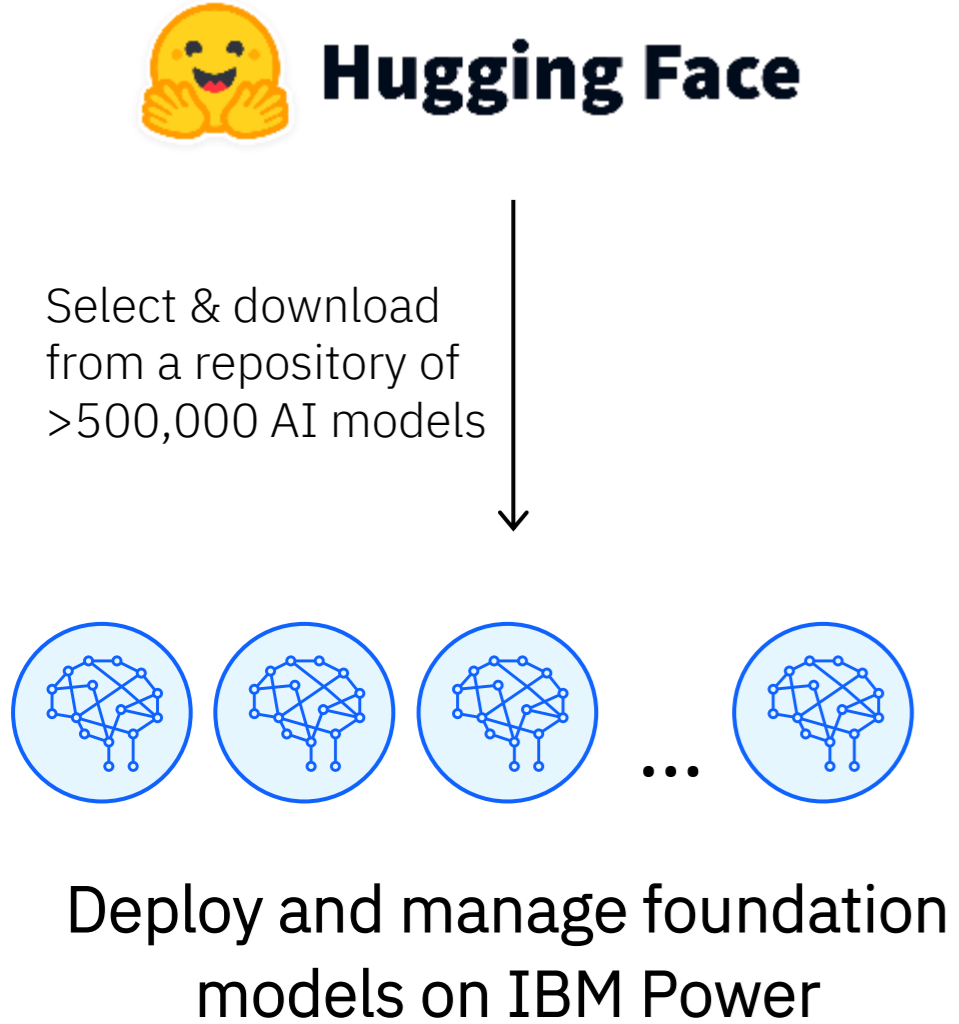
External AI services useful for IBM Power

Ansible Lightspeed
 → Generate playbooks for IBM i & AIX.

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

Get started with AI and watsonx with IBM Power.

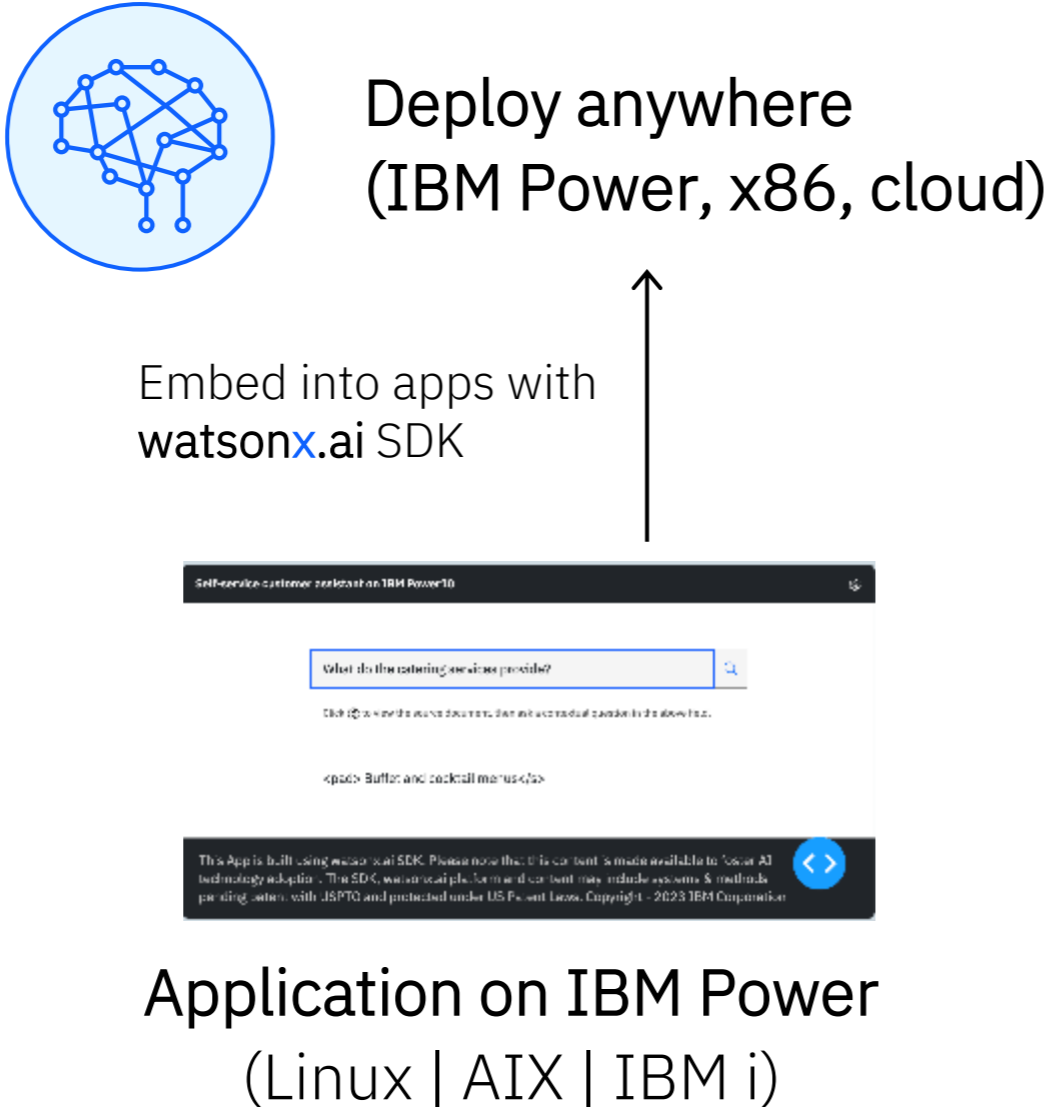
Deploy & manage foundation models securely.



Leverage best-of-breed open-source models and software technologies to build a scalable end-to-end AI workflow

- Customer service
- Knowledge worker
- Fraud reporting

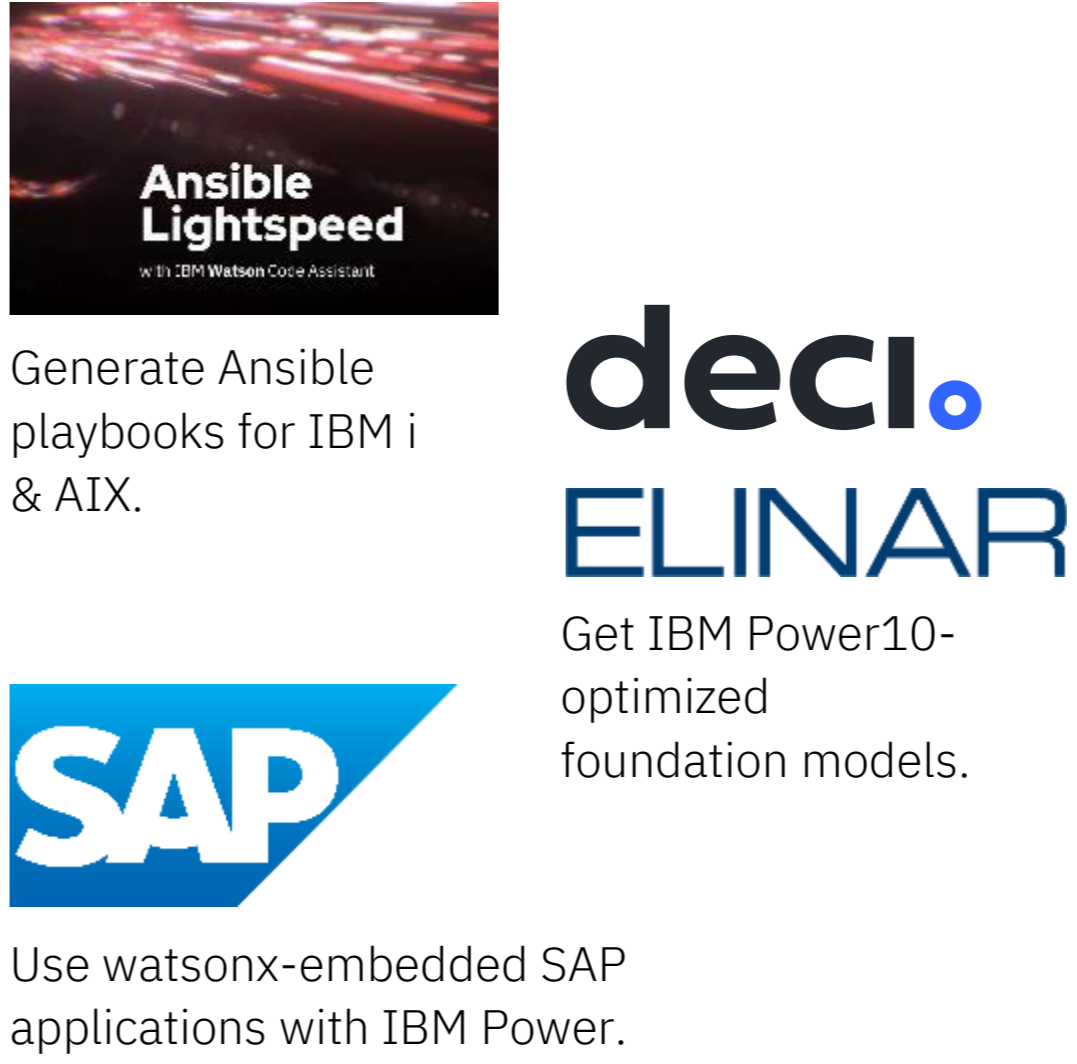
Embed foundation models into apps using the watsonx.ai SDK.



Embed AI quickly, in a secure and resilient environment, close to your mission critical data and transactions

- Report generation
- Citizen services
- Knowledge management

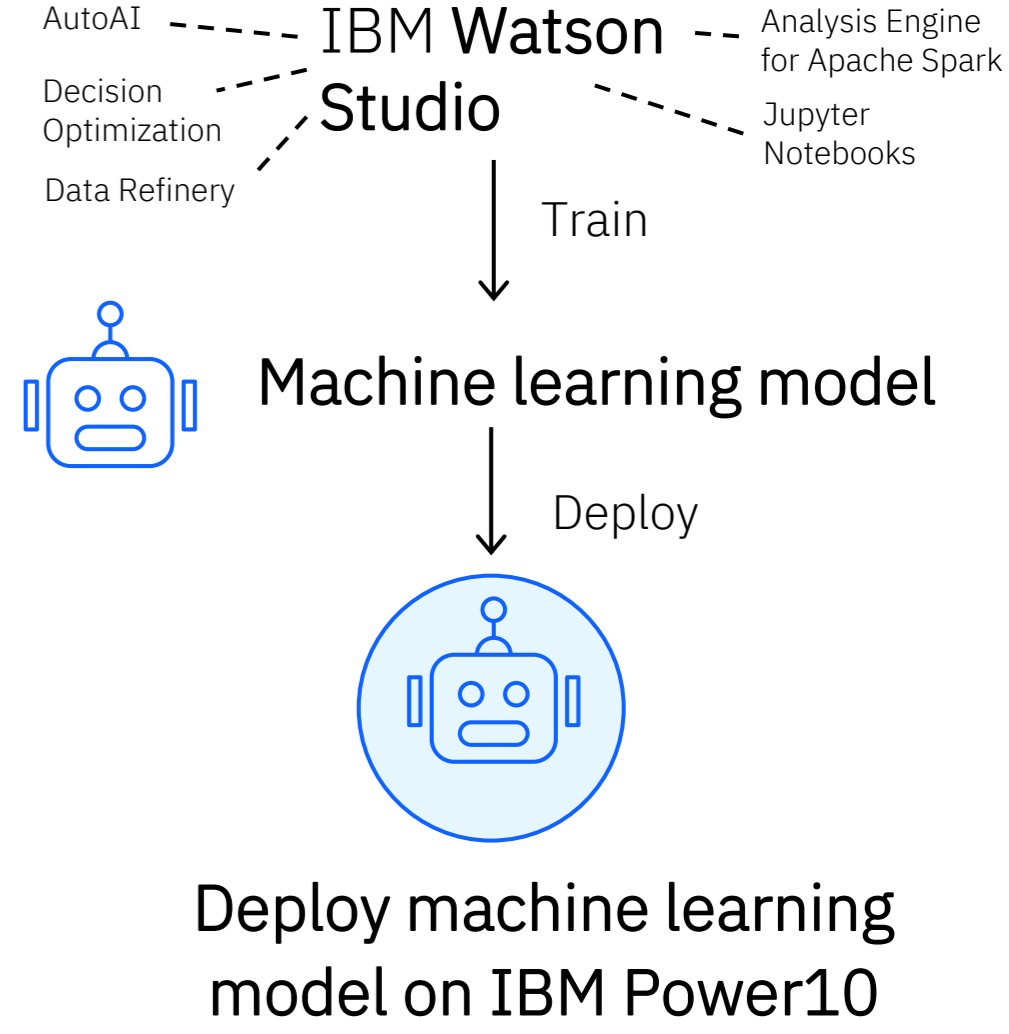
Consume watsonx services from customized ecosystem apps.



Deliver new services faster using generative AI capabilities embedded in familiar ecosystem apps

- Asset management
- Code generation
- Accounting automation

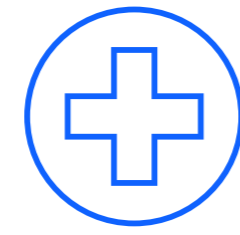
Train & deploy ML models within a single AI studio.



Train, tune, and inference machine learning models with on-chip acceleration without purchasing GPUs

- Fraud detection
- Risk underwriting
- Demand forecasting

Clients reinventing how work gets done.



A large hospital in Thailand tuned and deployed an AI solution for both speech-to-text and image analysis to automate manual pathologist workloads and treat patients faster.

- 25% faster inference vs. x86 server with NVIDIA T4 GPUs & double-precision inferencing



A financial institution in Europe improved its AIOps solution to optimize IT resources and forecast server outages and used AI to automate fraud detection.

- 3-4x speed-up of machine learning model training compared to Intel Xeon 6248 servers



A retailer in the US improved turnaround time for creating timely product-specific forecasts by training, tuning, and inferencing models at scale.

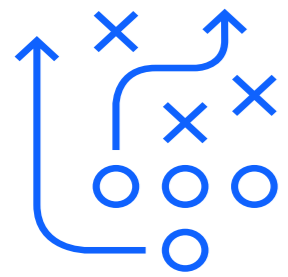
- Models trained and applied to goods improved from 40 to 207 models per second
- 8.5x faster inference for long-term forecasts compared to GPU clusters



An MSP in New Zealand tuned open-source models downloaded from Hugging Face to enable government clients to personalize and automate citizen services.

- Governance, control, and trust across full-stack
- Data-sovereign environments with optional access to an internet connection

Discover how **AI on IBM Power** can help transform your organization.



INFORM YOURSELF

Explore AI on IBM Power yourself with [blogs and demos](#) delivered through the [IBM Power Developer eXchange](#) community.

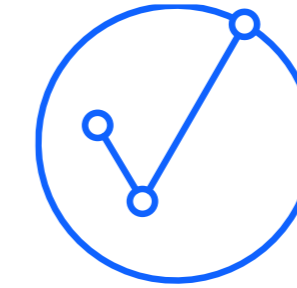
30 minutes or less



GET BRIEFED

Meet with an IBM AI expert for custom demonstrations of AI on IBM Power [capabilities, values, and roadmaps](#). Understand where AI can be leveraged to impact your organization.

2 – 4 hours



START PRODUCTIZING AI

Request a [use case alignment workshop](#) from your IBM sales representative to drive toward your first [proof-of-concept](#) in a few weeks.

1 – 4 weeks

IBM Power

AI on IBM Power: Use Case Alignment Workshop

Let's get your line of business and technical teams along with the IBM team in an interactive workshop experience to identify and understand your AI on Power use cases and align on the business outcome.

Workshop Outcomes

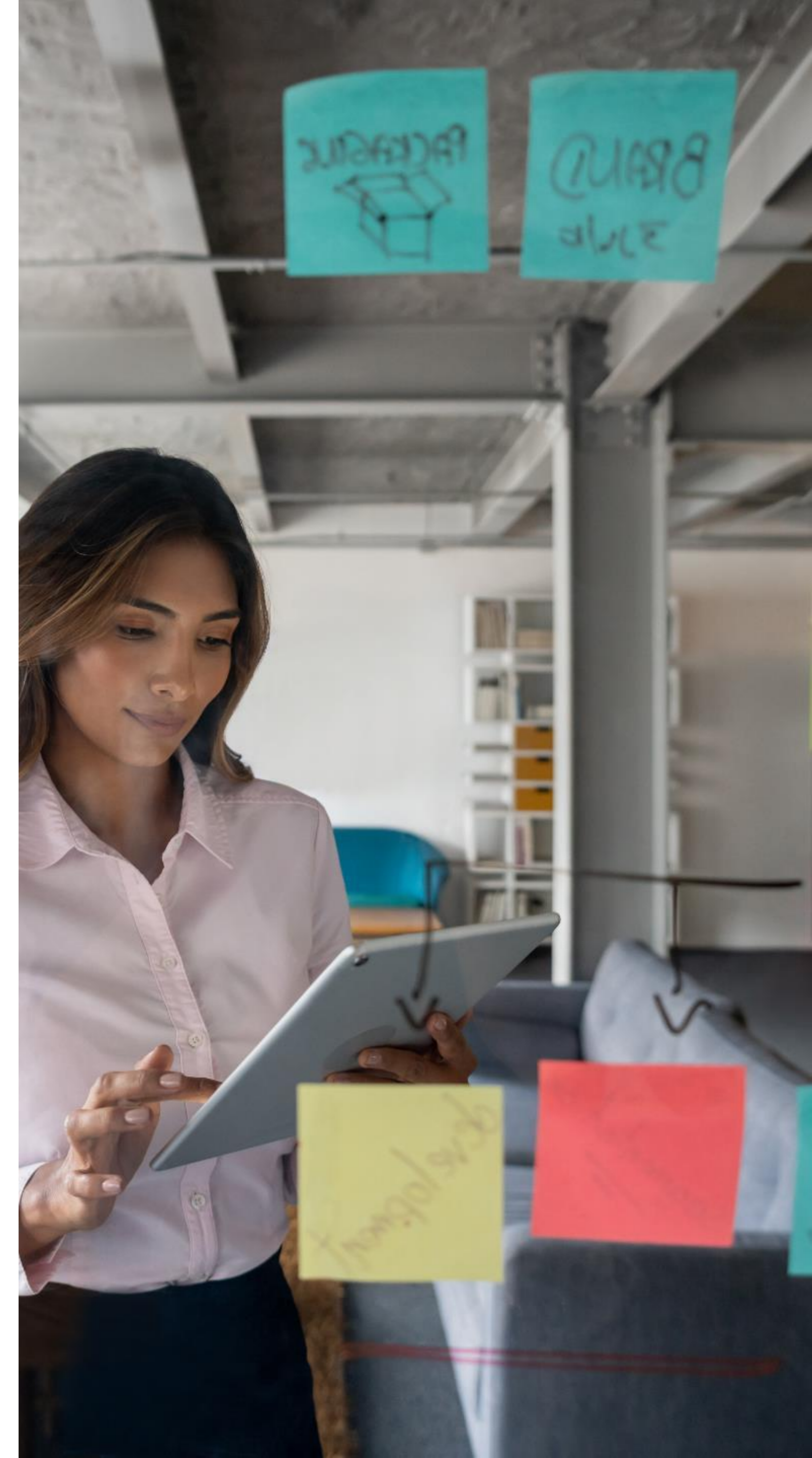
- Generate prioritized list of use-cases
- Align Business and IT on a single use-case
- Business opportunity statement and potential value

What we need from you:

- Ideas for leveraging AI to support business priorities
- 4 to 8 workshop participants (ideally) spanning Line of Business, IT/Development and Leadership

Potential Activities

- AI landscape
- Generative AI Creativity Matrix
- Business value prioritization
- Business & Technical Feasibility Assessment
- Use-case prioritization matrix
- Business opportunity statement



Thank You