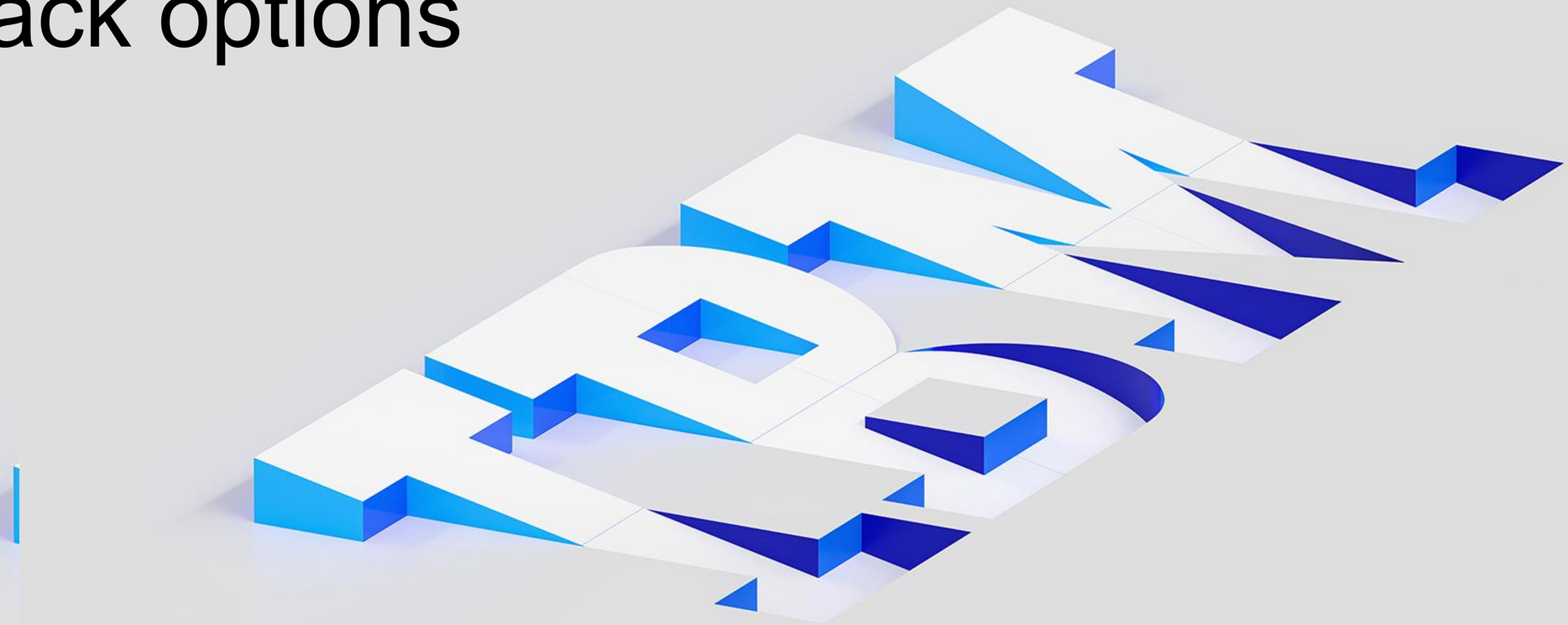
Module 1:
AI Stack options



Jesse Gorzinski
Senior Business Architect, IBM

Three main principles

Hardware

matériel informatique

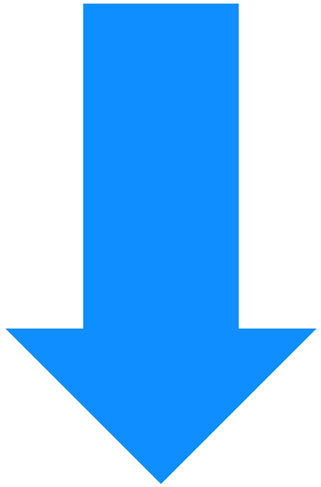
Software

logiciel

Integration

intégration

Three main principles



Hardware

matériel informatique

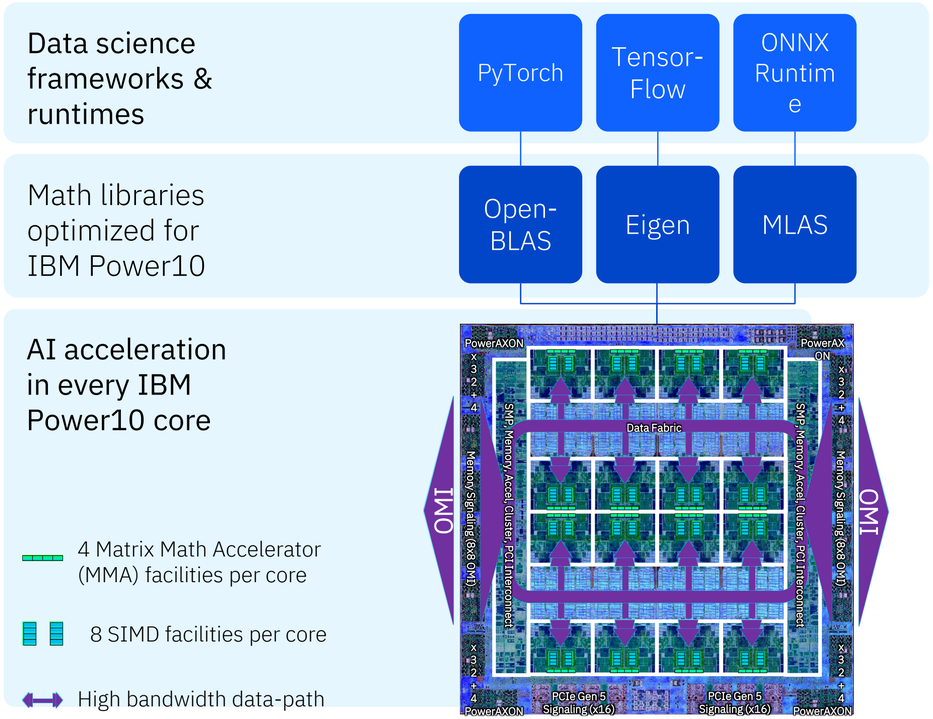
Software

logiciel

Integration

intégration

Accelerate AI Efficiently with AI Optimized Hardware



Each core has four MMA (Matrix Math Accelerator) facilities to accelerate matrix calculations that are used in many common AI workloads

Power 10 MMA Overview

Feature	AI Method	GPU	P10 with MMA
Training	Deep Learning	Best Fit (cost-perf)	Limited Benefit
	Machine Learning	Limited Benefit (cost-perf)	Best Fit (cost-perf)
	Foundation Model (like GenAI)	Best Fit (cost-perf)	Not Optimal
Re-training / Fine-tuning	Deep Learning	Best Fit (cost-perf)	Limited Benefit (cost-perf)
	Machine Learning	Not Applicable	Not Applicable
	Foundation Models (like GenAI)	Best Fit (cost-perf)	Limited Benefit (cost-perf)
Prompt Tuning (including RAG pattern)	Deep Learning	Not Applicable	Not Applicable
	Machine Learning	Not Applicable	Not Applicable
	Foundation Model (like GenAI)	Limited Benefit (cost-perf)	Best Fit (cost-perf)
Inference	Deep Learning	Limited Benefit (cost-perf)	Best Fit (cost-perf)
	Machine Learning	Limited Benefit (cost-perf)	Best Fit (cost-perf)
	Foundation Model (like GenAI)	Best Fit >3B depending on several factors (cost-perf)	Best Fit <3B depending on several factors (cost-perf)
SW Maintenance		Need to update GPU specific SW (CUDA, cuDNN, etc.)	Maintained by IBM / Partner

GPUs or Power 10 w/MMA*

*Please see speaker notes for details

Accelerate AI Efficiently on a high performant, sustainable platform

42%

more inferencing throughput(1)

Process up to **42% more** queries per second **with <1s latency** on IBM Power10 than compared x86 servers during peak times when using large language AI models.

Improved Performance

39%

More inferencing per watt (3)

Gain more insights by running **39% higher inferencing** per unit of energy Cloud Pak for Data on Power S1022 with MMA vs running on an x86 server with AMX

Improved CO2 Footprint

51%

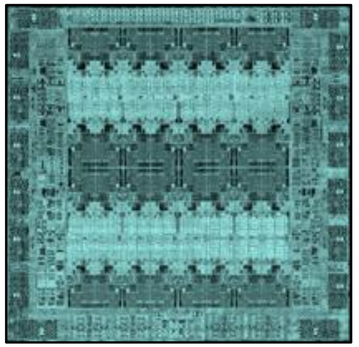
Lower solution costs over 3 years(2)

Achieve **51% lower 3-yr TCO** running highly parallel inferencing in Cloud Pak for Data on Power S1022 with MMA vs running on an x86 server with AMX

Improved Economics

Power 11 Processor Innovations + Spyre Accelerator + Growing AI Stack

Power 11 Processor Innovations

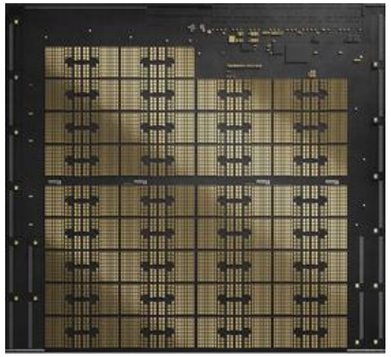


2.5D Stacking: Energy Optimization

Agnostic, 3x Pipes, 2x Capacity

Uptime, Energy Mgmt, Quantum Safe

IBM Spyre™ Accelerator



Enterprise Grade AI Chip

Perf/Watt competitive positioning

Enables generative capabilities on IBM Power

AI / MLOps

OpenShift AI – SOD in place

Data Fabric

Data Stage – SOD in place

Knowledge Catalog – SOD in Place

watsonx.data – SOD being worked

Consumption Model

PowerVS + watsonx – toolkits

On-Premise – OpenSource or Enterprise SW

Simplified

Artificial Intelligence (AI)

Human intelligence exhibited by machines

AI can be defined as a technique that enables machines to mimic cognitive functions associated with human minds – cognitive functions include all aspects of learning, reasoning, perceiving, and problem solving.

Machine Learning (ML)

Systems that learn from historical data

ML-based systems are trained on historical data to uncover patterns. Users provide inputs to the ML system, which then applies these inputs to the discovered patterns and generates corresponding outputs.

Deep Learning (DL)

ML technique that mimics human brain function

DL is a subset of ML, using multiple layers of neural networks, which are interconnected nodes, which work together to process information. DL is well suited to complex applications, like image and speech recognition.

Generative AI (LLMs)

Foundation Models

AI model built using a specific kind of neural network architecture, called a transformer, which is designed to generate sequences of related data elements (for example, like a sentence).

1950's 1980's 2010's 2020's

Artificial Intelligence (AI)

Human intelligence exhibited by machines

AI can be defined as a technique that enables machines to mimic cognitive functions associated with human minds – cognitive functions include all aspects of learning, reasoning, perceiving, and problem solving.

Machine Learning (ML)

Systems that learn from historical data

ML-based systems are trained on historical data to uncover patterns. Users provide inputs to the ML system, which then applies these inputs to the discovered patterns and generates corresponding outputs.

Deep Learning (DL)

ML technique that mimics human brain function

DL is a subset of ML, using multiple layers of neural networks, which are interconnected nodes, which work together to process information. DL is well suited to complex applications, like image and speech recognition.

MMA

Generative AI (LLMs)

Foundation Models

AI model built using a specific kind of neural network architecture, called a transformer, which is designed to generate sequences of related data elements (for example, like a sentence).

1950's 1980's 2010's 2020's

Artificial Intelligence (AI)

Human intelligence exhibited by machines

AI can be defined as a technique that enables machines to mimic cognitive functions associated with human minds – cognitive functions include all aspects of learning, reasoning, perceiving, and problem solving.

Machine Learning (ML)

Systems that learn from historical data

ML-based systems are trained on historical data to uncover patterns. Users provide inputs to the ML system, which then applies these inputs to the discovered patterns and generates corresponding outputs.

Deep Learning (DL)

ML technique that mimics human brain function

DL is a subset of ML, using multiple layers of neural networks, which are interconnected nodes, which work together to process information. DL is well suited to complex applications, like image and speech recognition.

MMA

Generative AI (LLMs)

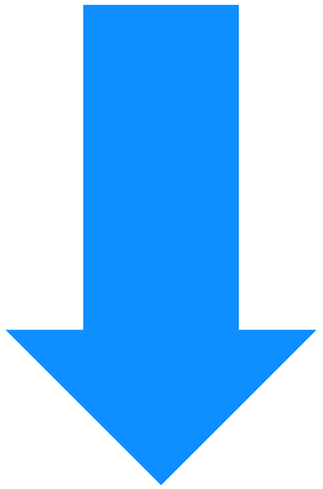
Foundation Models

AI model built using a specific e, called a transformer, which is designed to generate sequences of related data elements (for example, like a sentence).

Spyre

1950's 1980's 2010's 2020's

Three main principles



Hardware

matériel informatique

Software

logiciel

Integration

intégration

Software Options...

IBM i Native

Enterprise-supported open source and watsonx

Ecosystem Partners

Roll-your-own

IBM i Native solutions

Scikit-Learn (sklearn)

- Robust Machine Learning for Python

Llama.cpp

- Run large language models (LLMs)

Chainer

- Python-based neural networks

IBM i Native solutions: Tesseract OCR


```
"textAnnotations": [
  {
    "locale": "en",
    "description": "YMLU 621093 9\n22U1\nCAUTION\n07/08/2017\n",
    "boundingPoly": {
      "vertices": [
```

IBM i Native solutions

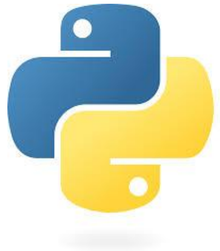
Why?

- Lowest barrier of entry
- Minimal architectural impact

Why not?

- Older versions
- Much fewer packages than available on Linux
- May need scalability of container platform such as Kubernetes or Red Hat’s OpenShift Container Platform

Enterprise-supported open source



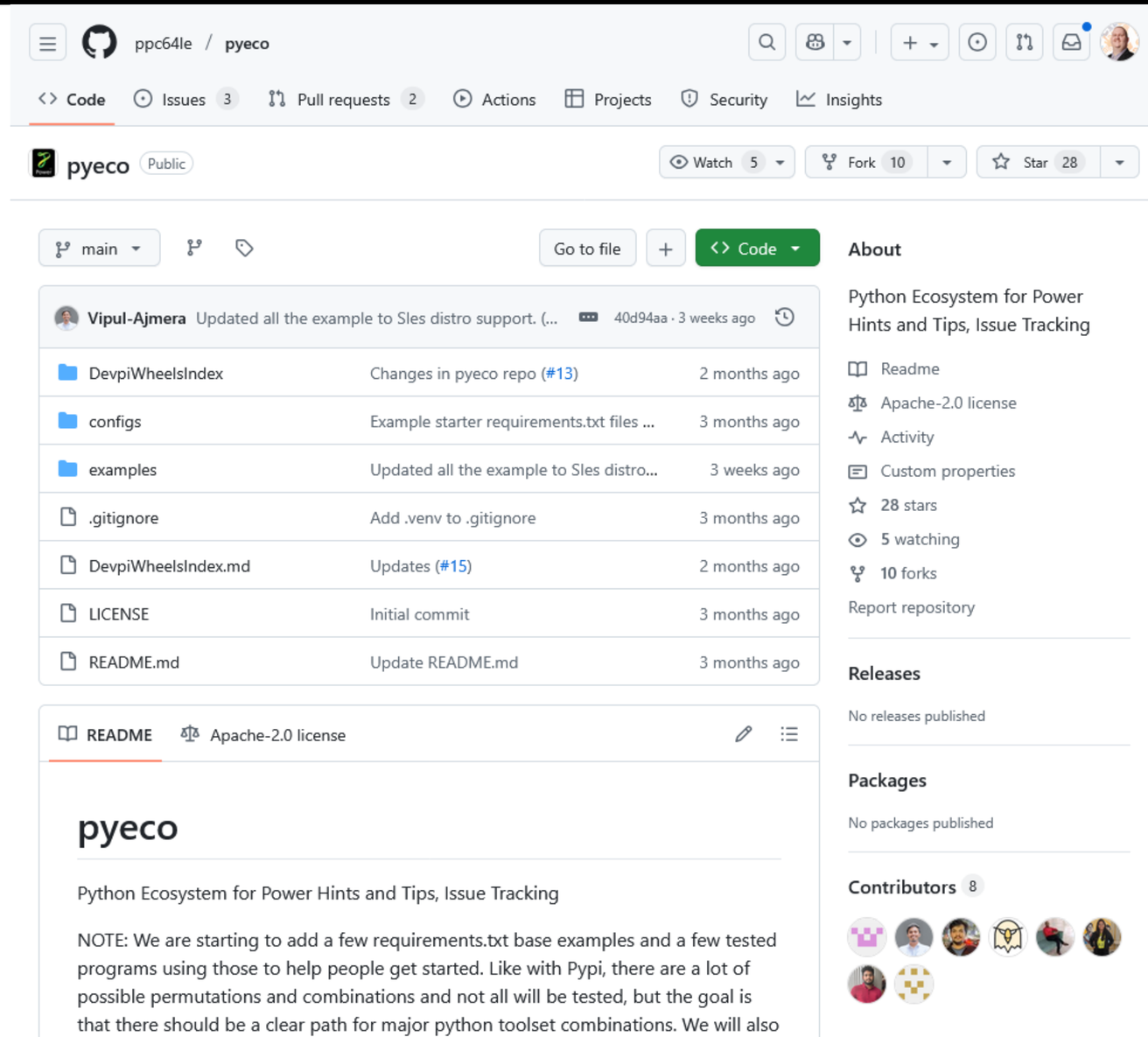
Red Hat OpenShift AI

Some options:

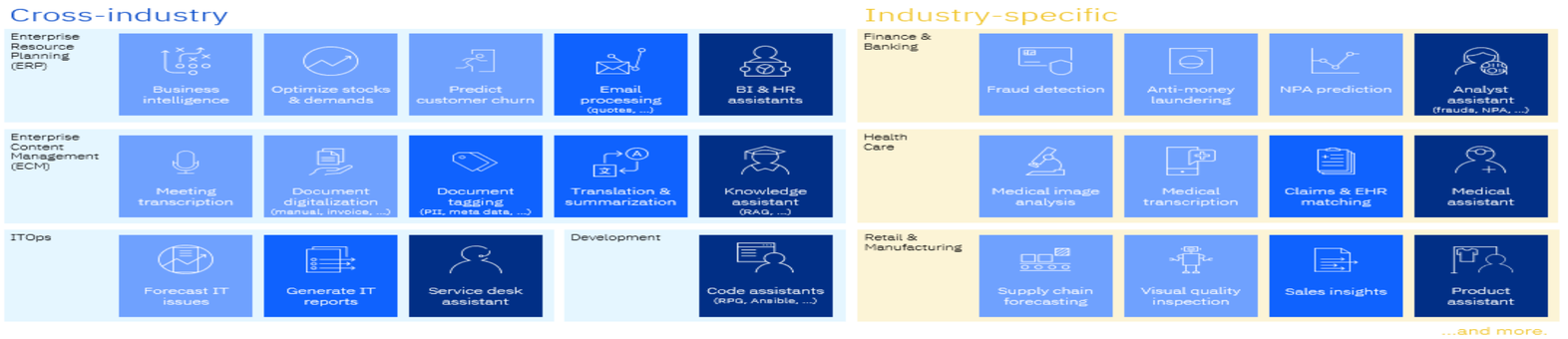
- Python Ecosystem for IBM Power
- OpenShift AI

Python Ecosystem for Power (“PyEco”)
ibm.biz/python-ai-power

- Successor to Rocket CE
- Packages optimized for Linux on Power
  - Exploit MMA and Spyre
  - Enterprise-level support
- Integrate with IBM i using Mapepire



Red Hat SW Stack for AI



AI use cases

Key use cases for IBM Power clients

Red Hat
Stacks

–Red Hat OpenShift AI (RHOAI)
– Red Hat AI Inference Server (RHAIIS)

Optimized AI foundation

→ *RHOAI to build, run and manage AI apps.*
→ *RHAIIS for simple, performant genAI inferencing on IBM Spyre*

Red Hat OpenShift Or RHEL

Operating system

→ *RHEL LPAR (RHAIIS)*
→ *Openshift – RHOAI or RHAIIS*

Math libraries optimized for on-chip (MMA)

Math libraries optimized for off-chip (Spyre)

Accelerated Infrastructure – Power 10/11

Optimized compute and libraries

for accelerating AI.

Red Hat AI overview for IBM Power

Red Hat AI

Red Hat AI Inference Server

Gen AI model Inference on RHEL/Linux or OpenShift/Kube

Red Hat OpenShift AI

Generative and Predictive AI platform for inference, training, tuning and GenAIOps

Trusted, Consistent and Comprehensive foundation

Hardware Acceleration IBM

Virtual

Private Cloud

Public Cloud

watsonx

Leverage foundation models to automate data search, discovery, and linking in watsonx.data

watsonx.ai

Train, validate, tune and deploy AI models

watsonx.data

Scale AI workloads, for all your data, anywhere

watsonx.governance

Enable responsible, transparent and explainable AI workloads

Leverage governed enterprise data in watsonx.data to seamlessly train or fine-tune foundation models

Direct, manage and monitor activities across the AI lifecycle, meeting risk and regulatory requirements with watsonx.governance

watsonx.ai: Prompt Lab

Experiment with foundation models and build prompts

Interactive prompt builder

Includes prompt examples for various use cases and tasks

Experiment with different prompts, save and reuse older prompts, use different models and vary different parameters

Experiment with zero-shot, one-shot, or few-shot prompting to get the best results

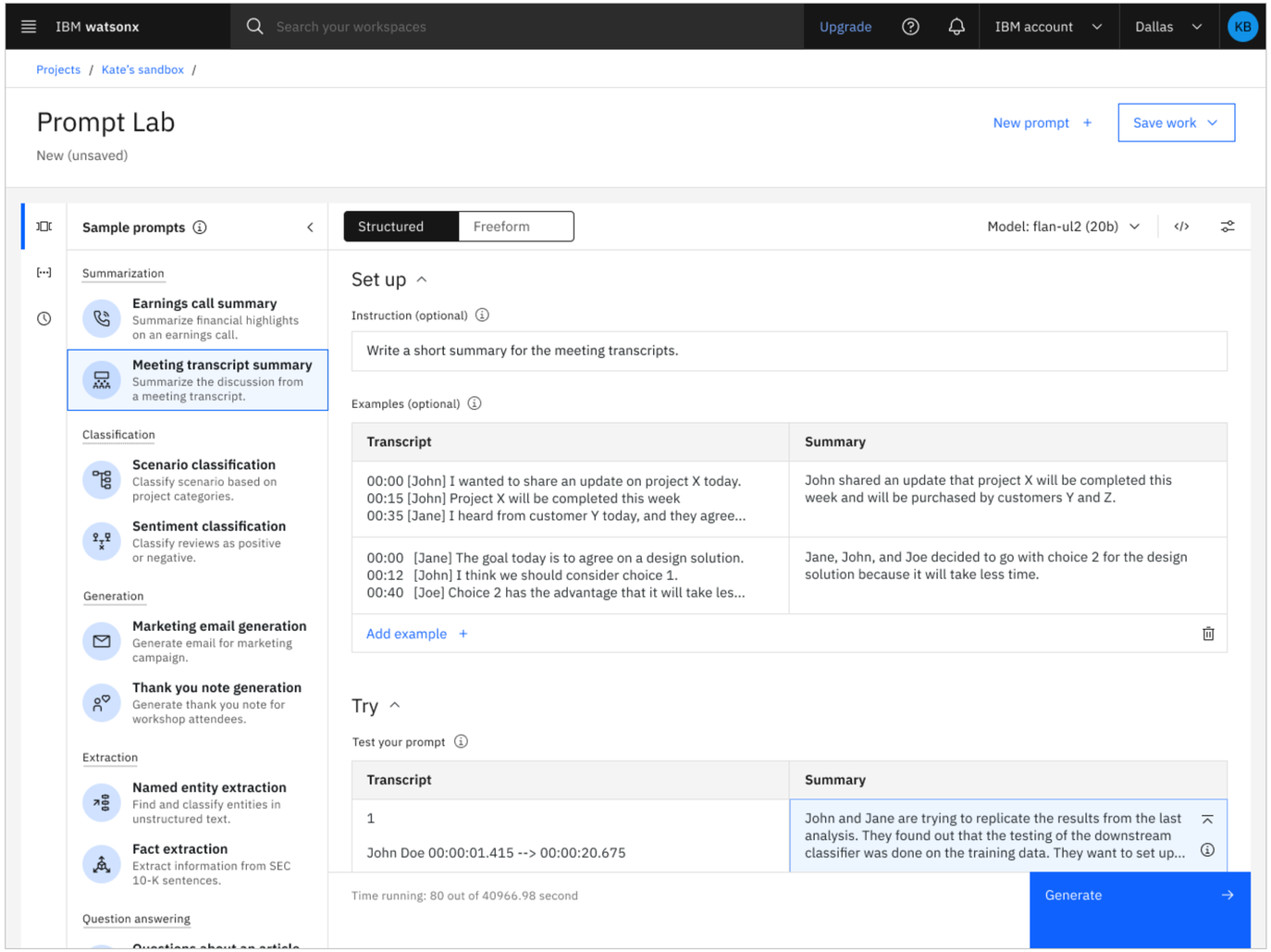
Experiment with prompt engineering

Choice of foundation models to use based on task requirements

Prevent the model from generating repeating phrases

Number of min and max new tokens in the response

Stop sequences – specifies sequences whose appearances should stop the model



watsonx.ai: Tuning Studio

Tune your foundation models with labeled data

Summary:

- Tool for performing PEFT and fine-tuning training techniques to optimize FM task performance
- Tuned model can be deployed and inferenced via the API or Prompt Lab

Initial tuning method at GA: Prompt-tuning

- **How it works:** creates an optimized sequence of values (called a soft-prompt vector) to add as a prefix to FM prompt to improve task performance
- **Technical origins:** The Power of Scale for Parameter-Efficient Prompt Tuning
- Subset of PEFT, similar to P-Tuning, LoRA, etc.

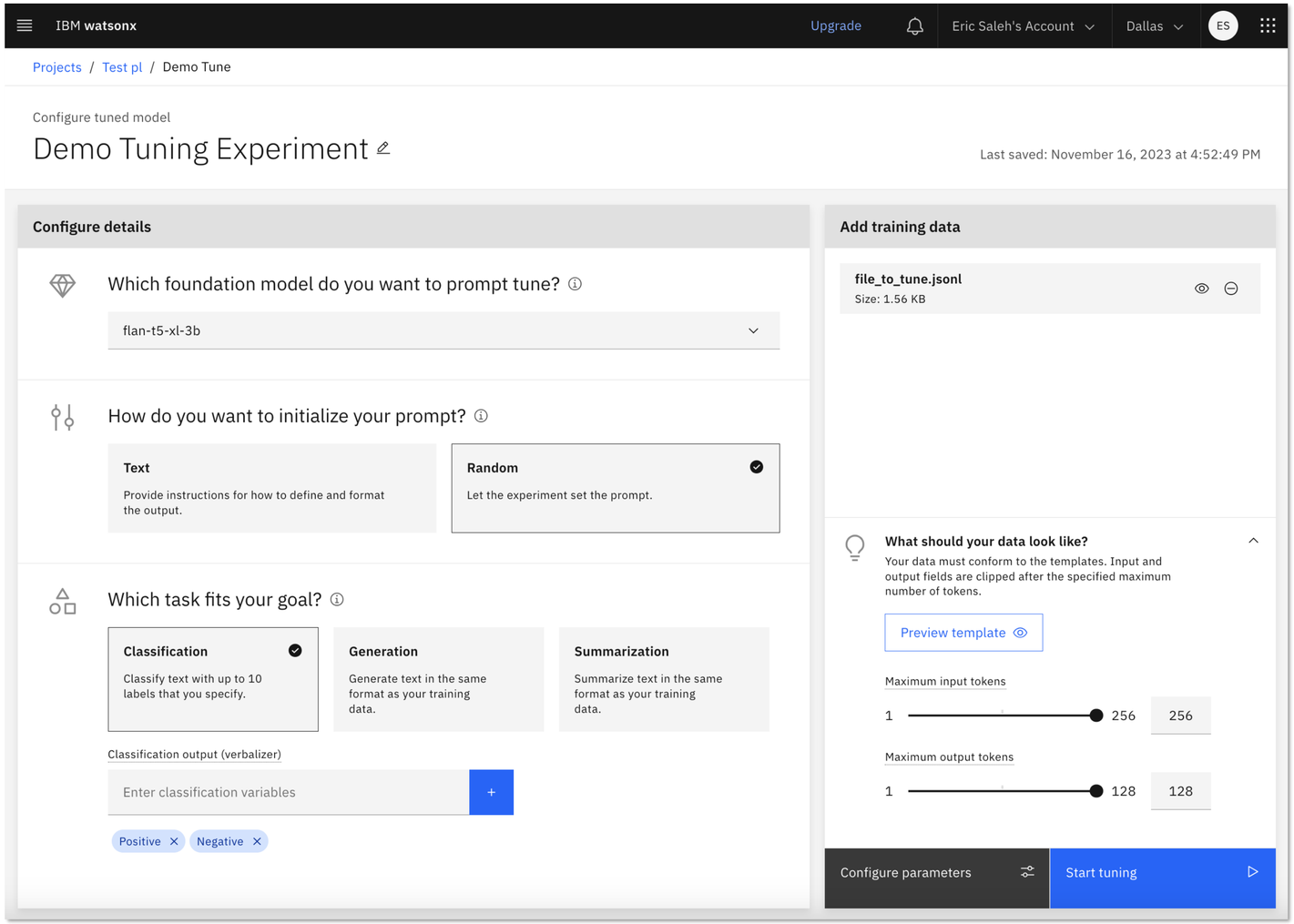
FMs eligible for prompt-tuning:

- flan-t5-xl-3b, llama-2-13b-chat, granite-13b-instruct-v2
- SaaS (Dallas, Tokyo, London, Frankfurt DC)
- Additional FMs currently in-development

Pricing:

- 43 capacity-unit-hours (CUH) rate per hour of active tuning
- Inference Resource-Unit price for deployed tuned model depends on FM inferencing class (learn more)

Product documentation



watsonx.ai: Data Science and MLOps

Build machine learning models automatically in the studio

Model training and development

Build experiments quickly and enhance training by optimizing pipelines and identifying the right combination of data

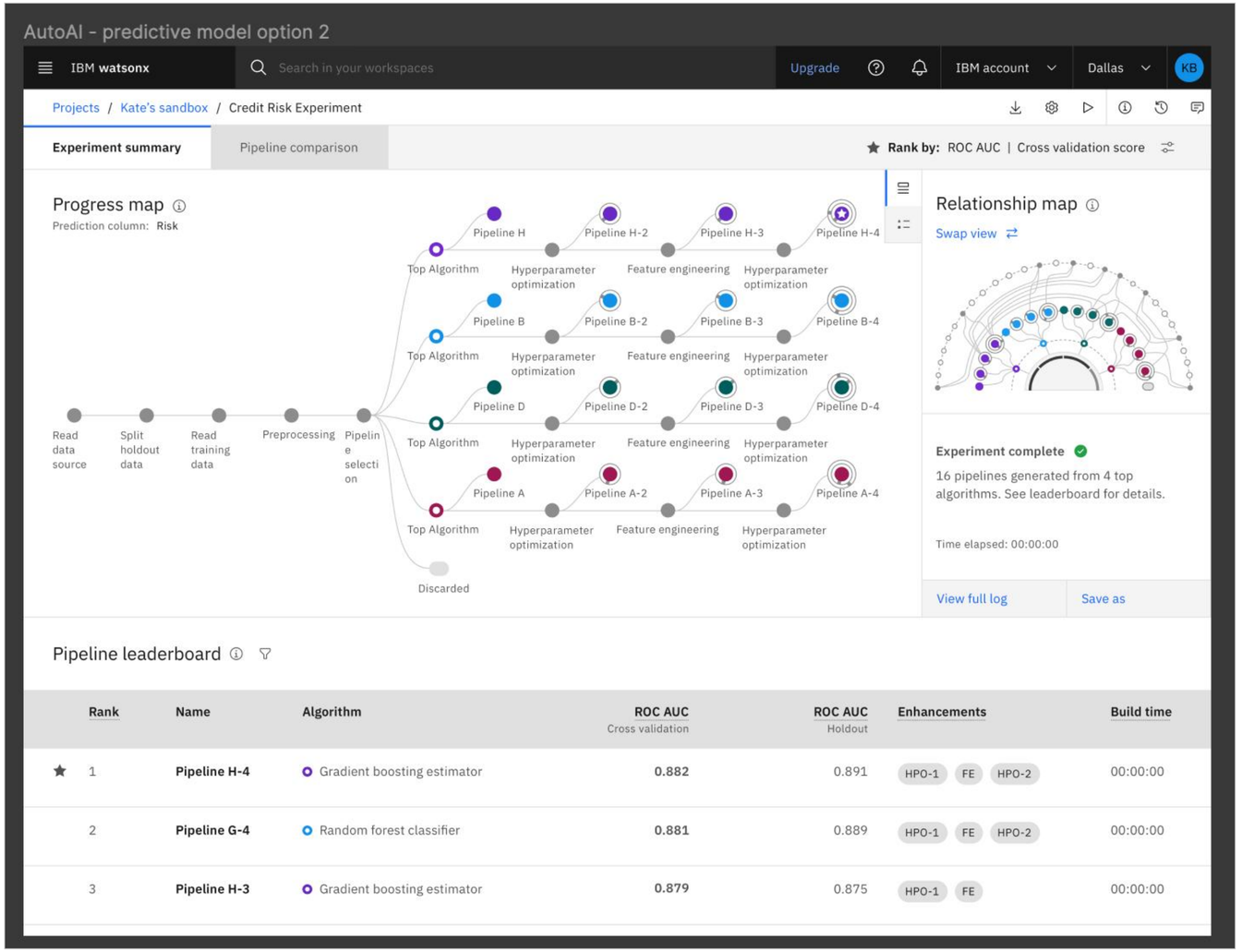
AutoAI, including preparing data for machine learning and generating and ranking candidate model pipelines

Use predictions to optimize decisions, create and edit models in Python, in OPL or with natural language

Integrated visual modeling

Prepare data quickly and develop models visually to help visualize and analyze enterprise data to identify patterns and trends, explore opportunities, and make informed, insightful business decisions

- Uncover correlations
- Insight for hypotheses
- Find relationships and connections within the data



watsonx.ai: Synthetic Data Generator

Generate synthetic tabular data to address your data gaps

Create synthetic data at scale

Unlock your valuable insights by using synthetic data.

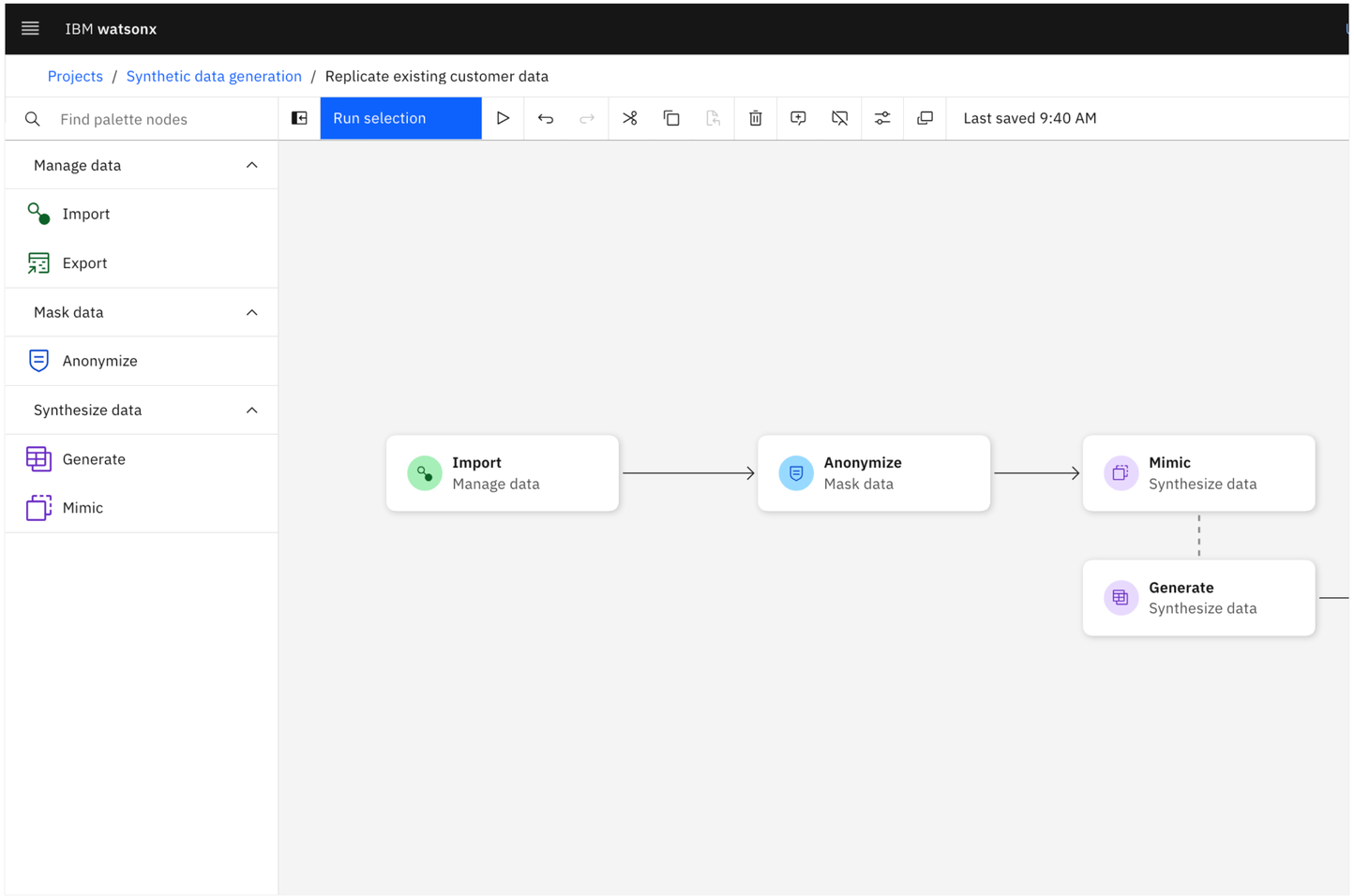
Create synthetic data using your existing data in a database or by uploading a file. If no data exists or can't be accessed, you can design your own data schema.

Address data gaps and create synthetic edge cases to expedite classical AI model training.

Select your model & privacy needs

Depending on your cost, fidelity, application, or data needs, you can select from multiple IBM models* to create your synthetic tabular data.

When using existing data, IBM models apply differential privacy to minimize your privacy risk and give you control over the level of privacy protection required for your organization.



Evaluation metrics available in Q3 2024

watsonx.ai value proposition

Improved performance

- Developing specialized models to produce better results for targeted tasks with lower infrastructure requirements to achieve improved price-performance, *(granite.13b for financial tasks).*

- Enhancement of models delivered through model refresh (granite.13b.V2), new models developed by IBM (e.g., granite.20b multilingual), or 3rd
- party models

3x Price-cuts

- granite.13b [3X less cost] available today at $0.0006 1,000 tokens (input/output)
- llama2.70b [2.7X less cost] available today at $0.0018 1,000 tokens (input/output)
- Llama2 13b [3X less cost] available today at $0.0006 1,000 tokens (input/output)

Multi-lingual support

Expanding language support beyond English through a combination of 3rd party model providers and IBM-developed multi-lingual models that support:

- English
- Japanese
- Spanish
- Portuguese
- French
- German

Differentiated Client Protection

IBM stands behind IBM-developed models and indemnifies the client against third-party IP claims.
IBM offers an additional peace of mind to clients by:

- not requiring them to indemnify IBM for their use of its models
- not capping its IP indemnification liability

IBM Power + watsonx.data

Anchoring on accessibility & quality, scale and data sovereignty

Ensure accessible, quality data
on scalable, reliable infrastructure to drive enterprise-wide AI

watsonx.data

<5 min to access data

Pre-packaged connectors to key Power data sources

Full control over data

Multiple built in features to establish governance and easily transform data.

Power

Scalable insights

Scale data transformation with performant infrastructure.

Data sovereignty

Keep all enterprise data on reliable infrastructure

IBM watsonx.data on IBM Power - *Scale your AI initiatives with confidence*

IBM Power
high-performant, always available, secure infrastructure

+

IBM watsonx.data
open, converged, enterprise-ready data platform

Challenge 1:

Relevant Data

Co-locate your data

Keep both transactional and analytical data in the same infrastructure

Access data quickly

Built-in connectors for key data sources

Challenge 2:

Security & Governance

Protect your data

Ensure compliance, security and business continuity by keeping it on IBM Power

Control data holistically

Built -in data governance capabilities

Challenge 3:

AI Reliability & Cost

Efficiently query your data

Efficiently process growing data volumes with higher throughput per core*

Optimize data querying

Packaged fit-for-purpose query engines to run queries at scale

Seamless support

with a single point of contact across the stack

** Based on IBM Internal Testing.*

Multiple early adopters, multiple geographies and multiple verticals

3 key examples.....

Government

Government telecom authority in Europe wants to offload/archive cold data from Db2 Warehouse to watsonx.data.

Optimize resources while keeping *critical enterprise data* on a *trusted and available* infrastructure.

Insurance

Life insurance company in North America running DB2 on AIX wants to leverage fit-for-purpose query engines (Presto and Spark) to run ETL at optimized costs

Optimized cost-performance to process data on a scalable infrastructure.

Heavy equipment

Heavy equipment rentals firm in North America running DB2 for i on IBM Power wants to have high performance Data Fabric using watsonx.data for multiple AI use cases.

Gain secure access to enterprise sensitive data to *quickly* enhance knowledge bases for genAI use cases.

Red Hat SW for MMA & Spyre – *Red Hat Openshift AI (RHOAI) & Red Hat AI Inference Server (RHAIIS)*

RHOAI

Client Value	*Build, run* and *manage* AI flexibly	*Easily integrate AI* with modernized applications	*Accelerate time-to-value* with fit-for-purpose hardware acceleration.
Key Capability	Single software for Model training*, inference and monitoring on-premise or on IBM PowerVS.	Run AI on the same Power platform as modernized apps	Unified software stack to seamlessly move workloads between MMA (on-chip) and Spyre (off-chip) accelerators for optimal performance.

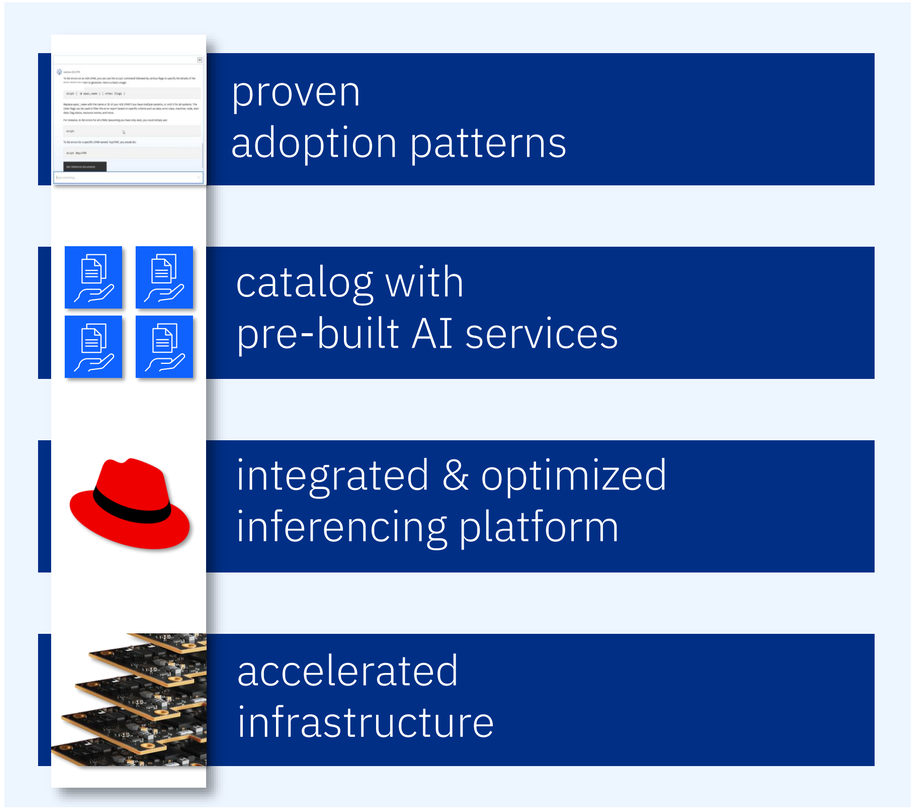
** Here training is only for Predictive AI use cases*

RHAIIS

Client Value	*Simple to get started*	*Flexible*	*Optimized*
Key Capability	Single container (vLLM) to run genAI use cases	Can run on RHEL or Openshift	Tuned to run on IBM Spyre for targeted genAI use cases

IBM Spyre™ for Power

Turnkey AI for enterprise workloads.



1 click

...to install AI services from the IBM-supported catalog.[1]

1 configuration

...to move AI services of the IBM-supported catalog between IBM Power & IBM Power Virtual Server.[2]

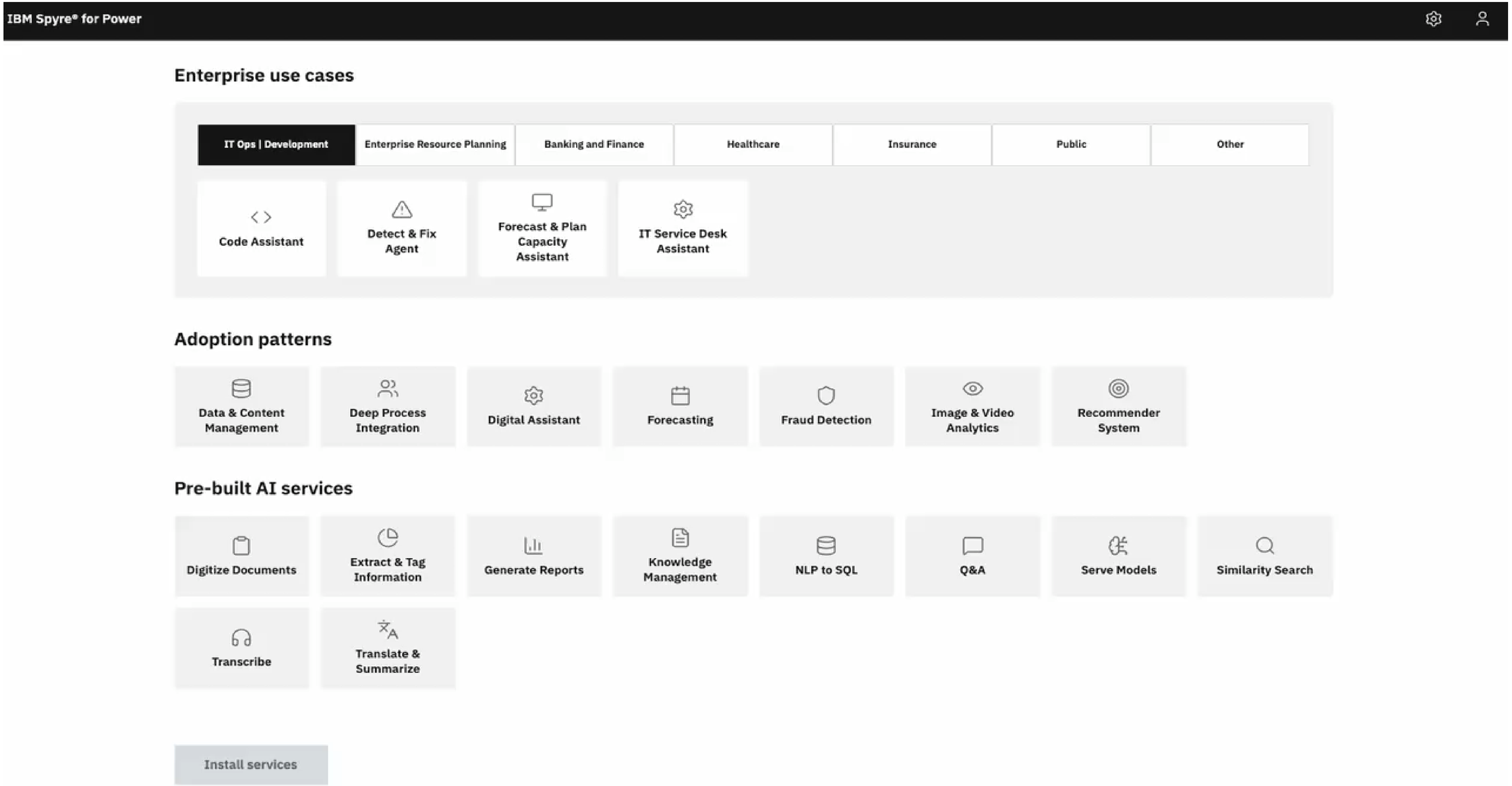
>8 million/hour

...document embeddings for knowledge base integration
using IBM Spyre™ Accelerator for Power with batch and prompt sizes of 128.[3]

Disclaimer: 1: AI services of the IBM-supported catalog are delivered as one or a set of containers that can be deployed with a single deployment command. The provided UI for the catalog executes such commands in the backend based on a single click within the UI page of the respective AI service. 2: A single configuration is enabled by exposed industry standard APIs to decouple services at the top and the backing inferencing service for all AI services that are part of the IBM-supported catalog. Any service that requires AI inferencing capabilities can connect inferencing services that provide OpenAI API or watsonx.ai API compliant inferencing endpoints (Spyre endpoint, RH AI Inferencing Server, IBM Cloud, OpenAI, Azure, AWS, GCP, ...). Services can run either on IBM Power or on IBM Power Virtual Server. 3: Based upon internal testing running 1M unit data set with prompt size 128, batch size 128 using 1-card container. Individual results may vary based on workload size, use of storage subsystems and other conditions.

AI Services for Power: GenAI Use cases

http://ibm.biz/aiservices



Leverage the appropriate AI-optimized software as needed to build your AI solutions.

Open-source
(with option of enterprise support)

Get started with AI use cases *quickly at minimal cost*

→ 1000’s of open-source packages and tools optimized for Power.
→ optional enterprise support

RedHat Openshift AI
(RHOAI)

End-to-end AI lifecycle *managed easily at scale*

→ Enhanced developer experience
→ Secure and compliant
→ Scalable across the enterprise

IBM Enterprise Data & AI
(watsonx.data, CP4D)

Enterprise-wide adoption of AI, *at scale*

→ Integrated platform for data analysis, organization and management.
→ Low-code/no-code AI development

ISV Solutions
(Industry-specific or AI use-case specific)

Jumpstart your journey with *pre-packaged AI solutions*

→ Enhance core applications (Core banking, ERP, ECM) with built-in AI capabilities.
→ AI-ready solutions for specific use cases.

AI options for IBM i

Cloud services

On-premise

On-platform

Off-the-shelf

Custom solution focus

AI Services for Power

watsonx™

AI Services for Power

watsonx™

ROCKETGRAPH

Wallaroo.AI

Red Hat AI

Python Ecosystem for IBM Power

LLaMA.cpp

Chainer

scikit learn

Thank you