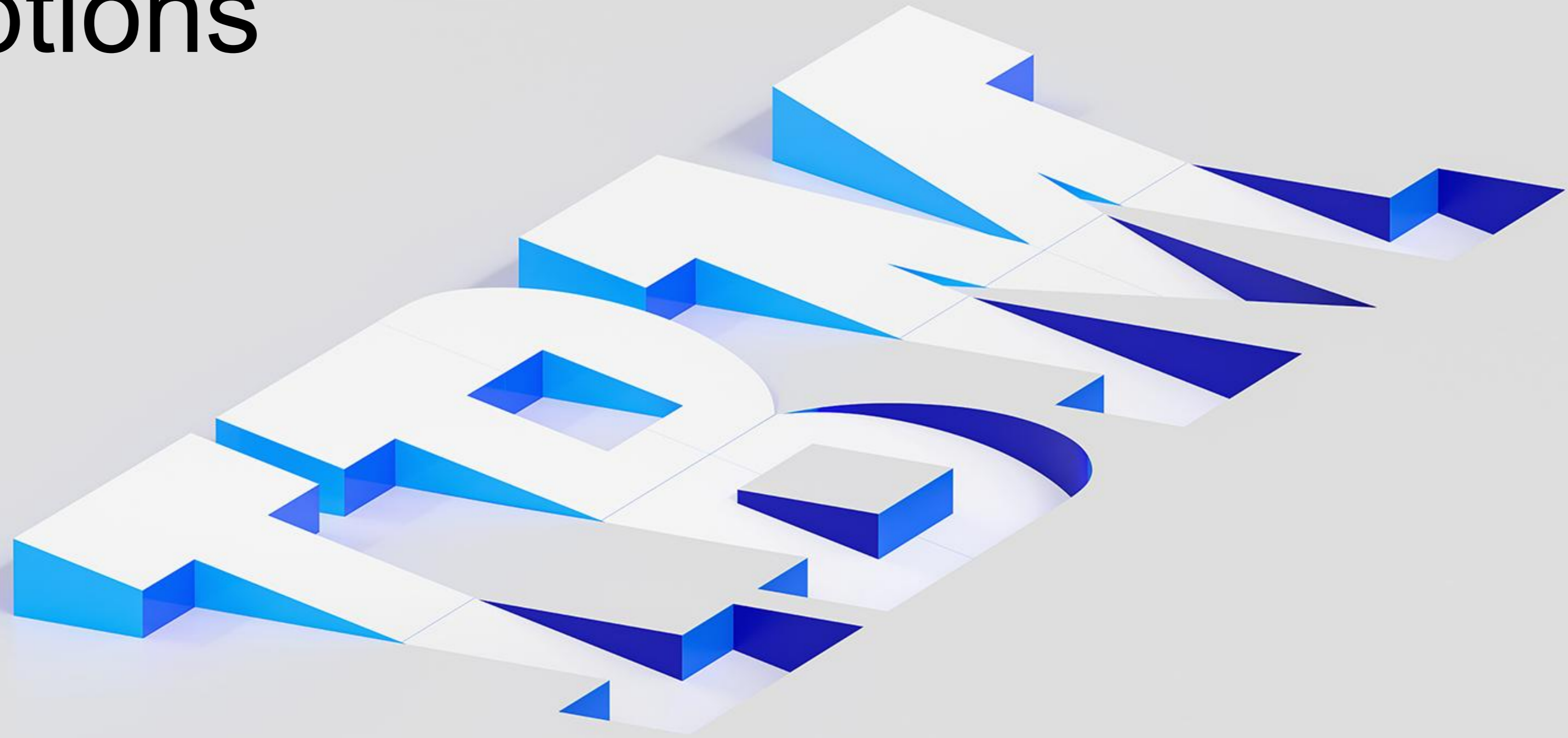


Module 1: AI Stack options



Three main principles

Hardware

matériel
informatique

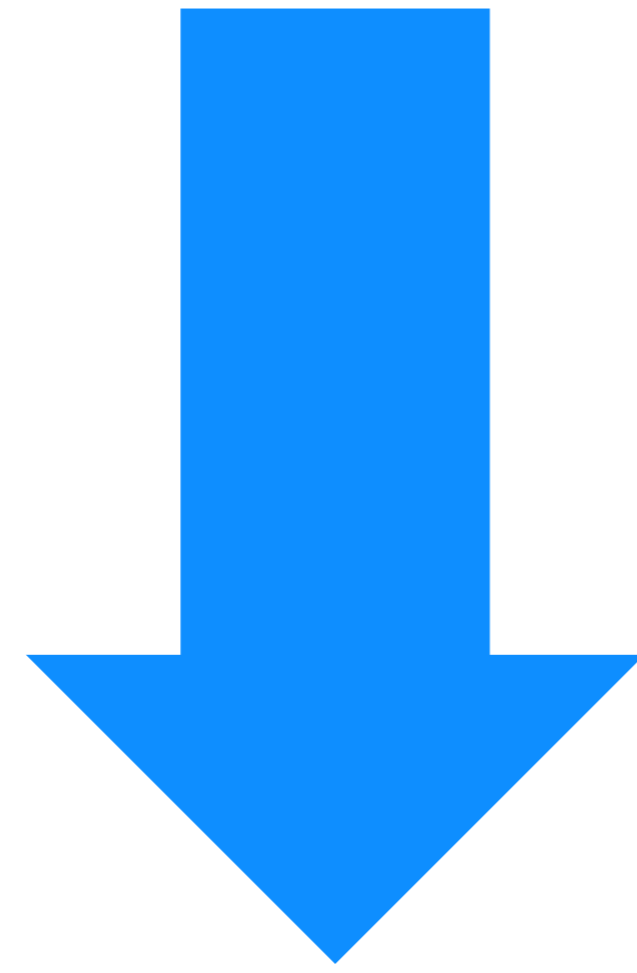
Software

logiciel

Integration

intégration

Three main principles



Hardware

matériel
informatique

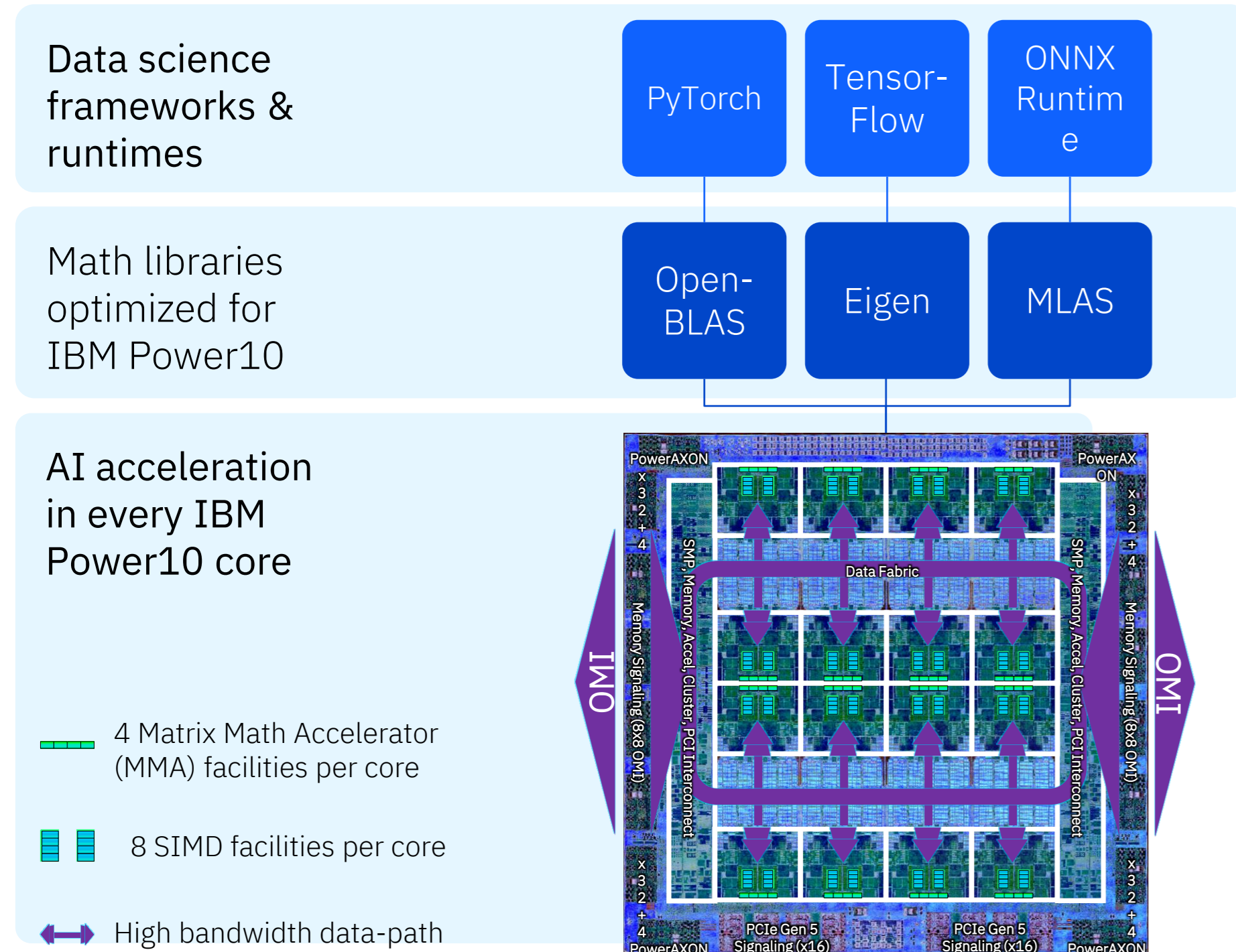
Software

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

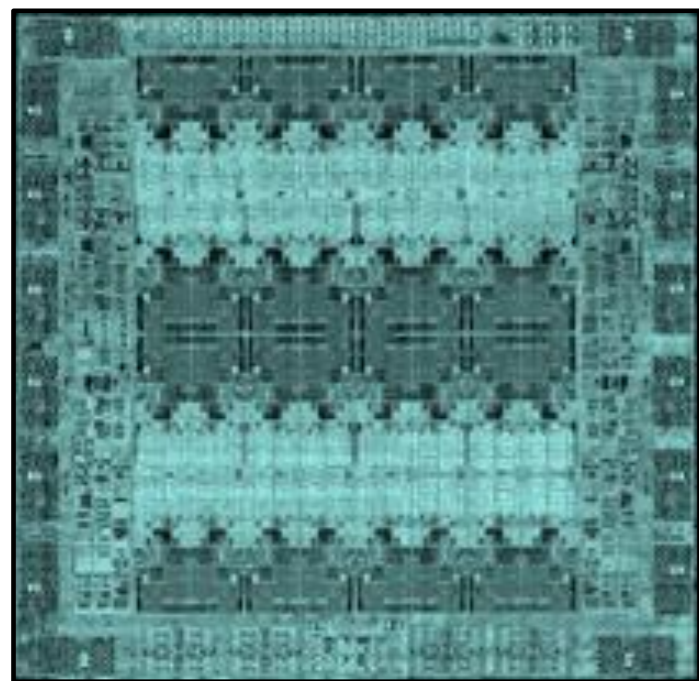
Feature	AI Method	GPU	P10 with MMA
Training	Deep Learning	Best Fit (cost-perf)	Limited Benefit
	Machine Learning	Limited Benefit (cost-perf)	Best Fit (cost-perf)
	Foundation Model (like GenAI)	Best Fit (cost-perf)	Not Optimal
Re-training / Fine-tuning	Deep Learning	Best Fit (cost-perf)	Limited Benefit (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

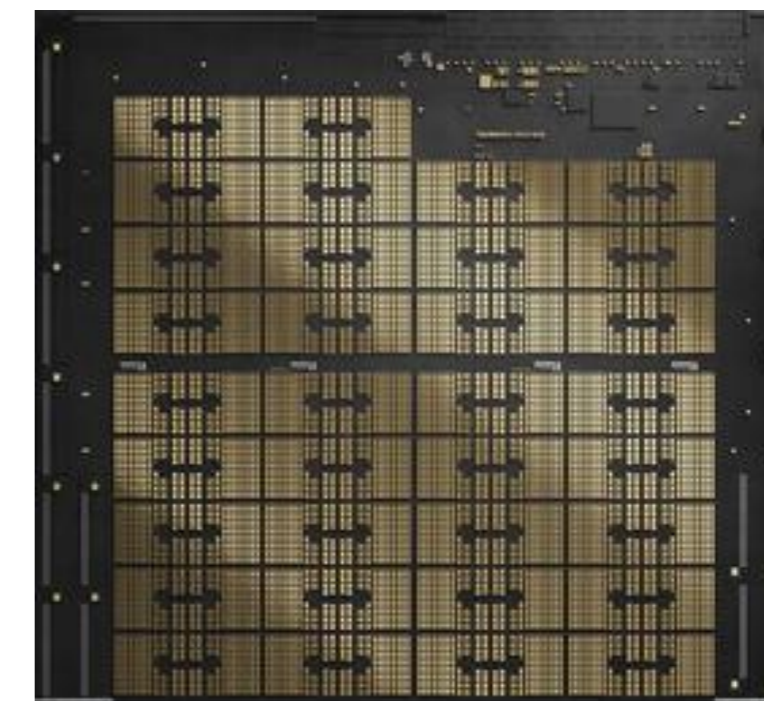
Power 11 Processor Innovations + Spyre Accelerator + Growing AI Stack

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

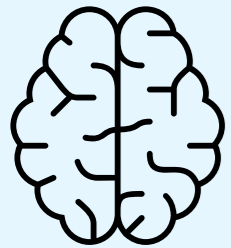
Data Stage – SOD in place
Knowledge Catalog – SOD in Place
watsonx.data – SOD being worked

Consumption Model

PowerVS + watsonx – toolkits
On-Premise – OpenSource or Enterprise SW
Simplified

Artificial Intelligence (AI)

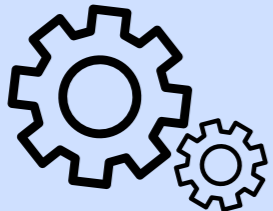
Human intelligence exhibited by machines



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)

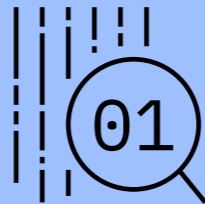
Systems that learn from historical data



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)

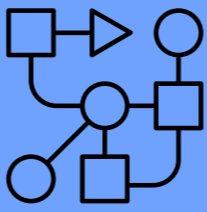
ML technique that mimics human brain function



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

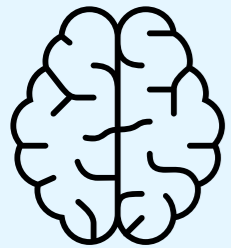


AI model built using a specific kind of neural network architecture, called a transformer, which is designed to generate sequences of related data elements (for example, like a sentence).



Artificial Intelligence (AI)

Human intelligence exhibited by machines

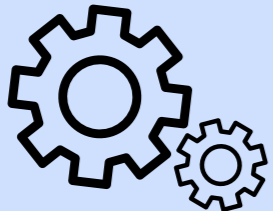


AI can be defined as a technique that enables machines to mimic cognitive functions associated with human minds – cognitive functions include all aspects of learning, reasoning, perceiving, and problem solving.

Machine Learning (ML)

Systems that learn from historical data

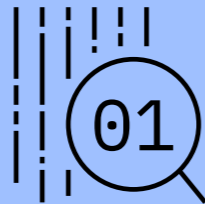
ML-based systems are trained on historical data to uncover patterns. Users provide inputs to the ML system, which then applies these inputs to the discovered patterns and generates corresponding outputs.



Deep Learning (DL)

ML technique that mimics human brain function

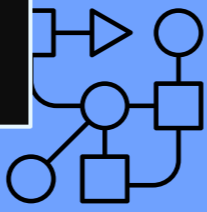
DL is a subset of ML, using multiple layers of neural networks, which are interconnected nodes, which work together to process information. DL is well suited to complex applications, like image and speech recognition.



Generative AI (LLMs)

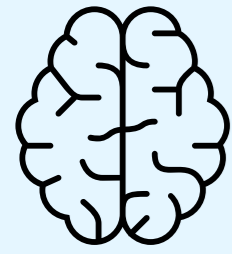
Foundation Models

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).



Artificial Intelligence (AI)

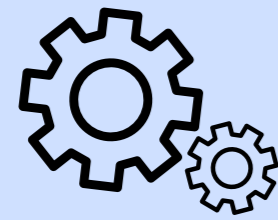
Human intelligence exhibited by machines



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)

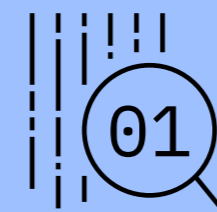
Systems that learn from historical data



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

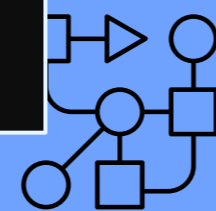


DL is a subset of ML, using multiple layers of neural networks, which are interconnected nodes, which work together to process information. DL is well suited to complex applications, like image and speech recognition.

MMA

Generative AI (LLMs)

Foundation Models



Spyre

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



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

The logo for LLaMA++, consisting of the text "LLaMA++" in a white, bold, sans-serif font on a dark rectangular background. The "++" is in orange.

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

The screenshot shows the GitHub repository page for 'pyeco' by user 'ppc64le'. The repository is public and has 5 watchers, 10 forks, and 28 stars. The repository contains several files and folders, including 'DevpiWheelsIndex', 'configs', 'examples', '.gitignore', 'DevpiWheelsIndex.md', 'LICENSE', and 'README.md'. The README file is currently selected and displays the following content:

pyeco

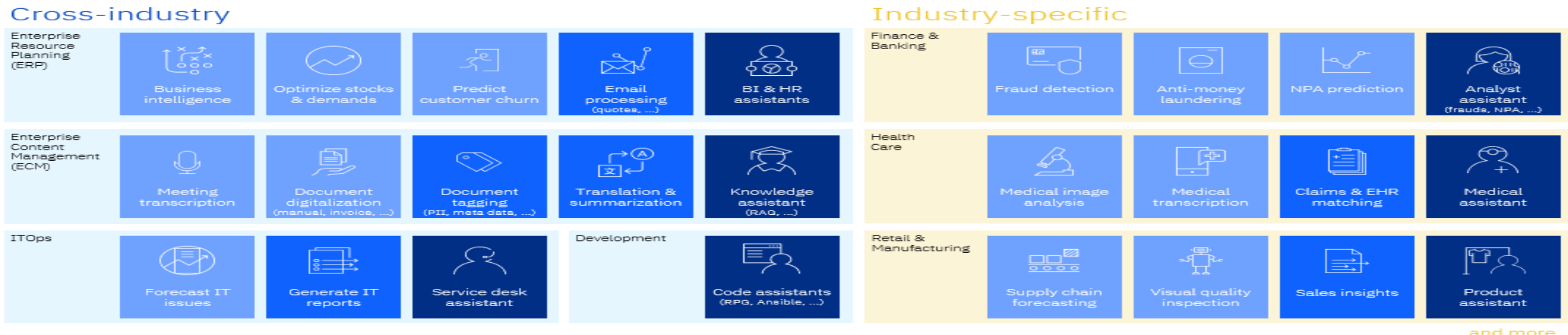
Python Ecosystem for Power Hints and Tips, Issue Tracking

NOTE: We are starting to add a few requirements.txt base examples and a few tested programs using those to help people get started. Like with Pypi, there are a lot of possible permutations and combinations and not all will be tested, but the goal is that there should be a clear path for major python toolset combinations. We will also

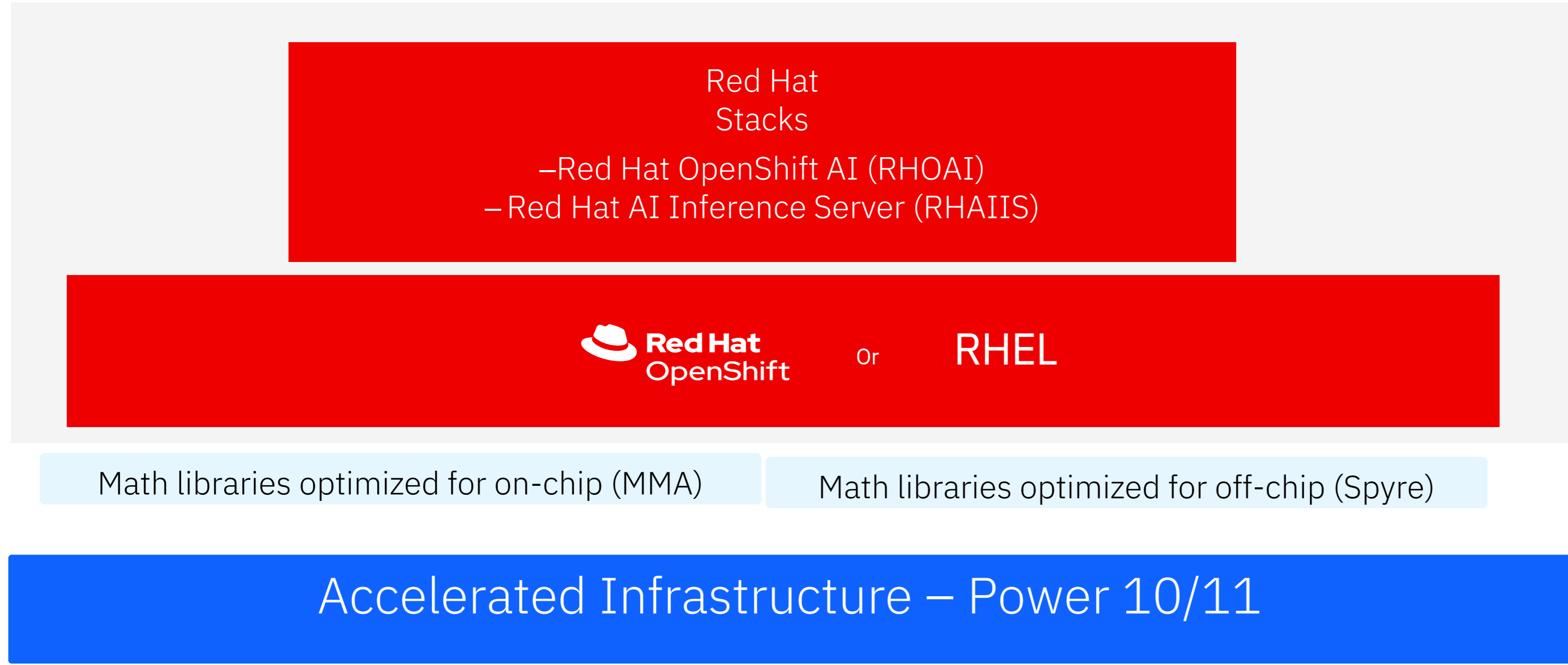
The right sidebar of the repository page shows the following information:

- About:** Python Ecosystem for Power Hints and Tips, Issue Tracking
- Readme:** Readme
- License:** Apache-2.0 license
- Activity:** Activity
- Custom properties:** Custom properties
- Stars:** 28 stars
- Watching:** 5 watching
- Forks:** 10 forks
- Releases:** No releases published
- Packages:** No packages published
- Contributors:** 8 contributors

Red Hat SW Stack for AI



AI use cases
Key use cases for IBM Power clients



Optimized AI foundation
→ RHOAI to build, run and manage AI apps.
→ RHAIS for simple, performant genAI inferencing on IBM Spyre

Operating system
→ RHEL LPAR (RHAIS)
→ Openshift – RHOAI or RHAIS

Optimized compute and libraries
for accelerating AI.

Red Hat AI overview for IBM Power



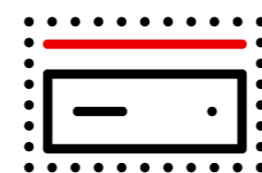
Gen AI model Inference on
RHEL/Linux or OpenShift/Kube



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

Trusted, Consistent and Comprehensive foundation

Hardware Acceleration



Virtual



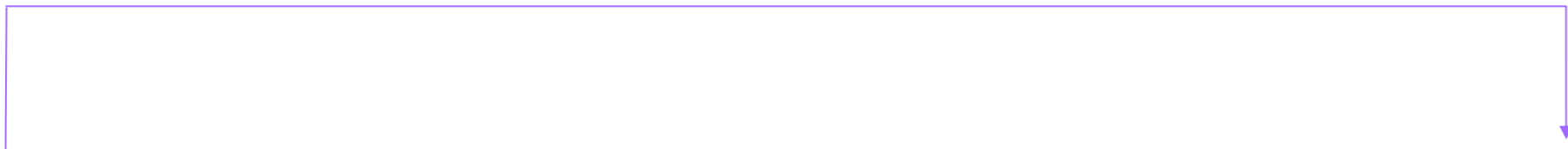
Private
Cloud



Public Cloud



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

The screenshot shows the IBM watsonx Prompt Lab interface. The top navigation bar includes the IBM watsonx logo, a search bar for workspaces, and user account information (IBM account, Dallas, KB). The main header displays 'Prompt Lab' and 'New (unsaved)', with buttons for 'New prompt +' and 'Save work'. The interface is divided into a left sidebar and a main workspace. The sidebar lists various prompt categories: Summarization (with 'Meeting transcript summary' selected), Classification (with 'Scenario classification' and 'Sentiment classification'), Generation (with 'Marketing email generation' and 'Thank you note generation'), Extraction (with 'Named entity extraction' and 'Fact extraction'), and Question answering. The main workspace is titled 'Set up' and shows a structured prompt configuration. The 'Instruction (optional)' field contains 'Write a short summary for the meeting transcripts.' The 'Examples (optional)' section displays a table with two columns: 'Transcript' and 'Summary'. The first example shows a transcript snippet and its corresponding summary. The second example shows another transcript snippet and its summary. Below the examples is an 'Add example +' button. The 'Try' section includes a 'Test your prompt' field and a table with 'Transcript' and 'Summary' columns. The transcript field contains '1' and 'John Doe 00:00:01.415 --> 00:00:20.675'. The summary field contains 'John and Jane are trying to replicate the results from the last analysis. They found out that the testing of the downstream classifier was done on the training data. They want to set up...'. At the bottom, a 'Generate' button is visible, along with a 'Time running: 80 out of 40966.98 second' indicator.

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](#)

The screenshot shows the IBM watsonx Tuning Studio interface for a "Demo Tuning Experiment". The top navigation bar includes the IBM watsonx logo, an "Upgrade" button, a notification bell, the user's account "Eric Saleh's Account", the location "Dallas", and a profile icon "ES". The breadcrumb trail is "Projects / Test pl / Demo Tune".

The main content area is titled "Configure tuned model" and "Demo Tuning Experiment". It is divided into two main sections: "Configure details" and "Add training data".

Configure details:

- Which foundation model do you want to prompt tune?** A dropdown menu is set to "flan-t5-xl-3b".
- How do you want to initialize your prompt?** Two options are shown: "Text" (Provide instructions for how to define and format the output.) and "Random" (Let the experiment set the prompt.), with "Random" selected.
- Which task fits your goal?** Three options are shown: "Classification" (Classify text with up to 10 labels that you specify.), "Generation" (Generate text in the same format as your training data.), and "Summarization" (Summarize text in the same format as your training data.). "Classification" is selected. Below this, there is a "Classification output (verbalizer)" field with a text input "Enter classification variables" and a "+" button. There are also "Positive" and "Negative" tags.

Add training data:

- A file named "file_to_tune.jsonl" is listed with a size of 1.56 KB.
- A section titled "What should your data look like?" provides instructions: "Your data must conform to the templates. Input and output fields are clipped after the specified maximum number of tokens." A "Preview template" button is available.
- Two sliders are shown: "Maximum input tokens" (set to 256) and "Maximum output tokens" (set to 128).

At the bottom, there are two buttons: "Configure parameters" and "Start tuning".

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

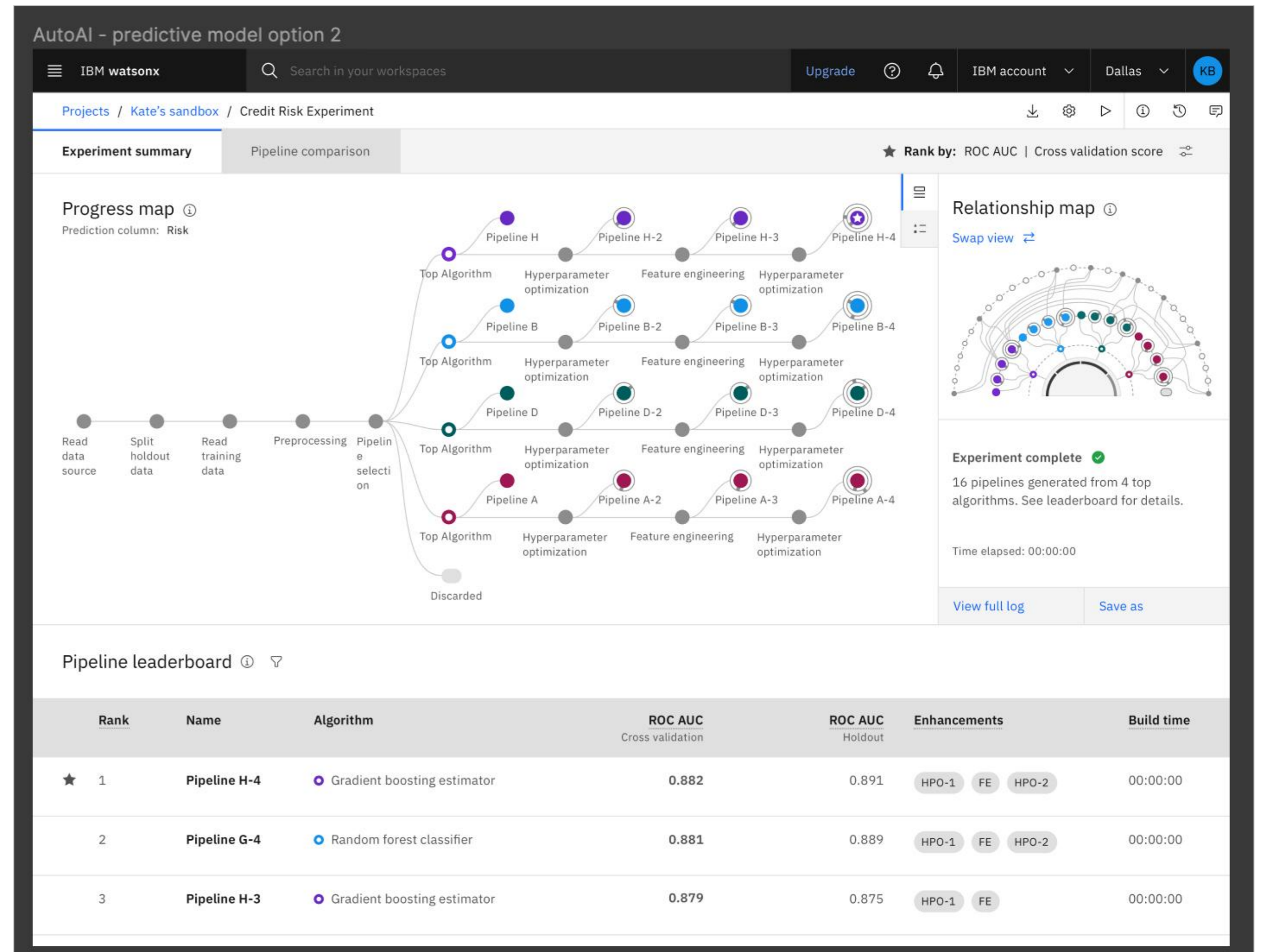
AutoAI, including preparing data for machine learning and generating and ranking candidate model pipelines

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.

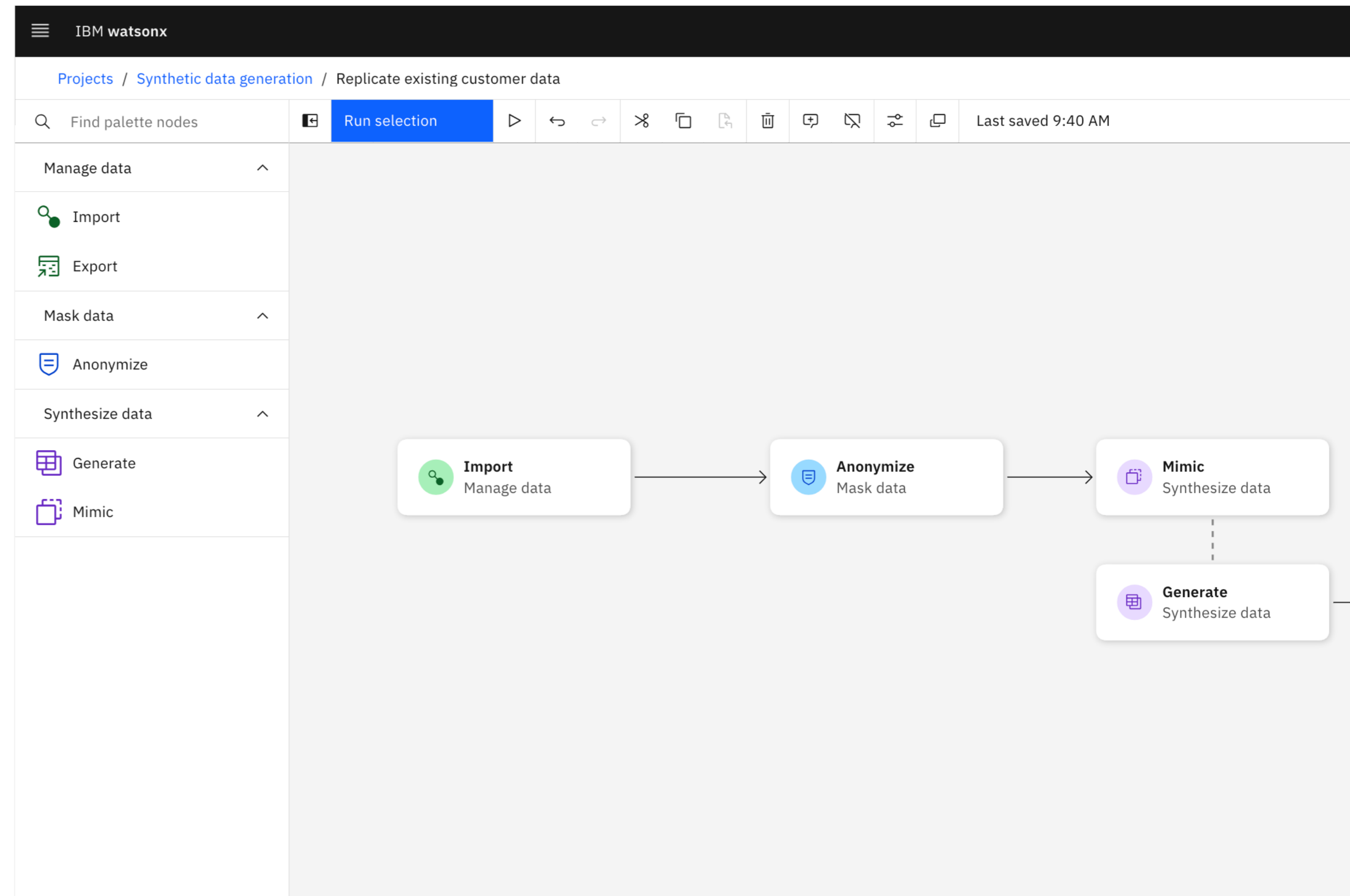
Create synthetic data using your existing data in a database or by uploading a file. If no data exists or can't be accessed, you can design your own data schema.

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

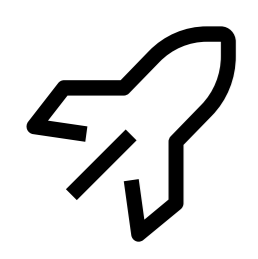
IBM stands behind IBM-developed models and indemnifies the client against third-party IP claims. IBM offers an additional peace of mind to clients by:

- 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



<5 min to
access data

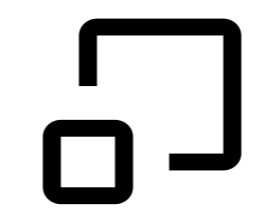
Pre-packaged
connectors to key
Power data sources



Full control
over data

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

watsonx.data



Scalable
insights

Scale data
transformation with
performant
infrastructure.

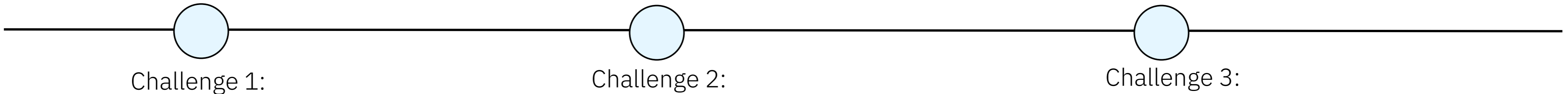
Power



Data
sovereignty

Keep all enterprise
data on reliable
infrastructure

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



Relevant Data

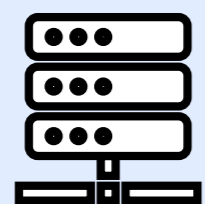
Security & Governance

AI Reliability & Cost

IBM Power
*high-performant,
always available,
secure infrastructure*




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




Co-locate your data

Keep both transactional and analytical data in the same infrastructure




Protect your data

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



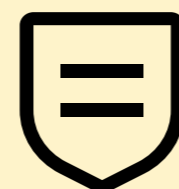
Efficiently query your data

Efficiently process growing data volumes with higher throughput per core*



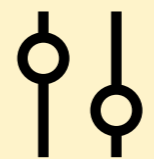
Access data quickly

Built-in connectors for key data sources




Control data holistically

Built -in data governance capabilities



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.

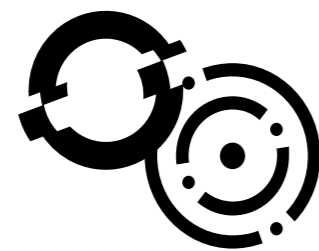
Gain secure access to enterprise sensitive data to *quickly* enhance knowledge bases for genAI use cases.



RHOAI

Client Value

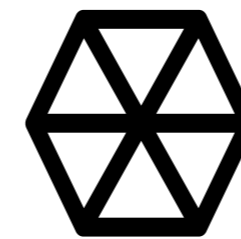
Build, run and *manage* AI flexibly



Key Capability

Single software for Model training*, inference and monitoring on-premise or on IBM PowerVS.

Easily integrate AI with modernized applications



Run AI on the same Power platform as modernized apps

Accelerate time-to-value with fit-for-purpose hardware acceleration.



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

Key Capability

Single container (vLLM) to run genAI use cases

Flexible




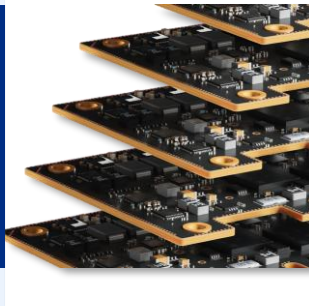
Can run on RHEL or Openshift

Optimized

Tuned to run on IBM Spyre for targeted genAI use cases

IBM Spyre™ for Power

Turnkey AI for enterprise workloads.

-  proven adoption patterns
-  catalog with pre-built AI services
-  integrated & optimized inferencing platform
-  accelerated infrastructure

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>











Enterprise use cases

IT Ops Development	Enterprise Resource Planning	Banking and Finance	Healthcare	Insurance	Public	Other
 Code Assistant	 Detect & Fix Agent	 Forecast & Plan Capacity Assistant	 IT Service Desk Assistant			

Adoption patterns

 Data & Content Management	 Deep Process Integration	 Digital Assistant	 Forecasting	 Fraud Detection	 Image & Video Analytics	 Recommender System
---	---	---	---	---	---	--

Pre-built AI services

 Digitize Documents	 Extract & Tag Information	 Generate Reports	 Knowledge Management	 NLP to SQL	 Q&A	 Serve Models	 Similarity Search
 Transcribe	 Translate & Summarize						

Install services

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
watsonxTM

AI Services for Power
watsonxTM



Thank you