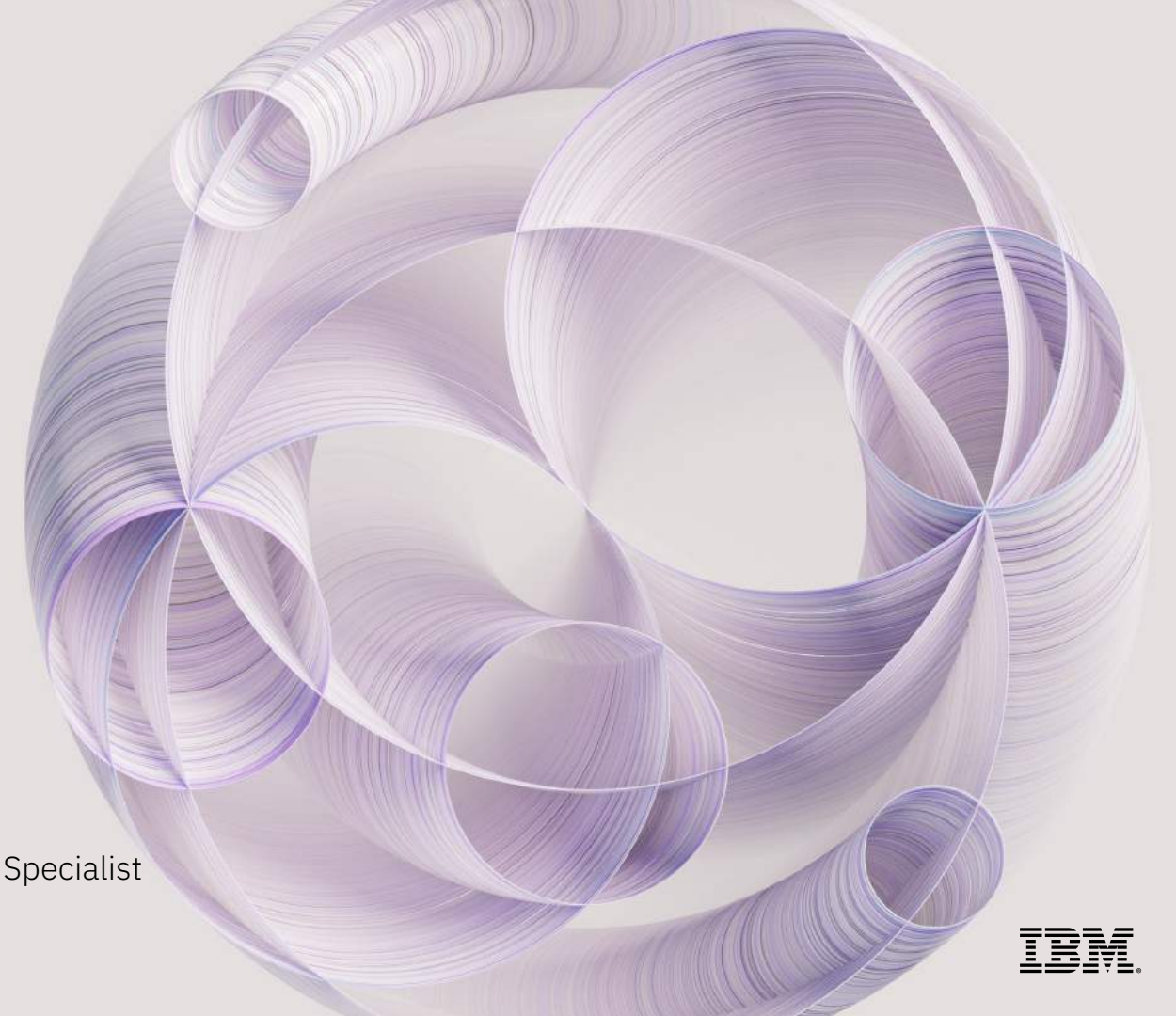
watsonx



Andrea Radovisc – Data & AI Technical Specialist
Andrea.Radovisc@ibm.com

IBM

The platform for AI and data

watsonx

Scale and accelerate the impact of AI across your business

watsonx.ai

Build, train, validate, tune and deploy AI models

A next generation enterprise studio for AI builders to build, train, validate, tune, and deploy both traditional machine learning and new generative AI capabilities powered by foundation models. It enables you to build AI applications in a fraction of the time with a fraction of the data.

watsonx.data

Scale AI workloads, for all your data, anywhere

Fit-for-purpose data store, built on an open lakehouse architecture, supported by querying, governance and open data formats to access and share data.

watsonx.governance

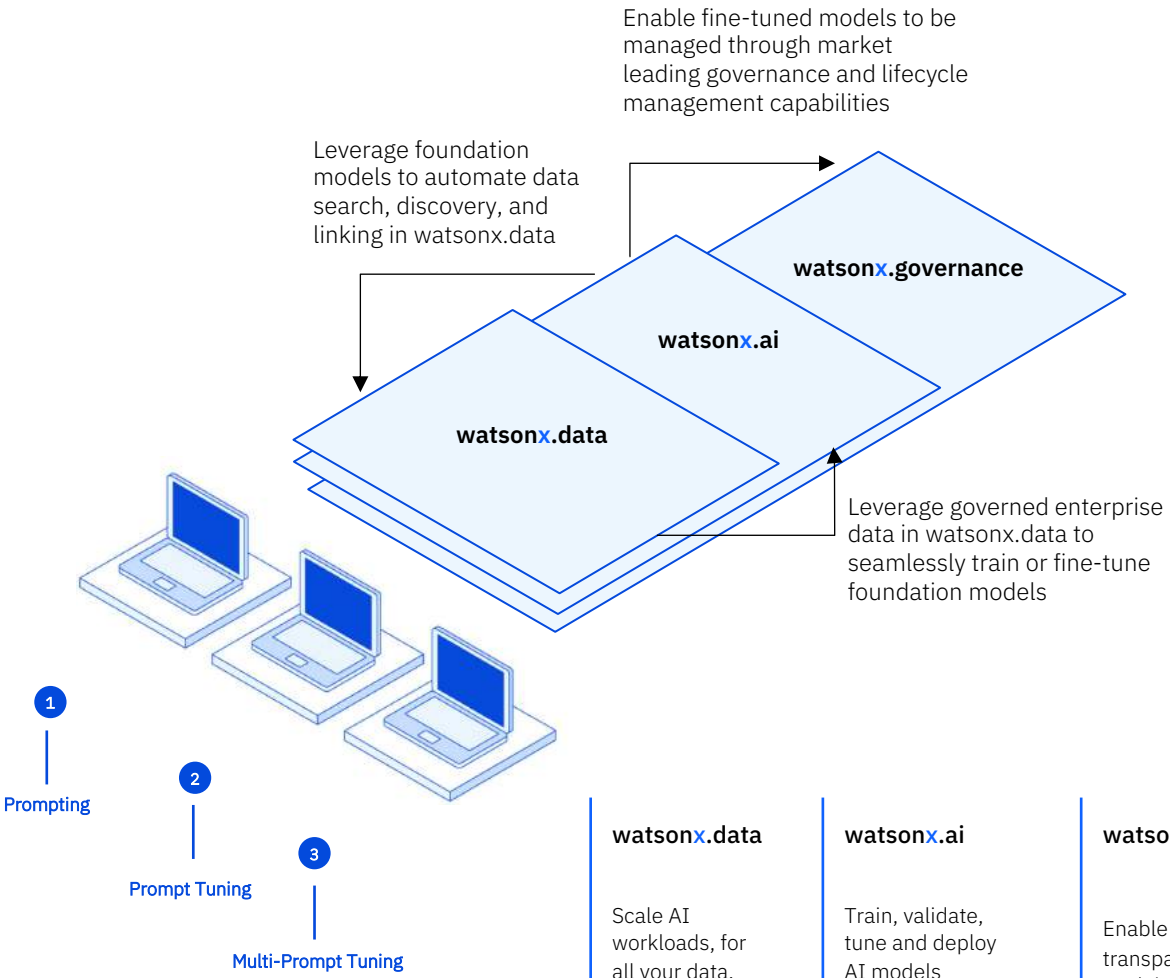
Accelerate responsible, transparent and explainable AI workflows

End-to-end toolkit for AI governance across the entire model lifecycle to accelerate responsible, transparent, and explainable AI workflows

The platform for AI and data |

watsonx

Scale and accelerate the impact of AI with trusted data.



watsonx.data

Scale AI workloads, for all your data, anywhere

watsonx.ai

Train, validate, tune and deploy AI models

watsonx.governance

Enable responsible, transparent and explainable AI workflows

The platform for AI and data

watsonx

Scale and accelerate the impact of AI across your business

watsonx.ai

Build, train, validate, tune and deploy AI models

A next generation enterprise studio for AI builders to build, train, validate, tune, and deploy both traditional machine learning and new generative AI capabilities powered by foundation models. It enables you to build AI applications in a fraction of the time with a fraction of the data.

watsonx.data

Scale AI workloads, for all your data, anywhere

Fit-for-purpose data store, built on an open lakehouse architecture, supported by querying, governance and open data formats to access and share data.

watsonx.governance

Accelerate responsible, transparent and explainable AI workflows

End-to-end toolkit for AI governance across the entire model lifecycle to accelerate responsible, transparent, and explainable AI workflows

What is **watsonx.ai?**

Train, validate, tune, and deploy AI models with confidence

Generative AI capabilities

Foundation Model Libraries

Prompt Lab

Tuning Studio

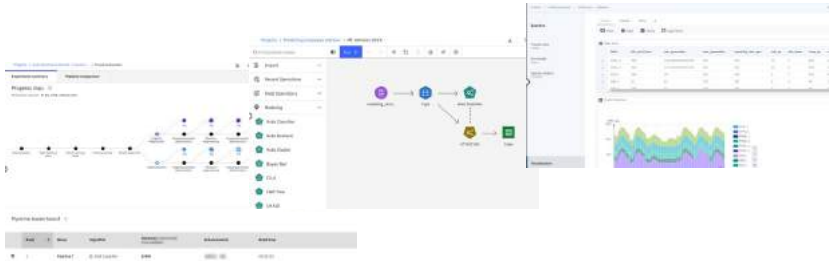
+ a **proven** studio for Machine Learning

ModelOps

Automated Development

Team Collaboration

Decision Optimization



watsonx.ai

Model strategy →

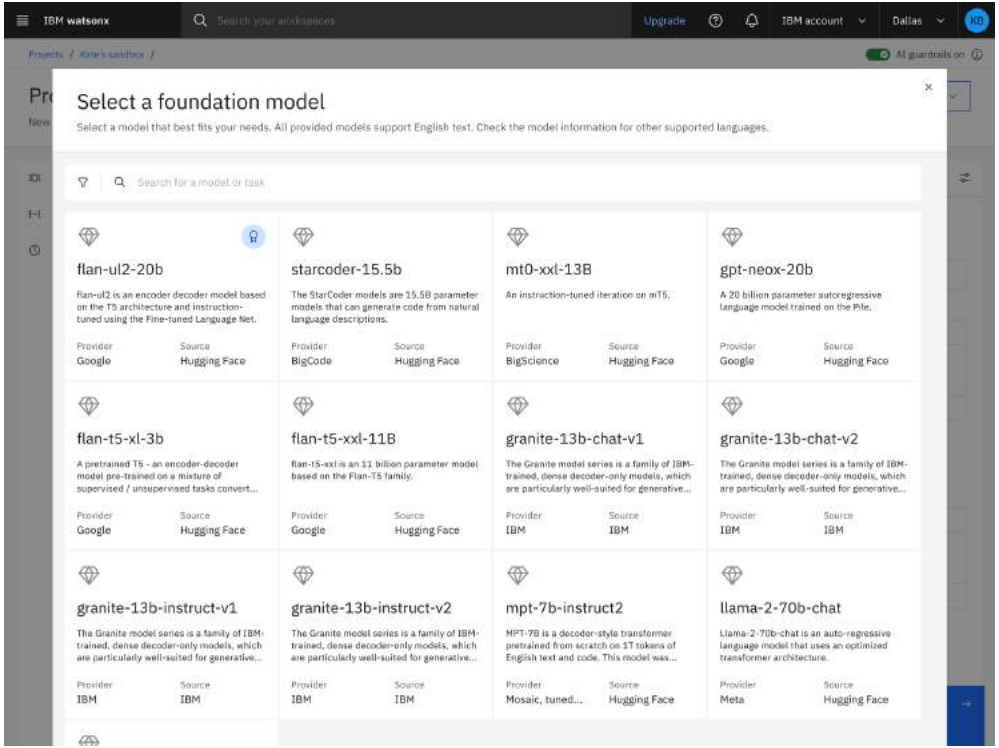
Multi-model

One model doesn't fit all use cases. We offer IBM-developed, open-source, third party, and BYOM.

Bigger is not always better. Specialized models can outperform general-purpose models with lower infrastructure requirements.

Hybrid, multi-cloud

Hybrid deployments. We provide the flexibility to deploy models on the platform of choice.



granite.20b.code is delivered through watsonx Code Assistant

watsonx.ai – Models available

granite-instruct/chat-V2
13 billion params
decoder only

- Q&A
- Generate
- Extract
- Summarize
- Classify
- RAG

llama-2*
13/70 billion params
decoder only

- Q&A
- Generate
- Extract
- Summarize
- Classify
- RAG

*Japanese-language Llama 2 model is available in the Tokyo region (elyza-Japanese-llama-2-7b-instruct)

flan-t5-xl-3b
3 billion params
encoder/decoder

- Q&A
- Generate
- Extract
- Summarize
- Classify
- RAG

flan-t5-xxl-11b
11 billion params
encoder/decoder

- Q&A
- Generate
- Extract
- Summarize
- Classify
- RAG

flan-ul2-20b
20 billion params
encoder/decoder

- Q&A
- Generate
- Extract
- Summarize
- Classify
- RAG

gpt-neox-20b
20 billion params
decoder only

- Generate
- Summarize
- Classify

mixtral 8x7b instruct
8x7 billion params
decoder only

- Q&A
- Generate
- Extract
- Summarize
- Classify
- RAG
- CodeGen

Product documentation link

Natural Language | Code

watsonx.ai: IBM **Granite** models

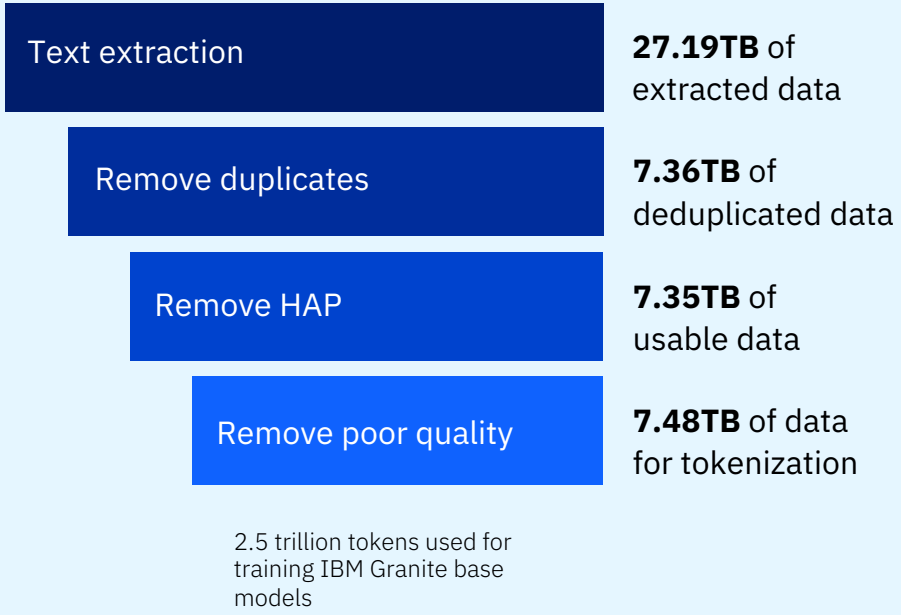
IBM's approach to AI model development is grounded in core principles of trust and transparency.

You can use them for...

- Summarization
- Insight extraction & classification
- Retrieval-Augmented Generation

These models have been trained on enterprise-relevant datasets across these domains:

- Internet
- Academic
- Code
- Legal
- Finance

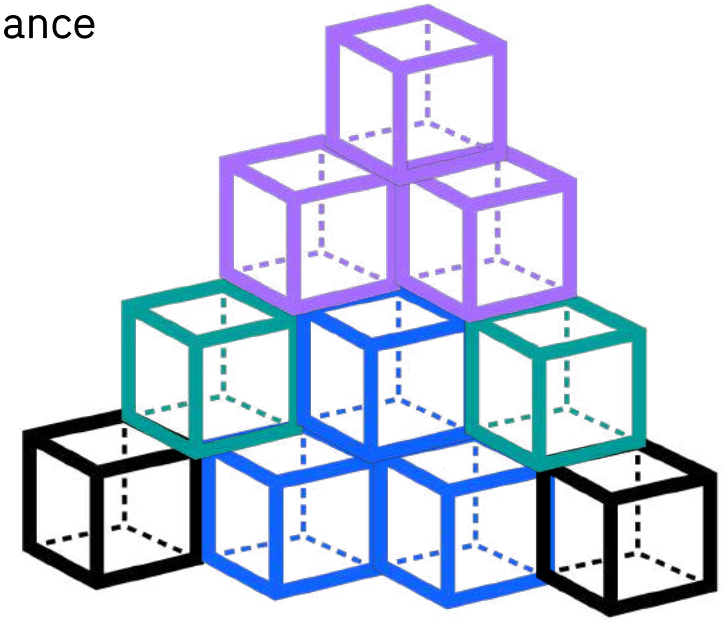


TB = Terabytes

Our approach to selecting **third-party models** in **watsonx.ai**

Technical considerations

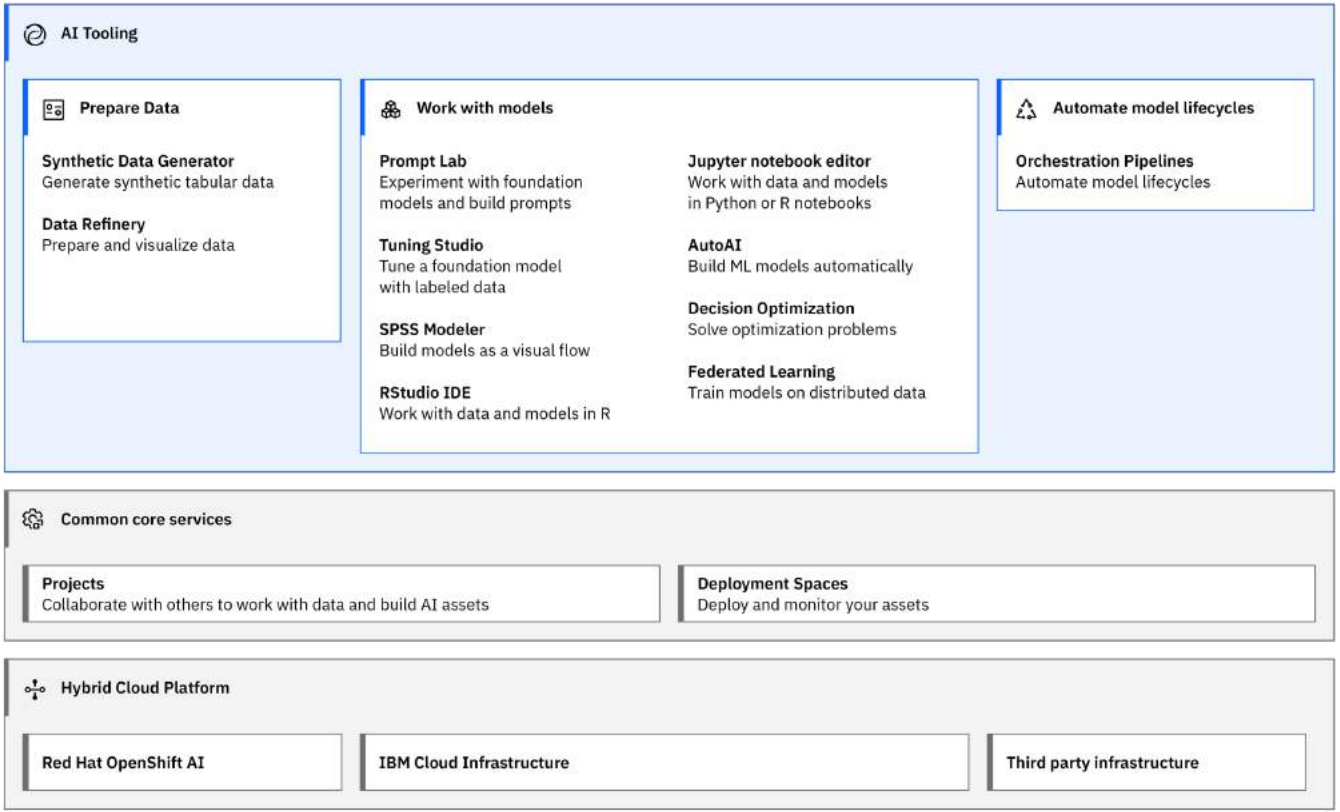
- Model performance
- Research
- Ethics
- Legal and data



Workflow

1. Review technical papers
2. Model Information
3. Performance Benchmark
4. Internal IBM use
5. Commercial Applicability
6. Licensing
7. Reputation
8. Use Case Alignment
9. Training Data
10. Infrastructure

IBM watsonx.ai architecture



Common core services

- Collaborative projects
- Deployment spaces
- Jobs
- Notifications
- Common connectivity
- Access and Authentication
- Resource management
- Central asset management system

The most common generative AI tasks implemented today

Retrieval-augmented generation (RAG)

Based on documents or dynamic content, create a chatbot or question-answering feature.

Building a Q&A resource from a broad knowledge base, providing customer service assistance

Summarization

Transform text with domain-specific content into personalized overviews that capture key points.

Conversation summaries, insurance coverage, meeting transcripts, contract information

Content generation

Generate text content for a specific purpose.

Marketing campaigns, job descriptions, blog posts and articles, email drafting support

Named entity recognition

Identify and extract essential information from unstructured text.

Audit acceleration, SEC 10K fact extraction

Insight extraction

Analyze existing unstructured text content to surface insights in specialized domain areas.

Medical diagnosis support, user research findings

Classification

Read and classify written input with as few as zero examples.

Sorting of customer complaints, threat and vulnerability classification, sentiment analysis, customer segmentation

The platform for AI and data

watsonx

Scale and accelerate the impact of AI across your business

watsonx.ai

Build, train, validate, tune and deploy AI models

A next generation enterprise studio for AI builders to build, train, validate, tune, and deploy both traditional machine learning and new generative AI capabilities powered by foundation models. It enables you to build AI applications in a fraction of the time with a fraction of the data.

watsonx.data

Scale AI workloads, for all your data, anywhere

Fit-for-purpose data store, built on an open lakehouse architecture, supported by querying, governance and open data formats to access and share data.

watsonx.governance

Accelerate responsible, transparent and explainable AI workflows

End-to-end toolkit for AI governance across the entire model lifecycle to accelerate responsible, transparent, and explainable AI workflows

AI needs governance

The process of directing, monitoring and managing the AI activities of an organization

watsonx.governance

Accelerate responsible, transparent, and explainable AI workflows

One unified, integrated AI governance platform to govern generative AI and predictive machine learning (ML)

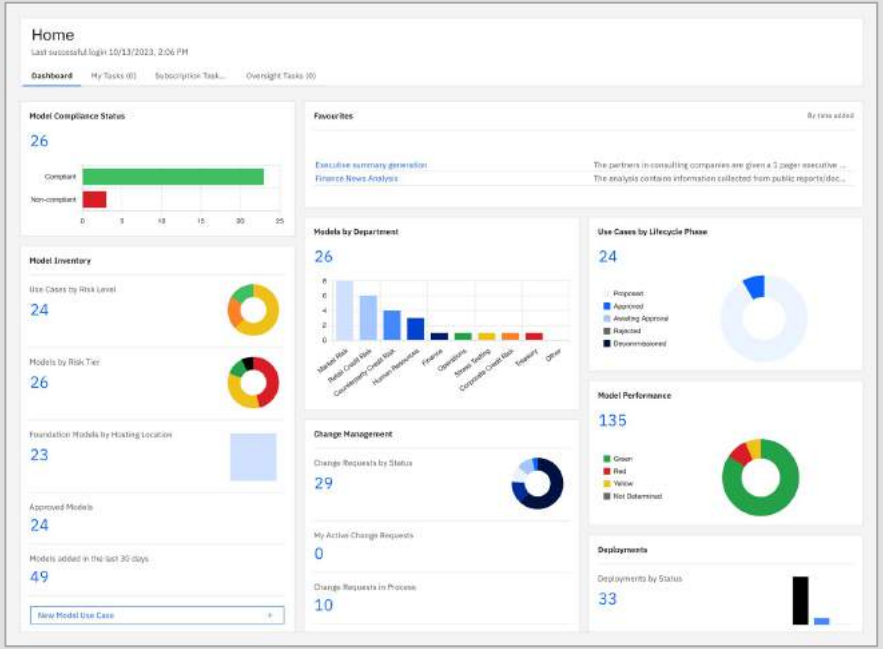
Compliance
Manage AI to meet upcoming safety and transparency regulations and policies worldwide—a “nutrition label” for AI

Risk
Proactively detect and mitigate risk, monitoring for fairness, bias, drift, and new LLM metrics

Lifecycle Governance
Manage, monitor and govern AI models from IBM, open-source communities and other model providers

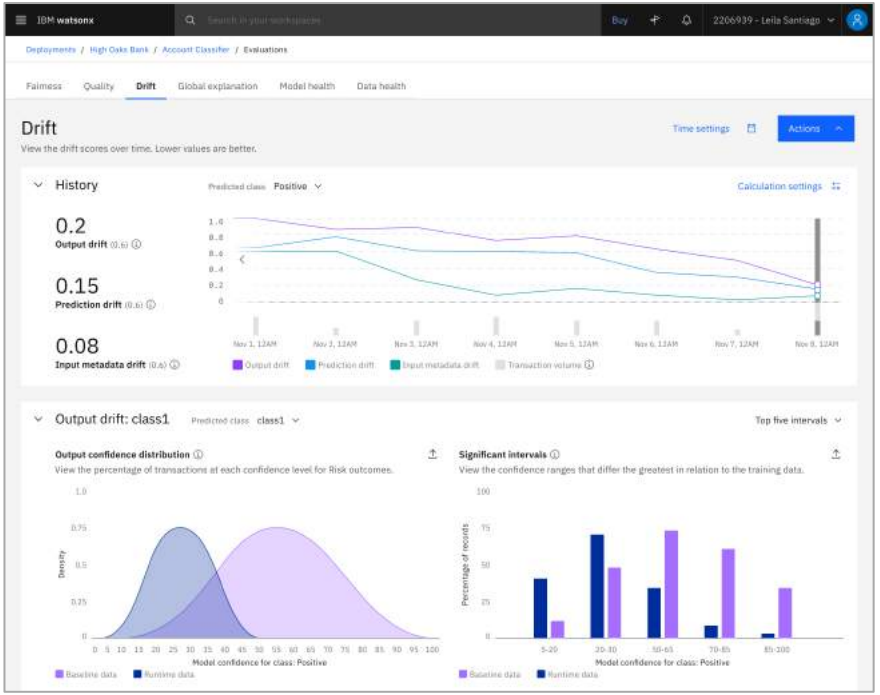
Compliance:
satisfy AI regulations

- Translate external AI regulations into **enforceable policies** for automated enforcement.
- Provide core services to help adhere to external AI regulations for audit and compliance
- Use **factsheets** for transparent model processes



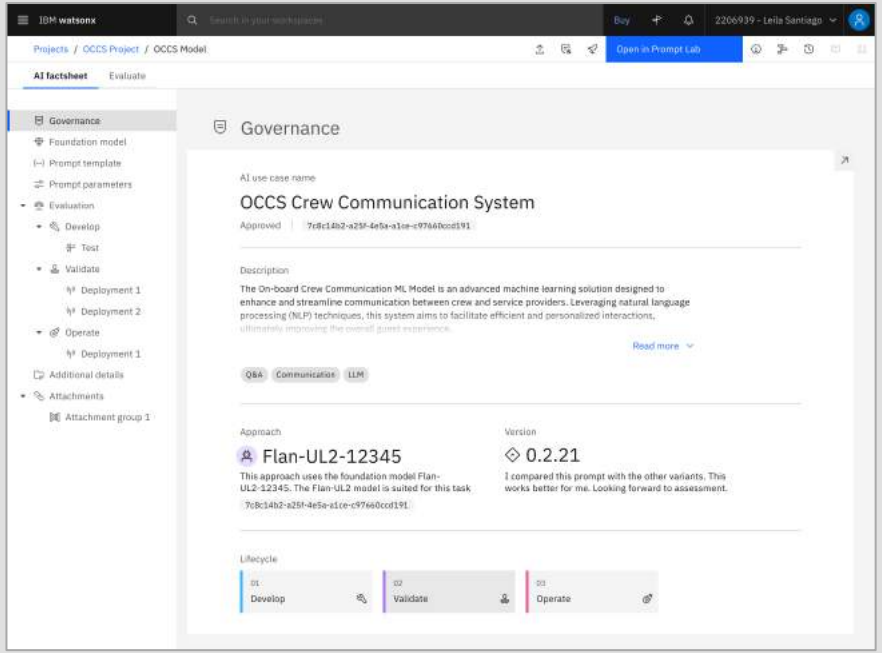
Trusted:
manage risk and protect reputation

- Preset **thresholds** for **alerts** when key metrics are breached
- Identify, manage and report on **risk and compliance** at scale
- Provide **explainable model** results in support of audits and to avoid fines



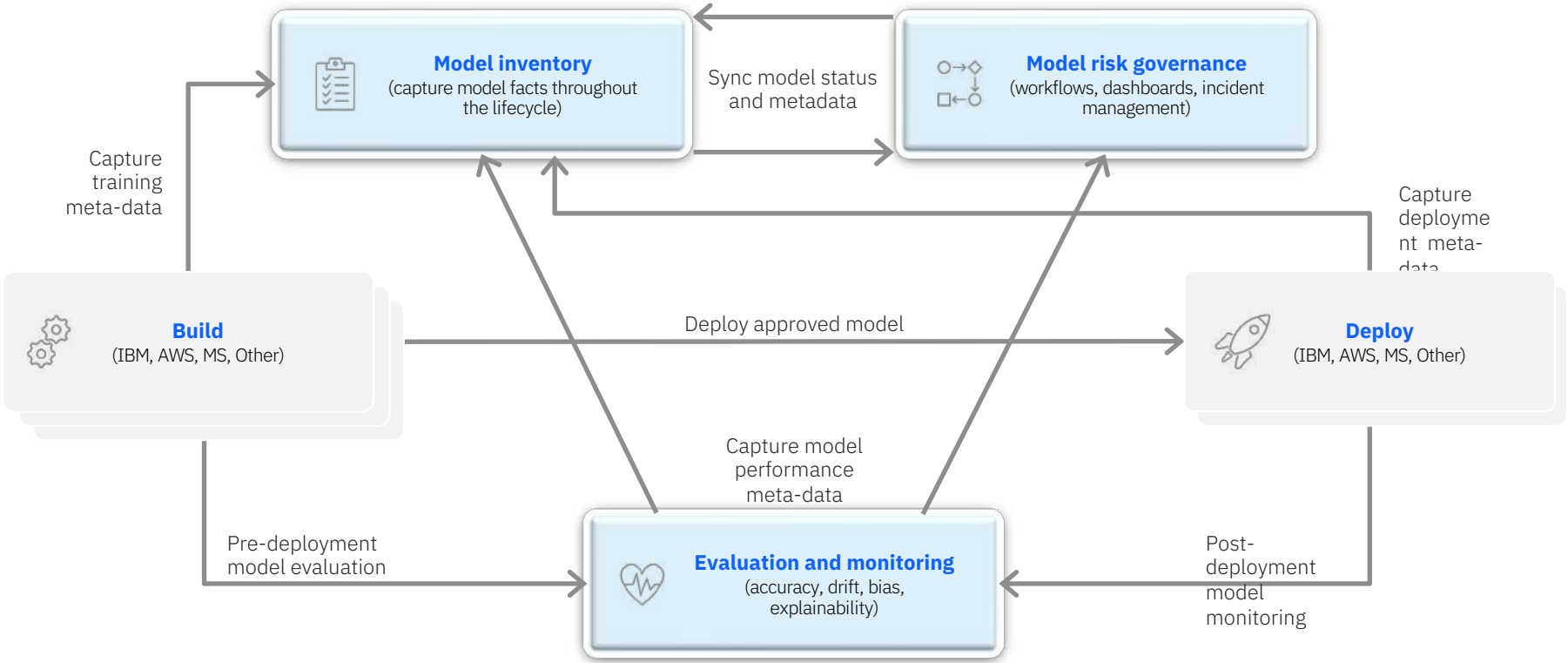
Lifecycle governance: operationalize AI with confidence

- Monitor, catalog, and govern models across the AI lifecycle
- Automate the **capture of model metadata** for to facilitate management and compliance
- **Oversee model performance** across the entire organization with dynamic dashboards and dimensional reporting



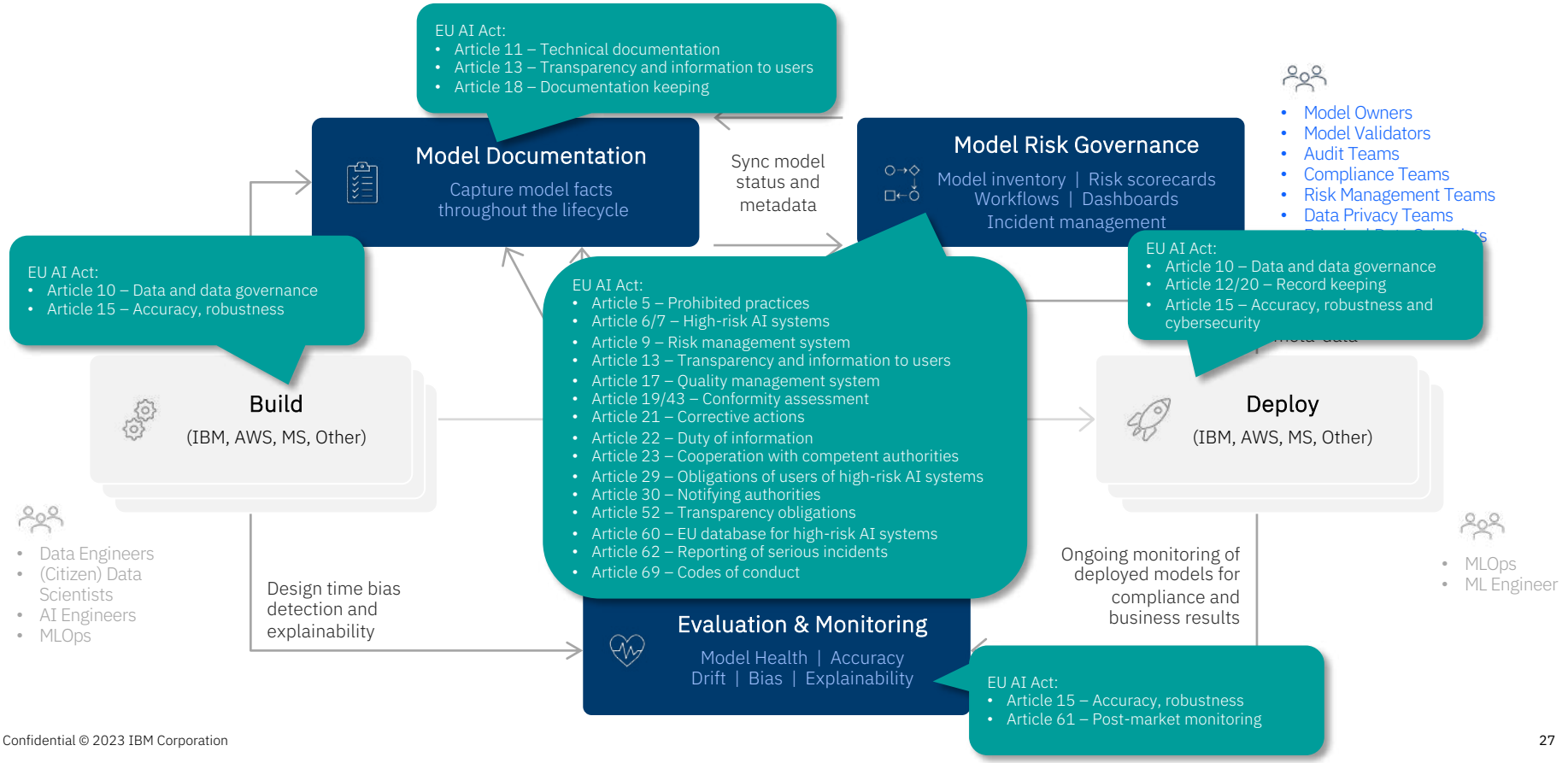
watsonx.governance

Govern across the AI lifecycle



watsonx.governance

In compliance with the EU AI Act©



The platform for AI and data

watsonx

Scale and accelerate the impact of AI across your business

watsonx.ai

Build, train, validate, tune and deploy AI models

A next generation enterprise studio for AI builders to build, train, validate, tune, and deploy both traditional machine learning and new generative AI capabilities powered by foundation models. It enables you to build AI applications in a fraction of the time with a fraction of the data.

watsonx.data

Scale AI workloads, for all your data, anywhere

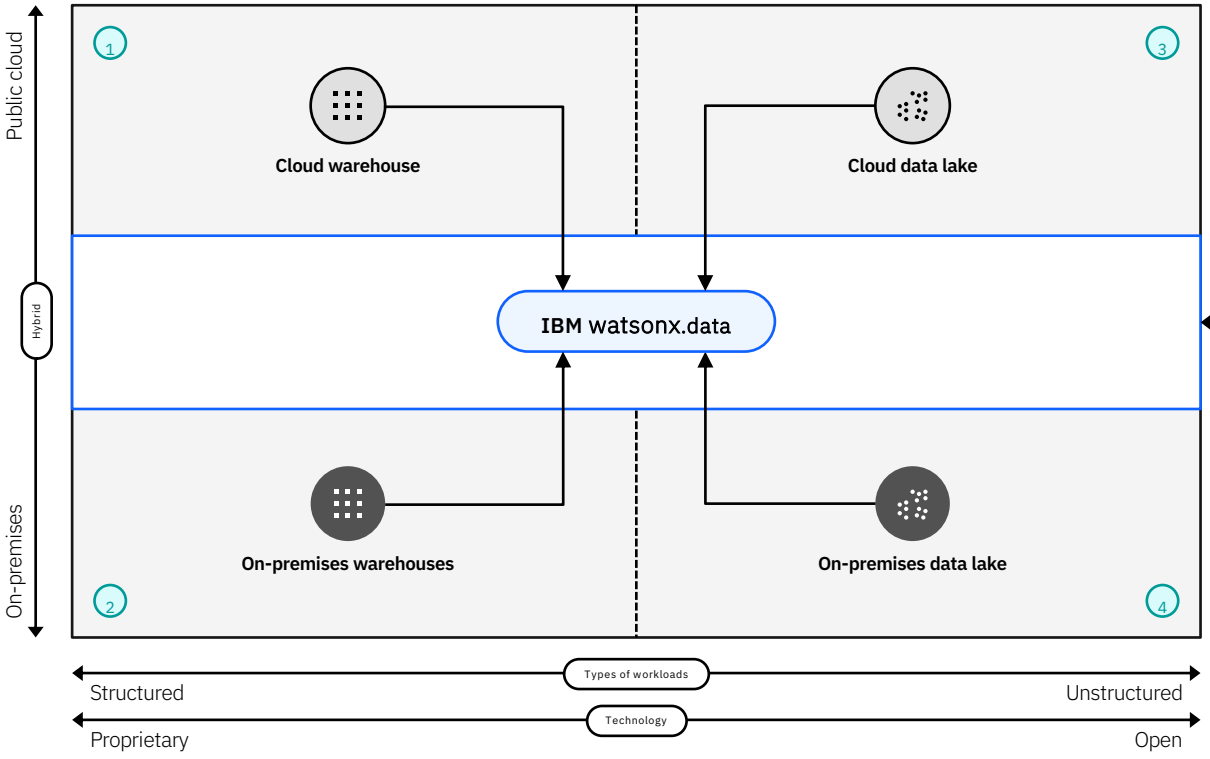
Fit-for-purpose data store, built on an open lakehouse architecture, supported by querying, governance and open data formats to access and share data.

watsonx.governance

Accelerate responsible, transparent and explainable AI workflows

End-to-end toolkit for AI governance across the entire model lifecycle to accelerate responsible, transparent, and explainable AI workflows

Access all your data, quickly and optimize your data architecture with multi-engine support and hybrid deployment of analytics and AI workloads



1
Optimize costly cloud warehouses
Make the most of fit-for-purpose query engines and compute resources

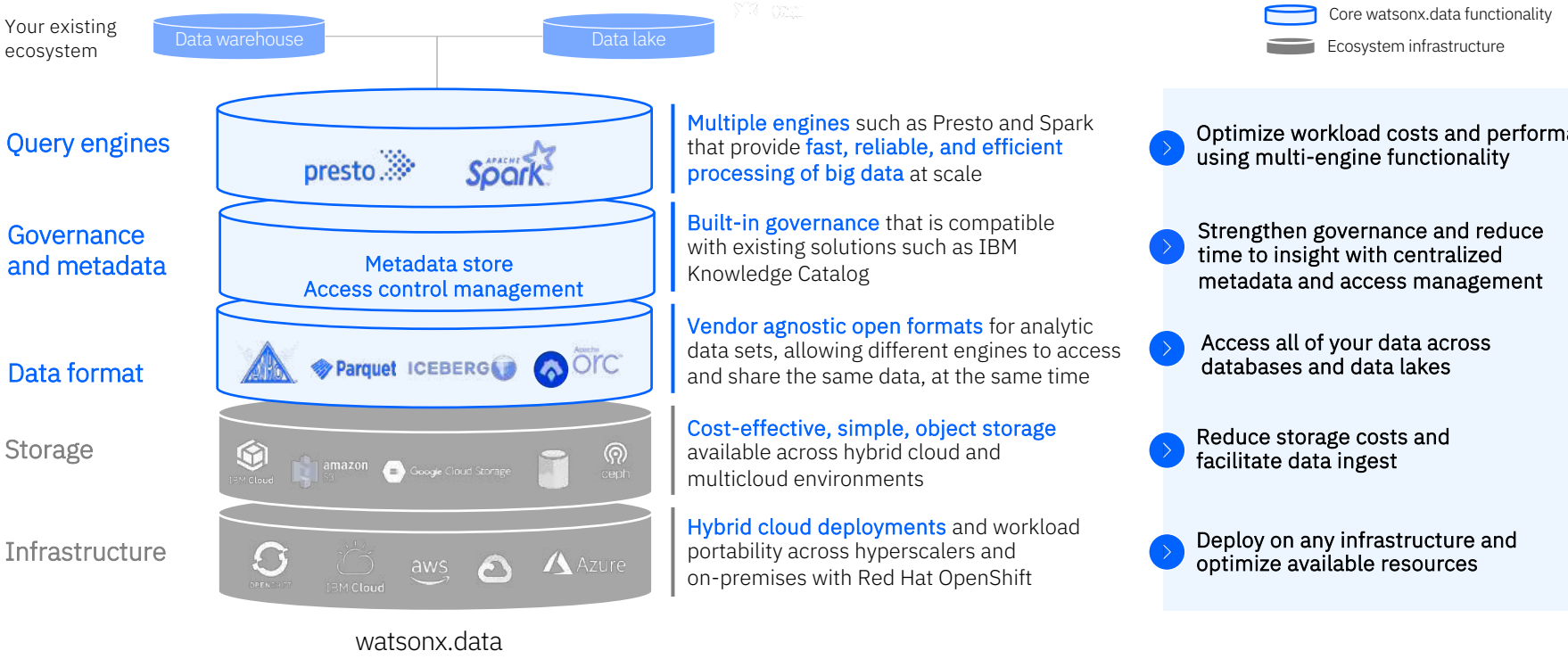
2
Optimize & access on-premises warehouses
Use low-cost object storage and fit-for-purpose engines

3 4
Modernize data lakes
Run existing reporting and enable new AI workloads without the cost and complexity of Hadoop

1 2 3 4
Deploy across hybrid cloud and multicloud
Seamlessly deploy to both the public cloud and to your existing on-premises investment

Optional

Overview of the key components of IBM watsonx.data: multiple query engines, open table formats, and built-in enterprise governance



IBM POV: Four core principles to tailor generative AI for enterprise

Open

→ Based on the best AI and cloud technologies available.

→ Giving access to the innovation of the open community and multiple models.

Targeted

→ Designed for targeted business use cases, that unlock new value.

→ Including curated models that can be tuned to proprietary data and company guidelines.

Trusted

→ Offering security and data protection.

→ Built with governance, transparency, and ethics that support increasing regulatory compliance demands.

Empowering

→ On a platform to bring your own data and AI models that you tune, train, deploy, and govern.

→ Running anywhere, designed for scale and widespread adoption to truly create enterprise value.

AI Alliance

Total annual R&D funding represented

>$80B

Students supported by these academic institutions

>400,000

Total staff members

>1,000,000

The AI Alliance

IBM has long recognized that when it comes to the future of AI, an open and transparent approach is the strongest path forward.

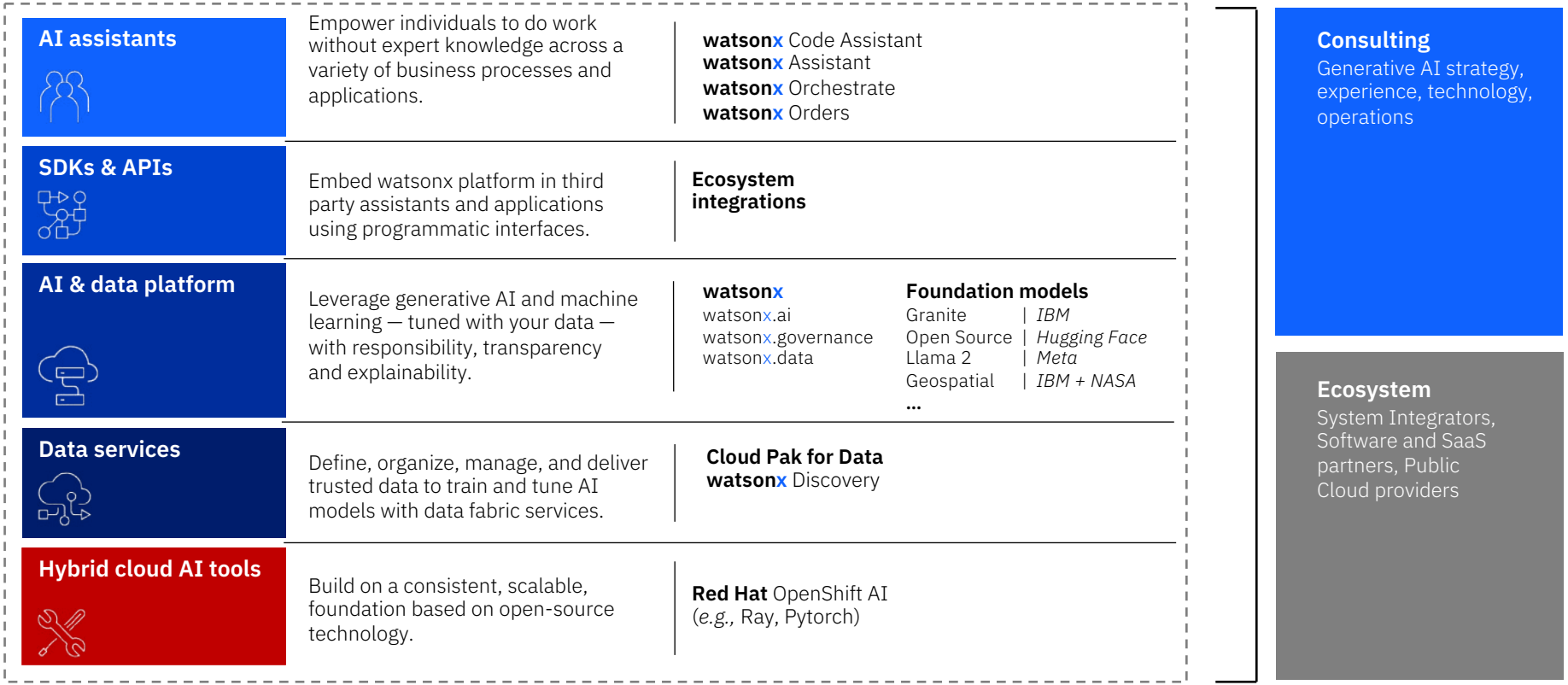
AI Alliance, launched by **IBM and Meta** with more than **50 leading industry**, academic, research and government organizations globally who will work together on creating actionable plans that advance responsible and inclusive AI that's rooted in **open innovation.**

Brings together a critical mass of compute, data, tools, and talent to **build** and support open technologies across software, models, and tools

Enable developers and scientists to understand, experiment, and adopt open technologies

Advocate for open innovation with organizational and societal leaders, policy makers and regulatory bodies, and the public

IBM’s generative AI technology and expertise



Analysts agree, IBM is a leader in the AI and MLOps market

IDC Marketscape: Leader in Worldwide Machine Learning Operations Platforms 2022 Vendor Assessment

A Leader in the 2023 Gartner® Magic Quadrant™ for Cloud AI Developer Services

(https://www.ibm.com/account/reg/us-en/signup?formid=urx-52236)

Forrester Wave: Multimodal Predictive Analytics and Machine Learning