NAND RESEARCH

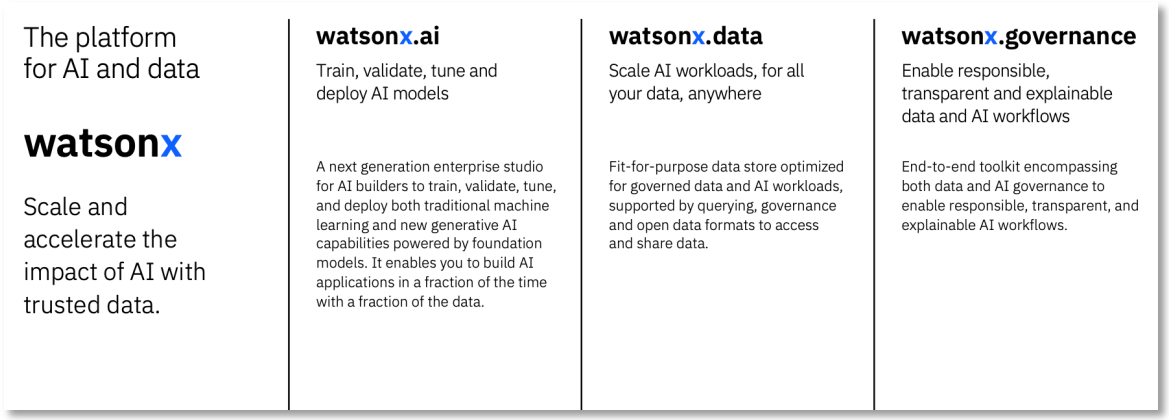
RESEARCH NOTE

INSIDE IBM WATSONX

STEVE MCDOWELL, CHIEF ANALYST
SEPTEMBER 11, 2023

BACKGROUND

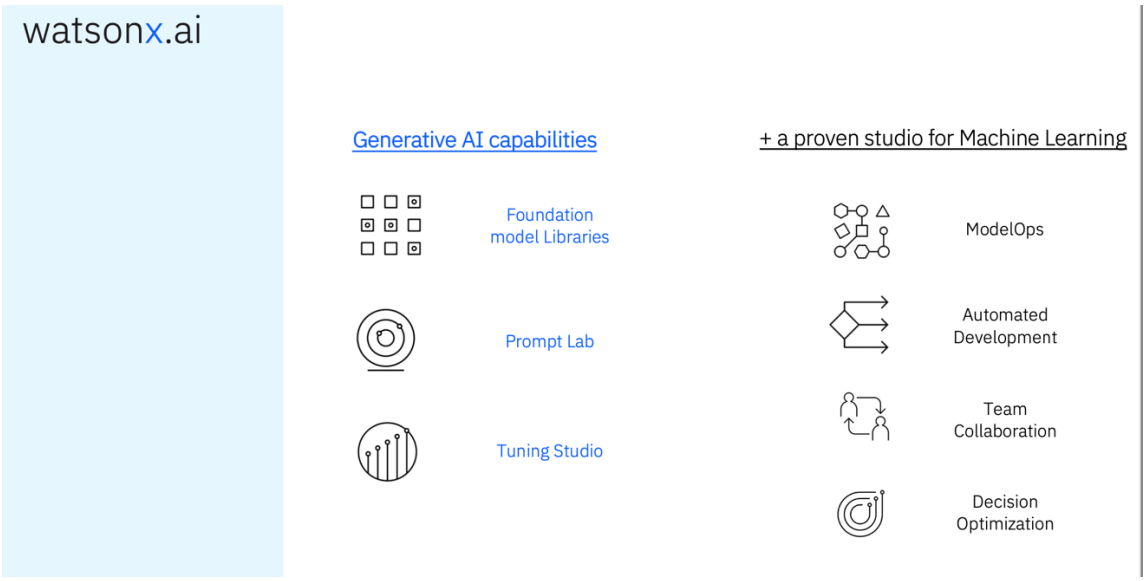
IBM recently unveiled its updated AI portfolio, “watsonx” The new set of offerings provides an enterprise-ready AI and data platform consisting of three intertwined solution stacks:



Let's look at each element within the watsonx portfolio.

WATSONX.AI

IBM watsonx.ai allows AI developers to harness models offered by IBM and the Hugging Face community to tackle a broad spectrum of AI development tasks. These models come pre-trained, geared to handle various Natural Language Processing (NLP) tasks, encompassing question answering, content generation, summarization, text classification, and data extraction.



Watsonx.ai includes the following capabilities:

- **Retrieval-Augmented Generation (RAG)**: Based on customer-specific content, RAG enables the development of context-aware chatbots and question-answering features.
- **Summarization**: Transform text with domain-specific content into personalized responses.
- **Content Generation**: Text generation for a specific purpose, such as marketing and other business content.
- **Named Entity Recognition**: Identifies and extracts essential information from unstructured text.
- **Insight Extraction**: Analyzes existing unstructured text content to generate insights in specialized domain areas.
- **Classification**: Reads and classifies written input without requiring examples.

Upcoming releases will expand the array of IBM-trained proprietary foundation models, facilitating efficient specialization in specific domains and tasks. IBM's watsonx also offers AI models from IBM and the Hugging Face community for various Natural Language Processing tasks.

Last month, IBM expanded watsonx.ai, announcing support for Meta's Llama 2-chat 70 billion parameter model in watsonx.ai studio. This collaboration builds on their joint work on open AI innovation, including projects like PyTorch and Presto.

Adding Llama 2 to watsonx.ai is a significant step in IBM's generative AI roadmap, with plans for more AI models and features. IBM prioritizes trust and security, allowing users to employ AI guardrails to remove harmful language.

WATSONX.DATA

IBM's watsonx.data is crafted to assist clients in overcoming challenges related to data volume, complexity, cost, and governance as they scale their AI workloads. The platform enables users to seamlessly access their data, whether stored in the cloud or on-premises, through a single entry point. This approach dramatically simplifies data access for non-technical users while ensuring security and compliance.

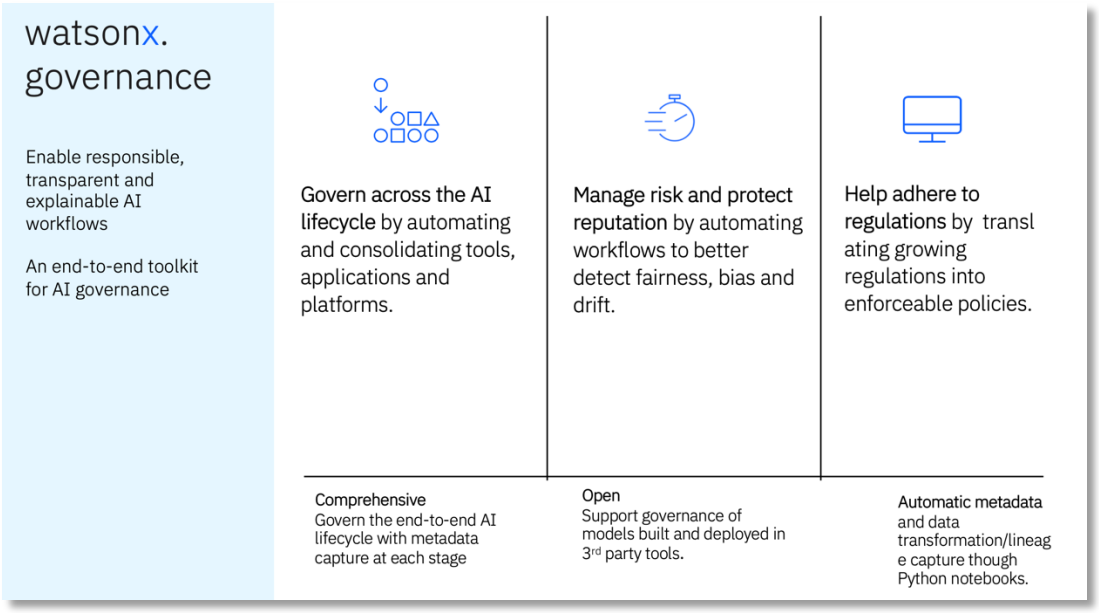
Furthermore, the significance of watsonx.data extends beyond data scientists and engineers. It empowers non-technical users by granting them self-service access to enterprise-grade, trustworthy data within a unified collaborative platform. Simultaneously, it reinforces security and compliance protocols through centralized governance and local automated policy enforcement.

Soon, watsonx.data will harness the capabilities of watsonx.ai foundation models to simplify and expedite user interactions with data. This will allow users to utilize natural language for tasks such as discovering, enhancing, refining, and visualizing data and metadata, creating a more conversational and user-friendly experience.

WATSONX.GOVERNANCE

As AI becomes increasingly integrated into everyday workflows, the necessity for proactive governance to ensure responsible and ethical decision-making within the organization grows.

Watsonx.governance leverages IBM's robust AI governance capabilities to assist organizations in implementing comprehensive end-to-end lifecycle governance, mitigating risks, and effectively managing compliance with the evolving landscape of AI and industry regulations.



Watsonx.governance empowers organizations to lead, oversee, and oversee their company's AI initiatives. The tool utilizes software automation to enhance your capacity to mitigate risks, handle regulatory mandates, and address ethical considerations, all without the need for costly transitions in your data science platform, even for models created using third-party tools.

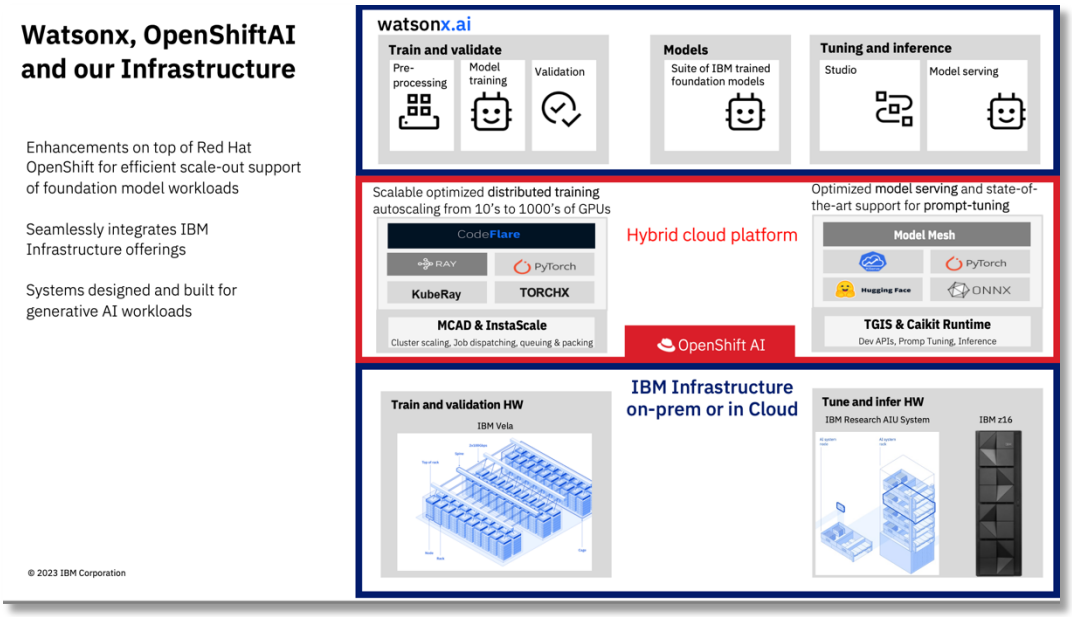
IBM's watsonx.governance is in tech preview today and is expected to be generally available later this year.

WATSONX INFRASTRUCTURE

The IT infrastructure of nearly every enterprise today is a hybrid-cloud infrastructure. IBM recognizes this, making the capabilities of watsonx available on-prem and in the cloud.

For example, IBM and Amazon Web Services (AWS) worked together to enable watsonx.data on AWS infrastructure. Enterprises can expedite cloud-based data modernization efforts by taking advantage of the openness, performance, and governance of IBM watsonx.data, while leveraging the scalability, agility, and cost-effectiveness of the AWS cloud infrastructure.

IBM delivers this flexibility by building watsonx atop an enhanced version of its Red Hat OpenShift technology, updated to ensure more efficient scale-out support for foundation model workloads. OpenShift also allows watsonx to seamlessly integrate with a broad range of IBM infrastructure offerings.



While cloud brings obvious benefits to IT organizations, the true benefit of AI in the enterprise blooms when using servers, storage, and accelerators designed for the unique requirements of AI workloads. IBM brings its considerable expertise in high-performance compute to the table with multiple AI-targeted infrastructure elements, including its Z-series mainframes and its IBM "Vela" AI supercomputer.

IBM is also looking to update its own IBM cloud offerings to support the demands of AI workloads better. The company tells us that in the coming months, it anticipates rolling out a comprehensive, high-performance, and adaptable AI-optimized infrastructure as a service through IBM Cloud, catering to the training and deployment needs of foundation models.

ANALYSIS

AI is no longer the exclusive domain of researchers and engineers. The technology is rapidly becoming a critical enabler for how business will be conducted moving forward. AI will become a competitive differentiator for those enterprises that quickly embrace the new technology. At the same time, AI doesn't look like anything we usually see in an enterprise data center.

Whether you're experimenting with the technology or deploying to production, AI remains a complex operation. Only with solutions such as IBM's watsonx can enterprises remove the pain and see accelerated time-to-value for their efforts.

Watsonx includes nearly everything you want in an enterprise-class AI solution. I've only just touched on its capabilities here. Watsonx also consists of various AI-assistants, including assistants for code generation and business task management,

along with development tools for prompt engineering and model tuning. The list goes on and on. No matter where you are in your AI journey, there's not a more comprehensive AI solution than watsonx.

There is also no company better positioned than IBM to bring AI out of the research labs and into the enterprise. After all, IBM has helped enterprises adapt and benefit from new technologies since its inception, AI included. It’s simply what IBM does.