



Unlocking AI with IBM Power

Barry Brown

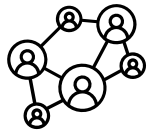
Technical Director, IBM Cloud & Power EMEA

March 11-12, 2026

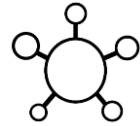
Budapest

1B

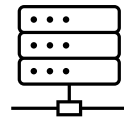
*new applications are expected by 2028¹
...and enterprises are drowning in complexity*



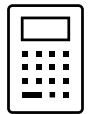
Silos



Data



Outages



*Cost
Optimizations*



*Skills
Shortages*

*...which makes keeping critical
systems continually running a
near impossible task...*

*...and these challenges will
only intensify as enterprises
strive to remain competitive in
the AI era*

The time for scaling AI is now
but enterprises face **obstacles**.



TOP 3 OBSTACLES



Data & integration complexity

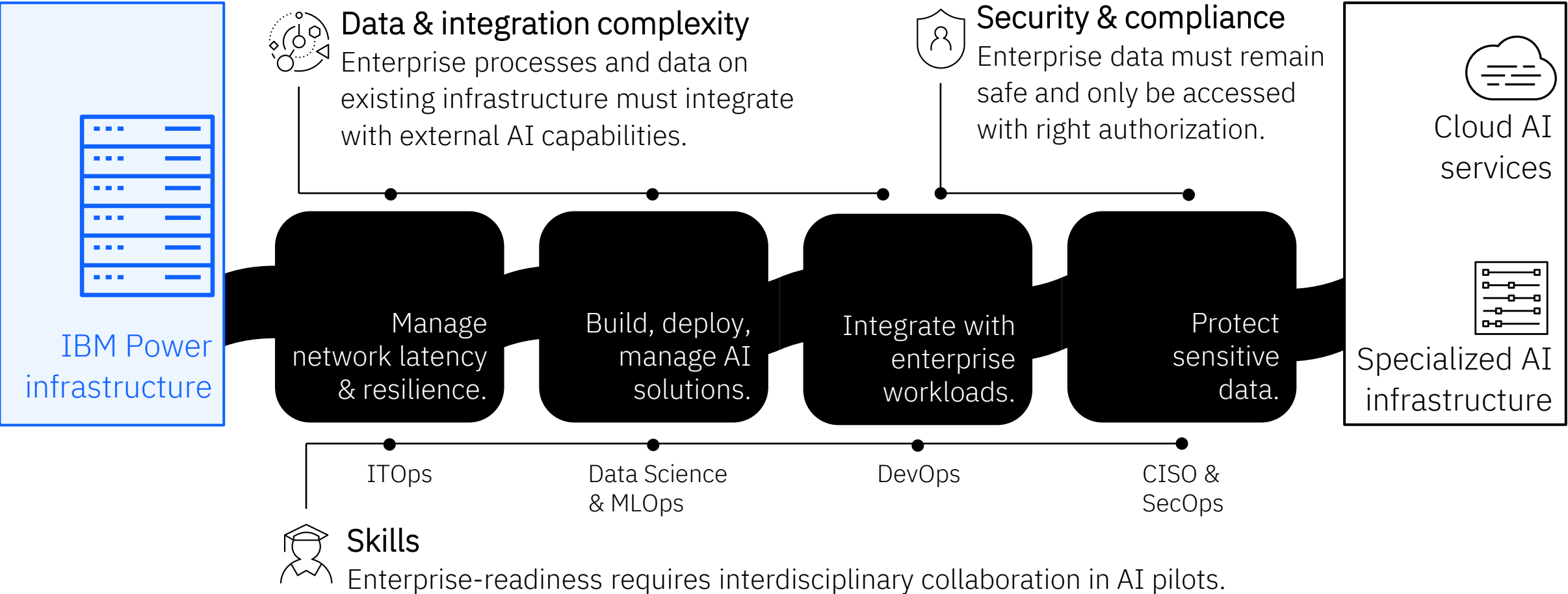


Skills



Security & compliance

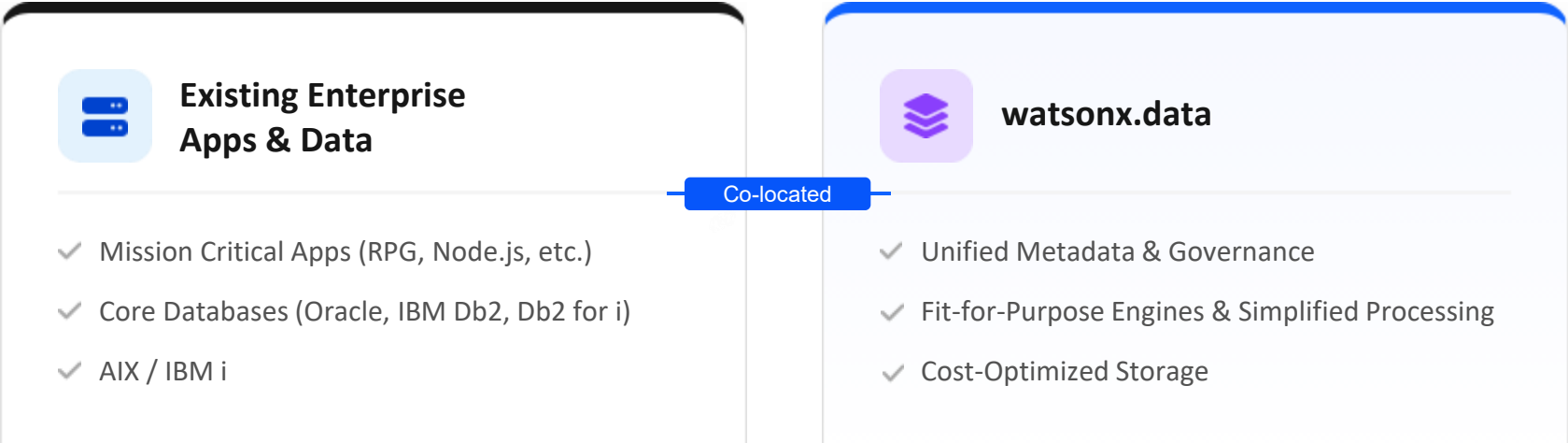
Core enterprise workflows get the largest ROI from AI, if obstacles are addressed.



One Platform. All Insights.

Securely co-located on the world's most reliable infrastructure.

Eliminate the risk and cost of moving data by bringing AI models directly to your enterprise core. Accelerate insights with a unified architecture that simplifies governance and maximizes ROI.



IBM Power
The Enterprise-Grade Foundation

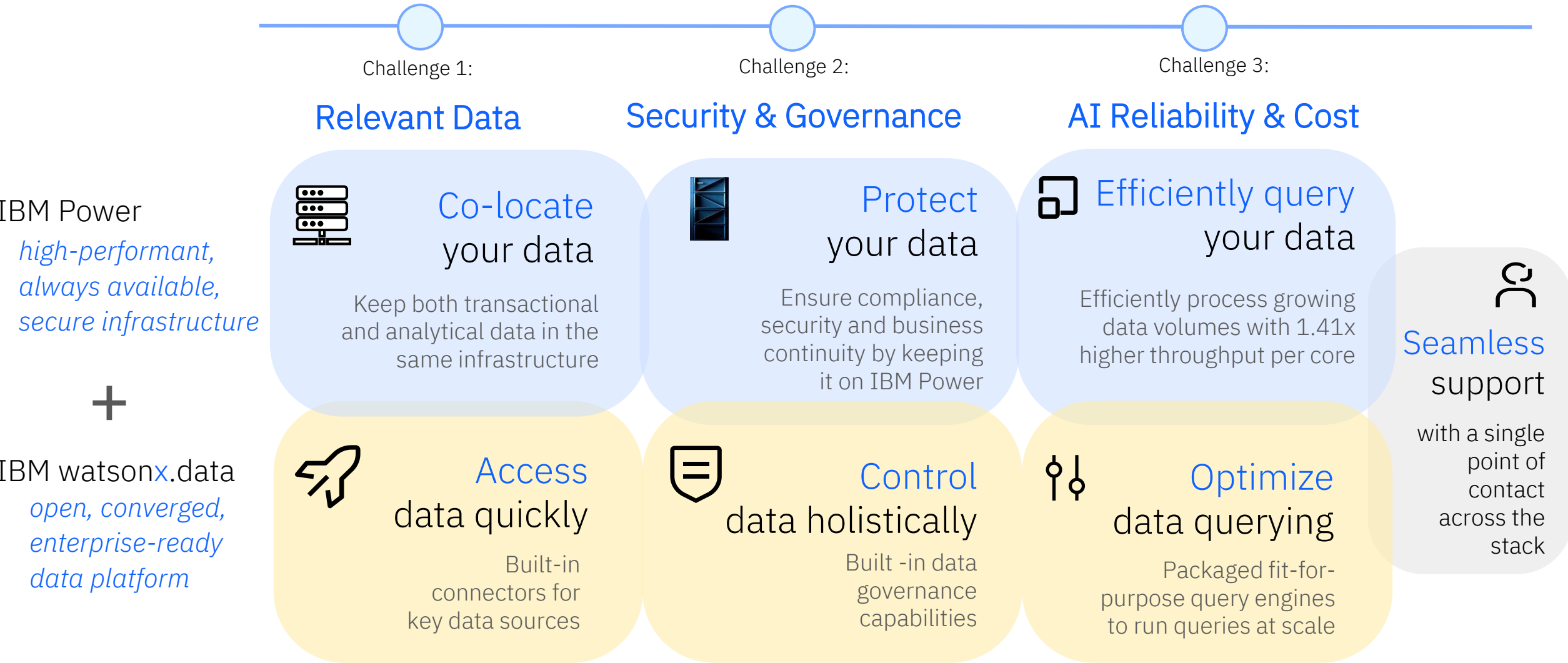
 **High Performance**
INFRASTRUCTURE

 **Scalable**
ARCHITECTURE

 **Availability**
ALWAYS ON

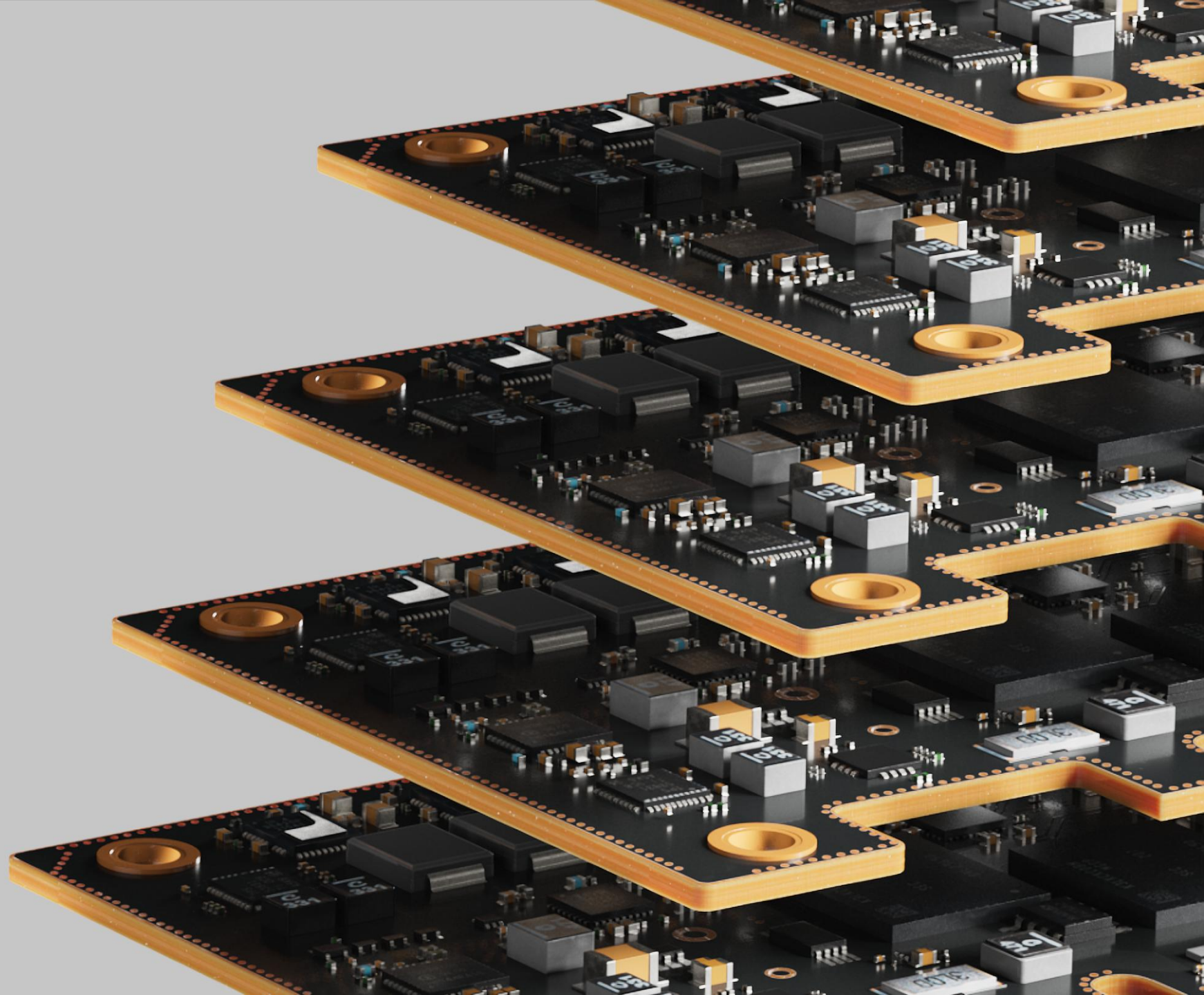
 **Secure**
BY DESIGN

Efficiently scale AI across your enterprise



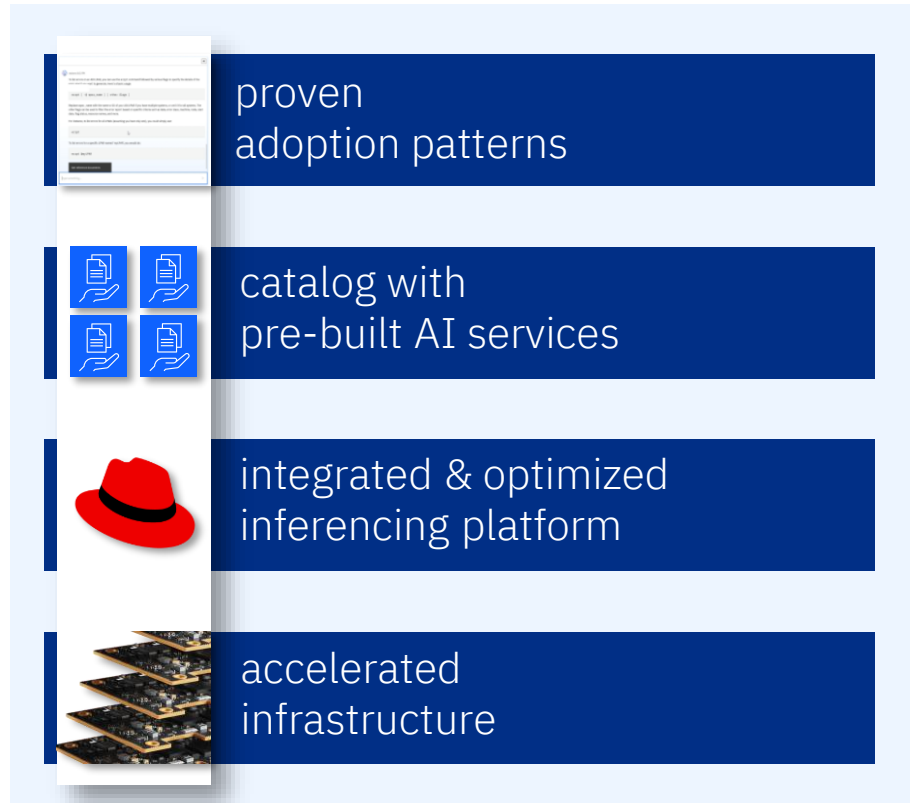
Introducing
IBM Spyre™
for Power...

Turnkey AI
for enterprise
workflows.



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.

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](#)

Real-life example:
Extract ordering
details in the order
process automatically

Hans Geis, a logistics provider
running their core ERP system
on IBM i, integrates AI for a...

*“5x increase in business
process rate with AI
integration into existing
enterprise workflows.”*

– Uwe Rempel, head of ZSI department, Geis Group¹

Disclaimer: 1: “We achieve a 5x increase in business process rate with AI integration into our existing processes on IBM Power leveraging on-chip accelerator (MMA and SIMD) technology.” - Uwe Rempel, head of ZSI department, Geis Group. Measured in staging system of Hans Geis. Without AI it takes a senior employee 2 minutes / mail = 120 seconds / mail; with AI, they can completely automate 80% with 0 minutes processing time: $(80\% \cdot 0 \text{ minutes} + 20\% \cdot 2 \text{ minutes}) / \text{mail} = 24 \text{ seconds} / \text{mail}$; leading to a factor of 5x: $120/24 = 5$.





Thank You