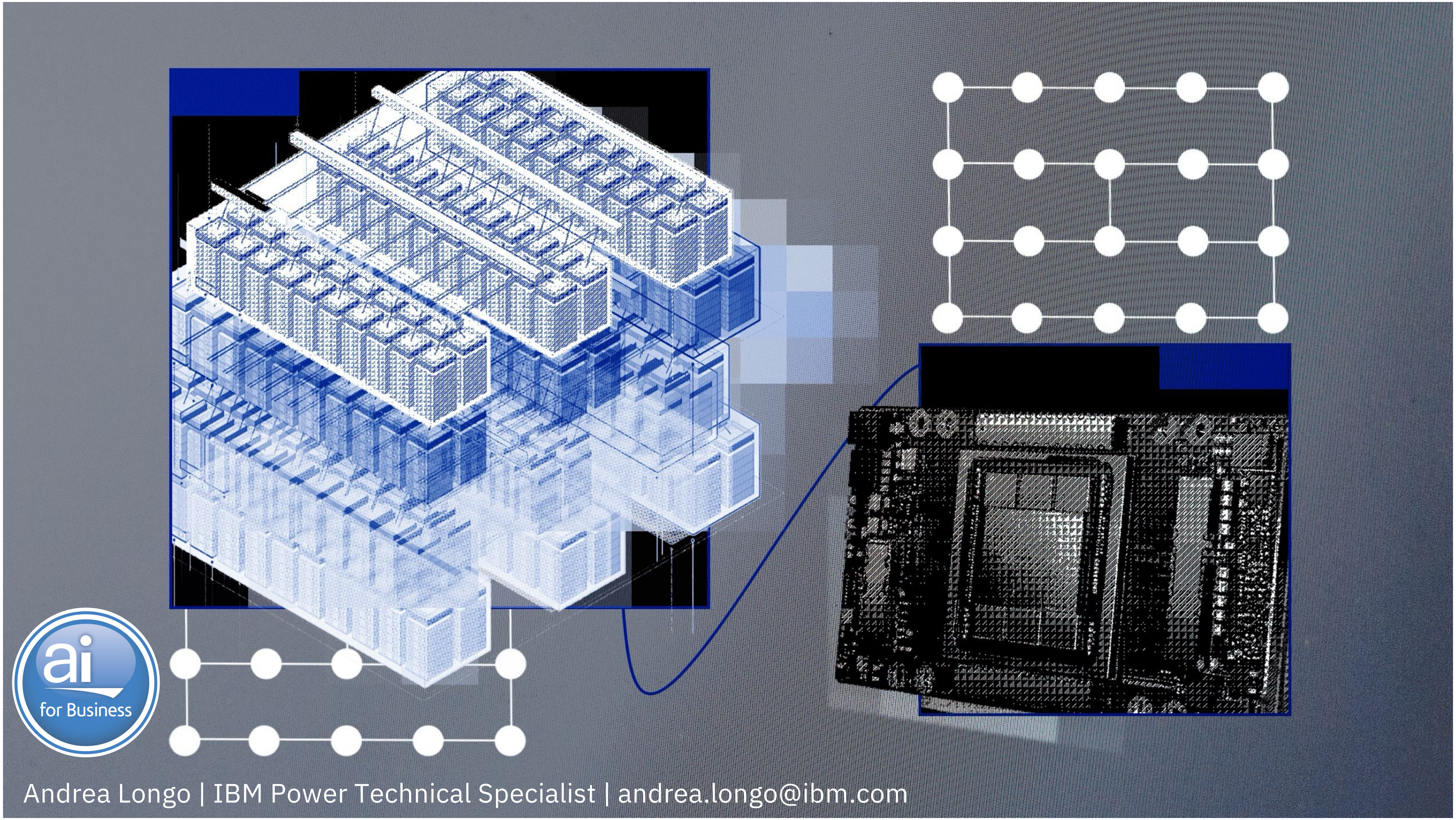
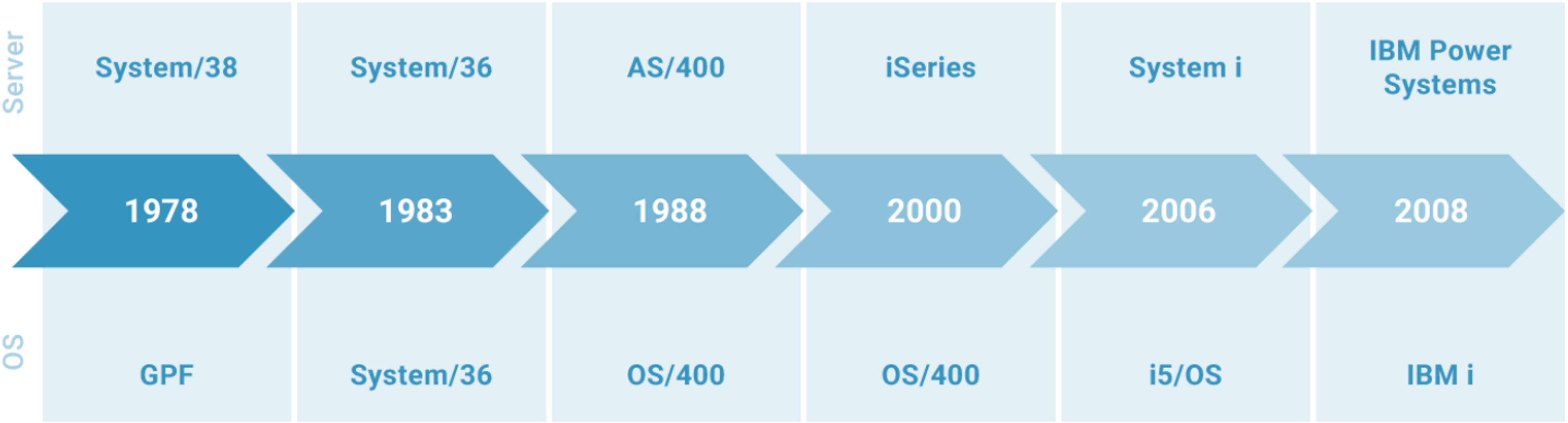


Andrea Longo | IBM Power Technical Specialist | andrea.longo@ibm.com



Server	System/38	System/36	AS/400	iSeries	System i	IBM Power Systems
	1978	1983	1988	2000	2006	2008
OS	GPF	System/36	OS/400	OS/400	i5/OS	IBM i

Great opportunity for both masters and apprentices

Foundation Models are bringing an inflection point in AI…

…but how enterprises adopt and execute will define whether they unlock value at scale

Generative AI has immense potential to accelerate digital transformation

Scale of impact points to swift adoption over next 3 years

$3-4T forecasted economic benefits to the global economy across industries

80% productivity gains across classes of knowledge workers and creative tasks

80% of enterprises will have incorporated Gen AI into their business processes

70% of software vendors will integrate Gen AI in their enterprise applications

Source: Gartner

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 Technology trends shaping the enterprise AI market

Open innovation

Two-thirds of models released in 2023 were open[1], with enterprises adopting both open and closed models; open models offer cost-effectiveness, flexibility, security, transparency, and opportunities for innovation.

Optimized models

85% of enterprises[2] are using fit-for-purpose models instead of defaulting to larger models, as fit-for-purpose models can offer cost-effective performance with tuning supported by larger models.

AI will be hybrid and on-premises

GenAI will be run across hybrid environments, with a significant share on-premises, where increasing hardware diversity makes on-premises inference more cost-effective.

Value of enterprise data

In 2024, 47% of enterprises aim to leverage their data to enhance model efficacy[3], yet less than 1% of enterprise data is currently represented in GenAI models[4], while almost all public data has been utilized.

AI middleware and platforms

Technical and developmental challenges are among the the top reasons why 80% of enterprises still do not have GenAI in production[5], robust GenAI middleware is needed to streamline the development process

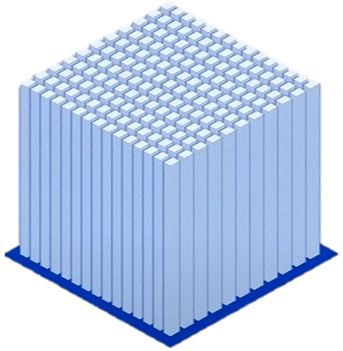
Sources: 1) Stanford Institute for Human-Centered AI (2024). *AI Index report 2024.* Stanford University. 2) Turner, M., et al. (2024, March). *Customer perspectives on AI-ready infrastructure priorities – 2024 Q1*. IDC. 3) McKinsey & Company. (2023). The state of AI. 4) IBM Research 5) Chandrasekaran, A., et al. (2024, June). GenAI trends and opportunities.

Value of enterprise data

To close the last mile of model efficacy, enterprises are looking to leverage their enterprise data

In 2024, 47% of enterprises said they are looking to do heavy model customization with enterprise data[1]

Current nearly all available *public data* is now represented in foundation models[2]



Less than 1% of all *enterprise data* is represented in foundation models[2]

52% of enterprise data is still in data centers[3]

Sources:1) McKinsey & Company. (2023). The state of AI 2)IBM Research 3) Flexera. (2024). 2024 State of the Cloud Report

Hybrid and on-premises AI

Gen AI On-premises will constitute 25% of the market, Edge 8%

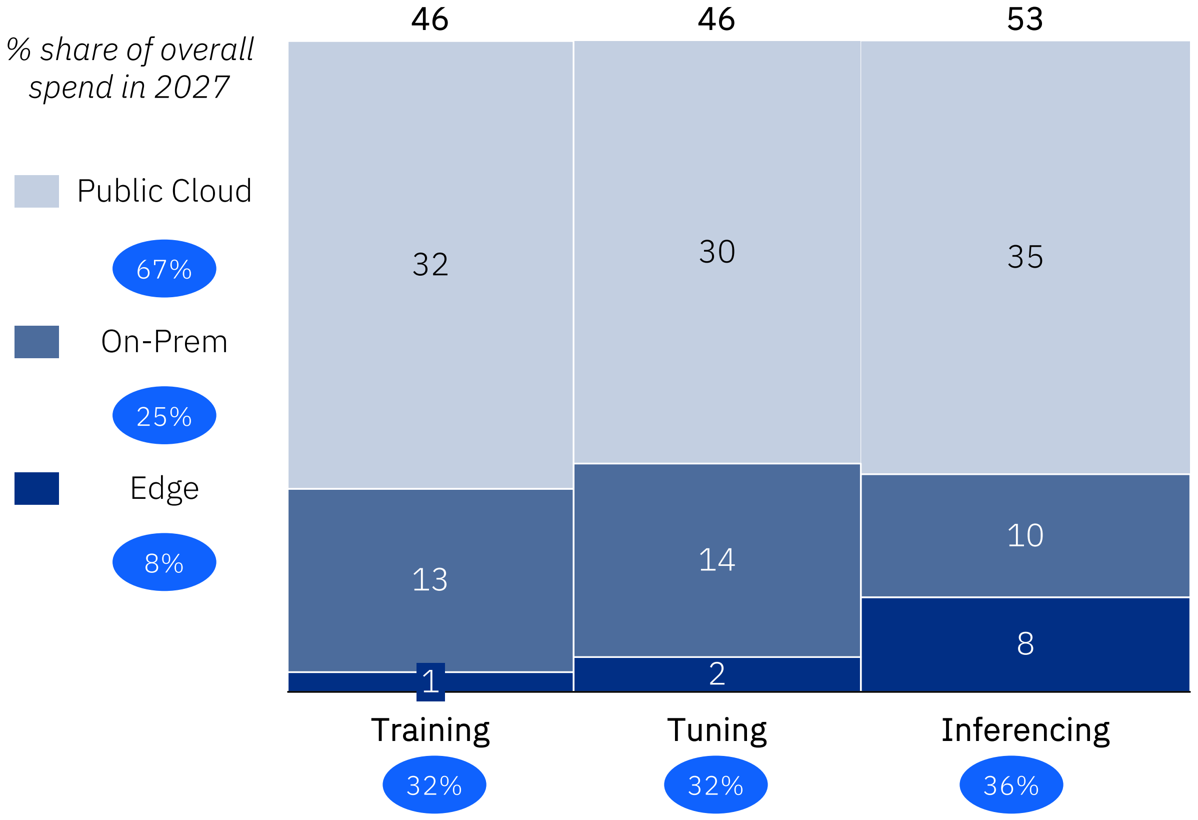
GenAI will be deployed across hybrid multi-cloud environments

48%
Of enterprises to deploy GenAI on-prem in the next year[1]

20%
Of enterprise to start edge deployments in the next year[1]

+38%
YoY growth GPU shipments to enterprises over the next three years [2]

GenAI market by deployment and split estimation by workload type by 2027 ($B)[3]



Sources: 1) Turner, M., et al. (2024, March). Customer perspectives on AI-ready infrastructure priorities – 2024 Q1. IDC. 2) Turner, N. (2023, May). Worldwide endpoint AI processor and accelerator forecast, 2023–2027. IDC.
3) Derived from Villars, R., et al. (2023, December). *Worldwide Core IT Spending for GenAI Forecast, 2023-2027*

IBM’s POV for AI-ready IT Infrastructure

Reliable performance

Scale AI inferencing for complex tasks like generative AI

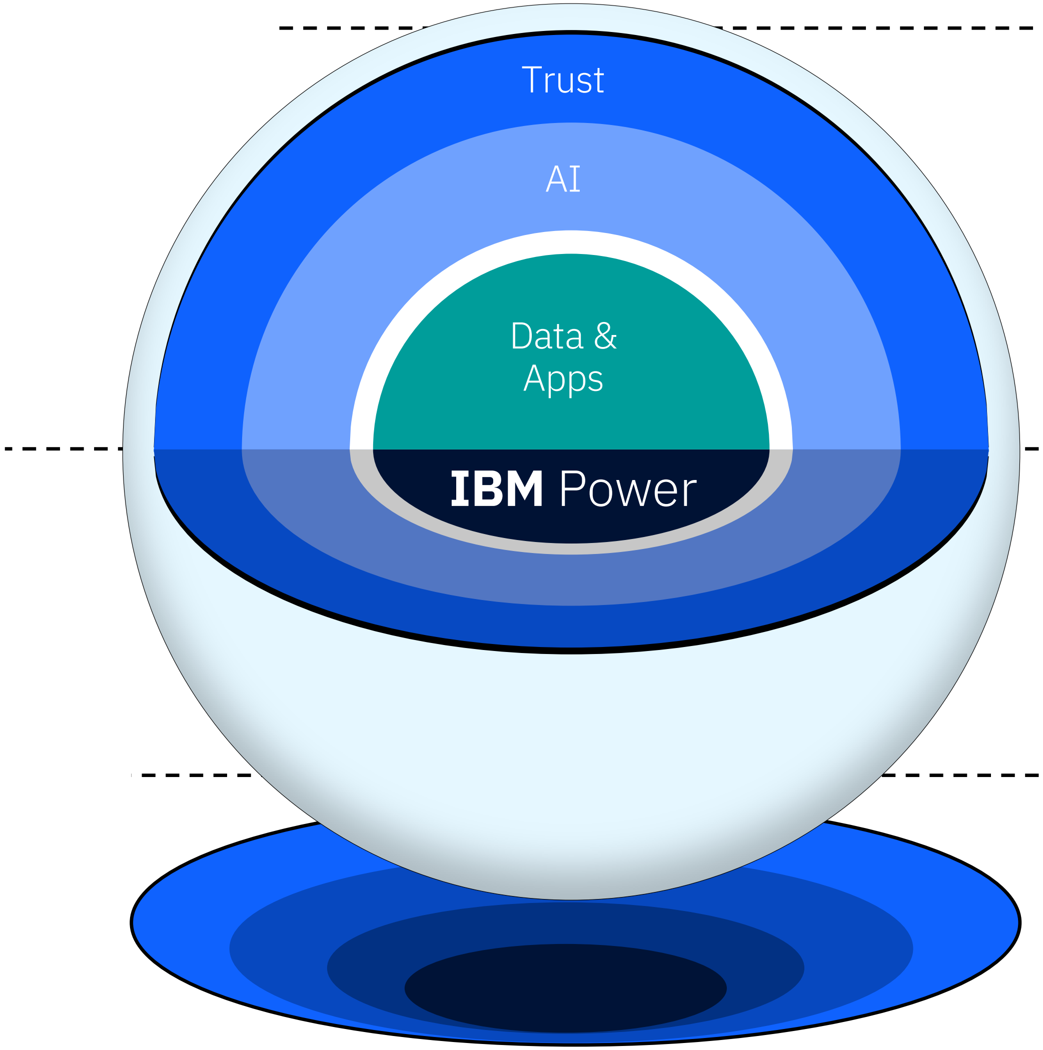
Hybrid flexibility

Create AI workflows based on where your data and applications reside

Secured insights

Offer security and data protection to promote trust and support compliance demands

AI for Business
with **IBM** Power.



AI-powered workflows.
Create new insights & value for LOBs, development, and operations in mission-critical contexts.

AI-ready infrastructure.
Accelerate data & AI workflows when and where needed – safely, reliably, efficiently, and easily.

AI-infused ecosystem.
Integrate seamlessly & flexibly with hybrid environments, data sources, and ISV solutions.

Accelerate AI Efficiently with IBM Power10 technical innovation

On-Chip AI acceleration

Per Core: 4x MMA* & 8x SIMD**

- Accelerate Matrix & Vector Math operations in the processor without GPUs
- Minimize data movement from processor to GPU and vice-versa
- Lower quantization for improved performance

Large Memory & Cache

TBs Main memory per core

- Store large AI models (like LLMs), large datasets (higher batch sizes), multiple AI models (parallel inferencing) in single memory (1TB-16TB)
- Large cache (4x) to speed up execution

Highly Parallel

4x more Threads per core

- Provides up to 8 threads per core (4x vs. Intel)
- Run parallel jobs for AI inferencing
- Improve throughput

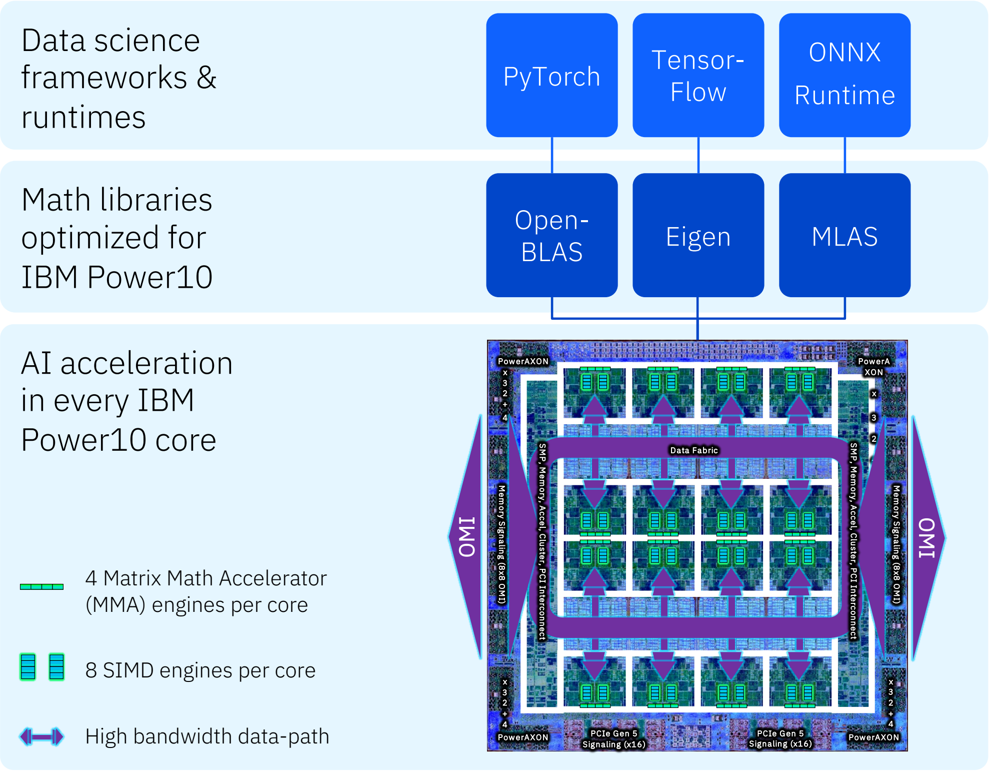
Enterprise Server Design

Lower CO2 footprint

- Same work with lower energy usage
- Consolidate more workloads due to higher guaranteed utilization

*MMA: Matrix Math Accelerator ****SIMD: Single Instruction Multiple Data

Accelerate AI efficiently.



Data science frameworks & runtimes

PyTorch

Tensor-Flow

ONNX Runtime

Math libraries optimized for IBM Power10

Open-BLAS

Eigen

MLAS

AI acceleration in every IBM Power10 core

▬ 4 Matrix Math Accelerator (MMA) engines per core

▮▮ 8 SIMD engines per core

↔ High bandwidth data-path

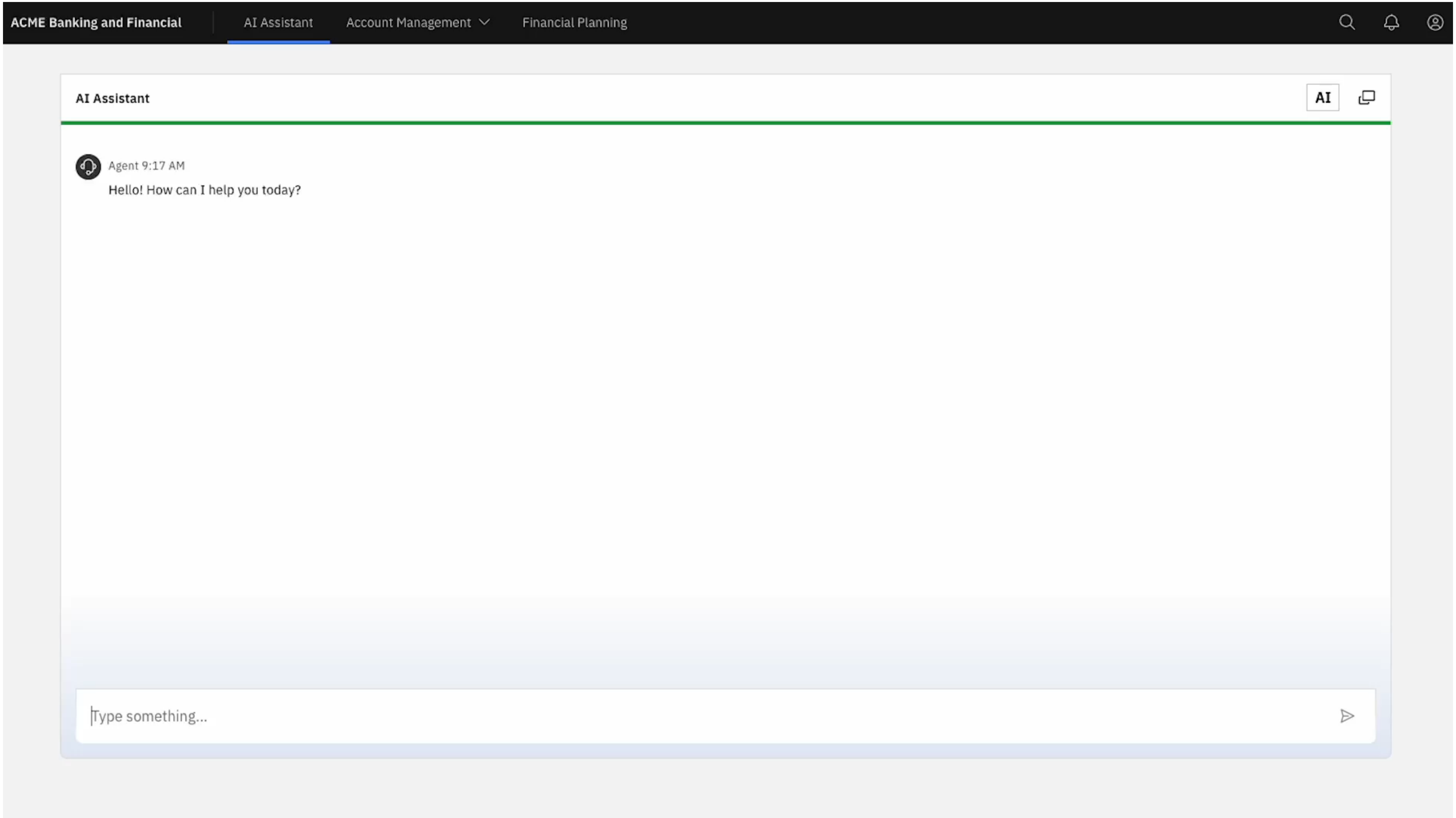
PowerAXON
PowerAXON
Data Fabric
SMP, Memory, Accel, Cluster, PCI Interconnect
SMP, Memory, Accel, Cluster, PCI Interconnect
Memory Signaling (8x8 OMI)
Memory Signaling (8x8 OMI)
OMI
OMI
PowerAXON
PCIe Gen 5 Signaling (x16)
PCIe Gen 5 Signaling (x16)
PowerAXON

Full stack optimization

Data scientists get out-of-the-box acceleration (no code changes required thanks to a full stack co-optimization), easing deployments of AI models such as large language models.

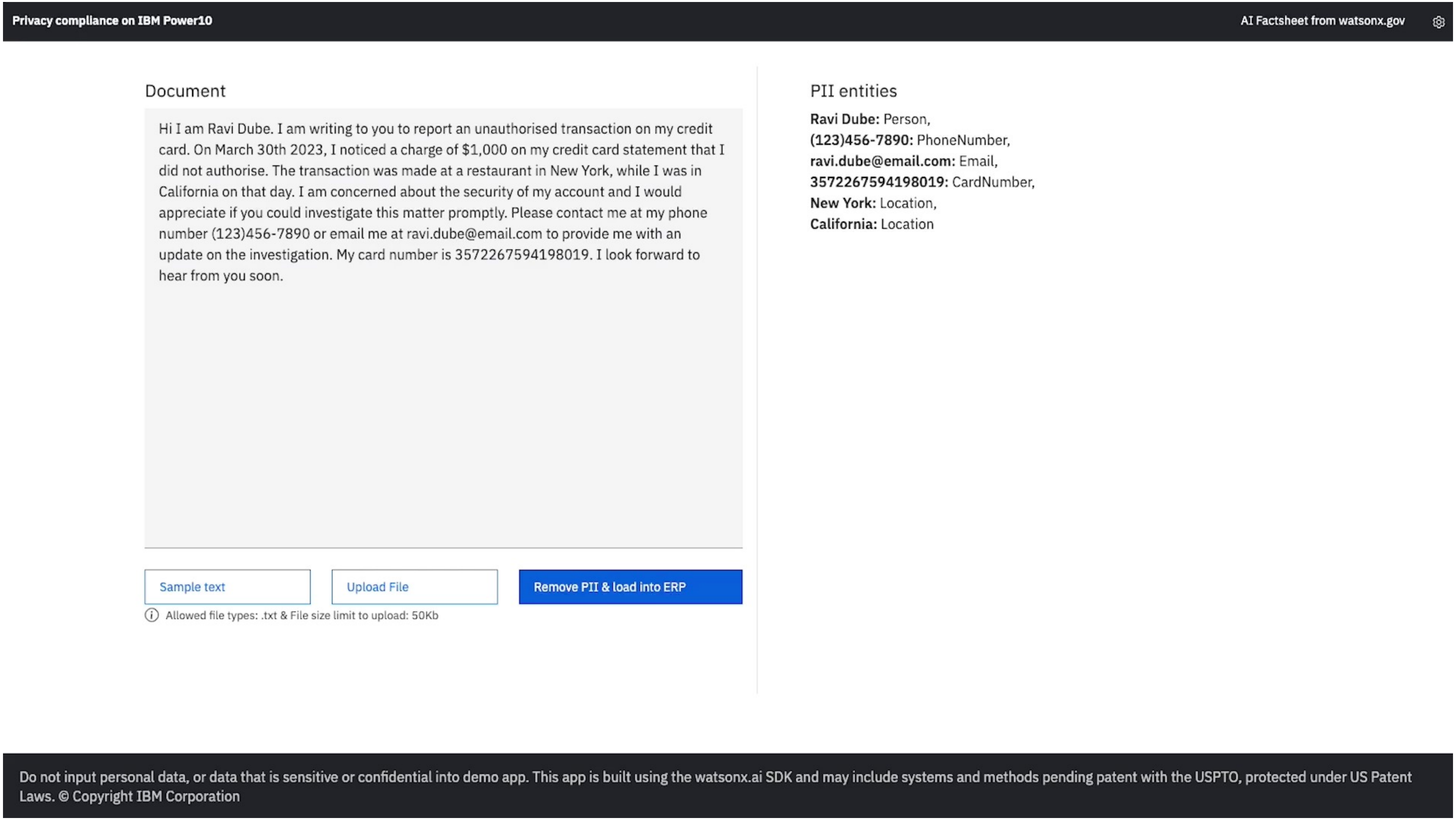
Sizing for large language models

- For best performance, try to maximize the number of cores per Power10 chip
- Plan with at least 1 dedicated chip per LLM; increase the number of chips when workload increases & replicate LLM
- Workload mainly depends on the number of concurrent model usages
- An LLM with 100B parameters may require ~100 GB RAM



Agent 9:17 AM

Hello! How can I help you today?


Privacy compliance on IBM Power10
AI Factsheet from watsonx.gov
Document
PII entities
Sample text
Upload File
Remove PII & load into ERP

Hi I am Ravi Dube. I am writing to you to report an unauthorised transaction on my credit card. On March 30th 2023, I noticed a charge of $1,000 on my credit card statement that I did not authorise. The transaction was made at a restaurant in New York, while I was in California on that day. I am concerned about the security of my account and I would appreciate if you could investigate this matter promptly. Please contact me at my phone number (123)456-7890 or email me at ravi.dube@email.com to provide me with an update on the investigation. My card number is 3572267594198019. I look forward to hear from you soon.

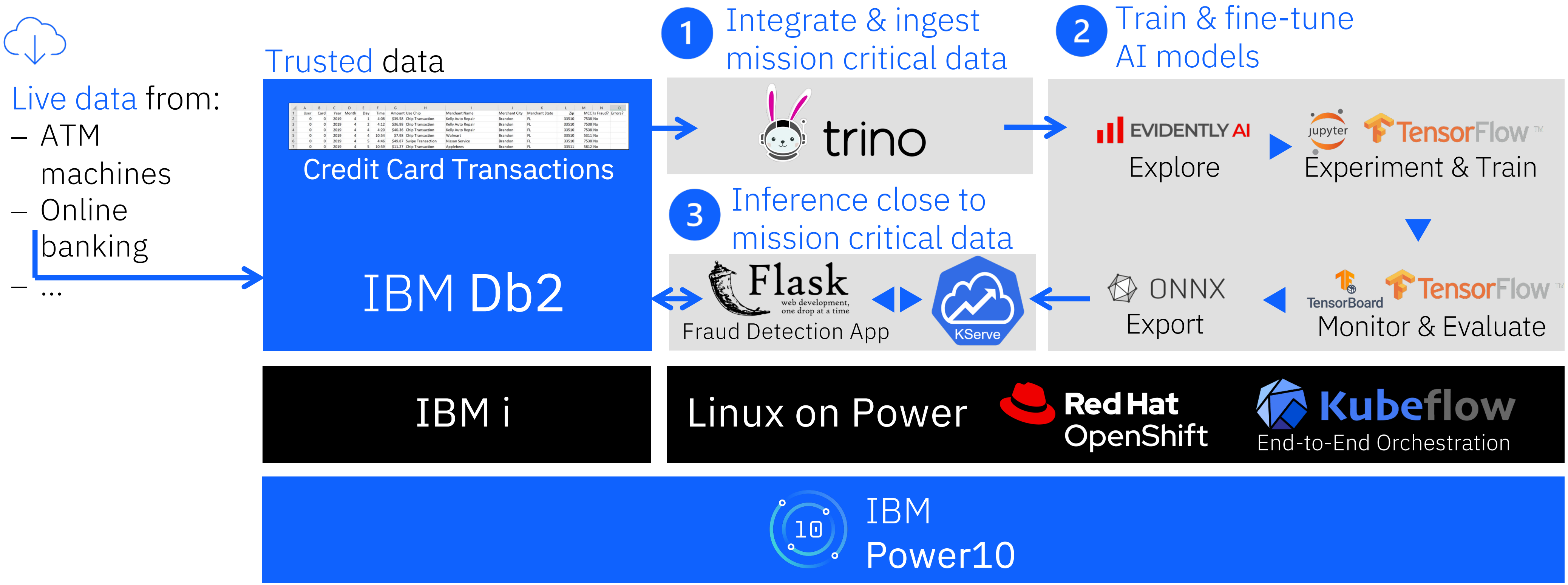
Ravi Dube: Person,
(123)456-7890: PhoneNumber,
ravi.dube@email.com: Email,
3572267594198019: CardNumber,
New York: Location,
California: Location

ⓘ Allowed file types: .txt & File size limit to upload: 50Kb

Do not input personal data, or data that is sensitive or confidential into demo app. This app is built using the watsonx.ai SDK and may include systems and methods pending patent with the USPTO, protected under US Patent Laws. © Copyright IBM Corporation

MLOps: Automating end-to-end AI workflows

Example: Real-time fraud detection close to IBM i data with Rocket AI Hub



Live data from:

- ATM machines
- Online banking
- ...

Trusted data

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	User	Card	Year	Month	Day	Time	Amount	Use Chip	Merchant Name	Merchant City	Merchant State	Zip	MCC	Is Fraud?	Errors?
2	0	0	2019	4	1	4:08	$39.58	Chip Transaction	Kelly Auto Repair	Brandon	FL	33510	7538	No	
3	0	0	2019	4	2	4:12	$36.98	Chip Transaction	Kelly Auto Repair	Brandon	FL	33510	7538	No	
4	0	0	2019	4	4	4:20	$40.36	Chip Transaction	Kelly Auto Repair	Brandon	FL	33510	7538	No	
5	0	0	2019	4	4	10:54	$7.98	Chip Transaction	Walmart	Brandon	FL	33510	5311	No	
6	0	0	2019	4	5	4:46	$49.87	Swipe Transaction	Nissan Service	Brandon	FL	33510	7538	No	
7	0	0	2019	4	5	10:59	$11.27	Chip Transaction	Applebees	Brandon	FL	33511	5812	No	

Credit Card Transactions

IBM Db2

1 Integrate & ingest mission critical data

2 Train & fine-tune AI models

Explore

Experiment & Train

Monitor & Evaluate

Export

3 Inference close to mission critical data

Fraud Detection App

IBM i

Linux on Power

End-to-End Orchestration

IBM Power10

Inferenced acceleration roadmap

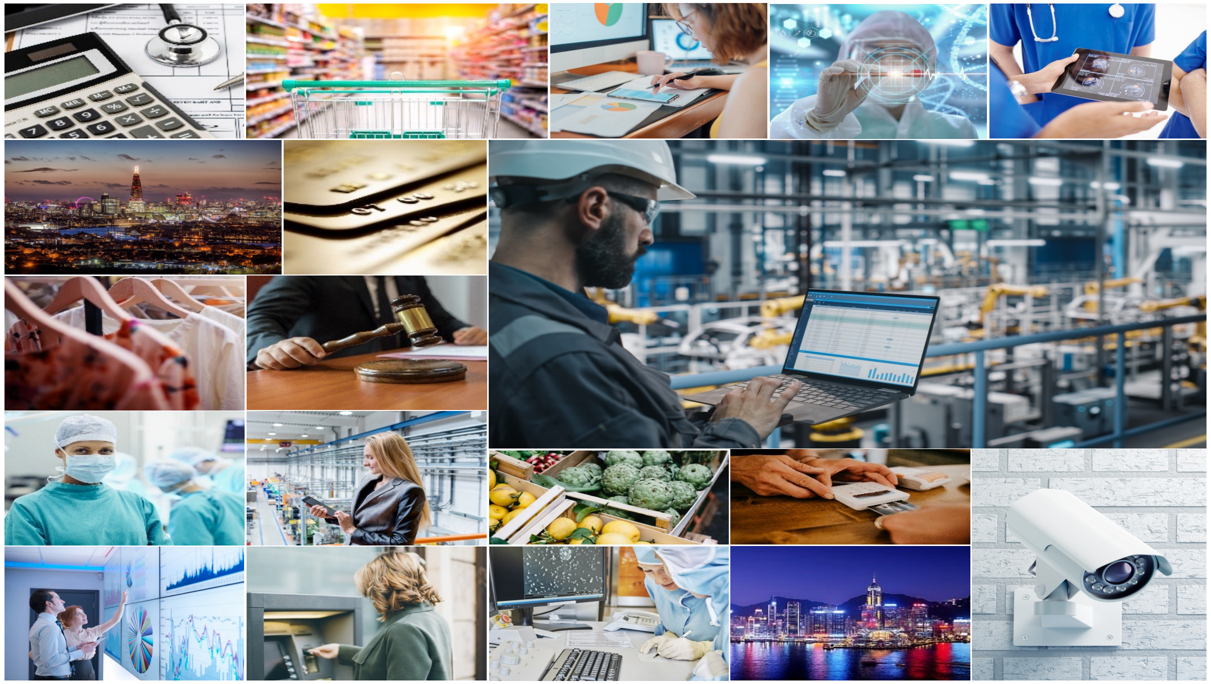
Scaling & throughput requirements →

IBM Power

- Traditional AI at scale
- Gen AI: Smaller single model targeted use cases, dev/test with 1-3 users, VectorDBs
- Computer vision use cases
- Audio processing use cases

2025 off-chip acceleration

- Gen AI with multiple large size models addressing a wide range of Enterprise AI use cases



Traditional & Small Language Model

Large Language Model

Model parameter size & accuracy →

IBM I & AI Strategy

Steve Will | IBM i CTO

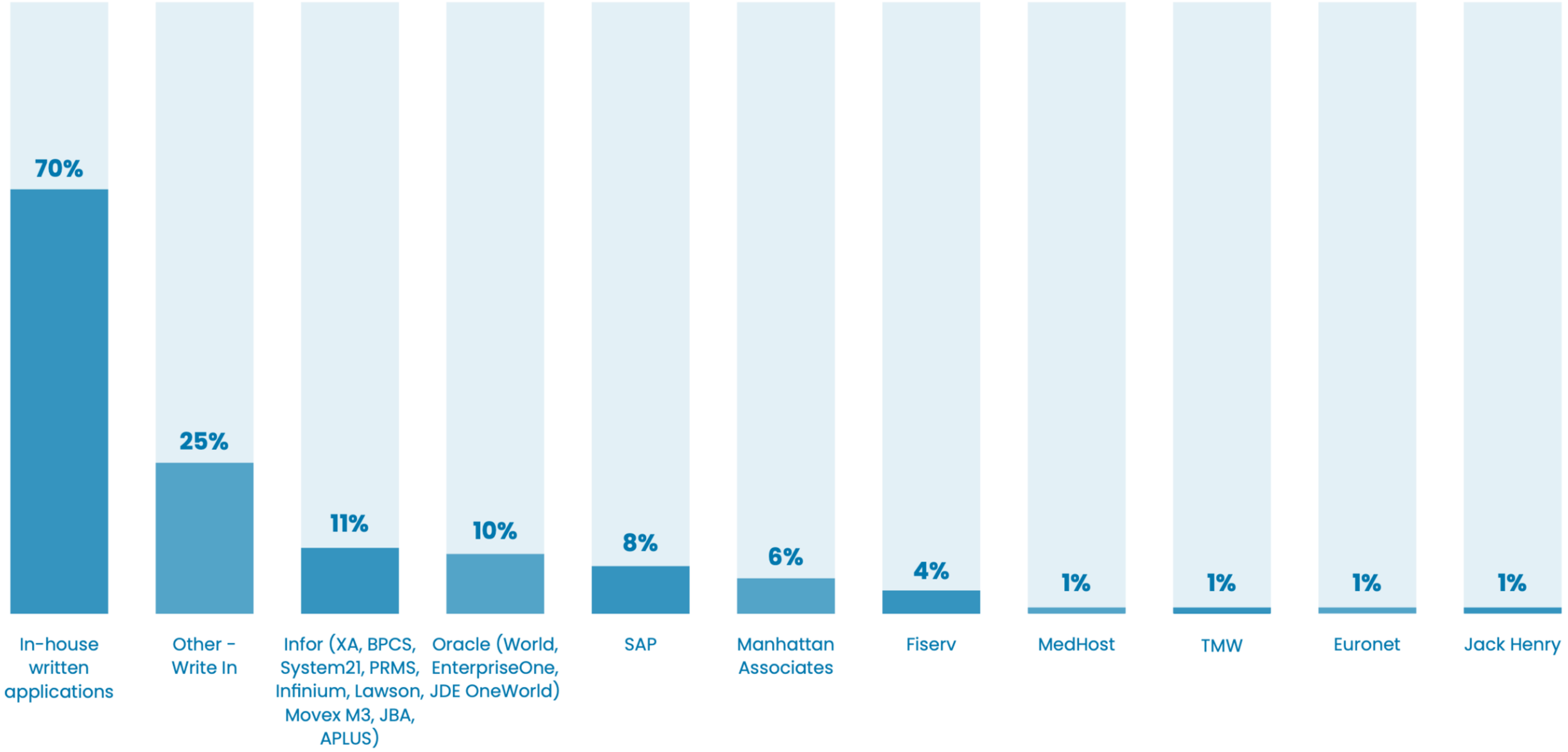
IBM i and AI: Three Use Cases, Today We Focus On One

Db2 Data Analytics
Trend Analysis
Anomaly detection

Operations: AIOps
Active Monitoring
Self-healing
INSTANA

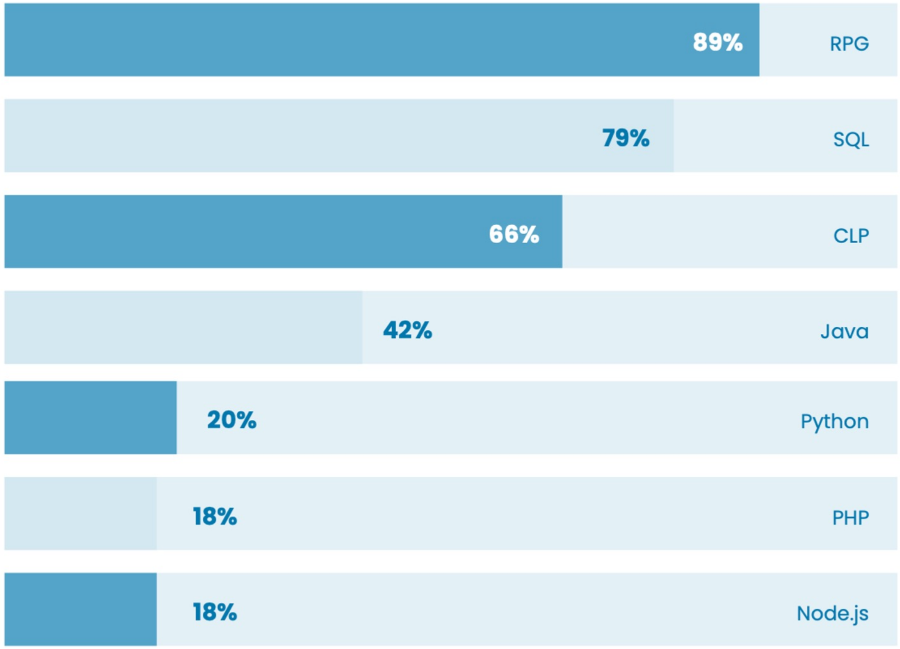
Developer
Experience

Which business applications are you running on IBM i?

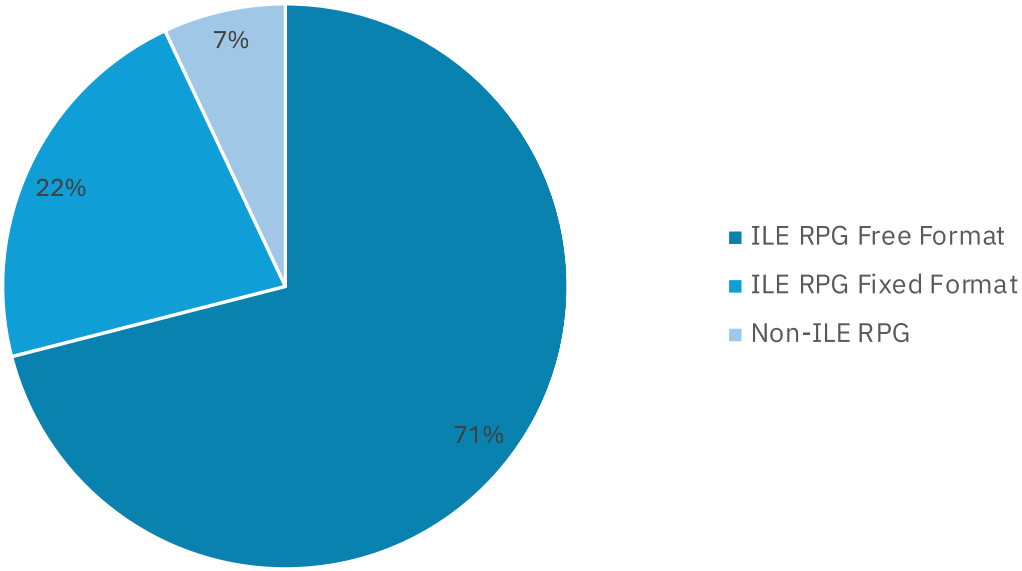


https://static.fortra.com/hs/pdfs/guides/marketplace/ibmi-marketplace-survey-results.pdf

Which development languages do you use today for new development on IBM i?

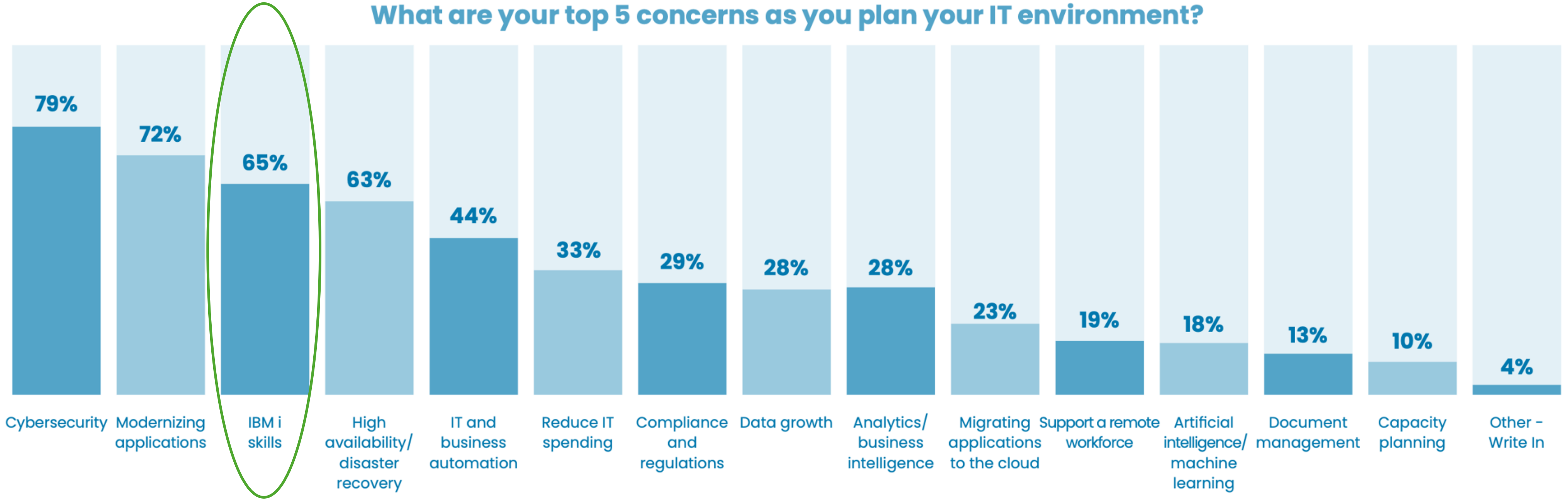


Primary Version of RPG for New Development



https://static.fortra.com/hs/pdfs/guides/marketplace/ibmi-marketplace-survey-results.pdf

What are your top 5 concerns as you plan your IT environment?



https://static.fortra.com/hs/pdfs/guides/marketplace/ibmi-marketplace-survey-results.pdf

Was aan, was af

Developer Experience

Help developer write code

Help understand code

How should the perfect RPG code assistant look like?

- Help programmers work with existing RPG

  Examine and explain existing RPG code

- Generate modern free-format ILE RPG based on a description

- Transform fixed-format ILE RPG into free-format ILE RPG

- Write test programs for RPG

How do you train such a model?

Using pairs

- block of code & explanation of block of code
  - Helps the model to understand English/SQL in relation to RPG code and create code according to a description
  - Generate code description from code itself
- block of code & block of code which tests the first block
  - How to test different blocks of code
  - How to transform old code into modern code

- block of old code & block of modern code achieving the same task

Using huge amount of non-paired code (more expensive and time consuming)

IBM i Approach: community based

- Use RPG code developed by IBM
- Use RPG code donated by IBM champions and experts
  - Susan Gantner, Jon Paris, Scott Klement, Jim Buck, Niels Liisberg, Yvonne Enselman, Mats Lidström, Koen DeCorte, Paul Tuohy, Steve Bradshaw, Hideyuki Yahagi
- Do you want to get involved?
  - Email alforIBMi@ibm.com
  - Agree to the license (utilisation of code to train LLM)
  - Submit code (https://ibm.github.io/rpg-genai-data) and decide whether you want the code to be accessible/not to outside IBM
  - Evaluate performance of the model once trained

Take home message

- 🤝 Be the Mr. Miyagi and the new generation will be grateful
- 🧑‍🏫 Oportunities to learn together
- 💥 Let’s make impact as a community for the RPG code assistant
- 🎉 Cheers to a 100 more years of IBM i

Andrea Longo | IBM Power Technical Specialist | andrea.longo@ibm.com