Université IBM i

uii2024
#ibmi
#uii2024

19 et 20 novembre 2024

IBM Innovation Studio Paris

S41 – AI et IBM Power

20 novembre 11:30 - 12:30

Marc Bouzigues
IBM Client Engineering EMEA
marc_bouzigues@fr.ibm.com

FRANCE

Université IBM i

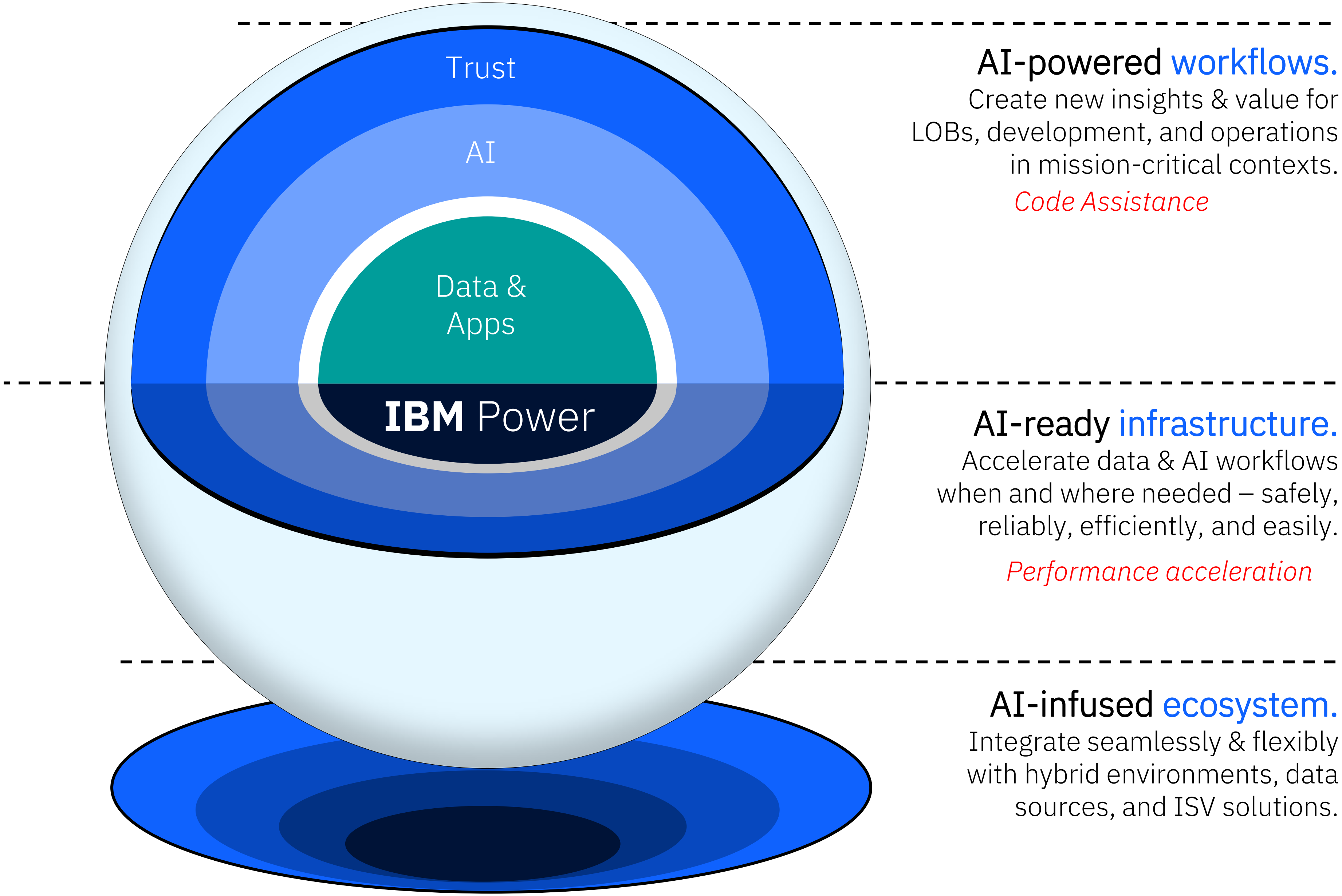
19 et 20 novembre 2024

IBM i

continuous innovation
continuous integration

AI et IBM Power

AI for Business
with **IBM** Power.



AI-powered workflows.

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

Code Assistance

AI-ready infrastructure.

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

Performance acceleration

AI-infused ecosystem.

Integrate seamlessly & flexibly with hybrid environments, data sources, and ISV solutions.

Market PoV

what our customers are asking for…

Top Industries:

- Banking
- Finance
- Healthcare
- Insurance
- Manufacturing
- Retail
- Public

Top GenAI Tasks & Use Cases:

- Q&A: Customer service & service desk, digital concierges, etc.
- Entity Extraction: Extract logistic information (addresses, products), medical information (diseases, treatments, medication), claim codes, locations, etc.
- Content Generation: Generate marketing briefs, reports for fraud analyses, IT issues & remediation steps, SQL for connecting to data on IBM Power, etc.
- Summarization: Summarize contracts, policies, regulations, medical reports, service tickets, etc.
- Process Flow Automation: Work order processing & PO processing using document digitalization & analysis, etc.

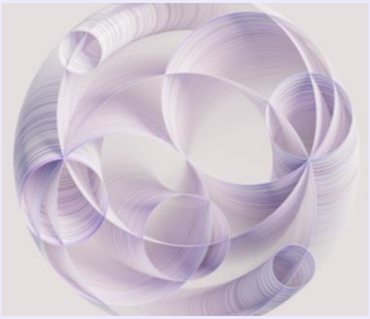
Top Classical ML Use Cases:

(time series analysis, regressions, decision trees, random forests, SVMs, clustering, …)

- Fraud & anomaly detection
- Demand forecasting
- Supply chain and inventory management
- Loan / investment risk analysis
- Predictive maintenance

Additional Top Use Cases:

- Computer Vision Inferencing
  - Manufacturing: visual quality inspection
  - Health care: Computer aided image analysis (e.g., cancer screening)
  - Law & Order: Security control, queue management, check out desks
  - Retail: Shelf stocking, produce spoilage
- Audio processing
  - Voice-to-Text & Text-to-Voice: improved digital concierges & audio insights



Power10 Processor Chip

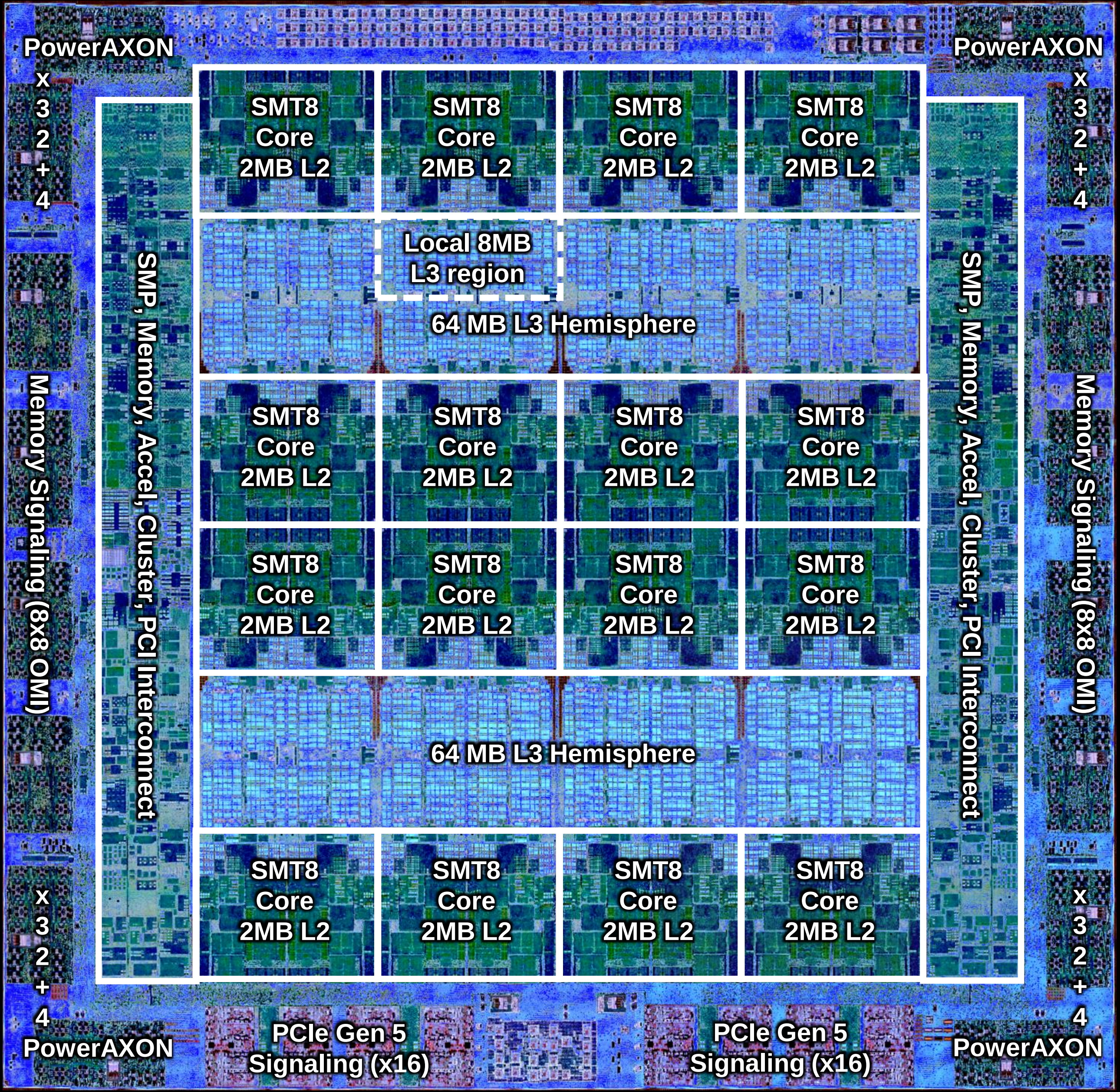
Technology and Packaging:
- 602mm² 7nm Samsung (18B devices)
- 18 layer metal stack, enhanced device
- Single-chip or Dual-chip sockets (New)

Computational Capabilities:
- Up to 15 SMT8 Cores (2 MB L2 Cache / core)
- Up to 120 MB L3 cache (low latency NUCA mgmt)
- Enterprise performance focus:
  - 1.3x core performance relative to POWER9
  - 1.2x thread strength relative to POWER9
  - 4x L2 cache, 4x MMU / core relative to POWER9
  - 4x crypto engines / core relative to POWER9

- AI computational focus MMA (Matrix Math Acceleration)
  - 2x general SIMD / core relative to POWER9
  - 4x matrix SIMD / core relative to POWER9
  - New AI instructions and data types

INFERENCE

Robust Data Plane:
- 2 TB/s raw (32 GT/s) PowerAXON + OMI signaling
- SMP interconnect for up to 16 sockets
- 2.2x OMI memory bandwidth relative to POWER9 (New)
- 64TB OMI DDR4 large system memory capacity (New)
- x64 PCIe Gen5 / DCM: 2x bandwidth relative to POWER9 (New

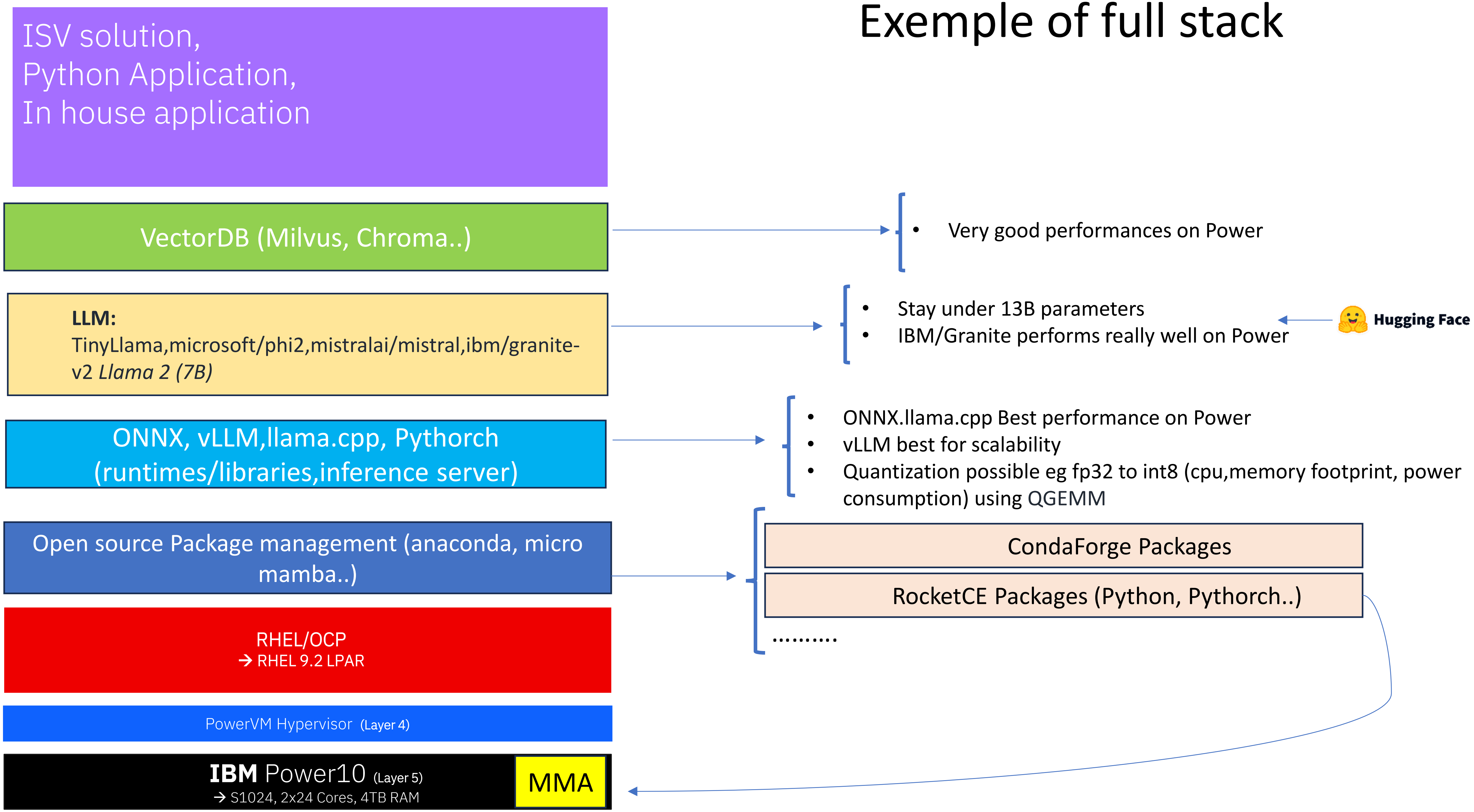


IBM

All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only

Vocabulary, notions and components related to AI and Gen AI

Exemple of full stack



ISV solution,
Python Application,
In house application

VectorDB (Milvus, Chroma..)

- Very good performances on Power

LLM:
TinyLlama,microsoft/phi2,mistralai/mistral,ibm/granite-v2 *Llama 2 (7B)*

- Stay under 13B parameters
- IBM/Granite performs really well on Power

🤗 **Hugging Face**

ONNX, vLLM,llama.cpp, Pythorch
(runtimes/libraries,inference server)

- ONNX.llama.cpp Best performance on Power
- vLLM best for scalability
- Quantization possible eg fp32 to int8 (cpu,memory footprint, power consumption) using QGEMM

Open source Package management (anaconda, micro mamba..)

CondaForge Packages

RocketCE Packages (Python, Pythorch..)

……….

RHEL/OCP
→ RHEL 9.2 LPAR

PowerVM Hypervisor (Layer 4)

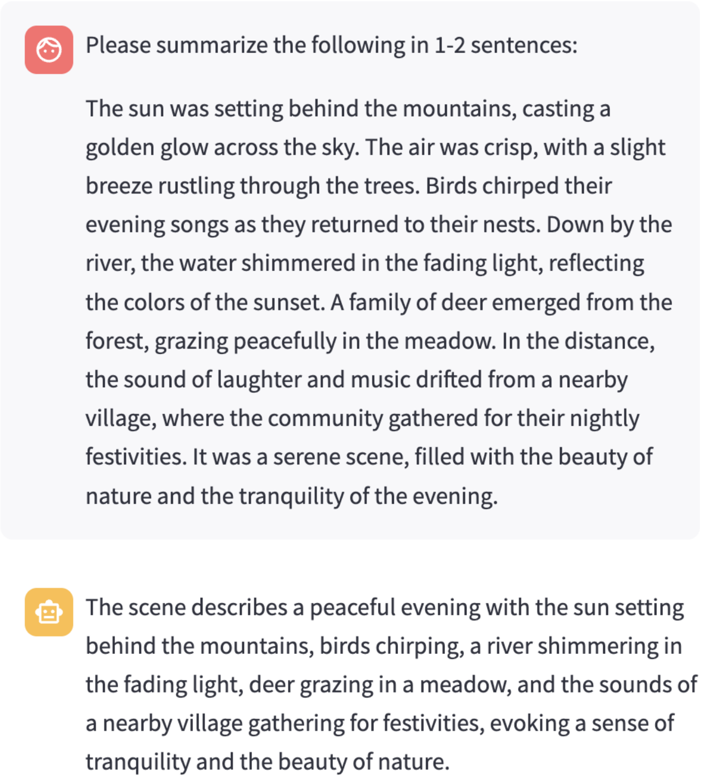
IBM Power10 (Layer 5)
→ S1024, 2x24 Cores, 4TB RAM

MMA

Gen AI differents scenarios examples

Scenario 1: Summarization, RAG

large input vs. small output



Please summarize the following in 1-2 sentences:

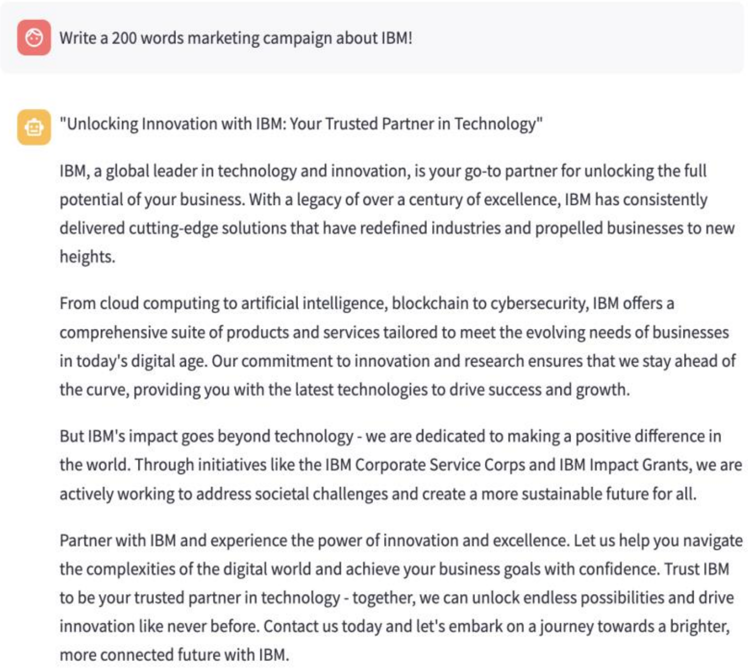
The sun was setting behind the mountains, casting a golden glow across the sky. The air was crisp, with a slight breeze rustling through the trees. Birds chirped their evening songs as they returned to their nests. Down by the river, the water shimmered in the fading light, reflecting the colors of the sunset. A family of deer emerged from the forest, grazing peacefully in the meadow. In the distance, the sound of laughter and music drifted from a nearby village, where the community gathered for their nightly festivities. It was a serene scene, filled with the beauty of nature and the tranquility of the evening.

The scene describes a peaceful evening with the sun setting behind the mountains, birds chirping, a river shimmering in the fading light, deer grazing in a meadow, and the sounds of a nearby village gathering for festivities, evoking a sense of tranquility and the beauty of nature.

Sequence Length (input): 1024 tokens
Generated Token (output): 200 tokens

Scenario 2: Code/Text Generation

small input vs. large output



Write a 200 words marketing campaign about IBM!

"Unlocking Innovation with IBM: Your Trusted Partner in Technology"

IBM, a global leader in technology and innovation, is your go-to partner for unlocking the full potential of your business. With a legacy of over a century of excellence, IBM has consistently delivered cutting-edge solutions that have redefined industries and propelled businesses to new heights.

From cloud computing to artificial intelligence, blockchain to cybersecurity, IBM offers a comprehensive suite of products and services tailored to meet the evolving needs of businesses in today's digital age. Our commitment to innovation and research ensures that we stay ahead of the curve, providing you with the latest technologies to drive success and growth.

But IBM's impact goes beyond technology - we are dedicated to making a positive difference in the world. Through initiatives like the IBM Corporate Service Corps and IBM Impact Grants, we are actively working to address societal challenges and create a more sustainable future for all.

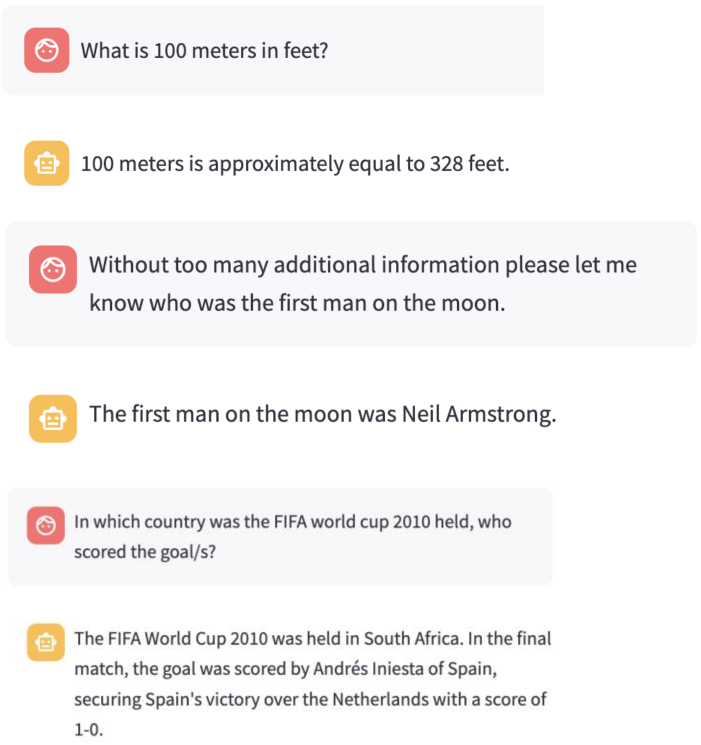
Partner with IBM and experience the power of innovation and excellence. Let us help you navigate the complexities of the digital world and achieve your business goals with confidence. Trust IBM to be your trusted partner in technology - together, we can unlock endless possibilities and drive innovation like never before. Contact us today and let's embark on a journey towards a brighter, more connected future with IBM.

Sequence Length (input): 32 tokens
Generated Token (output): 512 tokens

Gen AI differents scenarios

Scenario 3: Q&A, Translation

same input vs. same output



What is 100 meters in feet?

100 meters is approximately equal to 328 feet.

Without too many additional information please let me know who was the first man on the moon.

The first man on the moon was Neil Armstrong.

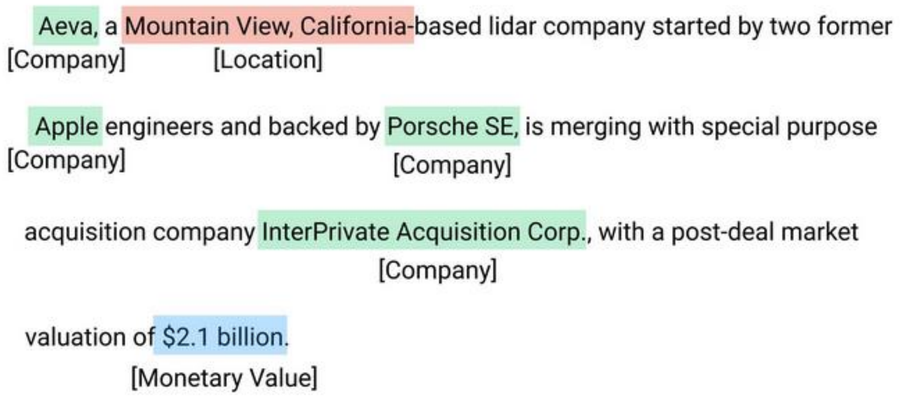
In which country was the FIFA world cup 2010 held, who scored the goal/s?

The FIFA World Cup 2010 was held in South Africa. In the final match, the goal was scored by Andrés Iniesta of Spain, securing Spain's victory over the Netherlands with a score of 1-0.

Sequence Length (input): 64 tokens
Generated Token (output): 64 tokens

Scenario 4: Entity Extraction

Large input vs. small output



Aeva, a Mountain View, California-based lidar company started by two former Apple engineers and backed by Porsche SE, is merging with special purpose acquisition company InterPrivate Acquisition Corp., with a post-deal market valuation of $2.1 billion.

Sequence Length (input): 512 tokens
Generated Token (output): 4 tokens

Understanding LLMs Key Metrics

1.Time To First Token (TTFT): How quickly users start seeing the model's output after entering their query. Low waiting times for a response are essential in real-time interactions, but less important in offline workloads. This metric is driven by the time required to process the prompt and then generate the first output token.

2.Time Per Output Token (TPOT): Time to generate an output token for *each* user that is querying our system. This metric corresponds with how each user will perceive the "speed" of the model. For example, a TPOT of 100 milliseconds/tok would be 10 tokens per second per user, or ~450 words per minute, which is faster than a typical person can read.

3. Latency Decode: The overall time it takes for the model to generate the output tokens.

4.Latency: The overall time it takes for the model to generate the full response for a user. Overall response latency can be calculated using the previous two metrics: latency = *(TTFT)* + *(TPOT)* * (the number of tokens to be generated). Latency = Prefill Latency + Latency Decode

5.Throughput: The number of output tokens per second an inference server can generate across all users and requests.

Phase 1 (Open Source stack)– Why using ONNX…

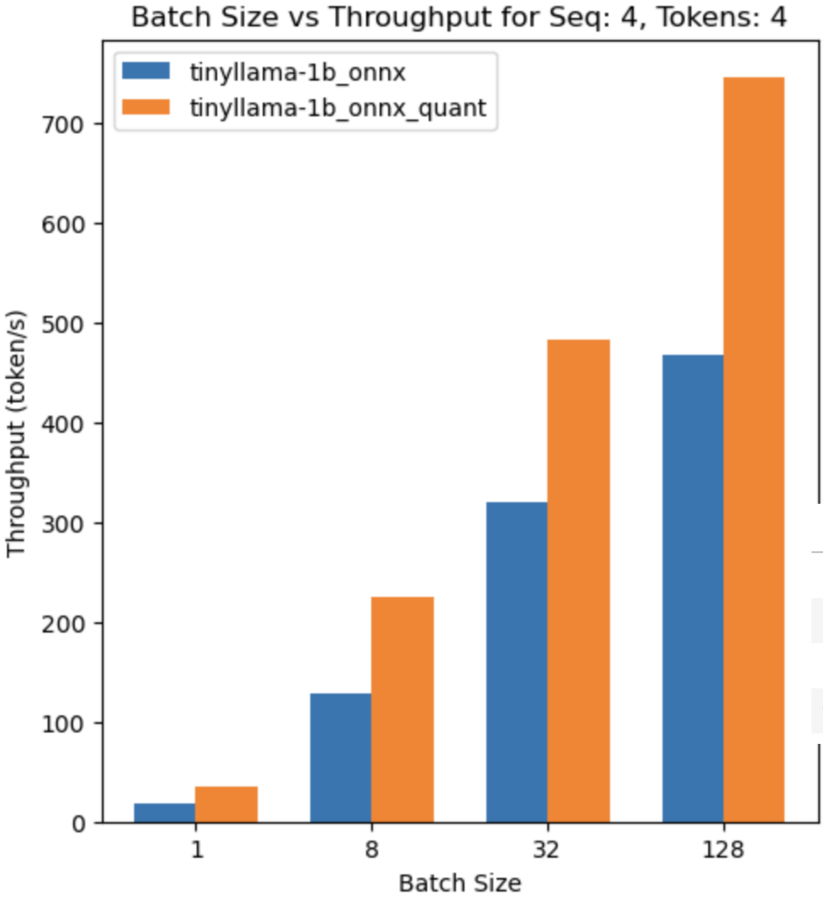
…and not pure PyTorch like most applications do?

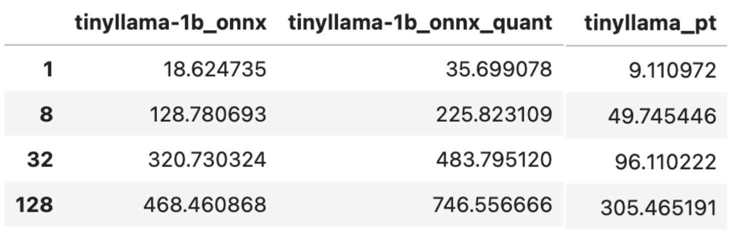
Pro

- **Performance** (see charts)
- Supported framework via RocketCE
- Possibility to quantize

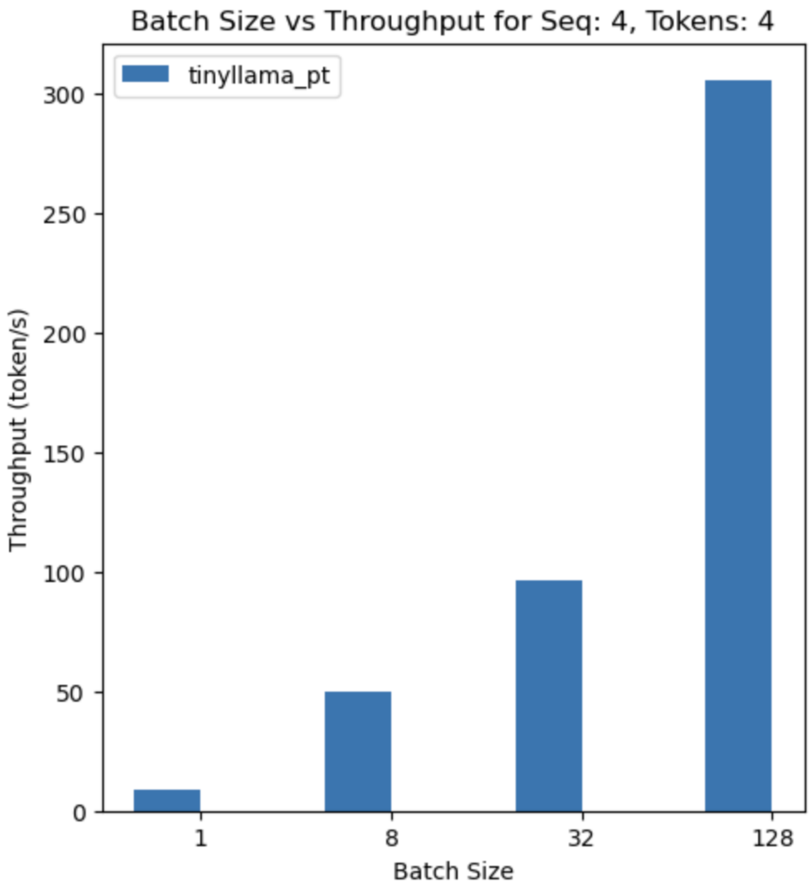
Contra

- Doesn‘t support all models (e.g. mixtral-8x7b currently unsupported)





	tinyllama-1b_onnx	tinyllama-1b_onnx_quant	tinyllama_pt
1	18.624735	35.699078	9.110972
8	128.780693	225.823109	49.745446
32	320.730324	483.795120	96.110222
128	468.460868	746.556666	305.465191



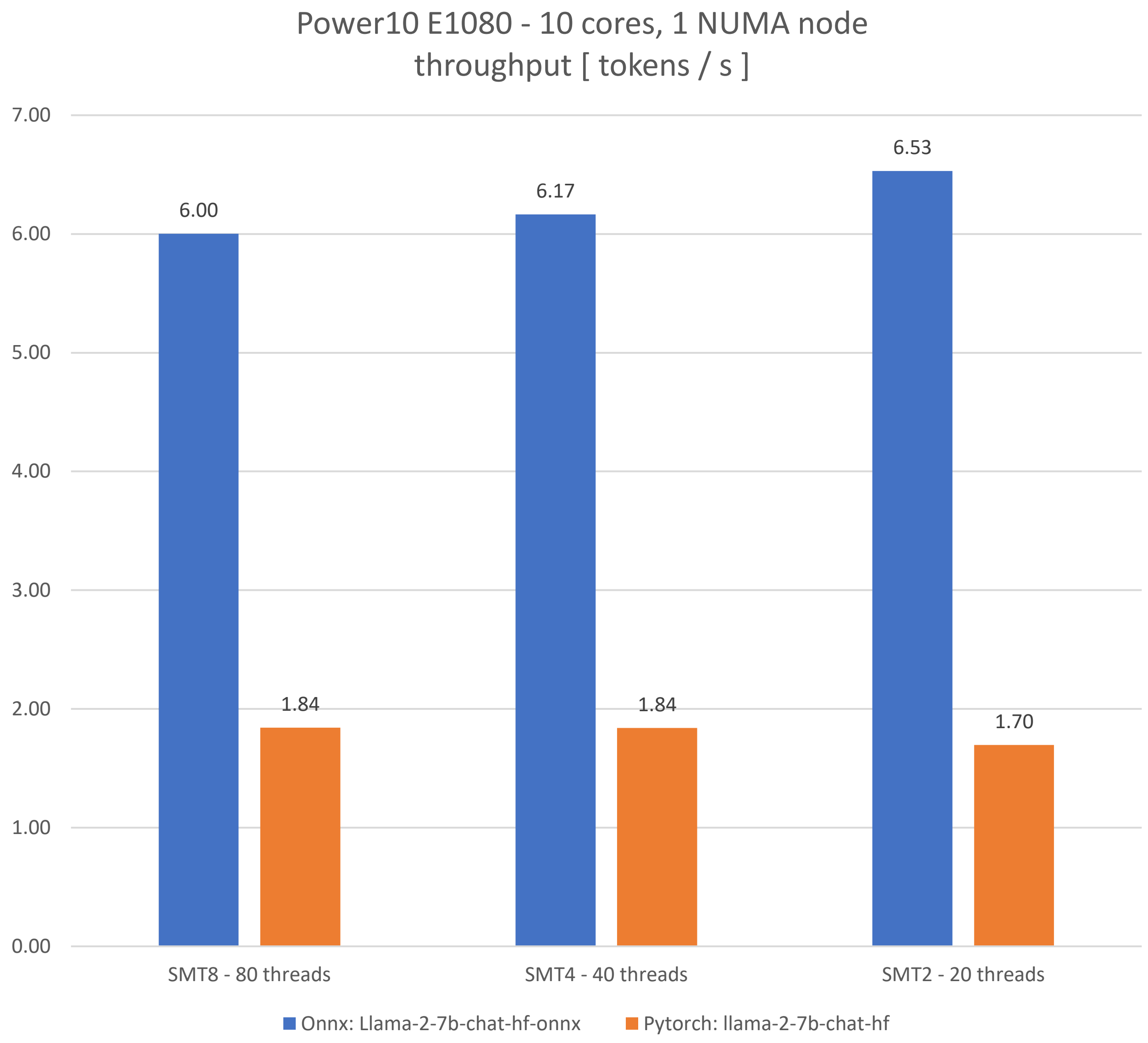
ONNX vs PyTorch runtime

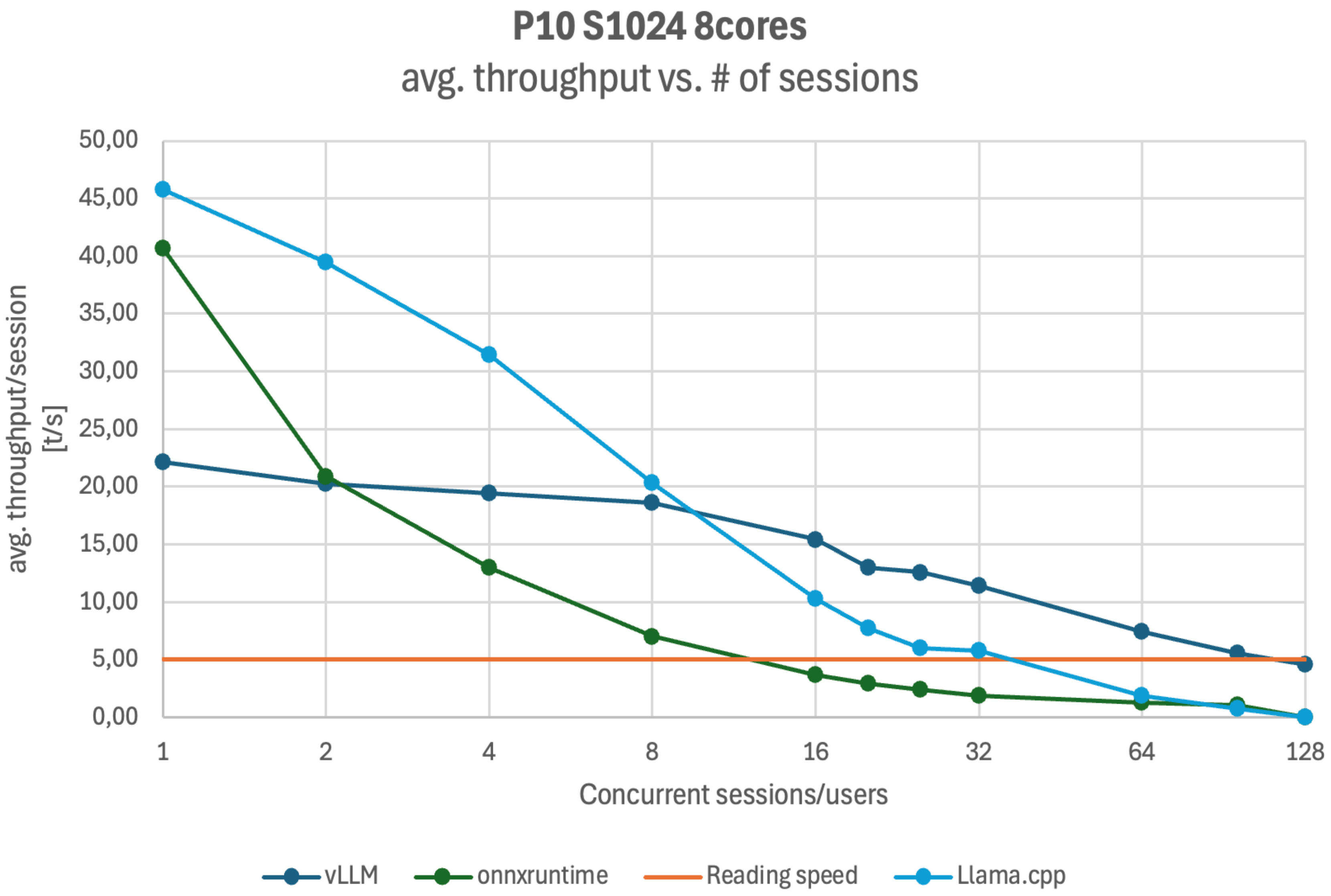
SITUATION

- Onnx and Pytorch runtime with MMA optimization
- Llama2-7B-chat model (FP32)
- Input prompt 7 tokens (batch_size=1)
- Output 128 tokens
  - Transformers and Optimum library
  - Onnx runtime, compile from source code or Conda
  - Pytorch from Conda Rocketce channel

Recommendation

- Use Onnx runtime and SMT=2 → higher throughput
- Onnx runtime utilize more MMA instructions
- For Onnx you can specify session options e.g number of threads for inference and thread affinity – NUMA

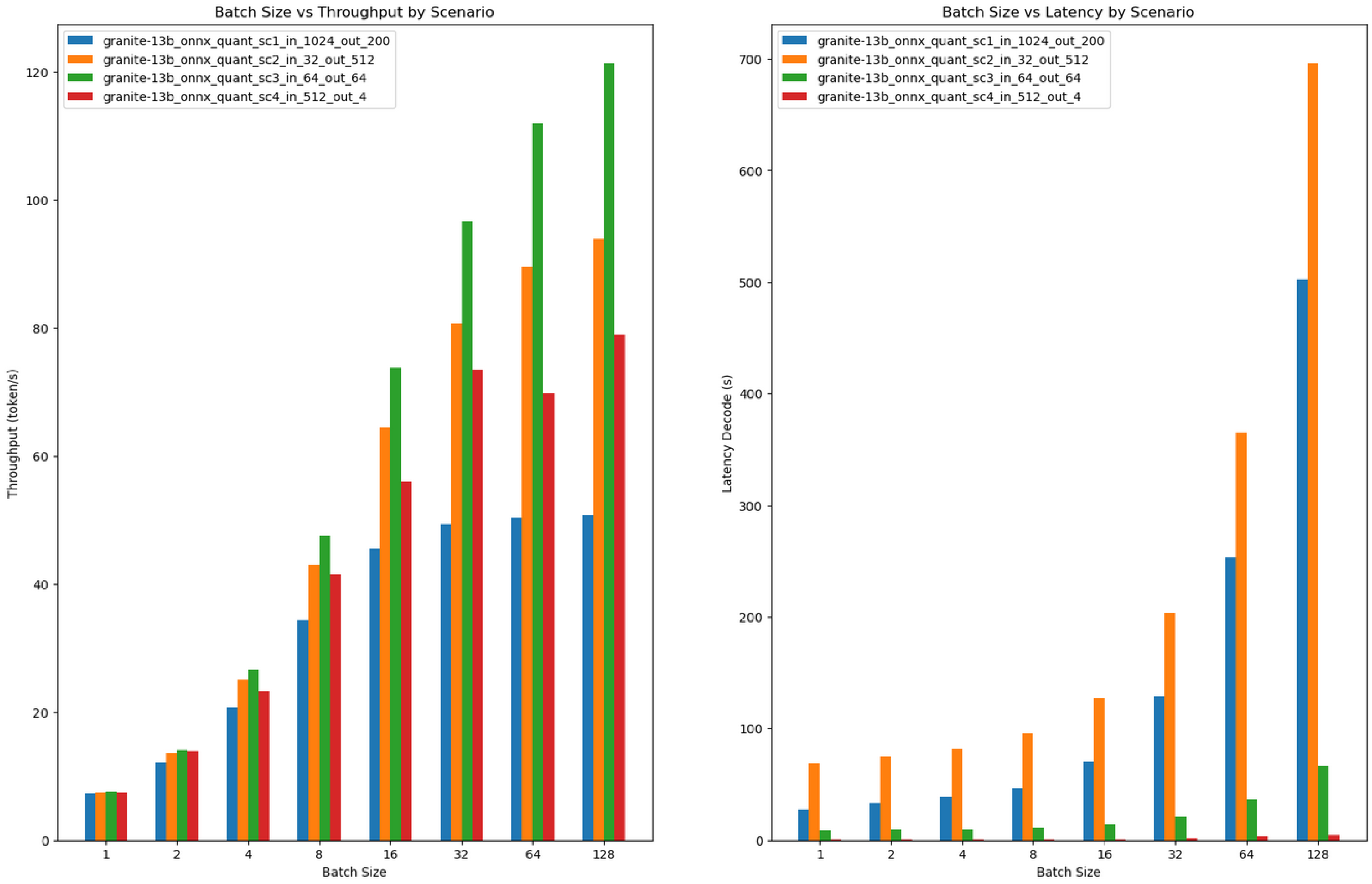




Key Observations

- Onnxruntime and llama.cpp have their sweetspot for single user/fewer users
- vLLM is the most stable inference engine and has the flattest decrease with increasing concurrency
- Focus on vLLM to align with IBM strategy as this will be the backend of watsonx.ai

Test results Granite-v2 13B



Scenario 1:
large input vs. small output
(Summarization, RAG)

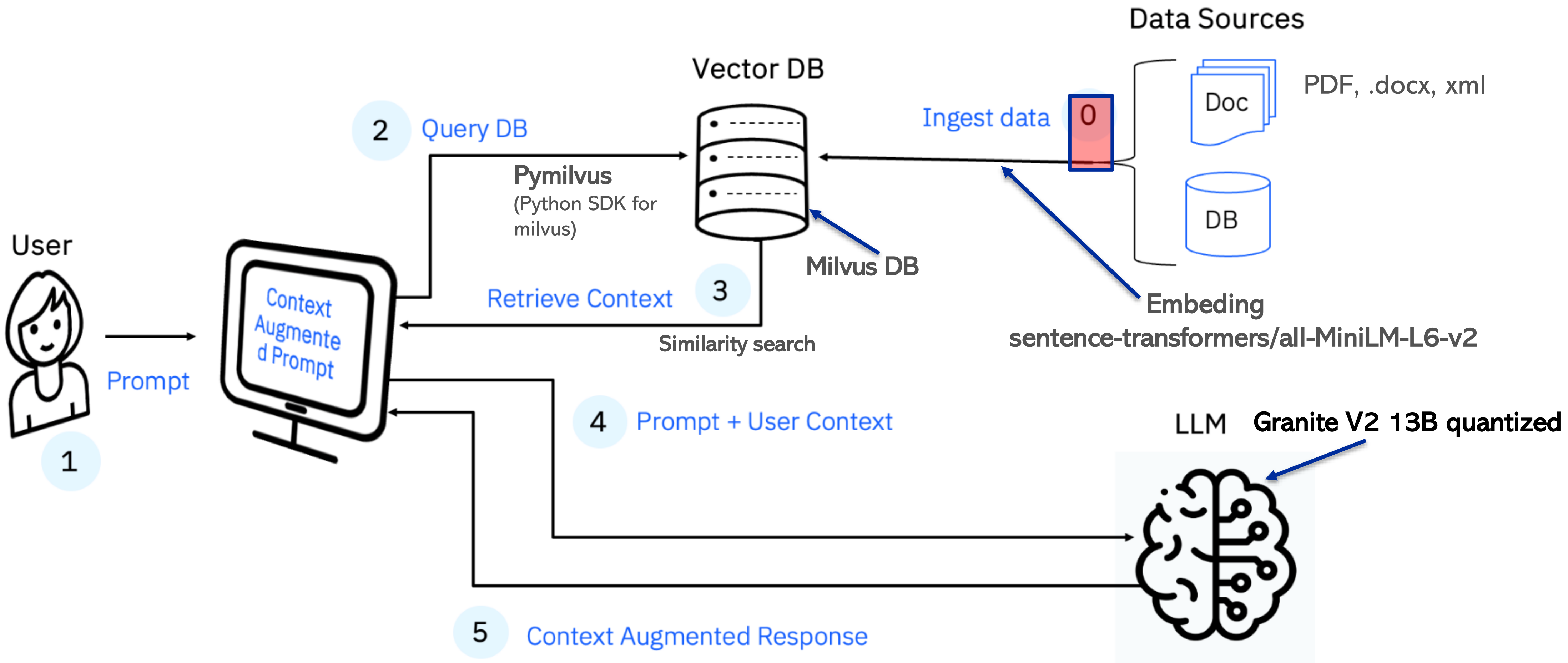
Scenario 2:
small input vs. large output
(Code/Text Generation)

Scenario 3:
same input/output sizes
(Q&A, Translation)

Scenario 4:
large input vs. small output
(Entity extraction)

Personalize GenAI without re-train a Model : RAG

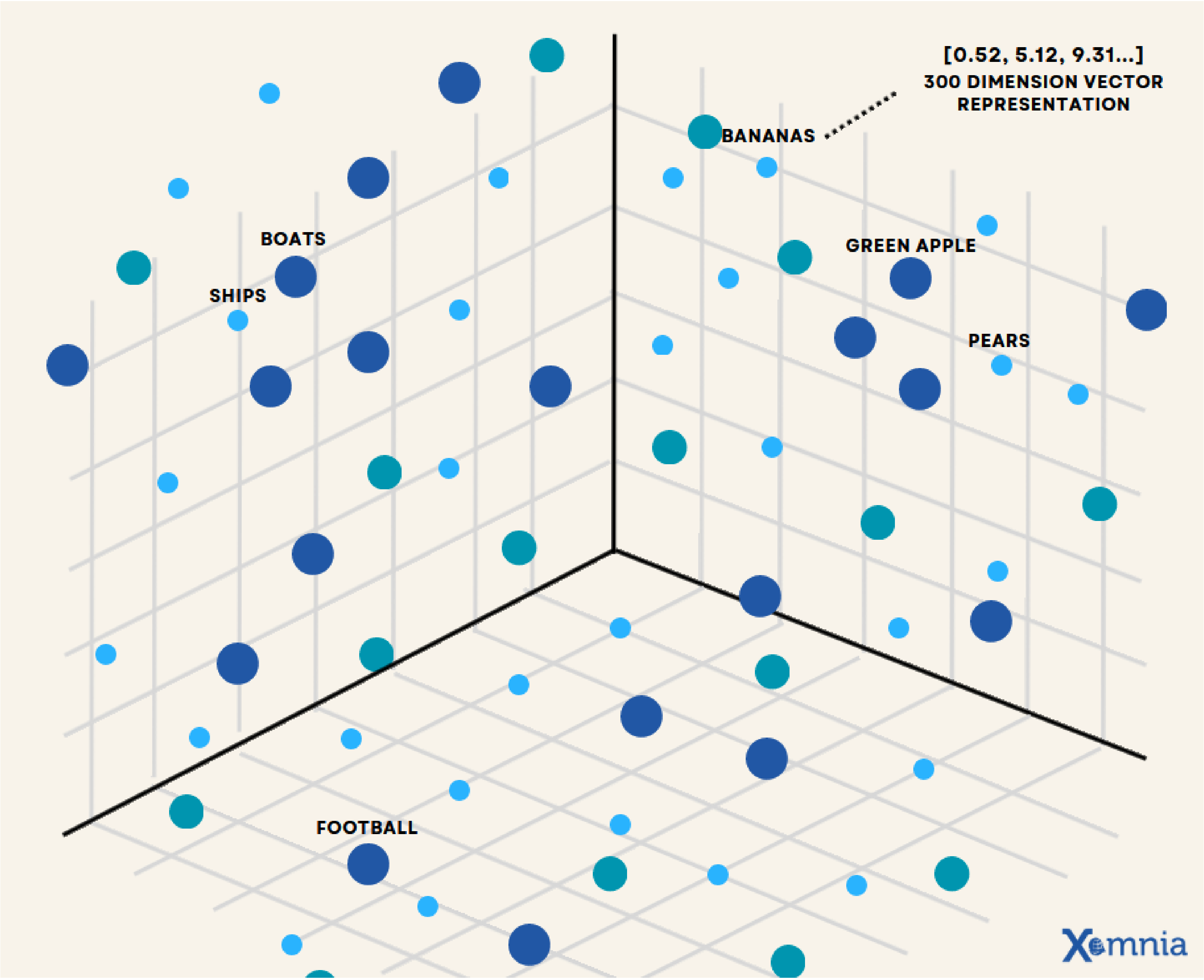
Adapt models to your domain via Retrieval Augmented Generation (RAG)



Vector Databases can make searching much more flexible

Traditional search typically represents data by using discrete tokens or features, such as keywords, tags or metadata. Traditional searches rely on exact matches to retrieve relevant results. For example, a search for "smartphone" would return results containing the word "smartphone.“

Vector representations enable similarity search. For example, a vector search for “smartphone” might also return results for “cellphone” and “mobile devices.”



https://www.xomnia.com/post/an-introduction-to-vector-databases-for-beginners/

Picking a vector database: a comparison and guide for 2023

	Pinecone	Weaviate	Milvus	Qdrant	Chroma	Elasticsearch	PGvector
Is open source	×	✅	✅	✅	✅	×	✅
Self-host	×	✅	✅	✅	✅	✅	✅
Cloud management	✅	✅	✅	✅	×	✅	(✔)
Purpose-built for Vectors	✅	✅	✅	✅	✅	×	×
Developer experience	👍👍👍	👍👍	👍👍	👍👍	👍👍	👍	👍
Community	Community page & events	8k☆ github, 4k slack	23k☆ github, 4k slack	13k☆ github, 3k discord	9k☆ github, 6k discord	23k slack	6k☆ github
Queries per second (using text nytimes-256-angular)	150 *for p2, but more pods can be added	791	2406	326	?	700-100 *from various reports	141
Latency, ms (Recall/Percentile 95 (millis), nytimes-256-angular)	1 *batched search, 0.99 recall, 200k SBERT	2	1	4	?	?	8

	Pinecone	Weaviate	Milvus	Qdrant	Chroma	Elasticsearch	PGvector
Supported index types	?	HNSW	Multiple (11 total)	HNSW	HNSW	HNSW	HNSW/IVFFlat
Hybrid Search (i.e. scalar filtering)	✅	✅	✅	✅	✅	✅	✅
Disk index support	✅	✅	✅	✅	✅	×	✅
Role-based access control	✅	×	✅	×	×	✅	×
Dynamic segment placement vs. static data sharding	?	Static sharding	Dynamic segment placement	Static sharding	Dynamic segment placement	Static sharding	-
Free hosted tier	✅	✅	✅	(free self-hosted)	(free self-hosted)	(free self-hosted)	(varies)
Pricing (50k vectors @1536)	$70	fr. $25	fr. $65	est. $9	Varies	$95	Varies
Pricing (20M vectors, 20M req. @768)	$227 ($2074 for high performance)	$1536	fr. $309 ($2291 for high performance)	fr. $281 ($820 for high performance)	Varies	est. $1225	Varies

Known to work on IBM Power
Known not to work on IBM Power (yet)

https://benchmark.vectorview.ai/vectordbs.html

Example of models running on Power (not an exhaustive list)

Models ranging from 1B…13B parameters are appropriates for Power

- LLMs for Inferencing
  1. TinyLlama (1B)
  2. microsoft/phi-2 (3B)
  3. mistralai/mistral (7B)
  4. ibm/granite (13b)
  5. *Llama 2 (7B)*
  6. *Lot more ….*

- *Sentence transformers (for embedding creation)*
  1. *ST/all-MiniLM-L6-v2*
  2. *….*

- Vector DB that runs on Power
  1. ChromaDB
  2. Milvus
  3. ….

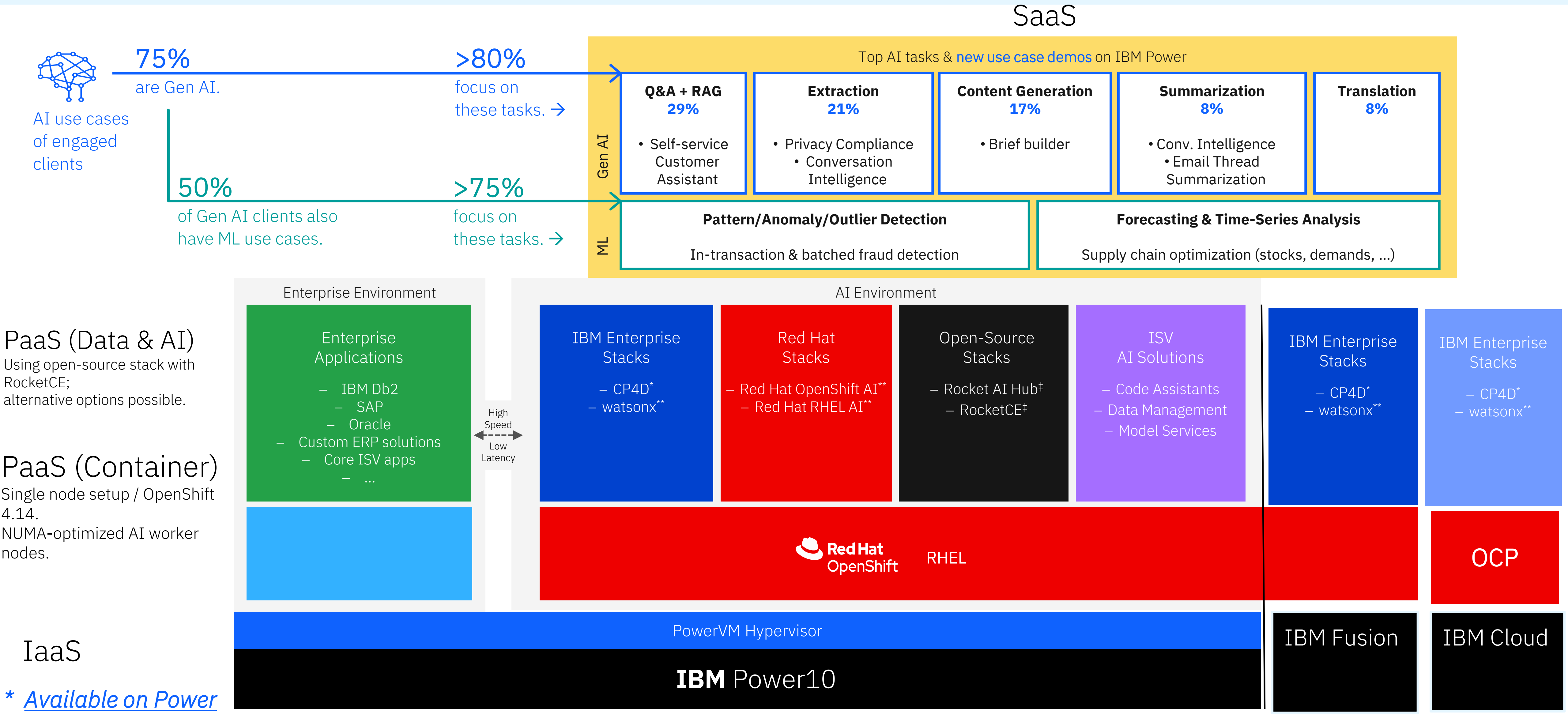
A model is a file module, that includes the parameters weight, model architecture the size depends on the parameter's numbers

Transformers : “Paris is a very beautiful city. I lived there for 3 years; it was a very enriching experience, and I loved it. I made great progress in history, but most importantly, it taught me how to speak …. ?”

indemnified models ?

sentence-transformers model: It maps sentences & paragraphs to a 384 dimensional dense vector space and can be used for tasks like clustering or semantic search.

Reference Architecture# for AI on Power, focusing on top client use cases



SaaS

Top AI tasks & new use case demos on IBM Power

AI use cases of engaged clients

75% are Gen AI.

>80% focus on these tasks. →

50% of Gen AI clients also have ML use cases.

>75% focus on these tasks. →

Gen AI

Q&A + RAG
29%
- Self-service Customer Assistant

Extraction
21%
- Privacy Compliance
- Conversation Intelligence

Content Generation
17%
- Brief builder

Summarization
8%
- Conv. Intelligence
- Email Thread Summarization

Translation
8%

ML

Pattern/Anomaly/Outlier Detection

In-transaction & batched fraud detection

Forecasting & Time-Series Analysis

Supply chain optimization (stocks, demands, ...)

PaaS (Data & AI)

Using open-source stack with RocketCE;
alternative options possible.

PaaS (Container)

Single node setup / OpenShift 4.14.
NUMA-optimized AI worker nodes.

IaaS

Enterprise Environment

Enterprise Applications
- IBM Db2
- SAP
- Oracle
- Custom ERP solutions
- Core ISV apps
- ...

High Speed ↔ Low Latency

AI Environment

IBM Enterprise Stacks
- CP4D*
- watsonx**

Red Hat Stacks
- Red Hat OpenShift AI**
- Red Hat RHEL AI**

Open-Source Stacks
- Rocket AI Hub‡
- RocketCE‡

ISV AI Solutions
- Code Assistants
- Data Management
- Model Services

Red Hat OpenShift RHEL

PowerVM Hypervisor

IBM Power10

IBM Enterprise Stacks
- CP4D*
- watsonx**

IBM Fusion

IBM Enterprise Stacks
- CP4D*
- watsonx**

OCP

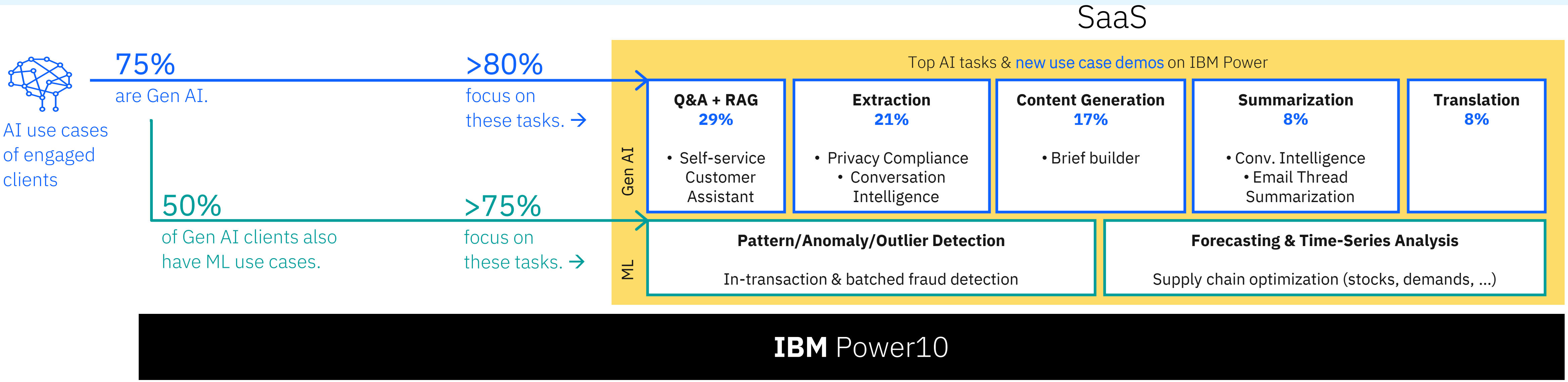
IBM Cloud

* *Available on Power*

** *In plan for Power*

# *can be done either on-premise or off-premise*

Reference Architecture# for AI on Power, focusing on top client use cases



SaaS

AI use cases of engaged clients

75% are Gen AI.

>80% focus on these tasks. →

50% of Gen AI clients also have ML use cases.

>75% focus on these tasks. →

Top AI tasks & new use case demos on IBM Power

Gen AI

Q&A + RAG
29%
- Self-service Customer Assistant

Extraction
21%
- Privacy Compliance
- Conversation Intelligence

Content Generation
17%
- Brief builder

Summarization
8%
- Conv. Intelligence
- Email Thread Summarization

Translation
8%

ML

Pattern/Anomaly/Outlier Detection

In-transaction & batched fraud detection

Forecasting & Time-Series Analysis

Supply chain optimization (stocks, demands, ...)

IBM Power10

Conversational Ai use cases
- output streaming
- SLO output adult reading rate 5 words/sec

Batchable Ai use cases (API)
- output streaming
- SLO output adult reading rate 5 words/sec

GenAI & ML Stack – Layer view

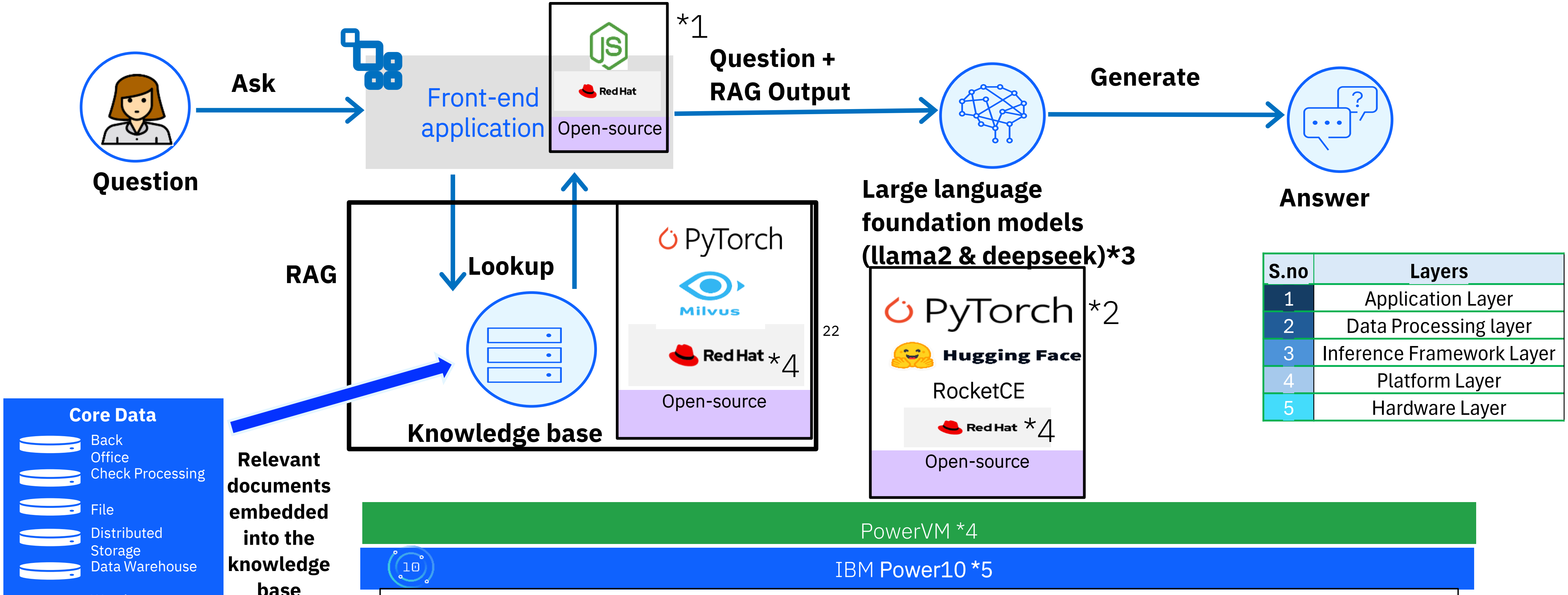
S.no	Layers	Use	Typical Components# for GenAI	Typical Components for Machine Learning (Open-source , Red Hat / IBM CP4D)
1	Application Layer	Application	Node.js, React.js, JAVA, Python	Node.js, React.js, JAVA, Python
		API Gateway	FastAPI (Python), 3Scale (RH)	FastAPI(Python), 3Scale / WatsonML
		Model Serving Service	FastAPI, Container Image	FastAPI, Container Image / WatsonML
2	Data Processing layer	Pre-processing	Python (RocketCE)	Python (RocketCE) / Watson Studio
		Tokenization (Hugging Face Tokenizers)	Huggingface (no pid)	Huggingface (no pid) / Analytics Engine for Apache Spark
		Post-processing	Python (RocketCE)	Python (RocketCE) / Watson Studio
3	Training (for ML) & Inference Framework Layer	Llama2 (7B/13B)**		
		Pytorch, ONNX, Tensorflow	RocketCE	RocketCE / Watson Studio (Jupyter or AutoAI)
4	Platform layer	Operating System	RHEL or Openshift	RHEL or Openshift / Openshift only
		Virtualization	PowerVM	PowerVM
5	Hardware Layer	Power10	Compute – per LPAR – 1 socket 12 cores/14 cores S10xx or 15 cores E10xX	Compute – per LPAR – 1 socket 2-6 cores S10xx or 1-2 cores E10xX / Use CP4D sizing tool **
		10 GbE network interface	Network -10 GB	Network -10 GB
		NVME SSD for storing model weights	Storage – 2x disks (excl OCP requirements)	Storage – 2x disks (excl OCP requirements)

Reference document for sizing CP4D on Power

# Community (Open-CE) and commercially supported (Rocket CE) options available

Client Use Cases – Chat Bot/Digital Assistant with GenAI (RAG)

Chatbot for helping customers with queries on housing data in APAC/Middle east



S.no	Layers
1	Application Layer
2	Data Processing layer
3	Inference Framework Layer
4	Platform Layer
5	Hardware Layer

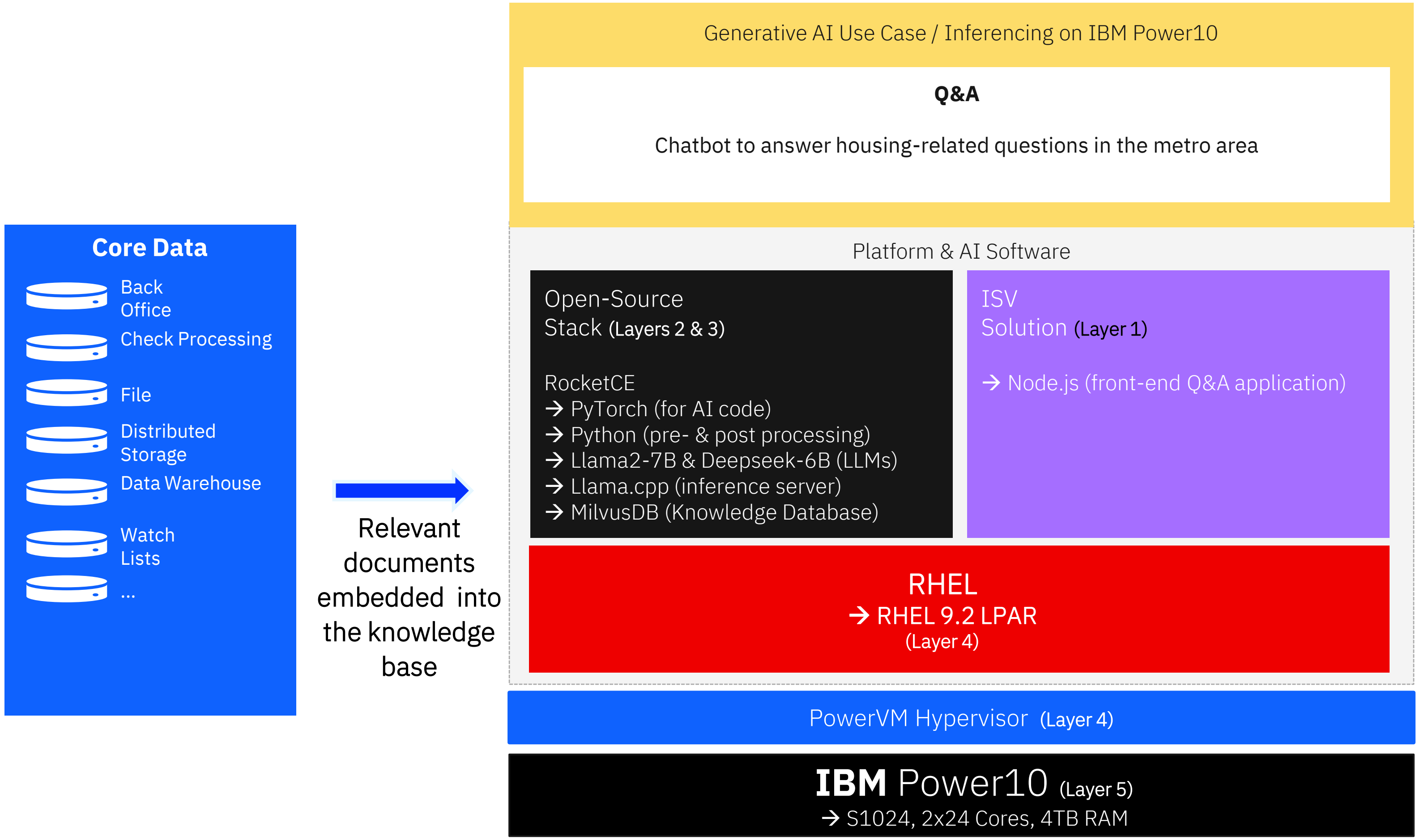
Why IBM Power10

- Data security on P10
- Run the inference close to the data - keep application and data secure
- Lack of GPU availability in the region

Government entity in APAC/middle east

Reference Architecture for specific AI use cases – Chat Bot

Government Agency in APAC/Middle east



Core Data

- Back Office
- Check Processing
- File
- Distributed Storage
- Data Warehouse
- Watch Lists
- ...

Relevant documents embedded into the knowledge base

Generative AI Use Case / Inferencing on IBM Power10

Q&A

Chatbot to answer housing-related questions in the metro area

Platform & AI Software

Open-Source Stack (Layers 2 & 3)

RocketCE
→ PyTorch (for AI code)
→ Python (pre- & post processing)
→ Llama2-7B & Deepseek-6B (LLMs)
→ Llama.cpp (inference server)
→ MilvusDB (Knowledge Database)

ISV Solution (Layer 1)

→ Node.js (front-end Q&A application)

RHEL
→ RHEL 9.2 LPAR
(Layer 4)

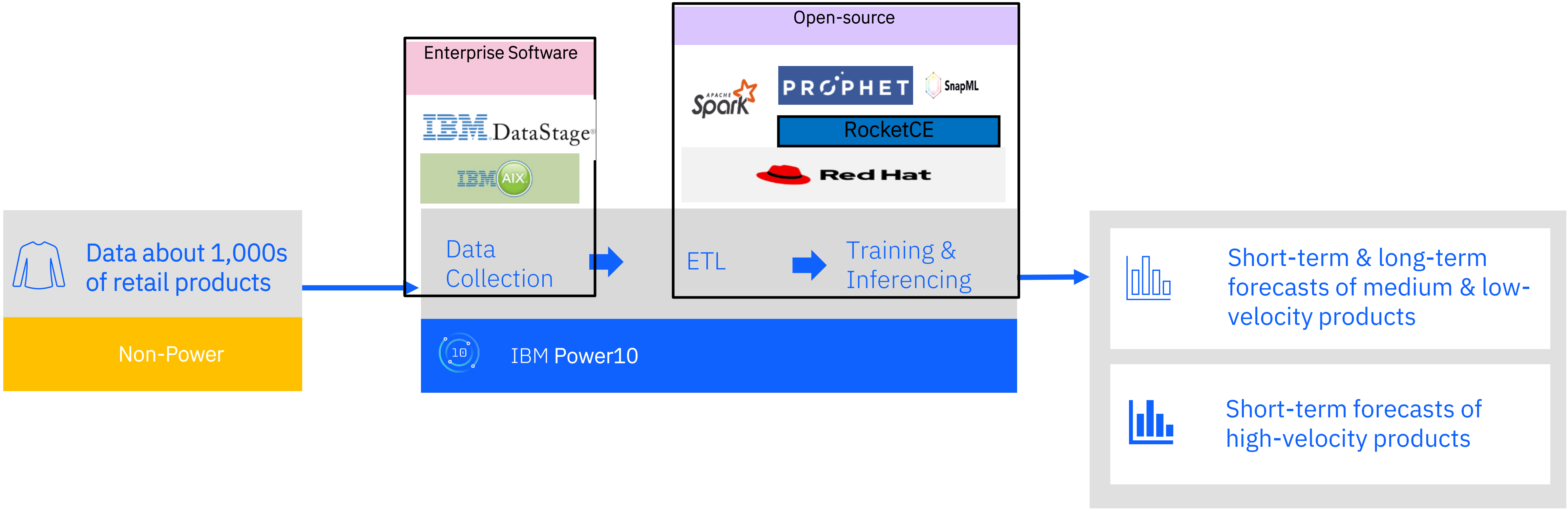
PowerVM Hypervisor (Layer 4)

IBM Power10 (Layer 5)
→ S1024, 2x24 Cores, 4TB RAM

S.no	Layers
1	Application Layer
2	Data Processing layer
3	Inference Framework Layer
4	Platform Layer
5	Hardware Layer

Client Use Cases – Demand forecasting

Forecasts for thousands of retail products & retail stores for stock planning



Data about 1,000s of retail products

Non-Power

Enterprise Software

IBM DataStage

IBM AIX

Data Collection

Open-source

Apache Spark

PROPHET

SnapML

RocketCE

Red Hat

ETL

Training & Inferencing

IBM Power10

Short-term & long-term forecasts of medium & low-velocity products

Short-term forecasts of high-velocity products

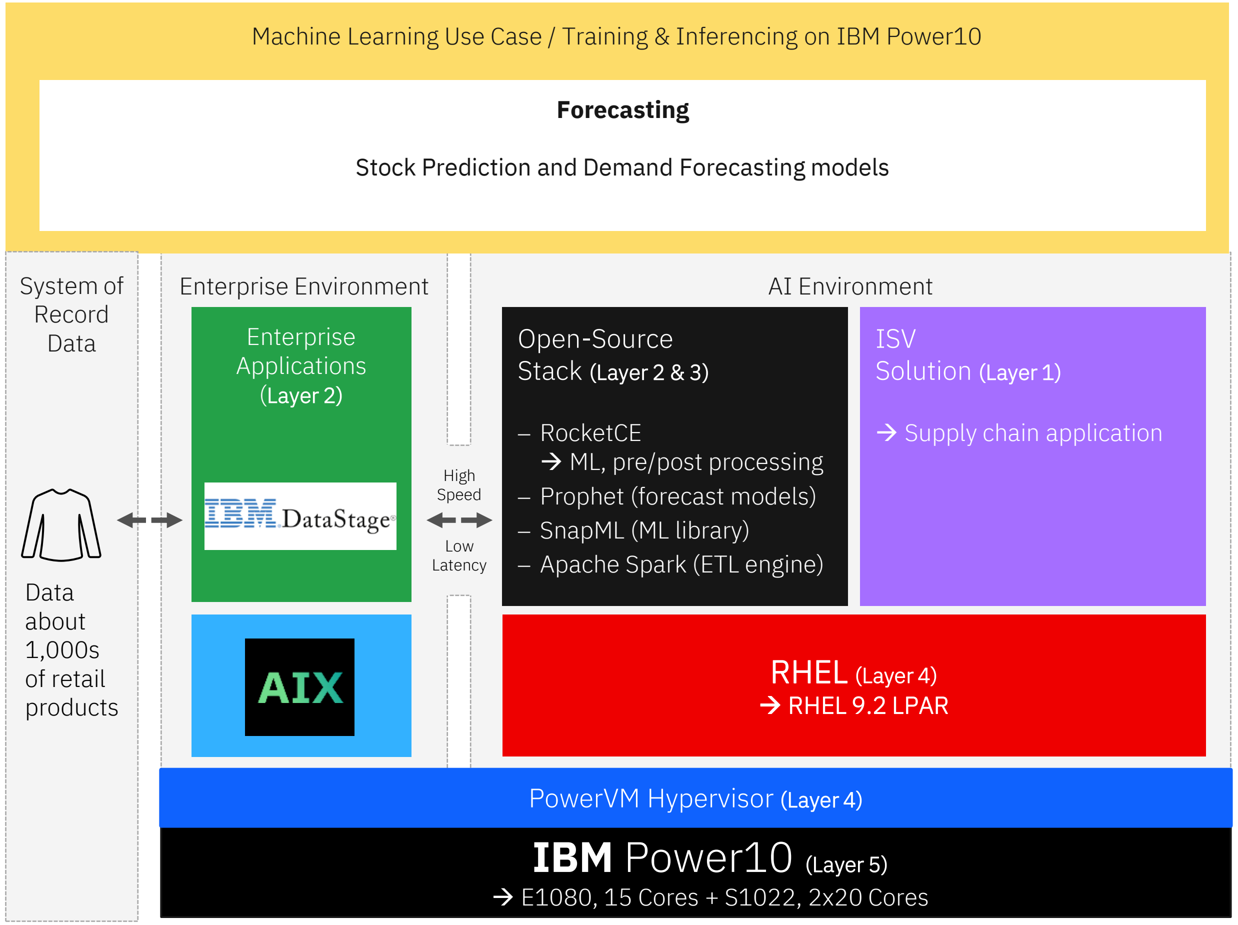
Why IBM Power10

- Very good throughput for timeseries-based forecasting
- Better cost/performance (vs. x86) for data collection due to consolidation

Large Retailer in Americas

Reference Architecture for specific AI use cases – Machine Learning

Retail Chain in North America



Machine Learning Use Case / Training & Inferencing on IBM Power10

Forecasting

Stock Prediction and Demand Forecasting models

System of Record Data

Data about 1,000s of retail products

Enterprise Environment

Enterprise Applications (Layer 2)

IBM DataStage

AIX

High Speed

Low Latency

AI Environment

Open-Source Stack (Layer 2 & 3)

- RocketCE
  → ML, pre/post processing
- Prophet (forecast models)
- SnapML (ML library)
- Apache Spark (ETL engine)

ISV Solution (Layer 1)

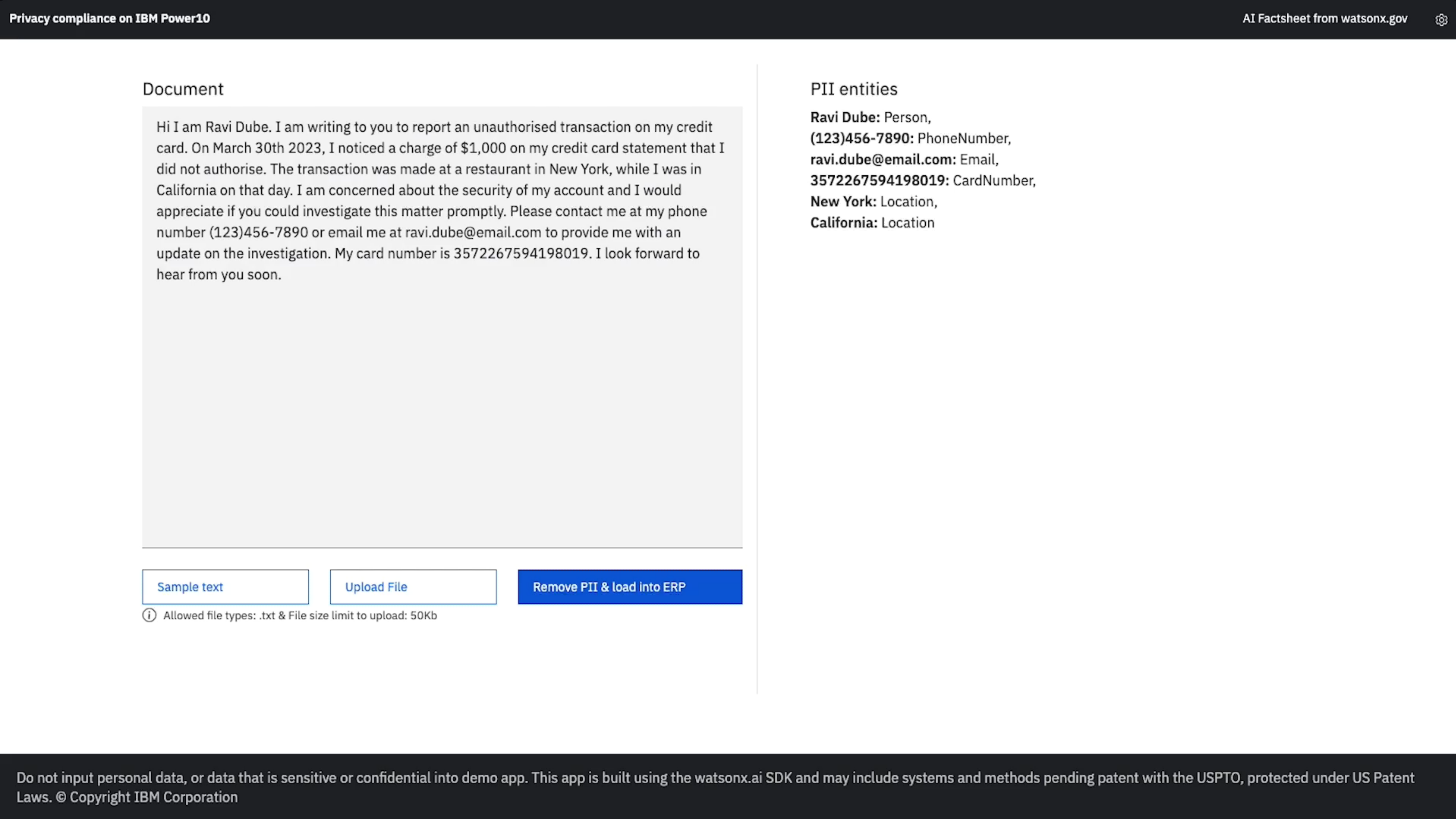
→ Supply chain application

RHEL (Layer 4)
→ RHEL 9.2 LPAR

PowerVM Hypervisor (Layer 4)

IBM Power10 (Layer 5)
→ E1080, 15 Cores + S1022, 2x20 Cores

S.no	Layers
1	Application Layer
2	Data Processing layer
3	Training and Inference (ML) Framework Layer
4	Platform Layer
5	Hardware Layer


Privacy compliance on IBM Power10
AI Factsheet from watsonx.gov

Document

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.

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

PII entities

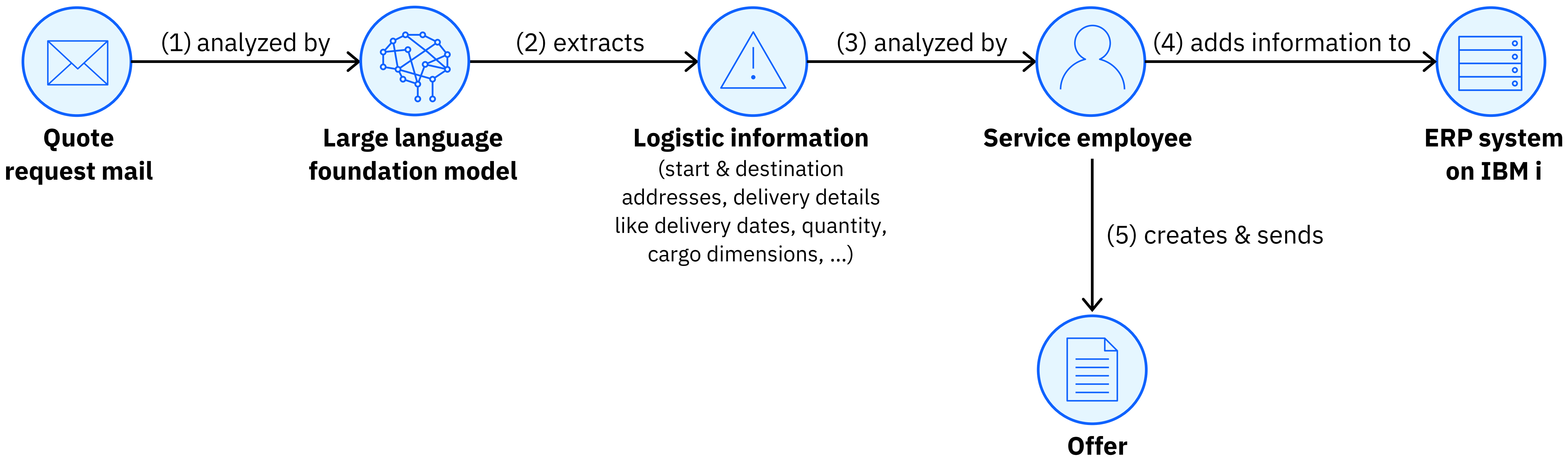
Ravi Dube: Person,
(123)456-7890: PhoneNumber,
ravi.dube@email.com: Email,
3572267594198019: CardNumber,
New York: Location,
California: Location

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

From mails to ERP system

IBM Power10 customer extracts information from mails asking for quotes, puts those information into an ERP system on IBM i, and creates an offer.
[“Erfahrungsbericht: Generative KI bei Hans Geis auf Power10”: www.sva.de/de/events/ifutureday]

Pattern 1:
Using Gen AI to transform unstructured data (text) into structured data (information) that is ingested to IBM Power.



Quote request mail

(1) analyzed by

Large language foundation model

(2) extracts

Logistic information
(start & destination addresses, delivery details like delivery dates, quantity, cargo dimensions, ...)

(3) analyzed by

Service employee

(4) adds information to

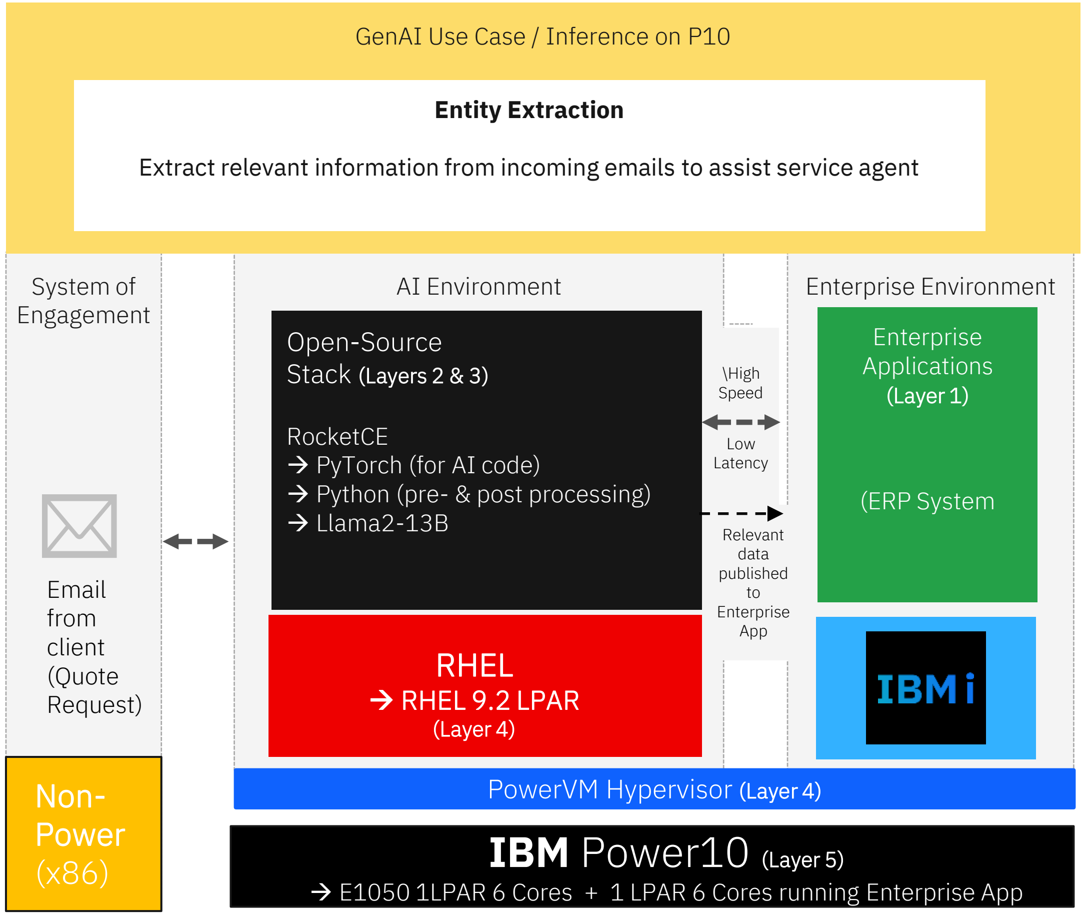
ERP system on IBM i

(5) creates & sends

Offer

Reference Architecture for specific AI use cases – Entity Extraction

Logistics Company in Europe



GenAI Use Case / Inference on P10

Entity Extraction

Extract relevant information from incoming emails to assist service agent

System of Engagement

Email from client (Quote Request)

Non-Power (x86)

AI Environment

Open-Source Stack (Layers 2 & 3)

RocketCE
→ PyTorch (for AI code)
→ Python (pre- & post processing)
→ Llama2-13B

RHEL
→ RHEL 9.2 LPAR
(Layer 4)

\High Speed

Low Latency

Relevant data published to Enterprise App

Enterprise Environment

Enterprise Applications (Layer 1)

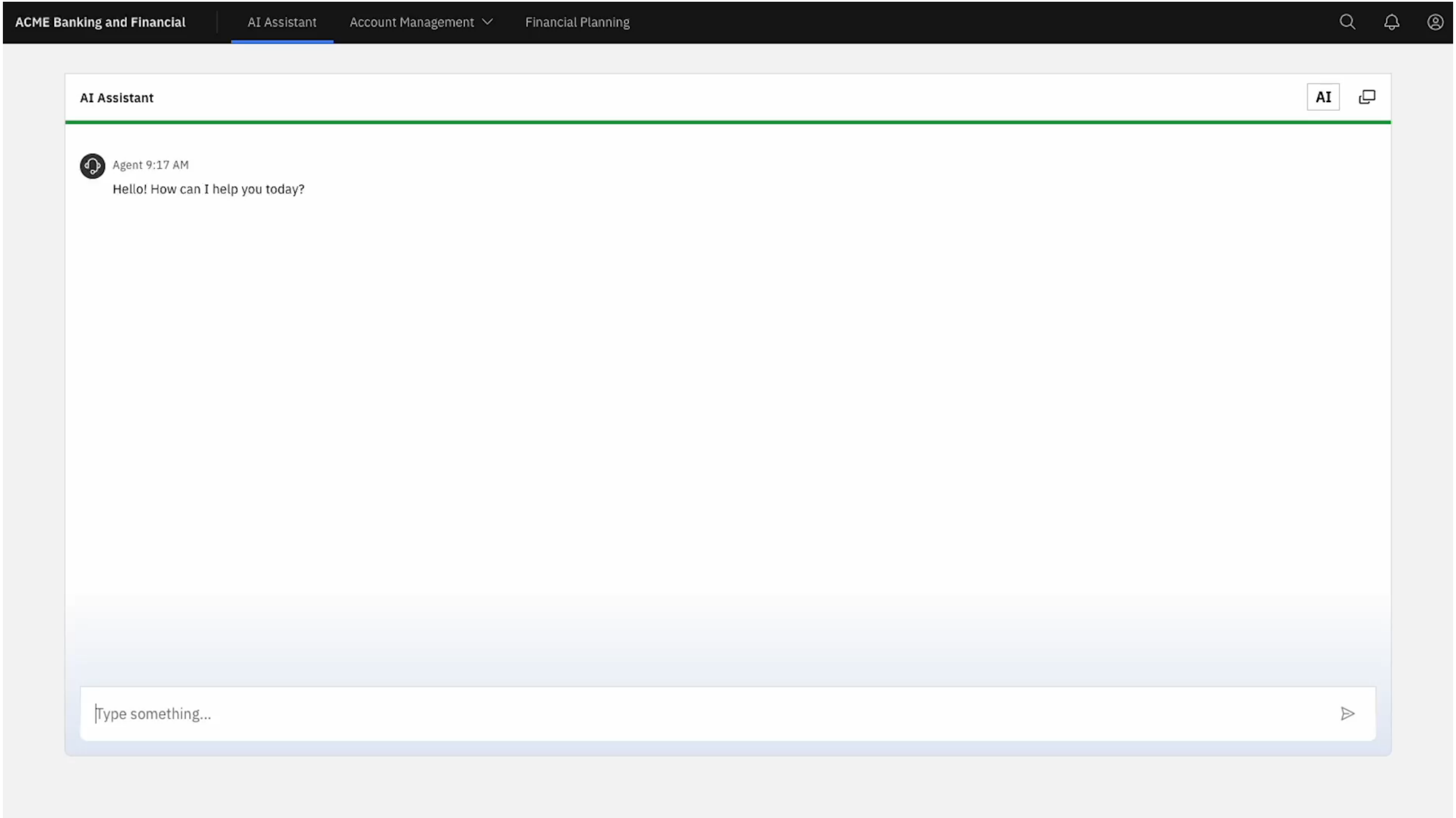
(ERP System

IBM i

PowerVM Hypervisor (Layer 4)

IBM Power10 (Layer 5)
→ E1050 1LPAR 6 Cores + 1 LPAR 6 Cores running Enterprise App

S.no	Layers
1	Application Layer
2	Data Processing layer
3	Inference (GenAI) Framework Layer
4	Platform Layer
5	Hardware Layer



Agent 9:17 AM

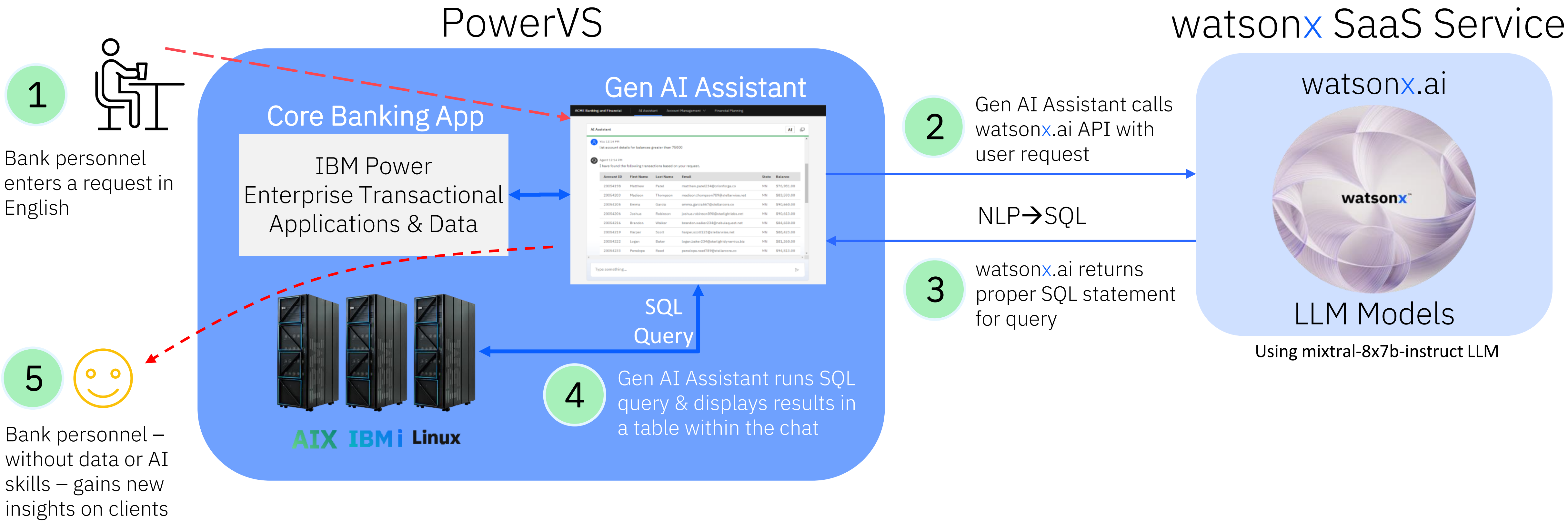
Hello! How can I help you today?

Demo Workflow – PowerVS + watsonx SaaS

Digital labor use cases:

- Bank fraud investigator using natural language to get fraud transaction details
- Bank marketing manager to find high value clients for an investment seminar

Pattern 2:
Using Gen AI to query information from IBM Power to gain insights easily.



PowerVS

watsonx SaaS Service

1 Bank personnel enters a request in English

Core Banking App

IBM Power Enterprise Transactional Applications & Data

Gen AI Assistant

You 12:14 PM
list account details for balances greater than 75000

Agent 12:14 PM
I have found the following transactions based on your request.

Account ID	First Name	Last Name	Email	State	Balance
20054198	Matthew	Patel	matthew.patel234@orionforge.co	MN	$76,981.00
20054203	Madison	Thompson	madison.thompson789@stellarwise.net	MN	$83,593.00
20054205	Emma	Garcia	emma.garcia567@stellarcore.co	MN	$90,660.00
20054206	Joshua	Robinson	joshua.robinson890@starlightlabs.net	MN	$90,613.00
20054216	Brandon	Walker	brandon.walker234@nebulaquest.net	MN	$84,650.00
20054219	Harper	Scott	harper.scott123@stellarwise.net	MN	$88,423.00
20054222	Logan	Baker	logan.baker234@starlightdynamics.biz	MN	$81,260.00
20054233	Penelope	Reed	penelope.reed789@stellarcore.co	MN	$94,513.00

2 Gen AI Assistant calls watsonx.ai API with user request

NLP→SQL

3 watsonx.ai returns proper SQL statement for query

watsonx.ai

LLM Models

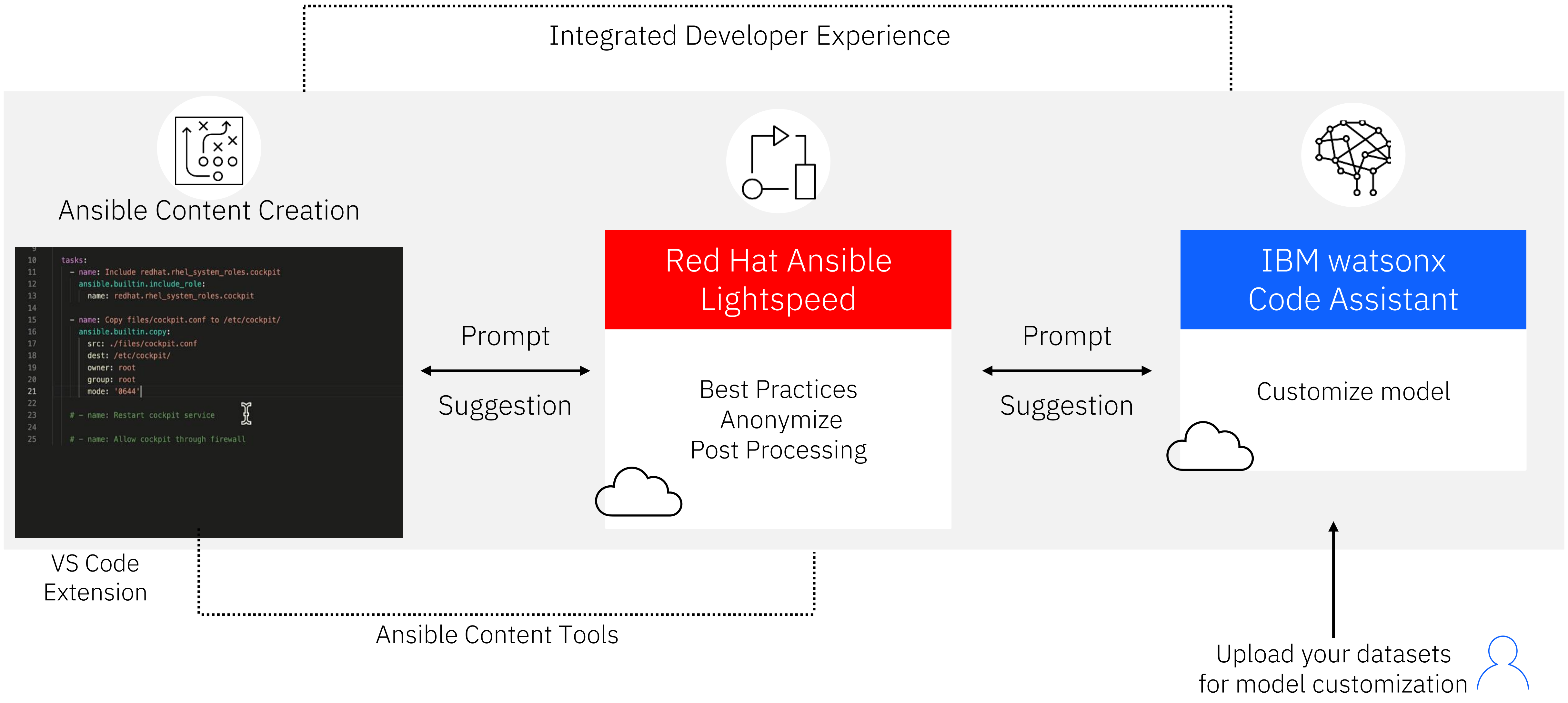
Using mixtral-8x7b-instruct LLM

SQL Query

4 Gen AI Assistant runs SQL query & displays results in a table within the chat

5 Bank personnel – without data or AI skills – gains new insights on clients

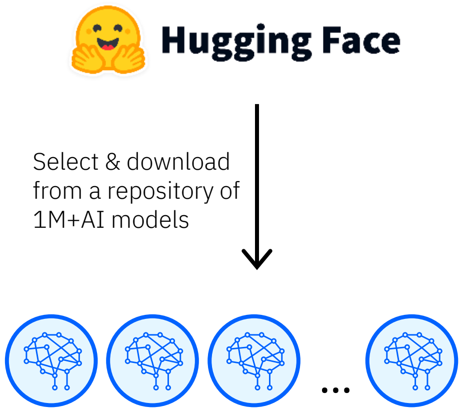
Automation develop Ansible playbook with Lightspeed and watsonx code assistant Power can be integrated and being “piloted”



```
  tasks:
    - name: Include redhat.rhel_system_roles.cockpit
      ansible.builtin.include_role:
        name: redhat.rhel_system_roles.cockpit

    - name: Copy files/cockpit.conf to /etc/cockpit/
      ansible.builtin.copy:
        src: ./files/cockpit.conf
        dest: /etc/cockpit/
        owner: root
        group: root
        mode: '0644'

    # - name: Restart cockpit service

    # - name: Allow cockpit through firewall
```

Get started with **AI** and **watsonx** with **IBM Power**

Deploy & manage **foundation models** securely.



Deploy and manage foundation models on IBM Power

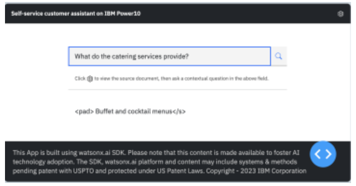
Leverage best-of-breed open-source models and software technologies to build a scalable end-to-end AI workflow

- Q&A Chatbot
- Email Summarization
- Entity Extraction

Embed **foundation models** into apps using the **watsonx.ai** SDK.

Deploy anywhere (IBM Power, x86, cloud)

Embed into apps with watsonx.ai SDK


Self-service customer assistant on IBM Power10

What do the catering services provide?

Click ⚙ to view the source document, then ask a contextual question in the above field.

<pad> Buffet and cocktail menus</s>

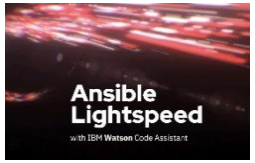
This App is built using watsonx.ai SDK. Please note that this content is made available to foster AI technology adoption. The SDK, watsonx.ai platform and content may include systems & methods pending patent with USPTO and protected under US Patent Laws. Copyright - 2023 IBM Corporation

Application on IBM Power
(Linux | AIX | IBM i)

Embed AI quickly, in a secure and resilient environment, close to your mission critical data and transactions

- Report generation
- Citizen services
- Knowledge management

Consume **watsonx** services from customized ecosystem apps.



Generate Ansible playbooks for IBM i & AIX.

deci. ELINAR

Get IBM Power10-optimized foundation models.

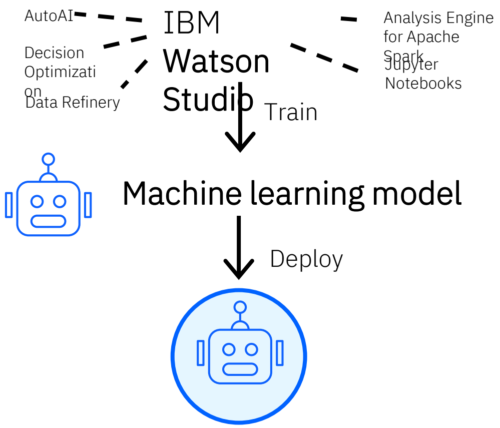
SAP

Use watsonx-embedded SAP applications with IBM Power.

Deliver new services faster using generative AI capabilities embedded in familiar ecosystem apps

- Asset management
- Code generation
- Accounting automation

Train & deploy **ML models** within a single AI studio.



Deploy machine learning model on IBM Power10

Train, tune, and inference machine learning models with on-chip acceleration without purchasing GPUs

- Fraud detection
- Risk underwriting
- Demand forecasting

Two recent announcement for AI and IBM Power

Code Assistant for RPG for IBM i
https://www.ibm.com/docs/en/announcements/statement-direction-code-assistant-rpg

“Statement of direction

IBM intends to deliver a code assistant for RPG - a generative AI tool which helps developers of IBM i software understand existing RPG code, create new RPG function using natural language description, and automatically generate test cases for RPG code.”

IBM Spyre off-chip accelerator on Power platform
https://www.ibm.com/docs/en/announcements/statement-direction-spyre-accelerator-power-platform

“Statement of direction

IBM intends to incorporate the IBM Spyre accelerator in future Power offerings to provide additional AI compute capabilities. Working together, IBM Power processors and IBM Spyre accelerator will enable the next generation infrastructure to scale demanding AI workloads for businesses.”

Disclaimer: Statements by IBM regarding its plans, directions, and intent are subject to change or withdrawal without notice at the sole discretion of IBM. Information regarding potential future products is intended to outline general product direction and should not be relied on in making a purchasing decision. The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code, or functionality. Information about potential future products may not be incorporated into any contract. The development, release, and timing of any future features or functionality described for IBM products remain at the sole discretion of IBM.

IBM Spyre Accelerator

The IBM Spyre Accelerator is a purpose-built enterprise-grade accelerator offering scalable capabilities for complex AI models and generative AI use cases. The new accelerator features 32 individual accelerator cores onboard, and each Spyre is mounted on a PCIe card.

Jointly designed by IBM Research and IBM Infrastructure, Spyre's architecture is designed for more efficient AI computation. Notably, the chip will send data directly from one compute engine to the next, leading to an efficient use of energy. This family of processors also uses a range of lower precision numeric formats (such as int4 and int8), to make running an AI model more energy efficient and far less memory intensive.

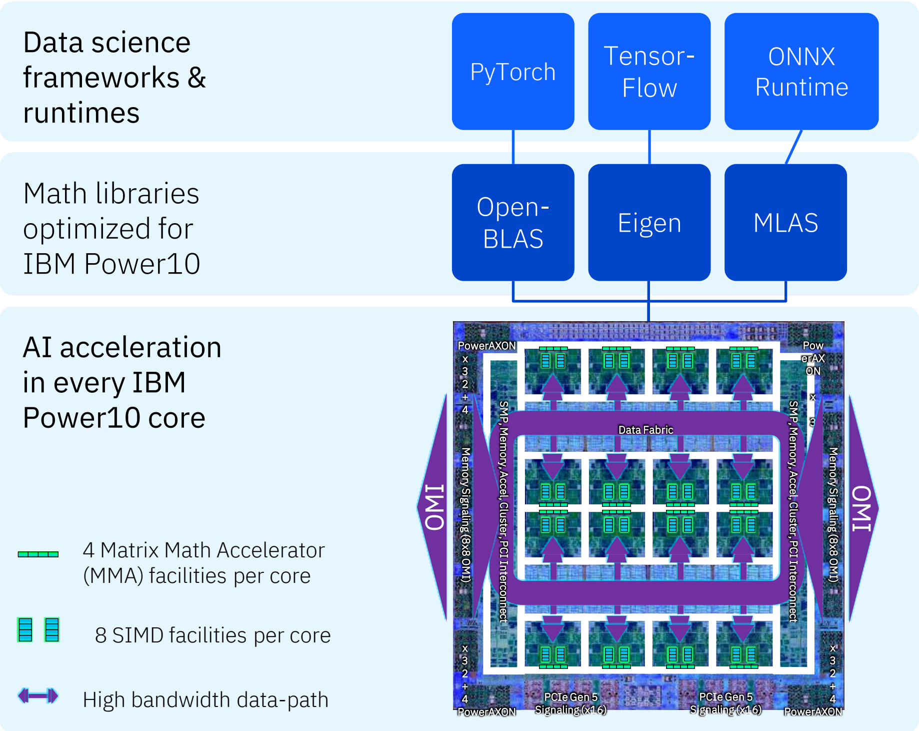
More details on our plans for the IBM Spyre Accelerator will be revealed in 2025.

https://www.ibm.com/docs/en/announcements/statement-direction-spyre-accelerator-power-platform

https://newsroom.ibm.com/blog-ibm-power-modernizes-infrastructure-and-accelerates-innovation-with-ai-in-the-year-ahead

© 2023 IBM Corporation

Accelerate AI Efficiently with AI Optimized Hardware



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) facilities per core

8 SIMD facilities per core

High bandwidth data-path

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

Each core has four MMA (Matrix Math Accelerator) facilities to accelerate matrix calculations that are used in many common AI workloads

Power10 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)	Limited Benefit (cost-perf)	Best Fit (cost-perf)
SW Maintenance		Need to update GPU specific SW (CUDA, cuDNN, etc.)	Maintained by IBM / Partner

GPUs or Power10 w/MMA*

*Please see speaker notes for details

What is RocketCE and Rocket AI Hub

RocketCE is *a distribution of over 200 Power-optimized packages* for AI, such as *TensorFlow, PyTorch, and Python.*

1. **OpenCE**: an AI open-source community championed by IBM for simplifying the build of 200+ Power-optimized packages for AI, including build script and actual community builds available via CONDA..

2. RocketCE
RocketCE leverages the base OpenCE packages and adds the P10 MMA specific libraries to create packages that are optimized for Power 10 (Tensorflow, PyTorch, Python,etc). RocketCE packages can be consumed via Conda or Red Hat Openshift Containers (UBI Images).

What is RocketAI Hub

Enhancement to the RocketCE repository by providing the Rocket P10 MMA SW in containers and adding additional tooling like Kubeflow that helps data science teams automate end-to-end data science workflows.

*** RocketCE and RocketAI Hub is available with Enterprise support from Rocket*

How does it help?

1. Allows Data Scientists on the customer or ISV side to use the same tools on Power for AI that they might already be familiar with on the non-Power platforms.
2. Make it easy to move AI code built on a non-Power platform onto Power if there is version compatibility. For ex: If the code is built using Tensorflow v2.x on Intel(x86), we can leverage the Tensorflow v2.x from RocketCE to run the same code on Power without making changes to the Python code.
3. RocketCE packages are specifically enabled to leverage the MMA on P10.

How does a customer access it?

To access these AI tools, visit the RocketCE channel on Anaconda and for Container images see Quay repository.

Reference Links:

RocketCE Announcement:
Announcement Letter Link

RocketAI Hub Announcement:
Announcement Letter Link

FIND Open Source Application availability on IBM Power link:
Find open source packages built for IBM Power

FIND ISV support on IBM Power link:

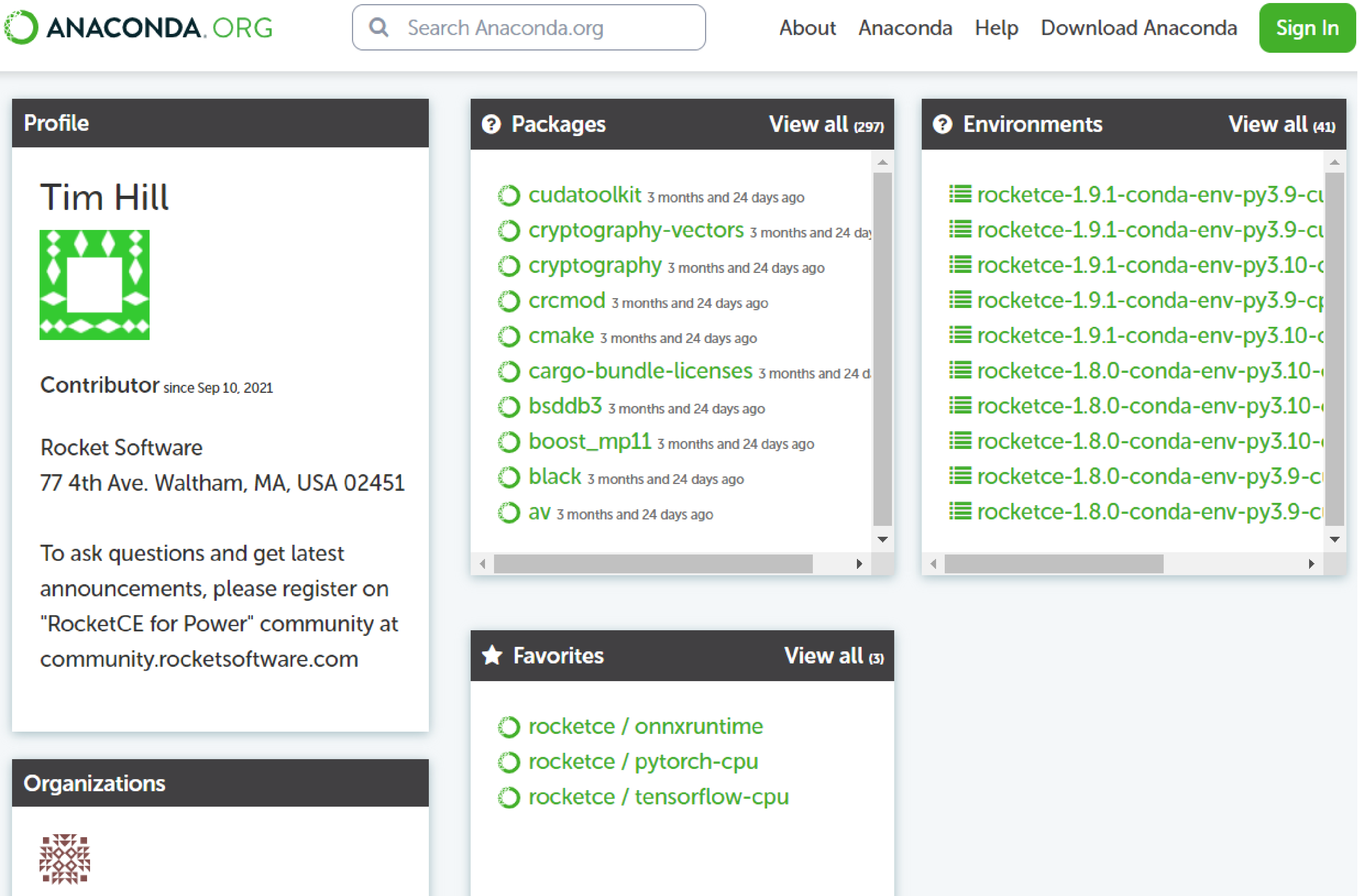
https://www.ibm.com/power/resources/isv/

RocketCE is prebuilt images, for Power

“Builds with Enterprise Production Support

Rocket Software hosts pre-built versions of the Open-CE at conda channel here. This channel provides packages for Power architecture(ppc64le).”

Enterprise Support available (more on that in a moment), but you can use these without support without charge.



ANACONDA.ORG

Search Anaconda.org

About Anaconda Help Download Anaconda Sign In

Profile

Tim Hill

Contributor since Sep 10, 2021

Rocket Software
77 4th Ave. Waltham, MA, USA 02451

To ask questions and get latest announcements, please register on "RocketCE for Power" community at community.rocketsoftware.com

Organizations

Packages View all (297)

- cudatoolkit 3 months and 24 days ago
- cryptography-vectors 3 months and 24 day
- cryptography 3 months and 24 days ago
- crcmod 3 months and 24 days ago
- cmake 3 months and 24 days ago
- cargo-bundle-licenses 3 months and 24 d
- bsddb3 3 months and 24 days ago
- boost_mp11 3 months and 24 days ago
- black 3 months and 24 days ago
- av 3 months and 24 days ago

Environments View all (41)

- rocketce-1.9.1-conda-env-py3.9-cu
- rocketce-1.9.1-conda-env-py3.9-cu
- rocketce-1.9.1-conda-env-py3.10-c
- rocketce-1.9.1-conda-env-py3.9-cp
- rocketce-1.9.1-conda-env-py3.10-c
- rocketce-1.8.0-conda-env-py3.10-
- rocketce-1.8.0-conda-env-py3.10-
- rocketce-1.8.0-conda-env-py3.10-
- rocketce-1.8.0-conda-env-py3.9-c
- rocketce-1.8.0-conda-env-py3.9-c

Favorites View all (3)

- rocketce / onnxruntime
- rocketce / pytorch-cpu
- rocketce / tensorflow-cpu

Credit to Marvin Gießing

This is Marvin Gießing, who used to work in IBM Power, but is now part of our IBM Client Engineering team.

Marvin wrote this article, as a demo at the IBM TechXchange event in Barcelona earlier this year: https://github.com/mgiessing/bcn-lab-2084

Another member of the IBM Power Global team, Ashwin Srinivas, then joined a group of us in Prague and took us through that lab.

I then went through it again later and have modified it slightly to work with IBM's Techzone environments. IBMers and IBM Business Partners can therefore follow my steps using the forked copy of Marvin's work I have here: https://github.com/DSpurway/RAG-with-Notebook

Customers can work with their IBMers and BPs to see this too!

Offering Priorities & Roadmap

For Sales, BPs & Clients under NDA Use Only

Currently Available
On-track

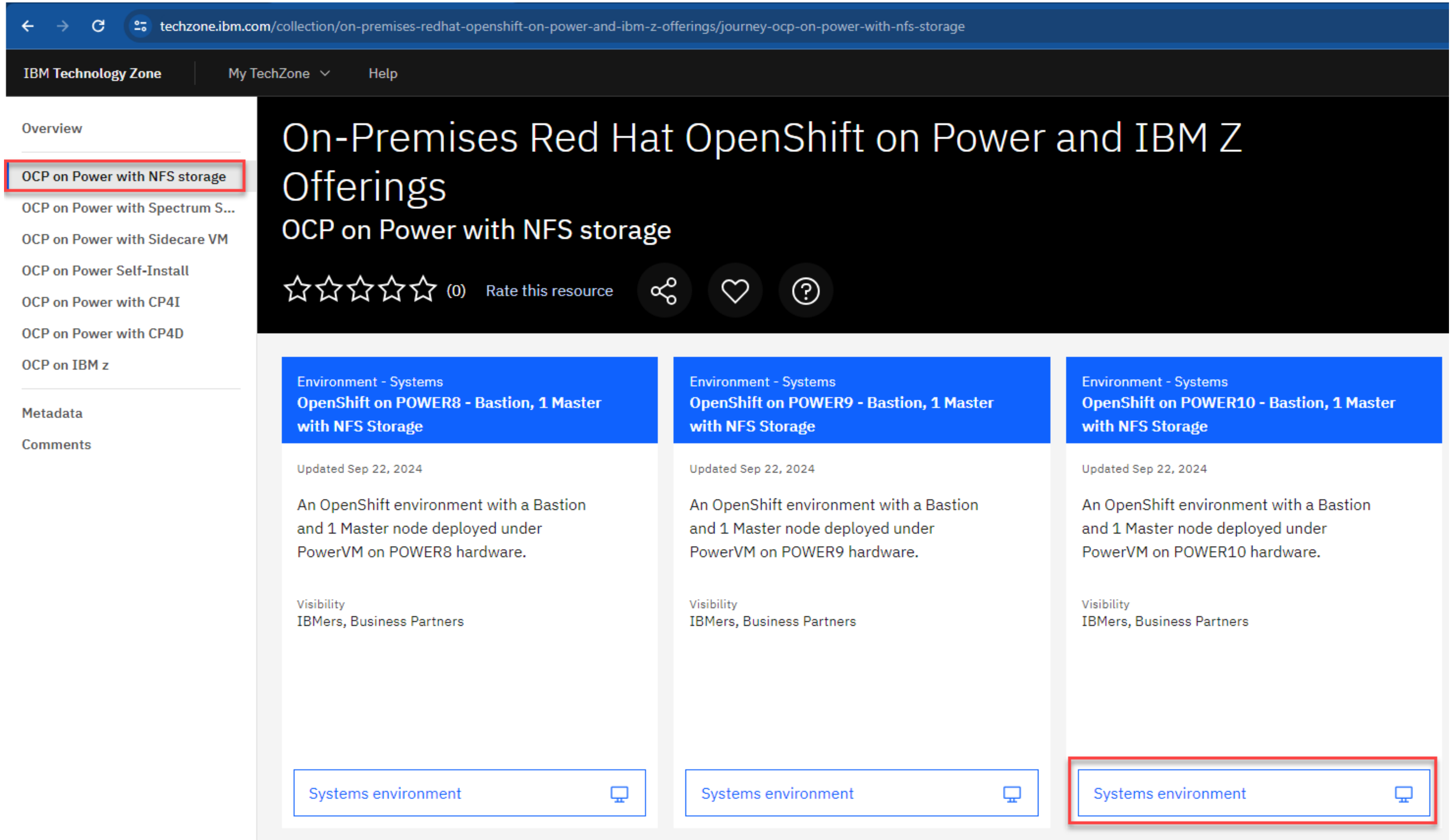
Uncommitted : Working it

	1H24	2H24	1H25	2H25	2025+
Scalable AI-ready Infrastructure					
- HW acceleration				P11 On-Chip Acceleration	Off-Chip Acceleration (AIU)
- Stack optimization & AI studio (FMs, SDKs, deployment services, etc.)	RocketAI Hub, RocketCE, CP4D *(WSL, WML, DB2W, AE4S, etc.)*	CP4D 5.0 (R-Studio) Open Data Hub	CP4D (IKC, DataStage Watson Pipelines) OpenShift AI	CP4D (Open Scale) Watsonx.ai	
- Simplified consumption				P11 AI Solution Environment	
Optimized e2e Hybrid Workflow					
- On-prem deployments		Fusion HCI (training, tuning) with IBM Power (inference)			
- Cloud deployments		watsonx.data (run with Power; connect to Power DBs)			
- Data democratization & governance		PowerVS + watsonx (toolkit for additional top GenAI use cases)			
AI infused ecosystem					
- Code Assistants	Ansible Lightspeed	RPG Code Assistant			
- Data management		MilvusDB (RocketCE)			
- Model Services	ElinarAI Deci	Sway AI			
- Core ISV apps	Equitus Vision Analytics Trovares	Infor			

IBM Techzone delivers “Show Not Tell” with IBM Technology

“Technology Zone is the single destination for our go-to-market teams and IBM business partners ecosystem to access on-demand and live environments to learn, build, show, and share the value of IBM solutions. Additionally, they can extend our certified base images and further customize them for test, education, demonstration, and pilot activities.”

So, we start here: https://techzone.ibm.com/collection/on-premises-redhat-openshift-on-power-and-ibm-z-offerings/journey-ocp-on-power-with-nfs-storage



techzone.ibm.com/collection/on-premises-redhat-openshift-on-power-and-ibm-z-offerings/journey-ocp-on-power-with-nfs-storage

IBM Technology Zone | My TechZone | Help

- Overview
- OCP on Power with NFS storage
- OCP on Power with Spectrum S...
- OCP on Power with Sidecare VM
- OCP on Power Self-Install
- OCP on Power with CP4I
- OCP on Power with CP4D
- OCP on IBM z
- Metadata
- Comments

On-Premises Red Hat OpenShift on Power and IBM Z Offerings

OCP on Power with NFS storage

☆☆☆☆☆ (0) Rate this resource

Environment - Systems
OpenShift on POWER8 - Bastion, 1 Master with NFS Storage

Updated Sep 22, 2024

An OpenShift environment with a Bastion and 1 Master node deployed under PowerVM on POWER8 hardware.

Visibility
IBMers, Business Partners

Systems environment

Environment - Systems
OpenShift on POWER9 - Bastion, 1 Master with NFS Storage

Updated Sep 22, 2024

An OpenShift environment with a Bastion and 1 Master node deployed under PowerVM on POWER9 hardware.

Visibility
IBMers, Business Partners

Systems environment

Environment - Systems
OpenShift on POWER10 - Bastion, 1 Master with NFS Storage

Updated Sep 22, 2024

An OpenShift environment with a Bastion and 1 Master node deployed under PowerVM on POWER10 hardware.

Visibility
IBMers, Business Partners

Systems environment