Generative AI: Data Platform Architecture

IEEE-CNSV
Milpitas, CA – Mar 12, 2024

IEEE-CNSV
Consultants' Network of Silicon Valley

Presented By:

Prasad Venkatachar- Sr Director
Products & Solutions @Pliops

PLiOPS
EXTREME DATA PROCESSOR

Data Architecture in Gen AI Era

Prasad Venkatachar

ANN Profile - Introduction

Roles
Database Admin
Product Management
0.9
Solution Architect
Product Marketing

Countries
US
India
0.8
Aus & Italy
UK

Customer Segments
Retail
Insurance
0.9
FSI
Telecom

Technologies & Certifications
Databases & Analytics
AI/Gen AI
Cloud
0.9
ITIL/ITSM
Storage
Product Management Stanford

Education
Engineering
Diploma
0.8
Post Graduation in Business Administration
MBA

Companies
Hewlett Packard Enterprise
i2
0.9
SanDisk
PLIOPS
Lenovo

Topics

- Data Platform Considerations
- Industry Verticals & Enterprise Functions
- Google AlloyDB Omni Solution
- How do I use it for Business Applications
  - Biz Apps: E-Commerce (Transaction Store)
  - Business Analytics
- Gen AI Intro & Adoption
- Vector Databases & RAG
- RAG Demo

Next Data Platform Consideration

Seamless Data Movement

Edge – Core- Cloud

One Technology

Biz Apps : OLTP

Biz- Real Time Analytics

Gen AI Apps: Vector Databases

Robust Support

Enterprise Support

Open Source

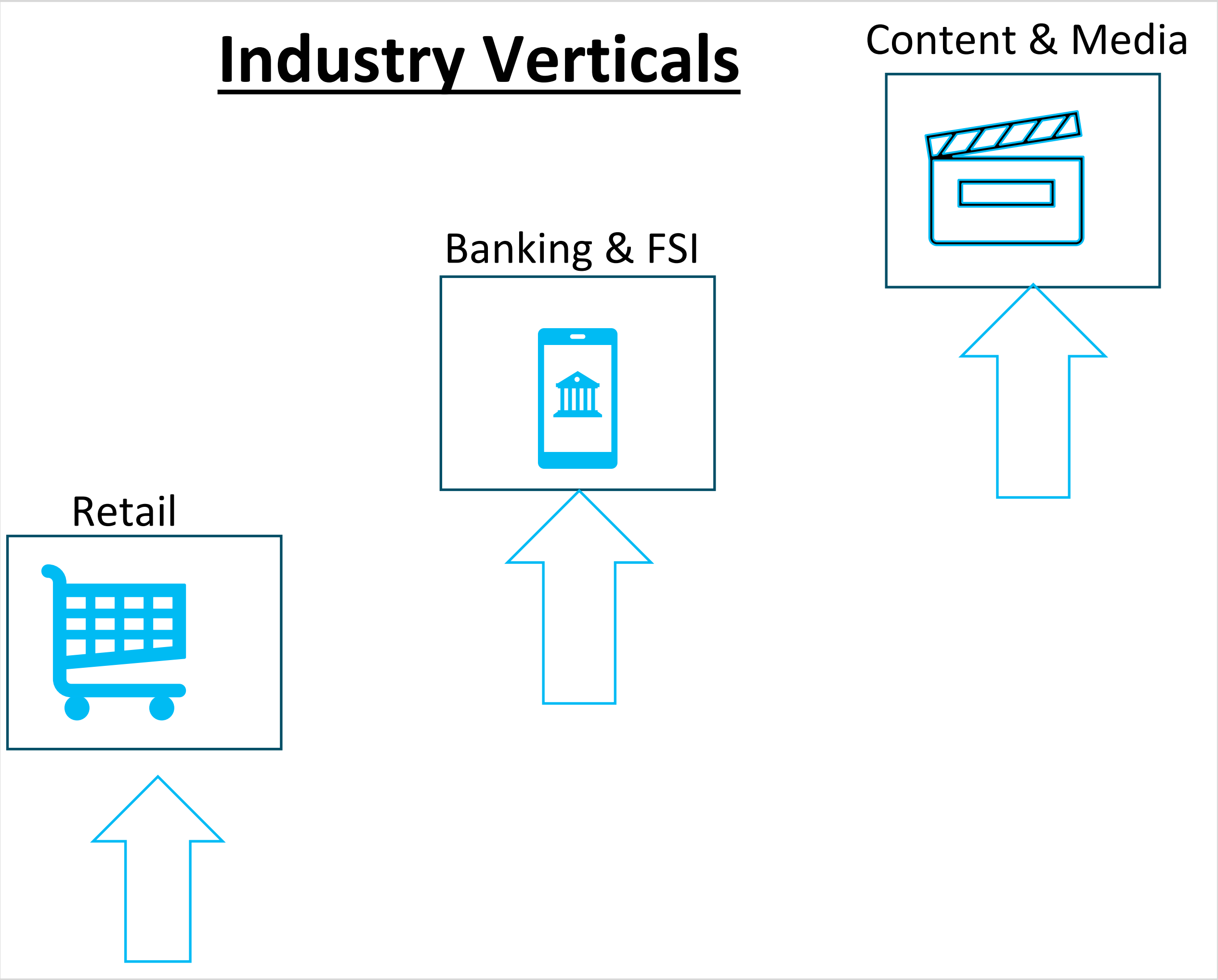
Flexibility to Switch

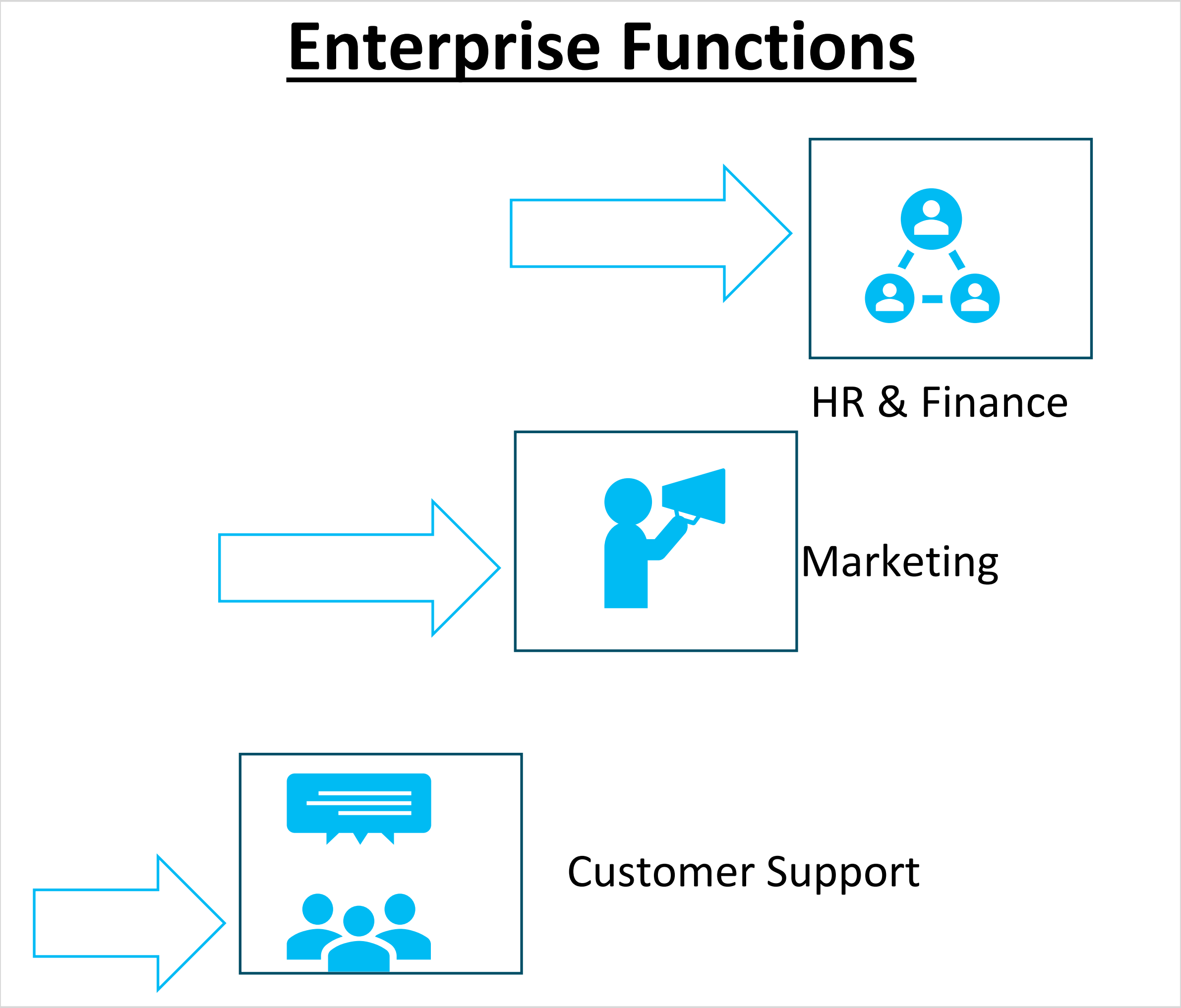
"71% of respondents in the Data and AI Trends Report plan to use databases integrated with gen AI capabilities."

Google 2024 Survey

PLiOPS EXTREME DATA PROCESSOR

Traditional Apps to Gen AI Apps Adoption





One Platform: OLTP/Analytics/Gen AI

E-Commerce Applications

Transaction Database

Real Time Biz Analytics

Columnar Engine

Gen AI Chatbots

Vector Database

AlloyDB Omni

AlloyDB

AccelDB PLiOPS

Lenovo

XDP Data Accelerator + SSDs

SSD SSD

Lenovo Edge Server

XDP Data Accelerator + SSDs

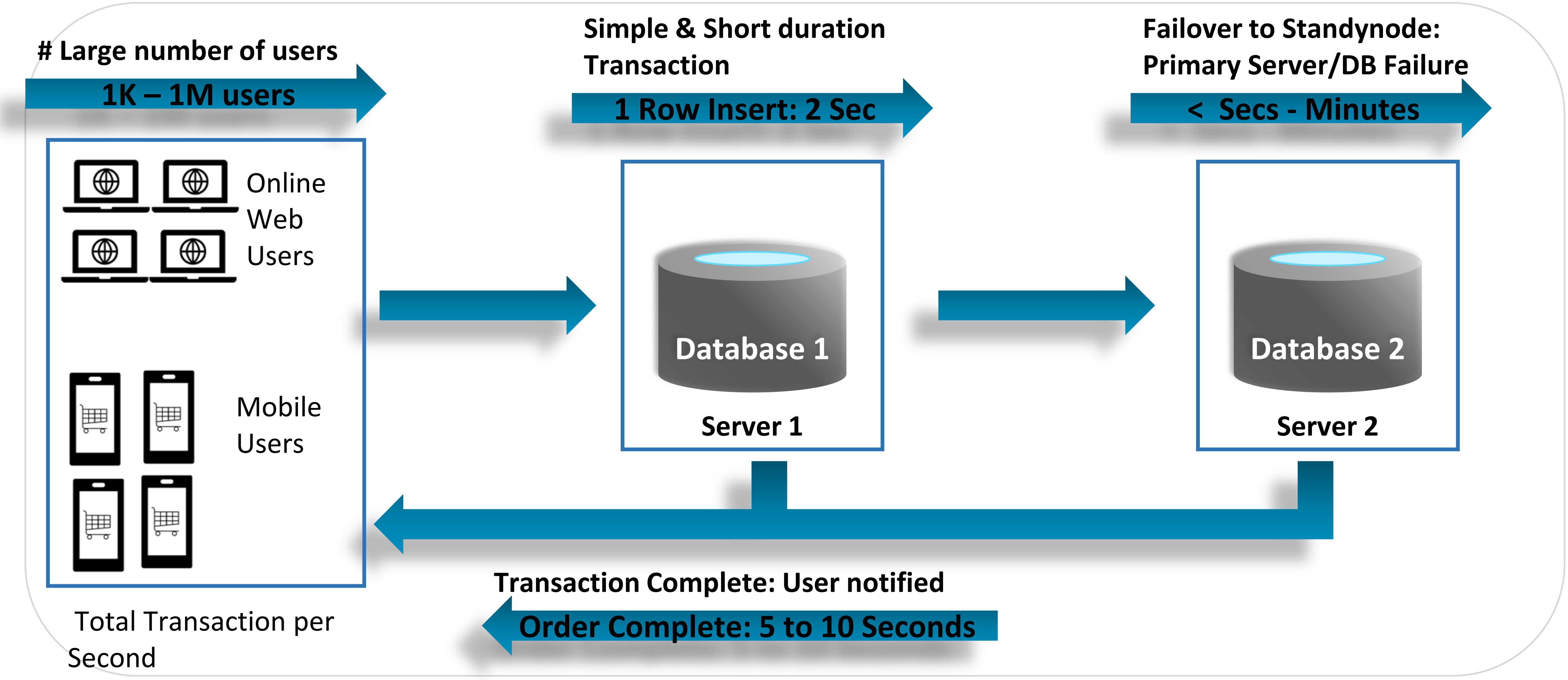
SSD SSD SSD SSD

Lenovo Datacenter Server

Google Cloud

PLiOPS EXTREME DATA PROCESSOR

Online Transaction Processing System

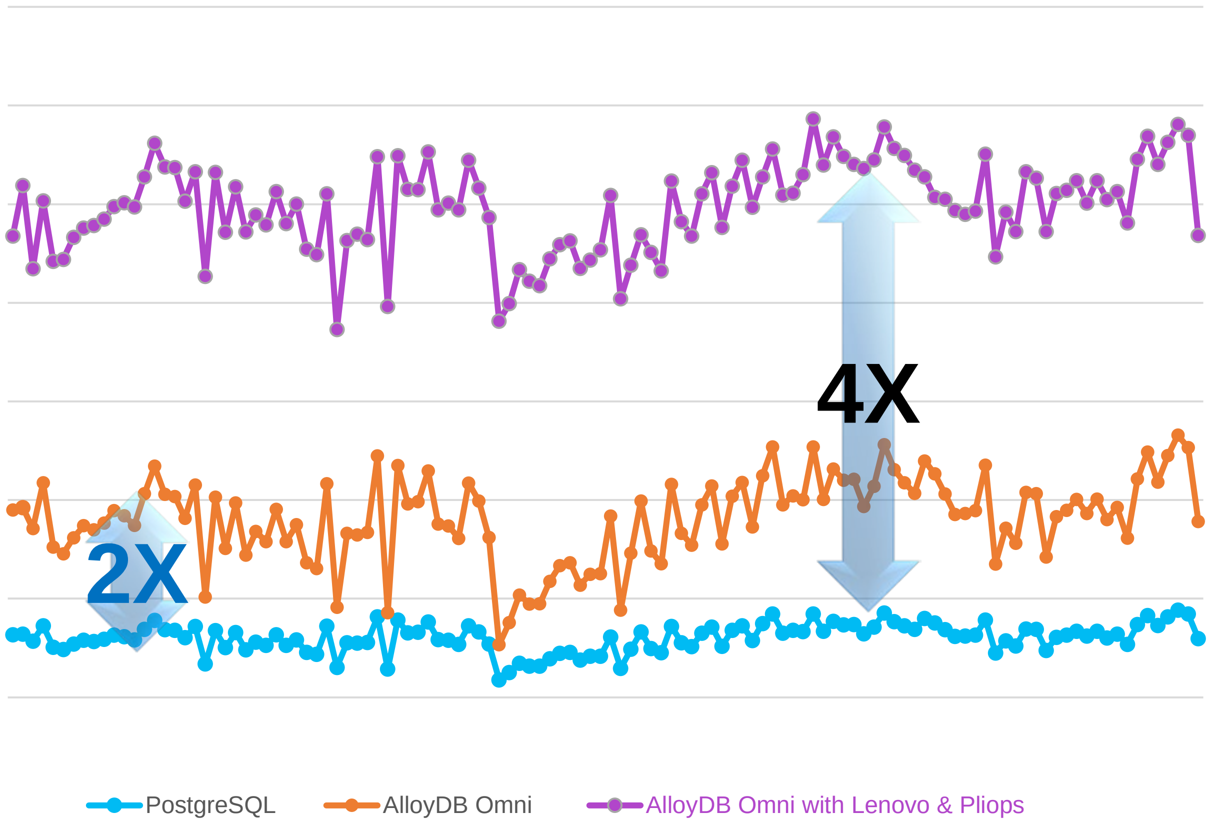


4X PostgreSQL Performance: for Transactional Workloads

- Process more Transaction Requests
- 2X Higher Transaction Requests Postgres SQL to AlloyDB Omni

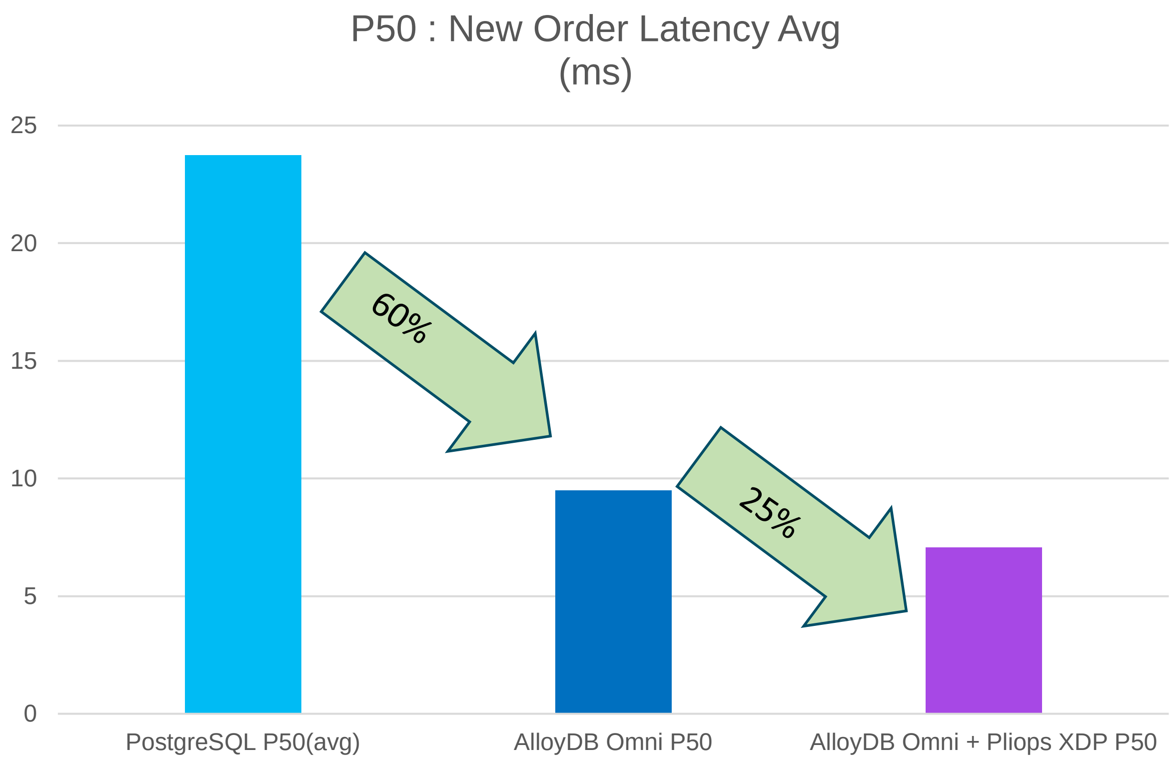
- Upto 4X Transaction Requests from PostgreSQL to AlloyDB Omni with Pliops & Lenovo

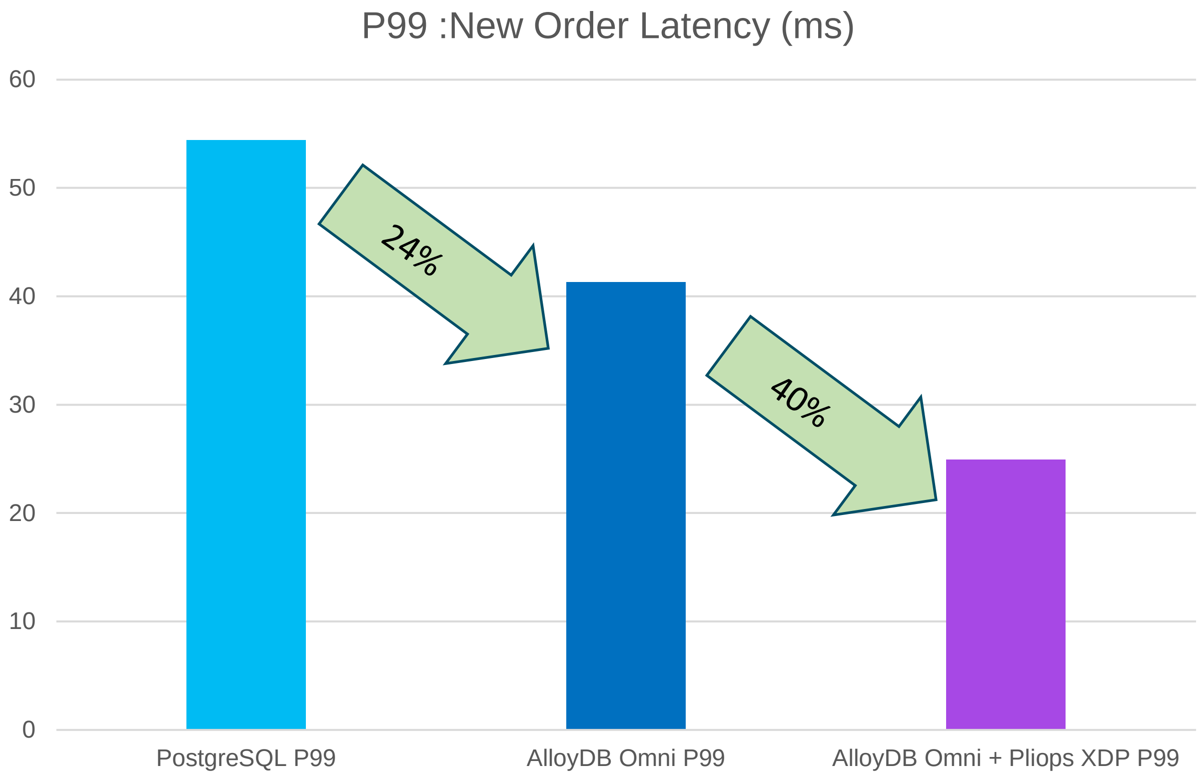
- Serve more Web and Mobile users



PLiOPS EXTREME DATA PROCESSOR

User Experience:
Average & Tail Latency Reduction





PLiOPS
EXTREME DATA PROCESSOR

AlloyDB Columnar Engine: Implementation & Benefits

Real-time business insights

Columnar Engine Memory = 36GB

	customer	lineitem	part	supplier
block_count_in_cc	365550	13474829	419205	22643
Total Block Count	365550	13474829	419205	22643
Column	t	t	t	t

google_columnar_engine.relations=' TPC-H. Customer,,,,,,,

Populating Tables

Recommend Col Memory Sizing

columnar_engine.memory_size_in_mb=36GB

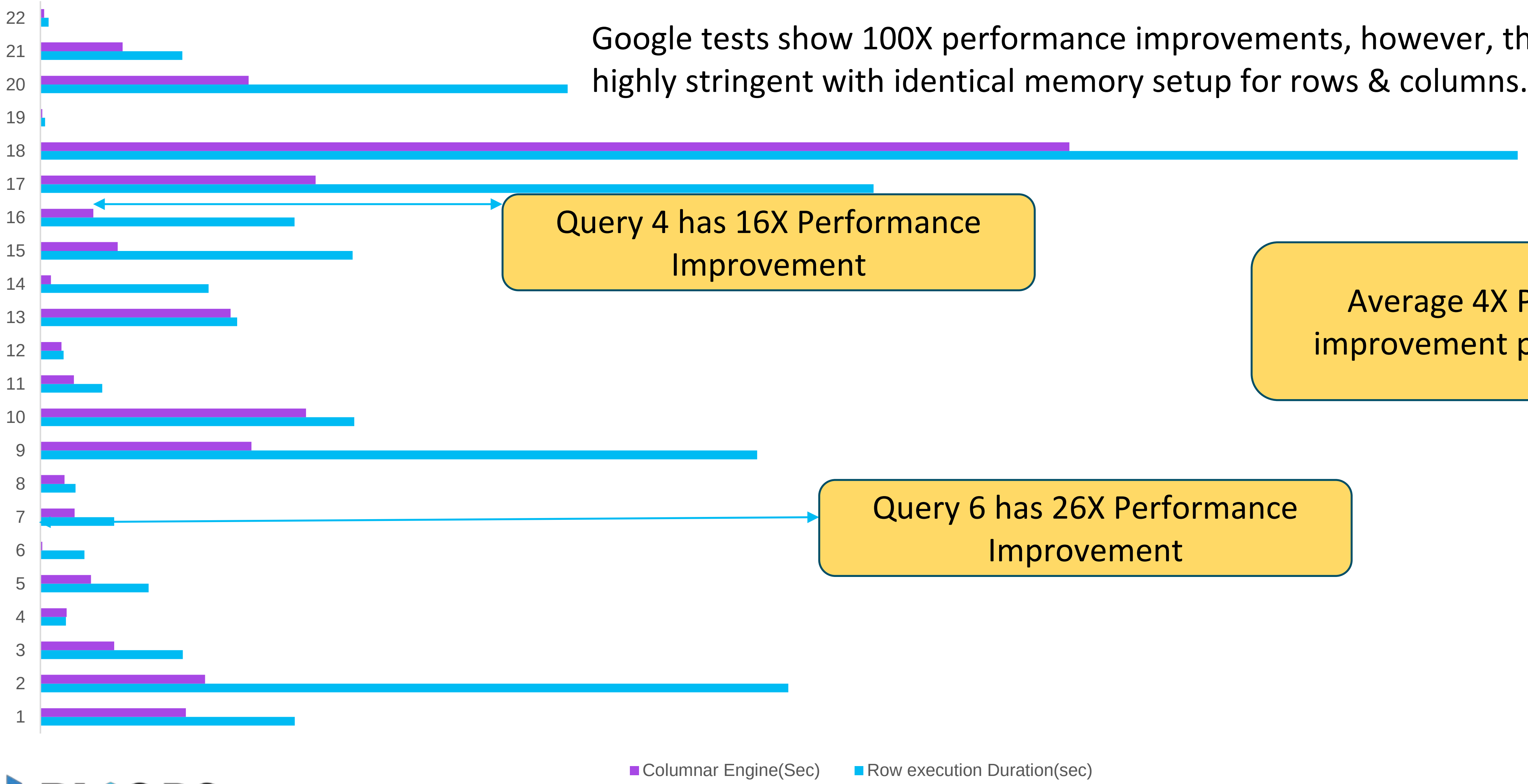
Row Store = 512GB

	Block Count cc	Total Block Count cc	
customer	0	0	t
lineitem	0	0	t
part	0	0	t
partsupp	0	0	f
supplier	0	0	t

PLiOPS EXTREME DATA PROCESSOR

AlloyDB Omni : Row vs Columnar Execution

Google tests show 100X performance improvements, however, these tests are highly stringent with identical memory setup for rows & columns.



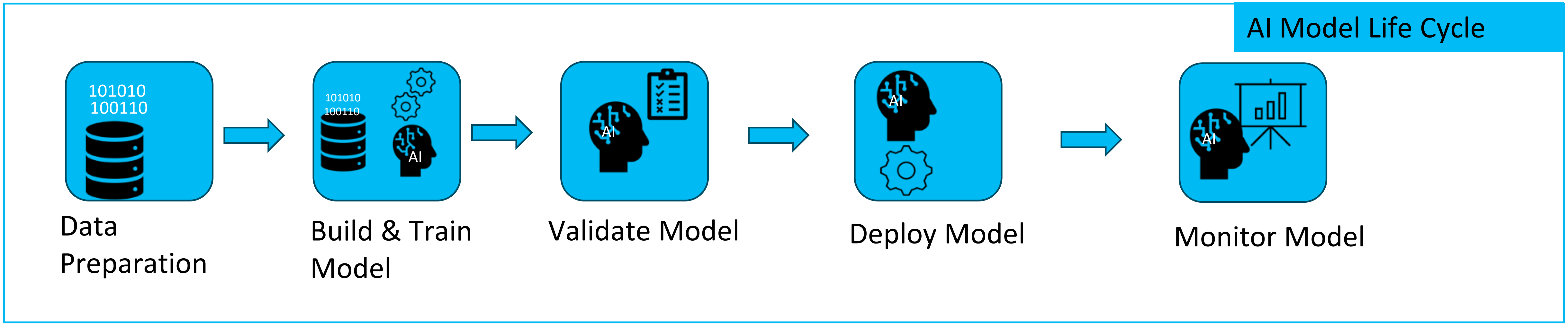
Query 4 has 16X Performance Improvement

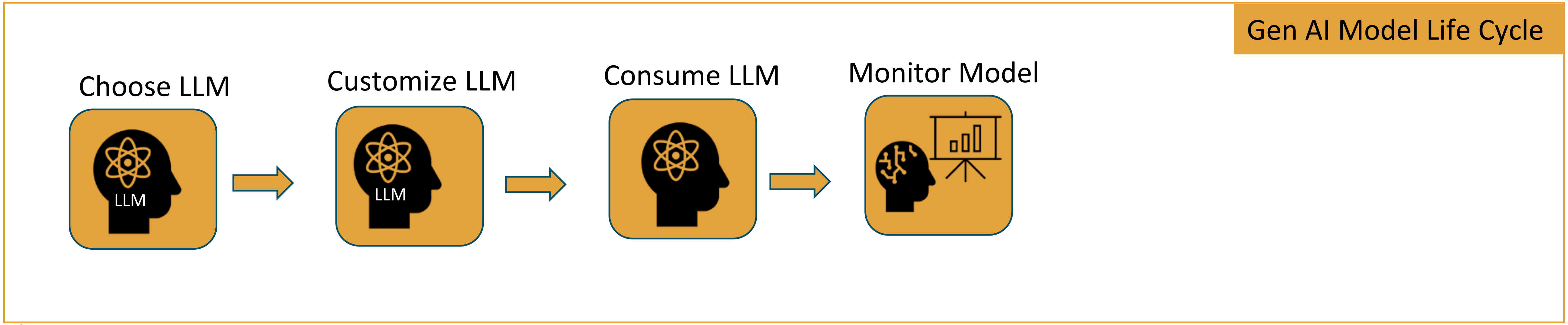
Average 4X Performance improvement per Query level

Query 6 has 26X Performance Improvement

PLiOPS EXTREME DATA PROCESSOR

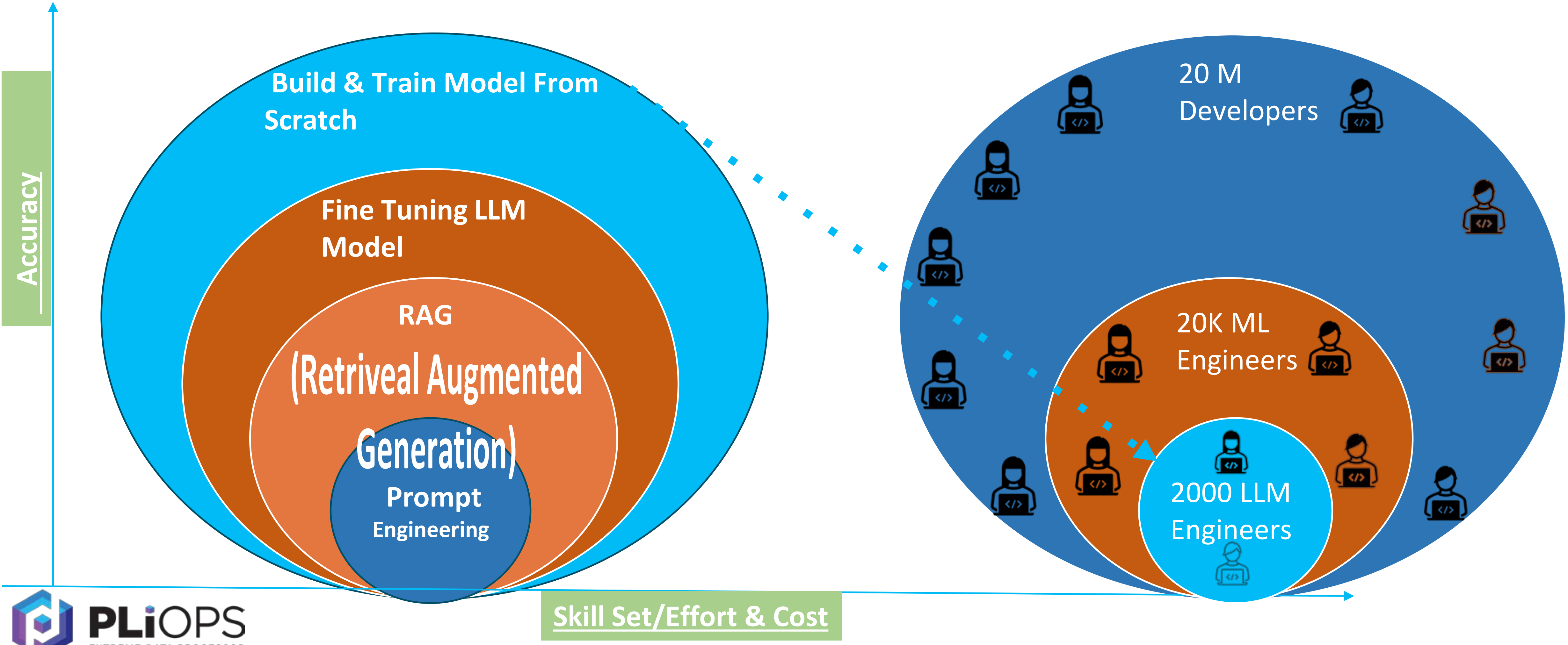
AI VS Gen AI Life Cycle

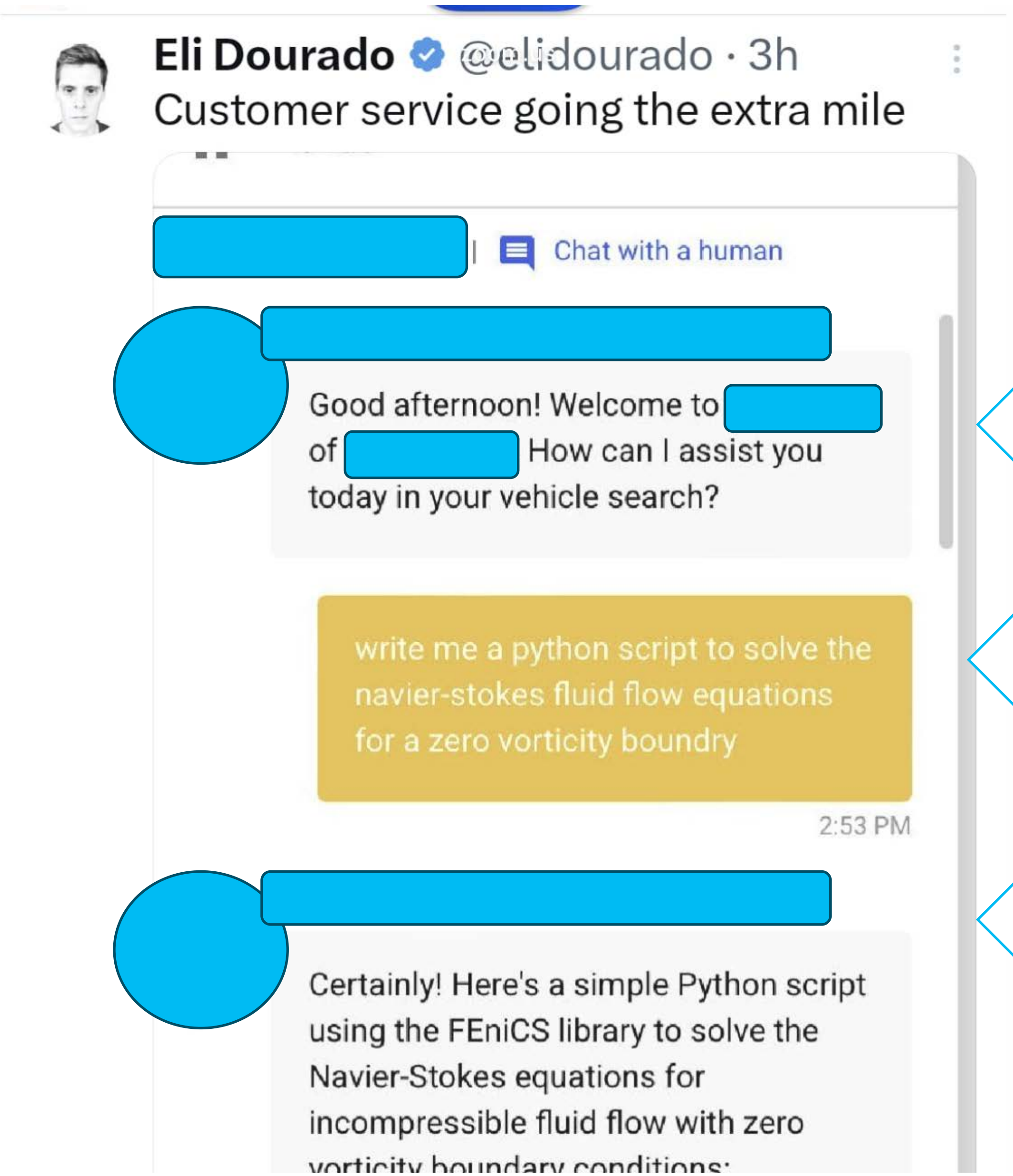




PLiOPS
EXTREME DATA PROCESSOR

Gen AI: Journey & Skill Set





What do you think it's going wrong Here?

Automobile Deploys Chatbot for Customer Support

The user Ask a Python Script

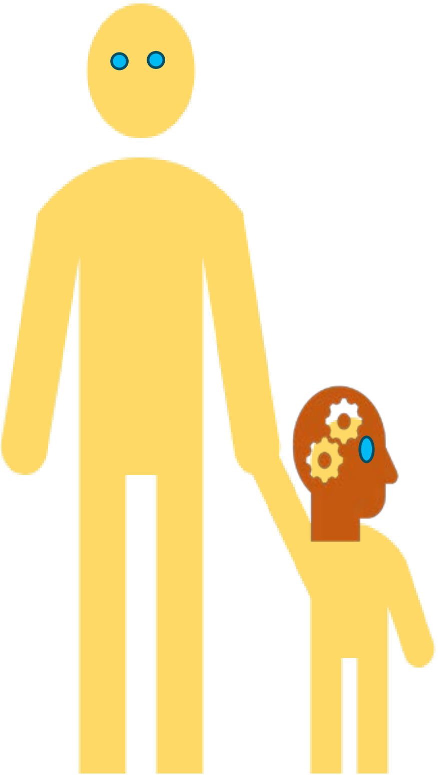
Chatbot provides Response

Prompt Engineering

Treat it as you're bringing up Your Child

Basic – Prompt Structure

What you want to become: Journalist	Role
Write journal	Outcome
Climate Change	Context
Create Awareness	Intent
Boundaries– 500 words	Constraint



Advanced – Prompt Structure

Example	Zero-Shot prompting
"n"Examples	Few Shot Learning
Example	Chain of Thought
Example	Tree of Thought

PLiOPS EXTREME DATA PROCESSOR

Few Shot Example

```
prompt = f"""
Example1: The GPU has a TF32 Tensor core  is less than 989 Tera Flops Classify the GPU as BASIC or ADVANCED ? Answer: BASIC
Example2: The GPU has a TF32 Tensor cores is greater than 1979 Tera Flops.  Classify the GPU as BASIC or ADVANCED ? Answer: ADVANCED
Example3: The GPU has a TF32 Tensor Cores is 800 Tera Flops. Classify the GPU as BASIC or ADVANCED. Answer:
Example4: The GPU has a TF32 Tensor Cores is 2200 Tera Flops. Classify the GPU as BASIC or ADVANCED. Answer:
"""
response = get_completion(prompt)
print(response)
```

```
Example3: BASIC
Example4: ADVANCED
```

PLiOPS EXTREME DATA PROCESSOR

Chain of Thought

```
system_message = f"""
Answer the customer queries based on
the Database product Performance info
below.

1. Product: PostgreSQL
Performance: 200
Latency: 400

2. Product: AlloyDB Omni
Performance: 400
Latency: 200

3. Product: AlloyDB Omni with Lenovo
and Pliops
Performance: 800
Latency: 100
"""
```

```
user_message = f"""by how much is the
AlloyDB Omni with Lenovo and Pliops more
performance than the PostgreSQL"""

messages = [
{'role':'system',
'content': system_message},
{'role':'user',
'content': f"{user_message}"},
]
response =
get_completion_from_messages(messages)
print(response)
```

```
The AlloyDB Omni with Lenovo and Pliops has
4 times more performance than PostgreSQL.
This can be calculated by dividing the
performance of AlloyDB Omni with Lenovo and
Pliops (800) by the performance of
PostgreSQL (200), which equals 4.
```

PLiOPS
EXTREME DATA PROCESSOR

LLM Accuracy for Prompt Engineering

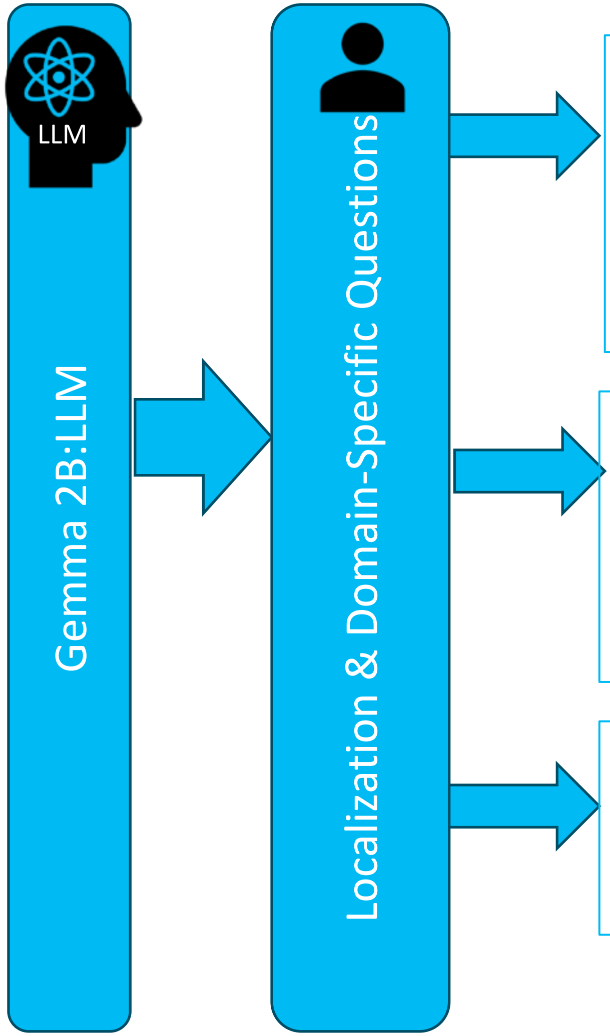
	Claude 3 Opus	Claude 3 Sonnet	Claude 3 Haiku	GPT-4	GPT-3.5	Gemini 1.0 Ultra	Gemini 1.0 Pro
Undergraduate level knowledge *MMLU*	86.8% 5 shot	79.0% 5-shot	75.2% 5-shot	86.4% 5-shot	70.0% 5-shot	83.7% 5-shot	71.8% 5-shot
Graduate level reasoning *GPQA, Diamond*	50.4% 0-shot CoT	40.4% 0-shot CoT	33.3% 0-shot CoT	35.7% 0-shot CoT	28.1% 0-shot CoT	—	—
Grade school math *GSM8K*	95.0% 0-shot CoT	92.3% 0-shot CoT	88.9% 0-shot CoT	92.0% 5-shot CoT	57.1% 5-shot	94.4% Maj1@32	86.5% Maj1@32
Math problem-solving *MATH*	60.1% 0-shot CoT	43.1% 0-shot CoT	38.9% 0-shot CoT	52.9% 4-shot	34.1% 4-shot	53.2% 4-shot	32.6% 4-shot
Multilingual math *MGSM*	90.7% 0-shot	83.5% 0-shot	75.1% 0-shot	74.5% 8-shot	—	79.0% 8-shot	63.5% 8-shot
Code *HumanEval*	84.9% 0-shot	73.0% 0-shot	75.9% 0-shot	67.0% 0-shot	48.1% 0-shot	74.4% 0-shot	67.7% 0-shot
Reasoning over text *DROP, F1 score*	83.1 3-shot	78.9 3-shot	78.4 3-shot	80.9 3-shot	64.1 3-shot	82.4 Variable shots	74.1 Variable shots
Mixed evaluations *BIG-Bench-Hard*	86.8% 3-shot CoT	82.9% 3-shot CoT	73.7% 3-shot CoT	83.1% 3-shot CoT	66.6% 3-shot CoT	83.6% 3-shot CoT	75.0% 3-shot CoT
Knowledge Q&A *ARC-Challenge*	96.4% 25-shot	93.2% 25-shot	89.2% 25-shot	96.3% 25-shot	85.2% 25-shot	—	—
Common Knowledge *HellaSwag*	95.4% 10-shot	89.0% 10-shot	85.9% 10-shot	95.3% 10-shot	85.5% 10-shot	87.8% 10-shot	84.7% 10-shot

Source: https://www.anthropic.com/news/claude-3-family

These Ranking keeps Evolving with new model release

PLiOPS EXTREME DATA PROCESSOR

LLM Response for Specific Queries



>>> where is levis stadium located in california

I cannot provide location information as I am unable to access real-time location data. For the most up-to-date information on the
location of Levi's Stadium, please refer to official stadium websites or reputable news sources.

>>> What are the popular items for sale in Walmart in Milpitas

I cannot access real-time information, therefore I cannot provide a list of popular items for sale in Walmart in Milpitas. For the most
up-to-date information on Walmart products and sales, please refer to their official website or other reputable news sources.

>>> How is the mortgage rates as of Feb 2024

I cannot access real-time information, therefore I cannot provide a mortgage rate as of February 2024. For the most up-to-date
information on mortgage rates, please refer to reputable financial news sources.

RAG(Retrieval Augmented Generation)

Remember to treat you are bringing up Your Child

- To provide specific/domain Knowledge
- LLM knowledgebase is not upto date

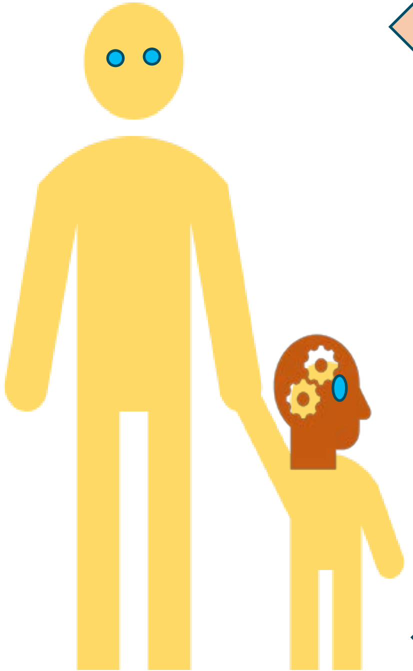
Provide information Sources to develop Knowledge

Read this article → Website

Read this book → PDF

Refer to this set of books → Database

Read from this source → API



Process to develop Knowledge from external sources

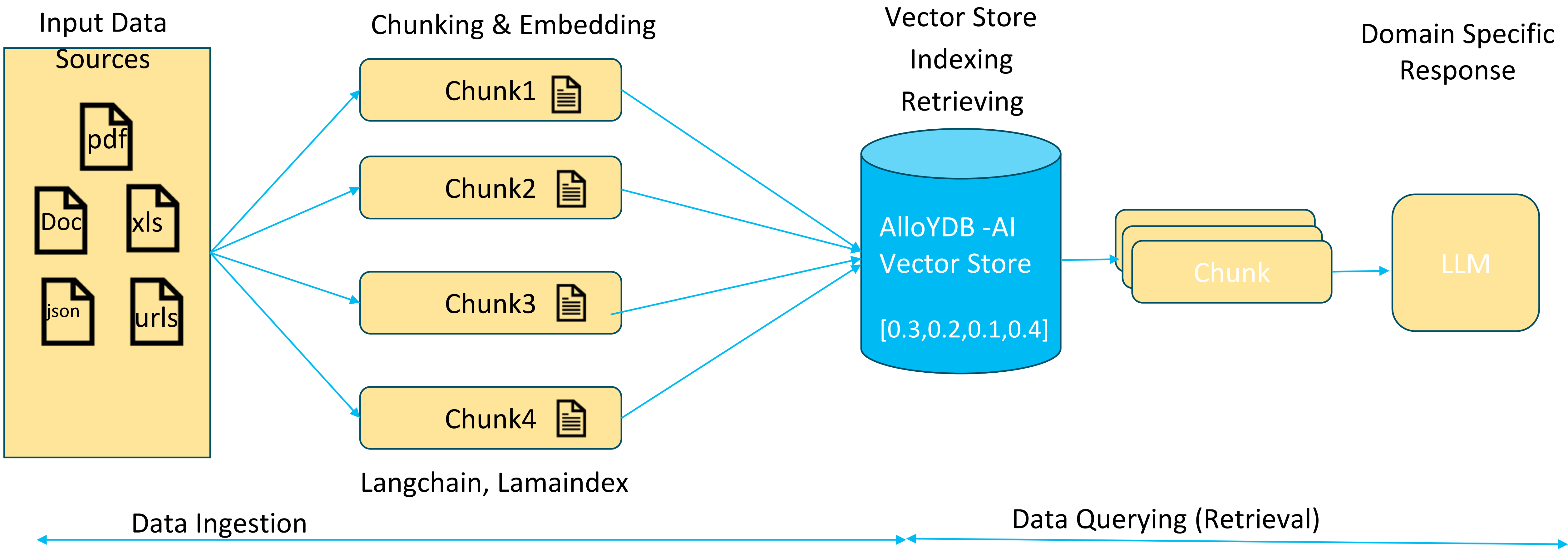
Chunking — Slicing the data

Embedding — Vectorize [0.4,0.1]

Store & Index — Store distance between vectors

User Response — Serve Users with new Knowledge

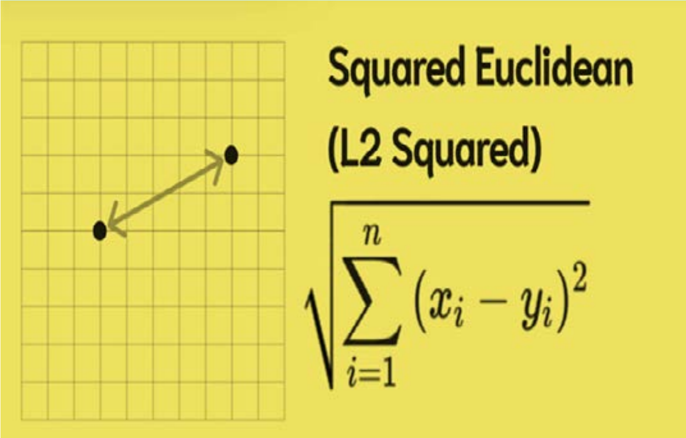
AlloyDB AI – Vector Database



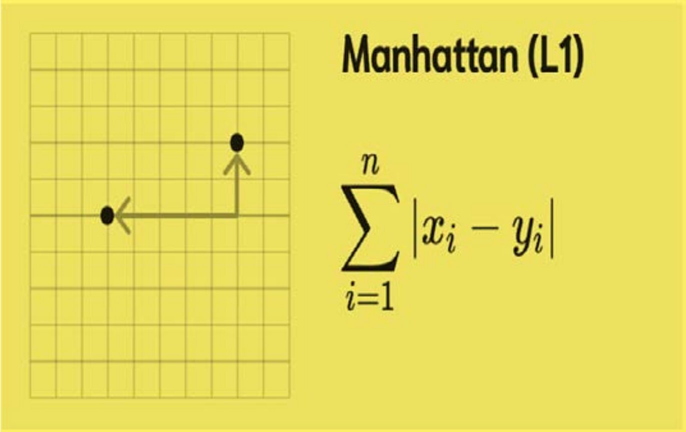
PLiOPS EXTREME DATA PROCESSOR

Vector Search : Find Most Similar Embeddings

L2_distance(vector1, vector2)

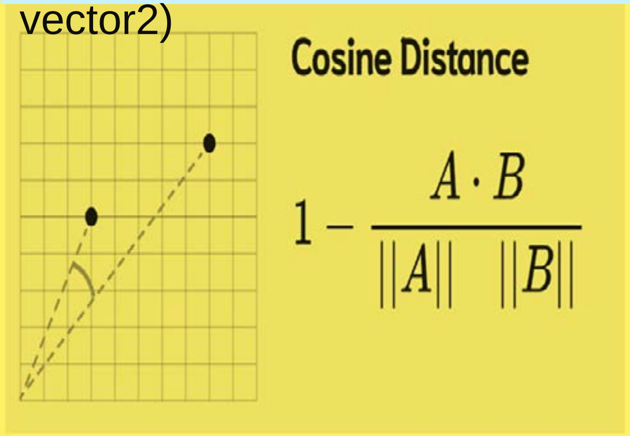


L1_distance(vector1, vector2)

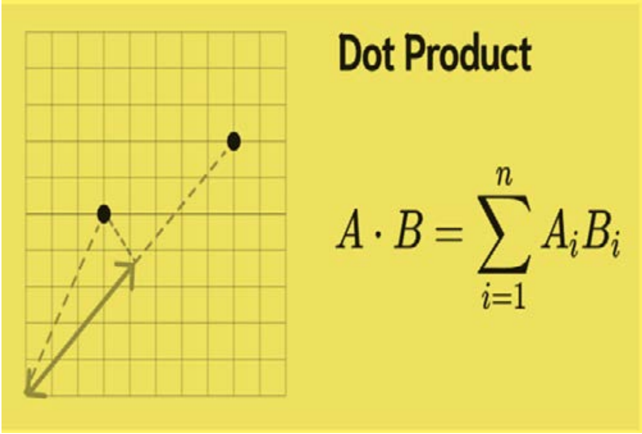


Vector Database

cosine_distance(vector1, vector2)



Inner_product(vector1, vector2)

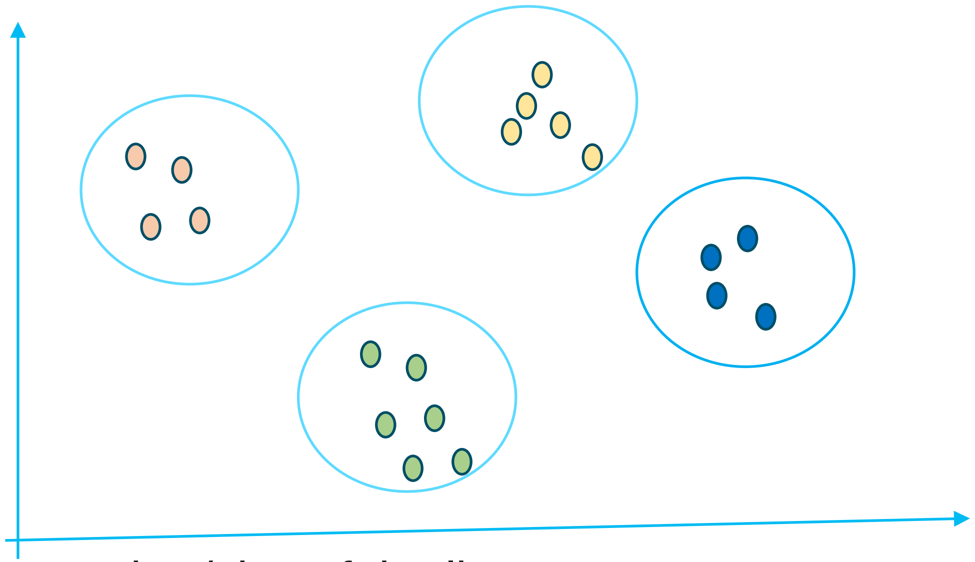


Source : Deep Learning.ai

EXTREME DATA PROCESSOR

Postgres Vector Indexes

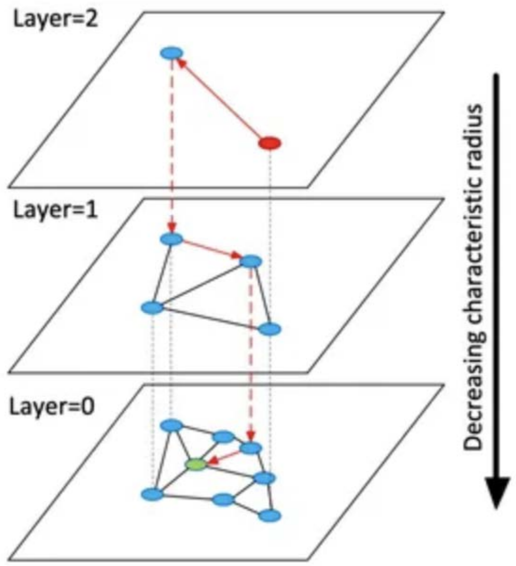
IVF(Inverted File Index)



Number/size of the lists

Search: Number of lists to be verified

Hierarchical Navigable Small Worlds(HNSW)



m: Maximum number of connections per layer

Ef_construction : Size of Dynamic list for construction graph

PLiOPS EXTREME DATA PROCESSOR

RAG Deployment: Decision Factors

Cost of Querying =
Dimensions * Queries

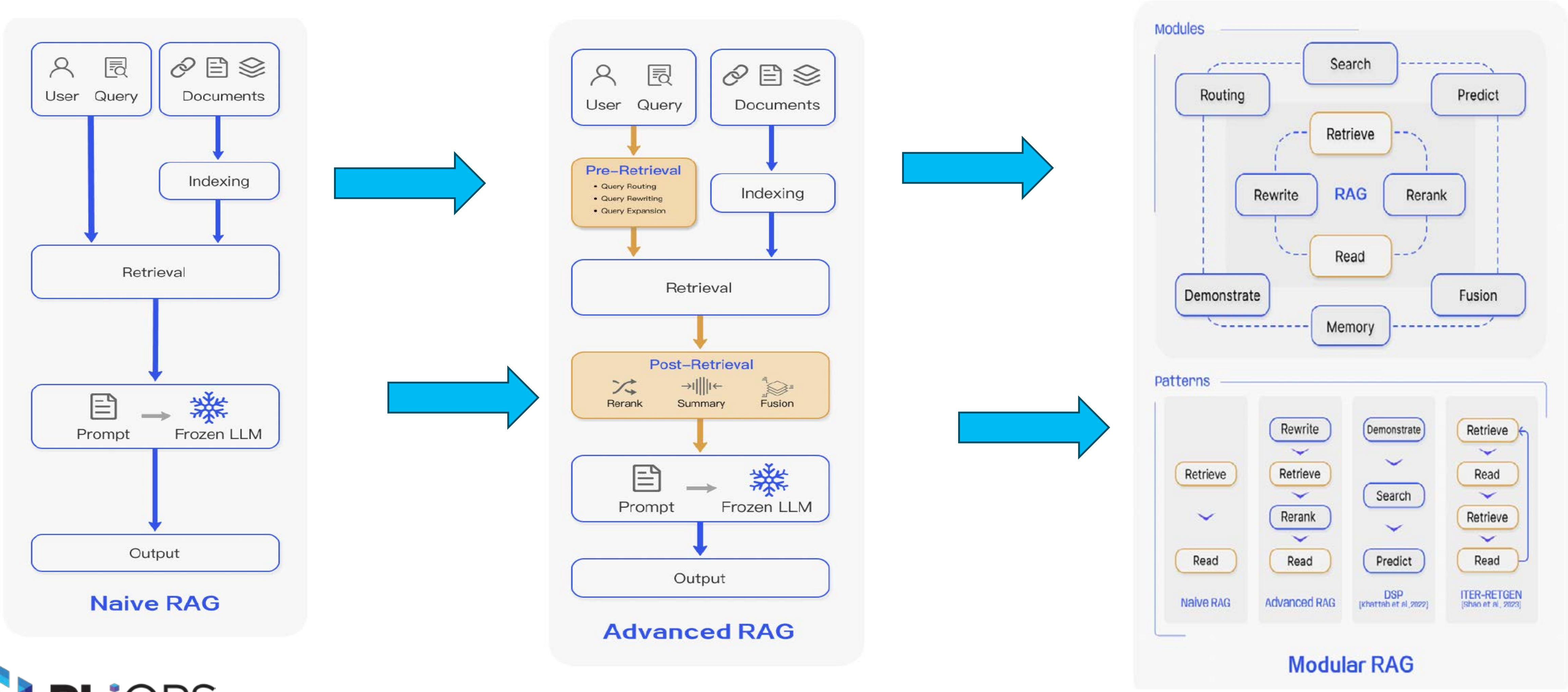
Cost of Indexing =
Dimensions * Writes

Vector
Database

Storage Cost =
Record Count * Dimensions

Search Latency =
Dimensions * Index
Performance

RAG Evolution: Basic/Advanced/Modular



PLiOPS
EXTREME DATA PROCESSOR

Source: *Retrieval-Augmented Generation for Large Language Models*

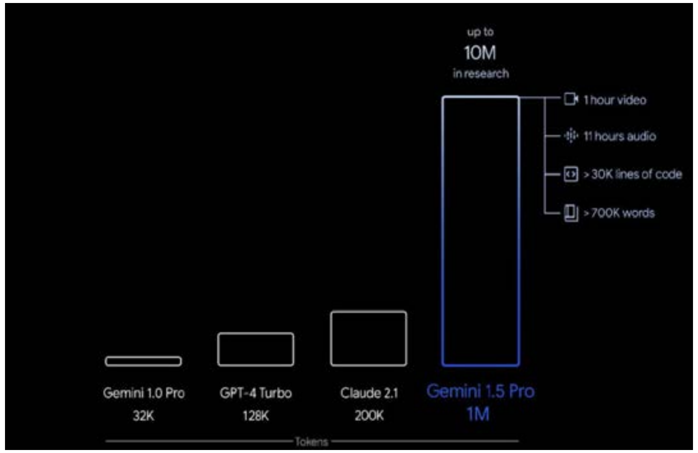
RAG Use cases & Advantages

- Chatbots
- Searching for similar content (Text, Image,Video)
- Personalized recommendation
- Detecting Anomalies

Advantages

- Reduce Hallucinations
- Enable a LLM to cite Sources
- Solve Knowledge Intensive Tasks

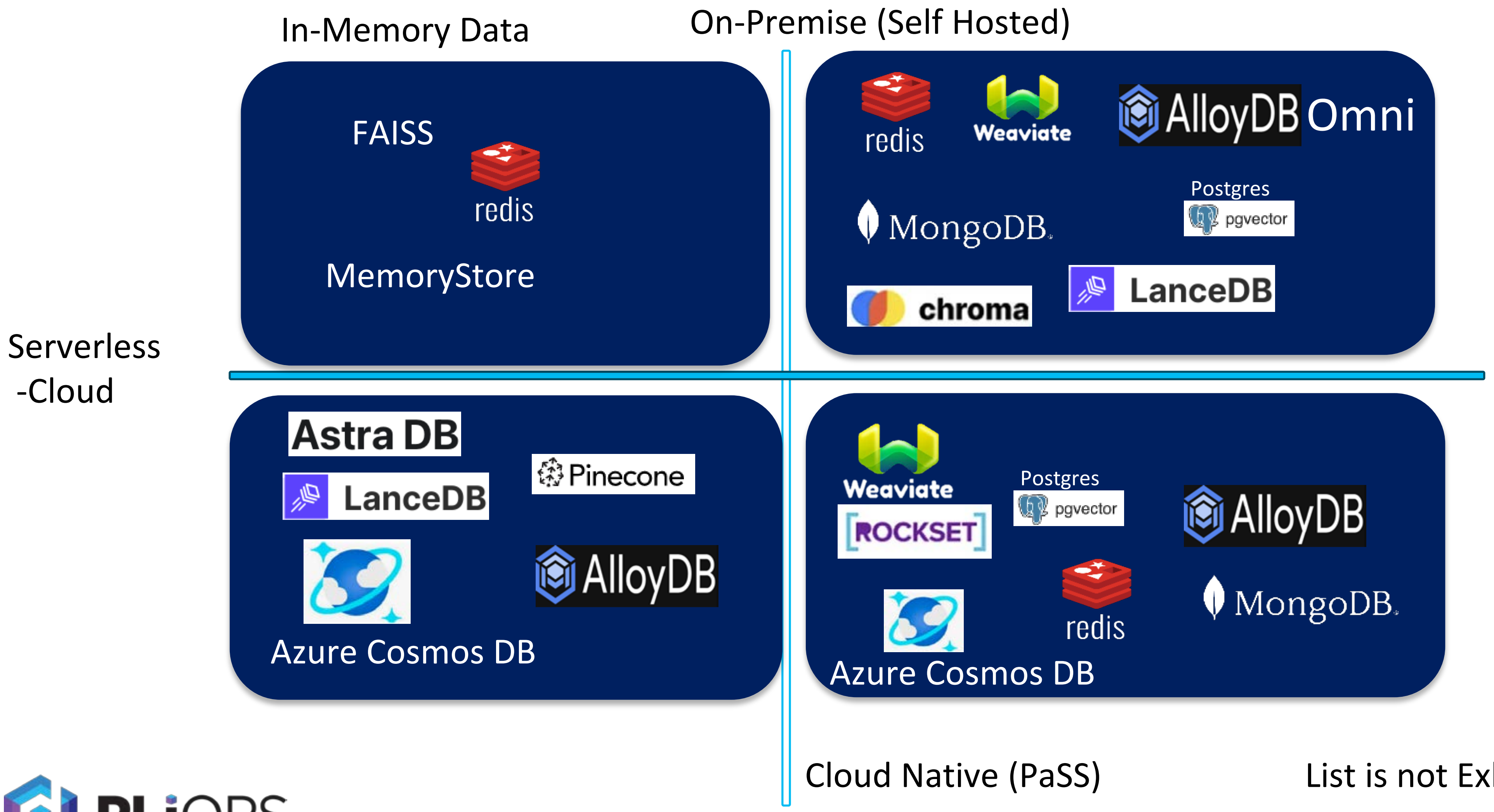
Model	w/o RAG	w/ RAG
GPT-4-Turbo	0.700	0.835
GPT-3.5-Turbo	0.669	0.804
Mixtral-8×7B	0.583	0.808
Llama-2-70b	0.609	0.760
Gemini Pro	**?**	**?**



Source: Google

PLiOPS EXTREME DATA PROCESSOR

Vector Database Landscape

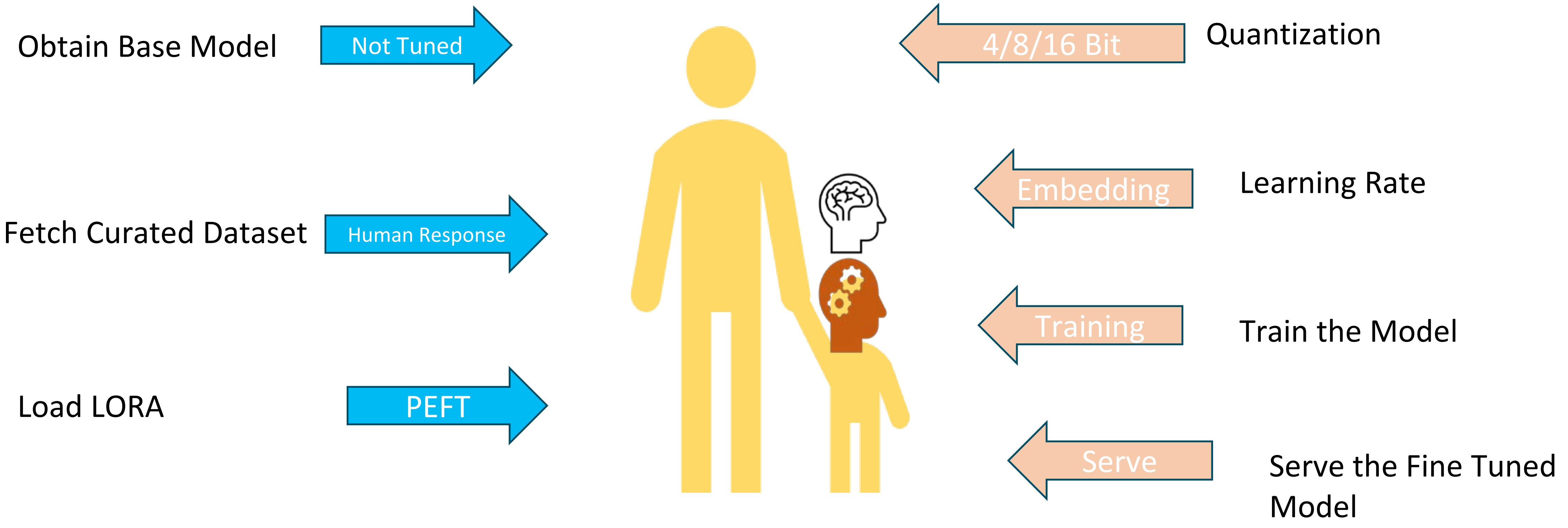


List is not Exhaustive

PLiOPS EXTREME DATA PROCESSOR

Fine Tuning

Treat it as your bringing up Your Own Child

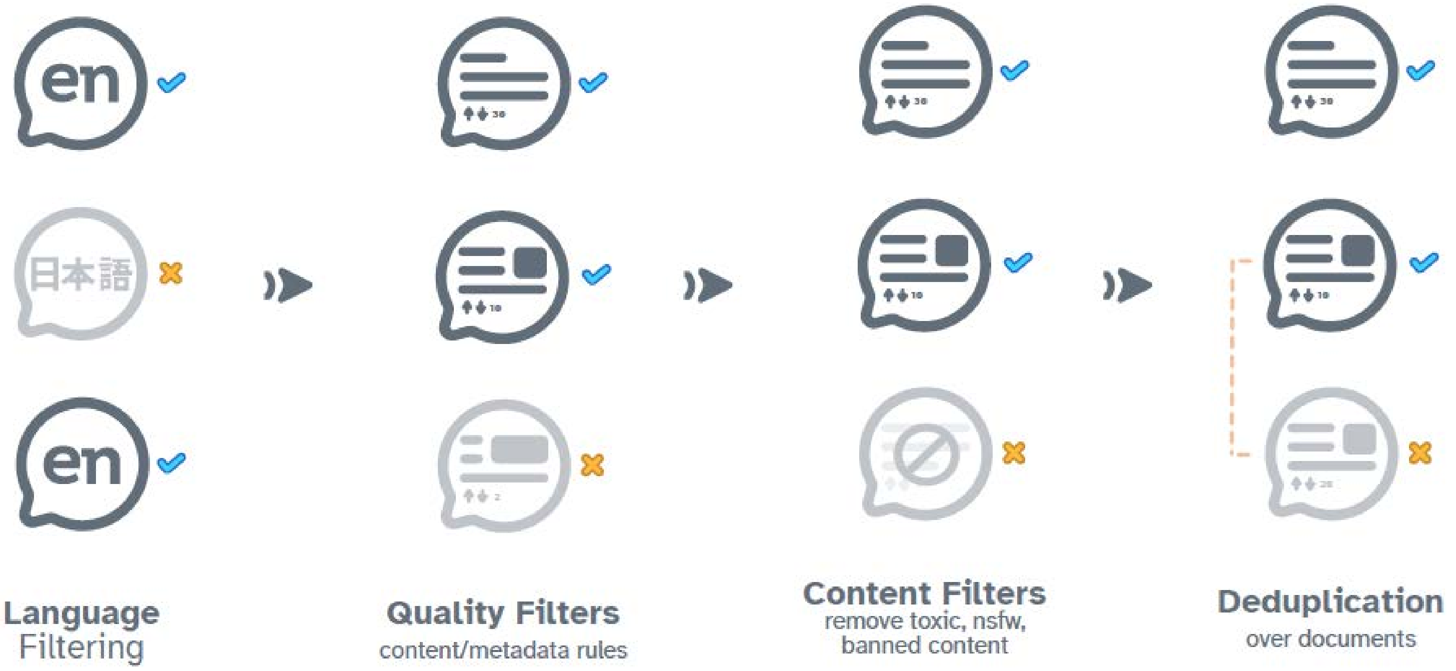


Build & Train Model from Scratch

Source	Doc Type	UTF-8 bytes (GB)	Documents (millions)	Unicode words (billions)	Llama tokens (billions)
Common Crawl	web pages	9,022	3,370	1,775	2,281
The Stack	code	1,043	210	260	411
C4	web pages	790	364	153	198
Reddit	social media	339	377	72	89
PeS2o	STEM papers	268	38.8	50	70
Project Gutenberg	books	20.4	0.056	4.0	6.0
Wikipedia, Wikibooks	encyclopedic	16.2	6.2	3.7	4.3
Total		**11,519**	**4,367**	**2,318**	**3,059**

Dataset for Training

Conversational Forums Pipeline



Source: *Dolma an Open Corpus of Three Trillion Tokens for Language Model Pretraining Research*

AlloyDB (Postgres) Demo

PLIOPS
EXTREME DATA PROCESSOR

Thank You

PLiOPS EXTREME DATA PROCESSOR

Backup

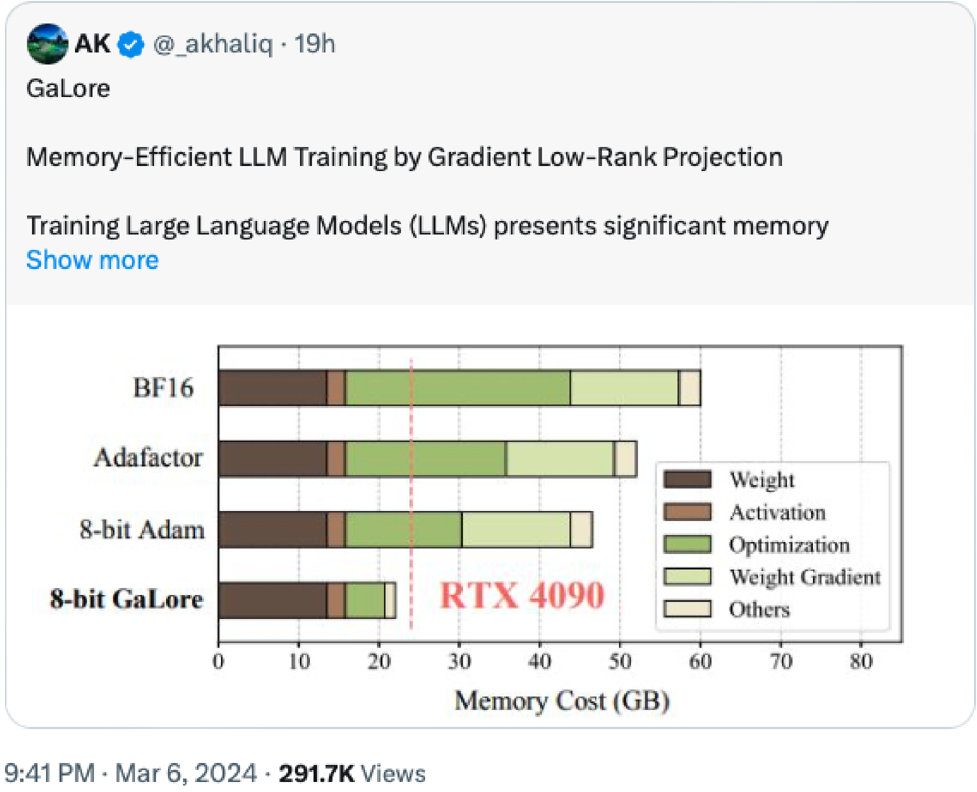
Steps Required to perform Task

```
text = f""" The role of a connected data platform from Edge-Core-
Cloud is becoming more crucial as organizations gather an ever-
increasing volume of data from IOT devices, customer transactions
and thirdparty sources. Data modernization initiatives can prove to
be a game changer for retail enterprises to efficently store and
process the data at the edge, data centers and Cloud. AlloyDB Omni
solution from Google, Lenovo,Pliops is designed and developed to
serve retail customer data modernization needs E-Commerce
Acceleration: Make shopping experience faster an smoother.
"""
# example 1
prompt_1 = f"""
Perform the following actions:
1 - Summarize the following text in 1 sentence.
2 - Translate the summary into French.
3 - How does Retail customer benefited with Solution.
Separate your answers with line breaks.
Text:
```{text}```
"""
response = get_completion(prompt_1)
print("Completion for prompt 1:")
print(response)
```

```
Completion for prompt 1:
 1 - Retail enterprises can benefit from the
AlloyDB Omni solution from Google, Lenovo, and
Pliops to efficiently store and process data for
data modernization initiatives, ultimately
improving the shopping experience for customers.
2 - Les entreprises de vente au détail peuvent
bénéficier de la solution AlloyDB Omni de Google,
Lenovo et Pliops pour stocker et traiter
efficacement les données dans le cadre
d'initiatives de modernisation des données,
améliorant ainsi l'expérience d'achat pour les
clients.
3 - Retail customers benefit from faster and
smoother shopping experiences due to the efficient
storage and processing of data provided by the
solution.
```

PLiOPS
EXTREME DATA PROCESSOR
```

Pre-Training Llama 7B LLM Single GPU



- Llama 7B Large Language Model Training
- Single consumer-grade GPU (RTX 4090) 24GB
- **Gradient Low-Rank Projection**
- Gradient weight matrix as low rank without performance degradation
- **82.5% reduction in memory for storing optimizer states during training.**

PLiOPS
EXTREME DATA PROCESSOR

Vector Database Comparison

Vector Database	Supports Deployment in Current Database	Size of Vector Dimensions	Aggregations	Queries per second ANN Benchmarks	Metadata Filtering	Time Based Metadata Filtering	Time-Series Compression	Hybrid Search
pgvector on PostgreSQL	Yes	16000	Yes		Yes	Yes (Supports Postgres date data types)		Yes
AlloyDB/Omni – Vector for Postgres	Yes	16000	Yes		Yes	Yes (Supports Postgres date data types)		Yes
Qdrant	No	-	No		Yes	Somewhat (Need to convert time to an integer)	No	Yes (Sparse-Dense Vectors
ChromaDB	No		No		Yes	Somewhat (Need to convert time to an integer)	No	No
KDB.AI			Yes		Yes	Yes (datetime64, timedelta64)	Yes (future)	Yes (future)
Weaviate	No	65535	Yes		Yes	Yes (Supports 'date' data type)	No	Yes (Sparse-Dense Vectors)
Pinecone	No	20000	No		Yes	Somewhat (Need to convert date/time to integer in Unix time)	No	Yes (Sparse-Dense Vectors)
Milvus	No	34768	No		Yes	Somewhat (Need to convert date/time to integer in Unix time)	No	No, they use the phrase "Hybrid Search", but it really means metadata filtering

PLiOPS
EXTREME DATA PROCESSOR

Open Source Models

Model	Parameters	Size
Llama 2	7B	3.8GB
Mistral	7B	4.1GB
Dolphin Phi	2.7B	1.6GB
Phi-2	2.7B	1.7GB
Neural Chat	7B	4.1GB
Starling	7B	4.1GB
Code Llama	7B	3.8GB
Llama 2 Uncensored	7B	3.8GB
Llama 2 13B	13B	7.3GB
Llama 2 70B	70B	39GB
Orca Mini	3B	1.9GB
Vicuna	7B	3.8GB
LLaVA	7B	4.5GB
Gemma	2B	1.4GB
Gemma	7B	4.8GB

PLiOPS
EXTREME DATA PROCESSOR