

AI w chmurze Oracle

Fundamenty nowoczesnych rozwiązań

Tomasz Żegleń

Account Cloud Engineering Manager



AI

- LLMs
- LLM Agents
- Vector Databases
- GPU Infrastructure
- AI Services
- Data Platform



Zrozumieć Oracle

OPROGRAMOWANIE
oraz
SPRZĘT

CHMURA

CHMURA
PUBLICZNA
GLOBALNA

CHMURA
PUBLICZNA
SUWERENNA
DLA UNII
EUROPEJSKIEJ

CHMURA
PRYWATNA
U KLIENTA

WYBRANE USŁUGI
CHMUROWE W DATA CENTER
KLIENTA

DUŻE REGIONY CHMUROWE W
DATA CENTER KLIENTA

Zrozumieć Oracle

OPROGRAMOWANIE
oraz
SPRZĘT

CHMURA

CHMURA
PUBLICZNA
GLOBALNA

CHMURA
PUBLICZNA
SUWERENNA
DLA UNII
EUROPEJSKIEJ

CHMURA
PRYWATNA
U KLIENTA

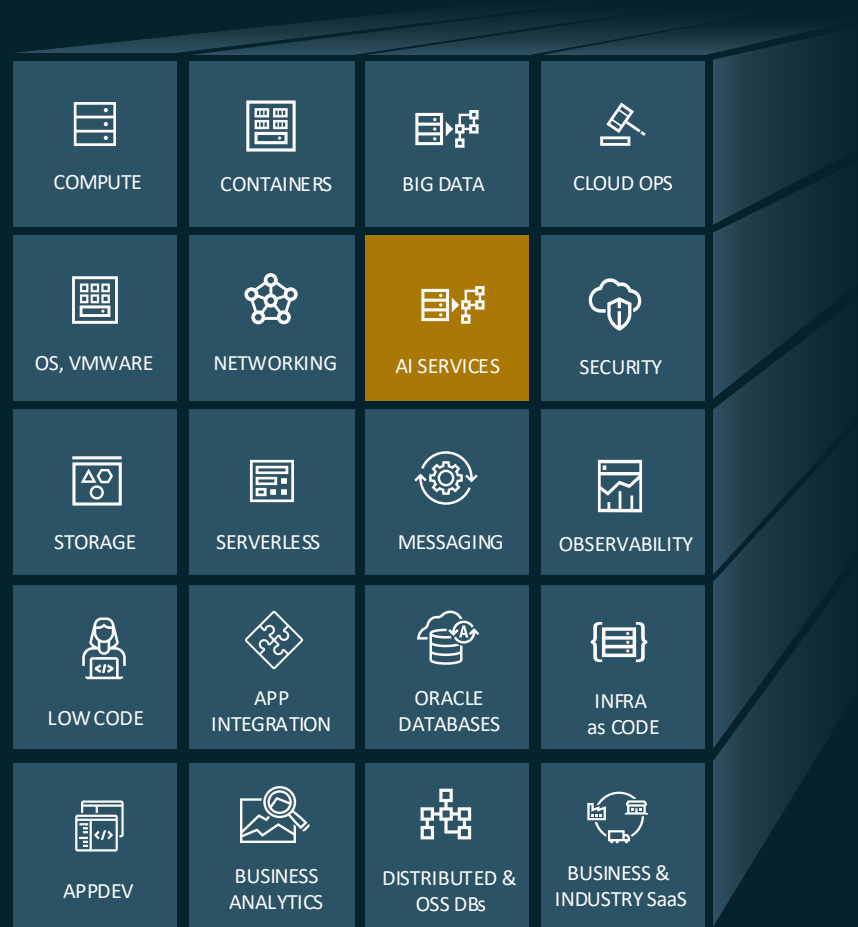
WYBRANE USŁUGI
CHMUROWE W DATA CENTER
KLIENTA

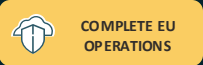
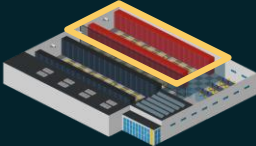
DUŻE REGIONY CHMUROWE W
DATA CENTER KLIENTA

AI



Oracle Cloud Infrastructure oferuje ponad 150 usług chmurowych w dostosowanych do klienta modelach wdrożeniowych w tym **Usługi AI**,



Globalna Chmura Publiczna	Data Center Oracle	
Europejska Suwerenna Chmura Publiczna	Data Center Oracle	 
Chmura Hybrydowa	Data Center Klienta	 Database Cloud IaaS Cloud Roving Edge
Chmura Dedykowana	Data Center Klienta	
Multicloud	Data Center Azure, GCP, AWS	  



AI

LLMs
LLM Agents
Vector Databases
GPU Infrastructure
AI Services
Data Platform



Jest wiele dróg do rozwiązania zdefiniowanego problemu do rozwiązania przez AI, kluczowym elementem jest wybranie właściwego podejścia i narzędzi

Klasyczny ML oparty na algorytmach

Przykład: Wykorzystanie algorytmów klasyfikacyjnych (np. SVM, drzewa decyzyjne) do analizy danych operacyjnych fabryki – np. klasyfikacja parametrów pracy instalacji w celu wczesnego wykrywania stanów odbiegających od normy, mogących prowadzić do awarii lub spadku wydajności.

- OML4PY w Oracle Data Science
- AutoML w Oracle Database 23ai
- Funkcje analityczne w Oracle Database 23ai

Analizy grafowe

Przykład: Analizy grafowe w sektorze dystrybucji towarów umożliwiają wykrywanie nieprawidłowości i nadużyć poprzez analizę relacji między sprzedawcą, dostawcami, logistyką – np. identyfikacja nietypowych wzorców tankowań lub tras dostaw, które mogą wskazywać na kradzieże paliwa lub manipulacje.

- OML4PY w Oracle Data Science
- Oracle Graph Studio
- Funkcje grafowe w Oracle Database 23ai

Sieci neuronowe (budowa własnych modeli)

Przykład: Opracowanie dedykowanego modelu CNN do analizy obrazu z kamer przemysłowych – np. automatyczne wykrywanie wycieków, dymu lub innych anomalii wizualnych w czasie rzeczywistym, wspierające systemy bezpieczeństwa i ochrony środowiska.

- Oracle Data Science
- OCI GPU Compute

Prompt Engineering

Przykład: Ekstrakcja kluczowych informacji (np. numerów partii, dat, lokalizacji, opisu usterki), przypisywanie spraw do odpowiednich działów oraz generowanie wstępnych odpowiedzi lub raportów z przebiegu reklamacji – przyspieszając cały obieg dokumentów i zwiększając jego spójność.

- Oracle Generative AI
- OCI GPU Compute – własne modele językowe np. Bielik

Dotrenowywanie modeli istniejących (Fine-tuning)

Przykład: Dotrenowanie modelu językowego do analizy raportów z kontroli jakości towarów. Automatyczne wykrywanie odchyłań od norm jakościowych i identyfikacja powtarzających się problemów w konkretnych partiach lub lokalizacjach.

- Oracle Data Science
- OCI GPU Compute
- OCI Generative AI
- OCI AI Services

Retrieval-Augmented Generation (RAG)

Przykład: System łączący wyszukiwanie dokumentów technicznych, norm branżowych i instrukcji operacyjnych z generowaniem zwięzłych podsumowań. Dostarczanie inżynierom i operatorom rafinerii aktualnych, kontekstowych odpowiedzi na pytania dotyczące eksploatacji instalacji, procedur awaryjnych czy zmian w regulacjach bezpieczeństwa.

- OCI Generative AI
- OCI Generative AI Agents
- Oracle Database 23ai / Vector Search for Embeddings

Klasyczne raporty i analizy (np. SQL)

Przykład: Analiza danych sprzedażowych ze stacji paliw przy użyciu zapytań SQL. Identyfikacja trendów zakupowych klientów, analiza sezonowości oraz monitorowanie skuteczności promocji i programów lojalnościowych, co wspiera planowanie działań marketingowych i zaopatrzenia.

- Oracle Analytics Cloud

Mapy i czasoprzestrzeń

Przykład: Analizy przestrzenno-czasowe wykorzystywane do monitorowania i optymalizacji tras dostaw oraz lokalizacji punktów sprzedaży. Identyfikacja opóźnień, przeciążeń logistycznych i luk pokrycia sieci w czasie rzeczywistym, co wspiera efektywniejsze planowanie dystrybucji oraz decyzje o budowie nowych punktów sprzedaży.

- Oracle Spatial Studio
- Oracle Spatial & Graph w ramach Oracle Database 23ai



Nasi Klienci wykorzystują rozwiązania AI w kilku modelach w zależności od przyjętej strategii, Oracle oferuje rozwiązania i narzędzia dopasowane do każdego podejścia

Provider-managed
Self-managed

AI Deployment Approaches¹

Consume Embedded in Apps	Embed APIs in a Custom App Frame	Extend Models via Data Retrieval	Extend Models via Fine-Tuning	Build Models from scratch
Application	Application	Application	Application	Application
Data retrieval and prompt engineering	Data retrieval and prompt engineering	Data retrieval and prompt engineering	Data retrieval and prompt engineering	Data retrieval and prompt engineering
Fine-Tuning	Fine-Tuning	Fine-Tuning	Fine-Tuning	Fine-Tuning
Foundation Model	Foundation Model	Foundation Model	Foundation Model	Foundation Model



Każde podejście do wdrażania wiąże się z kompromisami między korzyściami a ryzykiem.

Kluczowe czynniki wpływające na te kompromisy to:

- Koszty
- Wiedza organizacyjna i branżowa
- Możliwość kontrolowania bezpieczeństwa i prywatności
- Kontrola nad wynikami modelu
- Prostota wdrożenia

Consume

Work

Build

SaaS
Oracle Applications

AI Services -> IaaS & PaaS
Oracle Cloud Infrastructure
W modelu Hybrydowym oraz Dedykowanym w serwerowni klienta

1- Source: Gartner 2024, Get AI Ready — What IT Leaders Need to Know and Do, AI Deployment Approaches, GenAI taken as an example but can be extended to wider AI technologies



Przykład: Obliczenie Trójwymiarowego Rozkładu Dawki w ciele pacjenta, podczas Radioterapii

Problem:

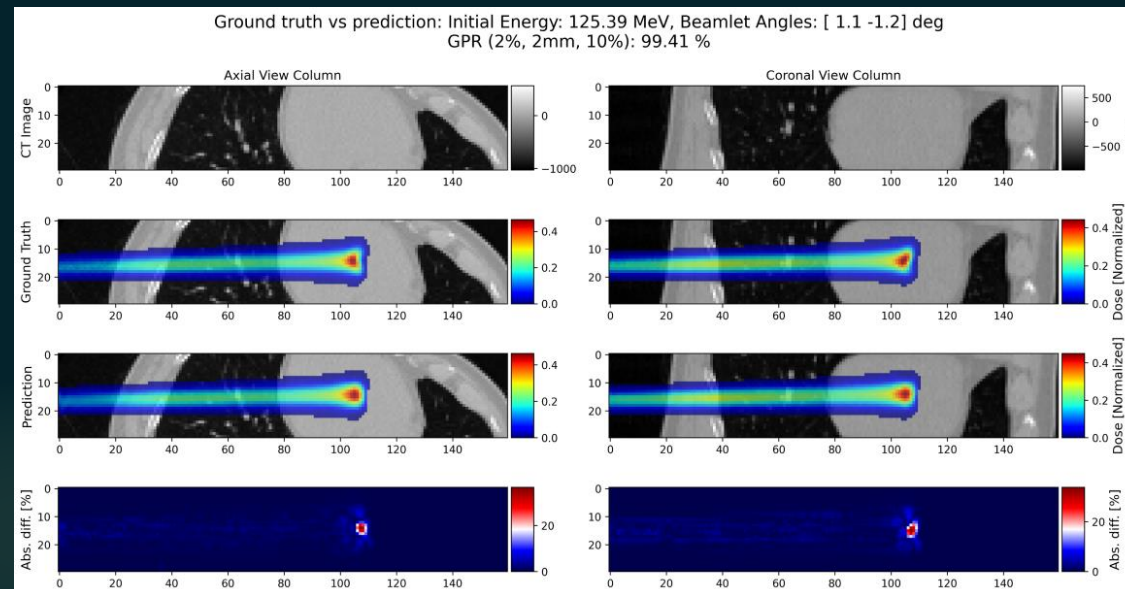
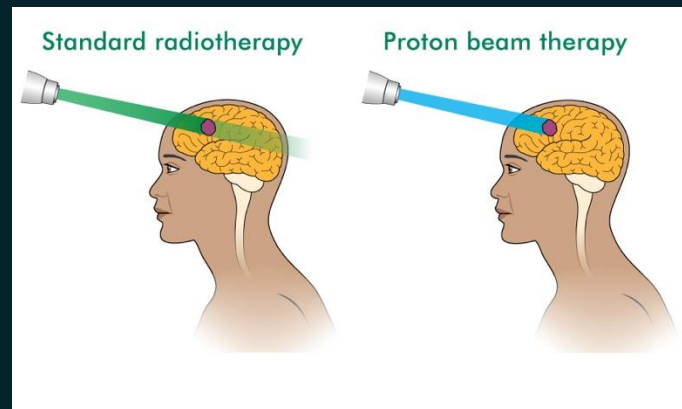
Metody do obliczenia Trójwymiarowego Rozkładu Dawki (z ang. 3-D dose distribution, 3DD) w ciele pacjenta, podczas Radioterapii na dzień dzisiejszy jest kosztownym obliczeniowo modelem, wykorzystującym metody stochastyczne, w szczególności Monte Carlo. Najszybsze oprogramowanie dostępne na rynku na dzień dzisiejszy, oblicza „3DD dose” dla pojedynczej wiązki w czasie około 60 sekund. Na cały plan pacjenta składa się 1000 takich wiązek.

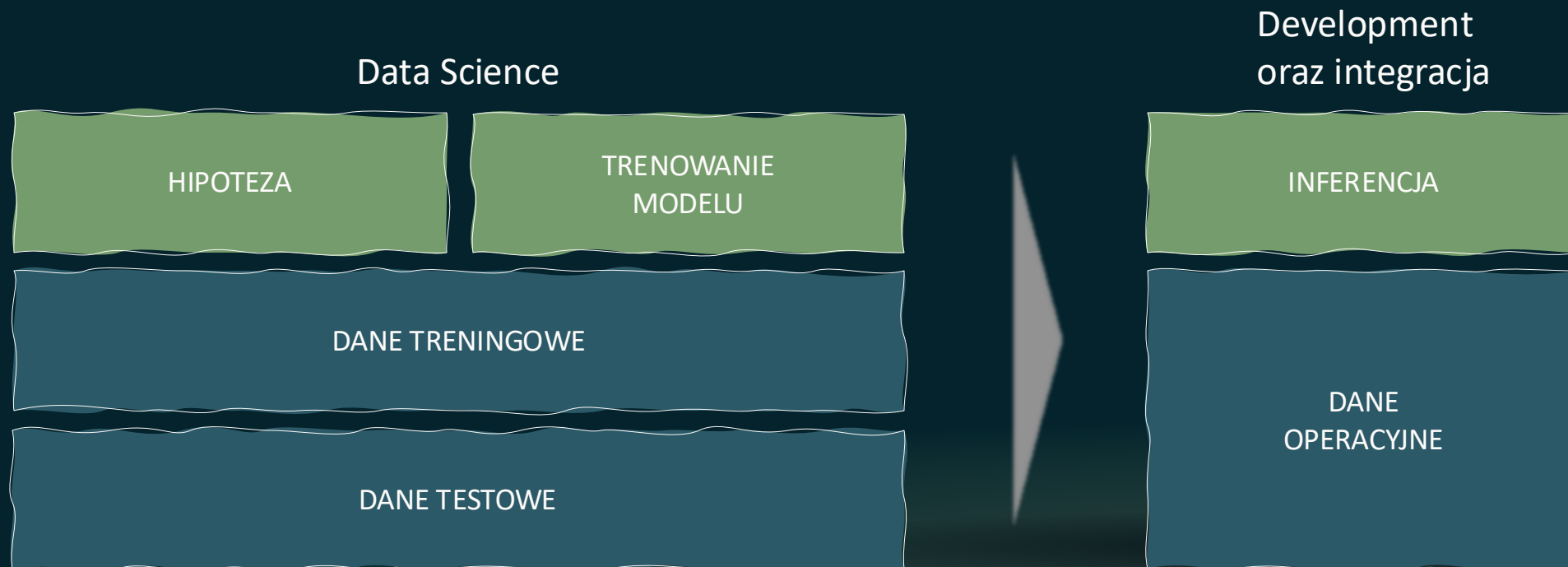
Rozwiązanie:

Metody Uczenia Głębokiego (DL), mają duży potencjał w akceleracji tych obliczeń. Tworzymy algorytm wykorzystujący głębokie sieci neuronowe do predykcji jaki będzie 3DD na podstawie skanu z tomografii komputerowej pacjenta, informacji w formie mapy wpływu jaki jest kierunek cząstek elementarnych niosących radiację (w naszym przypadku protonów) oraz energii tej wiązki wyrażonej w MeV.

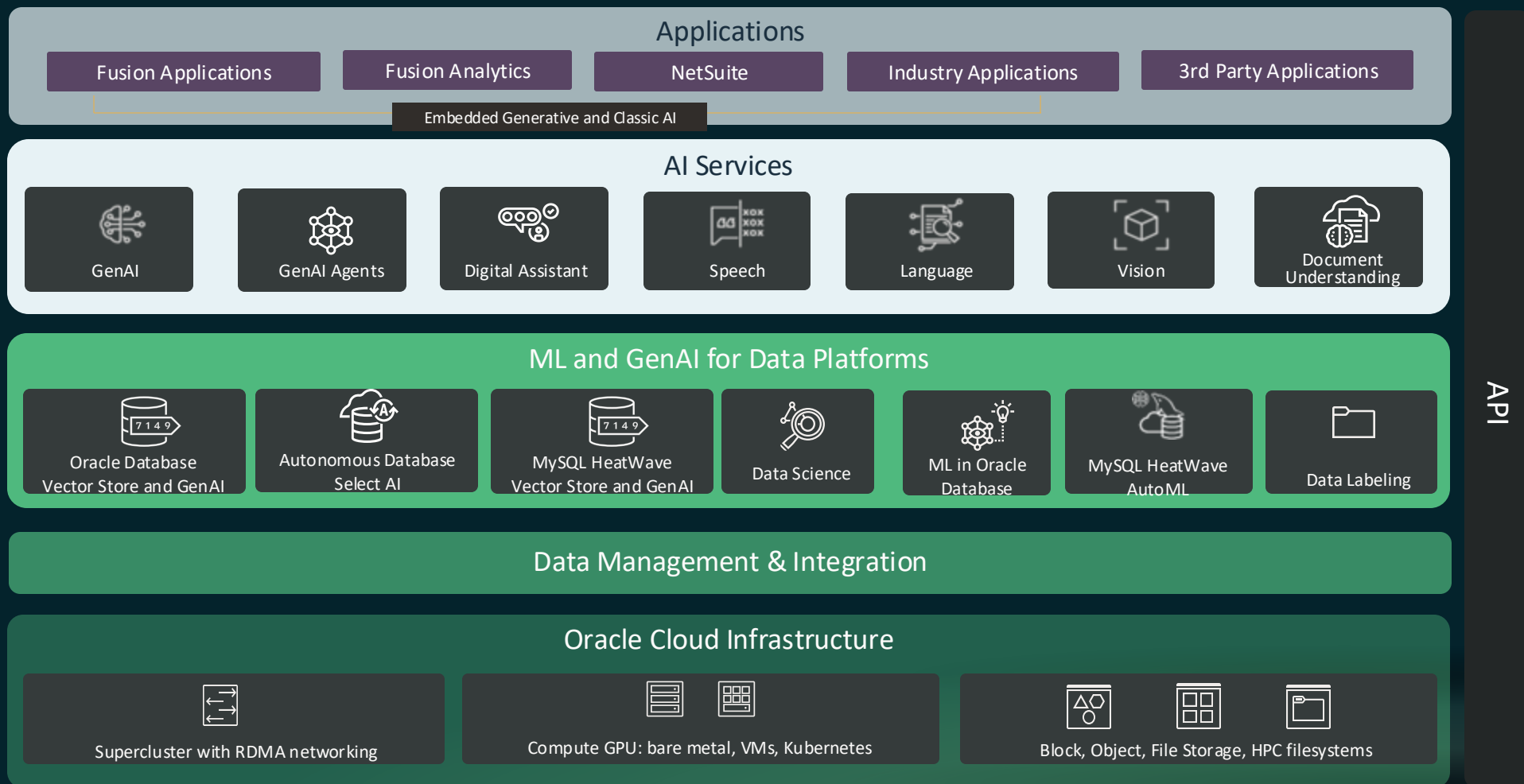
Aktualne wyniki:

Na ten moment algorytm osiąga obiecujące wyniki, mierzone jako wskaźnik GammaPassRate. Na ten moment średnia wartość GammaPassRate dla naszego modelu, mierzonej względem algorytmów Monte Carlo wynosi 94%. Nasz algorytm oblicza 3DD w czasie 8ms, **co oznacza akcelerację około 7500-krotną.**





Zbudowaliśmy kompletny ekosystem łączący sztuczną inteligencję w tym Gen AI na wszystkich warstwach chmury Oracle



> **Ready to Consume** for Business users

> **Ready to Work** for Developers, Data Scientists, ...

> **Ready to Build** for Data Engineering, Operations, ...

> Infrastructure for the **best price / performance**



Press Release

OpenAI Selects Oracle Cloud Infrastructure to Extend Microsoft Azure AI Platform

Austin, Texas—June 15, 2024

Oracle, Microsoft, and OpenAI are partnering to extend the Microsoft Azure AI platform to [Oracle Cloud Infrastructure \(OCI\)](#) to provide additional capacity for OpenAI.

OpenAI is the AI research and development company behind ChatGPT, which provides generative AI services to more than 100 million users every month.

"We are delighted to be working with Microsoft and Oracle. OCI will extend Azure's platform and enable OpenAI to continue to scale," said Sam Altman, Chief Executive Officer, OpenAI.

"The race to build the world's greatest large language model is on, and it is fueling unlimited demand for Oracle's Gen2 AI infrastructure," said Larry Ellison, Oracle Chairman and CTO. "Leaders like OpenAI are choosing OCI because it is the world's fastest and most cost-effective AI infrastructure."

OCI's leading AI infrastructure is advancing AI innovation. OpenAI will join thousands of AI innovators across industries worldwide that run their AI workloads on OCI AI infrastructure. Adept, Modal, MosaicML, NVIDIA, Reka, Suno, Together AI, Twelve Labs, xAI, and others use [OCI Supercluster](#) to train and inference next-generation AI models.

OCI's purpose-built AI capabilities enable startups and enterprises to build and train models faster and more reliably anywhere in Oracle's distributed cloud. For training large language models (LLMs), OCI Supercluster can scale up to 64k NVIDIA Blackwell GPUs or GB200 Grace Blackwell Superchips connected by ultra-low-latency RDMA cluster networking and a choice of HPC storage. OCI Compute virtual machines and OCI's bare metal NVIDIA GPU instances can power applications for generative AI, computer vision, natural language processing, recommendation systems, and more.



Customer References >

CERN openlab uses OCI for AI model training in its research

CERN openlab tests AI to support physics research scenarios, including building custom large language models with OCI Data Science Service.

Share: [f](#) [X](#) [in](#) [M](#)

"With Oracle, we can build and reuse AI models in a more environmentally friendly way. OCI Data Science makes it easier to share models for projects with multiple detectors and challenges as we experiment with even larger models."

Resato Cardoso, Data Scientist, CERN openlab



Products list

[Oracle Cloud Infrastructure](#)
[OCI Data Science](#)
[OCI AI Infrastructure](#)



Oracle signs up Meta for AI training deal

By Craig Hume published December 10, 2024

Meta projects to use Oracle Cloud for further AI training



When you purchase through links on our site, we may earn an affiliate commission. [Track.com](#) 2023.



Image credit: Oracle

- Oracle Cloud reveals new deal to train Meta's AI models on its services
- Oracle Cloud will look to enhance the training and deployment of Llama LLM family
- Partnership comes as revenue growth fails to meet targets

Press Release

Oracle Cloud Infrastructure Utilized by Microsoft for Bing Conversational Search

Austin, Texas—November 5, 2022

Oracle today announced a multi-year agreement with Microsoft to support the explosive growth of AI services. Microsoft is using [Oracle Cloud Infrastructure \(OCI\) AI infrastructure](#), along with Microsoft Azure AI infrastructure, for inferencing of AI models that are being optimized to power Microsoft Bing conversational searches daily. Leveraging the [Oracle Interconnects for Microsoft Azure](#), Microsoft is able to use managed services like Azure Kubernetes Service (AKS) to orchestrate OCI Compute at massive scale to support increasing demand for Bing conversational searches.

Bing conversational search requires powerful clusters of computing infrastructure that support the evaluation and analysis of search results that are conducted by Bing's inference model.

"Generative AI is a monumental technological leap and Oracle is enabling Microsoft and thousands of other businesses to build and run new products with our OCI AI capabilities," said Karan Batta, senior vice president, Oracle Cloud Infrastructure. "By furthering our collaboration with Microsoft, we are able to help bring new experiences to more people around the world."



Oracle United Kingdom > Customer References >

Goldman Sachs partners with OCI for enhanced high performance computing

Goldman Sachs Group, Inc., a global leader in financial services, leverages Oracle Cloud Infrastructure (OCI) to bolster high performance computing (HPC) capabilities for risk calculations.

Share: [f](#) [X](#) [in](#) [M](#)

"By leveraging OCI's capabilities as a part of our process, we are poised to enhance performance, scalability, elasticity, and resiliency in our high performance computing endeavors."

Atte Lahtiranta, Chief Technology Officer, Goldman Sachs

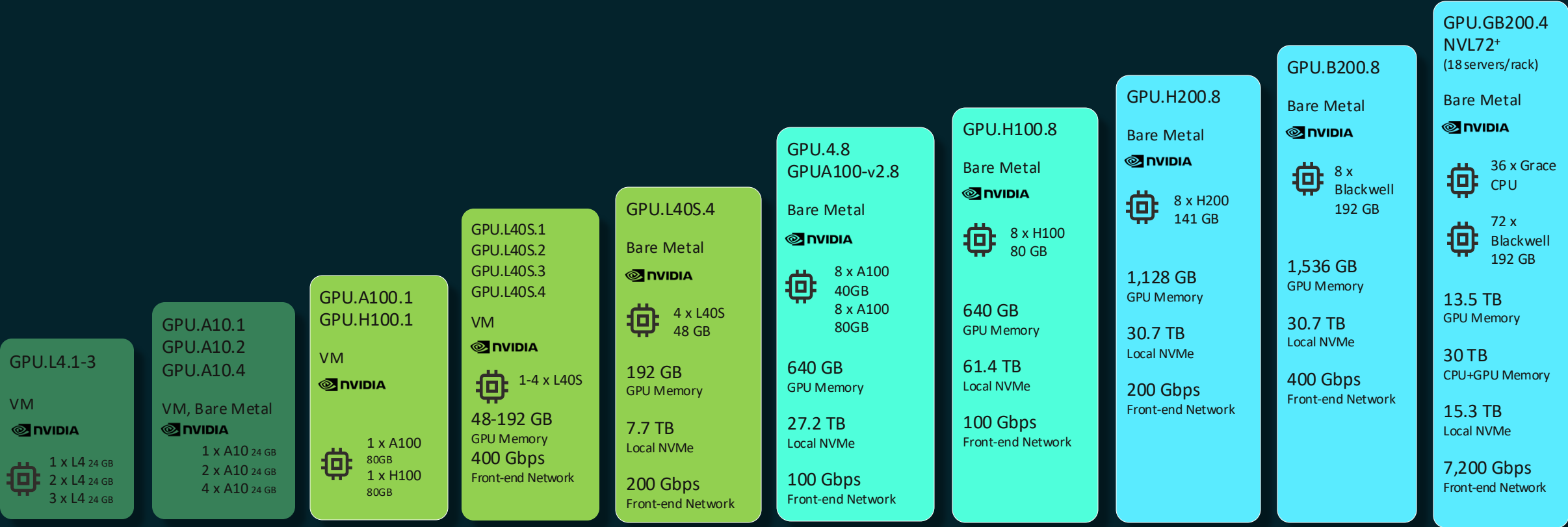
Goldman Sachs

Products list

[Oracle Cloud Infrastructure](#)



Infrastruktura GPU w Oracle Cloud



OCI Superclusters

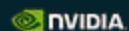
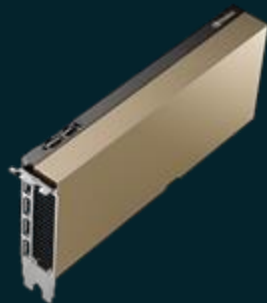
Max GPUs	3,840 GPUs	32,768 GPUs	16,384 GPUs	65,536 GPUs	131,072 GPUs	100,000+ GPUs
RDMA bandwidth per node	800 Gbps	3,200 Gbps	3,200 Gbps	3,200 Gbps	3,200 Gbps	28,800 Gbps

OCI offers NVIDIA GH200 Grace Hopper™ Superchip for testing
OCI also offers older generation NVIDIA P100 and V100 GPUs



Najnowsze GPU od NVIDIA i AMD w Oracle Cloud

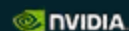
OCI Compute with L40S



L40S

4 x L40S 48 GB GPU
2 x 56c Intel Sapphire Rapids
1 TB DDR5
7 TB local NVMe
2 x 400 Gbps RDMA

OCI Compute with H200



H200

8 x H200 141GB GPU
2 x 56c Intel Sapphire Rapid
3 TB DDR5
30 TB NVMe
8 x 400 Gbps RDMA

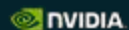
OCI Compute with GB200



GPU.GB200.4 NVL72+ (18 servers/rack)

4 x B200 192 GB GPU
2 x 72c NVIDIA Grace CPU
1 TB LPDDR5
15 TB local NVMe
4 x 400 Gbps RDMA

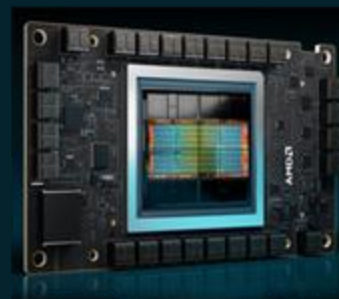
OCI Compute with B200



GPU.B200.8

8 x B200 192 GB GPU
2 x 64c Intel Emerald Rapids
4 TB DDR5
30 TB local NVMe
8 x 400 Gbps RDMA

OCI Compute with MI 300X

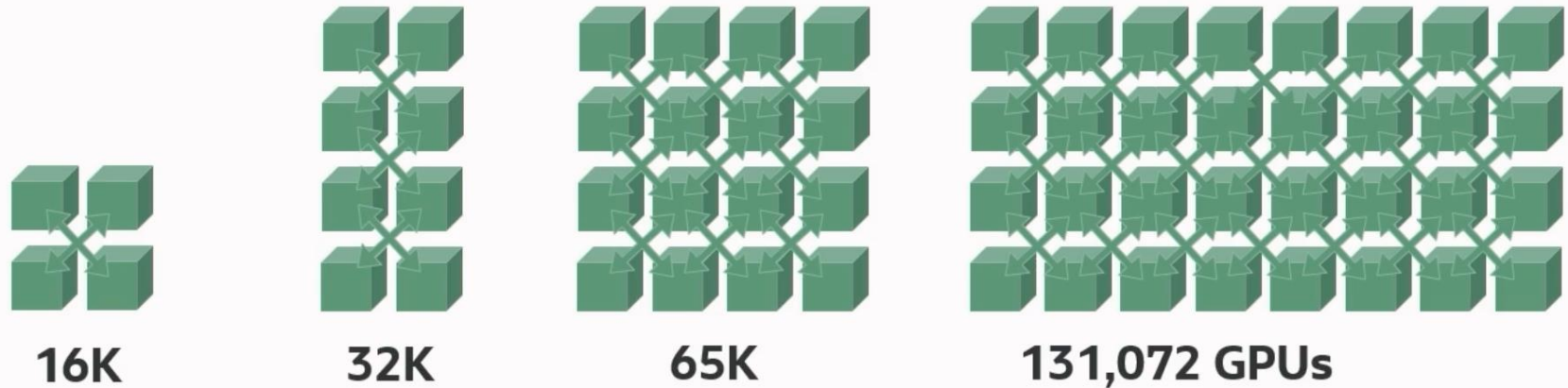


MI300X

8 x MI300X 192GB GPU
2 x 56c Intel Sapphire Rapids
2 TB DDR5
30 TB NVMe
8 x 400Gbps RDMA

Now Orderable

OCI continues to deliver high performance GPUs at increasing scale



Year of Availability

2022

2023

2024

2024-2025

of GPUs

16,384

32,768

65,536

131,072

Total Fabric BW (Pbps)

6.5

13

26

104

GPU to Backend Connectivity (Gbps)

400

400

400

800

Frontend Connectivity (Gbps)

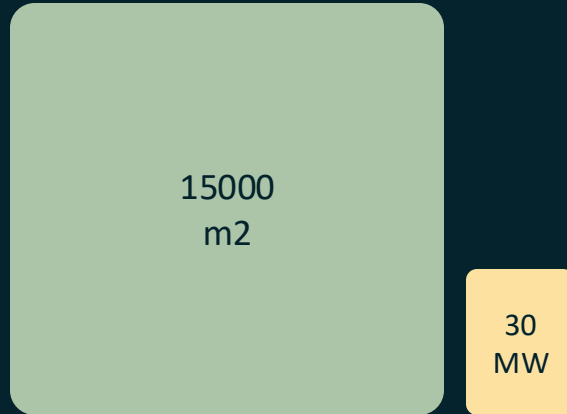
100

100

200

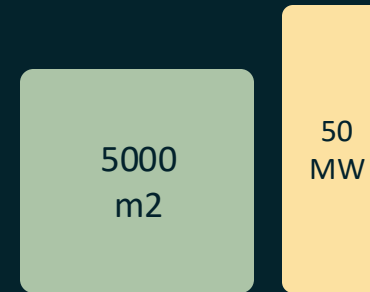
200 or 400

Tradycyjne Data Center



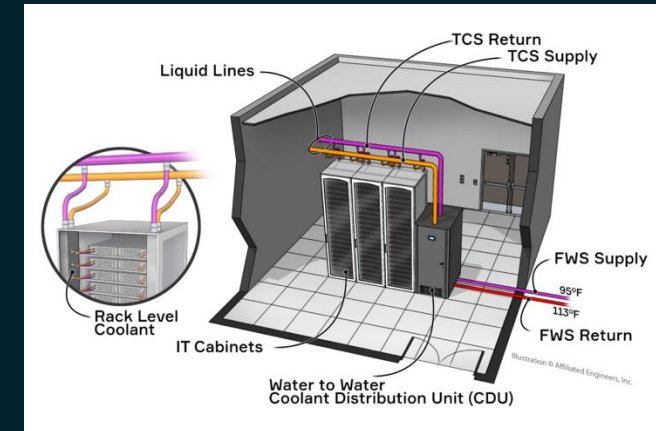
Standardowe szafy rack chłodzone powietrzem
500kg-600kg na szafę, 12 kW na szafę

GPU Data Center



GPU rack chłodzony cieczą
1400kg na szafę, 120 kW na szafę

12:1



<https://www.csemag.com/using-fluid-technology-to-address-cooling-limitations-in-data-centers/>

CHŁODZENIE CIECZĄ

- Bardziej efektywna energetycznie infrastruktura
- Chłodzenie cieczą jest 4x bardziej efektywne do chłodzenia powietrzem
- Mniej powierzchni oraz mniejsze zużycie energii



Redwood Shores, CA



Abilene, TX

B8

B7

B6

Redwood Shores, CA

B5

B1

B4

B2

B3

**Gas Turbine
Plant**

Existing 200MW Substation

New 1GW Substation

Abilene, TX





Oracle Cloud Infrastructure Deploys Thousands of NVIDIA Blackwell GPUs for Agentic AI and Reasoning Models

NVIDIA GB200 NVL72 systems with NVIDIA Quantum InfiniBand
networking now available from OCI.

April 28, 2025 by Ian Buck



<https://blogs.nvidia.com/blog/oracle-cloud-infrastructure-blackwell-gpus-agentic-ai-reasoning-models/?ncid=so-face-145015>



Generative AI

Modele LLM jako usługi as-a-service w Oracle Cloud Infrastructure

- ✓ Starannie dobrana biblioteka modeli LLM dla przedsiębiorstw
- ✓ Pomaga zwiększyć efektywność dzięki nowym GenAI Agents
- ✓ Umożliwia dostosowanie modeli do własnych potrzeb – fine-tuning

 Meta

Llama-3 70B

Llama-3.1 405B

 cohere

Command R+

Command R

Embed

JUŻ NIEDŁUGO NOWE MODELE AI



PRZYKŁAD: obsługa spraw ze wsparciem AI

AI wspiera organizacje w obsłudze spraw, automatyzując procesy, rozumiejąc dokumenty, kategoryzując dane, realizuje ekstrakcję kluczowych informacji, podsumowując treści i analizując spójność, co zwiększa efektywność i dokładność obsługi spraw oraz szybkość ich realizacji.

- ROZUMIENIE DOKUMENTÓW,
- AUTOMATYZACJA,
- KATEGORYZACJA,
- EKSTRAKCYJA INFORMACJI,
- PODSUMOWANIE,
- ANALIZA SPÓJNOŚCI DOKUMENTÓW

Sygn. akt I C 2060/21



**WYROK
W IMIENIU RZECZYPOSPOLITEJ POLSKIEJ**

Dnia 30 czerwca 2022 r.

Sąd Okręowy w Krakowie i Wydział Cywilny
w składzie następującym:
Przewodniczący : sędzia del. Michał Siemieniec
Protokolant: Izabela Kania

po rozpoznaniu w dniu 30 czerwca 2022 r. w Krakowie
na rozprawie
sprawy z powództwa [redacted]
przeciwko Bank S.A. z siedzibą w Warszawie
o ustalenie i zapłatę

I. ustala, że stosunek kredyt hipoteczny kursem CHF zawarty między [redacted] z jednym z innych zobowiązań z tytułu kredytu hipotecznego, zasądza od pozwanych:

II. zasądza od pozwanych:

Okręg wyborczy Nr 104
Okręgowa Komisja Wyborcza w [redacted]

UPOWAŻNIENIE

Obywatel [redacted] (nazwisko i imię, adres) [redacted]
jest uprawniony do pełnienia funkcji meza zaufania w wyborach do Sejmu Polskiej Rzeczypospolitej Ludowej w [redacted]
[redacted] (oznaczenie komisji wyborczej)

(pieczęć)

Przewodniczący
Okręgowej Komisji Wyborczej

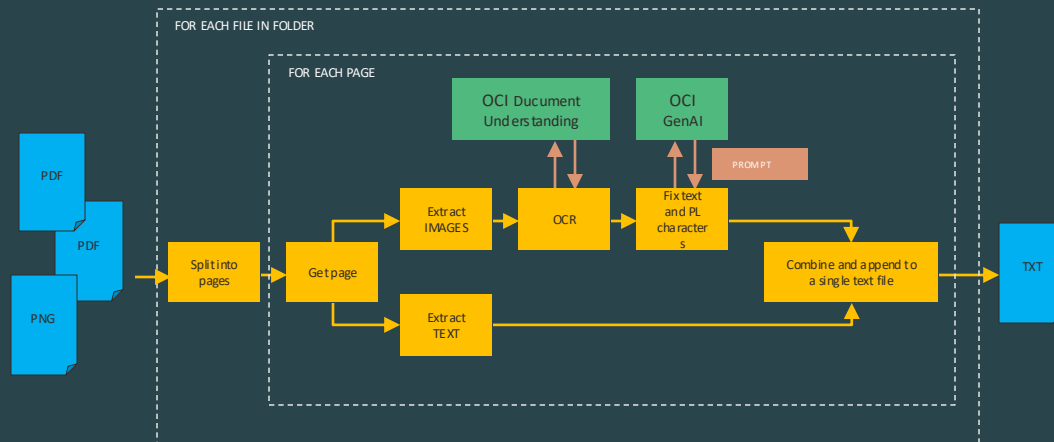
OCI GenAI

```
"file": "data/2024112233445/045_Pismo sądowe.txt",
"category": "Pismo sądowe",
"subcategory": "Wyrok nakazowy",
"summary": "Wyrok nakazowy wydany przez Sąd Rejonowy dla Warszawy Pragi Południe w sprawie",
"data": {
  "sygnatura_akt": "III K 839/21",
  "data_wydania": "14 października 2021 r.",
  "sąd": "Sąd Rejonowy dla Warszawy Pragi Południe w Warszawie III Wydział Karny",
  "skład": "Przewodniczący: SSR Jerzy Oździński",
  "oskarżony": "Jakub <UKRYTE>",
  "imię_oskarżonego": "Jakub",
  "nazwisko_oskarżonego": "<UKRYTE>",
  "imię_ojca_oskarżonego": "<UKRYTE>",
  "imię_matki_oskarżonego": "<UKRYTE>",
  "data_urodzenia_oskarżonego": "05.06.1992 r.",
  "miejsce_urodzenia_oskarżonego": "Warszawa",
  "data_szkody": "28 kwietnia 2021 r.",
  "miejsce_szkody": "Warszawa Wesoła, ul. Fabryczna 6a",
  "opis_szkody": "zabor elektronarzędzi: piła elektryczna Makita wartości 600 pln, dywan",
  "wartość_szkody": "2559 pln",
  "pokrzywdzony": "Adam <UKRYTE>",
  "artykuł_kodeksu_karnego": "art. 278§1 KK",
  "orzeczenie_i": "oskarżony uznany za winnego, wymierzona kara 1 roku ograniczenia v",
  "orzeczenie_ii": "zaliczenie na poczet kary okresu zatrzymania z 22.09.2021 r. (2 c",
  "orzeczenie_iii": "obowiązek naprawienia szkody w całości, zapłata 2559 złotych na",
  "orzeczenie_iv": "podanie wyroku do publicznej wiadomości przez wywieszenie na tabl",
  "orzeczenie_v": "zwolnienie oskarżonego od kosztów sądowych"
}
```

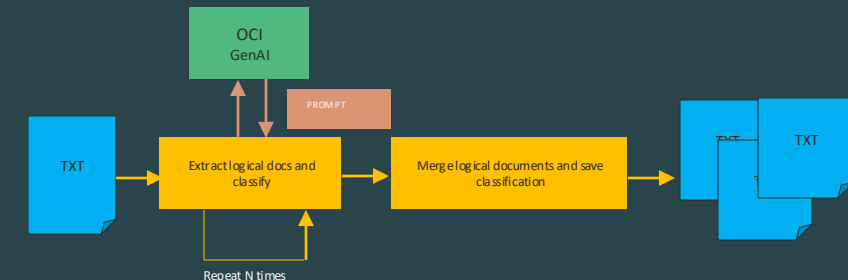


Agent orkiestrujący wywołania LLM i innych usług AI w OCI

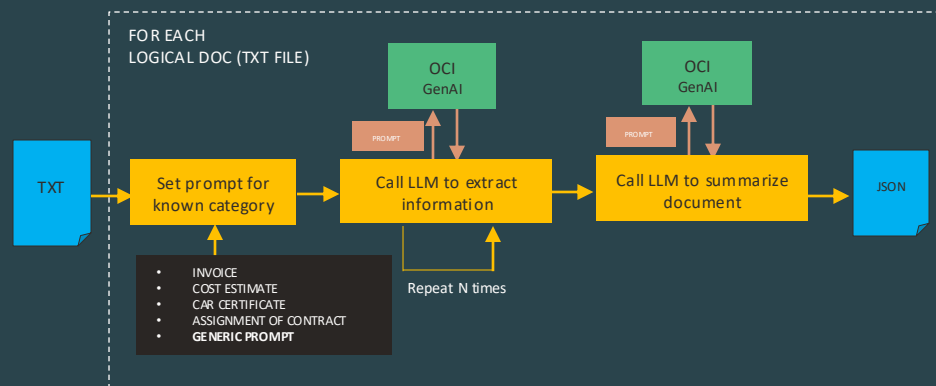
STAGE 1 / OCR and preprocessing



STAGE 2 / MERGE and CLASSIFY



STAGE 3 / EXTRACT



Example prompts used to analyze and process data

What documents do you see in this text?
Provide the page numbers from the original text while maintaining a logical order of pages and classify the document.
Respond only and exclusively in JSON format compliant with the provided schema.
Text:
{context}
Schema:
{schema}

Use the provided text to extract key information.
The text for analysis contains invoice data.
Identify information that constitutes tables or invoice items and include them in the resulting JSON document under "pozycje_faktury" in accordance with the schema.
Match the remaining information in the text to the required data in the schema.
Respond only and exclusively in JSON format compliant with the provided schema.
Text:
{context}
Schema:
{schema}

Language, Speech & Document Understanding

OCI Speech – transkrypcja, speech2text

Model Oracle

Model OpenAI Whisper

OCI Language

Detekcja języka

Klasyfikacja tekstu

Ekstrakcja podmiotu

Ekstrakcja kluczowych fraz

Analiza sentymentu

Maskowanie danych PII i PHI

Document Understanding - OCR

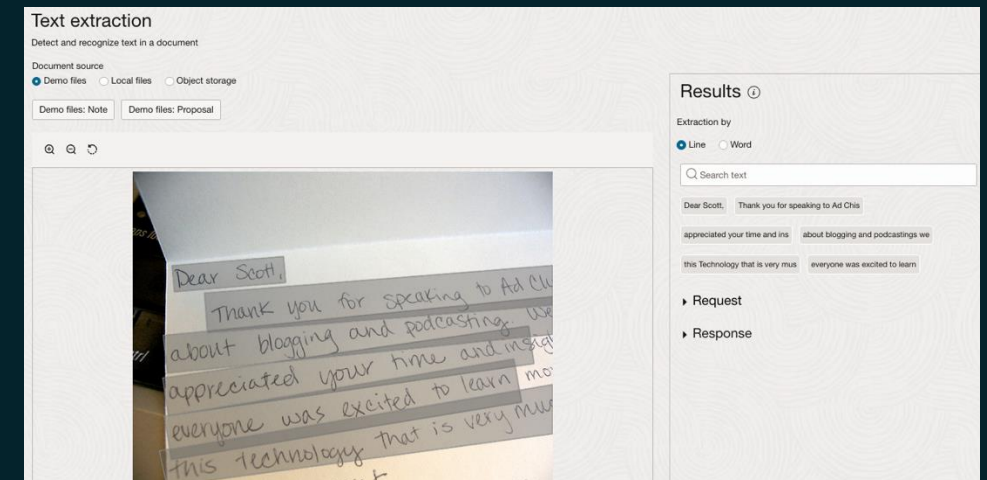
Rozpoznanie tekstu

Rozpoznanie tabel

Rozpoznanie wartości klucz-wartość

Klasyfikacja dokumentu

Możliwość fine-tuning



Agenci AI

Agent AI, po co nam to?

Bazowy model językowy (LLM) sam w sobie nadal nie wystarcza, aby być agentem AI.

"Jaki był mój przepływ gotówki w zeszłym tygodniu?"



Potrzebny dostęp do Twoich danych

"Wyłącz światło o 21:00"



Potrzebne wywołanie Twojego API, aby wykonać działanie

"Przeanalizuj ofertę w kontekście moich regulacji firmowych."



Potrzebne wnioskowanie kierowane Twoimi wytycznymi

"Wyślij zamówienie nr #1134"

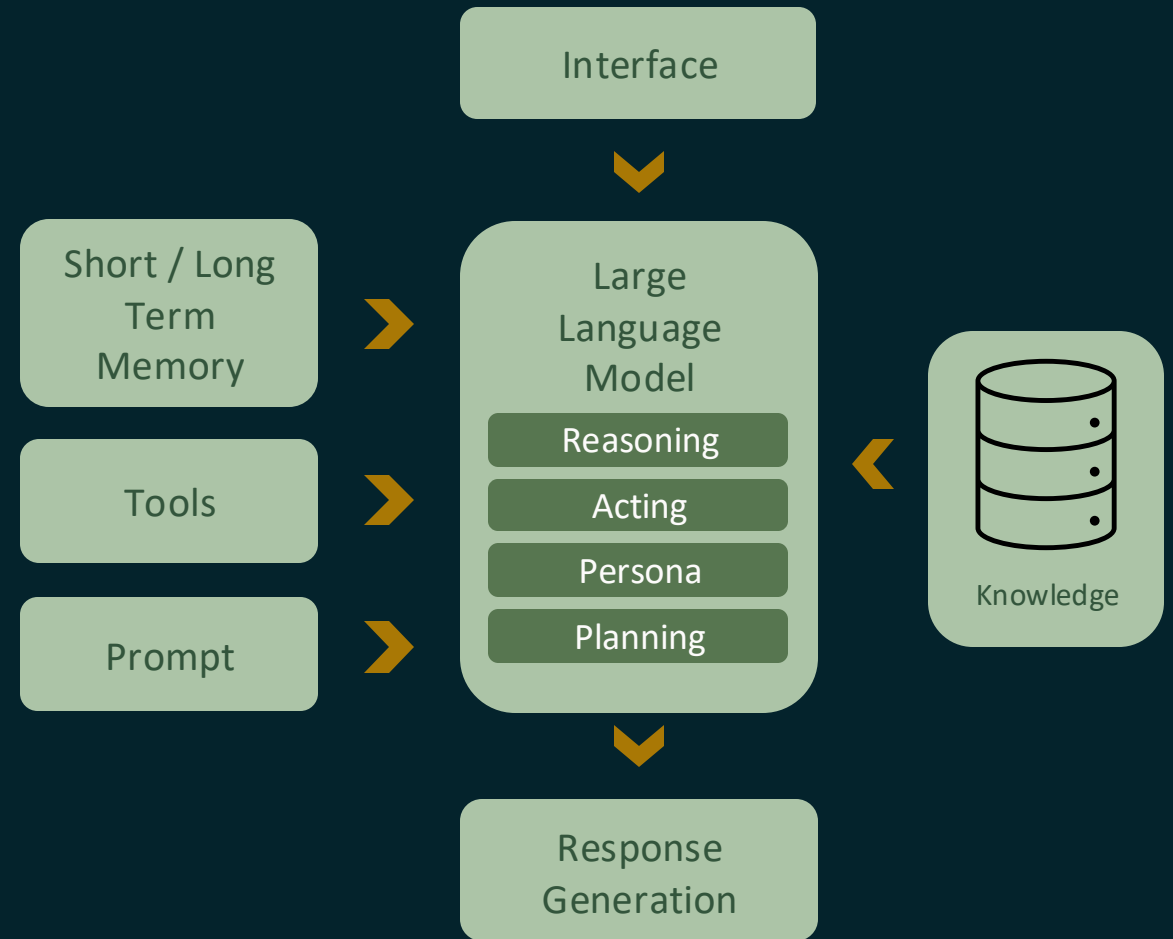


Potrzeba przestrzegania zdefiniowanego przez Ciebie procesu biznesowego

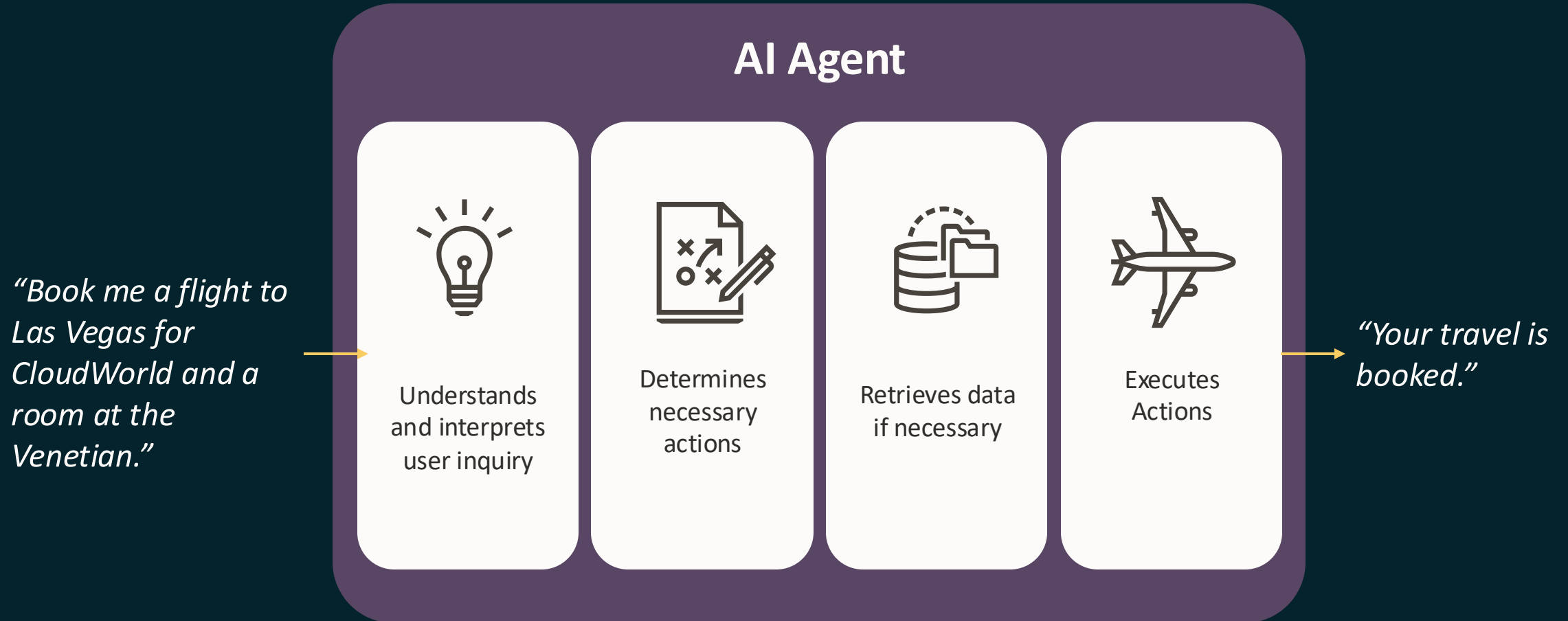
Czym jest Agent AI?

Ogólnie rzecz biorąc, AGENT AI to aplikacja oparta na LLM, która potrafi:

- Wykonywać złożone zadania
- Naśladować ludzki tok rozumowania
- Być skutecznym narzędziem do automatyzacji procesów
- Wykorzystywać zasoby wiedzy w Twojej firmie



Przykład prostego Agenta



OCI Generative AI Agent Platform Components



Agent Core

The "brain" of the Agent that orchestrates different components.



Tools

Interfaces that an Agent can use to retrieve information, generate content and interact with the world

Prebuilt



RAG tool



SQL tool

Custom



Functions



API Calling



- Fully managed cloud native agents
- Tools orchestration
- Multi-turn chat experience
- Context retention
- Custom instructions
- Human-in-the-loop capabilities
- Scalability and tight control over data security and governance

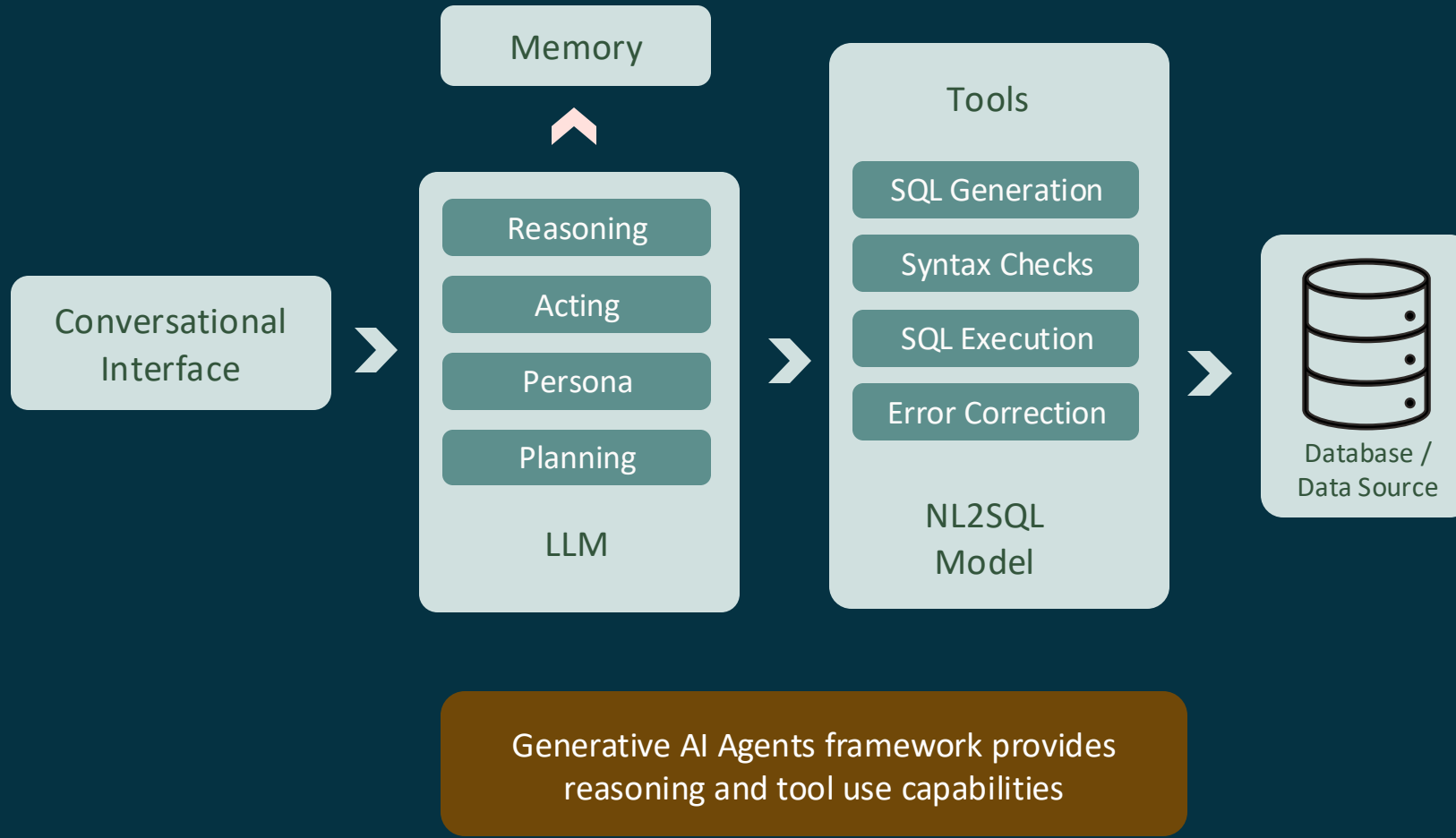
OCI Generative AI Agents: RAG tool

RAG (Retrieval-Augmented Generation) to technika, która łączy wyszukiwanie informacji z generowaniem języka, aby tworzyć bardziej kontekstowo trafne i oparte na faktach odpowiedzi.

Jak może być pomocna:

- **Zwiększona dokładność** – Pozyskuje istotne informacje ze źródeł zewnętrznych, co może zmniejszyć halucynacje i zwiększyć poprawność merytoryczną.
- **Lepsze zarządzanie kontekstem** – Wyszukuje dokumenty na podstawie zapytań, zapewniając modelowi językowemu bogaty kontekst do pracy.
- **Dostosowywalność i skalowalność** – Łatwo dostosować do wyspecjalizowanych dziedzin (np. prawo, medycyna, HR) poprzez wprowadzenie odpowiednich treści do baz wiedzy.

OCI Generative AI Agents: SQL tool



- SQL Generation
- SQL Explanation
- Self-correction
- SQL Execution
- Schema Linking
- In-context Learning
- Custom Instructions
- Additional Configs:
 - Table / Column Descriptions
 - Smaller / Larger Model Choice
 - Dialect (SQLite, Oracle SQL)

Agents on the AI Solution Hub



Chatbot with RAG using OCI Generative AI Agents

Improve insights to make smarter decisions by tapping into real-time data with Retrieval-Augmented Generation (RAG)

Try the OCI Generative AI RAG Agent on the [AI Solution Hub](#)

ORACLE