

Generative AI Part 2

Powering GenAI on Sovereign AI Clouds

Rise of GPU Server

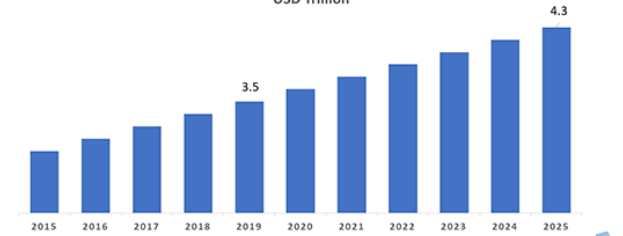
Worldwide Server Market Spending with a **32.1% growth** in Q1 2024 while expected to continue with a CAGR of 12.3% in a Five-Year Period, according to IDC

Unit demand began its recovery in the first quarter of 2024, growing 5.4% year over year, while server spending grew 32.1%, **driven by a continued shift in mix to GPU servers**, especially among hyperscalers and a handful of other large, mostly cloud, IT buyers. The market faces continued challenges throughout 2024, including remnants of impact from the pandemic, historically high inflation, slowdown in economic activity, supply chain disruption and geopolitical conflict.

However, the server market has proven resilient in recent years, as IT infrastructure has become increasingly mission critical for many organizations. An aging installed base is primed for refresh, and the launch of new generation processors late in 2023 will contribute to this cycle spreading into 2024, as the tech transition to new CPU platforms gradually moves forward. The market is expected to grow over the next five years at a 12.3% CAGR. The biggest impact and change to the forecast in this release is the **addition of detail for accelerated servers, including GPU servers**. While in 2023 large cloud service providers consumed most of these systems in an attempt to win the GPU arms race and stockpile inventory, it has become clear that these plans are not slowing down in 2024, and accelerated server adoption will continue growing at a high rate this year and beyond. This means that the server mix between very expensive accelerated servers and more moderately priced accelerated and non-accelerated servers will continue to shift towards the higher price

Global IT Spending Market

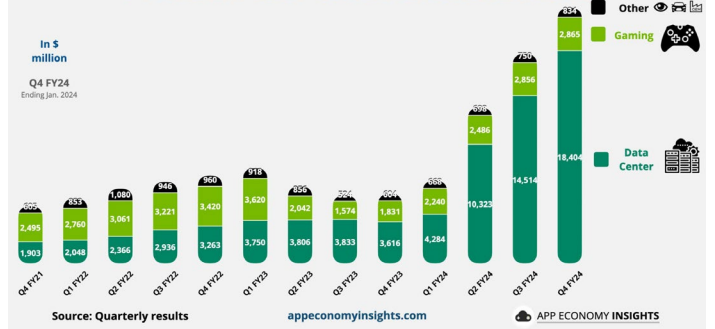
Historical Market and Forecast (2015-2025)
USD Trillion



Source: www.expertmarketresearch.com



NVIDIA Revenue Breakdown



Source: Quarterly results

appeconomyinsights.com

APP ECONOMY INSIGHTS

<https://www.idc.com/promo/servers>

The World Race to Electric Power



HOME > NEWS > THE DATA CENTER CONSTRUCTION CHANNEL

Here's the 5GW, 30 million sq ft data center pitch doc OpenAI showed the White House



Data center projects worth £14bn announced with new UK government AI plan



US data-center power use could nearly triple by 2028, DOE-backed report says

Africa's data centres aim for 2.5GW capacity amidst challenges

The continent faces hurdles such as inadequate electricity supply, regulatory hurdles, and insufficient infrastructure.



GMT+2 • Updated a month ago



SINES DC Becomes Europe's Largest and Most Sustainable Data Center Campus with Groundbreaking 1.2 GW IT Capacity



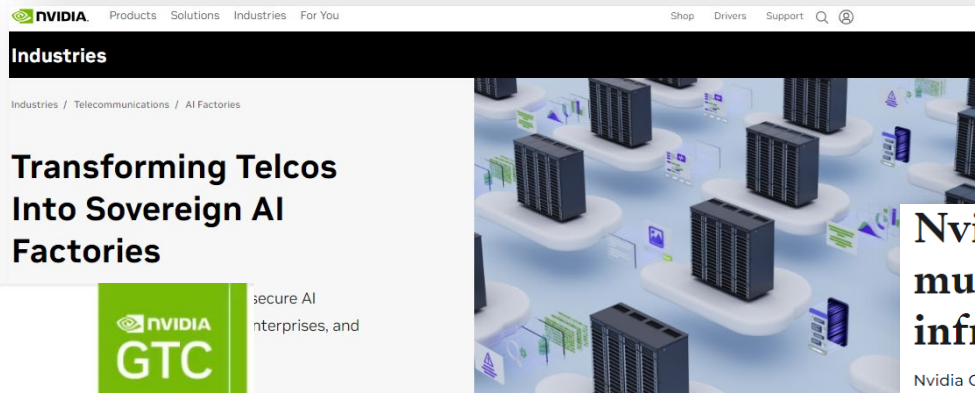
Models designed for a single GPU (Inference)



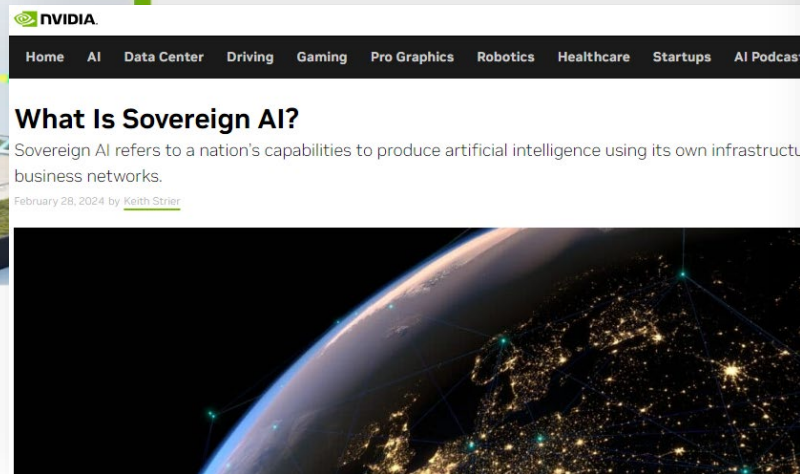
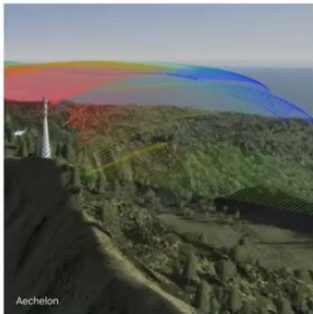
Introducing Gemma 3: The most capable model you can run on a single GPU or TPU



NVIDIA is talking about Sovereign AI



National Transformation With Sovereign AI



Nvidia CEO Huang says countries must build sovereign AI infrastructure

Nvidia CEO Huang says that fears about the dangers of AI are overblown, noting that other new technologies and industries such as cars and aviation have been successfully regulated.



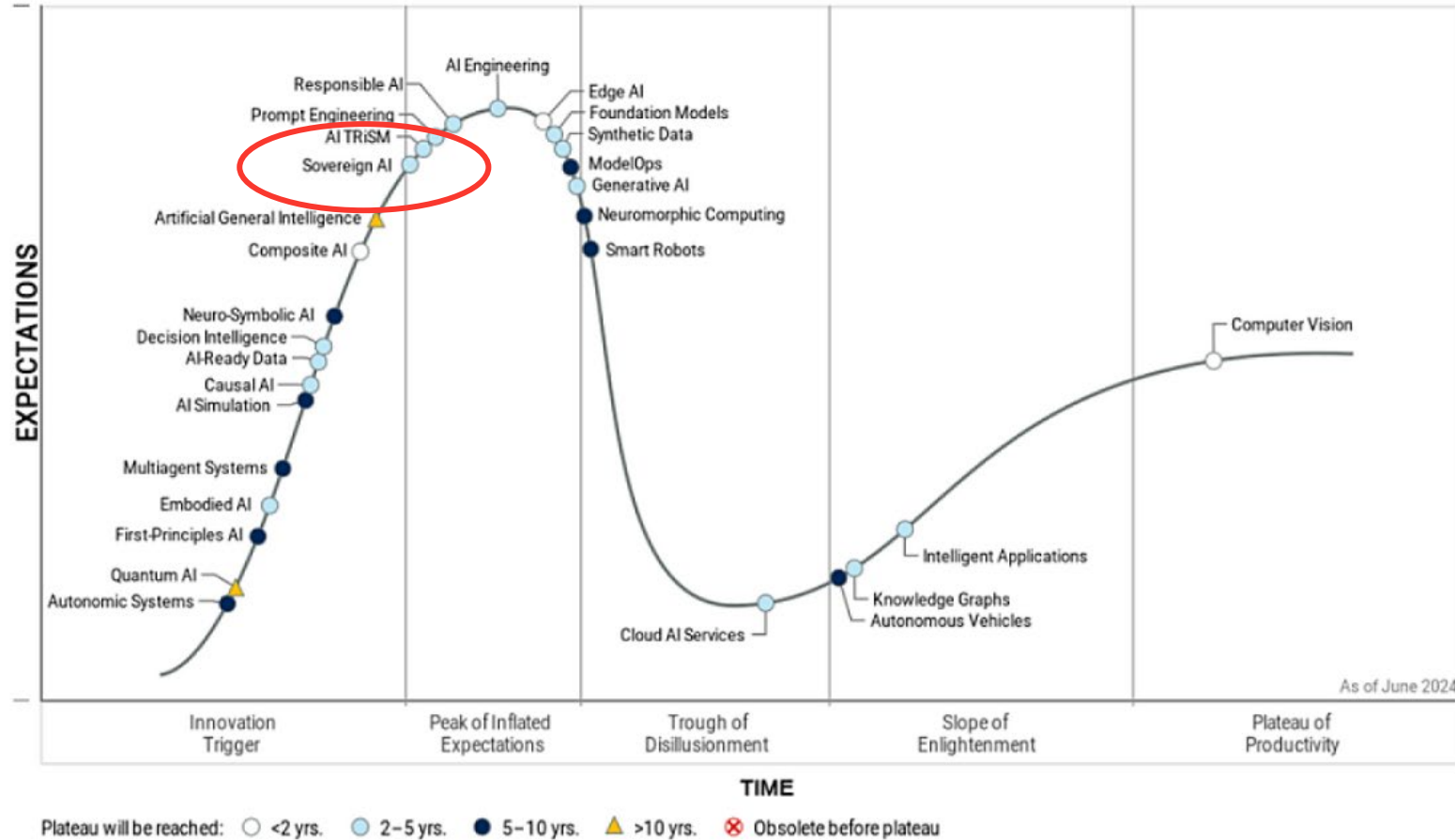
Reuters
Updated On Feb 12, 2024 at 04:01 PM IST



Nvidia CEO Huang

Hype Cycle for AI by Gartner

Hype Cycle for Artificial Intelligence, 2024



















“Larger enterprises and those that desire greater analysis or [use of their own enterprise data](#) with higher levels of security and IP and privacy protections will need to invest in a range of [custom](#) services.

This can include building licensed, customizable and proprietary models with data and machine learning platforms, and will require [working with vendors and partners](#). In this instance, costs can be in the millions of dollars.”

Gartner

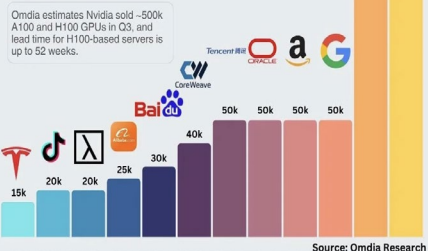
GPU/AI Cloud Providers and the Opportunity

TYPES OF GPU CLOUD

	Description	Companies	Example Products
Tier 1: Hyperscalers	Largest cloud computing providers that operate massive, globally distributed data centers and cloud infrastructure. These companies offer a variety of services and are not solely focused on ML/AI workloads but all types of computing workloads	    	<ul style="list-style-type: none">• <u>Cloud instances</u>: AWS EC2 P3, P4, P5, Azure NCv3-Series, NC 100 v4-Series, GCP Compute Engine,• <u>Cloud + software</u>: Amazon Bedrock, Azure AI Studio, Vertex AI Studio
Tier 2: Specialized Cloud Providers	New generation of specialized cloud providers that focus on providing GPU infrastructure for AI and high performance computing type of workloads	   	<ul style="list-style-type: none">• Crusoe: H100 SXM and A100 SXM instances• Coreweave: H100 HGX, H100 PCIe, A100 NVLink• Lambda: on-demand or reserved H100 instances
Tier 3: Inference-as-a-Service / Serverless Endpoints	Early to late stage startups who offer software abstraction (e.g., sometimes in the form of serverless endpoints) on top of GPU clouds for customers to finetune and deploy/serve models for inference more easily. Some also offer products that target distributed training (Together, Foundry)	Also offers training:       	<ul style="list-style-type: none">• Together Inference, Together Finetuning, Together GPU Clusters• AnyScale Endpoints, AnyScale Private Endpoints

Nvidia H100 GPU Shipments by Customer

Estimated 2023 H100 shipments by end customer.



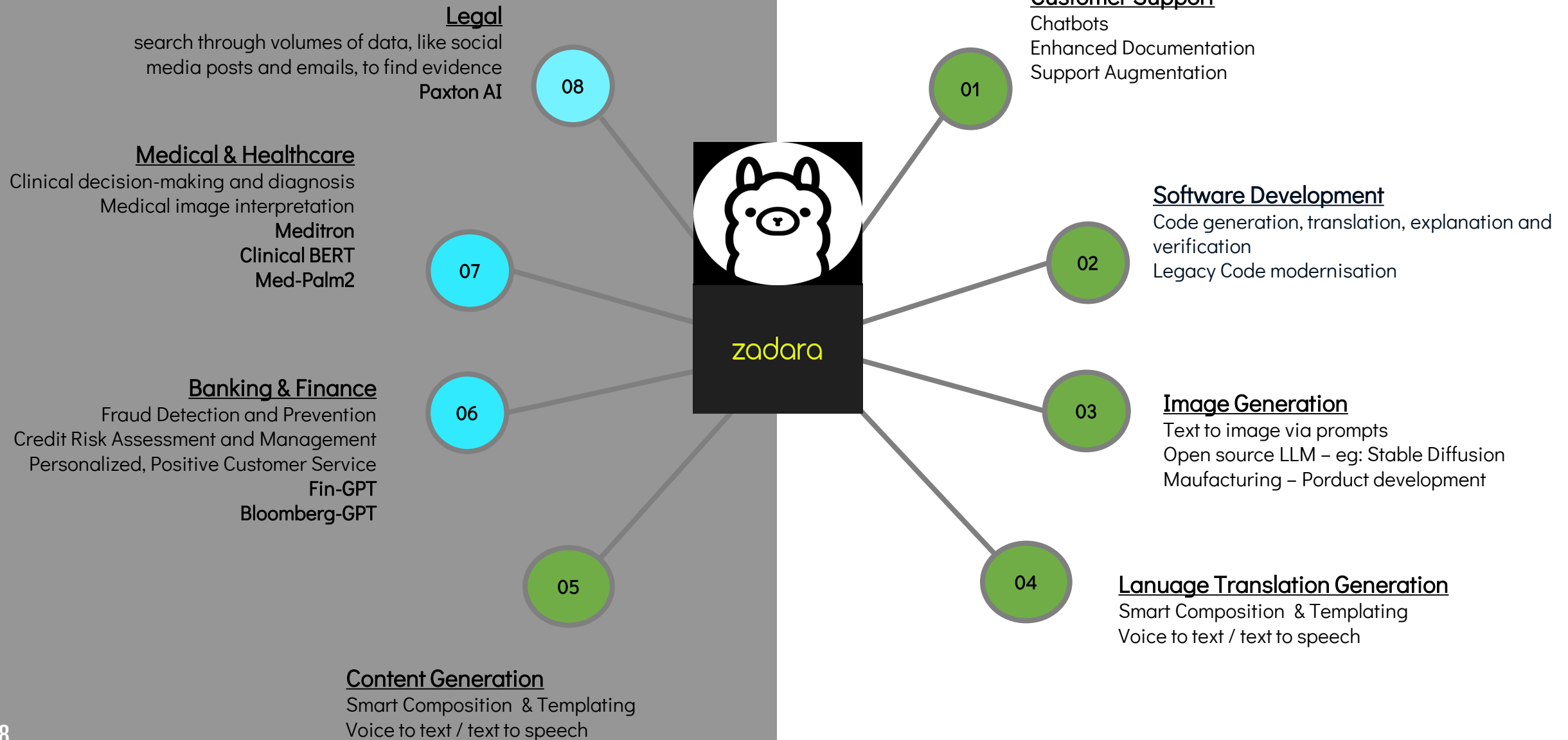
Multi-Tenancy

Inference-as-a-Service

The Opportunity

Domain Specific

Generic



Telco Sovereign AI Use Case

Customer Support

- Customer Agent QoS - Voice file analysis - Adherence to script, security and compliance
- FAQ - automated FAQ from RAG Assistants via API calls

System Monitoring & QoS

- Alert triggers to RAG Assistants
- Provide value-added alert content to Support Staff and Customers

Incident Management & Operations

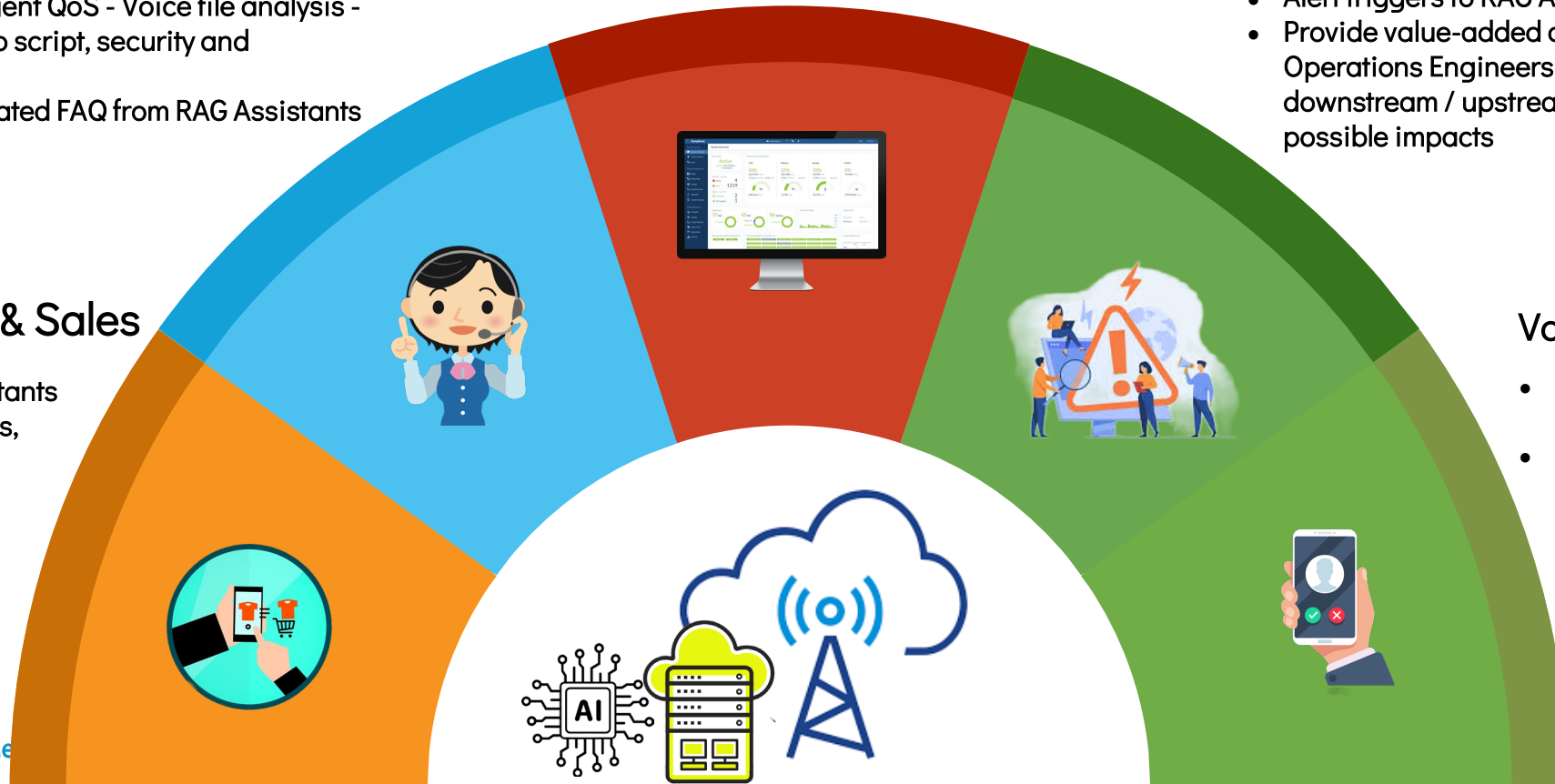
- Alert triggers to RAG Assistants
- Provide value-added alert content to Operations Engineers - Reference to downstream / upstream devices and possible impacts

Customer Product & Sales

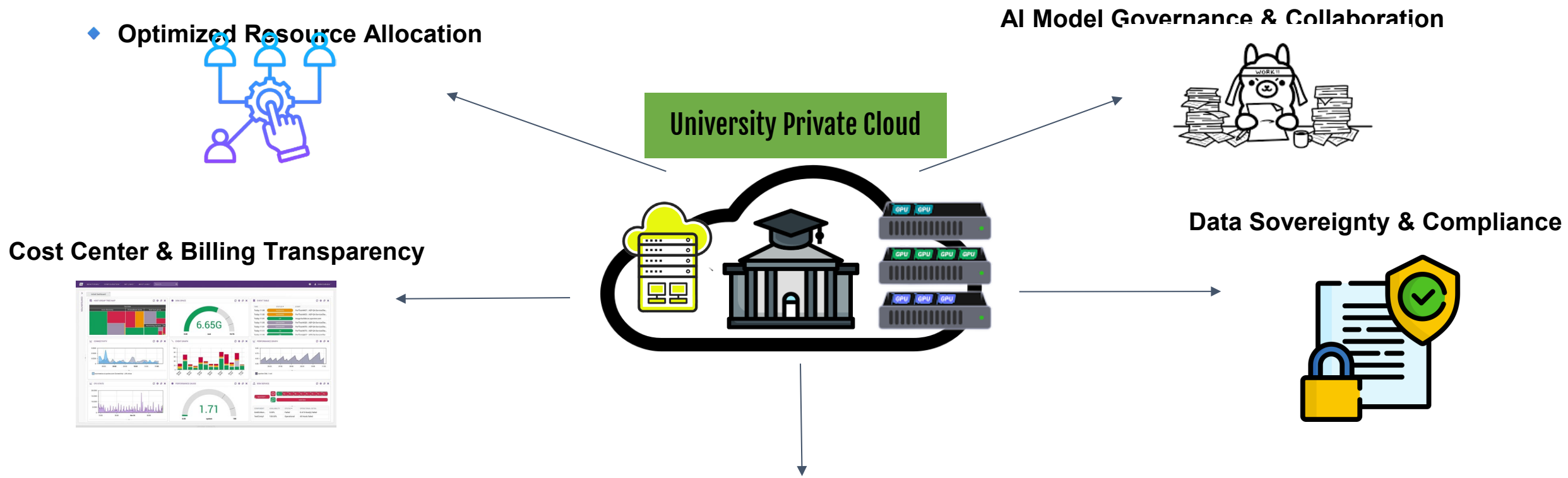
- Chat Agents to RAG Assistants re Products, specifications, pricing, ordering
- Automated Quotes

Voice / Call QoS

- User Calls - Voice file analysis
- Enhanced Customer experience



Sovereign AI in Education



Sovereign AI in Finance & Banking

Fraud Detection / Financial Crime Prevention



- Country Specific
- Enhanced Accuracy
- Payment Fraud
- Identity Theft (user behavior patterns)
- Insider Trading & Market Manipulation

Personalized Financial Services



- Analyze vast amounts of customer data
- Recommendation engines
- Chatbots and virtual assistants
- Predictive Analytics

Risk Assessment and Management

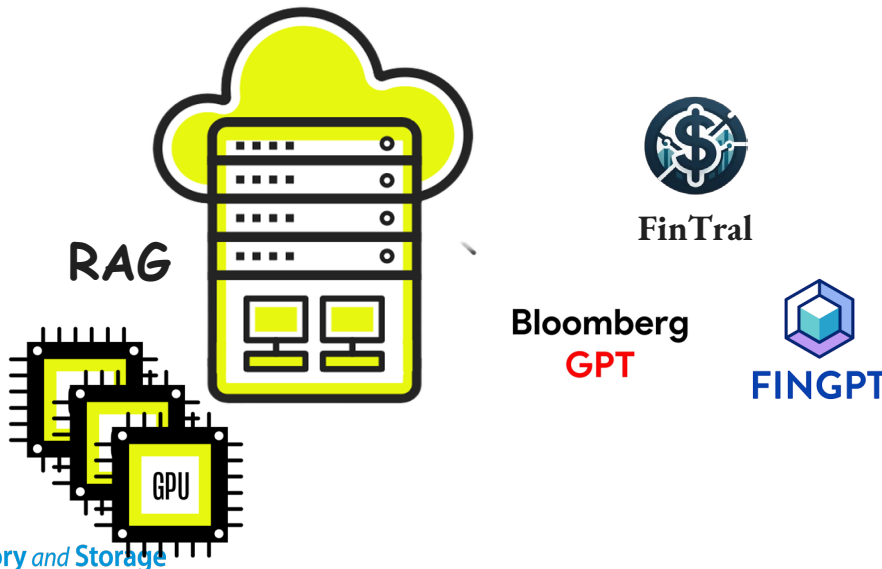


- Detect anomalies in real-time
- Better Investment & Portfolio Management
- Regulatory Compliance

Regulatory Compliance and Reporting



- Robotic Process Automation (e.g. KYC)
- AI for transaction monitoring
- Lower Compliance Costs & Operational Risks
- Audit Trail



Zadara Sovereign AI Platform Bundle

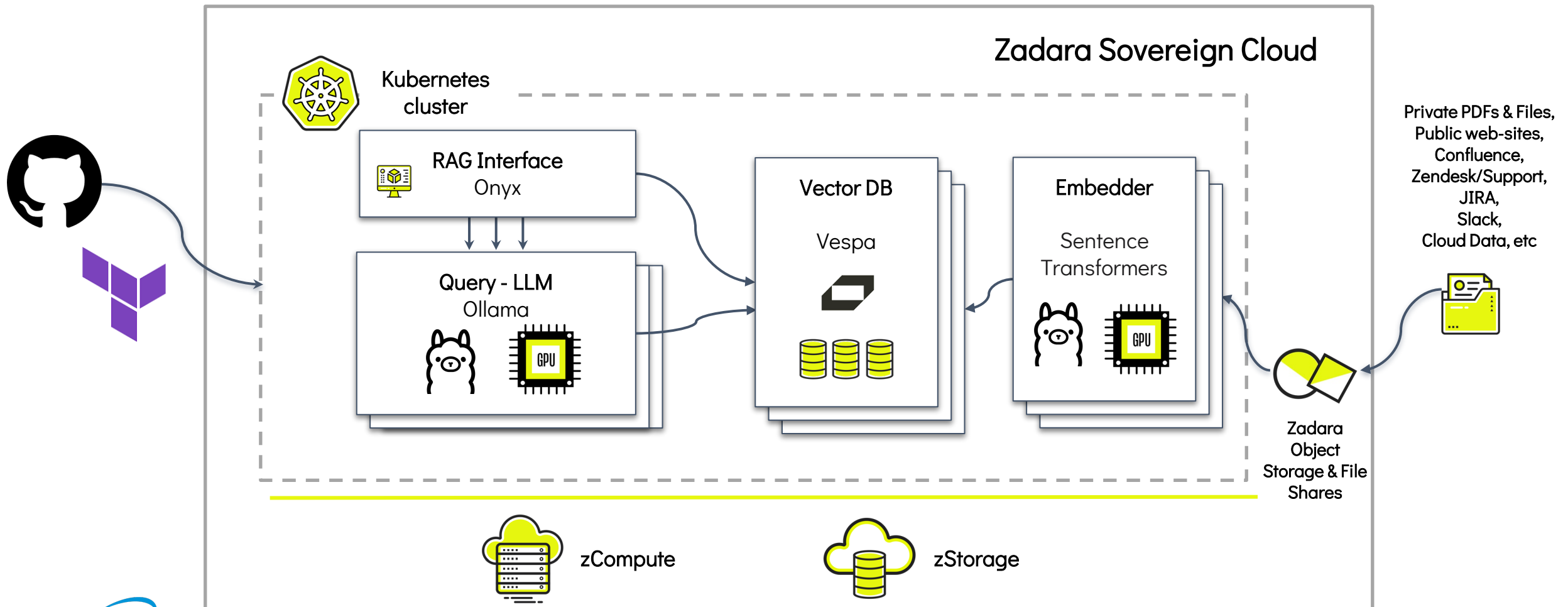
Zadara has decided to help any partner or customer that needs support to run their very own Sovereign AI [RAG](#) platform. This offering consists of the following components;

- [zCompute](#)
 - Including [GPUs](#)
- Deployed using [Terraform](#)
- A [Kubernetes](#) Cluster
- [Onyx RAG](#) Application & Connectors
- [Vespa Vector Database](#)
- [RAG](#) Platform Deployment / Installation
 - Terraform
 - Service
- [RAG](#) Platform Operations & Support
 - Upgrades, Incidents

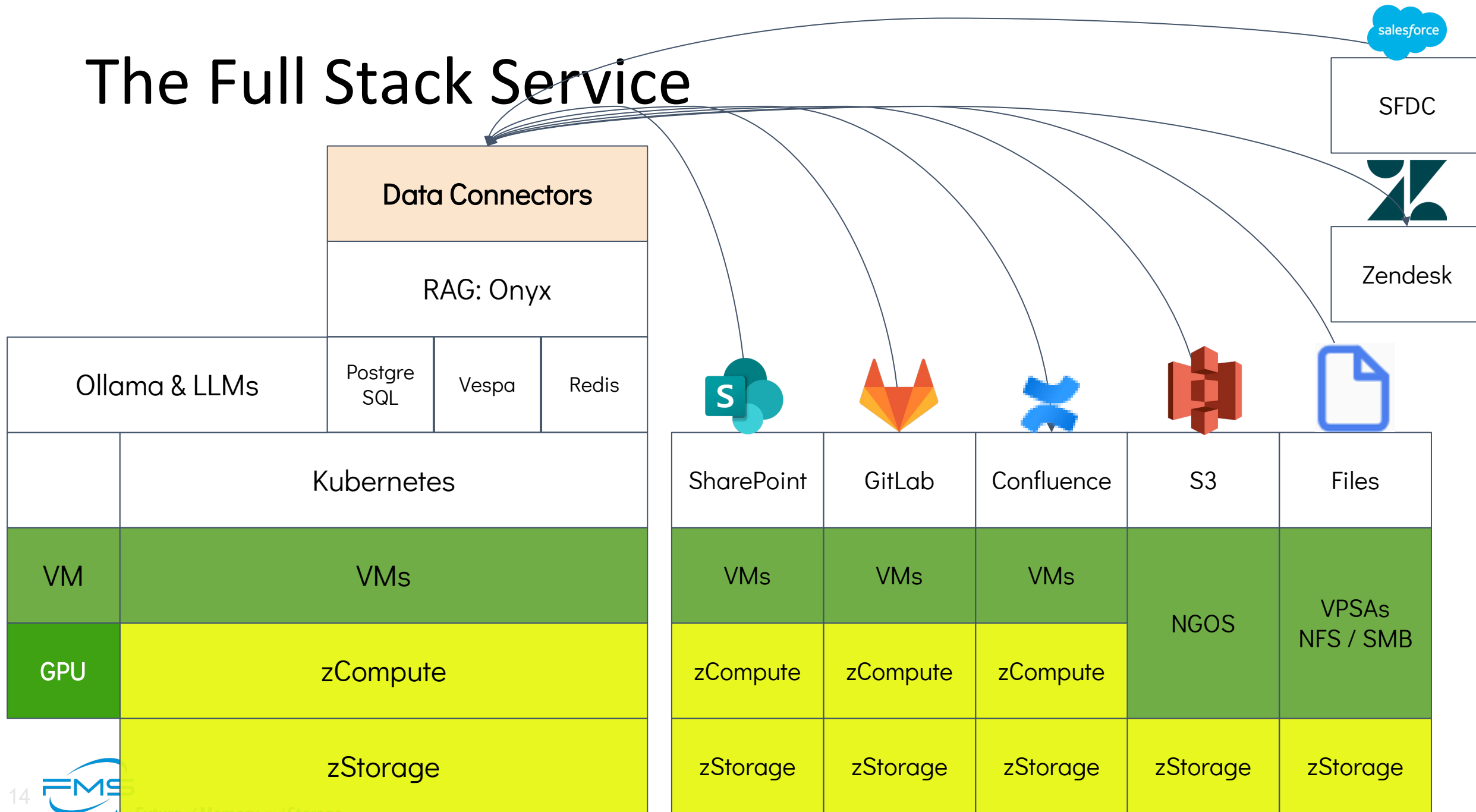
<u>Data Sources</u>	Config Data	<u>Vector DB</u>	<u>LLMs</u>
<u>Onyx</u>	Postgresql	<u>Vespa</u>	<u>Ollama</u>
<u>Kubernetes</u> Cluster K8s / EKS-D / Taikun			
VPC / AWS APIs		DVS	
<u>zCompute</u> + <u>GPUs</u>			
VPSA / EBS			
<u>zStorage</u>			

Zadara Customized Gen AI Framework for MSPs

From General Intelligence to Data Intelligence leveraging Zadara IaaS & tools, RAG, and AI Agents



The Full Stack Service



Customer Support

- Private Chat
 - Trained on all customer private + public data
- Support assistant
 - Generate preliminary proposed answer for new customer tickets

