

Asia's Largest

Al & Cloud Conference 2024

15 - 16, November 2024

Chennai Trade Center, Chennai



It's me

Lalit Rawat

Sr Solution Architect

Microsoft MVP/MCT

SR Solution Architect

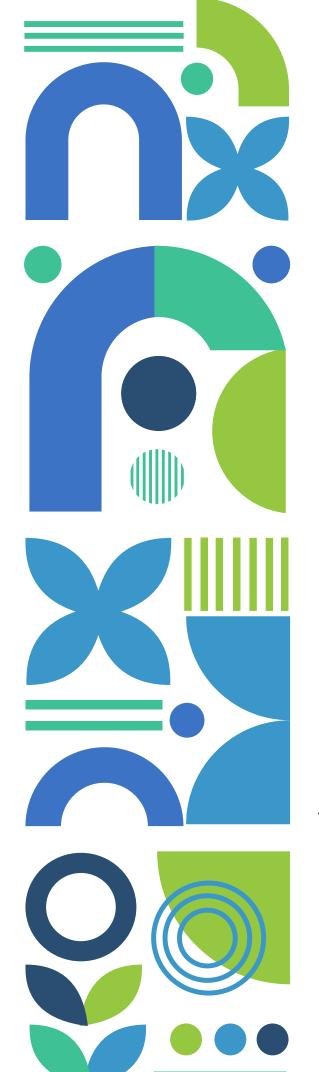
Public Speaker

Cloud Evangelist

Author

Carrier Counselor

Mentor/Coach



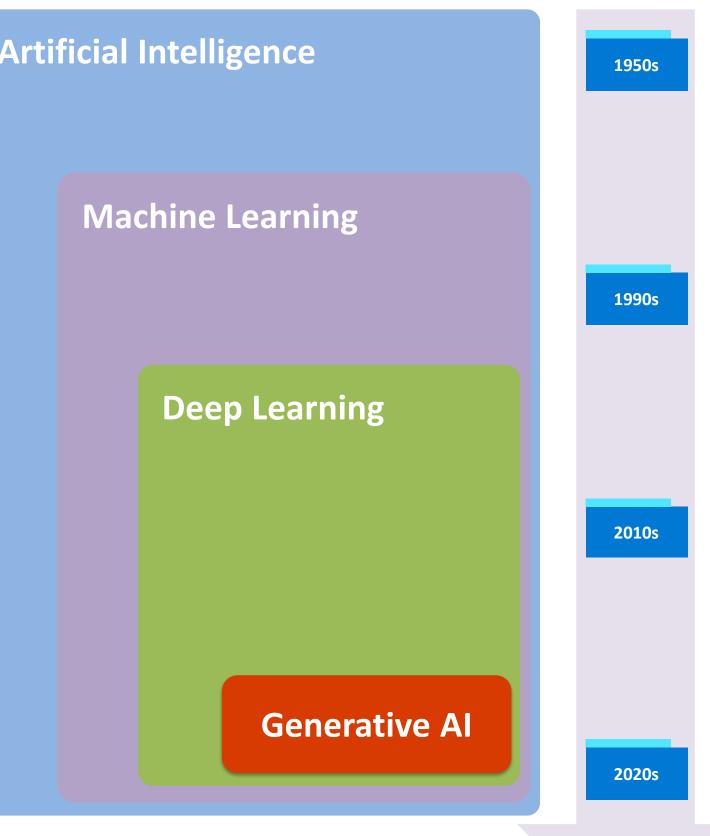
Join the community for Live conversation

https://teams.live.com/l/communit y/FEAxky034tnWMO65As





Generative Al History



Artificial Intelligence

the field of computer science that seeks to create intelligent machines that can replicate or exceed human intelligence

Machine Learning

subset of AI that enables machines to learn from existing data and improve upon that data to make decisions or predictions

Deep Learning

a machine learning technique in which layers of neural networks are used to process data and make decisions

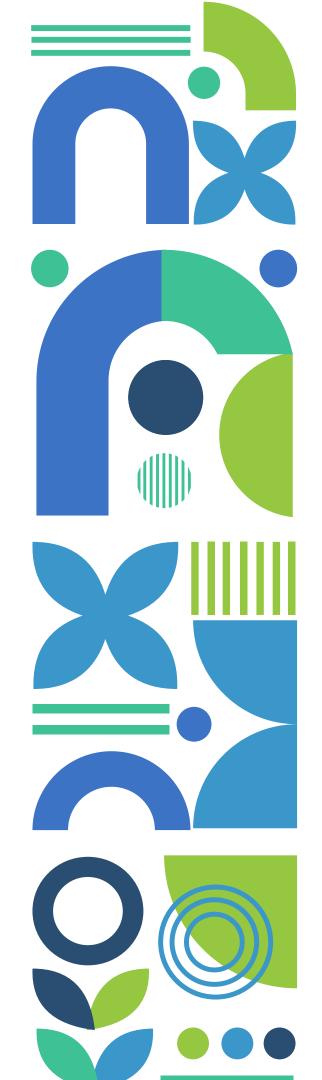
Generative Al

Create new written, visual, and auditory content given prompts or existing data.

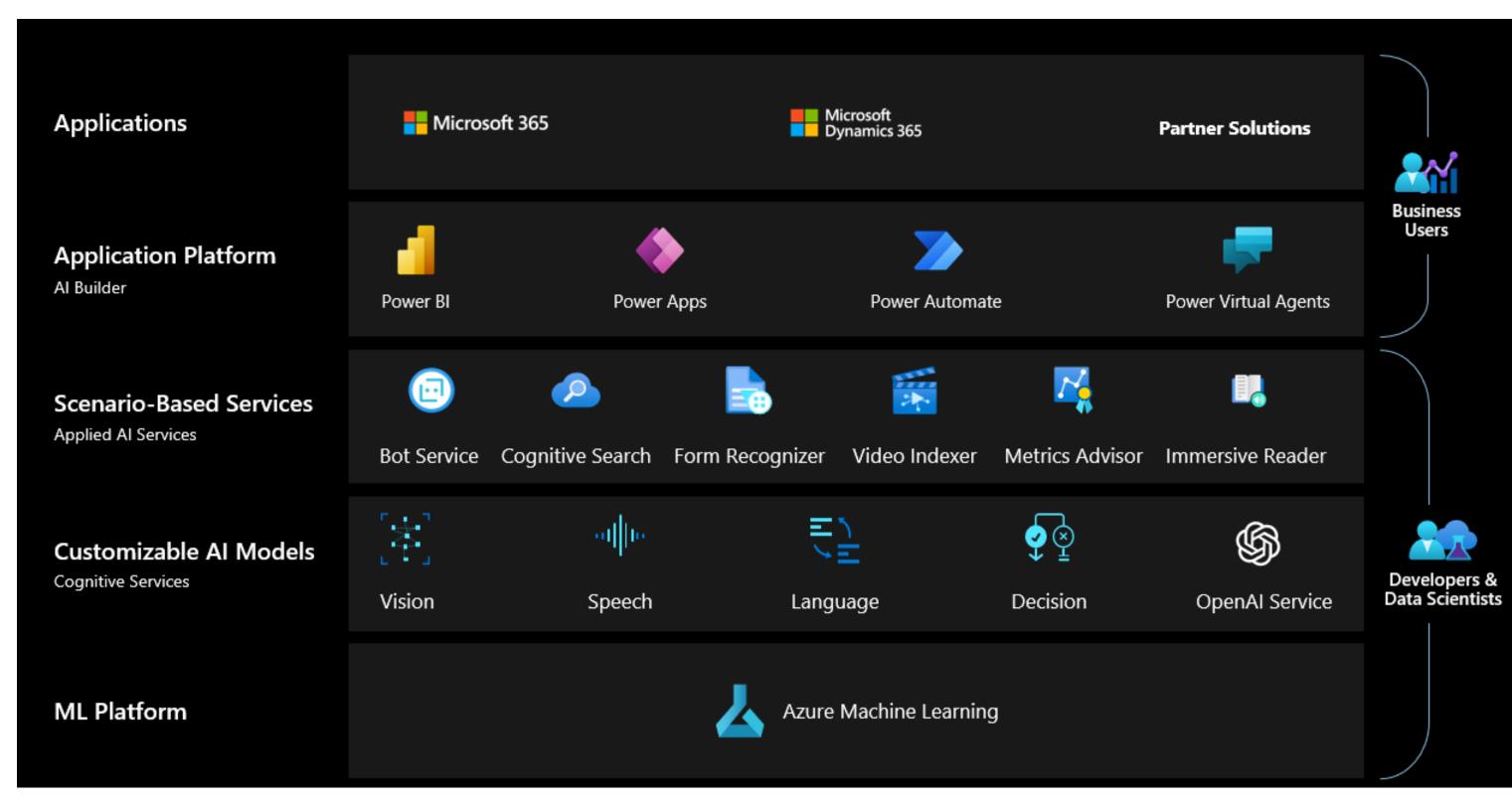
What is open Al

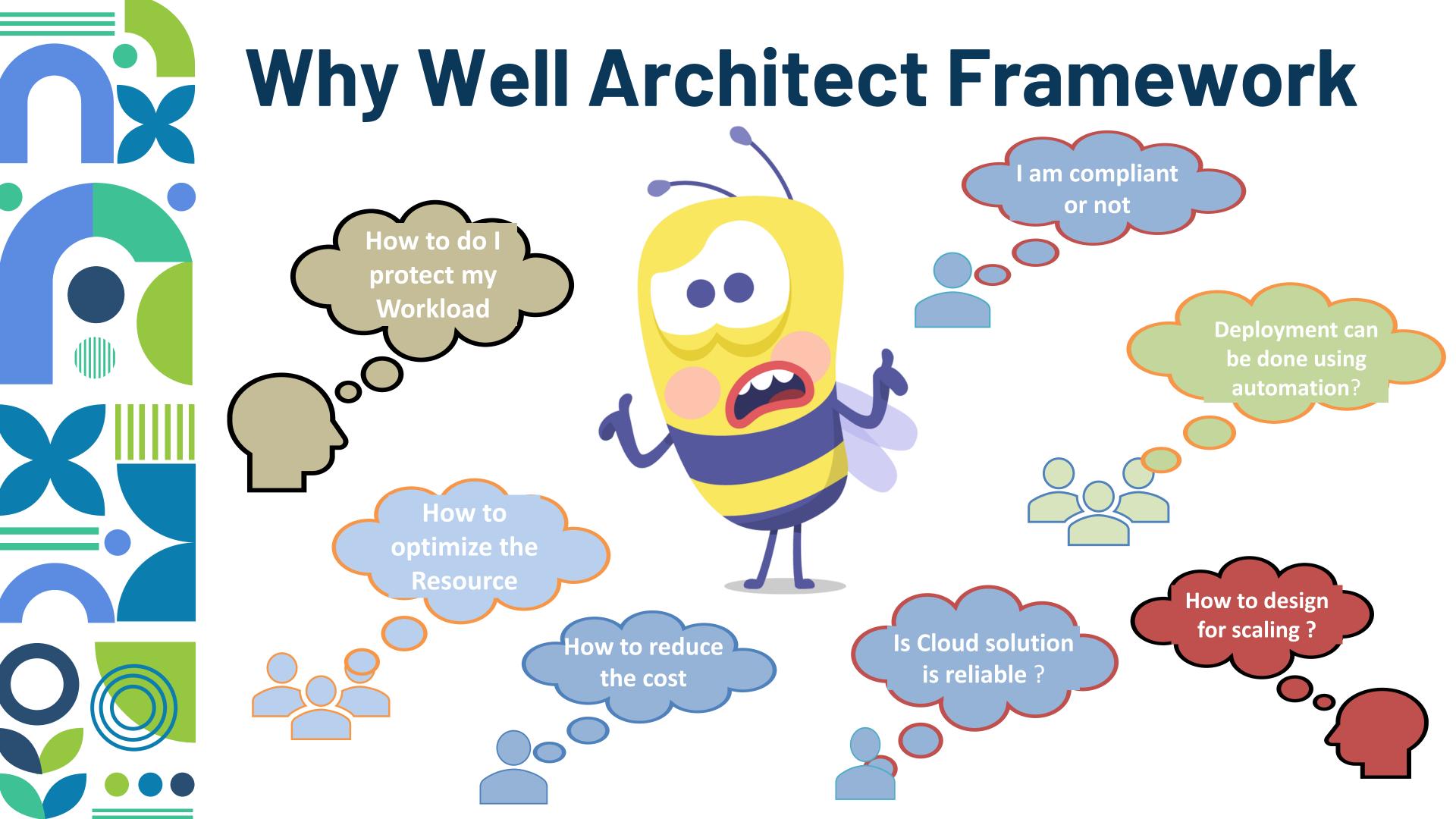
Azure OpenAl Service provides REST API access to OpenAl's powerful language models including o1-preview, o1-mini, GPT-4o, GPT-4o mini, GPT-4 Turbo with Vision, GPT-4, GPT-3.5-Turbo, and Embeddings model series. These models can be easily adapted to your specific task including but not limited to content generation, summarization, image understanding, semantic search, and natural language to code translation.





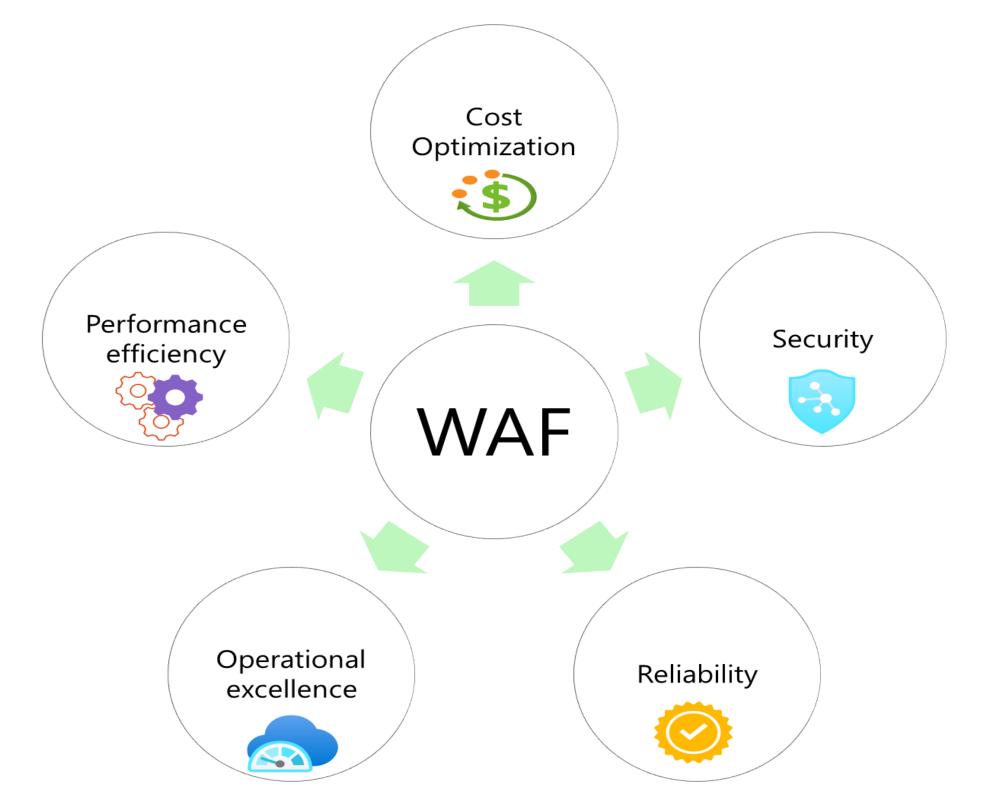
Azure Open Al-Integration



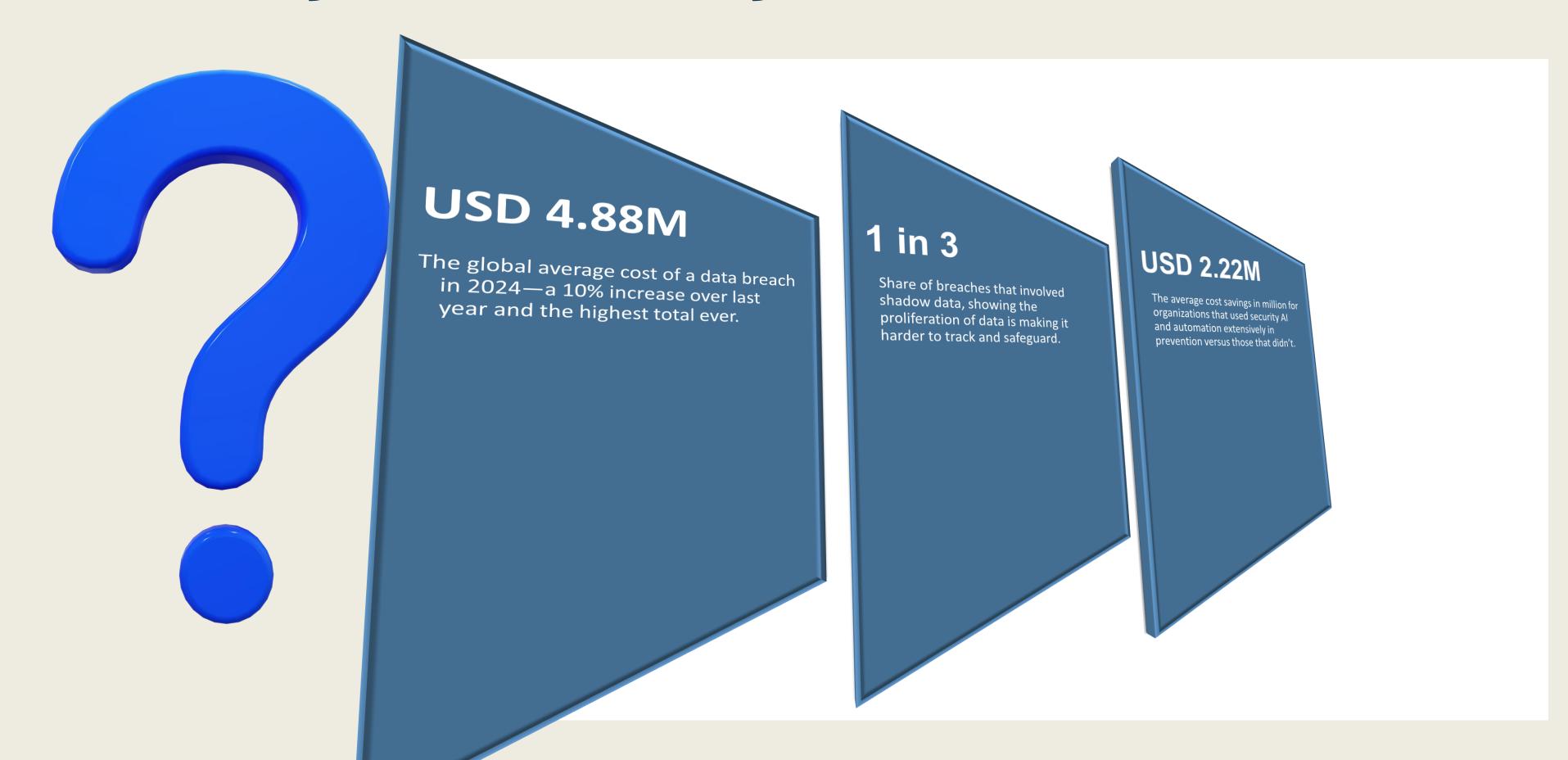




Importance of the WAF in Al Solution Design



Why Al Security is IMPORTANT





Potential Security Risk Associated with Al

Data Privacy Concerns

- Data Leakage
- Inadequate Data Anonymization
- Unauthorized Data Usage



Model Bias

- Bias in Training Data
- Amplification of Existing Biases
- Unequal Representation

Data Integrity Risks



- Tampered Training Data
- Data Poisoning
- Legal Liability

Risk

Adversarial Attacks

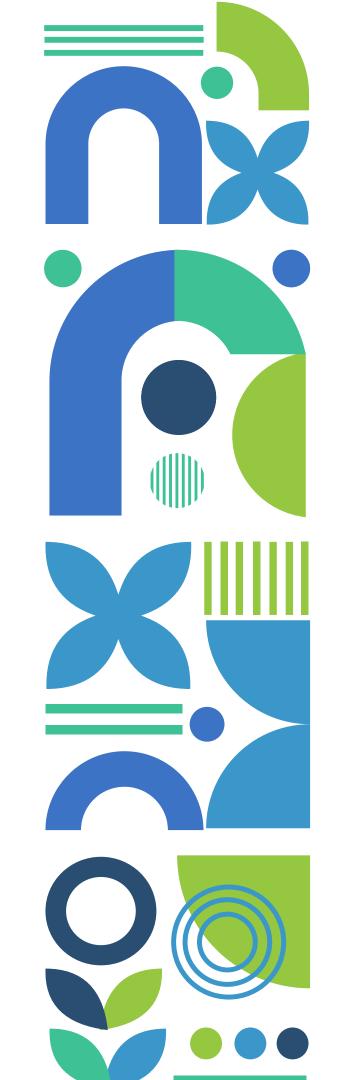
- Input Manipulation
- Model Exploitation
- Poisoning Attacks

Ethical Concerns

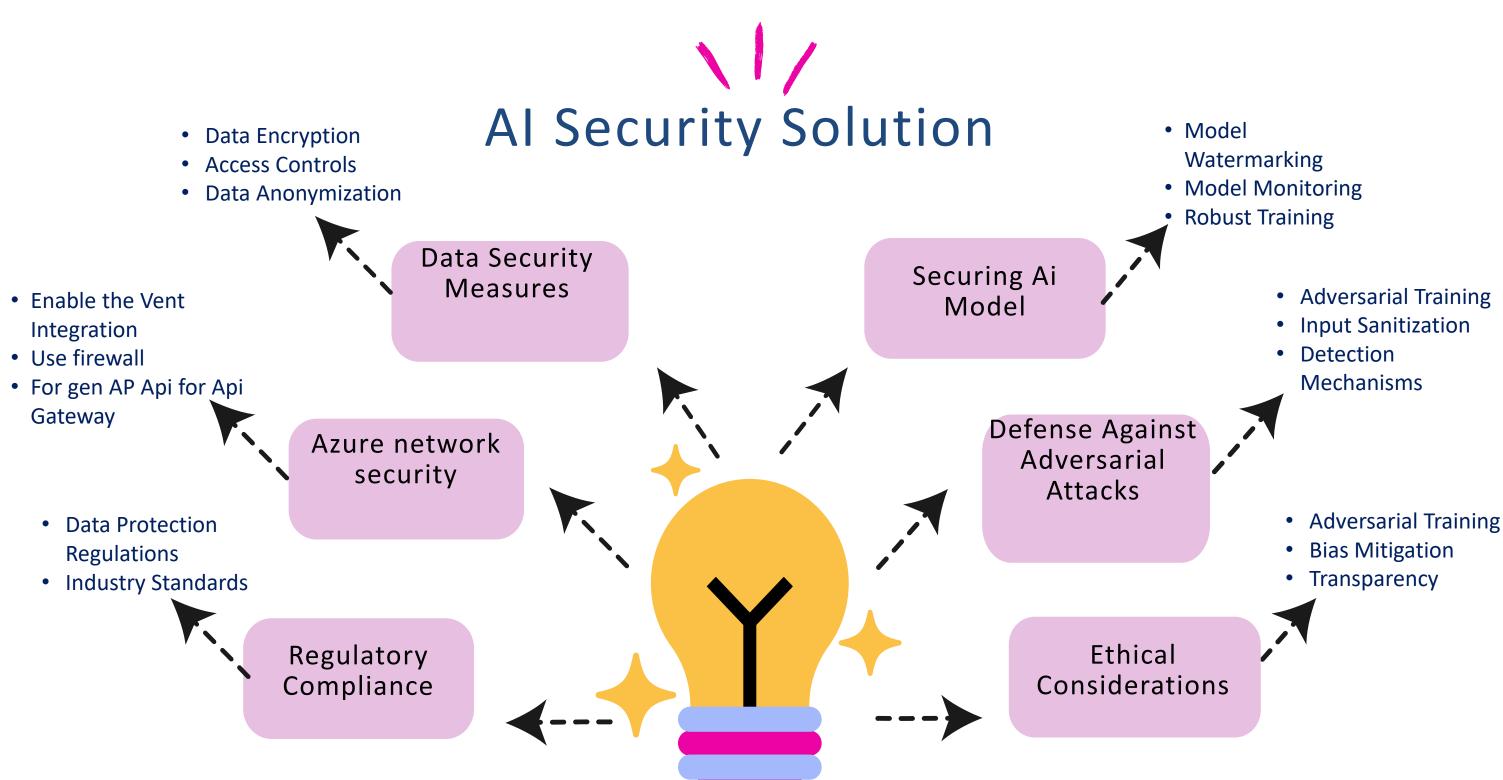
- Misuse of Generated Content
- Lack of Accountability

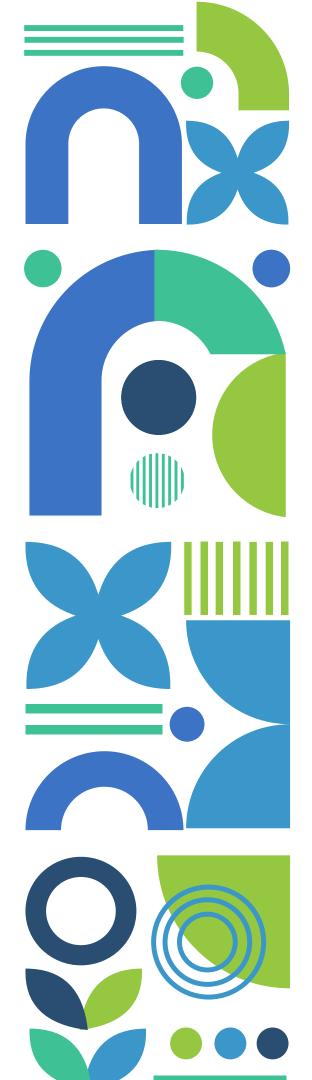
Intellectual Property Risks

- Infringement
- Lack of Attribution



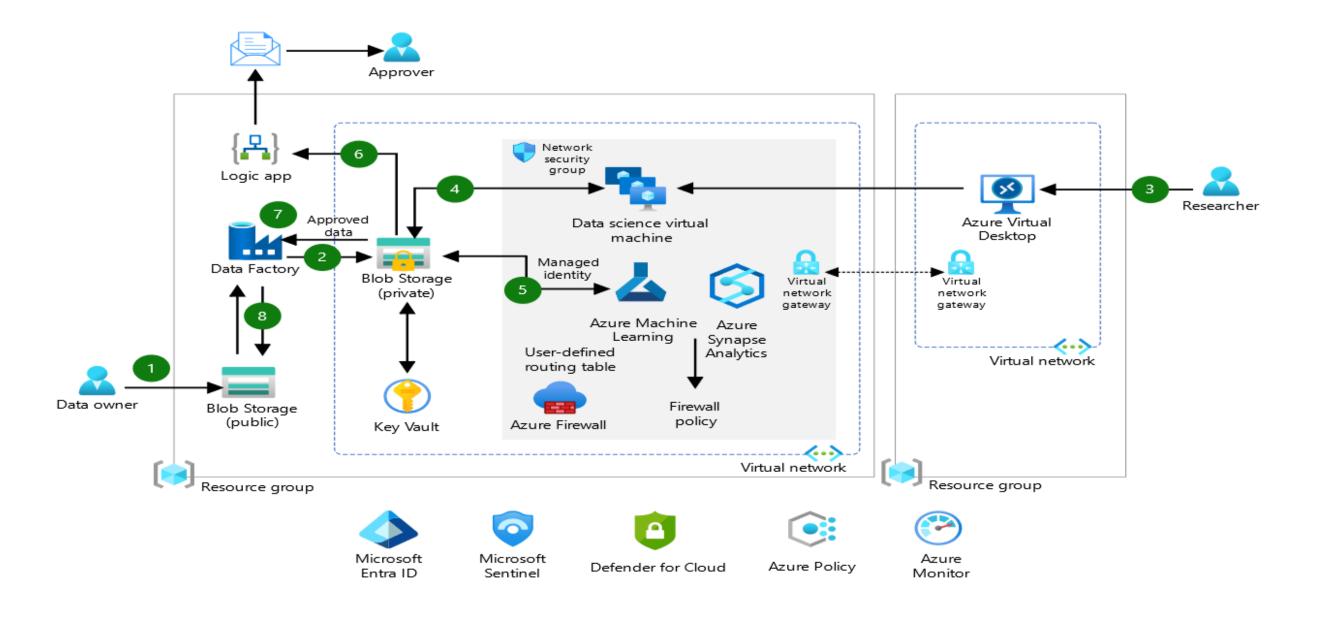
Al Security Solution

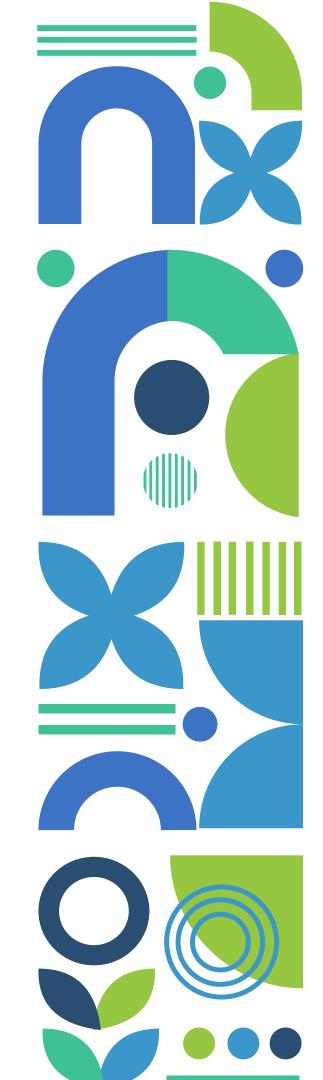




Microsoft

Secure research for regulated data





Azure Open Al Cost Optimization

Choose the Right Model:
Select smaller models (like
GPT-3 instead of GPT-4) when
they meet your needs, as
larger models are more
expensive.

Optimize Token Usage:
Reduce unnecessary tokens in prompts and responses to minimize usage and costs.

Set Usage Limits: Use Azure's quota and budget limits to control costs and prevent unexpected overages.

Leverage Batch Processing:
Process data in batches where possible, reducing the number of requests and optimizing throughput.

-3→

-7→

Scale Resources Dynamically: Use auto-scaling to adjust resources based on demand, avoiding over-provisioning. Monitor Usage Patterns:
Regularly review usage
metrics to identify and
address inefficiencies or
unnecessary usage.

Use Caching: Implement caching for repeated queries to avoid redundant processing and cost.

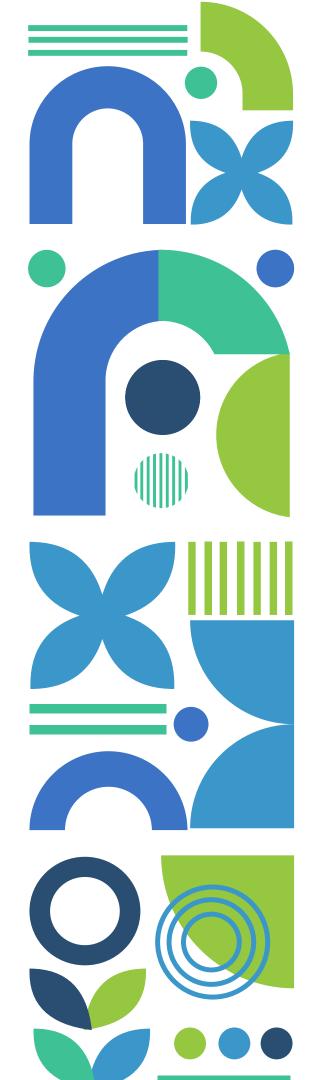
Optimize Prompt
Engineering: Design prompts
efficiently to get concise,
accurate responses with
fewer tokens.

Leverage Azure Discounts:

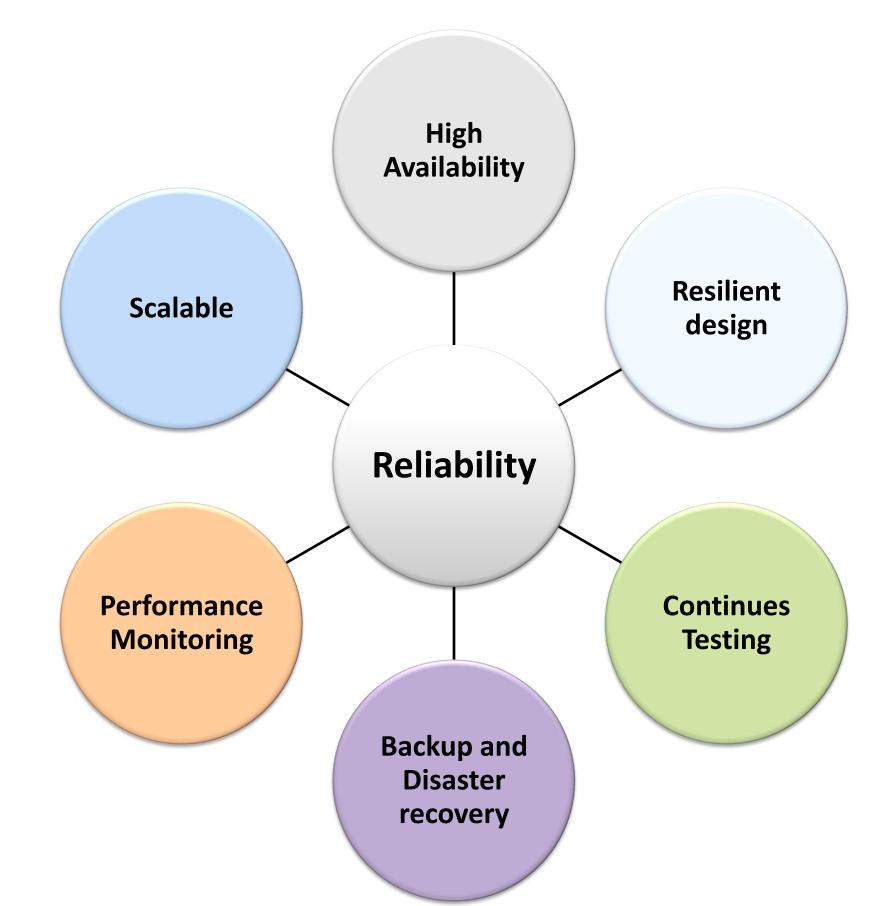
Take advantage of Reserved Instances, Enterprise Agreements, and Azure Hybrid Benefits to lower costs.

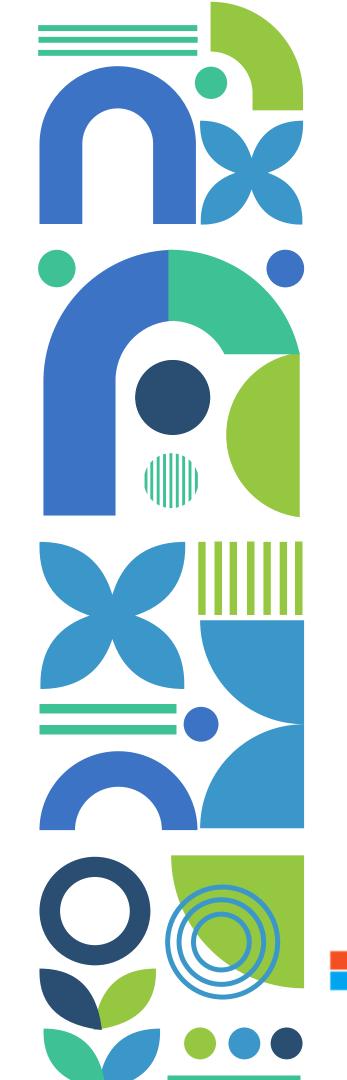
Experiment with LowerPriced Regions: Deploy
services in Azure regions with
lower pricing if it meets
compliance and performance
requirements.

-9-->

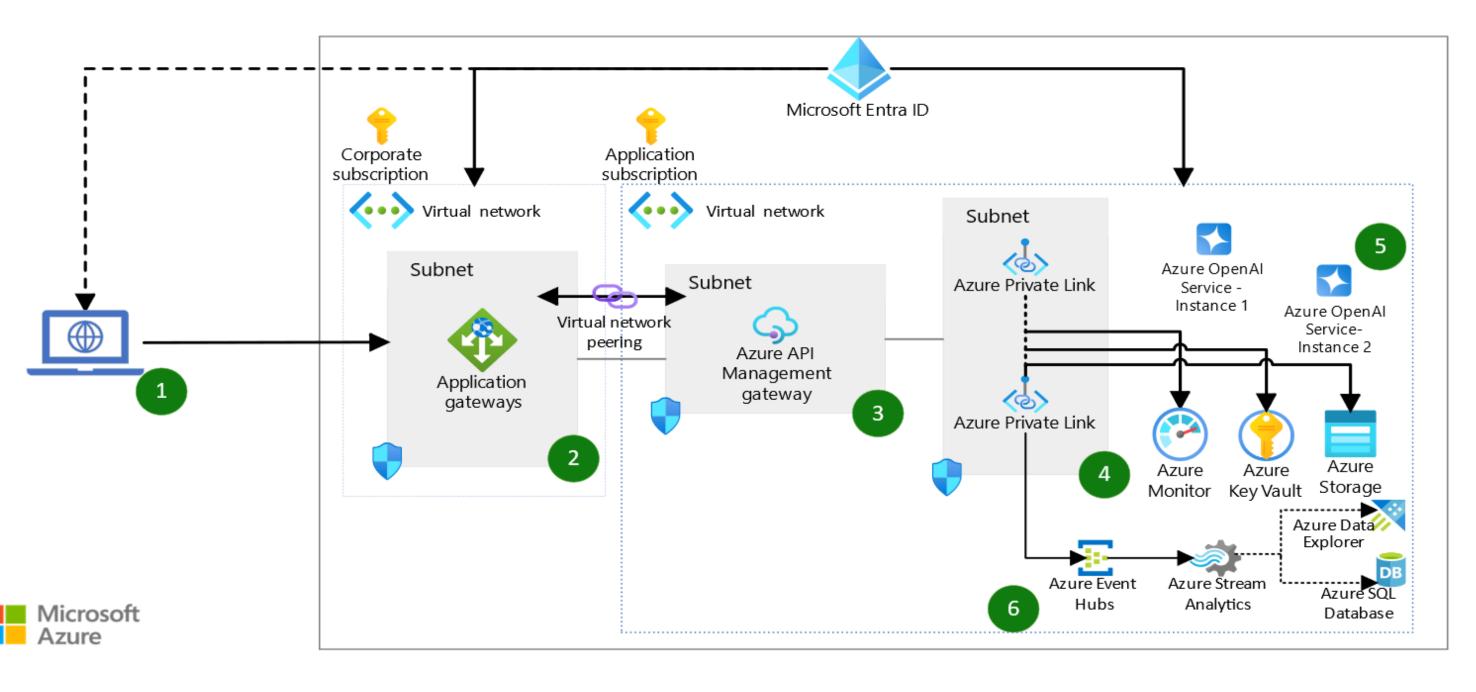


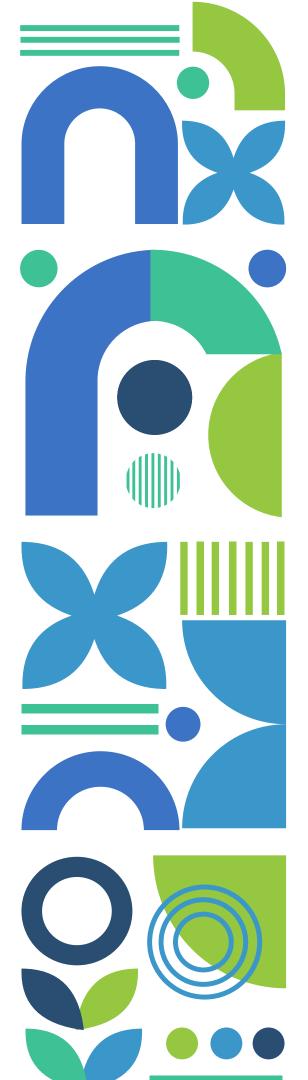
WAF-Solution For Reliability





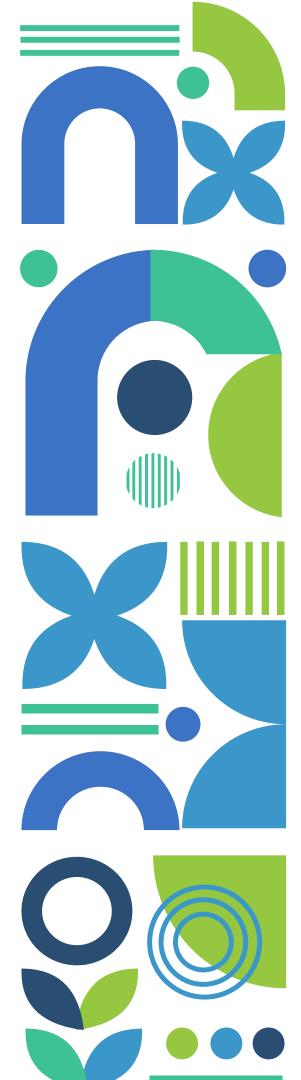
Monitoring for Azure OpenAl models





WAF Solution-Operation Excellence

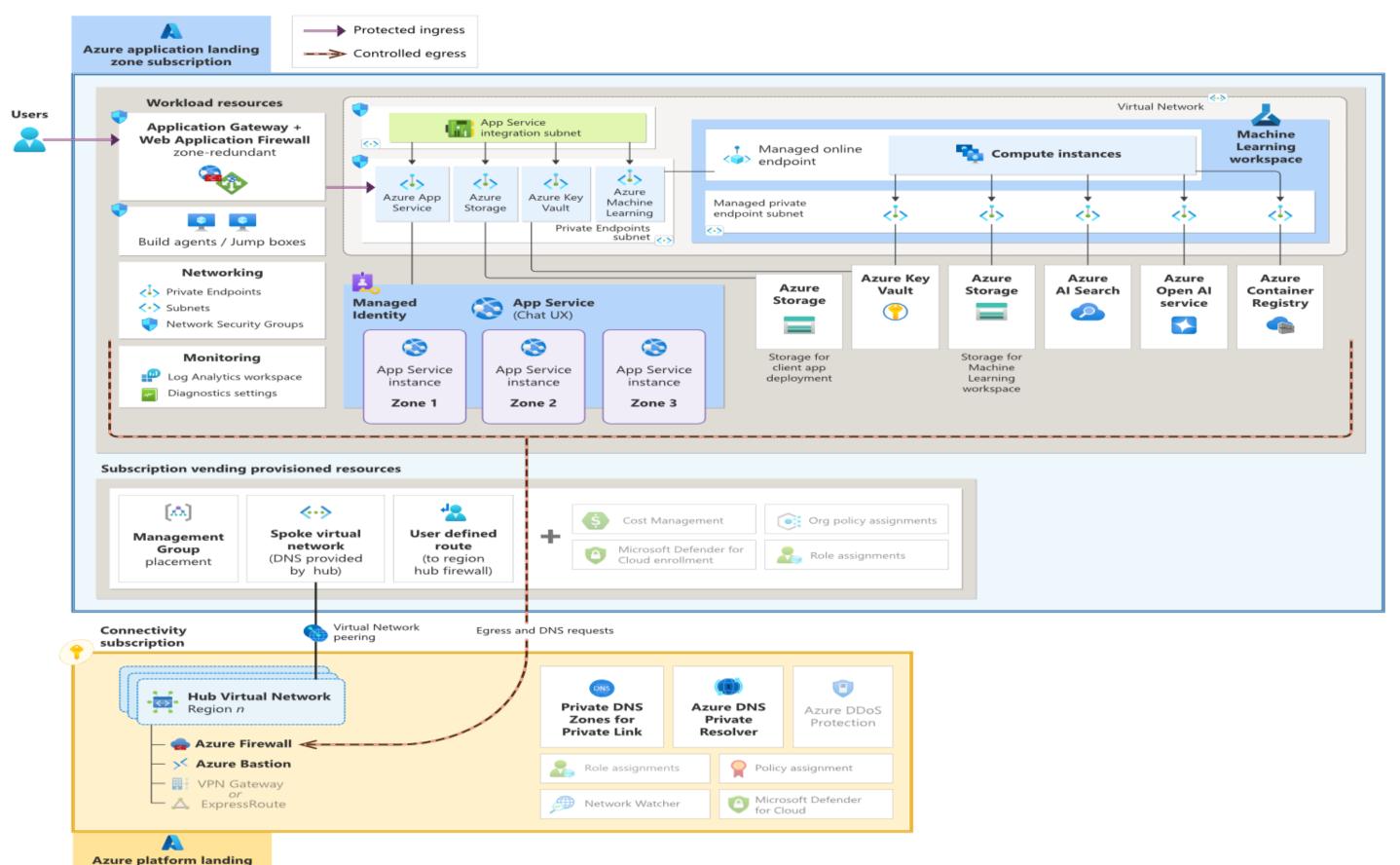




WAF Solution-Performance Efficiency

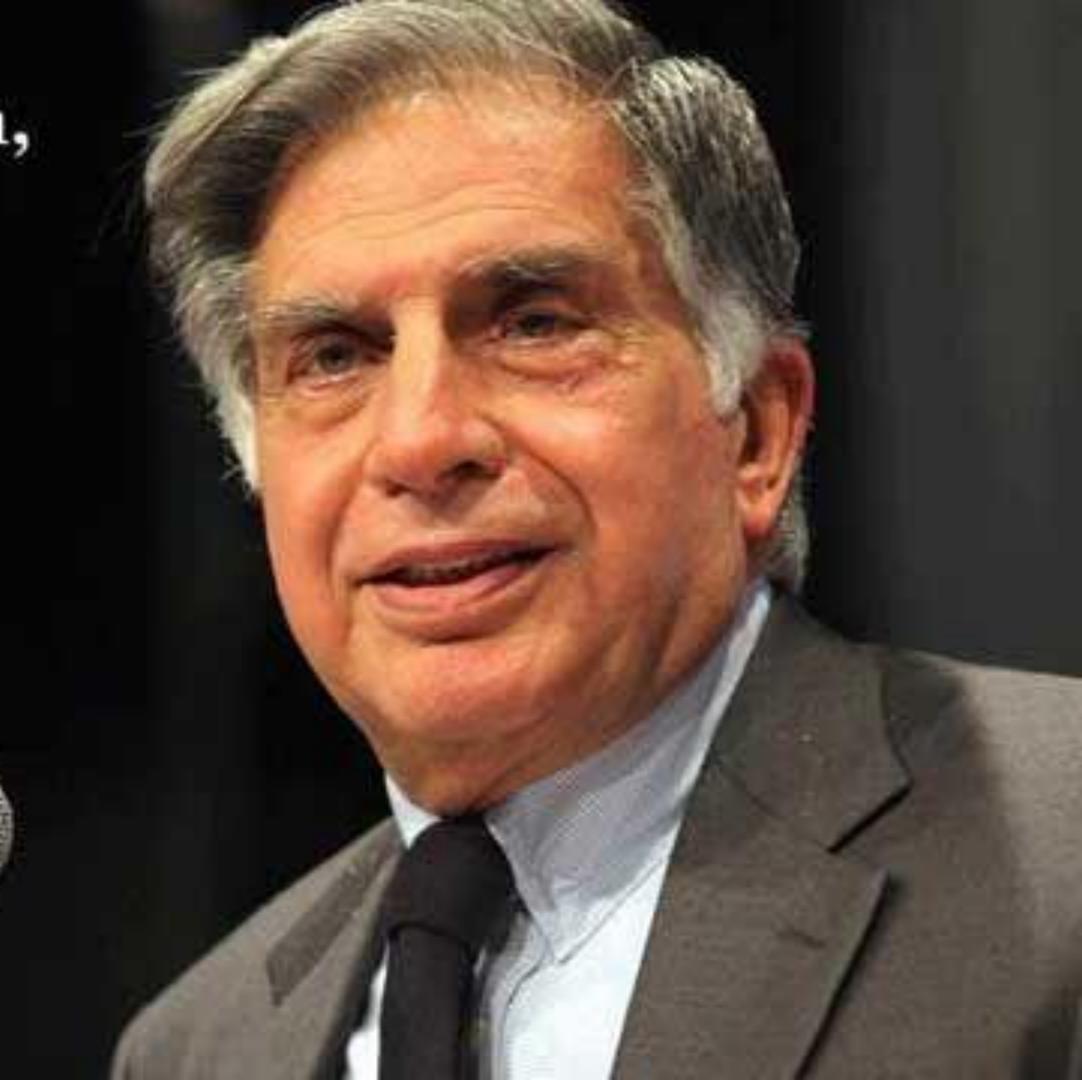
Right Sizing	
Autoscaling	
Al Model Optimization	
Parallel Processing	
Efficient Query Design	
Optimize Al Model usage	

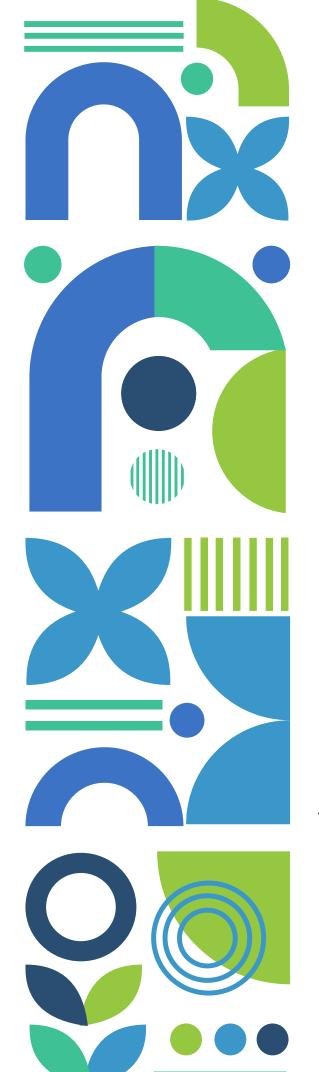
Azure Open Al Landing Zone



None can destroy iron, but its own rust can! Likewise none can destroy a person, but its own mindset can!

- Ratan Tata





Join the community for Live conversation

https://teams.live.com/l/communit y/FEAxky034tnWMO65As





Join Me on LinkedIn



Lalit Rawat Azure MVP-Evangelist, Auth...

Sr Cloud Solution Architect @ PwC | Azure Consultant, MVP, B...







THANK YOU!

Lalit Rawat

Let's build a Secure and Reliable Azure Open Al Solution

