Taiwan Cloud Summit 2025 臺灣雲端大會

ORGANIZED BY iThome

CLOUD FUTURE
BUILT TODAY

從雲原生架構看 AI 助理系統的效能、擴展性與韌性設計

Track J – 2025.7.2

Jerry Hsu

Jerry HSU

Expertise

Problem solving

SQL Server Performance tuning

CDN / WAF optimization

Hybrid cloud solution

Design Thinking coach

Experience

華碩電腦 技術服務部 lead

華碩電腦 系統開發課 lead

東森購物 資深資料庫管理師

淡江大學資訊處 技士

Community

2024 東吳大學 設計思維與AI

2024 SQL Summit Speaker

2022 SQL Summit Speaker

2021 SQL Saturday Speaker

籌辦台灣第一屆 SQL Saturday 年會

TW SQL PASS 8th 幹部

Agenda

我們想要建造的 AI Assistant 是？

雲原生架構如何支持這樣的需求？

- 系統韌性
- 高可用性
- 效能優化
- 安全防禦

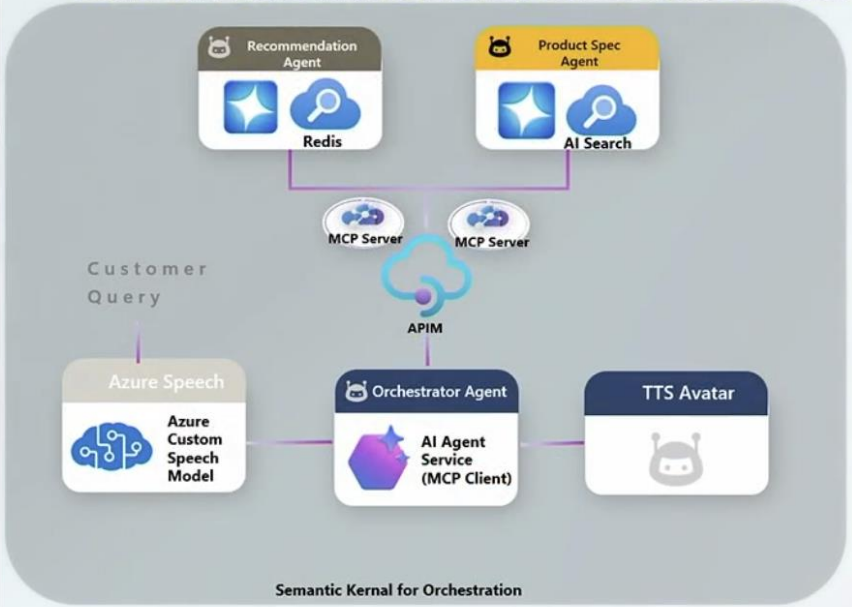
接下來的機會與挑戰

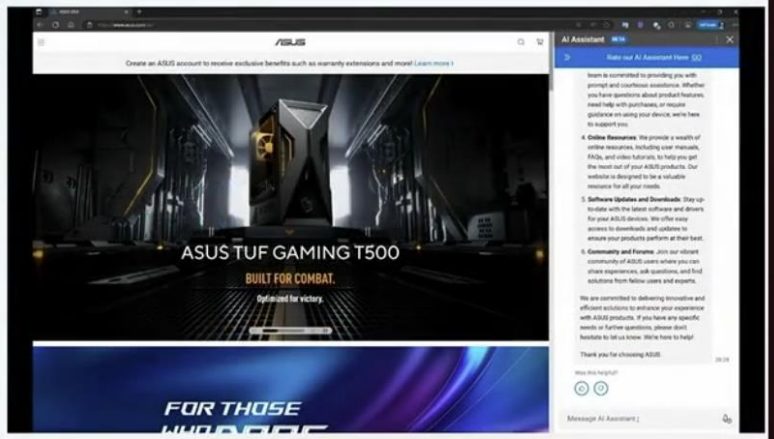
- Dataflow 優化
- Multi-Agent 急速發展下的治理難題
- MCP (Model Context Protocol) 與 A2A (Agent-to-Agent) 面臨的安全威脅

Microsoft | Microsoft Build Sessions Event guide Register now More

ASUS AI Assistant with Multi-Agent Architecture

Build on Azure with Semantic Kernel Orchestration and MCP-Connected Agents





"ASUS aims to transcend time and space limitations to engage consumers effectively. By leveraging AI assistants, we seek to significantly enhance customer satisfaction, boost revenue, and reduce service costs."

Marvin Hsu,
Vice President, ASUSTeK Computer Inc.

ASUS sought to improve its online customer engagement 24/7 by implementing an AI Assistant on its website. The **ASUS AI Assistant** led to a 6× increase in pageviews, 7× boost in interaction time, 600% rise in e-commerce conversion rate, 5× increase in purchases, and 20% higher average order value (AOV).

https://build.microsoft.com/en-US/sessions/BRK148

我們想要建造的 AI Assistant 是？

在一個面向消費者 (to C) 的網站，導入大語言模型，協助回答用戶有關：

產品規格、保固、維修、KB 諮詢、行銷優惠... 等問題。必須滿足以下要求：

理解語意、意圖

- Azure OpenAI GPT3.5 > GPT4o PTU > GPT4.1/4.1mini PTU and Google Gemini 2.5 Flash

回答嚴謹限制

- content filter / prompt engineering

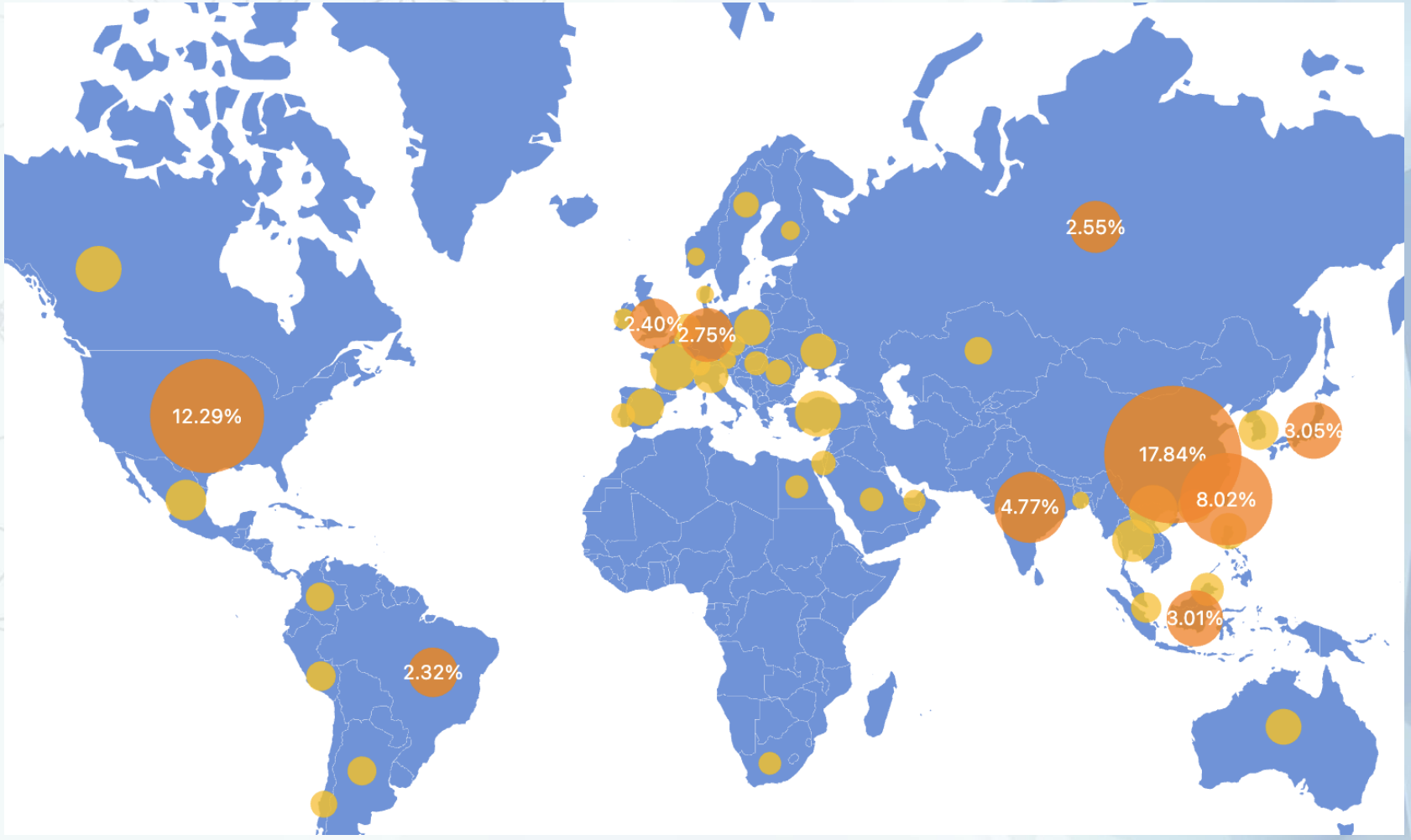
穩定精確

- embedding vector / AI search / prompt engineering

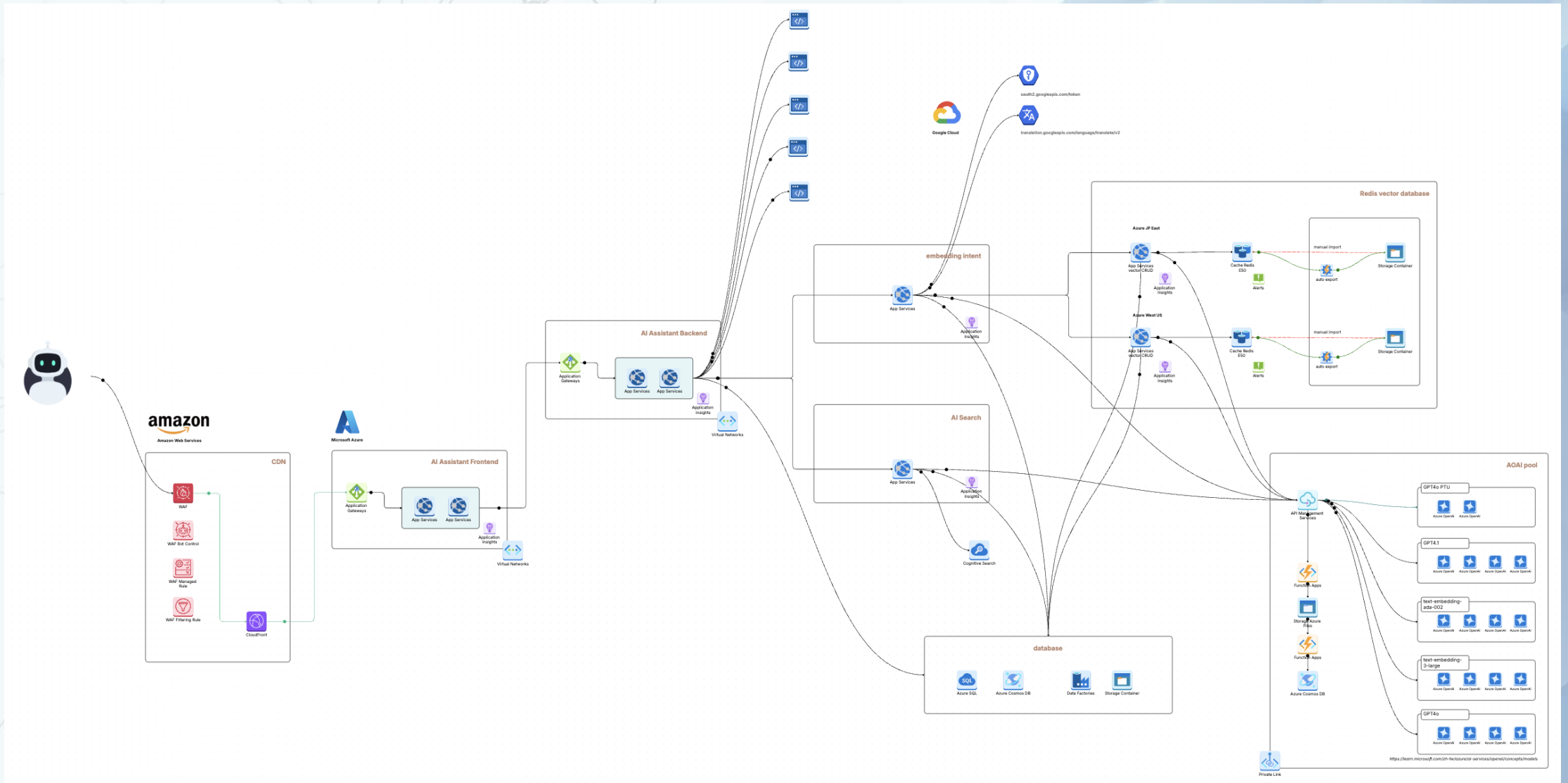
仿人語氣、品牌人設

- prompt engineering

來自全球每月超過 300TB 流量的網站服務



雲原生架構如何支持 AI Assistant 這樣的需求？



系統韌性（具備成本優勢的系統韌性）

關鍵系統面對突發事件與異常狀況時的應對能力；即使在系統部分故障下，仍能維持整體服務穩定，並快速恢復正常狀態。

系統韌性並非單一技術或機制，而是以下三大面向持續優化後的成果

- 高可用性（High Availability）：確保服務不中斷，避免單點故障擴散
- 效能調教（Performance Tuning）：確保資源高效運作，避免過載與延遲
- 安全防禦（Security Defense）：防止外部攻擊或異常行為消耗關鍵資源

三者缺一不可，共同構成系統面對複雜環境與風險時的韌性基礎

高可用性（High Availability）

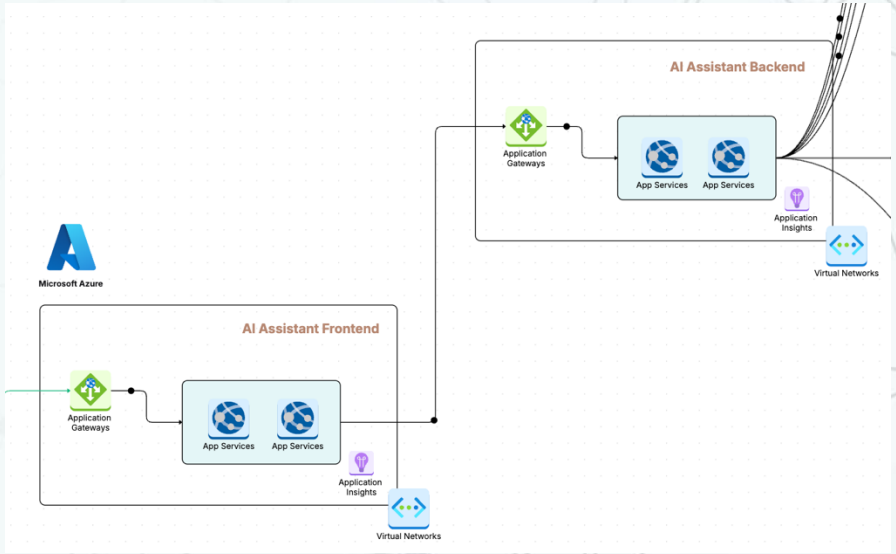
一系列確保系統功能持續可用的設計策略；透過冗餘部署、自動切換與容錯機制，降低單點故障影響，確保在異常發生時，服務不中斷、使用者體驗不受損。

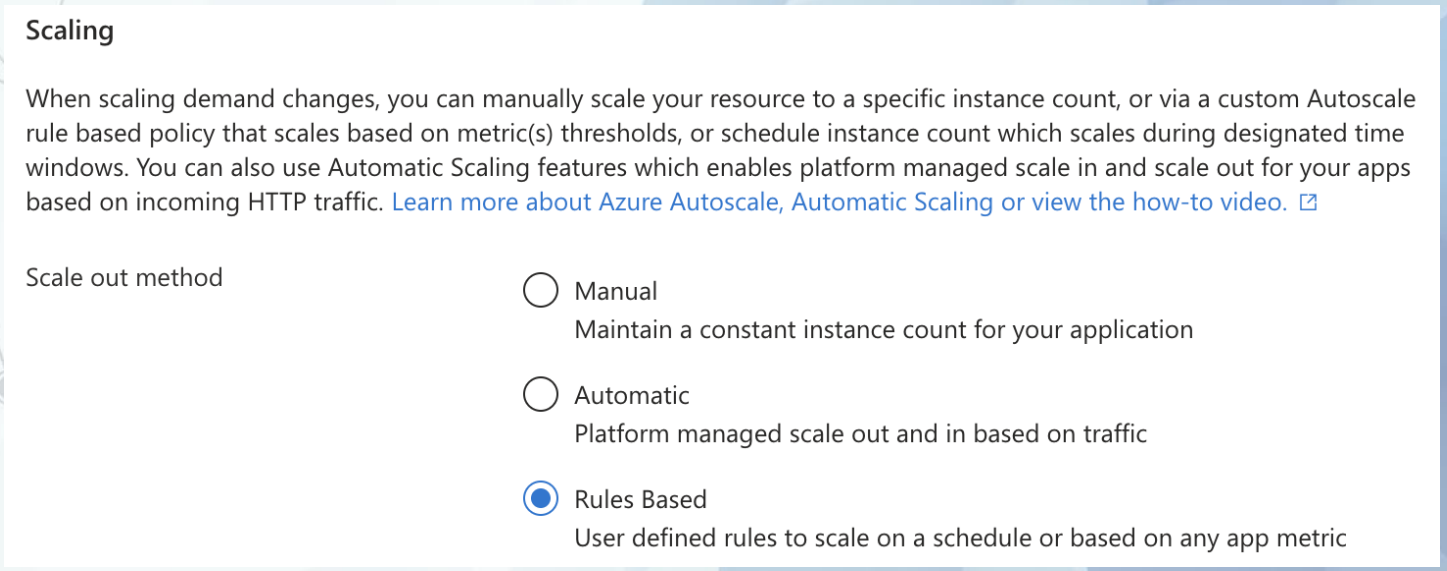
做最好的準備、最壞的打算？

高可用不代表系統永不故障，而是當故障發生時，服務仍能維持關鍵功能。我們需假設所有環節都可能失效，並透過自動擴展、跨區部署與自動復原等機制，將影響降至最低。

Things break; that's life. *Google SRE*

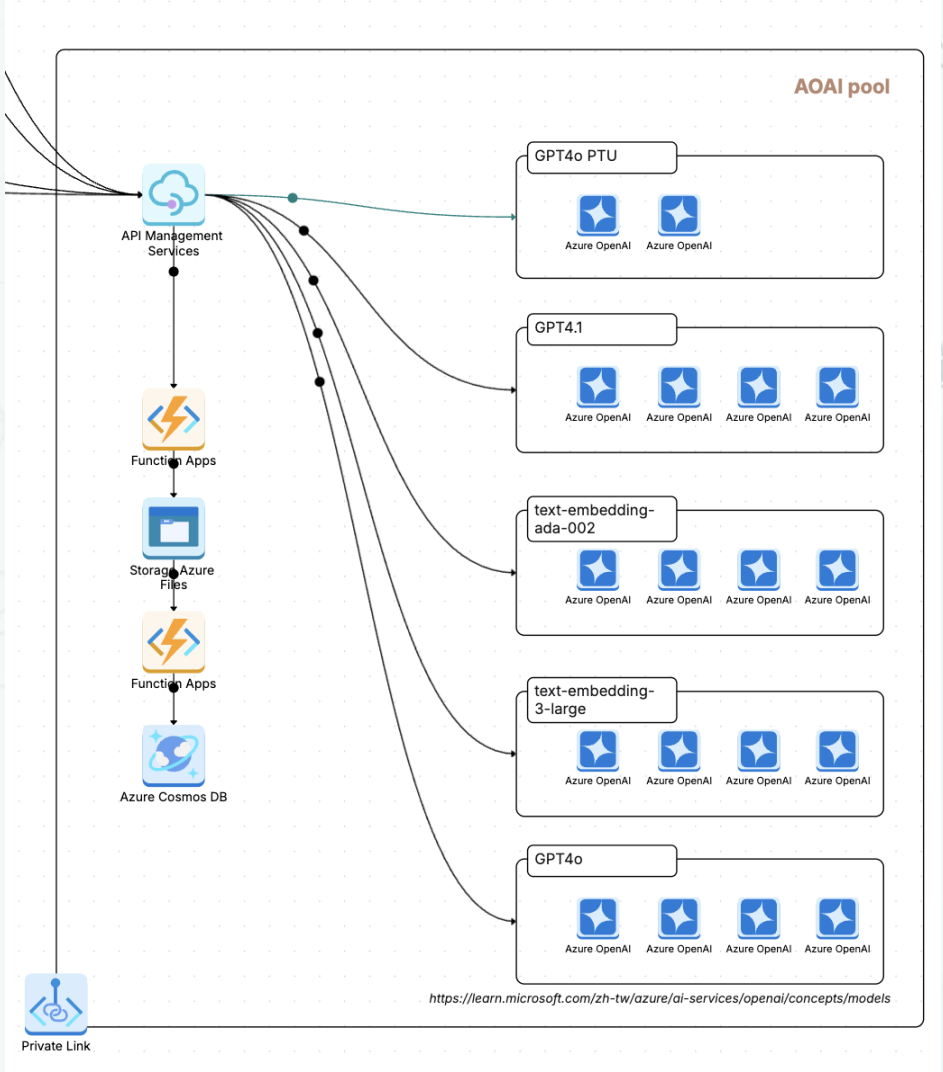
App Service plan HA 策略





- 閥值 window 錯開，輪流 rolling scale-out
- 啟用 zone redundant
- multi-region 部署

APIM 統一管理 AOAI endpoints，組成 AOAI pool 彈性擴充 LLM quota



- endpoints 封裝為內部 API，提供統一入口
- endpoints 改走 private link，安全性高
- 用戶（subscriptions）金鑰獨立控管
- 用戶流量個別限額
- 後端（backend）資源彈性定義
- 使用 Bicep 或 Terraform 進行管理

APIM inbound 分流

```
<inbound>
  <choose>
    <when condition="@(context.Request.IpAddress.StartsWith("203.0."))">
      <set-backend-service base-url="https://aoai-japaneast.openai.azure.com" />
    </when>
    <when condition="@(context.Request.IpAddress.StartsWith("198.51."))">
      <set-backend-service base-url="https://aoai-eastus.openai.azure.com" />
    </when>
    <otherwise>
      <set-backend-service base-url="https://aoai-centralus.openai.azure.com" />
    </otherwise>
  </choose>
</inbound>
```

APIM inbound 分流

```
<inbound>
  <set-variable name="region" value="@(context.Request.Headers.GetValueOrDefault("x-region", "default"))" />
  <choose>
    <when condition="@(context.Variables["region"] == "japan")">
      <set-backend-service base-url="https://aoai-japaneast.openai.azure.com" />
    </when>
    <when condition="@(context.Variables["region"] == "us")">
      <set-backend-service base-url="https://aoai-eastus.openai.azure.com" />
    </when>
    <otherwise>
      <set-backend-service base-url="https://aoai-southeastasia.openai.azure.com" />
    </otherwise>
  </choose>
</inbound>
```

APIM inbound rate limit

```
<inbound>
    <base />
    <choose>
        <when condition="@(context.Subscription.Name == "subscription-aia-prod")">
            <!-- 設定 Token 限額與容差 -->
            <set-header name="X-InputTokenLimit" exists-action="override">
                <value>200000</value>
            </set-header>
            <set-header name="X-InputTokenTolerance" exists-action="override">
                <value>1.0</value>
            </set-header>

            <!-- 設定速率限制：30 秒內最多 20 次 -->
            <rate-limit-by-key
                calls="20"
                renewal-period="30"
                counter-key="@(context.Subscription.Id)" />

            <!-- 指定後端服務 -->
            <set-backend-service backend-id="pool-gpt4o-ptu-only" />
        </when>
```

APIM backend pool

```
//pool
resource pool_gpt4o_ptu_only 'Microsoft.ApiManagement/service/backends@2023-09-01-preview' = {
  name: 'pool-gpt4o-ptu-only'
  parent: apim
  properties: {
    description: 'gpt-4o ptu-only pool'
    type: 'Pool'
    pool: {
      services: [
        {
          id: aoai_ptu_japaneast.id
          priority: 1
          weight: 1
        }

      ]
    }
  }
}
```

APIM backend pool

```
resource pool_gpt4o_ptu_first 'Microsoft.ApiManagement/service/backends@2023-09-01-preview' = {
  name: 'pool-gpt4o-ptu-first'
  parent: apim
  properties: {
    description: 'gpt-4o ptu-first pool'
    type: 'Pool'
    pool: {
      services: [
        {
          id: aoai_ptu_japaneast.id
          priority: 1
          weight: 1
        }
        {
          id: aoai_payg_japaneast.id
          priority: 2
          weight: 1
        }
        {
          id: aoai_payg_koreacentral.id
          priority: 3
          weight: 1
        }
        {
```

效能調教（Performance Tuning）

一系列用來優化系統運作效率的手段；針對影響系統效能的關鍵路徑與資源瓶頸，進行觀察、分析與優化，降低延遲、提升處理速度，確保系統不過度配置，在高負載下仍能穩定回應與擴展。

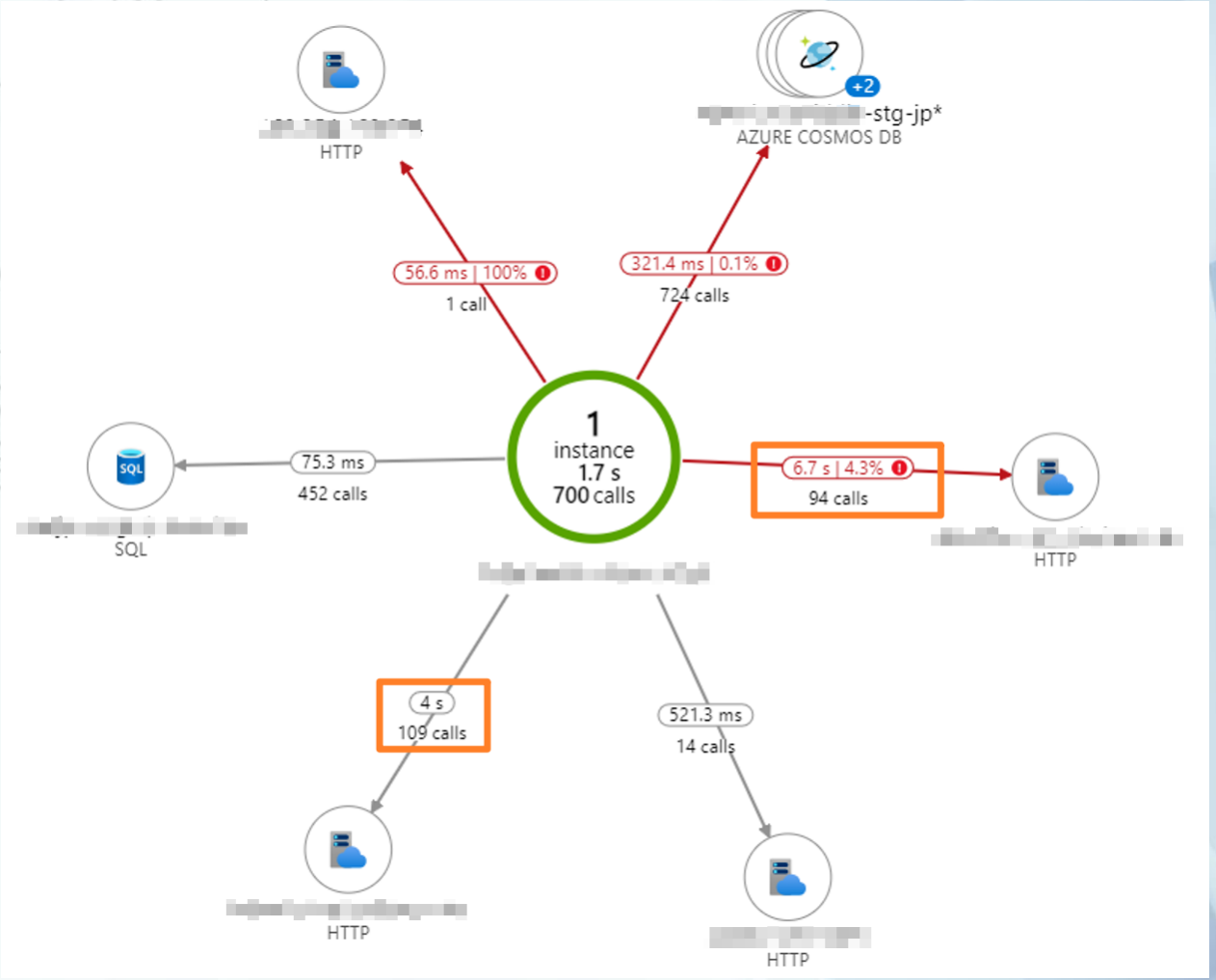
唯快不破？

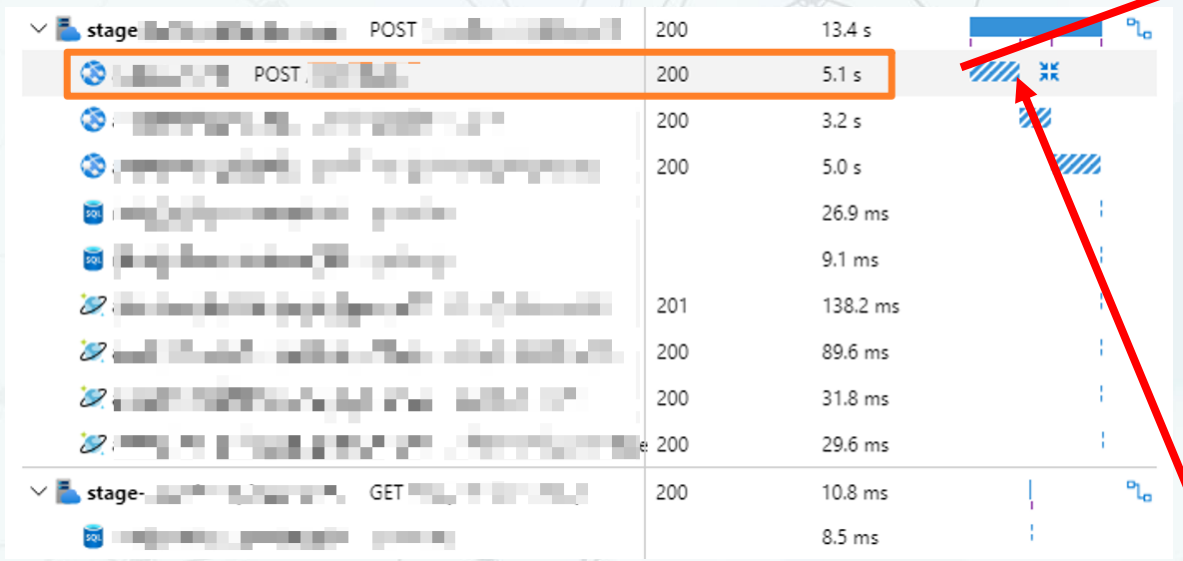
效能不是部署完成後才開始修補的議題，而是從設計之初就應內化於每一個決策細節。從資料模型、API 設計、快取策略到佇列機制，每一個小環節都應追求極致效率與節省。

Performance by design. *Tom Kyte*

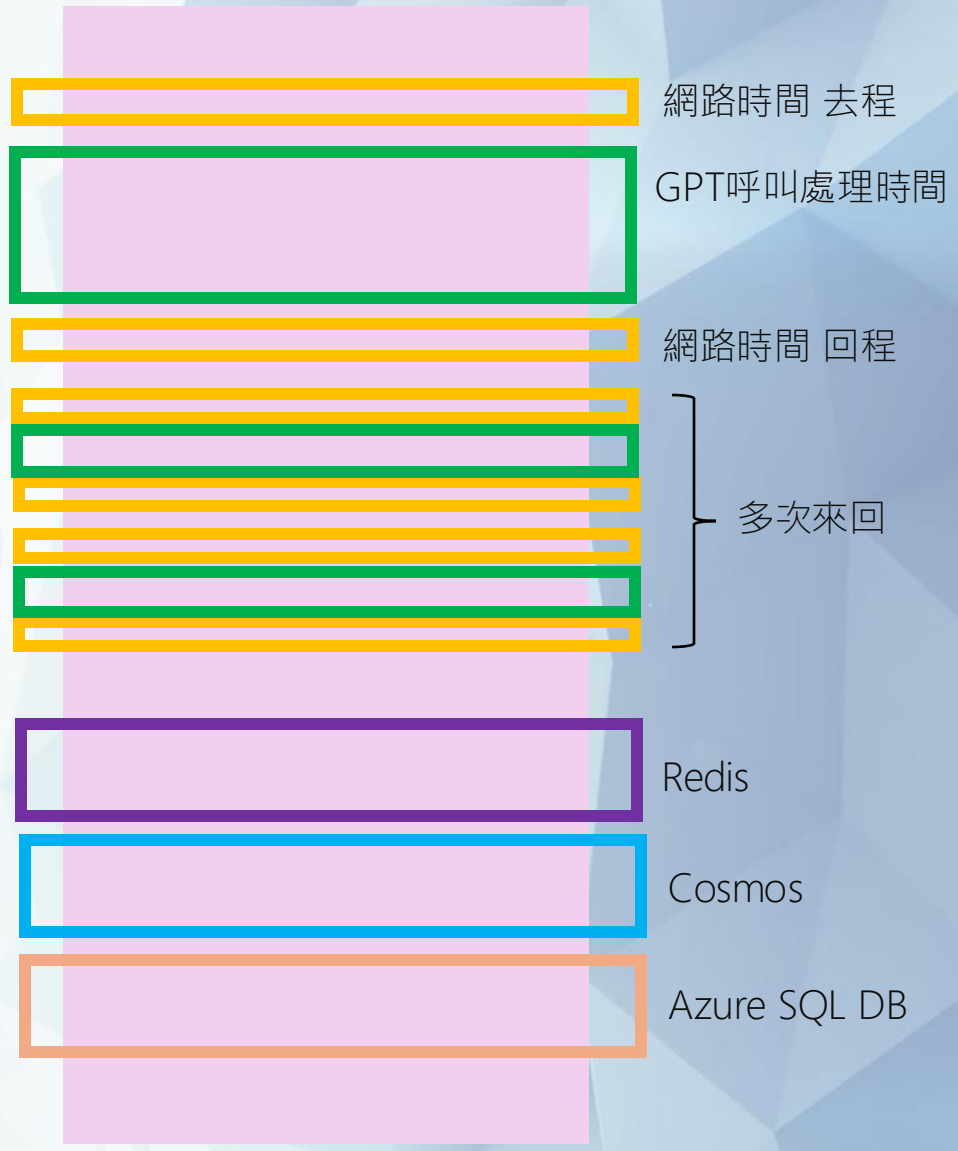
透過 Application Insight 追查 slow response

- Overview
- Activity log
- Access control (IAM)
- Tags
- Diagnose and solve problems
- Resource visualizer
- Investigate
  - Application map
  - Smart detection
  - Live metrics
  - Transaction search
  - Availability
  - Failures
  - Performance
- Monitoring
- Usage
- Configure
- Settings





重新思考 RTT (Round-Trip Time) 邏輯



數百萬筆的大型資料集，向量搜尋優化策略

- KNN（K Nearest Neighbors）查詢需慎重評估 top-K
- Vector Indexing methods
  - FLAT 逐筆比對所有向量，計算相似度後再排序選出最相近的結果，是最精確但最耗時的搜尋方式。
  - HNSW 使用多層次的圖結構，讓向量查詢像 GPS 導航一樣，從粗略跳躍到精確搜尋，大幅加快查詢速度。
- Redis query methods (RTT)
  - single queries 單筆查詢簡單直觀，但在高頻操作下會因多次網路往返導致效能低落。
  - batch queries (pipelining) 批次查詢能大幅減少延遲並提升吞吐量，但需多留意實作與錯誤處理的複雜性。

向量維度優化（儲存空間）

Model	text-embedding-ada-002	text-embedding-3-small	text-embedding-3-large
MTEB score	61.0 %	62.3 %	64.6 %
Dimension	1536	256 / 512 / 768 / 1536	256 / 512 / 768 / 1536 / 3072
Price (per 1K token)	$0.0001	$0.00002	$0.00013
Storage	~ 6 KB (1536 x float32)	256~ 1 KB 512~ 2 KB 1536~ 6 KB	256~ 1 KB 512~ 2 KB 1536~ 6 KB 3072~ 12 K
百萬向量儲存空間	~ 6 GB	~ 1GB to 6GB	~ 1GB to 12GB
適用場景	常規應用，語義搜尋	常規應用，語義搜尋	高準確搜尋（法律、醫療等）、 大規模語料壓縮與檢索

安全防禦（Security Defense）

一系列用來保護系統免受各種威脅的措施；識別服務中最關鍵、昂貴、擴展不易之組件，透過防禦機制，阻絕或降低資源消耗，提升系統韌性。

易攻難守？

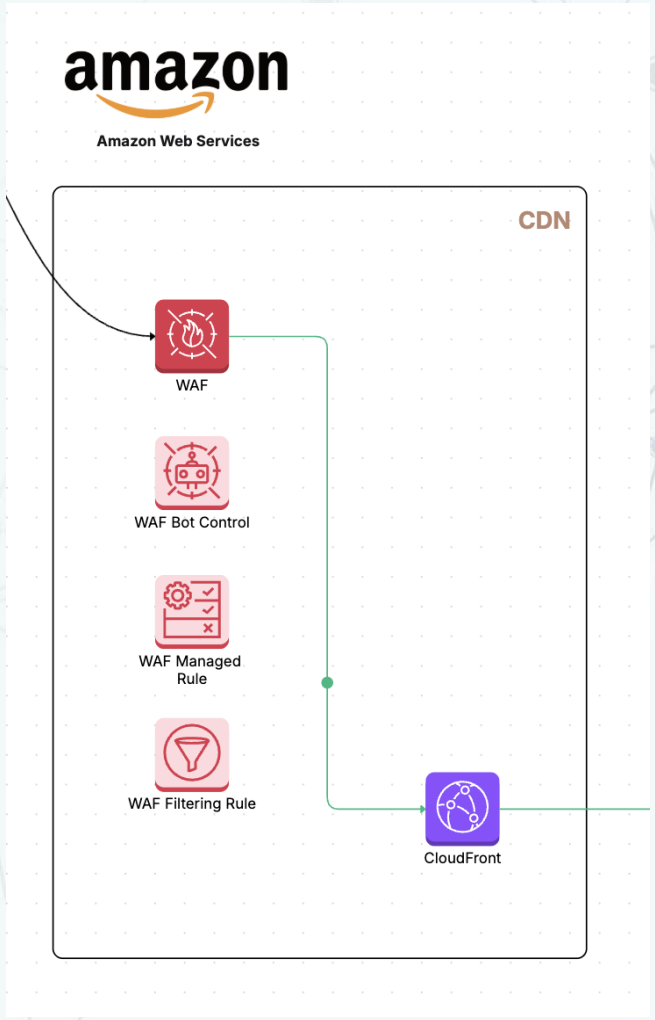
在高併發、高可用架構中，資料庫經常被視為最昂貴、最難擴展、也最易成為瓶頸的資源。

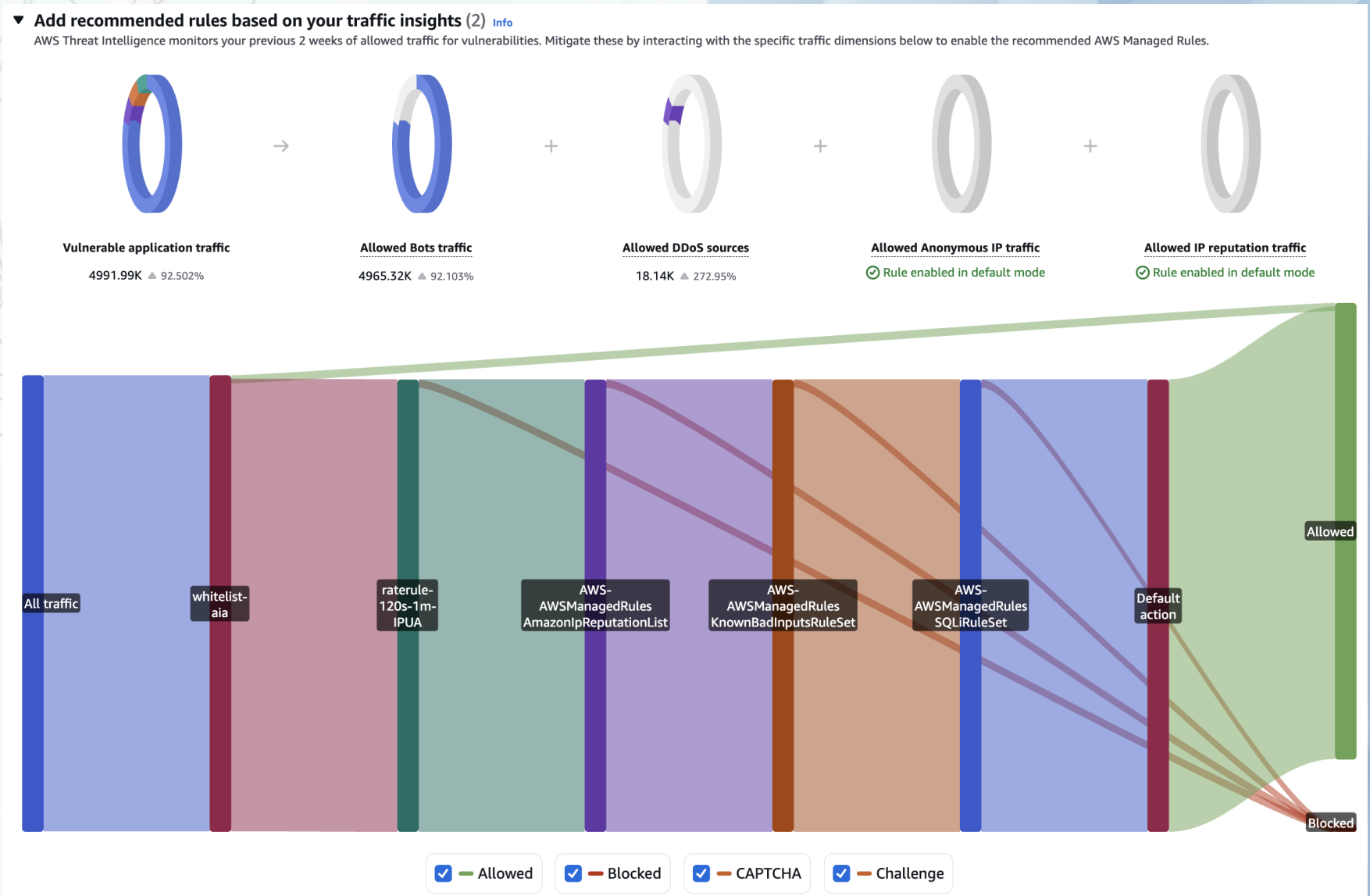
而今日，LLM 資源，也應被列入昂貴資源範圍。

OWASP Top 10 for LLM Applications 2025 *November 17, 2024*

- LLM10 無限資源耗盡（Unbounded Consumption）

AWS WAF


amazon
Amazon Web Services
CDN
WAF
WAF Bot Control
WAF Managed Rule
WAF Filtering Rule
CloudFront


Add recommended rules based on your traffic insights (2) Info
AWS Threat Intelligence monitors your previous 2 weeks of allowed traffic for vulnerabilities. Mitigate these by interacting with the specific traffic dimensions below to enable the recommended AWS Managed Rules.
Vulnerable application traffic
4991.99K ▲ 92.502%
Allowed Bots traffic
4965.32K ▲ 92.103%
Allowed DDoS sources
18.14K ▲ 272.95%
Allowed Anonymous IP traffic
Rule enabled in default mode
Allowed IP reputation traffic
Rule enabled in default mode
All traffic
whitelist-aia
raterule-120s-1m-IPUA
AWS-AWSManagedRules AmazonIpReputationList
AWS-AWSManagedRules KnownBadInputsRuleSet
AWS-AWSManagedRules SQLiRuleSet
Default action
Allowed
Blocked
Allowed
Blocked
CAPTCHA
Challenge

rate base with JA4

AI 摘要

JA4 是一種用於網路指紋辨識的技術，特別是針對TLS (傳輸層安全性) 協定。JA4+ 則是一個更廣泛的指紋辨識套件，涵蓋了TLS、HTTP 和SSH 等多種協議。 這些技術被設計為人類和機器都可解釋，以增強威脅檢測和安全分析的能力。

JA3 fingerprint

Count	JA3 fingerprint
31% (115)	73d9f91a3f97729a106e5ed93e1b4097
26% (95)	9a382d25aa597efd92d217d84b6eca00
19% (71)	1c789053a78bc0a3ddccac65f52a63c1
9% (34)	8d9f7747675e24454cd9b7ed35c58707
3% (12)	1226e212fe17254cdc21b941b0954350
3% (10)	1be4ecae27bde866e7802226c332b3da
2% (6)	b0031c60dcd922b8c8d343ecfa5cf91c

JA4 fingerprint

Count	JA4 fingerprint
52% (189)	t13d1517h2_8daaf6152771_b6f405a00624
26% (95)	t13d131000_f57a46bbacb6_e7c285222651
10% (35)	t13d4312h1_c7886603b240_b26ce05bbdd6
8% (28)	t13d1516h2_8daaf6152771_d8a2da3f94cd
3% (10)	t12d800600_15c493246000_a1e935682795
1% (4)	t13d1611h2_1711a4c0508c_6d021c4c45cd
0% (1)	t13d1515h1_8daaf6152771_cc38aef784ae

Action

Block

Rule name

300-1m-ja4-ip

Names must use 1-128 characters from A-Z, a-z, 0-9, hyphen, and underscore.

Statement

Rate limit

300

Rate limit must be between 10 and 2,000,000,000.

Evaluation window

1 minute

The default time span is 5 minutes.

Rule configuration

Request aggregation

- Source IP address
  Origin server IP address (most reliable, direct request source).
- IP address in header
  Header IP address (may be inconsistent or modified).
- Custom keys
  Create an aggregation key using one or more component types, selected from a list. You can combine an IP address option with other components, and you can create keys that don't use an IP address.
- Count all
  Count and rate limit all requests that match the rule's scope-down statement.

Request aggregation key

You can aggregate requests by one or more request components. AWS WAF first groups requests by the keys you provide, then counts each group and compares the counts against the rate limit.

JA4 fingerprint

Remove

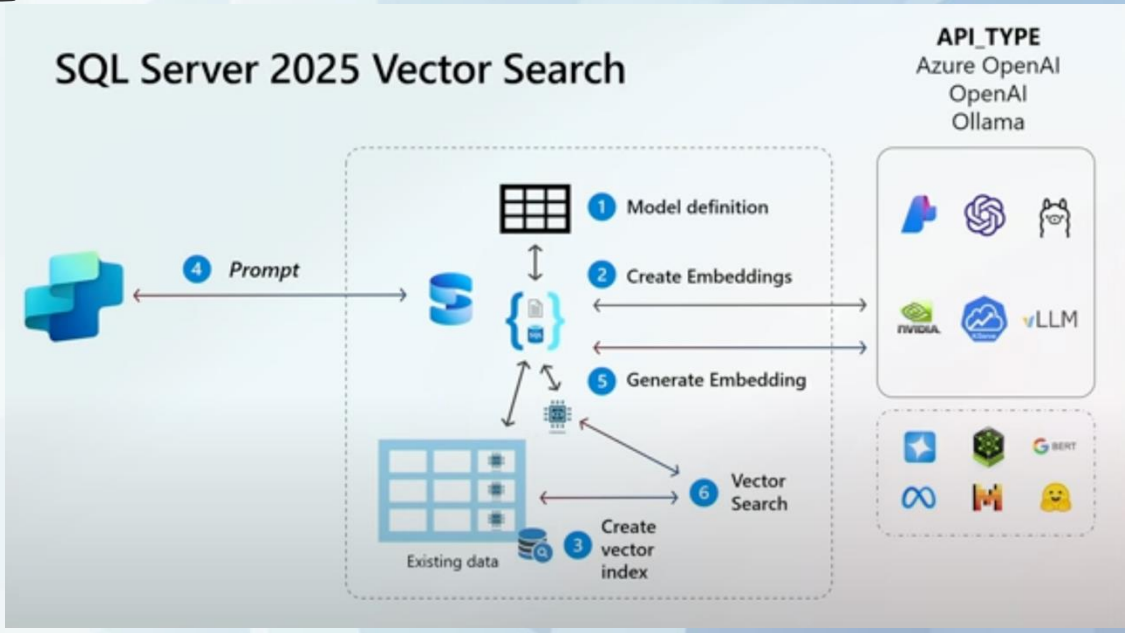
Fallback for missing JA4 fingerprint

- Match
  Treat the web request as matching the rule statement.
- No match
  Treat the web request as not matching the rule statement.

接下來的機會與挑戰 – Dataflow 優化

SQL 2025 / Azure SQL MI / Azure SQL database

Don't move your data to AI. Bring AI to your data.



SQL 原生向量功能

- 向量型別 column_name VECTOR({<dimensions>}) [NOT NULL | NULL]
- 向量函式 VECTOR_DISTANCE, VECTOR_NORM, VECTOR_NORMALIZE
- AI 函式 AI_GENERATE_EMBEDDINGS, AI_GENERATE_CHUNKS, CREATE EXTERNAL MODEL
- DiskANN-Based Vector Index *(GA soon)*

https://build.microsoft.com/en-US/sessions/BRK207

接下來的機會與挑戰 – Multi-Agent 急速發展下的治理難題

Agent 行為不穩定且難以測試驗證

在多 Agent 架構中，回應結果高度依賴 prompt、上下文、記憶體與工具回傳內容，自動化測試與驗證將更加複雜

A2A 呼叫鏈過長，錯誤難追蹤

Agent 之間互相呼叫會形成深度鏈結，只要一個環節出錯就可能引發非預期結果，完整的 trace id 更加重要

缺乏統一的 Agent Registry 與部署管理

若無 Agent 管理平台，開發者各自部署、自定功能與版本，造成維運混亂與衝突

接下來的機會與挑戰 – MCP 與 A2A 面臨的安全威脅

跨 Agent 呼叫鏈缺乏 Trace 與驗證機制（A2A Chain Tampering）

在 A2A 結構中，Agent 可能以 chain 方式呼叫彼此，但若每次呼叫未攜帶 trace id 或無驗證來源身份，會出現以下風險，如：呼叫來源不可驗證（可能遭模擬）、trace id 缺失導致無法追蹤、被插入惡意 Agent（中間人風險）

記憶體與向量資料存取缺乏隔離（Memory Access Leakage）

多 Agent 共用記憶體（如 Redis / Cosmos DB）時，若未劃分明確存取範圍，將可能導致：某 Agent 調用他人記憶，造成資訊洩漏、記憶污染回應混亂或模型誤導

工具呼叫未設限，導致濫用與資安暴露（Tool Invocation Risk）

若 Agent 可以無限制呼叫 Tool（API、查詢模組、外部服務），將可能產生：Tool 被濫用觸發高成本或敏感 API（如 DB Query、PII API）、工具服務成為外部攻擊點（如 SSRF）

Thank you for your time.