AWS Well-Architected Framework

效能達成效率支柱

Copyright © 2025 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

效能達成效率支柱: AWS Well-Architected Framework

Copyright © 2025 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

Amazon 的商標和商業外觀不得用於任何非 Amazon 的產品或服務，也不能以任何可能造成客戶混淆、任何貶低或使 Amazon 名譽受損的方式使用 Amazon 的商標和商業外觀。所有其他非 Amazon 擁有的商標均為其各自擁有者的財產，這些擁有者可能隸屬於 Amazon，或與 Amazon 有合作關係，或由 Amazon 贊助。

Table of Contents

摘要和介紹 1
簡介 1
效能效率 3
設計原則 3
定義 3
架構選擇 5
PERF01-BP01 了解和了解可用的雲端服務和功能 5
實作指引 6
資源 6
PERF01-BP02 使用雲端供應商或適當合作夥伴提供的指引，了解架構模式和最佳實務 7
實作指引 6
資源 6
0PERF01-BP03 將成本納入架構決策 9
實作指引 6
資源 6
PERF01-BP04 評估權衡如何影響客戶和架構效率 11
實作指引 6
資源 6
PERF01-BP05 使用政策和參考架構 12
實作指引 6
資源 6
PERF01-BP06 使用基準測試來推動架構決策 14
實作指引 6
資源 6
PERF01-BP07 使用資料驅動方法進行架構選擇 16
實作指引 6
資源 6
運算與硬體 19
PERF02-BP01 為您的工作負載選取最佳運算選項 19
實作指引 6
實作步驟 6
資源 6
PERF02-BP02 了解可用的運算組態和功能 22
實作指引 6

實作步驟 6
資源 6
PERF02-BP03 收集運算相關指標 25
實作指引 6
實作步驟 6
資源 6
PERF02-BP04 設定和正確大小的運算資源 27
實作指引 6
資源 6
PERF02-BP05 動態擴展運算資源 29
實作指引 6
資源 6
PERF02-BP06 使用最佳化的硬體型運算加速器 32
實作指引 6
資源 6
資料管理 35
PERF03-BP01 使用最能夠支援資料存取和儲存需求的專用資料存放區 35
實作指引 6
資源 6
PERF03-BP02 評估資料存放區的可用組態選項 43
實作指引 6
資源 6
PERF03-BP03 收集並記錄資料存放區效能指標 47
實作指引 6
實作步驟 6
資源 6
PERF03-BP04 實作策略以提高資料存放區中的查詢效能 50
實作指引 6
資源 6
PERF03-BP05 實作使用快取的資料存取模式 52
實作指引 6
資源 6
聯網與內容交付 55
PERF04-BP01 了解聯網如何影響效能 55
實作指引 6
資源 6

PERF04-BP02 評估可用的聯網功能 58
實作指引 6
資源 6
PERF04-BP03 VPN為您的工作負載選擇適當的專用連線能力 63
實作指引 6
資源 6
PERF04-BP04 使用負載平衡將流量分散到多個資源 65
實作指引 6
資源 6
PERF04-BP05 選擇網路通訊協定以提高效能 69
實作指引 6
資源 6
PERF04-BP06 根據網路需求選擇工作負載的位置 71
實作指引 6
資源 6
PERF04-BP07 根據指標最佳化網路組態 75
實作指引 6
資源 6
程序和文化 79
PERF05-BP01 建立用於測量工作負載運作狀態和效能的關鍵績效指標 (KPI) 80
實作指引 6
實作步驟 6
資源 6
PERF05-BP02 使用監控解決方案來了解效能最關鍵的領域 82
實作指引 6
資源 6
PERF05-BP03 定義改善工作負載效能的程序 84
實作指引 6
資源 6
PERF05-BP04 Load 測試您的工作負載 86
實作指引 6
資源 6
PERF05-BP05 使用自動化主動修復效能相關問題 88
實作指引 6
資源 6
PERF05-BP06 保留工作負載和服務 up-to-date 90

實作指引 6
實作步驟 6
資源 6
PERF05-BP07 定期審查指標 91
實作指引 6
資源 6
結論 94
貢獻者 95
深入閱讀 96
文件修訂 97
注意 99
AWS 詞彙表 100

效能達成效率支柱 – AWS Well-Architected Framework

發布日期：2024 年 11 月 6 日 (文件修訂)

本白皮書著重於介紹 AWS Well-Architected Framework 的效能達成效率支柱。本文提供了相關指引，可協助客戶在設計、交付和維護 AWS 環境時運用最佳實務。

簡介

AWS Well-Architected Framework 可協助您了解在 AWS 上建置工作負載時所做決策的優缺點。透過此架構，您將了解架構的最佳實務，以便在雲端設計和操作可靠、安全、有效率、經濟實惠且永續的工作負載。該架構可讓您根據最佳實務一致地量測架構，並識別需要改進的方面。我們相信，擁有 Well-Architected 工作負載可大幅提高企業成功的可能性。

此架構以六大支柱為基礎：

- 操作效能
- 安全
- 可靠性
- 效能效率
- 成本最佳化
- 永續性

本白皮書著重於如何將效能達成效率支柱原則應用到工作負載。在傳統的內部部署環境中，實現持久的高效能是一項充滿挑戰性的任務。使用本白皮書中的原則可協助您在 AWS 上建立架構，以長時間有效率地提供持續的效能。本文件中的指引和最佳實務分佈於五個關鍵重點領域，可做為在 AWS 上建置效能達成效率雲端解決方案的指導原則。這些重點領域包括：

- 架構選擇
- 運算與硬體
- 資料管理
- 聯網與內容交付
- 程序和文化

本白皮書適用於擔任技術職務的人員，例如技術長 (CTO)、架構師、開發人員和營運團隊成員。閱讀本白皮書之後，您將了解設計效能雲端架構時要使用的 AWS 最佳實務和策略。

效能效率

效能效率支柱包括能夠有效率地使用雲端資源，以滿足效能需求，並隨著需求變更與技術發展來保持該效率需求。

主題

- 設計原則
- 定義

設計原則

下列設計原則可協助您在雲端中達成和保持高效率工作負載。

- 讓進階技術變得更普及：將複雜的任務委派給雲端廠商，讓團隊更輕鬆地實作進階技術。與其要求 IT 團隊了解新技術的託管和執行方式，不如考慮使用技術即服務。例如，沒有任何SQL資料庫、媒體轉碼和機器學習都是需要專業知識的技術。在雲端，這些技術成為團隊可以使用的服務，讓團隊可專注於產品開發，而非資源佈建及管理。
- 幾分鐘內即可全球化：將工作負載部署到全球多個 AWS 區域可讓您以最低成本為客戶提供更低的延遲和更好的體驗。
- 使用無伺服器架構：採用無伺服器架構，您便無需執行和維護實體伺服器來完成傳統運算活動。例如，無伺服器儲存服務可以充當靜態網站 (因此無需 Web 伺服器)，而事件服務可以為您託管程式碼。如此一來，即可減輕管理實體伺服器的營運負擔，而且由於這些受管服務是在雲端規模上運行，因此還可以降低交易成本。
- 提高試驗頻率：使用虛擬及可自動化的資源，您可以使用不同類型的執行個體、儲存設備或組態，迅速完成比較測試。
- 考慮機械同感：使用最符合您目標的技術方法。例如，為工作負載選取資料庫或儲存時，請考量資料存取模式。

定義

專注於下列領域，以便在雲端實現效能達成效率：

- 架構選擇
- 運算與硬體

- 資料管理
- 聯網與內容交付
- 程序和文化

採取資料驅動的方法來建置高效能架構。從高階設計到選取和設定資源類型，收集架構各方面的資料。

定期檢閱您的選擇，確保您充分利用不斷發展的 AWS 雲端。監控可確保您能察覺預期效能發生的任何偏差情形。在架構中做出權衡以改進效能，例如使用壓縮或快取，或放寬一致性要求。

架構選擇

適用於特定工作負載的最佳解決方案各不相同，而解決方案通常會結合多種方法。Well-Architected 工作負載會使用多種解決方案，並採用不同的功能以提升效能。

AWS 資源有多種類型和組態，可讓您更輕鬆地找到最符合需求的方法。您還可以發現使用內部部署基礎設施不易實現的選項。例如，Amazon DynamoDB 這種受管服務，可提供全受管的 NoSQL 資料庫及任何規模下的十毫秒內延遲時間。

這個重點領域分享了如何選取高效率、高效能雲端資源和架構模式的各種指引和最佳實務。

最佳實務

- PERF01-BP01 了解和了解可用的雲端服務和功能
- PERF01-BP02 使用雲端供應商或適當合作夥伴提供的指引，了解架構模式和最佳實務
- 0PERF01-BP03 將成本納入架構決策
- PERF01-BP04 評估權衡如何影響客戶和架構效率
- PERF01-BP05 使用政策和參考架構
- PERF01-BP06 使用基準測試來推動架構決策
- PERF01-BP07 使用資料驅動方法進行架構選擇

PERF01-BP01 了解和了解可用的雲端服務和功能

持續了解並探索可用的服務和組態，有助您做出更完善的架構決策，並提升工作負載架構的效能效率。

常見的反模式：

- 您可以使用雲端作為並置資料中心。
- 移轉到雲端後，您不會將應用程式現代化。
- 對於需要保留的所有項目，您只使用一種儲存類型。
- 您使用的執行個體類型與目前標準最相符，但大於需求。
- 您會部署和管理可做為受管服務的技術。

建立此最佳實務的優勢：透過考慮新服務和設定，您可以大幅提升效能、降低成本並最佳化維護工作負載所需的工作量。它也可以協助您加速啟用 time-to-value雲端產品的 。

未建立此最佳實務時的曝險等級：高

實作指引

AWS 持續推出新的服務和功能，可改善效能並降低雲端工作負載的成本。繼續使用 up-to-date這些新服務和功能對於在雲端維持效能效能至關重要。將工作負載架構現代化也可協助您提升生產力、推動創新並釋放更多成長機會。

實作步驟

- 清查工作負載軟體和架構以存放相關服務。決定要深入了解的產品類別。
- 探索 AWS 產品，以識別並了解相關的服務和組態選項，協助您改善效能並降低成本和操作複雜性。
  - Amazon Web Services 雲端
  - AWS 學院
  - 有什麼新功能 AWS？
  - AWS 部落格
  - AWS 技能建置器
  - AWS 事件和網路研討會
  - AWS 培訓 和 憑證
  - AWS Youtube 頻道
  - AWS 研討會
  - AWS 社群
- 使用 Amazon Q 取得有關服務的相關資訊和建議。
- 使用沙盒 (非生產) 環境來學習和試驗新服務，而不會產生額外成本。
- 持續了解新雲端服務和功能。

資源

相關文件：

- Amazon Web Services 概觀
- Amazon EC2功能
- step-by-step使用 AWS 合作夥伴學習計劃學習
- AWS 訓練和認證

- 我成為 AWS 解決方案架構師的學習路徑
- AWS 架構中心
- AWS Partner Network
- AWS 解決方案程式庫
- AWS 知識中心
- 在 上建置現代應用程式 AWS

相關影片：

- AWS re：Invent 2023 - Amazon 的新功能 EC2
- AWS re：Invent 2022 - 使用 Amazon 降低您的營運和基礎設施成本 ECS
- AWS re：Invent 2023 - 使用 建立具有效率、敏捷性和創新的雲端 AWS
- AWS re：Invent 2022 - 部署 ML 模型，以高效能和低成本進行推論
- This is my Architecture

相關範例：

- AWS 範例
- AWS SDK 範例

PERF01-BP02 使用雲端供應商或適當合作夥伴提供的指引，了解架構模式和最佳實務

使用文件、解決方案架構師、專業服務或適當的合作夥伴等雲端公司資源，來引導您做出架構決策。這些資源可協助檢閱和改善架構，以實現最佳效能。

常見的反模式：

- 使用 AWS 作為常見的雲端供應商。
- 以非設計宗旨的方式使用 AWS 服務。
- 遵循所有指引，但未考量自身的業務環境。

建立此最佳實務的優勢：使用雲端供應商或適當合作夥伴的指引，可協助您針對工作負載做出正確的架構選擇，並讓您對決策充滿信心。

未建立此最佳實務時的曝險等級：中

實作指引

AWS 提供廣泛的指引、文件和資源，可協助您建置和管理高效雲端工作負載。AWS 文件提供程式碼範例、教學課程和詳細的服務說明。除了說明文件外，AWS 還提供訓練與認證計畫、解決方案架構師和專業服務，協助客戶探索雲端服務的不同層面，並在 AWS 上實作高效的雲端架構。

利用這些資源來深入了解寶貴的知識和最佳實務、節省時間並在 AWS 雲端 中取得更好的成果。

實作步驟

- 檢閱 AWS 文件和指引，並遵循最佳實務。這些資源可協助您有效選擇和設定服務，並取得更好的效能。
  - AWS 文件 (例如使用者指南和白皮書)
  - AWS 部落格
  - AWS 培訓 和認證
  - AWS Youtube 頻道
- 參加 AWS 合作夥伴活動 (例如 AWS 全球高峰會、AWS re:Invent、使用者群組和研討會)，向 AWS 專家學習使用 AWS 服務的最佳實務。
  - 透過 AWS 合作夥伴學習計劃逐步學習
  - AWS 活動和研討會
  - AWS 研討會
  - AWS 社群
- 當您需要其他指引或產品資訊時，請聯絡 AWS 尋求協助。AWS解決方案架構師和 AWS 專業服務會為解決方案實作提供指引。AWS合作夥伴提供 AWS 專業知識，協助您提升業務的靈活性和創新性。
- 如果您需要技術支援才能有效使用服務，請使用 支援。我們的支援計劃旨在為您提供適當的工具組合和專業知識，以便您在最佳化效能、管理風險以及控制成本的同時，在 AWS 上取得成功。

資源

相關文件：

- AWS 架構中心

- AWS Partner Network
- AWS 解決方案程式庫
- AWS 知識中心
- AWS 企業支援

相關影片：

- This is my Architecture
- AWS re:Invent 2023 - 使用 Amazon EventBridge 提升事件驅動型模式
- AWS re:Invent 2023 - 在 AWS 上實作分散式設計模式
- AWS re:Invent 2023 - 應用程式架構即程式碼

相關範例：

- AWS 範例
- AWS SDK 範例
- AWS 分析參考架構

0PERF01-BP03 將成本納入架構決策

將成本納入架構決策中，以提高雲端工作負載的資源使用率和效能效率。當您意識到雲端工作負載的成本影響時，就更有可能利用有效的資源並減少浪費的做法。

常見的反模式：

- 您只能使用一個執行個體系列。
- 您不會針對開放原始碼解決方案評估授權解決方案。
- 您不會定義儲存區生命週期政策。
- 您不會檢閱 的新服務和功能 AWS 雲端。
- 您只能使用區塊儲存。

建立此最佳實務的優勢：將成本納入到決策中可讓您使用更有效率的資源並探索其他投資。

未建立此最佳實務時的曝險等級：中

實作指引

優化工作負載成本可以提高資源利用率並避免雲端工作負載中的浪費。將成本納入架構決策中，通常包括適當調整工作負載元件大小以及啟用彈性，進而提高雲端工作負載效能的效率。

實作步驟

- 確立成本目標，例如雲端工作負載的預算限制。
- 找出造成工作負載成本增加的關鍵元件 (例如執行個體和儲存)。可使用 AWS 定價計算工具 和 AWS Cost Explorer 找出工作負載中的關鍵成本驅動因素。
- 了解雲端中的定價模式，例如隨需執行個體、預留執行個體、Savings Plans 和 Spot 執行個體。
- 使用 Well-Architected 成本最佳實務，針對成本最佳化這些關鍵元件。
- 持續監控和分析成本，以找出工作負載中成本最佳化的機會。
  - 使用 AWS Budgets 取得不可接受成本的警示。
  - 使用 AWS Compute Optimizer 或 AWS Trusted Advisor 得成本最佳化建議。
  - 使用 AWS Cost Anomaly Detection 取得自動化成本異常偵測和根本原因分析。

資源

相關文件：

- 什麼是 AWS Billing and Cost Management？
- 使用 進行成本最佳化 AWS
- 選擇 AWS 成本管理策略
- AWS 成本管理入門指南
- 成本智慧儀表板的詳細概要
- AWS 架構中心
- AWS 解決方案程式庫
- AWS 知識中心

相關影片：

- This is my Architecture
- AWS re：Invent 2023 - AWS 成本最佳化的新功能

- AWS re：Invent 2023 - 最佳化成本和效能，並追蹤緩解進度
- AWS re：Invent 2023 - AWS 儲存成本最佳化最佳實務
- AWS re：Invent 2023 - 最佳化多帳戶環境中的成本

相關範例：

- AWS Compute Optimizer 示範程式碼
- 成本最佳化研討會
- 雲端財務管理技術實作說明手冊
- 啟動最佳化：調整應用程式效能以實現最高效率
- 無伺服器最佳化研討會 (效能與成本)
- 擴充經濟高效的架構

PERF01-BP04 評估權衡如何影響客戶和架構效率

在評估與效能相關的改進時，判斷哪些選擇會影響客戶和工作負載效率。例如，如果使用鍵值資料存放區可提高系統效能，請務必評估此變更最終一致性本質對客戶的影響。

常見的反模式：

- 即使實作過程中有所取捨，您都假設應實作所有效能增益。
- 您只會在效能問題達到臨界點時才會評估工作負載變更。

建立此最佳實務的優勢：評估潛在的效能相關改善項目時，必須判斷技術變更的權衡是否符合工作負載要求。在某些情況下，您可能需要實作其他控制來彌補權衡。

未建立此最佳實務時的曝險等級：高

實作指引

根據效能和客戶影響，識別架構中的關鍵領域。確定如何進行改進、這些改進帶來的權衡，以及它們如何影響系統和使用者體驗。例如，實作快取資料有助於大幅提升效能，但需要明確的策略來確定更新或使快取資料失效的方式和時間，以防止不正確的系統行為。

實作步驟

- 了解工作負載需求和 SLA。

- 清楚定義評估因素。因素可能與工作負載的成本、可靠性、安全性和效能有關。
- 選擇可滿足需求的架構和服務。
- 進行實驗和概念驗證 (POC)，以評估權衡因素以及對客戶和架構效率的影響。通常，高可用性、高效能且安全的工作負載會耗用更多雲端資源，但能夠提供更完善的客戶體驗。了解工作負載複雜性、效能和成本的權衡。通常情況下，優先考慮其中兩個因素會以犧牲第三個因素為代價。

資源

相關文件：

- Amazon 建置者資料中心
- QuickSight KPI
- Amazon CloudWatch RUM
- X-Ray 文件
- 了解恢復模式和權衡取捨以便在雲端中高效進行架構

相關影片：

- 透過 Amazon CloudWatch RUM 優化應用程式
- AWS re:Invent 2023 - 容量、可用性、成本效率：挑選三項
- AWS re:Invent 2023 - 鬆耦合系統的進階整合模式和權衡

相關範例：

- 使用 Amazon CloudWatch Synthetics 測量頁面載入時間
- Amazon CloudWatch RUM Web 用戶端

PERF01-BP05 使用政策和參考架構

選擇服務和組態時，使用內部政策和現有的參考架構，以便在設計和實作工作負載提高效率。

常見的反模式：

- 您允許各種可能會影響公司管理開銷的技術。

建立此最佳實務的優勢：為架構、技術和供應商選擇制定政策，可讓您快速做出決策。

未建立此最佳實務時的曝險等級：中

實作指引

在選擇資源和架構方面擁有內部政策，提供在選擇架構時要遵循的標準和準則。這些準則可簡化在選擇合適的雲端服務時的決策過程，並有助於提高效能效率。使用政策或參考架構來部署工作負載。將服務整合到您的雲端部署，然後使用效能測試以確認您可以繼續滿足效能需求。

實作步驟

- 清楚了解雲端工作負載的需求。
- 檢閱內部和外部政策，以識別最相關的政策。
- 使用 AWS 提供的適當參考架構或您的業界最佳實務。
- 針對常見情況，建立包含政策、標準、參考架構和規範指引的連續體。這樣做可以讓您的團隊更快地行動。如果適用，為您的垂直發展量身打造資產。
- 針對沙盒環境中的工作負載，驗證這些政策和參考架構。
- 隨時 up-to-date掌握業界標準和 AWS 更新，確保您的政策和參考架構有助於最佳化您的雲端工作負載。

資源

相關文件：

- AWS 架構中心
- AWS Partner Network
- AWS 解決方案程式庫
- AWS 知識中心
- AWS 架構部落格

相關影片：

- This is my Architecture
- AWS re：Invent 2022 - 使用 SAP & AWS reference 架構為您的企業加速價值

相關範例：

- AWS 範例
- AWS SDK 範例

PERF01-BP06 使用基準測試來推動架構決策

對現有工作負載的效能進行基準化分析，以了解工作負載在雲端的效能，並根據該資料推動架構決策。

常見的反模式：

- 您倚賴不代表工作負載特性的常見基準。
- 您將客戶的意見回饋和看法作為唯一基準。

建立此最佳實務的優勢：對目前的實作進行基準測試可讓您衡量效能改進。

未建立此最佳實務時的曝險等級：中

實作指引

使用基準化分析搭配綜合測試，以評估工作負載元件的效能。與負載測試相比，基準化分析通常速度更快；要評估特定元件的技術時，會使用基準化分析。當您缺少執行負載測試的完整解決方案時，通常可在新專案開始時使用基準化分析。

您可以建置自己的自訂基準化分析測試，也可以使用產業標準測試，例如 TPC-DS，對工作負載進行基準化分析。比較環境時，產業基準化分析很有幫助。對於確定您希望在架構中進行的特定營運類型，自訂基準化分析非常實用。

基準化分析時，務必要預熱測試環境，以獲得有效結果。多次執行相同的基準化分析，以確認您已擷取到隨時間推移出現的任何變化。

由於基準化分析的速度通常比負載測試要快，因此可以在部署管道中盡早使用基準化分析，以便能更快提供有關效能偏差的回饋。當您評估元件或服務中的重大變更時，藉助基準化分析，您可以更快速地查看所做的變更是否合理。請務必使用基準化分析搭配負載測試，因為負載測試將告訴您工作負載在生產中的效能。

實作步驟

- 規劃和定義：

  - 為基準化分析定義目標、基準、測試案例、指標 (例如 CPU 使用率、延遲或輸送量) 以及 KPI。
  - 關注使用者體驗方面的使用者需求，以及回應時間和可存取性等因素。
  - 找出工作負載適用的基準化分析工具。可以使用 AWS 服務 (例如 Amazon CloudWatch)，或與工作負載相容的第三方工具。
- 配置並檢測：
  - 設定環境並配置資源。
  - 實作監控和日誌記錄以擷取測試結果。
- 基準化分析和監控：
  - 在測試期間執行基準化分析並監控指標。
- 分析並記錄：
  - 記錄基準化分析過程和調查結果。
  - 分析結果以找出瓶頸、趨勢和需要改善的領域。
  - 使用測試結果做出架構決策並調整工作負載。這可能包括變更服務或採用新功能。
- 最佳化並重複：
  - 根據您的基準化分析來調整資源配置和分配。
  - 調整後重新測試您的工作負載，以驗證改進。
  - 記錄您的學習，並重複此過程以確定其他有待改進的領域。

資源

相關文件：

- AWS 架構中心
- AWS Partner Network
- AWS 解決方案程式庫
- AWS 知識中心
- Amazon CloudWatch RUM
- Amazon CloudWatch Synthetics
- 基因體工作流程，第 5 部分：自動化基準測試
- 對 Amazon SageMaker AI JumpStart 中的端點部署進行基準測試和最佳化

相關影片：

- AWS re:Invent 2023 - 對 AWS Lambda 冷啟動進行基準測試
- 在雲端中對有狀態服務進行基準測試
- This is my Architecture
- 透過 Amazon CloudWatch RUM 優化應用程式
- Amazon CloudWatch Synthetics 的示範

相關範例：

- AWS 範例
- AWS SDK 範例
- 分散式負載測試
- 使用 Amazon CloudWatch Synthetics 測量頁面載入時間
- Amazon CloudWatch RUM Web 用戶端

PERF01-BP07 使用資料驅動方法進行架構選擇

為架構選擇定義清晰、資料驅動型方法，以確認是否使用正確的雲端服務和組態，來滿足特定業務需求。

常見的反模式：

- 您假設目前的架構是靜態的，且不應隨著時間而更新。
- 您的架構選擇基於猜測和假設。
- 您會隨時間導入架構變更，而且無需理由佐證。

建立此最佳實務的優勢：透過採用明確定義的方法來做出架構選擇，您可以使用資料來影響工作負載設計，並隨著時間的推移做出明智的決策。

未建立此最佳實務時的曝險等級：中

實作指引

使用雲端或外部資源 (例如已發布的使用案例或白皮書) 的內部經驗和知識，在架構中選擇資源和服務。您應擁有一個明確定義的流程，鼓勵對工作負載中可能使用的服務進行實驗和基準化分析。

關鍵工作負載的待辦項目不僅應包括可提供與業務和使用者相關的功能的使用者故事，還包括構成工作負載架構跑道的技術故事。這條跑道了解科技和新服務的新進展，並根據資料和適當理由採用這些技術和新服務。這證明該架構仍然面向未來，不會停滯不前。

實作步驟

- 與關鍵利益相關者互動，以定義工作負載需求，包括效能、可用性和成本考量。考慮工作負載的使用者數量和使用模式等因素。
- 建立架構跑道或技術待辦項目，系統會優先處理這些項目與功能待辦事項。
- 評價和評估不同的雲端服務 (如需詳細資訊，請參閱 PERF01-BP01 了解和了解可用的雲端服務和功能)。
- 探索符合效能需求的不同架構模式，例如微型服務或無伺服器 (如需詳細資訊，請參閱 PERF01-BP02 使用雲端供應商或適當合作夥伴提供的指引，了解架構模式和最佳實務)。
- 諮詢其他團隊、架構圖和資源，例如 AWS 解決方案架構師、AWS 架構中心 和 AWS Partner Network，以協助您為工作負載選擇正確的架構。
- 定義輸送量和回應時間等效能指標，以協助您評估工作負載的效能。
- 實驗並使用定義的指標來驗證所選架構的效能。
- 視需要持續監控並進行調整，以維持架構的最佳效能。
- 記錄您選擇的架構和決策，作為未來更新和學習的參考。
- 根據學習、新技術和指標 (其指出目前方法中需要的變更或問題)，持續檢閱和更新架構選擇方法。

資源

相關文件：

- AWS 解決方案程式庫
- AWS 知識中心
- 在 上建置 End-to-End資料驅動應用程式的架構模式 AWS

相關影片：

- This is my Architecture

- AWS re：Invent 2021 - 資料驅動的企業：從願景到價值
- AWS re：Invent 2022 - 提供永續、高效能的架構
- AWS re：Invent 2023 - 最佳化成本和效能，並追蹤緩解進度
- AWS re：Invent 2022 - AWS optimization：立即結果的可行步驟

相關範例：

- AWS 範例
- AWS SDK 範例

運算與硬體

特定工作負載的最佳運算選擇會根據應用程式設計、使用模式和組態設定而有所不同。架構會針對不同元件使用不同運算選擇，並採用不同功能以提升效能。若選錯運算資源，可能使架構的效能達成效率降低。

這個重點領域分享了如何識別和最佳化運算選項，以在雲端中實現效能達成效率的各種指引和最佳實務。

最佳實務

- PERF02-BP01 為您的工作負載選取最佳運算選項
- PERF02-BP02 了解可用的運算組態和功能
- PERF02-BP03 收集運算相關指標
- PERF02-BP04 設定和正確大小的運算資源
- PERF02-BP05 動態擴展運算資源
- PERF02-BP06 使用最佳化的硬體型運算加速器

PERF02-BP01 為您的工作負載選取最佳運算選項

為工作負載選擇最合適的運算選項，可讓您改善效能、減少不必要的基礎設施成本，並降低維護工作負載所需的作業工作量。

常見的反模式：

- 您使用曾用於內部部署的同一個運算選項。
- 缺乏對雲端運算選項、特徵以及解決方案，以及那些解決方案可以如何改善運算效能的認識。
- 您在替代運算選項更精確地符合工作負載特性時，過度佈建現有運算選項以符合擴展或效能需求。

建立此最佳實務的優勢：可以透過找出運算需求並根據可用選項進行評估，提高工作負載的資源效率。

未建立此最佳實務時的曝險等級：高

實作指引

為了最佳化雲端工作負載以取得效能效率，請務必針對您的使用案例和效能需求選擇最適當的運算選項。 AWS 提供各種運算選項，以因應雲端中的不同工作負載。例如，您可以使用 Amazon EC2 啟動

和管理虛擬伺服器、AWS Lambda執行程式碼，而不必佈建或管理伺服器、Amazon ECS 或 Amazon EKS 執行和管理容器，或平行AWS Batch處理大量資料。根據擴展和運算需求，您應該根據自己的情況選擇並設定最佳的運算解決方案。也可以考慮在單一工作負載中使用多種類型的運算解決方案，因為每種運算解決方案都有自己的優點和缺點。

下列步驟會引導您選取正確的運算選項，以符合您的工作負載特性和效能需求。

實作步驟

- 了解工作負載運算需求。需要考慮的關鍵需求包括處理需求、流量模式、資料存取模式、擴展需求和延遲需求。
- 了解適用於工作負載的不同 AWS 運算服務。如需詳細資訊，請參閱PERF01-BP01 了解和了解可用的雲端服務和功能。以下是一些關鍵的 AWS 運算選項、其特性和常見使用案例：

AWS 服務	重要特性	常用案例
Amazon Elastic Compute Cloud（Amazon EC2）	擁有專為硬體、授權要求、大規模選取的不同執行個體系列、處理器類型與運算加速器設計的選項	平移遷移、整合型應用程式、混合環境、企業應用程式
Amazon Elastic Container Service（Amazon ECS）、Amazon Elastic Kubernetes Service（Amazon EKS）	輕鬆的部署、一致的環境、可擴展	微型服務、混合環境
AWS Lambda	無伺服器運算服務可執行程式碼以回應事件，並自動管理基礎運算資源。	微型服務、事件驅動型應用程式
AWS Batch	有效且動態地佈建和擴展Amazon Elastic Container Service（Amazon ECS）、Amazon Elastic Kubernetes Service（Amazon EKS）和AWS Fargate運算資源，並根據您	HPC，訓練 ML 模型

AWS 服務	重要特性	常用案例
	的任務需求選擇使用隨需或 Spot 執行個體	
Amazon Lightsail	預先設定用於執行小型工作負載的 Linux 和 Windows 應用程式	簡易網路應用程式、自訂的網站

- 評估與每個運算選項相關聯的成本 (例如每小時費用或資料傳輸) 和管理開銷 (例如修補和擴展)。
- 在非生產環境中執行實驗和基準測試，以確定哪個運算選項最能滿足您的工作負載需求。
- 在您試驗和找出新的運算解決方案，請規劃遷移並驗證效能指標。
- 使用 Amazon CloudWatch 和最佳化服務等 AWS 監控工具AWS Compute Optimizer，根據實際使用情況模式持續最佳化運算資源。

資源

相關文件：

- 使用 AWS進行雲端運算
- Amazon EC2執行個體類型
- Amazon EKS Containers：Amazon EKS Worker Nodes
- Amazon ECS Containers：Amazon ECS Container Instances
- 函數：Lambda 函數組態
- 容器的規範性指引
- 無伺服器的規範性指引

相關影片：

- AWS re：Invent 2023 - AWS Graviton：工作負載 AWS 的最佳價格效能
- AWS re：Invent 2023 - 中的新 Amazon Elastic Compute Cloud 生成 AI 功能 AMS
- AWS re:Invent 2023 - Amazon Elastic Compute Cloud 的最新消息
- AWS re:Invent 2023 - 智慧型節約：Amazon Elastic Compute Cloud 成本最佳化策略
- AWS re:Invent 2021 - 為新一代 Amazon Elastic Compute Cloud 提供支援：深入研究 Nitro 系統

- AWS re：Invent 2019 - 最佳化 AWS 運算的效能和成本
- AWS re:Invent 2019 - Amazon Elastic Compute Cloud 基礎
- AWS re：Invent 2022 - 部署 ML 模型，以高效能和低成本進行推論
- AWS re：Invent 2019 - 最佳化 AWS 運算的效能和成本
- Amazon EC2基礎
- 以高效能和低成本部署機器學習 (ML) 模型以進行推論

相關範例：

- 遷移 Web 應用程式至容器
- 執行 Serverless Hello World
- Amazon EKS研討會
- Amazon EC2研討會
- 使用 Amazon Elastic Compute Cloud 自動擴展實現高效且彈性的工作負載
- 使用 Container Services 遷移至 AWS Graviton

PERF02-BP02 了解可用的運算組態和功能

了解運算服務的可用組態選項和特徵，有助您佈建適量的資源並提高效能效率。

常見的反模式：

- 您沒有根據工作負載特性，評估運算選項或可用的執行個體系列。
- 過度佈建運算資源以符合尖峰需求。

建立此最佳實務的優點：熟悉 AWS 運算功能和組態，以便您可以使用最佳化的運算解決方案來滿足您的工作負載特性和需求。

未建立此最佳實務時的曝險等級：中

實作指引

每個運算解決方案都有獨特的組態和功能，以支援不同的工作負載特性和需求。了解這些選項如何與您的工作負載互補，並確定哪種組態選項最適合您的應用程式。這些選項的範例包括執行個體系列、大小、功能（GPU、I/O）、爆量、逾時、函數大小、容器執行個體和並行。如果您的工作負載已使用相

同的運算選項超過四週，且您預期未來特性將保持不變，您可以使用 AWS Compute Optimizer來了解目前的運算選項是否適合 CPU和 記憶體角度的工作負載。

實作步驟

- 了解工作負載需求（例如CPU需求、記憶體和延遲）。
- 檢閱 AWS 文件和最佳實務，以了解有助於改善運算效能的建議組態選項。以下是一些需要考慮的關鍵組態選項：

組態選項	範例
執行個體類型	• 運算最佳化執行個體非常適合需要較高 vCPU 與記憶體比率的工作負載。 • 記憶體最佳化執行個體提供大量記憶體，以支援記憶體密集型工作負載。 • 儲存體最佳化執行個體專為需要對本機儲存體進行高連續讀取和寫入存取權（IOPS）的工作負載而設計。
定價方式	• 隨需執行個體允許您按秒數或時數來使用運算能力，無須簽訂長期合約。這些執行個體適合於超過效能基準需求的爆量。 • Savings Plans 可大幅節省隨需執行個體，以換取在一年或三年內使用特定運算能力的承諾。 • Spot 執行個體可讓您以折扣價利用未使用的執行個體容量，用於無狀態、容錯的工作負載。
Auto Scaling	使用 Auto Scaling 設定，使運算資源與流量模式相符。
規模調整	• 使用 Compute Optimizer 取得機器學習支援的建議，了解哪些運算組態最符合您的運算特性。

組態選項	範例
	• 使用 AWS Lambda Power Tuning 為您的 Lambda 函數選擇最佳組態。
硬體型運算加速器	• 加速運算執行個體會比 CPU型替代方案更有效率地執行圖形處理或資料模式比對等函數。 • 對於機器學習工作負載，請善用工作負載特有的專用硬體，例如AWS Trainium 、AWS Inferentia 和Amazon EC2 DL1

資源

相關文件：

- 使用 AWS進行雲端運算
- Amazon EC2執行個體類型
- Amazon EC2執行個體的處理器狀態控制
- Amazon EKS Containers：Amazon EKS Worker Nodes
- Amazon ECS Containers：Amazon ECS Container Instances
- 函數：Lambda 函數組態

相關影片：

- AWS re：Invent 2023 – AWS Graviton： AWS 工作負載的最佳價格效能
- AWS re：Invent 2023 – 中的新 Amazon EC2 生成 AI 功能 AWS Management Console
- AWS re：Invent 2023 – Amazon 的新功能 EC2
- AWS re：Invent 2023 – 智慧節省：Amazon EC2成本最佳化策略
- AWS re：Invent 2021 – 推動下一代 AmazonEC2：Nitro 系統的深度探索
- AWS re：Invent 2019 – Amazon EC2基礎
- AWS re：Invent 2022 – 最佳化 Amazon EKS 的效能和成本 AWS

相關範例：

- 運算最佳化工具示範程式碼
- Amazon EC2 Spot 執行個體研討會
- Amazon 的高效率和彈性工作負載 EC2 AWS Auto Scaling
- Graviton 開發人員研討會
- AWS 適用於 Microsoft 工作負載沉浸式日
- AWS 適用於 Linux 工作負載沉浸式日
- AWS Compute Optimizer 示範程式碼
- Amazon EKS研討會

PERF02-BP03 收集運算相關指標

記錄並追蹤與運算相關的指標，進一步了解運算資源的效能，並改善效能及使用率。

常見的反模式：

- 您只使用手動日誌檔案來搜尋指標。
- 您只會使用監控軟體記錄的預設指標。
- 您只會在有問題時審查指標。

建立此最佳實務的優勢：收集效能相關指標有助於使應用程式效能與業務需求保持一致，確保符合工作負載需求。這麼做也可以協助您持續改善工作負載中的資源效能和使用率。

未建立此最佳實務時的曝險等級：高

實作指引

雲端工作負載可以產生大量資料，例如指標、日誌和事件。在 中 AWS 雲端，收集指標是改善安全性、成本效益、效能和永續性的重要步驟。 使用 Amazon CloudWatch 等監控服務 AWS ，提供廣泛的效能相關指標，為您提供寶貴的洞見。CPU 使用率、記憶體使用率、磁碟 I/O 和網路傳入和傳出等指標可以提供使用率層級或效能瓶頸的洞察。將這些指標納入資料驅動的方法，以主動調整和優化工作負載的資源。 在理想的情況下，應該在單一平台中收集與運算資源相關的所有指標，並實作保留政策以支援成本和營運目標。

實作步驟

- 識別與您的工作負載相關的效能相關指標。您應該收集與資源使用率和雲端工作負載運作方式有關的指標 (例如回應時間和輸送量)。
  - Amazon EC2 預設指標
  - Amazon ECS 預設指標
  - Amazon EKS 預設指標
  - Lambda 預設指標
  - Amazon EC2記憶體和磁碟指標
- 為工作負載選擇並設定合適的日誌記錄和監控解決方案。
  - AWS 原生可觀測性
  - AWS Distro for OpenTelemetry
  - Amazon Managed Service for Prometheus
- 根據工作負載需求，為指標定義必要的篩選條件和彙總。
  - 使用 Amazon CloudWatch Logs 和指標篩選條件量化自訂應用程式指標
  - 使用 Amazon CloudWatch 策略標記收集自訂指標
- 為指標設定資料保留政策，以符合安全性和營運目標。
  - CloudWatch 指標的預設資料保留
  - CloudWatch 日誌的預設資料保留
- 如有必要，為指標建立警示和通知，可協助您主動回應效能相關問題。
  - 使用 Amazon CloudWatch 異常偵測建立自訂指標的警示
  - 使用 Amazon 為特定網頁建立指標和警示 CloudWatch RUM
- 使用自動化來部署指標和記錄彙總代理程式。
  - AWS Systems Manager 自動化
  - OpenTelemetry 收集器

資源

相關文件：

- 監控與可觀測性
- 最佳實務：使用 實作可觀測性 AWS

- Amazon CloudWatch 文件
- 使用 CloudWatch 代理程式從 Amazon EC2執行個體和內部部署伺服器收集指標和日誌
- 存取 的 Amazon CloudWatch Logs AWS Lambda
- 將 CloudWatch 日誌與容器執行個體搭配使用
- 發佈自訂指標
- AWS Answers：集中式日誌記錄
- AWS 發佈 CloudWatch 指標的服務
- 在 EKS上監控 Amazon AWS Fargate

相關影片：

- AWS re：Invent 2023 – 【LAUNCH】 現代工作負載的應用程式監控
- AWS re：Invent 2023 – 實作應用程式可觀測性
- AWS re：Invent 2023 – 建立有效的可觀測性策略
- AWS re：Invent 2023 – AWS Distro for 的無縫可觀測性 OpenTelemetry
- 上的應用程式效能管理 AWS

相關範例：

- AWS 適用於 Linux Workloads Immersion Day - Amazon CloudWatch
- 監控 Amazon ECS叢集和容器
- 使用 Amazon CloudWatch 儀表板進行監控
- Amazon EKS研討會

PERF02-BP04 設定和正確大小的運算資源

設定運算資源及適當調整其大小，以符合工作負載的效能需求，並避免未充分使用資源或過度使用資源的情況。

常見的反模式：

- 您忽略工作負載效能需求，導致過度佈建或佈建不足的運算資源。
- 您只選擇適用於所有工作負載的最大或最小執行個體。

- 為了方便管理，只能使用一個執行個體系列。
- 您可以忽略 AWS Cost Explorer 或 Compute Optimizer 的建議，以進行正確調整大小。
- 您未重新評估工作負載是否適用於新執行個體類型。
- 您只驗證組織的少量執行個體組態。

建立此最佳實務的優勢：透過避免資源的過度佈建和佈建不足，適當調整運算資源的大小可確保雲端中的最佳操作。適當調整運算資源的大小，通常可以提高效能和增強客戶體驗，同時降低成本。

未建立此最佳實務時的曝險等級：中

實作指引

適當調整大小可讓組織以有效率且符合成本效益的方式操作雲端基礎架構，同時滿足其業務需求。過度佈建雲端資源可能會導致額外的成本，而佈建不足可能會導致效能不佳和負面的客戶體驗。 AWS 提供 AWS Compute Optimizer和 等工具AWS Trusted Advisor，使用歷史資料提供建議，以正確調整運算資源的大小。

實作步驟

- 選擇最適合您需求的執行個體類型：
  - 如何為工作負載選擇適當的 Amazon EC2執行個體類型？
  - Amazon EC2 Fleet 的屬性型執行個體類型選擇
  - 使用屬性型執行個體類型選取範圍來建立 Auto Scaling 群組
  - 利用 Karpenter 整合來最佳化 Kubernetes 運算成本
- 分析工作負載的各種效能特性，以及這些特性與記憶體、網路和CPU用量之間的關係。使用此資料，可以選擇最適合您工作負載描述檔和效能目標的資源。
- 使用 Amazon 等監控工具來 AWS 監控資源用量 CloudWatch。
- 為運算資源選取適合的組態。
  - 對於暫時性工作負載，評估執行個體 Amazon CloudWatch 指標，例如 `CPUUtilization`，以識別執行個體是否使用不足或過度使用。
  - 對於穩定的工作負載，定期檢查AWS Compute Optimizer和 等工具 AWS 的許可化 AWS Trusted Advisor ，以識別最佳化和調整運算資源大小的機會。
- 在即時環境中實作之前，先測試非生產環境中的組態變更。
- 持續重新評估新的運算供應項目，並且根據工作負載需求進行比較。

資源

相關文件：

- 使用 進行雲端運算 AWS
- Amazon EC2執行個體類型
- Amazon ECS Containers：Amazon ECS Container Instances
- Amazon EKS Containers：Amazon EKS Worker Nodes
- 函數：Lambda 函數組態
- Amazon EC2執行個體的處理器狀態控制

相關影片：

- Amazon EC2基礎
- AWS re：Invent 2023 – AWS Graviton： AWS 工作負載的最佳價格效能
- AWS re：Invent 2023 – 中的新 Amazon EC2 生成 AI 功能 AWS Management Console
- AWS re：Invent 2023 – Amazon 的新功能 EC2
- AWS re：Invent 2023 – 智慧節省：Amazon EC2成本最佳化策略
- AWS re：Invent 2021 – 推動下一代 AmazonEC2：Nitro 系統的深度探索
- AWS re：Invent 2019 – Amazon EC2基礎

相關範例：

- AWS Compute Optimizer 示範程式碼
- Amazon EKS研討會
- 適當調整大小的建議

PERF02-BP05 動態擴展運算資源

為滿足需求，請使用雲端的彈性，來動態擴充或縮減運算資源，並避免為工作負載佈建過多或過少的容量。

常見的反模式：

- 您可以手動增加容量，對警示做出反應。

- 使用與內部部署相同的大小規模準則 (通常是靜態基礎設施)。
- 您在擴展事件之後維持增加容量，而不是縮減規模。

建立此最佳實務的優勢：設定和測試運算資源的彈性可協助您節省成本、維持效能基準，並隨著流量變化提升可靠性。

未建立此最佳實務時的曝險等級：高

實作指引

AWS 透過各種擴展機制，提供彈性來動態擴展或縮減資源，以滿足需求的變化。結合與運算相關的指標，動態擴展允許工作負載自動回應變更，並使用最佳運算資源集來實現目標。

您可以使用多種不同的方法達到資源的供需平衡。

- 目標追蹤法：監控您的擴展指標，並視需要自動增加或減少容量。
- 預測擴展：縮減每日和每週趨勢的預期。
- 基於排程的方法：按照排程來擴展可讓您根據可預測的負載變化來設定自己的擴展排程。
- 服務擴展：選擇可根據設計自動擴展的服務 (例如無伺服器)。

您必須確保工作負載部署可以同時處理向上擴展和縮減規模事件。

實作步驟

- 運算執行個體、容器和函數提供了彈性機制，可與自動擴展功能結合使用，或是作為服務功能提供。以下是自動擴展機制的幾個範例：

自動擴展機制	在哪裡使用
Amazon EC2 Auto Scaling	為確保您擁有正確數量的 Amazon EC2 執行個體，可處理應用程式的使用者負載。
Application Auto Scaling	自動擴展 Amazon 以外的個別 AWS 服務資源，EC2例如 AWS Lambda 函數或 Amazon Elastic Container Service （AmazonECS） 服務。
Kubernetes Cluster Autoscaler/Karpenter	自動擴展 Kubernetes 叢集。

- 擴展通常與運算服務相關，例如 Amazon EC2執行個體或 AWS Lambda 函數。請務必同時考慮非運算服務的組態 (例如 AWS Glue) 以符合需求。
- 確認用於擴展的指標符合要部署之工作負載的特性。如果您正在部署影片轉碼應用程式，則預期 100% CPU使用率，不應是您的主要指標。請改用轉碼任務佇列的深度。您可以將自訂指標用於擴展政策 (如有必要)。若要選擇正確的指標，請考慮下列 Amazon 指南EC2：
  - 指標應為有效的使用率指標，並說明執行個體的忙碌程度。
  - 指標值必須與 Auto Scaling 群組中的執行個體數成比例增加或減少。
- 對於 Auto Scaling 群組，確保使用動態擴展，而非手動擴展。我們也建議您在動態擴展中使用目標追蹤擴展政策。
- 確認工作負載部署可同時處理擴展事件 (擴充和縮減)。例如，您可以使用活動歷史記錄來驗證 Auto Scaling 群組的擴展活動。
- 評估工作負載以取得可預測模式，並在預計發生預測中的變化和隨需規劃變化時主動擴展。透過預測性擴展，可以消除過度佈建容量的需求。如需詳細資訊，請參閱使用 Amazon EC2 Auto Scaling 進行預測擴展。

資源

相關文件：

- 使用 進行雲端運算 AWS
- Amazon EC2執行個體類型
- Amazon ECS Containers：Amazon ECS Container Instances
- Amazon EKS Containers：Amazon EKS Worker Nodes
- 函數：Lambda 函數組態
- Amazon EC2執行個體的處理器狀態控制
- 深入探討 Amazon ECS Cluster Auto Scaling
- 介紹 Karpenter - 一個開放原始碼的高效能 Kubernetes Cluster Autoscaler

相關影片：

- AWS re：Invent 2023 – AWS Graviton： AWS 工作負載的最佳價格效能
- AWS re：Invent 2023 – AWS 管理主控台中的新 Amazon EC2 生成 AI 功能
- AWS re：Invent 2023 – Amazon 的新功能 EC2

- AWS re：Invent 2023 – 智慧節省：Amazon EC2成本最佳化策略
- AWS re：Invent 2021 – 推動下一代 AmazonEC2：Nitro 系統的深度探索
- AWS re：Invent 2019 – Amazon EC2基礎

相關範例：

- Amazon EC2 Auto Scaling 群組範例
- Amazon EKS研討會
- 透過在 上執行來擴展 Amazon EKS工作負載 IPv6

PERF02-BP06 使用最佳化的硬體型運算加速器

使用硬體加速器執行特定功能，比以 CPU 為基礎的替代方案更有效率。

常見的反模式：

- 在工作負載中，您尚未基準化分析一般用途執行個體和專用執行個體，而專用執行個體可以改善效能和降低成本。
- 您使用硬體型運算加速器來執行任務，比起使用以 CPU 為基礎的替代方案更有效率。
- 未監控 GPU 使用率。

建立此最佳實務的優勢：透過使用硬體型加速器，例如圖形處理單元 (GPU) 和現場可程式化閘道陣列 (FPGA)，您就可以更有效率地執行特定處理功能。

未建立此最佳實務時的曝險等級：中

實作指引

加速運算執行個體可讓您存取硬體型運算加速器，例如 GPU 和 FPGA。這些硬體加速器比基於 CPU 的替代品更有效地執行某些功能，例如圖形處理或資料模式匹配。許多加速的工作負載 (例如轉譯、轉碼和機器學習) 在資源使用方面變化很大。只在需要時執行此硬體，並在不需要時自動停用它們，以提高整體效能的效率。

實作步驟

- 確定哪些加速運算執行個體可以滿足您的需求。

- 針對機器學習工作負載，請利用專供工作負載使用的專用硬體，例如 AWS Trainium、AWS Inferentia 和 Amazon EC2 DL1。AWS與同類 Amazon EC2 執行個體相比，Inferentia 執行個體 (例如 Inf2 執行個體) 所提供的效能功耗比要高出 50%。
- 收集加速運算執行個體的用量指標。例如，可以使用 CloudWatch 代理程式為您的 GPU 收集 `utilization_gpu` 和 `utilization_memory` 等指標，如使用 Amazon CloudWatch 收集 NVIDIA GPU 指標中所示。
- 優化硬體加速器的程式碼、網路運作和設定，以確保系統會充分利用基礎硬體。
  - 最佳化 GPU 設定
  - 深度學習 AMI 中的 GPU 監控和最佳化
  - 將 I/O 最佳化以針對 Amazon SageMaker AI 中的深度學習訓練進行 GPU 效能調校
- 使用最新的高效能程式庫和 GPU 驅動程式。
- 使用自動化來釋出不使用的 GPU 執行個體。

資源

相關文件：

- 在 Amazon Elastic Container Service 上使用 GPU
- GPU 執行個體
- AWS Trainium 的執行個體
- AWS Inferentia 的執行個體
- 開始建構吧！使用自訂晶片和加速器來進行建構

- 加速運算
- Amazon EC2 VT1 執行個體
- 如何為工作負載選擇適當的 Amazon EC2 執行個體類型？
- 選擇最佳 AI 加速器和模型編譯來以 Amazon SageMaker AI 推斷電腦視覺

相關影片：

- AWS re:Invent 2021 - 如何選擇 Amazon Elastic Compute Cloud GPU 執行個體進行深度學習
- AWS re:Invent 2022 - [最新發佈！] 介紹基於 AWS Inferentia2 的 Amazon EC2 Inf2 執行個體
- AWS re:Invent 2022 - 使用 AWS Trainium 加速深度學習並加快創新速度

- AWS re:Invent 2022 - 透過 NVIDIA 在 AWS 上進行深度學習：從訓練到部署

相關範例：

- Amazon SageMaker AI 和 NVIDIA GPU Cloud (NGC)
- 搭配使用 SageMaker AI 與 Trainium 和 Inferentia，進行最佳化的深度學習訓練和推論工作負載
- 使用 Amazon SageMaker AI 中的 Amazon Elastic Compute Cloud Inf1 執行個體最佳化 NLP 模型

資料管理

特定系統的最佳資料管理解決方案會根據資料類型 (區塊、檔案或物件)、存取模式 (隨機或循序)、所需輸送量、存取頻率 (線上、離線、封存)、更新頻率 (WORM、動態) 及可用性和耐用性限制而有所不同。Well-Architected 工作負載會使用專用資料存放區，這些存放區採用不同的功能以提升效能。

這個重點領域分享了最佳化資料儲存、移動和存取模式，以及資料存放區效能達成效率的各種指引和最佳實務。

最佳實務

- PERF03-BP01 使用最能夠支援資料存取和儲存需求的專用資料存放區
- PERF03-BP02 評估資料存放區的可用組態選項
- PERF03-BP03 收集並記錄資料存放區效能指標
- PERF03-BP04 實作策略以提高資料存放區中的查詢效能
- PERF03-BP05 實作使用快取的資料存取模式

PERF03-BP01 使用最能夠支援資料存取和儲存需求的專用資料存放區

了解資料特性 (例如可共用、大小、快取大小、存取模式、延遲、輸送量和資料的持續性)，為工作負載選擇適合的專用資料存放區 (儲存或資料庫)。

常見的反模式：

- 由於具備某種特定類型資料庫解決方案的內部經驗和知識，您堅持使用某個資料存取區。
- 您假設所有工作負載都有類似的資料儲存和存取需求。
- 您未實作資料目錄以清查資料資產。

建立此最佳實務的優勢：了解資料特性和需求，可協助您判斷能滿足工作負載需求的最有效率且效能最高的儲存技術。

未建立此最佳實務時的曝險等級：高

實作指引

在選擇和實作資料儲存時，請確定查詢、擴展和儲存特性支援工作負載資料需求。 AWS 提供許多資料儲存和資料庫技術，包括區塊儲存、物件儲存、串流儲存、檔案系統、關聯性、索引鍵值、文件、記憶體內、圖形、時間序列和分類帳資料庫。每個資料管理解決方案都有為您提供的選項和組態，以支援您的使用案例和資料模型。透過了解資料特性和需求，您可以擺脫單片儲存技術，以及專注於適當管理資料的限制性 one-size-fits-all方法。

實作步驟

- 對您工作負載現有的各種資料類型執行清查。
- 了解並記錄資料特性和需求，包括：
  - 資料類型 (非結構化、半結構化、關聯式)
  - 資料量與成長
  - 資料耐用性：持續性、暫時性、臨時
  - ACID （原子性、一致性、隔離、耐久性） 需求
  - 資料存取模式 (大量讀取或大量寫入)
  - Latency (延遲)
  - 輸送量
  - IOPS （每秒的輸入/輸出操作）
  - 資料保留期間
- 了解適用於 工作負載的不同資料存放區 （儲存和資料庫服務） AWS ，其可滿足您的資料特性，如所述PERF01-BP01 了解和了解可用的雲端服務和功能。 AWS 儲存技術及其重要特性的一些範例包含：

類型	AWS 服務	重要特性
物件儲存	Amazon Simple Storage Service (Amazon S3)	具有不受限的可擴展性、高可用性，以及多個可存取性選項。對 Amazon S3 輸入和存取物件時，可以使用 Transfer Acceleration 或 Access Points 之類的服務來支援您的位置、安全需求和存取模式。

類型	AWS 服務	重要特性
封存儲存	Amazon S3 Glacier	專為資料封存而打造。
串流儲存空間	Amazon Kinesis Amazon Managed Streaming for Apache Kafka（Amazon MSK）	快速地擷取和儲存串流資料。
共用檔案系統	Amazon Elastic File System（Amazon EFS）	可供多種類型的運算解決方案存取的可掛載檔案系統。
共用檔案系統	Amazon FSx	以最新的 AWS 運算解決方案為基礎，支援四個常用的檔案系統： NetApp ONTAP、Open ZFS、Windows File Server 和 Lustre。FSxAmazonlatency、輸送量和 IOPS會因檔案系統而異，因此在選擇適合您工作負載需求的檔案系統時，應該考慮這一點。
區塊儲存	Amazon Elastic Block Store（Amazon EBS）	專為 Amazon Elastic Compute Cloud（Amazon ）設計的可擴展、高效能區塊儲存服務EC2。Amazon EBS包含交易IOPS型 密集型工作負載的 SSD後端儲存體，以及輸送量密集型工作負載的 HDD後端儲存體。

類型	AWS 服務	重要特性
關聯式資料庫	Amazon Aurora、Amazon RDS、Amazon Redshift 。	旨在支援 ACID（原子性、一致性、隔離、耐久性）交易，並維持參考完整性和強大的資料一致性。許多傳統應用程式、企業資源規劃（ERP）、客戶關係管理（CRM）和電子商務會使用關聯式資料庫來存放其資料。
鍵值資料庫	Amazon DynamoDB	已針對常見的存取模式進行最佳化，通常用於儲存和擷取大量資料。高流量 Web 應用程式、電子商務系統和遊戲應用程式是鍵值資料庫的典型使用案例。
文件資料庫	Amazon DocumentDB	旨在將半結構化資料儲存為 JSON類似 文件。這些資料庫可協助開發人員快速建置和更新應用程式，例如內容管理、目錄和使用者設定檔。
記憶體資料庫	Amazon ElastiCache、Amazon MemoryDB for Redis	適用於需要即時存取資料、最低延遲和最高輸送量的應用程式。您可以將記憶體資料庫用於應用程式快取、工作階段管理、遊戲排行榜、低延遲 ML 特徵存放區、微型服務簡訊系統，以及高輸送量串流機制

類型	AWS 服務	重要特性
圖形資料庫	Amazon Neptune	適用於此類應用程式：必須在高度連線圖形資料集之間，大規模導覽和查詢數百萬個關係，並且在過程中僅有毫秒延遲。許多公司使用圖形資料庫進行詐騙偵測、社交聯網和推薦引擎。
時間序列資料庫	Amazon Timestream	可快速地從隨時間變化的資料收集、合成和衍生洞見。IoT 應用程式 DevOps和工業遙測可以利用時間序列資料庫。
寬欄	Amazon Keyspaces (適用於 Apache Cassandra)	可使用表格、列和欄，但與關聯式資料庫不同，在同一個表格中，欄的名稱和格式會因列而異。您通常會在大規模工業應用程式中看到寬欄存放區，用於設備維護、叢集管理和路由優化。
總帳	Amazon Quantum Ledger 資料庫（Amazon QLDB）	可提供集中化且受信任的機構，為每個應用程式維護可擴展、不可變且以密碼編譯方式驗證的交易記錄。我們會看到用於記錄、供應鏈、註冊甚至銀行交易系統的總帳資料庫。

- 如果您要建置資料平台，請在 上利用現代資料架構 AWS 來整合您的資料湖、資料倉儲和專用資料存放區。
- 為工作負載選擇資料存放區時，需要考慮的關鍵問題如下：

問題	重要考慮事項
如何建構資料？	• 如果資料是非結構化的，請考慮 Amazon S3 之類的物件存放區或 Amazon DocumentDB 之類的無SQL資料庫 • 對於鍵值資料，請考慮 DynamoDB、Amazon ElastiCache（Redis OSS）或 Amazon MemoryDB
需要哪種層級的參考完整性？	• 對於外部金鑰限制，Amazon RDS 和 Aurora 等關聯式資料庫可以提供此完整性層級。 • 一般而言，在無SQL資料模型中，您會將資料還原成單一文件或要在單一請求中擷取的文件集合，而不是跨文件或資料表加入。
是否需要符合 ACID（原子性、一致性、隔離、耐久性）要求？	• 如果需要與關聯式資料庫相關聯的ACID屬性，請考慮關聯式資料庫，例如 Amazon RDS 和 Aurora。 • 如果沒有SQL資料庫 需要強烈的一致性，您可以搭配 DynamoDB使用強烈的一致性讀取。
儲存要求如何隨時間變更？ 這如何影響可擴展性？	• DynamoDB 和 Amazon Quantum Ledger Database（Amazon QLDB）等無伺服器資料庫將動態擴展。 • 關聯式資料庫在佈建的儲存體上有上限，而且一旦達到這些限制，通常必須使用碎片等機制進行水平分割。
讀取查詢與寫入查詢的比例是多少？ 快取可能改善效能嗎？	• 如果DAX資料庫是 DynamoDB ，則大量讀取工作負載可以從快取層中受益，例如 ElastiCache或 。 DynamoDB • 讀取也可以卸載，以讀取具有關聯式資料庫的複本，例如 Amazon RDS。

問題	重要考慮事項
儲存和修改（OLTP - 線上交易處理）或擷取和報告（OLAP - 線上分析處理）是否具有更高的優先順序？	• 對於高輸送量讀取即交易處理，請考慮無SQL資料庫，例如 DynamoDB 。 • 對於具有一致性的高輸送量和複雜的讀取模式（如聯結），請使用 Amazon RDS。 • 對於分析查詢，請考慮 Amazon Redshift 之類的資料欄式資料庫，或將資料匯出至 Amazon S3，並使用 Athena 或 Amazon QuickSight執行分析。
資料需要哪種層級的耐久性？	• Aurora 會自動跨區域內的三個可用區域複寫資料，這表示資料高度耐用且資料遺失的機會較低。 • DynamoDB 會自動跨多個可用區域複寫，具有高可用性和資料耐久性。 • Amazon S3 提供 11 個九的耐用性。許多資料庫服務，例如 Amazon RDS和 DynamoDB ，都支援將資料匯出至 Amazon S3 以進行長期保留和封存。
是否希望擺脫商務資料庫引擎、或授權成本？	• 考慮開放原始碼引擎，例如 PostgreSQL 和 MySQL on Amazon RDS或 Aurora。 • 利用 AWS Database Migration Service 和 AWS Schema Conversion Tool，從商務資料庫引擎遷移至開放原始碼
對資料庫的操作期望是什麼？ 移至受管服務是否為主要問題？	• 利用 Amazon RDS而非 Amazon EC2，以及使用 DynamoDB 或 Amazon DocumentDB 而非自我託管無SQL資料庫，可以降低營運開銷。

問題	重要考慮事項
目前如何存取資料庫？ 它是否僅存取應用程式，還是有商業智慧（BI）使用者和其他連線off-the-shelf的應用程式？	• 如果您依賴於外部工具，則可能必須保持與其所支援之資料庫的相容性。Amazon與其支援的引擎版本RDS完全相容，包括Microsoft SQL Server、Oracle、My SQL和Postgre SQL。

- 在非生產環境中執行實驗和基準測試，以確定哪個資料存放區最能滿足您的工作負載需求。

資源

相關文件：

- Amazon EBS磁碟區類型
- Amazon EC2 Storage
- AmazonEFS：Amazon EFS Performance
- Amazon FSx for Lustre 效能
- Amazon FSx for Windows File Server 效能
- Amazon S3 Glacier：S3 Glacier 文件
- Amazon S3：請求率和效能考量
- 使用 的雲端儲存 AWS
- Amazon EBS I/O 特性
- AWS的雲端資料庫
- AWS 資料庫快取
- DynamoDB Accelerator
- Amazon Aurora 最佳實務
- Amazon Redshift 效能
- Amazon Athena 10 大效能秘訣
- Amazon Redshift Spectrum 最佳實務
- Amazon DynamoDB 最佳實務
- 在 Amazon EC2和 Amazon 之間選擇 RDS

- 實作 Amazon 的最佳實務 ElastiCache

相關影片：

- AWS re：Invent 2023：改善 Amazon Elastic Block Store 效率並提高成本效益
- AWS re：Invent 2023：使用 Amazon Simple Storage Service 最佳化儲存價格和效能
- AWS re：Invent 2023：在 Amazon Simple Storage Service 上建置和最佳化資料湖
- AWS re：Invent 2022：在 上建置現代資料架構 AWS
- AWS re：Invent 2022：在 上建置資料網格架構 AWS
- AWS re：Invent 2023：深入探索 Amazon Aurora 及其創新
- AWS re：Invent 2023：使用 Amazon DynamoDB 進行進階資料建模
- AWS re：Invent 2022：使用專用資料庫現代化應用程式
- 深入探討 Amazon DynamoDB：進階設計模式

相關範例：

- AWS 目的建構資料庫研討會
- 專為開發人員打造的資料庫
- AWS 現代資料架構沉浸式日
- 在 上建置資料網格 AWS
- Amazon S3 範例
- 使用 Amazon Redshift 資料共用來最佳化資料模式
- 資料庫遷移
- MS SQL Server - AWS Database Migration Service（AWS DMS）複寫示範
- 資料庫現代化實際操作研討會
- Amazon Neptune 範例

PERF03-BP02 評估資料存放區的可用組態選項

了解並評估資料存放區可用的各種功能和組態選項，以最佳化工作負載的儲存空間和效能。

常見的反模式：

- EBS所有工作負載只使用一種儲存類型，例如 Amazon 。
- 您可以使用 佈建IOPS用於所有工作負載，而不需要針對所有儲存層進行實際測試。
- 您不知道所選資料管理解決方案的組態選項。
- 您完全依賴於增加執行個體大小，而不查看其他可用的組態選項。
- 您並不測試資料存放區的擴展特性。

建立此最佳實務的優勢：藉由探索和試驗資料存放區組態，您能夠降低基礎架構成本、改善效能，以及減少維護工作負載所需的工作量。

未建立此最佳實務時的曝險等級：中

實作指引

工作負載可以根據資料儲存和存取需求，使用一個或多個資料存放區。要優化效能達成效率和成本，您必須評估資料存取模式，以判斷適當的資料存放區組態。在探索資料存放區選項時，請考量各種層面，例如儲存選項、記憶體、運算、讀取複本、一致性要求、連線集區以及快取選項。嘗試使用這些不同的組態選項來改善效能達成效率指標。

實作步驟

- 了解資料存放區的目前組態 (例如執行個體類型、儲存體大小或資料庫引擎版本)。
- 檢閱 AWS 文件和最佳實務，以了解建議組態選項，以協助改善資料存放區的效能。要考慮的關鍵資料存放區選項如下：

組態選項	範例
卸載讀取 (例如讀取複本和快取)	• 對於 DynamoDB 資料表，您可以使用 卸載用於快取DAX的讀取。 • 您可以建立 Amazon ElastiCache （Redis OSS） 叢集，並將應用程式設定為先從快取讀取，如果請求的項目不存在，則會傳回資料庫。 • 諸如 Amazon RDS和 Aurora 等關聯式資料庫，以及佈建的SQL Neptune 和 Amazon DocumentDB 等資料庫都不支援新增僅供讀取複本來卸載工作負載的讀取部分。

組態選項	範例
	• 諸如 DynamoDB 等無伺服器資料庫將自動擴展。確保您已佈建足夠的讀取容量單位（RCU），以處理工作負載。
擴展寫入 (例如分區金鑰碎片或引進佇列)	• 對於關聯式資料庫，您可以增加執行個體的大小，以容納增加的工作負載，或增加佈建的 IOPs ，以允許基礎儲存的輸送量增加。 • 也可以在資料庫前面引入佇列，而不是直接寫入資料庫。此模式可讓您將擷取與資料庫分離並控制流速，這樣資料庫就不會不堪重負。 • 批次處理寫入請求，而不是建立許多短期交易，這有助於改善高寫入量關聯式資料庫的輸送量。 • DynamoDB 等無伺服器資料庫可以自動擴展寫入輸送量，也可以根據容量模式調整佈建的寫入容量單位（WCU）。 • 當您達到指定分割區索引鍵的輸送量限制時，仍然可能會遇到熱分割區的問題。透過選擇更均勻分佈的分割區索引鍵，或分片寫入分割區索引鍵，來緩解此問題。
使用政策來管理資料集的生命週期	• 可以使用 Amazon S3 生命週期，以在整個生命週期中管理物件。如果存取模式不明、會變化或是無法預測，則可以使用 Amazon S3 Intelligent-Tiering，讓其監控存取模式，並自動將未存取的物件移至成本較低的存取層。可以利用 Amazon S3 Storage Lens 指標，找出生命週期管理中的最佳化機會和差距。 • Amazon EFS生命週期管理會自動管理檔案系統的檔案儲存。

組態選項	範例
連線管理與集區	• Amazon RDS Proxy 可與 Amazon RDS和 Aurora 搭配使用，以管理資料庫的連線。 • 無伺服器資料庫 (例如 DynamoDB) 沒有與其相關聯的連線，但請考慮已佈建的容量和自動擴展政策來處理負載中的高峰。

- 在非生產環境中執行實驗和基準測試，以確定哪個組態選項能滿足您的工作負載需求。
- 完成試驗之後，請規劃遷移並確認效能指標。
- 使用 AWS 監控（例如 Amazon CloudWatch）和最佳化（例如 Amazon S3 Storage Lens）工具，使用實際用量模式持續最佳化您的資料存放區。

資源

相關文件：

- AWS的雲端儲存
- Amazon EBS磁碟區類型
- Amazon EC2 Storage
- AmazonEFS：Amazon EFS Performance
- Amazon FSx for Lustre 效能
- Amazon FSx for Windows File Server 效能
- Amazon S3 Glacier：S3 Glacier 文件
- Amazon S3：請求率和效能考量
- Amazon EBS I/O 特性
- AWS的雲端資料庫
- AWS 資料庫快取
- DynamoDB Accelerator
- Amazon Aurora 最佳實務
- Amazon Redshift 效能
- Amazon Athena 10 大效能秘訣
- Amazon Redshift Spectrum 最佳實務

- Amazon DynamoDB 最佳實務

相關影片：

- AWS re:Invent 2023：提高 Amazon Elastic Block Store 效率並更具成本效益
- AWS re:Invent 2023：使用 Amazon Simple Storage Service 最佳化儲存價格和效能
- AWS re:Invent 2023：在 Amazon Simple Storage Service 上建置和最佳化資料湖
- AWS re：Invent 2023： AWS 檔案儲存的新功能
- AWS re:Invent 2023：深入了解 Amazon DynamoDB

相關範例：

- AWS 目的建構資料庫研討會
- 專為開發人員打造的資料庫
- AWS 現代資料架構沉浸式日
- Amazon EBS Autoscale
- Amazon S3 範例
- Amazon DynamoDB 範例
- AWS 資料庫遷移範例
- 資料庫現代化研討會
- 使用 Amazon RDS for Postgress 資料庫上的參數

PERF03-BP03 收集並記錄資料存放區效能指標

追蹤並記錄資料存放區的相關績效指標，以了解資料管理解決方案的成效。這些指標可協助您最佳化資料存放區、確認是否符合工作負載需求，並提供工作負載執行方式的清晰概觀。

常見的反模式：

- 您只使用手動日誌檔案來搜尋指標。
- 您只會將指標發布到團隊使用的內部工具，而不會全面了解您的工作負載。
- 您只會使用所選監控軟體記錄的預設指標。
- 您只會在有問題時審查指標。

- 您只監控系統層級指標，而沒有擷取資料存取或用量指標。

建立此最佳實務的優勢：建立效能基準可協助您了解工作負載的正常行為和需求。異常模式可以更快地識別和偵錯，進而改善資料存放區的效能和可靠性。

未建立此最佳實務時的曝險等級：高

實作指引

要監控資料存放區的效能，您必須記錄一段時間的多個效能指標。這可讓您偵測異常情況，並根據業務指標衡量效能，以確認您是否滿足工作負載需求。

指標應包括支援資料存放區的基礎系統和資料庫指標。基礎系統指標可能包括CPU使用率、記憶體、可用磁碟儲存體、磁碟 I/O、快取命中率，以及網路傳入和傳出指標，而資料存放區指標可能包括每秒交易數、熱門查詢、平均查詢率、回應時間、索引用量、資料表鎖定、查詢逾時和開啟的連線數量。此資料對於了解工作負載的執行情況以及資料管理解決方案的使用方式至關重要。將這些指標納入資料驅動的方法，以調整和優化工作負載的資源。

使用工具、程式庫和系統來記錄與資料庫效能有關的效能測量值。

實作步驟

- 找出要追蹤的資料存放區關鍵效能指標。
  - Amazon S3 指標和維度
  - 在 Amazon RDS執行個體中監控 的指標
  - 使用 Amazon 上的 Performance Insights 監控資料庫負載 RDS
  - 增強型監視概觀
  - DynamoDB 指標和維度
  - 監控 DynamoDB Accelerator
  - 使用 Amazon 監控 Amazon MemoryDB CloudWatch
  - 應監控哪些指標？
  - 監控 Amazon Redshift 叢集效能
  - Timestream 指標和維度
  - Amazon Aurora 的 Amazon CloudWatch 指標
  - Amazon Keyspaces (適用於 Apache Cassandra) 中的日誌記錄和監控
  - 監控 Amazon Neptune 資源

- 使用核准的日誌記錄和監控解決方案來收集這些指標。Amazon CloudWatch 可以跨架構中的資源收集指標。您還可以收集和發布自訂指標以顯示業務或衍生指標。使用 CloudWatch 或第三方解決方案來設定警示，指出何時違反閾值。
- 檢查資料存放區監控是否能從可偵測效能異常的機器學習解決方案中獲益。
  - Amazon DevOpsGuru for Amazon RDS提供效能問題的可見性，並提出修正動作的建議。
- 在監控和日誌記錄解決方案中設定資料保留，以符合安全性和營運目標。
  - CloudWatch 指標的預設資料保留
  - CloudWatch 日誌的預設資料保留

資源

相關文件：

- AWS 資料庫快取
- Amazon Athena 10 大效能秘訣
- Amazon Aurora 最佳實務
- DynamoDB Accelerator
- Amazon DynamoDB 最佳實務
- Amazon Redshift Spectrum 最佳實務
- Amazon Redshift 效能
- 使用 的雲端資料庫 AWS
- Amazon RDS Performance Insights

相關影片：

- AWS re：Invent 2022 - Amazon RDS和 Aurora 的效能監控，採用 Autodesk
- 使用 Amazon DevOpsGuru for Amazon 進行資料庫效能監控和調整 RDS
- AWS re：Invent 2023 - AWS 檔案儲存的新功能
- AWS re：Invent 2023 - 深入探索 Amazon DynamoDB
- AWS re：Invent 2023 - 在 Amazon S3 上建置和最佳化資料湖
- AWS re：Invent 2023 - AWS 檔案儲存的新功能
- AWS re：Invent 2023 - 深入探索 Amazon DynamoDB
- 在 Amazon 上監控 Redis 工作負載的最佳實務 ElastiCache

相關範例：

- AWS 資料集擷取指標收集架構
- Amazon RDS監控研討會
- AWS 目的建構資料庫研討會

PERF03-BP04 實作策略以提高資料存放區中的查詢效能

實作策略以最佳化資料並改善資料查詢，以便為工作負載提供更高的可擴展性和更高效的效能。

常見的反模式：

- 您沒有分割資料存放區中的資料。
- 您在資料存放區中僅以一種檔案格式儲存資料。
- 您沒有在資料存放區中使用索引。

建立此最佳實務的優勢：最佳化資料和查詢效能可提高效率、降低成本並改善使用者體驗。

未建立此最佳實務時的曝險等級：中

實作指引

資料最佳化和查詢調整是資料存放區中效能效率的關鍵層面，因為其會影響整個雲端工作負載的效能和回應能力。未經過最佳化的查詢可能會使用更多資源和造成更大的瓶頸，進而降低資料存放區的整體效率。

資料最佳化包括數個技術，以確保高效的資料儲存和存取。這也有助於提高資料存放區中的查詢效能。關鍵策略包括資料分割、資料壓縮和資料去常規化，這些都有助最佳化資料的儲存和存取。

實作步驟

- 了解和分析在資料存放區中執行的重要資料查詢。
- 找出資料存放區中執行速度緩慢的查詢，並使用查詢計畫了解其目前狀態。
  - 分析 Amazon Redshift 中的查詢計畫
  - 在 Athena 中使用 EXPLAIN 和 EXPLAIN ANALYZE
- 實作策略以改善查詢效能。有些關鍵策略包括下列情況：
  - 使用單欄式檔案格式 (例如 Parquet 或 ORC)。

- 壓縮資料存放區中的資料以減少儲存空間和 I/O 作業。
- 資料分割可將資料拆分為較小的部分並縮短資料掃描時間。
  - 在 Athena 中分割資料
  - 分割區與資料分配
- 在查詢中對共同欄進行資料索引編制。
- 針對頻繁查詢使用具體化視觀表。
  - 了解具體化視觀表
  - 在 Amazon Redshift 中建立具體化視觀表
- 選擇正確的聯結作業以進行查詢。當您聯結兩個資料表時，請在聯結左側指定較大的資料表，並在聯結右側指定較小的資料表。
- 分散式快取解決方案可改善延遲並減少資料庫 I/O 操作的次數。
- 定期維護，例如清空、重新索引以及執行統計。

- 在非生產環境中實驗和測試策略。

資源

相關文件：

- Amazon Aurora 最佳實務
- Amazon Redshift 效能
- Amazon Athena 10 大效能秘訣
- AWS 資料庫快取
- 實作 Amazon ElastiCache 的最佳實務
- 在 Athena 中分割資料

相關影片：

- AWS re:Invent 2023 - AWS 儲存成本最佳化最佳實務
- AWS re:Invent 2022 - 使用 Amazon RDS 和 Aurora 進行效能監控，採用 Autodesk
- 使用新的查詢分析工具最佳化 Amazon Athena 查詢

相關範例：

- AWS 專用資料庫研討會

PERF03-BP05 實作使用快取的資料存取模式

實作可受益於快取資料的存取模式，以便快速擷取經常存取的資料。

常見的反模式：

- 快取頻繁變更的資料。
- 您依賴快取資料，就好像它是持久存儲並始終可用一樣。
- 您不考慮快取資料的一致性。
- 您不監控快取實作的效率。

建立此最佳實務的優勢：將資料儲存在快取中可改善讀取延遲、讀取輸送量、使用者體驗和整體效率，並降低成本。

未建立此最佳實務時的風險暴露等級：中

實作指引

快取是旨在存儲資料的軟體或硬體組件，以便更快或更有效地滿足未來對相同資料的請求。如果存儲在快取中的資料丟失，可以透過重複之前的計算或從另一個資料存放區中擷取來進行重建。

資料快取可能是改善整體應用程式效能並減輕基礎主要資料來源負擔的最有效策略之一。可以在應用程式的多個層級快取資料，例如在進行遠端呼叫的應用程式內 (稱為用戶端快取)，或使用快速次要服務來儲存資料 (稱為遠端快取)。

用戶端快取

透過用戶端快取，每個用戶端 (查詢後端資料儲存的應用程式或服務) 都可以在指定的時間內，在本機儲存其唯一查詢的結果。這可以先檢查本機用戶端快取，來減少網路對資料儲存的請求數量。如果結果不存在，則應用程式便可查詢資料儲存，並將這些結果儲存在本機。此模式允許每個用戶端將資料儲存在最接近的位置 (用戶端本身)，從而達到最低的延遲。當後端資料儲存無法使用時，用戶端也可以繼續提供某些查詢，從而提高整體系統的可用性。

這種方法的一個缺點是，當涉及多個用戶端時，它們可能會在本地存儲相同的快取資料。這會導致這些用戶端之間的重複儲存使用量和資料不一致。一個用戶端可能會快取查詢結果，一分鐘後，另一個用戶端可以執行相同查詢並獲得不同結果。

遠端快取

為了解決用戶端之間的重複資料問題，可以使用快速外部服務或遠端緩存來存儲查詢的資料。每個用戶端都會在查詢後端資料儲存之前檢查遠端快取，而非檢查本機資料存放區。此策略可實現用戶端之間更一致的回應、更好的儲存資料效率以及更高的快取資料量，因為儲存空間會獨立於用戶端進行擴展。

遠端快取的缺點是整個系統可能會遇到較高延遲，因為需要額外的網路跳轉來檢查遠端快取。用戶端快取可以與遠端快取一起用於多層級快取，以改善延遲。

實作步驟

- 識別可能受益於快取的資料庫APIs和網路服務。具有繁重讀取工作負載、比率高 read-to-write或規模昂貴的服務是快取的候選者。
  - 資料庫快取
  - 啟用API快取以增強回應能力
- 找出最適合您的存取模式的適當快取策略類型。
  - 快取策略
  - AWS 快取解決方案
- 遵循資料存放區的快取最佳實務。
- 針對平衡資料新鮮度和降低後端資料存放區壓力的所有資料，設定快取失效策略，例如 time-to-live（TTL）。
- 在用戶端中啟用自動連線重試、指數退避、用戶端逾時和連線集區等功能 (如果可用)，因為它們可以改善效能和可靠性。
  - 最佳實務：Redis 用戶端和 Amazon ElastiCache（Redis OSS）
- 監控快取命中率，目標為 80% 或更高。較低的值可能表示快取大小不足，或者無法從快取中受益的存取模式。
  - 應監控哪些指標？
  - 在 Amazon 上監控 Redis 工作負載的最佳實務 ElastiCache
  - 使用 Amazon 監控 Amazon ElastiCache（RedisOSS）的最佳實務 CloudWatch
- 實作資料複寫，將讀取卸載至多個執行個體，並提高資料讀取效能和可用性。

資源

相關文件：

- 使用 Amazon ElastiCache Well-Architected Lens
- 使用 Amazon 監控 Amazon ElastiCache（RedisOSS）的最佳實務 CloudWatch
- 應監控哪些指標？
- Amazon ElastiCache 白皮書的大規模效能
- 快取挑戰和策略

相關影片：

- Amazon ElastiCache Learning Path
- Amazon ElastiCache 最佳實務的成功設計
- AWS re：Invent 2020 - Amazon ElastiCache 最佳實務的成功設計
- AWS re：Invent 2023 -【LAUNCH】介紹 Amazon ElastiCache Serverless
- AWS re：Invent 2022 – 使用 Redis 重新構想資料層的 5 種絕佳方法
- AWS re：Invent 2021 - 在 Amazon 上深入探索 ElastiCache（Redis OSS）

相關範例：

- 使用 Amazon 提升我的SQL資料庫效能 ElastiCache（Redis OSS）

聯網與內容交付

工作負載的最佳聯網解決方案會根據延遲、輸送量需求、抖動和頻寬而有所不同。實體限制 (例如使用者或內部部署資源) 會決定位置選項。這些限制可能隨著邊緣節點或資源位置而有所差異。

在 AWS 上，聯網以虛擬化方式存在，並提供多種不同的類型和組態。如此就能更容易滿足您的聯網需求。AWS 提供了多種產品功能 (例如，增強型聯網、經 Amazon EC2 聯網最佳化的執行個體、Amazon S3 Transfer Acceleration 和動態 Amazon CloudFront)，可最佳化網路流量。AWS 還提供了聯網功能 (例如，Amazon Route 53 延遲路由、Amazon VPC 端點、AWS Direct Connect 和 AWS Global Accelerator)，可減少網路距離或抖動。

這個重點領域分享了在雲端中設計、設定和操作高效聯網和內容交付解決方案的各種指引和最佳實務。

最佳實務

- PERF04-BP01 了解聯網如何影響效能
- PERF04-BP02 評估可用的聯網功能
- PERF04-BP03 VPN為您的工作負載選擇適當的專用連線能力
- PERF04-BP04 使用負載平衡將流量分散到多個資源
- PERF04-BP05 選擇網路通訊協定以提高效能
- PERF04-BP06 根據網路需求選擇工作負載的位置
- PERF04-BP07 根據指標最佳化網路組態

PERF04-BP01 了解聯網如何影響效能

分析並了解網路相關決策如何影響您的工作負載，以提供高效的效能並改善使用者體驗。

常見的反模式：

- 所有流量都流經現有資料中心。
- 可以透過中央防火牆路由所有流量，而非使用雲端原生網路安全工具。
- 您佈建 AWS Direct Connect 連線時不了解實際用量需求。
- 定義聯網解決方案時，不會考慮工作負載特性和加密開銷。
- 對於雲端中的聯網解決方案，您可以使用內部部署概念和策略。

建立此最佳實務的優勢：了解聯網如何影響工作負載效能，有助於您識別潛在瓶頸、改善使用者體驗、提高可靠性並在工作負載變更時減少操作維護。

未建立此最佳實務時的風險暴露等級：高

實作指引

網路負責處理應用程式元件、雲端服務、邊緣網路和內部部署資料之間的連線，因此對工作負載效能可能有嚴重影響。除了工作負載效能外，使用者體驗也會受到網路延遲、頻寬、通訊協定、位置、網路擁塞、抖動、輸送量和路由規則的影響。

取得工作負載的已記錄在案的聯網需求清單，包括延遲、封包大小、路由規則、通訊協定和支援的流量模式。檢閱可用的聯網解決方案，確定哪些服務符合您的工作負載網路特性。雲端型網路可以快速重建，因此隨著時間演進您的網路架構是提高效能達成效率的必要條件。

實作步驟：

- 定義並記錄聯網效能需求，包括網路延遲、頻寬、通訊協定、位置、流量模式 (尖峰和頻率)、輸送量、加密、檢查和路由規則等指標。
- 了解金鑰 AWS 聯網服務，例如 VPCs、AWS Direct Connect、Elastic Load Balancing（ELB）和 Amazon Route 53。
- 擷取下列主要聯網特性：

特性	工具與指標
基礎聯網特性	• VPC 流程日誌 • AWS Transit Gateway Flow Logs • AWS Transit Gateway 指標 • AWS PrivateLink 指標
應用程式聯網特性	• Elastic Fabric Adapter • AWS App Mesh 指標 • Amazon API Gateway 指標
邊緣聯網特性	• Amazon CloudFront 指標 • Amazon Route 53 指標 • AWS Global Accelerator 指標

特性	工具與指標
混合聯網特性	• AWS Direct Connect 指標 • AWS Site-to-Site VPN 指標 • AWS Client VPN 指標 • AWS 雲端 WAN 指標
安全聯網特性	• AWS ShieldAWS WAF、 和 AWS Network Firewall 指標
追蹤特性	• AWS X-Ray • VPC 可連線性分析工具 • 網路存取分析器 • Amazon Inspector • Amazon CloudWatch RUM

- 基準測試並測試網路效能：
  - 基準網路輸送量，因為當執行個體位於相同的 時，某些因素可能會影響 Amazon EC2 網路效能VPC。測量相同 中 Amazon EC2 Linux 執行個體之間的網路頻寬VPC。
  - 執行負載測試，試驗聯網解決方案和選項。

資源

相關文件：

- Application Load Balancer
- EC2 Linux 上的增強型網路
- EC2 Windows 上的增強型網路
- EC2 置放群組
- 在 Linux 執行個體上使用 Elastic Network Adapter （ ENA ） 啟用增強型網路
- Network Load Balancer
- 使用 網路產品 AWS
- 轉換閘道
- 轉換到 Amazon Route 53 中的以延遲為基礎的路由

- VPC 端點

相關影片：

- AWS re：Invent 2023 - AWS networking 基礎
- AWS re：Invent 2023 - 聯網可以為您的應用程式做什麼？
- AWS re：Invent 2023 - 進階VPC設計和新功能
- AWS re：Invent 2023 - 開發人員的雲端聯網指南
- AWS re：Invent 2019 - 與 AWS 混合 AWS 網路架構的連線能力
- AWS re：Invent 2019 - 最佳化 Amazon EC2執行個體的網路效能
- AWS Summit Online - 改善應用程式的全球網路效能
- AWS re：Invent 2020 - 使用 Well-Architected Framework 建立網路最佳實務和秘訣
- AWS re：Invent 2020 - AWS networking 大規模遷移的最佳實務

相關範例：

- AWS Transit Gateway 和可擴展安全解決方案
- AWS 網路研討會
- 網路防火牆實際操作研討會
- 在 上觀察和診斷您的網路 AWS
- 在 上尋找和解決網路設定錯誤 AWS

PERF04-BP02 評估可用的聯網功能

評估雲端中可提升效能的聯網功能。透過測試、指標和分析來測量這些功能的影響。例如，利用可用的網路層級功能來降低延遲、網路距離或抖動。

常見的反模式：

- 您只在單一區域中活動，這是因為該區域是您總部的所在區域。
- 可以使用防火牆而非安全群組來篩選流量。
- 您TLS中斷流量檢查，而不是依賴安全群組、端點政策和其他雲端原生功能。
- 您只會使用子網路來分隔，而非採用安全群組的方式。

建立此最佳實務的優勢：評估所有服務功能和選項可提高工作負載效能、降低基礎架構成本、減少維護工作負載所需的人力，以及提升整體安全狀態。您可以使用 全域 AWS 骨幹為客戶提供最佳的聯網體驗。

未建立此最佳實務時的風險暴露等級：高

實作指引

AWS 提供 AWS Global Accelerator和 Amazon CloudFront 等服務，有助於改善網路效能，而大多數 AWS 服務具有產品功能（例如 Amazon S3 Transfer Acceleration 功能），可最佳化網路流量。

檢閱您可以使用哪些網路相關組態選項，及其對工作負載可能有何影響。效能最佳化取決於了解這些選項如何與您的架構互動，以及它們對衡量的效能與使用者體驗的影響。

實作步驟

- 建立工作負載元件清單。
  - 在建置統一的全球網路時，請考慮使用 AWS 雲端 WAN 來建置、管理和監控組織的網路。
  - 使用 Amazon CloudWatch Logs 指標 監控您的全域和核心網路。利用 Amazon CloudWatch RUM提供洞見，以協助識別、了解和增強使用者的數位體驗。
  - 檢視 AWS 區域 和可用區域之間以及每個可用區域內的彙總網路延遲，使用 AWS Network Manager來深入了解應用程式效能與基礎 AWS 網路效能之間的關係。
  - 使用現有的組態管理資料庫（CMDB）工具或服務，例如 AWS Config建立工作負載的清查及其設定方式。
- 如果這是現有的工作負載，則請識別並記錄效能指標的基準，並著重於瓶頸和要改善的領域。效能相關聯網指標會依據業務需求和工作負載特性而有所不同。首先，這些指標對於檢閱工作負載可能很重要：頻寬、延遲、封包遺失、抖動和重新傳輸。
- 如果這是新的工作負載，則請執行負載測試以識別效能瓶頸。
- 對於您找出的效能瓶頸，請檢閱您解決方案的組態選項，以找出改善效能的機會。查看下列主要聯網選項和功能：

改進機會	解決方案
網路的路徑或路由	使用網路存取分析器來識別路徑或路由。
網路通訊協定	請參閱PERF04-BP05 選擇網路通訊協定以提高效能

改進機會	解決方案
網路拓撲	在連接多個帳戶AWS Transit Gateway時，評估VPC對等和 之間的操作和效能權衡。AWS Transit Gateway 可簡化如何將所有互連VPCs，其可以橫跨數千個 AWS 帳戶和 內部部署網路。使用 在多個帳戶 AWS Transit Gateway 之間共用您的 AWS Resource Access Manager。 請參閱PERF04-BP03 VPN為您的工作負載選擇適當的專用連線能力
網路服務	AWS Global Accelerator 是一種聯網服務，使用 AWS 全球網路基礎設施將使用者流量的效能提升高達 60%。 Amazon CloudFront 可以改善全球工作負載內容交付和延遲的效能。 使用 Lambda@edge 執行函數，以自訂 CloudFront 提供更接近使用者的內容、減少延遲並改善效能。 Amazon Route 53 提供以延遲為基礎的路由、地理位置路由、地理位置鄰近性路由和以 IP 為基礎的路由選項，可協助您為全球使用者改善工作負載的效能。當工作負載分佈在全球範圍時，請檢閱工作負載流量和使用者位置，找出能夠最佳化工作負載效能的路由選項。

改進機會	解決方案
儲存功能資源	Amazon S3 Transfer Acceleration 是一項功能，可讓外部使用者受益於 的網路最佳化 CloudFront ，將資料上傳至 Amazon S3。這樣就可以更輕易地從與 AWS 雲端沒有專用連線的遠端位置輸送大量資料。 Amazon S3 多區域存取點可將內容複製到多個區域，並透過提供一個存取點來簡化工作負載。使用多區域存取點時，您可以使用可識別最低延遲儲存貯體的服務，來要求資料或將資料寫入 Amazon S3。
運算資源功能	Amazon EC2執行個體、容器和 Lambda 函數使用的彈性網路介面（ENA）會依每個流程限制。檢閱置放群組，以最佳化EC2您的網路輸送量。若要避免個別流程的瓶頸，請將應用程式設計為使用多個流程。若要監控並取得運算相關網路指標的可見性，請使用 CloudWatch 指標 和 ethtool。`ethtool` 命令包含在ENA驅動程式中，並將其他可發佈為自訂指標的網路相關指標公開給 CloudWatch。 Amazon Elastic Network Adapters（ENA）透過為叢集置放群組 內的執行個體提供更好的輸送量，進一步最佳化。 Elastic Fabric Adapter（EFA）是 Amazon EC2執行個體的網路介面，可讓您在 上執行需要大規模內部通訊的工作負載 AWS。 Amazon EBS最佳化執行個體使用最佳化的組態堆疊，並提供額外的專用容量來增加 Amazon EBS I/O。

資源

相關文件：

- Application Load Balancer
- EC2 Linux 上的增強型網路
- EC2 Windows 上的增強型網路
- EC2 置放群組
- 在 Linux 執行個體上使用 Elastic Network Adapter（ENA）啟用增強型網路
- Network Load Balancer
- 使用 網路產品 AWS
- 轉換到 Amazon Route 53 中的以延遲為基礎的路由
- VPC 端點
- VPC 流量日誌

相關影片：

- AWS re：Invent 2023 – 為下一步做好準備？ 設計網路實現增長和靈活性
- AWS re：Invent 2023 – 進階VPC設計和新功能
- AWS re：Invent 2023 – 雲端聯網開發人員指南
- AWS re：Invent 2022 – 深入探索 AWS 網路基礎設施
- AWS re：Invent 2019 – 與 AWS 混合 AWS 網路架構的連線能力
- AWS re：Invent 2018 – 最佳化 Amazon EC2執行個體的網路效能
- AWS Global Accelerator

相關範例：

- AWS Transit Gateway 和可擴展安全解決方案
- AWS 網路研討會
- 觀察並診斷您的網路
- 在 上尋找和解決網路設定錯誤 AWS

PERF04-BP03 VPN為您的工作負載選擇適當的專用連線能力

需要混合式連線來連接內部部署資源和雲端資源時，請佈建足夠的頻寬以滿足您的效能要求。預估混合工作負載的頻寬和延遲需求。這些數字將促進調整大小需求。

常見的反模式：

- 您僅針對網路加密需求評估VPN解決方案。
- 不會評估備份或備援連線選項。
- 您無法識別所有工作負載需求 (加密、通訊協定、頻寬和流量需求)。

建立此最佳實務的優勢：選擇和設定適當的連線解決方案將提高工作負載的可靠性並將效能最大化。透過識別工作負載需求、提前規劃和評估混合解決方案，您可以最大限度地減少昂貴的實體網路變更和操作額外負荷，同時增加您的 time-to-value。

未建立此最佳實務時的曝險等級：高

實作指引

根據頻寬需求開發混合式聯網架構。AWS Direct Connect 允許您私下將內部部署網路與 AWS連線。需要高頻寬、低延遲同時可達到一致效能時，適合這個選項。VPN 連線會透過網際網路建立安全連線。在以下情況使用它：當只需要臨時連線時、當成本是一個考慮因素時、或者在使用 AWS Direct Connect時等待建立彈性物理網路連線作為應急措施時。

如果您的頻寬需求很高，您可以考慮多個 AWS Direct Connect VPN或服務。流量可以在服務之間進行負載平衡，但VPN由於延遲和頻寬差異，我們不建議在 AWS Direct Connect 和 之間進行負載平衡。

實作步驟

- 預估現有應用程式的頻寬和延遲要求。
  - 對於要移至 的現有工作負載 AWS，請利用內部網路監控系統的資料。
  - 針對您沒有監控資料的新工作負載或現有工作負載，請諮詢產品擁有者以確定足夠的效能指標，並且提供良好的使用者體驗。
- 選取專用連線或 VPN 做為連線選項。根據所有工作負載需求（加密、頻寬和流量需求），您可以選擇 AWS Direct Connect 或 AWS VPN（或兩者）。下圖可協助您選擇適當的連線類型。
  - AWS Direct Connect 使用專用連線或託管連線，提供 AWS 環境的專用連線，範圍從 50 Mbps 到 100 Gbps。這可為您提供受管和受控的延遲以及佈建頻寬，因此您的工作負載可以有效率地連線

到其他環境。使用 AWS Direct Connect 合作夥伴，您可以從多個環境進行 end-to-end連線，提供具有一致效能的延伸網路。 AWS 提供使用原生 100 Gbps、連結彙總群組（LAG）或BGP同等成本多路徑（ECMP）擴展直接連線連線頻寬。
  - 提供支援網際網路通訊協定安全的受管VPN服務 AWS Site-to-Site VPN（IPsec）。建立VPN連線時，每個VPN連線都包含兩個通道，以實現高可用性。
- 請依照 AWS 文件選擇適當的連線選項：
  - 如果您決定使用 AWS Direct Connect，請為您的連線選擇適當的頻寬。
  - 如果您使用 AWS Site-to-Site VPN 跨多個位置的 來連線至 AWS 區域，請使用加速 Site-to-SiteVPN連線來改善網路效能。
  - 如果您的網路設計包含透過 的IPSecVPN連線AWS Direct Connect，請考慮使用私有 IP VPN 來提高安全性並實現分割。AWS Site-to-Site 私有 IP VPN會部署在傳輸虛擬介面（VIF）之上。
  - AWS Direct Connect SiteLink允許在全球各地的資料中心之間建立低延遲和備援連線，方法是透過AWS Direct Connect 位置 之間的最快路徑傳送資料，繞過 AWS 區域。
- 在部署到生產環境之前驗證連線設定。執行安全性和效能測試，以確保其符合您的頻寬、可靠性、延遲和合規性要求。
- 定期監控您的連線效能和使用情況，並視需要進行最佳化。

確定性效能流程圖

資源

相關文件：

- 使用 網路產品 AWS
- AWS Transit Gateway
- VPC 端點
- 建置可擴展且安全的多VPC AWS 網路基礎設施
- 用戶端 VPN

相關影片：

- AWS re：Invent 2023 – 使用 建置混合網路連線 AWS

- AWS re：Invent 2023 – 安全遠端連線至 AWS
- AWS re：Invent 2022 – 使用 Amazon 最佳化效能 CloudFront
- AWS re：Invent 2019 – 與 AWS 混合 AWS 網路架構的連線能力
- AWS re：Invent 2020 – AWS Transit Gateway Connect

相關範例：

- AWS Transit Gateway 和可擴展安全解決方案
- AWS 網路研討會

PERF04-BP04 使用負載平衡將流量分散到多個資源

在多個資源或服務之間分配流量，以讓您的工作負載能夠利用雲端提供的彈性。您也可以使用負載平衡來卸載加密終止，以提升效能、可靠性，以及有效管理和路由流量。

常見的反模式：

- 您在選擇負載平衡器類型時未考慮工作負載需求。
- 您不利用負載平衡器功能來進行效能最佳化。
- 工作負載在不使用負載平衡器的情況下，直接公開到網際網路。
- 您可以透過現有的負載平衡器路由所有網際網路流量。
- 您可以使用一般TCP負載平衡，並讓每個運算節點處理SSL加密。

建立此最佳實務的優勢：負載平衡器會處理單一可用區域中或跨多個可用區域的應用程式流量不同的負載，並實現高可用性、自動擴展及更充分利用您的工作負載。

未建立此最佳實務時的曝險等級：高

實作指引

負載平衡器會做為您的工作負載的進入點，從那裡將您的流量分散到後端目標，例如運算執行個體或容器，以提高利用率。

選擇正確的負載平衡器類型是最佳化架構的第一步。首先列出您的工作負載特性，例如通訊協定（例如 TCP、HTTP、 或 WebSockets）TLS、目標類型（例如執行個體、容器或無伺服器）、應

用程式需求（例如長時間執行的連線、使用者身分驗證或黏性），以及置放（例如區域、本機區域、Outpost 或區域隔離）。

AWS 提供多個模型，讓您的應用程式使用負載平衡。Application Load Balancer 最適合 HTTP和 HTTPS 流量的負載平衡，並提供進階請求路由，以交付現代應用程式架構為目標，包括微服務和容器。

Network Load Balancer 最適合需要極端效能的TCP流量負載平衡。它能夠每秒處理數百萬個請求，同時保持超低延遲性，並且還進行優化，可處理突發的和不穩定的流量模式。

Elastic Load Balancing 提供整合的憑證管理和SSL/TLS解密，可讓您靈活地集中管理負載平衡器SSL的設定，並從工作負載卸載CPU密集型工作。

選擇正確的負載平衡器之後，您可以開始利用其功能來減少後端為流量提供服務所需投入的工作量。

例如，同時使用 Application Load Balancer（ALB）和 Network Load Balancer（NLB），您可以執行 SSL/TLS 加密卸載，這是避免目標完成 CPU密集型TLS交握，以及改善憑證管理的機會。

當您在負載平衡器中設定SSL/TLS卸載時，它會負責加密來自用戶端的流量，同時將未加密的流量交付至後端，釋放後端資源並改善用戶端的回應時間。

Application Load Balancer 也可以提供 HTTP/2 流量，而不需要在目標上支援。這個簡單的決策可以改善您的應用程式回應時間，因為 HTTP/2 更有效率地使用TCP連線。

定義架構時，應該考慮您的工作負載延遲要求。例如，如果您有對延遲敏感的應用程式，您可能會決定使用 Network Load Balancer，以獲得極低的延遲。另外，您可能會決定讓工作負載更靠近您的客戶，也就是利用 AWS Local Zones 或甚至 AWS Outposts 中的 Application Load Balancer。

對延遲敏感的工作負載的另一個考慮是跨區域負載平衡。使用跨區域負載平衡，每個負載平衡器節點會將已註冊目標之間的流量分散到所有允許的可用區域中。

使用與您的負載平衡器整合的 Auto Scaling。效能效率系統的其中一個關鍵層面與適當調整後端資源大小有關。若要完成此操作，您可以利用後端目標資源的負載平衡器整合。使用與 Auto Scaling 群組整合的負載平衡器，目標會視需要從負載平衡器新增或移除，以因應傳入流量。負載平衡器也可以與 Amazon ECS 和 Amazon EKS 整合，用於容器化工作負載。

- Amazon ECS - 服務負載平衡
- Amazon 上的應用程式負載平衡 EKS
- Amazon 上的網路負載平衡 EKS

實作步驟

- 定義您的負載平衡需求，包括流量、可用性和應用程式可擴展性。
- 為您的應用程式選擇正確的負載平衡器類型。
  - 針對 HTTP/HTTPS 工作負載使用 Application Load Balancer。
  - 針對在 TCP或 上執行的非HTTP工作負載使用 Network Load BalancerUDP。
  - 如果您想要利用兩個 產品的功能，請使用兩者的組合ALB （ 作為 的目標NLB）。例如，如果您想要將 靜態 與來自 IPs的HTTP標頭型路由NLB搭配使用ALB，或者如果您想要將HTTP工作負載公開至 ，則可以這樣做AWS PrivateLink。
  - 如需負載平衡器的完整比較，請參閱ELB產品比較 。
- 如果可能，請使用 SSL/TLS 卸載。
  - 使用與 整合Application Load Balancer和網路Load Balancer來設定 HTTPS/TLS 接聽程式AWS Certificate Manager。
  - 請注意，某些工作負載可能因為合規原因而需要 end-to-end加密。在此情況下，必須允許在目標進行加密。
  - 如需安全最佳實務，請參閱SEC傳輸中的 09-BP02 強制執行加密。
- 選取正確的路由演算法 （ 僅限 ALB ）。
  - 路由演算法可以造成您的後端目標的妥善使用程度和它們影響效能程度的差異。例如， ALB 提供兩種路由演算法的選項：
  - 最低未解決請求：針對應用程式的請求因複雜性而異，或目標因處理功能而異的情況，用來讓負載更妥善地分散到您的後端目標。
  - 循環配置：當請求和目標類似，或是如果您需要在目標之間平均分散請求時使用。
- 考慮跨區域或區域隔離。
  - 針對延遲改善和區域失敗網域使用跨區域關閉 (區域隔離)。依預設，它會在 中關閉NLB，而在中ALB，您可以關閉每個目標群組 。
  - 使用跨區域開啟來增加可用性和彈性。根據預設， 的跨區域會開啟，ALB您可以在 中NLB為每個目標群組 開啟。
- 開啟HTTP工作負載的 HTTP keep-alives （ 僅限 ALB ）。透過此功能，負載平衡器可以重複使用後端連線，直到保持連線逾時到期為止，藉此改善您的HTTP請求和回應時間，並減少後端目標的資源使用率。如需如何為 Apache 和 Nginx 執行此操作的詳細資訊，請參閱使用 Apache 或NGINX作為後端伺服器的最佳設定為何ELB？
- 開啟負載平衡器的監控功能。
  - 開啟 Application Load Balancer 和 Network Load Balancer 的存取記錄。

- 要考慮的主要欄位ALB為 `request_processing_timerequest_processing_time`、 和 `response_processing_time`。
- 要考慮的主要欄位NLB為 `connection_time`和 `tls_handshake_time`。
- 請準備好在您需要日誌時進行查詢。您可以使用 Amazon Athena 查詢ALB日誌和NLB日誌 。
- 為效能相關指標建立警示，例如`TargetResponseTime`針對 ALB。

資源

相關文件：

- ELB 產品比較
- AWS 全球基礎設施
- 使用可用區域親和性改善效能並且降低成本
- 使用 Amazon Athena 逐步執行日誌分析
- 查詢 Application Load Balancer 日誌
- 監控 Application Load Balancer
- 監控 Network Load Balancer
- 使用 Elastic Load Balancing 在 Auto Scaling 群組的執行個體中分配流量

相關影片：

- AWS re：Invent 2023：聯網可以為您的應用程式做什麼？
- AWS re：Inforce 20：如何使用 Elastic Load Balancing 大規模增強您的安全狀態
- AWS re：Invent 2018：Elastic Load Balancing：Deep Dive 和最佳實務
- AWS re：Invent 2021 – 如何為您的 AWS 工作負載選擇正確的負載平衡器
- AWS re：Invent 2019：為不同工作負載充分利用 Elastic Load Balancing

相關範例：

- Gateway Load Balancer
- CDK 和 AWS CloudFormation 範例，用於使用 Amazon Athena 進行日誌分析

PERF04-BP05 選擇網路通訊協定以提高效能

根據對工作負載效能的影響，做出系統和網路間通訊協定的決策。

實現輸送量的延遲和頻寬之間存在關係。如果您的檔案傳輸使用傳輸控制通訊協定（TCP），較高的延遲很可能降低整體輸送量。有方法可以透過TCP調校和最佳化的傳輸通訊協定來修正此問題，但其中一個解決方案是使用使用者資料包通訊協定（UDP）。

常見的反模式：

- 無論效能需求為何，您都會TCP在所有工作負載中使用 。

建立此最佳實務的優勢：確認針對使用者與工作負載元件之間的通訊使用適當的通訊協定，可協助改善您的應用程式的整體使用者體驗。例如，無連線UDP允許高速，但不提供重新傳輸或高可靠性。TCP是全功能通訊協定，但需要更大的額外負荷來處理封包。

未建立此最佳實務時的曝險等級：中

實作指引

如果您有能力為應用程式選擇不同的通訊協定，而且您具備此領域的專業知識，請使用不同的通訊協定來最佳化應用程式和使用者體驗。請注意，這種方法有很大的困難，如果您已先用其他方法最佳化應用程式，才可嘗試。

改善您的工作負載效能的主要考慮是了解延遲和輸送量需求，然後選擇可最佳化效能的網路通訊協定。

何時考慮使用 TCP

TCP 提供可靠的資料交付，可用於工作負載元件之間的通訊，其中資料的可靠性和保證交付至關重要。許多 Web 型應用程式依賴 TCP型通訊協定，例如 HTTP和 HTTPS來開啟通訊TCP端，以便在應用程式元件之間進行通訊。電子郵件和檔案資料傳輸是也使用 的常見應用程式TCP，因為它是應用程式元件之間簡單且可靠的傳輸機制。TLS 搭配 使用 TCP可以為通訊增加一些額外負荷，這可能會導致延遲增加和輸送量減少，但它具有安全性的優勢。負擔主要來自交握處理的增加負擔，需要數個往返才能完成。一旦交握完成，加密和解密資料的負擔相對小。

何時考慮使用 UDP

UDP 是 connection-less-oriented通訊協定，因此適用於需要快速、高效傳輸的應用程式，例如日誌、監控和 VoIP 資料。此外，UDP如果您的工作負載元件可回應來自大量用戶端的小型查詢，請考

慮使用 ，以確保工作負載的最佳效能。Datagram Transport Layer Security （ DTLS ） UDP等同於 Transport Layer Security （ TLS ）。DTLS 搭配 使用 時UDP，額外負荷來自加密和解密資料，因為交握程序已簡化。DTLS 也會為UDP封包新增少量額外負荷，因為其中包含其他欄位，以指示安全參數並偵測竄改。

何時考慮使用 SRD

可擴展的可靠資料包 （ SRD ） 是一種針對高輸送量工作負載進行最佳化的網路傳輸通訊協定，因為它能夠跨多個路徑載入平衡器流量，並從封包捨棄或連結失敗中快速復原。SRD 因此，最適合需要運算節點之間高輸送量和低延遲通訊的高效能運算 （ HPC ） 工作負載。這可能包含平行處理任務，例如牽涉到在節點之間大量資料傳輸的模擬、建模和資料分析。

實作步驟

- 使用 AWS Global Accelerator 和 AWS Transfer Family 服務來改善線上檔案傳輸應用程式的輸送量。 AWS Global Accelerator 服務可協助您在用戶端裝置與 上的工作負載之間達到較低的延遲 AWS。透過 AWS Transfer Family，您可以使用 Secure Shell File Transfer Protocol （ SFTP ） 和 File Transfer Protocol over SSL（ FTPS ） 等 TCP型通訊協定，安全地擴展和管理檔案傳輸至 AWS 儲存服務。
- 使用網路延遲來判斷 是否適合工作負載元件之間的TCP通訊。如果用戶端應用程式和伺服器之間的網路延遲很高，則TCP三向交握可能需要一些時間，進而影響應用程式的回應能力。第一個位元組的時間 （ TTFB ） 和往返時間 （ RTT ） 等指標可用來測量網路延遲。如果您的工作負載為使用者提供動態內容，請考慮使用 Amazon CloudFront，這會為動態內容建立每個原始伺服器的持久連線，以移除可能拖慢每個用戶端請求的連線設定時間。
- 由於加密和解密的影響，TLS搭配 TCP或 使用 UDP 可能會導致工作負載的延遲增加和輸送量降低。對於此類工作負載，請考慮 SSL/TLS 卸載 onElastic Load Balancing，透過允許負載平衡器處理 SSL/TLS 加密和解密程序來改善工作負載效能，而不是讓後端執行個體執行。這有助於減少後端執行個體的使用CPU率，進而改善效能並增加容量。
- 使用Network Load Balancer （ NLB ） 部署依賴UDP通訊協定的服務，例如身分驗證和授權、記錄、DNS、IoT 和串流媒體，以提高工作負載的效能和可靠性。會將傳入UDP流量NLB分散到多個目標，讓您水平擴展工作負載、增加容量並減少單一目標的負荷。
- 對於高效能運算 （ HPC ） 工作負載，請考慮使用彈性網路轉接器 （ ENA ） Express 功能，該功能使用 SRD 通訊協定，透過為EC2執行個體之間的網路流量提供更高的單一流量頻寬 （ 25 Gbps ） 和更低的尾部延遲 （ 99.9 百分位數 ） 來改善網路效能。
- 使用Application Load Balancer （ ALB ） 在工作負載元件之間或 gRPC 用戶端和服務之間路由和負載平衡 gRPC （ 遠端程序呼叫 ） 流量。gRPC 使用 TCP型 HTTP/2 通訊協定進行傳輸，並提供效能優勢，例如更輕量的網路佔用、壓縮、高效的二進位序列化、對多種語言的支援，以及雙向串流。

資源

相關文件：

- 如何將UDP流量路由至 Kubernetes
- Application Load Balancer
- EC2 Linux 上的增強型網路
- EC2 Windows 上的增強型聯網
- EC2 置放群組
- 在 Linux 執行個體上使用彈性網路轉接器（ENA）啟用增強型網路
- Network Load Balancer
- 使用 網路產品 AWS
- 轉換到 Amazon Route 53 中的以延遲為基礎的路由
- VPC 端點

相關影片：

- AWS re：Invent 2022 – 在新一代 Amazon Elastic Compute Cloud 執行個體上擴展網路效能
- AWS re：Invent 2022 – 應用程式聯網基礎

相關範例：

- AWS Transit Gateway 和可擴展安全解決方案
- AWS 聯網研討會

PERF04-BP06 根據網路需求選擇工作負載的位置

評估資源置放的選項以減少網路延遲和提高輸送量，藉由減少頁面載入和資料傳輸時間來提供最佳的使用者體驗。

常見的反模式：

- 您可以將所有工作負載資源合併到單一地理位置。
- 您選擇的區域最接近您的位置，但不是最接近工作負載最終使用者。

建立此最佳實務的優勢：使用者體驗因使用者與您的應用程式之間的延遲而大受影響。透過使用適當的 AWS 區域 AWS 私有全域網路，您可以減少延遲，並為遠端使用者提供更好的體驗。

未建立此最佳實務時的風險暴露等級：中

實作指引

資源，例如 Amazon EC2執行個體，會放置在 AWS 區域、 AWS Local Zones 、 AWS Outposts或 AWS Wavelength 區域中的可用區域中。此位置的選擇會影響來自特定使用者位置的網路延遲和輸送量。Amazon CloudFront 和 等邊緣服務AWS Global Accelerator也可用於透過快取邊緣位置的內容，或透過 AWS 全球網路為使用者提供工作負載的最佳路徑來改善網路效能。

Amazon EC2提供置放群組以進行聯網。置放群組是執行個體的邏輯分組，用於減少延遲。搭配支援的執行個體類型和彈性網路轉接器（ENA）使用置放群組，可讓工作負載參與低延遲、減少抖動的 25 Gbps 網路。建議將置放群組用於受益於低網路延遲、高網路輸送量或兩者兼而有之的工作負載。

延遲敏感服務會使用 AWS 全球網路在邊緣位置提供，例如 Amazon CloudFront。這些邊緣位置通常提供內容交付網路（CDN）和網域名稱系統（）等服務DNS。透過在邊緣使用這些服務，工作負載可以對內容或DNS解決方案的請求做出低延遲的回應。這些服務還提供地理服務，例如內容的地理定位 (根據最終使用者的位置提供不同的內容)，或以延遲為基礎的路由，將最終使用者定向到最近區域的 (最小延遲)。

使用邊緣服務來減少延遲及啟用內容快取。正確設定 DNS和 HTTP/HTTPS 的快取控制，以從這些方法中獲益最多。

實作步驟

- 擷取與往返網路介面的 IP 流量有關的資訊。
  - 使用VPC流程日誌記錄 IP 流量
  - 如何保留用戶端 IP 地址 AWS Global Accelerator
- 分析您工作負載中的網路存取模式，以識別使用者如何使用您的應用程式。
  - 使用監控工具，例如 Amazon CloudWatch 和 AWS CloudTrail，收集網路活動的資料。
  - 分析資料以識別網路存取模式。
- 根據下列關鍵元素，為您的工作負載部署選取區域：
  - 資料所在位置：對於資料密集型應用程式 (例如大數據和機器學習)，應用程式碼執行時應盡可能接近資料。
  - 使用者所在位置：對於面向使用者的應用程式，請選擇接近工作負載使用者的一或多個區域。
  - 其他限制：考慮諸如成本和合規性之類的限制，如為工作負載選取區域時應考慮的事項中所述。

- 使用 AWS Local Zones 執行諸如影片轉譯等工作負載。Local Zones 可讓您因運算和儲存資源更接近最終使用者而獲益。
- AWS Outposts 適用於需要保持內部部署的工作負載，而您希望該工作負載能夠與 AWS中的其他工作負載無縫執行。
- 5G 裝置需要 ultra-low-latency高解析度即時影片串流、高擬真度音訊和擴增實境或虛擬實境（AR/VR）等應用程式。對於此類應用程式，請考慮在 5G 網路內AWS Wavelength AWS Wavelength 內嵌 AWS 運算和儲存服務，提供用於開發、部署和擴展 ultra-low-latency應用程式的行動邊緣運算基礎設施。
- 針對常用資產，使用本機快取或 AWS 快取解決方案以提升效能、減少資料移動以及降低環境影響。

服務	使用情況
Amazon CloudFront	使用 快取靜態內容，例如影像、指令碼和影片，以及動態內容，例如API回應或 Web 應用程式。
Amazon ElastiCache	用來快取 Web 應用程式的內容。
DynamoDB Accelerator	用來將記憶體內加速新增至 DynamoDB 資料表。

- 使用可協助您在更接近工作負載使用者的位置執行程式碼的服務，如下所示：

服務	使用情況
Lambda@Edge	用於在物件未經快取時起始的大量運算作業。
Amazon CloudFront Functions	用於可由短期函數啟動的簡單使用案例，例如 HTTP（s）請求或回應操作。
AWS IoT Greengrass	用來為連線的裝置執行本機運算、傳訊和資料快取。

- 某些應用程式需要藉由減少第一個位元組延遲和抖動並且增加輸送量，來獲得固定的進入點或較高的效能。這些應用程式可以受益於提供靜態廣播 IP 地址和在邊緣位置TCP終止的網路服務。 AWS Global Accelerator 可以為您的應用程式提升效能達 60%，並為多區域架構提供快速容錯移轉。AWS Global Accelerator 為您提供靜態廣播 IP 地址，作為託管於一或多個 的應用程式固定進入點 AWS 區域。這些 IP 地址允許流量盡可能接近您的使用者傳入 AWS 全域網路。透過在用戶端與最

接近用戶端的 AWS 邊緣位置之間建立TCP連線， AWS Global Accelerator 以減少初始連線設定時間。檢閱 的使用 AWS Global Accelerator ，以改善 TCP/UDP 工作負載的效能，並為多區域架構提供快速容錯移轉。

資源

相關的最佳實務：

- COST07-BP02 根據成本實作區域
- COST08-BP03 實作 服務以降低資料傳輸成本
- REL10-BP01 將工作負載部署到多個位置
- REL10-BP02 為您的多位置部署選取適當的位置
- SUS01-BP01 根據業務需求和永續性目標選擇區域
- SUS02-BP04 根據其聯網需求最佳化工作負載的地理定位
- SUS04-BP07 將網路之間的資料移動降至最低

相關文件：

- AWS 全球基礎設施
- AWS 本機區域 和 AWS Outposts，為您的邊緣工作負載選擇正確的技術
- 置放群組
- AWS 本機區域
- AWS Outposts
- AWS Wavelength
- Amazon CloudFront
- AWS Global Accelerator
- AWS Direct Connect
- AWS Site-to-Site VPN
- Amazon Route 53

相關影片：

- AWS Local Zones 解說器影片

- AWS Outposts：概觀和運作方式
- AWS re：Invent 2023 - 邊緣和內部部署工作負載的遷移策略
- AWS re：Invent 2021 - AWS Outposts：將 AWS 體驗帶到內部部署
- AWS re：Invent 2020： AWS Wavelength：在 5G 邊緣以超低延遲執行應用程式
- AWS re：Invent 2022 - AWS Local Zones：為分散式邊緣建置應用程式
- AWS re：Invent 2021 - 使用 Amazon 建置低延遲網站 CloudFront
- AWS re：Invent 2022 - 透過 改善效能和可用性 AWS Global Accelerator
- AWS re：Invent 2022 - 使用 建置您的全球廣域網路 AWS
- AWS re：Invent 2020：使用 Amazon Route 53 進行全域流量管理

相關範例：

- AWS Global Accelerator 自訂路由研討會
- 使用邊緣函數處理重新撰寫和重新導向

PERF04-BP07 根據指標最佳化網路組態

使用收集和分析的資料來做出有關優化網路組態的明智決策。

常見的反模式：

- 您假設所有效能相關問題都與應用程式有關。
- 您只能從靠近已部署工作負載的位置測試網路效能。
- 將預設組態用於所有網路服務。
- 您過度佈建網路資源來提供足夠的容量。

建立此最佳實務的優勢：收集必要的 AWS 網路指標並實作網路監控工具，可讓您了解網路效能並最佳化網路組態。

未建立此最佳實務時的曝險等級：低

實作指引

監控往返 VPCs、子網路或網路介面的流量，對於了解如何利用 AWS 網路資源和最佳化網路組態至關重要。使用下列 AWS 網路工具，您可以進一步檢查流量使用量、網路存取和日誌的相關資訊。

實作步驟

- 識別關鍵效能指標，例如要收集的延遲或封包遺失。 AWS 提供數種工具，可協助您收集這些指標。藉由使用下列工具，您可以進一步檢查流量用量、網路存取和日誌的相關資訊：

AWS 工具	在哪裡使用
Amazon VPC IP Address Manager 。	使用 IPAM 來規劃、追蹤和監控您 AWS 和內部部署工作負載的 IP 地址。這是最佳化 IP 位址用量和分配的最佳實務。
VPC 流程日誌	使用VPC流程日誌來擷取 中往返網路介面的流量詳細資訊VPCs。透過VPC流量日誌，您可以診斷過度限制或寬鬆的安全群組規則，並判斷往返網路介面的流量方向。
AWS Transit Gateway Flow Logs	使用 AWS Transit Gateway 流程日誌來擷取進出傳輸閘道的 IP 流量相關資訊。
DNS 查詢記錄	Route 53 收到的公有或私有DNS查詢的日誌資訊。使用 DNS 日誌，您可以透過了解請求的網域或子網域，或回應DNS查詢的 Route 53 EDGE位置來最佳化DNS組態。
Reachability Analyzer	Reachability Analyzer 可幫助您分析和偵錯網路可達性。Reachability Analyzer 是一種組態分析工具，可讓您在 中的來源資源和目的地資源之間執行連線測試VPCs。此工具可協助您確認您的組態符合您預期的連線能力。
網路存取分析器	網路存取分析器可協助您了解對資源的網路存取。您可以使用網路存取分析器來指定您的網路存取需求，並識別未符合您的指定需求的潛在網路路徑。藉由最佳化您的對應網路組態，您可以了解及確認網路的狀態，並且示範 AWS 上的網路是否符合您的合規要求。

AWS 工具	在哪裡使用
Amazon CloudWatch	使用 Amazon CloudWatch 並開啟網路選項的適當指標。請確定為您的工作負載選擇正確的網路指標。例如，您可以開啟 VPC Network Address Usage、VPCNATGateway AWS Transit Gateway、VPNTunle、 AWS Network Firewall Elastic Load Balancing 和 的指標 AWS Direct Connect。持續監控指標是觀察和了解您的網路狀態和用量的良好實務，並且可以協助您根據您的觀察來最佳化網路組態。
AWS Network Manager	使用 AWS Network Manager，您可以監控AWS 全球網路的即時和歷史效能，以用於操作和規劃。Network Manager 會在 AWS 區域 和 可用區域之間以及每個可用區域內提供彙總網路延遲，讓您更了解應用程式效能與基礎 AWS 網路效能之間的關係。
Amazon CloudWatch RUM	使用 Amazon CloudWatch RUM 收集指標，為您提供洞見，協助您識別、了解和改善使用者體驗。

- 使用 和 AWS Transit Gateway Flow Logs 識別熱門發言者VPC和應用程式流量模式。
- 評估和最佳化您目前的網路架構VPCs，包括 、子網路和路由。例如，您可以評估不同的VPC互連方式，或 AWS Transit Gateway 協助您改善架構中的聯網。
- 評估網路中的路由路徑，以確認始終使用目的地之間的最短路徑。網路存取分析器可協助您執行此操作。

資源

相關文件：

- 公開DNS查詢記錄
- 什麼是 IPAM？
- 什麼是 Reachability Analyzer？
- 什麼是網路存取分析器？

- CloudWatch 指標 VPCs
- 使用 Apache Parquet 格式的 VPC Flow Logs 來最佳化效能並降低網路分析的成本
- 使用 Amazon CloudWatch 指標監控您的全球和核心網路
- 持續監控網路流量和資源

相關影片：

- AWS re：Invent 2023 – 雲端聯網開發人員指南
- AWS re：Invent 2023 – 為下一步做好準備？ 設計網路實現增長和靈活性
- AWS re：Invent 2023 – 進階VPC設計和新功能
- AWS re：Invent 2022 – 深入探索 AWS 網路基礎設施
- AWS re：Invent 2020 – 使用 AWS Well-Architected Framework 建立網路最佳實務和秘訣
- AWS re：Invent 2020 – 監控和疑難排解網路流量

相關範例：

- AWS 聯網研討會
- AWS 網路監控
- 在 上觀察並診斷您的網路 AWS
- 在 上尋找和解決網路設定錯誤 AWS

程序和文化

在架構工作負載時，您可以採取一些原則和實務來協助您更有效率地執行高效能雲端工作負載。這個重點領域提供了各種最佳實務，以培養高效能雲端工作負載的文化。

打造這類文化時，請考慮以下重要原則：

- 基礎設施即程式碼：使用 AWS CloudFormation 範本等方法定義您的基礎設施即程式碼。使用範本可讓您將基礎設施與應用程式程式碼和組態一起置於原始檔控制中。這可讓您在基礎設施中套用開發軟體時所使用的相同做法，進而快速進行迭代。
- 部署管道：使用持續整合/持續部署 (CI/CD) 管道 (例如，原始程式碼儲存庫、建置系統、部署和測試自動化) 來部署您的基礎架構。這樣您就可以在反覆執行的過程中，採用可重複、一致且低成本的方式進行部署。
- 定義明確的指標：設定並監控指標以擷取關鍵績效指標 (KPI)。我們建議您同時使用技術和業務指標。對於網站或行動應用程式，關鍵指標是擷取第一個位元組或轉譯的時間。其他一般適用的指標包括執行緒計數、垃圾回收率和等待狀態。業務指標 (例如每個請求的彙總累計成本) 會提示您降低成本的方法。仔細考慮您計劃如何解釋指標。例如，您可以選擇最大值或第 99 個百分位數，而非平均值。
- 自動執行效能測試：在部署程序中，在成功通過快速執行測試之後，會自動啟動效能測試。自動化應建立一個新的環境，設定如測試資料之類的初始條件，然後執行一系列基準測試和負載測試。這些測試的結果應與組建版本綁定，方便您追蹤長時間的效能變化。對於長期執行的測試，您可以讓管道的這個部分與組建版本的其餘部分不同步。或者，您可以使用 Amazon EC2 Spot 執行個體在夜間執行效能測試。
- 負載產生：您應建立一系列的測試指令碼來複寫綜合性或預錄的使用者旅程。這些指令碼應該是冪等及非耦合的形式，而且您可能需要納入預熱型指令碼才能產生有效的結果。您的測試指令碼應盡可能地複寫生產環境中的使用行為。您可以使用軟體或軟體即服務 (SaaS) 解決方案來產生負載。可以考慮使用 AWS Marketplace 解決方案和 Spot 執行個體 — 這些可能是經濟實惠的負載產生方式。
- 效能可見度：關鍵指標應對您的團隊可見，尤其是針對每個組建版本的指標。這可讓您查看隨時間變化出現的任何顯著的正面或負面趨勢。您也應顯示錯誤或例外狀況數量的指標，以確保您測試的是可運作的系統。
- 視覺化：使用視覺化技術可以清楚指出何處出現效能問題、熱點、等待狀態或較低的利用率。在架構圖上重疊效能指標 — 呼叫圖表或程式碼有助於快速識別問題。
- 定期審查程序：架構效能不佳通常是效能審查程序不存在或中斷的結果。如果您的架構效能不佳，則實作效能審查程序可讓您不斷反覆進行改善。
- 持續優化：培養文化以持續優化雲端工作負載效能達成效率。

最佳實務

- PERF05-BP01 建立用於測量工作負載運作狀態和效能的關鍵績效指標 (KPI)
- PERF05-BP02 使用監控解決方案來了解效能最關鍵的領域
- PERF05-BP03 定義改善工作負載效能的程序
- PERF05-BP04 Load 測試您的工作負載
- PERF05-BP05 使用自動化主動修復效能相關問題
- PERF05-BP06 保留工作負載和服務 up-to-date
- PERF05-BP07 定期審查指標

PERF05-BP01 建立用於測量工作負載運作狀態和效能的關鍵績效指標 (KPI)

識別定量和定性衡量工作負載效能的 KPI。KPI 有助於測量與業務目標相關的工作負載的運作狀態和效能。

常見的反模式：

- 您只能監控系統層級指標，以深入了解工作負載，而不了解這些指標的業務影響。
- 假設 KPI 已做為標準指標資料發佈和共用。
- 沒有定義定量、可衡量的 KPI。
- 沒有將 KPI 與業務目標或策略保持一致。

建立此最佳實務的優勢：找出代表工作負載健康狀態和效能的特定 KPI，有助團隊以一致的標準排定優先事項並定義成功的業務成果。與所有部門共用這些指標可提供閾值、期望和業務影響的可見性和一致性。

未建立此最佳實務時的曝險等級：高

實作指引

KPI 可讓業務和工程團隊以一致的標準衡量目標和策略，以及將這些因素結合以產生商業成果的方式。例如，網站工作負載可能使用頁面載入時間，作為整體效能的指示。此指標將是衡量使用者體驗的多個資料點之一。除了找出頁面載入時間閾值外，您還應該記錄未符合理想效能時預期的成果或業務風險。較長的頁面載入時間會直接影響您的使用者，降低其使用者體驗等級，並可能導致客戶流失。當您定義

KPI 閾值時，請同時結合業界基準和最終使用者期望。例如，如果目前業界基準是網頁在兩秒內載入，但最終使用者期望網頁在一秒內載入，則您在建立 KPI 時應將這兩個資料點列入考慮。

團隊必須使用即時精密資料和歷史資料作為參考，來評估工作負載 KPI，並建立儀表板，針對 KPI 資料執行指標數學，以衍生營運和使用率洞察。KPI 應該加以記錄，並包含支援業務目標和策略的 KPI 和閾值，並應映射至受監控的指標。當業務目標、策略或最終使用者需求變更時，應該重新檢視 KPI。

實作步驟

- 識別利益相關者：識別和記錄關鍵業務利益相關者，包括開發和運營團隊。
- 定義目標：與利益相關者合作，以定義和記錄工作負載的目標。考慮工作負載的關鍵效能層面，例如輸送量、回應時間和成本，以及業務目標，例如使用者滿意度。
- 審核業界最佳實務：審核業界最佳實務，以找出符合您工作負載目標的相關 KPI。
- 識別指標：找出符合您工作負載目標的指標，可協助您衡量效能和業務目標。建立以這些指標為基礎的 KPI。範例指標是諸如平均回應時間或並發使用者數量等測量值。
- 定義並記錄 KPI：使用業界最佳實務和工作負載目標，為工作負載 KPI 設立目標。使用此資訊，來設定嚴重性或警示層級的 KPI 閾值。找出並記錄未達到 KPI 的風險和影響。
- 實作監控：使用監控工具 (例如 Amazon CloudWatch 或 AWS Config) 來收集指標並測量 KPI。
- 以視覺化方式傳達 KPI：使用 Amazon QuickSight 等儀表板工具，來視覺化 KPI 並與利益相關者溝通。
- 分析和最佳化：定期審核和分析 KPI，以找出需要改善的工作負載領域。與利益相關者合作以實作這些改進。
- 重新檢視和調整：定期審核指標和 KPI 以評估其有效性，尤其是在業務目標或工作負載效能變更時。

資源

相關文件：

- CloudWatch 文件
- 監控、日誌記錄和效能 AWS Partner
- AWS 可觀測性工具
- 大規模雲端遷移關鍵績效指標 (KPI) 的重要性
- 如何使用 KPI 儀表板追蹤成本最佳化 KPI

- X-Ray 文件
- 使用 Amazon CloudWatch 儀表板
- QuickSight KPI

相關影片：

- AWS re:Invent 2023 - 優化成本和效能並追蹤緩解措施的進度
- AWS re:Invent 2023 - 使用 AWS Health 大規模管理資源生命週期事件
- AWS re:Invent 2023 - Pinterest 的效能與效率：最佳化最新的執行個體
- AWS re:Invent 2022 - AWS 優化：立即見效的可操作步驟
- AWS re:Invent 2023 - 建立有效的可觀測性策略
- AWS Summit SF 2022 - 使用 AWS 獲得全堆疊可觀測性和應用程式監控
- AWS re:Invent 2023 - 針對前 1000 萬個使用者在 AWS 上進行擴展
- AWS re:Invent 2022 - Amazon 如何使用更好的指標來提高網站效能
- 為您的企業建立有效的指標策略 | AWS 活動

相關範例：

- 使用 QuickSight 建立儀表板

PERF05-BP02 使用監控解決方案來了解效能最關鍵的領域

了解並找出提高工作負載效能將對效率或客戶體驗產生正面影響的地方。例如，具有大量客戶互動的網站可受益於邊緣服務的使用，因為這樣可以將內容交付移至更接近客戶的地方。

常見的反模式：

- 您假設CPU使用率或記憶體壓力等標準運算指標足以解決效能問題。
- 您只會使用所選監控軟體記錄的預設指標。
- 您只會在有問題時審查指標。

建立此最佳實務的優點：了解效能的關鍵領域有助於工作負載擁有者監控KPIs和排定高影響改善的優先順序。

未建立此最佳實務時的曝險等級：高

實作指引

設定 end-to-end追蹤以識別流量模式、延遲和關鍵效能區域。監控您的資料存取模式，以確定是否有緩慢的查詢或分段及分割不佳的資料。使用負載測試或監控來找出工作負載受限領域。

透過了解架構、流量模式和資料存取模式，來提高效能效率，並確定延遲和處理時間。找出隨著工作負載的成長，可能會影響客戶體驗的潛在瓶頸。調查這些領域後，請審視自己可以部署哪個解決方案，來消除這些效能疑慮。

實作步驟

- 設定 end-to-end監控以擷取所有工作負載元件和指標。以下是 上監控解決方案的範例 AWS。

服務	在哪裡使用
Amazon CloudWatch Real-User Monitoring（RUM）	擷取來自實際使用者用戶端和前端工作階段的應用程式效能指標。
AWS X-Ray	透過應用程式層追蹤流量，並找出組成部分和相依性之間的延遲。使用 X-Ray 服務地圖，查看工作負載組成部分之間的關係和延遲。
Amazon Relational Database Service 效能洞見	檢視資料庫效能指標並找出效能待改善之處。
Amazon RDS增強型監控	檢視資料庫 OS 效能指標。
Amazon DevOpsGuru	偵測異常作業模式，以便在營運問題影響客戶之前識別。

- 執行測試，來產生指標、確定流量模式、瓶頸和關鍵效能區域。以下是如何進行測試的一些範例：
  - 設定CloudWatch合成 Canary，以程式設計方式使用 Linux cron 任務或速率表達式模擬瀏覽器型使用者活動，以隨時間產生一致的指標。
  - 使用 AWS 分散式負載測試解決方案，來產生尖峰流量或以預期成長速率測試工作負載。
- 評估指標和遙測，來找出關鍵的效能領域。與您的團隊一起檢閱這些領域，討論監控和解決方案，以避免瓶頸。
- 進行效能改善的實驗，並透過資料來衡量這些變更。例如，您可以使用CloudWatch Evidently 來測試對工作負載的新改進和效能影響。

資源

相關文件：

- re：Invent 2023 可 AWS 觀測性的新功能
- Amazon 建置者資料中心
- X-Ray 文件
- Amazon CloudWatch RUM
- Amazon DevOpsGuru

相關影片：

- AWS re：Invent 2023 - 【LAUNCH】 現代工作負載的應用程式監控
- AWS re：Invent 2023 - 實作應用程式可觀測性
- AWS re：Invent 2023 - 建立有效的可觀測性策略
- AWS Summit SF 2022 - 全堆疊可觀測性和應用程式監控 AWS
- AWS re：Invent 2022 - AWS optimization：立即結果的可行步驟
- AWS re：Invent 2022 - Amazon Builders 程式庫：Amazon 卓越營運 25 年
- AWS re：Invent 2022 - Amazon 如何使用更好的指標來改善網站效能
- 使用 Amazon CloudWatch Synthetics 視覺化監控應用程式

相關範例：

- 使用 Amazon CloudWatch Synthetics 測量頁面載入時間
- Amazon CloudWatch RUM Web Client
- X-Ray SDK for Python
- 上的分散式負載測試 AWS

PERF05-BP03 定義改善工作負載效能的程序

定義一個程序，以在新的服務、設計模式、資源類型和組態可用時對其進行評估。例如，對新的執行個體方案執行現有的效能測試，以判斷其是否可能改善工作負載。

常見的反模式：

- 您假設目前的架構是靜態的，且不會隨著時間而更新。
- 您會隨時間導入架構變更，而且無須指標佐證。

建立此最佳實務的優勢：定義進行架構變更的程序後，您就能使用收集的資料，以隨著時間影響工作負載。

未建立此最佳實務時的曝險等級：中

實作指引

工作負載的效能有一些關鍵限制。記錄這些內容，以便您知道哪種創新可以改善工作負載的效能。當新服務或技術可用時，請使用此資訊來找出緩解限制或瓶頸的方法。

識別工作負載的關鍵效能限制。記錄工作負載的效能限制，讓您知道哪些類型的創新可能會改善工作負載的效能。

實作步驟

- 識別 KPIs：識別 中KPIs概述的工作負載效能PERF05-BP01 建立用於測量工作負載運作狀態和效能的關鍵績效指標 (KPI)，以基準化工作負載。
- 實作監控：使用AWS 可觀測性工具收集效能指標並測量 KPIs。
- 執行分析：執行深入分析，以找出工作負載中效能不佳的區域 (例如組態和應用程式的程式碼)，步驟請參閱 PERF05-BP02 使用監控解決方案來了解效能最關鍵的領域。使用分析和效能工具，來確定效能改進策略。
- 驗證改進：使用沙盒或生產前環境，來驗證改進策略的有效性。
- 實作變更：實作生產中的變更，並持續監控工作負載的效能。記錄改進項目並與利益相關者溝通這些變更。
- 重新檢視和完善：定期檢視您的績效改善程序，以找出需要提高的領域。

資源

相關文件：

- AWS 部落格

- 新功能 AWS
- AWS 技能建置器

相關影片：

- AWS re：Invent 2022 - 提供永續、高效能的架構
- AWS re：Invent 2023 - 最佳化成本和效能，並追蹤緩解進度
- AWS re：Invent 2022 - AWS optimization：立即結果的可行步驟
- AWS re：Invent 2022 - 透過最佳實務指引最佳化 AWS 工作負載

相關範例：

- AWS Github

PERF05-BP04 Load 測試您的工作負載

對工作負載執行負載測試，以確認它可以處理生產負載並識別任何效能瓶頸。

常見的反模式：

- 可以對工作負載的個別部分進行負載測試，而非整個工作負載。
- 可以在與生產環境不同的基礎設施中進行負載測試。
- 您只對預期的 (而非超標) 負載進行負載測試，以協助預測未來可能發生問題的位置。
- 您可以在未諮詢 Amazon 測試政策的情況下執行負載EC2測試，並提交模擬事件提交表單。這會導致您的測試無法執行，因為它看起來像事件 denial-of-service。

建立此最佳實務的優勢：在負載測試過程中測量效能時，會顯示您將在負載增加到何種程度時受到影響。這可讓您能夠在工作負載受到影響之前預測所需的變更。

未建立此最佳實務時的曝險等級：低

實作指引

雲端中的負載測試是在實際條件下，以預期的使用者負載來衡量雲端工作負載效能的程序。此程序包括佈建類似生產環境的雲端環境、使用負載測試工具產生負載，以及分析指標以評估工作負載處理實際負

載的能力。必須使用生產資料的綜合或處理過的版本 (移除敏感或可識別身分的資訊) 執行負載測試。在交付管道中自動執行負載測試，並將結果與預先定義的KPIs閾值進行比較。此程序有助於您持續達到所需的效能。

實作步驟

- 定義測試目標：確定您要評估的工作負載效能層面，例如輸送量和回應時間。
- 選擇測試工具：選擇並設定適合您工作負載的負載測試工具。
- 設定您的環境：根據生產環境設定測試環境。您可以使用 AWS 服務來執行生產規模環境，以測試您的架構。
- 實作監控：使用 Amazon CloudWatch 等監控工具，收集架構中各項資源的指標。也可以收集和發布自訂指標。
- 定義方案：定義負載測試方案和參數 (如測試持續時間和使用者數量)。
- 進行負載測試：大規模執行測試方案。利用 AWS 雲端 來測試工作負載，以探索其無法擴展的位置，或它是否以非線性方式擴展。例如，使用 Spot 執行個體以低成本產生負載，並在生產中遇到瓶頸之前發現瓶頸。
- 分析測試結果：分析結果以找出效能瓶頸和需要改善的區域。
- 記錄和分享調查結果：記錄並報告調查結果和建議。與利益相關者分享此資訊，協助他們做出有關效能最佳化策略的明智決策。
- 不斷反覆執行：負載測試應定期執行，尤其是在系統更新變更之後。

資源

相關文件：

- Amazon CloudWatch RUM
- Amazon CloudWatch Synthetics
- 上的分散式負載測試 AWS

相關影片：

- AWS Summit ANZ 2023：透過 AWS 分散式負載測試，放心加速
- AWS re：Invent 2022 - AWS 為前 1，000 萬使用者擴展
- 使用 AWS 解決方案解決：分散式負載測試

- AWS re：Invent 2021 - 透過使用 Amazon 的終端使用者洞察最佳化應用程式 CloudWatch RUM
- Amazon CloudWatch Synthetics 示範

相關範例：

- 上的分散式負載測試 AWS

PERF05-BP05 使用自動化主動修復效能相關問題

使用關鍵績效指標（KPIs）結合監控和警示系統，主動解決與績效相關的問題。

常見的反模式：

- 您只讓操作人員有能力對工作負載進行操作變更。
- 您讓所有警示篩選到操作團隊，無須主動修復。

建立此最佳實務的優勢：主動修復警示動作能夠讓支援人員專注在無法自動採取行動的項目上。這有助於操作人員無須疲於處理所有警示，而僅專注於關鍵警示。

未建立此最佳實務時的曝險等級：低

實作指引

使用警示觸發自動化動作，盡可能修復問題。如果無法自動回應，則將警示上報給能夠回應的人員。例如，您可能有一個系統，可以預測預期的金鑰效能指標（KPI）值，並在違反特定閾值時發出警示，或者如果超出KPIs預期值，則工具可以自動停止或復原部署。

實作可在工作負載執行時提供效能可見度的程序。建置監控儀表板並建立效能預期的基準規範，以確定工作負載是否以最佳狀態執行。

實作步驟

- 識別修復工作流程：識別並了解可自動修復的效能問題。使用 AWS 監控解決方案，例如 Amazon CloudWatch 或 AWS X-Ray 來協助您進一步了解問題的根本原因。
- 定義自動化程序：建立 step-by-step可用於自動修正問題的修復程序。
- 設定啟動事件：將事件設定為自動啟動修復程序。例如，您可以定義觸發條件，在執行個體達到特定CPU使用率閾值時自動重新啟動執行個體。

- 自動化修復：使用 AWS 服務和技術來自動化修復程序。例如，AWS Systems Manager Automation 提供安全且可擴展的方式，來自動化修復程序。如果變更無法成功解決問題，則請務必使用自我修復邏輯來還原變更。
- 測試工作流程：在生產前環境中測試自動修復程序。
- 實作工作流程：在生產環境中實作自動修復。
- 制定說明手冊：制定並記錄說明手冊，其中概述了補救計畫的步驟，包括啟動事件、補救邏輯和採取的動作。確保培訓利益相關者，以協助他們有效地應對自動補救事件。
- 審查和完善：定期評估自動補救工作流程的有效性。如有必要，請調整啟動事件和補救邏輯。

資源

相關文件：

- CloudWatch 文件
- 監控、記錄和效能 AWS Partner Network 合作夥伴
- X-Ray 文件
- 在 中使用警示和警示動作 CloudWatch
- 建立卓越營運雲端自動化實務：最佳實務來自 AWS Managed Services
- 使用自動表格優化來自動調整您的 Amazon Redshift 效能

相關影片：

- AWS re：Invent 2023 - 自動化擴展、修復和智慧自我修復的策略
- AWS re：Invent 2023 - 【LAUNCH】 現代工作負載的應用程式監控
- AWS re：Invent 2023 - 實作應用程式可觀測性
- AWS re：Invent 2021 - 智慧自動化雲端操作
- AWS re：Invent 2022 - 在 AWS 環境中大規模設定控制項
- AWS re：Invent 2022 - 使用 自動化修補程式管理和合規 AWS
- AWS re：Invent 2022 - Amazon 如何使用更好的指標來改善網站效能
- AWS re：Invent 2023 - 卸載：診斷並解決 Amazon 的效能問題 RDS
- AWS re：Invent 2021 -{New Launch} 自動偵測並解決 Amazon DevOpsGuru 的問題
- AWS re：Invent 2023 - 集中您的操作

相關範例：

- CloudWatch 記錄自訂警示

PERF05-BP06 保留工作負載和服務 up-to-date

繼續使用 up-to-date新的雲端服務和功能，以採用有效率的功能、移除問題，並改善工作負載的整體效能效率。

常見的反模式：

- 假設您目前的架構為靜態，且不會隨著時間的推移而更新。
- 您沒有任何系統或定期規律可評估更新的軟體與套件是否與您的工作負載相容。

建立此最佳實務的優點：透過建立程序以持續 up-to-date使用新服務和產品，您可以採用新功能和功能、解決問題並改善工作負載效能。

未建立此最佳實務時的曝險等級：低

實作指引

在新服務、設計模式和產品功能推出時，評估提升效能的方法。透過評估、內部討論或外部分析，確定哪些方法可以提高工作負載效能或效率。定義程序來評估與工作負載相關的更新、新功能和服務。例如，建立使用新技術的概念證明或與內部小組協商。嘗試新的想法或服務時，執行效能測試以衡量其對工作負載效能的影響。

實作步驟

- 清查工作負載：清查工作負載軟體和架構，並識別需要更新的元件。
- 識別更新來源：找出與工作負載組成部分相關的新聞和更新來源。例如，您可以訂閱符合您工作負載元件的產品的 AWS 部落格最新消息。您可以訂閱RSS摘要或管理您的電子郵件訂閱 。
- 定義更新排程：定義排程以評估工作負載的新服務和特徵。
  - 您可以使用 AWS Systems Manager Inventory 從您的 Amazon EC2執行個體收集作業系統（OS）、應用程式和執行個體中繼資料，並快速了解哪些執行個體正在執行軟體政策所需的軟體和組態，以及哪些執行個體需要更新。
- 評估最新更新：了解如何更新工作負載的元件。利用雲端的靈活性快速測試新特徵對工作負載有何改善，藉以提高效能效率。

- 使用自動化：使用更新程序自動化，以減少部署新功能的工作量，並避免手動程序引起的錯誤。
  - 您可以使用 CI/CD 自動更新 AMIs、容器映像，以及與雲端應用程式相關的其他成品。
  - 可以使用 AWS Systems Manager Patch Manager 之類的工具來自動化系統更新流程，並使用 AWS Systems Manager Maintenance Windows 來排程活動。
- 記錄過程：記錄用於評估更新和新服務的過程。向擁有者提供所需的時間和空間，來研究、測試、試驗和驗證更新及新服務。參考文件化的業務需求KPIs，並協助排定哪些更新將對業務產生正面影響的優先順序。

資源

相關文件：

- AWS 部落格
- 新功能 AWS
- 使用自動化映像建置器管道實作 up-to-dateEC2映像

相關影片：

- AWS re：Inforce 2022 - 使用 自動化修補程式管理和合規 AWS
- 所有事項修補程式： AWS Systems Manager | AWS Events

相關範例：

- 庫存和修補程式管理
- 一個可觀測性研討會

PERF05-BP07 定期審查指標

作為日常維護的一部分或對事件或事故的回應，審查收集了哪些指標。透過這些審查來識別哪些指標是解決問題的關鍵，以及哪些其他指標 (如果被追蹤) 有助於識別、解決或預防問題。

常見的反模式：

- 您讓指標長時間持續處於警示狀態。
- 您建立自動化系統無法採取行動的警示。

建立此最佳實務的優勢：持續審查正在收集的指標，以確認指標正確識別、處理或防止問題發生。如果讓指標長時間持續處於警示狀態，指標也會變得過時。

未建立此最佳實務時的曝險等級：中

實作指引

不斷改進指標收集和監控。作為對事故或事件的回應的一部分，評估哪些指標有助於解決問題，哪些指標可以幫助解決問題但未被追蹤。使用此方法提高所收集指標的品質，從而可以防止事故發生或更快地解決將來的事故。

作為對事故或事件的回應的一部分，評估哪些指標有助於解決問題，哪些指標可以幫助解決問題但未被追蹤。使用此方法提高所收集指標的品質，從而可以防止事故發生或更快地解決將來的事故。

實作步驟

- 定義指標：定義與您的工作負載目標一致的關鍵效能指標以進行監控，包括回應時間和資源使用率等指標。
- 建立基準：設定各指標的基準和期望值。基準應提供參考點以識別偏差或異常。
- 設定規律：設定規律 (例如每週或每月一次) 以審核重要指標。
- 識別效能問題：每次審查期間都會評估趨勢，以及與基準值的偏差。查看是否有任何效能瓶頸或異常情況。對於已確認的問題，請展開深入根本原因分析，以了解問題背後的主要原因。
- 識別修正動作：使用您的分析來識別修正動作。這可能包括參數調整、修正錯誤和擴展資源。
- 記錄調查結果：記錄您的調查結果，包括已識別的問題、根本原因和修正動作。
- 反覆執行並改善：持續評估並改善指標審核過程。使用從以前的審核中學到的經驗教訓，隨著時間的推移提升程序。

資源

相關文件：

- CloudWatch 文件
- 使用 CloudWatch 代理程式從 Amazon EC2 執行個體和內部部署伺服器收集指標和日誌
- 使用 CloudWatch Metrics Insights 查詢您的指標
- 監控、日誌記錄和效能 AWS Partner Network 合作夥伴
- X-Ray 文件

相關影片：

- AWS re:Invent 2022 - 在 AWS 環境中大規模設定控制項
- AWS re:Invent 2022 - Amazon 如何使用更好的指標來提高網站效能
- AWS re:Invent 2023 - 建立有效的可觀測性策略
- AWS Summit SF 2022 - 使用 AWS 獲得全堆疊可觀測性和應用程式監控
- AWS re:Invent 2023 - 減輕負擔：使用 Amazon RDS 診斷和解決效能問題

相關範例：

- 使用 QuickSight 建立儀表板
- CloudWatch 儀表板

結論

為實現和維持效能達成效率，需要採取資料驅動的方法。您應積極考慮存取模式和權衡因素，以便讓您能夠執行最佳化進而獲得更高效能。使用以基準化分析和負載測試為基礎的審查程序，可讓您選取適當的資源類型和組態。將基礎設施視為程式碼可協助您快速安全地發展架構，同時可使用資料針對架構作出以事實為基礎的決策。結合使用主動監控和被動監控，可確保您架構的效能不會隨著時間的推移而降低。

AWS 會努力協助您建置高效執行的架構，同時提供商業價值。使用本白皮書中所討論的工具和技術以確保成功。

貢獻者

下列個人和組織為本文件作出了貢獻：

- Sam Mokhtari，Amazon Web Services 資深效率主管與解決方案架構師
- Josh Hart，Amazon Web Services 解決方案架構師
- Richard Trabing，Amazon Web Services 解決方案架構師
- Brett Looney，Amazon Web Services 首席解決方案架構師
- Nina Vogl，Amazon Web Services 首席解決方案架構師
- Eric Pullen，Amazon Web Services 解決方案架構師
- Julien Lépine，Amazon Web Services 專家 SA 經理
- Ronnen Slasky，Amazon Web Services 解決方案架構師

深入閱讀

如需其他協助，請參考下列資源：

- AWS Well-Architected 架構
- AWS 架構中心

文件修訂

若要收到此白皮書更新的通知，請訂閱 RSS 摘要。

變更	描述	日期
次要最佳實務更新	PERF03-BP04 的更新中增加了新的服務建議。	2024 年 11 月 6 日
已更新最佳實務指引	整個支柱的多個小更新。	2024 年 6 月 27 日
主要更新與重新建構	支柱經過重新建構，分成五個最佳實務領域 (原為八個)。內容已整合成五個領域並已更新。 新的最佳實務領域包括選擇架構、運算與硬體、資料管理、聯網與內容交付及程序和文化。	2023 年 10 月 3 日
次要更新	移除非包容性語言。	2023 年 4 月 13 日
新框架的更新	最佳實務已更新，納入了規範性指引，並增加了新的最佳實務。	2023 年 4 月 10 日
白皮書已更新	最佳實務更新了新的實作指引。	2022 年 12 月 15 日
白皮書已更新	已擴充最佳實務並新增了改善計畫。	2022 年 10 月 20 日
次要更新	已移除非包容性語言。	2022 年 4 月 22 日
次要更新	已更新連結。	2021 年 3 月 10 日
次要更新	將 AWS Lambda 逾時變更為 900 秒並更正了 Amazon	2020 年 10 月 5 日

	Keyspaces (適用於 Apache Cassandra) 的名稱。	
次要更新	修正了中斷的連結。	2020 年 7 月 15 日
新框架的更新	對內容進行了重大審查和更新	2020 年 7 月 8 日
白皮書已更新	語法問題的小幅度更新	2018 年 7 月 1 日
白皮書已更新	重新整理白皮書以體現 AWS 上的變更	2017 年 11 月 1 日
初次出版	效能達成效率支柱 – AWS Well-Architected Framework 已發布。	2016 年 11 月 1 日

注意

客戶有責任對本文件中的資訊進行自己的獨立評定。本文件：（a）僅供參考，（b）代表目前的 AWS 產品產品和實務，這些產品和實務可能隨時變更，恕不另行通知，且（c）不會從 AWS 及其附屬公司、供應商或授權方提供「原樣」的任何承諾 AWS 或保證，而沒有任何明示或暗示的保證、陳述或條件。 AWS 對其客戶的責任和責任受 AWS 協議控制，本文件不屬於 與其 AWS 客戶之間的任何協議，也未對其進行修改。

© 2023 Amazon Web Services, Inc. 或其附屬公司。保留所有權利。

AWS 詞彙表

如需最新的 AWS 術語，請參閱AWS 詞彙表 參考 中的AWS 詞彙表。