消除資料孤島：高效完成多系統查詢

IBM Cloud Pak for Data 的資料虛擬化

特色

- 以個別或集體方式，跨多個資料庫和大數據儲存庫進行查詢
- 集中化的存取控制和治理
- 讓眾多資料庫（即使分布在世界各地）以單一形式出現在應用程式面前
- 利用可擴充的強大平台來簡化資料分析

背景

資料無所不在，當今世上最佳的業務運作模式就是資料驅動。企業從越來越多且日益多元化的來源收集資料，然後分析資料並執行營運，而資料來源可能多達數千個或數百萬個。集中收集、治理、儲存、處理與分析該資料的複雜性、成本、時間及出錯風險呈指數成長。在此同時，作為資料來源的各種資料庫和儲存庫越來越強大，而且內建豐富的處理和資料儲存功能可用。

資料虛擬化概觀

IBM® Cloud™ Pak for Data（舊稱 IBM Cloud Private for Data）裡的資料虛擬化是獨家新技術，它可以將所有資料來源連接到單一自我平衡資料來源或資料庫集合（稱為*群集*）。請參閱圖 1。再也不用在集中的位置進行複製與儲存資料，然後執行分析查詢。分析應用程式可提交查詢，在資料來源所在的伺服器上執行查詢處理。查詢結果會合併在群集中，然後傳回原應用程式。不需複製任何資料，資料僅存在於來源。

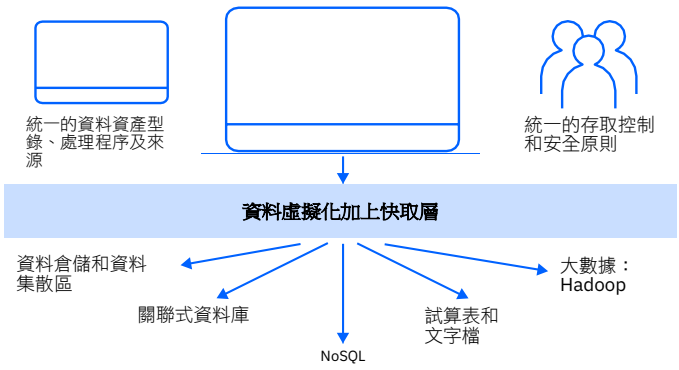

圖 1： Cloud Pak for Data 的資料虛擬化

資料虛擬化如何運作

應用程式連接到 IBM 資料虛擬化的方式與其連接到單一 IBM Db2® 資料庫類似。兩者一旦相連接，應用程式可以對系統提交查詢，就像它們在查詢單一資料來源資料庫一樣。工作負載會由與查詢資料有關的所有資料來源伺服器共同協作以分散運算資源。

重要功能

IBM 資料虛擬化中有多個重要功能可讓企業更有效地處理他們的資料。

協作運算

運用每個資料來源的處理能力，並且從每個資料來源實體儲存處存取資料，這樣可避免因移動與複製資料所發生的延遲。此外，所有儲存庫資料都可以即時存取，治理和資料錯誤問題幾乎不復存在。您不需要擷取、轉換、載入 (ETL) 與複製資料儲存體，因此可加快處理速度。此處理程序可以用比現有方法更快速、更可靠的方式，為決策制定應用程式或分析師提供即時洞察。它也可以與現有方法互補，在仍有需要為了歷史、歸檔或合規等用途而複製與移動部分資料時，與現有方式同時並存。

結構交疊 (Schema Folding)

在分散式資料系統中有一個常見實務範例，即眾多資料庫使用一個通用的結構 (Schema) 來儲存資料。舉例來說，您有多個用來儲存銷售資料或交易資料的資料庫，每個資料庫分屬不同租戶或地區。IBM 資料虛擬化可自動偵測跨不同系統的通用結構，並容許它們在資料虛擬化中以單一結構的形式出現，此處理程序稱為結構交疊 (schema folding)。例如，在 20 個資料庫中都有「銷售」表格，現在能夠以單一「銷售」表格出現，並可透過結構化查詢語言 (SQL) 將該表格當做一個虛擬表格進行查詢。

簡易的合併檢視工具

精巧的內嵌工具可讓您跨不同類型的資料庫（甚至分散在不同地理區）之間進行表格視圖的定義，如圖 2 所示。

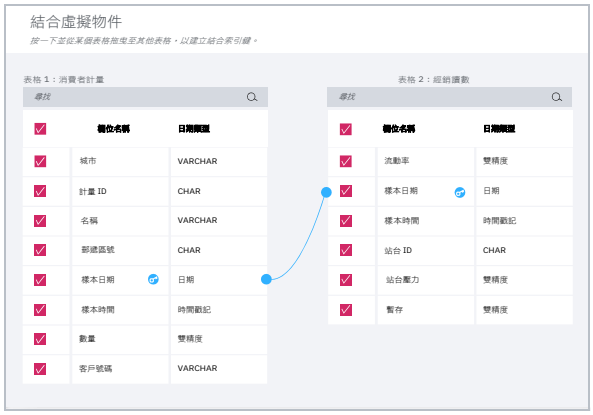

圖 2：直覺式介面讓合併表格視圖變簡單

安全性

在群集內部與傳回應用程式的所有通訊，全都會使用標準通訊協定，透過高度安全、健全且強大的 IBM 技術、『安全通訊協定 (SSL)』及『傳輸層安全性協定 (TLS)』來進行加密。

效能表現

IBM 資料虛擬化的對等式運算網格設計架構可提供遠勝於傳統聯合式架構的明顯好處。使用 IBM Research 提供的進階功能時，資料虛擬化引擎能夠利用進階平行處理與優化，快速從多重資料來源提供查詢結果。高度平行的協同運算模型可提供卓越的查詢效能，其查詢 100 TB 資料集的速度，比聯合式架構快了 430%[1]。IBM 資料虛擬化擁有無與倫比的多重查詢擴充能力，它可以同時結合與彙總十多個運作中的系統。

IBM 資料虛擬化不只速度快，還可以自動尋找資料庫和表格，讓多重資料來源的資訊查詢變簡單。查詢可以輕鬆結合多重來源的資料，包括關聯式資料庫、NoSQL 來源、試算表及純文字檔。

平台支援

IBM 資料虛擬化技術以單一 Db2 資料庫實例的形式與應用程式相連。因此，熱門的 Db2 連線用戶端和應用程式可以直接連接至 IBM 資料虛擬化，而且無需修改就能運作。即使接受查詢的資料來源集合包含許多類型的資料來源也一樣，例如：

- PostgreSQL
- Oracle
- Netezza®
- Microsoft SQL Server

IBM 資料虛擬化技術可以轉換成所有 SQL 用語，反之亦然。因此，您的應用程式可以自由撰寫 SQL、程序語言/SQL (PL/SQL) 及 SQL PL，就好像它們直接在 Db2 資料庫中運作，無需嘗試判斷目標資料系統是否支援該語法。例如，一些熱門工具可以與 IBM 資料虛擬化連接，無需任何修改或升級，包括

- IBM Cognos® Business Intelligence (BI) 軟體
- Tableau
- MicroStrategy
- Looker
- Plotly
- R
- Jupyter

應用程式所連接的資料虛擬化服務節點是 Cloud Pak for Data 當中的微服務。

Apache Hive	IBM Informix® 資料庫伺服器
Cloudera Impala	MariaDB
Db2 軟體	MySQL
IBM Db2 Big SQL	Netezza
IBM Db2 Event Store	Oracle
DerbyDB	PostgreSQL
Excel 和逗點分隔值 (CSV) 檔案	SQL Server
Hortonworks Data Platform (HDP) 搭配 Apache Hive	

表 1：支援的資料來源

最低硬體需求

Cloud Pak for Data 內的資料虛擬化需要下列配置：

- 16 (v) 核心處理器
- 64 GB 以上的實體隨機存取記憶體 (RAM)
- 建議 200 GB 的磁碟空間

IBM 資料虛擬化的常見實務範例

IBM 資料虛擬化很適合用來對高度分散的資料集執行分析，其中資料和分析結果對時間很敏感。它也很適用於對特定資料集執行單次分析。不僅如此，它也適用於業務運作對於分析結果的速度有高度要求的情況；而從資料來源批次複製的傳統方式往往會造成時間延遲。

許多組織會複製資料並建立新的資料儲存庫，以滿足業務單位 (LOB) 的分析需要。此處理程序需要配置實體資產，同時建立並維護新的 ETL，如此才能載入與轉換資料至那些儲存庫中。然而，當資料可供 LOB 使用時，資料往往已經過期。

現有方法已逼近許多 IT 組織的飽和點，但資料來源的數目和多元性還有分析需求日益增加，此方法已無法再擴充。IBM 資料虛擬化可提高 IT 組織的生產力，並為業務單位提供能夠存取企業層面資料的可擴充方法。

在許多實例中，複製或移動資料，例如個人資訊，會產生政策或法律問題；這些限制可能會妨礙獲取個人背景分析結果的商業需求。IBM 資料虛擬化將受保護的資料留在來源處，僅傳回個人背景查詢結果，藉此協助解決這類問題。

過去，資料科學家必須建立資料湖，從感興趣的來源複製資料並加以整合，然後才能透過分析檢驗假設。IBM 資料虛擬化消除資料湖需求，讓資料科學家可以將 IBM Watson® Studio 之類的工具直接連接至資料來源，然後聯合其所需資料來檢驗假設。

為關鍵分析專案提供敏捷性

IBM 資料虛擬化的簡易性可讓使用者以符合其分析需要的速度，隨時透過任何屬意方式取得可據以行動的統一資料。此技術可加快整合的速度與效能，並且提升決策制定以協助您適應瞬息萬變的商業需求。

IBM Cloud Pak for Data 的資料虛擬化支援廣泛的關鍵提案，包括：

- 更快速、更輕鬆的新型態互動系統 (System of Engagement) 現代化
- 為符合業務立即需求而執行即時分析
- 進行優化以減少存取組織資料的成本和複雜性

IBM 資料虛擬化能促進自助式 BI。可重複使用的虛擬化資料資產能提供對業務單位使用者友善的資料呈現方式，讓使用者能夠與資料進行互動，完全不必瞭解實體資料層的複雜性或資料儲存位置。此外，它還可以讓多種 BI 和報告工具從資料虛擬層取得資料。

IBM 資料虛擬化提供統一的 360 度視圖。虛擬化的資料資產可即時提供完整的資料檢視。虛擬資料層作為統一、整合的商業資訊視圖，可提高使用者瞭解與運用組織資料的能力。

IBM 資料虛擬化提供敏捷的服務導向架構 (SOA) 資料服務。資料虛擬化技術提供給 SOA 應用程式所需的資料服務層。它可以加速建立虛擬資產，無需觸碰基礎來源，這歸功於它能夠自動探索與對映封裝資料存取邏輯的元資料 (metadata)。資料虛擬化還能讓多個商業服務從中央位置取得資料，並在商業服務和實體資料來源之間提供虛擬連結。

IBM 資料虛擬化提供強化的資訊控制。它透過中央存取控制、健全的安全基礎架構及減少實體資料副本來提高資料品質，從而降低風險。元資料儲存庫藉由將以下項目編目來提供透明度和可視度：組織的資料儲存庫，以及不同資料儲存庫中資料之間的關係。

摘要：轉型並加速決策制定

Cloud Pak for Data 的資料虛擬化很適合尋求以下目標的組織

- 利潤、成長與降低風險
- 增進敏捷和生產力
- 優化現有 IT 投資

它改善現有伺服器和儲存投資的運用，同時減少不必要的資料抄寫，以及相關的複製和基礎架構管理成本。透過簡化管理和一組 SQL 應用程式設計介面 (API)，可讓您的公司從即時分析產生利益。

如需 IBM 資料虛擬化的相關資訊，請造訪：
https://www.ibm.com/analytics/data-virtualization

1. 效能指標測量透過位於 IBM 矽谷實驗室的受控制測試環境進行收集，其中對各種不同的 100 TB 資料來源使用 IBM 資料虛擬化。在 2019 年 5 月進行測量，效能表現的比較基礎是 IBM Federation。

IBM Corporation
New Orchard Road
Armonk, NY 10504

2019 年 1 月美國生產

本文件從初始發佈日期開始保持最新，IBM 得隨時變更。並非所有產品與服務都可在 IBM 每個營業的國家/地區使用。

其他任何產品或程式與 IBM 產品和程式一起使用時，使用者需自行負責評估與驗證。本文件中的資訊係「依現狀」提供，不含任何明示或默示之保證，包括且不限於可售性、特定目的之適用性及未涉侵權之保證。IBM 產品的保固是依據其所隨附合約的條款內容提供。

良好安全作法聲明：IT 系統安全涉及透過預防、偵測與回應您企業內外部不當存取來保護系統和資訊安全。不當存取可能導致資訊遭到更改、破壞、不當使用或濫用，或者造成毀損或濫用您的系統，包括用來攻擊其他對象。沒有任何 IT 系統或產品應視為完全安全，在預防不當使用或存取方面，也沒有任何一種產品、服務或安全措施可達到完全有效。IBM 系統、產品及服務的設計理念是相互組合成一種符合法規的綜合性安全方法，其中必然涉及其他作業程序，另外可能需要運用其他的系統、產品或服務才能發揮最大效用。IBM 不保證任何系統、產品或服務可以倖免或使您的公司倖免於任何一方的惡意或不法行為。

所有關於 IBM 未來方針或目的之聲明，隨時可能更改或撤銷，不再另行通知，且該等聲明僅代表目標與主旨。