資料湖倉的崛起：資料價值的新時代

查詢加速器如資料湖倉（data lakehouse）將資料倉儲及資料湖整合為單一的洞察系統，使企業能夠以較低的成本加速分析，並實現資料價值最大化。

文／Stan Gibson　譯／雲翻譯

COVID-19 疫情高峰期間，需要接種 6,500 萬劑疫苗，Walgreens 製藥更新兼醫療平台技術副總裁 Luigi Guadagno 需要了解疫苗該送往何處。為了尋找答案，Luigi Guadagno 查詢 Walgreens 的資料湖倉。資料湖倉在 Microsoft Azure 上透過 Databricks 技術執行。

Luigi Guadagno 表示：「我們藉由資料湖倉了解情況」。對 Guadagno 而言，從技術層面來看，將疫苗供應與患者需求匹配的需求出現的恰逢其時。這家製藥連鎖巨頭已建立資料湖倉應對上述挑戰，正如 Guadagno 所言，「在合適的地方為合適的患者提供合適的產品。」

過往，Walgreens 曾試圖利用資料湖完成這項任務，但卻面臨了兩項重大障礙：成本及時間。由於許多企業均試圖從龐大的資料中獲得分析知識，因此深明所面臨的種種挑戰。最終結局是，企業獲取洞見的方式出現了典範轉移，即仰賴一種新的技術類別，協助企業將資料的價值最大化。

進入資料湖倉

在傳統的模式中，企業維持兩組系統，作為資料策略的一部分：一組帳務系統，用於運行業務，以及一組洞察系統（如資料倉儲），用於收集商業智慧（BI）。隨著大數據問世，出現了第二種洞察系統資料湖，提供人工智慧及機器學習（AI/ML）洞察。然而，許多企業發現，這種依賴兩組獨立洞察力系統的典範並非長遠之道。

資料倉儲的截取、轉換及載入（ETL）過程耗時費日。該過程將資料從帳務系統轉換至資料倉儲，並將資料正規化、查詢並獲得答案。同時，非結構化資料將傾入資料湖中，接著由技術純熟的資料科學家使用 Python、Apache Spark 及 TensorFlow 等工具加以分析。

在 Guadagno 的領導下，總部位於美國伊利諾伊州迪爾菲爾德的 Walgreens 將洞察系統整合為單一的資料湖倉。 Guadagno 並非採用該方法的唯一人士。越來越多公司發現，資料湖倉（通常歸類為俗稱查詢加速器的產品類別）正在滿足一項關鍵需求。

Gartner 副總裁兼分析師 Adam Ronthal 表示：「資料湖倉彌補了部分資料湖的不足。這就是我們走到今天的關鍵。許多人無法從湖中獲得價值」。在 Databricks Delta Lake 資料湖倉的案例中，來自資料倉儲的結構化資料通常會添加至資料湖中。對此，資料湖倉加入了優化層，使資料能夠更廣泛地用於收集洞見。

根據 Gartner 的分析查詢加速器市場指南，Databricks Delta Lake 資料湖倉僅僅是在日益蓬勃的市場中的一個項目，其中還包括 Snowflake、Starburst、Dremio、GridGain、DataRobot 等

供應商，或許另有其他項目正在發展中，不勝枚舉。

一家私募股權公司 Moonfare 正由 AWS 上以 PostgreSQL 為基礎的資料倉儲過渡至 AWS 上用於商業智慧及預測分析的 Dremio 資料湖倉。於2022年秋季實施時，商業用戶將能在 AWS S3 的資料基礎上落實自主分析。查詢將包括挑選對客戶效益最佳的行銷活動，以及績效最卓越的基金經理。資料湖倉亦將有助於防止欺詐。

Moonfare 的資料工程師 Angelo Slawik 表示：「你可以憑直覺從資料湖中查詢資料。來自資料倉儲環境的用戶不應在乎資料的存放位置」。Angelo Slawik 認為：「至關重要的是，過程中省去了 ETL 作業」，更補充：「透過 Dremio，若資料在 S3 中，你可以查詢任何所需的資料」。

Moonfare 在與 AWS Athena 的概念驗證運行中選擇了 Dremio，AWS Athena 是一種互動式查詢服務，可對 S3 資料作出 SQL 查詢。根據 Slawik 的說法，由於 Dremio 具備高速的效能以及高度實用的用戶界面，用戶能夠憑直覺追蹤資料脈絡，因此經證實更有能力處理資料。不僅如此，Dremio 就安全及治理提供以角色為主的瀏覽及存取管制功能，有益於德國柏林的公司遵守 GDPR 法規。

在總部位於巴黎的法國巴黎銀行，分散的資料孤島被這家銀行巨頭的不同團隊用於 BI。獨立承包商 Emmanuel Wiesenfeld 重新架構資料孤島，創建了一個集中系統，因此商業用戶（如交易員）便可在「單一資訊來源」上運行分析查詢。

Wiesenfeld 表示：「交易團隊希望能夠協力合作，但資料過於分散。分析資料的工具也很分散，導致成本高昂且難以維護」，並解釋：「我們希望集中來自許多資料源的資料，以實現及時的狀態意識。如今，用戶可編寫指令運行資料」。

Wiesenfeld 使用 GridGain 的 Apache Ignite 技術，創建了記憶體內運算架構。Wiesenfeld 表示，新方法的關鍵是從 ETL 轉向 ELT，在執行運算的同時作

出轉換，藉此簡化整個流程，Wiesenfeld 認為，執行的結果是將延遲從數小時降低至數秒鐘。此後，Wiesenfeld 成立了一家名為 Kawa 的新創公司，為其他客戶，尤其是對沖基金提供類似的解決方案。

Starburst 採取網格方法，在 Starburst Enterprise 中利用開源的 Trino 技術改善對分散式資料的存取。網格方法並非將資料轉換至中央倉儲，而是允許在原來的位置存取資料。Sophia Genetics 正在其以雲端為基礎的生物資料學 SaaS 分析平台中使用 Starburst Enterprise。原因為：將敏感的醫療資料保存在特定的國家內，就監管因素而言至關重要。總部位於瑞士的 Sophia Genetics 資料服務總監 Alexander Seeholzer 在 Starburst 的案例研究中表示：「由於合規限制，我們根本無法部署從一個中心點存取所有資料的任何系統。」

新的查詢加速平台並沒有停滯不前。Databricks 及 Snowflake 已推出資料雲端及資料湖倉，功能乃針對特定行業的公司需求所設計，如零售業及醫療業。上述舉措與巨頭公司 Microsoft Azure、Google 雲端平台以及 Amazon Web Services 推出的特定行業雲端相互呼應。

資料湖倉成為最佳實踐

Gartner 的 Ronthal 認為，資料湖朝向資料湖倉的演化趨勢銳不可擋。Ronthal 表示：「我們正朝著資料湖倉成為最佳實踐的方向前進，但每個人的演化速度各自有別。「在大多數的情況下，湖並無能力提供生產需求」。

儘管資料湖倉供應商渴望將資料倉儲納入產品組合，但 Gartner 預測倉儲將會持續存在。「分析查詢加速器不大可能取代資料倉儲，但可透過實現符合業務及技術人員要求的效能，大幅增加資料湖的價值」，這總結了針對查詢加速器市場的報告。

Forrester Research 的副總裁兼首席分析師 Noel Yuhanna 並不同意該觀點，他斷言資料湖倉勢必將取代獨立倉儲及湖。

Yuhanna 表示：「我們確實看見倉儲及湖成為資料湖倉的未來，一組系統便已足夠。」Yuhanna 提到，由於資料湖倉使企業能夠在不同的資料位置落實聯合治理，對於擁有分散式倉儲及湖的企業而言，如同 Starburst 的網格結構將滿足其需求。

Yuhanna 認為，無論採用何種方法，公司都在追求更快速的從資料中獲得價值，並表示：「公司不希望在六個月後才獲得全方位的客戶資料，而是下週就能掌握資料。我們將這種情況稱為「快速」資料。資料一經創建後，你就能開始加以分析並獲得洞見」。

從洞察系統演變為行動系統

對 Guadagno 而言，發放疫苗是一項高調、拯救生命的任務，但 Walgreens 的資料湖倉在更平凡無奇卻不可或缺的零售任務方面也完成了出色的工作，例如發送處方提醒及產品優惠券。這些流程結合了對客戶行為與藥品供應及零售庫存的理解。他表示：「這過程可以變得非常複雜，需要具備非常個人化的洞見，使我們成為以客為尊的企業」。

對於踏上相似旅程的其他企業而言，Guadagno 建議：「盡快將所有資料放到資料湖倉中。不要著手任何冗長的資料建模或合理化工作，最好思考如何創造價值。將所有資料放置在存放庫中，透過治理及合作讓每個人都能存取資料。不要把錢浪費在整合及 ETL 上。」

在 Walgreens，Databricks 資料湖倉不僅僅是為了提升科技的效率，更是整體業務策略的關鍵。Guadagno 表示：「我們的任務是創造極度個人化的體驗，從零售點開始 — 需要何物以及何時需要。這就是資料的最終目的。今後再也不需要帳務系統及洞察系統，這是一個關乎行動的系統。」