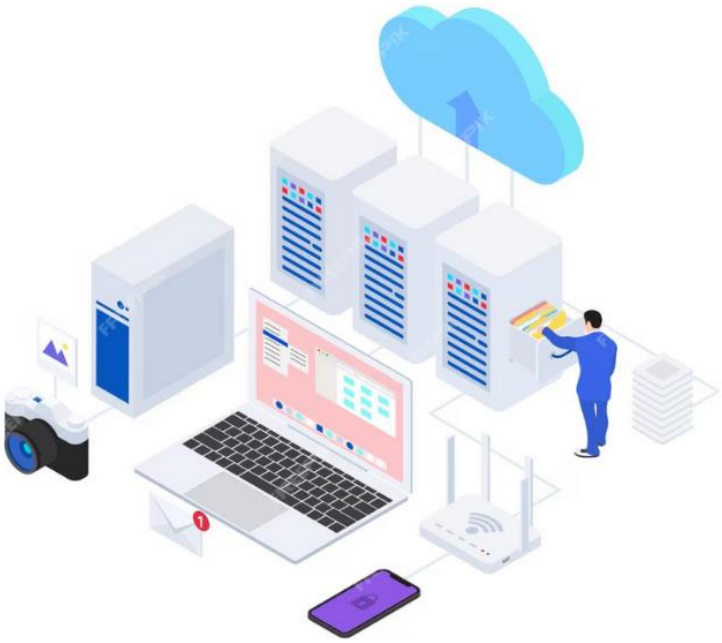
Omniwaresoft 歐立威科技

2024

ETL 全方位攻略

打造現代化資料流



ETL 重要觀念、應用場景、3 種常見資料流架構以及歐立威科技的獨家資料流生態系。幫助企業打造兼具即時性、高度資料品質和符合使用需求的現代化資流！

目錄

Omni waresoft 歐立威科技
全方位開源軟體解決方案建置商
Tel: (02) 2558-2656
mail: emarketing@omniwaresoft.com.tw

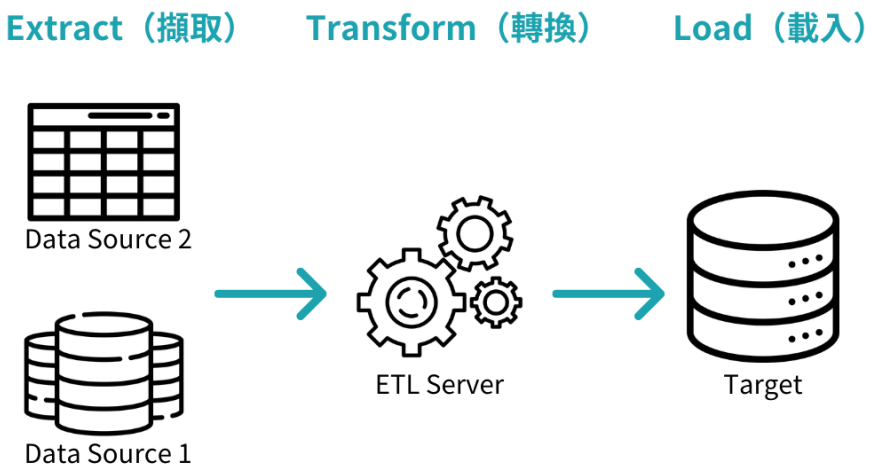
ETL 是什麼？

ETL 是一種資料處理模式，常見的資料處理模式可大致分為批次處理、近即時處理和即時處理模式，其中，ETL 被歸類為批次處理模式，ETL 作為一種核心資料處理策略，在業界被廣泛地採納與應用。

ETL - 擷取 (Extract) - 轉換 (Transform) - 載入 (Load) 是將多個資料源合併至資料倉儲的過程。此流程會從來源端將資料擷取，載入 ETL 伺服器，再經過不落地 (On-the-Fly) 運算後直接寫入目的端資料庫。

透過 ETL 解決方案，開發者可以在伺服器上轉換資料 (例如：資料運算、欄位合併和錯誤資料刪除等) ，將其整合並存入資料倉儲，最後透過 BI 工具在統一的資料集上產出報表，從而發掘高價值的商業洞見。

ETL 這種處理流程已行之有年，市面上的 ETL 工具種類繁多，在台灣常見的品牌有 Informatica、IBM Datastage、ODI、SSIS、Pentaho 等，開源的也有 Apache Nifi 等。



ETL 技術的演進與創新

ETL 解決方案起初由 ETL 單一軟體所執行，隨著資料庫運算能力提升，ETL 模式也因而得到發展與衍生，由於該模式在本質上仍屬於批次處理 (Batch Mode) 模式，因此它的即時性並不高。

之後，隨著企業更加要求時效性且希望降低資料重複傳輸。變更資料擷取（Change Data Capture, CDC）模式因而誕生，這種模式能以近即時的效率處理資料，因此又被稱為近即時資料 (Near Real-Time Mode) 處理模式。

後來，隨著串流資料 (Streaming data) 興起，加上企業要求更高的即時性，即時資料（Real-time mode）處理模式逐漸成為主流，透過結合以記憶體內運算 (In-Memory Processing) 為核心的產品，從而提升整個資料流的即時性。

上述處理模式雖以即時性進行劃分，但它們並無優劣之分，資料模式的選擇應取決於實際應用時的分析需求，以及企業在軟硬體、網路頻寬和成本上的考量。另外，來源端的運算負擔也是選擇過程中不可忽視的重要因素。

ETL 的運作原理：擷取、轉換、載入

擷取（Extract）

ETL 工具會從多種來源端（如：資料庫、API、檔案或網路服務）擷取資料，將其存放於暫存區後，再由擷取策略決定轉移至目標資料庫的頻率，以下為 3 種常見的資料擷取策略。

1. 完整擷取（Full Extraction）：
完整擷取會將資料從來源端完整地取出。在這個過程，資料不會經歷任何形式的轉換，而是保持其原有形式。這種方式的特點是它不追踪來源系統中的資料變更。因此，當系統無法識別或追蹤資料變更時，就會使用完整擷取。

每次執行擷取時，系統會將上次擷取的資料副本作為基準，並將其與新資料進行比對，從而識別和擷取新的變更。由於這種方式涉及大量的資料傳輸，因此完整擷取更適合處理規模較小的資料集。

2. 被動擷取（Partial Extraction- with update notification）：
被動擷取會在更動發生時自動發送通知，因此使用者只需擷取變更內容，而非整個資料集。另外，被動擷取也支援資料複製（Data Replication）功能，透過自動備份來確保資料的完整性。

被動擷取適合用於對資料變動極其敏感的業務環境或系統，例如：銀行可以透過被動擷取即時偵測並預防信用卡盜刷事件。

3. 主動擷取（Partial Extraction- without update notification）：
主動擷取會追蹤上次擷取時的資料變更。當系統無法自動發送通知時，主動擷取會指定時間內的資料變更，並將其檢索。

檢測變更的方式有很多，常見的做法是追蹤最後更新時間戳欄位，或在來源系統中建立變更表。系統會依據預設時程（如每週或每月）定期掃描和檢查資料變化，讓使用者只需擷取變更內容。

轉換（Transform）

轉換是資料整合的關鍵步驟，資料會在暫存區（Staging Area）中進行清洗、規則轉換和驗證等作業，以確保資料結構一致。

通常資料會在轉換階段經歷以下 5 個過程：
1. 標準化：標準化的目的是將資料格式進行統一，以協助分析，例如將 NULL 值定義為 0 或將「Male」或「Female」轉換成 M 和 F，並確保日期格式一致。
2. 資料過濾：資料過濾又稱資料清洗，目的是透過選擇特定的行與列來縮小資料集大小，從而提升資料處理效率。
3. 刪除重複資料：刪除重複資料的目的是識別並移除任何重複資料以確保準確性和一致。
4. 格式修訂：修訂資料格式的目的是根據企業的規範和標準來調整資料格式（如測量單位的轉換、日期和時間格式的轉換、字符集的轉換），讓資料符合法規或企業的標準格式。
5. 驗證：驗證的目的是檢查資料的完整性。在這個步驟中，系統會識別並標記異常資料。

載入（Load）

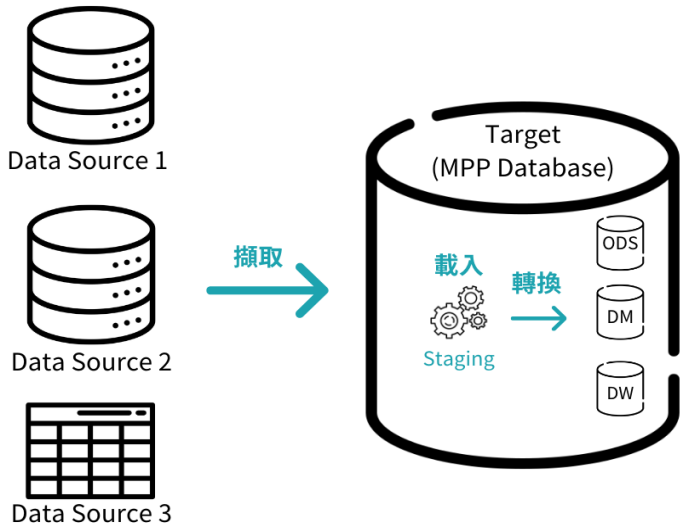
載入是 ETL 流程中的最後一步，目的是將資料存入目的端資料倉儲。由於資料倉儲在載入階段需處理龐大的資料集，因此確保資料架構的穩定非常重要。另外，根據資料性質和業務需求制定合適的策略也能提升搬遷效率，以下為兩種常見的資料載入策略：

1. 完整載入（Full Refresh）：
完整載入會載入來源端所有的資料。由於資料在進入倉儲前，目的端資料表會進行重置，因此這種方式也被稱為破壞性載入。雖然完整載入的做法直覺簡單，但在處理大規模資料集時會消耗較多時間和資源。

2. 增量載入（Incremental Load）：
這種策略只會在檢測到新的資料變更時建立錄。與完整載入相比，增量載入更容易管理。然而，如果系統發生故障就會導致資料不一致。

ELT 是什麼 ？

在 ELT 模式中，ETL 伺服器的作用變得相對次要，甚至在很多情況中僅扮演數據接口的角色。因為此模式主要依賴目的端資料庫的強大算力 (如 MPP 資料庫) 來進行資料轉換的工作，所以 ETL 伺服器（或 ETL 程序）在這個架構中，通常只負責資料的輸入 / 輸出，如下圖所示：



在上圖中，來源端資料，不論是異質資料庫、文字檔案、非結構化資料，都需經過落地 (Landing) 程序寫入目的端資料庫，如果以資料倉儲的概念來說，就是落地到暫存區。

而落地程序的方式有很多種，可以透過 ETL 軟體完成，也能透過由各種程式語言撰寫的程式，有許多目的端資料庫（如 Greenplum）甚至能利用介面（如 Platform Extension Framework, PXF）直接讀取外部資料，並在解讀完來源端資料後，以類似「Insert into」的語法直接寫入目的端資料庫。

而資料在落地後會依轉換的規則，以資料庫預存程式 (Stored Procedure) 的方式對資料做處理與轉換，最後寫入 ODS 或是 DW/DM 中。

該模式的特性如下：
1. 目的端為 MPP 資料庫，搭配高效的硬體，形成強大且可以橫向擴充（Scale-out）的架構。
2. ETL 軟體的影響力將減弱，在某些情況下它甚至只負責處理落地程序，如果來源端為 JSON、XML，或甚至是大型主機格式時，ETL 軟體就只負責原始資料的解譯，其他更複雜的處理和轉換任務則交由別的工具完成。
3. 如上所述，落地方式有很多種，它可以是 ETL 軟體、客製化程式，或目的端資料庫本身就具備解譯功能。

ETL 與 ELT 的比較

ETL 與 ELT 在資料處理策略上存在兩個顯著的區別：資料的轉換階段和使用場景。

1. 資料轉換階段

在 ETL 流程中，資料會先轉換再儲存，這種方式雖然能確保資料一致，卻會延長資料擷取時間，並限制原始資料後續被重新查詢和分析的可能。

相較之下，ELT 能將未經處理的原始資料存入資料倉儲，利用倉儲的運算能力進行清洗和轉換。由於原始資料未進行轉換，因此能在未來被重複使用和分析。

2. 使用場景

由於 ELT 能儲存未經預處理的資料，它非常適合處理大量非結構化和半結構化資料，如圖片、影片和 PDF 文件等。

相較之下，ETL 則適合處理複雜的資料轉換，例如資料遷移或讓交易密集型資料符合安控規範。由於 ETL 能夠在資料上傳到伺服器前完成所有轉換工作，因此它相較 ELT 更能確保資料的安全性。

ETL 與 ELT 的比對

比較項目	ETL	ELT
定義	資料從來源取出，在伺服器上進行轉換，最後載入資料倉儲	資料從來源取出，載入資料倉儲，並在倉儲內部進行轉換
擷取（Extract）	利用 ETL 工具	利用 ETL 工具或 DB 工具
轉換（Transform）	異質資料在伺服器上進行轉換	異質資料在倉儲內進行轉換
載入（Load）	資料經過轉換後才會載入資料倉儲	資料被擷取後直接載入資料倉儲
資料存取速度	較久，資料在伺服器上進行轉換，才會載入倉儲	較快，資料直接在倉儲內部進行轉換
維運成本	高，需要花更多成本維護用於執行資料轉換的伺服器	低，參與的系統相較 ETL 更少，因此維運成本更低
輸出資料	結構化資料、半結構化、非結構化資料	結構化、半結構化、非結構化資料
適用情境	資料規模較小且需要進行複雜的轉換程序	資料規模大並且注重資料處理效率

ETL 常見的使用案例

1. 資料處理

開發者能透過 ETL 執行多種資料處理管理任務，並與各種技術和工具相互整合，以下是 2 幾個常見的 ETL 應用場景：

1.1 資料系統整合

ETL 工具能將營運所需的重要資料（如姓名、地點和售價）與交易型資料（如銷售數據、提款和存款資訊、保險理賠資訊）相互整合，同時幫助企業將地端倉儲遷移至雲端。

另外，ETL 工具的即時資料處理能力，能幫助企業追蹤和分析即時資料流，從而提升決策效率。ETL 也能整合不同資料系統，例如在兩間公司的合併過程中，ETL 可以協助將不同資料系統整合至統一的資料倉儲。

1.2 利用 ETL 打造客製化服務

企業能夠擷取顧客的瀏覽紀錄、購買行為和網站點擊紀錄，並運用機器學習分析顧客的行為模式，再將分析結果整合到個人化推薦系統中，從而提供更符合顧客需求的客製化服務。

2. 自助式資料分析

2.1 利用 ETL 加速資料分析流程

過往，商業智慧模型需要依賴 IT 部門建立，且所有問題都只能透過靜態報告來解答，也就是說，每當有人對報告提出疑問時，IT 團隊就得重啟流程（從頭開始搜集、清理和分析數據）來進行解答，但這種方式不利於提高決策效率。

而 ETL 的自助式資料分析能簡化商業決策過程，雖然 IT 部門仍是控管資料存取權的核心角色，但現在各層級的使用者都能夠依據自身需求進行資料分析，從而加速決策進程。

2.2 利用 ETL 提升資料品質

ETL 工具能夠提升資料整合效率與資料品質的穩定性，這些工具還能標準化和自動化 ETL 流程，從而減少垃圾數據的產生。另外，ETL 工具也能整合有資料映射和血緣追蹤（Data Lineage）功能的管理工具。

2.3 利用 ETL 收集元資料（Metadata）

元資料是追蹤資料血緣的關鍵，因此掌握它的來源非常重要。當 ETL 工具從來源擷取資料並將其載入目標系統（雲端資料倉儲或資料湖泊) 時，它也會同時收集運行 BI 模型所需的元資料。

這樣的好處是開發者無需重新建模，因為元資料已被儲存於資源庫（Repository）以便後續進行調校、查詢和檢索。

3. IoT 資料整合

整合 IoT 設備資訊，作為資料分析的基石

整合 IoT 設備資訊對資料驅動的業務環境至關重要。ETL 工具能將 IoT 設備回傳的資料和帶有資料特徵（例如：時間戳和位置）的元數據一併整合和儲存，讓資料使用者對這些預整合的資料進行分析和調校。透過這些資料，企業也能對設備進行即時監控、預防性維護並優化營運策略。

ETL 在不同產業中的應用場景

1. 金融業的應用場景

金融機構需要透過收集大量結構化與非結構化資料來深入分析消費者行為，並將分析結果用於風險控管、優化金融服務和改善線上服務平台。

利用 ETL 工具，金融機構能夠選擇多種擷取模式，例如完整擷取、被動擷取和主動擷取，來收集多元異質資料，並透過修改資料缺失（Incomplete/Missing data）、雜訊（Noise）和離異值（Outliner），以確保資料的準確性。

2. 零售產業的應用場景

顧客會透過多元的渠道與品牌進行互動和交易，由於顧客在每個渠道中的行為模式都有所不同，且這些數據分散於多個平台和系統，企業很難獲得全方位的客戶視圖。

透過 ETL 收集和整合來自電子商務、社群、網站以及應用程式的資料有助於建立完整的資料脈絡，幫助提供個人化服務，並提高轉換率。

3. 機器學習與 AI 領域的應用場景

許多企業已開始探索如何將機器學習和 AI 應用於商業領域。由於機器學習和 AI 都需要大型資料倉儲來建立、分析並運行模型，因此選擇運算資源豐富的雲端架構是最合適的解決方案，透過 ETL 工具遷移並將數據轉換為可分析的資料有利於機器學習和 AI 的發展。

3 個 ETL 在不同職能的應用場景

具備 ETL 能力的職能必須對要操作的資料架構及內容有良好的掌握，才能透過 ETL 程序組合並取得相應的資料表及資料集，下方將列舉 3 種需要經常使用 ETL 的職場人員。

1. 資料科學家

資料科學家會針對特定主題進行分析，準備並梳理樣本資料，再形成資料模型，最後以直覺的方式呈現。同時，他們會在此過程中設計相應的 ETL 程序，並反覆地確認資料內容以及模型是否適合，再透過視覺化軟體呈現分析結果。

資料科學家往往需依賴功能兼具的平台以執行上述作業流程，因此最理想的工具必須同時具備 ETL、BI 和 ML 的功能。而 Pyramid Analytics 和 SAS Enterprise Guide 均能完成上述提及的所有工作。

2. 資料庫管理員（DBA）

DBA 人員在 ETL 程序中著重於擷取（Extract）和載入（Load）部分，首先，DBA 人員會從來源端擷取資料，並於該過程中執行備份/還原、根據資料溫度執行分流以及對資料進行清理，以確保其架構一致，再將資料載入目標資料倉儲。

DBA 人員能確保 ETL 程序能高效順暢地運行，並讓載入資料滿足報表製作、商業智能和資料分析需求。

3. 應用系統管理人員（AP）

應用系統管理人員(AP)的工作範疇包括執行資料流程驗證、資料分享(Data consuming) 和資料導入的前置作業。他們會識別擷取數據的品質，檢查數據是否有缺失、錯誤或與業務邏輯不符的情況，再來他們必須確保資料能在不同系統中安全地傳輸，以維護資料的隱私性，並控制使用者的存取權限。

AP 人員在 ETL 程序中負責維持資料的完整性、安全性以及持續監測和優化 ETL 中的轉換階段，讓數據持續適應多變的業務需求和環境。

導入 ETL 的 3 大挑戰

企業的資料量會隨時間推移變得更龐大和多元，這會增加整合異質資料、維護資料品質以及控制成本的難度。因為不同 ETL 工具適合處理的問題不盡相同，在選擇 ETL 工具前，必須將可能面臨的狀況納入參考，以下介紹企業常面臨的 3 大 ETL 挑戰。

挑戰 1 - 符合經濟效益

ETL 工具會定期更新資料流來確保資料倉儲維持最新狀態。然而隨著時間推移，當系統累積的技術債太多且沒有適時消化，就容易導致系統出現問題，這時就需要對 ETL 流程進行擴展來應對增加的工作量，例如：升級伺服器或倉儲的效能，但是升級硬體資源代表企業必須因資料遷移、停機時間而付出巨大的成本與時間，因此了解每個環節的詳細計價並提前進行成本分析，對於選擇合適的架構設計非常重要。

挑戰 2 - 提高資料品質

另一項挑戰是確保轉換後的資料維持完整性和正確性，因為在 ETL 流程中，手動寫入程式碼、執行變更以及在導入前未進行適當的測試和規劃容易造成錯誤（如重複載入和資料遺失）發生。
透過 ETL 的自動化流程能夠減少手動寫入程式碼時造成的錯誤，使用資料準確性測試(Data Accuracy Testing) 也能找出不一致和重複的資料。另外，ETL 工具的監測功能能幫助識別資料類型不相容的情況。

挑戰 3 - 全面的資料整合

由於資料的數量和複雜度都在不斷地增加，IT 部門需要整合上百種不同型態的資料。這些資料源包括結構化資料、半結構化資料、即時資料流、 資料檔、CSV、S3 儲存服務、串流資料等等，有些資料適合透過批次處理，其他則適合以串流處理進行轉換。

ETL 架構轉型

根據 IDC 的調查，2025 年全世界將產生 175 ZB 的新資料，IoT 設備也將在 2025 年前達到 557 億台，並產生 73.1 ZB 的資料量。

上段提及的 ETL 挑戰也將隨著急遽增加的資料量變得更加棘手，因此傳統的地端架構和企業營運模式必須做出相應的轉型。基於上述情境，我們推測 ETL 的未來將有以下 3 種可能的趨勢。

1. 立基雲端的基礎架構

面對指數成長的資料，傳統地端資料倉儲將因固定的硬體資源及僵固的架構而無法有效處理龐大資料，與地端環境相比，雲端架構可以帶來更多優勢，例如：利用雲端服務來執行異質資料的整合，從而消除硬體資源的維護成本。

另外，雲端架構能動態分配更多資源，如 CPU、記憶體和儲存空間 ，因此當資料量增加時，雲端架構能更靈活地擴展以滿足運算需求。

立基雲端的 ETL 架構對雲原生資料源及資料後續的應用效益巨大，同時還具有存取地端資料源的優勢。由於受限於本地法規要求，金融產業對於資安的要求又非常嚴謹，因此對於適法性的需求與雲端架構的靈活性，仍是雲架構要適應及調整的課題。

另外，當涉及到雲地整合的情況，網路頻寬及網路的穩定性也會是關鍵考量因素，因為這會直接影響資料的傳輸效率和可靠性。

2. 資料大眾化（Democratization of Data）

未來，資料分析將不再限於資料專家的工作範疇，各個企業部門都需要採用資料導向(Data Driven) 的決策模式。也就是說，企業必須將資料集中化、流程自動化，並針對不同部門的需求部署相應的 ETL 工具。

IT 部門可能會採用完整的資料轉換工具，而營運單位可能會採用資料流（Data Pipeline）工具，並依需求使用批次或串流功能。總而言之，如果企業能更自主地透過資料分析獲得洞見，競爭力將會顯著提升。

3. 提升串流資料與即時資料的效能

隨著資料增長，企業的營運模式和 IT 架構也會變得更加複雜。即時資料分析可以幫助企業發現營運瓶頸、監控系統效能以及調整資源分配，從而優化工作流程。

然而，批次處理模式因其延遲性無法滿足即時資料分析的需求。透過整合 ETL 工具、 CDC（異動資料擷取）和事件導向軟體架構(Event-Driven Architecture) 有助於提升串流資料與即時資料的效能。

歐立威的 ETL 解決方案

歐立威科技針對不同使用情境及需求提供 3 種 ETL 解決方案：批次處理模式、近即時處理模式和即時處理模式。

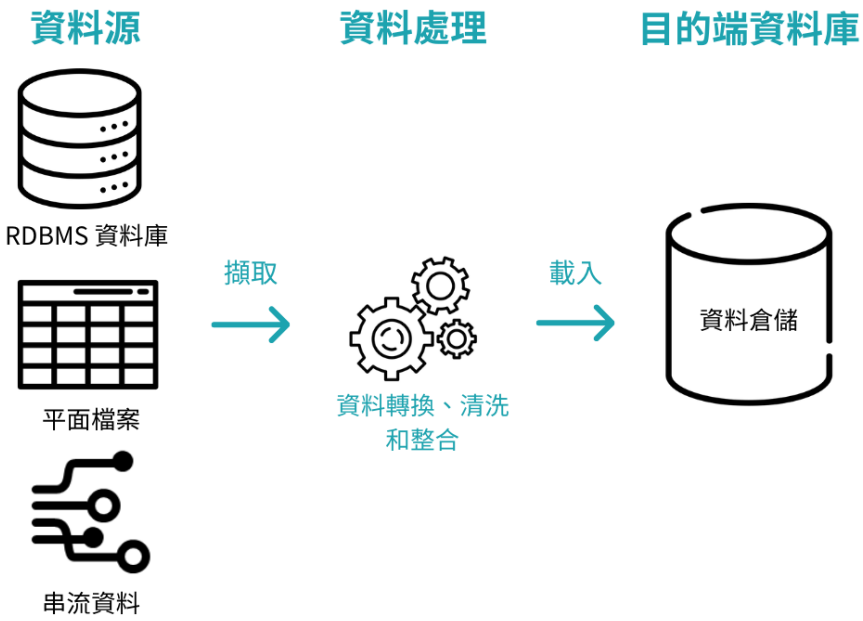
以下介紹 3 種解決方案的運作原理，讓使用者更準確地評估自身需求，唯有選擇合適的處理模式，才能達到工作流程自動化，減少手動錯誤並提高資料品質的成效。

批次處理模式

批次處理模式的原理是將大量資料分段處理，這種模式經常被用於執行定期或計畫性的任務，例如：定期資料輸入/輸出、定期報告生成等，當所有資料完成轉換後，資料會被全部存放於資料倉儲或暫存區。

Omni waresoft 歐立威科技
全方位開源軟體解決方案建置商
Tel: (02) 2558-2656
mail: emarketing@omniwaresoft.com.tw

相較其他模式，批次處理的資料流（Data Pipeline）運行頻率較低、持續時間較短，而且經常於低流量的離峰時段運行。正因為不要求即時性，所以通常會在離線環境中執行，從而降低營運成本，另外批次處理流程中的資料清理和驗證，能夠確保資料具有較高的準確性。

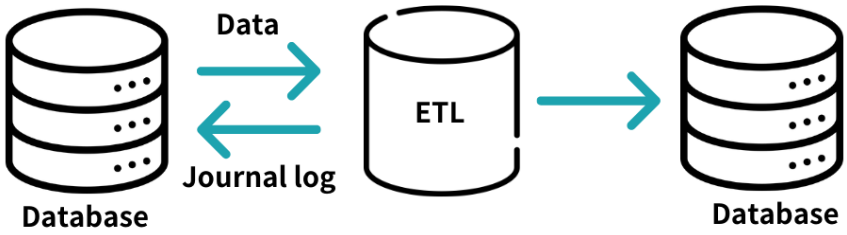


近即時處理模式

近即時處理模式又名為「低延遲」處理，目的是在分鐘內完成資料轉換。該模式會先識別來源端（大多數情況下是資料庫）的資料變動，接著它會即時地將變更更新到目的端資料庫；最後，系統會將這些變更項目進行歸檔，並清晰地標示出上次的資料更動。

這種處理模式不僅能即時處理資料，也能同時提高資料管理透明度。近即時處理模式能夠精準地在分鐘內獲得資料的最新狀態，也同時增加了資料處理流程的可追溯性。

這種模式通常會透過 ETL 軟體中的 CDC（Change Data Capture）元件並搭配來源端資料庫啟用的 Journal log 來執行以上工作，示意圖如下：



即時處理模式

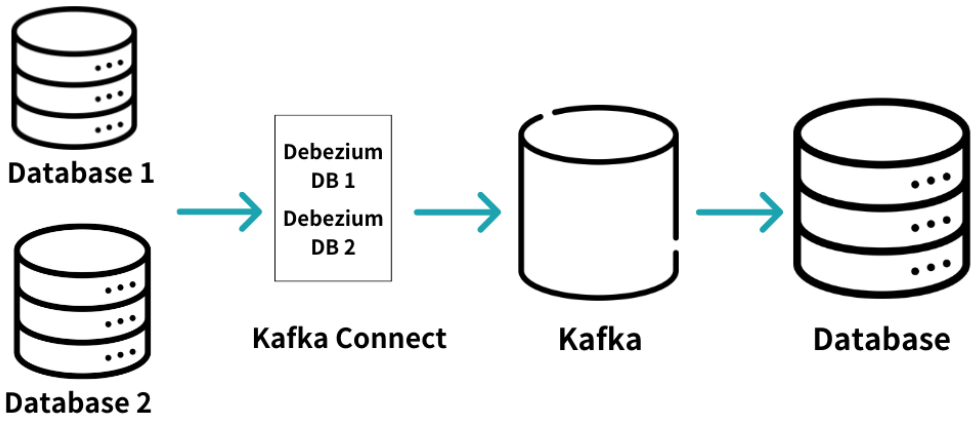
即時處理模式被用於處理連續不間斷的資料流，這種模式能接收來源端持續發送的大量資料，並利用資料倉儲強大的運算能力即時對資料進行處理和分析。

由於這種模式被用於處理動態且不斷更新的資料結構，它具有出色的處理速度和良好的容錯機制，這確保即使在資料部分丟失或傳輸順序錯亂的情境中，仍能被精確地處理。
另外，為了保持即時處理效益，這種模式需要穩定和低延遲的網路連通性，但對硬體的運算能力則沒那麼要求。

隨著 IoT 設備的普及，串流資料量逐年攀升，許多資料分析師或是資料科學家，都會透過串流資料運行模型或執行大數據分析。

在這種情況下，配合串流平台 (如 Kafka、RabbitMQ 等) 以及資料庫 (SQL 或是 NoSQL) 的 CDC 工具就應運而生。

這類型的 CDC 工具以 Kafka 的 Debezium 套件為代表，搭配 Flink 作為寫入資料庫的接口，達成串流 CDC 的目的，架構大致如下：



這種模式的即時處理效能高，只要硬體資源足夠，即時處理模式可以順暢地處理串流資料，並將其寫入目的端資料庫。

然而，當資料邏輯複雜，單靠硬體效能又無法達到即時性時，就需要透過完整的產品生態系來達成。因此，下方將介紹歐立威的即時處理 ETL 產品生態系。

歐立威科技的即時處理產品生態系

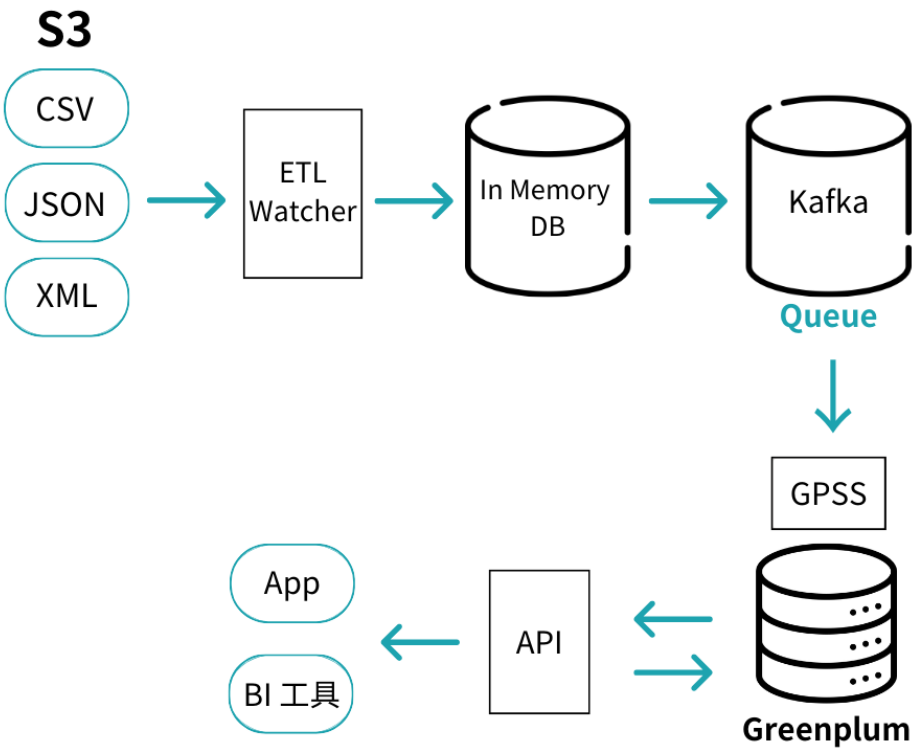
歐立威科技的解決方案整合了批次處理和即時處理的優勢，透過完整的產品生態系讓資料同時兼具高度準確性和即時性。

我們的處理模式是即時資料處理的延伸版本，其特性如下：

1. 來源端資料定期/不定期產生，可能是 XML/Json/CSV 等格式的文字檔案。
2. 資料需要進一步計算處理，故需要記憶體資料庫作暫存，以加快速度。
3. ETL 程式負責運算，其運算可以透過以下種方式執行：
   - 透過 ETL Server 中的專用程式或元件來進行
   - 利用 ETL 程式呼叫撰寫的程式 (如 Java 或 Python) 以執行特定任務
   - ETL 程式負責整體的處理流程，但具體的運算則交由記憶體資料庫的函式庫來完成
4. 已處理完畢的資料會按照預先設計好的模式寫入串流佇列中，並等待後續的操作，如插入、更新或刪除 (Insert/Update/Delete) 。
5. 利用資料庫專門用於接收串流資料 (Streaming data) 的接口，將資料依照預設的模式接收、整合並將其存入資料庫中。

此模式整合雲端的解決方案，多半取決於 ETL 對雲端資料源的支援程度，如果支援的接口種類齊全，則對雲端支援就相對完整。

即時資料處理流程的示意圖如下： ‘



此模式已經不僅限於 ETL 或是 ELT 的處理，而是透過整合整個資料產品生態系來達成串流資料的解譯、資料邏輯處理、即時寫入/更新大數據資料庫，並提供一套系統和流程來支援後端的資料服務，因此這個模式特別適合處理和分享巨量資料。

不過要在上述架構中達到一定的即時性，硬體上的搭配也非常重要，因此只有兼顧生態系中的軟硬體效能才能體現整個生態系統的優勢。

設計資料流需考慮的 3 大面向

為了滿足特定的使用目的，資料流的設計應根據使用者的需求進行客製。在建置資料流前，使用者必需選擇合適的架構並評估流程中的產品組合，因為這是影響效能的關鍵因素，對此，我們將在此段探討如何根據使用目的挑選合適的資料流，以及如何配置 ETL 工具來達到最佳化效能。

面向 3 - 如何選擇資料架構

1. 資料的使用目的

瞭解自身分析需求是建立任何資料流的基石，不同流程和架構皆有其獨特的使用情境，因此唯有選擇與使用與目的相符的資料流才能事半功倍。

如果企業主要目標為長期的趨勢分析，資料即時性就並非首要考量，批次模式在這種情況下更為合適；然而，若需求聚焦於作業流程的即時監控或系統架構的讀寫分離（例如大型主機負責寫入，而開放式主機則進行讀取），那近即時或即時資料處理模式就應該被優先考慮。

2. 預算

資料的即時需求和成本有直接的關聯，若追求越高的即時性，資料流的建置成本也會隨之上升。如果資料量達到大數據等級，例如每天有 TB 級的資料需處理，且這些資料必須被即時分析，除了選擇合適的 ETL 工具外，也需要選擇記憶體內資料庫（In-Memory）或 MPP 資料庫。而在實際操作中，可能還需採用叢集（Cluster） 架構以支援平行或分流運算。

在硬體配置方面，除了要有算力強大的 CPU 外，高規格記憶體和儲存設備也是必要條件。例如，為了追求更快速的讀寫效能，需要採用 NVMe SSD 這類先進的儲存裝置。同時，為了保證資料流在各元件間暢通無阻，還需要光纖級的高速網路傳輸能力。

3. 資料源的複雜度

資料轉換的複雜度以及資料庫供應商種類的多寡，是抉擇 ETL 或 ELT 模式的重要考量。假設來源資料的種類眾多，但轉換邏輯簡單，採用 ETL 模式就會較為合適。

如果轉換邏輯複雜，目的端資料庫就必須具備強大的運算能力，如果其運算能力不足，只有儲存功能，那麼採用 ETL 模式並搭配 ETL 軟硬體會是較好的選擇；然而，若目的端資料庫運算能力出色，DBA 團隊也夠厲害，並有完善的程式撰寫規範，選擇 ELT 模式則更為合適；但如果希望轉換效能不受撰寫程式人員的能力影響，選擇 ETL 工具則更為明智。

4. 現存環境的限制

資料架構是選擇資料流時的重要考量因素。多數公司的核心系統通常採用傳統大型主機系統，儘管這些主機支援電文交換的溝通模式，但在實際操作中，交換的資料量通常有限，而且就算有技術上的解決方案，出於維運考量，主機的管理者通常也不樂意讓外界直接介接擷取資料，因為這種做法會佔用主機資源。目前最常見的方式是定期將文字檔案上傳到特定檔案夾供外界擷取，在這種情況下採用批次模式會更為合適。

5. 其他考量

如果企業已建置元資料管理系統 (Metadata management, MDM)，採用大品牌的 ETL 工具會較為理想，因為 MDM 能讀取 ETL 軟體和 BI 軟體的資料庫，從而有效建立資料流血緣分析 (Lineage analysis)。

透過血緣分析，系統能更容易評估變更帶來的影響，即衝擊分析（Impact analysis）。由於資料轉換大多發生在程式中，如果採用 ELT 模式，而 MDM 軟體對解譯程式語法上的能力又存在差異，那會對 MDM`系統分析資料血緣的精準度造成影響。

此外，資料流中的作業流控制系統（Job Control System, JCS）也是一大考量要素，尚若採用強大的 JCS 系統來執行、監控、記錄各種作業的執行情況，對 ETL 軟體內建排程的需求就將大幅降低，甚至可以直接使用 ELT 模式，讓 JCS 全面控制所有執行流程。

上方列舉的面向能幫助使用者在選擇資料流時提供清晰的指引。接下來我們將探討在建構產品組合時需注重的 5 個要素。

面向 2 - 評估 ETL 產品/產品組合要素

評估 ETL 產品的適配性及產品組合，可以從這 5 個要點著手：

1. 對 SLA (Service Level Agreement) 的要求：SLA 是指建立 ETL 程序後，對資料更新頻率的需求。簡單來說，要求越高的即時性，建置成本（TCO, Total Cost of Ownership）也會相對增加。
2. 接管維運的能力：如果維運單位的 IT 人手不足，就會建議以架構單純且採用 ETL 工具為主的設計，以免造成維運上的障礙。
3. 產品的延續性：在挑選 ETL 產品時，除了考慮穩定性，也應評估其未來的更新能力。

4. 來源系統的介接方便與否：如果資料來源中有特殊的資料源 (如大型主機等)，又需要直接介接的話，就需要選擇特別的 ETL 工具，或是考量以文字檔拋接的方式介接。

5. 企業政策：選擇 ETL 工具時，也需考慮企業的內部政策和規範，例如與特定產品簽訂全球採購契約，或是有內部規範規定廠牌，都是左右 ETL 工具搭配的因素。

面向 3 - 雲端與地端的架構選擇

在規劃雲地整合方案過程中，除了考慮上述提及的雲環境要素外，選擇適合的雲端模式也至關重要，假如企業採用純雲端架構，它們必須在公有雲和私有雲間做出抉擇，若選擇公有雲方案，考量的面向則涉及是否透過單一供應商以獲取專屬服務，或採納多雲策略，透過結合多家供應商的服務來優化性能和提升成本效益，同時達到風險分散的效果。

以上各種設計該如何取決，與前述提及關於雲端資料源，或雲原生 ETL 系統存取地端資料源的能力息息相關，雖然越複雜的架構設計可能帶來更多的運作上彈性，但管理難度也會相應提升。

打造你的現代化資料流！

本攻略概要介紹 ETL 的基本概念與流程，然而在實際建置 ETL 架構時，企業往往會面臨專業人才短缺、流程複雜性高和操作困難等挑戰。

歐立威科技長期致力於最新的 ETL 技術，協助客戶解決任何 ETL 流程與建構資料流可能遭遇的疑難雜症。如果你有任何問題，歡迎與我們諮詢，讓我們一起打造現代化資料流！

Omniwaresoft 歐立威科技 全方位開源軟體解決方案建置商 | Tel: (02) 2558-2656 | mail: emarketing@omniwaresoft.com.tw