IBM DataStage

透過 IBM Cloud Pak for Data 的 DataStage
即時交付業務就緒的AI資料

經由資料整合交付業務就緒的資料

今日的數位企業在建立和使用資料的方式和以往大有不同。包括儲存在多種系統和資料庫有關於客戶的、交易的和員工的數據。這些資料所儲存之處遍及各種多雲、混合雲和資料湖等環境，因此各個組織正尋求可將這些分散來源和環境整合起來的方式，**利用人工智慧**獲取更快的洞察，以為其客戶傳遞有差異性、個性化的體驗。根據 Forrester 的研究，資料科學家花費 80% 的時間在準備和管理**人工智慧專案所需的資料**上。這些研究結果連同 IBM 的調查（91% 的企業組織並非有效使用資料）可說明：企業正苦於如何能從資料孤島中找出價值。這些用來從龐大資料量中即時存取資料並交付業務就緒資料的架構技術、實踐和工具稱為資料整合。企業運用靈活且可擴充的資料整合技術，透過在多種資料來源上擷取、轉換和載入 (ETL) 資料的方式，為下一個最佳產品、流失偵測與分析、供應鏈預測，和執行即時詐欺偵測等業務場景進行分析。

針對那些苦思在多雲或資料湖之間管理資料，又想縮短時間以建立和更新AI 模型與應用程式的高層主管、企業架構師或營運領導人，IBM® InfoSphere™ DataStage 是一套領先市場的資料整合解決方案，其功能超越 ETL而能交付可信賴的業務就緒資料功能，提供可擴充的多雲資料整合，並確保可信的業務就緒資訊能即時派上用場。DataStage 關鍵功能包括多雲執行時期支援，其採用一經設計即可在任何雲端上運行的技術，可在以自動工作負載平衡和低延遲平行引擎運作的同時，擴充工作負載。此外，也透過內建複製技術進行即時的資料交付，為開發人員支援持續整合及持續交付 (CI/CD) 以節省時間和成本；以及利用「自主整合設計」快速建立AI模型；利用內建資料品質驗證規則來自動偵測及解決資料問題。

DataStage 是 IBM DataOps 的其中一項功能，將持續不斷的高品質資料執行管理化以啟用人工智慧，並可從任何資料來源在正確的時間為正確的人提供自動化自助式資料處理。IBM InfoSphere DataStage 可在本地、IBM Cloud 和可部署於任何環境的 IBM® Cloud Pak™ for Data 等超融合平台上使用。IBM® Cloud Pak™ for Data 是完全整合的資料和 AI 平台，建立於 Red Hat® OpenShift®之上，提供 DataStage 完全雲原生架構，可隨您的企業擴充。其也為組織提供一個可以支援多資料交付樣式的平台，包括資料整合、資料複製和資料虛擬化，同時變更資料擷取 (CDC) 也能使用以 Kafka 為主的訊息列，在產生變更和交付資訊到雲端上目標資料庫和資料湖時，擷取到以紀錄為主的變更。

設計一次，在任何雲端上執行

根據 IDC 的一項研究，有 90% 的企業客戶都在使用多雲。多雲資料整合技術可以協助使用者將設計從執行時期分開，只要設計 ETL 工作一次後，即可在任何雲端環境上透過容器部署執行時期元件，以降低因處理大量資料而產生的延遲性。您可在本地建立和測試一項工作，然後在如 Microsoft Azure 的雲端環境中執行，以充分運用雲端上的 Azure 資料湖。工作參數和數值會透過 Kafka 訊息傳輸到 DataStage 的遠端實例。

多雲資料整合提供下列效益：

- 在本地和雲端環境之間整合資料
- 自動化工作設計體驗以簡化設計流程
- 執行遠端工作以將移出資料的成本降至最低
- 達成地域性政治要求
- 減少處理大量資料集所產生的延遲性，因可保留資料在原地不需移動。

工作負載平衡自動化和平行處理

有了一個完整的雲原生架構後，您可透過一個功能優異的平行引擎 (PX)，使用本地容器或 DataStage 的分享容器，動態擴充工作負載並優化龐大的資料集。使用者可選擇在 IBM DataStage Flow Designer 中建立平行、序列或 Apache Spark 工作。

您可在兩個執行時期引擎上進行 DataStage Flow Designer 工作：

- 類型為平行或序列型的工作僅可在平行引擎上執行。一般來說，資源需求高的工作會在平行引擎上執行，如此一來，使用平行處理完成複雜工作的平均時間為兩分鐘。
- 工作類型為 Spark 類型的工作可在 Spark 引擎上執行。

即時資料交付

DataStage 完整搭載可即時擷取的變更資料擷取 (CDC) 技術，若部署於容器上則可同時為資料整合和資料複製提供最佳效能。DataStage 可在 CDC 擷取日誌變更時允許大資料集的複雜轉換，利用複雜的轉換技術，並使用以 Kafka 為基礎的訊息佇列，將資料集交付到雲端的目標資料庫和資料湖中。DataStage 也允許依批次及依事件的大量資料轉換工作饋送到資料倉儲中。

透過 CI/CD 支援減少開發人員的時間和成本

為解決在不同的作業系統中管理容器化應用程式的挑戰，組織需要一個強大的開放原始碼工具，例如Cloud Pak for Data 上的 Red Hat OpenShift。Cloud Pak for Data 平台有助於擴充和配置容器，以支援關鍵IT 專案，例如微服務和雲端遷移策略。DataStage 容器允許為工作從開發到測試再到生產中的 CI/CD pipeline建立和自動化，並藉支援 GitHub 之類的來源控制工具來支援 CI/CD pipeline，以經常性發佈工作和正式上線。

可加速 AI的自動整合設計

藉自動探索和分類資產，並依據內建自訂轉換和品質規則建立整合流，以及偵測和保護敏感性資訊的方式，更快且更大規模地為AI 加速收集和整合資料。

快速評估：
透過自動工作設計

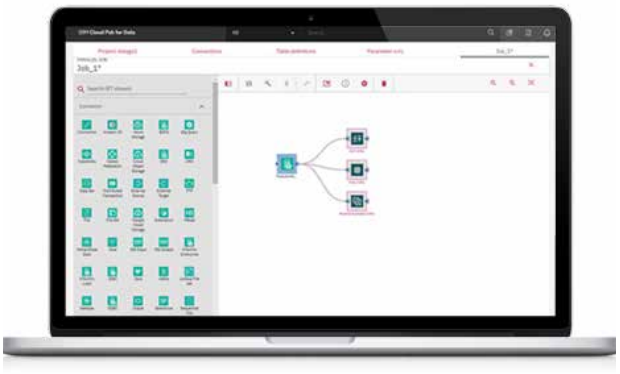

圖 1.含自動設計功能的 DataStage Flow Designer

IBM DataStage Flow Designer 是利用機器學習以網路為主提供給 DataStage 的 UI，可協助使用者（甚至是非技術人員）在工作中建立流程和階段。

DataStage Flow Designer 提供下列效益：

- 向下相容。無需轉移工作。許多公司在一個單一專案中就有成千上百的工作，他們依賴這些工作以維持 24 小時執行。因為在遷移過程中不允許有發生錯誤和運行中斷的可能性。這些公司可以將任何現有的 DataStage 工作在 IBM DataStage Flow Designer 中執行，因此不需將這些工作遷移到一個新位置。
- 增加開發人員生產力。IBM DataStage Flow Designer 功能包括內建搜尋、讓公司可立即上手的快速導覽、自動中繼資料傳播、智慧盤、建議階段和可同步顯示所有匯集錯誤的功能。開發人員可在使用這些功能設計工作的同時提高生產力，比傳統人工編碼的工作快上九倍。
- 操作器和連接性廣泛。除設計和開發能力外，DataStage 也提供數百種隨開即用的現成預建操作器。這些操作器可大幅降低開發人員花費在準備分析資料上的時間。隨每數週即新增的操作器，開發人員的生產力也同時獲得提高。

確保可信資料在交付時的使用品質和安全性。

DataStage 為資料整合提供單一使用者體驗，在資料對目標環境（例如資料湖）進行交付時使用執行資料驗證的 DataStage Flow Designer、標準化和匹配規則，以防止未經授權的使用者存取您的敏感性資料，而產生潛在性的品質和安全問題。資料品質這一個概念也可延伸到為資料倉儲 (DWH) 支援全面資料治理上。

總結

DataStage 提供：

- 設計一次、隨處執行：透過內建的自動化工作負載平衡、平行化和可擴充性來完成
- 可即時或依批次交付模式來擷取更新
- 內建復原力、易於操作且持續整合及持續交付
- 為人工智慧進行優化資料整合
- 使用機器學習功能自動進行工作設計
- 確保可信資料在交付時的使用品質和安全性

IBM 為混合多雲環境、本地或超融合系統（例如 IBM Cloud Pak for Data），或在任何所選的雲端平台上提供多種資料整合功能。這些不同的功能依照他們所選的部署模式，提供一套靈活且具擴充性的資料整合解決方案，使人工智慧可快速存取大量的高品質資料。

透過免費示範可了解更多有關
IBM InfoSphere DataStage的資訊

為何要選擇 IBM？
IBM DataOps 藉由提供領先市場的技術，搭配可運用人工智慧的自動

化、所融入的監管功能和強有力的知識目錄，在企業中為資料維持品質並可持續操作，協助建立一套企業可即用的分析基礎。增加資料品質，在正確的時間從任一來源為正確的人提供有效的自助式資料管路。

如欲瞭解更多關於 DataOps 的資訊，請造訪
ibm.com/dataops
如欲瞭解更多關於 IBM InfoSphere DataStage 的資訊，請造訪
ibm.com/products/infosphere-datastage
由下列連結造訪大數據和分析中心：
ibmbigdatahub.com

IBM Corporation
New Orchard Road, Armonk, NY 10504
Produced in the United States of America
April 2020