AWS 資料倉儲

2016 年 3 月

注意

本文資訊僅供參考，其內容為文件發佈日當時 AWS 的最新產品項目與實務方法，如有變更，恕不另行通知。客戶需自行獨立評估本文資訊，任何 AWS 產品或服務皆以「現狀」提供，不包含任何明示或暗示性保證。本文不提供任何來自 AWS、其附屬公司、供應商或授權人之任何保證、表示，契約承諾、條件或保證。AWS 對其客戶的責任與義務應由 AWS 協議管轄，本文並非 AWS 與其客戶之間的任何協議的一部分，也並非上述協議的修改。

目錄

摘要

全球各地企業的資料設計師、資料分析師及開發人員正準備將資料倉儲轉移至雲端，以提升效能並降低成本。本白皮書討論現代分析方法及資料倉儲架構，說明 Amazon Web Services (AWS) 提供的服務以建置此架構，並提供常見的設計模式以利用上述服務建立資料倉儲解決方案。

緒論

對於現代企業而言，資料與分析是不可或缺的。幾乎所有大型企業皆已建立資料倉儲以提供報告與分析，所利用的資料涵蓋各種來源，包括企業本身的交易處理系統以及其他資料庫。

但資料倉儲的建置與運作 (儲存來自一或多個資料來源的中央儲存庫) 一向是相當複雜且昂貴。大多數資料倉儲系統的建置皆非常複雜，前置軟體與硬體的支出成本高達數百萬美元，而且需要數個月的時間進行規劃、採購、建置及部署程序。在完成初期投資並建立資料倉儲之後，還必須僱用資料庫管理員工作團隊，以確保查詢的執行速度並保護資料以避免流失。

傳統資料倉儲也很難擴展。當資料量增加或想要讓更多使用者能使用分析與報告時，您必須忍受較慢的查詢速度，或投資時間與精力於昂貴的升級程序。事實上，有些 IT 團隊不鼓勵增加資料或新增查詢，以保護現有的服務水準協議。許多企業為了與傳統資料庫廠商維持友好的關係而陷入困境。這些企業通常會被迫升級受管系統的硬體或展開曠日費時的逾期授權談判。當他們的資料倉儲引擎達到擴展的限制時，將被迫轉移至相同廠商但不同 SQL 語法的其他引擎。

Amazon Redshift 改變了企業對於資料倉儲的想法，因為它大幅降低部署資料倉儲系統的成本與精力，並且不犧牲功能與效能。Amazon Redshift 是快速、完全管理的 PB 級資料倉儲解決方案，可讓您使用現有的商業情報 (BI) 工具，以具有成本效益的方式輕鬆分析大量的資料。使用 Amazon Redshift，您可以獲得直欄式資料倉儲引擎的效能，以十分之一的成本執行大量平行處理 (MPP)。您可以從每小時 0.25 USD 小量開始且無任何約束，然後擴展至數個 PB，每年每 PB 的費用為 1,000 USD。

自 2013 年 2 月推出以來，Amazon Redshift 已成為成長最快速的 AWS 服務之一，擁有數千個來自各種產業與企業規模的客戶。許多企業如 NTT DOCOMO、FINRA、Johnson & Johnson、Hearst、Amgen 及 NASDAQ 皆已轉移至 Amazon Redshift。因此，Amazon Redshift 已名列「*Forrester Wave：企業資料倉儲，2015 年第 4 季*」報告中的領導業者。[1]

在這份白皮書中，我們提供您必要的資訊，協助您在將資料倉儲從內部部署轉移到雲端的策略中獲益：

- 現代分析架構
- 在此架構中可選擇的資料倉儲技術
- 深入了解 Amazon Redshift 及其差異化功能
- 在 AWS 上以 Amazon Redshift 及其他服務建立完整資料倉儲系統的藍圖
- 有關從其他資料倉儲解決方案進行轉移，以及運用我們的合作夥伴生態系統的實用技巧

現代分析與資料倉儲架構

再次說明，*資料倉儲* 是儲存來自一或多個資料來源的中央儲存庫。資料一般會從交易系統及其他關聯式資料庫流入資料倉儲，通常包括結構化、半結構化及非結構化資料。這些資料的處理、轉換及獲取都以是規律的節奏進行。資料科學家、商業分析師、決策者等使用者，透過 BI 工具、SQL 用戶端及試算表存取資料。

為何要建立資料倉儲，為何不在記錄交易資料的線上交易處理 (OLTP) 資料庫上直接執行分析查詢？為了回答這個問題，讓我們看看資料倉儲與 OLTP 資料庫的不同之處。資料倉儲針對批次寫入操作與讀取大量資料進行最佳化，而 OLTP 資料庫則針對連續寫入操作與大量的小規模讀取操作進行最佳化。一般而言，由於資料傳輸量需求較高，資料倉儲通常採用去正規化結構描述，例如 Star 結構描述與 Snowflake 結構描述，而 OLTP 資料庫則採用高度正規化結構描述，其較適用於較高的交易傳輸量需求。Star 結構描述包含幾個大型事實資料表，這些事實資料表參照多個的維度資料表。Snowflake 結構描述是 Star 結構描述的延伸，包含已進一步正規化的維度資料表。

若要使用您的來源 OLTP 或其他來源系統，獲取利用資料倉儲做為獨立資料存放區進行管理的利益，建議您建立高效率的資料流程。上述流程可從來源系統擷取資料，並將資料轉換為適合資料倉儲的結構描述，然後載入資料倉儲。我們將在下一節討論分析流程的建構模塊，以及可用以建構流程的各種 AWS 服務。

分析架構

分析流程的設計可處理異質來源 (例如資料庫、應用程式及裝置) 傳入的大量資料串流。

典型的分析流程包含以下階段：

1. 收集資料。
2. 存放資料。
3. 處理資料。
4. 分析資料並將其視覺化。

如需圖示，請參閱以下圖 1。

圖 1：分析流程

資料收集

在資料收集階段，請考量您可能有不同類型的資料，例如交易資料、日誌資料、串流資料及物聯網 (IoT) 資料。AWS 為上述各種資料類型提供資料儲存體解決方案。

交易資料

交易資料，例如電子商務購買交易與金融交易，通常存放於關聯式資料庫管理系統 (RDBMS) 或 NoSQL 資料庫系統。資料庫解決方案的選擇需依據使用案例與應用程式特性而定。NoSQL 資料庫適用的情況為，資料的結構化不佳而無法配合已定義結構描述，或結構描述經常變更。另一方面，RDBMS 解決方案適用的情況為，交易跨越多個資料表列，而且查詢需要複雜聯結。Amazon DynamoDB 是完全受管的 NoSQL 資料庫服務，可做為您應用程式的 OLTP 存放區。Amazon RDS 可讓您為應用程式實作以 SQL 為基礎的關聯式資料庫解決方案。

日誌資料

可靠地擷取系統產生的日誌，有助於您使用存放於日誌中的資訊解決問題、進行稽核，以及執行分析。Amazon Simple Storage Service (Amazon S3) 是受歡迎的儲存解決方案，適用於可進行分析的非交易資料，例如日誌資料。Amazon S3 提供 11 個 9 的耐用性 (即耐用性 99.999999999%)，因此也是受歡迎的封存解決方案。

串流資料

Web 應用程式、行動裝置及許多軟體應用程式與服務會產生大量的串流資料，有時每小時高達數個 TB，而且必須持續收集、存放及處理這些資料。[2]利用 Amazon Kinesis 服務，即可以低成本輕鬆完成上述工作。

IoT (物聯網) 資料

遍布全球各地的裝置與感測器持續地傳送訊息。企業已看到持續成長的需求，才能擷取上述資料並從中獲取情報。利用 AWS IoT，連線裝置即可輕鬆安全地與 AWS 雲端進行互動。AWS IoT 可讓您更容易利用 AWS 服務，如 AWS Lambda、Amazon Kinesis、Amazon S3、Amazon Machine Learning 及 Amazon DynamoDB，來建置應用程式以便針對 IoT 資料加以收集、處理、分析及採取行動，而且無需管理任何基礎設施。

資料處理

收集程序可提供包含潛在有用資訊的資料。您可以分析擷取後的資訊以獲得情報，協助您的事業成長。例如，上述情報可能會提供您有關使用者行為以及您產品相對受歡迎程度的資訊。收集上述情報的最佳實務，是將原始資料載入至資料倉儲以執行進一步的分析。

有兩種處理工作流程類型可達到上述目標，分別是批次與即時。線上分析處理 (OLAP) 與 OLTP 是最常見的處理形式，它們各採用上述其中一種類型。線上分析處理 (OLAP) 通常採用批次方式。相對的，OLTP 系統則以即時處理為主，並且通常不太適合批次處理方式。如果您將資料處理與 OLTP 系統脫鉤，即可讓資料處理不影響您的 OLTP 工作負載。

首先，我們來看看批次處理涵蓋哪些程序。

擷取轉換負載 (ETL)

ETL 是將資料從多個來源取出以便載入資料倉儲系統的程序。ETL 通常是以定義完整的工作流程連續進行的程序。在此程序中，首先從一或多個來源擷取資料。然後將擷取的資料進行清理、強化、轉換，然後載入至資料倉儲。Hadoop 框架工具如 Apache Pig 與 Apache Hive 通常用於 ETL 流程以執行大量資料的轉換。

擷取負載轉換 (ELT)

ELT 是 ETL 的變形，其擷取的資料會先載入目標系統。資料載入資料倉儲之後，才執行轉換作業。如果您的目標系統足夠強大以處理轉換作業，則 ELT 通常可以順利運作。Amazon Redshift 在執行轉換方面效率很高，因此常用於 ELT 流程。

線上分析處理 (OLAP)

OLAP 系統以多維度結構描述存放累積的歷史資料。OLAP 系統廣為應用於資料採礦，可讓您擷取資料並在多維度上掌握趨勢。Amazon Redshift 已針對快速聯結進行最佳化，因此經常用於建置 OLAP 系統。

現在讓我們來看看資料即時處理涵蓋哪些程序。

即時處理

我們先前已討論串流資料，並提及以 Amazon Kinesis 做為擷取與存放串流資料的解決方案。您可以用逐筆記錄或在滑動時間間隔期間的頻率，循序並遞增處理資料，並將處理後的資料用於各種分析，包括關聯、彙總、篩選及取樣。此種處理類型稱為即時處理。透過即時處理獲得的資訊可提供企業檢視其商業與客戶活動的許多面向，例如服務使用情形 (可用於計量或收費)、伺服器活動、網站點擊次數，以及裝置、人員與實體商品的地理位置，並協助企業迅速因應新的情況。即時處理需要具備優異平行處理以及可擴展性的處理層。

您可以使用 AWS Lambda 即時處理串流資料。Lambda 可以從 AWS IoT 或 Amazon Kinesis Streams 直接處理資料。Lambda 可讓您執行程式碼，無需佈建或管理伺服器。

Amazon Kinesis Client Library (KCL) 是另一種處理 Amazon Kinesis Streams 資料的方式。KCL 提供比 AWS Lambda 更多的彈性，可讓您批次處理傳入的資料，以便進一步處理。您也可以使用 KCL，將各種轉換與自訂套用至您的處理邏輯。

Amazon Kinesis Firehose 是最容易將串流資料載入 AWS 的方法。它可擷取串流資料並自動載入 Amazon Redshift，可透過您目前使用的 BI 工具與儀表板提供接近即時的分析。您可以使用 Firehose 定義自己的批次規則，它將確實地批次處理資料並提供至 Amazon Redshift。

資料儲存體

您可以將資料存放於資料倉儲或資料市集，方法如下。

資料倉儲

如前所述，*資料倉儲* 是儲存來自一或多個資料來源的中央儲存庫。使用資料倉儲，您可以運用 BI 工具針對大量資料執行快速分析，並發掘隱藏在資料中的模式。資料科學家查詢資料倉儲以執行離線分析並發現趨勢。組織內部的使用者可透過特定 SQL 查詢、定期報告及儀表板來使用這些資料，以進行關鍵商業決策。

資料市集

資料市集 是簡易形式的資料倉儲，聚焦於特定的功能領域或主題。例如，您可以為組織內部的各個部門提供專屬資料市集，或依據地區提供區段資料市集。您可以從大型資料倉儲、運作存放區，或混合上述兩者，以建立資料市集。資料市集很容易設計、建置及管理。但是，由於資料市集聚焦於特定功能領域，跨功能領域的查詢會因為資料分佈四處而變得複雜。

除了建置資料倉儲之外，您也可以使用 Amazon Redshift 建置資料市集。

分析與視覺化

處理資料並使資料可供進一步分析之後，您需要適當的工具針對處理過的資料進行分析與視覺化。

在許多案例中，您可以使用與處理資料相同的工具執行資料分析。您可以使用 SQL Workbench 等工具，以 ANSI SQL 分析 Amazon Redshift 中的資料。Amazon Redshift 亦能搭配使用市面上的熱門第三方 BI 解決方案。

Amazon QuickSight 是一種快速、雲端驅動的商業情報 (BI) 服務，可以很容易地建立視覺化，執行隨選分析，並迅速從數據獲取商業洞見。Amazon QuickSight 已與 Amazon Redshift 整合，目前已開放預覽，計劃將於 2016 年全面推出。

如果您正在 Amazon S3 做為主要的儲存體，執行分析與視覺化的常見方法是在 Amazon Elastic MapReduce (Amazon EMR) 上執行 Apache Spark 筆記本。使用此程序，您可以有彈性地執行 SQL，或執行以 Python 與 Scala 等語言所撰寫的自訂程式碼。

另一種視覺化方式是 Apache Zeppelin，它是開放原始碼的 BI 解決方案，您可以在 Amazon EMR 中執行它，利用 Spark SQL 視覺化 Amazon S3 中的資料。您也可以使用 Apache Zeppelin 視覺化 Amazon Redshift 中的資料。

分析流程與 AWS 服務

AWS 提供多種服務以建立完整的分析平台。圖 2 顯示先前討論過的服務以及各項服務屬於分析流程的哪個部分。

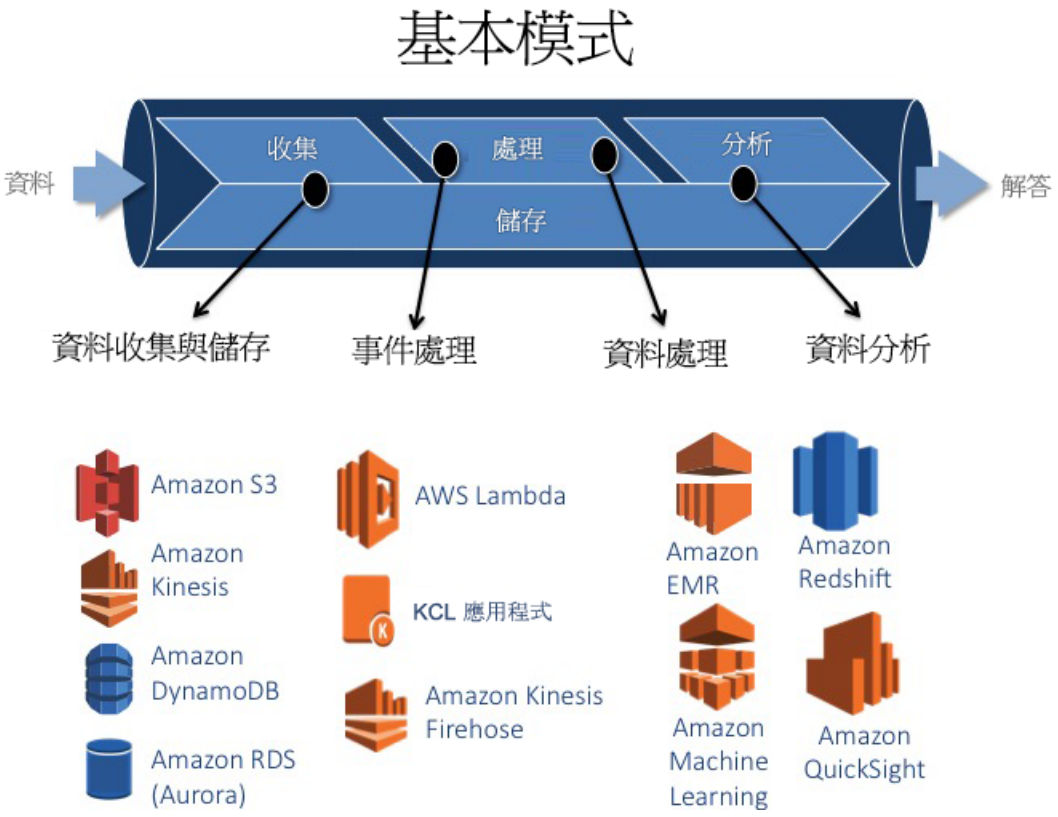

圖 2：分析流程與 AWS 服務

資料倉儲技術選擇

本節將討論建置資料倉儲時的選擇：橫列導向式資料庫、直欄導向式資料庫，以及大量平行處理架構。

橫列導向式資料庫

橫列導向式資料庫通常將所有橫列存放於實體區塊。透過次要索引達到高效能的讀取操作。Oracle Database Server、Microsoft SQL Server、MySQL 及 PostgreSQL 等資料庫都是橫列導向式資料庫系統。這些系統傳統上是用於資料倉儲，但它們比較適合用於交易處理 (OLTP) 而非分析。

為了最佳化做為資料倉儲使用的橫列導向式系統的效能，開發人員採用多項技術，包括建置實體化視圖、建立預累積匯總資料表、在每個可能的述詞組合上建立索引、實作資料分割區以透過查詢最佳化器運用分區劃分，以及執行以索引為基礎的聯結。

傳統橫列式資料存放區受限於單部機器可提供的資源。資料市集利用功能分片 (sharding)，稍微解決了上述問題。您可以將資料倉儲分散至多個資料市集，各個市集分別可滿足特定的功能領域。但是，當資料市集隨著時間而成長擴大，資料處理速度會變慢。

在橫列式資料倉儲中，每個查詢都必須在滿足查詢述詞的所有區塊中讀取所有直列的所有橫欄，包括您並未選取的直欄。上述方法會造成資料倉儲嚴重的效能瓶頸，您在倉儲中的資料表有較多的欄，但查詢僅使用其中少數欄。

直欄導向式資料庫

直欄導向式資料庫將每個欄整理至自己的實體區塊集，而非將所有列包裝成區塊。此功能提高了唯讀查詢的 I/O 效率，因為只需從磁碟 (或記憶體) 讀取查詢所要存取的欄。相較於橫列導向式資料庫，此方式使直欄導向式資料庫成為更好的資料倉儲選擇。

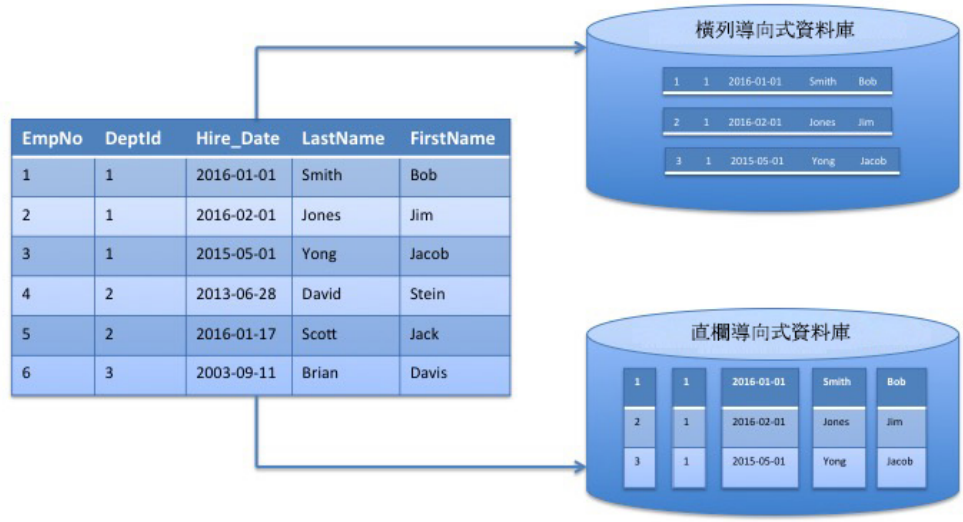

圖 3：橫列導向式與直欄導向式資料庫的比較

上方的圖 3 顯示橫列導向式與直欄導向式資料庫的主要差異。在橫列導向式資料庫中，各列包裝成自己的區塊，而在直欄導向式資料庫中，各欄包裝成自己的區塊。

除了 I/O 較快之外，使用直欄導向式資料庫的另一個最大利益是提升壓縮效能。由於每欄皆包裝成自己的區塊集，因此每個實體區塊皆包含相同的資料類型。當所有資料都有相同的資料類型時，資料庫即可使用效率極高的壓縮演算法。因此，所需的儲存空間少於橫列導向式資料庫。此方法也能大量減少 I/O，因為相同數量的資料存放於較少的區塊中。

有些資料倉儲採用直欄導向式資料庫，包括 Amazon Redshift、Vertica、Teradata Aster 及 Druid。

大量平行處理架構

MPP 架構讓您可利用叢集中所有可用的資源來處理資料，因此可大幅提升 PB 級資料倉儲的效能。只要在叢集中新增更多節點，即可提升 MPP 資料倉儲的效能。有些資料倉儲以 MPP 架構建置，包括 Amazon Redshift、Druid、Vertica、GreenPlum 及 Teradata Aster。Hadoop 與 Spark 等開放原始碼框架亦支援 MPP。

深入探討 Amazon Redshift

Amazon Redshift 採用直欄式 MPP 技術，為高效能、高成本效益的資料倉儲提供多項主要優點，包括高效率壓縮、減少 I/O 以及較低的儲存需求。它以 ANSI SQL 為基礎，因此幾乎或完全無需修改即可執行既有的查詢。因此，目前它已成為企業資料倉儲與資料市集的熱門選擇。本節將深入探討 Amazon Redshift，並進一步討論其功能。

Amazon Redshift 利用直欄式儲存以及將查詢平行處理並分佈至多個節點，為幾乎任何資料規模提供快速的查詢與 I/O 效能。它可自動執行與佈建、設定、監控、備份與保護資料倉儲等相關的常見管理工作，簡化了管理工作且成本低廉。利用上述自動化功能，您可在幾分鐘內實作 PB 級資料倉儲，而非傳統內部部署所需的數週或數個月。

效能

Amazon Redshift 採用直欄式儲存、資料壓縮及區域圖，可減少執行查詢時所需的 I/O 量。交錯排序方式可提供快速的效能，不會增加維護索引或投影所需的間接成本。

Amazon Redshift 採用 MPP 架構，透過平行與分散 SQL 操作以充分運用所有可用的資源。基礎硬體專為高效能資料處理而設計，利用本機連結儲存裝置在 CPU 與磁碟機之間達到最高的輸送量，而 10 GigE 網狀網路可使節點之間達到最高輸送量。可依據您的資料倉儲需求調整效能：AWS 提供具備固態硬碟的密集運算 (DC) 選項與密集儲存 (DS) 選項。透過持續部署軟體升級以持續提升效能，無需使用者介入。

耐用性與可用性

為提供最佳的資料耐用性與可用性，Amazon Redshift 會自動偵測並取代您資料倉儲叢集中任何故障的節點。它讓替換後的節點立即可用並先載入最常存取的資料，因此您可以很快地恢復查詢您的資料。Amazon Redshift 會鏡射整個叢集中的資料，因此會使用其他節點的資料來重建故障節點。上述叢集會處於唯讀模式，直到替換的節點完成佈建並新增至叢集，這通常只需幾分鐘。

Amazon Redshift 叢集位於一個可用區域內。[3]但是，如果您想要一個 Multi-AZ 設定的 Amazon Redshift，可以建立鏡射，然後自我管理複寫與容錯移轉。

您只要在 Amazon Redshift 管理主控台中按幾下滑鼠，即可透過 Amazon Redshift 設定強大的災難復原 (DR) 環境。您可以保留多個 AWS 區域的備份副本。若遇某個 AWS 區域發生服務中斷，即可從不同 AWS 區域的備份還原您的叢集。開始還原操作之後幾分鐘內，即可取得叢集的讀/寫存取權。

可擴展性與彈性

只要在主控台或API 呼叫中按幾下滑鼠，就能依據效能或容量需求的改變，輕鬆變更您資料倉儲中的節點數量與類型。[4]Amazon Redshift 可讓您從一個 160 GB 的小節點開始，然後利用多個節點，將容量一路擴展至 1 PB 以上的壓縮使用者資料。如需詳細資訊，請參閱關於叢集與節點，其收錄於 *Amazon Redshift 叢集管理指南* 中。[5]

在調整規模時，Amazon Redshift 會使現有叢集進入唯讀模式，然後佈建您所選容量的新叢集，再將資料從舊叢集平行複製至新叢集。在此過程中，您只需支付作用中 Amazon Redshift 叢集的費用。在佈建新的叢集時，您可以持續對舊叢集執行查詢。在您的資料複製到新的叢集之後，Amazon Redshift 會自動將查詢重新導向新的叢集，並移除舊叢集。

您可以利用 Amazon Redshift API 動作以程式方式啟動叢集、擴展叢集、建立備份、還原備份等。您可透過此方式，將上述 API 動作整合至您現有的自動化堆疊，或建立符合您需求的自訂自動化。

介面

Amazon Redshift 提供自訂的 Java Database Connectivity (JDBC) 與 Open Database Connectivity (ODBC) 驅動程式，您可以從主控台的**連接用戶端**索引標籤下載，這表示您可以使用各種您熟悉的 SQL 用戶端。您亦可使用標準 PostgreSQL JDBC 與 ODBC 驅動程式。如需 Amazon Redshift 驅動程式的詳細資訊，請參閱 Amazon Redshift 與 PostgreSQL ，其收錄於 *Amazon Redshift 資料庫開發人員指南*。[6]

您也可以找到許多經過驗證的整合範例以及許多受歡迎的 BI 與 ETL 廠商。[7]在這些整合中，載入與卸載會在各個運算節點中平行執行，讓您在多個資源之間能以最大的傳輸速率取入與匯出資料，這些來源包括 Amazon S3、Amazon EMR 及 Amazon DynamoDB。您可以使用 Amazon Kinesis Firehose 將串流資料輕鬆載入 Amazon Redshift，以現有的 BI 工具與儀表板執行接近即時的分析。您可以找到有關您 Amazon Redshift 資料倉儲叢集的運算利用率、記憶體利用率、儲存利用率及讀/寫流量等指標，只要利用主控台或 Amazon CloudWatch API 操作即可。

安全性

為協助提供資料安全性，您可以在 Amazon Virtual Private Cloud (Amazon VPC) 服務為基礎的虛擬私有雲端中執行 Amazon Redshift。您可以使用 VPC 的軟體定義聯網模型，定義防火牆規則以依據您設定的規則來限制流量。[8]Amazon Redshift 在您的用戶端應用程式與 Amazon Redshift 資料倉儲叢集之間支援 SSL 連線，使資料能夠在傳輸過程中受到加密。

Amazon Redshift 運算節點存放您的資料，但只能從叢集的領導者節點存取這些資料。這樣的隔離可提供另一層安全性。Amazon Redshift 整合 AWS CloudTrail，可讓您稽核所有 Amazon Redshift API 呼叫。[9]為確保您未使用之資料的安全，在各個區塊寫入至磁碟機時，Amazon Redshift 會利用硬體加速 AES-256 加密技術為每個區塊加密。此加密作業在 I/O 子系統中低層級執行；此 I/O 子系統會將寫入硬碟機的所有資料加密，包括中間查詢結果。各區塊會以原貌備份，這表示備份也會加密。依據預設，Amazon Redshift 會執行金鑰管理工作，但您可以選擇使用您自己的硬體安全模組 (HSM) 管理您的金鑰，或透過 AWS Key Management Service 管理您的金鑰。[10,11]

成本模型

Amazon Redshift 不需要長期投入或支付前期成本。此定價方式可針對您的需求，消除您規劃與購買資料倉儲容量的資本支出和複雜性。收費依據您的叢集中的節點大小與數量而定。

針對您已佈建儲存容量之高達 100% 的備份儲存容量，我們不會收取額外費用。例如，如果您有一個作用中的叢集，其中有兩個 XL 節點共計 4TB 儲存容量，AWS 將免費提供最多 4TB 的 Amazon S3 備份儲存容量。超過所佈建儲存容量的備份儲存容量，以及在您的叢集終止之後存放的備份，將以標準Amazon S3 費率計費。[12] 我們不會收取 Amazon S3 與 Amazon Redshift 之間通訊的數據傳輸費。如需詳細資訊，請參閱 Amazon Redshift 定價。[13]

理想的使用模式

Amazon Redshift 很適合使用您現有的 BI 工具進行線上分析處理 (OLAP)。諸多組織利用 Amazon Redshift 執行以下任務：

- 執行企業 BI 與報告
- 分析多項產品的全球銷售資料
- 存放股票交易歷史資料
- 分析廣告曝光率與點擊次數
- 匯總博弈資料
- 分析社會發展趨勢
- 衡量醫療照護的臨床品質、營運效率及財務績效

不適合的使用模式

Amazon Redshift 不太適合以下使用模式：

- **小型資料集** – Amazon Redshift 的建置適用於跨叢集進行平行處理。如果您的資料集小於 100 GB，將無法享受 Amazon Redshift 提供的所有利益，Amazon RDS 可能是比較理想的解決方案。

- **OLTP** – Amazon Redshift 是為了資料倉儲工作負載而設計的，提供極快速且低價的分析能力。如果您需要快速交易系統，您可選擇建立在 Amazon RDS 或 NoSQL 資料庫上的傳統關聯式資料庫系統，例如 Amazon DynamoDB。
- **非結構化資料** – Amazon Redshift 中的資料必須以已定義的結構描述進行結構化。Amazon Redshift 不支援各資料行的任意結構描述架構。如果您的資料為非架構化，您可在 Amazon EMR 上執行擷取、轉換及載入 (ETL)，使資料可供載入 Amazon Redshift。如果是 JSON 資料，您可以存放鍵值對，並在您的查詢中使用原生 JSON 功能。[14]
- **BLOB 資料** – 如果您計劃存放二進位大型物件 (BLOB) 檔案，例如數位視訊、影像或音樂，您可以考慮將這些資料存放於 Amazon S3 並在 Amazon Redshift 中參考其位置。在此情況下，Amazon Redshift 會追蹤有關二進位物件的中繼資料 (例如項目名稱、大小、建立日期、擁有者、位置等)，但大型物件本身則存放於 Amazon S3。

轉移至 Amazon Redshift

如果您決定從現有的資料倉儲轉移至 Amazon Redshift，您應依據以下幾個因素選擇您的轉移策略：

- 資料庫與資料表的大小
- 來源伺服器與 AWS 之間的網路頻寬
- 轉移及切換至 AWS 的作業，將以單一步驟完成或隨時間經過一連串的步驟才完成
- 來源系統的資料變更率
- 轉移過程中的轉換作業
- 您計劃用於轉移與 ETL 的合作夥伴工具

單步驟轉移

對於不需要連續操作的小型資料庫而言，單步驟轉移是理想的選擇。客戶可將現有資料庫擷取為逗號分隔值 (CSV) 檔案，然後使用 AWS Import/Export Snowball 等服務將資料集傳送至 Amazon S3，以便載入 Amazon Redshift。客戶接著可測試目標 Amazon Redshift 資料庫，確認資料是否與來源資料庫維持一致。通過所有驗證之後，即可將資料庫切換至 AWS。

雙步驟轉移

雙步驟轉移常用於各種大小的資料庫：

1. **初步資料轉移：**從來源資料庫擷取資料，最好在非尖峰時段進行以降低影響。接著資料會以前述單步驟轉移的方式，轉移至 Amazon Redshift。
2. **變更資料轉移：**在切換資料庫之前，必須將來源資料庫中在初步資料轉移之後發生變更的資料傳播至目標。此步驟將會同步來源與目標資料庫。當所有變更的資料完成轉移之後，即可驗證目標資料庫中的資料，執行必要的測試，如果通過所有測試，即可切換至 Amazon Redshift 資料倉儲。

資料庫轉移工具

有一些工具與技術可用於資料轉移。您可以交互使用這些工具，或使用市面上其他第三方或開放原始碼工具。

1. AWS Database Migration Service 支援前述的單步驟與雙步驟轉移程序。[15] 若要遵循雙步驟轉移程序，您可啟用補充日誌以擷取來源系統的變更。您可以在資料表或資料庫層級啟用補充日誌。
2. 其他資料整合合作夥伴工具如下：
   - Attunity
   - Informatica
   - SnapLogic
   - Talend
   - Bryte

如需有關資料整合及諮詢合作夥伴的詳細資訊，請參閱 Amazon Redshift 合作夥伴。[16]

設計資料倉儲工作流程

在先前章節中，我們討論了 Amazon Redshift 非常適合用於資料倉儲的一些功能。為了了解如何使用 Amazon Redshift 設計資料倉儲工作流程，我們來看看最常見的設計模式以及範例使用案例。

假設有一家跨國製衣廠擁有一千多家零售店，透過百貨公司與折扣商店銷售數個服裝產品線，並設有線上商店。從技術觀點而言，目前上述三個通路獨立營運。各通路擁有各自的管理、銷售點系統及會計部門。沒有單一系統合併所有相關資料集，以提供執行長整體事業的全方位檢視。

假設這位執行長希望縱觀公司三個通路的營運狀況，並且能夠執行以下的專案分析：

- 各個通路中存在何種趨勢？
- 哪些地區在各個通路中表現較佳？
- 公司的廣告與促銷宣傳的效益如何？
- 各服裝產品線之間存在何種趨勢？
- 何種外部力量會影響公司的銷售，例如失業率及天氣情況？
- 商店屬性如何影響銷售，例如員工與幹部的任職期間、小型購物商場相較於大型購物商城、商品在商店中的位置、促銷、通道末端商品展示架、廣告傳單以及店內展示方式？

一個企業資料倉儲就能解答上述問題。企業資料倉儲可收集來自三個通路的各種系統的資料，亦可收集公開的資料，例如天氣與經濟報告。各個資料來源每天會傳送資料，供資料倉儲運用。由於各個資料來源的結構可能不太一樣，所以會先執行擷取、轉換與載入 (ETL) 程序，將資料重新格式化為相同的結構。之後，即可針對所有來源的資料執行分析。我們採用以下資料流程架構來完成上述工作：

圖 4：企業資料倉儲工作流程

1. 此流程的第一個步驟是將不同來源的資料收集至 Amazon S3。Amazon S3 提供一個具有高耐用性、低成本及可擴展性的儲存平台，不同來源的資料能以極低的成本平行寫入此平台。

2. Amazon EMR 用於轉換及清理資料的來源格式，將資料轉換為目的格式。Amazon EMR 已內建整合於 Amazon S3，允許您的 Amazon EMR 叢集中的各個節點與 Amazon S3 之間平行進行多個資料傳送線程。

   資料倉儲通常會在夜間取得新資料。由於在深夜無需進行分析，因此上述傳送程序的唯一要求就是在早晨完成工作，以便執行長與其他商業使用者存取報告與儀表板。因此，您可以利用 Amazon EC2 競價市場進一步降低 ETL 成本。[17] 理想的競價策略是在午夜以極低的價格開始投標，然後隨著時間逐漸提高價格，直到容量獲得授予。隨著截止時間的逼近，如果競價投標不成功，您還是可以利用隨需價格確保在時間內完成。各個來源在 Amazon EMR 上可能有不同的轉換程序，但只要使用 AWS 隨收隨付制 (pay-as-you-go)，即可針對各項轉換建立個別的 Amazon EMR 叢集，並可精確調整適合的容量，完成所有資料轉換作業，不會爭奪其他作業的資源。

3. 各個轉換工作負載會將資料格式化、清理並傳送至 Amazon S3。這裡我們再次使用 Amazon S3，因為 Amazon Redshift 可使用各個叢集節點的多個線程，從 Amazon S3 平行載入資料。Amazon S3 亦提供歷史紀錄，並做為系統之間真正的格式化資源。如有其他需求，其他工具亦可使用 Amazon S3 的資料進行分析。

4. Amazon Redshift 將資料載入、排序、分發及壓縮至資料表，以便有效平行執行分析查詢。隨著資料量的增加及事業的擴充，您可以透過新增節點，輕鬆增加容量。

5. 若要將分析結果視覺化，您可以使用 Amazon QuickSight 或各種合作夥伴視覺化平台，這些平台利用 ODBC 或 JDBC 連接至 Amazon Redshift。此時，執行長與員工即可檢視報告、儀表板及圖表。現在，高階主管可藉由這些資料更有效運用公司資源，最終為股東提升收益與價值。

您可以在事業擴充、開闢新通路、推出顧客專用行動應用程式以及引進更多資料來源時，輕鬆擴大此彈性的架構。您只需要在 Amazon Redshift 管理控制台中按幾下滑鼠或呼叫幾個 API 即可完成。

結論

我們已看到資料倉儲策略的轉變，因為企業正紛紛將分析資料庫與解決方案從現場部署解決方案移轉至雲端，善加利用雲端提供的簡易性、高效能與成本效益。這份白皮書提供 AWS 資料倉儲現狀的完整說明。AWS 提供多樣化的服務與堅強的合作夥伴生態系統，協助您在雲端輕鬆建立及執行企業資料倉儲。成果是高性能、高成本效益的分析架構，可隨著您事業的成長在 AWS 全球基礎設施上不斷擴充。

作者群

協力完成這份文件的個人與組織如下：

- Babu Elumalai，Amazon Web Services 解決方案架構師
- Greg Khairallah，Amazon Web Services 首席 BDM
- Pavan Pothukuchi，Amazon Web Services 首席產品經理
- Jim Gutenkauf，Amazon Web Services 資深技術撰寫人員
- Melanie Henry，Amazon Web Services 資深技術編輯
- Chander Matrubhutam，Amazon Web Services 產品行銷

深入閱讀

如需其他協助，請參考以下資源：

- Apache Hadoop 軟體程式庫[18]
- Amazon Redshift 最佳實務[19]
- Lambda 架構[20]

備註

1https://www.forrester.com/report/The+Forrester+Wave+Enterprise+Data+Warehouse+Q4+2015/-/E-RES124041

2 http://aws.amazon.com/streaming-data/

3 http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/using-regions-availability-zones.html

4 http://docs.aws.amazon.com/redshift/latest/APIReference/Welcome.html

5 http://docs.aws.amazon.com/redshift/latest/mgmt/working-with-clusters.html#rs-about-clusters-and-nodes

6 http://docs.aws.amazon.com/redshift/latest/dg/c_redshift-and-postgres-sql.html

7 http://aws.amazon.com/redshift/partners/

8 https://aws.amazon.com/vpc/

9 https://aws.amazon.com/cloudtrail/

10 http://docs.aws.amazon.com/redshift/latest/mgmt/working-with-HSM.html

11 https://aws.amazon.com/kms/

12 http://aws.amazon.com/s3/pricing/

13 http://aws.amazon.com/redshift/pricing/

14 http://docs.aws.amazon.com/redshift/latest/dg/json-functions.html

15 https://aws.amazon.com/dms/

16 https://aws.amazon.com/redshift/partners/

17 http://aws.amazon.com/ec2/spot/

18 https://hadoop.apache.org/

19 http://docs.aws.amazon.com/redshift/latest/dg/best-practices.html

20 https://en.wikipedia.org/wiki/Lambda_architecture