2. Exploring Data Transformation with Google Cloud

2.1 The Value of Data

2.1.1 資料如何創造價值 (How data creates value)

- 資料是推動創新和差異化的重要因素，也是釋放 AI 價值的關鍵。
- 資料驅動 AI 業務洞察，幫助公司做出更好的即時決策，並作為公司構建和運行應用程式的基礎。
- 最有價值的洞察通常隱藏在來自各種來源和系統的非結構化資料點中，需要合適的工具、技能和策略相結合才能提取。
- ML 和 AI 使組織能夠從過去和現在的資料中產生洞察，並以新的方式感知、預測、推薦和分類資料。
- 智能資料雲: 是釋放更多業務價值的關鍵。

2.1.2 從資料中釋放商業價值 (Unlocking business value from data)

- 釋放資料價值是數位轉型的核心。
- 資料可分為三種類型：
    - 結構化(structured)： 高度組織化和明確定義，通常儲存在表格中，易於分析，例如 CRM 中使用的資料。
    - 半結構化(semi-structured)：介於結構化和非結構化資料之間，組織成層次結構，但沒有完全區分或任何特定順序，例如電子郵件、HTML、JSON 和 XML 文件。
    - 非結構化(unstructured)： 沒有預定義資料模型或未按預定義方式組織的信息，例如文本、圖像、聲音、影片和 IoT 傳感器的輸出資料。
- 非結構化資料佔所有新企業資料的 80% 到 90%，但傳統上難以分析。
- 雲技術和 API（例如 Google Cloud 的 Vision API）使企業能夠從非結構化資料中提取價值。

2.1.3 資料管理概念 (Data management concepts)

- 組織需要現代化的企業資料管理方法來管理大量資料。常見的選項包括資料庫、資料倉庫和資料湖。
- 資料庫(database)：有組織的資料集合，存儲在表中，可以電子訪問。
    - 關連式(relational)： 存儲和提供彼此相關的資料點的訪問，使用表格格式和 SQL 進行查詢和操作，例如 Cloud SQL 和 Spanner。
    - 非關連式(non-relational)： 結構較少，遵循靈活的資料模型，例如 Bigtable。

- 資料倉儲(data warehouses)： 用於分析和報告來自多個來源的結構化和半結構化資料的企業系統，例如 BigQuery。
- 資料湖(data lake)：用於攝取、存儲、探索、處理和分析任何類型或數量原始資料的儲存庫，例如 Cloud Storage。
- 資料倉儲和資料湖是互補工具，針對不同的用途進行了優化。資料倉儲用戶通常是業務智能分析師，而資料湖用戶包括資料工程師和資料科學家。
- 資料的民主化使用戶能夠更深入地了解業務情況。

2.1.4 資料在數位轉型中的角色 (The role of data in digital transformation)

- 組織可以訪問前所未有的資料，包括第一方資料（內部信息）和第二方/第三方資料（外部信息）。
- 第一方資料： 企業從客戶或受眾的交易和互動中收集的專有客戶資料集。
- 第二方資料： 來自另一個組織的第一方資料，例如合作夥伴或供應鏈中的其他企業。
- 第三方資料： 由不直接與組織的客戶或業務互動的組織收集和管理的資料集，例如來自政府、非營利組織或學術來源的資料，或來自行業特定來源的資料。
- 使用外部資料可以通過提供新的背景和見解大大提高資料的價值。
- 舉例說明：一家亞洲廉價航空公司通過分析目的地、航班時間和航班連接等資料，更準確地預測所需的餐食數量，從而改善客戶體驗和餐飲服務的盈利能力。

2.1.5 資料價值鏈 (The data value chain)

- 資料價值鏈描述了資料從原始資料轉化為可執行洞察的過程，包含以下步驟：
  - **生成 (Genesis)：** 資料的初始創建。
  - **收集 (Collection)：** 通過攝取將資料帶到裝配線上。
  - **處理 (Processing)：** 將原始資料轉換為可用於得出見解的形式。
  - **儲存 (Storage)：** 資料落地、可找到並準備進行分析和操作的地方，例如 NoSQL、資料倉庫和對象存儲。
  - **分析 (Analysis)：** 為業務導向的行動提供方向。
  - **激活 (Activation)：** 將分析結果推送給相關業務程序和決策者，以便採取行動。
- 最佳價值鏈會隨著技術進步、勞動力發展或所需輸出變化而變化。

2.1.6 資料治理 (Data governance)

- 資料治理是指設定適用於資料收集、儲存、處理和處置方式的內部標準（資料政策），管理誰可以訪問特定資料以及哪些資料受到治理，並遵守外部標準。
- 資料治理的好處：
  - 使資料更有價值，確保高質量資料，並使與利益相關者安全共享資料變得更容易。
  - 幫助用戶做出更好、更及時的決策，實現資料民主化。

    - 改進成本控制，消除資料重複。
    - 增強監管合規性，避免與不合規相關的風險。
    - 贏得客戶和供應商的更大信任。
    - 幫助管理風險，減少對敏感資料暴露的擔憂。
    - 允許更多人員訪問更多資料，同時確保資料安全。
- 資料治理的必要性:
    - 合規違規風險：沒有資料治理的組織可能面臨罰款、資料質量差、資料查找困難和 AI 模型訓練不佳等問題。
    - 資料的重要性：資料成為組織最有價值的資產，每個組織都需要資料治理。

2.2 Google Cloud Data Management Solutions

2.2.1 非結構化資料儲存 (Unstructured data storage)

- 應用程式需要儲存不同類型的資料，Google Cloud 提供多種核心儲存產品，包含 Cloud Storage、Cloud SQL、Spanner、BigQuery、Firestore 和 Bigtable。
- 對象儲存 (Object storage)： 將資料管理為「對象」，包含資料的二進制形式、元資料和唯一標識符 (URL)，適合儲存影片、圖片和音訊等非結構化資料。
- Cloud Storage：提供持久且高度可用的對象儲存服務，可儲存任意數量的資料並根據需要檢索。用途廣泛，例如提供網站內容、儲存用於歸檔和災難恢復的資料，以及分發大型資料對象。
- Cloud Storage 的主要儲存類別：
    - **標準儲存 (Standard Storage)**：適合頻繁訪問的“熱”資料，也適合短期儲存的資料。
    - **近線儲存 (Nearline Storage)**：適合不頻繁訪問的資料，如每月或更少次的讀取或修改，適用於資料備份、長尾多媒體內容或資料歸檔。
    - **冷線儲存 (Coldline Storage)**：低成本選擇，適合每 90 天最多讀取或修改一次的資料。
    - **歸檔儲存 (Archive Storage)**：成本最低，適合每年訪問不到一次的資料，適用於資料歸檔、在線備份和災難恢復。
- 所有儲存類別的共同特徵包括：
    - 無限制儲存且無最小對象大小要求
    - 全球可訪問性和多位置儲存
    - 低延遲和高耐久性
    - 統一的安全性、工具和 API 體驗
    - 多區域或雙區域儲存具有地理冗餘性，保護資料免受災難事件影響，並進行流量負載均衡以優化性能
- Autoclass： 根據每個對象的訪問模式自動將對象轉換為適當的儲存類別，簡化並自動化成本節省。

2.2.2 結構化資料儲存 (Structured data storage)

- 結構化資料由按預定義格式組織的數字和數值組成，易於在關係資料庫中搜索。
- **Cloud SQL：** 提供完全託管的關係資料庫，包含 MySQL、PostgreSQL 和 SQL Server 作為服務，負責應用補丁和更新、管理備份和配置複製等任務。
- **Spanner：** 一種完全託管、關鍵任務的關係資料庫服務，可水平擴展以應對意外的業務激增，特別適合需要 SQL 關係資料庫管理系統、內置高可用性、強大的全球一致性和高吞吐量的應用程式。
- **Cloud SQL 和 Spanner 的區別：** Cloud SQL 適合本地到區域的可擴展性，可用性超過 99.95%，DMS 可輕鬆遷移資料庫；Spanner 具有無限擴展性、強一致性和高達 99.999% 的可用性，適合需要全球資料和強一致性的情況。
- **BigQuery：** 一種完全託管的資料倉庫，包含從組織內各種來源收集的數 PB 資料，用於指導管理決策，提供儲存和分析兩項服務，具有內置功能，如機器學習、地理空間分析和商業智能，默認靜態加密，並與現有的合作夥伴生態系統無縫集成，可在多雲環境中工作，並具有內置機器學習功能，可與 Vertex AI 無縫集成。

2.2.3 半結構化資料儲存 (Semi-structured data storage)

- 半結構化資料包含結構化和非結構化資料的元素，通常包含一些組織屬性，如標籤或元資料，例如電子郵件。
- **Firestore：** 一種靈活的、水平可擴展的 NoSQL 雲資料庫，用於實時儲存和同步資料，可由移動和網絡應用程式直接訪問，以文檔的形式執行資料儲存，具有自動擴展和離線使用等特點。
- **Bigtable：** Google 的 NoSQL 大資料資料庫服務，支持許多 Google 核心服務，設計用於在一致的低延遲和高吞吐量下處理大工作負載，適合物聯網、用戶分析和金融資料分析等應用程式。
- 選擇 Bigtable 的情況：處理超過 1TB 的半結構化或結構化資料、資料快速且具有高吞吐量或資料變化迅速、處理 NoSQL 資料、資料是時間序列或具有自然排序、處理大資料並在資料上運行批處理或實時處理，或在資料上運行機器學習演算法。

2.2.4 選擇合適的儲存產品 (Choosing the right storage product)

- 選擇儲存產品取決於需要儲存的資料類型和業務需求。
- 如果資料是非結構化的，則 Cloud Storage 是最合適的選擇，需要決定儲存類別或是否使用 Autoclass 功能。
- 如果資料是結構化或半結構化的，則取決於工作負載是交易型 (OLTP) 還是分析型 (OLAP)，以及是否使用 SQL 訪問資料。
- 交易型
    - 需要 SQL：Cloud SQL 適合本地到區域的可擴展性，Spanner 適合全球擴展。
    - 不需要 SQL：Firestore 是最佳選擇。
- 分析型
    - 需要 SQL：BigQuery 是最佳選擇。

    - 不需要使用 SQL： Bigtable 提供可擴展的 NoSQL 解決方案，適合實時、高吞吐量且僅需要毫秒級延遲的應用程式。
- 選擇指南
    - **Cloud Storage**：非結構化資料（媒體、備份、歸檔）。
    - **Cloud SQL 和 Spanner**：結構化(structured)、交易型(OLTP)、SQL。
    - **Firestore**：半結構化(semi-structured)、交易型(OLTP)、No-SQL。
    - **BigQuery**：結構化(structured)、分析型(OLAP)、 SQL。
    - **Bigtable**：半結構化(semi-structured)、分析型(OLAP)、No-SQL。

2.2.5 資料庫遷移和現代化 (Database migration and modernization)

- 在遺留的本地資料庫上運行現代應用程式需要克服與延遲、吞吐量、可用性和擴展性相關的挑戰。
- **資料庫現代化的方法：**
    - **提升和轉移平台遷移：** 將資料庫從本地和私有雲環境遷移到由公共雲提供商託管的相同類型的資料庫，優點是變動最小，並由雲提供商管理資料和基礎設施。
    - **託管資料庫遷移：** 允許將資料庫遷移到完全託管的 Google Cloud 資料庫，需要仔細規劃，但完全託管的解決方案使您可以專注於更高優先級的工作。
- **DMS (Database Migration Service)：** 可以輕鬆將資料庫遷移到 Google Cloud。
- **Datastream：** 可用於在資料庫、儲存系統和應用程式之間同步資料。
- **Wayfair 的案例：** Wayfair 將其運行在 SQL Server 上的本地資料中心遷移到 Google Cloud，使用 Cloud SQL for SQL Server 進行遷移，並使用 GKE 和 Compute Engine VM 託管服務，使用 Pub/Sub 和 Dataflow 將運營資料發送到 BigQuery 中的分析儲存。

總而言之，Google Cloud 提供了多種資料儲存和管理解決方案，涵蓋了非結構化、結構化和半結構化資料，以及資料庫遷移和現代化的方法。選擇合適的解決方案需要根據資料類型、業務需求和工作負載類型等因素進行綜合考慮。

2.3 使資料有用且易於訪問 (Making Data Useful and Accessible)

2.3.1 使用 Looker 進行商業智能和洞察 (Business intelligence and insights using Looker)

- 發掘數資料中的見解需要相當大的努力和專業知識，商業智能解決方案可以幫助實現此目標。
- 組織常面臨的挑戰是找到合適的商業智能解決方案，有些方案過於複雜，有些方案只能使用部分資料進行分析。
- Looker：是一個 Google Cloud 商業智能 (BI) 平台，旨在幫助個人和團隊分析、可視化和共享資料，包括創建易於理解和共享的互動式儀表板和報告。

- Looker 具有可靠的業務資料權威，使團隊中的任何人都可以探索資料、提出並回答自己的問題，並創建可視化圖表，從而使組織不僅能夠發掘見解，還能夠採取行動。
- Looker 支持 BigQuery 以及超過 60 種不同的 SQL 數資料，提供豐富的互動式儀表板和報告，且不影響性能、規模、安全性或資料的新鮮度。
- Looker 也是 100% 基於網頁的，易於集成到現有工作流程中並與組織內的多個團隊共享。
- Diamond Resorts 的案例： Diamond Resorts 之前使用複雜的 Excel 工作簿和傳統 BI 工具來追蹤重要指標，導致沒有業務的共同視圖、重複的資料工程工作和不一致的項目優先級，且基礎設施未能滿足業務需求。他們遷移到雲端並開始使用 Looker 來幫助提高業務敏捷性，在不到 3 個月的時間內獲得實時見解，幫助他們應對 COVID 變化，並提供了 360 度的客戶視圖，並減少了手動報告的時間。

2.3.2 串流分析 (Streaming Analytics)

- 傳統上，資料是批量移動的，批量處理通常一次處理大量資料，但會有較長的延遲時間，不適用於需要流處理的時間敏感資料。
- 串流分析： 是指連續處理和分析資料記錄，而不是批量進行，通常適用於以小尺寸且連續生成資料的資料源。
- 流資料來源:
    - 設備傳感器(equipment sensors)
    - 點擊流(clickstreams)
    - 社交媒體提要(social media feeds)
    - 股市報價(stock market quotes)
    - 應用活動(app activity)
- 公司使用串流分析來實時分析資料，並提供廣泛活動的見解，例如計量、服務器活動、設備地理定位或網站點擊。
- 使用案例包括：
    - 電子商務(Ecommerce)
    - 金融服務(Financial services)
    - 投資服務(Investment services)
    - 新聞媒體(News media)
    - 公用事業(Utilities)
- Google Cloud 提供兩種主要的流分析產品：Pub/Sub 和 Dataflow。

2.3.3 Pub/Sub 和 Dataflow

- 資料管道的早期階段之一是資料攝取，這是接收大量流資料的地方。資料可能來自數千甚至數百萬個不同的異步事件流，例如 IoT 應用程式中的資料。

- **Pub/Sub：** 是一種分佈式消息服務，可以接收來自各種設備流的消息，例如遊戲事件、IoT 設備和應用程式流，意為向訂閱者發布消息。
- **Dataflow：** 創建了一個管道來處理流資料和批量資料。“處理”是指提取、轉換和加載資料 (ETL)。
- **Apache Beam：** 是一個開源的、統一的編程模型，用於定義和執行資料處理管道，包括 ETL、批處理和流處理。
- Dataflow 處理了基礎設施設置和維護的大部分複雜性，並建立在 Google 的基礎設施上，允許可靠的自動擴展以滿足資料管道需求。
- Dataflow 是無服務器和全託管的，這意味著軟件開發人員可以構建和運行應用程序，而無需配置或管理後端基礎設施，並且無需運營團隊即可部署、監控和管理軟件。
- 使用像 Dataflow 這樣的無服務器和全託管解決方案意味著您可以花更多時間分析資料集中的見解，而少花時間配置資源以確保管道成功完成下一個週期。

總而言之，Looker 提供了強大的商業智能功能，使組織能夠輕鬆分析、可視化和共享資料，並從中獲得洞察。串流分析則可以實時處理和分析資料，而 Pub/Sub 和 Dataflow 則提供了強大的資料管道解決方案，用於攝取、處理和分析流資料和批量資料，使資料從生成的那一刻起就更有用且易於訪問。