

Google 數據分析核心技術: BigQuery 深度介紹及初步教學

此篇白皮書介紹了 Google Cloud 資料分析服務 BigQuery, 一個架設在雲端、針對海量數據集的全託管交互式查詢服務。BigQuery 是 Google 提供給外界使用的核心技術之一, 其內部代號是 Dremel。本文討論了此項技術「大規模平行處理查詢引擎的獨特性」、「BigQuery 和 Dremel 之間的差異」、以及「BigQuery 與 MapReduce / Hadoop 和其他資料倉儲解決方案的比較」。

Google 如何營運大數據? 就靠 Dremel !

Google 無時無刻都在處理海量數據, 以提供消費者 Google Search、YouTube、Gmail、Google Docs 等服務。思考一下 Google 平常如何處理這些數據? 您可以想像以下場景:

有天 Director 突然問道:「昨天 Google Ads 在東京的廣告曝光次數是多少?」「快速畫一版 Google Ads 在某特定地區和特定時段的流量圖表。」如果你在 Google 工作, 要用什麼技術才能在短短幾分鐘內回答 Director 的問題? 答案就是 Dremel。

查詢服務 Dremel 允許您針對海量數據集進行類 SQL 語法查詢, 並在幾秒鐘內得到準確的結果。您只要會基本的 SQL 就可以靈活查詢超大型數據集。Google 的工程師、分析師、客戶經理, 每天都會多次使用 Dremel。

BigQuery: 開放給外界使用的 Dremel

在深入了解 Dremel 之前, 我們先簡單說明 Dremel 和 BigQuery 的差異。BigQuery 是 Google 提供給外界使用的 Dremel 版本, BigQuery 保留了 Dremel 的核心功能給第三方開發人員使用, 可以透過 REST API、命令行界面、Web UI、訪問控制來實作, 同時保留 Dremel 最突出且方便的查詢性能。此篇白皮書稍後將討論 Dremel 的底層技術, 然後將其與 BigQuery 和其他資料倉儲技術 (如: MapReduce/Hadoop 等解決方案) 比較。

架設於雲端的 Dremel 是一款大規模平行處理查詢服務, 它與 Google 共享相同的基礎架構, 可以在數十秒內掃描 350 億筆記錄未索引的數據, 因此它可以平行處理每個查詢並同時在數萬台伺服器上運行。

Google's Cloud Platform (GCP) 讓您可以以極具競爭力的性價比達到非常快速的查詢性能。同時您也不需為基礎設施負擔任何費用。讓我們透過以下 SQL 查詢作為範例，該查詢請求 Wikipedia® 中包含數字的內容標題：

```
select count(*) from publicdata:samples.wikipedia where REGEXP_MATCH  
(title, '[0-9]*') AND wp_namespace = 0;
```

相關資訊：

- 此「wikipedia」資料表包含維基百科文章內容的所有更改歷史記錄，涵蓋 3.14 億筆記錄，約 35.7 GB。
- 表達式 REGEXP_MATCH (title, '[0-9] +') 代表它對每個有更改記錄的標題執行一個匹配的正規表示式 (regular expression)，以提取標題中包含數字的記錄 (例如：「美國職業棒球大聯盟全壘打500強名單」或「2008 年美國總統大選」)。
- 最重要的是此資料表沒有預先準備好索引或任何預先統計的值。

當您在 BigQuery 上發出上述查詢時，在大多數情況下，您會得到「223,163,387」的結果，反應時間約 10 秒。您可以看到有大約 2.23 億筆記錄的維基百科標題 (含更改紀錄) 中包含數字，透過將正規表示式匹配至實際應用在資料表中的所有記錄來統計此結果。

Dremel 甚至可以在短短幾十秒內，在一個由大約 350 億筆記錄和 20 TB 組成的大型 logging 資料表上執行完複雜的正規表示式的文本匹配。Dremel 具有極高的擴展性，在多數情況下無論查詢的數據集多大，它都會在幾秒或幾十秒內產出結果。

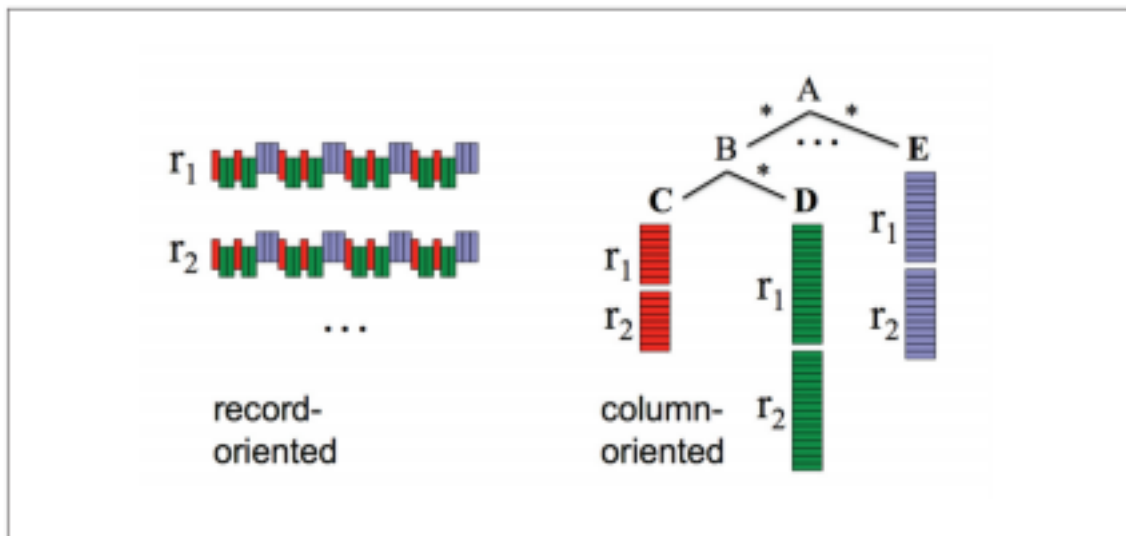
Dremel 的 Columnar Storage 和樹狀結構 (Tree Architecture)

為什麼 Dremel 可以那麼快速地產出結果？答案就是以下的兩個核心技術，這為 Dremel 帶來了極高的性能：

1. Columnar Storage: 數據以欄 (column) 的方式存儲，這使得可以實作出非常高的壓縮比和掃描 Columnar Storage 處理能力。
2. 樹狀結構能在幾秒內對數千台機發送查詢工作或收集結果。

Columnar Storage

Dremel 將數據存儲在 Columnar Storage 中，這意味著它將記錄 (也就是 row) 分離，並將每個值存儲在不同的 volume 上，而傳統數據庫通常將整個記錄存儲在一個 volume。



Columnar storage of Dremel

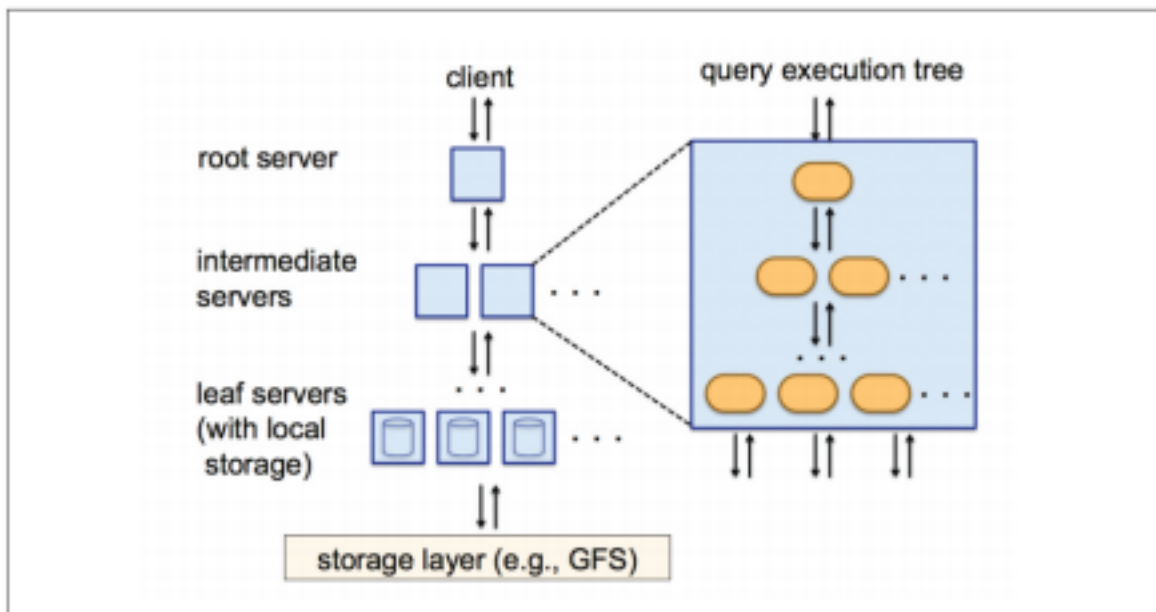
這種技術稱為 Columnar Storage，已用於傳統的資料倉儲解決方案。Columnar Storage 具有以下優點：

- 傳輸最小化：在執行查詢時，只掃描和傳輸每個查詢上所需的 column 值。例如：查詢 “SELECT top (title) FROM foo” 將僅訪問標題 column 值。在 Wikipedia 範例中，查詢將僅掃描 35.7 GB 中的 9.13 GB。
- 壓縮比更高：研究報告指出，Columnar Storage 可以達到 1:10 的壓縮比，而普通的 row-based storage 可以壓縮大約 1:3。因為每 column 具有相似的值，特別是如果 column 的基數 (可能 column 值的變化) 較低，則比 row-based storage 更容易獲得更高的壓縮比。

Columnar Storage 的缺點是更新現有記錄時缺乏效率。對於 Dremel，它根本不支援任何更新操作。因此該技術主要用於只讀 OLAP / BI 類型的使用場景。雖然該技術作為資料倉儲數據庫設計很受歡迎，Dremel 是首批透過上千台伺服器運算、以 columnar storage 為基礎的分析系統服務。

樹狀結構 (Tree Architecture)

Google 在設計 Dremel 時面臨的挑戰是如何在幾秒鐘內跨數萬台機器發送查詢和收集結果，而樹狀結構解決了這項挑戰。該結構形成一個大規模平行處理分散式樹狀圖，將查詢推送到樹中，然後以極快的速度收集來自葉子的結果。



Dremel 的樹狀結構

透過這種架構，Google 得以在 Dremel 上實踐分散式設計，並在雲端平台上大規模平行處理以 column 為基礎的資料庫。Columnar Storage 和樹狀結構是 Dremel 能在性能和成本上取得優勢的原因。

Dremel: 以「Google Speed」開展業務的關鍵


Google 自 2006 年以來就在使用和優化 Dremel。應用場景涵蓋：

- 抓取 web 文檔的分析
- 追蹤 Android Market 中應用程式的安裝數據
- Google 產品的當機報告
- Google Books 的 OCR 結果
- 垃圾郵件分析
- 為 Google Map 上的圖塊除錯
- 全代管式資料庫 Bigtable 中資料表的遷移
- Google 分散式構建系統上運行的測試結果
- 數十萬個磁碟的 Disk I/O 統計訊息
- 對 Google 數據中心的維運作業進行資源監控
- Google codebase 中的符號和相依性

從上述可以看出 Dremel 一直是 Google 的重要核心技術，讓 Google 可以在大多數的時候透過大數據加快運營腳步。

什麼是 BigQuery ？

Google 於 2012 年前後將 BigQuery 作為公開技術服務釋出給大眾，讓非 Google 的企業及開發人員可以利用 Dremel 的強大功能滿足其大數據處理要求。



The screenshot displays the BigQuery web interface. On the left, there is a sidebar with navigation options: Query history, Saved queries, Job history, Transfers, Scheduled queries, BI Engine, and Resources (+ ADD DATA). The main area is the Query editor, which contains a SQL query: `1 SELECT COUNT(*) FROM publicdata.samples.wikipedia`, `2 WHERE`, `3 REGEXP_CONTAINS(title, '[0-9]*') AND vp_namespace = 0`. Below the editor, there are buttons for Run, Save query, Save view, Schedule query, and More. The Query results section shows 'Query complete (2.3 sec elapsed, 9.1 GB processed)' and a table with one row:

Row	fd_
1	223163387

在 BigQuery 上查詢 Wikipedia 資料表 (如範例)，您只需註冊即可試用 BigQuery。

BigQuery 保留了 Dremel 的核心功能給第三方開發人員使用，它可以透過 REST API、command line、Web UI、控制存取、數據模式管理、與 Cloud Storage 整合來實作。

BigQuery 和 Dremel 共享相同的底層架構，使用者可以透過 BigQuery 使用到 Dremel 的功能，同時有效利用 Google 龐大的計算資源，包含：多次的跨區 (region) 複製、數據中心可擴展性高、不需開發人員管理的基礎設施。

BigQuery 與 MapReduce

接下來我們將針對 BigQuery 與其他大數據技術 MapReduce 和 Data Warehouse 等解決方案進行比較。

在 2004 年的研究中可以看出 Google 已經使用 MapReduce 處理大數據相當長的時間了，有些人應該已經聽過 MapReduce 及其開源版的 Hadoop，在此簡單說明

BigQuery 和 MapReduce 的區別：

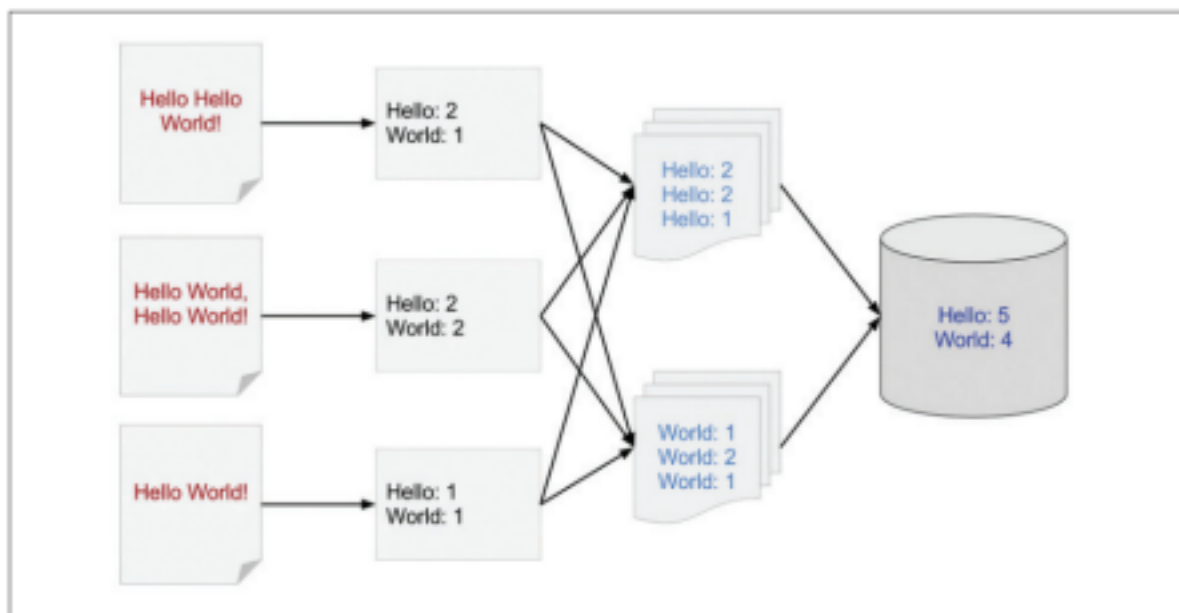
- Dremel 是大型數據集的互動式數據分析工具
- MapReduce 是批次處理大型數據集的程式框架

此外, Dremel 能在幾秒或幾十秒內完成大多數查詢, 甚至非程式人員也可以使用, 而 MapReduce 則花費更長的時間來完成數據查詢 (至少幾分鐘, 有時甚至幾小時或幾天)。

BigQuery 與 MapReduce 的比較

MapReduce 分散式計算技術允許您以程式實作自定義「mapper」和「reducer」功能, 並同時在數百或數千台伺服器上運行批次處理。下圖顯示了相關的 Data flow。

Mappers 從文字中提取單詞, reducers 會合併每個單詞的計數。



MapReduce Data Flow

透過 MapReduce, 企業可以以高度可擴展的方式, 經濟實惠且高效地在其大數據上平行處理資料, 無需從頭開始設計大型分散式計算集群或購買昂貴的高階關聯資料庫解決方案和設備。

在過去, MapReduce 的開源實作 Hadoop 一直是處理大數據的主流技術, 用於各種應用程式, 如: log 分析、社群媒體使用者分析、推薦引擎、非結構化數據處理、資料探勘、文字挖掘等。

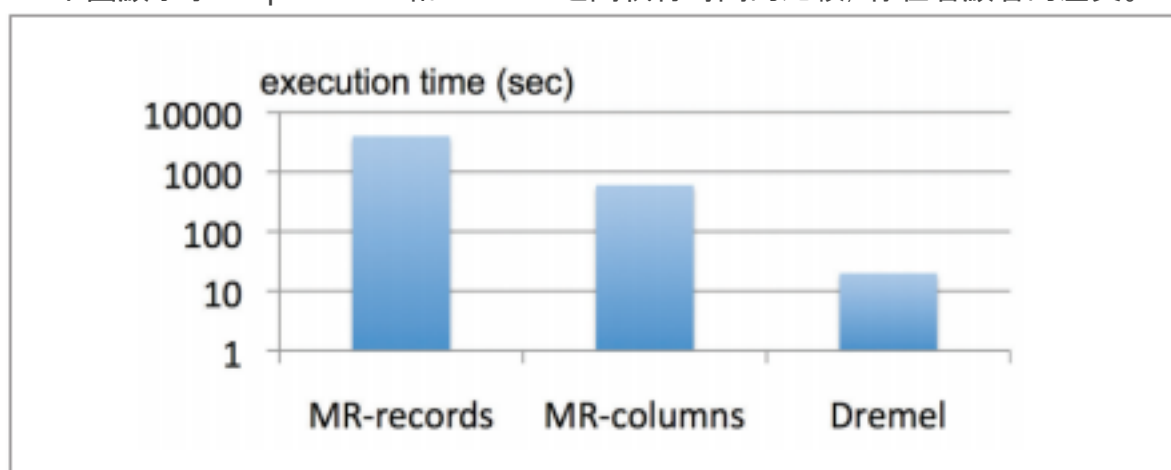
MapReduce 的限制

過去 Google Ads API 有時會使用 Google 內部的 MapReduce 前端：Tenzing(類似於 Hive, 因 Tenzing 可以做為 Hadoop SQL 前端) 分析流量, 能在極大的廣告資料表中執行多個 JOIN 操作。目標是在某些條件下合併和過濾它們, 以便為一組帳戶提取廣告列表。MapReduce 在這樣的場景中運行良好, 在合理的時間內 (例如：幾十分鐘) 提供結果。如果使用傳統的關聯資料庫技術, 那麼同樣的查詢將花費不合理的時間與成本, 或者根本不可能完成任務。

但是 MapReduce 只是一部分的解決方案, 能夠處理 $\frac{1}{3}$ 的問題。當我需要近乎即時的分析結果時, MapReduce 仍然太慢了。即使是相對蠻簡單的工作也需要幾分鐘才能完成, 更多時候還需要一天或甚至更長時間。此外如果在 MapReduce 中撰寫的程式碼有誤將導致結果不正確, 我將不得不修正錯誤並重新重新啟動作業。

MapReduce 被設計為批次處理框架, 因此不適用於臨時和試誤數據分析, 其運轉時間太慢, 不允許程式人員在大數據上執行反覆或一次性分析任務。也因此過去 Google Ads API 所有的分析任務中大約有三分之二上會使用 Dremel 而不是 MapReduce。

下圖顯示了 MapReduce 和 Dremel 之間執行時間的比較, 存在著顯著的差異。



MapReduce 和 Dremel 執行時間比較

此圖表是針對 850 億條記錄和 3,000 個節點 (node) 進行比較。「MR-records」是指在 MapReduce row-based storage 的存取作業, 而「MR-columns」是指基於 column 的儲存 MR 作業。MapReduce 和 Dremel 都是大規模平行處理計算的基礎架構, 但 Dremel 專門設計用於在幾秒鐘內對大數據執行查詢。

BigQuery 和 MapReduce 比較

BigQuery 和 MapReduce 是本質上不相同的技術，在不同的場景有不同的用法。下表對兩種技術進行了比較，並顯示了它們可應用的地方。

關鍵差異	BigQuery	MapReduce
它是甚麼？	查詢大型數據集的服務	用於處理大型數據集的程式模型
常見使用案例	大型數據集的臨時和試誤 互動查詢, 用於快速分析 和故障排除	批量處理大型數據集消耗 時間以進行數據轉換或聚合
應用實例		
適用於 OLAP/BI	是	否
適用於資料探勘	部份 (用於資料探勘的預先分析)	是
快速回應	是	否 (需要數分鐘到數天)
是否對非程式設計師來說 簡單易用 (例如: 分析師、技術支援人員)	是	否 (需要 Hive)
透過撰寫程式處理複雜的 邏輯	部份 (可以透過 User Defined Function 處理)	是
處理非結構化資料	部份 (可以使用正規表式或處理 JSON)	是
數據處理		
處理大量資料, JOIN 大型 資料表	是	是
更新既有資料	否	是

BigQuery 使用 SQL 處理結構化數據。例如: 您若需要在 BigQuery 定義一個表 (table), 您必須使用 column 定義, 然後將 CSV 中的數據導入 Google Cloud Storage, 接著導入 BigQuery。您還需要在 SQL 語句中表達查詢邏輯。BigQuery 適用於 OLAP (在線分析處理) 或 BI (商業智能) 用法, 其中大多數查詢都很簡單, 透過快速聚合和按 column 過濾來完成。

當您希望以程式處理非結構化數據時，MapReduce 是很好的選擇。Mapper 和 reducer 可以相容任何類型的數據並將其套用複雜的邏輯。MapReduce 可用於資料探勘等應用程式，您需要將複雜的統計算法或資料探勘算法應用於一大區塊文字或二進制數據。若需要輸出 GB 等級的數據，可能需要使用 MapReduce，就像合併兩張大表一樣。使用者可能希望應用這些標準來決定使用哪種技術：

使用 BigQuery

- 使用指定條件查詢的特定記錄。如：用帳戶 ID 檢索出的請求日誌。
- 透過動態變化的條件快速統計數據。如：獲取 Web 應用程式前一晚請求流量的統計，並將其繪製圖表。
- 試誤法分析數據。如：透過各種條件 (包括按小時，天等) 確定故障原因並彙總值等。

使用 MapReduce

- 在大數據上執行複雜的資料探勘需要多次疊代或多重路徑的程式演算法處理。
- 對大型數據執行複雜 JOIN。
- 輸出處理過的大量數據。

您可以整合並充分利用 BigQuery 和 MapReduce 這兩種技術來建立一個完整的解決方案。舉例來說：

- 使用 MapReduce 對大量數據進行複雜 JOIN 和數據轉換，接著利用 BigQuery 將結果數據快速聚合，並即時數據分析。
- 利用 BigQuery 快速分析數據，以達到預先檢查的目的，接著撰寫並執行 MapReduce 程式，以執行數據生成處理或資料探勘。

用於 OLAP/BI 的數據倉儲解決方案及設備

許多企業已開始使用數據倉儲解決方案和設備來處理 OLAP/BI 案例了。透過 BigQuery，您大致可以利用以下三種方法來增加大數據處理的效能：

- 關聯式 OLAP (ROLAP)
- 多維度 OLAP
- 全部掃描

關聯 OLAP (ROLAP)

ROLAP 是基於關聯式資料庫 (RDB) 的一種 OLAP 解決方案。為了讓 RDB 更快速，您常需要在執行 OLAP 查詢前先建立索引。如果沒有索引，在大數據執行查詢時，回應速度會非常慢。因此您必須事先替可能會需要的查詢建立索引，然而在許多情況下，您需要建立大量索引以覆蓋所有預期的查詢，這些索引的大小甚至可能比原始數據大。假如數據量真的非常龐大，有時候會需要更大、更複雜且更昂貴的硬體設備才能儲存整個數據集和索引。

多維 OLAP(MOLAP)

MOLAP 是一種 OLAP 解決方案，協助在設計階段根據預先定義的維度，建立資料立方體 (datacube) 或資料超市 (data mart)。舉例來說，假如您要將 HTTP 存取紀錄輸入到一個 MOLAP 解決方案中，您應該選擇像是「時間」、「請求 URL」和「User Agent」等維度，以便 MOLAP 建立包含這些維度和聚合數值的資料立方體。在這之後，分析人員和用戶便可以快速獲得查詢結果，像是「以小時為單位，計算特定用戶的請求數量是多少？」

然而 MOLAP 有一個缺點，在分析師分析數據前，BI 工程師需要先花費大量時間和金錢來設計和建立那些資料立方體和資料超市。有時候這些設計很脆弱，即使是微小的錯誤，都可能失敗導致需要重建整個流程。

全表掃描 (Full-scan) 速度是解決方案

如您所見，無論是 ROLAP 或 MOLAP，兩者都不適合即時查詢或反覆試驗數據分析，因為您必須在設計或輸入時定義所有可能的查詢。在現實世界中，即時查詢是 OLAP 需求的主要部分，如同 Googler 日常生活案例所見—你永遠無法想像未來會需要什麼樣的查詢。為了因應這些案例，您需要增加全掃描的速度 (或資料表掃描)，讓您不用透過索引或預先聚合數值就能直接從儲存空間存取所有紀錄。

如同前面所提到的，Disk I/O 吞吐量是影響全表掃描效能優劣的關鍵。傳統的資料倉儲解決方案和設備已經嘗試利用以下的技術，提升 Disk I/O 吞吐量：

- 記憶體資料庫或快閃儲存裝置：最普遍的解決方案是以記憶體模組和快閃儲存裝置 (SSD 固態硬碟) 組成資料庫設備，以處理大數據。如果您沒有成本限制，這是最好的解決方案。如果設備是用來儲存大數據，且是由 SSD 建構而成，費用可能會高達數十萬美元。

- Columnar Storage 技術會將每個紀錄的 column 值以不同的儲存量儲存。這可以讓壓縮比和 Disk I/O 效率更勝傳統的 row-based storage。自 90 年代以來，Columnar Storage 已經變成資料倉儲解決辦法的標準技術。BigQuery (Dremel) 透過進一步優化以充分利用它。
- 平行 Disk I/O: 提高磁碟吞吐量最後也是最重要的方法是平行 Disk I/O。多個磁碟驅動器平行運作，可以讓全表掃描的效能線性成長。有些資料倉儲設備提供特殊的儲存單元，可以在數十或數百個磁碟驅動器上平行處理您的查詢。但由於這些設備和儲存解決方案都需要部署在本地端，且為專有硬體產品，價格往往相當昂貴。

BigQuery 利用雲端平台的規模經濟解決平行 Disk I/O 的問題。您需要同時運行 10,000 磁碟和 5,000 個處理器，以便在一秒內執行 1TB 數據的全掃描。既然 Google 已經在自己的數據中心擁有大量磁碟，為何不善用它們實作大規模平行處理呢？

BigQuery 獨特的性能

因為 Dremel 有大規模平行查詢引擎的特殊組合，BigQuery 能為即時查詢提供極好的成本效益和全表掃描性能。

雲端大規模平行查詢服務

到目前為止，這種級別的查詢效能在沒有索引的情況下，數十秒內可以完全掃描 350 億筆記錄，只有非常昂貴的資料倉儲設備，或配有完整記憶體和快閃儲存設備的資料庫伺服器叢集整合，才有可能實現。

在 BigQuery 發布之前，企業必須花費數十萬美元，甚至更高的費用，才能有效查詢這樣的數據量。相較之下，使用 BigQuery 使成本大幅下降。想了解價格的差異，您可以參考我們在本文開頭所探討的維基百科案例。如果您在 BigQuery 上執行查詢，每個查詢僅需要花費 0.046 美元。

請注意，BigQuery 只會掃描查詢所需的 column 值，而非所有 column。截至 2019 年 10 月 1 日，在美國地區每 TB 查詢費用為 5 美金。這個查詢範例，所需要掃描的數據量為 9.13 GB，因此每個查詢的費用為 $0.00913\text{TB} * \$5 = \0.046 。更詳細的資訊請參考 [BigQuery 定價表](#)。

如何導入 (import) 大數據

在使用大數據時，如何將數據導入到 BigQuery 是第一個要克服的挑戰。可以透過以下兩個步驟完成：

1. 將您的數據上傳到 Google Cloud Storage。在這個階段會遇到的困難通常是用於此步驟網路頻寬。
2. 將文件輸入到 BigQuery。這個步驟可以透過命令列工具、Web 使用者介面或 API 執行，通常可以在半小時內輸入約 100GB 的數據量。

只要使用這些解決方案，就可以輕鬆地從舊有的資料庫中提取出大數據，轉換或清理後並將其輸入到 BigQuery。

為什麼要使用 Google Cloud

初期在將數據導入到雲端時，會需要一定的成本，但這些成本未來都會被 BigQuery 所帶來的強大優勢弭平。舉例來說，作為全託管服務 BigQuery 不需要您容量規劃、預先配置、24 小時全天候監控和操作，亦不需要手動安全程式修補和更新。您只需要將數據集上傳至您的 Google Cloud Storage，將數據輸入到 BigQuery 即可，其他部分就交給 Google 的專家管理。這會大幅降低您擁有數據處理解決方案的總體成本 (TCO)

。

不斷成長的數據集已經造成許多使用資料倉儲和 BI 工具的 IT 部門人員的負擔，工程師除了分析數據和解決問題外，還必須擔心很多問題。透過使用 BigQuery，IT 團隊可以更專注在像是建立查詢以分析高貢獻客群和效能數據。此外，BigQuery 的 REST API 讓您可以輕鬆建構 dashboard 和 mobile 前後端，讓您可以隨時隨地將有價值的數據交到需要的員工手中。

結論

BigQuery 是一種查詢服務，讓您可以在幾秒鐘內對數 TB 的數據執行類似 SQL 的查詢。MapReduce 和 Bigtable 都是 Google 內部重要的核心技術去處理各種分析任務。Google 推出的 Google BigQuery 是 Dremel 的進階版本，讓開發人員和企業能夠利用 Dremel 強大的功能來處理大數據，並同樣以快速的速度促進業務的發展。

雖然 MapReduce 適合需要長時間運作的批次處理 (如：資料探勘)，但 BigQuery 是需要快速獲得結果的即時 OLAP/BI 查詢的最佳選擇。BigQuery 是雲端大規模平行查詢資料庫，相較於過去資料倉儲解決方案和設備，其全表掃描查詢的效能和成本效益都有顯著的優勢。

(原文翻譯自 [Google Cloud](#)。)