# VMware Tanzu Greenplum

## 大規模平行分析數據平台

### 概觀

VMware Tanzu Greenplum 是專為複雜分析查詢所設計的關鍵任務企業資料倉儲，可因應 PB 規模的資料集。Greenplum 中的每個伺服器節點，都擁有且負責管理整體資料的不同部分。系統會使用大規模平行查詢最佳化工具進行協調，並透過高速軟體互連功能，在所有可用硬體之間自動分配資料及進行平行查詢工作負載。

VMware Tanzu Greenplum 能支援高度平行的大規模分析工作負載。我們採用的技術可提供高速資料載入和交易，且具有 ACID 保證。

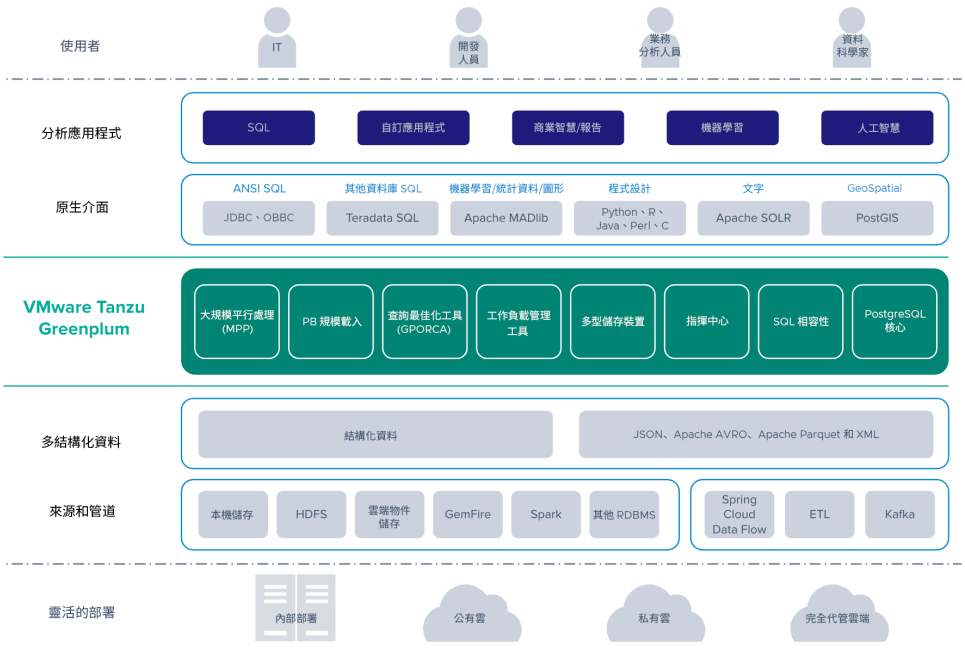
### 主要優勢

- 相容於 ANSI 和 PostgreSQL 的 SQL
- 可從 SMP 資料庫輕鬆移轉，以實現更大的規模，將容量從 TB 擴充至 PB
- 與資料湖相較，部署作業精簡且效率十足
- 高度並行的環境
- 高度可用的關鍵任務企業部署
- 支援豐富的資料科學工具
- 能搭配雲端儲存空間、Hadoop 和多語言系統建立數據聯邦

### 運用單一水平擴充環境來整合資料孤島

Greenplum 為大規模平行資料庫，可透過單一整合式叢集平台，出色管理數十、數百和數千 TB 的資料。在核心的關聯式資料庫功能之外，Greenplum 也提供一系列的整合功能、延伸模組、分析模組和擴充性方法，因此 Greenplum 不只是關聯式資料庫，更可做為資料儲存和分析平台。

儘管已採用資料湖，企業依舊在整合結構化和非結構化資料上吃盡苦頭。傳統的企業資料倉儲不擅於提供即時深入洞悉。而節節攀升的資料量，也會對基礎架構和資源造成壓力。VMware 精心打造的 Greenplum 為大規模平行關聯式資料庫管理系統，有助於解決上述挑戰。Greenplum 能執行不同用途的工作負載，包括傳統商業智慧、線上交易處理，以及採用 GPU 加速的深度學習。無論需要擷取串流、執行點查詢、探索資料科學，還是長時間執行分析查詢，Greenplum 都可助您一臂之力。由於 Greenplum 採用 PostgreSQL 為基礎，資料分析團隊能使用 SQL 來存取其強大功能。舉凡機器學習、地理空間查詢、圖形分析、文字分析，以及連結外部資料來源等作業，皆可透過熟悉的 SQL 介面進行。



**主要功能**

- MPP 無共享架構
- 關鍵任務高可用性
- 支援混合式交易處理工作負載
- 高速互連
- 規則式工作負載管理
- 平行查詢最佳化工具
- 可擴充的資料類型
- 聯邦式查詢處理
- 使用 S3 SELECT 技術加速查詢
- 多溫資料儲存
- 串流資料擷取
- 資料庫內的機器學習
- 可延展的圖形分析
- 採用 GPU 加速的深度學習
- 地理空間索引、搜尋和彙整
- Python、R、Java 和 C 擴充性
- 文字搜尋和分析
- 圖形化 DBA 管理
- 平行且可延展的資料備份
- 多叢集資料複本
- 順暢連接 Apache Spark、Apache Kafka 和 Apache Nifi

## 主要特點

**採用不限基礎架構的部署模式，避免受限於廠商**
Greenplum 可運用各種經濟實惠的選項，在資料中心或公有雲中執行，賦予您更高的便利性。無論在裸機、虛擬機、私有雲或公有雲上，Greenplum 皆可發揮出色的執行效能。

**輕鬆處理傳統商業智慧工作負載**
Greenplum 提供完善的 SQL-92 和 SQL-99 語言支援，並隨附 SQL 2003 OLAP 延伸模組，包括窗函數、Rollup、Cube，以及各種運算式功能。所有查詢都會在整個叢集內平行執行。Greenplum 能完整支援多款標準資料庫介面 (包括 PostgreSQL、SQL、ODBC、JDBC 和 OLEDB 等)，且已通過眾多商業智慧 (BI) 及擷取、轉換和載入 (ETL) 工具的認證。

**混合式儲存**
在資料儲存區和雲端服務的蓬勃發展下，現行資料不僅存放在許多不同的系統內，更採用多種格式。許多時候，資料會依據所在位置、執行的作業，以及存取頻率來分類，又可根據存取頻率劃分為即時或交易 (熱)、頻率較低 (暖) 或封存 (冷)。Greenplum 採用磁碟分割資料表機制，可讓使用者將「冷」和「熱」磁碟分割分別委派給資料行存放區和資料列存放區。從使用者的觀點來看，他們可在 Greenplum 中查詢採混合儲存機制的資料表，完全無需考量儲存類型為何。

**SQL 容器化**
Greenplum Resource Groups 提供資源隔離功能，可運用在查詢、多租戶和混合式工作負載等用途上。SQL 容器化作業會將 CPU 和記憶體資源 (以及並行交易) 匯集成群組，以確保每個項目都可分配到預先定義的數量。

**Apache MADlib 可提供嶄新的深度學習機會**
Apache MADlib 現在支援高度平行、採用 GPU 加速的處理作業，可用來進行深度學習。Greenplum 使用者可善加利用叢集硬體內嵌的 GPU，以實現達兩個數量級或優於純 CPU 處理作業的效能。

**分析圖形、地理空間和文字分析**
Greenplum 會透過 PostGIS (適用於 PostgreSQL 的空間資料庫延伸模組)，在資料庫中儲存及處理地理資訊系統 (GIS) 物件。Pivotal GPText 採用 Apache Solr 為基礎，能使用簡單的 SQL 介面處理原始文字資料，包括電子郵件和社交媒體摘要。圖形分析會透過 Apache MADlib 進行，這款採開放原始碼的資料庫具備多款圖形、統計資料和機器學習功能。

**使用 Apache Spark 大力推動分析**
Apache Spark 為記憶體內部資料處理引擎，能疾速運作。Greenplum Spark Connector 可在 Greenplum 和 Apache Spark 叢集之間進行高速且雙向的平行資料傳輸。如此一來，使用者就能使用儲存在 Greenplum 中的資料，快速執行記憶體內部分析、探索分析和 ETL 處理作業。

## 摘要

Greenplum 為採用開放原始碼的資料分析平台，可針對龐大的資料量進行強大且快速的分析。Greenplum 是專為機器學習和進階資料科學精心打造而成，能提供無與倫比的分析查詢效能來處理大量的資料，並與領先業界的分析資料庫和軟體堆疊緊密整合。如需有關 Greenplum 的其他詳細資訊，請參閱 https://tanzu.vmware.com/greenplum。歡迎前往 www.greenplum.org 下載開放原始碼版本的 Greenplum (Greenplum Database)。

**VMware, Inc.** 3401 Hillview Avenue Palo Alto CA 94304 USA Tel 877-486-9273 Fax 650-427-5001 vmware.com
11073 台北市信義區松仁路 100 號 台北南山廣場辦公大樓 38 樓之一 電話 +886-2-3725-7000 傳真 +886-2-3725-7010 www.vmware.com/tw