

VMware Greenplum

Massively Parallel Analytical Data Platform

AT A GLANCE

VMware Greenplum is a mission-critical enterprise data warehouse designed for complex analytic queries on data sets at petabyte scale. Each server node in Greenplum owns and manages a distinct portion of the overall data. Orchestrated by a massively parallel query optimizer, the system automatically distributes data and parallelizes query workloads across all available hardware via a high-speed software interconnect.

VMware Greenplum supports highly concurrent analytical workloads at scale. Our techniques provide high speed data loading and transactions with ACID guarantees.

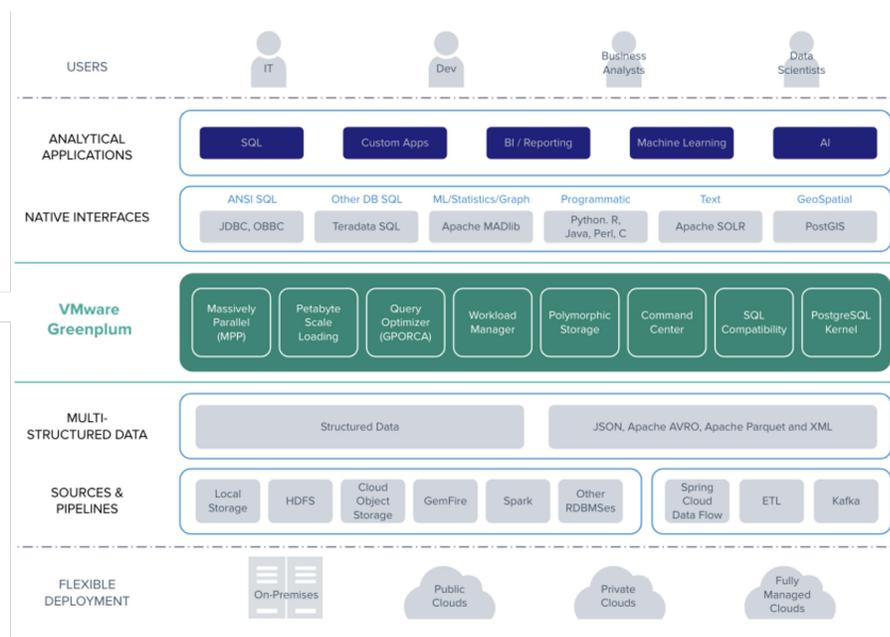
KEY BENEFITS

- ANSI and PostgreSQL compatible SQL
- Easy to migrate from SMP databases to achieve greater scale. Expand capacity from terabytes to petabytes
- Compact and efficient deployment when compared to data lakes
- High concurrency environment
- Highly available mission critical enterprise deployments
- Supports a robust suite of data science tools
- Data federations with cloud storage, Hadoop and polyglot systems

Consolidate data silos with a single, scale-out environment

Greenplum is a massively parallel database that excels managing 10's 100's and 1000's of terabytes in a single unified clustered platform. In addition to the core relational database features, Greenplum provides a series of integrations, extensions, analytic modules and extensibility methods that make Greenplum into not just a relational database but a Platform for data storage and analytics.

Despite the adoption of data lakes, enterprises continue to struggle with integrating structured with unstructured data. Traditional enterprise data warehouses are poor at delivering real-time insights. Data volume growth puts pressure on infrastructure and resources. For these challenges, VMware offers Greenplum, a massively parallel relational DBMS. Greenplum can run workloads from traditional business intelligence to on-line transaction processing to GPU accelerated deep learning. Whether it's streaming ingest, execution of point queries, data science exploration, or long-running analytic queries, Greenplum has you covered. Because Greenplum is based on PostgreSQL, data analytic teams use SQL to access its powerful features. Machine learning, geospatial queries, graph analytics, text analysis and connection into external data sources all use a familiar SQL interface.



Salient Features

MAJOR FEATURES

- MPP Shared Nothing Architecture
- Mission Critical High Availability
- Support Hybrid Transaction Processing Workloads
- High-speed Interconnect
- Rule Based Workload Management
- Parallel Query Optimizer
- Extensible Datatypes
- Federated Query Processing
- S3 SELECT Accelerated Queries
- Multi-temperature Data Storage
- Streaming Data Ingestion
- In-database Machine Learning
- Scalable Graph Analytics
- GPU Accelerated Deep Learning
- Geo-spatial Indexing, Search, and Aggregation
- Python, R, Java, and C Extensibility
- Text Search and Analysis
- Graphical DBA Administration
- Parallel and Scalable Data Backup
- Multi-cluster Data Copy
- Seamless connectivity to Apache Spark, Apache Kafka and Apache Nifi

Infrastructure-agnostic deployment minimizes lock-in

Enjoy the convenience of running Greenplum on a range of affordable options, in your data center or in the public cloud. Greenplum runs great on bare metal, virtual machines, private and public clouds.

Handle traditional BI workloads with ease

Greenplum offers comprehensive SQL-92 and SQL-99 language support with SQL 2003 OLAP extensions, including window functions, rollups, cubes and a wide range of other expressive functionality. All queries are executed in parallel across the entire cluster. Standard database interfaces (including PostgreSQL, SQL, ODBC, JDBC and OLEDB, etc.) are fully supported and certified with a wide range of business intelligence (BI) and extract, transform and load (ETL) tools.

Hybrid Storage

With the explosion of data stores and cloud services, data now resides across many disparate systems and in a variety of formats. Often, data is classified both by its location and the operations performed on the data, as well as how often the data is accessed: real-time or transactional (hot), less frequent (warm), or archival (cold). Greenplum's partition table mechanism enables users to delegate the "cold" and "hot" partitions to column and row stores respectively. From the user's perspective, they can query the table with mixed storage in Greenplum without being aware of the storage type.

SQL containerization

Greenplum Resource Groups provide resource isolation for query, multitenancy and mixed workloads. SQL containerization groups together CPU and memory resources—along with concurrent transactions—to ensure each is guaranteed a predetermined amount.

Apache MADlib offers new possibilities for deep learning

Apache MADlib now supports highly parallel, GPU-accelerated processing for deep learning. Greenplum users can take advantage of GPUs embedded in cluster hardware, achieving two orders of magnitude or faster performance from CPU-only processing.

Analyze graph, geospatial and text analytics

Greenplum allows geographic information system (GIS) objects to be stored and processed in the database via PostGIS, a spatial database extension for PostgreSQL. Pivotal GPText, based on Apache Solr, enables the processing of raw text data, including email and social media feeds, with an easy SQL interface. Graph analytics are delivered via Apache MADlib, an open-source library of graph, statistical and machine-learning functions.

Charge your analytics with Apache Spark

Apache Spark is an extremely fast, in-memory data-processing engine. The Greenplum Spark Connector provides a high-speed, bidirectional, parallelized data transfer between Greenplum and Apache Spark clusters. It enables users to run fast in-memory analytics, exploratory analytics and ETL processing with data persisted on Greenplum.

Summary

Greenplum is an open-source data analytics platform that provides powerful and rapid analytics on very large volumes of data. Uniquely geared toward machine learning and advanced data science, Greenplum delivers unmatched analytical query performance on large data volumes and tight integration with leading analytical libraries and software stacks. Additional details on Greenplum can be found at <https://www.vmware.com/products/greenplum>. An open-source version of Greenplum (Greenplum Database) is also available for download at www.greenplum.org.