

EMC GREENPLUM DATABASE

Driving the future of data warehousing and analytics

ESSENTIALS

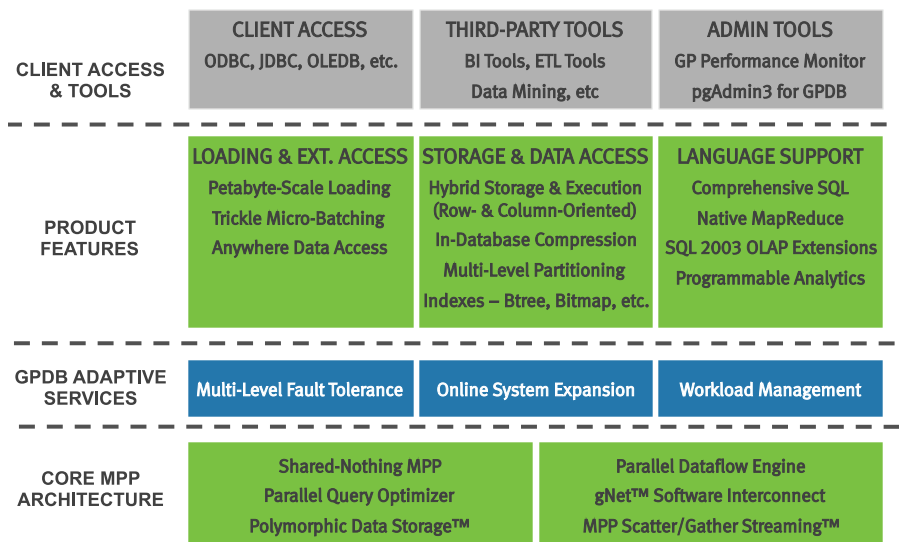
A shared-nothing, massively parallel processing (MPP) architecture supports extreme performance on commodity infrastructure

- Enables Big Data Analytics via the support of high-performance and flexible data exchange between Hadoop and Greenplum Database
- Designed with polymorphic storage (hybrid of row and column) to best fit the unique needs of each business intelligence and analytical use case
- Enables massive data storage, loading, and processing with unlimited linear scalability
- Provides automatic parallelization with no need for manual partitioning or tuning

MEETING THE CHALLENGES OF A DATA-DRIVEN WORLD

Rising IT costs, exploding data volumes, and ever-evolving competitive challenges have catalyzed new ways of thinking about effective systems for analytics. All these changes have driven radical changes in database technology, and collectively, these developments have led to a new approach for effective data exploitation. Decades-old legacy architecture for data management and analytics is inherently unfit for scaling to today’s dramatic increase in data volume.

The EMC® Greenplum® Database is a shared-nothing, massively parallel processing (MPP) architecture that has been designed for business intelligence and analytical processing. In this architecture, each server node acts as a self-contained database management system that owns and manages a distinct portion of the overall data. The system automatically distributes data and parallelizes query workloads across all available hardware.



The figure shown summarizes why the EMC Greenplum Database is the best platform on the market for mission-critical analytics. The core shared-nothing MPP architecture enables massive data storage, loading, and processing with unlimited linear scalability. Adaptive services provide worldwide enterprises with high availability, workload management, and online expansion of capacity. Key product features enable petabyte-scale loading, hybrid storage (row or column) to best fit the unique needs of each analytical use case, and embedded support for SQL, MapReduce, and programmable analytics. In addition, all major third-party analytic and administration tools are supported through standard client interfaces.

The core principle of the EMC Greenplum Database is to move the processing dramatically closer to the data and its users. This effectively enables the computational resources to process every query in a fully parallel manner, use all storage connections simultaneously, and flow data efficiently between resources as the query plan dictates. The result is that a wide variety of complex processing can be pushed down in close proximity to the data for maximum efficiency and incredible expressiveness.

The EMC Greenplum Database is already regarded as the most scalable mission-critical analytical database and in use at more than 200 leading enterprises worldwide.

GREENPLUM DATABASE FEATURES

DATABASE ARCHITECTURE

Core Massively Parallel Processing Architecture

The Greenplum Database architecture provides automatic parallelization of data and queries—all data is automatically partitioned across all nodes of the system, and queries are planned and executed using all nodes working together in a highly coordinated fashion.

Petabyte-Scale Loading

High-performance loading uses MPP Scatter/Gather Streaming™ technology. Loading speeds scale with each additional node to greater than 10 terabytes per hour, per rack. When loading a continuous stream, trickle micro-batching and re-usable table objects enable data to be loaded at frequent intervals (e.g., every five minutes), while maintaining extremely high data ingest rates.

Polymorphic Data Storage and Execution

Using Greenplum's Polymorphic Data Storage™ technology, the DBA can select the storage, execution, and compression settings that suit the way that table will be accessed. With this feature, customers have the choice of row- or column-oriented storage and processing for any table or partition. In addition, Greenplum Database also supports the placement of data on specific storage types, such as SSD media or network-attached storage (NAS) archival stores.

Anywhere Data Access

Anywhere data access enables queries to be executed from the database against external data sources, returning data in parallel regardless of location, format, or storage medium.

In-Database Compression

In-database compression uses industry-leading compression technology to increase performance and dramatically reduce the space required to store data. Customers can expect to see a three- to 10-time disk space reduction with a corresponding increase in effective I/O performance.

Multi-level Partitioning

Flexible partitioning of tables is based on date, range, or value. Partitioning is specified using DDL and enables an arbitrary number of levels. The query optimizer will automatically prune unneeded partitions from the query plan.

Dynamic Partitioning Elimination and Query Memory Optimization

Greenplum Database supports dynamic partition elimination and query memory optimization. Dynamic Partition Elimination disregards irrelevant partitions in a table and allows for significant reduction in the amount of data scanned and results in faster query execution times. The query memory optimization feature intelligently frees and reallocates memory to different operators during query processing, allowing for better memory utilization, higher throughput, and higher concurrency.

IN-DATABASE ANALYTICS

Native MapReduce

Greenplum natively runs MapReduce programs within its parallel engine and supports PL/Java, optimized C, and Java function support.

High-Performance gNet for Hadoop

Greenplum Database enables high-performance parallel import and export of compressed and uncompressed data from Hadoop clusters using gNet for Hadoop, a parallel communications transport with the industry's first direct query interoperability between Greenplum Database nodes and corresponding Hadoop nodes. To further streamline resource consumption during load times, custom-format data (binary, Pig, Hive, etc.) in Hadoop can be converted to GPDB Format via MapReduce, and then imported into Greenplum Database. This is a high-speed direct integration option that provides an efficient and flexible data exchange between Greenplum Database and Hadoop. gNet for Hadoop is available for both Greenplum HD Community Edition and Enterprise Edition.

Advanced Analytical Functions

Greenplum Database provides analytical functions (t-statistics, p-values, and Naïve Bayes) for advanced in-database analytics. These functions provide the needed metrics for variable selection to improve the quality of a regression model, as well as enhance the ability to understand and reason about the edge cases.

Programmable Analytics

A new level of parallel analysis capabilities for mathematicians and statisticians and support for R, linear algebra, and machine learning primitives is offered.

Greenplum Database Extension Framework and Turnkey In-Database Analytics

Greenplum Database delivers an agile, extensible platform for in-database analytics, leveraging the system's massively parallel architecture. Greenplum Database enables turnkey in-database analytics via Greenplum Extensions, which can be downloaded from EMC Subscribenet and installed using the new Greenplum Package Manager. This new Greenplum Database utility ensures automatic installation and updates of functional extensions like in-database GeoSpatial functions, PL/R, PL/Java, PL/Python, and PL/Perl. Greenplum Extensions dramatically simplify the task of enabling and managing advanced in-database functionality across a cluster. For example, extensions automatically get deployed on new nodes during expansions of Greenplum clusters.

DATABASE MANAGEMENT TOOLS

Online System Expansion

You can add servers to increase storage capacity, processing performance, and loading performance. The database can remain online and fully available while the expansion process takes place in the background. Performance and capacity increase linearly as servers are added.

Workload Management

With administrative control over system resources and their allocation to queries, users can be assigned to resource queues that manage the inflow of work to the database. Workload management also enables priority adjustment of running queries.

Dynamic Query Prioritization

Greenplum's Advanced Workload Management is extended with patent-pending technology that provides continuous realtime balancing of the resources of the entire cluster across all running queries. This gives DBAs the controls they need to meet workload service-level agreements in complex, mixed-workload environments.

Database Performance Monitor Tool

The Greenplum Database's Performance Monitor data collection agents gather metrics to help administrators analyze network patterns of Greenplum Database. Collecting these metrics allows system administrators to pinpoint the cause of network issues, and separate hardware issues from software issues.

Simple and Fast Parallel Installation

The parallel installation utility allows system administrators to install the Greenplum Database software on multiple hosts at once. When run as root, it also automates other system configuration tasks such as creating the Greenplum system user (gpadmin), setting the system user's password, setting the ownership of the Greenplum Database installation directory, and exchanging ssh keys between all specified host address names.

HIGH-AVAILABILITY, BACKUP, AND DISASTER RECOVERY SUPPORT

Self-Healing Fault Tolerance

Traditional MPP database fault-tolerance techniques were suitable for environments with less than 100 servers, but TCO has increased dramatically beyond that scale. Greenplum's fault-tolerance capabilities provide intelligent fault detection and fast online differential recovery, lowering TCO and enabling cloud-scale systems with the highest levels of availability.

Post-Recovery, Online Segment Rebalancing

After segment recovery, the EMC Greenplum Database segments can be rebalanced while the system is online. All client sessions remain connected to allow no downtime. The database remains functional while the system is recovered back into an optimal state.

Simpler, Scalable Backup with Data Domain Boost

Greenplum Database now includes advanced integration with EMC Data Domain[®] deduplication storage systems via EMC Data Domain Boost for faster, more efficient backup. This integration distributes parts of the deduplication process to Greenplum Database servers, enabling them to send only unique data to the Data Domain system. This dramatically increases aggregate throughput, reduces the amount of data transferred over the network, and eliminates the need for NFS mount management.

INTEROPERABILITY

Indexes—B-Tree, Bitmap, and More

Greenplum Database supports a range of index types, including B-Tree and Bitmap.

Comprehensive SQL Support

Greenplum Database offers comprehensive SQL-92 and SQL-99 support with SQL 2003 OLAP extensions and full standard support, including window functions, rollup, cube, and a wide range of other expressive functionality. All queries are parallelized and executed across the entire system.

Greenplum Database offers enhanced SQL support, including native support of 20+ Oracle functions, correlated sub queries, non-recursive WITH clause, and fixed format loader. These enhancements streamline support of third-party tools that generate such queries and make migration to Greenplum Database faster and simpler.

Client Access and Third-Party Tools

Greenplum supports standard database interfaces (PostgreSQL, SQL, ODBC, JDBC, OLEDB, etc.) and is fully supported and certified by a wide range of business intelligence (BI) and extract/ transform/load (ETL) tools.

pgAdmin3 for GPDB

pgAdmin3 is the most popular and feature-rich Open Source administration and development platform for PostgreSQL. Greenplum Database ships with an enhanced version of pgAdmin3 that has been extended to work with Greenplum Database and provides full support for Greenplum-specific capabilities.

XML Support

Greenplum Database provides support for XML, enabling high-performance, parallel load of XML documents into the database, support for XML data type and the XML Path language (xpath).

MAXIMIZE EMC GREENPLUM DCA BENEFITS WITH EMC GLOBAL SERVICES

EMC delivers a full complement of services for EMC Greenplum hardware and software to ensure that your system performs as expected in your environment, while minimizing risk to your business and budget. Expert planning, design, and implementation services help you quickly realize the value of your hardware and software in your environment—no matter how simple or complex. After implementation, EMC's data migration services can help you plan, design, and safely migrate your critical data over any distance to your new system. EMC will also help you integrate your new system into your information architecture and business intelligence and analytics applications (for example: SAS, MicroStrategy, Business Objects, Tableau, etc.), and manage your new environment when it is complete.

Extensively trained professional services personnel and project management teams, leveraging EMC's extensive data warehousing/business intelligence deployment best practices and guided by our proven methodology, accelerate the business results you need without straining the resources you have.

CONTACT US

For more information, contact your EMC Greenplum representative or visit www.greenplum.com.

EMC², EMC, Data Domain, Greenplum, and the EMC logo are registered trademarks or trademarks of EMC Corporation in the United States and other countries. All other trademarks used herein are the property of their respective owners. © Copyright 2011 EMC Corporation. All rights reserved. Published in the USA. 11/11 Data Sheet H8995