

# EMC GREENPLUM DATA COMPUTING APPLIANCE

## Driving the future of data warehousing

### ESSENTIALS

- Purpose-built, highly scalable data warehousing hardware and software that architecturally integrates Greenplum Database, Greenplum HD, and third-party applications as well as, compute, storage, and network into an enterprise-class, easy-to-implement system
- Optimized for fast query execution, unmatched data loading, and linear scalability
- Advanced “all-in-one” modular analytics solution for managing structured data, unstructured data, and ETL or BI processes
- Single platform for data warehousing, data marts, text mining, and statistical computing
- Enables greater insight and value from data with advanced analytics and unified data access
- Enterprise-level high-availability, storage, and disaster recovery using existing EMC solutions

### MEETING THE CHALLENGES OF A DATA-DRIVEN WORLD

Rising IT costs, exploding data volumes, and ever-evolving competitive challenges have spurred new ways of thinking about effective systems for data analytics. All these developments have led to radical changes in database technology and a new approach to exploiting data.

Decades-old legacy architecture for data management and analytics is inherently unfit for scaling today’s big data volumes. The EMC® Greenplum® Data Computing Appliance (DCA) offers the power of a massively parallel processing (MPP) architecture, while delivering the fastest data-loading rate and the best price/performance ratio in the industry—without the complexity and constraints of proprietary hardware. It is a purpose-built, highly scalable, parallel data analytics appliance that architecturally integrates database, compute, storage, and network into an easy-to-implement, enterprise-class system.

The Greenplum DCA is a unified big data analytics appliance—a modular solution for structured data, unstructured data, and Greenplum partner applications such as business intelligence (BI) and extract, transform, and load (ETL). Enterprises can start with a single, primary rack, which includes a Greenplum Database Module (Standard or High-Capacity) and expand the appliance in quarter-rack increments using Greenplum Database Standard Module, Greenplum Database High Capacity Module, Greenplum HD Module, or Greenplum Data Integration Accelerator Module in any order and amount, up to six racks total, as their demand for processing capacity grows. All modules are linked via a high-speed, high-performance, low-latency interconnect.

With the Greenplum DCA, your organization can embrace big data analytics quickly and easily. You can get results sooner by using an integrated appliance that offers optimized performance, ease of deployment, increased system monitoring and manageability, and a reduced footprint. The Greenplum DCA modules greatly simplify the expansion of capacity and performance of the Greenplum Database (analytic database) and Greenplum HD (Hadoop) portions of the systems. This data management appliance delivers maximum flexibility and scalability for organizations that are tackling terabyte- to petabyte-scale data opportunities.

### DATA COMPUTING APPLIANCE FEATURES

#### EXTREME AND PREDICTABLE PERFORMANCE WITH ELASTIC SCALABILITY

At the heart of the Greenplum Data Computing Appliance (DCA) is the Greenplum Database, with a shared-nothing, massively parallel processing (MPP) architecture that has been designed for business intelligence and analytical processing. The core principle of the



The EMC Greenplum DCA provides an ideal blend of flexibility, price, and performance, helping companies prevent delays in deploying actionable, intelligent applications for big data analytics.

Greenplum Database is to move processing dramatically closer to the data and its users. This effectively enables 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 a wide variety of complex processing that can be pushed down in close proximity to the data for maximum processing efficiency and unparalleled expressiveness.

### **SCATTER/GATHER STREAMING FOR DATA LOADING**

The Greenplum DCA manages the flow of data into all nodes of the appliance using the EMC Greenplum MPP Scatter/Gather Streaming™ (SG Streaming) technology. The system uses a parallel-everywhere approach to loading in which data flows from one or more source systems to every node of the database without any sequential choke points. As a result, the Greenplum DCA achieves loading speeds of more than 10 terabytes per hour per rack—two to five times faster than other appliance solutions.

### **CONFIGURATION, MASTER SERVERS, AND SEGMENT SERVERS**

You can expand the Greenplum DCA cluster by connecting up to six total cabinets with automatic data distribution and greater performance for analyst queries. Each primary rack contains two master servers and four segment servers. In a multi-rack configuration, the expansion racks do not have master servers. The master servers, as part of the Greenplum Database, are responsible for authentication, optimizing the query, balancing the workload among the different segment servers, and managing the fault tolerance mechanism of data.

### **ENTERPRISE HIGH AVAILABILITY**

The Greenplum DCA meets the reliability requirements of the most mission-critical enterprises by delivering multi-level, self-healing fault tolerance, which includes automated failover, fully online self-healing resynchronization, and multiple levels of redundancy and integrity checking. Data availability consists of a hardware RAID protection at the disk level, as well as data mirroring between the different segment servers. This system reliability ensures no data loss when a disk or server goes down.

### **RAPID DEPLOYMENT AND PREDICTABLE PERFORMANCE**

The Greenplum DCA is a purpose-built, open systems data appliance that architecturally integrates database, server, and storage into a single, easy-to-implement system that can be deployed and expanded in days—not weeks or months. You can expand the system in module increments to multi-rack. Appliance integration and pre-tuning ensures predictable performance while dramatically simplifying your data warehouse and analytics infrastructure—reducing your administrative overhead.

### **RELIABLE BACKUP AND DISASTER RECOVERY**

The Greenplum DCA uses EMC Data Domain® and EMC Symmetrix® to ensure robust and reliable remote data protection for the DCA data analytics environment. With EMC Data Domain's deduplication and backup technology, the Greenplum DCA can achieve fast, reliable data recovery with backup throughput speeds up to 14 TB/hour. Data Domain wide-area replication has also been qualified to remotely replicate a Greenplum database. The Greenplum DCA SAN mirror solution uses EMC Symmetrix VMAX™, EMC TimeFinder®/Snap, and Symmetrix Remote Data Facility (SRDF®) for advanced storage and data replication between two sites in synchronous mode.

#### GREENPLUM DB STANDARD MODULE

- Best price/performance ratio in the industry
- Supports linear scalability

#### GREENPLUM DB HIGH CAPACITY MODULE

- Can host multiple petabytes of data without taking up additional space, surging power consumption, or increasing costs
- Best price-per-unit data warehouse module

#### GREENPLUM HD MODULE

- The industry's first high-performance data co-processing Hadoop appliance
- Allows co-processing of structured and unstructured data

#### GREENPLUM DIA MODULE

- To host and to provide fast integration for partner analytics applications to Greenplum Data Computing

#### PROACTIVE EMC ONE SUPPORT STRUCTURE

EMC Customer Support Services provide the resources and services to quickly and proactively resolve solution-related issues and questions. This ensures business continuity and a highly available data environment. EMC's global maintenance and support is available around-the-clock via 24x7 online support tools, including live chat and online service request management, live telephone support, and onsite support through the industry's leading global field service organization.

In addition, the Greenplum DCA is enabled with Secure Remote Support (dial-home). Through this feature, the appliance provides around-the-clock remote and pre-emptive troubleshooting by automatically alerting the EMC Support Center of critical hardware and software errors. The EMC Support Center then remotely diagnoses the issue to prevent or shorten system downtime, and automatically dispatches customer engineers to accelerate hardware problem resolution.

#### DATA COMPUTING APPLIANCE MODULES

The Greenplum Data Computing Appliance (DCA) modules include:

- Greenplum Database Standard Module—A purpose-built, highly scalable data-analytics appliance module that architecturally integrates database, computing, storage, and network into an enterprise-class, easy-to-implement system. This module is the industry leader in price and performance.
- Greenplum Database High Capacity Module—A module designed to host multiple petabytes of data without taking up additional space, surging power consumption, or increasing costs. For businesses that require detailed analysis of extremely large amounts of data or those looking for a longer-term archive, this model offers the lowest cost-per-unit data warehouse.
- Greenplum HD Module—The world's first high-performance data co-processing Hadoop appliance module. The DCA marries Hadoop with the Greenplum Database, allowing the co-processing of both structured and unstructured data within a single, seamless solution.
- Greenplum Data Integration Accelerator (DIA) Module—A module designed to host and provide fast integration for partner analytics applications to the Greenplum Data Computing Appliance. For example, it is used to solve the challenges of data loading in a parallel and scalable model, to shorten batch loads or to implement micro-batch loading.

This table summarizes technical details of the four Greenplum DCA modules:

Module Type	Greenplum DB Standard Module	Greenplum DB High Capacity Module	Greenplum HD Module	Greenplum DIA Module
Software	Greenplum Database	Greenplum Database	Greenplum HD CE	Certified Partner Software
Segment Server	2 sockets/12 cores			
Total Memory	192 GB			
Storage Type	600 GB	2 TB	2 TB	2 TB
Total Number of Storage Drives	48			
Usable Capacity (uncompressed)	9 TB	31 TB	28 TB	70 TB
Usable Capacity (compressed)	36 TB	124 TB	112 TB	Not Applicable

Sample configurations of Greenplum DCA cluster with Greenplum Database and Greenplum Database High Capacity modules:

Module Type	GP DB Standard Module		GP DB High Capacity Module	
Number of Modules	4	24	4	24
Number of Racks	1	6	1	6
Usable Capacity (uncompressed)	36 TB	216 TB	124 TB	744 TB
Usable Capacity (compressed)	144 TB	864 TB	496 TB	2,976 TB
Scan Rate	24 GB/Sec	144 GB/Sec	14 GB/Sec	84 GB/Sec
Data Load Rate	10 TB/Hour	60 TB/Hour	10 TB/Hour	60 TB/Hour

## MAXIMIZE EMC GREENPLUM DCA BENEFITS WITH EMC GLOBAL SERVICES

EMC Global Services delivers a full range of support and services for EMC Greenplum hardware and software to ensure that your system will perform 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—no matter how simple or complex your environment may be. 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 as well as your business intelligence and analytics applications (such as SAS, Informatica, Micro Strategy, Business Objects, and Tableau), and will help you manage your new environment when it is complete.

Leveraging EMC’s comprehensive data warehousing and business intelligence deployment best practices, and guided by our proven methodology, our extensively trained professional services personnel and project management teams accelerate your business results without straining your resources.

## EMC GREENPLUM’S DATA COMPUTING PRODUCTS DIVISION

EMC’s Data Computing Products Division is driving the future of data warehousing and analytics with breakthrough products such as EMC Greenplum HD, EMC Greenplum Data Computing Appliance, EMC Greenplum Database, and EMC Greenplum Chorus—the industry’s first Enterprise Data Cloud platform. The division’s products embody the power of open systems, cloud computing, virtualization, and social collaboration—enabling global organizations to gain greater insight and value from their data than ever before.

### CONTACT US

To learn more about how EMC products, services, and solutions can help solve your business and IT challenges, contact your local representative or authorized reseller—or visit us at [www.EMC.com](http://www.EMC.com).

EMC<sup>2</sup>, EMC, Data Domain, EMC Greenplum, EMC Greenplum MPP Scatter/Gather Streaming, SRDF, Symmetrix, TimeFinder, VMAX, 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. 9/11 Solution Overview H7419.5