MEETING THE CHALLENGES OF A DATA-DRIVEN WORLD

Exploding data volumes, new data types and ever-growing competitive challenges have led to radical changes in analytical technologies and a new approach to exploiting data.

Decades-old legacy architectures for data management have reached scale limitations that make them unfit for analyzing big data. The fast-growing data assets, broad diversity in data type and structure, and the need for complex mathematics to unlock value from these data assets have overwhelmed traditional architectures.

The Pivotal™ Data Computing Appliance (DCA) is an integrated analytics platform that accelerates analysis of Big Data assets within a single integrated appliance. Integrating Pivotal Greenplum Database for analytics-optimized SQL on structured data, Pivotal’s HD distribution (Hadoop) and partner analytics, BI and ETL applications provides flexibility and expands the range of capabilities available to users. Delivery as a pre-configured appliance assures rapid deployment, simplified administration and industry-leading TCO.

MODULAR DESIGN FOR SCALABILITY & FLEXIBILITY

Designed as a modular platform, DCAs can be scaled at any time by adding new modules. Adding modules provides linear scalability of storage and compute capacity for database and Hadoop capabilities, while adding new module types expands the range of functionality available in a single appliance. DCAs can be initially configured with SQL database capabilities or Hadoop capabilities or both, and can be expanded after deployment by adding modules to support additional Hadoop, SQL, analytics, BI and ETL capabilities.

Combining the structured data processing of SQL in Pivotal Greenplum Database with the unstructured data analysis capabilities of Pivotal’s HD distribution (Hadoop) delivers maximum flexibility and scalability for organizations that require fast analysis of diverse, terabyte- to petabyte-scale data sets.

### AT-A-GLANCE

- Purpose-built, high-performance big data analytics appliance.
- SQL and Hadoop capabilities to analyze a variety of structured and unstructured data in a single appliance.
- Modular design provides un-equalled flexibility and scalability for integrating Pivotal Greenplum Database, Pivotal’s HD distribution (Hadoop) and partner analytics, BI and ETL applications.
- MPP Architecture accelerates execution through transparent parallelism.
- Pre-configured to accelerate deployment.
- Simplified management tools to reduce system administration burdens.

### OVERVIEW

MEETING THE CHALLENGES OF A DATA-DRIVEN WORLD

Exploding data volumes, new data types and ever-growing competitive challenges have led to radical changes in analytical technologies and a new approach to exploiting data.

Decades-old legacy architectures for data management have reached scale limitations that make them unfit for analyzing big data. The fast-growing data assets, broad diversity in data type and structure, and the need for complex mathematics to unlock value from these data assets have overwhelmed traditional architectures.

The Pivotal™ Data Computing Appliance (DCA) is an integrated analytics platform that accelerates analysis of Big Data assets within a single integrated appliance. Integrating Pivotal Greenplum Database for analytics-optimized SQL on structured data, Pivotal’s HD distribution for Hadoop-based processing of unstructured data and Pivotal partner analytics, BI and ETL applications provides flexibility and expands the range of capabilities available to users. Delivery as a pre-configured appliance assures rapid deployment, simplified administration and industry-leading TCO.

MODULAR DESIGN FOR SCALABILITY & FLEXIBILITY

Designed as a modular platform, DCAs can be scaled at any time by adding new modules. Adding modules provides linear scalability of storage and compute capacity for database and Hadoop capabilities, while adding new module types expands the range of functionality available in a single appliance. DCAs can be initially configured with SQL database capabilities or Hadoop capabilities or both, and can be expanded after deployment by adding modules to support additional Hadoop, SQL, analytics, BI and ETL capabilities.

Combining the structured data processing of SQL in Pivotal Greenplum Database with the unstructured data analysis capabilities of Pivotal’s HD distribution (Hadoop) delivers maximum flexibility and scalability for organizations that require fast analysis of diverse, terabyte- to petabyte-scale data sets.

### At-A-Glance

- Purpose-built, high-performance big data analytics appliance.
- SQL and Hadoop capabilities to analyze a variety of structured and unstructured data in a single appliance.
- Modular design provides un-equalled flexibility and scalability for integrating Pivotal Greenplum Database, Pivotal’s HD distribution (Hadoop) and partner analytics, BI and ETL applications.
- MPP Architecture accelerates execution through transparent parallelism.
- Pre-configured to accelerate deployment.
- Simplified management tools to reduce system administration burdens.

### Overview

MEETING THE CHALLENGES OF A DATA-DRIVEN WORLD

Exploding data volumes, new data types and ever-growing competitive challenges have led to radical changes in analytical technologies and a new approach to exploiting data.

Decades-old legacy architectures for data management have reached scale limitations that make them unfit for analyzing big data. The fast-growing data assets, broad diversity in data type and structure, and the need for complex mathematics to unlock value from these data assets have overwhelmed traditional architectures.

The Pivotal™ Data Computing Appliance (DCA) is an integrated analytics platform that accelerates analysis of Big Data assets within a single integrated appliance. Integrating Pivotal Greenplum Database for analytics-optimized SQL on structured data, Pivotal’s HD distribution for Hadoop-based processing of unstructured data and Pivotal partner analytics, BI and ETL applications provides flexibility and expands the range of capabilities available to users. Delivery as a pre-configured appliance assures rapid deployment, simplified administration and industry-leading TCO.

MODULAR DESIGN FOR SCALABILITY & FLEXIBILITY

Designed as a modular platform, DCAs can be scaled at any time by adding new modules. Adding modules provides linear scalability of storage and compute capacity for database and Hadoop capabilities, while adding new module types expands the range of functionality available in a single appliance. DCAs can be initially configured with SQL database capabilities or Hadoop capabilities or both, and can be expanded after deployment by adding modules to support additional Hadoop, SQL, analytics, BI and ETL capabilities.

Combining the structured data processing of SQL in Pivotal Greenplum Database with the unstructured data analysis capabilities of Pivotal’s HD distribution (Hadoop) delivers maximum flexibility and scalability for organizations that require fast analysis of diverse, terabyte- to petabyte-scale data sets.
PERFORMANCE ARCHITECTURE
The Pivotal DCA employs a massively parallel processing (MPP) architecture for fast SQL and MapReduce processing, plus the fastest data loading rates in the industry—without the complexity and constraints of proprietary hardware. DCAs are purpose-built for analytics, and provide scalable computation, storage and interconnect, delivered as a pre-configured appliance.

DATA COMPUTING APPLIANCE FEATURES
EXTREME ANALYTICAL PERFORMANCE
At the heart of the Pivotal Data Computing Appliance (DCA) are the Pivotal Greenplum Database, a shared-nothing, massively parallel SQL relational MPP database optimized for analytical and business intelligence (BI) processing and Pivotal’s HD distribution, a supported commercial distribution of Apache Hadoop. The core principle of the Pivotal DCA is to move SQL and MapReduce processing dramatically closer to the data, running analytics in parallel atop an MPP architecture while flowing data efficiently between resources as needed. The result is industry-leading performance for big data analysis at an affordable TCO.

The Pivotal DCA provides an ideal blend of flexibility, price and performance, helping companies prevent delays in deploying actionable, intelligent applications for Big Data Analytics.

INDUSTRY-LEADING FLEXIBILITY AND SCALABILITY
Pivotal DCAs are configured to match users’ needs, delivered ready to run and deployable within a few hours. Capabilities can be customized to the user’s needs, choosing from modules that support database, Hadoop and partner applications, such as analytics, data visualization, machine learning, BI and ETL.

Once deployed, DCA capacity can be scaled linearly by adding modules. New functions can also be added by adding new module types. DCAs can be configured from ¼ rack to 12 racks supporting a vast range of configurations by mixing module types.

COHERENT APPLIANCE-WIDE ADMINISTRATION
DCAs are easily administered regardless of configuration as all modules are managed and monitored through Pivotal Command Center, an appliance-wide administration tool. Integration with SNMP Network Management systems helps DCAs fit easily into most data center management frameworks.

PIVOTAL GREENPLUM DATABASE FEATURES
ANALYTICS-OPTIMIZED SQL PROCESSING
Database modules in the DCA run Pivotal Greenplum Database, an SQL database that has been completely optimized for analytical processing. By installing available algorithms that have been redesigned to run in parallel, Pivotal Greenplum Database provides 10x to 100x faster execution of statistical and analytical algorithms than traditional SQL databases.

INDUSTRY-LEADING PARALLEL DATA LOADING
The Pivotal Greenplum Database ingests data in parallel, unlike competing appliances, and achieves two to five times faster rates for loading and for Hadoop data ingest. Load and ingest rates scale linearly with system size, making Pivotal DCAs the industry leaders for data ingest.

LINEAR SCALABILITY OF PIVOTAL GREENPLUM DATABASE USING DCA MODULES
Pivotal Greenplum Database is designed for the Massively Parallel Processing (MPP) environment provided by the DCA. MPP allows users to linearly scale Pivotal Greenplum Database capacity, load rate and performance by adding modules. Modules are easily added, with service interruptions limited to a few minutes allowing data redistribution to be scheduled for maintenance intervals.

ENTERPRISE AVAILABILITY
The Pivotal DCA meets the reliability requirements of the most mission-critical applications by delivering multi-level, self-healing fault tolerance for storage, query processing and interconnect.

Data storage in the DCA is protected at three levels. Data is stored in RAID disk arrays that continue operating after a drive failure. Hot spare disks are provided and are swapped in automatically by RAID controllers in the event of a drive failure to avoid service interruptions. All database data is mirrored, with mirror copies of all data residing independently in the DCA. RAID, hot spares and mirrors support resynchronization processes that provide automated self-healing recovery for storage failures.
Query processing is also protected. All segment servers are redundant, protected by automatic failover to assure that a server failure does not result in a system outage. Redundant master servers with automatic failover assure that host nodes also present no single points of failure.

Fully redundant 10GB Ethernet switching and connections link segment processors and master servers. This redundant interconnect provides automatic failover eliminating single points of failure between nodes in the DCA.

RELIABLE BACKUP AND DISASTER RECOVERY

Database modules in the Pivotal DCA can be backed up using EMC Data Domain® for both back up and recovery as well as for remote disaster recovery using replicated data. EMC Data Domain’s de-duplication technology enables databases in the Pivotal DCA to achieve backup rates as fast 14 TB/hour. Once backed-up, Data Domain wide-area replication can remotely replicate Pivotal Greenplum Database data to remote sites for maintenance of warm standby systems to be used in the event of a disaster affecting the primary DCA.

PIVOTAL’S HD DISTRIBUTION (HADOOP) FEATURES

APACHE HADOOP WITH INCREASED PRODUCTIVITY

Pivotal’s HD distribution modules in the Pivotal DCA provide a fully-supported distribution of Apache Hadoop for analytics developers. Easily scaled by adding modules, HD affords DCA users the entire range of Hadoop capabilities, including Pig, Hive, HBase, Mahout, Zookeeper and other modules.

RAPID HADOOP DEPLOYMENT

Historically a platform built in-house, many Hadoop projects have faltered or been delayed by unforeseen complexities of configuring and optimizing shop-built Hadoop clusters. Pivotal’s HD distribution provides pre-configured, ready-to-run Apache Hadoop, slashing deployment time to days from weeks or perhaps months.

EFFICIENT HADOOP AND SQL INTEGRATION

Apache Hadoop users have at their fingertips a powerful platform, though Hadoop can be difficult to learn and use. Complex analysis of unstructured data on Hadoop typically requires MapReduce programming, often involving java development for even the most simplistic activities. As a result, systems designers are motivated to use Hadoop where needed, and seek SQL solutions to simplify more traditional data management tasks.

Unlike many vendors’ solutions, Pivotal integrates Hadoop and relational technologies in a single appliance. DCAs integrate Pivotal’s HD distribution modules with Pivotal Greenplum Database modules using the DCA’s high-speed interconnect, to provide fast parallel access to Hadoop HDFS data from Pivotal Greenplum Database modules.

Fast and easy integration of Hadoop and SQL cuts development effort, enhances performance, shortens schedules and provides the agility needed to adapt to changing requirements.

ENTERPRISE STORAGE OPTIONS

Pivotal’s HD distribution modules offer traditional HDFS storage based on direct-attached storage organized in RAID 0. This option will provide most users with sufficient disaster recovery protections. For mission-critical applications, Pivotal DCAs also support integration with EMC Isilon Scale-Out NAS to provide higher-reliability HDFS storage. Additional advantages include:

- Disaster recovery by using Isilon snapshots and replication to maintain geographically-diverse replicas
- Increased availability by distributing name node functionality in HDFS
- Reduced storage requirement by eliminating the 3x file replication within traditional HDFS
- Independent scaling of compute and storage resources
- Simplified loading data into HDFS using NFS or CIFS protocols to load Isilon.

DATA COMPUTING APPLIANCE UAP MODULES

Pivotal DCA UAP Edition offers an expanded and upgraded array of DCA Modules from which users can configure their desired analytics capabilities. Core modules for Pivotal Greenplum Database and Pivotal’s HD distribution (Hadoop) are described below and can be combined in a DCA:

- **Pivotal Greenplum Database Standard Module** — This module integrates Pivotal Greenplum Database with large capacity servers. Database modules can stand alone, be integrated together to scale the database and also be integrated with Hadoop and DIA modules. For those who require complex analysis of large amounts of structured data, this module offers the best cost/capacity trade-off.
- **Pivotal Greenplum Database Compute Module** — This module offers the same computational capacity and configurability as the Standard module above, with reduced storage capacity and cost. This module will suit users whose needs emphasize powerful analytical capabilities with modest storage capacity.
- **Pivotal’s HD distribution Module** — This module provides a high-performance Hadoop building block. Integrating Hadoop
software with CPUs, memory and direct-attach storage, these modules are designed to be clustered and can be integrated with Pivotal Greenplum Database modules to enable co-processing of structured and unstructured data.

- **Pivotal's HD distribution Compute Module** — Like the HD Module, this module offers Hadoop integrated with CPUs and memory, but does not include storage hardware, relying instead on integration with Isilon Scale-Out NAS for HDFS storage. HD Compute modules can be clustered to create Hadoop systems and can also integrated with Pivotal Greenplum Database for co-processing applications.

See tables below for DCA module specifications.

<table>
<thead>
<tr>
<th>Module Type</th>
<th>DB Compute Module</th>
<th>DB Standard Module</th>
<th>HD Module</th>
<th>HD Compute Module</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rackspace</td>
<td>8U (¼ rack)</td>
<td>8U (¼ rack)</td>
<td>8U (¼ rack)</td>
<td>2U</td>
</tr>
<tr>
<td>Software</td>
<td>Pivotal Greenplum Database</td>
<td>Pivotal Greenplum Database</td>
<td>Pivotal Greenplum Database</td>
<td>Pivotal’s HD Distribution</td>
</tr>
<tr>
<td>Number of Servers</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Number of CPU cores</td>
<td>64</td>
<td>64</td>
<td>64</td>
<td>32</td>
</tr>
<tr>
<td>Total Memory</td>
<td>256GB</td>
<td>256GB</td>
<td>256GB</td>
<td>128GB</td>
</tr>
<tr>
<td>Drive Type</td>
<td>300GB SAS</td>
<td>900GB SAS</td>
<td>3TB SATA</td>
<td>300GB SAS</td>
</tr>
<tr>
<td>Total Number of Storage Drives</td>
<td>96</td>
<td>96</td>
<td>48</td>
<td>6 (user data is stored in Isilon)</td>
</tr>
<tr>
<td>Usable Capacity (physical)</td>
<td>9 TB</td>
<td>275TB</td>
<td>36 TB</td>
<td>Not Applicable</td>
</tr>
<tr>
<td>Usable Capacity (User Data)</td>
<td>36 TB</td>
<td>110TB</td>
<td>144 TB</td>
<td>Not Applicable</td>
</tr>
</tbody>
</table>

Table 1: Pivotal Greenplum Database and Pivotal’s HD Distribution Module Specifications

### SYSTEM PERFORMANCE AND CAPACITY

Performance and capacity of the Pivotal DCA with Pivotal Analytic Database Compute and Pivotal Greenplum Database Standard modules scales according to the chart below:

<table>
<thead>
<tr>
<th>Module Type</th>
<th>Database Compute Module</th>
<th>Database Standard Module</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Racks</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>Number of Modules</td>
<td>4</td>
<td>48</td>
</tr>
<tr>
<td>Usable Capacity (uncompressed)</td>
<td>36 TB</td>
<td>432 TB</td>
</tr>
<tr>
<td>Usable Capacity (×4 compression)</td>
<td>144 TB</td>
<td>1728 TB</td>
</tr>
<tr>
<td>Scan Rate</td>
<td>40 GB/Sec</td>
<td>480 GB/Sec</td>
</tr>
<tr>
<td>Data Load Rate</td>
<td>16 TB/ Hour</td>
<td>TBD</td>
</tr>
</tbody>
</table>

Table 2: Pivotal Greenplum Database Capacity, Scan Rate and Load Rates

### HOSTING PARTNER APPLICATIONS

Pivotal DCAs can be extended to host partner applications to extend the functionality of Pivotal Greenplum Database and Pivotal’s HD distribution. These modules, called the Data Integration Accelerators (DIA) can be added to a DCA and provide a redundant server environment on which to host data loading, ETL analytics and BI products. DIAs are installed into the DCA and directly connected to the high-speed interconnect within the DCA to minimize floor space, maximum performance, facilitate easy administration using Pivotal Command Center.
PHYSICAL SPECIFICATIONS

Pivotal require modest power and cooling according to the table below, and simplify the task of provisioning DCAs into the data center.

<table>
<thead>
<tr>
<th>Physical Dimensions</th>
<th>DCA Quarter-Rack</th>
<th>DCA Half-Rack</th>
<th>DCA Full-Rack</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>75 in / 190 cm</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Width:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24 in / 61 cm</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Depth:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>42 in / 100 cm</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weight:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>700 lbs / 319 Kgs</td>
<td>1,100 lbs / 500 Kgs</td>
<td>1,600 lbs / 725 Kgs</td>
<td></td>
</tr>
<tr>
<td>Power VA:</td>
<td>3330</td>
<td>5440</td>
<td>9600</td>
</tr>
<tr>
<td>Cooling (BTU/HR):</td>
<td>11,300</td>
<td>18,500</td>
<td>32,700</td>
</tr>
</tbody>
</table>

Table 3: Pivotal DCA Environmental Specification for DCA Configurations

LEARN MORE

To learn more about our products, services and solutions, visit us at goPivotal.com.