EMC Greenplum Data Computing Appliance Enhances EMC IT’s Global Data Warehouse
Accelerating Big Data and Analytics

Abstract
This white paper illustrates a synergistic model for deploying EMC’s Greenplum Data Computing Appliance (DCA) with EMC IT’s incumbent Global Data Warehouse infrastructure. It illustrates the steps to migrate from EMC’s IT Global Data Warehouse, and the progress Greenplum DCA enabled EMC to make in addressing its business challenges of growing data (“Big Data”) and accelerated analytics.

August 2011
Contents

Executive Summary ......................................................................................................................... 4
  Problem Statement.......................................................................................................................... 4
  Value Statement............................................................................................................................ 4
  Audience....................................................................................................................................... 5

EMC IT’s Big Data Approach ......................................................................................................... 6
  EMC IT’s Information Reference Architecture .............................................................................. 6
  EMC IT’s Global Data Warehouse Success .................................................................................... 6
  EMC IT’s Big Data and Analytics Business Drivers ...................................................................... 7
  Greenplum Offerings .................................................................................................................... 8
  EMC Greenplum Database Software ............................................................................................ 8
  EMC Greenplum Data Computing Appliance (DCA) Product Family ........................................... 9
  Greenplum Enablers ...................................................................................................................... 9

Business POC .................................................................................................................................. 10
  Business Challenge....................................................................................................................... 10
  Business Success.......................................................................................................................... 11

EMC IT’s Greenplum POC Highlights ............................................................................................ 12
  Greenplum POC Platform Migration ............................................................................................ 12
  High-Level Migration POC Milestones ......................................................................................... 12
  How It Was Done: A POC ............................................................................................................. 12
  Greenplum DCA POC Deployment ............................................................................................... 12
  EMC IT Global Data Warehouse to Greenplum POC Migration .................................................. 14
  Migration Steps............................................................................................................................... 14

Business POC Benefits .................................................................................................................... 24
  Query Performance........................................................................................................................ 24
  New Analytics Abilities ................................................................................................................ 27
  EMC IT Global Data Warehouse to Greenplum POC: Lessons Learned ...................................... 30

Conclusion ....................................................................................................................................... 32
  Benefits......................................................................................................................................... 32
  References.................................................................................................................................... 33
Executive Summary

This white paper illustrates a synergistic model for deploying the Greenplum Data Computing Appliance (DCA) with EMC IT’s existing Global Data Warehouse (GDW) deployed on EMC IT’s Business Intelligence (BI) Grid infrastructure. It outlines the migration steps of moving the GDW to the DCA to solve business problems driven by growing data (“Big Data”), and the need for better, accelerated, and advanced analytics that could not be addressed by the previous Global Data Warehouse.

Problem Statement

As EMC has grown, so has our data. It is not unusual for us to analyze hundreds of millions of rows of data to gather insight into the health of our business.

As EMC has become more complex, so has the analysis that must be executed on this Big Data. Our business is transitioning from backward looking metrics (lagging indicators) to both backward and forward looking indicators (leading indicators), and is using statistical modeling to predict the impact that EMC’s current decisions will have on our future business model.

Until recently, EMC’s Global Data Warehouse had been struggling to meet the demand for Operational BI, Big Data, and advanced analytics. This resulted in shadow initiatives and labor-intensive processes to generate the information required by EMC’s business owners. For example, an EMC business unit data load on the legacy infrastructure took more than 6 days and created more than 650 batch reports that had to be reviewed by a business unit subject matter expert to create the needed information to do an analysis. (See Business POC section for details)

Greenplum is perfectly designed to satisfy this need for predictive analytics on Big Data. The introduction of the Greenplum DCA will reduce the need for these shadow activities, and will create a platform for EMC IT to develop and deploy an advanced analytics appliance that will serve as a platform to develop Business Intelligence as a Service (BlaaS).

Value Statement

This document illustrates the specific deployment steps and resultant benefits of the integration of the Greenplum DCA with EMC’s Global Data Warehouse, including:

The high-level deployment model of the Greenplum DCA into the existing EMC Global Data Warehouse infrastructure. This includes:

- The phased approach for moving EMC IT’s BI functions to the Greenplum DCA via the business proof of concept (POC).
- Improved data loading, performance, and functionality via the use of Greenplum DCA. Example:
• Ultra-fast loading -- The legacy infrastructure data load that took the business unit more than 6 days and 650 batch reports can be loaded in less than 29 minutes on the Greenplum DCA.

• A query running in 427 seconds on the legacy platform decreased to just 28.85 seconds on the Greenplum DCA for 99 million rows of data.

• A process that took 45 minutes on the legacy platform can now be completed in 3.46 minutes on the Greenplum DCA.

  – EMC IT increased its ability to use a tool set to create a dashboard and execute “on-the-fly” analytics instead of delayed batch reports.
  
  – The EMC / EMC IT “lessons learned” from the Greenplum DCA deployment.
  
  – The co-existence of both the EMC IT Global Data Warehouse (BI Grid) and the Greenplum DCA to address EMC’s data warehouse and BI challenges with Big Data and analytics to create a platform for Business Intelligence (BI) as a Service (BlaaS).

**Audience**

This white paper is intended for CIOs, data warehouse architects, Oracle architects, storage architects, Oracle DBAs, and server and network administrators.
EMC IT’s Big Data Approach

This section identifies the EMC Data Reference Architecture and EMC IT’s current infrastructure of the EMC IT Global Data Warehouse deployment.

EMC IT’s Information Reference Architecture

The diagram below illustrates EMC IT’s vision of the components and process for its Information Reference Architecture:

![Diagram of EMC IT’s Information Reference Architecture]

This logical architecture shows the need to have both the current Oracle Data Warehouse and the DCA be part of EMC IT’s Global Data Warehouse.

EMC IT’s Global Data Warehouse Success

This consolidated Oracle Global Data Warehouse deployed on EMC IT’s BI-Grid infrastructure (Oracle RAC) has delivered the following benefits for EMC and EMC IT (please review the Global Data Warehouse white paper in the References section for details):

- Consolidation of five data warehouses to one unified Global Data Warehouse infrastructure
- 10 times improvement in query performance from Day 1
- 180 percent improvement in batch job performance
• 50 percent reduction in data mart update times
• Two to three times performance improvement in reporting cube build times
• 200 percent improvement in dashboard rendering and drill-down performance

**EMC IT’s Big Data and Analytics Business Drivers**

Even with the greatly improved performance provided by the enterprise Oracle Data Warehouse, EMC IT was still not keeping up with:

• Data explosion
• Shadow systems sprawl
• Shrinking load windows
• Increasing complexity of the analytics required by the business
• Volumes of data required for predictive analytics and data mining
• Explosion of point solutions and shadow marts in an attempt to remediate the issues

Figure 2 is a high-level illustration of the challenges the data warehouse did not address:

![Diagram showing sources, manual analysis, slow moving data, static schemas, and Greenplum DCA]

Figure 2 illustrates the need to integrate the Greenplum DCA into the Global Data Warehouse infrastructure to address the business drivers that the Global Data Warehouse could not solve. These business drivers caused EMC IT to look at a new platform to deal with Big Data and analytics. EMC IT selected the Greenplum DCA for its ability to:
- **Analyze** data and information in real time
- **Apply** analytics and discern critical information
- **Unlock** the value of information from this Big Data

**Greenplum Offerings**

The following section gives a high-level overview of the Greenplum Database and the Greenplum DCA, and their ability to accelerate business value for EMC IT Big Data and analytics drivers.

**EMC Greenplum Database Software**

Built to support the next generation of Big Data warehousing and large-scale analytics processing, the EMC Greenplum Database manages, stores, and analyzes terabytes to petabytes of data. Users experience 10 to 100 times better performance over traditional RDBMS products – a result of the DCA’s shared-nothing MPP architecture, high-performance parallel dataflow engine, and advanced gNet software interconnect technology.

The Greenplum Database was conceived, designed, and engineered to enable customers to take advantage of large clusters of increasingly powerful, increasingly inexpensive commodity servers, storage, and Ethernet switches. EMC Greenplum customers can gain immediate benefit from deploying the latest commodity hardware innovations.

Greenplum’s shared-nothing architecture is optimal for fast queries and loads because processors are placed as close as possible to the data itself for faster operations with the maximum degree of parallelism possible.
EMC Greenplum Data Computing Appliance (DCA) Product Family
The Greenplum Data Computing Appliance family is a group of purpose-built, highly scalable, parallel data warehousing appliances that architecturally integrate database, compute, storage, and network into a single, easy-to-manage enterprise-class system. Designed for rapid analysis of data volumes scaling into petabytes, the Greenplum DCA is a powerful platform for unifying business intelligence and advanced analytics enterprise-wide. The Greenplum DCA has been proven to provide the industry’s fastest data loading capabilities, and, in turn, to rapidly deliver real business value to customers by reducing investments of money, time, and effort.

The Greenplum DCA product line is available in two models: the Standard DCA, which supports up to 144TB of compressed data; and the High Capacity DCA, which supports up to 496TB of compressed data. Regardless of your data computing requirements, there is a model in the DCA family to meet your needs. The Greenplum DCA family is offered in configurations ranging from a quarter-rack to multiple-rack appliances to achieve the maximum flexibility and scalability for organizations faced with terabyte to petabyte scale data opportunities. Because these appliances support the same state-of-the-art EMC Greenplum Database as their core, switching between the models can be effortless when a business grows and its requirements change.

The Greenplum DCA family provides a rapidly deployable, scalable, and cost-effective infrastructure so you can manage and analyze exploding data volumes while increasing performance and achieving greater business agility.

Greenplum Enablers
Greenplum is a strategic platform for EMC IT. At a high level, it provides the following benefits for EMC IT’s vision of Big Data and analytics:

- Ultra-fast loads and linear scalability
- Data aggregation with common taxonomy
- In database ad-hoc analytics
- Predictive analysis and proactive identification
Business POC
The EMC Customer Quality business unit focuses on EMC's customers, partners, and internal stakeholders, looking at quality in the broadest sense, including products, services, innovations, interactions, and processes. The following section highlights the EMC Corporate Quality business unit’s challenge using the current Oracle infrastructure.

Business Challenge

Prior to deploying the Greenplum DCA, the analytics process would begin with a 31-hour data load preparation process followed by a 4-to-5-day batch report run. At the conclusion of this process, Customer Quality would begin looking at product quality.

Customer Quality would then spend a week or more creating various reliability analyses. They would use other offline toolsets and spreadsheets, as well as many hours of labor-intensive analysis, to generate the root cause analysis of any identified issues and to predict expected future product quality.

The diagram below illustrates the process used on the EMC IT Global Data Warehouse and stand-alone reporting instance:

Figure 4: Previous Customer Quality process
**Business Success**
With hundreds of millions of records to sort through, obtaining the insight from the data using the existing platform was previously a time-consuming process. Now, with the Greenplum DCA, the Customer Quality business unit processes hundreds of millions of records in minutes. The diagram below illustrates the process used on the Greenplum DCA:

**Figure 5: New Customer Quality process with Greenplum**

With the Greenplum DCA, Customer Quality can perform quality analysis of all the products, versus only one product at a time in the Oracle environment.

Customer Quality can now drive business and customer value by using Greenplum’s speed.
EMC IT’s Greenplum POC Highlights

The following section outlines the highlights of the Greenplum POC platform migration and Greenplum POC architecture.

Greenplum POC Platform Migration

These sub-sections detail the high-level migration milestones and POC approach to migrate to the Greenplum DCA. They highlight the discovery and phased migration of a POC project to gain experience and knowledge on Greenplum, and describe how it could coexist with the existing Global Data Warehouse infrastructure.

High-Level Migration POC Milestones

The high-level migration milestones were:

- **February – March 2011**: Initial load tests run on the POC environment.
- **April – May 2011**: POC Greenplum DCA and initial procurement used to build a development (dev) and test POC environments.

Introducing the Greenplum DCA in EMC IT’s Global Data Warehouse infrastructure required planning, a knowledgeable IT staff, and business owners.

How It Was Done: A POC

EMC IT first got its knowledge and experience of deploying and migrating EMC Customer Quality data by migrating a few legacy data marts to the Global Data Warehouse infrastructures. The Customer Quality data was then migrated over to the Greenplum DCA for the POC to prove that the Greenplum DCA could address the real-world advanced analytics problem of the Customer Quality organization.

Greenplum DCA POC Deployment

The deployment of EMC IT’s Greenplum DCA for this POC included the introduction of the following hardware into the EMC IT Global Data Warehouse infrastructure:
Figure 6: EMC IT’s Greenplum DCA POC hardware layout

Table 1 shows the components of the separately deployed Greenplum appliances for the POC infrastructure:

<table>
<thead>
<tr>
<th>Components</th>
<th>GP1000 (Full Rack) DCA</th>
<th>GP100 (Half Rack) DCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Operating system</td>
<td>Red Hat Enterprise Linux 5 update 5 (RHEL 5.5).</td>
<td>Red Hat Enterprise Linux 5 update 5 (RHEL 5.5).</td>
</tr>
<tr>
<td>RDBMS Software</td>
<td>Greenplum 4.0.3</td>
<td>Greenplum 4.0.3</td>
</tr>
<tr>
<td>Segment Servers</td>
<td>16</td>
<td>8</td>
</tr>
<tr>
<td>Processors (per segment server)</td>
<td>(2) six-core (2.93GHz) 192 total cores</td>
<td>(2) six-core (2.93GHz) 96 total cores</td>
</tr>
<tr>
<td>Memory (per segment server)</td>
<td>48GB (768GB total)</td>
<td>48GB (384GB total)</td>
</tr>
<tr>
<td>Storage (hard drives per segment server)</td>
<td>(12) 600GB 15K SAS</td>
<td>(12) 600GB 15K SAS</td>
</tr>
<tr>
<td>Usable Capacity (uncompressed)</td>
<td>36TB</td>
<td>18TB</td>
</tr>
<tr>
<td>Usable Capacity (compressed)</td>
<td>144TB</td>
<td>73TB</td>
</tr>
<tr>
<td>Master Servers</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Processors (per master server)</td>
<td>(2) six-core X5680 (3.33 GHz) 12 total cores</td>
<td>(2) six-core X5680 (3.33 GHz) 12 total cores</td>
</tr>
</tbody>
</table>
Memory (per master server) & 48GB & 48GB \\
Storage (hard drives per master server) & (6) 600GB 10K SAS & (6) 600GB 10K SAS \\
Scan Rate & 24GB/s & 12GB/s \\
Data Load Rate & 10TB/Hr & 5TB/Hr \\

<table>
<thead>
<tr>
<th>Object Types</th>
<th>Number of Objects</th>
</tr>
</thead>
<tbody>
<tr>
<td>TABLES</td>
<td>3,634</td>
</tr>
<tr>
<td>TABLE PARTITIONS</td>
<td>485</td>
</tr>
<tr>
<td>TABLE SUBPARTITIONS</td>
<td>732</td>
</tr>
<tr>
<td>INDEXES</td>
<td>4,424</td>
</tr>
<tr>
<td>INDEX PARTITIONS</td>
<td>5,823</td>
</tr>
<tr>
<td>INDEX SUBPARTITIONS</td>
<td>1,924</td>
</tr>
<tr>
<td>MATERIALIZED VIEWS</td>
<td>34</td>
</tr>
<tr>
<td>LOB SEGMENTS</td>
<td>27</td>
</tr>
<tr>
<td>TRIGGERS</td>
<td>1,492</td>
</tr>
<tr>
<td>PACKAGES</td>
<td>70</td>
</tr>
<tr>
<td>PACKAGE BODY</td>
<td>55</td>
</tr>
<tr>
<td>PROCEDURES</td>
<td>194</td>
</tr>
</tbody>
</table>

Table 1: Greenplum DCA POC components

EMC IT Global Data Warehouse to Greenplum POC Migration

The following section details the method EMC IT used to migrate the EMC Global Data Warehouse to the Greenplum DCA platform.

Migration Steps

Here are the steps for migrating EMC IT’s Oracle Global Data Warehouse to the EMC IT POC Greenplum DCA:

Step 1. Assess the current Oracle environment.

1. Prepare a list of objects of the Oracle Global Data Warehouse (GDW)

The total database size in Oracle is 15TB.

Below are all application-related objects (counts) in the current Oracle environment.
2. Find the objects that do not need to be migrated to Greenplum:
   a. Materialized views
   b. Indexes – due to MPP architecture performance gains, indexes will not be needed

   Test to see if it is necessary to migrate certain objects, such as materialized views and indexes.

3. Prepare a list of schemas, with the size of the objects to be migrated.
4. Prepare the list of objects to be migrated to Greenplum.

<table>
<thead>
<tr>
<th>Object type</th>
<th>Count</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>TABLES and data</td>
<td>3,634</td>
<td>Tables and its data to be migrated</td>
</tr>
<tr>
<td>TABLE PARTITIONS</td>
<td>485</td>
<td>Created two versions: non-partitioned and then partitioned</td>
</tr>
<tr>
<td>TABLE SUBPARTITIONS</td>
<td>732</td>
<td>Tested all the options</td>
</tr>
<tr>
<td>INDEXES</td>
<td>4,424</td>
<td>Not migrated</td>
</tr>
<tr>
<td>INDEX PARTITIONS</td>
<td>5,823</td>
<td>Not migrated</td>
</tr>
<tr>
<td>INDEX SUBPARTITIONS</td>
<td>1,924</td>
<td>Not migrated</td>
</tr>
<tr>
<td>MATERIALIZED VIEWS</td>
<td>34</td>
<td>Not migrated</td>
</tr>
<tr>
<td>TRIGGERS</td>
<td>1,492</td>
<td>Not part of the POC</td>
</tr>
<tr>
<td>PACKAGES</td>
<td>70</td>
<td>Not part of the POC</td>
</tr>
<tr>
<td>PACKAGE BODY</td>
<td>55</td>
<td>Not part of the POC</td>
</tr>
<tr>
<td>PROCEDURES</td>
<td>194</td>
<td>Not part of the POC</td>
</tr>
<tr>
<td>SEQUENCES</td>
<td>720</td>
<td>Required for incremental loading. Not part of this POC</td>
</tr>
</tbody>
</table>

Table 3: Oracle Migrate/No Migrate object list

Step 2. Assess data types in the Oracle database.

Below are the data types and the number of columns for each.
### Data types

<table>
<thead>
<tr>
<th>Data types</th>
<th>Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>VARCHAR2</td>
<td>74,296</td>
</tr>
<tr>
<td>NUMBER</td>
<td>38,393</td>
</tr>
<tr>
<td>DATE</td>
<td>16,999</td>
</tr>
<tr>
<td>CHAR</td>
<td>845</td>
</tr>
<tr>
<td>FLOAT</td>
<td>564</td>
</tr>
<tr>
<td>TIMESTAMP</td>
<td>46</td>
</tr>
<tr>
<td>NCHAR</td>
<td>1</td>
</tr>
<tr>
<td>NVARCHAR2</td>
<td>25</td>
</tr>
</tbody>
</table>

**Table 4: Oracle Data Types**

1. Find the right equivalent data type available in Greenplum. Below are the different data types that are part of the tables to be migrated and their equivalent Greenplum (postgres) data types.

<table>
<thead>
<tr>
<th>Data types</th>
<th>Greenplum equivalent data type</th>
</tr>
</thead>
<tbody>
<tr>
<td>VARCHAR2, NVARCHAR2</td>
<td>VARCHAR</td>
</tr>
<tr>
<td>NUMBER</td>
<td>NUMERIC, NUMERIC (p,s), SMALLINT (2 bytes), INTEGER (4 bytes), BIGINT (8 bytes),</td>
</tr>
<tr>
<td>DATE</td>
<td>DATE or TIMESTAMP</td>
</tr>
<tr>
<td>CHAR, NCHAR</td>
<td>CHAR</td>
</tr>
<tr>
<td>FLOAT</td>
<td>NUMERIC (p,s), DECIMAL (p,s), REAL (4 bytes), DOUBLE (8 bytes),</td>
</tr>
<tr>
<td>TIMESTAMP</td>
<td>TIMESTAMP</td>
</tr>
</tbody>
</table>

**Table 5: Greenplum data type**

The following select statement assists in finding the precision and scale used for the NUMBER data types. It assumes statistics are up to date in Oracle:

```
select length(utl_raw.cast_to_number(HIGH_VALUE)) from all_tab_columns where table_name='TABLE_NAME' and column_name='COLUMN_NAME'
```

### Step 3. Identify and assess migration tools and their requirements.

Migration tools considered:
- Informatica
- Migration scripts
- Ora2pg freeware tool
- Web external tables
- Third-party utilities
Step 4. Prepare and deploy Greenplum DCA.

1. Site-specific requirements for DCA deployment.
   a. IP address on network for master, standby server, and network access.
   b. Choose passwords for root and gpadmin, and password for remote console access (iDRAC).
   c. Preferred locale, time zone for DCA servers.
   d. Preferred database character set encoding.

2. Installing Greenplum Data Computing Appliance (DCA).
   a. Appliance: Shipping state from Manufacturing.
      i. Appliance comes with racked, stacked, and cabled.
      ii. Software pre-installed, minimum on-site configuration required.

3. Initial deployment.
   a. Supply power to rack; inspect and validate rack hardware and cabling.
   b. Configure access to network (master and standby only).
   c. Refresh software if needed; run validation check.
   d. Configure locale, timezone, and NTP on all DCA servers.
   e. Run gpinitsystem to initialize database.
   f. Initialize standby database.
   g. Enable Greenplum performance monitor.
   h. Configure and connect EMC as needed.

*The installation time was only 8 hours for each DCA deployed (half and full rack) for a total of 16 hours.*
Step 5. Prepare Greenplum Database Environment in DCA.

1. Create a database in Greenplum DCA. Database properties will be inherited from template1 that was created during the DCA environment setup. Creating a database in Greenplum is as simple as shown in this screen shot:

![Figure 7: Greenplum database creation](image)

2. Preparing schemas.

   a. List all the schemas to be created, and create schemas in Greenplum using the below syntax.

![Figure 8: Greenplum database schemas](image)
**Step 6. Convert DDL from Oracle to Greenplum.**

1. Converting DDL from Oracle to Greenplum involves:
   a. Reading the table properties from the current Oracle environment.
   b. Converting the create table syntax from Oracle to Greenplum.
   c. Converting each data type in table definition from Oracle to Greenplum.
   d. Choosing the right distribution key.

2. Identify the distribution keys. EMC IT chose the distribution keys based on these criteria:
   a. If a table contains a primary key in Oracle, consider it a distribution key in Greenplum.
   b. If a table in Oracle contains no primary key but a unique key exists, then consider a unique key with fewer columns for a distribution key in Greenplum.
   c. Choose Random distribution when the table does not have a candidate column for the distribution key and the table is small.

3. No tool is perfect when it comes to converting the DDL from Oracle to any other MPP database.
   a. EMC IT leveraged a DDL conversion utility, which connects to Oracle and converts table DDLs for one schema, multiple schemas, or a single table.

   The EMC IT DDL conversion utility usage is as follows:

   Usage: `generate_gp_tables.sh -s <schema names separated by comma> -r <Optional:remap schema> -p <Optional:Path>`

   There are three switches in this script.
   -`-s` is mandatory and EMC IT needs to specify the schema name.
   -The EMC IT script can specify multiple schemas separated by comma.
   -The other two switches are optional.

   -`-p` switch is path for generating ddl scripts.
   -If this is not specified, ddl scripts will be generated in the current directory.

   -`-r` is for remap schema option.

   E.g.: `#/apps/oradump2/dump2/gpdump/scripts:oraedw1>/generate_gp_tables.sh -s extract,masterdata -r extract_data,mstr_data`
Directory created.
PL/SQL procedure successfully completed.
PL/SQL procedure successfully completed.
Directory dropped.

#/apps/oradump2/dump2/gpdump/scripts:oraedw1> ls -ltr
  total 496
  -rwxr-x--- 1 oracle dba 9861 Jun 6 01:50 generate_gp_tables.sh
  -rw-r--r-- 1 oracle dba 434555 Jun 6 02:04
cREATE_GPTABLES_EXTRACT_DATA.SQL
  -rw-r--r-- 1 oracle dba 39587 Jun 6 02:04
cREATE_GPTABLES_MSTR_DATA.SQL

The above process illustrates converting DDLs from Oracle to Greenplum for two schemas (Extract and MasterData).

Step 7. Create table structures in Greenplum.

1. Copy converted DDL files to the Greenplum master server script location.
2. Execute the script in Greenplum to create the DDL of all tables in the Greenplum database.

E.g.: psql -d gpcqdb -f create_gptables_EXTRACT_DATA.sql -L create_gptables_EXTRACT_DATA.log -o create_gptables_EXTRACT_DATA.out

Step 8. Create a gold environment as a data migration source.

1. Create a gold environment with migration scripts to perform the clone from the production Oracle Global Data Warehouse. This option will have no impact on the production database during migration of the data. The script:

   a. Shruts down the gold Oracle database and unmounts the Oracle asm diskgroups.
   b. Recreates the clone using EMC Timefinder technology.
   c. Keeps the Production Oracle instance in backup mode.
   d. Activates the created Timefinder clone.
   e. Takes the Oracle production instance out of backup mode.
   f. Mounts asm diskgroups on the gold environment and brings up the Oracle database on the gold environment to use as the Data Migration source.

Step 9. Leverage unloading and loading scripts from Oracle to Greenplum.

The challenge involved in EMC IT’s environment is to be able to migrate data from more than 3,600 tables, with approximately 132,000 columns. Each column of data should be able to be identified uniquely.
1. Unloading involves:
   a. Identifying the schemas and tables to be unloaded.
   b. Establishing the connection to the Oracle database and ASM.
   c. Choosing the type of unload, such as direct or conventional.
   d. Choosing the column separator (E.g.: ;).
   e. Choosing the Record separator (E.g.: \n).
   f. Choosing the Encloser (E.g. “‘).
   g. Choosing the Escape Enclose (E.g. \).

2. Loading to Greenplum involves:
   a. A YAML file for each table. An example YAML file is below. YAML is a human-friendly data serialization standard for all programming languages. In this example, this YAML file had the information about the target Greenplum environment (database name, user, hostname, port, input source information, column identifications in the files, and the delimiters used).

   ```
   VERSION: 1.0.0.1
   DATABASE: gpcqdb
   USER: gpadmin
   HOST: gpserver
   PORT: 5432
   GPLOAD:
   INPUT:
   - SOURCE:
     FILE:
     - /tmp/gp/W_TASK_NOTE_D.dat.
   - COLUMNS:
     - ROW_WID: numeric
     - W_SESSION_NUMBER: numeric
     - TASK_WID: numeric
     - NOTE_ID: numeric
     - NOTE_CRTE_BY_NM: text
     - NOTE_CRTE_BY_RSRC_ID: numeric
     - NOTE_CRTE_DT: timestamp
     - TASK_ID: numeric
     - NOTE_SUBJ: text
     - NOTE_TYPE_WID: numeric
     - NOTE_VISABILITY_FLG_WID: numeric
     - W_LAST_UPDATED_BY: text
     - W_INSERT_DT: timestamp
     - W_UPDATE_DT: timestamp
     - W_ROW_HASH: text
     - W_REPROCESSED_DT: timestamp
   - FORMAT: text
   - DELIMITER: ';'"
b. An input data source; a data file, pipe, or socket.
c. Gpload commands for all the tables involved to load the data from the input source to the Greenplum target database.

Here is an example gpload command:

```
gpload.py -f "/tmp/W_TASK_NOTE_D.ctl" -h gpserver -p 5432 -d gdcqdb -U gpadmin -l "/tmp/W_TASK_NOTE_D.log"
```

**Step 10. Migrate data from Oracle to Greenplum.**

1. Create sockets; unload the data using unload scripts, which unloads the data directly from Oracle ASM in parallel into PIPLE/SOCKET; compress the data; transfer to Greenplum servers; uncompress, and load data to Greenplum using the gpload utility.

Below are example commands to perform the data migration from Oracle to Greenplum using pipe/socket.

1. On Oracle server:

   
   ```
mknod /tmp/TABLE.dat p
   ```

2. On Oracle server:

   ```
gzip < /tmp/TABLE.dat | ssh -q gpadmin@gpmasterhost "gunzip > /tmp/TABLE.dat " &
   ```

3. On Greenplum DCA:

   ```
mknod /tmp/TABLE.dat p
gpload.py -f "/tmp/TABLE.ctl" -h gpmasterhost -p 5432 -d gpdb -U gpadmin -l "/tmp/TABLE.log"
   ```

4. On the Oracle server, start the unload process:

   a. Using this approach, it starts loading data into Greenplum as it unloads data from Oracle without the need for any intermediary storage.

You can also unload the data to an ASCII text file, read the data from this text file, and load into the Greenplum database. This approach requires the intermediary storage to hold the ASCII text files.
After the data has been migrated, the storage space needed on Greenplum is only 50 percent of the storage space needed in Oracle because it eliminates the need for most of the indexes and materialized views.

Here is a high-level illustration of steps 8 – 10:

**Figure 9: Oracle to Greenplum high-level migration steps**
Business POC Benefits

The following are illustrative examples of the benefits of using the Greenplum DCA.

Query Performance

Here are a few examples of queries that were executed in the Oracle environment, and the queries and reports on the Greenplum DCA.

![Figure 10: Oracle versus Greenplum Query 1 example](image)

As the diagram illustrates, a query that took 7 minutes, 7 seconds to execute in the Oracle environment now executes in 28.85 seconds (28855 ms) on more than 99 million rows.
Oracle sample Query 2:

```sql
SQL> set lines 200
13*
SQL> 1
   1 select rg.prmry_flg,
   2       pg.prod_grp_nm, pg.prod_sbg_nm, pg.prod_avail_flg, pg.prod_relbil_cd, pg.srvc_req_flg
   3       count(distinct instnc_id) cnt
   4   from masterdata.W_cq_report_group_REF rg,
   5   masterdata.W_cq_product_group_REF pg,
   6   gs.u_cq_install_history_F cih
   7   where rg.prmry_flg = 'Y'
   8   and prod_relbil_cd = 'B'
   9   and rg.prod_grp_wid = pg.row_wid
  10   and rg.item_num = cih.item_num
  11   group by rg.prmry_flg, pg.prod_grp_nm, pg.prod_sbg_nm, pg.prod_avail_flg, pg.prod_relbil_cd
  12   order by cnt desc
13*
SQL> /
```

```
P  PROD_GRP_NM
-----------------  PROD_SBG_NM
-----------------  P P S  CNT
 Y 3250VL2EU 3250VL2EU  N B N  49

51 rows selected.
Elapsed: 00:02:24.25
```
```sql
SELECT prmry_flg,
       prod_grp_nm,
       prod_sbg_nm,
       prod_avail_flg,
       prod_relbil_cd,
       srvc_req_flg,
       COUNT(DISTINCT instrc_id) cnt
FROM masterdata.W_cq_report_group_REF rg,
     masterdata.W_cq_product_group_REF pg,
     gs.W_cq_install_history_F ch
WHERE rg.prmry_flg = 'Y'
AND prod_relbil_cd = 'E'
AND pg.prod_grp_wid = pg.row_wid
AND rg.item_num = ch.item_num
GROUP BY rg.prmry_flg,
         prod_grp_nm,
         prod_sbg_nm,
         prod_avail_flg,
         prod_relbil_cd,
         srvc_req_flg
ORDER BY cnt DESC
;
```

<table>
<thead>
<tr>
<th>prmry_flg</th>
<th>prod_grp_nm</th>
<th>prod_sbg_nm</th>
<th>prod_avail_flg</th>
<th>prod_relbil_cd</th>
<th>srvc_req_flg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>DS24M2</td>
<td>DS24M2</td>
<td>N</td>
<td>E</td>
<td>N</td>
</tr>
<tr>
<td>Y</td>
<td>DS300B</td>
<td>DS300B</td>
<td>N</td>
<td>E</td>
<td>N</td>
</tr>
<tr>
<td>Y</td>
<td>DS220B</td>
<td>DS220B</td>
<td>N</td>
<td>E</td>
<td>N</td>
</tr>
<tr>
<td>Y</td>
<td>DS16M2</td>
<td>DS16M2</td>
<td>N</td>
<td>E</td>
<td>N</td>
</tr>
<tr>
<td>Y</td>
<td>NEX-5010</td>
<td>NEX-5010</td>
<td>N</td>
<td>E</td>
<td>N</td>
</tr>
<tr>
<td>Y</td>
<td>MP3000B</td>
<td>MP3000B</td>
<td>N</td>
<td>E</td>
<td>N</td>
</tr>
<tr>
<td>Y</td>
<td>NEX-5020</td>
<td>NEX-5020</td>
<td>N</td>
<td>E</td>
<td>N</td>
</tr>
<tr>
<td>Y</td>
<td>32SOVL2E</td>
<td>32SOVL2E</td>
<td>N</td>
<td>E</td>
<td>N</td>
</tr>
<tr>
<td>Y</td>
<td>ESS832E</td>
<td>ESS832E</td>
<td>N</td>
<td>E</td>
<td>N</td>
</tr>
<tr>
<td>Y</td>
<td>32SOVL2EU</td>
<td>32SOVL2EU</td>
<td>N</td>
<td>E</td>
<td>N</td>
</tr>
</tbody>
</table>

{51 rows}

Time: 13460.735 ms
```
Figure 12: Oracle versus Greenplum Query 1 and Query 2 example

The query graph illustrates that Query 1 was run in 427 seconds in Oracle and in 28 seconds on Greenplum DCA. Query 2 on Oracle executed in 144 seconds and 13 seconds on Greenplum.

New Analytics Abilities

Prior to the introduction of the Greenplum DCA, all analysis was done via batch reports. In the POC, the performance of the Greenplum DCA gave Customer Quality the ability to use a tool set to create a dashboard and execute on-the-fly analytics that was not available in the Oracle environment.

With the Greenplum DCA in place, Customer Quality can now process large data sets in record time.

Here are examples that illustrate the ability to view Big Data in a dashboard and do analytics on-the-fly via dashboard tools such as SAS or other third-party BI tools.
Greenplum now enables Customer Quality to provide management with a dashboard environment depicting many views of a product analysis, and, most importantly, to proactively address an issue before it is seen by the customer.

Figure 13: Big Data analysis dashboard
Customer Quality can now quickly analyze across a broad range of data to enable proactive remediation of quality concerns and ensure continued customer satisfaction.

The time saved using Greenplum has allowed enhanced predictive analytics to be performed, which will provide even more value internally and for EMC’s customers.
As the diagram illustrates, Greenplum gives accelerated insight into:

- Predictive models
- Data mining
- Dashboards
- Dynamic query reports

**EMC IT Global Data Warehouse to Greenplum POC: Lessons Learned**

Here are highlights of the lessons learned about Oracle from this migration.

<table>
<thead>
<tr>
<th>Challenge</th>
<th>Resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greenplum expertise</td>
<td>Greenplum reduces the complexity of Big Data analytics, enabling a simple retraining of the current EMC IT Oracle Data Warehouse team.</td>
</tr>
</tbody>
</table>
| Data type conversions    | There is no perfect tool for converting the data types from Oracle to any MPP database. The majority of the numeric values in Oracle were defined as NUMBER data type without precision and scale. To better manage storage space, we experimented with a few conversion options. Our testing showed the following as the best option:  
  1. Converted all the Oracle NUMBER columns to |
Greenplum NUMERIC data type.  
   a. NUMERIC data type consumes a minimum of 7 bytes in Greenplum and can expand if the data is larger. This has been a very good resolution so far. We had to remediate the load failures.  
   b. There is still a chance to save some space where there are columns that do not need 7 bytes of storage space. Greenplum native INT data types are more storage-efficient and perform better. This requires expensive 100 percent stats collection on Oracle or a lot of expertise on thousands of columns of space requirement.

<table>
<thead>
<tr>
<th>Data distribution</th>
<th>The distribution policy determines how to divide the rows of a table across the Greenplum segments.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1. Used hash distribution and a primary key as the distribution key where a primary key exists for a table in Oracle.</td>
</tr>
<tr>
<td></td>
<td>2. Used hash distribution and unique index as a distribution key where a primary key does not exist and a unique index exists.</td>
</tr>
<tr>
<td></td>
<td>3. Used random distribution for tables with no primary key and no unique index.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Staging space for unload files (for 10TB database)</th>
<th>For accelerated data movement, we leveraged a mounted NAS device between Oracle and Greenplum.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rather than using the compression algorithm during unload and the un-compression algorithm for load, we used PIPE to unload data with no compression and loaded to Greenplum via PIPE.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>OBIEE integration with Greenplum</th>
<th>EMC IT used the pgsqloDBC driver to successfully connect OBIEE with the Greenplum Database; it generates the SQL we wanted and expected.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>We also leveraged UDFs (user defined functions) in Greenplum to replicate some Oracle built-in functions.</td>
</tr>
</tbody>
</table>

| Generating a DDL from Oracle to Greenplum | There is no perfect tool out there. The big challenge is generating the DDL with the right distribution policy, right distribution key, and correct data type. Working with the experts at Greenplum, we leveraged a set of methodologies and scripts to convert the DDL from Oracle to Greenplum. |

Table 6: Lessons learned in EMC IT’s migration from Global Data Warehouse to Greenplum
Conclusion
This white paper illustrated a synergistic model for deploying the Greenplum DCA with the existing EMC IT Global Data Warehouse (BI Grid infrastructure), showing the high-level steps to migrate, business challenges the EMC’s Customer Quality business unit experienced in the migration from the Global Data Warehouse to Greenplum, and the improvements seen in Big Data and analytics processing with the Greenplum DCA for EMC IT.

Benefits
– Greenplum DCA reduces the need for labor-intensive offline analysis with dramatic results:
  • Customer Quality has seen a dramatic reduction in processing time.
  • Query examples that show a dramatic reduction in query execution time:
    • Query 1 example from 427 seconds executing in the Oracle environment down to 28.85 seconds for 99 million rows with the DCA.
    • Query 2 example from 144 seconds executing in the Oracle environment down to 13 seconds with the DCA.
  • Data load example from 6 days on the legacy platform to 29 minutes on the Greenplum DCA.
  • Introduction of third-party BI tools to accelerate visualization of Big Data and analytics.

– EMC IT created a platform to develop and deploy a Big Data and analytics appliance, and to use as a future platform for developing Business Intelligence as a Service (BlaaS).
References
The links below will direct you to detailed information on the following subjects:

EMC IT’s Global Data Warehouse

EMC IT’s Migration to the Open, Expandable Oracle BI Grid - Applied Technology

Greenplum Database and Data Computing Appliance

www.emc.com/bigdata

Greenplum DCA Data Sheet

Greenplum DCA Data Sheet

EMC Greenplum Data Computing Appliance architecture

EMC Greenplum Data Computing Appliance architecture

SAS interoperability with EMC Greenplum database

SAS interoperability with EMC Greenplum database

Acknowledgments
The author thanks EMC IT’s Global Data Warehouse, Greenplum, and the Customer Quality teams for assistance in creating this white paper.