70 TB of Disk Savings in 4 Days: EMC IT’s Informatica Data Subset Proof of Concept

Applied Technology

Abstract

This white paper illustrates the ability to reduce the data growth challenge seen with EMC’s Oracle Applications CRM implementation via a Proof of Concept (POC) sponsored by the EMC IT organization. This POC will demonstrate the rapid reduction of data for EMC’s Oracle E-Business Suite footprint for its environments via Informatica’s Data Subset solution.

April 2011
# Table of Contents

**Executive summary** ........................................................................................................... 4  
**Introduction** ....................................................................................................................... 4  
  Audience .............................................................................................................................. 4  
**Data storage needs** ........................................................................................................... 4  
**Technology overview** ....................................................................................................... 7  
  Informatica Data Subset architecture .................................................................................. 7  
  Two subsetting methods ....................................................................................................... 8  
  Informatica Data Subset Workbench .................................................................................... 10  
  Enterprise Data Manager ..................................................................................................... 11  
  EMC IT’s subset deployment infrastructure ......................................................................... 13  
  EMC storage components ................................................................................................. 15  
**EMC IT’s Proof of Concept (POC)** .................................................................................. 16  
  POC daily actions ............................................................................................................... 16  
    Day 0 .............................................................................................................................. 17  
    Highlights from the POC ............................................................................................... 19  
    Day 1 .............................................................................................................................. 19  
    Day 2 .............................................................................................................................. 19  
    Day 3 .............................................................................................................................. 20  
    Day 4 .............................................................................................................................. 20  
**Conclusion** ..................................................................................................................... 21  
  Acknowledgments ............................................................................................................. 22
Executive summary

Exponential “Data Growth” with enterprise applications such as Oracle E-Business Suite brings challenges for IT organizations to manage the physical (storage), operational processes (replication/archive), and financial (people/technology) costs of their data explosion.

An enterprise’s data footprint can be defined as the total storage needed by the organization to fulfill its business needs for the lifecycle of its Oracle Applications implementation infrastructures, in areas such as:

- Development
- Production
- Test
- Business continuity/disaster recovery

Introduction

The purpose of this paper is to illustrate a method and toolset to reduce the data growth challenges above via a Proof of Concept (POC) sponsored by the EMC IT organization. This POC will demonstrate the rapid reduction of data for EMC’s Oracle E-Business Suite footprint for development and test environments via Informatica’s Data Subset solution.

Audience

This white paper is focused on the CIO, system architect, Oracle architect, storage architect, and supporting staff, focusing on Oracle Applications DBAs, server administrators, and network administrators.

Data storage needs

EMC, like many large enterprises, has deployed enterprise-scale implementations of Oracle’s ERP and CRM solutions to enable its business in Manufacturing, Finance, Quoting, Customer Service, Professional Services, Sales, and Marketing.

Two enterprise-scale mission-critical systems support EMC's core revenue-generating functions ($15 billion in revenue for 2008):

- An ERP solution, supporting 20,000 employees with 2,000 concurrent users
- A CRM solution, supporting 36,000 named users worldwide with 3,500 concurrent users. This implementation is one of the top five Oracle Applications transactional systems in the world depending upon the modules that are used.

With time, EMC’s ERP production database has grown to 2 TB in size, and the CRM production database has grown about 7 TB. There are currently 15 instances of ERP and 19 instances of CRM databases serving to ensure efficient global application delivery and support.

These instances are supporting the following EMC business environments:

- Daily business (transactional)
- Reporting
- Downtime solution
- Test
- Development
- Projects
- Training
The total storage accounts for more than 320 TB, and it costs several million dollars of infrastructure to support these environments, people, and processes for EMC to maintain its competitive edge.

Figure 1 is an example of the data storage and growth needs that multiply via instance needs in the lifecycle of an Oracle Applications implementation.

Figure 1. The Oracle Applications lifecycle instance challenge - data growth

We can use three dimensions to visualize data growth. The “Doeswijk Data Model” in Figure 2 shows that growth.

The first dimension is primary production data growth, which applications staff try to estimate as best they can. The second is growth in copies or replicas, which very few people outside of the storage team have as a concern. The third dimension is retention data, which is point-in-time (static) data that should be archived (as needed for company or legal compliance). Total data size is the product of these three dimensions.

Primary data requires copies for many purposes like backup, development test, data mining, data warehouse processes (Extract/Translate/Load (ETL)), and data distribution. At some time, the majority of primary data ages but needs to be retained in an archive where backup is no longer required as long as there is at least one copy for recovery purposes.
If you visualize data in these terms, it will assist in understanding why enterprises are always running out of storage. Application users plan only for the production phase of their data and have no idea about the other environments needed after production go-live and about the number of copies needed to protect, analyze, and share their data. A change in any of these dimensions has a multiplying effect on the total data volume.

Data resides in storage; therefore, this model can illustrate storage capacity as a cube that contains data volume cubes. The dimensions of a storage capacity cube have a relation to the dimensions of the data volume cube. In some cases, these dimensions might be tiers of storage, which relate to corresponding data dimensions.
The EMC IT organization took on the challenge of a Proof of Concept to evaluate Informatica’s Data Subset solution. This solution has two components: an Enterprise Data Manager that has accelerators (Metadata) that speed the data objects needed in a subset solution, and the Workbench, which controls the subset process.

Objectives of EMC IT’s POC were the following:

- Use of EMC’s CRM “11.5.10” production instance
- Use the following five Oracle Applications modules to subset:
  - Installed Base Module (CSI)
  - Configurator Module (CZ)
  - Customer Service Module (CS)
  - Java Technology Foundation (JTF)
  - Custom (EMCCS)
- For performance and usability, understand and review the Informatica solution to have “ease of use” and have the ability for a performant tool when creating reduced database environments.
- Reduce the footprint of the test/development databases yet maintain referential integrity and usability.

**Technology overview**

The following were the technology components used in this POC:

**Informatica Data Subset architecture**

Any organization that uses an enterprise application maintains full size copies of the production database to meet development and testing requirements. This multiplicity of application environments presents an operational challenge (storage, process, and staff). The storage required to accommodate these numerous copies often stretches IT infrastructure to its limits. Intensifying this problem is the fact that numerous copies are unnecessarily full size as development and testing teams may only require relatively small samples of data to meet their objectives.

Informatica Data Subset enables enterprises to shrink non-production application databases using a highly scalable engine that removes data unnecessary for your team’s development and testing activities. It filters transaction data (which is usually the majority of all data) based on functional rules you define. Subsetted instances maintain data integrity and are available for immediate use.

Informatica Data Subset is designed for use on non-production instances. To maximize the reduction of data, Data Subset removes all transactions that do meet your criteria, regardless of their status (open, closed, and so on).

Informatica Data Subset enables organizations to eliminate this problem by creating smaller, targeted databases for project teams. This subsetted database maintains all application integrity while taking up only a fraction of the space.
Two subsetting methods
Informatica’s Data Subset solution offers two methods for creating smaller database copies: Select and Remove, and Select and Insert.

Select and Remove
This method enables policy administrators to identify data to keep, and Informatica Data Subset then deletes or truncates extraneous data. This method also typically requires the DBA to perform a reorganization of the affected tablespaces to fully reduce the footprint size of the new smaller database.

Figure 4 illustrates the Select and Remove method via time slice and functional slice.

![Diagram of Select and Remove Subset Method]

**Time Slice**
Production Database

**Database Subset**
Most Recent 6 months

**Functional Slice**
Production Database

**Database Subset**
Data from only 1 Business Unit

Figure 4. Select and Remove subset method
Select and Insert
This method creates an empty database template, and Informatica Data Subset then populates the database with only data specified in the retention policy. Both methods ensure that the resulting database copy maintains application integrity and requires less space than a full size clone.

Figure 5 illustrates the Select and Insert method via time slice and functional slice.

**Time Slice**
Production Database

Production Database

Copy only data to keep

Empty Database

Most Recent 6 months of Data

Clone

Copies of Subset for Dev, Test, Training, etc.

**Functional Slice**

Production Database

Copy only data to keep

Empty Database

Data Belonging to Only 1 Business Unit

Clone

Copies of Subset for Dev, Test, Training, etc.

**Figure 5. Select and Insert subset method**

How does it work?
Informatica engineers mine Oracle E-Business Suite, SAP, PeopleSoft Enterprise, and Siebel CRM enterprise applications to capture data structures, entity definitions, and business rules.

Informatica builds comprehensive accelerators, which power the Informatica engine to remove unnecessary transactional data with speed and integrity. The subset engine identifies transactions based on the policy definition and enables administrators to review the effect of the subsetting policy before actually removing the data. It is possible to use the Informatica Subset tool to extend the accelerators or to build support for custom modules.
Informatica Data Subset Workbench

The Workbench is where the action of subsetting is defined and executed. This software component gives the subsetting team members an easy-to-use interface that is the “control center” for the subsetting activities for the POC.

Figure 6 is a screenshot of the Workbench tool.

Figure 6. Informatica Subset Workbench
Enterprise Data Manager
The Enterprise Data Manager (EDM) is where the action of creating the “driving class” for the module to subset is defined. This software component gives the subsetting team member an easy-to-use module accelerator (metadata) for the relationship of business, module, and database objects.

Out-of-the-box accelerators for metadata
The Informatica Data Subset solution provides centralized management of all data growth policies from an integrated platform. Informatica Data Subset comes pre-populated with accelerators for major business applications such as Oracle E-Business Suite, SAP, PeopleSoft Enterprise, and Siebel CRM and provides the ability to extend Informatica’s functionality to custom and third-party applications. The solution includes a GUI-based extensibility component that enables administrators to examine the pre-packaged accelerators to satisfy any questions about definitions or logic. As business requirements dictate, administrators can modify the accelerators to accommodate any customizations or extensions to the business applications.

Figure 7 is a screenshot of the EDM tool.

![Figure 7. Informatica Enterprise Data Manager](image-url)
Figure 8 shows a driving class (Module Accounts Payable (AP)) that can be created with the Informatica Subset toolset.

**Figure 8. Driving class example**

Informatica Subset provides powerful filtering capabilities and intelligent built-in relational rules that administrators can define based on driving classes. A driving class includes a primary driving table (that is, a diagram above AP Invoices) within an enterprise application module and all the related tables to maintain referential integrity. Informatica Subset automatically analyzes the targeted data based on the defined rule and estimates how much data will remain in the database.
EMC IT’s subset deployment infrastructure

The following was the EMC POC 11.5.10 CRM deployment infrastructure. As Figure 9 shows, a two-server deployment contains the following:

- The Subset Database (SUBSET) resides on the server “crmtest1”. (This is a copy of production.)
- The Informatica Subset Workbench software (AMHOME) is installed on the server “crmtest2”.
- The Informatica Workbench/UI component resides on the server “crmtest3”.
- Informatica’s Enterprise Data Manager (EDM) resides on EMC IT’s POC subsetting personnel laptop.

![Figure 9. EMC’s POC subset deployment](image-url)
Figure 10 illustrates the components and processes for the selected subsetting method for the POC Select and Remove method.

Figure 10. The POC's subset components and process

The Select and Remove method was chosen for the POC for the following reasons:

- The size of the EMC CRM production instance was 7 TB.
- The decision was to not patch the POC infrastructure.
- It had the quickest/least amount of impact to the EMC POC subsetting team.
- It was the easiest method to promote to deployment.
EMC storage components

The following were the storage components used in the POC:

**EMC Symmetrix DMX™** - The Symmetrix DMX-4 system was used in the DMX series and extends EMC’s leadership in the high-end enterprise and storage market. The DMX-4 delivers immediate support for the latest generation of disk drive technologies, Flash drives for superior performance, 4 Gb/s Fibre Channel for high performance, and SATA II for high capacity.

The DMX-4 is based on Enginuity™ 5773, which provides investment protection that delivers performance gains along with information-centric security advancements via integration with RSA enVision®. With the DMX-4 and Enginuity 5773 all replication and security activities are easy to manage with the Symmetrix Management Console (SMC).

**EMC TimeFinder®** — TimeFinder allows users to nondisruptively create and manage point-in-time copies of data (local replication). This allows operational processes, such as backup, reporting, and application testing, to be performed independently of the source application to maximize service levels, without impacting performance or availability.

TimeFinder/Clone was used in this use case. It creates highly functional, high-performance, pointer-based, full-volume copies of Symmetrix DMX volumes that can be used as point-in-time copies for data warehouse refreshes, backups, online restores, and even volume migrations.

**EMC PowerPath®** — PowerPath works with the storage system to intelligently manage I/O paths, and supports multiple paths to a logical device. In this solution PowerPath manages four I/O paths and provides:

- Automatic failover in the event of a hardware failure. PowerPath automatically detects path failure and redirects I/O to another path.
- Dynamic multipath load balancing. PowerPath distributes I/O requests to a logical device across all available paths, thus improving I/O performance and reducing management time and downtime by eliminating the need to configure paths statically across logical devices.
EMC IT’s Proof of Concept (POC)

The EMC POC consists of three important components:

- **POC team** – EMC IT’s Oracle Applications DBA team, EMC IT’s Oracle Applications Development team (Modules), Subject Matter Experts (SMEs), and Informatica SMEs
- **Process and Subset Process** – In EMC IT’s case it was determined that with a large Oracle Applications instance size of 7 TB that the Select and Remove method would be used to subset.
- **Technology** - Informatica’s Subset Solution (EDM and Workbench UI)

**POC daily actions**

The following were the activities done each day for the POC candidate modules:

**Step 1.** Review EDM module accelerators – Use the EDM to define what data to use in the subset process.

**Step 2.** Confirm with the development/business team what the retention parameters are for each module. This defines the business retention rules (time/function) that create the driving tables for the subsetting step.

**Step 3.** Use Informatica’s Data Subset Workbench to start the Select and Remove Process. This includes three phrases: Pre-processing, data movement, post-processing.

- **Pre-processing Stage:** The Selected Entity or Module name should be referred here as a reference below (that is, Configurator or CZ)
  - Store index information on each subsetted table
  - Pre-rowcount of all rows of tables
  - Create a table stage (that is, if there are 40 tables in the Subset model then create 40 tables in the staging area. These 40 tables represent the number of metadata-driven tables affected by the Subset of the Configurator (CZ) module.).
  - Drop indexes on the source
  - Disable constraints
  - Disable triggers

- **Data Movement Stage:**
  - Move the data to stage. This is done in levels – starting with the top-level driving class table and then subsequent tables at the next level (Review the driving class example in Figure 8).
  - Truncate source table data
  - Move data back to the source from stage
  - Re-create indexes
  - Re-create PKeys
  - Post-rowcount of all rows of tables.

- **Post-Processing Stage:**
  - Re-enable triggers
  - Re-enable FK constraints
  - Drop staging tables in the Staging area

**Step 4.** Commence the subset via Informatica’s Workbench UI.

**Step 5.** Review Informatica’s Workbench (UI) subsetting results.
Day 0
Prior to the start of the POC, the Informatica Team conducted a Data Discovery session. This session used Informatica’s Data Growth Assessment Discovery questionnaire and the Data Growth Analyzer (DGA) tool set. A script executed on a copy of production and was then imported into the DGA spreadsheet.

The strength of this session/toolset was the following:

- Visualization of the data growth problem, thus allowing for establishment of data growth patterns in the modules
- Simulate the impact of the toolset, Informatica Data Subset, in reducing the data growth problem
- Illustrate the return on investment (ROI) if you decide to Subset/Archive, therefore a Go/No Go for a POC

Current state
Figure 11 displays EMC’s current 11.5.10 CRM production size state and potential future growth base on the Current State collected data for the DGA.
Figure 12 illustrates the current data growth in EMC’s CRM production instance and the top modules to review as candidates for subsetting.

### Production datafile (GB)

<table>
<thead>
<tr>
<th>Data</th>
<th>Free</th>
</tr>
</thead>
<tbody>
<tr>
<td>4,203</td>
<td>2,533</td>
</tr>
</tbody>
</table>

### Data by schema (GB)

<table>
<thead>
<tr>
<th>Schema</th>
<th>Size (GB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CZ</td>
<td>469.7</td>
</tr>
<tr>
<td>CSI</td>
<td>364.3</td>
</tr>
<tr>
<td>OTHER</td>
<td>362.3</td>
</tr>
<tr>
<td>CS</td>
<td>337.0</td>
</tr>
<tr>
<td>JTF</td>
<td>313.1</td>
</tr>
<tr>
<td>ON</td>
<td>294.6</td>
</tr>
<tr>
<td>PA</td>
<td>268.7</td>
</tr>
<tr>
<td>OBT_APPS_INTF</td>
<td>243.6</td>
</tr>
<tr>
<td>OSM</td>
<td>239.2</td>
</tr>
<tr>
<td>OKC</td>
<td>210.0</td>
</tr>
<tr>
<td>EMCOMP</td>
<td>163.0</td>
</tr>
<tr>
<td>ASO</td>
<td>151.0</td>
</tr>
<tr>
<td>AR</td>
<td>146.5</td>
</tr>
<tr>
<td>OKS</td>
<td>111.6</td>
</tr>
<tr>
<td>EMCCSI</td>
<td>115.7</td>
</tr>
<tr>
<td>QP</td>
<td>78.5</td>
</tr>
<tr>
<td>APPS</td>
<td>74.9</td>
</tr>
<tr>
<td>INV</td>
<td>71.2</td>
</tr>
<tr>
<td>H&amp;C</td>
<td>65.4</td>
</tr>
<tr>
<td>EMCOS</td>
<td>64.3</td>
</tr>
<tr>
<td>OTHER</td>
<td>58.3</td>
</tr>
</tbody>
</table>

Figure 12. EMC’s IT top subsetting candidate modules

Figure 13 illustrates the impact of subsetting (reduced size) to the Current State size of the POC Oracle Applications CRM production instance.

### Subsetting reduces non-production datafile size (GB)

- Clone of production datafile
- Subset
- Subset & truncate
- Subset & archive
- Subset, archive & truncate

Figure 13. Impact of subsetting (reduced size)
Figure 14 shows the DGA ROI for subsetting.

Figure 14. Savings of subsetting (reduced cost)

**Highlights from the POC**

The following highlights the dramatic reduction of data activity of each day of the Proof of Concept for EMC IT.

**Day 1**

This initial day was used to discuss the two methods to create a subset. The EMC POC team chose the Select and Remove method. The reason this method was selected was that EMC has such a large production instance of over 7 TB and the patch levels made this method easiest to do in the POC and to deploy post-POC.

The POC team started the action of the POC using both the EDM (metadata tool) and Workbench to start the first module to subset, which as seen in the DGA session was Configurator. The team developed the driving class from the metadata (CZ accelerator)/EDM and then the started the Select and Remove process of the CZ module via Informatica Workbench. This process took some time because of the size of the module (over 400 GB). (Please review the three process steps.) The EMC development team then created the retention parameters/“driving class” for the customer service module (CZ).

The following was accomplished on Day 1:

- Subset Architecture Presentation / Discussion – Select/Insert or Select/Remove
- Navigating Configurator (CZ) Accelerator/Metadata via EDM
- Develop retention parameters for CZ via EDM/EMC Development Team
- Started the CZ Subset process via Workbench

**Day 2**

Discussing the Customer Service (CSI) module was on the Day 2 agenda. The same POC team process took place:
• The team reviewed the Accelerator/Metadata needed for the CSI module via EDM.
• The POC development team created the retention parameters/“driving class” to use in the subsetting process (Select/Remove).
• The last action was to kick off the subsetting process via Workbench.

In summary, the following was accomplished:

• Reviewed a successful CZ (Configurator) Subset from Day 1.
• Started the Customer Service (CS) Subset
  ▪ Created a new driving class for CS via EDM
  ▪ Retention parameters were developed by POC development team members for CS
  ▪ Executed the CS Subset via Workbench
• Started the CSI (Installed Base) Subset
  ▪ Created a new driving class for CSI via EDM
  ▪ Retention parameters were developed by POC development team members for CSI
  ▪ Executed the CSI (Installed Base) Subset via Workbench

Day 3
A custom module (EMCCCS) schema was subsetted on Day 3. The same POC team process took place:

• Created the metadata needed for a custom module (EMCCCS) schema via EDM
• The POC development team created the retention parameters/“driving classes” to use in the subsetting process (Select/Remove)
• The last action was to kick off the subsetting process via Workbench

In summary, the following was accomplished on Day 3:

• CS was completed – CS_INCIDENTS_AUDIT_B table
• Created a custom subset against the EMCCS schema
• Began the JTF Subset as the next schema to be subsetted

Day 4
The JTF schema subsetting was completed on Day 4. The same POC team process took place:

• Created the metadata needed for the JTF schema via EDM
• The POC development team created the retention parameters/“driving classes” to use in the subsetting process (Select/Remove).
• The last action was to kick off the subsetting process via Workbench.

In summary, the following was accomplished on Day 4:

• Continued the JTF Subset
Conclusion
The following was the impact of the EMC IT Informatica Data Subset POC:

- The toolset was easy to deploy and use.
  - Toolset setup was easy.
  - EDM allows designers to understand and utilize existing supplied accelerators.
  - Ability to configure multiple application instances from one Informatica Data Subset infrastructure
- It was proven that the following EMC candidate modules could be subsetted rapidly (in a four-day POC):
  - (CZ, GL, CS, CSI, and JTF)
  - Created a custom subset, EMCCS
  - All work with the toolset done by EMC IT’s POC team member.
- There was substantial data savings from the POC.

The following shows the cumulative data savings both in total and via the subsetted modules.

PRE-SUBSET: the largest modules:
CZ (469.7 GB), CSI (364.3 GB) and CS (337 GB)

PRE-SUBSET:
Data = 4.203 TB
Freespace = 2.533 TB

POST-SUBSET:
Data = 2.989 TB
Freespace = 3.492 TB

That is 1.21 TB of space reduction!
In conclusion, when EMC IT used the savings shown above of 1.21 TB and multiplied this savings by the dimensions of data growth, production, replica, and archive in EMC’s Oracle Applications environments, then a total number of approximately 70 TB in savings was achieved in the four-day POC.

Additionally, the subsetting POC improves the efficiency of development, testing, and deployment via less people and process time. Because subsetted databases are referentially intact, they can be used for testing of patches, customizations, implementations of new features, and more. By reducing the size of the data set, users realize significant performance gains in backup and restore cycles via EMC’s TimeFinder technology, and see increased efficiency in EMC’s IT ongoing support of the Oracle Applications Lifecycle management (People/Process/Technology) savings.

**Acknowledgments**

The author would like to thank EMC IT’s POC team and IT development team, and Informatica’s POC team for assisting in the creation of this white paper.