EMC²
where information lives®

White Paper

60 TB of Savings in 4 Days: EMC IT's Informatica Data Archive 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 production environment via Informatica's Data Archive solution.

April 2011

Part Number h6539.1

Table of Contents

Executive summary **4**

Introduction **4**

Audience 4

Data storage needs **4**

Technology overview **7**

Informatica Data Archive Workbench User Interface 9

Enterprise Data Manager 10

EMC IT's archive deployment infrastructure 12

EMC storage components 13

EMC IT's Proof of Concept **13**

POC daily actions 13

Day 0 14

Highlights from the POC 16

Day 1 16

Day 2 16

Day 3 16

Day 4 16

Conclusion **17**

Acknowledgments 17

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 production environments via Informatica's Data Archive 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 growth since EMC IT's CRM production go-live to present to projected growth via Informatica's Data Growth Analyzer (DGA) toolset. This chart illustrates the need to reduce the size of EMC IT's CRM production and to archive (relocate transactional data) out of the production instance in the lifecycle of an Oracle Applications implementation.

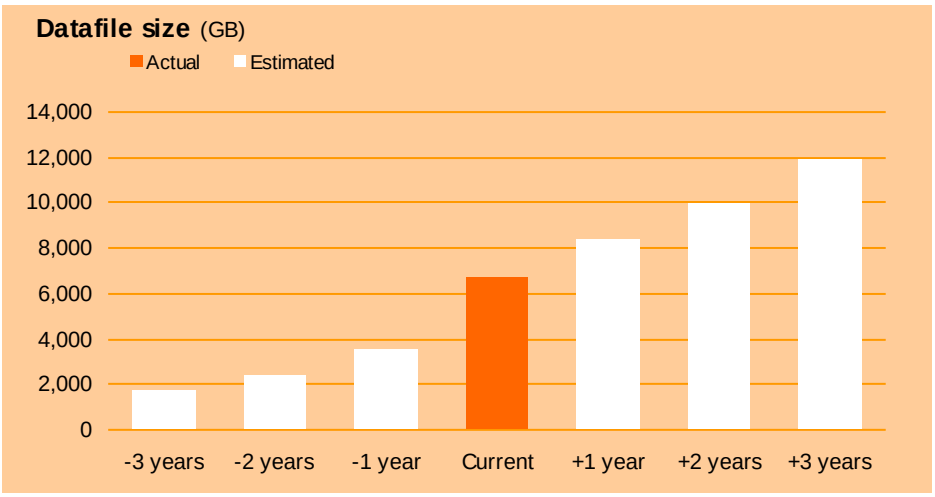

Figure 1. EMC IT's CRM Oracle Applications production instance data growth challenge

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 the applications team (infrastructure (storage/server) team, applications database administrators, and applications development team) tries to estimate as best they can. Growth in copies or replicas is the second, 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.

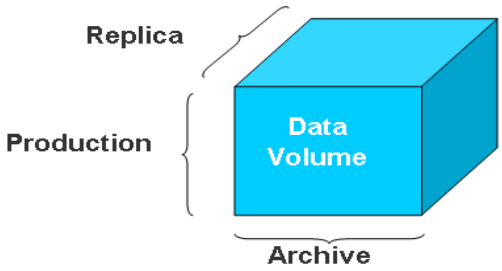

Figure 2. The Doeswijk Data Model shows the three dimensons of data growth

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.

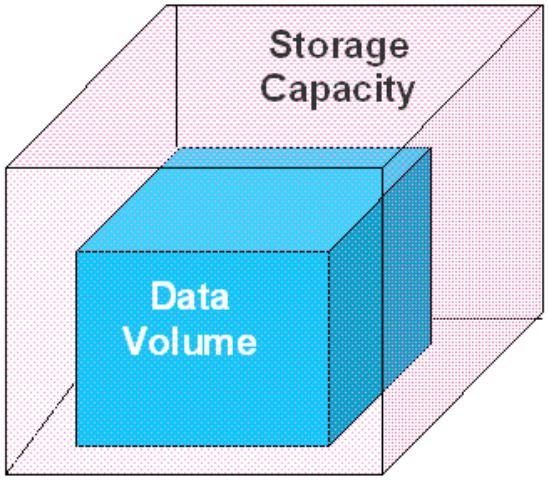

Figure 3. Data resides in the storage capacity cube

The EMC IT organization took on the challenge of a Proof of Concept to evaluate Informatica's Data Archive solution. This solution has two components: an Enterprise Data Manager that has accelerators (Metadata) that speed the data objects needed in an archive solution, and the Workbench, which controls the archive process.

Objectives of EMC IT's POC were the following:

- Use a copy of EMC's CRM "11.5.10" production instance
- Archive the following new modules:
  - Service Contracts (OKS)
  - Contracts Core (OKC)
- Use Informatica's Data Archive solution to archive the existing archived module Quoting (ASO/CZ) that now uses a third-party archive software solution. The goal was to demonstrate seamless access and the ability for an end user to access the archived (relocated) data as if it was still in the production instance.

- For performance and usability, understand and review the Informatica solution to have "ease of use" and have the ability for a performant tool to archive (relocate) inactive transactional data to nearline storage yet still maintain seamless accessibility

Technology overview

The following sections detail the technology components used in this POC.

Informatica Data Archive architecture

Figure 4 illustrates the components in the Informatica Data Archive architecture.

The production instance contains a staging area for the archive process (AM_STAGE). The Informatica tool repository schema (AM_HOME) and the online active archive (AM_HISTORY) relocated transactional data removed from the production instance.

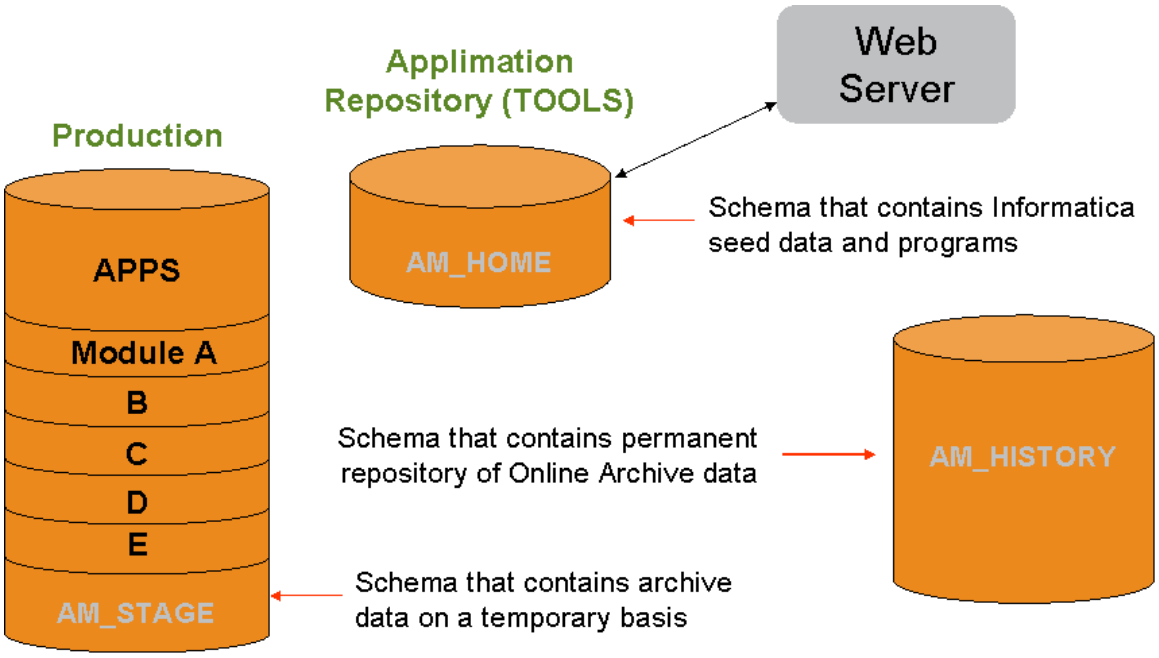

Figure 4. Informatica's Data Archive architecture

Archive method

There is a seven-step process used in the Informatica archive process:

1. Candidate Generation via the Informatica EDM tool
   - Uses the "top" or "driving" transaction table
   - Determines what is purgeable and what is not purgeable
   - Identifies what business rules a transaction failed
   - Used by archive/purge
2. Build Staging via the Informatica Archive Workbench tool
   - Creates tables in the staging schema, which temporarily houses archive data
   - Tables are exact copies of source tables
   - Dynamically built using the data dictionary
3. Archive – Copy to Stage via the Informatica Archive Workbench tool
   - Copies data from the source table to the staging table
   - The SQL statement used is INSERT INTO STAGE SELECT * FROM SOURCE
   - Middle tier is used for tables with special datatypes like LONG
4. Purge – Delete from Source via the Informatica Archive Workbench tool
   - Uses ROWID and PK to delete from the source table
   - If no PK, then the delete "metadata" is used
   - Parallel delete options
   - Oracle parallel delete
   - Applimation parallel delete
5. Xpress Detect - Validate the destination via the Informatica Archive Workbench tool
6. Xpress Merge - Copy to the destination via the Informatica Archive Workbench tool
7. Purge Staging via the Informatica Archive Workbench tool

Figure 5 illustrates these steps:

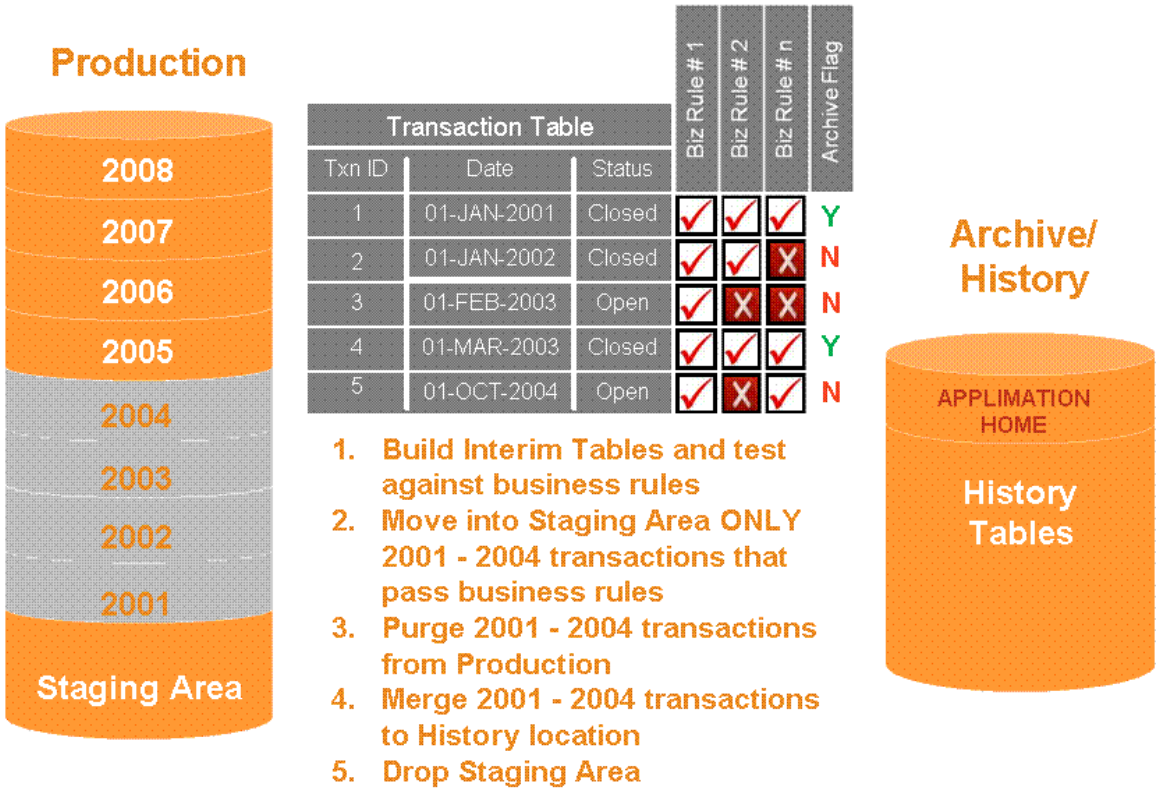

Figure 5. Archive process detailed method

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 archive engine identifies transactions based on the policy definition and enables administrators to review the effect of the archive policy before actually removing the data. It is possible to use the Informatica Archive tool to extend the accelerators or to build support for custom modules.

Informatica Data Archive Workbench User Interface

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

Figure 6 is a screenshot of the Workbench tool.

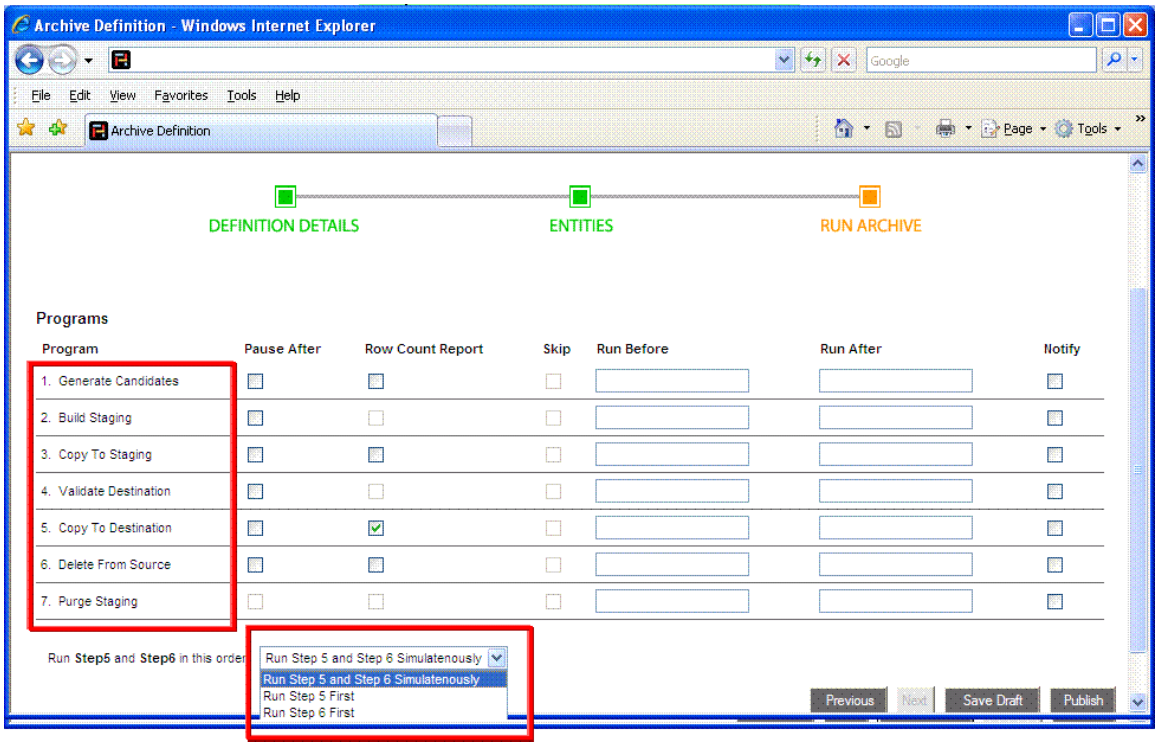

Figure 6. Informatica Data Archive Workbench

Enterprise Data Manager

Enterprise Data Manager (EDM) is where the action of creating the "driving class" for the module to archive is defined. This software component gives the archiving 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 Archive solution provides centralized management of all data growth policies from an integrated platform. Informatica Data Archive 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.

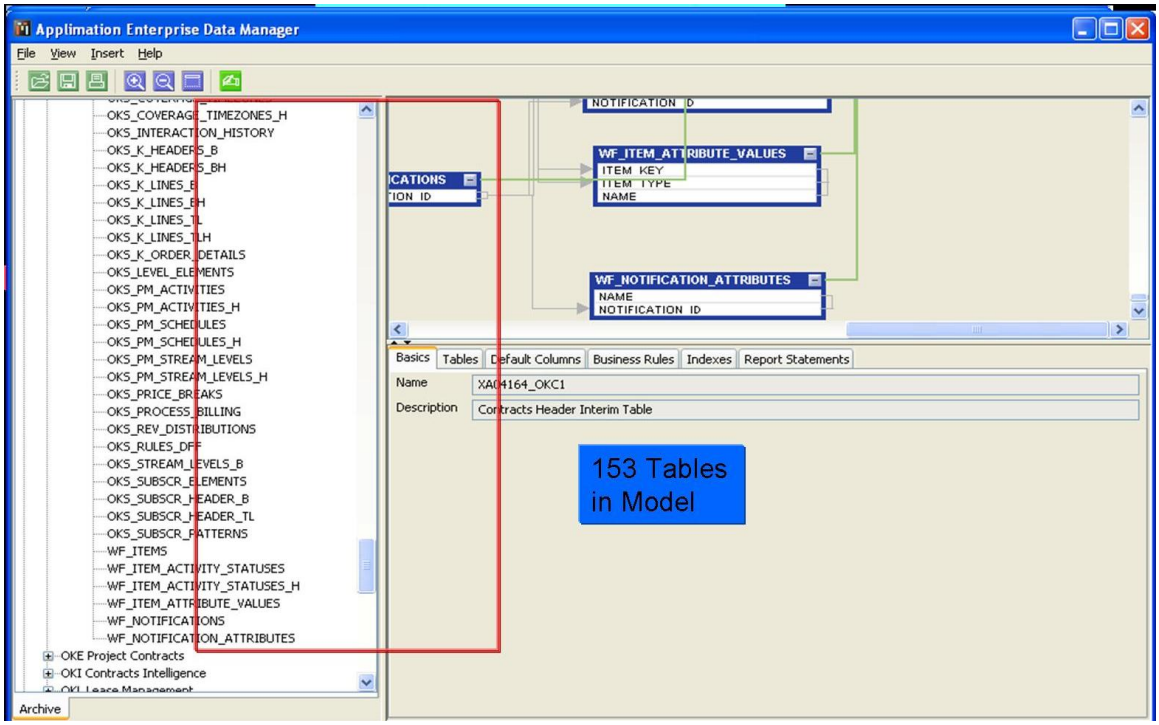

Figure 7. Informatica Enterprise Data Manager

Figure 8 shows a driving table (Header) that can be created with the Informatica Archive toolset.

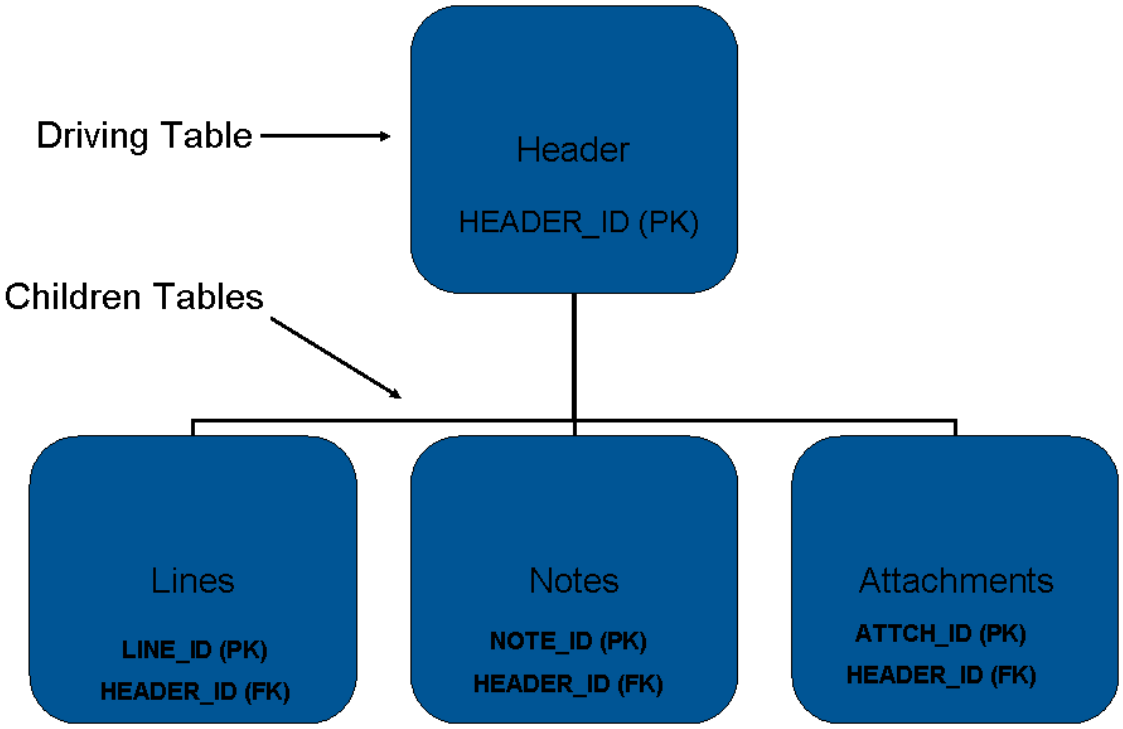

Figure 8. Driving class example

Informatica Archive 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, shown in this diagram as the Header table (Driving Table)) to Children Tables within an enterprise application module and all the related tables to maintain referential integrity. Informatica Archive automatically analyzes the targeted data based on the defined rule and estimates how much data will remain in the database.

EMC IT's archive deployment infrastructure

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

- A copy of EMC IT's CRM production database resides on "PROD".
- The Active Archive Database (HISTORY DB) resides on the server "HIST".
- The Informatica Archive Workbench software (AMHOME) is installed on the server "HIST".
- The Informatica Workbench/UI component resides on the server "CRMTST".
- Informatica's Enterprise Data Manager resides on EMC IT's POC archiving personnel laptop.

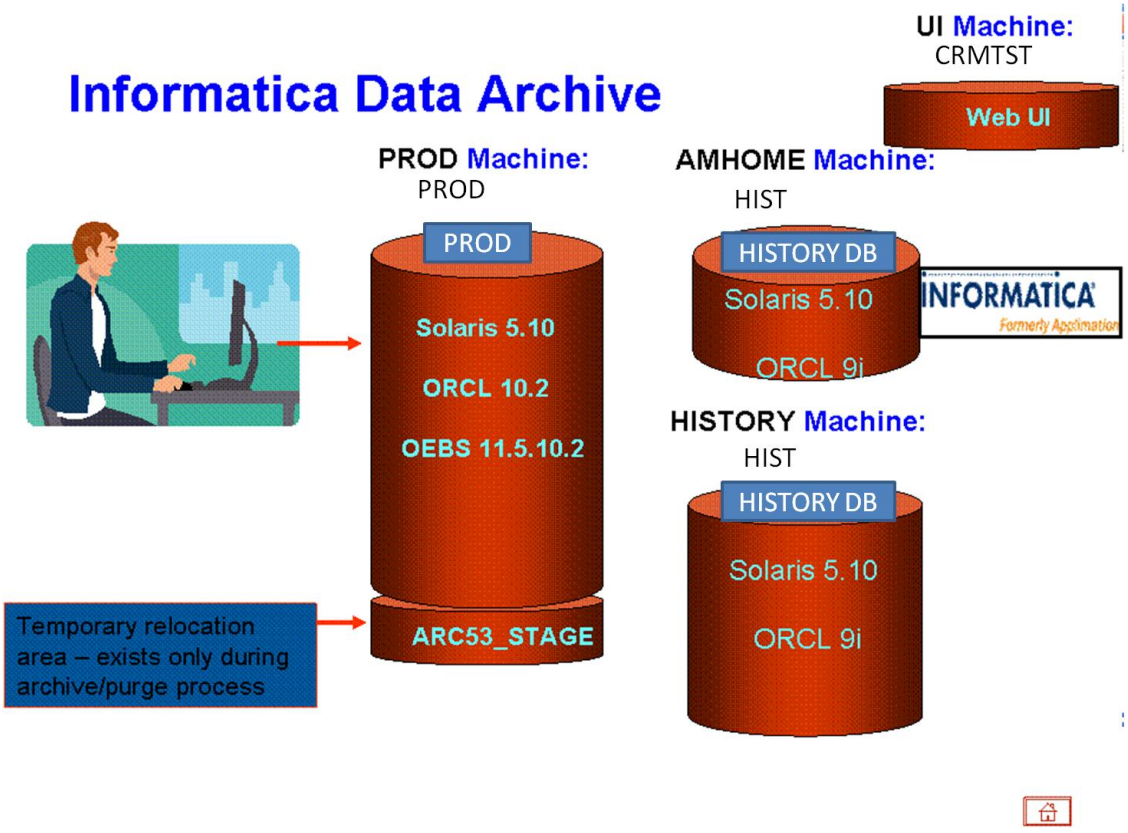

Figure 9. EMC's POC Archive deployment

.

EMC storage components

The following were the storage components used in the POC:

EMC Symmetrix DMX™ - The Symmetrix DMX-4 system was used from 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

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
- Archive process – In EMC IT's case it was determined — with a large Oracle Applications instance size of 7 TB and an incumbent third-party archive solution in place — to archive new modules with Informatica's Data Archive solution. Additionally, an already incumbent archived module (incumbent archiving solution) would be used with the Informatica Archive toolset to understand the process. Finally, the archive process would be tested both in the new Informatica Archive and in the incumbent third-party archive.
- Technology - Informatica's Data Archive 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 archive 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 creates the driving tables for the archiving step.

Step 3. Use Informatica's Data Archive Workbench to start the archive process. This includes three phrases: Pre-processing, data movement, post-processing.

Step 4. Commence the archive via Informatica's Workbench UI.

Step 5. Review Informatica's Workbench UI archiving 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) toolset. 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
- Simulated the impact of the toolset, Informatica Data Archive, in reducing the data growth problem
- Illustrated the return on investment (ROI) if you decide to archive, therefore making it a Go/No Go for a POC

Current state

Figure 10 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.

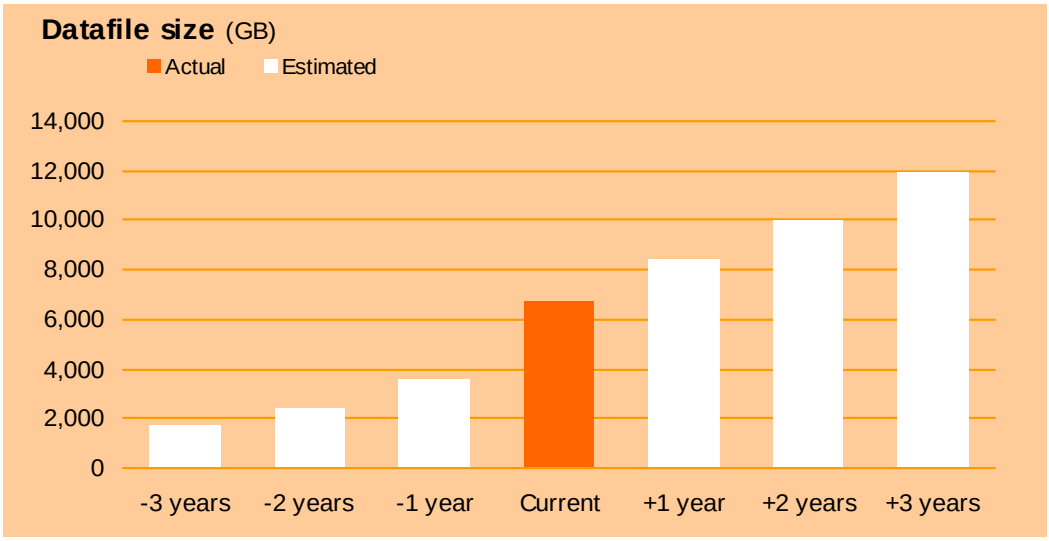

Figure 10. EMC's IT current and future datafile sizes

Figure 11 illustrates the impact of archiving (reduced size) to the Current State size of the POC Oracle Applications CRM production instance.

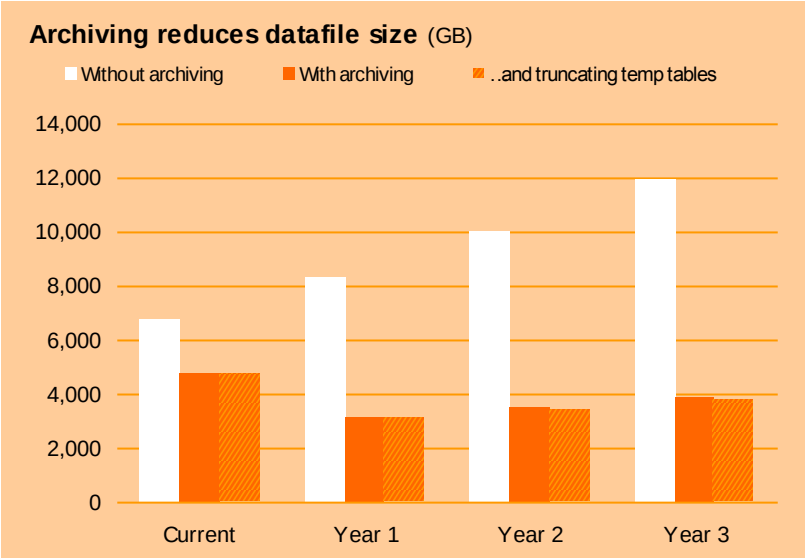

Figure 11. Impact of archiving (reduced size)

Figure 12 illustrates the DGA ROI for archiving.

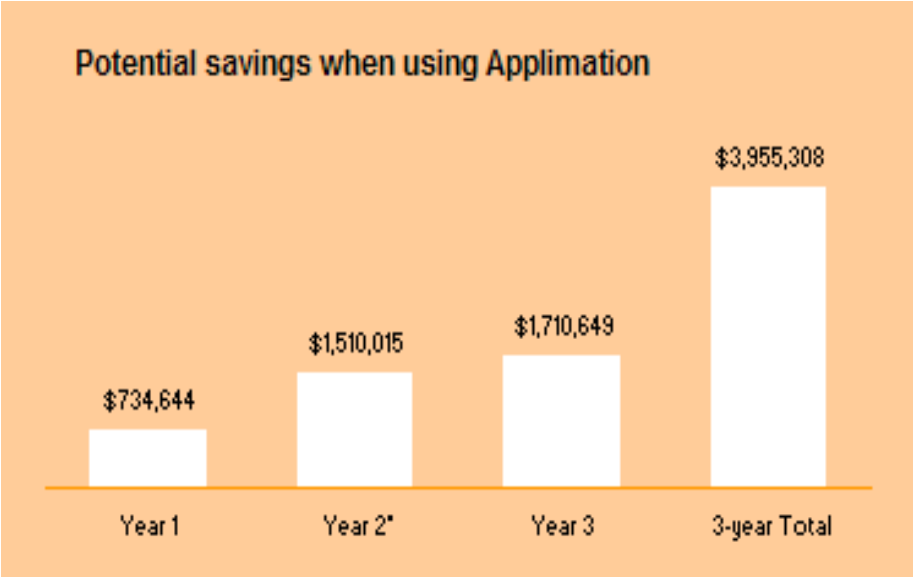

Figure 12. Savings of archiving (reduced cost)

Highlights from the POC

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

Day 1

The POC team started the action of the POC using both EDM (metadata tool) and Archive Workbench to start the first modules to archive, which were Service Contracts (OKS) and Contracts Core (OKC). Please review the archive process steps on page 8. The EMC development team then created the retention parameters/"driving class" for the OKS/OKC archive process.

The following was accomplished on Day 1:

- Navigated Service Contracts and Contracts Core (OKC/OKS) accelerators in the Enterprise Data Manager toolset to define the "driving class" for archiving parameters for each module
- Set up configuration with the Workbench UI
- Worked with the EMC POC development team on Business Rules for the OKC/OKS archive
- OKC/OKS archive started with Archive Workbench UI

Day 2

The POC tested the "Seamless Access" layer to Service Contracts and Contracts Core. In addition the POC team reviewed the Quoting Module, which was an incumbent archived module using another vendor's archive toolset.

The following was accomplished on Day 2:

- Tested the Seamless Access layer to OKC/OKS. EMC IT's POC development team tested the "Archive Only" view.
- Started analyzing the existing archived module "Quoting" (CZ/ASO)
  - Built a new "Custom" entity with EDM to match the current archived module
  - Archived exact tables now with the Informatica/Applimation Archive toolset
- Discussed the next options, which would be one of the following:
  - Informatica archive store into the incumbent third-party vendor's archive store
  - Incumbent third-party vendor's archive store into Informatica's archive store

Day 3

The POC team completed the test of the existing archived module, Quoting, archived with Informatica's Data Archive solution. The POC team reviewed where to store the archive, in either the incumbent third-party archive vendor's Archive History or Informatica's Archive History.

The following was accomplished on Day 3:

- Large Quoting module archive completed successfully
- (Optional next move) – Which repository to use?
  - Selected the incumbent third-party archive into the Informatica/Applimation archive
  - Some table differences were encountered between each vendor's archive in columns
- Tested "Duplication" functionality against Informatica's History

Day 4

The POC tested the incumbent third-party archive vendor's History with Quoting because the previous three days the team had success moving the archived data into Informatica's Archive History.

The following was accomplished on Day 4:

- Switched the direction testing of the History archive
  - Informatica's archived Quoting data went into the incumbent third-party archive vendor's History. The Informatica POC team moved data from 81 potential tables (not all populated) and it worked successfully.
- Discussed pros and cons of each methodology (Informatica Archive and incumbent third-party archive)

Conclusion

The following was the impact of the EMC IT Informatica Data Archive 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 Archive infrastructure.
- It was proven that the following EMC candidate modules could be archived rapidly (in a four-day POC):
  - Service Contracts (OKS)
  - Contracts Core (OKC)
  - Quoting (ASO/CZ) — an existing archived module via an incumbent archive software/module
- There was substantial data/cost savings from the POC. As Figure 10 illustrates, EMC can instantly improve performance and reduce EMC's IT burden by archiving transaction data. Informatica's Data Archive solution can immediately reduce datafile size by as much as *30%*, and slow its growth rate by *100%*. Manually truncating temporary tables can reduce datafile size by an additional *1%***.**

In conclusion, when EMC IT used the savings shown above 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 60 TB in savings was achieved in the four-day POC.

Additionally, the archiving POC improves the efficiency of an "active archive" on tiered storage. This allows end user the ability to "seamlessly" access the relocated data but at a much more lower cost of storage via less people and process time. By reducing the size of the production instance, users realize significant performance gains in the production instance and in backup and restore operational cycles (smaller production instance size) via EMC's TimeFinder technology, and see increased efficiency in EMC IT's 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.