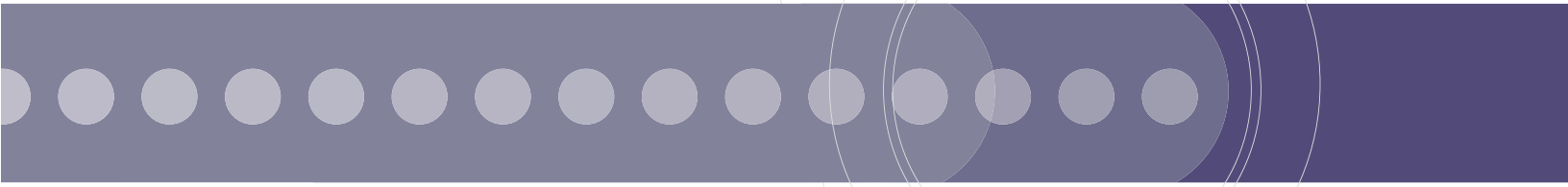
INFORMATICA®

CASE STUDY

Achieving High Performance with Informatica



This edition published February 2005

An In-Depth Study of Waste Management

Data volumes are getting larger. Batch windows are shrinking or becoming extinct with the advent of real- or near real-time computing. And performance is a critical factor that must be considered by companies requiring data integration activities.

Performance Improvements

Compared to the earlier data integration attempts using custom JDBC code, PowerCenter's and PowerExchange's native access and data compression capabilities enabled Waste Management to:

- Reduce load times from 90 hours to under four hours.
- Realize a 3X improvement in throughput of data volume as well as network volume.
- Reduce the sheer volume by 1.3 billion rows of data processed each week by utilizing patented change data capture technology provided by PowerExchange.

The general term "performance" encompasses the concepts of throughput and scalability, both of which are critical to the success of an enterprise data integration platform. Throughput measures the rate at which rows or bytes of data can be processed, and it describes how much computing power is required to process a given data integration scenario. Scalability is a measure of predictability and is used to describe how readily the computing environment can be augmented to handle larger and more complex data integration scenarios that are beyond the capacity of a minimal configuration.

According to Forrester Research, big data is a big problem that requires scalability, and complex data processing demands scalability, too.[1] In most large organizations, the focus is on managing all aspects of integration, from merging data—often from hundreds of disparate data sources—to cleansing and transforming the data, to providing users with a variety of ways to view the data. There are a variety of ways of accomplishing this, such as multiple, traditional warehousing implementations including operational data stores (ODS) and data mart creations, as well as migration, instance consolidation, and single view efforts. The bottom line is that, for many organizations, big data really is a big problem that must be handled efficiently and on a unified a platform of products that can most efficiently manage the performance levels required for big data.

This case study will take a detailed look at Waste Management Inc.'s use of Informatica products to achieve high levels of performance across all of their integration needs.

Waste Management, Inc. (WMI) is a leader in the waste collection, disposal, recycling, and waste-to-energy industry with over 52 percent market share, $12.2 billion dollars in revenue as of September 30, 2004, and $1.02 billion dollars in free cashflow.[2] Headquartered in Houston, Texas, the company serves more than 20 million residential and two million commercial customers in the U.S. and Canada. Its 52,000 employees, 26,000 collection vehicles, 293 landfills, and 16 waste-to-energy plants collect roughly 80 million tons of waste a year.

The problem that Waste Management faced was the lack of a central data repository for intelligently managing and serving key customers as well as powering business optimization efforts, such as more cost-effective truck routing. Their biggest challenge was integrating hundreds of data sources to present a holistic view of customer data. To do this would take time, effort, resources, and a strict IT discipline—but with a big pay off.

Today, Waste Management uses a centralized integration competency center (ICC) model to control data integration services using Informatica® products including PowerCenter®, PowerExchange™, PowerCenter Connect for Peoplesoft, and PowerAnalyzer™. The company processes more than 1 billion rows of data in daily batch and near real-time processing. There is a weekly process that loads another 470 million rows of data, for a total of approximately 6-7 billion rows of data processed each week. They maintain 35 downstream applications. This would not have been possible in part without the high performance data integration products provided by Informatica.

[1] How to Evaluate Enterprise ETL, by Philip Russom of Forrester Research, December 17, 2004.
[2] Source: Yahoo! Financials, February 2005

This paper will outline details of Waste Management's implementation and performance, and the challenges they faced while integrating multiple terabytes of data spanning hundreds sources into an enterprise solution that drives the business. It will highlight one of the most critical pieces of the architecture, the ODS, which facilitates several downstream applications that are used to solve key business problems and drive revenue for Waste Management. It will also spotlight a near real-time process to provide additional insight into the usage of Informatica products that spans across traditional warehousing and into support of operational and transactional processing.

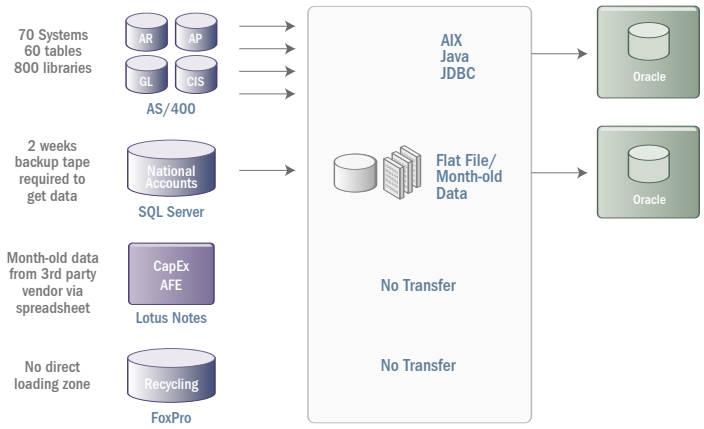


Figure 1
Initial picture of Waste Management's processing environment.[3]

The Early Days of Data Integration

The initial focus for Waste Management was to create a single view of customer. While this is a common project for many companies today, the challenge of consolidating customer data was daunting due to the numerous acquisitions of more than a thousand companies in the late 1990's. As a result, their data was located in more than 800 separate libraries spread across multiple IBM AS/400 servers—almost 70 of them. Waste Management needed to transform itself into a single, efficient company, and it had to start with the integration of data.

Dispersed billing information prevented Waste Management from obtaining the comprehensive view of its customer base required for optimum cash flow, targeted sales and marketing, and efficient business operations. But this wasn't just a business problem. The legacy accounts receivable application, Mid-America Systems (MAS), was implemented and in use across the company, but it required excessive personnel to maintain. A Java coding methodology was used to extract data from MAS, produce a flat file, and load to an Oracle table. The issue with that process was that it could oftentimes take five days to load one table of approximately 20 million rows—only one library at a time could be extracted, and timeouts were common. In addition, the lack of auditing and logging prevented Waste Management from performing backward checks to ensure data validity, thereby providing no confidence in the integrity of data. The space consumption for a full extract was in the hundreds of gigabytes. And there was no ability to incorporate data stored in Lotus Notes and FoxPro applications with that from the AS/400. Meanwhile, requirements for the data mounted for projects that were in the works to address solutions for routing and logistics, credit and collections, and sales and marketing. This demand furthered the need to implement a solution for extracting once and publishing as many times as necessary.

Then Informatica Entered the Picture

Once Waste Management decided on an automated solution, they evaluated and purchased Informatica PowerCenter to join together data from all of the disparate sources and load it into an ODS called "level0." The guiding principle of the ODS was to extract once and publish many times. The operating system on which PowerCenter was installed was IBM AIX. Oracle was used both for the repository database and the ODS. One of the immediate benefits realized was that PowerCenter allowed them to finally extract the Lotus Notes and FoxPro data via an easy-to-use ODBC interface. Because of the user-friendly design tools, the development time decreased when compared to the hand-coded Java method, and maintenance was much easier due to the reusability and object inheritance provided by PowerCenter—it was

"We knew our problems were solvable, but we also knew they could be potentially fatal if not addressed immediately. So, we started taking action. We tackled the information technology problems that were crippling the company."

Maury Myers,
Chairman and CEO of Waste Management, Inc.[4]

[3 & 4] Source: Presentation by Stephanie Gaines, Manager of Enterprise System Development, Waste Management

possible to make a change once and have it automatically propagate to anywhere that object was used.

With the addition of PowerExchange, Waste Management was able to treat the AS/400 files as a relational database against which SQL queries could be performed. The result was that extractions were significantly faster and easier to construct than before. Other benefits included elimination of timeouts, improved logging, and more readily available performance statistics. With respect to data integrity, any issue with inaccurate data could be addressed from a single place—the PowerCenter repository—allowing for quicker problem solving.

In terms of performance, significant time savings were realized compared to the previous method with the ability to run the extract, transform, and load in a single, end-to-end process combined with a compression ratio measured at 3:1 for data extractions from the AS/400. And not only was the timeline much shorter than the previous hand-coded method, but the fact that it was a single process provided fewer failure points with which they might have to contend. The initial test runs from 2001 on a 2-CPU IBM AIX server showed significant throughput improvements. A subset from prototype data is shown below in Figure 2. With these numbers, they expected to go from 5 days to 7-8 hours for a full nightly run on production-sized servers.

One reason for the throughput improvement was that Waste Management experienced better efficiencies with PowerCenter than with the Java code. The code was designed to go through each library on each network address to produce a flat file per network address. These files then had to be processed and loaded into Oracle tables. Java was causing too much of a slowdown with the number of processes that were required to do the job, and a tremendous amount of file input/output (I/O) resulting from the need to stage so many flat files. In addition, there were issues with the ability to perform delta processing due to the lack of date fields on each file. Using PowerCenter, many of the separate processes that were required by Java were eliminated, as was the file I/O. PowerCenter was able to perform everything in a single, end-to-end process without staging, and the delta processing could be easily addressed through in-memory lookups.

Next Phase

There were still some challenges with the initial implementation partly due to the fact that the AS/400 libraries were dynamic, often changing on a daily basis, and resulted in an inability to predict the extracts on a given day. Another problem was that WMI had reached the maximum CPU and memory limits that were available at the time. And, finally, because of the initial successes the company had, management began asking for more data in a quicker timeframe. Service level agreements had to be stepped up, despite the fact they were still dealing with billions of records—big data was still a big problem.

Iteration 1

Waste Management devised a load-balancing methodology by which they would perform extracts from the AS/400 based on time zone–workflows for eastern, central, mountain, and pacific were created to be run serially with each workflow consisting of several sessions running in parallel. While this approach provided some relief, the company was still hitting time constraints due to hardware limitations and the same data volumes as before, with the only difference being the organization of processing.

Iteration 2

In an attempt to reduce the number of rows being extracted on a nightly basis, and also for the purpose of providing more up-to-date information to the call centers, a change-only data capture methodology was developed using traditional AS/400 programming. However, Waste Management quickly realized the limitations of that solution which included additional hardware stresses on the AS/400 side, whereas before it had only been the AIX servers, as well as primary key corruption as a result of the hand-coded initiative.

Table	Row Count	Original Time	PowerCenter/ PowerExchange Time	% Improvement
1	6,728,325	23 hours	46 minutes	2,874%
2	8,423,399	6 hours	58 minutes	533%
3	13,189,731	7 hours	91 minutes	360%
4	18,964,746	8.4 hours	130 minutes	281%

Figure 2
Signs of performance improvement with PowerCenter/PowerExchange.[5]

[5] Source: POC results data from 2001, excerpts from internal Waste Management document

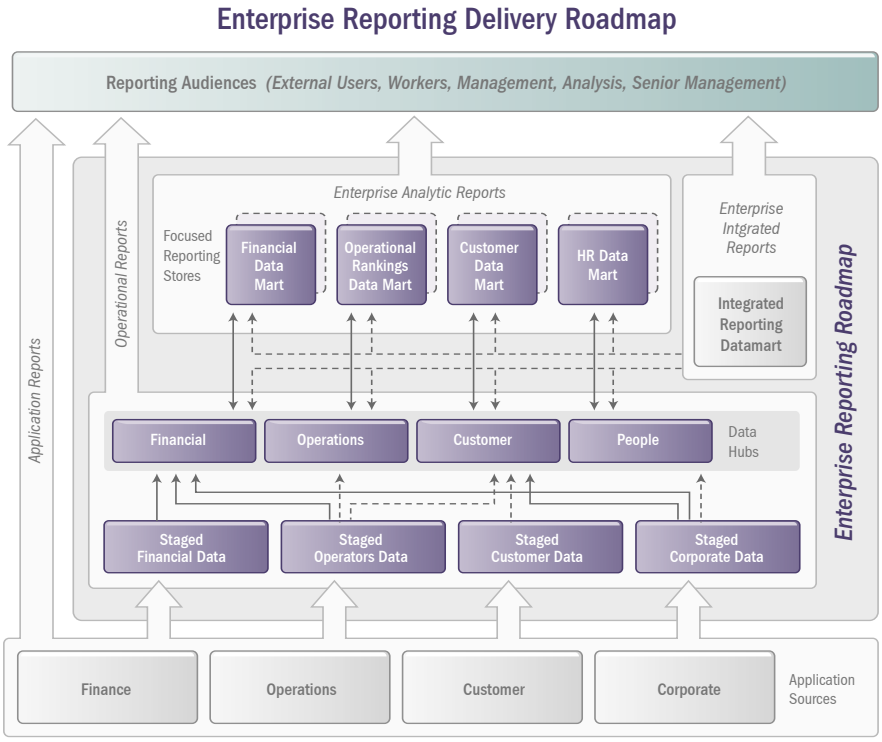


Figure 3
Today's environment at Waste Management
Note: Informatica PowerCenter and PowerExchange products are used for all data integration and movement throughout the Waste Management environment.

Current MAS Processing

Thanks to a critical hardware upgrade, Waste Management now has an IBM S85 running IBM AIX 5.1 with 24-CPU 650 Mhz, 64GB RAM, and 4TB disk space. The AIX server hosts PowerCenter 7.1.1, PowerCenter Connect for Peoplesoft 7.1, PowerExchange 5.1, and Oracle 9.2.0.3 that contains the repository database, data warehouse, and ODS and currently stores approximately 1.5TB of data.

Nightly throughput for the MAS batch process is calculated at approximately 6.9 MB/sec, with the weekly batch process running at 9.7 MB/sec. This is due to the built-in partitioning capabilities of PowerCenter that allow for streamlined, multi-threaded extract, transformation, and load processing. Additional partitioning capabilities automatically allow for proper data routing that ensures the integrity of data without the need for manual manipulations to bring results back together. Now Waste Management is assured not only of the quality and consistency of their data—but also of the unlimited throughput capabilities provided by PowerCenter.

Spotlight on Consolidated Customer View

The Consolidated Customer View, or CCV, is an application that is utilized by customer service representatives to quickly and efficiently access customer data. It consists of a browser-based user interface that provides access to customer data in MAS that is created by providing a near real-time data store from all customer and billing information. This is facilitated through the use of Informatica PowerExchange change data capture (CDC) technology that allows rapid access to customer information across all AS/400 libraries and logical partitions (LPARs) to improve the effectiveness of the customer service representative.

Approximately 500,000 rows of data are processed each day from 4am - 7pm in near real-time using the CDC capabilities provided by PowerExchange. This technology allows Waste Management to maintain business service level agreements of two minutes such that all data is updated and report-ready within two minutes of being posted to the transaction system on the AS400. This is done via a started task that captures journal entries of changes that have been logged since the last time of execution.

Results of Today

As mentioned earlier, Waste Management utilizes a central services ICC model where standards and process definitions, technology decisions, and all development work for integration initiatives are controlled by a centralized team. This team is responsible for maintaining the physical environment, operational data store, warehouse, and all 35 downstream applications that are run in batch nightly, weekly, and in real-time processes. Each application has its own service level agreements that must be met in order for critical business decisions to be made using the most accurate data in the timeliest manner.

As noted in Figure 3, several reporting audiences are serviced with four basic types of report requirements:

- Enterprise Integrated Reports: these are inter-day reports generated from the integrated reporting data mart using confirmed dimensions and standardized facts. These reports are delivered using Informatica PowerAnalyzer customer dashboards with indicators and alerts as well as summary reports linked by standard analysis paths.
- Enterprise Analytic Reports: these are inter-day reports that are also generated through analytic-oriented data cube technology.

- Operational Reports: these are inter-day reports that are generated from a combination of application databases or shared data hubs. The reports are provided in standardized format with aggregations for combined subject areas and data sources.
- Application Reports: these are intra-day reports that are generated from a single application database. These reports provide functional transaction validation and summarization.

PowerExchange CDC technology has replaced hand coding previously used for capturing changed-only data for consuming applications, including CCV as well as other MAS near real-time systems, which required a total of approximately three million rows of data on a daily basis. To support this, there are two servers dedicated for the near real-time processing—one has 12-CPU 600MHz with 16GB RAM and the other has 8-CPU 1.45GHz also with 16GB RAM. Both servers run IBM AIX 5.1 and the latest versions of PowerCenter and PowerExchange. These servers are utilized between 4am and 7pm daily to handle all near real-time data integration and movement needs.

MAS batch processing is still run using the time zone workflow methodology. It accounts for approximately 350 million rows of data loaded each night (Sunday - Friday) during a four and a half hour time period, with three and a half hours for the extract and an additional hour to rebuild indexes on the level0 ODS for approximate throughput of 28,000 rows/sec. The overall MAS batch process includes 718 total sessions with 42 sessions running in parallel at any given point during the batch window, and maximum CPU usage peaks at approximately 55 percent, indicating that there is much more capacity available in the existing hardware during the batch window period.

The IBM S85 server used for MAS processing is also used for all other batch processing taking place in the Waste Management environment. The rest of the data is loaded in additional batch processes that run via a combination of 73 workflows designed to load approximately 750 million rows each day sourcing a variety of additional data sources, including Peoplesoft application data. Data volumes are expected to grow by approximately 2TB over the next year as Waste Management incorporates more data from additional Peoplesoft applications, MAS, and various Web applications.

Conclusion

Waste Management has been very successful moving from a highly fragmented business that was encumbered with complexity and IT costs to a streamlined enterprise. One of the key enabling factors for this success was the introduction of a central integration competency center that utilizes Informatica products.

With the capabilities provided by PowerCenter and PowerExchange, Waste Management has realized a drastic improvement in the performance of nightly and weekly batch processing, and is able to provide data as it is needed to the business with the new ability to perform near real-time processing. They have been able to branch out from warehousing into other integration initiatives using the same technology and infrastructure, providing support for 35 production applications with still more capacity left for growth. The company has also been able to work smarter by rationalizing the architecture to support a single, end-to-end process rather than enduring the cost and overhead associated with flat file staging. And they have saved money, time, and effort and improved their overall return on investment by utilizing Informatica products that are built for performance.

INFORMATICA®

Worldwide Headquarters, 100 Cardinal Way, Redwood City, California 94063, USA
phone: 650.385.5000 fax: 650.385.5500 toll-free in the US: 1.800.653.3871 www.informatica.com

Informatica Offices Around The Globe: Australia • Belgium • Canada • China • France • Germany • Japan • Korea • the Netherlands • Singapore • Switzerland • United Kingdom • USA

J50061 6580 (03/21/05)