An Oracle White Paper
January 2010

Real-Time Data Integration for Data Warehousing and Operational Business Intelligence

ORACLE

Executive Overview

Data warehouses are used for more than just strategic reporting, analytics, and forecasting. Today, companies are investing significant resources to integrate valuable information contained in their data warehouse into their day-to-day operations. Incorporating business intelligence into operational decision making enables these organizations to optimize business performance throughout the day. However, to achieve these efficiencies, data must be provided in real time.

Why is real-time data so crucial? To support operational users and influence what should happen next, the enterprise data warehouse needs to know what is happening right now. There are many data integration technologies that serve the data acquisition needs of a data warehouse, and the demand for low-latency data is causing IT organizations to evaluate a range of approaches: intraday batch extract, transform, and load (ETL) processes; minibatches; enterprise application integration (EAI); extract, load, and transform (ELT) technologies; as well as real-time change data capture (CDC) techniques. The challenge is to determine what solution or combination of solutions will meet the need for current data, which will propel the move to operational data warehousing.

This white paper addresses the business reasons to move to real-time data warehousing and describes some of the common data integration approaches, with an emphasis on using real-time CDC capabilities.

The Data Warehouse and Real Time

Business time is increasingly moving toward real time. As organizations look to grow their competitive advantage, they are trying to uncover opportunities to capture and respond to business events faster and more rigorously than ever. Today, the majority of competitive advantage comes from the effective use of IT. Therefore, from that standpoint, the key to achieving faster business intelligence (BI) is a robust enterprise data warehouse combined with an enterprise analytics framework.

Across the enterprise, each facet of the business gathers data through an assortment of activities, and many organizations now deliver this data to a central data warehouse—where the data is captured, aggregated, analyzed, and leveraged to improve decision making. The quality of these decisions depends not only on the sophistication level of the analytics applications that run on the data warehouse, but also on the underlying data. Data has to be complete, accurate, and trusted. For that reason, it has to be timely: timely data ensures better-informed decisions.

The lifecycle of a data record through enterprise analytics starts with the capture of a business event in a data repository such as a database. Data acquisition technologies deliver the event record to the data warehouse. Analytical processing helps turn the data into information, and a business decision leads to a corresponding action.

To approach real time, the duration between the event and its consequent action needs to be minimized. As outlined in Figure 1, the initial data acquisition and delivery to the warehouse introduces the majority of the latency.

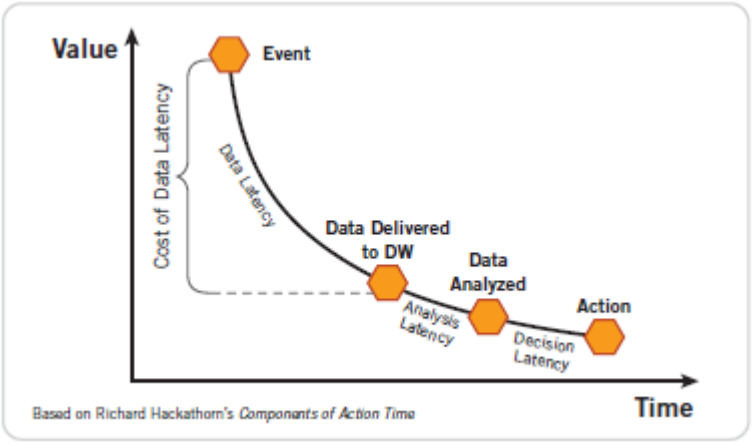


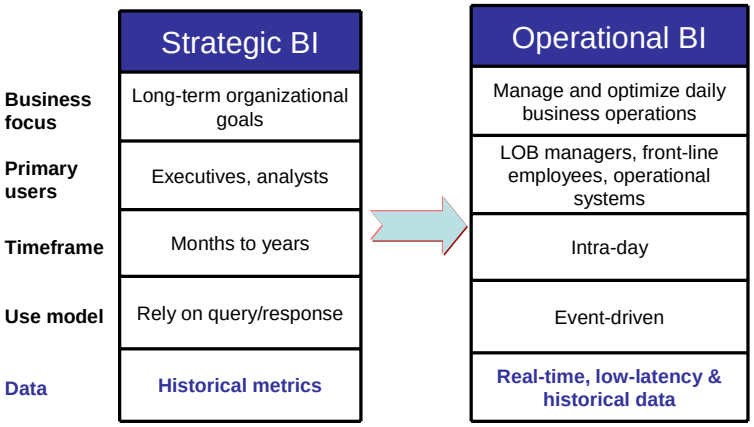
Figure 1. The longer it takes to capture and process data, the lower the value of the information.

Leading industry analysts have recently reported on this trend for BI and data warehousing based on the clear value achieved by companies that have deployed real-time capabilities. For example, in its *Overcoming Data Integration and Data Quality Challenges: End-User Survey Results* report (July

2008), IDC noted that "one of the emerging trends that stands out in the market is greater demand for real-time or near-real-time data integration." IDC's survey revealed that about 30 percent of organizations already employ some variation of real-time or near-real-time data integration. Another 40 percent are interested in doing so over the next 12 months. Similarly, in its *2009 Update: Evaluating Integration Alternatives* report, Forrester stated that "today's complex data integration requirements demand higher-quality data and more-robust metadata and auditability, with service level agreements (SLAs) requiring data delivery ranging from nightly batching to real-time services across heterogeneous IT ecosystems. To accommodate these changing needs, ETL software vendors have expanded their portfolios to include stronger metadata management, integrated data quality and profiling, as well as real-time data integration techniques including CDC and data federation."

The Data Warehouse Evolution to Operational Decision Support for Front-Line Users

Traditional data warehouses have focused on support for strategic BI—a resource for the small group of analysts and decision makers engaged in strategic planning that affects time horizons of months or years. Today, more and more companies maximize the value and competitive advantage of their data warehouse by using it in an operational role, adding mission-critical decision support to their workload. This new capability is referred to as operational BI. See Figure 2.



Based on Claudia Imhoff's "Differences between strategic, tactical and operational BI".

Figure 2. Data warehouses have moved beyond strategic planning into the realm of operational resources.

Long-term strategic decision making could be based on historical metrics derived from day-old or week-old data. In today's competitive business world, companies need to see ROI from their data warehouse and BI investments, not just in strategic planning but also in operational decision making. Particularly, front-line employees can provide more-responsive service and create efficiencies in their business functions if they have the most up-to-date information possible. By combining historical data with ongoing operational data, operational data warehouses enable a much-larger population of business users to make more-informed, proactive decisions.

Intelligence for operational execution includes product lookups, individualized customer offers, transaction exceptions, supply chain visibility, event detection, and notifications. Users numbering in the hundreds to tens of thousands can benefit from the information: gate agents, cashiers, dock workers, bank tellers, salespersons, customer service/call center agents, customers, and suppliers.

In operational data warehousing, the closer the warehouse is to real-time information, the more actionable it becomes for front-line users. These users need relevant information on what is happening right now and selected historical enterprise data as the organizational memory, to determine and influence what should happen next. For example, a retail store manager can respond to a sudden external event such as a snowstorm more proactively with up-to-the-minute inventory and pricing data than a competitor relying on yesterday's news.

The enterprise data warehouse helps personnel make excellent "small decisions" that collectively enhance competitive advantage in line with business strategy. For example:

- Should I expedite this package based on the criteria I see here?
- Should I extend a special offer or up-sell to this customer at this moment?
- Should I change our current marketing campaign based on this morning's results?

Ultimately, incorporating BI into operations facilitates automation, which improves efficiency. Using real-time data feeds, data warehouses can trigger business actions to automate more and more of these decisions based on predefined business rules.

High Availability and the Operational Data Warehouse

Because the operational data warehouse previously described is intricately woven into business operations, the highest level of availability is required for the data warehouse to support 24/7 operational decision making. Any downtime—planned or unplanned—now directly affects the business processes it supports. The underlying database must, therefore, support features at all levels to maximize availability in case of unplanned outages, such as hardware or software failures, and planned outages, such as database and application upgrade cycles. The data warehouse should also prevent data loss that typically happens due to user or application errors. These new requirements call for a significant shift in how enterprises think about data warehouses.

Data Acquisition Approaches for the Real-Time Enterprise

There are numerous technologies that serve data acquisition needs. One of the biggest differentiators among these solutions is the speed of data capture and delivery, as well as impact on the source systems. Only a few offer real-time data delivery with low system impact and no reliance on batch windows. Choosing the right solution requires a comprehensive understanding of organizational data requirements, including

- Data volume (size of data and number of updates)
- Date movement frequency
- Data integrity
- Transformation requirements
- Outage windows/business continuity

To clarify a common misconception, some data acquisition technologies often refer to "right-time" BI. *Right time* refers to the needs of the end users in accessing intelligence and can be different across different use cases. For operational data warehousing, the technology should deliver real-time capabilities and let the business user choose the right time to access the data.

Traditional Data Acquisition Approaches

Traditional data acquisition approaches include scripting, ETL, EAI, and real-time CDC. Scripts and ETL are batch oriented in data delivery, whereas EAI and real-time, log-based CDC support continuous data capture.

COMPARISON OF DIFFERENT DATA ACQUISITION APPROACHES

ATTRIBUTE	SCRIPTS	ETL	EAI	LOG-BASED CDC
Data volume	Medium	Very high	Low	High
Frequency	Intermittent	Intermittent	Continuous	Intermittent or continuous
Latency	Medium to high	Medium to high	Low	Low
Transactional	No	No	No	Varies by offering
Transformations	Intermediate	Advanced	Basic	Basic
Processing overhead	Intermittently high	Intermittently high	Continuous and medium	Continuous and low
Batch window	Yes	Yes	No	No

Scripts

Scripts are flexible and economical to develop, and almost every operating system can invoke scripts from their built-in scheduling facilities. However, scripts pose many challenges, such as being a drain on developer resource time and effort, as well as administrative challenges, such as manageability, documentation, and service-level agreement compliance.

Extract, Transform, and Load

ETL can be an ideal solution for the bulk movement of large volumes of data. Packaged ETL products also offer advanced transformation capabilities. As for data acquisition, ETL tasks are executed intermittently—typically during nightly maintenance windows when the datasources are quiesced, to ensure that datasources don't change during data acquisition and lead to inconsistencies across online transaction processing (OLTP) systems and the data warehouse.

To decrease data latency, some ETL products can perform—or be customized for—CDC capabilities. To support this configuration, ETL tools must store additional data in source tables, such as time stamps, to identify changed data since the last query. Most databases were not designed for such accommodations, and making changes to the database schema can create issues for the source applications. Such configurations could also place a burden on production systems, because complex queries would need to be run frequently across the whole database to identify changes. Further, capturing changed data with these systems does not necessarily lead to delivering data in real time. Typically with these solutions, the changed data is still staged on the ETL server, and eventually batch loaded into the target warehouse.

It is important to note that, in general, batch windows are disappearing. Significant growth in transactional data in OLTP systems, combined with the need to keep those systems and the operational data warehouse highly available at or near 24/7, leaves little to no time to allocate the batch window. At some point, a different data integration approach must be considered. Running several batch loads per day, such as minibatches or intraday batches, in an attempt to increase the frequency of refreshing the enterprise data warehouse, means overhead and business interruption trade-offs while never truly achieving a real-time infrastructure.

Enterprise Application Integration

Originally designed and intended for application integration, EAI solutions have evolved into a real-time data integration solution. EAI solutions continuously deliver data between source and target systems, provide fast data delivery, feature advanced workflow support, and facilitate basic data transformations. However, EAI imposes data volume constraints because these systems were designed to integrate applications, not data. EAI is designed to invoke applications and move instructions and messages and is an intrusive method for moving data from source systems. Nevertheless, with its ability to move data in real time, EAI solutions can support low-data volume environments for operational data warehousing needs.

Change Data Capture

CDC technologies identify and capture changes made to enterprise datasources, and then deliver those changes to target systems. As with other technology categories, not all CDC solutions are created equal. That is, not all are low impact or transaction aware. And the available offerings differ in terms of the overall solution overhead, latency, scalability, flexibility, data integrity, and recoverability. Although log-based CDC eliminates the batch window and heavy overhead on the source and provides data in subseconds, some CDC solutions still operate end to end in batch mode where an ETL product "pulls," or requests, periodically to receive a batch of all new changes made since the last request, and then performs transformations on batch data before loading the target system.

Real-time CDC solutions offer a continuous streaming, or "push," approach to delivering data. With such solutions, data changes are captured as they occur, and are then immediately pushed to the target data warehouse or the ETL system for performing the transformation. Total latency can be brought down to minutes or even seconds, making that data near-instantly available to drive operational BI and reporting. The real-time CDC solutions that capture the changed data from the database transaction logs do not impact the performance on the source systems, unlike offerings that use database triggers or table scanning.

Data Transformations—Where Do They Belong?

As data warehouses evolve and become more operational with the benefit of real-time data feeds, the requirements for transforming the data have also changed. As previously described, in traditional data warehousing, data acquisition tends to be batch oriented. Data moves between relational and multidimensional structures, and typically most of the transformations are handled on the chosen ETL engine.

As the data warehouse approaches real time, transformations tend to take place in the data warehouse. This is often called an ELT approach: extract, load, and then transform. The data warehouse stages and transforms the data to reduce data and analysis latency. This eliminates the need to aggregate changed data on a centralized server and removes an intermediate step from the overall data flow, as well as the associated costs of acquiring and managing the dedicated ETL server.

A major requirement for operational data warehouses that receive real-time data feeds is to handle both loading and querying workloads simultaneously. Enterprise data warehouses are increasingly being designed to support these mixed workloads so that the benefits of real-time data feeds can be fully realized. Leaders in data warehousing solutions, such as Oracle Exadata, support mixed workloads, enabling continuous data loading, dashboard updating, and prebuilt reporting with timely data. See Figure 3.

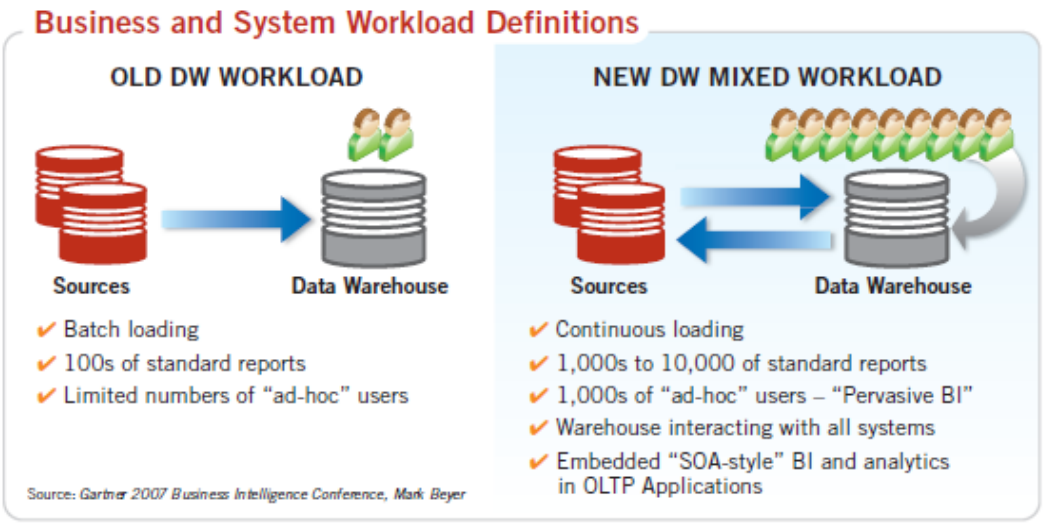


Figure 3. Data warehouses must support mixed workloads and ad hoc requests from thousands of active users.

Oracle GoldenGate for Real-Time Data Warehousing

To enable real-time data acquisition, Oracle GoldenGate uses log-based, real-time CDC capabilities to provide continuous capture and delivery of the most recently changed data between OLTP systems and the data warehouse. The application offers transactional, real-time data capture, routing, transformations, and delivery, using the push approach. As soon as a new database transaction is committed at the source system, that data is immediately captured via the database transaction logs and moved to the data warehouse where it can drive enhanced, strategic, and operational BI capabilities. Oracle GoldenGate can perform basic, row-level transformations at the point of capture or at the time of delivery. For heavy transformation requirements, Oracle GoldenGate can deliver the data to a staging area in the data warehouse for in-database transformations with Oracle Data Integrator Enterprise Edition (EE) to support an ELT architecture, or it can deliver the data to an existing ETL server (see the subsection titled "Combining Oracle GoldenGate with Extract, Transform, and Load Systems" for more details).

Oracle GoldenGate eliminates the need for batch windows, is extremely low impact, supports the movement of large data volumes, improves the ability to recover data in the event of a failure or outage, and moves read-consistent data with referential integrity.

As shown in Figure 4, Oracle GoldenGate and Oracle Data Integrator EE can both be used on the Oracle Exadata database. Oracle Exadata supports mixed workloads and enables very fast response times that front-line employees and line-of-business managers now demand.

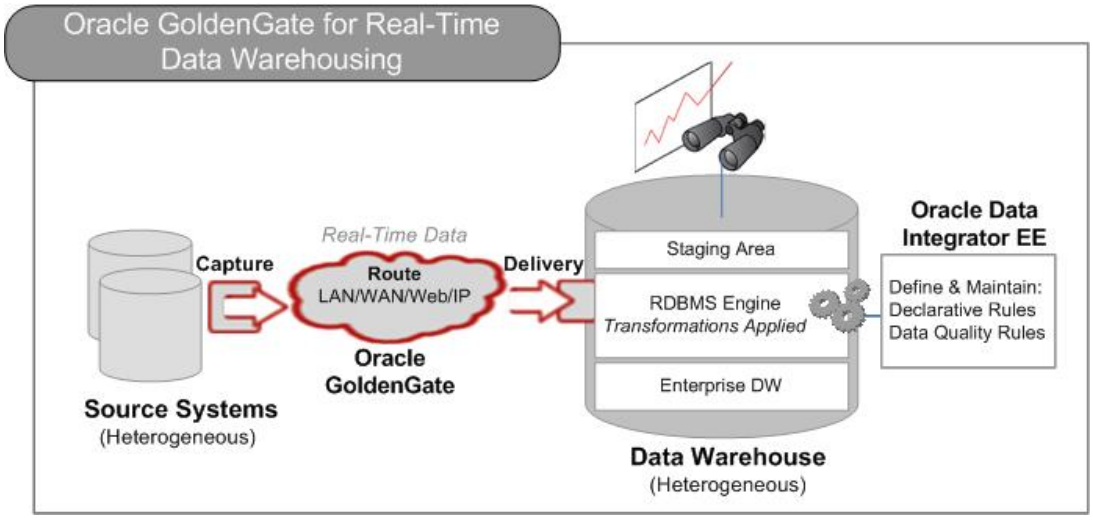


Figure 4. Oracle GoldenGate uses a componentized architecture to provide low-impact, real-time data capture and delivery to support operational BI demands.

Oracle GoldenGate Functional Overview

Oracle GoldenGate includes process modules for capturing, routing, transforming, and delivering transactional data in real time across heterogeneous environments. The application is designed to meet the needs of real-time data warehouse implementations.

Data Capture

The Oracle GoldenGate Capture module resides nonintrusively with the source database and continuously captures any new transactions. The new data is immediately moved into a database- and platform-independent universal data format called an Oracle GoldenGate Trail File. Trail Files not only enable heterogeneity but remove the risk of data loss or corruption, in the event of an outage at the source or target.

Data Delivery

Oracle GoldenGate's Delivery module continuously delivers all new data to the data warehouse, with end-to-end latency in subseconds. This means the most current data is always available for more-advanced, agile BI, actions, and reporting. In addition, because smaller sets of data are being moved at any given time—unlike batch methods—there is very little overhead imposed on the source and IT infrastructure. The Delivery module applies read-consistent data while maintaining referential integrity.

Transformation Support

Oracle GoldenGate provides built-in functions for row-level transformations. For complete, high-performance transformation requirements, it can be combined with Oracle Data Integrator EE. In this joint solution, users can perform set-based transformations inside the warehouse when higher end-to-end performance is desired. Due to this ELT architecture, no additional middle-tier server is needed.

Heterogeneity

Oracle GoldenGate supports log-based CDC for a wide range of database versions for Oracle Database, SQL Server, IBM DB2 OS/390 and LUW, Sybase ASE, Enscribe, SQL/MP and SQL/MX, and Teradata running on Linux, UNIX, Microsoft Windows, Oracle Solaris, and HP NonStop platforms. Oracle GoldenGate can deliver to a variety of data warehouses including SQL Server, Teradata, OracleDatabase, Netezza, Greenplum, HP Neoview and any warehouse running on an Open Database Connectivity–compliant database. Oracle GoldenGate can also be deployed with Oracle GoldenGate Application Adapters to deliver changed data to messaging systems.

Flexibility

Companies can quickly and easily involve new or different database sources and target systems to their data warehousing solutions by simply adding additional Capture and Delivery modules. This simplifies scalability and enables Oracle GoldenGate to extend solutions for moving data back to the OLTP system for any closed-loop activities, or for simultaneously sending data to reporting instances, data stores, backups, or other target systems. To further simplify the management of Oracle GoldenGate environments, users can deploy Oracle Management Pack for Oracle GoldenGate, a graphical user interface add-on product for speeding the deployment, monitoring, and reporting on all the Oracle GoldenGate process modules supported across the enterprise.

Combining Oracle GoldenGate with Extract, Transform, and Load Systems

For heavy transformation needs, Oracle GoldenGate can augment existing ETL implementations. In these configurations, Oracle GoldenGate handles the real-time, continuous data capture, or the "E" part of the ETL process, without impacting the source systems. There are several different ways Oracle GoldenGate can be used in combination with ETL or ELT systems.

- **Continuous feed to ELT staging area for maximum transformation performance.** In this best practice approach, Oracle GoldenGate is combined with Oracle Data Integrator EE to create an ELT architecture. Oracle Data Integrator EE delivers unique next-generation ELT technology that improves performance, reduces data integration costs, and works across heterogeneous systems. In this solution, Oracle GoldenGate delivers real-time data feeds to the target database's staging tables. Oracle Data Integrator EE extracts the data from the

staging area and loads the data into user tables, after performing transformations using the processing power of the database, which enables multifold transformation performance improvements over typical ETL architectures. In addition, avoiding a middle-tier transformation server also decreases the total cost of ownership of the data warehouse infrastructure. The savings can be used to configure the data warehouse for higher overall performance on all workloads, or can go straight to the bottom line.

- **Generate flat files for microbatches.** Oracle GoldenGate can provide real-time changed data in flat file format in minibatches or microbatches—such as every minute—to an existing ETL solution, which consumes the flat files to perform transformations and load the user tables. Rather than using a high-overhead ETL data extraction method, Oracle GoldenGate provides real-time, low-impact data acquisition for the transformation and loading processes. The end-to-end data latency for user tables depends on the transformation frequency and the speed of the existing ETL solution.
- **Delivery to enterprise messaging systems.** Oracle GoldenGate can also publish real-time data from source OLTP systems to a Java Message Service message queue or topic in XML or delimited text formats, from which the ETL systems can receive the changed data in real time before performing transformations and loading.

Combining Oracle GoldenGate's real-time CDC capabilities with ETL or ELT solutions enables the immediate, low-impact capture of new transactions on the source OLTP system. When compared to the overhead imposed by ETL's extract process, this combined solution provides a more cost-effective and efficient method for accessing timely information.

Oracle GoldenGate Customer Success with Real-Time Data Warehousing

There are numerous organizations that are seeing the benefits of real-time data warehousing today: leaders in banking and financial services, airline travel, telecommunications, manufacturing, retailers, and e-commerce businesses.

These organizations continue to push the envelope by turning their data warehouse into an integral part of the strategic and operational decision-making process. And they are achieving measurable gains in customer satisfaction levels as well as operational efficiencies—and ultimately, the bottom line.

The subsections that follow highlight two examples of this success. Overstock.com is a large e-business retailer experiencing exponential growth in customer base and data volumes. Montefiore Medical Center received an award from The Data Warehousing Institute for its use of Oracle GoldenGate to feed real-time patient clinical data into warehousing and reporting systems.

Overstock.com

Overstock.com is an online retailer offering discount brand name merchandise. Overstock.com offers customers the opportunity to shop smarter and more conveniently online for top-quality bargains In 2004, Overstock.com's revenue was up fivefold from 2002, and gross profits nearly quadrupled. With business growing at a rate of 100 percent a year, the company braced for continued growth and increasing customer demand.

Experiencing 14 to 18 million hits a month to its Website, Overstock.com recognized the need to both scale and streamline operations to better support its 24/7 customer transaction and reporting loads. Being an online retailer, there is zero tolerance for downtime. Overstock.com is always open, so its IT infrastructure must be continuously available. In addition, the company wanted to enable a real-time, single view of the customer to better understand purchasing habits, refine marketing efforts, and more-effectively drive business to its Website.

To achieve these objectives, Overstock.com decided to implement an enterprise data warehouse and customer analytics applications. The company selected Oracle GoldenGate to move customer data from the Oracle Database supporting its retail site into the data warehouse running the Teradata database.

Oracle GoldenGate allows access to data in real time, which enables Overstock.com to run reports around the clock, without putting additional strain on the operations system. In the past, the system could be tied up for long periods for a single data report, causing significant reduction in productivity. Moreover, Overstock.com was forced to treat all customers the same, whereas now the retailer can analyze customer behavior and purchase history to target marketing campaigns and service.

Overstock.com has already started to see the benefits of leveraging customer data in real time. Its vice president of data warehousing reporting and analytics says, "If we send out an e-mail campaign, we need to know if it's working. We need to know if consumers are clicking in the right place, if the e-mail is driving consumers to the site, and if those customers are making purchases. With Oracle GoldenGate, we can access this kind of data in real time, rather than waiting one, two, or even three days. The speed is phenomenal and the integrity of each transaction is preserved."

Montefiore Medical Center

The Data Warehousing Institute named Montefiore Medical Center, based in New York City, a winner of the 2006 Best Practices in Data Warehousing Awards. Montefiore won the award in the "Right Time Data Warehousing" category for its innovative use of Oracle GoldenGate to support real-time data warehousing.

Montefiore Medical Center, the University Hospital and Academic Medical Center for the Albert Einstein College of Medicine, ranks among the top 1 percent of all U.S. hospitals for investments in medical innovation and cutting-edge technology. Montefiore's unique combination of state-of-

the-art technology with "state-of-the-heart" medical and nursing care in a teaching and research environment offers its patients access to world-class medical experts, the newest and most innovative treatments, and the best medical center experience anywhere.

Montefiore was recognized for its pioneering approach to unlocking the value of historical patient and medical data to improve decision support. In this solution, Oracle GoldenGate is used to push critical patient data from the company's clinical information system (CIS) application, GE Healthcare's Centricity Enterprise, and feed the data in real time to a Sybase data warehouse and other servers for reporting purposes. From there, report-writing servers running the Clinical Looking Glass (CLG), a custom-built decision-support application, analyze patient data to help doctors and healthcare administrators make better, more-informed decisions around the delivery of care to hospital patients. Montefiore also uses Oracle GoldenGate for uninterrupted application and service availability of its CIS and physician order entry systems, which is especially critical in a paperless environment.

"The stakes are high when you're dealing with patient data," says Eran Y. Bellin, MD, director of outcomes analysis and decision support for Emerging Health Information Technology (EHIT), a wholly owned subsidiary of Montefiore. EHIT developed and continues to maintain the CLG decision-support software system at Montefiore Medical Center, under the direction of Dr. Bellin. "We believe real-time data warehousing for our Clinical Looking Glass system improves administration and decision making, and ultimately helps provide a higher quality of patient care."

Conclusion

Succeeding in today's competitive business environment requires good decisions. Operational data warehousing allows all users in the organization to access and respond to information in real time. Establishing and maintaining this real-time data warehouse requires a continuous low-latency data capture and delivery infrastructure. Oracle GoldenGate provides comprehensive functionality to continuously feed data warehouses with the most recent transactional data from OLTP systems using subsecond latency and without impacting the source systems.

Oracle GoldenGate provides the following key benefits:

- Real-time data for enabling more-advanced, agile BI
- Low-impact, high-performance data integration by reading database transaction logs
- Zero requirement for batch windows or using a middle-tier server
- Integration with Oracle Data Integrator EE for high-performance ELT architecture
- Support for large data volumes and heterogeneity
- Ability to augment existing ETL solutions with real-time, low-impact data acquisition

- Exceptional flexibility, easy implementation, and maintenance
- Robust data recovery after outages
- Ability to move read-consistent data with referential integrity

Organizations that leverage the most up-to-date BI in their day-to-day operations have seen significant improvements in operational quality, productivity, and customer service.

Real-Time Data Integration for Data Warehousing and Operational Business Intelligence
January 2010

Oracle Corporation
World Headquarters
500 Oracle Parkway
Redwood Shores, CA 94065
U.S.A.

Worldwide Inquiries:
Phone: +1.650.506.7000
Fax: +1.650.506.7200
oracle.com

Oracle is committed to developing practices and products that help protect the environment

0109