An Oracle White Paper
December 2009

Augmenting ETL Systems with Real-Time Change Data Capture

Executive Overview 1
Key Trends in Data Integration 2
Reducing Data Latency 2
Growing Data Volumes and Shortening Batch Windows 2
Improving Recoverability After a System Failure 3
Architectural Options 3
Perform Multiple Large Intraday Batch Extracts 4
Capture Changed Data in Minibatches 4
Augment ETL Systems with Change Data Capture Capabilities 4
Integrating Extract, Transform, and Load and Change Data Capture Technologies 5
Key Features and Functionality of Oracle GoldenGate 6
Augmenting Extract, Transform, and Load Systems with Oracle GoldenGate 7
Staging Tables 7
Direct Flat File Integration 8
Message Queue Infrastructure 9
Achieving Low-Cost, Real-Time Enterprise Data Warehousing 10
Solution Benefits 11
Proven Customer Success 13
Sabre Holdings 13
Overstock.com 13
Conclusion 14

Executive Overview

Competing effectively amid growing customer expectations requires business intelligence (BI) based on timely and accurate information. However, the traditional extract, transform, and load (ETL) data integration infrastructure in place at most organizations processes data in scheduled batches, which requires defined processing windows—typically during "off-business" hours— to transfer data to BI systems. Unfortunately, as businesses increasingly leverage online applications to operate 24/7 these processing windows are coming under severe pressure. Together, the exponential growth in transactional data volume and the need to provide continuous access to business systems, make these shrinking batch windows less and less effective at creating and maintaining a comprehensive and reliable BI environment.

This white paper discusses several options for addressing this challenge, with a particular focus on augmenting existing ETL systems with real-time change data capture (CDC) technology. Oracle GoldenGate provides real-time, log-based CDC capabilities that can coexist with ETL systems to simultaneously leverage existing investments while enabling fresh BI—without relying on batch windows.

Key Trends in Data Integration

Truly actionable BI requires the ability to effectively integrate data from a variety of sources. Some business operations require insight based upon the most current data to enable employees to take action with completely reliable and accurate information. To improve operational efficiency and effectiveness, companies are striving to create a robust BI infrastructure that uses timely operational information while retaining historical context.

Reducing Data Latency

Information is a time-perishable asset, especially for front-line employees who make many day-to-day tactical decisions, and its timeliness is directly correlated to its business value. As time passes, data decreases in value and becomes more out of sync with the events it represents. Businesses are beginning to realize the value of knowing about events that happened as recently as seconds ago. They are seeking more-timely information for improved operations and customer service, knowing that such improvements will in turn lead to a stronger competitive advantage and higher profits. For example, by monitoring the results of a price promotion campaign as it occurs, businesses can change course midday to improve results and increase revenues.

Growing Data Volumes and Shortening Batch Windows

Batch processing–oriented systems often scan source database tables and create a batch file to load onto the target database. This type of extract process creates high overhead on the source system and impacts transaction processing. In addition, to ensure read consistency during the extract process, this approach requires either extensive hand coding or accessing the source systems, which means that extraction must typically be performed during batch windows, most often at night.

For many years, batch processing was sufficient to meet BI requirements. However, with the growth of the internet, consumers transact business and expect access to services 24/7, which has effectively eliminated the concept of the "business day." As a result, the window available for batch data movement is shrinking, while simultaneously, the amount of data that must be processed is growing larger.

Faced with this dilemma, IT departments must employ highly efficient operations to process the data within batch windows, but many are finding it impossible to perform the required processes in the allotted time. (See Figure 1.)

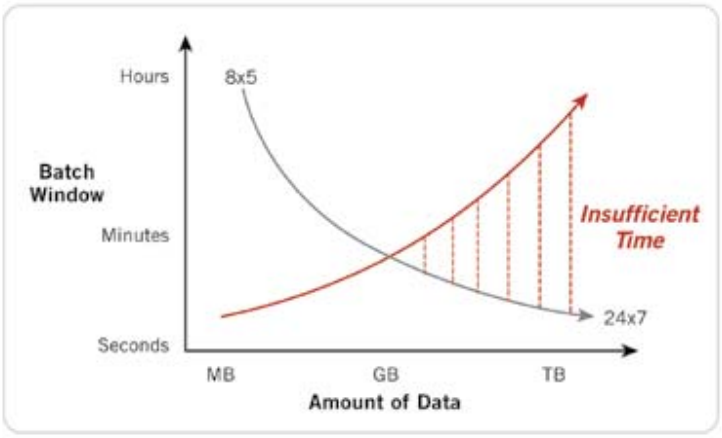


Figure 1. Data volumes are growing at a rate that exceeds the capacity of effective batch processes.

Improving Recoverability After a System Failure

If an extract or load process is interrupted, it might not have enough time after restart to complete the processing during the batch window. When an ETL system is configured to extract only the changes since the last batch via time stamps, this type of outage would result in an incomplete extraction due to interrupted access to the time stamps. The ETL system would process the next batch in the next batch window, capturing the data based on the new time stamps created. For data warehouses that operate according to a daily batch loading schedule, a day's worth of data would not be loaded into the data warehouse and would be unavailable for analysis.

In addition, if the ETL system has a process failure, or if the data warehouse is interrupted during the load process, the data warehouse might be left in an inconsistent state. After a data warehouse failure, a full recovery depends on whether the most recent batch files are still available for updating the tables. BI environments that support mission-critical operations need a better recovery solution.

Architectural Options

To access fresher data, organizations must gather it in a way that overcomes batch window restrictions and does not create an excess burden on online transaction processing (OLTP) servers. There are three basic architectural options:

- Customize the ETL system to extract multiple large intraday batches.
- Customize the ETL system to extract only the changed data in minibatches.
- Augment the ETL system with a dedicated solution for CDC.

Each of these options is described below along with the issues that organizations must consider when evaluating each option.

Perform Multiple Large Intraday Batch Extracts

With the option of customizing an ETL system to extract multiple large intraday batches, the ETL solution performs its full source extract processes several times throughout the day, such as every four to six hours. This strategy not only allows an organization to leverage its existing ETL infrastructure, but it also provides the enterprise with fresher data than a nightly batch-based system.

However, in this scenario, the ETL system must perform full scans of the production databases more frequently, which, due to the heavy overhead, can have a negative impact on mission-critical business systems, especially during peak hours. In addition, the overhead of full scan extracts might require costly hardware upgrades to the source systems. A frequency of every four to six hours is an improvement over once-daily batch routines, but it is still not real time.

Capture Changed Data in Minibatches

Some organizations have attempted to minimize data latency by building extended intelligence into the ETL system. In this approach, rather than extracting the entire contents of one database and delivering it to another, the system captures and delivers only the changed data. This results in data that is more current than data delivered via daily or intraday batch processing, and causes less network overhead because it moves less data at a time.

Nevertheless, this option requires considerable modification to the existing databases and ETL tasks. Often, additional data, such as time stamps, might have to be added to each table to identify whether there have been changes 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 also could place a burden on production systems, because complex queries would need to be run across the whole database to identify changes, and they would need to be run frequently throughout the day. Because of the time required to run queries for changed data and perform the rest of the processing through a system that is optimized to process data in batches, this method is not able to deliver changed data in real time.

Augment ETL Systems with Change Data Capture Capabilities

For companies that decide to leverage their existing ETL investments, augmenting their ETL system with a CDC solution can be the best of both worlds. The CDC solution can feed the ETL system with a continuous stream of real-time data, which not only decreases data latency for BI systems, but also eliminates the reliance on batch processing windows while allowing continuous operations for business-critical systems.

Although real-time data streaming technology is often perceived as being costly, recent studies show that, in many cases, this architecture results in a lower cost of ownership compared to modifying the existing ETL system for intraday batch or continuous minibatch loading.[1] As such, augmenting an existing ETL system with CDC technology is a viable option for the majority of organizations that operate ETL systems—it offers all the advantages of the previous options with none of the disadvantages.

Integrating Extract, Transform, and Load and Change Data Capture Technologies

When augmenting an ETL system with a CDC solution, the combined system should meet a list of stringent requirements to provide maximum performance and benefit.

- **High-performance, real-time change data capture.** The CDC system should be able to deliver data with subsecond speed, even with large data volumes, to enable continuous real-time data movement.
- **Support for heterogeneous data sources.** The CDC system should be able to support capture and delivery across different databases and platforms.
- **Compatible data formats.** The CDC system should be able to provide data in a format that the ETL system can easily read. Some are optimized to read batch files, others might read binary data, and others might work better with database tables.
- **Robust batch scheduling control.** The CDC system should allow users to control batch size, enabling them to set up batches as short as one minute. The system should be able to tell the ETL system immediately when a new batch is ready for processing.
- **Guaranteed delivery.** The CDC system should guarantee the delivery of changed data. It should avoid delivering duplicates and handle failures with fast, complete recovery. In addition, there should be a checksum mechanism that the ETL system uses to verify that all the data delivered by CDC has been read and consumed. With this step, users can ensure that all the data from source systems are delivered to the target environment.

[1] Richard Hackathorn and Jack Garzella, *Knowing Sooner than Later: Cost Analysis of Low-Latency Data in Enterprise Data Warehousing* (February 2007): download.101com.com/pub/tdwi/Files/Low-Latency_Data.pdf.

Key Features and Functionality of Oracle GoldenGate

Oracle GoldenGate supports the requirements for real-time data integration through continuous capture, routing, transformation, and delivery of transactional data between heterogeneous systems. The key application capabilities include

- **Real-time data movement.** Transactions are moved continuously from source to target systems with subsecond latency, removing the need for batch processing windows.
- **Transactional integrity.** Oracle GoldenGate moves transactions across heterogeneous databases with referential integrity, so that the true business history is kept intact.
- **High-performance, low-impact architecture.** The Oracle GoldenGate architecture only moves changed data from committed transactions by reading from transaction (or redo) logs, without creating extra tables. The application enables consistent high performance, supporting the delivery of thousands of transactions per second with negligible system impact or overhead.
- **Heterogeneous database support.** Transactional data can be integrated from and applied to many different databases—including Oracle Database, Sybase, Microsoft SQL Server, and IBM DB2—across all major platforms and operating systems. In addition, Oracle GoldenGate Application Adapters can be deployed to write transactional data to Java Message Service (JMS) queues and flat files.
- **Data transformation.** Oracle GoldenGate can apply row-level transformations and mappings within either its Capture or Delivery modules. It does not require a middle-tier server and supports table and row filtering based on user-defined criteria.
- **Continuous availability and reliability.** Oracle GoldenGate solutions operate without requiring system interruption or outage windows. They enable high availability and ensure no data will be lost or corrupted en route, even after an unexpected interruption or disaster.

Componentized Architecture

Oracle GoldenGate consists of decoupled components that can be configured to enable a variety of real-time data integration and continuous availability solutions (see Figure 2). This architecture allows each of the Oracle GoldenGate modules—Capture, Trail Files, and Delivery—to perform its tasks independently.

- **Capture.** The Capture module reads the result of insert, update, and delete operations by directly accessing the database's transaction (redo) logs and immediately captures the new data for distribution. The Capture component can send data to multiple destinations at once, which is also called a data pump configuration.

- **Trail Files.** Trail Files store the changed data in a transportable, platform-independent format outside the source and target databases to offer heterogeneity, improved reliability, minimal data loss, and negligible overhead.
- **Delivery.** The Delivery module applies Trail File data to any Open Database Connectivity (ODBC)–compliant database via native SQL, in the same order as it was committed at the source and with referential integrity.

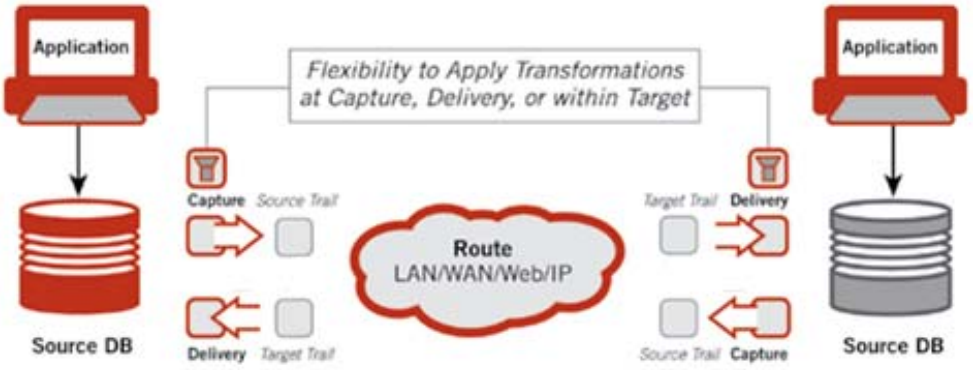


Figure 2. Oracle GoldenGate leverages a componentized architecture to rapidly and accurately replicate data in real time.

Augmenting Extract, Transform, and Load Systems with Oracle GoldenGate

Oracle GoldenGate offers a comprehensive selection of solutions for augmenting ETL systems with real-time, transactional CDC capabilities. These solutions help organizations capture the benefits of real-time BI while leveraging existing ETL investments.

Staging Tables

One approach to enhancing ETL infrastructure with real-time CDC is by staging tables. As shown in Figure 3, this configuration moves real-time transactional data from multiple production databases into staging tables within the data warehouse. The ETL system then pulls the data from the staging tables and loads the user tables after performing any necessary transformations. Using this approach, the ETL system never accesses the production databases directly, which eliminates any impact to the OLTP systems and enables near-real-time data—without depending on batch windows.

The staging tables can create a live copy of the transactional system to allow operational reporting to occur off of the primary OLTP system. The ETL system can gather data from the staging tables according to a schedule, or other user-defined criteria. Ultimately, the end-to-end data latency is determined by the ETL system's transformation schedule.

Because GoldenGate stores the changed data with transactional integrity and continues to capture changes even if there is an outage in the data warehouse, the staging tables can easily be restored to the latest point in source system transactions, minimizing the risk of data loss. This method fits best for ELT solutions, which perform the transformations within the target database, because it eliminates the need to move the data out of the target database.

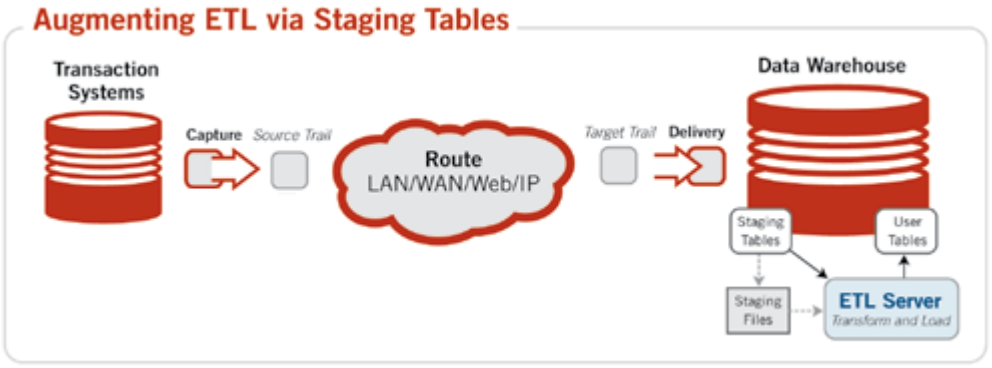


Figure 3. Using Oracle GoldenGate to load staging tables for ETL processes minimizes overhead and reduces data latency.

As a variation of this architecture, companies can use a separate database as the staging area, rather than staging tables in the primary data warehousing environment. In this method, if desired, the secondary database can act as a live standby system for the source database as well.

Direct Flat File Integration

Another method for augmenting ETL systems with real-time CDC is to have Oracle GoldenGate write the changed data into a flat file. The ETL system would then read from the flat file, perform the necessary transformations, and load the user tables of the data warehouse (see Figure 4). For many ETL systems, reading files is faster than scanning tables. For such systems, this method can handle higher data volumes than a staging database. It enables the user to decrease the data latency to minutes by configuring the frequency of microbatches, and it also eliminates batch windows. Similar to staging table approach, this architecture decouples the ETL processes from the source databases and allows the continuous capture of source database changes even if the ETL processes are interrupted. Because this method enables the processing of just changed data and does not use a staging database, it requires less system maintenance.

Oracle GoldenGate has the ability to provide data in a variety of formats, including delimited text files and binary files, to create the optimal feeding mechanism for existing ETL technology. In addition to delivering data in flat files, Oracle GoldenGate provides control files that can tell the ETL system when to read the data files. Users can set the size of the control files as well as the frequency of each microbatch period. Each file can then be rolled over every hour, every minute, or at any user-defined frequency. Due to the tracking allowed by the control file, the ETL server will always know what data it has to process—even after an interruption.

ETL systems can use Oracle GoldenGate's control files in two ways:

- The ETL system can continuously check for the presence of the control file in the designated folder, and when it is found, the system will know there is a new set of transactions to process.
- The ETL system can read the control file on a predetermined schedule and process the data that has been written since the last run.

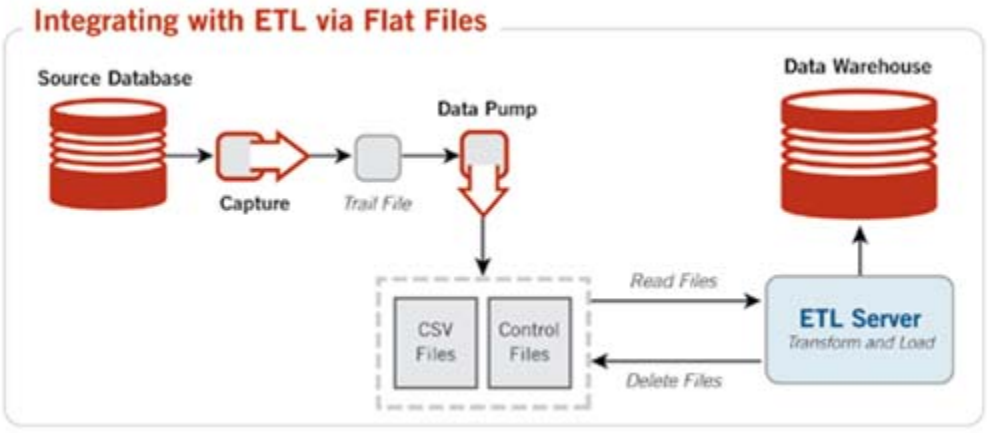


Figure 4. Oracle GoldenGate can write changed data to flat files for ETL systems to load a data warehouse.

Several organizations have successfully deployed Oracle GoldenGate to integrate with their ETL systems via flat files. One financial services organization chose this solution to accelerate its fraud detection routines, which it had been running every 24 hours. The company leverages Oracle GoldenGate to capture changed data from its source systems and write the transactions out to the flat files in near real time. The existing ETL system transfers the data to the data warehouse and into other decision-support systems, allowing automated fraud detection in minutes.

Another customer also relied on batch data processing every 24 hours. But for regulatory compliance purposes and to avoid legal fines, the company had to bring customer order information into the data warehouse on an hourly basis. By leveraging Oracle GoldenGate, the company is now able to meet this requirement with no impact on the source system.

Message Queue Infrastructure

Several major ETL systems now have the ability to read and process XML messages directly from JMS and other messaging queues. For organizations that have deployed a message queue infrastructure, Oracle GoldenGate can be implemented in conjunction with Oracle GoldenGate Application Adapters to publish real-time data from source OLTP systems to the message queue. The ETL system would then receive the changed data in XML format in real time. By using Oracle GoldenGate to publish transactions from source systems, companies can avoid the costly, time-consuming process of building different database or application adapters.

Note that with this method, the ETL system would still need to be configured to process XML data from that queue, be it JMS or another XML-based message queue format. Although this architecture provides real-time data to an ETL system without batch window outages, there is additional overhead in writing and processing XML, so it might not be best suited for instances that require extremely high transaction rates. However Oracle GoldenGate can publish transactions to a message bus in real time with subsecond latency (see Figure 5).

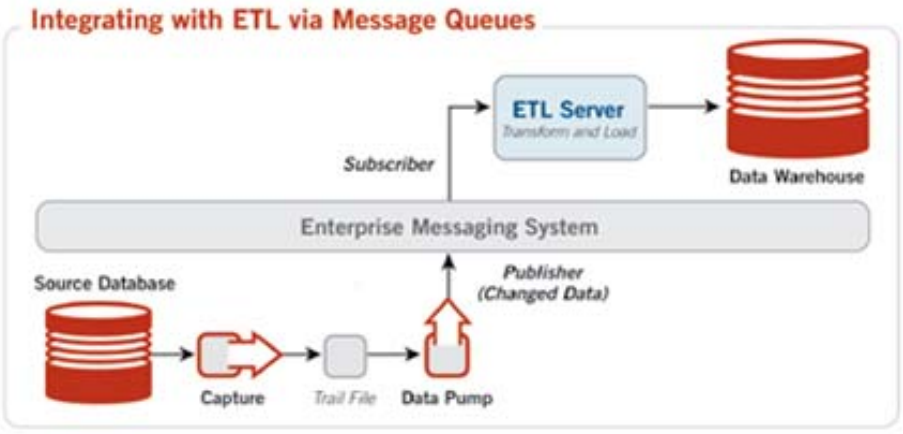


Figure 5. Oracle GoldenGate can publish transactional data to a message queue in real time for retrieval and processing by ETL systems.

Achieving Low-Cost, Real-Time Enterprise Data Warehousing

Deployed together, Oracle GoldenGate and Oracle Data Integrator Enterprise Edition create a high performance platform for real-time BI. Unlike traditional ETL data warehouse approaches, Oracle GoldenGate and Oracle Data Integrator Enterprise Edition enhance data integration performance by capturing data in real time and transforming that data inside the target system. Working in concert, these products eliminate the need to dedicate an additional ETL server—commonly required by traditional ETL approaches—and create a simpler and more-efficient architecture that reduces the TCO for real-time enterprise data warehousing initiatives (see Figure 6).

Each application offers unique value to create a powerful joint solution.

- **Oracle GoldenGate.** Provides real-time data acquisition for the data warehouse, by capturing data from disparate heterogeneous sources and delivering that data to the data warehouse instantaneously.
- **Oracle Data Integrator Enterprise Edition.** Delivers heterogeneous bulk data movement and high-performance data transformation with its unique E-LT (Extract, Load, Transform) architecture, which speeds implementation times and lowers infrastructure costs.

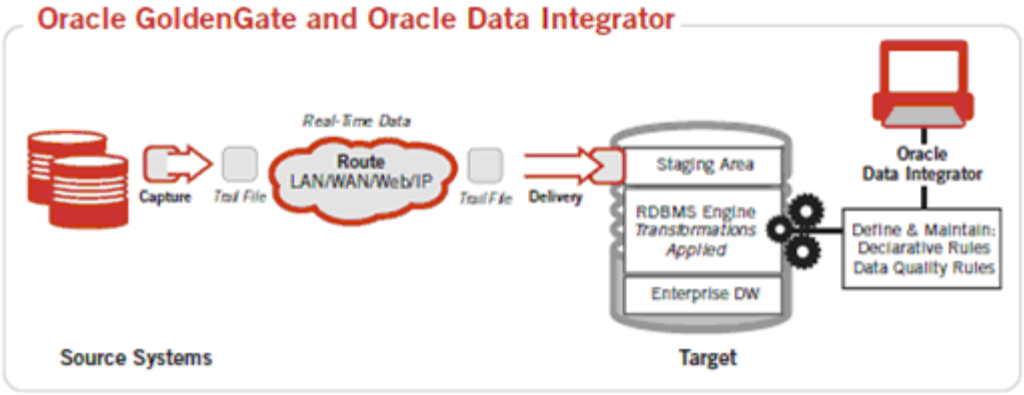


Figure 6. Oracle GoldenGate and Oracle Data Integrator Enterprise Edition create an efficient architecture to rapidly capture and transform enterprise data.

Solution Benefits

Deployed together, Oracle GoldenGate and Oracle Data Integrator Enterprise Edition create a compelling value proposition for organizations looking to optimize their data integration infractucture.

Create Tangible Business Value

The ability to accelerate data capture and transformation increases the competitive advantages of the enterprise data warehouse and improves BI. The combined architecture of these solutions enables organizations to improve several aspects of business information.

Eliminate Data Latency

Oracle GoldenGate captures new data transactions from one or more sources and instantly delivers that data to the target system. No additional intermediate server is required. Oracle GoldenGate continuously moves new data into the staging area of the data warehouse with only subsecond latency, meaning the most current data is always available for advanced, agile BI. Because smaller sets of data are being moved across the network at any given time—unlike batch processing methods—there is very little overhead imposed on the source systems and IT infrastructure.

Reduce the Time Required for Data Transformation

Oracle Data Integrator Enterprise Edition's E-LT architecture leverages disparate database engines to process and transform the data. This approach optimizes performance and scalability and lowers overall solution costs. Instead of relying on a separate, conventional ETL transformation server, Oracle Data Integrator Enterprise Edition's E-LT architecture generates native code for disparate RDBMS engines (such as SQL and bulk loader scripts). The E-LT

architecture extracts data from its sources, loads it into a target system, and transforms it using the power of the database. By leveraging existing databases and database expertise, Oracle Data Integrator Enterprise Edition provides unparalleled efficiency and lower cost of ownership. By reducing network traffic and transforming data in the database containing the target tables, E-LT architecture delivers the highest-possible data transformation performance.

Increase Information Technology Flexibility with Heterogeneous Database Support

Oracle Data Integrator Enterprise Edition supports all leading data warehousing platforms, including Oracle Database, Teradata, Netezza, and IBM DB2. This is complemented by the Oracle GoldenGate architecture, which decouples source and target systems, enabling heterogeneity of databases as well as operating systems and hardware platforms. Oracle GoldenGate supports a wide range of database versions for Oracle Database, SQL Server, DB2 z/Series and LUW, Sybase ASE, Enscribe, SQL/MP and SQL/MX, Teradata running on UNIX, Windows, and HP NonStop platforms. It can deliver data to any warehouse running on an ODBC-compliant database, including Oracle Exadata, Teradata, Netezza, and Greenplum. Companies can quickly and easily involve new or different database sources and target systems to their configurations by simply adding new Capture and Delivery processes.

Enhance Developer Productivity Through Declarative Design

Oracle Data Integrator Enterprise Edition shortens implementation times with its declarative design approach. Designers specify what they want to accomplish with their data, and then the solution generates the details of how to perform the task. Unlike conventional ETL designs, with Oracle Data Integrator Enterprise Edition, the business user or developer specifies the rules to apply to the integration processes. The solution automatically generates data flows, manages their complexity, and administers the correct instructions for the various source and target systems. With declarative design, the number and complexity of steps are greatly reduced, which in turn shortens implementation times. Automatic code generation reduces the learning curve for integration developers and streamlines access by non-IT professionals to the definition of their integration processes and data formats.

Improve Data Reliability

With data continuously and instantly being moved from the source systems to the data warehouse, it is critical that the data integration solution be capable of managing interruptions or outages. Oracle GoldenGate is designed for redundancy and reliability—ensuring no data loss or corruption. As previously discussed, captured data is immediately stored in Trail Files that can reside on both the source and target servers outside the databases. Once the systems are back online, the Trail Files retain the last capture of changed data, and immediately apply it to the target database.

Proven Customer Success

Many customers have deployed Oracle GoldenGate to enhance their existing ETL solutions with real-time CDC capabilities. Here are two of those implementations.

Sabre Holdings

Sabre Holdings is using Oracle GoldenGate and Oracle Data Integrator Enterprise Edition and has been able to reduce data latency and provide customized product offerings using fresh data drawn from customer interactions. In the company's implementation, Oracle GoldenGate captures changed data from various OLTP systems, such as ticketing and financial systems, and populates the staging tables within the data warehouse. Sabre Holdings uses Oracle Data Integrator Enterprise Edition to transform the data within the data warehouse and load the user tables. This combined solution supports the delivery of 200,000 to 230,000 records per hour to the data warehouse.

After adding Oracle GoldenGate to its ELT infrastructure, Sabre Holdings was able to minimize the amount of data movement, reduce the load on the original systems, and successfully consolidate data from multiple data marts to its enterprise data warehouse.

Overstock.com

Overstock.com deployed Oracle GoldenGate's real-time data integration solution to move customer transaction data from its Oracle 9*i* databases supporting its retail site into the Teradata Warehouse. Oracle Data Integrator Enterprise Edition is used for in-database complex transformations.

Overstock.com has seen the benefits of leveraging customer data in real time. When the company sends out an e-mail campaign, rather than waiting one, two, or three days, it immediately sees whether consumers are clicking in the right place, if the e-mail is driving consumers to the site, and if those customers are making purchases. By accessing the data in real time using Oracle GoldenGate, Overstock.com can immediately track customer behavior, profitability of campaigns, and ROI. The retailer can now analyze customer behavior and purchase history to target marketing campaigns and service in real time and provide better service to its customers.

In addition, in the past, the system could be tied up for long periods of time for a single data report, causing significant reduction in productivity. Because Oracle GoldenGate allows access to data in real time, Overstock.com can run reports around the clock, without putting additional strain on the operations system.

Conclusion

Oracle GoldenGate offers a set of robust solutions for augmenting existing ETL systems with the power of real-time, transactional, CDC technology. More-advanced, agile BI is achieved by combining Oracle GoldenGate with Oracle Data Integrator Enterprise Edition for real-time, higher-performance definition, execution, and maintenance of data transformations. Unlike traditional ETL approaches, this joint solution focuses on an E-LT architecture that drives down the latency of data, dramatically speeds the data integration process, simplifies the overall implementation, and delivers substantially better ease-of-use capabilities for active enterprise data warehouse solutions.

Through its support for a wide variety of topologies and architectures, Oracle GoldenGate allows organizations to leverage existing investments while simultaneously eliminating batch window dependency and overhead on source systems. This results in enhanced business intelligence and a stronger, sustainable competitive advantage.



Augmenting ETL Systems with Real-Time
Change Data Capture
December 2009

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

Copyright © 2009, Oracle and/or its affiliates. All rights reserved. This document is provided for information purposes only and the contents hereof are subject to change without notice. This document is not warranted to be error-free, nor subject to any other warranties or conditions, whether expressed orally or implied in law, including implied warranties and conditions of merchantability or fitness for a particular purpose. We specifically disclaim any liability with respect to this document and no contractual obligations are formed either directly or indirectly by this document. This document may not be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without our prior written permission.

Oracle is a registered trademark of Oracle Corporation and/or its affiliates. Other names may be trademarks of their respective owners.

0109