

WHITE PAPER

Oracle's Solution for Heterogeneous Data Integration

Sponsored by: Oracle Corporation

Steve McClure

August 2003

IN THIS WHITE PAPER

In this white paper, IDC examines the context of current business practices and organizational needs with respect to data integration and the approaches available to corporate IT to respond to those needs. We define data integration and discuss various factors to consider when developing and deploying data integration solutions. With this as an introduction, we then look at the various Oracle products and technology available for data integration and examine three case studies involving the use of Oracle's products.

SITUATION OVERVIEW

SEEKING COMPETITIVE ADVANTAGE IN A DIFFICULT ECONOMY

In 2002, IDC surveyed about 300 IT managers to investigate the IT issues they considered most important to them in the current economic environment, especially in the context of enterprise integration. Reflecting the importance of integration to their organizations, a majority of the respondents rated integration either extremely important or critical, and they indicated that the level of criticality would increase significantly in the next two years. This should come as no surprise, given the typical organization's collection of disparate, disconnected applications and associated databases.

Respondents were also asked to identify the most important business issues they had addressed or would address through enterprise integration. The top 3 most important business issues to be addressed through integration in rank order were:

1. Respond faster to changing business needs
2. Improve operational productivity of employees
3. Provide better service to customers

Essentially, companies are looking to integration to allow them to do more for less money.

At the core of most IT integration projects is data integration. Historically, this data integration has manifested itself in the form of database replication systems or data warehouses. These applications continue to be the major applications for data integration software.

More recently, IT systems have been called upon to support customer sales and service in near real time. Increasingly, IT projects have a requirement to provide real-time data integration so that comprehensive, up-to-the-minute customer information is

quickly available to a call center agent. The demands of business partners (e.g., within the supply chain) also continue to challenge the IT departments in their own and their partners' organizations. In situations such as these, data sources are almost always distributed and disparate, requiring data integration to enable the overall solution.

WHAT IS DATA INTEGRATION?

Data integration takes many forms, from simple file transfers to virtual database platforms. According to IDC, data integration software attempts to provide noninvasive, programmatic access to persistent structured data, whether in heterogeneous, homogenous, distributed, or centralized data sources. Organizations implementing data integration solutions would do so for one or more of the following reasons:

- Provide an integrated view of data.** Such a view may be in support of data usage beyond that governing the creation and maintenance of the data in its source. Data warehousing and business analytics fit here, as do enterprise information portals. The primary business benefit is a unified view of organizational information for business analytics and decision making.
- Allow multiple applications to behave cooperatively and harmoniously.** This is typically done using a messaging strategy where a consolidated resource queue publishes common messages, to which consumers of the queue can subscribe. The primary business benefit is a higher level of customer service and improved operational efficiency.
- Improve operational efficiency of the IT department.** Efficiency is achieved by consolidating the number of different data sources maintained, creating a unified virtual (or federated) database for use with new applications, or implementing a unified system for information sharing. The primary business benefit is improved developer efficiency, resulting in less investment in equipment and staff and more rapid response to changes in the business environment.

INTEGRATION CHALLENGES

Although, as stated above, a majority of IT managers rated integration as either extremely important or critical, adoption has been slow. In a survey completed in 2001, IDC asked technology and business professionals which of the following strategies they used to integrate ecommerce or call center applications with back-office or front-office systems:

- Standalone (i.e., no integration)
- File transfers
- File transfer with queued data
- Bidirectional replication
- Messaging
- Transaction messaging
- Transaction messaging with data synchronization

IDC found that 25% of the respondents had no integration at all; that is, these respondents had a standalone application unconnected to back-office or front office applications. Most other respondents were roughly split between a messaging strategy and strategies involving file transfer or bidirectional replication.

Although this survey is two years old, the responses to questions in the more recent 2002 IDC survey reveal that IT managers are still struggling with integration issues. In the 2002 survey, users were asked to identify the top 2 integration challenges facing them. Not surprising, they are:

1. Integration of new technology with legacy systems
2. Number and types of systems to be integrated

It is interesting to note that coordination with/between external partners and integration with externally hosted (ASP) applications rank at the bottom, indicating that organizations still primarily focus on integrating *internal* systems and data sources rather than connecting the organization externally.

Clearly the desire of users is to simplify the overall IT environment, including the collective set of applications being used by the business and supported by IT. But this has been the lure of every new wave of technology for the past two decades, including Common Object Request Broker Architecture (CORBA) a few years ago and XML and Web services today. IDC believes this survey data shows that organizations are facing what IDC has dubbed the "software complexity crisis," which is characterized by a large and highly complex software infrastructure accumulated over years of enterprise software deployments and which is now recognized to be in need of improved integration, organization, and management.

Further evidence of the software complexity crisis is revealed in another question asked in the 2002 survey, concerning top technology issues. The top 2 technology issues that will be addressed through integration in rank order are:

1. Support a technology architecture/standard (i.e., J2EE, XML)
2. Eliminate redundant technology/systems/applications

Clustered behind these two, in a second tier, are:

3. Unify multiple data structures/formats
4. Provide a common user interface across applications
5. Integrate heterogeneous systems and source codes

Based on our survey data, we offer the following observations regarding integration strategies:

- Requirements vary.** Projects involving data integration will have a wide variety of requirements. The application development and deployment tools and technology supplied by a vendor such as Oracle must be comprehensive enough to address all the requirements. For example, the technology must provide choices for connectivity and modularity so that users can select only required elements.
- Customizations are important.** Users still need the flexibility to develop and incorporate proprietary software to interface with their custom applications.

- ☒ **Approaches vary.** There is no dominant integration strategy employed by users for data integration. Some projects will require transaction messaging with data synchronization. Others can tolerate a more loosely coupled form of integration using message-oriented middleware.

INTEGRATION STRATEGIES

Clearly, given this range of requirements, there are a variety of different integration strategies, including the following:

- ☒ **Consolidated.** A consolidated data integration solution moves all data into a single database and manages it in a central location.
- ☒ **Federated.** A federated data integration solution leaves data in the individual data source where it is normally maintained and updated and simply consolidates it on the fly as needed. In this case, multiple data sources will appear to be integrated into a single virtual database, masking the number and different kinds of databases behind the consolidated view. These solutions can work bidirectionally.
- ☒ **Shared.** A shared data integration solution actually moves data and events from one or more source databases to a consolidated resource, or queue, created to serve one or more new applications. Data can be maintained and exchanged using technologies such as replication, message queuing, transportable table spaces, and FTP.

Oracle has extensive support for consolidated data integration and while there are many obvious benefits to the consolidated solution, it is not practical for any organization that must deal with legacy systems or integrate with data it does not own. Therefore, we will not discuss this type any further but instead concentrate on federated and shared solutions.

ARCHITECTURES: FEDERATED VERSUS SHARED

Federated data integration can be very complicated. This is especially the case for distributed environments where several heterogeneous remote databases are to be synchronized using two-phase commit. Solutions that do federated data integration access and maintain the data in place wherever it resides (e.g., in a mainframe data store associated with legacy applications). Data access is done "transparently" — for example, the user (or using application) interacts with a single virtual or federated relational database under the control of the primary relational database management system (RDBMS), such as Oracle. This data integration software is working with the primary RDBMS "under the covers" to transform and translate schemas, data dictionaries, and dialects of SQL; ensure transactional consistency across remote foreign databases (using two-phase commit); and make the collection of disparate, heterogeneous, distributed data sources appear as one unified database. The integration software carrying out these complex tasks must be tightly integrated with the primary RDBMS because, to be effective, the RDBMS must also provide all the other important RDBMS functions, including effective query optimization.

Data sharing involves the sharing of data, transactions, and events among various applications in an organization. It can be accomplished within seconds or over night, depending on the requirement. It may be done in incremental steps, over time, as individual one-off implementations are required. If one-off tools are used to implement data sharing, eventually the variety of data-sharing approaches employed begin to conflict, and the IT department becomes overwhelmed with an unmanageable maintenance burden.

What is needed is a comprehensive, unified approach that relies on a standard set of services to capture, stage, and consume the information being shared. Such an environment needs to include a rules-based engine, support for popular development languages, and compliance with open standards. GUI-based tools should be available for ease of development and the inherent capabilities should be modular to satisfy a wide variety of possible implementation scenarios.

The data-sharing form of data integration can be applied to achieve near real-time data sharing. While it does not guarantee the level of synchronization inherent with a federated data integration approach (e.g., if updates are performed using two-phase commit), it also doesn't incur the corresponding performance overhead. Availability is improved because there are multiple copies of the data.

CONSIDERATIONS WHEN CHOOSING AN INTEGRATION APPROACH

As noted above, there is a range of complexity in data integration projects — from relatively straightforward (e.g., integrating data from two merging companies that used the same Oracle applications) to extremely complex (e.g., managing replication across distributed, heterogeneous databases in multiple locales worldwide). For each project, the following factors can be assessed to estimate the complexity level. Pretend you are a systems integrator such as EDS trying to size a data integration effort as you prepare a project proposal.

- ☒ **Potential for conflicts.** Is the data source updated by more than one application? If so, the potential exists for each application to simultaneously update the same data.
- ☒ **Latency.** How synchronously must the data integration occur? Can it be an overnight batch operation such as the typical data warehouse? Must it be synchronous and with two-phase commit? Or, can it be quasi-real-time, where a two- or three-second lag is tolerable, permitting an asynchronous solution?
- ☒ **Transaction volumes and data velocity.** What are the expected average and peak transaction rates and data processing throughput that will be required?
- ☒ **Access patterns.** How frequently is the data accessed and from where?
- ☒ **Data source size.** Some data sources are going to be so big that backup and availability become primary issues.
- ☒ **Application and data source variety.** Are we trying to integrate two ostensibly similar databases following the merger of two companies that both use the same application, or did they each have different applications? Are there multiple data sources that are all relational databases? Or are we integrating data from legacy system files with relational databases and real-time external data feeds?
- ☒ **Data quality.** The probability that data quality adds to overall project complexity increases as the variety of data sources increases.

One point of this discussion is that the requirements of data integration projects will vary widely. Therefore, the platform used to address these issues must be a rich superset of the features and functions that will be applied to any one project.

SUMMARY OF APPROACHES

Table 1 summarizes the advantages and typical applications of each approach. The Oracle features and products noted will be discussed in detail later.

TABLE 1

FEDERATED VERSUS SHARED DATA INTEGRATION APPROACHES

Type	Oracle Features and Products Employed	Advantages	Typical Application
Federated (consolidating data on the fly by directly accessing data in place)	<ul style="list-style-type: none"> • Oracle Distributed SQL • Ultra Search 	<ul style="list-style-type: none"> • Multiple remote databases can each be accessed transparently at their sources • Applications can be run against another vendor's database without being recoded • Multiple heterogeneous data sources appear in a consolidated local virtual view for users and applications • Distributed transactions and Distributed SQL optimization • If integrated into DBMS as with Oracle, higher availability and higher performance • Immediate, real-time data synchronization 	<ul style="list-style-type: none"> • Federated database • Integration with legacy applications • Migration • Virtual data warehouse (typically where data either changes often or is rarely accessed) • Grid computing
Shared (moving the data source to the need)	<ul style="list-style-type: none"> • Oracle Streams • Transportable Table Spaces • Materialized Views 	<ul style="list-style-type: none"> • Data and other enterprise information can be consolidated in a unified queue • Enables sophisticated publish/subscribe distribution • Combinable with other automated procedures • Higher performance because transactions are performed on local databases • Higher availability because there can be multiple copies of the data 	<ul style="list-style-type: none"> • Event notification • Data warehouse • Message queueing • Replication • Supply chain integration • Grid computing

TABLE 1			
FEDERATED VERSUS SHARED DATA INTEGRATION APPROACHES			
Type	Oracle Features and Products Employed	Advantages	Typical Application
		<ul style="list-style-type: none"> • Each database has all the data required to support operations against it • Quick, efficient resync of disconnected, replicated databases 	

Source: IDC, 2003

DATA INTEGRATION SOLUTIONS PROVIDED BY ORACLE

Oracle continues to be the world's leading provider of database management software. Its relational database software has the largest market share in the worldwide relational database software market. The Oracle9i RDBMS is at the center of Oracle's support for data integration. Oracle9i includes the features, functions, and capabilities that enable an organization to integrate its data regardless of where or how it is maintained. With the exception of Oracle Transparent Gateway, the Oracle data integration features are integrated with Oracle9i, allowing an organization to efficiently adapt the Oracle9i capabilities to fit its specific needs. The result is the attainment of data integration benefits, such as faster time to market, with less development effort and lower total cost of ownership (TCO).

We break down our discussion of Oracle's products into four parts:

1. Implementing federated data integration
2. Implementing data integration for data sharing
3. Dealing with heterogeneous data sources
4. Enabling integrated search of data and content with Ultra Search

Oracle's data integration solutions are commonly adopted by organizations where the primary enterprise database involved in the integration project is an Oracle database, (e.g., Oracle9i). At the same time, other database management products such as DB2, Sybase, the SAP file system, flat files, Web services, or other data types may also be included in the mix of data sources to be integrated. In every case, the data integration solution will rely on functions and features of the Oracle9i RDBMS.

It should also be noted that in those cases where packaged enterprise applications are being integrated, a comprehensive enterprise application integration (EAI) platform, which uses features such as those found in Oracle9iAS, will be employed. EAI platforms are beyond the scope of this white paper. It is also worth noting that Oracle provides a substantial set of features for data consolidation, which is not covered in the discussion below.

ORACLE FEDERATED DATA INTEGRATION

Oracle's solution for federated data integration relies on the Oracle9i Distributed SQL feature, which allows Oracle9i to be used to implement a distributed database system composed of two or more databases. Using Distributed SQL, applications can access and modify data in all databases, whether local or remote. If remote, the environment optionally hides the location of the data, thus simplifying its access. Access and modification for all data is simultaneous and synchronous. The Distributed SQL feature makes multiple, distributed databases look like one Oracle database to any application. A user of a local database can access a database link to a remote database without being a user on the remote database. This transparent access simplifies the use and administration of both the databases and the applications.

A distributed transaction that alters data on multiple databases is particularly complicated. To manage this complexity, Oracle implements two-phase commit across all participating nodes in the distributed system in a way that is completely transparent to the user and the application. The user simply issues standard SQL statements such as COMMIT or ROLLBACK. Significantly, Oracle9i Distributed SQL supports this approach, even if non-Oracle databases are involved in the distributed transaction. This relieves the user of the significant burden of maintaining the integrity of the collection of disparate databases participating in the transaction.

Distributed query optimization is also provided to reduce unnecessary data movement when remote nodes are involved. Oracle's cost-based optimization, which can be guided by the user with parameters, is designed to arrive at the particular SQL statement that is optimized for the most efficient levels of data transfer and processing.

These Oracle9i features add up to a very powerful solution for a federated or virtual database application, since the features are fully integrated into the Oracle9i DBMS. This means that any application can take advantage of Oracle9i's underlying scalability, performance, integrity, and security while accessing and updating both Oracle and non-Oracle databases in one integrated, synchronous view. This reduces the TCO in two ways: It simplifies use and administration and lowers development and maintenance costs.

ORACLE DATA SHARING

Oracle's solution for data sharing is Oracle Streams, available with Oracle9i release 2. Oracle Streams enables the propagation and management of data, transactions, and events in a data stream either within a single database or from one database to another. Streams supports a publish/subscribe mechanism where the user has control over the creation, processing, flow, and termination of each transaction.

Oracle Streams has three elements: capture, staging, and consumption. Each element supports major languages and standards such as C/C++, PL/SQL, JMS, SOAP, or XML/SMTP; is governed by rules; is open and interoperable; and supports optional transformations. The Oracle rules engine can be employed in each phase to apply rules to various actions or activities.

- ☒ **Capture.** Streams captures events in two ways — enqueue and log capture. Enqueue is used to capture user messages and events from other systems. Log capture is used to capture Oracle RDBMS changes (Data Manipulation Language [DML] and Data Definition Language [DDL]) from one or more sources and process and queue them.
- ☒ **Staging.** All events are unified into a single stream and published in a staging area. Staging areas can subscribe to other staging areas according to rules. Once queued, the information can be processed (e.g., data can be quality checked, validated, enhanced, transformed, and reformatted), and then applied in several ways.
- ☒ **Consumption.** Events in the stream, under rules control, can be applied in several ways — directly by the receiving database engine, passed to a user-defined function, or dequeued to an application(s).

Oracle Streams is a powerful feature that can be flexibly employed and will easily support extensions as new requirements occur. It is a single, unified solution for information sharing, which allows more efficient development and deployment of data integration solutions. Oracle Streams is useful for almost any asynchronous solution. Some examples are:

- ☒ Message queuing
- ☒ Message-driven procedures
- ☒ Event notification
- ☒ Replication (by capturing logged changes that are then routed and applied)
- ☒ Data warehouse (extract, transform, load)
- ☒ Unified information sharing

TRANSPORTABLE TABLE SPACES

Transportable table spaces allow administrators to efficiently move large amounts of data from one database to another. Database data is stored in table spaces, and those table spaces map to physical files. The traditional means for moving data from one database to another is either a distributed query or an export followed by an import. With large amounts of data, this can be very time consuming. With a transportable table space, an administrator can export metadata about the transportable table space; physically move the files using efficient mechanisms, such as FTP (if needed); and then import the metadata to mount the table space at the destination. The metadata import/export is very fast, regardless of the size of the table space. Contrast this to a full import/export, which can be quite time consuming if there is a great deal of data. A user can also use the transportable table space feature to mark a table space read-only and then mount that table space on more than one database simultaneously.

MATERIALIZED VIEWS

Materialized views allow an administrator to maintain point-in-time copies of data. This is useful for disconnected systems, such as those on the laptops of a roving sales force, or any application where you only care about the latest version of the data, not the intermediate values. For example, if an organization had a reporting system that needed to refresh its data only every 24 hours, it could do so on a daily basis using the materialized view feature. Materialized views support fast refresh (change data only), as well as subsetting based on sophisticated criteria, including subqueries.

Oracle's support for data sharing allows users to develop applications that maintain replicated copies of data across geographic regions that can be updated anywhere and kept consistent in near-real time. This is considerably more flexible than traditional replication solutions. Both data and events can be managed with whatever transformations are required. Because it can handle data structure changes, the copies do not have to be identical, and their evolution causes fewer maintenance problems. This gives the user much greater control. These features and Streams' routing ability allow the resulting application to be more resilient, thus ensuring higher availability.

DEALING WITH HETEROGENEOUS DATA SOURCES

A heterogeneous environment for the purposes of this discussion is one involving one or more of the following: non-Oracle data sources, non-Oracle message-queuing software, or non-SQL applications — in other words, environments where Oracle software must interoperate with other vendors' software. To gain the promised benefits of Oracle's data integration solutions, this interoperability needs to be achieved as transparently as possible so that application developers don't have to customize their applications to deal with heterogeneous data sources (i.e., they can build on one consistent interface).

- ☒ **Non-Oracle data sources.** Oracle provides Transparent Gateways to achieve transparent interoperability with other major RDBMSs, such as DB2, SQL Server, and Sybase. In addition, it offers a Generic Connectivity feature for interoperability via ODBC and OLE DB. This allows for access to data stores for which Oracle does not have a Transparent Gateway.
- ☒ **Non-Oracle message-queuing software.** Oracle provides the Messaging Gateway feature to support communication between Oracle Streams and other non-Oracle message-queuing systems, such as IBM's MQ Series.
- ☒ **Non-SQL applications.** Oracle's approach to this requirement is to offer a variety of open interfaces with which users can interoperate with third-party applications or allow users to access Oracle9i databases from their own client applications.

TRANSPARENT GATEWAYS AND GENERIC CONNECTIVITY (TO NON-ORACLE RDBMS)

By creating synonyms for objects in non-Oracle databases, Oracle9i users can interoperate seamlessly and transparently with them. Thus, for example, the capability of Oracle Distributed SQL to make remote databases appear local can be extended to a remote DB2 data source as well. With this type of support, applications need only interoperate through one consistent Oracle interface. This greatly simplifies a seemingly complex development task because all the heavy lifting for heterogeneous data source interoperability can be relegated to

Transparent Gateways and Generic Connectivity. The heavy lifting to which we refer is the processing and management required to account for the differences between each vendor's RDBMS. These differences exist despite their nominal compliance with industry standards such as SQL. The automated transformations required to solve this problem are threefold:

- ☒ **SQL translations.** Differences between Oracle's implementation of SQL and that of each participating non-Oracle database are accommodated by translating between the two, no matter how subtle they may be.
- ☒ **Data Dictionary translations.** Users must have the capability to query the metadata of the remote, non-Oracle database — this is similar to SQL translation and typically involves SELECT statements.
- ☒ **Data type translation.** Situations arise when data types need to be translated transparently. An example, is translating a DB2 PACKED DECIMAL data type to an Oracle NUMBER data type

OPEN INTERFACES (FOR NON-SQL APPLICATIONS)

As mentioned earlier, Oracle offers various open interfaces that allow users both to interoperate with third-party applications and to access Oracle9i databases from their own client applications. Examples are:

- ☒ eXtensible Markup Language (XML)
- ☒ Java Messaging Service (JMS)
- ☒ Oracle Call Interface (OCI)
- ☒ Open Database Connectivity (ODBC)
- ☒ Java Database Connectivity (JDBC)
- ☒ Callouts to external procedures
- ☒ Web services
- ☒ XML Query Language (XQuery)

The objective of these interfaces is to provide reliable communications between loosely coupled components in a distributed environment. Note that in the case of Web services, a user can wrap the data returned by the Web service to make it look like a SQL data source. This enables Web services data to be used in SQL operations.

ULTRA SEARCH: ENABLING INTEGRATED SEARCH OF DATA AND CONTENT

Users and vendors alike have begun to take a more holistic view of information retrieval that transcends the traditional view organized around concern for specific formats or data types. This is in anticipation of the eventual merger of structured and unstructured data (sometimes referred to as content). To address this requirement of information integration, Oracle provides Ultra Search. For an in-depth discussion of Ultra Search, please refer to another recent IDC white paper written for Oracle, *Envisioning the Ultimate Enterprise Search System* found at <http://otn.oracle.com/products/ultrasearch/content.html>.

Oracle Ultra Search and the Oracle Text technology on which it is based are both features of Oracle9i and are free. Ultra Search offers current generation functionality that can perform integrated search of both structured and unstructured data across more than 150 file formats, including HTML, Microsoft Office, PDF, and XML documents. It provides standalone, Web-style search and browsing for intranet or extranet and is seamlessly integrated with Oracle9iAS Portal and Oracle Collaboration Suite. Ultra Search can build an index into an unstructured document that can then be incorporated into an SQL or XQuery query, thus providing a form of integration of structured and unstructured data. The result of the search is a set of citations of relevant documents and records.

The Oracle Text technology is heavily invested in XML, the semistructured middle ground that is becoming a lingua franca among content technologies. XML will allow Oracle to extend its expertise in structured database technology by adding structure to unstructured content, thus enhancing search precision. Oracle intends to use the XML orientation to drive more complex search syntaxes in the future, such as XML Query over diverse data sources. Thus, Oracle has laid the foundation to move toward information integration, something that today is supported by only a few small startups. Note that IDC defines information integration to involve a form of hybrid query that allows structured and unstructured information to interact and be combined.

The capabilities of Ultra Search (discussed in detail in the referenced white paper) include a variety of classification/categorization and clustering algorithms, relevance ranking, concept searching, support for 70 languages, rules-driven searching, proximity matching, ease of use, and customization. Like the database of which it is part, Ultra Search is designed to scale and integrate seamlessly with other Oracle9i capabilities, such as security features.

CASE STUDIES

BANCA IMI

Gruppo Sanpaolo d'Intermediazione Mobiliare is Italy's second largest bank and among the top 50 banks worldwide. The investment banking arm of the group is Banca d'Intermediazione Mobiliare (Banca IMI). In addition to servicing the other parts of the Sanpaolo Group, Banca IMI provides investment banking services to a wide range of institutions, including other banks, asset managers for major global corporations, and other financial institutions. In the process of buying and selling a variety of financial instruments, Banca IMI must communicate with all of the major exchanges (e.g., the New York Stock Exchange). The volume of its daily trades can frequently scale up to hundreds of thousands.

Because its business operations are global, its IT systems must operate 24 x 7, and transactions with both internal applications systems and external trading centers must be processed with minimum error and as close to real time as possible. This is a very demanding business application and a complex example of data integration. The main applications with which Banca IMI must communicate include its own back-end administrative systems and legacy applications, the financial exchanges, domestic and international customers, and financial networks. Not only do all of these have different protocols and formats, but there are also differences just within the financial exchanges themselves. The latter is handled via a marketing interface layer with custom software for each exchange. Domestic customers are connected via SNA, and FIX (financial information exchange standards) is used with international customers and other financial institutions.

Banca IMI uses Oracle Streams (Advanced Queuing) to coordinate transactions across all these stakeholders. Because Streams is a fully integrated feature of the Oracle database, and therefore can take full advantage of Oracle9's security, optimization, performance, and scalability, Streams can accommodate Banca IMI's very high level of transactions in close to real time and still perform all the transformations required to communicate in the various protocols needed. Scalability is very important to Banca IMI and the primary reason why it moved to Oracle Streams from a previous solution that relied on another vendor's product. "We're happy about what Advanced Queuing offers us," says Domenico Betunio, Banca IMI's manager of electronic trading. "Besides speed and scalability, we're impressed by messaging reliability, and especially auditing."

HONG KONG INSTITUTE OF EDUCATION

The Hong Kong Institute of Education (HKIEd) is a leading teacher education institution in the Hong Kong Special Administrative Region (HKSAR). The institute was formally established by statute in April 1994 by uniting the former Northcote College of Education, Grantham College of Education, Sir Robert Black College of Education, the Hong Kong Technical Teachers' College, and the Institute of Languages in Education, the earliest of which was started in 1939. The Institute plays a key role in helping the Hong Kong government fulfill its commitments: to develop new curriculum; to achieve its goal of an "all graduate all trained" teaching profession; and to provide for the continuous professional development of all serving teachers. The Institute is organized around four schools with a current enrollment of nearly 7,000 students in a variety of daytime and evening degree programs. The Institute staff exceeds 1,000, almost 400 of whom are teaching staff. Across the Institute, more than 200 funded research and development projects are being actively pursued. The two main languages the Institute must accommodate are English and Traditional Chinese.

Soon after its founding, HKIEd began in-house development of several administrative applications, all running in conjunction with Sybase databases. These applications included student admission, enrollment and profiling, human resources and payroll, smart card management, and a library interface. In 2002, HKIEd purchased a set of packaged applications: Banner, from SCT. SCT's Banner system runs on Oracle and is analogous to an enterprise resource planning system. It supports student services, admission, enrollment, and finance functions, some of which it took over from the in-house Sybase applications. The Sybase in-house applications still account for roughly 50% of the Institute's administrative applications, including classroom booking, HR, payroll, JUPAS student selection, smart card management, and the library INNOPAC system.

Since these vital Institute administrative systems are on two platforms, Sybase and Oracle, there is a requirement to keep the data consistent in more than 10 common data fields. They account for less than 5% of the total number of fields, which is still significant. Each database contains more than 10,000 records. The number of daily transactions affecting these data fields ranges from 200 to 5,000.

After experimenting with SQL*Loader scripts to update the common fields using a batch upload process, HKIEd switched to using the Oracle Transparent Gateway. Since January 2003 HKIEd has been using the gateway to access and update the Sybase database from Oracle to keep the common data fields in sync in real-time. HKIEd has created views in the Oracle database based on a distributed join of tables from the Oracle and Sybase databases. This enables SCT's Banner system to transparently access and update fields in the Oracle and Sybase databases. The system automatically performs a two-phase commit to preserve transactional consistency across the two databases. The Transparent Gateway has NLS support, enabling access to Sybase data in any character set.

INTERNET SECURITIES, INC.

This case illustrates the use of Oracle Streams for information sharing, load balancing and high availability.

Internet Securities, Inc. (ISI), a Euromoney Institutional Investor company is the pioneering publisher of Internet-delivered emerging market news and information. Internet Securities (www.securities.com) provides hard-to-get information through its network of 20 offices in 19 countries, covering 45 national markets in Asia, Central and Eastern Europe, and Latin America. Its flagship product, the Emerging Markets Information Service aggregates and produces unique company and industry information including financial, economic and political news, for delivery to professionals over the Internet. The subscription-based service enables users to access and search through a comprehensive range of unique business information derived directly from over 6800 leading local and international sources. Primarily because of its international clientele, the operations of ISI are run on a 7x24 basis. ISI has offices in 18 locales around the globe, with clients in each locale. Its provisioning operations are centralized and located with its headquarters in New York City.

ISI's content is also global and emphasizes information about emerging markets, which in this context means markets in countries like Romania, Brazil or China. The content being aggregated arrives in automated feeds of various forms at an average rate of 50,000 documents a day, with hourly arrival rates ranging from 100 to several thousand per hour. All documents, regardless of the source language, are converted to a single encoding standard (UTF8) when being loaded into the ISI document base.

One of ISI's competitive differentiators is that information is retained regardless of age. The size of the ISI content base has grown rapidly to over one terabyte and in tandem, the level of query activity has grown as well. Until recently, daily operations were run on NT-based systems using Oracle8i. The need for high scalability and availability while superceding performance prompted ISI to migrate its database operations onto Solaris using Oracle 9i. Oracle Streams was selected to achieve a major increase in availability and performance. Higher availability is obtained by fully replicating to a secondary server and through being able to perform much faster backups. The performance improvements come from load balancing between the servers and the upgraded hardware.

Overall, Oracle Streams is used for three databases supporting ISI operations. Each database is replicated to a secondary server. The three are:

- Backoffice database (100 Gb) – This database support the company's proprietary CRM and authentication systems.
- Document database (50 Gb) – This contains metadata about documents and true paths to the physical location of documents.
- Search database (1 Tb) – This is the database in which the documents are loaded and where the client queries are executed. Documents are stored as BLOBS. Each record is one document.

The replication for each database is such that either replica can service the functions of both, but for performance (load balancing) and administrative reasons both are typically serving different operational needs. For example, ISI call center agents primarily use "Backoffice A" while "Backoffice B" services all of the client activity

monitoring. In the context of the Document database*, "Documents A" is the production machine, and "Documents B" is the standby. In the case of the huge Search database, "Search A" is used to perform the document loading that occurs in hourly batches, and "Search B" is used to receive the users' queries that can be executed on either Search database server.

ISI expected to obtain benefits in two ways: performance and availability. It had evaluated other possible solutions, but these fell short in the performance dimension. Oracle Streams did not. With the Streams-based solution, queries are executing in half the time or better. With respect to availability, switchover in case of a failure is now instantaneous. With NT, ISI was using a physical standby that could be switched over in the best case in 7 to 8 minutes for read only, and 10 to 15 minutes for read/write transactions. ISI is happy with the improved level of activity and the speed of replication it has obtained with Streams and the Global Operations team is also complimentary about the support it has received from the Oracle services organization.

FUTURE OUTLOOK

KEY TRENDS

There are two important categories of trends that must be considered in the context of data or information integration. Both of these trends are informed and shaped by the trends in the business environment discussed earlier in this paper. Most of these trends have been emerging for past few years.

- ☒ **Data/information trends** describe the nature of data/information expected in the future.
- ☒ **Platform trends** describe the nature of computing environments expected in the future.

DATA/INFORMATION TRENDS

Data/information trends can be summarized in one word, "more." IDC expects there will be both more data types and more data. In a 1999 U.C. Berkeley study, the amount of information in the world was estimated to be one petabyte. Digital information accounted for roughly 60% of this total (nondigital contributors were film and X-rays, and a minor part for paper publications such as books), and growing at a phenomenal rate of 80% per year. As each year passes, more information is "born digital" (e.g., photographs). A fair amount of this data is in personal systems. The types of digital information are also increasing because it has become feasible to capture and maintain them. We can rest assured that organizations will find ways to use these new forms of digital information to gain competitive advantage, regardless of industry.

The implications of these trends are that any solution for data integration will need to add support for an ever-increasing diversity of data types and related metadata. Oracle has already begun to address some of this requirement using Ultra Search, with its object/relational database support and with its extensions to support XML. The rapidly increasing amount of data worldwide guarantees long-term market growth for Oracle and its data integration solutions.

* As of this writing the Document database is not yet utilizing Streams. It is expected to begin replicating via Streams in the fall of 2003.

PLATFORM TRENDS

Platforms, or computing environments, generally are also expected to evolve and diversify. IDC sees four interesting trends here.

- Grid computing
- Embedded and handheld
- Web services
- Linux

GRID COMPUTING

The trend toward grid computing is itself driven by innovations in hardware (computer blades and hardware clusters), OS trends (low-cost Linux, discussed below), the increasing acceptance of the view of computing as a utility, and the efforts of organizations, especially in this economy, to gain efficiencies in the use of their IT resources. Efficiencies come in the form of improved resource allocation, information sharing, and high availability. The most relevant efficiency to this white paper is information sharing.

In this context, information sharing means delivering information to a user or process regardless of where it resides. Because a grid will involve multiple systems, it is likely to be composed of heterogeneous systems — that is, file systems and databases from multiple vendors. All of the Oracle9i components and features discussed above are applicable in a grid environment. For example, companies can provision a database by using transportable table spaces to move a copy of the data to an idle instance. They can use Distributed SQL to access data they didn't move, and they can keep the data they did move in sync with Oracle Streams.

Oracle has specifically addressed grid developers by embracing the Globus Toolkit, the emerging standard for building grids. This is complemented with the Globus Resource Allocation Manager, which provides resource allocation and process creation, monitoring, and management services, including a plug-in to invoke PL/SQL routines or SQL commands specified in Globus Resource Specification Language. The portability required for the grid is ensured for Oracle9i because it runs on all major operating systems.

EMBEDDED AND HANDHELD

IDC believes that the future of the computer industry is not on the desktop. Computers will be ubiquitous. This trend is well under way with the success of handheld devices and telematics. It will be extended to include our everyday environment, rooms, furniture, appliances, and even clothing. Wireless is also playing a large role in this revolution, as has the Internet. Hence we already see practical applications for the mobile workforce.

From a database perspective, this means that there is a requirement to extract data subsets, supply them to the mobile worker, and later resync the transactions created while operating offline with the primary database. For example, a sales person, or service person would receive a set of daily assignments first thing in the morning along with all the relevant data for those calls. The person would record transactions about the calls during the day and then resync at the end of the day (or week). A doctor making rounds might do the same with the relevant set of his or her patients. Oracle's approach to information sharing ensures that data can be shared with the types of devices that will be used in these scenarios.

WEB SERVICES

IDC believes that Web services will be adopted in phases. What is occurring now is either limited to intranet use or to extranets between business partners with whom many technical details have been prearranged. Thus, Web services is part of the overall application being used, but in a carefully controlled context. Longer term, we believe it can be adopted for use in a broader context (e.g., exchanges), but even in this case, negotiations will be required to settle on a common set of definitions and an ontology for the domain in question. We believe that free-form use of Web services relying on Universal Description, Discovery, and Integration (UDDI) to "discover" relevant services automatically will be difficult to implement successfully because of issues related to semantics and semantic mediation. The exception may be those areas where there are more universally accepted semantics, such as currency conversion. Typically those cases will be limited to the more fundamental data types.

The impact of Web services on data integration is fundamental because the result of requesting a Web service will almost always be the return of data. In addition, some Web services will require the receipt of data as input. In many instances this will require extraction and transformation of data prior to the invocation of the Web service. Conversely, the data returned by the Web service may need to be integrated with local data. As we noted earlier, Oracle's Web services support allows users to wrap the data returned by the Web service to make it look like an SQL data source. This enables Web services data to be used in SQL operations.

LINUX

Linux has emerged as a serious contender in the operating environment space. As IT managers continue to look for cost savings in this difficult economic environment, Linux offers savings not only in the cost of the operating system but also in cost of the underlying hardware systems. Another benefit of Linux is that it is open source, which reduces the risk of vendor lock-in. However, it is only recently that Linux's support of enterprise-level performance has been brought up to acceptable levels. This is important for enterprise-level adoption beyond its role as a Web server or file/print server. It will help accelerate the availability of enterprise applications on Linux. Although only a small fraction of platforms worldwide, the number of Linux installations is growing faster than those of other operating systems.

ORACLE'S STRATEGY

It is Oracle's stated strategy that it views information integration as a key component in modern information management solutions. Oracle believes customers require the flexibility to choose among the various approaches for information integration (consolidated, federated, shared) and build solutions that possibly utilize all three approaches. Oracle intends to continue to provide capabilities integrated into the Oracle database and application server that enable customers to most efficiently and easily build a solution that meets their needs.

Oracle also says it is carefully monitoring the trends that affect information integration. As noted by IDC, the volumes and heterogeneity of data will continue to increase. Oracle's information integration tools will help customers better manage their vast stores of heterogeneous digital information across databases, file systems, and applications. Oracle plans to continue its support of all platforms, including SMP, commodity servers, and mobile devices. More important, Oracle sees its heterogeneous information integration capabilities as a complete solution for provisioning structured and unstructured data within the grid, and thus these capabilities will be key enablers in unlocking the potential of future computing architectures.

CONCLUSION

Data integration, already an important issue for IT managers, is widely expected to become a critical issue in the near future. Organizations that have been in existence for more than a year or two are challenged to provide an integrated view of the relevant information they need to support their operations because the information must be integrated from multiple, distributed, heterogeneous information sources. Typically, they will be integrating legacy files, relational data, and newer, nontraditional data. With rare exceptions, these organizations can expect that the size, velocity, and diversity of these data sources will increase.

Vendors are offering data integration solutions with benefits that can provide a substantial basis for convincing return on investment. Typically, these solutions provide some way to make the differences seem more transparent. The benefits include faster development and therefore faster response to business need, lower total cost, easier administration, and easier maintenance.

Oracle understands that users have a wide variety of data integration needs and that one solution does not meet every user's requirements. One size does not fit all. Oracle offers a very broad range of features in Oracle9i, from which a user organization can select the appropriate ones to fit a solution to its particular data integration needs.

COPYRIGHT NOTICE

External Publication of IDC Information and Data — Any IDC information that is to be used in advertising, press releases, or promotional materials requires prior written approval from the appropriate IDC Vice President or Country Manager. A draft of the proposed document should accompany any such request. IDC reserves the right to deny approval of external usage for any reason.

Copyright 2003 IDC. Reproduction without written permission is completely forbidden.