FOURTH QUARTER 2007

TDWI TECHNOLOGY MARKET REPORT

DATA INTEGRATION TOOLS

Comparison and Market Analysis

By Philip Russom
and Mark Madsen

tdwi
THE DATA WAREHOUSING INSTITUTE

DATA INTEGRATION TOOLS

Comparison and Market Analysis

By Philip Russom and Mark Madsen

Table of Contents

About the Authors

PHILIP RUSSOM is the senior manager of TDWI Research at The Data Warehousing Institute, where he oversees many of TDWI's research-oriented publications, services, and events. Prior to joining TDWI in 2005, Russom was an industry analyst covering BI at Forrester Research, Giga Information Group, and Hurwitz Group. He also ran his own business as an independent industry analyst and BI consultant, and was contributing editor with *Intelligent Enterprise* and *DM Review* magazines. Before that, Russom worked in technical and marketing positions for various database vendors. You can reach him at prussom@tdwi.org.

MARK MADSEN is president of Third Nature, a technology consulting and market research firm focused on business intelligence, data integration, and data management. Mark is an award-winning architect and former CTO whose work has been featured in numerous industry publications. He is a principal author of *Clickstream Data Warehousing* and frequently speaks at conferences and writes about business intelligence and emerging technology. For more information or to contact Madsen, visit http://ThirdNature.net.

About TDWI Research

TDWI Research provides research and advice for BI professionals worldwide. TDWI Research focuses exclusively on BI/DW issues and teams up with industry practitioners to deliver both a broad and deep understanding of the business and technical issues surrounding the deployment of BI/DW solutions. TDWI Research offers reports, commentary, and inquiry services via a worldwide Membership program and provides custom research, benchmarking, and strategic planning services to both user and vendor organizations.

About TDWI's Technology Market Reports

TDWI Technology Market Reports provide TDWI Members an annual overview of an important technology sector within the business intelligence (BI) market. Technology Market Reports highlight the major events in the sector for the previous 12 months, predict the segment's future direction, and provide a comparative review of the leading products in the sector as well as a summary description of niche segments and players. The reports aim to help business customers create a shortlist of products that they can evaluate in more depth before making a purchase, or to validate the direction and capabilities of an existing product.

Introduction

Data Integration Practices, Tools, Suites, and Platforms

As we'll see in this report, data integration (DI) is practiced in different ways, with different tools and techniques, in response to different technical and end user requirements. The dizzying array of options is itself a barrier to action. To help technical users clear the barrier, this report segments the leading practices, tools, related technologies, and suites for DI.[1] Based on the segmentation presented in this report, a technical user should be able to identify a DI practice that's appropriate to his/her organization, understand what combination of tools and technologies is required, then draft an evaluation list of vendor products that maps credibly to his/her requirements.

But first, we need to define key terms and concepts used in this report.

- **Data integration (DI).** In its complex manifestations, DI collects data from multiple sources, transforms and integrates this disparate data into a common data model, and loads the integrated data into a target database, application, or file. In its simple forms, DI merely extracts data from one source and copies it into a target. When done well, DI adds value to data by improving its content (which may require an additional data quality solution) or by creating data structures that wouldn't exist without DI (which is key to data warehousing).
- **DI practices.** The term *data integration* is a broad umbrella that includes multiple DI practices, namely: extract, transform, and load (ETL), enterprise information integration (EII), enterprise data replication (EDR), and enterprise application integration (EAI). Technical users implement these practices with hand-coding, vendor tools, or a mix of the two.[2]
- **DI-related practices.** Data integration is a data management practice, as are its multiple practices. To further complicate the matter, data integration is regularly practiced in tandem with other data management practices, including data quality, data profiling, metadata management, master data management, sort, and so on. Hence, a DI implementation of any size and maturity is today rather complex, involving a collection of DI and related practices, possibly with a unique vendor tool or hand-coded solution for each.
- **Data integration suites.** When technical users began applying multiple DI and DI-related practices to their DI initiatives, software vendors responded by building and acquiring more tools. As a result, several DI vendors (like Business Objects, IBM, and Informatica) now have suites of tools or modules—one each for individual DI and DI-related data management practices. The main benefit of a tool suite is that users have a single vendor to work with for the acquisition and support of multiple tools. The problem, however, is that tools and modules of the suite (especially when the vendor acquired them, instead of building them) do not integrate and interoperate deeply, if at all.

The most influential trend today is the evolution of DI tools into suites and (eventually) platforms.

[1] *Segmentation* is an analytic method that reveals the constituent parts of a thing, then sorts the parts by criteria like cost, complexity, type, approach, technologies required, or applicability to a known goal.

[2] For a detailed comparison of DI practices, see the TDWI Best Practices Report *Data Integration: Using ETL, EAI, and EII Tools to Create an Integrated Enterprise* (November 2005), available online at www.tdwi.org/research.

- **Data integration platforms.** The trend toward DI tool suites has influenced vendor product offerings deeply in recent years. Yet, it's now being superseded by a grander trend toward DI platforms. The DI platform seeks to correct the suite's lack of interoperability via the full integration of common tool elements, including those for deployment (metadata, servers, security, interfaces for data access or interoperability with other tools) and development (user interfaces for design, collaboration, management, security).

Note that DI platforms don't exist yet. A few years will pass before vendors finish assembling their suites and integrate them into true platforms. In its discussions of vendor products, this report focuses on the transition from DI suites to DI platforms, because this trend is the strongest defining criterion for DI product offerings today. In turn, it's an issue that confuses technical users who must select a vendor and a product. Suites and platforms make it difficult to sort out which vendor has which kind of tool, as well as which are best-of-breed and which interoperate appropriately. This report seeks to alleviate some of this confusion.

This report highlights ETL because it's the preferred approach for analytic DI, which involves business intelligence and data warehousing. And ETL is well on its way to becoming the preferred approach to operational DI, which involves database consolidations and migrations.

Data Integration Market Overview

Year in Review

The past 12 months have seen a lot of action in the DI tools market. Several major acquisitions accelerated the already brisk pace of market consolidation, while the advent of new low-cost providers, including open source DI vendors, brought more choices and increased pressure on product pricing. A number of vendors introduced innovations that will enable DI to penetrate organizations more deeply.

Key Events in the Past 12 Months

Here are the top-line industry events during the last year:

Market consolidation continues as large vendors buy small ones.

- **July 2007 – IBM acquires DataMirror.** IBM's Software Group continued its shopping spree (at least 30 acquisitions this decade) by acquiring DataMirror Corporation, which brings best-of-breed enterprise data replication (EDR) into IBM's portfolio. Although IBM already has replication tools and replication capabilities built into various products (like DB2), the DataMirror Integration Suite is more open to heterogeneous environments. Plus, it supports advanced functions not seen in most replication tools, like data transformation and bidirectional data synchronization. DataMirror's changed data capture and real-time operation should help bolster IBM's on-demand computing strategy. This acquisition reminds us that replication—though as old as computing itself—remains a valuable data integration practice.
- **Early 2007 – Talend Software launches.** Talend joins the small but growing community of vendors—including Apatar and Pentaho—that offer commercially supported open source ETL software. Now that open source has arrived for DI tools, users have more options to consider. Though not completely free (support still has a price), open source DI tools cost

considerably less than their proprietary cousins. And it's now possible to start with an open source ETL tool and consider moving to more expensive proprietary software in the future, if requirements demand it.

- **Early 2007 – HP acquires Knightsbridge.** This regional system integrator is known for its well-respected data warehouse practice. But Knightsbridge also has a DI practice that's considered one of the best for "big data"—that is, integrating terabyte-scale volumes of data. As no coincidence, HP announced the acquisition of Knightsbridge just before announcing NeoView, HP's new multi-terabtye-scale data warehousing appliance.

New DI vendors and products continue to emerge.

- **Late 2006 – IBM launches Information Server.** This is an important milestone on IBM's path to a unified DI platform, because Information Server pulls together several servers and tool user interfaces to make working with multiple DI and DI-related products from IBM more seamless. In many ways, IBM Information Server is a result of the product integration initiative Ascential Software started before being acquired by IBM in 2005. IBM's DI platform isn't finished or fully unified, but IBM has shown a commitment to making it so.
- **October 2006 – Oracle acquires Sunopsis.** Oracle has been a DI vendor for years, offering Oracle Warehouse Builder (OWB), a batch-oriented ETL tool designed largely for use with Oracle databases. Sunopsis is complementary, with its focus on ELT and real-time operation. Curiously, however, Oracle has folded Sunopsis into its Fusion Middleware product line, where it provides near-time DI in the context of an application integration suite. After all, most enterprise application integration (EAI) tools are weak on query and DI, which are useful when application integration is data-intense.
- **May 2006 – IBM acquires Unicorn.** This move gained IBM an independent tool for enterprise metadata management. At the time, IBM representatives described Unicorn as the eighteenth integration- or process-oriented acquisition since 2001. This acquisition highlights how important metadata management is to data management in general and DI specifically.
- **Mid-2006 – Business Objects redefines EIM.** The concept of enterprise information management (EIM) has been around for years with a focus on database administration. But Business Objects' redefinition puts DI and DI-related practices at its core, along with their close ties to BI platforms for reporting and data analysis. Business Objects' EIM product offering includes most of the modules this report requires of a DI platform, which makes Business Objects—well known as a major BI vendor—also a leading DI platform vendor.

Key Trends in the Past 12 Months

Below is a summary of the top trends in the DI tools market:

- **Vendor offerings are evolving from individual tools to suites and platforms of tools.** As pointed out earlier, the trend that most defines vendor products has been, for several years, the movement toward suites of tools for DI and DI-related tasks (like data quality, profiling, and master and metadata management). The trend toward DI suites is currently evolving into a trend toward DI platforms. The difference is that a suite is a collection of largely autonomous tools, whereas a platform unifies them into fewer servers and a common user interface for development and deployment across all tools. Although this trend is about vendor products, it's driven by the rising user practice of applying multiple DI tools to single initiatives.

The leading driver for DI vendor acquisitions is the build-up of DI suites and platforms.

- **Market consolidation continues as vendors acquire each other.** DI suites and platforms require multiple DI and DI-related tools, and there's a limit to how quickly a software vendor can build new tools. Hence, the trend toward DI suites and platforms is the leading driver for most of the acquisitions made this decade in the DI vendor community. This is especially true of acquisitions made by Business Objects, IBM, and Informatica, which are at the forefront of DI platform development. The ramification of market consolidation is that users have fewer vendors to buy tools from. This is good for users who wish to consolidate suppliers, but not so good for users who desire an independent DI vendor.

- **Integration approaches are converging.** For example, EAI, EII, and replication continually improve their data transformation and bulk data capabilities such that they more closely resemble ETL. At the same time, ETL is improving near-time operation to resemble the other approaches. As vendors combine overlapping tools in suites and platforms, more convergence occurs. This makes it harder for users to choose the best DI approach for a project, but allows them to stretch the use of a DI tool to cover some of the capabilities of other tools.

The two leading challenges to DI success today are data volume and real time.

- **The volume of data continues to increase.** Scaling up to terabyte-scale data volume is the leading challenge to DI today. To achieve scalability, both user-built solutions and vendor tools rely heavily on parallel processing, distributed DI architectures, clusters or grids of servers, and the massive addressable memory of 64-bit servers.

- **Data collection occurs more frequently, pushing toward real time.** While the 24-hour cycle is still the norm for running DI jobs that refresh target databases, DI is more and more asked to access, collect, and integrate data multiple times a day. The consequence is that DI must still support older batch functions, but also newer, real-time ones. As users embrace time-sensitive business practices like on-demand computing, zero latency, and performance management, they need DI to operate in multiple speeds or frequencies, which a multi-tool DI suite or platform does.

- **Federated approaches to DI continue to gain ground.** Enterprise information integration (EII) is by nature federated, as are database functions like database views (whether materialized or not). The point of federation is to leave data where it originated and access its most recent value on an as-needed basis. The use of federation in DI solutions continues to increase, but at a glacial pace. Note that federation is one of the capabilities you should look for in a DI suite or platform.

- **Hub-and-spoke is the most common DI architecture, soon to be joined by services.** First, recognize that DI merits architecture, just as other IT systems do. Without architecture, DI deteriorates into a tangle of unorganized one-off interfaces. Second, hub-and-spoke is still the basis of most DI architectures (regardless of DI practice), although when multiple DI practices are applied, each may have a hub that interoperates with other DI hubs. Third, most users today apply Web services as independent interfaces that contradict DI architecture. As more users learn how to organize a true service-oriented architecture (SOA), expect to see service hubs for DI as well as DI solutions exposed via service hubs.

Hub-and-spoke is the preferred architecture for integration implementations.

The Year Ahead

Continuance of Established Trends

Of the trends just mentioned, scalability and real-time are the most pressing for users because these are now considered standard requirements, yet are still difficult to achieve. Hence, when users plan new DI work or updates to old work, they should allocate ample man-hours and new technology acquisitions. The slow adoption of federated DI and changes to DI architecture are not so pressing. As vendors' suites continue to evolve into platforms, users of DI tools will need to decide carefully which additional tools to acquire, which upgrades to apply, and what design changes to make in DI solutions.

New or Emerging Trends

You can expect these trends to continue to gain momentum in the next 12 months, just as they have done for years now. But these are also joined by new or emerging practices and technologies:

- **Operational DI.** DI isn't just for BI anymore. Analytic DI—usually manifested as ETL in support of data warehousing—continues to grow as an established practice. But its blue-collar sibling—operational DI—is growing even faster, as DI is regularly applied to operational database and application consolidations, migrations, synchronizations, and upgrades.
- **Collaborative DI.** As the practices of analytic DI and operational DI have grown, so have the number of DI specialists in data warehousing teams, data integration competency centers, and on other teams. Very recently, mildly technical business users (like brand managers and business analysts) are demanding hands-on access to DI projects and their development artifacts. As the DI team gets larger and more diverse, DI tools must provide more collaborative functions for these people.
- **DI high availability.** Data integration is being asked to operate more frequently per day, as well as in real time. This is needed to support business methodologies that demand fresh data, like operational BI, on-demand computing, and performance management. These methodologies can't manage a business without fresh data delivered reliably via real-time DI, so DI must be continuously available to enable them. Hence, when you cross the line into real-time DI, you also cross into DI high availability as a new requirement, which is met by fault-tolerant hardware and software or a cluster of DI servers that supports failover.

Real-time DI requires DI high availability, an often overlooked requirement.

- **Cross-business DI.** Long caged by the corporate firewall, DI is now unchained and roaming the Internet. As evidence, note the many DI and BI vendor tools that have recently added connectors for Salesforce.com, the quintessential extra-enterprise application. Since a lot of cross-business communications pass via EDI and XML-based documents, support for these and other semi-structured data standards (and translations among them) is a rising requirement for DI solutions.
- **External data and the Web gain importance.** Many organizations provide access to internal data but have difficulty meeting the user demand for external data. Few are bringing external data from outside sources or Web sites into their environments. Early adopters are seeing benefits from incorporating this external data, mainly by leveraging specialized integration tools that evolved for use in an Internet environment. Expect demand for outside data to increase and for this trend to continue.

Expect to extend DI with text analytics in the next three years or so.

- **Unstructured data is the new frontier for DI.** An enterprise data warehouse seeks to be a "single version of the truth" upon which most organizational decision-making is based. However, it's not the whole truth unless it represents information from the mass of unstructured data—typically in documents of mostly text, like Microsoft Office files and e-mails—that all organizations have. The catch is that a specialized technology like text analytics is required to find and translate text-based information into SQL-accessible database records that data warehouses and BI tools can use. In the next few years, expect to expand your DI solutions to include text analytic capabilities.

Defining the DI Platform

The Modules of a DI Platform

As mentioned, a few software vendors are acquiring and building multiple DI and DI-related tools, then packaging them in an integrated suite or DI platform. Before we look at the product offerings of leading DI software vendors to see how they compare in terms of comprehensive platforms, let's list the seven tools (or modules) you can expect to find in the ideal DI suite or platform:

- **Extract, transform and load (ETL).** ETL is the core engine of a data integration platform. Historically, ETL tools focused on cross-platform movement and transformation of data in a batch processing model. Recent product updates are including more features that allow for transformation and loading in a near-real-time or streaming model.
- **Enterprise data replication (EDR).** Replication is the most commonly used data integration technology today. The basic replication utilities built into most databases simply copy data one-way from one database to another in real time or batch mode with no data transformation other than type conversions. But EDR tools are more feature-rich in that they can handle bidirectional transfers across different brands of databases. EDR tools may also support advanced features like data transformation and changed data capture.

- **Data federation / enterprise information integration (EII).** Federation is a method for on-demand data access. Unlike data movement technologies, federation leaves the data in place at the sources. This makes federation appropriate for a different set of integration problems, such as providing current data for on-demand reporting or making live data from several systems appear as if it were from a single table.
- **Data profiling.** Profiling is a loose term that describes automated data analysis used to gain insight into the data being integrated. Data profiling ranges from basic features, like counting distinct values or nulls in columns, to advanced abilities, such as relating data from different sources based on the patterns and values in the fields. Most DI products provide basic profiling features in the development environment but charge for full-featured profiling.
- **Data quality.** The primary purpose of data quality tools is to standardize data elements and provide consistent verification and validation rules. The roots of most data quality tools are in name-and-address cleansing and other customer data issues. Modern tools have extended those features to address other data types (like product, location, and employee data) and provide features for generic pattern matching, dictionary and synonym lookups, and standardization to various industry formats.
- **Metadata repository.** Metadata is everywhere, so almost every data management tool includes a metadata repository and other functions for managing metadata. For example, most DI products provide basic metadata reporting services as part of the development environment. Metadata repositories sold as separate modules include features like import and versioning of metadata from separate modeling and business intelligence environments, tracing data lineage from source to the point of usage, and end user metadata reporting.

  In many tools the metadata repository also manages non-metadata entities, like development objects, project documents, and team communications (like annotations and threads associated with objects and documents). Since team members tend to collaborate through these entities, the repository enables a form of collaborative DI.

Most multi-tool technology stacks are stitched together with metadata, as are DI suites and platforms.

- **Master data management (MDM).** MDM is the practice of defining and maintaining consistent definitions of business entities (like customer, product, employee), then sharing them via data integration and application integration techniques across multiple IT systems within an enterprise and sometimes beyond to partnering companies or customers. More simply put: MDM is the practice of acquiring, improving, and sharing master data.

Other Features of a DI Platform

Aside from the modules just listed, other key elements are part of a DI platform. These are not separate modules, but components or features inherent in the platform's design.

- **Shared metadata.** Metadata support and use is at the core of these products. Even when a vendor provides a metadata product, that does not mean their own modules interoperate on a shared metadata framework; yet this is key to a unified platform. Without it you have a set of standalone modules.
- **Centralized management and administration.** As with metadata, the ideal platform should provide for centralized management of all the modules. Even though modules may be logically or physically separate, there should be centralized logging, monitoring, and control of services. It's not a platform if all the pieces have to be managed independently.
- **Scheduling and monitoring.** Every product should have the ability to monitor, start, suspend, and stop jobs, show their status, and allow an administrator to see errors. This capability should be available from a single point, yet provide a view across operating systems, sources, targets, and servers.

Vendor Comparisons

Goals of Software Vendors Relevant to DI Platforms

The DI platform is a goal for only a few vendors, and it will be years before these are complete:

DI suites are moving targets, and DI platforms haven't truly emerged. Expect them to improve.

- **DI platforms don't really exist, yet.** It's important to note that not all DI vendors currently support all of the modules and features listed here as required or desirable for a DI suite or platform. Even when these exist today, the degree of functionality of competing products differs dramatically, as does the degree of interoperability among the modules and features of a suite or platform. Hence, you should think of every DI platform as a work in progress. The list of modules and features per vendor will increase regularly, and the amount of interoperability between modules of the same platform will improve over time.
- **Not all vendors will build a DI platform.** The DI platform is today the goal of only six vendors, namely Business Objects, IBM, Informatica, Microsoft, Oracle, and SAS. Each of the other DI vendors focuses mostly on a particular type of DI product, instead of a suite of multiple products. (See the section "Niche Vendors" later in this report.)

A DI platform may be a subset of a larger integration platform or the complement of a BI platform.

- **For some vendors, the platform goal is broader than data integration.** For example, Business Objects is aiming at fully unified platforms for both BI and DI, with significant integration between the two. Likewise, SAS is pursuing a platform strategy similar to that of Business Objects by extending integration across its analytics applications. The broadest strategy is probably that of IBM's WebSphere and Information Management product lines, which are evolving toward a unified platform for data integration, application integration, and process integration.

DI Vendors by DI Platform Breadth

One way to segment the community of DI vendors is by the breadth of the suites they're producing, as well as whether they'll produce a suite or stay focused on one DI approach:

- **Horizontal.** Vendors with a horizontal strategy view data integration as a practice that underlies both analytic and transaction processing systems and forms a core IT infrastructure component. Their approach is to expand to fill all different types of data integration needs, but to stay below the applications that consume or produce data. DI platform vendors in this category include IBM, Informatica, and Microsoft.
- **Vertical.** Another perspective is that analytic processing is architecturally distinct from transaction processing, in that requirements drive very different system and application designs. Therefore, a DI platform should support these needs in the same way that EAI platforms support OLTP applications. Vendors taking a vertical approach are focused on creating a DI platform that favors the unique needs of the BI environment. This means that the DI tools must have features that extend up and down into other layers of the BI technology stack. DI platform vendors in this category include Business Objects, Oracle, and SAS.
- **Focused.** Not every vendor aspires to own your entire data integration infrastructure. Some vendors choose to be the best-of-breed product in a specific technology segment like EII. Others focus on the needs of specific vertical markets. Both approaches recognize that there will always be a need to provide specialized capabilities for some customers that large platform vendors can't provide. Some users prefer to acquire DI tools from independent vendors, thus avoiding relentless up-sell attempts from a large vendor's sales department. Representative vendors in this category include Composite Software (focused on EII), GoldenGate (focused on EDR), and ETI (focused on ETL).

DI Platform Modules per Leading Vendor

Table 1 lists the six leading vendors that provide some kind of DI suite or platform, namely Business Objects, IBM, Informatica, Microsoft, Oracle, and SAS. This report considers these vendors to be leaders because each is a large, established company with a mature ETL tool, plus other integration and data management tools that fill out a suite. And all are pursuing a strategy that will eventually unify the suite into a platform.

Six leading vendors are pursuing a DI platform strategy.

One way to segment the leading vendors is to identify the DI platform modules currently offered by each vendor. Table 1 does this, with columns for the seven DI platform tools and modules defined earlier, correlated with rows for the six leading DI platform vendors.

Technology Portfolio of Leading Data Integration Vendors

	ETL	EII	EDR	Profiling	Quality	Metadata	MDM
Business Objects	Yes	Yes (1)	No	Yes	Yes	Yes	No
IBM	Yes	Yes (1)	Yes (1)	Yes	Yes	Yes	Yes (1)
Informatica	Yes	Yes (2)	No	Yes	Yes	Yes	No
Microsoft	Yes	No	Yes (1, 3)	Yes (3)	No	No	No (4)
Oracle	Yes	Yes (3)	Yes (1, 3)	Yes	Yes (3)	Yes	Yes (1)
SAS	Yes	No	No	Yes	Yes	Yes	No

Table 1. *This table identifies the modules that the leading DI platform vendors currently offer.*[3]

Notes on Table 1:

(1) The vendor provides this module, but it is not tightly integrated with other modules in the DI product line.

(2) EII functionality is enabled by an embedded product from Composite Software.

(3) Provided but limited in functionality.

(4) Microsoft purchased master data management (MDM) vendor Stratature recently, and Microsoft will release its own version of an MDM product at a future date.

Table 1 confirms that all the leading vendors are well on their way in terms of collecting multiple tools and modules into a suite. But a suite is not a fully integrated DI platform. For example, IBM has the most comprehensive suite at the moment, and the new IBM Information Server does much to integrate diverse products with common underlying metadata and user interfaces. The EDR and MDM modules are not included in this integration but are planned in future releases. Likewise, Informatica lacks EDR at the moment but otherwise has a rich suite. Since most of Informatica's tools and modules were built in-house, interoperability and UI integration among these is often better than with other DI vendors. Business Objects has in recent years acquired Firstlogic (data quality and profiling), Medience (EII), and Inxight (BI search and text analytics). These products have diverse user interfaces, so Business Objects is working on tighter UI and metadata integration in the next product release.

Some DI platform modules are more important than others.

The list of modules supported by DI platforms is a bone of contention. Vendors still disagree about what some of the core modules should be, although most agree that covering the underlying data access and movement technologies is a requirement. Although the list of required modules presented in this report is comprehensive and accurate, some of the items are problematic. For instance, users tend to use changed data capture and EDR separately from other DI and DI-related tasks, and master data management is sometimes applied as a separate application (or as part of a packaged application). Thus, the inclusion of these in the DI platform is not as pressing as other modules.

Regardless of who has what today, DI suites and platforms are moving targets, and the list of required and desired modules will shift over time. Expect application-oriented vendors (like IBM, Oracle, and Microsoft) to pad the list with application integration and process integration modules. In early October 2007, Business Objects released new products for BI search and text analytics,

[3] Table 1 lists vendors but not product or module names, due to space limitations and the narrow scope of this report.

based on its acquisition of Inxight a few months prior. Hence, it's possible that other BI-oriented vendors will extend the list to include BI search and text analytics. Users must keep pace with their preferred software suppliers and other vendors before deciding whether a platform approach would work well in their organizations, as well as which platform would eventually fit best.

Profiles of Leading DI Vendors

Below are short profiles of the leading DI vendors. The profiles are based on hands-on evaluations of the products and extensive interviews with company executives and customers, as well as the authors' long-standing study of data integration products and practices.

Business Objects[4]

For the last five years, Business Objects has been moving beyond its core BI tools and investing heavily in data integration products. Many people are surprised to find Business Objects listed as a leading data integration vendor because the company is best known for business intelligence and still gets little recognition as a provider of enterprise data integration tools.

Business Objects positions its ETL, federation, metadata management, and data quality products under the umbrella term "enterprise information management." Over the last few years, the company has successfully moved its core integration product from small, single-project ETL to a competitive enterprise data management suite.

Business Objects views the DI platform as a vertical technology stack from the transaction sources to the consumption of data by end user reporting and analysis tools. It treats BI as a distinct class of applications that requires its own data integration tools, in the same way that enterprise applications require their own specialized integration technologies.

After a number of acquisitions in this space, Business Objects is now focused on bringing the discrete products together into a unified platform and moving the DI platform to an enterprise level from its project-specific and departmental roots.

Differentiators

- Close links between the integration and business intelligence tools provide deeper visibility and management of the entire data warehouse stack by both BI and ETL administrators. Even so, many customers use Business Objects' DI tools without Business Objects' BI platform.
- Strong data quality offerings in the customer data space, integrated with the ETL tool.
- Solid products with high usability and low cost compared to other enterprise vendors.

Primary Challenge

- Business Objects does not position products to support operational DI. It needs to rationalize the vertical focus of its products with broader data integration needs, since DI extends outside the data warehouse stack.

[4] As this report was going to press, SAP AG announced its intention to acquire Business Objects SA. Business Objects is the leading BI vendor and a leading DI vendor. If this acquisition goes through (and TDWI believes it will), SAP will suddenly be a leader in both BI and DI.

IBM

IBM is the largest of the DI providers and offers a comprehensive suite of products, although not all of these are provided within a uniform framework. IBM is far along the path to a DI platform with the announcement of IBM Information Server (IIS), which puts ETL, EII, profiling, metadata, quality, and service components together in the unified WebSphere and Information Management products.

The core components of IIS have been redesigned to operate on common metadata services and a unified user interface, both of which were problems with earlier versions of the product line. As the DI platform evolves, expect IBM to extend capabilities using data modeling and process tools from its Rational Software product lines.

The company strategy is horizontal in nature, with the goal of providing data integration to support both business intelligence and transaction processing environments. This vision runs into challenges due to overlap between DI products for the data management environment and DI products for the application development environments. Application-oriented DI is provided by Websphere Business Integration products, which are managed separately from the Information Server platform. To provide application-oriented DI services, IBM provides the WebSphere Information Services Director component within IIS. IBM is one of few vendors to attempt to bridge this gap between the discrete integration needs of both BI and transaction processing.

Differentiators

- Offers integration products in every area, and in some cases multiple products with similar functionality to meet needs in specific environments or markets. The overlap of functionality across multiple, similar products may confuse some people, but a large and diverse product portfolio of this type is necessary when a technology provider has a large and diverse customer base, as IBM does.
- Solid data movement and transformation tools with a long history in the market.
- Products extend beyond data integration to application integration and services, with the long-term goal of supporting these through the same underlying infrastructure.

Primary Challenge

- Overcoming customer confusion due to multiple DI marketing messages and multiple overlapping products that are managed by different groups within the company.

Informatica

Informatica is the largest pure-play DI vendor, although it is dwarfed by larger competitors like IBM, Microsoft, and Oracle. From the perspective of customers and revenue based on data integration, Informatica is usually considered the market leader. The company has the best-known brand in the ETL market, and has been working hard to leverage that visibility to expand the use of its products outside the data warehouse.

One of the original ETL tool providers, Informatica has extended its product line to provide data integration and management tools that cover most of the functionality needed in a DI platform. One advantage the company has as a pure-play DI vendor is an easier time staying neutral and thus working with a broad range of partners. This advantage may decrease over time as larger technology partners introduce integration tools and DI platforms of their own.

In the latest version of PowerCenter, Informatica is leading the market with advanced integration features like dealing with complex data and on-demand integration services. Competitors are following closely in this space, but Informatica still maintains a lead in its earlier adoption of some of these features and practices. For example, Informatica was the first DI vendor to emphasize on-demand and outside-the-firewall integration with business partners and service providers.

Differentiators

- Company focus is only on data integration, without motivation to cross-sell products like BI platforms or database management systems, and with no distractions due to other product lines.
- Comprehensive product portfolio, including most of the tools required of a DI platform. Built as an integrated platform, though acquisitions sometimes throw a wrench in this. Informatica is better than its competitors at unifying the UI and platform.
- Reputation for a strong service and developer network.

Primary Challenge

- Providing application-oriented and operational data integration to address a broader range of DI platform requirements.

Microsoft

SQL Server Integration Services (SSIS) was a major improvement over Data Transformation Services (DTS), which SSIS replaced in SQL Server 2005. Unlike most other integration offerings, SSIS is narrowly focused on batch-oriented ETL. Microsoft offers few of the modules TDWI Research considers part of an enterprise DI platform.

SSIS is more developer-oriented than most ETL tools, which makes sense, since Microsoft has historically focused on application developers and third parties for the bulk of its business sales. The downside of this approach is the need to know more about the Microsoft stack in order to fully leverage the tool. The upside is that, for the Microsoft-literate, SSIS provides a powerful tool able to leverage any .NET language, component, or service in the Windows environment. Due to this flexibility, it is possible to leverage SSIS for some application-oriented data integration by linking to BizTalk Server.

Microsoft's DI strategy differs from most of its competitors in that it runs only on a Windows platform and comes only bundled with SQL Server. This constrains the use of SSIS to homogeneous, Microsoft-centric customers or project-specific use in larger organizations.

Differentiators

- SSIS provides a development platform capable of taking advantage of all parts of the Microsoft technology stack. However, access to data sources outside the Microsoft stack is limited, generally requiring third-party connectors.
- Highly customizable and flexible tool.
- The basic version of SSIS is available at no extra charge with the SQL Server 2005 Standard license, and the advanced version comes with an Enterprise license.

Primary Challenge

Lack of company focus on data integration. DI is subjugated as an add-on to the database management system and not tied in to other development efforts or related products.

Oracle

Oracle is the leading data warehouse platform vendor, so adding the ETL tool Oracle Warehouse Builder (OWB) was a logical extension. OWB is surprisingly good, given that a basic version of it is available at no additional charge to customers with certain Oracle Database licenses.

Oracle sells an enterprise option of OWB that provides additional features like more extensive metadata management and support for large environments (multiproject deployment, reusability, and so on). It also sells ERP/CRM connectors and data quality as separate options for OWB. With these options, it is offering a unified set of tools for data warehouse support in an Oracle database environment.

Confusing matters is the acquisition last year of Sunopsis, a company providing a standalone ETL tool whose strength was its strong support for many different databases. The Sunopsis ETL tool has been rebranded Oracle Data Integrator as part of the Fusion middleware product line. Both OWB and Data Integrator are being delivered today, making Oracle the only vendor to sell two competing ETL product lines.[5]

Like Microsoft's SSIS, deployment options for Oracle's OWB are limited because it requires the vendor's database management system. This makes it hard to use SSIS or OWB outside a homogeneous environment. Using Fusion Middleware tools to avoid the Oracle-only limitation is challenging because of the application-development focus of this middleware and its lack of understanding of DI platform issues. Future positioning and integration of Oracle Data Integrator are expected to address this problem.

[5] Oracle Corporation recently offered a bid to acquire BEA Systems, a leading provider of tools for enterprise application integration (EAI) and other middleware approaches. If this acquisition goes through, it will significantly change the mix of products in the Oracle Fusion product line, which may affect the future of Oracle Data Integrator.

Differentiators

- OWB takes advantage of native database capabilities in a highly integrated fashion, providing features other vendors don't.
- Best-in-class features support dimensional models.
- OWB is bundled at no additional charge with every Oracle database.

Primary Challenge

- OWB is limited to running atop Oracle databases, with limited data access outside Oracle environments. ODI is supposed to address this issue; however, it does not have the profiling or quality modules of OWB.

SAS

SAS is the largest privately held software company in the world, with $1.9 billion in revenues in 2006, and is known primarily for analytics and BI. SAS customers have been using the SAS programming language for years as a custom ETL coding tool. The SAS language is designed to manipulate data, so it has been a logical choice for extracting data in SAS environments.

SAS entered the tool-based ETL market with Enterprise Data Integration Server. The product is built on top of the company's existing technology, which allows SAS programmers to leverage their expertise. At the same time, there are still ties into the underlying environment, so customers must buy base SAS products in addition to the ETL product and have more to learn. This means SAS will appeal most to customers who are already using SAS products.

SAS's view of a DI platform is similar to that of Business Objects, with a focus on a DI platform that is primarily aimed at the data management and BI market rather than a horizontal offering. SAS covers the data management modules of a DI platform with offerings for data quality, data profiling, and metadata. Its gap is in delivering basic data access technologies like federation and replication. Based on the DI vision, expect continued investment by SAS in building out an enterprise-level DI platform offering.

Differentiators

- Powerful data processing language underlying SAS products allows for complete customization and extension of the tools.
- Fully integrated with all other SAS components, allowing full vertical use of all products within the DI context.
- Long experience in the BI and analytics market where most DI tools are used.

Primary Challenge

- Moving from a tightly integrated set of base technology and analytics products to a broader, tools-focused DI market.

Niche Vendors

There are many niche segments within the DI market, and the seven modules of the ideal DI platform represent the leading niches. Besides the leading DI platform vendors listed earlier in this report, a number of software vendors provide DI and DI-related tools outside the context of a DI platform (though sometimes in the context of some kind of smaller suite). The following list contains representative vendors and products for these and other niches.[6]

- **Extract, transform, and load (ETL).** Representative products include Co>Operating System from Ab Initio, ETI Solution from ETI, DataMigrator from iWay Software (a subsidiary of Information Builders), Data Integrator from Pervasive Software (acquired from Data Junction), and Sybase ETL from Sybase, Inc. (acquired from Solande).
- **Open source ETL.** Of special note are the relatively new ETL tools based on open source, including Apatar Community Edition from Apatar, Inc., Pentaho Data Integration from Pentaho Corporation, and Talend Integration Suite from Talend SA.

EII and data quality vendors are targets of acquisitions as large vendors build their suites.

- **Enterprise information integration (EII).** Data federation or EII was a short-lived market segment, represented by a dozen or so software vendors early this decade. Since then, most of these vendors have been acquired or have gone out of business. Today, EII is usually found embedded in platforms for DI, BI, or database management, as we saw earlier in the discussion of DI platforms. Even so, a handful of standalone EII tools linger, including Composite Information Server from Composite Software (which is also embedded in platforms by Cognos and Informatica), Federation Server from IBM, Ipedo XIP from Ipedo, and iWay Data Hub from iWay Software (a subsidiary of Information Builders).
- **Enterprise data replication (EDR).** Basic, table-based replication is built into almost all brands of relational database management systems. More advanced replication tools—suited to heterogeneous enterprise use—include Transformation Server from DataMirror Corporation (currently being acquired by IBM), Transactional Data Management Solutions from GoldenGate Software, and Replication Server from Sybase.
- **Data quality.** Many of the leading data quality tools have been acquired this decade, and are now part of larger DI suites and platforms (though usually still sold standalone, too). For example, Business Objects acquired Firstlogic, IBM got QualityStage from the Ascential acquisition, Informatica acquired Similarity Systems, and SAS acquired DataFlux. Even so, the data quality vendor community is still rather large, and it includes well-known data quality tools like i/Lytics Data Quality from Innovative Systems, Customer Data Quality and other tools from Group 1 (a subsidiary of Pitney Bowes), and Trillium Software System from Trillium Software (a subsidiary of Harte-Hanks).[7]

[6] The vendors and products mentioned here are representative, and the list is not intended to be comprehensive.

[7] For a complete survey of data quality vendors and tools, see the TDWI Technology Market Report *Enterprise Data Quality Tools* (Q2 2006), available to TDWI Members online at www.tdwi.org/research.

- **Metadata management.** Some kind of metadata repository and functions for metadata management are built into almost every tool for ETL, EII, data quality, reporting, data modeling, and so on. In addition to these, Meta Integration Model Bridge from Meta Integration Technology, Inc. (MITI) is a tool for exchanging and integrating metadata among various tools, platforms, and applications. Standalone metadata repositories for multitool enterprise use include ASG-Rochade from Allen Systems Group and Advantage Repository from Computer Associates.
- **Master data management (MDM).** The majority of MDM solutions are home-grown using a variety of tools (for ETL, data quality, data modeling, etc.) and systems (databases, metadata repositories, packaged applications). Among these, ETL is the preferred tool, so MDM has become a module of ETL-based DI platforms. Despite the preponderance of home-grown solutions, a variety of packaged applications for MDM are available, including NetWeaver MDM from SAP, LensBuilder from Silver Creek, Siperian Hub from Siperian, and WebSphere Product Center and WebSphere Customer Center from IBM.[8]
- **Sort.** Many DI solutions sort data repeatedly as they process it, so it makes sense to use a standalone, high-performance sort tool to speed up DI processing. The two leading tools in this niche are CoSort from Innovative Routines International and DMExpress from Syncsort. Though both vendors started with a focus on high-speed sort enabled by parallel processing, both have evolved to offer capabilities for DI and data quality, too.
- **Text analytics.** User organizations are just now starting to use a technology called text analytics, which (among other things) finds entities and facts in documents of mostly natural language text (e.g., word processing files and e-mail) and converts these facts to structured database records that a data warehouse or BI tool can understand. Eventually, text analytics may become a required module for DI platforms. Representative tools include Text Analytics Suite from Attensity, Content Mining Platform from Clarabridge, and Text Analytics Suite from ClearForest (recently acquired by Reuters).[9]

Think of text analytics as ETL for text. Hence, it's for DI, not content management.

Recommendations

Technical users responsible for selecting and maintaining tools for data integration and related data management practices should keep the following points in mind:

- **Recognize that leading DI vendors now offer suites of tools.** These vendors still sell most of their DI suite modules individually, so you can follow a best-of-breed strategy by acquiring tools from multiple vendors—evaluating tools singly, as you need them. However, if you choose to follow a single-vendor strategy, you'll need to evaluate entire suites to ensure that the suite you choose has the modules and features you will need both today and in the future.

Like it or not, you're now shopping for DI suites, not DI tools.

[8] For a complete survey of tools for MDM, see the TDWI Technology Market Report *Segmenting Master Data Management Solutions* (Q4 2006), available to TDWI Members online at www.tdwi.org/research.

[9] For more information about text mining, see the TDWI Best Practices Report *BI Search and Text Analytics* (Q2 2007), available online at www.tdwi.org/research.

- **Watch the DI tool suites to see when and if they evolve into unified platforms.** Again, a suite is not a platform. Both have multiple DI and DI-related tools or modules, but the platform has far deeper interoperability across tool servers, as well as a seamless integration of tool user interfaces. As you evaluate suites, ask the vendor for a "road map" detailing the eventual integration of the suite into a platform. Also, once you're committed to a suite, press the vendor to give priority to interoperability features that you need in the future platform.
- **Expect more market consolidation.** As vendors fill out their suites, they will continue to acquire niche vendors. The good news is that most tools improve under new ownership, thanks to the greater resources of a larger vendor. The bad news is that, in the short term, releases are delayed and FUD abounds as the new owner redefines the tool's direction.

DI is more than ETL, so evaluate tools accordingly.

- **Remember that data integration is multiple practices.** This is why DI suites have multiple tools. It's not just ETL, but also EII and EDR, as well as related practices like data profiling and data quality. Expect to use more of these together and with greater interoperability.
- **Expect ETL to retain its hegemony as the preferred DI method for BI.** After all, BI and data warehousing require complex transformations of large datasets, which ETL handles better than other approaches do.
- **Expand your concept of DI beyond ETL.** Other approaches have advantages. For instance, data quality addresses many of the problems and opportunities that DI exposes, and EII and replication provide real-time operation for small amounts of data that complements the massive but latent datasets of ETL.
- **Use the vendor and tool information in this report.** The profiles of leading vendors, descriptions of niche tools, and modules of DI suites and platforms explained in this report should help you create an evaluation list that maps realistically to your requirements.

TDWI RESEARCH

TDWI Research provides research and advice for BI professionals worldwide. TDWI Research focuses exclusively on BI/DW issues and teams up with industry practitioners to deliver both broad and deep understanding of the business and technical issues surrounding the deployment of business intelligence and data warehousing solutions. TDWI Research offers reports, commentary, and inquiry services via a worldwide Membership program and provides custom research, benchmarking, and strategic planning services to user and vendor organizations.

tdwi
THE DATA WAREHOUSING INSTITUTE

1201 Monster Road SW
Suite 250
Renton, WA 98057

T 425.277.9126
F 425.687.2842
E info@tdwi.org

www.tdwi.org