White Paper

Data Integration:
Creating a Trustworthy Data Foundation for Business Intelligence

Business Objects™

Author: MAS Strategies

Contributors: Darren Cunningham, MaryLouise Meckler, Jennifer Meegan, David Nguyen, Philip On

Audience: Report developers, data warehouse managers, IT director, CIO

Contents

Executive Summary

To make sound decisions and comply with governmental reporting requirements, an organization must first establish a solid data foundation. This foundation must combine historical data with current values from operational systems in order to provide a single version of the truth that can be then used to identify trends and predict future outcomes. Data integration technology is the key to consolidating this data and delivering an information infrastructure that will meet strategic business intelligence (BI) initiatives and tactical and governmental reporting requirements. Data integration is the enabling technology for providing trustworthy information, enhancing IT and end-user productivity, and helping organizations achieve and maintain a competitive edge. Data integration enables mid-size and large organizations to effectively and efficiently leverage their data resources in order to satisfy their analysis and reporting requirements.

While a homegrown data integration effort frequently yields a quick and dirty solution that may initially appear inexpensive, any upfront savings are often soon lost as demands on resources and personnel change. Vendor-supported packaged solutions, on the other hand, have withstood the test of time. Since they include capabilities such as metadata integration, ongoing updates and maintenance, access to a wider variety of data sources and types, and design and debugging options rarely offered by in-house solutions, they serve to increase the productivity of the IT organization. This is an important advantage as few organizations have unlimited resources and most are under constant pressure to do more with less. Additionally, most homegrown data integration solutions are almost never integrated with an organization's BI tools. Such integration is, however, available with commercial offerings either by adherence to industry standards and/or through integration with the BI tools in the data integration vendor's total product portfolio.

This paper will discuss the importance of data integration and help you identify the key challenges of integrating data. It will also provide you with an overview of data warehousing and its variations, as well as summarize the benefits and approaches to integrating data.

Introduction

Imagine you work with one of your organization's mission-critical operational systems. Your organization considers you the go-to person for any query or reporting request associated with this system.

Can you reference historical values?

Can you comply with the reporting requirements of Sarbanes-Oxley?

Is important data trapped inside proprietary applications?

Can you combine data from several departmental systems?

What if your CEO were to ask you to modify one of your year-end reports to compare this year's numbers to those from the two previous years? Unfortunately, the operational system you are reporting off of only stores current-year detail records and prior-year summary balances. The summary balances from two years ago were purged from the system at the beginning of this year. When you try to explain this to the CEO, all you hear is, "So how long will it take to find the lost data?"

Your medium-sized company is publicly held and under the Sarbanes-Oxley act your CEO and CFO must certify the accuracy of its financial statements based on data from several internal systems and spreadsheets. In a recent conversation, the CFO asked you to confirm the roll-ups in these reports were trustworthy; that they were timely, auditable, and not based on "adding apples to oranges."

In addition to the requests from the CEO and CFO, you receive another request to produce a report from a commercial enterprise application software package your company has recently implemented. Unlike the production system you've worked with for years, this system contains seemingly strange files that appear to contain both system and user data. As you start to learn more about these files and try to decipher the meaning of apparently incomprehensible acronyms, you wonder just how you will access the necessary data to solve this request.

The chief marketing officer asks you to identify the 100 customers who produced the most revenue for your company last year. These companies would be placed on a "preferred customer list" and their requests given special handling and top priority. As your company's sales and service departments each tracked customer revenues in their own departmental systems, you had to first match customers and add these revenues together. When you proudly present the list to the CMO, he gives you a strange look and asks why a company he expected to rank among the top 25 was not even on the list.

Are these scenarios familiar to you?

Why Is Data Integration Important?

To be successful, your organization—large or small—must run its operations effectively and efficiently, which requires the ability to analyze operational performance. If you can't see how you're performing, how do you know you are making the right business decisions? For an organization to thrive, or perhaps even survive, operations and analysis must work together and reinforce each other. This is especially important in small-to-medium size organizations, which—in order to grow and expand—need to focus their limited resources and take appropriate actions to build upon their successes while quickly identifying and resolving operational problems.

To draw valid conclusions, an organization needs to be able to analyze both current and historical data from multiple disparate sources. With a bit of luck, the organization can consolidate the data from these disparate sources without resorting to "desperate measures."

Without the entire picture, it's difficult to make sound and dependable business decisions. That's because good decision-making requires a complete and accurate view of data. The ability to access and integrate all of your data sources is the start to getting the complete picture—and the key to not compromising your decision-making process.

Though your organization needs a complete view of operations, the data you need often resides in a variety of application systems that do not necessarily all use the same database management system. Furthermore, these application systems may only contain current data values. They may not store prior data values needed to provide historical context and to discover trends.

Data integration allows an organization to consolidate the current data contained in its many operational or production systems and combine it with historical values. And the creation of a data warehouse (or, on a more limited scale, a single-subject data mart) facilitates access to this data. Collecting and consolidating the data needed to populate a data warehouse or data mart and periodically augmenting its content with new values while retaining the old is a practical application of data integration.

Approaches for Data Integration

An organization can integrate its data through a variety of methods including:

- Enterprise-class data warehouse implementations hosting massive amounts of historical data
- Departmental or subject-level data marts focused on a single organizational unit or functional area
- Operational data stores containing current values of data extracted from several operational systems
- Enterprise information integration (EII) deployments that provide a direct, real-time view of data residing in multiple operational systems

Hybrid approaches are common and include, for example, departmental data marts populated from enterprise data warehouses or an EII deployment that access a data warehouse for historical data and operational systems for the latest values. Most organizations use a combination of methods as part of their overall information architecture. Whatever the form (see Appendix for additional details), the intent is to create a data platform for analytical purposes. By consolidating, standardizing, and, in many cases, summarizing the data contained in multiple operational systems, an organization can analyze the combined data to achieve a "single and trustworthy version of the truth."

By integrating data, organizations can more effectively use this data for analytical purposes.

Data Warehouse

There are a multitude of benefits resulting from integrating operational data within a data warehouse or data mart. You can build these to:

- ►Integrate recent and historical data values
- ►Combine data from disparate sources
- ►Create a data foundation for analytical purposes
- ►Improve data quality
- ►Establish consistency throughout the organization
- ►Facilitate the adoption of corporate data standards without having to modify existing operational systems
- ►Provide historical breadth and enable trend analysis

Operational Data Store

Many organizations have created operational data stores to consolidate current data from multiple operational systems. While a data warehouse typically collects historical values, an operational data store focuses on current values. You can build an operational data store to:

- ►Obtain a complete view of your customer
- ►Integrate current financial data for government reporting and compliance purposes
- ►Consolidate current information from multiple sources

Enterprise Information Integration (EII)

EII allows real-time access to data in multiple systems making it appear as if it came from a single system. While some vendors have let marketing get ahead of reality by claiming that EII eliminates the need to build a data warehouse, this is not usually the case. EII complements a data warehouse and should be considered one component of an organization's overall enterprise information architecture.

Consider, for example, the analogy of someone with two checking accounts. An operational data store can be used to determine the total current balance while a data warehouse can be used to track a given expenditure over the last several years. An EII solution would allow you to do both, assuming the data warehouse and the operational data store were both being accessed. In the business world, EII can be used to simultaneously query multiple inventory sites to see if there is sufficient stock on hand to immediately satisfy an incoming order. It could also be used to identify products with excess inventory and be linked to a system that would send email offers, with special price incentives, to targeted prospects to encourage additional purchases of these items.

You can use enterprise information integration to:

- Provide an integrated view across all sources—production systems, operational data stores, data warehouses, and data marts
- Obtain a real-time view of data spread across federated (perhaps one at each manufacturing location) operational systems
- Enable operational, or real-time, business intelligence by accessing historical values in data warehouses or data marts and the real-time values in operational systems
- Jump-start data integration efforts by first deploying an EII solution, perhaps to quickly satisfy an important user requirement, and then deciding if the data should ultimately be extracted to a data warehouse, data mart, or operational data store

WARNING SIGNS: Does Your Organization Suffer from Poor Data Integration?

The following situations could benefit from trustworthy data integration.

- **No single version of the truth.** Managers are arguing about the fact that analyses results differ—even though the data came from the same operational system.
- **Inability to comply with governmental reporting requirements.** The CEO and CFO are uncomfortable signing off on the company's financial statements because there is no way to trace the numbers back to its original source. The Sarbanes-Oxley Act requires isolated financial data be integrated and that the CEO and CFO certify, subject to penalties that include imprisonment, the accuracy of their company's financial statements.
- **Incomplete data foundation.** Presentations that include an analysis prefaced by a statement such as, "…except for the data that we were unable to obtain from…" Or worse, a presentation that begins with, "Due to the discovery of data not included in last period's analysis, we are reversing our decision…"
- **Poor audit trail and data lineage.** An analyst alerts management to a potential problem discovered while running a query against the data in an operational system. The analyst cannot, however, answer the follow-up question, "How long has this problem existed?"
- **Inability to consolidate data from multiple sources.** As a result of an out-of-stock condition for a critical part, an organization must expedite an order and purchase the item at a premium price. Once the order arrives, the organization discovers another division had an excess quantity of the same part and was trying to sell it at a discount to balance its inventory.
- **Poorly integrated, stovepipe operational systems.** While analysts use a variety of business intelligence tools to generate reports from application systems, they re-enter relevant summary values into a spreadsheet for any analyses requiring data from more than one application.
- **Lack of common data definitions.** With a series of very convincing charts and graphs, an executive presents what appears to be a thorough analysis of the cause of a particular problem. However, while the format of the presentation qualifies as a work of art, the executive's credibility suffers greatly when someone says, "That's not what that data means, where in the world did you get that?"
- **Historical values not retained in a data warehouse or data mart.** An analyst runs the same report each week against an application system. However, in order to see period-to-period comparisons, the analyst maintains a spreadsheet. Each week, he must manually add a new column and enter that week's report values.

- **Lack of an integrated 360° view.** The CEO of one of your largest customers has called your company's CEO to complain that when his people contact your call center for support, they are not receiving the attention he believes they deserve.
- **High cost of maintaining in-house "one-time" code.** Your company is in the process of developing a new order entry system that, when deployed, promises to provide a significant advantage over your competitors. Things are going smoothly until, six months into the project, the lead programmer is called away to "patch some extract code" that no longer seems to work with the latest version of the ERP system from which the data is sourced. Because she last modified her code three years earlier and she has not kept up with the new version of the ERP system, this task takes much longer than anyone anticipated and the deployment of the new order entry system is now behind schedule.

The Benefits of Data Integration

An organization can reap many benefits from data integration. These include the ability to:

Provide a Single View of the Organization

Historically, operational systems, especially legacy ones, were created to solve a particular set of needs, each evolving into an independent island of information. These needs included order entry, shipping and receiving, payroll, manufacturing and inventory control, and customer support. In the case where a company deploys the same operational system in several locations, each location might still have its own database sometimes with differing value lists and even data definitions—a separate isle of information. The islands and isles must be consolidated in order to obtain a single view of the entire organization.

By integrating data across disparate operational systems, an organization can increase the effectiveness of its data access and analysis capabilities.

Deliver Trusted Information

While the quality of any decision is highly dependent on the quality of data upon which it was based, governmental compliance regulations and associated reporting requirements makes the need for trustworthy data even more essential.

For data to be trustworthy it must be of the highest quality. Data profiling and data quality tools can be used to ensure this. Data profiling can be used to identify problems and anomalies in the source data as, for example, telephone or social security numbers that don't match their expected format or pattern, new orders with requested delivery dates in the prior century, the number of unique values in a field and a count of suspicious values such as "99999" or blank, and gender code fields with eight different values. It can also be used to examine inter-record dependencies such as sales orders for products not on the product master file. Data profiling and data quality tools complement each other; once data profiling identifies an issue, data quality tools can then be used to facilitate its resolution. Data quality tools can be used to eliminate duplications, verify and correct addresses, standardize data values by substituting corporate standards for departmental variations as exemplified in the previously mentioned region code example, and even augment records with additional data such as geocodes, credit ratings, or census and other demographic information.

In order to be trustworthy, the lineage of the data (i.e., where it originated and how it was transformed) must also be known and auditable. It is also important to be able to perform impact analysis to see what reports and processes are dependent on a given data element.When ultimately standardized, the long-term benefits of having a common set of business rules and common set of definitions and terms can greatly improve efficiency and effectiveness.

Analyze Current Values and Trends

In an operational environment, organizations deploy query and reporting tools along with production reports to determine current status. They generally summarize the data and only maintain historical data values for a limited time, if at all. Though an operational system provides

the most current values, these values may not be appropriate for tracking and analyzing how something has changed over time. A data warehouse or data mart is usually needed if access to historical values is required.

In a production environment, values are constantly changing, as most transactions typically update one or more data values. There is nothing more frustrating than performing an analysis against an operational system only to find ten minutes later, as a result of new transactions having hit the system, you now get different results. You can avoid this problem by capturing a data snapshot and storing it in a data warehouse. While the values stored may not be up-to-the second, they are usually collected at well-defined, cut-off cycles (e.g., monthly, weekly, daily, hourly). And this ensures the validity of period-to-period comparisons.

A single observation does not a trend make! Just as it takes two points to determine a straight line, it takes a series of values collected over a period of time to determine a trend.

Treat Data as a Corporate Asset

While many organizations speak of their data as a corporate asset, its quantity, unlike other assets, is not necessarily limited. Data is the one asset that can grow and reproduce almost without limit, while often mutating in the process. As it happens, despite the widespread use of data flow diagrams, data doesn't really flow from one system to another. Rather, a copy (perhaps somewhat transformed and/or summarized) is sent to the second system while the first system still retains the data. This can lead to inconsistent data in each system and as a result, inconsistent decision making.

Using data integration to consolidate the data from the various operational systems serves to create a "single version of truth" so you can treat data as the enormous asset it is. To do this effectively, the lineage of the data, including its origin and/or derivation, must be readily available and not lost.

Get a Complete or 360° View of Your Business

To obtain a complete view of your organization it may be necessary to consolidate data from several individual units. Consider multiple divisions of an organization. Each division has its own purchasing system and wants to maximize the discount it receives from its vendors. Although each of the separate divisions most likely has the purchasing details it needs to negotiate a discount with each vendor, the organization as a whole could likely negotiate better discounts if it were able to aggregate the total amount it purchased from each vendor across all of its divisions. For example, consider a three-division company. If each division places a $1 million purchase order each year with the same vendor, it could instead base its discount negotiations on the fact that the company as a whole spends $3 million a year with the vendor.

Integrating data from multiple systems creates an environment where the whole is worth far more than the sum of the individual parts.

Because it can now determine the total amount it purchases from the vendor, it would likely receive a higher percentage discount than each division would have received by negotiating independently.

Discover and Reconcile Differing Data Definitions and Business Rules

Every department understands its own data. It's the data from other departments that always seems to be wrong and in need of reconciling to fit the individual department's needs.

A great benefit of any data integration effort is the discovery that different parts of the same organization do not necessarily speak a common language or use the same business processes. When ultimately standardized, the long-term benefits of having a common set of business rules and common set of definitions and terms can greatly improve efficiency and effectiveness.

For example, when determining departmental productivity using "cost per employee" as a metric, do two part-time employees, each working a four-hour day, count as one employee or two? The answer is likely to differ by department and unless an organization-wide definition is established, departmental comparisons are not meaningful.

Once an organization recognizes differing definitions and standardizes on enterprise definitions, data integration can facilitate their implementation. It may be impractical to modify every operational system to reflect the enterprise standard. However, it is possible to transform the data extracted from each operational system to conform to the enterprise standard definitions and value lists as the data is loaded into the warehouse or, with an EII approach, in the process of accessing it.

Metadata matters! Minimize communication errors by ensuring every department speaks the same business language and uses the same data definitions.

Data definitions are an example of metadata. Metadata is nothing more than "data about data" and data definitions are but one example. Other examples include standardized field names and column headings, computations for derived data (e.g., profit equals revenue less expenses), data element value lists (e.g., NY is the code for New York; allowable values for region code are N for North, S for South, E for East, W for West, and M for Mid-West), etc. In addition, to allowing the entire organization to speak the same language and understand the algorithm behind a computed value, it provides a valuable audit trail for compliance purposes, especially if the data lineage or data source and associated transformations are also captured.

Take Incremental Steps Rather than Attempting to Do Everything at Once

The planning and implementation of a full-scale enterprise data warehouse does not occur overnight. While the end results will almost certainly justify the effort, organizations can develop an effective information architecture by taking small, incremental steps. For example, a sales data mart can be used to track and analyze customer purchases and provide valuable insights for spotting trends and recognizing cross-selling or up-selling opportunities. Enterprise information

integration (EII) can be used to access and combine real-time data residing in multiple operational systems; once the data warehouse is deployed, EII can be used to access it as well and thus, provide a historical perspective. Some organizations have used EII to provide a quick view of operational data in order to decide if a more formal effort should then be undertaken to add this data to an existing data warehouse.

Don't try to boil the ocean. A phased, incremental approach to an overall enterprise information management architecture can begin with a data mart or an enterprise information integration (EII) solution.

Create and Maintain Organization-Wide Reference Files

All organizations have data used across the several departments. Examples of these "reference data" files include customer data, product data, employee data, vendor data, and even financial data such as the company's chart-of-accounts. In many organizations. individual departments maintain their own reference files and problems frequently arise when different departments use different identifiers or keys for the same customer, making it difficult, if not impossible, to accurately combine. For example, if a customer's revenues from both the sales and the service departments can't be accurately combined, the total value of that customer's account would be understated.

Every organization has data, such as customer and product files, that are used across the organization. These reference files facilitate the organization's ability to create a "360 degree view" of the subject they reference.

While the term "Master Data Management" is receiving a tremendous amount of attention, it is simply an extension of the reference file concept, a concept behind the use of centralized Rolodex files even before the common business use of computers. Data integration technology, combined with data quality software, is the underlying technology for creating organization-wide reference files and master data management solutions.

Reference files are a subset of metadata management; for example the definition and allowable values of the data elements collected for each customer or product are examples of metadata.

Maintain the Response and Performance of Operational Systems

The days of having to "submit queries and run reports against the production databases only between noon and 1 pm or after 6 pm" are hopefully long past. Yet running queries or reports against the database used by an online application can still negatively impact the performance and user response time of that application. Performance counts! If an analysis request negatively impacts the response of an operational system, the analysis request will be deferred, perhaps permanently! With a data warehouse or data mart, you offload the query to an environment where the period can be optimized for this purpose.

- ►Deliver a complete view of a customer
- ►Offload the processing burden on operational systems
- ►Standardize business processes and data definitions
- ►Combine current and past values from disparate sources in order to see the big picture

Approaches to Implementing a Data Integration Solution

Once you recognize the benefits of and need for data integration, you have to determine how best to move forward. The two basic approaches to data integration are:

1. Develop and build your own in-house solution
2. Acquire a commercial offering

You should carefully consider the pros and cons of each.

In-House Development

Organizations that develop their own data integration solutions frequently do so in a somewhat piecemeal fashion, without any overall data integration strategy. They generally assign an analysis request that requires access to data from multiple sources to the IT department. A programmer then writes the code necessary to access and integrate all of the data.

If the programmer is fortunate, the source systems are well documented, the content of the data fields conform to the documentation, and each of the individual systems use the same value lists and code sets to represent the individual values of common data elements. If this is not the case, the programmer's task quickly expands to include data value transformations. This frequently causes the schedule to slip, especially if the data mappings are not simple one-to-one transformations.

Satisfying the initial consolidation requirement is only the beginning of the overall integration effort. As any experienced programmer knows, the initial coding effort is followed by ongoing support and maintenance especially if a new analysis request requires additional data fields or the file structure of the source systems changes. One of the givens in any applications environment is the ongoing need to respond to change; another is that "quick and dirty" one-time coding efforts frequently evolve into scheduled production jobs.

Moreover, a series of uncoordinated, individual integration tasks, even if each one were successfully accomplished, ultimately result in an assortment of uncoordinated (and usually undocumented) solutions that collectively, quickly become unmanageable. The problem is further compounded if a different programmer is responsible for each individual data integration solution—as most programmers have their own individual programming idiosyncrasies and may have even used different programming languages

Programmer turnover is another factor to consider. While programming the initial extract program may involve creativity, future maintenance of these programs is often a thankless task. In general, programmers prefer new challenges and the original authors of the extract program may no longer be available to maintain them. And even if they are, they may not go out of their way to mention their initial involvement in the creation of the extract programs.

Purchasing Commercial Data Integration Solutions

While an initial data integration request can lead to an initial decision to develop the code in-house, a commercial data integration solution, due to the need to appeal to a broad audience, provides a wide range of capabilities generally not incorporated into a homegrown solution. These include:

- ►Support for a variety of data types, sources, and targets
- ►Integration with many commercial application software packages
- ►An extensive library of data transformation functions
- ►Data quality functionality
- ►Metadata integration with the other tools
- ►Data lineage tracking and impact analysis
- ►Documentation and audit trails
- ►Ability to satisfy both current and future requirements
- ►A variety of packaging options and price points

Support for a Wide Variety of Data Types, Sources, and Targets

With the possible exception of data integration tools available from some database vendors, most commercial data integration software populates a wide variety of target databases. Some database vendors have limited the scope of their data integration software to only populate, or work best with, their own databases. Any organization contemplating using a database vendor's data integration offerings should recognize the potential for platform lock-in. Data integration offerings from non-database vendors usually offer more flexibility and will likely not constrain the future choice of databases and/or operating systems. Many are also designed to work, out-of-the-box, with a wide variety of data types, not just those that are SQL-based. In addition to relational structures, these include mainframe legacy data structures, XML data structures, and message queuing systems.

Integration with Commercial Application Software Packages

A commercial data integration solution that can work directly with third-party packaged software applications minimizes, or even avoids, many of the problems associated with continually modifying and retesting homegrown integration programs. This retesting of an in-house developed solution is required whenever there are changes to the packaged application software. Commercial data integration solutions usually do this as part of their normal maintenance. Even if your organization is currently using homegrown applications software, it is likely to use enterprise application software sometime in the future as it grows and expands. A good commercial data integration software offering should be able to integrate data from these packaged applications and facilitate the population a data warehouse or data mart. Some data integration vendors also offer easy-to-deploy yet highly customizable data marts, designed to quickly integrate with a wide variety of enterprise application software packages.

An Extensive Library of Data Transformation Functions

By including a library of pre-packaged, but extensible, data transformation functions, leading commercial data integration products are capable of performing data transformations and aggregations. This minimizes the need for custom coding and code maintenance. A robust offering will have interactive debugging facilities that allow the data integration staff to monitor the data flowing through each transformation, establish conditional breakpoints, and view and profile live data flows.

Data Quality Functionality

As the quality of the data is a major factor in a successful data integration implementation, it needs to be part of the overall solution. In some cases the data integration vendor will OEM and support best-of-breed, third-party data cleansing software as a part of its offering, making the fact that it was developed by another vendor relatively transparent to the deploying organization.

Metadata Integration with Other Tools

Most packaged data integration solutions are also designed to leverage and integrate metadata. This is accomplished by conforming to standards such as the Object Management Group Common Data Warehouse Metamodel (OMG – CWM) thus allowing the data integration software's metadata repository to exchange metadata with other CWM-compliant metadata repositories used by third-party design and business intelligence tools. When a single vendor supplies a range of data integration and business intelligence tools, it can easily share metadata across all of its products and greatly facilitate the overall ease of integrating the individual components.

Data Lineage Tracking and Impact Analysis

Impact analysis, or the ability to determine how a change to a source system data field can affect a business intelligence report or analysis, is only possible through metadata integration and the resultant ability to track end-to-end data lineage. Data lineage is especially important when a target field is derived from multiple source system fields. A good commercial data integration solution facilitates change data management by providing strong impact analysis capabilities including "what-if" developer scenarios.

Documentation and Audit Trails

An often-overlooked benefit of a commercial data integration solution is it serves to document the underlying data transformation processes and data lineage. This is far from a minor consideration, as many organizations initially attempting to deploy in-house developed code have unfortunately discovered. When these organizations later tried to ascertain how data sourced from legacy systems was transformed prior to being integrated with other data, they discovered the documentation, assuming it even existed, was woefully inadequate, the programmer who wrote the original programs was no longer employed by the company, and the production version of source code could not be found. The ability to document the transformation processes and the data lineage is much more than a technical issue, it is an essential element of any audit trail and is required to ensure that any reported results are trustworthy. And trustworthiness is an absolute requirement for complying with governmental reporting requirements.

Ability to Satisfy Future Needs

A commercial data integration solution needs to meet both current and future performance requirements. This may be accomplished through a variety of mechanisms such as parallel processing technology and workload balancing. Change data capture techniques can track and extract only those changes that have occurred to relevant fields in the source data files since the prior extract.

A Variety of Packaging Options and Price Points

Ideally, a commercial data integration solution will have several editions, with varying price points dependent on functionality (e.g., varying by number of servers, number of supported databases) to address the economic constraints of departmental- and enterprise-scale deployments. If multiple editions are available, they should be upwardly compatible with each other.

Data Integration Build Versus Buy—the Bottom Line

As a general rule, unless the data integration task is truly a "one-time" effort, organizations should strongly consider a packaged data integration solution. The short-term initial costs associated with an in-house programming effort are likely to be less than the acquisition cost of a packaged product. But on-going maintenance costs and the indirect costs associated with an inability to respond quickly to change will just as quickly consume the initial cost savings.

In almost all situations, a packaged solution is significantly less expensive in the intermediate—and long-term.

In addition, the productivity resulting from the ability of most commercial data integration packages to integrate and share metadata with other data warehouse tools is something most in-house solutions simply do not consider or provide. Of particular importance is the ability to share metadata with modeling and design tools and the business intelligence tools that will access the data warehouse. Commercial data integration packages are also likely to be integrated with or include data quality and data profiling technology—functionality frequently overlooked by homegrown, in-house development efforts.

When considering data integration tools offered by a database vendor, it's important to recognize one of the major strengths of a database vendor's own data integration product can also be one of its major weaknesses. That is, a vendor often optimizes its solution for populating its own database. In fact, some offerings, with the possible exception of also generating flat files, can only populate a vendor's own database.

As organizations grow, their data integration needs tend to multiply and a commercial data integration solution is usually acquired. Organizations anticipating this should consider deploying a commercial data integration solution early on. While it may be tempting to try and solve each data integration challenge with an in-house band-aid approach, the deployment of a commercial data integration solution is an investment that will yield both immediate and future benefits for both IT and the user communities.

Build Versus Buy Decision Criteria

Data Integration Considerations	Build	Buy
Initial Start-up Cost	Lower	Higher
Continuing Operational Cost	Higher	Lower
Ongoing Support and Maintenance	In-house responsibility	Vendor responsibility
One-time "quick and dirty" task	Consider	May be overkill unless "one-time" task becomes ongoing request
IT staff requirements	Higher	Lower
IT productivity	Detracts from	Contributes to
Data sources / data targets	Single / single	Multiple/multiple, Multiple/single, Single/multiple
Data Sources include third-party enterprise application software	Changes to data source will require IT effort	Changes to data source likely handled by data integration vendor
Complex transformations	Limited; IT must create custom code	Comprehensive; customized user code can usually be added to vendor-supplied library
Impact Analysis and data Lineage	Limited	Usually included
Auditable, documented transformation process	Unlikely	Yes
Data quality functionality	Frequently overlooked or relatively limited	Usually included, sometimes via OEM partnerships
Metadata sharing with third-party tools	Frequently overlooked	Likely to conform to industry standards
Integration with vendor's other products	Not applicable	By design; also provides single source of support

Conclusion

Reliable data is the basis for sound decision making. And data integration is also the key to delivering trusted information —do users of business intelligence tools feel they are basing their decisions on trustworthy data? The best tools are of little value if the data they analyze is not complete, accurate, and trustworthy.

Operational and analytical systems complement each other. Organizations must effectively deploy both in order to succeed. For analytic purposes such as trend analysis and forecasting, it's necessary to collect time-stamped data values from multiple sources in a data warehouse or data mart. For operational purposes, it's frequently necessary to have real-time access to data resident in operational systems. Organizations can use an operational data store to consolidate current data values from multiple operational systems. They can use an enterprise information integration solution to combine current operational and historical data warehouse data and/or to directly access data spread across several operational systems.

Data integration technology is used to bring this data together. In fact, data integration and data quality solutions are the keys to achieving trusted information. While some organizations choose to develop their own in-house data integration solutions, those that use packaged software solutions can benefit from the vendor's expertise and experience in working with multiple, and sometimes esoteric data sources. This also frees up their staffs for more productive tasks that help gain a competitive advantage. Additionally, commercial data integration products usually provide metadata interoperability with other tools and track data lineage and provide impact analysis. Regardless of how obtained, data integration enables data warehouses, data marts, and operational data stores—which all provide organizations with the means to make reliable business decisions and comply with government reporting requirements. Successful data integration is a key factor for an organization's ultimate business intelligence success. It is the cornerstone of any successful enterprise information management architecture.

Appendix

Many consider Bill Inmon the father of data warehousing. In his book Building the Data Warehouse , he defined a data warehouse as, "a subject-oriented, integrated, nonvolatile, time-variant collection of data in support of management's decisions."

Data Characteristic	Production Application	Data Warehouse
Data use	Operational	Analytical
Level of detail	Detailed	Detailed and summary
Data currency	Real-time, latest value	Multiple historical generations
Longevity	Relatively brief	"Forever"
Stability	Dynamic	Static
Scope of definition	Application-wide	Enterprise-wide
Orientation	Application	Subject
Data operations	Capture/Update	Read
Data per transaction	Limited	Large
Database optimized	For update	For access

Source: Updated from "Data Warehouse—Concepts and Implementation Strategies" presentation, M. Schiff.

While these characteristics are not meant as absolutes for each environment, they represent general statements as to what is typical of each environment. For example, although data warehouse content is obviously updated with new values each time a new snapshot is added, the general use of the data in the warehouse is for read-only analysis purposes. A data warehouse typically adds new, time-stamped values of existing data elements; a production system usually modifies existing values. For example, a production application for payroll might contain the salary of each employee; a data warehouse might contain the salary history for each employee. When an employee receives a salary change, the new value would replace the old value in the payroll system while an additional record, containing the new salary and effective date, would be added to the data warehouse content where it would reside along with the previous salary and quite likely all past salary amounts (or a least a reasonable history) for each employee as well.

There are also times when an organization needs to collect data from several operational systems for additional operational purposes such as determining current part quantities across all of its inventory control systems. This data warehouse variant is commonly referred to as an operational data store. While it differs from the classic data warehouse as it stores relatively current values and minimal history, the process of bringing this data together is another classic example of data integration.

Bill Inmon and Claudia Imhoff highlighted this difference in their book, Building the Operational Data Store , when they defined an operational data store as a "subject oriented, integrated, current valued data store, containing only corporate detailed data."

Appendix (continued)

Enterprise Information Integration or EII is a somewhat hybrid approach that directly accesses data contained in a several operational systems in order to provide a transparent view that makes appear as if the data resided in a single source. Assuming that the data values are compatible, EII can be of value in operational or real-time BI environments or to enable quick analysis of data that has not yet been incorporated into a data warehouse. An EII solution is especially useful when it can access both operational systems and a data warehouse, as it can then provide both real-time and historical values.

[1]W.H. Inmon and Claudia Imhoff, 1996, John Wiley & Sons, Inc.

About MAS Strategies

Michael A. Schiff is the founder and principal analyst of MAS Strategies. MAS Strategies specializes in helping vendors market and position their business intelligence and data warehousing products in today's highly competitive market. Typical engagements include SWOT analyses, market research, due diligence support, technology white papers, public presentations, and helping organizations evaluate tactical and strategic product and marketing decisions. MAS Strategies also assists user organizations in data warehouse procurement evaluations, needs analysis, and project implementations.

With over 30 years of industry experience as a developer, consultant, vendor, industry analyst, and end-user, Michael, is an expert in developing, marketing, and implementing solutions that transform operational data into useful decision-enabling information. Michael was the Vice President of the Data Warehousing and Business Intelligence service at Current Analysis, Inc., an industry analyst firm where he provided tactical market intelligence and analysis while managing the company's E-Business analyst team.

Michael was the Executive Director - Data Warehousing and Advanced Decision Support for Oracle Corporation's Public Sector Group and Director of Software AG's Data Management program where he was one of the industry's earliest proponents of the data mart concept. In 1984, while at Digital Equipment Corporation, he formulated the architecture for one of the first successful data warehouse implementations. In previous positions as IT Director and Systems and Programming Manager he acquired practical, first-hand, knowledge of the technical, business, and political realities that must be addressed for any successful systems implementation or product launch.

Michael earned his Bachelor and Master of Science degrees from MIT's Sloan School of Management where he specialized in operations research as an undergraduate, and in information systems as a graduate.

For further information about MAS Strategies, visit its web site at: www.mas-strategies.com.

About Business Objects

Business Objects is the world's leading business intelligence software company. Business intelligence enables organizations to track, understand, and manage enterprise performance. The company's solutions leverage the information that is stored in an array of corporate databases, enterprise resource planning (ERP), and customer relationship management (CRM) systems.

Popular uses of BI include enterprise reporting, management dashboards and scorecards, customer intelligence applications, financial reporting, and both customer and partner extranets. These solutions enable companies to gain visibility into their business, acquire and retain profitable customers, reduce costs, optimize the supply chain, increase productivity, and improve financial performance.

In December 2003, Business Objects completed the acquisition of Crystal Decisions, the leader in enterprise reporting. The combined product line includes software for reporting, query and analysis, performance management, analytic applications, and data integration. In addition, Business Objects offers consulting and education services to help customers effectively deploy their business intelligence projects.

Business Objects has more than 24,000 customers in over 80 countries. The company’s stock is traded under the ticker symbols NASDAQ: BOBJ and Euronext Paris (ISIN: FR0004026250 - BOB). It is included in the SBF 120 and IT CAC 50 French stock market indexes. Business Objects can be reached at +1 800 877 2340 and www.businessobjects.com.

Printed in the United States – December 2005 PT# WP2084-D.

▶ **www.businessobjects.com**

For a complete listing of our sales offices, please visit our web site.

Business Objects™