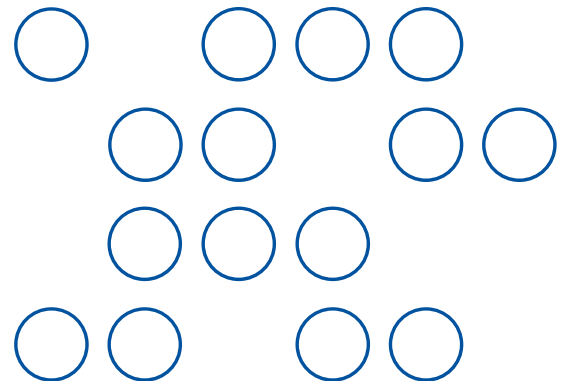




Data Quality and Data Integration: The Keys for Successful Data Warehousing





Overview

Many organizations have successfully implemented data warehouses to analyze the data contained in their multiple operational systems to compare current and historical values. By doing so, they can better, and more profitably, manage their business, analyze past efforts, and plan for the future. When properly deployed, data warehouses benefit the organization by significantly enhancing its decision-making capabilities, thus improving both its efficiency and effectiveness.

However, the quality of the decisions that are facilitated by a data warehouse is only as good as the quality of the data contained in the data warehouse - this data must be accurate, consistent, and complete. For example, in order to determine its top ten customers, an organization must be able to aggregate sales across all of its sales channels and business units and recognize when the same customer is identified by multiple names, addresses, or customer numbers. In other words, the data used to determine the top ten customers must be integrated and of high quality. After all, if the data is incomplete or incorrect then so will be the results of any analysis performed upon it.

Concepts Underlying a Data Warehouse

The data warehouse concept originated in an effort to solve data synchronization problems and resolve data inconsistencies that

resulted when analysts acquired data from multiple operational or production systems. One of the most important functions of a data warehouse is to serve as a collection point for consolidating and further distributing data extracts from an organization's production systems. The data warehouse also must ensure that this data is uniform, accurate, and consistent and thereby serves as a "single version of truth" for the enterprise.

However, this is much more complicated than it might first appear, especially since each production system was developed to satisfy a particular operational need. Consequently, each application system was designed with its own data standards and thus was poorly integrated with other systems. This integration is particularly challenging when dealing with legacy systems that were implemented before any real effort was made to establish enterprise data standards or even common data definitions.

Even if we lived in a world with enough disk space and CPU resources to allow time-stamped data values from each transaction associated with every production system to be saved forever, year-end data purges never took place, and computers could quickly read and aggregate all this data for analysis, data warehouses would still be desirable. At a minimum, the data warehouse would be needed to integrate the data in each system and establish a common format. Moreover, not all of the data

Web-based Retailer

Challenge: Automate business processes associated with consumer purchases:

Solution: The organization deployed a data warehouse to enable information to flow effectively through the company by automating the entire supply chain. From the receipt of a customer order via the Internet, through sending the order to the manufacturer and finally to organizing delivery logistics and invoicing – this retailer dramatically enhanced its end-to-end business processes.

Benefit: The acquired functionality included the ability to:

- Execute order placements
- Monitor the books
- Provide decision support
- Analyze transactional data, product performance, sales and earnings statistics, and information on customer experiences.



an organization requires for analysis purposes is stored in its operational systems. Consequently, data warehouses are frequently augmented with data from third-party content providers. This content might, for example, include customer demographics and lifestyle data, credit information, or geographic data used to determine distances from firehouses, telephone company central offices, or even tax jurisdictions. Data warehouses are also likely to contain derived data fields and summary values resulting from the consolidation of data contained in one or more operational systems.

Even when organizations developed data standards, it was unlikely that they modified the existing operational systems to reflect these standards; rather these standards were applied only when developing and implementing new systems. Consequently, when the data residing in these operational systems was needed to populate a data warehouse, it was often necessary to first transform the data from each source to be consistent with the enterprise data standards prior to loading it into the data warehouse. The data warehouse was sometimes the first attempt and often the first place that the data actually conformed to corporate standards!

Data integration and data quality are the two key components of a successful data warehouse as both completeness and accuracy

of information are of paramount importance. Once this data is collected it can be made available both for direct analysis and for distribution to other, smaller data warehouses.

Variations on a Theme: Data Warehouse, Data Mart, Operational Data Store, EII:

The need to bring consistent data from disparate sources together for analysis purposes is the basic premise behind any data warehouse implementation. Based on this need, various data warehouse architectures and implementation approaches have evolved from the basic concept as originally formulated by Bill Inmon in his book “Building the Data Warehouse” (W.H. Inmon, 1992, John Wiley & Sons, Inc.). Inmon stated, “a data warehouse is a subject oriented, integrated, nonvolatile, time variant, and nonvolatile collection of data in support of management’s decisions.”

There are now a variety of approaches to data warehousing including enterprise data warehouses, data marts, operational data stores, and enterprise information integration. However, most organizations deploy a hybrid combination with each

Components Manufacturer

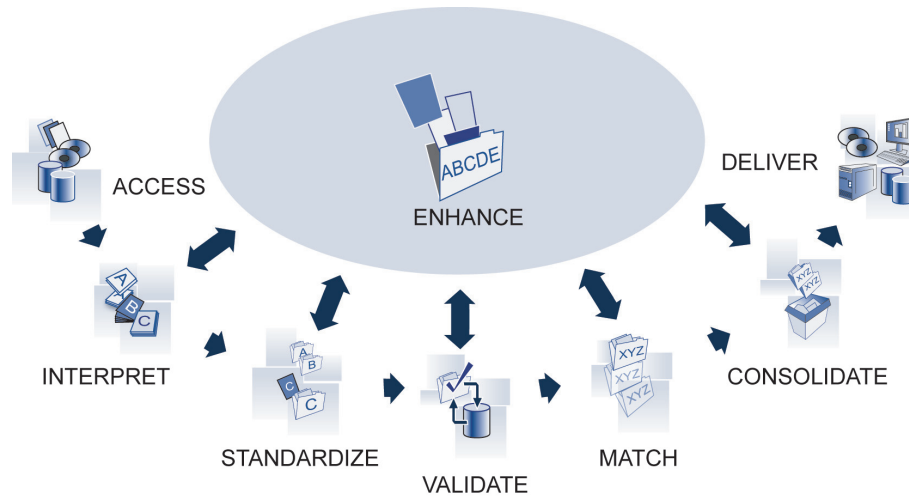
Challenge: Provide access to information on production efficiency, sales activities and logistics and transform it into useful intelligence. Enable users to query data sources without having to rely on IT assistance.

Solution: Implemented a Web-based data warehousing solution, providing the ability to:

- Access and analyze data from anywhere via the Web.
- Analyze group sales profits down to the customer or individual product level.

Benefit:

- Reliability of product shipments
- Reduced manufacturing costs



approach complementing the others. Although they may differ in content, scope, permanency or update cycle they all have two characteristics in common: the need to integrate data and the need for this data to be of high quality.

Data Warehouses

From a conceptual perspective, data warehouses store snapshots and aggregations of data collected from a variety of source systems. Data warehouses encompass a variety of subject areas. Each of these source systems could store the same data in different formats, with different editing rules, and different value lists. For example, gender code could be represented in three separate systems as male/female, 0/1, and M/F respectively; dates might be stored in a year/month/day, month/day/year, or day/month/year format. In the United States “03062004” could represent March 6, 2004 while in the United Kingdom it might represent June 3, 2004.

Data warehouses involve a long-term effort and are usually built in an incremental fashion. In addition to adding new subject areas with each iteration, the breadth of data content of existing subject areas is usually increased as users expand their analysis and their underlying data requirements.

Users and applications can directly use the data warehouse to perform their analysis. Alternately, a subset of the data warehouse data, often relating to a specific line-of-business and/or a specific functional area, can be exported to another, smaller data warehouse, commonly referred to as a data mart. Besides integrating and cleansing an organization’s data for better analysis, one of the benefits of building a data warehouse is that the effort initially spent to populate it with complete and accurate data content further benefits any data marts that are sourced from the data warehouse.

Data Marts

A data mart, if populated from a data warehouse, contains a subset of the data from the data warehouse. If this is the case, then it is generally considered to be a dependent data mart and can be implemented relatively quickly as the data has already been collected and integrated within data warehouse. The quality of its content is directly dependent upon the contents of the data warehouse. Independent data marts are those that are developed without regard to an overall data warehouse architecture, perhaps at the departmental or line-of-business level, typically for use as a temporary solution.



Operational Data Stores

As the independent data mart cannot rely on an existing data warehouse for its content, implementation will take longer than a dependent data mart, assuming, of course, that the data warehouse used to populate the dependent data mart already existed. Just because a data mart operates independently of any other data mart or data warehouse, it is nonetheless still important that the data within it be complete and accurate. If not, erroneous analysis is likely to occur and invalid conclusions drawn.

Pragmatically, an independent data mart may be the only viable approach when the existing enterprise warehouse is being built incrementally and the data needed by the data mart is not yet available from the warehouse. Building a corporate data warehouse on a “subject by subject” approach is certainly a reasonable and proven strategy. Many organizations that have tried to populate their enterprise data warehouses with data for all requested subject areas prior to initial rollout have found that this was akin to attempting to trying to “boil the ocean,” the task was simply too overwhelming to be realistically accomplished in anything other than a phased approach.

It is reasonable to assume that an organization’s independent data marts will ultimately be combined. Eventually they will lose their independence as individual data needs are ultimately satisfied through an enterprise data warehouse.

Combining the content requirements of these independent data marts to determine the contents of the enterprise data warehouse will be significantly easier if each data mart contains high quality, complete data. This “bottoms up” approach of using the requirements of existing independent data marts to then determine the requirements of a data warehouse from which they will be populated has been effective in organizations where several departments first needed to quickly implement their own solutions. These organizations could simply not wait for their “top down” data warehouse to first be built.

A common problem that exists in many organizations is the inability to quickly combine operational data about the same entity such as a customer or vendor that exists in multiple systems. A classic example occurred when banking institutions first started adding new service offerings such as investment accounts to their more traditional savings and checking account offerings. Many of these new services were supported by systems that existed independently. When the bank needed to see all of the current financial information it had about a customer, it needed to combine and consolidate data from all of these systems, assuming of course it could identify that a customer whose account information resided in several systems, was the same customer. As this need became increasingly more important, the operational data store (ODS) came into vogue.

A primary difference between data warehouses and operational data stores is that while a data warehouse frequently contains multiple time-stamped historical data snapshots, with new snapshots being added on a well-defined periodic schedule, an operational data store contains current values that are continually in flux. A data warehouse adds new time-stamped data values and retains the old ones; an operational data store updates existing data values. While the initial load and continual updating of the operational data store are classic examples of data integration, the ability to identify and link different accounts each captured from a different system, as belonging to the same customer is also a classic example of data quality. This underscores the importance of, and interdependence between, data quality and data integration, when solving real-world business problems.

Enterprise Information Integration (EII):

While not necessarily a new concept, the idea of enterprise information integration, or EII, has received much publicity in the past few years. Simply stated, it involves quickly bringing together data from multiple sources for analysis purposes without necessarily storing it in a separate database. Some vendors even have gone so far as to claim that an EII approach can replace a



Data Warehousing Variants				
	ODS	DATA WAREHOUSING	DATA MART	EII
VALUES	Current	Historic, possibly current	Historic, possibly current	Current
RIGOROUS DATA QUALITY	Usually	Yes	Dependent-Yes Independent-Unlikely	Unlikely
PRIMARY USE	Operational & Tactical	Tactical & Strategic	Tactical & Strategic	Operational & Tactical
ORGANIZATIONAL SCOPE	Functional or Line-of-Business	Enterprise	Functional or Line-of-Business	Depends on Sources
IMPLEMENTATION TIMEFRAME	Intermediate-Term	Long-term	• Dependent-Short • Independent-Intermediate	Short-term
LEVEL OF DETAIL	Detailed	Detailed & Summary	Detailed & Summary	Depends on sources
DATA VOLATILITY	High -values added	Low-values added	Low-values added	Depends on sources
RELATIONSHIP TO OTHER VARIANTS	Complementary	Complementary	Complementary	Complementary
An organization's data warehousing architecture can consist of a variety of components that co-exist and compliment each other.				

traditional data warehouse or data mart with a “virtual data warehouse” by eliminating the need to extract and store the data into another database. However, the ramifications associated with this approach (e.g., such as the underlying data changing, or even being purged, between analysis) must be not be overlooked.

That said, an EII solution certainly complements the other data warehousing variants and can be a valuable resource for those wishing to perform quick, perhaps ad hoc, analysis on current data values residing in operational systems. It can help alleviate the backlog of requests which are a constant struggle for any IT staff.

Organizations must, however, recognize that the data in these operational systems may not be consistent with each other, that the data quality of each source may vary widely, and that historical values may not be available. This is a risk many users are willing to take for “quick and dirty” analysis when the needed data is not contained in a formal data warehouse or data mart. In fact, many organizations use an EII approach to establish processes and programming logic that enable their

users to transform and pull together data from multiple sources for purposes that include desktop analysis. Of course, if the quality of the data in the underlying operational systems were high, the EII analysis would obviously benefit.

EII solutions can also be successfully used to prototype or evaluate additional subject areas for possible inclusion in a data warehouse. Some organizations have initially deployed EII solutions when the data warehouses or data marts did not contain the needed data and later added this data to their data warehouse content. In order to combine current and historical values, organizations can include an existing data warehouse or data mart as one of their sources and thus combine historical and current data values.

Data Integration and Data Quality

While estimates vary, it is generally agreed that data integration and data quality represent the majority of the cost of implementing a data warehouse. Lack of data integration and poor data quality are the most common causes of post-implementation



data warehouse failures. First impressions count; if decision-makers find that the data warehouse contains incomplete or incorrect data, they are likely to seek their data elsewhere.

At a very simple level, data integration and data quality are concerned with collecting accurate data, transforming it into a common format with a common set of values, providing appropriate aggregations or summary tables, and loading it into the data warehouse environment. This sounds simple enough but there are many complicating factors that must be considered.

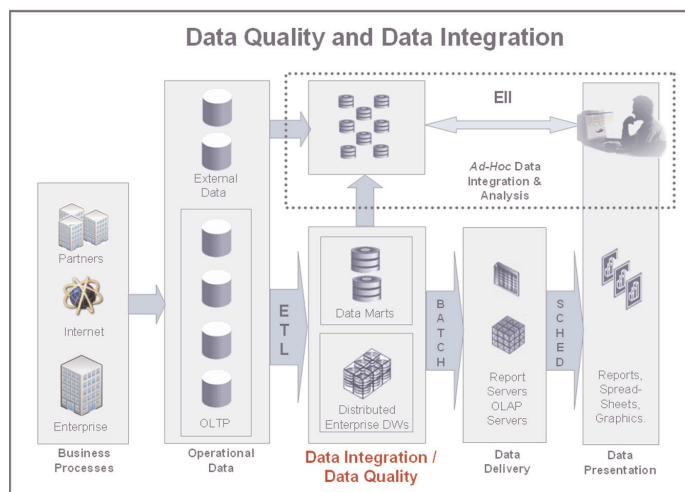
Multiple Data Sources

The requisite data is likely to be stored in a variety of systems, in a variety of formats, and on a variety of platforms. Assuming the required data resides in a computer system, not on paper in a file cabinet, the data sources may be relational databases, XML documents, legacy data structures (such as Adabas, IDMS, IMS, VSAM, or even sequential EBCDIC files). It may be contained in packaged enterprise application systems such as SAP or Siebel where knowledge of the business logic is necessary to understand and access the underlying data. The data may reside on a wide variety of computing platforms including mainframe, Unix, Windows, and Linux environments.

Data Transformation

The detailed data residing in the operational systems must frequently be consolidated in order to generate and store, for example, daily sales by product by retail store, rather than storing the individual line-item detail for each cash register transaction.

Complex arithmetic or statistical calculations frequently need to be applied to the source data in order to perform, for example, percent of sales calculations or “top x” rankings, especially if items that of low value are grouped into an “others” category before being loaded into the warehouse.



Data Linking

In many cases data records concerned with the same object (for example, a given customer, employee, or product), resides in multiple source systems. These records must first be linked together and consolidated prior to being loaded into the data warehouse. Integration with data quality software is often the only realistic way of matching these records, especially when trying to deal with the nuances involved identifying customers or vendors. Each system could contain its own variation, in format or spelling, of the same customer name and/or address. Once again, data quality software can greatly facilitate this task.

Data Augmentation

Data warehouses and data marts are not exclusively populated from the data contained in an organization’s operational systems. It is frequently desirable to augment this data with information from outside sources. While organizations can try and collect this on their own, there are many commercial sources of company-centric and people-centric data that can be used to augment data warehouse content.



Data Volumes

It is frequently necessary to load very large data volumes into the warehouse in a short amount of time, thereby requiring a parallel processing and memory-based processing. While the initial data loads are usually the most voluminous, organizations have a relatively long load window in which to accomplish this task since the initial load is done prior to the data warehouse being opened for production. After the data warehouse is in use, new data content must be loaded on a periodic basis. The load volume can be reduced if change data capture techniques are employed to capture only data that has changed since the prior data load. In some cases, Enterprise Application Integration (EAI) technology, frequently involving message queues, can be used to link enterprise applications to the data warehouse data integration processes in order to capture new data on a near-real-time basis.

Collaborative User/IT Development Efforts

Many data warehouses have been implemented only to quickly discover that the data content was not what the users had in mind. Much finger pointing and general ill will can be avoided if the data integration staff can work collaboratively with the end-user analysts. They should be able to view the results of the data transformation process on real data rather than trying to interpret somewhat abstract data flow diagrams and transformation descriptions. The ability to view live data and the associated transformation processes involved in the data integration process can help avoid nasty surprises when, for example, a field in the source system thought to contain telephone numbers actually contains text data.

Energy Provider

Challenge: Collect and analyze large volumes of data in different formats from both internal and external sources to optimize business processes and make predictive calculations for planning purposes.

Solution: A data warehouse that provided the infrastructure to manage high volumes of data critical to decision-making processes:

- Hourly forward and historic energy and capacity positions; including mark-to-market (exposure price volatility) for two years into the future. The positions can be viewed hourly, daily, weekly, monthly and yearly.
- Retail choice analysis and statistics, including alternate supplier load and capacity obligation, which can be viewed by load aggregator, supplier, zone, and rate class
- Customer shopping statistics that provide forward and historic views of customer demand, including the measurement of customer churn
- Weather analysis and statistics for historical and forward views of temperature, dew point, and wind speed
- Portfolio performance reporting that measures the impact of the business decisions over time
- Short-term energy deal "what-if" analysis.

Benefit:

- Ability to manage its data in all of its different formats
- Develop tools for analysis
- Ultimately deliver it to the browsers of market analysts and managers in its organization

Consequently, the company was able to make the best decisions possible using the best information available.



Changing Requirements

A successful data warehouse builds user momentum and generates increased user demand that results in a larger user audience and new data requirements. The data integration processes must be able to quickly respond to new data sources, or changes to the underlying file structure of the existing source systems, without compromising existing processes or causing them to be rewritten. Increased user demand often translates into a narrower data warehouse load window, especially if new users are in geographic areas that now require access to the data warehouse at times during which there was previously little or no user demand.

Metadata Integration

Most data integration tools store the metadata (or data about data) associated with its sources and targets in a metadata repository that is included with the product. At a minimum, this metadata includes information such as source and target data formats, transformation rules, business processes concerned with data flows from the production systems to the data warehouse (i.e., data lineage), and the formulas for computing the values of any derived data fields. While this metadata is needed by the data integration tool for use in defining and creating appropriate data transformation processes, its

value is enhanced when shared with other tools utilized in designing the data warehouse tables and business intelligence tools that access the warehouse data. If the metadata also includes information about what analysis program uses which data element, it can be a valuable source for analyzing the ramifications of any change to the data element (i.e., impact analysis).

Additional Thoughts on Data Quality

Data quality is involved throughout the entire data warehousing environment and is an integral part of the data integration process. Data quality involves ensuring the accuracy, timeliness, completeness, and consistency of the data used by an organization while also making sure that all parties utilizing the data have a common understanding of what the data represents. For example, does sales data include or exclude internal sales and is it measured in units or dollars, or perhaps even Euros?

In most data warehousing implementations data quality is applied in at least two phases. The first phase is concerned with ensuring that the source systems themselves contain high quality data while the second phase is concerned with ensuring that the data extracted from these sources can then be combined and loaded into the data warehouse. As mentioned earlier, even if the

Trends in Data Warehousing

Several trends are developing in the data warehouse market, many of which are directly concerned with data integration and data quality. These include:

- EAI and ETL, will continue to converge due to the need to update the data warehouse with the recent transactions.
- The use of “active” data warehouses that directly feed analytical results back to operational systems will grow
- Pragmatic hybrid approaches to data warehousing will continue to win-out over insistence on architectural purity
- Data quality will be recognized as an up-front requirement for both operational and analytical systems efforts, rather than an after-the-fact fix
- EII will succeed when marketed as complementary to, not a replacement for traditional data warehouses and data marts
- Disparate data integration tools will give way to end-to-end data integration platforms that provide end-to-end data integration functionality
- Data integration platforms will be callable both through direct application programming interfaces and as Web services



data residing in each of the sources is already accurate and clean, it is not simply a matter of directly combining the individual sources as the data in each source could exist in a different format and use a different value list or code set. One system might use of the alphabetic codes (S,M,D) to represent “single,” “married,” and “divorced” while another might represent them with the numeric codes (1, 2, 3). The data loaded into the warehouse must conform to a single set of values; data cleansing and data transformation technology must work together to ensure that they do. Of course, duplicate occurrences of the same customer or vendor across multiple systems, or even in the same system, with different variations of the same name and/or address, is a well-known example of a data quality issue that was previously discussed.

Approaches to DQ/DI The “Do it Yourself” Approach

Many organizations have attempted to access and consolidate their data through in-house programming. After all, how difficult can it be to write a few programs to extract data from computer files? Assuming for a moment that the files are documented (and the documentation up-to-date), the programming team has in many cases succeeded, although usually after the originally estimated completion date. Unfortunately, the initial extract and load is usually the easy part!

It is a fact of life and systems that “things change.” Even when the initial data integration programs work as required, there will be a continuing need to maintain them and keep them up-to-date. This is one of the most overlooked costs of the “do it yourself” approach to data integration and one that is frequently ignored in estimating the magnitude of any in-house data integration effort. This approach frequently does not consider data quality. Even when it does, only the most obvious data quality issues are considered as the organization’s programming staff does not have the time or experience to build strong data quality tools that are comparable to those offered by commercial data quality vendors.

Additionally, most packaged data integration software has a metadata repository component that allows for sharing of metadata with other data warehouse components such as database design and business intelligence tools. However, in-house software frequently does not provide for sharing its own the metadata or leveraging the metadata collected by other data warehouse components. In fact, the metadata collected in the “do it yourself” approach is usually rather limited and may only be contained in COBOL file descriptions for the input and output formats or in the actual program code for the transformation and aggregation logic. In general, metadata residing in “home-grown” software cannot be readily shared with other data warehouse tools.

Commercial Data Integration Tools

Fortunately, the industry has recognized the need for data integration tools and a variety of offerings are commercially available. An appropriate tool should be capable of accessing the organization’s data sources and provide connectors for packaged enterprise application software systems and XML data sources. It should include a powerful library of built-in data transformation and data aggregation functions that can be extended with the additional of new functions developed by the deploying organization. It must, of course, be able to populate the target data warehouse databases, be they relational databases or proprietary OLAP cubes.

The tool should also have sufficient performance not only for the initial data requirements, but also for the additional content, or additional generations of current content. This can be accomplished perhaps through the use of change data capture techniques that can be expected in the future as the data warehouse subject areas grow. Sufficient headroom should be available to be able to handle not only the current update frequency and data volumes but also anticipated future requirements. Both batch ETL loads and (should future plans require this capability) event-driven “near-real-time” EAI transactions should be supported.



Commercial Data Quality Tools

The tool should provide a simple, yet powerful, user interface allowing users to visually perform the data integration tasks without having to write low-level code or even utilize a high level programming language or 4GL. It should be able to be used by a variety of audiences, not just data integration specialists. Users and data integration specialists would be able to collaborate in an iterative process, and ideally view the transformation process against live data samples prior to actually populating the data warehouse.

The data integration tool should itself be easy to integrate with other technology in use by the organization and should be callable through a variety of mechanisms including application programming interfaces (APIs) and Web services. The technology should also be deployable as a standalone tool. Of particular importance is the ability to perform data cleansing and validation or to be able to be integrated with tools that can provide data quality. Strong metadata capabilities should be included, both for use by the data integration tool itself, and to facilitate the sharing of metadata with the business intelligence tools that will access the integrated data. If directly licensed the product should operate in a “lights out” environment by setting up appropriate jobs steps and event-driven conditional error handling routines, with the ability to notify the operations staff if an error condition is encountered that cannot be resolved. It should be offered with several pricing models so it can be economically deployed, such as through an application services provider (ASP) for utilization by small organizations, or smaller units of large enterprises.

Many organizations have attempted to resolve data quality issues through their own devices as well. While edit routines can be built into programs to check for proper formats, value ranges, and field value dependencies in the same record, these are relatively simple when compared, for example, to ensuring that a customer address is a valid and up-to-date.

Data quality vendors, once best known for name and address correction and validation, have significantly expanded their capabilities in terms of the scope of the data they can handle

(i.e., non-name-and-address data), the volumes they can support, and the databases they maintain in order to validate data. Many have expanded their offerings to include both software licensing and ASP delivery.

Data Integration Tools Supplied by Database Vendors

Most database vendors now offer data integration tools packaged with, or as options for, their database offerings. Although these tools should certainly be considered when evaluating data integration products, it is important to recognize that database vendors want their own database to be at the center of their customers’ data universe.

Consequently, a data integration tool from a database vendor could very well be optimized to populate the vendor’s own databases by, for example, taking advantage of proprietary features of that database. If used to populate other databases, they may not perform as well, or even at all.

That said, if an organization has standardized on a particular database for all of its data warehousing projects, the data integration tools offered by that database vendor could be used as the basis against which other data integration tools are compared.

Summary

Today almost all organizations recognize the significant advantages and value that data warehousing can provide both for pure analysis and as a complement to operational systems. While data warehouses exist in many forms, including enterprise-scale centralized monoliths, dependent and independent data marts, operational data stores, and EII implementations, they all benefit from complete, consistent, and accurate data. After all, what is the value of any analysis that are based upon faulty or incomplete information?



While an organization's overall data warehouse architecture can encompass a variety of forms, each organization must decide what is right for its own purposes and recognize that implementing a successful data warehousing environment is a continuous journey, not a one-time event.

Whatever the choice, two things are certain: data integration and data quality will be key components of, if not the enabling technology for, the organization's data warehousing success. Data integration is an ongoing process that comes into play with each data load and with each subject area extension; the quality of the data in the warehouse must be continually monitored to ensure its accuracy. Organizations that ignore these requirements must be careful that instead of building a data warehouse that will be of benefit to their users, they do not inadvertently wind up creating a repository that provides suboptimal business value.

Many data warehousing industry vendors can provide robust data integration and data quality solutions. In addition to developing and marketing products, these vendors offer a wealth of experience and expertise that they can share with their customers. As a result, an organization is best served when it deploys a commercial, fully supported and maintained set of tools rather than trying to develop and maintain such a technology on its own.



Group 1 Software
4200 Parliament Place • Suite 600
Lanham, MD 20706-1844
1-800-368-5806

For more information about our products and services, please log onto our website: www.g1.com

For more information about Pitney Bowes, visit www.pb.com

Group 1 is a registered trademark of Group 1 Software, Inc. Pitney Bowes is a registered trademark and the Pitney Bowes Process Bar Design is a trademark of Pitney Bowes Inc. All other trademarks are the property of their respective companies.