Data Integration

Best Practices, Methodologies and Emerging Trends

Data Meaning

Table of Contents

Introduction

In business, the only thing more dangerous than no information is inaccurate information. Information is the cornerstone of sound business decision-making. Corporate planning and strategy development rely on the accuracy of the Business Intelligence (BI) provided to business leaders by various corporate divisions and departments. Within operational systems, inaccurate information can halt production if vital raw material is out-of-stock or spare parts cannot be located. Marketing strategies rely on accurate analysis of customer behavior and trends. Unfortunately, many organizations struggle to produce high-quality, meaningful information in a timely manner. Too often reports generated by different sections of the organization are confusing and contradictory because of different terminology, time-scales, and dimensions.

Fragmentation of corporate information happens in a number of ways. Departments may purchase software applications in isolation, or budgetary constraints may result in the implementation of partial solutions. Developments in technology may lead to additional demands for information that must be fulfilled rapidly, resulting in a quick-fix approach. Data may be downloaded and manipulated in user-designed spreadsheets which do not link to any other business software. Uncontrolled spreadsheets proliferate into spreadmarts and become the basis for an organization's BI.

The problem with fragmented information is that business leaders cannot be certain that the data driving their planning and decision-making is reliable, robust, and comprehensive. Decisions based on incomplete or inaccurate data may be flawed, inconsistent, and lead to costly mistakes.

From an operational point of view, the manual keying and interventions required to transfer data from one application to another are time-consuming and increase the risk of error and miscalculation.

To draw relevant information from multiple databases, legacy systems, and enterprise applications, companies must identify and implement data integration solutions that best suit their business needs. This paper examines current data integration best practices and emerging trends.

The Importance of Data Integration

Business information systems can only be effective if they are underpinned by an infrastructure which integrates and analyzes relevant information from the entire range of corporate software. The bulk of an organization's information is likely to be stored within unconnected databases and legacy systems. Data integration applications are required to extract data from multiple sources and convert it into usable information, from which meaningful, insightful business reports can be generated and trends identified.

Operational and transactional systems, such as procurement, inventory control, and customer relationship management, also benefit from data integration. For example, a company can only maximize its buying power if it clearly understands the extent of its relationship with a supplier. If several locations are placing small orders, individual discounts may be minimal. If the combined orders are significant, the company can leverage greater price reduction across the organization. Ensuring that data is integrated across the organizations delivers savings in cost and time while increasing the reliability and usability of business information.

Data Warehouse

One of the most common strategies in data integration is to create a data warehouse. This offers an organization the ability to combine, standardize, and analyze data drawn from numerous operational systems within one location. Integrating operational data within a data warehouse can offer significant benefits to an organization, such as the ability to compare historical and current data. Also, corporate data standards can be established without modifying existing operational systems, and performance of operational systems can be preserved.

Operational applications tend to focus on the current status to retain limited historical information. Values change constantly as new transactions are recorded. Lack of historical data makes it difficult to track changes and analyze trends within operational systems. By capturing and combining current and historical information, data warehouses enable the period-to-period comparisons essential in trend analysis and forecasting.

When data is stored and analyzed in a variety of different applications, the reports produced can be diverse and even contradictory. None of them are necessarily incorrect, but the use of different definitions, terminology, and logic can lead to different results, sometimes referred to as multiple versions of the truth. These differences are rectified within the data warehouse where data is transformed to corporate standard definitions. The result is consistent, accurate reporting which enables business leaders and decision-makers to have confidence in the information supplied to them.

Depending on the amount of information to be processed, running queries or reports against an operational database can adversely impact user response time and performance of the application. Running queries and reports against the information contained in the data warehouse avoids the potentially negative impact by offloading the processing burden to a separate environment.

Data Integration Methods

Data integration technology offers a range of tools from which companies can build an infrastructure that best suits their requirements. Data integration tools cleanse and reconcile data drawn from multiple sources, enabling the assimilated data to be used for reporting, analysis, and other critical business functions.

Extract, Transform and Load (ETL) systems gather data from a variety of data sources, provide it with a common structure, and place it in a single data store. These data stores can be a data warehouse, data mart, or repository, and processing is usually carried out in batches. ETL offers large-scale efficiencies through the automation of integration processes, including data migration, configuration, synchronization, upgrading, and archiving.

The Extract, Load and Transform (ELT) method is similar to ETL, the main difference being that data is transformed after it has been loaded into the target database or repository. ELT is also a batch-processing method.

Data federation, or data visualization, allows data from multiple sources to be viewed by the data integration application without the requirement to move or copy data. Query processing capability creates a snapshot of the information required while leaving the data unchanged in the source repository. Data federation is most useful when users are looking for a complete view of an entity at a single moment in time. For example, if a salesperson is contacting a customer, he may use data federation technology to view all of the information held in various applications that relate to that customer in real-time while on the call. Data federation applications have limitations as they do not provide the historical data required to support tracking and trending.

Real-time data integration identifies changes in a data source and updates the organization's data warehouse or BI solution in real time. The relative merits of real-time integration compared to batch updates are examined below.

The data integration solution a company chooses to implement will depend on which method best suits its business needs. Companies may even use a combination of methods depending on the purposes for which data integration is required.

Real-Time Integration vs. Batch Updates

With the proliferation of mobile communication devices, such as BlackBerrys and iPhones, individuals have come to expect information to be provided instantly. The pressure on IT departments to respond rapidly to requests for information has increased demand for real-time integration. Change data capture (CDC) tools and technologies recognize when a change is made to one data source and transmit the change to the organization's data warehouse or BI solution in real-time.

Real-time integration is important for businesses that need to make instant decisions. It has a place in short-cycle businesses that require immediate analysis of their supply chain. It is also useful when synchronizing operational applications. However, the trade-off for the availability of real-time data integration is the possibility of poor data quality.

Batch updates remain the most widely used method for updating data integration applications. Data integration processes which use batch updates have plenty of time to cleanse data to ensure consistency before it reaches its target destination. Batch updates take place at regular intervals determined by the business.

Companies should match their integration methods to their data latency thresholds. An organization which analyzes data on a weekly basis does not require real-time integration. Their goals can be met by integrating batch data, which will ensure data quality.

In-House Development vs. Vendor-Supported Solutions

In a study of 359 IT executives, managers, and staff, conducted by the BeyeNETWORK in 2011, the third most commonly cited data integration tool was "hand-coded/in-house solutions," after Microsoft and Oracle. Hand-coded/in-house solutions were selected by 29 percent of respondents.

Internal development of data integration systems often appears to be an appealing method of resolving the issue of fragmented information. Organizations may initially see some benefit in a rapidly-implemented, inexpensive solution developed in-house. However, these benefits can be outweighed by the inability of such a solution to respond to changing demands.

In-house solutions may be developed in a disjointed manner. Often the requirement to integrate data from multiple sources is driven by a need to fulfill a specific request for information. A programmer writes the code to extract and transform the necessary data. Soon this quick-fix becomes an established production report on which vital decisions are made. Follow up requests or changing requirements result in further data fields being added. As a result, the task of supporting and maintaining the integration solution becomes complex and time-consuming.

Personnel changes may result in an in-house solution becoming unsupported as the development knowledge is lost. The programming language and approach taken may be difficult for another programmer to follow, if he has not been involved in the initial design. In large organizations, there may be a number of programmers writing data integration solutions for specific requests. The result may be an undocumented and unmanageable series of solutions which do not communicate with each other.

Vendor-supported solutions may appear to be more expensive initially. However, purchasing a tried-and-tested commercial data integration package avoids the need for the business to develop, modify, and continually retest internal solutions. Specialist developers working full-time on similar projects ensure that vendor-supported solutions contain the most up-to-date business solutions. Ongoing updates and product maintenance are usually offered as standard, reducing the burden on the organization's IT department.

Commercial data integration products can perform data transformations and aggregations, avoiding the need for custom coding and code maintenance. Built-in features allow businesses to cleanse data, ensure consistency and manage error handling and performance monitoring.

Emerging Trends

There are challenges ahead for IT departments as they manage, manipulate, and process business information. Organizations are demanding data be extracted from escalating volumes of unstructured data at the same time as shrinking response times. Increasingly complex integration tools are required to combine new technology with existing and legacy systems.

With email, video-clips, and web-based analytics becoming established as business communication tools, the requirement to pull usable data from unstructured sources becomes more pressing. In the BeyeNETWORK 2011 study, 70 percent of the respondents agreed with the statement "Growing data volumes and shrinking processing windows are significant challenges in our organization."

Software as a Service (SaaS) applications offered less technical departments the ability to subscribe to niche software which could be set up quickly with no large capital investment up front. Users' principal concerns at purchase were likely to be functionality, security, and reliability. Often purchased in isolation, departments increasingly want to share information between multiple SaaS applications and between SaaS applications and legacy systems. Integrating SaaS applications is a challenge which requires the functional department and IT professionals to work together.

Summary

In summary, organizations must take time to examine their data integration needs before committing to a particular route. For many organizations, a combination of data integration methods will be necessary to address all their current and future requirements.

The need for real-time data integration must be balanced against the potential adverse impact on information quality. Unless the organization has a requirement for low latency or high concurrency, the quality benefits of batch processing tend to outweigh the loss of quality.

For any data integration solution that is more than a one-off response to a specific request, organizations are likely to be better served by a vendor-supported commercial data integration package. Short-term cost-savings resulting from internal solutions are likely to be outweighed by the longer-term issues around maintenance, development, and in-house testing.

Data integration must continue to develop to incorporate new types and sources of data if it is to ensure that robust, accurate, and timely information is supplied to meet changing business needs.

About Our Group

Data Meaning utilizes cutting-edge technologies to build innovative and effective Business Intelligence and Data Warehousing solutions. Our experienced, professional staff can design and deliver pioneering reporting systems to give you a unique perspective to your data and an edge in your decisions.

Along with the world-class consulting services Data Meaning offers, they are also an official licensed reseller of the award-winning MicroStrategy Business Intelligence Reporting Suite, a fully integrated BI platform that makes Business Intelligence faster, easier, and more user-friendly. Data Meaning has MicroStrategy certified consultants available to help you deploy MicroStrategy with ease. For your BI and DW design, install and implementation and training needs please visit us at **www.datameaning.com** or email **info@datameaning.com.**