Delivering information you can trust
September 2006

IBM Information Management software

Designing an on demand data warehouse

ON DEMAND BUSINESS™

Contents

In an increasingly online world, customers, partners and other companies have made an enormous investment in data warehousing to help produce the best possible information for business decisions. Although the analysis of this data is largely constrained to historical trend analysis for forecasting and strategic planning, it has the potential to deliver much more value to the organization if it is brought closer to operational processes. This is where the information can be used to make better operational decisions—programmatically. For example, a detailed, historical market trend analysis made available to an automated demand planning process can help produce better decisions and lower costs. On demand data warehousing is focused on this concept, creating a data warehouse that can provide current analytical information directly into business processes.

This paper describes how to apply the concepts of a Service Oriented Architecture (SOA) to your existing data warehouse investment to deliver information as a service to the business. This approach allows your data warehouse to assume a more operational role within your company. It also provides a mechanism to reduce data latency and help ensure that decisions are being made on the most current and accurate information. By leveraging SOA and providing information as a service, you can broaden the reach of the data warehouse to a much wider audience within your business, providing all applications, processes and users with the right information, in the right context, at the right time.

Data in the data warehouse provides high value to your business

When businesses need to make strategic decisions, they usually turn to a data warehouse for the richest and most complete information available. Data warehouses earn this distinction by taking data from the silo context of each source system and merging it together to provide better information transparency across the business.

The value of the data within the data warehouse is evidenced by the rise in spending on data warehousing projects. According to a recent TDWI–Forrester Research survey, 41 percent of practitioners expect spending on data warehousing to increase by at least 11 percent in the next budget cycle.[1] Similarly, an IDC survey found that the total number of business intelligence projects involving data integration is expected to grow 23 percent in the next 24 months versus the previous 24 months.[2]

To support this growth, data within data warehouses has several characteristics that provide value above and beyond data found in operational systems. These characteristics include:

Complete. Data in the data warehouse represents a superset of data across multiple operational applications. Because this data is not limited to the context of a specific function or application, it automatically provides a more complete view of information. The inherently broad scope of the information in the data warehouse cannot be achieved in individual transactional systems.

Accurate. Most data is cleansed and validated before it is entered into the data warehouse. This process includes de-duplication and validation vis-à-vis business rules and standards. When data is cleansed, the result is significantly higher quality data, which provides business users with confidence in their decision making and results in better decisions.

Aggregated. While operational systems focus on individual transactions, data warehouses have a much broader scope. They are capable of aggregating information across many transactions and systems. This aggregation is one aspect of data warehouses that makes them particularly good for strategic analysis.

Enriched. In addition to providing a comprehensive (complete) view of data across source systems, data warehouses often provide enrichment of that data from external sources. Although it is applicable across all domains, enrichment is typically associated with customer data, where data services like those provided by Acxiom, Experian and Dun & Bradstreet can provide verification and enhancement of individual and company data. This enrichment not only improves the quality and accuracy of the data, but it also provides valuable context for marketing purposes.

Auditable. Data warehouses are capable of expressing changes within information across time. They also typically track metadata about where data came from and what happened to it along the way. These characteristics make data warehouses an excellent source of audit information.

Operational processes should be linked to the data warehouse

Data warehouse implementations in most companies support only historical or strategic analytics and reporting activities; they are not linked into operational processes and applications. Yet, substantial benefits can be gained by providing analytical feedback into upstream processes for "closed-loop" processing. According to Forrester Research, "Combining active data warehousing with complete [point of sale] data is enabling merchandisers to close the loop between demand and sourcing, capturing literally billions of dollars previously left on the table."[3] Two technical inhibitors, however, prevent most companies from implementing closed-loop processing: data latency and limited data access.

Data latency

Most data warehouses contain a snapshot of data that is repopulated only during batch cycles. In fact, according to a TDWI (The Data Warehouse Institute) survey, only 6 percent of respondents claimed to load their data warehouses in near-real time.[4] While this delay is fine for strategic analysis, it introduces a synchronization issue for closed-loop processing, because data used to make a decision may not align with the state of data in transactional and operational systems.

For closed-loop processing to work, data latency in the data warehouse must be within an acceptable and practical range for the particular type of data. Many processes that would benefit from closed-loop analytics are very sensitive to the exact current state of information. For example, inventory allocation processes need to have an accurate, up-to-the-minute view of data to be effective. A single transaction that contains a large order at one location can have an impact on how allocation decisions are made.

The good news is that most companies recognize the value of reducing analytical latency. According to the TDWI survey cited above, 19 percent of the same respondents expect their data warehouses to be refreshed in near-real time within 18 months of that report.[5]

Limited data access

Data warehouses are typically accessed using business intelligence tools such as ad hoc query, online analytical processing (OLAP) and data mining tools. These tools are optimized for the design of the data within data warehouses and provide an easy way to derive valuable information from the data. These tools are perfect for strategic analysis, since they provide all the functionality necessary to analyze and explore data from many different angles. However, these tools are tailored for human analysis and often limited in their distribution within an enterprise. This means many processes that could benefit from the data simply do not have access to it. Increasing access to information within the data warehouse is vital to creating effective closed-loop processes using this data.

Most processes that could benefit from closed-loop analytical data are designed into applications or enterprise integration technologies. These technologies communicate using specific protocols that are not supported by most business intelligence tools. In addition, as soon as analytical data is inserted into operational processes, that data becomes even more mission-critical. Considerations like high availability, performance under heavy loads and fault tolerance become extremely important. Most data warehousing and business intelligence infrastructures are not designed to meet these operational demands.

SOA provides a framework to deliver information

Service-oriented architecture provides a framework for overcoming these two technological inhibitors. Within an SOA, functional components are loosely coupled to each other, allowing one component to easily call others using simple, open interfaces. SOA provides an excellent way to package existing functions so they are easy to catalog and find, easy to publicize and socialize, and easy to integrate into enterprise integration technologies.

By applying this concept of delivering information as a service to data integration, functions related to creating and delivering data from the data warehouse can be packaged as services that can be easily called by processes and applications. This architecture can directly overcome the technological inhibitors described above.

From a data latency perspective, the ETL (extract, transform, load) processes that load the data warehouse can be repackaged as services so they can be called on demand. Instead of waiting for batch cycles to update the data warehouse, the data can be "trickle" fed into the data warehouse throughout the day as events occur in source systems. The effect is that the data warehouse can contain near-real-time data, using exactly the same business rules and validations used today to load it. Of course, one requirement is that the ETL tool must be capable of publishing logic as services, and these services must be able to handle the demands of on demand data processing.

To expand access to the data beyond business intelligence tools, services can be created that deliver the data upon request to other applications. These services package analytical data from the data warehouse and deliver it to processes and applications that request it. Again, a Service Oriented Architecture is ideal because it provides that information as a service to the requester in a predictable way that is easily plugged into processes and applications.

IBM calls the combination of these two concepts "on demand data warehousing." Both the creation and consumption of data warehouse information can be accomplished on demand via a Service Oriented Architecture.

On demand data warehousing can provide extraordinary benefits

The benefits of on demand data warehousing and delivering information as a service can be extraordinary. Tying rich and accurate analytical data into the applications and processes that automate decisions in standard, reusable services can provide great opportunities for achieving operational efficiencies. These efficiencies can have an enormous impact across many different activities in multiple industries. The following three examples show how IBM customers have applied these concepts.

A large U.S. pharmaceutical company created an on demand data warehouse around its financial data. The data warehouse is loaded by the trickle feed process throughout the day to provide detailed financial metrics to management. At the same time, this information is published through shared services that can be plugged into processes and applications related to clinical trials. These applications use the data to optimize their activities and control costs within the clinical trials process.

A major U.S. automobile manufacturer created an on demand data warehouse for its sales data. The data warehouse is trickle fed by ETL services to alleviate pressure on the batch window. On the consumption side, delivery services were created that plug valuable vehicle sales data by region and by type back

into operational applications using enterprise application integration (EAI) technology. This information is now used to optimize inventory orders for parts and accessories at the dealers by providing guidance based upon the actual vehicle distribution in that region. This process allows the automaker to help the dealer optimize inventory and improve accessory sales.

A leading international pharmaceutical company created an on demand data warehouse for its pharmacokinetic data produced during pre-clinical testing of new compounds. The data warehouse is loaded by trickle feed during the day, rather than during batch cycles, to speed the flow of information from study to study and provide immediate analytical feedback on study results. The result is a reduction in the time required to conduct pre-clinical trials and move promising compounds into the clinical trials process. This translates to tangible revenue benefit by extending the commercial life of new drugs.

These examples demonstrate cost-saving and revenue-producing opportunities that were capitalized by these companies using on demand data warehousing and closed-loop processing. Each company plans to extend its application of on demand data warehousing into other areas to further improve efficiencies and augment revenue opportunities.

A model for on demand data warehousing enables queries through open, standard services

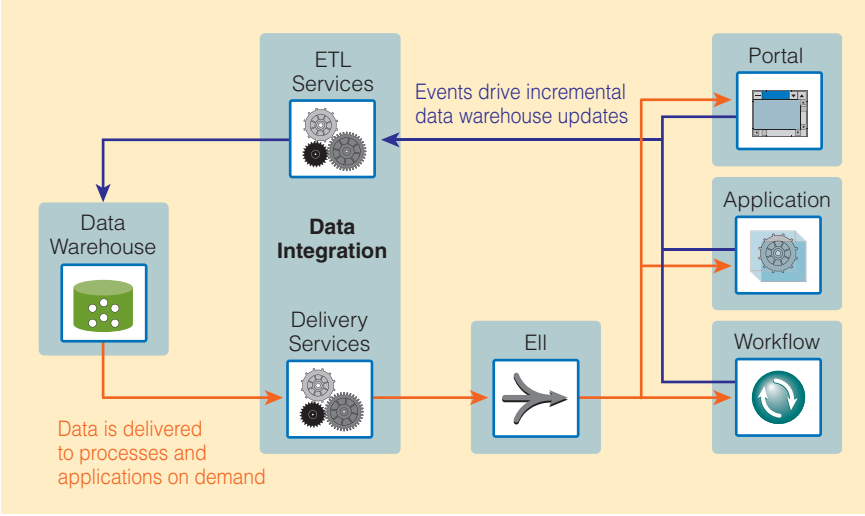
On demand data warehousing creates a service-oriented wrapper around a data warehouse, allowing it to be loaded or queried on demand through open, standard services. The goal is to allow the data warehouse to be plugged directly into business process events, some that trigger the creation of data that needs to be put into the data warehouse and some that consume data from the data warehouse to make real-time decisions. Typically, these processes are found within enterprise applications or created inside information integration products such as IBM® Information Server, as shown in Figure 1.

Creating ETL services

ETL services are simply ETL processes that have been repackaged as services. Since most ETL processes are designed to perform a bulk extract from one or more data sources and a bulk load of data into a data destination, this design often requires some rethinking in the services-oriented context. The ideal solution is to reuse the core logic of an ETL process on an individual message basis. Reusing this logic rather than re-creating it can help to recapture the initial investment, speed time to market and help ensure a single point of maintenance. Essentially, this reuse of logic removes the "E" (extraction) from ETL and replaces it with a message request.

The benefits of creating ETL services are reducing the latency of creating analytical data and linking the warehouse more tightly into enterprise architectures. Specific fact tables can be updated as data changes—driven from events in applications or EAI platforms. The data warehouse can become the basis for closed-loop processing, making the best information available to these processes. For example, each sales order can be loaded into the data warehouse as it is received, providing up-to-the-minute information on sales performance by product. The result is a consistent set of rules applied to the data entering the data warehouse, whether it is created on a message-driven basis or in a batch load.

Figure 1: On demand data warehousing with IBM Information Server



An additional benefit of incremental updates is the reduction of processing required during batch windows. Since much ETL processing happens intra-day, the batch process load is substantially lightened. For companies whose batch windows are becoming pressured, trickle feeding of the data warehouse can offer relief by spreading that processing across a larger portion of the day.

Some ETL tools provide mechanisms for easily repackaging the core logic of ETL processes as shared services. Several factors are important when choosing an ETL technology. It should have the ability to reuse core ETL logic on a message-driven basis and to publish it as a service. It should also be easy to redefine the input and output data formats for the services without affecting the original process, since the format requirements within the service-oriented setting may be substantially different from what is designed into the ETL process. An ETL technology should enable these capabilities without maintaining two sets of logic—one for batch and one for services.

Creating delivery services

Delivery services focus on broadening access to data warehouse information, particularly to other applications and processes. The key to the success of delivery services is to pre-package access in a way that does not force the application developer to understand the complex schema of the data warehouse. Essentially, the people who understand the data warehouse structure build consumable services that are easier for the groups who want to use the data to understand.

Delivery services are created based upon the demands of the business. Specific processes can benefit from specific facts, dimensions and aggregates within the data warehouse. Data delivery services are created by analyzing these specific requirements and packaging that data into shared services. These services can be easily called by applications and processes, which can use the data—when they need it—for closed-loop processing or other purposes. These services can also be called directly by enterprise information integration (EII) products like IBM WebSphere® Information Integrator, which allows developers to assemble data together across multiple sources and include the benefits of caching and ad hoc querying.

The benefits of delivery services are centered on information availability. Applications and operational business processes get access to the best data within the enterprise for making decisions. For example, an inventory allocations process can be supplemented with current and historical information on sales for each product within each region, resulting in a better basis for allocation decisions. In addition, processes get the added value of auditability and historical context for the data. Standards for requirements such as security policies, data quality, transformations or formats can also be consistently enforced in these services. This approach reduces the requirement for spawning new datamarts for every group that wants to use that specific data. By decoupling the information from data sources, companies increase the flexibility of how information is utilized and create consistent, reusable mechanisms to deliver trusted information as a service to the business.

Some data integration technologies provide mechanisms to easily produce delivery services based on data warehouse schemas. Data integration technology should have the ability to easily create data delivery services based on known metadata structures. Additional benefits can be gained from integration technologies that dynamically link both technical and business metadata to the information it handles while making that metadata available to users. Associating metadata to the delivery services greatly facilitates its reuse by providing a common understanding of what the data means to both technical and business users. The metadata will help ensure that the information is correctly applied to the request and will also facilitate control and governance over the data. In addition, data integration technology should be able to produce services that can be consumed by both application and integration infrastructures.

Outlining design considerations

As organizations evaluate data integration technologies, there are several factors to consider when designing a data warehouse.

Dimensions. Since most data warehouses store data in a dimensional structure like a star or snowflake schema, this structure must be considered whenever data is loaded. Data within a dimensional structure can be very sensitive to rapid updates because of the extensive indexing requirement. Organizations that implement near-real-time updates usually handle this issue by partitioning their warehouse into different segments for batch and real-time data and using a lower degree of indexing on the real-time partitions.

Aggregates. Aggregates provide significant insight into the data within a data warehouse by summarizing and deriving data across dimensions. Calculating aggregates in an on demand data warehouse is complicated by splitting real-time and static data partitions. Companies typically calculate them directly within the ETL service and store them in separate transaction-grain aggregate tables to accommodate this issue.

Integration. To be useful in the context of closed-loop processing, the ETL and data delivery services created for an on demand data warehouse must be easily plugged into applications and business processes; that is, they must support the protocols native to these infrastructures. Web services are a good choice since most synchronous application development and integration technologies support them. However, Web services do not inherently support guaranteed delivery and asynchronous processing. Since many business processes require this functionality, other transport protocols like Java™ Messaging Service (JMS) can be considered. JMS provides asynchronous messaging and guaranteed delivery, plus it easily integrates with most integration technologies. Ideally, the flexibility to choose the right protocol for the specific requirements of different applications and processes is the best approach.

Data quality. Failure to cleanse data in real-time partitions introduces the risk that this data will corrupt information integrity. The best way to handle the need for quality is to embed cleansing processes directly into the ETL logic and likewise into the ETL services, allowing the same logic to be used in both batch and real time.

Data security. Because many processes, applications and users may have access to services, data delivery services may need to be controlled based on security credentials.

The batch cycle. Trickle-feed processing can reduce strain on the batch window by spreading ETL processing throughout the day, effectively reducing the amount of work that must be accomplished during the batch cycle. However, when both batch and real-time processing are heavy during concurrent periods, the batch window can be strained. The impact of this can be controlled by throttling real-time processing during the batch window, employing parallel processing to speed up batch processing time or separating batch and on demand ETL activities on different hardware.

Query performance. Poor performance can quickly create a bottleneck for transactional processes and potentially affect the business. Several factors can affect performance: the size of individual data sets within transactions, the frequency of data delivery service requests and the volume of transactions that involve calls to ETL services. These factors can be managed by implementing parallel processing within ETL services when

data sets are large, using data caching together with data delivery services to reduce calls back to the data warehouse, and load balancing the high volume of transactions across ETL servers to remove hardware bottlenecks.

High availability. When the usage model for ETL changes from batch to one where transactional processes rely on ETL and data delivery services, the requirement for high availability becomes vital. New services must be designed to eliminate any single points of failure.

Data warehouses can provide cost efficiencies and revenue opportunities

Data warehouses provide great benefit for strategic decision-making by creating the "best data" within a business. Technical issues have restricted the use of this data to strategic analysis, despite a measurable business benefit for this data within operational systems. Service-oriented architectures offer a strong basis for overcoming these technical issues.

On demand data warehousing focuses on capturing this opportunity by providing a data warehouse that can provide current analytical information directly into business processes—delivering information as a service. Armed with this better and more complete information, processes and applications can make more informed automated decisions. The result is better operational efficiency, lower operational costs and increased revenue opportunities.

On demand data warehousing is enabled through the selection and implementation of the right tools and technologies. The technologies that you choose determine the effectiveness of the effort, particularly since the technical challenges associated with on demand processing and access differ from those associated with batch processing. The data warehouse ultimately must be able to support both, so appropriate design is a critical factor to the success of the project.

When implemented properly, the on demand data warehouse can provide enormous business benefits, both in cost efficiencies and revenue opportunities. Companies that can find and capitalize upon these opportunities can achieve an immediate competitive advantage.

IBM Information Integration Solutions has helped companies to achieve success in their on demand data warehousing efforts by providing the technologies, methodology and best practices for quickly and successfully deploying these projects.

For more information about IBM Information Integration Solutions, please visit our Web site at **ibm.com**/software/data/integration

IBM Software Group
Route 100
Somers, NY 10589
U.S.A.

Produced in the United States of America
September 2006

[1] Agosta, Lou. Data Warehouse Spending Expectations Strong. Forrester Research, July 12, 2004.

[2] McClure, Steve. Data Integration Software Market Assessment. IDC custom study for Ascential Software, September 10, 2004.

[3] Agosta, Lou. Third Generation Data Warehouses Enable Integrated Business Intelligence. Forrester Research, June 14, 2004.

[4] Eckerson, Wayne and Colin White. Evaluating ETL and Data Integration Platforms. TDWI Report Series, 101communications. 2003.

[5] Ibid.

TAKE BACK CONTROL WITH **Information Management**

GC34-2532-01