International Engineering Journal For Research & Development

Vol.2
Issue 6

E-ISSN NO:-2349-0721

Impact factor : 2.25

STUDY OF REAL-TIME DATA WAREHOUSE

Sohail Ansari

Assistant Professor, IT Deptt., S.V.P.C.E.T ,Nagpur Maharashtra, India
mohd.s.ansari@gmail.com

ABSTRACT :

In today's fiercely competitive marketplace, companies have an insatiable need for information. Key to maintaining a competitive advantage is understanding what your customers want, what they need and the manner in which they want to receive your products or services. It is becoming increasingly clear that companies poised to experience the greatest success will be those firms that can effectively leverage their data to meet organizational needs, build solid relationships with stakeholders and above all, meet the demands of today's customers (Schroeck, 2000).The global economy of today demands that organizations adhere to the constantly changing needs of the customer. Additionally, the speed and dynamic nature of business often negates the time required for long-term planning and time-consuming implementations in order to stay ahead. Because of this, organizations must implement solutions that can be deployed quickly and in a cost-effective manner (Zicker, 1998). So, how does an organization meet these ever-changing, complex requirements? An effective real-time business intelligence infrastructure that leverages the power of a data warehouse can deliver value by helping companies enhance their customer experiences. Furthermore, a real-time data warehouse eliminates the data availability gap and enables organizations to concentrate on processing their valuable customer data. By designing a data warehouse with the end user in mind, you multiply your chances of better understanding what your customer needs and what you need to help that customer achieve his or her goals (Haisten, 2000). The ability to access meaningful data in a timely, efficient manner through the use of familiar query and analysis tools is critical to realizing competitive advantages. Equally important is the moving and sharing of data throughout an organization, between departments, offices and business partners. But with the proliferation of mixed-system environments that must somehow be integrated with decision support systems, data marts and warehouses, electronic business solutions and enterprise applications, the challenges increase. When customer information is disjointed and spread across the organization, the challenges can become insurmountable. Data – customer data, financial data, and Internet click-stream data – is a powerful asset provided it can be integrated and utilized to enhance customer experiences. Customers have become more complex and expectations are higher than ever before to meet their needs effectively. Today's businesses are under extreme pressure from both traditional and new rivals. Only those organizations that deliver the best customer experience will thrive. That means improving business management, providing market diversity

and generating competitive advantages. A successful real-time data warehouse can be the silver bullet your organization needs to prosper in the Internet era – that is, if you can avoid the common data warehousing pitfalls. In fact it has been said that data warehousing is e-Business. As we move forward, it is becoming clear that without the support of a Considerations for Building a Real-time Data Warehouse Data Mirror Corporation White Paper Page 2 data warehouse, companies cannot successfully implement their e-Business strategies (Schroeck,2000). The following pages offer an informative approach to evaluating real-time replenishment software for feeding a data warehouse. We will outline real-time data transformation and integration requirements for the most functionally rich data warehouse and highlight how your business can experience positive results quickly to enable you to exceed the ever-changing needs and expectations of your customers.

I. INTRODUCTION

Companies use data warehouses to store information for marketing, sales and manufacturing to help managers run the organization more effectively. The ability to manage and effectively present the volume of data tracked in today's business is the cornerstone of data warehousing. But when the data warehouse is replenished in real-time it empowers users by providing them with the most up-to-date information possible. Almost immediately after the original data is written, that data moves straight from the originating publisher to the data warehouse. Both the before and after image of a record is available in the data warehouse memory, thereby supporting easy and efficient processing for query and analysis at any time. Given the benefits of real-time data warehousing, it is difficult to understand why the "snapshot" copy process has prevailed. Currently, the dominant method of replenishing data warehouses and data marts is to use extraction, transformation and load (ETL) tools that "pull" data from source systems periodically – at the end of a day, week, or month – and provide a "snapshot" of your business data at agiven moment in time. That batch data is then loaded into a data warehouse table. During each cycle, the warehouse table is completely refreshed and the process is repeated no matter whether the data has changed or not. Historically, best practices have been hampered by problems with integrating diverse production systems with the data warehouse. Snapshot copy was deemed "right" because it was next to impossible to get real-time, continuous data warehouse feeds from production systems. As well, since query tools were relatively unsophisticated and complex to debug, it was also difficult to get consistent, reliable results from query analyses if warehouse data was constantly changing. In the Internet era, more people are beginning to realize the limitations that snapshot copy replenishment presents and demand better alternatives. Snapshots do not involve entire database movement but simple captures of parts of database tables; for example, specified columns. As well, not each individual change is made to a record between copy processes. In this light, the snapshot process can be likened to looking at last week's newspaper or using last week's stock market results to trade stock today. The Internet era is about having absolutely current and up-to-date business intelligence information. Data is a perishable commodity: the older it is, the less relevant. Businesses need tools that can provide real-time business intelligence and an absolutely current and comprehensive picture of their organization and their customers – not last week or last month, but right now.

II. COMPONENTS OF REAL-TIME DATA WAREHOUSING

An up-to-the-second view of customer data, once an ideal, is fast becoming a reality for businesses wishing to implement real-time business intelligence solutions. But how does the data warehouse actually operate? An intelligent warehousing solution and framework can commonly be divided into three fundamental tiers with data flows between them. The three layers are Presentation Layer, Architecture Layer, and Middleware Layer. These tiers or layers must be seamlessly integrated and function as one to ensure the immediate success and long-term benefits of a data warehouse.

II.A Presentation Layer

The presentation layer manages the flow of information from the warehouse to the analyst, providing an interface that makes it easier for the analyst to view and work with the data. This layer is where graphical user interface (GUI) tools are most important. Front-end query tools should provide an easy and efficient way to visually represent data for decision making in two or more dimensions. Pattern recognition and analytic algorithms can highlight areas for close human analysis, but in the end humans still have an edge in improvisation, gut feeling and trend forecasting.

Warehousing assists users in the analysis of sales data so they can make informed decisions that have real-time impact on company performance. The presentation layer's ability to store and present multidimensional views or summaries of data is one reason why multidimensional databases and query tools are popular at this level of the warehouse.

II.B ARCHITECTURE LAYER (STRUCTURE, CONTENT/MEANING)

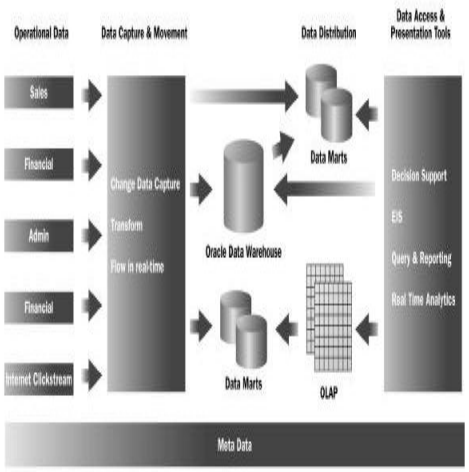
The architecture layer describes the structure of the data in the warehouse. An important component of the architecture layer is flexibility. The level of flexibility is measured in terms of how easy it is for the analyst to break out of the standard representation of information offered by the warehouse in order to do custom analysis. Custom analysis is where semantic thickness becomes important. Semantic thickness is the degree of clear business meaning embedded in both the database structure and the content of the data itself. Field names such as “F001” for customer number and obscure numbers such as “01” to indicate “Backorder” status are considered semantically thin, or ambiguous and difficult to understand. In contrast, field naming standards such as "Customer_Name" containing the full customer name and "Order_Status" containing the complete description "Backorder&” are semantically thick, meaningful and easily understood. In other words, data structure and content must be clear to the analyst at the presentation layer of the data warehouse. The underlying data schema for the warehouse should be simple and easily understood by the end user of the data.

II.C Middleware Layer (Interfaces and Replenishment)

The middleware layer is the glue that holds the data warehouse together. It integrates the data warehouse with production and operational systems. Data needed for warehouse applications often must be copied to and from computers of different types in different locations. Warehousing often implies transformational data integration. Production data needs to be secure and is frequently not in the format needed for

warehousing. Real-time integration and replenishment tools that help businesses deal with the data management issues of implementing a data warehouse can add real value. The rest of this paper will focus on how a real-time integration and replenishment solution – or a capture, transform and flow (CTF) tool can contribute to the simplicity and efficiency of a real-time data warehouse.

III. CAPTURE, TRANSFORM AND FLOW (CTF)



III.A Change Data Capture

Today, more and more businesses using a data warehouse are beginning to realize they cannot achieve point-in-time consistency without continuous, real-time change data capture. There are several techniques used by data integration / replenishment software to move data. Essentially, integration tools either push or pull data on an event driven or polling basis. Push integration is initiated at the source for each subscribed target. This means that as changes occur, they are captured and sent, or "pushed" across to each target. Pull integration is initiated at the target by each subscribed target. In other words, the target system extracts the captured changes and "pulls" them down to the local database. Push integration is more efficient as it can better manage system resources. As the number of targets increases, pull integration becomes resource draining on the source system, especially if that machine is a production machine that may already be overworked. Event driven integration is a technique that involves events at the source initiating capture and transmission of changes. Polling involves a monitoring process that polls the status to initiate capture and application of database changes. Event driven integration conserves system resources as integration only occurs after preset events whereas polling requires continuous resource utilization by a monitoring utility. But in order to compete with an information-driven Internet era, organizations must employ solutions that offer the option of updating databases as incremental changes occur, reflecting those changes to subscribed systems. With advanced CTF solutions, every time an add, change or delete occurs in the production environment, it is automatically captured and integrated or "pushed" in real-time to the data warehouse. By significantly reducing batch window requirements and instead making incremental updates, users regain computing time once lost. Beyond real-time integration, change data capture can also be done periodically. Data can be captured and then stored until a predetermined integration time. For example, an organization may schedule its refreshes of full tables or changes to tables to be integrated hourly or nightly. Only data that has changed since the previous integration needs to be transformed and transported to the subscriber. The data warehouse can therefore be kept current and consistent with the source databases.

III.B Transformation

The way companies think about data, and the way it is represented in databases, has changed

significantly over time. Obscure naming conventions, dissimilar coding for the same item (e.g. number representation as well as character based codes), and separate architectures are all commonplace. Software that can transform data across multiple computing environments and databases can remedy these problems while consolidating the information in your data warehouse. Companies are beginning to realize the benefits of sharing data between enterprise resource plannin(ERP) systems and relational data stores housed in databases including Oracle, Sybase, DB2/UDB, and Microsoft SQL Server. The problem is that ERP systems use proprietary data structures that need to be cleansed and reformatted to fit conventional database architectures. Rows and columns may have to be split or merged depending on the database format. For example, an ERP system may require that "Zip Code" and "State" is part of the same column while your company's data structure may have the two columns separated. Similarly, your company may have "Product Type" and "Model Number" in an inventory database as one column, whereas the ERP system requires them to be split. Data transformation and integration software can accommodate these requirements in order to make your data – and consequently, your data warehouse – more useful and meaningful to users.

Publisher	Transformation process	Subscriber
Smith, M. $10.00 U.S. 20" 1B	Two-field Consolidation and Rearrangement Euro Conversion Unit Conversion Value Substitution	Mary Smith 9.333 Euros 50.8 cm In Stock

Figure 2: Sample data transformations.

Other applications of data transformation software include changing data representation (U.S. dollars converted to British Sterling, metric to standard, character columns to numeric, abbreviations to full text, number codes to text), visualization (aggregate, consolidate, summarize data values), and preparation for loading multidimensional databases. Transformational data integration software can conduct individual tasks such as translating values, deriving new calculated fields, joining tables at source, converting date fields, and reformatting field sizes, table names and data types. All of these functions allow for code conversion, removal of ambiguity and confusion associated with data, standardization, measurement conversions, and consolidating dissimilar data structures for data consistency.

III.C Flow

This refers to replenishing the feed of transformed data in real-time from multiple operational systems to one or more subscriber systems. Whether a data warehouse or several data marts, the flow process is a smooth, continuous stream of bits of information as opposed to the batch loading of data performed by ETL tools.

IV. CURRENT ISSUES IN REAL-TIME DATA WAREHOUSING

IV.A Time to Market

In today's competitive economy, time to market means everything for data warehousing projects. Why do so many data warehousing projects fall behind schedule or even fail? One major reason for data warehouse failures is that many data warehouses are populated with operational data that is poor in quality. Raw operational data needs to be selected, filtered and transformed before consolidating it in warehouse tables for business

intelligence purposes. Operational data is typically stored in multiple tables and consists of codes and abbreviations, making it difficult to access for decision support. A simple invoice, for example, may contain data from over a dozen different tables. More often than not, operational systems also contain inconsistent data. An inventory system may store data as "Male" and "Female," while a system used by sales stores the same information as "M" and "F." Given the circumstances, most would agree that unleashing end users on a data warehouse without first cleansing or transforming the raw transactional data that populates the warehouse would be a poor idea. Data quality can also seriously affect data warehouse performance. With data warehouses and data marts, the analogy is: "garbage in, garbage out." You won't find the trends and relationships you're looking for in the data unless you feed your query and analysis tools with the right information. A little knowledge, or the wrong knowledge, can be a dangerous thing. It can give you an incomplete or flawed picture of your customer or your business. This problem of data quality can be avoided by selecting a replenishment solution that offers advanced capture, transform and flow (CTF) technology. CTF tools enable you to capture raw data from multiple operational databases and flow the data in real-time into data warehouse tables while transforming the data on-the-fly into meaningful information. Users are empowered through the means to translate values, derive new calculated fields, reformat field sizes, table names and data types. CTF tools help accelerate time to market while adding value to business intelligence information by keeping the data clean, current and in a format conducive to query and analysis. Through a combination of best practices and best-of-breed solutions such as capture, transform and flow tools, companies can reasonably expect to have end users querying the data warehouse within a short time frame. In addition, data warehouse projects can quickly get out of hand in terms of size and scope because it is difficult for IT managers to deny requests from users and executives to change or expand the design. This can make it difficult or even impossible to deliver BI systems on time and under budget. Large data warehouses need to integrate data from across the organization—different hardware, operating systems, databases and applications—and integration efforts can be time-consuming and costly. A successful BI application takes good advance planning, organization-wide sponsorship and input, and doing your homework to choose the right tools and the right vendors to partner with. The remainder of this white paper will explore other issues that should be considered when the goal is to implement a data warehouse project on schedule and on budget.

CONCLUSION AND FUTURE SCOPE

When deciding how to build a data warehouse, there is a tendency to get caught up in the technology and forget the basics that contribute to its creation. The questions that your business needs to ask are how fast does your company respond to new ideas, new competitive pressures and new opportunities? How can it do great work fast? Can each branch of the organization stay informed about customers, competitors and business trends? The competitive and customer pressures of the new economy have created an insatiable demand for up-to-the-second business information. It is no longer acceptable for businesses to make decisions based on day-old or week-old data. Employees, decision-makers, partners and customers alike need access to information while it is still relevant. Real-time data warehousing essentially allows organizations to

combine long-term planning tactics with up-to-the minute decision-making to ensure enhanced customer experiences, influence customer loyalty and increase the bottom line (Brobst and Venkatesa, 1999). The Internet has significantly changed the scope and expectations for data warehousing implementations. Developing a robust, scalable data warehouse with consistent, real-time data and the ability to answer all user-information needs is not an easy goal to achieve. Keeping data current is one of the most difficult and time-consuming challenges in managing data warehouses and data marts. But if organizations fail to take advantage of real-time data capabilities for business intelligence, they will lose the opportunity to respond quickly to changing market trends. These businesses will be less agile and will have difficulty meeting the customers' rising expectations for 24/7 service. Predictably, they will lose market share to truly customer-centric businesses that do invest in real-time data warehousing. By understanding the relevant issues and staying informed of current practices in data warehouse development, your organization can harness the power of real-time business intelligence and attain a cost-efficient, profitable computing environment. In a very short time, your business could be well on its way toward meeting customer needs and generating a competitive edge in today's competitive economy.

References

[1] IEEE Computer. Special Issue on Heterogeneous Distributed Database Systems, 24(12), December 1991.

[2] J. Gray, A. Bosworth, A. Layman, and H. Pirahesh.Data Cube: A relational operator generalizing group-by, cross-tabs and sub-totals. IEEE Transactions on Knowledge and Data Engineering, 1995. To appear.

[3] A. Gupta, H.V. Jagadish, and I.S. Mumick. Data integration using self-maintainable views. Technical memorandum, AT&T Bell Laboratories, November 1994.

[4] A. Gupta and I.S. Mumick. Maintenance of materialized views: Problems, techniques, and applications. IEEE Data Engineering Bulletin, Special Issue on Materialized Views and Data Warehousing, 18(2):3{18, June 1995.

[5] John Vandermay Vice President of Development DataMirror Corporation : Considerations for Building a Real-time Data Warehouse 2001.

[6] Brobst, Stephen A. and Venkatesa, AVR, "What is Active Data Warehousing?," Abridged from "Active Warehousing," Teradata Review Spring 1999 <http://www.ncr.com/repository/articles/data_warehousing/what_is_active_dw.htm>.

[7] Haisten, Michael, "Real-Time Data Warehouse: What is Real Time about Real-Time Data Warehousing?," DM Review Online August 2000 <http://www.dmreview.com/master_sponsor.cfm?NavID=68&EdID=2589>.

[8] Schroeck, Michael J., "E-Analytics—The Next Generation of Data Warehousing," DM Review August2000 <http://www.dmreview.com/master.cfm?NavID=55&EdID=2551>.

[9] Zicker, John, "Real-Time Data Warehousing," DM Review March 1998 <http://www.dmreview.com/master_sponsor.cfm?NavID=55&EdID=676>.

[10] Jennifer Widom, "4th Int'l Conference on Information and Knowledge Management (CIKM), Nov. 1995"