International Journal of Computer (IJC)
ISSN 2307-4523 (Print & Online)
https://ijcjournal.org/index.php/InternationalJournalOfComputer/index

Construction of Real Time Data Warehouse

Sohail Irfan*

Solution Architect/Software Engineer, Saudi Electricity Company, Riyadh, Saudi Arabia
Email: Sohailch40@live.com

Abstract

Data warehouse is a copy of transaction data specifically structured for querying, analysis and reporting Data warehouse is a database application that actually store and collect the data from any particular business domain for decision support system. There are different ways to implement and design the data warehouse. Some processes are involved the data extraction, transformation and loading. Integration of the data from various sources in to the data warehouse is major concept. So, while the design of the data warehouse such that to entertain all these processes accurately then we can guarantee the data purity. Approach of recent times is becoming very much famous now days that is real time data warehouse. Real time data warehouse actually load the data from the transactional and operational data stores when in real time. As soon the data is coming up in the external data source it will appear into the real time data warehouse so this paper will cater the discussion of structure of real time data warehouse. Real time data warehouse approach has major deficit of extraction and loading process. It could be very efficient source of decision support system if we can eliminate the deficiencies.

Keywords: dimensionality; extraction; integration; cleansing; optimization; DataMart; transactional; analytical.

1. Introduction

Until recently, there were few viable tools to provide real-time data warehousing nor an absolutely current picture of an organization's business and customer. But if businesses have survived without continuous, asynchronous, multi-point delivery of data in the past, why then would such solutions become so critical to business today. 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.

Received: *7/4/2024*
Accepted: *9/4/2024*
Published: *9/14/2024*

---

* Corresponding author.

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 a given 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.

2. Architecture of Real Time Data Warehouse

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 [4].

- Presentation layer
- Architecture layer
- Middleware layer

2.1. 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 [2,13].

2.2. Architecture Layer

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 [5].

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 as shown in Figure 1.

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 [6].

2.3. Middleware Layer

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.

3. Efficient Business Intelligence Replenishment

Before the Internet existed, only a few dozen users – usually hardcore data analysts – accessed most data warehouses. But the democratization of information access over the last few years has created new challenges. Web-based architectures must routinely handle large volumes of concurrent requests while maintaining consistent query response times, and must scale seamlessly as the data volume and number of users grows over time. In addition, data warehouses need to remain available 24 hours a day because the web makes it cost effective for global corporations to provide data access capabilities to end users located anywhere in the world. This is where data replenishment and resiliency tools come in to provide real-time access [1].

Even if your organization does not wish to implement real- time data warehousing, it is still important to consider how efficient your current extract tool is at replenishing the data warehouse. Most ETL tools are batch processors, not real-time engines. This method can be both resource intensive and time consuming. The larger the data warehouse, the longer it takes to replenish with this method. In some cases, the volume of data being loaded into the warehouse begins to exceed the batch window allotted for it. This process is inherently

inefficient. In a typical environment where 20 per cent of the data on productions systems is changing every week or month, why would you choose to refresh the entire data warehouse every week or month? Why send 20 gig of data when you could send only 20 per cent of 20 gigs – or 4 gigs – that has changed since the data warehouse was last replenished.

Due to the growth in business and the related increases in data, companies may find it difficult to fit its "batch job" into a periodic time window of eight hours, and may as a result cut into normal usage hours. In contrast to standard ETL tools, consider an advanced CTF solution that instead captures, transforms and flows data in real-time into an efficient, continuously replenished data warehouse [10].

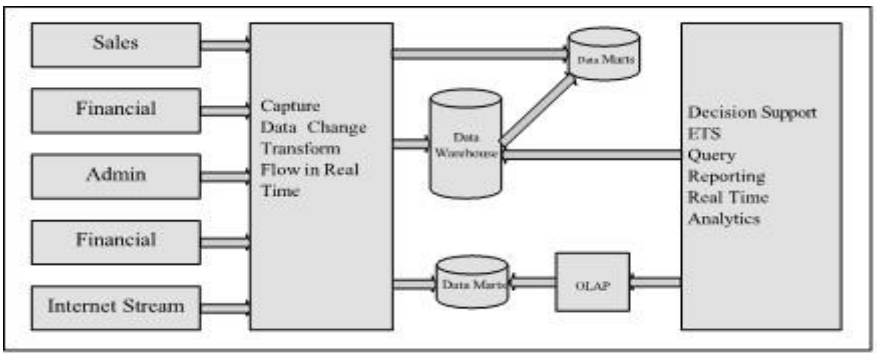

Figure 1: Real time data warehouse

4. Capture, Transform and Flow (CTF)

While it is important to select a database, hardware and operating system platform is must, then selection of an ETL tool is extremely considerable, but it's not a must. While evaluation ETL tools, it should have following characteristics so that it may pay back exactly what we are looking for:

4.1. 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 extracted integration turns into resource demanding on the source system, especially if that system is a production system that might be overworked before. Event driven integration implies a technique that entails events at the source starting capture and transmission of changes. Polling includes a process monitoring, which polls the status to initiate capture and application of database changes. Event driven integration protects system resources as entailing integration occurs after preset events whereas polling requires continuous resource consumption by a monitoring tool [15].

4.2. Transformation

It is the process and the way the organizations keep the data, and the way it is stored in databases, has changed thoroughly over time span. Uncertain naming conventions, different coding for the similar item (e.g. numeric representation as well as string-based codes), and separate architectural designs are all dissimilar. An application that can transform data from number of computing environments and databases can get you to rid of these problems whereas it might consolidate the information in the data warehouse. Companies are beginning to realize the benefits of distributing data between enterprise resource planning (ERP) systems and relational data stores housed in databases.

Table 1: Transformation

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

The dilemma is that ERP systems use proprietary data structures that need to be cleaned and reformatted in order to be compatible with conventional database architectures. Rows and columns may have to be split or merged depending on the database format. For instance, an ERP system may require that “ABC information” and “XYZ information” is part of the same column while your company’s data structure may have the two columns separated. Data transformation and integration software can accommodate these requirements to make your data more useful and meaningful to users that will consequently make your data warehouse more practical.

4.3. Flow

It refers to refilling the feed of altered data in real-time from several operational systems to one or more subscriber systems. The flow process is a smooth, continuous stream of bits of information as opposed to the batch loading of data performed by ETL tools either for a data warehouse or several data marts [12].

5. Evaluation Criteria for CTF Solution

More robust and powerful capture, transform and flow (CTF) software exists that can facilitate the real time delivery of meaningful information to subscribed systems, movement among mixed platforms and databases, and the selecting and filtering of the data broadcasted. In the changing Internet era, companies should strive to select real-time CTF solutions that propose the following features and facilities.

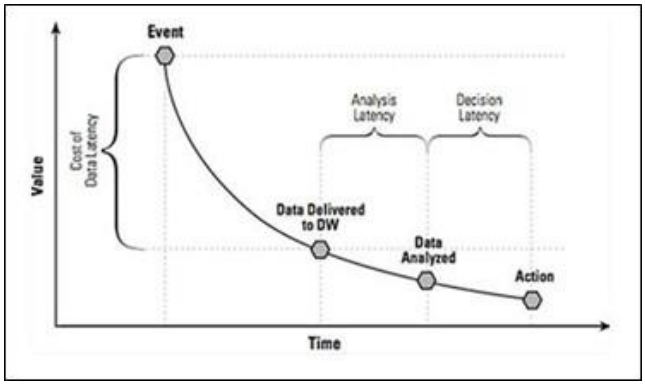

Figure 2: Real time data warehouse analysis graph

For a business executive it is time taken job to get customized report such as - What are the most products have been sold to the customers between the ages 16 to 35 to assess the popularity of that particular product.

5.1. Selectivity

Business solutions like data marts and warehouses require the ability to select and filter which data is moved. Well- organized CTF software offers an array of features including transformation capabilities and selection functions in addition to data enhancement and built-in data filtering. This allows source data to be selectively filtered by row and/or column before refilling the data warehouse.

5.2. Support for Heterogeneous Environments

Because of changes in technology, corporate mergers, and purchasers, most organizations now bear multiple computing platforms and databases, each storing separate pockets of information. This information may be entirely incompatible to the next.

Most data integration solutions focus on moving data only between uniform database software and systems. Some integration tools, however, are capable of moving data among a wide range of databases and systems. Number of vendors currently offers transformational data integration tools to consolidate and synchronize heterogeneous data into a data mart or warehouse or to integrate data from legacy production systems with data from newer web-based systems. However, many of these tools still do not offer synchronized abilities [9].

5.3. Ease of Administration

Some integration software can be uncomfortable to install and set up. A serious consideration when evaluating transformational data integration tools is the work required for setup and implementation of the software. IT staff and end users seek an “out-of-the-box” experience from data warehousing utilities. Organizations should ensure that no programming changes to existing databases and applications are required. A good thing to keep in mind is that the solution software is meant to avoid resource-intensive, costly custom extract programming, and time-consuming. Vendors that offer portable graphical user interfaces and tools for web or wireless management

of integration software can empower businesses and ease the strain of constantly being on-site to control integration status and inspection. Organizations can therefore focus on driving their business.

5.4. Meta-data Management Capabilities

Fundamentally, examining meta-data enhances the end user's understanding of the data they are using. It can also facilitate valuable "what if" analysis on the impact of changing other elements and data schemas. For administrators, meta-data helps them ensure data precision, consistency and integrity [7].

During data transport, advanced capture, transform and flow (CTF) replacement solutions store metadata in tables located in the subscriber and/or publisher database. This is an attractive feature to companies wanting to share meta-data among diverse databases and applications. Most databases and tools manage meta-data in a different way. They store the meta-data in distinct formats and use proprietary meta-data to perform definite tasks. An open CTF solution allows organizations to distribute meta-data in different formats using published industry principles.

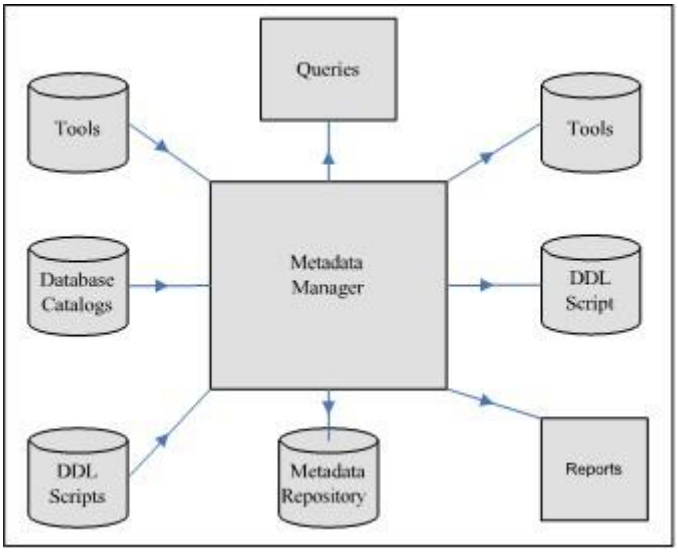

Figure 3: Meta data manager

Using CTF technology based on open standards such as the Open information Model (OIM), Meta-data Coalition Standard for Meta-data Integration or XML Interchange Format (XIF), meta-data can be easily integrated to another repository. This deals with the challenge of standardizing meta-data. Without this functionality, companies often must remedy to custom development of a tool capable of entering meta-data in a variety of formats e.g., a development task that often proves time-consuming, difficult to accomplish, and may negatively impact time to market for the business intelligence solution [14].

5.5. Scalability

Scalability refers to the ability of a computer system or a database to operate proficiently with significant data. While it is possible to forecast some of the future information needs of an organization, a large portion will be

erratic. In the Internet era, it is certain that as business environments change, so too will the kinds of decisions that need to be made and the information that manipulates them. Therefore, it is required for business needs to think in advance and search for a scalable solution when looking for an efficient data integration tool. Organizations should not look at data warehouse development as having a beginning, middle and end, but as a continuous process that must evolve as the organization grows. Failing to do so will result in the loss of information necessary for future competitive advantage and strategic decision making.

6. Current issues in real-time data warehousing

Data refreshing in data warehouses normally related with data loading in data ware house as these are not similar things both are separate process and both have their own importance. While Data ware house start to import the data into it start the process of data loading initially. Further many of other routines and procedures are pre-written into the data ware house architecture so those needs to be started.

These sorts of routines and processes such as stored procedures, trigger, functions, views, etc. (database units) are the part of data stimulating into the data ware house. This data stimulant is most significant in data retrieval time, result oriented outputs and run time queries optimization. All these functionalities are to be pre-written to sum up the loaded data and aggregation on the basis of most recurrent queries to data ware house [8].

6.1. Time to Market

In today's competitive economy era, time to market replicates everything for data warehousing projects. Many data warehousing projects fall behind schedule or even fail! One main reason for data warehouse failures is that many data warehouses are populated with operational data that is deprived in eminence. Raw operational data needs to be filtered, transformed, and selected before consolidating it in warehouse tables for business intelligence rationale. Operational data is typically stored in multiple tables and consists of codes and abbreviations so that it can be tricky to access it for decision support. For example, q simple invoice, may contain data from over a dozen different tables. Operational systems may 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 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 [3].

6.2. Buy versus Build

Despite the buy-versus-make recommendations of most data warehousing analysts, some companies still choose custom coding to handle the replenishment and transformation of data. In this situation, businesses can write customized programs to integrate data to the data warehouse. However, given the productivity gains of using a mapping-based integration tool over scripting, it is difficult to justify allocating the time and resources required for developing custom applications. Mapping-based tools involve built-in administration utilities that provide quick and easy definition of the integration process. Changes in the integration schema are simply re-mapped using the same front-end that defined the original process. Some companies who choose to bring the task in-house and custom code extraction routines soon discover that it is a difficult problem with its own challenges. Why allocate a programmer to do punishing extract programming when your organization can use an easy-to-implement mapping-based tool? Package solutions can provide businesses with many advantages including flexibility, scalability, and upgrade support for the latest database versions upon their release.

6.3. Aggregation

Aggregation refers to the gathering of information in separate sets from two or more sources. Often, this data is stored in a data warehouse in a summarized form. For example, an organization may wish to summarize the data by various time periods. Aggregates are used for two primary reasons. One is to save storage space; data warehouses can get very large, and the use of aggregates greatly reduces the space needed to store data. The second reason is to improve the performance of business intelligence tools. When queries run faster, they take up less processing time and the users get their information back much more quickly. However, there is also a negative side to data aggregation. The process can result in the loss of time sensitive linear data.

If a business is trying to compile a complete customer profile in order to understand its customers, aggregation is not beneficial. Some data warehouses store both detailed information and aggregated information. This approach may take up even more space, but gives users the possibility of looking at all details while still having good query performance when looking at summaries. As well, user exit support exists today in some software that facilitates aggregation while also providing the ability to create a complete customer profile.

7. Conclusion

Real time data warehouse is an up-to-the-second view of operational data, once an ideal, is fast becoming a reality for businesses of an organization wishing to implement real-time business intelligence solutions. Real time warehouse can cater all the process to suffice the data warehouse requirements. Practically this approach is the best among other implementation approaches of data warehouse but as it seeks real time data from the operational data stores it may have major challenge while retrieving, extraction, transformation and loading

process are being executed.Real time data warehouse is the approach that can provide accurate information for decision making as we know the importance of critical decisions in an organization, these can direct effect on business and net effects as whole. So only the real time data warehouse is the way that can provide real picture of the transactional data and can have significant effect on business decisions.

References

[1] B. H¨usemann, J. Lechtenb¨orger, G. Vossen, "Conceptual data warehouse modeling," Proc. DMDW 2000.

[2] W. Lehner, J. Albrecht, H. Wedekind, "Normal Forms for Multidimensional Databases," Proc. 10th SSDBM 1998, 63–72.

[3] G. Ozsoyoglu, D. A. Singer, and S. S. Chung, "Anti-tamper databases: Querying encrypted databases," in Proceedings of the 17th Annual IFIP WG 11.3 Working Conference on Database and Applications Security, Estes Park, Colorado, Aug. 4-6 2003.

[4] Tony R. Sahama, Peter R. Croll, "A data warehouse architecture for clinical data warehousing", ACM International Conference Proceeding Series; Vol. 249, Queensland University of Technology, Brisbane, Queensland, 2007.

[5] Mohammad Rifaie, Keivan Kianmehr, Reda Alhajj, Mick J. Ridley, "Data modelling for effective data warehouse architecture and design", International Journal of Information and Decision Sciences 2009 - Vol. 1, No.3 pp. 282 - 300, ' School of Informatics, Bradford University, West Yorkshire, UK, 2009.

[6] Khurram Shahzad, "Semi-star schema for operational and analytical requirements of SMEs", International Journal of Management and Decision Making 2009 - Vol. 10, No.1/2 pp. 33 - 52, Department of Computer and Systems Sciences, Royal Institute of Technology (KTH)/Stockholm University, Forum 100, SE 164 40 Kista, Stockholm, Sweden, 2009.

[7] V.Breazu-Tannen and R. Subrahmanyam. "Logical and computational aspects of programming with Sets/Bags/Lists," In LNCS 510: Proceedings of 18th International Colloquium on Automata, Languages, and Programming, Madrid, Spain, July 1991, pages 60{75. Springer Verlag, 1991.

[8] William H. Inmon, "Building the Data Warehouse", John Wiley & Sons, Inc., New York, NY, 2005.

[9] Ronald Gage Allan, "Data Models for a Registrar's Data Mart", Business Intelligence Journal, Office of Student Financial Services Georgetown University, Washington, D.C. 2005.

[10] Vasant Dhar, Roger Stein, seven methods for transforming corporate data into business intelligence, Prentice-Hall, Inc., Upper Saddle River, NJ, 1997.

[11] Friedman, J. M. "Data Mining and Statistics: What's The Connections?" 29th Symposium on the Interface in Data Mining and the analysis of large data sets, Houston, TX. 1997.

[12] Swapnil Gorhe, "ETL in Near-Real Time Environment: Challenges and Opportunities.

[13] G. Graefe. "Query evaluation techniques for large databases." ACM Computing Surveys, 25(2):73–170, 1993.

[14] Senda Bouaziz1, Ahlem Nabli1, Faiez Gargouri "From Traditional Data Warehouse To Real Time Data Warehouse" Sfax University, Faculty of Sciences, BP 1171, Tunisia, MIRACL.

[15] Swapnil Gorhe. "ETL in Near-Real Time Environment: Challenges and Opportunities" Computers and Information Science Auckland University of Technology Auckland, New Zealand.