International Journal of Engineering Innovation & Research
Volume 1, Issue 2, ISSN : 2277 – 5668

Reducing ETL Load Times by a New Data Integration Approach for Real-time Business Intelligence

Darshan M. Tank
Department of Information Technology,
L.E.College, Morbi-363642, India
dmtank@gmail.com

***Abstract*—Reducing business latency is essential in today's competitive and demanding environments. This means responding immediately to new information as it arrives, and having the right information in time to make the best decision. Integrating, processing and delivering results in real-time is a huge challenge, particularly as data volumes continue to increase dramatically and sources of data are ever more distributed and varied.**

The decision making process in traditional data warehouse environments is often delayed because data cannot be propagated from the source system to the data warehouse in time. The typical update patterns for traditional data warehouses on an overnight or even weekly basis increase this propagation delay. Keeping data current by minimizing the latency from when data is captured until it is available to decision makers in this context is a difficult task. A real-time data warehouse aims at decreasing the time it takes to make business decisions and tries to attain zero latency between the cause and effect of a business decision.

An ETL process that periodically copies a snapshot of the entire source consumes too much time and resources. Alternate approaches that include timestamp columns, triggers, or complex queries often hurt performance and increase complexity. What is needed is a reliable stream of change data that is structured so that it can easily be applied by consumers to target representations of the data.

To keep up with the market competition, there is increased need to minimize ETL load times. In this paper I have introduced an approach for data integration that deals with reducing ETL load times.

***Key words*— Business Intelligence, Dynamic Warehouse, Real-time Data Integration, Change Data Capture**

I. Introduction

The amount of information available to large-scale enterprises is growing rapidly. New information is being generated continuously by operational systems. In order to support efficient analysis and mining of such diverse, distributed information, a data warehouse collects data from multiple, heterogeneous sources and stores integrated information in a central repository. The data warehouse needs to be updated periodically to reflect source data updates [1]. The operational source systems collect data from real-world events captured by computer systems. The observation of these real-world events is characterized by a propagation delay. The update patterns (daily, weekly) for data warehouses and the data integration process (extract-transform-load) result in increased propagation delays.

Traditionally, there is no real-time connection between a data warehouse and its data sources, because the write-once and read-many decision support characteristics would conflict with the continuous update workload of operational systems and result in poor response time. Existing models lack built-in mechanisms for handling change and time.

The design of an active data warehouse has to deal with two sorts of propagation delays in data warehouse environments.

(1) Delays in capturing real-world events by operational systems, and

(2) Delays in loading and integrating data into the data warehouse.

The aim of our proposed approach is to shrink ETL load times by mean of CDC (Change Data Capture) technique and thereby delivering greater value from information. It also enables meaningful analyses across time even when data changes under the source system.

II. Latency In Business Intelligence

The business value of an action decreases with the amount of time elapses from the occurrence of the event to taking action. However, the data about that transaction is stored within the warehouse environment only after some time-windows. Afterwards, the data is analyzed, packaged, and delivered to the user-application. This process also took time to be accomplished [4]. Therefore, only after a time-window, the decision based on these analysis results and the relevant action can be performed.

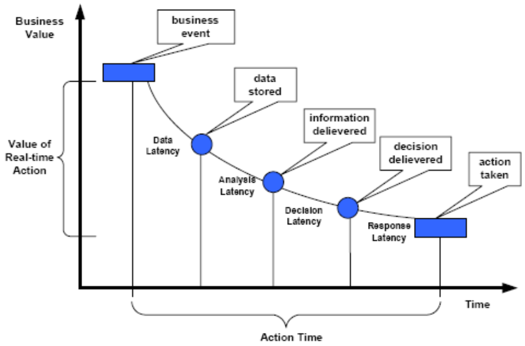

Fig. 1 Business value and action time

The end-to-end time or elapsed time required to respond by taking action in response to the business transaction in an intelligent manner is called action time and can be regarded as the latency of an action. Action time comprises four components namely data latency, analysis latency, decision latency and response latency. Data latency is the time from the occurrence of the business event until the data is stored and ready for analysis. The time from data being available for analysis to the time when information is generated out of it is called analysis latency. Decision latency is the time it takes from the delivery of the information to selecting a strategy in order to change the business environment [5]. Response latency is the time needed to take an action based on the decision made and to monitor its outcome.

III. Real Time Business Intelligence

In the competitive world of business, real-time information is fast becoming a requirement. The goal of business intelligence (BI) systems is to enable better, more informed, and faster decisions [12]. BI can help with the critical issues of a company, such as finding areas with the best growth opportunities, understanding competition, discovering the major profit and loss areas, recognizing trends in customer behavior, determining their key performance indicators, and changing business processes to increase productivity.

Many essential operational decisions need some actual yet integrated and subject-oriented data in or near real-time. However, the direct real-time operational/tactical decision support is not achieved by traditional Business Intelligence Systems [13]. These types of analytical applications are generally completely disconnected from operational IT systems. The decisions are executed by communicating them as a command or suggestion to humans, thus always cause latency. The real-time analysis requirements demand a set of service levels that go beyond what is covered by a traditional Business Intelligence System.

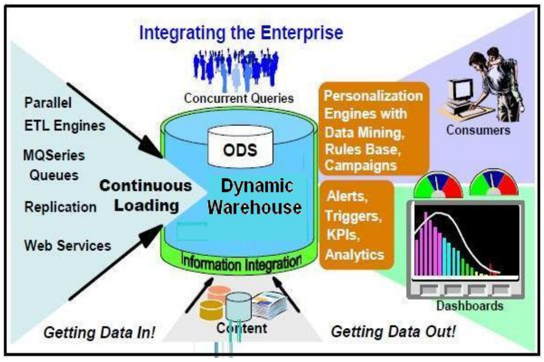

Figure. 2 Real-time Business Intelligence

Fig. 2 shows an overall view of real-time BI. On one side, it shows the various techniques for getting data into the data warehouse from the various data sources and integrating it and on the other side, out data provides the information that is current, real-time, or near real-time and enables management to make more informed decisions.

Real-time business intelligence is having access to information about business actions as soon after the fact as is justifiable based on the requirements. This enables access to the data for analysis and its input to the management business decision-making process soon enough to satisfy the requirement.

IV. Dynamic Warehouse

Dynamic or Real-Time Data Warehousing (RTDW) is referring to the technical aspects that timely perform automatic updates in a Data Warehouse. It implies that any data change occurring in a source system is automatically and instantaneously reflected into the Data Warehouse. All changes in the Data Warehousing environment take place simultaneously with the change in the source system. RTDW concepts include physical modifications to the database schema and the database environment, movement of data across the enterprise, ETL processes, and modification of downstream processes, alerts, creation of extracts, cubes and data marts [2].

Real-time Data Warehouse delivers the right information to the right people just in time. Many essential operational decisions (e.g. promotion effectiveness, customer retention, key account information) need some actual yet integrated and subject-oriented data in or near real-time. However, the direct real-time operational or tactical decision support is not achieved by traditional Business Intelligence Systems. These types of analytical applications are generally completely disconnected from operational IT systems. The decisions are executed by communicating them as a command or suggestion to humans, thus always cause latency. The real-time analysis requirements demand a set of service levels like data freshness, continuous data integration, analytical environments, active decision engines, adaptive platform for the event stream processing that go beyond a traditional Business Intelligence System [3].

Dynamic warehousing is part of the next generation of technology that enables organizations to gain more business insight and deliver relevant information on demand. It enables businesses to provide a more complete and accurate picture to users at any given point in time. Traditional data warehouses can make it difficult to keep up with today's fast-paced environments, dynamic warehousing delivers immediate and integrated information.

V. Design Considerations For RTDW

The design of an RTDW has to consider technical aspects: scalability, high availability, frequent (i.e. just-in-time or continuously) data loading, mixed workload, etc. as well as the integration of active mechanisms which deal with the two sorts of propagation delays in Data Warehouse environments [3]:

1. Delays in capturing real world events by the operational systems and

2. Delays in loading and integrating data into the Data Warehouse. The business requirements for RTDW are

1. Performance
   - Within seconds
2. Scalability
   - Support for large data volumes, mixed workloads and concurrent users
3. Availability
   - 7 **X** 24 **X** 365
4. Data Freshness
   - Accurate, up to the minute data

VI. ETL Process

Extract Transform Load (ETL) is a common terminology used in data warehousing which stands for extracting data from source systems, transforming the data according to the business rules and loading to the target data warehouse. ETL is a process in data warehousing responsible for pulling data out of the source systems and placing it into a data warehouse.

ETL systems move data from OLTP systems to a data warehouse, but they can also be used to move data from one data warehouse to another. A heterogeneous architecture for an ETL system is one that extracts data from multiple sources. The complexity of this architecture arises from the fact that data from more than one source must be merged, rather than from the fact that data may be formatted differently in the different sources [6].

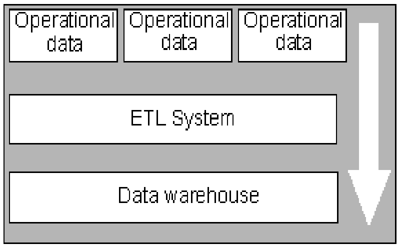

Fig. 3 ETL System Architecture

The ETL process is not a one-time event; new data is added to a data warehouse periodically. Typical periodicity may be monthly, weekly, daily, or even hourly, depending on the purpose of the data warehouse and the type of business it serves.

VII. Problems With The Current ETL System

Traditionally, ETL processes run on a periodic basis (weekly, daily). With the increasing popularity of data warehouses and data marts, the ability to refresh data in a timely fashion is more important than ever. Current ETL systems will completely rebuild the data warehouse periodically to ensure that information used for reporting was current. As the data warehouse increases in complexity and the demand for more up-to-the-minute data increases, the possibility of maintaining the data warehouse in this fashion becomes intractable [9]. The following are the weaknesses of the current ETL system.

1. Current ETL process is inefficient when the data under the source system is not changed frequently.
2. Cost of recompilation can be expensive since the degree of modification to base tables or relations is normally small.
3. It increases response time (latency), network traffic, wasteful to CPU or memory utilization.
4. It maximizes resource requirements and bulk transfer.
5. It is intrusive to the source databases.

VIII. Proposed Approach - CDC (Change Data Capture)

To bring changed data across from source systems into your data warehouse instead of loading in all of the source data and doing a complete refresh, you would normally look for columns in your source data or source files to indicate the creation and modified date of a row of data. Your process would then load in only those rows of data that were new or modified since your last load date. This mechanism enabled automatic feeds of new or changed database records through to your data warehouse [7].

Change data capture is an approach to data integration, based on the identification, capture, and delivery of only the changes made to operational/transactional data systems. By processing only the changes, CDC makes the data integration, and more specifically the 'Extract' part of the ETL process more efficient. When done correctly, it also reduces the 'latency' between the time a change occurs in the source systems and the time the same change is made available to the business user in the data warehouse.

Next generation Data Integration and ETL tools need to support Change Data Capture, a technology that enables to identify, capture, and move only the changes made to enterprise data sources. No longer can the entire source data be moved. Implementing CDC makes data and information integration in real-time significantly more efficient, and delivers data at the right-time [7].

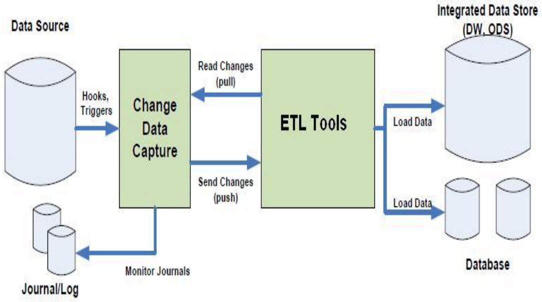

Fig. 4 Working of CDC in conjunction with ETL tools

A common case for using CDC is in conjunction with ETL tools for faster and more efficient data extract in data warehouse implementations. A key goal of CDC is to improve efficiency by reducing the amount of data that needs to be processed to a minimum [14]. Therefore if the business requirements are for only certain changes to be captured, then it would be wasteful to transfer all changes. The most advanced CDC solutions therefore provide filters that reduce the amount of information transferred, again minimizing resource requirements and maximizing speed and efficiency.

A. CDC Methodologies

System developers can set up CDC mechanisms in a number of ways as mentioned below.

1. Timestamps on rows
2. Version Numbers on rows
3. Status indicators on rows
4. Time/Version/Status on rows
5. Triggers on tables
6. Transaction log files on databases

IX. Conclusion

This paper proposes a new data integration approach; one that identifies the differences or changes from operational systems and transferring them only. It offers an effective solution to the challenge of efficiently performing incremental loads from data source. An ETL application incrementally loads change data from operational systems to a data warehouse or data mart.

It will eliminate the need to perform costly data refreshes or data snapshot comparisons, enable changes to be processed in real time, improves the efficiency of the entire ETL process, deliver timely information to end-users at a low cost, minimize system impact, and reduces the associated costs including CPU cycles, storage, network bandwidth, human resources and enables timely consistent analysis of changed information.

References

[1] Robert M. Bruckner, A M. Tjoa (2002). "*Capturing Delays and Valid Times in Data Warehouses - Toward Timely Consistent Analyses*". Journal of Intelligent Information Systems (JIIS), Vol. 19(2), pp. 169-190, Kluwer Academic Publishers, September 2002.

[2] Robert M. Bruckner, A M. Tjoa (2001). "*Managing Time Consistency for Active Data Warehouse Environments*". In Proceedings of the Third International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2001), Springer LNCS 2114, pp. 254-263, Munich, Germany, September 2001

[3] Robert M. Bruckner, Beate List, Josef Schiefer (2002). "*Striving Toward Near Real-Time Data Integration for Data Warehouses*". In Proceedings of the Fourth International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2002), Springer LNCS 2454, pp. 317-326, Aix-en-Provence, France, September 2002.

[4] Josef Schiefer, Jun-Jang Jeng, Robert M. Bruckner (2003). "*Managing Continuous Data Integration Flows*". Decision Systems Engineering Workshop (DSE'03); co-located with 15th Conference on Advanced Information Systems Engineering (CAiSE'03), CEUR Workshop Proceedings, Velden, Austria, June 2003.

[5] Josef Schiefer, Robert M. Bruckner, (2003). "*Container-managed ETL Applications for Integrating Data in Near Real-time*". In Proc. of the International Conference on Information Systems (ICIS 2003), AIS Publishing, pp. 604-616, Seattle, WA, USA, Dec. 2003.

[6] W.H. Inmon and Dan Meers "*Maximizing the "E" in Legacy Extract, Transform & Load (ETL)*" December 2003

[7] White Paper by Attunity Ltd. "*Efficient and Real Time Data Integration with Change Data Capture*" February 2009 Available: http://www.attunity.com

[8] E. Schallehn, K. U. Sattler, and G. Saake, "*Advanced Grouping and Aggregation for Data Integration*". CIKM- Atlanta, GA, 2007.

[9] Jorg, T., Dessloch, S. "*Towards generating ETL processes for incremental loading*" IDEAS, 2008

[10] Jorg, T., Dessloch, S. "*Formalizing ETL Jobs for Incremental Loading of Data Warehouses*" BTW, 2009

[11] Kimball, R., Caserta, J. "*The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data*" John Wiley & Sons, 2004

[12] Samuel S. Conn "*OLTP and OLAP Data Integration: A Review of Feasible Implementation Methods and Architectures for Real Time Data Analysis*" 2005 IEEE.

[13] N. Kannan, "*Real-Time Business Intelligence – Building Block for Business Process Optimization*", DM Review Online. July 2004

[14] I. Ankorion. "*Change Data Capture-Efficient ETL for Real-Time BI*". Article published in DM Review Magazine, January 2005 Issue.

Author's Profile

Darshan M. Tank
dmtank@gmail.com

Educational Background				
Degree	**Board/University**	**Year of passing**	**Percentage**	**Class Obtained**
M.E. (CE)	DDU, Nadiad	2009	64.00	First
B.E. (IT)	Saurashtra University, Rajkot	2004	66.00	First (Distinction)

Work Expr (Industry - 3.0 yrs)			
Sr	**Company**	**Designation**	**Duration**
1	Marwadi Shares and Finance Ltd. Rajkot	Sr. Programmer	Oct 06 to May 07
2	Creative Infotech Pvt. Ltd, A'bad	Software Developer	Jan 05 to Aug 06
3	ACT Computer Education, Morbi	Jr. Programmer	June 04 to Jan 05

Work Expr (Teaching - 3.7 yrs)			
Sr	**Institute**	**Designation**	**Duration**
1	L.E. COLLEGE, MORBI (Information Technology Dept)	Asst. Prof.	9th May 2011 to till date
2	CHARUSAT - Charotar Institute of Technology – Changa	Asst. Prof.	1st July 2008 To 7th May 2011

My area of research interest is Real-time Business Intelligence Using Dynamic Data Warehouse Environment. During my dissertation, I had worked on “DYNAMIC WAREHOUSING AND MINING” and the same I have implemented using Microsoft SQL Server 2005 and IBM DB2 V9.0.

Paper Published In International Conference:		
Sr.	**Name of the Conference**	**Title**
1.	Information Technology and Business Intelligence – 2009 at IMT Nagpur	Timely Consistent Analysis with Minimized Latency for Data Propagations using Dynamic Data Warehouse Environment
2.	Advances in Recent Technologies in Communication and Computing, ARTCom 2010 at Kottayam Kerala,	Speeding ETL Processing in Data Warehouses Using High-Performance Joins For Changed Data Capture (CDC)