International Journal of Computer Theory and Engineering, Vol. 2, No. 5, October, 2010
1793-8201

Empirical Study on Dynamic Warehousing

C K Bhensdadia*, Yogeshwar P Kosta**

***Abstract*— In large enterprises, huge amounts of data are generated and consumed, and only substantial fractions of the data change rapidly. Decision makers need up-to-date information to make timely and sound business decisions. Unfortunately, conventional decision support systems do not provide the low latencies needed for decision making in this rapidly changing environment.**

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.

***Index Terms* – Dynamic Warehousing, Data Warehouse, latency.**

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. 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, 2. Delays in loading and integrating data into the data warehouse.

The aim of our approach is to shrink propagation delay 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.

A. Latency In Data Warehousing

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. Therefore, only after a time-window, the decision based on these analysis results and the relevant action can be performed.

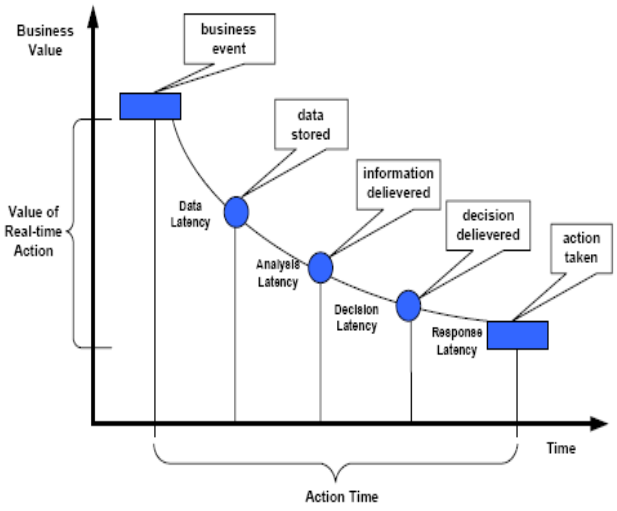

Figure 1. Business value and action time]

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. 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. This type of latency mostly depends on the time the decision makers need to decide on the most appropriate actions. Response latency is the time needed to take an action based on the decision made and to monitor its outcome.

* Professor & Head, Dept. of Computer Science & Engineering, Dharmsinh Desai University, Nadiad, Gujarat (INDIA) (ckbhensdadia@yahoo.co.in)

** Dean, Faculty of Technology (ypkosta@yahoo.com) (IEEE Member), Charotar University of Science Technology (CHARUSAT), Education Campus, Changa – 388421, Ta – Petlad, Dist – Anand, Gujarat (INDIA)

IACSIT International Association of Computer Science and Information Technology WWW.IACSIT.ORG

B. 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.

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 [1]. 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.

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.

C. Traditional VS. Dynamic Data Warehouse

TABLE 1: DIFFERENCE BETWEEN TRADITIONAL AND DYNAMIC DATA WAREHOUSE

Traditional Warehouse	Dynamic Warehouse
Strategic • Passive • Historical trends	Tactical • Focuses on execution of strategy
Batch • Offline analysis	Real-Time • Information on demand • Most up-to-date view of the business
Isolated • Not interactive	Integrated • Integrates data warehousing with business processes
Best effort • Guarantees neither availability nor performance	Guaranteed • Guarantees both availability and performance

D. 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: 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

E. Building Blocks for Operational BI

Operational BI goes beyond traditional BI implementation in two key areas: (1) expanding the use of BI to virtually anyone (Information Democracy) and (2) access to the most current information for decision making. This demands that the technology and architecture that support BI systems be pervasive, productive, and efficient. In addition, these BI systems need to ensure on-demand, or right time information availability (i.e. what is needed, when is it needed, where is it needed). There are three fundamental building blocks for an Operational BI system:

1) Information Delivery – Efficient, Pervasive, and Productive Delivery of Information to the User
2) Information Serving – A database platform that makes information available for delivery
3) Information Integration – Real-time, On-demand (or just-in-time) integration

These building blocks are depicted below as layers in an Operational BI solution followed by a detailed discussion of each.

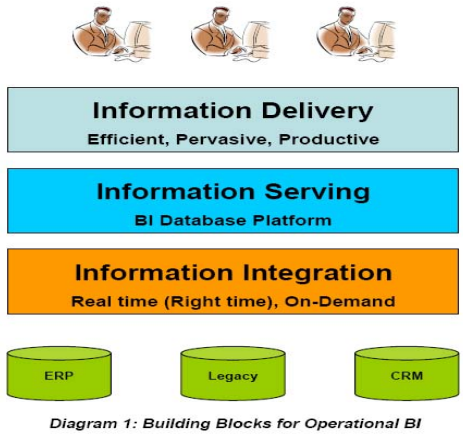

Figure 2.

A. Information Delivery

Operational BI means delivering more information to more people in support of tactical as well as strategic decision making. This requires solutions to be productive, simple and ubiquitous. Tools such as dashboards, reports, and portals need to be seamless and intuitive for all users. Furthermore, solutions must lend themselves for rapid changes,

customization, and self-configuration to allow for adaptation to market needs as well as personal needs and preferences.

B. Information Serving

Typical technologies used to serve information for BI purposes include the Data Warehouse (DW), Data Marts, Operational Data Store (ODS), and databases that offload operational data for reporting purposes. Traditionally, these systems are updated periodically, on a weekly or even nightly basis. For Operational BI, these data stores need to be updated more frequently, to guarantee information currency. For example, consider a dashboard that provides access to key performance indicators (KPI). In Operational BI, we would like to have some KPI updated multiple times within the course of a single day, or even continuously. In addition to serving integrated information in a physical database, a virtual database can be used for delivering the most current results. This approach is based on virtual data federation technology.

C. Information Integration

As mentioned earlier, data integration is a key component of BI. For Operational BI, data integration deals with the following: 1. Users need current information. The more recent the data the more valuable it is. Yesterday's data is often dated and is used only for historical analysis. 2. Data Volumes are exploding - Meaning that it is becoming more challenging to integrate, move and prepare the data so it is ready for delivery to the business users to be used as INFORMATION. 3. Batch windows are shrinking - Nightly batch operations used to move data into BI databases are becoming obsolete. Given the above business requirements, there is a need to improve the process of data and information integration and a move to a real-time environment. No longer can the entire source data be moved and massaged. Next generation Data Integration and ETL (Extract-Transform-Load) tools need to support Change Data Capture (CDC), a technology that enables to identify, capture, and move only the changes made to enterprise data sources. Implementing CDC makes data and information integration in REAL TIME significantly more efficient, and delivers data at the right-time.

II. Background Study

Over 85% of the data in an enterprise are unstructured and it needs to be managed. The most significant impact on warehousing has been the need for real-time warehouse data via WCS (warehouse control systems). In today's economy, distribution centers need to be more dynamic to meet the ever changing demands of the global economy. They must constantly re-invent themselves, whether it is simply expanding an existing footprint, adding new operational processes such as value added services, or finding better ways to fulfill orders quicker. Warehouses cannot remain stagnant.

The ability of a warehouse to be dynamic depends on the configurability and scalability of the WCS. The warehouse control system enables an automated warehouse or distribution center to reach peak operating performance. These new technologies remove the inefficiencies commonly associated with under or over utilized labor and material handling equipment. As an element of lean manufacturing and elimination of waste, a warehouse control system pulls product through an automated warehouse or distribution center increasing overall productivity and throughput.

Some solutions offer that the key to the optimization of material flow by warehouse automation is tracking key performance indicators such as the current and anticipated workloads at workstations in order to make key material routing decisions; inbound and outbound order tasks to make key material release decisions. Most organizations realize that the key to success lies in how well they manage data and the banking industry is no exception. From customer statistics to strategic plans to employee communications, financial institutions are constantly juggling endless types of information. Not only does this data provide the basis for major corporate moves, it also impacts business on a more granular level by helping to maintain customer loyalty and improve staff productivity. Simply put, a bank's information is its lifeline. That's why it's critical for financial institutions to be able to access relevant data when it's needed most.

Yet many banks struggle to access in-context data in a timely manner because it's often disorganized and inaccessible. In an effort to gather customer and operations information, many banks have one or more data warehouses or data marts. All too often, however, traditional data warehousing systems don't have the flexibility to help managers organize and integrate information into business processes, so some tasks may become inefficient, disorganized or unfocused.

Information may be structured, where it's similar in form and consistently stored. Or it may be unstructured, where it appears in diverse forms and is often scattered across disparate repositories and silos. The combination of structured and unstructured data makes it difficult for banks to find, analyze and use information to accomplish daily tasks. And even when the data is accessible, it may be outdated or irrelevant.

III. The Information Latency Continuum and IT Projects

The timeliness of information, or what we refer to as the information latency continuum, can span a wide range, depending on how fresh the information must be to provide the most business value. The latency of information used for analytical purposes can typically range from weeks to days, as is required for business intelligence (BI) reports based on historical data. On the other hand, information used in operational scenarios such as inventory management, straight-through processing, customer on-boarding and online order confirmation typically must be delivered within hours, minutes or seconds.

Different business scenarios have specific tolerance levels for receiving analytical and operational information. For example, the tolerance level for receiving order confirmation information after placing online orders on e-commerce Web sites cannot be greater than a few minutes, whereas it may be

IACSIT International Association of Computer Science and Information Technology WWW.IACSIT.ORG

quite acceptable to send an inventory restocking notification to a supplier within hours. More exacting scenarios such as online self-service banking portals can drive the demand for more current or "live" customer account information to seconds.

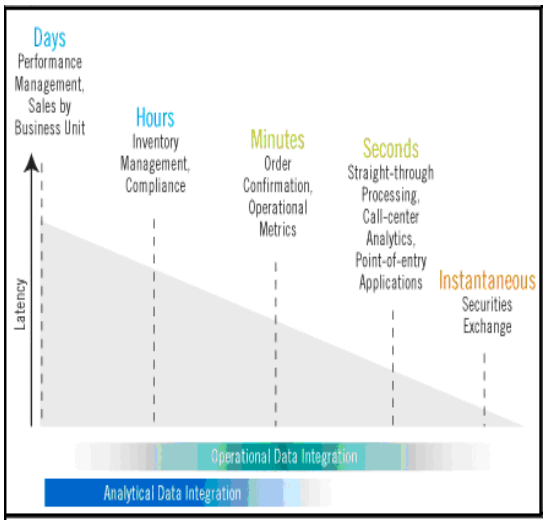

Figure 3. The Information Latency Continuum

In order to cost-effectively access and deliver timely and trusted data to support both analytical and operational scenarios, IT organizations typically undertake a variety of data integration projects. The success of a data integration project typically depends on the ability to meet an IT organization's service level agreements (SLAs) related to data latency, data completeness and data accuracy.

IV. Need For Research

Traditional data warehouses are increasingly being challenged by demands for real-time data access, analysis of structured and unstructured data, and the need to synchronize core customer and product information across operational systems to create a single view of the enterprise. Dynamic warehousing is an approach that enables organizations to deliver more dynamic business insights by integrating, transforming, harvesting and analyzing insights from structured and unstructured information. Capable of processing large amounts of information, a dynamic warehousing infrastructure can enable organizations to respond on demand to unscheduled analysis requests, and as events trigger the need for information throughout the day. It outlines the options and decision criteria for selecting the right components that define the extended infrastructure for dynamic warehousing, including how to:

1) Access and integrate source system data.
2) Gather requirements and implement changes to the data warehouse model.
3) Synchronize master data for key business entities across operational systems.

Combining the right warehouse platform and a comprehensive industry data model with a unified platform for data integration that all share business concepts, transformation rules and metadata can enable the deep collaboration between business analysts and IT that is required to deliver information on demand. Add to this a data warehouse platform that is capable of handling the mixed workloads, real-time analysis and advanced analytics, and organizations can deliver a truly dynamic warehousing infrastructure that delivers the right information, whenever it is required, wherever it is needed.

The real-time analysis requirements demand a set of service levels that go beyond a traditional Business Intelligence System.

1) Data freshness
2) Continuous data integration
3) Highly available analytical environments based on an analysis engine
4) Active decision engines
5) An adaptive platform for the event stream processing
6) High availability and scalability

V. Goal & Objective

The goal of a dynamic warehouse is to deliver information on demand to people and processes, which can help to dramatically shorten the lag between events and actions while improving the quality of decisions. A dynamic warehouse model should be tightly integrated with the data integration tools that manage and load source system data and the data modeling tools that implement changes to reflect new requirements. A dynamic warehouse must leverage an infrastructure that will:

1) Deliver real-time access to data in context for each application
2) Embed analytics as part of a business process
3) Extract information from unstructured as well as structured information
4) Support multiple applications with changing requirements and increasing volumes of data.

Dynamic warehousing is part of the next generation of technology that enables organizations to gain more business insight and deliver relevant information on demand. An effective solution can help financial institutions access and analyze existing data and ultimately boost business in a highly competitive market. Whereas traditional data warehouses can make it difficult to keep up with today's fast-paced banking environments, dynamic warehousing delivers immediate, integrated information. With this information at their fingertips, employees can take action and make timely decisions. The ideal dynamic warehousing solution can help:

1) Leverage information in innovative ways to optimize business processes.
2) Access relevant, real-time data for a single version of the truth.
3) Improve productivity by enabling employees and customers to retrieve information where and when they need it.
4) Build better customer relationships and implement more effective marketing campaigns.
5) Ease compliance requirements for regulations and risk management.

VI. THE DEMAND FOR REAL-TIME INTEGRATION

According to a Forrester survey of more than 600 technology decision makers and enterprises across North America, improving integration between applications is one of the top priorities for organizations today. The following are some of the business challenges that organizations face in today's competitive market.

1) Increasing demand for real-time information for reporting and analytics. Traditionally, reporting was done from warehouses which were updated on a daily or weekly basis. For many types of reports, that data is current enough. For others, though, nothing short of up-to-the-minute would suffice, such as inventory data where product inventory is very high or very low, or billing information where billing is done by the minute or every fraction of a day.
2) Large volumes of information are difficult to handle in a batch window. As more information is gathered – such as online transaction data, inventory data, and customer information – the effort involved in moving it to the warehouse increases drastically. Many organizations are finding that an eight hour batch window is no longer sufficient for traditional ETL tools to integrate all of the needed data.
3) Necessity to conduct business 24/7 is reducing batch windows. As more business is done across time zones and over the web, many organizations are faced with the problem of shrinking batch windows, making it more difficult for traditional ETL tools to extract data in the short time available.
4) Growing need to detect and react to business events as they happen. Many organizations are looking for ways to detect business events in production systems and have those events trigger a response in another system. For example, a cell phone company would like to send a text message to a customer running low on minutes asking if him if he would like to purchase more.
5) The need to track all changes for auditing purposes. Organizations need to comply with regulations, which often require them to continuously track all changes to data and not just the net result of those changes.
6) Increasing need to keep data in sync across the enterprise. Customers want up-to-the minute access to order, payment and inventory data so they can buy products, pay bills and check delivery status online. Employees need much of the same so they can better service customers and make wise business decisions.

Organizations want to deploy new applications using data on legacy systems without paying for an increase in workload. Organizations want new applications running on new platform to avoid this cost, but integrating the data from those legacy systems without increasing the load on them is a key challenge.

A. Real-Time Data Integration

Permanently integrating data from different operational sources addresses the time liness issue discussed in the previous section. Feeding data continuously into a data warehouse minimizes the average latency from when a fact is first captured in an electronic format somewhere in an organization until it is available for a knowledge worker who needs it. Besides technical challenges (mixed workload caused by concurrent updates and analytical queries, scalability, performance, minimized scheduled downtimes, etc.) there are issues which directly affect the analytical environment.

1) Analysis results may change unexpectedly from the analyst's perspective during the repetition of an identical analytical query if the result set was affected by newly integrated data in the meantime. This is a really critical situation, because it confuses analysts that are accustomed to the stable snapshot paradigm for data warehouses. It is very difficult for them to determine the cause for such an unexpected change: is it newly integrated data, or is there a data quality problem. Thus, there is a need for providing a data warehouse model that copes with continuously integrated data and nonetheless is able to reflect a stable view of the data as it exists at some point in time.
2) Keeping aggregates current. Aggregates are intended to provide better performance for analytical queries which return results at some aggregated level, rather than all the detailed data. This is a common situation in analytical environments using OLAP. In a traditional data warehouse all the aggregates are updated at the end of every update window. However, in a continuous loading environment this is not feasible. We need a model that is able to reflect multiple versions of aggregates regarding the same dimension hierarchy levels.
3) Active analytical environments. Active data warehouses need a powerful scheduling mechanism for the mixed workload of queries and continuous data updates. The response time requirement for tactical queries is very critical compared with complex strategic analyses, where a decrease in priority will only have minor effects on the response time.

A dynamic warehouse model should be tightly integrated with the data integration tools that manage and load source system data and the data modeling tools that implement changes to reflect new requirements. In Real-time Data Integration, we should consider two key aspects of timeliness.

1) Response time: The time it takes for a response to be provided back to the user.
2) Latency: It refers to the currency of the information that is "how old is it".

Included with SQL Server 2005, SQL Server Integration Services (SSIS) offers a comprehensive, powerful, and competitive solution for data integration and ETL. The successor to Data Transformation Services (DTS), SSIS is a platform for a new generation of high performance data integration technologies. The main goal of real-time data extraction, transformation and loading tools are,

1) Keep warehouse refreshed
2) Minimal delay

IACSIT International Association of Computer Science and Information Technology WWW.IACSIT.ORG

The following are the issues related with data integration tools.

1) How does the system identify what data has been added or changed since the last extract
2) Performance impact of extracts on the source system

B. Change Data Technology & Real-Time Data Integration

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. This latency can typically be configured to be as near to real time as is practical, opening up new business opportunities of faster decisions, faster reaction times, and faster business execution.

Next generation Data Integration and ETL (Extract Transform and Load) tools need to support Change Data Capture (CDC), 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.

A key goal of CDC is to improve efficiency by reducing the amount of data that needs to be processed to a minimum. 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. CDC has many benefits and is complementary to traditional ETL tools.

1) Changes are captured in real-time so information is always up-to-date. CDC captures changes continuously as they occur. The resultant information is always up-to-date rather than being only as current as the last batch window.
2) No impact on the performance of production systems. CDC reads database log files rather than querying the database directly.
3) No requirement for batch windows. With changes captured, transformed and applied continuously, there is no need to take the systems down to extract data.
4) Easily scales to very large databases and large numbers of transactions. Only changes are replicated rather than all of the data in the changed tables. The result is much greater scalability through less data being moved across.
5) Does not require changes to the source system. Because CDC is only reading the log, it does not require changes directly to the source database, yet it can detect all transactions including descriptive information about the change.
6) Logs all changes to the system, not just the net results of those changes. For audit and compliance, all insert, update and delete actions are recorded rather than just the net results of those actions.
7) Complements ETL tools. Many corporations combine the strengths of CDC and ETL tools. By using CDC to flow changes in real-time to ETL tools for transformation customers get the best of each product.

VII. 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.

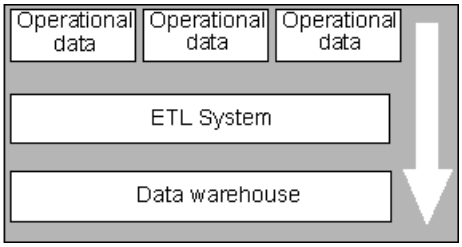

Figure 4. ETL System Architecture]

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.

A. 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.

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.

B. On-Demand Data Warehouse

Data Warehouses (DW) are used to support reporting, analysis and data mining applications. They are typically updated on weekly or nightly. Today, DW are starting to become a source of information that needs to support operational type of applications and as such need to provide re-fresh data. That is the On-Demand DW, where information is available at the right time, meaning that different information items are updated at different times as required by the applications that a given DW is required to support. Once in the DW, tools like SQL Server Reporting Services or Analysis Services can seamlessly be used in order to delivery the right information to the right people, at the right time.

The following Blue Print describes how to build an Active DW where information can be updated in different frequencies. The Blue Print recommends using Change Data Capture regardless of frequency due to the significant improvements in efficiency.

C. Solution BluePrint using Microsoft and Attunity software:

3.1 Software

1) Microsoft SQL Server 2005 – scalable and reliable DW Server
2) Microsoft SQL Server Integration Services (SSIS) – ETL for replicating and transforming the data
3) Attunity Connect – bulk extraction of enterprise data, including relational access to legacy data
4) Attunity Stream – periodic of continuous extraction of data changes

3.2 Process

1) Identify the source tables and what latency requirements you have.
2) Define a target schema in the DW SQL Server.
3) For First time loading for the DW:
4) Configure data source in Attunity Connect. For non-relational data sources Attunity will automatically create a relational schema.
5) Define an SSIS process that uses the configured Attunity Connect data source for extracting the source data, and then transforming and loading it into SQL Server.
6) For periodic updates to the DW (daily, hourly, every few minutes):
7) Configure a CDC (change data capture) agent in Attunity Stream. Attunity Stream provides access to the changes in the form of a new data source (i.e. "change source").
8) Define an SSIS process that uses the configured Attunity Steam change source for extracting the changes on the source data, and then transforming and loading it into SQL Server.

VIII. 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.

Change data capture (CDC) technology has become a strategic component of many data warehouse and business intelligence (BI) architectures. Given today's ever increasing struggle for more immediate access to real time data and information, constant pressure on efficiency costs, and the exponential growth of underlying data volumes, CDC is now a "must have" for the modern BI and data warehouse project.

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 (Extract Transform and Load) tools need to support Change Data Capture (CDC), 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.

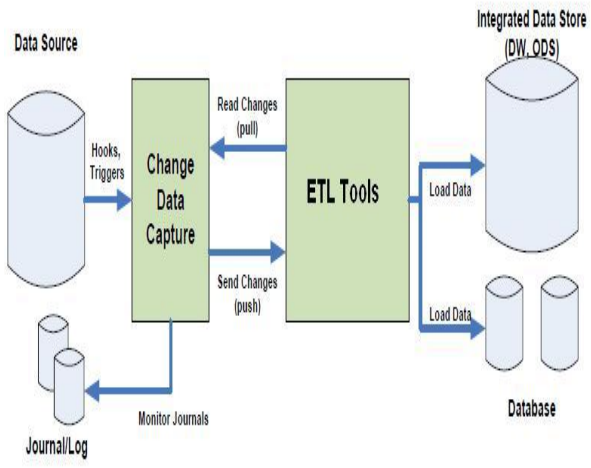

Figure 5. working of CDC in conjunction with ETL tools]

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

IACSIT International Association of Computer Science and Information Technology WWW.IACSIT.ORG

improve efficiency by reducing the amount of data that needs to be processed to a minimum. 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

B. The Value of CDC

The value benefits of using CDC technology can be many and varied. However, when used in conjunction with ETL tools such as Microsoft's SSIS it is normally adopted for one of the following key reasons:

- Delivering data on-demand and in near real time – when business requirements demand much more current and/or real time access to information from within their BI systems, CDC provides the means to deliver it.
- Dramatically increasing efficiencies – when processing and/or network resources are at a premium, using CDC to move only the data that has changed rather than periodical bulk data transfers, is a much more efficient way of addressing the growth in data volumes head-on.
- Eliminating the need for batch windows - data is captured and processed while the underlying systems keep working. The continuous data stream of a CDC solution can virtually eliminate the "batch window" bottleneck.
- A strategic solution - a comprehensive platform that works with other integration products and can be used for many initiatives, including BI/DW, CPM, BAM, migrations, consolidations, and more.
- Cost savings - CDC can substantially reduce IT operational costs in terms of human resources required for a given integration project, as well as on-going costs related to system and storage requirements.

Change data capture is an innovative new software technology that is changing the data integration landscape. With today's organizations striving for ever more immediate access to key business information in order to remain competitive, and IT departments struggling to maintain the exponentially growing data volumes being generated within their data based information systems, CDC technology can represent a very strategic solution to solving the problems.

C. Change Data Capture – Supported Platforms and Data Sources

1) Mainframe: Adabas, DB2 , IMS/DB , VSAM
2) Unix: Oracle 9i , Oracle 10g , Adabas
3) AS/400: DB2/400
4) HP Nonstop (Tandem): Enscribe, SQL/MP
5) Windows : SQL Server 2000,: SQL Server 2005, Oracle 9i, Oracle 10g, Adabas

IX. CONCLUSION

Dynamic Warehouse is implemented using Business Intelligence Development tools provided by SQL Server 2005. The main constraints we have considered here is to minimize propagation delay. Microsoft SQL Server 2005 Integration Services provides a platform for building high performance data integration solutions, including extraction, transformation, and load packages for data warehousing.

The major issue related with the current system is processing overhead. We can reduce this processing overhead by mean of CDC (Change Data Capture) technique. Instead of loading all of the source data and doing a complete refresh, we need to bring only changed data across from source systems into our data warehouse. Change Data Capture, a technology that enables to identify, capture, and move only the changes made to enterprise data sources.

X. RESEARCH ISSUES & FUTURE WORK

1) Search and text analytics capabilities so knowledge can be extracted from unstructured data.
2) Support for real-time access to aggregated, cleansed information.
3) Process management capabilities that leverage analytics for improved decision-making.
4) A high-performance, scalable system that can support operational needs.
5) Comprehensive data profiling and analysis.
6) Data cleansing, matching, and standardization.
7) Pre-built transformations and mappings to standard sources and links to target data model.
8) Real-time change data capture to mainframe and distributed systems.
9) Business glossary to link business terms to technical data definitions.
10) Universal connectivity to databases, applications, and syndicated information.
11) Ability to deploy integration processes as services callable by any application on demand.
12) Support for federated queries that access multiple sources without requiring source data movement.
13) Ability to track data lineage and perform impact analysis on proposed changes.

ACKNOWLEDGMENT

We are thankful to The Omnipotent GOD for making us able to do something. We express our gratitude to the management of DDU for providing us research opportunities and their wholehearted support for such activities. Finally, our acknowledgement can not end without thanking to the authors whose research papers helped us in making this research.

REFERENCES

[1] The dynamic warehousing infrastructure: Establishing a foundation to meet new information requirements by IBM

[2] From conflicting, unintegrated historical data to actionable insight – An introduction to dynamic warehousing from IBM
[3] Advances in Data Warehouse Performance – White Paper by Winter Corporation
[4] Zero-Latency Data Warehousing: the State-of-the-art and experimental implementation approaches by THO, M.N
[5] An Overview of Data Warehousing and OLAP Technology by Surajit Chaudhuri and Umeshwar Dayal
[6] Sense & Response Service Architecture (SARESA): An Approach towards a Real-Time Business Intelligence Solution and its use for a Fraud Detection Application
[7] Modeling Data Warehouse Refreshment Process as a Workflow Application by M. Bouzeghoub, F. Fabret, M. Matulovic-Broque
[8] A Method for Demand-driven Information Requirements Analysis in Data
[9] Warehousing Projects by Robert Winter and Bernhard Strauch - 2002 IEEE
[10] Enhanced Business Intelligence - Supporting Business Processes with Real-Time Business Analytics by Andreas Seufert and Josef Schiefer
[11] Container-Managed ETL Applications for integrating data in near real-time by Schiefer & Bruckner from Twenty-Fourth International Conference on Information Systems
[12] Designing an ETL Solution Architecture Using Microsoft SQL Server 2005
[13] Integration Services
[14] OLTP and OLAP Data Integration: A Review of Feasible Implementation Methods and Architectures for Real Time Data Analysis by Samuel S. Conn from IEEE Xplore
[15] Managing Continuous Data Integration Flows by Josef Schiefer, Jun-Jang Jeng, Robert M. Bruckner
[16] Enabling Operational Business Intelligence with Real-Time Data Integration
[17] Solutions Using Microsoft SQL Server 2005
[18] Real Time Data Integration Using Change Data Capture Technology with Microsoft SSIS by Attunity Ltd. – July 2008
[19] Striving towards Near Real-Time Data Integration for Data Warehouses by Robert M. Bruckner, Beate List, and Josef Schiefer
[20] Evaluating real-time data integration solutions by IBM Corporation April 2008

IACSIT
International Association of Computer Science and Information Technology
WWW.IACSIT.ORG