Speeding ETL Processing in Data Warehouses Using High-Performance Joins For Changed Data Capture (CDC)

Darshan M. Tank[1], Amit Ganatra[2], Y P Kosta[3]
[1,3]Department of IT, [2]Department of CE
Charotar Institute of Technology (Faculty of Technology)
Charotar University of Science and Technology, Changa
Anand-388421, India
[1]dmtank@gmail.com, [2]amitganu@yahoo.com,
[3]ypkosta@yahoo.com

C K. Bhensdadia
Department of CE
Faculty of Technology
Dharmsinh Desai University
Nadiad-387001, India
ckbhensdadia@yahoo.co.in

***Abstract*—In today's fast-changing, competitive environment, a complaint frequently heard by data warehouse users is that access to time-critical data is too slow. Shrinking batch windows and data volume that increases exponentially are placing increasing demands on data warehouses to deliver instantly-available information. Additionally, data warehouses must be able to consistently generate accurate results. But achieving accuracy and speed with large, diverse sets of data can be challenging.**

Various operations can be used to optimize data manipulation and thus accelerate data warehouse processes. In this paper we have introduced two such operations: 1. Join and 2. Aggregation – which will play an integral role during preprocessing as well in manipulating and consolidating data in a data warehouse. Our approach demonstrate how we can save hours or even days, when processing large amounts of data for ETL, data warehousing, business intelligence (BI) and other mission critical applications.

***Keywords*—Business Intelligence, Near Real-Time Data Warehousing, Change Data Capture (CDC), Extract-Transform-Load (ETL)**

I. INTRODUCTION

The widespread use of the Internet and related technologies in various business domains has accelerated the intensity of competition, increased the volume of data and information available, and shortened decision-making cycles considerably. Typically, in a large organization, many distributed, heterogeneous data sources, applications, and processes have to be integrated to ensure delivery of the best information to the decision makers. In order to support effective analysis and mining of such diverse, distributed information, a data warehouse (DWH) collects data from multiple, heterogeneous (operational) source systems and stores integrated information in a central repository.

There are numerous challenges facing IT departments today as they deal with reduced budgets, smaller staffs, and ever-increasing demands on the development of business critical applications. Within the factory, there are often large volumes of data that have to be processed from application to application or from repository to repository, creating a flow of data across the enterprise [1].

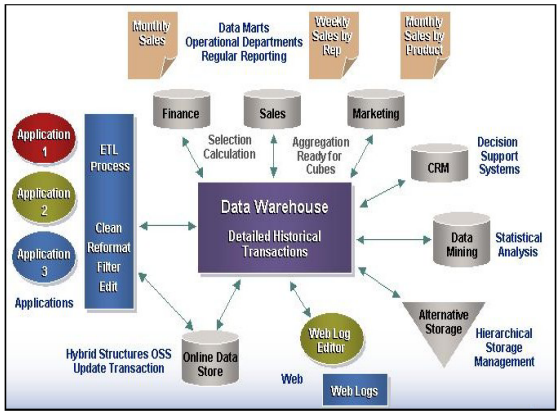

Figure 1. Technical architecture of Corporate Information Factory.

From above Fig., it's easy to see that data volumes can easily grow while demands on the data increase as well. This can occur when processing the data and putting it into another form or database. The management of all of this data movement and the time it takes to process all of the information creates a new set of IT challenges [5].

II. 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 [3].

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

978-0-7695-4201-0/10 $26.00 © 2010 IEEE
DOI 10.1109/ARTCom.2010.63

from the fact that data may be formatted differently in the different sources.

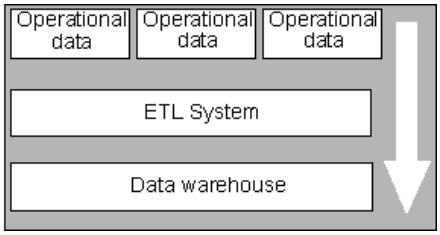

Figure 2. ETL System Architecture.

The ETL process is not a one-time event; new data is added to a data warehouse periodically [4]. 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.

III. DATA PROCESSING CHALLENGES

Staying competitive in today's real-time business world demands the capability to process ever-increasing volumes of data at lightning speed. In today's fast-changing, competitive environment, a complaint frequently heard by data warehouse users is that access to time-critical data is too slow. Shrinking batch windows and data volume that increases exponentially are placing increasing demands on data warehouses to deliver instantly-available information.

Additionally, data warehouses must be able to consistently generate accurate results. But achieving accuracy and speed with large, diverse sets of data can be challenging [5]. There has been increased need to update the underlying data infrastructure to improve access to, and the quality of, real-time business intelligence (BI) and other mission-critical operations, including data marts, data mining, online data stores, and online transaction processing (OLTP) systems.

One of the primary factors affecting data warehouse performance is the underlying physical distribution of the data in the databases, which must be consolidated in various ways [6]. An inability to perform data intensive operations efficiently will inevitably impede business analyses and subsequent decision-making.

IV. KEY COMPONENTS OF DATA WAREHOUSE PROCESSING

Various operations can be used to optimize data manipulation and thus accelerate data warehouse processes. Two operations in particular, join and aggregation play an integral role during preprocessing as well in manipulating and consolidating data in a data warehouse.

Deriving useful information from raw data within a data warehouse involves joining factual and dimensional information before aggregating it to produce a report or output for downstream analysis. However, joins and aggregations are data-intensive and time-intensive, which prevent data warehouses with insufficiently fast ETL software from performing them frequently; but frequency is key to providing data for timely business analysis [7].

For processing data for BI analytics and other mission-critical applications, the usual go-to software products include various extract, transform, load (ETL) and data warehousing solutions [12]. However, these products do not always provide the speed required to deliver data on time. It must support the Advanced Data Management (ADM) capabilities to reduce processing time of data-intensive operations through the use of high-performance joins and high-performance aggregations.

A. Joins

Joins are used to combine information from two or more data sources, such as database tables, and place it into a new data source suitable for downstream processing or reports. Joins are particularly powerful because they enable rapidly changing data to be organized into categories for subsequent report preparation through the use of matching keys. Joins are used to pre-process data, to improve the efficiency of queries, and to accelerate changed data capture (CDC).

B. Aggregations

Aggregations play a key role at various stages of data warehouse processing. From preprocessing of data prior to its entering the data warehouse, to dimensional data analysis used for conducting queries and generating reports, aggregations are critical to efficiently preparing, configuring, and analyzing large volumes of data.

1) Using High-Performance Aggregations for Preprocessing: By aggregating data prior to loading it into the data warehouse, queries, database loads, and other downstream processing can be performed much faster. ADM should quickly summarize factual data to the minimum level of granularity required by the data warehouse. A common application removes redundant transaction data from multiple sources for faster queries and database loads [17].

2) Pre-Calculated Aggregations Speed Queries: Data warehouse experts agree that aggregates are the best way to speed data warehouse queries. According to data warehousing expert Ralph Kimball, "Every data warehouse should contain precalculated and prestored aggregation tables". Aggregate operations yield the greatest performance benefits when they are used for input to data analysis.

In the case of a six million row query, reducing the number of rows read by creating aggregations across dimensions can vastly accelerate processing time [13]. A query answered from base-level data can take hours and involve millions of data records and millions of calculations. With pre-calculated aggregates, the same query can be answered in seconds with just a few records and calculations.

3) Using Multi-Level Hierarchal Aggregations: Sophisticated aggregation schemes recognize dimensional hierarchies and build higher-level aggregations from more granular aggregations. For instance, a daily aggregation of sales for a region could be used to build a monthly aggregation of sales by region, and thereby avoid an aggregation of the more granular daily sales totals. Such aggregate-awareness can significantly increase speed and enhance reporting capacity.

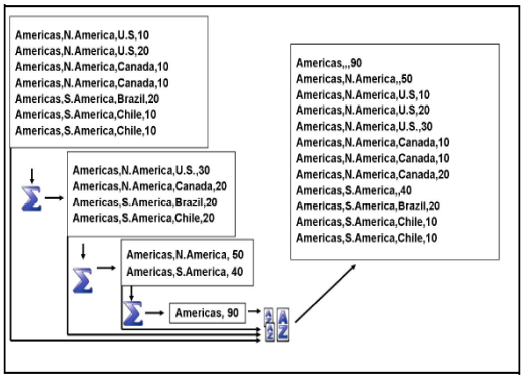

Figure 3. Multi-Level Hierarchal Aggregation.

Aggregations can also be used to replace fact data with rolled up versions of themselves. ADM should also perform multi-level hierarchal aggregation (Fig. 3) to help build cubes for more advanced dimensional data analyses.

V. CDC (CHANGE DATA CAPTURE)

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. CDC solutions occur most often in data-warehouse environments since capturing and preserving the state of data across time is one of the core functions of a data warehouse, but CDC can be utilized in any database or data repository system. 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 (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 [6].

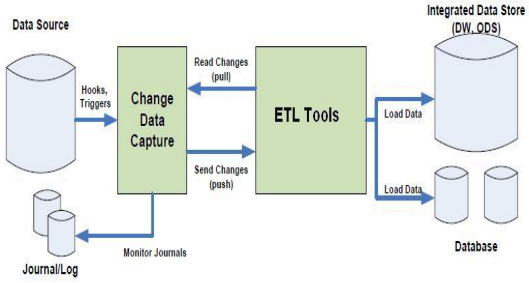

Figure 4. Working of CDC in conjunction with ETL tools

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 [6]. 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.

VI. HIGH-PERFORMANCE JOINS FOR CHANGED DATA CAPTURE (CDC)

To reduce the time needed to retrieve information, data must be preprocessed into the proper form for the dimensional data warehouse. High-speed joins are critical to this process, which can include lookups of legacy values for appropriate replacement, cleansing and validating data, identifying and eliminating mismatching values, and pre-aggregation. Using high-performance join for data-intensive operations, descriptive information can be combined with factual data late in a processing sequence so that storage and throughput requirements are minimized.

CDC is an increasingly important pre-processing function. By loading only new, updated, and deleted records into the data warehouse, CDC significantly conserves time and resources when carrying out data-intensive processes [16].

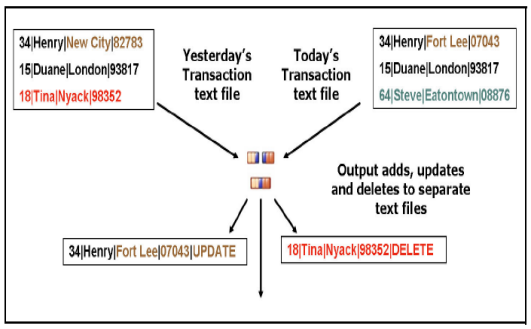

Figure 5. Joins Used for Changed Data Capture.

Rather than replacing the information in the data warehouse with the data in the entire online transactional database, a join will match the primary key of the previously loaded record with its corresponding new record and then compare the data portions of the two records to determine if they've changed. In this way, only added, deleted, and altered records are updated, which significantly reduces elapsed time of database loads. By using a high-performance join for CDC, data warehouse updates can be performed with far greater efficiency.

VII. CONCLUSIONS

Aggregations and joins are widely used in nearly every industrial sector where large volumes of data must be analyzed and to significantly accelerate data processing. It speeds processes like ETL, staging data for a data warehouse, and database loading. It also minimizes resource consumption, making it possible to consolidate hardware for significant cost savings.

ACKNOWLEDGMENT

The authors' wishes special thanks to the Management of Charusat for their moral support and continuous encouragement.

REFERENCES

[1] Imhoff and Claudia "The Corporate Information Factory" DM Review, December, 1999 (http://dmreview.com)

[2] Bruckner, R. M., and Tjoa, A M. "Capturing Delays and Valid Times in Data Warehouses—Towards Timely Consistent Analyses," Journal of Intelligent Information Systems, September 2002.

[3] Josef Schiefer & Robert M. Bruckner "Container-Managed ETL Applications for integrating data in near real-time" Twenty-Fourth International Conference on Information Systems (2003)

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

[5] White Paper by Syncsort Incorporated "Solving the Challenges of Exponential Data Growth" 2009 (http://syncsort.com)

[6] White Paper by Attunity Ltd. "Efficient and Real Time Data Integration with Change Data Capture" February 2009 (http://attunity.com)

[7] Robert M. Bruckner, A M. Tjoa "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.

[8] Robert M. Bruckner, Beate List, Josef Schiefer "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.

[9] Josef Schiefer, Jun-Jang Jeng, Robert M. Bruckner "Managing Continuous Data Integration Flows". Decision Systems Engineering Workshop (DSE'03), Velden, Austria, June 2003.

[10] Thomas JÄorg and Stefan Dessloch "Near Real-Time Data Warehousing Using State-of-the-Art ETL Tools", 2007

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

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

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

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

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

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

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