International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 3 Issue: 1 | Jan 2016 www.irjet.net p-ISSN: 2395-0072

Accelerating ETL Efficiency: A Cutting-Edge Data Integration Framework for Real-Time Business Insights

Robin Verma[1]

[1]*Independent Researcher, Pune, Maharashtra, India -411014*

Abstract - *In today's fast-paced, competitive environments, reducing business latency is critical. This entails responding instantly to new information as it arrives and having the right data at the right time for optimal decision-making. The challenge of integrating, processing, and delivering real-time results becomes even greater as data volumes grow and sources become more distributed.*

Traditional data warehouse environments often suffer from delayed decision-making processes due to the lag between capturing data from the source and integrating it into the data warehouse. Typically, updates in such environments occur daily, or even weekly, exacerbating the latency issue. In these scenarios, keeping data up to date and minimizing the time gap between data capture and decision-making is a challenging task. The real-time data warehouse is a solution aimed at reducing decision-making time by minimizing latency and striving to achieve near-zero delay between business action and corresponding analysis.

The conventional ETL process, which periodically takes snapshots of entire source systems, is often time-consuming and resource-intensive. Alternatives such as using timestamp columns, triggers, or complex queries can worsen performance and increase system complexity. Therefore, what is required is a steady, dependable stream of change data, formatted in such a way that it can be easily consumed and applied to target data structures.

With growing market competition, the need to minimize ETL load times becomes increasingly urgent. This paper introduces a data integration approach designed to reduce ETL load times while maintaining data integrity and responsiveness.

Key Words: Business Intelligence, Data Engineering, Dynamic Warehouse, Real-Time Data Integration, Change Data Capture.

1.INTRODUCTION

This document is template. We ask that authors follow some simple guidelines. In essence, we ask you to make your paper look exactly like this document. The easiest way to do this is simply to download the template, and replace(copy-paste) the content with your own material. Number the reference items consecutively in square brackets (e.g. [1]). However, the authors name can be used along with the reference number in the running text. The order of reference in the running text should match with the list of references at the end of the paper.

The volume of data that large enterprises manage is rapidly expanding, with new information being generated continually by operational systems. To support efficient analysis and extraction of insights from this distributed, varied information, data warehouses gather data from multiple heterogeneous sources and store integrated information in a centralized repository. However, data warehouses must be updated periodically to reflect new data from these source systems [1]. Operational systems capture real-world events, but a delay occurs during this data propagation process.

Traditional data warehouses operate with periodic updates, such as daily or weekly intervals, creating significant latency in propagating data from the source systems to decision-makers. The nature of traditional data warehousing, which often involves a 'write-once, read-many' approach, is incompatible with the continuous update cycles required for real-time responsiveness. Additionally, most legacy models lack the in-built mechanisms necessary for handling real-time changes, leading to delays that slow down business insights.

The architecture of an active data warehouse must address two critical types of latency:

1. The delay in capturing real-world events in operational systems.
2. The delay in loading and integrating data into the data warehouse.

The purpose of the proposed approach is to leverage Change Data Capture (CDC) techniques to reduce ETL load times, ultimately ensuring that information can be delivered more quickly and with higher relevance. This enables analyses that can account for real-time changes in source systems, improving business outcomes.

2. Latency In Business Intelligence

The value of a business decision diminishes as the time between the occurrence of an event and the response increases. Unfortunately, in traditional environments, data is often stored in data warehouses long after the relevant event has occurred. This delay impacts the timeliness of decision-making processes. Once data is captured and stored, it undergoes analysis and packaging before being delivered to business users, creating additional delays [4]. This results in the ability to take action only after a significant time lag.

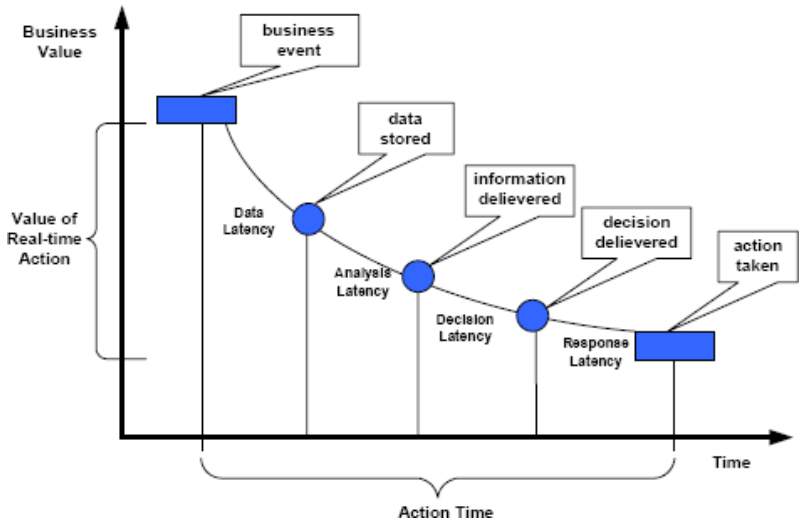

Fig. 1 Business value and action time

The total elapsed time from the occurrence of an event to the moment a decision is made and action is taken is referred to as action time. Action time can be broken down into several components:

1. **Data latency**: The time between when a business event occurs and when the data is captured and ready for analysis.
2. **Analysis latency**: The duration from when data is available for analysis until meaningful insights are generated.
3. **Decision latency**: The time required to deliver insights to decision-makers and determine the appropriate course of action [5].
4. **Response latency**: The time it takes to act on the decision and observe the outcome.

In traditional BI environments, these various forms of latency can lead to suboptimal decision-making, as the window of opportunity for the most impactful actions may have passed.

3. Real Time Business Intelligence

In the business world, having real-time information is becoming essential. The purpose of business intelligence (BI) systems is to enable more informed and faster decision-making [12]. BI systems can address critical business concerns, such as identifying growth opportunities, analyzing competition, understanding customer behavior trends, and monitoring key performance indicators.

Real-time business decisions require timely, integrated, subject-oriented data, but traditional BI systems are not designed for real-time operational support [13]. These analytical systems typically function independently of operational IT systems, leading to latency due to human intervention. The need for real-time analysis creates new service level demands, surpassing the capabilities of traditional BI systems.

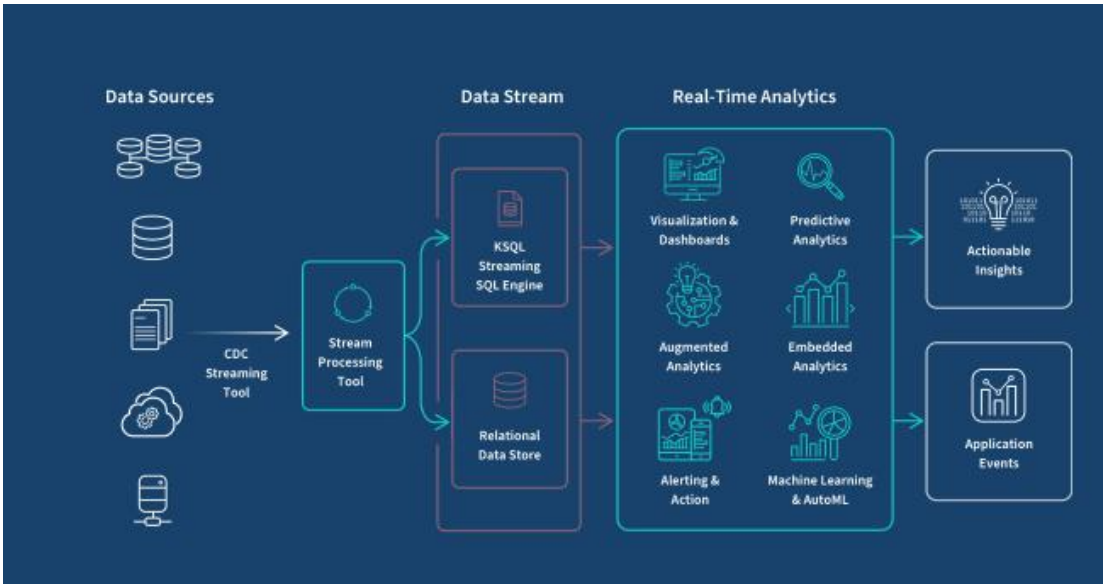

Figure. 2 Real-time Business Intelligence

Fig. 2 illustrates the concept of real-time BI. On one side, it shows the data ingestion process from various sources, and on the other, it shows how this data is utilized in real-time or near real-time to support decision-making.

Real-time BI delivers actionable data quickly enough to meet business demands, allowing for timely decision-making.

4. Dynamic Warehouse

Dynamic or real-time data warehousing (RTDW) refers to the technical processes that keep data warehouses updated in near real-time. This ensures that any change in a source system is instantly reflected in the data warehouse. RTDW encompasses database modifications, data movement, ETL processes, and updates to downstream processes such as data marts [2].

RTDW enables the timely delivery of relevant information to the right people. Essential operational decisions, like evaluating promotion effectiveness or customer retention, require integrated, up-to-date data.

However, traditional BI systems fall short in providing real-time decision support, causing latency. RTDW sets new expectations for data freshness, continuous integration, and real-time decision engines, which traditional BI systems cannot meet [3].

Dynamic warehousing represents the future of technology, providing businesses with more insightful data and timely information delivery. Traditional data warehouses struggle to meet today's fast-paced demands, while dynamic warehousing offers immediate access to integrated information.

5. Design Considerations For RTDW

Designing a real-time data warehouse (RTDW) involves several technical considerations, such as scalability, high availability, continuous data loading, and workload balancing. It must also address the two types of propagation delays[3]:

1. Delays in capturing real-world events.
2. Delays in data integration into the warehouse.

RTDW business requirements include:

1. **Performance**—Response times within seconds.
2. **Scalability**—Handling large data volumes and mixed workloads.
3. **Availability**—24/7 uptime, 365 days a year.
4. **Data Freshness**—Up-to-date information.

6. ETL PROCESS

ETL (Extract, Transform, Load) is a core process in data warehousing, responsible for pulling data from source systems, transforming it according to business logic, and loading it into the target warehouse. This process is repeated periodically, such as monthly, weekly, daily, or even hourly, depending on the warehouse's purpose [6].

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].

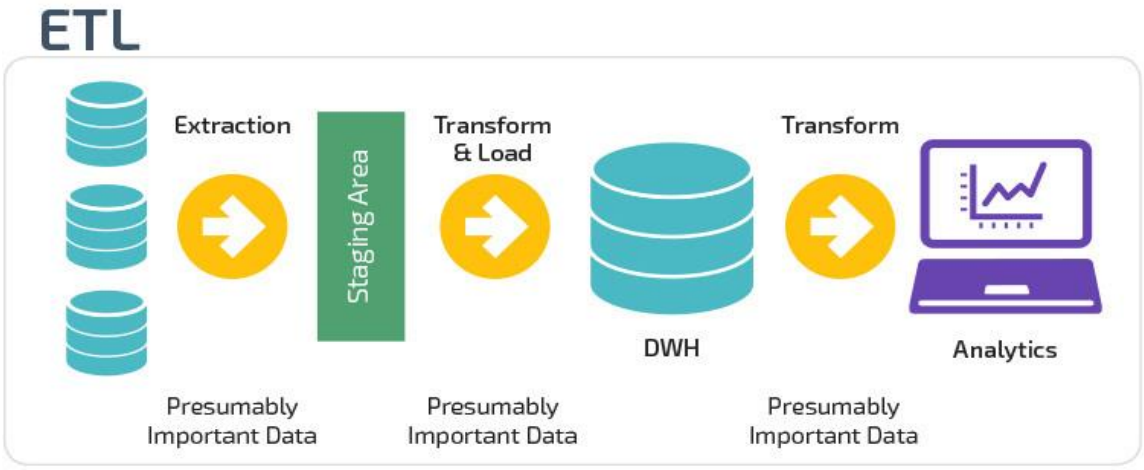

Fig. 3.1 ETL System Architecture

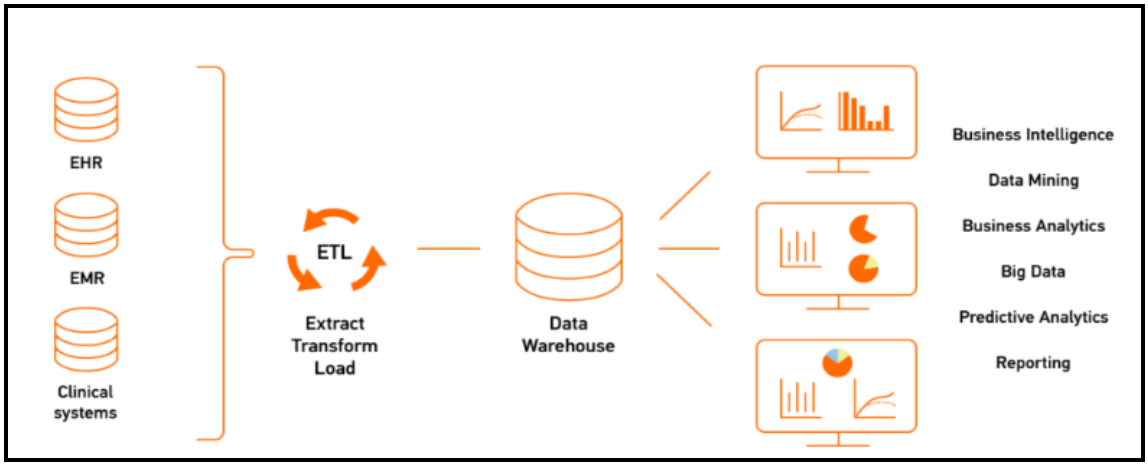

Fig. 3.2 ETL System Architecture

ETL systems are often used to move data between OLTP systems and data warehouses, or even between different data warehouses. The complexity of this architecture arises from merging data from different sources, which often have unique formats and structures.

7. Problems with the Current ETL System

Traditional ETL processes operate on periodic schedules (daily, weekly). As data warehouses grow in complexity and the need for more timely data increases, the traditional model becomes less viable [9]. Current ETL systems face several issues:

1. They are inefficient when source data changes infrequently.
2. The cost of recompilation is high, even when changes to base tables are minimal.

3. Latency, network traffic, and resource consumption increase significantly.
4. Bulk data transfers demand substantial system resources.
5. These processes are intrusive to source databases.

8. Proposed Approach – CDC (Change Data Capture)

To transfer changed data from source systems to a data warehouse, rather than loading all the source data and performing a full refresh, columns indicating the creation and modification dates of data rows are typically utilized. The process then imports only those rows that are new or modified since the last load date. This method allows for automated feeds of updated or new database records into the data warehouse [7].

Change data capture (CDC) is a data integration technique that identifies, captures, and delivers only the changes made to operational/transactional data systems. By processing just the changes, CDC optimizes the data integration process, especially the 'Extract' phase of ETL, and reduces the latency between the occurrence of changes in source systems and their availability to business users in the data warehouse.

Modern Data Integration and ETL tools must incorporate CDC, a technology that identifies, captures, and transfers only the changes made to enterprise data sources. Moving all source data is no longer viable. Implementing CDC dramatically enhances real-time data integration and ensures timely data delivery [7].

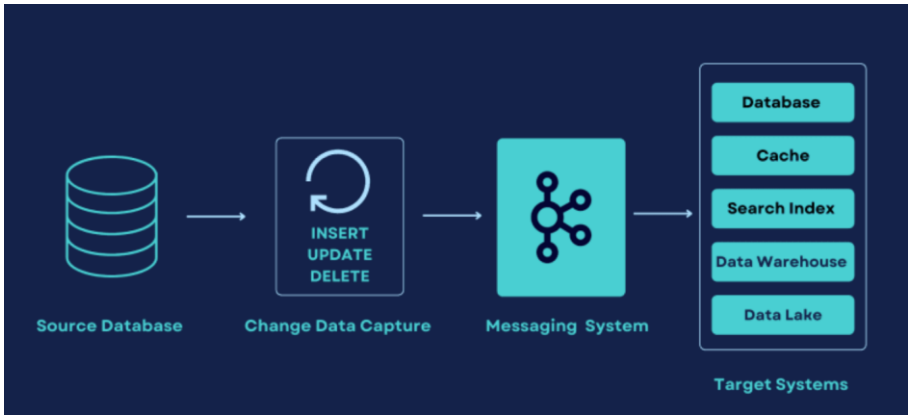

Fig. 4 Working of CDC in conjunction with ETL tools

CDC is frequently used in conjunction with ETL tools for more efficient data extraction in data warehouse implementations. Its main goal is to improve efficiency by minimizing the amount of data processed [14]. If only certain changes need to be captured based on business requirements, it is wasteful to transfer all changes. Advanced CDC solutions provide filters to limit the information transferred, further optimizing resources and enhancing speed and efficiency.

A. CDC Methodologies

CDC Methods CDC mechanisms can be established through various approaches:

1. Row Timestamps
2. Row Version Numbers
3. Row Status Indicators
4. Combination of Time/Version/Status
5. Database Table Triggers
6. Database Transaction Log Files

9. CONCLUSIONS

This paper presents a new and advanced approach to data integration, one that specifically identifies and transfers only the changes or differences from operational systems, rather than moving the entire dataset. This method addresses the common challenge of efficiently managing incremental data loads from various sources into a data warehouse or data mart. By focusing on just the change data, it eliminates the need for costly and time-consuming full data refreshes or cumbersome snapshot comparisons. The proposed approach ensures that changes are processed in real time, significantly improving the efficiency of the ETL (Extract, Transform, Load) process. It also allows for the timely delivery of updated and relevant information to end-users, all while keeping operational costs at a minimum. This includes lowering CPU cycles, storage demands, and network bandwidth usage, as well as reducing the need for extensive human resources. By streamlining the process, the approach minimizes system impact and ensures more timely, consistent, and accurate analysis of updated data, providing a scalable and cost-effective solution for data-driven businesses.

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.

BIOGRAPHIES

Robin Verma is an Expert in Data Engineering, Analytics and data Science. He is an active contributor in the industry to educate innovators and engineers with new ideas and possibilities in his expertise and have contributed to many real time data engineering projects across the globe. His innovative ideas and experience have been acknowledged, referred, and implemented by other experts in their fields and it significantly helped the wider industry to solve toughest business problems.