The Benefits of Change Data Capture in Enhancing Data Availability in the Digital Transformation Era

Azizah Zakiah[1], Rahadian Yusuf[2] and Ary Setijadi Prihatmanto[3]

[1]Informatics Engineering, Faculty of Engineering, Widyatama University, Cibeunying Kidul 204A, Bandung 40125
[1]Doctoral Program School of Electrical Engineering and Informatics, Institut Teknologi Bandung, Jalan Ganesha 10, Bandung, 40132, Indonesia
[2,3]School of Electrical Engineering and Informatics, Institut Teknologi Bandung, Jalan Ganesha 10, Bandung, 40132, Indonesia
[1]azizah.zakiah@widyatama.ac.id,[1]33222011@mahasiswa.itb.ac.id,[2]yrahadian@itb.ac.id,[3]ary.setijadi@itb.ac.id

Abstract. This research aims to analyze the key benefits of implementing Change Data Capture (CDC) in enhancing real-time data availability, particularly in the context of digital transformation. CDC enables instant data updates, which play a crucial role in reducing data latency and improving organizational operational efficiency. One of the primary impacts of CDC is the reduction of reliance on traditional batch processing, which typically causes delays in data access and excessive resource usage. Additionally, CDC minimizes the load on source systems by only capturing and transferring changes, thus enhancing data processing efficiency. The implementation of CDC also supports consistent data synchronization across various systems and databases, which is vital for organizations with distributed operations. The findings of this research indicate that CDC not only improves real-time data availability but also accelerates decision-making and ensures better operational efficiency within organizations.

Keywords: Change Data Capture, CDC, Digital Transformation, ETL, Data Streaming.

1 Introduction

In the era of digital transformation, data has become a critical asset for organizations across all sectors. The ability to access and process real-time data is increasingly important for operational efficiency, decision-making, and gaining a competitive edge. However, maintaining data consistency and availability across various systems in real time remains a significant challenge. This is where Change Data Capture (CDC) comes into play. CDC is a technology used to identify and capture changes made to a database, enabling businesses to synchronize data efficiently without the need for full database replication. It ensures that changes are tracked continuously, providing up-to-date information across multiple systems.

V. Mardiansyah and B. A. Prasetyo (eds.), *Proceedings of the Widyatama International Conference on Engineering 2024 (WICOENG 2024)*, Advances in Engineering Research 252,
https://doi.org/10.2991/978-94-6463-618-5_32

One of the key benefits of CDC is its ability to enhance data availability. Data availability refers to the extent to which data is accessible and usable at the right time. In a world where digital operations run 24/7, ensuring real-time access to updated data is crucial. For instance, businesses that rely on customer data, inventory management, or financial records need immediate access to any changes made within their systems to make informed decisions. CDC enables organizations to maintain this level of data integrity and accessibility without disrupting ongoing operations.

Additionally, CDC improves operational efficiency by reducing the need for complex batch processing. Traditional data integration methods often involve running batch jobs at set intervals, which can lead to delays in data updates. With CDC, these delays are minimized, as it continuously monitors and updates changes. This not only enhances performance but also optimizes resource utilization, as organizations can focus on incremental updates rather than full data loads. Moreover, CDC supports the implementation of real-time analytics, empowering businesses to respond proactively to market trends and customer behavior.

As organizations continue to embrace cloud computing and big data technologies, the demand for CDC solutions grows. Cloud environments, in particular, require continuous data synchronization across distributed systems. CDC offers a scalable solution for this, allowing data migration and replication to occur in real-time without causing downtime. In a study conducted by Wang et al. (2021), organizations using CDC to support cloud migration projects experienced a significant reduction in data transfer times and improved overall system performance. This demonstrates the role of CDC in not only maintaining data availability but also in enabling smoother transitions to cloud-based infrastructures.

Change Data Capture is an essential technology for organizations looking to improve data availability in the digital transformation era. Its ability to provide real-time updates, enhance operational efficiency, and support cloud-based infrastructures makes it a valuable tool for modern data management strategies. As data continues to grow in volume and complexity, implementing CDC will be crucial for businesses seeking to maintain a competitive edge [1]. Future research should explore the integration of CDC with emerging technologies like artificial intelligence and machine learning to further enhance its capabilities.

In the era of digital transformation, many organizations face significant challenges in ensuring real-time data availability and consistency across various systems they use. The reliance on traditional methods, such as batch processing, often leads to delays in data updates, negatively impacting decision-making, operational efficiency, and the organization's ability to adapt to rapid changes in the business environment. Additionally, with the increasing adoption of cloud computing and big data, there is a need for solutions that can efficiently manage large volumes of data and maintain synchronization across systems without causing disruptions or downtime. While

Change Data Capture (CDC) offers promising solutions to these challenges, many organizations still do not fully understand its implementation benefits and how this technology can be integrated with modern infrastructures to enhance data availability optimally.

The objectives of this research are to analyze the key benefits of implementing Change Data Capture (CDC) in enhancing real-time data availability in the context of digital transformation and to identify the positive impacts of CDC on organizational operational efficiency, particularly in reducing reliance on traditional batch processing

2 Related Work

Log-based Change Data Capture (CDC) is an effective technique for monitoring and recording data changes in real-time data warehousing. This method entails scanning transaction log files to capture modifications without altering the structure of the source system [2]. It can be utilized across various database systems, including both traditional relational databases and NoSQL document stores [3]. Log-based CDC provides several benefits, such as minimal disruption to source system performance, enhanced data freshness, and facilitation of real-time decision-making analysis [2]. Recent studies have aimed at optimizing the Extract, Transform, and Load (ETL) process through CDC, allowing for near-real-time data updates [4] and minimizing unnecessary ETL executions [5]. Although challenges exist, such as ensuring the reliability of log file readings, log-based CDC continues to be a viable solution for efficient data integration and real-time data warehousing across diverse database systems.

Has conducted research on methodologies for structuring change tables in real-time Business Intelligence (BI) , which is essential for making data-driven decisions in today's environment. It emphasizes the smooth integration of Online Transaction Processing (OLTP) and Online Analytical Processing (OLAP) systems, underlining the significance of change tables in this integration. The study reviews various approaches for developing and maintaining change tables, including triggers, asynchronous triggers, and log data capture techniques, while assessing the benefits and drawbacks of each within different database contexts[6].

Change Data Capture (CDC) is a key technique for enhancing the efficiency of Extract, Transform, and Load (ETL) processes in data warehousing. Log-based CDC methods track transaction logs to capture and publish data changes, facilitating near real-time updates and optimizing ETL operations [5]. However, the performance of CDC approaches can differ based on the structure and type of data source. A workload-aware CDC framework that integrates trigger-based, timestamp-based, and log-based methods has been proposed to enhance service quality in on-demand data warehousing [7]. Experimental findings indicate that the best CDC approach varies

according to different data sources and structures [8]. While log-based CDC offers benefits such as real-time data access and efficient resource use, it is crucial to evaluate the specific needs and characteristics of the data environment when choosing the most suitable CDC strategy for supporting real-time data warehouse systems.

3 Benefits of Change Data Capture

Conventional Extract-Load-Transform (ETL) processes [9] are no longer adequate and must be redesigned to handle streaming data, heterogeneity, usability, extensibility (custom processing), and continuous validation. Striim is an innovative, end-to-end distributed platform for streaming ETL and intelligence, designed to facilitate the fast creation and deployment of streaming applications. Its real-time ETL engine is built from the ground up to allow both business users and developers to efficiently build and implement streaming applications [10].

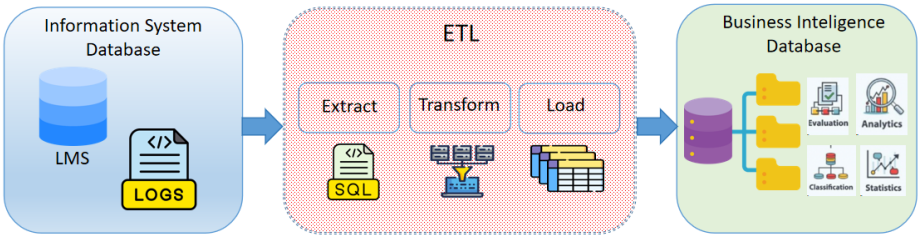

Fig. 1. Architecture of Conventional Extract-Load-Transform (ETL).

The execution of the ETL process is scheduled at specific times and in an environment separate from the production environment. As a result, data in the data warehouse is not available in real-time. Additionally, the scheduling operates without considering changes to the source data, meaning the ETL process is still executed even when there are no changes in the source data [5].

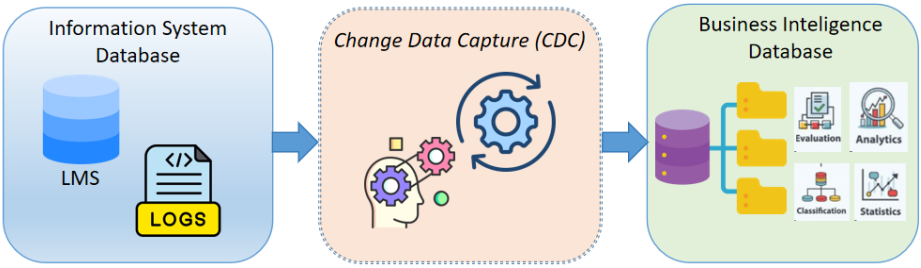

Fig. 2. Architecture of Change Data Capture (CDC)

The key benefits of implementing Change Data Capture (CDC) in enhancing real-time data availability are:

1) Real-Time Data Updates: CDC allows data to be updated instantly when changes occur, ensuring the data is always up-to-date without needing to wait for batch processes.
2) Minimized Impact on Source Systems: CDC reduces the load on source systems as it does not require reprocessing of the entire datasets. only the changes are captured and transferred.
3) Faster Decision-Making: With real-time data availability, organizations can make faster and more accurate decisions based on the most current information.
4) Improved Operational Efficiency: CDC eliminates the need to run time-consuming traditional ETL processes, speeding up data integration and enhancing efficiency.
5) Data Synchronization Across Systems: CDC ensures consistent data synchronization across various systems and databases, which is critical for distributed business operations.

The positive impacts of Change Data Capture (CDC) on an organization's operational efficiency, particularly in reducing reliance on traditional batch processing, include:
1) Reduced Data Latency: With CDC, data changes are captured and applied in real-time, eliminating the need to wait for batch processing cycles. This accelerates access to the latest data and improves operational speed.
2) Resource Savings: Traditional batch processing often requires significant resources in terms of time and computational power. CDC only captures significant data changes, reducing the resources needed to process large amounts of data.
3) More Efficient Data Processing: CDC allows for incremental data updates, avoiding the need to reprocess or transmitted entire datasets. This improves the efficiency of data flows within the system.
4) Increased System Uptime: With CDC, systems don't need to be paused or slowed down for batch processing, which is typically done at scheduled intervals. Operations can continue more smoothly without interruptions from large ETL processes.
5) Higher Responsiveness to Changes: CDC enables organizations to react quickly to data changes, enhancing flexibility in adapting to shifts in business or operational environments.

Acknowledgments. Thanks to Windyatama University for funding this conference.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

[1] M. Vaithianathan and D. Seenivasan, "Real-Time Adaptation: Change Data Capture in Modern Computer Architecture," *International Journal of Advancements in*

Computational Technology, vol. 1, pp. 49–61, 2023, doi: 10.56472/25838628/IJACT-V1I2P106.

[2] J. Shi, Y. Bao, F. Leng, and G. Yu, "Study on Log-Based Change Data Capture and Handling Mechanism in Real-Time Data Warehouse," in *2008 International Conference on Computer Science and Software Engineering*, IEEE, 2008, pp. 478–481. doi: 10.1109/CSSE.2008.926.

[3] K. Ma and B. Yang, "Log-based change data capture from schema-free document stores using MapReduce," in *2015 International Conference on Cloud Technologies and Applications (CloudTech)*, IEEE, Jun. 2015, pp. 1–6. doi: 10.1109/CloudTech.2015.7336969.

[4] E. Henke, Y. Peng, I. Reinecke, M. Zoch, M. Sedlmayr, and F. Bathelt, "An Extract-Transform-Load Process Design for the Incremental Loading of German Real-World Data Based on FHIR and OMOP CDM: Algorithm Development and Validation," *JMIR Med Inform*, vol. 11, 2023, doi: 10.2196/47310.

[5] F. M. Imani, Y. D. L. Widyasari, and S. P. Arifin, "Optimizing Extract, Transform, and Load Process Using Change Data Capture," in *2023 6th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, 2023, pp. 266–269. doi: 10.1109/ISRITI60336.2023.10468009.

[6] A. Morozova, L. Bielova, I. Meniailov, and V. N. Karazin, "Approaches to Organization of Change Tables for Real-Time Business Intelligence," 2023.

[7] W. Qu, X. Liu, and S. Dessloch, "A Workload-Aware Change Data Capture Framework for Data Warehousing," 2021, pp. 222–231. doi: 10.1007/978-3-030-86534-4_21.

[8] H. Chandra, "Experimental Results on Change Data Capture Methods Implementation in Different Data Structures to Support Real Time Data Warehouse," *Int J Bus Inf Syst*, vol. 1, no. 1, p. 1, 2020, doi: 10.1504/IJBIS.2020.10020890.

[9] F. de Assis Vilela, V. C. Times, A. C. de Campos Bernardi, A. de Paula Freitas, and R. R. Ciferri, "A non-intrusive and reactive architecture to support real-time ETL processes in data warehousing environments," *Heliyon*, vol. 9, no. 5, May 2023, doi: 10.1016/j.heliyon.2023.e15728.

[10] S. Smys, T. Senjyu, and P. Lafata, "Second International Conference on Computer Networks and Communication Technologies Lecture Notes on Data Engineering and Communications Technologies 44," May 2019. [Online]. Available: http://www.springer.com/series/15362