International Journal of Computer Trends and Technology
ISSN: 2231 – 2803 / https://doi.org/10.14445/22312803/IJCTT-V71I2P103

Volume 71 Issue 2, 15-18, February 2023
© 2023 Seventh Sense Research Group®

Original Article

An Implementation of Change Data Capture (CDC)

Hemadri Lekkala

IT Program Manager, IBM Consulting, India/USA

Received: 18 December 2022 Revised: 21 January 2023 Accepted: 02 February 2023 Published: 11 February 2023

Abstract *- This paper gives an overall picture of the implementation of Change Data Capture (CDC) from Mainframe Systems(z/OS) to downstream systems in a heterogeneous environment to reduce the latency to near real-time. This paper describes how the implementation can be done for data movement in terms of data extraction, load, and transformation to downstream systems. The data extraction captures the changed data based on the log file systems on Mainframes (z/OS). The processed data are pushed into the queue system on a first come basis, and ultimately, processed data loads into real-time Cloud based heterogeneous systems. The data Capture implementation framework supports data trigger and replication using standard data capture methods. This implementation supports larger organizations in making real-time decisions in terms of data for better analytics.*

Keywords *- Change Data Capture, Real Time Data Processing, Extract, Transform and Load, AWS Cloud Data Transformation.*

1. Introduction

The majority of enterprise Applications generally have relational databases as their persistent data store. Operations on these applications lead to creating, updating, and deleting data from tables within relational databases. Downstream applications look at change logs to discover the source systems' changes. The incremental data is essential in processing very specific changes to business data, thereby avoiding the full processing of entire data within operational databases.

1.

A new era of applications provides publish and subscribe-based services for publishing the changes for the relevant business data. Initial data loads must be performed before sending the captured changes from source systems. However, some legacy application database replication tools cannot be used without needing substantial changes to legacy applications. Downstream systems are then forced to load entire content from the Operational database to process and generate incremental meaningful information as part of overnight batch jobs. This approach increases overall batch execution time and increases strain on infrastructure due to additional load processed on a period (daily / weekly / monthly) basis.

A custom change data capture solution is needed to capture the data via a change log pointing to specific change events generated by the source system on z/OS using IMS as a database and transform the data into near real-time to the downstream consumers. Latency has been reduced from days to near real-time with the proposed solution.

2. Change Data Capture

This white paper's purpose is to provide a solution for a custom Change Data Capture (CDC) Application for identifying daily incremental changes on legacy/mainframe systems. The main objective of this implementation is to provide a real-time data replication solution to make IMS data on z/OS available to modernized applications running on web or mobile-based platforms. COBOL programs on z/OS are currently used to extract IMS data and transfer it to databases running on remote servers.

CDC tool captures database changes to the z/OS IMS database segments, and that data can be published and landed on target database tables. The CDC data is very application-centric and low-level. The implementation created a foundation to migrate from batch data processing to near real-time data pipelines (PostgreSQL) and APIs, one of the key characteristics of modern data architecture.

3. Solution Overview

A solution has been proposed to get the near real-time data from Mainframe z/OS using IMS relational database. Configuration must be completed via Classic Data Architect (CDA) and Management Console for the initial data load from source systems to target systems using the "refresh" method with the CDC tool. Then capture the data for the changes on the source systems by using "continuous mirroring" and publish the data to downstream consumers on a continuous basis".

The following diagram, Fig 3. gives a solution overview of a real-world implementation of Change Data Capture from z/OS to PostgreSQL Database. The same design approach can be used for any other database application; however, the drivers may need to be installed as per target and source systems. Below are the solution components. Please see Fig 3 below for details.

- IBM Infosphere Classic CDC for IMS
- IBM Infosphere CDC for FlexRep
- IBM Infosphere Classic Data Architect

3.1. Source System

Source Operational System provides business data to downstream systems. The operational System has a relational database referred to as the Source system in the above diagram. The application on the source system is on mainframes with multi technologies like IMS, COBOL, MQ, etc. The data on the mainframe (z/OS) is predominantly on the IMS parent-child relationship.

3.2. Target System

The consumers of mainframe data would be published to the PostgreSQL database, which is on the AWS cloud environment.

3.3. Initial load

Data capture subscriptions for the respective IMS data segments to primary PostgreSQL using a method called “refresh”.

3.4. Incremental Load

Once the initial load is completed onto PostgreSQL, then the incremental changes will be captured thru CDC and published to respective tables on the primary PostgreSQL database.

4. Detailed Design and Data Flow

Following the flow diagram, **Fig 4** gives a rough view of the flow of activities conducted as part of the execution of the Custom CDC Application. Data related to ADDS, INSERTS, UPDATES and DELETES have been captured. Custom CDC Application commits all changes to Change Database. Please see Fig4 for details on the data flow.

5. CDC Techniques

There are several methods/techniques available for handling the change data capture process depending on the source system’s data. Most of the database systems have a transaction Log file that records all the changes and modifications for each transaction at the source systems. The operational, transactional database will not get the effect when CDC uses a transactional log. The “Refresh” method would be used to get the initial load from Mainframe IMS applications to downstream systems via built-in transformation. This process is called “Initial Load”. The subsequent changes for the transactions would be loaded as an incremental load to the downstream consumers. This process is called “Mirroring”.

6. Downstream Systems / PostgreSQL

6.1. PostgreSQL Primary

Initial load and subsequent/incremental changes are loaded to PostgreSQL on AWS Cloud. PostgreSQL is the primary database to be used to send the data to various downstream consumers. This database will be appended with a new copy of the source database as soon as the change data capture application completes its processing on mainframe / mid-range systems.

6.2. PostgreSQL Secondary / READ replica

PostgreSQL database (N+1) gets the sync copy of the Source System as a read replica. This read replica database will serve as a secondary database and be used to verify the checks and balances of the database in terms of record counts.

7. Subscriptions

Table mapping will be completed using subscriptions. IMS table mappings are completed using custom or ordinal positions with respective target table and column mappings. User exists added as some of the data loaded as it is in EBCDIC (Extended Binary Coded Decimal Interchange Code) format. Data type conversion must be done before publishing the data to target PostgreSQL tables.

8. Monitoring and Error Handling

8.1. Monitoring Subscriptions

Subscription and data monitoring can be done thru the Monitoring tab in IBM Data Replication Management Console. Management Console is an administration application that allows you to configure and monitor replication. Management Console allows you to manage replication on various servers, specify replication parameters, and initiate refresh and mirroring operations from a client workstation. After defining the data that will be replicated and starting replication, you can close Management Console on the client workstation without affecting data replication activities between source and target servers. Custom CDC Solution will abort its processing and will not commit any changes to Change Database if any error is encountered either during the comparison of the latest and prior copy of the Source System (operational database) or during recording changes in the respective Change databases.

8.2. Error Handling

Subscription and event failures are notified using an event handling mechanism. Custom notifications can be set up in the event of subscription failures or data store failures. An automated email will be generated when the notifications are configured.

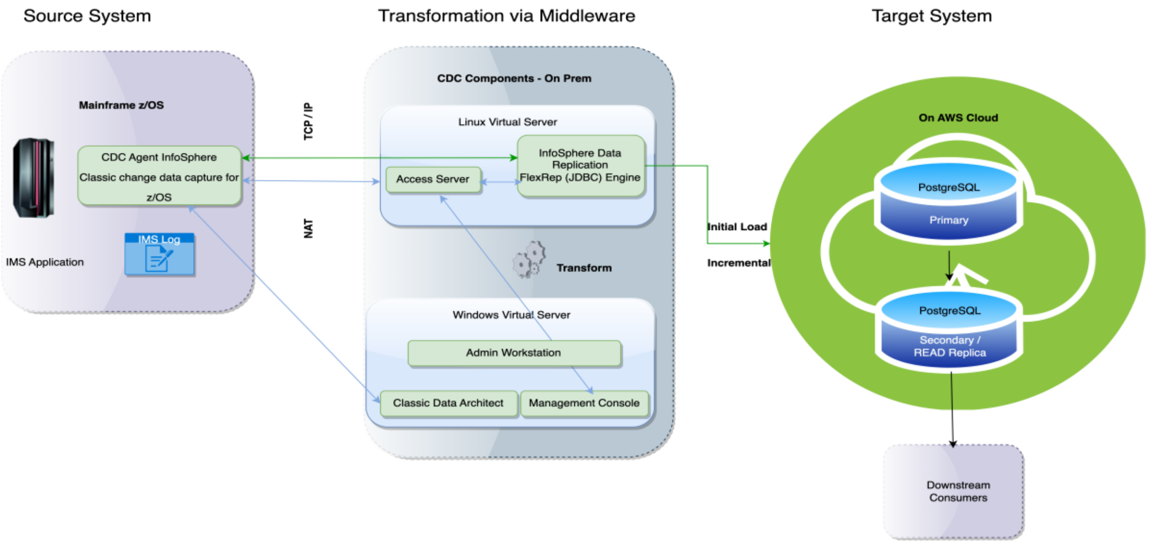

Fig. 3 Overall Architecture

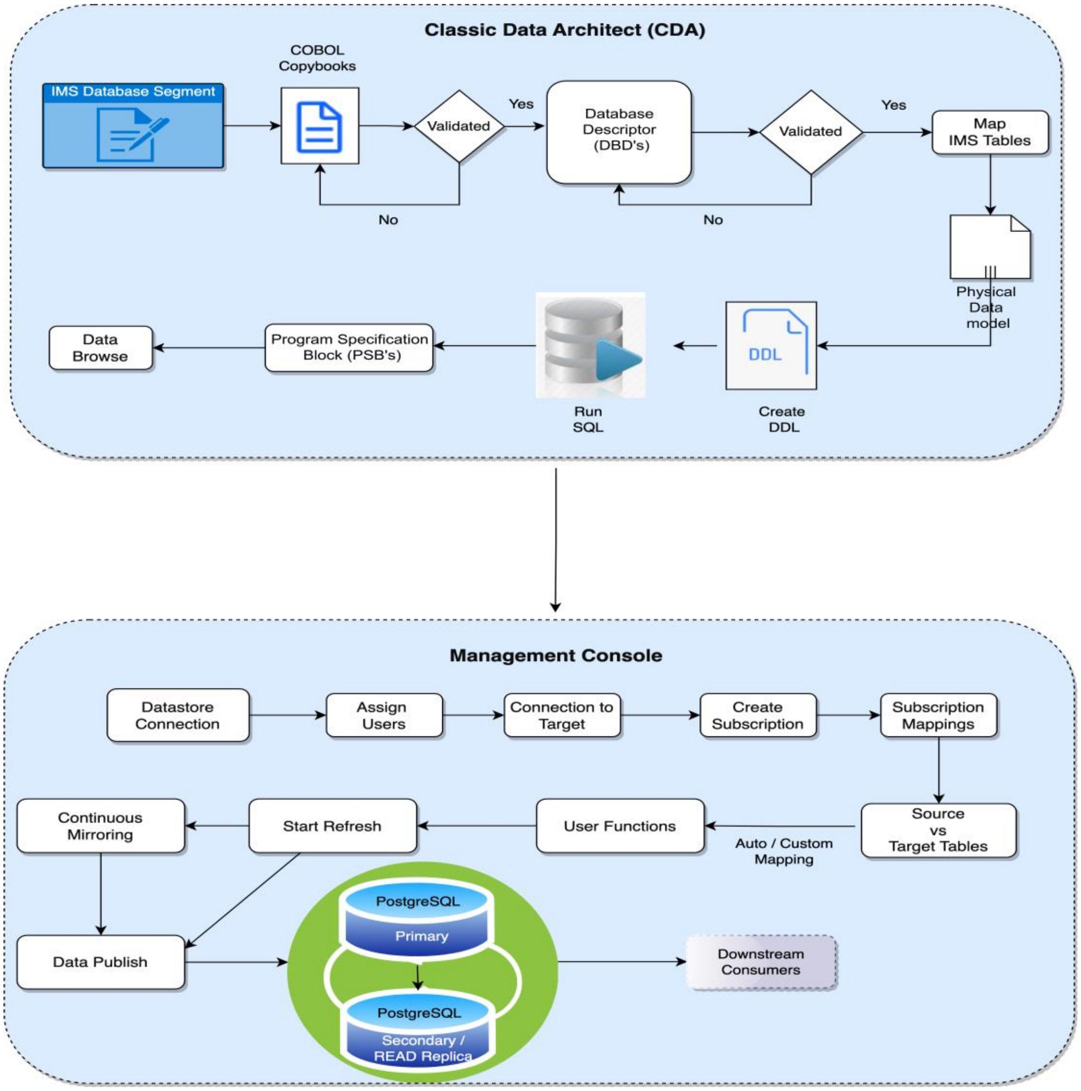

Fig. 4 Process Flow

Conflicts of Interest

I have declared (s) that there is no conflict of interest regarding the publication of this paper.”

Appendix

Please see the pictures below for the respective sections.

Conclusion

Change Data Capture (CDC) can be used as a tool to reduce data processing in Legacy Applications and share the data changes with Target systems in real time. It can be used to create a data pipeline in streaming services.

Acknowledgements

I would like to thank the following great reviewers who helped me in reviewing the content of the document and for their valuable suggestions.

Prajyot Dhaygude, Associate Delivery Partner, IBM Consulting.

Sandeep Sahu, Technical Manager, IBM Consulting.

References

[1] Website, 2022. [Online]. Available: https://www.ibm.com/docs/en/idr/11.4.0?topic=change-data-capture-cdc-replication

[2] V. Amarnath et al., "Review on Energy Efficiency Green Data Centers," *International Journal of Recent Engineering Science,* vol. 5, no. 2, pp. 21-26, 2018. *Crossref,* https://doi.org/10.14445/23497157/IJRES-V5I2P105

[3] Jatti Mounika, and Nagaveni B. Birada, "P&ID Data Analysis using Hadoop Map Reduce," *International Journal of Recent Engineering Science*, vol. 5, no. 1, pp. 8-10, 2018. *Crossref*, https://doi.org/10.14445/23497157/IJRES-V5I1P102

[4] Nisar Ahmed et al., "How to Protecting Kernal Code and Data," *International Journal of Computer & organization Trends (IJCOT)*, vol. 5, no. 4, pp. 20-23, 2015. *Crossref,* https://doi.org/10.14445/22492593/IJCOT-V23P301

[5] Abhishek Gupta et al., "Information Assurance via Big Data Security Analytics," *International Journal of Computer & organization Trends (IJCOT)*, vol. 5, no. 2, pp. 85-91, 2015.

[6] Khaled Elbehiery, and Hussam Elbehiery "Cloud Computing Technologies; Principals and Fundamentals," *International Journal of P2P Network Trends and Technology,* vol. 9, no.2, pp. 1-5, 2019. *Crossref,* https://doi.org/10.14445/22492615/IJPTT-V9I2P401

[7] Dr. Khalaf Khatatneh, Osama Nawafleh, Dr.Ghassan Al-Utaibi, "The Emergence of Edge Computing Technology over Cloud Computing," *International Journal of P2P Network Trends and Technology*, vol. 10, no. 2, pp. 1-5, 2020. *Crossref,* https://doi.org/10.14445/22492615/IJPTT-V10I2P401

[8] Dr. P.K.Rai, Rajesh Kumar Bunkar, "Architectural Data Security in Cloud Computing," *International Journal of Computer & organization Trends (IJCOT)*, vol. 4, no. 4, pp. 23-27, 2014.

[9] N. Madhavi Latha, and Y.Pavan Narasimha Rao, "A Novel Secured Data Transmission Model with Load Balancing for Cloud Computing," *International Journal of Computer & organization Trends (IJCOT)*, vol. 6, no. 1, pp. 45-50, 2016. *Crossref,* https://doi.org/10.14445/22492593/IJCOT-V29P304

[10] Dr. D.Shravani, "Model Driven Architecture based Agile Modelled Layered Security Architecture for Web Services Extended to Cloud, Big Data and IOT," *International Journal of Computer & organization Trends (IJCOT)*, vol. 6, no. 4, pp. 55-64, 2016.

[11] Konna Sirisha, and E. Deepthi, "Provide Privacy of Shared Data and Calculate Performance of Data Management System in Multiple Clouds," *International Journal of Computer & organization Trends (IJCOT)*, vol. 8, no. 1, pp. 24-28, 2018.

[12] Suri Nilima, G. Sivalakshmi, "An Improved Signature Schema for user Authentication and Privacy of Shared Data for Dynamic Groups in Cloud", *International Journal of Computer & Organization Trends (IJCOT)*, vol. 8, no. 2, pp. 39-44, 2018.

[13] Khaled Elbehiery, and Hussam Elbehiery, "5G as a Service (5GaaS)," *SSRG International Journal of Electronics and Communication Engineering*, vol. 6, no. 8, pp. 22-30, 2019. *Crossref,* https://doi.org/10.14445/23488549/IJECE-V6I8P104

[14] Khaled Elbehiery, and Hussam Elbehiery, "Millennial National Security's Cornerstones 5G, Cloud Technology, and Artificial Intelligence," *SSRG International Journal of Electronics and Communication Engineering,* vol. 6, no. 8, pp. 44-54, 2019. *Crossref,* https://doi.org/10.14445/23488549/IJECE-V6I8P107

[15] F. Twinkle Graf, and P. Prema, "Secure Collaborative Privacy in Cloud Data With Advanced Symmetric Key Block Algorithm," *SSRG International Journal of Computer Science and Engineering*, vol. 2, no. 2, pp. 40-44, 2015. *Crossref,* https://doi.org/10.14445/23488387/IJCSE-V2I2P109

[16] S. Manjunatha, L. Suresh, "A Study on Consolidation of Data Servers in Virtualized Cloud Atmosphere," *SSRG International Journal of Computer Science and Engineering*, vol. 6, no. 11, pp. 47-50, 2019. *Crossref,* https://doi.org/10.14445/23488387/IJCSE-V6I11P110