Middleware ETL with CDC based on Event Driven Programming

I Gede Adnyana[1], Made Sudarma[2], and Wayan Gede Ariastina[3]
[1]Department of Electrical and Computer Engineering, Post Graduate Program, Udayana University.
[2,3]Department of Electrical and Computer Engineering, Udayana University
*Email: adnyana.nakkuta@gmail.com

Abstract— To achieve the real time data warehouse is strongly influenced by the process in the data warehouse known as Extract, Tranform, Loading (ETL). One way to optimize the ETL process is processing only data that undergoes changes on the On Line Transaction Processing (OLTP) system. This technique is known as Change Data Capture (CDC) which is designed to maximize the efficiency of the ETL process. In this research middleware was built as a place where the ETL process will be carried out, transaction data from the OLTP system will be captured and sent directly to the middleware for further processing. The method used to capture changes in OLTP systems is Change Data Capture (CDC) based on Event Driven Programming, where this technique relies on events that occur in OLTP in capturing data changes. Functional testing is done by making a simulation of the insert and update processes in test applications namely OLTP CRM system.The results of the research obtained are (1) Change Data Capture (CDC) based on Event Driven Programming can capture changes in data that occur in OLTP CRM database;(2) ETL process to load data from Normalized Data Store (NDS) to data warehouse with Timestamp technique can load data that only undergoes changes that are processed to be loaded in Data Warehouse;(3) An increase in the amount of data that is processed has an effect on increasing processing time. Other factors that affect the value of process time are execution plan and cache memory.

Index Terms— change data capture, event driven programing, middleware, timestamp

I. INTRODUCTION

IN information management, a system is needed that combines humans, tools or technology, media, procedures and controls known as information systems. The application of information systems in a company can provide benefits, it can saving operational costs because of the automation that can accelerate work time. The information system can also store data and process it accurately so as to minimize the risk of human error.

Problems occur when information systems that manage data cannot facilitate the decision making process. The reason is that large volumes of data within a company are spread into different systems, multiple databases and diverse structures. This condition is getting worse by finding inconsistent data between various systems that are done partially [1]. Another difficulty that arises is when the report / information is integrated because data consistency between various source systems is difficult to achieve. Therefore, it is necessary to have an integrated data processing model that can process heterogeneous transactional data called Data Warehouse (DWH).

The current trend, a company needs the latest information in decision making. Data warehouse is required to provide information in the latest (real time). To achieve the real time data warehouse is strongly influenced by the process in the data warehouse known as Extract, Tranform, Loading (ETL). This process involves extracting data from outside sources, changing the data to suit operational needs and loading it as the final target. ETL's quality and performance are not the same, thus optimizing ETL processes for real time decision making is very important [2]. The information generated by the data warehouse through the ETL process is actually not exactly real time but still has a delay in processing. This time delay in the ETL process which is close to real time is what triggers the use of the term Nearly Real Time Data Warehouse (NRTDWH). One way to optimize the ETL process is processing only data that undergoes changes to the On Line Transaction Processing (OLTP) system. This technique is known as Change Data Capture (CDC) which is designed to maximize the efficiency of the ETL process.

In this research a middleware will be built as a place where the ETL process will be carried out, transaction data from the OLTP system will be captured and sent directly to the middleware for further processing. The method used to capture changes in OLTP systems is Change Data Capture (CDC) based on Event Driven Programming, where this technique relies on events that occur in OLTP in capturing data changes. In this research the data warehouse design was carried out in conjunction with the OLTP system design.

(p-issn: 2579-5988, e-issn: 2579-597X)

II. METHOD

A. Extract, Transform, Load (ETL)

Extract, Transform, Loading (ETL) is a very important process in the data warehouse, with ETL data from operations can be entered into the data warehouse. ETL can also be used to integrate data with pre-existing systems.

The goal of ETL is to collect, filter, process and combine relevant data from various sources to be stored in the data warehouse. The result of the ETL process is the production of data that meets the data warehouse criteria such as historical, integrated, summarized, static data, and has a structure designed for the purposes of the analysis process [3].

B. Change Data Capture (CDC)

Change Data Capture (CDC) is an innovation approach to data integration, based on identification, capture, and sending changes made by source data. By processing only changes, CDC makes the process of data integration more efficient and reduces costs by reducing latency [4].

There are two CDC scenario models :

1) Batch-Oriented CDC (Pull CDC) : interesting set of data that only changes periodically in high or low frequencies
2) Live CDC (Push CDC) *:* is sending data changes to the ETL tool after changes occur. Can be done with the mechanism of event-delivery or messaging middleware.

C. Event Driven Programming

Event driven programming is a programming technique where all program execution flow is determined by an event. When the program starts, it will wait for input from the user. For each event that appears, the program will run the syntax to respond. Program execution flow is determined by the sequence of events that appear [5].

Event is a response from a program when a user takes action on a particular GUI (Graphical User Interface) in the application. In order for the GUI component to produce an event when there is interaction between the user and the GUI, such as pressing a button, moving the mouse, etc., a listener is needed to do it. With this listener, events generated from the GUI can respond from an interaction directed to a program instruction.

D. Incremental Extraction with Timestamp

This method focuses on the extraction process by extracting data that only undergoes changes to minimize the time to carry out the extraction process[6]. From the update process carried out in the extraction stage, the incremental extraction with timestamp method can minimize the update process compared to similar methods, namely incremental extraction with identity column.

In the incremental extraction with timestamp the update process is done on one data line to change the Last Successful Extraction Time (LSET) and Current Extraction Time (CET) values in the metadata table. While the incremental extraction method with identity column is made a change in status / flag, for example the status has not been processed to be processed in all new data extracted.

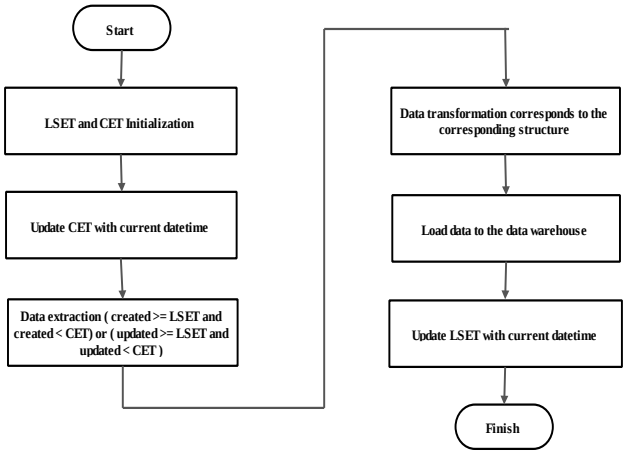


Fig 1. Flowchart of Incremental Extraction with Timestamp

The process begins with the Last Successful Extraction Time (LSET) and Current Extraction Time (CET) initialization. Then update the CET value to current datetime. Extract the latest data by taking data (created> = LSET and created <CET) or (updated> = LSET and updated <CET). If the latest data has been obtained, then the data transformation process is carried out according to the corresponding structure and the process of loading to the data warehouse. After the ETL process is complete, then update the LSET value to current datetime. Checking new data or data experiencing changes is done with a certain frequency that is set by the scheduler.

III. RESULT AND DISCUSSION

A. Insert Process on OLTP CRM Database

Testing in this phase was carried out to prove that the CDC process based on Event Driven Programming managed to capture changes in data made on the OLTP CRM system when the insert process was carried out and all changes to this data were successfully sent from the OLTP CRM database to the Normalized Data Store (NDS) database and from NDS database to the Data Warehouse (DWH) database so that the NRTDWH Customer Relation Management Information System was formed.

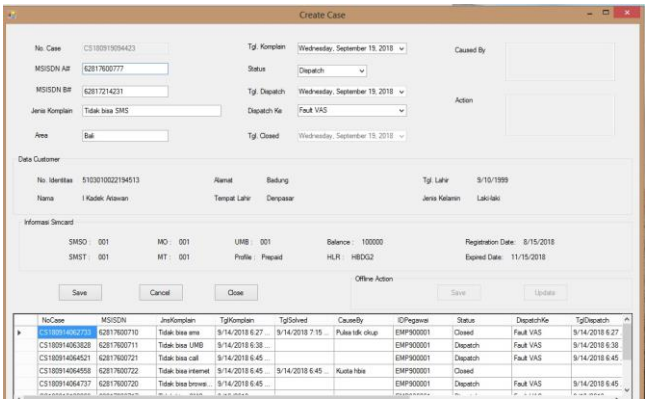


Fig. 2. Insert Process on Customer Complaint Form

On Fig. 2 there is a process of adding customer complaint data with customer number '62817600777'. This insert process adds new data to the TKomplain table can be seen in Fig. 3

NoCase	MSISDN	JnsKomplain	TglKomplain	TglSolved	CauseBy	IDPegawai	Status	DispatchKe
CS180919094423	62817600777	Tidak bisa SMS	9/19/2018 9:44 ...			EMP900001	Dispatch	Fault VAS

Fig. 3. Results of the Insert Process in the TKomplain Table

The insert process in the TKomplain table triggers the CDC Event Driven Programming process. When the create case button on the form is clicked, besides inserting the TKomplain table on the OLTP database but also inserting it into the NDS database. The results of adding new data to the TNDSKomplain table on the NDS database can be seen in Fig. 4

	Kuncikomplain	nocase	MSISDN	JnsKomplain	TglKomplain	DispatchKe	TglDispatch	Active
1	11	CS180919094423	62817600777	Tidak bisa SMS	2018-09-19 00:00:00.000	Fault VAS	2018-09-19 00:00:00.000	1

Fig. 4. Results of the Insert Process in the TNDSKomplain Table

Every 30 second interval, the scheduler on the middleware ETL module performs a new data check in the NDS database. If new data is found, data is added to the DWH database by loading new data contained in the NDS database. The results of adding new data to TFakta_Komplain table can be seen in Fig. 5

	Kuncikomplain	nocase	MSISDN	JnsKomplain	TglKomplain	DispatchKe	TglDispatch	Effective_Date	Active
1	11	CS180919094423	62817600777	Tidak bisa SMS	2018-09-19 00:00:00.000	Fault VAS	2018-09-19 00:00:00.000	2018-09-19 09:57:52.050	1

Fig. 5. Results of the Insert Process in the TFakta_Komplain Table

Calculation of process time is done by calculating the difference from initial data retrieval process time from the NDS database with the end time of the data process loaded in the DWH database. From the calculation results, the processing time for the insert is 30 milliseconds, shown in Fig. 6

	IdTransaksi	NamaTabel	JumlahInsert	JumlahUpdate	JumlahDelete	Start	Finish	Waktu proses
1	49	TNDSKomplain	1	0	NULL	2018-09-19 09:57:52.050	2018-09-19 09:57:52.080	30

Fig. 6. Time Calculation Results of the Insert Process

B. Update Process on OLTP CRM Database

Testing in this phase is carried out to prove CDC Event Driven Programming can capture changes that occur when the update process is performed on OLTP CRM system and all changes to this data were successfully sent from the OLTP CRM database to the NDS database and from the Normalized Data Store (NDS) database to the Data Warehouse (DWH) database so that the NRTDWH Customer Relation Management Information System was formed.

The TKomplain table update process is done in the update case form, this form is used to change the status of a case complaint. An update is made if a case has been resolved, the case is closed and given information about the cause and the action taken to solve the case of the customer complaint. The

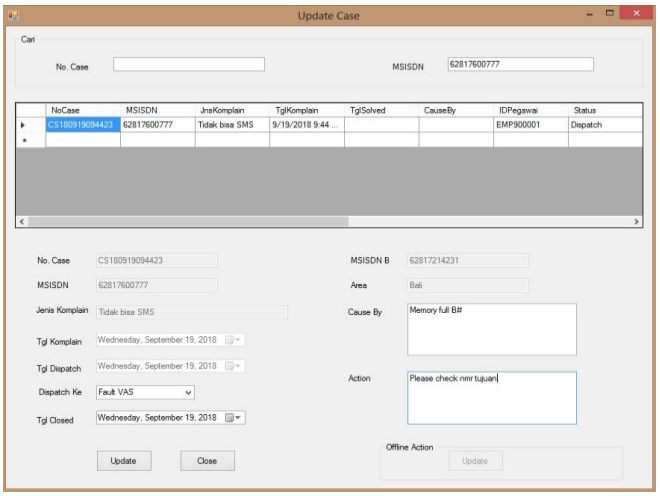


Fig. 7. Update Process on Case Update Form

TKomplain table update process can be seen in Fig. 7

In Fig. 7 there is an update process of customer complaint data with MSISDN '62817600777'. The case on the customer complaint has been solved so that the case update can be done by giving the case status to closed. The update

NoCase	MSISDN	JnsKomplain	TglKomplain	TglSolved	CauseBy	IDPegawai	Status
CS180919094423	62817600777	Tidak bisa SMS	9/19/2018 9:44 ...	9/19/2018 10:06...	Memory full B#	EMP900001	Closed

Fig.8. Results of the Update Process in the TKomplain Table

process in the TKomplain table is as shown in Fig. 8

The update process in the TKomplain table triggers CDC Event Driven Programming to capture data changes in the status column. The CDC process then changes the data in the TNDSKomplain table in the NDS database. The update process on TNDSKomplain tables is done by updating the active status of the old data and inserting new data according to the data that changes. The results of updates to the TNDSKomplain table in the NDS database can be seen in

	Kuncikomplain	nocase	MSISDN	JnsKomplain	TglKomplain	TglSolved	Status	Active
1	11	CS180919094423	62817600777	Tidak bisa SMS	2018-09-19 00:00:00.000	NULL	Dispatch	0
2	12	CS180919094423	62817600777	Tidak bisa SMS	2018-09-19 09:44:23.000	2018-09-19 10:06:20.350	Closed	1

Fig. 9. Results of the Update Process in the TNDSKomplain Table

Fig. 9

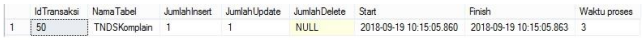
Every 30 second interval, the scheduler on the middleware ETL module checks for new data changes in the NDS database. If new data is found, data changes are made to the DWH database by loading new data contained in the NDS database. The results of adding new data to TFakta_Komplain table can be seen in Fig. 10

Calculation of process time is done by calculating the difference from initial data retrieval process time from the NDS database with the end time of the data process loaded

	Kuncikomplain	nocase	MSISDN	JnsKomplain	TglKomplain	TglSolved	Status	Effective_Date	Active
1	11	CS180919094423	62817600777	Tidak bisa SMS	2018-09-19 00:00:00.000	NULL	Dispatch	2018-09-19 09:57:52.050	0
2	12	CS180919094423	62817600777	Tidak bisa SMS	2018-09-19 09:44:23.000	2018-09-19 10:06:20.350	Closed	2018-09-19 10:15:05.860	1

Fig. 10. Results of the Update Process in the TFakta_Komplain Table

in the DWH database. From the calculation results, the processing time for the update is 3 milliseconds, shown in



	IdTransaksi	NamaTabel	JumlahInsert	JumlahUpdate	JumlahDelete	Start	Finish	Waktu proses
1	50	TNDSKomplain	1	1	NULL	2018-09-19 10:15:05.860	2018-09-19 10:15:05.863	3

Fig. 11. Time Calculation Results of the Update Process

Fig. 11

An increase in the amount of data that is processed has an effect on increasing processing time. Other factors that affect the value of process time are execution plan and cache memory. When the same query is run and as long as the execution plan is still stored in memory it can shorten the processing time.

IV. CONCLUSION AND FUTURE WORK

Change Data Capture (CDC) based on Event Driven Programming can capture changes in data that occur in OLTP CRM database that are triggered by insert and update activities on OLTP CRM system. Changes to data captured by the CDC are then loaded into the Normalized Data Store (NDS) database.

The ETL process for loading data from the Normalized Data Store (NDS) to the Data Warehouse with the Timestamp technique can load data that only undergoes changes that are processed to be loaded in NRTDWH. The ETL process with the Timestamp technique is run with a certain time interval set in the scheduler.

An increase in the amount of data that is processed has an effect on increasing processing time. Other factors that affect the value of process time are execution plan and cache memory

In future research, query optimization techniques can be developed in the ETL process so that it can shorten ETL processing time to be created Nearly Real Time Data Warehouse (NRTDWH).

REFERENCES

[1] Ponniah, Paulraj. 2001. Data Warehousing Fundamentals. John Wiley & Sons, Inc. USA.

[2] Kakish, Kamal dan Kraft, A.T. 2012. ETL Evolution for Real-Time Data Warehousing. Proceedings of the Conference on Information Systems Applied Research. ISSN: 2167-1508.

[3] Kimball, R., Caserta, J. 2004. *The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleanin*. Canada : John Wiley & Sons Inc.

[4] Attunity Ltd. 2004. Next Generation ETL. Attunity White Paper.

[5] Perkins, Hal dan Hotan, Michael. 2013. GUI Event Driven Programming. Software Design & Implementation Slides.

[6] Inmon, W.H .2005. Building the Data Warehouse 4th Edition. John Wiley & Sons, Inc. USA.