# Change Data Capture on OLTP Staging Area for Nearly Real Time Data Warehouse base on Database Trigger

I Made Sukarsa
Departement of
Information Technology
Faculty of Engineering
Udayana University,
Bali, Indonesia

Ni Wayan Wisswani
Departement of
Informatic Manajemen
Politeknik Negeri Bali,
Bali, Indonesia

I K. Gd. Darma
Putra
Departement of
Information Technology
Faculty of Engineering
Udayana University,
Bali, Indonesia

Linawati
Departement of
Electrical Engineering
Faculty of Engineering
Udayana University,
Bali, Indonesia

## ABSTRACT

A conventional data warehouse use to produce summary from an organization information system in a long time period. This condition will make the management unable to get the most up to date data every time it needed. Therefore a nearly real time data warehouse which will manage the ETL process with a more compact data and a shorter period is needed.

The design of nearly real time data warehouse in this research is implemented in two steps. The first step is done by data collection technique modeling to make a more compact ETL data managed. This step is done by putting the staging area on an Online Transactional Processing (OLTP). It can minimize the failure of data movement process from the OLTP to the staging area. Besides that, the CDC method is also had applied on the OLTP. This method will be implemented with a trigger active database. The trigger will capture of the data changing on the OLTP, transform it and then load it to the staging area in one time. The second step is the synchronization process of the data movement from the staging area to the nearly real time data warehouse. This process is done by mapping the movement which is ran by the SQL Yog. The mapping result will accomplished by the windows task scheduler

## General Terms

Modelling System, Data Warehouse

## Keywords

Nearly real time data warehouse, Change Data Capture, Surrogate key, Trigger.

## 1. INTRODUCTION

Data warehouse is a need for an organization. Data warehouse (DWH) capable to be the data sources to all integrated report making process which are needed in prompting the decision making process. [1]. Data source from various OLTP processed through the various stages that consist of Extract, Transform and Loading (ETL). ETL is built on a tier that is placed between the source data and DWH and known also as a staging area[2]. Extract part relied on to take data from multiple sources within a specific time period to be taken to DWH. Data is cleaned, integrated and transformed into a specific format by the transform and then moved to the DWH by Loading component.

Conventional ETL machine will work on time variant. This machine will save the data periodically in accordance to the organization business process flow [3]. This characteristic made the DWH unable to give the most up to date information from every event on the transactional system. The fact is data

warehouse which is real time is really needed in decision making which is need the highest level of up to date information. [4].

Real time data warehouse will able to show the ETL working result in an exact time according to the transactional time on a number system [5]. But ETL as the core of data warehouse [6] cannot really work on real time [7]. This happens because of the ETL need some time to process the data from various sources in a large amount, and has to go through some communication component [8]. The delay time is needed by ETL to process this summary, which trigger the term Nearly Real Time Data Warehouse (NRTDWH) [7].

To produce NRTDWH, ETL therefore can be implemented by applying Change Data Capture (CDC) [9]. CDC is used to know the changing on the data sources and then capture it to be given to the database destinations which need it [10]. This ability made CDC able to capture data changing efficiently [11] therefore NRTDWH will be easier to be implemented.

Based on the above explanation, therefore the effort to create NRTDWH by CDC modeling becomes really important to be implemented.

## 2. RELATED WORK

Some researches on the development of CDC modeling and real time data warehouse have been done [12]. The modeling of CDC processes uses the log analysis while it introduces the architecture of semi real time DWH to make real time data warehouse by using the CDC mechanism which have owned by Oracle.

[10] Modeling of the data changing capture process by using a set of web service. Captured modeling use the web service is also done by [13] and to facilitate real time data warehouse is introduce an architecture of multi level real time data cache. Meanwhile [8] modeling of the ETL for real time DWH with using schedule algorithm to balance the query and updates thread control trigger based on ETL machine.

In our research will be develop a trigger based CDC modeling which will capture data changing on different sources system. The same trigger will transform the capture result in one time and then load it to the staging area which is placed on the OLTP.

The capture, transform and load (CTL) which has designed, made the DWH able to receive the data summary faster. It happen because the ETL process a smaller amount data and the CTL process result is the final data which is accordance to

the DWH structure. This condition made the synchronization process of the whole data sources to DWH doesn't need a more advanced transformation.

.

# 3. CAPTURE, TRANSFORM AND LOAD

## 3.1 CTL Framework

The CTL model architecture for NRTDWH which will be developed on this research is visualize like the following figure 1:
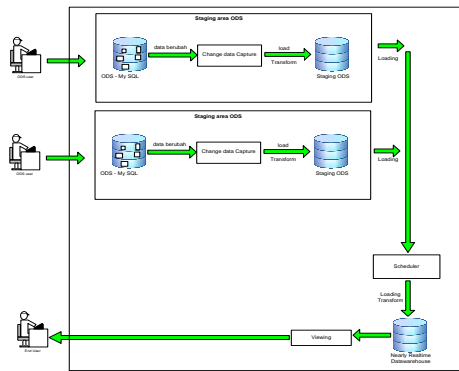


Figure 1. General architecture of the system

In this model, transform and load process will be conducted by each OLTP engine so as to reduce the time delay due to the staging area located at each OLTP and do not need to build a new staging area as in the models that already exist. The integration process has been completed on the OLTP so the data warehouse will receives the final data.

NRTDWH on this research is produced from the CTL process on different OLTP sources. This model is starting to work when a user enters new data, change or delete a record or some field on the OLTP.

Event insert, will make a trigger capture the inserted data and then save it as a new record on a table in staging area which is appropriate. An update to one or some field on a record, make a trigger captured the changing which is made. The result will be used to updating data or being save as a new record on a table accordingly on a staging area. On the other hand, if the deleted process happens, therefore deleted data will change some field on the active record in staging area. The delete

proscess can make a trigger inserted as a new data to the appropriate table on the staging area. CTL will work like figure 2, in the following
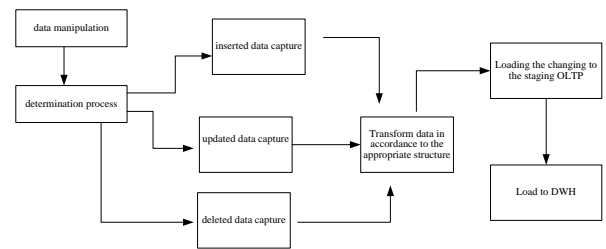


Figure 2. CTL process flow

When the transform of the capture result is done, trigger might do one of these processes:

1. Simple Transform Process. This process will do some field adjustment and formatting data between captured data with the structure on the staging area. This process happens if the information on the related topic on a staging area is the information which comes from one table and doesn't need relation with other table.
2. Leveled Transform Process. This process is completed with advance query joint operation process and other operation which has look up characteristic. This is done if the information comes from some tables on the OLTP.

All saved CTL process result on the staging area then move to NRTDWH by task scheduler based on the metadata mapping design. This metadata will be the basic rule to do join data from every OLTP sources to NRTDWH. In order to make the data warehouse easier to understand, therefore the data on data warehouse will be shown through a data mart application.

## 3.2 Dimensional Modelling

On this research, all of the OLTP uses the same MySQL platform database. OLTP will give the data that NRTDWH needed, while staging area will load the CTL results into dimensions and facts tables which are ready to be joined to NRTDWH. Through the figure 3, will be shown the star schema which will be put on each staging area on OLTP and the dimensional modeling on the data warehouse.
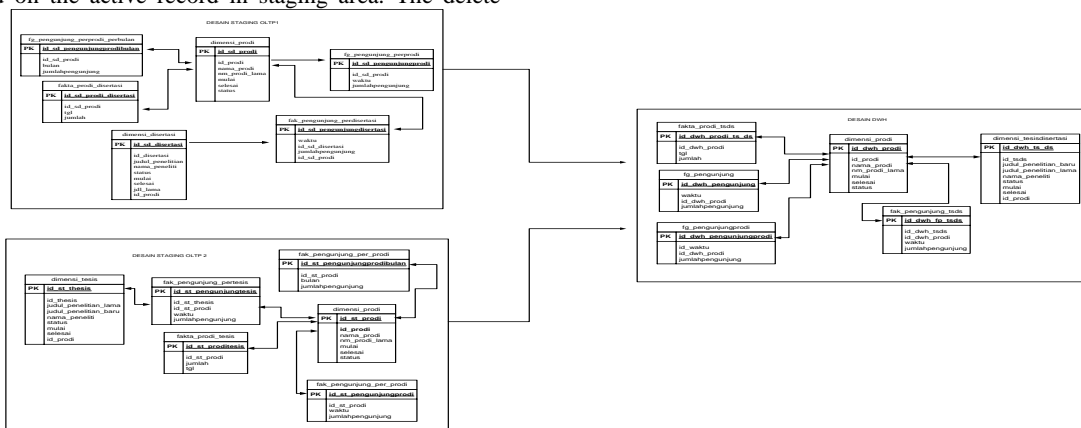


Figure 3. Dimensional Modeling

Even though comes from a different sources, the join data process from the staging area to NRTDWH, doesn't need an advanced transformation process to form a new surrogate key on every dimensions and facts. Even though so, all data on NRTDWH will be able to be differentiated. This happen because the surrogate key on this research is designed to keep the characteristic from OLTP source. The surrogate key model on this research is also able to prevent the failure of joining data process because of the same data.

## 3.3 Nearly Real time data warehouse

The effort to make the NRTDWH on this research is done by some way, which are:

a. Staging area design which unite with the OLTP database.

This is done to shorten the time of data capture changing process from the OLTP to the staging area. Therefore the transform process can be done immediately. This model is also to minimize the communication failure. It is because of the data source and target put on the same host. The staging area placement on the OLTP is also to make the synchronization process into the NRTDWH become easier. It is because of the whole data process is done in the OLTP, therefore all of the save data in the staging area are the final data in accordance to the structure which NRTDW wanted.

b. Shorten the data load time span to NRTDWH with a trigger.

The effort to shorten the load process is done by using a trigger. Trigger will make the capture can be done in a short time period of time if it is compare with the other CDC method. The shorter capture process surely will influence the time which is needed for the transform and load process on the staging area.

c. Join the Transform on the Change Data Capture

The CTL process which is done in one time by using the same trigger surely will minimize the delay between capture and transform. This will immediate the load process to the staging area, therefore the synchronization will also be organized to be shorter.

d. The use of a trigger, function and procedure as the transform engine.

On this research, all of the capture process, transform and load which take place will be run by PL/SQL trigger, function and procedure. Trigger is chosen because all the process will works faster and all daily transaction capable to work without disturbance. This happen because PL/SQL works on DBMS. Trigger also can be known events that make the record in accordance in OLTP changing. Therefore the changing data will have the transform process directly without comparison with previous data which are have already save on the DWH. This will help NRTDWH easier to achieve.

## 3.4 The Synchronization Process

The synchronization process is done by moving and joining the data processes result which is load on the staging area on each OLTP. This process consists of two main components. The first component will do the metadata mapping which will be done by SQL Yog Ultimate. The metadata will use as the basic rules when the synchronization process happened. All these mappings are saving on a job file which is different for every source. The second component is the scheduler which contains of the data moving time span schedule to DWH. This process will run the job file on a metadata scheme which has

made. The making scheduler is done by a windows operation system which is scheduled in every one minute.

## 3.5 Testing and Results

The testing of CDC modeling on this research use three testing application: the thesis system and the Dissertation system which act as an OLTP, and the data mart of Udayana university application. The testing is done by manipulating some data dummy which is spread on each OLTP. The data manipulation is done only to some tables on OLTP which might be the source of the DWH.

The testing on this research is done by two phase. The First phase is done to know that the CTL process on the staging area is done successfully. The second testing is done to prove that the synchronization from the staging area on each OLTP to NRTDWH is successfully done by the scheduler.

### 3.5.1 Capture, Transform and Load Testing

Trigger will do the CTL process before and after insert, update and delete happen on an OLTP. These manipulation processes will influence the facts and dimensions tables of each staging area. One of the CTL processes which will be observed is one of it the manipulation process of insert, update and delete on the th_thesis table. The insert process of th_thesis table through a form visualize on the following figure 4:
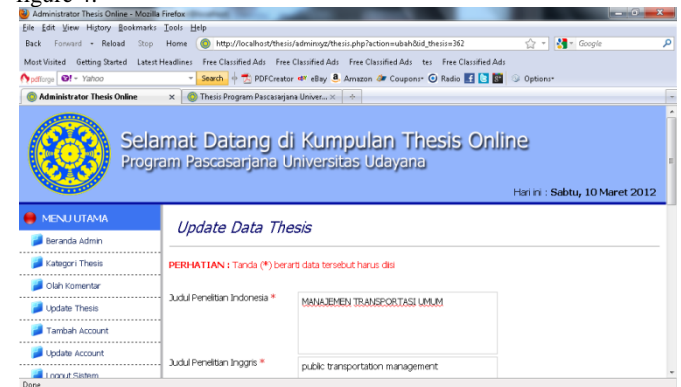


Figure 4: Insert Form of the thesis OLTP system.

When the insert happen through above form, the CTL on th_thesis table will work to insert a new row to the table dimension and fact on the staging area. It caused the dimension table will be like the figure 5.



Figure 5. Inserts data result to the thesis dimension

While the While the fact table pengunjung_pertesis will like the figure 6



Figure 6. Data insert result to the fact table pengunjung_pertesis

Other fact table is also influenced by this process is the prodi_tesis table. When CTL succeed therefore the table will be like figure 7.



Figure 7. Inserts result to prodi_tesis

The pengunjung_prodi fact table will also have some changes when the insert to th_tesis is done. The result of CTL process on this table will be like figure 8



Figure 8. The inserts result to pengunjung_prodi table on thesis OLTP system.

Insert to the th_thesis table will also influence the pengunjung_prodi_perbulan fact table. This process will caused the table changes like figure 9.



Figure 9. Inserts result to the table of pengunjung_prodi_perbulan on thesis system.

The update process which will influence the dimension and the facts is done by two means: First, update the th_thesis table which is done trough a form like the figure 10.



Figure 10. Updates Form of the thesis OLTP System.

Update on this form, is done to the name of the researcher field, the title of the research or the id_prodi field. This will trigger CTL to work and influence the dimension and facts table on the staging area. If the change happened only on the field of name and the title of the inputed data research, therefore CTL will caused the thesis dimension change like figure 11.



Figure 11. Update result of thesis dimension

If this changing is done on the id_prodi field, therefore prodi dimension will change like the figure 12.



Figure 12. Updates result of id prodi field on prodi dimension

The changing of id_prodi field, can influence the fact table on the staging area. The fact table which is change is the pengunjung_pertesis table. This changing will be shown on figure 13.



Figure 13. Update result pengunjung_per_thesis

After the CTL working, the prodi _tesis fact table will be like the figure 14.



Figure 14. Updates Result of the id_prodi field on the prodi_tesis table.

Because of these process, the pengunjung_prodi fact table will be like figure 15.



Figure 15. Updates result of pengunjung_prodi table of the thesis system

Then the changed of the other facts is fg_pengunjungprodibulan. It will change like figure 16.



Figure 16. Updates result of the pengunjung_prodi perbulan table on the thesis system.

The second update method to th_thesis is done through a form like figure 17.
.



Figure 17. Update Form on the lihat field on the table thesis.

User activity through this form, caused the value of lihat field which is save on the th_thesis table will change. This change caused CTL work, therefore the pengunjung_per_tesis table will be like figure 18 on the following.



Figure 18. Update result of the lihat field on the pengunjung_pertesis table

Other table will also change because of this process is the pengunjung prodi fact table. The results will look like figure 19.



Figure 19. Update Result of the pengunjung_prodi table.

CTL process which is trigger by the lihat field is also change the pengunjung_prodi_perbulan table. The changing on this table will be like figure 20.

Figure 20. Update result of the pengunjung_prodi_perbulan table.

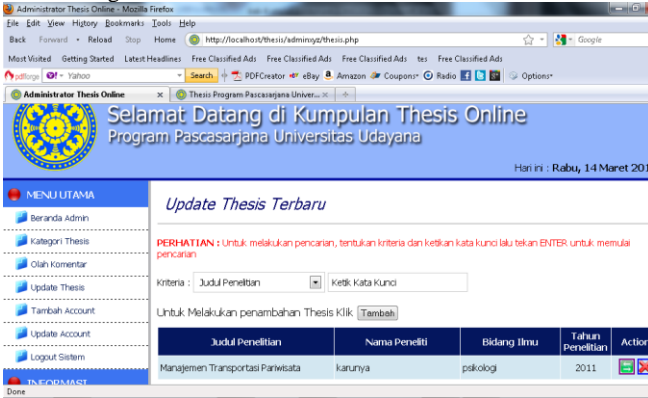The delete process on the thesis system is done through a form like figure 21.


Figure 21. Delete form of the th_thesis table on OLTP thesis system.

The delete activity through this form, triggers the CTL process to work. It makes some change on the record on some tables in a staging area. The first table which will change is the thesis dimension table. Changing on this table is shown like the figure 21.


Figure 21. Delete result on the thesis dimension table.

Other table which also will change is the fakta_prodi_tesis. Due to this process this table will be look like this following figure.


Figure 22. Delete results of the study program thesis

While the pengunjung_per_prodi table will be look like figure 23.


Figure 23. Delete result on the pengunjung_prodi table

## 3.5.2 Data Synchronization Process to Data Warehouse

The data synchronization process from OLTP source to NRTDWH is done by a scheduler. Its work in according to the scheme which has designed. Data which is successfully moved from staging area will be joined into NRTDWH based on the metadata which is shown on table 1 on the following.

Table 1. *staging area* Metadata of *DWH*

| Source staging area | Source tables | Destination table on *NRTDWH* |
|---|---|---|
| *DWH* disertasi | Dimensi disertasi | Dimensi_ts_ds |
| *DWH* disertasi | Dimensi prodi | Dimensi_prodi |
| *DWH* disertasi | Fak_pengunjung_perdisertasi | Fak_pengunjungtsds |
| *DWH* disertasi | Fakta_prodi_disertasi | Fakta_prodi_tsds |
| *DWH* disertasi | Fg_pengunjung_prodi | Fakta_pengunungprodi |
| *DWH* disertasi | Fgpengunjungprodibln | Fakpengunjungprodbln |
| *DWH* thesis | Dimensi tesis | Dimensi_ts_ds |
| *DWH* thesis | Dimensi prodi | Dimensi_prodi |
| *DWH* thesis | Fak_pengunjung_pertesis | Fak_pengunjungtsds |
| *DWH* thesis | Fakta_prodi_tesis | Fakta_prodi_tsds |
| *DWH* thesis | Fg_pengunjung_prodi | Fata_pengunungprodi |
| *DWH* thesis | Fg_kunjungprodibulan | Fakkunjungprodbulan |

The above metadata will be the rule base of the synchronization process. Figure 24 on the following, show the succeed synchronization history of the capture job scheduler.


Figure 24. Job scheduler history

One of the succeed synchronization process which is shown on figure 25 on the following


Figure 25. The synchronization result to the prodi_tesis_disertasi table on the NRTDWH

The synchronization result which is saving on the dimension and fact table on NRTDWH is shown through a data mart application. It make the data on the NRTDWH easier to read and help the end user to get a whole meaning. Trough this application, the data on NRTDWH has to going through masking process first. This process is done by syncronize the prodi dimension table with the related fact. One of the masking processes is done between the values on prodi dimension table which is shown like figure 26 with the record value on figure 25.

Figure 26. Data on the prodi dimension table on NRTDWH

Based on this, therefore the masking process result on the testing application will give a result like figure 27.



Figure 27. Masking result of the dimension and fact table

Through this application, the masking of a result is also can be seen by using graphics. The graphics which is get from the data on the prodi_disertasi fact is like the following figure 28.
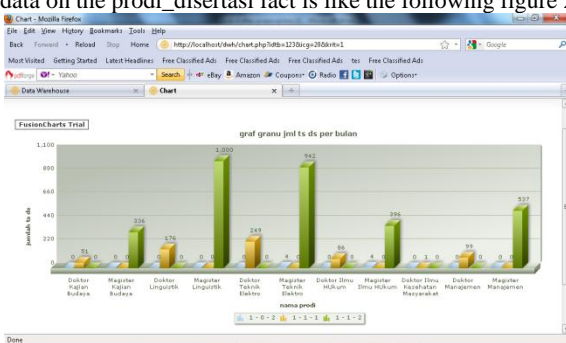


Figure 28. Masking Graphic Result

## 4. CONCLUSION AND THE FUTURE WORK

On this research has developed a method to create nearly data warehouse which comes from some different OLTP with the same platform. NRTDWH is done by implementing CTL based on trigger. It will run the transform and load process in one time on the staging area which is put on the OLTP. This future research is able to be done by applying CTL to create nearly real time data warehouse for form different platform data sources and perform measurements on the OLTP performance because of the extra burden of staging machine. Data integration issues also need special attention to meet a more dynamic modeling. If further research can be done will be obtained data warehouse implementation model that is more real time by cutting processing time in the staging area.

## 5. ACKNOWLEDGMENTS

## 6. REFERENCES

[1] Robert M. Bruckner, Beate List, and Josef Schiefer, Striving towards Near Real-Time Data Integration for Data Warehouses , Data Warehousing and Knowledge Discovery Lecture Notes in Computer Science, 2002, Volume 2454/2002, 173-182, DOI: 10.1007/3-540-46145-0_31

[2] Javed, Dr.Muhammad Younus. , Nawaz, Asim. ,*2010. Data Load Distribution by Semi Real Time Data Warehouse, I*n: Computer and Network Technology (ICCNT), 2010 Second International Conference On page(s): 556 - 560

[3] Inmon, W.H. 2005. *Building The Data Warehouse Fourth Edition*. Canada : Wiley Publishing.Inc.

[4] Simitsis, A.; Vassiliadis, P.; Sellis, T.;, *Optimizing ETL Processes in Data Warehouses*.In Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on  Digital Object, Page(s): 564 – 575

[5] Vandermay, John., 2001. *Considerations for Building a Real-time Data Warehouse*

[6] Savitri, F.N. , Laksmiwati, H. ,Study of localized data cleansing process for ETL performance improvement in independent datamart, Electrical Engineering and Informatics (ICEEI), 2011 International Conference on, [diunduh : 13 Agustus 2011]

[7] Langseth ,Justin., 2004,  *Real-Time Data Warehousing: Challenges and Solutions.*

[8] Jie Song; Yubin Bao; Jingang Shi; 2010, *A Triggering and Scheduling Approach for ETL* . Computer and Information Technology (CIT), 2010 IEEE 10th International Conference on , Page(s): 91 – 98.

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

[10] Mitchell J Eccles, David J Evans and Anthony J Beaumont, True Real-Time Change Data Capture WithWeb Service Database Encapsulation, 2010, 2010 IEEE 6th World Congress on Services

[11] Attunity Ltd , 2009, *Efficient and Real Time Data Integration With Change Data Capture*, Tersedia di http://www.attunity.com/cdc_for_etl

[12] Jingang Shi, Yubin Bao, Fangling Leng, Ge Yu.2008,*Study on Log-Based Change Data Capture and Handling Mechanism in Real-Time Data Warehouse*. In International Conference on Computer Science and Software Engineering, CSSE 2008, Volume 4: Embedded Programming / Database Technology / Neural Networks and Applications / Other Applications, December 12-14, 2008, Wuhan, China. pages 478-481, IEEE Computer Society, 2008.

[13]Liu Jun; Hu ChaoJu; Yuan HeJin. 2010. *Application of Web Services on The Real-time Data Warehouse Technology*, Advances in Energy Engineering (ICAEE), 2010 International Conference on  , Page(s): 335 – 338