RULE BASED ETL (RETL) APPROACH FOR GEO SPATIAL DATA WAREHOUSE

Norhaira Nordin[1],Azman Yasin[2]and Mazni Omar[3]

[1]*Universiti Utara Malaysia(UUM), norhairanordin@gmail.com*
[2]*Universiti Utara Malaysia(UUM), yazman@uum.edu.my*
[3]*Universiti Utara Malaysia(UUM), mazni@uum.edu.my*

ABSTRACT

This paper presents the use of Service Oriented Architecture (SOA) for integratingmultisource heterogeneous geospatial data in order to facilitate geospatial data warehouse. In this study, Real Based ETL (RETL) concept is adapted in order to extract, transform and load data from a variety of heterogeneous data sources. ETL will transform datato schematic format and loading data into the Geo spatial data warehouse. By using a rule-based technique, the distribution of parallel ETL pipeline will enhance and perform more efficient in large scale of data and overcome data bottleneck and performance overhead. This can ease the disaster management and enables planners to monitor disaster emergency response in an efficient manner.

Keywords:Geographical Information System, Service Oriented Architecture, Extract Transform Load, Pipeline, Rule Based.

I INTRODUCTION

In recent times,weoftenseeextremechangesinmeteorologicalphe nomenasuch as floods, landslide, and tsunami. This extreme change has contributed to the utilization of information technology to facilitate disaster management and enables planners to work primarily from monitoring form to disaster emergency response more effective.The critical task of disaster management system is the geographical information (GI) itself which consists of large amount of data collected from satellite, airborne, and ground based imaging systems from multi-platform, multi- source, multi-temporal image and data manipulation to facilitate decision making(Song, Fang, Chen, & Yang, 2011; Yvan Bédard & Han, 2009).

In order to deal with the integration of multisource heterogeneous geospatial data, the design for data integration is based on a Service Oriented Architecture (SOA)(Niu, Guo, Lu, & Chen, 2011; Wenfeng, 2010; Zheng, He, Xiong, Liu, & Framework, 2011; Zheng, Huang, & Zhang, 2010). The architecture of SOA mainly composed of data layer, services layer and display layer (Wang, Liu, Wang, & Xu, 2011) which can offer the flexibility to handle different task and provide servicesand the separationofbusiness logics from the foreground andbackground.(Feng & Lee, 2010; Zheng, Huang, & Zhang, 2010). In Figure 1, it is shown SOA framework and spatial information servicesin GI (Ying, Yang, Ya-fu, & Li-zhou, 2010)

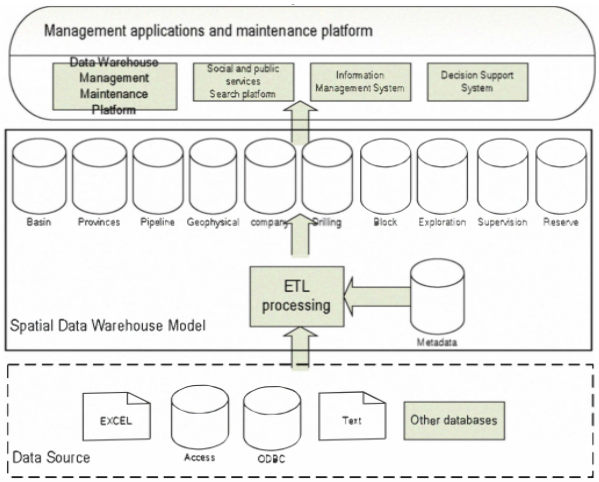

Figure 1. SOA framework for GI System

In the study of spatial data warehouse, preparing and handling the spatial data warehouse in the data warehousing industry takes about 60% of the data warehouse developmenttime(Rifaie, Blas, Muhsen, Mok, & Ridley, 2008).GI data can come from the various databases such as Oracle, SQL Server and ARCGIS(Oracle, 2010; Microsoft, 2010; Tedrick, 2013).Extract, Transformation and Load (ETL) is a backbone for solving the multiple sources of spatial data, expression scale, consistency, attribute and so on(Naeem et al., 2008; Ying et al., 2010).First issue in geo spatial data warehouse is to handle large amount of structured data and unstructured data that can come to terabyte level and will turn to data organization, management for GI application bottleneck (W. Song, Chen, & Yang, 2011). Second issue is the integration of data for each data loading window can cause performance overhead where there are many extra ETL component during transformation data to data warehouse such as aggregation, conditional split routes, lookup, data conversion

transform,derived column transform and so on (Tok, Parida, Masson, Ding, & Sivashanmugam, 2012) . Thus, to increase the data integration performance of geospatial data warehouse, this study proposed the use of rule-based concept of ETL pipeline (RETL) in Geo Spatial SOA Framework.In this study, RETL is focused on performance in parallel ETL pipelines for extracting data and actions from operational databases and loading them into enterprise geospatial data warehouse. The remainder of this paper is organized as follows. In the section II, we present SOA spatial data integrationand following section III presents the proposed RETL concept. Sections IVthe expected result and section V will summarize our conclusion and action for future work.

II SERVICE ORIENTED SPATIAL DATA INTEGRATION

Spatial data warehousing for geographic knowledge often have several departmental or application-oriented independent databases which may overlap in content. By using Service Oriented Architecture (SOA), the integrated platform can supports various distributed resource from several geographical server(Wang et al., 2011). SOA is particularly consisting of data layer, service layer, and display layer. In Figure 2, it is shown Proposed SOA framework and Spatial Data Warehouse.

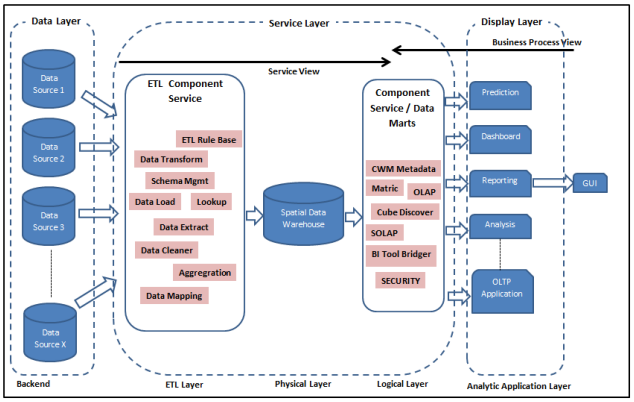

Figure 2. Proposed SOA framework and Spatial Data Warehouse

A. Data Layer

The data layer mainly includes the fundamental geographicinformation, the Aerial photo, satellite image, and theinterpretation information of remote sensingsuch as population data, socio-economic data, water data, and weather data and so on.

B. Service Layer

The service layers mainly distribute andmanage the various types of services, establish the links between application layer and services, datalayer, and accessthedistributed multi-source heterogeneous data. The ETL system are used for extracting business data, transform and cleansing it into an analytical format.

C. Display Layer

Display layer is designed to visualized information from multisource spatial data directly use to end user and the corresponding function.

Taking into account on the complexity of spatial data, mass characteristics, data bottleneck, and volume of GI data,the execution efficiency need to be improve.Next, we will focus on SOA service layer for spatial data integration by enhancing ETL pipeline job.

III PROPOSED RULE BASED ETL (RETL)

One of the most difficult proses during the creating any data warehouse is extracting, transforming, cleansing and loading data proses from data source to warehouse destination (Santos, Bernardino, & Vieira, 2011).

A. ETL Technology

The Extract-Transform-Load (ETL) system is the foundation of the data warehouse(Kimball & Caserta, 2004). ETL is typically used to integrate data from a variety of applications, developed and supported by different server or hosted on separate computer hardware. It integrates the scattered, disorder and heterogeneous data into the target data warehouse, which is used by different business systems. The ETL process start with retrieves data from data sourcethen processes the data by enforces data quality and consistency standards, conforms data so that separate sources can be used together and finally loads it into the database. The following steps are process in ETL:In Figure 3, it is shown detail ETL process (Kimball & Caserta, 2004).

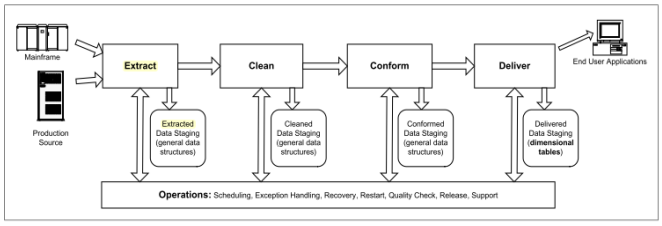

Figure 3. ETL Data Flow Thread

B. Rule Based ETL (RETL) Architecture

The current ETL system allows parallel data processing and aggregation which can scale up to the most demanding systems. In many studies, the parallel execution is distributed for different ETL task process (J. Song, Bao, & Shi, 2010). The proposed

RETL will introduce the motivated ideas and the back ground of ETL Rule Based and Parallel Pipeline.

C. RETL Architecture

The proposed RETLcontains fivemain components:

- Listener
- Triggering Rules
- ETL Job Dispatcher
- ETL Rule Based
- ETL Engine

In figure 4, it is shown proposed RETL architecture.

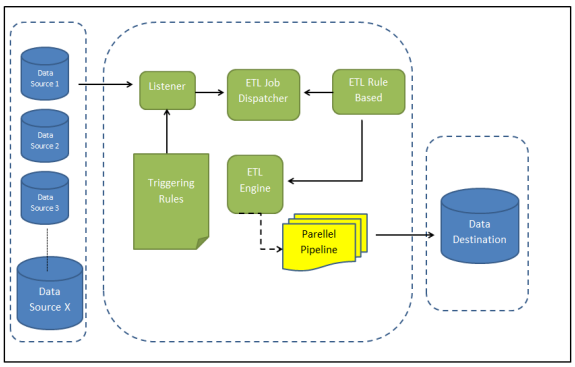

Figure 4. RETL Architecture

RETL starts when the **triggering rules**or messaging mechanismnotifiesthe **listener**to analysing new data source to be extract. The triggering rules will identify the ETLjobs defined by users, analysing the syntax jobsand checking the trigger method rules of ETL job such as BEGIN, CHECK, INIT and END so the RETL system can define the next phase of ETL flow. Next, the **job dispatcher** will determine which ETL pipeline task is used according to user-define pipeline process. **The Rule Based** is responsible for dividing data in a batch and set how many pipelines will be used to run the ETL task in parallel. Our proposed component will be ETL Rule based and Parallel Pipeline.

The monitoring RETL system is based on the analysis of logs generated SQL Integration module, so the status of RETL system can be monitored when deploys it on production environment.

D. ETL RULE BASED

ETL Rule based will optimize the distribute resource in parallel pipeline. There are maximum *N* thread access data in parallel pipeline.(Basco, Gerardo, Dofitas Jr., & Byun, 2012; Mendes & Falcone Sampaio, 1998)

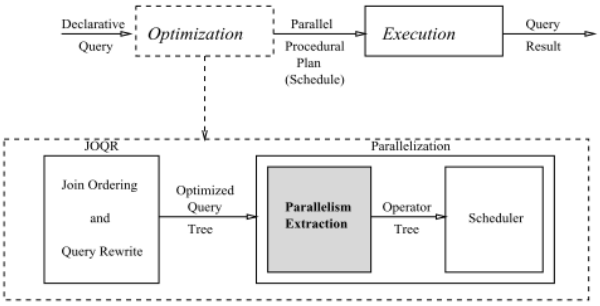

Figure 5. Parallel Optimization Diagram

Step 1: We have to specify the pipeline task can work parallel or not. It depends on GI data whether it dependent or independent data.

Step2:A set of rules was used based on the individual core's capabilities and how many threads, total size of threads.Below in Figure 6 is an illustration of the rules of how a thread is assigned to a processor or core (Basco et al., 2012).

Step 3: Rule based will be sent to the ETL to run the ETL system

Check if thread size + Processor1.taskload ≤ SafeLoadLimit

If thread size size + Processor1.taskload < SafeLoadLimit, Processor 1 will receive the thread, else the thread will be given to Processor 2 to be evaluated.

If thread size size + Processor2.taskload < SafeLoadLimit, Processor 2 will receive the thread, else the thread will be given to Processor 3 to be evaluated.

If thread size size + Processor3.taskload < SafeLoadLimit, Processor 3 will receive the thread, else the thread will be given to Processor 4 to be evaluated.

If thread size size + Processor4.taskload < SafeLoadLimit, Processor 4 will receive the thread, else the thread will be given back to Processor regardless condition # 2.

Figure 6. Example of rules assign to a processor

E. PARALLEL PIPELINE

Parallel processing divides a large task into many smaller tasks, and executes the smaller tasks concurrently on several nodes. As a result, the larger task completes more quickly.There are three points need to be considering while setting the rule base for parallel processing:

- The effectiveness for main memory and CPU usages.
- Optimization problem that relates the volume of data to be propagated from a data source towards the data processing area.
- In several operations might be mutually exclusive and don't have any dependencies

among each other. These operations can easily be parallelized.

Loading or querying massive datafrom into one table will result time costly. Splitting large data file into small data can improve the efficiency in RETL(Sun, 2012).

$$C = \sum_{i=1}^{n} c_i ; T \gg \sum_{i=1}^{n} t_{c_i}$$

Where C is the total counts of the records reside in the data file, and C_i is the counts of the records reside in one ofpartitions, and T is the total time cost of loading the data file, and the $t_{ci}$is the time cost of loading one of partitions. Theclause above means that the time cost of loading one large file at a time is significant larger than the sum of time cost of loading its partitions separately.

IV EXPECTED RESULT

In sequential processing, the query is executed as a single large task. In parallel processing, the query is divided into multiple smaller tasks, and each component task is executed on a separate nod.

Figure 4 and Figure 5 contrast sequential processing of a single parallel query with parallel processing of the same query (Oracle, 1997).

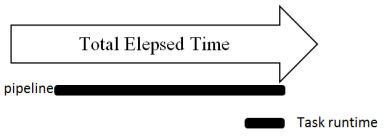

Figure 7. Sequential Processing of a Large Task

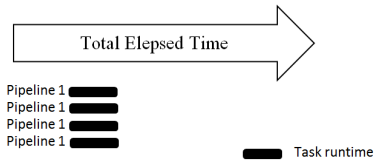

Figure 8. Parallel Processing: Executing Component Tasks in Parallel

Basic performance goals of parallel system can be measure in terms of two important properties (Oracle, 1997):

- Speedup= $\frac{Time\_Original}{Time\_Parallel}$
- Scaleup = $\frac{Volume\_Parallel}{Volume\_Original}$

Assuming during T time, N_u ETL tasks and N_q queries have been executed. The approximate radio between executing time of single query and single update is δ. Average executing time is the average time consuming for executing one SQL (update) during time T, it can be calculated as (J. Song et al., 2010).

$$AET = T \cdot (\sum_{i=1}^{N_u} ET_i.count + N_q \cdot \delta)^{-1}$$

AET is a relative value to measure the performance and it will reach a stable value when the system is full-load. Basically, the smaller the AET is, the more efficient RETL is.

V CONCLUSION

In this paper, data integration in geo spatial data warehouse is studied to resolve the problem of preparing and handling the spatial data warehouse in the data warehousing industry. The large amount of structured and unstructuredGI data will result indataorganization andmanagementofGIapplicationbottlenecks. The ETL plays an important role in spatial data warehouse (Chen & Chi, 2004). However the integration of data for each data loading window can cause performance overhead when there are many extra ETL component during transformation data to data warehouse such as aggregation, conditional split routes, lookup, data conversion transform,derived column transform and others. An integrated SOA geo spatial data is introduced that incorporating a rule based technique to distribute parallel ETL pipeline so that the transformation performance can be enhanced. In the future, we still need to do some research works that expand the spatial data dependency, spatial data complexity and ETL rule based algorithm that can integrates with multisource heterogeneous geospatial data.

ACKNOWLEDGMENT

"The authors wish to thank the Ministry of Education, Malaysia for funding this study under the Long Term Research Grant Scheme (LRGS/b-u/2012/UUM/Teknologi Komunikasi dan Infomasi)"

REFERENCES

Basco, J. R., Gerardo, B. D., Dofitas Jr., C., & Byun, Y.-C. (2012). A Rule-based Parallel Processing to Speed-Up an Application. *2012 IEEE 14th International Conference on Commerce and Enterprise Computing*, 144–146. doi:10.1109/CEC.2012.32

Chen, X., & Chi, Z. (2004). Applying dp to etl of spatial data warehouse, (August), 26–29.

Feng, Y.-H., & Lee, C. J. (2010). Exploring Development of Service-Oriented Architecture for Next Generation Emergency Management System. *2010 IEEE 24th International Conference on Advanced Information Networking and Applications Workshops*, 557–561. doi:10.1109/WAINA.2010.198

Kimball, R., & Caserta, J. (2004). *The Data Warehouse ETL Toolkit*. Wiley Publishing, Inc.

Microsoft. (2012).*Integration Services (SSIS) - SQL Server*. (n.d.). Retrieved from http://msdn.microsoft.com/en-us/sqlserver/cc511477.aspx

Mendes, S. F., & Falcone Sampaio, P. R. (1998). Rule-based parallel query optimization for OQL using a parallelism extraction technique. *Proceedings Ninth International Workshop on Database and Expert Systems Applications (Cat. No.98EX130)*, 705–710. doi:10.1109/DEXA.1998.707485

Naeem, M. A., Dobbie, G., Weber, G., Bag, P., Street, P., & Zealand, N. (2008). An Event-Based Near Real-Time Data Integration Architecture Keywords :

Niu, H., Guo, P., Lu, Y., & Chen, H. (2011). Disaster Recovery, 1–4.

Oracle. (1997).*Parallel Programming with Microsoft .NET*. (2010, January). Retrieved from http://docs.oracle.com/cd/A58617_01/server.804/a58238/ch1_unde.htm

Oracle. (201 0). Oracle Warehouse Builder. Retrieved 3/2/20 10, from http://www.oracle.com/technetwork/developer-tools/warehouse/overview/index.ht~nl.

Rifaie, M., Blas, E. J., Muhsen, A. M., Mok, T. T. H., & Ridley, M. J. (2008). Data Warehouse Architecture for GIS Applications, (178), 178–185.

Santos, R. J., Bernardino, J., & Vieira, M. (2011). 24 / 7 Real-Time Data Warehousing : A Tool for Continuous Actionable Knowledge. doi:10.1109/COMPSAC.2011.44

Song, J., Bao, Y., & Shi, J. (2010). A Triggering and Scheduling Approach for ETL in A Real-time Data Warehouse, (Cit), 91–98. doi:10.1109/CIT.2010.57

Song, W., Chen, J., & Yang, Y. (2011). OPERATIONAL DATA STORE OF MULTI-PLATFORM , MULTI-SOURCE , MULTI- SCALE , MULTI-TEMPORAL DATA SETS, (68), 2–5.

Sun, K. (2012). SETL : A Scalable And High Performance ETL System, 8–11.

Tedrick J. (2013). *The Best of ArcGIS 10.2*. Retrieved from http://msdn.microsoft.com/en-us/sqlserver/cc511477.aspx

Tok, W., Parida, R., Masson, M., Ding, X., & Sivashanmugam, K. (2012). *Microsoft SQL Server 2012 Integration Service*.

Wang, X., Liu, J., Wang, Y., & Xu, S. (2011). Service-oriented Integration and Application of Earthquake Emergency Information, (7771014).

Wenfeng, Z. (2010). Spatial information integration of earthquake disaster prevention and reduction based on SOA. *The 2nd International Conference on Information Science and Engineering*, 1–4. doi:10.1109/ICISE.2010.5689162

Ying, L., Yang, W., Ya-fu, C., & Li-zhou, F. (2010). The Application of Spatial Data Warehouse Technology in The National Petroleum Resources Database, 257–260.

Yvan Bédard, & Han, J. (2009). Fundamentals of Spatial Data Warehousing for Geographic Knowledge Discovery, 45–68.

Zheng, Z., He, S., Xiong, Z., Liu, Z., & Framework, A. S. O. A. (2011). Hydrologic Prediction and Visualization Integrated Mobile System Based on Service Oriented Architecture, (200805016).

Zheng, Z., Huang, D., & Zhang, J. (2010). A SOA-based Decision Support Geographic Information System for Storm Disaster Assessment, (200805016).