International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-5S3, July 2019

Issues and Handy Solutions Addressed at Every Stage in Real Time Data Warehousing, I.E. ETL (Extraction, Transformation & Loading)

Arif Ali Wani, Bansi Lal Raina

Abstract— In the standard ETL (Extract Processing Load), the data warehouse refreshment must be performed outside of peak hours. It implies that the functioning and analysis has stopped in their all actions. It causes the amount of cleanness of data from the data Warehouse which isn't suggesting the latest operational transections. This issue is known as data latency. The data warehousing is employed to be a remedy for this issue. It updates the data warehouse at a near real-time Fashion, instantly after data found from the data source. Therefore, data latency could be reduced. Hence the near real time data warehousing was having issues which was not identified in traditional ETL. This paper claims to communicate the issues and accessible options at every point in the near real-time data warehousing, i.e. The issues and Available alternatives are based on a literature review by additional Study that focus on near real-time data warehousing issue.

Index Terms: Business Intelligence, Data Latency, Data Warehouse, Data Warehousing, ETL, Near Real Time Data Warehousing.

I. INTRODUCTION

The data warehouse is upgraded all the way through ETL (Extract, Transform, and Loading) procedure. ETL has the accountability to detect related shift data, extract it into the staging area, and change it into a Proper format, and then finally load it into the data warehouse table[1]. Conventionally, the data warehouse is updated occasionally by ETL [2]. Which means the data in the data warehouse is not relevant to the current condition [3], where there is a real-time data between the two procedures that are updating . Therefore, it makes a less precise analysis outcome. The traditional ETL should be not be done at peak hours which is an another issue that need to be addressed [2].When means that the analytical as well as operational action must stop all [1]–[4]. This causes a very serious problem for a machine which is running 24 hours a day and 7 days in a week [5]. Based on these issues there must be a mechanism for updating data warehouse instantly even after a small change found in data so that the least data should be fulfilled by the user. This phenomena is known as near real-time data warehousing[6]–[8]. Other conditions are usable data warehouse [5]or real-time ETL. Loading data procedure into the data warehouse is performed incessantly, on the close real-time data warehousing. This approach is different from traditional approach which is employed occasionally.[9]. Previously near real-time data warehousing have many issues which was not found on the traditional ETL. In the related study our main focus is to study on near real-time data warehousing issues[8], its not clear how the failure occurred at Extraction, transformation or loading stages. The aim of this paper is to suggest nomenclature which consists of Available solutions and issues at each point. These available solutions and issues originate from the literature review by various other researchers who draw light on near real-time data warehousing issues. With this nomenclature, it is anticipated that the additional research will get easier to address the problem and participation that will be evolved.

Manuscript published on 30 July 2019.
* Correspondence Author (s)
Arif Ali Wani, Computer Science and Engineering Department, Glocal University, Saharanpur, India.
Bansi Lal Raina, Computer Science and Engineering Department, Glocal University, Saharanpur, India.

II. RELATED WORK

Few studies which have clustered theses issues into real-time data warehousing and these problems can be broken into five impeaches, specifically enabling real-time ETL, simulating real-time truth tables, altering data versus OLAP query, Scalability and question emptiness, and real-time alerting[10]. On the other hand another study split this into three classes, specifically enabling the real-time ETL system, data flow management and enabling real-time Business Intelligence [11]

I. MAINTAINING THE INTEGRITY OF THE SPECIFICATIONS AVAILABE SOLUTIONS AND PROBLEMS IN REAL TIME DATA WAREHOUSING

Transformation stage

Transformation is a process to adjust the data obtained from the data source into a predetermined format. There are two Troubles in the transformation phase as follows:

A. *Master data overhead*

Data saved in the data warehouse can be divided into two components, specifically, master data and transactional data[8]. Master data is Data that is not often changed. By Way of Example,

is merchandise or Client? From the data warehouse, master data is executed from the measurement table. Transaction data is often changing Data in line with this transaction happened in the data source. For illustration the revenue transaction[6]. In data warehouse, transaction Data is employed from the table.

Every data warehouse refreshment process is predicated on Transaction data created. However, this process also requires master data. Master data will be utilized for transaction data connect process. Therefore, exactly the identical master data will be often extracted. Its Issue is known as master data overhead[4].

To solve this, master data is placed on a cache, even while real time Data is set in the database queue. Moreover, the link is performed between real-time data on the database queue together with master Data on cache or every single transaction. This mechanism is exemplified in Figure 1 as follows [12]:

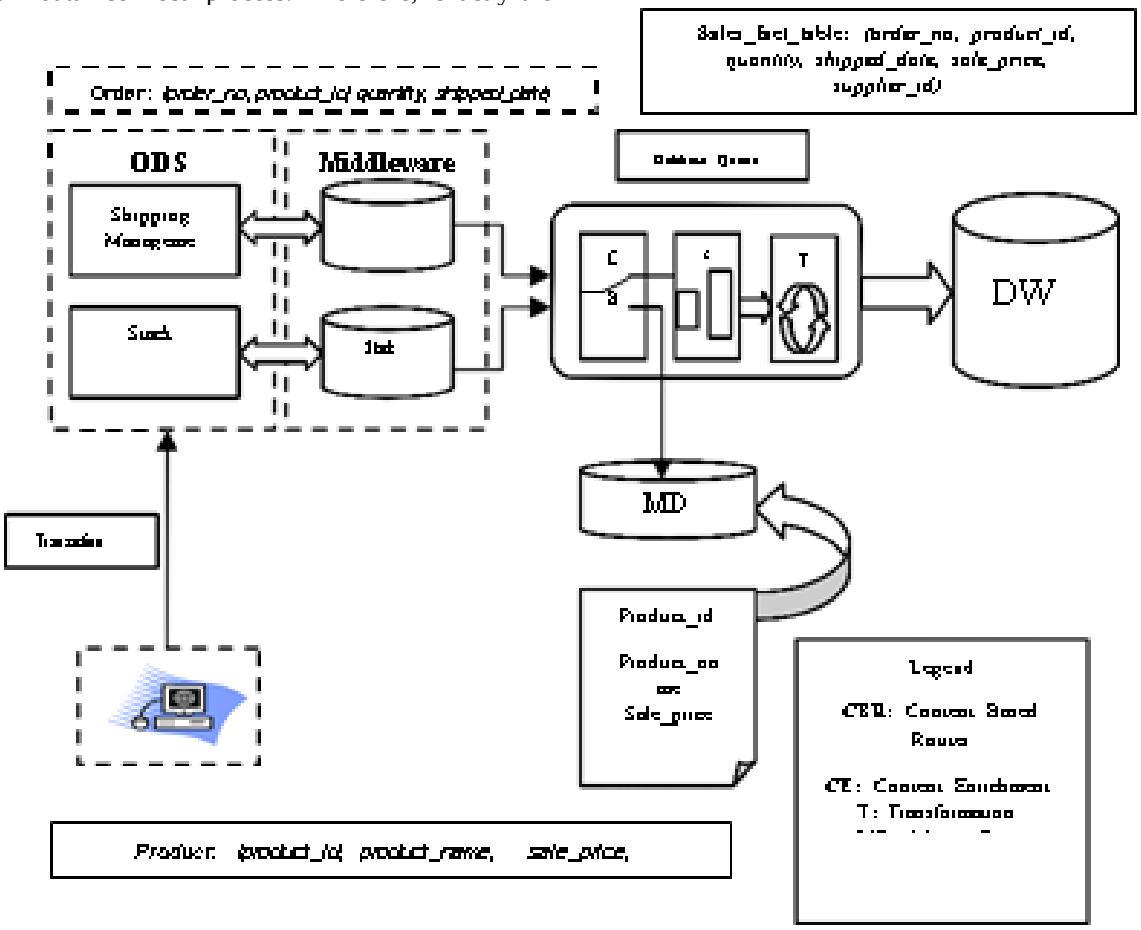

Fig.1. Mechanism for handling master data overhead

B. *Require midway server to achieve data aggregation*

The transformation process is performed before data is loaded to the data warehouse. In the traditional ETL, transformation processes a Set of data from the staging area with ETL tools. On the close Real-time data warehousing, every data warehouse refreshment Process only conveys a couple of tiny quantities of data. This Resulted the transformation process cannot be achieved on every data warehouse refreshment cycle. To solve this, a Method using the title ELT (Extract Load Change) may be used. With ELT, the transformation process is implemented in the data warehouse [6], [7] as shown Figure 2 [9].

Based on Figure 2, extracted data in the data source will be loaded right into the data warehouse. The resulting data Warehouse includes an unclean backup of the data origin. As a result, the transformation process to reshape data into an proper format is necessary.

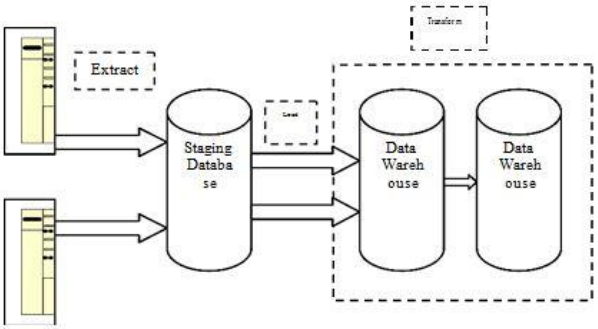

Fig.2. ELT method

C. *Loading stage*

Loading is the last phase from the ETL process, in which blank data in the conversion process is loaded to the right data warehouse table. The issue occurred when you can find Trades during OLAP analysis. Transaction data will overlap with OLAP process [2]. Consequently, there'll be performance degradation in the analysis process. Another outcome is that the occurrence of OLAP inner inconsistency.

a. Degradation in performance

To minimize the performance degradation from the analysis process, staging tables (temporary tables) may be applied as a solution. Staging table is a table that has exactly the exact Same format with data warehouse destination tables. These staging tables will be utilized to stores and receives real-time data briefly.

Data warehouse destination table will be upgraded from this Staging table occasionally. This solution is called by trickle and reverses [3]. This solution has been developed into a multistage Trickle and reverse shown in Figure 3 [13].

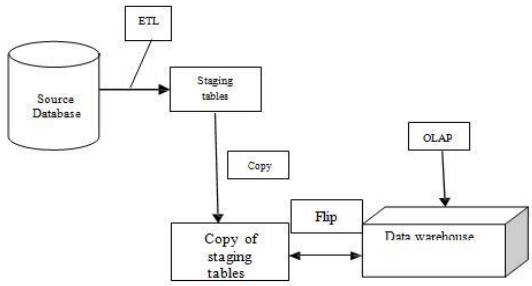

Fig. 3. Data loading using 'Trickle & Flip'

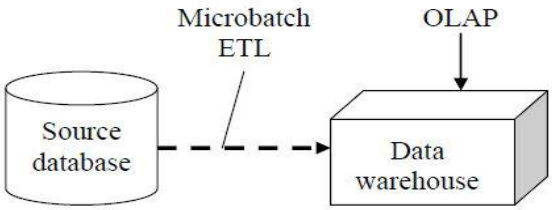

Fig. 3.1. Near real-time warehousing with near real-time ETL

b. Internal inconsistency at OLAP

OLAP is intended to work with inactive data. There's no Mechanism to stop data alteration to data used by an OLAP process. When alterations of data occur in exactly the same Time using an OLAP action that uses this data, OLAP will Difficulty inconsistent outcome. This Issue is called by OLAP Inner inconsistency[14]. This Issue can happen in OLAP Operation like roll-up – drill down. Inconsistencies will happen between aggregation and detail effect. To prevent this, analysis proses on table solution can be performed outside the refreshment period of data warehouse table Another solution is taking a picture of data in the data Warehouse table and also Using RTDC (Real Time Data The photo data is utilized for the analysis process, Whereas the data warehouse table is utilized to store real-time data Permanently Forever}. {RTDC is the use of an external cache between the data source and data warehouse. This cache will be accustomed to temporarily save real-time data[12]. The cache will be read occasionally to transfer its content into the data warehouse table. If Required, JIT (Just in Time) process may be used to unite Data in the data warehouse table and cache within an analysis process [7], [11].

Another solution is mixing row lock with layer - view that is generated dynamically. This method is termed by layer-based opinion as shown in Figure 4 [4], [8]

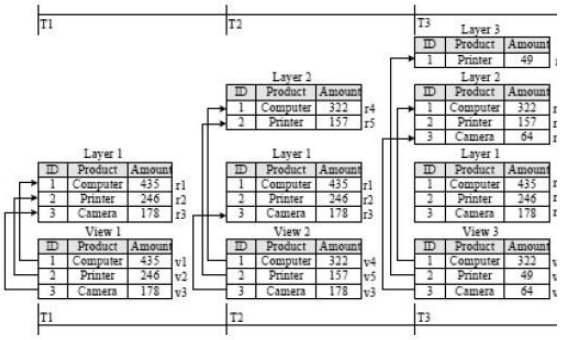

Fig.4. Layer-base view

Based on Figure 4, when rows in a table are utilized for the analysis process, these rows will probably have row lock. If in the same Time these rows be a goal to get a data source trade, a new Coating will be created automatically to keep those Shift For every analysis process, an opinion which involves Made layer will be generated. The Goal of this opinion would be to Supply the most recent data for every analysis process[9], [11].

c. Available solutions and Problem nomenclature results

Solutions which were developed from the method of extraction, Transformation, and loading near real-time data warehouse could be summed up into taxonomy as shown in Figure 5 as follows:

Based on Figure 4, when rows in a table are utilized for the analysis process, these rows will probably have row lock. If in the same Time these rows be a goal to get a data source trade, a new Coating will be created automatically to keep those Shift For every analysis process, an opinion which involves Made layer will be generated. The Goal of this opinion would be to Supply the most recent data for every analysis process[9], [11].

d. Available solutions and Problem nomenclature results

Solutions which were developed from the method of extraction, Transformation, and loading near real-time data warehouse could be summed up into taxonomy as shown in Figure 5.

e. Data Extraction stage

Extraction is the process of accessing data from the data source. There are just two problems in the extraction phase as follows: The data source can be divided into two components, namely, stored data Place and data stream. The stored data collection is data Which Can Be used Over and over again, and rare upgrading procedure. The data Stream is data the usage not and continuously shifting [6], [7]. An instance of this data stream is in the fund Program, network traffic monitoring, click stream net Program, detector, email, and phone call data detail [6], [11]. To deal with data stream, the stream processor can be used [9], [12]. Whereas to tackling stored data collection, CDC -- Change Data Grab is utilized [11]. To incorporate both data, stream Chip and CDC are attached to a message queue which Connected at data integration tool [9].

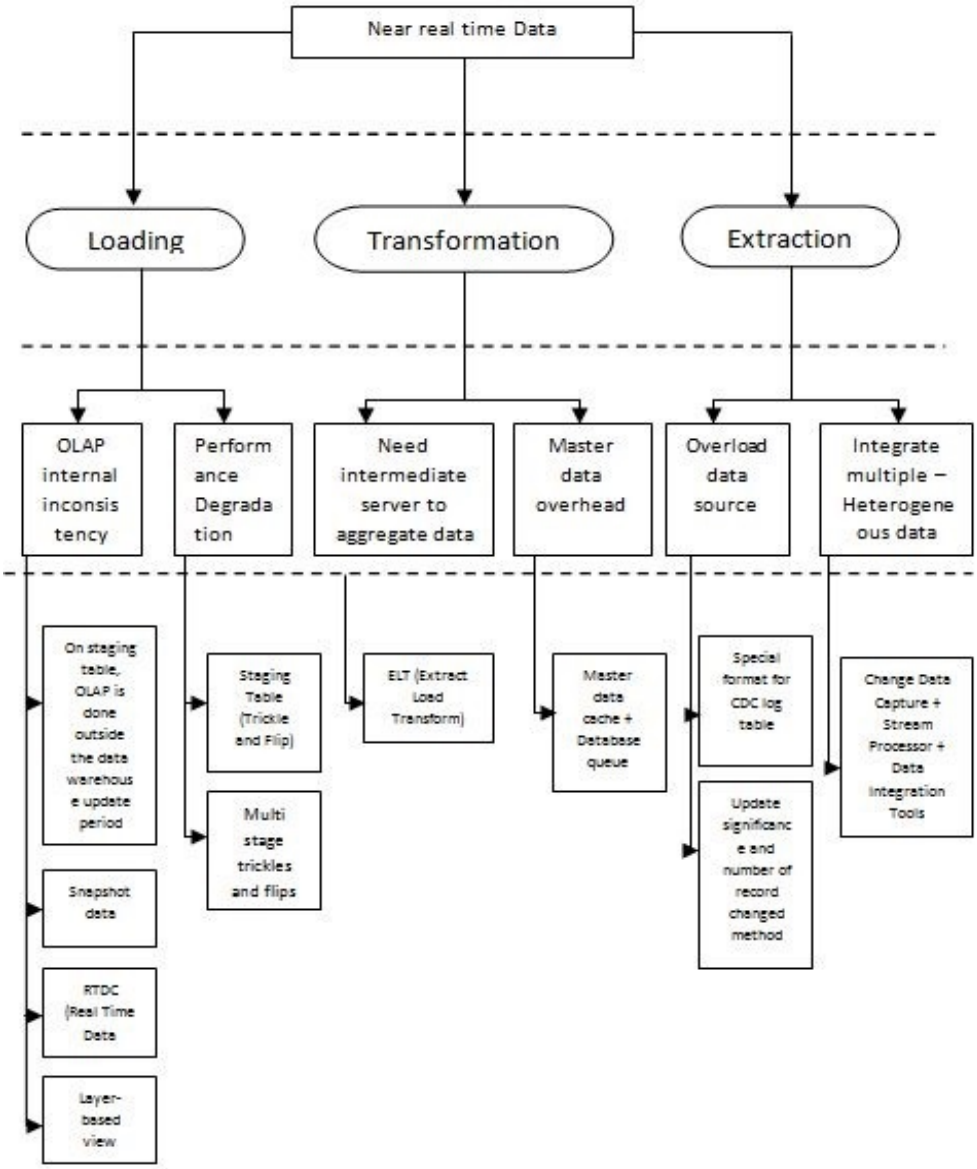

Fig.5. ETL solutions and problem Real Time Data Warehouse taxonomy result CDC-trigger [7].

f. Data source overload

Reading data source Always make overload which disturbs Operational actions. Due as well as serving the operational Trades, the data source should also function as a reading by CDC. To solve this, more effective extraction method from the CDC has developed, which can be called by upgrade significance and amount Of record altered method. This method aims to receive priority Data to be performed on each extraction. Change data that don't match the priority group is going to be pulled using conventional ETL this method is displayed in Figure 6. The other way is to create a Special format to your log table in

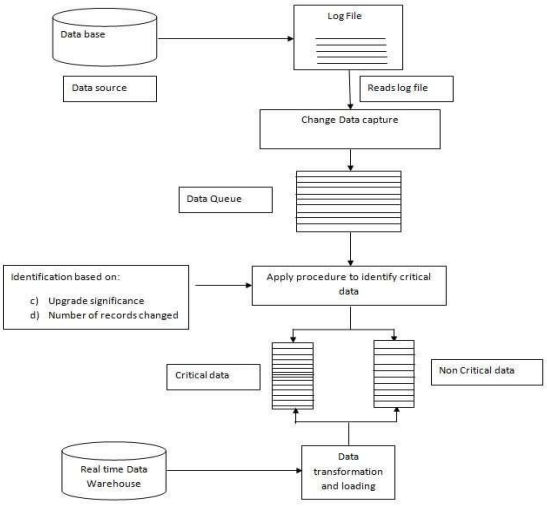

Fig.6. Significant Update and methods to change number of record.

III. CONCLUSION

This research paper has suggested a taxonomy which includes the alternatives and various issues over each stage of an ETL i.e., extraction, transformation, and loading at the real-time data warehousing according to literature review this effect, further study related to real-time data, warehousing may use it to have the attention Difficulties and their participation which will be given.

REFERENCES

1. E. Low, L. No, B. Windows, and L. Costs, "Efficient and Real Time Data Integration With Change Data Capture," *Integr. Vlsi J.*, no. Cdc, pp. 1–20, 2009.
2. R. J. Davenport, "ETL vs ELT," no. June, 2008.
3. G. Swetha, D. Karunanithi, and K. A. Lakshmi, "Data Integration Models for Operational Data Warehousing," vol. 3, no. 2, pp. 508–516, 2014.
4. R. S, S. Balaji. B, and N. K. Karthikeyan, "From Data Warehouses to Streaming Warehouses: A Survey on the Challenges for Real-Time Data Warehousing and Available Solutions," *Int. J. Comput. Appl.*, vol. 81, no. 2, pp. 15–18, 2013.
5. A. A. Wani and B. L. Raina, "Data in Data Warehouse and its Qualities Issues," no. 9, pp. 1753–1756, 2019.
6. K. Kakish and T. A. Kraft, "ETL Evolution for Real-Time Data Warehousing," *Proc. Conf. Inf. Syst. Appl. Res.*, p. 12, 2012.
7. N. Rahman, "Refreshing Data Warehouses with Near Real-Time Updates," *J. Comput. Inf. Syst.*, vol. 4417, no. Spring, p. 70, 2007.
8. A. A. Wani, U. Chandra, P. Bansi, and L. Raina, "Security Challenge in Big Data for Behaviour Analytics," vol. 5, no. 7, pp. 578–581, 2018.
9. R. J. Santos and J. Bernardino, "Real-time data warehouse loading methodology," p. 49, 2008.
10. A. A. Wani, A. Khan, A. Jamal, and P. K. Gupta, "Cost Efficient Media Cloud Storage and Systematic Risks Involved in the Cloud Computing," no. 9, pp. 2466–2469, 2019.
11. A. A. Wani, "Discovery of knowledge by using Data warehousing as well as ETL processing."
12. M. A. Naeem, G. Dobbie, and G. Weber, "An event-based near real-time data integration architecture," *Proc. - IEEE Int. Enterp. Distrib. Object Comput. Work. EDOC*, no. April, pp. 401–404, 2008.
13. D. Agrawal, "The reality of real-time business intelligence," *Lect. Notes Bus. Inf. Process.*, vol. 27 LNBIP, pp. 75–88, 2009.
14. J. Zuters, "Near real-time data warehousing with multi-stage trickle and flip," *Lect. Notes Bus. Inf. Process.*, vol. 90 LNBIP, pp. 73–82, 2011.

AUTHORS PROFILE

Arif Ali Wani received his Bachelor's degree in Information and Technology from Model Institute of Engineering and Technology (MIET) affiliated to Jammu University, Jammu India. During the 2008 and M.Tech in Computer Science and Engineering from Gurgaon College of Engineering affiliated to Maharshi Dayanand University Rohtak, in the year 2013. He is having 9 years of teaching experience, his area of business is Data Warehouse and Data mining, Computer Network. He has published and his Research papers in peer reviewed International Journals, Book Chapters, international and national level conferences.

Bansi Lal Raina Backed by an exceptionally brilliant academic record, Prof. Raina has been engaged in administration, teaching & research for nearly 35 years now. He was awarded prestigious national fellowship of "TATA INSTITUTE OF FUNDAMENTAL RESEARCH" (T.I.F.R), Bombay, INDIA wherein he spent four years of research work and then proceeded to USA on an International fellowship to obtain his M.Tech (Computer Science Engineering & PhD. From 'USC', USA. Did he not only write an exemplary research paper at an early age of his career of 10+2 standard published by reputed 'American Mathematical Society' (January 1969 page 48-51), but his paper (part of which is noted below just for reference) was also widely acclaimed and often cited (e.g., See A. Del Cintel, 2008-SPRINGER) which in a dramatic development helped various eminent Scientists like Prof. ANDRE WILE then at PRINCETON UNIVERSITY, to draw a vital connection between the ELLIPTIC CURVES and MODULAR FORMS (See Ribet: Tanahama-Shimura Conjecture, 1986) leading him eventually to the famous solution in 1995 of even more famous CONJECTURE (See Annals of Mathematics, 142 (1995), which was unsolved for the last 350 years, earned Prof. Wiles a well-deserved 'KNIGHT HOOD' & the most prestigious award of 'FIELDS MEDAL'. Dr. Raina's above cited results have also immensely helped in the development of many subjects and more recently in 'CALABI-YAU' spaces & 'STRING Theory' in ASTRONOMY, thereby unifying the theories of 'NEWTON'S Gravitation, QUANTUM Physics & EINSTEIN'S Relativity.