A Review on Traditional ETL Process for Better Approach in Business Intelligence

Erarda Vuka[1]
Dept. Of Informatics
University of Tirana
erarda.vuka@yahoo.com

Olta Petritaj[2]
Dept. Of Applied Informatics
University College “Logos”
olta.petritaj@gmail.com

Abstract

Data Warehouse is a repository of strategic data from many sources collected over a long period of time. Traditional DW operations mainly consist in extracting data from multiple sources, transforming these data into a compatible form and loading them to DW schema for further analysis, this is the key point of building an ETL process. The Extract, Transform and Load (ETL) process of Data Warehousing is critical in determining the capabilities of Business Intelligence (BI) systems. The demand for fresh data in data warehouses has always been a strong requirement from the BI users but the traditional ETL process does not support real time necessities, as the refreshment of data warehouses has been performed in an off-line batch mode. To resolve this issue, near-real time ETL approaches come out as a solution. In case of the traditional ETL approaches used by data warehouses which do not support real time data, this paper with the guidance of the existing literature, will evaluate an extended discussion about the ETL process in DW and its importance in BI development, also elaborate on the motives that increase the need for near real time ETL in Business Intelligence Systems.

Key words: Data warehouse, ETL, near real-time ETL, Business Intelligence

Introduction

Business Intelligence (BI) Systems have been serving organizations in better decision making and improved performance by collecting enormous amount of data that the organizations today have about their employees, suppliers, customers, their preferences etc. The need for fresh new data with data warehouses has become a substantial desideratum. Data warehouse refreshment (integration associated with new data) is traditionally done in the off-line manner. Because of this, even though processes for updating the data area are executed, OLAP users and programs are not able to access any data. This specific list of activities usually takes place in a pre-specified loading time window, to prevent overloading the operational OLTP source systems with the extra workload of this workflow. Active Data Warehousing is the term for a brand new tendency in which DWs usually are updated as frequently as possible, because of the higher demands associated with users for fresh files. This idea is also termed as Real-Time Data Warehousing.

One of the key components of BI system is Data Warehouse. Data warehouses are designed to consolidate data from disparate databases and to better support strategic and tactical decision making needs. The querying or mining abilities of BI system depend on the efficiency of its data warehousing architecture, particularly the way in which the ETL (Extract, Transform and Load) process is performed.

The ETL process takes care of detecting relevant changes in the data from operational databases, extracts it into staging area, transforms it into appropriate formats and loads it into data warehouse. The ETL process, conventionally refreshes the data in the data warehouse in an offline and batch mode [Vas09]. This causes the level of freshness of data in the data warehouse not indicating the latest operational transactions and leads to an issue referred in the scholarly literature as Data Latency [Wib15]. To address this issue, near-real time ETL approaches have emerged as the most promising solution. Therefore the main contributions of this paper are as follows:

1. An extended discussion about the ETL process in DW and its importance in BI development.
2. A brief overview of the techniques used and the problems of traditional ETL processes which induce data latency in the Data Warehouses.
3. Elaborate on the motives that increase the need for near real time ETL in Business Intelligence Systems, we discuss interesting challenges and research issues for this area.
4. The different approaches developed to perform real time ETL and the key problems associated with real time ETL.

1. The role of data warehouse in Business Intelligence

Business intelligence is a broad category of applications and technologies for gathering, providing access to, and analyzing data for the purpose of helping enterprise users make better business decisions. The term implies having a comprehensive knowledge of all factors that affect a business, such as customers, competitors, business partners, economic environment and internal operations, therefore enabling optimal decisions to be made [Kos16]. It leverages technologies that focus on counts, statistics and business objectives to improve business performance. BI applications generally use data gathered from a data warehouse and converts it into actionable insights.

Data warehouse acts as a middleware in Business application architecture, which are essential as direct accessing to operational and transactional data for decision support applications is infeasible. A Data Warehouse (DW) is simply a consolidation of data from a variety of sources that is designed to support strategic and tactical decision making [Pas14]. Data warehouse helps in achieving various goals.

Such as:

A. Data Integrity

The data warehouse concept seeks to integrate data across time and across different subject areas in such a way that users of the warehouse can easily obtain facts about their business.

B. Normalization of data

Normalization includes the logical analysis of the determination of the most simplest, stable data that can be stored in database and makes warehouse more understandable.

C. Information consistency

The metadata layer of the data warehouse enforces information consistency by allowing data within the data warehouse to be defined in business terms as opposed to using database jargon.

2. Traditional ETL

We need to load data warehouse regularly so that it can serve its purpose of facilitating business analysis and keep updated. The process of extracting data from source systems and bringing them into the data warehouse is commonly called ETL [Bin15]. In the context of Data Warehousing, ETL refers to a core process that facilitates congregation of data from disparate sources and gives a single version of truth in an enterprise.

ETL stands for Extract, Transform and Load Definition - A process is used to enable companies to move data from multiple sources, reformat and cleanse them, and then load the data into another area for analysis or operational system for support of the organizations business process.

a) Extracts data from homogeneous or heterogeneous data sources.

In this activity are performed:

• Mapping of the data source that will be needed to meet the information needs of the BI project, whether they are in some ERP or legacy systems.

• Design and execution of the extraction process, which can be performed with the execution of SQL queries next to the databases and / or with the support of own programs for this purpose [Fer16].

b) Transforms the data for storing it in proper format or structure for querying and analysis purpose. In some cases it is necessary to perform an equalization of the data, especially with regard to codes or data that are different in different databases of origin. Thus, data transformations are performed so that there is a uniqueness of these codes [Fer16].

c) Loads it into the final target (database, more specifically, operational data store, data mart, or data warehouse). If the data are all equalized then the data load is performed in the BI solution, either in the Staging Area (where the data are stored and then loaded into the three-dimensional bases (cubes) or simply in two-dimensional tables [Fer16].

All the three phases usually execute in parallel since the data extraction takes time, so while the data are being pulled, another transformation process executes, processing the already received data and prepares the data for loading and as soon as there are some data ready to be loaded into the target, the data loading starts without waiting for the completion of the previous phases [Min16].

As mentioned ETL is a very required process for making different data source to pull at one end. If it is conducted not effectively then there is too much cost wastage behind that.

3. The necessity for near real-time ETL

Traditionally, ETL processes have been responsible for populating and updating the data warehouse both for the bulk load at the initiation of the warehouse and incrementally, throughout the operation of the warehouse in an off-line mode. ETL process updates data warehouse periodically. This implies that data warehouse is not relevant to the current condition, where there is real time data between two updating process. Thus, it makes less accurate analysis result [Wib15].

The other problem is that traditional ETL should be performed at off peak hours. It means operational and analysis activities must be stopped. Based on these problems, there should be a mechanism for updating the data warehouse proximately after the change in the data source, so that the enterprise's needs connected to the latest data can be met. This requirement needs a different approach of ETL to simplify the loading of data into data warehouse in a continuous way unlike the periodic manner used in the traditional ETL approach.

The long term vision for near real time warehousing is to have a self-tuning architecture, where user requirements for freshness are met to the highest possible degree without disturbing the administrators' requirements for throughput and availability of their systems. Clearly, since this vision is founded over completely controversial goals, a reconciliation has to be made:

A more pragmatic approach involves a semi-automated environment, where user requests for freshness and completeness are balanced against the workload of all the involved sub-systems of the warehouse (sources, data staging area, warehouse, data marts) and a tunable, regulated flow of data is enabled to meet resource and workload thresholds set by the administrators of the involved systems. [Vas09]

4. The General Architecture for the near real-time ETL process

Near real time ETL deviates from the traditional conception of data warehouse refreshment, which is performed off-line in a batch mode, and adopts the strategy of propagating changes that take place in the sources towards the data warehouse to the extent that both the sources and the warehouse can sustain the incurred workload [Hal12].

Based on the general architecture of a near real time ETL suggested by Panos Vassiliadis [Vas09], the data warehouse consists of the following components:

1) Data Sources hosting the data production systems that populate the data warehouse.

2) An intermediate Data Processing Area (DPA) where the cleaning and transformation of the data takes place and

3) The Data Warehouse (DW).

Each source can be assumed to comprise a data store and an operational data management system. Changes that take place at the source side have first to be identified as relevant to the ETL process and subsequently propagated towards the warehouse, which typically resides in a different host computer.

Based on this architecture, each source host a Source Flow Regulator (SFlowR) module that is responsible for the identification of relevant changes and transmits them towards the warehouse at periodic or convenient intervals, depending on the policy chosen by the administrators. This period is significantly higher than the one used in the current state-of-practice and has to be carefully calculated on the basis of the source system's characteristics and the user requests for freshness.

Also, the Data Processing Flow Regulator (DPFlowR) module is responsible of deciding which source is ready to transmit data. Once the records have left a certain source, an ETL workflow receives them at the intermediate data processing area. The primary role of the ETL workflow is to cleanse and transform the data in the format of the data warehouse. Apart from these necessary cleansings and transformations, the role of the data processing area is versatile: (a) it relieves the source from having to perform these tasks, (b) it acts as the regulator for the data warehouse, too (in case the warehouse cannot handle the online traffic generated by the source) and (c) it can perform various tasks such as check pointing, summary preparation, and quality of service management. Once all the ETL processing is over, data are ready to be loaded at the warehouse.

A Warehouse Flow Regulator (WFlowR) orchestrates the propagation of data from the DPA to the warehouse based on the existing workload from the part of the end users pretention queries and the QoS "contracts" for data freshness, ETL throughput and query response time.

Many other alternative architectures for near real time ETL have been suggested this year by various researchers, for example: Enterprise Application Integration (EAI), Fast transformations via Capture - Transform - Flow (CTF) processes, Fast loading via micro batch ETL, On-demand reporting via Enterprise Information Integration (EII), which can be used depending on system requirements.

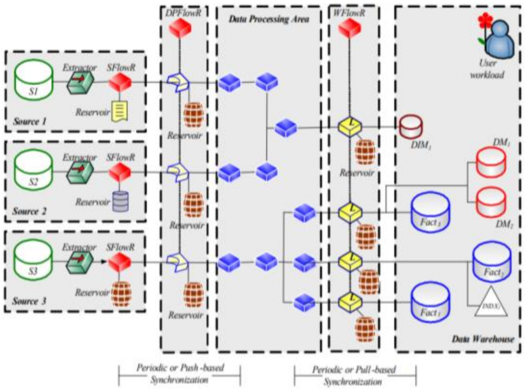

Figure 1: Architecture of near real time data warehouse [Vas09]

5. Some factors that affect near real time processes

Now we present the main factors that may influence near real time abilities in the traditional data warehouse architecture, given the fact that the data integration process has a high cost in those systems. In a traditional DW data integration occurs with the system offline, at dead hours, and a supposedly very large set of data coming from different data sources is extracted, transformed and loaded. After transforming, that data is stored in batches for loading. When one wants to get to near-real time capabilities, this is done by dropping the interval between executions of two consecutive ETL processes, for instance, they can be run with a 5 minutes interval between them. The size of the batches is therefore smaller, but the cost of the ETL process is high even when ran offline. In a near-real time context this process runs with a much higher frequency, over much smaller batches and most probably while the system remains online and serving queries. Another factor is the dropping and recreation of indexes, which is not possible in a near-real time system if it stays online, and may sustain in undesirable delays if the system is taken offline to allow it. As a consequence, an online near real time process is going to be quite slow inserting new data. Therefore is a need for fresh information at any time and permanent online availability of the data warehouse, then it is preferable to use live data warehouse technologies, which we can present in future works.

Conclusions

The topic of this article has been to review basic concepts of ETL process in DW and its importance in the new approaches of business intelligence architecture.

The data in organizations are exponentially increasing and the necessity for timely and accurate insights from the Business Intelligence (BI) systems used in many organizations is also continually growing. Nowadays BI vendors are adopting agile practices to produce mountains of organization's data to facilitate the right information to the right people at the right time and using a proper ETL process might drastically affect the business outcome. The traditional approach of ETL is not appropriate in the changing scenery of business. Today's businesses demand information that are fresh as possible as the value of the business visions declines as it gets older. But given the huge dimensions that BI is expanding, market-based data warehousing requirements, and optimization of ETL processes, there is future scope for further research and development of better approaches towards real-time ETL.

A section has also been dedicated to analyze the influence of a set factors on near-real time capabilities of a traditional data warehouse architecture, we concluded regarding the difficulty of a traditional data warehouse architecture to enforce updates in near real time. The proposal for a solution using real time data warehouse may be the scope of future work.

References

[Kos16] R. Kosaraju, *Business Intelligence and Data warehousing,* http://maxiltechnology.com/Business_Intelligence_and_Datawarehousing.pdf

[Pas14] K. Passi, *Review on role of data warehouse in Business Intelligence*, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3 Issue 7, July 2014.

[Bin15] P. Bindal , P. Khurana, *ETL Life Cycle*, Purnima Bindal et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (2) , 2015, 1787-1791, ISSN:0975-9646

[Fer16] S. Fernandes, *ETL-etapa fundamental de um projeto de BI*, June 2016, http://www.fiveforces.com.br/blog/business-intelligence/etl-etapa-fundamental-de-um-projeto-de-bi.html

[Min16] M. Minhaj, *An Exploratory Study of Near-Real Time ETL Approaches for the Design of Agile Business Intelligence*

Infrastructure. http://sdmimd.ac.in/SDMRCMS/articles/CRM2016/2.pdf

[Hal12] R. Halenar, *Real Time ETL Improvement*, International Journal of Computer Theory and Engineering, Vol. 4, No. 3, June 2012

[Man11] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). *Big data: The next frontier for innovation, competition, and productivity.*

[Vas09] Vassiliadis, P., & Simitsis, A. (2009). *Near real time ETL. In New trends in data warehousing and data analysis* (pp. 1-31).Springer US.

[Wib15] Wibowo, A. (2015, May). *Problems and available solutions on the stage of Extract, Transform, and Loading in near real-time data warehousing* (a literature study).In Intelligent Technology and Its Applications (ISITIA), 2015 International Seminar on (pp. 345-350). IEEE.