


Chapter 14

Challenges and Solutions of Real-Time Data Integration Techniques by ETL Application


Neepa Biswas

 <https://orcid.org/0000-0003-2790-1768>
Narula Institute of Technology, India

Sudarsan Biswas

RCC Institute of Information Technology, India

Kartick Chandra Mondal

 <https://orcid.org/0000-0003-3647-5799>
Jadavpur University, India

Suchismita Maiti

Narula Institute of Technology, India

ABSTRACT

Business organizations are trying to focus from the traditional extract-transform-load (ETL) system towards real-time implementation of the ETL process. Traditional ETL process upgrades new data to the data warehouse (DW) at predefined time intervals when the DW is in off-line mode. Modern organizations want to capture and respond to business events faster than ever. Accessing fresh data is not possible using traditional ETL. Real-time ETL can reflect fresh data on the warehouse immediately at the occurrence of an event in the operational data store. Therefore, the key tool for business trade lies in real-time enterprise DW enabled with Business Intelligence. This study provides an overview of ETL process and its evolution towards real-time ETL. This chapter will represent the real-time ETL characteristics, its technical challenges, and some popular real-time ETL implementation methods. Finally, some future directions towards real-time ETL technology are explored.

DOI: 10.4018/979-8-3693-0413-6.ch014

INTRODUCTION

Traditional Data warehouse (DW) use to store static data. In DW, strategic analysis is performed on Business data which is integrated from heterogeneous data source. The data is captured, aggregated, cleaned and analyzed for deriving better decisions (Wrembel, 2006; Hong et al., 2009). The analytical decision depends not only on data processing applications, but also on the derived data. The data should be accurate, relevant and timely in nature. The more timely processed data ensure the more better decision making. The decision making process is often delayed in traditional DW due to late propagation of data from the source system to DW in time. Presently, organizations want to access up-to-date data for decision making. So, the concept of real-time ETL technique is introduced (Biswas et al., 2020).

In traditional batch processing ETL, DW refreshment is performed in offline mode in daily, weekly or monthly basis (Vassiliadis, 2009; Biswas et al., 2019). Data is extracted from different types of sources, then it is cleaned and transformed and at last loaded into the DW. These activities are generally performed at night during the warehouse downtime. Any interference is unwanted during the loading and query processing on the DW. These historical data is stored for future analysis purpose.

The way of organizations accessing data is rapidly changing. Nowadays organizations want to access real-time transactional data for taking immediate decision. Currently many industries such as stock exchange, e-commerce, air traffic control, telecommunication etc. have the requirements of correct report based on fresh data in DW as operational decision can be made speedy.

This can't be performed on the status report of yesterday.

Nowadays, the web is considered as important source. In this case, the transactional data that emerge at the source side are not always possible to collect later, if off- line refreshment is performed on the warehouse. Besides, the volume of data for analysis is becoming very high and the response time is shortening. So, the demand is increasing for superior ETL tool. Therefore, the time window is shortening for loading in DW. So the main focus for Business Intelligence (Waas et al., 2013) lies in the DW and the ETL process for supporting continuous data flow (Tho and Tjoa, 2003; Polyzotis et al., 2007) and decreasing downtime.

So, how to define fresh data? Freshness is signifying from minutes to seconds or sub-seconds of data flow delay. The trends of “near real-time” (J'org and Dessloch, 2009; Chen et al., 2010; Wibowo, 2015) or “real-time” (Cuzzocrea et al., 2014; Patel and Patel, 2022) is going to be the new challenges in technological solutions. Some commercial systems like Informatica PowerCenter, Infosphere Datastage, Oracle Data Integrator, Talend Open Studio, Azure Data Factory etc. are working towards getting fresh data in DW.

In traditional ETL, data warehouse refreshment is done periodically in off-peak time. Due to periodical updation fresh data could not be accessed for analysis purpose. The late arrival of data in warehouse is termed as data latency. Another problem of traditional ETL is off-peak hour loading when all operational and analysis process should be paused. With the introduction of web applications, e-commerce, retail, banking, network traffic monitoring etc. demand running system in 24x7 hours along with latest updated data. To address these issue “near real-time” and further “real-time” mechanism is introduced. Here the main focus is to minimize data latency by identifying the changes immediately in the operational data source and propagate it to warehouse quickly.

Moving towards achieving “real time” implementation, the simple way is to shorten the refreshment cycle. Generally, five to fifteen minutes latency can be categories as “near real time” approach (Vassiliadis and Simitis, 2008). This solution does not require much changes in the existing operational system.

22 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:

www.igi-global.com/chapter/challenges-and-solutions-of-real-time-data-integration-techniques-by-etl-application/336357

Related Content

Multi-Label Classification

Jesse Readand Albert Bifet (2014). *Encyclopedia of Business Analytics and Optimization* (pp. 1581-1584). www.irma-international.org/chapter/multi-label-classification/107350

Towards Private-Public Research Partnerships Combining Rigor and Relevance in DWH/BI Research: The Competence Center Approach

Anne Cleven, Robert Winterand Felix Wortmann (2010). *International Journal of Business Intelligence Research* (pp. 60-71). www.irma-international.org/article/towards-private-public-research-partnerships/43682

Business Intelligence in the Music Industry Value Chain: Ensuring Sustainability in a Turbulent Business Environment

Hanne Russ, Jean-Pierre Kuilboerand Noushin Ashrafi (2014). *International Journal of Business Intelligence Research* (pp. 50-63). www.irma-international.org/article/business-intelligence-in-the-music-industry-value-chain/108012

Business Intelligence is No 'Free Lunch': What We Already Know About Cost Allocation – and What We Should Find Out

Johannes Epple, Robert Winter, Stefan Bischoffand Stephan Aier (2018). *International Journal of Business Intelligence Research* (pp. 1-15). www.irma-international.org/article/business-intelligence-is-no-free-lunch/203654

A Combined Multi-Criteria Decision-Making Framework for Process-Based Digitalisation Opportunity and Priority Assessment (DOPA)

Nihan Yldrm, Birden Tulu Siyahi, Ouz Özbek, mran Ahioluand Almira Selin Kahya (2022). *International Journal of Business Analytics* (pp. 1-22). www.irma-international.org/article/a-combined-multi-criteria-decision-making-framework-for-process-based-digitalisation-opportunity-and-priority-assessment-dopa/298018