Data Integrity in Real-time Datawarehousing

Imane LEBDAOUI, Ghizlane ORHANOU, Said EL HAJJI

Abstract— **Information freshness and integrity are the main pillar of making sound decision. Real-time datawarehousing is a trend of delivering fresh information to decision making processes in "Real-time". However, to flow up real-time data into the datawarehouse, systems may skip some necessary treatments. Thus, data integrity may be threatened. This paper discusses data integrity issues towards the need of accessing to Real-time datawarehouses and introduces an IA-RTDWg model that preserves both integrity and availability.**

***Index Terms*— datawarehouse, integrity, real-time data-warehousing.**

I. INTRODUCTION

Datawarehouses (DW) gather data that flow up from various heterogeneous data-sources into one integrated repository and feed dedicated decision making systems. Sound decisions rely on fresh data that respect CIA-triad (Confidentiality, Integrity and Availability).

A DW can be updated according to fixed frequency or continuously in real-time. It is observed that when increasing refreshment frequency and shortening the refreshment intervals, some anomalies appear [1]; data integrity may be threatened consequently.

Fig. 1. Impact of increasing refreshment frequency on data integrity

According to our bibliographic study, no attempt has been yet dedicated to identify the best balancing between real-time datawarehousing (RTDWg) requirements and CIA rules especially integrity. We introduce, in the present papaer, a new model: IA-RTDW that will fulfill RTDWg while insuring data integrity and data availability.

The remainder of this paper is structured as follows: Section 2 presents real-time datawahousing and data integrity basis. A comparison between data integrity principles and real-time datawahousing requirements are given in section 3. Section 4 is dedicated to present a new IA-RTDWg model. Conclusion and perspectives are given in section 5.

Manuscript received March 16, 2013; revised April 06, 2013.

Imane LEBDAOUI, Laboratory of Mathematics, Computing and Applications, Faculty of Sciences, University of Mohammed V–Agdal, BP.1014 RP. Rabat, Morocco (e-mail: imane.lebdaoui@gmail.com).

Said EL HAJJI, Laboratory of Mathematics, Computing and Applications, Faculty of Sciences, University of Mohammed V–Agdal, BP.1014 RP. Rabat, Morocco (e-mail: elhajji@fsr.ac.ma).

Ghizlane ORHANOU, Laboratory of Mathematics, Computing and Applications, Faculty of Sciences, University of Mohammed V–Agdal, BP.1014 RP. Rabat, Morocco (e-mail: ghizlane.orhanou@gmail.com).

II. REAL-TIME DATAWAHOUSING AND DATA INETGRITY BASIS

A. Processing data in Real-time

When data move from operational sources to destination (DW), it is amenable to many processing: after being taken by Changes Data Capture systems (CDC), data are extracted, transformed and loaded into the appropriate tables in the DW.

Because of the increasing demands for fresh and just-updated information, the 24x7 operations and the recent trends of business globalization, enterprises are more oriented to RTDWg. Implementing this system involves high technologies especially those related to ETL systems since they are the cornerstone of any DW solution. However, data freshness should always be driven by the business requirements, not the technology itself [3].

RTDWg implies **real-time access** to **real-time** data.

- "Real-time access" comprises accessibility and availability of data in the DW in real-time,
- "Real-time data" means that data in the DW must have the same level of freshness of data that have just been modified by operational systems.

B. Types of data

In order to handle data change in real-time, real-time DW uses specific criteria or measurements: real-time tables, real-time partition and smart query tools to prioritize data integration according to their importance [3], thus data are processed differently. We give the following categorization of data.

Data can be static or dynamic.

- Static data: are data that might receive no important and frequent changes since their first loading into the DW (Ex: Social Code, national Code)
- Dynamic data: are living-data which change frequently or continuously. Here, we distinguish between:
  - Real-time data (RT-data): are timely data whose value may be used to achieve real-time reports, analysis or decisions. This type is the most important for real-time decision making systems including RTDWg.

ISBN: 978-988-19252-9-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

- Non Real-time data (NRT-data): are dynamic data whose value may be changed but are not used for real-time purpose.

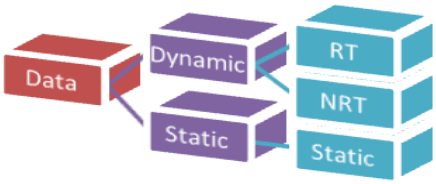

Fig. 2. Types of data

C. RT-Datawarehousing Challenges

Besides traditional DW challenges, RT DW has to face the problem of applying data change in a timely fashion to a datawarehouse that also remains available for query [5]. This definition implies two major challenges the RTDWg must face: data quality and data availability.

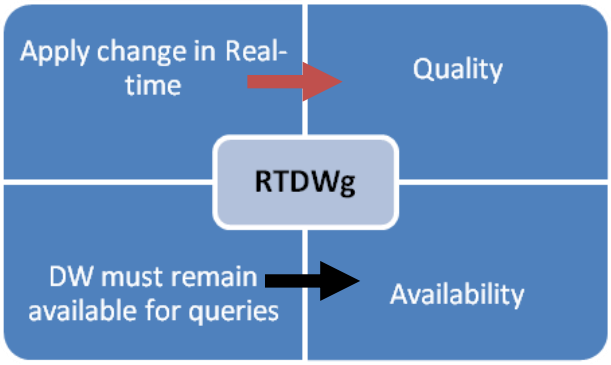

Fig. 3. RTDWg principal major challenges

Because data change undergo several treatments when flowing up from sources to destination, it is necessary that quality must be checked whenever data move from one component to the next one. Moreover, in real-time datawarehouses (RTDW), data must attain destination according to organization temporal requirements (timely manner or with some latency).

one of the barrier of adoption analytics solutions, particulary datawarehousing, is data quality. In fact, a survey conducted in 2012, [8], shows that 46% of respondents consider that data quality problems are the biggest barrier to Enterprisewide BI/Analytics Adoption.

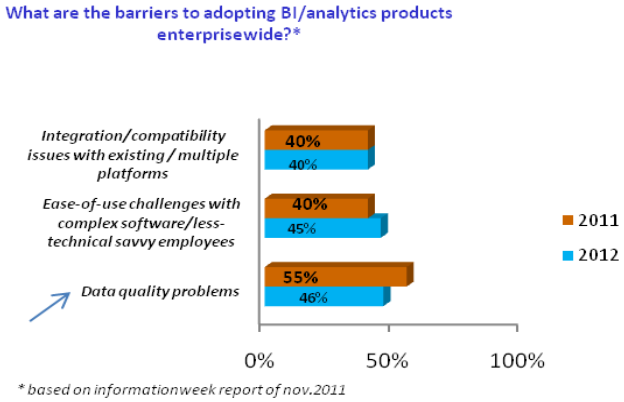

Fig. 4. Barriers to Enterprisewide BI/Analytics Adoption

Data quality is a set of six creteria (fig. 4). One criterion is data integrity that is the focus of this paper.

Fig. 5. Data Quality Dimensions [5]

D. Data integrity basis

Integrity refers to the trustworthiness of the data resources. If data are missing important relationship linkages and are unable to link related records together, then they may actually introduce duplication across all systems [4]. Data integrity is the assurance that the data can be accessed or modified by those authorized. Besides, it is one important criteria of ensuring data quality (fig. 5) which is one big challenge for RTDWg [8].

III. Data Integrity versus Real-time Datawahousing

A. I-A dilemna in RTDWg

When discussing integrity issues throughout RTDWg requirements, many points have to be considered:

- To fulfill integrity, (I), data must undergo many controls mostly based on hard-coded systems. Thus, additional time is needed to control and validate data. In such circumstances, information may be not available in the destination DW in timely fashion, thus the real-time requirement is not totally respected.

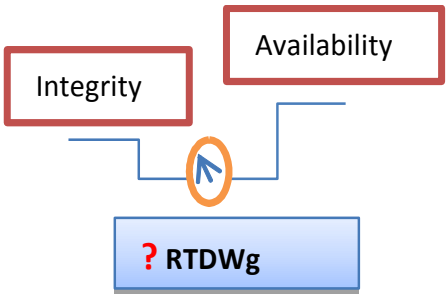

Fig. 6 Prioritize integrity in RT –DW

- To fulfill data availability, (A), in real-time fashion, many systems avoid cleansing and verifying data. Data integrity is consequently compromised. In practice, to achieve the maximum of data availability

ISBN: 978-988-19252-9-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

and accessibility, systems are designed to accept unverified data [7].

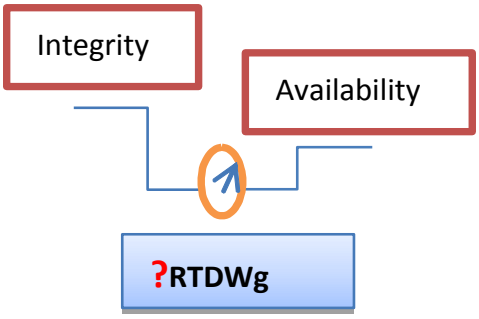

Fig. 7. Prioritize availability in RT –DW

Consequently, there is a real need to correctly balance between IA (Integrity and Availability) corners to ensure the effective RTDWg.

In one hand, RTDWg systems must guarantee data integrity. If someone accesses data when they are just captured by CDC-mechanisms or transformed by ETL tools and modifies them in illegal way, integrity is consequently despoiled. If a program or a tool in a DW solution didn't process data according to predefined rules and if a step of datawarehousing is ignored or skipped, the integrity may also be corrupted.

In the other hand, RTDWg requires that data must be available in the DW in real-time or at least within the allowed time interval. Therefore, the availability of data is required to be insured inside the DW, when a user or program needs to use the changed data for decision purpose. Thus, it can be verified and measured just once data reach the destination table in the DW.

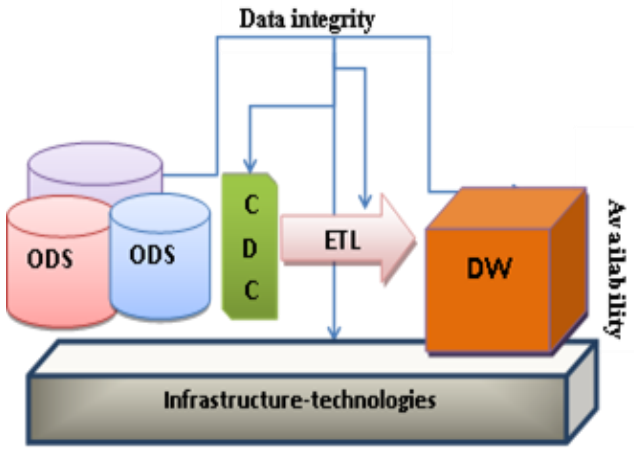

Fig. 8. Places where to ensure I-A corners in RT DW

B. Integrity violation in RTDWg

Although data integrity must guarantee correctness and accuracy of data within datawarehouses or databases, to ensure some level of worthiness, we explain how combining RTDWg challenges and data integrity constraints might cause violations or conflicts.

a. Temporary duplicated DW schema without constraints

To handle data in real-time, one solution suggests the replication of DW structure. In this replicated structure, tables have no contents or indexes or primary keys or constraints of referential integrity [2]. This approach is based on the following principle: new insertion into empty tables with no constraints is performed much faster than in large sized tables [2]. When conditions are convenient, data in the duplicated schema are replicated into the original one. This approach makes data available for decision making processes although it threatened integrity constraints. Furthermore, it requires additional technical resources (hardware, CPU, storage space...) to enable new creation of extra DW schema.

b. Partitioning and parallelism

Partitioning and parallelism mechanisms are widely used by real-time systems (RT-ETL, RT-partition) [6]. The concept relays on processing small data volume through nodes to enable real-time and fast data treatment. Parallelism and partitioning imply extra processors and techniques of hashing and merging. These techniques allow high availability of data in the destination. However, there is a need to schedule the simultaneous loading of different constructs or parts of them [6]. Thus, data integrity may be hampered when a partition does not reach destination in the correct scheduled way.

IV. IA-RTDWG MODEL

For a given real-time DW solution, we consider that data integrity constraints are totally respected inside operational sources. In normal situation, when there are data changes, they are detected and caught by real-time Change Data Capture then transformed and loaded by ETL tools. Thus, in case of big amount of data change and important DW size, it is necessary to override integrity constraints brakes in order to get data available for decision support systems in real-time.

Problem statement: Given that queries and decision making systems need that data must be accurate and available in timely manner in the DW, we present our IA-RTDWg model based on duplicating fact table's schema.

IA-RTDWg is to fulfill RTDwg with respect of data inetrity while assuring data availability in the DW in real-time.

The duplication concerns only dynamic real-time data and occurs when the DW is busy and could not integrate new data change immediately.

The duplicated fact tables:

- have no contents but keep the same original integrity constraints;
- receive the fresh RT-data when the DW is busy;
- are joined with the original fact tables and feed RT queries and RT dashboards.
- When the DW is judged not busy, the duplicated fact tables are replicated into to original fact tables, thereafter, are emptied.

ISBN: 978-988-19252-9-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

- Every duplicated fact table is linked with a witness table that contains primary keys of the duplicated fact table and two extra attribute: a flag, that mentions whether the change was integrated into the real DW, and a date attribute, that shows when the change was made.

  For example, let's consider the following star schema of a simplified DW, which contains one fact table and two dimension tables.

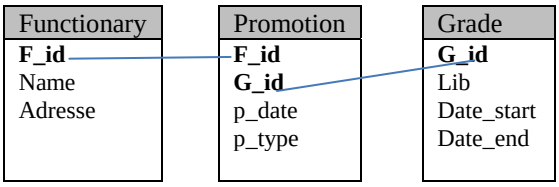

Fig. 9 Sample of star schema

When a data change occurs and concern the fact table of Fig. 9, IA-RTDWg model creates the following duplicated fact table with the related witness table:

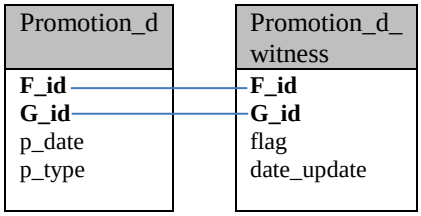

Fig. 10 Sample of duplicated fact table

Besides, the IA-RTDWg model duplicates dimension tables once they are concerned by data change and when the DW is judged as busy to receive new data.

If for workload reasons, the DW is not able to integrate data from the temporary fact table; IA-RTDWg model creates a secondary duplicated fact table and handle data like they are in one DW with one fact table. In case of the DW is always unable to receive extra data, the secondary table is replicated into the primary one and is emptied thereafter and becomes ready to receive further data changes. Once the primary table is replicated into the DW, it will be emptied and the secondary fact table deleted, and so on.

V. Conclusion and Perspectives

RTDW whose data integrity isn't continuously guaranteed is a system that leads to incorrect decisions. When data integrity is first established and implemented through constraints, it must be respected along all processes involved. Furthermore, integrity must be insured each time data move to the next level in the DW solution and across infrastructure technologies. It must safeguard the whole system against unauthorized access to data and the data history.

In this work, we have discussed data integrity violation in regards of RTDWg temporal constraints. We have also presented our model that is based on a temporary duplication of fact tables that are concerned by data change while conserving their integrity constraints.

Future work involves the test of IA-RTDWg model under real world systems, the analysis of results and comparison with similar approaches. In the future, we plan to measure the impact of IA-RTDWg model on the performance of involved system.

References

[1] Thomas JÄorg and Stefan Dessloch, Near Real-Time Data Warehousing Using State-of-the-Art ETL Tools. BIRTE 2009: 100-117,2009

[2] Ricardo Jorge Santos and Jorge Bernardino, Real-Time DataWarehouse Loading Methodology", IDEAs'o8, ACM 978-1-60158-188-0/08/09, 2008.

[3] Li Chen , Wenny Rahayu , David Taniar, Towards Near Real-Time Data Warehousing, Proceedings of the 2010 24th IEEE International Conference on Advanced Information Networking and Applications, p.1150-1157, April 20-23, 2010

[4] Kamal Kakish, Theresa A. Kraft, ETL Evolution for Real-Time Data Warehousing, Proceedings of the Conference on Information Systems Applied Research ISSN: 2167-1508 New Orleans Louisiana, USA,2012

[5] Rittmanmead , Realtime Data Warehouse Challenges, http://www.rittmanmead.com/2010/05/realtime-data-warehouse-challenges-part-1/, 2010

[6] Panos Vassiliadis, Alkis Simitsis, Near Real Time ETL. AoIS Vol.3, 2008 ISBN 978-0-387-87430-2,2008

[7] J.Korsunsky, http://www.dbta.com/Articles/Editorial/Trends-and-Applications/Preserving-Data-Integrity-73015.aspx, 2011

[8] Doug Henschen, reports.informationweek.com (november 2011), Report ID: R3551111, 2011

ISBN: 978-988-19252-9-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)