A Survey of Real-Time Data Warehouse and ETL

Fahd Sabry Esmail Ali

Abstract— Data warehouses (DWH) are typically designed for efficient processing of read only analysis queries over large data, allowing only offline updates at night. The current trends of business globalization and online business activities available 24/7 means DWH must support the increasing demands for the latest versions of the data. Real-Time Data Warehousing aims to meet the increasing demands of Business Intelligence for the latest versions of the data. Informed decision-making is required for competitive success in the new global marketplace, which is fraught with uncertainty and rapid technology changes. Decision makers must adjust operational processes, corporate strategies, and business models at lightning speed and must be able to leverage business intelligence instantly and take immediate action. Sound decisions are based on data that is analyzed according to well-defined criteria. Such data typically resides in a Database Warehouse for purposes of performing statistical and analytical processing efficiently. Achieving Real-Time Data Warehousing is highly dependent on the choice of a process in data warehousing technology known as Extract, Transform, and Load (ETL). This process involves: 1) Extracting data from outside sources; 2) Transforming it to fit operational needs; and 3) Loading it into the end target (database or data warehouse). Not all ETL's are equal when it comes to quality and performance. As such, optimizing the ETL processes for real time decision making is becoming ever increasingly crucial to today's decision-making process. An effective ETL leads to effective business decisions and yields extraordinary decision- making outcomes. This study overviews the theory behind ETL and raises a research vision for its evolution, with the aim of improving the difficult but necessary data management work required for the development of advanced analytics and business intelligence.

Index Terms— Data Warehousing, Extract, Transform, Load, ETL, Data Warehouse Loading, Real-Time

1 INTRODUCTION

Data warehousing provides architectures and tools for business executives to systematically organize, understand, and use their data to make strategic decisions. With the explosive advent of the Internet, broadband communication, mobile computing, and access to cloud computing, the past couple of decades gave a new meaning to the phrase "information overload". Companies need to consider how to adopt and utilize real-time data and information into the fabric of their decision-making or risk falling behind their competitors. Indeed, this is an era of unprecedented data copiousness and accumulation. The utter variety of new information available on diverse platforms of electronic data sources has changed the way we live, collaborate, conduct business, undertake research, and make decisions; however, the increased reliance upon networked data has introduced unique data quality challenges. Organizations demand for quick access to new insights has led to predictive analytics for forecasting emerging demands, risks and opportunity. Advanced analytics apply statistical and predictive algorithms to forecasting, correlation, and trend analysis. In contrast, traditional data Warehousing and Business Intelligence have typically been associated with historical analysis and reporting. Advanced statistical and predicative analysis takes advantage of the large data sets (big data) stored within data warehouses to foresee risk, anticipate customer demand, and formulate more successful product and service offerings.

- *Fahd Sabry Esmail Ali is currently pursuing masters degree program in Information System in Arab Academy, Demonstrator in Management Information Systems department in Modern Academy for Computer Science and Information Technology, Cairo, Egypt. E-mail: fahdsabry985@gmail.com*

The advanced analytics are highly dependent on access to the most recent, real-time business data; hence, data warehouses must have instantaneous real-time access to business transactions. As the advanced analytics requirements get closer to real- time, the software applications must tolerate some amount of data incompleteness or inaccuracy, as it is not financially or technically feasible to provide 100% of the data within such strict time requirements (Henschen, 2011). According to William H.Inmon, a leading architect in the construction of data warehouse systems, "*A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision making process*".

2 DATA WAREHOUSE CHARACTERISTICS

The major features of a data warehouse. The four keywords, *subject-oriented*, *integrated*, *time-variant*, and *nonvolatile*, distinguish data warehouses from other data repository systems, such as relational database systems, transaction processing systems, and file systems.

2.1 Subject-oriented

A data warehouse is organized around major subjects, such as customer, supplier, product, and sales. Rather than concentrating on the day-to-day operations and transaction processing of an organization, a data warehouse focuses on the modeling and analysis of data for decision makers. Hence, data warehouses typically provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process.

2.2 Integrated

A data warehouse is usually constructed by integrating multiple heterogeneous sources, such as relational databases, flat files, and on-line transaction records.

2.3 Time-variant

Data are stored to provide information from a historical perspective (e.g., the past 5-10 years). Every key structure in the data warehouse contains, either implicitly or explicitly, an element of time.

2.4 Nonvolatile

A data warehouse is always a physically separate store of data transformed from the application data found in the operational environment. Due to this separation, a data warehouse does not require transaction processing, recovery, and concurrency control mechanisms.

3 STARS, SNOWFLAKES, AND FACT CONSTELLATIONS

A data warehouse, however, requires a concise, subject- oriented schema that facilitates on-line data analysis. The most popular data model for a data warehouse is a multi- dimensional model. Such a model can exist in the form of a star schema, a snowflake schema, or a fact constellation schema.

3.1 Star Schema

The most common modeling paradigm is the starschema, in which the data warehouse contains a large central table (fact table) containing the bulk of the data, with no redundancy, and a set of smaller attendant tables (dimension tables).

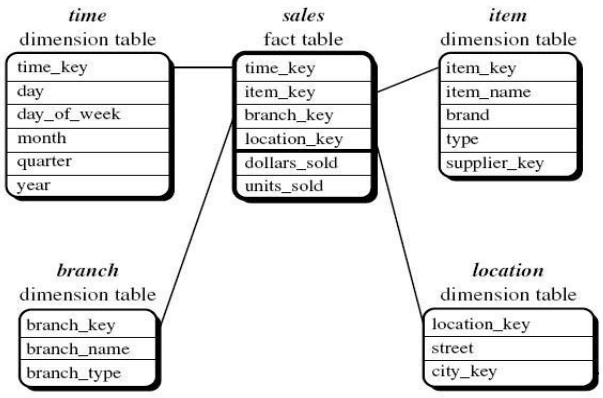

Fig. 1 A simple star schema

3.2 Snowflake Schema

The snowflake schema is a variant of the star schema model, where some dimension tables are normalized, thereby further splitting the data into additional tables. The resulting schema graph forms a shape similar to a snowflake.

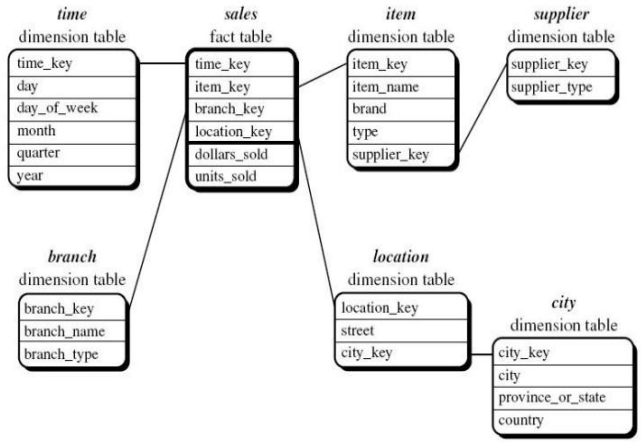

Fig. 2 A simple snowflake schema

3.3 Galaxy Schema

Sophisticated applications may require multiple fact tables to share dimension tables. This kind of schema can be viewed as a collection of stars, and hence is called a galaxy schema or a fact constellation.

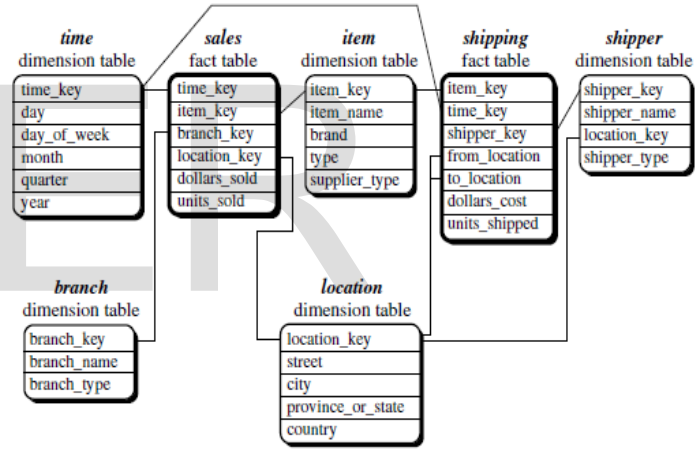

Fig. 3 Galaxy schema of a data warehouse for sales and shipping

4 DATA WAREHOUSE ARCHITECTURE

Data warehouses often adopt a three-tier architecture.

4.1 The Bottom Tier

The bottom tier is a warehouse database server that is almost always a relational database system. Back-end tools and utilities are used to feed data into the bottom tier from operational databases or other external. These tools and utilities perform data extraction, cleaning, and transformation.

4.2 The Middle Tier

The middle tier is an OLAP server that is typically implemented using either a relational OLAP (ROLAP) model, that is, an extended relational DBMS that maps operations on multidimensional data to standard relational operations; or a multidimensional OLAP (MOLAP) model, that is, a special- purpose server that directly implements multidimensional data and operations.

4.3 The Top Tier

The top tier is a front-end client layer, which contains query and reporting tools, analysis tools, and/or data mining tools, extraction, cleaning, and transformation.

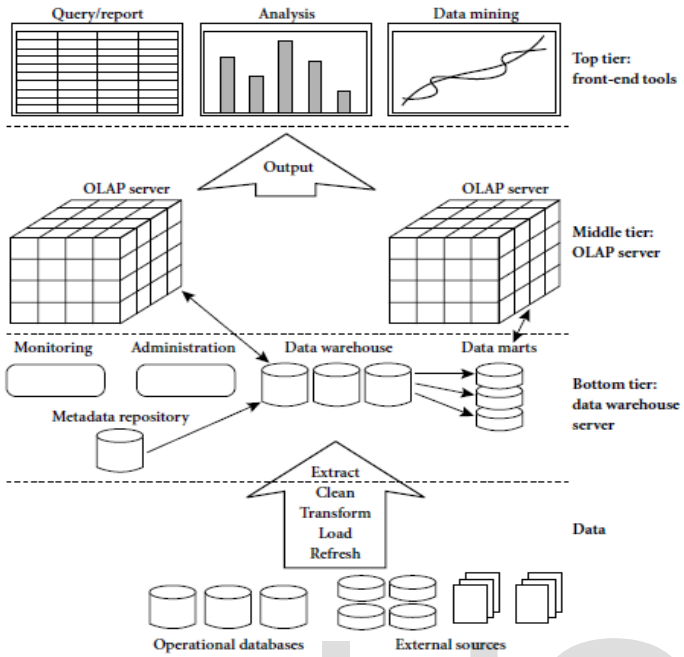

FIG. 4 A THREE-TIER DATA WAREHOUSING ARCHITECTURE

Four different views regarding the design of a data warehouse must be considered: the top-down view, the data source view, the data warehouse view, and the business query view.

4.4 Top-Down View

The top-down view allows the selection of the relevant information necessary for the data warehouse for matching the current and future business needs.

4.5 Data Source View

The data source view exposes the information being captured, stored, and managed by operational systems. This information may be documented at various levels of detail and accuracy, from individual data source tables to integrated data source tables. Data sources are often modeled by traditional data modeling techniques, such as the entity-relationship model or CASE (computer-aided software engineering) tools.

4.6 Data Warehouse View

The data warehouse view includes fact tables and dimension tables. It represents the information that is stored inside the data warehouse, including pre-calculated totals and counts, as well as information regarding the source, date, and time of origin, added to provide historical context.

4.7 Business Query View

The business query view is the perspective of data in the data warehouse from the viewpoint of the end user.

5 DATA WAREHOUSE DESIGN

A data warehouse can be built using a top-down approach, a bottom-up approach, or a combination of both. The top- down approach starts with the overall design and planning. It is useful in cases where the technology is mature and well known, and where the business problems that must be solved are clear and well understood. The bottom-up approach starts with experiments and prototypes. This is useful in the early stage of business modeling and technology development. It allows an organization to move forward at considerably less expense and to evaluate the benefits of the technology before making significant commitments. In the combined approach, an organization can exploit the planned and strategic nature of the top-down approach while retaining the rapid implementation and opportunistic application of the bottom- up approach. In general, the warehouse design process consists of the following steps:

A. Choose a Business Process to Model.
B. Choose the Grain of the Business Process.
C. Choose the Dimensions.
D. Choose the Measures.

6 DATA WAREHOUSE & ETL

The typical ETL-based data warehouse uses staging, integration, and access layers to house its key functions. The staging layer or staging database stores raw data extracted from each of the disparate source data systems. The integration layer integrates the disparate data sets by transforming the data from the staging layer often storing this transformed data in an operational data store (ODS) database.

The integrated data is then moved to yet another database, often called the data warehouse database, where the data is arranged into hierarchal groups often called dimensions and into facts and aggregate facts. ETL plays an important role in data warehousing architecture since these ETL processes move the data from transactional or sources systems to data staging areas and from staging areas into the data warehouse. Demands for real-time data warehousing result from the recent trends of business globalization, 24x7 operations, ever increasing data volumes, competitive pressures, demanding customers and increased demand by decision makers for real- time data. (Ankorion, I., 2005). These current trends require business to have access to the most updated data for analysis and statistical purposes, which necessitates a requirement for building real-time data warehousing and ETL. Techniques to achieve real-time data warehousing include the Change Data Capture (CDC) technique and the integration of change data capture with existing ETL processes to maximize the performance of ETL and achieve real time ETL (Jörg, T., Deßloch, S., 2008). The CDC integration with existing ETL tools provides an integrated approach to reduce the amount of information transferred while minimizing resource requirements and maximizing speed and efficiency (Tank, D,. M. et al., 2010). In contrast, migrating the data into data warehouse using conventional ETL tools has a latency problem with the large volumes of data sets because ETL

processes consume substantial CPU resources and time for large data sets.

7 ETL OPERATIONS

7.1 Data Extraction

A Taking out the data from a variety of disparate source systems correctly is often the most challenging aspect of ETL. Objective: Convert the data into a single format which is appropriate for transformation processing.

Types:

Full extraction: In this type of extraction data from the source system is completely extracted.

Incremental extraction: In this type of extraction only the changes made to the source systems. Example) Change data capture (CDC) is mechanism that uses incremental extraction.

7.2 Data Transformation

This converts data from legacy or host format to warehouse format using similar Steps*:*

- Selecting only certain columns to load.
- Translating coded values.
- Encoding free-form values.
- Deriving a new calculated value.
- Sorting.
- Joining data from multiple sources.
- Generating surrogate-key values.
- Splitting a column into multiple columns.
- Aggregation.
- Applying forms of simple, complex data validation.

7.3 Data Load

Loads the data into the end target, usually the data warehouse. Some data warehouses may overwrite existing information with cumulative information; frequently updates with extracted data are performed on hourly, daily, weekly, or monthly basis.

Mechanisms to load a DWH include:

- SQL loader: generally used to load flat files into data warehouse.
- External tables: this mechanism enables external data to be used as a virtual table which can be queried and joined before loading into the target system.
- Oracle Call Interface (OCI) and direct path Application Programming Interface (API): are methods used when the transformation process is done outside the database.
- Export/Import: this mechanism is used if there are no complex transformations and data can be loaded into target data warehouse

8 EVOLUTION OF ETL

With the evolution of Business intelligence, ETL tools have undergone advances and there are three distinct generations of ETL tools.

The First-generation ETL tools were written in the native code of the operating system platform and would only execute on the native operating system. The most commonly generated code was COBOL code because the first generation data was stored on mainframes. These tools made the data integration process easy since the native code performance was good but there was a maintenance problem.

Second generation ETL tools have proprietary ETL engines to execute the transformation processes. Second generation tools have simplified the job of developers because they only need to know only one programming language i.e. ETL programming. Data coming from different heterogeneous sources should pass through the ETL engine row by row and be stored on the target system. This was a slow process and this generation of ETL programs suffered from a high performance overload.

Third Generation ETL tools have a distributed architecture with the ability to generate native SQL. This eliminates the hub server between the source and the target systems. The distributed architecture of third generation tools reduces the network traffic to improve the performance, distributes the load among database engines to improve the scalability, and supports all types of data sources. Third Generation ETL uses relational DBMS for data transformations. In this generation the transformation phase does processing of data rather than row by row as in second generation ETL tools. "In the ETL architecture, all database engines can potentially participate in a transformation—thus running each part of the process where it is the most optimized. Any RDBMS can be an engine, and it may make sense to distribute the SQL code among different sources and targets to achieve the best performance. For example, a join between two large tables may be done on the source" (De Montcheuil, Y., 2005). RDBMS have power for data integration; ETL tools are taking the advantage of this feature of the RDBMS to improve their performance.

9 REAL-TIME DWH AND ETL TECHNIQUES

Increasingly there is a need to support and make business decisions in near real-time based on the operational data itself.

Typical ETL architectures are batch update oriented and cause a significant lag in the currency of information at the data warehouse.

We question the performance effectiveness of typical batch ETL architectures and near real-time based updates on operational data and raise the questions to address instant-time research in order to address the timely business-decision making process. We raise the issue that the current ETL process needs to move away from periodic refreshes to continuous updates; however online updating of data warehouses gives rise to challenges of view synchronization and resource allocations. "To cope with real-time requirements, the data warehouse must be able to enable continuous data integration, in order to deal with the most recent business data" (Santos, R., J., Bernardino, J., 2009).

View synchronization problems arise when views are composed of data derived from multiple data sources being updated indiscriminately. Resource challenges result when there are conflicting resource demands of long-term analysis queries

in the presence of concurrent updates. In traditional ETL tools, loading is done periodically during the downtime and during this time no one can access the data in data warehouse. The separation between querying and updating clearly simplifies several aspects of the data warehouse implementation, but has a major disadvantage that the data warehouse is not continuously updated. (Polyzotis, N., et al. 2007).

Traditional ETL tools are not capable enough to handle such continuous inserts or updates with no data warehouse down time. In real time data warehousing loading is done continuously as opposed to a periodic basis in traditional approaches. One approach to the general architecture of a near real time data warehouse consisting of the following elements:

(a) Data Sources hosting the data production systems that populate the data warehouse, (b) an intermediate Data Processing Area (DPA) where the cleaning and transformation of the data takes place and (c) the Data Warehouse (Vassiliadis, P., Simitsis A., 2008).

The role of the data processing area (DPA) is to: a) cleanse and transform the data in the format required by the DW; b) act as the regulator for the data warehouse (in case the warehouse cannot handle the online traffic generated by the source); and c) perform various tasks such as check pointing, summary preparation, and quality of service management.

"A Warehouse Flow Regulator (WFlowR) orchestrates the propagation of data from the DPA to the warehouse based on the current workload from end users posing queries and the requirements for data freshness, ETL throughput and query response time." (Vassiliadis, P., Simitsis A., 2008). Third generation ETL tools are using techniques to achieve real time data warehousing without causing downtime. Some of the real time ETL techniques are found in the research of J. Langseth (Langseth, J., 2008) that include:

Near real time ETL: The cost effective solution for applications that do not have a high demand for real time data is to just increase the frequency of loading, for ex: from daily to twice a day.

1) Direct Trickle feed: In this approach true real time data can be achieved by continuously moving the changed data from the source systems by inserting or updating them to the fact tables. There is a scalability problem with this approach because complex queries don't perform well with continuous updates. Constant updates on tables, which are being queried by reporting or OLAP tools leads to degrading the query performance of the data warehouse.

2) Trickle and flip: In this approach, data is inserted or updated into staging tables which are in the same format as target tables. The real time data is stored in staging tables, which have same format as historical data in target tables. The data warehouse can access fresh data instantly by getting a copy from the staging tables into the fact tables, the time window for refreshing the data warehouse can vary from hours to minutes.

3) External real time data cache: In this approach real time data is stored outside data warehouse in an external real time data cache (RTDC). The function of RTDC is to load the real time data into database from source systems. It resolves the query contention and scalability problem by directing the queries to RTDC which access real time data. With this approach, there is no additional load on the data warehouse as the real time data lies on separate cache data base. It provides up-to-the-second data and users don't wait for queries to get executed because they are so quick (Langseth, J., 2008).

The physical data residing in OLAP is in its de-normalized form for query processing while relational online analytical processing (ROLAP) needs data to be in 3rd Normal form because it uses the relational queries for processing the data. Multidimensional analytical processing (MOLAP) can be used because data is built from a data cube, which is separate from transactional data.

10 REAL-TIME DATA BUSINESS INTELLIGENCE

One proposal for real-time business intelligence architecture requires that the data delivery from the operational data stores to the data warehouse must occur in real-time in the format referred to data streams of events (D. Agrawal, 2009). The usage of real-time data event streams eliminates the reliance on batched or offline updating of the data warehouse.

This architecture also introduces the middleware technology component, referred to as the stream analysis engine. This stream analysis engine performs a detailed analysis of the incoming data before it can be integrated into the data warehouse to identify possible outliers and interesting patterns. The goal of the stream analysis process is to "extract crucial information in real-time and then have it delivered to appropriate action points which could be either tactical or strategic" (D. Agrawal, 2009). Complex Event Processing Engines (CEP engines) such as Stream-base enable business users to specify the patterns or temporal trends that they wish to detect over streaming operational data known as events. Decision makers can then take appropriate actions when specific patterns occur.

The origins or CEP engines were in the financial domain where they were applied to algorithmic stock trading. More recently they are being applied to make decisions in real-time such as the click stream analysis of manufacturing process monitoring (S. Chaudhuri, et. al 2011).
The arrival of events from the input streams trigger the query processing, and the queries are performed continuously as long as events arrive in the input stream. One major technical challenge is that the continuous running queries may reference data in the data base and impact near real-time requirements. A major challenge is that algorithms which require multiple passes over the data are no longer feasible for streaming data.

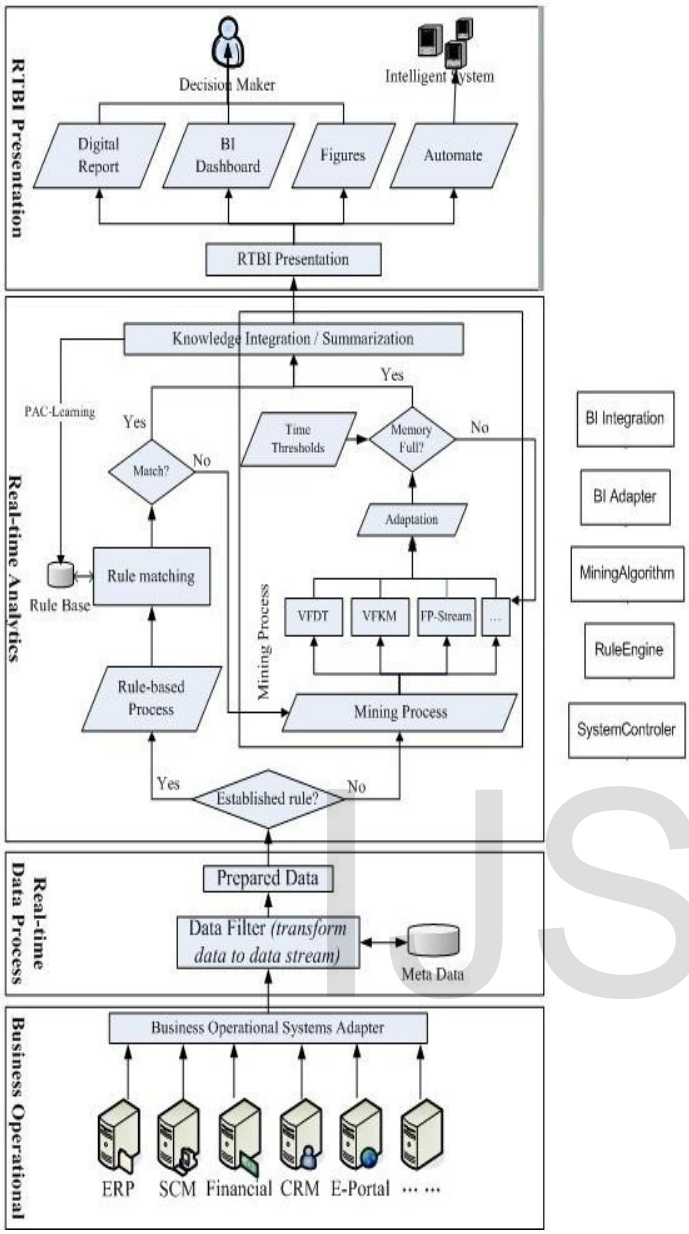

FIG. 5 REAL-TIME BI ARCHITECTURE

11 CONCLUSION

As the role of enterprises becomes increasing real-time such as real-time BI will be increasing important to such companies. In the traditional ETL approach, the most current information is not available. With the increase demands by businesses for real-time Business Intelligence and Predictive Analytics, there is a need to build ETL tools, which provide real-time data into data warehouses. Not every analysis task warrants real-time analysis. The trade-off between the overhead of providing real-time business intelligence and data warehousing, and the need for such an analysis calls for serious research and consideration. Otherwise, or the resulting system may have prohibited costs associated with it (D. Agrawal, 2009). The underlying technology components and custom solutions are prohibitively expensive. The importance, complexity and criticality of such an environment make real-time BI and DW a significant topic of research and practice; therefore, these issues need to be addressed in the future by both the industry and the academia (Vassiliadis, P., Simitsis A., 2008).

REFERENCES

[1] Agrawal, D., (2009), The Reality of Real-Time Business Intelligence, Proceedings of the 2nd International Workshop on Business Intelligence for the Real Time Enterprise (BIRTE 2008), Editors: M. Castellanos, U. Dayal, and T. Sellis, Springer, LNBIP 27, 75-88.

[2] Ankorion, I. (2005). Change data capture: Efficient ETL for Real-Time BI. Information Management, 15(1), 36-36. Retrieved May 29, 2012.from:http://search.proquest.com/docview/214690875?accountid=14584

[3] Athanassoulis, M., Chen, M., S., Ailamaki, A., Gibbons, P. B., and Stoica R., (2011), MaSM: Efficient Online Updates in Data Warehouses, In Proceedings of the 2011 International Conference on Management of Data (SIGMOD '11, ACM, New York, NY, USA, 865-876. DOI=10.1145/1989323.1989414 Retrieved June 14, 2012 from http://doi.acm.org/10.1145/1989323.1989414

[4] Behrend, A., Jörg, T., (2010), Optimized incremental ETL Jobs for Maintaining Data Warehouses. In Proceedings of the Fourteenth International Database Engineering & Applications Symposium (IDEAS '10). ACM, New York, NY, USA, 216-224. DOI=10.1145/1866480.1866511 Retrieved May 29, 2012 from http://doi.acm.org/10.1145/1866480.1866511

[5] Bergamaschi, S., Guerra, F., Orsini, M., Sartori, C., Vincini, M., (2011), A Semantic Approach to ETL Technologies, Data & Knowledge Engineering, 70(8), 717-731.

[6] Bloomberg BusinessWeek Research Services, (2011), The Current State of Business Analytics: Where Do We Go From Here? A white paper produced in collaboration with SAS.

[7] Chaudhuri, S., Dayal, U., Narasayya, V., (2011) An overview of Business Intelligence Technology, Communications of the ACM, 54(8), 88-98. DOI= 10.1145/1978542.1978562 Retrieved June 29,2012 from http://doi.acm.org/10.1145/1978542.1978562

[8] Clikeman, P. M. (1999). Improving information quality. Internal Auditor, 56(3), 32-33.

[9] Conn, S., S., (2005), OLTP and OLAP Data Integration: a Review of Feasible Implementation Methods and Architectures for Real Time Data Analysis, Southeast Con, Proceedings IEEE , 515- 520. DOI = 10.1109/SECON.2005.1423297 Retrieved May 14,2012from URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1423297&isnumber=30732

[10] De Montcheuil, Y., (2005) Lesson – Third Generation ETL: Delivering the Best Performance, What Works: Best Practices In Data Warehousing and Business Intelligence, 20, p. 48.

[11] Evans, B., (2005), Improving the data warehouse with selected data quality techniques: Metadata management, data cleansing and information stewardship. Capstone Report, University of Oregon, Applied Information Management Program. Retrieved June 14, 2012 from URL: http://hdl.handle.net/1794/7814.

[12] ETL Data Extraction Methods - Part Two Retrieved June 14, 2012 from http://www.best-business-intelligence.com/2011/09/etl- continued-data-extraction-methods.html.

[13] Gerber, M., Von Solms, R., (2008) Information security requirements – Interpreting the Legal Aspects, Computers and Security, 27, 124 - 135.

[14] Huang, K., T., Lee, Y., W., Wang, R., Y., (1999), Quality Information and Knowledge, NJ: Prentice-Hall. Huang, D. , L., Luen, P., Rau, P., Salvendy, G., (2010), Perception of Information Security, Behavior & Information Technology, 29(3), 221 - 232.

[15] Henschen, D., (2010), Agile Business: 2010 BI and Information Management Survey, Information Week, Report ID: R1921110, Retrieved June 29, 2012 from http://analytics.informationweek.com

[16] Henschen, D., (2011), 2012 BI and Information Management Trends, Information Week, Report ID: R335111, Retrieved June 29, 2012 http://reports.informationweek.com
[17] Inmon, W. H., (1996), Building the Data Warehouse, 1st edition, Indiana: Wiley Publishing Inc.
[18] Jarke, M., Vassiliou, V., (1997) Data Warehouse Quality: A Review of the DWQ Project. Retrieved May 14, 2012 from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.36.4346
[19] Shaker, H., El-Sappagh, A., Abdeltawab, M., Hendawi, A., Hamed, A., Bastawissy, E., (2011), A Proposed Model for Data Warehouse ETL Processes, Journal of King Saud University - Computer and Information Sciences, 23(2), 91-104. Retrieved May 14, 2012 http://www.sciencedirect.com/science/article/pii/S131915781100019X.
[20] Schneider, D., A., (2007), Practical Considerations for Real-Time Business Intelligence, Berlin: Springer Simitsis, A., Skoutas, D., Castellanos, M., (2010) Representation of Conceptual ETL Designs in Natural Language Using Semantic Web Technology, Data & Knowledge Engineering, 69(1), 96-115. Singh, R., Singh, K., (2010), A descriptive classification of causes of data quality problems in data warehousing. IJCSI International Journal .