ETL and its impact on Business Intelligence

Nitin Anand

AIACT&R, New Delhi

Abstract- Business Intelligence (BI) is considered to have a high impact on businesses. Research activity has risen in the last years. An important part of BI systems is a well performing implementation of the Extract, Transform, and Load (ETL) process. In typical BI projects, implementing the ETL process can be the task with the greatest effort.

Index Terms- Business Intelligence, Data Warehouse, Decision Making, ETL, Operational Data, Metadata

I. INTRODUCTION

Business intelligence (BI) has gained wide recognition in the last years. It also got high business impact and is seen as a key enabler for increasing value and performance" [1].

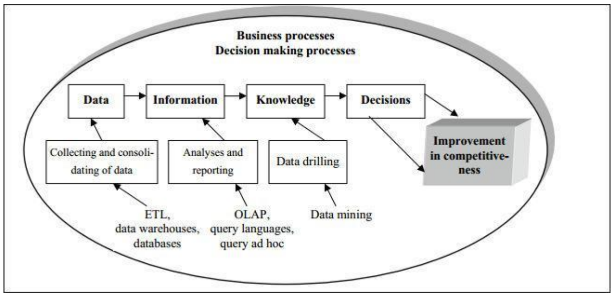
Business processes
Decision making processes
Data
Information
Knowledge
Decisions
Improvement in competitive-ness
Collecting and consoli-dating of data
Analyses and reporting
Data drilling
ETL, data warehouses, databases
OLAP, query languages, query ad hoc
Data mining

Figure 1: The role of BI systems in decision making [18]

Unsurprisingly, the progress of BI is monitored by management and IT consultants [2]. It is recognized as having a high relevance for the profit of businesses [3]. It is agreed that a strategic business intelligence approach will be needed [4].

At the same time, Business intelligence is a rather new discipline with a lot of research activity. Even though the term has been coined in 1958 [5], the number of published papers has risen considerably in the last few years. The rapid progress has also brought a high level of heterogeneity [6]; this causes both problems for businesses and offers research opportunities. It is possible to grasp the current state of BI [1] and practitioner's literature tries to lay out a roadmap on how to implement BI in a company [7]. There is no reliable roadmap for BI progress, though.

One important component of BI is the Extract, Transform and Load (ETL) process. It describes the gathering of data from various sources (extract), its modification to match a desired state (transformation) and its import into a database or data warehouse (load). ETL processes take up to 80% of the effort in BI projects [8]. A high performance is thereby vital to be able to process large amounts of data and to have a up-to-date database. The term ETL is known for a while [9] and the relevant market is already divided by a number of major players [10]. A data warehouse is predominantly used to store detailed summary data and metadata. Detailed data concern, for instance, sales or production volume in a given period. In order to increase effectiveness of queries, data in a data warehouse are subject to aggregation. Data e.g. on sales may be aggregated in a geographical dimension, a time period dimension or a product line dimension, etc. On the other hand, metadata include information on data themselves. They facilitate a process of extracting, transforming and loading data through presenting sources of data in the layout of data warehouses. Metadata are also used to automate summary data creation and queries management.

Furthermore, existing BI architectures typically feature a unidirectional communication flow between different components. The architectures proposed in [12] and [13] are good examples where they only feature a one-way data flow from data sources to data warehouse. The limitation of unidirectional data flow (i.e., no backward data flow from data warehouse to data sources) is that no adjustment or correction is allowed on data source even if an error is found. This may lead to the garbage-in-garbage out situation. If organizations want to correct the error, they have to repeat the entire BI process especially that of the cleansing procedures again. To overcome these problems, a two-way data integration flow [14] is suggested whereby the cleansed data can be sent back to data sources to improve accuracy and reduce cleansing work.

Another issue with existing BI architectures is the lack of support on metadata management. A good BI architecture should include the layer of metadata. A metadata repository is essential for business users to store and standardize metadata across different systems. By having a well-structured metadata, organizations will be able to track and monitor data flows within their BI environment [15]. In addition, they will be able to ensure the consistency of definitions and descriptions of data that support BI components and thus avoid misunderstanding and misinterpretation of data.

Aside from that, some of the architectures do not include operational data store (ODS) within the BI environment. For instance, Watson's BI architecture (2009) [16] contains only data warehouse and data marts whereas [17] include only data warehouse. In order to address operational data needs of an organization, it is essential to implement ODS to provide current or near current integrated information that can be accessed or updated directly by users. Through this way, decision makers will be able to react faster to changing business environment and requirements. Furthermore, it is necessary to consider data staging area in the ETL (Extract Transform-Load) process. As most of the data from data source require cleansing and transformation, it is important to create a temporary storage for data to reside prior to loading into ODS or data warehouse. Without building this staging area, the process of working on the data [27]

II. Business Intelligence Architecture

This paper describes a framework of a five layered BI architecture (see Figure 1), taking into consideration the value and quality of data as well as information flow in the system. The five layers are data source, ETL (Extract-Transform-Load), data warehouse, end user, and metadata layers. The rest of this section describes each of the layers.

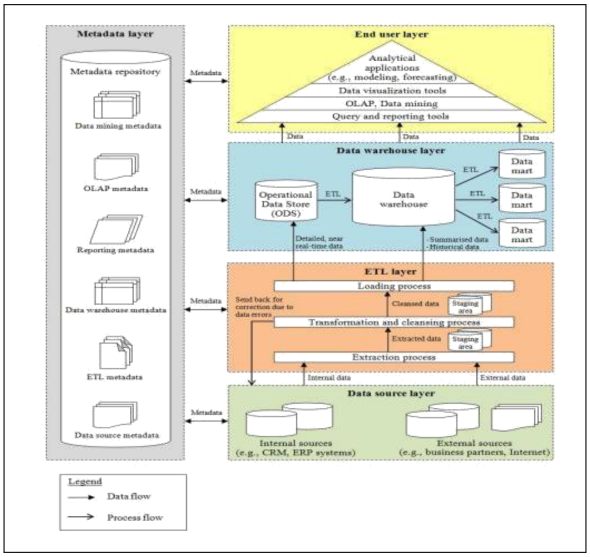

Fig 2: BI Architecture [12]

III. Data Source Layer

Nowadays, many application domains require the use of structured data as well as unstructured and semi-structured data to make effective and timely decision [12]. All these data can be acquired from two types of sources: internal and external.

Internal data source refers to data that is captured and maintained by operational systems inside an organization such as Customer Relationship Management and Enterprise Resource Planning systems. Internal data sources include the data related to business operations (i.e., customers, products, and sales data). These operational systems are also known as online transaction processing systems because they process large amount of transactions in real time and update data whenever it is needed. Operational systems contain only current data that is used to support daily business operations of an organization. Generally, operational 0mainly on specific business operations such as sales, accounting, and purchasing [19]

External data source refers to those that originate outside an organization. This type of data can be collected from external sources such as business partners, syndicate data suppliers, the Internet, governments, and market research organizations [20]. These data are often related to competitors, market, environment (e.g., customer demographic and economic), and technology [21] .

It is important for organizations to clearly identify their data sources. Knowing where the required data can be obtained is useful in addressing specific business questions and requirements, thereby resulting in significant time savings and greater speed of information delivery. Furthermore, the knowledge can also be used to facilitate data replication, data cleansing, and data extraction [26]

IV. ETL (Extract-Transform-Load) Layer

This layer focuses on three main processes: Extraction, Transformation and Loading [17]. Extraction is the process of identifying and collecting relevant data from different sources Usually, the data collected from internal and external sources are not integrated, incomplete, and may be duplicated. Therefore, the extraction process is needed to select data that are significant in supporting organizational decision making. The extracted data are then sent to a temporary storage area called the data staging area prior to the transformation and cleansing process. This is done to avoid the need of extracting data again should any problem occurs. After that, the data will go through the transformation and the cleansing process.

Transformation is the process of converting data using a set of business rules (such as aggregation functions) into consistent formats for reporting and analysis. Data transformation process also includes defining business logic for data mapping and standardizing data definitions in order to ensure consistency across an organization [

V. Data Warehouse Layer

There are three components in the data warehouse layer, namely operational data store, data warehouse, and data marts. Data flows from operational data store to data warehouse and subsequently to data

VI. Operational Data Store

An operational data store (ODS) is used to integrate all data from the ETL layer and load them into data warehouses.

ODS is a database that stores subject-oriented, detailed, and current data from multiple sources to support tactical decision making It provides an integrated view of near real-time data such as transactions and prices. In addition, the data stored in ODS is volatile, which means it can be over-written or updated with new data that Blow into ODS [22]. As such, ODS does not store any historical data. Generally, ODS is designed to support operational processing and reporting needs of a specific application by providing an integrated view of data across many different business applications [23]. It is normally used by middle management level for daily management and short-term decision making [24]. Since the data stored in ODS are updated frequently (i.e., in minutes or hours), it is useful for reporting [i]types that require real time (within 15 minutes) or near time (updated in 15 minutes to 1 hour) information [25]

VII. End User Layer

The end user layer consists of tools that display information in different formats to different users. These tools can be grouped hierarchically in a pyramid shape (as shown in Figure 1). As one moves from the bottom to the top of the pyramid, the degree of comprehensiveness at which data are being processed increases.

VIII. Conclusion

This paper has described a framework of five-layered BI architecture with various components. BI architecture plays an important role in affecting the success of a BI implementation. To have a smooth BI operation, organizations can benchmark their architectural plan against the framework proposed here. By having a good BI architecture, organizations will be able to maximize the value from their BI investments, and thereby meet their business requirements and improve business performance. However, at this point, the framework proposed in this paper remains conceptual in nature. Though it is built based on existing literature, the framework still needs to be validated using real-life BI cases to affirm its usability. Future research therefore can go along this line to validate the framework.

References

[1] H. J. Watson and B. H. Wixom. The current state of business intelligence. Computer, 40(9):96-99

[2] Gartner, Inc. Press release: Gartner reveals five business intelligence predictions for 2009 and beyond, January 2009

[3] S. Williams and N. Williams. The Profit Impact of Business Intelligence. Morgan iiKaufmann, San Francisco, CA, 2006. Online, 2007.

[4] Z. Panian. Business Intelligence in support of business strategy. In Proc.MCBE'06, pages 19-23, StevensPoint, 2006. WSEAS.

[5] H. P. Luhn. A business intelligence system. IBM J.Res. D ev., 2(4):314{319, 1958.

[6] Z. Panian. Expected progress in the field of business intelligence. In Proc. AIKED'09, pages 170-175,Stevens Point, 2009. WSEAS.

[7] L. T. Moss and S. Atre. Business Intelligence Roadmap: The Complete Project Lifecycle for Decision-Support Applications. Addison-Wesley Longman, Boston, MA, USA, 2003

[8] W. H. Inmon. Building the Data Warehouse. Wiley, New York, NY, USA, 3rd edition, 2002.

[9] R. Kimball, L. Reeves, W. Thornthwaite, M. Ross, and W. Thornwaite. The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing and Deploying Data Warehouses. Wiley, New York, NY, USA, 1998.

[10] Gartner, Inc. Press release: ETL magic quadrant update: A market in evolution, May 2002.Online:http://www.gartner.com/reprints/informatica/106602.html.

[11] Gartner Inc. Magic quadrant for data integration tools, September 2008.

[12] Baars, H. & Kemper, H.-G. (2008). “Management Support with Structured and Unstructured Data: An Integrated Business Intelligence Framework,” Information Systems Management, 25(2). 132-148.

[13] Shariat, M. & Hightower Jr, R. (2007). 'Conceptualizing Business Intelligence Architecture,' Marketing Management Journal, 17(2). 40-46.

[14] Dayal, U., Castellanos, M., Simitsis, A. & Wilkinson, K. (2009). “Data Integration Flows for Business Intelligence,” Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology, 1-11.

[15] Pant, P. (2009). “Essential Components of a Successful BI Strategy,” [Online], [Retrieved January 21, 2011], http://www.informationmanagement.com/specia lreports/2009_155/business_intelligence_bi-10015846-1.html?pg=1

[16] Watson, H. J. (2009). “Tutorial: Business Intelligence – Past, Present, and Future,” Communications of the Association for Information Systems, 25, 487-510.

[17] Sen, A. & Sinha, A. P. (2005). “A Comparison of Data Warehousing Methodologies,” Communications of the ACM, 48(3). 79-84.

[18] Olszak, C. M.,& Ziemba, E. (2004). Business intelligence systems as a new generation of decision support systems. Proceedings PISTA 2004, International Conference on Politics and Information Systems: Technologies and Applications. Orlando: The International Institute of Informatics and Systemics.

[19] Hoffer, J. A., Prescott, M. B.& McFadden, F.R. (2007). Modern Database Management, Pearson/Prentice Hall, Upper Saddle River, New Jersey.

[20] Ranjan, J. (2009). “Business Intelligence: Concepts, Components, Techniques and Benefits”, Journal of Theoretical and Applied Information Technology, 9(1). 60-70.

[21] Haag, S., Cummings, M. & Philips, A. (2007). Management Information Systems for the Information Age, McGraw-Hill, Boston.

[22] Imhoff, C., Galemmo, N. & Geiger, J. G. (2003). Mastering Data Warehouse Design: Relational and Dimensional Techniques, John Wiley & Sons, Indianapolis, Indiana.

[23] Chan, J. O. (2005). “Optimizing Data Warehousing Strategies,” Communications of the IIMA, 5(1). 1-13.

[24] Li, Z., Huang, Y. & Wan, S. (2007). “Model Analysis of Data Integration of Enterprises and E-commerce Based on ODS,”International Conference on Research and Practical Issues of Enterprise Information Systems (CONFENIS 2007). 1, 275-282.

[25] Walker, D. M. (2006). “Overview Architecture for Enterprise Data Warehouses,” [Online], [Retrieved December 1, 2010], http://www.datamgmt.com/index.php?module=d ocuments&JAS_DocumentManager_op=downlo adFile&JAS_File_id=29

[26] Reinschmidt, J. & Francoise, L. (2000). Business Intelligence Certification Guide, IBM, San Jose, California

[27] Davenport, T. H. & Harris, J. G. (2007). “The Architecture of Business Intelligence,” [Online], [Retrieved December 5,2010], http://www.accenture.com/NR/rdonlyres/15DCF F6A-4DE0-44D8-B778-630BE3A677A2/0/ArchBIAIMS.pdf.

Authors

First Author – Nitin Anand, AIACT&R, New Delhi, proudtobeanindiannitin@gmail.com