Data Quality Management For Data Warehouse Systems: State Of The Art

Hamid Naceur BENKHALED[1] and Djamel BERRABAH[2]

EEDIS Laboratory, Djilali Liabes University, Sidi Bel Abbes, ALGERIA
benkhalednaceur@gmail.com[1]
djberrabah@gmail.com[2]

Abstract. During the last years, Data Warehouse (DW) systems have been considered as the most effective tool for decision support making. Most of the enterprises are obliged to implement their own Data Warehouse systems in order to use their collected data, make decisive decisions out of it and have a place in the market. However, most of the DW projects are interrupted due to poor Data Quality (DQ) problems like missing values, duplicate values and referential integrity issues. DQ problems can decrease customer satisfaction and increase the cost of the data warehouse projects. At the same time, the arriving of Big Data puts new requirements on the traditional DW systems and specifically on the ETL (Extract, Transform, Load) process, which is responsible for data collecting, cleansing and loading. These requirements can be summarized into the real time analyzing and the need of collecting the most recent data. This paper will include two important points: (1) a survey of the existing approaches in the literature for managing data quality in the traditional data warehouse systems, (2) a survey about the existing approaches for adapting traditional DW systems to the new requirements of Big Data.

Keywords: Data Warehouse · ETL · Data Quality · Big Data.

1 Introduction

Organizations all around the world are implementing their own Data warehouse (DW) systems in order to use their collected data and extract useful information from it to make a decisive decision. In the last years, DW systems have proven their efficiency by giving the enterprises a step ahead in the market competition. William H. Inmon who is considered as the father of Data Warehouse defines a data warehouse as "a collection of Integrated, Subject-Oriented, Non Volatile and Time Variant data in support of managements decisions" [10].

Despite all these advantages, DW systems can sometimes fail to meet the stakeholders expectations. Many DW projects have been interupted due to Data Quality (DQ) problems and according to the Data Warehousing Institute the estimated annual losses in USA are around 600 billions dollars because of poor DQ, the same study shows that 15% to 20% of the stored data in most of the

organization is erroneous or unusable [6]. As a result, the stakeholder can lose its trust in the efficiency of the DW and that can cause customers dissatisfaction and increase the cost of the DW projects. Knowing now the importance of data quality inside the DW, proposing a data quality management system is very important to keep the users trust in the DW system and to make correct decisions. Many approaches were proposed in the literature, some of these approaches focus on integrating a data quality management system into the DW life cycle like [9]. A quality metadata model for managing data quality was proposed also in [13] and a Data warehouse development life cycle to manage data quality was proposed in [16] and others. Some of these approaches will be discussed in details in this paper as well as a comparative study.

In the other hand, we are witnessing the arriving of the Big Data era, which puts new requirements on the traditional DW systems and specifically on the ETL process. The ETL process is considered as a time consuming process but the old DW systems were not sensitive to the latency presented by this workflow [4]. For example, one of the Big Data applications IoT (Internet of Things) need to execute near-real time analyzing and use the most recent collected data and that was not the case of the traditional DW systems [15]. So adapting DW systems to the new Big Data requirements is very challenging.

In order to adapt DW systems to the new Big Data requirements, a number of approaches were proposed in the literature, some of these approaches focus on proposing an architecture that integrates the two technologies (DW, BG) like in [21]. Others focus on adapting the ETL architecture to the new streaming requirements [15]. We can also find approaches which propose an ontology based data quality framework in order to manage data quality for the streaming application which is the case of Big Data [7]. Moreover, a semantic ETL was proposed in order to integrate perfectly heterogeneous sources [1].

The rest of this paper is organized as follows: In section 2 a background about ETL, Big data and Data Quality is included. In section 3 a detailed description of the existing approaches for managing Data Quality in the DW systems, Section 4 is about the existing approaches for adapting DW systems to the new requirements of Big Data and Section 5 concludes the paper with the indication of possible future researches.

2 Background

2.1 ETL

The collection of data from multiple sources in different formats, the cleaning and the transformation of the collected data in order to be loaded correctly in the data warehouse is known as the ETL process (Extraction, Transformation, Loading) [12] [24] , This process is considered as the most important process in the data warehouse life cycle, ETL represents 70% of the efforts in the data warehouse projects [14]. ETL usually deals with a huge amount of data and that is what makes it an extremely time consuming process [4], it is implemented as

a workflow where data processors are connected by data flows [18]. The most important phase in the ETL process is the transformation phase, also called data staging area (DSA) [25]. Most of the data cleaning tasks are performed at this stage in order to improve DQ, but generally ETL tools do not include advanced cleaning capabilities [24], as a result, poor DQ problems can appear causing serious issues in decision making by giving wrong conclusions.

2.2 Data Quality

Data quality management is defined in [6] as the process that includes the definition of policies and the attribution of roles in order to collect, maintain and diffuse data. The process can't be accomplished without a partnership between the business and the technology groups.

Data quality dimensions Data quality dimensions are used to assess and to measure the value of DQ, the major Data Quality dimensions were summarized in [2], where accuracy, completeness, currency and consistency are considered as the principal DQ dimensions in addition to other secondary dimensions like accessibility and interpretability. For each dimension one or two metrics are provided. Two types of accuracy are cited, syntactic accuracy and semantic accuracy. Syntactic accuracy focuses on whether a value V is one of the values in the attribute definition domain or not. Several functions exist to measure accuracy like Edit distance, similar sounds and character transposition. The second type of accuracy is semantic accuracy, which is more complex to measure comparing to the first one because it is defined as how close a value V it is to the real world value V. In the relational world completeness describes how much a table extents and covers the associated real world; completeness is described by the presence of null values in the tuples. Four types of completeness are defined: value completeness, tuple completeness, attributes completeness and relation completeness. Three time-related dimensions were defined in the book. Currency is defined as how quickly the stored data is updated, it can be measured using the meta-data of last update. Volatility depends on the type of data, it is considered high if the data changes frequently and low if the data is stable like date of birth. Timeliness describes whether the current data is useful for the current task or not. The consistency is a dimension to cover the violations of the defined semantic rules in the database or files. Mostly, these semantic rules are expressed using integrity constraints and data edits.

2.3 Big Data

Many organizations tried to give a definition to the Big Data term like the definition of Oracle in [5] and the definitions of Microsoft and Intel in [20] [11], but the most accepted and used definition by the Big Data community is given by the Gartner Group in 2001, which define Big data using 4 Vs (Volume, Variety,Velocity and Veracity) where (1) the term volume is used to refer to a huge

amount of data collected from different sources (mobiles, social media, sensors . . . Etc.) [3], (2)Variety because the type of the collected data can be structured like traditional relational data bases or can be semi-structured (XML files) or unstructured like text files, (3) Velocity is defined as the speed, which the data arrives with to an enterprises and how much time it took to be analyzed and well understood and finally (4) Veracity that represent data suitability and credibility for the target audience.

Analyzing Big Data using technologies like Hadoop gives us a great possibility for extracting useful and hidden information and use it to take good decisions. But with big data comes big errors, all the research that were based on erroneous data give bad results in term of authenticity and accuracy, so underestimating DQ can drive us to bad conclusions.

As mentioned in the introduction section the arriving of Big Data puts new requirement on the traditional DW systems and specifically on the ETL process, which is responsible of data extraction from multiple source, data transformation and loading into the DW, but with Big Data the ETL process can take too much time and that what can be an obstruction of the real time analyzing process which is the main goal of the Big Data analyzes. A number of solutions were proposed in the literature. They are discussed in section 4.

3 Data quality in Data warehouse Systems

Many organizations around the world are implementing Data warehouses in order to explore their collected data and analyze it to get the right decisions. However, many data warehouses project have been cancelled due to Data Quality problems [6] . So proposing a DQ management system for data warehouses can increase the effectiveness of the DW and increase the customers satisfaction, a number of approaches were proposed in the literature. In this section, some of these approaches will be discussed.

A meta-data based Data Quality system for managing Data Quality in data warehouses was proposed in [8], the authors started the paper by mentioning the important of total quality management (TQM) inside a typical enterprise which focus on the customer demands and quality problems for all the stake-holders in the data warehouse system. Using a proactive DQ management can ensure regular quality improvement and that's by (1) quality planning, which allows building quality specifications and (2) quality control by assuring that the delivered data conforms to the fixed specifications. Two DQ factors were studied in this paper, quality of design and quality of conformance, Quality of design allows the transformation of quality requirements into specifications and the goal of quality of conformance is to make sure that the processed data in the warehouse is compliant with the user requirements. A meta data management component is integrated into the the data warehouse life cycle which contains all the major information c[8] concerning DQ this component is composed of : (1) Rule Base which contain all the needed rules to measure Data Quality in addition to the time schedules of executions. (2) Notification Rules: the role of

this component is to decide who should be informed in case of quality rule violation. (3) Quality statement : responsible for delivering the quality results to the end-users. The paper also includes some metrics for measuring Data Quality dimensions like plausibility, timeliness and usefulness using data mining techniques and descriptive statistics to extract data characteristics, which can be used to define constrains for DQ measurements. The proposed architecture was implemented in a Swiss bank database and all the quality rules used in the system were defined using SQL statements. The feed-backs from the end-users show that the Data Quality controlled by the metadata based quality system is acceptable.

J.Chankaranarayanan proposes in [22] a new framework for the management of Data Quality in decisional environments and specifically in data warehouses, the author mentioned that most of the existence approaches concerning the quality of data in the warehouses focuses on fixing quality goals than translate them to analysis queries. But, it is important that decision makers should be able to gauge DQ in the desired contest. As a result, the proposed framework allows the communication of the quality information and give the ability to the decision maker to gauge Data Quality not only at the final stage but also in all the stages of the processing, in this article accuracy is chosen by the author as a quality dimension to show how the framework can integrate DQ and how it can be measured. The proposed framework is based on the Information Product Map (IPMAP) and IP approach, which allows managing information as a product and tracing a quality problem to its sources and identifying all the impacted stages. The paper also provides the necessary meta-data requirement for the management of the Data Quality in a DW. For the sake of improving DQ, the meta-data for each IPMAP construct is enriched with meta-data that includes: identifier of the stage, responsible of the stage and 6 other information. The use of IPMAP allows the implementation of a total DQ management by offering 3 majors potentials, the first one is estimating of the delivery time using techniques like PERT or Critical Path Method, IPMAP also provides reachability which can help in identifying all the infected stages with quality problem once detecting one stage. Tractability is also possible with IPMAP; using the meta-data associated with each stage we can identify the responsible department of the Data Quality problems.

In order to manage perfectly DQ in data warehouses, a simplified approach for quality management was proposed [13], the authors mentioned that to guarantee Data Quality in a global way, the development team has to understand DQ problems for all the entities involved in the data warehouse system from the decision makers to the executive manager, each entity has its own point of view for DQ. The proposed framework is composed of multiple steps, where the first step is to define a Quality Council, which is responsible for the identification and the evaluation of the quality parameters; in addition, the Council is also responsible for the formulation of quality policies and a quality system. The next step is to define quality parameters, for each parameter a measured agent must be fixed; a set of DQ parameters and its corresponding metrics were mentioned

in this paper. The authors also said that for each Data Quality parameter, an acceptable value should be also fixed in order to compare it to the calculated value, if the calculated value is in the range of the acceptable value than the quality of data in the warehouse is acceptable. In the other case, if the calculated DQ value is not in the acceptable range than the quality of data has to be improved using error detection and correction techniques. However it's better to prevent these errors from the beginning by building data processes from scratch and re-designing the existence ones by introducing error controls and quality control using meta data. A quality meta-data model is also proposed in this paper where each stakeholder have its own quality goal imposed on DW object and achieved by quality query which is evaluated using quality metrics.

The authors in [17] proposed an meta-data quality architecture for managing DQ in the DW systems, their architecture is based on quality planning where the users have to specify their quality requirements, than these quality requirements will be introduced to the meta-data of the the warehouse as quality statement. The proposed architecture allows controlling Data Quality during all the phases of the data warehouse processes. A framework for managing Data Quality in Data Warehousing was proposed in [16]. Knowing that in the most of cases, DQ problems don't appear until during the data warehouse project. So as a result, the proposed framework was based on a data warehouse development life cycle (DWDLC) where all the phases of the data warehouse project are included from the planning to the implementation and maintenance. Seven data quality dimension were included in the proposed DWDLC (Accuracy, Completeness, Timeliness, Integrity, Consistency, Conformity and record duplication), each one or two dimensions are associated to a layer. The proposed DWDLC is composed from 7 layers where the most important layers are the Analysis and Development layers. Data Accuracy and completeness were associated to the analysis layer since the data profiling should be done at this phase. In the development layer consistency and conformity dimensions should be verified.

Other works: Beside the discussed approaches above, other papers discussed the data quality problems in DW systems, for example the authors in [23] proposed a descriptive taxonomy of all the stages where data warehousing is affected with data quality problems (data sources, data profiling, ETL phase, issues related to the schema design). The authors in [19] provided an overview about the problems of data cleaning and their solutions and they presented a classification of these problems based on if it's a single or a multiple source problem.

Discussion : The approach proposed in [8] was implemented in a Swiss bank and the users were satisfied from the delivered data quality. However, The authors used only SQL statements to define quality rules and they didn't use users defined functions, this approach does not cover all data quality dimensions and it's not metioned if there is a possibility of extension. In [13] the authors proposed a framework for managing data quality in data warehouse but the paper does not include how to improve the data quality in the case where the measured value is not acceptable. In [16] the authors proposed a Data warehouse

Development Life Cycle associated with quality dimensions but no data quality metrics were mentioned in the paper. In [22] the author based the approach on the IPMAP and what helped covering only Three data quality dimensions. The following Table shows how much each proposed approach cover the quality dimensions discussed in the background section.

Table 1. Data Quality dimension covering

References	**Accuracy**	**Completness**	**Consistecy**	**TimeRelated-Dimensions**
[8]	YES	NO	YES	YES
[22]	YES	YES	YES	YES
[13]	YES	YES	YES	NO
[17]	YES	NO	YES	NO
[16]	YES	YES	YES	YES

4 Adapting DW systems to the new Big Data requirements

This section is dedicated to the proposed approaches in the literature for adapting the traditional DW systems to new requirements of Big Data. For example the authors in [21] proposed a new architecture for integrating the two technologies while the authors in [15] proposed a new ETL architecture for data streaming applications which is the case of Big Data.

A comparison between Big Data and data warehouse has been made in [21], The authors of this article thinks that big data still a young field under development while the large utilization of data warehouses in organizations and research fields make it a mature technology. Multilayer architecture also has been proposed in the paper in order to integrate the two technologies. The results of the research summarized the major differences between Big Data and data warehouse technologies, where the principal data sources used in data warehouse are usually transnational databases while big data use generally social networks, sensors, emails and more as sources. Another important difference is the scope of use, Data warehouses are generally used in decision support and OLAP (Online Analytical Processing) while Big Data is usually used in discovering knowledge from huge amount of data. The principal actors in the data warehouse are business analysts without any knowledge of data technologies while in Big Data the users are generally data scientist and analysts. The proposed architecture is composed of three principal layers: Data upload, Data processing and storage, data analysis. The data upload layer is for storing data according to its type where structured data is directed for pre-processing and the unstructured data is stored as raw data. In the processing layer the structured data is aggregated and stored in the aggregate data area where OLAP can be done. The unstructured data stored as raw data can be loaded into a contextualized data area after

applying some filleting techniques on it. The filtered data can be also loaded to the related data area after the process of patterns finding. Finally, the data analysis layer is where OLAP analysis and business Intelligence are done in order to support decision-making. Using Traditional ETL systems in Big Data analytics is a problem to execute real time analyzing and to make fast decisions. The authors of [15] saw that the best way to solve this problem is to create a new ETL architecture based on stream processing systems. They divided the requirement for a streaming ETL system into three majors categories: ETL requirements, Streaming requirements and infrastructure requirements.

Four components architecture was proposed in the paper. The first one is a Data collector, the principal tasks of this component is to make sure that all the tuples are routed to the right destination while keeping receiving new tuples at the same time. The data collector must be also scalable in order to serve more clients in the case of augmentation in the number of data sources. The authors chose to use Apache Kafka as a data collector. The second principal component in the proposed architecture is a streaming ETL engine that receives data as batches from the data collector, all the transformations and data cleaning operations are done inside the streaming ETL engine which is equipped with full ETL traditional tools, the cleaned data is stored in the ETL engine in order to be transferred later to the warehouse. S-Store is chosen as a streaming ETL engine. The next component is composed from two principal parts: one or several OLAP Engines and a query processor. The OLAP engine must contain a data warehouse with a delta data warehouse that allows faster queries. The streaming ETL engine send its data to the delta data warehouse via a data migrator, and the OLAP engine takes care of merging the new data with the full data warehouse (periodically). In the other hand the query processor must allow the user to execute queries on the staging are of the ETL engine. Postgres was chose as a back end database in their experimentation. The last principal component in their proposed architecture was a data migrator that allows transferring data between the streaming ETL engine and the OLAP Backend without losing any information. In order to test their new architecture, the authors experiments two types of configurations. The first one is based on push technique, which means that the streaming ETL engine pushes the newly cleaned data to the warehouse and the second one is based on pull technique, which means that the warehouse pulls the new processed data from the streaming ETL engine at the start of an analytical query. The experimentation results showed that pulling new data from the ETL engine is the best choice regarding staleness; the results also showed that if the priority is the query execution time than the best technique is to push data from the streaming ETL engine to the warehouse.

An anthology-based framework for managing data quality in different dimensions was proposed in the field of data streams applications in [7], the proposed architecture is composed of three main services: (1) query based quality service which serve for analyzing the query and identifying to operators that can have an impact on the data quality value, (2) Content based quality service, the role of this service is to compute data quality value depending on the existing data

in the stream and the evaluation of the defined semantic rules in the ontology , finally (3) application based quality service which allow the user to add data quality values to the streamed data directly from the user defined functions. It's also mentioned in this paper that most of the existing approaches focus on a limited number of data quality dimensions, so the proposed architecture has to be extensible and should be also optional to turn it on/off in case of memory overhead. In order to link that data stream elements (Window, Attribute) with the data quality dimensions and metrics in a suitable way, an Ontology was proposed. The authors used in their experimentation two categories of DQ dimensions; Application based DQ dimensions and system based DQ dimension, some of these dimensions (Completeness, Data Volume, Timeliness, Accuracy, Consistency and confidence) can belong to one or both categories. The proposed DQ ontology use DQ factors to link DQ dimensions and metrics to the data stream element (Window and attribute). The system performance experimentation showed that using a DQ framework in a DSMS required more CPU power just in the initialization phase where the DQ ontology have to be load, after the initialization phase the CPU power and the used memory is the same in both cases (with and without a DQ framework).

From the discussed approaches in sections 4 and 3, we can see that metrics used to assess data quality in the traditional DW systems need to be improved in order to guarantee a good data quality in the case of Big Data. Specifically, concerning the need of real time analyzing which is a big impediment for the traditional metrics. As a result of that, using some Big Data techniques like MapReduce in evaluating the data quality dimensions can be a possible solution.

5 Conclusions

This paper provides a survey of Data Quality management in the data warehouse systems, we have discussed the huge impact of poor DQ problems on the efficiency of the DW systems and we saw some of the proposed approaches for managing DQ. The paper also includes the problem of adapting the traditional DW to the new requirements of Big Data, which is considered very challenging due to the latency of the ETL process. As future works we are aiming to improve DQ management in the data warehouse systems by exploring the Semantic Web technologies and Linked Data.

References

1. S. K. Bansal and S. Kagemann. Integrating big data: A semantic extract-transform-load framework. *Computer*, 48(3):42–50, 2015.
2. C. Batini and M. Scannapieco. *Data and information quality: dimensions, principles and techniques.* Springer, 2016.
3. G. Bello-Orgaz, J. J. Jung, and D. Camacho. Social big data: Recent achievements and new challenges. *Information Fusion*, 28:45–59, 2016.

4. N. Berkani, L. Bellatreche, and S. Khouri. Towards a conceptualization of etl and physical storage of semantic data warehouses as a service. *Cluster computing*, 16(4):915–931, 2013.
5. J. P. Dijcks. Oracle: Big data for the enterprise. *Oracle white paper*, page 16, 2012.
6. J. G. Geiger. Data quality management, the most critical initiative you can implement. *Data Warehousing, Management and Quality, Paper*, pages 098–29, 2004.
7. S. Geisler, S. Weber, and C. Quix. An ontology-based data quality framework for data stream applications. In *16th International Conference on Information Quality*, pages 145–159, 2011.
8. M. Helfert and C. Herrmann. Proactive data quality management for data warehouse systems. In *DMDW*, volume 2002, pages 97–106, 2002.
9. M. Helfert, G. Zellner, and C. Sousa. Data quality problems and proactive data quality management in data-warehouse-systems. *Proceedings of BITWorld*, 2002.
10. W. Inmon. Building the data warehouse, qed technical pub. *Group*, 1992.
11. Intel. Intel peer research on big data analysis.
12. P. T. T. C. Jensen, C.S. Synthesis lectures on data management. *San Rafael*, 2010.
13. V. Kumar and R. Thareja. A simplified approach for quality management in data warehouse. *arXiv preprint arXiv:1310.2066*, 2013.
14. X. Liu, C. Thomsen, and T. B. Pedersen. Mapreduce-based dimensional etl made easy. *Proceedings of the VLDB Endowment*, 5(12):1882–1885, 2012.
15. J. Meehan, C. Aslantas, S. Zdonik, N. Tatbul, and J. Du. Data ingestion for the connected world. In *CIDR*, 2017.
16. R. R. Nemani and R. Konda. A framework for data quality in data warehousing. In *International United Information Systems Conference*, pages 292–297. Springer, 2009.
17. R. B. Palepu and D. Rao. Meta data quality control architecture in data warehousing. *International Journal of Computer Science, Engineering and Information Technology*, pages 15–24, 2012.
18. P. Patil, S. Rao, and S. B. Patil. Data integration problem of structural and semantic heterogeneity: data warehousing framework models for the optimization of the etl processes. In *Proceedings of the International Conference & Workshop on Emerging Trends in Technology*, pages 500–504. ACM, 2011.
19. E. Rahm and H. H. Do. Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, 23(4):3–13, 2000.
20. W. Redmond. The big bang: How the big data explosion is chan-ging the world, 2012.
21. S. O. Salinas and A. C. N. Lemus. Data warehouse and big data integration. *Int. Journal of Comp. Sci. and Inf. Tech*, 9(2):1–17, 2017.
22. G. Shankaranarayanan. Towards implementing total data quality management in a data warehouse. *Journal of Information Technology Management*, 16(1):21–30, 2005.
23. R. Singh, K. Singh, et al. A descriptive classification of causes of data quality problems in data warehousing. *International Journal of Computer Science Issues*, 7(3):41–50, 2010.
24. J. Trujillo and S. Luján-Mora. A uml based approach for modeling etl processes in data warehouses. In *International Conference on Conceptual Modeling*, pages 307–320. Springer, 2003.
25. P. Vassiliadis, A. Simitsis, and S. Skiadopoulos. Conceptual modeling for etl processes. In *Proceedings of the 5th ACM international workshop on Data Warehousing and OLAP*, pages 14–21. ACM, 2002.