Data warehouse and data quality – an overview

Helena Brajkovic, Danijela Jaksic, Patrizia Poscic
University of Rijeka
Department of Informatics
Radmile Matejcic 2,
51000 Rijeka, Croatia
helena.brajkovic2@gmail.com, {danijela.jaksic, patrizia}@inf.uniri.hr

Abstract. *Almost every company today is collecting and storing data for later analysis and business decision making in some kind of repository, usually a data warehouse. In order for decision to be timely and accurate (content-wise), it should be based on high quality data. Data quality policies and standards can be applied to every data warehouse. However, every data warehouse is unique and there is a need to consider a lot of criteria to achieve the best outcome. The goal of this paper is mostly to give an overview of data quality issues in a data warehouse, explain different criteria and categories for data quality and show some causes and consequences of saving low quality data in the data warehouse. The main contribution of this paper is, consequently, an overview of data quality issues in the context of data warehouses and an insight into some new and emerging research areas in the field of data warehouse quality.*

Keywords. data warehouse, data quality, database, ETL, data model

1 Introduction

Amount of data that business organizations and their business processes generate nowadays is bigger than ever. Additionally, every person on this planet generates some data on a daily basis. A need to store (and later process) all that data in an efficient and useful way becomes a necessity.

A data warehouse (DW) is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision-making process (Inmon, 2002) and it is one of the approaches to store this integrated and processed data in one place, for a later business analysis.

However, this data needs to be of a high quality in order to serve as a good basis for business decision making. This is the motivation behind this paper – the authors wanted to get an overview of data quality (DQ) approaches and issues, in the context of data warehouses. It is also the main focus of this paper.

Low-quality data represents incorrect, untimely or inconsistent data – however, data that is correct, complete, readable and available can also be a candidate for a low-quality data (Vuk et al., 2015). Bad data is the one with inadequate semantics; the one that can't be interpreted correctly. That data can still be processed and used but the user will receive limited or wrong information. With that in mind, data quality (its policies and standards) must be defined and maintained in every data warehouse.

The main research question of this paper is “What are the most common data quality issues in a data warehouse?”. Related to the question, the goal of this paper is to give an overview of data quality issues in a data warehouse and of some aspects which demand additional attention (like categories which might contain data of poor quality). The main contributions of this paper are, thusly, an overview of data quality issues in the context of data warehouses and additionally, an insight into some new and emerging research areas in the field of DW data quality.

The paper is organized as follows: in Section 2 the data quality problem is presented, in Section 3 criteria for defining data quality is explained, in Section 4 most common problems that every data warehouse faces are described, in Section 5 most common requirements for data quality are mentioned, in Section 6 some phases in building a data warehouse where the mistakes can happen and thus produce low quality data are mentioned, in Section 7 different aspects and user roles in data quality are shown, in Section 8 some consequences of bad quality data are discussed, in Section 9 some recent research approaches in the field are shown and finally in Section 10 are some conclusions.

2 The data quality problem

Before the appearance of a data warehouse (DW) as an integrated system for business analysis, data was stored in databases (DB). Even now, databases are used on a daily basis for storing all kinds of data. The goal of a data warehouse is to integrate all those databases that store daily operational data and to serve as a comprehensive, historical and central repository for further analysis. Because of that, most of the data quality approaches in the context of a DW can be related to previously researched data quality problem in the context of a DB (broadened with some DW-specific issues).

The most popular and widely used database underlying model was relational model – it is so even nowadays, for operational everyday tasks. In a relational model tables are mutually connected with relations and designed to store different types of data. One of the biggest problems in databases is redundancy, because of which the same value is being stored into multiple rows (Chen et al., 2011). To avoid data redundancy tables are normalized, that is, specific rules are followed to store data from one table in multiple (related) tables (Chen et al., 2011). Multiple categories of the normal forms can be distinguished, each of them representing one step forward to better data structure in order to remove redundancy in a database. If, for any reason, relationships between the tables need to be changed and re-defined in a database, it needs to be understood how that change will affect data model and if new tables need to be created and normalized.

Changing a data model is putting at risk data integrity and increases the risk of creating data redundancy in the tables (Chen et al., 2011). Changing the database model after it's already in use is not easy and it requires a lot of planning and testing before the change is implemented. Data warehouse is much more prone to changes in its environment – changes in the data model of its data sources, changes in the users' requirements, technology changes and system upgrades – it all affects the DW. Redundancy is still one of the biggest data issues but more effective methods exist to find the errors in data and their manipulation after they are already saved and used in a data warehouse.

Three approaches to data quality in a data warehouse can be seen in literature: intuitive, theoretical and empirical (Wang et al., 1996).

Intuitive approach is based on a previous experience and intuition of the researcher. It is based on accuracy and reliability of the data in a DW. With this method it's easier to make decisions about which criteria is the most important for data quality and quality of the data system.

Theoretical approach is based on researching the process which, gradually, turns good data into bad.

Empirical approach is more concentrated on data as a product that transforms, loads and is being processed.

However, these three approaches could be seen through a much wider view of the complete system for which data quality is being measured. In this context, four categories of data quality can be distinguished (Wixom et al., 2001):

1. Internal quality,
2. Quality of availability,
3. Contextual quality,
4. Representational data quality.

Internal quality is observing accuracy, objectivity, credibility and privacy of the data source.

Quality of availability is making sure of data availability. It is being implemented by introducing security measures into online services and applications where the data is being originally inserted.

Contextual quality is referring to events that data consumers are performing. Data must match context the consumer is currently in, and data must bring the consumer information he needs. It includes following criteria: data relevancy, added value, timeliness, complexity and data quantity.

Representational data quality means that consumers must easily understand the data in front of them. Data must be consistent, interpretative and presented in an intuitive way.

Table 1 contains an overview of these categories and key criteria related to each category.

Table 1. Data quality categories

Internal quality	**Quality of availability**	**Contextual quality**	**Representational data quality**
Accuracy	Availability	Relevancy	Consistency
Objectivity	Security measures	Timeliness	Interpretability
Credibility		Complexity	Intuitive representation
Privacy		Data quantity	

There can be four additional categories, when observing data from information system point of view with the goal to own data whose quality is controllable, that is, making sure that with bigger volume of data in a DW its quality doesn't decrease. These categories are: timeliness, data quality, expense and worth.

Each data has its lifecycle, it becomes inaccurate and inconsistent with time. It is important to serve the right information at the right time to the user.

Data quality is a result of invested human resources, money and time. If the person who is managing the data knows the process and its advantages and disadvantages, extracting necessary information will be cost-effective and faster. Expenses can be referred to consequences of supporting low quality data over time – re-entry and re-control of data, making wrong business decisions and how those decisions affect company business and expenses for improving data quality. That contains cost of investing in employee experience, applying standards of data quality, cost of analysis and repair costs.

The user is determining value that every data is bringing to him. Since the data has previously been through a series of transformations, its value should be as high as possible.

Table 2. Data quality criteria

Data Completeness	**Consistency**	**Value**	**Conformity**	**Accuracy**	**Integrity**
Available at any time	Data format from different sources	Data accuracy	Consistent data keeping	Real world data	Table connections
Data usability	Saving the same fact throughout whole data warehouse	Data reasonableness	Ease of reading information	Actual events	Loosing memory space
Incomplete data					Loosing speed of work

3 Measuring data quality in a data warehouse

Data quality is being measured according to three criteria (Calero et al., 2010): DBMS quality, data quality and data model quality.

DBMS (Database Management System) quality – to achieve the higher quality of DBMS, it is best to use international standards.

Data quality consists of three parts, quality of data definition (does the data represent real business image), quality of data values (values that are saved to the tables must represent real world, they have to be accurate and must contribute to business requirements based on which they were saved to the data warehouse) and quality of data presentation (format in which the data is being saved must be clear and intuitive to the user).

Data model quality – DW designer must choose tables, processes and data partitions so they represent logical structure of a data warehouse and that they serve its purpose.

4 Bad data indicators

Many business systems have multiple data sources. To coordinate all that data and save it to a unified schema, ETL[1] processes are being used to transform and insert data into a data warehouse. This step is critical since it is necessary to detect in advance all the possible irregularities in the source and find a way to diminish or remove these irregularities. Problems that are most often encountered are:

- Inability to compensate source records that are incomplete,
- NULL values,
- Process of manual inserting data is not being controlled on the user's side,
- Rules of data structure are not being followed,
- Errors when transferring data from one system to another (exporting data from one database to another even in the same format leads to errors in data values),
- Additional integration of the data from external systems that have different data structure.

5 Data quality requirements

Every data warehouse is specific and contains different issues that can be improved, that is how it can be contributed to data quality. To improve data quality, the problems must be previously known in order to handle them properly. Possible consequences that bad data will make to the business processes must also be known in advance - so that it could be measured if the problem is resolved or not.

Six different criteria can be distinguished according to which it can be decided if the data in a DW is of high quality: completeness, consistency, value, conformity, accuracy and integrity (Singh et al., 2010). An overview of the criteria is shown in Table 2, together with key factors for every criterion.

Data completeness: is referred to data availability at any time, if the data is usable, if any data is incomplete or missing. Example: let's take customer data and products that every customer bought. It contains description of every bought product and its price but is missing quantity of products bought for some customers. That data is not complete and will be discarded for every customer that is missing quantity.

Consistency: ideally, data from every source should be in the same format and if that is not possible then the data must be transformed in order to save it to a unified data warehouse and the same data item should be stored the same way, no matter the source format. Example: tracking if a user is currently an active customer - it can be marked with numbers 0 and 1, or with character string 'Active' or 'Inactive'. It is

[1] ETL (Extract, Transformation, Load) is everything that connects source records and data warehouse, that is, representational layer of a warehouse (Kimball et al., 2013).

important to define one format of every information that will be saved in a data warehouse.

Value: is referred to value of every fact that is saved to a data warehouse, its accuracy and reasonableness. Example: being careful when inserting records that describe specific product since it is easy make a typing mistake. Wrong information that is saved to a data warehouse won't be clear to the user.

Conformity: every fact that is saved to multiple places must be in the same format in every table. Multiple formats can be specified for saving specific data type and it is important to be consistent with the defined formats. Example: saving data of type DATE throughout data warehouse. Type DATE has a lot of predefined formats and one format must be chosen that will be used in a DW, so later a faster and easier search on the necessary data can be done and more time won't be lost to transform existing data to a unified format.

Accuracy: data must represent events from a real world, that is, data saved to a DW must have happened or must exist in the real world. Example: address of every user is an existing address. Data that is incorrect must be removed from a data warehouse since it can lead to bad business decisions.

Integrity: Data is saved throughout many tables and there must exist a way to mutually connect the tables to have complete image of the saved data. If it's not possible to connect necessary tables, data will be duplicated in all tables that can't be connected - with that memory capacity and velocity of retrieving and processing the data will be lost.

Example: there are two tables that hold data about customers and products and a DW is used for different reports. There is a need to change products table because for one of the products there are new values that need to be saved in a DW. The existing table cannot be changed since it is already in use, but instead of that, a new instance of the same table with a few changes in the attributes can be created.

When creating a data warehouse, all the criteria from above must be considered. If one of them is skipped, there will be data losses in a DW and the information the user is receiving won't be consistent which will result with the user being unsatisfied with the results.

6 Classification of losing data quality

Except having to consider all the criteria for data quality loss, it must be clear where exactly bad data is located so specific processes can be improved to increase data quality in the data warehouse.

Data quality can be compromised in different phases of a DW development cycle and in different layers of a DW architecture: in the data sources, in data integration, ETL processing or data model.

6.1 Source data

Most of the data warehouses today are collecting data from multiple (and heterogeneous) data sources. Some of that sources are included into a DW design from the beginning and some of them are being added as new business requirements are changing or progressing. Data in the data sources are often typed or selected by the users with a help of different applications and information systems or they are generated from some devices. Each of that device or application has its own way of saving the data. A security level must be ensured and embedded into the applications and devices so not everyone can generate data that is later processed. If the data generator is not known and under control, data becomes unreliable and that is affecting data quality directly. Some of the issues that affect low quality in data in this phase are: bad choice of source data, different time of loading the data from different sources, lack of checking and limiting data in the source applications/devices, using different formats in the source systems, bypassing defined standards, possibility of inserting empty values, additionally adding columns to the tables or input fields in the applications, using special signs inconsistently, using default values for missing data, incomplete metadata.

6.2 Data integration

This is a phase of loading data from multiple data sources and checking its type and format to ensure that data that is being inserted to a data warehouse is consistent. After it has been decided which data sources will be used, a data profile for every source needs to be defined in order to decide which data is necessary or not, and to define a format of all data types that will be saved to a DW. These steps are often skipped or developers don't have enough time to define all details, since in the beginning of the project it is more important to have a stable system that will send the data, process it and have a stable infrastructure for a data warehouse. If the data is not profiled right away, later during maintenance process data will have to be filtered and transformed more often to have more accurate and meaningful information that represents real business events. Issues that affect data quality in the integration phase are: selecting inappropriate profiling tool, insufficient analysis of the source data, not limiting input boxes to a specific character length, insufficient analysis of total records that are being inserted, inaccurate data parsing and applying standards, not knowing the data model and relationships between the tables.

6.3 ETL

In this phase data is imported from the sources to a data warehouse. Different transformations are being applied to data in order to adjust it to a DW model and standards and then store them there. This is one of the most important phases of handling the data warehouse. Here, data is saved and is not being checked unless an anomaly is noticed and must be corrected. It is important to plan all problems in the source data, make specific actions to standardize that data and import to the data warehouse only the data that is necessary for business analysis. In an ETL process, every record from the source data is being checked and data model must be well known by now (its formats and standards) in order to create a quality ETL process that will fix or discard all the data not suitable for the DW. Before this phase, a source to target analysis must be made, the data must be integrated and data model known so it can properly map all the data to appropriate tables. Process of data refresh is a process which is applying changes from a data source to a data warehouse. This process is being executed through different cleaning phases, integration and filtering phases (Vassiliadis et al., 2000). These functions can be sorted consecutively with many different combinations and they need to be carefully applied so none of the necessary data is lost. Every system is unique, the order that transformations are applied, helps owning a higher quality data warehouse.

In an ETL process, data quality depends of three factors (Vassiliadis et al., 2000):

- Data coherence – maintaining data in the same metrics system leads to data of higher quality.
- Data Completeness – percent of data found in a data warehouse, compared to amount of data that is necessary. Also referred to having all the necessary data in the moment it is being needed, not having missing or incomplete data.
- Data age – for older data in the data warehouse, it is possible that it’s no longer accurate. This is referring to the data that has an end date and new records are inserted with new and valid values.

Major problems that produce low quality data in ETL phase are: different business rules in the source data, inability to define one schedule to refresh data, bad detecting of only new or changed data to save to a warehouse, not knowing ETL standards to manipulate data, wrong selection of data that will be saved to slowly changing dimensions, wrong source to target mapping, incorrectly manipulating with NULL values, lack of data documentation, wrong handling insert or update dates, incorrect usage of insert, update and delete functions, inconsistent naming of the ETL processes (sessions, transformations, flows, jobs) losing records that don't meet specific criteria, inability to restart an ETL job from specific checkpoint without losing data (Singh et al., 2010), (Golfarelli et al., 2017).

6.4 Data model

Schema of the data model must be designed and prepared to save the data for specific business of the company. This is one of the first steps when building a data warehouse and it's important to dedicate enough time and watch all details of the business requirements. There are possible problems with data quality that are affected by data model: failure to satisfy project criteria, delay in source data, wrong data granularity in the tables, inability to follow standards caused by wrong or incomplete tables definitions or relationships between them (Singh et al., 2010), (Golfarelli et al., 2017).

7 Different aspects of data quality

Data quality can be observed through different criteria, phases of data manipulation and category of users. The last approach depends on the user and his purpose in using the data. Based on that, the same data can have different level of quality to people with different roles and different purpose of work (Jarke et al., 2006).

Decisions maker is a person that is using different data analysis tools to extract the information he needs. He is also taking care of quality of the data that is already saved to a warehouse, time the data is inserted there, last time it was refreshed and ease of writing a query over necessary tables.

Data warehouse administrator's job is to define methods to detect data anomalies, find out what is the cause of every anomaly and what consequence that anomaly has made on existing data. Administrator is also taking care of data age, anomalies reports, functionality of tools that build the reports and last, metadata and its availability.

Data warehouse designer is in charge of metrics saved to a data warehouse, complete warehouse data model, changes that occur in the model and how are they applied in practice. To a designer, metadata is important since it is describing which table is storing which data and how are the tables mutually connected. DW designer also needs to take care of tools for DW maintenance - they need to be tested to make sure they meet all the conditions for a specific business data warehouse.

Data warehouse programmers develop tools and programs for data warehousing and they need to know what kind of data will be kept in the DW in order to build specific function to manipulate the data and its tables. Also, they need to know standards for managing and implementing a data warehouse so they can integrate those standards within the tool.

8 Consequences of bad quality data

A person that is using a data warehouse daily will best describe the problems he or she encounters, omissions in the data and if he or she eventually sees the cause of the problems. With that information the mistakes can be analyzed and removed and new standards implemented in order to have a higher quality data in a data warehouse.

Except taking improving guidance from people that are working on a DW on a daily basis, bad data can be detected with cost analysis, that is, if expenses of maintaining the DW have risen and the amount of data didn't increase, quality issues exist.

If the data is used as support to making business decisions, and if some of the decisions ended badly, it's possible that data that affected that decisions is incomplete or incorrect. Decisions made based on faulty data can reflect employee dissatisfaction.

Customer dissatisfaction is also one of the indicators of bad data. This is happening while using online services and generating personalized messages based on customer interest and previous purchases. If the customer sees suggestions he or she doesn’t like, he/she will use other services in the future.

9 Related work

Data quality in a DW is a widespread issue, influenced by a lot of events and business processes and can be managed in a lot of ways. Many authors dedicated their time to investigate causes of the data quality problems, and some widely known research was already mentioned in the paper.

Wixom and Watson (Wixom et al., 2001) conducted a research about implementation success factors of a data warehouse. They concluded that strong connection between system quality and data quality, good management and organized resources with high quality skills are associated with higher quality data.

Calero, Piattini, Pascual and Serrano (Calero et al., 2010) researched metric that is affecting data quality: table metrics, star metrics and schema metrics. They described metrics that anyone can use to measure data quality by separate categories.

Vuk, Cirikovic and Suk (Vuk et al., 2015) explored different approaches to data quality and factors defining data quality. Their main research question was how can multiple data sources affect data quality with time, life expectancy of specific data stored in the warehouse, and when it becomes inconsistent.

Vassiliadis, Bouzeghoub, Jarke and Quix (Vassiliadis et al., 2000), (Quix, 1999) and (Jarke et al., 2006) pointed out importance of metadata and how it affects overall content quality. They presented quality factors such as new source integration and how they affect other dependent tables and metrics, and of course, importance of good data warehouse maintenance.

R. Singh and Dr. K. Singh (Singh et al., 2010) introduced stages of data warehousing that can affect data quality and they classified problems in each stage. The summary of most common problems was: lack of standards, varied data sources, missing and inconsistent data and inadequate data quality testing.

Additionally, some newer research and emerging trends in a field of DW data quality will be mentioned here. It can be seen that some new data quality issues appeared in the context of data stream management systems, data lakes, big data, view maintenance quality in a DW and in general data quality related to different DW design methodologies.

In (Jarke et al., 2017) the authors researched the role of conceptual models, their formalization and implementation as knowledge bases, and the related metadata and metamodel management. They traced this evolution from traditional database design, to data warehouse integration, to the recent data lake architectures.

In (Berti-Equille et al., 2016) the authors discussed data quality problems in the context of Big Data applications and the challenges to data quality regarding the unprecedented volume, large variety, and high velocity of data.

In (Geisler et al., 2016) the authors researched data quality in Data Stream Management Systems (DSMS). DSMS provide real-time data processing in an effective way, but there is always a tradeoff between data quality and performance. They proposed an ontology-based data quality framework for relational DSMS that included data quality measurement and monitoring in a transparent, modular, and flexible way.

In (Gosain et al., 2019) the authors focused on view maintenance models’ quality in the context of a DW data quality. In this context they mentioned problems such as definition of views, composition of views and maintenance of views but focused their research on the problem of the judicious use of materialized views so as to achieve the best combination of good query performance and low view maintenance in terms of cost efficiency.

In (Di Tria et al., 2017) the authors provide metrics for evaluating the quality of multidimensional schema in reference to the effort spent in the design process and the automation degree of the data warehousing methodology. They theirour evaluation to the major emerging hybrid methodologies for data warehouse schema design, as a case study.

In (Gautam, 2018) the author argued that the formal representation of metadata can enhance the quality of information required for decision making. The author introduced three formal models to properly manage metadata with different types of operation and to provide qualitative information to user.

In (Golfarelli et al., 2017) the authors show a very good overview of more than 20 years of research on data warehouse systems, from their early relational implementations (still widely adopted in corporate environments), to the new architectures solicited by Business Intelligence 2.0 scenarios, and up to the new challenges of integration with big data settings.

In (Homayouni,2019) the authors propose ADQuaTe, an automated data quality test approach that uses an unsupervised machine learning technique to discover constraints in the DW that may have been missed by experts.

In (Heine, 2019) the authors describe a prototype of an automated data quality monitoring system. The focus is on the aspect of expressing advanced data quality rules such as checking whether data conforms to a certain time series or whether data deviates significantly in any of the dimensions within a data cube.

In (Mann, 2020) the authors summarized various metrics proposed for conceptual model of Data Warehouse and their formal and empirical validation to prove their correctness and practical utility. It is a good state of the art paper on this topic.

10 Conclusion

Data quality is an important issue, nowadays more than ever. In order to make a good business decision, the companies need to have timely, accurate and realistic data.

Data quality in a data warehouse (DW) always needs to be in focus since it has direct consequences to the business and the customers. Processes can always be modified and adjusted to the new standards, but it's important to invest enough time, money and educated employees to build a good and quality data warehouse. If data quality control is skipped in the data warehouse building phase, additional analysis and process implementation will cost much more - it will demand more people and eventually cause down time in the moment when the changes are being implemented.

Business model is changing as the market demands change, in order to serve best possible product and make the best profit. Changing business rules means changing the data warehouse model. Every model holds its advantages and disadvantages and it is important that disadvantages are well known from the beginning and that people are working on reducing or removing them, that standards are being applied and data quality is being tested constantly. The data warehouse structure is constantly being replaced with a newer, better models, but developers and DW architects must not forget to check the data quality from different perspectives, many categories and various criteria, in order to have a high-quality data in a data warehouse.

The main research question of this paper was “What are the most common data quality issues in a data warehouse?” and the goal of the paper was to give an overview of data quality issues in a data warehouse. The main contributions of this paper are: a) an overview of data quality issues in the context of data warehouses and b) an insight into some new and emerging research areas in the field of DW data quality.

It can be noted that the research based on a traditional DW methodologies and architectures is very comprehensive and very well defined – there is a little need to further research this area. However, some new and emerging trends in data influenced further (and future) research in a data quality filed in some new contexts and perspectives. Some of these are definitely fields of data stream management systems, data lakes, big data, view maintenance quality in a DW and in general data quality related to different DW design architectures and methodologies.

Acknowledgements

This paper is based upon work supported by the University of Rijeka under projects “uniri-drustv-18-182”.

References

Berti-Equille, L., Quix, C., Gudivada, V., Hai, R. & Wang, H. (2016). Quality in Databases. *The 11th International Workshop on Quality in DataBases in conjunction with VLDB*.

Calero, C., Piattini, M., Pascual, C. & Serrano, M. A. (2010). Towards Data Warehouse Quality Metrics. *ALARCOS Research Group*.

Chen, T. X., Meyer, M. D., Ganapathi, N., Liu, S. & Cirella, J. M. (2011). *Improving Data Quality in Relational Databases: Overcoming Functional Entanglements*. TRI Press.

Di Tria, F., Lefons, E. & Tangorra, F. (2017). Cost-benefit analysis of data warehouse design methodologies. *Information Systems*.

Gautam, V. (2018). Qualitative model to enhance quality of metadata for data warehouse. *International Journal of Information Technology*.

Geisler, S., Quix, C., Weber, S. & Jarke, M. (2016). Ontology-based data quality management for data streams. *Journal of Data and Information Quality (JDIQ)*.

Golfarelli, M. & Rizzi, S. (2017). From star schemas to big data: 20+ years of data warehouse research. *A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years,* Springer.

Gosain, A., Sabharwalb, S. & Guptac, R. (2019). Validating Quality metrics for the View Maintenance Models of Data Warehouse. *International Journal of Applied Engineering Research*.

Heine, F., Kleiner, C. & Oelsner, T. (2019). Automated Detection and Monitoring of Advanced Data Quality Rules. *Database and Expert Systems Applications (DEXA 2019), Lecture Notes in Computer Science*, Springer.

Homayouni, H., Ghosh, S. & Ray, I. (2019). ADQuaTe: An Automated Data Quality Test Approach for Constraint Discovery and Fault Detection. *IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI)*. Los Angeles, CA, USA.

Inmon, W. H. (2002). *Building the Data Warehouse (Third Edition)*. John Wiley Press, NY, USA.

Jarke, M., Jeusfeld, M. A., Quix, C. & Vassiliadis, P. (2006). Architecture And Quality In Data Warehouses An Extended Repository Approach. *Other publications TiSEM*, Tilburg University, School of Economics and Management.

Jarke, M. & Quix, C. (2017). On warehouses, lakes, and spaces: the changing role of conceptual modeling for data integration. *Conceptual Modeling Perspectives*, Springer.

Kimball, R. & Ross, M. (2013). *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling (Third Edition).* Wiley Press, NY, USA.

Mann, S. & Siwach, M. (2020). Data Model Quality Metrics of Data Warehouse: A Survey. *Proceedings of the International Conference on Innovative Computing & Communications (ICICC)*.

Quix, C. (1999). Repository Support for Data Warehouse Evolution. *DMDW,* 6-28.

Singh, R. & Singh, K. (2010). A Descriptive Classification of Causes of Data Quality Problems in Data Warehousing. *International Journal of Computer Science Issues (IJCSI)*, 7(3).

Vassiliadis, P., Bouzeghoub, M. & Quix, C. (2000). Towards Quality-Oriented Data Warehouse Usage snd Evolution. *Information Systems*, Pergamon, 25(2), 89-115.

Vuk, D., Ciriković, E. & Suk, D. (2015). Kvaliteta podataka i njen značaj danas. *Visoka škola Virovitica*, 54-58.

Wang, R. Y. & Strong, D. M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, 5-33.

Wixom, B. H. & Watson, H. J. (2001). An empirical investigation of the factors affecting data warehousing success. *MIS Quarterly*, 17-41.