Effective Big Data Management and Opportunities for Implementation

Manoj Kumar Singh
Adama Science and Technology University, Ethiopia

Dileep Kumar G.
Adama Science and Technology University, Ethiopia

A volume in the Advances in Data Mining and Database Management (ADMDM) Book Series

Information Science REFERENCE
An Imprint of IGI Global

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue
Hershey PA, USA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com

Library of Congress Cataloging-in-Publication Data

Names: Singh, Manoj Kumar, editor. | Kumar G., Dileep, 1982- editor.
Title: Effective big data management and opportunities for implementation / Manoj Kumar Singh and Dileep Kumar G, editors.
Description: Hershey : Information Science Reference, 2016. | Includes bibliographical references and index.
Identifiers: LCCN 2016004454| ISBN 9781522501824 (hardcover) | ISBN 9781522501831 (ebook)
Subjects: LCSH: Big data. | Database management.
Classification: LCC QA76.9.D3 E335 2016 | DDC 005.7--dc23 LC record available at https://lccn.loc.gov/2016004454

This book is published in the IGI Global book series Advances in Data Mining and Database Management (ADMDM) (ISSN: 2327-1981; eISSN: 2327-199X)

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

Chapter 5
The Challenges of Data Cleansing with Data Warehouses

Nigel McKelvey
Letterkenny Institute of Technology, Ireland

Kevin Curran
Ulster University, UK

Luke Toland
Letterkenny Institute of Technology, Ireland

ABSTRACT

Data cleansing is a long standing problem which every organisation that incorporates a form of data processing or data mining must undertake. It is essential in improving the quality and reliability of data. This paper presents the necessary methods needed to process data at a high quality. It also classifies common problems which organisations face when cleansing data from a source or multiple sources while evaluating methods which aid in this process. The different challenges faced at schema-level and instance-level are also outlined and how they can be overcome. Currently there are tools which provide data cleansing, but are limited due to the uniqueness of every data source and data warehouse. Outlined are the limitations of these tools and how human interaction (self-programming) may be needed to ensure vital data is not lost. We also discuss the importance of maintaining and removing data which has been stored for several years and may no longer have any value.

1. INTRODUCTION

Processing and analysing data has become increasingly important to organisations in recent years. As companies are growing and adapting, the ability to retrieve current and correct data is of key importance. Data cleansing, cleaning or scrubbing is the process of identifying and removing or modifying incorrect entries or inconsistencies in a dataset to improve the overall quality (Rahm et al, 2000). Data

DOI: 10.4018/978-1-5225-0182-4.ch005

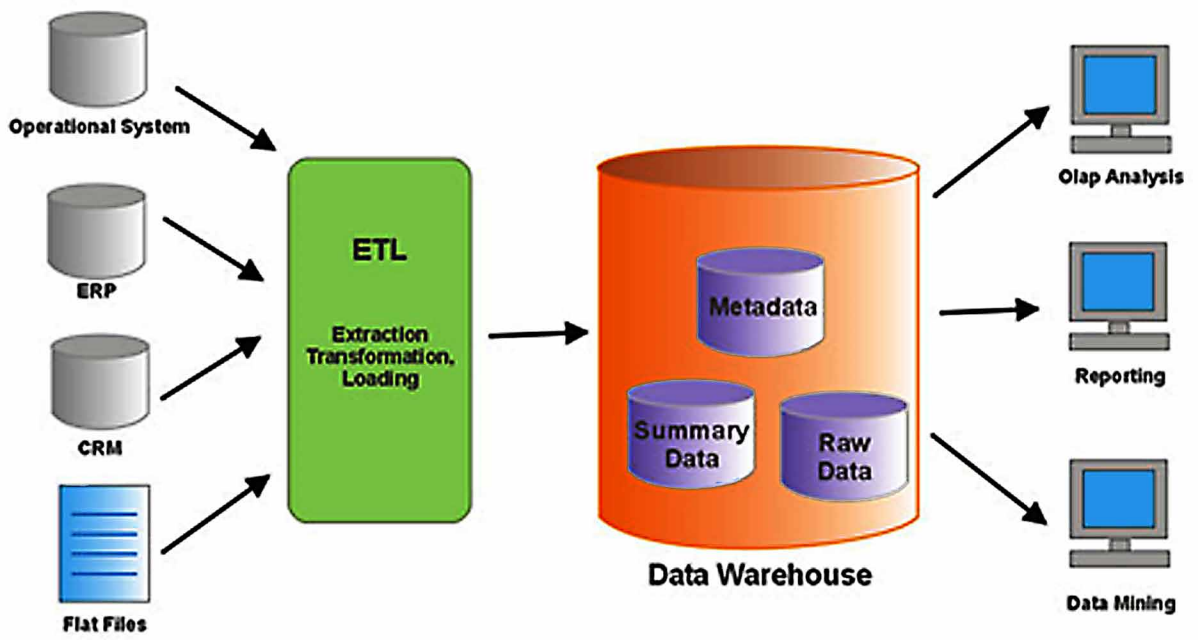
warehousing is the concept of storing data in a relational database which is designed for query and analysis rather than transaction processes (Docs.oracle.com, 2014). It is also referred to as an organisation's "single source of truth". It is designed to provide management with a large amount of data from multiple sources within the organisation, which is vital in strategic decision making. For data to be stored in a data warehouse, it is crucial that it is cleansed. This process becomes more difficult as retrieving data from multiple sources increases the amount of "dirty data" and may also introduce an inconsistency in the way in which the data is represented.

Figure 1 describes the typical flow and layout of a data warehouse. Extraction, Transformation and Loading is the process reliable for the initial loading and refreshing the contents of the data warehouse. The probability of this data being incomplete or incorrect is quite high as it has been retrieved from multiple sources, therefore the data is processed through a number of methods, which include instance extraction and transformation, instance matching and integration, filtering and aggregation. Data cleansing is normally performed in a separate area before data is loaded into the data warehouse. The sheer volume of data being processed means that writing a successful tool to complete this task is very difficult.

2. DATA QUALITY

Data auditing is the first step in the data cleansing process. Its purpose is to process through the data and outline any data anomalies that are found (Muller et al, 2003). Using statistical and parsing methods, this process derives information such as value range, frequency of values, variance, uniqueness, occurrence of null values, typical string patterns, also detecting any functional dependencies and association rules in the complete data collection (Muller et all, 2003). Data quality refers to the standard, reliability and

Figure 1. Data warehouse model



efficiency of data to inform and evaluate decisions (Karr et al, 2003). For data to be processed as fast and efficiently as possible, data must adhere to a certain standard. Data which adheres to this standard is said to be of high quality. To measure the quality of a data collection, scores are assessed. The result of these scores will identify the need to data cleanse and to which level of data cleansing is performed.

Before data is entered into a database, it usually is passed through a number of phases which include human interaction and computation. Naturally, data errors will occur through typographic or formatting errors or misunderstanding of the data source. To ensure data quality, during its lifespan, certain data will undertake iterative processes involving collection, transformation, storage, auditing, cleaning and analysis. This may be spread across multiple organisations and or agencies, potentially over large amounts of time. On average, U.S organizations believe that 25% of their data is inaccurate (Qas.com, 2014). Quality assurance is a major part of a data warehouse model. Ensuring that data is kept to a high standard when stored in a data warehouse is a time consuming and expensive task. Using processes such as Extraction, Transformation and Loading (ETL) can keep content current and correct. Data cleansing is executed in the data staging area. After data is collected from its source, it is gathered in this area and prepared for analysing. ETL and staging are considered to be the most important sequence of events to occur in the data warehouse (Singh et al, 2010). It is an important area for tracking down errors and undertaking audits to validate data quality.

3. PROBLEMS

Data cleansing in a data warehouse is a problematic task. Each data warehouse and each source can be unique in its own way. This means that each source can create a new conflict, which may not have been planned or foreseen. The degree to which each problem occurs is largely based on the schema level and instance level assigned to each data source. This section will explore the problems associated with both sourcing methods. Due to data cleansing being an expensive and time consuming task, it is just as important to reduce the cleaning problem before analysing data.

3.1. Semantic Complexity

Semantic complexity is described as the user's representation of what data represents in a given database. This is a common problem as different users may have different conceptions of what the data represents. For example, a two databases containing a list of drivers and their cars are to be merged. There may be a data conflict where by a user will use the driver's licence number as a primary key and the other user may use the driver's social security number as a primary key. This can lead to an occurrence of missing values, meaning any data that is requested may be incorrect.

3.2. Multi-Source Problems

Problems which may be present at single-source level are only more emphasised at multi-source. Data from each source may contain certain attributes which may be represented differently and can overlap or contradict. This may be because each source is developed uniquely and could be tailored to suit different applications, and when combining several sources this results in a large degree of heterogeneity.

At schema level, the main issues are schema translation and schema integration, specifically being naming conventions and structural conflicts in the databases. Naming conflicts arise when attributes in different databases are assigned the same name and represent the same object in different sources. Structural conflicts arise when there are variations in the representation of data in different sources. Due to these conflicts, data is merged from different sources may be entered more than once creating duplicate entries and contradicting records.

In an ideal scenario, data from different sources would complement each other without overlapping or creating "dirty data". Thus, the main challenge of cleaning data retrieved from different sources is identifying the overlapping data. This problem is often referred to as object identify problem, the merge/purge problem, or the duplicate problem (Hernandez et al, 1998). Although this problem can be overcome, there are a few issues which may arise:

- The dataset may be too large to reside in main memory at a single time. Therefore, the main dataset may have to exist in external memory allowing as few pass overs as possible to solve the problem.
- The incoming data may be corrupted which makes it difficult to compare matches.

In a multiple source scenario, each source may also have different platforms. As an example, IBM mainframe may contain files on a server using an Oracle database, while another IBM mainframe may contain files on a server using SQL. The exact amount of sources needed for a data warehouse are based on the needs of the organisation and the specifics of the implementation.

3.3. Single-Source Problems

Since data warehouses are mostly built using multiple sources, single-source problems are uncommon, but may still arise. At instance level, errors that occur are largely similar to errors which occur at instance level from multiple sources; entry errors, misspellings, duplicate entries. At schema level, problems facing data cleansing are slightly different to multiple source problems. Some of which are poor schema design, referential integrity and lack of integrity constraints.

3.4. Data Transformation

Data transformation is the process of converting the format of one data set into another dataset (Techopedia2.com, 2014). It is usually carried out after the data to be transformed has been verified. Verification ensures that the data to be transformed is evaluated to ensure that it will be necessary to cleanse.

There are currently many tools which are available to complete this task, but are expensive and can be ineffective. For these reasons it is vital that the organization has a working knowledge of their data sources. This method is also used to refresh current information in the data warehouse to ensure it is of high data quality.

3.5. Data Maintenance and Improving Performance

After data has been cleansed, it is important to ensure that it is being managed and updated/refreshed on a regular basis. In an ideal situation, a data warehouse would be able to retain an organisation's information for several years, but in order to retain the system's performance a portion of the data is going to have to

be extracted. Data purging is a term that is commonly used to describe methods that permanently erase and remove data from a storage space (Techopedia.com, 2014). Typically organisations will come to a decision on the data to be removed from the data warehouse.

Purging a database can be executed by a program which will use circular buffer methodology, meaning the oldest data must make way for the newest data. This is an inefficient method as the importance of data should not be ranked on its age but rather its use to the organisation. Therefore, it can be better practice to manually purge a database. This will involve a single or several administrators manually selecting data to be removed from the database, ensuring that important data is retained.

4. CONCLUSION

Data cleansing is a fundamental process in every data warehouse. Data cleansing can be defined as a list or sequence of operations which improve overall data quality. Data warehouses can supply organisations with endless information and statistical observations. Therefore, to retrieve the most current and correct information each data collection should be cleansed before it is entered into the data warehouse. In this paper, the author has provided a rough description of the problems associated with data cleansing from multiple and single sources while trying to provide a focus on schema-level and instance-level problems. We have discussed the importance of maintaining already cleansed data, specifically the benefits to manually reviewing data to be removed over automatically removing data using a tool. It has also become evident that creating a tool to cleanse a data source is a tough task. Data is represented in many different formats, and a single tool may not be efficient in cleansing a number of data sources. It has become clear that although these tools can prove useful in a number of areas, such as removing duplicates, some human intervention may be needed to review data and users may need to self-program these tools to suit the circumstances.

REFERENCES

Bradji, L., & Boufaida, M. (2011). Open User Involvement in Data Cleaning for Data Warehouse Quality. *International Journal of Digital Information and Wireless Communications*, *1*(2), 536–544.

Docs.oracle.com. (2014). *Data Warehousing Concepts*. Retrieved from: http://docs.oracle.com/cd/B10500_01/server.920/a96520/concept.htm

Helfert, M. & Herrmann, C. (2002). *Proactive data quality management for data warehouse systems*. Academic Press.

Hernandez, M. & Stolfo, S. (1995). *The merge/purge problem for large databases*. Academic Press.

Hernandez, M., & Stolfo, S. (1998). Real-world data is dirty: Data cleansing and the merge/purge problem. *Data Mining and Knowledge Discovery*, *2*(1), 9–37. doi:10.1023/A:1009761603038

Karr, A., Sanil, A. & Banks, D. (2003). *Data Quality: A Statistical Perspective*. Academic Press.

Lee, M., Lu, H., Ling, T., & Ko, Y. (1999). Cleansing data for mining and warehousing. Academic Press.

Monge, A., Elkan, C., & Associates. (1996). The Field Matching Problem: Algorithms and Applications. Academic Press.

Muller, H., & Freytag, J. (2003). *Problems*. Methods, and Challenges in Comprehensive Data Cleansing.

Qas.com. (2014). *Contact Data Management Software and Services | Experian Data Quality*. Retrieved from: http://Qas.com

Rahm, E., & Do, H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, *23*(4), 3–13.

Singh, R., & Singh, K. et al. (2010). A descriptive classification of causes of data quality problems in data warehousing. *International Journal of Computer Science Issues*, *7*(3), 41–50.

Techopedia2.com. (2014b). *What is Data Transformation? - Definition from Techopedia*. Retrieved from: http://www.techopedia.com/definition/6760/data-transformation

Techopedia.com. (2014a). *What is Data Purging? - Definition from Techopedia*. Retrieved from: http://www.techopedia.com/definition/28042/data-purging