Global Journal of Computer Science and Technology: G Interdisciplinary
Volume 23 Issue 2 Version 1.0 Year 2023
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Coping with Data Inconsistencies in the Integration of Heterogenous Data Sources

By Joshua Edem Agomor & Meda Saawah Appiah

Abstract- This research examines the problem of inconsistent data when integrating information from multiple sources into a unified view. Data inconsistencies undermine the ability to provide meaningful query responses based on the integrated data. The study reviews current techniques for handling inconsistent data including domain-specific data cleaning and declarative methods that provide answers despite integrity violations. A key challenge identified is modeling data consistency and ensuring clean integrated data. Data integration systems based on a global schema must carefully map heterogeneous sources to that schema. However, dependencies in the integrated data can prevent attaining consistency due to issues like conflicting facts from different sources. The research summarizes various proposed approaches for resolving inconsistencies through data cleaning, integrity constraints, and dependency mapping techniques. However, outstanding challenges remain regarding accuracy, availability, timeliness, and other data quality restrictions of autonomous sources.

Keywords: data, data amalgamation, data inconsistency, data dependencies, integrity constraints, schema.

GJCST-G Classification: FOR Code: 0806

Strictly as per the compliance and regulations of:

Coping with Data Inconsistencies in the Integration of Heterogenous Data Sources

Joshua Edem Agomor [α] & Meda Saawah Appiah [σ]

Abstract- This research examines the problem of inconsistent data when integrating information from multiple sources into a unified view. Data inconsistencies undermine the ability to provide meaningful query responses based on the integrated data. The study reviews current techniques for handling inconsistent data including domain-specific data cleaning and declarative methods that provide answers despite integrity violations. A key challenge identified is modeling data consistency and ensuring clean integrated data. Data integration systems based on a global schema must carefully map heterogeneous sources to that schema. However, dependencies in the integrated data can prevent attaining consistency due to issues like conflicting facts from different sources. The research summarizes various proposed approaches for resolving inconsistencies through data cleaning, integrity constraints, and dependency mapping techniques. However, outstanding challenges remain regarding accuracy, availability, timeliness, and other data quality restrictions of autonomous sources. Additional research is needed to develop more automated ways of reconciling inconsistencies from source data with the requirements of the global schema. The ability to provide high-quality integrated data is crucial for organizations to maximize the value of their information assets. This research aims to promote further investigation into semi-automated remediation of inconsistencies and leveraging source data quality metrics to aid the integration process. Overcoming inconsistencies is critical to enabling unified views and meaningful analytics from merged cross-organizational data.

Keywords: data, data amalgamation, data inconsistency, data dependencies, integrity constraints, schema.

I. Introduction

Data is raw facts (M. Chen et al., 2009). Having the same data in different formats and in many tables causes inconsistent data. Data integration, also known as data amalgamation, is the act of merging data from several sources into cohesive sets of information for operational and analytical reasons (Lenzerini et al., 2014). One of the fundamental components of the entire process of data management is integration of data, its primary goal is to create clean, consistent, and consolidated data sets that meet the information requirements of various organization end users. An integrated view of diverse databases is referred to as a global schema. However, an important part of creating a global schema is identifying common types of information from the various local schemas. The world wide web has facilitated a prevalent access to autonomous, distributed, and dissimilar sources of data and has gotten worse. However, external users can now access an escalating quantity of records or databases, particularly the publicized ones on the internet and media. When converting user requests to queries across multiple data sources with varying data quality, this process does not take the quality of the data sources into consideration (Hariri, Fredericks, & Bowers, 2019).

In terms of dealing with inconsistent data in information amalgamation or integration, there are essentially two methods (DeCastro-García, Muñoz Castañeda, Fernández Rodríguez, & Carriegos, 2018). The first method in nature which is also based on domain-specific data cleaning, however, it is bureaucratic, and alteration methods used on sources' data that have been obtained. Declarative is used as the second strategy (DeCastro-García et al, 2018). In essence, several studies propose methods for giving insightful responses even when a database does not adhere to its integrity criteria (Cao, Lu, & Wen, 2019).

One of the most difficult aspects of data amalgamation is dealing with discrepancies or data being inconsistent (Lenzerini, Salaria, & Roma, 2014). The data integration systems in this work are differentiated by having a global schema-based design and a variety of sources. The sources include the relevant data, but the global schema conceals, integrates, and displays a virtual picture of the underlying sources. A mapping establishes connections between data sources and global schema components. Inconsistency may occur because sources may contain data that, when combined with other sources, may contradict constraints, and the global schema typically contains integrity constraints. Since the ability of a data integration system to respond to queries in terms of the global schema is one of its primary objectives, and because the response to a query is based on the data stored in the sources, inconsistency has a substantial impact on the system's capacity to respond in a meaningful way.

The global schema frequently includes integrity requirements, and sources may contain data that violates integrity constraints when combined with data from other sources. Since responding to inquiries in terms of the global schema is one of the primary goals

Author α: Ghana Institute of Management and Public Administration, School of Technology. e-mail: joeagomor@gmail.com
Author σ: American University, Kogod Business School. e-mail: medappiah@gmail.com

of a data amalgamation system. How to model the consistency of the data and, consequently, specify and establish that the data is clean, is one of the most crucial concerns in relation to data cleaning (Angeles & MacKinnon, 2004). Data amalgamation has been grappling with the difficult task of resolving structural, syntactic, and semantic heterogeneities between source and target data for several years.

II. Review of Literature

Different departments within a large, contemporary company will very certainly create, store, and search for their vital data using various platforms. However, the company can only fully understand the value of the data they hold by merging the information from these diverse platforms. One method of integrating data is called database federation, in which a relational database management system is used as middleware to give users uniform access to a variety of disparate data sources. According to (Haas et al., 2002), they went through the fundamentals of database federation, introduce a few different types, and specify the circumstances in which each type of partnership should be employed.

Angeles & MacKinnon in 2004 contend that user quality priorities, data inconsistencies, and data quality differences among the participating sources have not been adequately addressed by the processes and optimization of information integration, such as query processing, query planning, and hierarchical structuring of user results. They suggested creating a data quality manager to manage semantic heterogeneity and data quality by establishing communication between the information integration process, the user, and the application. To specify the quality standards, metrics, and evaluation procedures, data quality manager will include a reference model, a measurement model, and an Assessment Model. By taking into consideration, data quality estimates to discover the ideal combinations for the execution plan, data quality manager will also aid in query planning. Data quality may also be utilized to resolve data inconsistency after query execution and inconsistent data detection. This method will result in the integration and rating of query results using the user-defined quality criteria. (Angeles & MacKinnon, 2004).

In this paper, we provide our initial set of findings on lowering uncertainty during data integration. We contend that certain guidelines must be followed when handling uncertainty at each of three different levels by data integration systems. First, because there may be too many of them to create and maintain or because the proper mappings might not be clear in certain fields, the mappings between the data sources and the mediated schema may only be generally correct (such as bioinformatics). Second, if information extraction methods are utilized to get the data from the sources, inaccurate data could be generated. Third, inquiries may be sent to the system using keywords rather than needing to follow a prescribed syntax.

a) Data Quality

Any inconsistent data and integration of data into systems would not be complete without hitting on how quality the data is. This could be the first challenge to affect the consistency of the data. Any challenge along one or more quality dimensions that renders data totally or partially inappropriate for use is what we refer to as a “data quality problem” (Strong et al., 1997). The term "high data quality" refers to data that is appropriate for use by data consumers and is handled regardless of the environment in which it is created and consumed. (Angeles & MacKinnon, 2004). According to (Angeles & MacKinnon, 2004). Accuracy, completeness, consistency, and timeliness are some examples of quality criteria or dimensions that have been used to define data quality.

b) Data Integration

Applications that need querying across several independent and heterogeneous data sources encounter a common difficulty called data integration. Progress in large-scale scientific projects where numerous researchers independently create data sets, improved collaboration between agencies from the government having a high-quality search and their sources of data across multiple sources on the internet all depend on how their data has been amalgamated. (Ioannou & Staworko, 2013)

Amalgamation of data could also be seen from diverse angles, however it cannot be said without considering the volume thus, not only may each data source have a substantial amount of data, but there are now tens of thousands of data sources, even for a single field. (Sena et al., n.d.)Also said Velocity, thus, many of the data sources are particularly dynamic because of the rate at which data is being collected and regularly made available. For instance, there are several data sources that offer close to real-time, constantly updating data on the stock market, such as bid and ask prices, number of shares traded, etc. Traditional data amalgamation techniques are unable to provide an integrated view of stock market data from all these data sources. Veracity, thus, there are substantial disparities in the coverage, quality, and timeliness of the data given by different data sources (even within the same topic).Variety, thus, when it comes to how data is formatted at the schema level and how they represent the same real-world object at the instance level, data sources (even within the same domain) are incredibly diverse(Debattista et al., 2015). They show a great deal of variation even for quite comparable entities.

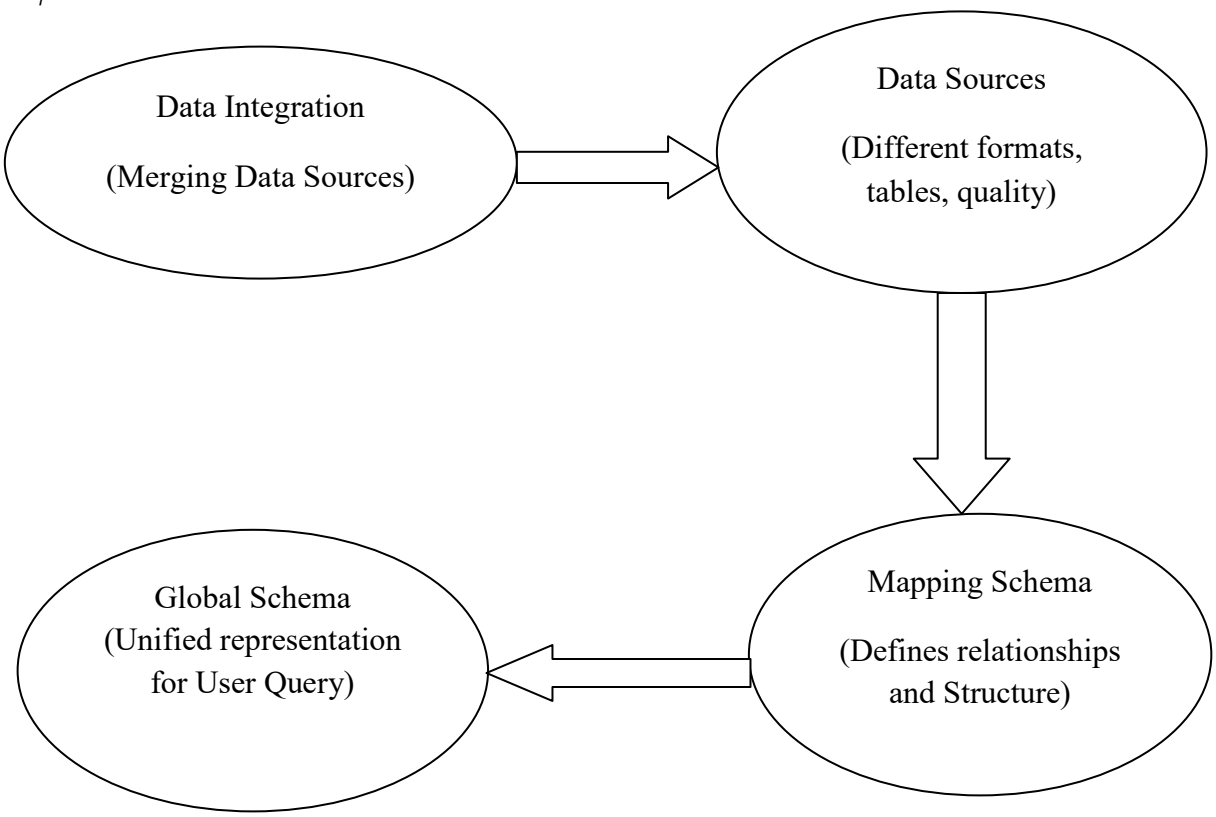
Uncertain mappings in the schema; Data amalgamation solutions employ schema mappings to

specify the semantic relationships between the text in the mediated schema and the data in the sources. Schema mappings, however, may not be trustworthy. Accurate mappings between different data sources in applications are frequently challenging to create and maintain (Dong et al., 2009).

c) *Inconsistencies between data integration and data sources*

In fact, if the data obtained from the sources in a data integration system does not meet the integrity criteria, then there is no global database and query response is useless. When two sources' data contradict one other, this scenario results. This scenario is often handled by applying the proper transformation and cleaning techniques to the data that the sources have acquired(Coelho et al., 2010) (Bouzeghoub, M., & Lenzerini, M. (2001). This paper approaches the issue from a more theoretical standpoint in this section.

Conceptual framework



III. Methodology

In this method, data amalgamation is going to be based on the global schema. We speak about data amalgamation systems, whose goal is to combine data from many sources and give the user a single picture of that data. A representation of this unified view is provided by the global schema, which also provides a reconciled view of all data in a user-query form. It goes without saying that connecting the data sources to the global schema is one of the most crucial aspects of creating a data integration system(Pham et al., 2014). This mapping needs to be appropriately considered when formalizing a data integration system. However, in data amalgamation systems, there are components which are supposed to be used, thus the sources, the mapping, and the global schema itself. Mapping heterogeneous data sources to a unified global schema is one of the most critical aspects of data integration, yet also one of the most complex. The global schema provides an abstracted, integrated view of the sources, enabling users to query across sources as if they were a single repository. However, creating accurate mappings between sources and the global schema poses many challenges; sources often contain overlapping, redundant, or conflicting data representations. For example, two sources may have different definitions or formats for a customer entity. Resolving these discrepancies requires in-depth analysis of the source schemas and data.

Sources are dynamic, changing over time as new data arrives. Keeping mappings synchronized requires ongoing governance. Outdated mappings will propagate errors during integration according to (Y. Chen et al., 2020).

Sources may contain bugs, errors, inconsistencies, or missing values that get propagated

through mappings. Cleansing and transforming source data is usually required.

The global schema offers a cohesive, acquiescent, and simulated representation of the primary sources, whereas the source schema naturally displays the structure of the sources where the real data are situated. The mapping's presumptions show how the components of the global and source schemas relate to one another. The global schema and source schemas serve different purposes in data integration system architecture. The global schema provides a consolidated, integrated view that abstracts the complexity of the sources. It creates a single logical interface that users can query to access data from multiple sources. The global schema structures the data into the forms and relationships needed to support business reporting and analysis. It is optimized for flexibility, performance, and ease of use.

In contrast, source schemas directly reflect the underlying structure and semantics of the original data sources. They model the physical storage, formats, and data elements within each source. Source schemas preserve the quirks and nuances of how each system represents data. They may contain duplicative or overlapping data elements. Source schemas favor accuracy and completeness oversimplification.

These differing purposes lead to key distinctions which are.

Global Schema: Unified view spanning sources, simplified data model, unified semantics, optimized for querying and analysis.

Source schema: System-specific view, matches source storage structure, preserves source peculiarities, optimized for accuracy.

On the other hand, the local as view approach is based on the idea that each source's content should be described from a modeling perspective in terms of a view over the entire schema. At the point when the information incorporation framework is established on a model or a metaphysics, this kind of circumstance is critical (Gruninger, M. 2002). When the data integration system is built on a global schema that is dependable and well-known within the company, this idea works best.

IV. Discussions and Findings

Due to the possibility of interdependence in the data supplied by the Analyzed Database, consistency may not have been attained. However, it is impossible to collect all the data while avoiding null values, this issue has persisted. Accuracy, reliability, availability, timeliness, and other data restrictions specific to each autonomous component database are also addressed. There are several methods that have been used to resolve data base inconsistencies. (Angeles & Mac Kinnon, 2004). The presence of null values in source data can undermine the accuracy and reliability of integrated data sets. Null values typically indicate missing data - facts that should exist but were not captured or stored by the source system. This absence of data leads to incomplete snapshots of business entities, lacking critical attributes needed for analysis. Null values also reduce confidence in the correctness of integrated data. A missing value provides no actual evidence that can be checked or validated. Questions arise over why data is missing and whether the absence itself implies inaccurate representations.(Dong et al., 2009) also says during integration, nulls can introduce ambiguities when linking records across sources. If a key attribute is null, determining matches across sources becomes far more difficult. Nulls also complicate joins and data aggregation. Once integrated, large volumes of nulls make quality assurance and issue diagnosis challenging. Pinpointing the root causes of data gaps requires tracing nulls back to the upstream sources and transformation logic. For certain types of analysis, such as mathematical calculations or machine learning, nulls must be imputed or substituted for proxy values (Agomor & Agomor, 2023). This can introduce estimation errors if not done carefully.

V. Conclusion

In data integration, it may not be possible to reconcile the data collected from the sources with the mapping and limitations of the global schema in a way that's acceptable to both, as defined earlier. For example, this happens when the tuples returned by the view related with connections break a key imperative to the given connection for a worldwide pattern, as the presumption of sound perspectives does not let us to tuples with duplicate keys should be ignored (Lenzerini, 2014). We need a different description of the mapping if we are not to conclude that there is no global database that is appropriate in this context. We specifically need a categorization that allows query processing even when the sources' data is corrupt. does not adhere to the global schema's integrity constraints. Challenges arise from inconsistencies, redundant data, and constraints imposed by the global schema. Organizations must reconcile schema limitations, integrity constraints, and defective source data that violates business rules. Thoughtful planning, extensive validation, iterative enhancements, and continuous data monitoring are imperative for flexible, scalable data integration with minimal disruption. Further innovations in machine learning, automation, and data provenance tracing will aid future integration initiatives.

VI. Recommendations

In light of our findings, we recommend implementing a comprehensive data quality framework

Standardizing data formats, leveraging ETL tools to transform and validate data, defining business rules and integrity checks, implementing master data management, profiling integrated data, automating standardization and cleansing, establishing data governance practices, treating integration as an iterative process, and continually monitoring and enhancing the process based on identified issues are critical ways to address inconsistent data in integration. An iterative approach that focuses on upfront planning, leverage machine learning for pattern recognition and identification of anomalies, validation throughout the pipeline, governance, and continuous improvement will allow organizations to effectively tackle data inconsistencies.

VII. Future Works

In future research, we envision the development of an advanced data quality framework that incorporates machine learning and natural language processing to automatically detect and rectify real-time data quality issues, alongside the exploration of semantic data integration methods for improved alignment of diverse data sources, including unstructured data. To streamline data integration, user-friendly interfaces for mapping management will be created, while robust data governance practices and compliance considerations will be integrated throughout the process. We will also focus on optimizing scalability and performance for large-scale datasets and examine the integration of data residing in cloud environments, emphasizing standardized protocols and exploring the inclusion of machine learning models directly into the data integration process to enhance accuracy and efficiency.

References Références Referencias

1. Agomor, J. E., & Agomor, K. S. (2023). The Effect of COVID-19 on Tertiary Students in Ghana: The Case of the Ghana Institute of Management and Public Administration (GIMPA). *Public Policy*.
2. Angeles, P., & MacKinnon, L. M. (n.d.). *Detection and Resolution of Data Inconsistencies, and Data Integration using Data Quality Criteria*. 8.
3. Chen, M., Ebert, D., Hagen, H., Robert, S., & Van, R. (2009). *Data, Information, and Knowledge in Visualization*.
4. Chen, Y., Avitabile, P., & Dodson, J. (2020). Data Consistency Assessment Function (DCAF). *Mechanical Systems and Signal Processing*, *141*, 106688. https://doi.org/10.1016/j.ymssp.2020.106688.
5. Coelho, P. S., Popovič, A., & Jaklič, J. (2010). The Role of Business Knowledge in Improving Information Quality Provided by Business Intelligence Systems. In J. E. Quintela Varajão, M. M. Cruz-Cunha, G. D. Putnik, & A. Trigo (Eds.), *enterprise Information Systems* (Vol. 110, pp. 148–157). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-16419-4_15.
6. Debattista, J., Lange, C., Scerri, S., & Auer, S. (2015). Linked "Big" Data: Towards a Manifold Increase in Big Data Value and Veracity. *2015 IEEE/ACM 2nd International Symposium on Big Data Computing (BDC)*, 92–98. https://doi.org/10.1109/BDC.2015.34.
7. Dong, X. L., Halevy, A., & Yu, C. (2009). Data integration with uncertainty. *The VLDB Journal*, *18* (2), 469–500. https://doi.org/10.1007/s00778-008-0119-9.
8. Haas, L. M., Lin, E. T., & Roth, M. A. (2002). Data integration through database federation. *IBM Systems Journal*, *41*(4), 578–596. https://doi.org/10.1147/sj.414.0578.
9. Ioannou, E., & Staworko, S. (2013). *Management of Inconsistencies in Data Integration* [Application/pdf]. 9 pages. https://doi.org/10.4230/DFU.VOL5.10452.217.
10. Lenzerini, M. (n.d.). *Data Integration: A Theoretical Perspective*. 14.
11. Lenzerini, M., Salaria, V., & Roma, I.-. (2014). *Data Integration: A Theoretical Perspective Data Integration: A Theoretical Perspective*. *June*. https://doi.org/10.1145/543613.543644.
12. Pham, M. T., Rajć, A., Greig, J. D., Sargeant, J. M., Papadopoulos, A., & McEwen, S. A. (2014). A scoping review of scoping reviews: Advancing the approach and enhancing the consistency. *Research Synthesis Methods*, *5* (4), 371–385. https://doi.org/10.1002/jrsm.1123.
13. Sena, B., Garcés, L., Allian, A. P., & Nakagawa, E. Y. (n.d.). *Investigating the Applicability of Architectural Patterns in Big Data Systems*.
14. Strong, D. M., Lee, Y. W., & Wang, R. Y. (1997). Data quality in context. *Communications of the ACM*, *40* (5), 103–110. https://doi.org/10.1145/253769.253804.

This page is intentionally left blank