THE COMPUTERTECH
(An International Peer Review Journal)
VOLUME 3; ISSUE 2 (JULY-DEC); (2017)
WEBSITE: THE COMPUTERTECH

Data Cleansing Strategies, Enabling Reliable Insights from Big Data

Bharath Kishore Gudepu[1]
[1]Computer Information Systems, University of Central Missouri, 511 S Holden St, Warrensburg, MO 64093

Abstract

Large amount of data available for the organization for which by which the same can influence its business decision. As large amount of data has been collected from distinguish resources and as the data has not been filtered which ultimately affect the accuracy of prediction result. Therefore, data cleansing offers the best quality by which the organization will be in great position to make its decision and the data is absolutely ready for analyzation. As the business is getting bigger and bigger therefore everyone organization data is also enlarged with the passage of time and conventional methods are just futile exercise. However, with the concept of data cleansing the large data can be accurately analyzed as well as shall remove the anomalies etc. Data cleansing is the process of identifying the errors, detecting the errors and make them correct as each department and organization gets fail which they do not have correct data and if the you don't have the correct data you would be not in a position to make a correct data base decision. Thus, most of the originations loses their business position. However, data cleansing is a bit long process and can take sufficient time but is very essential for big data to be correctly analyzed and all the possible anomalies be removed by applying this method.

Keywords: Data Cleansing, Big Data, Data Quality, Reliable Insights, Data Governance, Data Management, Data Profiling, Metadata, Compliance, Data Accuracy, Data Discovery, Analytics, Enterprise Data, Data Integrity, Data Strategy.

Introduction

As the growing enthusiasm for data-driven decision-making has created the importance of accurate and precise prediction over the previous years. The sharp development of the data driven has find out new opportunity for the business and the procedure of analyzing the big data quietly so as to get the essential result. Baldy, if the data handled incorrectly and unreliable the information would lead to a dirty decision. Data cleansing or data cleaning is the process by which the data can be improved by identifying and removing errors and inconsistencies. Insufficient information will cause uncertainties during the data analysis and this should be controlled in the data cleansing stge. Bugs and missing values in the dataset will cause a required result and might affect the business decision too. The data must be accurate to avoid losses, problems and additional cost due to poor quality of data. For example, according to the "Price Waterhouse Coopers" survey conducted in 2001, 75% of 599 companies have suffered losses due to data quality issue. Thereafter, these businesses merely believe on data like customer relationship management and supply chain management, therefore it is so essential for an organization to have better quality of data in order

to gain more precise and useful result. Outstanding data can only be provided by data cleansing as the data collected from the various sources might be dirty. Data quality means to fulfill all the requirements needed for business. It is achieved through people, technology and processes. It ensures compliance and consistency particularly when data from different databases are combined. Without appropriate data best management, even a minor error might cause revenue loss, Thus, data quality and data cleansing always linked together as ensuring data quality is critical and necessary before sharpening of analytic focus can occur. This paper aims at reviewing the available data cleansing methods specifically for big data. Since data cleansing framework needs to meet data quality criteria and fulfill big data characteristics, therefore this paper will identify the data cleansing challenge in big data. Data cleansing methods will be explained in brief along with the weaknesses and strengths of each method [1].

Data cleansing strategies

There are number of strategies by which the big data can be cleaned such as

De-duplication

Duplication means when there are more than one entry made in the system and the same has been repeated in different locations in the system. De-duplication system includes

1. *Identifying duplication:* Sorting out repeated records using advanced techniques like fuzzy matching, which applies machine learning to recognize similar but not identical data entries. Our intelligent system ensures thorough duplicate detection while minimizing false positives.
2. *Merging or Purging Duplicates*: this strategy is used to find out whether to club duplication record into a single or no. it also provides the option whether to remove the entry completely or make accurate entry if required so. Our sophisticated merging algorithm preserves the most reliable data while eliminating redundancy.
3. *Error detection and correction*: this is another key element of data cleansing it ensure to find out the error and correct it accordingly. It includes the key steps such as spotting anomalies and correcting errors. Spot unusual data patterns, such as extreme outliers or conflicting values, using advanced algorithms that analyze trends and flag inconsistencies for further review. Adjust misspellings, correct formatting inconsistencies, and resolve numerical discrepancies to improve data accuracy.
4. *Standardizing of data*: This formats ensures consistency across different systems and datasets, making it easier to analyze and integrate. This is particularly crucial for structured fields like dates, phone numbers, and addresses, where variations can be confusing. Convert diverse data formats into a consistent structure, such as ensuring all phone numbers include country codes or all dates follow the same. Align data values to a standard reference, such as converting all monetary values into a single currency or ensuring measurements use the same unit.

5. *Missing Data Handling.* Incomplete datasets can lead to inaccurate analysis and decision-making. Addressing missing data requires strategies to either estimate missing values or mark incomplete records for further action. Key options include:

Data Imputation: Use statistical techniques to estimate and fill in missing values based on historical data and contextual clues [2].

Removing or Flagging Data: Determine whether to delete records with substantial missing information or mark them for follow-up and review.

6. *Data Enrichment*

Enhancing raw datasets with additional information improves their value and depth. Organizations can gain a more comprehensive view of customers, products, or business operations by incorporating external or supplemental data. Key strategies include:

Completing Missing Information: Fill in gaps by appending relevant details, such as completing addresses with missing ZIP codes [3].

Integrating External Sources: Integrate third-party data, such as demographic insights or geographic details, to provide more context and improve analysis.

7. *Data Parsing and Transformation*

Raw data is often unstructured and difficult to analyze. Parsing and transformation techniques refine and organize this data, making it more accessible and useful for business intelligence and reporting [4].

Data Parsing

Break down complex text strings into distinct elements, such as extracting a full name into separate first and last name fields.

Data Transformation

Convert data from one format (e.g., Excel spreadsheet) to another, ensuring it is ready for use.

Challenges in Data Cleansing for Big Data

Data cleansing for big data comes with unique challenges that stem from the scale, diversity, and velocity of data generation. Some of the key challenges include:

Volume: Big data sets often consist of millions or even billions of records, making it challenging to process and cleanse them efficiently. Traditional data cleansing methods may need help to handle such massive volumes of data effectively [5].

Variety: Big data is characterized by its diverse sources, formats, and structures. Data may originate from structured databases, semi-structured formats like JSON or XML, or unstructured sources like text documents and social media feeds. Cleaning such varied data types requires versatile tools capable of handling different data formats [6].

Velocity: Big data is generated at an unparalleled speed, with constant data streams flowing in from various sources in real-time. The rapid pace of data creation complicates the data cleansing process, as organizations must cleanse and analyze data on the fly to derive timely insights.

Quality: Maintaining data quality is paramount in big data environments. However, data sources' sheer volume and diversity often result in poor data quality, including missing values, inconsistencies, and inaccuracies. Identifying and rectifying these issues without compromising processing speed is a significant challenge [7].

Scalability: Traditional data cleansing tools may need more scalability to handle big data effectively. Organizations need scalable solutions capable of processing and cleansing data across distributed computing environments as data volumes grow.

Addressing these challenges requires advanced data cleansing tools with machine learning algorithms, parallel processing capabilities, and distributed computing frameworks. By overcoming these hurdles, organizations can ensure their ample data assets' accuracy, consistency, and reliability, paving the way for data-driven insights and decision-making [8].

Solutions and Best Practices

In addressing the challenges of data cleansing for big data, several solutions and best practices can help organizations streamline their processes and ensure data accuracy:

Automated Data Quality Checks: Implement automatic data quality checks at various data pipeline stages to detect and correct errors early on. This includes validating data formats, identifying missing values, and flagging outliers [9].

Standardization and Normalization: Standardize data formats and values across different sources to ensure consistency and compatibility. Normalizing data can reduce redundancy and improve data integrity.

Deduplication: Identify and remove duplicate records from datasets to prevent data redundancy and maintain data accuracy. Utilize algorithms and techniques, such as fuzzy matching, to identify similar records for deduplication.

Data Profiling: Conduct thorough data profiling to understand the dataset's structure, quality, and relationships. Data profiling helps identify anomalies, outliers, and inconsistencies that require cleansing.

Scalable Infrastructure: Invest in scalable infrastructure and technologies that can manage the volume and velocity of big data. Distributed computing frameworks like Apache Spark and Hadoop enable parallel processing of large datasets, facilitating efficient data cleansing.

Data Governance Framework: Establish a robust framework that defines policies, processes, and responsibilities for managing data quality. Implement data stewardship roles and workflows to ensure accountability and ownership of data quality issues.

Continuous Monitoring and Improvement: Implement mechanisms for constantly monitoring data quality metrics and performance indicators. Regularly evaluate and refine data cleansing processes based on feedback and evolving business requirements.

User Training and Education: Provide training and education to users involved in the data cleansing process to ensure they understand best practices, tools, and techniques. Foster a culture of data quality awareness and collaboration across the organization.

Discussion

To discuss the main ideas of data cleansing. It is the most essential and power of tool to be sued by the bigger organization for the best and precise data as it provides the best and quality data of the business organization so that they can take positive decision based on the data cleansing. It is the modernized era of business and most of the organization relies upon the computerized data as well as data which has been already analyzed. Therefore, data cleansing offers the best quality by which the organization will be in great position to make its decision and the data is absolutely ready for analyzation. As the business is getting bigger and bigger therefore everyone organization data is also enlarged with the passage of time and conventional methods are just futile exercise. However, with the concept of data cleansing the large data can be accurately analyzed as well as shall remove the anomalies etc. Data cleansing is the process of identifying the errors, detecting the errors and make them correct as every department and organization gets final which they do not have correct data and if the you don't have the correct data you would be not in a position to make a correct data base decision. Thus, most of the originations loses their business position. Data cleansing is one of the most brilliant technique by which the world class business organizations can get their desire result [10-13].

Conclusion

To sum off all the perspective of data cleansing it has been proved that data cleansing has enough role in business organizations so that they can take good decision as well as would have accurate data by which they get better decision and could have fire wall for business losses. This method also provides the plan by which the organization can get their business activities at smooth pace and shall have chance to plan their businesses as per their requirement. The sharp development of the data driven has find out new opportunity for the business and the procedure of analyzing the big data quietly so as to get the essential result. Baldy, if the data handled incorrectly and unreliable the information would lead to a dirty decision. Data cleansing or data cleaning is the process by which the data can be improved by identifying and removing errors and inconsistencies. Insufficient information will cause uncertainties during the data analysis and this should be controlled in the data cleansing stge. Bugs and missing values in the dataset will cause a required result and might affect the business decision too. This method also determines the future plan for a bigger organization so that accurate data would help them to sustain their business. This formats ensures consistency across different systems and datasets, making it easier to analyze and integrate. This is particularly crucial for structured fields like dates, phone numbers, and addresses, where variations can be confusing. Convert diverse data formats into a consistent structure, such as ensuring all phone numbers include country codes or all dates follow the same. Align data values to a standard reference, such as converting all monetary values into a single currency or ensuring measurements use the same unit.

References:

[1] Alam, K., Mostakim, M. A., & Khan, M. S. I. (2017). Design and Optimization of MicroSolar Grid for Off-Grid Rural Communities. Distributed Learning and Broad Applications in Scientific Research, 3.

[2] Agarwal, A. V., & Kumar, S. (2017, November). Unsupervised data responsive based monitoring of fields. In 2017 International Conference on Inventive Computing and Informatics (ICICI) (pp. 184-188). IEEE.

[3] Mishra, M. (2017). Reliability-based Life Cycle Management of Corroding Pipelines via Optimization under Uncertainty (Doctoral dissertation).

[4] Agarwal, A. V., & Kumar, S. (2017, October). Intelligent multi-level mechanism of secure data handling of vehicular information for post-accident protocols. In 2017 2nd International Conference on Communication and Electronics Systems (ICCES) (pp. 902-906). IEEE.

[5] Ramadugu, R., & Doddipatla, L. (2022). Emerging Trends in Fintech: How Technology Is Reshaping the Global Financial Landscape. Journal of Computational Innovation, 2(1).

[6] Malhotra, I., Gopinath, S., Janga, K. C., Greenberg, S., Sharma, S. K., & Tarkovsky, R. (2014). Unpredictable nature of tolvaptan in treatment of hypervolemic hyponatremia: case review on role of vaptans. Case reports in endocrinology, 2014(1), 807054.

[7] Shakibaie-M, B. (2013). Comparison of the effectiveness of two different bone substitute materials for socket preservation after tooth extraction: a controlled clinical study. International Journal of Periodontics & Restorative Dentistry, 33(2).

[8] Gopinath, S., Janga, K. C., Greenberg, S., & Sharma, S. K. (2013). Tolvaptan in the treatment of acute hyponatremia associated with acute kidney injury. Case reports in nephrology, 2013(1), 801575.

[9] Shilpa, Lalitha, Prakash, A., & Rao, S. (2009). BFHI in a tertiary care hospital: Does being Baby friendly affect lactation success?. The Indian Journal of Pediatrics, 76, 655-657.

[10] Singh, V. K., Mishra, A., Gupta, K. K., Misra, R., & Patel, M. L. (2015). Reduction of microalbuminuria in type-2 diabetes mellitus with angiotensin-converting enzyme inhibitor alone and with cilnidipine. Indian Journal of Nephrology, 25(6), 334-339.

[11] Gopinath, S., Giambarberi, L., Patil, S., & Chamberlain, R. S. (2016). Characteristics and survival of patients with eccrine carcinoma: a cohort study. Journal of the American Academy of Dermatology, 75(1), 215-217.

[12] Swarnagowri, B. N., & Gopinath, S. (2013). Ambiguity in diagnosing esthesioneuroblastoma--a case report. Journal of Evolution of Medical and Dental Sciences, 2(43), 8251-8255.

[13] Swarnagowri, B. N., & Gopinath, S. (2013). Pelvic Actinomycosis Mimicking Malignancy: A Case Report. tuberculosis, 14, 15.