THE COMPUTERTECH
(An International Peer Review Journal)
VOLUME 4; ISSUE 1(JAN-JUNE); (2018)
WEBSITE: THE COMPUTERTECH

The Role of Data Profiling in Improving Data Quality

Bharath Kishore Gudepu[1], Divya Sai Jaladi[2]

[1]Senior Informatica Developer, Transamerica, 10100 N Central Expy Ste 595, Dallas, TX 75231

[2]Senior Lead Application Developer, SCDMV, 10311 Wilson Boulevard, Blythewood, SC 29016, UNITED STATES

Abstract

Data health pertains to the comprehensive quality, usefulness, and value of data inside an organization. Inadequately handled data can obstruct decision-making, squander resources, and result in lost opportunities. Data profiling is the essential initial step for evaluating and enhancing data quality. It entails analyzing the structure, linkages, and characteristics of data to detect inconsistencies, redundancies, or inaccuracies. Profiling systematically organizes and sanitizes data, enhancing its usability. Research indicates that just 3% of data adheres to quality criteria. This is a remarkable figure that underscores the difficulties businesses have owing to faulty data. Erroneous or insufficient data may lead to misguided judgments or redundant efforts. The time allocated to seeking, verifying, and rectifying data detracts from productive endeavors. Unexploited insights inside raw data hinder firms from achieving competitive advantages or fostering successful innovation. Robust data is accessible: It is readily identifiable and accessible to people in need. It is conveyed in a coherent and structured format, facilitating straightforward interpretation of its content. It fulfills a function, whether for analysis, decision-making, or operational enhancement. Improve data integrity by identifying and rectifying mistakes such as duplication, inconsistencies, or absent values. Guarantee adherence to data rules (e.g., GDPR) by appropriate structure and transparency. Analyze datasets to discern patterns and trends, transforming raw data into valuable business insight. Enhance cooperation by increasing data accessibility for team members across all departments. Ultimately, profiling converts raw data into a strategic resource, facilitating informed decision-making and fostering creativity. Upon profiling and organizing data, it may be assimilated into analytical procedures. By employing technologies such as dashboards or visualizations, enterprises acquire insights into their operations, market trends, or consumer behavior.

Keywords: Business Metadata, Decision-Making, Data Governance, Data Management, Data Quality, Metadata Management, Compliance, Data Profiling, Analytics, Enterprise Data, Data Discovery, Data Integrity, Business Intelligence, Data Strategy, Big Data.

Introduction

Data profiling entails the examination, analysis, and generation of informative summaries of data. The approach provides a comprehensive picture that facilitates the identification of data quality concerns, hazards, and overarching patterns. Data profiling offers essential insights into data that organizations may subsequently utilize to their benefit.

Data profiling meticulously analyzes data to ascertain its authenticity and quality. Analytical algorithms identify dataset attributes like mean, minimum, maximum, percentile, and frequency to scrutinize data meticulously. It subsequently conducts analysis to reveal metadata, encompassing frequency distributions, principal connections, potential foreign key candidates, and functional dependencies. Ultimately, it utilizes this information to reveal how these elements correspond with your business's standards and objectives. Data profiling may eradicate expensive inaccuracies prevalent in client datasets. These mistakes encompass null values (unknown or absent values), inappropriate values, values exhibiting anomalously high or low frequency, values deviating from anticipated patterns, and values beyond the typical range [1-3].

The Strategy for Data Profiling

Historically, attaining market leadership mostly relied on providing the appropriate goods at the optimal moment. Nonetheless, the industrial and technological revolutions revolutionized the marketplace. Organizations commenced the production of analogous items, necessitating market leaders to innovate by creating goods that were more cost-effective, superior, and expedited. This progress, along with diminished entry barriers, inundated the market with rivals providing indistinguishable items. As a result, market leaders declined, facilitating the emergence of a commodity-driven economy [4-7].

In an environment characterized by narrow profit margins and fierce competition, corporations acknowledged that a singularly outstanding product no longer ensured success. In the past decade, emphasis has transitioned to process optimization as an essential approach for development. In the current economy, profitability relies equally on managing costs and producing income. To realize significant cost reductions, enterprises have used two primary applications: ERP (Enterprise Resource Planning) and CRM (Customer Relationship Management).ERP systems seek to optimize operational operations and reduce processing costs, providing effective expenditure management. Conversely, CRM systems emphasize the development of lucrative client relationships, tackling the substantial expenses linked to customer acquisition and retention. In conjunction with these tools, corporations have used data warehouses to enhance strategic decision-making, optimize expenditures, and maximize savings throughout the firm. The foundation of ERP and CRM systems is the data that drives their operations. Accurate data on inventories, suppliers, customers, vendors, and other organizational components is crucial for success. In the absence of precise and trustworthy information, these implementations are prone to failure, as they are fundamentally dependent on the concept of "garbage-in, garbage-out."

Types of data profiling

There are three primary categories of data profiling:

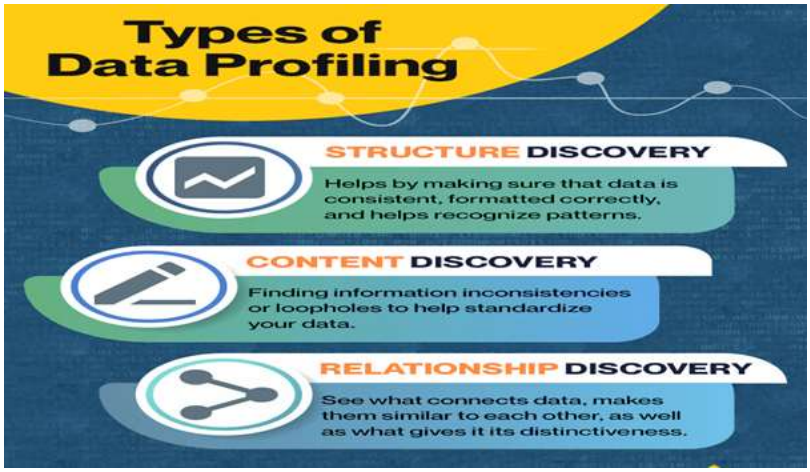
Discovery of Structure

Verifying that data is coherent and properly structured, and doing mathematical validations on the data (e.g., summation, minimum, or maximum values). Structure discovery aids in assessing the organization of data, such as the proportion of phone numbers that lack the requisite amount of digits.

Examining individual data records to identify inaccuracies. Content discovery determines the precise rows in a table that exhibit flaws and finds systematic faults within the data, such as phone numbers without an area code.

Exploration of Relationships

Identifying the interconnections among data components. For instance, essential connections between database tables and references among cells or tables within a spreadsheet. Comprehending linkages is essential for data reutilization; interconnected data sources must be consolidated or imported in a manner that maintains significant relationships [8-11].



Advantages And Disadvantages Of Data Profiling

In general, there are minimal to no drawbacks associated with data profiling. Having a substantial quantity of data is one aspect, but the quality is paramount, which is where data profiling becomes essential. Standardized data that is meticulously prepared minimizes the likelihood of dissatisfied clients or misinterpretation.

The obstacles are mostly systemic, since disaggregated data complicates the process of retrieval. However, the use of certain data tools and apps should mitigate such issues and can only enhance a company's decision-making process. Let us examine further significant advantages and obstacles.

Benefits

Data profiling provides a unique high-level picture of data that surpasses other tools. Specifically, you can anticipate:

Enhanced Analytical Precision: Comprehensive data profiling will guarantee enhanced quality and increased data credibility. Accurately profiling your data helps enhance the understanding of the relationships across various data types and sources, hence facilitating data governance protocols.

Maintains Centralized Information: Through data profiling, the examination and analysis of your data will significantly enhance its quality and organization. The examination of source data will rectify inaccuracies and identify the regions with the most significant problems. It will thereafter generate insights and organizing that optimally centralizes your data.

The Function of Data Profiling

Data profiling guarantees that the data is robust and suitable for its designated use. This is really the initial phase in the management and utilization of data. Data profiling can reveal many data quality concerns, including absent data, duplication, and errors. It also emphasizes patterns, regulations, and trends within the data. This information is essential as it enables firms to enhance data quality, optimize data transformation, and facilitate informed decision-making. Obstaclesp challenges in data profiling generally arise from the intricacy of the tasks involved. Specifically, you can anticipate:

Costly and labor-intensive: Data profiling may become very intricate when attempting to execute a successful program, partly due to the substantial number of data amassed by a normal firm. Hiring qualified specialists to examine results and make judgments without the appropriate tools may be both costly and time-consuming. Insufficient resources: To begin the data profiling process, a corporation need all its data consolidated in a single location, which is frequently not achieved. When data is dispersed across many departments and a qualified data professional is absent, it becomes exceedingly challenging to conduct a comprehensive data profile of the organization [12-15].

Procedure of Data Profiling

Phases of Data Profiling Data profiling is a systematic method designed to comprehend and enhance the quality of gathered information. The phases encompass:

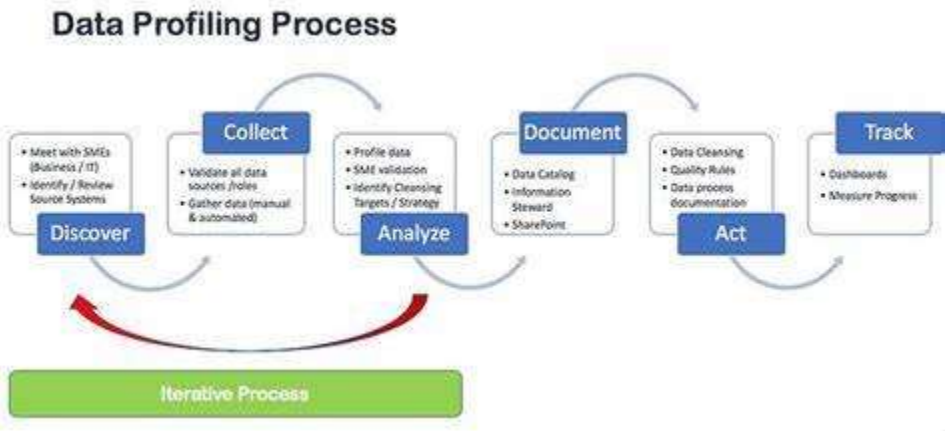
Data Discovery: Initiate the collection of data from diverse sources, including website visitor metrics, click counts, and session lengths.

Assessing Data Integrity: Evaluate the precision and comprehensiveness of the gathered data to detect inaccuracies, omissions, or discrepancies.

Data Analysis: Examine the specifics to identify patterns, trends, and correlations among data points for enhanced understanding.

Cataloging Insights: Maintain a record of findings and analysis to enhance dissemination and comprehension.

Improving Data Quality: Rectify detected issues by fixing inaccuracies, addressing deficiencies, or acquiring more data [16-18].



Discussion

The enhancement in imaging throughput, novel analytical frameworks, and advanced computer resources create new opportunities for data-intensive phenotypic profiling of small compounds in drug development. Image-based profiling assays evaluating single-cell characteristics have been employed to investigate mechanisms of action, target effectiveness, and toxicity of small compounds. Technological advancements in generating extensive data sets, along with novel machine learning methodologies for analyzing high-dimensional profile data, present prospects to enhance many stages of drug discovery. This paper examines the application of machine learning methodologies in functional profiling procedures, emphasizing chemical genetics, as demonstrated by recent investigations. Although their usefulness in image-based screening and profiling is clearly apparent, instances of innovative discoveries that transcend conventional understanding through the application of machine learning techniques are just now starting to surface. Future research must provide approaches that reduce the entry barriers to high-throughput profiling experiments by optimizing image-based profiling assays and facilitating the use of advanced learning technologies, such as readily deployable deep neural networks.

Conclusion

This study examines the significant impact of data quality on data analytics, emphasizing essential factors including correctness, completeness, consistency, dependability, and timeliness. Addressing these characteristics is crucial for realizing the whole potential of analytics tools and approaches. The paper emphasizes problems such as data integration, complications in purification, and the evolution of data sources, while proposing best practices like data profiling, cleansing, and standardization to address discrepancies. Empirical case studies highlight the correlation between superior data quality and the efficacy of analytics, illustrating advantages such as augmented decision-making, elevated consumer happiness, and optimized operations. In contrast, the dangers of inadequate data quality—defective tactics, inaccurate models, and monetary losses—are highlighted. The study promotes a proactive strategy, encouraging firms to invest in comprehensive data governance, sophisticated technologies, and proficient individuals to ensure consistent data

quality. By doing so, enterprises may foster innovation, secure a competitive advantage, and attain sustainable growth.

References

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

[2] Agarwal, A. V., Verma, N., & Kumar, S. (2018). Intelligent Decision Making Real-Time Automated System for Toll Payments. In Proceedings of International Conference on Recent Advancement on Computer and Communication: ICRAC 2017 (pp. 223-232). Springer Singapore.

[3] Pasham, S.D. (2017) AI-Driven Cloud Cost Optimization for Small and Medium Enterprises (SMEs). The Computertech. 1-24.

[4] Chen, D., & Zhao, H. (2012). Data security and privacy protection issues in cloud computing. International Conference on Computer Science and Electronics Engineering, 647-651.

[5] Garg, P., Verma, D., & Kaushal, V. (2018). A study on data migration techniques for cloud computing. International Journal of Advanced Research in Computer Science, 9(1), 45-52.

[6] Sai, K.M.V., M. Ramineni, M.V. Chowdary, and L. Deepthi. Data Hiding Scheme in Quad Channel Images using Square Block Algorithm. in 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI). 2018. IEEE.

[7] Pasham, S.D. (2018) Dynamic Resource Provisioning in Cloud Environments Using Predictive Analytics. The Computertech. 1-28.

[8] Ahmed, T., & Smith, M. (2018). Cloud data migration: Challenges, solutions, and future directions. Journal of Cloud Computing, 7, 12-29.

[9] Tallon, P. (2013). Corporate data migration strategies: Managing risks and maximizing benefits. MIS Quarterly, 37(4), 1125-1147.

[10] Grolinger, K., Higashino, W. A., Tiwari, A., & Capretz, M. A. M. (2013). Data management in cloud environments: NoSQL and NewSQL data stores. Journal of Cloud Computing: Advances, Systems and Applications, 2(1), 1-24.

[11] Inmon, W. H. (2005). Building the data warehouse (4th ed.). Wiley.

[12] Khine, P. P., & Wang, Z. (2018). Data lake: A new ideology in big data era. Proceedings of the 2018 IEEE 6th International Conference on Future Internet of Things and Cloud Workshops, 37-42.

[13] Kimball, R., & Ross, M. (2013). The data warehouse toolkit: The definitive guide to dimensional modeling (3rd ed.). Wiley.

[14] Dageville, B.,andDias, K. (2006). Oracle's Self-Tuning Architecture and Solutions. IEEE Data Eng. Bull., 29(3), 24-31

[15] 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.

[16] 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).

[17] 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.

[18] 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.