International Journal for Research Publication and Seminar
ISSN: 2278-6848 | Vol. 15 | Issue 4 | Oct-Dec 2024 | Peer Reviewed & Refereed

Enhancing Data Quality through Automated Data Profiling

Nandish Shivaprasad
Independent Researcher, USA.

Abstract
Data profiling is now a popular solution for automating data accuracy and data quality and is characterized by increased reliability of datasets. This paper briefly discusses the difficulties of achieving high data quality, the importance of automation in overcoming these difficulties, and the methods and procedures of data profiling. Automated profiling through data validation thus leads to improved decision making, especially through the unearthing of gaps and contradiction as well as supporting data management as a critical component of compliance. The paper also demonstrates through the use of interesting case examples and illustrating applications how profiling can open up the full utility of organisational data resources.

Keywords
Data quality, Auto data profiling, data management, anomaly detection

Introduction

Due to the current incorporation of data in organizational processes and systems, the accuracy or the reliability of such data greatly influence the decision, strategy and innovation of the organization. The problems that result from poor data quality are that wrong decisions will be made, that many processes will not run properly, and that regulations can be violated.

Automated data profiling solves this problem by making it possible for organizations to understand the structure and content of a data set hence noting the problems that need to be fixed. This paper considers the possibilities of automated data profiling, its advantages, and the tools and techniques in automated data profiling with an emphasis on data quality. Due to recent advances in automation and novel technologies, data profiling stands as the fundamental workhorse for solid data governance and high-quality outcomes of data-driven approaches.

Data Quality Challenges

The issues of data quality are as omnipresent as the use of data in decision making, company operations and, in general, the overall business performance. In this age of big data, correct, comparable, and credible information is crucial in managing and supporting key operations in organizations. However, managing quality of data poses great challenges due to variance in data source, data format, and data volume.

One of the main issues that need be addressed is data missing which misleads the results of certain calculations. Data that is missing may be due to something as simple as data not being entered fully, to difficulties in system compatibility. This in its turn results in gaps which disrupt the quality of datasets and their relevance for analytics or decision-making. However, the format and structure of the data can become the aggravating factor in quality problems.

When data collected from different sources are compiled together there are common issues of standardization and compatibility these include issues with field names, record formats or classification rules. Such a model is less consistent, and it becomes difficult to make further analysis or gain insights since a lot of efforts may be needed to reconcile such data. Another vital challenge is that of duplicate records, which again go to the heart of any data quality framework.

There may be scenes where similar entries are not identified because the systems do not recognize the duplicity because of the difference in the manner in which information is keyed. Redundancy of data is not healthy because when two or three departments have the same data, this creates traffic and also processing workload, all of which is costly (Ehrlinger & Wöß, 2022). For instance, in customer relationship management (CRM) systems, data duplication leads to customers' records splitting thus if

a business intends to personalize servicing a certain customer or even measuring the behavior of the client, it is going to encounter significant challenges.

Old or stale data is also an issue that hits the quality of products or services offered heavily. Unfortunately, when data is old and irrelevant, it persists in databases, giving analyses noise that can hide trends and future projections. Managers who use data that is old and stale stand to lose a great deal since they base their decision on information that is no longer accurate to the firm's environment.

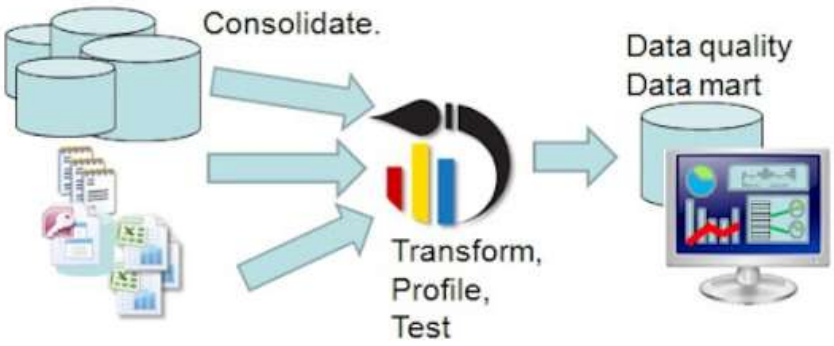

Figure 1 Automating data profiling (Datamartist.com, 2023)

Data accuracy itself is also another core issue challenge Since the accuracy of various data is very important for a company's performance, there is enough evidence that show that it is very difficult to achieve this on a consistent level. Wrong information as a result of errors made by people, technology breakdown, or, inefficient knowledge flow, is dangerous.

For instance, improper numbers in the financial sector may lead to fines with the healthcare sector; a numerical inaccuracy may be a threat to patients' lives. It is therefore quite important to maintain accuracy especially in these large complex databases and this means going through validation procedures which if it was done manually, they would take a lot of both money and time. However, the amount of data produced today presents its own problem in terms of volume with regards to data big data.

Managing and processing big data became a major challenge that organisations have to deal with. Conventional practices either to store or process the corresponding quantities of data are inadequate in the context of AI, resulting in clogs that negativize quality. It also raises challenges, given the complexity of the data ecosystems, in the current world. Different technologies and platforms are implemented in organizations and as they grow, their data environments are heterogeneous. An implication of this is that data is segmented into silos, which hampers its use across sectors which is its strength.

In addition, real-time data integration emerges as a challenge because the activities of syncing data between systems in real time call for enhanced infrastructural investment and sophisticated systems. Accompanying this remains the problem of the protection of the data and their admissibility. Laws such as GDPR and CCpa have compelled organizations to find that thin line between effectiveness of data and data privacy.

Anyone who violates security measures, or is not compliant, weakens the trust required and can lead to financial and reputational losses (Taleb et al., 2021). One could thus not downplay the task of updating the data standards as well as the requirements hence the need to embrace the change. Over the lifespan of industries, criteria of 'high quality' of data also emerge, shift and transform.

This means that organisations need to constantly bring into alignment their working practices to these standards in a process that may involve rewiring data governance arrangements. This element in turn requires staff training, adoption of further technical developments, and staff commitment to ensure high data quality. Solving these problems demands a systematic approach to enhance technologies, increase the reliability of the processes, and develop a vision on data as a key resource.

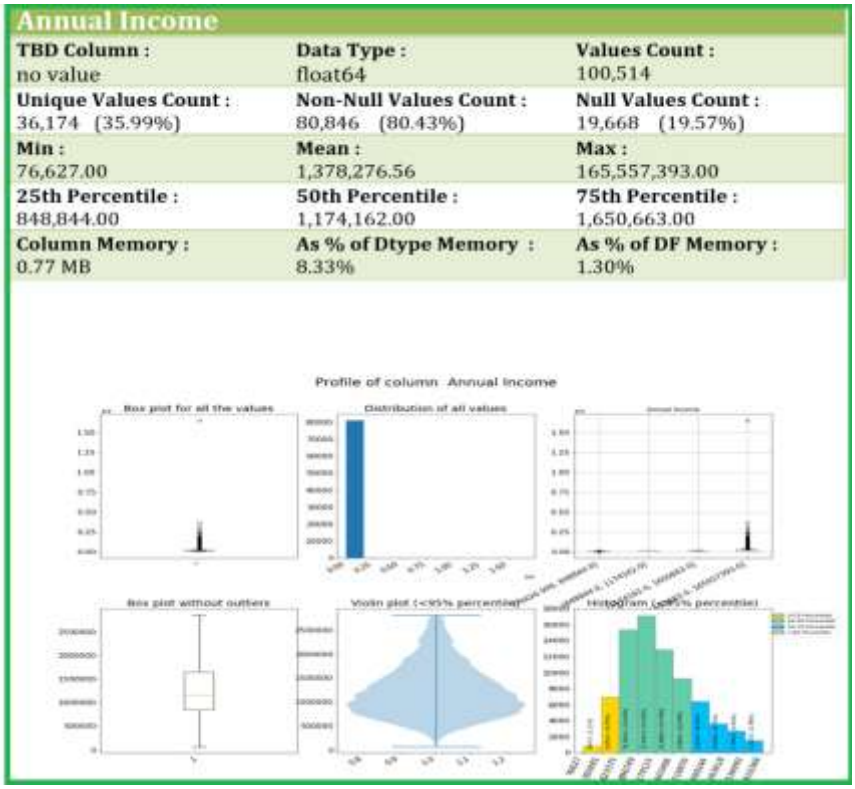

Figure 2 Automated Data Profiling Using Python (Towards Data Science, 2023)

Automated Data Profiling

Data profiling is an imperative function in today's data management involving algorithms and technologies used to analyse data sets for patterns. Data auditing is a crucial activity in achieving the quality of data, to provide an organized approach to meet challenges that may affect an organization's decision-making on data.

Automated data profiling is different from traditional methods of data profiling using paper and pencil which is time taking and can involve human errors. This automation helps organizations to work with ever growing levels of complexity with the data, environments that manual methods cannot manage.

In its basic form, automated data profiling means the process of drifting through datasets to produce metadata at varying levels of granularity, which include but is not limited to summaries of data descriptive information, statistical summary of the data, relationships between fields, and any other anomalies that may be present in the data. This metadata gives a summary of the data and hence; is very useful when a user does not want to deal with all records personally.

There are several activities that are within the sphere of automated data profiling; one of the most important of them is the identification of missing or incomplete values (Jakubik et al., 2024). Contact details and professions usually have many missing values in datasets and profiling tools are capable of identifying these fields, estimate the probability of missing entries and provide recommendations to the problem. Blank spaces are a common problem that can either contribute to methodological errors or cause skewed conclusions which mostly in the fields like health or finance.

While the presence of missing data problems is detected with automated data profiling it also presents patterns such as is any field always missing or is missing data conditional on any conditions. The same applies for automated data profiling as it also focuses on finding out irregularities across data sets.

Anomalies like changing in font style, formatting or differences in data in the entry usually occur after compiling data from different sources. Profiling tools filter such discrepancies based on set rules, or in others, Machine Learning algorithms, to help ensure the dataset is normalized, or brought into line with a proper format. The third important feature of automated data profiling also pertains to different methods of linking data throughout data sets.

Almost all datasets contain related variables that must satisfy certain requirements or constraints, for example, referential integrity within a database table. These relationships are then further evaluated by automated profiling tools for violations of constraints, or the presence of potentially error-prone unnatural correlations.

In this way, data profiling helps data governance and compliance processes to reach the goal of compliance with data regulatory rules and organizational needs. For example, profiling tools can ensure that data correspond to certain standards as defined by certain sectors such as GDPR or HIPAA, thereby, giving companies the assurance they need to use their data in audits, reporting or predictive analytics.

Automated data profiling is also of significant importance in regard to the identification of data distributions. Profiling tools can find values that occur more often or less frequently than the rest, or values that are spread out in a particular way within a dataset or find that there might be points that distort a result or there may be other hidden things that might justify further scrutiny or necropsy. For instance, in a retail setting, profiling could show that certain regions or product types recorded high sales values other than was being anticipated.

This capability is especially useful in data-driven prediction, as the plausibility of such models is influenced by input data quality and relevancy. Automated profiling makes it easier to retain only customizable sets with no irregularity that may affect the reliability of the observed predictions. Automation of data profiling for big data therefore affords big data environments immense values based on scalability.

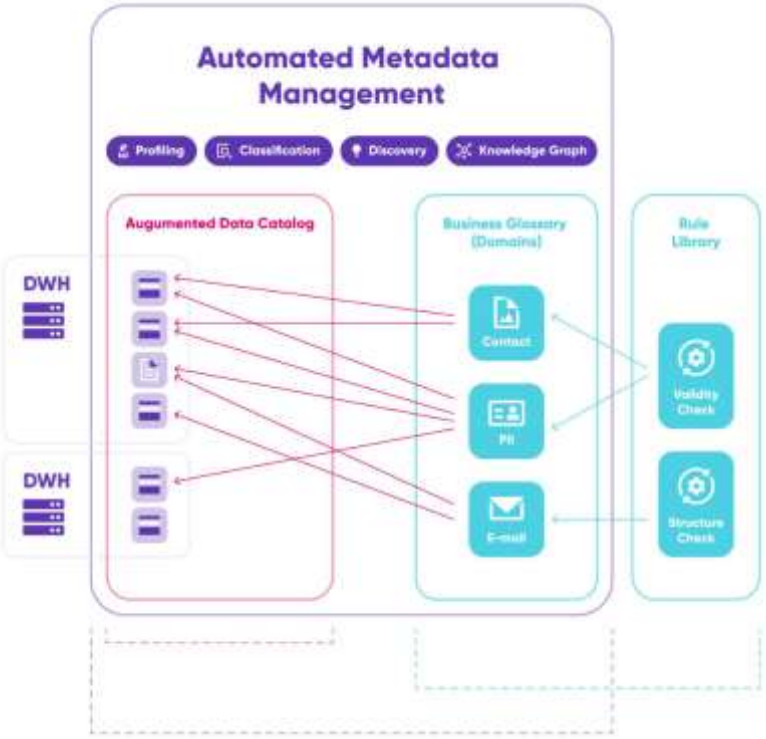

Figure 3 Automated data quality monitoring for data observability (Ataccama, 2023)

With organizations including a plethora of IoT device-generated data and data from social media and other transactional systems, the demand for software tools that can address the complexity that is observed has grown (Radivojević et al., 2020). Data profiling tools are expected to accommodate such demands, and perform profiling on millions of records in the shortest time possible compared to manual work.

It also applies to profiling integration with data pipelines which is the ability of tools to profile data while in transit. It also enables organizations to assure data quality at each step of the process since the profiling of data is possible in real-time.

However, it is noteworthy that even automated data profiling has some issues. Profiling tools might take up a large part of IT investment when being introduced with corresponding training taking their fair share as well. Organizations then have to guarantee that the people who are working on it know how to use it and analyse the findings correctly.

Furthermore, automated profiling is highly dependent on the effectiveness of the rules and algorithms used in the process. Lack of well-defined rules may result in over identification of problems or failure to detect problems when they exist in the profiling process. However, alongside the positive effects that automation has in terms of cutting short the time taken in performing manual activities, it has its flaws

in that it cannot independently draw conclusion on the need for using pull instead of push techniques in an organization.

Certain aspects of data quality are limited, including context and business rules, cannot be determined by the technical validation of data quality measures. It is therefore relevant to find the means and ways of achieving an optimal level of automation, which coupled with human supervision would be beneficial to data profiling practices (Jang et al., 2019). One more factor is the moral and legal issues connected with the usage of data profiling provided by automation.

Many profiling tools analyse personal information that is often in violation of standards in data protection and privacy. Profiling must be done ethically and legally compliant, whereby organisational leaders develop measures that prevent the leakage of data to the wrong parties.

For instance, different levels of anonymization may be used to point data that does not reveal personal information allowing best analyses and avoiding privacy threats at the same time. Actually, the utilization of automated data profiling in the future is most successfully implemented with the help of artificial intelligence and machine learning. These technologies supplement profiling in ways that allow tools to continually refine their processes and outputs.

The profiling tools should be constantly and easily updated to accommodate new structures: machine learning algorithms can learn about new structures in the data and find new problems the previous part did not. Moreover, the inclusion of the natural language processing enables profiling tools to perform profiling on unstructured data in form of text, audio or video among others. Using these technologies, automated data profiling will occupy an ever-increasing importance in managing data assets and realizing their value while ensuring data quality.

Techniques and Tools for Automated Data Profiling

Data profiling itself and techniques/ tools for automated data profiling are an innovative area of technology and method focusing on the issues of data quality. These solutions have become increasingly sophisticated to address the modern data complexity provided with excellent accuracy and performance. Automated data profiling is about identifying metadata that defines the structure, content; and quality of a data set.

Methods used for this purpose are intended to provide maximal coverage of informational content, from elementary mathematical means of summarizing information to artificial neural networks giving an opportunity to investigate findings beyond the observation level. Applications of these techniques include diverse tools that support the management, validation, and optimization of the data in different environments and uses within an organization.

The first and probably one of the most basic approaches to automated data profiling is descriptive statistics (Articles & Articles, 2024). This includes raw score measurements like the averages-mean, mode, standard deviation, and distribution of number and categorical data. These serve the purpose of giving a first impression of the summary statistics of the dataset and their dispersion.

Furthermore, common profiling tools comprise assertions regarding data shape, including measures like skewness and kurtosis that may reveal the presence of certain forms of systematic bias or peculiarity within distributions. Profiling tools are not limited to univariate methods but go up to bivariate and multivariate that examines relationships between values.

Figure 4 Data profiling charts (Collibra, 2023)

These capabilities include correlation analysis, covariance matrices, and chi-square tests enable tools to detect the dependency or interaction of attributes to provide essential relational integrity and data consistency checks. Analysis of format anomalies and irregularities is another of the crucial skills inherent to automated data profiling, and pattern recognition is critical to this work.

Patterns and predefine pattern libraries are employed often to check if the data being entered is in line with required formats like email format, phone numbers format or even Social Security numbers. For example, in entries that are analysed numerically, the tool can underline components that do not correspond to these patterns and may suggest correction.

Machine learning algorithms have improved the ability of profiling systems to recognize patterns, and to develop methods and heuristics for recognizing new patterns of data that may not have been considered in the creation of systems. In the same way, missing values could be imputed or the data type of a certain value can be determined using an AI to enhance the reliability and adaptability of profiling mechanisms.

Profile enhanced tools are complemented in turn by anomaly detection that helps to define the values that differ from norms. Such algorithms work with statistical thresholds, clustering, or machine learning in the form of isolation of forests as well as autoencoders to identify outliers. They may be caused by errors, but outliners can be valuable, for instance, to trace a new trend or special events.

Selective tools enable users to assess these anomalies against the backdrop of noise to make a sensible distinction about when an anomaly is significant. Moreover, the anomaly detection algorithms can be applied for real time mode and perform recursive calculations of the data streams for detecting any changes (Epperson et al., 2024). Another important approach of automated data profiling is data redundancy analysis, a goal of which is identification of duplicate records and repeated information.

Through using both exact matching and fuzzy matching algorithms, tools are able to determine which records are duplicates, despite this there can be small differences such as spelling or formatting. For exact matching, measures such as Levenshtein distance or Jaro-Winkler similarity can be applied while for probabilistic record linkage predictive methods calculate the extent of records' similarity based on several fields.

Profiling tools make it possible to control the dataset size and avoid such pathologies as excessive data volume that may distort analyses. Conversely, constraint validation is one of the traditional methods used to guarantee that datasets confirm to specific rules or specifications. All the impositions can be from simple tasks such as defining ranges for numerical values, ensuring mandatory fields are completed or from more complicated relational database structures such as foreign keys.

Checks employed by automated profiling tools confirm these constraints, where violations that might affect the dependability of the dataset are detected. Advanced tools superimpose business constraints; it utilizes Knowledge-based or best practices for analysing the constraints. For example, a profiling tool developed for healthcare data may incorporate patients for clinical coding systems such as ICD-10 or SNOMED; thereby making the data conform to selected medical terminologies.

In order to strengthen these techniques, the broad selection of automated profiling tools has been developed to meet the requirement and range of the enterprises. Informatica Data Quality and Talend Data Preparation are the examples of the tools used to perform profiling at an enterprise level because they are designed to support big data environments.

These tools interface with the other data platforms, consisting of profiling, cleansing and transformation, in the same environment. Free solutions such as Apache Griffin and DataCleaner are available for those organisations who want to develop complex profiling solutions at a relatively low cost.

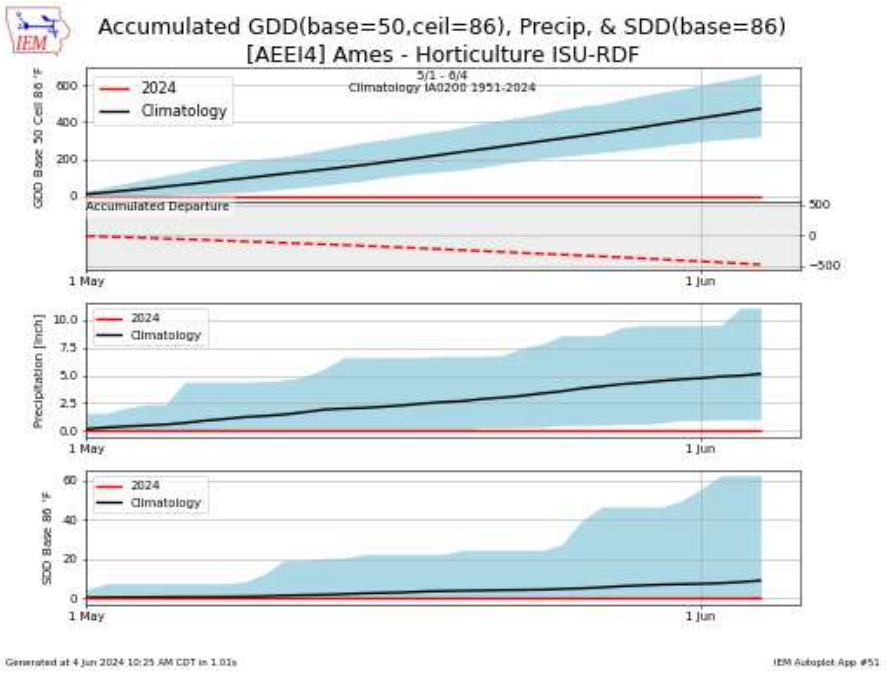

Figure 5 Automated Data Plotter (Mesonet, 2023)

This is because these tools use community development to ensure innovative practices are adopted within various implementations while at the same time allowing for flexibility (Mitropoulos et al., 2021). Profiling tools or services like AWS Glue DataBrew and Google Cloud Data Catalogue are available because they are more scalable and easily adaptable to the concept of cloud-based data architectures.

These tools offer instant profiling services, thereby avoiding the necessity for an enormous infrastructure within the organization. Also, integrations with other cloud ecosystems make them very friendly to work with; users are not exposed to issues moving their data from profiling to analytics or visualization. Some state-of-art tools like Alteryx and Trifacta use machine learning for profiling or present algorithm for suggestion to clean or enrich the data.

```
# Install necessary libraries
# !pip install pandas dataprep

import pandas as pd
from dataprep.eda import create_report

# Sample dataset
data = {
    "Name": ["Alice", "Bob", "Charlie", "David", "Eva"],
    "Age": [25, 30, None, 35, 40],
    "Salary": [50000, 60000, 70000, None, 80000],
    "Department": ["HR", "IT", "Finance", "HR", None],
}

# Create a DataFrame
df = pd.DataFrame(data)

# Generate a profiling report
report = create_report(df)

# Display the report
report.show_browser()
```

In contrast, these tools not only identify problems but also recommend courses of action for addressing those problems, making it easier to prepare data. The future of automated data profiling is in combining these techniques with relatively recently developed concepts like natural language process and blockchain.

Profiling based on unstructured data, as text or speech, being an approach available through the help of NLP-powered profiling tools. Opportunities exist therefore in the use of BCT in maintaining data provenance as well as detecting any alterations potentially arising from profiling processes to guarantee data integrity.

These technologies are still in their early phases but as these technologies grow or develop there is highly developed profiling tools that enough comprehensive to handle modern data environment challenges while at the same time protecting quality and security of data. Automated profiling tools are capable but must be properly deployed and managed to achieve the intent. Profiling initiatives must therefore have objective set that allow the organisation to contain such within a broader data management framework.

Moreover, reliability of profiling results depends on the quality of algorithms and rules used also. These rules must be update and maintained often to reflect changing standards of data and requirements of the organization. In addition, profiling tools must be tied-up with data governance frameworks in order to meet the legal requirement and ethical standards.

Used systematically and implemented by employing the most sophisticated technologies, automated data profiling enables organisations throughput the hidden value and use of their data for the creation of novel solutions and more effective decision making.

Improving Data Quality Through Profiling

Data profiling for improving quality is a technical approach of verifying all compiled data to ensure that they are correct, consistent, and suitable for use. Data profiling serves as a diagnostic function that turns up discrepancies and serves as the basis for remediation attempts. A primary way that rises in profiling improves the quality is through increasing the problem of different norms within sets.

Profiling tools describe the structure and content of the datasets and report on any differences to the prescribed values such as duplicated records, out-of-range values, or values in the wrong format. These issues, for instance, mean that organizations can focus on cleaning them and using approaches like record merging, or formatting irregularities, to ensure their records are clean once again.

Automated profiling takes this even step further in addition to identifying errors, intelligent algorithms are able to suggest or even perform corrections autonomously, saving enormous amounts of time and

increasing productivity (Yayik et al., 2022). Data profiling also plays an important role in pointing out weaknesses in the data and, in particular, cases when some data is missing.

Data gaps are often returned in tables due to the variability in data sources or if data is entered manually by a third party. Profiling tools identify missing data points so that problems identified can be solved using methods such as imputation, interpolation or well-informed techniques of estimation. In the same way, it is possible to define anomalies, which are values that differ from others or display some other irregularity, as defects, which could be manifested as outliers or, on the contrary, as extraordinary occurrences.

Profiling serves to facilitate examination of these anomalies as well as to identify their nature, origin, and respond to them appropriately. Aside from that, profiling helps in data management from any more mistakes by having rules and constraints implemented in profiling. Real-time tracking of data by automated tools makes it easier to track compliance to certain set standards each time hence lowering the rate of recurrent problems.

Data profiling also help in meeting the data governance and compliance on information used in organizations. Legal requirements including GDPR or HIPAA demand that there is quality data with the aim of preserving the data in question and using it in the right manner. It is done so by profiling tools that assess the datasets against the compliance criteria and pinpointing the violations along with the compliance of the information.

As this capability highlights legal implications that can harm an organization, it also helps to build trust between organizations and their stakeholders, improving the companies' image. Moreover, it is possible to use these profiling tools in order to ensure that different teams and systems adhere to the same set of best practices in regard to data. This harmonization is particularly beneficial in large organizations or industries due to the realization that data is locked in several silos hindering the organization's overall decision-making and flow.

As such, profiling helps to centralize the issue of data quality and make organizations capable of targeting problems and applying resources as necessary effectively. In this sense, using profiling is not limited to correcting inaccuracies in the data pool in the short term, profiled data is also useful to address more distant long-term objectives, such as data integration and data analysis.

Business Analytics, Machine learning and Business Intelligence all need high quality data to take it forward in the best possible manner. Profiling makes certain that the datasets gathered are sound and amenable for consistency with higher brain models that produces more accurate prognosis and more detailed outcomes.

Profiling is useful in integration tasks since it helps place data from different sources into a common structure to ensure compatibility (Scarcella, 2019). This is particularly important where shared data can inform key decisions within the business as is the case within the health care or financial sectors. Managing organisations benefit from enhanced efficiency and competitive advantage when profiling tools are used to unlock the hidden potential of data assets and enhance them into resources with strong innovation value.

It is within this context that integrating profiling into data management processes allows organizations to effectively build an organizational culture for drawing on information in making value creating decisions today and evolving to meet the challenges of the future.

Conclusion

Automated data profiling tool is an absolutely vital in modern data environment as it helps to control the quality of the data. As the means for identifying errors, correcting inconsistencies, and facilitating compliance, profiling enables organisations to achieve their goal to unleash the value of data.

The connectivity with higher advanced analytics and governance techniques opens paths to enhanced decision-making and operation orientations. Various profiling technologies shall always help in the

advancement of performance since they will help in addressing the future more so the present hitches and gaps in an organization's performance. Through the use of automated data profiling, organizations throughout the globe can set the cornerstone for success in the fast-developing environment of data applications.

References

Articles, Z., & Articles, Z. (2024, January 26). UNDERSTANDING PROFILING AND AUTOMATED DECISION-MAKING UNDER GDPR: IMPLICATIONS AND PRACTICAL APPLICATIONS - Zedroit. *Zedroit - Ensuring Privacy, Securing Business*. https://www.zedroit.com/understanding-profiling-and-automated-decision-making-under-gdpr-implications-and-practical-applications/

Ehrlinger, L., & Wöß, W. (2022). A survey of data quality measurement and monitoring tools. *Frontiers in big data*, *5*, 850611. https://doi.org/10.3389/fdata.2022.850611

Jakubik, J., Vössing, M., Kühl, N., Walk, J., & Satzger, G. (2024). Data-centric artificial intelligence. *Business & Information Systems Engineering*, 1-9. https://doi.org/10.1007/s12599-024-00857-8

Jang, W. -J., Lee, S. -T., Kim, J. -B., & Gim, G. -Y. (2019). A Study on Data Profiling: Focusing on Attribute Value Quality Index. *Applied Sciences*, *9*(23), 5054. https://doi.org/10.3390/app9235054

Mitropoulos, P., Patroumpas, K., Skoutas, D., Vakkas, T., & Athanasiou, S. (2021). BigDataVoyant: Automated Profiling of Large Geospatial Data. In *EDBT/ICDT Workshops*. http://star.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-2841/BigVis_1.pdf

Radivojević, T., Costello, Z., Workman, K., & Garcia Martin, H. (2020). A machine learning Automated Recommendation Tool for synthetic biology. *Nature communications*, *11*(1), 4879. https://doi.org/10.1038/s41467-020-18008-4

Scarcella, L. (2019). Tax compliance and privacy rights in profiling and automated decision making. *Internet Policy Review*, *8*(4). https://ssrn.com/abstract=3933264

Taleb, I., Serhani, M. A., Bouhaddioui, C., & Dssouli, R. (2021). Big data quality framework: a holistic approach to continuous quality management. *Journal of Big Data*, *8*(1), 76. https://doi.org/10.1186/s40537-021-00468-0

W. Epperson, V. Gorantla, D. Moritz and A. Perer, (2024). "Dead or Alive: Continuous Data Profiling for Interactive Data Science" in IEEE Transactions on Visualization & Computer Graphics, vol. 30, no. 01, pp. 197-207. 10.1109/TVCG.2023.3327367

Yayik, A., Aybar, V., APIK, H. H., Içöz, S., Bakar, B., & Güngör, T. (2022). Deep learning-aided automated personal data discovery and profiling. *Turkish Journal of Electrical Engineering and Computer Sciences*, *30*(1), 167-183. https://doi.org/10.3906/elk-2102-54