ISSN- 245-1142

International Journal of Humanities and Information Technology Volume 6, Issue 4, 2024

Autonomous Data Quality Management via ML in Cloud Warehouses

Aditi Namdeo*

Independent Researcher, Northeastern University, Boston, USA

Abstract

With this volume of data, velocity and variety, companies must be prepared to make a Degree of Autonomous Data Quality Management (ADQM) in their cloud data warehouse. If data is deployed to cloud or distributed environments with high frequency of change, defining the data quality rules subject-wise is not sufficient because they will not provide a reliable data, likewise if these rules are defined schema-wise, there can be multiple data schemas with different attributes or pattern which is also distributed. Traditional rule based data quality approaches, which used to be effective for ensuring data quality, are ineffective when schema and data pattern changes occur very often and extremely rapidly in cloud and distributed environments, e.g., data can be added with missing values and duplicated in the process of data quality improvement. To address the problems of data quality in real-time, a machine learning (ML) automatic data quality management mechanism for locating, categorizing and rectifying real time data quality problems inside the cloudwarehouses is proposed. Each of these four layers has built-in with 4 ingestion providers that support a variety of data types, ranging from structured such as relational tables, to unstructured such as JSONs, text files and more; data quality vendors that provide data statistics; built-in models for anomaly detection, supervised classifiers and various clustering models to find inconsistencies; and models built on reinforcement-learning feedback loops to automatically correct data using imputation, deduplication and schema alignment. It is employed in cloud warehouse systems that have scalability so that monitoring and adaptation to changing are possible. Experimental evaluations show that the system is more accurate in detecting and decreases the human effort in comparison to conventional ETL-based quality systems. The goal of the desired solution is to provide more trust on the data, lower operational costs, provide trusted analytics in the Enterprise cloud environment. Future work covers introducing the LLM within the framework for semantic DQ reasoning, adopting a domain transfer learning approach to tackle different variations of cloud applications scenarios, ensuring the privacy and security of the DQ process in multi-tenant cloud warehouses, introducing federated learning techniques, and the more powerful explainable, adaptable DQ process in realistic enterprise requirements and use cases at scale and in production in multi-tenant cloud settings.

Keywords: Autonomous Data Quality, Machine Learning, Cloud Data Warehouses, Data Profiling, Anomaly Detection, Data Governance, Reinforcement Learning

International Journal of Humanities and Information Technology (2024) DOI: 10.21590/ijhit.06.04.14

Introduction

In the Digital Transformation era, that cloud based Data Warehouse has turned to be one of the most important aspects of the way organisations operate and evolve as it stores, processes and analyzes a ton of structured and semi-structured data sources [1] [2]. These attributes of data-driven decision making in enterprises that are critical in enterprise solutions are provided by scalable storage, elastic computing and near-real-time analytics that are typical features of cloud warehouses [2] [3]. However, as the volumes of data increase and become more complex, data quality assurance becomes a very difficult task [5]. Various information sources tend to include inconsistent information as can be outdated information, data ambiguity with respect to its semantics, as well as data duplication, data missing, schema gap and others [5] [6]. All this has a consequence on the downstream analytics, models of machine learning and business intelligence solutions that results in wrong conclusions and sub-optimal decisions [6][7].

Most traditional methods for overcoming data quality problems so far have been by using a set of user-defined rules, manual inspection, as well as static pipelines that are used to migrate the data, so called Extract, Transform, Load (ETL) pipelines [7]. But these are not suitable for the changing conditions in today's cloud environment where conditions change rapidly and irrationally [7] [8]. Cloud Data Warehouses are characterized by rapidly (fast) changing schemas of data, rapid (fast) data ingestion stream and the addition of new and update of data stream [3]. The rule based systems, if they were used on a large scale with the objectives of efficiency/

commodore were not going to be scalable, and would also be expensive to maintain and be able to respond to if they encountered a problem they weren't made for [4]. Besides, none of the current systems are capable of establishing a new set of data anomaly patterns, nor deal with data environments that are very dynamic and complex [5].

Recent focus has been in moving away from these disadvantages towards automated and intelligent data quality management with the help of machine learning (ML) techniques [4]. Otherwise, in the case where there is no rule based knowledge, patterns can be learned from a set of historical data, anomalies can be detected and systems can adapt to the changing distribution of data using ML based techniques [4],[6]. Among these techniques are anomaly detection, clustering, classification/classifier, reinforcement learning and other methods have been proven to be effective in detecting and correcting data quality issues in more scalable and adaptive ways [6]. Assessments for the behavior of the ML-based approach can be similarly generalised if the knowledge and set of rules is not known from the start, and can be refined over time, which makes them applicable in cloud-native application architectures, in which the data characteristics tend to change over time [3].

But none of the existing solutions delivers any type of integrated end-to-end solution; they are semi-automated or standalone solutions for which they lack the cloud-based data warehouse environment in which they are supposed to work. Most of the techniques reported only for detecting the DQ problem, and this is without taking any steps toward either solving or automating the problem [4]. Others, however, limit their data quality solutions to specific data quality challenges: missing information imputation, duplicate detection and so on; without providing a shared set of tools for managing data quality in general [1]. Moreover, in several aspects, such as model drift, model interpretability, model scale and privacy issues [5], it is challenging to achieve complete object detection. An environment that used to be viewed as a horizontal solution started to form in the minds that will now control the data quality in each cloud in an independent and continuous way – in a world of multi-cloud/hybrid-cloud. All looking for a one-size-fits-all that can grow them both quickly and consistently, ensure high quality data, and – of course – isn't cloud or other provider dependent. For enterprises to adapt to a multi-cloud/hybrid-cloud environment, having an all-in-one and completely independent cloud orchestration solution that ensures guaranteed data quality, essentials or absolute idempotency between different clouds across the cloud's lifecycle is no easy feat.

In this perspective, a new paradigm which arises in consideration of data quality is Autonomous Data Quality Management (ADQM) based on the machine-learning technology [4]. ADQM's goals is to create autonomous systems capable of identifying, diagnosing and fixing the data quality issues with little or no effort [4] [2]. When the adapter is interfaced with the cloud datastores of data warehouses, the ADQM systems can continuously learn and take corrective actions on data streams with patterns changing overtime [2]. Apart from improved data sources to be worked upon these systems also are very effective to reduce manpower and time required to perform all the Data Governance functions to a great extent [3]. Moreover, the use of technology of the 'Cloud Native' concept enable Elasticity, Fault tolerance, Scalability and enterprise scale deployment.

This paper are presented common idea of the ADQM for the CDW using ML method. All the steps involved in data quality, data ingestion, data profiling, detection of anomalies and correction have also been considered in the framework. Benefits from supervised, unsupervised and reinforcement learning, and is capable of reaching adaptive and continuous improvement. Should be cloud native, scalable and real time interactive. The framework enables seamless data centrifugation, that can be performed without any outer pipelines at all, plus seamlessly integrate data analytics with the layer of the data warehouse.

The first aim to achieve with this project is to fill the gap of legacy data quality management solutions to full and autonomous, ML-driven solutions. The proposed approach is expected to achieve these goals and enhance accuracy of the Data problem Detection, reduce manual job on the correcting process and reduces the time of Data problem Detection. Besides this, it emphasizes on feedback-oriented learning processes, and adjusts the system according to users' feedback and learning historical corrections made with regard to the learning process. This kind of flexibility is in essence vital in keeping long-lasting data integrity within active cloud settings.

In conclusion, Autonomous Data Quality Management with machine learning is a crucial step forward with regards to present day data engineering procedures. Organisations are being deluged with data – both in terms of quantity and importance – and this information need to be trusted, consistent and reliable. The planned data environment will become a smarter and self healing data environment, which will speed up the development of next generation analytics, and Artificial Intelligence (AI) solutions and the efficiency of the data environment operations.

Requirements of Autonomous Data Quality Management via ML in Cloud Warehouses

For us it is called Autonomous Data Quality Management (ADQM) and we require the technical, architectural and operational capabilities for our cloud data warehouses to achieve reliable, scalable, intelligent data quality management. The requirements can be functional specifications, non-functional specifications and ML-specific ones.

Functional Requirements

Functional requirements are features that need to be supported by the system; among them, those that are the most important are:

Data Ingestion and Integration

Data feeds from multiple data sources with and without structures and different data types can be provided to the system in a variety of data structures or through different APIs, and also as data streams from IoT sensors, transactional data sources or feed from third party application. It should ensure it's compatible with cloud DWs and will not break the pipelines.

Data Profiling and Monitoring

It is important to continually review the quality in this instance, and do so using the quality indicators Completeness, Consistency, Uniqueness, Validity and Timessiness. Data quality-report generation should be automatic as should be data deviation monitoring in real time using QOS.

Anomaly Detection and Classification

Several issues with data quality need to be identified: data in-missing, duplicate data, mismatched schemas and outlier data. It should classify the anomalies by severity & anomaly type in order to the occurrence and it should assist to take up proper corrective measures.

Automated Data Remediation

Imputation, deduplication, normalisation and schema aligning are examples of processes that can be separately implemented within the framework for the task of correcting data. The actions should automatically be taken based on the results of the ML model.

Non-Functional Requirements

Non-functional attributes are attributes that ensure the efficiency, reliability and scalability of a system:

Scalability

It needs to be able to ingest huge volumes of data, close to "cloud native" data warehouse scale, with high data throughput, and scale out quite easily to distributed computing configurations.

Real-Time Performance

If feedback and correction of the data quality is required in near-real time or real time, then a low latency process needs to be implemented within the timescales of near-real time or real time.

Fault Tolerance and Reliability

It should not fail to be able to deal with nodes failures, malfunctioning data pipelines or squeezing in errors when building out the data models without crashing and losing data during run time.

Security and Compliance

Handling will be done in compliance with the clearly defined security processes and laws regarding privacy among which the General Declaration on Data Protection Regulation (GDPR).

Machine Learning-Specific Requirements

These requirements focus on intelligent and adaptive capabilities:

Adaptive Learning Models

In the context of Cloud Computing, the ML models are constantly called to be updated with new: distributions of data, the concept-drift and changing patterns of data.

Explainability and Interpretability

System's output should be clear with respect to what type of anomalies this system would capture and what trust and transparency it would need to take to gain it.

Feedback Mechanism

Creating a more practical model implementation with Reinforcement learning that utilizes feedback loop is needed to validate user and correction of the model.

Model Deployment and Versioning

The framework should enable the seamless migration, monitoring / upgrading of the models for machine learning / AI to the cloud-based data warehouse.

A Framework for Autonomous Data Quality Management via ML in Cloud Warehouses

"Autonomous Data Quality Management (ADQM)" framework for cloud DWs that automatically checks the quality of the data and automatically identifies/corrects the data quality issues (if feasible), with very little or no human involvement. It is supposed to be utilized in cloud conditions where the data is dynamic, distributed and noteworthy. It is a flow of data engineering flows and intelligent "learning" ML components that can adapt and learn from the historical data which can be used today and is prepared to handle the future.

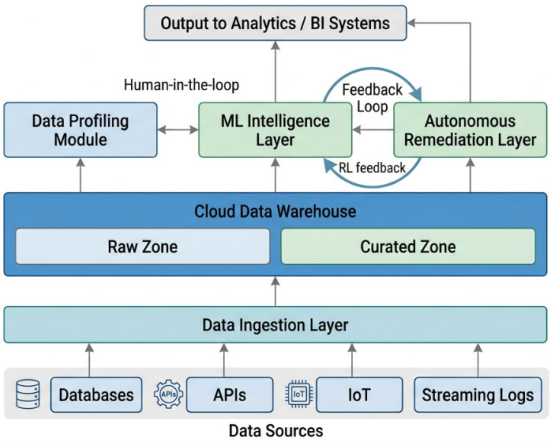

Figure 1: Overall Architecture of Autonomous Data Quality Management (ADQM) Framework

It consists of five layers; (1) Data Source and Ingestion Layer; (2) Data Storage and Warehouse Layer; (3) Data Profiling and Feature Engineering Layer; (4) Machine Learning–based Data Quality Intelligence Layer; and (5) Autonomous Remediation and Feedback Layer. All these are essential in E2E, Data quality management.

Data Source and Ingestion Layer

Data comes from a number of diverse (heterogeneous) data sources in the Data Source & Ingestion Layer. Some of the systems they live in are: Relational databases, Enterprise applications, APIs, IoT devices, Logs and streaming platforms. They do have hybrid ingestion ways: batch and real time ingestions to the cloud ecosystem.

This layer allows for a seamless and proper generation to cloud and for an easy and responsible ingestion via scalable ingestion tools with messaging. There are several types of ingestion constraint that can be provided: lighter schema conforming, data type validation and mandatory fields check syntax constraints. But with ML modules it is put off (deferred) downstream – i.e. in the downstream – which means you will need to ensure that the module is valid for use.

When it comes to ingestion fault tolerance, the layer is extremely useful as it ensures that if it is ingesting massive amount of data and data at high speed, it won't lose any. Ingestion pipelines also include metadata from the data ingested – e.g., data delivered, origin, time ingested, ingestion batch Id etc – used to provide quality analysis of the ingested data, as well as tracking the traces of the data.

Data Storage and Cloud Warehouse Layer

This data is then put in a scalable cloud data warehouse world. Here typical analytic services are acquired with which the ML engine gets structured, high-quality analytics. Distributed storage, columnar data format and elastic computing scaling are used to efficiently handle large amounts of data in the warehouse by the power of the universe.

The idea behind the raw zone is to retain the data as is; no matter how it has been edited; the latter is called curated zone which will hold the (manually) edited and transformed data in its current format without the edit and/or change. By separating data-provenance and rollback it will be possible to increase the data-provenance in case of malformed transformations.

Moreover, this layer will be managing a set of metadata management systems, where the Schema definition, Data Lineage Graph, various Data Quality measures information will be stored for each layer. The quality trends of data in the long term that are important to use are other metadata.

Data Profiling and Feature Engineering Layer

It is achieved by performing the raw data processing and extracting following valuable statistical and structural features in the Data Profiling and Feature Engineering Layer used for the ML based assessment of data quality: the missing

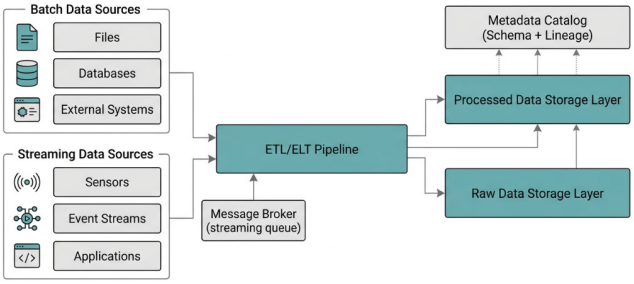

Figure 2: Data Ingestion and Storage Pipeline in Cloud Warehouse

value ratio, the uniqueness scores, distribution, correlation structures and schema consistency data quality metrics.

Profiling is performed when ingesting as well as periodically when on stored datasets (in case some might be delayed or hidden anomalies). Further, each data record/dataset will generate its feature vectors and these will be used as data sets for machine learning models.

The feature engineering contributes to the improvement of the models' accuracy into one of the key parts of the process. It provides means of encoding, manipulation/transformation of the data (e.g., aggregations, summarisations etc.,), semi-structure data or textual data encoding etc., normalisations, extract the data from time-series using advanced data processing methods.

The layer is vital for producing high-value and organized data for delivering it to the ML layers that can have a significant impact on accuracy in an anomaly detection and remediation.

Machine Learning–Based Data Quality Intelligence Layer

The proposed framework is built on the most solid foundation: Machine Learning Based Data Quality Intelligence Layer. The goal is to detect, classify and predict data quality issues using supervised and unsupervised learning algorithms and reinforcement learning approaches, in this layer.

Anomaly Detection Module

Abnormality Detection using Unsupervised Learning technique (cluster, autoencoder & Density based model, outlier detection & identification). Completely especially, it comes in handy in situations in which labeled information is not accessible. The model acquires the knowledge of data's normal distribution and abnormal data are labeled as potential abnormal data.

Classification Module

If there is a labelled data set, then these models (supervised) are used to classify the known problem types of DQ (e.g., missing values, scheme violations, etc.) to one of several classes as well as multiple types of DQ to one out of several different classes.It is designed to convert a set of categories

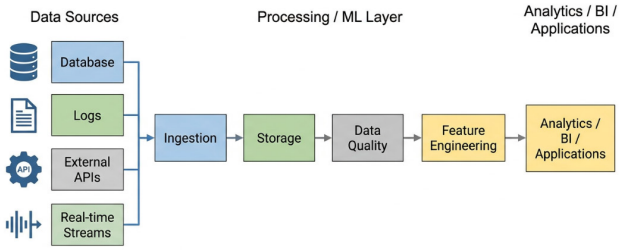

Figure 3: Data Profiling and Feature Engineering Process

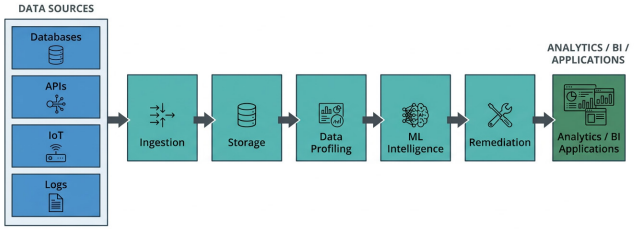

Figure 4: Machine Learning–Based Data Quality Intelligence Layer

of DQ issues like missing data, duplicated data, schema violations and similar categories to one of the several categories which supervised learning models (such as coordinate descent machine, random forests, decision trees) will recognize. To facilitate decisions about what remediation is needed.

Drift Detection Module

Another challenge is concept drift among data dynamically changing over time as is the case for cloud data. For this module, changes of data distribution over time are real-time tracked. Changes in the transformer are detected using several statistical methods, multiple ML based models of drift detection and the model is re-trained/re-adapted when changes are detected.

Predictive Quality Assessment

This has been used to predict future events of this kind and to deduce the cause of the event, based on the past rules. Specialised models for time-series forecast and learning models can be used to provide anticipatory estimates, including the warning of abnormalities before they happen, and taking action before they come to pass.

Autonomous Remediation and Feedback Layer

This engine is used to take the result of the ML intelligence layer, and based on the information provided, perform corrective actions such as:

To betters reinforce the remediation strategy, a reinforcement learning (RL) mechanism is employed for the reinforcement to reinforce strategies/punish strategies overtime. This system is given feedback based on the user validation, accuracy of analytics results at the downstream and compliance of business rules. Reward will be given for desired/appropriate behavior, and penalties for non-correct/sub-optimal behavior. In this way, the necessary correction policies are adaptively optimised so that they will present necessary corrections in an optimum way.

Self-supported, supervised and man-controlled seamlessly reached and configured for critical actions. Data Engineer/Data Analyst who understands and know what changed is right or adequate to add more labels. The feedback is passed to the ML models further increasing its accuracy and getting rid of any future or future error.

System Integration and Cloud Deployment

All the above is being implemented on a cloud-native (micro-services, containerisation technologies) architecture. The lessons learned are that each single thing per layer is our individual service, which implies that we can fail, be scalable and flexible. APIs can be considered as the primary medium of communication between various components and ensure the transfer of data between different components of a system is carried smoothly.

It works decently with modern day cloud data warehouses, and can run in the cloud data warehouse, or be translated into the cloud data warehouse. Because of continuous integration and continuous deployment (CI/CD) pipelines, updating the ML models and system components happens continuously which results to no downtime for the system.

Performance Evaluation

A concept called Autonomous Data Quality Management (ADQM) was proposed and analysed the capability of data quality detection, classification and correction for CDW environment. Emphasis placed on Correctness, Efficiency, Scalability and Adaptability of the assessment in a variety of data contexts. The outcome that showed, when compared with results from the traditional remediation system of rules, the adoption of autonomous remediation systems (including the one developed since, based on machine learning algorithms (MLA)) is significant in terms of the quality of the results.

Experimental Setup

Distributed computing, together with an environment in the cloud provides the data warehouse platform. A set of real and artificial data was created for testing this system that simulated the presence of a number of real data quality issues: outlier data value, mixed schemas, missing data and duplicate data. These were wrapped around this framework in such a way that they would become scalable, and could be tested independently. The two approaches to the baseline are: Use of a Rule-based approach (in the classic ETL scenario); Use of a standard Anomaly Detector approach (standard ML features of the rule, plus no usage remediatation).

Evaluation Metrics

Several performance evaluation metrics were used for measuring the performance. The precision, recall and F1

score were used to evaluate the performance of the detection system. Data Quality Correction was used to perform Good Data Assessment based on the Data Quality Improvement Index (DQII) that was used to perform robust assessment of the extend of inconsistencies in the data that can be corrected once the data was remediated. The effectivity of the system was measured on parameters such as: Latency (time for detection & correction) and Throughput (records per second) data delivery of the system. A series of data set sizes was used to measure scalability, and stability and response time was measured.

Anomaly Detection Performance

This intelligence layer with the help utilise the ML data quality issue outcome and outcome of the intelligence layer data quality was acceptable. The Autoencoders models were able to learn high complex data distributions, the Clustering models featured in the unsupervised learning methods along with these contributed to the high precision and high recall of the anomaly detection module. The system, which was proposed, was able to expose the anomalies identified that were not detected during previous testing and resulted in higher F1 scores than a mock based system where initial solution levels are set on all test sets. To ensure high accuracy it has been further enhanced to classify the anomalies that are picked up in the classification module and then take targeted remediation action.

Remediation Effectiveness

In the use of an autonomous remediation layer the overall quality of the data has significantly increased. Data was imputated (returned from missing relationships) and predictive data (boosting questions to yield similar data) was used to complete missing information, while similarity-based matching techniques were used to remove any duplicates. Any new inconsistencies in schema were dealt with using automated schema alignment strategies. Also, a reinforcement learning feedback system was beneficial because it learned to best correct errors in the long run by adjusting for errors from both the user and the system by considering the feedback (information) given by the user during the process. Overall data quality improvement index revealed a significant improvement from the baseline data quality system which means that data quality and consistency has been improved after processing.

Scalability and System Performance

The more data values in the scale tests the higher the stability. The system was designed as a Cloud-native Microservices architecture that results in a drop in the Latency level, and a virtually undetectable horizontal (growth) scaling. The detection and remediation pipeline is still able to scale to see an increase in dataset by order of magnitude. It has been used in research of appropriateness of the data in the enterprise cloud related to big data, big volume and big speed.

The outcomes of the evaluation show that the proposed ADQM system is more accurate, flexible and scalable as compared to the traditional ones. But with machine learning, real-time detection and fixing of data quality problems can be made possible with autonomous remediatization - saving manual effort and operating overhead. What's more, the reinforcement learning mechanism will help the system learn and improve, over time. My overall verdict: it's a pretty valuable quality data structure for use in today's data warehouses and certainly DWs in the cloud.

Future Research Directions and Opportunities

Within the proposed Autonomous Data Quality Management (ADQM) there are numerous exciting research and development challenges to explore. The next generation of the cloud data quality ecosystem should aim to make the interpretation, scaling, security and intelligence of cloud-based Data Quality greater, either by expanding their size, complexity and heterogeneity or by augmenting these with other capabilities.

Integration of Large Language Models for Semantic Data Quality

Leverage-Large Language Models (LLMs) to improve semantic understanding of Data Quality issues is one of the crucial future components of it. Existing ML models have achieved a very high accuracy regarding statistical knowledge of 'anomalousness' but failed to provide knowledge and contextual knowledge of 'anomaly.' When used properly, LLM can improve the understanding of the semantics of the column, a discovery of semantic inconsistencies, and create an alert in the event of errors that are domain-specific. This translates to being able to be smarter in how to enforce business rules, and even influence automated business rule corrections, on enterprise data that can be complicated.

Federated Learning for Privacy-Preserving Data Quality Management

More and more rules and regulations focusing on data privacy and privacy become recommended and federated learning is gaining popularity as an attractive solution for AI based notifications. Distributed cloud nodes can be used in the training of models, without sensitive data having to be transferred to a centralized location when models are deployed. In this way it ensures the privacy but at the same time enhance the worldwide model. Other publications on this topic can improve the quality synchronization mechanism of Federated and cross warehouses anomaly detection or by optimizing it.

Real-Time Streaming and Edge Data Quality Processing

Many organizations today are using real-time analytics, and many others are looking to do so in the future, making future ADQM systems that do not just use batch processing, but

support streaming data pipelines, a necessity. By leveraging the edge-computing, a close proximity with the data source that integrates seamlessly with the data source themselves could be a solution to these data quality issues. This is not only going to minimize latency, but it's also going to make things more responsive for applications that are time-sensitive as in IoT systems, financial trading and smart city infrastructure.

Explainable and Trustworthy AI for Data Quality

Moreover, decision making that relates to data quality associated with an ML process (which would be made more explainable) is one aspect that should be addressed. Current models are considered as a 'black-box,' on which data engineers have to trust auto-corrections. Future studies should focus on develop and implement the explainable AI (XAI) to allow man to understand what has happened and how the remediation actions are to be carried out. This will enable users of the autonomous systems to be more transparent, audible and trust worthy.

Self-Evolving and Zero-Touch Data Quality Systems

The upcoming generation of the ADQM can be purely independent and/or lightly-managed systems. Real-time feedback loops can in addition be used, automatically, to create optimization amongst detection thresholds, remediation strategies and model configurations using the reinforcement learning. This "no-touch" would be the icing on the cake when it comes to bringing more efficient, streamlined data operations and help advance the already highly independent big cloud data management.

Cross-Cloud and Multi-Cloud Data Quality Orchestration

One of the fundamental problems with solving a problem using multi-cloud solutions is it won't have a single source of truth.One of the important factors to keep in mind when fixing businesses headed towards multi-clouds is the uniformity in data. In the future a greater effort should be focussed on the cross cloud orchestration mechanisms presented in this document that link the rules that the data quality Association provides across different cloud providers. Requires frequent sharing of metadata, easy-to-interoperate ML models and a single monitoring dashboard to keep governance process well managed.

Conclusion

In this paper an overall solution using machine learning techniques to cope with the need of Autonomous Data Quality Management (ADQM) in cloud DW is proposed. Organizations now have to deal with more painful issues regarding data quality because it's available in the cloud. Manual designed data quality management (DQM) systems and standard algorithms are excellent alternatives for managing data quality in relatively small systems with fixed data schemas and sources, but they do not work well in the case of big data distributed systems and changing data. To handle the issues mentioned, the proposed architecture will be focused on adding a machine learning-based intelligence and a cloud native data warehouse architecture to support automated data quality management that will be able to scale and meet evolving demands.

In the future "ADQM" is broken up into pipes: data ingestion, data storage, data profiling, data quality machine-learning intelligence and automatic data quality remediation with feedback loops. With such a multi layered architecture it becomes easier to identify and rectify issues such as data quality problems like lack of consistency, missing data, data duplication, second/third system schema, and/or anomalies in real-time. Can also use supervised, unsupervised or reinforcement learning for knowledge/signed recognition of system's known and unknown DQ problems and can learn from feedback loops.

The framework functionality in terms of effectiveness (remediation) and accuracy (detection), scalability and operation efficiency is sufficient in comparison to traditional methods admittedly. There are less manual efforts and better analytic results with integrated automated corrections systems. Similarly, the new deployment – all in the cloud – retains a high workload, low latency and highly resilient characteristic of the systems.

The new ADQM will be a small step towards intelligent and self-managed data settings. It also provides Endogenous Quality of Data (QOD) right at the meeting point and enabling Data Governance in the contemporary business organisations with their own autonomy. The System is expected to evolve as newer technologies come into the picture of embedding AI, such as but not limited to Federated Learning, Explainable AI and Large Language Models. To conclude, we hope that we've provided you with an understanding of how an autonomous data quality engine can help you reach scalable, efficient and trustworthy data interaction in next-gen cloud-based DWs.

References

[1] C. Batini and M. Scannapieca, *Data Quality: Concepts, Methodologies and Techniques*. Berlin, Germany: Springer, 2006.

[2] P. A. de Freitas, A. C. de Souza, and R. S. Mello, "Information governance, big data and data quality," in *Proc. 2013 IEEE 16th Int. Conf. Comput. Sci. Eng. (CSE)*, pp. 1142–1143, 2013.

[3] M. Al-Ruithe, E. Benkhelifa, and K. Hameed, "Key dimensions for cloud data governance," in *Proc. 2016 IEEE 4th Int. Conf. Future Internet of Things and Cloud (FiCloud)*, pp. 379–386, 2016.

[4] I. T. Taleb et al., "Big data quality framework: A holistic approach to continuous quality management," *Journal of Big Data*, vol. 8, no. 76, 2021.

[5] L. Cai and Y. Zhu, "The challenges of data quality and data quality assessment in the big data era," *Data Science Journal*, vol. 14, pp. 2, 2015.

[6] B. T. Hazen, C. A. Boone, J. D. Ezell, and L. A. Jones-Farmer, "Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications," *Int. J. Prod. Econ.*,

vol. 154, pp. 72–80, 2014.
[7] R. R. Nelson, P. A. Todd, and B. H. Wixom, "Antecedents of information and system quality: An empirical examination within the context of data warehousing," *J. Manage. Inf. Syst.*, vol. 21, no. 4, pp. 199–235, 2005.
[8] M. Helfert and C. Herrmann, "Introducing data-quality management in data warehousing," in *Information Quality*, Routledge, 2014, pp. 135–150.
[9] N. Miloslavskaya and A. Tolstoy, "Big data, fast data and data lake concepts," *Procedia Computer Science*, vol. 88, pp. 300–305, 2016.
[10] H. Mehmood et al., "Implementing big data lake for heterogeneous data sources," in *2019 IEEE Int. Conf. Data Eng. Workshops (ICDEW)*, pp. 37–44, 2019.
[11] C. Moreno, R. A. Carrasco, and E. Herrera-Viedma, "Data and artificial intelligence strategy: A conceptual enterprise big data cloud architecture," 2019.
[12] P. Ponniah, *Data Warehousing Fundamentals for IT Professionals*. Hoboken, NJ, USA: Wiley, 2011.