Implementing of Data Quality in KYC using Cloud Environment

Sowmya M C*, Manjuprasad B, S. Meenakshi Sundaram

Dept. of Computer Science and Engineering, GSSS Institute of Engineering and Technology for Women, Mysuru.

DOI: https://doi.org/10.21467/proceedings.1.38

* Corresponding author email: Sowmyasonu381@gmail.com

Abstract

Data Quality is a cross-disciplinary and often domain specific problem due to the importance of fitness for use in the definition of data quality metrics. Existing model and methodologies capabilities are restricted to the structured data type and limit its ability to assess data quality in web and big data. Online banking is getting popularity due to location independence, 24/7 services and responsiveness. Financial services through the internet are running under various threats like phishing, malware, Man-In-The Middle (MITM) attack and the evolving sophistication of compromise techniques. One-time password (OTP) in online banking system alleviate the risk and make it secure. Know Your Customer (KYC) and sanctions requirements continues to be a key focus area for financial institution (FI) management, and firms must ensure they are following appropriate compliance procedures to meet the increasing regulatory demands.

Index Terms- Data Quality, OTP, KYC

1 INTRODUCTION

The data used in business analytics can be small or big. Term small data is a synonym for traditional data. Traditional data is defined as electronic data that is stored in databases, data warehouses or legacy systems. the definition of data includes digital data measurements, raw digital values, processed digital values and met vales. The emergence of an era of big data attracts the attention of industry, academics, and government. The use and analysis of big data must be based on accurate and high-quality data, which is a necessary condition for generating value from big data. Therefore, we analyzed the challenges faced by big data and proposed a quality assessment framework and assessment process for it. , effective fraud detection techniques and models are needed to improve the quality and reducing the cost of health care services, for which expertise domain knowledge is required[1] .

Achieving high data quality has become an important element in managing data within an organization. Possessing high data quality could help an organization to formulate better business strategy and unveil business pattern for decision making. Failure in providing high data quality to the organization have brought various issues such as false decision due to incorrect data, high cost of operation and lack of customer satisfaction.

Know Your Customer (KYC) information verification technique has been introduced as Challenge Question (CQ) during login using user ID and Password in order to verify user more intensively. In that case KYC must be privatized with widespread dynamic user input. The KYC database enriches from account opening initial data, user interaction and dynamic update through the application; on the other hand user can add more confidential information or random question/questions with answer/answers to the KYC database to make the authentication process much stronger and secured. Ranking on the KYC information will also be considered to be used as CQ; CQ will be asked to the user during login after success in user ID and Password verification. One or more CQ will be assigned to ask the user based on the risk factors assessment result. Top ranked CQ will be asked to the user when the risk assessment result is comparatively higher; on the other hand low ranked CQ will be asked for lower risk. The main endeavor of the banks and non bank financial institutions are to provide a consistent, secured and high available process of authentication to their customers with minimizing potential avenues of attack, especially attacking vectors beyond the control of either the customers or the Financial Institutions (FIs).

A lot of progress has been established in data quality research which are not only limited to the adoption of surveys and questionnaires as mentioned before. Thus, we urge to answer questions regarding progress in data quality research through a review of data quality research articles that has been published before this. Our main intention in doing this review is to highlight potential issues in data quality research and to discuss potential unfilled research gap in data quality research especially in managing data quality within the organization. Furthermore, this review is intended to facilitate data quality implementation within the organization by discussing the strengths and weaknesses of existing data quality management model and data quality assessment methods [2].

2 The challenges of data quality

The diversity of data sources brings abundant data types and complex data structures and increases the difficulty of data integration. The data generated from their own business systems, such as sales and inventory data. But now, data collected and analyzed by enterprises have surpassed this scope. Big data sources are very wide, including: 1) data sets from the internet and mobile internet; 2) data from the Internet of Things; 3) data collected by various industries; 4) scientific experimental and observational data physics experimental data, biological data, and space observation data. These sources produce rich data types. One data type is unstructured data, for example, documents, video, audio, etc. The second type is semi-structured data, including: software packages/modules, spreadsheets, and financial reports. The third is structured data. The quantity of unstructured data occupies more than 80% of the total amount of data in existence.

Data volume is tremendous, and itis difficult to judge data quality within a reasonable amount of time. After the industrial revolution, the amount of information dominated by characters doubled every ten years. After 1970, the amount of information doubled every three years. Today, the global amount of information can be doubled every two years. In 2011, the amount of global data created and copied reached 1.8 ZB. It is difficult to collect, clean, integrate, and finally obtain the necessary high-quality data within a reasonable time frame. Because the proportion of unstructured data in big data is very high, it will take a lot of time to transform unstructured types into structured types and further process the data. This is a great challenge to the existing techniques of data processing quality. Data change very fast and the "timeliness" of data is very short, which necessitates higher requirements for processing technology.

Due to the rapid changes in big data, the "timeliness" of some data is very short. If companies can't collect the required data in real time or deal with the data needs over a very long time, then they may obtain outdated and invalid information. Processing and analysis based on these data will produce useless or misleading conclusions, eventually leading to decision-making processing and analysis software for big data is still in development or improvement phases; really effective commercial products are few. No unique and approved data quality standards have been formed in China and abroad, and research on the data quality of big data has just begun. In order to guarantee the product quality and improve benets to enterprises, in 1987 the International Organization for Standardization (ISO) published ISO 9000 standards. Nowadays, there are more than 100 countries and regions all over the world actively carrying out these standards. This implementation promotes mutual understanding among enterprises in domestic and international trade and brings the benet of eliminating trade barriers. By contrast, the study of data quality standards began in the 1990s, but not until 2011 did ISO published ISO 8000 data quality standards ( ). At present, more than 20 countries have participated in this standard, but there are many disputes about it. The standards need to be mature and perfect. At the same time, research on big data quality in China and abroad has just begun and there are, as yet,few results .

3 LITERATURE SURVEY

Many authentication methods have been developed. There are a variety of technologies and methodologies FIs can use to authenticate transaction of the customers. These methods include the use of customer passwords, personal identification numbers (PINs), digital certificates using a public key infrastructure (PKI), physical devices such as smart cards, one-time passwords (OTPs) [4][5][8], USB plug-ins or other types of "tokens", transaction profile scripts, biometric identification, and others. The level of risk protection afforded by each of these techniques varies.

An alternative method of authentication for financial services through the internet. It is a method to minimize financial fraud forgery on online financial network. The main challenge to avoid fraudulent activity in the financial network is to keep the system away from unauthorized person. The proposed method presented here includes sensitive personal information, which is called KYC information, to verify the actual owner of the account for online financial activity.

It considers all the known and upcoming possible ways to theft information and unauthorized entry into the online financial system. The dimension to the risk of the hacking and information stealing is unlimited and tendency for these illegal operations is evolving from time to time; the method proposed is a way to extend the KYC database as required by risk assessment in a certain interval. As the FIs already preserved customers' KYC information, so it is effective and fruitful to continue the reuse of KYC data for authentication purpose rather than to bear additional cost involving mechanism [3].

In many data analytics scenarios the path from data to decisions is unclear and not everything can be automated. Typically, answering an analytical question leads to further questions about the data, so an exploratory or investigative analysis on the data is necessary. The evolving security threat can be minimized in a significant way. Moreover, the model performance is preserved and improved eventually in each user activity. Authentication measures are dynamically assigned to make the system more reliable and keep the unauthorized user out of the whole process.

4 3. Tools of KYC:

XAMPP stands for
Cross-Platform (X),
Apache (A),
MySQL (M)
PHP (P) and
Perl (P)

4.1 Components of XAMPP:

- **Apache:** Apache is the real mesh server’s tender that measures besides transports of mesh material for the system. Apache will be largely well-known mesh server’s web based, controlling around 54percent of altogether sites.
- **MySQL**: Each work delicate, basic or else confounded needs the record to secure accumulated facts. MySQL, exposed foundation will be ecosphere's largely notable databank organization. This shows the whole thing as of pro locales to capable stages like Word Press.
- **PHP:** PHP leftovers to Hypertext Pre-processor. This will be server-side scripting vernacular so as to control without a doubt the most surely understood destinations on the planet, including Word Press and Face book. It is exposed foundation, for the most part easy towards study; works faultlessly through MySQL, settle to predominant choice in favour of network engineers.
- **Perl**: Perl will be irregular states active encoding lingo used generally to sort out encoding, structure overseer. Likewise less standard for network progression purposes, Perl has a huge amount of forte applications [4].

5 QUALITY ASSESSMENT PROCESS FOR BIG DATA

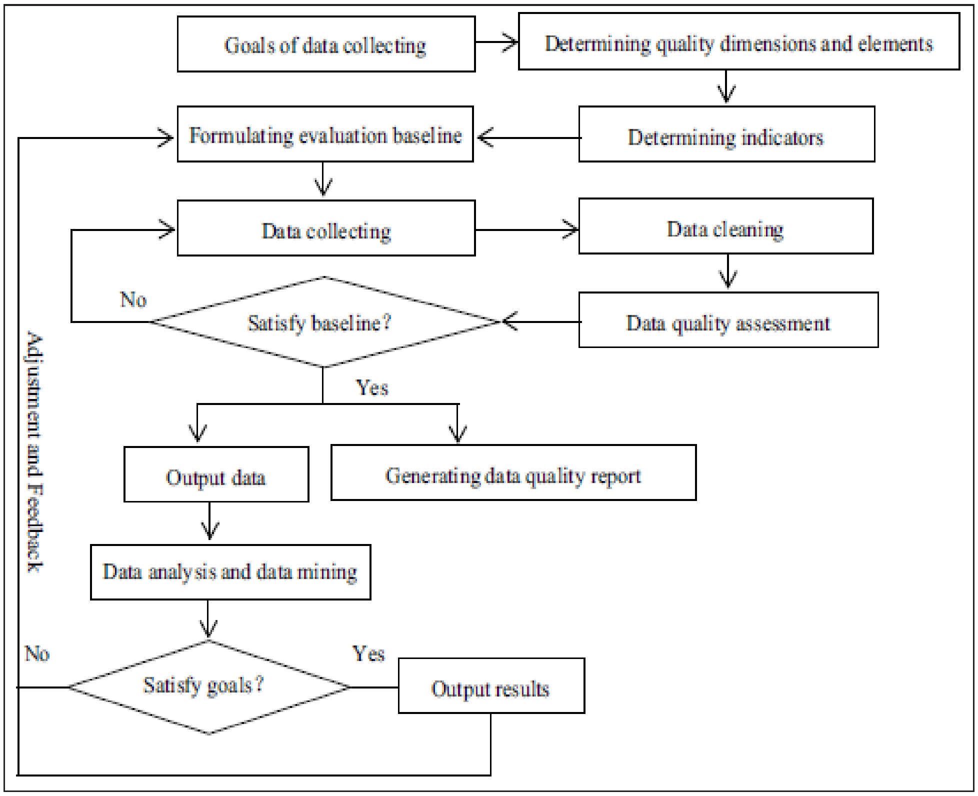

Figure 1: Quality Assessment process for big data

Determining the goals of data collection is the first step of the whole assessment process. Big data users rationally choose the data to be used according to their strategic objectives or business requirements, such as operations, decision making, and planning. The data sources, types, volume, quality requirements, assessment criteria, and specifications as well as the expected goals need to be determined in advance. In different business environments, the selection of data quality elements will differ. For example, for social media data, timeliness and accuracy are two important quality features. However, because it is difficult to directly judge accuracy some additional information is needed to judge the raw data and other data sources serve as supplements or evidence. Therefore, credibility has become an important quality dimension. However, social media data are usually unstructured, and their consistency and integrity are not suitable for evaluation. The field of biology is an important source of big data. However, due to the lack of uniform standards, data storage software and data formats vary widely. Thus, it is difficult to regard consistency as a quality dimension, and the needs of regarding timeliness and completeness as data quality dimensions are not high. In order to further quality assessment, we need to choose specific assessment indicators for every dimension. These require the data to comply with specific conditions or features. The formulation of assessment indicators also depends on the actual business environment. Each quality dimension needs different measurement tools, techniques, and processes, which leads to differences in assessment times, costs, and human resources. In a clear understanding of the work required to assess each dimension, choosing those dimensions that meet the needs can well define a project's scope. The preliminary assessment results of data quality dimensions determine the baseline while the remaining assessment as a part of the business process is used for continuous detection and information improvement.

Data analysis and data mining do not belong to the scope of big data quality assessment, but they play an important role in the dynamic adjustment and feedback of data quality assessment. We can use these two methods to discover whether valuable information or knowledge exists in big data and whether the knowledge can be helpful for policy proposals, business decisions, scientific discoveries, disease treatments, etc. If the analysis results meet the goal, then the results are outputted and fed back to the quality assessment system so as to provide better support for the next round of assessment [5].

6 KYC Information Database:

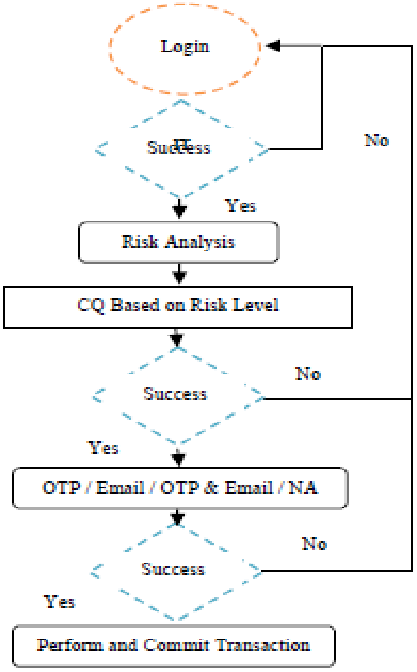

Figure 2: KYC information database

Model of authentication for financial application, some techniques of comprehensive risk factor judgment have been proposed to identify a suspicious login over many existing risk evaluations by analysis of user's historical activities data, and also proposed a new authentication method based on KYC information for authorizing the user during login as well as before performing transaction. In this method CQs are choosing from a collection of

CQs where the level of the CQ is defined by the risk factors calculation result. Risk factors are calculated considering Credential Risk (New user, Failed login attempt, User with no/very few identity information, Changing nature of transaction), Behavioural Risk Changed regular transaction purpose, Called to a new URL, which did not call by this particular user Transaction Risk (Exceeded regular transaction limit, Exceeded profile transaction limit), and Location Risk For appropriate and secure transaction using application, proposed model performs two operations. First one is risk factor calculation and second one is assigning the CQ based on the risk assessment result. The brief idea of the model is depicted in the Figure. In the figure the initial step is login of the user with user ID and Password verification like other online applications. The forwarding stage is risk analysis for the login succeeded user. Next stage assigns one or more CQ based on risk level formed by the result of prior stage. Final stage of the verification is OTP / EMAIL / OTP & Email confirmation if it is indicated by the result of risk analysis. In some other cases confirmation stage may not be applicable where CQ is in final stage before performing and committing a transaction. The CQ based authentication applies immediately after login and before performing and committing a transaction to verify the user rigorously [6].

7 Conclusion

In this paper, it has been proposed dynamic KYC based MFA authentication method to secure access of the financial services through the internet. Dynamic KYC based transaction authorization method to ensure secure and flawless financial access to the actual account holder of the online bank. Analysis and simulation results show that the proposed method provides equal control as existing OTP authorization minimizing some dynamic risk of being stolen and delay delivery of SMS. Critical data quality dimensions, systematic data quality management and data quality assessment methods have been successfully answered. As new technology become available, data in organizations is no longer limited to what are stored in the database.

References

[1] J. Liu, J. Li, W. Li, and J. Wu, “Rethinking big data: A review on the data quality and usage issues,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 115, pp. 134–142, May 2016.

[2] S. E. Madnick, R. Y. Wang, Y. W. Lee, and H. Zhu, “Overview and Framework for Data and Information Quality Research,” *Journal of Data and Information Quality*, vol. 1, no. 1, pp. 1–22, Jun. 2009.

[3] O.B. Lawal, A. Ibitola, O.B. Longe, “Internet banking authentication methods in Nigeria Commercial Banks,” African Journal of Computing & ICT,Vol 6. No. 1, March 2013.

[4] S. L. Lim, D. Damian, and A. Finkelstein, “StakeSource2.0: using social networks of stakeholders to identify and prioritise requirements,” in *Proc. ICSE*, 2011, 1022–1024.

[5] J. A. McCarty and M. Hastak, “Segmentation approaches in datamining: A comparison of RFM, CHAID, and logistic regression,” *J. of Business Research*, 60(6), 656–662, 2007.

[6] Syeda Farha Shazmeen, Shyam Prasad "A practical approach for secure internet banking based on cryptography," International Journal of Scientific and Research Publications, Volume 2, Issue 12, December 2012.