iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** IX **Month of publication:** September 2021

DOI: https://doi.org/10.22214/ijraset.2021.38226

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue IX Sep 2021- Available at www.ijraset.com

Enhancement of Data Quality in the Information Systems within Organization

Nishita Shewale
Atharva College of Engineering, Mumbai

Abstract: ***To introduce unified information systems, this will provide different establishments with an insight on how data related activities take place and there results with assured quality. Considering data accumulation, replication, missing entities, incorrect formatting, anomalies etc. can come to light in the collection of data in different information systems, which can cause an array of adverse effects on data quality, the subject of data quality should be treated with better results. This paper inspects the data quality problems in information systems and introduces the new techniques that enable organizations to improve their quality of data.***

Keywords: Information Systems (IS), Data Quality, Data Cleaning, Data Profiling, Standardization, Database, Organization

I. INTRODUCTION

As Information systems have a significant impact on financial Institutions, Education, Public and Private Organizations to provide up-to-date insights of their activities, processed records and manage information on their business, products, as well as help them into increasing revenue by foreseeing future trends and business. For Analysts, they offer opportunities to collect, categorize & use the information, to come up with proof of concepts to provide solutions, design new applications & products, for preparation of data, to automate the process which generates reports , or for the external presentation of their research and scientific expertise. The data quality takes precedence. Only correct data can offer resilient, beneficial outcomes and permit for a profound knowledge of the data to an establishment which is constantly up-to-date. The completeness, correctness and timeliness of the data are hence important for successful operational processes. The errors or anomalies in the data may affect different aspects and will make entire analytical process unreliable for an establishment. Therefore, the aim of this paper is to define and classify the problems of the quality of data that may arise in the information systems, and then present new techniques that are used to recognize, quantify, resolve data quality problems in information systems, to improve their Data quality.

II. INFORMATION SYSTEM (IS) - OVERVIEW

An information system (IS) is centralized databases that can collect, manage, and disseminate information about individual information, activities & Identifiers. The data considered here contains metadata about employee details such as Personal Identifiable Information, Employment details, Third party details. The IS architecture typically comprises the following components (see Figure 1):

1) Data Access Layer (DAL)
2) Application Layer (AL)
3) Presentation Layer (PL)

The diagram below depicts the individual processes and shows which components belong to which process step:

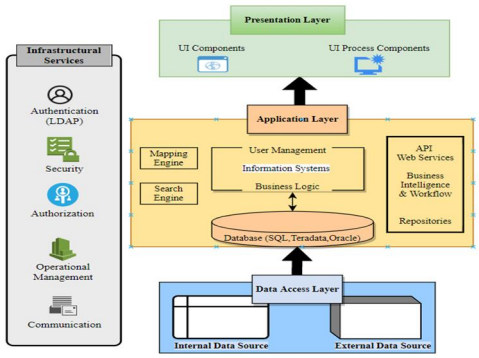


Figure 1: Information System Architecture

The Internal and external data sources are a part of data access layer. This level includes, for example, databases from administrations or human resources and external recruiter repositories, identifiers such as employee social data. The application layer houses the information system and its applications, which combine, manage, and analyses the data stored at the underlying level. The presentation layer depicts for the user the target group-specific preparations and representations of the analysis results. Beside various reporting options, you can also fill out portals and websites for the establishment here. There are Infrastructural Services, which are orthogonal to the described layers and provide overlapping services for the entire information system, such as security, authentication, authorization, communication, and so on. This data model describes the entities and their relationships with one another.

III. ORGANIZATIONAL LEVEL INTEGRATED ETL PROCESS FOR DATA QUALITY MAINTENANCE

Data warehouses [5][7] cause and provide extensive data cleaning support. They load and continuously refresh massive amounts of data from a variety of sources, increasing the likelihood that some sources contain "dirty data." Furthermore, data warehouses are used for decision making, so the accuracy of their data is critical to avoiding incorrect conclusions. For example, duplicated or missing data will cause an inaccurate or misleading statistics (“garbage in, garbage out”).

Data cleaning is one of the most intractable problems in data warehousing because of the wide range of data inconsistencies and the sheer volume of data. Further data transformations deal with schema/data translation and integration, as well as filtering and aggregating data to be stored in the warehouse, during the ETL process (extraction, transformation, loading), as illustrated in Figure 1. As shown, all data cleaning is typically done in a separate data staging area prior to loading the transformed data into the warehouse. To assist with these tasks, a wide range of tools with varying functionality are available, including Big Data, Pyspark, Ab initio, Informatica, etc. among others. Often, a significant portion of the cleansing rules and transformation work must be performed manually by developers or analytics that are simple to write and maintain.

Data transformation steps for federated database systems and web-based information systems are similar to those for data warehouses. There is typically a wrapper for extraction and a mediator for integration per data source [11][12]. So far, these systems have only provided limited support for data cleaning, instead focusing on data transformations for schema translation and schema integration.

Data is not pre-integrated as it is in data warehouses, but must be extracted from multiple sources, transformed, and combined during query execution. The resulting communication and processing delays can be significant, making acceptable response times difficult to achieve. The effort required for data cleaning during extraction and integration will further increase response times, but it is necessary to achieve useful query results.

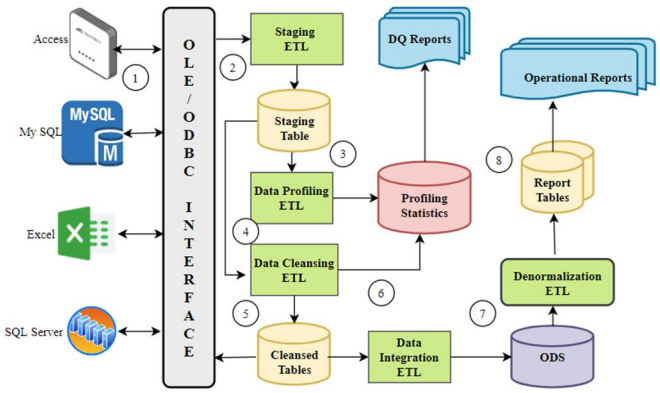


Figure 2: Organization level ETL Process

In general, data cleaning comprises of several stages:

A. Data Analysis

A detailed data analysis is required to determine which types of errors and inconsistencies must be removed. Besides a manual examination of the data or data samples, analysis programs should get metadata about the data properties and detect data quality issues.

B. Definition of Transformation Workflow and Mapping Rules

A large number of data transformation and cleaning steps may be required depending on the number of data sources, their degree of heterogeneity, and the "dirtiness" of the data.

A schema translation is sometimes used to map sources to a common data model; a relational representation is typically used for data warehouses.

Early data cleaning steps can help to resolve single-source instance issues and prepare data for integration. Later steps address schema/data integration and cleaning multi-source instance issues such as duplicates. Control of the data flow for these transformations and cleaning steps should be specified within a workflow that defines the ETL for data warehousing.

Schema-related data transformations and cleaning steps should be specified as much as possible using a declarative query and mapping language to enable automatic generation of transformation code. Furthermore, during a data transformation workflow, user-written cleaning code and special-purpose tools should be able to be invoked. The transformation steps may ask for user feedback on data instances that lack built-in cleaning logic.

C. Verification

The correctness and effectiveness of a transformation workflow and the transformation definitions should be tested and evaluated, for example, on a sample or copy of the source data, in order to improve the definitions as needed. Multiple iterations of the analysis, design, and verification steps may be required, for example, because some errors become apparent only after applying certain transformations.

D. Transformation

The transformation steps are carried out either by running the ETL workflow for loading and refreshing a data warehouse or by answering queries from multiple sources.

E. Backflow of Cleaned Data

After (single-source) errors are removed, the cleaned data should replace the dirty data in the original sources in order to provide improved data to legacy applications and avoid redoing the cleaning work for future data extractions. The cleaned data is available from the data staging area for data warehousing (Fig. 1).

In the following part issues related to data sanity has been described.

IV. PROBLEMS RELATED TO DATA QUALITY

A data warehouse is a collection of large amounts of data from various sources that have been consolidated. Enormous amount of data does not automatically guarantee quality. And with high volumes, it becomes more important to focus on the quality in order to derive some meaningful insights out of the available data. In most circumstances, the worth of the data is determined by its ‘fitness of use’, This criteria of data are the central part to determine the necessity or mandate for organizations to invest in data quality.

In each organization, data is often stored in a database. It can be personally identifiable information (PII), Information about business, projects and products. For further processing & management, this data needs to be of good quality, so that users can get better results. The quality of data is often defined as the condition of data based on factors such as accuracy, completeness, consistency, reliability and whether it's up to date. Different stakeholders can set requirements, In the IS context, e.g. especially by users of an IS, but also by the IS administrator. Poor quality data includes data errors, typographical errors, missing values, incorrect formatting, etc. Such quality issues in database need to be analyzed and then rectified by data transformation and data cleansing. The following are the typical issues of data in the context of an IS (see figure 3):

A. Missing values (features completeness)
B. Incorrect information caused by input, measurement, or processing errors (characteristic correctness)
C. Duplicates in the dataset (feature redundancy-free)
D. Unevenly represented data (feature uniformity)
E. Logically contradictory values (consistency characteristic)

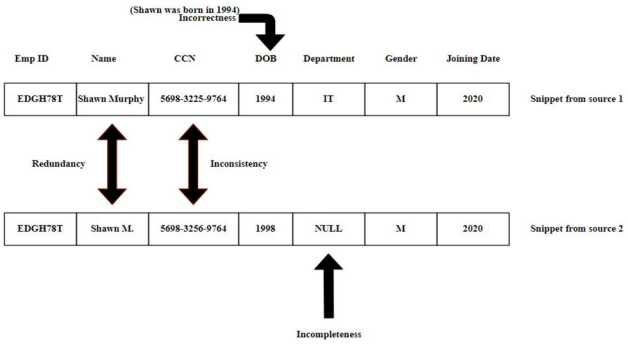


Figure 3: Most Common Data Quality Issues example in IS

Let's look at the possible improvements & approaches with examples in order to achieve quality & enrichment of the data.

V. DATA QUALITY IMPROVEMENTS

In organizations, heterogeneous data sources are used to feed their Business Intelligence (BI) and analytics. In order to maintain the sanity of this data in-house integrated data warehouse from both internal and external sources establishments needs to overcome the problems stated in Chapter III. At this stage, it is crucial to find the root causes and to enhance the data quality. The process of identifying and correcting errors and inconsistencies to increase the quality of given data sources in IS; is referred to as data cleansing (or "Data scrubbing") [4]. Data Cleansing includes all necessary activities to clean rogue data (inaccurate, incomplete or inconsistent, erroneous, etc.)
The following is a rough outline of the data cleaning process [6]:

1) Identifying and defining the actual issue
2) Locate and identify erroneous instances
3) Correction of discovered flaws

Data cleansing can be done using a variety of specialized methods and technologies within the process. [10] Mainly subdivides them into the following phases (see figure 4):

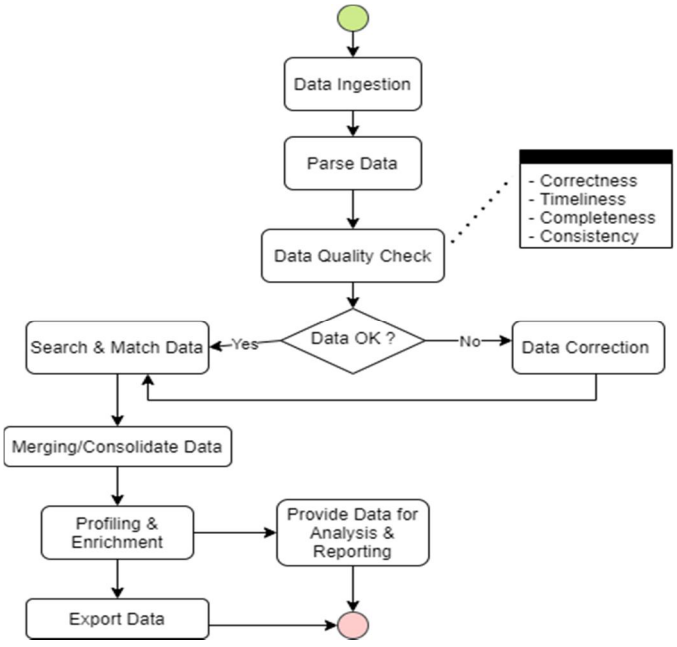


Figure 4: Data Cleansing Process Flow

Parsing is the first and most crucial component of data cleansing. A parser works as a simulator or interpreter component that breaks data into smaller elements for easy to understand, which helps the user to understand and transform the attributes more accurately. The process identifies tokens within a data instance and finding for recognizable patterns & segregating data into smaller elements. For this process, e.g. for names, mobile number, addresses, zip code and city. Application-specific rules & transformations can be created using these generated tokens. The biggest problems here are different field or entity formats that have to be detected. More often than not, parsing aids in the identification of data that requires cleaning, as well as the ability to customize or extend the rule library with user-defined rules for specific needs.

Correction & Standardization is further necessary to check the parsed data for correctness, and then make the alterations afterwards. Standardization is the prerequisite for successful search & match and there is no way around using a second reliable data source. For address data, a postal validation is recommended likewise for Telephone/Mobile number, a country code, area code validation is proposed.

Enhancement or data enrichment is the process that enhances existing data. Additional data is added to close existing information gaps with data from other sources. Most common enrichment values are demographic, geographic or address information.

Consolidation (merge) intends to match data entities with similar contexts are recognized by bringing them together.

Search and Match comprises of different types, such as for de-duplicating, matching to different datasets, consolidating or grouping. The adaptation allows the same data to be recognized. Redundancies, for example, can be identified and condensed for more information.

Data profiler profiles data and calculates a variety of statistics about it, both for the entire dataset and for individual fields or columns. Data profiling is capturing technological structures, analyzing them to discover data quality issues, and identifying business inconsistencies in order to understand the data and Profilers make no adjustments. They are only used to analyze data. Profilers are frequently employed at the start of a data study, but they can also be useful for better illustrating data after Consolidation.

For establishing and sustaining optimum data quality in information systems, every process is necessary. The cleansing eliminates errors in integrating multiple data sources in IS.

Table 1 shows how data cleansing processes (cleaning up, standardization, enrichment, matching, and merging) can improve the quality of data sources by identifying records with incorrect names on an employee list.

The cleansing procedure adds missing entries to the database according to a specific format & fields are automatically adjusted to establish guidelines.

A. Source Data Before Cleansing

Data Source					
Employee ID	**Name**	**CCN**	**Birth Date**	**Address**	**Mobile Number**
478630	Shawn Murphy	0001-2788-0323-2108	30-09-1994	140 J, Keneddy Road, Mumbai, 400002	9320557748
478630	Shawn John Murphy	0000-0000-0000-0000	30.09.1994	Keneddy Road, Mumbai, 400002 140J	+91 9320557748
478630	Mr. Shawn Murphy	0001-2788-0323-2108	126540	24 Galaxy 400002 Road	93 20 55 7748
478631	S. Murphy		12/34/67	10 Dexter Ave 400002	9456639064/9320557748
478631	Murphy Shawn	0001-2788-0323-2108	30.02.1994		000009320557748
	John Burphy	1278803232108	30.03.1990	Road C.,400002 140 J,IN	###&#455555
500236	Lea Steven	0235-7539-8756	02-04-1988	840-4678,Raymond street Nagpur 400045	91 7865445324
500236	Steven Lea	023575398756	1988	4678 Raymond Nagpur 400045	7865445324

B. Data After Cleansing

In this example, the missing pin code is determined based on the addresses and added as a separate field. Enrichment rounds off the content by comparing the information against external source, such as geographic and demographic factors, and dynamically expanding and optimizing it with attributes.

Cleansed Data						
Employee ID	**FirstName**	**LastName**	**CCN**	**Birth Date**	**Address**	**Mobile Number**
478630	Shawn	Murphy	0001-2788-0323-2108	1994-09-30	400002;MH; Mumbai; 140 J. Keneddy Road	9320557748
478630	Shawn	Murphy		1994-09-30	400002;MH; Mumbai; 140 J. Keneddy Road	9320557748
478630	Shawn	Murphy	0001-2788-0323-2108		400002;MH; Mumbai; 25 Galaxy Road	9320557748
478631	Shawn	Murphy			400002;MH; Mumbai; 10 Dexter Ave	9456639064, 9320557748
478631	Shawn	Murphy	0001-2788-0323-2108	1994-02-30		9320557748
	Shawn	Murphy		1994-03-30	400002;MH; Mumbai; 140 J. Keneddy Road	
500236	Lea	Steven	0235-7539-8756	02-04-1988	400045;MH;Nagpur; 840-4678 Raymond street	7865445324
500236	Lea	Steven	0235-7539-8756		400045;MH;Nagpur; 840-4678 Raymond street	7865445324

C. Data Before Enrichment

Cleansed Data

Employee ID	FirstName	LastName	CCN	Birth Date	Address	Mobile Number
478630	Shawn	Murphy	0001-2788-0323-2108	1994-09-30	MH; Mumbai; 140 J. Keneddy Road	9320557748
478630	Shawn	Murphy		1994-09-30	MH; Mumbai; 140 J. Keneddy Road	9320557748
478630	Shawn	Murphy	0001-2788-0323-2108		MH; Mumbai; 25 Galaxy Road	9320557748
478631	Shawn	Murphy			MH; Mumbai; 10 Dexter Ave	9456639064, 9320557748
478631	Shawn	Murphy	0001-2788-0323-2108	1994-02-30		9320557748
	Shawn	Murphy		1994-03-30	MH; Mumbai; 140 J. Keneddy Road	
500236	Lea	Steven	0235-7539-8756	02-04-1988	MH;Nagpur; 840-4678 Raymond street	7865445324
500236	Lea	Steven	0235-7539-8756		MH;Nagpur; 840-4678 Raymond street	7865445324

D. Data After Enrichment

The following example shows how the reconciliation and merge process works. Because related entries within a system or across systems can be automatically recognized and linked, tuned, or merged, merging and matching promote consistency.

Enriched Data

Employee ID	FirstName	LastName	CCN	Birth Date	Address	Pin Code	Mobile Number
478630	Shawn	Murphy	0001-2788-0323-2108	1994-09-30	MH; Mumbai; 140 J. Keneddy Road	400002	9320557748
478630	Shawn	Murphy		1994-09-30	MH; Mumbai; 140 J. Keneddy Road	400002	9320557748
478630	Shawn	Murphy	0001-2788-0323-2108		MH; Mumbai; 25 Galaxy Road	400002	9320557748
478631	Shawn	Murphy			MH; Mumbai; 10 Dexter Ave	400002	9456639064
478631	Shawn	Murphy			MH; Mumbai; 10 Dexter Ave	400002	9320557748
478631	Shawn	Murphy	0001-2788-0323-2108	1994-02-30			9320557748
	Shawn	Murphy		1994-03-30	MH; Mumbai; 140 J. Keneddy Road	400002	
500236	Lea	Steven	0235-7539-8756	02-04-1988	MH;Nagpur; 840-4678 Raymond street	400045	7865445324
500236	Lea	Steven	0235-7539-8756		MH;Nagpur; 840-4678 Raymond street	400045	7865445324

E. Matching

This example looks for entries that are related to Shawn Murphy and Lea Steven. Despite the similarities in the datasets, not all information is duplicated. The adjustment functions scrutinize the data in the individual records to determine which are redundant and which are independent.

Cleansed Data

Employee ID	FirstName	LastName	CCN	Birth Date	Address	Pin Code	Mobile Number
478630	Shawn	Murphy	0001-2788-0323-2108	1994-09-30	MH; Mumbai; 140 J. Keneddy Road	400002	9320557748
478630	Shawn	Murphy		1994-09-30	MH; Mumbai; 140 J. Keneddy Road	400002	9320557748
478630	Shawn	Murphy	0001-2788-0323-2108		MH; Mumbai; 25 Galaxy Road	400002	9320557748
478631	Shawn	Murphy			MH; Mumbai; 10 Dexter Ave	400002	9456639064
478631	Shawn	Murphy			MH; Mumbai; 10 Dexter Ave	400002	9320557748
478631	Shawn	Murphy	0001-2788-0323-2108	1994-02-30			9320557748
	Shawn	Murphy		1994-03-30	MH; Mumbai; 140 J. Keneddy Road	400002	
500236	Lea	Steven	0235-7539-8756	02-04-1988	MH;Nagpur; 840-4678 Raymond street	400045	7865445324
500236	Lea	Steven	0235-7539-8756		MH;Nagpur; 840-4678 Raymond street	400045	7865445324

F. Consolidation

The merge creates a comprehensive data set from the reconciled data. In this example, Shawn Murphy's duplicate entries are merged into a single record that contains all of his information also his previous employment ID 478631 has different DOB (wrong information given- human error) which after cross checking with existing data can be removed along with his earlier Address

Cleansed Data

Employee ID	FirstName	LastName	CCN	Birth Date	Address	Pin Code	Mobile Number
478630	Shawn	Murphy	0001-2788-0323-2108	1994-09-30	MH; Mumbai; 140 J. Keneddy Road	400002	9320557748
478630	Shawn	Murphy		1994-09-30	MH; Mumbai; 140 J. Keneddy Road	400002	9320557748
500236	Lea	Steven	0235-7539-8756	02-04-1988	MH;Nagpur; 840-4678 Raymond street	400045	7865445324
500236	Lea	Steven	0235-7539-8756		MH;Nagpur; 840-4678 Raymond street	400045	7865445324

Golden Record

Employee ID	FirstName	LastName	CCN	Birth Date	Address	Pin Code	Mobile Number
478630	Shawn	Murphy	0001-2788-0323-2108	1994-09-30	MH; Mumbai; 140 J. Keneddy Road	400002	9320557748
500236	Lea	Steven	0235-7539-8756	02-04-1988	MH;Nagpur; 840-4678 Raymond street	400045	7865445324

When it comes to the production of statistics and data reports, the frontend of the IS could be checked with the profiling process. Profiling makes it simple to assess the overall state of the data, identify, prioritize, and correct errors, and address the root cause of quality issues. After creating a profile, a facility can respond to quality issues by continuously monitoring profile-related parameters.

VI. ELIMINATING DATA REJECTION

In figure 5, after passing the data through parser, there might be rejected records which didn't pass the rule tests or not following the criteria. These records can be fixed the frequent problems at the internal or external data sources.

If provision is made to identify the reason of rejections with the Error Message & Error codes user can store this data in the database & draw trends to showcase the main rejection reasons to the third party or clients & work on eliminating the frequent and huge amount of rejected data henceforth, helping the establishment to nurture the database with more data to feed Business Intelligence.

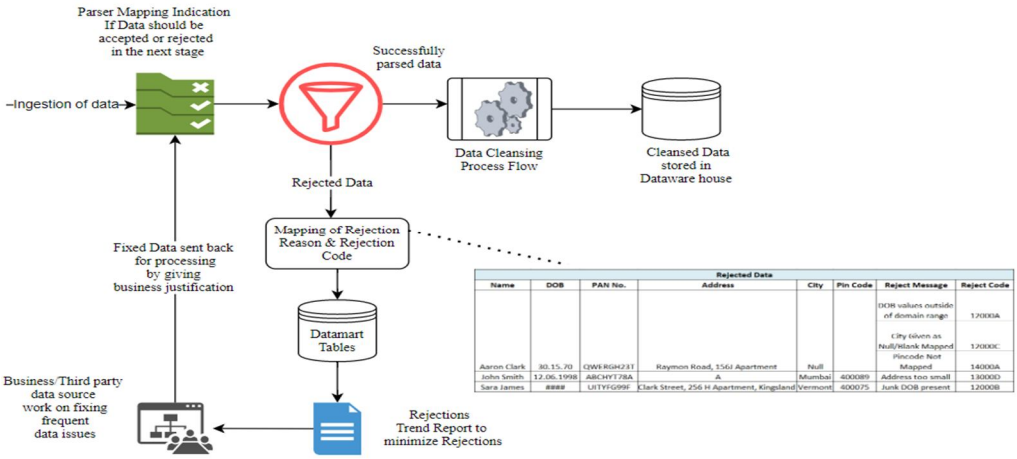


Figure 5: Process Flow for Rejected Data Salvation

VII. DISCUSSION

A data cleaning strategy must meet several requirements. To begin with, it should detect and eliminate all major errors and inconsistencies in individual data sources, as well as when integrating multiple sources. Tools to reduce manual inspection and programming effort should aid the approach, and it should be easily extensible to cover additional sources. Furthermore, data cleaning should be performed concurrently with schema-related data transformations based on comprehensive metadata. Mapping functions for data cleaning and other data transformations should specify declaratively and reused for other data sources and query processing. A workflow infrastructure, particularly for data warehouses, should support in order to execute all data transformation steps for multiple sources and large data sets in a reliable and efficient manner.

While there is a large body of research on schema translation and schema integration, the research community has paid little attention to data cleaning. Several authors, for example, [13][14][15][19][2][3], concentrated on the problem of duplicate identification and elimination. Some research groups focus on general problems that are not limited to data cleaning but are relevant to it, such as special data mining approaches [25][24] and data transformations based on schema matching [8][21]. Recently, several research efforts have proposed and investigated a more comprehensive and uniform treatment of data cleaning, which includes several transformation phases, specific operators, and their implementation [13][19][9]. The essential deliverables in data quality research are data quality management models, methodologies, and data quality assessment methods. Furthermore, data quality can be managed systematically within an organization by implementing an appropriate data quality management model and data quality assessment methods. In this review, we discussed several available models, methodologies, and assessment methods; however, more research can be done to fill the gaps identified in this review, such as the quality assessment of unstructured data types. Furthermore, in this day and age of big data technology, models, methodologies, and data quality assessment methods that can support unstructured data are critical.

VIII. CONCLUSION

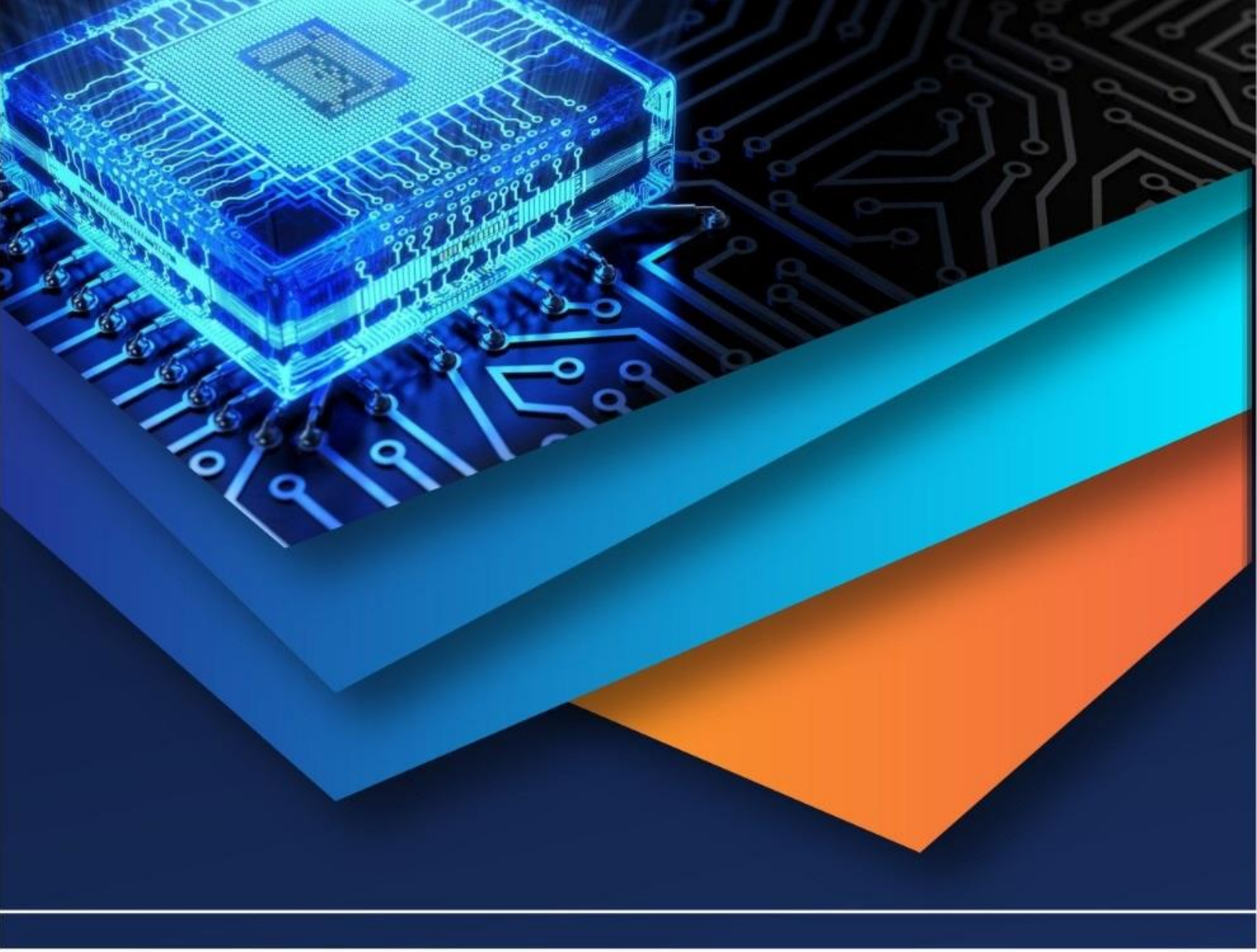
The question addressed were in this paper was which quality problems can occur in information systems and how to fix and improve those using new techniques or methods, such as Data Cleansing & how to minimize the rejection of records. As a result, it was demonstrated that data quality can be improved at various stages of the data cleansing process in any information system, and that high data quality can be obtained in financial institutions and organizations in order to successfully operate; for example, information systems. In this regard, the review and improvement of data quality is always a priority. The illustrated concept can be used as a starting point for utilizing facilities. It provides a suitable procedure and a use case, on the one hand, for better evaluating data quality in IS, better prioritizing problems, and preventing them in the future as soon as they occur.

These data errors must be corrected and improved using Data Cleansing. "The sooner quality defects are detected and corrected, the better," it says.

Already during the acquisition phase, the author or a downstream control authority can correct software errors such as typos, missing values, incorrect formatting, contradictions, and so on. There are numerous tools available to assist universities and all organizations in implementing data cleansing processes. All facilities can use these tools to improve; the completeness, correctness, timeliness, and consistency of their key data, as well as successfully implement and enforce formal data quality policies. Data-cleaning tools are primarily commercial in nature, and they are available for both small application contexts and large data integration application suites. In recent years, a market for data cleansing as a service has emerged on organization level.

REFERENCES

[1] Rahm, E. and Do, H.H. Data cleaning: Problems and current approaches. IEEE Bulletin of the Technical Committee on Data Engineering, 23(4), 2000.
[2] Monge, A. E. Matching Algorithm within a Duplicate Detection System. IEEE Techn. Bulletin Data Engineering 23 (4), 2000 (this issue).
[3] Monge, A. E.; Elkan, P.C.: The Field Matching Problem: Algorithms and Applications. Proc. 2nd Intl. Conf. Knowledge Discovery and Data Mining (KDD), 1996.
[4] Naumann, F. and Leser, U. Informationsintegration: Architekturen und Methoden zur Integration verteilter und heterogener Datenquellen. Dpunkt Verlag, 1. Edition, Oktober 2007.
[5] Chaudhuri, S., Dayal, U.: An Overview of Data Warehousing and OLAP Technology. ACM SIGMOD Record 26(1), 1997.
[6] Helmis, S. and Hollmann, R. Web-based data integration, approaches to measuring and securing the quality of information in heterogeneous data sets using a fully web-based tool, 1. edition © Vieweg+Teubner | GWV Fachverlage GmbH, Wiesbaden 2009.
[7] Jarke, M., Lenzerini, M., Vassiliou, Y., Vassiliadis, P.: Fundamentals of Data Warehouses. Springer, 2000
[8] Abiteboul, S.; Clue, S.; Milo, T.; Mogilevsky, P.; Simeon, J.: Tools for Data Translation and Integration. In [26]:3-8, 1999.
[9] Raman, V.; Hellerstein, J.M.: Potter's Wheel: An Interactive Framework for Data Cleaning. Working Paper, 1999. http://www.cs.berkeley.edu/~rshankar/papers/pwheel.pdf.
[10] Rahm, E. and Do, H.H. Data cleaning: Problems and current approaches. IEEE Bulletin of the Technical Committee on Data Engineering, 23(4), 2000.
[11] Tork Roth, M.; Schwarz, P.M.: Don't Scrap It, Wrap It! A Wrapper Architecture for Legacy Data Sources. Proc. 23rd VLDB, 1997.
[12] Wiederhold, G.: Mediators in the Architecture of Future Information Systems. Computer 25(3): 38-49, 1992
[13] Galhardas, H.; Florescu, D.; Shasha, D.; Simon, E.: Declaratively cleaning your data using AJAX. In Journees Bases de Donnees, Oct. 2000. http://caravel.inria.fr/~galharda/BDA.ps.
[14] Galhardas, H.; Florescu, D.; Shasha, D.; Simon, E.: AJAX: An Extensible Data Cleaning Tool. Proc. ACM SIGMOD Conf., p. 590, 2000
[15] Hernandez, M.A.; Stolfo, S.J.: Real-World Data is Dirty: Data Cleansing and the Merge/Purge Problem. Data Mining and Knowledge Discovery 2(1):9-37, 1998
[16] Bernstein, P.A.; Dayal, U.: An Overview of Repository Technology. Proc. 20th VLDB, 199
[17] R. Y. Wang, V. C. Storey, and C. P. Firth, "A framework for analysis of data quality research," IEEE Transactions on Knowledge and Data Engineering, vol. 7, no. 4, pp. 623–640, 1995.
[18] E. Pierce, "Assessing data quality with control matrices," Communications of the ACM, vol. 47, no. 2, pp.82–86, 2004
[19] Lee, M.L.; Lu, H.; Ling, T.W.; Ko, Y.T.: Cleansing Data for Mining and Warehousing. Proc. 10th Intl. Conf. Database and Expert Systems Applications (DEXA), 1999.
[20] E. Pierce, "Assessing data quality with control matrices," Communications of the ACM, vol. 47, no. 2, pp. 82–86, 2004
[21] Milo, T.; Zohar, S.: Using Schema Matching to Simplify Heterogeneous Data Translation. Proc. 24th VLDB, 1998.
[22] Monge, A. E. Matching Algorithm within a Duplicate Detection System. IEEE Techn. Bulletin Data Engineering 23 (4), 2000 (this issue).
[23] Monge, A. E.; Elkan, P.C.: The Field Matching Problem: Algorithms and Applications. Proc. 2nd Intl. Conf. Knowledge Discovery and Data Mining (KDD), 1996.
[24] Savasere, A.; Omiecinski, E.; Navathe, S.: An Efficient Algorithm for Mining Association Rules in Large Databases. Proc. 21st VLDB, 1995.
[25] Srikant, R.; Agrawal, R.: Mining Generalized Association Rules. Proc. 21st VLDB conf., 1995



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY