Critical Data Consolidation in MDM to Develop the Unified Version of Truth

Ms. Dupinder Kaur*, Dr. Dilbag Singh
Department of Computer Science and Engineering
Chaudhary Devi Lal University, Sirsa, India

***Abstract*—Organization seeking growth and competitive lead should use Master Data Management (MDM) as a foundation for efficient decision making. An MDM framework creates a trusted and reliable continuous record of customers, products, suppliers and other shared data sets. In master data, the critical data is consolidated to portray essential business entities into a Unified version of Truth. To create trusted view of master data challenges like quality, identity resolution, analytics and investment are faced. In proposed research, a technique has been designed to generate Master Data to assist the policy maker to address the said issues. In this paper, four steps have been taken for master data creation namely: Data Enrichment, Data Matching, Data Merging and Data Governance. To achieve legitimate data quality TALEND open studio has been used for data pre-processing and enrichment. An algorithm is designed to match and merge the master records. To validate the designed approach, results are evaluated using Pandas Data Frame on Python platform. This paper will assist the policy makers of the organizations in formulating the business strategies.**

***Keywords*—Master data management (MDM); master record; TALEND; data matching and merging**

I. INTRODUCTION

Data Management is concerned, with the entire lifecycle of a data resource from the creation to retirement, with advances and changes. Data Management comprises the procedures, practices, ideas and measures for proficient utilization of data resources. An enlargement of the data size, in many folds, makes the organizations to adopt the data management practices to maintain the quality. The transformation, integration and cleaning of data is needed for efficient handling of data [1]. The new challenge for maintaining the quality of most essential assets of business demands Master Data Management (MDM) program. MDM comprises of technologies, processes and disciplines to incorporate the cleaning, administration and controlling all shared data assets. An MDM arrangement formulates a single accurate set of data populated across different frameworks by ensuring the consistency and accuracy of organization's shared data assets. [2]. The core benefits of MDM includes informed decision making, reducing data duplication, better data compliance and handling change requests.

Master data deals with the critical data entities of a company such as clients, items and resources that are shared across value-based applications. It additionally enables the development of a 'single version of truth' [3]. The technical operations assisted by MDM activities include Data quality improvement, Master Data creation and Data engineering [4]. MDM is significant for a wide range of applications to make a complete beginning to end plan that drives progression and achieves better business results. A viable MDM execution enables better usage of basic data present in organization [5]. Relevant to Master data Management, various data taxonomies used in an enterprise are:

- Transactional Data: The inner or outer exchange or transaction that happen including sales, orders, purchase orders, card payment, etc.
- Reference Data: It addresses set of qualities that are referred by frameworks, applications, information store just as by transaction and master data, such as status codes, state contractions, segment fields and so forth [6].
- Reporting Data: The aggregated data compile for the purpose of analytic and reporting. For example: order status (Accept or Reject).
- Meta Data: The data that describe label or characterize other data. For example: properties of media file: its size, type, resolution and author, etc.
- Master Data: The persistent, non transactional data that defines primary business entities such as customer, product, employees and inventory. Master data is the unified version of common data that are often duplicated across the enterprise. The figure below depicts various categories of Master Data [7].

Master Data of an organization can include all entities involved in financial structure, transactions done by organization or locations such as address. If master data quality does not meet expectations, it can affect the efficiency of business operations [9]. Thus, there is need to manage this data. Master Data Management ensures complete, consistent, and accurate data in different areas of organization's activities MDM includes process of data collection, cleansing, consolidation and distribution in the organization ensuring the control of use in various analytical and operational applications [10]. The underlying driver for execution of MDM is to deal with data quality issue frequently emerges across data entities and information systems [13]. MDM is implemented in most of organization with aim to guarantee that master data includes reference details that reflects the present state of business [11].

*Corresponding Author.

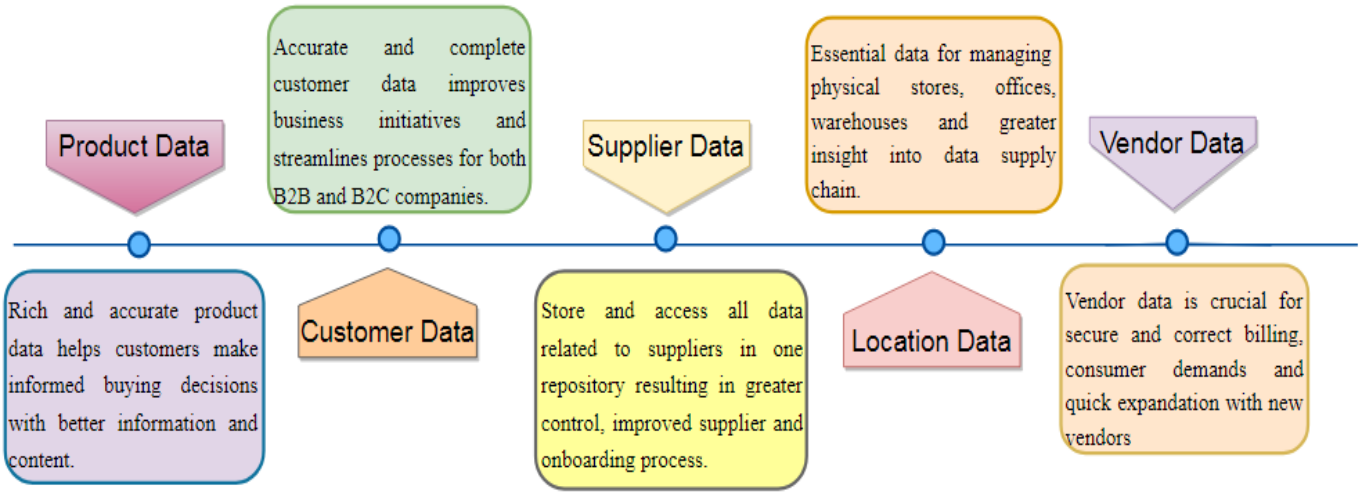

Fig. 1. Categories of Master Data [9].

Fig. 1 illustrates various categories of Master data. On the basis of need of an enterprise, Master data can be an employee data, party data, asset data or ledger data [8].

A Master Data Management program may provide an option to the industry with new ways to handle the data quality issues that the industry has struggled for few years and help prevent the "information rich and data poor" dilemma [14].

The main activities required for implementing MDM in an organization consist of following activities:

- Discover the need of master data management in organization and reference data.
- Categorize the sources and contributors of data.
- Define and maintain data architecture [12].

In light of the examination and investigation, following are the major contributions of this paper.

- To study the management, categories and related terms of Master Data.
- Designing of an algorithm for generating Master Record.
- Estimation and fixing of Data Quality using TALEND MDM tool.
- Use of Pandas Data Frame for creating Master Record.

The remaining part of the paper is organized as follows:

In Section 2, the related literature is reviewed. Section 3 elaborates the research methodology used for present research work. Section 4 is about the design of the proposed algorithm for generating Master records. Section 5 presents the Input dataset. Section 6 shows implementation and results of the proposed algorithm. Last section discusses the findings of implementation of the proposed method for creating qualitative and accurate master data to assist the organizations in decision making.

II. Literature Review

Radachirkova [1] et al. analyzed the usability of the existing data transformation tools for their utilization to achieve the desired quality characteristics in business measures. The focus was to enhance the data quality using existing tools instead of adopting a novel data transformation procedure for every data analytics task. Tending to this issue; Data cleaning, migration and transformation tools were summarized as black box procedure by exposing some properties such as applicability requirements, portion of data modified and constraints satisfied by data over applied procedure. The formal study revealed that these primary outcomes could be applied for accomplishing desired data quality outcomes in data analytics.

PanagiotisLepeniotis [2] integrated the MDM platform with Business Transformation Programme for decision making. The examination revealed that the Management of the Master Data and the never-ending confirmation of the Data Quality are crucial for any organization despite of having a BTP or not. An MDM impacted BTP decision model has been presented in the research. On the basis of case study audits and interviews, the research identified an improved indulgent into decision making process of a BTP concerning MDM and the way, these decisions affect the fruitful execution of a BTP.

Fernando Gualo [3] et al. assessed the "Functional Suitability" of MDM applications by considering useful necessities from section 100 to 140 of ISO 8000 and proposed a solution by considering the appraisal and affirmation of Functional Suitability of MDM applications. In addition to basic requirements, test cases are also designed required for the evaluation. In order to infer the prerequisites from ISO 8000, all the parts are covered except of ISO 8000-115, ISO 8000-116 and ISO 8000-150. Application of assessment method for existing master data based application is also featured.

Shreya Mrigen [4] et al. proposed a method for singular data management in pharmaceutical company. Hierarchies and type of problem faced in managing MDM solution in

pharmaceutical company has been examined in this research. The examination revealed that MDM is an essential activity for creating singularity, improving data consistency and data processing and market examination in pharmaceutical companies.

Dilbag Singh [5] et al. provided an MDM solution for building frameworks. A strategy to pilot the implementation process of MDM has been provided with a Framework, Roadmap and a DFD design. This strategy described complete description of step wise approach to have a better view point on generating, handling, validating and monitoring master data. A study on existing work and Gartner's Hype cycle has been performed to know the latest technical trends in MDM.

Chun Zhao [6] et al. designed a model for assessing the viability and rationality of master data network. Set Pair Analysis (SPA) design of MDM has been presented in the research. Data network based on master data and data keys are established using MDM system. In association with master data, the main concern was to assess the effectiveness and rationality of the network. Contextual analysis showed convenient update of information, distribution and active response are significant elements in cloud fabricating climate.

On the basis of literature review, it can be concluded that designing of policies and standards are required to evaluate the functional suitability of MDM based applications. Further, the identity resolution, business rules and MDM solution is required for data consistency and security. Data governance and stewardship is of utmost importance for managing the data.

III. Research Methodology

Identity resolution technique is significant in MDM as it integrates two or more data identities into one object. To allow data citizens to access the right information, a productive identity resolution technique is required. In MDM, Identity resolution finally determines the master record. To realize this need, a methodology has been proposed in the present research for identity resolution in MDM. Research begins with exploring, analyzing and enrichment of input data set on TALEND tool for legitimate quality of data. The characteristics of relevant attributes of input data sets like matching score, threshold value and merging condition are considered. Finally, the experimental study has been performed to generate Master Table using Pandas data frame. Hence, an exploratory, descriptive and experimental research methodologies are used in this research. This concrete motivation for generating an MDM solution arises from everyday needs of global identification, linking and synchronization of master data across heterogeneous data sources.

IV. Proposed Algorithm for Consolidating Critical Data Entities

In the present research; data is collected from various sources in different formats, therefore, the consolidation has been carried out. Consolidation amalgamates all the data, remove redundancies and inaccuracy before combining it into single place. Critical data consolidation enables 360-degree-view of data assets, efficient plan, implementation and execution of a business process. Data consolidation in MDM initiates with data collection from the significant sources, utilizing business rules to build up a unique data source, data governance and transmission to the concerned departments. To consolidate critical data entities over collected data set, an algorithm has been proposed in this research as shown in Fig. 2.

Input: Data Set collection from relevant sources.
Output: Master Table - Unified version of Truth.
Steps:

1. Collect data from multiple sources
2. Build a standardized data model for integration as per data set requirements.
3. Evaluates data quality with predefined standards and polices.
4. If data quality is not legitimate as per defined rules, do
5. Apply data cleaning and standardization techniques
6. End if
7. Design a model for resultant Master Table
8. Select the desired attributes to be matched from input data files.
9. For each record in input do
10. Apply fuzzy matching technique on corresponding attributes of input files
11. Set m = similarity score of attributes
12. Set k= threshold value
13. If m>=k
14. Set the record for best match and automatic merge.
15. End if
16. If m>=0 && m<k
17. Set the record for manual merge using data stewardship
18. Else
19. If m<0
20. Set the record as worst match and not pertinent for merge.
21. End if
22. End if
23. End for
24. Generate Master Table as T
25. Perform automatic and manual merge of records into T
26. Select a record as Instance I
27. If I is new record
28. Add it to T by following step 3.
29. Else
30. Repeat step 8
31. Apply Data Governance technique on T
32. Deploy T as final master table: Unified version of truth

Fig. 2. Algorithm for Generating Master Data through Consolidation.

The above algorithm is proposed for creating master data. Initially, data is gathered from heterogeneous sources. On the basis of the data set requirements, Data Model-1 is created. The quality of consolidated data is verified by keeping into account the policies and rules of the organization concerned. In case the data is adequate, it will be processed further, failing which quality improvements techniques like Data Cleaning, Enrichment or scrubbing are applied to ensure the data consistency, accuracy and timeliness. Once the data is

enriched, the attributes to be matched are selected from the input files. For each record, a similarity score (m) is generated. The corresponding attributes are matched and a score is generated by using fuzzy matcher. A threshold value (k) is selected on the basis of matching score for categorizing the matching and merging process. The attributes of the input files are compared to find out the match. For each record, a matching score is calculated by applying fuzzy matcher.

On the basis of matching score, a threshold value is selected. Matching score of a record, greater than or equal to threshold value, will lead to the best match and the data is automatically merged. However, if matching score is greater than or equals to zero or less than threshold, it is considered as average match and processed for manual merge using Data stewardship. For worst match, matching score is negative and thus the data is not considered for a merge. In the proposed strategy, master table is generated by applying automatic and manual merge. In case there is a new record, it will be directly added to Master table and if a record already exists in the table then the entire process of matching and merging is again carried out. Data Governance polices are applied continuously on master data to ensure the quality of the data.

The resultant Master record should correctly match and merge using identity Resolution technique. Variation in data will adversely affect the search process and quality of data. For example: A Person may use Ritesh as a name at one place but Ritesh Kumar or Retish K. at another places. The variation in names may be due to use of nick names sometimes, aliases or initials. Change in address of roads, areas, billings, mailing etc may lead to variations. Such variations are overcome while merging the data using proposed algorithm. Table I describes range of variations for designed Identity resolution.

Table I explains the range of Variation for identity Resolution using designed algorithm. Identity matching allows correct result focusing on standard and quality of key data elements such as address, phone and email address. It allows enrichment of customer profiles to improve the accuracy of matching.

TABLE I. RANGE OF VARIATION FOR DESIGNED IDENTITY RESOLUTION TECHNIQUE

Type of Data	Potential Variants on a Data Type	Priority
Name Variation	Ashwani, Ashwini, Ash, Aswani	High
Abbreviation	Mohammad, Muhammad, Mohd, Mhd	High
Phonetics /Spelling Variation	Ritesh, Reetesh, Ritish	High
Date Format	9/01/2020, 09/01/2020, Jan 09, 2020, 09 Jan 2020, 01/09/2020	High
Suffix Variation	Ranjeet Singh, Ranjeet S. , Ranjeet	Medium
Null Values	Not known, Unknown, ?, 000, N/A, [Blank],NaN	Medium
Organization Name	Chaudhary Devi lal University, CDLU, Chaudhary devilal University	Medium
Department Name	Department of Computer Science & Engineering, DCSE, Deptt. of Compt. Sc. &Engg.	Medium
Titles	Mr. Pawan, Dr. Pawan, Pawan MD	Low

V. INPUT DATASETS

To implement the proposed work, data set has been collected from University Computer Science & Engineering department, Library section and Account section. The information of student is submitted to department during admission of a course while a separate form is filled by student for issuing the books in Library. As for master data, two or more sources of same information is required The figure below represents the screenshot of input data files.

	D_Reg_No	D_Name	D_Father_Name	D_Email_Id	Subject	Category	Fees	D_Contact
0	130032	MOHIT	VED PARKASH	rrmohit77@gmail.com	MCA-32	GENERAL	8230 Rs/-	96******0229
1	130034	AKSHAY BANSAL	MR.KULBHUSHAN BANSAL	akbansal77@gmail.com	MCA-32	NaN	8230	99******0230
2	130035	PARDEEP KUAMR	JOGENDER SINGH	kumarpar258@gmail.com	MCA-32	SC	SC	99******0231
3	130036	PALLAVI	MR PAWAN KUMAR	pallavi1992@gmail.com	MCA-32	SC	750/-	92******0232
4	130036	PALLAVI	MR PAWAN KUMAR	pallavi1992@gmail.com	NaN	NaN	750/-	93******0233
5	130037	kawal Preet	Balraj Singh	singhkawal97@gmail.com	MCA-32	SC	750/-	99******0234

Fig. 3. Dataset1- Input_Flow_1 File.

Fig. 3 shows the input dataset1 as Input_flow_1 file containing the data collected from department. The attributes D_Reg_no, D_Name, D_Father_Name represents the Registration No, Name and Father Name of student as recorded in department and so on.

	L_Reg_No	L_Name	L_Father_Name	L_Email_id	L_Contact
0	130004	Amritpal	Tehal Singh	dhaliwal19362@gmail.com	79******0202
1	130032	MOHIT	VED PARKASH	rrmohit77@gmail.com	96******0229
2	130031	SUMANDEEP KAUR	LAKHWINDER S.	KAURDA061@GMAIL.COM	75******0228
3	130037	Kawal Preet	BalraJ SINGH	singhkawal97@gmail.com	99******0234
4	130016	Monika Verma	Stapal Singh Verma	monuverma1577@gmail.com	99******0213
5	130030	Vikas	Jitender	vikurampura@gmail.com	77******0227

Fig. 4. Dataset1- Input_Flow_2 File.

Fig. 4 describes second dataset as Input_File_2 taken from library. The attributes L_Reg_no, L_Name, L_Father_Name represents the Registration No, Name and Father Name of student as recorded in Library and so on.

VI. IMPLEMENTATION AND RESULTS

The schema diagram designed for present work comprises of four steps: Data Enrichment, Data Matching, Data Merging and Data Governance as shown below.

Fig. 5 illustrates the schema design for data consolidation process in MDM. A sequential flow of execution from Data enrichment to Data Governance has been employed in this research.

For the implementation, the quality of input data is verified first. As Fig. 3 addresses the problem of Data redundancy (Duplicate records for Registration No 130036), domain integrity constraint violation ('SC' value in FEES attribute) and incompleteness (Null entries in Subject and Category attribute). Before creating Master data, all these issues must be resolved to make data clean. Thus, in present research, data enrichment is performed on TALEND tool. The following sections describe the modules Data Enrichment, Data Matching, Data Merging and Data Governance in detail.

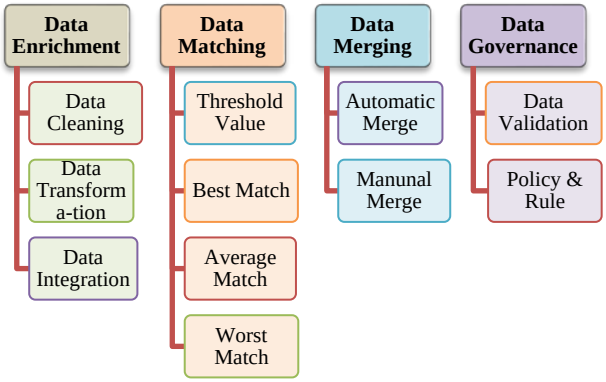

Fig. 5. Schmea for Methodology used in Research.

A. Module 1: Data Enrichment

Data Enrichment is characterized as enhancing or appending the collected data by supplementing incomplete, missing or inaccurate data. The subsequent enriched data empowers organizations in effectively customizing the data. To implement Data Enrichment, TALEND open studio for MDM has been used in present research as shown below.

Fig. 6 illustrates TALEND window with two input data files as Dept_data and Lib_data taken from University department and library. The results of TALEND tool are shown through Python platform for bringing together the results with further implemented result.

1) Data cleaning: Data cleaning is the process of removing or fixing inaccurate, corrupt, outdated, and incorrect formatted, duplicate entries from data sets. Data Cleaning is performed on input files Dept_data and Lib_data using predefined components in TALEND. The following figure represents the screenshot of Clean Data in Output_Flow_1 file.

Fig. 7 signifies that Unique Row and Replacement components are used for removing redundancy; null values, inaccurate and domain integrity constrain violation on input data source. The contents of Output_Flow _1 file shows that the above said problems are resolved in this step. Similarly, the same procedure is applied on library data and clan data is stored in Output_Flow_2 file.

2) Data transformation: The way toward changing over data from one format or construction into another is known as Data Transformation. In the present research, Python language has been used to execute the MDM solution. Python, being case sensitive language; lower case sentences are required to be converted into upper case sentences. Hence, data transformation is performed under mapping section of TALEND tool.

Fig. 8 indicates the Mapping process performing both Data Transformation and Integration. A mapping component takes two input as Main (department) and Lookup (Library) as shown above. The records are mapped on the basis of registration number of student. The attributes: Student Name, Father Name and Email id are converted into upper case under this section.

3) Data integration: Data Integration refers to the process of consolidating data from multiple sources into single or unified view. This technique helps analytic tools to generate effective and actionable business intelligence.

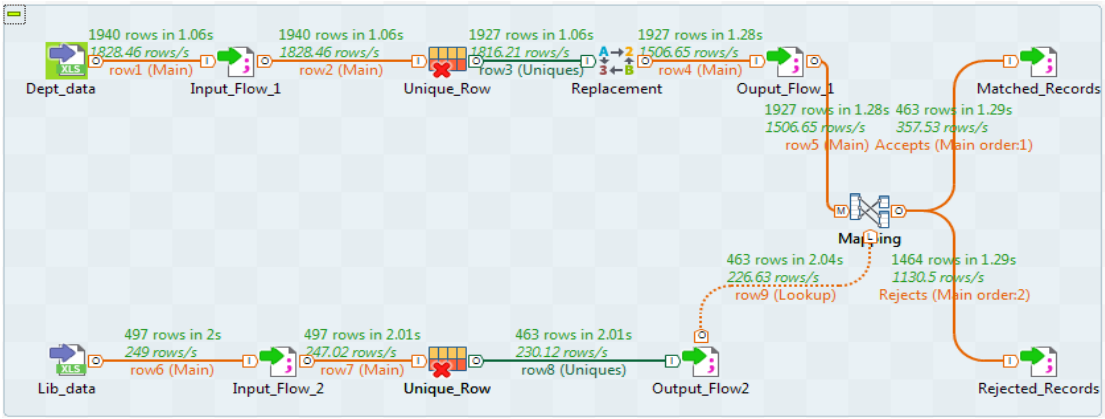

Fig. 6. Data Enrichment using TALEND.

	D_Reg_No	D_Name	D_Father_Name	D_Email_Id	Subject	Category	Fees	D_Contact
0	130032	MOHIT	VED PARKASH	rrmohit77@gmail.com	MCA-32	GEN	8230.0	96******0229
1	130034	AKSHAY BANSAL	MR.KULBHUSHAN BANSAL	akbansal77@gmail.com	MCA-32	GEN	8230.0	99******0230
2	130035	PARDEEP KUAMR	JOGENDER SINGH	kumarpar258@gmail.com	MCA-32	SC	750.0	99******0231
3	130036	PALLAVI	MR PAWAN KUMAR	pallavi1992@gmail.com	MCA-32	SC	750.0	92******0232
4	130037	kawal Preet	Balraj Singh	singhkawal97@gmail.com	MCA-32	SC	750.0	99******0234
5	130038	SUNIL KUMAR	KRISHAN KUMAR	kumarsunil123@gmail.com	MCA-32	SC	750.0	89******0235

Fig. 7. Clean Dataset- Output_Flow_1.

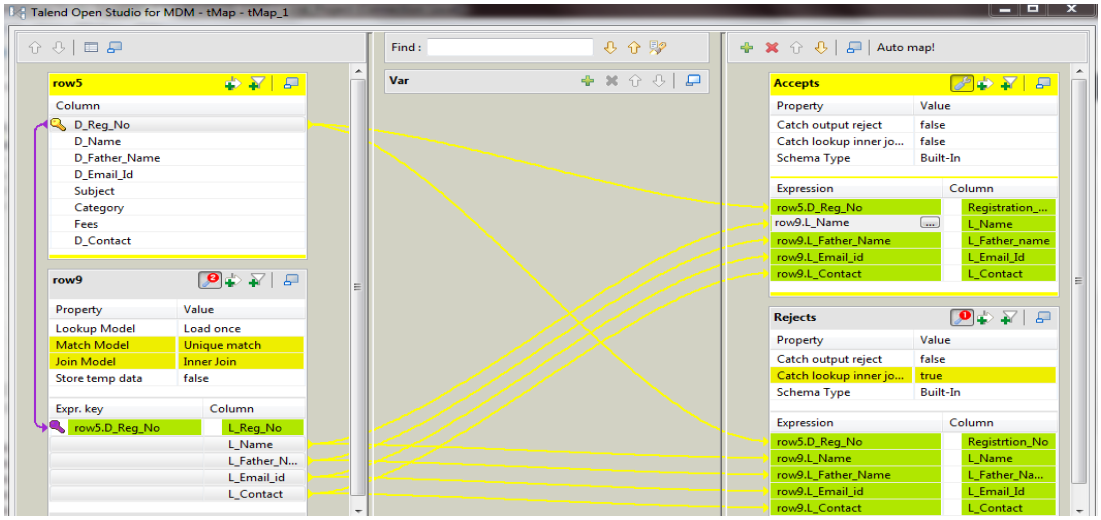

Fig. 8. Data Transformation using Talend Mapping Window

Fig. 9 describes the integrated result of mapping component as Matched records common to both department and library. In mapping, department and library records are mapped to generate a single copy of matched records by considering "Registration No". In above table, the attribute: registration number taken from department, whereas Name, father Name, Email and contact are taken from library record. The matched records are common records to library and department. Rejected records are not taken for analysis in present study.

Fig. 10 shows the Rejected records. The records of student which are not present in the library are rejected.

B. Module 2: Data Matching

Matching refers to the method of comparing different data sets and matching them against each other. The objective is to find the data that refers to same entity. To implement matching process, Python Fuzzy Match library is used to evaluate the matching score of records. A fuzzy match represents a match that is not exact. This technique identifies two elements of string, text or entities that are approximately same but not exact. In order to find a match score, four attributes: Name, Father Name, Email Id and Contact are considered for comparison in present research. Following values are evaluated on the basis of matching score of records.

1) ThresholdValue: Threshold value is considered on the basis of maximum matching of attributes. The Figure below illustrates the threshold value taken in the present research for merging of records in Final master Table.

In Fig. 11, the attribute best_match_score represents the matching score of the records. Matching score value 0.561963 is taken as threshold value. The records above or on threshold are considered as best match whereas below threshold and positive match score are average matched.

	Registration_No	L_Name	L_Father_name	L_Email_Id	L_Contact
0	130029	KAVITA	SUNDER LAL	KAVITABISHNOI97209@GMAIL.COM	89******0226
1	130030	VIKAS	JITENDER	VIKURAMPURA@GMAIL.COM	77******0227
2	130031	SUMANDEEP KAUR	LAKHWINDER S.	KAURDA061@GMAIL.COM	75******0228
3	130032	MOHIT	VED PARKASH	RRMOHIT77@GMAIL.COM	96******0229
4	130037	KAWAL PREET	BALRAJ SINGH	SINGHKAWAL97@GMAIL.COM	99******0234
5	130042	ASHOK KUMAR	MAHENDER SINGH	AGODARA7711@GMAIL.COM	96******0239

Fig. 9. Integrated Matched Records after Mapping.

	Registrtion_No	L_Name	L_Father_Name	L_Email_Id	L_Contact
0	130034	NaN	NaN	NaN	NaN
1	130035	NaN	NaN	NaN	NaN
2	130036	NaN	NaN	NaN	NaN
3	130038	NaN	NaN	NaN	NaN
4	130039	NaN	NaN	NaN	NaN
5	130040	NaN	NaN	NaN	NaN

Fig. 10. Rejected Records after Mapping.

	best_match_score	D_Name	L_Name	D_Father_Name	L_Father_Name	D_Email_Id	L_Email_Id	D_(
775	0.596529	MAMTA RANI	MAMTA	BHOOP SINGH	BHOOP SINGH	MAMTAK1207@GMAIL.COM	MAMTAK1207@GMAIL.COM	98***
170	0.593535	PARDEEP KAUR	PARDEEP KAUR	KULVINDER SINGH	KULWINDER SINGH	PARDEEPKAUR1313RD@GMAIL.COM	PARDEEPKAUR1313RD@GMAIL.COM	93***
243	0.561963	PRIYANKA RANI	PRIYANKA	KRISHAN KUMAR	KRISHAN KUMAR	PRIYAKAMBOJABC@GMAIL.COM	PRIYAKAMBOJABC@GMAIL.COM	79***
997	0.561095	GAURAV WADHWA	GAURAV WADHWA	PREK KUMAR	PREK KUMAR	GAURAV99@GMAIL.COM	GORUWADHWAW1@GMAIL.COM	81***
999	0.537010	LALITA DEVI	LALITA DEVI	OM PARKASH	OM PARKASH	LALTA89@GMAIL.COM	LALLTINEW01@GMAIL.COM	80***

Fig. 11. Data on the basis of Threshold Value.

2) Best match: The record for which the matching scores value is greater than or equal to 0.561963 are categorized as best match records as shown below.

Fig. 12 shows the best matched records. The highest matching score is 0.947167. All the records above than or equal to threshold value are categorized as best matched records.

3) Average match: The record for which the matching scores value is greater than zero but less than 0.561963 are categorized as average match records as shown below.

Fig. 13 represents the screenshot of records below threshold value and greater than zero. There is need to manually determine which records are to be merged. Hence, these are categorized as Average matched records.

4) Worst match: The record for which the matching scores value is less than zero are categorized as worst match records as shown below.

Fig. 14 describes the worst matched records. All the records with negative matching score are considered as worst match records.

C. Module 3: Data Merging

Merging process allows merging the source data record into destination data record. Two base object records are merged to form a single consolidated base object. For merge operation, the attributes: Name, Father Name, Subject, Category, Fees are taken from department table whereas Email id, Contact are taken from library table.

Considering, k=Threshold Value, m=Matching score

Merge Condition =
- m>=k — Best Match → Automatic Merge
- m>0&&m<k — Average Match → Manual Merge
- m<=0 — Worst Match → Rejected

1) Automatic merge: The records for which matching scores is greater than or equal to threshold value are categorized as best match and considered for Automatic merge. Automatic merge is simply performed by merging corresponding attributes from two datasets. The result is Final Master Table containing Master record or Unified version of Truth for each record.

2) Manual merge: The records for which matching score is below than threshold value or greater than zero are categorized under average match and considered for manual merge. Data Steward helps to manually determine which records should be considered for a match. A data steward is accountable for carrying out data usage and security policies as determined through organization. The resultant records are stored into Final Master Table.

Fig. 15 depicts the Master Table after applying data enrichment, matching and merging. The content of the master table represents the Golden copy of each record stored in a single table. The attributes of table D_Name, D_father_Name, Subject, category, Fees are taken from department and L_Email_Id, L_Contact are taken from library. Thus, it represents a unified version of truth consolidated from multiple sites. It is essential to adopt an MDM strategy as master data represents the most valuable business objects and agreed upon the information that is shared among the organization.

D. Module 4: Data Governance

Data Governance is the set of policies, roles and standards that guarantees high quality throughout the lifecycle of data. It is the process of governing how data is used, by whom and when. Data Governance ensures that data accessibility, security and management rules are followed consistently, every working day. Services provide by data governance are: aligning rules and policies, establishing accountability and decision rights, specifying data quality requirements and performing data stewardship. An optimized data governance program will underpin the business transformation toward operating on digital platform at many levels. MDM requires Data governance activities for agreement on business and technical resources, data to be mastered, business rules and policies, consolidation rules and data quality. Thus, implementation of MDM is irrelevant without doing the significant bit of data governance.

1) Data validation: The purpose of Data Validation is to envelope accuracy: "The degree of similarity of an action to a standard or a genuine worth" and validity: “the degree to which data conforms to defined business rules”. It ensures that information present in the system is correct. Thus, in this study, the scope of Data validation is specified to figures out what data to validate, when to validate data and how to manage data that fails validation. Thus, Data Validation being an integral part of governance improves data quality and efficiency of the system.

	best_match_score	D_Name	L_Name	D_Father_Name	L_Father_Name	D_Email_Id	L_Email_Id	D_Conta
840	0.947167	RANJEET SINGH	RANJEET SINGH	HARPAL SINGH	HARPAL SINGH	BTECH.SULEKHA@GMAIL.COM	BTECH.SULEKHA@GMAIL.COM	99******02
1001	0.929684	ASHWANI AHUJA	ASHWANI AHUJA	DILBAG RAI	DILBAG RAI	ASHWANIAHUJA2009@GMAIL.COM	ASHWANIAHUJA2009@GMAIL.COM	90******02
116	0.920330	SAHIL NARANG	SAHIL NARANG	SHIV DAYAL NARANG	SHIV DYAL NARANG	SAHILN321@GMAIL.COM	SAHILN321@GMAIL.COM	95******02
1000	0.901804	VIJAY KUMAR	VIJAY KUMAR	BACHANA RAM	BACHANA RAM	VIJAYVAID77@GMAIL.COM	VIJAYVAID77@GMAIL.COM	89******02
838	0.896942	DIKSHA JASUJA	DIKSHA JASUJA	BHAGWAN DASS	BHAGWAN DASS	ABHIMONARK@GMAIL.COM	ABHIMONARK@GMAIL.COM	98******02

Fig. 12. Best Match Records.

	best_match_score	D_Name	L_Name	D_Father_Name	L_Father_Name	D_Email_Id	L_Email_Id	D_Cont
243	0.561963	PRIYANKA RANI	PRIYANKA	KRISHAN KUMAR	KRISHAN KUMAR	PRIYAKAMBOJABC@GMAIL.COM	PRIYAKAMBOJABC@GMAIL.COM	79******0
997	0.561095	GAURAV WADHWA	GAURAV WADHWA	PREK KUMAR	PREK KUMAR	GAURAV99@GMAIL.COM	GORUWADHWAW1@GMAIL.COM	81******0
999	0.537010	LALITA DEVI	LALITA DEVI	OM PARKASH	OM PARKASH	LALTA89@GMAIL.COM	LALLTINEW01@GMAIL.COM	80******0
225	0.476170	SAPNA RANI	SAPNA RANI	JARNAIL SINGH	JARNAIL SINGH	SAPNAKAMBOJ9696@GMAIL.COM	SAPNAKAMBOJ9696@GMAIL.COM	93******
998	0.463360	RAJENDER SINGH	RAJENDER SINGH	SHER SINGH	SHER SINGH	RAJSINGHSERA56@GMAIL.COM	RAJUSINGH77@GMAIL.COM	82******0

Fig. 13. Average Match Records.

	best_match_score	D_Name	L_Name	D_Father_Name	L_Father_Name	D_Email_Id	L_Email_Id	D_Con
784	-0.517985	NITIN KUMAR SACHDEVA	RITESH KUMAR	BINDER SACHDEVA	RAVINDER	BAJAJ7201@GMAIL.COM	RITESHKUMAR66@GMAIL.COM	99******0
672	-0.540876	SUNDER	ASHOK KUMAR	KAILASH	MAHENDER SINGH	SUNDER2233@GMAIL.COM	AGODARA7711@GMAIL.COM	88******0
1064	-0.620879	SOUJANYA UPPAL	AJAY	RAKESH UPPAL	RAMESH KUMAR	SOUJANYAUPPAL99@GMAIL.COM	WAYASIJA23@GMAIL.COM	92******0
62	-0.665611	MOHD SOAIB	POOJA	MOHD ILYAS	JAGDISH KUMAR	SOABMOHD9191@GMMIL.COM	POOJAWALECHA490@GMAIL.COM	75******0
355	-0.695725	AKSHAY BANSAL	MOHIT	MR.KULBHUSHAN BANSAL	VED PARKASH	AKBANSAL77@GMAIL.COM	RRMOHIT77@GMAIL.COM	99******0

Fig. 14. Worst Match RecordsBest Match Records.

	D_Name	D_Father_Name	Subject	Category	Fees	L_Email_Id	L_Contact
0	PRIYANKA	DINESH KUMAR	MCA-32	GEN	8230	AGYATT@GMAIL.COM	90******0201
1	POOJA	JAGDISH KUMAR	MCA-32	GEN	8230	POOJAWALECHA490@GMAIL.COM	77******0207
2	SAHIL NARANG	SHIV DAYAL NARANG	MCA-32	GEN	8230	SAHILN321@GMAIL.COM	95******0210
3	MONIKA VERMA	SATPAL SINGH VERMA	MCA-32	GEN	8230	MONUVERMA1577@GMAIL.COM	99******0213
4	PARDEEP KAUR	KULVINDER SINGH	MCA-32	GEN	8230	PARDEEPKAUR1313RD@GMAIL.COM	93******0214
5	KARAMJEET KAUR	JAGJEET SINGH	MCA-32	GEN	8230	KHUNDAL07@GMAIL.COM	92******0215
6	MANPREET KAUR	RANJEET SINGH	MCA-32	SC	750	MK9821752@GMAIL.COM	94******0216
7	GURJASHAN SINGH	GURTEJ SINGH	MCA-32	GEN	8230	GURJASHAN80@GMAIL.COM	93******0223
8	AJAY	RAMESH KUMAR	MCA-32	GEN	8230	WAYASIJA23@GMAIL.COM	99******0225
9	VIKAS	JITENDER	MCA-32	GEN	8230	VIKURAMPURA@GMAIL.COM	77******0227

Fig. 15. Master Table.

2) Policies and rules: The significance of Data governance policy is tied straightforwardly to the significance of a solid data governance program. A committee or governance team establishes policies and rules over collection, storage, usage and security mechanism of organization's data programs. Following functions are articulated in this study for making a governance policy.

- Reliable, proficient and successful administration of the information resources all through the organization and over the long run.
- Designing of Laws and regulation specifically designed for organization's data program.
- Proper assurance and security levels for various classifications of information as set up by the governance team.

In present paper, an algorithm has been designed to create Master data, with input data sources containing number of records: 1940 and 497, respectively. The records of these data sets are matched to find the similarity score using fuzzy matcher. On the basis of score and threshold considered, the input data set is classified into three types as shown.

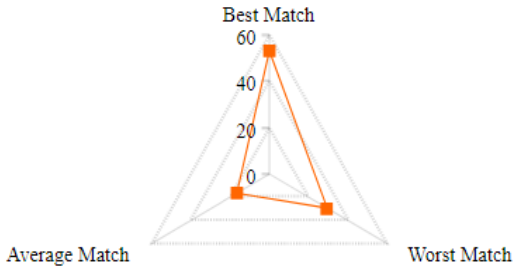

Fig. 16. Threshold based Classification of Input Data Sets.

The above Fig. 16 categories the input data set as: Best match, Average match and Worst match records. Matching is performed using fuzzy matcher to find a similarity score of a record. On the basis of match score, threshold value is considered to define the merge condition. The records above the threshold are taken as best matched records whereas below threshold and greater than zero are average matched records. Matching score with negative value are categorized as worst matched records. It is examined that the records in best, average and worst matching are: 53.2 %, 16.6% and 29.48% respectively for above said input. Master Table has been created using best and average matched records by automatic and manual merge process. It ultimately helps in improving data quality, business operation and analytics.

VII. Conclusion

In present study, an algorithm for critical data consolidation in MDM has been designed for small organizations. As Small and mid-sized organizations do not have the required capabilities to support best practices in MDM. They often underestimate complexity, cost, the flexibility to easily add attributes, and level of collaboration required for MDM program. Thus, to realize this target, a holistic approach for resolving and managing critical data entities in MDM has been presented in the present paper. In addition to data consolidation, Identity Resolution technique is also designed which helps in assessment of records while making "Unified version of truth". To discover and assess the quality of the identifying attributes, TALEND tool for MDM has been used in this research. Data quality over collected dataset is enriched with TALEND tool. By using fuzzy matcher, the matching score of corresponding attributes of input datasets is calculated. On the basis of threshold value, records are categorized as: Best match, Average match and Worst match. Best and average matched records are merged to generate Master Table while worst match records are rejected. Thus it is concluded that this approach is an optimal fit solution in small organization which enables users to integrate and circulate single standard view of master data across the frameworks like ERP, CRM, Apps, and systems. An effective MDM technique helps the business over wide variety of activities like reporting, up-selling and cross-selling of decision making and observance. The chances of being business successful, increases with significant implementation of master data.

References

[1] Rada chirkova, Jon Doyle, Juan Reutter, “Ensuring Data Readiness for Quality Requirements with Help from Procedure Reuse,” Journal of Data and Information Quality, ACM digital library, vol 13, issue 3,April 2021.

[2] Panagiotis Lepeniotis,“Master Data Management: Its importance and reasons for failed implementations,” Sheffield Hallam University, Ph.D thesis, Jan 2020.

[3] Fernando Gualo, Ismael Caballero, Moises Rodriguez, “Towards a software quality certification of master data-based applications,” Software Quality Journal, 28(3), 1019-1042, 2020.

[4] Shreya Mrigen, Dr. Vikram N B, “Relevance of Master Data Management in Pharmaceutical Industries,” International Journal for Research in Applied Science & Engineering Technology, 8(6), 190-197, 2020.

[5] Panagiotis Lepeniotis Master Data Management: Its importance and reasons for failed implementations, Sheffield Hallam University, Ph.D thesis, Jan 2020.

[6] Dilbag Singh, Dupinder kaur, “A Master Data Management solution for building frameworks: a constructive way to pilot the implementation,” in 2nd international conference on Data Analytics and Management (ICDAM), Springer 2021.

[7] Chun Zhao, Lei Ren, Ziqiao Zhang, Zihao Meng, “Master data management for manufacturing big data: a method of evaluation for data network,” World Wide Web, Springer, 23, 1407–1421, 2019.

[8] https://blog.semarchy.com/backtobasics_data_classification last accessed on 05/08/2021.

[9] https://www.stibosystems.com/what-is-master-data-management last accessed on 05/08/2021.

[10] Aditya Rahman A, Gusman Dharmaet al. “Master Data Management Maturity Assessment: A Case Study of a Pasar Rebo Public Hospital. 2019,” International Conference on Advanced Computer Science and information Systems (ICACSIS), IEEE, 497-504 Bali, Indonesia, 2019.

[11] Igor Prokhorov, Nikolai Kolesnik “Development of a master data consolidation system model (on the example of the banking sector),” Post proceedings of the 9th Annual International Conference on Biologically Inspired Cognitive Architectures, BICA, 142, 412-417, Elsevier, Prague, Czech Republic, 2018.

[12] F. G. Pratama et al.: Master Data Management Maturity Assessment: A Case Study of Organization in Ministry of Education and Culture. International Conference on Computer, Control, Informatics and its Applications (IC3INA), 1-6, IEEE, Tangerang, Indonesia, 2018.

[13] Z. Murti et al,: Master Data Management Planning: (Case Study of Personnel Information System at XYZ Institute). International Conference on Information Management and Technology (ICIMTech), 160-165, IEEE, Jakarta, 2018.

[14] S. Thomas Ng et al.: A Master Data Management Solution to Unlock the Value of Big Infrastructure Data for Smart, Sustainable and Resilient City Planning. Procedia Engineering, 196, 939-947, Elsevier 2017.