International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8
DOI: https://doi.org/10.17762/ijritcc.v11i8.7947
***Article Received**: 03 May 2023 **Revised**: 30 June 2023 **Accepted**: 26 July 2023*

Data Quality Optimization for Decision Making Using Ataccama Toolkit: A Sustainable Perspective

Aatif Jamal[1], Mohatesham Pasha Quadri[2], Mohd Rafeeq[3]
[1]Department of Computer Science & Information Technology
Maulana Azad National Urdu University
Hyderabad-500032
Email- aatif.jamal0017@gmail.com
[2]Department of Computer Science & Information Technology
Maulana Azad National Urdu University
Hyderabad-500032
Email- syedmohtesham21@gmail.com
[3] Department of Computer Science & Information Technology,
Maulana Azad National Urdu University,
Hyderabad, India – 500032,
Email: rafeeq@manuu.edu.in

Abstract- The world of internet has given us the access to explore the different domains of datasets and its usage. We have lots of heterogeneous data available on the digital platform which is meaningless unless we do not make the valuable use of it. What if we say that we can use these datasets in our need for the business requirements? The critical data can be delivered evenly and shared through master data management and integration techniques. However, given the cross-domain and heterogeneous nature of the data, it is still difficult to assess effectiveness and rationality. In this paper we have developed a pipeline using Ataccama and implemented the way of how we can yield the data, synthesize and optimize it using integration and Master data management (MDM) tools. These tools were assessed based on the performance characteristics and types of data quality problems addressed. We have tried to simplify the complexity and used various dictionaries and lookup along with the ruleset to fetch the required data from the dataset via the MDM application. Profiling the dataset and its validation based on different parameters. In results, it is found that the efficiency and the quality of data has been improved and optimized after using the integration techniques.

Keywords: Integration tool, MDM Tool, Data Quality, Optimization, Lookup, Profiling, Ataccama.

I. INTRODUCTION

We have a large volume of heterogeneous data that was generated by many departments and domains. The master data management strategy has been utilized to achieve data exchange and connection among supply chains [1]. The critical data can be delivered evenly and shared through master data management and integration techniques. However, given the cross-domain and heterogeneous nature of the data, it is still difficult to assess effectiveness and rationality [2]. Numerous mathematical techniques and optimization models have been presented in order to process big data and enhance the capabilities of big services in manufacturing environments [9].

However big data is dispersed throughout every stage of the manufacturing process, in numerous phases, and across multiple departments. Master data (MD) is the mapping of the reference data in the subsystems. The master data management approach has been utilized to achieve data exchange and connection among supply chains [2]. The critical data can be delivered evenly and shared through master data management and integration techniques. However, since these cross domains and heterogeneous data still lacks to evaluate the effectiveness and rationality is still a challenge [1].

Numerous mathematical techniques and optimization models have been presented in order to process big data and enhance the capabilities of big services in manufacturing environments. However, big data is dispersed throughout every stage of the manufacturing process, in numerous phases, and across numerous departments. The mapping of reference data in subsystems is known as master data and it can be used to represent business entities and connect cross-system business processes [1]. ETL tools are essential for supporting data integration plans because they let businesses collect data from various sources and combine it in a single, centralized location [10]. Depending on the data characteristics in the cloud manufacturing environment, the method of master data can be employed to manage the core data [11]. In order to actualize a complete and speedy supply chain deployment, the data of all phases in the cloud manufacturing environment can be opened.

***Article Received**: 03 May 2023 **Revised**: 30 June 2023 **Accepted**: 26 July 2023*

ETL tools also enable the collaboration of various data kinds. In addition, there are other more integration methods that are available, but most of them are comparatively less effective and lack data validation features [4]. So we have used the latest Integration cum MDM tool to optimize the data quality and to check its data validation percentage gain.

II. RELATED WORK

Zhao, et al. (2019) proposed a method for evaluating data network quality for master data management (MDM) in manufacturing big data environments. The authors claimed that a well-designed data network is essential for effective MDM, and existing approaches for evaluating data network quality are not sufficient for manufacturing big data environments. Their approach combines graph theory with domain-specific knowledge to evaluate the quality of data networks in terms of data availability, data consistency, and data reliability. Overall, the paper makes a valuable contribution to the field of MDM in manufacturing big data environments [1].

Ikbal Taleb, et al. (2021) developed a holistic framework for managing the quality of big data that includes four key components: data acquisition, data integration, data processing, and data usage. The authors discussed that traditional approaches to data quality management are insufficient for big data environments, which require a more comprehensive and continuous approach to quality management. The paper identifies gaps in the existing literature and discusses the benefits and limitations of the proposed framework. Overall, the paper makes a valuable contribution to the field of big data quality management by providing a comprehensive framework for managing data quality in big data environments [2].

Anders Haug, Jan Stentoft (2011) proposed a holistic framework for managing the quality of big data. It identifies gaps in traditional approaches to data quality management and discussed that big data environments require a more comprehensive and continuous approach to quality management. The framework consists of four key components and includes specific quality criteria and metrics to monitor and improve data quality throughout the big data lifecycle. Overall, the paper makes a valuable contribution to the field of big data quality management [3].

The paper "Big Heterogeneous Data Integration and Analysis" by Stella Vetova (2021) proposes a framework for integrating and analysing big data from diverse sources. The author states that big data integration is challenging due to the heterogeneity of data sources and the need for specialized tools and techniques. The proposed framework includes three main components: data ingestion, data integration, and data analysis. The research provides a valuable contribution to the field of big data integration and analysis by providing a framework that can be used to integrate and analyse data from diverse sources [4].

Li Cai (2015) explores the unique challenges of assessing data quality in the context of big data. The authors argue that traditional approaches to data quality assessment are insufficient for big data environments due to the sheer volume and complexity of data. The paper discusses the challenges of data quality assessment in the big data era and proposes a set of guidelines for assessing data quality in these environments. Overall, the paper provides a valuable contribution to the field of data quality assessment in the context of big data [5].

Rodríguez-Mazahua, L., et al. (2016) provides a broad overview of the field of big data. The authors discuss the various applications of big data across industries, including healthcare, finance, and transportation. They also examine the tools and technologies used in big data management, such as Hadoop and Spark. The paper addresses the challenges associated with big data, including issues related to data quality and privacy, and provides an overview of the latest trends in big data research. Overall, the paper provides a valuable contribution to the field of big data by providing a comprehensive perspective on the subject [6].

Sidi, Fatimah, et al. (2013) provides an overview of the various dimensions of data quality. The authors represent that data quality is a critical factor in decision-making and that a comprehensive understanding of the dimensions of data quality is necessary for effective data management. The paper surveys the existing literature on data quality and provides a taxonomy of data quality dimensions, including accuracy, completeness, consistency, timeliness, and relevance. Overall, the paper provides a valuable contribution to the field of data quality management by providing a comprehensive overview of data quality dimensions [7].

III. INTEGRATION AND MASTER DATA MANAGEMENT TOOLS

Software that facilitates the data integration process on the data source is known as a data integration tool (Lisa Ehrlinger et al.(2022)). Data cleansing, mapping, and transformation are all accomplished using master data management tools (Yasir Arfat et al. (2019)). The use of data governance and data quality technologies together with data integration tools is now possible. When an organization creates a single source of truth for all of its crucial data, they do so using master data management techniques and technologies. A company can distribute correct and consistent master data across the entirety of its operation by using master data management [10]. Below are some important and most trendy tools described:

1. Ataccama ONE

Ataccama is a complete augmented data management platform that provides a number of functions, such as data discovery and profiling, metadata management, data catalogue, data quality management, administration of master and reference data, and big data processing and integration [17]. The platform is intended to be completely integrated while remaining modular., accommodating various data types, users, domains, and deployment scenarios. It incorporates text analytics, machine learning capabilities, and the ability to enrich data using external sources. It also includes data lake profiling functionality [17].

What sets Ataccama apart is its ability to unify Data Governance, Data Quality, and Master Data Management into a single, AI-powered fabric that operates seamlessly across hybrid and cloud environments. This allows organizations to drive innovation at an accelerated pace while ensuring data trust, security, and governance. Notably, Ataccama serves as both a data integration tool and a master data management solution, with native compatibility on major Big Data platforms such as Spark, AWS, Databricks, MapR, Azure, and Hortonworks [17]. Below is an explanation of some of its features in more depth.

1.1Data Quality

Ataccama harnesses the power of AI to enhance data quality operations. It offers automatic anomaly detection capabilities, allowing for real-time identification of data irregularities[17]. Business rules can be assigned to data automatically, streamlining the process. The platform enables fast data processing regardless of the dataset's size. Customization of data quality rules is made simple, eliminating the need for coding by offering various pre-built rules. Ataccama excels in duplicate detection, enforcing standardization rules, and sending change notifications. Moreover, it facilitates the identification of data patterns and enables corresponding actions to be taken[17].

1.2. Master Data Management

Ataccama offers rapid model development capabilities facilitated by its data discovery and profiling features [17]. It provides AI-suggested matching rules and includes a comprehensive data quality processing component. The platform also offers a feature-rich data steward web application. Ataccama's augmented, multidomain Master Data Management (MDM) hub is well-suited for mission-critical deployments. Notably, it excels in data cleansing, matching, and merging, providing enhanced property and customization options. Users can easily implement flexible data and logical models, as well as workflows.

1.3. Data Catalogue

Ataccama incorporates automated data discovery and change detection capabilities within this component. It enables automated data quality calculation, anomaly detection, and enforcement of data policies [17].

1.4. Reference Data Management

It manages all reference data in a single app and gives users and systems accurate reference data. It guarantees data governance and record hierarchies. It offers operational usage and meta-driven development. [17].

1.5. Data Integration

In order to meet different use cases, it extracts, loads, and transforms data. With built-in data quality features, it can improve integrated data and orchestrate, manage, and monitor pipelines. It can also combine traditional, cloud, and big data sources [17].

1.6. Data Stories

It works directly with data to present information is made possible by this. To keep your story current, select from a variety of charts and highlight key information. It also has potential to share results with others with video exports and embed [17].

2. Informatica Multidomain MDM

Informatica offers a modular Master Data Management (MDM) solution that empowers users to establish a unified and reliable view of their data [4,6]. This tool makes it possible to develop an extensive and accurate representation of business-critical data, even when it originates from diverse, duplicated, or contrasting sources. It incorporates machine and AI learning capabilities including features like data security, business process management, data integration, and control of the quality of the data. Additionally, the solution allows for seamless enrichment of master data records through integration with external data providers[18]. It can be used or deployed either on-premises or in cloud environments.

3. SAP Master Data Governance

Enterprise Master Data Management (MDM) capabilities is offered by SAP through its SAP Master Data Governance product [8]. This solution offers flexible deployment options, allowing users to implement it either on-premises or in the cloud. With SAP Master Data Governance, organizations can effectively consolidate and centrally manage their master data. The software offers pre-built data models, business rules, workflow capabilities, and user interfaces to support multiple master data domains and implementation styles [8]. Additionally, it gives users the ability to create, verify, and keep track of business rules to ensure that master data is prepared and to evaluate the effectiveness of data management procedures.

4. IBM InfoSphere Master Data Management

It is an all-encompassing solution that manages all aspects of crucial company data, regardless of the system or model, and shows it to application users through a unified view [19]. This solution provides a flexible architecture that enables hybrid cloud systems and assures adherence to data governance rules and standards. Both the standard and advanced editions of InfoSphere MDM are available for on-premises implementation as well as fully managed cloud solutions [19].

5. Innovative Systems Synchronos

It is an enterprise MDM solution utilized for analytical or operational needs is Synchronos platform. The product is available for on-premises, cloud, or hybrid deployment [8]. Synchronos offers data profiling, data discovery, and data monitoring, together with a 360-degree view that enables customers to learn more about deeper linkages in data. Customers can create and change workflows using the functionality of workflow management, while navigational tools and a graphical display are made possible by hierarchy management [8].

6. Semarchy xDM

This tool is offered by Semarchy. The software makes use of machine learning methods to facilitate sophisticated matching, survival, curation, and classification as well as stewardship. The tool has a native data model that makes visible lineage, audibility, and governance possible. To combine the data hub with current applications and business processes, xDM can integrate any data source via batch and real-time APIs.[8]

7. Integrate.io

Integrate.io is a cloud-based solution that enables seamless integration, processing, and preparation of data for analytics purposes [4]. It serves as a comprehensive toolkit for constructing efficient data pipelines. What sets Integrate.io apart is its user-friendly approach, accommodating individuals with varying levels of technical expertise through its no-code and low-code options. The platform also offers advanced customization capabilities through its API component. With Integrate.io's package designer, users can easily implement various data integration use cases, including simple replication, complex data preparation, and transformation tasks [6]. The intuitive graphic interface allows for the implementation of Extract, Transform, Load (ETL), Extract, Load, Transform (ELT), Extract, Transform, Load, Transform (ETLT) or replication processes.

8. Altova MapForce

Altova MapForce is a highly efficient and scalable data integration tool known for its lightweight nature. Compared to expensive data management products, the MapForce Platform offers similar capabilities and features at a significantly lower cost. Its user-friendly visual interface includes a built-in function builder and data mapping debugger, simplifying the integration process.[8] MapForce supports automated data integration, making it an excellent middleware solution for connecting distributed applications within local enterprises, web-based workflows, and cloud architectures.

9. Talend

Talend offers an open-source solution that enables seamless integration and can be customized by users. Renowned for its high performance, Talend is specifically designed to meet analytical data requirements [20]. It provides a cost-efficient method for connecting data and adopts a unique analytical data-oriented approach to deliver optimal business analysis capabilities [20]. Talend facilitates bulk development processes, ensuring faster data migration [20]. Additionally, Talend features a smart data migration mechanism that efficiently maps and migrates data based on specific criteria.

10. Pentaho

Pentaho offers distinctive capabilities for generating actionable data insights by tracking user usage and tailoring insights accordingly. It provides a comprehensive end-to-end analytics solution that spans across integrated applications. The platform enables real-time access to analytical reports at any time [20]. Pentaho incorporates custom-built components that facilitate data transformation and integration into a unified data source. Integration of unstructured data is made simple with minimal coding efforts [20]. Additionally, Pentaho Data Insights includes data profiling functionalities for in-depth analysis and ensures comprehensive data quality. The platform also constructs intelligent business rules to optimize performance.

11. OpenText

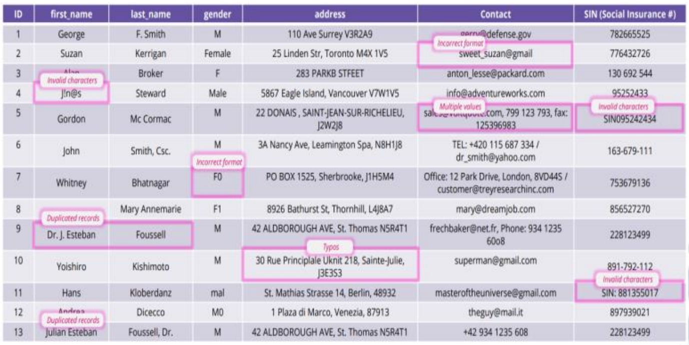
The platform assists organizations in consolidating traditional data and enterprise content into a unified platform [20]. It enables efficient evaluation of data interpretation and streamlined management of people and resources. With its agile integration approach, it supports parallel processing and rapid migration. By streamlining business operations, it enhances productivity [20]. The platform also promotes transparency among participants, facilitating optimal business processes. It effectively manages and streamlines the flow of information. The most appropriate data were obtained after the data was processed and all data rules, conditions, and filters were applied. These data were then saved in a database and could be utilized

***Article Received**: 03 May 2023 **Revised**: 30 June 2023 **Accepted**: 26 July 2023*

for business analysis and decision-making. A pictorial flow chart has been shown in figure 1.

IV. DATASET DESCRIPTION

We have used the secondary insurance dataset that includes the list of names of those who have purchased insurance. There are multiple columns out of which we have extracted some feature columns from the dataset and they are first_name,last_name(Concatenated as src_name), gender, address, email_address as contact, src_sin (Social Insurance Number), birth date, primary_ key. There are 1,39,317 rows in this dataset with null values, duplicates, invalid characters, and confusing or unclear values. A snapshot (figure 1. Input Data set Quality) highlights the issues has been shown below. We have merged multiple columns according to our need for processing.



ID	first_name	last_name	gender	address	Contact	SIN (Social Insurance #)
1	George	F. Smith	M	110 Ave Surrey V3R2A9	gerry@defense.gov	782665525
2	Suzan	Kerrigan	Female	25 Linden Str, Toronto M4X 1V5	sweet_suzan@gmail	776432726
3	Alan	Broker	F	283 PARKB STFEET	anton_lesse@packard.com	130 692 544
4	J!n@s	Steward	Male	5867 Eagle Island, Vancouver V7W1V5	info@adventureworks.com	95252433
5	Gordon	Mc Cormac	M	22 DONAIS , SAINT-JEAN-SUR-RICHELIEU, J2W2J8	sales@[illegible].com, 799 123 793, fax: 125396983	SIN095242434
6	John	Smith, Csc.	M	3A Nancy Ave, Leamington Spa, N8H1J8	TEL: +420 115 687 334 / dr_smith@yahoo.com	163-679-111
7	Whitney	Bhatnagar	F0	PO BOX 1525, Sherbrooke, J1H5M4	Office: 12 Park Drive, London, 8VD44S / customer@treyresearchinc.com	753679136
8		Mary Annemarie	F1	8926 Bathurst St, Thornhill, L4J8A7	mary@dreamjob.com	856527270
9	Dr. J. Esteban	Foussell	M	42 ALDBOROUGH AVE, St. Thomas N5R4T1	frechbaker@net.fr, Phone: 934 1235 60o8	228123499
10	Yoishiro	Kishimoto	M	30 Rue Principlale Uknit 218, Sainte-Julie, J3E3S3	superman@gmail.com	891-792-112
11	Hans	Kloberdanz	mal	St. Mathias Strasse 14, Berlin, 48932	masteroftheuniverse@gmail.com	SIN: 881355017
12	Andrea	Dicecco	M0	1 Plaza di Marco, Venezia, 87913	theguy@mail.it	897939021
13	Julian Esteban	Foussell, Dr.	M	42 ALDBOROUGH AVE, St. Thomas N5R4T1	+42 934 1235 608	228123499

Table 1. Input data set

V. METHODOLOGY

We have used the Ataccama One MDM cum Integration application to construct the pipeline and workflows. We used a dataset that is unstructured, disorganised, and filled with incorrect and undesirable data. The ruleset was then defined using the master data management programme. If we look at Table 1, we can see that there are typos, duplicate entries, poor formatting, invalid characters, multiple values, etc. These all values have all been spotted and recognised during profiling. Then, we have used cleanse to get rid of incorrect characters and merging numerous records into one, we applied transformation to enhance and enrich the data quality.

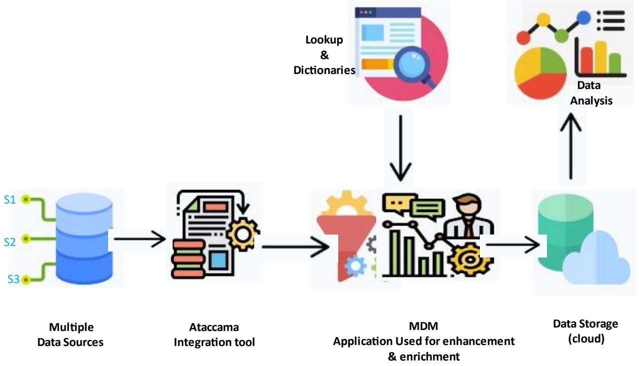


Figure 1. Data Processing using Ataccama

Then, we created lookups or dictionaries, which are in charge of replacing records with accurate ones drawn from a list of available alternatives. Here, we've utilised leet code to define a few special characters and a list of alternatives that may be used to retrieve records from a lookup based on matching in order to connect (group) the present record with another. Additionally, we conducted a validity check to eliminate values with duplicates, invalid data formats, or mistakes.

We gathered the most appropriate data after processing the data and applying all filters, conditions, and data rules. Figure 2 shows the cleansing process which removes the invalid character by fetching the most suitable data from the lookup.

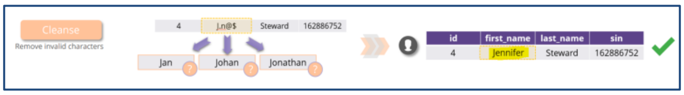


Figure 2: Cleansing Process

Figure 3 shows profiling, which explains the stats of all the columns present in the dataset in the tabular and charts form. It gives the categorical and overall idea of data present in the dataset. A key factor in this is data enrichment. The majority of MDM systems are merely capable of transforming and sanitizing data, but this tool went beyond that and used automated enrichment based on a defined lookup. Using the most significant and effective method to retrieve the closest data based on the established matching rules is one of its AI features.

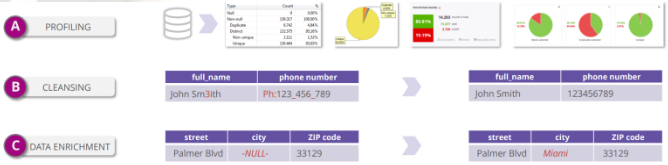


Figure 3. Profiling, Cleansing and Data Enrichment Process

Figure 4 shows the Pipeline which we have created to perform all the activities. It has various component used for different operations. Here First we have used The Text file reader which consist of our dataset. Then we have used multiplicator to use other two components in parallel. Component is the combination of Column assigner, Simple Scoring, Mapper,

DOI: https://doi.org/10.17762/ijritcc.v11i8.7947
***Article Received**: 03 May 2023 **Revised**: 30 June 2023 **Accepted**: 26 July 2023*

cleanse and Normalizer where cleansing, Transformation, matching and enrichment are taking place. We have used transliterate which consist of lookup and predefined leet code used for some special characters. For each row it can transform value given by an expression and store it into an output column. There can be more input expressions and output columns. The transformation tries to find substrings (given by the property Rules) in the input expressions and substitutes them by string given in the matching rule. When there are two rules with the same prefix, the longest possible rule is always used (i.e. considered as matched). Then we have cleansing which consist of set of rules in it, then we have column assigner in which we have assigned output columns which we need to validate. Simple scoring plays a vital role here. It is responsible to Scores and flags input records according to defined rules. Passes through all defined scoring cases and evaluates their conditions. If the condition is true, it appends the explanation to the explanation Column and increments the value in scoreColumn.

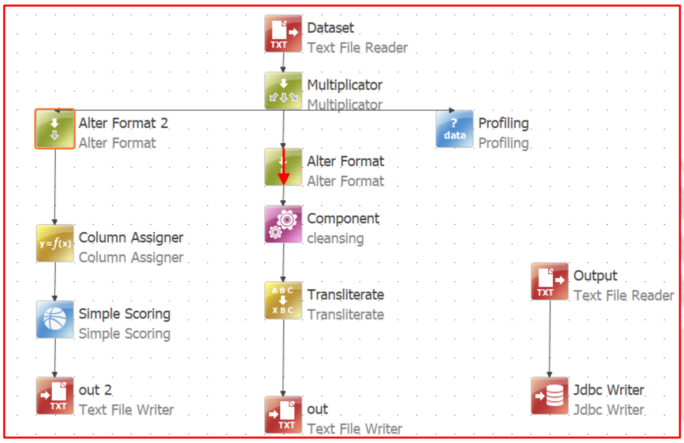


Figure 4. Ataccama Pipeline

Making of Mater Data:

In the dataset, we could see that it provide data inconsistant ,often reffering to a single instance but differ in format, structure and quality.So Master data aims to find a way to select the best value across all data sources and creata a single truth called “Master data” shown in Figure 5. The resulting master data is then saved in an Oracle database, where all business users can access the results.

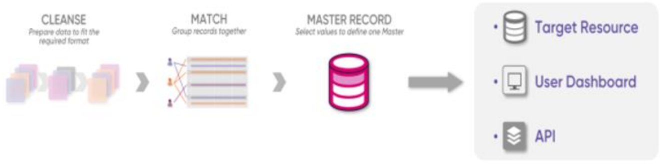
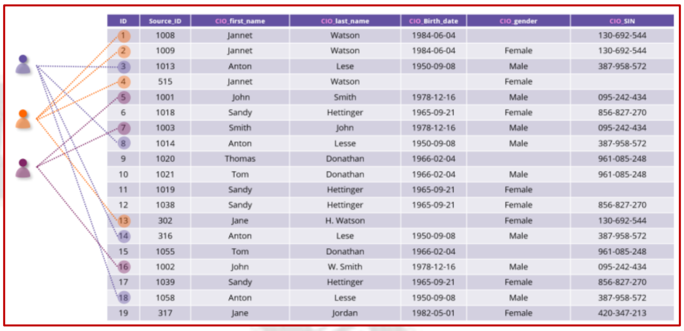


Figure 5. Cleanse, Match and Creation of Master Record

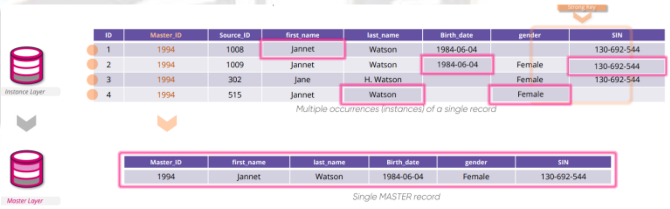
Matching and grouping are crucial. Once the matching and merging is complete, we will group the results to create a master data set. Each entry has been given a source_ID that can be used to create the master data. Figure 6 demonstrates that IDs 1, 2, 4, 13, and 38 all have the same social security number, which is CIO_SIN (the name used for the output column). It has been demonstrated that each cluster is represented with a vibrant colour to represent grouping of each source_id.



ID	Source_ID	CIO_first_name	CIO_last_name	CIO_Birth_date	CIO_gender	CIO_SIN
1	1008	Jannet	Watson	1984-06-04		130-692-544
2	1009	Jannet	Watson	1984-06-04	Female	130-692-544
3	1013	Anton	Lese	1950-09-08	Male	387-958-572
4	515	Jannet	Watson		Female	
5	1001	John	Smith	1978-12-16	Male	095-242-434
6	1018	Sandy	Hettinger	1965-09-21	Female	856-827-270
7	1003	Smith	John	1978-12-16	Male	095-242-434
8	1014	Anton	Lesse	1950-09-08	Male	387-958-572
9	1020	Thomas	Donathan	1966-02-04		961-085-248
10	1021	Tom	Donathan	1966-02-04	Male	961-085-248
11	1019	Sandy	Hettinger	1965-09-21	Female	
12	1038	Sandy	Hettinger	1965-09-21	Female	856-827-270
13	302	Jane	H. Watson		Female	130-692-544
14	316	Anton	Lese	1950-09-08	Male	387-958-572
15	1055	Tom	Donathan	1966-02-04		961-085-248
16	1002	John	W. Smith	1978-12-16	Male	095-242-434
17	1039	Sandy	Hettinger	1965-09-21	Female	856-827-270
18	1058	Anton	Lesse	1950-09-08	Male	387-958-572
19	317	Jane	Jordan	1982-05-01	Female	420-347-213

Figure 6. Grouping of Data

Various rules are outlined for building the master data. The master data created with the name master_id is displayed in Figure 7. If you look at this figure, you'll notice that the last name has three names that are identical to it, while the first name is tallied three times. Therefore, it gathered those data and created a master data set based on the number of maximum values discovered in the columns.


Strong Key
Instance Layer
Master Layer

ID	Master_ID	Source_ID	first_name	last_name	Birth_date	gender	SIN
1	1994	1008	Jannet	Watson	1984-06-04		130-692-544
2	1994	1009	Jannet	Watson	1984-06-04	Female	130-692-544
3	1994	302	Jane	H. Watson		Female	130-692-544
4	1994	515	Jannet	Watson		Female	

Multiple occurrences (instances) of a single record

Master_ID	first_name	last_name	Birth_date	gender	SIN
1994	Jannet	Watson	1984-06-04	Female	130-692-544

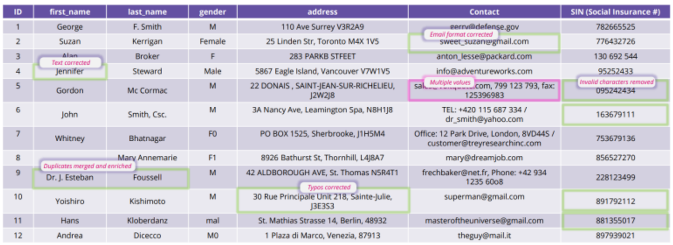
Single MASTER record

Figure 7. Master Data

VI. RESULT AND DISCUSSION

In this research,we have tried to improve the data quality and consistency. A methodology was put into place utilising the Ataccama toolset. The major goal of this research is to determine the degree to which this tool can improve the accuracy, enrichment, uniqueness, and consistency of data. The growth in data quality's enrichment, enhancement, and uniqueness has been demonstrated by actual application. Additionally, it demonstrates how search may be used to retrieve alternative suggestions to enrich records and combine them together to create master data, as well as how data quality can be improved. The outcomes have demonstrated are consists of a collection of rules that stated to produce the desired results.

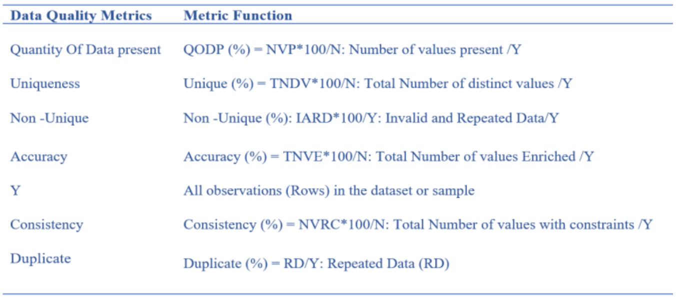
DOI: https://doi.org/10.17762/ijritcc.v11i8.7947
Article Received*: 03 May 2023* ***Revised****: 30 June 2023* ***Accepted****: 26 July 2023*



ID	first_name	last_name	gender	address	Contact	SIN (Social Insurance #)
1	George	F. Smith	M	110 Ave Surrey V3R2A9	gerry@defense.gov	782665525
2	Suzan	Kerrigan	Female	25 Linden Str, Toronto M4X 1V5	sweet_suzan@gmail.com	776432726
3	Alan	Broker	F	283 PARKB STFEET	anton_lesse@packard.com	130 692 544
4	Jennifer	Steward	Male	5867 Eagle Island, Vancouver V7W1V5	info@adventureworks.com	95252433
5	Gordon	Mc Cormac	M	22 DONAIS , SAINT-JEAN-SUR-RICHELIEU, J2W2J8	sal[illegible].com, 799 123 793, fax: 125396983	095242434
6	John	Smith, Csc.	M	3A Nancy Ave, Leamington Spa, N8H1J8	TEL: +420 115 687 334 / dr_smith@yahoo.com	163679111
7	Whitney	Bhatnagar	F0	PO BOX 1525, Sherbrooke, J1H5M4	Office: 12 Park Drive, London, 8VD44S / customer@treyresearchinc.com	753679136
8		Mary Annemarie	F1	8926 Bathurst St, Thornhill, L4J8A7	mary@dreamjob.com	856527270
9	Dr. J. Esteban	Foussell	M	42 ALDBOROUGH AVE, St. Thomas N5R4T1	frechbaker@net.fr, Phone: +42 934 1235 60o8	228123499
10	Yoishiro	Kishimoto	M	30 Rue Principale Unit 218, Sainte-Julie, J3E3S3	superman@gmail.com	891792112
11	Hans	Kloberdanz	mal	St. Mathias Strasse 14, Berlin, 48932	masteroftheuniverse@gmail.com	881355017
12	Andrea	Dicecco	M0	1 Plaza di Marco, Venezia, 87913	theguy@mail.it	897939021

Table 2. Output Data Set

Table 2 shows how the redundant data were consolidated and improved. and more unneeded data was eliminated. Some text are corrected which consists of special characters or other invalid string that are specified in the lookup. The format of the contact(Table 3) has been adjusted, invalid characters have been eliminated, and some mistakes have been fixed. Profiling the input and output datasets revealed that many null values were obtained with the proper data from the lookup indicated, duplicates were eliminated and enriched, non-unique data were improved, and the accuracy of unique columns had significantly risen.The end result was produced using the Ataccama master data center and the data Quality indicators shown in table 3, which are taken from [2] research.

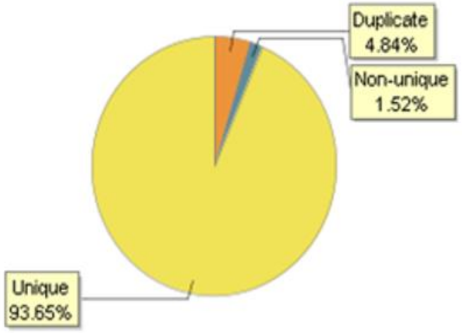
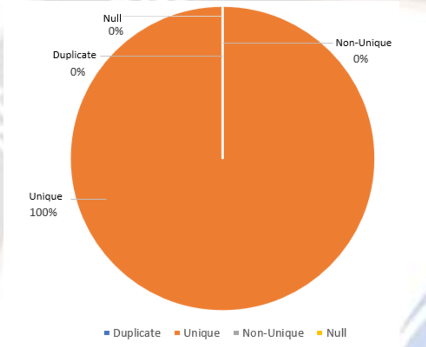
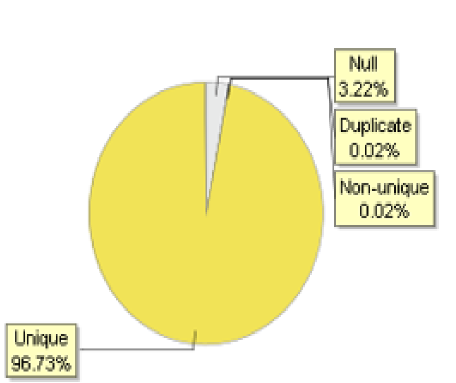
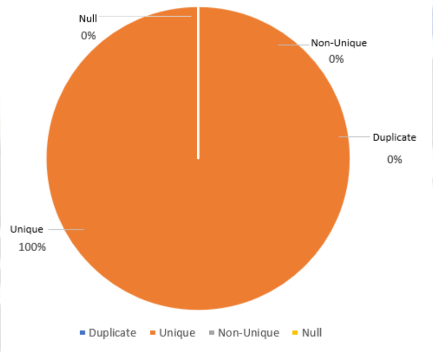


Data Quality Metrics	Metric Function
Quantity Of Data present	QODP (%) = NVP*100/N: Number of values present /Y
Uniqueness	Unique (%) = TNDV*100/N: Total Number of distinct values /Y
Non -Unique	Non -Unique (%): IARD*100/Y: Invalid and Repeated Data/Y
Accuracy	Accuracy (%) = TNVE*100/N: Total Number of values Enriched /Y
Y	All observations (Rows) in the dataset or sample
Consistency	Consistency (%) = NVRC*100/N: Total Number of values with constraints /Y
Duplicate	Duplicate (%) = RD/Y: Repeated Data (RD)

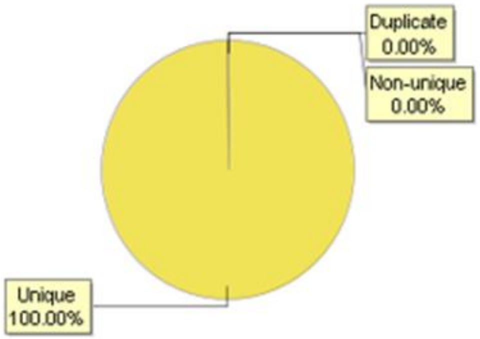
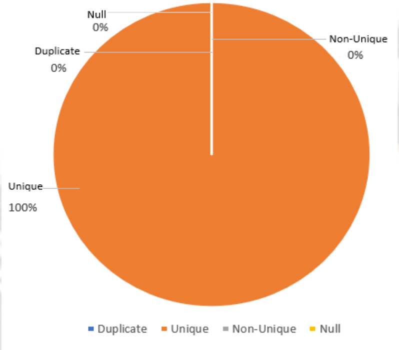
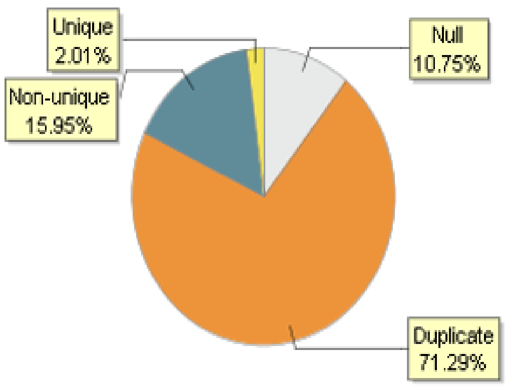
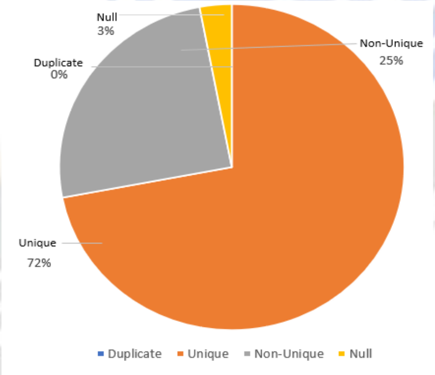
Table 3. Data Quality Metrics and Function

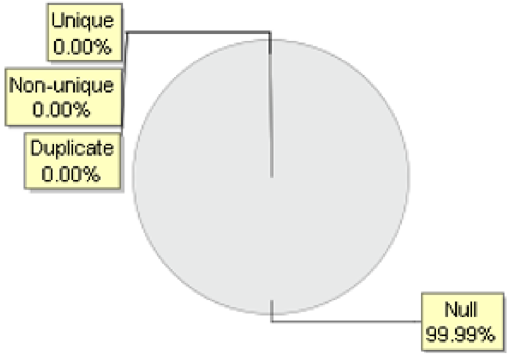
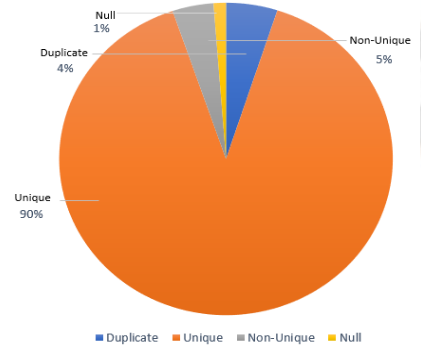
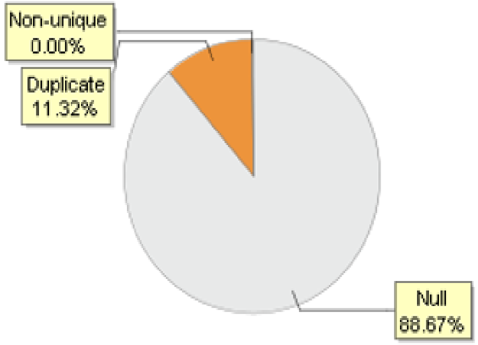
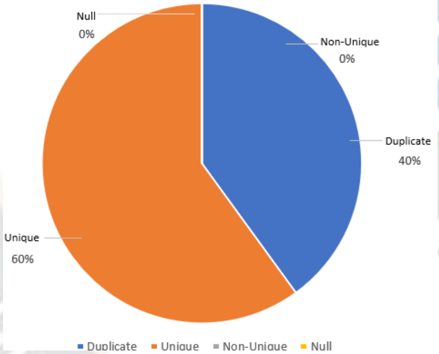
We have also tried to show the comparison on the basis of each column which we have used.Detailed explanation has been shown in the Table 5.

Table 4. Column Wise Categorization

INPUT DATA PROFILE Coulmns Names	OUTPUT DATA PROFILE	EXPLANATION
1. SRC_NAME  	 	The SRC_NAME column contains 4.84% duplicate values, 1.52% non-unique values, and 93.65% unique values. After processing the data using Ataccama, we were able to remove the duplicate data. Additionally, any invalid data such as special characters, numbers, and ambiguous data in the column names were dropped, resulting in the retention of only the unique data.
2. SRC_SIN(Social Insurance)  	 	The SRC_SIN column represents the source social insurance data and contains 0.02% duplicate values, 0.02% non-unique values, 3.22% null values, and 96.73% unique values. After applying rules, conditions, and checks on the dataset using Ataccama, we successfully eliminated the duplicate data. Furthermore, any invalid data such as special characters, numbers, or ambiguous entries in the column were removed, resulting in a dataset that solely consists of unique and valid data.
3. SRC_PRIMARY_KEY		The SRC_PRIMARY_KEY column serves as the source primary key, specifically designated to identify certain records for creating the master record. This column contains 100% unique values, ensuring that each record is distinct and non-duplicated.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8
DOI: https://doi.org/10.17762/ijritcc.v11i8.7947
***Article Received**: 03 May 2023 **Revised**: 30 June 2023 **Accepted**: 26 July 2023*

<table>
<tr>
<td>
</td>
<td>
</td>
<td></td>
</tr>
<tr>
<td>4. SRC_BIRTH_DATE

</td>
<td>
</td>
<td>The SRC_BIRTH_DATE column represents the birth date of the SRC_NAME column and contains 71.29% duplicate values, 15.95% non-unique values, and 2.01% unique values. After processing the data using Ataccama, the duplicate data has been eliminated completely. Additionally, the presence of invalid data, such as special characters, has been reduced by 25%. Ambiguous data in the column names has been dropped, resulting in optimized data.</td>
</tr>
<tr>
<td>5. SRC_EMAIL</td>
<td></td>
<td>The SRC_EMAIL column represents the email ID and initially contained 99.99% null values. To address this, we utilized a lookup value and conditions based on the name column to predict the email ID using the first name and last name. This process was</td>
</tr>
</table>

		implemented to facilitate decision-making and provide additional utility. The data underwent validation using system configuration to ensure accuracy and reliability..
6. SRC_GENDER 		The SRC_GENDER column represents the source gender information. Initially, it contained 11.32% duplicate values, 0% non-unique values, and 88.67% null values. After processing the data using Ataccama, the system assigned genders based on the input provided during the pipeline creation. The null values were replaced with valid gender data, such as Male or Female. The duplicates observed in the data were resolved, ensuring that repeated occurrences of Male, Female, or other genders were eliminated.

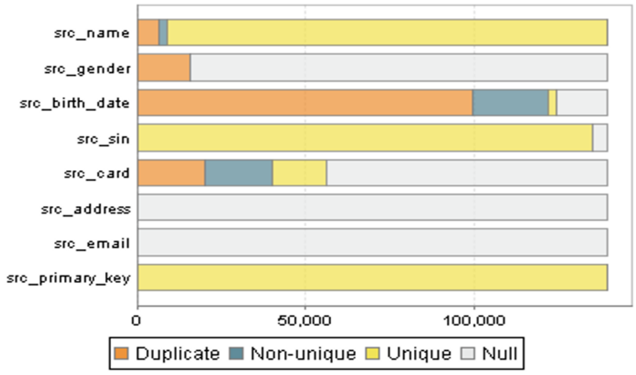
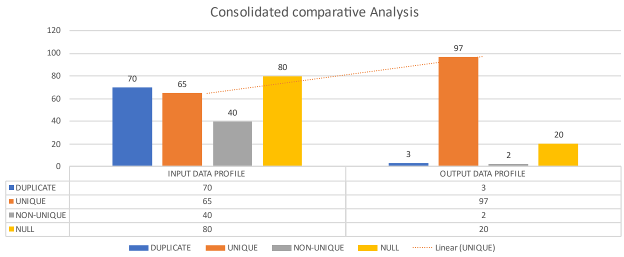


Figure 8. Master Input Data Graph



	INPUT DATA PROFILE	OUTPUT DATA PROFILE
DUPLICATE	70	3
UNIQUE	65	97
NON-UNIQUE	40	2
NULL	80	20

Figure 9. Consolidated comparision

The master graph for all of the input columns is shown in figure 8. We attempted to demonstrate a comprehensive comparative study of input data vs output data acheived in figure 9. We have seen an improvement in the quality of the input data, and an overall Unique value has been attained. Linear displays the amount of unique value achieved between 65% and 97%, which is the target range. Apart from this , we could see that duplicate percentage has been drooped from 70% to 3% which means that the huge amount of redundent data has been eliminated. Non Unique has been decreased to 2% from 40% and null values has been populated to 60% on the basis on lookup and dictionaries. From the above result, it states that the efficiency and the quality of data has been highly improved using Ataccama Application.

CONCLUSION

We have completed the comprehensive study and execution for data quality optimization and learned how the data uniqueness can be achieved and optimized with the help of Ataccama application which can be further used to business for decision-making. This toolset may be used to optimize massive data and has the potential to study a lot about them. Additionally, it can be utilized in the future to overcome the barriers associated with huge data and heterogeneity. With the use of this tool, we can create a customized query that meets our specific data needs. In the future, we will need such powerful AI-based solutions that have the capability of making recommendations for potential data checks. Big data still presents challenges to overcome. The quantity of optimization yield from larger and unstructured datasets can be improved to a greater degree with these tools. Improved data effectiveness and efficiency can lead to better decision-making and analysis for the organization. The outcome of this work will add the value to research in the area of Data quality optimization.

REFERENCES

[1] C. Zhao, L. Ren, Z. Zhang, and Z. Meng (2020), “Master data management for manufacturing big data: a method of evaluation for data network,” World Wide Web, vol. 17, no. 2, pp. 1407–1421, 2020, doi: 10.1007/s11220-019-00707-8.

[2] I. Taleb, M. A. Serhani, C. Bouhaddioui, and R. Dssouli (2021), “Big data quality framework: a holistic approach to continuous quality management”, vol. 8, no. 1. Springer International Publishing, 2021. doi: 10.1186/s40537-021-00468-0.

[3] Haug, A., Arlbjørn, J.S (2011), “Barriers to master data quality”, J. Enter. Inf. Manag.

[4] Vetova, Stella. (2021). Big heterogeneous data integration and analysis. AIP Conference Proceedings. 1733. 030007. 10.1063/5.0043621.

[5] Cai, Li & Zhu, Yangyong (2015), “The Challenges of Data Quality and Data Quality Assessment in the Big Data Era”, Data Science Journal. 14. 10.5334/dsj-2015-002.

[6] Rodríguez-Mazahua, L., Rodríguez-Enríquez, CA., Sánchez-Cervantes, J.L. et al (2016), “A general perspective of Big Data: applications, tools, challenges and trends”, J Supercomput 72,(2016).3073–3113 https://doi.org/10.1007/s11221-015-1501-1.

[7] Johansson Anna, Maria Jansen, Anna Wagner, Anna Fischer, Maria Esposito. Machine Learning Techniques to Improve Learning Analytics. Kuwait Journal of Machine Learning, 2(2). Retrieved from http://kuwaitjournals.com/index.php/kjml/article/view/189

[8] Sidi, Fatimah & Hassany Shariat Panahy, Payam & Affendey, Lilly & A. Jabar, Marzanah & Ibrahim, Hamidah & Mustapha, Aida. (2013). Data quality: A survey of data quality dimensions. 10.1109/InfRKM.2012.6204995.

[9] Yasir Arfat, Sardar Usman, Rashid Mehmood & Iyad Katib (2019), “Big Data Tools, Technologies, and Applications: A Survey”,DOI: 10.1007/978-3-030-13705-2_19.

[10] N. R. Sabar, J. Abawajy and J. Yearwood (2017), "Heterogeneous Cooperative Co-Evolution Memetic Differential Evolution Algorithm for Big Data Optimization Problems", IEEE Transactions on Evolutionary Computation, vol. 21, no. 2, pp. 315-321, April 2017, doi: 10.1109/TEVC.2016.2002260.

[11] Dreibelbis, A., Hechler, E., Milman, I., Oberhofer, M., et al. (2008), “Enterprise Master Data Management (Paperback)”, An SOA Approach to Managing Core Information. Pearson Education.

[12] A. K. Sangaiah, A. Goli, E. B. Tirkolaee, M. Ranjbar-Bourani, H. M. Pandey and W. Zhang (2020), "Big Data-Driven

Cognitive Computing System for Optimization of Social Media Analytics," in IEEE Access, vol. 8, pp. 82215-82220, 2020, doi: 10.1109/ACCESS.2020.2391394.

[13] Kaur, Prableen & Sharma, Manik& Mittal, Mamta. (2018), “Big Data and Machine Learning Based Secure Healthcare Framework. Procedia Computer Science” 132. 1049-1059. 10.1016/j.procs.2018.05.020.

[14] Haldorai, Anandakumar&Arulmurugan, R. & Chow, Chee Onn. (2019), “Big Data Analytics for Sustainable Computing. Mobile Networks and Applications”, 18. 10.1007/s11036-019-01393-6.

[15] Roy, Chandrima& Pandey, Manjusha &SwarupRautaray, Siddharth. (2018). A Proposal for Optimization of Data Node by Horizontal Scaling of Name Node Using Big Data Tools. 1-6. 10.1109/I2CT.2018.8523795.

[16] Kumar Pramanik, K. K., Neha, R. ., Limkar, S. ., Sule, B. ., Qureshi, A., & Kumar, K. S. . (2023). Accurate Classifier Based Face Recognition using Deep Learning Architectures by Noise Filtration with Classification. International Journal of Intelligent Systems and Applications in Engineering, 11(3s), 179–183. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/2558

[17] Wang, Hui & Wang, Wenjun & Cui, Laizhong& Sun, Hui & Zhao, Jia & Wang, Yun &Xue, Yu. (2017), “A Hybrid Multi-Objective Firefly Algorithm for Big Data Optimization. Applied Soft Computing”, 69. 10.1016/j.asoc.2017.06.023.

[18] J. Feng, L. T. Yang, R. Zhang, S. Zhang, G. Dai and W. Qiang (2020) "A Tensor-Based Optimization Model for Secure Sustainable Cyber-Physical-Social Big Data Computations", IEEE Transactions on Sustainable Computing, vol. 5, no. 2, pp. 217-174, 1 April-June 2020, doi: 10.1109/TSUSC.2018.2281466.

[19] Lisa Ehrlinger1,2*,Wolfram Wöß(2022) “A Survey of Data Quality Measurement and Monitoring Tools” Volume 5 - 2022 | https://doi.org/10.3389/fdata.2022.850611.

[20] R.Mukherjee and P. Kar (2017), "A Comparative Review of Data Warehousing ETL Tools with New Trends and Industry Insight," IEEE 7th International Advance Computing Conference (IACC), Hyderabad, India, 2017, pp. 943-948, doi: 10.1109/IACC.2017.0192.

[21] Wang, Shuliang & Yuan, Hanning. (2014),“Spatial Data Mining: A Perspective of Big Data. International Journal of Data Warehousing and Mining”. 10. 50-70. 10.4018/ijdwm.201410010

[22] Singh, Saravjeet & Singh, Jaiteg. (2022). A Survey on Master Data Management Techniques for Business Perspective. 10.1007/978-981-16-4284-5.