International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

IJISAE

ISSN:2147-6799 www.ijisae.org Original Research Paper

Comprehensive Approach for data Integration in Legacy System based Distributed Web Information Domain Using Machine Learning Techniques

Jinduja. S[1], Narayani. V[2]

Submitted: 28/05/2023 **Revised**: 17/07/2023 **Accepted**: 29/07/2023

Abstract: Legacy systems in data integration are the outdated hardware or software systems acts as the hurdle towards the effective data integration operation. The legacy systems stored their data bases information which are not feasible to link with the recent distributed data information integrated system due to its ineffectiveness and compatibility issues. The new technologies and open standards are not easily coping up with the existing legacy systems if it is not handled with proper care. Most of the past data integration schema dealt with these legacy systems which are common in use but they will act as an essential component in the future references. The existing methodologies target towards the integration of legacy data without care about its infrastructure and data classification resulted in failed integrated information collections. This research article proposes a machine learning approach to deal with the legacy systems in distributed web information domain with proper machine learning approach with its unique characteristics. In future this research paper will be extended with the implementation of genetic algortihm based distributed web information system.

Keywords: *Legacy system, Machine learning, web data, distributed data, heterogeneous data*

I. Introduction:

Legacy system:

A legacy system is outdated computing software and/or hardware that are still in use. The system still meets the needs it was originally designed for, but doesn't allow for growth.

Machine Learning:

Machine Learning is the field of study that gives computers the capability to learn without being explicitly programmed. This amazing technology helps computer systems learn and improve from experience by developing computer programs that can automatically access data and perform tasks via predictions and detections.

Heterogeneous data:

Heterogeneous data are any data with high variability of data types and formats. They are possibly ambiguous and low quality due to missing values, high data redundancy, and untruthfulness. It is difficult to integrate heterogeneous data to meet the business information demands.

Distributed System:

A distributed system is a system whose components are located on different networked computers, which communicate and coordinate their actions by passing messages to one another. Distributed computing is a field of computer science that studies distributed systems.

II. Methodology

There are 5 stages in the proposed methodology for the comprehensive approach for data integration in distributed web information domain using machine learning techniques. They are,

Stage-1: Rigid coupling communication

a. Facilitate communication through middleware entity

b. Routing and connection service monitoring.

c. Application & messaging services management.

d. Communication protocol maintenance.

Stage-2: Flexi coupling communication

a. Facilitate communication through face to face mode.

b. Perform data transfer through mapping.

c. Queue based message routing.

Stage-3: Sleeve coupling communication

a. Facilitate communication through application level.

b. Data abstraction, automation, and transfer between resources.

[1]Research Scholar, Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu, India.
Email: jindujaseelan@gmail.com
[2]Assistant Professor, Department of Computer Science,
St Xavier’s College (Affiliated to Manonmaniam Sundaranar University), Tirunelveli, Tamil Nadu, India.

c. Bi direction transfer.

Stage-4: Muff coupling communication

a. Facilitate communication through automatic mode.

b. Map data, database and data warehouse

c. Transfer and synchronize the data.

Stage-5: Flange coupling communication

a. Facilitate communication through cloud based services.

b. Perform integration through framework connectors.

The proposed methodology of comprehensive approach for data integration in legacy systems based distributed web information system using machine learning techniques is as follows in Fig-1.

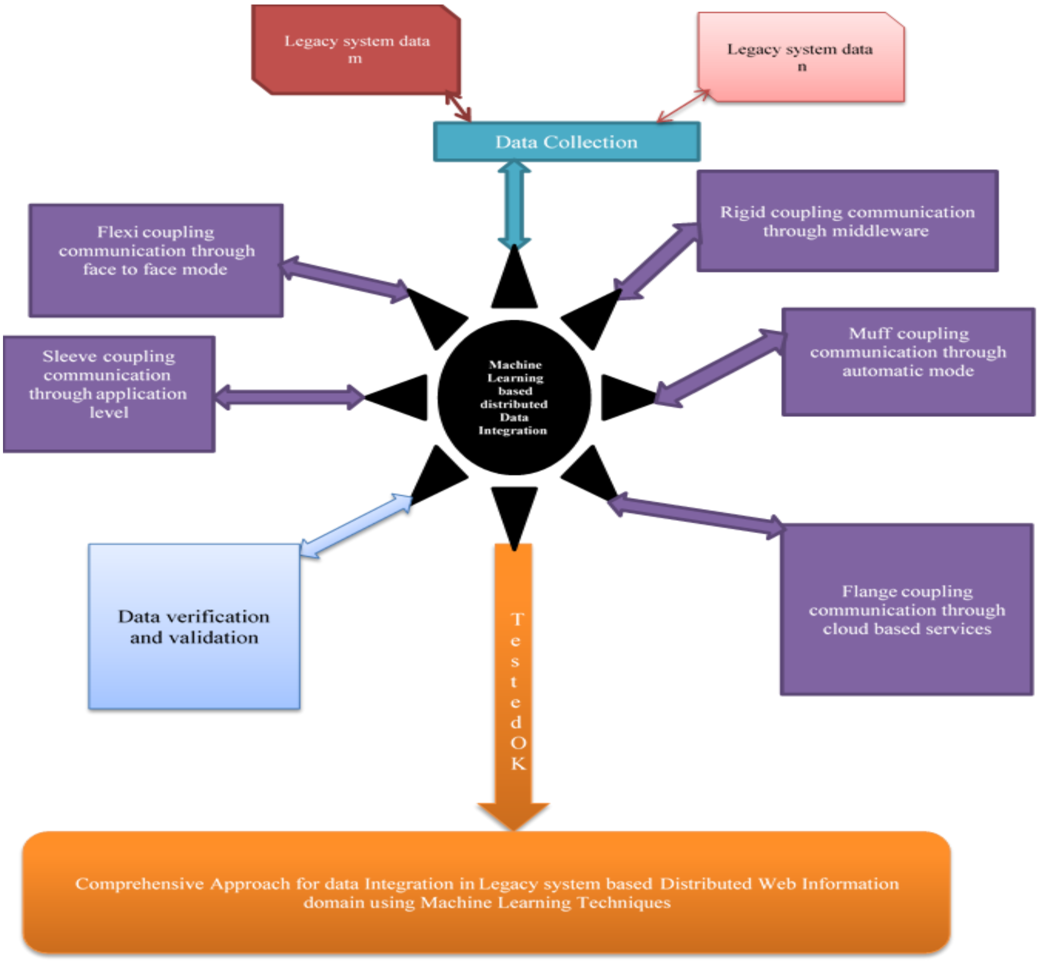

Legacy system data m
Legacy system data n
Data Collection
Rigid coupling communication through middleware
Flexi coupling communication through face to face mode
Muff coupling communication through automatic mode
Sleeve coupling communication through application level
Machine Learning based distributed Data Integration
Data verification and validation
Tested OK
Flange coupling communication through cloud based services
Comprehensive Approach for data Integration in Legacy system based Distributed Web Information domain using Machine Learning Techniques

Fig-1: Proposed Comprehensive approach for data integration

The flow chart for the comprehensive approach for data integration in legacy system based distributed web information domain using machine learning techniques is as follows,

```
Start
Input: Legacy system based web informatics data collection for data integration
Step-1: Identify the coupling level of resources based on data complexity in different data formats.
Select case (Coupling level)
Case (Rigid):     Goto Step-2;
Break;
Case (Flexi):     Goto Step-3;
Break;
Case (Sleeve):    Goto Step-4;
Break;
Case (Muff):      Goto Step-5;
Break;
Case (Flange):    Goto Step-6;
Break;
End Select
```

```
Step-2: Perform Machine learning based Middleware communication approach
Goto Step-7
Step-3: Perform Machine learning based face to face communication approach
Goto Step-7
Step-4: Perform Machine learning based application level communication approach
Goto Step-7
Step-5: Perform Machine learning based automated communication approach
Goto Step-7
Step-6: Perform Machine learning based cloud based communication approach
Goto Step-7
Step-7: Verification and validation
If all the above are success go to end
Else go to step-1
End if
End
```

III. Implementation

Step-1: Identify the coupling level of resources based on data complexity in different data formats.

The following machine learning based decision making tree acts as the supportive logic for identifying the coupling level of resources.

```
If data complexity level=high and data format accessibility=hard then
    Coupling level=rigid
Else if data complexity level=low and data format accessibility=easy then
Coupling level=flexi
        If data architecture =supportable then
        Coupling level=application level
        Else if data architecture =modifiable then
        Coupling level=automated
        End if
        If readymade pipeline communication=supportable then
        Coupling level=cloud based
        End if
End if
```

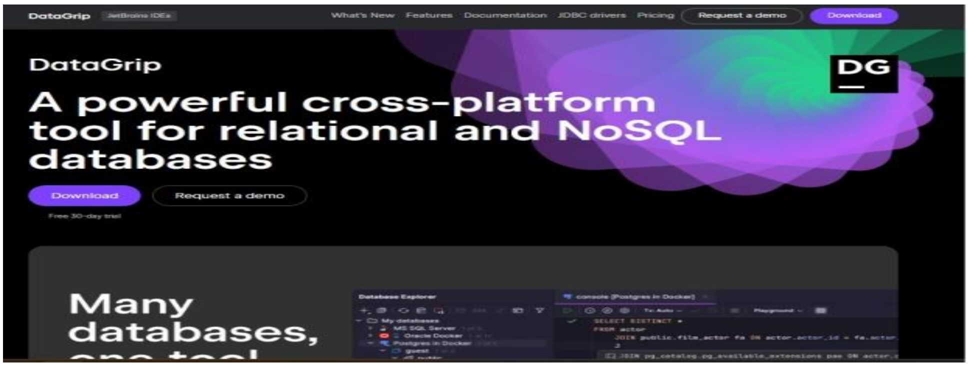
Step-2: Perform Machine learning based Middleware communication approach

a. Facilitate communication through middleware entity

The middleware communication focuses on the following,

i. Database communication

The legacy system databases are managed with proper communication tool for machine learning based mapping the data formats and establish communication for future correspondence. The tools such as Data grip [7]; Azure and tobino etc. are used for valid communication as in Fig-2.



DataGrip

A powerful cross-platform tool for relational and NoSQL databases

Free 30-day trial

Many databases, one tool

```
SELECT DISTINCT *
FROM actor
    JOIN public.film_actor fa ON actor.actor_id = fa.actor
```

Fig-2: Data grip tool for implementation

ii. Message oriented communication

The XML based XMPP [8] protocol , message queue services and message transmitter are the different tools acts as the supporting tool for message oriented communication among the legacy systems as in Fig-3.

Fig-3: XMPP tool for implementation

iii. Transaction processing support

The middleware transactional tools for implementation are as follows

- XML,
- REpresentational State Transfer(REST),
- Java Script Object Notation and Simple Object Access Protocol

These tools are used for maintaining and monitoring the transaction among the legacy systems.

b. Routing and connection service monitoring.

The routing and connection monitoring process includes the following machine learning procedures,

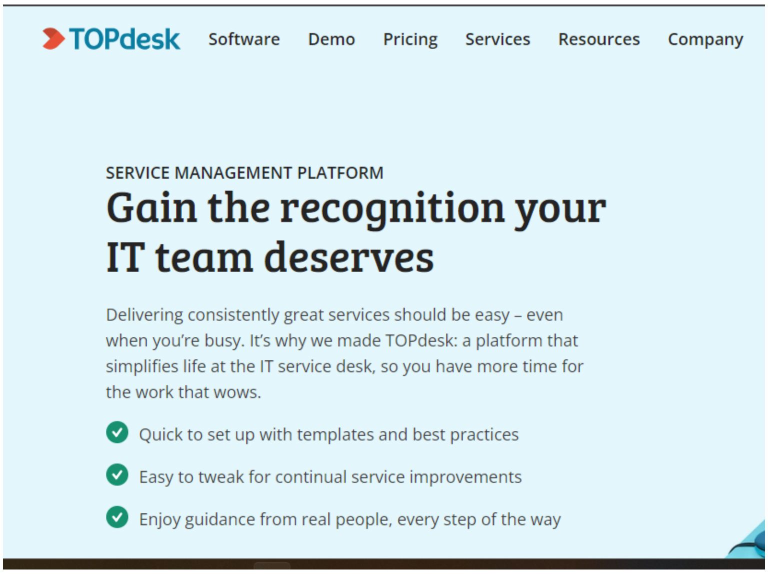
- Identify the route between the source and destination using the graph neural networks approach of shortest path method.
- Change and update the routing rule based on transaction progress.
- Access the metadata for interface related information.
- Monitor the connection until unit data transfer completed.

c. Application & messaging services management.

The application and messaging services focuses on the quality of service with the target performance achievement among the legacy systems communication. The main functionalities are,

- Detection of Quality of Service(QoS)
- Diagnose the QoS
- Remedy measures
- Report the resulting QoS.

The tools such as service now, service desk plus, top desk [9] etc. are playing the vital role in application services management as in fig-4.



SERVICE MANAGEMENT PLATFORM

Gain the recognition your IT team deserves

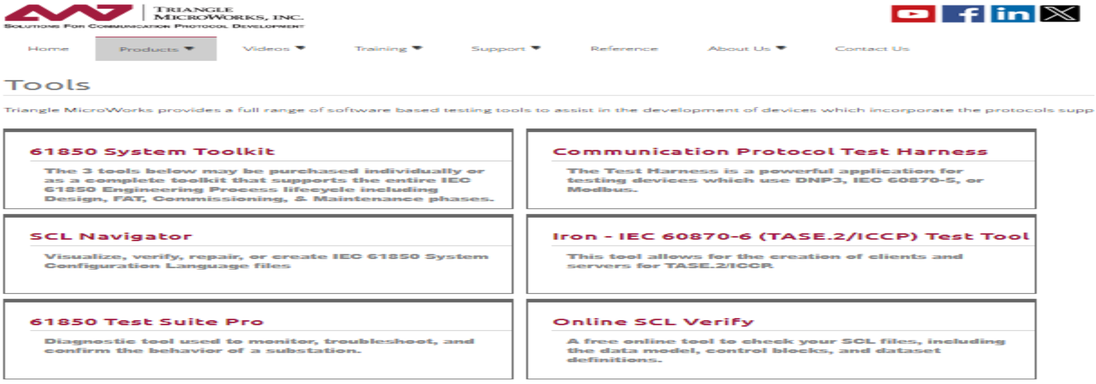
Delivering consistently great services should be easy – even when you're busy. It's why we made TOPdesk: a platform that simplifies life at the IT service desk, so you have more time for the work that wows.

- ✓ Quick to set up with templates and best practices
- ✓ Easy to tweak for continual service improvements
- ✓ Enjoy guidance from real people, every step of the way

Fig-4: Top-desk tool for implementation

d. Communication protocol maintenance.

The set of all rules and policies that defines the communicating among the legacy systems plays the vital role in effective data integration. Lab view and micro works [10] are the tools used for communication protocol maintenance as in fig-5.


Triangle MicroWorks, Inc.
Solutions For Communication Protocol Development
Home
Products
Videos
Training
Support
Reference
About Us
Contact Us

Tools

Triangle MicroWorks provides a full range of software based testing tools to assist in the development of devices which incorporate the protocols supp

61850 System Toolkit

The 3 tools below may be purchased individually or as a complete toolkit that supports the entire IEC 61850 Engineering Process lifecycle including Design, FAT, Commissioning, & Maintenance phases.

Communication Protocol Test Harness

The Test Harness is a powerful application for testing devices which use DNP3, IEC 60870-5, or Modbus.

SCL Navigator

Visualize, verify, repair, or create IEC 61850 System Configuration Language files

Iron - IEC 60870-6 (TASE.2/ICCP) Test Tool

This tool allows for the creation of clients and servers for TASE.2/ICCP.

61850 Test Suite Pro

Diagnostic tool used to monitor, troubleshoot, and confirm the behavior of a substation.

Online SCL Verify

A free online tool to check your SCL files, including the data model, control blocks, and dataset definitions.

Fig-5: Micro works tool for implementation

Step-3: Perform Machine learning based face to face communication approach

a. Facilitate communication through face to face mode.

Face to face mode communication can be done through by writing the customizable code for data communication among the legacy systems or establish connection through system interface if possible.

MySQL, SQL server are the tools used for implementing the face to face communication.

b. Perform data transfer through machine learning based mapping.

Data transfer through mapping focuses on the setting up of correspondence relation between the data, format, and protocols.

i. Data mapping:

The machine learning based data format mapping includes the following steps:

Procedure-1: Define data with its frequency and train the resultant data.

Instead of DOB the required field is age with the frequency of all records.

Procedure -2: Match the source and destination fields.

Source: DOB

Destination-Age

Resultant field: Age

Procedure -3: Test the sample data with proper computation.

If Name=Jindhu and DOB=08/10/2021 then Age=1 as on 08/10/2022

Procedure -4: Apply supervised learning for the successful test results.

Age=Integer [(current date-DOB)/365]

Procedure -5: Maintain and update the trained data if needed.

Age in terms of years is quotient 365, months is quotient 12

ii. Data format mapping

Data format mapping represents the source and destination with the same unique data format transformation.

The DOB format in source legacy system may be YYYY/MM/DD and the destination legacy system is DD/MM/YYYY then the required conversion is the universal format of DD/MM/YYYY for the integrated data format mapping. The process includes the extractions of first 4 digits of year and relocates it in the 7 to 10th position of the destination date data format followed by the 6 and 7th with 4 and 5.Finally 9 and 10th with 1 and 2nd position in the destination data format.

2023/08/12 will be converted to 12/08/2023

iii. Protocols mapping

The protocol mapping of old TCP/IP protocol for data transmission with new Modbus, old HTTP protocol Universal access protocol in new IoT based device data communication for integration supports the effective data integration in legacy systems.

c. Queue based message routing.

The micro service architecture based legacy systems uses the queue based message routing for data integration process. The machine learning based message routing in queue handling mechanism performs the data integration task in an effective way. The functions are as follows, Function-1: Maintain a message buffer with head and tail information.

Function -2: Train the legacy system data sets for message validity using the head information.

Function -3*: Perform the following branching function*

If the message = request then source head is the target

Else if message=response then destination tail is the target

Else if message= error then both are target

Else if message=plain info then both are target

Function -4: Apply supervised learning for the message transformation.

Function -5: Based on head and tail details identify the Producer that adds info to the queue and consumer that extracts info from the queue.

Function -6: Halt the process until Queue length=empty.

Step-4: Perform Machine learning based application level communication approach

a. Facilitate communication through application level.

The following processes are involved in this integration task,

i. Identify the application interface for information exchange.

ii. Transform, consolidate, refine, and finally transfer the data through application user interface.

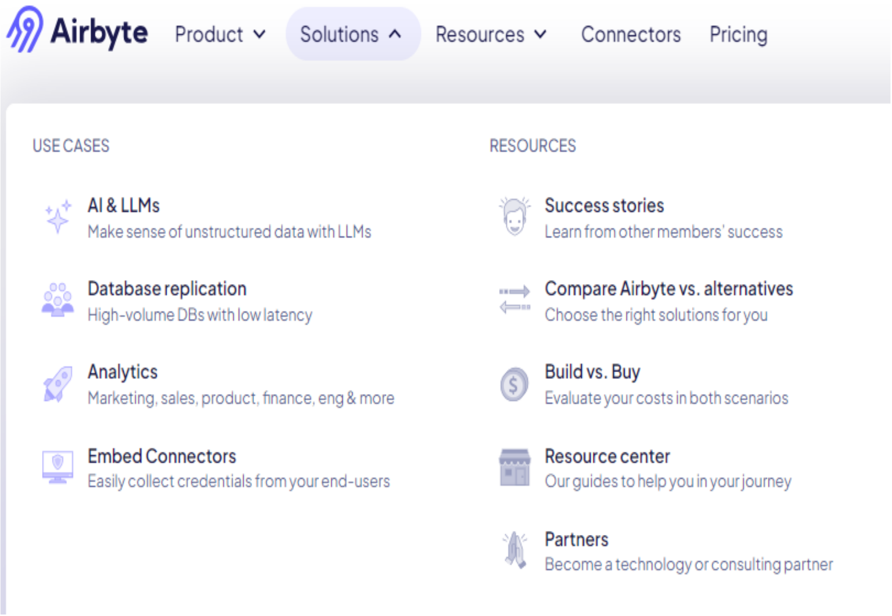
iii. Use semi supervised learning for storing and capturing the data formats.

iv. Handle the exception of downtime and error mapping.

v. Monitor and report the results to the application server.

b. Data abstraction, automation, and transfer between resources.

The tools such as air byte [11], snowflake, and apache Kafka are used for the data abstraction, automation and transfer between resources as in fig-6.



USE CASES

AI & LLMs
Make sense of unstructured data with LLMs

Database replication
High-volume DBs with low latency

Analytics
Marketing, sales, product, finance, eng & more

Embed Connectors
Easily collect credentials from your end-users

RESOURCES

Success stories
Learn from other members' success

Compare Airbyte vs. alternatives
Choose the right solutions for you

Build vs. Buy
Evaluate your costs in both scenarios

Resource center
Our guides to help you in your journey

Partners
Become a technology or consulting partner

Fig-6: Air byte tool for implementation

c. Bi direction transfer.

The 2-way flow of communication between the legacy systems for information exchange and data integration process is required for effective integration. The IO-Link wireless protocol plays the vital role in bi directional communication between legacy hardware's and software's.

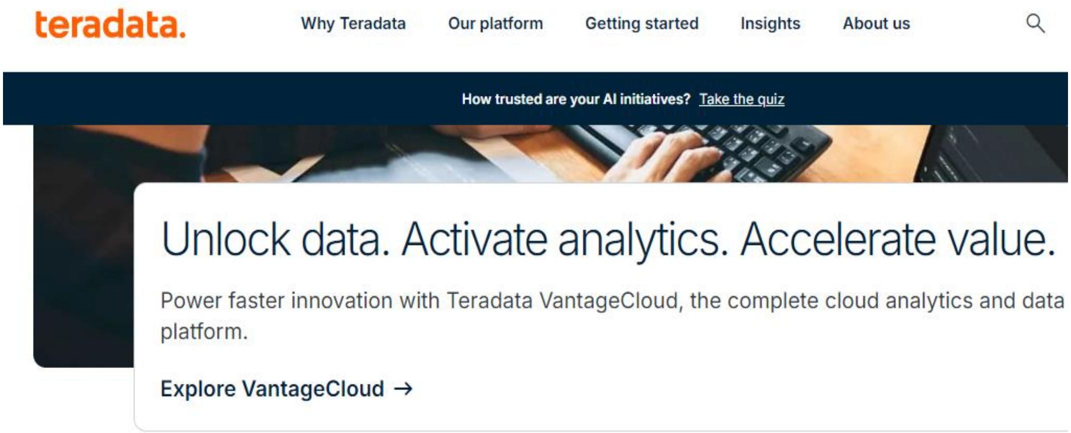
Step-5: Perform Machine learning based automated communication approach

a. Facilitate communication through automatic mode.

The process of integration based on automated database integration, data warehouse integration and data lake integration with automated synchronization using unified user interface with the power of universal data modifications.

b. Map data, database and data warehouse

The tools used for mapping are amazon redshift, tera data vantage [12] and oracle autonomous warehouse as in fig-7.



Unlock data. Activate analytics. Accelerate value.

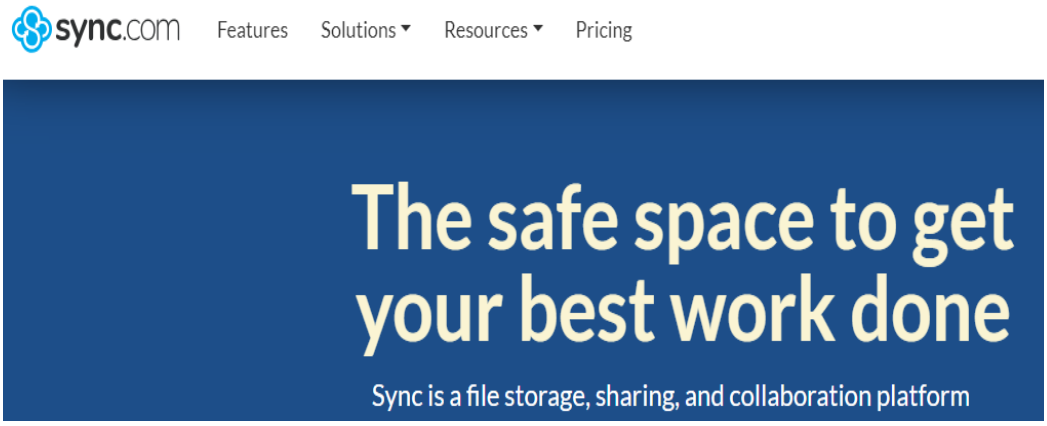
Power faster innovation with Teradata VantageCloud, the complete cloud analytics and data platform.

Fig-7: Teradata tool for implementation

c. Transfer and synchronize the data.

The synchronization of data is essential for efficient data integration in the distributed web information domain.

The tools used for the implementation are sync log [13] as in fig-8, auditor, and storage analyzer etc.



The safe space to get your best work done

Sync is a file storage, sharing, and collaboration platform

Fig-8: Sync log tool for implementation

Step-6: Perform Machine learning based cloud based communication approach

a. Facilitate communication through cloud based services.

The machine learning based data integration of legacy systems can be done through cloud based support by the following operations.

Operation-1: Build the trained framework for handling the data.

The framework acts as the base for data integration.

Operation-2: Establish communication using the connectors.

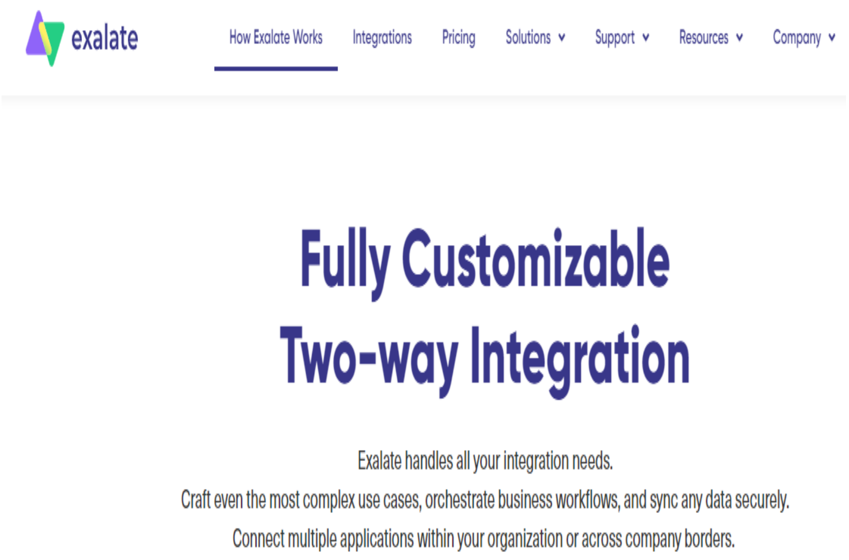
The connectors are software tools for data transfer.

Operation-3: Facilitate the automated cloud tools for entire integration.

The open source is app purchase tools are used for implementation.

b. Perform integration through framework connectors.

The tools used for cloud based legacy systems data integration using machine learning approaches are, zapier, workato and exalate [14] as in fig-9 etc.



Fully Customizable Two-way Integration

Exalate handles all your integration needs.
Craft even the most complex use cases, orchestrate business workflows, and sync any data securely.
Connect multiple applications within your organization or across company borders.

Fig-9: Exalate tool for implementation

Step-7: Verification and validation

Loop back verification and source to source mapping along with proof reading are used for validation and verification for effective legacy system based data integration process.

IV. Results and Discussion

Consider the legacy system data collection from Kaggle standard data set [14] for MS Access and github [15] for dbase with a collection of 15 legacy data sources.

The proposed methodology provides the better results by applying the proposed comprehensive legacy system based data integration approaches using machine learning techniques.

This research module produces 93% (14 out of 15 legacy system data sets) of success rate for the proposed comprehensive legacy system based data integration approaches using machine learning techniques.

The comparison metrics for parameter variation between existing and proposed methodologies with precision, recall etc. are given in Table-1 as follows,

Table-1: Proposed methodology parametric comparisons

No	Approach	Accuracy	Precision	Recall	F1 score value
1	Data mining based data integration approach.	66%	0.67	0.68	0.66
2	Comprehensive approach for data integration in legacy system based distributed web information domain using machine learning techniques.	93%	0.94	0.92	0.93

The following diagram as in fig-10 shows the execution comparison between the proposed and existing methods in the legacy system based data integration process.

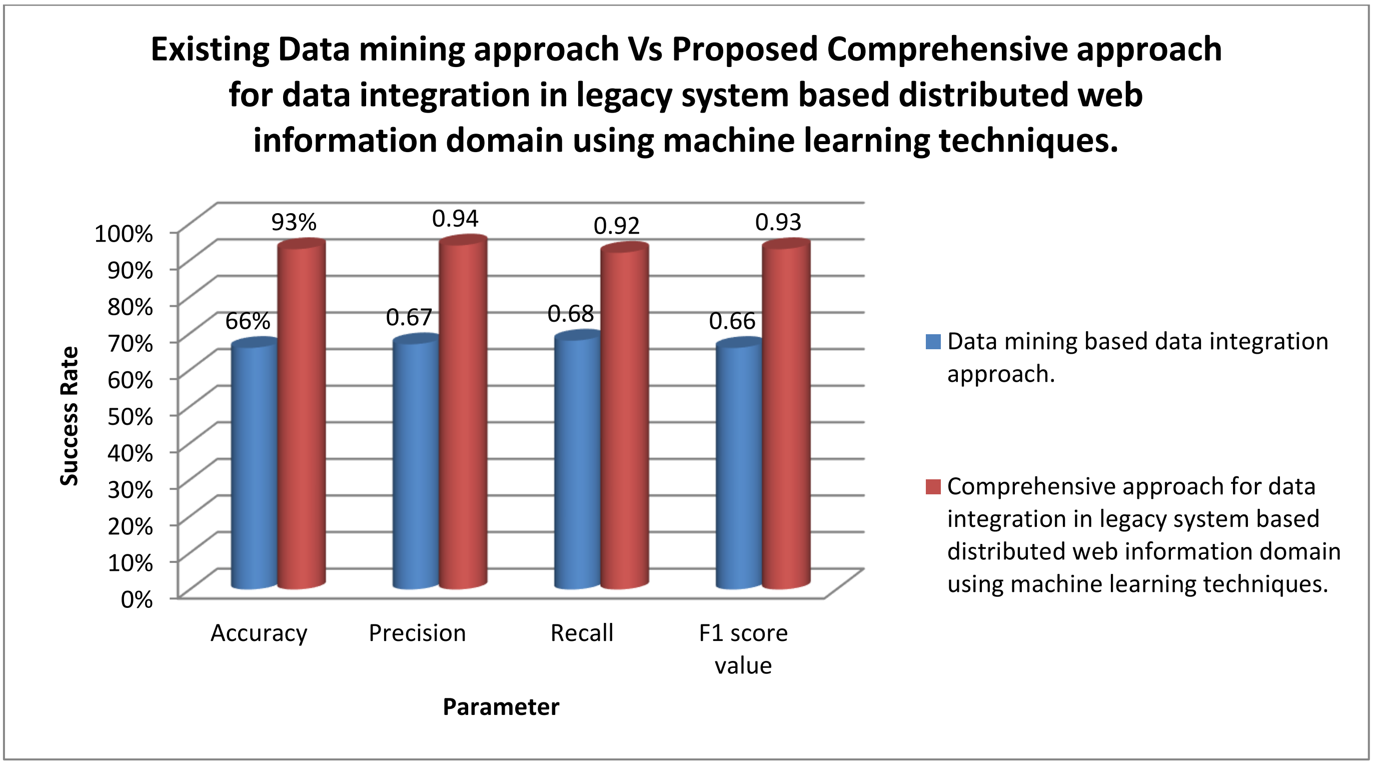


Fig-10: Proposed vs. existing methodology execution comparisons

V. Conclusion:

The legacy system based data integration is the complex and nurturing procedure due to its drawbacks of inappropriate data and with its relational functionalities. The data integration of legacy systems is very essential to cope up with the real time data processing system in all the recent domains.

This research module initially concentrates on the identification of legacy data systems coupling type such as rigid ,flexible, application level, automated and cloud based , followed by the different ways of dealing in machine learning based coupling and finally with the verification and validation procedures.

The proposed methodology provides 93% of success rate when compared with the existing data mining based data integration approach which produced only 66% success.

In future this research methodology extends its way through genetic algorithm based implementation towards the best data integration approach.

References:

[1] Alden, D.L. and Nariswari, A., 2017. Brand Positioning Strategies during Global Expansion: Managerial Perspectives from Emerging Market Firms. In The Customer is not Always Right? Marketing Orientations in a Dynamic Business World (pp. 527-530). Springer, Cham.

[2] Boso, N., Hultman, M. and Oghazi, P., 2016, July. The impact of international entrepreneurial-oriented behaviors on regional expansion: Evidence from a developing economy. In 2016 Global Marketing Conference at Hong Kong (pp. 999-1000).

[3] Boso, N., Oghazi, P. and Hultman, M., 2017. International entrepreneurial orientation and regional expansion. Entrepreneurship & Regional Development, 29(1-2), pp.4-26

[4] Nasrin JOKAR, Reza Ali HONARVAR, Shima AgHAMIRZADEH, and Khadijeh ESFANDIARI, "Web mining and Web usage mining techniques," Bulletin de la Société des Sciences de Liège, vol. 85, pp.321 - 328, 2016. Anurag Kumar and Kumar Ravi Singh, "A Study on Web Structure Mining," International Research Journal of Engineering and Technology (IRJET), vol. 04, no. 1, pp. 715-720, January 2017

[5] Dutton, T. An Overview of National AI Strategies. Available online: http://www.jaist.ac.jp (accessed on 8 January 2020).

[6] https://www.jetbrains.com/datagrip/

[7] https://xmpp.org/

[8] https://www.topdesk.com/en/

[9] https://www.trianglemicroworks.com/

[10] https://airbyte.com/

[11] https://www.teradata.com/platform

[12] https://www.sync.com/

[13] https://exalate.com/

[14] https://www.kaggle.com/datasets