International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 November 2023*

Systematic Approach for Modern data Integration in Hybrid Data System based Distributed Web Information Domain Using Machine Learning Techniques

Jinduja.S[1], Narayani.V[2]
[1]Research Scholar
Department of Computer Science
St.Xavier's College (Autonomous), Palayamkottai – 627 002
[2]Assistant Professor
Department of Computer Science
St.Xavier's College (Autonomous), Palayamkottai – 627 002
Affiliated to Manonmaniam Sundaranar University, Abishekapatti, Tirunelveli – 627 012

Abstract

The data integration process evolving from one state to another based on the data systems which are currently in use or may be updated in the future. The modern systems stored their data bases information in terms of its stream flow and big data formats are handled with data lakes and cloud support for further processing. The outdated technologies and traditional standards are not feasible to access the modern hybrid data systems if it is not handled with proper approaches and tools. The proposed data integration components for current use will also focus on the adoption for future trends and techniques which is the main draw back in the existing data integration approaches. The existing data integration methodologies mainly focus on the data content integration but not with the additional requirements, structure, and proper tool selection which leads the methodologies to deviate it from the goal achievements. This research article proposes a systematic approach for modern data integration in hybrid data system based distributed web information domain using machine learning techniques. In future this research paper will be incorporated with the implementation of automata theory based distributed web information integration system.

Keywords: Modern system, Machine learning, web data, distributed data, heterogeneous data

I. Introduction:

Modern data integration:

Data, information, and intelligence are the essential components for any modern data integration more of which must be handled in real-time.

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 6 phases in the proposed methodology for the systematic approach for modern data integration in hybrid data system based distributed web information domain using machine learning techniques. They are,

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 November 2023*

Phase-1: Requirements phase
a. Modern interface support
b. Scaling support.
c. Dynamic speed for access.
d. Ready to update any time.
e. Hybrid system support.
f. Proper framework for connectivity.
g. Authorization based monitoring.
h. Flexible to handle.
i. Remote access support.
j. Security.

Phase-2: Design phase
a. Remove user/component based timely codes.
b. Focus on transaction fields than entire old design templates.
c. Support stream and batch processing.
d. Perform initial data ordering.
e. Static storage process.
f. Multiple and variable system support.
g. Transparent monitoring.
h. Quality is better than quantity.
i. Integrated Development Environment support.
j. Feasible for service portability.

Phase-3: Content Phase
a. Multiple source and destination.
b. Capture data at once.
c. Focus on updated data alone.
d. Flexible UI.
e. Universal structured data converter.
f. Resource availability.
g. Scalability.
h. Data access restrictions.
i. Cloud content support.
j. Backup support.

Phase-4: Path phase
a. Prototype creation.
b. Real-time data integration test.
c. Perform data access service.
d. Modern platform implementation.
e. Deal all types of data components.
f. Cloud based service.

Phase-5: Tools phase
a. Enterprise or open source based tools.
b. Extraction or Workflow based tools.

Phase-6: Verification phase
a. Storage and processing.
b. Integration.
c. Transformation.
d. Analytics.
e. Observance.

The proposed methodology of systematic approach for modern data integration in hybrid data system based distributed web information domain using machine learning techniques is as follows in Fig-1.

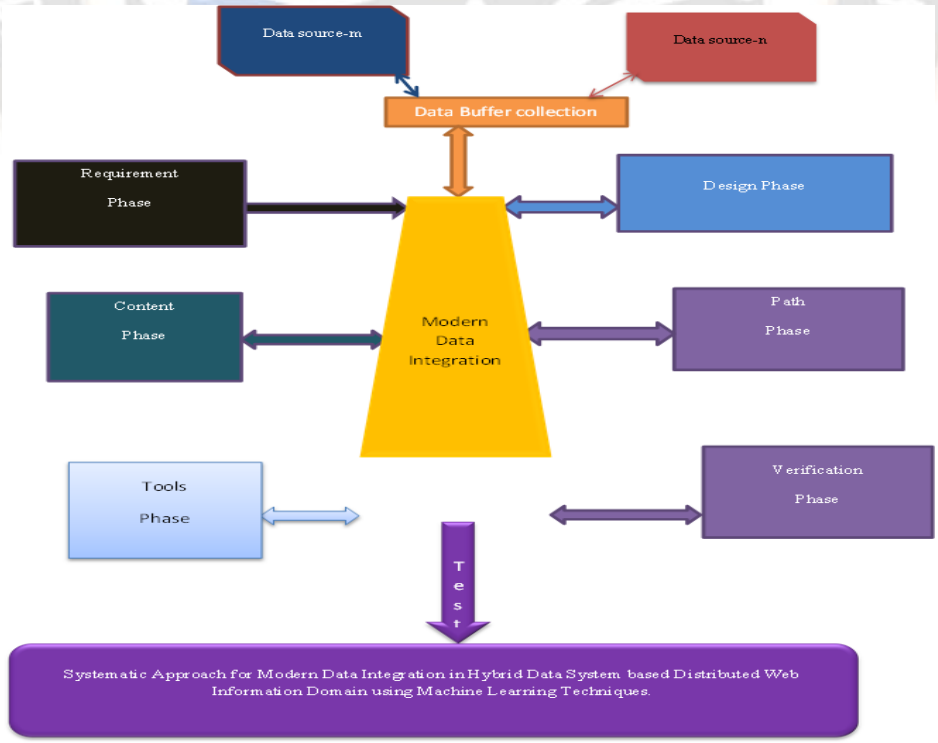

Fig-1: Proposed Systematic approach for data integration

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 November 2023*

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

Start

Input: Modern system based web informatics data collection for data integration

Step-1: Defining the requirement's phase.

a. Modern interface support
b. Scaling support.
c. Dynamic speed for access.
d. Ready to update any time.
e. Hybrid system support.
f. Proper framework for connectivity.
g. Authorization based monitoring.
h. Flexible to handle.
i. Remote access support.
j. Security.

Step-2: Apply Machine learning based design features based on requirements

a. Remove user/component based timely codes.
b. Focus on transaction fields than entire old design templates.
c. Support stream and batch processing.
d. Perform initial data ordering.
e. Static storage process.
f. Multiple and variable system support.
g. Transparent monitoring.
h. Quality is better than quantity.
i. Integrated Development Environment support.
j. Feasible for service portability.

Step-3: Specify Machine learning based data content characteristics based on requirements

a. Multiple source and destination.
b. Capture data at once.
c. Focus on updated data alone.
d. Flexible UI.
e. Universal structured data converter.
f. Resource availability.
g. Scalability.
h. Data access restrictions.
i. Cloud content support.
j. Backup support.

Step-4: Perform Machine learning based path for integration.

a. Prototype creation.
b. Real-time data integration test.
c. Perform data access service.
d. Modern platform implementation.
e. Deal all types of data components.
f. Cloud based service.

Step-5: Identify the tools for implementation.

a. Enterprise or open source based tools.
b. Extraction or Workflow based tools

Step-6: Verification and validation

a. Storage and processing.
b. Integration.
c. Transformation.
d. Analytics.
e. Observance.

End

Phase-1: Defining the requirement's phase.

The requirements phase acts as basic component in this research module due to the foundation architecture which lead to the achievement of the proposed Systematic Approach for Modern data Integration in Hybrid Data System based Distributed Web Information Domain Using Machine Learning Techniques.

The following 10 requirement phase components create the entire architecture for the data integration in modern data using machine learning approaches.

a. Modern interface support

The quick and handy application programming interface requirement satisfies the advanced data access requirement of any time anywhere by anybody policy.

b. Scaling support.

The role of big data and vast amount of resources usage in data handling process requires the scaling capability to deal with any amount of data with proper care.

The progressive nature of databases to data warehouses and then to data lakes is the solid proof for scaling support requirement in modern data integration techniques.

c. Dynamic speed for access.

The modern database integration requires the change of speed in access based on availability, resource management, and security.

d. Ready to update any time.

The sudden change in the participated entity will definitely affect the data integration if not properly updated as a whole, which lead to the anytime modifications over the duly process.

e. Hybrid system support.

The generic data system, legacy system or modern system must be dealt with proper care as a single hybrid system for future use.

f. Proper framework for connectivity.

The proper framework and integration interface requirement is needed for the efficient connection of different sets of data resources in the integration process.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 November 2023*

g. Authorization based monitoring.

The data access levels must be clearly defined for the entire data resources for maintaining and monitoring the entities for the integration process based on different authorization hierarchy.

h. Flexible to handle.

The flexibility in handling the number of transactions from high to low on each batch ensures the smooth transaction in data integration process.

i. Remote access support.

The modern era requires the data access in any place at any time which enables the availability of remote access support facility to all the data access points.

j. Security.

The entire data integration function must support security features in data protection with different policy sets for the sensitivity of data present in the data integration module.

III. Implementation

Phase-2: Apply Machine learning based design features based on requirements

a. Remove user/component based timely codes.

The machine learning approach uses the decision trees for the code maintenance as follows in fig-2,

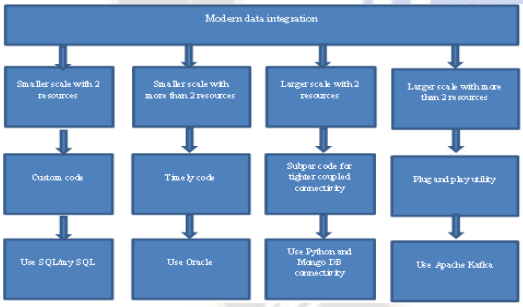

Fig-2: Machine learning based code constructs

The apache Kafka tool acts as the plug and play utility for the data integration process without caring about the coding part for any time of data users as in the fig-3.

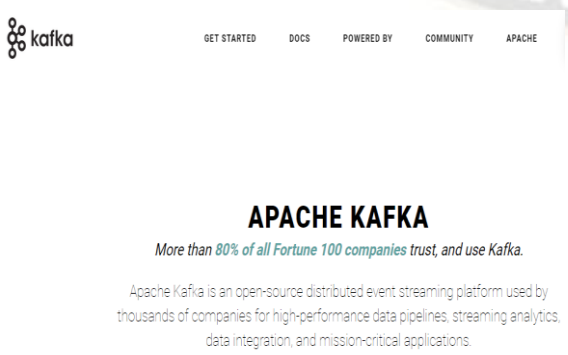

Fig-3: Apache Kafka tool page

b. Focus on transaction fields than entire old design templates.

The main focus on modern data integration is the data content integration rather than connecting all the fields. The transactional fields are the important fields for the source and destination resources in the data integration process. The linear regression in machine learning method identifies the dependent fields for the data transaction process for predicting the data integration success. The supporting fields for data integration process acts as the independent fields in the data transaction process.

Based on linear regression in data integration process,

The dependent fields: Responsible for reliable data transaction

The independent fields: support for data transaction irrespective of the data resource design.

c. Support stream and batch processing.

The machine learning based association rules structure decides the data transaction flow types as follows,

If application=Traditional then
Apply Batch processing data flow
Else if application=Custom appliances then
Apply Stream based data flow
Else if application=Modern then
Apply Hybrid data flow
Else
Apply Customized data flow
End if

d. Perform initial data ordering.

The logistic regression in machine learning sets the data with the probability of its placement prediction to its position based on the corresponding event occurrence in data ordering. The data ordering improvises the data integration process with proper references for future analysis. The data ordering is done through

- Standardization
- Normalization
- Ranking
- Sorting
- Rolling

e. Static storage process.

The modern data integration efficiency relies on the machine learning based static storage process with dynamic storage facility.

The storage process concentrates on the following as in fig-4,

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 November 2023*

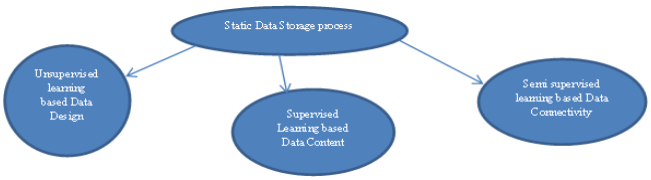

Fig-4: Storage process

f. Multiple and variable system support.

The machine learning based system support for modern data integration relies on the following constraints,

If source=legacy system then
Apply hybrid system support
Else if source=generic system then
Apply cloud system support
Else if source=modern system then
Apply multi cloud system support
Else
Apply feasible system support
End if

g. Transparent monitoring.

The monitoring components in the data transaction processing improves the operational process and controls the data are as follows in table-1

Table-1: Monitor table

Sl.No	Component
1	Source validity
2	Destination validity
3	Best source
4	Data validity
5	Integration point
6	Data route
7	Data connection
8	Volume
9	Throughput
10	Delay and gaps

The transparent monitoring includes the following priorities based on fuzzy membership values as follows in table-2.

Table-2: Monitor ordered table

Sl.No	Component	Fuzzy membership value	Priority	Final component ordering
1	Source and destination validity	1.0	1	Source and destination validity
2	Data report	0.1	10	Data connection
3	Best source	0.5	6	Data validity
4	Data validity	0.8	3	Volume
5	Integration point	0.4	7	Throughput
6	Data route	0.3	8	Best source
7	Data connection	0.9	2	Integration point
8	Volume	0.7	4	Data route
9	Throughput	0.6	5	Delay and gaps
10	Delay and gaps	0.2	9	Data report

h. Quality is better than quantity.

The process of collecting data is inferior to the process of collecting valid and relevant data. The quality of data is discriminated by the machine learning based association rules.

Rule-1: Data=Raw and safe
Process: Apply data cleansing and classification for data integration

Rule-2: Data=Generic
Process: Apply Generic or legacy data integration process

Rule-3: Data=Sharable
Process: Apply Hybrid data integration process.

Rule-4: Data=Personalized
Process: Apply Regional based policies.
Rule-5: Data=Sensitive
Process: Apply Data access level authorization.

Rule-6: Data=Classified
Process: Apply Restricted data mode with the highest level of security.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 November 2023*

i. Integrated Development Environment support.

The IDE support for modern data integration design using machine learning approach uses the supervised learning approach for the identified data systems is as represented in the following table-3.

Table-3: IDE tool selection

Sl.No	System type	Supervising learning based IDE tool selection
1	Standalone traditional system	Apache Sqoop
2	Traditional distributed system	Apache Flume
3	Modern system	Apache Kafka

j. Feasible for service portability.

The size of the cloud is cost expensive. The choice of cloud service must be directly depends on the data size and complexity levels.

Cloud service selection Cs=Data Size (N) x Complexity level (L) x Cost(C) ---------- (1)

Phase-3: Specify Machine learning based data content characteristics based on requirements

a. Multiple source and destination.

The machine learning based data content specifications for multiple source and destination are as in table-4,

Table-4: Content specifications

Same type	Different type
Same platform	Library function
Same interface	Connector codes
Same resources	Common interface

b. Capture data at once.

The machine learning based constraint structures for capturing data are,

If data type=Non transactional then
Support recapture
Else
Support Capture data at once.
End if

c. Focus on updated data alone.

The update at regular intervals by using the machine learning based dynamic programming model is better than single final updating process.

The dynamic programming model supports low latency and effective load balancing. Riggers are used for each sub parts of updating in the data modules.

d. Flexible UI.

The user interface must support the following features

- Inner and outer joins.
- Data lookup.
- Transformation.
- Duplication and
- Aggregation.

The scalable and flexible user interface such as apache airflow, IBM infosphere, oracle integrator, and aws glue are used as in fig-5.

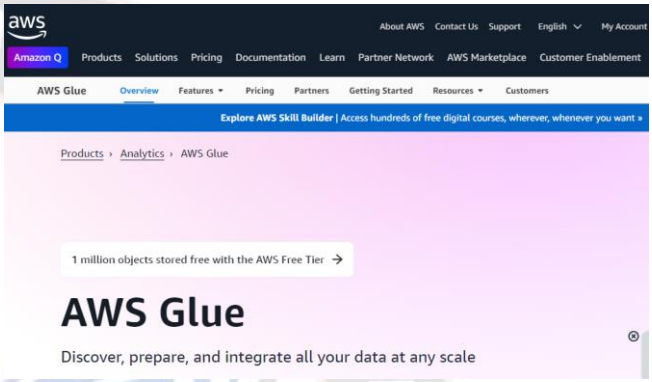

Fig-5: AWS Glue tool page

e. Universal structured data converter.

The following functions with the supervised learning approach supports the data conversion process in modern data integration.

- Change Data capture
- Extract, Transform and Load
- Extract, Load and Transform

f. Resource availability.

Implement scheduling approaches such as Fair scheduling developed by Facebook or capacity scheduling developed by yahoo based machine learning approaches for maintaining resource availability.

Fair scheduling: Equal bandwidth for all components.
Capacity scheduling: first queue service based with deficit queue earns more.

The unsupervised learning approach is used to identify and maintain the scheduling approaches.

g. Scalability.

The reinforcement learning approach based on machine learning is the key for scalability feature in modern data integration.

If workload increase then
Add new node
Else if workload decrease then

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 November 2023*

Remove unused node
Else
No change in nodes
End if

h. Data access restrictions.

The following supervised learning policy steps withhold the data access restrictions.

Step-1: Recognize the data resource type
Step-2: Extract data Roles and responsibilities
Step-3: Find Data sensitivity
Step-4: Categorize data access levels

- ✓ *NO ACCESS*
- ✓ *READ ONLY*
- ✓ *WRITE ONLY*
- ✓ *FULL ACCESS*

Step-5: Develop the manual
Step-6: Perform frequent monitor and review.

i. Cloud content support.

The cloud tools such as machine learning based Jitter bit, Pentaho, snap logic and mule soft are used for cloud content support implementation.

j. Backup support.

The tools such as amazon web services, informatica, boomi, five Tran etc. are used for back up support in modern data integration.

Phase-4: Perform Machine learning based path for integration.

a. Prototype creation.

The following steps are included in the machine learning based prototype creation for modern data integration.

Step-1: Validate the existing integration approaches.
Step-2: Create the basic modernization key areas.
Step-3: Identify the integration platform based on fuzzy logic approach.
Step-4: Implement the execution with proper tool selection.

b. Real-time data integration test.

The following functions are included in the machine learning based prototype creation for modern data integration.

Function-1: Select stream flow data approach for any real-time data.
Function -2: Select appropriate IDE.
Function -3: Analyze data based on genetic approach.
Function -4: Test with on cloud and off cloud scenarios.

c. Perform data access service.

The data access service includes the following operations.

- *Data preparation*
- *Data analysis and*
- *Data reports creation and*
- *Data presentation.*

d. Modern platform implementation.

The integration of data from SQL, Oracle etc. to Cloud requires the modern platform implementation. The machine learning based decision making schema plays the vital role in this component.

Decision Making Schema (Decision-value)
{
Decision-1: Upgrade Hardware.
Decision-2: Upgrade Software.
Decision-3: Apply cloud support.
Decision-4: Migrating data.
Decision-5: Apply advanced data integration approach.
}

e. Deal all types of data components.

The different data structures for modern data integration based on machine learning approach using classification techniques are

i. Type of data

- *Structured*
- *Semi Structured*
- *Unstructured*
- *Multi structured*
- *Raw data*

ii. Data generators

- Audi visual media
- Social media
- Emails
- Real-time events
- IoT sensors
- Triggers

iii. Data integrity process features

- Combine vast data
- Process the data
- Value added data
- Context creation
- Presentation

f. Cloud based service.

The selection of appropriate cloud based services such as Azure, IBM, and Amazon etc. produces the good results in modern data integration using cloud services.

Phase-5: Identify the tools for implementation.

a. Enterprise or open source based tools.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 November 2023*

i. Informatics Power center:

This tool is an enterprise supporting tool for modern data integration as in Fig-6

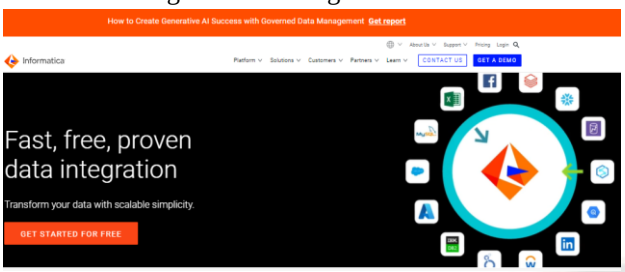

Fig-6: Enterprise tool page

b. Extraction or Workflow based tools

The sample workflow based tool is Prefect tool as in Fig-7

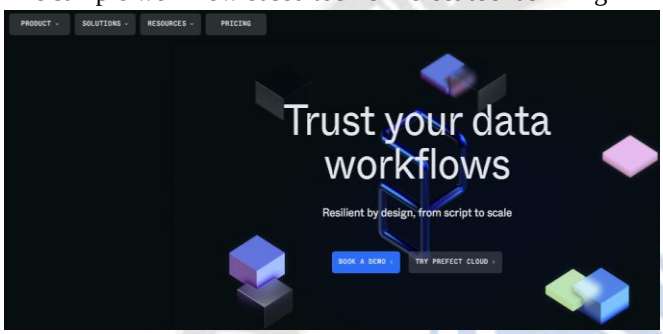

Fig-7: workflow tool page

Phase-6: Verification and validation

a. Storage and processing.

The optimal selection of data storage verification and validation through fuzzy based machine learning approach provides 3 entities.

- *Data warehouses*
- *Data Lakes*
- *Data mesh*

b. Integration.

The data integration are verified and validated through two types of data processing,

- *Batch processing*
- *Stream processing*

c. Transformation.

The data transformation verification and validation are evaluated through 3 tests.

- *Data organization test*
- *Data quality test and*
- *Usability test.*

d. Analytics.

Data analytics tools using statistical methods based on machine learning approaches are verified and validated through

- *Microsoft power BI [11]*
- *Zoho Analytics [12]*
- *Qlik [13]*

e. Observance.

The data observability test for verification and validation using machine learning approach includes the following stages

Stage-1: Track the data integration
Stage-2: Monitor the integrated data.
Stage-3: Store historical records.
Stage-4: Set Triggers and alarms.
Stage-5: Check the final presentation.

The implementation phases include various components and appropriate tool selection is based on the feasibility, availability and efficiency targeted procedures.

IV. Results and Discussion

Consider the modern system data collection from Kaggle standard data set [14], github [15], and real-time data set with a collection of 58 data sources.

The proposed methodology gives the best result through the implementation of systematic approach for modern data integration in hybrid data system based distributed web information domain using machine learning techniques.

This research module produces 98.3% (57 out of 58 hybrid system data sets) of success rate for the proposed systematic approach for modern data integration in hybrid data system based distributed web information domain using machine learning techniques.

The comparison metrics for parameter variation between existing and proposed methodologies with precision, recall etc. are given in Table-5 as follows,

Table-5: Proposed methodology parametric comparisons

No	Approach	Accuracy	Precision	Recall	F1 score value
1	Data mining based data integration approach.	62%	0.59	0.58	0.61
2	Systematic approach for modern data integration in hybrid data system based distributed web information domain using machine learning techniques.	98.3%	0.98	0.97	0.99

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 November 2023*

The following diagram as in fig-8 shows the execution comparison between the proposed and existing methods in the modern system based data integration process.

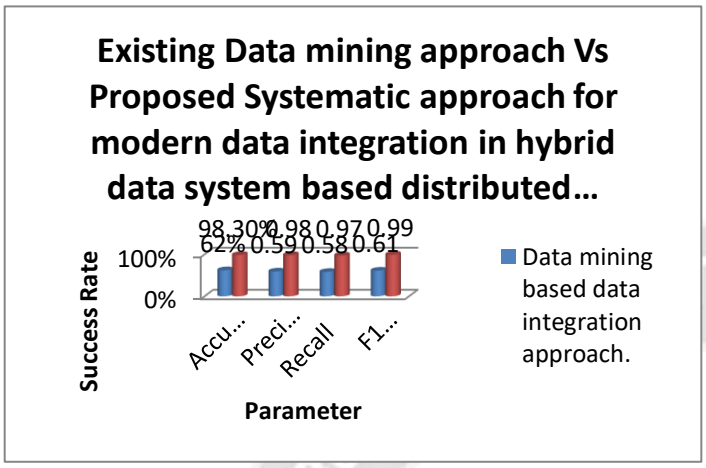

Fig-8: Proposed vs. existing methodology execution comparisons

V. Conclusion:

The modern system based data integration requires updated and layer based procedures due to its immense impact of remote access which are not handled with proper care by the existing data mining based data integration approaches. The data integration of modern systems plays the vital role in the real time data processing system for the current domains.

This research component initially focuses on the modern data integration requirements followed by the data modern data design, then with the modern data content specifications followed by the modern data integration path discretions and then finally with the proper tools selection with machine learning based verification and validation procedures.

The proposed methodology provides 98.3% of success rate [57/58 data sets] when compared with the existing data mining based data integration approach which produced only 62% success [36/58 data sets].

In future this research methodology will concentrate on the neural automation 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.
5. 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
6. Dutton, T. An Overview of National AI Strategies. Available online: http://www.jaist.ac.jp (accessed on 8 January 2020).
7. https://kafka.apache.org/
8. https://aws.amazon.com/glue/
9. https://docs.informatica.com/data-integration/powercenter.html
10. https://www.prefect.io/
11. https://www.microsoft.com/en-us/power-platform/products/power-bi
12. https://www.zoho.com/analytics/
13. https://www.qlik.com/us
14. https://www.kaggle.com/datasets
15. https://github.com/