International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8
***Article Received:** 25 June 2023 **Revised:** 12 July 2023 **Accepted:** 30 August 2023*

Deterministic Approach for data Integration in Distributed Web Information System Using Machine Learning Techniques

Jinduja.S[1], Narayani.V[2],
[1]Research Scholar, Manonmaniam Sundaranar University, Tirunelveli,
Email: jindujaseelan@gmail.com
[2]Assistant Professor, Department of Computer Science, St Xavier's College, Tirunelveli

Abstract: The day today digital life is incorporated with data in which the collection of data describes each individual in the exact form of their occurrence in our current digital world. The process of handling data from one source is not in a practical condition nowadays. Each resource handles its own form of data so that their information system dependent data is always ready for their use and moreover it's feasible for them to provide security to their data. The distributed web information system is collection of data with different formats which requires more effort to handle it in an efficient manner. The art of collecting the data, ordering the data, and integrating the data in a distributed web information system is a complex process to implement. The existing methodologies focuses on the integration of results but resulted with improper classification and duplication of information collections. This research article proposes a machine learning approach for handling heterogeneous data in distributed web information system with proper classification of data along with its unique characteristics. In future this research paper will be improved with the implementation of artificial intelligence based distributed web information system.

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

I. Introduction:

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.

Web information System:

Web information system, or web-based information system, is an information system that uses Internet web technologies to deliver information and services, to users or other information systems/applications.

II. Methodology

There are 3 stages in the proposed methodology for the deterministic approach for data integration in distributed web information system using machine learning techniques. They are,

Stage-1: Cleaning with Data Extraction

i. The data resources are

a. Databases

b. Applications

c. Files

ii. Basic Extraction methods

a. Querying

b. Scrapping

iii. Data cleaning methods are,

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8
Article Received: *25 June 2023* ***Revised:*** *12 July 2023* ***Accepted:*** *30 August 2023*

a. Remove redundancy

b. Remove incorrect data

c. Remove irrelevant data

Stage-2: Machine learning based data integration utilities incorporation

a. Handling Isolated data
b. Manage inconsistent data
c. Dealing heterogeneity
d. Scalability maintenance
e. Security handling

Stage-3: Verification and Validation

a. Loop back verification
b. Source to source mapping verification
c. Double entry check
d. Proof reading

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

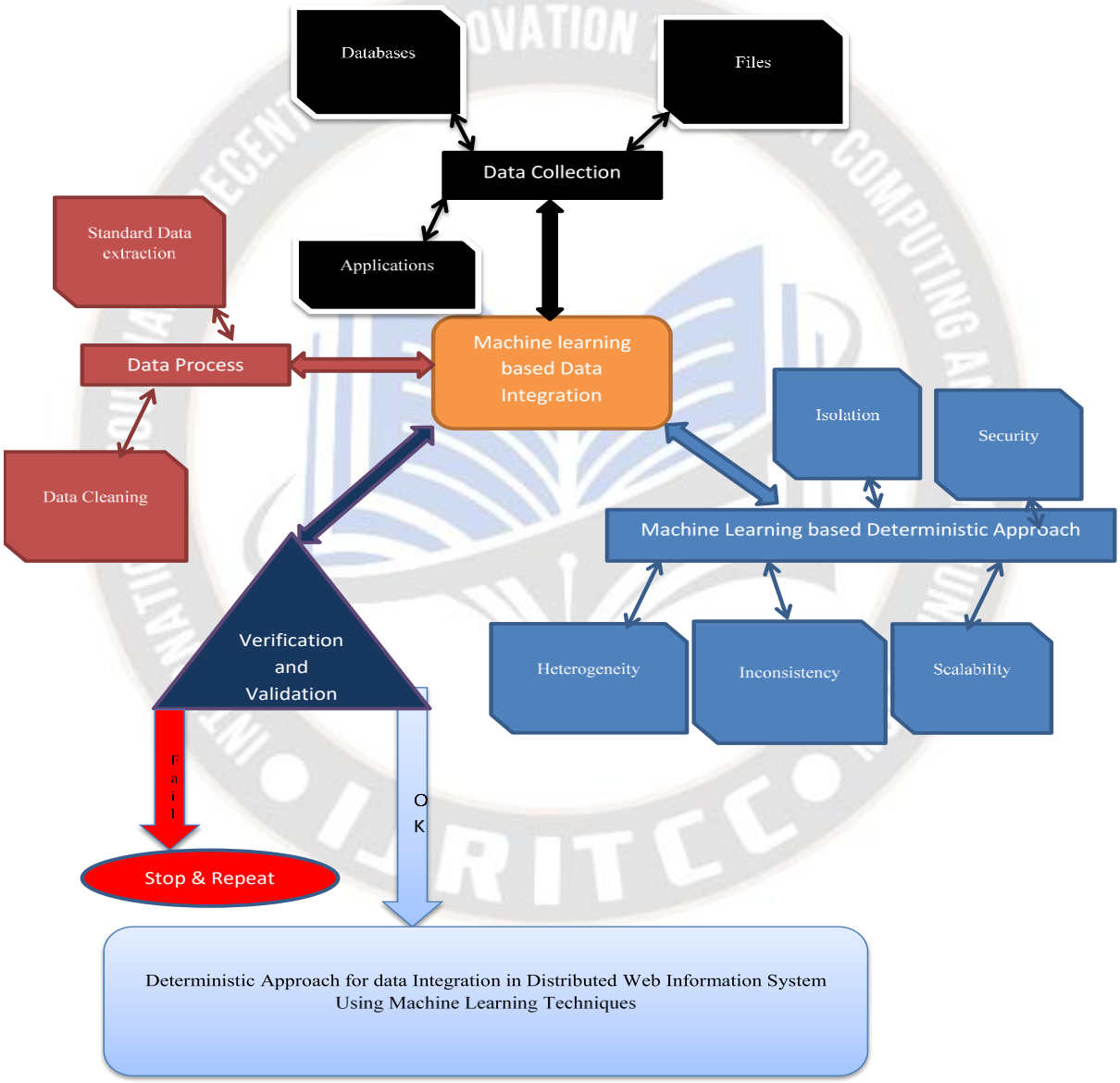

Fig-1: Proposed Deterministic approach for data integration

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

Start
Input: Heterogeneous web informatics data collection for data integration
Step-1: Apply Data extraction approaches

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8
***Article Received:** 25 June 2023 **Revised:** 12 July 2023 **Accepted:** 30 August 2023*

a. Querying
b. Scrapping
Step-2: Perform Data cleaning approaches
a. Duplicate data removal
b. Wrong data removal
c. Non-related data removal.
Step-3: Deterministic data integration approach using machine learning
1. Lone data handling mechanism.
2. Multiple resource data handling.
3. Identify and remove inconsistency.
4. Perform feasible scalability.
5. Maintain security.
Step-4: Verification and validation
Perform
1. Loop backs'.
2. Source mapping.
3. Double entry check.
4. Proof reading
If all the above are success go to end
Else go to step-1
End if
End

III. Implementation

Stage-1: Cleaning with Data Extraction
The data resources are
a. Distributed Databases:
Instead of single computers databases are stored in multiple distributed computers as distributed database (DB) in which the server client approach handles the entire process of storing and retrieving the data. The global schema defines the data transaction policies within the distributed database system as in fig-2.

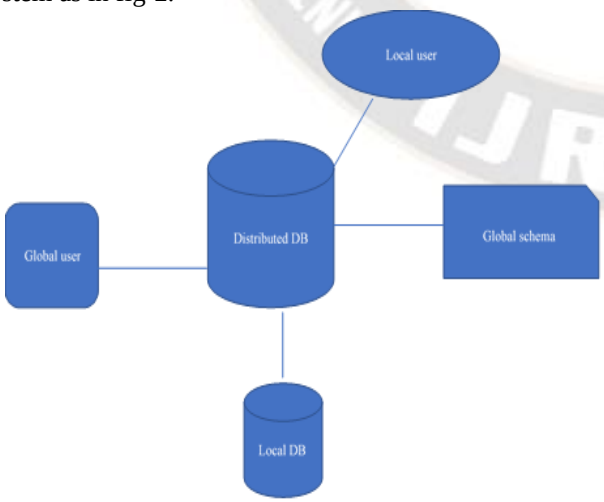

Fig-2: Distributed database set-up for proposed methodology

b. Applications:

The application's stored the data in the format of SQLite relational databases in terms of rows and columns along with images and videos. The configuration information is stored in conf databases as in fig-3.

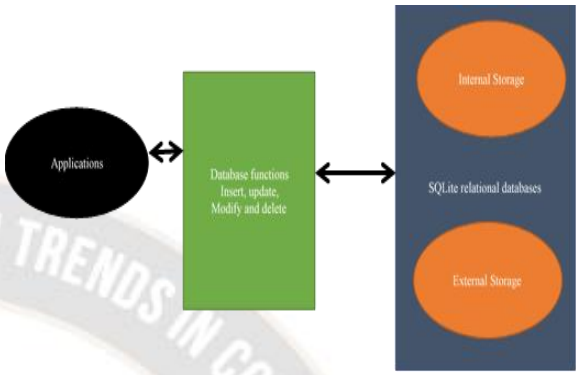

Fig-3: Application database set-up for proposed methodology

c. Files:

It contains the storage information, content, and description about the internal/external storage path along with the actual references.

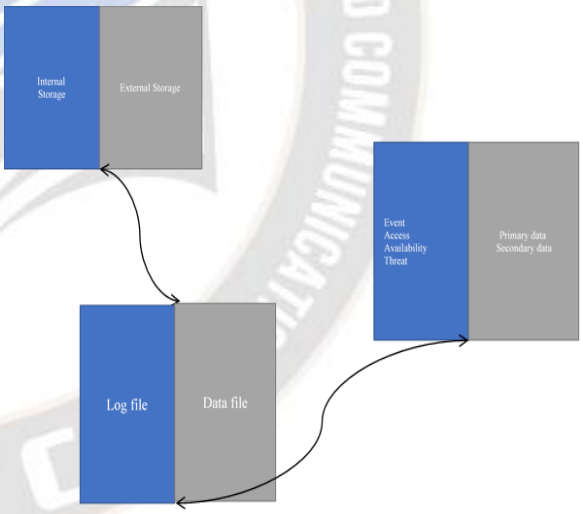

Fig-4: Files database set-up for proposed methodology

Basic Extraction methods

a. Querying:

The Structured Language Queries such as

Select,

Insert,

Update and

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8
***Article Received:** 25 June 2023 **Revised:** 12 July 2023 **Accepted:** 30 August 2023*

Delete operations are used for extracting the information's from a web informatics distributed database.

b. Scrapping

The web scraper tools are available for extracting the web information's from any data resources with the proper installation features. Some sample web scrapping tools for implementation re as follows,

- ✓ Scrappy
- ✓ Proweb scrapper
- ✓ Bright data
- ✓ Parse hub office
- ✓ Apify
- ✓ Octo parse
- ✓ Dex

The following fig-5 shows the sample web scrapping page using web scrapper tool.io [9].

	A	B	C	D	E	F	G	H
1	web-scraper-order	web-scraper-start-url	category-link	category-link-href	subcategory-link	subcategory-link-href	product-link	product-link-href
2	1520967326-47	http://webscraper.io/test-sites/e-commerce/allinone	Computers	http://webscraper.io/test-sites/e-commerce/allinone/computers	Tablets	http://webscraper.io/test-sites/e-commerce/allinone/computers/tablets	Amazon Kindle	http://webscraper.io/test-sites/e-commerce/allinon
3	1520967394-79	http://webscraper.io/test-sites/e-commerce/allinone	Computers	http://webscraper.io/test-sites/e-commerce/allinone/computers	Laptops	http://webscraper.io/test-sites/e-commerce/allinone/computers/laptops	ThinkPad X230	http://webscraper.io/test-sites/e-commerce/allinon
4	1520967406-85	http://webscraper.io/test-sites/e-commerce/allinone	Computers	http://webscraper.io/test-sites/e-commerce/allinone/computers	Laptops	http://webscraper.io/test-sites/e-commerce/allinone/computers/laptops	Lenovo IdeaPad M...	http://webscraper.io/test-sites/e-commerce/allinon
5	1520967359-62	http://webscraper.io/test-sites/e-commerce/allinone	Computers	http://webscraper.io/test-sites/e-commerce/allinone/computers	Laptops	http://webscraper.io/test-sites/e-commerce/allinone/computers/laptops	Lenovo ThinkPad...	http://webscraper.io/test-sites/e-commerce/allinon
6	1520967351-58	http://webscraper.io/test-sites/e-commerce/allinone	Computers	http://webscraper.io/test-sites/e-commerce/allinone/computers	Laptops	http://webscraper.io/test-sites/e-commerce/allinone/computers/laptops	Lenovo Legion Y7...	http://webscraper.io/test-sites/e-commerce/allinon
7	1520967381-73	http://webscraper.io/test-sites/e-commerce/allinone	Computers	http://webscraper.io/test-sites/e-commerce/allinone/computers	Laptops	http://webscraper.io/test-sites/e-commerce/allinone/computers/laptops	Toshiba Portege...	http://webscraper.io/test-sites/e-commerce/allinon
8	1520967315-42	http://webscraper.io/test-sites/e-commerce/allinone	Computers	http://webscraper.io/test-sites/e-commerce/allinone/computers	Tablets	http://webscraper.io/test-sites/e-commerce/allinone/computers/tablets	IdeaTab S5000	http://webscraper.io/test-sites/e-commerce/allinon
9	1520967293-33	http://webscraper.io/test-sites/e-commerce/allinone	Phones	http://webscraper.io/test-sites/e-commerce/allinone/phones	Touch	http://webscraper.io/test-sites/e-commerce/allinone/phones/touch	Nokia 123	http://webscraper.io/test-sites/e-commerce/allinon
10	1520967379-72	http://webscraper.io/test-sites/e-commerce/allinone	Computers	http://webscraper.io/test-sites/e-commerce/allinone/computers	Laptops	http://webscraper.io/test-sites/e-commerce/allinone/computers/laptops	MSI GL62VR 7RFX	http://webscraper.io/test-sites/e-commerce/allinon
11	1520967330-49	http://webscraper.io/test-sites/e-commerce/allinone	Computers	http://webscraper.io/test-sites/e-commerce/allinone/computers	Tablets	http://webscraper.io/test-sites/e-commerce/allinone/computers/tablets	Memo Pad HD 7	http://webscraper.io/test-sites/e-commerce/allinon
12	1520967404-84	http://webscraper.io/test-sites/e-commerce/allinone	Computers	http://webscraper.io/test-sites/e-commerce/allinone/computers	Laptops	http://webscraper.io/test-sites/e-commerce/allinone/computers/laptops	Acer Predator He...	http://webscraper.io/test-sites/e-commerce/allinon
13	1520967322-45	http://webscraper.io/test-sites/e-commerce/allinone	Computers	http://webscraper.io/test-sites/e-commerce/allinone/computers	Tablets	http://webscraper.io/test-sites/e-commerce/allinone/computers/tablets	IdeaTab A8-50	http://webscraper.io/test-sites/e-commerce/allinon
14	1520967361-63	http://webscraper.io/test-sites/e-commerce/allinone	Computers	http://webscraper.io/test-sites/e-commerce/allinone/computers	Laptops	http://webscraper.io/test-sites/e-commerce/allinone/computers/laptops	Lenovo ThinkPad...	http://webscraper.io/test-sites/e-commerce/allinon
15	1520967283-28	http://webscraper.io/test-sites/e-commerce/allinone	Phones	http://webscraper.io/test-sites/e-commerce/allinone/phones	Touch	http://webscraper.io/test-sites/e-commerce/allinone/phones/touch	Ubuntu Edge	http://webscraper.io/test-sites/e-commerce/allinon
16	1520967398-81	http://webscraper.io/test-sites/e-commerce/allinone	Computers	http://webscraper.io/test-sites/e-commerce/allinone/computers	Laptops	http://webscraper.io/test-sites/e-commerce/allinone/computers/laptops	Dell Latitude 54...	http://webscraper.io/test-sites/e-commerce/allinon
17	1520967357-61	http://webscraper.io/test-sites/e-commerce/allinone	Computers	http://webscraper.io/test-sites/e-commerce/allinone/computers	Laptops	http://webscraper.io/test-sites/e-commerce/allinone/computers/laptops	Toshiba Portege...	http://webscraper.io/test-sites/e-commerce/allinon
18	1520967408-86	http://webscraper.io/test-sites/e-commerce/allinone	Computers	http://webscraper.io/test-sites/e-commerce/allinone/computers	Laptops	http://webscraper.io/test-sites/e-commerce/allinone/computers/laptops	Toshiba Portege...	http://webscraper.io/test-sites/e-commerce/allinon
19	1520967305-37	http://webscraper.io/test-sites/e-commerce/allinone	Computers	http://webscraper.io/test-sites/e-commerce/allinone/computers	Tablets	http://webscraper.io/test-sites/e-commerce/allinone/computers/tablets	Galaxy Note	http://webscraper.io/test-sites/e-commerce/allinon
20	1520967375-70	http://webscraper.io/test-sites/e-commerce/allinone	Computers	http://webscraper.io/test-sites/e-commerce/allinone/computers	Laptops	http://webscraper.io/test-sites/e-commerce/allinone/computers/laptops	ThinkPad X240	http://webscraper.io/test-sites/e-commerce/allinon
21	1520967309-39	http://webscraper.io/test-sites/e-commerce/allinone	Computers	http://webscraper.io/test-sites/e-commerce/allinone/computers	Tablets	http://webscraper.io/test-sites/e-commerce/allinone/computers/tablets	MeMo PAD FHD 10	http://webscraper.io/test-sites/e-commerce/allinon
22	1520967346-56	http://webscraper.io/test-sites/e-commerce/allinone	Computers	http://webscraper.io/test-sites/e-commerce/allinone/computers	Laptops	http://webscraper.io/test-sites/e-commerce/allinone/computers/laptops	Asus ROG Strix G...	http://webscraper.io/test-sites/e-commerce/allinon
23	1520967355-60	http://webscraper.io/test-sites/e-commerce/allinone	Computers	http://webscraper.io/test-sites/e-commerce/allinone/computers	Laptops	http://webscraper.io/test-sites/e-commerce/allinone/computers/laptops	Asus ASUSPRO B94...	http://webscraper.io/test-sites/e-commerce/allinon
24	1520967328-48	http://webscraper.io/test-sites/e-commerce/allinone	Computers	http://webscraper.io/test-sites/e-commerce/allinone/computers	Tablets	http://webscraper.io/test-sites/e-commerce/allinone/computers/tablets	Asus MeMO Pad	http://webscraper.io/test-sites/e-commerce/allinon
25	1520967279-26	http://webscraper.io/test-sites/e-commerce/allinone	Phones	http://webscraper.io/test-sites/e-commerce/allinone/phones	Touch	http://webscraper.io/test-sites/e-commerce/allinone/phones/touch	Iphone	http://webscraper.io/test-sites/e-commerce/allinon
26	1520967363-64	http://webscraper.io/test-sites/e-commerce/allinone	Computers	http://webscraper.io/test-sites/e-commerce/allinone/computers	Laptops	http://webscraper.io/test-sites/e-commerce/allinone/computers/laptops	Apple MacBook Ai...	http://webscraper.io/test-sites/e-commerce/allinon
27	1520967387-76	http://webscraper.io/test-sites/e-commerce/allinone	Computers	http://webscraper.io/test-sites/e-commerce/allinone/computers	Laptops	http://webscraper.io/test-sites/e-commerce/allinone/computers/laptops	Dell Latitude 54...	http://webscraper.io/test-sites/e-commerce/allinon
28	1520967291-32	http://webscraper.io/test-sites/e-commerce/allinone	Phones	http://webscraper.io/test-sites/e-commerce/allinone/phones	Touch	http://webscraper.io/test-sites/e-commerce/allinone/phones/touch	LG Optimus	http://webscraper.io/test-sites/e-commerce/allinon
29	1520967373-69	http://webscraper.io/test-sites/e-commerce/allinone	Computers	http://webscraper.io/test-sites/e-commerce/allinone/computers	Laptops	http://webscraper.io/test-sites/e-commerce/allinone/computers/laptops	Hewlett Packard...	http://webscraper.io/test-sites/e-commerce/allinon
30	1520967303-36	http://webscraper.io/test-sites/e-commerce/allinone	Computers	http://webscraper.io/test-sites/e-commerce/allinone/computers	Tablets	http://webscraper.io/test-sites/e-commerce/allinone/computers/tablets	iPad Mini Retina	http://webscraper.io/test-sites/e-commerce/allinon
31	1520967391-78	http://webscraper.io/test-sites/e-commerce/allinone	Computers	http://webscraper.io/test-sites/e-commerce/allinone/computers	Laptops	http://webscraper.io/test-sites/e-commerce/allinone/computers/laptops	Asus ROG Strix G...	http://webscraper.io/test-sites/e-commerce/allinon
32	1520967307-38	http://webscraper.io/test-sites/e-commerce/allinone	Computers	http://webscraper.io/test-sites/e-commerce/allinone/computers	Tablets	http://webscraper.io/test-sites/e-commerce/allinone/computers/tablets	Galaxy Note	http://webscraper.io/test-sites/e-commerce/allinon
33	1520967344-55	http://webscraper.io/test-sites/e-commerce/allinone	Computers	http://webscraper.io/test-sites/e-commerce/allinone/computers	Laptops	http://webscraper.io/test-sites/e-commerce/allinone/computers/laptops	Asus ROG Strix S...	http://webscraper.io/test-sites/e-commerce/allinon
34	1520967367-66	http://webscraper.io/test-sites/e-commerce/allinone	Computers	http://webscraper.io/test-sites/e-commerce/allinone/computers	Laptops	http://webscraper.io/test-sites/e-commerce/allinone/computers/laptops	Dell Latitude 54...	http://webscraper.io/test-sites/e-commerce/allinon
35	1520967396-80	http://webscraper.io/test-sites/e-commerce/allinone	Computers	http://webscraper.io/test-sites/e-commerce/allinone/computers	Laptops	http://webscraper.io/test-sites/e-commerce/allinone/computers/laptops	Asus ZenBook UX5...	http://webscraper.io/test-sites/e-commerce/allinon
36	1520967389-77	http://webscraper.io/test-sites/e-commerce/allinone	Computers	http://webscraper.io/test-sites/e-commerce/allinone/computers	Laptops	http://webscraper.io/test-sites/e-commerce/allinone/computers/laptops	Apple MacBook Ai...	http://webscraper.io/test-sites/e-commerce/allinon
37	1520967371-68	http://webscraper.io/test-sites/e-commerce/allinone	Computers	http://webscraper.io/test-sites/e-commerce/allinone/computers	Laptops	http://webscraper.io/test-sites/e-commerce/allinone/computers/laptops	Apple MacBook Pr...	http://webscraper.io/test-sites/e-commerce/allinon
38	1520967400-82	http://webscraper.io/test-sites/e-commerce/allinone	Computers	http://webscraper.io/test-sites/e-commerce/allinone/computers	Laptops	http://webscraper.io/test-sites/e-commerce/allinone/computers/laptops	Asus VivoBook Pr...	http://webscraper.io/test-sites/e-commerce/allinon
39	1520967311-40	http://webscraper.io/test-sites/e-commerce/allinone	Computers	http://webscraper.io/test-sites/e-commerce/allinone/computers	Tablets	http://webscraper.io/test-sites/e-commerce/allinone/computers/tablets	Galaxy Tab	http://webscraper.io/test-sites/e-commerce/allinon
40	1520967287-30	http://webscraper.io/test-sites/e-commerce/allinone	Phones	http://webscraper.io/test-sites/e-commerce/allinone/phones	Touch	http://webscraper.io/test-sites/e-commerce/allinone/phones/touch	Nokia X	http://webscraper.io/test-sites/e-commerce/allinon
41	1520967369-67	http://webscraper.io/test-sites/e-commerce/allinone	Computers	http://webscraper.io/test-sites/e-commerce/allinone/computers	Laptops	http://webscraper.io/test-sites/e-commerce/allinone/computers/laptops	Dell Latitude 55...	http://webscraper.io/test-sites/e-commerce/allinon
42	1520967340-54	http://webscraper.io/test-sites/e-commerce/allinone	Computers	http://webscraper.io/test-sites/e-commerce/allinone/computers	Tablets	http://webscraper.io/test-sites/e-commerce/allinone/computers/tablets	Lenovo IdeaTab	http://webscraper.io/test-sites/e-commerce/allinon
43	1520967353-59	http://webscraper.io/test-sites/e-commerce/allinone	Computers	http://webscraper.io/test-sites/e-commerce/allinone/computers	Laptops	http://webscraper.io/test-sites/e-commerce/allinone/computers/laptops	Asus ROG Strix G...	http://webscraper.io/test-sites/e-commerce/allinon
44	1520967320-44	http://webscraper.io/test-sites/e-commerce/allinone	Computers	http://webscraper.io/test-sites/e-commerce/allinone/computers	Tablets	http://webscraper.io/test-sites/e-commerce/allinone/computers/tablets	MeMO Pad 7	http://webscraper.io/test-sites/e-commerce/allinon
45	1520967285-29	http://webscraper.io/test-sites/e-commerce/allinone	Phones	http://webscraper.io/test-sites/e-commerce/allinone/phones	Touch	http://webscraper.io/test-sites/e-commerce/allinone/phones/touch	Sony Xperia	http://webscraper.io/test-sites/e-commerce/allinon
46	1520967301-35	http://webscraper.io/test-sites/e-commerce/allinone	Computers	http://webscraper.io/test-sites/e-commerce/allinone/computers	Tablets	http://webscraper.io/test-sites/e-commerce/allinone/computers/tablets	Galaxy Note 10.1	http://webscraper.io/test-sites/e-commerce/allinon
47	1520967334-51	http://webscraper.io/test-sites/e-commerce/allinone	Computers	http://webscraper.io/test-sites/e-commerce/allinone/computers	Tablets	http://webscraper.io/test-sites/e-commerce/allinone/computers/tablets	Galaxy Tab 3	http://webscraper.io/test-sites/e-commerce/allinon
48	1520967313-41	http://webscraper.io/test-sites/e-commerce/allinone	Computers	http://webscraper.io/test-sites/e-commerce/allinone/computers	Tablets	http://webscraper.io/test-sites/e-commerce/allinone/computers/tablets	Galaxy Tab 4	http://webscraper.io/test-sites/e-commerce/allinon

Fig-5: Sample scrapper tool implementation

Data cleaning methods are,

a. Remove redundancy

The machine learning based redundancy removal approach includes the following operations

Step-1: Apply reinforcement learning for trial and error mapping

Step-2: Break the bigger distributed tables to smaller ones.

Step-3: Apply second Normal form 2NF such that each row is entirely dependent on the primary key alone.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8
***Article Received:** 25 June 2023 **Revised:** 12 July 2023 **Accepted:** 30 August 2023*

Step-4: If error occurs learn it for redundancy remove the record otherwise continue

Step-5: Continue steps 1to4 until all the smaller tables are satisfied.

b. Remove incorrect data

The machine learning based incorrect data removal approach includes the following operations

Step-1: Apply supervised learning with prior knowledge of correct data format.

Step-2: Split the bigger distributed tables to smaller table of rows and columns.

Step-3: Create valid file and invalid file lists along with discrepancies list.

Step-4: Apply Search function for discrepancies list.

Step-5: If error occurs place it in invalid file otherwise place it in valid file list.

Step-6: Continue steps 1to 5 until all the tables are classified.

c. Remove irrelevant data

The machine learning based irrelevant data removal approach includes the following operations

Step-1: Apply unsupervised learning approach with accept or delete modes.

Step-2: Split the bigger distributed tables to smaller table of rows and columns.

Step-3: Apply filters with proper criteria or conditions for data component necessity.

Step-4: If unnecessary data found then delete else accept.

Step-6 : Continue steps 1to 4 until all the tables are removed with irrelevant informations.

Stage-2: Machine learning based data integration utilities incorporation

a. Handling Isolated data

The machine learning based isolated data handling approach includes the unsupervised learning structure by performing the following operations.

i. Operation-1:

Perform data centralization

The isolated database information's are analyzed with unsupervised learning information's for type, format, and content segregation. The servers transfer the data to the centralized storage after the proper database file creation using web informatics as in fig-6.

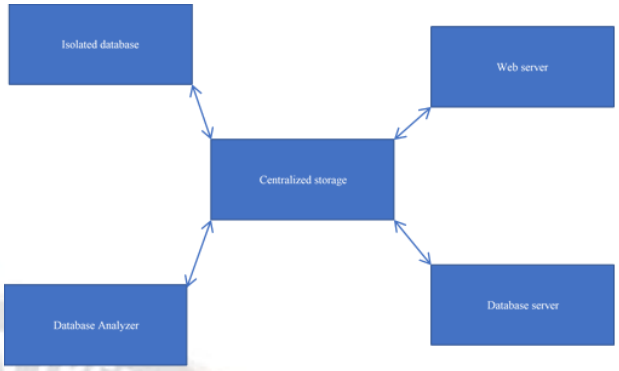

Fig-6: Centralized data storage set-up for proposed methodology

ii. Opertaion-2:

Use script's to transfer isolated unit to warehouse

The individual scripts are used to transfer data from isolated restricted databases to centralized data storage. The tools used are,

- SQL script
- Python
- Perl
- Java script

iii. Operation-3:

Apply ETL (Extract, Transform, and Load) tools to manage the data in a centralized data warehouse.

The familiar ETL tools used for centralized data ware house are as follows,

- Talend
- Stitch
- Informatics power center
- AWS Glue

The sample stitch interface [10] is as represented in the following fig-7

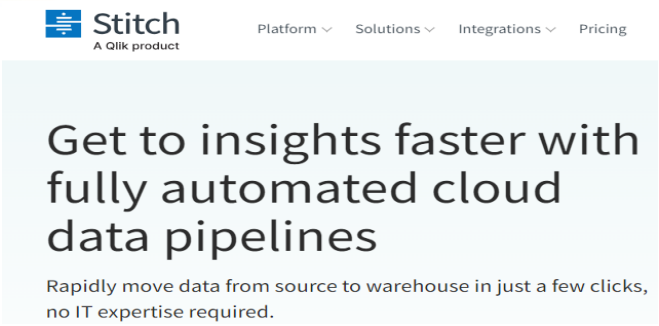

Fig-7: Sample scrapper tool page

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8
***Article Received:** 25 June 2023 **Revised:** 12 July 2023 **Accepted:** 30 August 2023*

Operation-4: Build Data Governance framework

The centralized data collection after the ETL tool implementation leads to data governance framework which is the main responsible unit to implement the policies, rules, and constraints to every data organization in the distributed web resource environment.

b. Manage inconsistent data

The tackling processes used for managing inconsistent data are,

- Track the missing data
- Replace missing data
- Remove outliers
- Perform data standardization

The tools used for the processes are,

- ✓ ArcESB
- ✓ Xplenty
- ✓ Automate.io

c. Dealing heterogeneity

The heterogeneous data handling is easily implemented through his machine learning based pattern mapping with proper data format matching.

Databases used for handling the data heterogeneity are,

- ❖ Hadoop,
- ❖ Sparkle and
- ❖ NOSql

d. Scalability maintenance

The increased demand, user growth and data volume plays the vital role in scalability issues in data integration.

The supervised learning in machine learning approach maintains the scalability issue with the following functions,

i. Function-1:

The architecture optimization operation process initializes the application and resource management with proper planning for its future growth.

ii. Function-2:

The distributed database load balancing strategy plays the vital role in scalability in data integration.

iii. Function-3:

Shading and replication handling are the processes used for reduce the scaling impact.

iv. Function-4:

The vertical scaling of increasing the servers and horizontal scaling of upgrading the servers are supporting functions for maintaining the scalability.

e. Security handling

The Policy Based Access Control (PBAC) technique is used for the security in data integrated centralized web resources. Policy creation, authorization, auditing, and implementation are centralized in the integrated data domain.

Stage-3: Verification and Validation

The verification and validation process includes the basic functionalities which act as a tool to test the integration results are correct or not.

The following four processes confirm the verification and validation of the proposed methodology results.

a. Loop back verification

Information extracted from one system matches with the results collected form the other system which is in use for further processing without any issues.

b. Source to source mapping verification

Join the two data sources together and perform comparison for any differences which provides the source to source mapping verification.

c. Double entry check

The password checking process for double entry acts as the similar role in this data validation process.

d. Proof reading

The randomized data blocks are checked for errors and issues if not then another block of data is selected for the validation process. One third of data are checked for every data transmission process.

The sample software tools used for the data verification and validation are as follows,

Software Tools:

- ✓ DataDeck
- ✓ Panoply

IV. Results and Discussion

Consider the heterogeneous data collections from Kaggle standard data set [7] and Paper code [8] with a collection of 12 data resources.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8
***Article Received:** 25 June 2023 **Revised:** 12 July 2023 **Accepted:** 30 August 2023*

The proposed methodology gives the better data integration results by applying the proposed deterministic data integration approaches using machine learning techniques.

This research article gives 91.66% (11 out of 12 heterogeneous data sets) of success rate for the proposed deterministic data integration approaches using machine learning techniques.

The parametric comparison between existing and proposed methods with precision, accuracy etc. are represented in the below Table-1 format,

Table-1: Proposed methodology parametric comparisons

No	Approach	Accuracy	Precision	Recall	F1 score value
1	Data mining based data integration approach.	58.33%	0.58	0.57	0.59
2	Deterministic approach for data integration in distributed web information system using machine learning techniques.	91.66%	0.90	0.92	0.91

The following fig-8 shows the performance comparison between the proposed and existing methodologies.

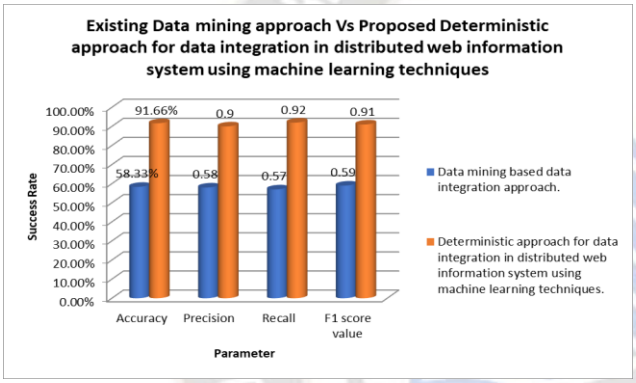

Fig-8: Proposed vs. existing methodology performance comparisons

V. Conclusion:

Data integration from heterogeneous data sets is an interesting process if every integrating components works in a proper chain of planned events in order to achieve the optimal results otherwise it produces hazard of data with more complexities. This research module proposed several stages of implementation starting with the data cleaning, followed by data balancing then with the machine learning based data integration and finally it includes the process of verification and validation. The proposed methodology gives 91.66% of success rate when compared with the existing data mining based data integration approach which produced only 58.33% success. Heterogeneous data/mixture analysis technology is expected to play a significant role in almost all the domains. In future this research methodology will be incorporated with artificial intelligence based machine learning implementing 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:www.kaggle.com/datasets/

[8] https://paperswithcode.com/datasets?data

[9] https://webscraper.io/

[10] https://www.stitchdata.com/