ISSN No:-2456-2165

A Framework for the Integration of Big Data from Heterogenous Sources to a Repository

[1] Trust. A. C. [2] Bennett. O. E. [3] Matthias. D.
Department of Computer Science
Rivers State University, Port Harcourt, Nigeria

Abstract:- Heterogeneity is one of the essential features of big data. Integrating heterogeneous data sources with problems such as data-value conflicts, data representation conflicts, data-unit conflicts, and data precision conflicts, and data movement from heterogeneous sources into a central database are challenging. The proposed framework uses extract, transform and load technologies to integrate and publish data from multiple sources. The system supports a wide range of database sources and targets to potentially transform the data to a unified knowledge base. The System adopts an Object-Oriented Design Method (OODM) and provides flexible and independent design and management of ETL processes in order to solve the data integration problems. The System has the ability to transform the contents of the databases, its representation as to their data fields, resolve data type incompatibilities, semantic incompatibilities to a unified view for a user.

I. INTRODUCTION

Data integration started from the get-go during the 1980s in which joining the various combination of information sources is generally view as data Storehouses [10]. The main strategy for data integration was completed by college of 'Minnesota' in the year 1991 and was then used for 'Integrated Public Used Microdata services (IPUMS). This applied the customary information stockroom innovation and ETL way to deal with arranged information into unified schema [8]. In the year 2009, the methodology of data integration was changed to get to and recovering information straight from the source by applying the Service Oriented Architecture which depends mainly on mapping. This method of mapping is done in two approaches which are the GAV and LAV approach. The justification of the difference in the approach was mainly because of changes to organization requests in getting to continuous information through an interceded means [7].

In 2010, semantic clashes emerged and specialists looked to determine this issue between heterogeneous sources. One basic approach to determine this issue was applying the utilization of ontologies, which is addressed as cosmology based information integration [7]. In 2011 information seclusion issue emerged which came about to the improvement of urgent information model. One strategy proposed to dispose of this issue was improved information representation. This model was intended to redesigned information model by increase and by organizing meta-information in a standard type of information elements. The term heterogeneous data are any data that has high variability of data formats and types. They are ambiguous and low quality as a result of missing values, high data redundancy, and untruthfulness. It is much more difficult to integrate heterogeneous data in order to meet the demands of business information.

Data integration is an aspect of research that solves a pervasive challenge faced in application that will need to query over multiple autonomous data sources [5]. AutoMed is a technique to heterogeneous data integration and transformation based on the use of a reversible schema transformation sequences, which gives the capability to carry out data integration over heterogeneous sources. The most applied and mature technique to data integration is the Extract, Transform & Load Technology. Though it has some deficiencies in that it does not support types of data sources such as websites, spatial, or biomedical. ETL tools are only used for data integration and do not provide an analytical environment.

Integrating heterogeneous data sources is challenging. Thus, the problem of combining data coming from different sources, and providing a user with unified view of these data needs to be addressed. Some of such problems include:

i. Semantic mismatch between schemata is also a problem as different schemas may represent the same information in different ways.
ii. Determining which data should be merged may not be clear at the outset
iii. The unique identifiers between records of two different datasets often do not exist.

The aim of this research is to develop a framework for the integration of big data from heterogeneous data sources to a single repository.

The Objectives to achieving this aim are to:

i. Resolve Semantic mismatch between schemata.
ii. Define the type of data to be merged.
iii. Provide access to integrated data and methods of creating connections between specific data elements.
iv. Integrate any type of data which will fit in the operational needs of the user.
v. Test and evaluate that data sources belonging to disparate domains have a unified view.

ISSN No:-2456-2165

II. RELATED WORKS

The paper 'Representational Sate Transfer (REST) framework' works by applying logical data warehouse to information mix for examination in which explicit data can then be gotten from singular data set by means of sending queries to the supporter using HTTP Concept in which the client's data set embraces REST interface.

This therefore means that, the utilization of REST interface on data set for integration will then be a middle person between the information source and the customer information stockroom where the cycle of combination is done. The REST view innovation makes it conceivable to distinguish data without replicating the whole data from data set which indicates a logical data warehouse. The work endeavours to utilize REST interface on any of the devices or tools for data integration in which data can overcome a specific inquiry obtained using HTTP concept. The inquiry that a client sends is finished by explicit aggregated data and is done when required, along with some degree of arrangements between the benefactor and the client. Logical data warehouse that is another data Management is utilized for series of investigation. With consistent data warehousing diverse enterprise data will then be viewed as one particular data store house for huge scope data.

Utilizing this system, data can be abandoned in its unique source instead of replicating the information into a focal database; the structure will permit one to obtain data by means of a mediator (Restif going about as miniature service) that makes a line of correspondence along these lines leaving data in their unique sources. It additionally lessens intricacy in information displaying and this is on the grounds that Pyrrho Database which has REST view innovation doesn't have to re-compose programming code.

It advances data transparency since there will be an arrangement between the client and the supporter in giving the perspectives as indicated by SLA and along these lines, the client works with the perspectives given. This system can likewise permit one to examine live data. This structure lessens data redundancy since it doesn't urge one to duplicate all data from various sources to one focal data vault.

The paper "A federated ontology-driven query-centric approach" portrays the information integration area of 'Intelligent Data Understanding System' (INDUS) – a deliberate, extensible, stage free climate for information reconciliation, information driven and information acquired from heterogeneous, circled, independent data sources. This work, when outfitted with AI algorithms for ontology guided information acquired can motivate the speed of uncovering in emerging information rich spaces (e.g., organic sciences, environmental sciences, financial matters, safeguard, humanistic systems) by enabling specialists and leaders rapidly and deftly research and look at immense proportions of data from various sources.

As checked above, INDUS executes a unified, query driven technique to manage data extraction and incorporation gotten from various heterogeneous, circled and independent data sources. The structure utilizes a 3-layer design including the actual layer, ontological layer then the UI layer. The actual layer permits the system to talk with information sources. It relies upon united informational collection engineering. The Ontological layer has global ontology demonstrated by customers and its mapping to close ontologies related with the information sources. It normally changes inquiries conveyed with respect to thoughts in a global ontology to plans for execution. The plans portray what information to isolate from each of the data source and then how to integrate its results. Finally, the UI layer enables a user to collaborate with the system, describe ontologies, post inquiries and find solutions. The multifaceted nature related with its route toward social event the information is stowed away from the last user. Every data source is presented as a vault of examples of concepts related with the data source. Every concept is basically a combination of occasions or records in a relational database for instance a bunch of tables, and a bunch of connection between the pair of tables. Input from a user (analyst) incorporates: an ontology which associates the distinctive data sources from the user's viewpoint, executable code which has the ability to perform express estimations (in case they are not maintained directly from the data sources) and an inquiry conveyed in regards to user 'specified ontology'. This enables a scientist to concentrate and integrate data from different data sources and store up its results in a relational database that is coordinated by their ontology and can be controlled using an application program or a relational database task.

In the Paper 'Integrating heterogeneous data in the web of data', the Resource Description Framework (RDF) [2] is progressively embraced as the turn design for integrating heterogeneous data sources. A few reasons can be called attention to clarify this pattern. In the first place, RDF gives a unified data model (facilitated named graphs). Also, it grants developing interminable area information formalizations, as thesauri, vocabularies and also ontologies that can freely be reused and expanded. Lastly, RDF-based information coordination structures benefit by the thinking capacities presented by the Semantic Web Technology which are supported by expansive theoretical works. Eventually, RDF brings the possibility to utilize the colossal data base tended to by Web of Data, thus uncovering up freedoms to find related informational collections, improve data, and make added esteem by squashing up of the information.

Regardless, RDF-based information integration requirements are all things considered with legacy data which are not privately stored as RDF. The Web of Data is consistently emerging overtime as more institutions, organizations circulate their data observing the Linked Open Data guidelines [1 yet enormous measure of information still locked in storage facilities where they cannot open to the Web or to data integration structures. These information, habitually insinuated as the significant Web [4], normally

ISSN No:-2456-2165

contain legacy relational databases and records in various information designs, as they are queried by means of Web services or Web forms. That kind of data sources are not actually associated with each other and are hardly recorded by means of web search apparatuses. In this way, obtaining these legacy information sources to either perform RDF-based integration or publish Linked Data on the Web of Data anticipates that techniques should make a translation of heterogeneous data into a RDF representation. Ever since the mid 2000's, much research has inspected such systems. Somewhat, they all depend upon the unequivocal or evident description of the mapping which teaches how to decipher each data thing from its novel course of action into a RDF representation. In this regard, relational databases (RDB) have grabbed a lot of attention because of their prevailing position [11] [9] [6].

III. DESIGN METHODOLOGY

The proposed framework has a place with an update-driven methodology. The primary thought is to utilize a few components from designs of data warehouse as well as mediated frameworks. These sources require unique functionalities for information extraction. Therefore, the arrangement has a coverings layer, which happens in the mediated systems. Integrated data and meta-data are stored in a relational data set.

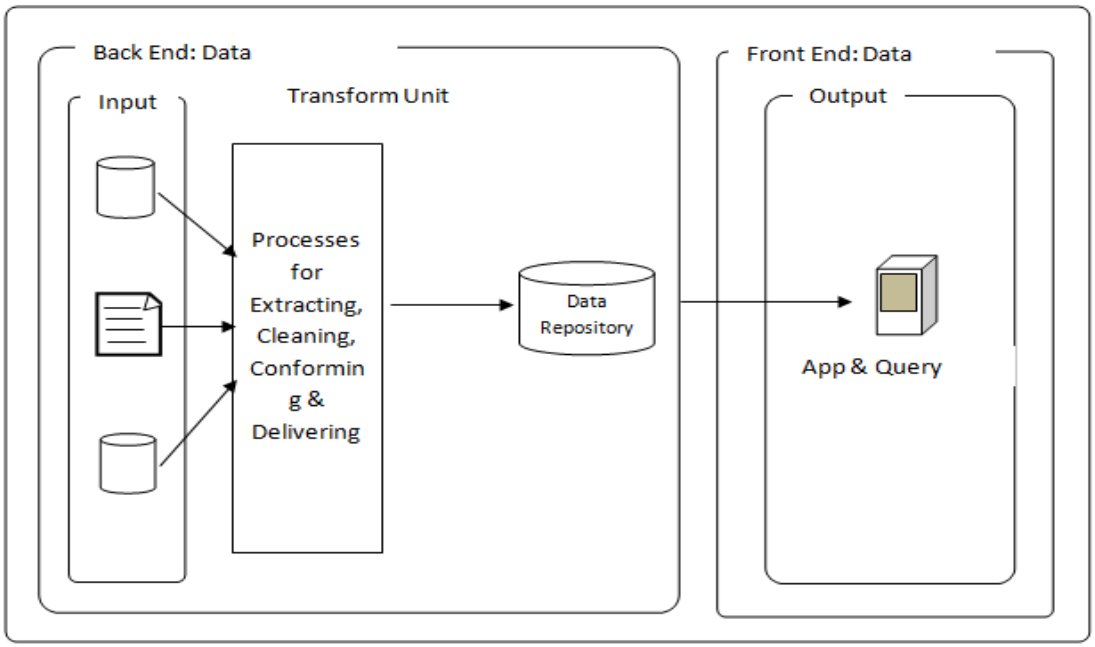

Fig 3.1 Proposed System Architecture

A data integration system is defined in terms of Schemata. A Unified view that was produced from a processed query is identified as global schema. The format of the various data sources and how they relate to one another is identified as Source Schema. The way in which the global and source schemata interrelate is known as mapping.

Input Unit (Source schema)

The System first pulls all the data from the various sources. Some of these sources are relational databases which include: Social media database, Customer database, Web database etc. Relational databases use tables to store information. The standard fields and records are represented as columns (fields) and rows (records) in a table.

Transform Unit (Mapping)

There is no universal approach to data integration; however, data integration solutions typically involve a few common elements which are:

i. A network of data sources
ii. A master server
iii. Clients accessing data from the master server

The Clients sends a request to the Master Server for data. The Master server then intakes the needed data from the sources. The Data is extracted from the sources, converts all the data into a common format so that one set of data is compatible with another. Then it loads this new data into the repository.

Output Unit (Global schema)

When you submit your query, the repository locates the data, retrieves it and presents it in an integrated view.

Functional Design

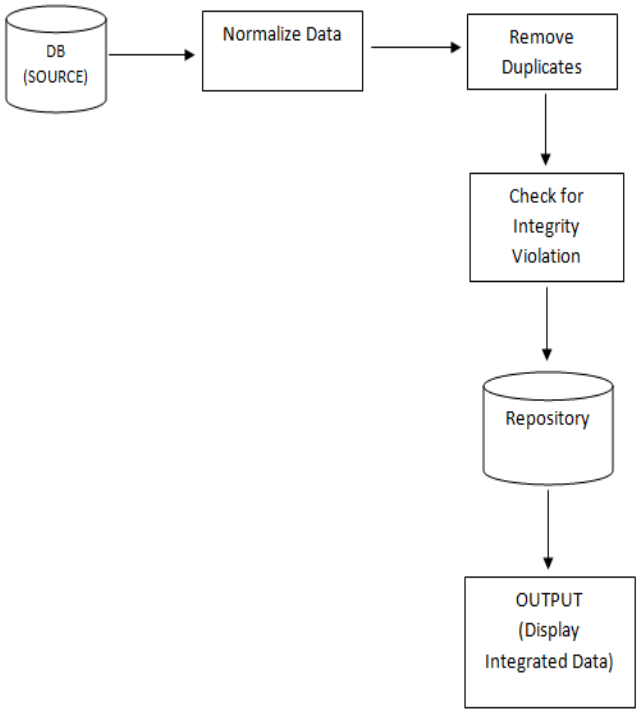

Fig 3.2 Functional Design

Extracting. The raw data coming from the source systems is first uploaded into the system for integration.
Cleaning.

"The level of data quality acceptable for the source systems is different from the quality required by the repository. Data quality processing may involve many discrete steps, including checking for valid values, ensuring consistency across values, removing duplicates, and checking whether complex business rules and procedures have been enforced." Below is an example of a record containing errors.

Table 3.1 Records with inconsistent values

First Name	Last Name	Phone	E-mail	Age	Address
John	Eze	johneze@gmail.com	080321 41884	28	

Conforming. Data conformity is required whenever at least two data sources are converged in the repository. Separate data sources can't be queried together except if we take care of the entirety of the issues of syntactic-and semantic-data heterogeneity.

Table 3.2 System conformation

Process	Requirements
Extraction Transformation	Record must not contain null values Records must contain valid values Records must contain right formats Duplicates must be eliminated Records must be consistent Integrity violation must be checked

Delivering. The general purpose of "the back room is to prepare the data for querying. The last step is to change the data structure to meet this functionality. This is regularly a set of basic symmetrical schemes known as multi-dimensional models."

IV. RESULTS AND DISCUSSION

This system of integration of big data from heterogeneous sources into a repository is realized using JAVASCRIPT and MYSQL (Database Server). The user must to press the 'integrate data' button to start the data integration process and then can start working with data using other perspectives. After the integration process is completed, the user can view the integrated data in the global database by clicking the 'view data' button. The system also gives room for a user to upload a database containing information which may have some errors. This database will then further undergo the necessary integration processes such as Normalization, Removing duplicates, Integrity violation checks.

The system contains useful buttons such as Upload, Integrate, and View. Information record such as Name, Age, E-mail, Phone number, Address, Interest, Website etc, in tabular form stored in Relational database and uploaded from the various heterogeneous sources. The System Considers 3 sources from which information are uploaded into the system which are:

i. Social media
ii. Customer source
iii. Web source

The individual's information gotten from the sources and integrated should be a complete record for viewing. There are various data samples that have been used to test the performance of the system and necessary results obtained.

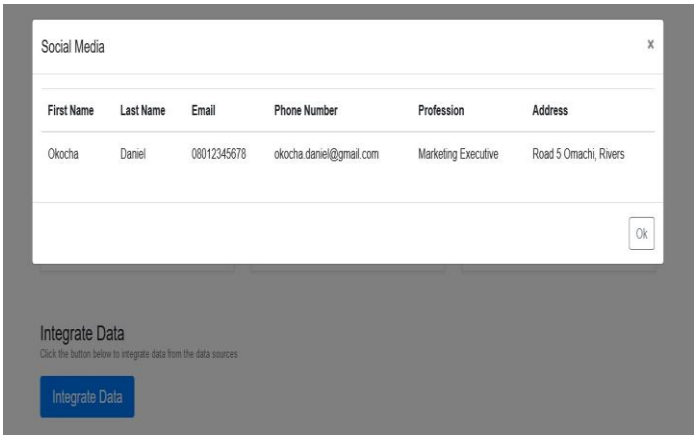

Fig. 4.1 Sample Records from Social Media

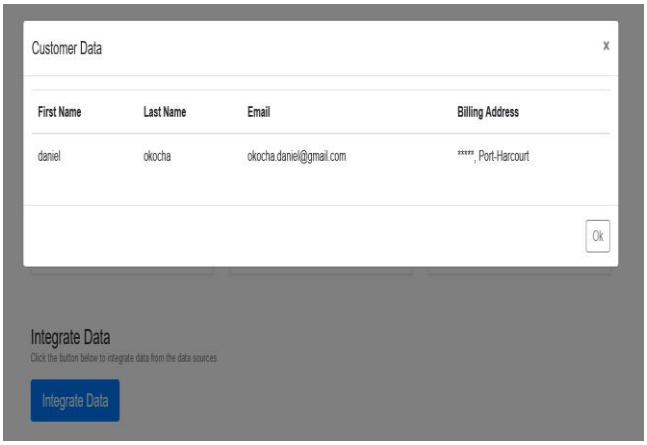

Fig. 4.2 Sample Records from Customer Data

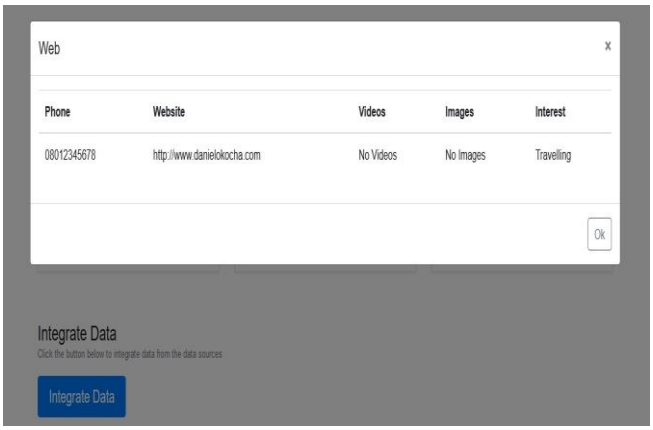

Fig. 4.3 Sample Records from Web

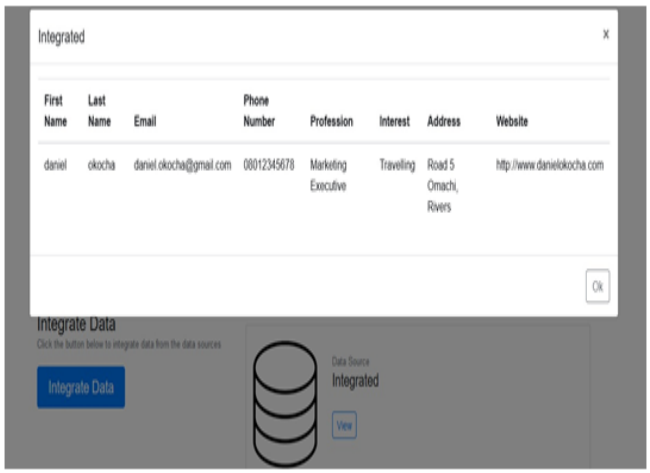

Fig. 4.4 Sample of Integrated Data

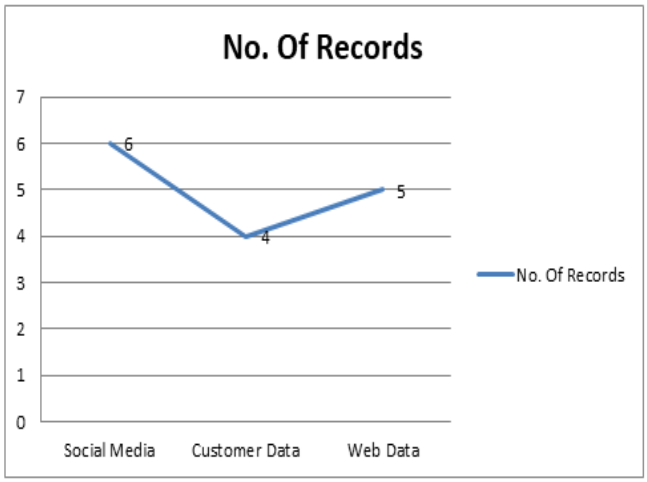

Fig. 4.5 Chart showing records as against expected record

V. CONCLUSION AND FUTURE SCOPE

This research work uses the Common Data Storage technique or the Physical Data Storage technique. Regular Data Storage is a framework, which will in general duplicate the information from the source frameworks to another framework. This way unified gathered information is stored and overseen by those new frameworks rather than the first source. All the more normally called Data Warehouse; this method helps in information collection from different sources, consolidating them to a focal space and management (Database documents, centralized servers, and level records). The immense volumes, nonetheless, requires separate information reconciliation frameworks. A graphical user interface of the prototype is a web-based application. The user must to press the 'integrate' button to start the data integration process and then can start working with data using other perspectives. Database can be uploaded into the system then after data load, the user can integrate data in the global database (with integration) by pressing the 'integrate' button. The user can also execute SQL queries on the global database.

Additionally, it would be important that information integration and information analysis are made continuously, on the grounds that it is unfathomable these days and particularly for Internet applications to settle on choices dependent on generally old information.

In every one of the frameworks surveyed, little of them consider multi-media information sources like Image or Video types. It is fascinating to propose an intervention or shared design which would include in their local information sources these arrangements joined with different configurations, with assessing the advancements brought by the semantic Web in the semantic depiction of these information. The serious issues for this situation will be the cross examination, the integration and the ordering of these information.

REFERENCES

[1]. Berners-Lee T. (2006). Linked Data, in Design Issues of the WWW.
[2]. Cyganiak R., Wood D. & Lanthaler M. (2014): RDF 1.1 Concepts and Abstract Syntax. *W3C Recommendation*
[3]. Das S., Sundara S. & Cyganiak R. (2012). R2RML: RDB to RDF Mapping Language.
[4]. He, B., Patel, M., Zhang, Z., Chang, K. C.-C. (2007). Accessing the Deep Web. *Communications of the ACM* 50(5), 94–101.
[5]. Koch, C. (2001). Data Integration against Multiple Evolving Autonomous Schemata. *CERN-THESIS-2001-036.*
[6]. Michel, F., Montagnat, J., Faron-Zucker. C. (2014). A survey of RDB to RDF translation approaches and tools.
[7]. Ray, S., Bandyopadhyay, S. and Pal, S. (2009). Combining Multisource Information Through Functional-Annotation-Based Weighting: Gene Function Prediction in Yeast. *IEEE Transactions on Biomedical Engineering*, 56(2), pp.229-236.
[8]. Ruggles, S., Hacker, J. and Sobek, M. (1995). General Design of the Integrated Public Use Microdata Series. Historical Methods: *A Journal of Quantitative and Interdisciplinary History*, 28(1), pp.33-39
[9]. Sequeda, J., Tirmizi, S. H., Corcho, Ó., Miranker, D. P. (2011). Survey of directly mapping SQL databases to the Semantic Web. *Knowledge Eng. Review* 26(4), 445–486.
[10]. Smith, J., Bernstein, P., Dayal, U., Goodman, N., Landers, T., Lin, K. and Wong, E. (1981).Multibase. *Proceedings of the May 4-7, 1981, national computer conference on* - AFIPS '81.
[11]. Spanos, D. E., Stavrou, P. Mitrou, N. (2012). Bringing Relational Databases into the Semantic Web: A survey. Semantic Web Journal 3(2), 169–209.