Data Integration Process of Business Intelligence – A Review

Efozia N.F., Anigbogu S.O., Asogwa D.C. and Anigbogu G.N.
Prototype Engineering Development Institute (PEDI), National Agency For Science and Engineering Infrastructure (NASENI), Federal Ministry of Science and Technology (FMST), Ilesa, Osun State, Nigeria
Department of Computer Science, Faculty of Physical Sciences, Nnamdi Azikiwe University, Awka, Nigeria
Computer Science Department Nwafor Orizu College of Education, Nsugbe, Nigeria
fenngo31@yahoo.com, so.anigbogu@unizik.edu.ng, dormatfk@yahoo.com, anigbogugloria@yahoo.com

Abstract— Integration of the huge data from various heterogeneous structured, semi-structured and unstructured sources in a Business Intelligence (BI) system is quite complex and tasking especially when done manually. Enhanced and automated design of data integration process in Business Intelligence (BI) is therefore necessary. The paper reviewed some data integration techniques for BI process with the aim of determining which would be better in the implementation for the design of BI system. The objective is to ensure reliable, scalable, real-time and efficiency in the delivery of a more accurate and actionable evidence-based decision making for management usage. And from the reviewed literature, the hybrid of any two or more data integration techniques such as ontology-based data integration (OBDI) and virtual data integration (VDI), proved to be of more benefit as it resolved most of the weakness in either of the data integration techniques. The comparison of the data integration techniques and their application (it was deduced) showed that a hybrid model yielded an enhanced data integration process for Business Intelligence system.

Index Terms—Data Integration, Data warehouse, OBDI, OLAP, Wrappers, VDI

1 INTRODUCTION

Business Intelligence (BI) is considered as the set of strategies, processes, applications, data, products, technologies and technical architectures which are used to support the collection, analysis, presentation and dissemination of business information [1].

Today a large amount of data presently termed big data is generated every day, thus making data integration (DI), storage and processing extremely challenging, for organization to analyze and utilize for competitive advantage.

Business Intelligence (BI) as a concept and technology has significant potential in transferring data from distributed and heterogeneous sources into an integrated enterprise view for supporting organizational decision making, management and strategic planning. As a knowledge management tool, it has direct impact on business performance of an enterprise [2]. The main purpose of Business Intelligence (BI) is to enable interactive and easy access to diverse data, enable manipulation, transformation and transportation of these data, and provide business managers and analysts the ability to conduct appropriate analyses and perform actions.

The process takes cyclical nature and includes stages of information needs, definition, information collection, information processing, analysis, information dissemination, information utilization and feedback. The cycle of the received feedback helps to re-evaluate or re-define information needs. In Business Intelligence (BI) process, there is usually no clear concentration on a specific topic or problem. The resources of a BI system are used for constant monitoring of internal and external business environment.

In order words, the systems serve common information needs of keeping users informed about the state of business environment, often combing a monitoring function with alerts, exception reports and other tools to draw attention to changes or inconsistencies. Therefore, an important feature of Business Intelligence (BI) system is their ability to produce a complete composite view that would assist in avoiding surprises [3]. Fig. 1 shows the generic Business Intelligence process model.

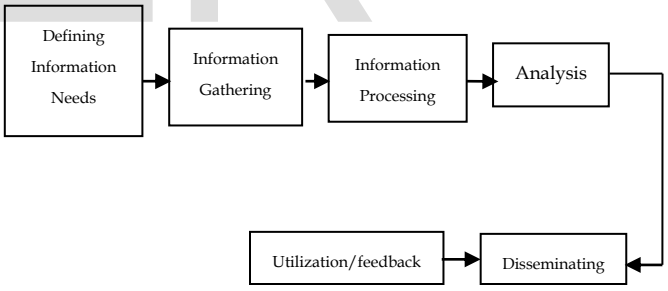

Fig. 1: A Generic Business Intelligence Process Model (Source: [3]**)**

The data integration (DI) and transformation layer include processes for transforming data from operational and external sources into a form suitable for database storage. It is termed Extract Transform Load (ETL) or Extract Load Transform (ELT) processes. This is because data integration (DI) is a vital process feature in all Business Intelligence (BI) system. Ladjel, et al (2003) had noted that a Business Intelligence (BI) system is a technique that can consistently bring about conflict resolution, intelligence, adaptive and automatic data integration (DI) model it can also increase speed of processing and reduction in cost of business. Data integration process as a heterogeneous data from different sources (and is user-friendly) would be of high demand by business decision makers as it would help them to be ahead of their competitors, and plan their budget

in the right direction.

The concept of data integration (DI) is becoming more relevant while the magnitude of available data sources increases steadily while data creation is occurring at a record rate. Files of data are useless to businesses if there is no good way to retrieve, observe, and interact with that data, in order to generate tangible information and practical knowledge. A strong, vital, and reliable data integration strategy can help companies to harness information coming in from every direction and use it to support their business goals. In fact, Mathijs (2016) had observed that the advent of micro-service architecture, has made data integration (DI) gain a renewed wave of attention.

Traditional data warehouse (DW) forms yet another data silo in organization, which leads to slow degradation of the data warehouse (DW) benefits of providing optimized integrated data delivery. But with the integration of data becoming important in many areas, it would provide decision makers and analysts' uniform access to 'distributed and heterogeneous' sources of data such as Relational Databases, XML Documents, and Semantics Databases [4].

Furthermore, data integrators need to have a global overview of all the resources spread over the source system, as experts need to be involved and consulted every time a change is introduced in a source system. All these create bottleneck in the development of the systems. Therefore, new ways of integrating data sources are required, since traditional ways of integrating simply required too much modeling, maintenance and coordination among the owners of data sources [5].

Furthermore, Business Intelligence systems should closely correspond to business objectives of enterprises. Therefore, the most important motives that support implementation of Business Intelligence systems in enterprises may include the following as noted by Celina and Ewa [6] to include:

1. Transitioning from instinct and intuition decision making to objectivism that is based on the analysis of facts, indexes, balanced score cards, managerial cockpits, and so on.
2. Forecasting enterprise development along with customers' and suppliers' behaviour.
3. Matching operational activities with realisation of strategic objectives (measuring development in the realisation of strategies, monitoring of business process effectiveness, matching budgets and investments with corporate strategies);
4. Implementing standards that are used as the basis for repetitive, regular and cyclical business processes within organisations.
5. Unifying informational transfers in order to make them more transparent and unifying roles of individuals who participate in decision making processes.
6. Rapid detecting of information that deviates from commonly accepted standards and procedures and that suggests some possibilities that new threats will emerge (dishonest customers, inflated material or energy usage, etc.)
7. Reducing time that is necessary to analyse information, and decreasing a number of participants who are involved in analysing and processing of information.
8. Automatic and rapid reporting and preparing of plans and forecasts.

Hence, data integration is a fundamental, yet deceptively challenging, component of any organization's Business Intelligence (BI) and data warehouse (DW) strategy. Existing data integration methodologies do not assure quality of data thereby making it difficult to access. The paper tried to review some data integration (DI) approach in resolving the issue of quality of data and its accessibility in Business Intelligence (BI).

2 METHODOLOGY

Olga and John [7] had opined that every term must be identified by its name, definition, its use, the time-frame and domain of validity and various rules associated with it. Here under, we have the main modalities of data integration as well as its conceptual model as depicted in fig. 2.

1. Data means 'what is given' and refers to what we can perceive experience or register with our senses or devices.
2. Concept represents an abstract idea generalized from particular instances.
3. Model is a set of concepts with defined relationships.
4. Information is created when data are interpreted based on a set of concepts.
5. Informational value of data is defined by the ability of data to provide us with useful and reliable information for our decisions and actions.
6. Data source is a database, a website, a publication, or any other collection of data, which is constructed upon a set of concepts and models.
7. Data element (DE) is an atomic unit of data collection that is unambiguously defined in the controlled vocabulary of a project. A collection of Data Elements defines a structure of a data set.
8. Authoritative source contains the most reliable values for a specific Data Element.
9. Data set is a collection of data that is produced from the data source at a moment in time based on a defined structure and a set of rules.
10. Data value is a specific value of a Data Element. ''A body temperature of 101°F'' employs Data Element ''body temperature'' measured in Fahrenheit and its value ''101''.
11. Data pool is a collection of all data sets gathered for co-processing.
12. Data transformation is any action performed on data which can be simple selecting all Data Elements (SELECT *) or any combination of sophisticated statistical and analytical operations.
13. Project is a temporary endeavour undertaken to create a unique product, service, or result.

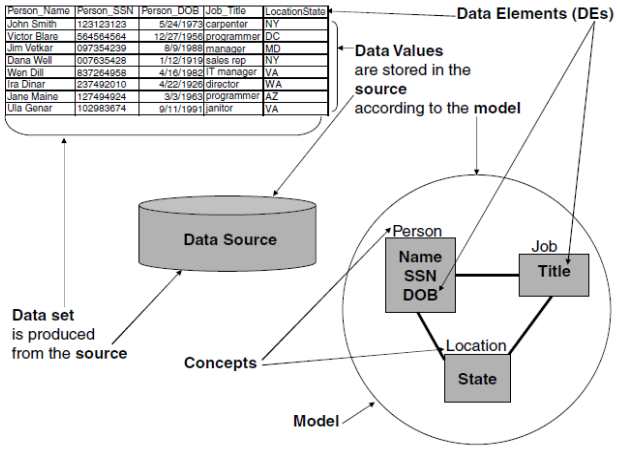

Fig. 2: Illustration of Main Modalities of Data Integration (Source: [7]**)**

Integration is pursued in one of two ways; the first way is when an exact question is known, and an answer and data to the answer is available. The aim here is at finding reliable sources and pulling the needed fields into a database designed for the purpose. The second way is aimed at understanding the power of data at large when a number of data sources of varying reliability are available. In both approaches, analyzing of the same data with various methods and techniques and then positive results should be presented according to requirements of various user groups.

The layers at which integration occurs are Data Sources, Data Elements (DEs), Data Sets and Data values. The integration that occurs in these layers include; integration of concepts, models, controlled vocabularies, and methods of data acquisition, frequencies of updates, as well as units and formats of records. The most crucial part of the integration is choosing the integration keys. The models of data integration include federated databases, data warehouse, mediation, materialized, virtual, peer to peer and ontology-based data integration among others.

In the virtual data integration approach, the sources contain the real data and one or several virtual view(s) contain reconciled integrated schemas over these sources. The autonomous and heterogeneous data sources are queried by these homogeneous views. For this purpose, the system needs mappings that describe the relationships and semantic dependencies between integrated and source schemas. This method is characterised by approaches such as mediation, peer-to-peer, ontology-based, etc.

Ontology-based data integration has been frequently used in the Semantic Web. It is a form of virtual approach data integration. It is an approach used to obtain an automated access to the information contained in various sources. The information items are described by means of metadata provided by ontology. While there are many definitions for ontology in different categories, in the context of data integration, it is considered as a formal description of a domain of interest intended for sharing among different applications. In other words, ontology describes the semantics of the data in order to provide a uniform way to make different parts understand each other. One of the first projects in which a mediation system has been developed using ontology is InfoSleuth. More so, an ontology-based data integration system is a system that provides a conceptual view on top of pre-existing information sources. Here, ontology is used for the identification and association of semantically corresponding information concepts of data sources [8].

4 SYSTEM DESIGN AND IMPLEMENTATION

The goal of a data integration system is to develop a homogeneous interface for end users to query several heterogeneous and autonomous sources. Building such a homogeneous interface raises many challenges among which are the heterogeneity of data sources, the fragmentation of data, the processing and optimization of queries which appear to be the most important. The heterogeneity of the sources can be present at different levels such as hardware system, database management system (DBMS), data language or query language. It can also be present at a semantic level. The schemas of sources can be heterogeneous.

Data integration is a fundamental, yet deceptively challenging, component of any organization's business intelligence and data warehousing strategy. Data integration involves combining data residing in different data repositories and providing business users with a unified view of this data. In addition, companies face a challenge of ensuring that data being reported is current and up-to-date. The following are the components of any data integration system as devised by Helena and Paulo [9] to enable us fully understand how the system works; Data sources, Mediated schema, Source description, Semantic mappings, Wrappers, Reformation, Plan generator, and Execution engine.

There are two basic design approaches to data integration challenges, termed Procedural and Declarative.

- In the procedural approach, data are integrated in an ad-hoc manner with respect to a set of pre-defined information needs. Here the basic issue is to design suitable software modules that access the sources in order to fulfill the pre-defined information requirements. Most data integration project (virtual and materialized), such as TSIMMIS (The Stanford-IBM Manager of Multiple Information Sources), SQUIREL, and WHIPs follow this idea. They do not require an explicit notion of integrated data schema, but rely on two kinds of software components; wrappers that encapsulate sources, converting the underlying data objects to a common data model, and mediators which obtain information from one or more wrappers or other mediators, refines the information by integrating and resolving conflicts among the pieces of information from the different sources and provide the resulting information either to the user or to other mediators. The main idea here is to have one mediator for every query pattern required by the user. And generally, there is no constraint on the consistency of the results of different mediators.
- The declarative approach provides a crucial advantage over the procedural one, despite the fact that

building a unified representation may be costly, it does allow for maintaining of consistent global view of the information sources, which represents a reusable component of the information integration systems [10].

The following technologies and/or techniques can be applied in resolving the main data integration (DI) approach challenges (Procedural and Declarative);

a. Resource Description Framework (RDF): This is used for integrated schema description as well as providing a unified view of data. It has a well-defined syntax and data type and it has reasonable processing complexity.
b. Description Logic (DL): This is used to find any contradiction in the integrated schema (satisfactory of concepts in DL terms). It has well-defined semantics and decidable routines for basic services satisfactory, which makes it suitable for knowledge representation and reasoning in this domain. An example is the web ontology language (OWL).
c. Resource Description Query Language (RDQL): This is used for reformation of queries. It is a query language for RDF and it provides a data-oriented query model.

It is noted that every data integration system are built with the aim of building an integrated view of the data defined in various sources and develop a mechanism for data extraction from it.

Hence, design methodology for an integration process consists of four steps;

- Pre-integration: This has to do with analyzing schemas before integration to determine the integration technique, order of integration and to collect additional information.
- Schema Comparison: Here concepts are compared; conflicts and schemas properties are searched for.
- Conforming Schemas: This Solves schema conflicts.
- Merging and Restructuring: This merges and restructures the schemas so that they conform to certain criteria.

Again, application of data integration design layers in Business Intelligence (BI) implies considering whether data needs to be physically moved or whether a virtual or "in-place" approach to accessing and aggregating data makes more sense. Primary DI layers include:

1. Data Integration and storage layer
2. Data Analysis Layer
3. A reporting Layer
4. Administration Layer

Thomas Jorg [11] had opined that integration systems are aimed at resolving heterogeneity and data quality problems thereby providing a single-system image to users. The fundamental data integration (DI) approaches are Virtual and Materialized (Data warehouse) integration.

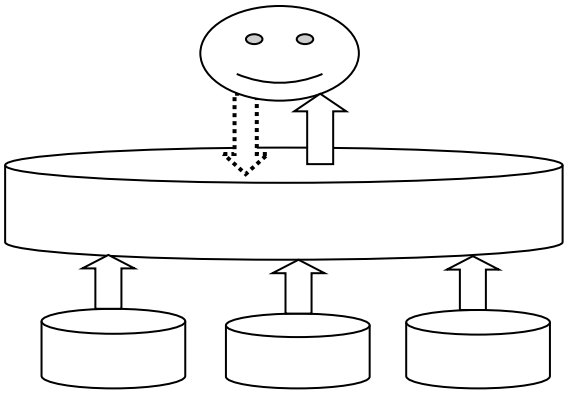

Fig. 3a: Materialized Data Integration (Source: [11])

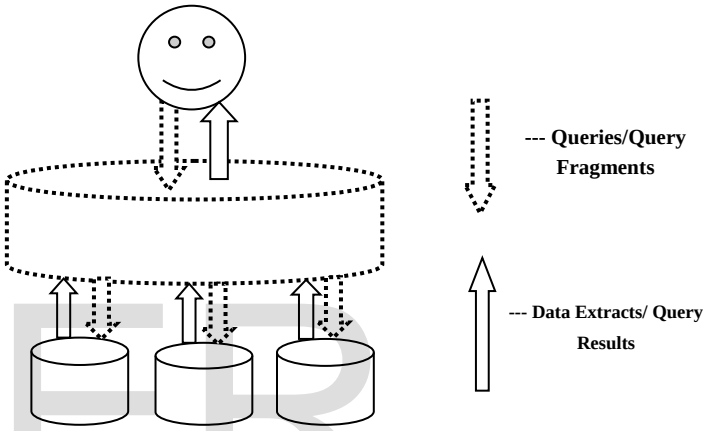

Fig. 3b: Virtual Data Integration (Source: [11])

The main difference between materialized and virtual data integration (MDI and VDI) as shown in fig. 3a and b, has to do with the place where data to be integrated is stored. And for materialized data integration (MDI), source data (data warehouse–DW) is copied to a central repository. But for the virtual data integration (VDI), all data remains in the source systems which provides for real-time as up-to-date data is always retrieved for querying purpose. Figure 3.0a and 3.0b is a high level abstraction diagram. It indicates that the solid lines and arrows depict the flow of data, while dashed lines and arrows depict the flow of queries.

In the materialized data integration (MDI), data is continuously moved from the sources to the integration system. And data movement is performed either periodically in an asynchronous manner or triggered by updates at the source systems. More so, query evaluation is done locally at the integration system and does not require any interaction with the sources.

In contrast, virtual data integration (VDI) systems do not store data centrally but provide an integrated view of source data. In the evaluation of user's query, the integration system identifies query fragments that involve the particular source. These query fragments are sent to the respective source systems and processed locally. Next, the fragment query results are sent back to the integration system where the overall query result is then assembled. With this explanation, it therefore implies

that materialized data integration (MDI) is continuously performed in the background, while virtual data integration (VDI) is performed on the fly at the time of query evaluation. Furthermore, the typical architecture of an integration system is described in terms of two types of modules; wrappers and mediators. Wrapper's aim is to access a source, extract the relevant data, and present such data in a specified format. While mediator's aim is to collect wrappers (or mediators) so as to meet specific information (data) need of the integration system. The core problem in the design of an integration system as observed by Diego, et al, [12] is the specification and realization of mediators. Figure 3.0c shows the architecture of data integration (DI).

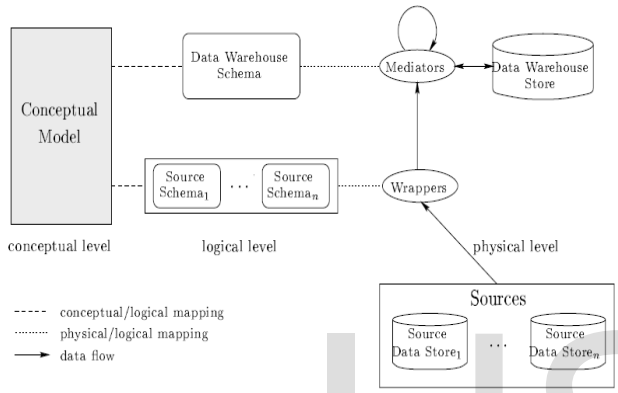

Fig. 3c: Architecture for Data Integration (Source: [12])

4 DISCUSSION

Business Intelligence allows an organization to gain a better understanding of their clients, the market, supply, as well as competitors in order to make smart business decisions. The best integration solution is the combination of virtual and physical approaches.

Data integration (DI) is an essential process for Business Intelligence (BI) application. Traditional BI approach physically moves data from original data sources to specialized target data stores after going through data cleansing, transformation and de-normalization. Ladjel, et al, [4], had noted that the process is termed Extract Transform Load (ETL) or Extract Load Transform (ELT). In other words, it is either the data is transformed before loading to the data warehouse (DW) or is loaded before transformation. The following are the roles of data integration (DI) in Business Intelligence (BI) as opined by Nitika, et al, [13];

- It provides a consistent single version of the truth coming from multiple heterogeneous sources of data.
- It resolves the issue of data related problem and extra cost to reconciliation of data.
- It is needed to feed BI application, as its efforts are worth the efforts by meeting the needs of business people in time and at low cost.
- It is the centralization of authority and governance over the integrated systems.
- Data Integrators need to have a global overview of all the resources spread over the source systems.
- Experts need to be involved and consulted every time a change is introduced in source system. This creates bottlenecks in the development of these systems.
- The traditional ways of integration involves too much modeling, maintenance and coordination among the owners of the data sources.

Again, owners of data sources need to be able to collaborate without any central authority or global standardization. This implies challenges like scheme matching will need to be tackled in a much more parallelizable and scalable way, since the ability to scale up to more sources requires redesign architectures to exploit the power of large clusters. Therefore, the paper is of the view that it would be better to combine or hybridize the virtual and/or materialized approach to data integration in a Business Intelligence data integration process in order to annex the features in both approaches for an enhanced Business Intelligence process which would in turn give rise to achieving better decision making for management in enterprises of any domain industry.

5 CONCLUSION

In concluding this review, we recall that Business Intelligence (BI) is aimed at enabling business users to easily access and analyzes relevant enterprise information so that they can make timely and fact-based decisions. The major issue today is that crucial information is scattered throughout the separately developed data sources, in a way that makes the **big picture** is difficult to obtain. Data integration process in Business Intelligence system as noted by Munmun and Nashreen [14] presents a unified virtual view of all these scattered data within a domain, allowing users to pose queries across the complete integrated schema as if they are interacting with a single data source. The integration of different sources of information is a crucial necessity today. Hence, Enterprises, Ecommerce, Bioinformatics, the Semantic Web and many other sectors in our business environments cannot continue to exist without integrating their information. Shokoh [8] had observed that different challenges exist and several technologies have been proposed to integrate and manage heterogeneous sources in a homogeneous way. Therefore, this paper has been able to review the data integration process of Business Intelligence and proposed that a hybrid model would be a better approach in order to be able to harness the benefits of the main approaches to Data Integration (Materialized and Virtual mediation) as well as ontology-based and virtualization, which can assist management of business organizations and the experts, in making intelligent, reliable, scalable and efficient decisions in the domain sector it is applied.

REFERENCES

[1] M. Aneesha and K. C. Balajkarthik "Qualitative Analysis of BI in the Budgeting Process" *MSc degree student, (2017)* at Halmsted University, Sweden.

[2] P. Kavassalts "Computer Applicants in the Modern Enterprise" *www.atlantis-group.gr, (2015)*

[3] R. Skyrius, G. Kazakeviciene, and V. Bujausas "From Management Information System to Business Intelligence: The Development of Management Information Needs" *International Journal of Artificial In-*

telligence and Interactive Multimedia, Vol 2, No 3.DOI: 10:9781/ ijimai 2013.234

[4] L. Bellatreche, G. Pierra, N. X. Dung and D. Hondjack "An Automated Information Integration Technique using Ontology – based DB Approach" *Roc of concorrat Engineering (CE 2003), Special Track DI in Engineering, Madeira*, Portugal 26-30/07/2003, 217-224.

[5] M. G. Dabrock "A Distribution Semi-Automatic Approach'' *Master of Business Informatics Thesis, (2016)* Department of Information and Computing Sciences, Utrecht University.

[6] M. O. Celina and E. Ziemba "Approach to Building and Implementing BI Systems" Published in *Interdisciplinary Journal of Information, Knowledge and Management Vol.2 (2007)* Retrieved from www.elsevier.com/...pdf on Dec. 5, 2012

[7] O. Brazhnik and J. F. Jones "Anatomy of Data Integration" *Journal of Biomedical Informatics* Vol.40 pg 252-269, (2007). Retrieved from www.elsevier.com/locate/yjbin on Dec. 3, 2012

[8] S. Kermanshahani "IXIA (IndeX-based Integration Approach) A Hybrid Approach to Data Integration" *Networking and Internet Architecture [cs.NI].* Universite Joseph-Fourier - Grenoble I, 2009. https://tel.archives-ouvertes.fr/tel-00407575

[9] H. Galihardas and P. Carreira "Virtual Data Integration" *https://fenix.tecnico.ulisboa.pt/downloadFile/.../VirtualDataIntegration-Parte-2.pdf, (2015)* retrieved on May, 2016

[10] A. Amineh, H. Saboohi, N. Nematbakhsh "A RDF-based Data Integration Framework" *NEEC 2008* www.1211.6273.pdf/ retrieved on May 23, 2016

[11] T. Jorg, "Incremental Recomputations in Materialized Data Integration", *Doktor der Ingenieurwissenschaften (Dr.-Ing.)* genehmigte Dissertation von, vom Fachbereich Informatik der Technischen Universit¨at Kaiserslautern zur Verleihung des akademischen Grades, 18-01-2013 D 386. www.dissertation_thomas_joerg.pdf. Retrieved August 13, 2017.

[12] C. Diego, G. Giuseppe De, L. Maurizio, N. Daniele and R. Riccardo "Data Integration in Data warehousing" *International Journal of cooperative information system Vol.10, No.3 Pg.237-271, (2001)* Retrieved from www.calv.etal-IJCIS-2001.pdf/ on Dec.6, 2012

[13] N. Arora, U. Devi, Preeti and Pratibha "Data Virtualization Technology" *International Journal of Innovations and Advancement in Computer Science IJIACS* ISSN 2347-8616 volume 6, issue 4 April 2017

[14] M. Bhattacharya and N. Nesa "Study on Theoretical Aspects of Virtual Data Integration and its Applications" *Munmun Bhattacharya Int. Journal of Engineering Research and Applications* ISSN: 2248-9622, Vol. 6, Issue 2, (Part - 1) February 2016, pp.69-74 www.ijera.com, retrieved on May 14, 2016