A Review on Big Data Integration

B. Arputhamary
Assistant Professor,
Department of Computer Applications,
Bishop Heber College,
Trichy, Tamil Nadu, India.

L. Arockiam, Ph.D.
Associate Professor
Department of Computer Science,
St.Joseph's College,
Trichy, Tamil Nadu, India.

,

ABSTRACT
Big Data technologies are becoming a current talk and a new "buzz-word" both in science and in industry. Today data have grown from terabytes to petabytes and now it is in zeta bytes. Increased amount of information increases the challenges in managing and manipulating data. Data integration is a main issue in large data sets which is managed by Extract, Transform and Load (ETL) tools such as Data Warehouses. Data Warehouse is the process of transforming all multiple data formats into a single format and consolidating them in one place. Now days, data generated from social networks, web server logs, sensors used to gather climate information, stock market data, e-mails, transaction records, web click streams, etc. Most of these data are in unstructured or semi structured forms. Today organizations' are trying to find new solutions such as ETLs to manage the situation. The existing data warehousing tools and techniques were inefficient to handle unstructured and semi structured data. This paper presents the issues and challenges of data integration in Big Data environment and techniques for big data integration. A new ETL framework is proposed open problems for future research of data integration are identified in big data environment.

General Terms
Big Data Integration, Data Integration

Keywords
Big Data, Integration, Data Warehouse, Hadoop.

1. INTRODUCTION
The term 'Big Data' initially named from the big or large volume of data. According to Gartner[18], Big data is defined as high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making. There is no proper definition for Big Data. It was defined as a situation where the volume, velocity and variety of data exceed an organization's storage or compute capacity for accurate and timely decision making. Storage of these big data can be done by introducing multiple data centers, where as utilizing these data in an effective manner is a tedious one which is to be considered carefully. Today, organizations invest more in data manipulation and most of the time the stored data are unused and they are not retrieved and utilized in a proper way. Enterprises and organizations are spending more in data processing. Actually big data is not a problem whereas it is a big asset to the organizations. Data are from different sources and integrating such data is very important and available data warehousing tools are used for doing integration. But in big data environment data from different sources are of different formats and existing data warehousing techniques in data mining are inefficient to handle such situation. This paper aims at presenting the issues related to data integration in big data environment and techniques available for Big Data Integration.

2. RELATED WORKS
The paper [1] converts the XML based schemas to ROLAP (relational online Analytical Processing) to multiple data warehouse schemas such as star schema, snowflake schema and fact constellation. A framework is proposed to detect more than one data warehouse schemas from the given XML schema. A new data structure is defined to construct schema graph. In this paper, some steps could be added to transform the semi structured data to XML. The main disadvantage of this model is, if the source XML changes, then the schema structure needs to be reconstructed.

In paper [2] ,Rabah Alshboul performed explorative study on Data warehouses are made in different organizations namely: Hallmark and American Airlines and discovered critical success factors (CSF). Some of the crucial factors in implementing data warehouses were identified. This paper presented the five major critical successful factors for the successful implementation of Data Warehouse Architecture.

An Oracle White paper [3] has introduced oracle Data Integrator 12C (ODI12C) to bridge the Enterprise and Big Data World. To provide efficient connectivity between Oracle Database and Hadoop, Oracle Data Integrator is introduced. In this paper, the issues related to data integration in Big Data world such data quality and data correlation were highlighted. The complete solutions for big data integrations are suggested as, combining traditional technologies with Hadoop, Integrated Design Tools, Integrated platforms and real time analytics.

In paper [4], Rajni Jindal et al. performed comparative study on various Data Warehouses based on its design characteristics and proposed a new object oriented framework for multidimensional data warehouses. And the framework has two major parts namely requirement level and design level. An integrator component of this framework gets information from different sources. In designer level, UML designer helps to construct multidimensional model represented in the form or star and snowflake schema. This framework satisfies all the requirements of data warehouse design and it is achieved by using UML and star, snowflake schema's also considered in this model.

A tutorial paper [5], Xin Luna Dong et al. outlined the progress on schema mapping, record linkage and data fusion by data integration community. It addressed the challenges of Big Data Integration and identified a range of open problems for future research community. This paper also presents how Big Data Integration differs from Big Data Integration in 5 V's.

Andreas Schultz et al. [6] provide a framework to integrate and cleanse Linked data from web. Mainly focus on web data access, data translation, identity solution, provenance tracking, data quality and assessment. In LAID framework, data integration is performed at five levels. And this framework has been implemented in three runtime environments namely single machine/ in-memory, RDF store version and Hadoop version. The performance of LDIF is evaluated in all three environments. It provides better result in Amazon EC2.

In paper [11], a knowledge as a Service (KaaS) framework is proposed for cloud data management. The main objectives of this paper are, (i) Storing large amounts of disaster related data from diverse sources, (ii) Facilitating search, (iii) Supporting interoperability and integration. The authors have proposed a Disaster-CDM Architecture which Consists of two parts, Acquiring Knowledge and Knowledge Delivery. Disaster CDM is an ongoing work. Simulation of data acquisition process is completed. Remaining part of the framework will need to be implemented. Critical factors to be considered in knowledge acquisition are identified as

(i)Integration of heterogeneous technologies, (ii) The Criteria for optimal storage and Semantic Integration of diverse data sources identified as a research challenge in knowledge delivery.

The authors [12], proposed Big Graph for enterprises to manage Big Data and facilitate data integration. Big Graph is a tool which facilitates enterprises to create and link their data. Layers of Big Graph are identified as,

Semantic Layer: Interact with users and submits user queries to the RDF processor.

RDF Layer: Interface with client applications. Responsible for add/remove triples and converting text based query into a set of operations.

Triple Adapter Layer: TAL is responsible for converting triples into a key that can be stored in KVS layer.

KVS Layer: Implements a distributed ordered key value store and responsible for distribution and retrieval of the data from other servers.

Communication Layer: Communicate with other servers.

The **Data Aggregation Layer** currently on development which Enables to access, extract, and aggregate together different data from different source and to convert those data to RDF triples that can then stored. (RDF: Resource Description Framework). Big Graph is an on-going project. Future works: Development of data reconciliation components to reconcile the heterogeneity among various RDF resources.

The authors[13], proposed cloud based warehouse for masFlight airlines. The authors [14], has proposed a model Adaptive Data Service Co-ordinator (ADSC) model. The model aims at improving data service in cloud. It is a supplementary model and part of DaaS. The model [14] has implemented on the framework of Hadoop MapReduce. It has five functional modules. ADSC is used to improve efficiency for big data access with content sensitive capability and adaptive query optimization. ADSC is planned to benefit enterprise cloud application with more efficient big data and big table operation.

In paper [15], an efficient multi dimensional fusion algorithm was proposed. The algorithm has two phases. (i) Partition of big data with higher dimensions into certain number of relatively smaller data subsets. (ii) The discernible matrixes of all data subsets are computed to obtain their core attribute sets. Methodologies used are (i) attribute reduction and (ii) rule extraction. The author has proposed an efficient fusion algorithm based on rough set theory. This algorithm performs better than the original algorithm in a simulated distributed computing environment.

The authors [16] Xiongpai QIN, et al. performed comparative study on RDBMS and MapReduce. A Unified System for Big Data Analytics is proposed by fusing RDBMS with MapReduce and some critical issues are still unresolved. In this paper, two works with experiment results were presented. A cost model is used to route user queries to different data layouts.

The authors [17] have developed DIVE (Data Intensive Visualization Engine), which makes big-data VA approaches accessible to scientific researchers. DIVE employs an interactive data pipe- line that's extensible and adaptable. It encourages multiprocessor, parallelized operations and high-throughput, and structured data streaming. DIVE can act as an object-oriented data- base by joining multiple disparate data sources. And, although the authors have presented bioinformatics applications here, DIVE can handle data from many domains.

In this work [19], authors made a study on the big data solution for predicting the 30-day risk of readmission for the CHF patients. The proposed solution leverages big data infrastructure for both information extraction and predictive modeling. The study on the effectiveness of the proposed solution with a comprehensive set of experiment, considered quality and scalability. As ongoing work, the authors [19] aim at leveraging big data infrastructure for our designed risk calculation tool, for designing more sophisticated predictive modeling and feature extraction techniques, and extending our proposed solutions to predict other clinical risks.

In paper [20], The Ophidia analytics framework, a core part of the Ophidia research project, has been presented. The analytics framework is responsible for atomically processing, transforming and manipulating array- based data, by providing a common way to run on large clusters analytics tasks applied to big datasets. The paper[20] has highlighted the design principles, the algorithm and the most relevant implementation aspects associated to the Ophidia analytics framework. In future work the authors [20] planned to develop an extended set of parallel operators to support new scientific use cases. Array- based primitive's extensions, a data analytics query language and an optimized query planner will be considered to support more complex operators and dataflow driven requests. A comprehensive analytics benchmark will be also defined and implemented to further evaluate the performance of the system.

3. BIG DATA ANALYTICS

Big Data analytics means identifying hidden patterns and business information from large amounts of data, to make the business more agile. This data comes from everywhere: social network activity, web server logs, sensors used to gather climate information, stock market data, e-mails, transaction records, web click streams, etc[18].

The three V's of Big Data are - Volume, Velocity, and Variety[18]. When most people hear the term Big Data, they assume it to be a massive transactional data set. However, volume is only the first dimension of Big Data, and is potentially the least important among all three dimensions [10][26].

3.1. Volume

Volume derives the amount of data from terabytes to Peta bytes. Volume is the most important and distinctive feature of Big Data which impose additional and specific requirements to all traditional technologies and tools currently used. Big Data volume includes such features as size, scale, amount, dimension for tera and exascale data recording either data rich processes, or collected from many transactions and stored in individual files or databases – all needs to be accessible, searchable, processed and manageable.

3.2. Velocity

Big Data are often generated at high speed, including also data generated by arrays of sensors or multiple events, and need to be processed in real-time, near real-time or in batch, or as streams (like in case of visualization).

3.3. Variety

Variety deals with the complexity of big data and information and semantic models behind these data. This is resulted in data collected as structured, unstructured, semi-structured, and a mixed data. Data variety imposes new requirements to data storage and database design which should dynamic adaptation to the data format, in particular scaling up and down.

3.4. Veracity

Veracity dimension of Big Data includes two aspects: data consistency (or certainty) what can be defined by their statistical reliability; and data trustworthiness that is defined by a number of factors including data origin, collection and processing methods, including trusted infrastructure and facility. Big Data veracity ensures that the data used are trusted, authentic and protected from unauthorized access and modification. The data must be secured during the whole their lifecycle from collection from trusted sources to processing on trusted compute facilities and storage on protected and trusted storage facilities.

3.5. Value

Value is an important feature of the data which is defined by the added-value that the collected data can bring to the intended process, activity or predictive analysis/hypothesis. Data value will depend on the events or processes they represent such as stochastic, probabilistic, regular or random. Depending on this the requirements may be imposed to collect all data, store for longer period[22][23].

4. BIG DATA CHALLENGES

Today organizations are struggling to manage and manipulate these unstructured and structured data every day[9].

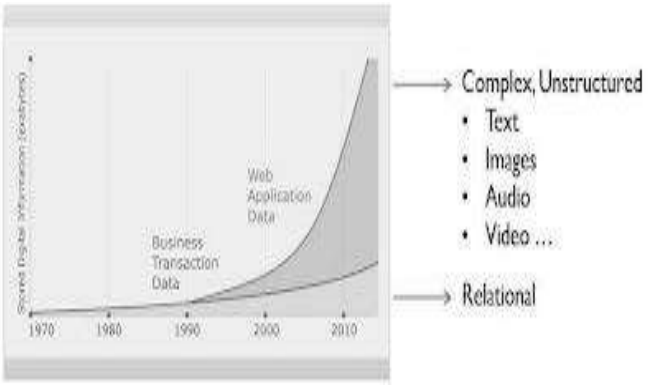

Figure 1. Graph for Unstructured data

The primary obstacles in maintaining these data are

- Difficult to handle the increased amount of data (Sizes range from terabytes to peta bytes now it is in zeta bytes)
- Storing and Managing the Big Data
- Integrating variety of data and to extract

  Knowledge from large scale and high dimensional data sets.
- Inability to utilize the data without wasting it.
- Costs involved in dealing with Big Data

5. DATA WAREHOUSE

Data Warehouse is the process of transforming all multiple data formats into a single format and consolidating them in one place. According to Bill Inmon [4] data warehouse was defined as "A data ware house is a subject oriented, integrated, time variant, non volatile collection of data in support of management's decision making process". Advantages of data warehouse are, (i) helps to gain overall perspective from various parts of the organization, (ii) Consolidate old and historical data, (iii) faster and easier to retrieve, (iv) Data presentation and reporting are very flexible powerful and reliable. Data Warehouses play important role in data integration where the existing data warehousing tools are inefficient in handling Big Data. The following section describes Data Warehouses in Big Data[24][25].

6. DATA INTEGRATION

Data integration [5][7] combines large data from different sources and converts it into a single format.

Data Integration perform important role in organization to make right decisions at right time. Data Warehouses with On Line Analytical Processing (OLAP) tools are used to integrate and analyze data. The traditional Data Warehouses and data mining techniques perform integration on large volumes of data only. They are inefficient to perform integration on streaming and unstructured or semi structured data.

7. BIG DATA INTEGRATION

Big data integration (BDI) means linking or fusing large volumes of heterogeneous data from many dynamic data sources. The main differences of BDI from traditional Data Integration are [5][7],

(i) The number of data sources even for a single domain has grown to be in the tens of thousands.

(ii) Many of the data sources are dynamic.

(iii) The data sources are extremely heterogeneous in their structure.

(iv) The data sources are of widely differing qualities with significant differences in the coverage, accuracy and timeliness of data provided.

7.1. Categories of Big Data Integration

There are two broad categories in Big Data Integration. They are,

1. The integration of multiple big data sources in big data environments.
2. Integration of unstructured big data sources with structured enterprise data.

8. GENERAL STEPS IN BIG DATA INTEGRATION

Step1: Data from different sources are extracted.

Step2: Conversion of unstructured/semi structured data into structure data.

Step 3: Transforming structured data into Data warehouse

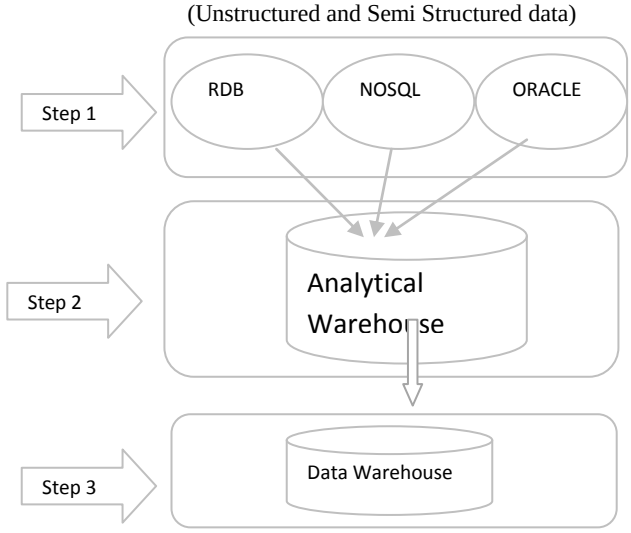

Figure 2. Framework for BDI

The general steps to perform data integration in Big Data environment are identified and it differs from the traditional materialized warehouse. In traditional data warehouses structured data from different sources are collected and loaded. A new analytical warehouse is introduced in this framework, which collects unstructured and semi structured data from different sources and it performs filtering and validations on unstructured and semi structured data. Finally the structured data produced by analytical warehouse are stored in Data Warehouse.

9. CHALLENGES IN BIG DATA INTEGRATION

The following techniques are used in Data Warehouses to perform Data Integration. These techniques raise some issues in Big Data environment. The schema mapping, record linkage and data fusion are challenging areas of data integration in Big Data environment[5].

9.1. Schema mapping:

Schema mapping in traditional warehouses extract and integrate structured data from web tables and web lists. The number of data sources also increases the variety of data and the traditional schema mappings are inefficient and infeasible to integrate data from different sources. Hadoop has introduced an approach "schema on read" which gives freedom to define the schema after the data has been stored [1].

9.2. Record Linkage:

Record linkage refers to the task of identifying records that refer to the same logical entity across different data sources, especially when they may or may not share a common identifier across the data sources. Record Linkage has traditionally focused on linking a static set of structured records that have the same schema. In BDI, record linkage raises some challenges because of 5 V's. The main challenges of record linkage in BDI's are,

a) Data sources are heterogeneous in their structure and many sources provide unstructured text data.

b) Data sources are dynamic and continuously evolving.

9.3. Data Fusion:

Data fusion refers to resolving conflicts from different sources and finding the truth that reflects the real world. Unlike schema mapping and record linkage, data fusion is a new field that has emerged only recently. Its motivation is exactly on the veracity of data [2][3][4].

10. TECHNIQUES FOR MEETING BIG DATA INTEGRATION CHALLENGES

This section describes the existing techniques to perform data integration and what are the problems they raised in Big Data Integration[5][8].

10.1. Techniques on Schema mapping:

10.1.1. Schema on read:

Hadoop[21] community has introduced "schema on read" mapping techniques which gives the freedom to define the schema after the data has been stored.

10.1.2. Data Space Systems:

Data space systems is to provide best effort services such as simple keyword search over the available data sources at the beginning, and gradually evolve schema mappings and improve search quality over time.

10.2. Techniques on Record Linkage:

10.2.1. Parallel record linkage using map reduce:

Technique for adaptive blocking and to balance load among different nodes.

10.2.2. Clustering Techniques:

To manage dynamic and continuously evolving data sources, clustering techniques have been proposed.

10.2.3. Linking text to structured data:

This technique has been proposed to tag and match free text to structured data.

10.3. Techniques for Data Fusion:

In Data fusion numbers of techniques have been proposed particularly to solve the veracity related challenges [11][12].

10.3.1. Online fusion:

Online fusion aims at finding the single truth from conflicting values and to find multiple truths.

10.3.2. Fusion in a Dynamic World:

This approach discovers the truth from dynamic data.

10.3.3. Combining fusion with linkage:

This approach combines the data fusion with record linkage to discover the truth.

11. OPEN PROBLEMS IN BDI

This section describes the open problems in Big Data Integration which will help the researchers to know the cutting edge in big data integration [7][20].

a) Integrating crowd sourcing data

b) Integrating data from data markets.

c) Providing an exploration tool for data sources.

d) Optimal Data Layout for different types of data.

e) Indexing techniques to improve data access performance.

f) Bridging structured and unstructured data together for analysis.

12. CONCLUSION

Today data are being generated, collected and analyzed at an unprecedented scale. Big Data Integration is the one of the major issues in Big Data Environment. This paper reviewed the techniques for data integration in addressing the challenges raised by Big Data including volume, velocity, variety and veracity. From the study of Big Data Integration, it is identified as the existing techniques and approaches are inefficient to handle the problems of data heterogeneity. Therefore new frameworks, techniques and algorithms are expected in future to manage this situation. In future a mechanism can be proposed to handle the data integration issue in Big Data environment. And also this paper identified some open problems in big data integration for future research.

13. REFERENCES

[1] Soumy sen, Ranak Ghosh, Debanjali, Nabendu Chaki,2012.“Integrating XML Data into Multiple ROLAP Data Warehouse Schemas”, International Journal of Software Engineering and Application (USEA), Vol 3,No.1, Jan 2012.

[2] Rabah Alshboul, 2012. “Data Warehouse Explorative Study”, Applied Mathematical Sciences, Vol.6, 2012, No.61, 3015-3024.

[3] An Oracle White Paper, Sep 2013. Big Data and Enterprise Data: Bridging Two worlds with Oracle Integrator12C(ODI12C).

[4] Rajni Jindal, 2012. “Comparative study of Data Warehouse Design Approaches: A Survey”, International Journal of Database Management Systems (IJDMS), vol.4, No.1, Feb 2012.

[5] Xin Luna Dong, Divesh Srivastava, 2013. “Big Data Integration”, ICDE conference 2013.

[6] Andreas Schultz,, Andrea Matteini, Robert Isele, 2012. “LDIF- A framework for Large-Scale Linked Data Integration” www 2012 Developer Track, Apr 18-20, 2012, Lyon, France.

[7] Sachchidanand Singh, Nirmala Singh, 2012. “Big Data Analytics”, International Conference on Communication, Information & Computing Technology (ICCICT), Oct 19-20, 2012.

[8] Gueyoung Jung, Nathan Gnanasambandam, Tridib Mukherjee, 2012. “Synchronous Parallel Processing of Big-Data Analytics Services to Optimize Performance in Federated Clouds”, 2012 IEEE Fifth International Conference on Cloud Computing, IEEE.

[9] Yuri Demchenko, Paola Grosso, Cees De Laat, Peter Membrey, 2013. “Addressing Big Data Issues in Scientific Data Infrastructure”, IEEE.

[10] Jiaqi Zhao, 2014. “ A Security Framework in G-Hadoop for big data computing across distributed cloud data centers”, Journal of Computer and System Sciences 80(2014) 994-1007.

[11] Katarina Grolinger, Mirriam A.M.Capretz , 2013. “Knowlegde as a service Framework for Disaster Data Management”, 2013 workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises,2013.

[12] Aisha Naseer, Loredana Laera, Takahide Matsutsuka, 2013. ”Enterprise BigGraph”, 2013 46th Hawaii International Conference on System Sciences.

[13] Dr.Tulinda Larsen, 2013. “Cross-Platform Aviation Analytics Using Big Data Methods”,IEEE.

[14] Chih-Wei Lu1, 2013.“An Improvement to Data Service in Cloud Computing with Content Sensitive Transaction Analysis and Adaptation”**,** 2013 IEEE 37th Annual Computer Software and Applications Conference Workshops, 2013.

[15] Jin Zhou, 2013. “ An Efficient Multidimensional Fusion Algorithm for IoT Data Based on Partitioning”,

[16] Xiongpai QIN,Huiju WANG, Furong LI, Baoyao ZHOU 2012. “Beyond Simple Integration of RDBMS

and MapReduce-Paving the Way toward a Unified System for Big Data Analytics: Vision and Progress", 2012 Second International Conference on Cloud and Green Computing, IEEE.

[17] Steven J. Rysavy, Dennis Bromley, and Valerie Daggett, 2014. "DIVE: A Graph-Based Visual-Analytics Framework for Big Data" March/April 2014 Published by the IEEE.

[18] David Loshin, "Big Data Analytics", Elsevier, 2013.

[19] Kiyana Zolfaghar, Naren Meadem, Ankur Teredesai, Senjuti Basu Roy, Brian Muckian, 2013. "Big Data Solutions for Predicting Risk-of-Readmission for Congestive Heart Failure Patients ",2013 IEEE International Conference on Big Data, IEEE.

[20] Sandro Fiore, Cosimo Palazzo, Alessandro D'Anca, Ian Foster, Dean N. Williams, Giovanni Aloisio, 2013. "A big data analytics framework for scientific data management" 2013 IEEE International Conference on Big Data.

[21] A White Paper, 2013. "Aggregation and analytics on Big Data using the Hadoop eco- system".

[22] A White Paper, 2013. "SAAS Institute in USA, "Big Data Meets Big Data Analytics".

[23] C.N.Hofer and G.Karagiannis, 2011. "Cloud Computing services: taxonomy and Comparison".

[24] Yi Yuan, Haiyang Wang, Dan Wang, Jiangchuan Liu, 2012. "On Inference- aware provisioning for cloud-based Big data Processing".

[25] Raymond Gardiner Goss and Kousikan Veeramuthu, 2013. "Heading Towards Big Data Building A Better Data Warehouse For More Data, More Speed, And More Users".

[26] Yuri Demchenko, Paola Grosso, Cees de Laat, Membray, 2013." Addressing Big Data Issues in Scientific Data Infrastructure".