A Review on Integration of Big Data Base

Rashmi Shrivastava[1], Gaurav Kumar Saxena[2], Kailash Patidar[3]
P. G. Scholar[1], Assistant Professor[2], H.O.D.[3](Computer Science)
[1,2,3]Shri Satya Sai University of Technology & Medical Science, Sehore, M. P., India
Email – [1]rashu25j@yahoo.co.in, [2]gaurav.saxena18@rediffmail.com, [3]kailashpatidar123@gmail.com

Abstract: *In all over the world technologies and there challenges to manage these technologies are increased day by day. Data have grown from zetabytes to petabyte every month. The term big data is basically the high volume of data. Basically integration of data is the process of transferring the data from source format to destination format. Data is collected from different type of social sites like e-mail, web server, social networking sites etc. The integration of this huge amount of data collected from different sources is very important. The collection of these data is in unstructured and semi structured form. Because we are using huge amount of data transfer and problem arise to transfer that data 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 by different scholars.*

Key Words: *Data warehouse, Hadoop, Big Data integration, Integration, Information center.*

1. INTRODUCTION:

Big theory requires more technique to transfer huge amount of data daily. The requirement of this technique is to manage this transfer of data from source to destination. But in big data environment data from different sources are of different formats and we have some technique to manage these data and handle such situation. Many data warehousing and data management approaches have been supported by integration tools such tool as ETL- Extracts, Transform and Load. Extract, transform and load are three database function combined in one place to another. ETL comes from data warehouse but data warehouse technique is not handling or manages the big data integration. ETL- Extract, Transform and load performs the process of transferring the data from the source to destination and manage the information of data.

- **Extract** – The task of extract is to collect the data from external source understand and assess the quality of data. The purpose of extract is to understand the format of data, assess the overall quality of the data and to extract the data from its source so that it can be manipulated in next task.
- **Transform-** The purpose of transform the source format into destination data format understanding the source format and assess the destination format.
- **Load-** Store the data into target system, understand the target system and access the fault tolerance.
- Data warehouse is process of transforming all data format into a single format. Date warehouse is manually coded integration programs using java map reduce and Hadoop's. Data are generated from different type of social networking sites, web server, e-mail, web browser etc. all are in unstructured form. To solve these problem organization are try to find out different types of new technologies. Big integration of data has big high volume, high velocity, high variety, high veracity and high value of information.
- **Volume**-Volume is the most important feature of big data theory which adds some additional technologies and tool. Volume contains not only the large amount of data but also the number of sources. Volume has some features as scale, size, amount for data processes and stored in files or database.
- **Velocity**- Big data is generated at high velocity also data generated from different events, arrays sensors in real time.
- **Variety-** Data variety include new requirement of data storage and data format. Variety deals with the complexity of big data and information.
- **Veracity-** Data veracity has security that data must be trusted, original, secure from unauthorized attack. During the whole life cycle data must be secured from trusted source to trusted compute and stored in a trusted and protected storage.
- **Value-** Value is an important feature of data integration defined as added value that collected from different source.

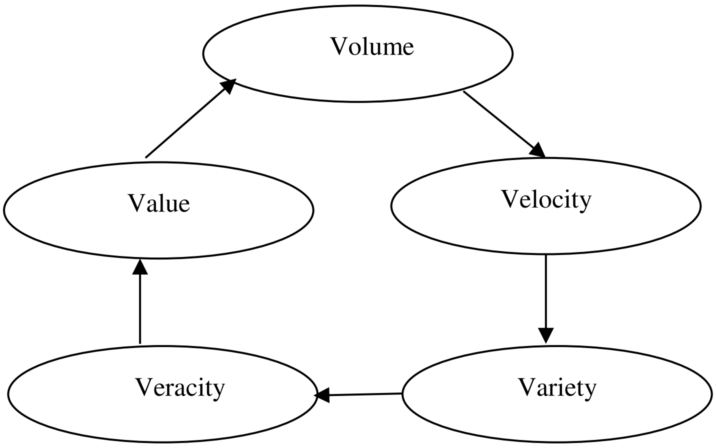

Figure 1: Flow chart of big integration of data

Data collected from different type of social sites such as web server, different browser, e-mail, etc. As a result data are collected in semi structured, unstructured, structured from. To manage these problems organization are try to find different technologies. So solve these problem we are using some tool like –ETL (Extract, transform and load) such as Data warehouse. In computing Extract, Transforming and load (ETL) refers to a process in database usage and especially in data warehousing. Extract, transforming and load is short for extract transform, load three database function that are combined into one tool to pull data out of one database and place to another .First the extract read the date from specified source database and extracts a desired subset of data. To convert it to the desired position we are using some rules or lookup tables. ETL(Extract, transform and load) be a temporary subset of data for report requirement of more date for other purpose as data warehouse; conversion data from one database type to another database type To manage all three database refer to three separate function combined into a single programming. Data warehouse is the process of transforming all information of data format into a single. The challenges are common on big data project in many of the cases the integration of the data process is likely to become more complicated to manage as all encompassing data warehouse and rigid ETL routines give the way to more dynamic environment involving a variety of different system.

As the increase of database day by day, the requirement of modern IT Database is the ability to handle that amount of the large volume of data and the original impulsion behind the NoSQL movement. But the ability to handle that amount of data or scale is something that the all databases categorize as NoSQL share. Example along with the large number of user who use data base on the regular bases of petabytes and perform thousand or lakhs of operation per second and cluster exceed thousand of node. However in modern IT database it must to do more than large scale.

2. BIG DATABASE:

The big database of the big environment usually includes data sets with the biggest sizes beyond the ability which is commonly used for the software tools to manage, curate, capture, manage, and process data within the tolerable elapsed time. Extremely large amount of data sets that may be analyzed computationally to expose patterns, trends, and associations, it is especially related to the human behaviour and there interaction. Investment of IT is going towards managing and maintaining big data"

The term big data is describes the large volume of data – both structured and unstructured – that inundates a business on a day-to-day basis. But it's not the amount of data that's important. It's what organizations do with the data that matters. Big data can be analyzed for insights that lead to better decisions and strategic business moves**.**

Data growth challenges and opportunities as being three-dimensional, i.e. increasing volume, velocity, and variety. Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization." Gartner's definition of the 3Vs is still widely used, and in agreement with a consensual definition that states that "Big Data represents the Information assets characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value.

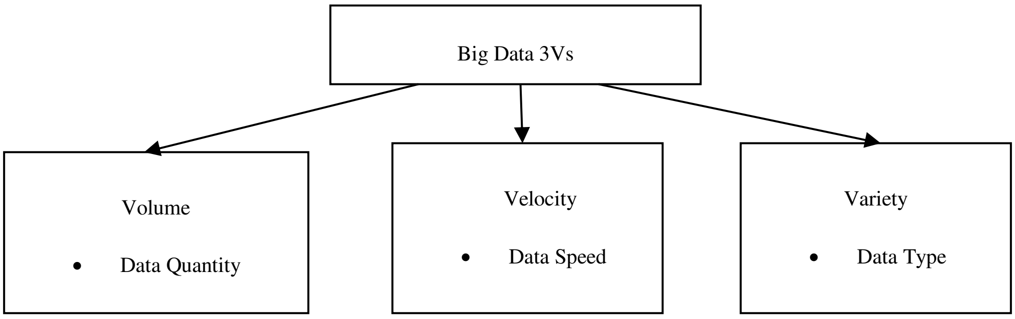

Figure 2: Block Diagram of Big Data

3. BIG DATABASE ARCHITECTURE CAPABILITIES :

Data Integration Capability, Statistical Analysis Capability, Database Capability, Storage and Management Capability, Processing Capability, Big Data Architecture Capabilities:

1. Across the cluster Write multiple copies for the fault tolerance, processing scalable highly parallel batch, implementation of Leading Map Reduce, Apache Hadoop are able to distribute the data under workloads across the thousands of the nodes. Breaking problem into the smaller sub-problems has Map Reduce Processing Capability.
2. With the parallel data optimized processing import/export of the data. For SQL processing Connects Hadoop to relational databases. Exports the MapReduce results for Hadoop, and exports other targets to the RDBMS, the Data Integration Capability.
3. With no modification of. Programming Oracle R Enterprise allows reuse pre-existing R scripts for analysis the language for statistical capability.

4. BIG DATA REFERENCE ARCHITECTURE :

In the reference architecture of the big data most of the Big Data projects use variations of the Data. Provides the good background for understanding of the Big Data at the view of high level of this reference architecture which provides the good background for understanding the Big Data architecture and to understand how it complements the existing analytics, databases, BI and systems. One-size-fits-all approach this architecture of the big data is not fixed. Each component of the big data reference architecture has at least several numbers of alternatives with its own to get the advantages and disadvantages for the particular workload of data architecture. In this architecture companies are often start with the subset of the patterns of the architecture of data, and as in insight they realize value for gaining the key business outcomes and they expand the breadth of use of the data architecture. Figure below shows the big data reference architecture.

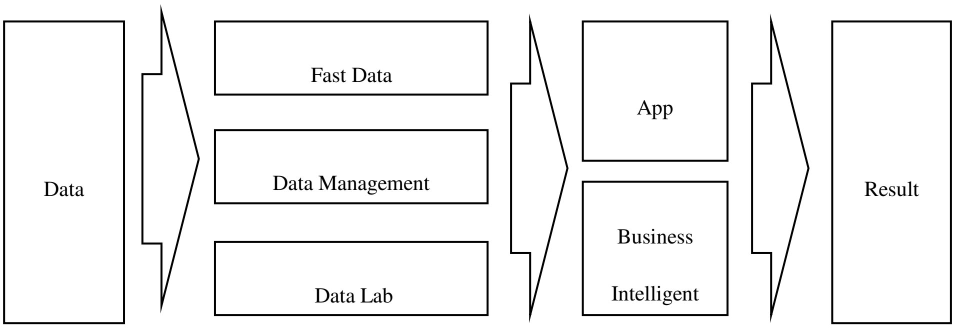

Figure 3: Big Data Reference architecture

5. REVIEW:

A number of scholars has performed work on the various aspects related to big data applications. Some works are reviewed in short in this section. Reference [1] present the paper on technology depended of reference architecture of big data base system. This paper shows the work on technology and design of architecture of product and services in the big data system. When we construct a big data system associated classification and reference architecture are aimed to facilitate selection of technology and architecture. Reference [2] presents the Hadoop and NoSQL technology tool to manage evolves of big data in ecosystem. This system can provides the real value by disparate data access APIs. To access data in multiple stores unified query system which allow single query.

Reference [3] presents a unified the approach of spatial data query in GIS (Geographic Information system). The paper presents the framework for integrating of data information from the stored dataset value. The paper solves the problem of development in Geographic Information system (GIS) application in that there is no interoperability exists among the different database. Reference [4] presents the review in which many scholars present different techniques for the issues and challenges of data integrating in big data environment. This paper solves the problem on future research of data integration in big data environment.

Reference [5, 10] paper presents framework to convert XML schema to ROLAP data warehouse schema. The paper solves the problem on concentrating the technique of converting XML to the relational model and the increment on the challenges because of unstructured data such as in XML. Reference [6] shows the importance of IC’s enhancing end user deals with the satisfaction of data warehouse in big data environment. Paper solves the problem data warehouse application usually takes more time to perfect and develop. Reference [7, 9] paper presents the ETL technique to handle the challenge to manage the data environment. This solve problem of big data integration are sketched to proceed on research in future in big data environment.

Reference [8] presented a paper on different approaches and schema used for the design of data warehouse in big data environment at different level. We also offer Object Oriented framework to design the data warehouse. Finally object oriented help to solve the problem of data warehouse in big data environment. Reference [11, 12] present the full text

Information Retrieval (IR) to manage the demand of application in increment of information and able to the huge amount of existing collection of web page. The paper gives the solution of the problem to re-index the existing collection and operation of whole collection with cost proportion to the size of collection.

Reference [13] present the paper which propose a new model for searching of semi structured data set with simple keyword search and interactively query. This model evaluates the new requirement of the date base system. This paper evaluates the technical uses of the database system. Reference [14] present the garlic project on which the aim of this project is to associated and develop the tool to manage large database from multimedia environment which are in different format so that it is capable of integrating of different variety of data. This paper solves the problem in distributed database system and we will work on that area.

Reference [15] presents paper on new data structured called Banian which analyze and manage the big data system. It solves the problem of limitation of storage of data and the data query with the large scale storage management of data. Reference [16] this paper describe the sematic technology for the management of the system provides the data to access the layer and made the relationship between graphical user interfaces. This prototype system built for water's power generation and GE power. This gives the result to solve the problem of expenditure and productivity saving. Reference [17] represents the technical topic HANA integration for collaboration with Hadoop based infrastructure. This paper is to manage the result and pattern design of data architectural with SAP HANA database in big data infrastructure.

Reference [18] presents the architecture on cloud computing to test the different level of data base management of the system based on Hadoop. In this paper we solve the problem of complex data application because cloud computing technology is still in its beginning. Reference [19] present the value of big data in market today and shows the difference of big data collection in past and manage database the challenges in present includes in antitrust enforcement. This paper present inquiry on big data collection and define the analysis on aware of antitrust enforcement in coming year. Reference [20] present the design on query language known as VQuel .the goal is to support intermediate and final result of unified query language. This paper discuss about the design of key language and challenge of implantation for moving forward to the design.

6. CONCLUSION:

Now a day's data are being generated form collected and analyzed at an exceptional scales. Big Data Integration is a major issue in Big Data Environment. This paper reviewed the different techniques for big data integration in addressing the challenges raised by Big Data including volume, velocity, variety and veracity. From the different study of Big Data Integration, it is recognized that the existing techniques and approaches are inefficient to handle the problems of data. Therefore new 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.

7. ACKNOWLEDGMENT:

The authors would like to thank Ms. Rashmi Patwa, Research Engineer, for their many constructive comments and suggestions in this paper. They would like to thank the contributions of Innovative Technology Design and Training Centre (ITDTC), BHOPAL, Directed by Mr. Piyush Jain for their Support in Technology Data.

REFERENCES:

1. Pekka Pääkkönen, and Daniel Pakkala, "Reference Architecture and Classification of Technologies, Products and Services for Big Data Systems", Elsevier, Feb.2013, Pg. 166-186.
2. Jeanne W. Ross, and Peter Weill, and David Robinson, "unified Query on Big Data Management system", Oracle white paper, March 2016, Vol.
3. Mohammed Abdalla, and Hoda M. O. Mokhtar, and Mohamed Noureldin, "A Unified approach for spatial Data query", International Journal of Data Mining & Knowledge Management process, November 2013, Vol.3. No.6.
4. B. Arputhamary, and L. Arockiam, "A Review on Big Data Integration", International Journal of Computer Applications, 2014.
5. Soumya Sen, and Ranak Ghosh, and Debanjali Paul, and Nabendu Chaki, "International Journal of Software Engineering & Applications, January 2012, Vol.3., No.1.
6. Lei-da Chen, and Khalid S. Soliman, and En Mao, and Mark N. Frolick, "Measuring user satisfaction with data warehouses: an exploratory study", Elsevier, 2000. Vol no., Issue no., Pg no.103-110.
7. B. Arputhamary, and L. Arockiam, "Data Integration in Big Data Environment", Bonfring International Journal of Data Mining, February 2015. Vol. 5, No. 1.
8. Rajni Jindal, and Shweta Taneja, "Comparative study of data warehouse Design approach: A Survey" International Journal of Database Management Systems, February 2012, Vol.4, No.1.

9. Xin Luna Dong, and Divesh Srivastava, "Big Data Integration", (International conference on engineering 2013 Seminar), April 2013, Issue no. 1063-6382, Pg.1245-124.
10. Sriram Raghavan, and Hector Garcia-Molina, "Integrating Diverse Information Management Systems: A Brief Survey", Infolab publication server, December 2008.
11. Eric W. Brown, and James P. Callan, and W. Bruce Croft, "Fast Incremental Indexing for Full-Text Information Retrieval", International Conference on Very Large Data Bases, 1994, Pg. 192-202, Setpember.
12. Sergey Melnik, and Sriram Raghavan, and Beverly Yang, and Hector Garcia-Molina, "Building a Distributed Full-Text Index for the Web", ICDESA, July 2001, Vol. no.19, Issue no.3, Pg.217-241.
13. Roy Goldman, and Jennifer Widom' "Interactive Query and Search in Semistructured Databases", The world wide web database, 1998, Vol. no. 1590, Issue no., Pg.52-62.
14. Michael J. Carey, and Laura M. Haas, and Peter M.Schwarz, and Manish Arya, and William F. Cody, and Ronald Fagin, and Myron Flickner, and Allen W. Luniewski, and Wayne Niblack, and Dragutin Petkovic, and John Thomas, and John H, andWilliams and Edward L. Wimmer. "Towards Heterogeneous Multimedia Information Systems: The Garlic' Approach", 5th Int'l Workshop on Research Issues in Data **11** Engineering: Distributed Object Management, 1995, Pg 124-131.
15. Tao Xu, and Dongsheng Wang, and Guodong Liu," Banian: A Cross-Platform Interactive Query System For Structured Big Data", Tsinghua Science and Technology February 2015, Volume 20, No.1, Issue no.8, Pg. 62-71.
16. Jenny Weisenberg Williams, and Paul Cuddihy, and Justin McHugh, and Kareem S. Aggour, and Arvind Menon, and Steven M. Gustafson, and Timothy Healy, "Semantics for Big Data Access & Integration: Improving Industrial Equipment Design through Increased Data Usability", IEEE International Conference, 2015, Pg.1103-1112.
17. Norman May, and Wolfgang Lehner, and Shahul Hameed P., and Nitesh Maheshwari, andCarsten Müller, and Sudipto Chowdhuri, and Anil Goel, "SAP HANA –From Relational OLAP Database to Big Data Infrastructure", open proceeding, July.2015., Vol. no.15, Issue no.2 , Pg.141-152.
18. Junbin Duan, and Pengcheng Fu, and an Gong, and Zhengfan Zhao, "Design of Test Data Management System Architecture Based on Cloud Computing Platform", International Conference on Circuits and Systems, Vol. no., Issue no., Pg., May.2015.
19. Shepard Goldfein, And James A. Keyte, "Antitrust and 'Big Data': New Terrain for Inquiry?" Newyork lam journal, March.2016. Volume 255. No. 43.
20. Amit Chavan, and Silu Huang, and Amol Deshpande, and Aaron J. Elmore, and Sam Madden, and Aditya Parameswaran, "Towards a Unified Query Language for Provenance and Versioning", USENIX Association Berkeley, July.2015, Pg. 1-5.