OPEN ACCESS at journalijcir.com

Review Article

ISSN: 2395 -5775

IMPLEMENTATION OF UNIFIED QUERY FOR BIG DATABASE USING SQL ON ORACLE PLATFORM

Rashmi Shrivastava[1]., Gaurav Kumar Saxena[2] and Kailash Patidar[3]

Abstract

Today data have sprout from petabytes (10^{15} bytes) to zettabytes (10^{21} bytes). Day-by-day the increment in technologies all over the world is leading to an increase in information exchange and the challenges of data processing. The information, often called as "data", is integrated with the help of different tools and software. Data integrating is the process of transferring the data from one format (also referred as "source format") to another format (also referred as "destination format"). A very large amount of data, also referred as "big data", is always followed by the problem of transferring the data. ETL (Extract, Transform and load) Tools are the best options to manage the data transfer. In ETL three database functions are combined in a single tool using the process called "data warehousing". Data warehouse transforms all data formats into a single format for the purpose of data transfer. Sometimes, date warehouse is achieved by manually coded integration programs using java, map reduce, SQL, Hadoop, or any other software language. Since the data is generally generated from different sources like social networking sites, web servers, e-mails, web browsers etc. so they are in comparatively unstructured form. To solve the problem of unstructured data transfer using single software tool the market organizations are always in need of an efficient solution using the various technologies to handle the big data.

Key Words: Big Data, Data integration, Data warehouse, ETL Tool, SQL.

INTRODUCTION

Data sets are sprout up rapidly in all over the world because they are increasingly accumulated by enormous information. The technologies in the world per capita capacity to store information has increase or double every year or month since 1980 to now. So big data in big environment has increased or rapidly or constantly form few tera-bytes(10^{12} bytes) to many peta-bytes (10^{15} bytes). Big data theory requires more techniques and technologies to manage data information from source data format to destination data format. Big integration of data has big high volume, high velocity, high variety, high veracity and high value of information. In fig.1 shows the flow chart of big data integration.

1. *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, and amount for data processes and stored in files or database.
2. *Velocity*- Big data is generated at high velocity also data generated from different events, arrays sensors in real time.
3. *Variety*- Data variety include new requirement of data storage and data format. Variety deals with the complexity of big data and information.
4. *Veracity*- Data veracity has security that data must be trusted, original, and 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.
5. *Value*- Value is an important feature of data integration defined as added value that collected from different sources.

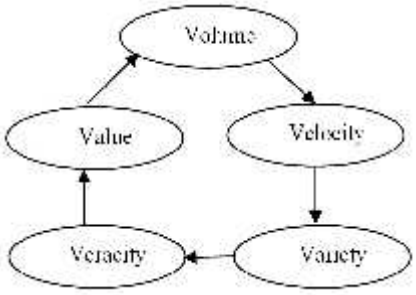

Fig. 1 Flow chart of a simple model of big data integration for data transfer

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.

[1,2,3]Department of Computer Science, Shri Satya Sai University of Technology & Medical Science, Sehore, M.P., India

Correspondence and Reprint Requests: Rashmi Shrivastava

Received: January 15, 2016 | **Accepted:** February 4, 2017 | **Published Online:** March 28, 2017

Conflict of interest: None declared | **Source of funding:** Nil

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. Fig 2 shows the diagram of big data environment that can require a big shift in the big data management principle and procedure covering the data integration. Data volume is continuously increasing from social media and from internet. It transfers the data among the nodes even in the state of node failure without any interruption.

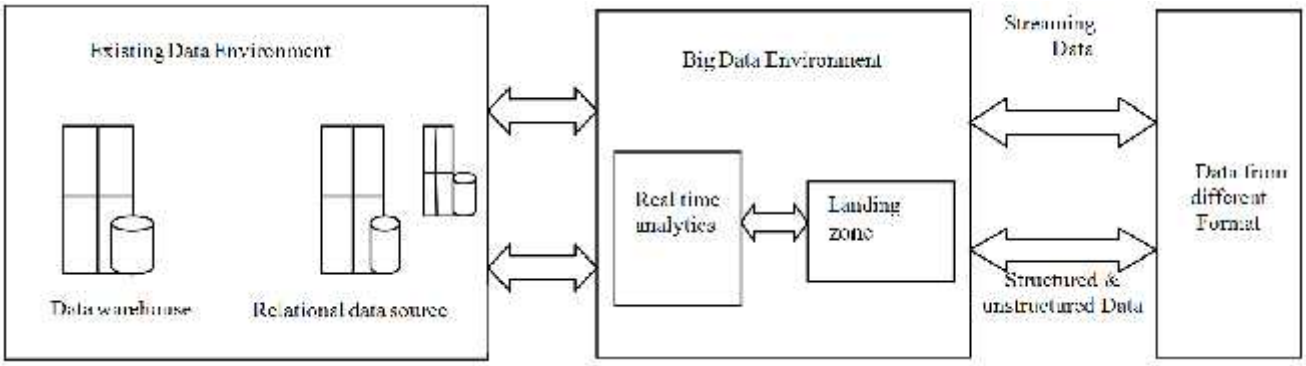

Fig 2- Data Flow in a Database System

Fig 3 shows the communication between the client and application server. The three clients i.e., Client 1, client 2 and client 3 they send their request to application server there can be n number of application server. Application servers 1 send the query to database server and same like application server 2 and application server 3. Database server optimized and processed the query and returns the output in data format to application server.

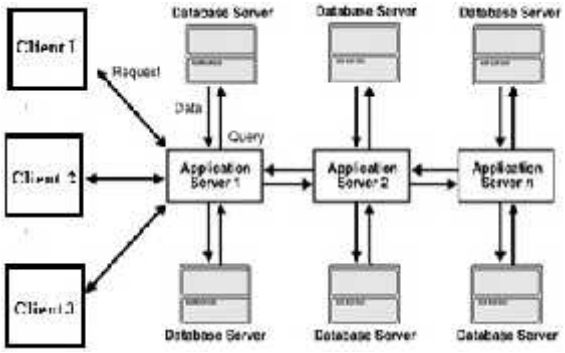

Fig 3 Communication System Architecture in Big Data Environment

LITERATURE REVIEW

Reference [1] presents the paper on technology depended of reference architecture of big data base system. This paper concentrated on reference architectures, commercial products/services and technologies for big data systems. Especially, the survey focused on stream pro-cessing, graph modelling, business intelligence and visualization technologies, big data benchmarks, virtualization and cloud solutions, new technology frameworks, and commercial products/ser-vices. It works 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). This paper introduces the issues of data interoperability, advantages of Geo-Graphic metadata, and its mechanism for data interoperability. In this paper we proposed an interoperable framework for spatial data query.

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 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. This paper solves the problem on future research of data integration in big data environment. Reference [5, 10] presents the 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], [8] paper present 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 [9] 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. Paper proposed and compared different schemes for storing and managing inverted less using an embedded database system. We showed that an intelligent scheme for packing inverted lists in the storage structures of the database can provide performance and storage Efficiency comparable to tailored inverted less implementation. Reference [13] presents the paper which proposes 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. This paper systematically introduces Banian and describes the associated system architecture, as well as the design principle and working flow of each module 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.

IMPLEMENTATION OF UNIFIED QUERY ON BIG DATABASE USING SQL

The high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making is defined as Big data. A single declarative statement can access and analyse data in many disparate stores is known as Unified query. In this paper Following are the steps that are followed to develop a unified query on big database with different architectures. Select the oracle database control for the recovery of the services. Check the status of services, which services has the status "automatic started" and which has "disabled" in the oracle services. Now go to the Oracle SQL developer to show the database table format as shown the fig. 4(a) –

The three databases have been created in Oracle SQL developer as in fig.4 (a): Scott, Scott1, and Scott2. The view of Scott database in Oracle SQL developer is shown in Fig 4(b). It contains Tables, views, Editioning Views, indexes, packages, procedure, function, queues, queues tables, triggers, type, sequences, materialized views, materialized views logs, synonyms, public synonyms, database link, public database links, directories, editions, application express, java, xml schemas, xml DB Repository, Recycle bin, jobs, other users

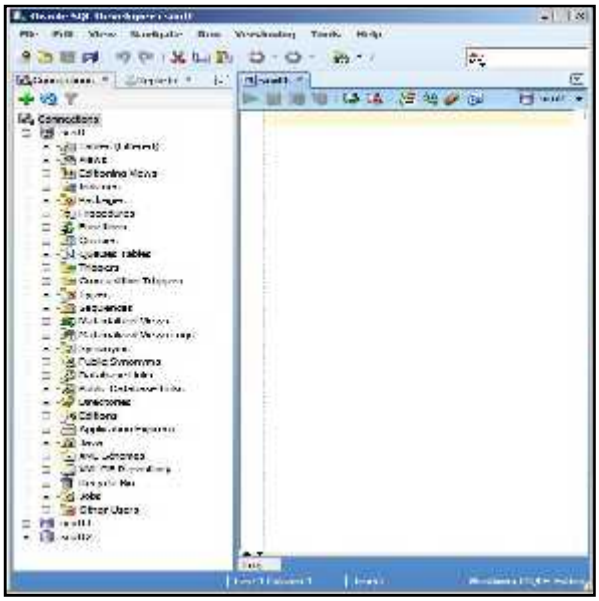

Fig 4 (a) Oracle SQL Developer (b) Big Data Environment

Fig 5 shows "IT_PROJ_ORG" Tables view which contains Column_ID, Column_name, Data_type, nullable, Data_default, comments. Column_ID is auto incremented value. Column_name like ACT_MTH, MOBILENUMBER, REC_OCT, CIRCLE, FCD, RECHARGE_AMOUNT, CATEGORY_NEW etc. These fields has character type of datatype like varchar2 (7byte), varchar2 (12byte), varchar2

(7byte), varchar2 (6byte), varchar2 (18byte), varchar2 (16byte), varchar2 (12byte) respectively with null able values. The three different times analysis of three different table is can be measured as shown in fig. 6, fig. 7 fig. 8.

In the present work the time based analysis of the data access execution for three different data bases is recorded in following table:

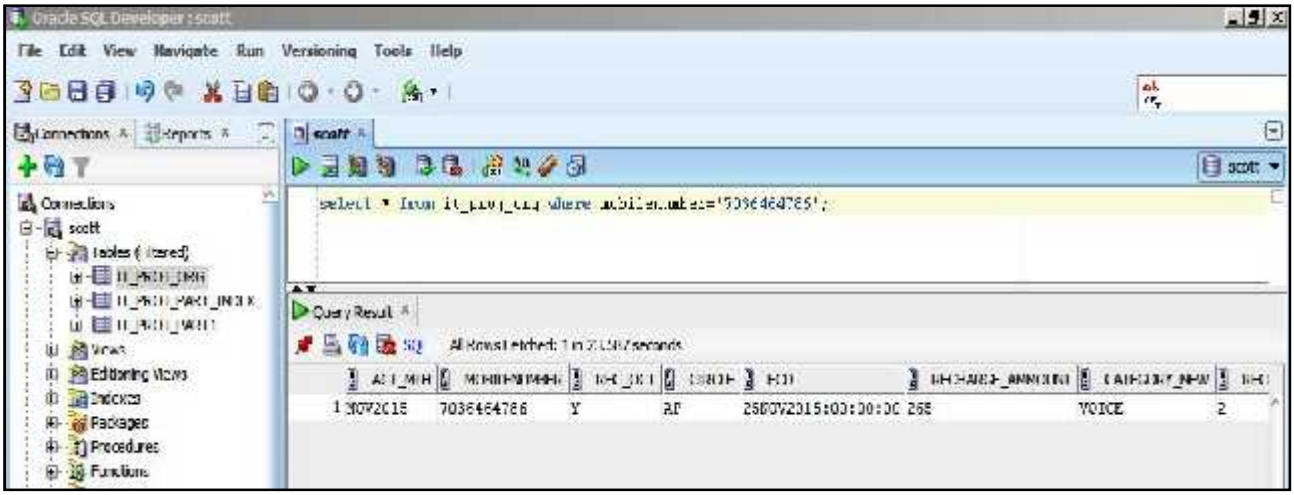

Fig. 6 Oracle SQL Developer - IT_PROJ_ORG

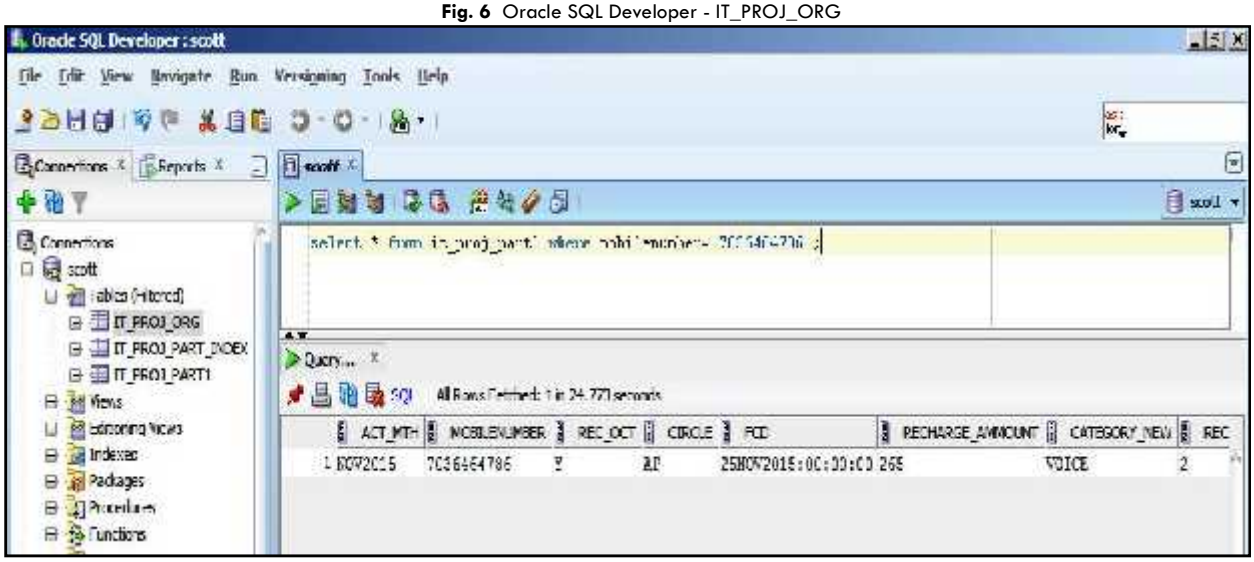

Fig. 7 - Oracle SQL Developer - IT_PROJ_PART1

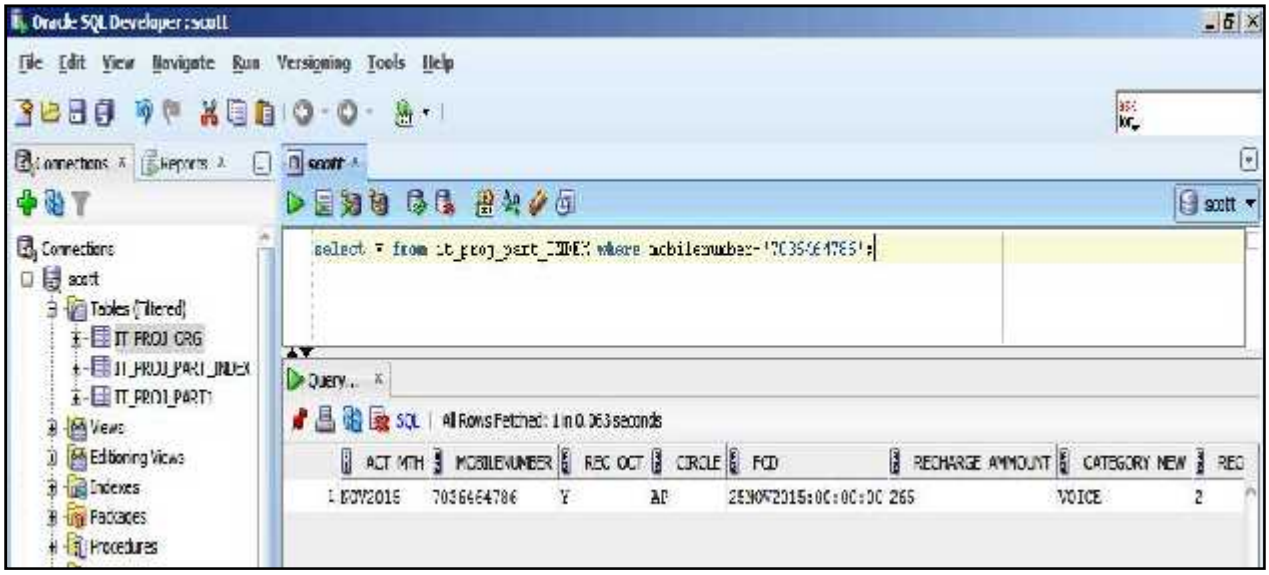

Fig. 8 - Oracle SQL Developer - IT_PROJ_PART_INDEX

TABLE I TIMING PERFORMANCE OF UNIFIED QUERY IN DATABASES

Data Type	Column Name	Table	Time (Seconds)
Varchar2(12 byte)	Mobile number	IT_PROJ_PART1	24.773
Varchar2(12 byte)	Mobile number	IT_PROJ_ORG	23.587
Varchar2(12 byte)	Mobile number	IT_PROJ_PART_INDEX	0.063

Table I describes the calculated time for running the query in different tables of Scott database. In table IT_PROJ_PART1 takes 24.773 Seconds for running the query and in same query for table IT_PROJ_ORG takes 23.587 Seconds and when indexing has been done in table IT_PROJ_PART_INDEX then execution timing of same query has been reduced to 0.063 Seconds. Fig 9 shows the time performance graph.

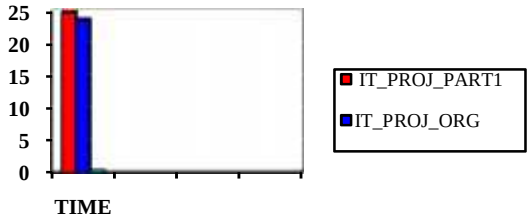

Fig.9 – Timing performance graph

CONCLUSIONS

Today data is generated from different type of social networking sites like e-mail, web server, web browser etc. which are in unorganized form. As the increment of date in both science and industrial field there is also increment in amount of information of data and managing their challenges. The integration of big data is major issues to manage in big data environment. By the study of data integration it is identified that the techniques are inefficient to manage the problem. Therefore new techniques are expected to find out in future to handle this problem of big data environment. This paper helps to find the problem and provide a solution for that which would help to overcome the problem of big data integration for future.

ACKNOWLEDGMENT

The authors thank Rashmi Patwa and Ankita Jain (Research Engineers at Innovative Technology Design and Training Centre (ITDTC) and Varun Suman (SAS-MOBILITY, TCS, Mumbai, India), Bhopal, India) for their many constructive comments and suggestions in this paper.

REFERENCES

[1] Pekka Pääkkönen, and Daniel Pakkala, "Reference Architecture and Classification of Technologies, Products and Services for Big Data Systems", Elsevier, Vol. 2, Issue no. 4, Pg. 166-186, Feb.2013.

[2] Jeanne W. Ross, and Peter Weill, and David Robinson, "unified Query on Big Data Management system", Oracle white paper, March 2016.

[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*, Vol.3. No.6. Issue no., Pg. no. 55-71, November 2013.

[4] B. Arputhamary, and L. Arockiam, "A Review on Big Data Integration", *International Journal of Computer Applications*, Issue no. 5, Pg. 21-26, Feb 2015.

[5] Soumya Sen, and Ranak Ghosh, and Debanjali Paul, and Nabendu Chaki, "Integrating XML Data into multiple ROPAL Data warehouse schemas", *International Journal of Software Engineering & Applications*, Vol.3., No.1, Pg. 197-206, January 2012.

[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*, Vol. 37, Issue 3, Pg no.103-110, April 2000.

[7] B. Arputhamary, and L. Arockiam, "Data Integration in Big Data Environment", *Bonfring International Journal of Data Mining*, Vol. 5, No. 1, Pg. 1-5, February 2015.

[8] Xin Luna Dong, and Divesh Srivastava, "Big Data Integration", International conference on engineering 2013 Seminar, April 2013.

[9] Rajni Jindal, and Shweta Taneja, "Comparative study of data warehouse Design approach: A Survey" *International Journal of Database Management Systems*, Vol.4, Issue 1. Pg. 33-45, February 2012.

[10] Sriram Raghavan, and Hector Garcia-Molina, "Integrating Diverse Information Management Systems: A Brief Survey", Infolab publication server, vol. 24, Issue 4, Pg. 44-52, 2001.

[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*, Pg. 192-202, Setpember.1994.

[12] Sergey Melnik, and Sriram Raghavan, and Beverly Yang, and Hector Garcia-Molina, "Building a Distributed Full-Text Index for the Web", ICDESA, Vol. 19, Issue 3, Pg.217-241, July 2001.

[13] Roy Goldman, and Jennifer Widom, "Interactive Query and Search in Semistructured Databases", The world wide web database, Pg.1-7, 1998.

[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", IEEE, Pg 1-22, 1995.

[15] Tao Xu, and Dongsheng Wang, and Guodong Liu," Banian: A Cross-Platform Interactive Query System For Structured Big Data", Tsinghua Science and Technology Volume 20, No.1, Issue no. 1007-0214, Pg. 62-71 February 2015.

[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, Pg.1103-1112, 2015.

[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, Vol.15, Issue2 , Pg.141-152, July.2015.

[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, Pg. 362-364, May.2015.

[19] Shepard Goldfein, And James A. Keyte, "Antitrust and 'Big Data': New Terrain for Inquiry?" New york lam journal, Volume 255—No. 43, March.2016.

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, Pg. 1-5, July.2015.

****