International Journal of Notural and Engineering Sciences
Uluslararası Doğa ve Mühendislik Bilimleri Dergisi
E-ISSN: 2146-0086, 11 (3): 23-28, 2017, www.nobel.gen.tr

Handling Data Integration Analytics for Students of UAF by using Big Data and SPSS

Badarqa SHAKOOR[1*] Ahsan Raza SATTAR[1]
[1]Department of Computer Science, Faculty of Sciences, University of Agriculture, Faisalabad, Pakistan

*Corresponding Author
E-mail: badarqa25@gmail.com

Received: September 20, 2017
Accepted: November 30, 2017

Abstract

Data Integration refers to associating data from various sources. Most of the organizations store their information in multiple databases, retrieving data from different sources and the integration of data provides a unified view. The problem is maintenance of real time datasets in UAF and more important, handling the heterogeneity of student's records when they are fetched from the different data sources, during integration the data from multiple sources shouldn't be in same format this cause the quality issues. The aim of this study is to handling data integration, analytics for students of UAF by using big data, application help us in saving the required information from the huge amount of data. As per our survey integration tool gave us the opportunity to manage, share and synchronize the large data sets in a same format. The interface of the tool is straightforward for user and there is no need to create any data models. Consistency and scalability of providing datasets have been improved by using these tools. We have implement big data for analysis of student record stores in different databases. Important data is extracted and then loads it into record and building of records, then update and store data on a regular base. In view of our consequences and analysis we summarized that infrastructure of data integration is helpful and increase the value of data through unified system.

Keywords: Big Data; Data Integration; Data Quality; SPSS

INTRODUCTION

There is a huge data of student that management is dealing but unless the maintenance and quality will be a serious issue. Data integration plays a key role in determining the potency of a corporation, be it as the amount of back end systems integration or integration of body processes, tasks, and databases. Integration of student's data leads various issues during admission process or maintaining fee record because the data is coming from heterogeneous data sources in this result the quality of information integration emphasizes on the degree of information storage, structure and therefore the levels at that the info may be integrated and operated as one entity. Grouping and maintaining the massive knowledge sets is expensive, so organizations tend to adapt to cloud methodologies for storing the info and recycle [1].

Data fusion is the strategy for a mix and translation information from very surprising sources to determine data of a fresh out of the plastic new quality. Coordination of databases into a run of the mill storehouse has turned into a quest subject for quite a long while Data fusion could be a horribly propelled drawback, and has pertinence in numerous fields, similar to data re-building, Data Warehouse, web information Systems, E-business, Scientific Databases, and so on the matter of irregularity have conjointly as of late been consideration of enthusiasm inside of the space information stockrooms (DW) as a DW could be a store of coordinated data from dispersed, self-governing, and most likely heterogeneous, sources. Old inquiry frameworks work by coordinating the term that is being looked with the qualities keep inside of the relating data. On the off chance that the information contained in databases is conflicting [4].

Big data is a theoretical idea because various concerns, logical and mechanical ventures, research researchers, information investigators, and specialized professionals have diverse meanings of enormous information. In the late 1970s, the idea of "database machine" rose, which is an innovation exceptionally utilized for putting away and breaking down information. In the 1980s, individuals proposed "share nothing," a parallel database framework, to take care of the demand of the expanding information volume [3]. On June 2, 1986, a turning point occasion happened when Teradata conveyed the main parallel database framework with the capacity limit of 1TB to Kmart to help the vast scale retail organization in North America to grow its information stockroom [2].

To run analysis and visualization the developers usually expend extra time for extraction, reformatting, and integration of data. Due to their complexity, it is hard to write the data reshaping programs. Thus, they are still required because getting data in desired form through analytical tool and different operations like cleaning, extraction, transforming and integration. In a context of big data, the sources have huge volume and heterogeneous data due to this problem the developer cannot go over all the data. Data quality will be effected and preparation of data is capable of being scaled on large data sets [6].

Every Organization must integrate the data of every department so as per university data the data of students among different department can be integrated the problem faced during the integration process is maintenance of real time data set of students which are in the different sources and handling of students record which are heterogeneous in multiple sources. Required information from single data sources cannot be satisfied with the current technological world. Now a day's integration of information will be done by multiple data sources across every organization.

The integration of data employed for combining the data which is residing in heterogeneous and independent data source and a unified global schema is provided to users. The permanent uniting of the sources would be a most useful form of integration. Proposed system sources of university lead to many problems: The problem may be in the operating system and hardware, Faulty software of data management, User interface or business rules and constraint of integrity [5].

DATA INTEGRATION REQUIREMENTS

To fuse heterogeneous information sources requires more than a device for orchestrating information into an ordinary etymological structure. Integration of data is a capricious activity that incorporates bargain at various levels.

Model Of data

Information sources can tremendously differentiate concerning the structures they use to address information (data instance, tables, records, and so forth. Trade off heterogeneous information models requires a regular information model to guide information starting from the distinctive information sources.

Composition of data

When we have settled upon a normal information appear, the issue rises of obliging unmistakable representations of the same component or property. For example, two sources may use different names to address the same thought ("payment" and "cost"), or the same name to address different thoughts ("assignment" to mean both the endeavor an agent is taking a shot at and an endeavor for which a laborer is the examiner), or two courses for going on the same information ("date of birth" and "age"). In addition, information sources may have the same information using unmistakable information structures.

How much data integration infrastructure would help you?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Helpful	13	32.5	43.3	43.3
	More Helpful	11	27.5	36.7	80.0
	Don't know	6	15.0	20.0	100.0
	Total	30	75.0	100.0	
Missing	System	10	25.0		
Total		40	100.0		

Table1: Percentage that found data integration infrastructure helpful

Occurrence of Data

At the event level, reconciliation issues fuse making sense of whether "things" starting from different sources address the same honest to goodness substance and selecting a source while clashing information is found in different information hotspots for case, various birth dates for the same person [7].

TECHNIQUES FOR DATA INTEGRATION

Information systems can be depicted by using a layered outline on the most astounding layer, customer's entrance data and organizations through various interfaces that continue running on top of different applications. Applications may use middleware like trade taking care of exchanged handling (EH) screens, middleware arranged message (MAM), SQL-middleware, thus on to get to data by method for a data access layer. The data itself is administered by a data stockpiling structure. Database organization structures are used to join the data access and limit layer. All things considered, the mix issue can be tended to on each of the presented structure layers. In figure 1 the detailed architectural level [12] of data integration techniques is shown and how we can collaborate the data of two users through these techniques [11].

Manual Integration

In this technique, the customers clearly collaborate with every single huge data structure and physically arrange picked data. That is, customers need to oversee customer interfaces and request tongues. Additionally, customers need bare essential data on territory, sensible data representation, and data semantics.

Common User Interface

For this situation, the client is supplied with a typical client interface (e.g., a web program) that gives a uniform look and feel. Information from pertinent data frameworks is still independently exhibited so that homogenization and joining of information yet must be finished by the clients.

Integration by Applications

This methodology utilizes joining applications that entrance different information sources and return incorporated results to the client. This arrangement is useful for a little number of segment frameworks. Be that as it may, [5] applications turn out to be progressively quick as the quantity of framework interfaces and information organizations to homogenize and incorporate develops.

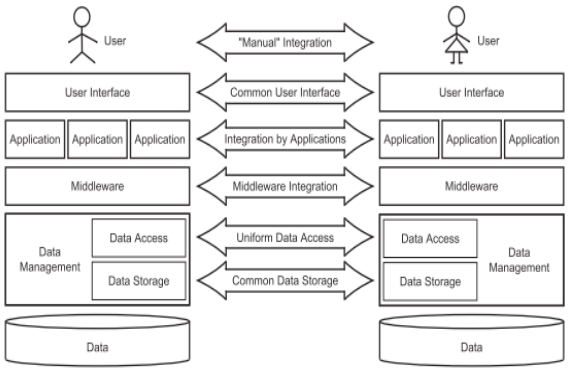


Figure 1: Architectural levels of Data Integration Techniques

Integration by Middleware

Middleware gives reusable usefulness that is for the most part used to settle devoted parts of the incorporation issue, e.g., as done by SQL-middleware. While applications are calmed from executing regular reconciliation usefulness, mix endeavors are still required in applications. Moreover, distinctive middleware devices ordinarily must be

consolidated to manufacture coordinated frameworks.

Uniform Data Access

For this situation, a coherent combination of information is expert at the information access level. Worldwide applications are furnished with a bound together worldwide perspective of physically circulated information, however just virtual information is accessible on this level. Neighborhood data frameworks keep their self-governance and can bolster extra information access layers for different applications. Be that as it may, worldwide procurement of physically coordinated information can be tedious since information access, homogenization, and combination must be done at runtime.

Common Data Storage

In this technique, physical information incorporation is performed by exchanging information to another information stockpiling; neighborhood sources can either be resigned or stay operational. As a rule, physical information coordination gives quick information access. In any case, if nearby information sources are resigned, applications that entrance them must be relocated to the new information stockpiling also. If nearby information sources stay operational, periodical reviving of the regular information stockpiling should be considered.

BIG DATA INTEGRATION

In the time of enormous information certain result of our ability to consider, accumulate and store propelled learning on an unequal scale, and our organized needs to research and think cost from this data in settling on data driven choices to change all parts of society. From a generous variety of fields, the data is being accumulated. Huge Data Integration is exceptionally astonishing from the ordinary data blend since it joins the virtual mix and the rose data dissemination focus. In BDI there are different data sources which contain considerable measure of data and steady availability of data in perspective of component data sources [8].

Present days enormous information reconciliation advancing and it contrasts from conventional information joining. The Big Data time frame is the certain result of our ability to create, assemble and store propelled information at a remarkable scale, and our comparing desiring to separate and separate worth from this information in settling on information driven decisions to alter all parts of society. Huge Data goes with an extensive measure of sureties. The information is being assembled today in an inconceivable collection of spaces. Representations fuse Web substance and reports, Web logs, broad scale e-business, casual associations, sensor frameworks, cosmology, genomics, remedial records, perception, et cetera. Since the estimation of information impacts when it can be associated and consolidated with other information to make a united representation, integration of big data (BDI) is essential to comprehension the certification of Big Data.

RESEARCH OBJECTIVES

The objective of this research is to handle the data integration, analytics for students of UAF by using big data. To bring data together will be from multiple data sources and providing relevant information a unified view to achieve the goal of user will be the main purpose of the study.

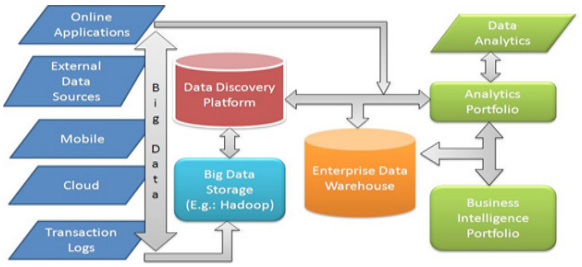


Figure 2: Data Integration Ecosystem for big data and Analytics

PROBLEM SOLUTION USING SPSS AND VISUAL STUDIO

With the immeasurable number of independent data sources accessible on the Internet today, clients have entry to an extensive assortment of data sources. Data integration frameworks are being created to give a uniform interface to many data sources, question the significant sources naturally and rebuild the data from various sources. So, SPSS has an ability to interrogate the data in an organized manner.

Gathering the Data

Gathering a data is itself a huge problem but it's normal for every situation. As per research perspective we took the data of students from different department and analyzes that issues of integration occurred. Every year approximately 5000 or more students in hostel and 3000 or more day scholars took admission in this university. Maintenance of this huge record is still done in register system and somehow use the computer to store that record in the result of this, it is impossible for university administration to remove the problem of heterogeneity and data quality. Probably this university has ability to bring out all the capabilities of students and at every semester university award a merit scholarship to student who deserve. Maintaining the record at one place without heterogeneity we made a small application in Visual Studio that every department will save their student record with their picture as well the result of this we can easily find the student with their registration number. Numerous organizations that create business programming applications have perceived the need to permit their customers to tweak their applications, as indicated by client inclinations or by expanding the current components. These abilities enhance the possibility of the application being acknowledged and received by various classifications of end clients.

Microsoft Visual Studio .NET

VS.NET is a blend of improvement instruments for building distinctive sorts of complex programming arrangements: desktop applications, ASP web applications, XML web benefits, and even versatile applications. VS.NET bolsters four principle programming dialects: Visual C++ .NET, Visual C# .NET, Visual Basic .NET [9], and Visual J# .NET, all of which utilize the same IDE and influence the usefulness of the .NET system. Additionally, other outsider dialects, (for example, COBOL, Eiffel, FORTRAN, and so on.) might be added to manufacture .NET arrangements. .NET system is a Common Language Infrastructure that facilitates building programming applications, including the advancement of Web administrations. An application of student management system can be created utilizing more than one dialect, because Microsoft Intermediate Language

(MSIL), which is a transitional representation supporting all dialects.

Descriptive Statistics:

After gathering the data our next step to conclude a result from our data and suggesting the institution which one would be the best tool or feature of data integration. As per survey which would works better in this institution, half of all surveyed research use SPSS for analyzing the result that which attribute will be best for an organization. Regression analysis help us to predict the future that how data integration would help the organization. A key driver of upgraded profitability in business and fast monetary progression around the world amid the 20th century was the regular utilization of factual devices in assembling and in addition benefit enterprises.

MATERIALS AND METHODS

It is more vital that approach for the research framework is for improving the quality of data by integrating the data of student. The first approach is ETL of relevant data so, it can therefore isolate student data from the unmistakable arrangements, and change the distinctive outlines to the united graphs as showed by SPSS. Finally, it stacks the planned data. The second approach is about removing the issues of integration behind every integration issues are different using the best suited technique will cover these issues.

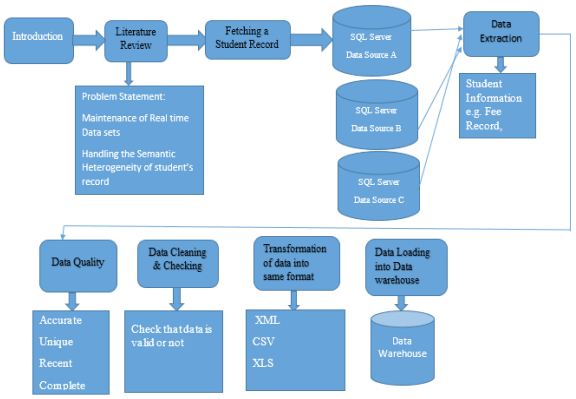


Figure 3: Research Framework

Retrieving Student Data from SQL Server

Working on dummy data of students and then creating a database for student's record in university when he took admission various tables are maintained in management system. To utilize information from a SQL Server database, characterizing a SQL Server information source and one or more report datasets. When you characterize the information source, you should indicate an association string and qualifications with the goal that you can get to the information source from the customer PC.

Data Processing Through ETL

Apparently, to populate a DW, a dimensional information show must be envisioned from some time recently. After that, we must execute a data stream plan called Extraction, Transformation and Loading (ETL) [13], to an excessive step categorically appreciated in the information management query about field. In this procedure, information is picked, removed, changed taking after the dimensionally presented data, and set away into a DW for efficient information examinations. Besides, basic perspective should be highlighted.

Data Quality

Data integration and data quality [10] are interconnected concepts. Data integration always get advantage from data quality, it is instinctive that most information quality issues get to be clear when information in one source are contrasted and comparative information put away in a di□erent source. When they are recognized, there is the requirement for suitable instruments that permit an information coordination framework to play out the question preparing capacity. These procedures are the conflict determination strategies, which assume the significant part of supporting inquiry handling in virtual information incorporation frameworks.

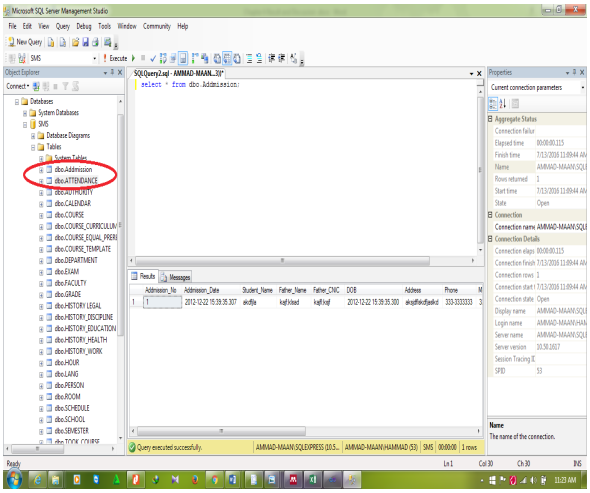


Figure 4: Retrieving Student Data

Data Sets

Responses that we got from respondent were enter in SPSS like this, questionnaire is appropriate way to connect with another user.

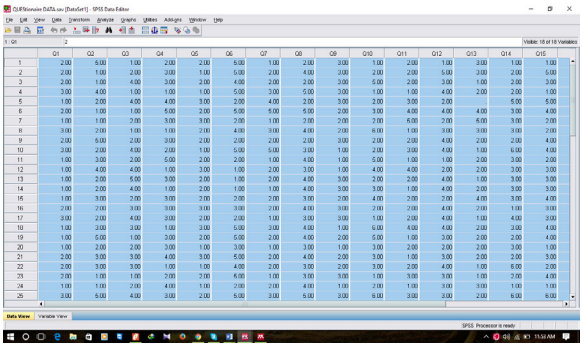


Figure 5: Data Sets in SPSS

Student Management System

The application which is capable to store the data of student along their pictures so it is easy recognize the student.

Challenges of Data integration

Incorporating divergent information has dependably been a troublesome assignment, and given the information blast happening in many associations, this errand is not getting any less demanding. More than 50% of respondents

to our study appraised data integration issues as either a high or high inhibitor to actualizing new applications. The three principle information mix issues recorded by respondents were information quality and security, absence of a business case and deficient financing, a poor data integration framework and metadata management.

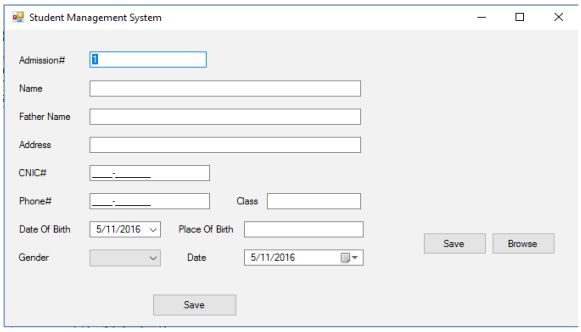


Figure 6: GUI of Application

To what extent, do you agree that data quality and security issues are improved through integration of data?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Strongly Agree	7	17.5	23.3	23.3
	Agree	5	12.5	16.7	40.0
	Uncertain	12	30.0	40.0	80.0
	Disagree	4	10.0	13.3	93.3
	Strongly Disgaree	2	5.0	6.7	100.0
	Total	30	75.0	100.0	
Missing	System	10	25.0		
Total		40	100.0		

Table2: Percentage that found quality and security issues

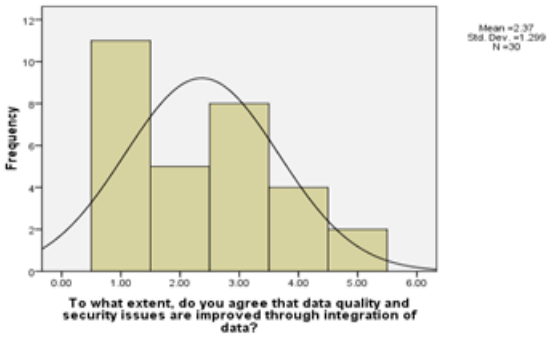
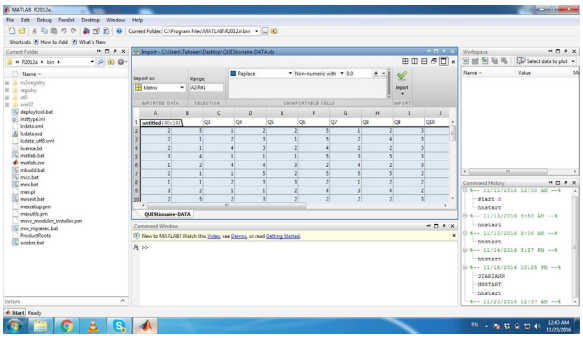


Figure 7: Responses of Respondents for quality issues

RESULTS AND DISCUSSIONS

Through survey, the applications needing access multi-source heterogeneous information regularly require incorporated question handling, compose bound together inquiry expressions for a wide range of information sources through the supplier. Supplier be utilized as a part of heterogeneous information coordination, it tackled the issue that software engineer must utilize distinctive access designs, and even diverse dialects to get to heterogeneous information sources previously. To begin with, the framework must give an effortlessly traversable client interface that gives productive access to huge volumes of test information. Clients must have the capacity to characterize questions adaptably, assigning items and ascribes to recover and limitations under which to recover them. The framework must have the capacity to scale to bolster a lot of time arrangement information, beginning in the several terabytes range.



Analysis of Responses through neural fitting tool (NFT) in MATLAB.

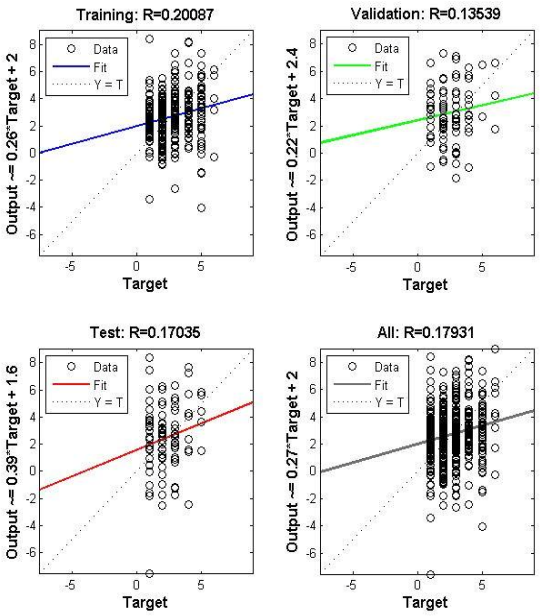


Figure 8: Analysis of Responses using NFT

CONCLUSION AND FUTURE WORK

The student management system which helps the institution in making the best use of resources and integration of data made easy by this application. Secondly, we gave an overview of issues face in field of data integration and important methodologies in the territory of integration seen from a database point of view. Even though data integration is one of the more seasoned research points in the database area. The most troublesome integration issues are data quality and security as per survey. Data quality that can be described through exactness, fulfillment, opportuneness, and consistency of information, is of major enthusiasm for the convenience of integrated data. Improving these issues, we suggested the techniques of data integration would the best. The target of this study was to contrast ascription procedures for managing and missing information in the MATLAB by embracing NFT tool. Future work likewise incorporates investigating continuous information and how semantic ETL can help with its combination. Applications can be worked for different spaces, for example, education

REFERENCES

[1] Kadadi, A., R. Agrawal, C. Nyamful and R. Atiq., "Challenges of data integration and interoperability in big data", In *Big Data (Big Data), International Conference on IEEE*, **1**(1): 38-40(2014).

[2] Walter, T., "Teradata past, present, and future", *UCI ISG lecture series on scalable data management* **1**(1): 44-48(2009).

[3] DeWitt, D., and Gray, J., "Parallel database systems: the future of high performance database systems", *Communications of the ACM*, **35**(6): 85-98(1992).

[4] Lujan-Mora, S. and M. Palomar., "Reducing inconsistency in integrating data from different sources", *In Database Engineering and Applications, International Symposium on IEEE*, **1**(1): 209-218 (2001).

[5] El-Demerdash, K. K. and F. Amer., "Data integration framework with an application for Ministry of Interior", *In Informatics and Systems (INFOS), 8th International Conference on IEEE*, **1**(1): 18-28(2012).

[6] Knoblock, C. A. and P. Szekely., "Semantics for Big Data Integration and Analysis", *In the Proceedings of the AAAI Fall Symposium on Semantics for Big Data*, 1(1): 28-31(2013).

[7] Bertino, E. and E. Ferrari., "XML and data integration", Internet Computing, IEEE, 5(6): 75-76(2001).

[8] Dong, X. L. and D. Srivastava., "Big data integration", In Data Engineering (ICDE), 29th International Conference on IEEE, 1(1): 1245-1248(2013).

[9] Yau, S. S., and Y. Yin., "A privacy preserving repository for data integration across data sharing services", IEEE Transactions on Services Computing, 1(3): 130-140(2008).

[10] Peralta V., "Data quality evaluation in data integration systems", 1(1): 10-163(2006).

[11] Ziegler, P. and K. R. Dittrich., "Data integration: problems, approaches and perspectives", In Conceptual Modelling in Information Systems Engineering on Springer Berlin Heidelberg, 1(1): 39-58(2007).

[12] Dittrich, K. R., and D. Jonscher., "All Together Now: Towards Integrating the World's Information Systems", In JISBD ,1(1): 1-7(2000).

[13] Rodzi, N. A. H. M., M. S. Othman., and L. M. Yusuf.,"Significance of data integration and ETL in business intelligence framework for higher education", In 2015 International Conference on Science in Information Technology (ICSITech) on IEEE, 1(1): 181-186(2015).