International Journal of Technical Innovation in Modern Engineering & Science (IJTIMES)
Impact Factor: 5.22 (SJIF-2017), e-ISSN: 2455-2585
Volume 4, Issue 8, August-2018

Development and Implementation of Modern Data Warehouse for Educational Institution

Sarvani. Vinnamala
Assistant Professor, Computer Science Engineering,
Priyadarshini College of Engineering, Sullurpeta, Nellore District, Andhra Pradesh.

Abstract— ***Big Data provides an opportunity to educational Institutions to use their Information Technology resources strategically to improve educational quality and guide students to higher rates of completion, and to improve student persistence and outcomes. Data warehouse tools and technologies can't deal with the load and scientific procedure of data into significant form for top administration. Big data innovation ought to be executed to broaden the current data warehouse arrangements. Institutions officially gather tremendous measures of data so the academic data of institute has been developing fundamentally and turn into a major academic data. These datasets are rich and developing. Institutional managements need tools to create data from the records. The created data is relied upon to support the basic procedure of best level management. This paper investigates how big data innovation could be executed with data warehouse to help basic leadership process. In this system, we propose Hadoop as large data analytic devices to be implemented for data ingestion organizing.***

Keywords— Data Warehouse; Big Data; Academic Institution; Design; Analysis and Testing.

INTRODUCTION

Now a day, the educational institutes have to generate funds for their research and other operational activities as the government funding has been limited to aided institutes. Utilizing a decision support system is a proactive way to use data to manage, operate, and evaluate educational institute in a better way. Depending on the quality and availability of the underlying data, such a system could address a wide range of problems by distilling data from any combination of education records maintenance system.

A data warehouse system gathers heterogeneous data from several sources and integrates them into a single data store [1]. Data warehouses are used for reporting and data analysis and form the core component of business intelligence (BI) [2]. The goal of data warehouses is to help researchers and data analysers perform faster analysis and make better decisions [3]. Data warehousing also makes it possible to do data mining, which is the science of discovering patterns in the data for further decision-making, such as predictions or classifications [4]. Data warehouses often use large-scale (petabyte) data stores to keep archival as well as current data to enable data analysers to find precise patterns based on long-term changes in the data.

Educational institutions are working in a more and more complex and competitive environment. They have to compete with other institutions to answer to national and global economic, political and social changes. Moreover, different stakeholders are expecting higher education institutions to produce right solution in a timely manner to these demands. To overcome this condition, higher education needs to produce the right decisions required for dealing with these rapid changes by analysing vast data sources that have been generated.

Further utilization of the information contained in the data warehouse is the activity of data analysis using certain techniques and methods. There are several algorithms for knowledge data discovery, like classifying, clustering and mining [5]. The data contained in the data warehouse can used as input for the application system for example like a dashboard. With the existence of this dashboard is expected to be a solution for the learning process to monitor the academic condition and then could take the right decision. However, organizations are recognizing that traditional data warehouse technologies are dying to meet new business requirements, especially around streaming data, real-time analytics, and large volumes of unstructured and complex data sets.

To solve this problem, this paper aims to design and implement a modern data warehouse for academic information system to support decision making process. The designed system accommodates Hadoop platform, a powerful analytical tools which is able to produce a graph that displays the student data information statistically. To support parallel and distributed processing of large volumes of data, most solutions involve Hadoop technology. Hadoop is capable to perform analysis of large heterogeneous datasets at unprecedented speeds

As a result, top management will have a dashboard to monitor the existing condition of the academic atmosphere of university. The reporting dashboard itself will cover operational, strategic and analytical dashboard. The operational dashboards will tell us what is happening now, while strategic dashboards will track key performance indicators in academic process. Moreover, analytical dashboards will process data to identify trends.

The main contributions of this paper are as follows: (1) the designed system enables the communication among different platform and datasets, including smart phones, web, and desktop application whether it is structured, semi-structured and unstructured data. 2) The system provides solution to the top level management in order to know the academic condition in their university. 3) The proposed system could be implemented to other university who need a decision support system for big data.

The remaining part of this paper is organized as follows. Section 2 presents the background and the related work. Section 3 presents the design of the system and Section 4 presents the testing of the proposed system. Finally, the conclusions are drawn in Section 5.

MODERN DATA WAREHOUSE

Data warehouse is the combination of concepts and technologies that facilitate organizations to manage and maintain historical data obtained from operational and transactional applications [6]. It helps knowledge workers (executives, managers, analysts) to make quicker and more informed decisions. Data warehouse is a new paradigm in strategic decision making environment. Data warehouse is not a product but an environment in which users can find strategic information [7]. Data warehouse is a place to store information that is devoted to help make decisions [8]. The Data warehouse contains a collection of logical data separate from the operational database and is a summary. Data warehouse allows the integration of various types of data from a variety of applications or systems. This ensures a one-door access mechanism for management to obtain information and analyse it for decision making. Data warehouse has several characteristics [8, 9]: subject-oriented, integrated data, non-volatile, time-variant, and not normalized.

Data warehouse used data modelling technique called dimensional modelling technique. Dimensional modelling is a call-based model that supports high-level query access. Star Schema is a form of dimensional modelling scheme that contains a fact table at its centre and dimensional tables. Fact table contains descriptive attribute that is used for query and foreign key process to connect to dimension table. Decision analysis attributes consist of performance measures, operational metrics, aggregate sizes, and all other metrics needed to analyse organizational performance. Fact table shows what is supported by data warehouse for decision analysis. The dimension table contains attributes that describe the entered data in the fact table.

Extract, Transform, and Load (ETL) is a data integration process that extracts data from outside sources, transforms the data according to business needs, and stores it into data warehouse [7]. The data used in the ETL process can come from a variety of sources including enterprise resource planning (ERP) applications, flat files, and spread sheets.

Data warehouse support decision support system. Decision Support Systems (DSS) is a computer-based system that helps decision makers use the data and models available to solve problems [10]. DSS functions combine the resources of each individual with the ability of the computer to improve the quality of the decision. DSS requires data coming from various sources to solve the problem. Every problem needs to be solved and every opportunity and strategy requires data. Data is the first component of the DSS architecture. The data relate to a state that can be simulated using a model that is the second component of the DSS architecture. Some systems also have knowledge which is the third component of the DSS architecture. The fourth user interacts with the system through a user interface which is the fifth component in the DSS architecture. In building the DSS, it is necessary to plan a mature system accompanied by the preparation and incorporation of components well.

Data warehouse is widely implemented, including in the education industry. It is possible to implement data warehouse for typical university information system [11]. Academic data warehouse supports the decisional and analytical activities regarding the three major components in the university context: didactics, research, and management [12]. Data warehouse has important role in educational data analysis [13].

With the arriving of big data, traditional data warehouse cannot handle large amount of data [14]. In the past, educational data has been gathered mainly through academic information system and traditional assessments. However, it is increasingly being gathered through online educational systems, educational games, simulations and social media now. Huge workload, concurrent users and data volumes require optimization of both logical and physical design. Therefore, data processing must be in parallel. Moreover, traditional data warehouse cannot extract unstructured data that has varying data structure into information. Traditional data warehouse was design with the purpose of integrating structured data from transactional sources that is supported by OLAP-based analysis. It is the opportunity for big data technology to solve the problem. The integration between big data technology such as Hadoop and data warehouse is very important. To support parallel and distributed processing of large volumes of data, most solutions involve Hadoop technology [15, 16]. Hadoop is capable to perform analysis of large heterogeneous datasets at unprecedented speeds.

Table 1 summarizes the characteristics of traditional data warehouse and modern data warehouse, from the several points of views like the purpose, data sources, scope, architecture, technology, and end-user.

Table 1. The characteristic of traditional data warehouse and modern data warehouse

Characteristics	Traditional Data Warehouse	Modern Data Warehouse
Purpose	Treatment of collected data for a specific business area that is integrated, non-volatile and time-varying. It supports decision- making process.	Processing of structured, semi-structured, and unstructured data, from diverse sources and the volume of data exceeds the ability of traditional tools to capture, store, manage and analyse them
Data source	Usually transactional and operational databases.	Various sources and data types (social media, sensors, blog, video, and audio).
Scope	The integrated structured data to support Business Intelligence (BI) and Online Analytical Processing (OLAP).	Analyse and discover knowledge from large volumes of data characterized by the 4Vs (volume, velocity, variety and veracity)
Architecture	Oriented to processes of extraction, transform and load (ETL). Star schema is the appropriate solution for the architecture	The architecture is depends on the problem. There is still no reference architecture or standardized terminology. They are some proprietary and product-oriented architectures from the vendor
Technology	The technology is mature and tested tools in large amount applications, both free and large amount applications, both free and licensed software.	The technology is still growing. Hadoop is one of the open-source software framework used for distributed storage and processing of dataset of big data
End-user	Business analysts or top managements who do not require specific knowledge of technologies or data exploration.	Data scientists with knowledge in technologies, algorithms, mathematics and statistics.

SYSTEM DESIGN

ETL is the main process in traditional data warehouse technology which cannot handle unstructured data. In this system, we need a flexible ETL process which can handle several data quality issues, as for instance duplicated data, inconsistency data, and garbage data. The proposed system can be seen in Fig 1. In the system, there is a combination between Hadoop and RDBMS. Hadoop can enhance RDBMS as data ingestion/staging tool, but also as data management and data presentation platform.

Fig. 1. The proposed system.

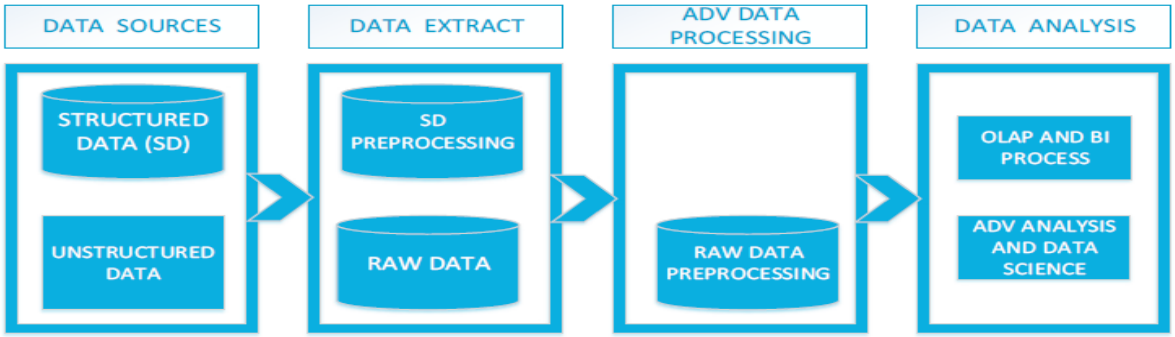

Fig. 2. The architecture of system

As is shown in Fig. 2, the architecture of proposed system was presented. Structured data are aggregated into our schema, while unstructured data is unpredictable data, and usually does not have an easily computer-recognizable format. The examples of unstructured data are free- text, images, videos, webpage's, RSS feeds, Meta data and web server logs. In our proposed system, the unstructured data will be processed by performing categorization and filtering and then it will store in the contextualized data. The uncategorized data will be remaining in the raw data. Next, through a process of searching for relationships or patterns, the data in the contextualized data will store into related data. Then, the related data that are explored data. Finally, it is possible to integrate between explored data and aggregate data to be analysed using OLAP techniques and business intelligence.

IMPLEMENTATION AND TESTING

In this section will be discussed about the implementation of the system in accordance with the analysis and system design. The structured data comes from PostgreSQL databases, while unstructured data comes from social media such as Facebook, twitter and LinkedIn.

Fig. 3 shows the analysis page. In this application, users could create new analysis so the report can be customized as they need. In every analysis, it is possible to produce some graphs or charts to support the generated report. Some advanced users need OLAP Navigator and MDX Query Editor to create powerful report.

The sample chart can be seen in Fig. 4. User can customised the type of the chart, so the generated report will be more meaningful for the reader.

The more advanced interface for analytical feature can be seen in Fig. 5. Advance users could drag and drop item dimensions in the left panel, then put into the column, row or filter in the right panel to produce the insightful report.

Fig. 3. Analysis page of DSS application

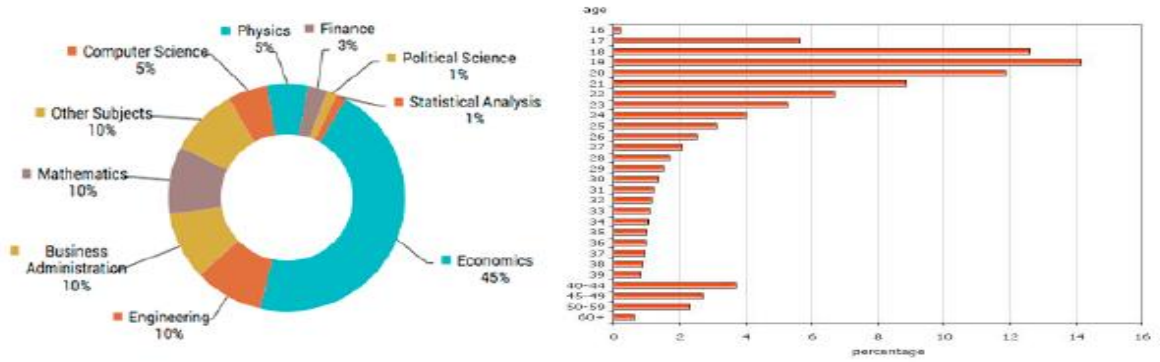

Fig. 4. Chart page of DSS application.

.

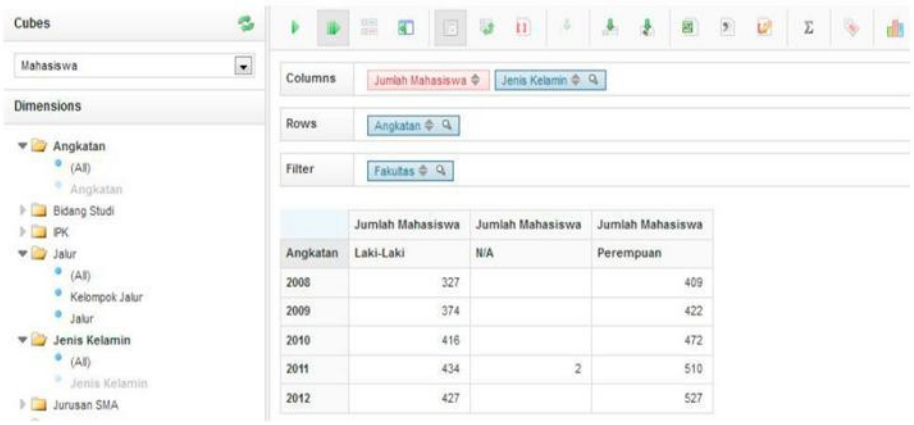

Fig. 5. Advanced analysis of DSS academic

Questionnaires were distributed among thirty university staffs which cover from top management, middle management and bottom management. Rector and vice rectors are categorized as top management. In the middle management, it contains dean, vice dean and their staffs. Head of departments are grouped into bottom management. The assessed indicators include application interface, graphic customization features, ease of use of the application, ability to meet the user needs, and overall application. Detailed assessment of application usage can be seen in Table 2.

Table 2. Assessment of application usage

Indicators	1	2	3	4	5
User Experience	-	-	8	16	10
Graphics customization feature	-	-	4	22	9
Ease of using	-	-	5	18	12
Applications meet the needs	-	-	3	25	7
Overall application	-	-	2	27	5

Description of the rating scale:

Value 1: Very bad

Value 2: Bad

Value 3: Enough

Value 4: Good

Value 5: Very good

CONCLUSION

This paper has investigated modern data warehouse which could substitute conventional data warehouse which can't deal with enormous data in educational system. The big data innovation approach to deal with data warehouse will help reduce challenges related with conventional data analysis. Additionally, this has the capability of enhancing the educational system with new learning ways, and settling on decision making by policy producers more efficient and effective. This paper concludes by laying out future directions relating with the improvement and execution of an institutional undertaking on Big Data.

REFERENCES

M.P. Mathen, (2010) "Data warehouse testing, Infosys Developer IQ Magazine" 1–8.

N. Dedic, C. Stanier, (2017) "An evaluation of the challenges of multilingualism in data warehouse development, in: 18th International Conference on Enterprise Information Systems, Rome, Italy", ISBN: 978-989-758-187-8, pp. 196–206.

V. Rainardi, (2008) "Building a Data Warehouse with Examples in SQL Server", first ed., Apress, ISBN: 1590599314, 9781590599310, pp. 477–489.

M.J. Berry, G. Linoff, (1997) "Data Mining Techniques: For Marketing, Sales, and Customer Support", second, John Wiley & Sons, Inc. ISBN: 978-0-471-17980-1.

Santoso, Leo Willyanto (2011) "Classifier Combination for Telegraphs Restoration." In: International Conference on Uncertainty Reasoning and Knowledge Engineering, Denpasar - Indonesia. DOI: 10.1109/URKE.2011.6007844.

Golfarelli, Matte and Stefano Rizzi. (2009) "A Survey on Temporal Data Warehousing." Int Journal of data Warehousing & Mining 5(1): 1–17.

Ponniah, Paulraj (2010) "Data Warehousing: a Comprehensive Guide for IT Professional." 2nd ed., New York: The McGraw-Hill Companies.

Reddy, G. Satyanarayana, Rallabandi Srinivasu, M. Poorna Rao, and Srikanth R. Rikkula. (2010) "Data Warehousing, Data Mining, OLAP and OLTP Technologies Are Essential Elements to Support Decision-Making Process In Industries". Int. Journal on Comp. Sci. and Eng. 2(9): 2865-2873.

Gour, Vishal, S.S. Sarangdevot, G.S. Tanwar, and A. Sharma. (2010) "Improve Performance of Extract, Transform and Load (ETL) in Data Warehouse". Int. Journal on Comp. Sci. and Eng. 2(3): 786-789.

Turban, Efraim, Jay E. Aronson, Ting-Peng. Liang, and R. Sharda (2004) "Decision Support and Business Intelligence Systems", 10th ed.,New Jersey: Pearson Education, Inc.

Youssef, Bassil (2012) "A Data Warehouse Design for A Typical University Information System", Journal of Comp. Sci. & Research 1(6): 12-17, Dec. 2012.

Dell'Aquila, Carlo, Francesco Di Tria, Ezio Lefons, and Filippo Tangorra. (2007) "An Academic Data Warehouse", Proceedings of the 7th WSEAS Int. Conf. on Applied Informatics and Communications, Athens, Greece, August 24-26.

Mirabedini, Shirin (2014) "The Role of Data warehousing in Educational Data Analysis", Journal of Novel Applied Sciences 3(5): 1439-1445.

Salinas, Sonia Ordonez and Alba C.N. Lemus. (2017) "Data Warehouse and Big Data integration" Int. Journal of Comp. Sci. and Inf. Tech. 9(2): 1-17.

Leo Willyanto Santoso*, Yulia. (2017) "Data Warehouse with Big Data Technology for Higher Education," in 4th Information Systems International Conference 2017, ISICO 2017, 6-8 November 2017, Bali, Indonesia.

J. Nandimath, E. Banerjee, A. Patil, P. Kakade, and S. Vaidya. (2013) "Big data analysis using Apache Hadoop," 2013 IEEE 14th Int. Conf. Inf. Reuse Integr., pp. 700–703, 2013.