Data warehousing with IBM cloud db2 warehouse

Abstract

Nowadays, data warehouse tools and technologies cannot handle the load and analytic process of data into meaningful information for top management. Big data technology should be implemented to extend the existing data warehouse solutions. Universities already collect vast amounts of data so the academic data of university has been growing significantly and become a big academic data. These datasets are rich and growing. University's top-level management needs tools to produce information from the records. The generated information is expected to support the decision-making process of top-level management. This paper explores how big data technology could be implemented with data warehouse to support decision making process. In this framework, we propose Hadoop as big data analytic tools to be implemented for data ingestion/staging. The paper concludes by outlining future directions relating to the development and implementation of an institutional project on Big Data.

1. Introduction

Higher educations 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 analyzing vast data sources that have been generated. Most of higher education institution invest enormous resources in information technology to implement data warehouse system .

The development of data warehouse is a way to extract the important information from the scattered data in some information systems into a centralized integrated storage and support the need for data history. This integrated data can be utilized for information delivery activities that can be reviewed from various dimensions and can be set the level of detail.

Further utilization of the information contained in the data warehouse is the activity of data analysis using certain techniques and methods. There are several algorithm for knowledge data discovery, like classifying, clustering and mining . 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, 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, semistructured 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 present the testing of the proposed system. Finally, the conclusions are drawn in Section 5.

2. Traditional data warehouse and modern data warehouse

This section describes about traditional data warehouse and modern data warehouse. The differences between them are also discussed.

Data warehouse is the combination of concepts and technologies that facilitate organizations to manage and maintain historical data obtained from operational and transactional applications . 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 . Data warehouse is a place to store information that is devoted to help make decisions . 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 analyze it for decision making. Data warehouse has several characteristics : subject-oriented, integrated data, nonvolatile, time-variant, and not normalized.

Data warehouse used data modeling technique called dimensional modeling technique. Dimensional modeling is a call-based model that supports high-level query access. Star Schema is a form of dimensional modeling scheme that contains a fact table at its center 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 analyze 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 . The data used in the ETL process can come from a variety of sources including enterprise resource planning (ERP) applications, flat files, and spreadsheets.

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 . 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 . Academic data warehouse supports the decisional and analytical activities regarding the three major components in the university context: didactics, research, and management . Data warehouse has important role in educational data analysis.

With the arriving of big data, traditional data warehouse cannot handle large amount of data . 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. Hadoop is capable to perform analysis of large heterogeneous datasets at unprecedented speeds.

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

3. 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 enhances RDBMS as data ingestion/staging tool, but also as data management and data presentation platform.

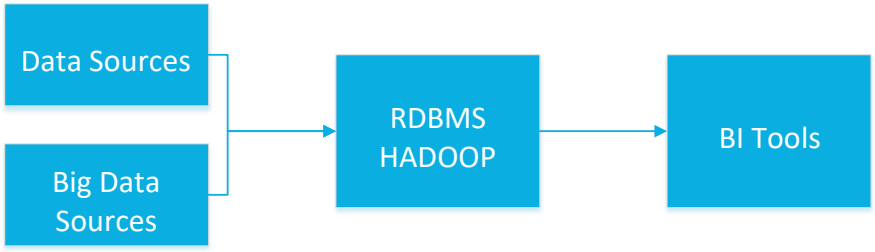

Fig. 1. The proposed system.

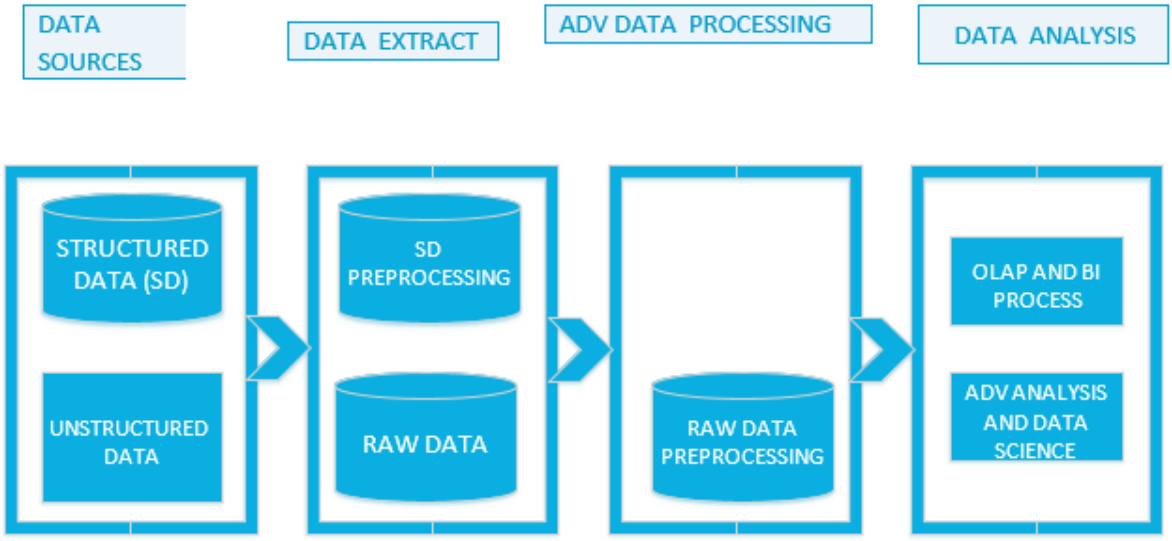

Fig. 2. The architecture of system.