© 2017 IJSRST | Volume 3 | Issue 6 | Print ISSN: 2395-6011 | Online ISSN: 2395-602X
Themed Section: Science and Technology

Quality Assurance in Knowledge Data Warehouse

Sri Haryati[1], Ali Ikhwan[2], Diki Arisandi[3], Fadlina[4], Andysah Putera Utama Siahaan[5]

[1,5]Faculty of Computer Science, Universitas Pembangunan Panca Budi, Medan, Indonesia
[2]Faculty of Science and Technology, Universitas Islam Negeri Sumatera Utara, Medan, Indonesia
[3]Faculty of Engineering, Department of Informatics, Universitas Abdurrab, Pekanbaru, Indonesia
[4]Department of Informatics Management, AMIK STIEKOM, Medan, Indonesia
[2,5]Ph.D. Student of Universiti Malaysia Perlis, Kangar, Malaysia

ABSTRACT

Knowledge discovery is the process of adding knowledge from a large amount of data. The quality of knowledge generated from the process of knowledge discovery greatly affects the results of the decisions obtained. Existing data must be qualified and tested to ensure knowledge discovery processes can produce knowledge or information that is useful and feasible. It deals with strategic decision making for an organization. Combining multiple operational databases and external data create data warehouse. This treatment is very vulnerable to incomplete, inconsistent, and noisy data. Data mining provides a mechanism to clear this deficiency before finally stored in the data warehouse. This research tries to give technique to improve the quality of information in the data warehouse.

Keywords : Data Mining, Knowledge, Data Warehouse

I. INTRODUCTION

Data is the valuable information. The right data produce the accurate information. It will provide useful functions for a client. The daily operational activities of a data warehouse can exchange information as much as infinity. This information will be processed to produce a decision. It is used to determine the strategic steps of an organization. This activity is usually done by large companies [1]. Data from the company's daily operations are stored on an operational database using a relational database architecture. The data must be tested, updated, detailed, and normalized. It is to eliminate duplicate data and speed up data access.

The latest, detailed, and normalized data may not necessarily produce accurate and complete information. For this data to be better, historical data is required. The process of extracting or adding knowledge from a large amount of data is called knowledge discovery in the database. It is a database technology developed to address the business needs of a technology that can store significant amounts of data, process data, and provide data and information, to be used in the analysis and decision-making process. The problem that often happens is how to keep the data is not damaged and remain qualified [2]. Quality needs to be kept so that data is not modified. It also works to keep data free from the threat of crime.

II. Theories

2.1 Data Warehouse

Data warehouse is a collection of technology for decision-making that aims to help corporate knowledge workers to store and organize data systematically. They use the data in the process of analysis and strategic decision making for the company. Data warehouses are built by integrating data from multiple sources, containing duplicate data, different forms of data representation, writing errors, as well as incomplete and inconsistent data. It also produces accurate, consistent, complete, and qualified knowledge [3]. The transformation and incorporation of different data representations, as well as elimination of data duplication, become an important and necessary process. In the data warehouse, the process of loading and refreshing large amounts of data, coming from various operational databases, occurs on an ongoing basis. The probability of data to be stored in the data warehouse contains untested data. Data warehouse is used for the decision-making process, so data accuracy is an important factor to ensure the quality of information and knowledge generated.

IJSRST173658 | Received : 01 August 2017 | Accepted : 06 August 2017 | July-August-2017 [(3) 6: 239-242]

Data warehouses collect and link data collected from multiple sources stored using the same scheme. It provides both historical data and compressed data. Long-term trends and patterns of data can be detected to support analytical and decision-making activities within the company on historical data. Concise data supports efficient access in large quantities and can reduce the size of the database [4][5]. Data warehouse also connects data from several operational databases and data from outside the company. Because incorporated with external data, incomplete data problems, noisy, and often inconsistent. It will result in low-quality data. To improve the efficiency of the process of knowledge discovery required quality data. Data quality can be improved by pre-processing data before it is stored in the data warehouse. Applying the data mining technique is to improve the quality of data to be stored in a data warehouse.

2.2 Data Mining

Data Mining is the processes associated with improving data quality. It is to find the added value of the information during embedded. It is performed manually from a data warehouse by specifying a data pattern. The goal is to manipulate data into more valuable information obtained by extracting and recognizing relevant or interesting patterns of data contained in the database. Data mining provides a mechanism for improving data quality in a data warehouse. Data pre-processing techniques from data mining can be used to improve data quality and efficiency of knowledge discovery process, ultimately can improve the quality of knowledge resulting from the process of knowledge discovery. It determines the quality of data [6].

Data analysis and business intelligence are essential tools for manipulating data for presenting information according to the needs of users with the aim of assisting in the analysis of behavioral observation collections. The definition of data mining can be interpreted as follows:

- The process of discovering patterns from large amounts of stored data.
- The extraction of potential information of data stored in large amounts data.
- Exploration of automated or semi-automatic analysis of large amounts data to find meaningful patterns and rules.

III. Result and Discussion

3.1 Praprocessing

Preprocessing is the initial process of cleaning data before data is extracted to obtain a qualified data pattern. This process can produce valuable knowledge. There are several preprocessing techniques, namely cleaning, integration, transformation, and reduction. Data cleaning is related to erroneous and inconsistent detection and erasing of data. Data integration combines data from multiple sources into a coherent data storage medium, such as a data warehouse. Data transformation includes the transformation or consolidation of data into the proper form. Data reduction is a process for reducing data size, using aggregation functions or summarizing data, selecting data features that can represent data as a whole, eliminating duplicate data, or clustering data. Data cleaning is a major step in efforts to improve data quality. Data cleaning is the process of filling in empty or missing data values, detecting and eliminating errors in data, noise, outliers, and inconsistent data fixing, to improve data quality.

3.2 Incomplete

Several factors cause the incomplete data. They are unavailable attributes required; attributes are not recorded at data entry because they are considered unimportant, data logging function errors, data deleted because they do not match other data, and errors when modifying the data. Some of the techniques used to fill in missing or missing values on incomplete data are as follows:

1. Ignoring incomplete data. This technique is very ineffective, and can only be used on data with some attribute values missing.
2. Change the attribute values manually. This technique can repair incomplete data well, but it takes a long time to fill in missing attribute values and not appropriate for large databases.
3. Automatically populate attribute values.
   - Use global constants. The missing attribute values are filled with global constants such as unknown infinity for numeric attributes. The use

of global constants can lead to a false interpretation of these constants; Computer programs can assume global constants are attributes that have their concepts and meanings.

- Use the average value of all data attributes.
- Use the average value of data attributes in the same group with incomplete data.
- Use a possible value. The most probable values can be estimated using the regression method, the Bayes algorithm, or the decision tree.

3.3 Noisy

Noisy is data that has an incorrect or improper attribute value, generally in the attribute which is the result of the measurement variable. Noise on the common data is the outlier. Outliers are data with very different properties from other information in the same group. Noise on data is caused by errors in instruments or data collection tools, data recording errors due to human factors or computer errors, data transmission errors, and limitations of hardware and database technology. Outliers can be detected by statistical, distance, and deviation-based techniques while the noise can be fixed with smoothing technique.

Some outliers and noise removal techniques are described as follows:

1. Statistical-based. Detection of outliers with a statistical approach using probability model of data distribution. Outliers are identified using a discordancy test, based on data distribution, distribution parameters (average and variance), and some outliers to detect. There are two outlier detection procedures, i.e., block procedures and consecutive procedures. Block procedures assume all data is an outlier, or all data is not an outlier. In consecutive procedures, the first test is performed on the least visible data as an outlier. If the test results state that the data is an outlier, then other information is also expressed as an outlier.
2. Distance-based. Outliers are detected based on the distance between objects or data. Outliers are data that do not have enough neighbors; Neighbor is determined by the distance of data with certain data, which is calculated using the distance function such as Euclidean Distance, Manhattan Distance, and others. The commonly used distance-based outlier detection algorithm is the index-based algorithm and cell-based algorithm. Index-based algorithms use multidimensional index structures, such as tree structures, to search for each data neighbors within a radius d around the data.
3. Deviation-based. This technique detects an outlier by determining the main characteristics of an object or data in a group. Data that is not by the group character is considered an outlier. Sequential exception and OLAP data cube are included in the deviation based detection technique. Sequential exception techniques simulate the way humans distinguish unusual objects from among similar objects. Inequality or dissimilarity is measured at each succession of sections in order. For each subset, the dissimilarity of the subset is determined by the previous subsets. The dissimilarity function is any function that gives a set of objects as inputs and returns a value, high or low. If the objects in a subset have a high degree of similarity, then the function returns a low value; Otherwise, the function returns a high value. Outliers are objects whose values of dissimilarity are high. Outlier detection with OLAP approach uses data cube computing to identify areas containing anomalies in multidimensional data. The process of detecting outliers is done simultaneously with the data cube computing process. A data cube is a form of multidimensional data representation in the data warehouse, which consists of facts and dimensions. Data cube computing is done by using aggregation functions, to summarize data based on dimension levels. The aggregation function is calculated at each multidimensional point, or cell in the data cube, in the form of a dimension-value pair.
4. Binning. The binning method performs smoothing on the ordered data, by dividing the data into groups with the same amount of data. Based on a neighborhood or next value, smoothing can be done by replacing the value of the data attribute with the mean, median, or boundaries of each bin.
5. Regression. It can be done by matching data on the regression function. The regression function is used to find a mathematical equation that matches the data and can soften the noise. Regression functions include linear regression and multiple linear regression. Linear regression includes a straight line search separating two variables, so one variable can be used to predict other variables. Multiple linear regression is the development of linear regression

function, which allows the variable to be modeled as a linear function for multidimensional vectors.

3.4 Inconsistent

In the data warehouse, inconsistency data can occur during the integration process of multiple operational databases, where an attribute can have different names or formats in various databases, table structure differences, and so on. Inconsistency due to the integration of operational database can occur at the data level and scheme level. Inconsistency at the data level can be fixed manually or using knowledge engineering tools. Manually, the data is repaired using external sources, such as performing a transaction log record. Knowledge engineering tools are used to detect errors in data whose constraints are known, for example, functional dependencies between apathetic attributes are used to find the values whose constraint functions conflict. Inconsistency at the schematic level is fixed in the data transformation phase. Data transformation includes the process of smoothing, aggregation, generalization, normalization, and construction attributes. Smoothing is the process of removing noise from data; Is part of the data cleaning phase. Aggregation is the process of summarizing data, using the aggregation function. Examples of data aggregation: daily sales data can be summarized to calculate total monthly or yearly sales. Data generalization replaces data representation at low or primitive levels into high-level representations, using hierarchical concepts.

IV. Conclusion

To improve the efficiency of the knowledge discovery process in the data warehouse and the quality of information generated from the knowledge discovery process requires high-quality data. In fact, the data to be stored in the data warehouse tends to be incomplete, noisy, and inconsistent. Data mining provides a preprocessing mechanism to eliminate dirty data and improve data quality. Data cleaning is the initial phase of the process of preprocessing data, to fill the value of data is empty or lost, detect and eliminate errors in data. Inconsistency can occur at the data level and schema level. Inconsistency at the data level can be improved by tracking transaction records manually or using knowledge engineering tools, such as functional dependencies; While at the schematic level is improved in the data transformation phase, including aggregation, generalization, normalization, and attribute construction. Data quality not only affects the results of knowledge discovery but also affects the process of knowledge discovery itself. To ensure proper decision making the quality of data in the warehouse must be good, that is complete, accurate, and consistent.

V. REFERENCES

[1]. M. J. Berry dan G. Linoff, Data Mining Techniques: For Marketing, Sales, and Customer Support, New York: John Wiley & Sons, Inc, 1997.

[2]. D. T. Larose, Data Mining Methods and Models, Canada: A John Wiley & Sons, Inc, 2006.

[3]. C. D., Discovering Knowledge in Data: An Introduction to Data Mining, Canada: John Wiley & Sons, 2014.

[4]. Zakea Il-Agure; Mr.Hicham Noureddine Itani, “LINK MINING PROCESS,” International Journal of Data Mining & Knowledge Management Process, vol. 7, no. 3, pp. 45-51, 2017.

[5]. M. Mertik dan K. Dahlerup-Petersen, “Data engineering for the electrical quality assurance of the LHC - a preliminary study,” International Journal of Data Mining, Modelling and Management, vol. 9, no. 1, pp. 65 - 78, 2017.

[6]. L. Nunez-Letamendia, J. Pacheco dan S. Casado, “Applying genetic algorithms to Wall Street,” International Journal of Data Mining, Modelling and Management, vol. 3, no. 4, pp. 319 - 340, 2011.