The Development of Data Warehouse to Support Data Mining Technique for Traffic Accident Prediction

Wiwik Budiawan[1,*], *Singgih* Saptadi[1], and *Ary* Arvianto[1]

[1]Department of Industrial Engineering, Diponegoro University, Semarang - Indonesia

Abstract. Traffic accidents are one of the major health problems that cause serious death in the world and ranks 9th in the world. Traffic accidents in Indonesia ranks 5th in the world. One effort to improve traffic safety is to design traffic accident prediction models. Prediction models will utilize accident-related data in traffic through data mining processing. The data warehouse offers benefits as a basis for data mining. Building an effective data warehouse requires knowledge and attention to key issues in database design, data acquisition and processing, as well as data access and security. This study is the first step in the development of data mining accidents based prediction system. The output of this initial stage is the design of data warehouses that can provide periodic and incidental data to the data mining process, especially in the prediction of accidents. The method used to design data warehouse is Entity Relationship Diagram (ERD).

Keywords: **data warehouse,;data mining; accident prediction; ERD**.

1 Introduction

Based on the Article 93 of Government Ordinance No. 43 / 1993 are unexpected and unintentional events involving vehicles with or without other road users resulting in human casualties or property losses. Accidents are one of the causes of death in Indonesia. The number of deaths from traffic accidents in Indonesia reached 28-38 thousand per year. Accident also often creates another traffic jam because of mobilizing the victim and vehicle involved. The road users sometimes are also curious and they stopped a while to see what happened, particularly in Indonesia. The traffic jam will make carbon emission increase since the efficiency of combustion is not optimal condition. Therefore, accident has impacts on not only human safety and equipment losses, but also environment. Predicting traffic accident means preventing the accident from being occurred.

The toll road between Semarang and Solo is a toll road that connects the cities of Semarang, Salatiga and Surakarta. PT Trans Marga Jateng (PT TMJ), Bawen branch manages 2 sections of Semarang-Solo Toll Road namely section 1 (Banyumanik-Ungaran) and section 2 (Ungaran-Bawen). Although toll roads have relatively more ideal conditions than arterial roads in general, accidents still occurred on the toll road. According to PT TMJ, the Banyumanik-Bawen toll road during 2014 there were 75 cases of accidents that occurred with the vehicle flows reaching 600 thousand vehicles each year [1]. Based on this, efforts to improve traffic safety in toll roads need to be carried out.

Factors causing accidents according to [2] consists of three groups, namely human factors, road factors, and vehicle factors. Identifying the factors that cause accidents, the relationship between the occurrence of accidents and its cause can be assessed. Beshah and Hill [3] explained that traffic safety on toll roads can be improved by reducing the factors that cause accidents. Identification of the factors causing the accident and its relationship to the occurrence of this accident can be done by making a prediction model.

Development of prediction models can be done using Data Mining algorithm. Prasetyo [4]; Sowmnya and Ponmuthuramalingam [5]; Beshah and Hill [3] defined data mining as a process to get information from a set of data that helps in decision making. The study in Data Mining consists of Classification, Clustering, Estimation, and Association. Based on this, it is necessary to pay attention to the appropriate Data Mining study area to make accident prediction models.

The quality of the prediction process is determined by the availability of information (related to the cause of the accident) that is reliable, fast, and accurate. Information becomes an asset that is very influential for the success of the accident investigation and prediction process. Therefore, a system support is needed that can help decision makers work well in the investigation and prediction process. Consequently, the decision makers can obtain results that are in accordance with quality information (precise, accurate, and relevant). By using appropriate information technology, a quality information can be produced.

Data warehouse is a form of database that has large-scale data [6]. Data warehouse is not an operational database, but a database that contains data in a certain time dimension that is very useful for the purposes of

* Corresponding author: wiwikbudiawan@ft.undip.ac.id

evaluation, analysis and planning carried out by management in a company. Functions of Data Warehouses among others: report generation, Online Analytical Processing (OLAP), Data Mining, and Executive Information Processes [7]. This study aims to design the data warehouse structure to support the process of predicting accidents in the toll road using the Entity Relationship Diagram (ERD) method. The application of a data warehouse requires a large distribution of data so that the information displayed can be diverse and can form potential information patterns.

2 Methods

Data warehouse is a collection of data obtained from various sources that are used to support the management decision-making process within the company. According to Williams [7], data warehouses have uses, among others

The characteristics of the data warehouse include [8] [6]:

1. Subject oriented. The data warehouse is compiled based on the main subjects of a database (such as customers, products, sales). Each subject's physical area is implemented as a collection of tables related to the data warehouse and not oriented to the specific application process or function. Subject orientation is different from Online Transaction Processing (OLTP).
2. Integrated. Data are taken from many sources to a data warehouse. The data are changed, reformatted, rearranged, summarized, and so on.
3. Time Variant. The data warehouse uses a time stamp to represent historical data. The time dimension is critical to identifying trends, predicting future operations, and managing operating goals.
4. Non Volatile. Unlike records in operational databases which are usually always accessed and manipulated, the data in the data warehouse has different characteristics.

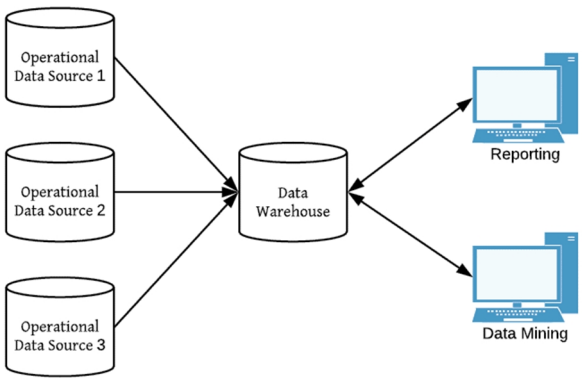

Fig. 1. Data Warehouse Architecture (Connoly & Begg [8])

The data warehouse architecture is a framework designed by understanding how data is moved in the system. The architectural characteristics of the data warehouse are [8].:

a. Data are taken from existing information systems, databases, and files.
b. Data are integrated and transformed before being stored in the data warehouse.
c. Data warehouse is a read-only database created to make decisions.
d. Users access the data warehouse through the front-end tool or application.

Data Warehouse design methodology according to Kimbal [9] there are 9 stages: process selection, grain selection, identification and dimensional adjustment, selecting facts, storing initial calculations in fact table, completing dimension tables, selecting the duration of the database, slowly tracking dimensional changes, and specify query priority and query mode. Whereas Nelson had previously divided the data warehouse design stage into five. The stages in this study will adopt the methodology from Nelson [10]. This is due to simpler method and in accordance with the data availability on the toll road. Nelson [10] described the stages in designing a data warehouse, namely:

1. Understanding the demands of stakeholders
2. Understanding the availability of data sources
3. Designing a data warehouse model
4. Defining data mapping (Source-Target Mapping)
5. Convert Source Mapping Targets to Metadata Targets

The data collection process was carried out with interviews, secondary data and literature study on Data Warehouses. Interviews were carried out with operational parts and parts related to accidents on the toll road. In the same section then the collection of operational data was carried out both related to accidents and those not. This was done to explore various information related to the prediction of vehicle accidents on the toll road. Literature study was carried out through reference searches (books, journals, research reports) relating to the development of a data warehouse.

The next stage was to design a data warehouse to predict vehicle accidents using the Entity Relationship Diagram (ERD) technique. ERD is a technique used to model data requirements of an organization, usually by a system analyst as the requirements on analysis phase of a system development project. While as if diagramming techniques or props provide the basis for the design of a relational database that underlies the information system developed. ERD together with supporting details is a data model which in turn is used as a specification for the database [11].

There are three components in ERD formation, including: Entities, Relationships / Relationships, and Attributes. In the relation component, there is a maximum limit of relationships between entities with each other. Brady and Loonam [11] described the maximum limits of these relationships, including: One to One (1: 1), One to Many (1: M), and Many to Many (M: M).

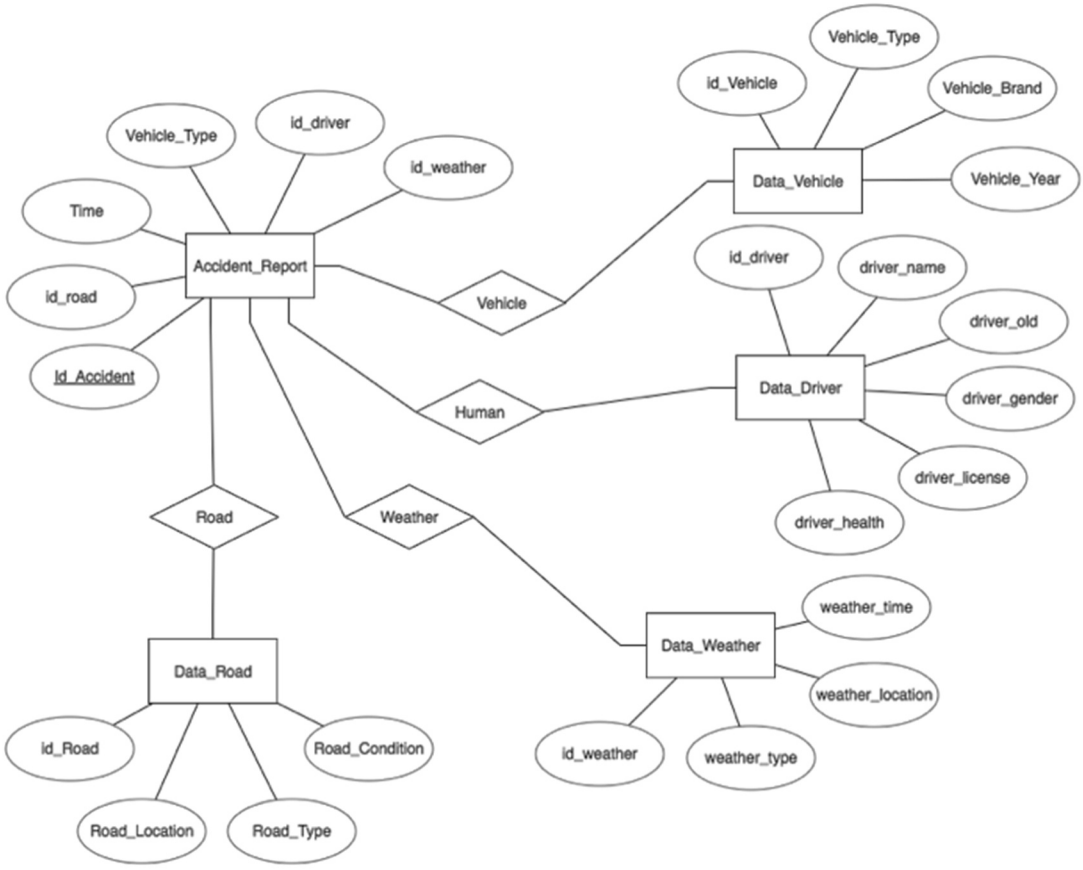

Fig. 2. ERD Data Technique for Traffic Accident Prediction

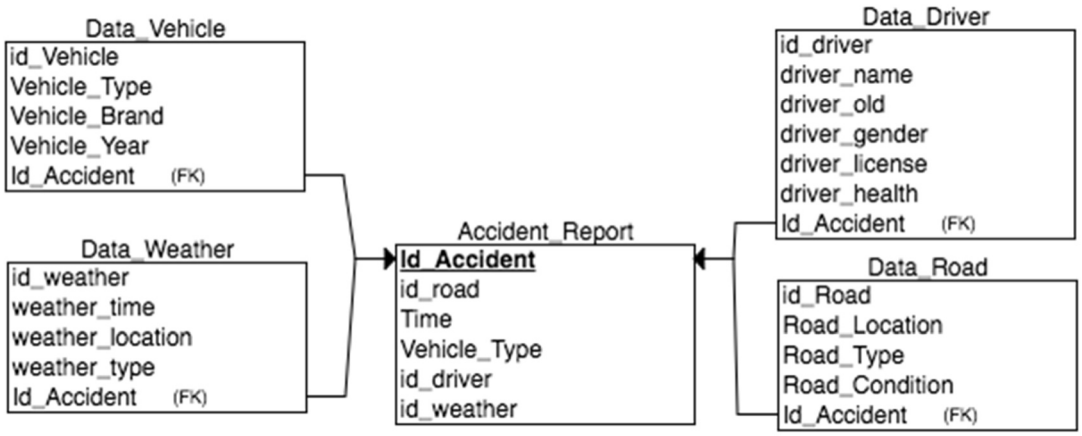

Fig. 3. Relational Table of Data Warehouse

3 Results and Discussion

3.1 Understanding the demands of stakeholders

As the operators of Toll Roads in Semarang City and Semarang Regency, PT Jasa Marga and PT TMJ tried to reduce the number of accidents. The method for reducing the number of accidents is to build an accident prediction model. One of the success factor of the prediction model is the availability of Data Warehouses. Data Warehouse was designed so that it can immediately be used for accident prediction. Therefore, it can be simplified that the demand from stakeholders was the availability of data and in accordance with the prediction model that will be made.

3.2 Understanding Data Source Availability

Factors that cause accidents according to [2] generally consists of three groups, namely human factors, road factors, and vehicle factors. Beshah and Hill [3] suggested that traffic safety on the toll road can be improved by

reducing the factors that cause accidents. Identification of the factors causing the accident and its relationship to the occurrence of this accident was used in determining the data needs in building a Data Warehouse. Operational data in the field includes:

1. Traffic accident reports
2. Weather condition data
3. Road condition data
4. Vehicles type data
5. Driver condition data

3.3 Designing Data Warehouse Model

Connolly & Begg [8] explained that the implementation of a data warehouse can provide benefits for an agency or company. These benefits include: large potential gains in investment, competitive advantage, and increased productivity of decision making. In the process of designing a data warehouse, one of the theories that can be used is a star scheme. The star scheme consists of a central data table and is connected to one or more dimension data tables. Star schemes become the standard for designing a data warehouse [8].

Based on operational data in the field, then the data were connected in ERD. The resulting ERD was then converted to Relational Scheme (RS).

The ERD and RS results showed the data needed in the accident prediction process. Data requirement in the accident prediction process was supported by four dimension data (Weather Conditions, Road Conditions, Vehicle Type, and Driver Conditions).

4 Conclusions

Based on the design, four dimension data can be prepared to support the accident prediction process. These four dimension data merge into one centralized data as Accident Report Data. The results of this study are the initial stages in building the Toll Road Accident Prediction Model. The use of ERD and RS is very helpful for developers in building a Data Base System. Then, from ERD and RS, it is easy to convert into Data Base System form.

References

1. TMJ, “Traffic Report,” 2015.
2. A. Pakgohar, R. S. Tabrizi, M. Khalili, and A. Esmaeli, “The role of human factor in incidence and severity of road crashes based on the CART and LR regression: A data mining approach,” *Procedia Comput. Sci.*, vol. 3, pp. 764–769, 2011.
3. T. Beshah and S. Hill, “Mining Road Traffic Accident Data to Improve Safety: Role of Road-Related Factors on Accident Severity in Ethiopia.,” in *The 2010 AAAI Spring Symposium*, 2010.
4. E. Prasetyo, *Data Mining Concept and Application with Matlab (in Bahasa)*. Andy Offset, 2012.
5. M. Sowmya and P. Ponmuthuramalingam, “Analyzing the Road Traffic and Accidents with Classification Techniques,” *Int. J. Comput. Trends Technol.*, vol. 5, no. 4, 2013.
6. W. H. Inmon, *Building the Data Warehouse*. Wiley, 2002.
7. G. Williams, M. Hegland, and S. Roberts, “Data Mining,” Australia, 1998.
8. T. Connolly and C. Begg, *Database Systems :A Practical Approach to Design, implementation and management*. England: Addision Wesley, 2002.
9. R. Kimball and M. Ross, *The Kimball Group Reader: Rentlessly Practical Tools for Data Warehousing and Business Intelligence*. Wiley, 2010.
10. R. R. Nelson, P. A. Todd, and B. H. Wixom, “Antecedents of Information and System Quality: An Empirical Examination Within the Context of Data Warehousing,” *J. Manag. Inf. Syst.*, vol. 21, no. 4, pp. 199–235, 2014.
11. M. Brady and J. Loonam, *Exploring The Use of Entity-relationship Diagraming as a Technique to Support Grounded Theory Inquiry*. Emerald Group, 2010.