Improving The Data Management: ETL Implementation On Data Warehouse At Indonesian Vehicle Insurance Industry

Michael Abhinaya Bagioywono¹, Jansen Wiratama²*

¹, ²Information Systems Department, Universitas Multimedia Nusantara, Tangerang, Banten, 15810, Indonesia
*Corresponding Author:
Email: jansen.wiratama@umn.ac.id

Abstract

Unknowingly, risk is an essential part of an individual's life. Every day individuals have the potential to be threatened by possibilities that can produce something that has fatal social, human, and financial consequences. Insurance can help individuals to relieve the financial burden caused by unwanted things by transferring individual losses to insurance companies. This transfer of losses will distinguish individuals from possible bankruptcy and financial security. Automotive insurance is a liability for loss or damage to motorized vehicles. This time we will take the example of an insurance company specializing in automotive insurance, namely Top Gear Insurance (TGI). This company uses customer data storage using an ordinary manual database that does not yet use a data warehouse system. This ordinary manual database makes it difficult for TGI to retrieve data for reprocessing and makes data inaccessible from anywhere, asynchronous, concise, and inefficient. There is a solution to the TGI problem of creating a data warehouse with a star schema approach for storing and processing data. The data warehouse is likely to make the data within the company more accessible, efficient, simple, and understandable so that TGI can develop its business through data analysis from the data it already has. Datawarehouse has many business advantages, such as increasing Business Intelligence, data quality and consistency, saving time, and supporting historical data analysis and queries. The data warehouse consists of datamart, OLTP, OLAP, and Star Schema. Using Mondrian as a visualization showed that TGI can get information about customer data, policies, and claims easily, quickly, and concisely. That can also help TGI create customer profiles and targeted marketing and company evaluation based on the visualization provided.

Keywords: Data Warehouse, ETL, Mondrian, Star Schema, and Vehicle Insurance.

I. INTRODUCTION

Without realizing it, risk is an essential part of an individual's life. Every day individuals have the potential to be threatened by possibilities that can produce something that has fatal social, human, and financial consequences. That gives rise to interrelated human emotions, namely emotion, fear, and hope. Insurance can help individuals to relieve the financial burden caused by unwanted things by transferring individual losses to insurance companies. This transfer of losses will distinguish individuals from possible bankruptcy and financial security [1]. An insurance policy is an agreement between two parties, namely the party providing the insurance and the insured, which makes a partnership in which the insured will pay a certain amount of money as collateral to protect against an uncertain loss, which can mitigate the higher financial consequences of insurance, given [2]. Insurance, in general, is divided into two parts: health insurance and general insurance [3]. General insurance consists of vehicle, home, goods, and electronics insurance. Automotive insurance is a liability for loss or damage to motorized vehicles. Automotive insurance includes two guarantees, namely losses to individual vehicles and liability, namely legal responsibility to bear third-party losses related to vehicles [4].

Top Gear Insurance is an insurance company specializing in automotive insurance. Richard Hammond founded Top Gear Insurance in the early 1990s and is one of society's most trusted automotive insurance companies. This company covers motor vehicle insurance such as cars, motorcycles, buses, and trucks. This company uses customer data storage using an ordinary manual database that still needs to use a data warehouse system. This manual database makes it difficult for Top Gear Insurance to retrieve data for reprocessing and makes data inaccessible from anywhere, asynchronous, concise, and inefficient. To
overcome the problems that exist in Top Gear Insurance, it is needed to create a data warehouse with a star schema approach for storing and processing data. The data warehouse is required to make the company's data more accessible, efficient, simple, and understandable so that Top Gear Insurance can develop its business through data analysis from the data it already has.

II. STUDY OF LITERATURE

a) Insurance

Insurancetrafact or agreement of compensation between two individuals, namely an identity contract, the insurer or the insurer and the party that binds the insured to compensate for the losses contained in the insurance work in a way that the insurer will be willing to cover costs incurred from the insured as long as the insured pays a premium to the insurer and the things covered are things that are still in the contract [5], [6].

b) Vehicle Insurance

Motor vehicle insurance is a contract between an individual and an insurance provider that protects the individual from situations that allow the individual to have a loss in the context of a vehicle, such as an accident or theft. The insurance company will bear all forms of responsibility in the event of an accident or unwanted things on the condition that the individual will pay the premium to the insurance company. Vehicle insurance includes three parts, namely property, damage related to vehicles or objects from individuals, liability related to individual responsibility to make compensation or payments for losses or damage experienced by other people, and medical, which will cover all medical expenses needed when sick caused by accident [7].

c) Data Warehouse

The data warehouse was discovered first by Bill Inmon in 1990 with the information warehouse. The data warehouse is subject-oriented, integrated, time-variant and unchanging data that supports management in making significant decisions within a company or organization [8]. A data warehouse can also collect data integrated with different sources to produce multidimensional data to support corporate decision-making [9]. The creation of a data warehouse can generally be classified into two categories, data-based and needs-based approaches. The data-based approach will analyze the data in the database and convert it into multidimensional data. In contrast, the needs-based approach will analyze organizational needs and create models based on the requirements. Datawarehouse has many business advantages, such as increasing Business Intelligence, data quality and consistency, saving time, and supporting historical data analysis and queries. The data warehouse consists of datamart, OLTP, OLAP, and Star Schema [10].

d) Datamarts

Datamart is part of the data warehouse, which has the contents of a small series of data warehouses that support the needs at the company level, such as departments. Datamart is different from data warehouses because datamart focuses on the needs of company users in a business, does not contain detailed data, and contains less information than data warehouses [11].

e) Online Analytical Processing (OLAP)

E.F. Codd, the inventor of the relational database, first discovered the Online Analytical Processing (OLAP) concept. OLAP technology allows analysts, managers, and executives to access data simultaneously quickly, consistently and efficiently for users [12]. OLAP was created to design and support complex analytical processes to reveal market trends and essential factors in business. Olap has several operations: Drill UP, Drill DOWN, Slicing, and Pivot.

f) Extract, Transform, and Loading (ETL)

In creating a data warehouse, combining all data from different sources is needed using Extraction - Transformation - Loading (ETL). ETL will extract data from different sources and convert the data into a form that suits your needs which will eventually be stored in the data warehouse [13].

http://ijstm.inarah.co.id
The use of ETL is to collect, manipulate and combine relevant data from various company sources, which will produce data that meets the criteria of a data warehouse, namely historical, integrated, encapsulated and static (cannot change) [14].

g) Star Schema

Star schema is a multi-dimensional model that provides an overview of an organizational structure database optimized for use in data warehouses and business intelligence. Star schema was first introduced by Ralph Kimball in 1990 as an efficient schema that can reduce data duplication and make data more aggregated. Data can be filtered in the data warehouse [15]. The star schema consists of 2 tables, namely the fact table, which will store transactional data or data that can be measured, and the dimensional table, which will store the attributes of the data. The fact table is a table at the centre of the entire schema, and the dimension table is a supporting table that surrounds the schema [16].

III. METHODS

Several steps must be carried out in making and designing data warehouses for companies. The first step will be the Extract, Transform, Load (ETL) Extract step to retrieve data from a different database; load is the transformation of data into appropriate data, and load produces data according to the desired output. From the ETL step that has been made, a Star Schema will be created, which contains the fact table and dimension table, and at the final stage, an output will be made in the form of a report from the existing data.

a) Data collection methods

The type of data that will be used for this research is quantitative data; this is because the data we have is primarily numerical data, which has different amounts, but in the dataset used, there is a lot of categorical data that will be used to graph the company's data output. The research will use secondary data.

http://ijstm.inarah.co.id
Secondary data: The data we have comes from a community website that provides free data collection and usage, namely kaggle.com, with 28,000 data using the following link. https://www.kaggle.com/datasets/girishvutukuri/insurance-fraud. The Data obtained from kaggle.com is in comma-separated values (CSV) format, which will be uploaded to phpMyAdmin using the XAMPP application. The total data tables that will be used to create a data warehouse are three tables, namely train_policy, train_demographic, and train_claim.

b) Data Understanding

The data obtained for this research is customer data in the TopGear Insurance company, which consists of customer demographic data, customer claim data, and data from existing policies for each customer.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Type</th>
<th>Data</th>
<th>Section</th>
</tr>
</thead>
<tbody>
<tr>
<td>InsuredGender</td>
<td>String</td>
<td>Male, Female</td>
<td>Data Demographic</td>
</tr>
<tr>
<td>InsuredAge</td>
<td>Int</td>
<td>Numeric</td>
<td></td>
</tr>
<tr>
<td>InsuredOccupation</td>
<td>String</td>
<td>Categorical</td>
<td></td>
</tr>
<tr>
<td>SeverityOfIncident</td>
<td>String</td>
<td>Categorical</td>
<td></td>
</tr>
<tr>
<td>DateOfIncident</td>
<td>Date</td>
<td>Date</td>
<td>Data Claim</td>
</tr>
<tr>
<td>.TypeOfIncident</td>
<td>String</td>
<td>Categorical</td>
<td></td>
</tr>
<tr>
<td>DateOfPolicyCoverage</td>
<td>Date</td>
<td>Date</td>
<td>Policy</td>
</tr>
<tr>
<td>InsuredRelationship</td>
<td>String</td>
<td>Categorical</td>
<td></td>
</tr>
</tbody>
</table>

c) ETL

To create a data warehouse from TopGear Insurance, data cleaning must be carried out so that the data accessed is easier to understand and can be utilized for the company's needs later. To clean and transform data so that it is used to produce graphics for the company, the Extract – Transform – Load (ETL) process must be carried out. The tool that will be used to create an ETL from TopGear Insurance is Pentaho, also known as Pentaho Data Integration Tools (PDI). PDI will be used to create a PDI Transformation to run ETL and a PDI job to run many ETLs simultaneously in a predetermined order.

Fig 3. ETL process from TopGear insurance

Extract is the stage of selecting and retrieving data in a company database, which has many formats, for example, .csv, .xlsx, or accurate live data, which is fetched into the database at regular intervals. At the TopGear Insurance transformation stage, data will be collected using Microsoft CSV Input, taken from individual computer files or files from the PHP Myadmin database on the company server. Transform is the stage where data will be changed into more structured data where the data will be normalized. At this stage, the data will be copied, converted according to needs, and reformatted according to the needs of the company database. The quality of the data in the database will be maintained and maintained so that it is not damaged during the transformation process. Load is the final process where data changed by performing ETL will be loaded into the data warehouse. In PDI transformation, the data will be loaded into XLSX format. From the ETL created, a star schema will be made using PDI transformation, forming a fact table whose results will be used to create a workbench schema.
d) **Schema Workbench**

Schema workbench will be used to create results from star schema designs using UML diagrams. The results are in Figure 4. Created a fact table and dimension table containing train_claim, train_demographic, and train policy. The data taken is stored in the phpMyAdmin database, made with the Xampp application's help.

![Fig 4. Class Diagram of TopGear Insurance](image)

**Schema Workbench Diagram**

e) **Mondrian**

Mondrian will display BI in Online Analytical Processing (OLAP). Mondrian will display a visualization of the query measure carried out in the schema workbench to be communicated to the user in the browser.

Mondrian examples:

- TopGearFuel - Hitung Pembayaran Premi
- TopGearFuel - Visualisasi Status Relasi Customer
- TopGearFuel - Visualisasi keuangan Customer Asuransi
- TopGearFuel - Visualisasi Penyediaan Kesehatan keolahragan
- TopGearFuel - Visualisasi Pembayaran Premi
- DQLab-Foodmart - Cube Penjualan
- DQLab-Foodmart - Cube Penjualan
- PHL_W9 - ProductCube
- PHL-Minimart - Sales Cube
- iPivot pivot table
- iPivot pivot table by XMLA
- iPivot with 4 hierarchies
- iPivot with role ‘California Manager’ set
- iPivot with arrows
- iPivot with colors
- Various queries formatted using the Mondrian tag-library
- Basic interface for ad hoc queries
- XML for Analysis tester

Other links:

- Mondrian home page
- Mondrian project page
- iPivot home page
- iPivot project page

![Fig 5. Results from Modrian's creation](image)

**Fig 5. Results from Modrian's creation**

IV. **RESULTS AND DISCUSSION**

In this study, several processes will be carried out to get the desired final result, namely the star schema. First, ETL will be carried out to select data and produce more structured and understandable data. Then, a star schema will be created using Pentaho, which will later be made as a workbench schema to deliver business intelligence visualizations using Mondrian.

[http://ijstm.inarah.co.id](http://ijstm.inarah.co.id)
a) ETL Process

![Image of ETL Process diagram]

Fig 6. Demographic Distribution Indication of No Fraud

Figure 6 is the process for sorting and filtering rows, which will group indications of customers who do not commit insurance fraud to get money deliberately. First, data will be taken from CSV, which will take two tables from the local database on the computer, namely demographic data and vehicle target data, which contains indications of whether the customer has signs of fraud. Then, the data you want to retrieve will be separated using stream lookup. Then, it will select the values used for the table. After that, row filtering will be carried out for indications of fraud, which will be sorted based on customer ID and saved as a .xlsx output file.

![Image of data table]

Fig 7. The results of the KTR demographic division are not fraudulent

Fig 8. Demographic Distribution of Indications of Fraud

Figure 8 is the process for sorting and filtering rows, which will group indications of customers who commit insurance fraud to get money deliberately. First, data will be taken from CSV, which will take two tables from the local database on the computer, namely demographic data—and vehicle targets, which contain indications of whether the customer has signs of fraud. Then, the data you want to retrieve will be separated using stream lookup. Then, it will select the values used for the table. After that, row filtering will be carried out for indications of fraud, which will be sorted based on customer ID and saved as a .xlsx output file.
**Fig 9. KTR results of the demographic division of fraud**

In Figure 9, PDI jobs will run to carry out transformations from customers who have indications of fraud and those who do not. This job has a demographic and target file check before running the job. If there is no file, it will be written to the log.

**Fig 10. ETL Process For the division of fraud classification**

In Figure 10, PDI jobs will run to carry out transformations from customers who have indications of fraud and those who do not. This job has a demographic and target file check before running the job. If there is no file, it will be written to the log.

**Fig 11. Creating a Star Schema**

Figure 11 is an image of creating a star schema from Top Gear Insurance using Pentaho Data Integration. Star Schema will be made to combine several tables, which will become a dimension table, and the main table will produce a fact table. First, each table combined to create a star schema will be called into Pentaho using the CSV File Input. First, demographic tables will be combined with policies using stream lookup; the data collected is on insurance policy numbers in the form of foreign keys, policy annual premiums, and policy deductibles. Then, it will be combined with the claims table, which is combined with the stream lookup value, to retrieve the Claim ID, date of incident, number of vehicles, and authorities contacted. The results of all these combinations will form the fact table of Top Gear Insurance.

http://ijstm.inarah.co.id
Figure 12 displays the results of the fact table created using Pentaho. In the Schema Workbench, a data warehouse will be created, which will later be used in the model for data visualization. The first thing to do in the workbench is to retrieve data that Etl has done into the TopGearInsurance database; this data is stored in phpMyAdmin, assisted by XAMPP. After that, you can create a multidimensional database. Making Cube has three dimensions; each has a hierarchy containing tables and levels. In the cube section, a dimension usage table will be made, and calculations will be made based on the numerical data in the fact table; this is illustrated in Figure 13.
The previously created Schemaworkbench will be used in Modrian, using MDX queries to generate tables. From the results of the created workbench, there were tables of the results from Mondrian.

<table>
<thead>
<tr>
<th>Measures</th>
<th>All Demographic</th>
<th>Hitung Pembayaran Asuransi</th>
</tr>
</thead>
<tbody>
<tr>
<td>FEMALE</td>
<td>19,516,128.1</td>
<td></td>
</tr>
<tr>
<td>MALE</td>
<td>16,649,319.5</td>
<td></td>
</tr>
<tr>
<td>NA</td>
<td>38,969</td>
<td></td>
</tr>
</tbody>
</table>

**Fig 14. Comparison of Total Premium Payments for Men and Women**

<table>
<thead>
<tr>
<th>Measures</th>
<th>Rata-rata mobil yang kecelakaan</th>
</tr>
</thead>
<tbody>
<tr>
<td>Major Damage</td>
<td>4,041,006</td>
</tr>
<tr>
<td>Minor Damage</td>
<td>5,485,281</td>
</tr>
<tr>
<td>Total Loss</td>
<td>4,335,131</td>
</tr>
<tr>
<td>Trivial Damage</td>
<td>1,339,764</td>
</tr>
</tbody>
</table>

**Fig 15. Total Accident Damage Based on Type**

<table>
<thead>
<tr>
<th>Measures</th>
<th>Perhitungan Keuntungan</th>
</tr>
</thead>
<tbody>
<tr>
<td>FEMALE</td>
<td>364,074,400</td>
</tr>
<tr>
<td>MALE</td>
<td>300,411,600</td>
</tr>
<tr>
<td>NA</td>
<td>561,600</td>
</tr>
</tbody>
</table>

**Fig 16. Company Profit Calculations**

<table>
<thead>
<tr>
<th>Measures</th>
<th>Total Insurance Holders</th>
</tr>
</thead>
<tbody>
<tr>
<td>FEMALE</td>
<td>15,644</td>
</tr>
<tr>
<td>MALE</td>
<td>13,162</td>
</tr>
<tr>
<td>NA</td>
<td>30</td>
</tr>
</tbody>
</table>

**Fig 17. Total Holders / Users of TopGearInsurance services**

<table>
<thead>
<tr>
<th>Measures</th>
<th>Calculation of insurance holder</th>
</tr>
</thead>
<tbody>
<tr>
<td>Husband</td>
<td>5,002</td>
</tr>
<tr>
<td>Not.in-family</td>
<td>5,222</td>
</tr>
<tr>
<td>Other-relative</td>
<td>5,153</td>
</tr>
<tr>
<td>Own.child</td>
<td>5,242</td>
</tr>
<tr>
<td>Unmarried</td>
<td>3,993</td>
</tr>
<tr>
<td>Wife</td>
<td>4,224</td>
</tr>
</tbody>
</table>

**Fig 18. Demographic status of insurance holders**

Figure 14 is a visualization that depicts insurance payments based on premiums obtained from all Top Gear Insurance customers based on male and female gender. Figure 15 shows the damage to the customer's car when the accident occurred. Damage is divided into three parts: Major Damage, Minor Damage, Total Loss, and trivial damage. Figure 16 is a visualization that depicts the total calculation of profits obtained from Top Gear Insurance customers, with a total profit of around 664 million US Dollars. Figure 17 is a visualization of data on total insurance customers at TopGear Insurance with a total of 15,664 and 13,162 female insurance holders, which indicates that insurance users with female gender use Top Gear Insurance more than men. Figure 18 is a visualization of the relationship between Top Gear insurance holders. From the visualization, insurance holders who have children have the highest number, with 5,242.
V. CONCLUSION

The conclusion drawn from this research is that implementing a data warehouse within a company can help companies make decisions by providing essential and understandable information. Using Mondrian as a visualization showed that Top Gear Insurance could easily, quickly, and concisely get information about customer data, policies, and claims. This can help Top Gear Insurance to create customer profiles and targeted marketing and company evaluation based on the visualization provided.

VI. ACKNOWLEDGMENTS

We would like to acknowledge the invaluable support received from the Big Data Laboratory of the Information Systems Department at Universitas Multimedia Nusantara. Their contribution played a significant role in the successful completion of this research. We extend our sincere gratitude to our colleagues from the Information Systems Department, whose insightful input and expertise greatly assisted us throughout the research process. Their valuable contributions helped shape the direction of this study and enhance its overall quality.

REFERENCES


http://ijstm.inarah.co.id


