Data warehouse for analysing music sales on a digital media store

Kiefer Stefano Ranti[1*], Deyanara Tuapattinaya[1], Calvin Chang[1], and Abba Suganda Girsang[1]

Computer Science Department, BINUS Graduate Program - Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia

***kiefer.ranti@binus.ac.id**

Abstract. Nowadays, every company knows that when making a decision that has a potential in affecting their assets, an accurately processed report is necessary in order to support the reasoning behind their decision. Generating a report for stakeholders quickly and accurately is highly required in assisting them making a data-driven decision. By developing a data warehouse, it is possible for a company to do a data-driven decision making to appeal to their customer segments. This paper proposes a data warehouse model design to analyse the sales data contained in the database. The method that was implemented for this particular data warehouse development is the nine-step methodology designed by Kimball. The results are then presented in pdf form and an interactive dashboard.

1. Introduction

Almost every company that exists now is using a database to store their data. They realized that data stored in a database are very valuable, as it can provide an up-to-date and accurate information on how business is doing [1]. The stored data in the database needed to be processed to produce a result of information that is useful for analysis purposes. In processing the data, it is necessary to create a reporting system that is not only fast, but also accurate.

A reporting can be done by creating a long, detailed, and complex SQL Query, but doing so will take too much time compared to developing a data warehouse. A data warehouse is much more efficient than querying because it processes data into information and from information into a report without taking a lot of time reading any unnecessary data [2] [3]. All the data that is processed in a data warehouse will provide relevant information to management and executives in a company to help them make a business decision [4].

In this particular case study, the company already used a database to store their transactional data, but because they did not implement a data warehouse in their system, they are unable the generate a report data automatically and efficiently. Currently, reporting the sales and analysing them is very time-consuming as they need to either export the database into a spreadsheet or make a complex query then analysing them further. Therefore, the development of a data warehouse is going to be very beneficial for this company, as the result will be the data displayed in various reports that will assist the company in making a data-driven decision.

The methodology that was used for the data warehouse development is the 9-step Kimball method. This method is conceived around the 1980s and still being used until now. It has been adopted by many companies and is a mainstream industry standard practice [5].

2. Related Works

2.1. Data Warehouse Concept

A data warehouse is a relational database that was created for a query process that was meant to aid the process of analysis and reporting [6]. It is developed by integrating data from multiple sources that support analytical reporting and decision making. Nowadays data has become the prerequisites in making a decision thus causing the interests in creating analytics reports using a data warehouse increased exponentially [7]. The main purpose of a data warehouse is to assists companies in performing strategic planning and decision making based on long-term data storage to make a quick and accurate decision [8]. Developing a data warehouse requires data cleaning, data integration, and data consolidations [5].

2.2. Data Warehouse Design

The process in designing and developing a data warehouse can be done by first implementing the nine-step Kimball methodology, then using a modelling dimension create a dimensional model which in this case, is a star schema [9] [10]. An Entity relationship diagram (ERD) is used to draw the Online transactional process (OLTP). Then the most important process, ETL (extract, transform, load) are done. This process extract data from the source, transform the extracted data into something useful for the data warehouse requirement and load a function to input data by running a script periodically. After the ETL process, the results can be presented in the form of a report or a dashboard to for easier usability.

3. Proposed Method

3.1. Kimball Methodology

In developing a data warehouse using the Kimball methodology, there are nine steps that need to be followed:

1. **Choosing the process**. The business process that was used as a model to design the data warehouse is this: to buy a song or album from the store, a user should register an account first. After that, the user can log in and proceed to the purchasing process. In this process, the user can choose the items they would like to buy, and then adding it to their cart, where the system will automatically calculate the amount that user should pay. The user can then proceed to checkout, choose their preferred payment type, then complete the transaction by paying. Items that the user just bought can then be downloaded from the library in their account.

2. **Choosing the grain**. In song sales, the analysis covers the number of songs sold and song total sales. The analysis will be done for each country, genre, track, playlist, and period (months, quarter, year).

3. **Identifying and conforming the dimension**. Table 1 shows the relationship between dimensions and grain in a matrix table.

Table 1. Relationship between dimension and grain.

	Field	Description
Time	X	X
Country	X	X
Genre	X	X
Track	X	X
Playlist	X	X

4. **Choosing the fact**. The facts that were chosen in this data warehouse is Invoices Fact. The fact consists of 'CustomerID', 'OrderDate', 'EmployeeID', 'GenreID', 'PlaylistID' dan 'TrackID'.

5. **Storing pre-calculation in the fact table**. Star schema approach was chosen here in order to get the total transaction from various dimensions. Figure 1 shows the star schema of the digital media store.

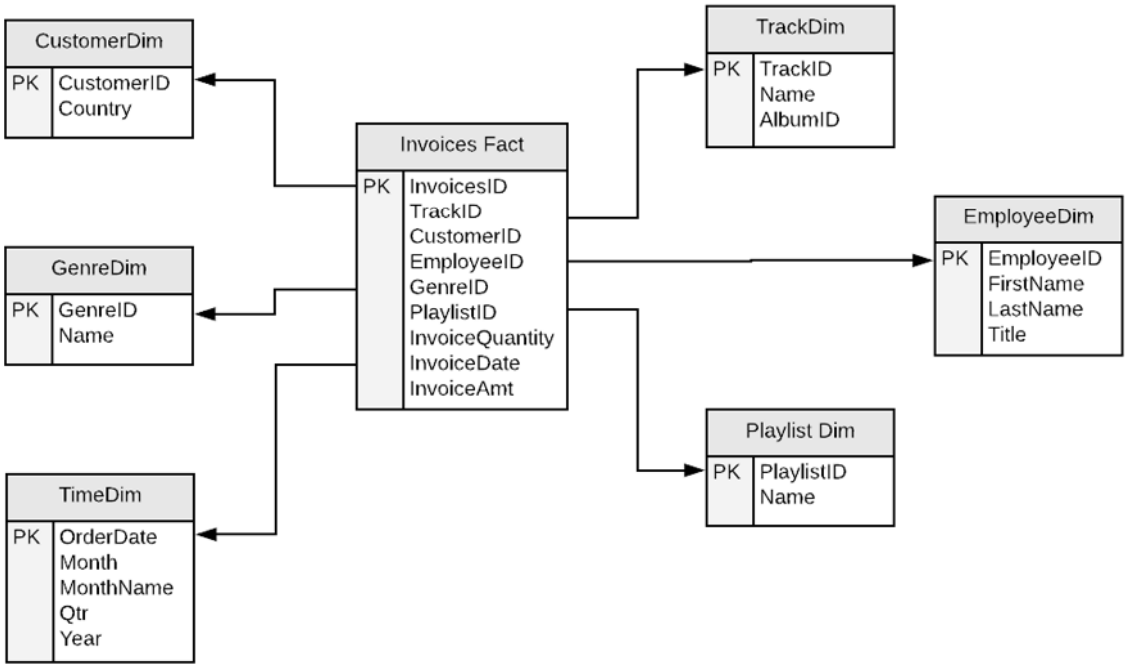

Figure 1. Star schema result.

6. **Rounding out the dimensions table**. The descriptions for each dimension tables are shown in table 2.
7. **Choosing the duration of database**. The data that will be processed into the data warehouse is the sales history data. The duration of the data is from when the database was created, 2009 until 2013, which means 5 years' worth of data is used.

8. **Tracking slowly changing dimension**. In this particular data warehouse design, the slowly changing dimension type is type 1, overwrite. Which means it overwrites old data with new data and does not track any historical data.

Table 2. Dimension tables details.

Dimension	Field	Description
Time	Month, Quarter, Year	The report can be viewed sorted by month, quarter, or year.
Country	Country	The report can be viewed sorted by country
Genre	GenreID	The report can be viewed sorted by genre
Track	TrackID	The report can be viewed sorted by track
Playlist	PlaylistID	The report can be viewed sorted by playlist

9. **Decide the physical design**. This step discusses the ETL process. The ETL process is done every day to ensure the data is up to date to ensure the information that will be received are accurate.

3.2. Extract, Transform, Load Process

This is the most important step in the data warehouse development, where the source data from an online transactional process is integrated into the data warehouse. Pentaho data integration was used in this ETL process [11]. Figure 2 shows the ETL process for every dimension table. Figure 3 shows the ETL process for the fact table, InvoicesFact. Figure 4 shows the successfully executed ETL job.

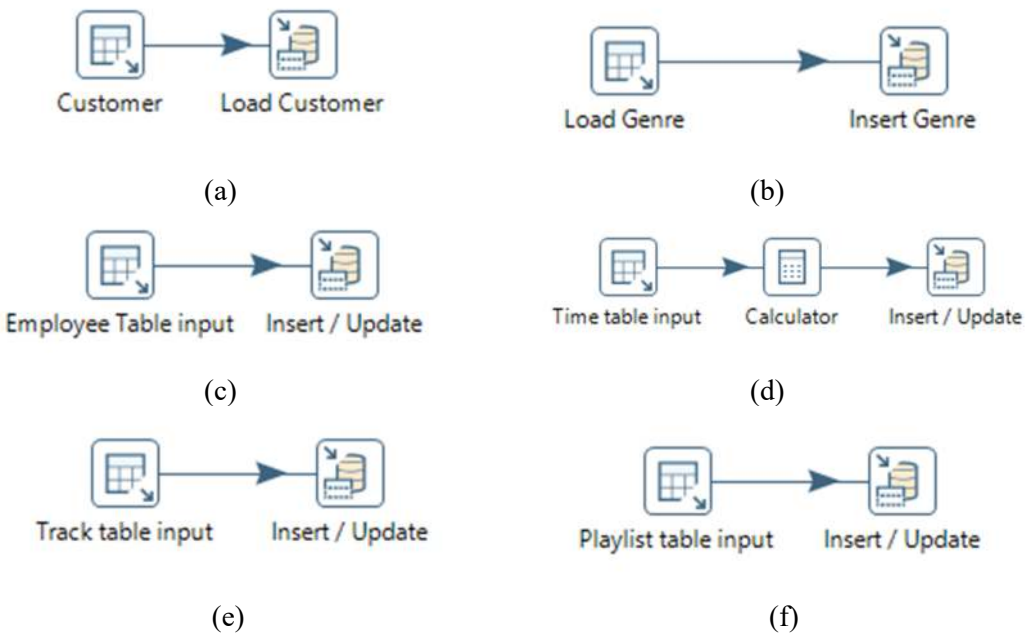

Figure 2. ETL process for dimension table (a) CustomerDim, (b) GenreDim, (c) EmployeeDim, (d) TimeDim, (e) TrackDim, (f) PlaylistDim.

Figure 3. Loading and generating data for InvoicesFact table.

Figure 4. ETL job.

4. Results

After successfully executing the ETL process on Pentaho, utilizing the data warehouse is now possible. Generating reports on the data warehouse is much faster, and the report can be generated based on a various parameter. As an example, table 3 shows how many songs are sold for each genre and table 4 shows how many times a track has been bought.

A dashboard is also made to generate an overall report. In this case, the tool that was used to create the dashboard is Qlik Sense [12]. Figure 5 shows a pie chart of the total customer for each genre, and it is clear that the majority of customer bought a rock genre song. Figure 6 is a bar chart for the customer's country, and the chart shows that most of the customers are from the USA. Footnotes Footnotes should be avoided whenever possible. If required they should be used only for brief notes that do not fit conveniently into the text.

Table 3. Number of songs sold for each genre.

Genre Name to Total Customer Table

Genre Name	Total
Totals	**59**
Alternative	4
Alternative & Punk	50
Blues	23
Bossa Nova	7
Classical	14
Comedy	4

Table 4. Number of track sales.

Track Name to Total Customer Table

Track Name	Total
Totals	**2240**
The Trooper	5
Eruption	4
Hallowed Be Thy Name	4
Sure Know Something	4
The Number Of The Beast	4
Untitled	4

Genre Name to Total Customer Pie Chart

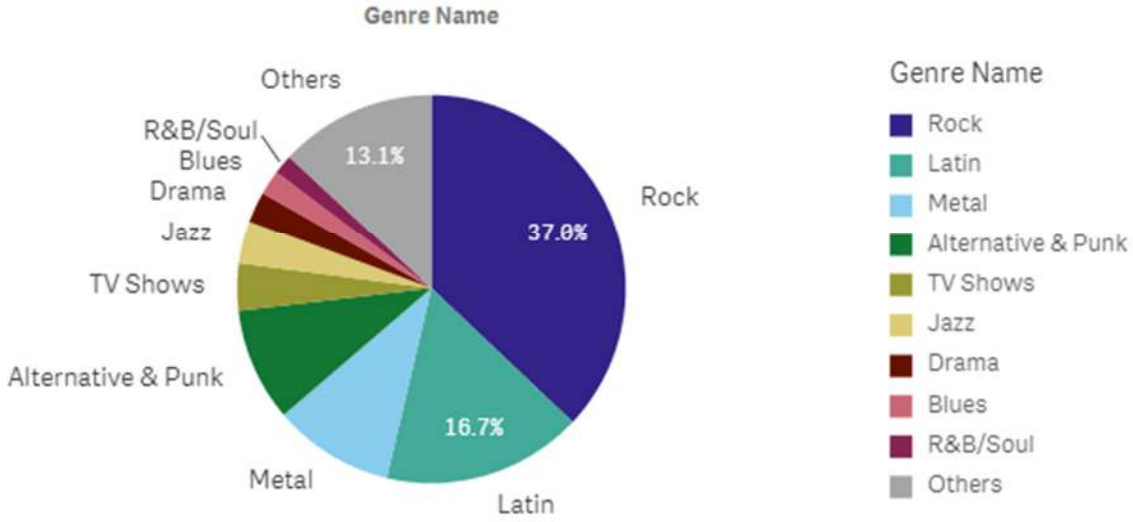

Figure 5. Pie chart of total customers for each genre.

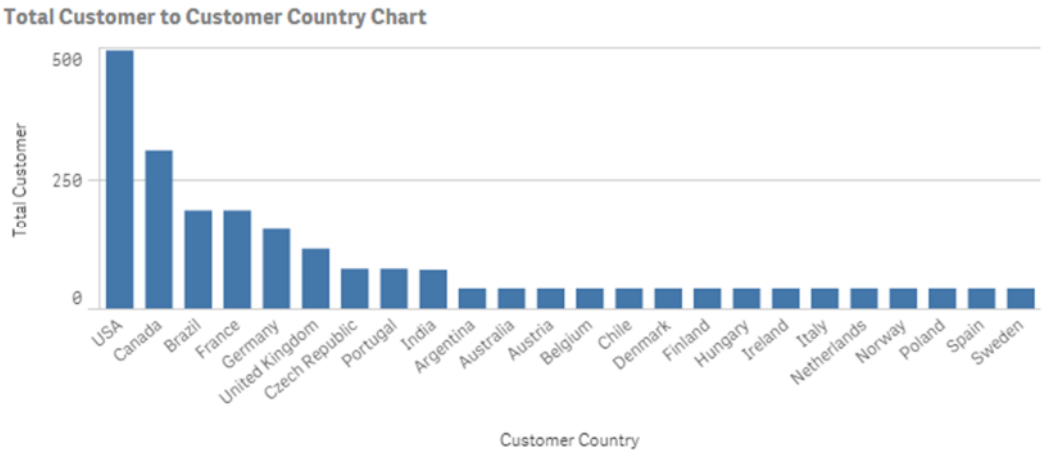

Figure 6. Bar chart of total customers for each country.

5. Conclusions

By developing a data warehouse, generating a report can be done easier and faster than querying SQL manually. Reports can be generated in many forms, such as a dashboard, making it more readable and easier to understand for the management and executives in a company. It can quickly show any sales data and history, customer preferences, and which products is the best-selling one. These reports can help the company management and executives in making a fast and accurate data-driven decision.

References

[1] Andersen, O., Thomsen, C. and Torp, K., 2018. SimpleETL: ETL Processing by Simple Specifications. In *20th International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data co-located with 10th EDBT/ICDT Joint Conference International Workshop on Data Warehousing and OLAP*.

[2] Dedić, N. and Stanier, C., 2016. An evaluation of the challenges of multilingualism in data warehouse development.

[3] Snijders, C., Matzat, U. and Reips, U.D., 2012. " Big Data": big gaps of knowledge in the field of internet science. *International Journal of Internet Science*, *7*(1), pp.1-5.

[4] Kimball, R. and Ross, M., 2011. *The data warehouse toolkit: the complete guide to dimensional modeling*. John Wiley & Sons.

[5] Kimball, R., Ross, M., Becker, B., Mundy, J. and Thornthwaite, W., 2015. *The kimball group reader: Relentlessly practical tools for data warehousing and business intelligence remastered collection*. John Wiley & Sons.

[6] Breslin, M., 2004. Data warehousing battle of the giants. *Business Intelligence Journal*, *7*, pp.6-20.

[7] Dehne, F.K.H.A., Kong, Q., Rau-Chaplin, A., Zaboli, H. and Zhou, R., 2015. Scalable real-time OLAP on cloud architectures. *Journal of Parallel and Distributed Computing*, *79*, pp.31-41.

[8] Bizarro, P. and Madeira, H., 2002, September. Adding a performance-oriented perspective to data warehouse design. In *International Conference on Data Warehousing and Knowledge Discovery* (pp. 232-244). Springer, Berlin, Heidelberg.

[9] Maliappis, M.T. and Kremmydas, D., 2015. An Online Analytical Processing (OLAP) Database for Agricultural Policy Data: a Greek Case Study. In *HAICTA* (pp. 214-225).

[10] Sidi, E., El Merouani, M. and El Amin, A.A., 2016. Star Schema Advantages on Data Warehouse: Using Bitmap Index and Partitioned Fact Tables. *Star*, *134*(13).

[11] Bouman, R. and Van Dongen, J., 2009. *Pentaho solutions: business intelligence and data warehousing with Pentaho and MySQL*. Wiley Publishing.

[12] Ilacqua, C., Cronstrom, H. and Richardson, J., 2015. *Learning Qlik Sense®: The Official Guide*. Packt Publishing Ltd.