Data Warehouse Design for Construction Material Purchase Order Management

Preprint · January 2025

DOI: 10.13140/RG.2.2.20819.28961

6 authors, including:

Bhimaja Goonatillaka
Esoft Metro Campus
6 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Amila Sandaruwan
4 PUBLICATIONS **0** CITATIONS
SEE PROFILE

PPG Dinesh Asanka
University of Kelaniya
99 PUBLICATIONS **135** CITATIONS
SEE PROFILE

Data Warehouse Design for Construction Material Purchase Order Management

W.A.D.B.C. Goonatillaka
Faculty of Graduate Studies and Research Sri Lanka
Institute of Information Technology
Colombo, Sri Lanka
goonatillakac@gmail.com

B.W.G.A. Sandaruwan
Faculty of Graduate Studies and Research Sri Lanka
Institute of Information Technology
Colombo, Sri Lanka
asruwan@gmail.com

W.A.S. Wickramarachchi
Faculty of Graduate Studies and Research Sri Lanka
Institute of Information Technology
Colombo, Sri Lanka
geethwiki@gmail.com

B.A. Munasinghe
Faculty of Graduate Studies and Research Sri Lanka
Institute of Information Technology
Colombo, Sri Lanka
buddhi.munasinghe1995@gmail.com

R M S D Ratnayake
Faculty of Graduate Studies and Research Sri Lanka
Institute of Information Technology
Colombo, Sri Lanka
sathilad@gmail.com

P.P.G. Dinesh Asanka
Faculty of Graduate Studies and Research Sri Lanka
Institute of Information Technology
Colombo, Sri Lanka
dineshasanka@gmail.com

Abstract—A data warehouse is a comprehensive framework that can be used for business analysis and intelligence. The management can perform ad-hoc queries on the warehouse database with ease to fulfil their business specific requirements. This is done with minimum obstructions to the most important operational databases. This phenomenon is applicable in the construction industry as well, where there is a considerable degree of wastage associated with resource management in their construction projects. The three main resources in this industry are categorized as manpower, material, and machinery. In this study Microsoft SQL server data analysis tools are used to facilitate the generation of information to fulfil the functional requirements of a selected construction company in their material purchase order handling. Microsoft Power BI is used in data visualization, which is a tool that provides a unified and scalable platform that can be used to bridge the gap between data and the decision-making process. This prototype proposes a framework to discover the best possible suppliers for a certain domain specific circumstance that is best known to the management of the construction company.

Keywords — Construction data warehouse, MS SQL Server, MS Power BI, Star Schema

I. INTRODUCTION

Construction companies can be categorized into three levels, namely large scale, medium scale, and small scale. One of the main tasks that the management needs to monitor is the resources. In the field of construction manpower, material, and machinery are the three main resources. When the business grows bigger and handles large scale projects the resource management becomes more complicated. The wastage of resources may lead to projects failing to meet their budget and the deadline. As a result, the construction company may lose their business opportunities and target profits. Hence, resource management is one of the most important aspects of construction management. Among many other aspects, it is vital for them to keep track of the best suppliers for a certain material or a group of materials. The best suppliers may be determined by considering multiple factors. The quality of the material supplied, and the cost of that material are two main factors to consider. However, the delays in the purchase order delivery also needs to be considered. If the delivery is too early, the stores and warehouses may not be capable of holding them as expected. On the other hand, when the delivery is delayed, the construction process is also delayed. As a result, the actual work done

may not match with the project work plan. As mentioned earlier, this may lead to a failure in the construction project.

In this study, a construction company dataset is used to design a data warehouse for the purchase order placement of materials. The operational database is developed by a certain selected software house [1], that provides cutting edge project management software solutions to small and medium construction companies. Their operational database offers practical and innovative solutions to the operational business transactions. Currently, this software house does not offer a data warehouse for historical data analysis.

A. *The Need for a Data Warehouse*

The operational databases store and maintain the primary data for daily and short-term operational transactions. Transformed secondary data is required for medium and long-term decision making. For this purpose, organizations use data warehouses to structure data, taken from various formats of operational data like structured, unstructured, and semi-structured. Having organized historical data in one place enables long term data analysis more efficient [2]. All gathered data is stored in one format, which allows the stakeholders to make accurate decisions based on uniform data. Future trends can be identified as data warehouses store a large volume of historical data. This also enables us to evaluate the success or failure of past decisions made. In addition, a complete audit can be performed on the stored data with respect to their sources. Most importantly, security also can be facilitated, allowing only authorized users to gain access to data while blocking others. Data warehouses can store a large amount of historical data that is used to identify future trends, which can be used in decision making [3]. This involves read operations of many records. If the same database is used, it can interfere with the operations of the organization. So, a separate storing repository is required away from the transactional database. This is another reason why a separate data warehouse framework is needed for business intelligence and data analytics. Apart from that, the operational databases are optimized for write operational performance, whereas data warehouses should be optimized for read operations. Therefore, while the operational databases need to be normalized, the data warehouses need to be denormalized.

II. LITERATURE REVIEW

Construction is a field comprising projects with large scales and complex structures. Management of these projects deals with many machineries, labours, materials, finance and equipment. Simultaneously, the enterprise should be concerned about the aspects of schedule, site safety, quality and cost of projects in the complex economic, social and natural atmospheres. All these factors add complexity to the projects. For an enterprise receiving projects synchronously, the ability of management and decision making is a predominant part of the business process[6]. Additionally, the data within a company in the construction field is different from organization to organization. And also, these organizations of construction are comparatively hesitant to accept changes especially in the field of information technology. Many of the national level construction companies operate with the emphasis of refurbishment, development, and maintenance of the contracted locations. Moreover, they operate multi offices and multi sites established in different locations which incorporate requirements and activities of diverse range of stakeholders including suppliers, customers, consultants. Etc. One of the business processes that needs an improvement is the purchasing process. It can be a powerful means of gaining a competitive advantage to the company and raising the overall profitability. The organizations also deal with various data sources.

Data integration is a critical part of many organizational processes. Since it is time consuming and requires an expensive endeavour, the outcome should be well analyzed, organized, and designed. During the process, data warehousing (DW) is a critical component in the data strategy even of a construction company, especially in the process of placing the stock purchase orders. Simply a DW is a relational database that uses multidimensional modelling technique (Star-schema) to organize data from the perspective of multi dimensions against the typical two-dimensional database modeling modelling process. It is essential to follow accurate measures when designing a data warehouse. There are several ways to design a data warehouse. Stock purchase order management is a functional system which generates a huge amount of operational data. Fact and dimension are the two major types of tables used while designing a data warehouse. The fact table is completed with measures of columns and the surrogate keys used to link the dimension tables. Values available in the measure columns are used to measure business fact. These stored values are analyzed with the dimension attributes[7]. Dimensions are the entities aligned according to the purpose of keeping records. Dimensions may exist in the form of snowflakes schema, star schema or fact constellation schema.

A. Non- Functional requirements

These are considered as the requirements with equal importance together with functional requirements. Three of the prominent non-functional requirements are listed as follows.

Integrity- analysts should follow necessary procedures to manipulate the quality of data migrating from the operational systems. Additionally, the controls assuring the consistent data consolidation and presentation should be taken into the consideration. Data cleansing is a data integration technique that satisfies the integration well.

Multidimensionality- one of the aspects unique to the data warehousing system. This is identified as a kind of technique used to model information as facts and dimensions. For the quality enhancement, it needs more than a collection of facts and dimensions. In the multi dimensionality, internal and external factors are accessed under strict quality and time control constraints. Here the integration of row information takes place to derive strategic information.

Performance-This is one of the vital factors influencing the DW design. To the extent that a database showcases a high performance, the DW meets the current and future requirements of the organization. Performance also affects the enormous amount of data that is extracted, transformed, and loaded on to the repository. Since the DW usually deals with the aggregation of huge amounts of data, both the backstage and query performance are essential factors to be considered.[8]

B. 5Vs

The five major and innate characteristics of big data are identified as the five Vs. These five Vs assist data scientists to derive more value from the data and to be customer centric. In the early stage of the century, big data was defined in terms of 3Vs (Velocity, Variety and Volume). With time two additional Vs were added to the list to articulate and communicate the critical characteristics of big data. The list of characteristics are as follows.

Volume: The first out of the 5Vs refers to the amount or the initial size of data collected. If the volume seems to be large enough, that is considered as big data.

Velocity: This refers to the speed that the data is generated and moved. This aspect is demanded by the organizations to accelerate the data flow

from sources such as machines, networks, social media and smartphones to make the decisions in the nearest real time.

Variety: The third aspect refers to the diversity of data types. Data may be directed from multiple sources prevailing within and outside of the organization. This data can be structured, semi-structured and unstructured.

Veracity: The fourth facet of 5Vs refers to the accuracy and quality of data. The collected data may consist of missing, incomplete, or inaccurate data. Veracity defines the trustworthiness of the collected data.

Value: The final V implies what an organization can perform with the available data, merely what is the value that big data can add to the outcome is considered.

III. METHODOLOGY

A data warehouse is a complete framework that aids key people of an organization to perform ad-hoc queries to retrieve much needed information for business intelligence and data analytics. There are two main types of schemas that can be used in a data warehouse. Namely, star schema and snowflake schema. This study is based on the star schema to implement the data warehouse for the selected construction business to facilitate their decision-making process. The star schema design contains duplicated data as it is denormalized. Aggregated data is encouraged to be used to reduce the usage of resources like CPU time and memory during the reporting and data analysis stage.

The data set is received as a MySQL backup file from the construction company, which is in its native structured format. For simplicity, the stock purchase order is selected as the schema and only the related tables are extracted to csv files for further analysis. After removing some irregular records, the data set is imported to Microsoft SQL server platform. There, certain data types are changed to suit the new Platform. This Microsoft SQL server database is used as the source file for the data warehouse.:

1. *DATA WAREHOUSE DESIGN*

A. StockPurchaseOrderFact Table

The fact table consists of surrogate keys and measurements. Apart from that, stockPurchaseOrderId, which is the primary key of the purchase order table of the operational database is also added as an audit column to use in troubleshooting purposes. Measurements like transport cost, VAT, discount, delivery delay, rate and line total are used to analyze data for the business needs. Line total is an aggregated column. The grain or the level of detail in the fact table is individual materials being purchased by the construction company.

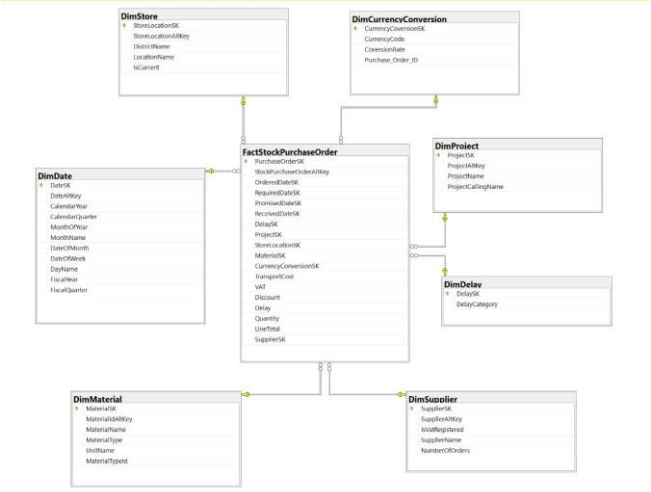

Fig.1. Proposed Design.

B. Dimension Tables

There are seven-dimension tables. Namely, dimDate, dimMaterial, dimProject, dimCurrencyConversion, dimStore, dimSupplier, and dimDelay. All dimension tables are directly linked to the fact table using surrogate keys. These surrogate keys are of integer data type. In all dimension tables all columns are SCD type 1, except for the dimStore table. There, a type 2 slowly changing dimension is used to keep track of

the data when store related data tend to change. These changes are possible, for example when the company changes the location of a store or discontinues use of any of the stores. In dimDate a natural hierarchy is used. It is designed and used as a role-playing dimension as multiple date columns from the fact table are linked to the dimDate. Namely, receivedDateSK, promisedDateSk, and requiredDateSk. dimDelay is used to categorize the delay in delivery to make the analysis more meaningful and efficient. No coded data but meaningful verbose column data is used in dimension tables. The dimension tables are not directly related with each other and hence they are not normalized as it adheres to the star schema design. All dimension tables are linked to the fact table with one-to-many relationships. The delay column of dimDelay dimension is categorized as given in table 1. The delay is calculated by subtracting requiredDate from receivedDate.

Delay (in days)	**Category**
More than 5	OverDelay
Between 1 and 5	Delay
0	OnTime
Between -1 and -5	Early
Less than -5	OverEarly

Table 1. Categorical table of dimDelay

C. Star Schema Implementation.

The data warehouse star schema design is implemented by using a series of steps. See figure 2. This is done after considering the plan and design documented under the above Data Warehouse Design section. As the design is updated several times, a step is added to drop tables if it already exists.

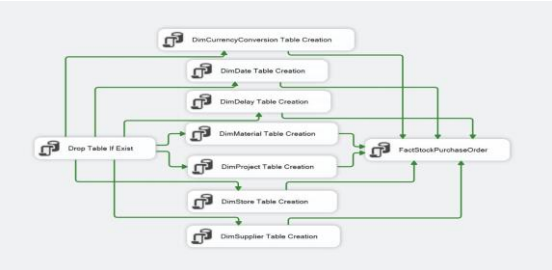

Fig.2. Steps of star schema table creation

D. Extract Transform Load Process.

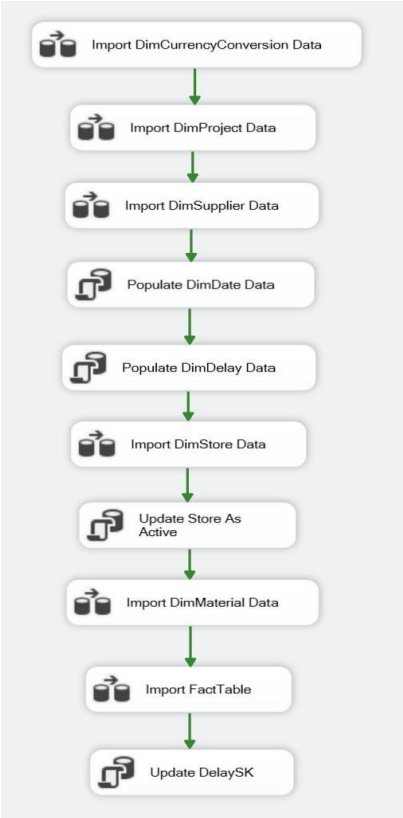

Fig. 3.Application of ETL in summary on all tables

The extract transform Load (ETL) process is first applied to the dimension tables. See fig.3. above. When importing data, the choice of columns is selective, where all irrelevant columns for analytical purposes are ignored. The DimMaterial table is derived from multiple tables of the source database. Namely, material, ma_units, and material_type tables. The supplier primary key from the operational database is used to sort and merge when importing data into the dimSupplier dimension. This is also to import the number of orders registered per supplier as aggregated data. See fig. 4. below.

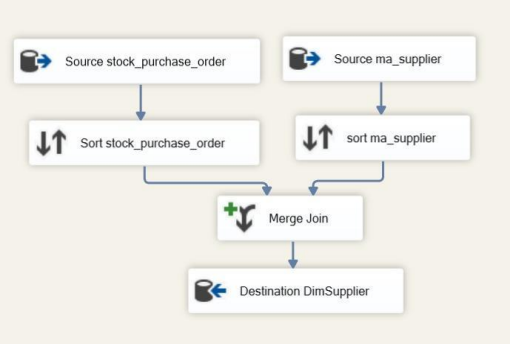

fig. 4. ETL on dimSupplier

After successful ETL on all dimension tables, the ETL is applied to the fact table. See fig. 4. The orderDateSk is derived as an integer value after removing the separators within the date format given in the operational database. The line total is derived by multiplying the Quantity and rate columns. The requiredDateSk, PromisedDateSk, and recievedDateSK columns are also derived following the same technique used for the orderDateSK. All other surrogate keys of the fact table are derived using simple lookups used. For example, to retrieve projectSK, the project_code column from the merged column set is mapped with the projectALTKey of the dimProject table. A list of derived columns are transformed by replacing null values if any with a zero.

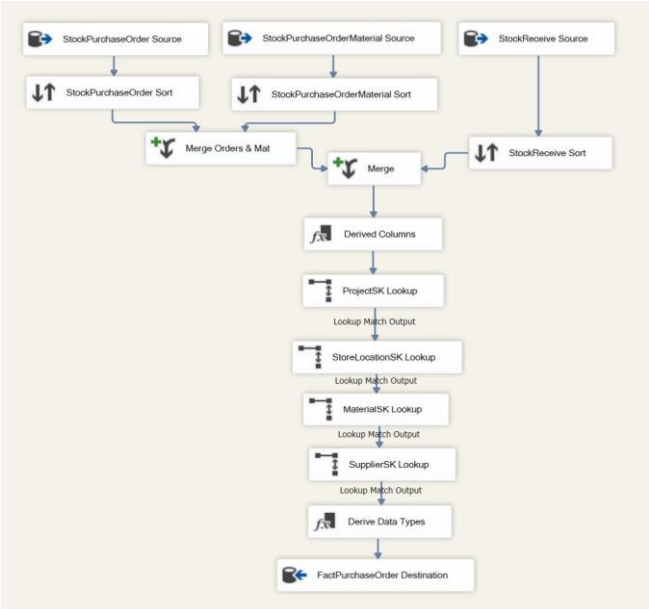

Fig.5 ETL on fact table

IV. ANALYSIS AND REPORTING

Ad-hoc queries can be performed using Microsoft Power BI in par with the on-demand requirements of the management. See fig.6, for a summary report on the projects registered with the company.

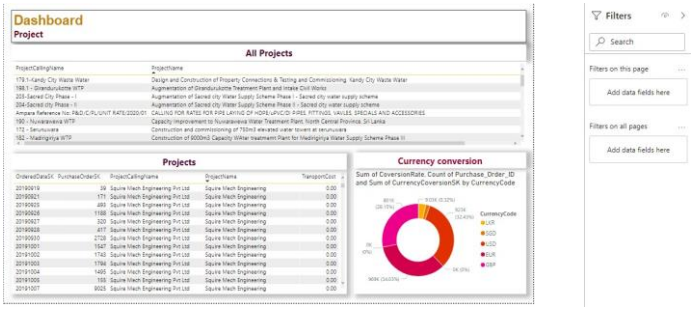

Fig.6.Projects registered with the company.

Supplier related summary can be retrieved to visualize the top suppliers with the highest number of order placements registered with the construction company. See fig.7. below.

Fig.7. Top suppliers.

A summary report on the suppliers supplying a certain selected material can be retrieved together with the store location where those materials are received. The best supplier for a certain material can also be revealed based on the past data. See fig. 8.

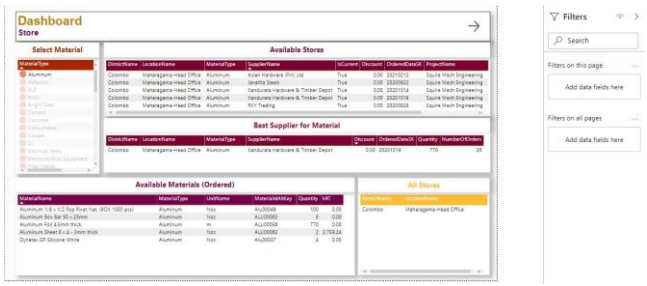

Fig.8. The best supplier for a selected material.

Further Improvements

The purchase order returns can also be considered as an improvement to the current data warehouse design, which can reveal a better mechanism to figure out the efficiency of suppliers. The data warehouse design can be improved to incorporate currency conversion impact on the projects on going. This is somewhat complicated as different currency conversions may be applied on different material purchased under the same purchase order.

As another improvement by analyzing the ordered date and the delivered date of the historical purchase orders, it would be possible to identify the lead time of the supplier and the required safety stock to manage the process until the new stock reach the location.

In addition, weather records can be incorporated into the dimDate dimension, which can be obtained from a reputed weather reporter. This is important to understand how prices of a certain material may be affected, for example, during a rainy season the price of sand may increase. With this a staging area can be used to improve the ETL process as resources may be taken from multiple sources and formats. The ETL process can be further improved, when 'date added' columns can be used to keep track of the changes made to operational records and thereby extract only those changes. This will make the ETL process more efficient.

V. CONCLUSION.

The development of technology has affected almost every business field in many ways. It has eased our daily lives and developed the ability to help in making effective and efficient decisions. By analyzing results of the designed data warehouse, the management of a construction company can get a summarized view of the supplier and purchase details to make effective decisions on material purchases related to each construction project which would directly affect the final project cost, achieving the given timelines and the company profit.

With the analysis we were able to identify the suppliers that delivered the stocks to the requested locations on time, later than the promised date or earlier. In the construction field, delivering the stock on the required date is highly important as delivering the stock over early or a delay could affect the quality of materials, storage facilities, daily operations of workers and the project timeline.

The cost of material and material quality is equally essential when considering the purchase orders. Since purchase cost directly affects final project cost and company profit, the cost of the materials was also considered during the analysis of data warehousing so that the management or purchasing team could use the analysis to reduce the material cost and wastage.

The designed data warehousing model will help consumers to get quantitative and transparent information, organized and delivered in a way to make the effective decisions to reduce the material cost and improve the customer satisfaction.

REFERENCES

[1]. S. Ltd, "Sky Management Systems (Pvt) Ltd", Skymanagementsystems.com, 2021. [Online]. Available: https://skymanagementsystems.com/. [Accessed: 10- Nov- 2021].

[2]. "What is a Data Warehouse and Why Do You Need One?", Dew Solutions, 2021.

[Online]. Available: https://www.dewsolutions.in/why need-data-warehouse/. [Accessed: 10- Nov- 2021].

[3]. JanbaskTraining, 2021. [Online]. Available: https://www.janbasktraining.com/blog/data-warehouse. [Accessed: 10- Nov- 2021].

[4]. ZHANG, T. MA and Q. SHEN, "Application of Data Warehouse and Data Mining in Construction Management", 2021, pp. 1-12.

[6]. Kimball, R. and Merz, R. (2000), "The Data Webhouse Toolkit: Building the Web-enabled Data Warehouse", Industrial Management & Data Systems, Vol. 100 No. 8, pp. 406-408. https://doi.org/10.1108/imds.2000.100.8.406.1

[7]. "Enhancing Data Warehouse Design with the NFR Framework", CEP 50740-540, Recife, PE, Brazil, 2021.

[8]. "The 5 V's of Big Data", *SearchDataManagement*, 2021. [Online]. Available: https://searchdatamanagement.techtarget.com/definition/5-Vs-of-big-data?amp=1. [Accessed: 20- Nov- 2021].