

2014801_Dylan

by Dylan Thomas

Submission date: 22-Mar-2024 02:38PM (UTC+0000)

Submission ID: 227602075

File name: 24885_Dylan_Thomas_2014801_Dylan_565908_263485620.docx (4.39M)

Word count: 4027

Character count: 23496

Assignment 1: Data Warehousing

Author: Dylan Thomas 2014801

Data Warehousing and Data Mining (ACCA6026)

School of Applied Computing

Wales Institute of Science and Art

22/03/24 Date of Submission

Contents

1. Introduction	3
2. Task 1	3
2.1 OLTP and Data Warehousing	3
2.2 Warehouse Architecture and ETL	4
2.3 Data Marts	5
2.3.1 Managing Director	5
2.3.2 Finance Director	6
2.3.3 Sales and Marketing Director	6
2.4 Dimensional vs Normalised	6
2.4.1 Dimensional Database	6
2.4.2 Normalised Database	7
3. Task 2	8
3.1 Source Systems	8
3.2 System Schema	9
3.3 Data Marts	11
3.4 ETL Plan	11
3.4.1 Extraction	11
3.4.2 Transformation	12
3.4.3 Loading	12
3.5 Data Quality and Governance	13
4. Conclusion	13
5. Reference List	14

1. Introduction

Data Management is imperative for success, regardless of sector. This applies to Cambrian Real Estate, where the implementation of a robust data warehousing solution is needed to meet the dynamic landscape of real estate. This report will look to provide clarity on effective solutions for the executive team on the queries posed, aiming to streamline operations and improve the decision-making processes.

An Estate Agents like Cambrian deals with a variety of data including financial and PII data which needs to be considered within the solution discussions. The proposed solution of a data warehouse aims to organise and consolidate data into a centralised repository. This report will delve into the intricacies of designing and implementing a data warehousing solution. The solution should take into consideration a variety of factors like data architecture, ETL processes, data modelling, industry standard compliance. All these factors should come together to provide Cambrian with a data-driven solution providing data intelligence and the adaptability to meet changing demands.

2. Task 1

2.1 OLTP and Data Warehousing

According to Oracle, OLTP (Online Transactional Processing) is *"a type of data processing that consists of executing a number of transactions occurring concurrently"*. [1]

The OLTP databases are optimised for the storing of normalised transactional data within a RDBMS (Relational Database Management System). OLTP Databases are typically prioritised for transactional processing and individual record access. OLTP systems follow a normalised data model that's optimised for data storage and integrity given the minimal redundancy.

According to IBM, a Data Warehouse *"pulls together data from many different sources into a single data repository for analytics and decision support"*. [2]

On the other hand, data warehouses prioritise effective querying and analytical processing over data redundancy. In a data warehouse, the data is sorted into star schemas utilising facts and dimensions. Dimensions store the context and background of transactions, such as strings or character fields like market regions, while the facts store numerical data, such as dates or costs as defined by *Kimball* [3]. Data Warehouses are usually used by analysts and BI tools to analyse the data for decision making purposes.

When it comes to reporting output, implementing a Data Warehouse has the following advantages over an OLTP solution:

Data Quality and Governance: Cleansing and checking the data is necessary before transformation using ETL or ELT process. This allows improved governance and greatly increases the quality of the data. Data dictionaries including important metadata should be present in every database. This allows the tracking of personally identifiable fields (PII), the tracing of data lineage, and logging of data sources.

Enhanced BI + Historical Data: Analysis and a coherent picture of the business can be produced by gathering and combining data from different sources. An OLTP system by itself would not be able to achieve this. This provides a unified view of the data, with the integration simplifying analysis and

reporting. As the Warehouse can store a vast amount of data efficiently, storing of historical data allows trend analysis and forecasting activities to be undertaken.

Optimized Performance and Efficiency: The Data Warehouse can handle complex queries, which not only increases query performance and speed but decreases the strain on the OLTP system. Additionally, real-time data handling is not required of data warehouses, reducing the impact on operating systems.

2.2 Warehouse Architecture and ETL

ETL is the data processing method that links all the databases in a data warehouse, which is a singular repository for a collection of databases. The acronym ETL stands for extract, transform and load. The data landscape has changed in recent years due to the emergence of new technologies, which also resulted in the transformation and load processes being replaced by ETL.

Extract: Data is processed in a data warehouse using three steps. The first step being data extraction from identified source systems, which could be from an external system or one of the internal OLTP platforms. The data is subsequently transferred to a staging server or database, based on the volume of data for processing. The data is loaded into a staging area so that the OLTP server won't have to be used, thus improving system performance, in the event of a full data refresh. The staging area houses the data, which is identical to that of the source system.

Transform: After Extraction, the data can be combined and loaded into the data warehouse. Depending on the business requirements, the data is changed. It is possible to gather aggregates, combine data from other sources to provide business insights, and to provide operational or strategic views of the data.

Load: Data marts, which are essentially small, focused data clusters that address certain business needs or questions, are utilised to store data. By creating a data mart and linking it to other databases can allow niche reporting and complex forecasting.

Following the ETL process, the data needs to be visualised in a way that's aesthetically pleasing and easily understandable, this usually requires some thought and consideration. Reporting tools like PowerBI and Tableau would be beneficial for providing the reporting. This will facilitate business decision-making for the stakeholders, enabling them to digest the data immediately according to *Segall et al. [4]*

Data warehouses are usually loaded once a day, processing data from the previous day. Real-time or close to real-time reporting is achievable with new cloud technology and stream analytics, but at a great cost and should only be considered if there is a genuine business need.

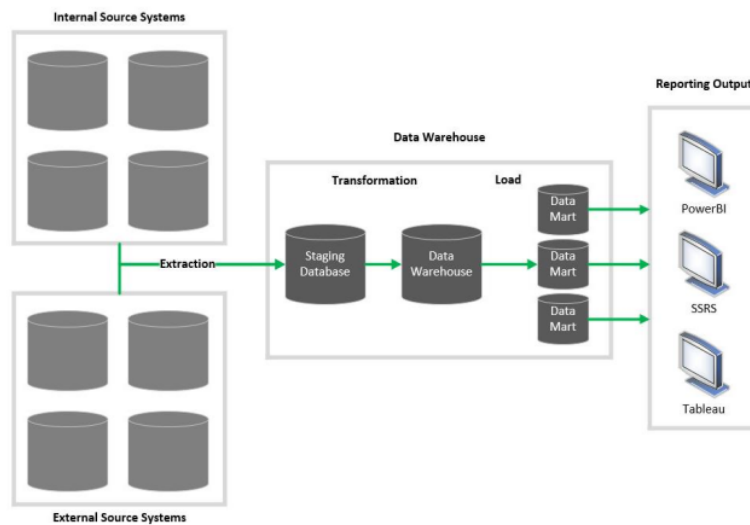


Figure 1 – ETL Process Example

2.3 Data Marts

According to IBM, a Data Mart is “a subject-oriented relational database that stores transactional data in rows and columns, which makes it easy to access, organize, and understand”. [5]

There’s a variety of advantages to BI users from the implementation of data marts, for the users specifically offering direct access to relevant data to that business function, while allowing customisation of the data models and subsequent queries. By providing just the relevant data this helps improve performance of reports, having to call upon a more focussed dataset. By having a data mart for specific business functions or departments it makes access permissions and usage policies easier to implement.

There are negatives to data marts, although having just the relevant data is focussed, it can be deemed a negative as it causes data silos. When designing the marts consideration should be taken to the data architecture, as it can quite easily create duplicated and redundant marts which are costly overheads with little benefit. There are also resource overheads, with a requirement to support and monitor moving forward.

2.3.1 Managing Director

The managing director is usually the head of an organisation, as such needs a high-level summary of the organization's performance in every area. Key performance indicators (KPIs) from all departments, including sales, finance, human resources, will need to be combined into a data mart for the managing director. Aggregated data of important parameters including revenue, profitability, market share, and overall business performance may be included in this data mart. All this data should aid the individual in evaluating the state of the company to make appropriate strategic decisions.

2.3.2 Finance Director

The Finance Director would be interested in a finance focussed data mart to assist with forecasting, budgeting and financial health needs. The mart should include finance ratios like Financial ratios, income statements, balance sheets, cash flow statements, budget variance analysis. To assess performance, this data mart may include comparisons with industry benchmarks and using historical financial patterns for forecasting. The analytics tools within the mart should allow the finance director to perform financial modelling, scenario analysis to aid decision-making.

2.3.3 Sales and Marketing Director

The director of sales and marketing would be interested in understanding market trends, sales performance, customer behaviour, and marketing efficiency. Data from a CRM database, sales database, and external data sources like market research and social trends need integration into the sales and marketing data mart. Metrics like customer sales pipeline analysis, competition analysis Return on Investment (ROI) should be available from this data mart, along with predictive and other modelling technique advanced analytics being utilised.

2.4 Dimensional vs Normalised

Dimensional and Normalised database modelling are the two most popular techniques preferred by the Data Engineering community as explained by *Teorey et al* [6]. While query performance and analytical flexibility is given priority by a data warehouse's dimensional structure, data integrity and transaction processing efficiency is the priority of an OLTP system's normalised structure. Transactional data from OLTP systems is transformed by data warehouses into a format that is ideal for business intelligence and decision support tasks.

2.4.1 Dimensional Database

Dimensional database modelling is a technique for creating databases optimised for querying large volumes of data efficiently. For this technique, the data is separated into Dimension and Fact Tables. Strings and other descriptive data are stored within dimension table, whereas Keys, integers and other quantifiable metrics that are held in Fact tables.

Dimensional modelling is the most efficient for query optimisation, it is simple, flexible, and intuitive. However, the focus on analytical querying can affect performance, and encounter complexities when updates to the schema are required.

The database model below is a basic example using the "star schema" method, a focal fact table surrounded by dimensions resembling a star. While some more sophisticated models employ a snowflake schema, which allows a dimension to be linked to another dimensions, it negatively affects efficiency.

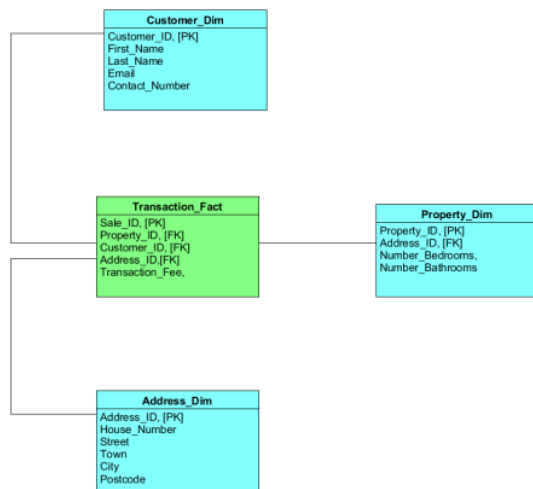


Figure 2 – Dimensional Database Example

2.4.2 Normalised Database

Normalisation is the process of setting up a model so that data can be adequately stored in an information database. The outcome is all tedious data is removed, leaving attribute-related data only in the table. Isolating an information base into at least two tables and outlining the relationships between them are standard components of normalisation. The objective is to separate the data so that adjustments, removals, and additions can be done in a single table.

There are five levels of normal form, however the desired outcomes can usually be reached when meeting third normal form.

First Normal Form (1NF): Eliminating repeatable groups, each cell holding singular values rather than a set of values and any repetition of groups and columns are removed.

Second Normal Form (2NF): The tables must meet the rules set in 1NF, and partial dependencies are now removed. Any attributes not identified as a key are now functionally dependent on the PK.

Third Normal Form (3NF): The tables must meet the rules set in 2NF, and elimination of transitive dependencies. There should be no dependencies for a non-key attribute on another, only to key attributes.

Additional levels include Fourth Normal Form (4NF), Fifth Normal Form (5NF) and Boyce Codd Normal Form (BCNF).

Normalisation unlike dimensional modelling is flexible when updates are required to the schema, it also requires less storage due to the integrity and fewer instances of redundant data within the modelling. This technique does however have overheads for performance, resources, queries, and manpower on occasions. The normalised data can require overly complicated queries, and multiple joins, and this can stem from the database schema being over-normalised.

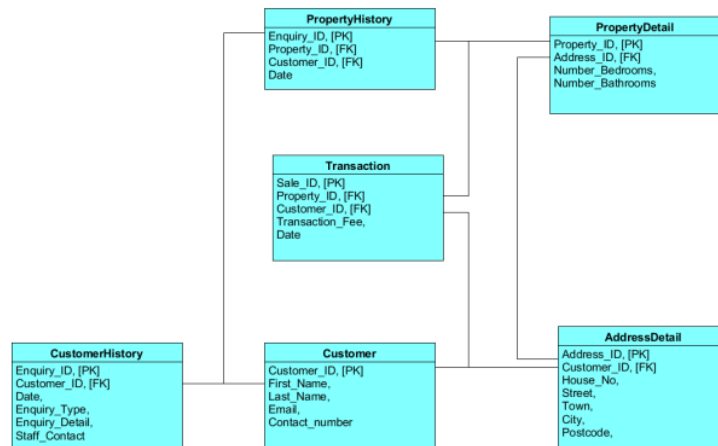


Figure 3 – Normalised 3NF Database Example

3. Task 2

3.1 Source Systems

It's possible to divide the source systems into internal and external. The external sources could be government statistics or open data sources like online listing portals and market research, whereas the internal data sources come from the Cambrian Real Estate-managed systems including Customer, Accounting and Accounting related data. An ETL system would merge these source systems into a staging area. In the event of a full system failure, it is simple to reload the data because staging of the data is done independently of the source system. By accessing directly from the staging rather than the source, the ETL processing time can be optimised, depending on the approach.

Within the brief, Cambrian Real Estate have identified the following sources which have been split into Internal and External sources, along with some further suggested data sources.

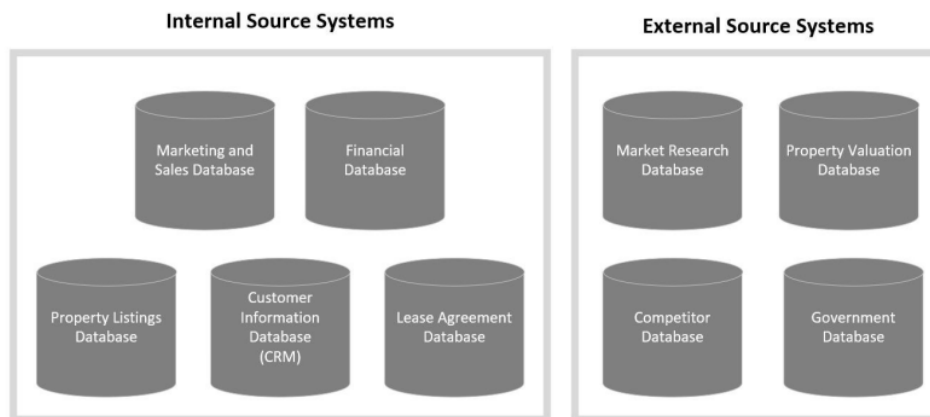


Figure 4 – Cambrian Real Estate Source Systems

3.2 System Schema

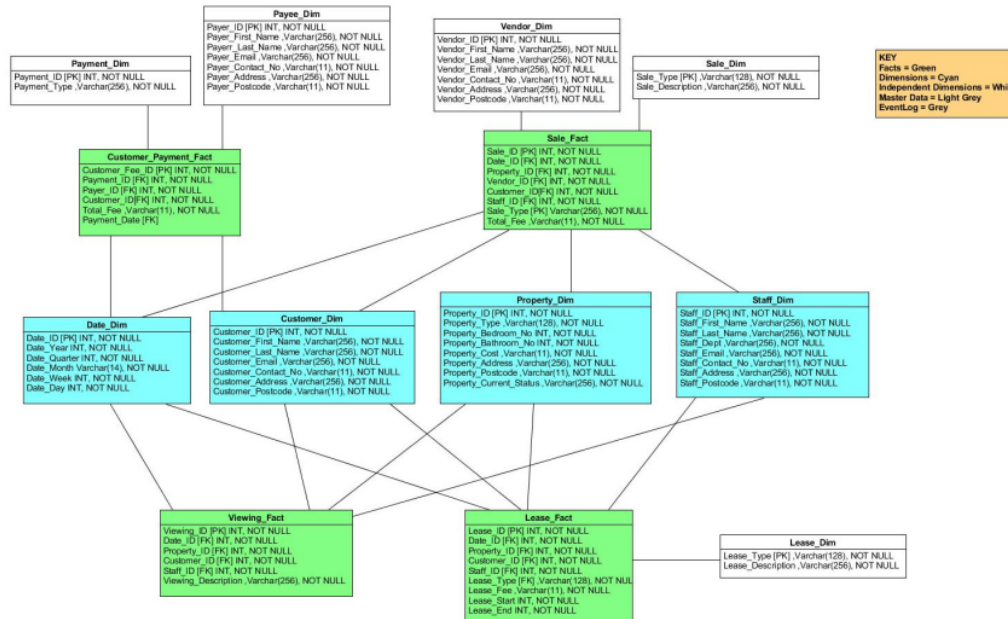


Figure 5 – Cambrian Real Estate Warehouse Star Schema

As can be seen by the above figure, the design has needed some consideration. For consistency, all attributes and tables are named with an upper camel case with underscore throughout. Additional documentation has been included with keys labelled, Null constraints and variable characters like VARCHAR and INT identified, along with maximum string lengths. Although Primary and Foreign Keys have been identified, this could've been improved with the labelling of surrogate keys.

This schema has the characteristics of a typical star schema, in that each table is a solitary join from data tables, and both the dimensions and independent dimensions provide groupings of business context. It was designed with consideration for optimisation and the ACID design principles. The acronym stands for Atomicity, Consistency, Isolation and Durability [7].

Atomicity: Commands should all be viewed as a single entity that either succeed or fail. The database's integrity should be preserved in the case of unforeseen disruption.

Consistency: All changes should follow constraints of the database, including rules, datatypes, triggers.

Isolation: Tables should be able to run in isolation, allowing commands to affect specific tables as required without affecting the whole database.

Durability: CRUD functionality persists within the database even during system outage or failure.

Aggregated Tables: The tables provide a summary of data produced from comprehensive base data taken from FACT tables. Precalculated totals and averages of a granular event are stored in these tables. As a result of processing less data to provide a response, query performance is enhanced.

Performance is best achieved by using distinct aggregate Fact tables rather than using existing granular Fact tables. End user queries aren't processed ad hoc, so pre-aggregating the data as part of the ETL process speeds up processing. This is also helpful when processing very large volumes of data, as real-time searches impede performance.

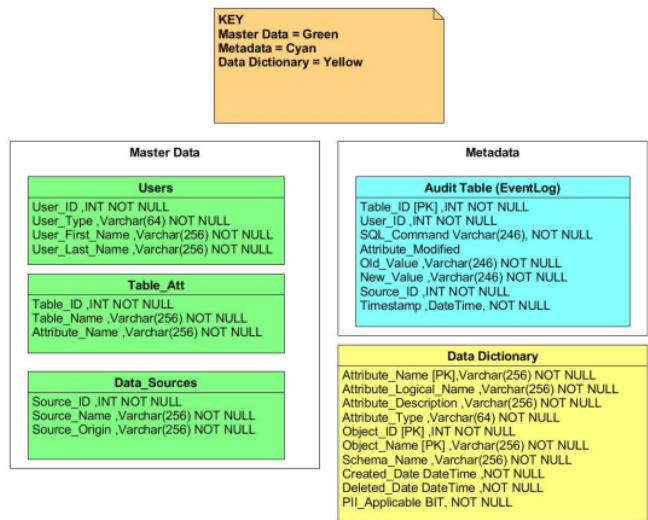


Figure 6 – Cambrian Real Estate Master data, Metadata and Data Dictionary

Metadata: Data that offers details about other data is referred to as metadata. To put it simply, it's data about data. It provides further specifics about information's structure, like it's format, usage, and characteristics. Metadata plays a role in managing, organising, and understanding the data stored within the Warehouse. Large volumes of data can be better categorised, data search and retrieve operations improved, data integrity ensured, and data integration can be facilitated with effective metadata.

Audit Table: Is an audit table trail for all key actions made on the data warehouse. When a given event takes place, a database trigger will execute a series of instructions. In this instance, a table will record the results of any Data Definition or Manipulation operations performed within the warehouse. This table supports real-time monitoring, security, compliance, and auditing by showing what changes were made, and by whom, tracing unacceptable changes to the source.

Data Dictionary: A data dictionary is a centralised metadata repository that offers detailed information on the data items used in the databases and information systems of an organisation as explored by *Buchanan et al* [8]. It acts as a manual or reference for comprehending the relationships, structure, description, and use of the various data items. It is particularly important in specifying whether a data item is Personally Identifiable Information (PII), which would require said items to be encrypted or pseudonymised.

The Dictionary is useful for both the business and IT teams, by documenting the items including their meanings, uses, it promotes consistency across the organisation, which is appreciated by end users down the line when digesting reports and the development teams when making changes to datasets.

3.3 Data Marts

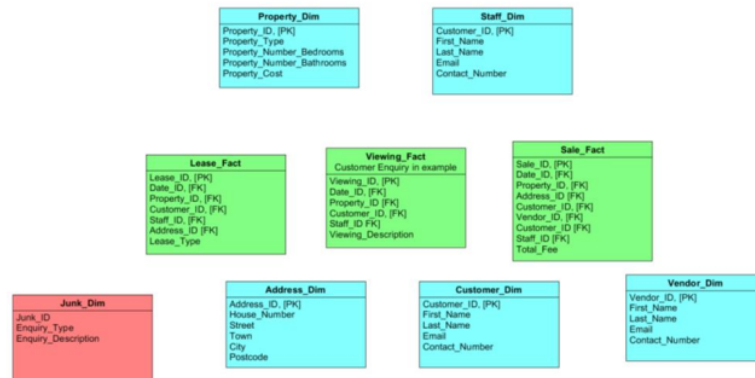


Figure 7 – Data Mart

For Cambrian Real Estate there are likely to be multiple data marts required, a strategic Mart with aggregated data of need-to-know metrics for the Managing Director, Financial and Marketing Marts were all discussed previously. Alongside these Marts it's likely that a HR mart for Staff specific reporting.

The initial star schema design lends itself to data marts, the tables identified are likely to be present in other data marts, known as conformed dimensions. These might have extra properties on the table designating it as a conformed dimension, or they could be within a conformed schema, meaning it is part of various ETL operations. An enhancement for improved queries has been to split the Property_Dim, taking the address and creating its own Dimension for any geographical specific queries. A junk Dimension has been added, it is best practice to include for capturing flags or indicators. It can be included per warehouse or per mart, when there are instances of large free key data it would be beneficial to include as it can improve overall query performance.

Additionally, by consolidating the facts down to lease, sales, and customer with data at the lowest level of transactional data granularity, it could be feasible to service multiple Marts with a single overarching schema. It could serve as the foundation for aggregated tables; however, it would be advisable to develop reporting views. Views sit on top of the fact tables and contain business logic. A view is recommended for basic searches that do not process vast volumes of data.

The effectiveness of Data Marts when used against large volumes of stored data for the purpose of building reporting output and its effectiveness was investigated by *Vladimir et al* [9].

3.4 ETL Plan

3.4.1 Extraction

The ETL process run time can affect performance issues or failures. Scheduling nightly during

unsociable hours not only keeps the warehouse up to date, but failures and performance issues have minimal impact. The impact on operational services can be minimised by accessing previously identified internal databases including those containing financial transaction data, and listings during off-peak hours. The Larger databases could be incrementally imported daily, with new and updated data being checked for deltas. External sources may require direct database connections or API calls.

3.4.2 Transformation

This stage of the process aims to cleanse and standardise the data, as imported data isn't always clean when landed into the Data Warehouse. In this phase the data should be checked for missing values, inconsistency, and inaccuracies. Effective data validation rules will improve data quality, however for consistency data should still be standardised, and formatted appropriately for reliable reporting with accuracy when visualised. Examples of transformation activities include,

Property Data: Product names converted into a consistent format; all naming variants stored in a consistent format. If necessary, the data can be enriched by adding an additional column for Property_Category, with a number then allocated for each.

Lease and Sale Data: Date of sales are already captured as date and time rather than string, however total sums can be calculated for each transaction type using lease_fee and total_fee along with transaction types for further Sales metrics. Similarly staff names are already split into first and last name avoiding an additional normalisation step, allowing the transaction data and staff data to be aggregated for provision of sales by staff metrics.

Customer Data: Data standardisation of customer addresses will be required as currently address is captured within one attribute. The data can then be cleansed, removing duplicate records where necessary, as a vendor could have a One-to-Many relationship with properties. The addressing reporting could then be enriched using the postcode for an additional region column for ease of use.

3.4.3 Loading

Once transformed the modified data must be imported into the data warehouse. Prior to populating the Fact tables, dimension tables must be loaded with the cleansed data, allowing all foreign key relationships to be present. Checking and loading dimensions according to type is best practice. The Fact tables can be loaded after all dimensions have been loaded. An Example of Loading process below,

Target Table: Sale_Fact Table, Contains sales transaction data including sales amount, date, customer_ID, Vendor_ID, staff_ID and Sale_Type.

Customer_Dim Table, Contains customer information including Customer_ID, and demographic information like name, address, region, etc.

Loading Process: Transformed Sales data is loaded into the Sales Fact Table. Transformed customer data loaded into the Customer Dimension Table, removing duplication. Property data is loaded into the Property_Dim Table, ensuring that only active property data is loaded.

Error Handling: Mechanisms should be implemented to log any errors within the ETL process, with failed records being logged for review and potential manual interventions.

3.5 Data Quality and Governance

The following have been identified as data quality and governance topics to be considered for the Data Warehouse moving forward, explored by *Yeoh et al* [10].

Data Privacy: Sensitive information should be safeguarded, compliance with data privacy regulations, such as GDPR (General Data Protection Regulation) as a minimum expectation. A key activity involves anonymizing or pseudonymizing personal PII data, retaining data for no longer than is required and obtaining consent for customer data.

Data Security: Data should be from unauthorized access, only those with correct credentials should be able to access. Activities include implementing access controls and auditing mechanisms to meet regulatory compliance requirements.

Data Accuracy: Data loaded into the Warehouse should strive to be error free and as accurate as possible. Source data should be validated against predefined rules, data cleansing activities implemented where necessary to remove inaccuracies.

Data Governance: Cambrian Estates must implement policies and procedures for managing data throughout the lifecycle, during collection, storage, usage, and disposal. It would be advisable to define roles and responsibilities, assign data ownership and implement appropriate access controls.

4. Conclusion

In conclusion, the implementation of a data warehouse at Cambrian Real Estate will only enhance the capacity and provision of data for analysis, decision-making, and operational effectiveness. The storage and use of large volumes of data requires the use of ETLs and data warehousing. Given the real estate market is a saturated one, with plentiful competition it's key for Cambrian to stand out by making informed decisions earlier than their competitors to gain the upper hand over their rivals. While providing the benefits in the market, a data warehouse solution implemented appropriately will enhance the company's data quality and governance. The data warehouse provides a platform for business intelligence and analysis through its data marts approach.

5. Reference List

- [1] Anonymous, "What is OLTP?", 2024. [Online]. Available: [https://www.oracle.com/uk/database/what-is-oltp/#:~:text=What%20is%20OLTP%3F-,OLTP%20defined,sending%20text%20messages%2C%20for%20example](https://www.oracle.com/uk/database/what-is-oltp/#:~:text=What%20is%20OLTP%3F-,OLTP%20defined,sending%20text%20messages%2C%20for%20example.). [Accessed: 21-Mar-2024]
- [2] Anonymous, "What is a Data Warehouse?", 2024. [Online]. Available: <https://www.ibm.com/topics/data-warehouse> [Accessed: 21-Mar-2024]
- [3] R. Kimball and M. Ross, "The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modelling", 2008. [Online]. Available: <https://www-vlebooks-com.ezproxy.uwtsd.ac.uk/Product/Index/20119?page=0&startBookmarkId=-1> [Accessed: 21-Mar-2024]
- [4] R. S. Segall, *Handbook of research on big data storage and visualization techniques*. Hershey, Pennsylvania: Engineering Science Reference, 2018. [Online]. Available: <https://www-igi-global-com.ezproxy.uwtsd.ac.uk/gateway/book/179829> [Accessed: 21-Mar-2024]
- [5] Anonymous, "What is a Data Mart?", 2024. [Online]. Available: <https://www.ibm.com/topics/data-mart> [Accessed: 21-Mar-2024]
- [6] T. J. Torey, "Database modeling and design logical design", 2011. [Online]. Available: <https://web-p-ebsscohost-com.ezproxy.uwtsd.ac.uk/ehost/detail/detail?vid=0&sid=6aac81e-30d6-4ca1-8974-d567efc58581%40redis&bdata=JnNpdGU9ZWwhvc3QtGl2ZQ%3d%3d#AN=356330&db=nlebk> [Accessed: 21-Mar-2024]
- [7] Anonymous, "What is an ACID Transaction?", 2024. [Online]. Available: <https://www.databricks.com/glossary/acid-transactions> [Accessed: 21-Mar-2024]
- [8] E. M. Buchanan *et al*, "Getting Started Creating Data Dictionaries: How to Create a Shareable Data Set," 2021, [Online]. Available: <https://journals.sagepub.com/doi/full/10.1177/2515245920928007> [Accessed: 21-Mar-2024]
- [9] V. Belov, A. N. Kosenkov, and E. Nikulchev, "Experimental Characteristics Study of Data Storage Formats for Data Marts Development within Data Lakes," 2021 [Online]. Available: <https://www.proquest.com/docview/2576378772?accountid=130472&parentSessionId=ehEXdeAFrsjXcfYYTF4ygm8GFK0nkb0paKk0KcZyA%3D&pq-origsite=primo&sourcetype=Scholarly%20Journals> [Accessed: 21-Mar-2024]
- [10] W. Yeoh, *et al*, "Information quality and governance for business intelligence". 2014. [Online]. Available: <https://www-igi-global-com.ezproxy.uwtsd.ac.uk/gateway/book/78926> [Accessed: 21-Mar-2024]

Figures

Figure 1 – ETL Process Example

Figure 2 – Dimensional Database Example

Figure 3 – Normalised 3NF Database Example

Figure 4 – Cambrian Real Estate Source Systems

Figure 5 - Cambrian Real Estate Warehouse Star Schema

Figure 6 - Cambrian Real Estate Master data, Metadata and Data Dictionary

Figure 7 – Data Mart

FINAL GRADE

GENERAL COMMENTS

81 /100

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

PAGE 11

PAGE 12

PAGE 13

PAGE 14

PAGE 15