

Assignment 1 Data Warehousing

by Sameeran Raveenthiran

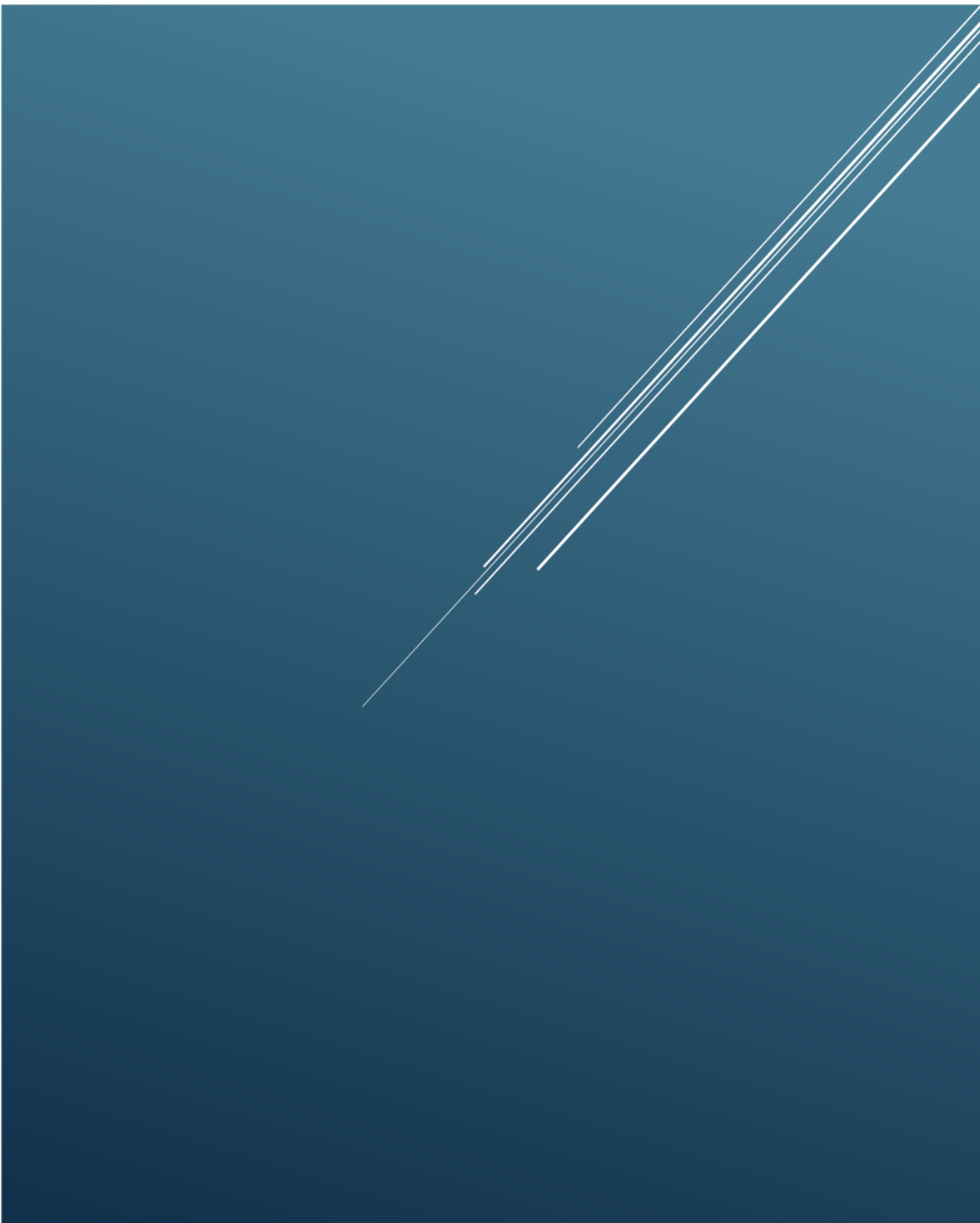
Submission date: 22-Mar-2024 12:40PM (UTC+0000)

Submission ID: 227578108

File name: 33405_Sameeran_Raveenthiran_Assignment_1_Data_Warehousing_565908_1386966180.docx
(1.06M)

Word count: 3466

Character count: 22233



DATA WAREHOUSING

Sameeran Raveenthiran (2114454)

Contents

Introduction:	2
Task 1:	2
Difference and Similarity between OLTP System and Data Warehouse	2
Deciphering Data Warehouse Architecture: Integrating Internal and External Data Source	3
What is ETL?	6
Empowering Management Decisions: Utilising Data Marts to Meet Directorial Needs	7
Differentiating OLTP and Data Warehouse Structures for IT Directors	7
Task 2:	9
Identification of Internal and External Source System:	9
Implementation Optimal Star Schemas and Guide to Data Mart Identification and Extraction Schemas:	10
Strategic ETL Blueprint: Sourcing and Transforming data for Efficient Data Warehouse Loading	11
Discussion of Data Quality and Governance in ETL and Warehouse Operations.	12
Conclusion:	12
References:	13

Introduction:

Based in Swansea, UK, Cambrian Real Estate Company has been a significant player in residential and commercial real estate transactions across multiple UK cities for over three decades. Recently, it expanded its reach to include properties in the Republic of Ireland, managing a diverse portfolio that includes private homes, apartment buildings, offices, and shopping centres. The company operates various departments crucial to its success, including Sales, Property Management, Customer Service, Finance, and Marketing.

Presently, Cambrian utilises an Operational Transactional Processing (OLTP) system to efficiently manage transactions across departments, handling customer interactions, property listings, management data, leases, and financial transactions. However, despite its efficiency in daily operations, this system lacks comprehensive data analysis and reporting capabilities. Recognising the potential for improved decision-making and business intelligence, CEO Mary Thomas has explored the implementation of a data warehouse solution. This initiative aims to enhance the existing OLTP system, enabling Cambrian to effectively leverage its data. This report aims to educate the board on data warehousing, design an appropriate data warehouse structure, identify relevant data marts, and outline a simplified ETL plan. This strategic approach aligns with Cambrian's business strategy, aiming to drive smarter decision-making, improve customer service, and gain a competitive edge in the real estate market.

Task 1:

Difference and Similarity between OLTP System and Data Warehouse

OLTP and Data Warehouse have some similarities and differences. First, OLTP and Data Warehouse are the same because they both maintain high data quality and integrity as it is important because inaccurate data can cause many problems in operational systems as well as in analytical reports. Additionally, in both cases, there might be a requirement for data transformation, even though it will be for different purposes. OLTP systems will have the need to transform data for transaction processing, relatively Data Warehousing involves aggregating and transforming data for analytical purposes.

On the other hand, OLTP and Data Warehouse have some differences. First, the purpose of OLTP integration is to perform day-to-day, transactional operations. It targets on efficiently handling and processing a large volume of small, routine transactions in real-time. Such as, online banking, order processing, and reservation systems, but Data Warehousing lies towards supporting analytics and business intelligence, and this includes organising, consolidating, and summarising data from many derived sources to provide a foundation for decision-making. Second, OLTP typically has a normalised data model, which effectively minimises redundancy to guarantee efficient transaction processing, although Data Warehouse often employs

star schema or denormalization structure to ease quick querying and reporting, this enables more efficient analysis. Third, in terms of performance OLTP prioritises fast response times for each individual transaction. In contrast, Data Warehouse is capable of processing large volumes of data to support complex analytical queries. Clearly, OLTP and Data Warehouse have similarities and differences.

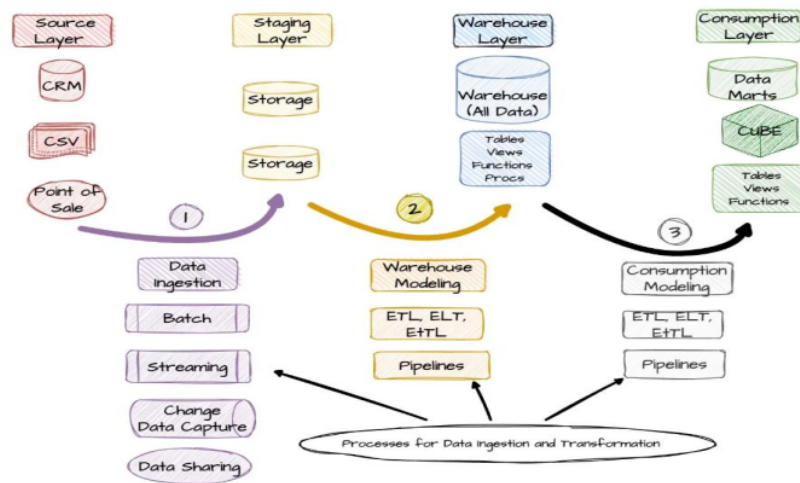
The benefits of Data Warehouse Investment

In addition, investing in a data warehouse system can bring several benefits to an organisation, enhancing its ability to manage and leverage data effectively for strategic decision making. For example, data warehouse provides a consolidated and reliable source of data. This leads to better informed decision making, through analysing aggregated and historical data, as this process ensures a much clearer picture of trends, business performance, and patterns. Furthermore, as mentioned above, data warehouse tends to be more scalable because it can withstand much larger volumes of data, as a company grows and accumulates more data, a scalable data warehouse can accommodate increased data storage and processing requirements. When it comes to an organisation being cost efficient is extremely crucial, while initial investment costs may be involved in implementing a data warehouse, the long term benefits it produces outweighs the expenses. Such as, improved efficiency in reporting and analysis will potentially lead to cost savings, and the ability to make informed decisions can positively impact the bottom line.

Deciphering Data Warehouse Architecture: Integrating Internal and External Data Source

The role of the data warehouse architecture is to allow decision-making, and fact-driven process. Which can be used by everyone within the organisation, especially when organisations utilise self-service analytics to their advantage. Data Warehouse Architecture is a subsystem, and intentional design of data services that combines different data sources into a single repository for the purpose of business intelligence, and analysis. The architecture is arranged by a set of logical services that creates the backbone of a data warehouse system, that contributes to a structured way of managing, storing, and retrieving large amounts of data.

Layers of Data Warehouse Architecture:



Data

warehouse consists of various functional layers, and each have their own specific capabilities. Source, staging, warehouse, and consumptions are the most common data warehouse architecture layers.

Source Layer: Referred to as the logic layer of all systems of record (SOR) contributing data to the warehouse such as point-of-sale, CRM, ERP systems, or marketing automation. Each SOR source adheres to a distinct data format, and the data capture may vary based on the specific requirements of that format.

Staging Layer: A designated space for data originating from the sources SOR serves as a landing zone. A recommended best practise for data staging involves ingesting information from the SOR without implementing business logic or transformations. It is imperative to emphasise that staging data should not be utilised in production data analysis, as the information within this area is yet to undergo processes such as cleansed, standardised, modelled, governed, and verified.

Warehouse Layer: This layer serves as the storage repository for all data, The warehouse data is now subject-oriented, integrated, time-variant, and non-volatile. Within this layer, the necessary components, including physical schemas, views, tables, stored procedures, and functions, are present to facilitate access to the modelled data in the warehouse.

Consumption Layer: This is also referred to as the analytics layers, and this layer is where the user is expected to model data for the purpose of consumption using analytics tools. Such as, ThoughtSpot, data scientists, business users, and data analysts.

Source System: A data warehouse source system designates the original or external system from which the data is eventually stored in a data warehouse. The source system contains the raw data produced by an organisation operational

processes and transactions. Subsequently, the data is extracted from these source systems, undergoes transformation, and is then loaded into the data warehouse, serving analytical and reporting purposes.

Key characteristics of data warehouse source systems include:

Operational Systems: Source system frequently consists of operational databases that underpin an organisations day to day functions. These databases encompass transactional systems for finance, sales, human resources, inventory, and other business processes.

Diversity of Sources: Source systems exhibit diversity, encompassing various types of databases, applications, and external data sources. Examples of source systems span from traditional relational databases to more recent data sources like NoSQL databases, web services, or flat files.

Enterprise Applications: Commonly, source systems include enterprise applications such as Enterprise Resource Planning (ERP), Supply Chain Management (SCM), and Customer Relationship Management (CRM). These applications diligently gather and store data pertaining to different facets of an organisation's operations.

External Data Sources: Beyond internal systems, organisations frequently integrate external data sources, including market research data, industry reports, social media feeds, thereby enriching their comprehension of the business environment.

Structured and Unstructured Data: Source systems can accommodate both structured data, meticulously organised in tables, and unstructured data, such as text documents, multimedia, or images. Consequently, data warehouses often grapple with the integration of these disparate data types.

Filling a data warehouse encompasses the extraction of data from diverse internal and external sources, modifying it as needed, and then loading it into the data warehouse to support analytical objectives. Let's provide a summary of how various sources can be employed within this procedure.

Internal Sources:

Internal Sources encompass data generated within the organisation through its day-to-day operational processes. For example, common internal sources include transactional databases, enterprise applications such as ERP, SCM, CRM, internal data warehouses, and other systems integral to routine business operations.

Furthermore, the data originating from internal sources inherently and managed by the organisation. It encompasses important information about finances, sales, and various internal processes. Internal data sources often exhibit a high degree of integration into the organisations existing systems. This tight integration facilitates a

relatively seamless process for extracting, transforming, and loading (ETL) data into the data warehouse.

External Sources:

External Sources involve data originating outside the organisation. This data is collected from entities or system external to the organisational internal operations. Furthermore, external sources comprise market research reports, social media feeds, industry publications, external databases, web services, and any information obtained from sources external to the organisation.

The characteristics of external data provides valuable context beyond the organisation's internal operations. It includes information about market trends, competitor activities, and various external factors influencing the business environment. Integrating external data poses challenges due to differences in formats, structures, and data quality. Processing and validating data from external sources may require additional efforts.

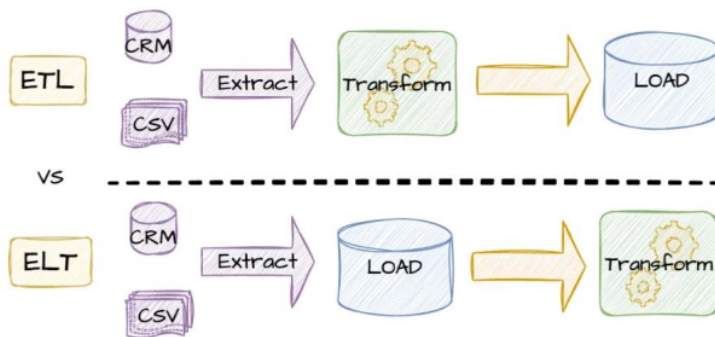
What is ETL?

The ETL process, short for extract, transform, load, integrates data from various sources into a unified dataset for a data warehouse or similar systems. Emerging in the 1970s with the rise of databases, ETL has become vital for data integration, computation, and analysis in warehousing projects. It sets the stage for analytics and machine learning by cleansing and organizing data according to business needs, improving, reporting, facilitating advanced analytics.

Organisations frequently employ ETL to:

- Extract data from legacy systems.
- Cleanse data to enhance data quality and ensure consistency.
- Load data into a designated target database.

This robust process is indispensable for achieving data-driven insights and maintaining data integrity within organisations.



Implementing ETL procedures offers numerous benefits, including enhanced data integrity, streamlined analytics, and improved decision-making. ETL ensures precision by extracting data from various sources, transforming it into a uniform format, and loading it into a central repository. ETL seamlessly integrated with day-to-day operations, minimising disruptions through efficient scheduling. Additionally, it automates data integration tasks, freeing up resources for strategic initiatives and expediting data processing for timely insights and actions.

In conclusion, ETL implementation signifies a strategic investment in data management that synergises with existing operations, fortifying the organisation's analytical capabilities seamlessly.

Empowering Management Decisions: Utilising Data Marts to Meet Directorial Needs

Data Marts are tailored repositories designed to meet the unique needs of organisational departments like the managing director, finance director, and sales and marketing director. They provide precise information crucial for decision-making. For the managing director, data marts offer insights into organisational metrics like revenue, profitability, and market share. Finance director benefit from financial data focused on revenue trends, expense analysis, and budget variances. Sales and marketing directors gain insights into sales performance, customer behaviour, and market trends. By consolidating relevant data, data marts facilitate strategic planning, cost optimisation, and marketing strategy enhancement, empowering directors to steer the company toward its objectives.

Through the utilization of data marts, each director gains access to timely, pertinent, and actionable information tailored to their specific functional domain, thereby enabling them to make well-founded management decisions that drive organizational success.

Differentiating OLTP and Data Warehouse Structures for IT Directors

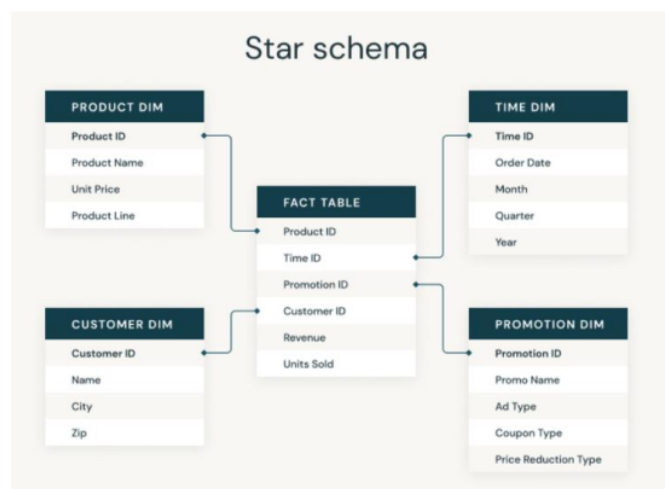
While both normalised structures in OLTP databases and dimensional structures in data warehouses serve as means of organising and storing data, they differ fundamentally in design, purpose, and usage.

In OLTP databases, the normalised structure places a premium on data integrity and minimizes redundancy. Data is partitioned into separate tables, and relationships are established using foreign keys. This ensures that each information is stored only once, minimising the likelihood of data anomalies and inconsistencies. However, this normalisation strategy may result in intricate joins and queries, impacting query performance, particularly with analytical queries spanning multiple tables.

Conversely, the dimensional structure of a data warehouse focuses on facilitating ease of querying and analytical processing. It revolves on a central fact table containing business metrics or measures, surrounded by dimensions tables that provide context and descriptive attributes. This denormalised arrangement simplifies queries and reporting by reducing the necessity for complex joins and expediting data retrieval. Moreover, the dimensional modelling is tailored for analytical queries commonly employed in decision support and business intelligence applications.

Moreover, while the normalised OLTP structure is geared towards transactional processing, ensuring the accurate capture and maintenance of real-time data, the dimensional structure of a data warehouse is geared towards analytical processing. It accentuates historical data analysis, trend identification, and decision-making support.

To Summarise, normalised OLTP structures prioritise data integrity and transactional processing efficiency, dimensional structures in data warehouses cater to optimised analytical querying and decision support. They offer a denormalized, simplified schema conducive to efficient data analysis.



A star schema, a prevalent dimensional

modelling method for data warehouse, features a central fact table surrounded by dimensional tables, forming a visual representation resembling a star. The fact table consists of quantitative data or facts, while dimension tables contain descriptive attributes providing context to the facts. For instance, in the image above, the fact table would contain sales transaction data, whereas the dimension tables would offer details about products, customer, time, geographical aspect, and promotion. This design facilitates efficient querying and analysis, enabling users to gain valuable insights into business operations through various dimensions.

Task 2:

Identification of Internal and External Source System:

Internal Source System:

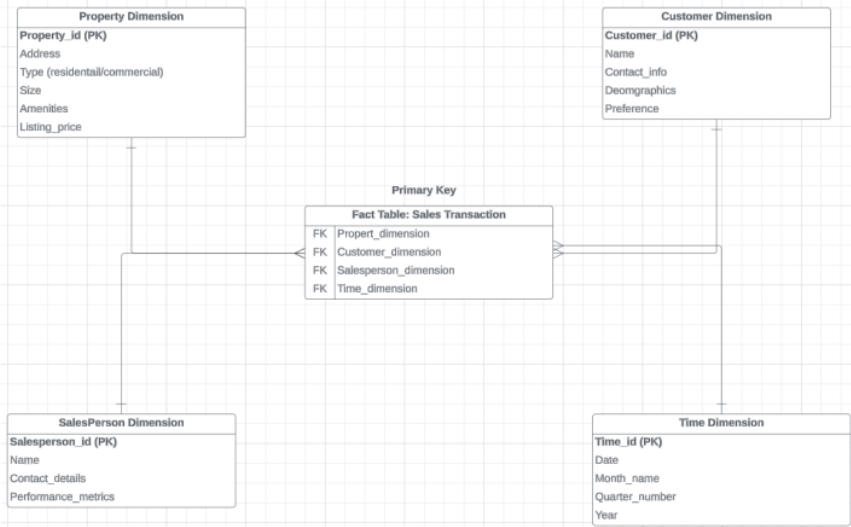
- **Customer Database:** This includes information about the customers, contact details, preferences, and purchase history.
- **Property Listings Database:** Saves details regarding the available properties, addresses, specifications, and pricing.
- **Lease Agreement Database:** Oversees lease agreements for both commercial and residential properties, this includes rental amounts, terms, and lease durations.
- **Financial Database:** Monitors financial transactions, such as expenses, sales revenue, and profit margins.

External Source System:

- **Market Trends Data:** Offers insights into market trends, such as pricing fluctuations, property demands, and regional preferences.
- **Property Valuation Data:** Provides information regarding the property valuations, by aiding the assessment valuation of Cambrian's real estate portfolio and probable investment opportunities.
- **Demographic Data:** Delivers demographic information for various locations, purpose of assisting target market analysis and property pricing strategies.
- **Competitor Data:** Provides insights into competitor activities, pricing strategies, market positioning, and allowing Cambrian to discover competitive advantages and areas for improvement.

Implementation Optimal Star Schemas and Guide to Data Mart Identification and Extraction Schemas:

Main Star Schema:



Additional Star Schema:

In the proposed main star schema, we advocate for the establishment of several data marts to effectively categorise and analyse the diverse datasets of Cambrian Real Estate Company. The sales Data Mart will be a central hub for scrutinising sales transactions and performance metrics, featuring the pivotal Fact Table "Sales Transactions". This mart will be enriched by Dimension Tables such as "Property Dimension", "Customer Dimension", "Salesperson Dimension", and "Time Dimension".

Additionally, the Financial Data Mart will diligently monitor financial transaction and metrics, comprising the core Fact Table "Financial Transactions". Supporting this table will be Dimension Tables including "Property Dimension", "Time Dimension", and "Expense Type Dimension."

Moreover, the Customer Relationship Data Mart will concentrate on customer behaviour analysis, with the Fact Table "Customer interactions" at its core. Dimension Tables such as "Customer Dimension" and "Time Dimension" will augment this mart.

In the additional star schema, the meticulously crafted data marts will be tailored to specific business objectives, ensuring a nuanced understanding of sales, financial, property management, and customer relationship data. This structured approach is poised to empower Cambrian Real Estate Company with the insights necessary for informed decision-making and strategic planning, thereby propelling business growth and bolstering its competitive edge.

Strategic ETL Blueprint: Sourcing and Transforming data for Efficient Data Warehouse Loading

In devising a simplified ETL strategy for the company's data warehouse, it is crucial to outline the data sourcing and transformation process for each table within the proposed star schema. The Fact Table "Sales Transactions" in the Sales Data Mart, for instance, will primarily source data from the company's existing OLTP system, capturing vital sales-related details such as transaction dates, sales amounts, and customer information. Similarly, Dimension Tables like "property Dimension" and "Customer Dimension" will draw data from internal databases such as the property listings database and customer database. These sources will provide comprehensive insights into properties listed by Cambrian and customer profiles, encompassing contact details and preferences.

In the Financial Data Mart, the Fact Table "Financial Transaction" will gather data from the company's financial database, encompassing revenue, expenses, and profits. Additionally, Dimension Tables like "Property Dimension", and "Expense Type Dimension" will acquire contextual data from internal systems. The Fact Tables "Property Listings" and "Lease Agreements" in the Property Management Data Mart will source data from internal databases. Transformations involve data cleansing, eliminating duplicates, rectifying errors, and standardising data formats. Data aggregation and enrichment enhance data quality. Adhering to this ETL plan enables Cambrian Real Estate Company to seamlessly load data into the data warehouse, facilitating advanced analytics informed decision-making.

Discussion of Data Quality and Governance in ETL and Warehouse Operations

Implementing a new data warehouse system requires careful attention to data quality and information governance. During the ETL process, ensuring data accuracy and integrity is essential. This involves rectifying errors, eliminating duplicates, and maintaining consistency across data sources. Data security and privacy are also critical, requiring measures to safeguard against unauthorised access and compliance with regulations like GDPR or HIPAA.

During normal operations, ongoing monitoring and maintenance are necessary to ensure data accuracy and relevance. Data governance frameworks establish accountability and ownership of data assets, while data stewardship programs oversee quality initiatives and compliance.

Data lineage and documentation facilitate transparency and auditability, enabling stakeholders to understand the context and reliability of analytical insights. Regular data profiling and quality assessments detect anomalies and trends, facilitating proactive remediation.

Effect management of data quality and information governance is vital for informed decision-making and strategic objectives. Cambrian Real Estate can leverage these practises to drive continuous improvement and achieve success in its data initiatives.

Conclusion:

In conclusion, Cambrian Real Estate Company's pursuit of heightened decision-making and business acumen through the adoption of a data warehouse solution underscores its dedication to leveraging data strategically. Recognising the constraints of their OLTP system, the company aims to bolster its capabilities with a tailored data warehouse framework that caters to its diverse requirements. By addressing the distinct needs of key stakeholders like the Managing, Director, Finance Director, Sales and Marketing Director, and IT Director, Cambrian

endeavours to unlock insights, streamline operations, and secure a competitive advantage in the real estate sector. With a focus on amalgamating internal and external data source, the envisioned data warehouse solution promises to empower informed decision-making, elevate customer service, and discern evolving market trends. Through this strategic endeavour, Cambrian positions itself for sustained growth and prosperity in the ever-evolving real estate landscape.

References:

Task 1:

OLTP:

1. <https://www.ibm.com/topics/oltp>
2. <https://www.oracle.com/uk/database/what-is-oltp/>

Data Warehouse:

3. <https://www.ibm.com/topics/data-warehouse>

Data Warehouse Architecture:

4. <https://www.thoughtspot.com/data-trends/data-modeling/data-warehouse-architecture#DWA1>

Internal Sources & External Sources:

5. Inmon, W. H., & Hackathorn, R. D. (1994). "Using the Data Warehouse."
6. Kimball, R., & Ross, M. (2002). "The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling."
7. Turban, E., Sharda, R., & Delen, D. (2019). "Business Intelligence: A Managerial Perspective on Analytics."

ETL:

8. <https://www.ibm.com/topics/etl>

Data Marts:

9. Kimball, R., & Ross, M. (2013). The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modelling. John Wiley & Sons.
10. Inmon, W. H., & Hackathorn, R. D. (2007). Using the Data Warehouse. John Wiley & Sons.

11. Vaisman, A., & Zimányi, E. (2014). Data Warehouse Systems: Design and Implementation. Springer Science & Business Media.

“Normalised” OLTP & “Dimensional” Data Warehouse:

12. Kimball, R., & Ross, M. (2013). The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling. John Wiley & Sons.
13. Inmon, W. H., & Hackathorn, R. D. (2007). Using the Data Warehouse. John Wiley & Sons.
14. Date, C. J., Darwen, H., & Lorentzos, N. A. (2006). Temporal Data and the Relational Model: A Detailed Investigation into the Application of Interval and Relation Theory to the Problem of Temporal Database Management. Morgan Kaufmann.
15. https://www.academia.edu/748815/The_data_warehouse_lifecycle_toolkit?uc-g-sw=19222417

Task 2:

Main Star Schema:

1. Inmon, W. H., & Hackathorn, R. D. (1994).
2. Using the Data Warehouse. John Wiley & Sons.

Data Marts:

3. Kimball & Ross, 2013; Inmon & Hackathorn, 1994

ETL:

4. <https://www.linkedin.com/advice/0/what-impact-does-data-quality-have-etl-processes>
5. <https://www.matillion.com/blog/what-is-etl-the-ultimate-guide>

Discussion:

6. Kimball, R. and Ross, M. (2013) The Data Warehouse Toolkit: The Complete Guide to Dimensional Modelling. 3rd Edition, John Wiley & Sons, Inc., Indianapolis.

Assignment 1 Data Warehousing

GRADEMARK REPORT

FINAL GRADE

GENERAL COMMENTS

54/100

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