

Data Warehousing and Data Mining

by Karen Windsor

Submission date: 21-Mar-2024 10:34AM (UTC+0000)

Submission ID: 227374267

File name: 24482_Karen_Windsor_Data_Warehousing_and_Data_Mining_565908_1698261072.docx (2.89M)

Word count: 5120

Character count: 29543

Data Warehousing and Data Mining

ACCA6026 CAMBRIAN REAL ESTATE COMPANY
KAREN WINDSOR (1918576)

Table of Contents

1	Introduction	3
2	Task 1	3
2.1	The benefits of a data warehouse over an Online Transaction Processing (OLTP) system..	3
2.2	Data Warehouse Architecture	4
2.2.1	Architectural Layers.....	4
2.2.2	Data ingestion and transformation	5
2.2.3	Data Warehouse Architecture Components	6
2.3	Data Marts	7
2.3.1	What is a Data Mart?.....	7
2.3.2	Purpose of a Data Mart	7
2.3.3	Data Mart formation and operations.....	7
2.4	Difference between a 'Normalised' OLPT system and a 'Dimensional' structure of a Data Warehouse.....	8
2.4.1	Normalised Model.....	8
2.4.2	Dimensional Structure	8
3	Task 2	10
3.1	Internal and External Sources	10
3.2	Star Schema	12
3.3	Data Marts	17
3.3.1	Sales Data Mart.....	18
3.3.2	Rental/Lease Data Mart	19
3.3.3	Maintenance Data Mart	20
3.3.4	Finance Data Mart.....	21
3.4	ETL Plan.....	22
3.4.1	Data Source/cleansing and destination.....	24
3.5	Data Governance and Data Quality within ETL	25
3.5.1	Data Quality.....	25
3.5.2	Data Governance	25
4	Conclusion	25
5	References	26
6	Appendix	27
6.1	OLTP Structure.....	27
6.2	Data Warehouse Structure	27
6.3	Data Mart Examples.....	28
6.4	Metadata Business Rules.....	28
6.5	Data Cleansing	29
6.6	Data Quality Life Cycle	30
6.7	Sensitive Data	30

Figure 1 Data Warehouse Architecture	4
Figure 2 Ingestion and Transformation	5
Figure 3 ETL Process	6
Figure 4 Mirroring.....	6
Figure 5 Data Mart.....	7
Figure 6 Normalisation Process	8
Figure 7 Star Schema	9
Figure 8 Property Sales Star Schema.....	12
Figure 9 Property Lease/Rental Star Schema.....	13
Figure 10 Maintenance Star Schema.....	14
Figure 11 Financial Star Schema	15
Figure 12 Marketing Star Schema	16
Figure 13 Conformed Dimension MetaData Tables and Master Data Tables.....	16
Figure 14 Sales Data Mart	17
Figure 15 Sales Data Mart	18
Figure 16 Lease or Rental Data Mart.....	19
Figure 17 Maintenance Data Mart.....	20
Figure 18 Finance Data Mart	21
Figure 19 Transformation Processes	22
Figure 20 ETL flowchart	23
Figure 21 OLTP Structure	27
Figure 22 Data Warehouse Structure.....	27
Figure 23 Data Quality Life Cycle	30
Table 1 Internal and External Data Sources.....	10
Table 2 Source/cleanse/destination	24
Table 3 Data Mart Examples	28
Table 4 Data Cleansing	29
Table 5 Sensitive Data	30

1 Introduction

This report comprises of 2 tasks. The first will demonstrate the differences between an Online Transaction processing database and a data warehouse, the architecture of a data warehouse and the use of data marts to cater to the differing requirements of the directorates. The second task designs suitable star schema and data marts for Cambrian referencing potential data sources and finally explains the development of a simple extract, transformation and load plan for the data warehouse. Additionally, the report will address how Cambrian can assure data compliance and trustworthiness by implementing data quality and governance processes.

2 Task 1

2.1 The benefits of a data warehouse over an Online Transaction Processing (OLTP) system.

The present OTPL system has served Cambrian well over the years as it can manage rapidly changing data and it has the functionality to support day-to-day operations, providing information on:

- i. Customer transactions
- ii. Property listings
- iii. Property management
- iv. Leases
- v. Financial transactions
- vi. Staff

It does have a few issues that need to be resolved if Cambrian is going to expand. Kimball [1] states that OLTP systems need to track numerous separate transactions daily and encourage the use of very similar processing requests, according to Bhatia [2] processing is repetitive and it only allows for one record to be processed at any one time. E.g. a record of a customer who might be phoning a branch requesting to view a property. OLTP does not have sufficient processing power to perform complex analysis on clients and property records although it does cater for a high level of transaction throughput when processing simplistic queries. ([OLTP Structure](#))

Transitioning to a data warehouse approach will support business intelligence to:

- i. Improve analytical capabilities allowing analysts to perform complex analytics to uncover valuable insights and trends that are buried within the data, e.g. market analysis, creating predictive pricing models and developing demand forecasting to identify investment opportunities.
- ii. By consolidating historic data over time the data warehouse can provide a comprehensive view of Cambrian's business operations to determine property values over time, analyse market trends for properties in all regions and improve client services.
- iii. Enhance scalability and performance as the data warehouse will optimise query techniques to ensure efficient processing of all queries, especially complex analytics, improving client services.
- iv. By utilising centralised data storage and standardised data models the data warehouse can simplify regulatory compliance and meet reporting needs by ensuring accurate and consistent data retrieval/analysis across departments and branches.
- v. To create data marts for efficient analysis and report creation. ([Data Warehouse Structure](#))

2.2 Data Warehouse Architecture

It is important to understand what a 'Data Warehouse' actually is and how it is structured, Inmon, one of the founding fathers of data warehouse design states that:

Data warehouse architecture refers to a subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision-making process. [3]

It is equally important to understand what role a data warehouse house can perform for Cambrian and what advantages implemented one will deliver. Rivera [4] who describes himself as an Analytics Evangelist describes the function of a data warehouse as a logic service that provides a structured and coherent method of storing, managing, manipulating and retrieving vast amounts of data.

When deciding on what level of data warehouse to implement Hammergren and Simon [5] suggest that the cost of implementation and level of complexity required should be deliberated as different levels can be employed:

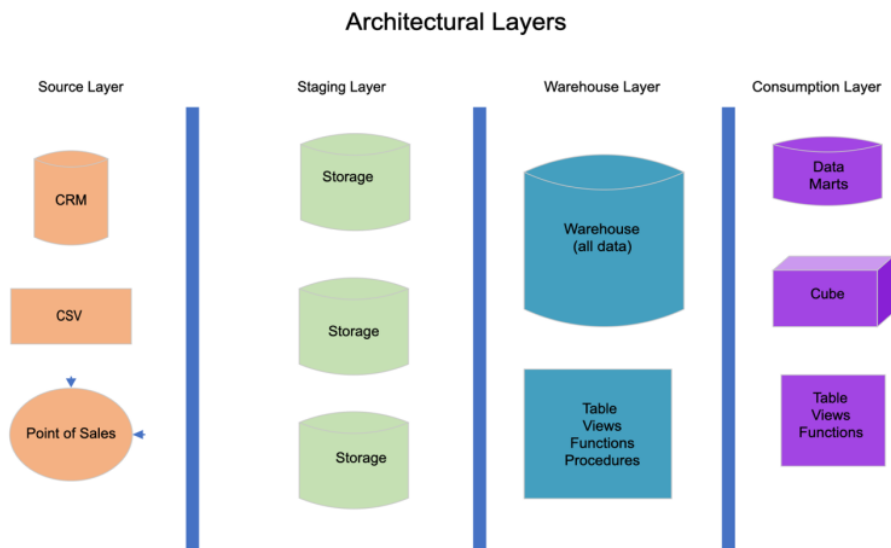
- i. Lite
Minimal investment and suitable for smaller companies – so not suitable for Cambrian.
- ii. Deluxe
Mid-sized investment, enterprise scope which aims to assist a company to integrate its business needs into a data warehouse – Cambrian's complex analytic requirements have surpassed this model
- iii. Supreme Enterprise does come at a higher cost but will permit Cambrian to drive strategy and integrate business intelligence, benefiting both internal staff and stakeholders whilst providing improved client service.

2.2.1 Architectural Layers

Within a data warehouse there are several functional layers:

- i. Source Layer feeds into the data warehouse
- ii. Staging Layer – landing area for source data, here data is uncleansed, unstandardised, unmodelled, ungoverned or verified.
- iii. Warehouse Layer – physical schema, tables, views, stored procedures and warehousing functions.
- iv. Consumption Layer – analytics layer.

Figure 1 Data Warehouse Architecture



Created in Visio

CRM – Cambrian's Software solution to manage sales/marketing and customer support

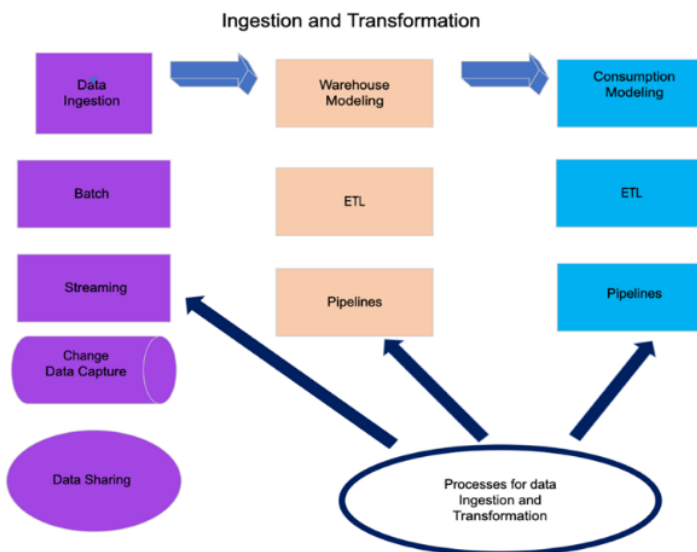
2.2.2 Data ingestion and transformation

Data ingestion is the process of getting data into the source layer from different systems, files and records, of which there are several categories:

- i. Batch Processing.
One of the most common approaches, it handles data in large chunks at specified times. Drawbacks include data latency, inflexible scheduling, intense costly CPU and disc usage and integration difficulties.
- ii. Data Streaming.
Real-time processing and it is perfect for applications that require timely responses and for real-time analytics.
- iii. Change Data Capture.
Works to ensure that only modified data is ingested by using the database transaction log to capture any created, modified and/or deleted. This methodology is great for synchronising operational data stores and staging data from the transactional systems.
- iv. Data sharing.
Utilises cloud providers like Snowflake or DataBricks and does not require data duplication and would provide Cambrian with real-time data sharing, to improve data sharing and communications between Cambrian's branches.

Each ingestion method would need to be evaluated by examining data sources and Cambrian's business requirements to ensure the right methodology is chosen to provide the most cost-effective and efficient implementation method.

Figure 2 Ingestion and Transformation



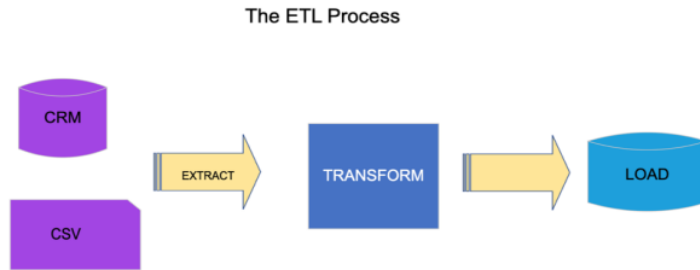
Created in Visio

2.2.3 Data Warehouse Architecture Components

2.2.3.1 Extract, Transform, and Load (ETL)

This process moves, transforms and loads data, and it can according to Sabtu et al, [6] be used by data warehouses to manoeuvre operational structuring and transactional data. ETL uses a defined set of rules that can be set up by data warehouse teams to clean and organise raw data from source systems and prepare it for storage, data analytics and even machine learning to discover new patterns and trends.

Figure 3 ETL Process



Created in Visio

2.2.3.1.1 Extract

Extracts source data from internal system(s) or external source(s) to be loaded onto a receiving database or server, often referred to as a staging area.

2.2.3.1.2 Transform

The next step is the transformation of data in the staging area, this could involve an aggregation or mapping process depending on the business rules or requirements.

2.2.3.1.3 Load

Data can be loaded into either smaller databases, tables or data marts depending on Cambrian's analytic and reporting requirements. The Finance Department might need to report on the effectiveness of a marketing campaign in one of the regions, this could be done by linking property sales and marketing tables to generate a data mart with historical sales data, with data during and after the campaign.

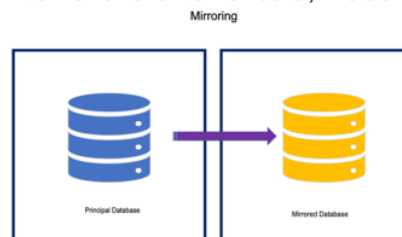
2.2.3.2 Data Warehouse processes

Figure 4 Mirroring

The data warehouse will be able to source all of Cambrian's internal and external data, including data created within the CRM platform and external files uploaded directly from Excel CSV (spreadsheet) files or through SQL-2-SQL upload.

These processes can be automated to run at regular intervals using Microsoft SQL Server Integration Services (SSIS) [7], a database software used to perform data migration tasks. Software packages like SSIS are capable of database mirroring which Cambrian could utilise to ensure data/information availability to all end-users. It is important to

ensure that business rules used to dictate how the ETL runs are robust, and a Validation at the Source process is created to prevent errors in data sources and ETL pipelines from corrupting Cambrian's data.



2.3 Data Marts

2.3.1 What is a Data Mart?

Cambrian needs up-to-date information to make predictions across the UK and the Republic of Ireland. Scriney et al [8] explain how information processed in Cambrian's data warehouse can be used to create Data Marts to perform complex analytics.

Oracle [9] documents that a Data Mart is a form of data focused on a single theme, subject or department. Implementing data marts will give the Finance, Sales and Marketing teams access to data specific to their business requirements. This will speed up information queries, analytics and decision-making as it removes the requirement to search the whole of the data warehouse or manually aggregate the data from several sources.

2.3.2 Purpose of a Data Mart

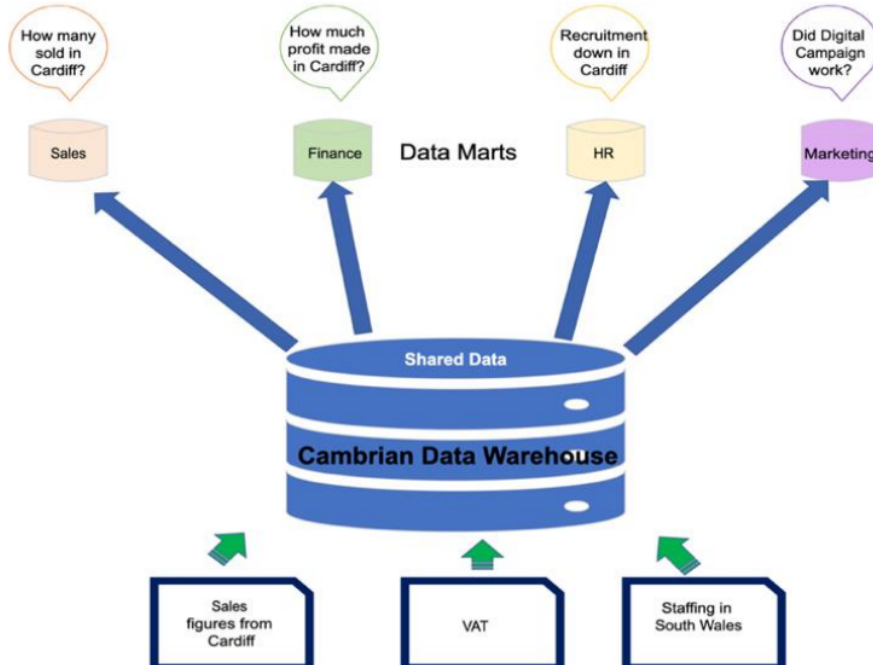
- i. To provide departments with access to data they use most often.
The Sales Department would need information on properties currently for sale/lease, however, this information would not be useful to the Human Resources (HR) department, so a data mart could be built just containing this information solely for the Sales Department.
- ii. Improved response time.
The data mart will be smaller and specific to the Department and will access a reduced volume of data, making the process more streamlined and user-specific. This will save processing time and reduce running costs.

[Table of example possible Data Marts](#)

2.3.3 Data Mart formation and operations

The data warehouse is where data sources are consolidated and instead of users having to extract their information from the warehouse, they can go straight to their departmental data marts. The data marts would order/sort the data from the warehouse and data would be available for the users to analyse.

Figure 5 Data Mart



Created in Visio

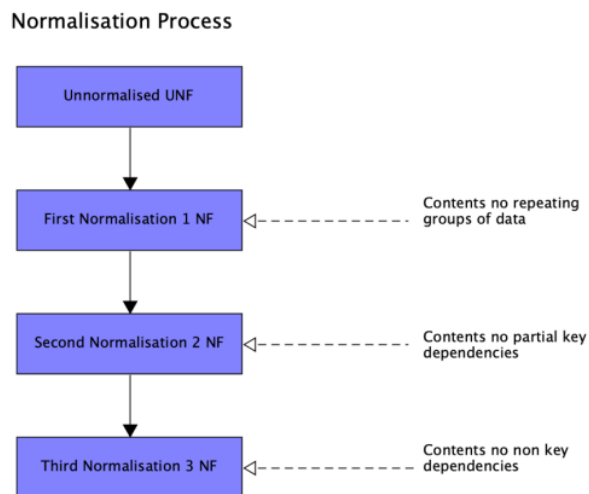
2.4 Difference between a 'Normalised' OLTP system and a 'Dimensional' structure of a Data Warehouse

The Company's normalised OLTP system is designed primarily for transaction processing, and it focuses on data integrity, consistency, and relegating data redundancy. The data is organised into multiple related tables and diminishes redundancy by normalising the data ensuring that data is stored only in one place. However, this methodology is not optimal for the type of complex analysis that Cambrian now needs, as it requires complex joins across multiple tables to retrieve all the necessary data. This can lead to performance issues as the queries can take a significant amount of processing time and resources to run.

Conversely dimensional warehouse models are built and optimised for complex analytical queries and reporting. They employ a de-normalised schema and are comprised of fact, dimension, aggregate, master data and metadata tables [10].

2.4.1 Normalised Model

Figure 6 Normalisation Process



2.4.2 Dimensional Structure

2.4.2.1 Fact tables

These contain the core metrics or measures of organisational performance:

- Property sales
- Property leases
- Property viewing
- Advert costs
- Property Maintenance/Repairs
- Wages
- Outgoings

Dimension tables surround fact tables and typically contain the foreign keys that are primary keys in linked Dimension tables.

2.4.2.2 Dimensional tables

These tables provide the context and descriptive attributes for data in fact tables. They represent how data can be sliced and diced for analysis. An example of a dimension table would be a property table which would contain:

- Property_ID – primary key for the dimension table and the corresponding foreign key in the fact table

- Property_Type
- Property_location
- Property_size
- Property_status – for sale or lease

One thing to note about the dimension tables is that they can join to different fact tables though all the fields must be represented and consistently formatted in both tables.

2.4.2.3 Metadata tables

Tables containing data about data, that is information about data that makes auditing and working with the data easier. E.g. where the data is sourced, where it is sourced, and what it is called. [Business rules](#) can also be part of metadata held by the company, stating what each attribute contains and a brief description of data held in each of the attributes.

2.4.2.4 Aggregated tables

Aggregation of data in act tables, using functions like average, count, sum, min and max.

2.4.2.5 Master data tables

Collects the definite collection of historical data that can be used to develop a Risk Model for Cambrian.

2.4.2.6 Star schemer

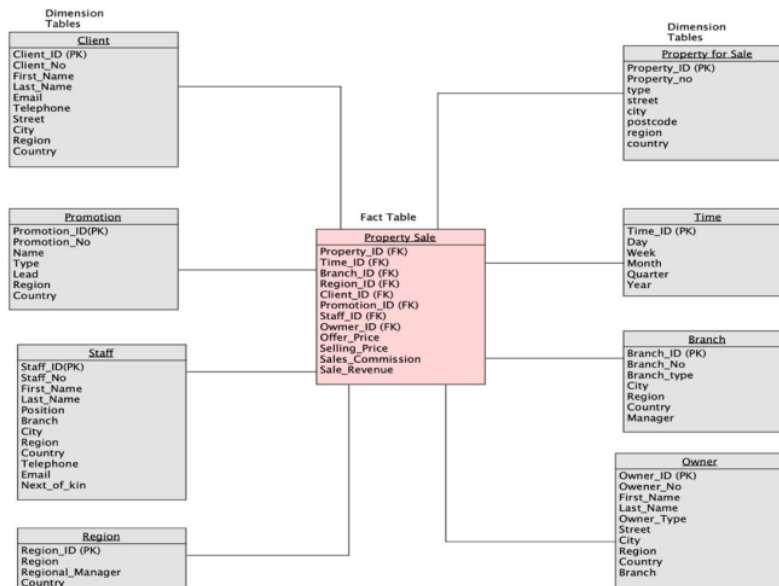
A star schemer according to Databricks [11] is a method of displaying data in an organised multi-dimensional model so that 'humans can understand and analyse the data. Star schemas can be applied to:

- Data warehouses
- Databases
- Data marts

They are exceptional for analysing large data sets so they will be perfect to meet Cambrian's data complex data analytic and reporting requirements.

Dimension tables within a star schema can be attached to several different fact tables, however, their format and specifications must be common to all the fact tables they pertain to.

Figure 7 Star Schema



3 Task 2

3.1 Internal and External Sources

For the data warehouse to be a success there is a need to plan track and document all the different sources of information that will make up the main data feeds. There are 2 different types of data sources:

- i. **Internal Data Sources**
According to Ventiv [12] internal data comprises data and information that will come directly from Cambrian's systems and are specific to Cambrian's requirements.
- ii. **External Data Sources**
External data comes from systems or organisations which are external to Cambrian, it will however provide Cambrian with information about the competition, and economic and social trends. It will also provide Cambrian with legal documents that the company needs to provide the best possible service to its clientele.

To ensure that the planning of the data sources is comprehensive a table containing both the internal and external sources that will feed into Cambrian's new data warehouse has been compiled.

Table 1 Internal and External Data Sources

Source	Description	Internal or External
Properties listed database and lease agreement database	Information about the current properties that Cambrian has on its books. Which would include sale/lease/rent, location, size, price, amenities and what type of property it was commercial, flat, detached. It would also have a field to identify the owner(s). For lease/rentals there would need to be information on length of lease/rental.	Internal – generated on Cambrian's Enterprise Resource Planning system through the sales/rental teams
Transaction Record	Containing information on the property transactions, including sale/rental/lease price, date of transaction, buyer and seller/owner details and a link to the property listing. It would also include information on profit margins.	Internal from the ERP through the sales/rental teams
Customer Record Database	Details about potential and substantive buyers/owners/renters, including contact information and for buyers/renters' preferences	Internally collected and managed through Cambrians customer relationship management system CRM
Financial Record Database	Relating to property transactions – invoices, receipts, payment records and any legally binding financial data – VAT paid. Would also contain information pertaining to the costs of running the company – staff, equipment, offices	Created internally by the finance team using Cambrian's
Maintenance and repairs record database	Maintenance activities – any repairs carried out on leased/rental building on behalf of clients	Generated internally by the maintenance team using maintenance logs and Cambrian's work order system
Staff records database	Staff's personal details, plus employment details, Tax Codes, starting/end dates, Personal development reviews, line manager, and performance review.	Generated internally by the human resource team using their bespoke staff management system
Marketing and advertising database	Reports analysing the market trends, property prices, demand supply dynamics and forecasts for UK and Ireland. The marketing will need to link in with finance as there will be a cost to any external	Internally generated by the marketing team but some of the data will come from external marketing surveys

	surveys or market research performed outside of the team. Feedback on how advertising campaigns had increased or had no effect on company income.	and from social media and websites.
Government Databases	Land registry, local authorities, planning departments to prove ownership of property and council tax and property tax assessments	External sources by the legal team.
Real estate portals – competitor data	Online estate agents like Zoopla and Rightmove, where multiple agents list their properties	Externally sourced by the marketing team to discover like properties for sale/rent for market trend analysis and current property valuation
Legal documents	Relating to ownership, property deeds	Externally sourced by the legal team
Property Surveys	Information on property condition, valuation, structure and geographical information which is especially useful in ex-mining communities.	External sourced by the legal team from qualified surveyors.
Census data	Demographical information to enrich client tables	External sourced used by both the marketing and sales teams

3.2 Star Schema

Figure 8 Property Sales Star Schema

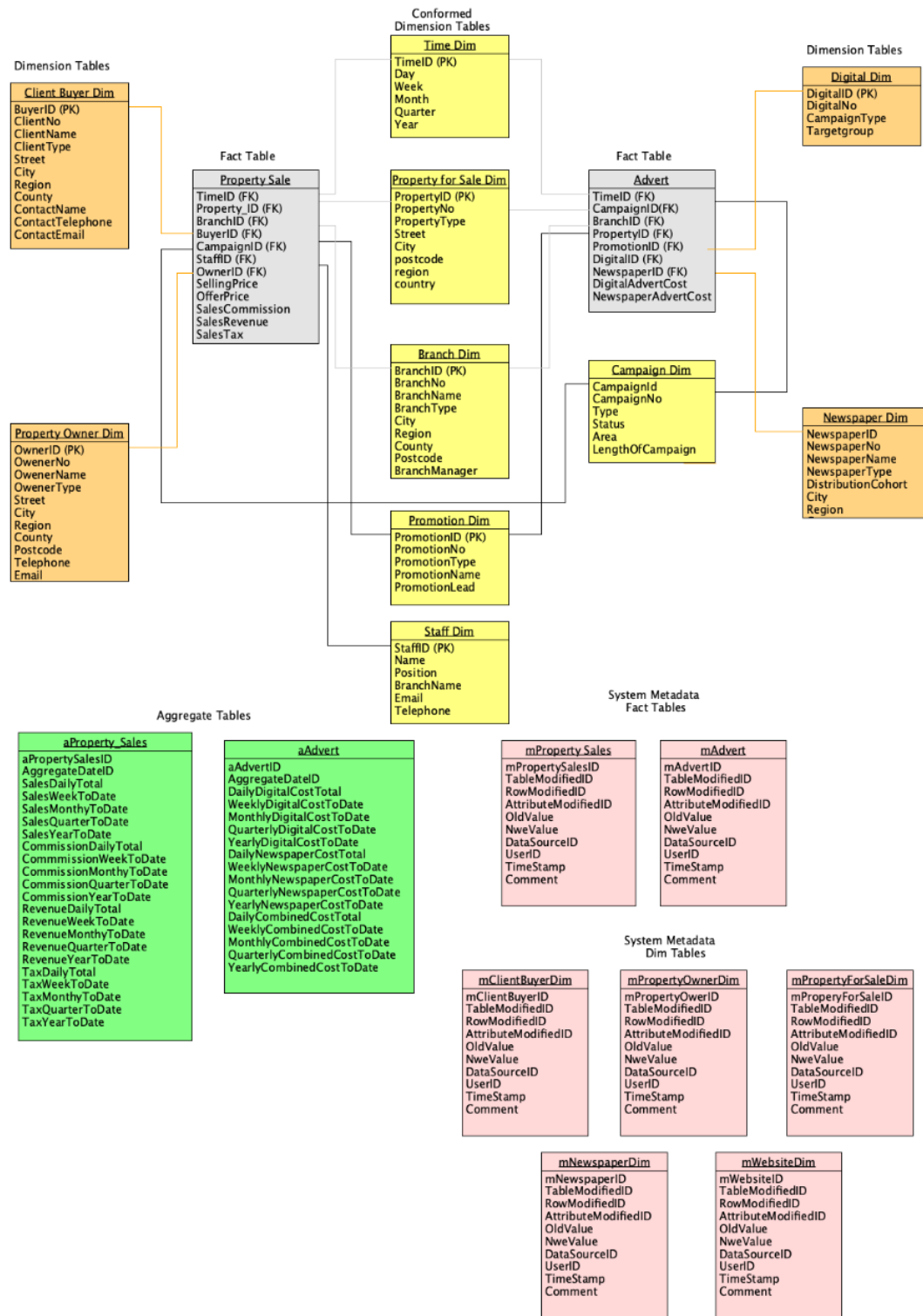


Figure 9 Property Lease/Rental Star Schema

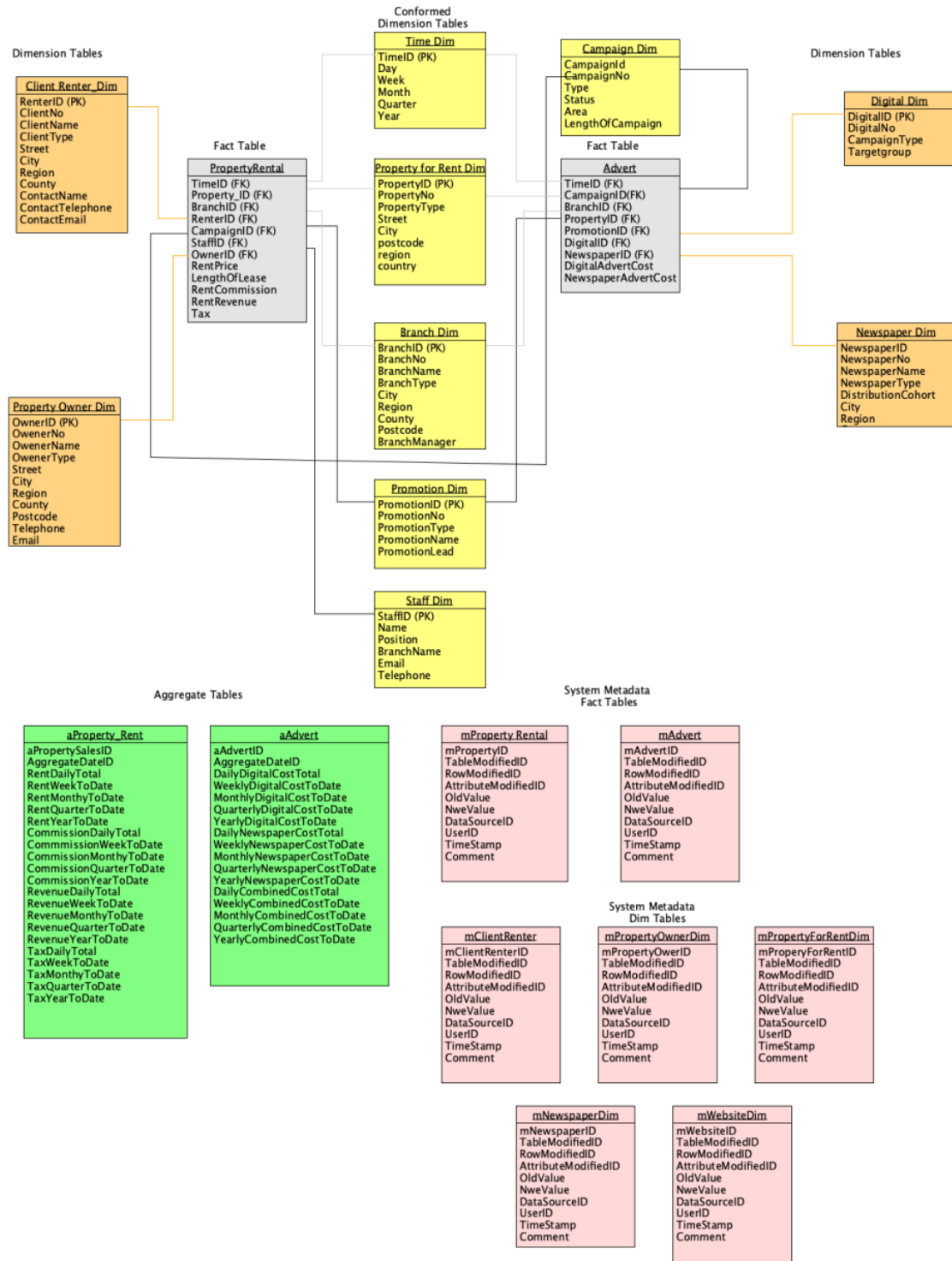


Figure 10 Maintenance Star Schema

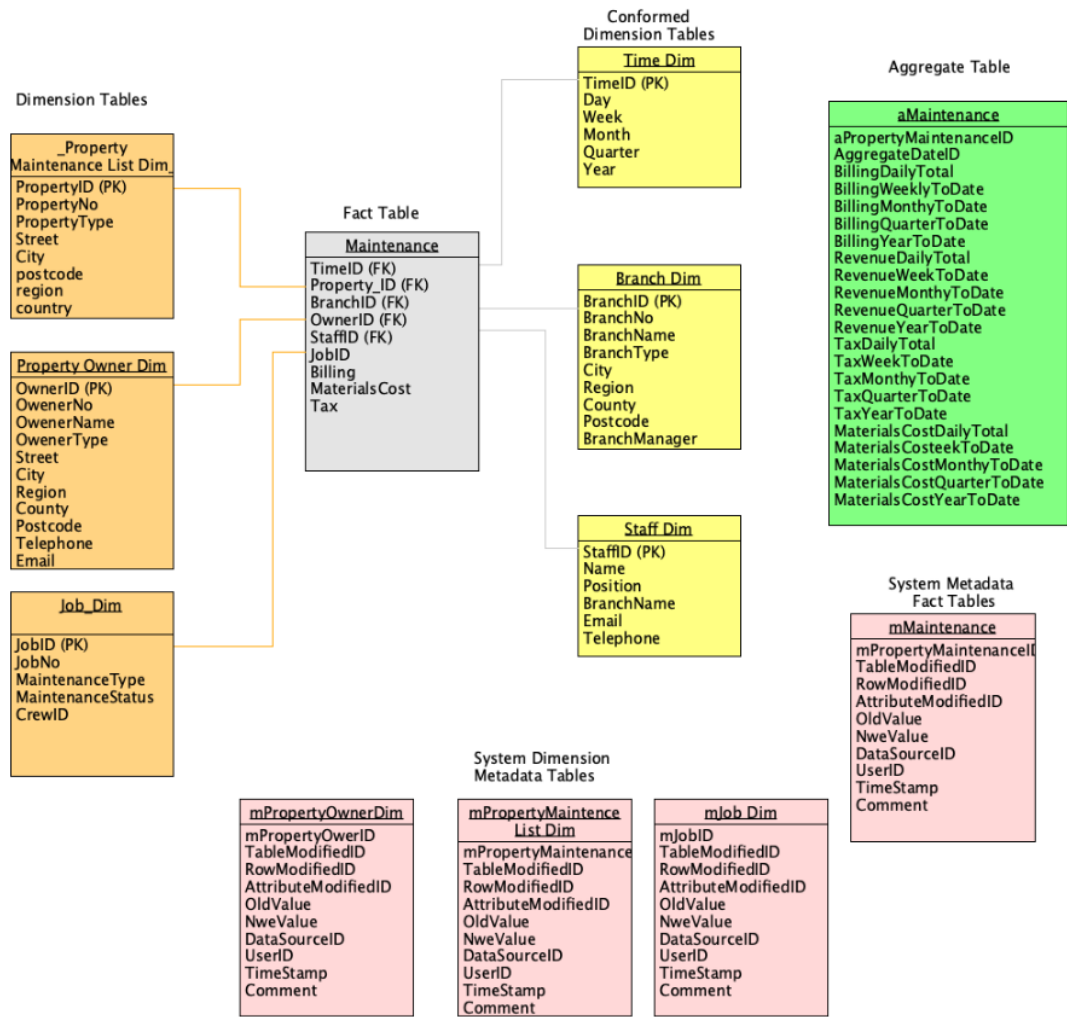


Figure 11 Financial Star Schema

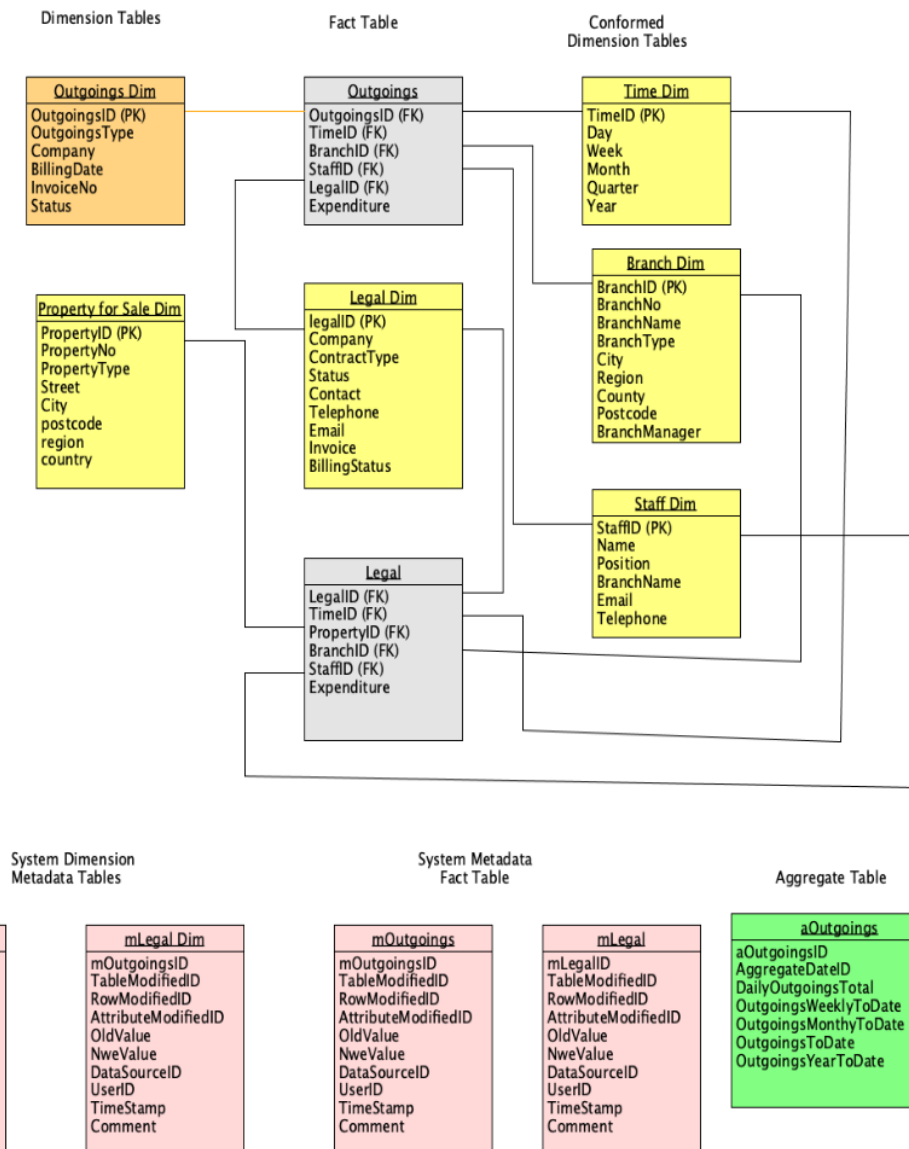


Figure 12 Marketing Star Schema

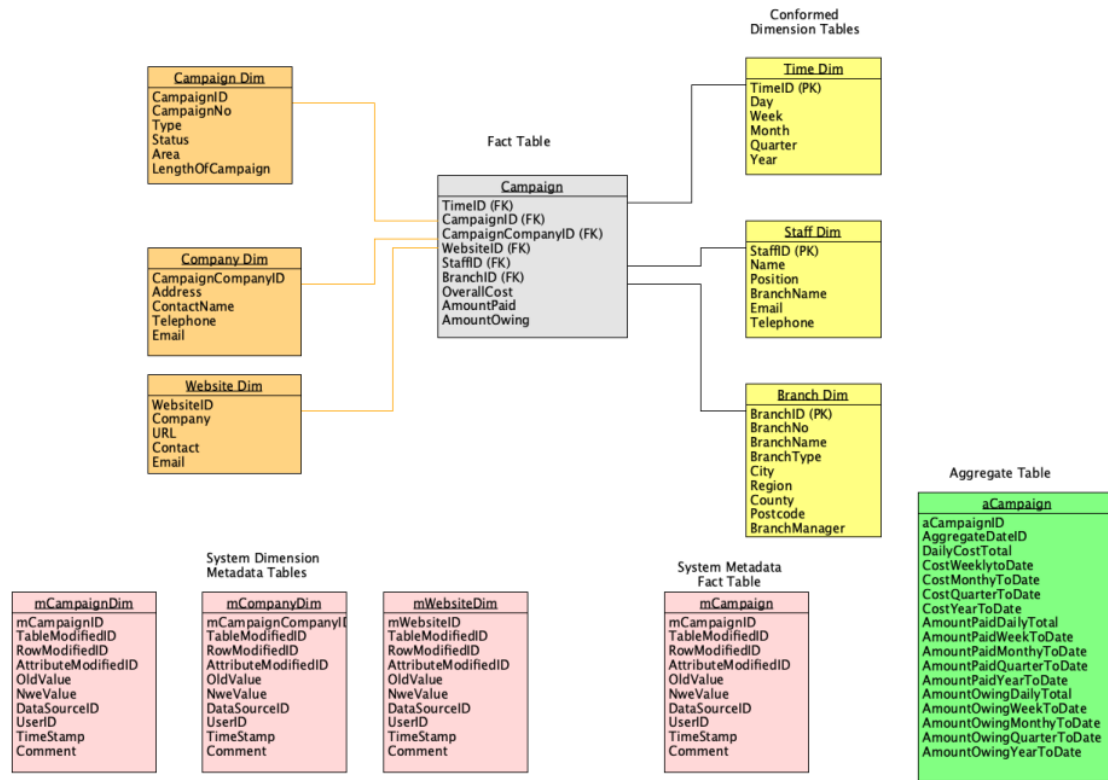
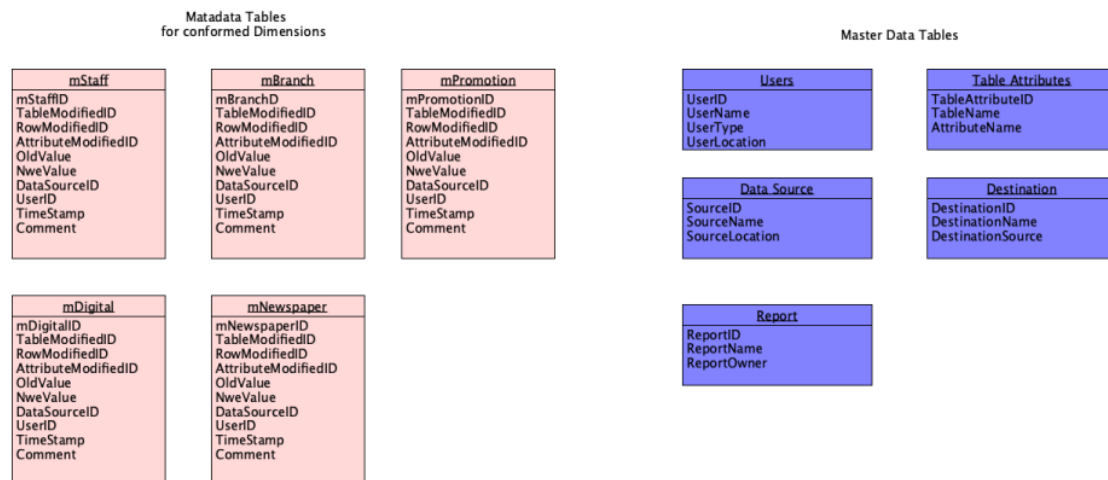


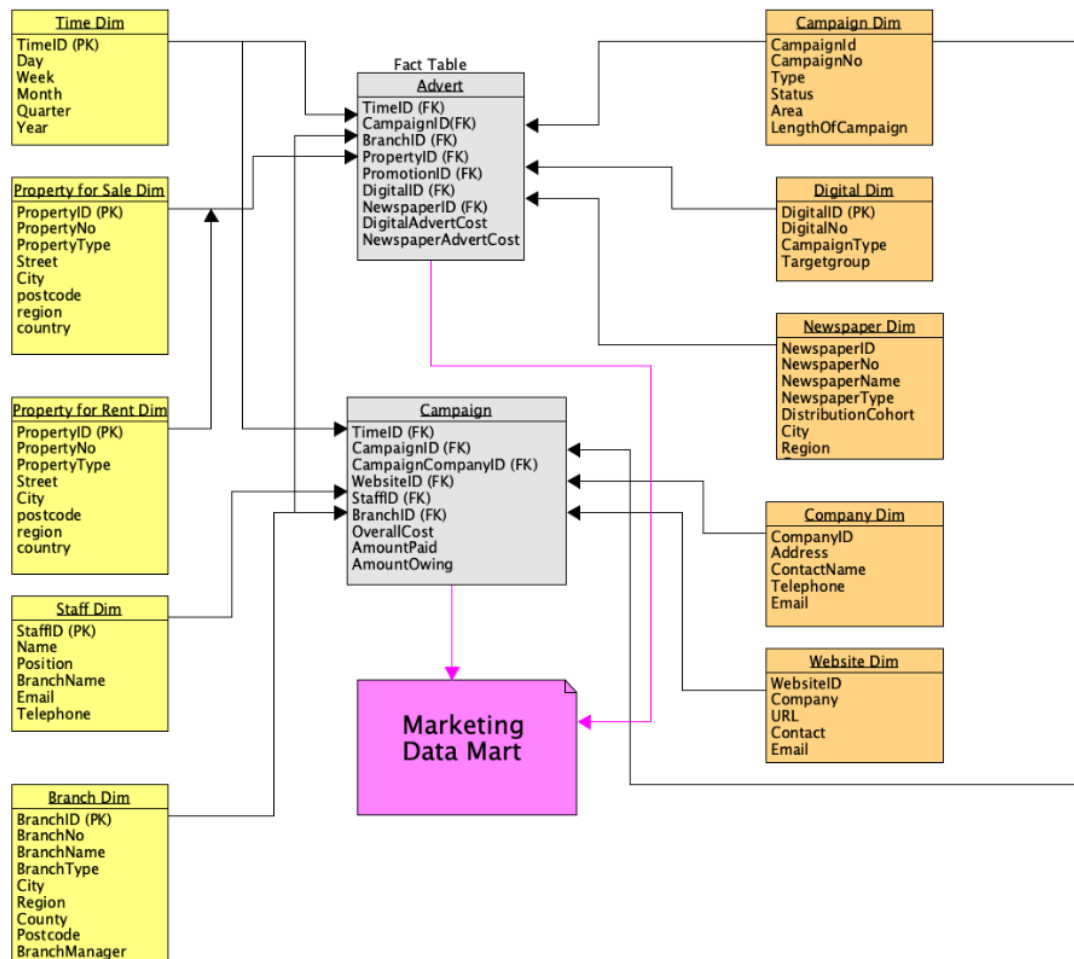
Figure 13 Conformed Dimension MetaData Tables and Master Data Tables



3.3 Data Marts

Data marts in a data warehouse perform a similar function as a database view [13], with the benefit of focusing on a single subject or departmental operation. Using data marts will provide Cambrian directors and their teams with access to data they need to gain insights quickly and without having to perform complex searches on the whole data warehouse. The data warehouse can hold several different data marts to meet the differing requirements of all the directorates within Cambrian. All of the data marts created for Cambrian will hold information about the staff and branches involved with each transaction. The Company Director would have access to a summary of all the Data Marts which then could be disseminated to the board members.

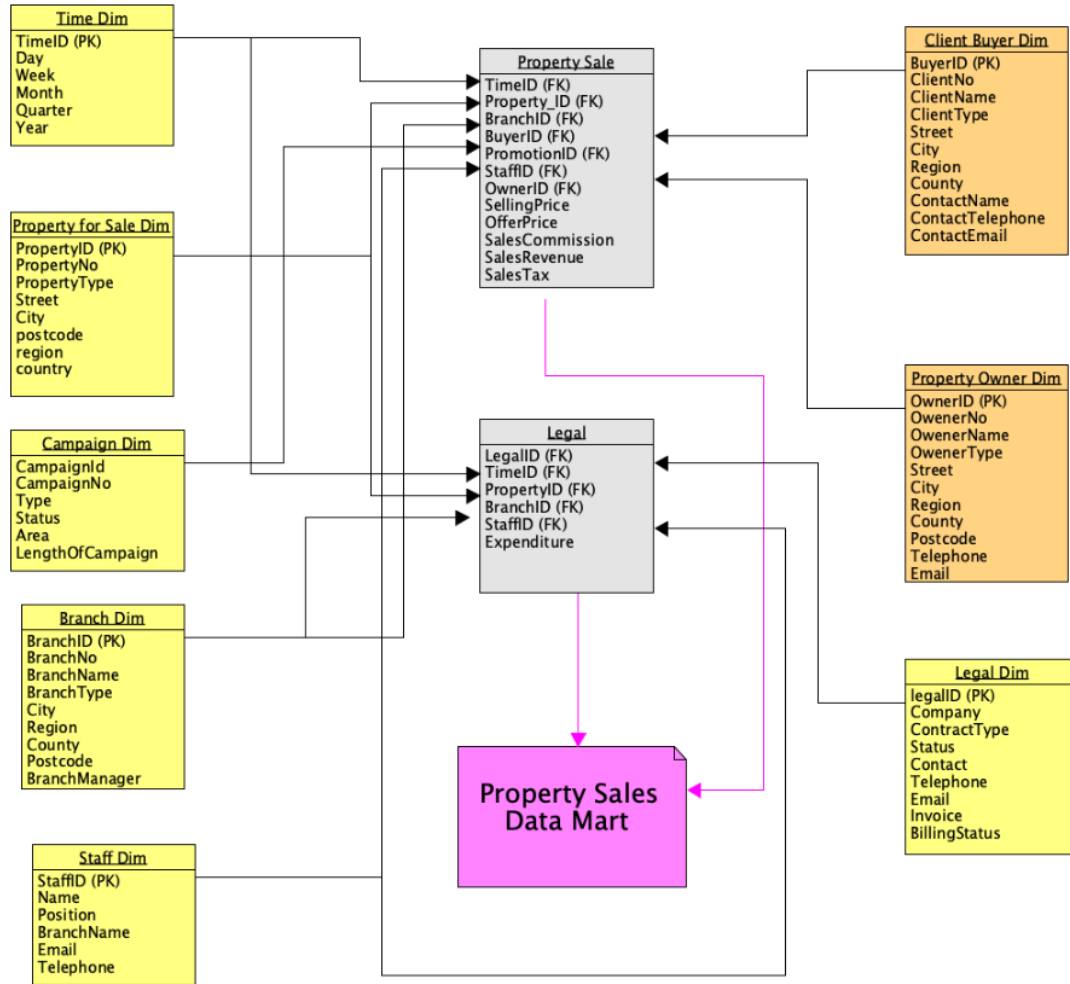
Figure 14 Sales Data Mart



The marketing director will need a data mart to be created that will contain all the data about the activity within their department, which will include all adverts that have been placed digitally and in newspapers, plus any campaign's past, present and those in the pipeline to be run in the future. With this information, they will be able to analyse the performance and impact of all the campaigns allowing the department to make strategic short-, mid- and long-term plans to increase Cambrian's exposure in the property market. The Website dimension table will include competitor data so that the marketing team can keep a close watch on how the main competition is performing in specific regions.

3.3.1 Sales Data Mart

Figure 15 Sales Data Mart



The marketing director will need a data mart that provides information about Cambrian's marketing activities, including:

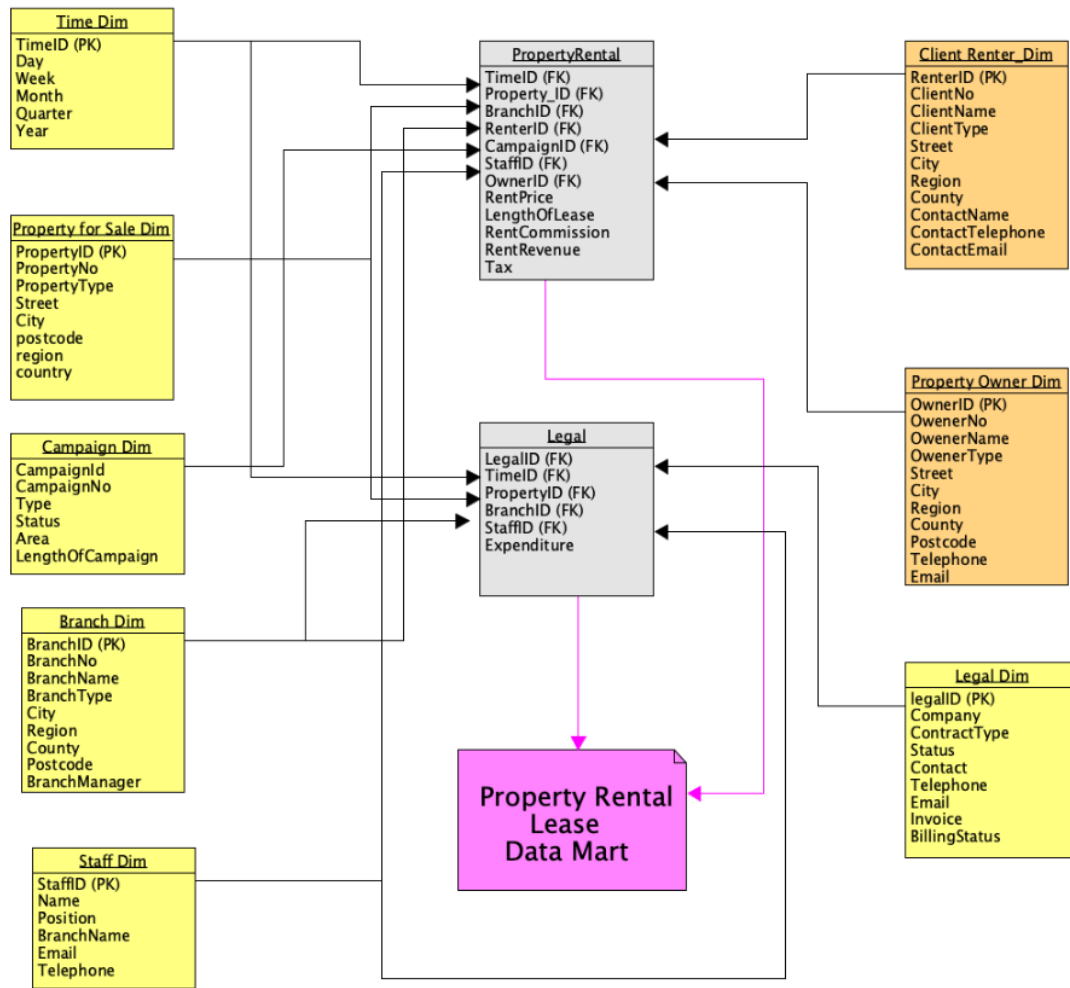
- i. All digital and print advertisements
- ii. Past, present, and future marketing campaigns
- iii. Campaign performance and impact data

Providing them with information to perform:

- i. Strategic short/mid/long term planning
- ii. Monitor performance against competitors
- iii. Enhance Cambrian's presence in property market

iv. Rental/Lease Data Mart

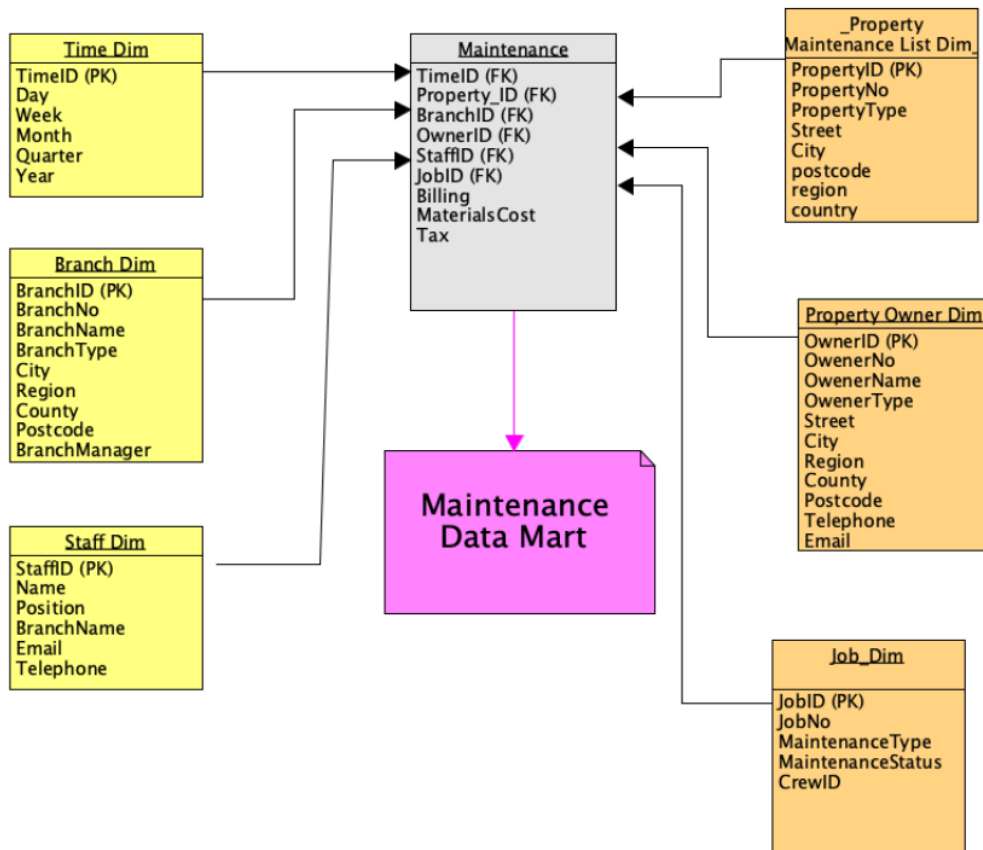
Figure 16 Lease or Rental Data Mart



The directorate in charge of rentals and leases will need the same information as that of the Sales directorate with the slight difference that they will need information on the current and past lease/rental agreements.

3.3.2 Maintenance Data Mart

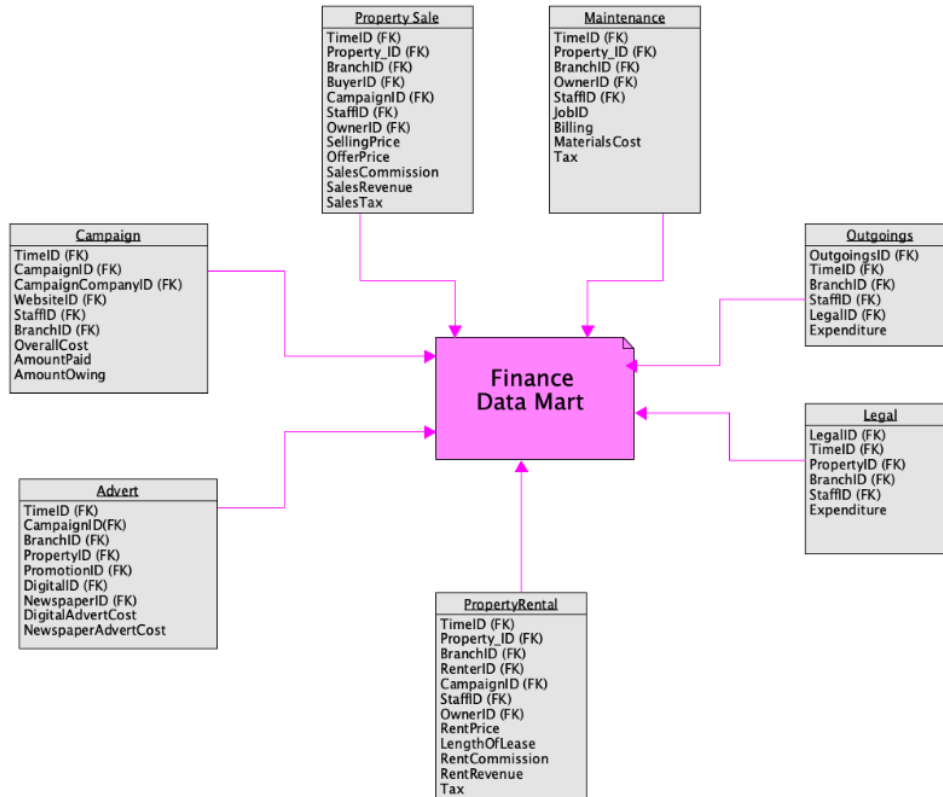
Figure 17 Maintenance Data Mart



This data mart ensures that the property manager has up-to-date information on all the properties that need repairs and what crew has/will be working on them.

3.3.3 Finance Data Mart

Figure 18 Finance Data Mart



The finance director will need a data mart that shows the financial position of Cambrian, with the functionality to drill down to Countries, Regions, Cities and branches. This data mart will track all the income and outgoings of the company. This data mart will provide the Financial Director and the board with all the information they need to make long-term plans for the Company.

3.4 ETL Plan

The creation of an ETL plan for Cambrian is essential to ensure the efficient transfer of data from source systems into the data warehouse or staging destination whilst ensuring data is consistent. It complies with all the business rules in metadata tables. The ETL plan involves three stages [14] :

- i. Extraction
Identifying which source system data is coming from.
Employment of data extraction tools like SQL Server Management Studio or other extraction software to extract data from source systems.
Data is 'profiled' to understand its structure, quality and relationships.
- ii. Transformation
Cleansing data by removing duplicates and special characters, correcting errors and standardising formats.
Transforming data into a consistent format suitable for analysis and reporting.
Data Enrichment enhances data by adding calculated fields, aggregation or joining with other datasets.
- iii. Loading
Destination selection to choose the target destination into which to load transformed data.
Loading transformed data into destination tables using either a bulk load technique or an incremental update.
Validating the loaded data to ensure accuracy and completeness before making it available for analysis.

Figure 19 Transformation Processes

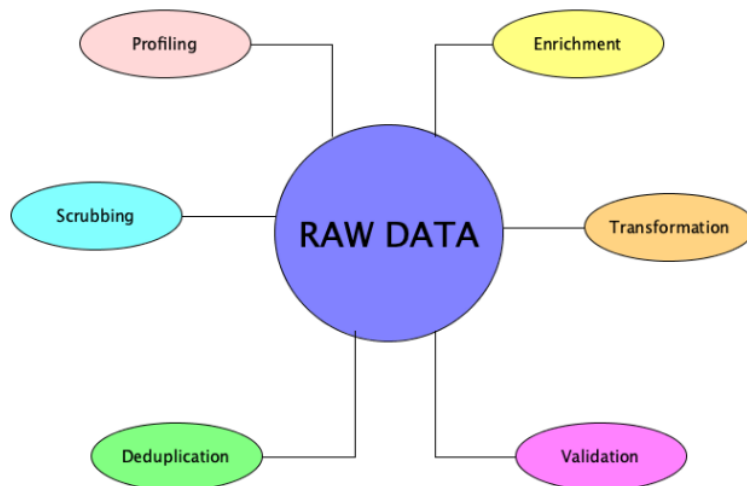
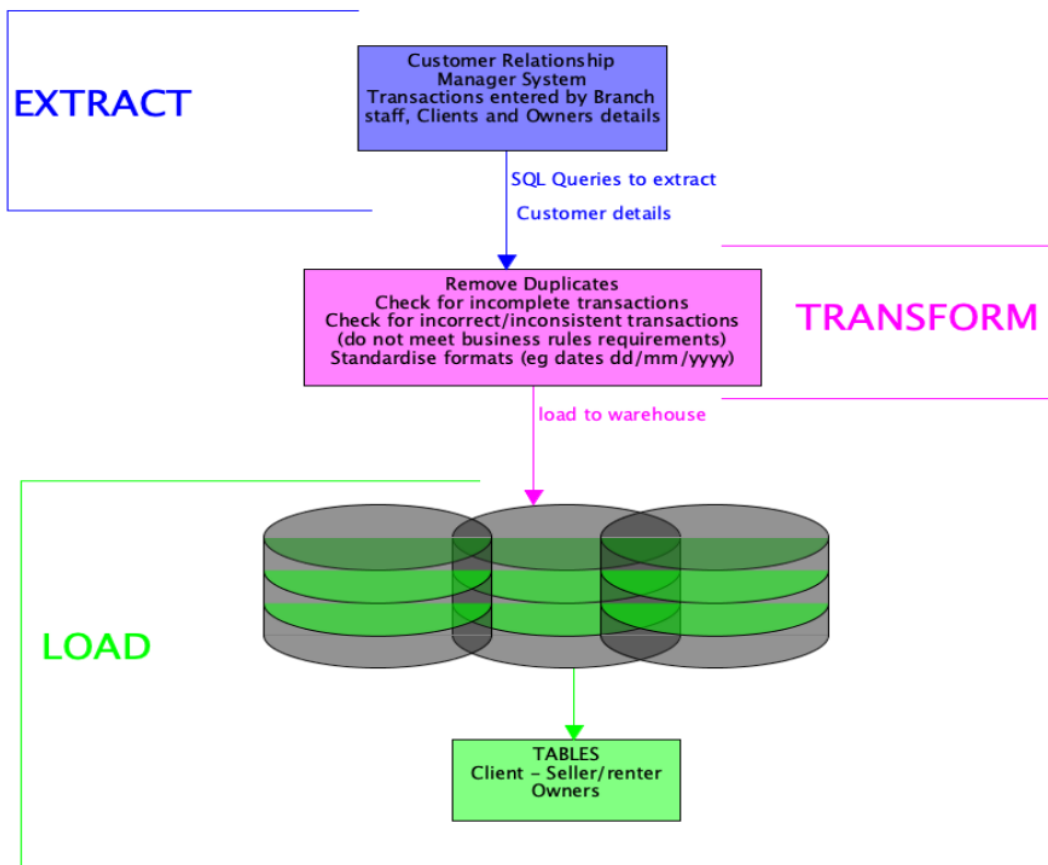


Figure 20 ETL flowchart

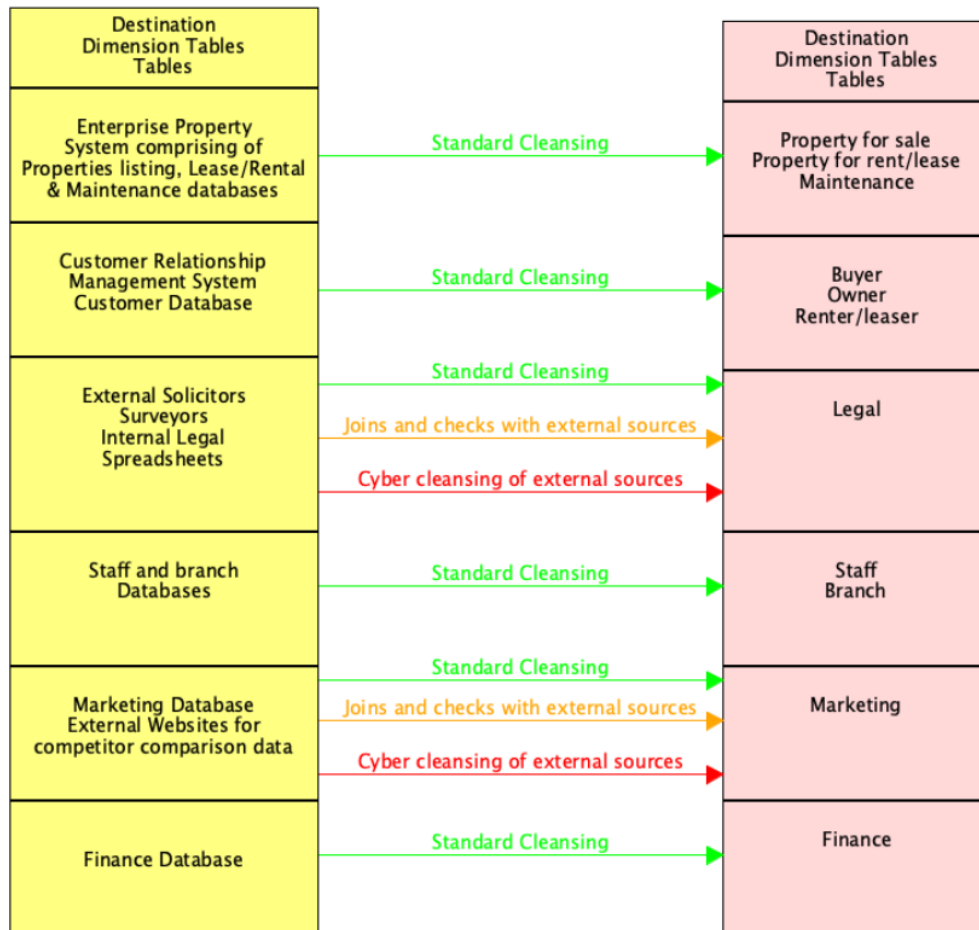


3.4.1 Data Source/cleansing and destination

All customer data will be sourced from the Customer database (CRM), which is updated by branch staff. However, before the transactions are loaded to the data warehouse's 'client' and 'owners' dimension tables they will need to be cleansed and transformed. [15]

All the data will go through a similar transformation process before being loaded to their corresponding dimension tables. However, any dimension tables that are receiving data from external sources will need additional checks to ensure data validity and that files are free from any threats.

Table 2 Source/cleanse/destination



3.5 Data Governance and Data Quality within ETL

Effective data warehouse processing is dependent on ensuring high-quality data throughout the entire [data quality lifecycle](#) from extract to analysis. During the ETL process, there are several important stages of data quality and data governance that Cambrian must consider [16]

3.5.1 Data Quality

- i. Data profiling is vital to identify data quality issues, e.g. missing values, inconsistent formats
- ii. The implementation of robust data validation (business) rules and checks during the ETL process to ensure the accuracy, completeness and consistency of data being loaded into the data warehouse
- iii. The data transformation (mapping) needs to be carefully designed and thoroughly tested to maintain data integrity and prevent invalid data from being loaded into the warehouse.
- iv. To understand data provenance it is necessary to capture and manage metadata which should include data lineage, definitions and transformation rules.
- v. The establishment of clear error handling, data remediation and exception management will help reduce the impact of data quality issues within the warehouse and identify potential issues promptly.
- vi. Continuous data monitoring and the reporting of any data quality issues.

3.5.2 Data Governance

- i. Implementation of a comprehensive data governance framework which clearly defines roles, responsibilities and policies to ensure data quality and appropriate use of the data.
- ii. The appointment of data stewards to monitor, maintain and improve data quality to ensure sustainable data quality.
- iii. Data monitoring and auditing to implement data quality metrics and identify any potential data quality issues.
- iv. Data lineage and audit trails to assess and comprehend data origins, transformation and usage which are vital for data assurance, trust and regulatory compliance.
- v. Data security and privacy to ensure that appropriate access controls and data privacy are in place for [sensitive information](#) and comply with the General Data Protection Regulation [17]

By addressing all the data quality and data quality considerations Cambrian can ensure the reliability, assurance and appropriate use of their data assets.

4 Conclusion

The implementation of a data warehouse will provide Cambrian with the opportunity to enhance its decision-making capabilities, improve operational efficiency and gain an advantage over its competitors. By utilising the power of dimensional modelling and data marts to provide targeted high-quality information to the board and stakeholders. The star schema design, data marts and ETL process will ensure the appropriate use of Cambrian's data assets and will empower the leadership to be able to make more informed decisions and long-term plans.

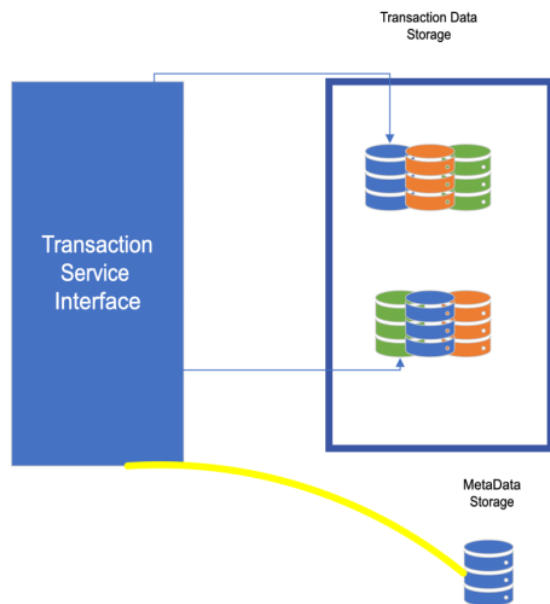
5 References

- [1] R. Kimball and M. Ross, *The Kimball Group Reader*, Indianapolis: Wiley, 2010.
- [2] P. Bhatia, *Data Mining and Data Warehousing*, Cambridge: Cambridge University Press, 2019.
- [3] B. Inmon, *Building the Data Warehouse*, Indianapolis: Wiley, 2005.
- [4] S. Rivera, "The Fundamentals of data warehouse architecture," *Thoughtspot*, 16th September 2023. [Online]. Available: <https://www.thoughtspot.com/data-trends/data-modeling/data-warehouse-architecture#>. [Accessed 1st March 2024].
- [5] T. Hammergren and A. Simon, *Data Warehouseing for Dummies*, Indianapolis: Wiley, 2009.
- [6] A. Sabtu, N. Azmi, N. Sjarif, S. Ismail, O. Yusop and Sarkan, "The Challenges of Extract, Transform and Loading (ETL) system implementation for near real-time environment," in *International Conference on Research and Innovation in Information System ICIIIS*, Langkawi, 2017.
- [7] "SQL Server Integration Services," Microsoft, [Online]. Available: <https://learn.microsoft.com/en-us/sql/integration-services/sql-server-integration-services?view=sql-server-ver16>. [Accessed 2nd March 2024].
- [8] M. Scriney, S. McCarthy, A. McCarren, P. Cappellari and M. Roantree, "Automating Data Mart Construction from Semi-structured Data Sources," *Computer journal*, vol. 62, no. 3, pp. 394-413, 2019.
- [9] "What is a Data Mart," Oracle United Kingdom, [Online]. Available: <https://www.oracle.com/uk/autonomous-database/what-is-data-mart/>. [Accessed 3rd March 2024].
- [10] W. Inmon and R. Hackathorn, *Using the Data Warehouse*, New York: John Wiley & Sons, 1994.
- [11] "What is a star schema," Databricks, 2024. [Online]. Available: <https://www.databricks.com/glossary/star-schema#:~:text=What%20is%20a%20star%20schema,for%20querying%20large%20data%20sets..> [Accessed 4th March 2024].
- [12] Ventiv Technology, "The Difference Between Internal and External Data," Ventiv Technology, 11th April 2022. [Online]. Available: <https://www.ventivtech.com/blog/whats-the-difference-between-internal-and-external-data>. [Accessed 6th March 2024].
- [13] J. Ramakolote, J. Van der Poll and C. Dongmo, "Towards a Conceptual Framework for Data Management in Business Intelligence," *Inf*, vol. 14, p. 547, 2023.
- [14] S. Hosmane, "Overview of Data Warehouse," 2021. [Online]. Available: <https://api.semanticscholar.org/CorpusID:235803544>. [Accessed 17th March 2024].
- [15] M. Hernandez and S. Stolfo, "Real-world Data is Dirty: Data Cleansing and Merge/Purge Problem.," *Data Mining and Knowledge Discovery*, vol. 2, pp. 9-37, 1998.
- [16] D. Loshin, *The Practitioners Guide to Data Quality Improvement*, San Francisco: Morgan Kaufmann, 2011.
- [17] "General Data Protection Regulation," Information Commissioner's Office, [Online]. Available: <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/>. [Accessed 21 March 2024].

6 Appendix

6.1 OLTP Structure

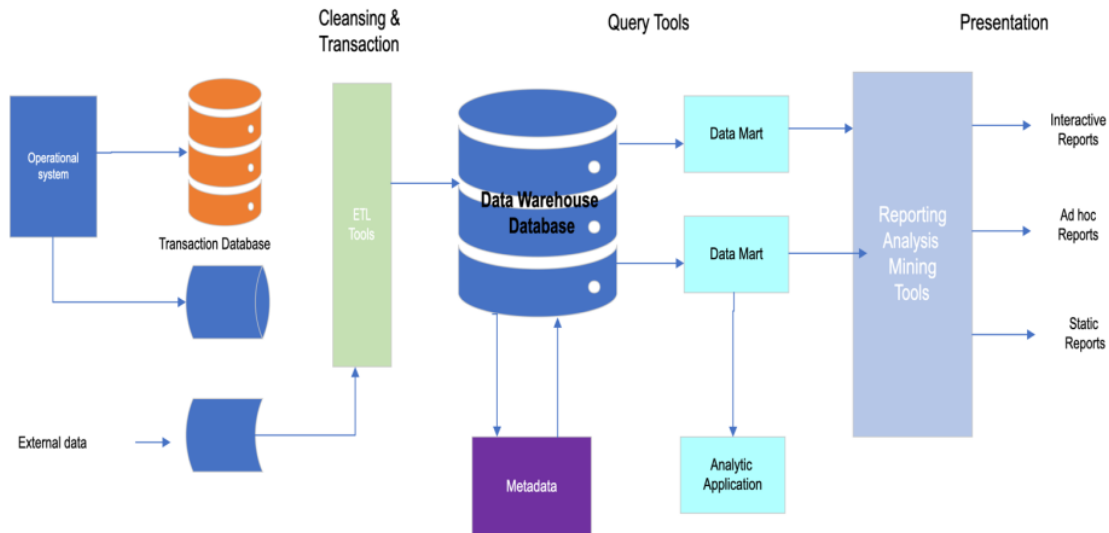
Figure 21 OLTP Structure



Created in Visio

6.2 Data Warehouse Structure

Figure 22 Data Warehouse Structure



Created in Visio

6.3 Data Mart Examples

Table 3 Data Mart Examples

User Group	Data Mart	Description
Managing Director	Performance Metrics	Key Performance Indicators - revenue, profit margins, market share, customer satisfaction index
	Market Analysis	Over view of market trends, competitor analysis and regional performances
	Operational Efficiency	Tracking metrics like agent productivity, property turnover rates, average time to close deals
Board	Financial	Financial overview, metrics, budget vs actual spends, incomes, cash flow analysis, investment analysis
	Strategic Planning	long term strategic goals, market expansion, risk assessments
	Compliance	Legal, audits, governance status
Sales Department	Property Inventory	available properties, location, type, size, price, status
	Lead Management	Leads, conversions, sales pipelines, sales forecasts
	Customer Relationship	customer profiles, preferences, interactions, feed back for the CRM
Finance Department	Financial	Aggregated financial data, balance sheets, income statements, cash flow statements
	Expense management	Tracks expenses, cost allocations, vendor payments, cost saving opportunities
	Revenue Analysis	Revenue Streams, pricing strategies, commission structures, profitability by property type, region
HR Department	Employee Performance	Employee productivity, sales performance, training, career development
	Recruitment	recruitment pipelines, applicant tracking, hiring metrics, retention, diversity
	Employee Engagement	Employee satisfaction, turnover rates, strategies for improvement
Marketing Department	Campaign	Effectiveness, conversion rates, return investment
	Digital	website analytics, social media engagement, email marketing performance, online advertising
	Market Segmentation	demographical information on customer behaviours, market strategy performance

6.4 Metadata Business Rules

Attribute Name	Data Type	Description
Day	Int	Dates are added as integers and then converted to calendar date
Week	Int	Dates are added as integers and then converted to calendar date
Month	Int	Dates are added as integers and then converted to calendar date
Quarter	Int	Dates are added as integers and then converted to calendar date
Year	Int	Dates are added as integers and then converted to calendar date
PropertyType	Varchar(50)	Type of property, commercial, flat, freehold, detached, house, bungalow
Street	Varchar(50)	Type of property, commercial, flat, freehold, detached, house, bungalow
City	Varchar(100)	City name
Postcode	Varchar(10)	Postcode comprising of letters , numbers and spaces
Region	Varchar(100)	County area
Country	Varchar(50)	England, Scotland, Wales, Northern Ireland, Republic of Ireland
BranchName	Varchar(50)	Name of individual branch
BranchType	Varchar(50)	Head Office, Regional Branch, Area Branch, Local Branch
BranchManager	Varchar(50)	Manager of individual branch
PromotionType	Varchar(50)	National, Regional, Local – where the promotion is running
PromotionName	Varchar(100)	What the promotion is called
PromotionLead	Varchar(100)	Name of staff member leading the project
Name	Varchar(200)	First and last name, is used for Staff/Owners/Clients
Position	Varchar(100)	Position held within Cambrian
Email	Varchar(100)	Current email
Telephone	Varchar(20)	Contact telephone number – including/excluding spaces
OwnerType	Varchar(30)	Private or commercial
OwnerName	Varchar(100)	Individual's first and last name or Company name

ClientType	Varchar(30)	Private or commerical
CampaignType	Varchar(100)	Social media, website, emails
TargetGroup	Varchar(100)	Region or population demographic
NewspaperName	Varchar(100)	Newspaper where the advert has been placed
NewspaperType	Varchar(100)	National, regional, free,
MaintenanceType	Varchar(100)	Electrical, structural, plumbing
MaintenanceStatus	Varchar(100)	Ongoing, finished, awaiting
CrewID	Varchar(100)	Crew working on property – internal, external
Company	Varchar(200)	In website dim is the name of the company's website
URL	Varchar(200)	Websites address
Contact	Varchar(100)	First and last name of person who is the lead contact for the company
Status	Varchar(100)	In Campaign Dim and Legal dim - current , pending or finished
Area	Varchar(100)	Where was campaign run – city, region, country
LengthOfCampaign	Int	How long the campaign has or will run
ContractType	Varchar(100)	Rental agreement, lease, deeds of ownership
BillingStatus	Varchar(100)	Paid, owing, not due
Invoice	Int	Invoice number

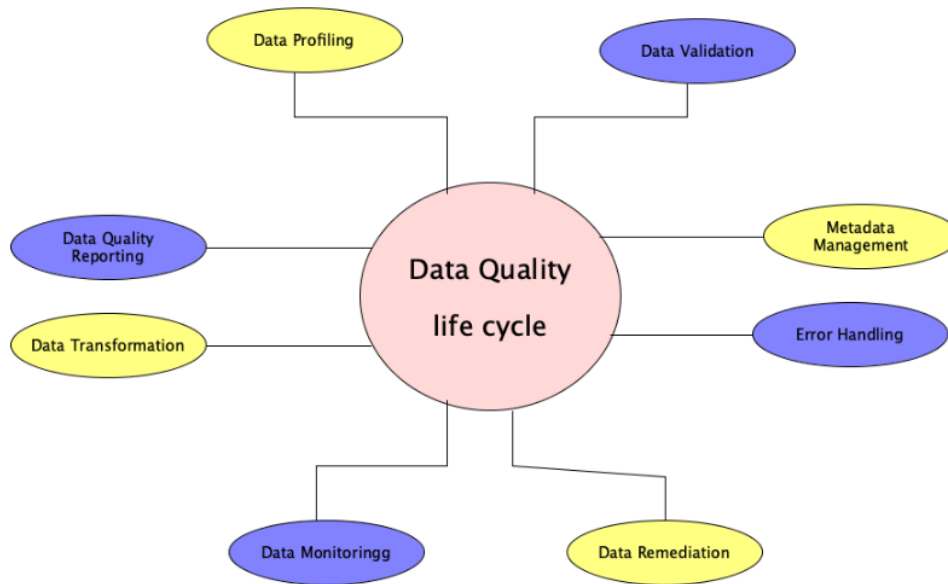
6.5 Data Cleansing

Table 4 Data Cleansing

Process	Description
Deduplicate	Remove any duplicates
Irrelevant	Remove any data that is not relevant to the table or analysis
Structure	Correct any formatting or typo within the data and ensure consistency
Incomplete/missing data	Use appropriate methodology to manage missing data, i.e. removal
Standardisation	Convert data into consistent format – date dd/mm/yyyy
Mapping	Matching external data values to internal values
Enrichment	Supplementing internal sources with external data. i.e. census to improve depth of data
Validation	Create validation at source rules to perform data quality checks i.e. ensuring that each customer has First and Last names.
Special characters	Removal of special characters that could cause an issue when performing analysis
Cyber	Scan any external source for potential threats

6.6 Data Quality Life Cycle

Figure 23 Data Quality Life Cycle



6.7 Sensitive Data

Table 5 Sensitive Data

Type	Description
Personal Data	Names, contacts, email addresses, phone numbers, physical addresses, identification numbers (passport numbers, driver's license, NI number) Location data – this includes staff and client's data
Special Category Data	Racial or ethnic origin, Political opinions, member of trade unions, religious beliefs, sex orientation
Confidential Business Data	Intellectual property, commercially sensitive information, company plans or campaigns

Data Warehousing and Data Mining

GRADEMARK REPORT

FINAL GRADE

GENERAL COMMENTS

75/100

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

PAGE 16

PAGE 17

PAGE 18

PAGE 19

PAGE 20

PAGE 21

PAGE 22

PAGE 23

PAGE 24

PAGE 25

PAGE 26

PAGE 27

PAGE 28

PAGE 29

PAGE 30

PAGE 31
