

BG Data warehousing

by Benjamin Gaylard

Submission date: 02-Apr-2024 06:41PM (UTC+0100)

Submission ID: 228392116

File name: 23453_Benjamin_Gaylard_BG_Data_warehousing_565908_1933221019.docx (187.96K)

Word count: 2826

Character count: 18476

Table of Contents

Task One:	2
OLTP Databases:	2
Data Warehouses:	2
Key Differences and Benefits:	2
Architecture of a Data Warehouse and Source Systems:.....	3
Components of Data Warehouse Architecture:	3
Benefits of ETL Processes:	4
Data Marts for Targeted Information Provision:	6
Key Characteristics and Benefits of Data Marts:.....	6
Example Use Cases of Data Marts:	6
Difference Between Normalised OLTP Structure and Dimensional Data Warehouse Structure:	8
Example Star Schema for the Company's Data:	8
Task Two:	9
Internal and external source systems	9
Identification of Internal Source Systems:	9
Identification of External Source Systems:.....	9
Star Schema	10
Data Governance.....	Error! Bookmark not defined.
References	11
Figure 1 - A Diagram of Data Warehouse Architecture and ETL Process [12].....	4

Task One:

OLTP Databases:

OLTP databases are the backbone of day-to-day business operations, handling high-volume, real-time transactions efficiently. Whether it's processing customer orders, managing inventory, or tracking interactions in customer relationship management (CRM) systems, OLTP databases ensure smooth and responsive transaction processing. These systems are optimized for fast data retrieval and updates, emphasizing the importance of transactional consistency and reliability [1].

In OLTP databases, data is typically organized using normalized data structures to minimize redundancy and maintain data integrity. This design enables efficient storage and retrieval of transactional data, supporting tasks such as insertions, updates, and deletions with low-latency requirements [2]. Examples of OLTP systems include e-commerce platforms, banking systems, and airline reservation systems, where real-time transaction processing is essential for business operations. Additionally, OLTP databases are often designed to ensure ACID properties (Atomicity, Consistency, Isolation, Durability) to guarantee data reliability in a high-concurrency environment [3].

Data Warehouses:

Data warehouses, on the other hand, are designed for data analysis and reporting rather than transactional processing. They serve as centralized repositories of historical data, aggregating information from OLTP systems and other sources to support business intelligence (BI) initiatives. Unlike OLTP databases, data warehouses prioritize query performance and analytical capabilities, enabling organizations to extract insights from large volumes of data [4].

Data warehouses often employ denormalized or star-schema designs to facilitate complex queries and analytical processing. Before being stored in the data warehouse, data undergoes Extract, Transform, and Load (ETL) processes to clean, transform, and integrate it from various sources. This ensures data consistency and quality, making it suitable for analysis and reporting purposes [4].

Organizations use data warehouses for a wide range of analytical tasks, including trend analysis, predictive modelling, and decision support. Business intelligence tools, executive dashboards, and data mining platforms leverage the capabilities of data warehouses to provide stakeholders with actionable insights for strategic decision-making [4].

Key Differences and Benefits:

The primary difference between OLTP databases and data warehouses lies in their purpose and optimization. OLTP databases focus on high-volume, real-time transactions, while data warehouses prioritize data analysis and reporting. This distinction is reflected in their design, with OLTP databases emphasizing fast data retrieval and updates and data warehouses excelling at complex queries and data mining [5].

Data warehouses provide several key benefits to organizations. They enable improved decision-making by providing a centralized repository of historical data for analysis [6]. Data warehouses also enhance operational efficiency by identifying inefficiencies and bottlenecks in business processes [7]. They contribute to enhanced customer service by tracking customer interactions and identifying trends. Additionally, data warehouses support new product development by revealing market opportunities and informing product development strategies.

Architecture of a Data Warehouse and Source Systems:

The architecture of a Data Warehouse comprises multiple components that work together to facilitate data storage, processing, and analysis.

Components of Data Warehouse Architecture:

1. **Operational Source Systems:** These are the primary transactional systems, such as OLTP databases, CRM (Customer Relationship Management) systems, ERP (Enterprise Resource Planning) systems, and external data sources, from which data is extracted for loading into the Data Warehouse [8].
2. **ETL Processes:** ETL processes encompass Extracting data from source systems, transforming it to conform to the Data Warehouse schema and business rules, and loading it into the Data Warehouse [9].
3. **Staging Area:** The staging area serves as an intermediate storage space where extracted data is temporarily stored before being transformed and loaded into the Data Warehouse. It facilitates data cleansing, integration, and validation processes [10].
4. **Data Warehouse Database:** The Data Warehouse database is the central repository that stores integrated, historical data from multiple source systems. It typically consists of fact tables and dimension tables organised in a star or snowflake schema to support analytical querying and reporting [9].
5. **OLAP (Online Analytical Processing) Engine:** OLAP engines provide multidimensional analysis capabilities, enabling users to perform complex queries, drill-down analysis, and data visualisation on the Data Warehouse data [11].
6. **Front-end Tools:** Front-end tools, such as reporting tools, dashboards, and BI (Business Intelligence) applications, provide user interfaces for querying, analysing, and visualising Data Warehouse data [10].

Populating the Data Warehouse: The process of populating a Data Warehouse involves several steps, as illustrated in the diagram below:

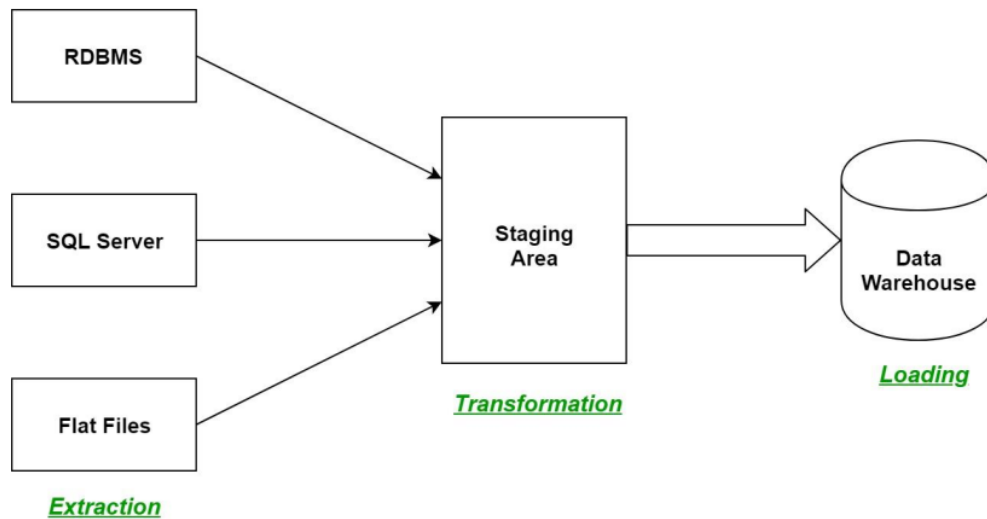


Figure 1 - A Diagram of Data Warehouse Architecture and ETL Process [12]

1. **Data Extraction:** Data is extracted from operational source systems using ETL tools and techniques. This involves identifying relevant data sources, extracting data subsets based on predefined criteria, and transferring the extracted data to the staging area.
2. **Data Transformation:** Extracted data undergoes transformation processes to conform to the Data Warehouse schema, including data cleansing, data integration, data enrichment, and data aggregation. Transformations ensure consistency, accuracy, and compatibility of data across different source systems.
3. **Data Loading:** Transformed data is loaded into the Data Warehouse database, where it is stored in fact tables and dimension tables according to the dimensional modelling schema. Loading processes may involve full loads, incremental loads, or real-time data streaming, depending on the frequency and volume of data updates.

Benefits of ETL Processes:

- **Data Quality:** ETL processes ensure data quality by cleansing, standardising, and validating data before loading it into the Data Warehouse. This improves the accuracy and reliability of analytical results and decision-making [13].
- **Data Integration:** ETL processes integrate data from disparate source systems, consolidating it into a unified and consistent format within the Data Warehouse. This enables cross-functional analysis and facilitates a holistic view of organisational data [13].
- **Operational Efficiency:** ETL processes automate data movement and transformation tasks, reducing manual effort and increasing operational efficiency. This allows organisations to streamline data processing workflows and focus resources on value-added activities [13].
- **Scalability and Flexibility:** ETL processes are scalable and flexible, handling large volumes of data from diverse sources and adapting to changing business requirements. This ensures that the Data Warehouse can accommodate growth and evolution in data volume, variety, and complexity over time [13].

- **Data Governance:** ETL processes enforce data governance policies and procedures, ensuring compliance with regulatory requirements and internal data standards. This enhances data security, privacy, and integrity within the Data Warehouse environment [13].

Data Marts for Targeted Information Provision:

Data marts are subsets of a data warehouse that are tailored to meet the specific needs of individual departments or user groups within an organisation. They serve as specialised repositories for storing and accessing relevant data related to specific business functions, such as finance, sales, marketing, or operations. By focusing on a specific subject area or business function, data marts provide targeted information and analytical capabilities to support decision-making and performance analysis within departments [14].

Key Characteristics and Benefits of Data Marts:

1. **Subject-Oriented:** Data marts are organised around specific subject areas or business functions, such as sales, marketing, or finance. This enables users to access and analyse data relevant to their departmental requirements without being overwhelmed by irrelevant information [15].
2. **Agility and Flexibility:** Data marts are designed to be agile and flexible, allowing for rapid deployment and customisation to meet changing business needs. They can be developed incrementally, focusing on high-priority areas and expanding over time to incorporate additional data sources and analytical capabilities [16].
3. **Performance Optimisation:** Data marts are optimised for performance, enabling fast query response times and efficient data retrieval for analytical purposes. By pre-aggregating and indexing data based on departmental requirements, data marts enhance query performance and improve user productivity [16].
4. **Data Governance and Security:** Data marts enforce data governance policies and access controls, ensuring that departmental users have appropriate permissions to access and manipulate data relevant to their roles and responsibilities. This enhances data security, privacy, and compliance with regulatory requirements [16].
5. **Decision Support:** Data marts provide decision support capabilities, empowering departmental managers and analysts to derive insights, monitor performance, and make informed decisions based on relevant data. They facilitate ad-hoc querying, reporting, and data visualisation to support exploratory analysis and hypothesis testing [16].

Example Use Cases of Data Marts:

1. **Sales Data Mart:** A sales data mart may contain transactional data related to customer orders, product sales, and revenue performance. It enables sales managers to track sales trends, analyse customer behaviour, and forecast future sales opportunities.
2. **Marketing Data Mart:** A marketing data mart may contain data on campaign performance, customer demographics, and market segmentation. It allows marketing managers to evaluate marketing effectiveness, target specific customer segments, and optimise marketing strategies.
3. **Finance Data Mart:** A finance data mart may contain financial data, such as income statements, balance sheets, and cash flow statements. It provides finance managers with insights into financial performance, budget variance analysis, and cost allocation across departments.

4. **Inventory Data Mart:** An inventory data mart may contain data on inventory levels, stock movements, and supply chain operations. It enables operations managers to optimise inventory management, minimise stockouts, and improve supply chain efficiency.

Difference Between Normalised OLTP Structure and Dimensional Data Warehouse Structure:

The structure of an OLTP database is typically normalised, meaning that data is organised into multiple related tables to minimise redundancy and dependency. This normalisation reduces data duplication and ensures data integrity, making it suitable for transactional processing and operational efficiency. However, normalised OLTP structures may be complex and less conducive to analytical querying and reporting [15].

In contrast, the structure of a data warehouse is dimensional, emphasising the efficient retrieval and analysis of data for decision support and business intelligence purposes. Dimensional modelling involves organising data into two types of tables: fact tables and dimension tables.

- **Fact Tables:** Fact tables contain quantitative, numerical data (facts) that represent business metrics or measurements, such as sales revenue, product quantities, or customer orders. Fact tables typically have foreign key relationships with dimension tables and serve as the central focus for analytical queries and analysis [14]
- **Dimension Tables:** Dimension tables contain descriptive attributes (dimensions) that provide context and additional information about the facts stored in the fact tables. Dimensions represent categorical or textual data, such as product categories, customer segments, or periods. Dimension tables serve as reference tables for querying and filtering data in fact tables [15].

Example Star Schema for the Company's Data:

Consider a hypothetical scenario where the company operates in the retail industry, selling products through online and offline channels. A simplified star schema for analysing sales data within the company's data warehouse may include the following components:

- **Fact Table:** The central fact table could be the "Sales" fact table, containing measures such as sales revenue, quantity sold, and discount amount. It would have foreign key relationships with dimension tables representing various aspects of the sales transaction, such as:
 - **Product Dimension:** Contains attributes related to products, such as product ID, product name, category, and brand.
 - **Customer Dimension:** Contains attributes related to customers, such as customer ID, name, address, and demographics.
 - **Time Dimension:** Contains attributes related to time, such as date, month, quarter, and year, facilitating time-based analysis of sales trends.
 - **Store Dimension:** Contains attributes related to store locations, such as store ID, store name, address, and geographical information.

Task Two:

Internal and external source systems

what is Internal source systems & external source systems

Internal source systems is information that come directly from the company/business whereas external source systems is information that come from an different company/business which is not associated to the company [18].

Identification of Internal Source Systems:

1. Customer Database: Contains information about customers, including their contact details, preferences, purchase history, and interactions with Cambrian's services.
2. Property Listings Database: Stores details about properties available for sale or rent, including address, type, size, price, amenities, and photos.
3. Lease Agreement Database: Manages lease agreements for properties managed by Cambrian, including terms, rental amounts, lease durations, and renewal dates.
4. Financial Database: Records financial transactions, including revenue from property sales and rentals, expenses, payroll, taxes, and other financial metrics.

Identification of External Source Systems:

1. Market Trends Data: External data sources provide insights into real estate market trends, such as average property prices, market demand, housing inventory, interest rates, and economic indicators.
2. Property Valuation Data: External databases or services offering property valuation estimates based on various factors like location, property features, market conditions, and recent sales data.
3. Demographic Data: External sources providing demographic information about different locations, including population demographics, income levels, employment rates, education levels, and household composition.
4. Competitor Data: Information on competitors in the real estate market, including their property listings, pricing strategies, market share, customer satisfaction ratings, and marketing tactics.

Star Schema



Figure 2 - Star Schema

Data Governance and Data Quality within ETL

The company/business must carefully consider various crucial stages of data quality and data governance throughout the entire data lifecycle, from extraction to analysis, to ensure effective data warehouse processing [19].

Data Quality

Data profiling plays a crucial role in uncovering potential data quality concerns, such as the presence of missing values or inconsistent formats. A robust approach to data validation through the implementation of stringent business rules and checks during the Extract, Transform, Load (ETL) process is essential. This ensures the accuracy, completeness, and consistency of the data being integrated into the data warehouse. Careful design and thorough testing of data transformation procedures are imperative to uphold data integrity and prevent the inclusion of invalid data within the warehouse environment. Effective understanding of data provenance necessitates the meticulous capture and management of metadata, encompassing elements such as data lineage, definitions, and transformation rules. Establishing clear protocols for error handling, data remediation, and exception management is vital in mitigating the impact of data quality issues within the warehouse while enabling prompt identification of potential concerns. Continuous data monitoring and the prompt reporting of any identified data quality issues further fortify the integrity of the data environment.

Data Governance

The deployment of a comprehensive data governance framework is imperative to delineate clear roles, responsibilities, and policies aimed at guaranteeing data quality and the judicious utilization of data resources. This includes the designation of data stewards tasked with the continuous monitoring, maintenance, and enhancement of data quality standards, thus ensuring the sustainability of data integrity. Furthermore, the implementation of robust data monitoring and auditing procedures is essential to enforce data quality metrics and promptly identify any potential data quality discrepancies. Incorporating data lineage and audit trails facilitates the thorough examination and comprehension of data origins, transformations, and utilization patterns, thereby bolstering data assurance, fostering trust, and ensuring compliance with regulatory standards such as the General Data Protection Regulation (GDPR). Additionally, stringent measures about data security and privacy must be instituted to enforce appropriate access controls and safeguard sensitive information in adherence to GDPR guidelines [20].

References

- [1] R. Elmasri and S. B. Navathe, Fundamentals of Database Systems (7th ed.), Addison-Wesley., 2016.
- [2] C. J. Date, Introduction to Database Systems (14th ed.), Pearson Education, 2020.
- [3] R. Ramakrishnan and J. Gehrke, Database Management Systems (3rd ed.). McGraw-Hill., McGraw-Hill, 2000.
- [4] R. Kimbal and M. Ross, The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling, 3rd Edition, Wiley, 2013.
- [5] AWS, "What's the difference between OLAP and OLTP?," n.d.. [Online]. Available: <https://aws.amazon.com/compare/the-difference-between-olap-and-oltp/>. [Accessed 23 March 2024].
- [6] Fivetran, "Top 6 benefits of data warehouse," 22 May 2023. [Online]. Available: <https://www.fivetran.com/learn/benefits-of-data-warehouse>. [Accessed 24 March 2024].
- [7] N. Gemes, "7 Practical Benefits of Data Warehousing (+ Use Cases)," May 10 2023. [Online]. Available: <https://whatagraph.com/blog/articles/data-warehousing-benefits>. [Accessed 24 March 2024].
- [8] javatpoint, "Difference between Operational Database and Data Warehouse," n.d.. [Online]. Available: <https://www.javatpoint.com/operational-database-vs-data-warehouse>. [Accessed 29 March 2024].
- [9] W. H. Inmon, Data Warehousing Fundamentals: Designing, Building, and Maintaining Data Warehouses, Wiley, 2002.
- [10] R. Kimball and M. Ross, The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling, ETL, and Business Intelligence, Wiley, 2002.
- [11] J. Han and M. Kamber, Multidimensional Databases and Data Warehousing, 2011.
- [12] geeksforgeeks, "ETL Process in Data Warehouse," n.d.. [Online]. Available: <https://www.geeksforgeeks.org/etl-process-in-data-warehouse/>. [Accessed 23 March 2024].
- [13] codeless platforms, "What is ETL? – A Comprehensive Guide to Extract, Transform, and Load," 23 January 2023. [Online]. Available: <https://www.codelessplatforms.com/blog/what-is-etl/>. [Accessed 29 March 2024].
- [14] IBM, "Fact tables and entities," 08 March 2021. [Online]. Available: <https://www.ibm.com/docs/en/ida/9.1?topic=models-fact-tables-entities>. [Accessed 02 April 2024].
- [15] J. Stagnitto and L. Corr, Agile Data Warehouse Design: Collaborative Dimensional Modeling, from Whiteboard to Star Schema, DecisionOne Press, 2011.

- [16] AWS, "Summary of differences: OLAP vs. OLTP," n.d.. [Online]. Available: <https://aws.amazon.com/compare/the-difference-between-olap-and-oltp/#:~:text=OLTP%20uses%20relational%20databases.&text=OLAP%20uses%20star%20schema%2C%20snowflake,uses%20normalized%20or%20denormalized%20models..> [Accessed 23 March 2024].
- [17] IBM , "What is OLTP?," n.d.. [Online]. Available: <https://www.ibm.com/topics/oltp>. [Accessed 30 March 2024].

BG Data warehousing

GRADEMARK REPORT

FINAL GRADE

GENERAL COMMENTS

50/100

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