Data Modeling Techniques for Data Warehousing

The evolution of the concept of data warehousing

A data warehouse is a system that holds data extracted from sources in the operational systems, transformed into appropriate style for storage. This aids in making informed decisions.
Earlier methods of data storage have captured data efficiently however, changes were made to unexpected problems encountered, especially integration problems. Examples of these earlier data storage models include punch cards in the 1950s by huge companies for storing records of customers, employees and competitors (Heide, 2009), Database Management Systems in 1966 for manipulating data in databases (Rouse, 2024), Personal Computers and 4L technology for storing private information and the idea of simple programming for beginners conceptually and Relational databases for creating queries through Structured Query Language (Foote, 2023)
Data warehouses are the latest form of all the above systems discussed. These systems have negatively resulted in inefficient integration due to extension of databases and application systems and inaccurate data. Data warehousing was developed to address the issues of integration and consistency as well as improve the decision-making systems. At some point in 2008, A system known as non-relational database or NoSQL was created with the function of expanding storage and work across business and possibly disposing of SQL because of their lack of scalability. However, that has recently changed, and SQL is more adaptable to change than NoSQL and as a result, SQL has still become the most widely used tool.
Alternatives of Data Warehouses could be data lakes. These alternatives are popular due to their high flexibility and apparently, more information is captured compared to data warehouses. (Foote, 2023)

The difference between Data Marts and Data Warehouse

Data warehouses are used by business to make cautions decisions as they provide strategic information for enterprises whereas Data marts are used to make tactical decisions for the business on taking correct actions that satisfy the goals or objectives of the business.
Data Warehouses give an integrated work environment as it provides a single combined view of the enterprise, usually joined by pre-solved mini tasks of the business. Data marts are departmental as they separate smaller parts of data from the huge piece aid customers in gaining access to the data.
The design of data warehouses is more complex than the design of a single data mart as it is often unclear where to start in data warehouses in achieving business goals, making it more complicated to make the decisions of achieving those goals. Data marts are more simple because they are much smaller compared to data warehouses, making the decision process more straightforward.
The information in the data warehouse comes from multiple sources, therefore this would be easier for data warehouses to get additional information. Data marts retrieve their

information from limited sources, including operational systems and external sources, mainly due to how small their space is compared to data warehouses.
Data warehouses make use of Top-Down models where a huge complex problem is broken down into smaller more manageable activities, aiding the process of making decisions on activities/tasks to be performed aligning with enterprise objectives. Data marts use bottom-up estimates where these mini, manageable tasks are solved and brought together to form the final product or make the final decision. (Karan, 2022)

Data Analysis in Data Warehousing

Data Analysis is a process that involves the following actions: collection, cleaning, transforming, modelling, and interpreting data in a data warehouse. (Eldridge, 2024)
Data collection is described as a process of reading the available data present in the data as well as storing that data to provide strategic information to aid decision making.
Data cleaning is the process of improving the quality of data by checking for potential errors present in the data item interfering with the decision-making system on making strategic decisions. Data cleaning provides certain types of data repairing techniques recommending updates that could add improvement to detected erroneous areas (Ilyas & Chu, 2015).
Transforming data into a data warehouse includes converting a data item into an appropriate format according to the data warehouse. An example of a required format, the data would be required to transform to would be an XML document, database file or Excel spreadsheet. (Pratt & Berstein, 2022)
Data Modelling is the process of providing descriptions of the following: information held in the data warehouse but more efficiently, a simple database, the relationship between those data items and their constraints. (Watt, 2014)
Interpreting data items can be beneficial for warehouses as they can gain some reviews from that data and come up with improvement areas on making better decisions. However, given by how large data warehouses are, it would take time to gather all the feedback needed, including the time to break down the main enterprise view into smaller tasks, as well as getting them reassembled.

Data Warehouse Architecture and Implementation choices

The structure of data warehouse can be referred to its formal definition as follows:
“Data warehouses are described as a collection of subject-oriented, integrated, non-volatile and time-variant data in support of management’s decisions” This definition was inspired by the father of data warehousing: Bill Inmon.
Subjects are data items that that are applicable to real world, such as a manufacturing company holding shipment and delivery as their practical subjects. Subject-oriented data enables a merged perspective on a data item.
Integration is used to combine all data from different source systems to decrease the inconsistency of data elements before they are stored in the data warehouse. Naming conventions could be an example of simplifying disparate data, making understanding equal with the level of the element.

Time-variance of data in the data warehouse stores current and historical values which are captured by taking snapshots. This will be beneficial in interpreting the changes and development of certain enterprises over time.
Data warehouses have non-volatile data as information in these systems are to be remain constant to assure trustworthy analysis. However, new data can still be added to the warehouse, just no changes are applicable to that data once captured. Besides analysis, queries are performed by non-volatile data (Sonny, 2023)

Suitable Approaches and techniques for architecting the data in the data warehouse.

The important methods to take for structuring data in a data warehouse, includes the Inmon approach which consists of centralized Data Warehouse, Data normalization, data marts, High initial investment, and the bottom-up model. (Fedynshyn, 2023)
The Inmon approach is a method that was established by Bill Inmon and can be identified as the Top-Down Model. Judging by its name that was discussed earlier, 'top-down' it breaks the main problem into smaller manageable tasks. (Fedynshyn, 2023)
According to this approach, a centralized data warehouse is initiated to perform as a combined architecture of single data marts. All business data is accessed and stored into the data warehouse including queries. (Fedynshyn, 2023)
Data normalization is another major key aspect of architecting data in the warehouse which involves breaking the data in the data warehouse into independent data marts. Each data mart is assigned a certain individual task to complete; this can reduce inconsistency thus enduring data integrity. (Fedynshyn, 2023)
Research shows that Bill Inmon's approach is a high initial investment as it in demand of a considerable upfront interest despite the limitations of the payoff. (Fedynshyn, 2023)
Ralph Kimball's approach to architecting data in a data warehouse is identified as the bottom-up model which would be described as the opposite of the Inmon model as it involves integrating all the departmental tasks into a single enterprise view. (Fedynshyn, 2023)
Instead of starting with the data warehouse, the model starts with the data marts which can also be identified as building blocks. Each building block is designed for a certain business activity and can be developed quickly to align with objectives more efficiently. (Fedynshyn, 2023)
Data from each building block are organized using a simple structure known as a star schema. Due to its simple structure, making it relatively easy for the user to understand, it is appropriate for quick, iterative analytics. (Fedynshyn, 2023)
Starting with data marts enables users to use data elements soon due to its decision-making system easier than that system within the data warehouse. This offers the business benefits of being complete sooner. (Fedynshyn, 2023)
Once all data marts of the model are complete, they are scaled up and integrated to form the final data warehouse allowing this model to be prone to change (Fedynshyn, 2023)

Data modelling in a data warehouse

Data modelling is the process of representing the needs and requirements of a business in a model for individuals involved in the organization can understand these requirements visually. This visual representation will also show the data structures of the organization and how these various data structures will also relate to one another (Morris, 2021).
Data modelling is a significant process in a data warehouse as information within the data warehouse needs to be presentable. A data model assists a data warehouse to have proper information by improving its consistency, accuracy, and organization, as well as facilitate the data warehouse's integration.
With data modelling, accuracy can be improved by establishing data quality rules, as well as rules of extracting data and converting that data to a suitable format in preparation to load it. Defining data quality rules keeps you informed of what is required out of the data, especially in terms of its quality.
Once the data quality rules are defined choosing the most relevant data sources. However, before choosing these sources, it is essential to be completely aware of your needs, to understand fully which sources would be most suitable for retrieving data for the data warehouse.
Once data has been selected, their quality will be validated using data validation check to ensure the quality of the data is aligned with the defined quality rules.
Therefore, data models improve accuracy, as well as consistency.
Data consistency can be improved by data modelling by ensuring individuals in the organization are on track, by using the same terminology, aligning understanding amongst individuals. Initially, the model will initiate data standards explaining how data will be collected, captured, and used for the business. These standards are to be aligned with policies, procedures and guidelines of a framework defined by the business.
Once the standards are set, it is up to the data owner to ensure the alignment of these standards. This is also done by using data validation checks, including data profiling and data standardization.
Research has shown how the data modelling process aids data consistency and now data organization shall be discussed.
Data modelling improves data organization by creating a conceptual model of how data will be organized in the data warehouse. One of the main steps this process has in organizing data is by creating a data architecture, outlining the various data types that are to be collected as well as how it will be stored and processed. Just like with consistency, a framework consisting of governing data can be used to define policies on creating rules for naming conventions and defining data standards. Once data in the data warehouse has been set, a metadata management activity can be used to manage the data captured and keep track of where it came from and how it us used over time. With regards to this data, data modelling clearly assists data organization.
The biggest aspect that makes data modelling critical for a data warehouse is the facilitation of integration of data warehouse. As data warehouses are mainly about integration, data modelling can facilitate this process of data warehouses. Defining requirements for integrating data in the warehouse is the first step so that individuals involved in the

organization are aware of which data elements of a certain type are too be combined in the data warehouse. Once requirements are identified, users should use ETL tools to receive data from different source systems and then convert that extracted information to a format that is suitable according to the integration requirements. Once data is transformed from to appropriate format, the process of data virtualization implemented to concatenate all converted data in the warehouse.
With regards to this discussion, data modelling is clearly a crucial activity in a data warehouse in terms of integration, accuracy, inconsistency, and organization (Lehnerdt, 2023).

Process Model for data warehouse modelling

The process of process modelling is defined as the graphical representation of business activities via the workflow. This model is an important component of automation so it would automatically define activities and determine the optimized path for the flow of work. The benefits given by process models are the improvement of efficiency, as workflow is already planned and organized for workers to understand what skills they need to develop in the training program and what activities they need to perform to meet objectives of the organization (Vanner, 2020).
Process modelling consists of the following key steps to designing a data warehouse, defining requirements, data staging, data storage and data integration.
Defining business requirements focuses on identifying rules based on what users need for their business. This component is aimed at solving the user's potential issue and providing strategic information.
After the requirements of the business has been identified, the next phase of the process model is to retrieve information from various sources outside of the business. Once those sources have been extracted, they must transform that data into appropriate data format and type according to the business requirements in preparation of loading it into the data warehouse. Once the conversion is complete, it is fully prepared to be loaded and is applied to the data warehouse.
The repositories of data in the data warehouse can consist of current values and historic values. These values, especially historic values, are captured via the usage of snapshots. With data warehouses, data can be kept together as a combined set of all data stored using various integration strategies. A method used to integrate data in a data warehouse is Ralph Kimball's bottom-up approach, which combines all independent or dependent marts together.

Core Data Modelling Techniques for the data warehouse development process

The core data model is the model that is responsible for identifying and defining data types. The aim of this type of data model is to join data from multiple source systems., as well as the historization of that data. The following techniques of core data modelling used on data

warehouses include relational model, entity-relationship model, hierarchal, network, dimensional, object-oriented database, and object-relational models.
In the relational model, data is architected in tables with columns and rows, fields, and attributes respectively. A table includes a primary key that is used to distinctly identify a row, as well as foreign keys that are used to describe the relationship between each table; either one-to-many, many-to-one, one-to-one or many-to-many. Relational models are applicable to any structure that relies on attributes and fields.
Entity relationships are used to represent the relationships between certain tables, which are referred to as tables. This model does not link relationships with entities, as that's the function of foreign keys as mentioned before, but rather displays which tables is related. These techniques are used in systems that enable a process of huge complex structures to be broken down into smaller, more manageable, also known as top-down approach. This could be in content management systems or file systems.
Hierarchical Model functions as a tree structure consisting of data being organized with parent and child elements. Each record in this structure contains one parent and multiple children. The parent could be representation of the primary key although research here does not state otherwise. One drawback listed in this research is complexity of the interaction between the data elements. These techniques can be used in the following systems: Extensible Markup Language and Geographic Information Systems.
Network Technique contains data structured in a graph, entities being positioned as nodes and relationships among edges. While this benefits in finding a solution to complex data relationships, it has the potential of forming a more difficult data base structure. Applications of network models include systems of grouping information into separate attributes with ease; this would include social networks, educational systems, and customer management systems (Anon., 2023).
Dimensional Models are are techniques that arrange information into fact tables and dimension tables. Dimension tables are basic information fields holding certain attributes related to the organization. Fact tables are the tables that hold units of measurements that can be applied to the attributes of the dimension. Dimensional models are the most applicable techniques for data warehouses. (Anon., 2023)
The second-to last model is object-oriented, which works with the data as objects. Each object represents and entity that is summarized as a combination of attributes. Objects that share similar characteristics fall under forms known as classes. The model is useful in complex system software including Computer Aided Design (Anon., 2023).
Lastly, object-relational Model is like a combination of relational and object-oriented Models. It stores objects in a table filled with attributes and fields. This technique would be most applicable to huge applications with advanced data management skills, such as enterprise resource planning systems (Anon., 2023).

An overview of a Function a data model must support for modelling a data warehouses.

A data modeler is an individual who works at creating various data models that identify the organization integration and management of data. Data warehouses are functioned by integration as they are a combination of data marts or independent groups within an

enterprise. A good example of data model could be the Oracle SQL Developer Data Model toot. The data model tool contains significant features with deep integration with Oracle products being one of them. This could be suitable for Oracle database users as this data model involves deep integration and as we have mentioned before, data integration is the important function of supporting data warehouse that shall be discussed. Another example of a suitable data model tool would be IBM InfoSphere Data Architect. Research has shown reported that all aspects of a project are managed through its Integrated Development Environment, in short IDE. IDE consists of features that enable code management which provides the benefits of potentially disposing of errors, and most possibly the integration process. (Team, 2022)

Populating the data warehouse or data mart

Data warehouses can be done by using the ETL process, especially after data has been extracted from source systems and transformed into a suitable format that is required by the business requirements and business rules for extraction. The normalized data store is populated with data after data has gone through a fire wall for verification of data quality meeting with the requirements of the business and becoming more aligned with the stakeholders' and business owner's expectations. (Rainardi, 2008)

Bibliography

Gorunescu, F., 2011. *Data Mining.* 1st Edition ed. s.l.:Springer Berlin, Heidelberg.
Singh, K., 2018. *Data Cube and its basic functionality.* [Online]
Available at: https://www.linkedin.com/pulse/data-cube-its-basic-functionality-kishan-singh
[Accessed 26 February 2024].
Fivetran, 2023. *Data movement: the ultimate guide.* [Online]
Available at: https://www.fivetran.com/learn/data-movement
[Accessed 26 February 2024].
Heide, L., 2009. *Punched-Card Systems and the Early Information Explosion, 1880-1994.* Baltimore: John Hopkins University Press.
Foote, K. D., 2023. *A Brief History of the Data Warehouse.* [Online]
Available at: https://www.dataversity.net/brief-history-data-warehouse/
[Accessed 28 February 2024].
Rouse, M., 2024. *DBMS(Database Management System).* [Online]
Available at: https://www.techopedia.com/definition/24361/database-management-systems-dbms
[Accessed 28 February 2024].
Karan, R., 2022. *Difference Between Data Mart and Data Warehouse.* [Online]
Available at: https://www.shiksha.com/online-courses/articles/difference-between-data-mart-and-data-warehouse/
[Accessed 28 February 2024].
Eldridge, S., 2024. *data analysis.* [Online]
Available at: https://www.britannica.com/science/data-analysis
[Accessed 28 February 2024].
Pratt, M. K. & Berstein, C., 2022. *data transformation.* [Online]
Available at: https://www.techtarget.com/searchdatamanagement/definition/data-transformation
[Accessed 28 February 2024].
Ilyas, I. F. & Chu, X., 2015. Trends in Cleaning Relational Data: Consistency and Deduplication. *Foundations and Trends,* 5(4), pp. 283-393.
Watt, A., 2014. *Chapter 5 Data Modelling.* 2nd Edition ed. s.l.:s.n.
Sonny, R., 2023. *The fundamentals of data warehouse architecture.* [Online]
Available at: https://www.thoughtspot.com/data-trends/data-modeling/data-warehouse-architecture#:~:text=%22Data%20warehouse%20architecture%20refers%20to,on%20the%20world%20of%20data.
[Accessed 16 September 2023].
Fedynshyn, R., 2023. *Building a data warehouse: A step-by-step guide.* [Online]
Available at: https://www.n-ix.com/building-a-data-warehouse/
Morris, A., 2021. *Data Modelling Explained: Types & Benefits.* [Online]
Available at: https://www.netsuite.com/portal/resource/articles/data-warehouse/data-modeling.shtml
Lehnerdt, R., 2023. *The importance of Data Modelling for your DWH.* [Online]
Available at: https://www.analyticscreator.com/blog/the-importance-of-data-modeling-for-your-dwh
Team, O. M. E., 2022. *Top Data Modelling tools you must know.* [Online]
Available at: https://www.onlinemanipal.com/blogs/top-data-modelling-tools-you-must-know

Anon., 2023. *What is Data Modelling? Types, Processes and Tools.* [Online]
Available at: https://www.altexsoft.com/blog/data-modeling/
Vanner, C., 2020. *What is Process Modelling? 6 Essential Questions Answered.* [Online]
Available at:
https://www.google.com/search?q=Define+a+process+model+author+and+date&sca_esv=6e8c819a630d1523&sxsrf=ACQVn09sN3C4oKgjM5bjm1ziRiTJ5NkzdQ%3A1709307905171&ei=AfjhZcWICoiphbIPp5yZoAI&ved=0ahUKEwiFjf-TtNOEAxWIVEEAHSdOBiQQ4dUDCBA&uact=5&oq=Define+a+process+mod
Rainardi, V., 2008. *Populating the Data Warehouse.* [Online]
Available at: https://link.springer.com/chapter/10.1007/978-1-4302-0528-9_8