Data Warehouse – Design Factors

Agenda

Here, we will cover:

- Meta Data
- Accessing Data Warehouse
- Modeling Techniques Requirement
- Analysis Modeling
- Tool Selection

Meta Data

- Metadata is data about data. In other words, it describes the various aspects of data, such as its structure, meaning, relationships, and origin.
- It helps define and document the data warehouses components such as data sources, tables, columns, transformation, and business rules.
- It provides a common vocabulary and understanding among stakeholders, such as business analysts, developers, and administrators.
- It supports the data integration, quality, and governance processes by identifying and resolving data inconsistencies, redundancies, and errors.
- It facilitates data analysis and reporting by providing context, data lineage, and history of data.

Meta Data

- Metadata can be classified into two categories – Technical and Business
- Technical metadata describes the technical aspects of data, such as format, structure, storage, and processing.
- Business metadata describes the business aspects of data, such as its meaning, context, usage, and ownership.

Examples of Technical Meta Data

- Data dictionary: A repository of data definitions, attributes, and relationships.
- Source-to-target mapping: A document that shows how the data is transformed and loaded from source systems to the data warehouse.
- ETL process metadata: A record of the ETL jobs, scripts, parameters, and dependencies.
- Database schema: A visual representation of the database objects, such as tables, views, indexes, and constraints.

Examples of Business Meta Data

- Business glossary: A collection of business terms, definitions, and rules
- Data lineage: A record of the data origins, transformation, and destination
- Data ownership: A definition of the business units, roles, and responsibilities for the data
- Data quality rules: A set of criteria for data accuracy, completeness, consistency, and timeliness

Accessing Data Warehouse

- Once the data warehouse is designed and populated with data, it needs to be accessed by users to support decision-making processes.
- There are multiple access patterns to consider while designing a Data Warehouse to support the needs of users.
- The most common access patterns are: Ad-hoc Query and Analysis, Online Analytical Processing (OLAP), Reporting, and Data Mining.
- Each pattern has its own requirement and characteristics, and the Data Warehouse design needs to take these into account.

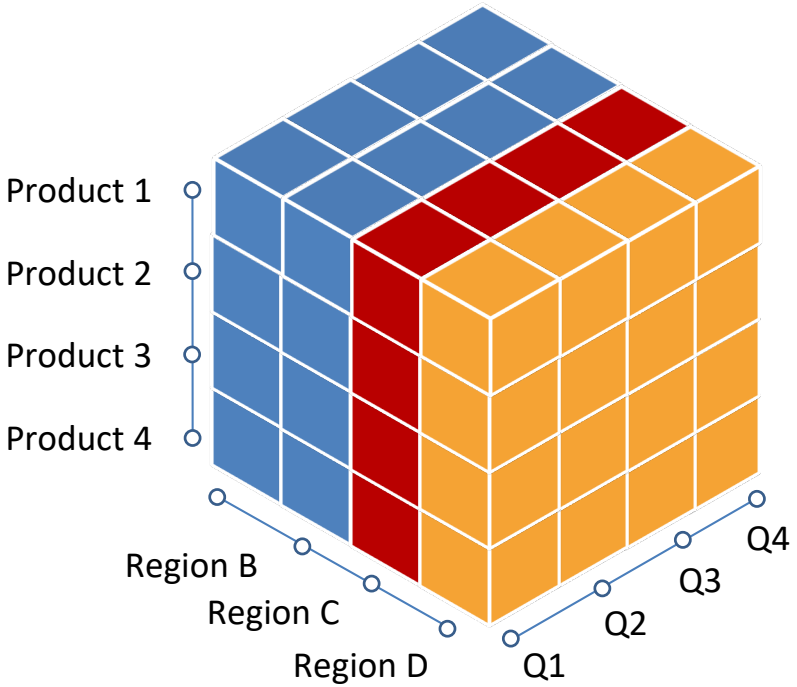
Accessing Data Warehouse: Ad-hoc Query and Analysis

- This query requires flexible and powerful query tools that allow for the exploration and analysis of data in a free-form manner.
- The Data Warehouse design needs to ensure that the data is organized in a way that supports ad-hoc query and analysis by providing a rich set of dimensions, measures, and hierarchies that can be used to slice and dice the data in different ways.
- The design should take into account the performance requirements by optimizing database schema, indexing the data appropriately, and caching frequently accessed data.

Accessing Data Warehouse: OLAP

- OLAP (Online Analytical Processing) is a technology to perform complex, multidimensional data analysis in real-time.
- OLAP works by organizing data into multidimensional structures called cubes, allowing users to analyze a large amount of data quickly and efficiently via web-based interfaces or specialized OLAP software such as Microsoft SQL Server Analysis Services, Oracle OLAP, Tableau, MicroStrategy, Pentaho Mondrian.
- These cubes contain measures (quantitative data such as sales figures) and dimensions (qualitative data such as product categories or time periods).
- The technology provides support for a wide range of analytics functions, such as trend analysis and forecasting.

3-dimensional Cube in Data Warehousing - Illustration



- Summarizing sales by specific product, by time, and by store location.
- These are data cube dimensions.

Accessing Data Warehouse: Data Mining

- Data Mining is the process of discovering patterns, relationships, and insights using techniques from statistics, machine learning, and artificial intelligence.
- Data mining process, along with data warehousing, enables organizations can make data-driven decisions to improve business performance.
- Some of the use cases of data mining include identifying customer buying patterns for customer segmentation, detecting fraud in financial transactions, market basket analysis, predicting equipment failures in the manufacturing plant (Predictive maintenance), and recommender systems.

Data Mining Techniques

- Association Rule Mining - To identify relationships between variables in large data sets. It is often used in market basket analysis to identify which items are frequently purchased together.
- Text mining - Extract insights and knowledge from unstructured data such as text documents. Example use cases include social media and sentiment analysis.
- Regression: Algorithm to predict/forecast a continuous numerical value based on one or more input variables.
- Clustering: Algorithm to group similar items together based on their attributes or characteristics. Example use case includes segmentation and profiling.
- Neural Networks: This algorithm is used to simulate the functioning of the human brain and learn from data. It is often used for pattern recognition and prediction.

Modeling Techniques Requirement

- The success of a data warehouse project depends heavily on the quality of its data models.
- The two most commonly used techniques – Dimensional modeling and Entity Relationship (ER) modeling.
- Dimensional Modeling is a technique that organizes data into a series of interrelated fact and dimensions tables with a star or snowflake schema. It allows for optimized querying and analysis.
- ER modeling is a technique used to represent the relationships between data entities and attributes associated with those entities. ER models are often used in operational databases to support transaction processing and application development.

Modeling Techniques Requirement

- Scalability: Ability to handle a large amount of data (data warehouses typically store historical data and can grow to be very large).
- Performance: Allows for efficient querying and analysis of data to handle complex queries on large datasets.
- Flexibility: To handle changes in the data sources, business requirements, and user needs.
- Usability: Intuitive and easy to use.
- Maintainability: Allows for easy maintenance, updates, and modifications to the data warehouse over time.

Modeling Techniques Requirement

- Accuracy: Ensures the accuracy and integrity of the data stored.
- Consistency: Ensures that the data is consistent across different data sources and is stored in a standardized format.
- Integration: Allows for easy integration with other systems and applications within the organization.

Analysis Modeling

- Modeling refers to the process of defining and organizing the data in a way that is optimized for querying and reporting.
- Analysis modeling involves identifying key performance indicators (KPIs) and metrics that are important for business.
- It involves designing a schema that reflects the way data will be used in analysis rather than how it is stored in the operational system.
- Analysis modeling helps business users and analysts to gain insights, make data-driven decisions, and achieve their organizational goals.

Tool Selection

- The selection of the appropriate tool is a critical aspect of designing a successful data warehouse.
- Factors considered: Business requirements, cost, scalability, ease of use, speed, and compatibility with existing systems.
- Variety of tools available: ETL, Business Intelligence, and DBMS.
- Examples of popular tools include Informatica PowerCenter, IBM DataStage, Microsoft SQL Server, Oracle Data Integrator, Tableau, and Microsoft Power BI.

Summary

A brief recap:

- Metadata describes other data, including information about data sources, data types, relationships
- Accessing Data Warehouse is the process of retrieving data from a data warehouse for use in reporting, analysis, and other tasks
- Modeling technique requirements are the necessary skills, processes, and methods for developing effective data models in a data warehouse environment.
- Modeling and Analysis Modeling is the process of creating a data model that can be used for both storage and analysis purposes in the Data Warehouse.
- Tool Selection is the process of selecting the appropriate software tools to support data warehouse development, including ETL, modelling, reporting, and analysis tools.