Data Warehousing - Modeling

Agenda

Here, we will cover:

- Schemas
- Data Aggregation
- Data Explosion
- Dimensional Modeling
- Star and Snowflake Schema
- Transaction and Snapshot Schema
- Staging Area Modeling views of Bill Inmon and Ralph Kimball
- Hub and Spoke Approach

Schemas

- A schema is a logical structure used to organize data in a database.
- In data warehousing, schemas are used to organize data in a way that supports efficient querying and analysis.
- The three main types of schemas used in data warehousing: Star Schema, Snowflake Schema, and Galaxy Schema.
- Each schema has its own benefits and drawbacks, and the choice of schema depends on the specific requirement of the data warehousing project.
- Choosing the right schema is an important part of a successful data warehouse, as it can greatly impact the efficiency and effectiveness of data analysis.

Data Aggregation

- Data aggregation is the process of collecting and summarising data from multiple sources into a single dataset for analysis.
- It is used in data warehousing to improve efficiency and reduce storage requirements.
- In combination with general Aggregation techniques, it uses techniques such as data compression, dimensionality reduction, and summarization.
- Aggregation techniques, in general, focus on manipulating and presenting data at a higher level of granularity at the same time allowing for drill-downs, while data aggregation techniques in combination with aggregation techniques are focused on reducing the amount of data needed to be stored or processed for faster querying and efficiency.

Data Explosion

- As data volumes grow exponentially, data warehousing is facing new challenges in terms of storing, processing, and analyzing massive amounts of data.
- The emergence of big data, which includes data from various sources, including social media and IoT devices, has significantly contributed to this data explosion.
- To effectively manage data explosion in data warehousing, data architects need to adopt new approaches that allow handling of volume, variety, and velocity of data generated.
- Strategies such as data compression, data archiving, data virtualization, and cloud computing can help manage data explosion and ensure data warehouse remain efficient and effective.

Dimensional Modeling

- Dimensional Modeling is a technique that organizes data into a series of interrelated facts and dimension tables such as star or snowflake schema. It is different from Entity-relationship (ER) modeling, which involves techniques to represent relationships between data entities and attributes associated with those entities.
- Fact tables contain measurements or metrics, such as sales and transactions, and are typically large. Dimension tables contain descriptive information (attributes and characteristics) about the data in the fact tables, such as time, location, and product information, and are typically smaller in size.
- The goal of Dimensional modeling is to create a flexible and scalable data model that is optimized for querying and analysis, thus being more effective for data warehousing and business intelligence applications.

Star and Snowflake schema

- Star and snowflake schema is a type of dimensional modelling used in the data warehouse to organize and represent data.
- The schema model resembles a star, with one central table (fact table) surrounded by several related tables (dimension tables).
- Snowflake schema is an extension of star schema and involves the normalization of dimension tables. The schema model resembles a snowflake due to the presence of multiple related tables branching out from a central fact table.

Star and Snowflake schema

- Star schema is simpler to understand and provides better query performance because of its denormalized structure, which reduces the number of joins required to retrieve data.
- Snowflake schema allows for improved data accuracy, better accommodation of complex business requirements, and more efficient use of storage space.
- Star schema is easier to maintain because it has fewer tables and simple relationships.
- The choice between star and snowflake schema depends on specific business requirements and data characteristics – such as the complexity of the data model, query performance requirements, and maintenance considerations.

Transaction schema

- Transaction schema is a type of schema used in data warehousing that is optimized for handling high-volume transactional data.
- Tables in transaction schema are usually normalized to eliminate data redundancy. The schema involves a large number of tables with many relationships between them.
- In transaction schema updates, inserts and deletes are performed on individual records rather than on large groups of records. This design is effective for managing large amounts of data that are frequently updated, inserted, or deleted.
- The transaction schema is commonly used in operational databases (OLTP) and real-time systems. In other words, they are suited for situations where there is a high volume of transactions and queries are typically limited to small sets of records.

Snapshot schema

- Snapshot schema is a type of data warehouse schema used to store historical data that captures a point-in-time view of the business process.
- The schema contains a snapshot fact table, a current dimension table, and a history dimension table.
- The schema is designed to answer how the business looked at a specific point in time, thus allowing comparison between different periods.
- It is commonly used for tracking changes in dimensions, where a dimension represents a business entity such as a product, customer, or employee. It is best used for answering queries such as the analysis of sales data for a specific quarter along with the customer demographics and product information for that period.

Staging Areas Modeling views of Bill Inmon

- Bill Inmon, also known as the “father of data warehousing”, believes in building a single, integrated data warehouse that stores all the data in a normalized form.
- In this view, the staging area should be designed as a normalized operational data store (ODS) to integrate data from various sources. This is called the “top-down” approach.
- This data is then transformed into a common format and moved to the data warehouse for reporting and analysis. The data marts are created as subsets of the enterprise data warehouse
- The approach requires more time and resources to design and maintain but provides for a flexible and consistent view of data across the enterprise.

Staging Areas Modelling views of Ralph Kimball

- Ralph Kimball believes in building a dimensional data warehouse that stores data in a denormalized form for easy querying and analysis.
- In this view, the staging area should be designed as a dimensional staging area that mirrors the structure of the data warehouse. In other words, it emphasizes on building dimensional data marts first and then integrating them into a larger enterprise data warehouse. This approach is called the “bottom-up” approach.
- Data is extracted, transformed, and loaded (ETL) into the staging area and then transformed into dimensional structures for analysis and reporting.
- This approach takes less time and resources to design and maintain, but it can result in duplicated data and inconsistent views across the enterprise.

Comparison of Inmon and Kimball approach

- Inmon’s approach places a strong emphasis on data quality, consistency, and integration and is usually used in large, complex organizations with significant data integration needs.
- Kimball’s approach places a strong emphasis on user requirements and business needs and is often used in smaller or mid-sized organizations, requiring more agility and flexibility in their data warehousing solutions.
- Some organizations might find a hybrid approach that combines elements of both Inmon and Kimball. The approach is sometimes referred to as the “hub-and-spoke” approach as it combines the central hub model of the Inmon approach with the fact and dimension tables of the Kimball approach.

Hub and Spoke Approach

- In this hybrid approach, data is first extracted from various sources and transformed into a centralized hub, which is modeled using the Inmon approach. This hub is then used to feed multiple data marts, which are modeled using Kimball's approach. The data marts are designed around specific business functions or departments and contain fact tables and dimensions.
- The advantage of the approach is – it is a flexible and scalable framework that can adapt to changing business needs over time. The centralized hub provides a single source of truth for the organization, while data marts can be added or modified as needed.
- The approach requires a higher level of complexity in the design and implementation process. It can be more challenging to maintain data consistency and accuracy across the different components of the system.

Summary

A brief recap:

- Schemas are logical structures to represent data in a database or a data warehouse.
- Data Aggregation consists of groups of techniques to combine data from multiple sources or records into a single summarised view.
- Data Explosion is a phenomenon where the volume of data being generated and stored increases at an exponential rate.
- Dimensional Modeling technique is used to organize data in the data warehouse in a way that is optimized for querying and analysis.
- Star and Snowflake Schemas are two types of dimensional modeling schemas used in data warehousing, with Star Schema being simpler and Snowflake Schema being more normalized.

Summary

- Transaction and snapshot schema are types of schema to organize time-related data in a warehouse
- Staging Area Modeling is a process of structuring and organizing data in a staging area before loading it into a data warehouse.
- Inmon and Kimball have provided different approaches to data warehousing, with Inmon advocating a top-down approach and Kimball advocating a bottom-up approach.