Explore Data Warehouses

2023-08-01

Question 1:

Data warehouses are specialized databases that are designed for analytical processing and reporting. They are used to store large amounts of historical data from various sources, such as transactional databases, logs, and external systems. To construct a data warehouse in a relational database, two key concepts are often used: Fact tables and star schemas.

Fact Tables: A fact table is a central table in a data warehouse that contains quantitative data (facts) related to a specific event or business process. It typically consists of foreign keys referencing dimension tables and numerical measures. For example, in a retail data warehouse, a fact table could contain sales data with foreign keys referencing dimensions like product, time, and store. Fact tables store aggregated data that can be used for generating reports and performing data analysis efficiently.

Star Schemas: A star schema is a type of database schema commonly used in data warehouses. It consists of a central fact table connected to multiple dimension tables, forming a star-like shape when visualized. The fact table and dimension tables are linked through foreign key relationships. The advantage of a star schema is that it simplifies queries and enables fast data retrieval for analytical purposes. Each dimension represents a specific attribute or characteristic, and the fact table contains the numerical data related to those dimensions.

Regarding using a transactional database for OLAP (Online Analytical Processing), it is generally not recommended. Transactional databases are designed for efficient and reliable data storage and processing of day-to-day business operations, emphasizing data integrity and consistency. OLAP, on the other hand, focuses on complex data analysis and reporting. OLAP queries require aggregations and extensive joins, which can significantly impact the performance of a transactional database, leading to slower response times for operational tasks. It's best to separate OLAP and transactional databases to ensure optimal performance for both types of operations.

Question 2:

Data Warehouse: A data warehouse is a large, centralized repository that stores structured, historical data from different sources for reporting and analysis. It is designed to support decision-making processes and business intelligence activities. Data warehouses often use techniques like ETL (Extract, Transform, Load) to gather data from various sources and transform it into a consistent format suitable for analysis. Examples of data warehouses include Amazon Redshift, Google BigQuery, and Microsoft Azure Synapse Analytics.

Data Mart: A data mart is a subset of a data warehouse that focuses on a specific business function or department within an organization. Unlike a data warehouse, which caters to the entire enterprise, data marts are smaller, more focused databases. They provide data to a specific group of users, such as sales, marketing, or finance, and are designed to support the analytical needs of that particular group. Data marts are often created to address specific business requirements and provide faster access to relevant data. For example, an organization may have separate data marts for sales analytics, marketing analytics, and finance analytics.

Data Lake: A data lake is a storage repository that holds a vast amount of raw, unstructured, and structured data. Unlike data warehouses and data marts, data lakes do not require a predefined schema, allowing data to be stored as-is. Data lakes are designed to store data from various sources, including logs, social media, IoT devices, and more. The data can be used for both batch processing and real-time analytics. Data lakes

provide a flexible and scalable solution for data storage and analysis. Examples of data lakes include Amazon S3, Azure Data Lake Storage, and Hadoop-based systems.

Example Use Case: Suppose a retail company wants to analyze sales data to gain insights into customer behavior and product performance. They could have a data warehouse to store historical sales data from various channels, a data mart for sales analytics accessible to the sales team, and a data lake to store raw data from social media and customer feedback.

Here's an article that explains the differences in detail:

embed_url("https://www.youtube.com/watch?v=GHpcLEkkmLc")

Question 3:

For the bird strike database in Practicum I, we can design a fact table to analyze the frequency and impact of bird strikes at different airports. The fact table could be named "BirdStrikeFact" and contain the following columns:

1. BirdStrikeID (Primary key): A unique identifier for each bird strike incident.
2. AirportID (Foreign key): References the Airport dimension table, representing the airport where the bird strike occurred.
3. DateID (Foreign key): References the Date dimension table, representing the date when the bird strike occurred.
4. BirdSpeciesID (Foreign key): References the BirdSpecies dimension table, representing the species of the bird involved in the strike.
5. AircraftID (Foreign key): References the Aircraft dimension table, representing the aircraft involved in the strike.
6. WildlifeSize: A numerical measure representing the severity of the bird strike (e.g., minor, moderate, severe).
7. DamageExtent: A numerical measure representing the extent of damage caused by the bird strike.

The fact table will store individual bird strike incidents, and the foreign keys will link to the respective dimension tables for detailed information about the airport, date, bird species, and aircraft involved. This design allows for efficient querying and analysis of bird strike data by various dimensions, such as time, location, bird species, and aircraft type.

Please note that the above design is just one possible approach to constructing a fact table for bird strike analytics. The specific design and attributes may vary based on the requirements and business goals of the analysis.