Explore Data Warehouses

Bhavitha Kandru

Question 1:

Data warehouses are often constructed using relational databases. Explain the use of fact tables and star schemas to construct a data warehouse in a relational database. Also comment on whether a transactional database can and should be used to house an OLAP. Lastly, why would an organization use a relational database with a star schema rather than a dedicated NoSQL database specifically engineered for data warehousing and analytics?

Answer

Data warehouses, as specialized repositories, excel in optimizing read-heavy operations, handling complex queries across vast datasets, and conducting time-series analyses. Their design is tailored to facilitate decision-making through analytical queries and data analysis. To efficiently organize data for analytical purposes, data warehouses often employ structured schemas and methodologies that leverage relational databases. This approach, with its focus on efficiency, instills confidence in the decision-making process. Two key concepts in this realm are fact tables and star schemas.

Fact Tables and Star Schemas:

A fact table is the central table in the star schema of a data warehouse. It contains measurable, quantitative data for analysis, typically transactional or event data. Fact tables store different metrics, known as facts, which can be additive (like sales amount, which can be summed up) or non-additive (like temperature). These tables also contain foreign keys to dimension tables, which provide descriptive attributes related to the facts.

The star schema, a popular data modeling approach used in data warehousing, features a central fact table surrounded by dimension tables. Dimension tables, linked to the fact table through foreign keys, contain descriptive attributes or dimensions that can be used for filtering, grouping, and labeling the facts (e.g., date, product, customer). This structure, with its intuitive nature, makes it easier for analysts to understand the data layout, fostering a sense of comfort and familiarity.

Transactional Database for OLAP:

A transactional or Online Transaction Processing (OLTP) database is optimized for handling a high volume of transactions, such as insert, update, and delete operations. While technically possible, using a transactional database to house an Online Analytical Processing (OLAP) system is not recommended due to differing optimization goals. OLTP systems are optimized for fast, atomic transactions and data integrity in operational tasks, supporting many short transactions. In contrast, OLAP systems are optimized for read-heavy operations, complex queries, and aggregations over large datasets, requiring a different performance tuning than OLTP systems.

Relational Database with Star Schema vs. NoSQL for Data Warehousing:

For several reasons, organizations may choose a relational database with a star schema over a dedicated NoSQL database for data warehousing and analytics. Relational databases have been around longer and offer a mature suite of tools for data management, including sophisticated mechanisms for data integrity, transactions, and security. SQL is a powerful and widely used language for data analysis; relational databases support SQL natively, making it easier to perform complex queries and analyses without learning new query languages or paradigms. Some organizations prefer the balanced relational databases offer. They can handle both transactional and analytical workloads to a certain extent, even if it means sacrificing some performance for analytical queries compared to dedicated NoSQL solutions. Organizations with existing relational databases and in-house expertise may find extending their current systems to support data warehousing more cost-effective than investing in new technologies and training.

However, dedicated NoSQL databases engineered for data warehousing and analytics offer a distinct advantage. They provide superior performance, scalability, and flexibility for specific use cases, particularly those involving unstructured data or requiring horizontal scaling beyond what relational databases can efficiently manage. This flexibility empowers organizations to make technology choices that align with their specific requirements, existing infrastructure, and strategic goals.

Question 2:

Explain the difference between a data warehouse, a data mart, and a data lake. Provide at least one example of their use from your experience or find out how they are generally used in practice. Find at least one video, article, or tutorial online that explains the differences and embed that into your notebook.

Answer:

Data warehouses, data marts, and data lakes are all systems for storing and managing vast volumes of data, but they differ in data type and organization. A data warehouse, a data mart, and a data lake are examples of storage solutions that handle and store enormous amounts of data. While they have some similarities, there are significant variances in their purpose, design, and usage.

Data warehouses:

Data warehouses are centralized, integrated data repositories for corporate information and decision-making. It usually entails taking data from several sources, converting it into a common format, and loading it into a structured database schema. The data is structured around established dimensions and topic areas, and it is designed for advanced querying and analysis .For example, A retail organization may utilize a data warehouse to examine sales patterns across several items, geographies, and periods.

Data mart:

A data mart is a subset of a data warehouse intended to support a specific business unit or department. This subset of data from the data warehouse is customized to particular user needs. Data marts are easier to construct and manage than full-scale data warehouses because they require less complicated data integration and can be tailored to specific use cases. For example, A marketing department may utilize a data mart to assess consumer behavior and campaign efficacy.

Data Lake:

Structured data refers to data that is organized into a specific format, making it easy to search and analyze. Unstructured data, on the other hand, does not have a predefined format and can include things like text, images, and videos. A data lake is a storage repository that enables businesses to store and handle massive amounts of both structured and unstructured data at scale. Unlike a data warehouse or a data mart, a data lake does not impose a preset schema or data structure and can store data in its raw form. This makes it more adaptable and scalable than traditional data storage options, but more difficult to administer and query. Data lakes are commonly used for exploratory analytics and machine learning, which involve analyzing massive volumes of data to discover new insights and patterns. For example, A healthcare institution may utilize a data lake to store and analyze patient information, medical pictures, and other health data to improve results and create novel therapies.

Click here to watch the video

Click here to read the article

Data Warehouses, Data Marts, and Data Lakes are three data storage solutions i.e;

Data Warehouses are primarily used for business intelligence and visualizations. They offer fast query performance and highly curated data for analytical insights.

Data Marts support specific business units with tailored data access, optimizing query and reporting performance.

Data Lakes provides a flexible and cost-effective solution for storing extensive data. They are ideal for machine learning, exploratory analytics, and big data projects.

Most large organizations combine these solutions to meet their diverse data storage and analysis requirements.

Question 3:

After the general explanation of fact tables and star schemas, design an appropriate fact table for Practicum I's bird strike database. Of course, there are many fact tables one could build, so pick some analytics problem and design a fact table for that. For example, you might want to build a fact table for "birdstrike facts", such as the average and total number of bird strikes by time periods (weeks, months, years), or bird strikes by region or airport and, again, broken down by time period. This might be useful for time series analysis or to determine seasonality for bird strikes which might then inform notices to pilots. Be sure to explain your approach and design reasons.

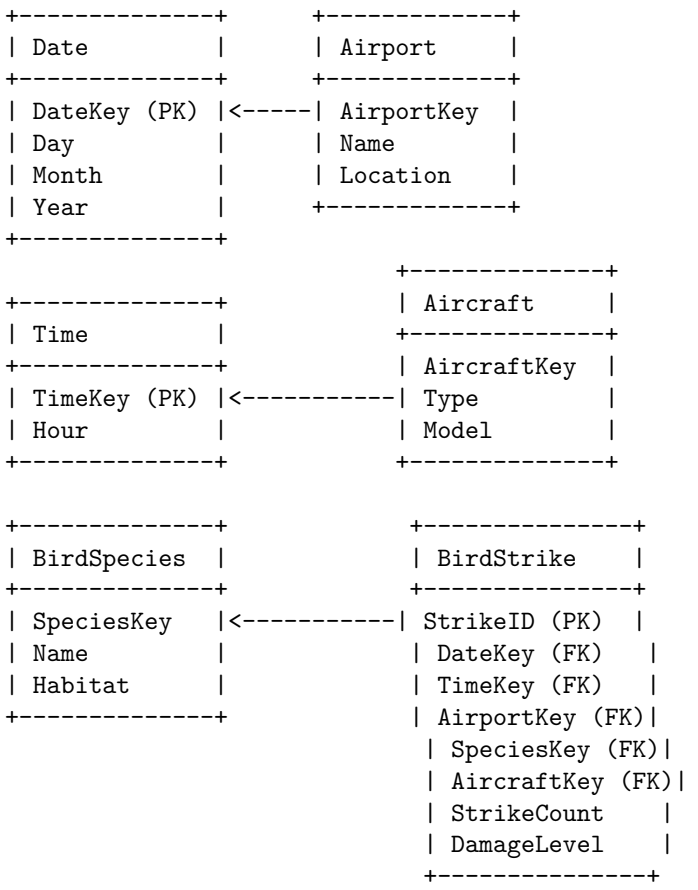
Just design it (ideally by creating an ERD for it); you do not need to actually implement it or populate it with data (of course, you may do so if you wish in preparation for the next practicum).

Answer:

I have designed a fact table for the bird strike database focusing on analytics to understand patterns and seasonality. The fact table captures essential data for each bird strike incident. Additionally, several dimension tables provide context and details for more in-depth analysis.

The design approach aims to facilitate time series analysis and regional comparisons crucial in predicting bird strike risks and implementing preventive measures. The fact table is linked to dimension tables such as date, time, airport, bird species, and aircraft involved. This is done through shared keys, allowing for multifaceted analysis across different dimensions.

For instance, linking the fact table to a date dimension allows us to analyze time series to identify trends and seasonality in bird strike incidents. The airport and region dimensions allow for analyzing bird strikes by location and identifying high-risk areas or airports. Understanding which bird species are most involved in strikes and how different aircraft types are affected can help tailor preventive measures.



This Entity Relationship Diagram (ERD) guides organizing and examining bird strike data. The BirdStrike fact table is the central element of this schema, which contains information about individual bird strike incidents. The Date, Time, Airport, BirdSpecies, and Aircraft tables are dimension tables that provide descriptive details about each dimension. This schema can handle various queries, from simple counts and aggregations to more advanced analyses involving temporal patterns, geographic distribution, and interactions between species and aircraft types. This design enables the extraction of valuable insights that can be used to improve aviation safety and reduce the risks associated with bird strikes.