Explore Data Warehouses

Q1

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 OLAP.

Answer for Q1

A fact table is a table in a data warehouse that contains the measurements, metrics, or facts of a business process, such as sales revenue, quantity sold, or profit. Fact tables also contain foreign keys that link to the dimension tables in a star schema.

A star schema is a type of database schema where a fact table is at the center and connected to multiple dimension tables through foreign keys(like a star). The dimension tables contain descriptive information, such as date, product, store, and customer, that provide context for the facts in the fact table. Moreover, the dimension tables are expected to be much smaller in size compared to fact tables.

The star schema is a popular design for a data warehouse because it simplifies querying and reporting. The schema allows users to easily aggregate and analyze data across multiple dimensions. Because the schema is denormalized, it can also improve performance by reducing the number of joins required to retrieve data.

Transactional databases are optimized for running production systems—everything from websites to banks to retail stores. These databases excel at reading and writing individual rows of data very quickly while maintaining data integrity. Although a transactional database can technically be used for OLAP, it shouldn't be used to OLAP. There're several reasons: (1) OLAP requires denormalized schema to improve reading large volume of data(by reducing the joining operation), while transactional databases usually contain normalized schemas. (2) Transactional databases usually overwrite previous values with the most current version of a transaction, and thus it is hard to track changes over time, which is important for analytics.

Q2

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 how you believe they might be used in practice. Find at least one video, article, or tutorial online that explains the differences and embed that into your notebook.

1. Data Warehouse

- Description: A data warehouse is a centralized repository that stores data(predefined structured data) from various sources, organizes it, and prepares it for analysis. It is designed to support business intelligence activities, such as reporting, data mining, and data analysis. Data warehouses are typically used by large enterprises with complex data management needs, with a focus on data consistency and accuracy.
- Example: Amazon may use a data warehouse to store and analyze data from various sources, such as sales data, customer data, and inventory data. With a data warehouse, the company can store the structured sales data in a predefined schema that is optimized to perform complex queries and analysis,

which allows the company to gain insights into customer behavior, inventory management, and sales performance across multiple stores and regions.

2. Data Mart

- Description: A data mart is a subset of a data warehouse that focuses on a specific department, function, or business unit within an organization. It is designed to provide a more targeted view of data and enable faster decision-making. Data marts are typically used by smaller organizations or departments within large enterprises.
- Example: A marketing department within Amazon may use a data mart to store and analyze data related to customer behavior, campaign performance, and lead generation. The data mart would allow the department to quickly identify trends and make data-driven decisions to improve marketing effectiveness.

3. Data Lake

- Description: A data lake is a large repository that stores raw, unstructured, and semi-structured data. It is designed to support big data analytics, machine learning, and other advanced analytics applications. Data lakes uses low-cost storage compared to data warehouses, with a focus on scalability and flexibility.
- Example: A social media company wants to collect and analyze large volumes of customer data, including data from social media platforms, website clickstream data, and customer service logs. The data comes in a variety of formats and is constantly evolving as new sources of data emerge. In this scenario, a data lake would be preferable to use than a data warehouse. With a data lake, the company can store all the customer data in its raw form, without having to define a schema in advance. This allows the company to easily add new types of data as they become available, and to store the data at a lower cost compared to a structured data warehouse. The company can then use big data tools and technologies, such as Apache Spark or Hadoop, to analyze the data and derive insights.

References

- Article: https://aws.amazon.com/compare/the-difference-between-a-data-warehouse-data-lake-and-data-mart/?nc1=h_ls
- Video: https://www.youtube.com/watch?v=hYP8xfGpKHs

Q3

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. Be sure to explain your approach and design reasons. Just design it (perhaps draw 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 for Q3

ERD: https://i.imgur.com/NJbDKFV.jpg

The factBirdStrikes table can be used as the fact table, as it contains measurement(incident_count) and contains foreign keys that link to the dimension tables in a star schema, which will provide additional details. A sample record for factBirdStrikes would be (1,3,4,5, "2000-08-14", 5), which represents there're 5 incidents happened on 2000-08-14, while the related sky__condition was 5, flight__phase was 4 and aircraft__type was 3.

The fact table's measure would be the count of incidents for each date, aircraft type, flight phase and sky condition combination. Notice that I remove some fields such as airport__name, state and warning, since they're irrelevant to the analysis problem. This can improve the performance of queries and reduce storage requirements, since the star schema is typically denormalized and optimized for query performance.

This fact table can help answer questions such as:
1. What's the relationship between the frequency of incidents and the flight phase;
2. Which quarter had the most incidents in year 2010;
3. Which sky condition has the least incidents when the aircraft type is military.