CS5200.ExploreDataWarehouses.LuC

Chenhao Lu

2023-04-10

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 OLAP.

A data warehouses are databases that consolidate data from multiple sources, augmented with summary information, and historical data over a long time period. They are primary queried and non-volatile. They contain data records for a long time and provide users with a single consolidated interface to data, enabling decision-support systems.
A fact table in a data warehouse is a table that contains all the facts, measurements or metrics of the attributes. We use it to store the primary / foreign keys, which is necessary in most relational database.
Star schema is the database organization structure in the dataware house. It stores the attribute of the database constructs teh entity of relational database.
For a transactional database, we can but in many cases should avoid use OLAP. Because transactional database store the table rows together and OLAP store columns together. So they are for different purposes.

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 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.

Data warehouse, as discussed in question 1, is database that consolidates data from multiple sources and can be applied to different types of use-cases. For example, in a large retail corporation, data warehouse is used for consolidating and storing data from various sources, such as inventory, sale information and customer information. And it will be used for generating reports, conducting inventory management, and so on.
Data mart, on the hand, is actually a subset of data warehouse. It is geared towards the needs of a particular user. For example, in a large retail corporation, sales department could have a data mart to store all their department's data such as inventory and invoice data.
Data lake is the collection of all forms of data either structured or unstructured from various sources. Basically, all the data are in their raw form, without any cleaning or consolidation. For example, for a large tech company collects all customer information, log files, social media data, images, videos in a data lake for future usage.
In industry and real design scenario, data warehouses, data marts, and data lake also have differences in size, data detail, and so on.
AWS has a well-explained article regarding the difference between a data warehouse, data lake, and data mart.
Link: https://aws.amazon.com/compare/the-difference-between-a-data-warehouse-data-lake-and-data-mart/#:~:text=It%20is%20a%20central%20repository,raw%20data%20and%20unstructured%20data.

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. Be sure to explain your approach and design reasons.

Based on the bird strike database, a research team wants to find out which bird species casued the most accidents for future studies. Therefore, they came up a fact table based on their needs:

FactTable with two attributes: birdStrikeId, altitudeFt and four foreign keys for four dimension tables.
Dimension table 1: speciesDetail, PK sId, attribute: species.
Dimension table 2: flightDetail, PK fId, attribute: flightPhase, flightDate, damage.
Dimension table 3: landingDetail, PK lId, attribute: originAirport.
Dimension table 4: flightCondition, PK cId, attribute: skyCondition.
Fact table and dimension tables only consist all the required attribute for the analytical questions. Therefore, it can save some time for analysis and database management.