Explore Data Warehouses

Yangli Liu

2023-04-12

Question 1

According to the course material, data warehouses are databases that consolidate data from multiple sources, augmented with summary information, and historical data over a long time period. Fact tables and star schemas are two commonly used techniques to construct a data warehouse in a relational database.

Fact tables contain the fundamental measurements of the enterprise, and they are the ultimate target of most data warehouse queries. These measurements are usually numeric and can be aggregated over different dimensions, such as time, location, product, or customer. Fact tables typically have many foreign keys to dimension tables, which contain descriptive information about the measurements. Star schemas are a type of data warehouse schema that consists of a central fact table connected to multiple dimension tables, which are arranged in a star-like shape. The fact table is at the center of the star, while the dimension tables contain the descriptive attributes that provide context to the measures.

It is not recommended to use a transactional database for OLAP (Online Analytical Processing). The reason is based on: 1. The transactional databases are optimized for fast write operations and high concurrency while OLAP requires complex read operations and aggregation. 2. OLAP usually will use a large amount of data and sometimes need to perform some complex procedures like sorting, grouping or filtering. These procedures all likely impact the performance of the transactional database.

Question 2

Data warehouse, data mart, and data lake are all different approaches to storing and managing data for analytics purposes. I shall explain each of them and provide a example:

1. Data warehouse: A data warehouse(DW) is a repository of suitable operational data(data that documents the everyday operations of an organization) gathered from multiple sources, stored under a unified schema, at a single site. It can successfully answer any ad hoc, complex, statistical or analytical queries. The data once gathered can be stored for a longer period allowing access to historical data. The data warehouses provide the user a single consolidated interface to data, which makes decision-support queries easier to write. The example can be amazon. This company's retail department wants to analyze its sales data to gain some insights about customer behavior, product trends and regional performance. Then Amazon can use a data warehouse to store data from their sale systems, customer relationship management software or their marketing departments. Then this data warehouse can be arranged to a star schema, with a fact table contains sales data and dimension tables contains customer, product and location data. Afterwards, Amazon analytical department will be able to perform some complext analysis based on those tables.

2. Data mart: A data mart is a subset of data warehouse that is designed for a particular department of an organization such as sales, marketing, or finance. It can be optimized for specific bussiness needs or reporting some special requirements. Based on Amazon again. If the company would like to know some specific reporting, Amazon can builds a data mart that is focusing on marketing data. For example,

the campaign performance, open email rates and their social media engagement. Those marketing data can then be optimized for specific reporting requirements of their marketing department.

3. Data lake: A data lake is a large, centralized repository of raw, unstructured, and semi-structured data that is stored in its native format. The purpose of a data lake is to enable data scientists, analysts, and other users to explore and analyze data without predefined schemas or structures. Data lakes are typically built using big data technologies. Amazon could build a data lake that stores raw, unstructured or semi-structured data. For example, they can store data like clickstream data or website logs. Then some experts can use the data lake to explore and analyze data using some big data tools or techniques. This will help the company to understand some user behavior patterns or the trendy products in their website.

Here's an article link that you can click on for explaining the differences between a data warehouse, a data mart, and a data lake: https://www.indeed.com/career-advice/career-development/data-lake-vs-data-warehouse-vs-data-mart

Question 3

Based on the bird strike database schema, a suitable fact table for this database could be the "incidents_fact" table.

The "incidents_fact" table would contain foreign keys to the dimension tables, "airports_dim" and "conditions_dim". The primary key of the table would be the "rid" column from the "incidents" table.

The columns in the "incidents_fact" table would be measures related to bird strikes. Some possible measures include the species of birds involved in the incident, the altitude and flightPhase of the incident, the extent of damage to the aircraft(is it heavy or not). The "incidents_fact" table could be used to analyze patterns and trends related to bird strikes. For example, the table could be used to determine frequency of bird strikes by (airline, aircraft, flightPhase), which bird species caused the heavy incidents(species, heavyFlag), and how different conditions (e.g. heavyFlag, conditions, altitude, flightPhase) affect the Severity of a bird strike.

Overall, the "incidents_fact" table would provide a central location for storing and analyzing data related to bird strikes, allowing stakeholders to make informed decisions about how to minimize the risk of bird strikes and their impact on flights. An ERD for the bird strike database schema and "incidents_fact" table is shown below:

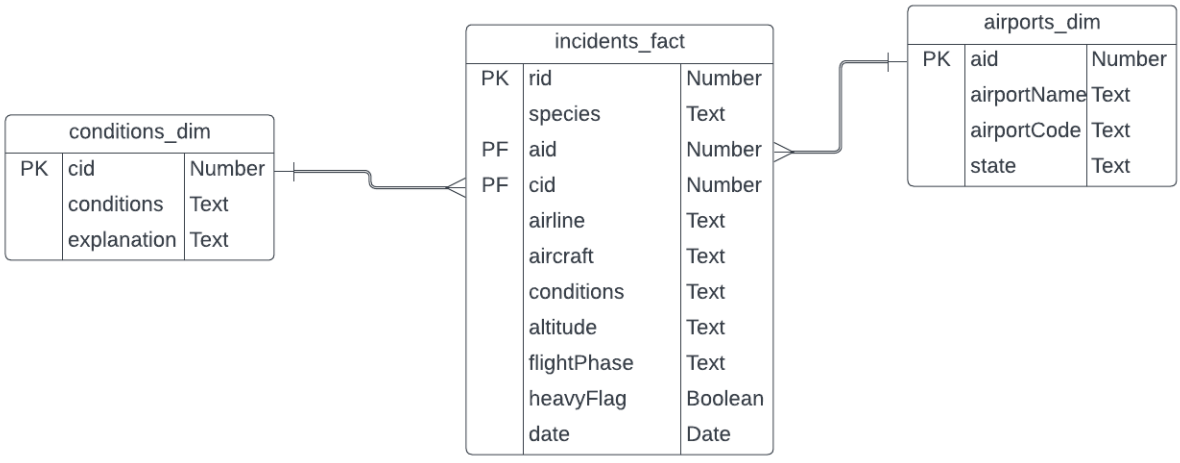

Figure 1: incidents_fact