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.

A1) A data warehouse is a collection of large amount of data that can be used to make informed decisions by the organization. It is a single source of data truth. A single source of data truth is structuring of the best quality of data in one place.

Features Of A Data Warehouse

1) Subject Oriented - The data is related to a particular subject

2) Integrated - Different people have different conventions when it comes to naming conventions. Hence common standards are described so that the data warehouse picks the best quality data.

3) Time Variant - As the data warehouse is used for analysis and reporting we should know the insights of the previuos data hence the data warehouse should store hostorical data.

4) Non - Volatile - The data flows in the data warehouse as it is. Once in the data warehouse it cannot be changed or deleted.

5) Summarized - The data is aggregated or segmented so that it can be easily used for data analysis.

Data flows into a data warehouse from transactional systems, relational databases, and other sources.

Fact Table

A Fact table in a Data Warehouse system is a table that contains all the facts, measurements or metrics of the attributes. Fact tables work with dimension tables. The fact tables store the numeric value and the dimension attribute value which is the foreign keys.

The Three Types of Fact Tables

1) Transactional – Transactional fact table is the most basic one that each grain associated with it indicated as "one row per line in a transaction"

2) Periodic snapshots – Periodic snapshots fact table stores the data that is a snapshot in a period of time.

3) Accumulating snapshots – The accumulating snapshots fact table describes the activity of a business process that has a clear beginning and end.

Fact table is necessary in the data warehouse as it stores all the foreign keys or primary keys in some cases and these keys are necessary to identify the data present in the table.

Star schema

It is database organization structure for use in the data warehouse. The fact table is present in the center of the star schema and the dimension tables form the points of the star. It stores the attributes of the data. Star schema is necessary for the data warehouse as it contains the attributes which describe the entity of the database.

Transactional databases store the table rows together which is helpful when you are constantly accessing whole rows and OLAP databases store table columns together which is used when we aggregate field values.

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.

A2) Differences between Data Warehouse, Data Mart and Dat Lake

Data Warehouse

It stores data that is already structired and modeled. The data presnet can be used for all types of use-cases.

Example : The product information - the manufacturing, the purchase, the sales, the cost, etc all details are maintained in a data warehouse in a structured manner.

Data Mart

It is basically a mechanism to extract data from the data warehouse. It is basically a subset of a data warehouse. It contains summarized data collected for the analysis of a specific subject.

Example : The sales department of a company

Data Lake

A data lake contains all forms of data both stuructured and unstructred data is present in a data lake. All the data is collected and later analyzed what has to be done with it. It is a cheaper way to store different tyes of data in large quantities.

Example - Twitter in the B2C space (They have text (Tweets), Images, Videos, Links, Direct Messages, Live Streams, etc.)

A Video explanig the difference between a database, data warehouse, data mart and a data lake.

```
library(vembedr)
embed_url("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).

Fact and Dimension Tables ERD

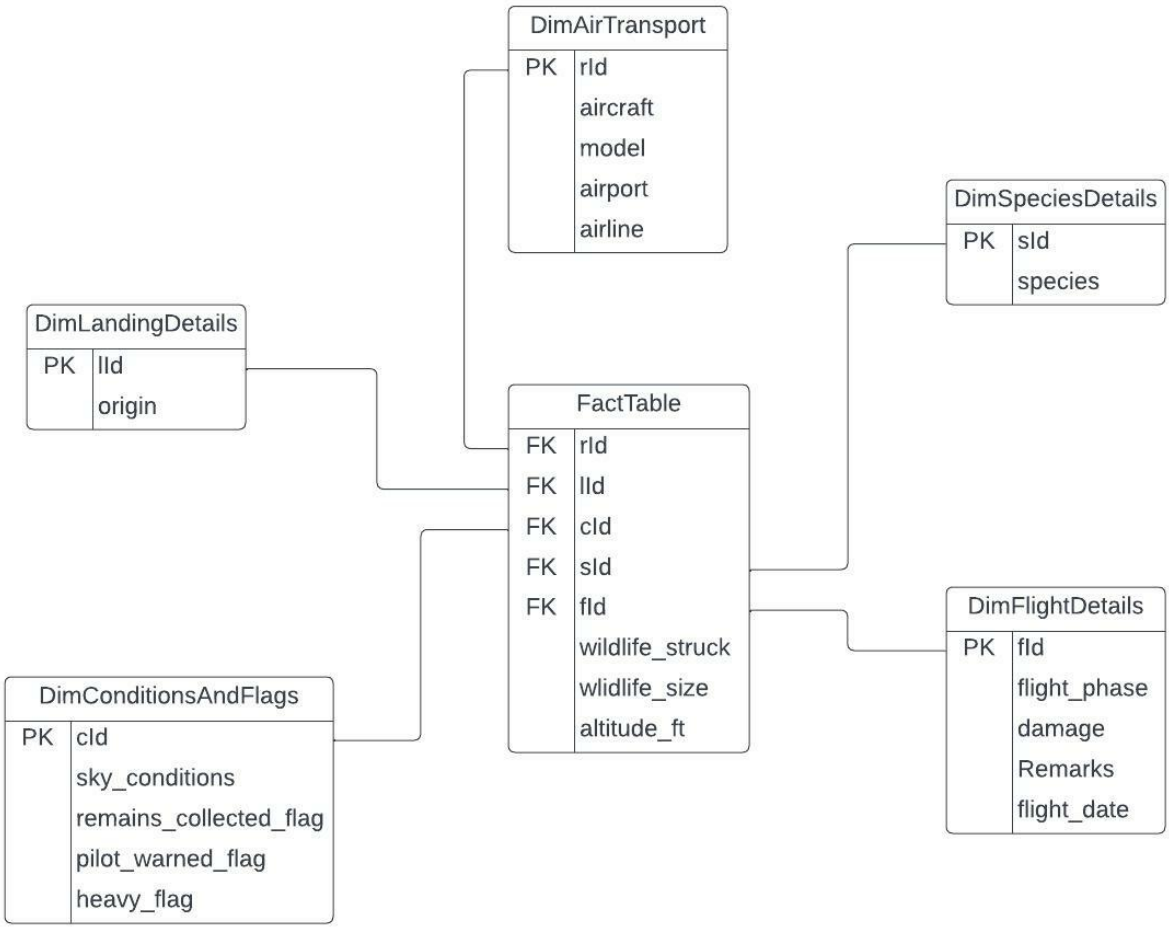

Figure 1: Fact Table ERD

The fact table has been created for thr Bird Strike CSV (Practicum 1).I have build a fact table along with it's dimension tables. The fact table has all the foreign keys which are primary keys of the dimension table. The fact table has the number of wildlife struck which is a quantity, wildlife size, altitude_ft which are all quantities and measure and hence should be present in the fact table and not the dimension table. The remaining attributes are present in different dimension tables.