Do: Explore Data Warehouses Assignment

Mansi Pravin Thanki (002128043)

Question 1 (30 Points)

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:

1. The main goal of data warehouse is to provide a holistic view while making some business-oriented decisions.
2. The multi-dimensioned nature of data which is present in data warehouses can be represented using star schema.
3. Fact tables are present at the center of the star schema and they are quite heavily loaded with data.
4. Fact Table: The tables that store the multidimensional data are known as fact tables.
5. Fact tables can be said to be the foundation or the building block of data warehouse.
6. The fact tables are large and it contains the primary information in a data warehouse
7. The fact table is a repository of the numerical values that are taken during the measurement event
8. The star schema consists of a fact table with a single table for each dimension (which are also known as lookup tables)
9. These dimension tabled (lookup tables) contain the attributes which are descriptive in nature and are used for the filtering purpose.
10. The fact tables contains all the keys that were present in the different dimension tables connecting to it, and additionally contains attributes that contains aggregated values.
11. Major advantages of star schema and fact tables are:

- Star schema provides a streamlined and faster access for the Online Analytical Processing (OLAP) queries. This helps in improved support for the Business Intelligence.
- The presence of redundancy in star schema is not very high, but it is enough to make querying process faster to provide support for data mining.
- Precomputing of data
- Recommended to use whenever query revolves around a particular event
- The sophisticated end users can use star schema in the analytical tools for generating custom reports
- the fact tables contain pre-joined rows and the data is already aggregated.
- Tasks like finding the average, or sum or count and join-minimization is made easier using star schema.

Can and should transactional database be used to OLAP:

- No, transactional databse should not be used to OLAP
- Transactional databases are updated on frequent and real-time basis. However, fact tables are unsuitable for Online Transaction Processing (OLTP) due to these frequent updates.
- For OLAP, for the purpose of analysis, denormalized database is used. However, transactional databases are usually normalized to provide efficiency.
- Therefore, transactional database is not recommended to be used to OLAP.

References:

1] https://northeastern.instructure.com/courses/110675/pages/5-min-read-kimball-r-2008-fact-tables-in-dimensional-modeling?module_item_id=7344777

2] https://onedrive.live.com/view.aspx?resid=99666BAF021553F1!70430&ithint=file%2cpptx&authkey=!AMjat-Smd7pWtmQ

Question 2 (30 Points)

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.

Answer:

The common property for data mart, data warehouse and data lake is the fact that all three of them store data. However, a lot of difference when it comes to the purpose that each of them address.

A] DATA SOURCES:

i) Data Warehouse: A data warehouse is an amalgamation of data from numerous sources.
ii) Data Lake: A data lake contains data from various sources and in various formats
iii) Data Marts: The source of data is usually from Data warehouse and addresses a particular use case

B] USAGE PURPOSE:

i) Data Warehouse: It helps in making decision-oriented decisions as the data is queried and is not volatile
ii) Data Lakes: Since amount of data is huge in data lakes, the data is used for big data analytics
iii) Data Marts : Data marts are nothing but subsets of data warehouse which serves purpose to a particular use case/business function. The data is used by a particular department for analytics.

C] DATA FORMAT:

i) Data Warehouse: Data stored is in multidimensional structured format
ii) Data Lakes: A data lake contains a lot of raw data which can be in unstructured format/semi-structured/structured format. Data is stored regardless of it's format.
iii) Data Marts: Data is stored in structured format

D] VOLUME OF DATA:

i) Data Warehouse: Stores large volume of data, but the volume is less than data lakes
ii) Data Lakes: Stores extremely high volume of data
iii) Data Marts: Stores less volume of data than data warehouses as it sticks to department specific data

E] COST:

i) Data Warehouse: Cost is less than that of data lakes.
ii) Data Lakes: Stores extremely high volume of data so the cost is VERY EXPENSIVE
iii) Data Marts: Cost is lesser compared to data warehouses

EXAMPLE:

1. DATA WAREHOUSE:

Social Media industries can use data warehouses for targetted marketing, so they can track user clicks data, video watched timestamps etc and analyse buyers behaviour

2. DATA LAKES:

Social media based companies like Instagram, Facebook etc need to store the posts data, comments, likes, videos etc. A data lake would be useful to store such vast amounts of data and derive analytical insights.

3. DATA MARTS:

A sales data mart can be used by sales department which would contain month-round, year-round sales data, sales team performance data, location based sales etc.

Video Link For Concepts:

https://www.youtube.com/watch?v=hYP8xfGpKHs

References:

1] https://onedrive.live.com/view.aspx?resid=99666BAF021553F1!70430&ithint=file%2cpptx&authkey=!AMjat-Smd7pWtmQ

Question 3 (40 Points)

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:

Analytical Problem:

Finding the number of incidents for each airline for every month sorted according to their impact and damage.

The impact and damage columns are not numerical (they are categorical) and hence will be first converted into numerical format based on their impact and damage intensity.

For eg: No Damage -> 0, Damage caused -> 1 The summation of damage caused across every months will give out the required information.

Link to the ERD Diagram (Star Schema):

https://lucid.app/lucidchart/f8b95ae4-99b9-49e3-af9b-2c3fdc73a971/edit?invitationId=inv_f6ddb715-dee8-4c7e-bf30-b7ffcc7a57d1

The fact table is IncidentCountFact. The dimension tables of star schema are:

1. Airport_Dim
2. Airline_Dim
3. Model_Dim
4. Date_Dim

dateid, airlineid, airportid, modelid are the foreign keys in the fact table for the above mentioned dimension tables.