**Jennifer Faer**

**IS 5313 Structured Data and Querying**

**Dr. Bill Rosener**

**Fall 2024**

[**https://billrosener.com/teaching/is4293/assignments/assignment7/assignment7-description.html**](https://billrosener.com/teaching/is4293/assignments/assignment7/assignment7-description.html)

**Chapter 7: Review Questions**

1. What are BI systems?

Business Information (BI) systems are information systems that assist managers

and organizations to help support decision-making by an analysis of current and past activities. (Page 496).

1. How do BI systems differ from transaction processing systems?

BI systems help with the analysis of past data and prediction of future events, but do not process routine transactions. (page 496)

1. Name and describe the two main categories of BI systems.

Reporting: sort, filter, group, and do basic calculations on operational data.

Data mining: perform complex analyses on data.

(page 497)

1. What are the three sources of data for BI systems?

Operational databases, extracts of operational databases, and purchased data (page 546).

1. Summarize the problems with operational databases that limit their usefulness for BI applications.

Performance: databases are good at frequent, small transactions, but BI apps are typically dealing with large or complex queries and calculations that slow down a typical database.

Capability: databases don’t typically have the ability to perform the complex calculations BI apps can handle.

Latency: BI apps tend to be real-time snapshots, but databases prioritize transaction integrity (which can sacrifice speed on a large scale).

(no source, I just read about this a lot.)

1. What is an ETL system, and what functions does it perform?

Extract, transform, and load (ETL) systems extract data from operational systems; transform the data and load them into data warehouses; and maintain metadata that describe the source, format, assumptions, and constraints of the data. (page 546)

1. What problems in operational data create the need to clean data before loading the data into a data warehouse?

Cleaning data helps the data warehouse correctly store data for future access and processing. Reasons for cleaning include consistency (NY vs New York), duplicate records, missing or incomplete data (for example, a city and state but no zip code), redundancy, inaccurate or incorrect data (for example, if a client relocates and the db is not updated), irrelevant data (stuff we just don’t need in the data warehouse), integration issues (for example, two different systems with two different customer IDs), and issues with data types (for example, a date stored as a string rather than a datetime or numeric data type). (page 507)

1. What does it mean to transform data? Give an example other than the ones used in this book.

Transforming data converts it into a consistent and usable format for the data warehouse. Examples include consistent name formatting (Title Case or UPPERCASE), using consistent units of measurement (don’t mix and match pounds and kilograms), and cleaning data in ways that handle missing or erroneous data. (page 546)

1. Define data mining.

Data mining is the use of statistical and mathematical techniques to find patterns in database data. (glossary)

1. Explain the difference between unsupervised and supervised data mining.

Supervised data mining creates a predictive model based on past experiences and guided by experts, while unsupervised data mining uses the data alone to create a model with no preconceived assumptions or categories. (Glossary)

*7.11 Name five popular data mining techniques.*

 Classification, clustering, market basket, regression, and decision trees. (pages 495-500)

*7.12 Explain how a decision tree classifies a new record.*

Decision trees navigate a series of simple hierarchical questions (decisions) based on an algorithm. (page 501-504)

* 1. Why are data warehouses necessary?

Data warehouses address the limitations of operational databases, by helping manage very large data sets, data integration, and complex analyses and reporting. (page 506)

7.14 Give examples of data warehouse metadata.

Metadata includes the data’s source, format, assumptions, constraints. (page 507)

* 1. Explain the difference between a data warehouse and a data mart. Give an example

other than the ones used in this book.

A data mart is a collection of data that is smaller than the data warehouse that addresses a specific component or functional area of the business. (page 508) An example might be the banking industry, using a data mart for marketing data, with customer segementing and marketing campaign effectiveness.

7.16 What is the enterprise data warehouse (EDW) architecture?

EDW is a central repository that combines the data warehouse with the data mart, where “the data warehouse maintains all enterprise BI data and acts as the authoritative source for data extracts provided to the data marts.” (page 509)

* 1. Describe the differences between operational databases and dimensional databases.

The differences are summarized in Figure 7-15 below:

|  |  |
| --- | --- |
| **Operational Database** | **Dimensional Database** |
| Used for structured transaction data processing | Used for unstructured analytical data processing |
| Current data are used | Current and historical data are used |
| Data are inserted, updated, and deleted by users | Data are loaded and updated systematically, not by users |

 (page 509)

* 1. What is a star schema?

A relational database that stores quantitative data. It gets its name from its starlike shape. (page 510)

7.19 What is a fact table? What type of data are stored in fact tables?

A fact table stores quantitative data (like sales numbers) which are called measures, with foreign keys linking to dimension tables; used for BI/analytics. (page 549)

* 1. In relation to fact tables, what is a measure?

See above. A measure is a unit of data held in a fact table.

* 1. What is a dimension table? What types of data are stored in dimension tables?

A dimension table holds the attributes used in queries for analysis; it connects to a fact table in a star schema dimensional db. (page 549)

* 1. What is a slowly changing dimension?

A column in a dimensional database with data whose values change occasionally and irregularly over time, such as addresses or last names. (page 549)

* 1. Why is the time dimension important in a dimensional model?

This is what allows the data to be analyzed over time. (page 549)

* 1. What is a conformed dimension?

A dimension table that has a relationship with two or more fact tables. (page 549)

* 1. What does OLAP stand for?

Online analytical processing; a data processing technique for analyzing measures and dimensions. (page 549)

**SCREEN CAPTURE of SQL Exercises**

****