Data Warehousing, Analysis, and Mining: 21 Common Questions

---

1. What is a data warehouse and why is it used?

A data warehouse is a centralized repository that stores structured data from various sources. Its primary use is for reporting and data analysis, offering a unified historical view of a company's data.

---

2. Can you explain the differences between OLAP and OLTP?

- OLAP (Online Analytical Processing) is optimized for complex queries and historical data analysis. It's designed for read-heavy operations, such as generating reports, visualizations, and trend analysis. It handles large data sets and typically uses a Star or Snowflake schema.
- OLTP (Online Transaction Processing) focuses on real-time transaction management, such as processing orders or recording customer payments. It is optimized for fast write-heavy operations. It deals with small, real-time transactions and usually employs a normalized schema.

---

3. What is a dimension table and a fact table?

These are the building blocks of a data warehouse schema.

- Dimension tables contain descriptive attributes (e.g., customer names, product categories) that provide context to the data, helping answer questions like "who, what, where, and when."
- Fact tables contain quantitative data (e.g., sales figures, transaction amounts), which are the focus of analysis. Fact tables often reference dimension tables for a deeper understanding of metrics.

---

4. What are the stages of ETL in data warehousing?

The ETL (Extract, Transform, Load) process is fundamental for transforming raw data into a structured, ready-to-analyze format, ensuring data accuracy and reliability.

- Extract: Data is gathered from multiple sources like relational databases, APIs, or flat files.

- **Transform: Data is cleaned, formatted, and reshaped to match the data warehouse schema. This step may involve removing duplicates, calculating new fields, or applying business rules.**
- **Load: The processed data is loaded into the data warehouse, becoming accessible for querying and analysis. A more modern approach, ELT, involves loading raw data as is, with the transformation happening within the data warehouse.**

---

5. Describe the star schema and snowflake schema. Which is better and why?

Schemas provide a framework for organizing data in a data warehouse.

- **Star Schema: Features a central fact table surrounded by denormalized dimension tables. It is simple, intuitive, and optimized for quick queries, making it suitable for most business intelligence use cases. It uses more storage space but offers faster query performance due to fewer joins.**
- **Snowflake Schema: A normalized version of the star schema, where dimension tables are split into additional tables to reduce redundancy. While it saves storage space, it can complicate queries and slow performance due to more joins. It is ideal for scenarios requiring minimal redundancy.**

The choice between them depends on the use case; star schemas are better for simplicity and faster queries, while snowflake schemas are ideal for minimizing redundancy.

---

6. How would you design a data warehouse for a large-scale organization?

Designing requires careful planning for scalability, performance, and specific business needs. Key steps include:

- **Requirement Gathering: Understanding business objectives, KPIs, and data sources.**
- **Data Modeling: Choosing an appropriate schema design (e.g., star, snowflake) based on reporting needs and data relationships.**
- **Technology Stack: Selecting tools and platforms (e.g., Snowflake, Redshift, BigQuery) that align with scalability and budget.**
- **ETL/ELT Processes: Designing pipelines to handle high data volumes while ensuring quality.**
- **Performance Optimization: Implementing indexing, partitioning, and caching strategies for fast query execution.**

---

7. How do you maintain data quality in a data warehouse?

Poor data quality can lead to incorrect analyses and decisions. Measures include:

- **Validating data during the ETL process to check for errors or inconsistencies.**
- **Implementing data profiling to understand data patterns and identify anomalies.**
- **Setting up automated monitoring and alerts for data discrepancies.**
- **Regularly cleaning and deduplicating data to increase accuracy and consistency.**

---

8. What challenge might arise when mining data from heterogeneous sources for a supermarket located at an airport departure lounge?

Integrating inconsistent data formats, such as different currencies or time zones, can lead to misleading patterns unless the data is first cleaned and standardized.

---

9. How can a retail chain use data mining to optimize product placement?

By analyzing customer purchase patterns, a retail chain can use association rule mining to place frequently bought items together, thereby increasing cross-selling opportunities.

---

10. How does the model building phase in the analytics life cycle help a bank detect fraud?

It allows the bank to create predictive models using historical transaction data to flag unusual or suspicious transactions in real-time.

---

11. Why is the operationalized phase critical for healthcare analytics?

It involves deploying models into hospital systems to assist doctors with real-time patient monitoring, leading to faster diagnosis and improved care.

---

12. How do you handle schema changes in a data warehouse?

Schema changes are inevitable and must be handled efficiently to minimize disruptions and enhance data integrity. Strategies include:

- **Schema Versioning: Maintaining multiple schema versions and migrating data incrementally.**

- **Backward Compatibility: Ensuring new changes do not break existing queries by retaining legacy fields or creating views.**
- **Automation Tools: Using tools like DBT or Liquibase to automate migration and rollback processes.**
- **Impact Analysis: Identifying and updating dependencies (queries, reports, downstream systems) affected by changes.**
- **Testing: Validating schema changes in a staging environment before production deployment.**

---

13. How would you design a data warehouse for an e-commerce business?

For e-commerce, a design might include:

- **Data Sources: Integrating data from transactional databases, web analytics platforms, CRM systems, and inventory systems.**
- **Schema Design: Using a star schema with fact tables for sales transactions and dimensions for customers, products, and time.**
- **ETL Process: Developing pipelines to handle large data volumes, including incremental loading for transaction updates.**
- **Performance Optimization: Partitioning the sales fact table by date and using materialized views for common aggregations (e.g., daily revenue, top-selling products).**
- **Analytics and Reporting: Ensuring support for dashboards showing metrics like sales trends, customer retention, and inventory levels.**

---

14. How does Snowflake handle concurrency issues?

Snowflake's multi-cluster architecture supports high concurrency by automatically spinning up additional compute clusters during peak demand.

---

15. What is the significance of Exploratory Data Analysis (EDA)?

Exploratory Data Analysis (EDA) helps to better understand the data. It builds confidence in the data before engaging machine learning algorithms, allows refinement of feature variable selection for model building, and helps discover hidden trends and insights.

---

16. How can a manufacturer use a star schema to analyze production efficiency?

The central fact table can hold production metrics (e.g., units produced), while dimension tables categorize data by time, machine, location, etc.

---

17. What is the benefit of the three-tier data warehouse architecture for large enterprises?

The three-tier data warehouse architecture separates storage, processing, and presentation layers, enhancing the scalability, maintainability, and security of the system.

---

18. How can a hospital apply classification techniques in diagnostics?

Classification models can categorize patients into risk groups (e.g., high or low risk) based on symptoms and medical history.

---

19. What clustering application can help in urban planning?

Clustering household data helps segment neighborhoods based on income, consumption, or lifestyle for better resource allocation.

---

20. Why might a bank prefer clustering over classification for new customers?

Clustering helps discover unknown customer groups based on behaviors without predefined labels, making it ideal for exploratory analysis, whereas classification requires predefined labels.

---

21. A supermarket wants to analyse customer buying behaviour using an Apriori algorithm.

The dataset below represents transactions where each row shows the list of items bought together. Find frequent item sets using Apriori with a minimum support count of two. Generate association rules with minimum confidence of 40%.

- **The solution involves putting item-wise lists and making associations of two items and three items.**
- **Confidence is calculated as support(I1 ∩ I2 ∩ I3) / support(I1 ∩ I2)**