UNIT–2 : BI – Data Mining & Warehousing (PYQ Solution)

AKTU · Business Intelligence and Analytics (BCDS051) · Solutions by **Chirag Jain**, Ycotes

SECTION A Answers (≈70 Words Each)

Q1. What do OLTP and OLAP stand for? (Jan 2023)

OLTP stands for Online Transaction Processing: it handles day-to-day, short, frequent insert/update transactions (banking, ticketing) using current detailed data. OLAP stands for Online Analytical Processing: it supports complex queries, multidimensional analysis and reporting over large historical data in a data warehouse, for decision making.

Q2. What is a data repository? (Jan 2023)

A data repository is a centralised place where data is stored, organised and maintained for analysis and reporting. Examples are databases, data warehouses, data marts, data lakes and metadata repositories. It keeps data consistent, secure and easy to access for many users and applications.

SECTION B Answers (≈250 Words Each)

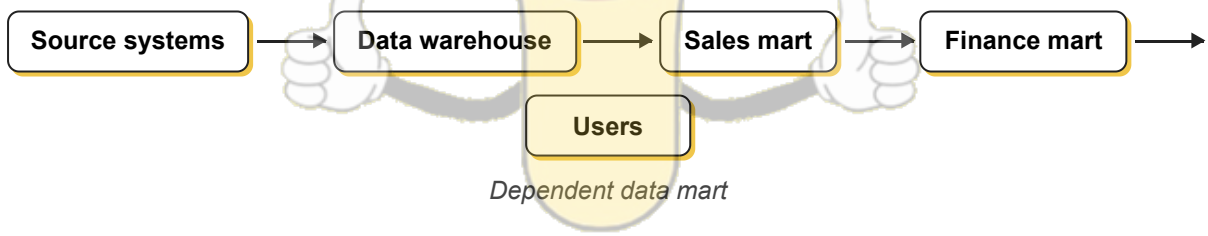
Q1. Explain the different types of data mart. Also define fact table and mapplet. (Jan 2023)

A **data mart** is a small, subject-oriented subset of a data warehouse serving one department (sales, finance, HR).

Types:

- **Dependent data mart:** built from an existing enterprise data warehouse (top-down). Data is consistent and integrated.
- **Independent data mart:** created directly from operational or external sources without a warehouse (bottom-up). Quick and cheap, but data may be inconsistent and siloed.
- **Hybrid data mart:** combines warehouse data with other operational sources.

Basis	Dependent	Independent
Source	Data warehouse	Operational systems
Consistency	High	Low
Cost and time	Higher	Lower



Dependent data mart

Fact table: the central table of a star or snowflake schema. It stores **measurable, numeric data** (sales amount, quantity) called measures, along with **foreign keys** to dimension tables (date, product, customer). It is usually very large and growing; the grain defines what one row means.

Mapplet: a **reusable set of transformations** (in ETL tools such as Informatica) built once and used in many mappings. It contains input and output transformations and hides the logic, which saves development effort and gives consistency.

SECTION C Answers (≈500 Words Each)

Q1. How do the Star and Snowflake schemas differ? Explain with a suitable example. (Jan 2023)

Both are dimensional models for a data warehouse, with a central **fact table** linked to **dimension tables**.

Star schema: dimensions are **denormalised**, each directly joined to the fact table, so the diagram looks like a star.



Star schema: fact table at centre

Snowflake schema: dimensions are **normalised** into sub-dimensions. Example: Product dim → Category dim and Brand dim; Store dim → City dim → Country dim.

Example: Sales fact (date_id, product_id, store_id, units, amount). In a star, Product dim holds product, category and brand together. In a snowflake, Product dim holds category_id and brand_id referring to separate tables.

Basis	Star	Snowflake
Dimensions	Denormalised	Normalised
Joins	Few, fast queries	More joins, slower
Storage	More redundancy	Less redundancy, less space
Design	Simple, easy	Complex
Maintenance	Harder to update	Easier to update
Best for	Fast reporting, BI tools	Large, complex hierarchies

Conclusion: star is preferred for query speed and simplicity; snowflake for storage saving and data integrity.

Q2. What are the steps of the ETL process? How is ETL useful in data

warehousing? (Jan 2023)

ETL (Extract, Transform, Load) moves data from sources into the warehouse.



ETL process

Steps:

1. **Extract:** read data from databases, files, ERP, CRM and APIs into a **staging area**; may be full or incremental extraction.
2. **Clean and validate:** remove duplicates, fix errors, handle missing values.
3. **Transform:** convert formats and data types, apply business rules, aggregate, join, derive new columns, standardise codes and units.
4. **Load:** write the transformed data into warehouse tables (initial full load, then incremental loads); build indexes.
5. **Monitor and log:** audit trails, error handling and scheduling.

Usefulness in data warehousing:

- Integrates data from **heterogeneous sources** into one consistent view.
- Improves **data quality** and trust.
- Keeps historical data and supports time-based analysis.
- Reduces load on operational systems (processing is done outside).
- Automates refresh, making reporting faster.

Conclusion: ETL is the backbone that fills the warehouse with clean, integrated, analysis-ready data.

Q3. What are different data mining techniques and how do businesses use them? (Jan 2023)

Data mining discovers patterns and knowledge from large data sets.

Technique	Idea	Business use
Classification	Assign records to known classes (decision tree, Naive Bayes)	Credit scoring, spam and churn detection
Clustering	Group similar records without labels (k-means)	Customer segmentation
Association rules	Find items that occur together (Apriori)	Market basket analysis, cross-selling
Regression / prediction	Estimate a numeric value	Sales and demand forecasting
Anomaly (outlier) detection	Find unusual records	Fraud detection
Sequential patterns / time series	Patterns over time	Stock trends, buying sequences
Summarisation / visualisation	Compact description of data	Dashboards, reports
Neural networks	Learn complex patterns	Recommendation, image and speech tasks

Business benefits: targeted marketing, better retention, risk and fraud reduction, inventory optimisation, personalised recommendations and improved decisions.

