Data Warehousing and Data Mining - UNIT III

1. Data Marts

1.1 Definition

A **Data Mart** is a focused subset of a data warehouse designed for a specific business line or department, such as sales, finance, or marketing. It contains subject-specific data that helps users retrieve relevant information quickly without searching the entire data warehouse.

> **Example:** A Sales Data Mart will include data relevant to sales transactions, customer purchases, and regional revenue breakdowns.

1.2 Characteristics / Features

- Subject-specific (e.g., Marketing, HR, Finance)
- Faster access due to smaller size
- Easier to maintain and manage
- Improves decision-making for specific departments
- Can serve as building blocks for larger Data Warehouses

2. Types of Data Marts

2.1 Dependent Data Mart

- Created from a centralized data warehouse
- Data is first integrated into the warehouse, then fed to the data mart
- Ensures consistency and central control
- Common in enterprise-scale data architecture

2.2 Independent Data Mart

- Standalone system created without a central data warehouse

- Pulls data directly from operational or external sources
- Easier and quicker to implement, but may lead to inconsistency
- Suitable for small departments or early-stage businesses

2.3 Hybrid Data Mart

- Combines features of both dependent and independent data marts
- May extract data from both the central data warehouse and external sources
- Offers flexibility while maintaining some level of consistency

---

3. Metadata

3.1 Definition

Metadata is data that describes other data. It provides details such as data origin, structure, transformation rules, and business meaning. It plays a critical role in managing, maintaining, and utilizing a data warehouse.

3.2 Types

- **Technical Metadata:** Database structures, ETL mappings, schemas
- **Business Metadata:** Definitions, business rules, KPIs
- **Operational Metadata:** Data refresh times, audit logs, lineage

3.3 Importance

- Enables data understanding and usability
- Helps track data lineage
- Crucial for documentation, auditing, and compliance
- Facilitates ETL operations and BI tools

---

4. Data Transformation

4.1 Definition

Data Transformation is the process of converting extracted data from source systems into a format suitable for analysis and reporting in the data warehouse.

4.2 Steps in Data Transformation

- **Selection:** Choosing relevant data fields
- **Cleansing:** Removing inconsistencies and errors
- **Integration:** Merging data from multiple sources
- **Deduplication:** Removing duplicate records
- **Standardization:** Ensuring data is in a uniform format (e.g., date, currency)
- **Encoding:** Replacing values with standardized codes or keys
- **Aggregation:** Summarizing data for higher-level analysis

4.3 Importance

- Ensures data accuracy and integrity
- Prepares data for efficient querying
- Makes data analysis meaningful and reliable

---

5. Hardware Architecture

5.1 Process Hardware

- Supports data movement (ETL), transformation, and job scheduling
- Includes ETL engines and workflow managers
- Critical for automating and orchestrating data pipelines

5.2 Server Hardware

- Central component that hosts the data warehouse
- Requires high processing power, memory, and scalable storage

- Must support high availability and redundancy

5.3 Network Hardware

- Facilitates fast data transfer between data sources, servers, and clients
- Includes switches, routers, and load balancers
- Network bottlenecks can severely degrade performance

5.4 Client Hardware

- End-user devices used for accessing and analyzing warehouse data
- May include web-based tools, BI dashboards, or thick client applications

---

6. Database Concepts in Data Warehouse

6.1 Subject-Oriented

- Organized by business subjects (e.g., customer, product)
- Simplifies data analysis and reporting

6.2 Integrated

- Combines data from heterogeneous sources into a unified view
- Requires data cleansing and transformation

6.3 Time-Variant

- Data is stored with timestamps
- Enables historical analysis and trend reporting

6.4 Non-Volatile

- Once data enters the warehouse, it is not updated or deleted
- Supports stable and consistent querying over time

---

7. Database Structures and Layout

7.1 Star Schema

- Central Fact Table connected to Dimension Tables
- Simplified and optimized for read access

7.2 Snowflake Schema

- Normalized dimension tables (sub-divided)
- Reduces redundancy but increases complexity

7.3 Fact Constellation / Galaxy Schema

- Multiple Fact Tables sharing Dimension Tables
- Suitable for complex, multi-subject analysis

7.4 Additional Design Features

- **Partitioning:** Improves query performance by breaking large tables
- **Indexing:** Helps in fast data retrieval
- **Materialized Views:** Precomputed summaries for faster access
- **OLAP Cubes:** Multidimensional data structures for analytical queries

---

8. File Systems in Data Warehousing

8.1 Definition

A file system manages how data is stored, accessed, and retrieved on physical storage. In data warehousing, efficient file systems ensure reliable data storage and access.

8.2 Common File Systems

- **NTFS:** Secure, used in Windows servers
- **EXT4:** Linux-based, commonly used in open-source environments

- **HDFS:** Distributed system for big data storage (Hadoop)

8.3 Importance

- Determines performance, fault tolerance, and scalability
- Essential for supporting large-scale analytics
- Directly impacts backup, recovery, and replication capabilities