Data Warehousing

Course Learning Objectives	● To know the details of data warehouse Architecture ● To understand the OLAP Technology ● To understand the partitioning strategy ● To differentiate various schema ● To understand the roles of process manager & system manager
Course Outcomes	● Design a data warehouse architecture that is suited for different problems ● Create a data warehouse that utilizes OLAP technology for various purposes. ● Examine and apply the partitioning strategy in Data Warehouse. ● Differentiation of different schemas for a specific problem. ● Create the roles for both the process manager and system manager.

Course Duration: 45 Hours

Unit I - INTRODUCTION TO DATA WAREHOUSE

What is Data warehouse? Data warehouse Architecture Data warehouse Introduction - Data warehouse components- operational database Vs data warehouse – Data warehouse Architecture – Three-tier Data Warehouse Architecture - Autonomous Data Warehouse- Autonomous Data Warehouse Vs Snowflake - Modern Data Warehouse

Lab Component

Data exploration and integration with WEKA

Outcome

Weka Tools and Database Installation using Software.

Unit II - ETL AND OLAP TECHNOLOGY

What is ETL – ETL Vs ELT – Types of Data warehouses - Data warehouse Design and Modeling - Delivery Process - Online Analytical Processing (OLAP) - Characteristics of OLAP - Online Transaction Processing (OLTP) Vs OLAP - OLAP operations- Types of OLAP- ROLAP Vs MOLAP Vs HOLAP.

Lab Component

- Apply Weka tool for data validation.
- Plan the architecture for real-time application.

Outcome

- Know the ETL Process and Online Transaction Processing (OLTP).

Unit III - META DATA, DATA MART AND PARTITION STRATEGY

Meta Data – Categories of Metadata – Role of Metadata – Metadata Repository – Challenges for Meta Management - Data Mart – Need of Data Mart- Cost Effective Data Mart- Designing Data Marts- Cost of Data Marts- Partitioning Strategy – Vertical partition – Normalization – Row Splitting – Horizontal Partition

Lab Component

- Write the query for schema definition.
- Design data warehouse for real-time applications.

Outcome

- Query For Schema Definition.
- Design data warehouse for real-time.

Unit IV - DIMENSIONAL MODELING AND SCHEMA

Dimensional Modeling- Multi-Dimensional Data Modeling – Data Cube- Star Schema- Snowflake schema- Star Vs Snowflake schema- Fact constellation Schema- Schema Definition - Process Architecture- Types of Data Base Parallelism – Datawarehouse Tools

Lab Component

● Analyse the dimensional Modeling.

● Case study using OLAP.

● Case study using OTLP.

Outcome

● Get More Ideas about Data Modeling.

● Implement the Concept OLAP and OLTP.

Unit V - SYSTEM & PROCESS MANAGERS

Data Warehousing System Managers: System Configuration Manager- System Scheduling Manager - System Event Manager - System Database Manager - System Backup Recovery Manager - Data Warehousing Process Managers: Load Manager – Warehouse Manager- Query Manager – Tuning – Testing

Lab Component

● Implementation of warehouse testing.

Outcome

● Getting Data Warehousing System Handling Knowledge and Testing.

Test Projects:

Use Cases:

1. **BUSINESS INTELLIGENCE DASHBOARD**

   Task 1: Open Weka's Explorer interface.

   Task 2: Data Collection and Preparation

   Task 3: Design and implement a data warehouse to support a business intelligence dashboard.

   Task 4: Integrate data from sales, marketing, and finance.

   Task 5: To provide a unified view for decision-makers.

2. DATA COLLECTION AND PREPARATION

Task 1: Gather structured raw data (CSV, databases) related to Auditing and compliance

Task 2: Navigate to the "Preprocess" tab.

Task 3: Load dataset using "Open file" (CSV, ARFF).

Task 4: Examine attributes, statistics, distributions.

Task 5: Use "Attribute Statistics" or "Visualize" in Weka.

3: Data Warehouse Implementation

Task 1: Healthcare Data Warehouse – create or get the structured data model as per requirements in csv.

Task 2: implement a data warehouse for a retail company focusing on inventory management

Task 3: use the Partition and Normalization strategy to increase the data retrieving speed by query.

Task 4: Initiate the ETL process

Task 5: Examine attributes, statistics, distributions.

4. Data Modeling and Dimensional Design

Task 1: Gather structured raw data (CSV, databases) related to Auditing and compliance for Sales and Product Dimension.

Task 2: create Sales and Product Dimension by using dimensional Modeling

Task 3: Load the data by ETL

Task 4: Data Transformation and Standardization - Normalize or standardize data for consistency and comparability using Weka's filters.

Task 5: Data Exploration and Visualization, Use Weka's Explorer tools to visualize data attributes, distributions, and trends.

5.   Data Warehouses for Compliance

Task 1: Data Preprocessing - Handle missing values using tools like "Replace Missing Values".

Task 2: Normalize or standardize data via "Normalize" or "Standardize".

Task 3: Feature Selection and Transformation

Task 4: Perform feature selection with "Attribute Selection".

Task 5: Choose relevant attributes.

6.   Data Warehouses to Address Data Quality Issues

Task 1: Data Assessment- Load data into Weka, visualize attributes, detect issues like missing values, outliers, and inconsistencies.

Task 2: Handling Missing Values - Use Weka's "Replace Missing Values" filter to manage missing data through imputation or deletion.

Task 3: Outlier Detection and Treatment - Identify outliers using Weka's tools and decide on their treatment, such as removal or transformation.

Task 4: Data Transformation and Standardization - Normalize or standardize data for consistency and comparability using Weka's filters.

Task 5: Error Correction and Validation - Correct inconsistencies and validate data against rules for accuracy using Weka's functionalities.

7.   Data Warehouses for Data-Driven Decision-Making

Task 1: Data Collection and Integration - Gather and integrate relevant data sources into Weka-compatible formats (e.g., CSV, ARFF).

Task 2: Data Exploration and Visualization, Use Weka's Explorer tools to visualize data attributes, distributions, and trends.

Task 3: Preprocessing and Cleaning - Preprocess data using Weka to handle missing values, outliers, and inconsistencies for accurate analysis.

Task 4: Model Selection and Building - Choose appropriate data mining algorithms from Weka's toolkit (e.g., decision trees, neural networks) based on decision-making objectives.

Task 5: Model Evaluation and Validation - Evaluate model performance using Weka's evaluation metrics to ensure reliability and accuracy, cross-validation or testing.

8. **Retail Sales Analytics**

Task 1: Requirements Gathering - Identify key business questions related to sales performance and Determine data sources, including POS systems, inventory databases, and customer records.

Task 2: Data Modeling - Design a dimensional data model with dimensions such as time, product, and customer.

Task 3: ETL Processes - Implement ETL processes to extract sales data from different sources. Load the transformed data into a central data warehouse.

Task 4: Data Warehouse Implementation - Set up a data warehouse using a suitable database management system (DBMS). Establish relationships between tables

Task 5: Querying and Reporting, Write SQL queries to retrieve insights into sales performance.

9. **Healthcare Data Warehouse**

Task 1: Implement a data warehouse for a healthcare organization.

Task 2: Initiate the ETL process to Integrate patient records, treatment data, and diagnostic information for analytical purposes.

Task 3: Model Building - Select suitable algorithms (Classify/Cluster). Task

4: Evaluate performance using metrics (accuracy, precision, etc.).

Task 5: Visualization and Interpretation- Visualize results using Weka's tools.

10. Develop Model Data Warehouse Application

Task 1: Gather structured raw data (CSV, databases) related to Auditing and compliance

Task 2: ETL Processes - Implement ETL processes to extract sales data from different sources. Load the transformed data into a central data warehouse.

Task 3: Data Warehouse Implementation - Set up a data warehouse using a suitable database management system (DBMS). Establish relationships between tables.

Task 4: Model Evaluation for Evaluate performance using metrics (accuracy, precision, etc.). Employ cross-validation.

Task 5: Visualization and Interpretation to Visualize results using Weka's tools for Decision Making for Business. Training Hours - 15 Hrs (Theory) + 30 Hrs (Practical) = 45 Hrs.

Student Assessment Plan:

Each of the above-mentioned test projects will be divided into tasks by the training partner for each specific institution. Such tasks will be jointly evaluated by the faculty and the training partner and the following weightage is to be followed.

- 70% weightage to the external practical assessment.
- 30% weightage to the internal assessment.

Final Test Project/External Assessment Plan:

The Final Test Project will be chosen from the list given above, jointly by the college faculty and the Training Partner. The Final Test Project will be assessed on the following tasks, for 70%

Details	Marks
Task: 1	20
Task: 2	20
Task: 3	20
Task: 4	20
Task: 5	20

Employment Potential:

This course shall enable Computer Science Engineers to get employment in various engineering sectors.