CS-103(New): Distributed Database Concepts

Pre-requisites: Students should be well-versed with the basic and advanced concepts of RDBMS

Objectives:

Main objective is to understand the principles and foundations of distributed databases. This course addresses architecture, design issues, integrity control, query processing and optimization, transactions, and concurrency control & distributed transaction reliability.

Unit 1. Distributed databases: An overview **[2]**

1.1 Features of distributed Vs centralized databases Chapter 1 from Book 2
1.2 Why DDB? DDBMS
1.3 Promises / problem areas in implementing a DDB Section 1.3,1.5 from Book 1

Unit 2. DDBMS Architecture **[4]**

2.1 DBMS Standardization Chapter 4 from Book 1
2.2 Architectural models for DDBMS
2.3 DDBMS architecture
2.4 Distributed catalog management Section 21.8 from Book 3

Unit 3. Distributed database design **[10]**

3.1 Alternative design strategies Chapter 5 from book 1
3.2 Distributed design issues
3.3 Concepts of join graphs Section 4.2.1.2 from book 2
3.4 Fragmentation and allocation Chapter 5 from Book1

Unit 4. Overview of Query processing **[4]**

4.1 Query processing problems
4.2 Objectives of query processing Chapter 7 from book 1
4.3 Complexity of relational algebra operators
4.4 Characterization of query processors
4.5 Layers of query processing

Unit 5. Query decomposition & data localization **[2]**

5.1 Query decomposition
Chapter 5.2 Localization of distributed data 8 from book 1

Unit 6. Optimization of distributed queries **[10]**

6.1 Query optimization
Centralized query optimization Join ordering in fragment queries. Distributed query optimization algorithms Chapter 9 from book1
6.2 Centralized query optimization
6.3 Join ordering in fragment queries
6.4 Distributed query optimization algorithms

Unit 7. Management of distributed transactions **[2]**

7.1 Framework for transaction management Chapter 7 from book 2
7.2 Supporting atomicity of distributed transactions
7.3 Concurrency control of distributed transactions
7.4 Architectural aspects of distributed transactions

Unit 8. Concurrency control **[6]**

8.1 Foundations of distributed concurrency control Chapter 8 from book 2
8.2 Distributed deadlocks
8.3 Concurrency control based on timestamps
8.4 Optimistic methods for distributed concurrency control

Unit 9. Distributed DBMS reliability **[8]**

9.1 Reliability concepts & measures
9.2 Failures & fault tolerance in distributed systems from book 1
9.3 Failures in DDBMS
9.4 Local reliability protocols
9.5 Distributed reliability protocols
9.6 Dealing with site failures
9.7 Network partitioning

Reference Books:

1. Principles of Distributed Database Systems; 2nd Edition By M. Tamer Ozsu and Patrick Valduriez Publishers: Pearson Education Asia ISBN: 81-7808-375-2
2. Distributed Database; Principles & Systems By Stefano Ceri and Giuseppo Pelagatti Publications: McGraw-Hill International Editions ISBN: 0-07-010829-3
3. Database systems (2nd edition) By Raghuramakrishnan and Johannes

CS-203(New): Data Mining and Data Warehousing

Unit 1. **Introduction to Data Mining** **[4]**

- Basic Data Mining Tasks
- DM versus Knowledge Discovery in Databases
- Data Mining Issues
- Data Mining Metrics
- Social Implications of Data Mining
- Overview of Applications of Data Mining

Unit 2. **Introduction to Data Warehousing** **[4]**

- Architecture of DW
- OLAP and Data Cubes
- Dimensional Data Modeling-star, snowflake schemas
- Data Preprocessing – Need, Data Cleaning, Data Integration & Transformation, Data Reduction
- Machine Learning
- Pattern Matching

Unit 3. **Data Mining Techniques** **[4]**

- Frequent item-sets and Association rule mining: Apriori algorithm, Use of sampling for frequent item-set, FP tree algorithm
- Graph Mining: Frequent sub-graph mining, Tree mining, Sequence Mining

Unit 4. **Classification & Prediction** **[16]**

- Decision tree learning: [3 hrs]
  Construction, performance, attribute selection
  Issues: Over-fitting, tree pruning methods, missing values, continuous classes
  Classification and Regression Trees (CART)
- Bayesian Classification: [6 hrs]
- Bayes Theorem, Naïve Bayes classifier,
- Bayesian Networks
- Inference
- Parameter and structure learning
- Linear classifiers [4 hrs]
- Least squares, logistic, perceptron and SVM classifiers
- Prediction [3 hrs]
- Linear regression
- Non-linear regression
-

Unit 5 **Accuracy Measures** **[4]**

Precision, recall, F-measure, confusion matrix, cross-validation, bootstrap

Unit 6. **Software for data mining and applications of data mining** **[4]**
R, Weka, Sample applications of data mining

Unit 7. **Clustering** **[4]**

- k-means
- Expectation Maximization (EM) algorithm
- Hierarchical clustering, Correlation clustering

Unit 8. **Brief overview of advanced techniques** **[**4]

- Active learning
- Reinforcement learning
- Text mining
- Graphical models
- Web Mining

Reference Books:

1. Data Mining: Concepts and Techniques, Han, Elsevier ISBN:9789380931913/ 9788131205358
2. Margaret H. Dunham, S. Sridhar, Data Mining – Introductory and Advanced Topics, Pearson Education
3. Tom Mitchell, ―Machine Learning‖, McGraw-Hill, 1997
4. R.O. Duda, P.E. Hart, D.G. Stork. Pattern Classification. Second edition. John Wiley and Sons, 2000.
5. Christopher M. Bishop, ―Pattern Recognition and Machine Learning‖, Springer 2006
6. Raghu Ramkrishnan, Johannes Gehrke, Database Management Sysstems, Second Edition, McGraw Hill International
7. Ian H.Witten, Eibe Frank Data Mining: Practical Machine Learning Tools and Techniques, Elsevier/(Morgan Kauffman), ISBN:9789380501864
8. [Research-Papers]: Some of the relevant research papers that contain recent results and developments in data mining field