Lectures for the course: Data Warehousing and Data Mining (406035)

Week 1

Lecture 1

- Discussions on the need for data warehousing
- How DW is different from OLTP databases

Week 2

Lecture 2

- Evaluation norms were announced
- Class Test date was announced
- Expectations from Term paper and Term Project were announced
- What is a data warehouse?

Lecture 3

- Why a data warehouse is required?
- Components of a data warehouse – Source System, Data Staging Area, Presentation server and User Interface.

Lecture 4 (a) and (b)

- 3 tier Data warehouse architecture
- OLAP – ROLAP, MOLAP and HOLAP
- Multidimensional Data Model (MDM)
- 2-D and 3-D representations of Data in MDM
- Slicing, Dicing, Roll-up and Drill down OLAP operations
- Hierarchy in a dimension

Week 3

Lecture 5

- OLAP operations re-visited
- Concept Hierarchy – Schema Hierarchy and Set-grouping Hierarchy defined
- Total order and partial order among concept hierarchy levels explained
- Roll-up and Drill down using concept hierarchy and dimension reduction

Lecture 6

- Data Cube defined as a lattice of cuboids
- DMQL statement for cube definition
- Types of measure explained – Distributive, Algebraic and Holistic

Lecture 7 (a) and (b)

- ERD and normalization revisited
- Dimensional Modeling
- Fact tables and dimension tables
- De-normalization and its effect on data warehouse design
- Report generation from Dimensional Model
- Star schema and Snowflake Schema - Definitions
- How star schema provides symmetric entry into the fact table from the dimension tables

Week 4

Lecture 8

- Fact table and Dimension table revisited
- Size estimate of Fact and Dimension tables
- Four main steps in Data warehouse design – Identify business process, Define grain, Identify dimensions and Identify facts
- Data marts
- Flexibility of dimensional models – How dimensional model can handle new measures and new dimensions in the Fact tables. How old records are updated with default values for new columns
- New attributes in the dimension table
- Details of date dimension
- Gray et al paper circulated

Lecture 9

- Snowflake and Fact constellation schemas
- Factless fact table
- Degenerate dimensions
- Sparse fact tables
- Multiple fact tables with new dimensions
- Addition of new dimensions in existing fact tables e.g., freq_cust_id in sales fact tables
- Retail Sales fact table and Promotion Fact tables – How they operate together

Week 5

Lecture 10

- Recap of old concepts – Star schema, Snowflake schema, Fact constellation, Factless fact table
- De-normalized fact table
- Introduction to Inventory Business process

Lecture 11

- Inventory Periodic Snapshot Fact table schema
- Additive, Semi-additive and Non-additive facts
- Size estimate of periodic snapshot schema
- Limitations of periodic snapshot schema
- Introduction to inventory transactions

Lecture 12 (a) and (b)

- Inventory Transactions
- Inventory Accumulating Snapshot

Week 6

Lecture 13

- Data Warehouse Bus Architecture
- Data Warehouse Bus Matrix
- Conformed Dimensions – Identical Table, Sub-set Dimension and Roll-up Dimension

Lecture 14

- Slowly Changing Dimensions – Type 1, Type 2, Type 3 and Type 6 response
- Customer Relationship Management Data Mart
- Aggregated and Segmentation Attributes of Customer Dimension
- Rapidly Changing Large Dimensions
- Outrigger Dimension
- Minidimension
- Effect of Minidimension on Dimension-only queries

First Class Test was held here

Week 7

Lecture 15

- Recap of Changing Dimensions
- Slowly changing and Rapidly Changing Dimensions

Lecture 16

- E-Commerce Data Warehouse
- Basics of Browser-Web Server Interaction
- Clickstream Analysis and Web Log as Source of Information
- Challenges in using web log data – Identification of Visitor Source, Visitor, Session, Proxy Servers and Browser Caches
- Unique Dimensions in E-Commerce Data Warehouses – Session, Page, Event and Referral
- Fact Tables – Complete Session Facts and Page Event Facts
- E-Commerce Data Warehouse Sizing

Lecture 17 (a) and (b)

- View Materialization – Full Materialization, No Materialization and Partial Materialization
- Which Views to Materialize?
- How to Use Materialized Views for Optimizing Queries?
- How to Efficiently Update and Refresh Materialized Views?
- Efficient Implementation of Materialized View Calculation in MOLAP.
- Data Cube Chunks and Order of Visiting Data Chunks for Efficient Calculation of Aggregation. Reduction in Memory Requirement for Partial Sum Storage.

Week 8

Lecture 18

- Materialized view selection – Paper by Harinarayan, Rajaram and Ullman – Greedy Algorithm

Lecture 19

- Continued with the greedy algorithm and further examples
- Indexing OLAP Data – Paper by Sarawagi
- Multidimensional Array indexing – sparse dimension issues
- Bitmap Indexing

Lecture 20 (a) and (b)

- Multidimensional Indexing – R Tree
- Join Indexing
- A brief introduction to Business Dimensional Life Cycle for Data Warehousing Projects

Week 9

Repeat of First Class Test was held here for students who missed due to campus interviews.

Mid-Sem Exam was held here.

Week 10

Lecture 21

- Introduction to Data Mining
- KDD and Data Mining
- SQL and Data Mining
- Mining Association Rules – Why they are Important?
- Itemsets, Frequent Itemsets, Infrequent Itemsets
- Downward and Upward Closure Properties

Lecture 22 (a) and (b)

- A priori Algorithm for Association Rule Mining
- Generation of Frequent Itemsets
- Extraction of Association rules using Confidence Measures
- How Association Rules may be used in Data Warehouses

Week 11

Lecture 23

- Partitioning Algorithm for Association Rule Mining

Lecture 24

- Dynamic Itemset Counting Algorithm for Association Rule Mining proposed by Brin et al.

Lecture 25 (a) and (b)

- FP-tree Algorithm for Association Rule Mining proposed by Han et al.

Week 12

Lecture 26

- Clustering Algorithms
- What is Clustering?
- Why we need Clustering?
- K-means Clustering

Lecture 27

- K-medoid Clustering – PAM algorithm

Week 13

Lecture 28

- Clustering Algorithm – CLARA and CLARANS

Lecture 29

- Hierarchical Clustering – Agglomerative and Divisive
- Intra-cluster and Inter-cluster distance measures
- Dendrogram and its representation
- Agglomerative hierarchical clustering using average inter-cluster distance

Lecture 30 (a) and (b)

- Clustering Algorithm – BIRCH
- Clustering Feature and Clustering Feature Tree

Week 14

Lecture 31

- Classification - What is Classification?

- Training, Testing and Using a Classifier
- Confusion Matrix
- Multilayer Perceptron as a Classifier

Lecture 32

- Back Propagation for MLP Training
- Other Learning methods for MLP

Lecture 33 (a) and (b)

- Use of Decision Tree for Classification
- ID3 Algorithm
- Information Gain
- Fuzzy Classification

Week 15

Lecture 34

- Overview of Text and Web Mining

Lecture 35

- Summary and Feedback

Lecture 36 (a) and (b)

- Term Project Demo