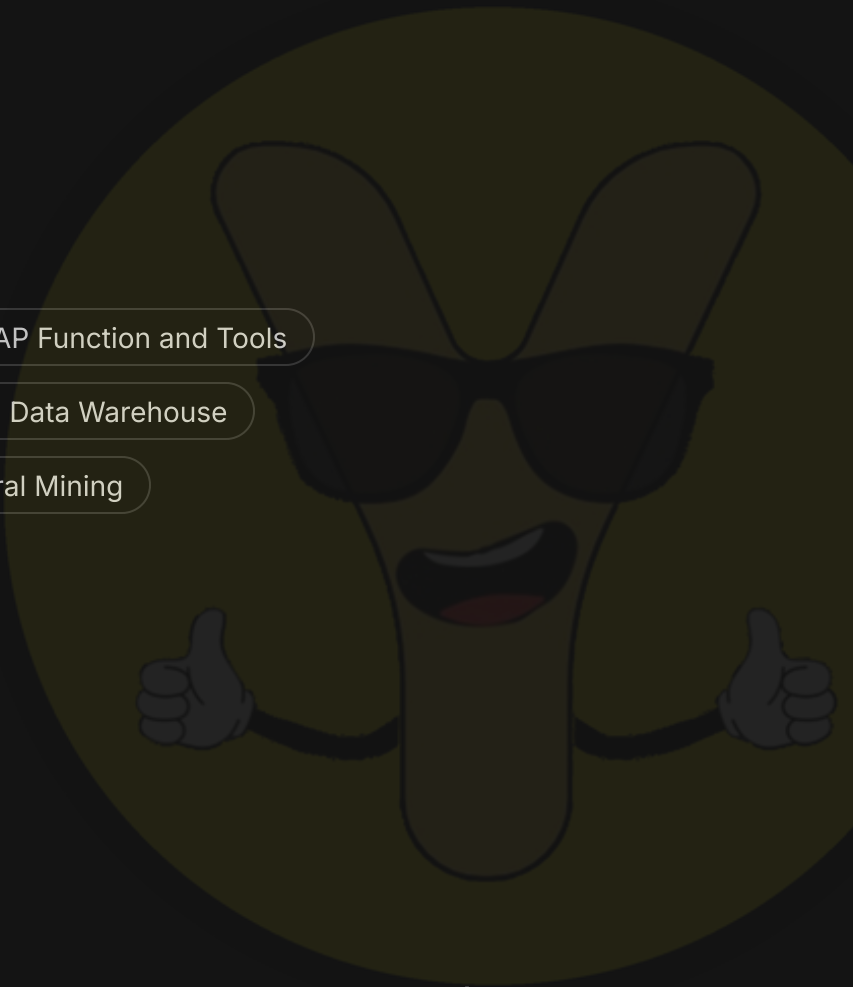
AKTU · 3RD YEAR · SEMESTER 5 · BCS058

Data Warehouse & Data Mining

Unit 5 — Data Visualization, OLAP and Recent Trends

- Data Visualization and Overall Perspective
- Aggregation
- Historical Information
- Query Facility
- OLAP Function and Tools
- OLAP Servers: ROLAP, MOLAP, HOLAP
- Data Mining Interface
- Security, Backup and Recovery
- Tuning Data Warehouse
- Testing Data Warehouse
- Types of Warehousing Applications
- Web Mining
- Spatial Mining
- Temporal Mining



DECODED BY

Chirag Jain

Founder, Ycotes · Gold Medalist in Artificial Intelligence, IIT Madras · taught 10,000+ students

theycotes.com

Data Visualization in Data Mining

DATA VISUALIZATION

The visual representation of complex multidimensional data structures and data mining results using graphical formats such as charts, maps, and 3D cubes to enable intuitive pattern discovery.

- Transforms abstract numerical data into visual patterns.
- Allows domain experts to evaluate data mining results visually.
- Facilitates interactive exploration of large-scale data warehouses.

Data Visualization Techniques

Visual data mining techniques convert high-dimensional data into intuitive graphical forms.

- Pixel-oriented: Maps each data value to a colored pixel to display massive datasets.
- Geometric projection: Project multi-dimensional space onto 2D/3D (e.g., Parallel Coordinates).
- Icon-based: Represents data dimensions using visual features like Chernoff faces or stick figures.
- Hierarchical: Divides data space into sub-spaces mapped to nested visuals (e.g., Tree-maps).
- Model visualization: Displays mining models directly such as Decision Trees and Association rules.

Aggregation in Data Warehouse

DATA AGGREGATION

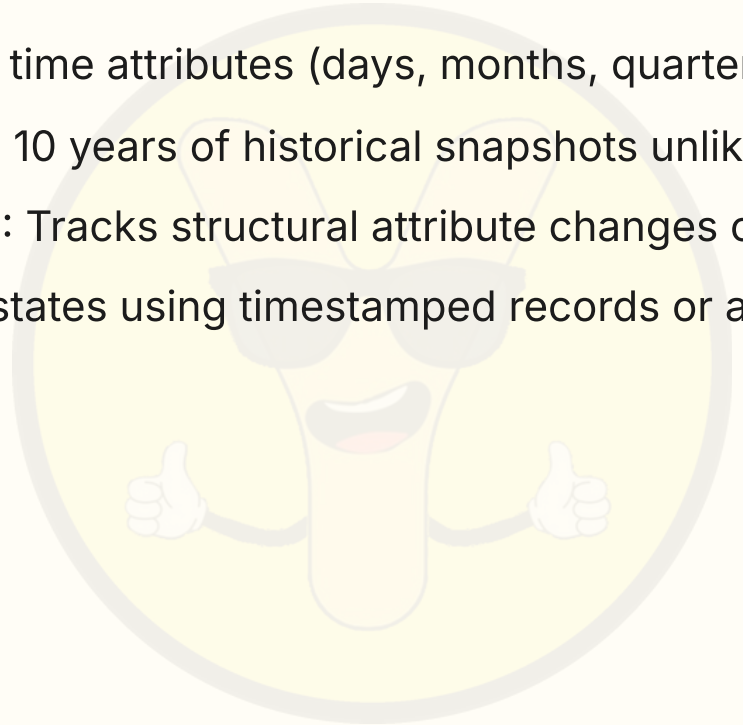
The process of consolidating and summarizing detail-level transactional data into higher-level granular metrics along dimension hierarchies using aggregate functions.

- Reduces query processing time by pre-computing common summaries.
- Utilizes SQL functions such as $SUM()$, $AVG()$, $COUNT()$, $MIN()$, and $MAX()$.
- Supports OLAP operations through hierarchical grouping levels.

Historical Information Management

Data warehouses maintain time-variant historical data over extended periods to support trend analysis.

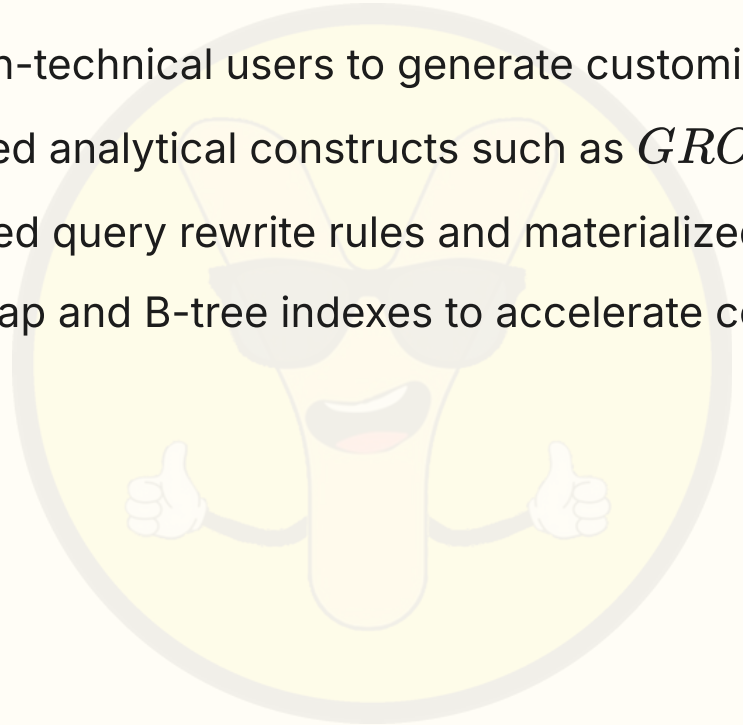
- Time-variant design: Stores explicit time attributes (days, months, quarters, years) in every fact table.
- Long retention periods: Retains 5 to 10 years of historical snapshots unlike operational databases.
- Slowly Changing Dimensions (SCD): Tracks structural attribute changes over time (Type 1, Type 2, Type 3).
- Delta storage: Preserves historical states using timestamped records or audit logs.



Query Facility in Data Warehouse

A robust query facility enables efficient ad-hoc analytical querying and data retrieval.

- Ad-hoc Query Support: Enables non-technical users to generate customized reports.
- SQL Extensions: Integrates advanced analytical constructs such as $GROUPBYCUBE$ and $ROLLUP$.
- Query Optimization: Uses cost-based query rewrite rules and materialized view selection.
- Specialized Indexing: Employs Bitmap and B-tree indexes to accelerate complex search predicates.



OLAP Functions and Multidimensional Analysis

Online Analytical Processing (OLAP) enables interactive analysis of multidimensional data.

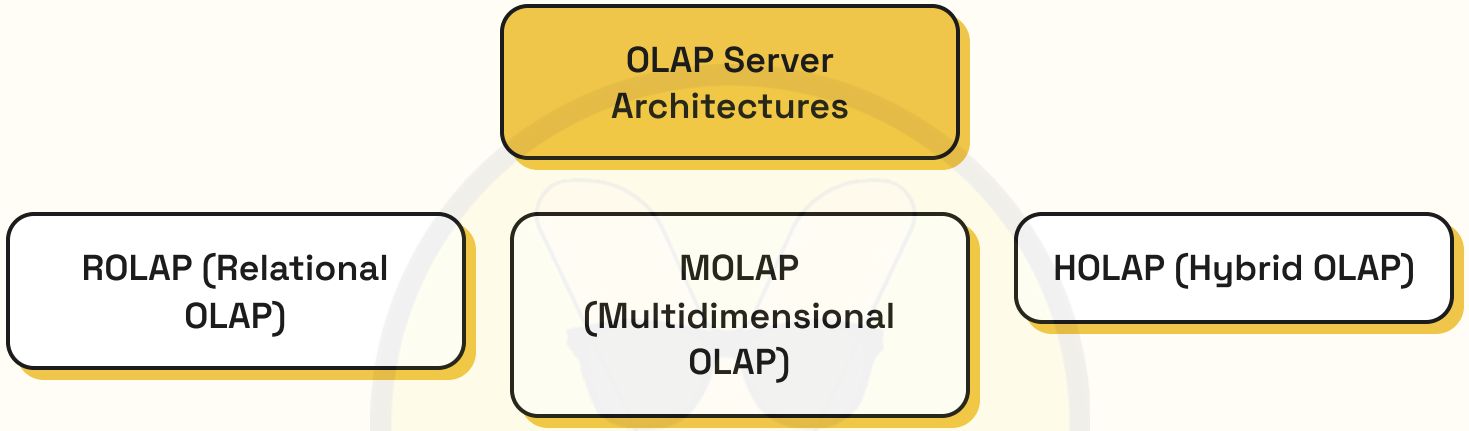
- Multidimensional Data Cube: Modeled as N-dimensional space with facts and dimensions.
- Roll-up: Performs aggregation by climbing up a dimension hierarchy or reducing dimensions.
- Drill-down: Navigates from summarized data to detailed data by stepping down hierarchies.
- Slice and Dice: Slice isolates a single dimension; Dice selects a sub-cube across multiple dimensions.
- Pivot (Rotate): Rotates data axes to provide alternative graphical perspectives of the cube.

Execution Flow of OLAP Cube Operations



Sequential flow of multidimensional OLAP operations on data cubes

Taxonomy of OLAP Server Architectures



Classification of OLAP server implementation architectures

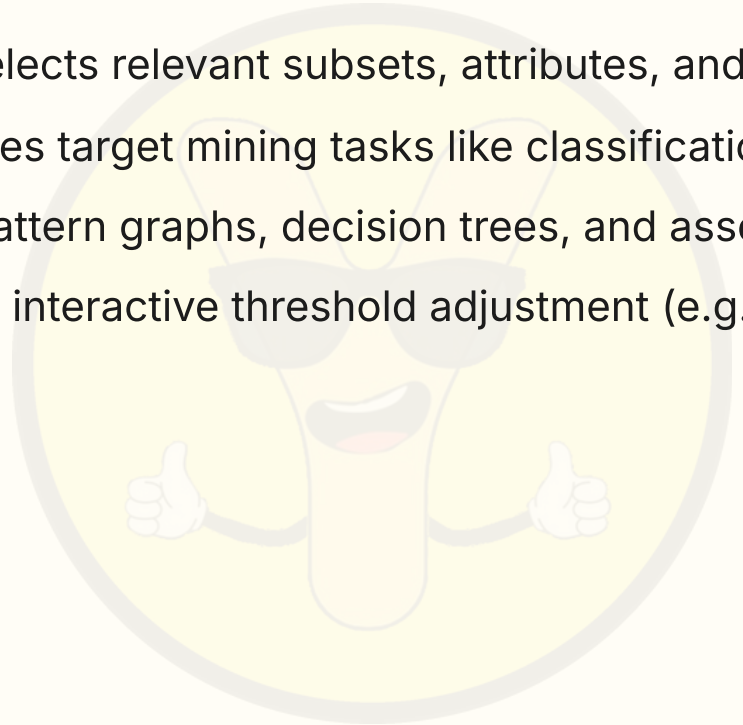
Comparison of ROLAP, MOLAP, and HOLAP

Feature	ROLAP	MOLAP	HOLAP
Storage Structure	Relational Database (Star/Snowflake)	Proprietary Array/Cube Structures	Relational Detail + Array Aggregates
Query Performance	Medium (SQL join dependent)	Fast (Direct array indexing)	Fast for Summary, Medium for Detail
Scalability	High (Scales to Terabytes)	Medium (Limited by cube size)	High (Combines ROLAP + MOLAP)
Build / Maintenance	Fast initial load, slower queries	Long cube rebuild time	Balanced maintenance overhead
Storage Overhead	Low (Relational tables + indexes)	High (Sparse array overhead)	Moderate storage requirement

Data Mining Interface Design

A Data Mining Interface bridges end users and data mining execution engines.

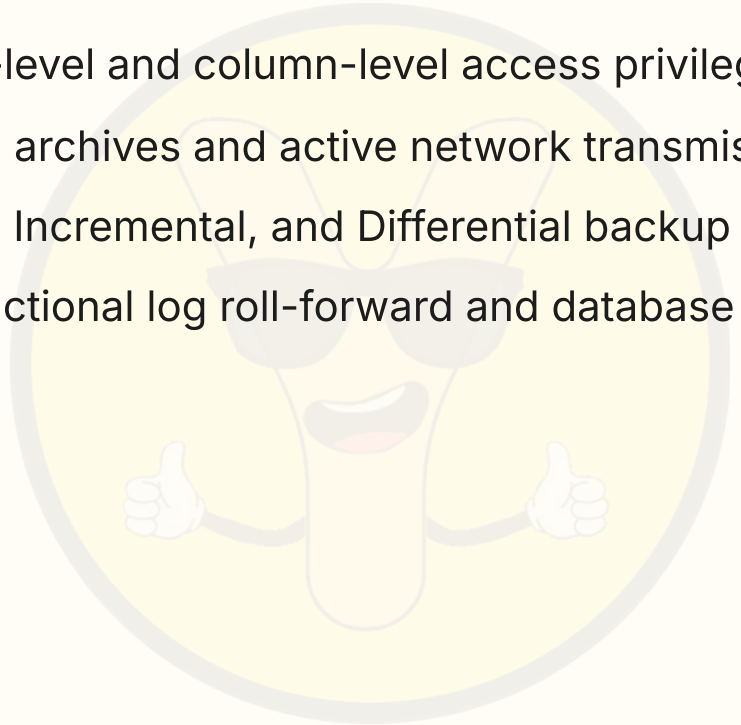
- Task-relevant data specification: Selects relevant subsets, attributes, and constraints.
- Mining primitive selection: Configures target mining tasks like classification or clustering.
- Interactive visualization: Displays pattern graphs, decision trees, and association rules.
- Graphical model refinement: Allows interactive threshold adjustment (e.g., minimum support).



Security, Backup, and Recovery in DW

Maintaining data integrity, confidentiality, and availability in enterprise data warehouses.

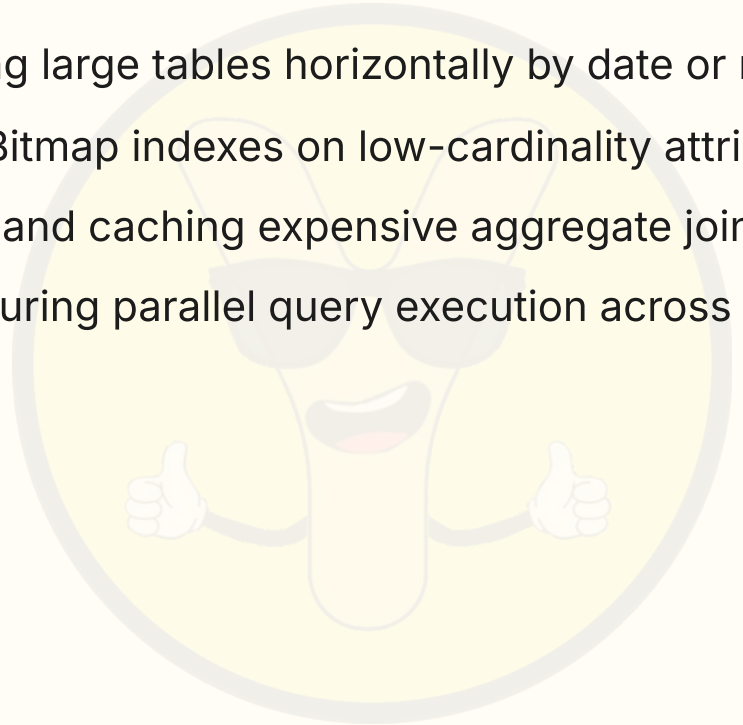
- Role-Based Security: Enforces row-level and column-level access privileges for user roles.
- Data Encryption: Encrypts historical archives and active network transmissions.
- Backup Strategies: Implements Full, Incremental, and Differential backup cycles for enterprise data.
- Recovery Mechanisms: Uses transactional log roll-forward and database mirroring for fast fault recovery.



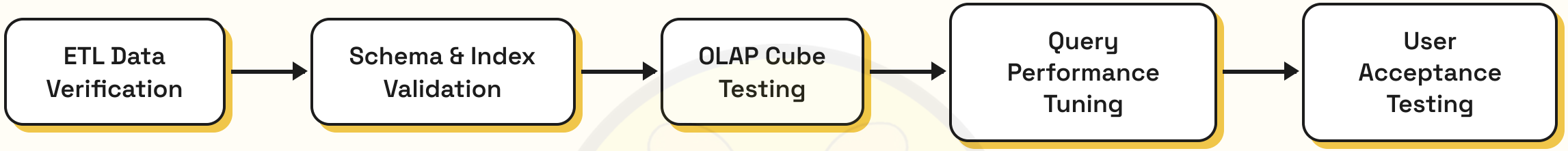
Data Warehouse Tuning

Data warehouse tuning optimizes hardware resources, query response times, and storage.

- Physical Schema Tuning: Partitioning large tables horizontally by date or region.
- Indexing Strategies: Implementing Bitmap indexes on low-cardinality attributes.
- Materialized Views: Pre-computing and caching expensive aggregate joins.
- Memory and I/O Parallelism: Configuring parallel query execution across CPU cores.



Data Warehouse Testing Lifecycle

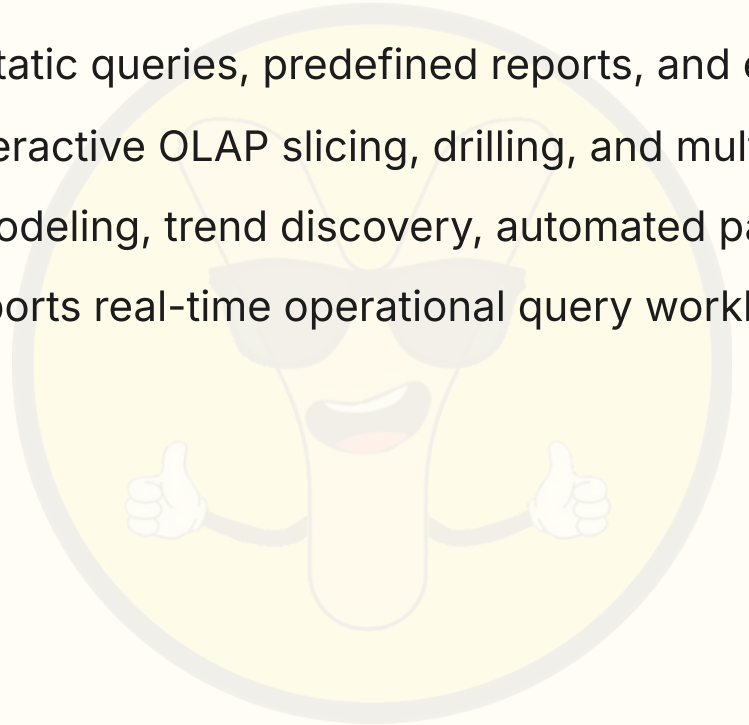


Structured lifecycle steps for data warehouse testing and validation

Types of Data Warehousing Applications

Data warehousing applications are categorized based on analytical workload and user goals.

- Information Processing: Supports static queries, predefined reports, and executive dashboards.
- Analytical Processing: Supports interactive OLAP slicing, drilling, and multidimensional analysis.
- Data Mining: Supports predictive modeling, trend discovery, automated pattern recognition.
- Operational Data Store (ODS): Supports real-time operational query workloads on fresh data.



Classification of Web Mining



Taxonomy of Web Mining domains and sub-fields

Web Mining: Content, Structure, and Usage

Web mining extracts implicit knowledge from Web documents, links, and user clickstreams.

- Web Content Mining: Extracts unstructured text, images, and HTML metadata from web pages.
- Web Structure Mining: Analyzes hyperlink graph structures using algorithms like PageRank and HITS.
- Web Usage Mining: Analyzes web server access logs, clickstream paths, and browser session histories.
- Applications: Target marketing, website personalization, search engine optimization, recommendation systems.

Spatial Data Mining

SPATIAL DATA MINING

The extraction of implicit knowledge, spatial relationships, or non-trivial topological patterns from spatial databases containing geometric and location data.

- Processes vector (points, lines, polygons) and raster (grid imagery) spatial data types.
- Uses spatial clustering algorithms like DBSCAN and OPTICS.
- Extracts spatial association rules (e.g., proximity to highways implies high property value).

Temporal Data Mining

TEMPORAL DATA MINING

The process of extracting implicit, non-trivial, and actionable temporal patterns, time-series sequences, and trends from datasets with explicit time stamps.

- Discovers sequential patterns (e.g., event A followed by event B within time interval Δt).
- Performs time-series similarity matching and trend analysis.
- Analyzes periodicity, temporal association rules, and seasonal variations.

Exam Corner: High-Yield Questions

- **Q1** Differentiate between ROLAP, MOLAP, and HOLAP server architectures with neat diagrams and structural comparison.
- **Q2** Explain the core OLAP operations (Roll-up, Drill-down, Slice, Dice, Pivot) using a 3D data cube example.
- **Q3** What is Web Mining? Compare Web Content Mining, Web Structure Mining, and Web Usage Mining in detail.
- **Q4** Define Spatial Data Mining and Temporal Data Mining. Explain their key features and real-world applications.
- **Q5** Discuss the essential techniques for Data Warehouse Tuning, Data Security, Backup, and Recovery mechanisms.
- **Q6** Explain the Data Warehouse testing lifecycle and how materialized views optimize the query facility.

Unit 5 done. You are exam-ready.

Revise the Exam corner once more, then open Unit 1 on theycotes.com. Every AKTU subject, unit by unit, free. Notes by Chirag Jain, Ycotes.

