

# Focal Points Data Warehouse Technologies

## Fundamentals

- Terms: Data Warehousing, Data Warehouse, DWH Requirements (FASMI/Codd)
- OLAP vs. OLTP
- Distinction DWH vs. Big Data vs. Parallel Systems vs. Federated Systems

Fast Analysis on Shared Multidimensional Information  
4 dimensional, User, TAPS, Unstructured, Flexible reporting

## Architecture

- Sketch of DWH Architecture
- Definition/Characteristics of all Components of DWH Architecture

point aggregate  
↓ entire dataset  
Distribution → SUM, COUNT, MIN, MAX  
↓ computed from intermediate values  
Algebraic → AVG, STDDEV  
Holistic → Median, Rank, Percentile  
↓ entire dataset is needed

## Multidimensional Data Model

- Terms: Cube, Operations on Cube, Summation Types, Types of Aggregate Functions, Types of Hierarchies → Simple, Parallel
- Star vs. Snowflake Schema
- Slowly Changing Dimensions
- ETL
- Bulk Loading, ETL, ELT
- Monitoring and Extraction Techniques for Different Source Types
- Differential-Snapshot Problem
- Duplicate Detection
- Transformation on Example
- Identify Data Quality Issues (via SQL)

Simple Hierarchies – Present a straightforward classification with each level aggregating directly into the next. For instance, "Product Category → Product Family → Product Group → Item" shows successive levels leading up to a top node.

Parallel Hierarchies – Within a single dimension, multiple, independent paths coexist without hierarchical dependence, such as a time dimension that includes separate paths for "Year → Quarter → Month → Day" and "Week," allowing flexible aggregation levels - can use roll up or drill down to navigate in between the levels

## OLAP Queries

- Terms: CUBE, ROLLUP, GROUPING SETS, OVER-clause, Special Aggregate Functions
- Cube on Example
- Iceberg Cube → optimized cube only necessary aggregations fit into memory

```
SELECT A, B, C, COUNT(*), SUM(X) FROM R GROUP BY CUBE(A, B, C) HAVING COUNT(*) >= N;
```

notes Page 36

Relational  
• ROLAP  
• Sparse Storage  
• Relational tables only non empty cells  
• Storage proportional to non empty cells and # of each record (both # of rows and # of columns)

MOLAP  
• denormalized storage  
• empty cells are included in the structure  
• data stored in compressed multidimensional array  
• Storage depends on cube size and potential filling degree  
• NOLAP S2E  
• cube size < log<sub>2</sub> dimension # of values  
• cube size < log<sub>2</sub> dimension # of values

## Storage Structures

- Terms: Partitioning, Materialization Strategies → Hybrid
- ROLAP vs. MOLAP vs. HOLAP → highly compressed version of multidimensional cube Ex ②
- Data Dwarf on Example
- Column vs. Row Stores

Query Execution  
OLAP usage  
comprison R → C  
Early or Late materialization  
Early or Late materialization

CDSID  
Clustering  
Dimensionality  
Symmetry  
Tuple references  
Dynamic behaviour

## Index Structures

- Characterization of Index Structures
- B-Tree Specialties
- Grid File
- Bitmap Indexes
- UB-Tree
- R-Tree

Notes  
Pg 152

log(n)  
search  
IX  
PDF



## Query Optimization

### Star Join

- Invariant Grouping/Early Pre-Grouping
- Data Cube Lattice & Pipesort

aggregate as early as possible

Data Analysis (BA)

Olap

DW

ETL

operational systems

## Business Intelligence

- Terms: Business Intelligence, Business Intelligence Process, Report Types
- MCQ: K-Means Clustering, Association Rule Mining, Decision Trees

analytical process that transforms company knowledge into action oriented knowledge

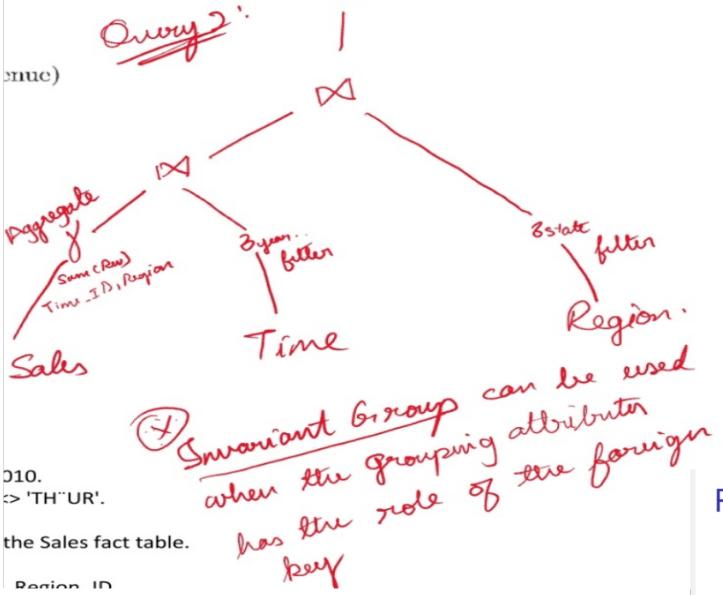
create clusters based on centroid

stem, cluster, support, confidence

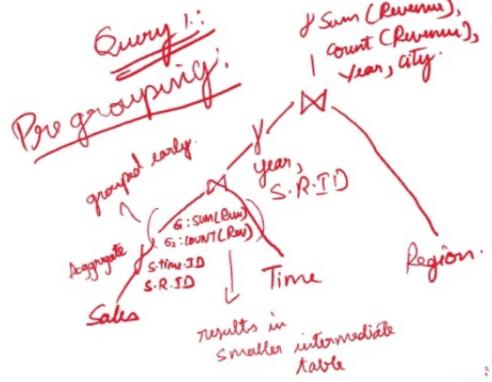
no. of items / total deviation

Ex @ one note

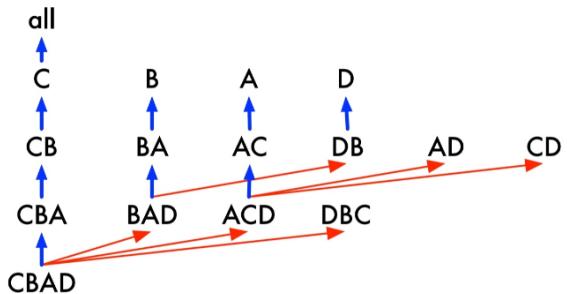
create boxes from the sales table to



## Early/Pre grouping



## PipeSort: Sort Plan



## Assignment 4:

- Goal: find a subgraphs with minimal sum of edge costs.
- edge costs: Sorting: Sort data based on order.
- Pipelining: Additional grouping on subsets of attributes.

### Example:

