Week 4 Summary: Data Cubes, Star Schemas, OLAP, and Iceberg Cubes

Z3003 Data Mining: Scaling to Large Data · 7–9 October 2026 · Dr. Innocent Nyalala

Lecture 10: Multidimensional Data Model and Star Schemas

A data cube models facts (one farm-season), additive or derived measures (yield, revenue), and grouping dimensions (location, time, crop) across $200,000$ farm-seasons ($10,000$ farms $\times 10$ years $\times 2$ seasons). Without hierarchies, n dimensions define 2^n cuboids; with L_i hierarchy levels per dimension, the lattice contains $\prod_{i=1}^{n}(L_i + 1)$ cuboids.

- **Star Schema:** Central fact table joined directly to denormalized dimension tables (contrasted with normalized snowflake schemas and multi-fact constellations).
- **Measure Classes:** Distributive (`SUM`, `COUNT`, `MIN`, `MAX`), Algebraic (`AVG = SUM/COUNT`, standard deviation), and Whole-Table Non-Decomposable (`MEDIAN`, `RANK`). Never average group averages; aggregate `SUM` and `COUNT` separately.

Lecture 11: OLAP Operations and Bitmap Indexing

Analysts navigate cubes via **Roll-up** (coarsen hierarchy), **Drill-down** (add detail; e.g., confirming the 2017 drop from 2.7 t to 2.22 t hit both Masika and Vuli seasons across all regions), **Slice** (fix one dimension), **Dice** (filter ranges across multiple dimensions), and **Pivot** (rotate axes). Low-cardinality columns such as 5 crops and 5 regions use compact bitmaps (25 KB each for $200,000$ rows) to evaluate compound `AND`/`NOT` predicates via bitwise operations.

Lecture 12: Cube Computation at Scale and the Bridge to Pattern Mining

Full cube materialization explodes exponentially ($2^{20} > 10^6$ cuboids). Partial materialization pre-computes high-benefit cuboids greedily, and **iceberg cubes** store only cells meeting `HAVING COUNT(*) >= min_sup`. Replacing dimensions with 22 agro-dealer items produces $2^{22} - 1 = 4,194,303$ itemsets, motivating Apriori.

Worked Check: Weighted Average & Iceberg Basket Rule

District A (100 farms, mean 2.0 t) and District B (300 farms, mean 3.0 t) have true overall mean $(200 + 900)/400 = 2.75$ t, not $(2 + 3)/2 = 2.5$ t. Across $120,000$ agro-dealer baskets, rice seed $\rightarrow$ urea achieves 77.2% confidence and lift 6.80, whereas rice seed $\rightarrow$ hoe shows 54.8% confidence but lift 0.91 because `hoe` appears in 60% of all baskets.