

Indian Institute of Technology Madras
ZANZIBAR CAMPUS

Z3003 · DATA MINING: SCALING TO LARGE DATA

Cube Computation at Scale — and the Road to Pattern Mining

Lecture 12 · Friday 9 October 2026

Dr. Innocent Nyalala | BS Semester 5 (DSAI) | IIT Madras Zanzibar

Session Outline

IIT Madras Zanzibar

1. (0–4 min) Recap: cubes, cuboids, OLAP
2. (4–14 min) Why the full cube explodes: 2^n and $\prod(L_i + 1)$
3. (14–22 min) Materialization: none, full or partial — and iceberg cubes
4. (22–30 min) From cubes to baskets: transactions and itemsets
5. (30–42 min) Support, confidence and lift on 120,000 real baskets
6. (42–50 min) Exercise, summary, Week 5 preview

Learning Outcomes

IIT Madras Zanzibar

By the end of this session, you will be able to:

Explain	why computing every cuboid becomes impossible as dimensions grow
Compare	no, full and partial materialization, and choose which cuboids to pre-compute
Define	an iceberg cube and connect it to a minimum-support threshold
Compute	support, confidence and lift of a rule by hand from basket counts
Explain	why checking every itemset is hopeless — the problem Apriori solves

Cube Computation at Scale — and the Road to Pattern Mining

IIT Madras Zanzibar

→ **Why the Full Cube Explodes**

Materialization: What to Pre-Compute

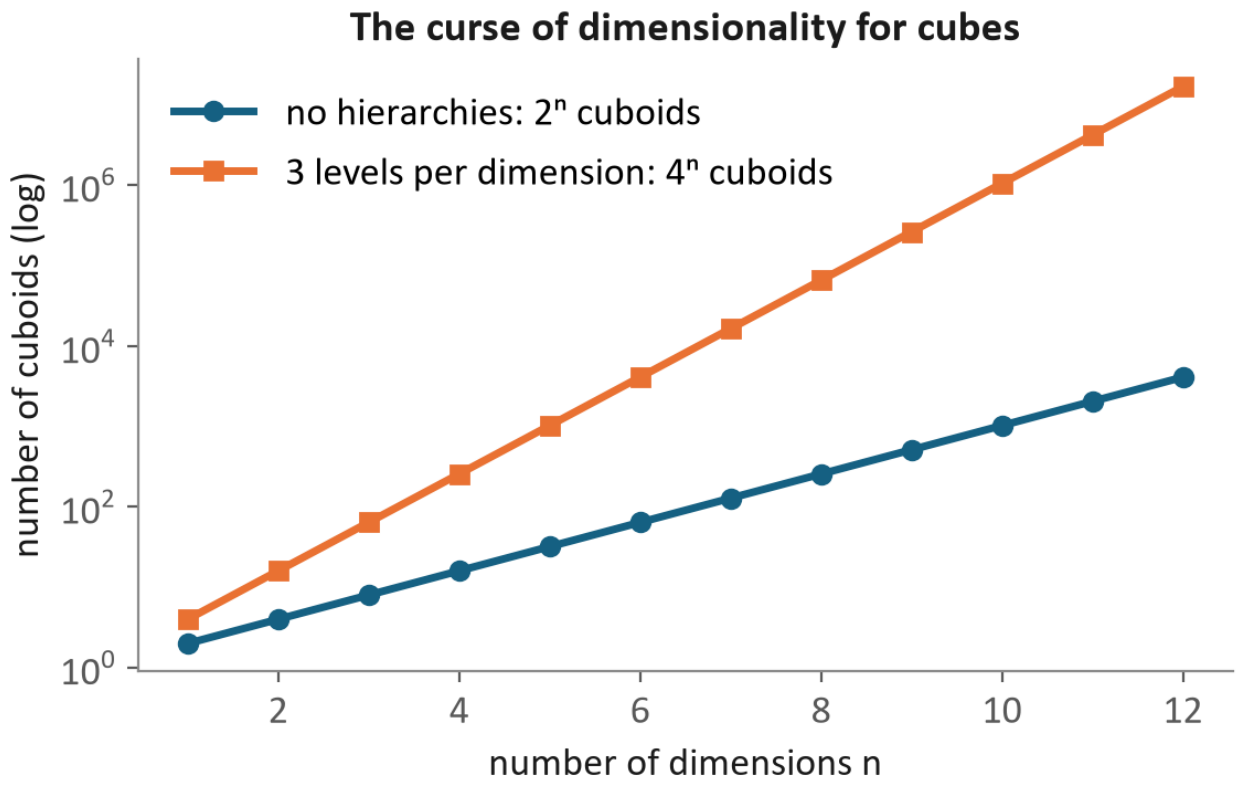
Iceberg Cubes

From Cubes to Baskets: Itemsets

Support, Confidence and Lift

In-Class Exercise

How Many Cuboids? The Count Explodes



dimensions n	2^n	4^n (3 levels each)
3	8	64
5	32	1,024
10	1,024	1,048,576
20	1,048,576	$\approx 1.1 \times 10^{12}$

Plus: every cuboid holds many cells — the base cuboid alone can be as big as the fact table.

Conclusion: we cannot pre-compute every cuboid of a real warehouse. We must choose.

Worked Example: Our Warehouse's Cube

IIT Madras Zanzibar

Dimension	levels	choices
location	district, region	3
time	season, year	3
crop	crop, group	3
farm_type	farm_type	2
Total		**3 × 3 × 3 × 2 = 54**

1 **Base cuboid cells**
$11 \times 5 \times 10 \times 2 \times 4 = 4{,}400$

2 **Add 5 yes/no dimensions**
$54 \times 2^5 = 1{,}728$ cuboids

3 **Real warehouses (20+ dims)**
millions of cuboids

Lesson: a few dimensions are fine; every extra dimension multiplies the work. Choosing what to pre-compute is essential.

Cube Computation at Scale — and the Road to Pattern Mining

IIT Madras Zanzibar

Why the Full Cube Explodes

→ **Materialization: What to Pre-Compute**

Iceberg Cubes

From Cubes to Baskets: Itemsets

Support, Confidence and Lift

In-Class Exercise

Materialization: Pre-Cook, or Cook to Order?

IIT Madras Zanzibar

No materialization

compute every query from the fact table

+ no extra storage

− slow answers

Full materialization

pre-compute every cuboid

+ instant answers

− impossible storage & refresh

Partial materialization

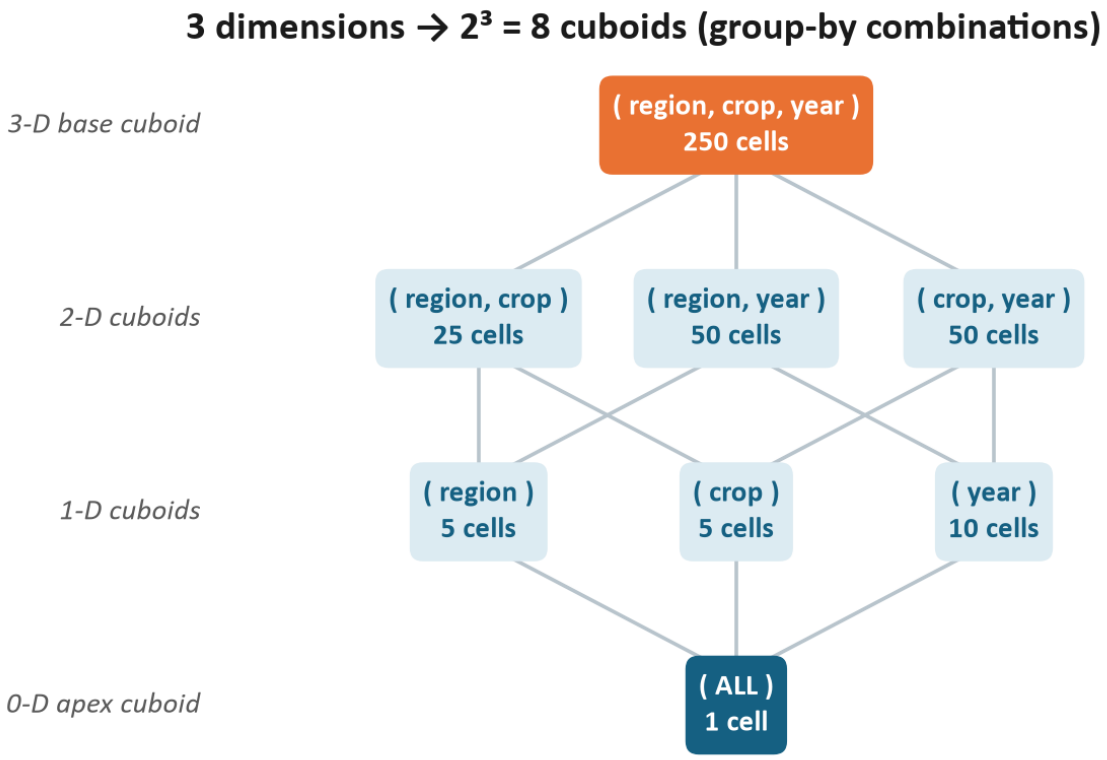
pre-compute a chosen few

+ fast where it matters

− must choose wisely

Restaurant analogy: cook to order (none), cook every dish at dawn (full), or pre-cook the popular dishes and cook the rest on demand (partial). Real systems choose partial.

Which Cuboids to Pre-Compute?



1. **Asked often**
   the Minister's dashboard: region × crop × year

2. **Derives many others**
   store region × crop → region and crop in 25 additions

3. **Greedy choice**
   add the cuboid that saves the most query time; repeat

Why it works: SUM and COUNT are distributive — a coarser cuboid is just a sum over a finer one already stored.

Cube Computation at Scale — and the Road to Pattern Mining

IIT Madras Zanzibar

Why the Full Cube Explodes

Materialization: What to Pre-Compute

→ **Iceberg Cubes**

From Cubes to Baskets: Itemsets

Support, Confidence and Lift

In-Class Exercise

Iceberg Cubes: Keep Only the Cells That Matter

IIT Madras Zanzibar

```
SELECT district, crop, year, season, COUNT(*)
FROM farm_seasons
GROUP BY district, crop, year, season
HAVING COUNT(*) >= 100;
```

1 **Cells in the base cuboid**
11 × 5 × 10 × 2 = 1,100

2 **Cells with ≥ 100 records**
1,000 kept — 100 dropped

Iceberg analogy

Only the tip above water matters to a ship.

Only cells with enough data matter to an analyst.

With many dimensions, most cells are nearly empty — dropping them saves most of the storage.

Remember this phrase: 'keep what occurs at least min_count times' — it is the definition of a FREQUENT pattern.

Cube Computation at Scale — and the Road to Pattern Mining

IIT Madras Zanzibar

Why the Full Cube Explodes

Materialization: What to Pre-Compute

Iceberg Cubes

→ **From Cubes to Baskets: Itemsets**

Support, Confidence and Lift

In-Class Exercise

A New Kind of Data: Shopping Baskets

IIT Madras Zanzibar

transaction	items bought
1	rice seed, urea, herbicide, hoe
2	clove seedlings, shade net, drying mat
3	hoe, sacks
4	sprayer, pesticide, gloves
…	120,000 baskets in total

1. **Transaction (basket)**
   one visit — like one receipt

2. **Item**
   one product; 22 distinct items

3. **Itemset**
   any set of items, e.g. {rice seed, urea}

4. **The question**
   which itemsets occur together often — and matter?

Everyday version: your supermarket receipt; Jumia's 'customers who bought this also bought'; Netflix's 'because you watched'.

Baskets ARE a Cube

IIT Madras Zanzibar

basket	rice seed	urea	hoe	shade net	…
1	1	1	1	0	…
2	0	0	0	1	…
3	0	0	1	0	…
4	1	1	0	0	…

1. **Each item = a yes/no dimension**
   22 items → 22 binary dimensions

2. **Itemset count = one cube cell**
   rice seed = 1 AND urea = 1

3. **Frequent itemset = iceberg cell**
   count ≥ minimum support

Same problem, new clothes: 22 yes/no dimensions → $2^{22} - 1 = 4{,}194{,}303$ possible itemsets, just as cuboids exploded.

Cube Computation at Scale — and the Road to Pattern Mining

IIT Madras Zanzibar

Why the Full Cube Explodes

Materialization: What to Pre-Compute

Iceberg Cubes

From Cubes to Baskets: Itemsets

→ **Support, Confidence and Lift**

In-Class Exercise

Support, Confidence and Lift: What Each One Tells Us

IIT Madras Zanzibar

Support

How often does it happen?

Definition

The share of all baskets that contain the itemset.

Why it matters

Rare patterns are noise. A minimum support keeps only frequent itemsets, so the search stays small and the results are worth acting on.

$$\text{baskets with A} \;/\; \text{all baskets}$$

Confidence

How reliable is the rule?

Definition

Of the baskets that contain A, the share that also contain B.

Why it matters

It shows how far we can trust 'if A, then B' as a prediction, for example stocking urea next to rice seed.

$$\text{support}(A \cup B) \;/\; \text{support}(A)$$

Lift

Better than chance?

Definition

Confidence divided by how common B is across all baskets.

Why it matters

A popular item looks linked to everything. Lift removes that effect, so only real associations remain.

$$\text{confidence}(A \rightarrow B) \;/\; \text{support}(B)$$

Use all three: support filters out rare patterns, confidence ranks how reliable a rule is, lift confirms the link is real.

Reading Lift: Below 1, Equal to 1, Above 1

IIT Madras Zanzibar

Lift below 1
Negative association

What it means

A makes B less likely than usual. The two items tend to avoid each other.

Example

rice seed → hoe: lift 0.91. Rice buyers buy hoes slightly less often than the average customer.

Ignore the rule

Lift equal to 1
Independent

What it means

A tells us nothing about B. B appears as often with A as without it.

Example

B is in 60% of all baskets and in 60% of baskets with A. Lift = 0.60 / 0.60 = 1.

No rule to act on

Lift above 1
Positive association

What it means

A makes B more likely than usual. The higher the number, the stronger the link.

Example

rice seed → urea: lift 6.80. Rice buyers are almost seven times as likely as average to buy urea.

Act on it: stock or bundle

Rule of thumb: lift 2 means twice as likely as average; lift 0.5 means half as likely.

Support: How Often Does It Happen?

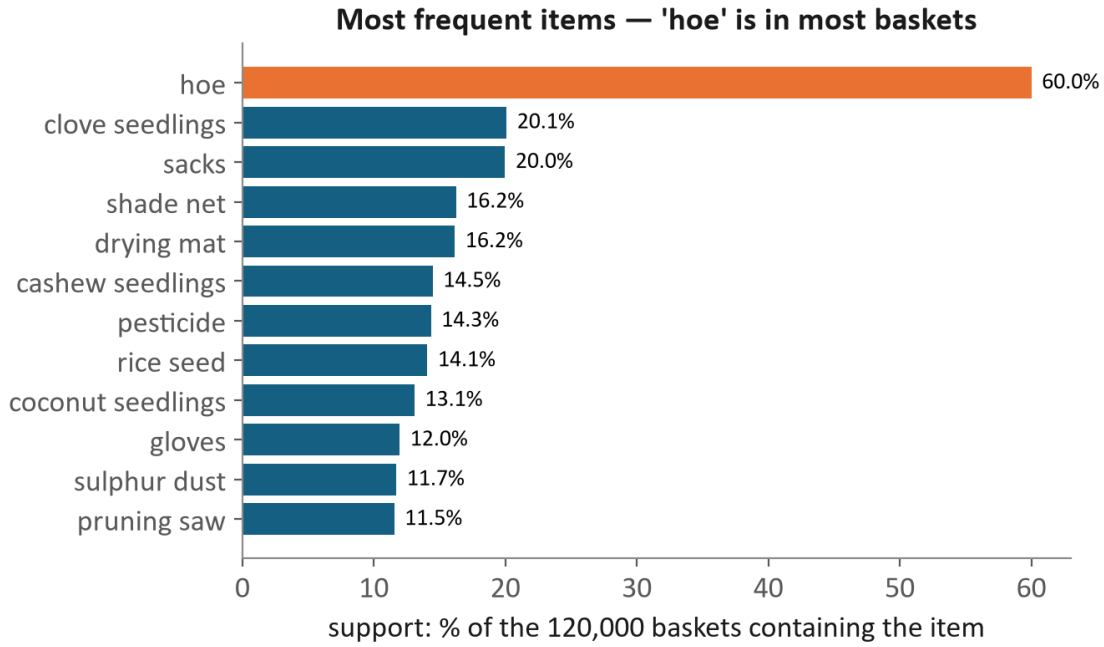
$$\text{support}(A) = \text{baskets containing } A \;/\; \text{all baskets}$$

frequent $\Leftrightarrow$ *support* $\geq$ *min_sup*

1 **support(rice seed)**
16,889 / 120,000 = 14.1%

2 **support({rice seed, urea})**
13,043 / 120,000 = 10.9%

min_sup = 5%? {rice seed, urea} at 10.9% is frequent.



Confidence: How Reliable Is the Rule?

$$\text{confidence}(A \rightarrow B) = \text{support}(A \cup B) \,/\, \text{support}(A)$$

of the baskets containing A, the share that also contain B

1 **rice seed → urea**
13,043 / 16,889 = 77.2%

2 **rice seed → hoe**
9,250 / 16,889 = 54.8%

Both look strong … 77% and 55% confidence. But hoe is in 60% of ALL baskets. Is 55% impressive — or less than we'd expect anyway?

Lift: Is the Rule Better Than Chance?

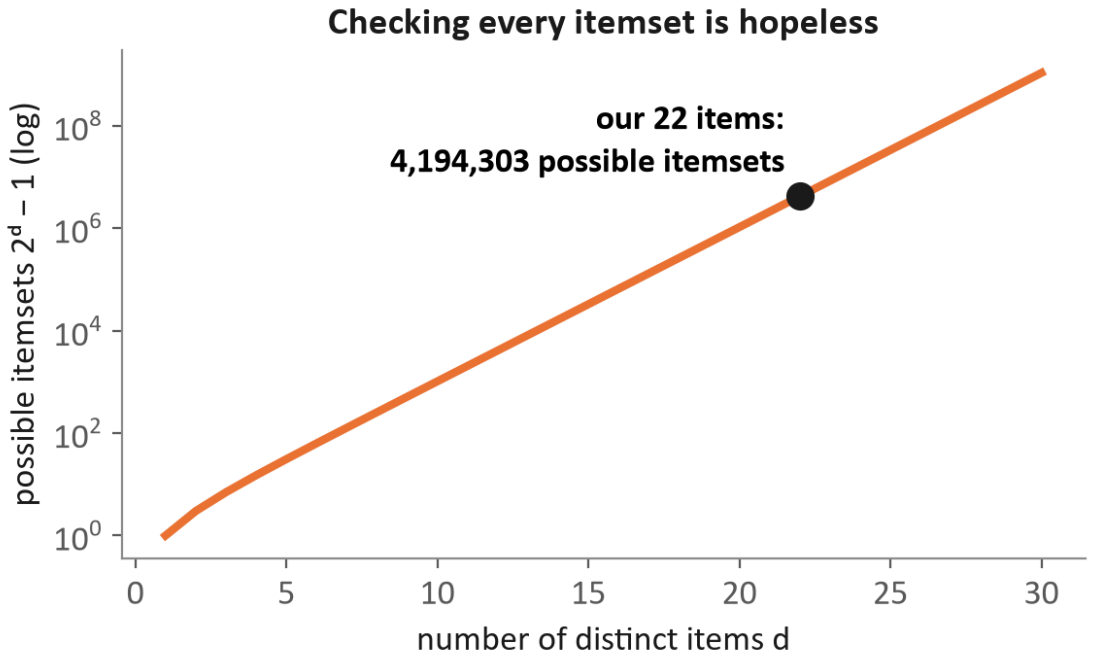
$$\text{lift}(A \rightarrow B) = \text{confidence}(A \rightarrow B) \,/\, \text{support}(B)$$

> 1 real association · = 1 independent · < 1 negative

Rule	confidence	support(B)	lift	Verdict
rice seed → urea	0.772	0.114	**6.80**	strong real association
sprayer → pesticide	0.635	0.143	**4.43**	real association
rice seed → hoe	0.548	0.600	**0.91**	an illusion: hoe is just popular

Lesson: high confidence is not enough. Always check lift — a very popular item makes any rule toward it look strong.

Why Not Just Count Every Itemset?



1 **Our 22 items**
$2^{22} - 1 = 4{,}194{,}303$ itemsets

2 **× 120,000 baskets**
$\approx 5 \times 10^{11}$ checks

3 **A supermarket (10,000 items)**
2^{10000} — impossible

Hint for Week 5 (Apriori): if {rice seed, urea} is NOT frequent, can {rice seed, urea, hoe} be frequent? No — a bigger set never occurs more often than a smaller set inside it. That one idea prunes almost everything.

Pattern Mining in the Real World

IIT Madras Zanzibar

1 Supermarkets

shelf placement and promotions from receipts

2 E-commerce (Jumia, Amazon)

'customers who bought this also bought'

3 Streaming (Netflix, YouTube)

'because you watched …'

4 Mobile money

unusual combinations of transactions flag fraud

5 Healthcare

symptoms or drugs that occur together

6 Agro-dealers (our data)

stock urea next to rice seed; bundle offers

Common thread: find what goes together, then check with lift that it is more than popularity.

Cube Computation at Scale — and the Road to Pattern Mining

IIT Madras Zanzibar

Why the Full Cube Explodes

Materialization: What to Pre-Compute

Iceberg Cubes

From Cubes to Baskets: Itemsets

Support, Confidence and Lift

→ **In-Class Exercise**

In-Class Exercise (6 minutes, individually)

IIT Madras Zanzibar

IN-CLASS TASK

1,000 baskets: 200 contain bread; 400 contain milk; 150 contain both bread and milk.

1. Support of {bread, milk}?
2. Confidence of bread → milk?
3. Lift of bread → milk? Is the rule interesting?
4. A cube has 6 dimensions, each with 2 levels. How many cuboids?

Solution

IIT Madras Zanzibar

1. **Support**
   150 / 1,000 = 15%

2. **Confidence**
   150 / 200 = 75%

3. **Lift**
   0.75 / 0.40 = 1.875 > 1 → yes, bread buyers are almost twice as likely to buy milk

4. **Cuboids**
   $(2 + 1)^6 = 3^6 = 729$

Summary & Exam Checkpoints

IIT Madras Zanzibar

Cubes explode

2^n or $\prod(L_i + 1)$ cuboids — pre-computing everything is impossible.

Partial materialization

store what is asked often and what derives many others; iceberg cubes keep only cells with count ≥ threshold.

Support & confidence

support = share of baskets with the itemset; confidence = P(B | A).

Lift

confidence / support(B): > 1 real, < 1 illusion. Rice seed → hoe: lift 0.91.

What's Next: Week 5

IIT Madras Zanzibar

Wed **Apriori**
the pruning idea, step by step, on our baskets

Thu **FP-Growth**
mining without generating candidates; tutorial timing both

Fri **Pattern evaluation**
lift, χ^2, and other interestingness measures

Mini-project: dataset approval checkpoint is next Friday, 16 October.

References & Further Reading

- Han, J., Pei, J. & Tong, H. (2023). Data Mining: Concepts and Techniques, 4th ed. — Chapter 4 (cube computation) and Chapter 6 (frequent patterns).
- Harinarayan, V., Rajaraman, A. & Ullman, J. D. (1996). Implementing data cubes efficiently. Proc. ACM SIGMOD, 205–216.
- Beyer, K. & Ramakrishnan, R. (1999). Bottom-up computation of sparse and iceberg CUBEs. Proc. ACM SIGMOD, 359–370.
- Agrawal, R., Imieliński, T. & Swami, A. (1993). Mining association rules between sets of items in large databases. Proc. ACM SIGMOD, 207–216.