

Indian Institute of Technology Madras
ZANZIBAR CAMPUS

Z3003 · DATA MINING: SCALING TO LARGE DATA

The Multidimensional Data Model: Data Cubes & Schemas

Lecture 10 · Wednesday 7 October 2026

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

Session Outline

IIT Madras Zanzibar

1. (0–5 min) Recap: the warehouse, and this week's new 200,000-row dataset
2. (5–15 min) From a flat table to a cube — you already use cubes every day
3. (15–30 min) Facts, measures and dimensions; cuboids and the lattice
4. (30–40 min) Concept hierarchies: district → region, season → year
5. (40–58 min) Star, snowflake and fact-constellation schemas — designed for our data
6. (58–65 min) Distributive, algebraic and holistic measures
7. (65–75 min) Exercise, summary, questions

Learning Outcomes

IIT Madras Zanzibar

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

Explain	what a data cube is, using a table you already know (school results, M-Pesa statements)
Identify	the fact, measures and dimensions in a dataset, and count its cuboids
Draw	concept hierarchies and use them to count cuboids: $\prod (L_i + 1)$
Design	a star schema for our farm data, and compare it with snowflake and fact constellation
Classify	measures as distributive, algebraic or holistic — and explain why it matters at scale

Recap: The Warehouse Is Built — Now Organize It

IIT Madras Zanzibar

Friday

WHY a warehouse; SITN; ETL fills it

Today

HOW data is organized inside: cubes and schemas

Tomorrow

HOW analysts move through it: OLAP operations

Analogy: a warehouse without a model is a library with books piled on the floor. The cube is the shelving system: every book can be found by subject, author and year.

This Week's Dataset: 200,000 Farm-Seasons

IIT Madras Zanzibar

farm	year	season	region	district	crop	size (ha)	rain (mm)	yield (t)	price	revenue ($)
1	2016	Masika	Pemba North	Micheweni	Cashew	6.27	1,209	1.60	1.90	3,035
1	2016	Vuli	Pemba North	Micheweni	Cashew	6.27	1,096	1.77	1.89	3,384
1	2017	Masika	Pemba North	Micheweni	Cashew	6.27	1,315	1.24	2.13	2,640
…	…	…	…	…	…	…	…	…	…	…

1. **10,000 farms** same crops and prices as Weeks 1–3
2. **× 10 years** 2016–2025, including a 2017 drought
3. **× 2 seasons** Masika (long rains), Vuli (short rains)
4. **= 200,000 rows** one farm in one season per row

Files: datasets/zanzibar_farm_seasons.csv (200,000 rows), zanzibar_farms.csv (10,000 farms), zanzibar_input_purchases.csv (120,000 shopping baskets — Friday). Synthetic teaching data; see DATA_DICTIONARY.md.

The Multidimensional Data Model: Data Cubes & Schemas

IIT Madras Zanzibar

→ **Foundations: From a Table to a Cube**

Facts, Measures & Dimensions

Concept Hierarchies

Warehouse Schemas: Star, Snowflake, Constellation

Measures: Which Totals Can Be Combined?

In-Class Exercise

You Already Use Cubes: Your Exam Results

IIT Madras Zanzibar

	Maths	Physics	English
Amina	78	65	82
Baraka	54	71	60
Chausiku	91	88	75

One term's marks: student × subject. Stack one sheet per term → a cube.

Same marks, many views. A data cube stores data so that every view — any slice, any total — is quick to get.

1. **Class teacher**
   one TERM: all students × all subjects
2. **Maths teacher**
   one SUBJECT: all students × all terms
3. **Parent**
   one STUDENT: all subjects × all terms (the report card)
4. **Head teacher**
   AVERAGE per subject per term: all students rolled into one number

Our farms: region × crop × year instead of student × subject × term.

From a Flat Table to a Cross-Tab

SUM(revenue), $ millions, 2016–2025 — one face of the cube

	Clove	Cashew	Rice	Coconut	Cassava
North Unguja	162.2	15.3	15.2	9.2	5.1
South Unguja	119.5	14.3	9.3	8.8	8.1
Zanzibar Urban	140.5	24.9	10.2	8.3	5.6
Pemba North	125.4	27.3	12.4	8.1	10.7
Pemba South	196.0	19.1	16.8	6.4	7.2

```
SELECT region, crop,
       SUM(revenue_usd)
FROM farm_seasons
GROUP BY region, crop;
```

1. **200,000 rows → 25 numbers**
   each cell = one group's total

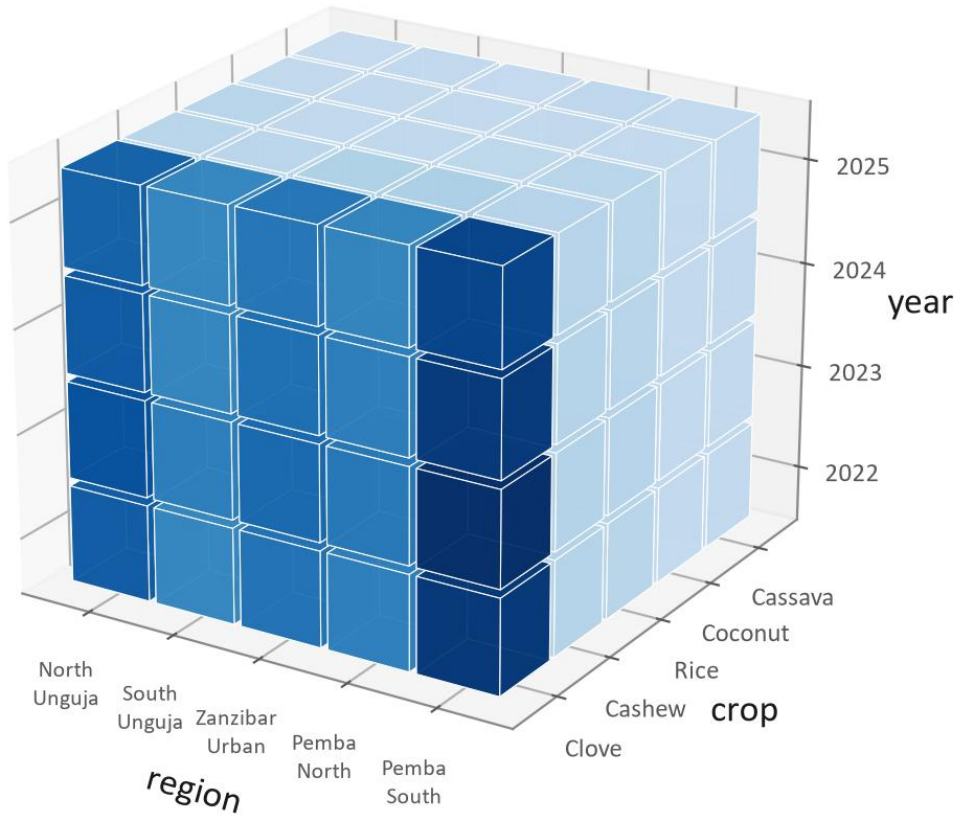
2. **Top cell**
   Pemba South × Clove ≈ $196 M

Key idea: a cross-tab is one face of a cube. Add a third grouping column (year) and the table becomes a cube.

What Is a Data Cube?

IIT Madras Zanzibar

Each small cube = total revenue for one (region, crop, year)



Definition

A data cube stores a measure (e.g. total revenue) for every combination of values of several dimensions.

1 **One cell**

Pemba South × Clove × 2023 = $23.2 M

2 **Size of this cube**

5 regions × 5 crops × 10 years = 250 cells

More than 3 dimensions? same idea — a 'hypercube' we cannot draw.

The Multidimensional Data Model: Data Cubes & Schemas

IIT Madras Zanzibar

Foundations: From a Table to a Cube

→ **Facts, Measures & Dimensions**

Concept Hierarchies

Warehouse Schemas: Star, Snowflake, Constellation

Measures: Which Totals Can Be Combined?

In-Class Exercise

Facts, Measures and Dimensions

IIT Madras Zanzibar

Fact

one business event we record

e.g., one farm's harvest in one season

Measure

a number about the fact that we add up or average

e.g., yield_tonnes, revenue_usd, rainfall_mm

Dimension

a way to group or filter facts

e.g., region, district, crop, year, season

Example	Fact	Measure	Dimensions
M-Pesa statement	one transaction	amount (TZS)	date · type (send / withdraw / pay) · agent
Our farm data	one farm in one season	yield, revenue, rainfall	region · district · crop · year · season

Quick test: can you meaningfully ADD it up? Revenue — yes, a measure. Region — no, a dimension.

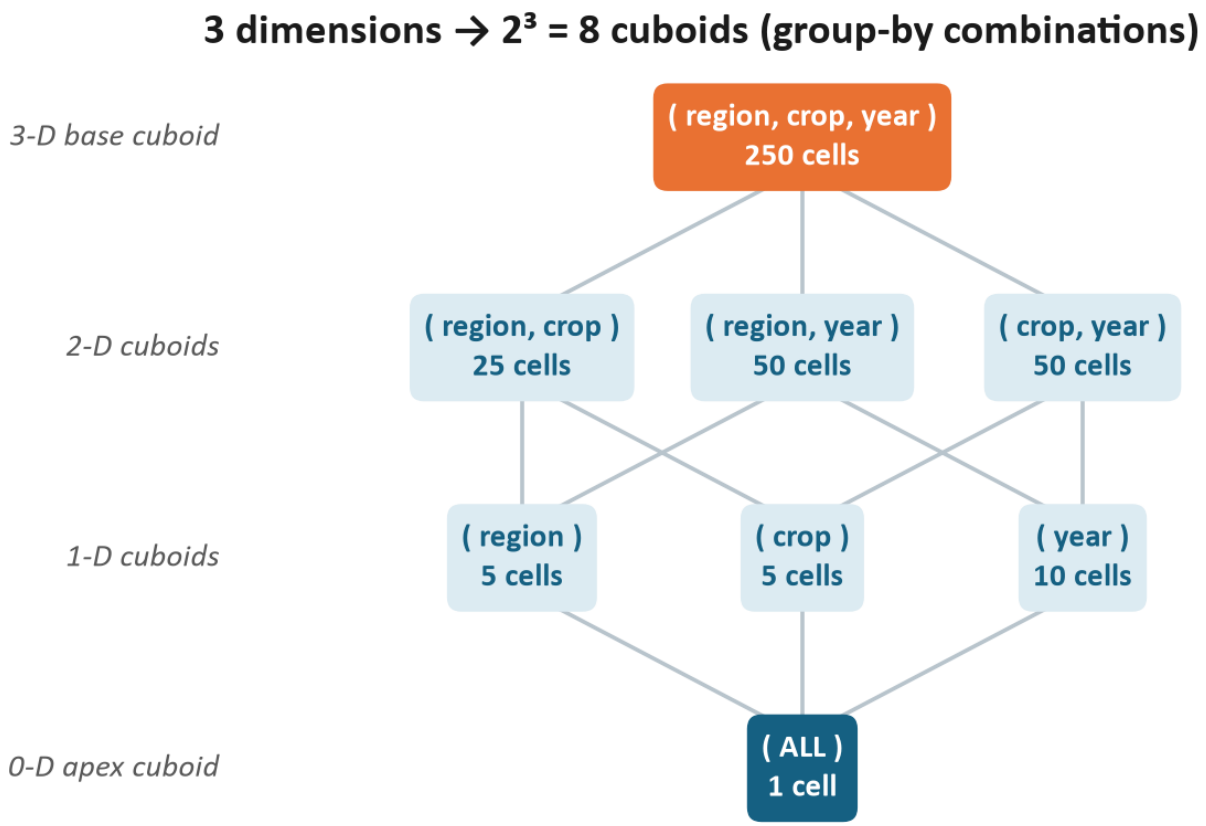
Classify Every Column of Our Dataset

IIT Madras Zanzibar

Column	Role	Why
farm_id, record_id	key	identifies a row — never add it up
year, season	dimension (time)	group by when
region, district	dimension (location)	group by where
crop, farm_type	dimension	group by what / who
fertilizer_used, irrigated	dimension	group by practice (Yes/No)
yield_tonnes, revenue_usd	measure	adding up makes sense
rainfall_mm, price_per_kg	measure (average)	averaging makes sense; summing does not
farm_size_ha	farm attribute (or measure)	describes the farm, not the season

Watch out: a number is not automatically a measure — IDs and codes are numbers you must never sum.

Cuboids: Every Way to Group the Data



$$\text{number of cuboids} = 2^n$$

each of the n dimensions is either in or out

1. **Base cuboid**
   grouped by ALL dimensions (most detail)
2. **Apex cuboid**
   grouped by NONE: the grand total
3. **Lines**
   a lower cuboid can be computed from the one above

Check: 3 dimensions → 2 × 2 × 2 = 8 cuboids. 10 dimensions → 2^{10} = 1,024 cuboids.

Three In-or-Out Choices Give 2 × 2 × 2 = 8 Cuboids

IIT Madras Zanzibar

Region	Crop	Year	Cuboid (group by)	Cells
✓	✓	✓	**(region, crop, year)**	250
✓	✓	—	(region, crop)	25
✓	—	✓	(region, year)	50
—	✓	✓	(crop, year)	50
✓	—	—	(region)	5
—	✓	—	(crop)	5
—	—	✓	(year)	10
—	—	—	**(ALL) grand total**	1

Step 1
Region: in or out → **2**

Step 2
Crop: in or out doubles it → **4**

Step 3
Year: in or out doubles it → **8**

$$2 \times 2 \times 2 = 2^3 = 8$$

n dimensions → 2^n cuboids

Every new dimension doubles the count. 10 dimensions → 2^{10} = 1,024 cuboids.

The Multidimensional Data Model: Data Cubes & Schemas

IIT Madras Zanzibar

Foundations: From a Table to a Cube

Facts, Measures & Dimensions

→ **Concept Hierarchies**

Warehouse Schemas: Star, Snowflake, Constellation

Measures: Which Totals Can Be Combined?

In-Class Exercise

Concept Hierarchies: Levels Inside a Dimension

IIT Madras Zanzibar



Crop groups
cash crops: clove, cashew · food crops: rice, cassava, coconut

Everyday version: Google Maps — street → town → district → country. Zooming out = moving up the hierarchy.

Each Dimension Offers Its Levels Plus ALL: That Is the +1

IIT Madras Zanzibar

Location		Time		Crop
district (11)		season (20)		crop (5)
region (5)	×	year (10)	×	crop group (2)
ALL (1)	+1	ALL (1)	+1	ALL (1) +1
2 levels + 1 = 3 choices		2 levels + 1 = 3 choices		2 levels + 1 = 3 choices

3 × 3 × 3 = 27 cuboids

Why +1? ALL means the dimension is rolled up completely and left out of the group-by.
No hierarchy: $L_i = 1$, so $L_i + 1 = 2$ (in or out). That is why 2^n is the special case.

Counting Cuboids With Hierarchies

IIT Madras Zanzibar

$$\text{cuboids} = \prod (L_i + 1)$$

L_i = levels in dimension i (not counting ALL); $\prod$ = multiply

Dimension	Levels L_i	Choices $L_i + 1$
Location	2 (district, region)	3
Time	2 (season, year)	3
Crop	2 (crop, crop group)	3
Total		**3 × 3 × 3 = 27**

1. **No hierarchies**
   2 × 2 × 2 = 8 cuboids

2. **With hierarchies**
   3 × 3 × 3 = 27 cuboids

3. **Add a 4th dimension (2 levels)**
   27 × 3 = 81 cuboids

Why it matters: cuboids multiply fast. Friday: we cannot pre-compute them all, so we must choose.

Worked Example: M-Pesa Transactions Have 24 Cuboids

IIT Madras Zanzibar

Step 1 List each dimension's levels, from most detail up to ALL. *Cube: transaction amount by date × type × agent*

Date	Transaction type	Agent
day	type	agent
month	ALL	district
year		ALL
ALL		
+1	+1	+1

Step 2 Choices per dimension = levels + 1.

3 levels + 1 = 4	×	1 level + 1 = 2	×	2 levels + 1 = 3

Step 3 $4 \times 2 \times 3 = 24$ cuboids

Type has one level, so it gives 2 choices (in or out), just like the 2^n case.

The Multidimensional Data Model: Data Cubes & Schemas

IIT Madras Zanzibar

Foundations: From a Table to a Cube

Facts, Measures & Dimensions

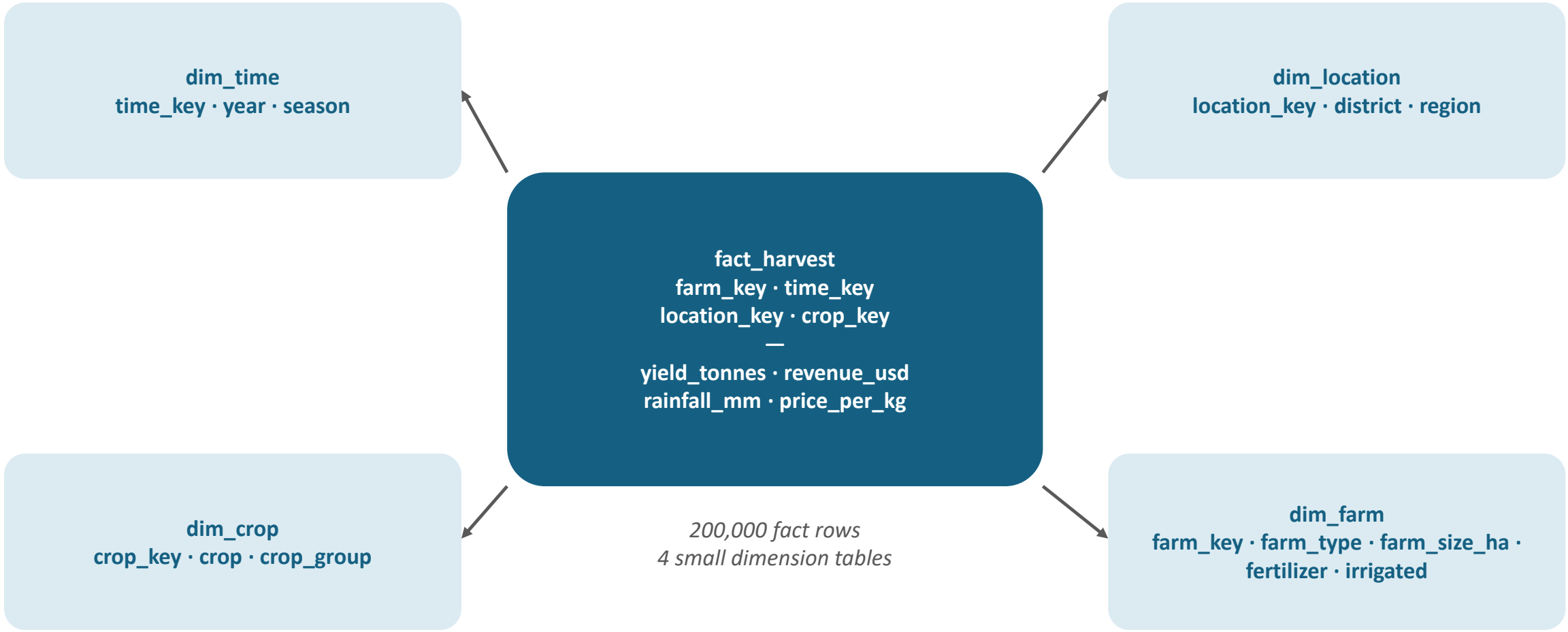
Concept Hierarchies

→ **Warehouse Schemas: Star, Snowflake, Constellation**

Measures: Which Totals Can Be Combined?

In-Class Exercise

The Star Schema, Designed for Our Farms



Star schema: one central fact table (measures + keys) joined to one table per dimension.

Why the Star Shape? A Query in One Join per Dimension

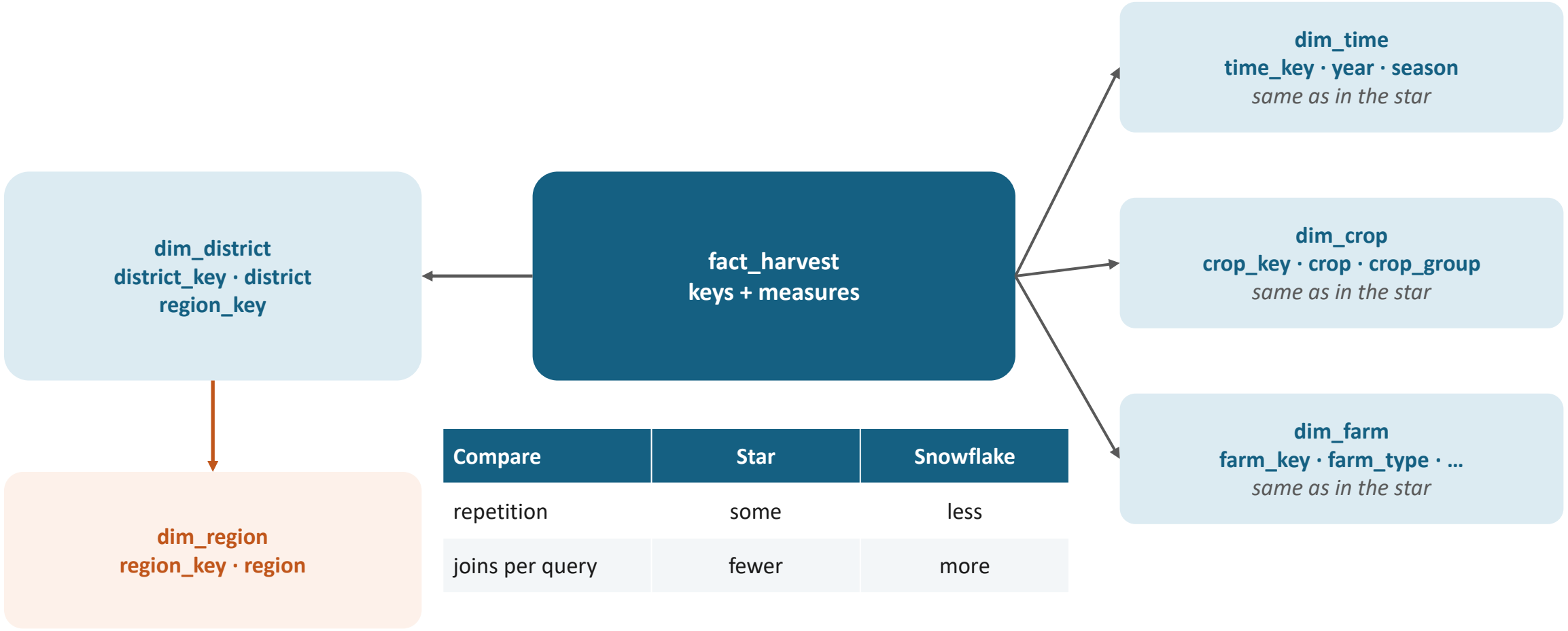
IIT Madras Zanzibar

```
SELECT l.region, SUM(f.revenue_usd)
FROM fact_harvest f
JOIN dim_location l ON f.location_key = l.location_key
JOIN dim_crop     c ON f.crop_key     = c.crop_key
JOIN dim_time     t ON f.time_key     = t.time_key
WHERE c.crop = 'Clove' AND t.year = 2023
GROUP BY l.region;
```

1. **Simple pattern**
   one join per dimension, every time
2. **Small dimensions**
   11 districts · 5 crops · 20 seasons: joins are fast
3. **Less storage**
   keys instead of repeating 'Pemba North' 40,000 times
4. **One place to fix**
   rename a district once, in dim_location

Tools love stars: Power BI, Tableau and Excel PivotTables generate exactly this query pattern when you drag fields onto a chart.

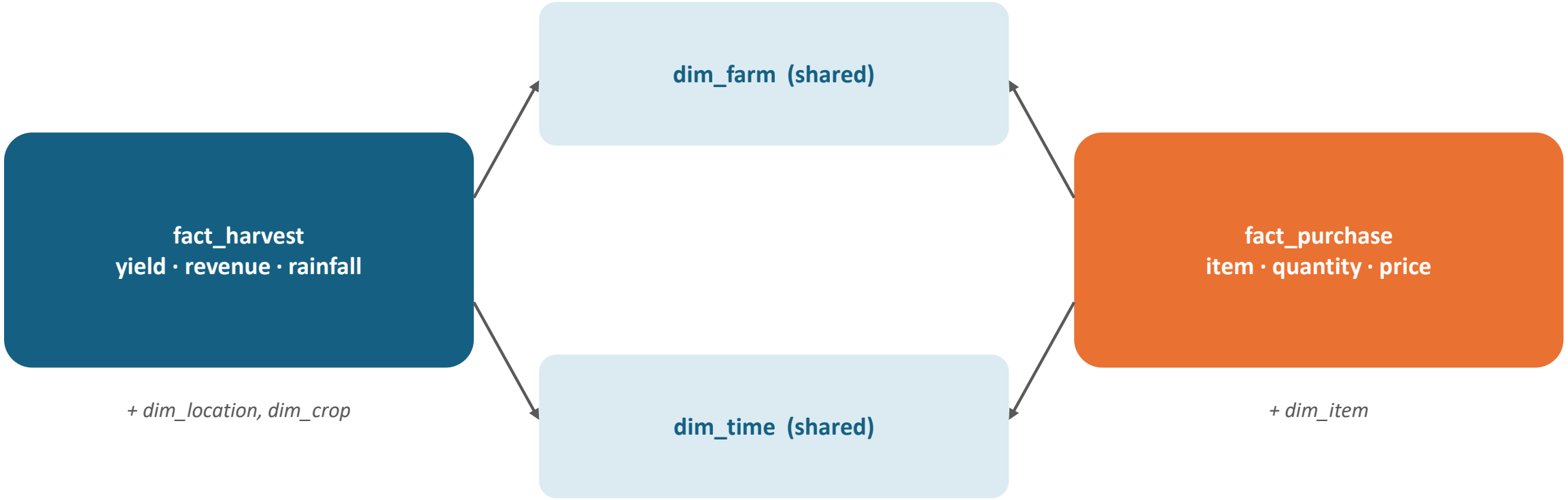
Snowflake Schema: Normalize the Dimensions



Compare	Star	Snowflake
repetition	some	less
joins per query	fewer	more

Rule of thumb: dimension tables are small, so most designers accept the repetition and keep the simpler star.

Fact Constellation: Two Facts Sharing Dimensions



New questions become possible: 'do rice farms that buy urea get higher yields next season?' — joining two processes through shared, conformed dimensions.

Choosing a Schema

IIT Madras Zanzibar

Schema	Shape	Strength	Weakness	Use when
Star	1 fact + 1 table per dimension	simple, fast, tool-friendly	some repetition in dimensions	the usual default
Snowflake	dimensions split into sub-tables	less repetition	more joins, harder to read	a dimension is very large
Fact constellation	several facts share dimensions	cross-process analysis	more design and governance	several business processes

Our warehouse: a constellation — fact_harvest and fact_purchase — each built as a star around conformed farm and time dimensions.

Data Cubes in the Real World

IIT Madras Zanzibar

1 M-Pesa / mobile money

amount by date × transaction type × agent

2 Supermarkets

sales by product × branch × day → restocking

3 DHIS2 health dashboards

cases by facility × disease × month

4 Telecom networks

data usage by tower × hour × plan

5 University records

grades by course × cohort × semester

6 Excel / Power BI

every PivotTable drag-and-drop is a cube query

Common thread: one measure, viewed along several dimensions, at several levels of detail.

The Multidimensional Data Model: Data Cubes & Schemas

IIT Madras Zanzibar

Foundations: From a Table to a Cube

Facts, Measures & Dimensions

Concept Hierarchies

Warehouse Schemas: Star, Snowflake, Constellation

→ **Measures: Which Totals Can Be Combined?**

In-Class Exercise

Three Kinds of Measures

Distributive

combine the parts' results directly

e.g. SUM, COUNT, MIN, MAX

Algebraic

combine a FIXED number of distributive values

e.g. AVG = SUM / COUNT; standard deviation

Holistic

needs ALL the raw values

e.g. MEDIAN, MODE, RANK

✓ **Distributive, on our data**

Pemba South revenue = Chake Chake 123.0 + Mkoani 122.5 = 245.6 $M

Holistic trap: the median of Pemba is NOT the median of its districts' medians — you must go back to the raw rows.

Worked Example: Averaging Averages Goes Wrong

IIT Madras Zanzibar

District	farms (COUNT)	avg yield	SUM = avg × count
A	100	2.0 t	200 t
B	300	3.0 t	900 t
Both	**400**	?	**1,100 t**

✗ **Average of averages**
(2.0 + 3.0) / 2 = 2.5 t — wrong

✓ **SUM ÷ COUNT**
1,100 / 400 = 2.75 t — right

Why?

District B has three times as many farms, so it must count three times as much.

Cube design rule: store SUM and COUNT in every cell; compute AVG = SUM / COUNT only when you display it.

Everyday version: your semester GPA weights each course by its credits — it is not the average of course grades.

The Multidimensional Data Model: Data Cubes & Schemas

IIT Madras Zanzibar

Foundations: From a Table to a Cube

Facts, Measures & Dimensions

Concept Hierarchies

Warehouse Schemas: Star, Snowflake, Constellation

Measures: Which Totals Can Be Combined?

→ **In-Class Exercise**

In-Class Exercise (10 minutes, pairs)

IIT Madras Zanzibar

IN-CLASS TASK

A supermarket chain records every sale: date, branch (branch → city → country), product (product → category), quantity and amount.

1. Name the fact, the measures and the dimensions.
   Hint: the fact is one event; measures are numbers you add up; dimensions are ways to group.

2. Draw a star schema: fact table columns and each dimension table's columns.
   Hint: keys and measures go in the fact table; each dimension table lists its levels as columns.

3. Branch A: 200 sales averaging TZS 5,000; Branch B: 50 sales averaging TZS 20,000. Average sale across both?
   Hint: do not average the two averages. Add the total amounts first, then divide by the number of sales.

4. Time has levels day → month → year. How many cuboids does the cube have (branch, product, time hierarchies)?
   Hint: for each dimension count its levels + 1 (for ALL), then multiply.

Solution

IIT Madras Zanzibar

1

Q1

fact = one sale; measures = quantity, amount; dimensions = date, branch, product

2

Q2

fact_sale(date_key, branch_key, product_key, quantity, amount); dim_date(day, month, year); dim_branch(branch, city, country); dim_product(product, category)

3

Q3

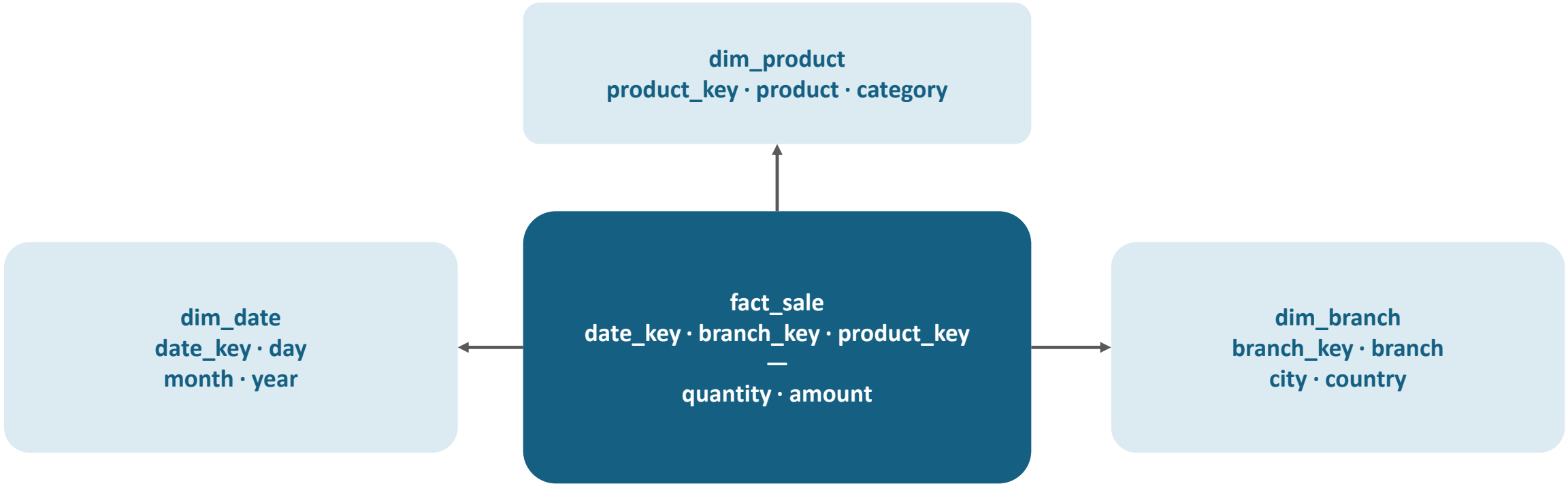
SUMs: 200 × 5,000 = 1,000,000 and 50 × 20,000 = 1,000,000 → 2,000,000 / 250 = TZS 8,000 (not 12,500)

4

Q4

branch 3 levels → 4 choices; product 2 → 3; time 3 → 4; 4 × 3 × 4 = 48 cuboids

Q2 Solution: Star Schema for the Supermarket



1 row per sale
3 small dimension tables

Hierarchies sit inside the dimensions: day → month → year; branch → city → country; product → category.

Summary & Exam Checkpoints

IIT Madras Zanzibar

Cube = measure × dimensions

Every cell is one group's total. Base cuboid = all dims; apex = grand total.

2^n and $\prod(L_i + 1)$

Without hierarchies each dimension is in or out: 3 dimensions give $2^3 = 8$. With hierarchies each gives levels + 1 (for ALL) choices: $3 \times 3 \times 3 = 27$.

Star schema

Fact table (keys + measures) in the middle, one table per dimension. Snowflake splits dims; constellation shares them.

Measures

Distributive (SUM), algebraic (AVG = SUM/COUNT), holistic (MEDIAN). Never average averages.

Common Questions

IIT Madras Zanzibar

Q **Is a data cube a separate file or a picture?**

A Neither: it is the set of all group-by totals, stored in tables or a special OLAP engine. The cube picture is a way to think.

Q **Why not keep region inside the fact table?**

A You could, but repeating text in 200,000 rows wastes space and makes renaming error-prone. Keys fix both.

Q **Is MEDIAN never allowed in a cube?**

A It is allowed, but it cannot be built from smaller cells — you must keep or re-read the raw values.

Q **Is the star schema the same as normalization from databases?**

A The opposite: a star deliberately keeps some repetition (denormalized) to make reading fast.

What's Next

IIT Madras Zanzibar

Thu **OLAP operations**
roll-up, drill-down, slice, dice, pivot on our 200,000 rows; bitmap indexes

Tut **Tutorial 4**
build a star schema and a cube in pandas; answer OLAP questions

Fri **Cube computation at scale**
what to pre-compute; iceberg cubes; the bridge to pattern mining

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

References & Further Reading

- Han, J., Pei, J. & Tong, H. (2023). Data Mining: Concepts and Techniques, 4th ed. — Chapter 3 (data cube, schemas, measures).
- Kimball, R. & Ross, M. (2013). The Data Warehouse Toolkit, 3rd ed. Wiley — dimensional modelling and star schemas.
- Gray, J. et al. (1997). Data cube: a relational aggregation operator generalizing group-by, cross-tab, and sub-totals. Data Mining and Knowledge Discovery, 1, 29–53.