香港城市大學
City University of Hong Kong

專業 創新 胸懷全球
Professional·Creative
For The World

Data Warehouse Modeling:

Data Cube and OLAP

CS5483 Data Warehousing and Data Mining

Motivation

- Suppose you want to know the sales information of a supermarket chain.

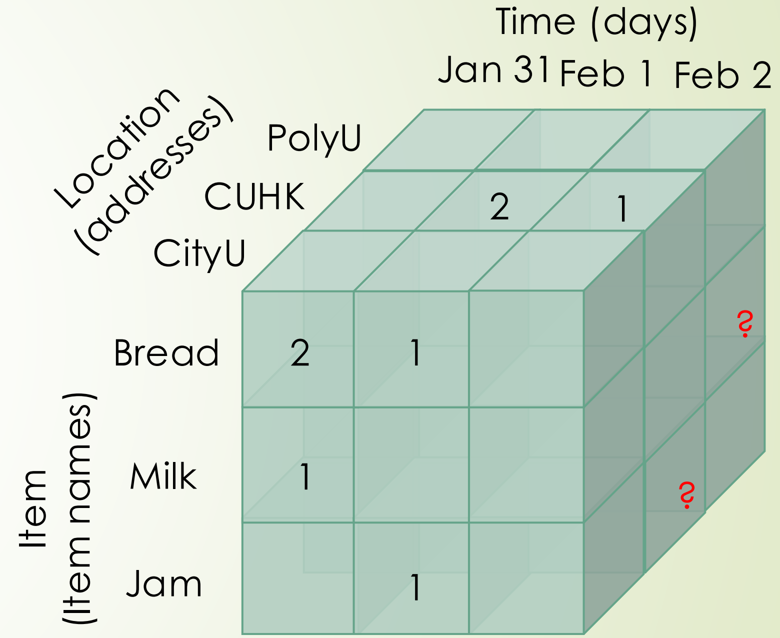
TID	Time	Location	Item (quantity,unit price)
1	Jan 31	CityU, Kln.	bread (2, HK$5), milk (1, HK$10)
2	Feb 1	CUHK, N.T.	bread (2, HK$5)
3	Feb 1	CityU, Kln.	bread (1, HK$5), jam (1, HK$5)
4	Feb 2	CUHK, N.T.	bread (1, HK$5), jam (1, HK$5)
5	Feb 2	PolyU, Kln.	milk (2, HK$10)

- From the above transactional data, which store has the best sales performance?
  - A maximum of ___________ units sold in ______________.
  - A maximum of ___________ dollars sold in _____________.
- How to design a data warehouse for efficient analysis?

Dimension modeling

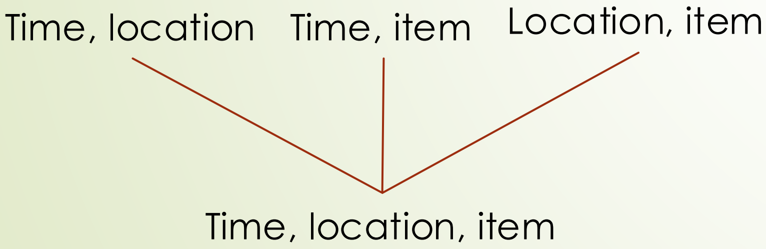
TID	Time	Location	Item (quantity,unit price)
1	Jan 31	CityU, Kln.	bread (2, HK$5), milk (1, HK$10)
2	Feb 1	CUHK, N.T.	bread (2, HK$5)
3	Feb 1	CityU, Kln.	bread (1, HK$5), jam (1, HK$5)
4	Feb 2	CUHK, N.T.	bread (1, HK$5), jam (1, HK$5)
5	Feb 2	PolyU, Kln.	milk (2, HK$10)

- **C**_______________: Multi-dimensional array of **c**_______
  - containing the **f**______ (units sold, dollars sold)
  - indexed using the **d**____________ (time, location, item).
- How to summarize?



Units sold according to time, location, and item.

Dimension reduction

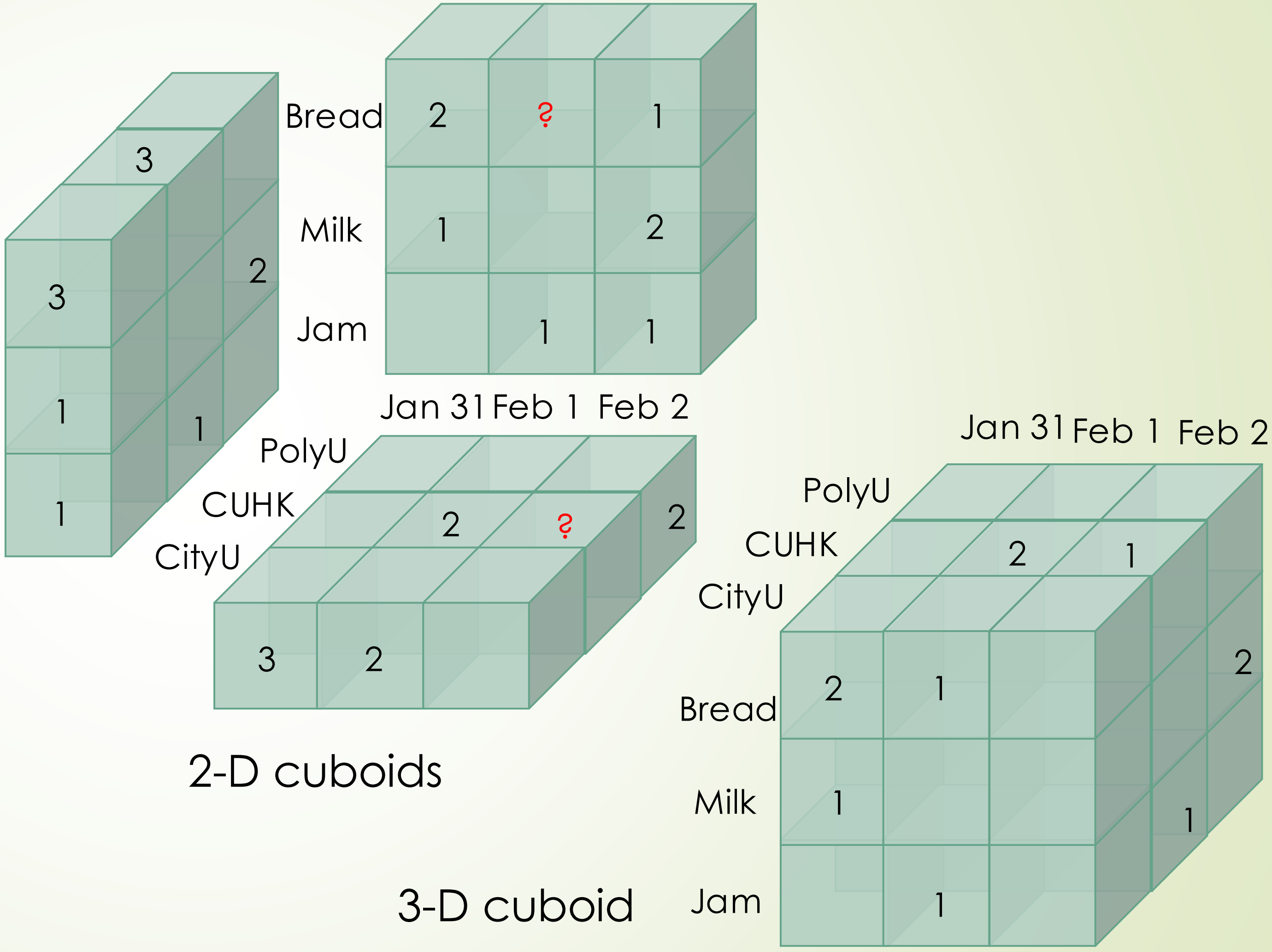


Time, location

Time, item

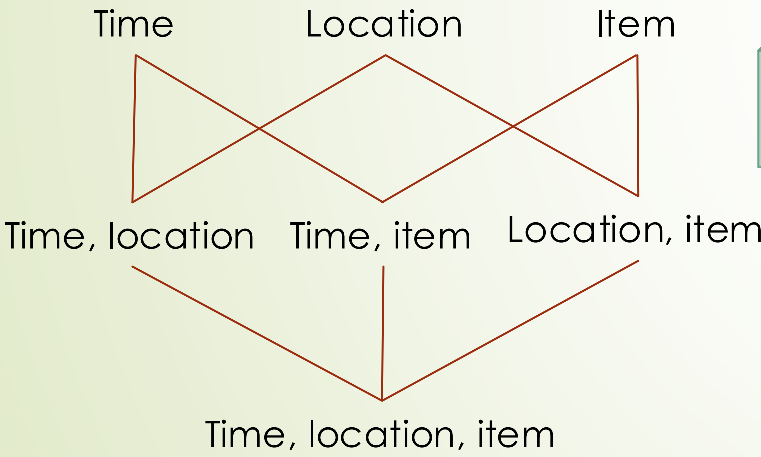
Location, item

Time, location, item



2-D cuboids

3-D cuboid



Time

Location

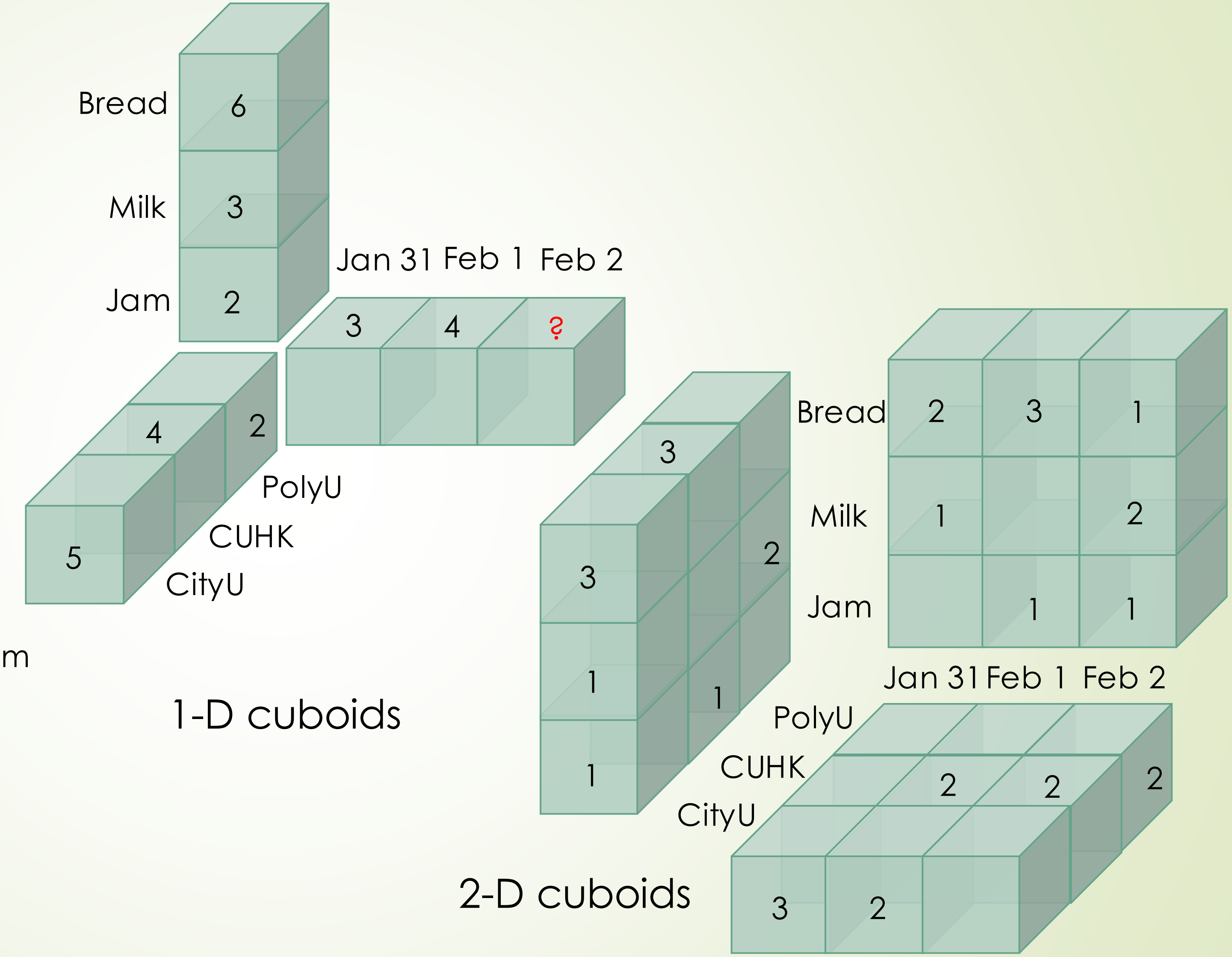
Item

Time, location

Time, item

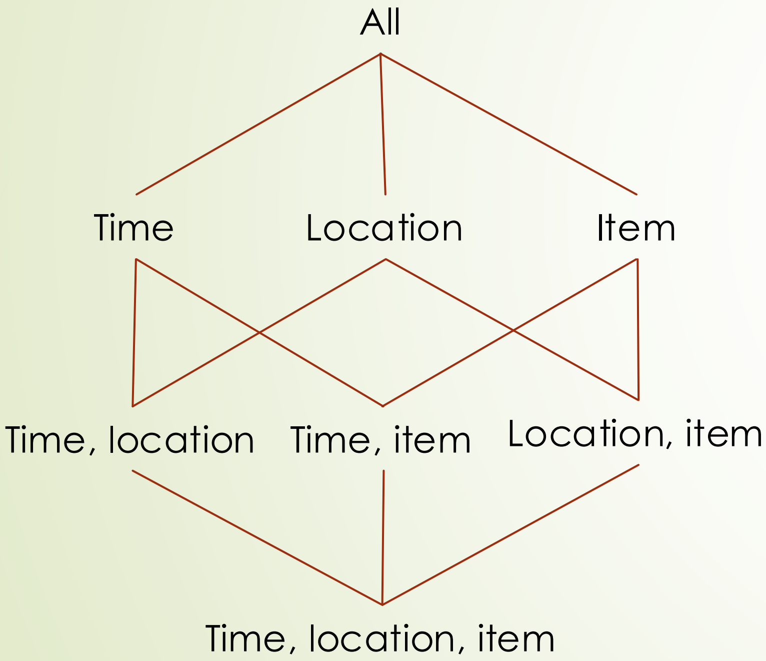
Location, item

Time, location, item



1-D cuboids

2-D cuboids

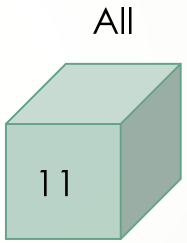


All

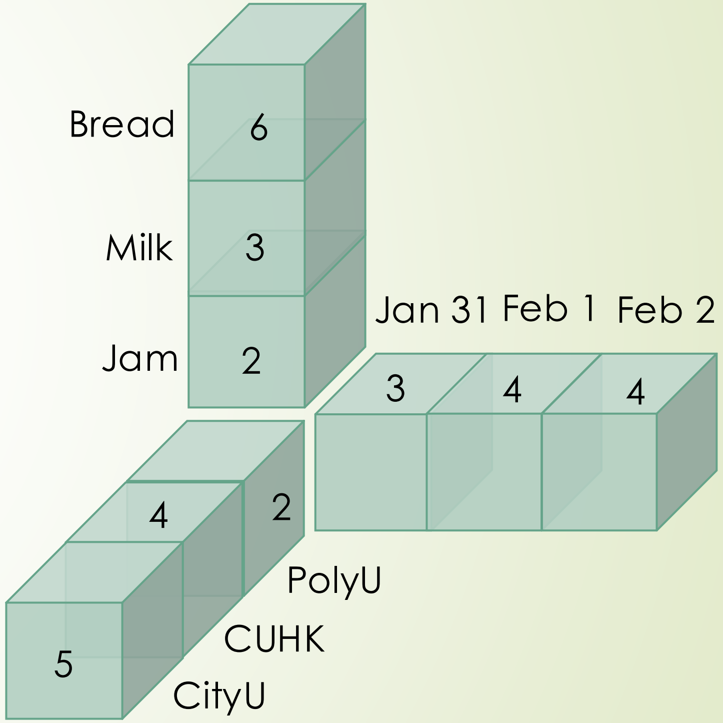
Time Location Item

Time, location Time, item Location, item

Time, location, item

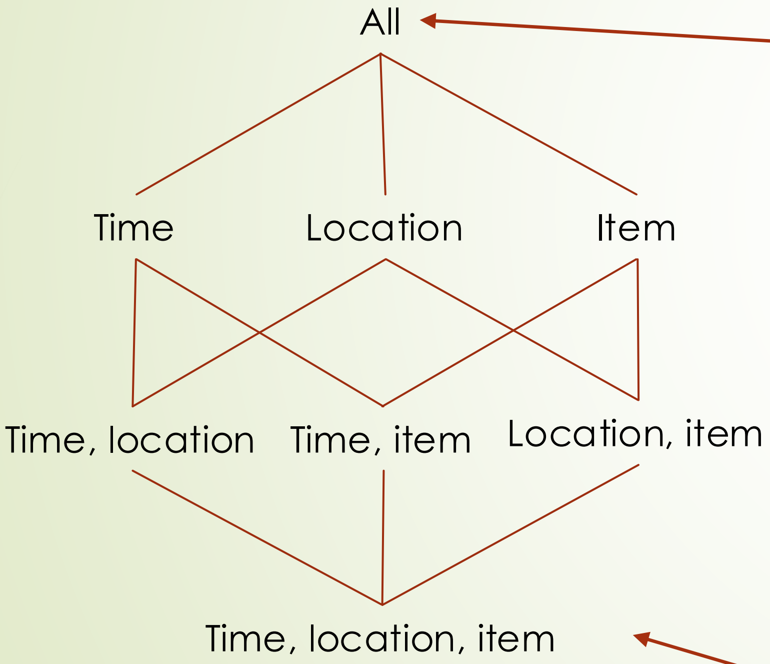


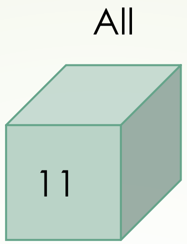
0-D cuboid



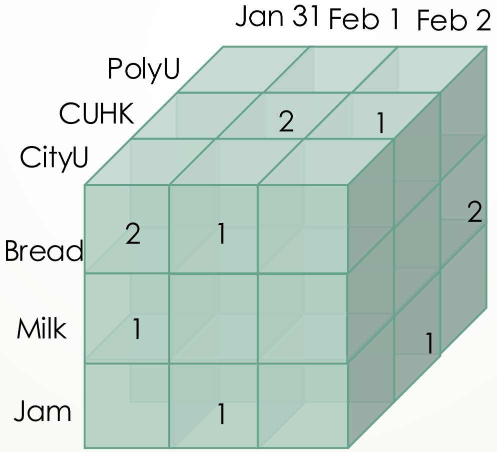
1-D cuboids

Lattice structure

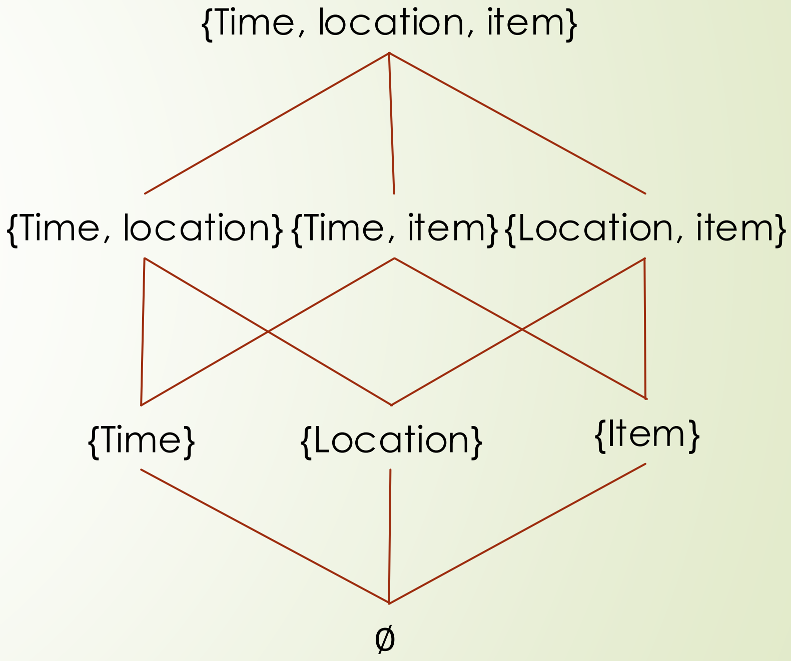




A____cuboid:
Highest level of summarization

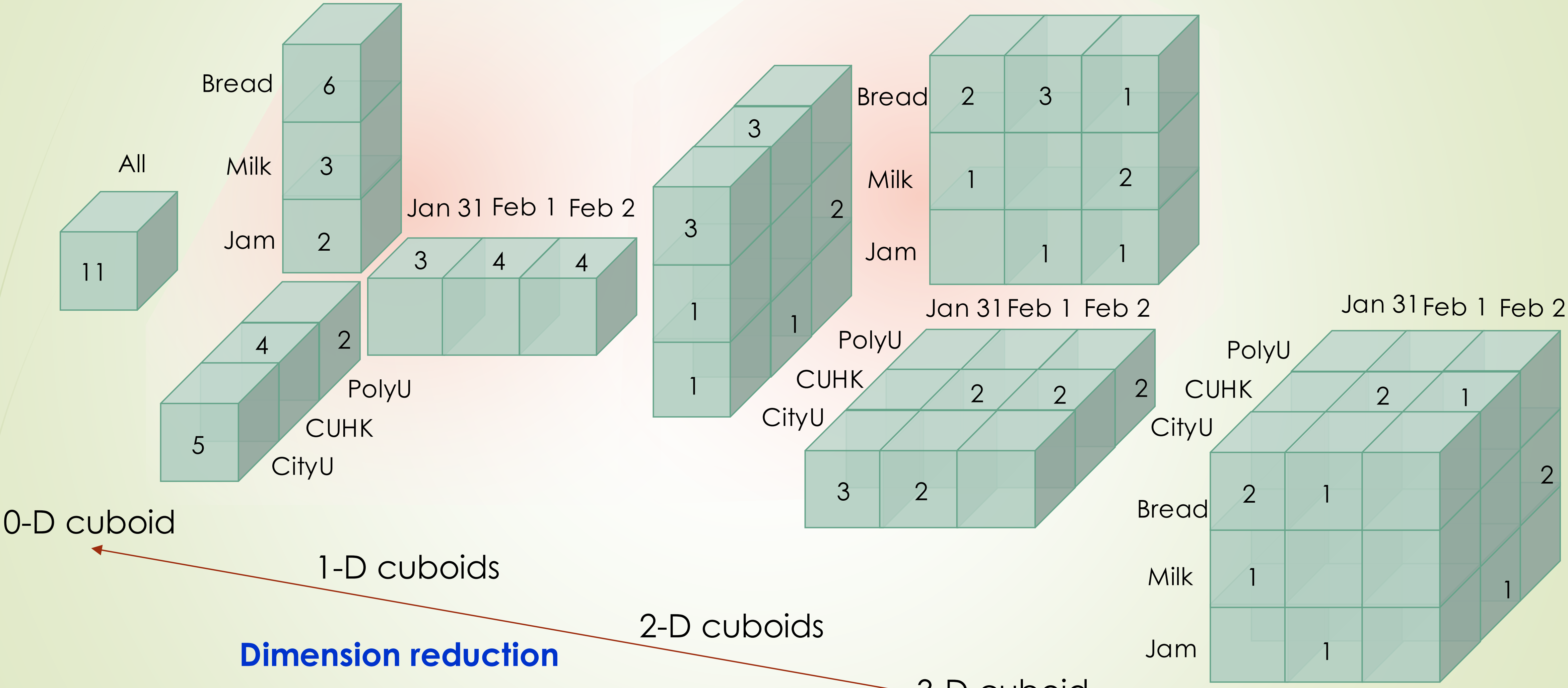


B____ cuboid:
Lowest level of summarization



Hasse diagram for
the Boolean lattice

Data Cube: The lattice of all cuboids



0-D cuboid

1-D cuboids

2-D cuboids

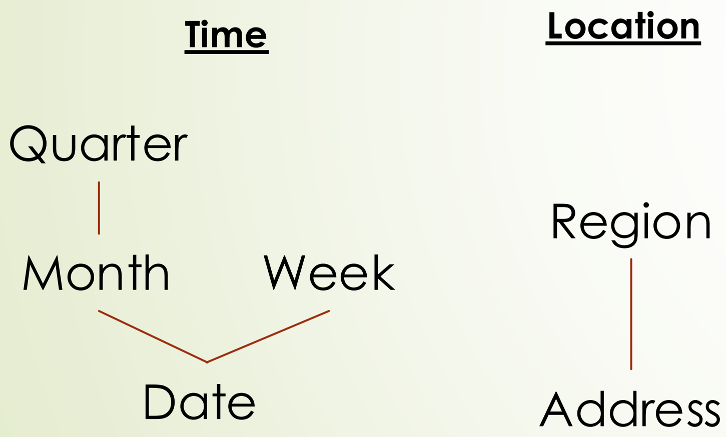
3-D cuboid

Dimension reduction

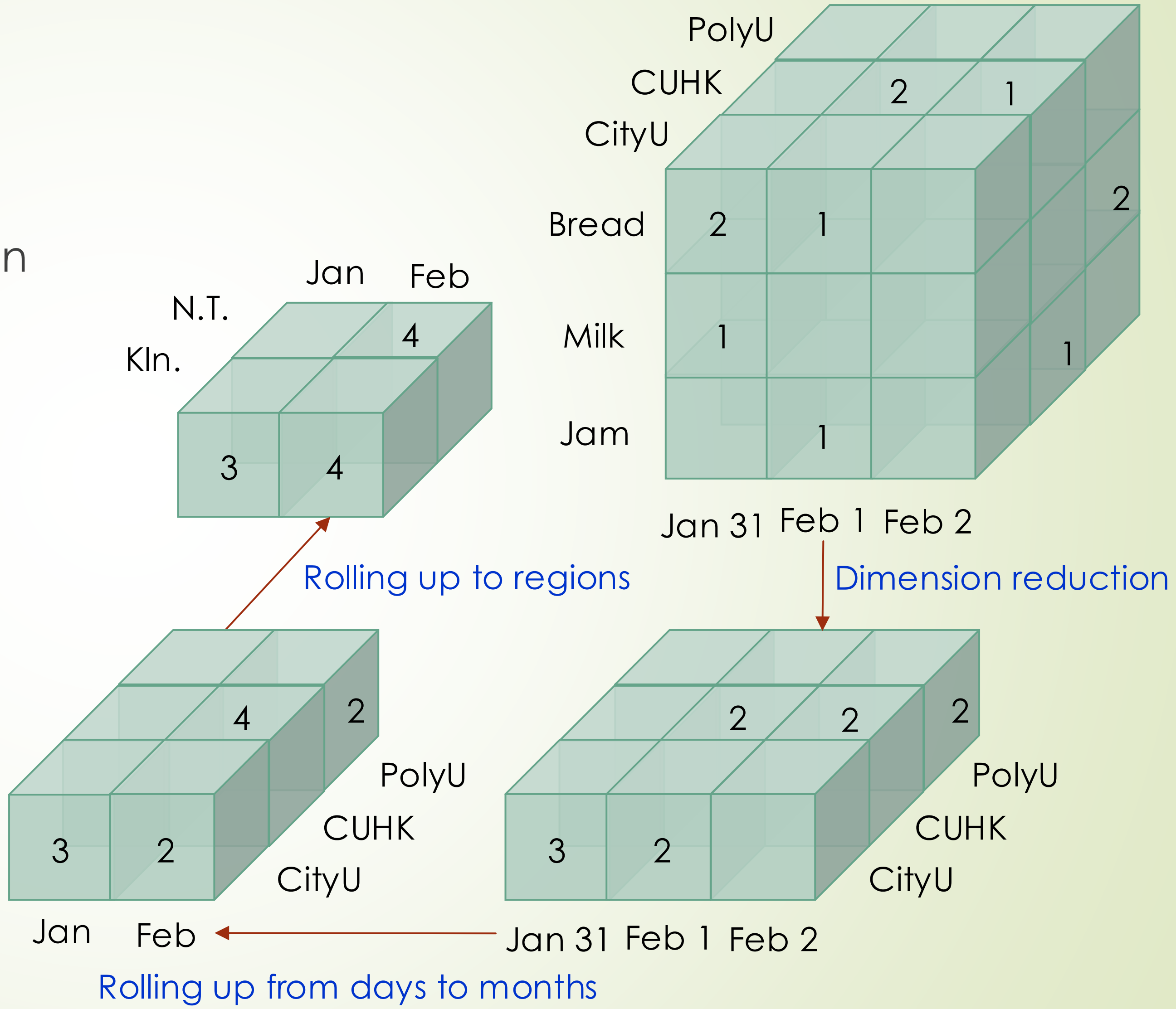
Any other operations of interest?

Roll-up operations

- How many units of items was sold in different months and regions?

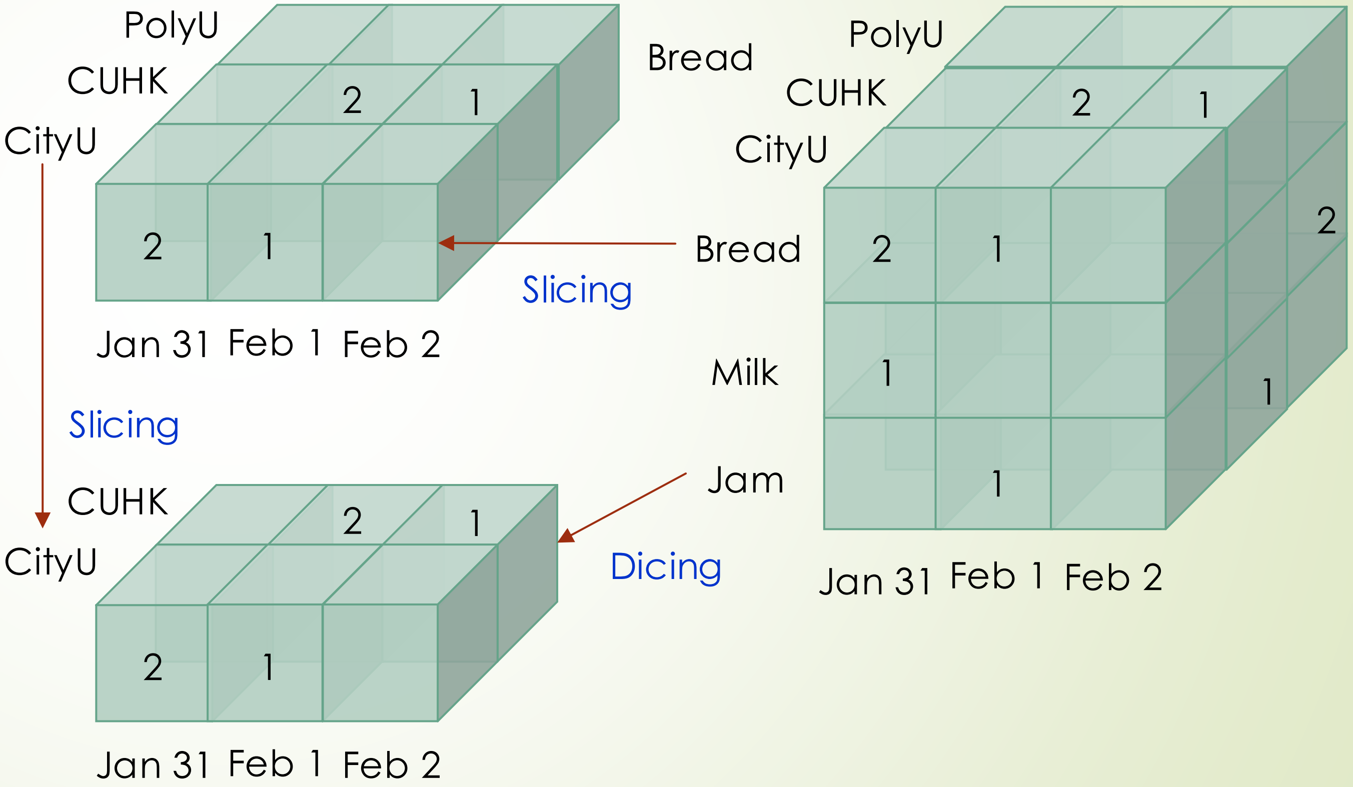


- **C__________ hierarchies**



Selection operations

- How many units of bread was sold in CityU or CUHK on different days?
- **Slice**:
  - Selection on 1 dimension.
- **Dice**:
  - Selection on multiple dimensions and/or with multiple values.



OLAP operations

- **O______ a________________ processing**
  - Roll-up
    - Dimension reduction
    - Climbing up a concept hierarchy
    - Reverse operation: **Drill-down**
  - Slice and Dice
    - Selection on one or more dimensions
  - Others (optional): **Pivot**/rotate, **drill-across**, **drill-through**
- In contrast with **o_____ t_____________ processing (OLTP)**: insert, update, delete.

The curse of dimensionality

- With d dimensions, how many cuboids can be obtained by dimension reduction? ____________
- If dimension i has L_i levels of concepts, how many cuboids can be obtained by the roll-up operation? ____________________
- How to store and compute the data cube?
  - **F_______ materialization**: Compute and store all the cuboids.
  - **N_______ materialization**: Store only the base cuboid and compute other cuboids on the fly.
  - **P_______ materialization**: Compute and store some parts of the data cube.
- How to store the base cuboid for efficient computation of other cuboids?

References

- 4.2 Data Warehouse Modeling: Data Cube and OLAP
- Optional
  - Hands-on tutorial of ETL using Pentaho