CONSTRUCTOR UNIVERSITY

OLAP Databases

Aalborg University,
adapted from Torben Bach Pedersen,
Man Lung Yiu and Dimitra Vista

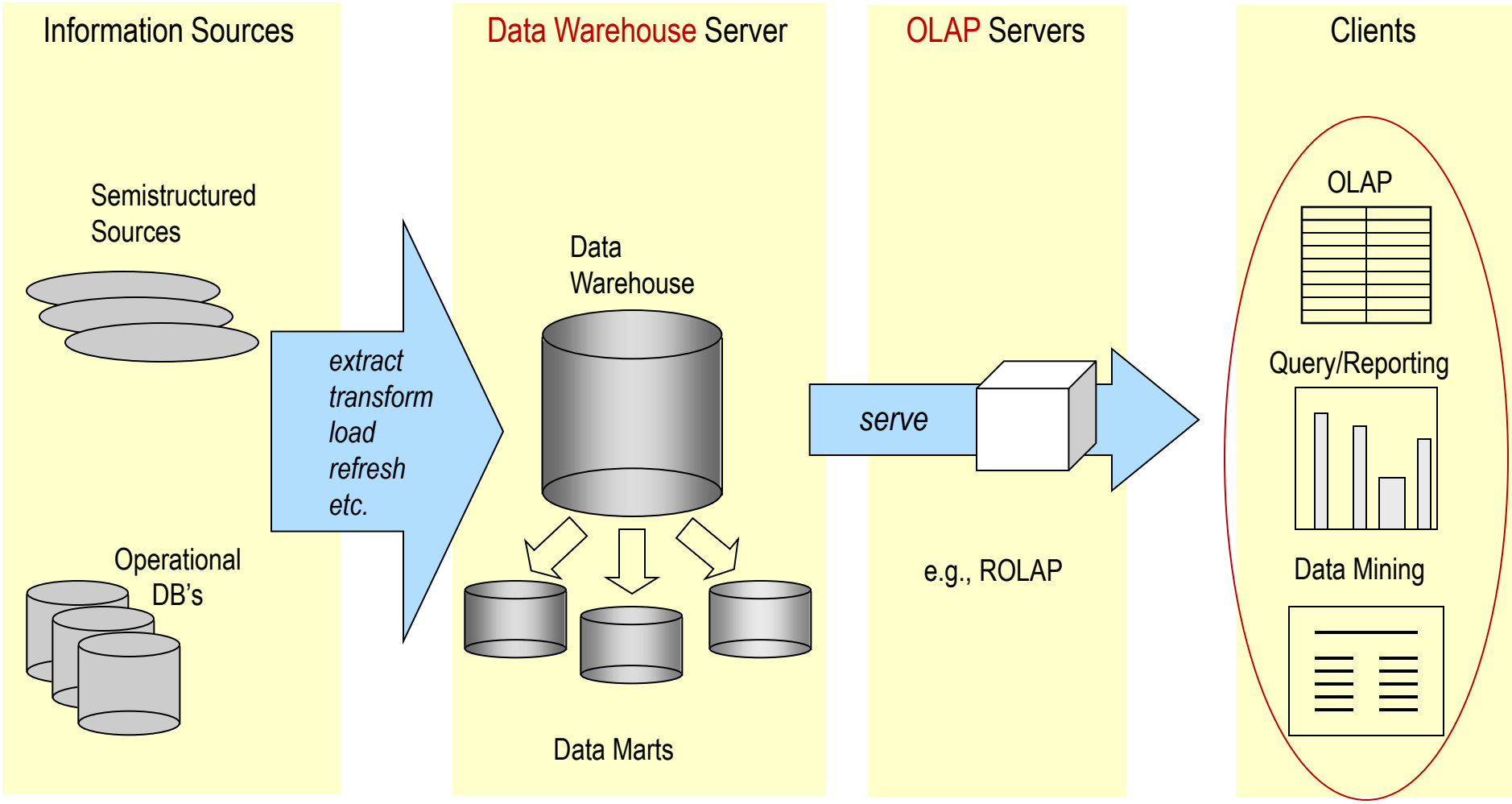
Overview

- Data Warehousing & Decision Support
- Datacubes, Dimension Hierarchies
- ROLAP & MOLAP
- ETL
- Summary

Decision Support Systems (DSS)

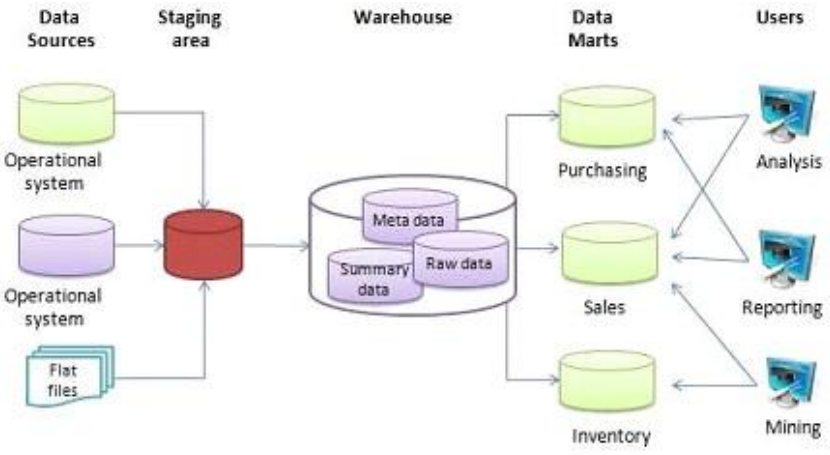
- Support business decisions
  - OLAP vs OLTP
- Examples of high-level analytical questions:
  - *What products have been most profitable for the company this year?*
  - *Is it the same group of products that were most profitable last year?*
  - *How is the company doing this quarter versus this same quarter last year?*
- Examples of data used for making decisions
  - Retail sales transaction details
  - Customer profiles (income, age, sex, etc.)
  - logs

DSS: Architecture



Data Warehousing: Informal

- Problem: critical enterprise information disparate, unavailable
  - locations, representations, storage, accessibility, completeness, ...
- Data Warehouse
  = system for reporting & data analysis
  - one or more disparate sources
    → central, integrated repository
  - current + historical data
  - creating analytical reports
    →core component of business intelligence



[soha jamil / Wikipedia]

- data cleansing: extract, transform, load (ETL)

Data Warehousing: Definition

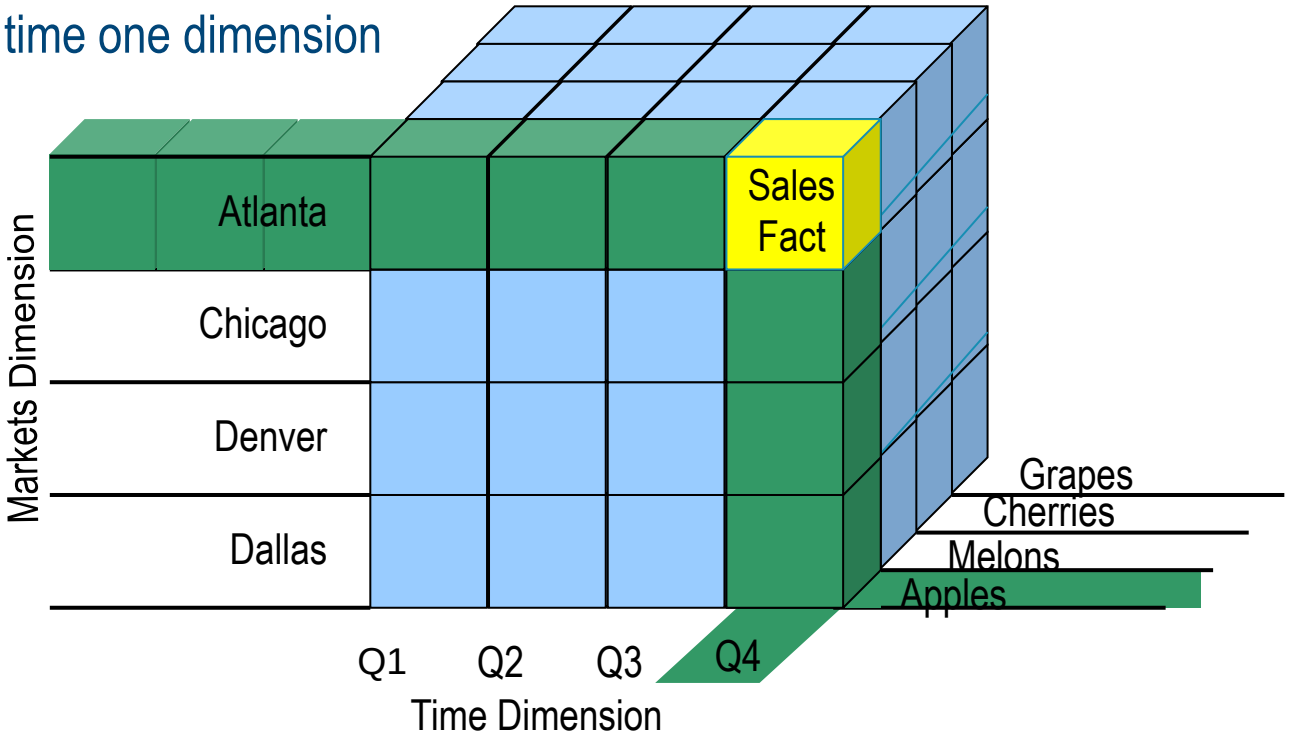
- "A warehouse is a subject oriented, integrated, time-variant, and non-volatile collection of data in support of management decision making process"
  - Bill Inmon, 1990
- Key features:
  - Subject Oriented: particular subject instead of company ongoing operations
  - Integrated: gathered from a variety of sources, merged into a coherent whole
  - Time Variant: particular time period
  - Non-Volatile: data, never removed

OLAP

- OLAP = Online Analytical Processing
  - Edgar Codd, 1994
  - Differentiated against OLTP = Online Transaction Processing
- software category motivated by industry, introducing advanced data analysis
  - decision making, business modeling, operations research, …

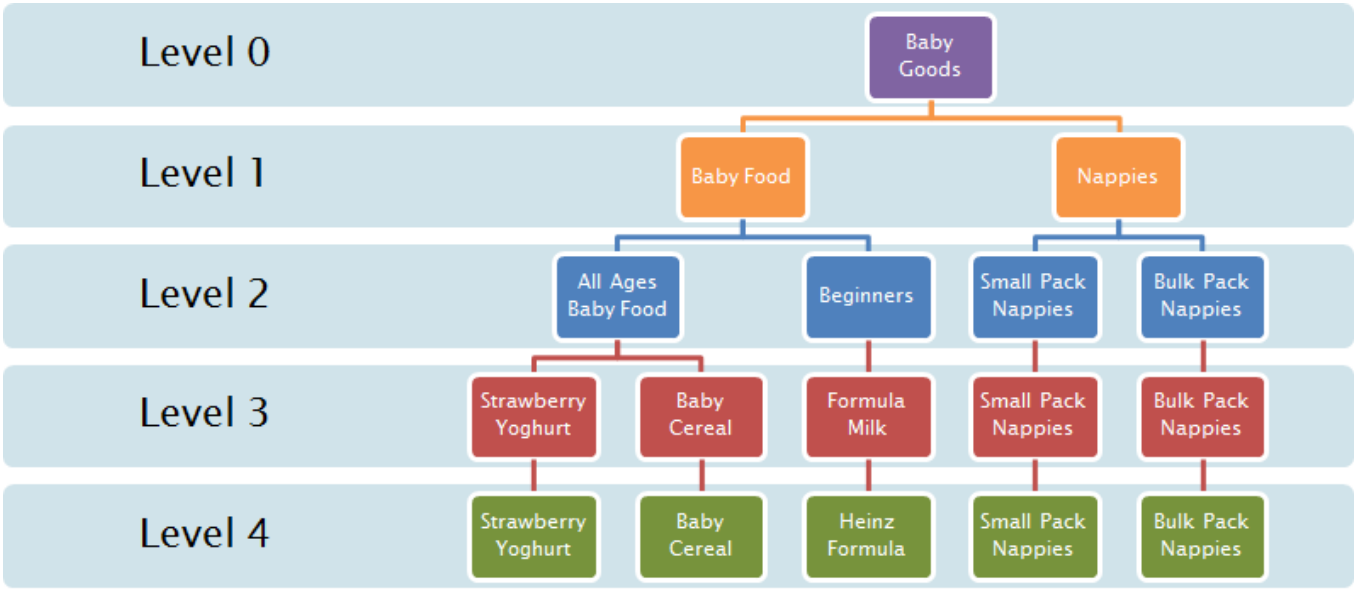
Datacubes

- Data structure for fast analysis along different views („dimensions“), on all levels of detail
  - Technically: multi-dimensional array + metadata
  - Typically, time one dimension



Dimension Hierarchies

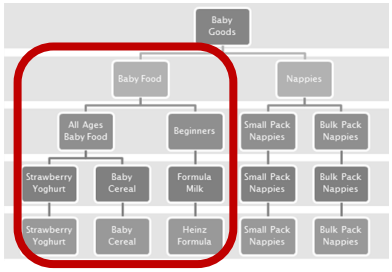
- Dimension enumerates values along an axis
  - Ex: time (predefined, ordered), product (custom, unordered)
- Dimension hierarchy = generalization levels of a dimension
  - „zoom levels“ into datacube



[OBIEE]

Dimension Hierarchies

- Dimension enumerates values along an axis
  - Ex: time (predefined, ordered), product (custom, unordered)
- Dimension hierarchy = generalization levels of a dimension
  - „zoom levels“ into datacube
  - Roll-up done based on hierarchies
- Strict nesting: Lower bins roll up neatly into higher bins
  - Not always strict! ex: week vs month



location:

city ⟵ region ⟵ country

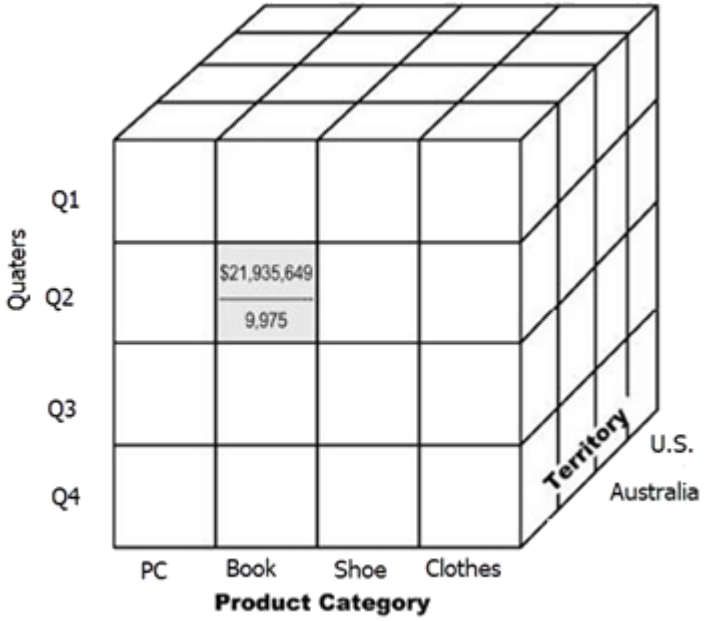
time:

day ⟵ month ⟵ year

day ⟵ week ⟵ year

Datacube Operations

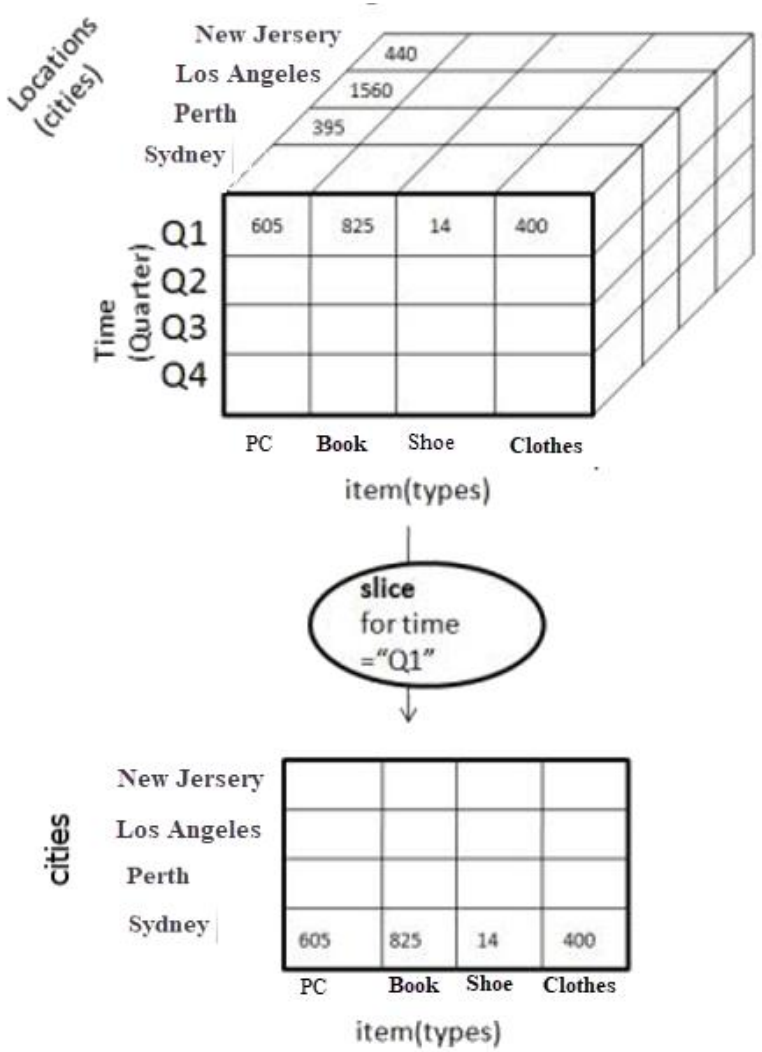
- Extraction + aggregation + combinations:
  - Slice
  - Dice
  - Roll-Up
  - Drill Down
  - Pivot
- Later, with arrays,
  we will want to do more



[guru99.com]

Operations: Slicing

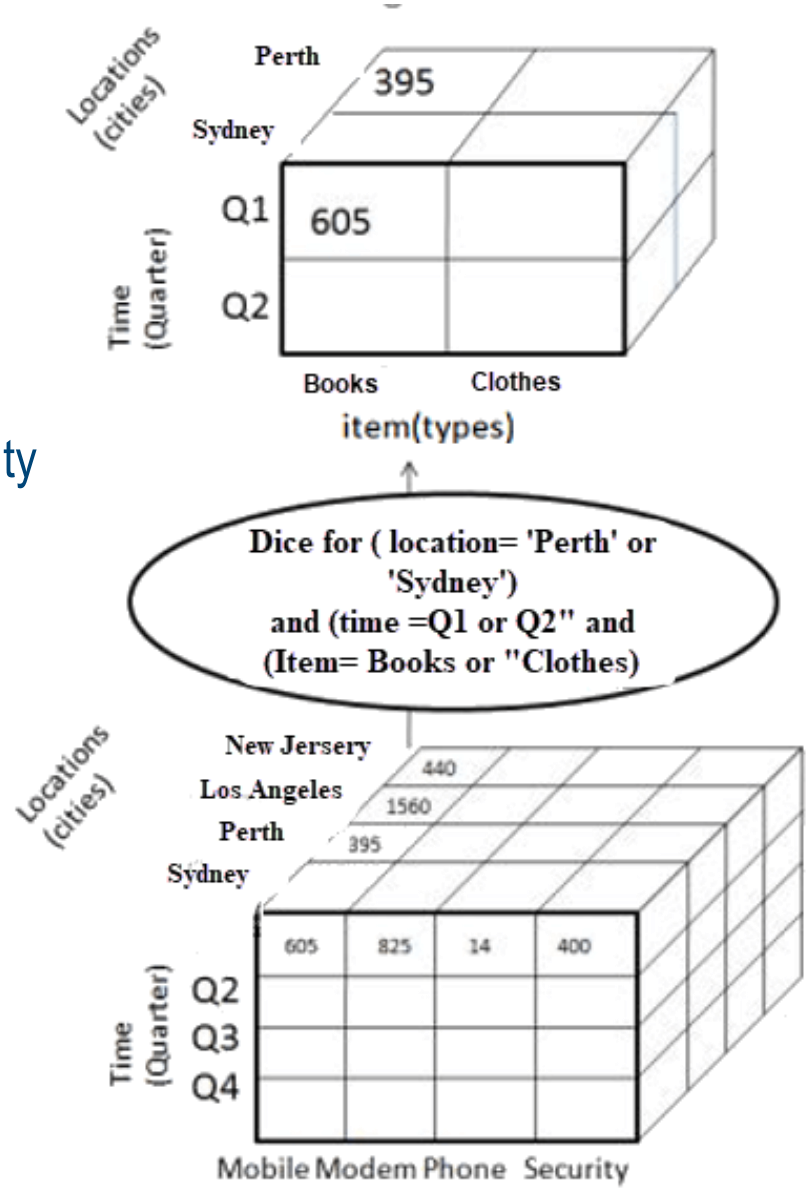
- Slicing = Select sub-cube by selecting dimension values to fewer points
  - Result cube has less dimensions
- Ex: select particular time slice



[guru99.com]

Operations: Dicing

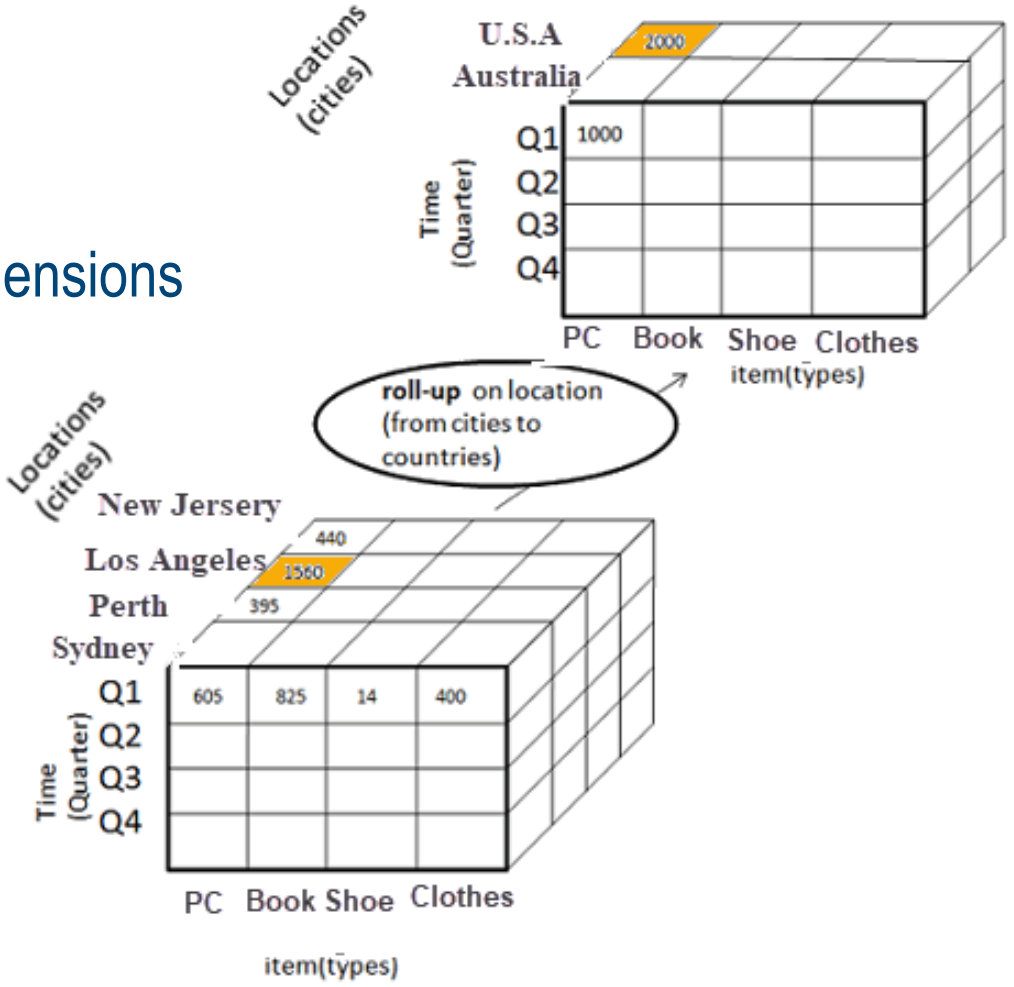
- Dicing = subsetting
  - „thicker slices“, not reducing dimensionality
- Ex: derive subcube
  by selecting along location, time, item
  simultaneously



[guru99.com]

Operations: Roll-Up

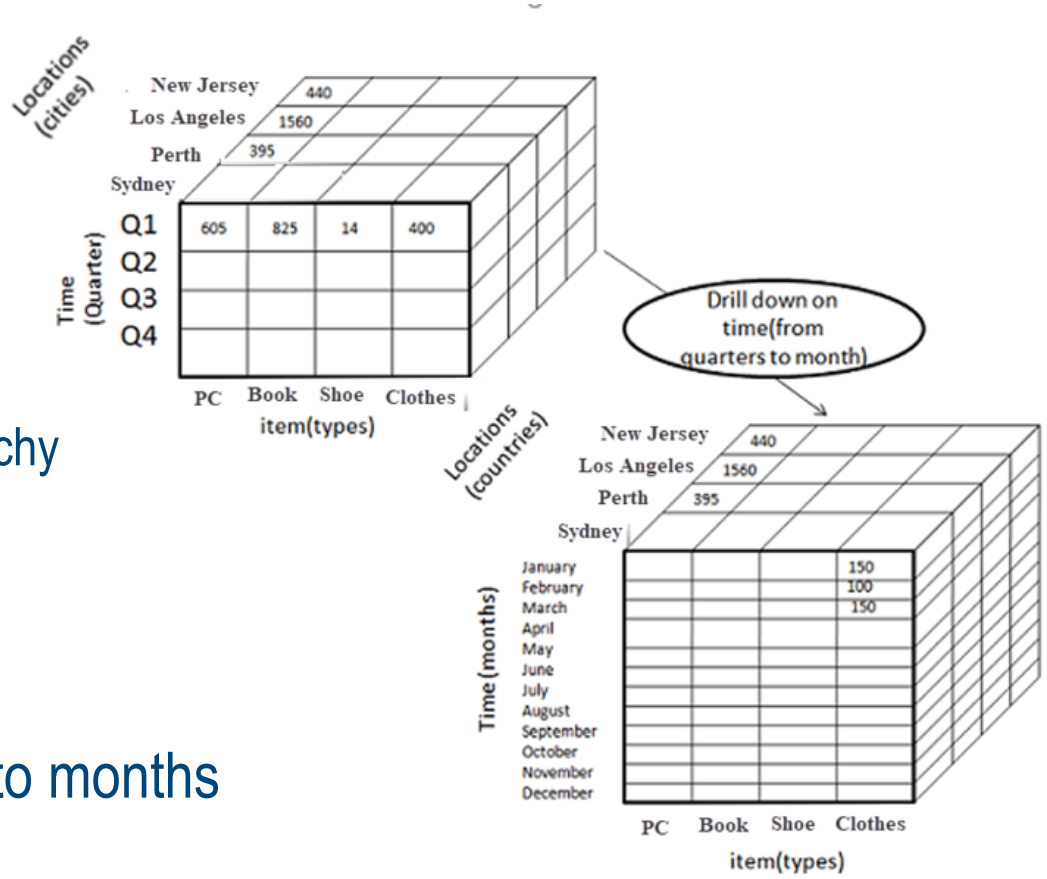
- Roll-Up = aggregation along dimensions
  - also: „consolidation“
  - collapsing a dimension hierarchy
  - „climbing up“ concept hierarchy
- Ex: consolidating
  from cities to countries



[guru99.com]

Operations: Drill-Down

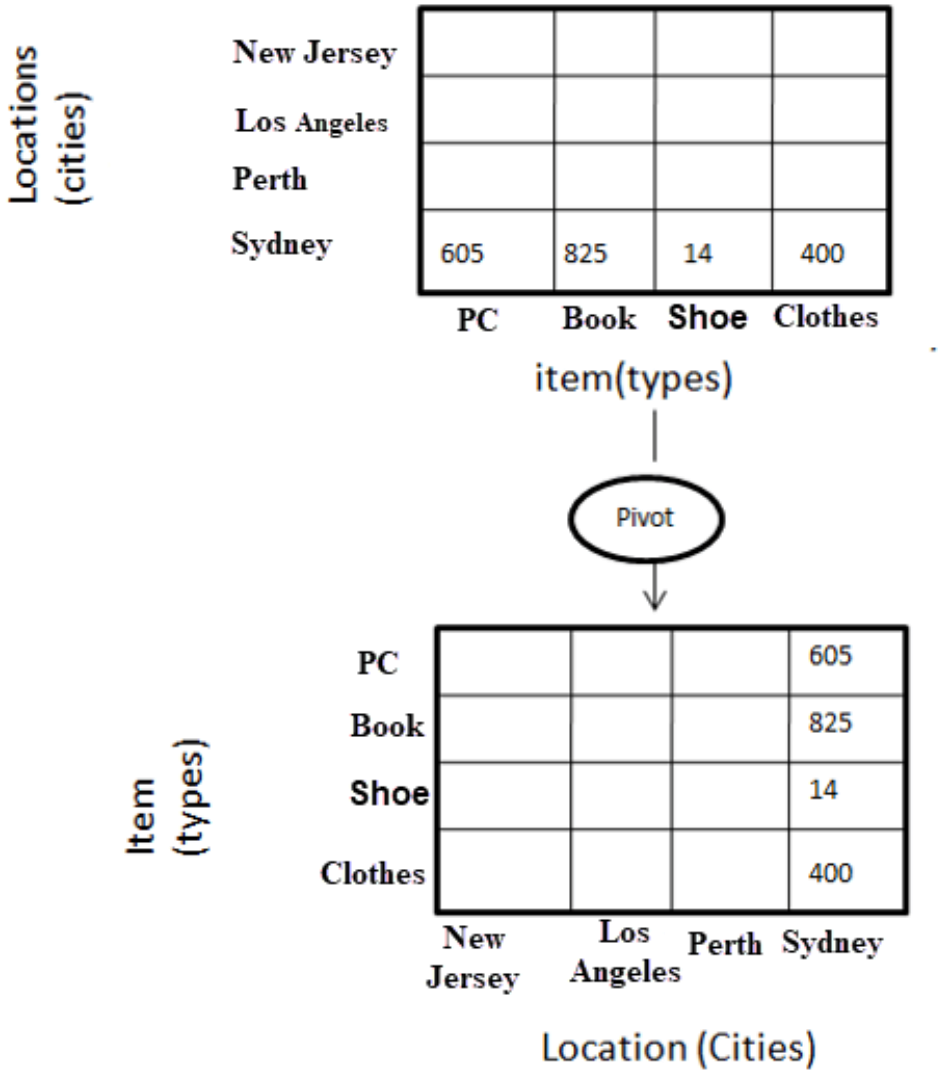
- Drill-Down = fragment data into smaller parts
  - Moving down concept hierarchy
  - Expanding some dimension
- Inverse of roll-up
- Ex: detailing from quarters to months



[guru99.com]

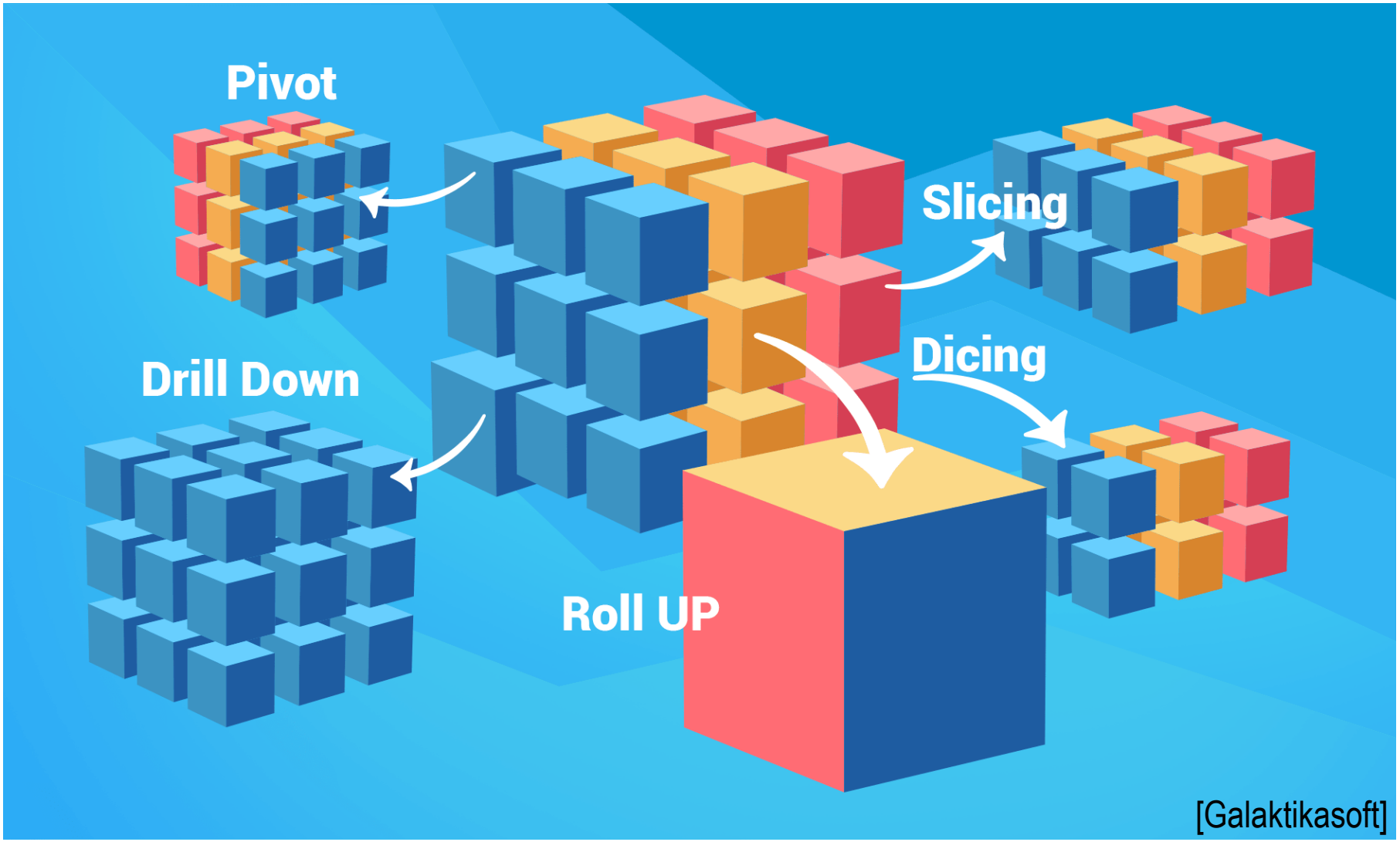
Operations: Pivot

- Pivot = rotate axes
  - show another view
  - Ex: swap rows & columns
- Ex: swap cities ↔ product types



[guru99.com]

Visual Summary: Datacube Ops



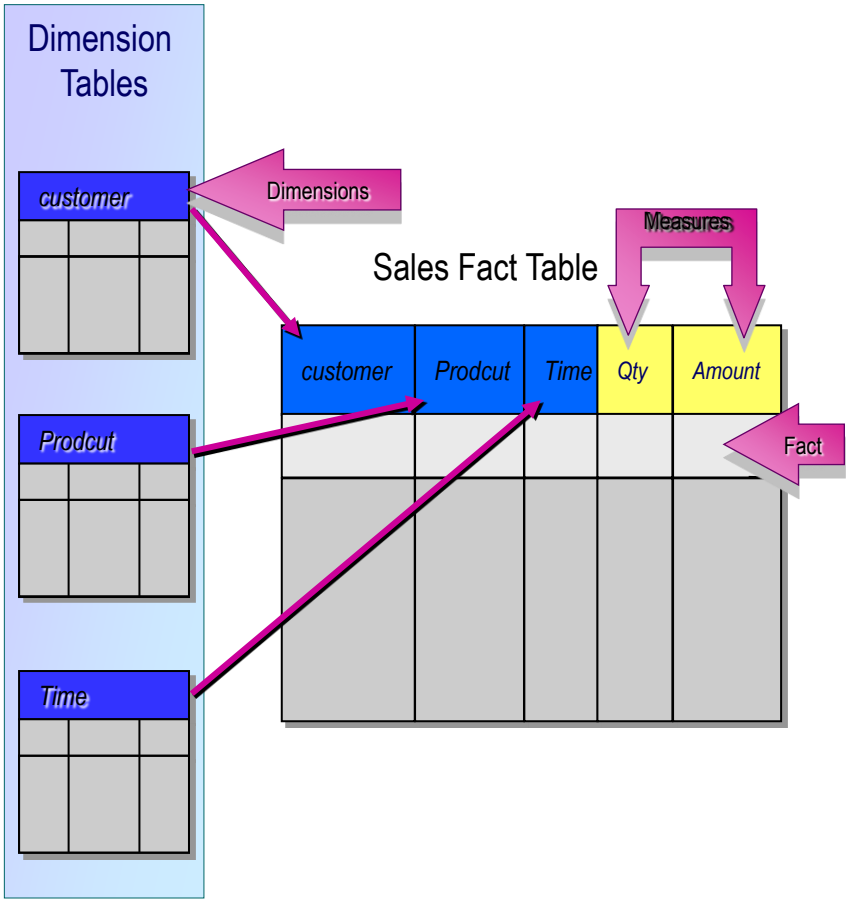
OLAP Datacube Querying

- ISO SQL does not directly support cubes
  - changing with SQL/MDA
- Multidimensional Expressions (MDX) = query language for OLAP
  - Microsoft 1997, also adopted by other vendors
    - https://docs.microsoft.com/en-us/sql/mdx/multidimensional-expressions-mdx-reference?view=sql-analysis-services-2017
  - Ex:

```
SELECT
  { [Measures].[Store Sales] } ON COLUMNS,
  { [Date].[2002], [Date].[2003] } ON ROWS
FROM Sales
WHERE ( [Store].[USA].[CA] )
```

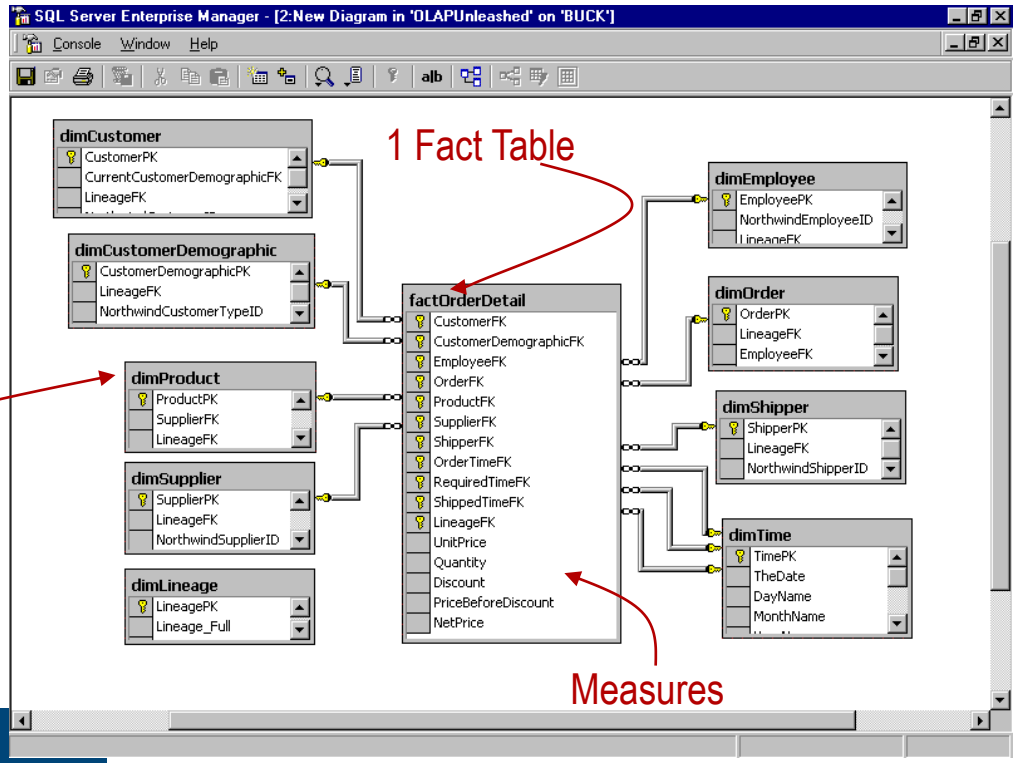
Datacubes in ROLAP: Facts & Dimensions

- Mapping datacubes to relational table schema ?
- Central fact table = tuples + n-D "coordinate" attributes
  - foreign keys
  - non-keys = measure
- Dimension = table(s)
  - with coordinates + descriptions („metadata")
- One step of normalization: keys → dimension tables



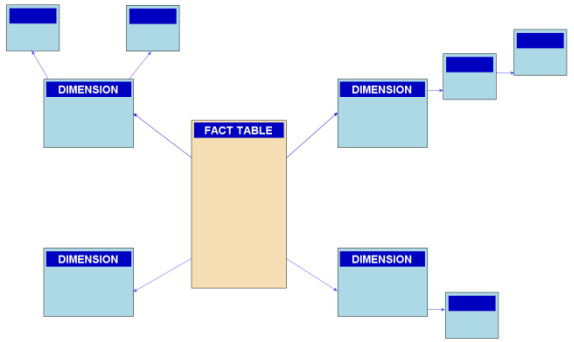
Star Schema

- star schema = multidimensional data structure in relational database
  - Dimension hierarchies = aka lookup tables around fact table
- MS SQL Server Enterprise Manager:

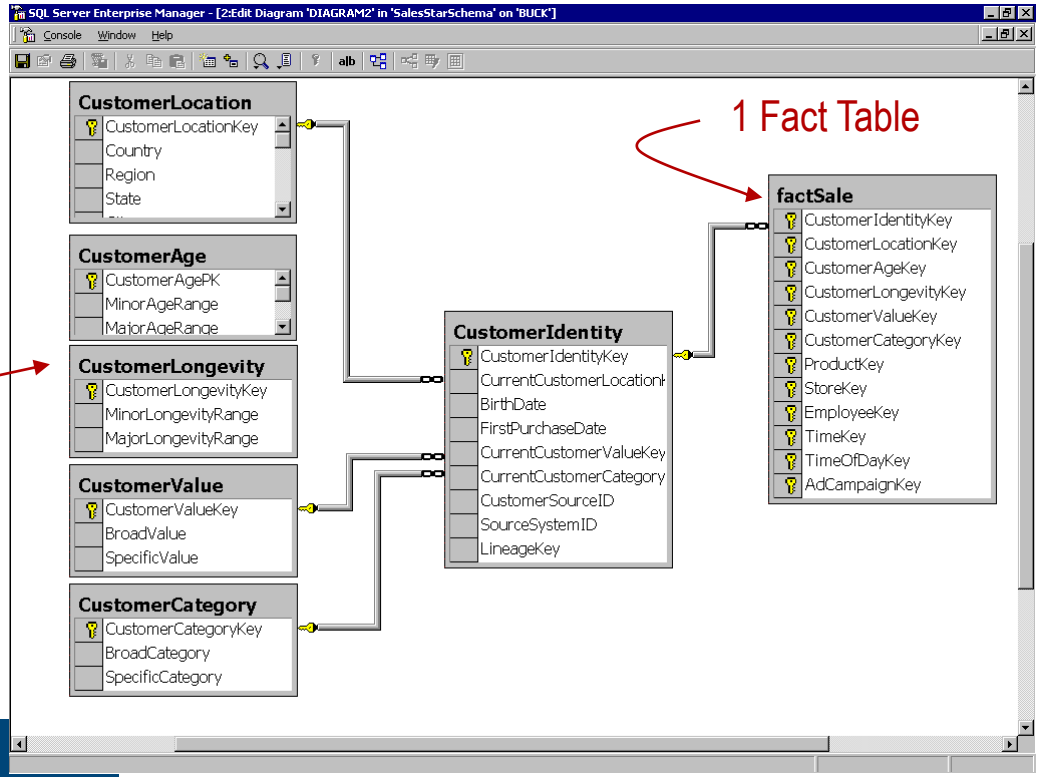


Snowflake Schema

- snowflake schema = refinement of star schema
  - Normalizing dimension tables
  - Ex:
    - *Year → Month → Day*
    - *Week → Day*
- MS SQL Server Enterprise Manager:



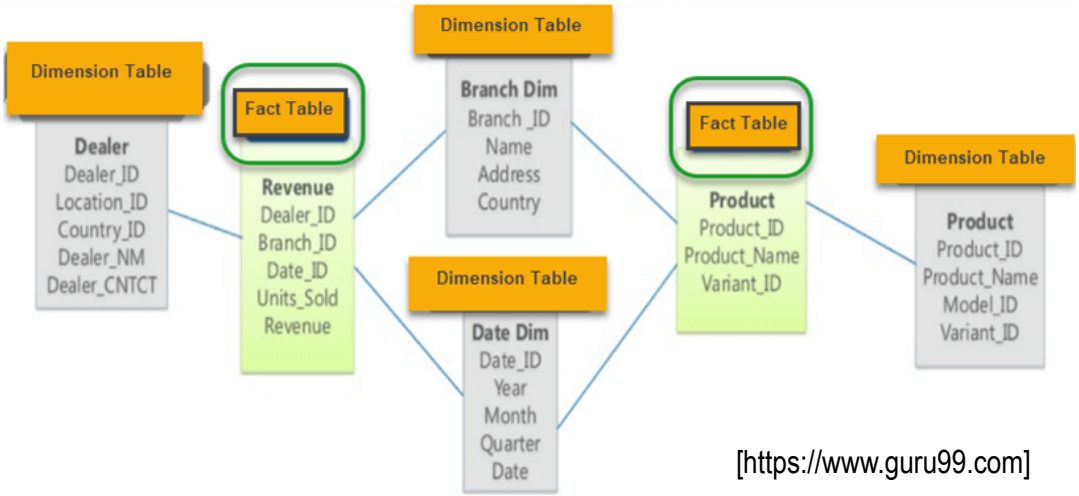
[SqlPac @ Wikipedia]



normalized Dimension Tables

Galaxy Schema

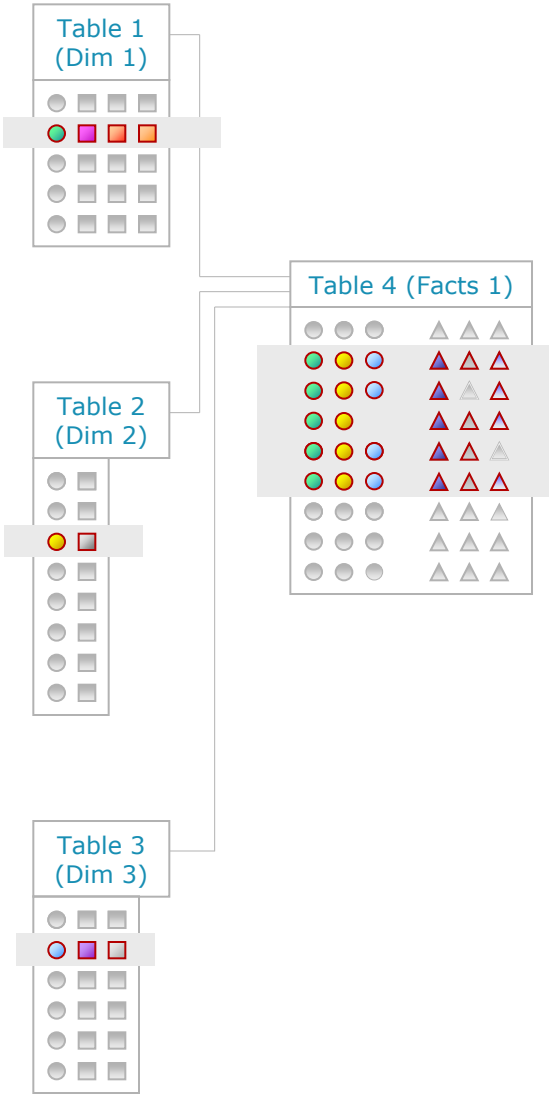
- Galaxy schema = Constellation schema = linked datacubes
  - Sharing dimension(s)



[https://www.guru99.com]

- helpful for aggregating fact tables

A Query in ROLAP



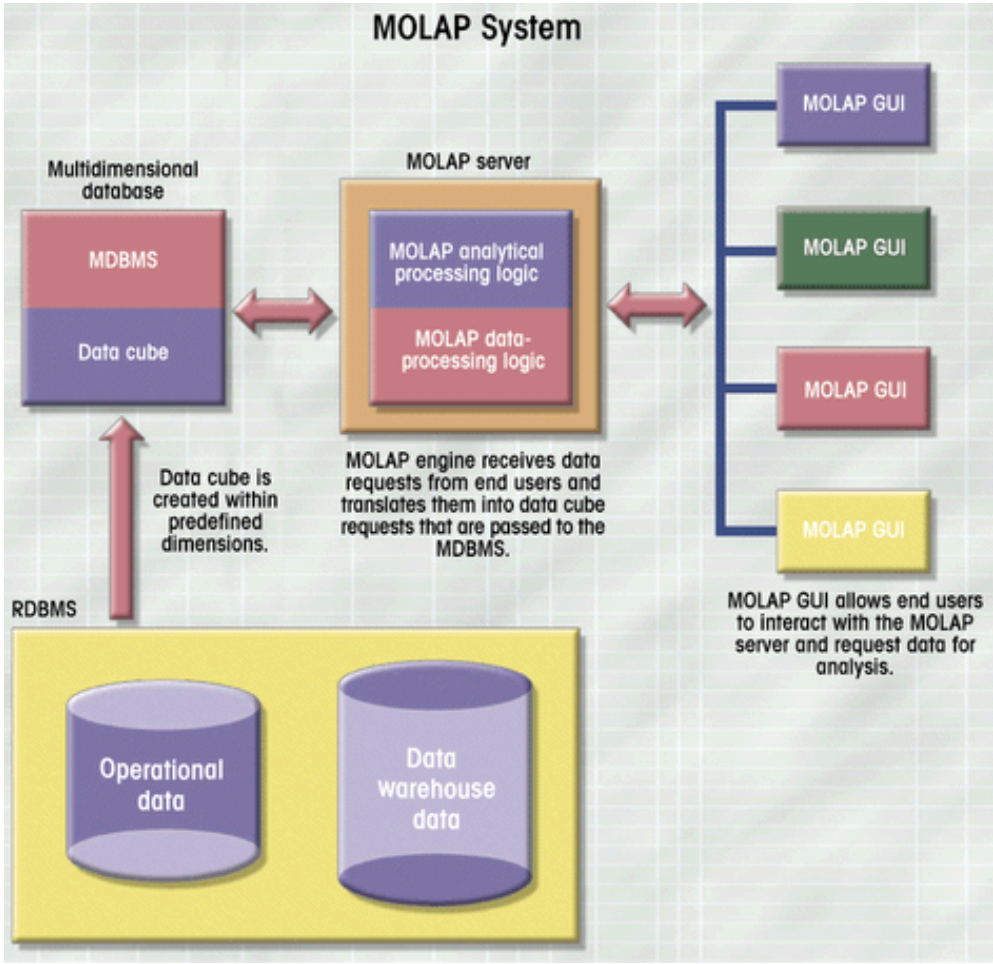

```
SELECT Fact_Column_1
      ,Fact_Column_2
  FROM Table 4     T4   -- Fact
      ,Table 1     T1   -- Dim 1
      ,Table 2     T2   -- Dim 2
      ,Table 3     T3   -- Dim 3
 WHERE T4.Dim_Col_1 = T1.Dim_Col_1
   AND T4.Dim_Col_2 = T2.Dim_Col_1
   AND T4.Dim_Col_3 = T3.Dim_Col_1
   AND T1.Dim_Property_2 = 'Product 1'
   AND T2.Dim_Property_1 = 'City 1'
   AND T3.Dim_Property_1 = 'Salesman 1'
```

Performance of ROLAP methods

- ~ 70% of the time spent on CPU, rest on I/O
- Most of the CPU time spent in sorting intermediate results
  - ~ 10-20% is spent on copying data
- I/O composed of read/write into large tables

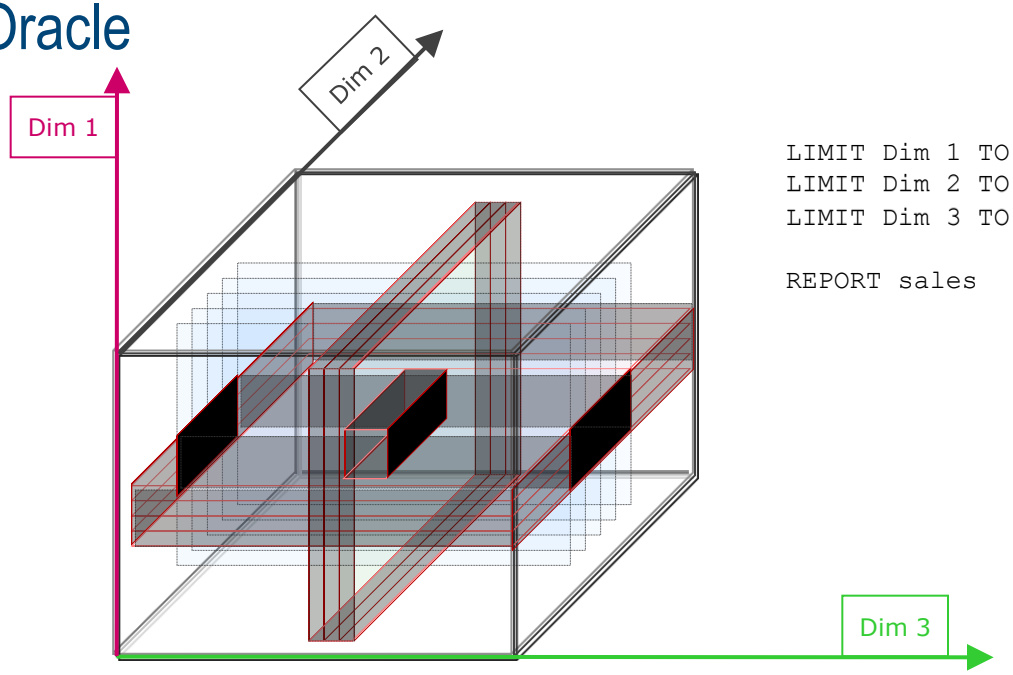
The MOLAP Approach

- Native datacube
  = multidimensional array
  - plus metadata
- Fast position-based computation
  - cell values stored in fixed positions determined by dimension values
- Often used for data marts



A Query in MOLAP

- Proprietary QLs
- Ex: Oracle

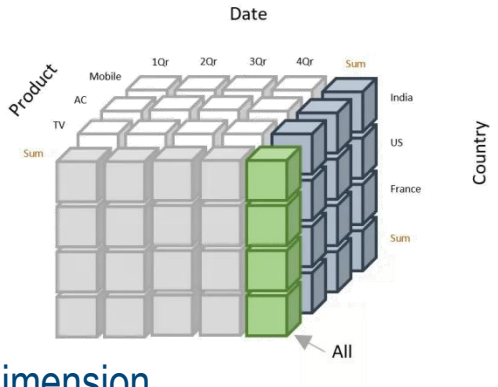



```
LIMIT Dim 1 TO 'Soap'
LIMIT Dim 2 TO 'Bremen'
LIMIT Dim 3 TO 'John Doe'

REPORT sales
```

Summary: Data Warehousing Terminology

- Typically, warehouse data is multidimensional, resembling large datacubes
- Datacube: a structure containing values of same type, each identified by a coordinate tuple along the dimensions of the cube
  - Ex: Cube of sales figures with 3 dimensions region, product, time



- Fact: The subject of analysis
  - Ex: sales revenues
- Measures: The coordinates of the specific data points, one taken from each dimension
  - Ex: (“Bremen”, “Apples”, “2025-12-01”) might identify the fact containing the revenues made from selling apples that day
- Dimension: a set of values providing coordinates that can act as measures for locating facts
  - Ex: region, product, time
- Hierarchies: Nested grouping of values within dimensions, used for aggregation (“roll-up”) etc.
  - Ex: time = day / week / year; region = Bremen / Germany / Europe