OLAP Databases

Instructor: Peter Baumann
email: p.baumann@jacobs-university.de
tel: -3178
office: room 88, Research 1

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

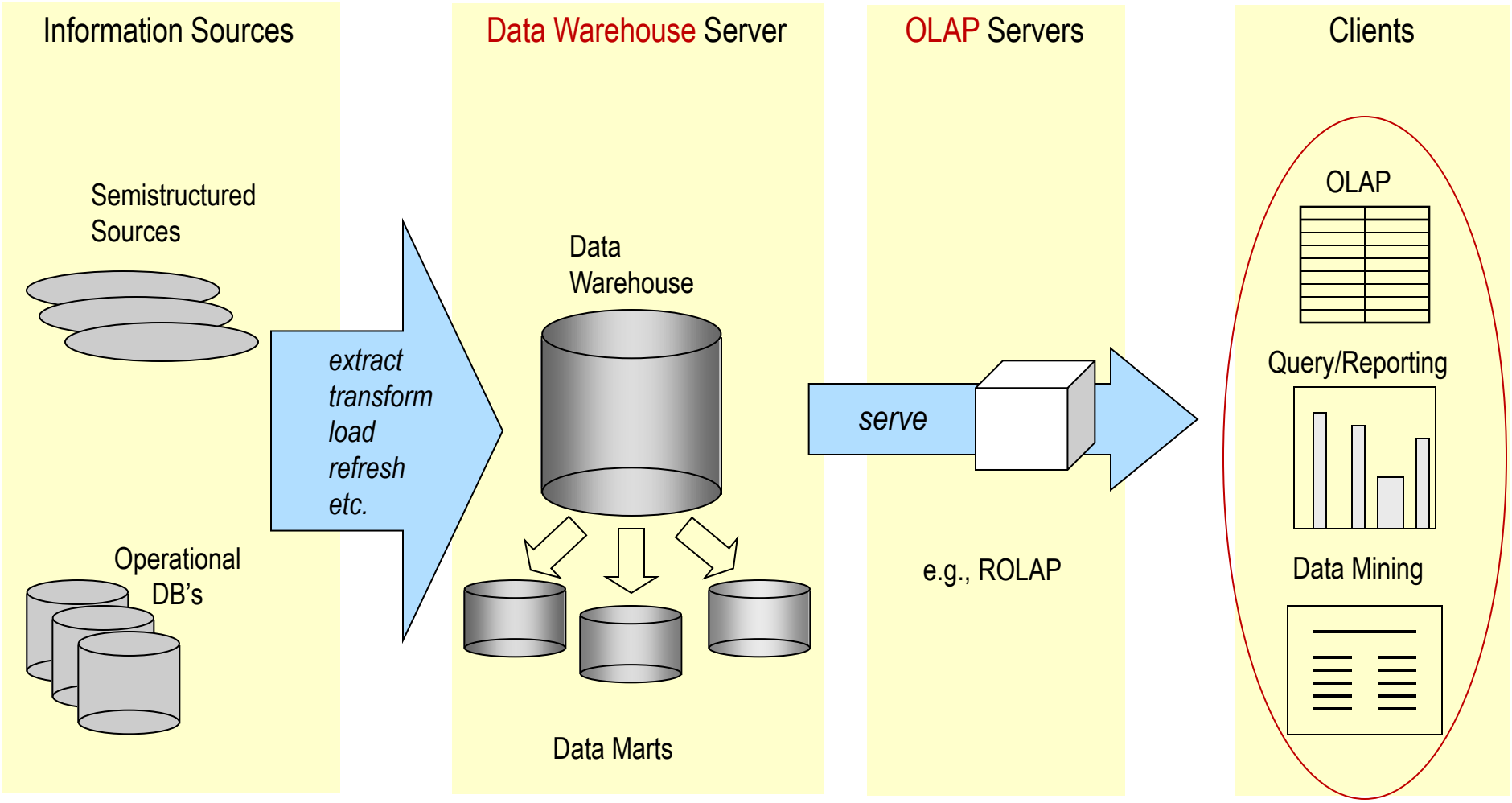
Overview

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

Desicion Support Systems (DSS)

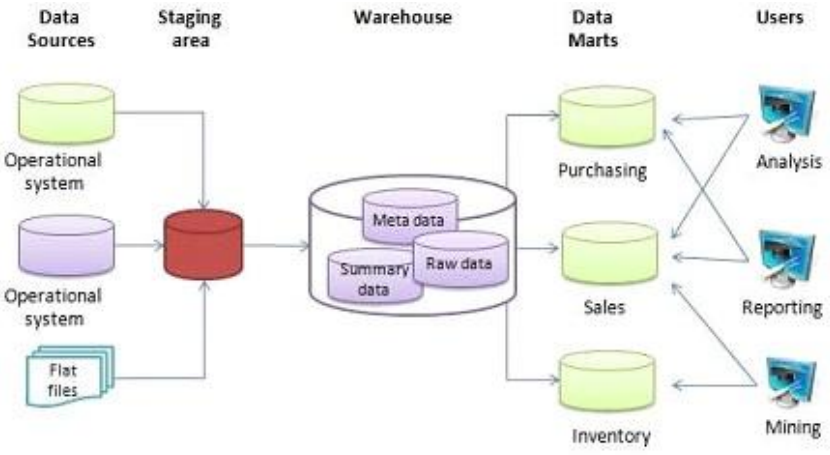
- Support business decisions
  - OLAP vs OLTP
- Examples of high-level analytical questions:
  - *What products have been most profitable for the company this year?*
  - *Is it the same group of products that were most profitable last year?*
  - *How is the company doing this quarter versus this same 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, transfer, 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

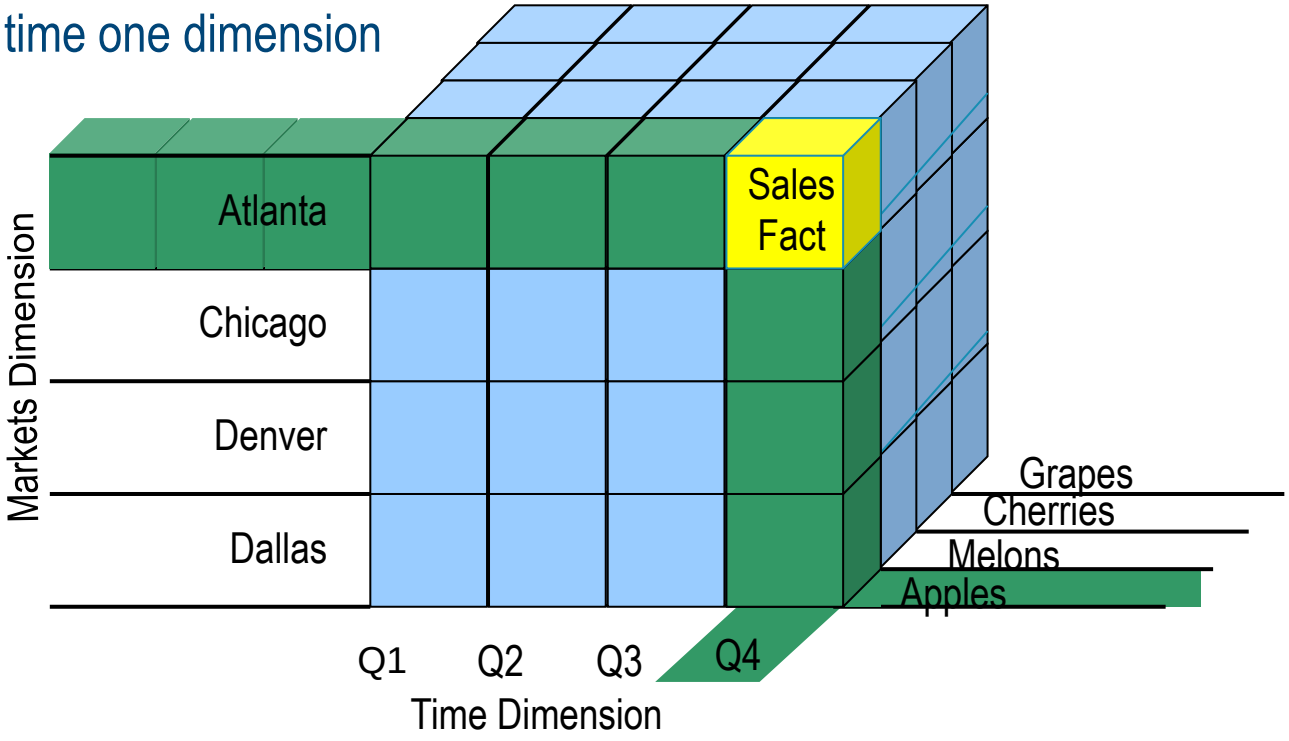
- OLAP = Online Analytical Processing
  - Edgar Codd, 1994
  - Differentiated against OLTP = Online Transaction Processing
- software category motivated by industry, introducing advanced data analysis
  - decision making, business modeling, operations research, …
- enables analysts to extract & view business data from different points of view
  - dimensions
- OLAP Characteristics
  - multidimensional data analysis techniques
  - Strong use of aggregate functions for summarizing large volumes of data
  - advanced database support
  - easy-to-use end-user interfaces (spreadsheet type)
  - client/server architecture

Overview

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

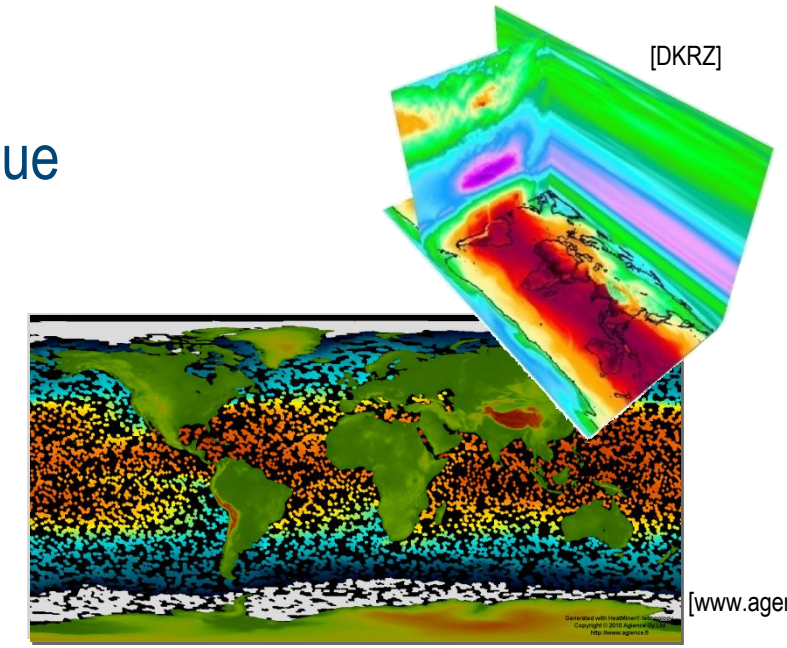
Datacubes

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



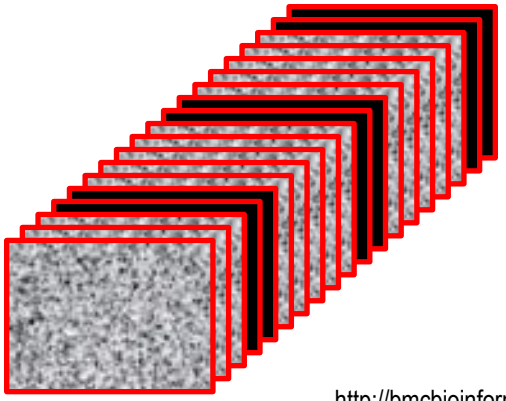
Dense vs Sparse Datacubes

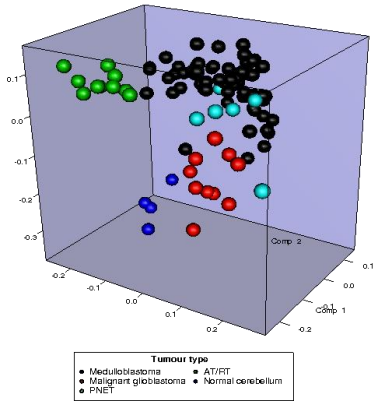
- Dense = every cell has meaningful value
  - Ex: climate simulation
- Sparse = some values null
  - Clustered data
  - Empty regions
    - *Ex: retail – open Mon thru Fri*



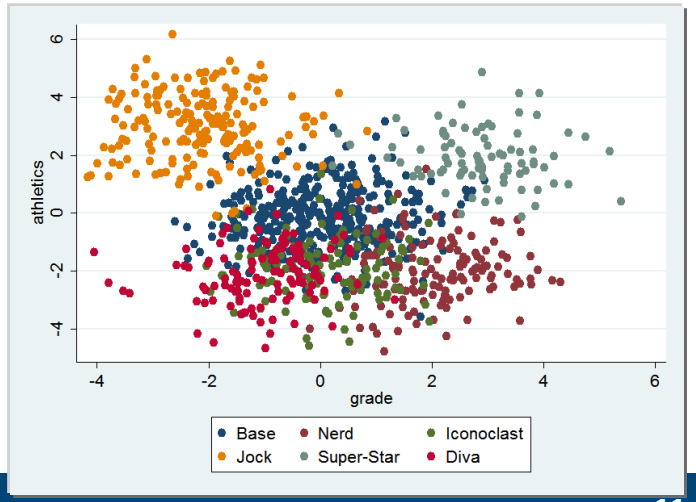
[DKRZ]

[www.agencie.fi]





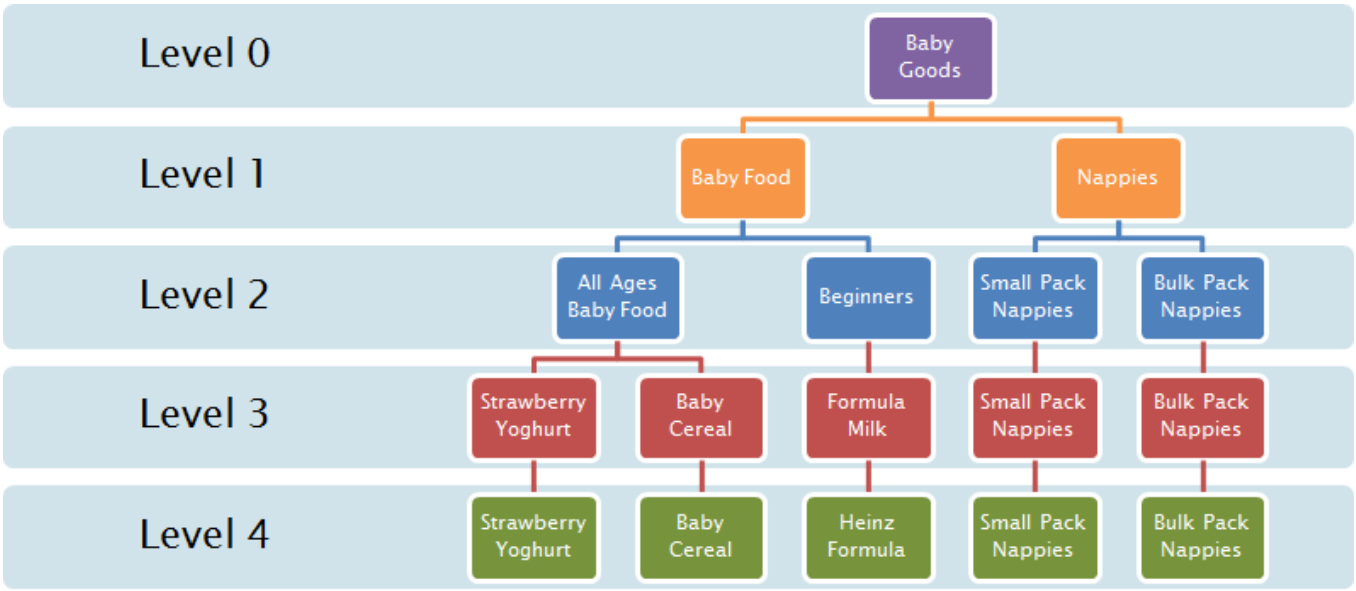
http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-253



http://lookfordiagnosis.com/mesh_info.php?term=cluster%20analysis&lang=1

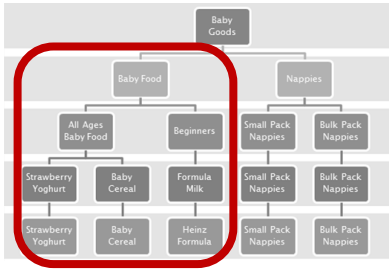
Dimension Hierarchies

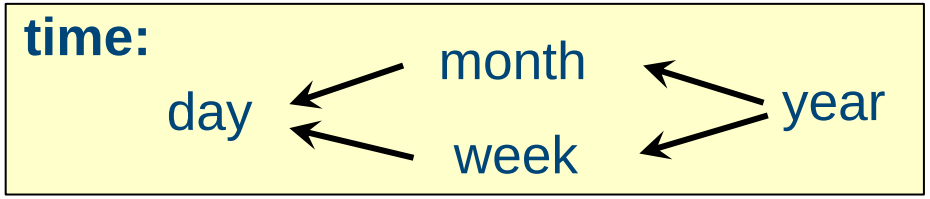
- Dimension enumerates values along an axis
  - Ex: time (predefined, ordered), product (custom, unordered)
- Dimension hierarchy = generalization levels of a dimension
  - „zoom levels“ into datacube



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

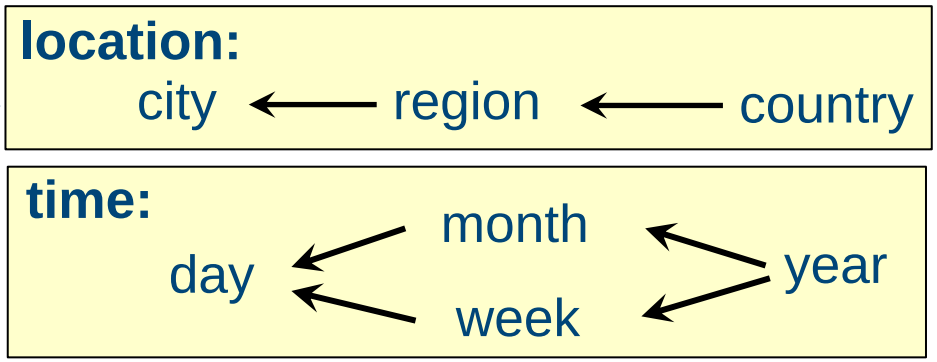




Datacubes

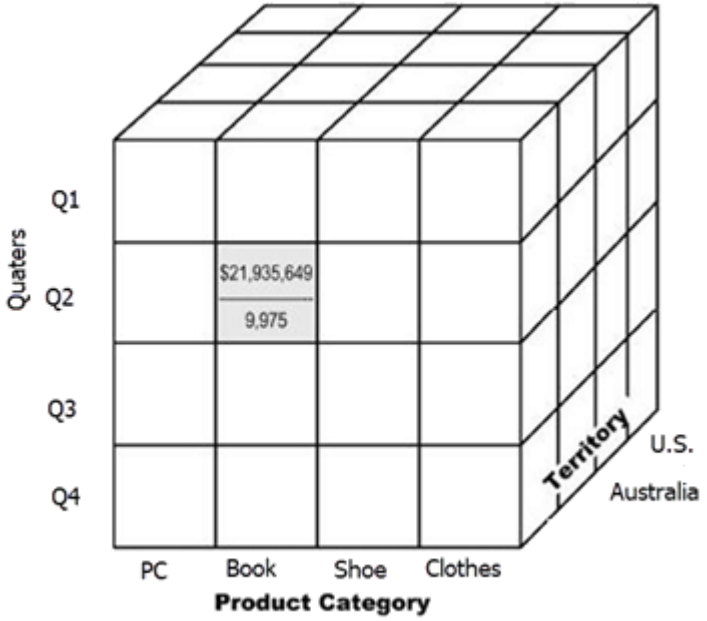
- Normalizing dimensions
  → dimension hierarchies

- Datacube = collection of fact &



Datacube Operations

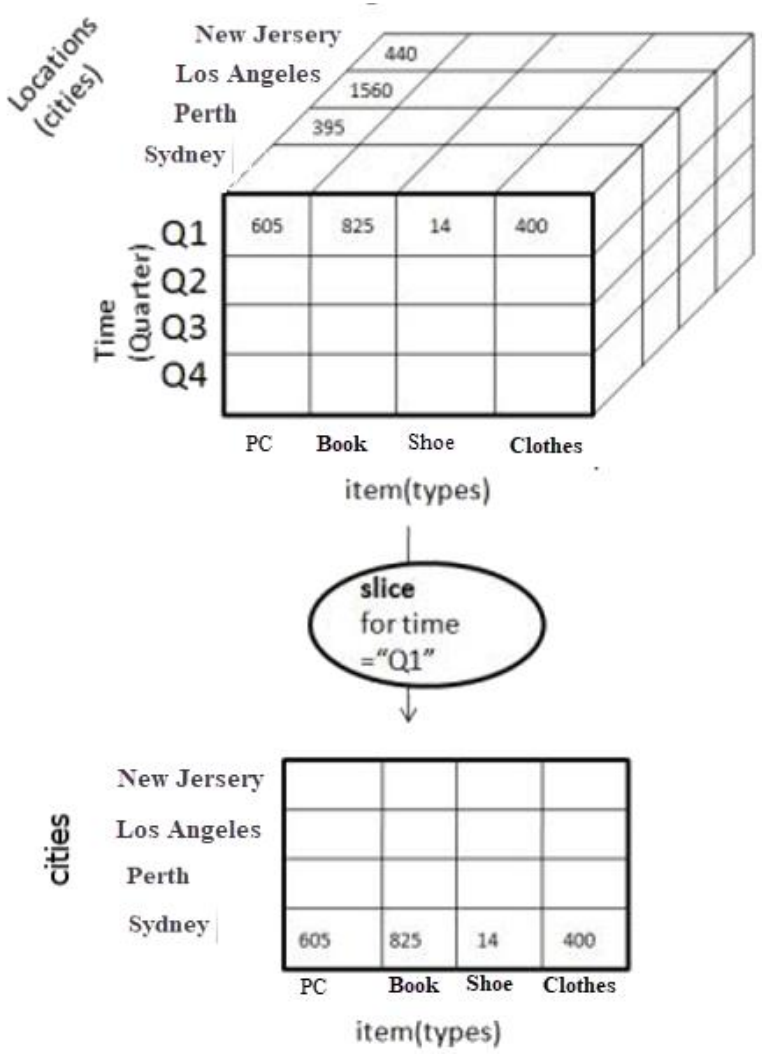
- Extraction + aggregation + combinations:
  - Slice
  - Dice
  - Roll-Up
  - Drill Down
  - Pivot
- Later, with arrays,
  we will want to do more



[guru99.com]

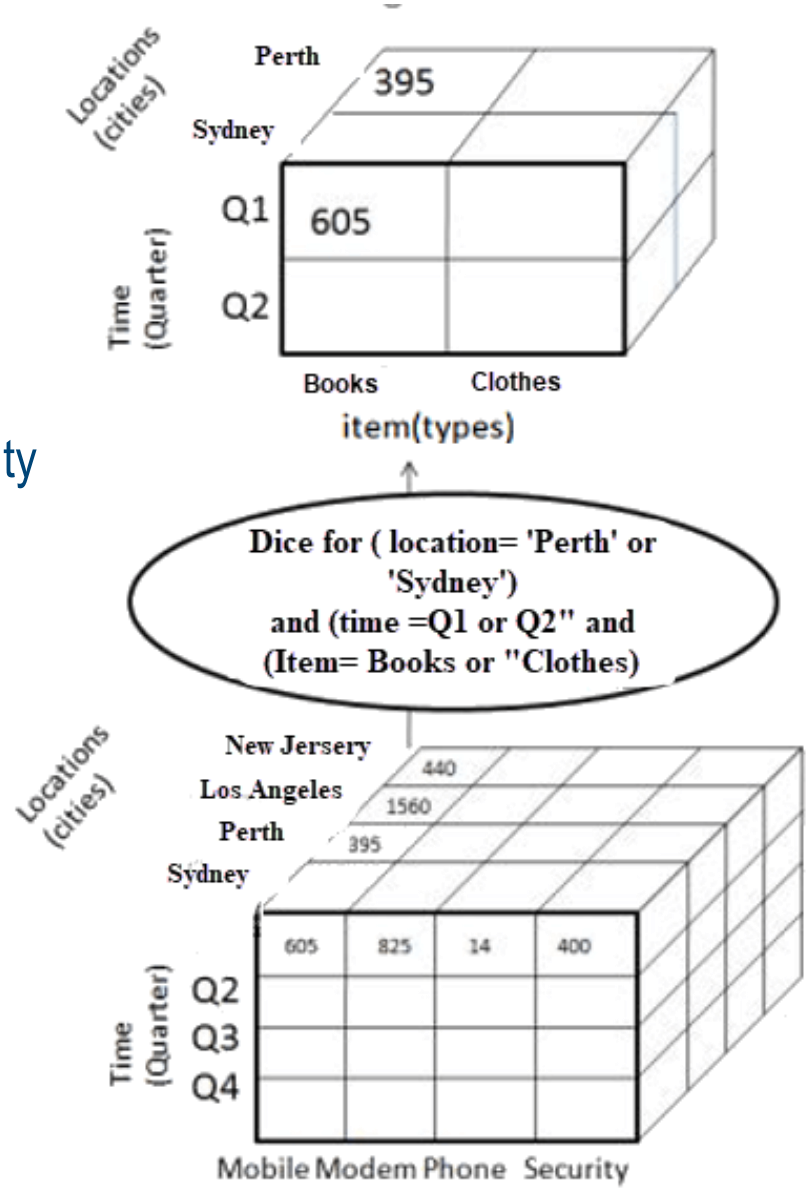
Operations: Slicing

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



Operations: Dicing

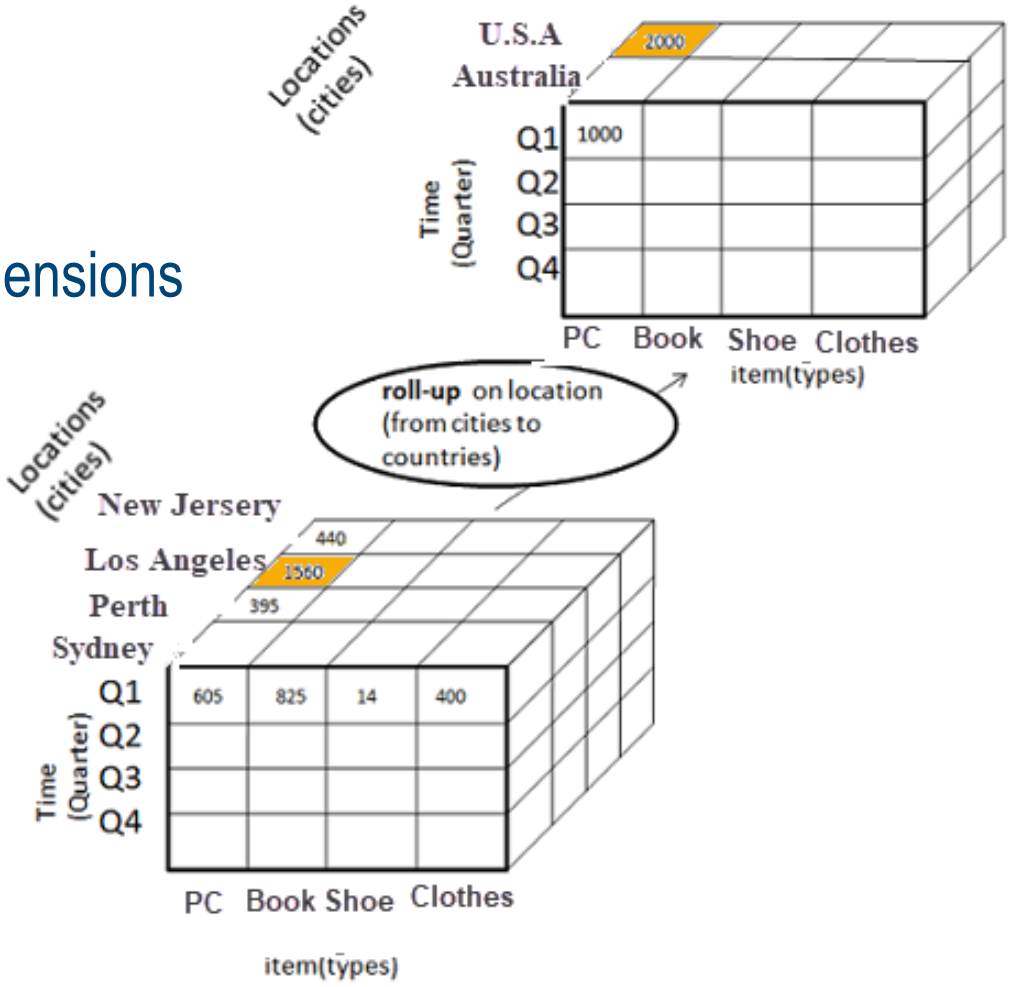
- Dicing = subsetting
  - „thicker slices“, not reducing dimensionality
- Ex: derive subcube
  by selecting along location, time, item
  simultaneously



[guru99.com]

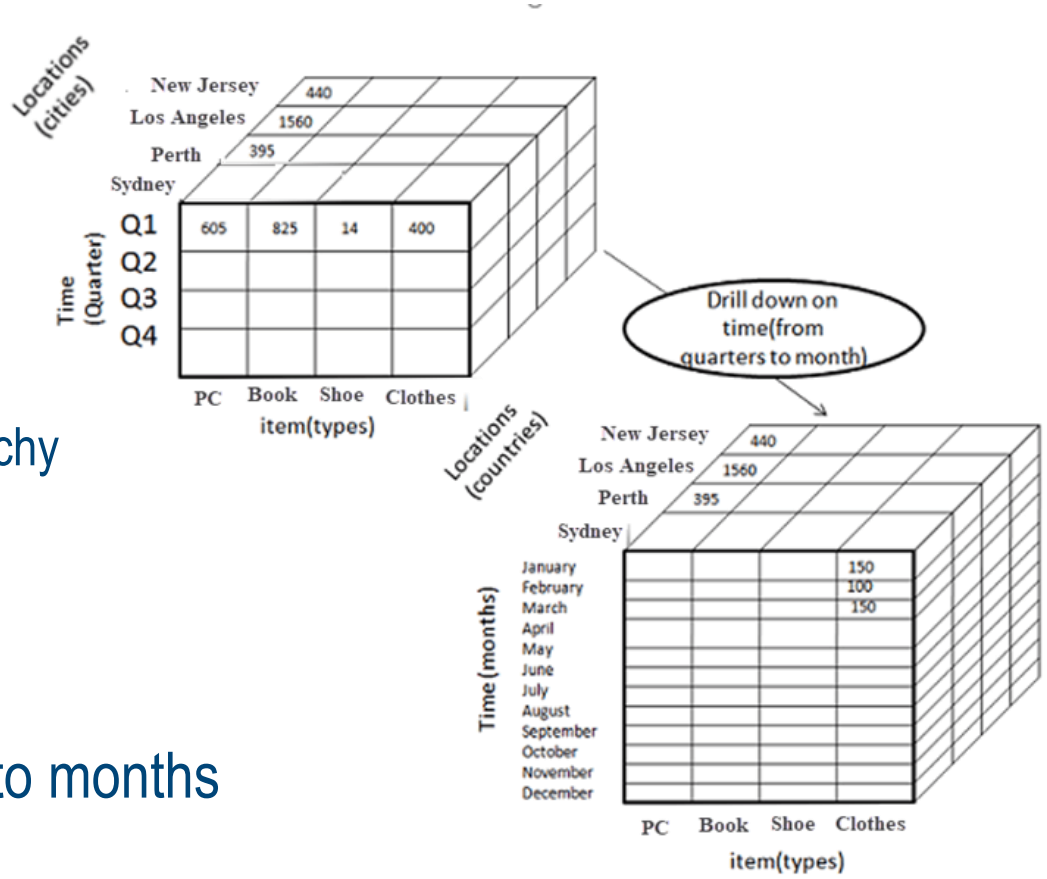
Operations: Roll-Up

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



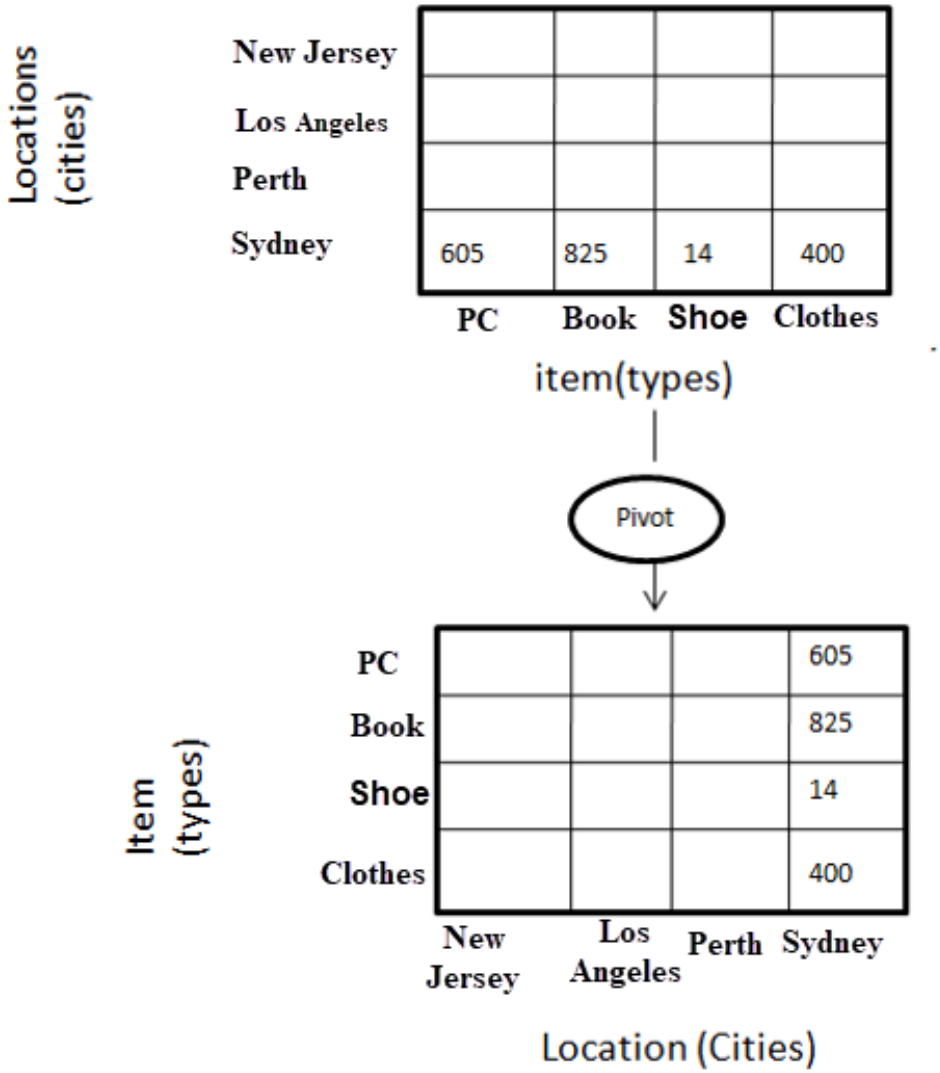
Operations: Drill-Down

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

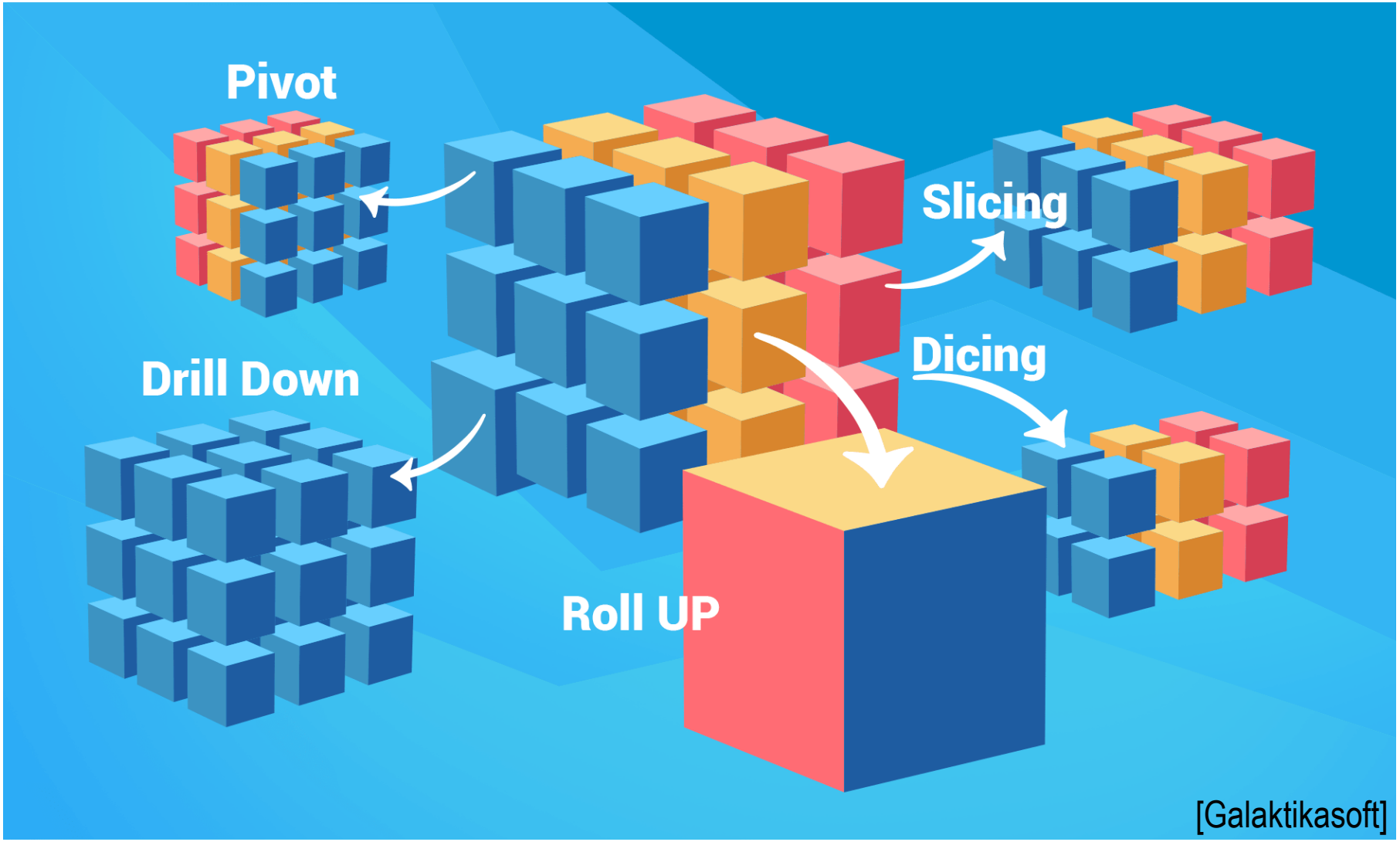


Operations: Pivot

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



Visual Summary: Datacube Ops



Overview

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

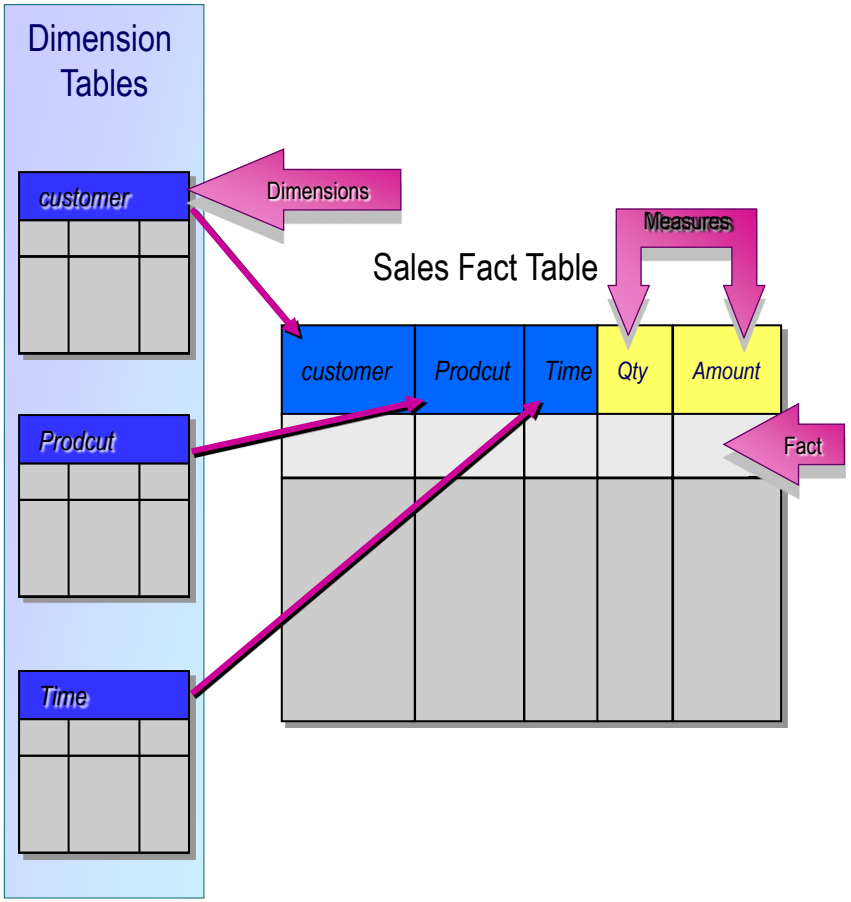
OLAP Datacube Querying

- 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 (Wikipedia):

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

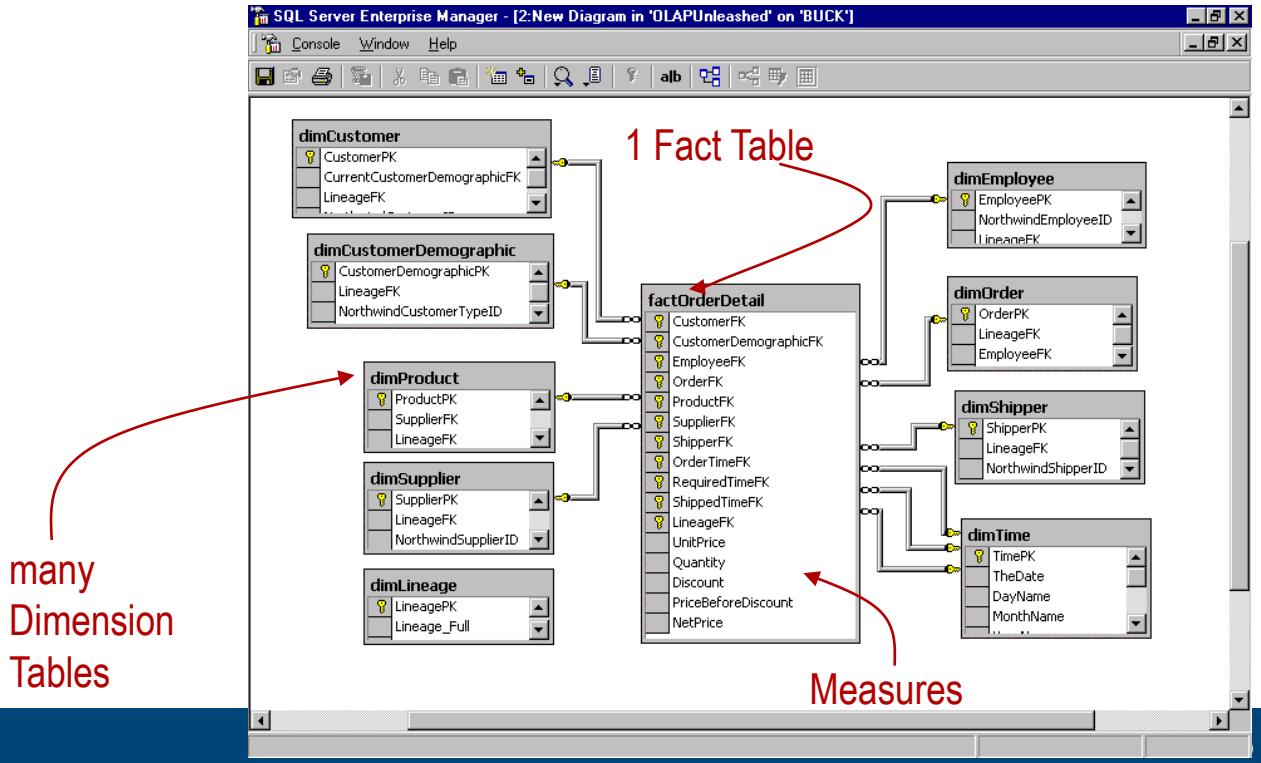
Datacubes in ROLAP: Facts & Dimensions

- 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



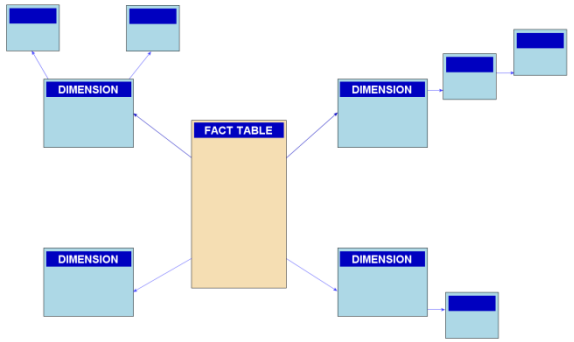
Star Schema

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

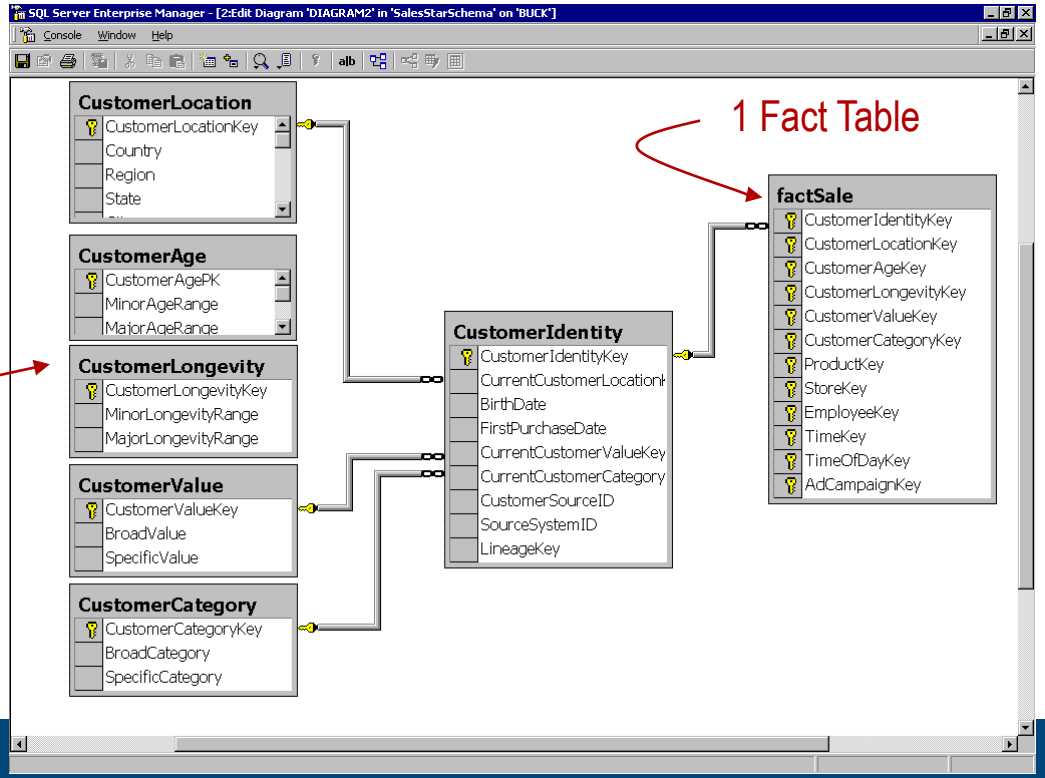


Snowflake Schema

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



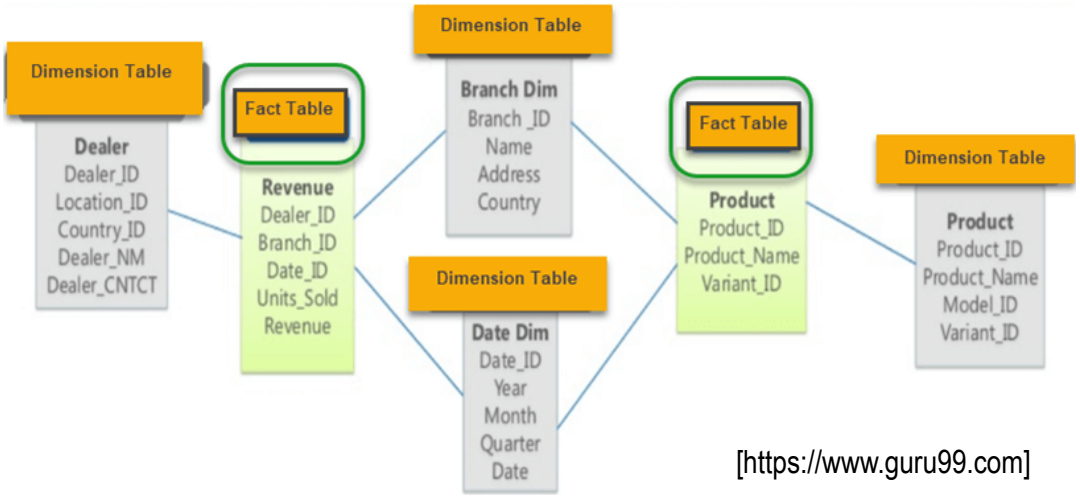
[SqlPac @ Wikipedia]



normalized Dimension Tables

Galaxy Schema

- Galaxy schema = combined datacubes
  - Sharing dimension(s)

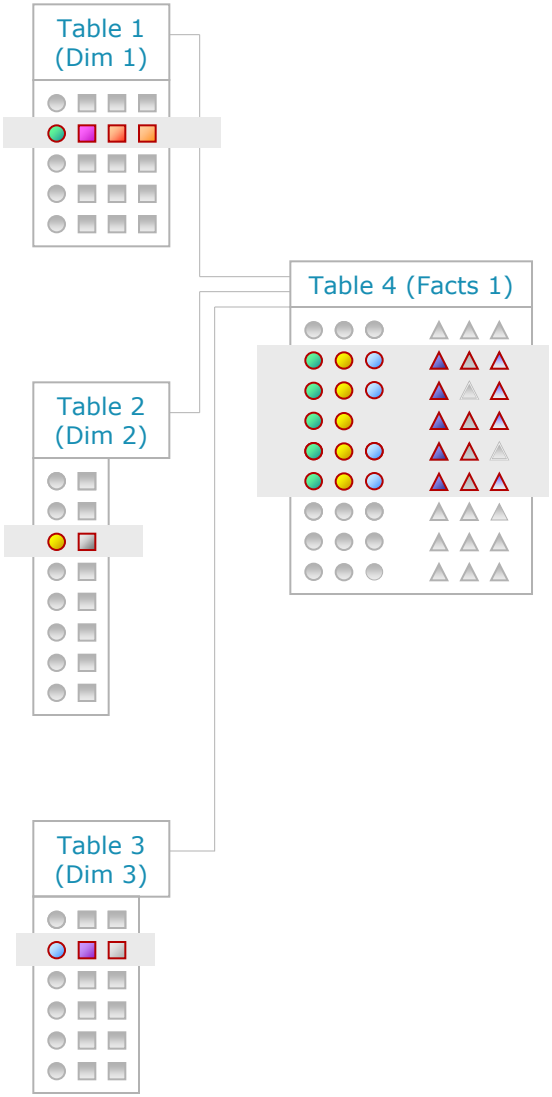


[https://www.guru99.com]

- helpful for aggregating fact tables

- also called „Fact Constellation Schema“

A Query in ROLAP



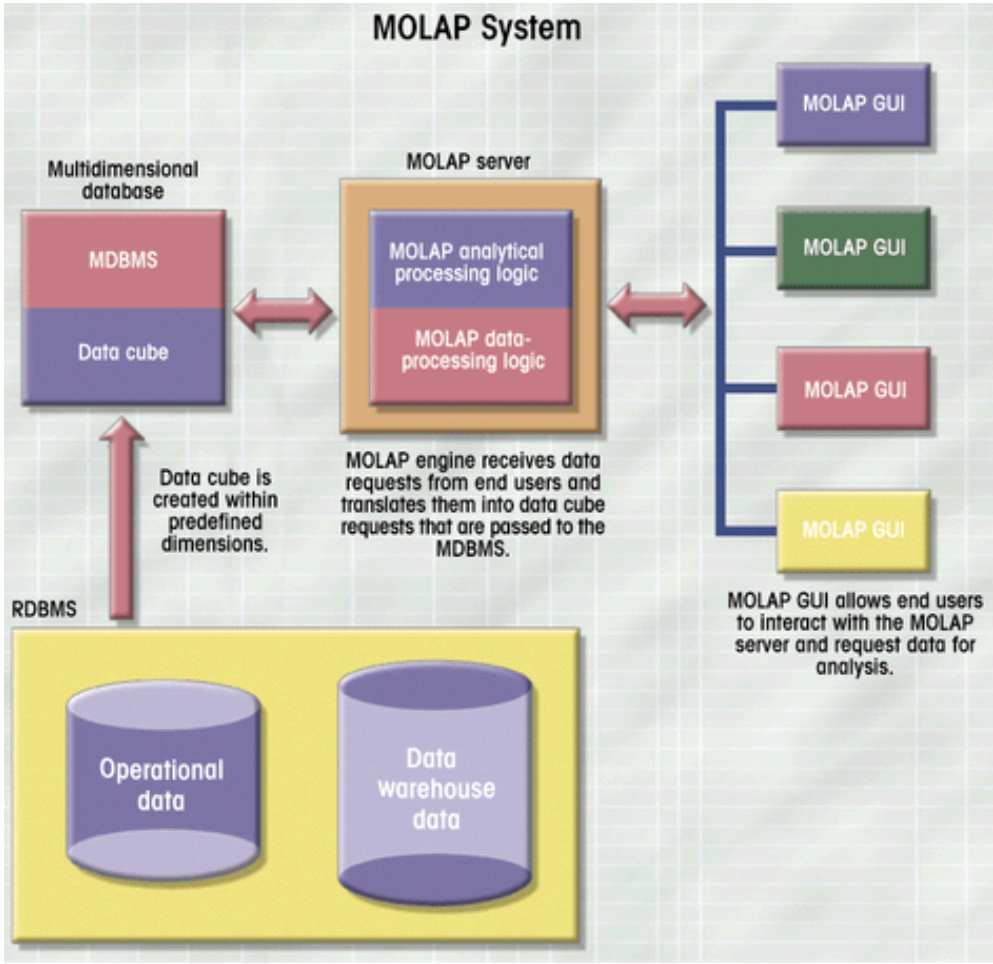

```
SELECT Fact_Column_1
      ,Fact_Column_2
  FROM Table 4     T4   -- Fact
      ,Table 1     T1   -- Dim 1
      ,Table 2     T2   -- Dim 2
      ,Table 3     T3   -- Dim 3
 WHERE T4.Dim_Col_1 = T1.Dim_Col_1
   AND T4.Dim_Col_2 = T2.Dim_Col_1
   AND T4.Dim_Col_3 = T3.Dim_Col_1
   AND T1.Dim_Property_2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

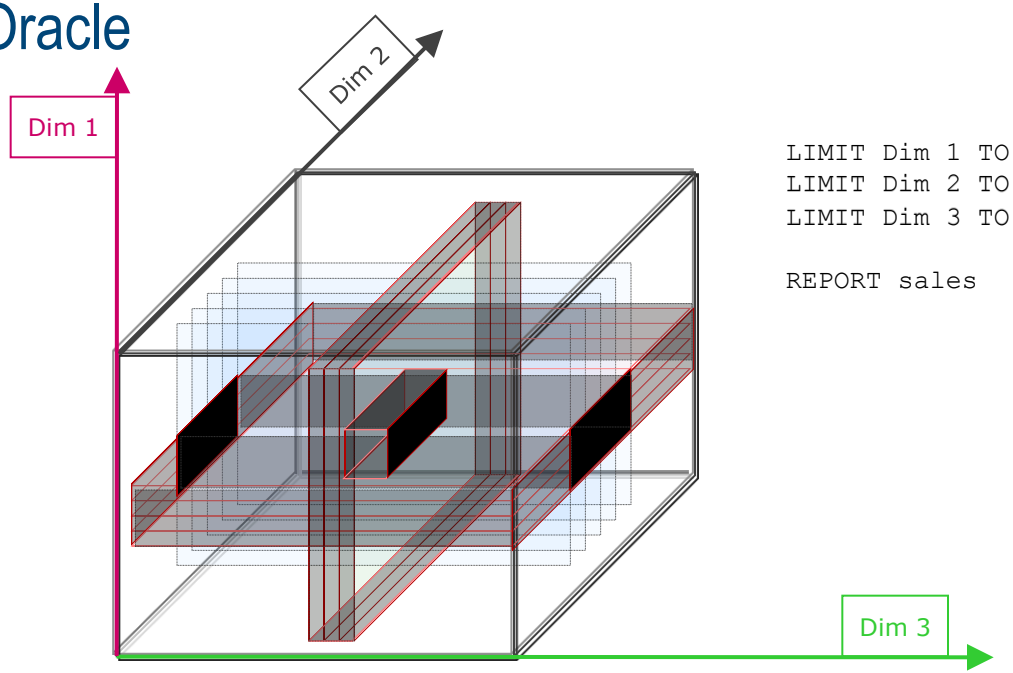
The MOLAP Approach

- 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
```

Overview

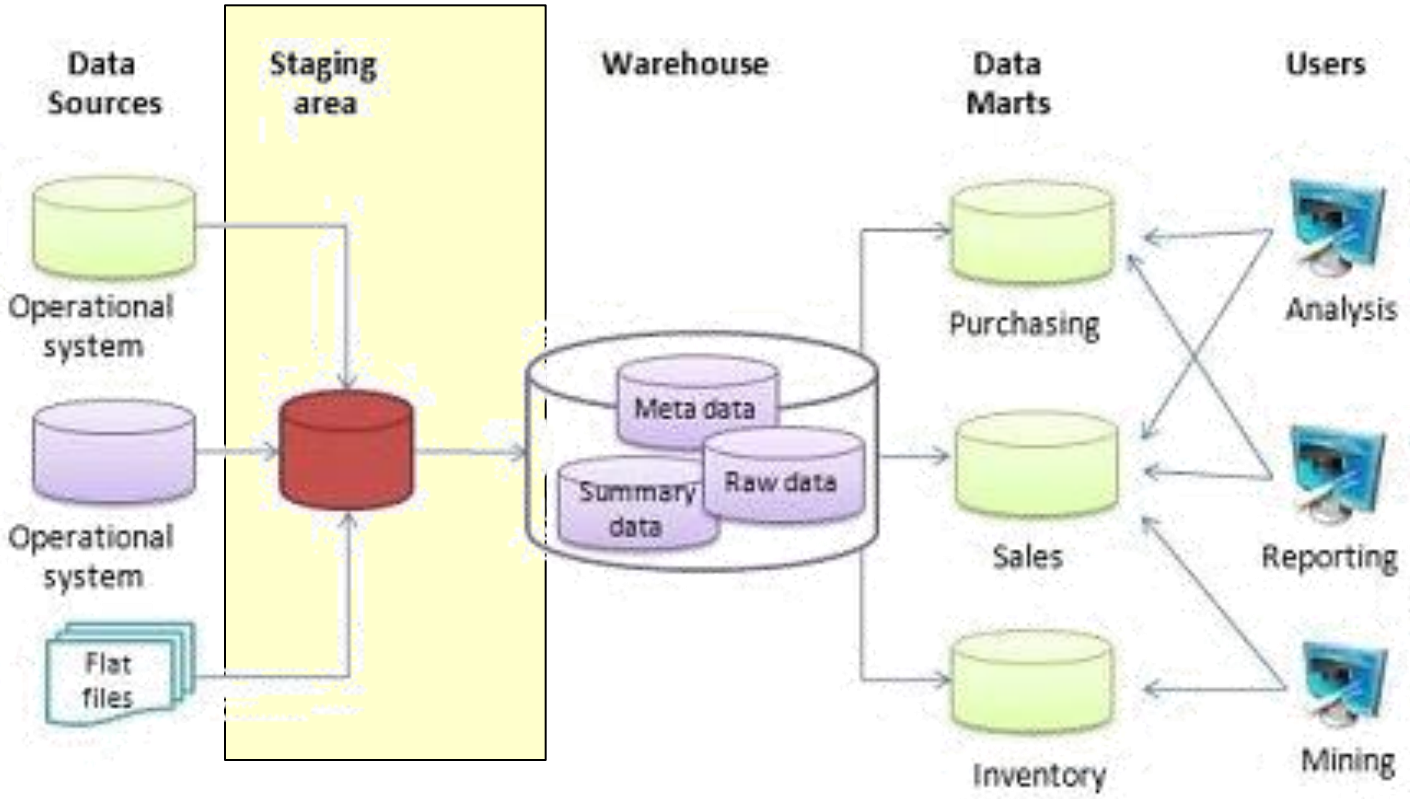
- Data Warehousing & Decision Support
- Datacubes, Dimension Hierarchies
- ROLAP & MOLAP
- ETL
- Summary

ETL

- Extract
  - Extract relevant data
- Transform
  - Transform data to DW format
  - Build keys, etc.
  - cleaning of data
- Load
  - Load data into DW
  - Build aggregates, etc.

- most underestimated process in DW development
- most time-consuming process in DW development
  - 80% of development time spent on ETL!

ETL in Data Warehouse Architecture



[soha jamil / Wikipedia]

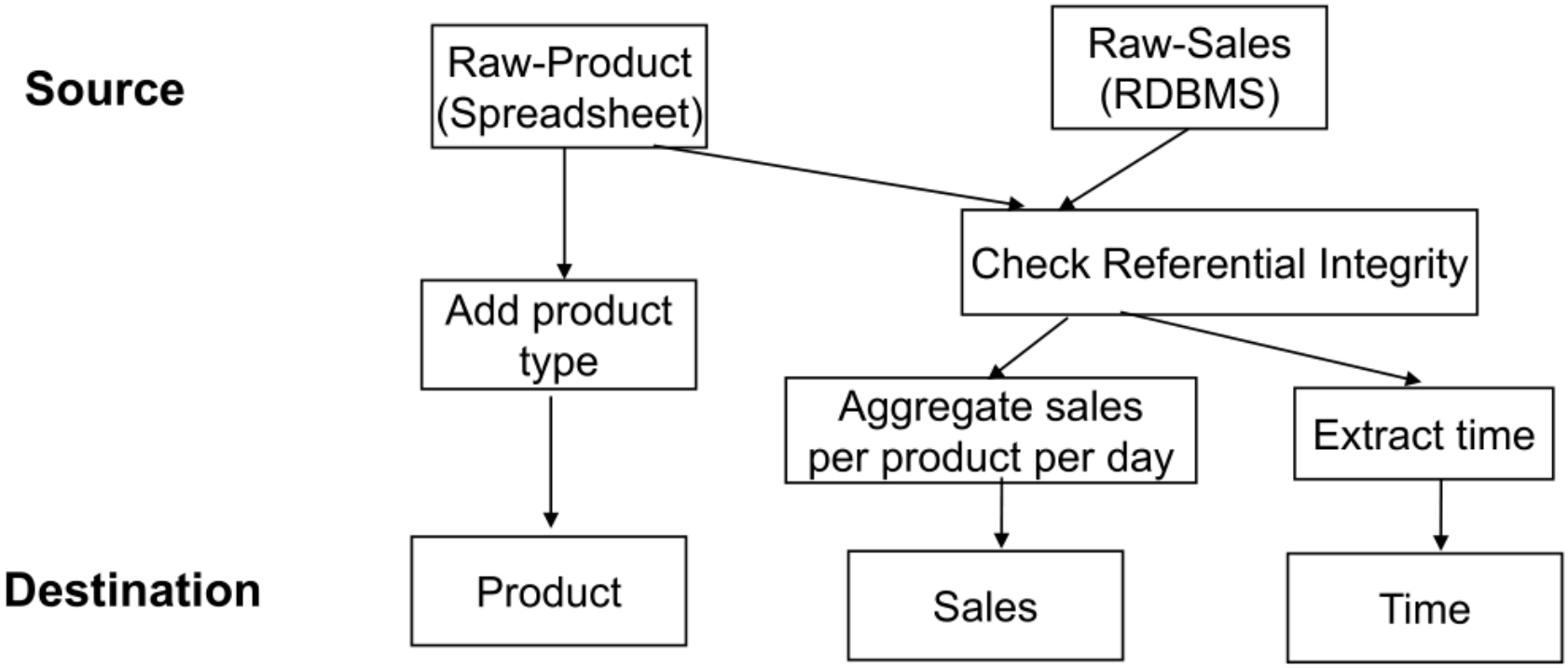
Common Transformations

- Data type conversions
  - EBCDIC → ASCII/UniCode
  - String manipulations
  - Date/time format conversions
    - *Ex: unix time 1201928400 = what time?*
- Normalization/denormalization
  - To desired DW format
  - Depending on source format
- Building keys
  - Table matches production keys to surrogate DW keys
  - Correct handling of history - especially for total reload

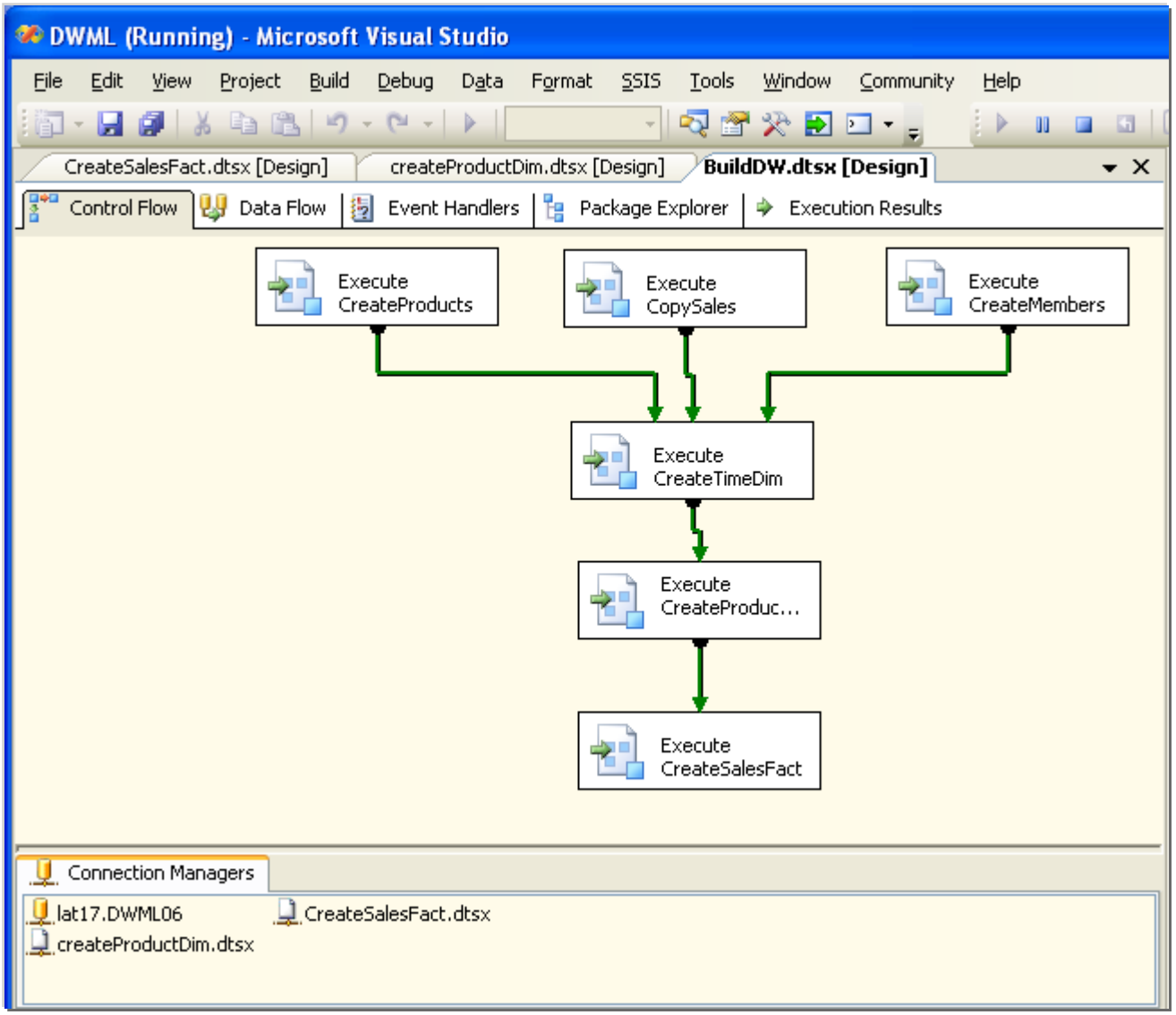
Data Cleansing: Why?

- Garbage In - Garbage Out
- BI does not work on “raw” data
  - Pre-processing necessary for BI analysis
- Handle inconsistent data formats: Spellings, codings, …
- Remove unnecessary attributes: Production keys, comments,…
- Replace codes with text (Why?)
  - City name instead of ZIP code, e.g., Aalborg Centrum vs. DK-9000
- Combine data from multiple sources with common key
  - E.g., customer data from customer address, customer name, …

Sample High-Level Extract Diagram



Ex: Microsoft BI Dev Studio



Overview

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

Summary: Data Warehousing Terminology

- Typically warehouse data is multidimensional, with very large fact tables
- Fact table
  - The subject, focus of analysis
- Measures
  - The specific elements of analysis
- Dimension
  - An object that allows to explore the measures from different perspectives
- Hierarchies
  - Classification of dimensions, useful for data exploration and aggregation
- Granularity
  - Level of detail of the stored data

Summary

- Data warehouse ≠ software product or application,
  but information processing system architecture geared at decision making
  - OLAP vs OLTP
- OLAP
  - Multi-dimensional, timeline, integrated, aggregated
  - ROLAP vs MOLAP
  - Star vs Snowflake vs Galaxy schema
- Part of bigger BI plot
  - ETL, Data Warehousing, OLAP, Data Mining, ...
- Recently: Data Lakes