

# 5- Data warehousing

---

## Data Management

---

*Michele Mastroianni*

[Michele.mastroianni@unicampania.it](mailto:Michele.mastroianni@unicampania.it)

[mmastroianni@unisa.it](mailto:mmastroianni@unisa.it)

## References

1. World Conservation Monitoring Centre. 1996. "**Guide to Information Management in the Context of the Convention on Biological Diversity**". United Nations Environment Programme, ISBN: 92-807-1591-5, Nairobi, Kenya. <http://www.mekonginfo.org/assets/midocs/0003032-utilities-communications-guide-to-information-management-in-the-context-of-the-convention-on-biological-diversity.pdf>
2. Rowley, J. (2007). **The wisdom hierarchy: representations of the DIKW hierarchy**. Journal of information science, 33(2), 163-180.  
[https://journals.sagepub.com/doi/abs/10.1177/0165551506070706?casa\\_token=kZHC0hnp354AAAAA:Om5KYFRjQ7YI0BHaOWYu\\_lazeb24ezb631\\_kja5Rc0C-P7-\\_HwH0tE2jA1Bb\\_vQ2KBw72GDJf3fl](https://journals.sagepub.com/doi/abs/10.1177/0165551506070706?casa_token=kZHC0hnp354AAAAA:Om5KYFRjQ7YI0BHaOWYu_lazeb24ezb631_kja5Rc0C-P7-_HwH0tE2jA1Bb_vQ2KBw72GDJf3fl)
3. Lenhardt, W C et al 2014 "**Data Management Lifecycle and Software Lifecycle Management in the Context of Conducting Science**". Journal of Open Research Software, 2(1): e15, pp. 1-4, DOI:  
<http://dx.doi.org/10.5334/jors.ax>
4. Allard, S. (2012). DataONE: Facilitating eScience through collaboration. Journal of eScience Librarianship, 1(1), 3.  
<https://pdfs.semanticscholar.org/91e6/472248cef044b7720b21353451fb7779895f.pdf>
5. FAUNDEEN, John L., et al. **The United States geological survey science data lifecycle model**. US Department of the Interior, US Geological Survey, 2013. <https://pubs.usgs.gov/of/2013/1265/pdf/of2013-1265.pdf>
6. KHAN, Nawsher, et al. **Big data: survey, technologies, opportunities, and challenges**. The scientific world journal, 2014, 2014. <https://downloads.hindawi.com/journals/tswj/2014/712826.pdf>
7. El Arass, M., Tikito, I., & Souissi, N. (2017, April). "**Data lifecycles analysis: towards intelligent cycle**". In 2017 Intelligent Systems and Computer Vision (ISCV) (pp. 1-8). IEEE. <https://ieeexplore.ieee.org/document/8054938>



Università  
degli Studi  
della Campania  
*Luigi Vanvitelli*

# What is a Data Warehouse?

Defined in many different ways, but not rigorously.

- A decision support database that is maintained **separately** from the organization's operational database
- Support **information processing** by providing a solid platform of consolidated, historical data for analysis.

“A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process.”—W. H. Inmon

# Data Warehouse is Subject-Oriented

- Organized around major subjects, such as **customer, product, sales**
- Focusing on the modeling and analysis of data for decision makers, not on daily operations or transaction processing
- Provide **a simple and concise** view around particular subject issues by **excluding data that are not useful in the decision support process**

# Data Warehouse is Integrated

- Constructed by integrating multiple, heterogeneous data sources
  - ■ relational databases, flat files, on-line transaction records
- ■ Data cleaning and data integration techniques are applied.
  - Ensure consistency in naming conventions, encoding structures, attribute measures, etc. among different data sources
    - E.g., Hotel price: currency, tax, breakfast covered, etc.
  - When data is moved to the warehouse, it is converted.

# Data Warehouse is Time Variant

- The time horizon for the data warehouse is significantly longer than that of operational systems
  - Operational database: current value data
  - Data warehouse data: provide information from a historical perspective (e.g., past 5-10 years)
- Every key structure in the data warehouse
  - Contains an element of time, explicitly or implicitly
  - But the key of operational data may or may not contain “time element”

# Data Warehouse is Nonvolatile

- A **physically separate store** of data transformed from the operational environment
- Operational **update of data does not occur** in the data warehouse environment
  - Does not require transaction processing, recovery, and concurrency control mechanisms
  - Requires only two operations in data accessing:
    - *initial loading of data* and *access of data*

# Two different approaches

- **OLTP:** On-Line Transaction Processing
  - Is oriented to transaction-based procedures (selling, ticket booking,....)
- **OLAP:** On-Line Analytical Processing
  - Is oriented to interactive analysis of business data
- Both OLAP and OLTP may based on the same data



# OLTP vs. OLAP

- Sell a product by an online seller
- Sell a ticket for train Victoria Station-Oxford
- How many products have been sold from Sport shoes Division in May?
- How many ticket have been sold by the train stations and for which location?

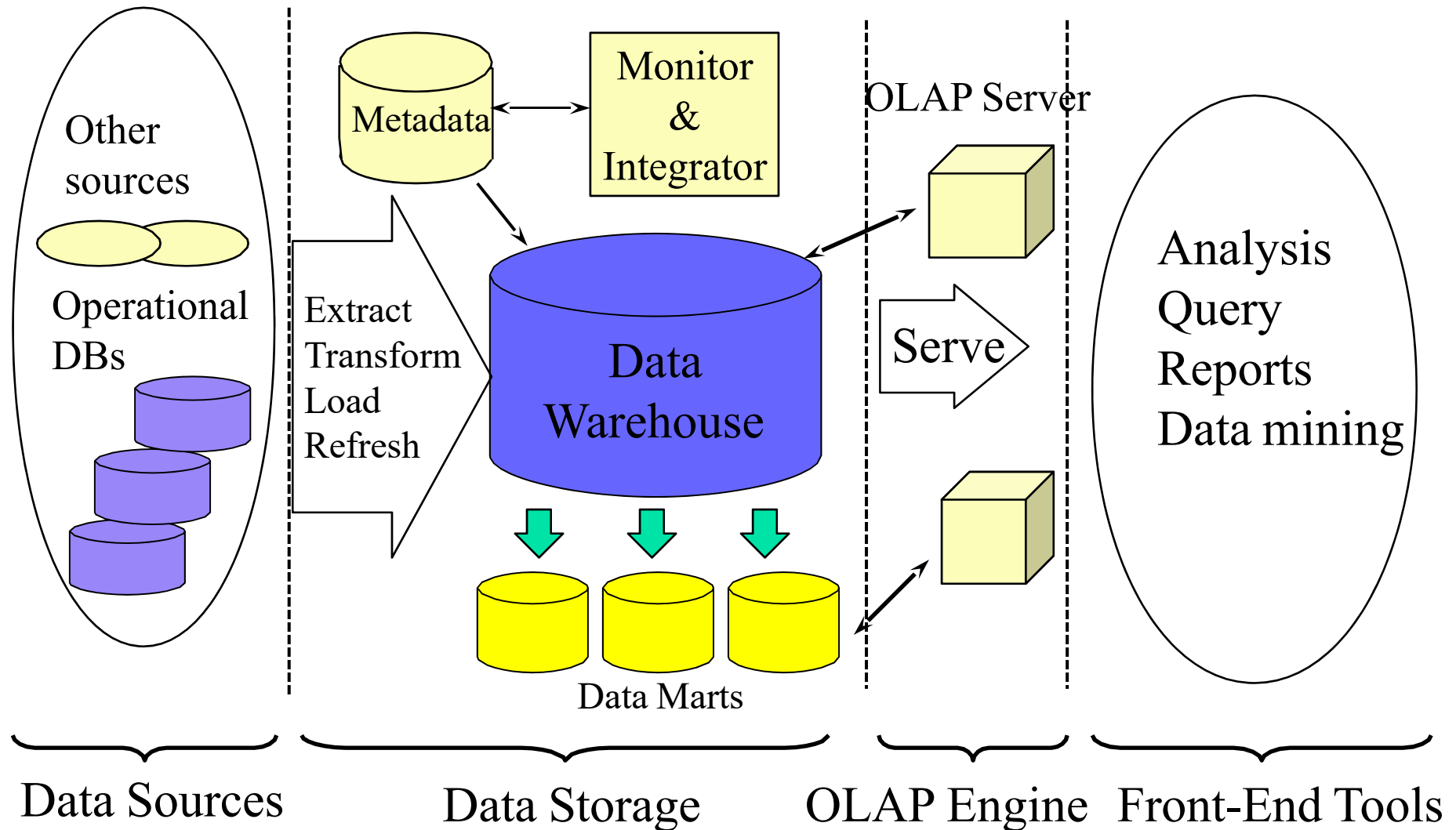
# OLTP vs. OLAP

	<b>OLTP</b>	<b>OLAP</b>
<b>users</b>	clerk, IT professional	knowledge worker
<b>function</b>	day to day operations	decision support
<b>DB design</b>	application-oriented	subject-oriented
<b>data</b>	current, up-to-date detailed, flat relational isolated	historical, summarized, multidimensional integrated, consolidated
<b>usage</b>	repetitive	ad-hoc
<b>access</b>	read/write index/hash on prim. key	lots of scans
<b>unit of work</b>	short, simple transaction	complex query
<b># records accessed</b>	tens	millions
<b>#users</b>	thousands	hundreds
<b>DB size</b>	100MB-GB	100GB-TB
<b>metric</b>	transaction throughput	query throughput, response

# Why a Separate Data Warehouse?

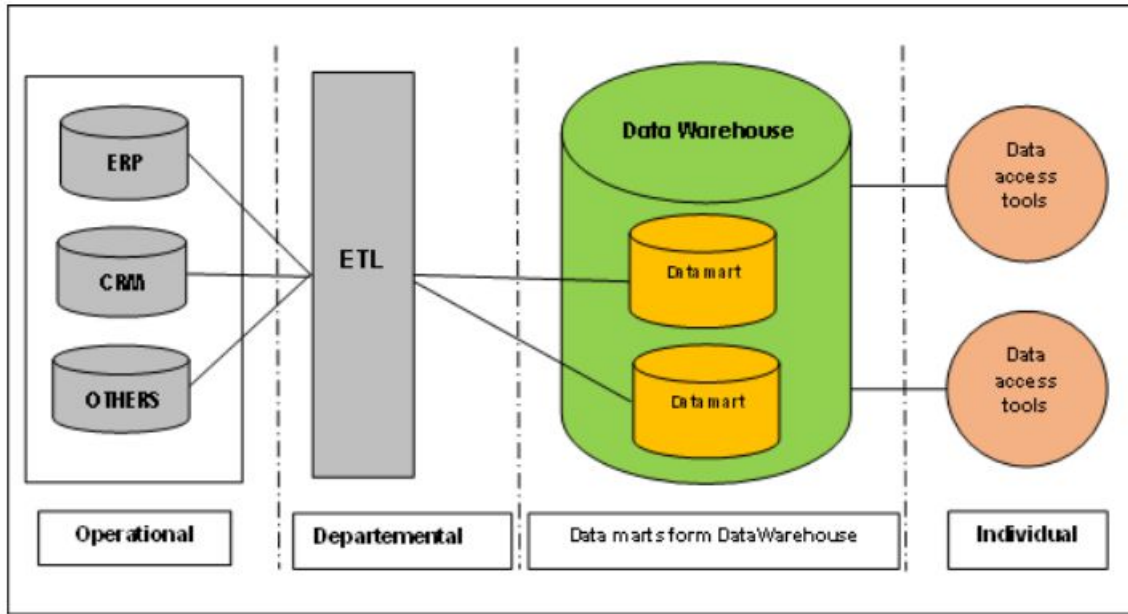
- High performance for both systems
  - DBMS – tuned for OLTP: access methods, indexing, concurrency control, recovery
  - Warehouse – tuned for OLAP: complex OLAP queries, multidimensional view, consolidation
- Different functions and different data:
  - missing data: Decision support requires historical data which operational DBs do not typically maintain
  - data consolidation: DS requires consolidation (aggregation, summarization) of data from heterogeneous sources
  - data quality: different sources typically use inconsistent data representations, codes and formats which have to be reconciled
- Note: There are more and more systems which perform OLAP analysis directly on relational databases

# Data Warehouse: A Multi-Tiered Architecture

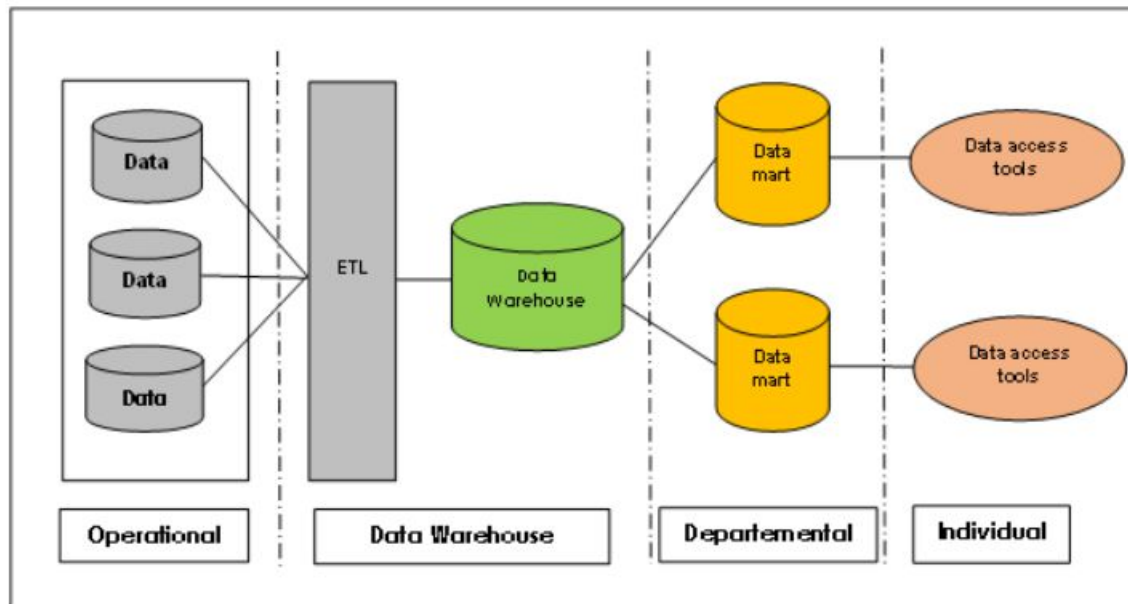


# Three Data Warehouse Models

- **Enterprise warehouse**
  - collects all of the information about subjects spanning the entire organization
- **Data Mart**
  - a subset of corporate-wide data that is of value to a specific groups of users. Its scope is confined to specific, selected groups, such as marketing data mart
    - Independent (captured from external sources)  
Dependent (directly from company datawarehouse)
- **Virtual warehouse**
  - A set of views over operational databases
  - Only some of the possible summary views may be materialized



Inmon model



Kimball Model

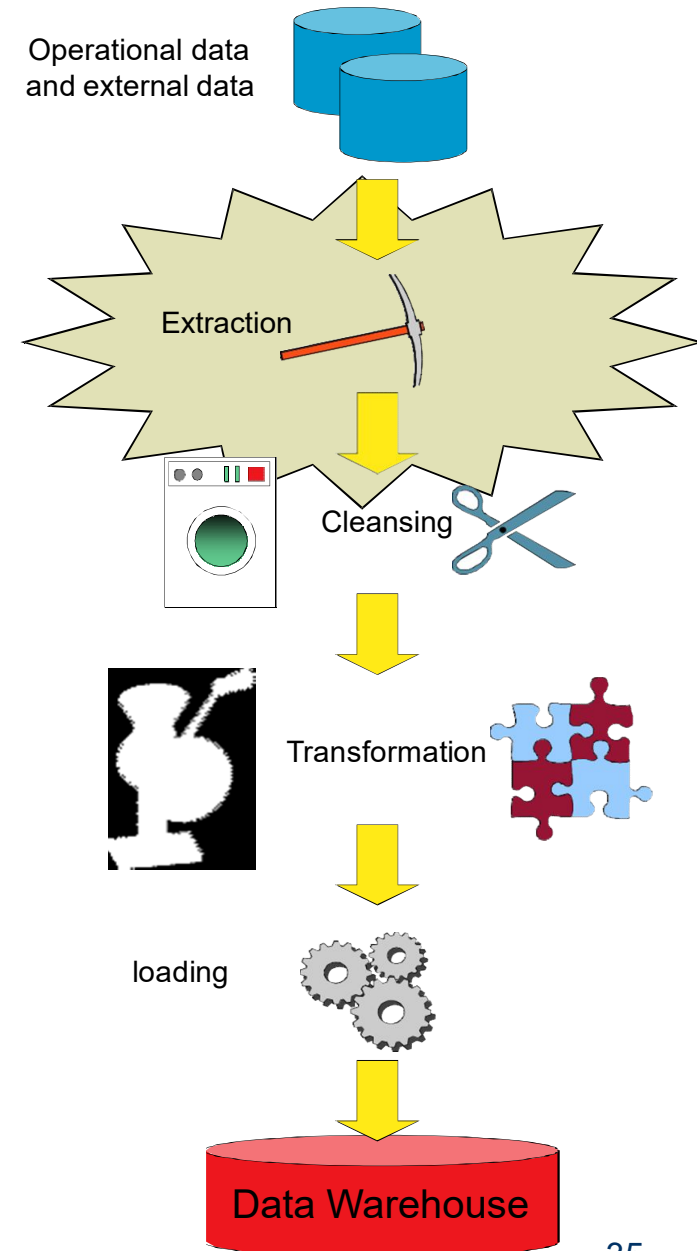
# Extraction, Transformation, and Loading (ETL)

- ETL processes extract, integrate, and clean data from operational sources to eventually feed the data warehouse.
- At the abstract level, ETL processes produce a single, high-quality, detailed data source, that in turn feed the DW (*reconciliation*)
- Depending on the architecture, this data source is physical (reconciled data layer) or virtual. In the former case ETL processes are physically directly connected to the reconciled layer, in the latter to the (primary) DW or to the datamarts.
- Reconciliation takes place in two occasions: when a data warehouse is populated for the first time, and every time the data warehouse is updated.
- ETL consists in four phases:
  - ✓ *extraction*
  - ✓ *cleansing (or cleaning)*
  - ✓ *transformation*
  - ✓ *loading*

**Note:** Cleansing and transformation are not always considered as separate phases in the literature. Here we say that cleansing is essentially devoted to correct *values*, whereas transformation is mainly devoted to correct *formats*.

# Extraction

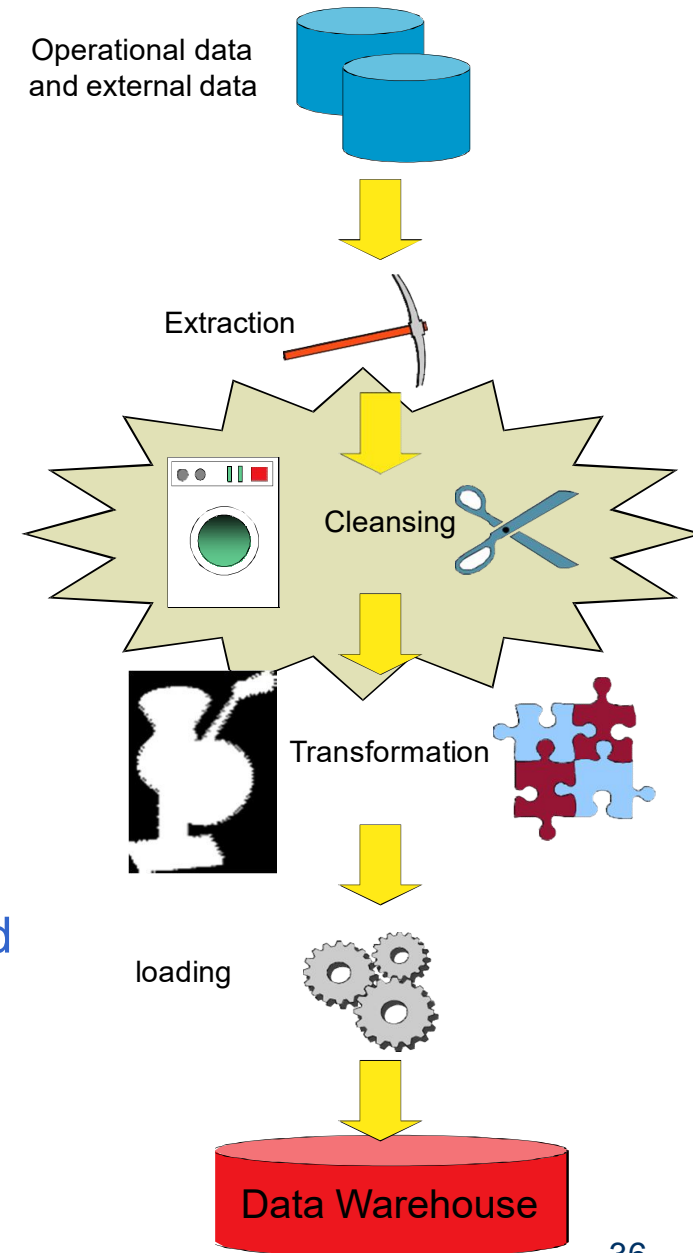
- Relevant data is obtained from sources:
  - ✓ *Static extraction* is performed when a data warehouse needs populating for the first time. Conceptually speaking, this looks like a snapshot of operational data.
  - ✓ *Incremental extraction* is used to update data warehouses regularly, and seizes the changes applied to source data since the latest extraction.
    - often based on the log maintained by the operational DBMS
    - based on time-stamp
    - source-driven
- The data to be extracted is mainly selected on the basis of its quality.





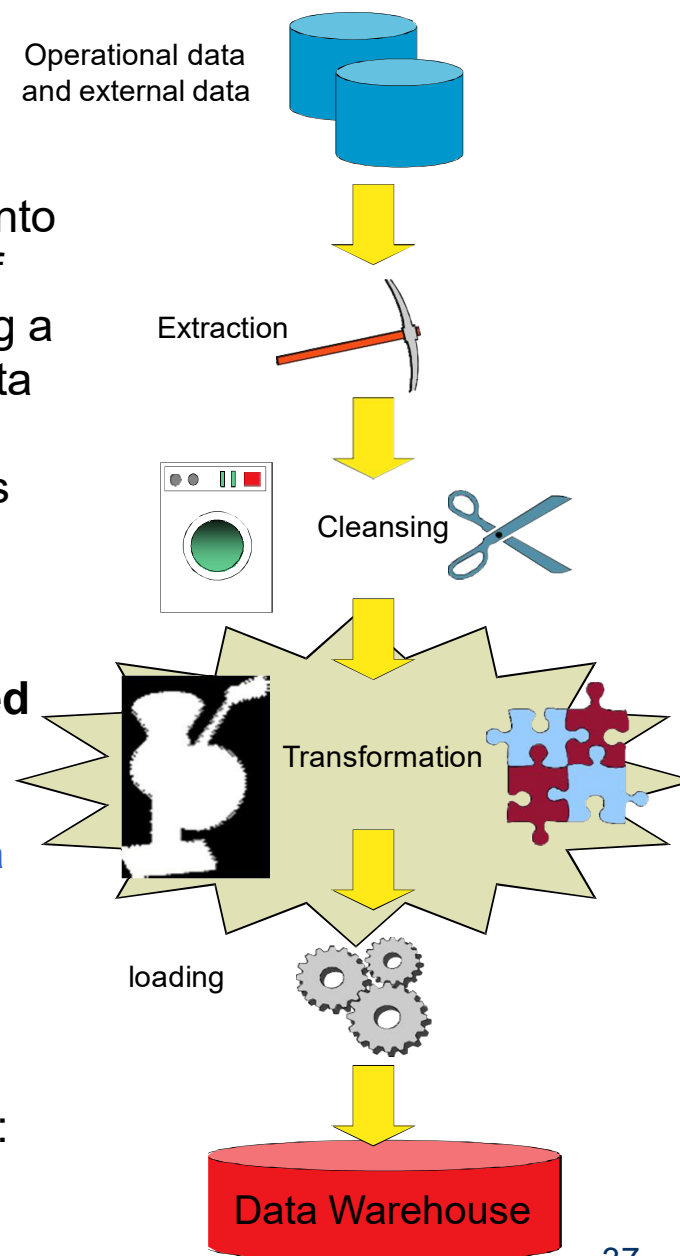
# Cleansing

- It is supposed to improve data quality—normally quite poor in sources. The most frequent inconsistencies are due to
  - ✓ Duplicate data
  - ✓ Inconsistent values that are logically associated (*e.g., cities and zip codes*)
  - ✓ Missing Data
  - ✓ Unexpected use of fields
  - ✓ Impossible or wrong values
  - ✓ Inconsistent values for a single entity because different practices were used (*e.g., due to use of abbreviations*)
  - ✓ Inconsistent values for a single entity because of typing mistakes



# Transformation


- It converts data from its operational source format into a specific data warehouse format. Independently of the presence of a reconciled data layer, establishing a mapping between the source data layer and the data warehouse layer is generally made difficult by the presence of many different, heterogeneous sources
  - ✓ There may be loose texts that can hide valuable information
  - ✓ Different formats can be used for individual data
- main transformations for **populating the reconciled data layer**:
  - ✓ conversion and standardization that operate on both storage formats and units of measure to uniform data
  - ✓ matching that associates equivalent fields in different sources
  - ✓ selection that reduces the number of source fields and record
- main transformations for **populating the DW layer**:
  - ✓ normalization is substituted by denormalization
  - ✓ Aggregation is used to sum up data



# Cleansing and Transformation

John White  
Downing St. 10  
TW1A 2AA London (UK)

*Formatting*

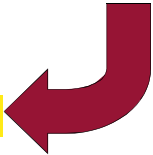


```
firstName: John  
lastName: White  
address: Downing St. 10  
ZIPcode: TW1A 2AA  
city: London  
country: UK
```

```
firstName: John  
lastName: White  
address: 10, Downing Street  
ZIPcode: TWA1 2AA  
city: London  
country: United Kingdom
```

*Standardization*

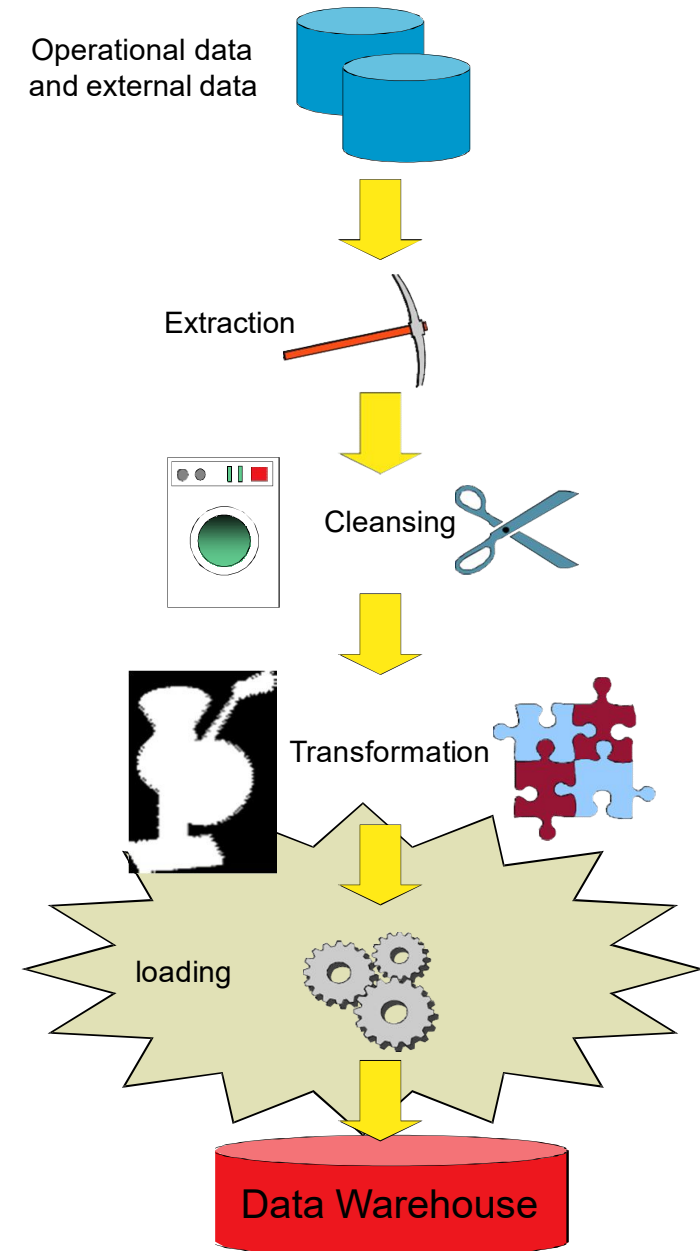
*Correction*



```
firstName: John  
lastName: White  
address: 10, Downing Street  
ZIPcode: SWA1 2AA  
city: London  
nation: United Kingdom
```

# Loading

- Loading data into the DW
  - ✓ **Refresh:** Data warehouse data is completely rewritten. This means that older data is replaced. Refresh is normally used in combination with static extraction to initially populate a data warehouse.
  - ✓ **Update:** Only those changes applied to source data are added to the data warehouse. Update is typically carried out without deleting or modifying preexisting data. This technique is used in combination with incremental extraction to update data warehouses regularly.



# Multidimensional Model

“What is the total amount of receipts recorded last year per state and per product category?”

“What is the relationship between the trend of PC manufacturers’ shares and quarter gains over the last five years?”

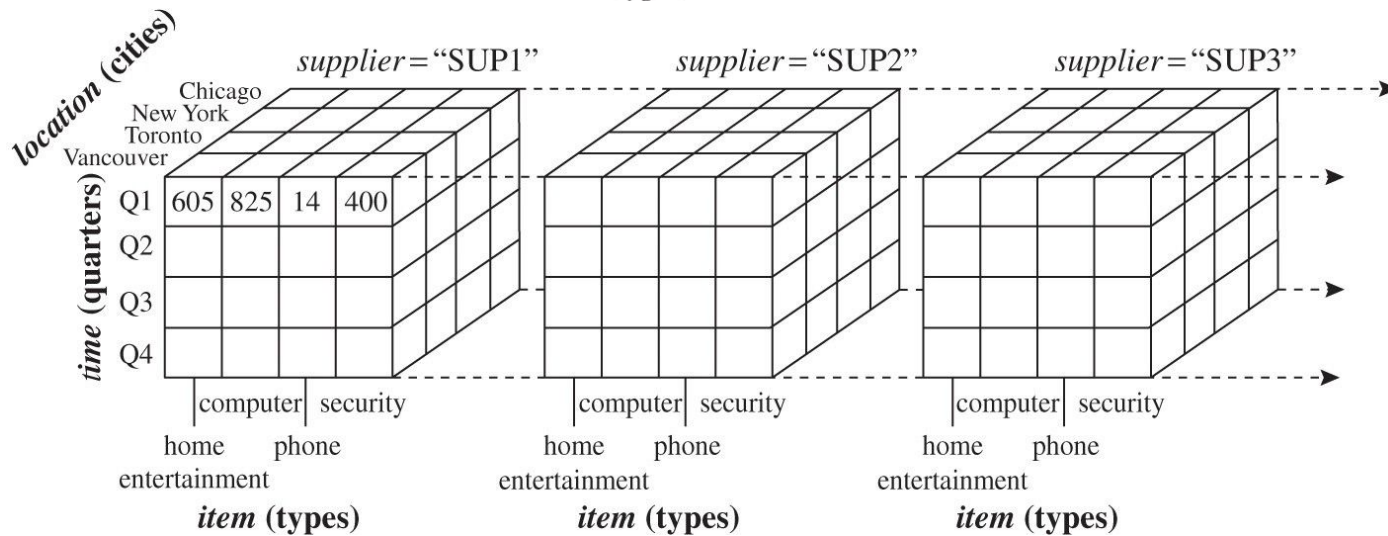
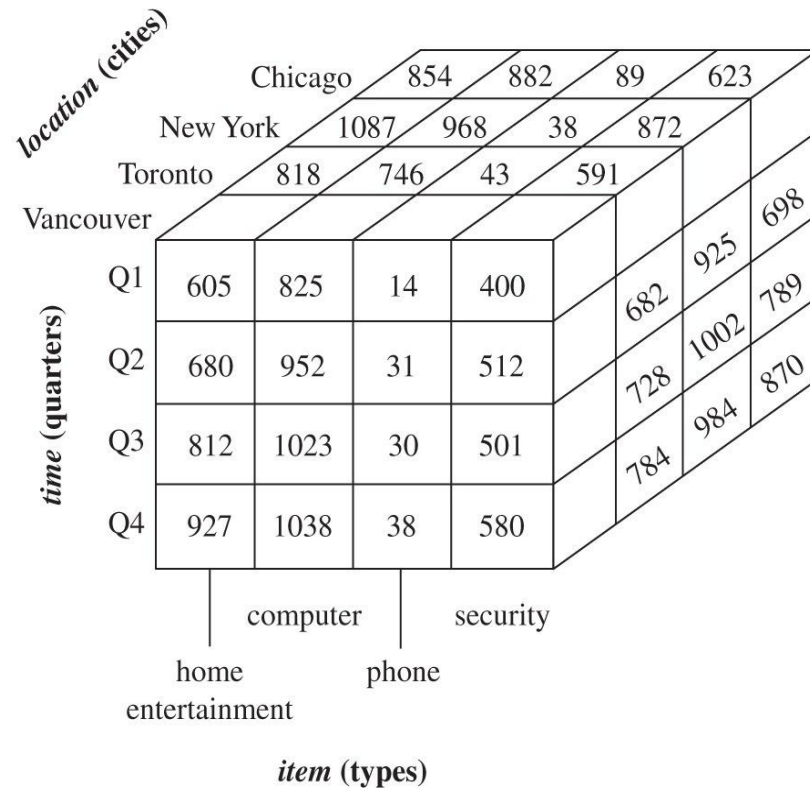
“Which orders maximize receipts?”

“Which one of two new treatments will result in a decrease in the average period of admission?”

“What is the relationship between profit gained by the shipments consisting of less than 10 items and the profit gained by the shipments of more than 10 items?”

# From Tables and Spreadsheets to Data Cubes

- A **data warehouse** is based on a **multidimensional data model** which views data in the form of a data cube
- A data cube, such as **sales**, allows data to be modeled and viewed in multiple dimensions
  - **Dimension tables**, such as **item** (**item\_name**, **brand**, **type**), or **time**(**day**, **week**, **month**, **quarter**, **year**)
  - **Fact table** contains **measures** (such as **dollars\_sold**) and keys to each of the related dimension tables
- In data warehousing literature, an n-D base cube is called a **base cuboid**. The top most 0-D cuboid, which holds the highest-level of summarization, is called the **apex cuboid**. The lattice of cuboids forms a **data cube**.

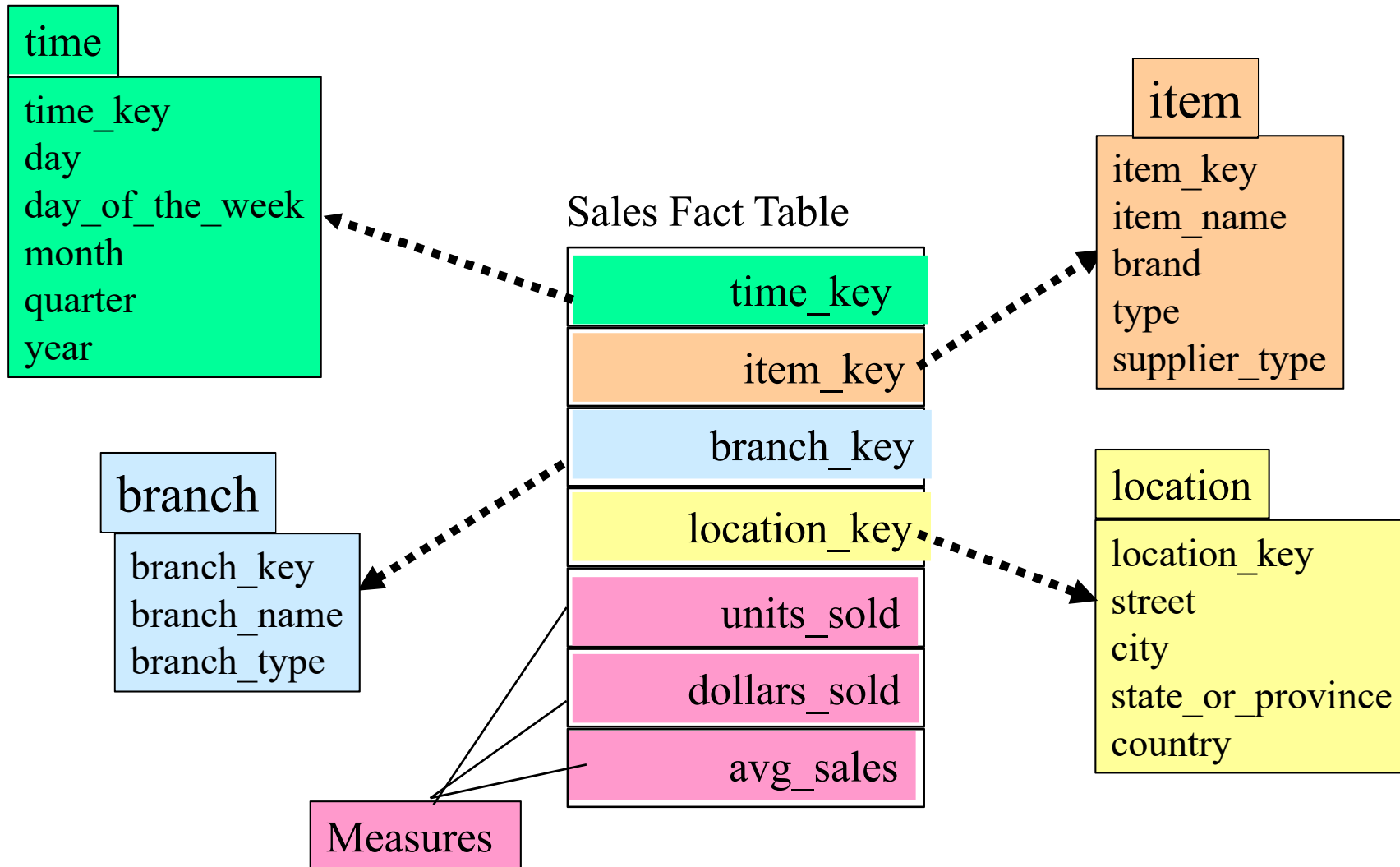


# Modeling of Data Warehouse

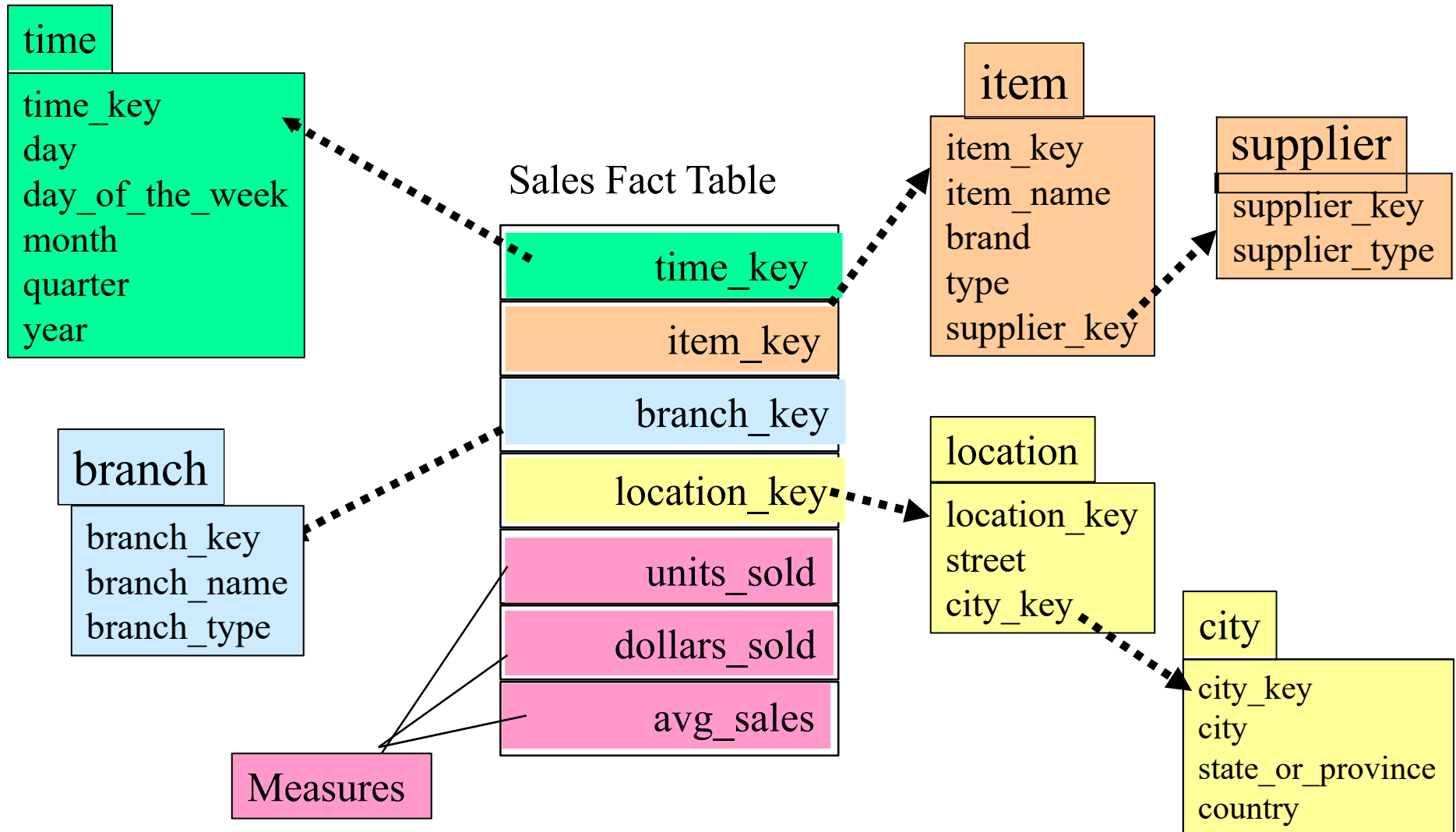
- Modeling data warehouses: dimensions & measures
  - Star schema: A fact table in the middle connected to a set of dimension tables
  - Snowflake schema: A refinement of star schema where some dimensional hierarchy is **normalized** into a set of smaller dimension tables, forming a shape similar to snowflake
  - Fact constellations: Multiple fact tables share dimension tables, viewed as a collection of stars, therefore called **galaxy schema** or fact constellation



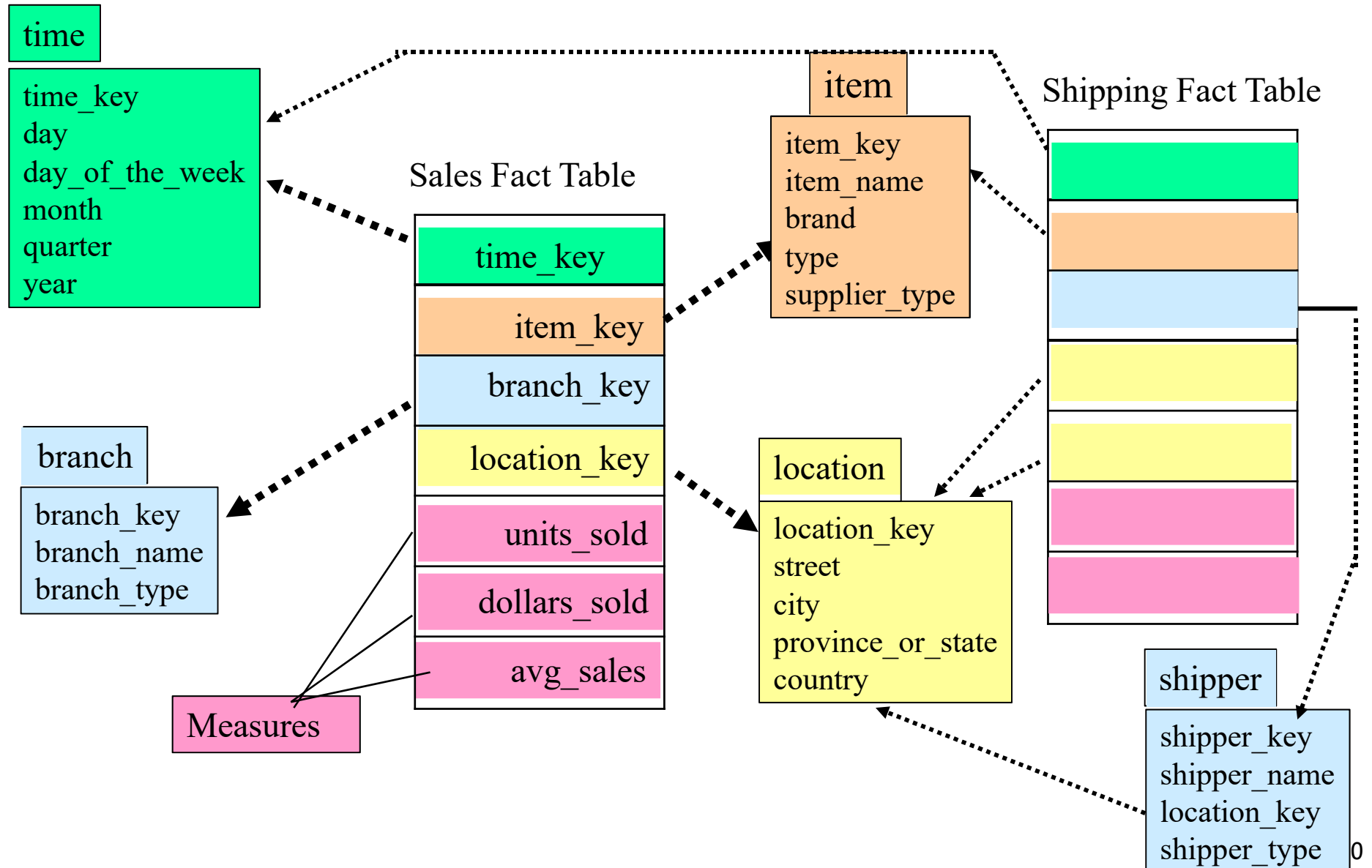
# Example of Star Schema



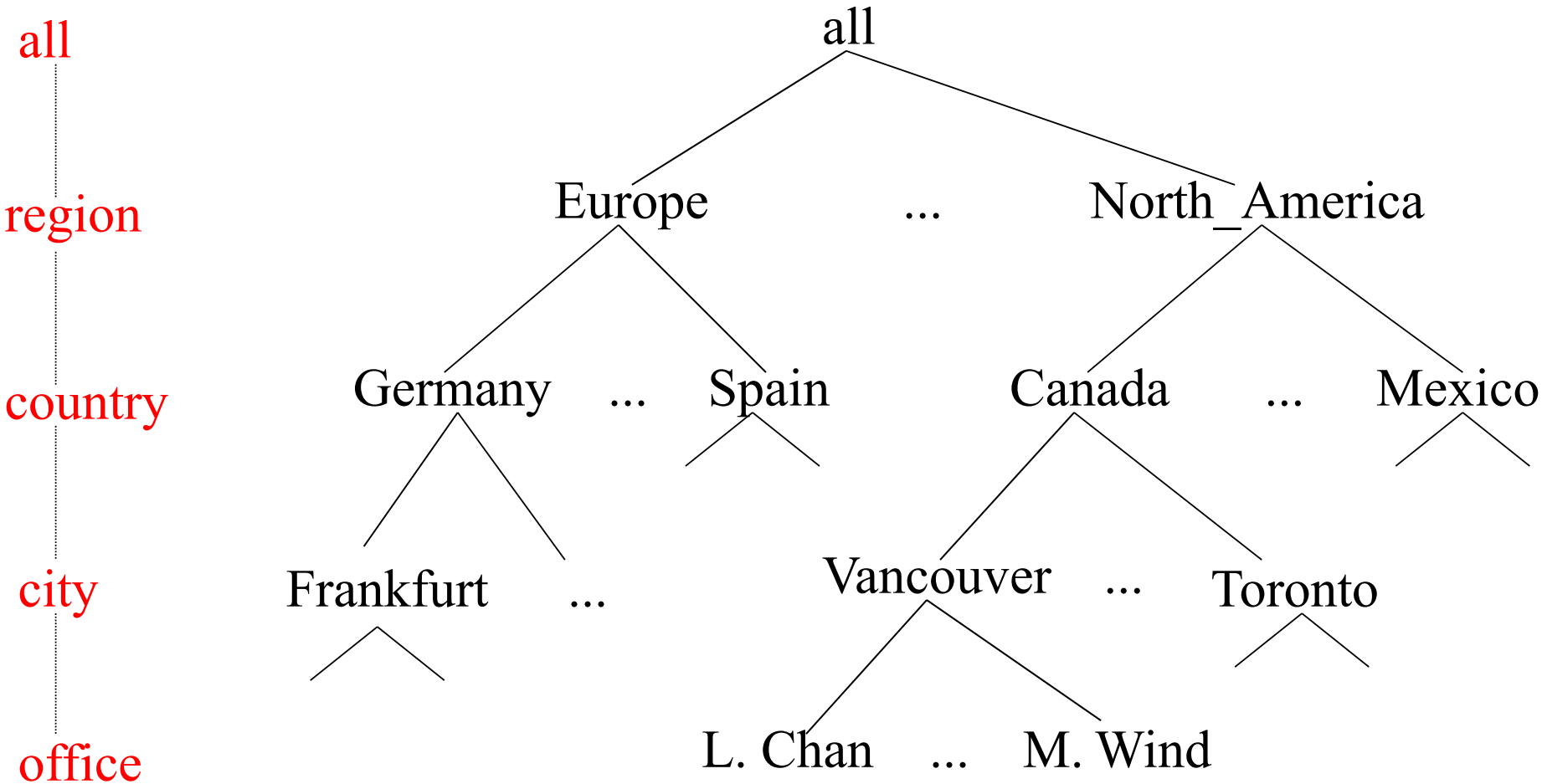
# Example of Snowflake Schema



# Example of Fact Constellation



# A Concept Hierarchy: Dimension (location)



# Data cube measures

- Measure: a numeric function that can be evaluated at each point in the data cube space:
  - Fact
  - Aggregation of facts

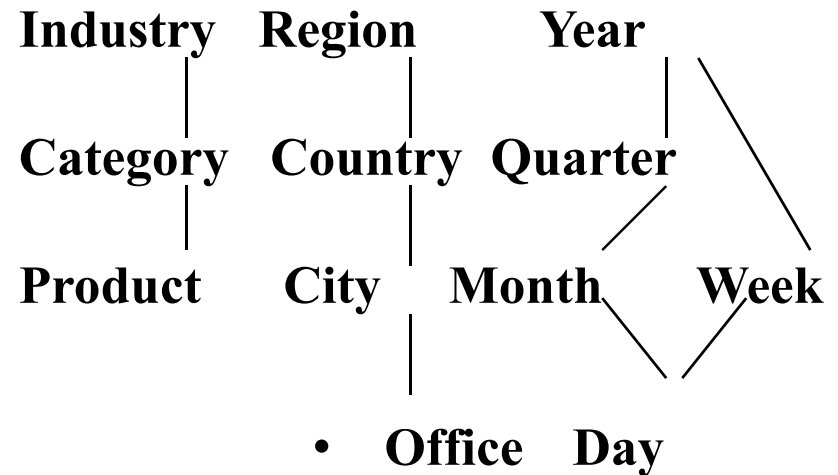
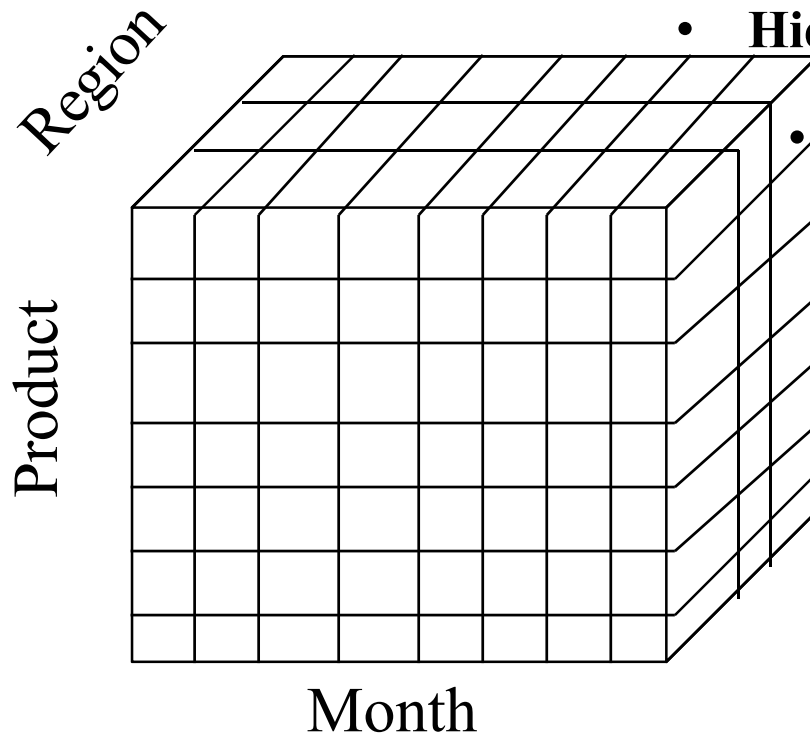
# Data Cube Measures: Three Categories

- **Distributive**: if the result derived by applying the function to  $n$  aggregate values is the same as that derived by applying the function on all the data without partitioning
  - E.g., `count()`, `sum()`, `min()`, `max()`
- **Algebraic**: if it can be computed by an algebraic function with  $M$  arguments (where  $M$  is a bounded integer), each of which is obtained by applying a distributive aggregate function
  - E.g., `avg() = sum() / count()`, `min_N()` ...
- **Holistic**: if there is no constant bound on the storage size needed to describe a subaggregate.
  - E.g., `median()`, `mode()`, `rank()`

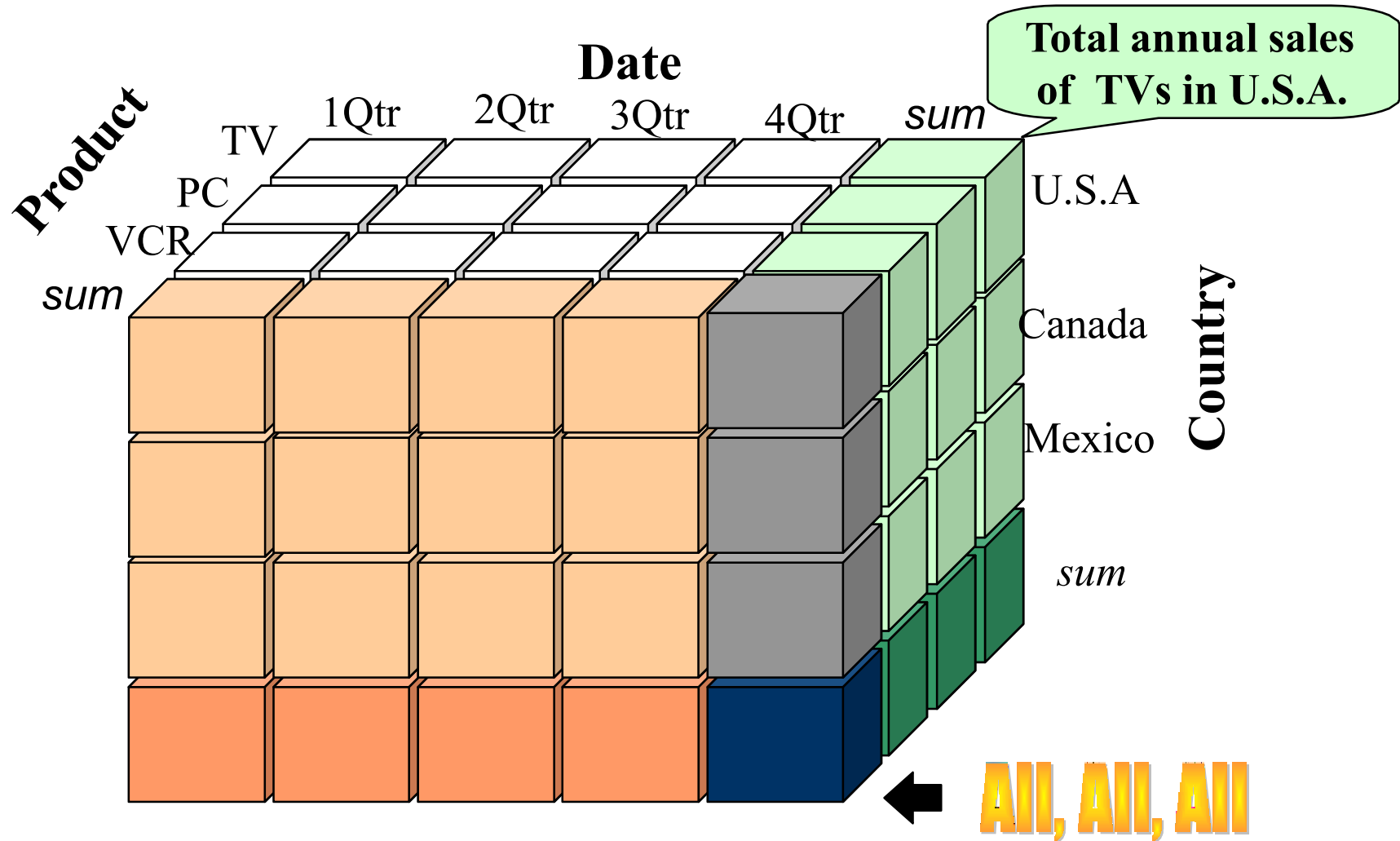
# Multidimensional Data

- Sales volume as a function of product, month, and region

- Dimensions: *Product, Location, Time*
- Hierarchical summarization paths



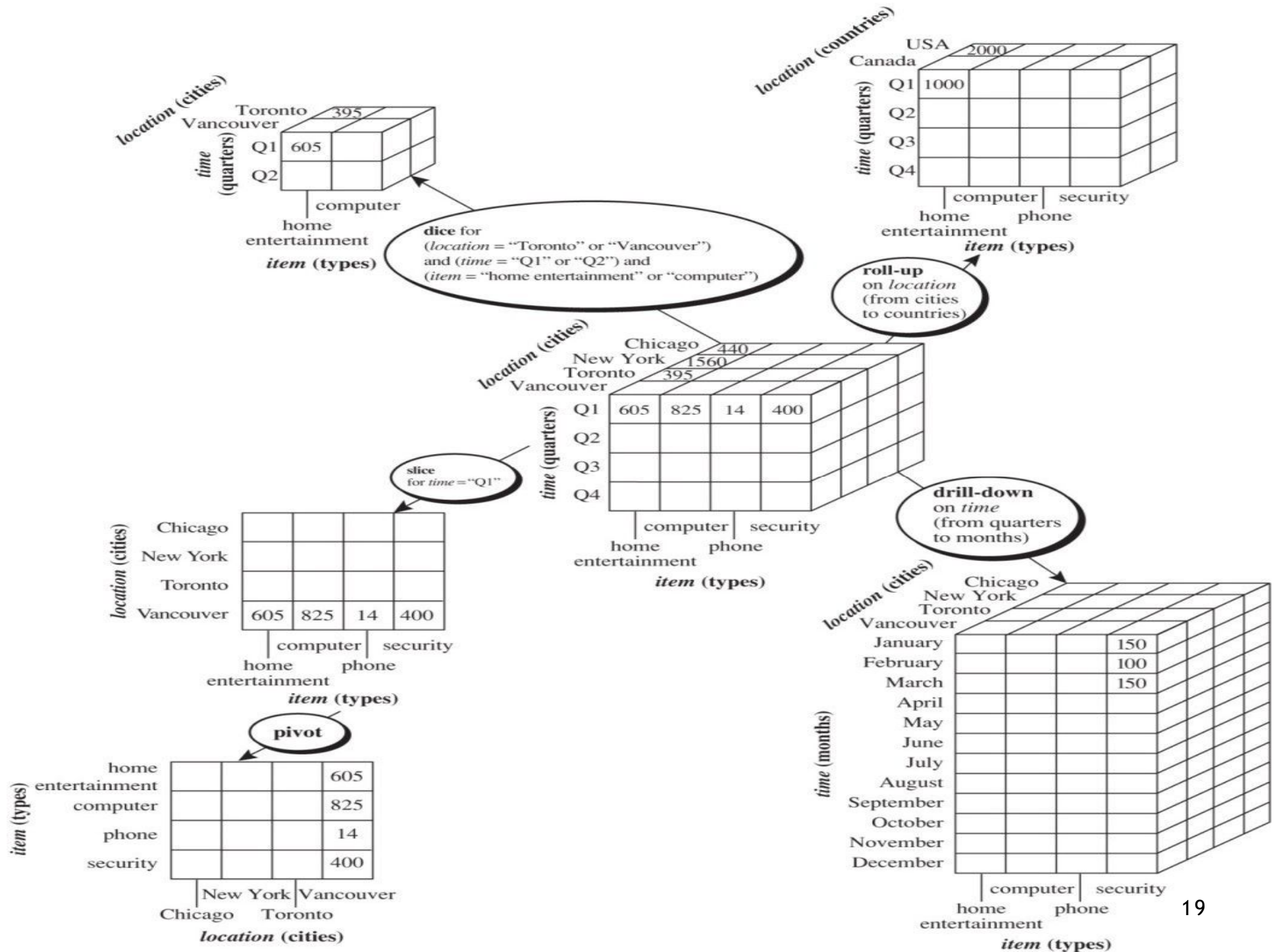
# A Sample Data Cube

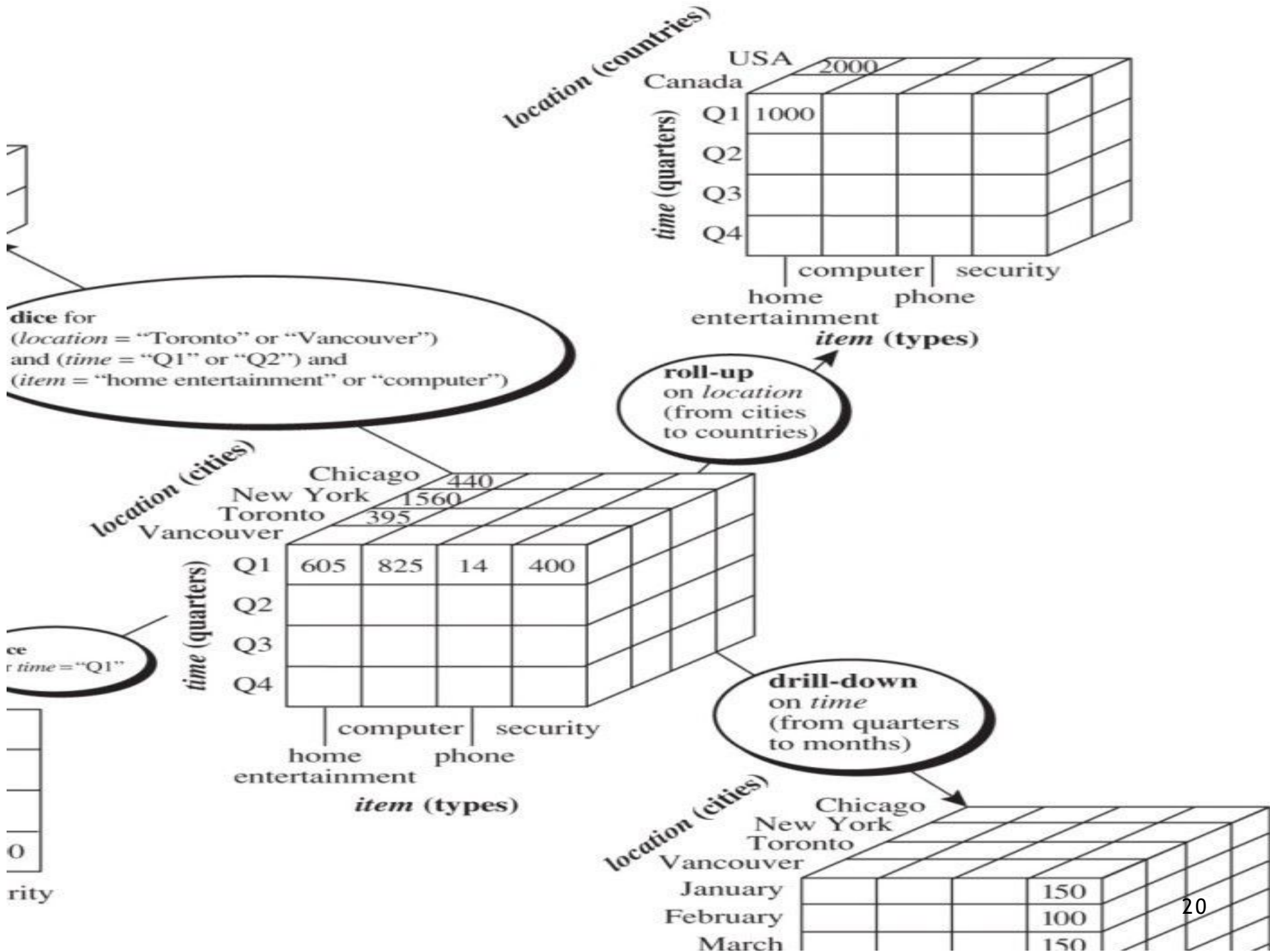




# Typical OLAP Operations

- **Roll up (drill-up):** summarize data
  - *by climbing up hierarchy or by dimension reduction*
- **Drill down (roll down):** reverse of roll-up
  - *from higher level summary to lower level summary or detailed data, or introducing new dimensions*
- **Slice and dice:** *project and select*
- **Pivot (rotate):**
  - *reorient the cube, visualization, 3D to series of 2D planes*
- Other operations
  - **drill across:** *involving (across) more than one fact table*





(item = "home entertainment" or "computer")

drill-down on location (from cities to countries)

location (cities)

time (quarters)

	Chicago	440		
	New York	1560		
	Toronto	395		
	Vancouver			
Q1	605	825	14	400
Q2				
Q3				
Q4				
	computer	security	home entertainment	phone

slice for time = "Q1"

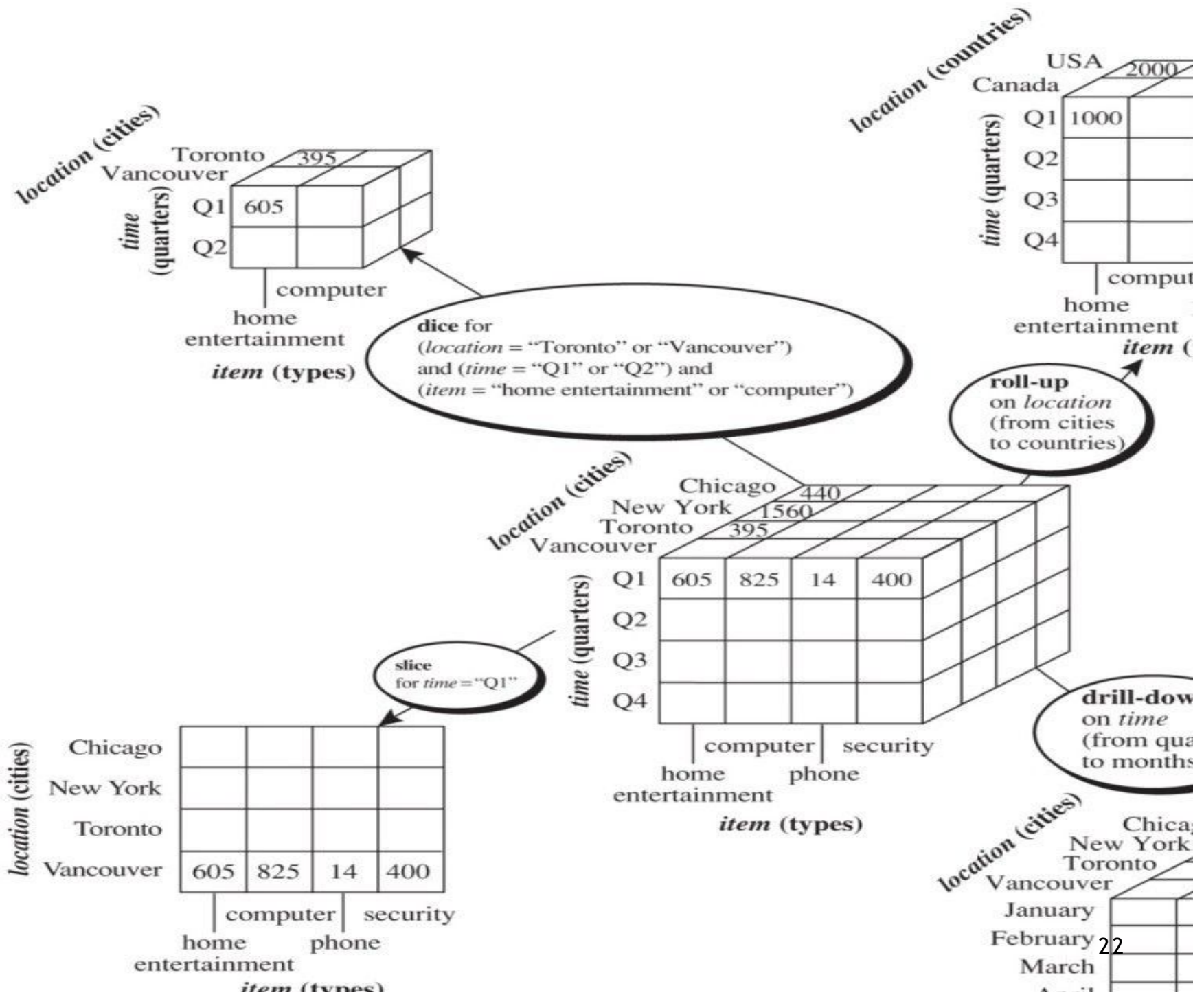
drill-down on time (from quarters to months)

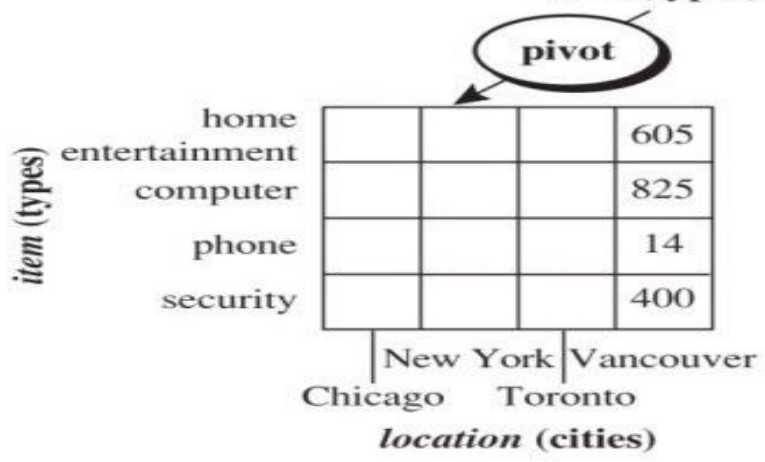
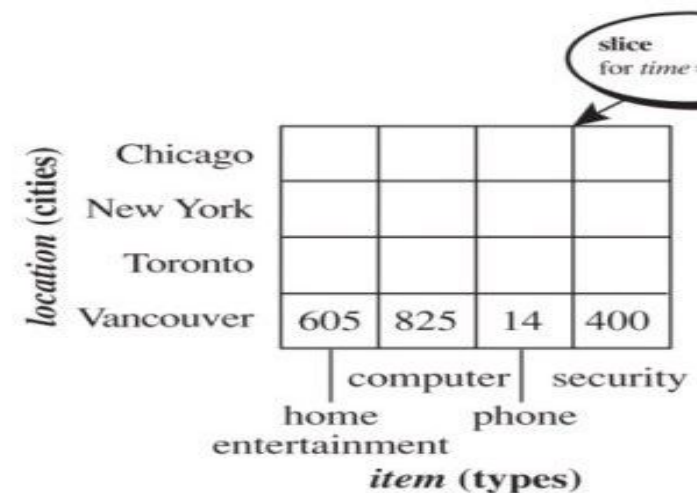
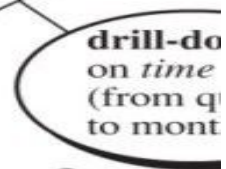
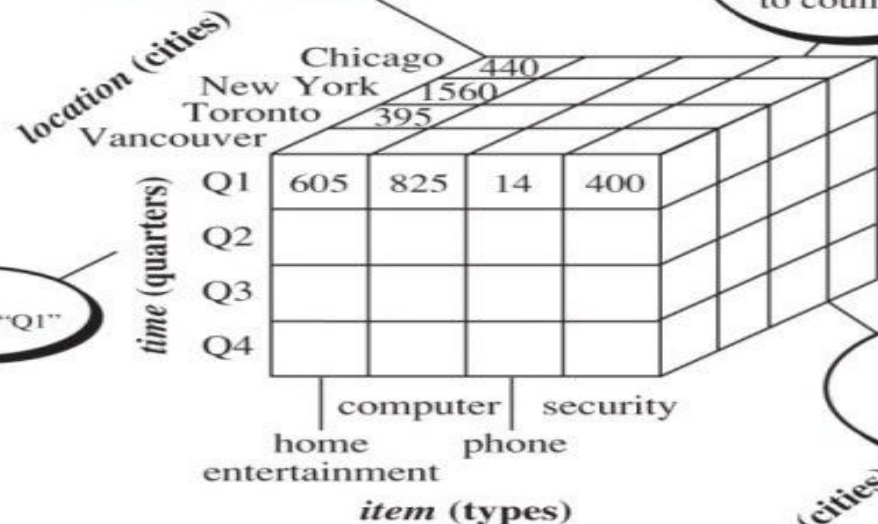
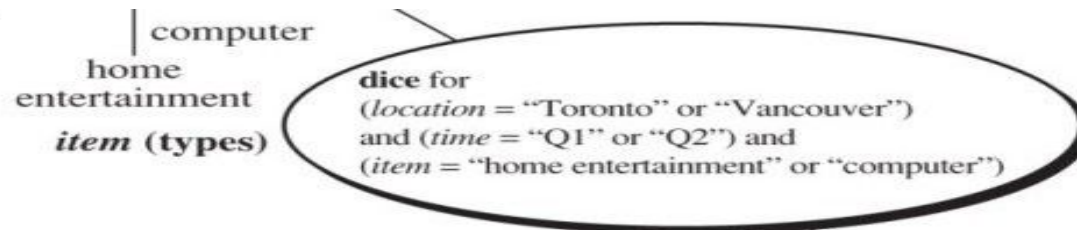
400
security

location (cities)

time (months)

	Chicago			
	New York			
	Toronto			
	Vancouver			
January				150
February				100
March				150
April				
May				
June				
July				
August				
September				
October				
November				
December				
	computer	security	home entertainment	phone



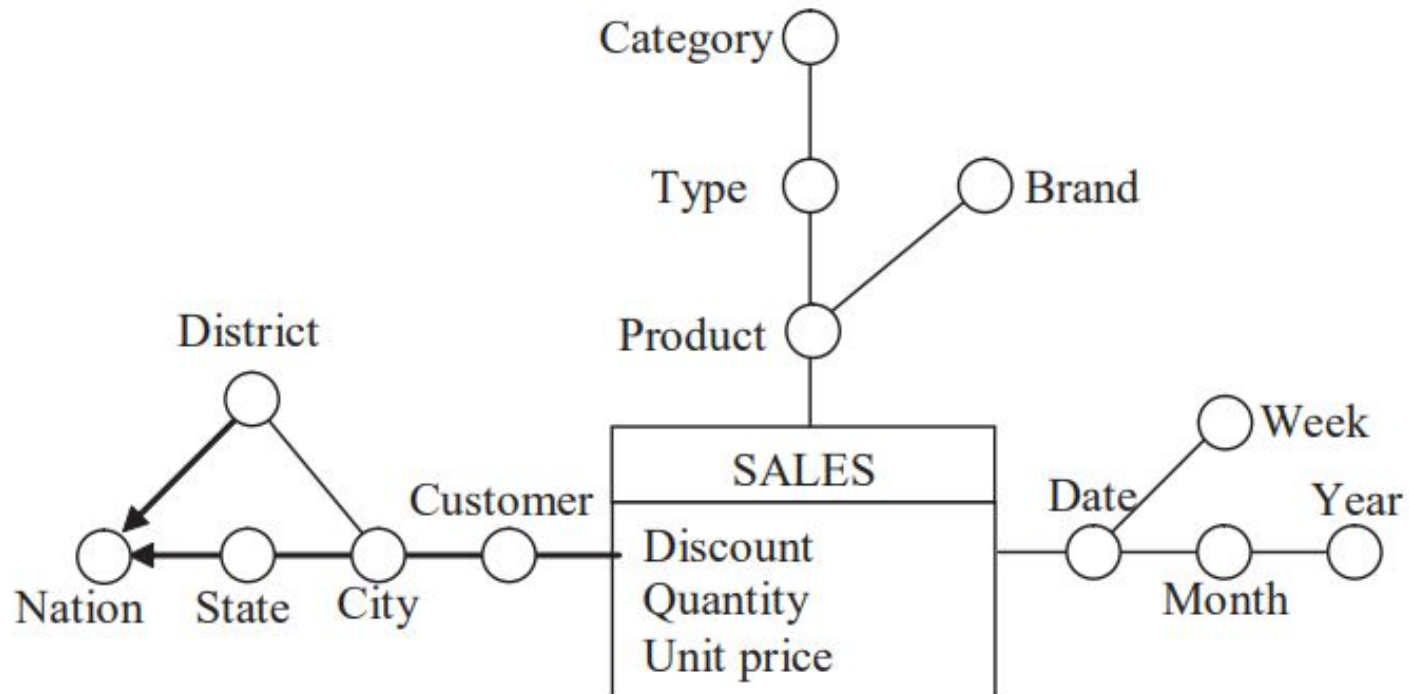


# Conceptual Fact Modeling

- The Star schema is to be considered a «logical» modeling of Facts
- It is possible to use «conceptual» modeling also for Facts
- It is possible to use E-R as a Fact conceptual model
- We will focus on «Dimensional Fact Modeling» (DFM) by Golfarelli and Rizzo (more suited)

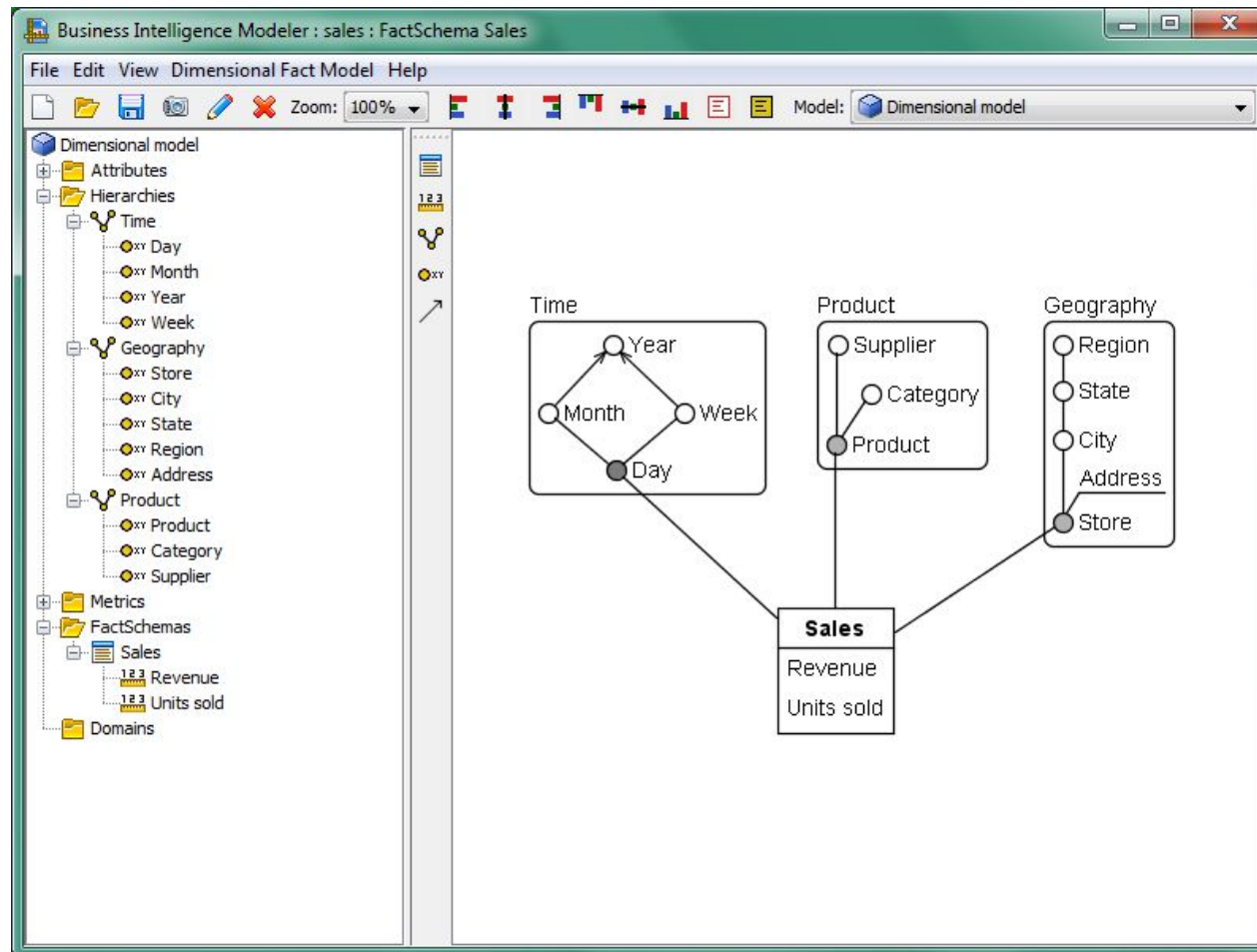
# Example

*A basic fact schema for the SALES fact*





# CASE Tool: BIModeler



<http://www.bimodeler.com/>

## References

1. Ponniah, P. **Data Warehousing Fundamentals for IT Professionals**, 2nd Edition, wiley, 2010, ISBN: 978-0-470-46207-2
2. Golfarelli, M. (2009). **DFM as a Conceptual Model for Data Warehouse**. In Encyclopedia of Data Warehousing and Mining, Second Edition (pp. 638-645). IGI Global.

Italian mother-tongue could prefer

1. Pighin, M., Marzona, A., **Sistemi informativi aziendali - ERP e sistemi di data analysis**, 3a edizione, Pearson, ISBN: 9788891907677.