数据挖掘导论

Introduction to Data Mining

Data Warehouse

数据智能实验室
DATA INTELLIGENCE LABORATORY
DI-LAB
浙江大学
Zhejiang University

Chapter 4: Data Warehousing and On-line Analytical Processing

- Basic Concepts
- Data Modeling
- OLAP Operations
- Design & Implementation
- Summary

A Typical Scenario

- A large company, with numerous branches, whose managers want to evaluate the contribution of each branch to the overall business performance of the company

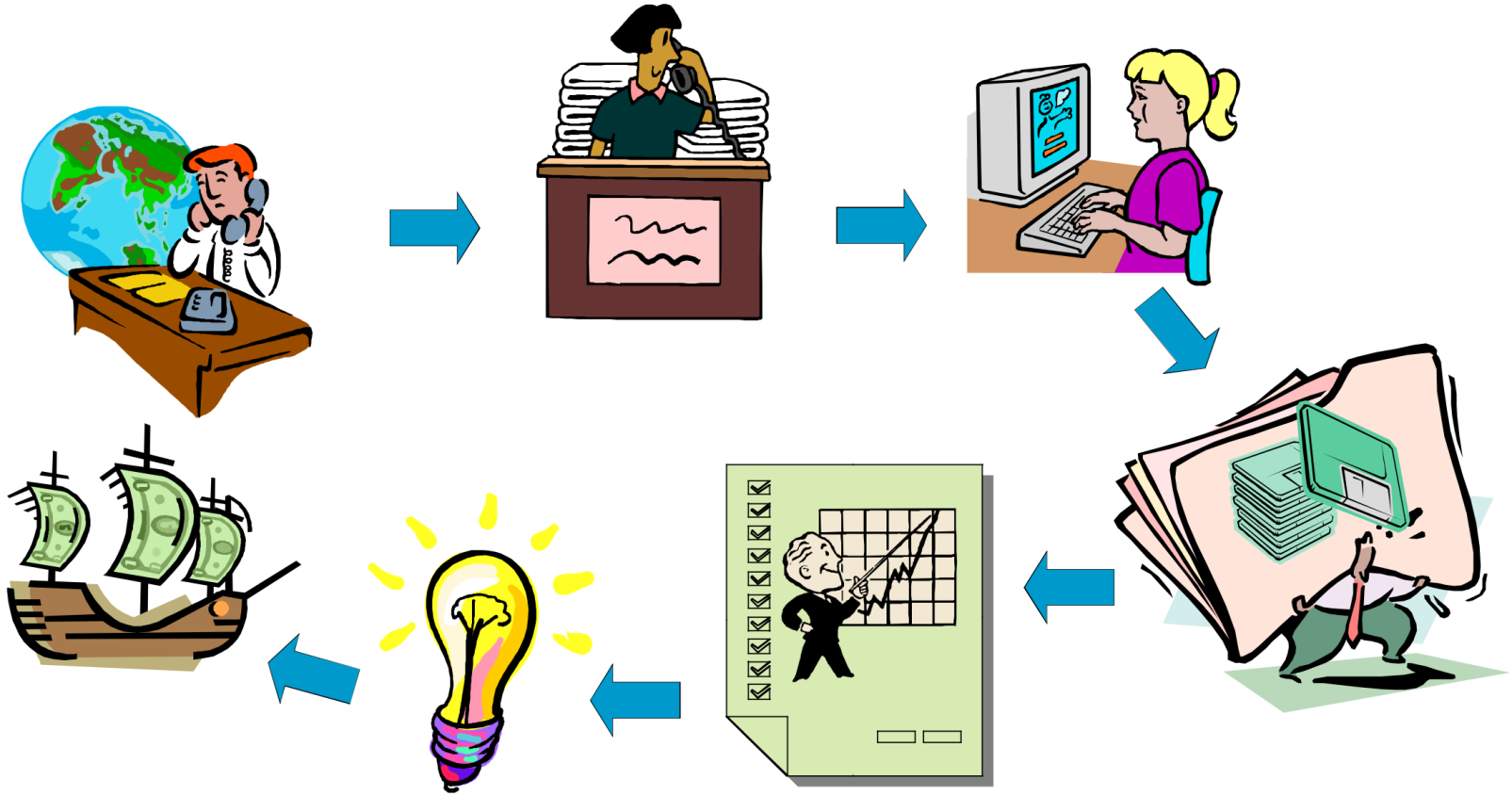
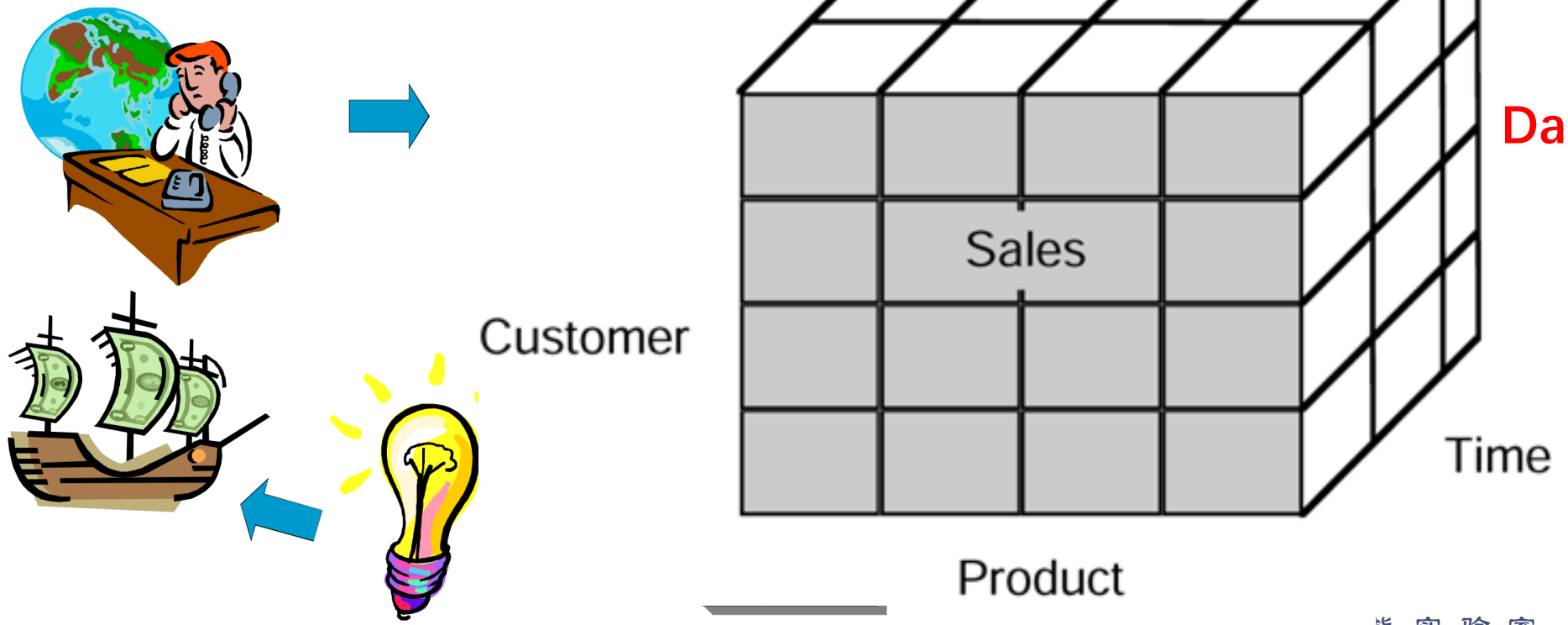


Image source: http://www.diag.uniroma1.it//~rosati/dmds-1516/datawarehousing.pdf

A Typical Scenario

An information repository that integrates and reorganizes data collected from sources of various kind and makes them available for analysis and evaluations aimed at planning and decision-making



Data Warehouse

What is a Data Warehouse?

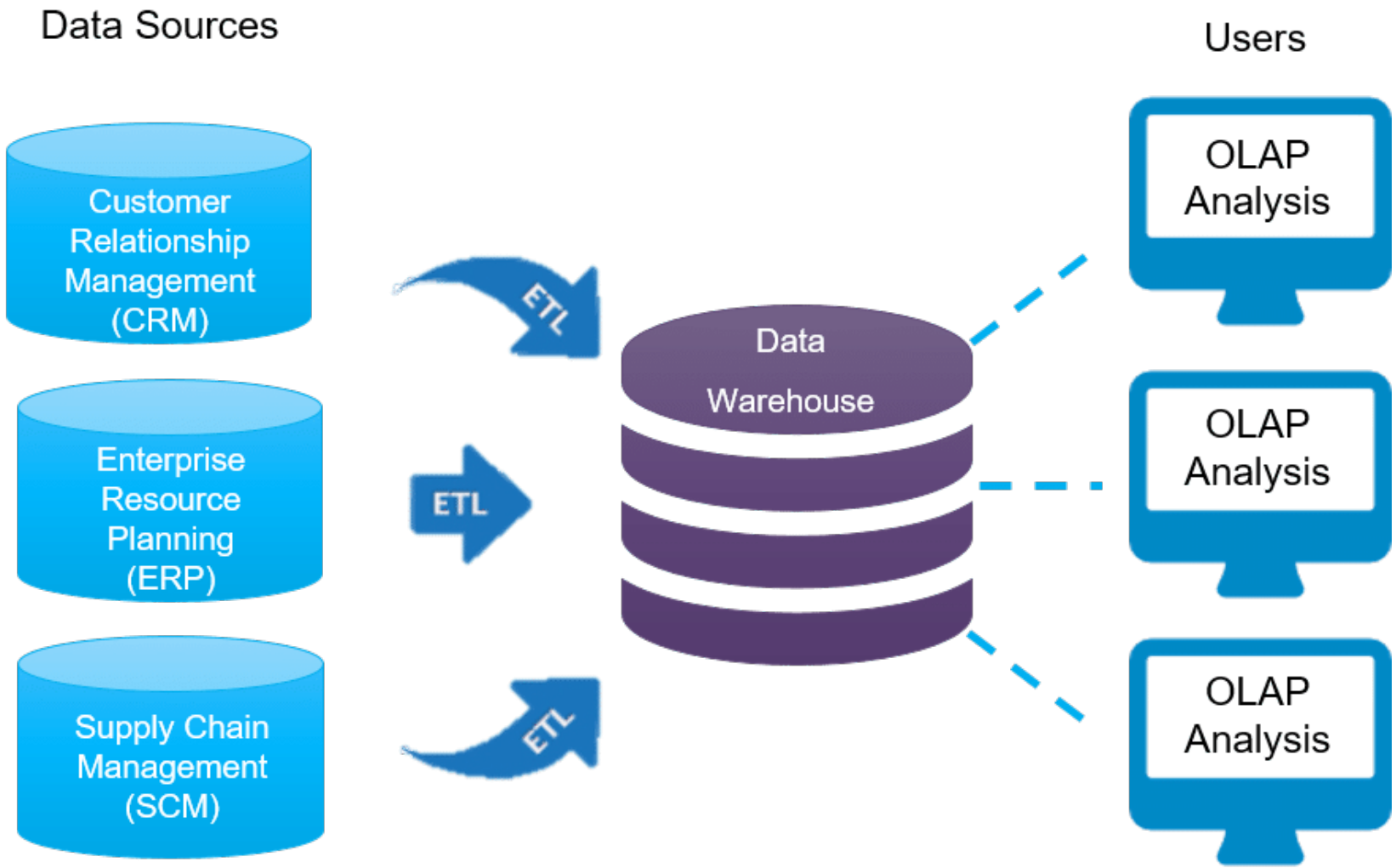


Image source: https://ibutechnologies.com/services/data-warehousing/

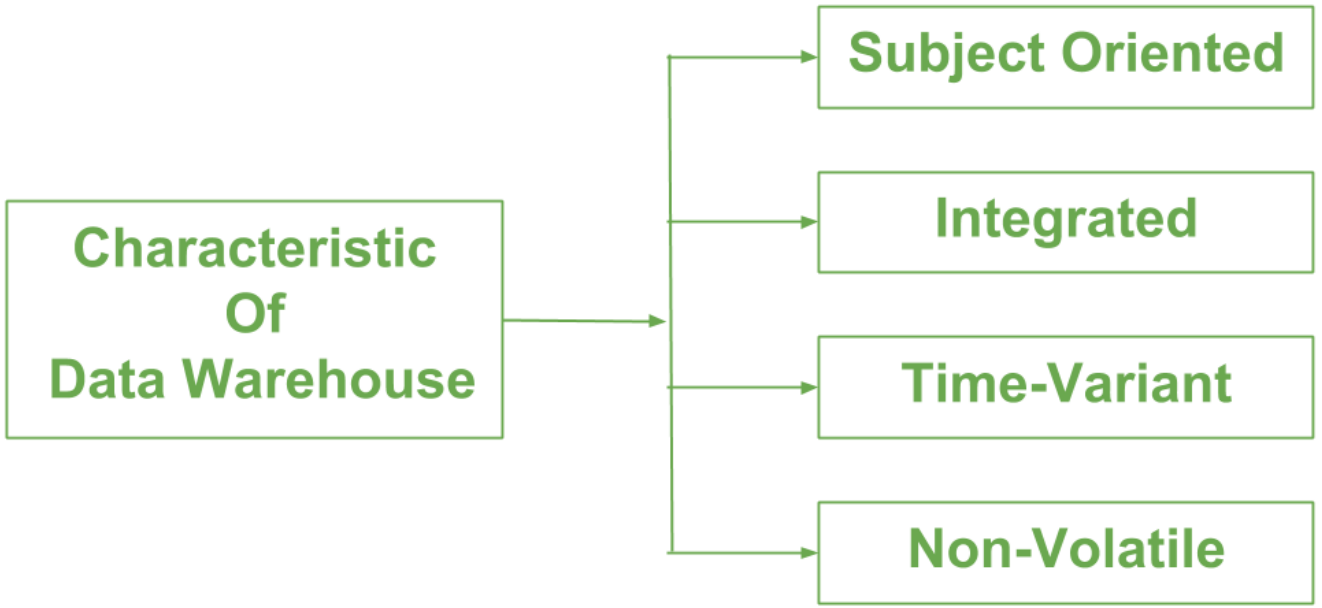
OLTP vs. OLAP

	OLTP (OnLine Transaction Processsing)	OLAP (OnLine Analytical Processing)
users	clerk, IT professional	knowledge worker
function	day to day operations	decision support
DB design	application-oriented	subject-oriented
data	current, up-to-date detailed, flat relational isolated	historical, summarized, multidimensional integrated, consolidated
usage	repetitive	ad-hoc
access	read/write index/hash on prim. key	lots of scans
unit of work	short, simple transaction	complex query

Features of Data Warehouse?

- "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

- Provide a simple and concise view around particular subject, such as customer, product, or sales

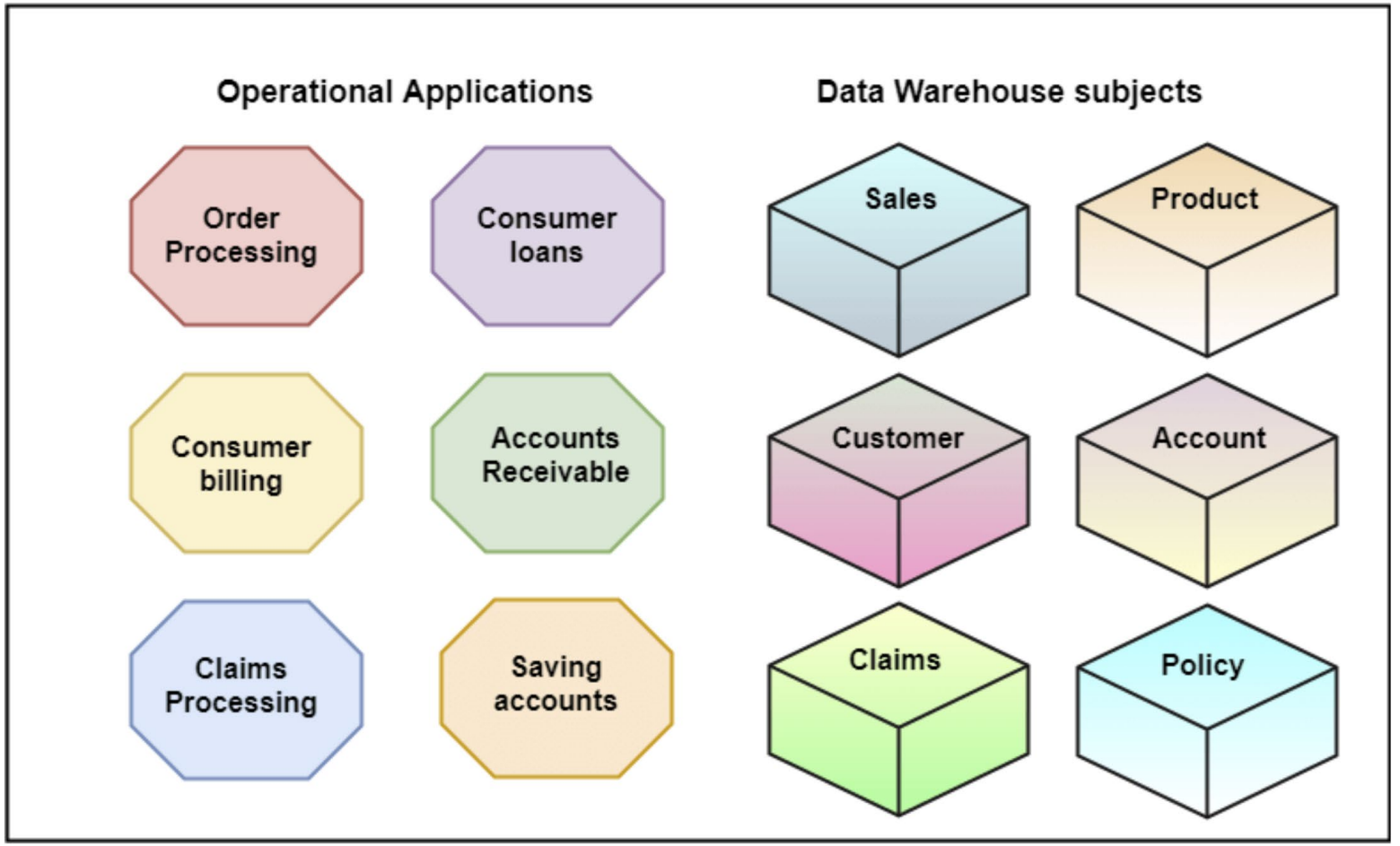
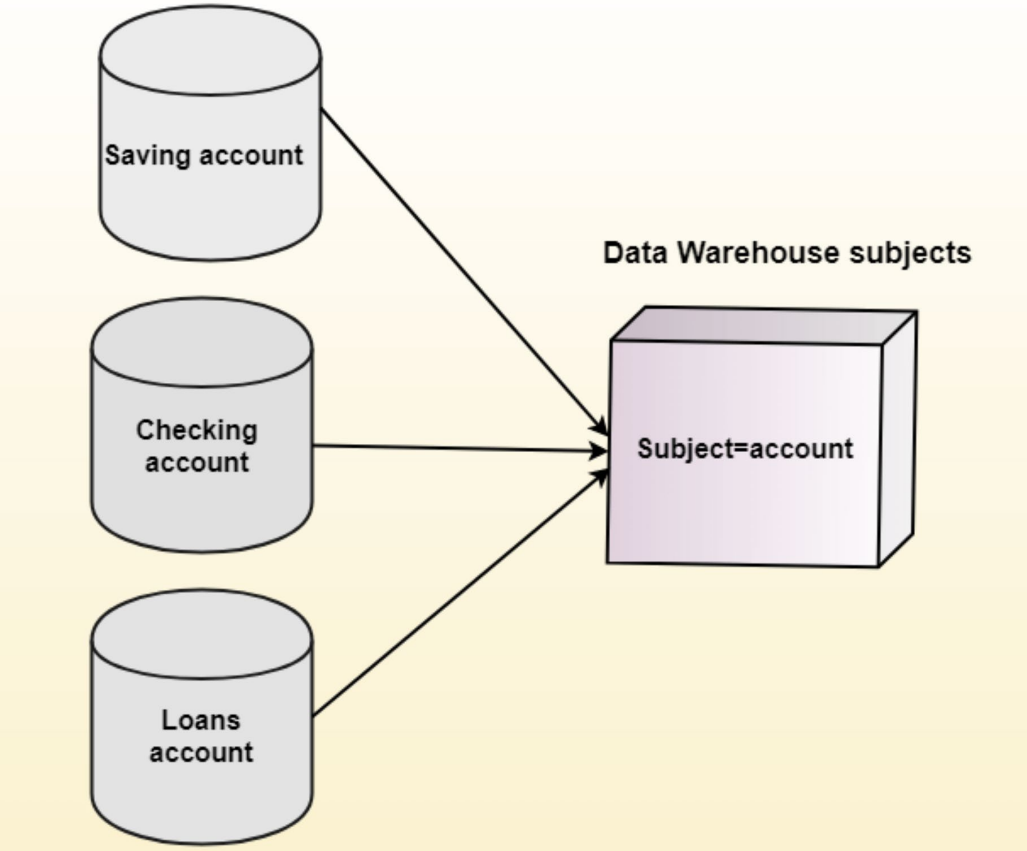


Image source: https://www.javatpoint.com/data-warehouse

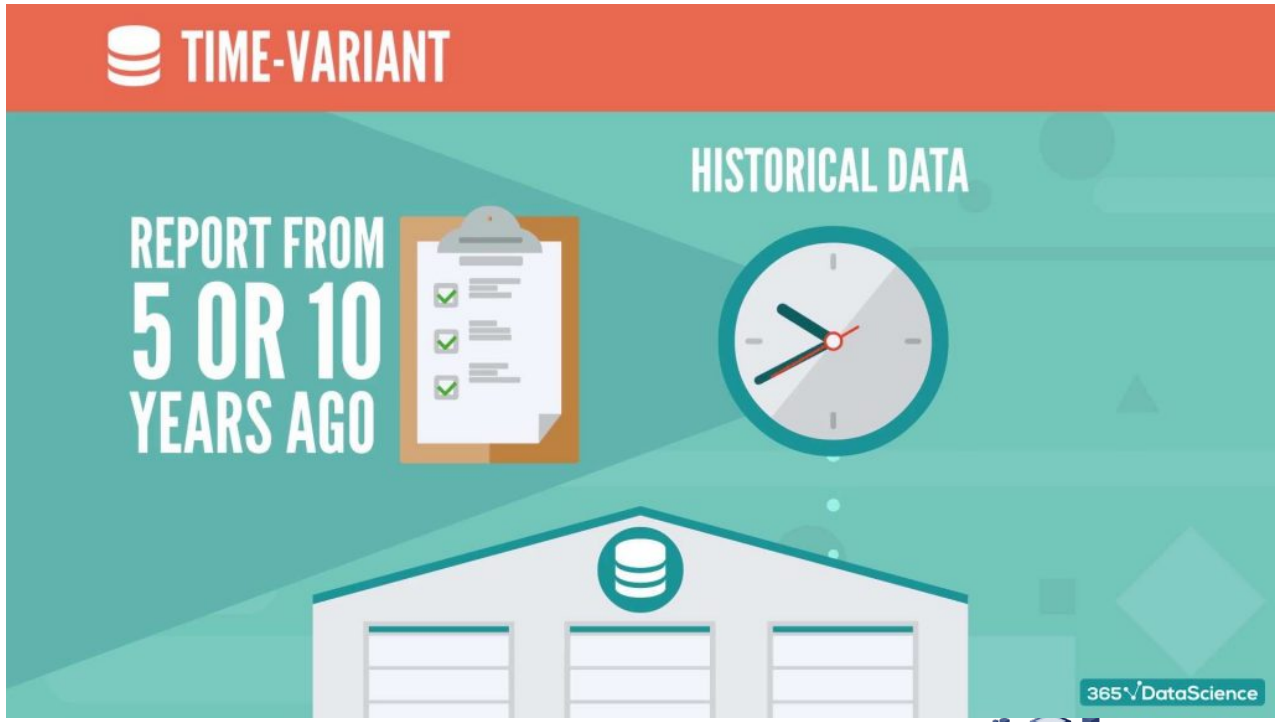
Data Warehouse is Integrated

- Constructed by integrating multiple, heterogeneous data sources
- Data cleaning and data integration techniques are applied



Data Warehouse is Time Variant

- **Historical information is kept in a data warehouse**
  - ✓ Operational database: current value data
  - ✓ Data warehouse data: provide information from a historical perspective (e.g., past 5-10 years)

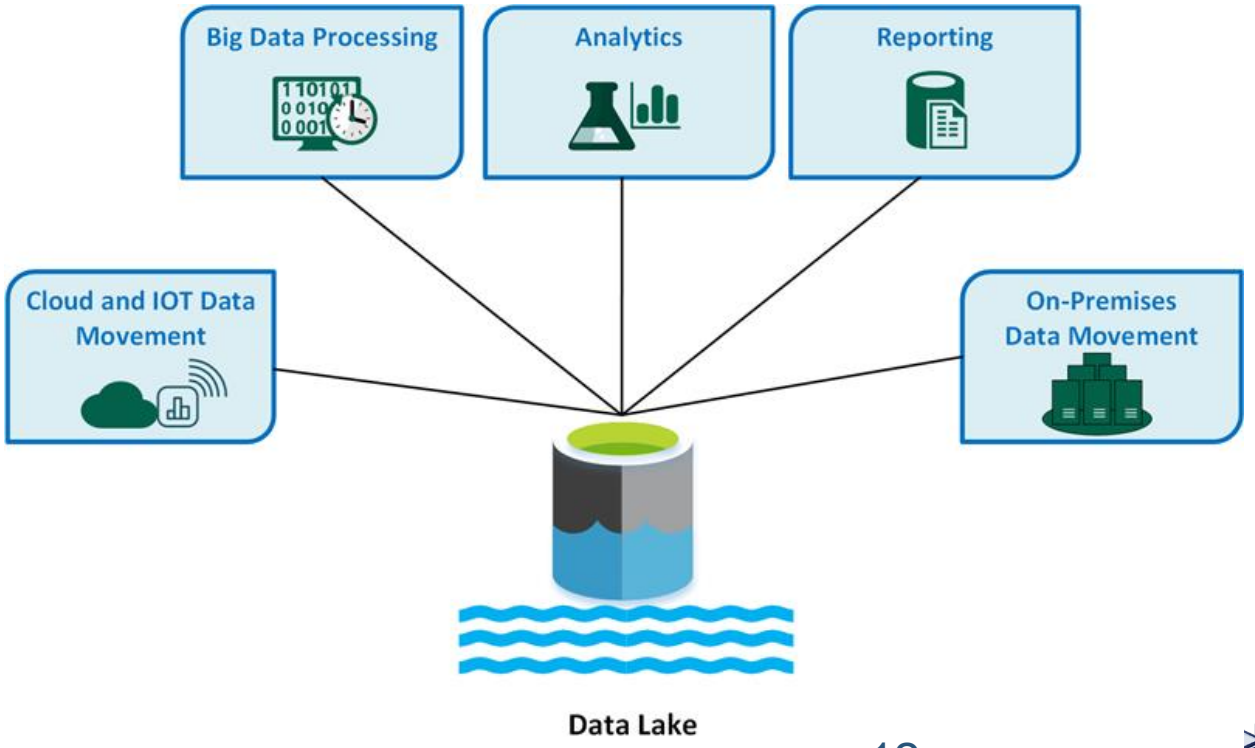


Data Warehouse is Non-volatile

- Non-volatile means the previous data is not erased when new data is added to it
- Operational update of data does not occur in the data warehouse environment
  - ✓ Requires only two operations accessing: initial loading of data and access of data

Data Lake

- **A data lake is a centralized repository that allows you to store all your structured and unstructured data at any scale**
  - ✓ You can store your data as-is, without having to first structure the data, and run different types of analytics—from dashboards and visualizations to big data processing, real-time analytics, and machine learning to guide better decisions.



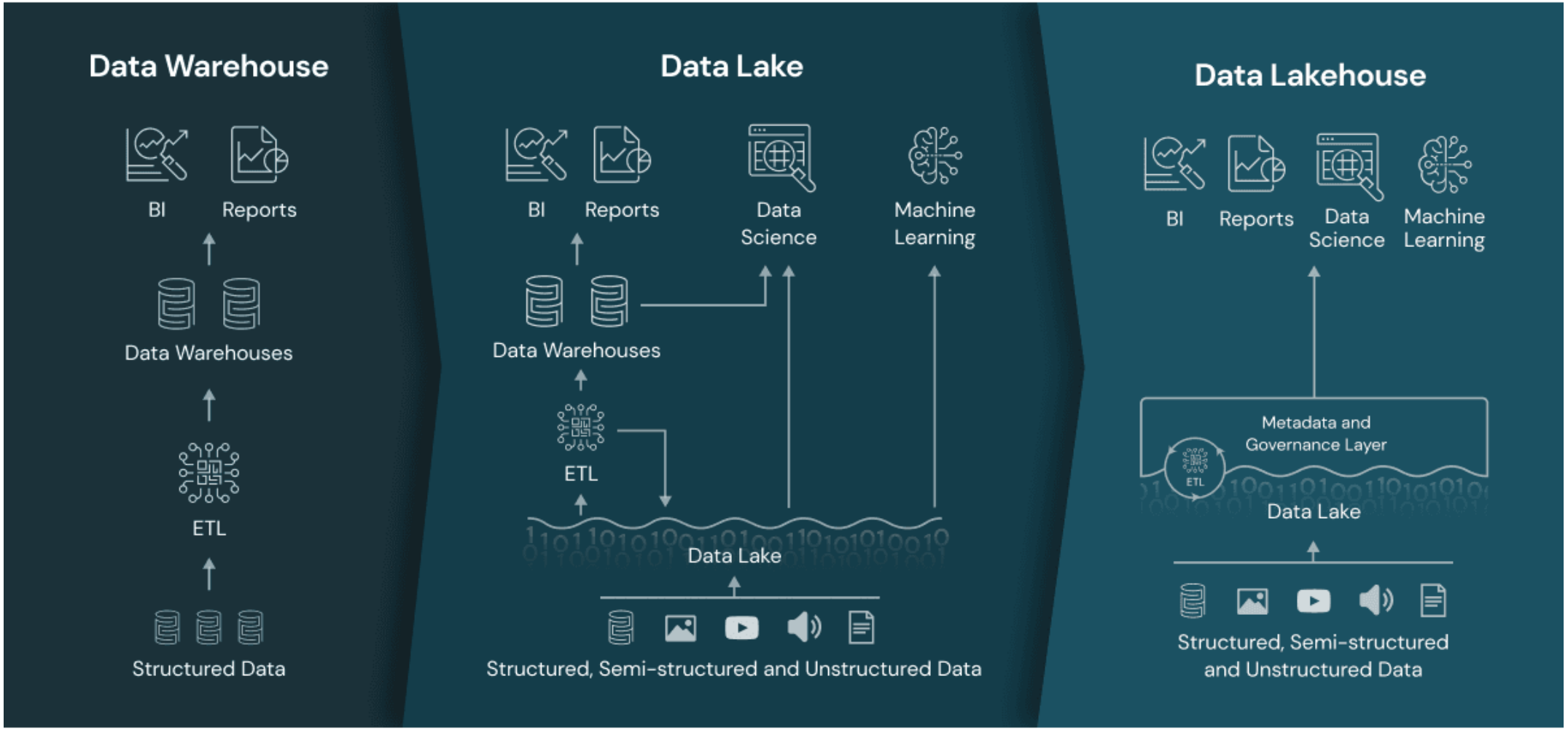
Data Lake v.s. Data Warehouse

Characteristics	Data Warehouse	Data Lake
Data	Relational from transactional systems, operational databases, and line of business applications	Non-relational and relational from IoT devices, web sites, mobile apps, social media, and corporate applications
Schema	Designed prior to the DW implementation (schema-on-write)	Written at the time of analysis (schema-on-read)
Price/Performance	Fastest query results using higher cost storage	Query results getting faster using low-cost storage
Data Quality	Highly curated data that serves as the central version of the truth	Any data that may or may not be curated (ie. raw data)
Users	Business analysts	Data scientists, Data developers, and Business analysts (using curated data)
Analytics	Batch reporting, BI and visualizations	Machine Learning, Predictive analytics, data discovery and profiling

Data Lakehouse（湖仓一体）

- A data lakehouse is a new, open data management architecture that combines the flexibility, cost-efficiency, and scale of data lakes with the data management and ACID transactions of data warehouses, enabling business intelligence (BI) and machine learning (ML) on all data.

Data Lakehouse（湖仓一体）

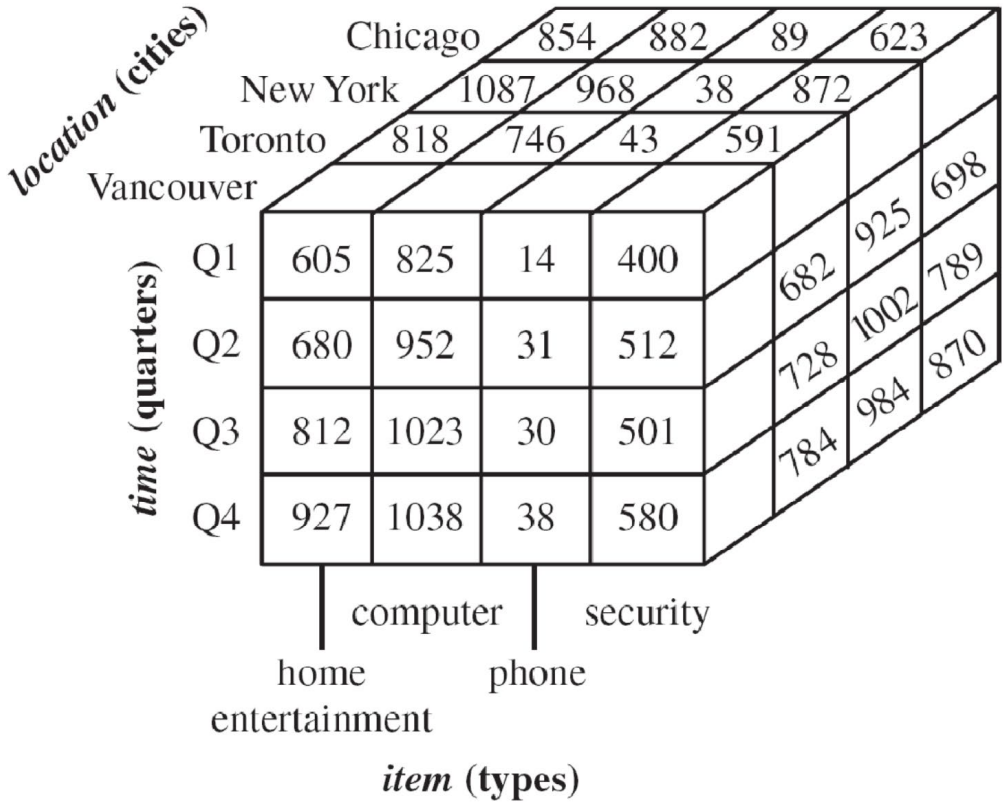


Chapter 4: Data Warehousing and On-line Analytical Processing

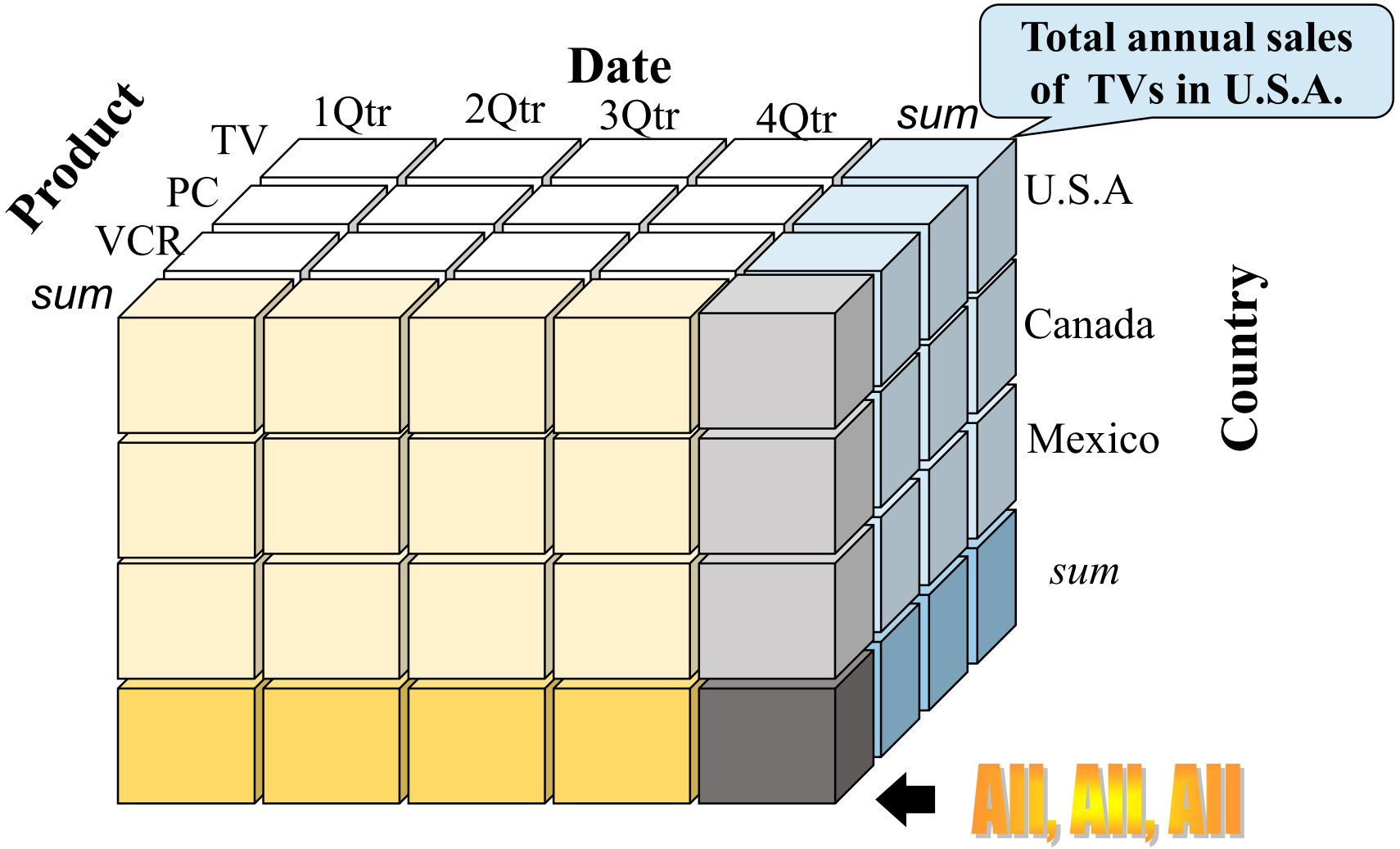
- Basic Concepts
- Data Modeling
- OLAP Operations
- Design & Implementation
- Summary

Data Cube

- A data warehouse is based on a **multidimensional data model** (a.k.a. data cube)
- A data cube allows data to be modeled and viewed in multiple dimensions



Another Example of Data Cube



Types of Aggregator

- **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()
- **Holistic:**
  - ✓ if there is no constant bound on the storage size needed to describe a subaggregate
    - E.g., median(), mode(), rank()

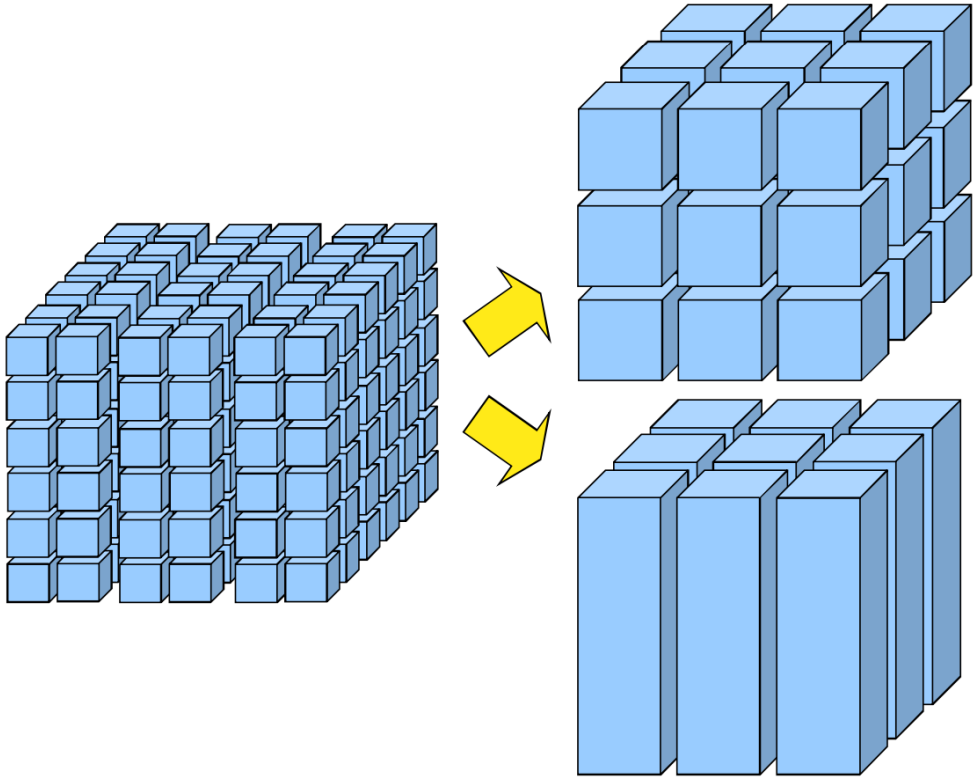
(province, item) + max aggregator + Map Visualization

Chapter 4: Data Warehousing and On-line Analytical Processing

- Basic Concepts
- Data Modeling
- OLAP Operations
- Design & Implementation
- Summary

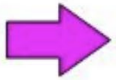
Roll-up

- The roll-up operator causes an increase in data aggregation or removes a detail level from an attribute hierarchy



Roll-up

sale	prodId	storeId	date	amt
	p1	c1	1	12
	p2	c1	1	11
	p1	c3	1	50
	p2	c2	1	8
	p1	c1	2	44
	p1	c2	2	4



sale	prodId	date	amt
	p1	1	62
	p2	1	19
	p1	2	48

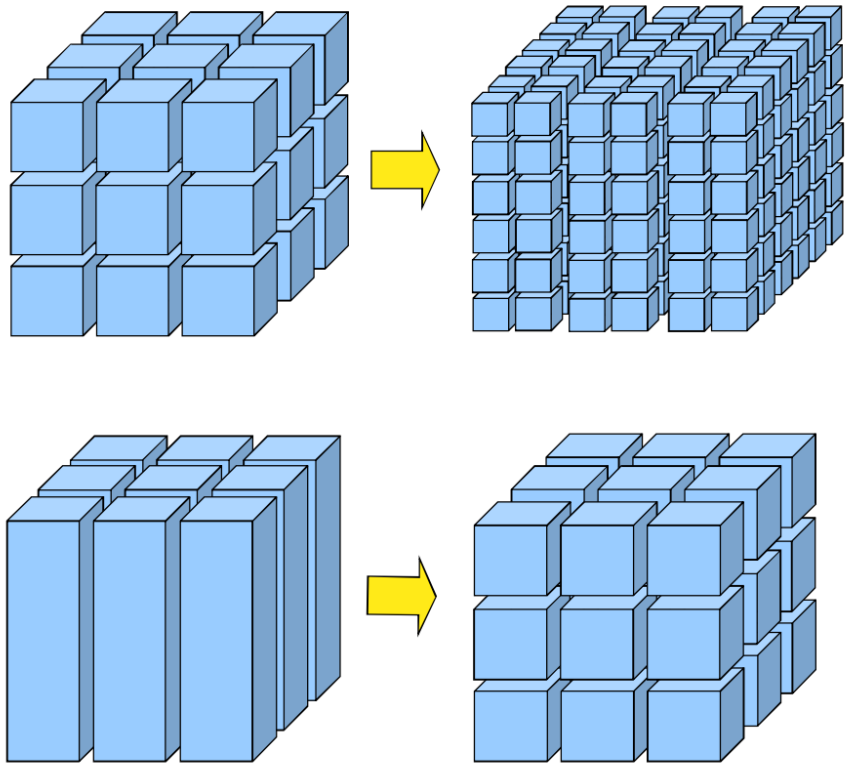
rollup →

← drill-down

Image source: https://slideplayer.com/slide/10455148/

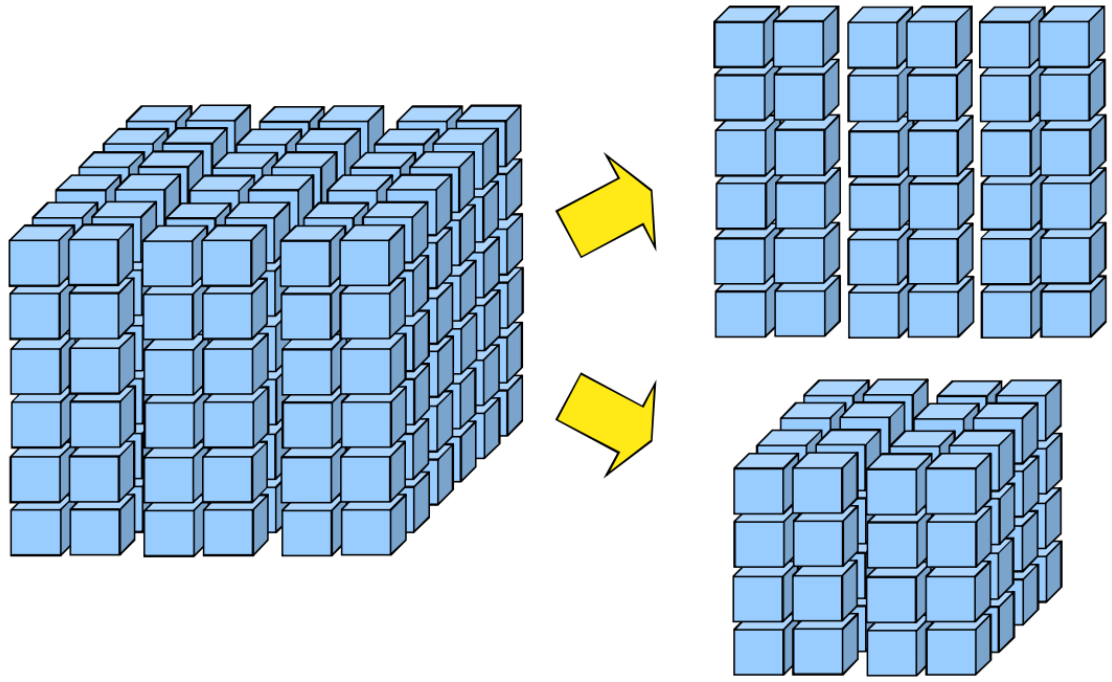
Drill-down

- It is the inverse of roll-up. We use it to reduce the level of aggregation in the data



Slice and Dice

- It produces “sub-cubes” starting from cubes.
- Slicing: reduces the number of cube dimensions after setting one of the dimension
- Dicing: reduces the set of data being analyzed by a selection criterion



Slice and Dice

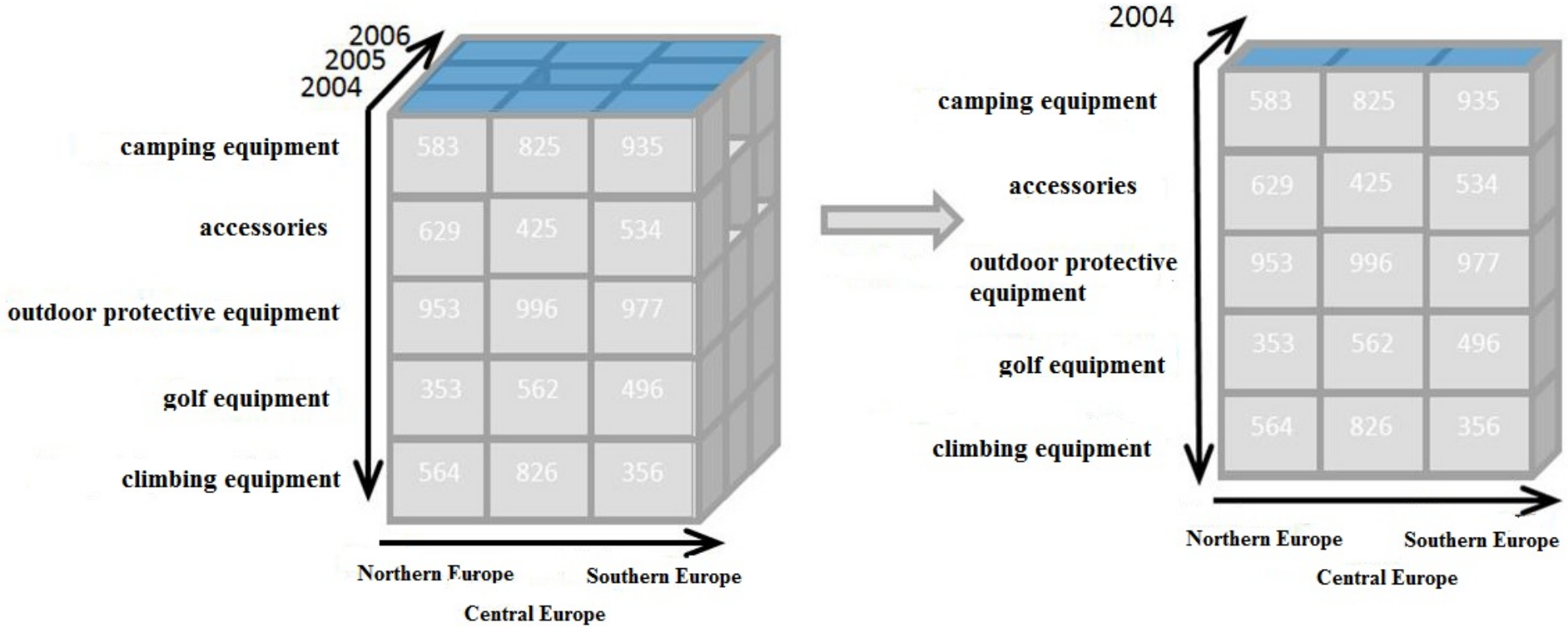


Image source: https://pediaa.com/what-is-the-difference-between-slice-and-dice-in-data-warehouse/

Slice and Dice

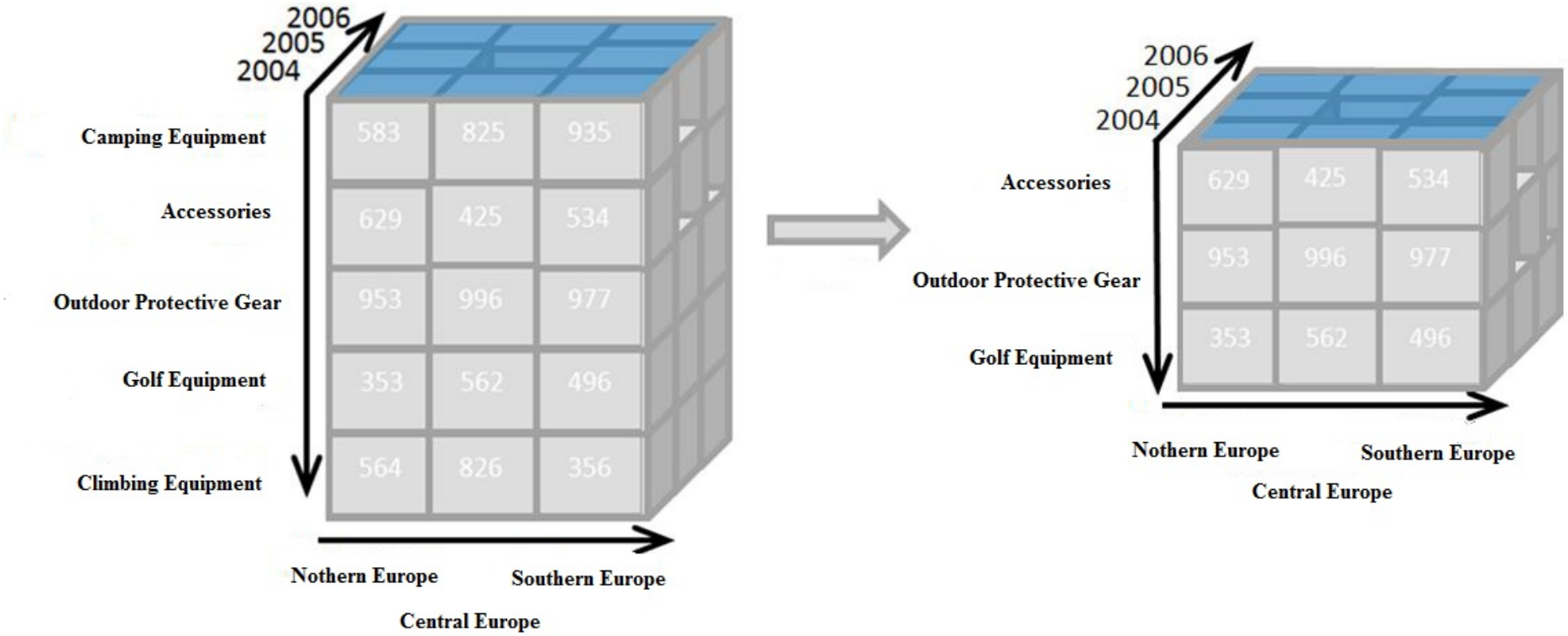
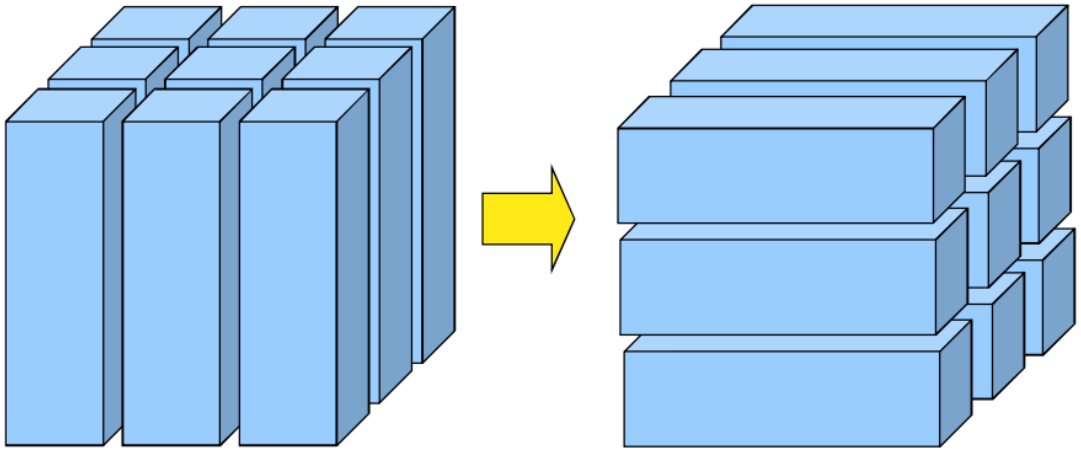
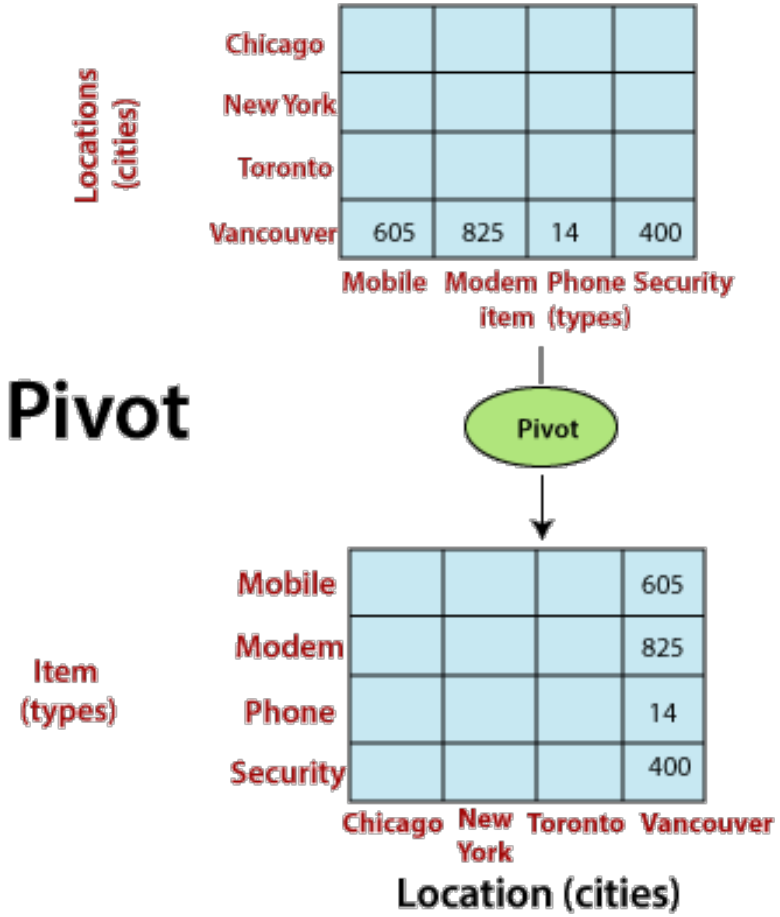


Image source: https://pediaa.com/what-is-the-difference-between-slice-and-dice-in-data-warehouse/

Pivoting

- It is used to change the structure of the cube, that is, when a different view of the data in the cube is needed.



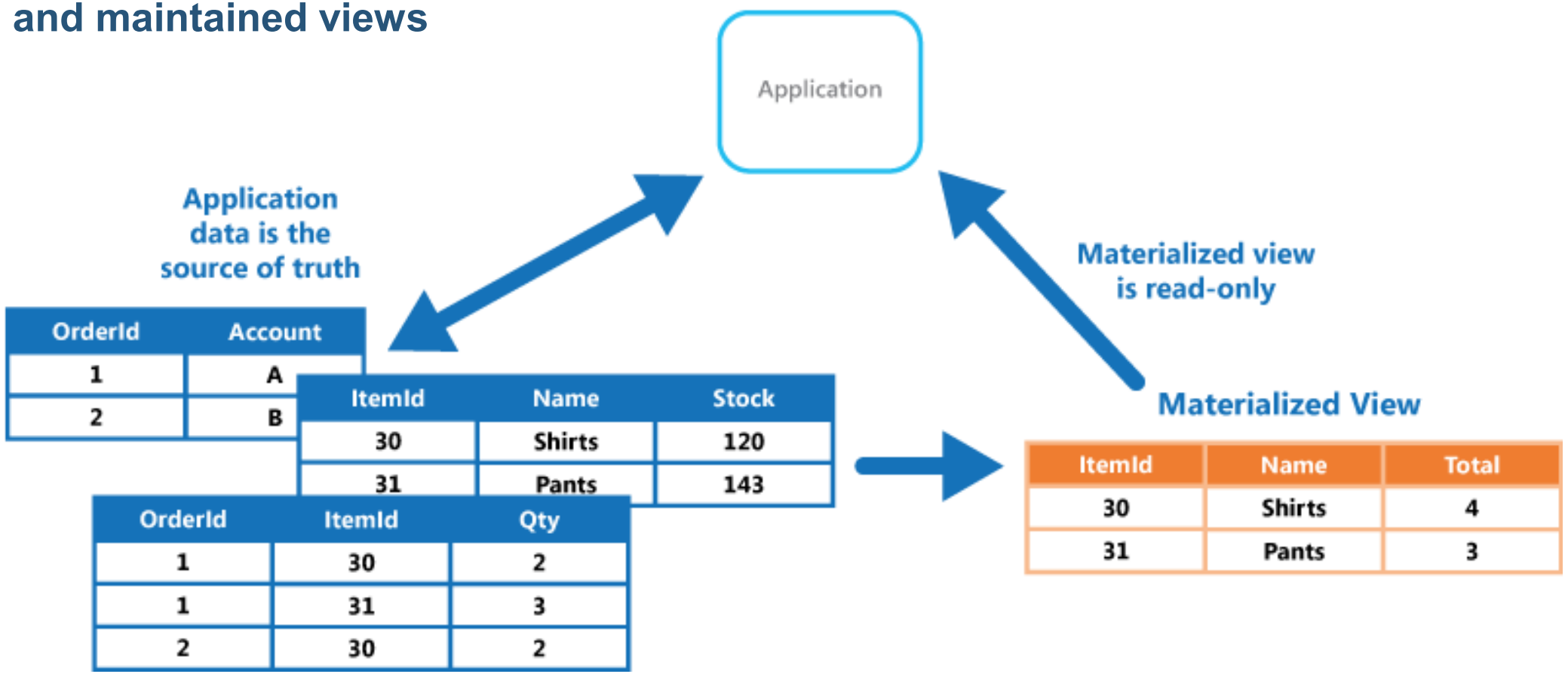


Chapter 4: Data Warehousing and On-line Analytical Processing

- Basic Concepts
- Data Modeling
- OLAP Operations
- Design & Implementation
- Summary

View Materialization

- Warehouses can be thought of as a collection of asynchronously replicated tables and maintained views



OrderId	Account
1	A
2	B

ItemId	Name	Stock
30	Shirts	120
31	Pants	143

OrderId	ItemId	Qty
1	30	2
1	31	3
2	30	2

ItemId	Name	Total
30	Shirts	4
31	Pants	3

View Materialization

- **Maintenance policy: Controls when we do refresh**
  - ✓ Immediate: As part of the transaction that modifies the underlying data tables
  - ✓ Deferred: Some time later, in a separate transaction

Types of Deferred Materialization

- **Lazy**
  - ✓ Delay refresh until next query on view; then refresh before answering the query (slows down queries than updates)
- **Periodic**
  - ✓ Refresh periodically (e.g. once in a day)
  - ✓ Queries possibly answered using outdated version of view tuples
  - ✓ Widely used, especially for asynchronous replication in distributed databases, and for warehouse applications
- **Event-based or Forced**
  - ✓ E.g., Refresh after a fixed number of updates to underlying data tables

Data Cube Materialization

Data cube can be viewed as a lattice of cuboids

- ✓ The bottom-most cuboid is the base cuboid
- ✓ The top-most cuboid (apex) contains only one cell
- ✓ How many cuboids in an n-dimensional cube with L levels?

$$T = \prod_{i=1}^{n} (L_i + 1)$$

Materialization of data cube

- ✓ Materialize every (cuboid) (full materialization), none (no materialization), or some (partial materialization)
- ✓ Selection of which cuboids to materialize
  - Based on size, sharing, access frequency, etc.

Efficient Processing OLAP Queries

- **Determine which operations should be performed on the available cuboids**
  - ✓ Transform drill, roll, etc. into corresponding SQL and/or OLAP operations, e.g., dice = selection + projection
- **Determine which materialized cuboid(s) should be selected for OLAP op.**
  - ✓ Let the query to be processed be on {brand, province_or_state} with the condition "year = 2004", and there are 4 materialized cuboids available:
    1) {year, item_name, city}
    2) {year, brand, country}
    3) {year, brand, province_or_state}
    4) {item_name, province_or_state} where year = 2004

  Which should be selected to process the query?

Indexing OLAP Data: Bitmap Index

- **OLAP queries are typically aggregate queries**
  - ✓ Precomputation is essential for interactive response times
- **Index on a particular column**

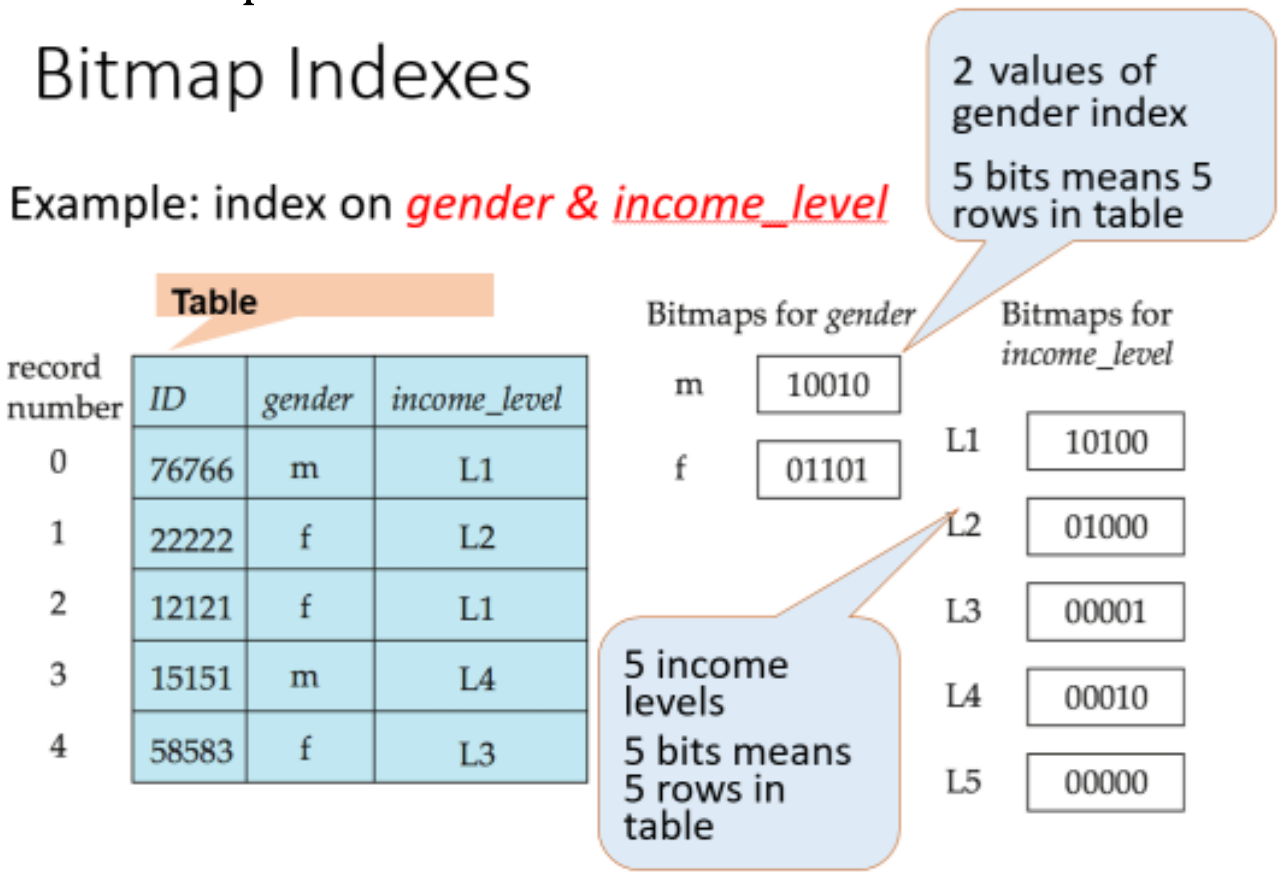


Image source: https://mjromeo81.com/2017/01/28/introduction-to-bitmap-indexes/

Chapter 4: Data Warehousing and On-line Analytical Processing

- Basic Concepts
- Data Modeling
- OLAP Operations
- Design & Implementation
- Summary

Summary

- **Basic Concepts**
  - ✓ Subject-oriented, integrated, time-variant, non-viotile
- **Data Modeling**
  - ✓ Data Cube
- **OLAP Operations**
  - ✓ Drilling, rolling, slicing, dicing and pivoting
- **Design & Implementation**
  - ✓ Efficiency issues (view materialization, bitmap index)
  - ✓ SAP Data Warehouse Cloud