ADVANCED DATABASE

Data warehouse

Dr. NGUYEN Hoang Ha

Email: nguyen-hoang.ha@usth.edu.vn

Contents

I. Context

II. DW concepts

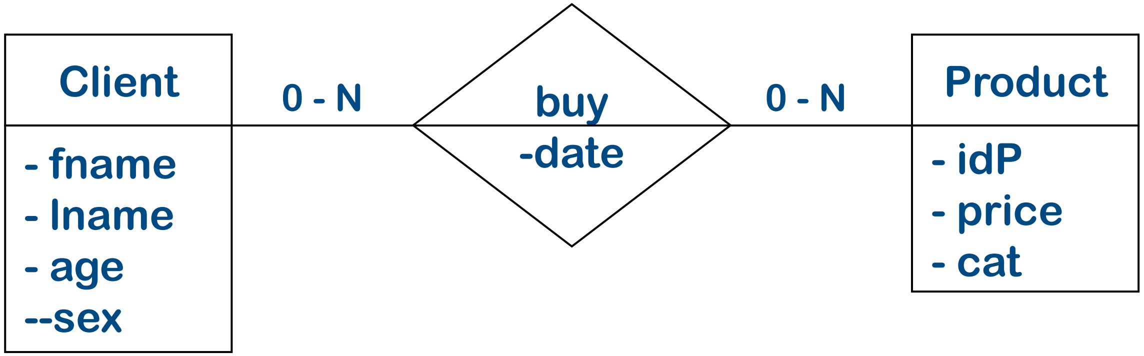
III. DW architecture

IV. Data mart

V. Data warehouse queries

Modelization of a DB

- **Products sold to clients**



- **Record every sale without aggregation (e.g., by month, by male client, ...)**
- **Important = Non redundant/consistency/efficiency**

Not suitable for datawarehouse

Modelization Entity/Relationship

- Advantages:
  - Normalization (redundancy/consistency)
  - Optimization of transactions
  - Reduce the storage space

- Disadvantages for a manager:
  - Schema too complete:
    - Tables/column not useful for analysis
  - No graphical interface to use the E/R schema
  - Not suitable for analysis

Context

- A manager want to knows

Who are my best clients?

Why and how sales have evolved?

Which French people like fish?

What is the amount of my sales by day?

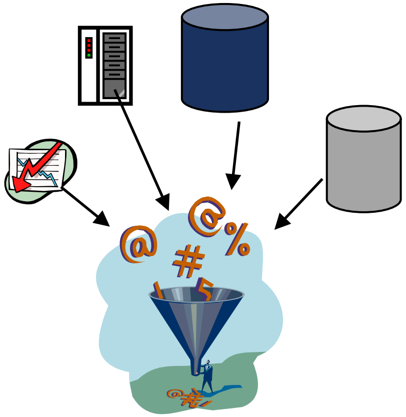
Data is everywhere yet BUT

- I can't find the data I need
  - data is scattered over the network
  - many versions, subtle differences

- I can't get the data I need
  - need an expert to get the data
- I can't understand the data I found
  - available data poorly documented
- I can't use the data I found
  - results are unexpected
  - data needs to be transformed from one form to other

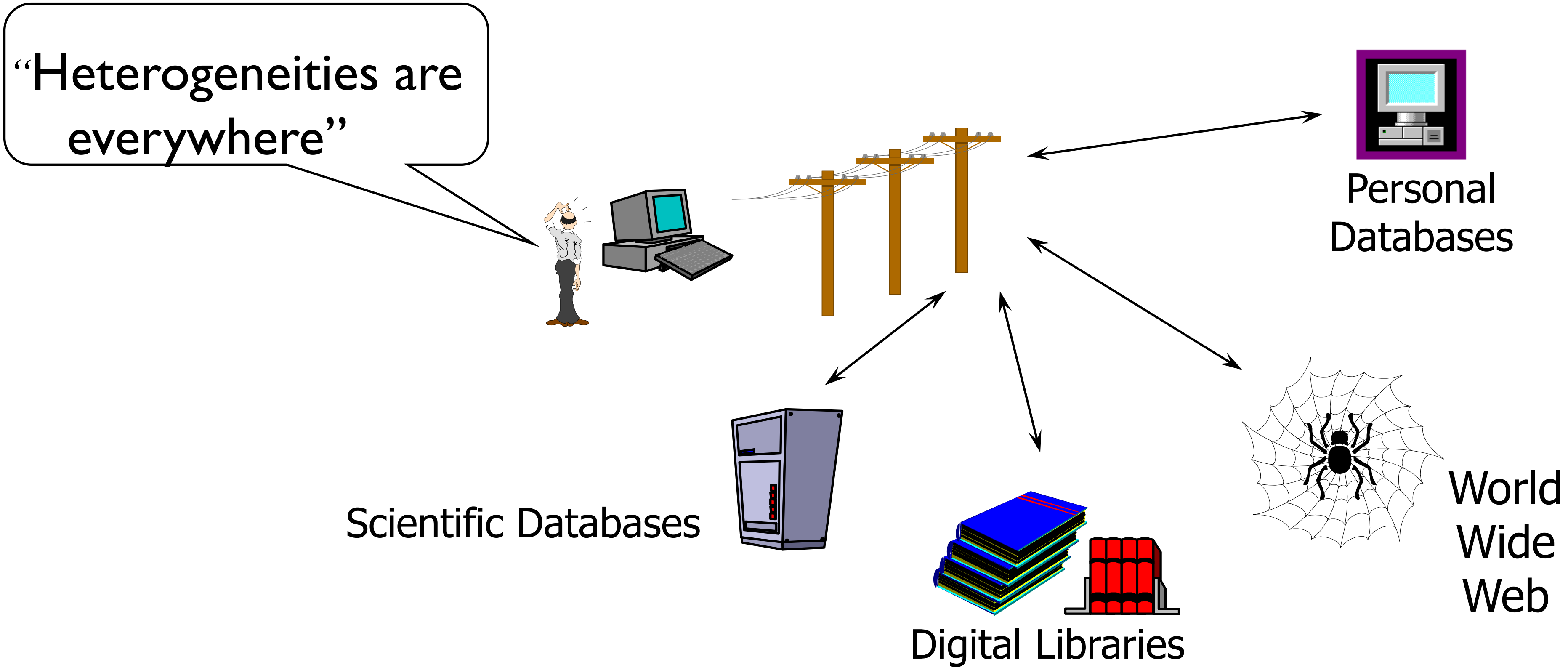
Available data

- Operational data
  - Databases (Oracle, SQL Server)
  - Files (XML, Excel, HTML, …)
  - …



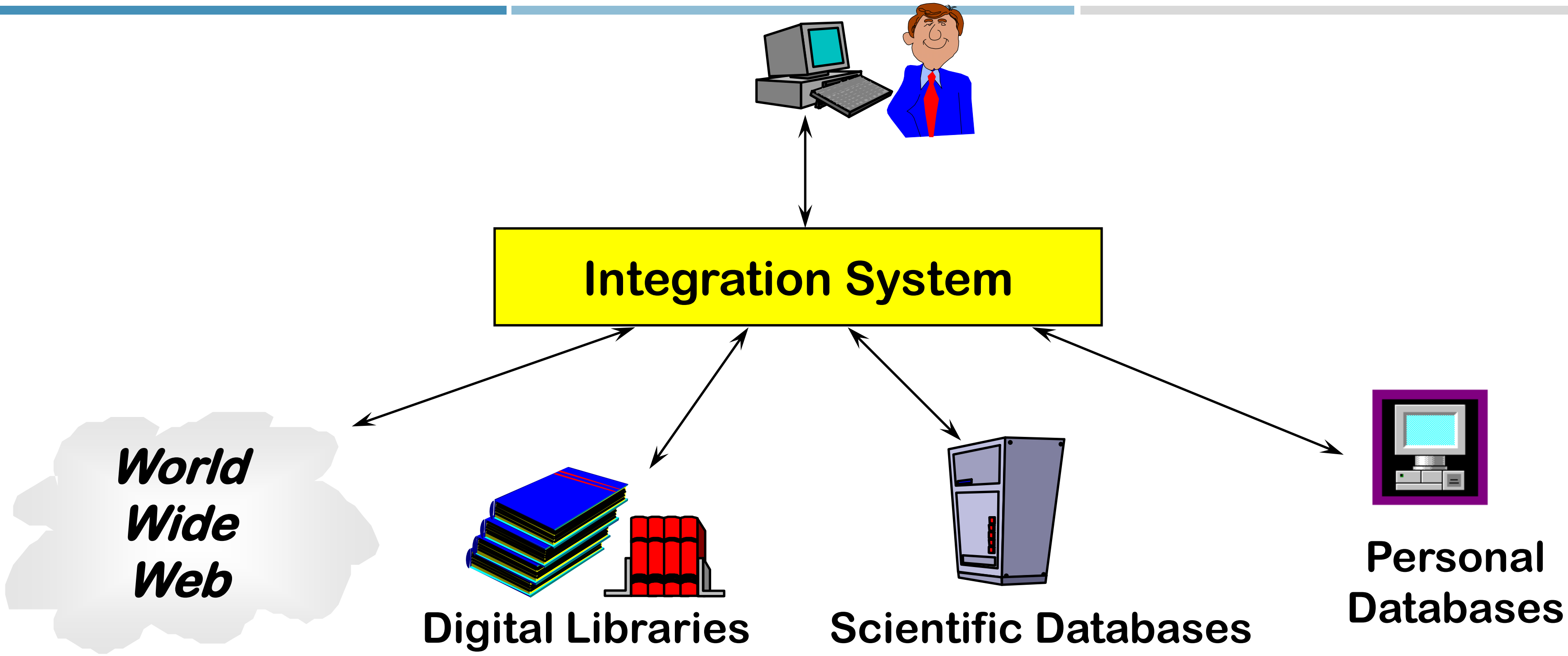
- Characteristics :
  - Distributed
  - Heterogeneous (different data structures)
  - Detailed (often too detailed for analysis)
  - Not adapted for analysis (the production must not be blocked)
  - No time information

Problem: Heterogeneous Sources



- Different interfaces
- Different data representations
- Duplicate and inconsistent information

Goal: Unified Access to Data



- Collects and combines information
- Provides integrated view, uniform user interface
- Supports sharing

DATA WAREHOUSE CONCEPTS

What is Data Warehousing?

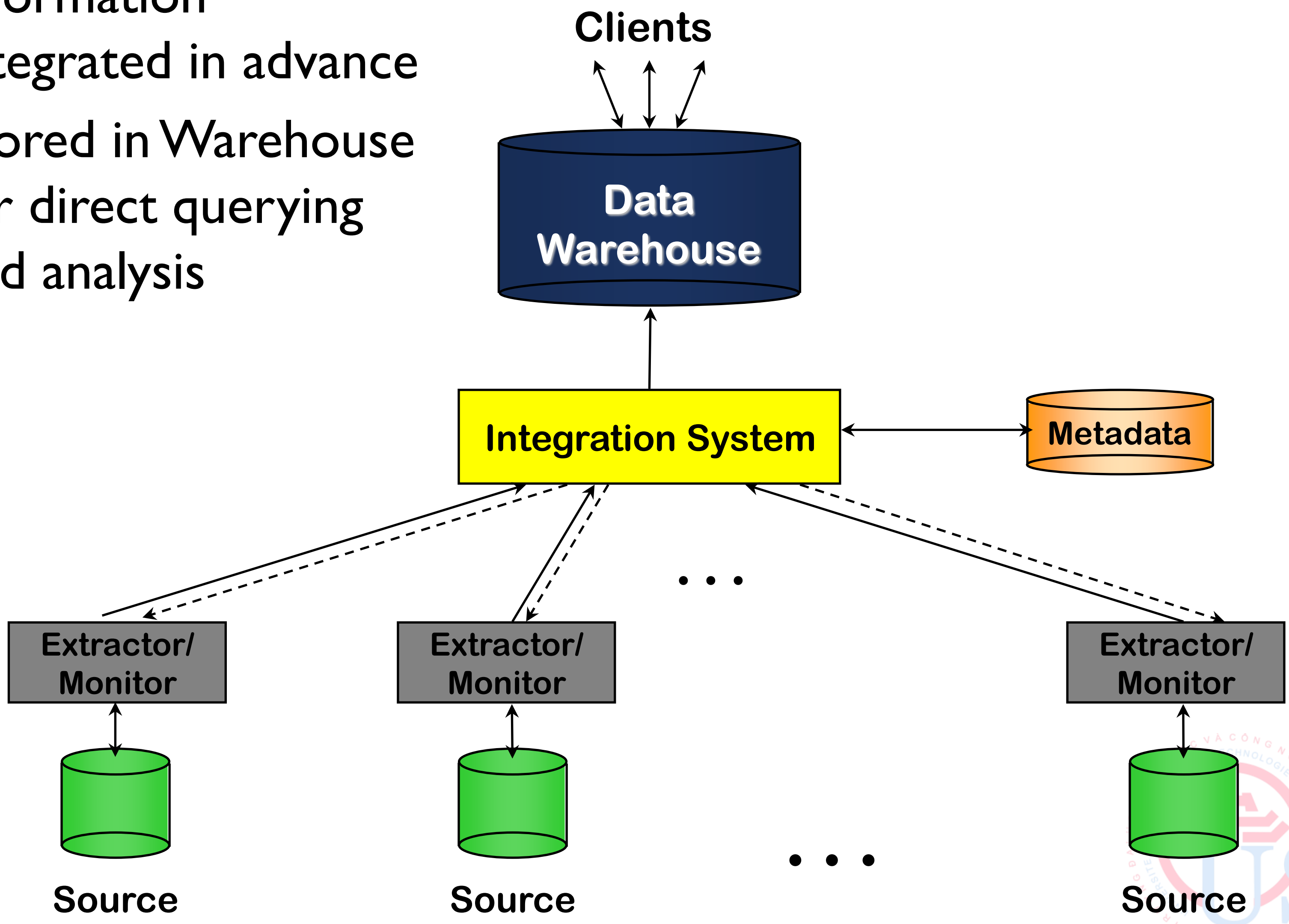
Information
Data

A process of transforming data into information and making it available to users in a timely enough manner to make a difference

[Forrester Research, April 1996]

The Warehousing Approach

- Information integrated in advance
- Stored in Warehouse for direct querying and analysis



Advantages of Warehousing Approach

- High query performance
  - But not necessarily most current information
- Doesn’t interfere with local processing at sources
  - Complex queries at warehouse
  - OLTP at information sources
- Information copied at warehouse
  - Can modify, annotate, summarize, restructure, etc.
  - Can store historical information
  - Security, no auditing

What is a Data Warehouse?

- Practitioners Viewpoint

"A data warehouse is simply a single, complete, and consistent store of data obtained from a variety of sources and made available to end users in a way they can understand and use it in a business context."

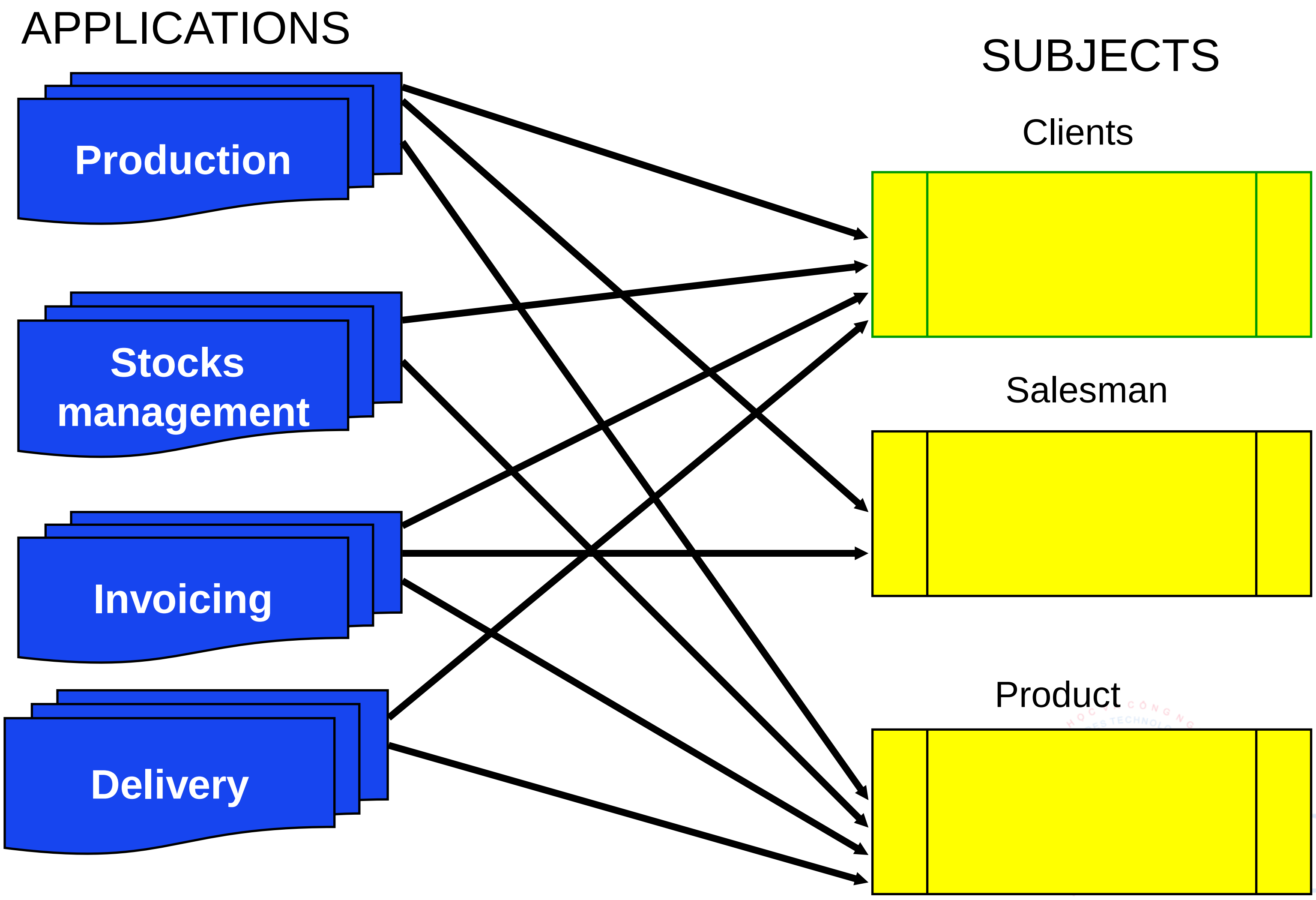
-- Barry Devlin, *IBM Consultant*

- An Alternative Viewpoint

"A DW is a subject-oriented, integrated, time-varying, non-volatile collection of data that is used primarily in organizational decision making."

-- W.H. Inmon, Building the Data Warehouse, 1992

Subject oriented



Time variant data

- Each data is associated to a date
- The time play a key role in DW

Operational databases

In May 2012 Contact

Name	Town
Dupont	Paris
Durand	Lyon

In July 2013 Contact

Name	Town
Dupont	Marseille
Durand	Lyon

DW

Calendar

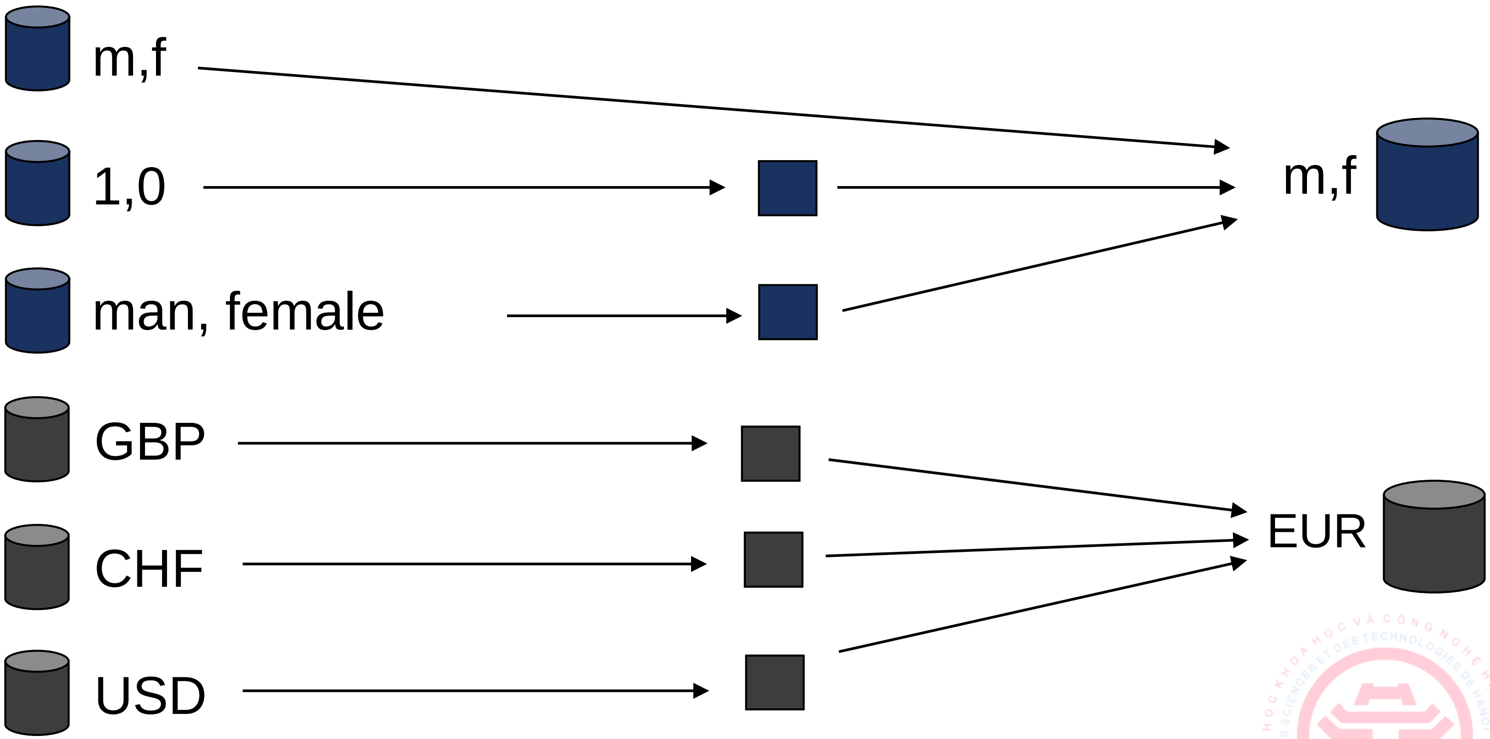
Code	Year	Mon.
1	2012	May
2	2013	July

Contact

Code	Name	Town
1	Dupont	Paris
1	Durand	Lyon
2	Dupont	Marseille

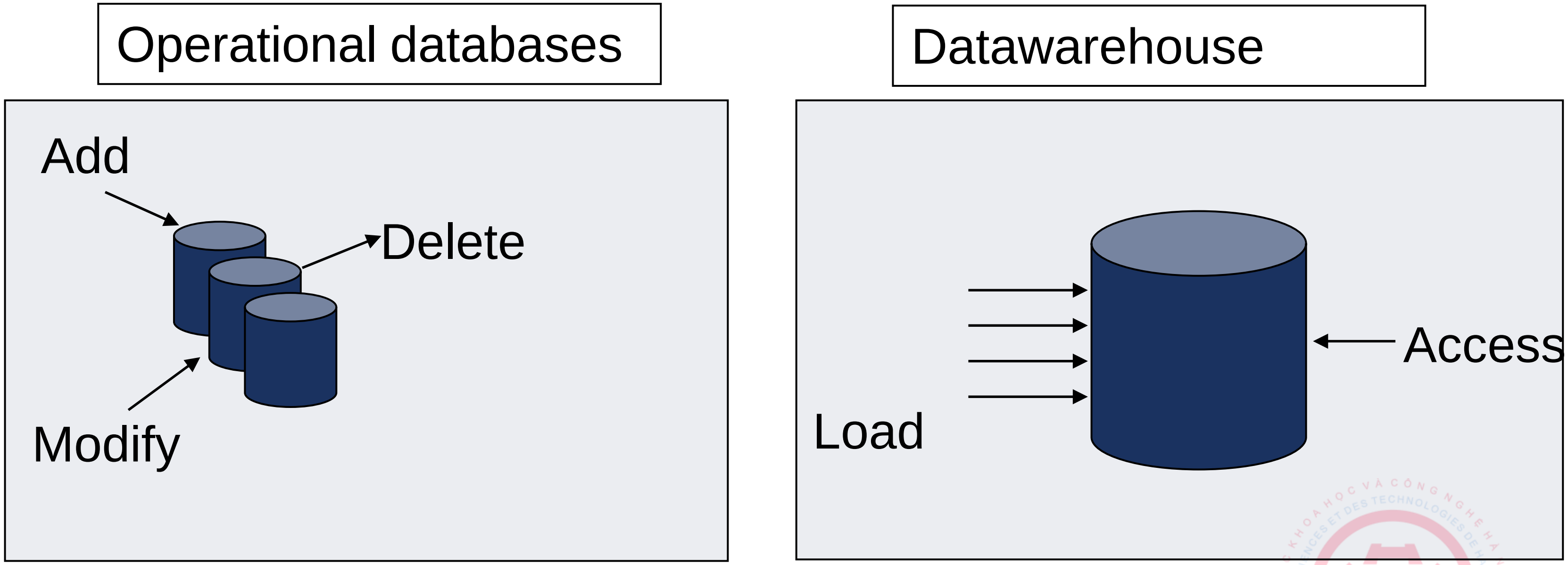
Integrated Data

- Data Normalization
- A unique referential



Non volatiles

- Copy of production data
- Adding only (traceability)



Operational databases

Add

Delete

Modify

Datawarehouse

Load

Access

Very Large Databases

- Terabytes -- 10^12 bytes: Walmart -- 24 Terabytes
- Petabytes -- 10^15 bytes: Geographic Information Systems
- Exabytes -- 10^18 bytes: National Medical Records
- Zettabytes -- 10^21 bytes: Weather images
- Zottabytes -- 10^24 bytes: Intelligence Agency Videos

Usage of the DW

- Business Intelligence:
  - Visualize and exploit a huge amount of complex data
  - «Business Intelligence is a set of methodologies, processes, architectures, and technologies that transform raw data into meaningful and useful information used to enable more effective strategic, tactical, and operational insights and decision-making. »

- 3 main tools:
  - OLAP: On-Line Analytical Processing
  - Data mining
  - Query and Visualization tools

OLTP vs Data warehouse

- OLTP Systems are used to "*run*" a business

- The Data Warehouse helps to "*optimize*" the business

OLTP vs. Data Warehouse

- OLTP systems are tuned for known transactions and workloads while workload is not known a prior in a data warehouse
- Special data organization, access methods and implementation methods are needed to support data warehouse queries (typically multidimensional queries)
  - e.g., *average amount spent on phone calls between 9AM-5PM in Pune during the month of December*

OLTP vs. Data Warehouse

- OLTP
  - Application Oriented
  - Used to run business
  - Detailed data
  - Current up to date
  - Isolated Data
  - Clerical User

- Warehouse DW
  - Subject Oriented
  - Used to analyze business
  - Summarized and refined
  - Snapshot data
  - Integrated Data
  - Knowledge User (manager, analyst)

OLTP vs. Data Warehouse

OLTP

- Performance Sensitive
- Few Records accessed at a time (tens)
- Read/Update Access
- No data redundancy
- Database Size 100MB -100 GB

Data Warehouse

- Performance relaxed
- Large volumes accessed at a time(millions)
- Mostly Read (Batch Update)
- Redundancy present
- Database Size 100 GB - few terabytes

OLTP vs Data Warehouse

- OLTP
  - Transaction throughput is the performance metric
  - Thousands of users
  - Managed in entirety

- Data Warehouse
  - Query throughput is the performance metric
  - Hundreds of users
  - Managed by subsets

Summary: OLTP vs. Data warehouse

Characteristics	OLTP (standard DB)	Data warehouse
Use	Day to day management	Decision making
User type	Employees (eg. Clerical)	Analysts, managers
Number of user	More (thousands, millons)	Less (hundreds)
Operations	A lot update, some read (simple and short query) Many transactions	Mostly read (long and complex query) Almost no transaction
Time	Current snapshot	Time variant
Changed speed	Up-to-second	Later
Perception	Bidimensionnal	Multidimentional
Normalization	Frequent	Rare
Derived data	Low, rare	High, common
Size	Smaller (MB-TB)	Bigger (TB, PB, EB)

Commercial DW solutions

Business Objects

SPSS

COGNOS

Hyperion

Microsoft

sas

ORACLE

FRANCE

Ab InITIO

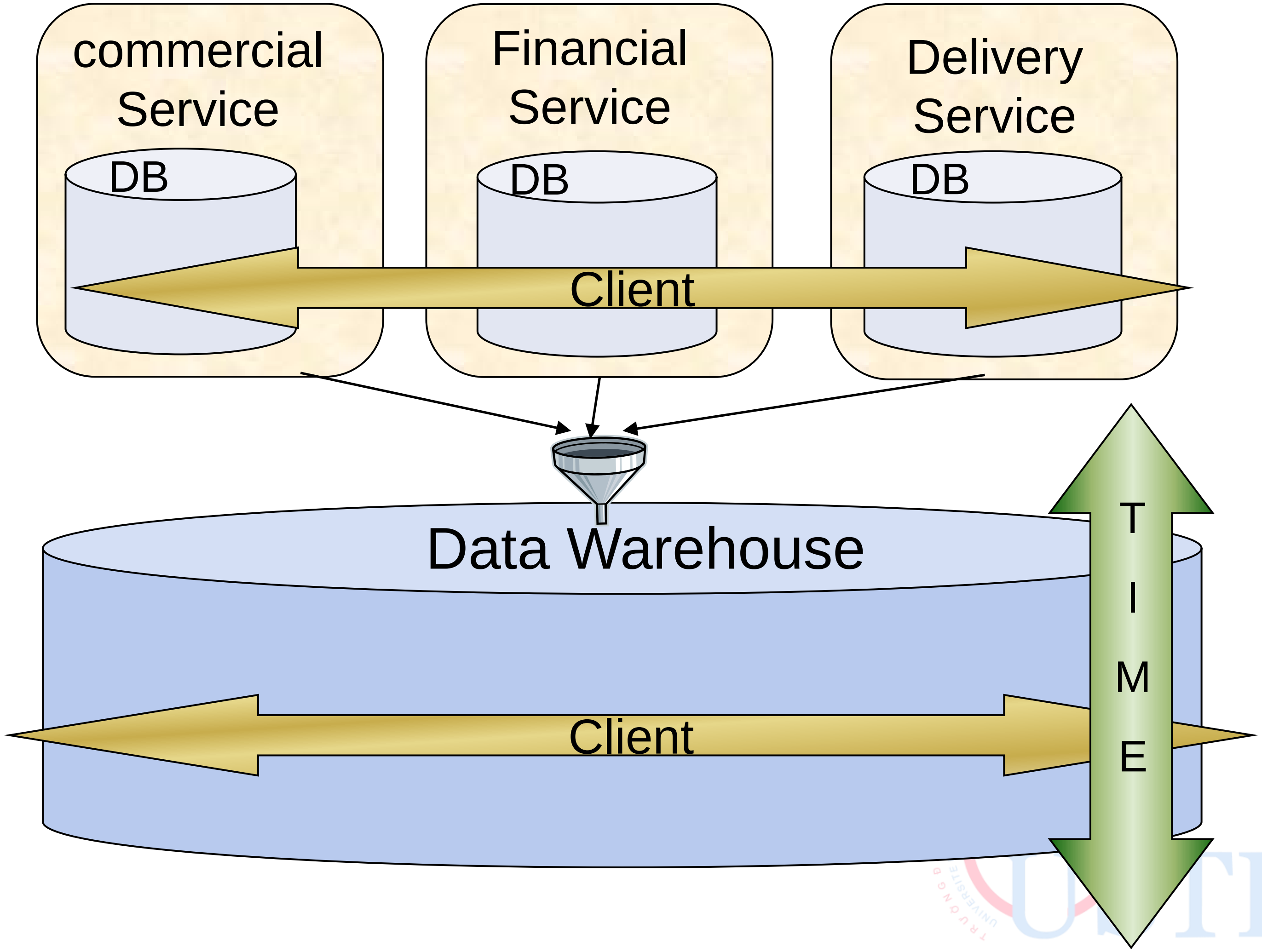
IBM

ARCHITECTURE

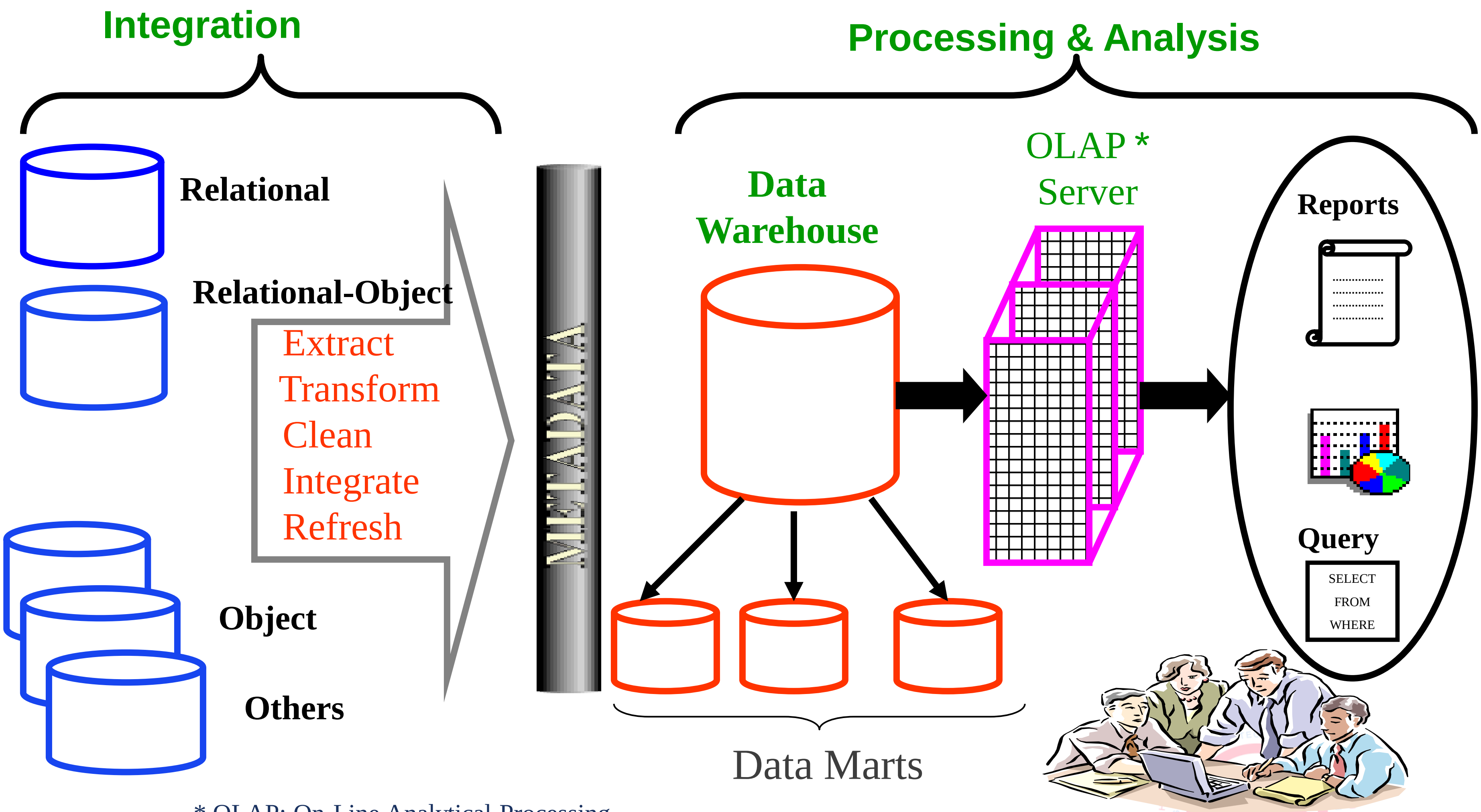
DB and DW: Illustration

OLTP: On-Line Transactional Processing

OLAP: On-Line Analitical Processing



Architecture of a Data warehouse



Integration

Processing & Analysis

Relational

Relational-Object

Object

Others

Extract
Transform
Clean
Integrate
Refresh

METADATA

Data Warehouse

OLAP *
Server

Reports

Query

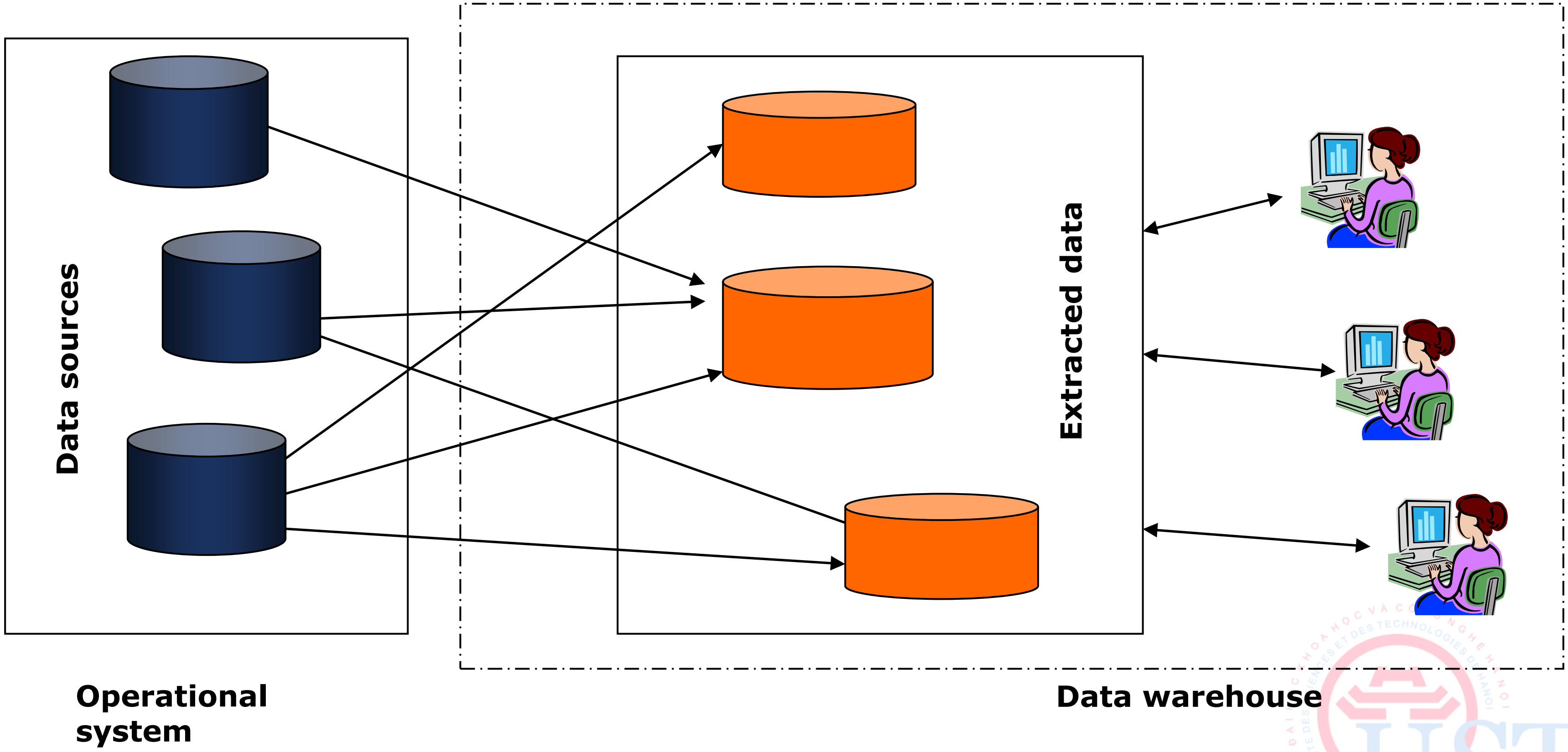
SELECT
FROM
WHERE

Data Marts

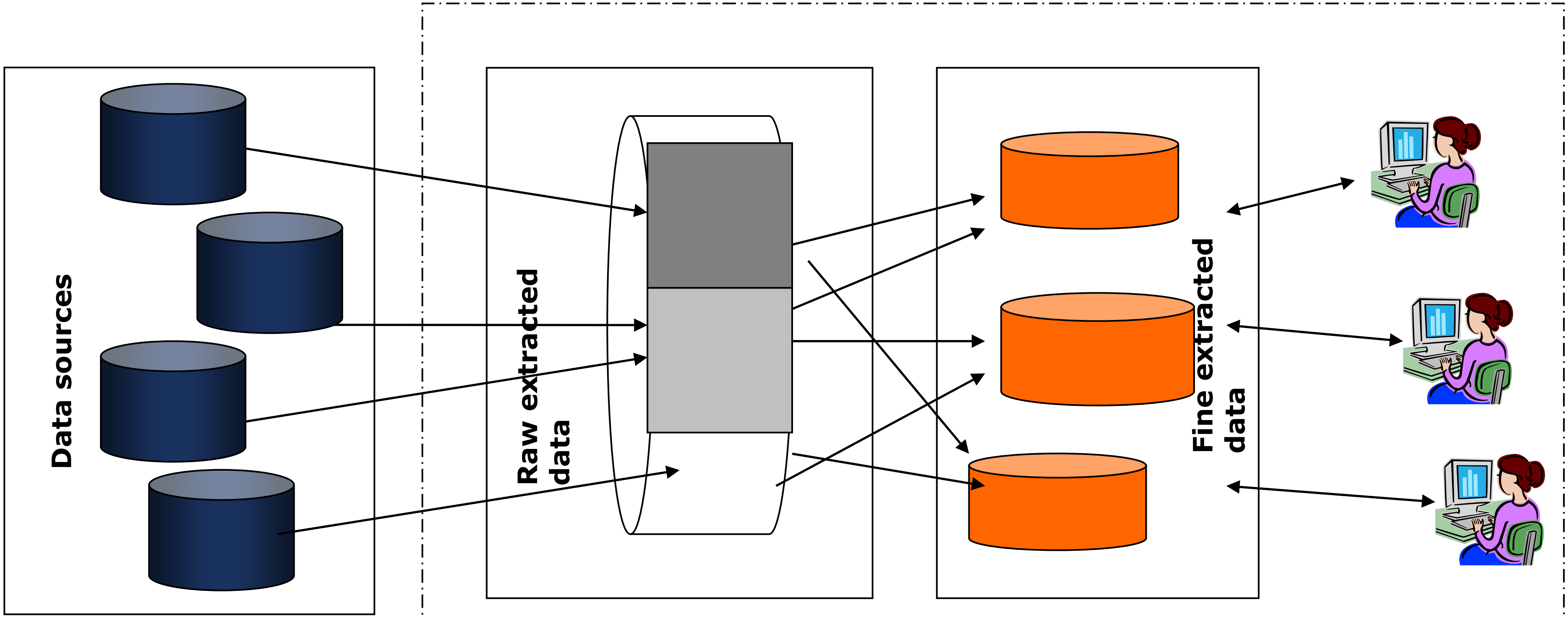
* OLAP: On-Line Analytical Processing

Data architecture – 2 layers

- 2 layers



Data architecture – 3 layer



Operational system

Data warehouse

Refresh

- Propagate updates on source data to the warehouse
- Issues:
  - when to refresh
  - how to refresh -- refresh techniques

When to Refresh?

- Periodically (e.g., every night, every week) or after significant events
- Every update: not warranted unless warehouse data require current data (up to the minute stock quotes)
- Refresh policy set by administrator based on user needs and traffic
- Possibly different policies for different sources

Refresh Techniques

- Full Extract from base tables
  - Read entire source table: too expensive
  - Maybe the only choice for legacy systems
- Update on changes

How To Detect Changes

- Create a snapshot log table to record ids of updated rows of source data and timestamp
- Detect changes by:
  - Defining after row triggers to update snapshot log when source table changes
  - Using regular transaction log to detect changes to source data

Data Extraction and Cleansing

- Extract data from existing operational and legacy data
- Issues:
  - Data quality at the sources
  - Sources of data for the warehouse → Merging different data sources
  - Data Transformation
  - How to propagate updates (on the sources) to the warehouse
  - Terabytes of data to be loaded

Schema Design

- Database organization
  - must look like business
  - must be recognizable by business user
  - approachable by business user
  - Must be *simple*
- Schema Types
  - Star Schema
  - Fact Constellation Schema
  - Snowflake schema

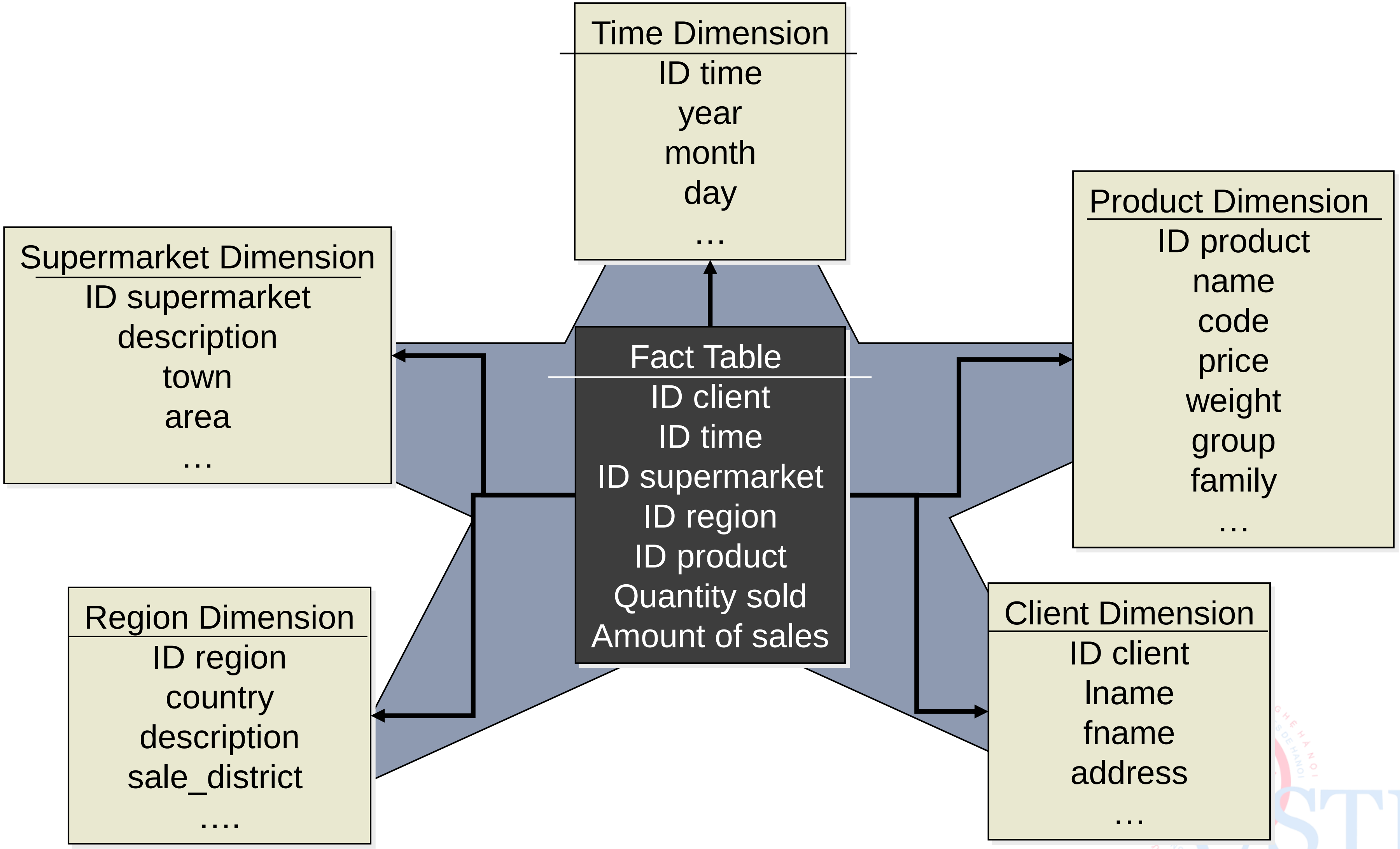
Dimension Tables

- Dimension tables
  - Define business in terms already familiar to users
  - Wide rows with lots of descriptive text
  - Small tables (about a million rows)
  - Joined to fact table by a foreign key
  - heavily indexed
  - typical dimensions
    - time periods, geographic region (markets, cities), products, customers, salesperson, etc.

Fact Tables

- Is the central table
- mostly raw numeric items
- narrow rows, a few columns at most
- large number of rows (millions to a billion)
- Access via dimensions

Star Schema



Time Dimension
ID time
year
month
day
…

Supermarket Dimension
ID supermarket
description
town
area
…

Product Dimension
ID product
name
code
price
weight
group
family
…

Fact Table
ID client
ID time
ID supermarket
ID region
ID product
Quantity sold
Amount of sales

Region Dimension
ID region
country
description
sale_district
….

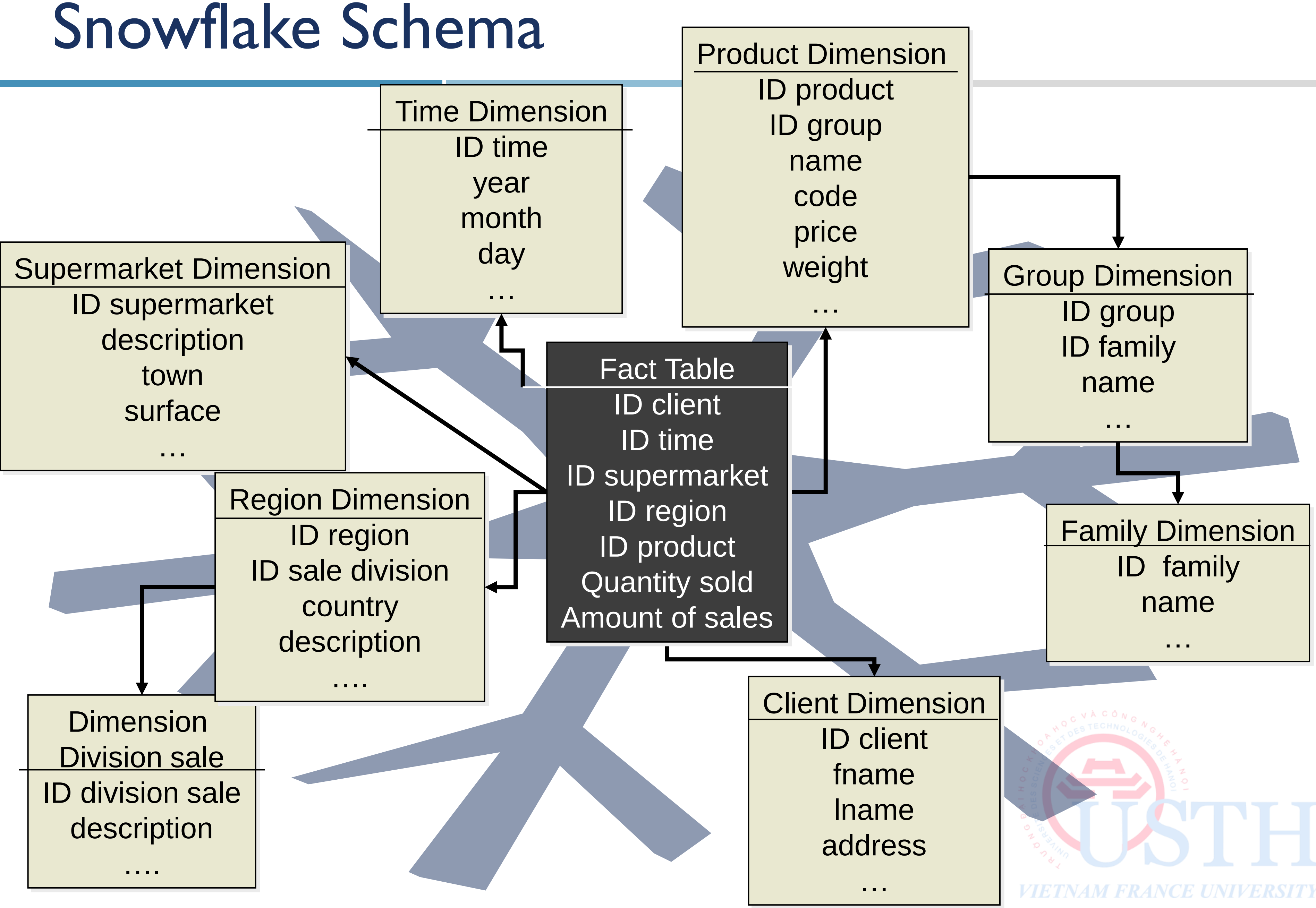
Client Dimension
ID client
lname
fname
address
…

Advantages / Disadvantages

- simple
- more used !!!

- redundancy (dimension tables may not be normalized)
- size of dimension

Snowflake Schema



Time Dimension
ID time
year
month
day
…

Product Dimension
ID product
ID group
name
code
price
weight
…

Supermarket Dimension
ID supermarket
description
town
surface
…

Fact Table
ID client
ID time
ID supermarket
ID region
ID product
Quantity sold
Amount of sales

Group Dimension
ID group
ID family
name
…

Region Dimension
ID region
ID sale division
country
description
….

Family Dimension
ID family
name
…

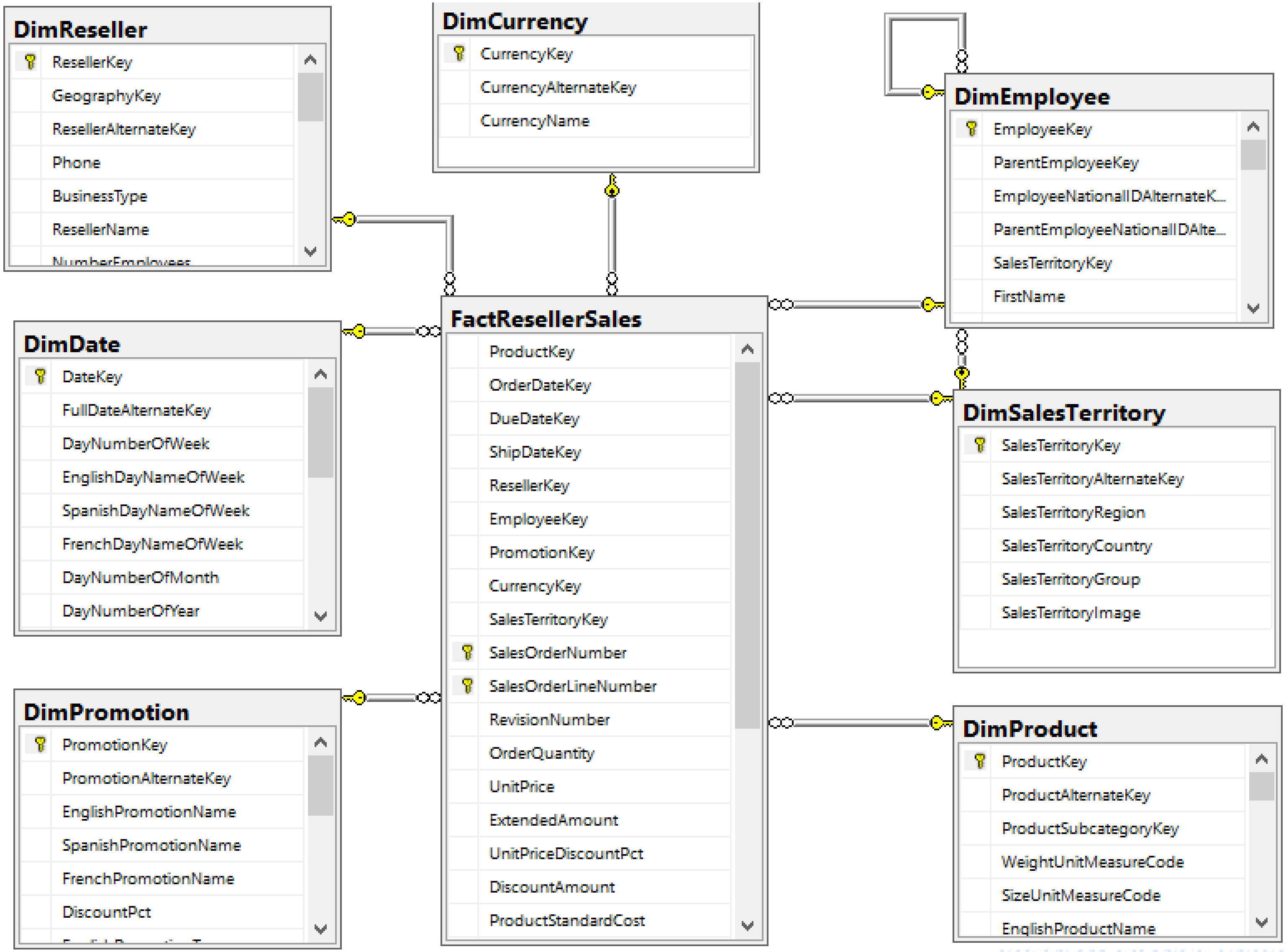
Dimension Division sale
ID division sale
description
….

Client Dimension
ID client
fname
lname
address
…

Snowflake Schema

- Variation of the star schema
- Dimension tables are normalized
- Less redundancy but slower execution of queries (joins)
- Mixed approach
  - Some tables are normalized some not

A part from Adventure Work DW



DimReseller
- 🔑 ResellerKey
- GeographyKey
- ResellerAlternateKey
- Phone
- BusinessType
- ResellerName
- NumberEmployees

DimCurrency
- 🔑 CurrencyKey
- CurrencyAlternateKey
- CurrencyName

DimEmployee
- 🔑 EmployeeKey
- ParentEmployeeKey
- EmployeeNationalIDAlternateK...
- ParentEmployeeNationalIDAlte...
- SalesTerritoryKey
- FirstName

DimDate
- 🔑 DateKey
- FullDateAlternateKey
- DayNumberOfWeek
- EnglishDayNameOfWeek
- SpanishDayNameOfWeek
- FrenchDayNameOfWeek
- DayNumberOfMonth
- DayNumberOfYear

FactResellerSales
- ProductKey
- OrderDateKey
- DueDateKey
- ShipDateKey
- ResellerKey
- EmployeeKey
- PromotionKey
- CurrencyKey
- SalesTerritoryKey
- 🔑 SalesOrderNumber
- 🔑 SalesOrderLineNumber
- RevisionNumber
- OrderQuantity
- UnitPrice
- ExtendedAmount
- UnitPriceDiscountPct
- DiscountAmount
- ProductStandardCost

DimSalesTerritory
- 🔑 SalesTerritoryKey
- SalesTerritoryAlternateKey
- SalesTerritoryRegion
- SalesTerritoryCountry
- SalesTerritoryGroup
- SalesTerritoryImage

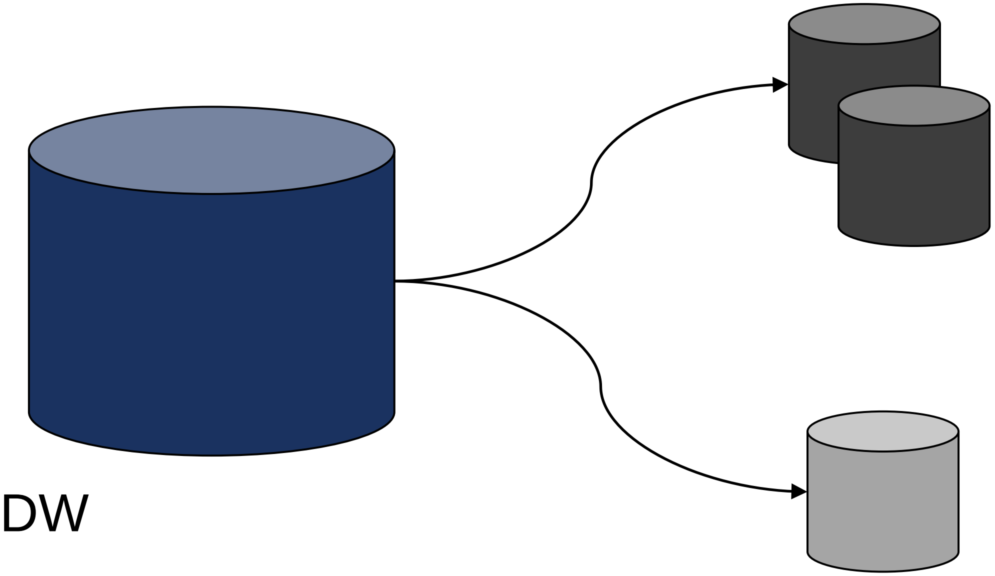
DimPromotion
- 🔑 PromotionKey
- PromotionAlternateKey
- EnglishPromotionName
- SpanishPromotionName
- FrenchPromotionName
- DiscountPct

DimProduct
- 🔑 ProductKey
- ProductAlternateKey
- ProductSubcategoryKey
- WeightUnitMeasureCode
- SizeUnitMeasureCode
- EnglishProductName

DATA MART

Data mart

- Subset of a DW
- Specific needs of a service/function
- View according to a specific jobs



Datamarts of the Marketing service

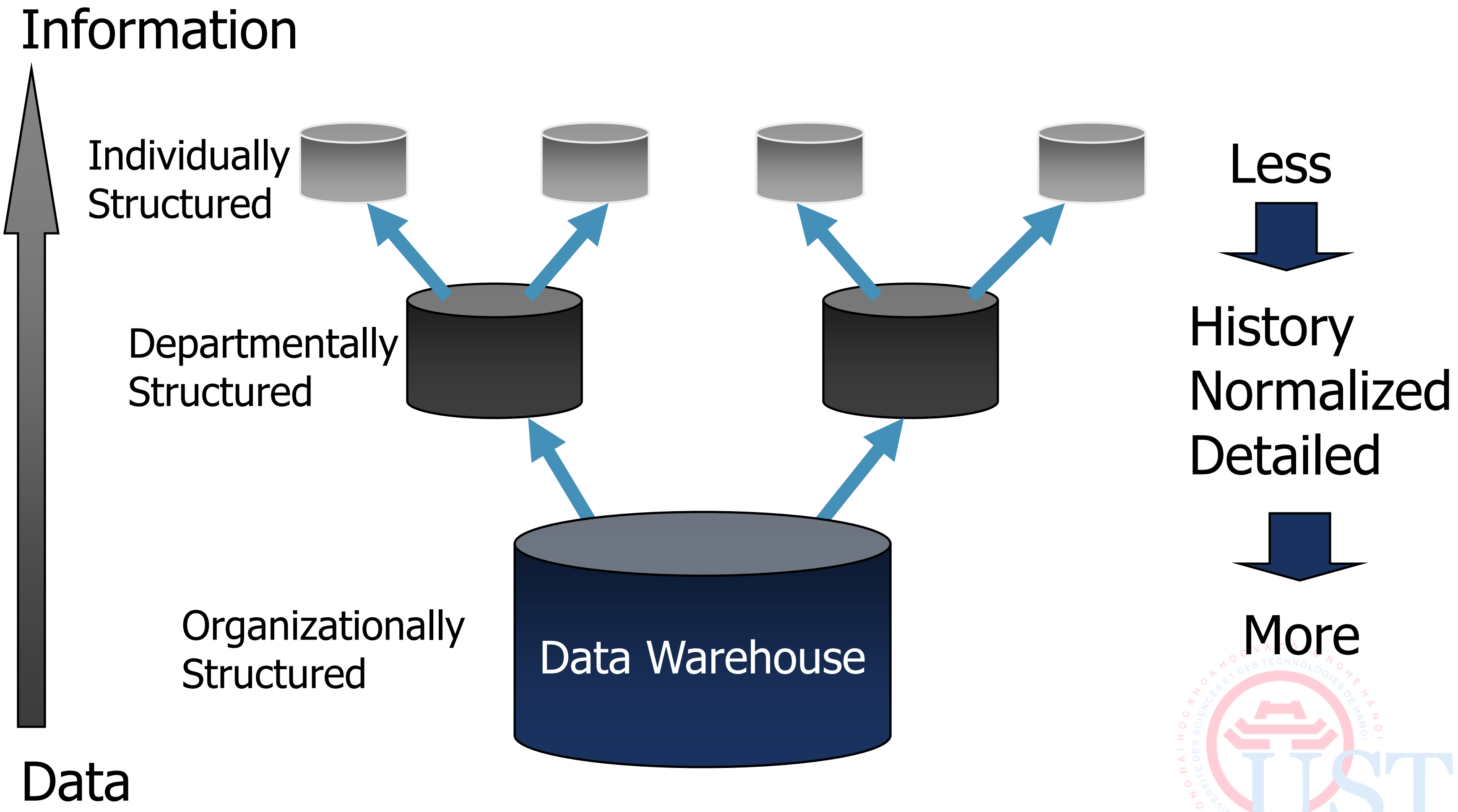
Datamart of the HR service

DW

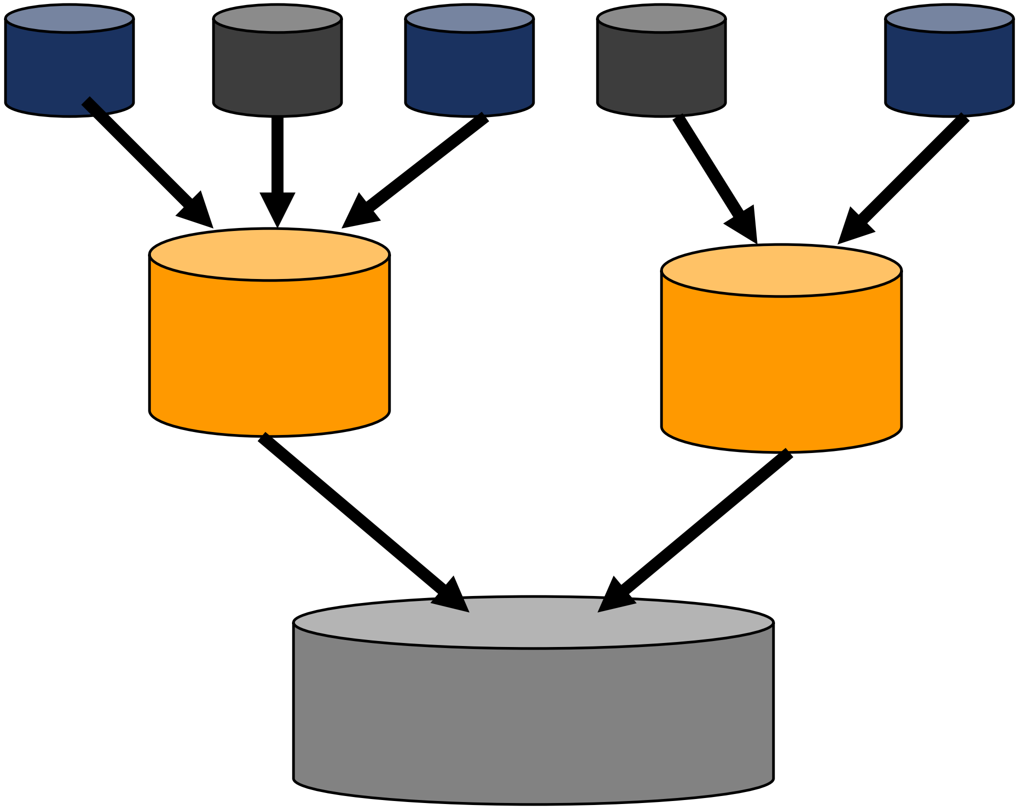
Interest of datamarts

- Structured environment
  - according to a job needs
  - According to a specific usage
- Less data than DW
  - Ease the manipulation and understanding ot the Data
  - Improve query response time
- Targeted users
  - DM more easy to define

From the Data Warehouse to Data Marts



Data Mart Centric

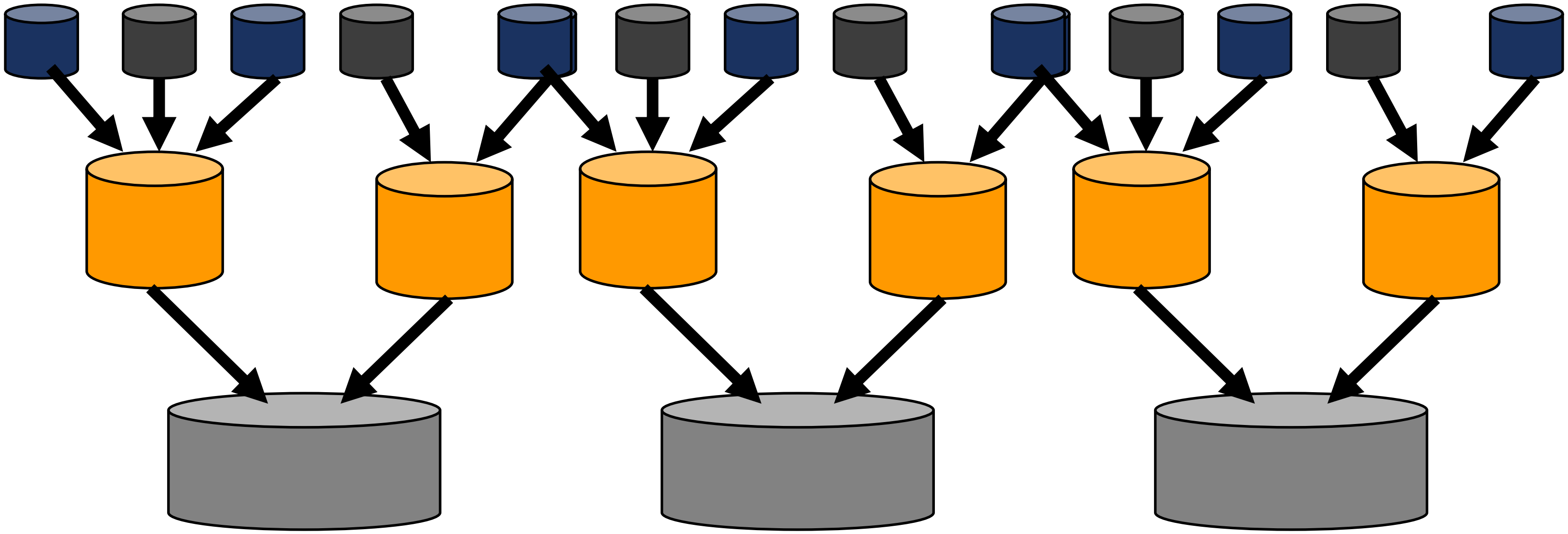


Data Sources

Data Marts

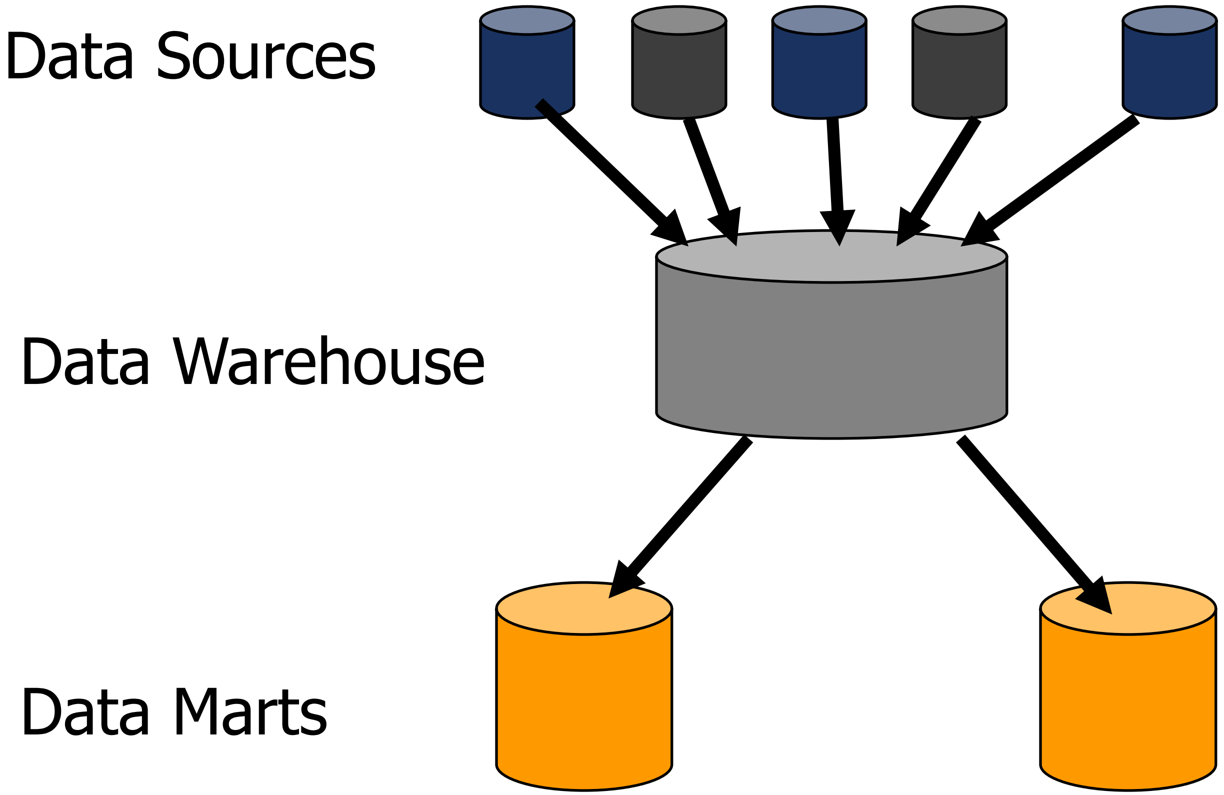
Data Warehouse

Problems with Data Mart Centric Solution



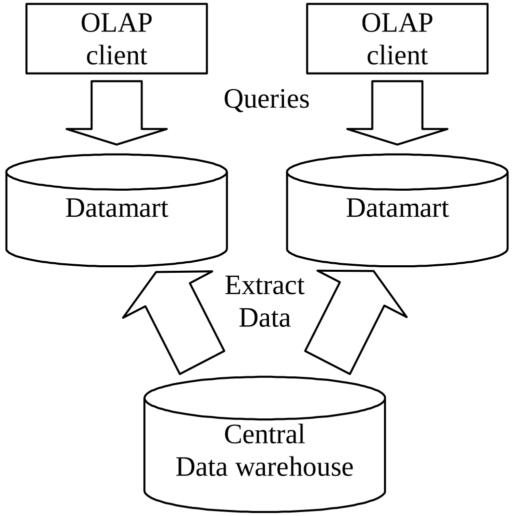
If you end up creating multiple warehouses, integrating them is a problem

True Warehouse

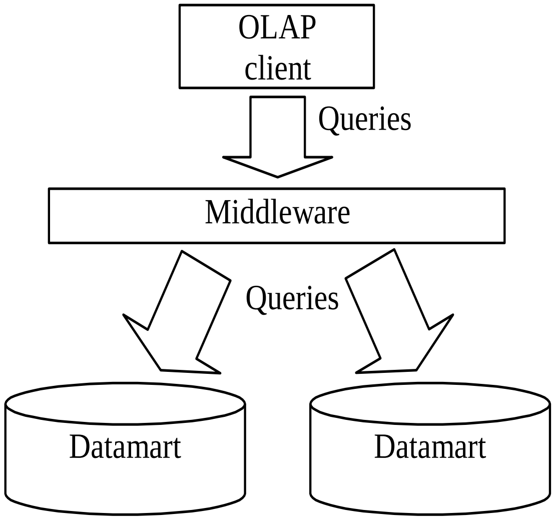


Data Marts

- A data mart (departmental data warehouse) is a specialized system that brings together the data needed for a department or related applications.



Centralized



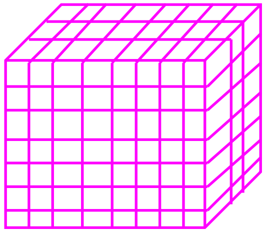
Decentralized

DATA WAREHOUSE QUERIES

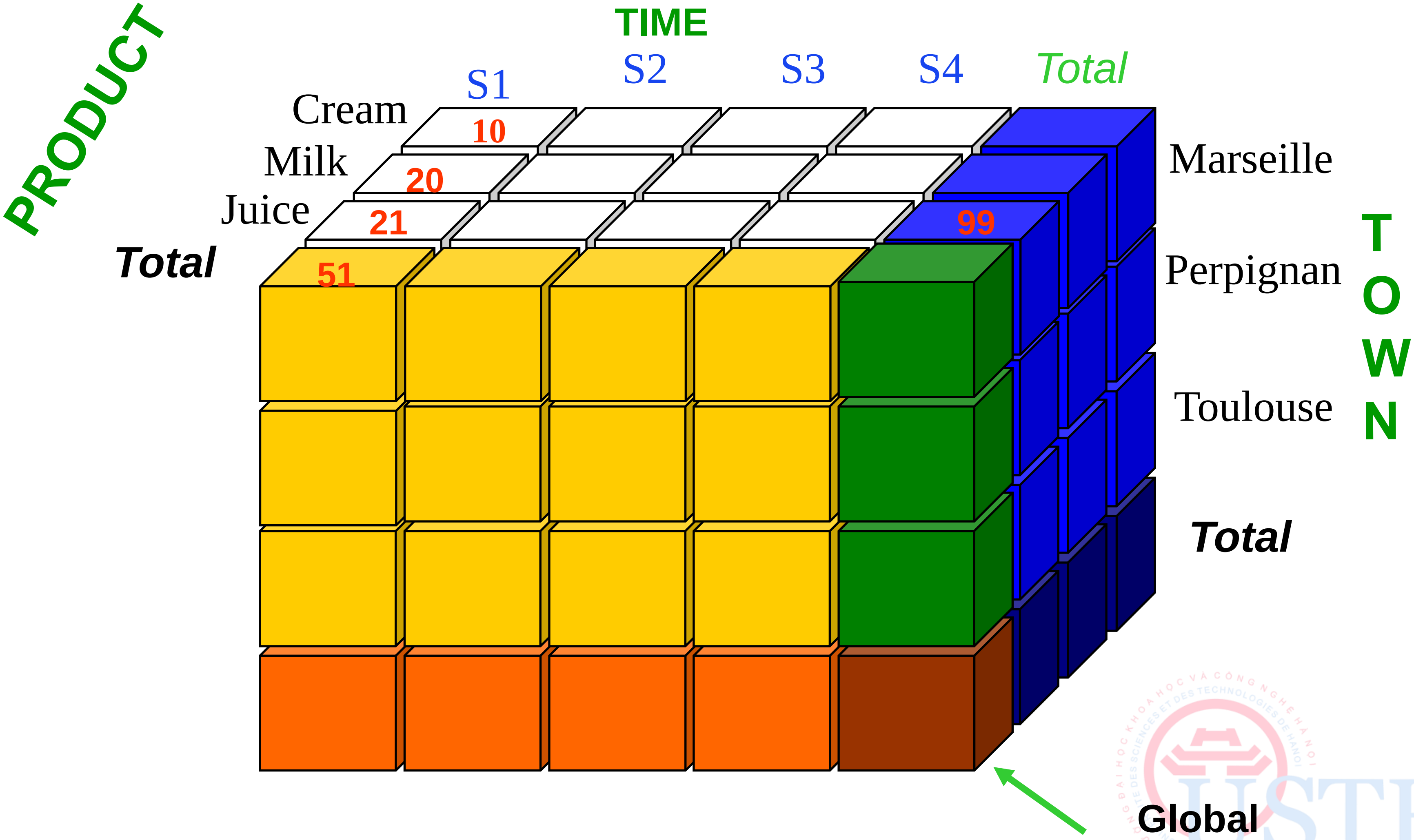
OLAP Hyper cube

- Online Analytical processing
- Objectives
  - Get information already agregated according to users needs
  - Representation of information in one hyper cube at N dimensions

- OLAP Operations
  - Fonctionnalities used to facilitate the multidimensional analysis: operations on the hyper cube



Example of a data cube

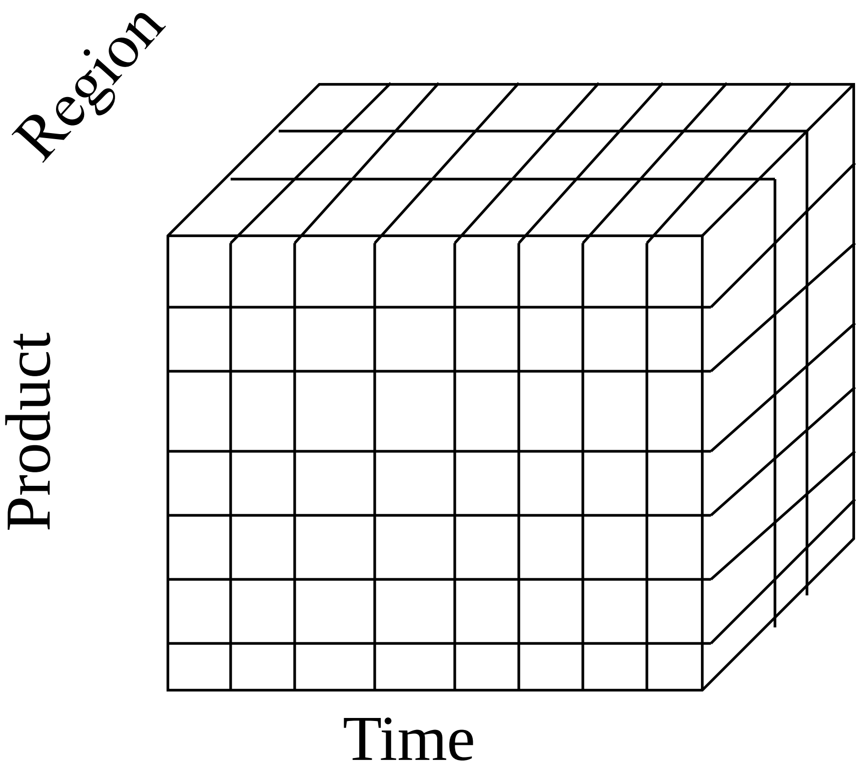


Dimension

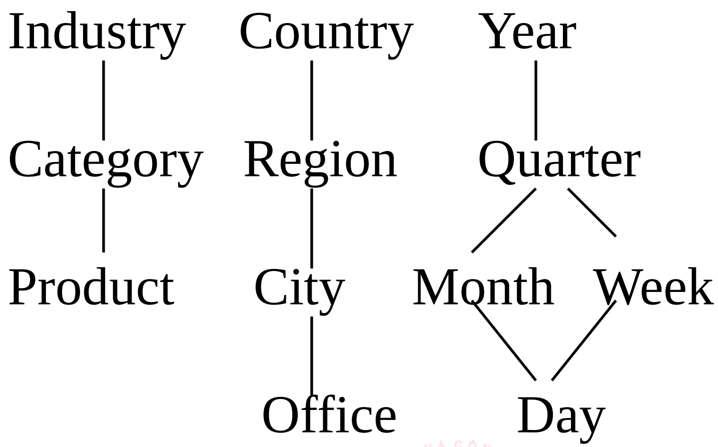
- **Dimension** is a data element that categorizes each item in a data set into non-overlapping regions
  - Eg: time, town, product
- Roles: to provide filtering, grouping and labeling.
- Each dimension in a data warehouse may have one or more hierarchies applied to it.
  - Time:
    - Day > Month > Year
    - Day > Week > Year
    - Day > Month > Quarter > Year

Multidimensional View of Data

- Sales volume (measure) as a function of product, time, and geography (dimensions).



Dimensions: Product, Region, Time
Hierarchical summarization paths



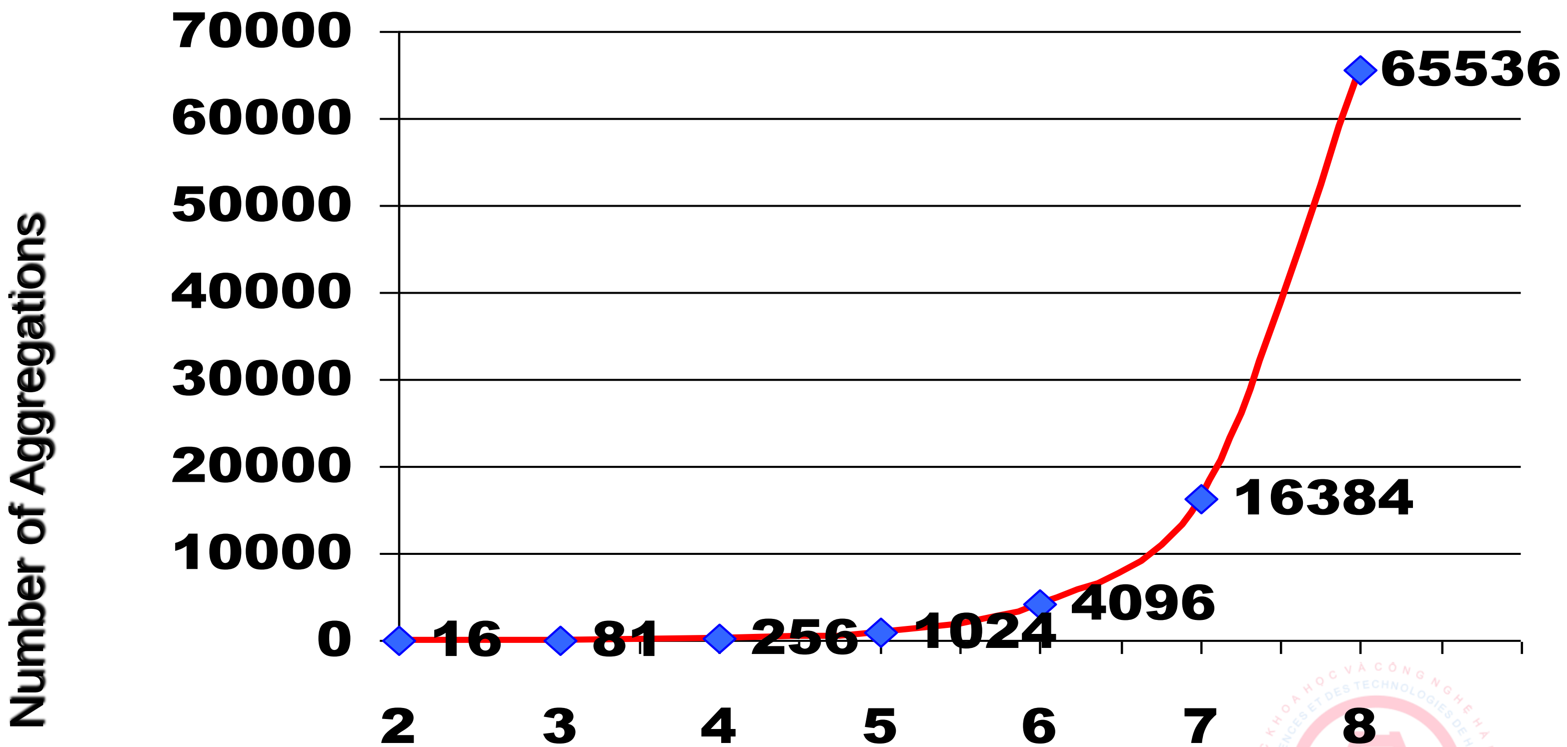
Industry
Category
Product

Country
Region
City
Office

Year
Quarter
Month
Week
Day

Typical Cube Problems: Data Explosion

Data Explosion Syndrome



(4 levels in each dimension)

Number of Dimensions

Microsoft TechEd'98

Storage of the data cube

- ROLAP: Relational On-Line Analytical Processing
  - Using relational tables

- MOLAP: Multidimensional On-Line Analytical Processing
  - Storage in a n-dimension array
    (a new data structure)

MOLAP

- Difference ROLAP - MOLAP
  - Storage Model
    - MOLAP : direct (n-dimension array)
    - ROLAP : indirect (relational tables)

- Advantages/Disadvantages of MOLAP

  + Direct access for queries

  - If sparse data => waste disk space
  - No standard

MOLAP example

Time	Trim1			Trim2			Trim3			Trim4			Total			
Product Town	M	P	T	M	P	T	M	P	T	M	P	T	M	P	T	Tot
Cream	**8**			**4**			**6**			**9**			**27**	**10**		
Milk	**22**		**10**	**23**			**19**			**29**			**93**			
Juice	**21**			**24**		**10**	**25**			**29**			**99**			
Total	**51**															

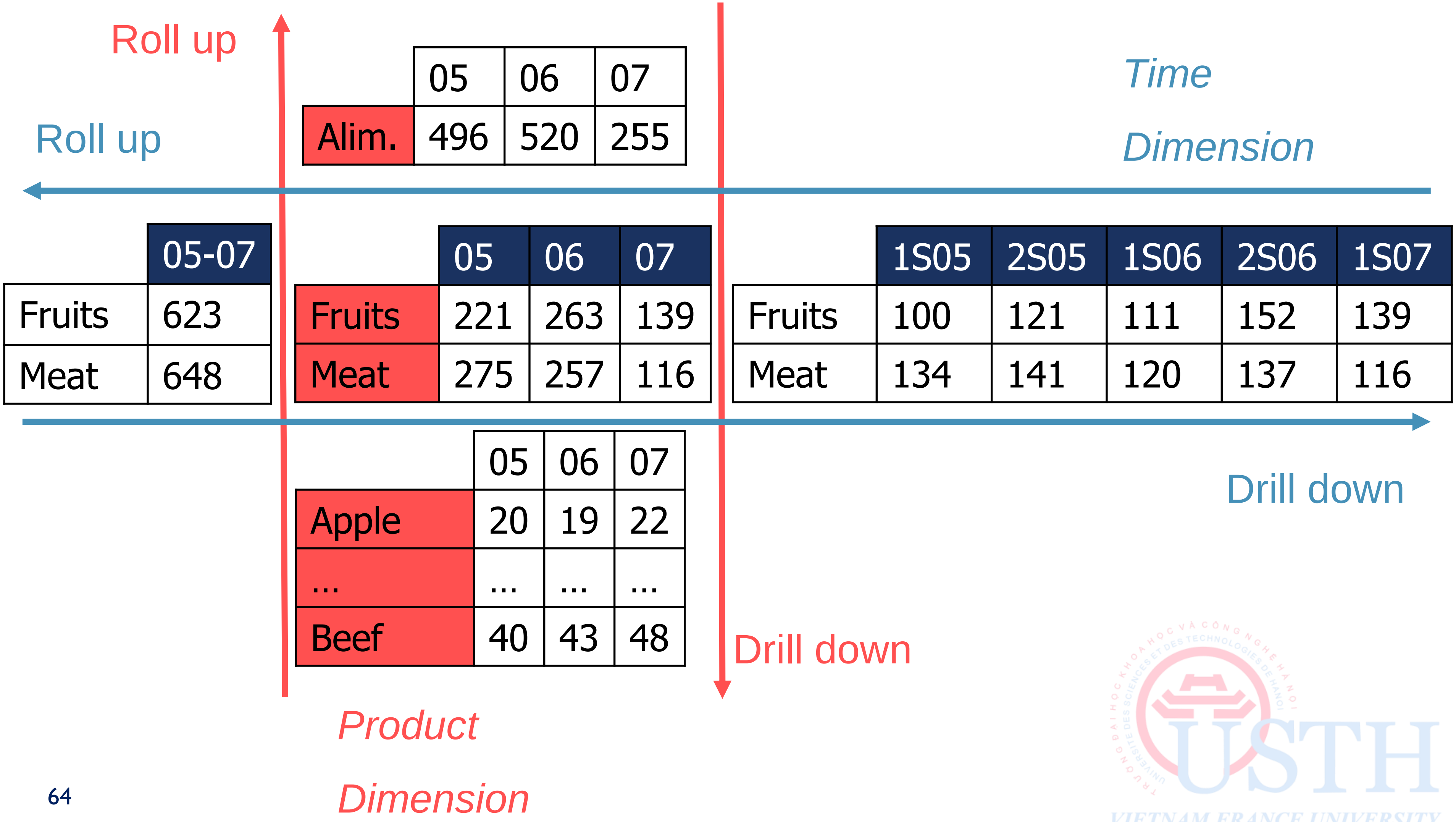
Sales by product, time and town

M: Marseille, P: Perpignan, T: Toulouse

Cube Algebra

- Roll up :
  - Agregate on a dimension
    - Week ➔ Month
    - Operators: ROLLUP, CUBE, GROUPING SETS
- Drill down :
  - Detail on a dimension
    - Month ➔ Week
- Slice & Dice :
  - Selection and projection on 1 dimension
    - Month = 04-2003 ; Project(Region, Product)
- Rotate:
  - Move the cube to visualize a face
    - (Region,Product)➔(Region, Month)

Roll-up, Drill-down



Roll up

Roll up

	05	06	07
Alim.	496	520	255

Time

Dimension

	05-07
Fruits	623
Meat	648

	05	06	07
Fruits	221	263	139
Meat	275	257	116

	1S05	2S05	1S06	2S06	1S07
Fruits	100	121	111	152	139
Meat	134	141	120	137	116

Drill down

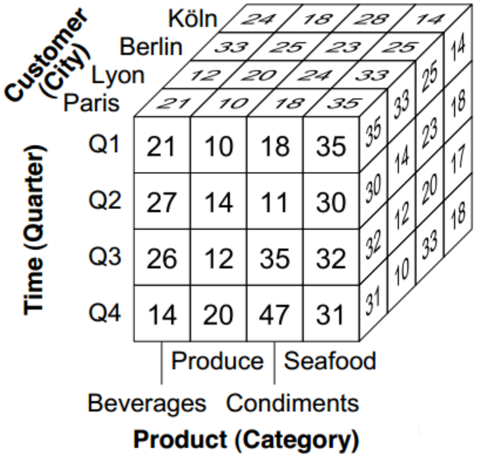
	05	06	07
Apple	20	19	22
…	…	…	…
Beef	40	43	48

Drill down

Product

Dimension

Roll-up, Drill-down example



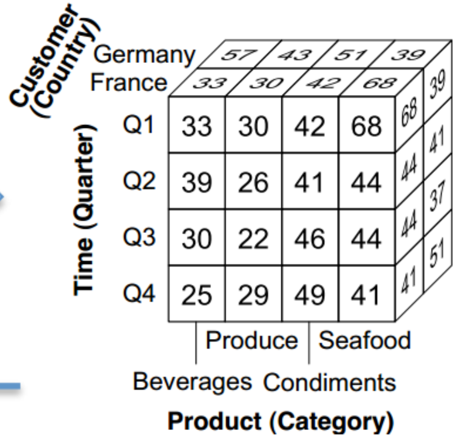
	Beverages	Produce	Condiments	Seafood
Q1	21	10	18	35
Q2	27	14	11	30
Q3	26	12	35	32
Q4	14	20	47	31

Data cube for 2012

To compute the sales quantity by countries

Roll-up to the country level

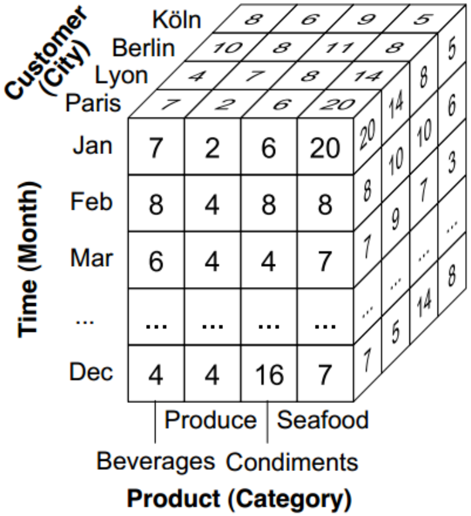
Drill-down to the city level



	Beverages	Produce	Condiments	Seafood
Q1	33	30	42	68
Q2	39	26	41	44
Q3	30	22	46	44
Q4	25	29	49	41

Drill-down to the month level

To see why sales of seafood in Q1 is higher than other products



	Beverages	Produce	Condiments	Seafood
Jan	7	2	6	20
Feb	8	4	8	8
Mar	6	4	4	7
...	...	...	...	...
Dec	4	4	16	7

ROLLUP Example

Input

Animal	Loc	Quantity
Dog	Paris	12
Cat	Paris	18
Turtle	Rome	4
Dog	Rome	14
Cat	Naples	9
Dog	Naples	5
Turtle	Naples	1

```
SELECT Animal, Loc, SUM(Quantity)
                          AS Quantity

FROM Animals
GROUP BY ROLLUP (Animal, Loc)
```

Output

Animal	Loc	Quantity
Cat	Paris	18
Cat	Naples	9
Cat	–	27
Dog	Paris	12
Dog	Naples	5
Dog	Rome	14
Dog	–	31
Turtle	Naples	1
Turtle	Rome	4
Turtle	–	5
–	–	63

CUBE Example

Input

Animal	Loc	Quantity
Dog	Paris	12
Cat	Paris	18
Turtle	Rome	4
Dog	Rome	14
Cat	Naples	9
Dog	Naples	5
Turtle	Naples	1

```
SELECT Animal, Loc, SUM(Quantity)
AS Qty
FROM Animals
GROUP BY CUBE (Animal, Loc)
```

Output

Animal	Loc	Quantity
Cat	Paris	18
Cat	Naples	9
Cat	–	27
Dog	Paris	12
Dog	Naples	5
Dog	Rome	14
Dog	–	31
Turtle	Naples	1
Turtle	Rome	4
Turtle	–	5
–	–	63
–	Paris	30
–	Naples	15
–	Rome	18

GROUPING SETS Example

Input

Animal	Loc	Quantity
Dog	Paris	12
Cat	Paris	18
Turtle	Rome	4
Dog	Rome	14
Cat	Naples	9
Dog	Naples	5
Turtle	Naples	1

```
SELECT Animal, Loc, SUM(Quantity) as Qty
FROM Animals
GROUP BY GROUPING SETS (Animal, Loc, ())
```

Output

Animal	Loc	Qty
Cat	–	27
Dog	–	31
Turtle	–	5
–	–	63
–	Paris	30
–	Naples	15
–	Rome	18

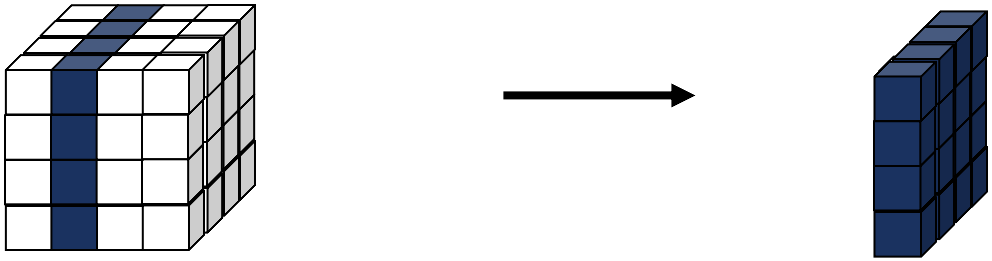
Slice

- Slice ⇔ projection

		05	06	07
Eggs	Vn	220	265	284
	Fr	225	245	240
Meat	Vn	163	152	145
	Fr	187	174	184

→

		06
Eggs	Vt	265
	Fr	245
Meat	Vt	152
	Fr	174



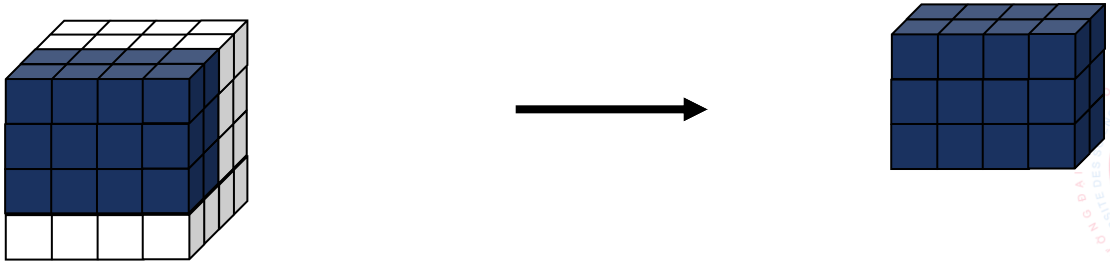
Dice

- Dice ⇔ Selection

		05	06	07
Eggs	Vt	220	265	284
	Fr	225	245	240
Meat	Vt	163	152	145
	Fr	187	174	184

→

		05	06	07
Egg	Vt	220	265	284
	Fr	225	245	240



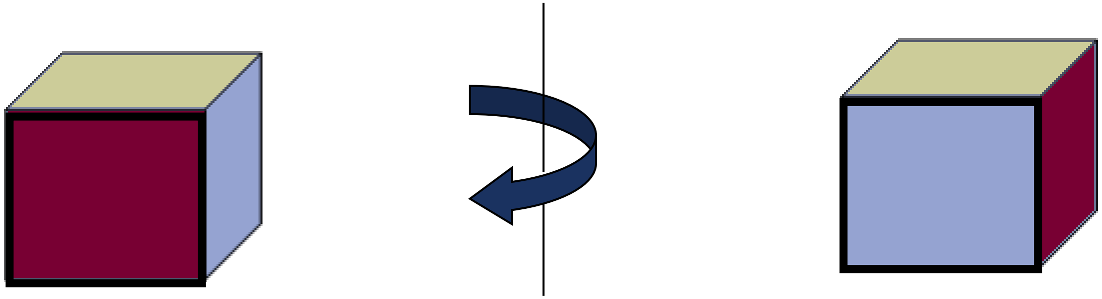
Pivot

Rotate

	05	06	07
Eggs	221	263	139
Meat	275	257	116

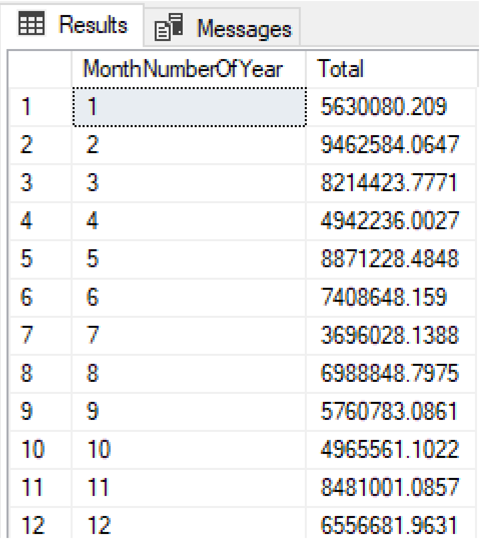
→

	05	06	07
Vt	101	120	52
Fr	395	400	203



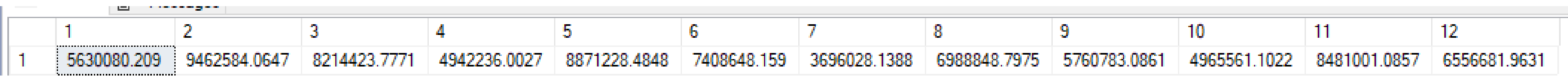
Challenge: Pivot table

```
USE AdventureWorksDW2014

SELECT MonthNumberOfYear,
SUM(UnitPrice * OrderQuantity) Total

FROM FactResellerSales F INNER JOIN
DimDate D ON F.ShipDateKey =
D.DateKey

GROUP BY MonthNumberOfYear
```



Results | Messages

	MonthNumberOfYear	Total
1	1	5630080.209
2	2	9462584.0647
3	3	8214423.7771
4	4	4942236.0027
5	5	8871228.4848
6	6	7408648.159
7	7	3696028.1388
8	8	6988848.7975
9	9	5760783.0861
10	10	4965561.1022
11	11	8481001.0857
12	12	6556681.9631



	1	2	3	4	5	6	7	8	9	10	11	12
1	5630080.209	9462584.0647	8214423.7771	4942236.0027	8871228.4848	7408648.159	3696028.1388	6988848.7975	5760783.0861	4965561.1022	8481001.0857	6556681.9631

```
SELECT *

FROM( SELECT MonthNumberOfYear Month, UnitPrice * OrderQuantity SubTotal

FROM FactResellerSales F INNER JOIN DimDate D ON F.ShipDateKey = D.DateKey) Tb

PIVOT( SUM(SubTotal)  FOR [Month] IN ([1],[2],[3],[4],[5],[6],[7],[8],[9],[10],[11],[12])) PT
```

Exercise

The rector of the USTH would like to observe the facts that could influence the rate success of the students. To do so, he requires a DW that could answer the following queries

- What is the exam success rate with respect to the course and year?
- What is the exam success rate for a mandatory course for the year 2016?
- What is the exam success rate with respect to the sex and the year?
- How many 22 year old students have succeed the advance database exam?
- What is the number of succeeding students during winter semester 2015?

To construct this DW, the data source is the following: we know the name, age, sex of the student, the course name, if it is mandatory or not, the exam date, the given mark, and a success “Boolean”.

Propose a star scheme DW.