POLITECNICO
MILANO 1863

Data Warehouse

Cinzia Cappiello
A.A. 2023-2024

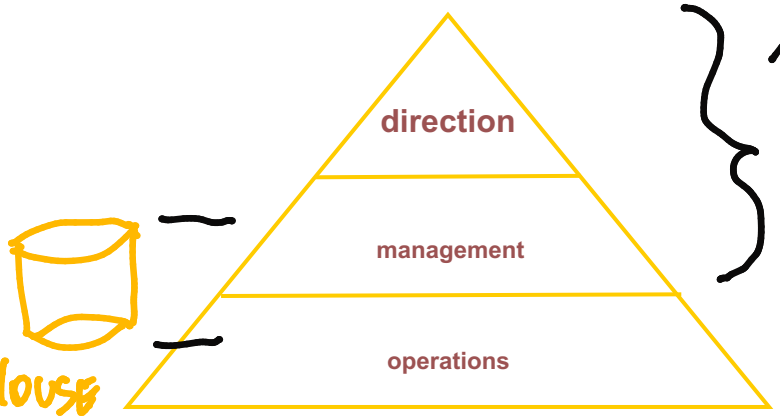
Outline

- What is a Data Warehouse?
- Data Warehouse Architecture
- Data Warehouse operations

Introduction

3

Business Processes' Pyramid



"ROUTINE TASKS" PERFORMED BY MANAGERS, DIRECTORS...

OPERATIONS PERFORMED IN THE DB (CCRUD,...)

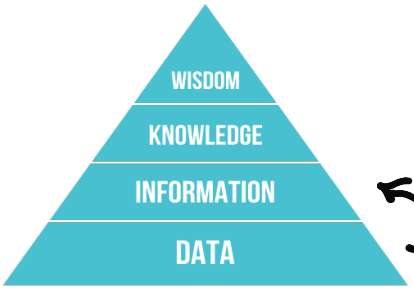
↓

USERS HAVE ACCESS TO ALL DATAS AND ALL INFORMATIONS

DATA WAREHOUSE

ROLE OF PREPARING DATAS FOR THE "DECISION TEAM" THAT WILL BASE THE DECISIONS ON THE GIVEN INFORMATIONS

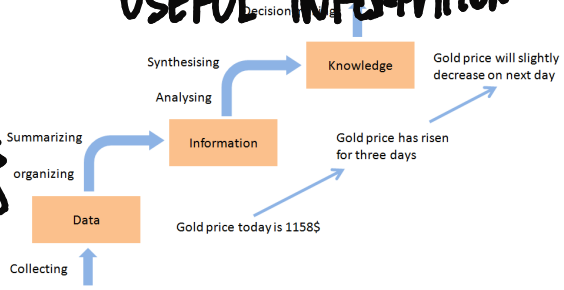
Knowledge pyramid



[wikipedia]

DATA WAREHOUSE TRASFORMS DATAS IN USEFUL INFORMATION

! REMEMBER:
INFORMATION= DATAS WITH A CONTEXT AND A MEANING



5

scope of decisions

	Operational		Strategic
Accuracy	High	⟷	Low
Level of detail	Detailed	⟷	Aggregate
Time horizon	Present	⟷	Future
Frequency of use	High	⟷	Low
Source	Internal	⟷	External
Scope of information	Quantitative	⟷	Qualitative
Nature of information	Narrow	⟷	Wide
Age of information	Present	⟷	Past

Carlo Vercellis 2006

6

ONCE DIFFERENT SYSTEMS ARE BUILT, GENERALLY THEY ARE BASED ON DIFFERENT REQUIREMENTS AND MEANINGS
⇒ THEY ARE SURELY DIFFERENT

OPERATIONAL LEVEL — DECISIONAL LEVEL

OLTP and OLAP systems

ONLINE TRANSACTION PROCESSING — ONLINE ANALYTICAL PROCESSING

OLTP (Standard DB)	OLAP
Mostly updates	Mostly reads
Many small transactions	Queries are long and complex
Current snapshot	History
Raw data	Summarized, reconciled data
Thousands of users (e.g., clerical users)	Hundreds of users

EXAMPLE: A CHANGE IN A VALUE IS PERMANENT
THE OLD VALUE IS NO MORE AVAILABLE AFTER AN UPDATE

THE TWO SYSTEMS ARE SEPARATED BECAUSE:
- OPERATIONS ON OLAP DO NOT HAVE TO BE APPLIED TO THE MAIN DATABASE, OTHERWISE IT WOULD BE AFFECTED IN SPEED
- WE WANT OLTP TO GUARANTEE A FASTER EXECUTION OF TRANSACTION

7

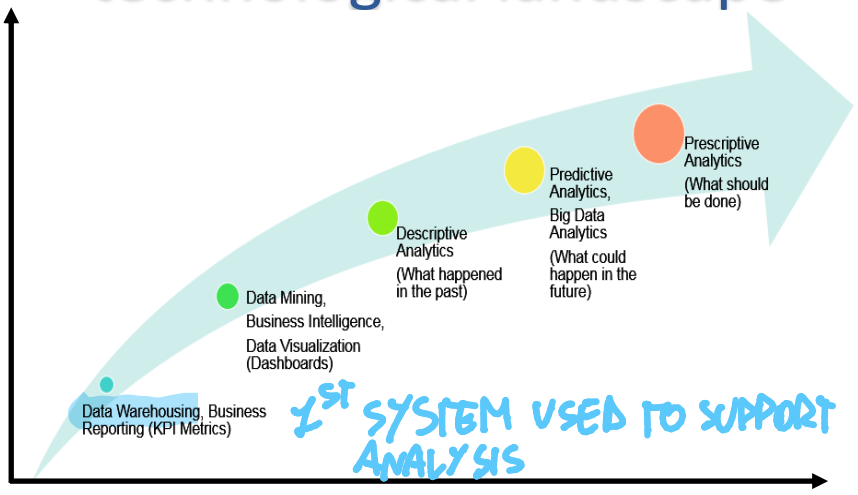
OLAP properties

- FASMI
  - Fast → FAST ANSWERS
  - Analytical → SUPPORT THE ANALYSIS THAT THE USER WANTS TO DO
  - Shared → DIFFERENT USERS CAN ACCESS THE DATA
  - Multidimensional → USE OF A M.D. MODEL
  - Informational → FOCUS ON RELEVANT INFORMATIONS

8

8

Data analysis – the technological landscape



9

Data Warehouse

10

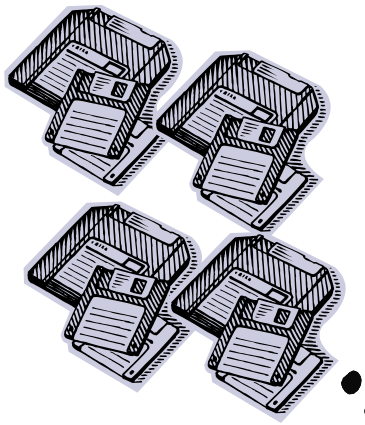
What is a Data Warehouse?

- Data should be **integrated across the enterprise(s)**
- **Summary data** provide real value to the organization
- **Historical data** hold the key to understanding data over time
- What-if capabilities are required

↓

CAN PROVIDE SOME TOOLS TO MAKE SIMULATIONS

What is a Data Warehouse?



A single, complete and consistent store of data obtained from a variety of different sources made available to end users, *so that they can understand and use it in a business context.*

[Barry Devlin]

GENERALLY, WE CAN HAVE TWO DIFFERENT BUT VALID DEFINITIONS:

- BIG REPOSITORIES WITH DATAS THAT SUPPORT DECISION
- PROCESS TO TRANSFORM DATAS IN INFORMATIONS

SUMMARIZE → DATA WAREHOUSING: PROCESS
DATA WAREHOUSE: REPOSITORY

An alternative definition of Data Warehouse

Data

A data warehouse is a process for *transforming data into information* and for making it available to users *in a timely enough manner to make a difference*.

[Forrester Research, April' 96]

13

Data Warehouse

As a dataset: decision support database maintained separately from the organization's operational database

As a process: technique for assembling and managing data from various sources with the purpose of answering business questions. Thus making decisions that were not previously possible

14

14

OLAP Properties

- OLAP systems are characterized by FASMI properties:
  - Fast
  - Analytical
  - Shared
  - Multidimensional
  - Informational

15

Data Warehouse

AN OLAP WITH SOME ADDITIONAL PROPERTIES:

- A Data Warehouse is a
  - subject-oriented: "the data contained in a data warehouse are primarily concerned with the main entities of interest for the analysis, such as products, customers, orders and sales" → DESIGN OF DATA WAREHOUSES AFFECTED BY USERS' REQUESTS → NOT FLEXIBLE AT ALL
  - Integrated: "The data originating from the different sources are integrated and homogenized as they are loaded into a data warehouse" → UNION OF DIFFERENT SOURCES
  - time-variant: "All data entered in a data warehouse are labelled with the time period to which they refer"
  - non-volatile (persisent): "Once they have been loaded into a data warehouse, data are usually not modified further and are held permanently "
- collection of data that is used primarily in organizational decision making.

[Bill Inmon, Building the Data Warehouse, 1996]

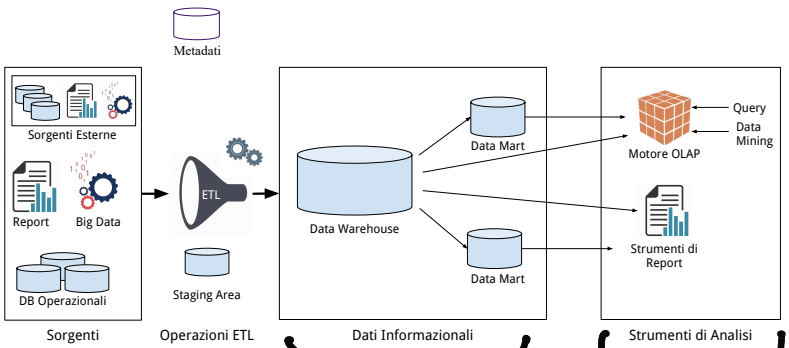
THE REAL DESIGN IS OBTAINED AFTER THE DATA INTEGRATION

16

Dimensions of a Data Warehouse

- Data warehouses are **very** large databases
  - Terabytes (10^{12} bytes)
  - Petabytes (10^{15} bytes): e.g. Geographic Information Systems
  - Exabytes (10^{18} bytes): e.g. National Medical Records
  - Zettabytes (10^{21} bytes): e.g. Weather reports, including images
  - Yottabytes (10^{24} bytes): e.g. Intelligence Agency Videos

Data Warehouse - architecture



SOURCES (INTERNAL; EXTERNAL)

ETL MODULE *

DATA WAREHOUSE → BIG-REPOSITORY

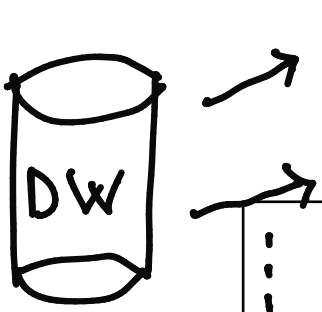
ANALYTIC TOOLS:
- REPORTING TOOLS
- DATA MINING TOOLS
- MACHINE LEARNING TOOLS

SHARED RESOURCES

PROBLEM: LIMIT THE RESPONSE TIME

*SEE NEXT PAGE

DISTRIBUTED ARCHITECTURE



DATA MART, TO STORE A REDUCED SET OF DATAS USEFUL FOR SOME SPECIFIC FIELDS

Where is a DW useful

- **Commerce**: sales and complaints analysis, client fidelization, shipping and stock control
- **Manufacturing plants**: production cost control, provision and order support
- **Financial services**: risk and credit card analysis, fraud detection
- **Telecommunications**: call flow analysis, subscribers' profiles
- **Healthcare structures**: patients' ingoing and outgoing flows, cost analysis

19

ETL (Extraction, Transformation, Loading)

- **Extraction**: data are extracted from the available internal and external sources.
- **Transformation**: the goal of the transformation phase is to improve the quality of the data extracted. Some of the main operations that are executed are:
  - Data Cleaning
  - Reconciliation, Entity Matching
  - Data standardization
  - Deduplication
- **Loading**: after being extracted and transformed data are loaded into the data warehoise

FOCUS ON RELEVANT DATA

DATAS IN THE SOURCE CHANGES ↓ PERIODIC EXTRACTION

DURING ETL, DATAS CHANGE THEIR REPRESENTATION

21

Metadata

- Metadata contain the following data:
  - Information about the data warehouse structure (e.g., dimensions, hierarchies, fact)
  - Information about values stored in the data warehouse: each attribute is characterized by its provenance, e.g., which is the data sources from which data were extracted and the transformations to which they have been subjected
  - Usage statistics of the data warehouse, e.g. number of accesses to a field
  - Description of the application domain and related data properties, data ownership and loading policies

Examples of data warehouse queries

- Show total sales across all products at increasing aggregation levels for a geography dimension, from state to country to region, for 2017 and 2018.
- Create a cross-tabular analysis of our operations showing expenses by territory in South America for 2017 and 2018. Include all possible subtotals.
- List the top 10 sales representatives in Asia according to sales revenue for automotive products in year 2018, and rank their commissions.

OLAP-oriented data models

- must support sophisticated analyses and computations over different dimensions and hierarchies
- Must guarantee fast responce time even to complex queries
- Most appropriate data model: multidimensional model

24

Dimensional Fact Model

- Allows one to describe a set of fact schemata
- The components of a fact schema are:
  - Facts
  - Measures
  - Dimensions
  - Dimension Hierarchy

25

Dimensional Fact Model

- A **fact** is a concept that is relevant for the decisional process; typically it models a set of events of the organization
- A **measure** is a numerical property of a fact → HOW TO MEASURE THE KPI
- A **dimension** is a fact property defined w.r.t. a finite domain; it describes an analysis coordinate for the fact, it is a perspective for analysing data
- **Dimension Hierarchy**: relate low-level (detailed) concepts to higher-level (general concepts)
  - Example: Store – City – Region/Province – Country

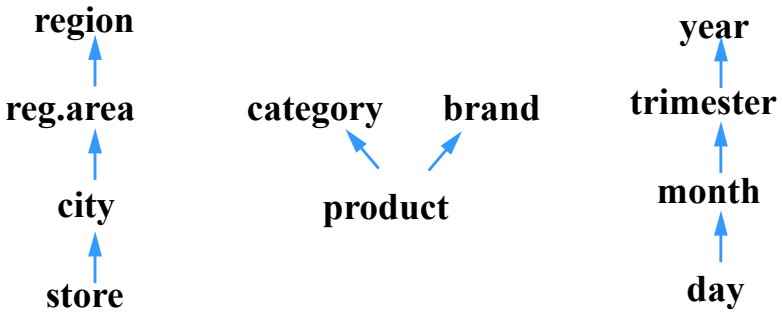
WHAT THE USER WANT TO MONITOR ↓ KPI

Dimensional Fact Model

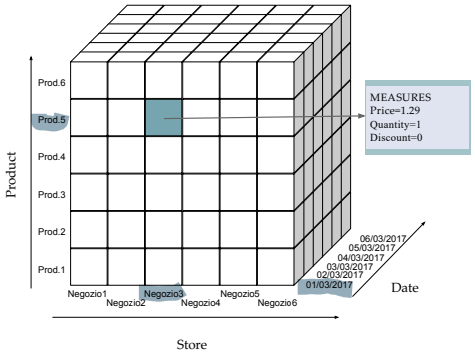
- The multidimensional view of data is represented as a data cube or an hypercube
- Cube dimensions are the search keys
- Each dimension may be hierarchical
  - DATE {DAY-MONTH-TRIMESTER-YEAR}
  - PRODUCT {BRAND - TYPE - CATEGORY}

    (e.g. LAND ROVER - CARS - VEHICLES)
- Cube cells contain metric values

Dimensions and hierarchies



Example sales



* ESSELUNGA S.p.A. *
VIA CORRIDONI n.30 - BERGAMO (BG)
P.I.: 04916380159

	EURO
GSK DENT. AQUAFRESH	1,19
GSK DENT. AQUAFRESH	0,99
GSK DENT. AQUAFRESH	0,99
GSK DENT. AQUAFRESH	0,99
GSK DENT. AQUAFRESH	0,99
SCHIACCIATE OLIV&MARI	1,29
CRACKERS OLIVIA&MARIN	1,79
GRISS POMODORO OL&MAR	1,49
8 x 2,99	
BONDUE. COCCOLE SPINA	23,92
SCONTO FIDATY 30%	7,20-S

sconTomaggio.com

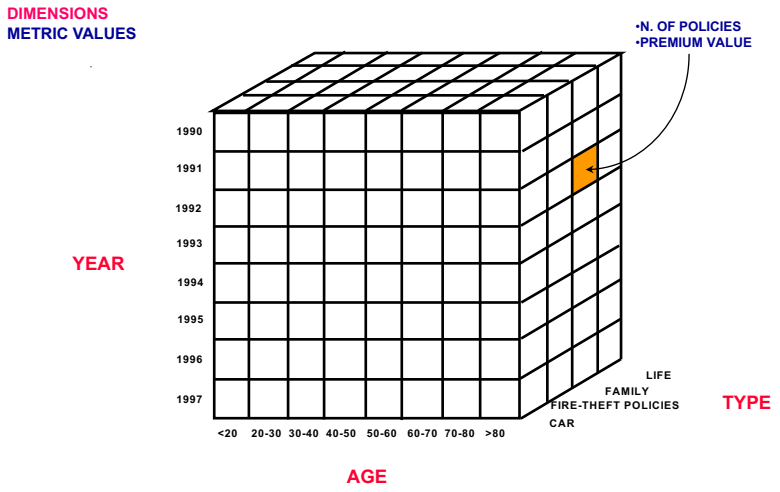
TOTALE EURO	26,44 *
PAGAMENTO SCONTO EURO	0,04
PAGAMENTO BUONI SCONTO	9,50
PAGAMENTO BANCOMAT	16,90
RESTO	0,00 *

N. CARTA FIDATY: 040********92

---------- NUOVA RACCOLTA PUNTI ----------

SALDO AD OGGI PUNTI	4.744
PUNTI SULLA SPESA	42
TOT.PUNTI FIDATY	42
NUOVO SALDO PUNTI	4.786

Example: An Insurance Company Data Cube



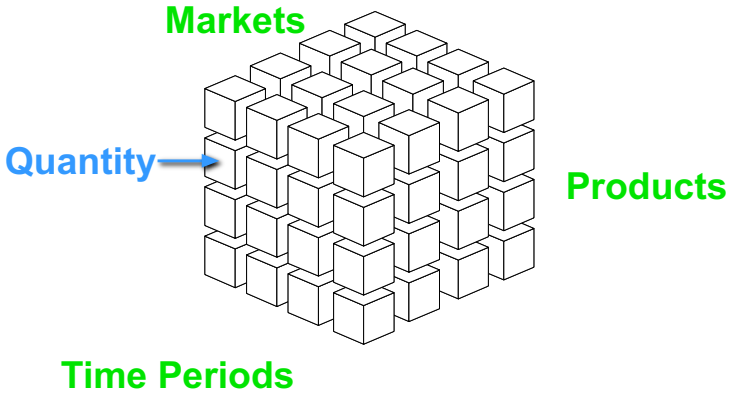
30

Examples

- Store chain
  - Fact: sales
  - Measures: sold quantity, gross income
  - Dimensions: product, time, zone
- Telecom Operator
  - Fact: phone call
  - Measures : cost, duration
  - Dimensions: caller subscriber, called subscriber, time

31

Multidimensional Representation



Sales Cube

32

FROM A DETAILED TO A SPECIFIC AN

OLAP operations

- Slice/Dice → COLLECT ALL THE FACTS WITHIN A CERTAIN PARAMETER
- Roll up/Drill down

↓

FROM A DETAILED TO A SPECIFIC ANALYSIS

EXAMPLE: SALES SINGLE STORE

↓

WHOLE NATION

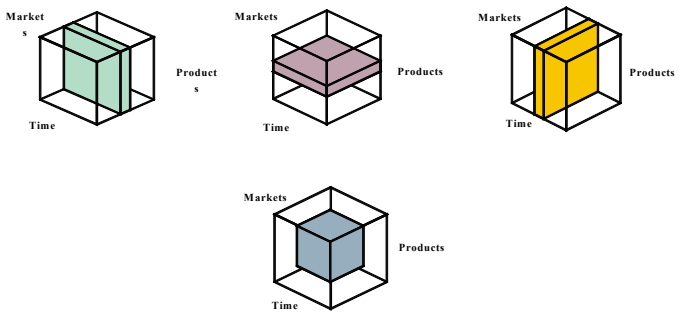
33

Typical Olap Operations

- **slice-and-dice**
  - The **slice** operation performs a selection on one dimension of the given cube, resulting in a subcube
  - The **dice** operation defines a subcube by performing a selection on two or more dimension
- **roll-up**
  - Aggregates data at a higher level – e.g. last year's sales volume per product category and per region
- **drill-down**
  - De-aggregates data at the lower level – e.g. for a given product category and a given region, show daily sales
- **pivoting**
  - Selects two dimensions to re-aggregate data (cube re-orientation)
- **ranking**
  - Sorts data according to predefined criteria
- traditional operations (select, project, join, derived attributes, etc.)

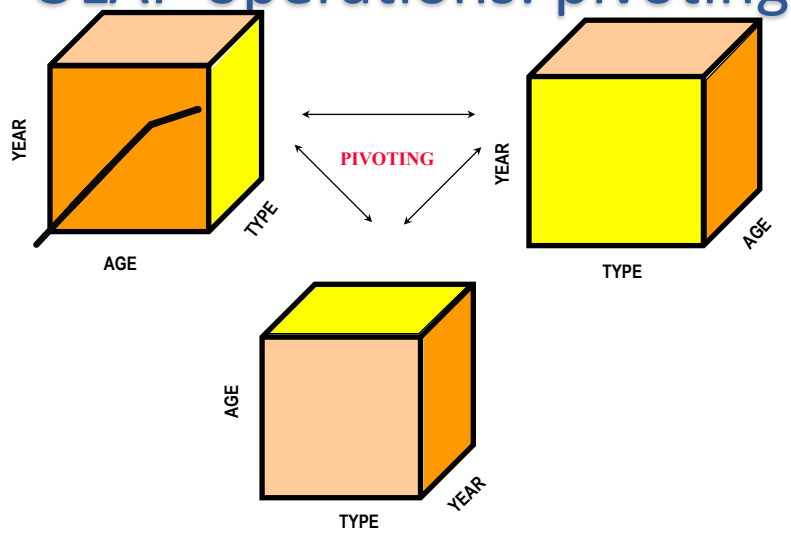
34

Slice/Dice operations



35

OLAP operations: pivoting



Pivoting

Fact table view:

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

Multi-dimensional cube:

day 2

	c1	c2	c3
p1	44	4	

day 1

	c1	c2	c3
p1	12		50
p2	11	8	

	c1	c2	c3
p1	56	4	50
p2	11	8	

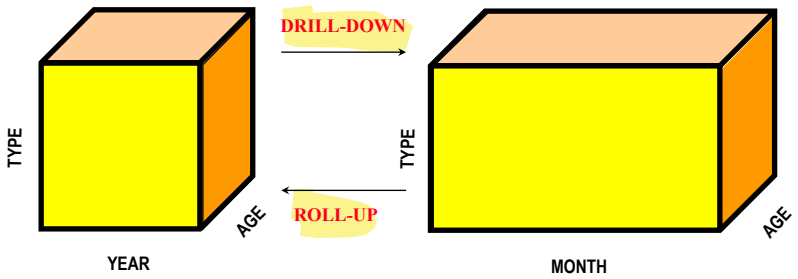
Hector Garcia Molina, Data warehouse and OLAP

Category	Year	Metrics	Dollar Sales
Electronics	1997		$ 10.616
	1998		$ 29.299
Food	1997		$ 5.300
	1998		$ 5.638
Gifts	1997		$ 16.315
	1998		$ 20.047
Health & Beauty	1997		$ 6.042
	1998		$ 5.665
Household	1997		$ 38.383
	1998		$ 50.391
Kid's Korner	1997		$ 2.559
	1998		$ 2.943
Travel	1997		$ 4.497
	1998		$ 4.792

	Metrics	Dollar Sales	
	Year	1997	1998
Category			
Electronics		$ 10.616	$ 29.299
Food		$ 5.300	$ 5.638
Gifts		$ 16.315	$ 20.047
Health & Beauty		$ 6.042	$ 5.665
Household		$ 38.383	$ 50.391
Kid's Korner		$ 2.559	$ 2.943
Travel		$ 4.497	$ 4.792

38

OLAP OPERATIONS



39

Time

ID	Day	Month	Trimester	Year
T1	31/3/2018	March	1	2018
T2	02/4/2018	April	2	2018

Product

ID	Brand	Type	Category
P1	M1	Milk	Food
P2	M2	Bread	Food

Sales

Time	Product	Store	Price
T1	P1	PV1	3,5
T1	P2	PV2	4
T1	P1	PV3	4
T2	P2	PV3	5
T2	P2	PV1	4
T2	P1	PV2	3

Store

ID	Address	City	Region	State
PV1	Via Roma 1	Milan	Lombardia	Italy
PV2	Via Milano 4	Rome	Lazio	Italy
PV3	Via Torino 5	Milan	Lombardia	Italy

40

Sales Roll Up

Time	Product	Store	Price
T1	P1	PV1	3,5
T1	P2	PV2	4
T1	P1	PV3	4
T2	P2	PV3	5
T2	P2	PV1	4
T2	P1	PV2	3

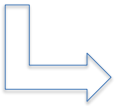
T1	P1	Lombardia	7,5
T1	P2	Lazio	4
T2	P2	Lombardia	9
T2	P1	Lazio	3

Amount of sales per region

41

Roll up

T1	P1	Lombardia	7,5
T1	P2	Lazio	4
T2	P2	Lombardia	9
T2	P1	Lazio	3



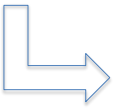
2018	P1	10,5
2018	P2	13

Amount of sales per year and product without considering the store

42

Roll up

2018	P1	10,5
2018	P2	13



2018	23,5

Amount of sales per year

43

Drill down

2018	23,5

2018	P1	10,5
2018	P2	13

Amount of sales per year and product

Drill down

2018	P1	10,5
2018	P2	13

1/2018	P1	7,5
1/2018	P2	4
2/2018	P1	3
2/2018	P2	9

Amount of sales per trimester

Roll-up

Month	Metrics / Customer Region	North-East	Mid-Atlantic	South-East	Central	South	North-West	South-West	England	France	Germany
	Dollar Sales										
Jan 97		$ 620	$ 753	$ 30	$ 660	$ 2.405	$ 1.312	$ 440	$ 1.002	$ 1.002	$ 383
Feb 97		$ 258	$ 252	$ 800	$ 975	$ 160	$ 582	$ 744	$ 310	$ 799	$ 118
Mar 97		$ 648	$ 244	$ 148	$ 250	$ 1.085	$ 2.961	$ 650	$ 1.240	$ 119	$ 142
Apr 97		$ 787	$ 588	$ 447	$ 486	$ 226	$ 506	$ 601	$ 119	$ 550	$ 85
May 97		$ 1.350	$ 245	$ 936	$ 159	$ 664	$ 626	$ 107	$ 135	$ 200	$ 177
Jun 97		$ 842	$ 582	$ 1.281	$ 937	$ 240	$ 774	$ 176	$ 1.139	$ 652	$ 254
Jul 97		$ 652	$ 690	$ 486	$ 1.293	$ 605	$ 303	$ 818	$ 103	$ 124	$ 173
Aug 97		$ 1.783	$ 304	$ 1.032	$ 170	$ 398	$ 356	$ 432	$ 190	$ 241	$ 407
Sep 97		$ 581	$ 778	$ 3.558	$ 587	$ 440	$ 1.652	$ 1.071	$ 315	$ 210	$ 202
Oct 97		$ 2.291	$ 1.840	$ 600	$ 656	$ 1.300	$ 718	$ 1.210	$ 427	$ 220	$ 520
Nov 97		$ 39	$ 1.602	$ 1.082	$ 1.187	$ 842	$ 759	$ 745	$ 232	$ 101	$ 1.037
Dec 97		$ 381	$ 1.588	$ 343	$ 118	$ 1.459	$ 635	$ 2.021	$ 259	$ 210	$ 119
Jan 98		$ 311	$ 1.174	$ 2.634	$ 3.130	$ 954	$ 2.083	$ 1.351	$ 747	$ 426	$ 447
Feb 98		$ 2.518	$ 702	$ 1.123	$ 1.336	$ 1.227	$ 3.887	$ 545	$ 268	$ 277	$ 282
Mar 98		$ 2.459	$ 1.523	$ 1.178	$ 4.708	$ 1.420	$ 3.514	$ 1.948	$ 1.705	$ 276	$ 1.168
Apr 98		$ 407	$ 841	$ 524	$ 712	$ 133	$ 2.486	$ 49	$ 390	$ 1.298	$ 221
May 98		$ 667	$ 1.721	$ 440	$ 148	$ 80	$ 1.310	$ 303	$ 104	$ 657	$ 65
Jun 98		$ 699	$ 1.096	$ 898	$ 353	$ 902	$ 839		$ 230	$ 155	$ 105
Jul 98		$ 586	$ 1.897	$ 412	$ 226	$ 406	$ 361	$ 1.628	$ 267	$ 1.011	$ 41
Aug 98		$ 894	$ 326	$ 792	$ 1.832	$ 1.199	$ 295	$ 1.816	$ 277	$ 102	$ 118
Sep 98		$ 338	$ 3.179	$ 505	$ 427	$ 99	$ 2.976	$ 885	$ 135	$ 85	$ 1.110
Oct 98		$ 544	$ 413	$ 1.467	$ 209	$ 679	$ 706	$ 556	$ 480	$ 485	$ 99
Nov 98		$ 671	$ 459	$ 1.471	$ 2.066	$ 701	$ 716	$ 986	$ 1.127	$ 154	$ 440
Dec 98		$ 836	$ 2.096	$ 1.726	$ 3.642	$ 395	$ 1.740	$ 1.943	$ 1.143	$ 366	$ 307

Quarter	Metrics / Customer Region	North-East	Mid-Atlantic	South-East	Central	South	North-West	South-West	England	France	Germany
	Dollar Sales										
Q1 1997		$ 1.526	$ 1.249	$ 978	$ 1.885	$ 3.650	$ 4.855	$ 1.834	$ 2.552	$ 1.920	$ 643
Q2 1997		$ 2.979	$ 1.415	$ 2.664	$ 1.582	$ 1.130	$ 1.906	$ 884	$ 1.393	$ 1.402	$ 516
Q3 1997		$ 3.016	$ 1.772	$ 5.076	$ 2.050	$ 1.443	$ 2.311	$ 2.321	$ 608	$ 575	$ 782
Q4 1997		$ 2.711	$ 5.030	$ 2.025	$ 1.961	$ 3.601	$ 2.112	$ 3.976	$ 918	$ 531	$ 1.676
Q1 1998		$ 5.288	$ 3.399	$ 4.935	$ 9.174	$ 3.601	$ 9.484	$ 3.844	$ 2.720	$ 979	$ 1.897
Q2 1998		$ 1.773	$ 3.658	$ 1.862	$ 1.213	$ 1.115	$ 4.635	$ 352	$ 724	$ 2.110	$ 391
Q3 1998		$ 1.818	$ 5.402	$ 1.709	$ 2.485	$ 1.704	$ 3.632	$ 4.329	$ 679	$ 1.198	$ 1.269
Q4 1998		$ 2.051	$ 2.968	$ 4.664	$ 5.917	$ 1.775	$ 3.162	$ 3.485	$ 2.750	$ 1.005	$ 846

46

Roll-up

Category	Year	Metrics / Customer Region	North-East	Mid-Atlantic	South-East	Central	South	North-West	South-West	England	France
		Dollar Sales									
Electronics	1997		$ 138	$ 1.774	$ 384	$ 138	$ 2.346	$ 2.554	$ 2.184	$ 566	$ 199
	1998		$ 1.184	$ 4.529	$ 1.892	$ 7.232	$ 651	$ 9.488	$ 476	$ 2.683	$ 462
Food	1997		$ 759	$ 682	$ 729	$ 262	$ 588	$ 469	$ 807	$ 156	$ 615
	1998		$ 530	$ 925	$ 959	$ 677	$ 213	$ 1.503	$ 261	$ 165	$ 175
Gifts	1997		$ 2.532	$ 1.355	$ 1.854	$ 1.413	$ 2.535	$ 2.132	$ 1.904	$ 908	$ 375
	1998		$ 1.955	$ 2.785	$ 2.800	$ 2.695	$ 1.813	$ 2.844	$ 1.778	$ 1.158	$ 717
Health & Beauty	1997		$ 624	$ 640	$ 1.317	$ 647	$ 588	$ 754	$ 654	$ 143	$ 292
	1998		$ 611	$ 887	$ 566	$ 382	$ 499	$ 1.162	$ 1.044	$ 273	$ 72
Household	1997		$ 5.354	$ 4.112	$ 5.410	$ 4.446	$ 3.058	$ 3.974	$ 2.654	$ 3.545	$ 2.875
	1998		$ 5.787	$ 5.320	$ 5.416	$ 6.812	$ 4.334	$ 5.008	$ 7.588	$ 2.139	$ 3.649
Kid's Korner	1997		$ 201	$ 398	$ 485	$ 186	$ 409	$ 323	$ 396	$ 105	$ 34
	1998		$ 247	$ 422	$ 441	$ 380	$ 221	$ 592	$ 290	$ 198	$ 19
Travel	1997		$ 624	$ 505	$ 564	$ 386	$ 300	$ 978	$ 416	$ 48	$ 38
	1998		$ 608	$ 559	$ 1.096	$ 611	$ 464	$ 316	$ 573	$ 257	$ 198

Category	Year	Metrics	Dollar Sales
Electronics	1997		$ 10.616
	1998		$ 29.299
Food	1997		$ 5.300
	1998		$ 5.638
Gifts	1997		$ 16.315
	1998		$ 20.047
Health & Beauty	1997		$ 6.042
	1998		$ 5.665
Household	1997		$ 38.383
	1998		$ 50.391
Kid's Korner	1997		$ 2.550
	1998		$ 2.943
Travel	1997		$ 4.497
	1998		$ 4.792

47

Drill-down

Metrics	Dollar Sales	
Year	1997	1998
Category		
Electronics	$ 10.616	$ 29.299
Food	$ 5.300	$ 5.638
Gifts	$ 16.315	$ 20.047
Health & Beauty	$ 6.042	$ 5.665
Household	$ 38.383	$ 50.391
Kid's Corner	$ 2.559	$ 2.943
Travel	$ 4.497	$ 4.792

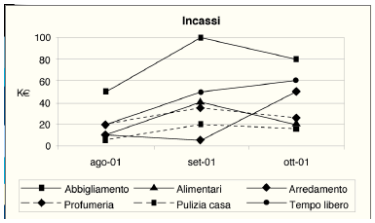
Metrics	Dollar Sales											
Customer Region	North-East		Mid-Atlantic		South-East		Central		South		North-West	
Year	1997	1998	1997	1998	1997	1998	1997	1998	1997	1998	1997	1998
Category												
Electronics	$ 138	$ 1.184	$ 1.774	$ 4.529	$ 384	$ 1.892	$ 138	$ 7.232	$ 2.346	$ 651	$ 2.554	$ 9.488
Food	$ 759	$ 538	$ 682	$ 925	$ 729	$ 959	$ 262	$ 677	$ 588	$ 213	$ 469	$ 1.503
Gifts	$ 2.532	$ 1.955	$ 1.355	$ 2.785	$ 1.854	$ 2.800	$ 1.413	$ 2.695	$ 2.535	$ 1.813	$ 2.132	$ 2.844
Health & Beauty	$ 624	$ 611	$ 640	$ 887	$ 1.317	$ 566	$ 647	$ 382	$ 588	$ 499	$ 754	$ 1.162
Household	$ 5.354	$ 5.787	$ 4.112	$ 5.320	$ 5.410	$ 5.416	$ 4.446	$ 6.812	$ 3.058	$ 4.334	$ 3.974	$ 5.008
Kid's Corner	$ 201	$ 247	$ 398	$ 422	$ 485	$ 441	$ 186	$ 380	$ 409	$ 221	$ 323	$ 592
Travel	$ 624	$ 608	$ 505	$ 559	$ 564	$ 1.096	$ 386	$ 611	$ 300	$ 464	$ 978	$ 316

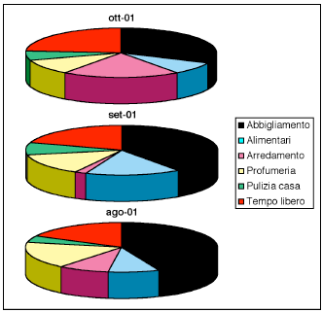
48

Visualization and Reports

- Data may be visualized graphically, in an Excel-like format: tables, hystograms, graphics, 3D surfaces, etc.

incassi (K€)	Ottobre 2001	Settembre 2001	Agosto 2001
Abbigliamento	80	100	50
Alimentari	20	40	10
Arredamento	50	5	10
Profumeria	25	35	20
Pulizia casa	15	20	5
Tempo libero	60	50	20





51

Aggregate Queries

Examples:

- Total sales per product category, per supermarket, per day
- Total monthly sales for all the products, per supermarket
- Total monthly sales per category per supermarket
- Avg. monthly sales per category, for all supermarkets

POLITECNICO
MILANO 1863

Data Warehouse design

Cinzia Cappiello
A.A. 2023-2024

The problem

WE CAN'T USE TRADITIONAL DATA BASES BECAUSE THEY ARE DESIGNED FOR TRANSACTION PURPOSE, WHILE WE NEED ANALYTICAL INSTRUMENTS

Relational DBs have the following problems:

Complexity o f the applications

High response time for answering to complex queries

Consequences

Raw data are used at the operations level

Raw data are scarcely used at the strategic level

Data Warehouse design

MODEL → MULTIDIMENSIONAL → DECREASE RESPONSE TIME → DESIGN BASED ON USER REQUIREMENTS

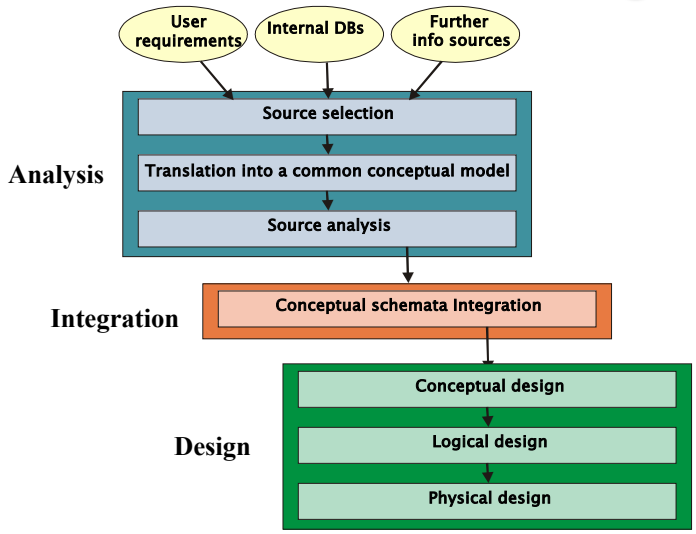
- The design of a data warehouse is different from the design of a traditional db
  - Data have different characteristics
  - Design is based on the available data sources
  - Design is driven by different criteria
- The design of a data warehouse aims to maintain a low number of entities but high coverage

↓

INCLUDE ALL RELEVANT DATAS BUT MINIMIZE ENTITIES

3

Data Warehouse Design

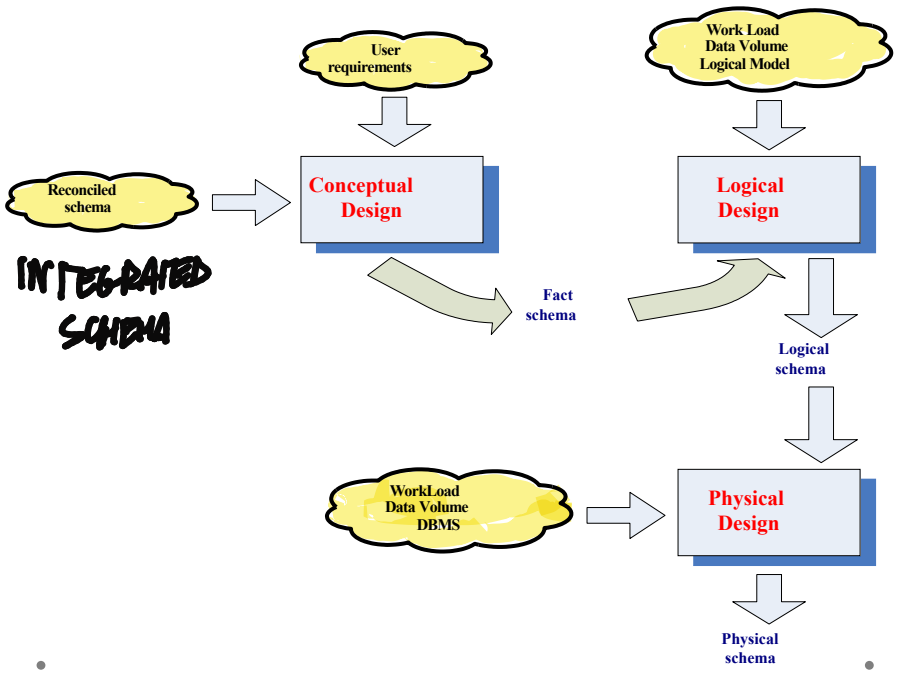


4

Data Warehouse Design

- Data Warehouses are based on the **multidimensional model**
- A standard conceptual model for DW does not exist
- The Entity/Relationship model cannot be used in the DW conceptual design

5



6

Requirements elicitation

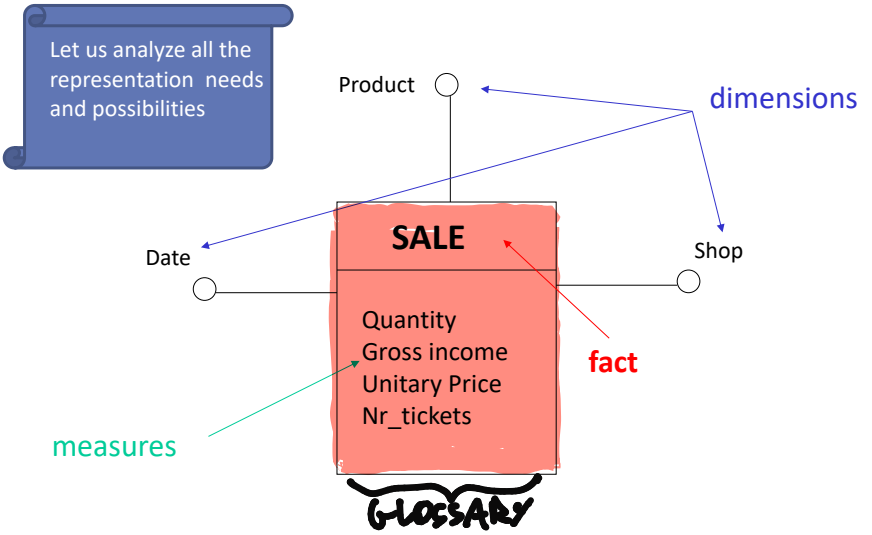
- In order to select facts it is important to understand which are the users requirements AND WHICH KPI ARE INTERESTING FOR HIM
- Requirements elicitation is conducted by interviewing the people that have to perform the analysis

Conceptual Model

DOMANDA ESAME FEBBRAIO 2017

Fact Schema

Let us analyze all the representation needs and possibilities



9

INPUT: CONCEPT MODEL OF DB
↓
IDENTIFY FACT, GENERALLY ENTITIES ASSOCIATED WITH N:N RELATIONSHIP

From E/R to Dimensional Fact Model (DFM)

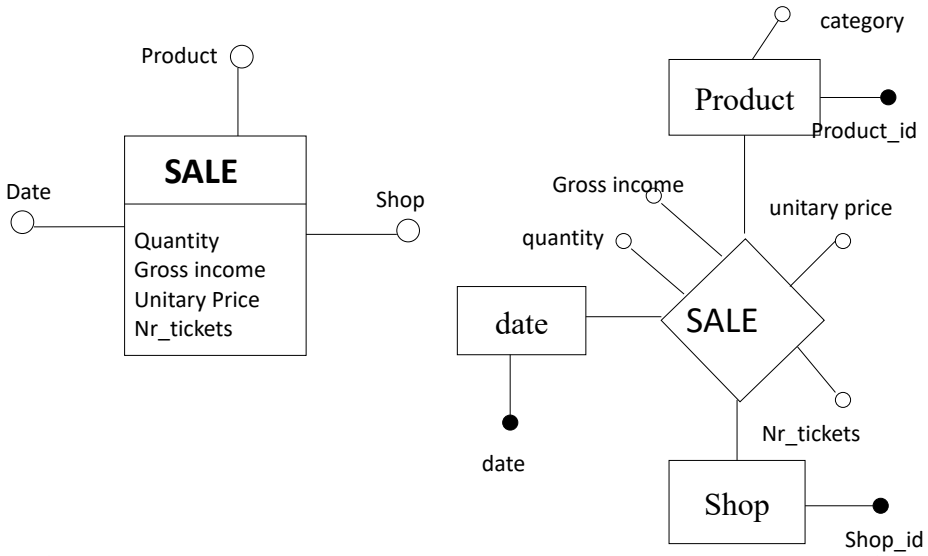
- A fact describes an entity or an N to M relationship among its dimensions. Entities that are often updated (e.g., sales) are good candidate for being transformed in facts.
- The fact value must uniquely determine the value of each dimension, e.g. a sale uniquely determines the day in which it has been done. This is represented as

  sale → day, month, year

- Naming convention: the dimensions of a same fact schema must have distinct names

10

DFM and E/R



11

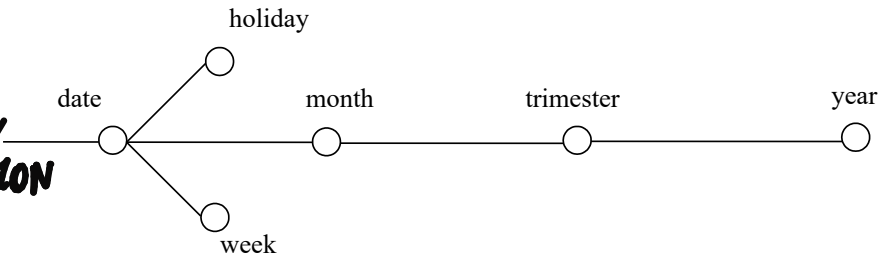
Dimensional attribute

- A dimensional attribute must assume discrete values, so that it can contribute to represent a dimension
- Dimensional attributes can be organized into hierarchies

12

Hierarchy

- A dimensional hierarchy is a directional tree where
  - Nodes are dimensional attributes
  - Edges describe n:1 associations between pairs of dimensional attributes
  - Root is the considered dimension



LOWEST GRANULARITY OF AN INFORMATION

REPRESENTATION OF A 1:N RELATIONSHIP

Events and aggregations

- A primary event is an occurrence of a fact; it is represented by means of a tuple of values
  - ✓ On 10/10/2001, ten 'Brillo' detergent packets were sold at the BigShop for a total amount of 25 euros

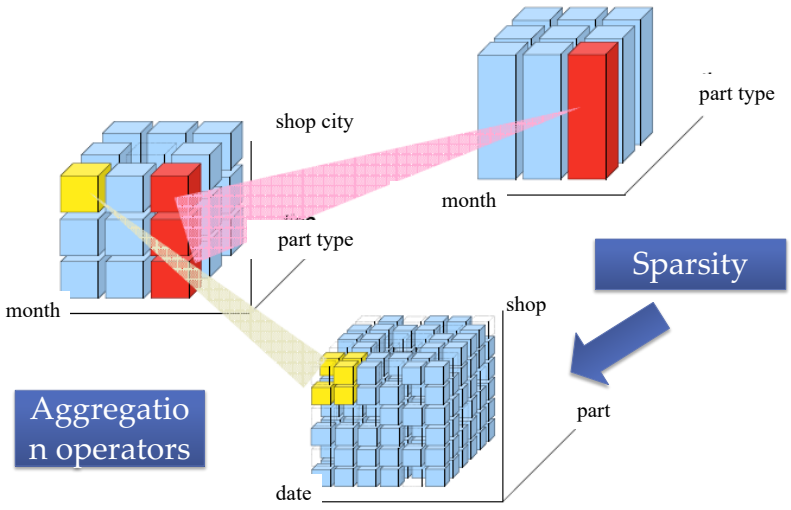
ALL THE ONES STORED IN THE DATABASE

Events and aggregations (2)

- A hierarchy describes how it is possible to group and select primary events
- The root of a hierarchy represents the finest aggregation granularity present in the warehouse (e.g.sales one by one, or by day, or by week, depending on what the designer deems appropriate)

15

Events and aggregations



16

Events and aggregations (3)

- Given a set of dimensional attributes (pattern), each tuple of their values identifies a secondary event that aggregates (all) the corresponding primary events
- For each dimensional attribute, a value is associated with the secondary event; this value summarizes the values taken by the corresponding measure in the primary events
- For example the sales can be grouped by Product and Month:
  - ✓ in October 2001, 230 'Brillo' detergent packets were sold at the BigShop for a total amount of 575 euros

17

Secondary event

- The sales can be further grouped by Product, Month, and City
- If we consider city, product and month as dimensional attributes, the tuple

  (city: 'Rome', product: 'Brillo', month: 10/2001)

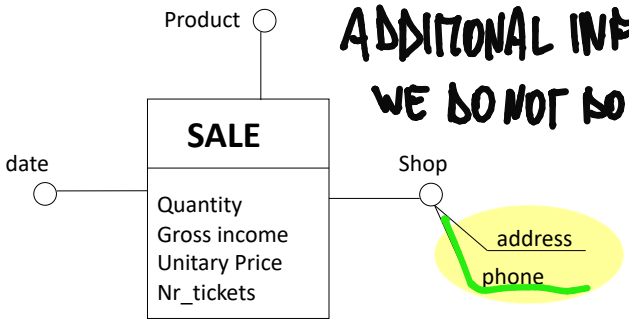
  identifies another secondary event
- It aggregates all the sales related to the product 'Brillo' in shops of 'Rome' during the month October 2001

OTHER EVENTS THAT CAN BE STORED USING DIFFERENT FACTS

18

Descriptive attributes

- A descriptive attribute contains additional information about a dimensional attribute
- They are uniquely determined by the corresponding dimensional attribute
- They are relevant for analytical purposes only as selection predicates

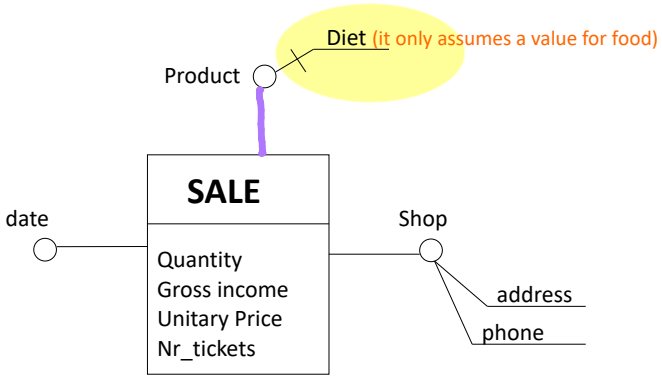


ADDITIONAL INFORMATIONS IN WHICH WE DO NOT DO ANALYSIS

19

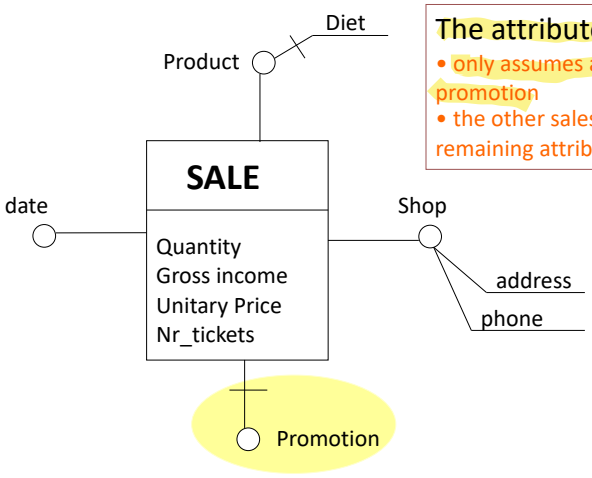
Optional edges

- Some edges of a fact schema could be optional



20

Optional dimensions



The attribute Promotion:
- only assumes a value for products in promotion
- the other sales are characterized by the remaining attributes

Cross-dimensional attributes

- A cross-dimensional attribute is a dimensional or a descriptive attribute whose value is obtained by combining values of some dimensional attributes
  - ✓ For example, IVA (VAT) is computed based on the product category and the state

Convergence

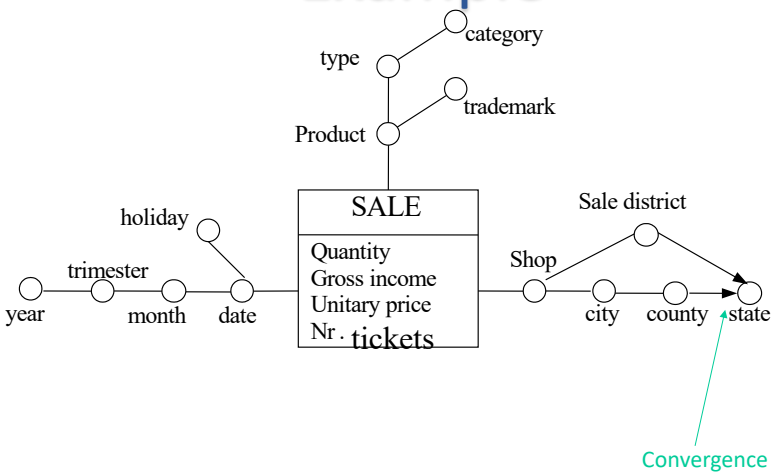
- It is related to the structure of a hierarchy
  - ✓ Two dimensional attributes can be connected by more than two distinct directed edges
  - ✓ For example:

Shop → city → county→ state

or

Shop→ sale district → state

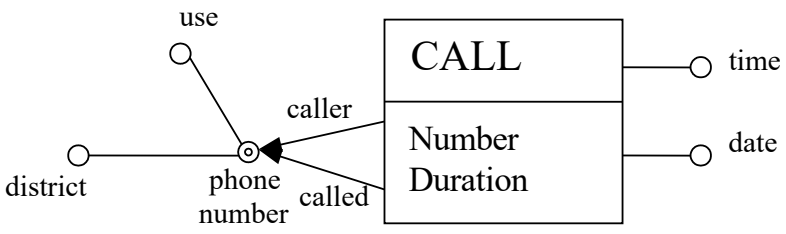
Example



Hierarchy Sharing

- In a fact schema, some portions of a hierarchy might be duplicated
- As a shorthand we allow hierarchy sharing
- If the sharing starts with a dimension attribute, it is necessary to indicate the roles on the incoming edges
- Necessary condition: the unicity of the value must hold on both branches

Hierarchy Sharing

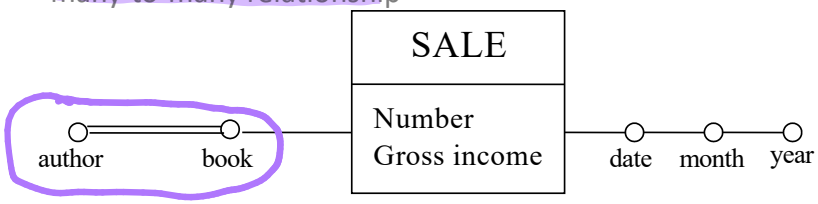


It is in fact a shorthand to represent
the duplication of the whole hierarchy

N:N RELATION BETWEEN 2 NODES, REPRESENTED BY A DOUBLE EDGE

Multiple edges

- Recall: the dimension values must be uniquely determined by the fact
- Some attributes, or some dimensions, may be related by a many-to-many relationship



- we denote them by multiple edges
- they are dealt with in a special way at logical design time

27

Measure Aggregation

- Aggregation requires to specify an operator to combine values related to primary events into a unique value related to a secondary event (e.g. sum of sold quantity aggregated by month)
- A measure is additive w.r.t. a given dimension iff the SUM operator is applicable to that measure along that dimension

28

DOMANDA ESAME LUGLIO 2017

Measure Classification: Additivity

SUM, AVG, MIN, MAX OVER BOTH TEMPORAL (AND NOME) HIERARCHIES

DEFAULT

- Additive measures (flow or rate measures): Can be meaningfully summarized using addition along all dimensions
  - E.g., sales amount can be summarized when the hierarchies in Store, Time, and Product dimensions are traversed

→ RELATED TO TIME PERIOD

AVG, MIN, MAX AS BEFORE
SUM ONLY OVER NON-TEMPORAL HIERARCHIES

- Semiadditive measures (stock or level measures): Can be meaningfully summarized using addition along some (not all) dimensions
  - E.g., inventory quantities, can be aggregated in the Store dimension, but cannot be aggregated in the Time dimension

→ PARTICULAR TIME INSTANT

- Nonadditive measures (value-per-unit measures): Cannot be meaningfully summarized using addition along any dimension
  - E.g., item price, cost per unit, exchange rate

→ PARTICULAR TIME INSTANT (BUT THEY ARE RELATIVES)

ONLY AVG, MIN, MAX

Elzbieta Malinowski & Esteban Zimányi 2008

29

29

The n.of tickets is non-additive (and in general non-aggregable) w.r.t. the product

- By n. of tickets we mean the n. of "buyings" i.e. the ticket count
- The association between product and ticket is many-to-many
- E.g. by summing up the ticket count on the product type we count the same type twice if it is the type of products that are in the same ticket

Ticket	Product	Type
S1	P1	T1
S1	P2	T1
S2	P1	T1
S2	P3	T2

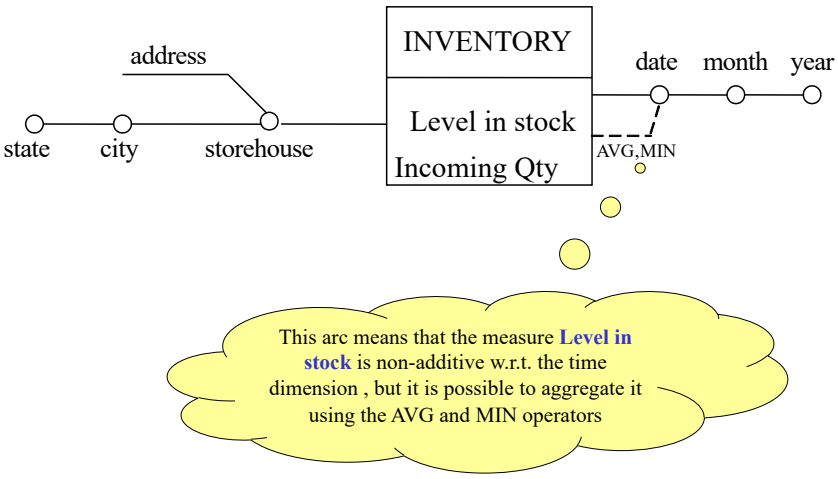
how many tickets containing p1 ? → 2
how many tickets containing p2 ? → 1
how many tickets containing p3 ? → 1
how many tickets with products of type t1 ? → 2

BUT

Sum(tickets with type(product) =t1) = 3 !!!

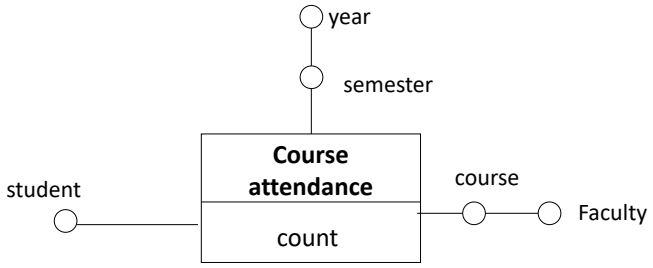
30

Aggregability



35

Empty fact schemata



A fact schema is empty if there are no measures.
In fact, the default measure is the count

EXAMPLE: "SALE = N.PRODUCT * AMOUNT"

37

Conceptual design

Conceptual design

- Conceptual design takes into account the documentation related to the integrated, reconciled input database
  - Conceptual schema (e.g. Entity/Relationship)
  - Logical schema (e.g. relational, XML…)

INPUT:
- USER REQUIREMENTS
- INFORMATIONS FROM THE INTEGRATED SCHEMA

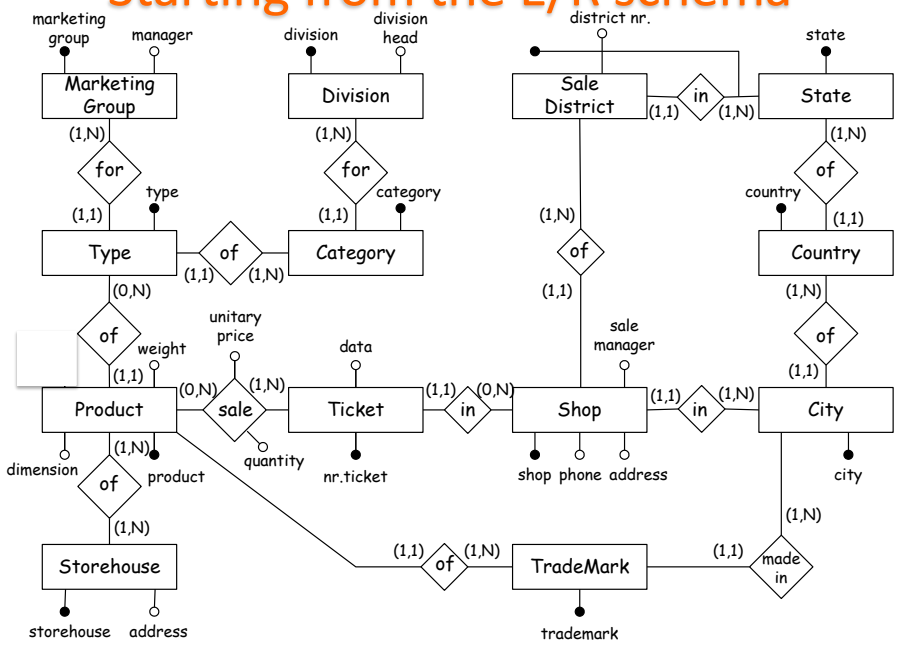
Top-down methodology

1. Fact definition (a subject oriented collection of data !!)
2. For each fact:
   1. Attribute tree definition
   2. Attribute tree editing
   3. Dimension definition
   4. Measure definition
   5. Fact schema creation

HOW TO IDENTIFY FACTS:

- n:n RELATIONSHIP
- DYNAMIC ENTITIES (ONES WITH HIGH AMOUNT OF UPDATES)
- WEAK ENTITIES

Starting from the E/R schema



Starting from the Relational Schema

Product(product,weight,dimension,trademark:TradeMark,type:Type)
Shop(shop,address,phone,salemanager,(ditrictnr,state):District,city:City)
Ticket(nrticket,date,shop:Shop)
Sale(product:Product,nrticket:Ticket,quantity,unitaryprice)
Storehouse(storehouse,address)
City(city,country:Country)
Country(country,state:State)
State(state)
District(district,state:State)
Prod_Storehouse(product:Product,storehouse:Storehouse)
TradeMark(trademark,madein:City)
Type(type,marketinggroup:MarketingGroup,category:Category)
MarketingGroup(marketinggroup,manager)
Category(category,division:Division)
Division(division,divisionhead)

Fact definition

- Facts correspond to events that dynamically happen in the organization
  - In an E/R schema, it can correspond to an entity F or to an association among *n* entities $E_1, E_2, ..., E_n$
  - In a relational schema, a fact corresponds to a relation (table) R

Fact definition

- Good fact candidates: entities or relationships representing **frequently updated data**
- Static archives: **NO!**
- **Remark**: when a fact is identified, it becomes the root of a new fact schema

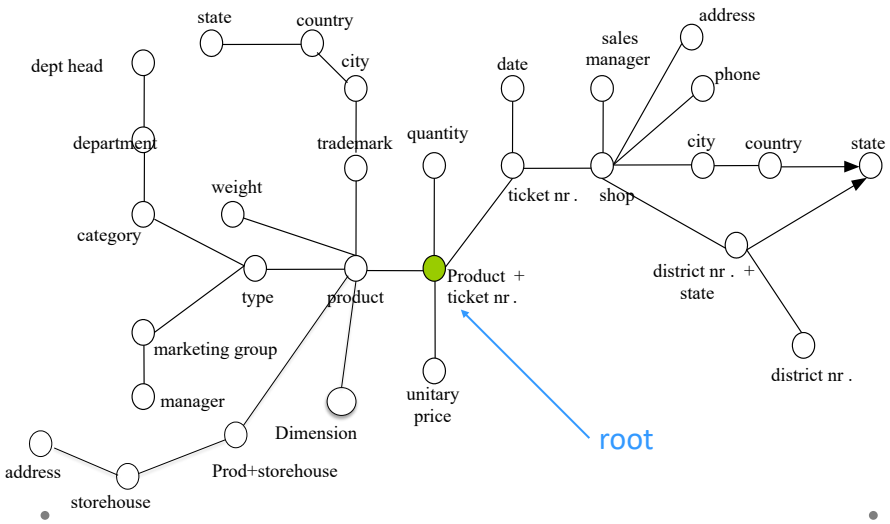
44

Attribute tree definition

- The attribute tree is composed by:
  - Nodes, corresponding to attributes (simple or complex) of the source schema
  - Root, corresponding to the primary key of the fact F
  - For each node, the corresponding attribute uniquely determines its descendant attributes

45

Attribute tree: example

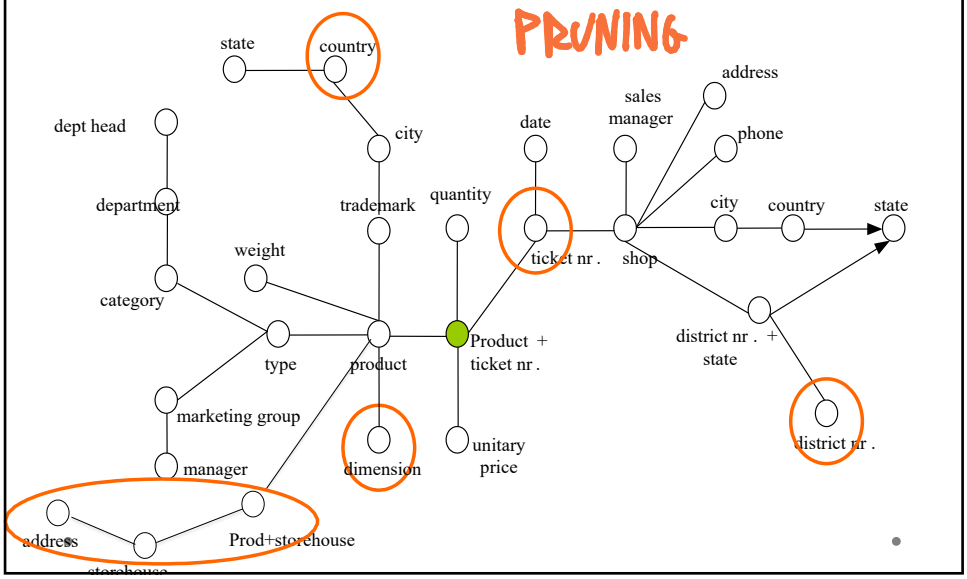


Attribute tree editing

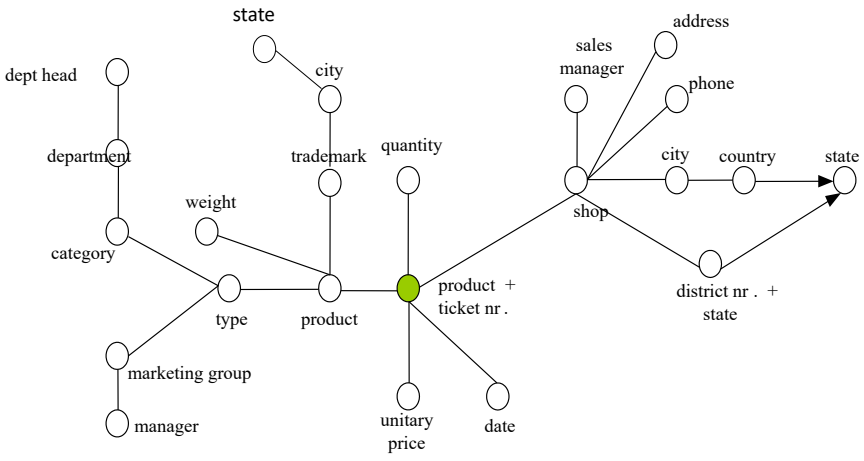
- The editing phase allows to remove some attributes which are irrelevant for the data mart
  - Pruning of a node v: the subtree rooted in v is deleted
  - Grafting of a node v: the children of v are directly connected to the father of v

WHICH DATAS ARE RELEVANT?
WHAT CAN WE DELETE?

Attribute tree editing: example



Attribute tree editing: example

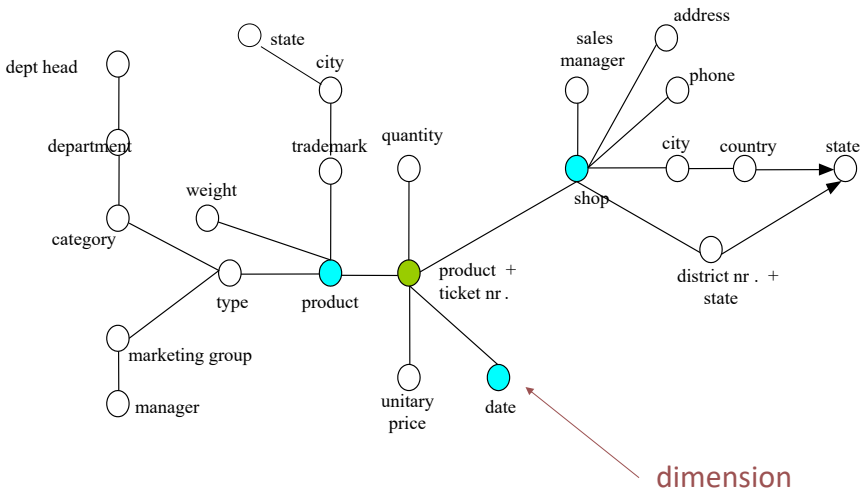


Dimension definition

- Dimensions can be chosen among the children of the root
- Time should always be a dimension
  - Historical source: time is an attribute
  - Snapshot source: not always time is directly represented. In this case it is necessary to add time.

IMPORTANT FOR THE EXAM: IF YOU MISS IT YOU'LL 100% FAIL :-)

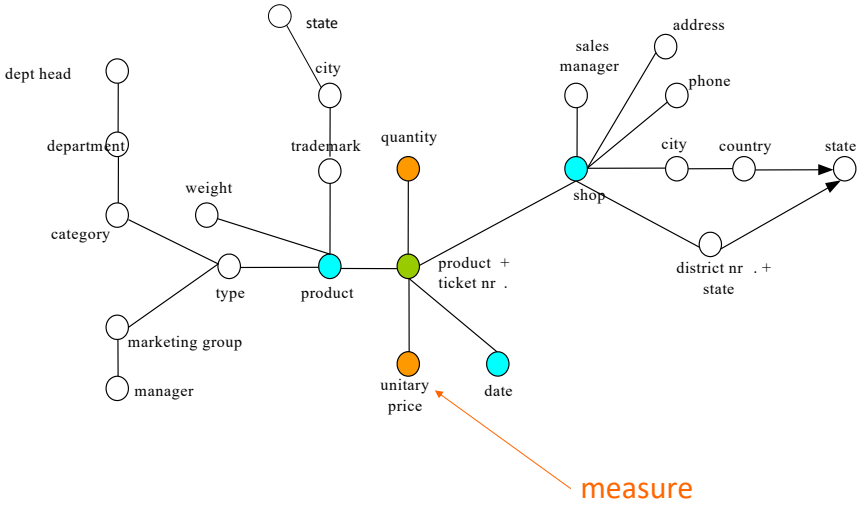
Dimensions definition: example



Measure definition

- If the fact identifier (set of attributes) is included in the set of dimensions, then numerical attributes that are children of the root (fact) are measures
- Further measures are defined by applying aggregate functions to numerical attributes of the tree
  - Generally: sum, average, min, max, count
- It is possible that a fact has no measures (empty)

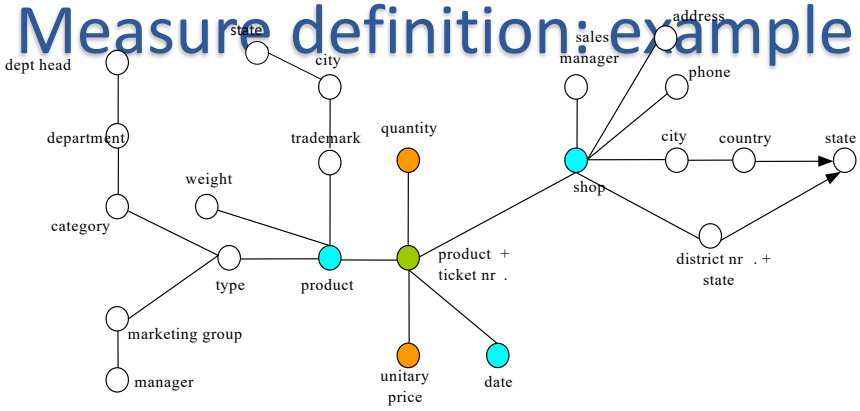
Measure definition: example



Glossary

- In the glossary, an expression is associated with each measure
  - The expression describes how we obtain the measure at the different levels of aggregation starting from the attributes of the source schema

Measure definition: example

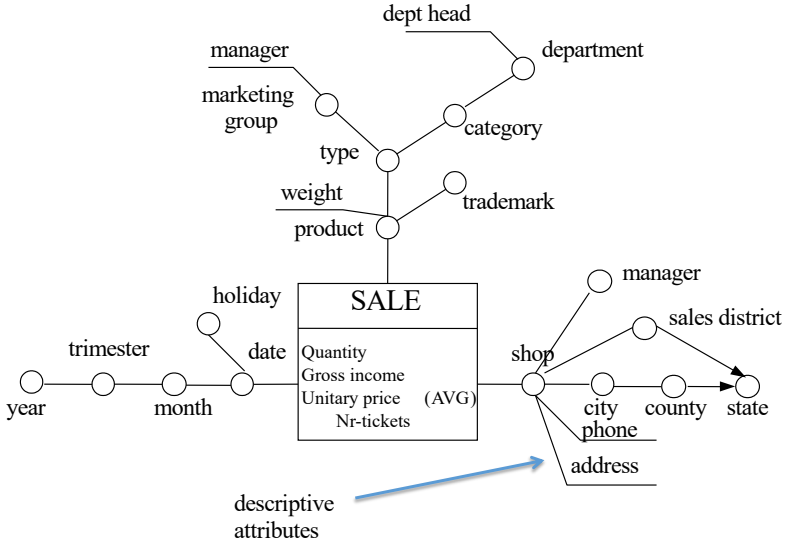


Quantity = SUM(Sale.quantity)
Gross income=SUM(Sale.quantity*Sale.unitaryprice)
Unitary price=AVG(Sale.unitaryprice)
Nr-tickets=COUNT(*)

Fact schema creation

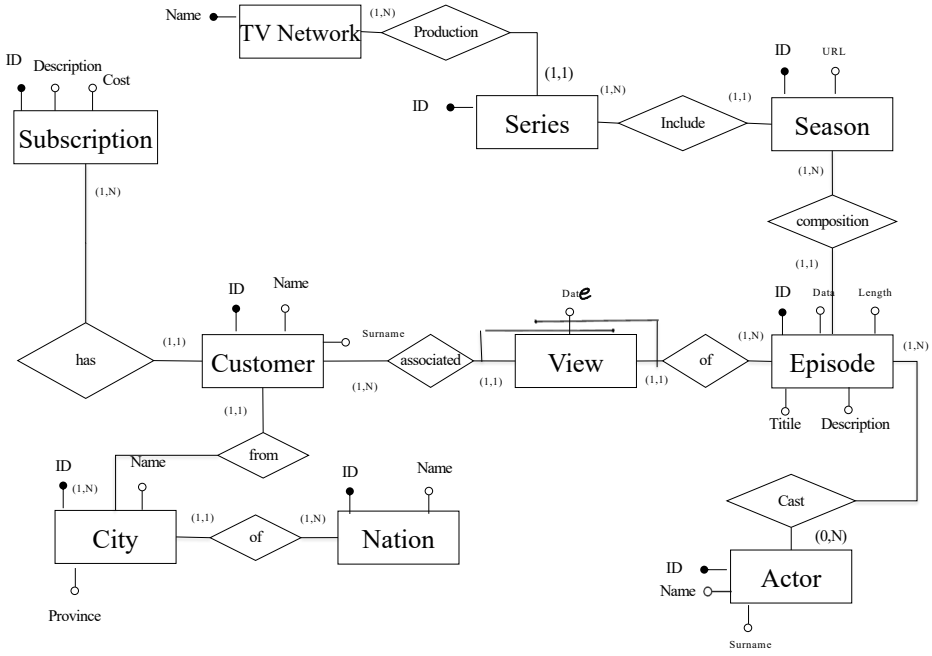
- The attribute tree is translated into a fact schema including dimensions and measures
  - Dimension hierarchies correspond to subtrees having as roots the different dimensions (with the least granularity)
  - The fact name corresponds to the name of the selected entity

Fact schema creation: example



Exercise

- The ER schema is a portion of a database related to a video content streaming service. Starting from this DB, we want to build a DW to make decisions regarding the catalog of contents for the following season and advertising to customers.
- In particular, we want to analyze:
  - Which are the TV series that have been preferred in the last year (highest number of views); it is requested also the possibility to have details about the individual seasons or single episodes;
  - Which are the most successful series (highest number of views) for a type of customer or a geographical area



SOLUTION

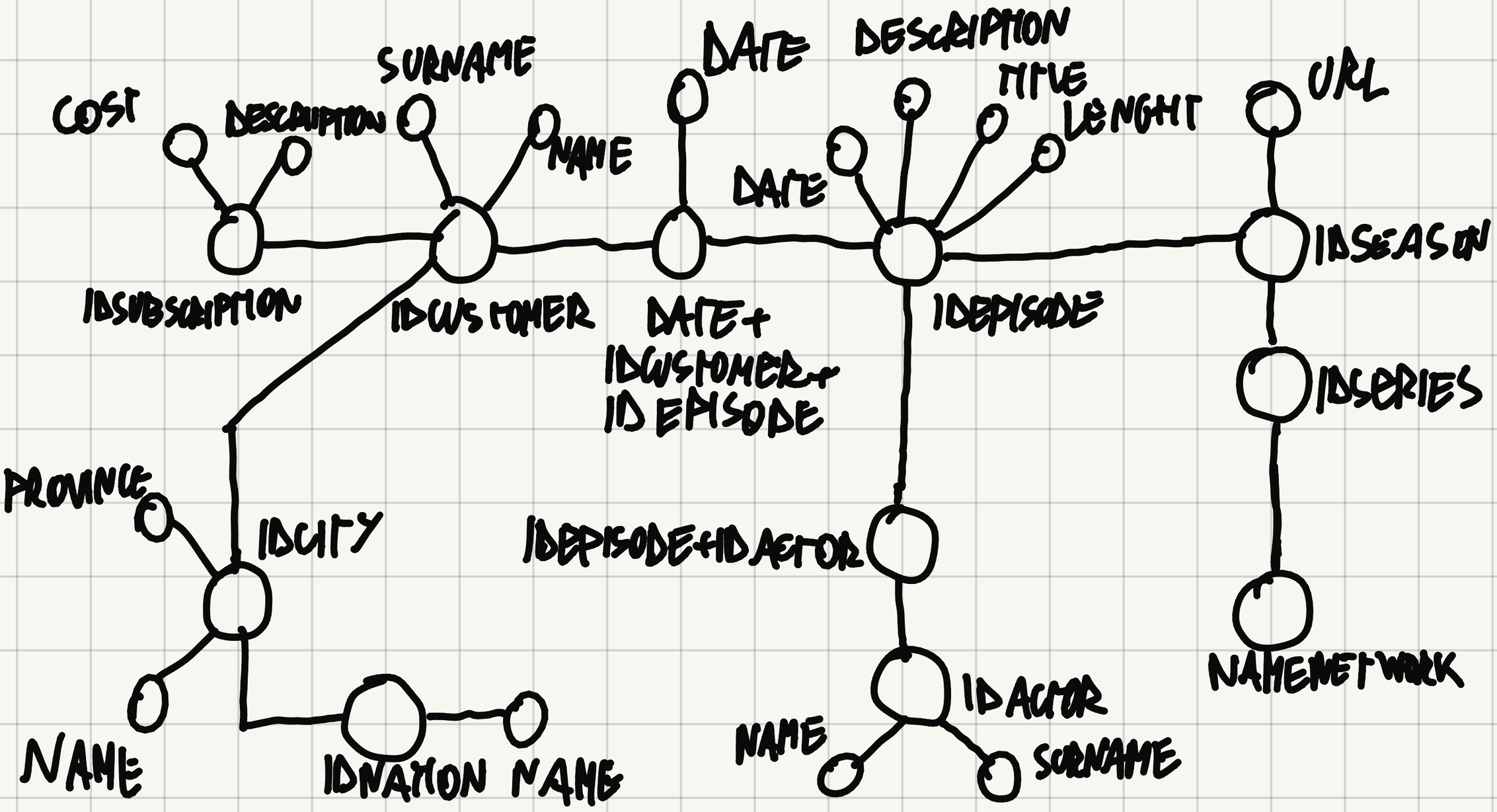
1) FIND THE RELEVANT FACTS

DYNAMIC POWER:

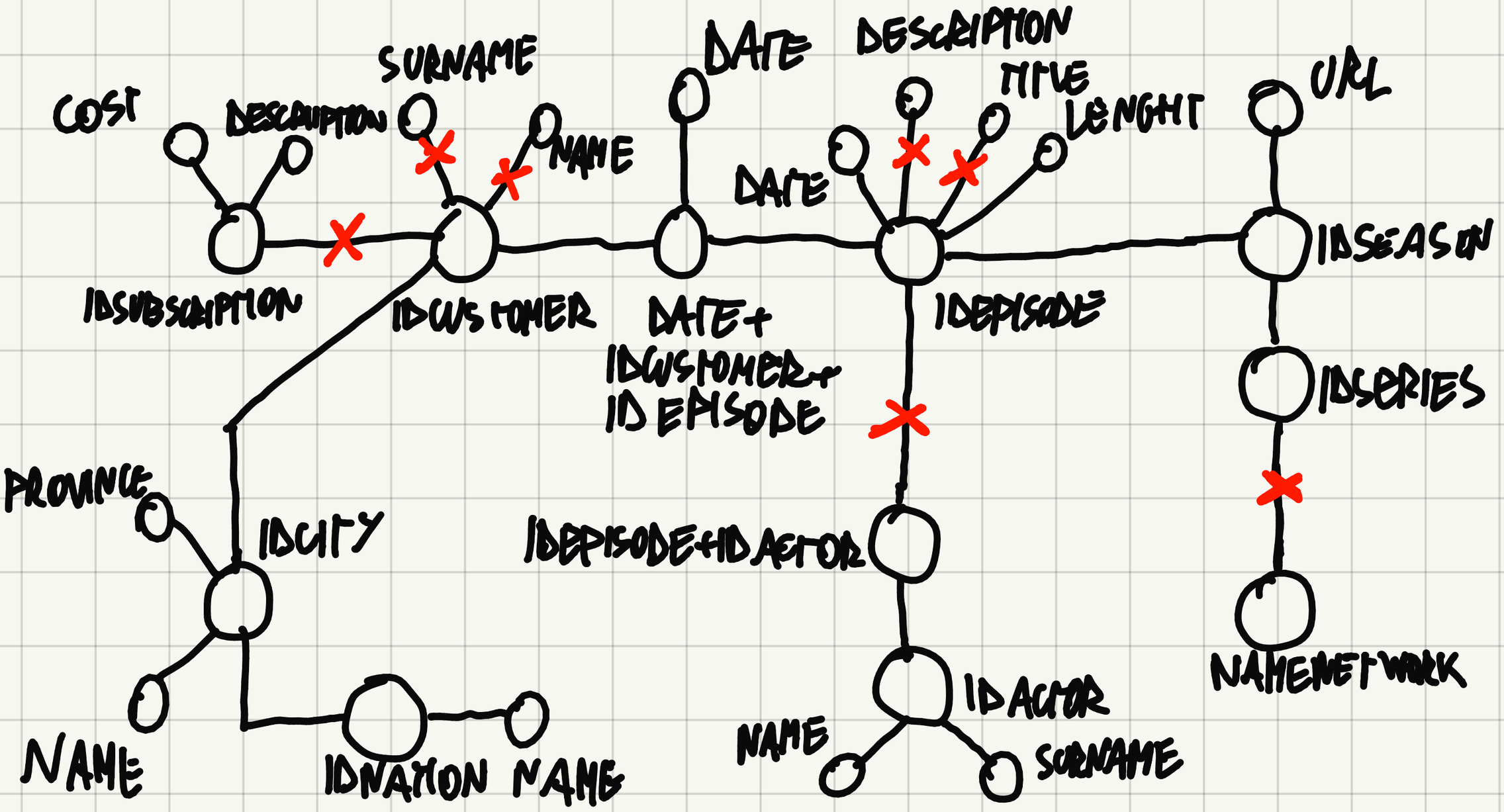
- VIEW ("DATE" AS ONLY ATTRIBUTE + EMPTY FACT BECAUSE WE CAN'T DO NOTHING BUT COUNTING)

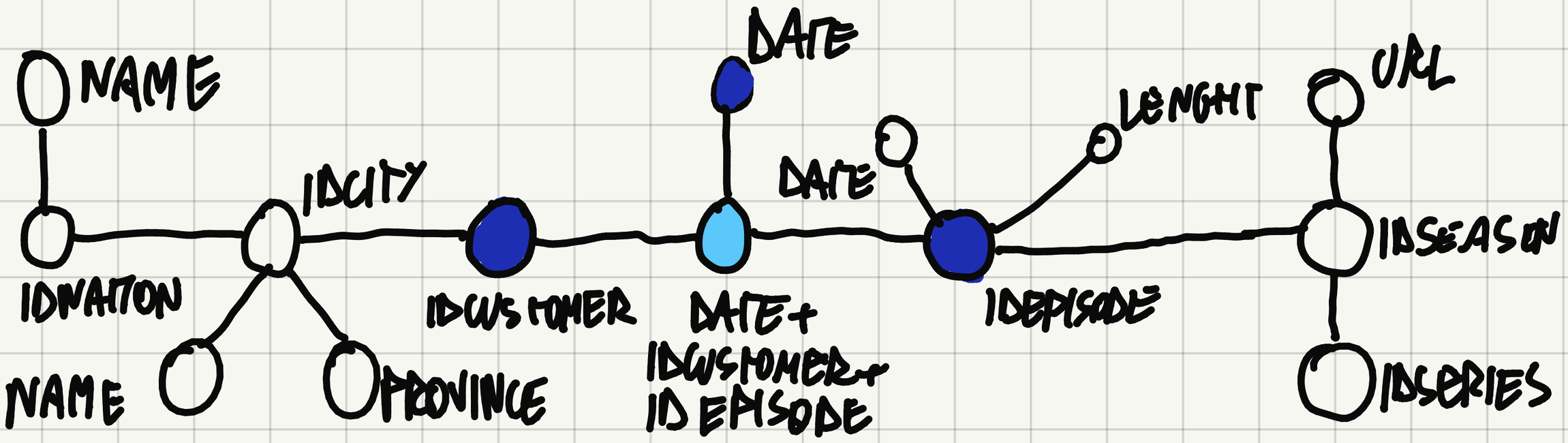
- KEY: IDCUSTOMER + DATE + IDEPISODE

2) CONSTRUCT



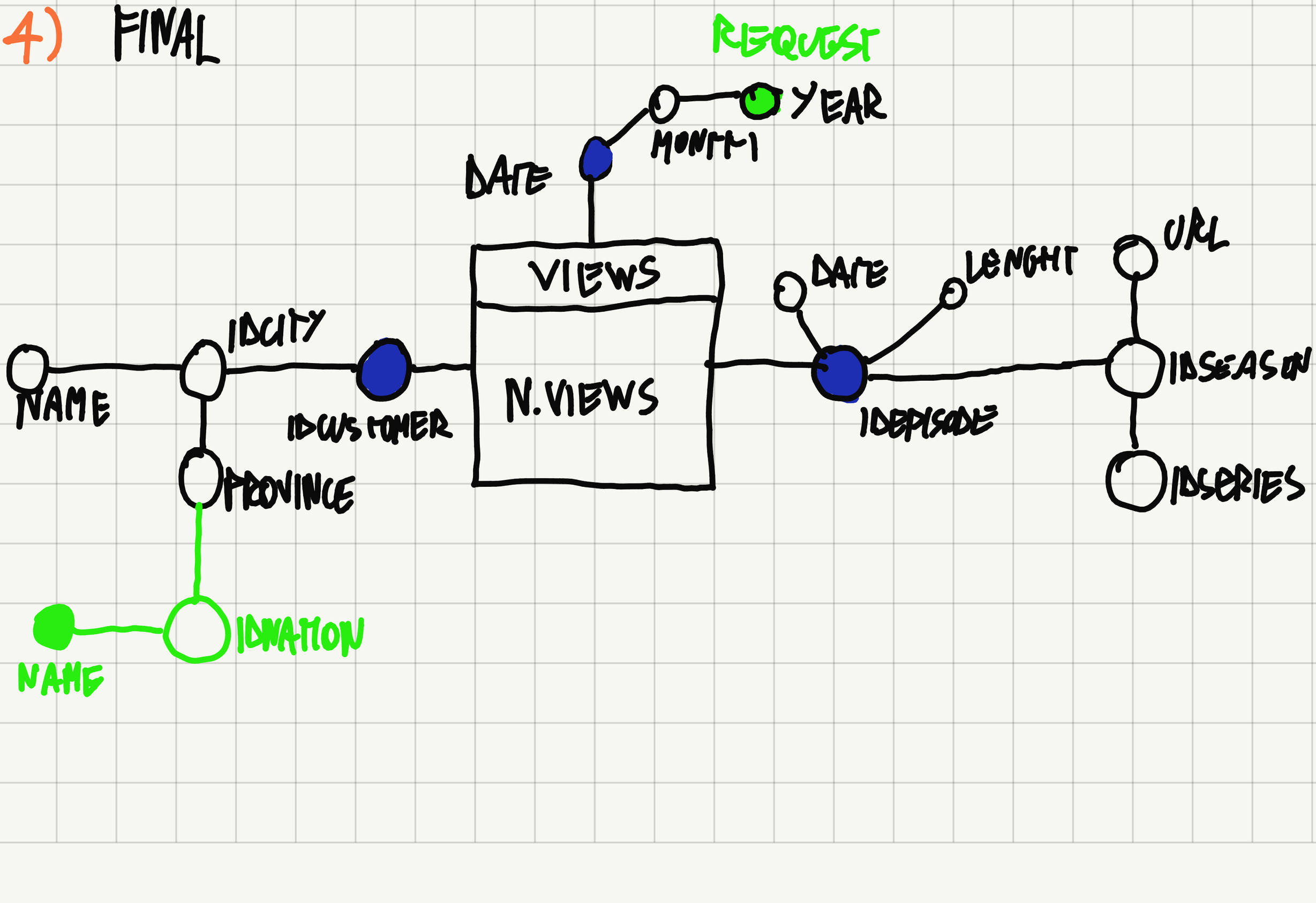
- WE WANT TO CONSIDER THE MINIMUM AMOUNT OF DATAS. IN GENERAL, WE WILL CONSIDER ONLY THE REQUESTED DATA





3) IDENTIFY FACTS AND DIMENSION

4) FINAL

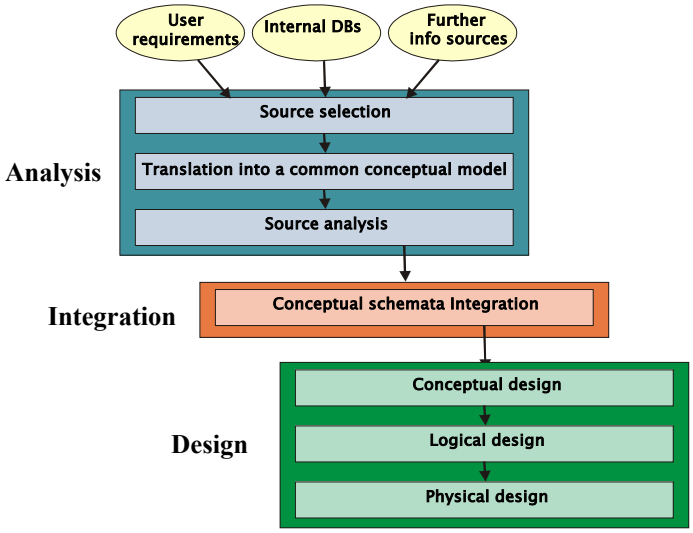


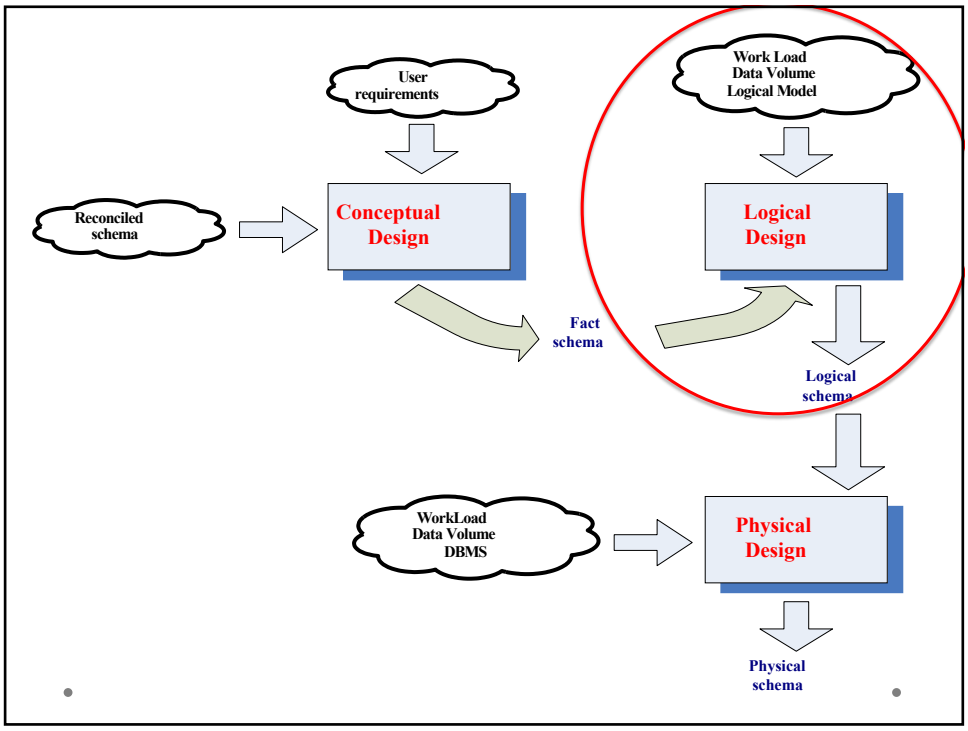
POLITECNICO
MILANO 1863

Data Warehouse design

Cinzia Cappiello
A.A. 2023-2024

Data Warehouse Design





DECIDE HOW TO REPRESENT THOSE INFORMATIONS IN THE SCHEMA

3

Logical Models

WE HAVE THREE DIFFERENT MODELS

- MOLAP
- ROLAP
- HOLAP

4

Data Mart logical models

MOLAP stands for Multidimensional OLAP. In MOLAP cubes the data aggregations and a copy of the fact data are stored (materialized) in a multidimensional structure on the computer. It is best when extra storage space is available on the server and the best query performance is desired. MOLAP local cubes contain all the necessary data for calculating aggregates and can be used offline. MOLAP cubes provide the fastest query response time and performance but require additional storage space for the extra copy of data from the fact table.

→ NOTE: REQUIRES ADDITIONAL INVESTMENTS!!!

* ROLAP stands for Relational OLAP. ROLAP uses the relational data model to represent multidimensional data. In ROLAP cubes a copy of data from the fact table is not necessarily made, and the data aggregates are stored in tables, separately or in the source relational database. A ROLAP cube is best when there is limited space on the server and query performance is not very important. ROLAP local cubes contain the dimensions and cube definitions but normally aggregates are computed when needed. ROLAP cubes require less storage space than MOLAP and HOLAP cubes.

HOLAP stands for Hybrid OLAP. A HOLAP cube has a combination of the ROLAP and MOLAP cube characteristics. It does not necessarily create a copy of the source data; however, data aggregations are stored in a multidimensional structure on the server. HOLAP cubes are best when storage space is limited but faster query responses are needed.

WE WANT TO STORE OLAP IN A RELATIONAL DB

MOST USED STRATEGY

ALL POSSIBLE ANALYSIS ARE ALREADY CALCULATED

ESPECIALLY FOR HIGH QUANTITIES OF DATAS TO STORE

5

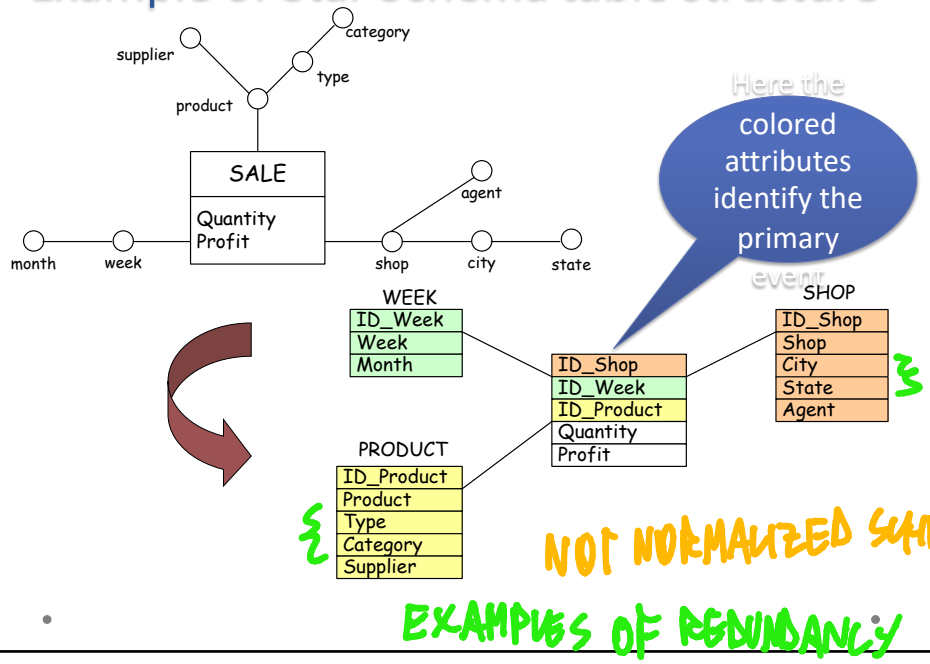
ROLAP

- It is based on the Star Schema
- A star schema is :
  - A set of relations DT1, DT2, …DTn - dimension tables - each corresponding to a dimension.
  - Each DTi is characterized by a primary key di and by a set of attributes describing the analysis dimensions with different aggregation levels
  - A relation FT, fact table, that imports the primary keys of dimensions tables. The primary key of FT is d1 d2 … dn ; FT contains also an attribute for each measure

6

DAY MONTH YEAR
15/11/23 11/23 23
04/01/21 01/21 21
⋮ ⋮ ⋮

Example of Star Schema table structure



NOT NORMALIZED SCHEMAS

EXAMPLES OF REDUNDANCY

7

! IN THIS CONTEXT, WE ACCEPT THE REDUNDANCY AS IT PREVENT US TO MAKE DIFFICULT COMPLEX OPERATIONS (LIKE JOINS) AND HAVING A DIRECT ACCESS TO THE REQUESTED INFORMATION. OBVIOUSLY, THAT MEANS MORE DATA TO STORE

It is possible to define different variants of the star schema to manage aggregate data, e.g. in a unique fact table

1° row represents sale values for the single shop, 2° row represents aggregate values for Roma, 3° row represents aggregate values for Lazio, etc...

SALE

Shop_key	Date_key	Prod_key	qty	profit	…
1	1	1	170	85	…
2	1	1	300	150	…
3	1	1	1700	850	…
…	…	…	…	…	…

SHOP

Shop_key	shop	city	region	…
1	COOP1	Bologna	E.R.	…
2	-	Roma	Lazio	…
3	-	-	Lazio	…
…	…	…	…	…

8

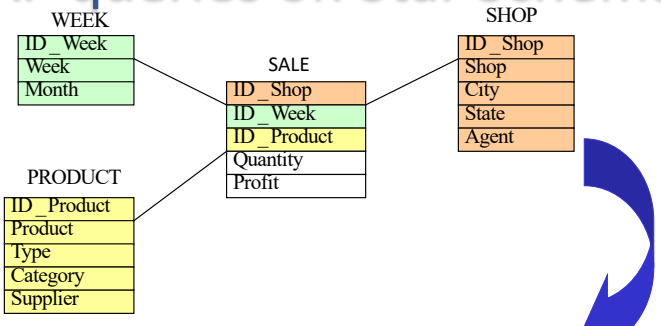
Star schema: considerations

- Dimension table keys are surrogates (i.e. generated ids), for space efficiency reasons
- Dimension tables are de-normalized, i.e. they contain redundancy: note that

  product → type → category

  means that for each different product all the info related to type is repeated, and the same for the category
- De-normalization introduces redundancy, but fewer joins to do
- The fact table contains information expressed at different aggregation levels

OLAP queries on Star Schema

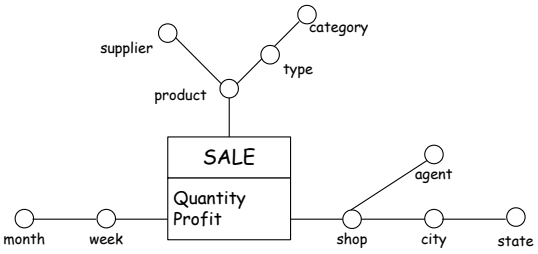



```
select City, Week, Type, sum(Quantity)
from   Week, Shop, Product, Sale
where  Week.ID_Week=Sale.ID_Week and
       Shop.ID_Shop=Sale.ID_Shop and
       Product.ID_Product=Sale.ID_Product and
       Product.Category = 'FoodStuff'
group by City,Week,Type
```

SOLUTION TO AVOID REDUNDANCY

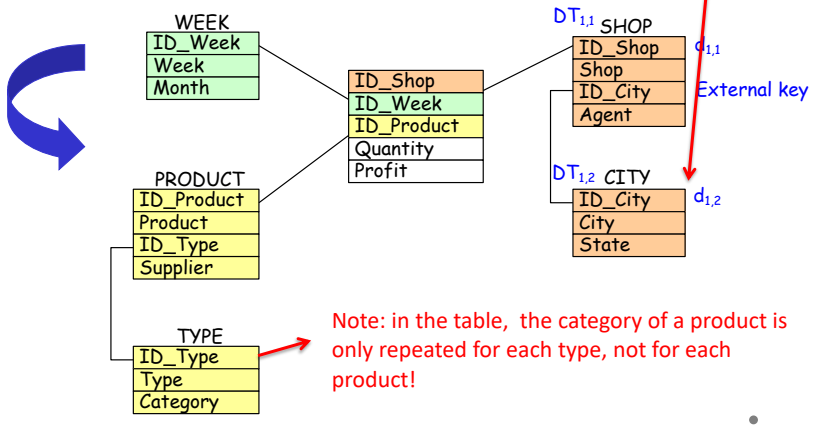
Snowflake schema

- The snowflake schema reduces the de-normalization of the dimensional tables DTi of a star schema
- Dimensions tables of a snowflake schema are composed by
  - A primary key di,j
  - A subset of DTi attributes that directly depend on di,j
  - Zero or more external keys that allow to obtain the entire information
- In a snowflake schema
  - Primary dimension tables: their keys are imported in the fact table
  - Secondary dimension tables



Snowflake schema

Note: in the table, the state of a shop is only repeated for each city, not for each shop!



Note: in the table, the category of a product is only repeated for each type, not for each product!

Snowflake schema: considerations

- Reduction of memory space
- New surrogate keys
- Advantages in the execution of queries related to attributes contained into fact and primary dimension tables

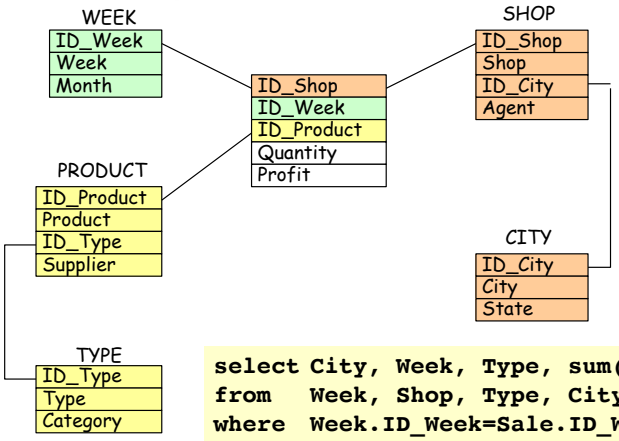
13

Normalization & Snowflake schema

- Attributes uniquely determined (transitively or not) by the snowflake attribute are placed in a new relation

14

OLAP queries on snowflake schema




```
select City, Week, Type, sum(Quantity)
from   Week, Shop, Type, City, Product, Sale
where  Week.ID_Week=Sale.ID_Week and
       Shop.ID_Shop=Sale.ID_Shop and
       Shop.ID_City = City.ID_City and
       Product.ID_Product=Sale.ID_Product and
       Product.ID_Type=Type.ID_Type and
       Product.Category = 'FoodStufs'
group by City,Week, Type
```


Views

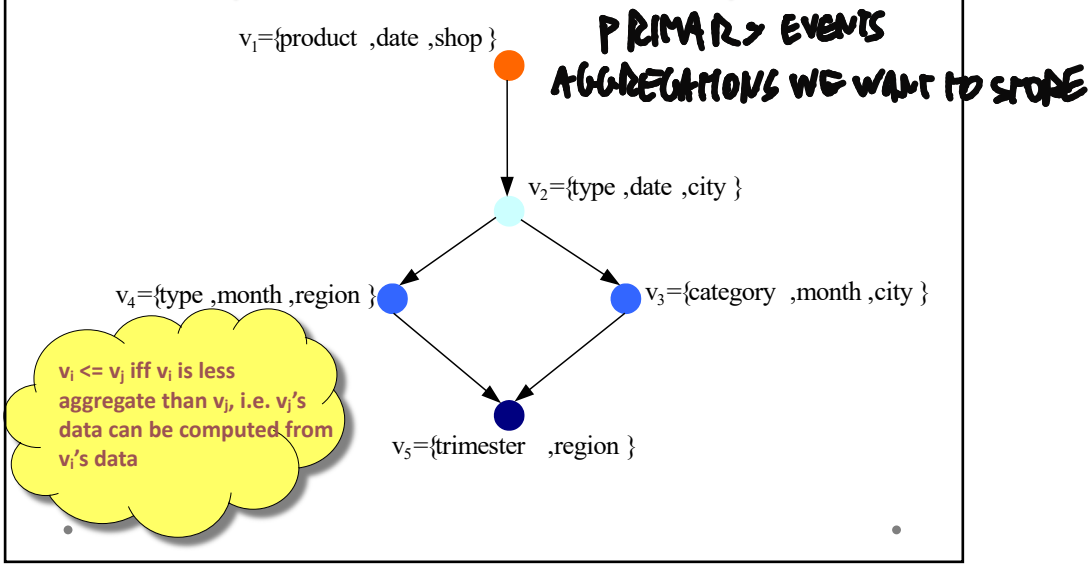
- Aggregation allows to consider concise (summarized) information
- Aggregation computation is very expensive → pre-computation (materialization)
- A view denotes a fact table containing aggregate data
- We can pre-compute views to make computation more efficient

Views

- A view can be characterized by its aggregation level (pattern)
  - Primary views: correspond to the primary aggregation levels
  - Secondary views: correspond to secondary aggregation levels (secondary events)

Views

(MultiDimensional Lattice)



Partial aggregations

- Sometimes it is useful to introduce new measures in order to manage aggregations correctly
  - Derived measures: obtained by applying mathematical operators to two or more values of the same tuple

Partial aggregations

Profit=Quantity*Price

Type	Product	Quantity	Price	Profit
T1	P1	5	1,00	5,00
T1	P2	7	1,50	10,50
T2	P3	9	0,80	7,20

22,70
(total profits)

SUM AVG

Type	Quantity	Price	Profit
T1	12	1,25	15,00
T2	9	0,80	7,20

22,20

We can't just sum up profits as before!!

The correct solution consists in the aggregation of data on the primary

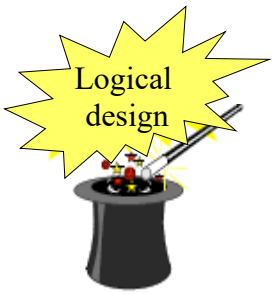
Logical design
Rolap

29

Logical modelling

- Sequence of steps that, starting from the conceptual schema, allow one to obtain the logical schema for a specific data mart

INPUT
Conceptual Schema
WorkLoad
Data Volume
System constraints



OUTPUT
Logical Schema

30

Workload

- In OLAP systems, workload is dynamic in nature and intrinsically extemporaneous
  - Users' interests change over time
  - Number of queries grows when users gain confidence in the system
  - OLAP should be able to answer any (unexpected) request
- During requirement collection phase, deduce it from:
  - Interviews with users
  - Standard reports

31

Workload

- Characterize OLAP operations:
  - Based on the required aggregation pattern
  - Based on the required measures
  - Based on the selection clauses
- At system run-time, workload can be desumed from the system log

32

Data volume

- Depends on:
  - Number of distinct values for each attribute
  - Attribute size
  - Number of events (primary and secondary) for each fact
- Determines:
  - Table dimension
  - Index dimension
  - Access time

33

Logical modelling: steps

- Choice of the logical schema (star/snowflake schema)
- Conceptual schema translation
- Choice of the materialized views
- Optimization

34

From fact schema to star schema

- Create a fact table containing measures and descriptive attributes directly connected to the fact
- For each hierarchy, create a dimension table containing all the attributes

35

Guidelines

- Descriptive attributes (e.g. color)
  - If it is connected to a dimensional attribute, it has to be included in the dimension table containing the attribute (see slide n. 14, snowflake example, agent)
  - If it is connected to a fact, it has to be directly included in the fact schema
- Optional attributes (e.g. diet)
  - Introduction of null values or ad-hoc values

36

Guidelines

- Cross-dimensional attributes (e.g. VAT)
  - A cross-dimensional attribute b defines an N:M association between two or more dimensional attributes $a_1, a_2, \ldots, a_k$
  - It requires to create a new table including b and having as key the attributes $a_1, a_2, \ldots, a_k$

37

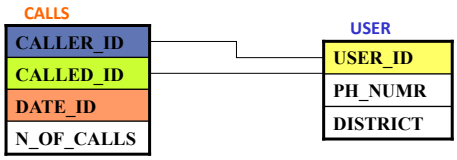
Guidelines

- Shared hierarchies and convergence
  - A shared hierarchy is a hierarchy which refers to different elements of the fact table (e.g. caller number, called number)
  - The dimension table should not be duplicated
  - Two different situations:
    - The two hierarchies contain the same attributes, but with different meanings (e.g. phone call → caller number, phone call → called number)
    - The two hierarchies contain the same attributes only for part of the hierarchy trees

38

Shared hierarchies and convergence

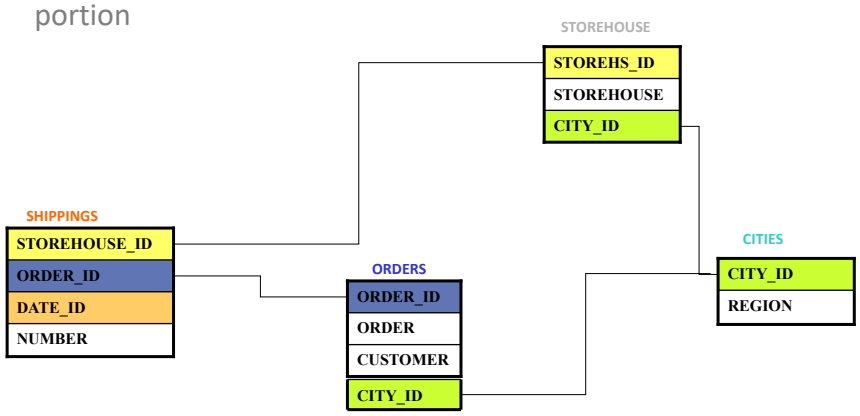
- The two hierarchies contain the same attributes, but with different meanings (e.g. phone call → caller number, phone call → called number)



39

Shared hierarchies and convergence

- The two hierarchies contain the same attributes only for part of the trees. Here we could also decide to replicate the shared portion

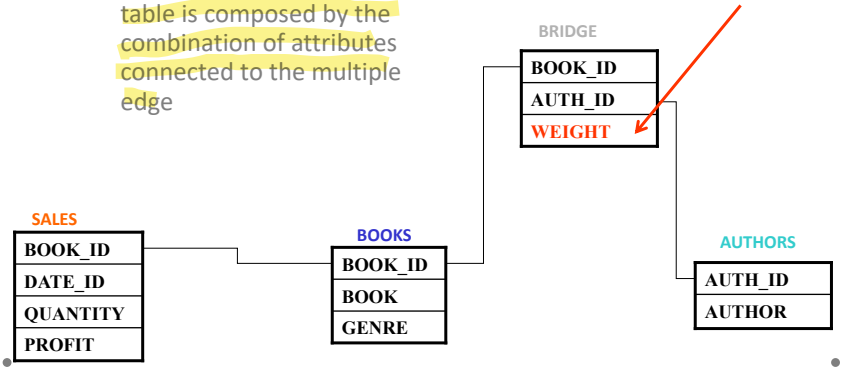


40

Guidelines

- Multiple edges
  - A bridge table models the multiple edge
    - the key of the bridge table is composed by the combination of attributes connected to the multiple edge

The weight of the edge is the contribution of each edge to the cumulative relationship



Guidelines

- Multiple edges: bridge table
  - Weighed queries take into account the weight of the edge

Query computing the profit for each author

```
SELECT AUTHORS.Author,SUM(SALES.Profit * BRIDGE.Weight)
FROM AUTHORS, BRIDGE, BOOKS, SALES
WHERE AUTHORS.Author_id=BRIDGE.Author_id
AND BRIDGE.Book_id=BOOKS.Book_id
AND BOOKS.Book_id=SALES.Book_id
GROUP BY AUTHORS.Author
```

Guidelines

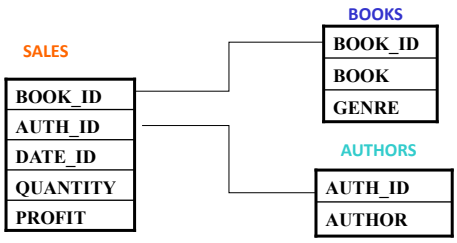
- Multiple edges: bridge table
  - Impact queries do not take into account the weight of the edge

Query computing the copies sold for each author

```
SELECT AUTHORS.Author, SUM(SALES.Quantity)
FROM AUTHORS, BRIDGE, BOOKS, SALES
WHERE AUTHORS.Author_id=BRIDGE.Author_id
AND BRIDGE.Book_id=BOOKS.Book_id
AND BOOKS.Book_id=SALES.Book_id
GROUP BY AUTHORS.Author
```


Alternative solution: keep the star model (only one level after the fact)

Multiple edges with a star schema:
add authors to the fact schema



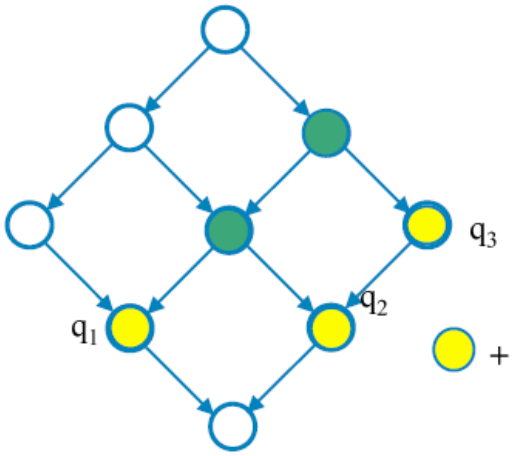
Here we don't need the weight because the fact table records quantity and profit per book and per author

Secondary-view precomputation

- The choice about views that have to be materialized takes into account contrasting requirements:
  - Cost functions' minimization
    - Workload cost
    - View maintenance cost
  - System constraints
    - Disk space
    - Time for data update
  - Users constraints
    - Max answer time
    - Data freshness

45

Materialized views (MD lattice)



= **exact views:**
They solve exactly the queries

= **less aggregate views:**
They solve more than one query

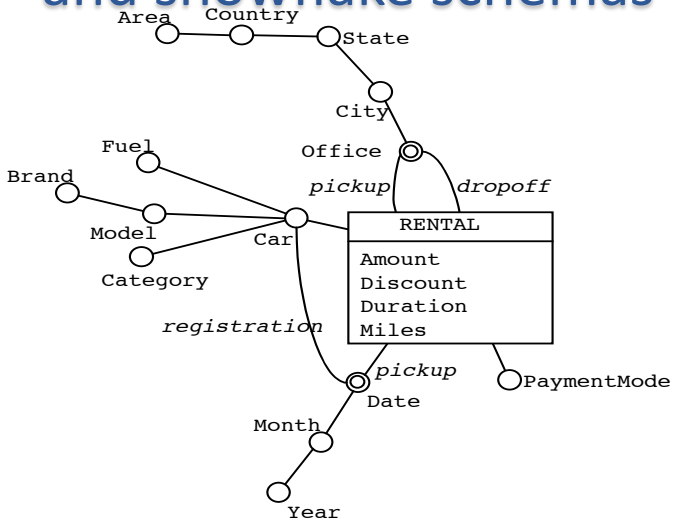
+ = **candidate views**
They could reduce elaboration costs

46

Materialized Views

- It is useful to materialize a view when:
  - It directly solves a frequent query
  - It reduce the costs of some queries
- It is not useful to materialize a view when:
  - Its aggregation pattern is the same as another materialized view
  - Its materialization does not reduce the cost

Exercise: from the DFM to Star and snowflake schemas

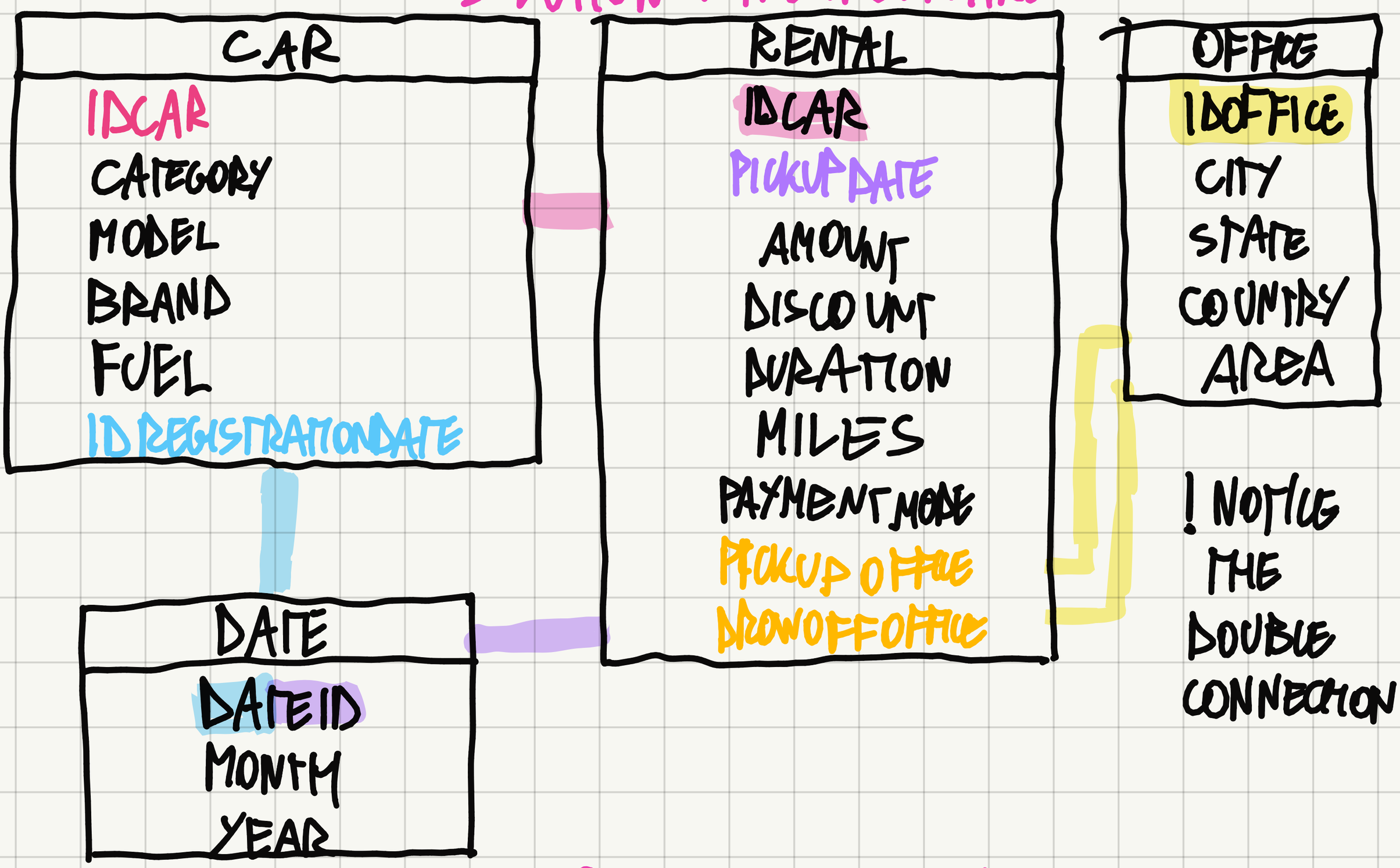


Exercise: from the DFM to Star schema

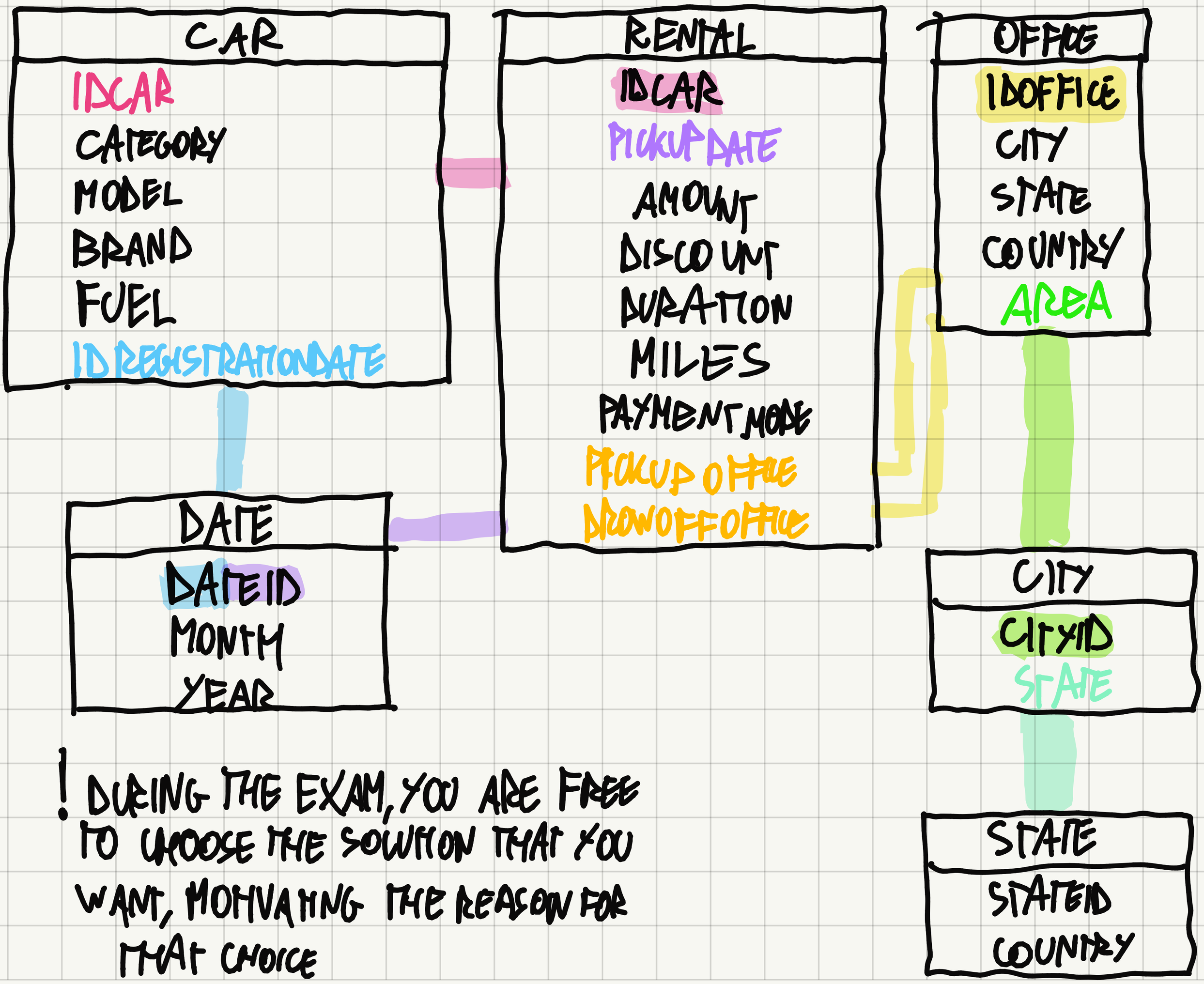
References

- Stefano Rizzi : Data Warehouse Design: Modern Principles and Methodologies McGraw-Hill, 2009
- M. Golfarelli, S. Rizzi: Data Warehouse: teoria e pratica della progettazione McGraw-Hill, 2002.
- Matteo Golfarelli: Data Warehouse Life-Cycle and Design. Encyclopedia of Database Systems 2009: 658-664
- Stefano Rizzi: Business Intelligence. Encyclopedia of Database Systems 2009: 287-288
- Ralph Kimball: The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses John Wiley 1996.
- On the Internet: Oracle® Database Data Warehousing

SOLUTION USING SNOWFLAKE



SOLUTION USING STAR



! DURING THE EXAM, YOU ARE FREE TO CHOOSE THE SOLUTION THAT YOU WANT, MOTIVATING THE REASON FOR THAT CHOICE

POLITECNICO
MILANO 1863

Data Warehouse – queries

Cinzia Cappiello
A.A. 2023-2024

The data cube in SQL

- It expresses all the possible tuple aggregations of a table
- It uses the new polymorphic value ALL

Data cube in SQL (Relational OLAP)

Considerinonly the red cars of Fiat and Ford models in years 1994 and 1995, find the total sales for each model, year and color of the car. Include in the answer also the aggregations computed using only one or two of the three attributes.

```
select Model, Year,
       Color, sum(Sales)
from Sales
where Model in {'Fiat','Ford'}
  and Color = 'Red'
  and Year between 1994 and 1995
group by Cube(Model, Year, Color)
```

CALCULATE ALL THE POSSIBLE AGGREGATION

3

Relevant Facts

model	year	color	sales
fiat	1994	red	50
fiat	1995	red	85
ford	1994	red	80

4

RESULT OF THE CUBE:

ALL THE POSSIBLE AGGREGATION A USER CAN ASK

All the data in the cube

model	year	color	sum (sales)
fiat	1994	red	50
fiat	1995	red	85
fiat	1994	ALL	50
fiat	1995	ALL	85
fiat	ALL	red	135
fiat	ALL	ALL	135
ford	1994	red	80
ford	1994	ALL	80
ford	ALL	red	80
ford	ALL	ALL	80
ALL	1994	red	130
ALL	1995	red	85
ALL	ALL	red	215
ALL	1994	ALL	130
ALL	1995	ALL	85
ALL	ALL	ALL	215

5

Roll up

ROLLUP enables a SELECT statement to calculate multiple levels of subtotals across an ordered set of dimensions.

Using the ROLLUP operator instead of the CUBE operator eliminates the results that contain ALL only in one column (except for the last), thus the aggregations only by model or only by year are not computed.

```
select Model, Year,
       Color, sum(Sales)
from Sales
where Model in {'Fiat','Ford'}
  and Color = 'Red'
  and Year between 1994 and 1995
group by Rollup (Model, Year, Color)
```

6

The data after roll-up

model	year	color	sum(sales)
fiat	1994	red	50
fiat	1995	red	85
ford	1994	red	80
fiat	1994	ALL	50
fiat	1995	ALL	85
ford	1994	ALL	80
fiat	ALL	ALL	135
ford	ALL	ALL	80
ALL	ALL	ALL	215

→ It is a kind of "progressive aggregation"

7

WAREHOUSE	PRODUCT	SUM(QUANTITY)
San Fransisco	Samsung	300
San Fransisco	iPhone	260
San Jose	Samsung	350
San Jose	iPhone	300

```
SELECT warehouse, product,
SUM(quantity) FROM inventory
GROUP BY CUBE(warehouse,product)
ORDER BY warehouse, product;
```

WAREHOUSE	PRODUCT	SUM(QUANTITY)
San Fransisco	Samsung	300
San Fransisco	iPhone	260
San Fransisco	(null)	560
San Jose	Samsung	350
San Jose	iPhone	300
San Jose	(null)	650
(null)	Samsung	650
(null)	iPhone	560
(null)	(null)	1210

```
SELECT warehouse, product,
SUM(quantity) FROM inventory
GROUP BY ROLLUP (warehouse ,
product);
```

warehouse	product	SUM(quantity)
San Fransisco	iPhone	260
San Fransisco	Samsung	300
San Fransisco	NULL	560
San Jose	iPhone	300
San Jose	Samsung	350
San Jose	NULL	650
NULL	NULL	1210

https://www.sqltutorial.org/sql-rollup/

8

Rollup

- Difference between Cube and rollup:

CUBE evaluates aggregate expression with all possible combinations of columns specified in group by clause, whereas the Rollup evaluates aggregate expressions only relative to the order of columns specified in group by clause.