Exercises *Bases de Dades Avançades* - Data Warehousing

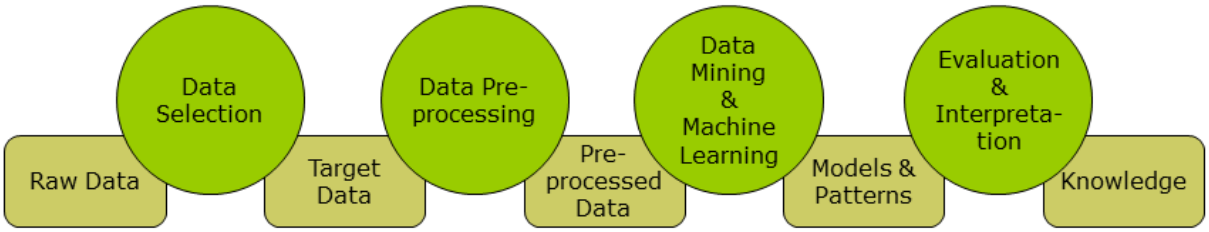
Database Technologies and Information Management (DTIM) group
Universitat Politècnica de Catalunya (BarcelonaTech), Barcelona
September 10, 2021

Contents

0.1 Data Warehousing Architectures 5
0.2 OLAP and the Multidimensional Model 7

0.1 Data Warehousing Architectures

1. Give examples of metadata that can be used at every step of the Knowledge Discovery in Databases.



Data Selection

Data Pre-processing

Data Mining and Machine Learning

Evaluation and Interpretation

2. Enumerate the advantages of having separate files (e.g., CSV) to be analyzed and compare it against the advantages of having a DBMS.

Advantages of using a DBMS	**Advantages of using separate files**

3. In the current post-industrial information society, data is one of the most valuable resources. However, data is not useful unless it can be processed and turned into information that allows enterprises a competitive advantage. Moreover, our ever-more efficient information systems are collecting ever increasing amounts of data. Further, while data may be out there, it may be in many disparate forms and formats, essentially making it unusable.

   Pfizer, a large multi-national pharmaceutical giant, has a plethora of clinical trials for a number of drug projects. Moreover, data collection and analyses operations are spread across the world, making it harder to enforce data standards. Even harder to enforce was the programming and validation standards that are required of pharmaceutical companies. Clinical trials are run in scores of countries and the collected data is needed by clinicians and statisticians from every site for analysis and product defense.

   In a typical clinical trial, especially large and/or multicenter ones, there are many sources of data, including electronic data transfers from sites, central labs and Contract Research Organizations (CROs). Moreover, with multiple trials going on across many projects for the same compound/drug, the issue becomes how to manage all these data and at the same time, not repeat data collection. Thus, to summarize briefly, the data is received in house, is cleaned up, required transformation routines are applied to massage and reformat the data, stored in an appropriate data format, and reports/analyses are run. So, what are we to do with these mountains and islands of data to help us test hypotheses, derive conclusions, and identify trends and opportunities?[1]

4. In the arena of Health Insurance, millions of claims are generated everyday, including hospitals, physicians, and pharmacies. A number of clearing houses process these claims and route them to the appropriate payer. Thus, these clearing houses sit on a huge amount of data that runs in 10s of gigabytes everyday but other than routing them, no other insights are gleaned from these data.

   The company gets data from the claims clearing houses and Pharmacy Benefit Managers (PBM) on a weekly basis. It processes the data and completes analysis requirements specified by the drug companies and the health insurance companies.

   Thus, such a data warehouse can provide invaluable intelligence in terms of:

[1] https://www.lexjansen.com/nesug/nesug99/iw/iw043.pdf

- Drug Positioning Information
- Patient Population Characteristics
- Indications drug is being used for
- Prescribing Physician Characteristics
- Regional Preferences
- Prevalence of Diseases
- Preferred Drugs for Diseases
- Procedures being performed
- Disease related Information

Enumerate some functionalities/benefits that you can get in such environment.[2]

5. Explain in which sense each of the four characteristics of a DW in W. Inmon's definition increases the amount of data in the company.

   1)
   2)
   3)
   4)

[2] https://www.lexjansen.com/nesug/nesug99/iw/iw043.pdf

0.2 OLAP and the Multidimensional Model

1. Identify factual and dimensional information in the following CSV sample file.

ATTESTATION OF TRAFFIC ACCIDENTS

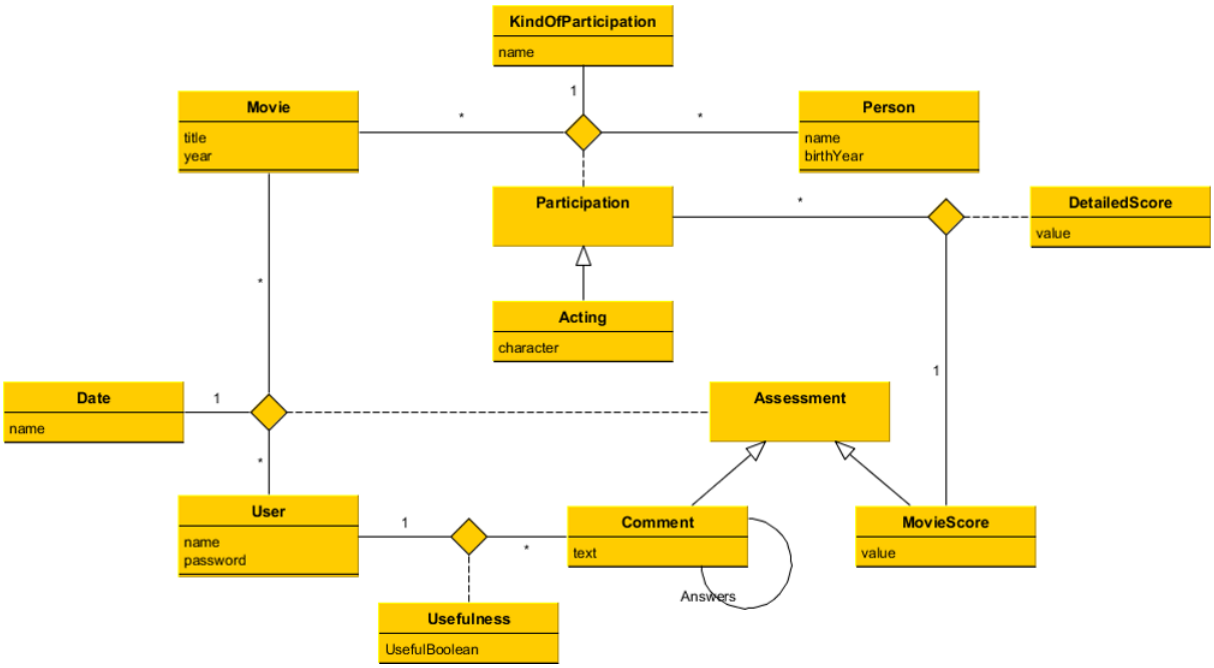
Month	Month name	Year	Police Region (RP)	Transit Regional Area (ART)	Number
1	Gener	2011	RP PIRINEU OCCIDENTAL	ART PIRINEU OCCIDENTAL	3
1	Gener	2011	RP GIRONA	ART GIRONA	10
1	Gener	2011	RP PONENT	ART PONENT	2
1	Gener	2011	RP CENTRAL	ART CENTRAL	7
1	Gener	2011	RP METROPOLITANA NORD	ART METROPOLITANA NORD	20
1	Gener	2011	RP METROPOLITANA SUD	ART METROPOLITANA SUD	10
1	Gener	2011	RP TERRES DE L'EBRE	ART TERRES DE L'EBRE	4
1	Gener	2011	RP CAMP DE TARRAGONA	ART TARRAGONA	5
2	Febrer	2011	RP PIRINEU OCCIDENTAL	ART PIRINEU OCCIDENTAL	3
2	Febrer	2011	RP GIRONA	ART GIRONA	6
2	Febrer	2011	RP PONENT	ART PONENT	4
2	Febrer	2011	RP CENTRAL	ART CENTRAL	1

2. Identify factual and dimensional information in the following CSV sample file.

CITIZEN SECURITY

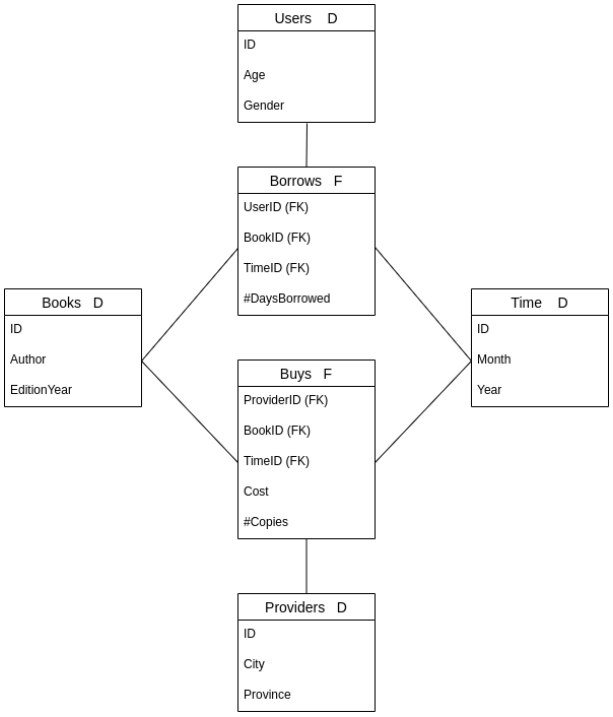
DISTRICTS	RELATED TO PEOLPLE	RELATED TO PROPERTY	WEAPON POSESSION	DRUG POSESSION	DRUG CONSUMPTION
CENTRO	48	127	10	190	82
ARGANZUELA	61	45	3	10	6
RETIRO	0	19	4	12	0
SALAMANCA	18	58	0	7	0
CHAMARTÍN	12	29	0	16	9

3. Identify some multidimensional schema (i.e., fact subject of analysis and its corresponding analysis dimensions) in the next UML class diagram.



4. Simplify the multiple star schemas corresponding to different CSV files in a single constellation schema. You can use a “bus Matrix” (i.e., Fact×Dimension) as an intermediate step.

5. Write multidimensional queries using the following constellation schema and the steps required in each exercise:



1)
- A:=Roll-up(borrows, Users.Gender, Sum)
- B:=Roll-up(A, Time.Month, Sum)
- R:=Selection(B, Time.Month=’January’)

2)
- A:=Roll-up(buys,Providers.AllProviders,Sum)
- B:=Roll-up(A, Books.Allbooks,Sum)
- C:=Roll-up(B, Time.AllDays,Sum)
- D:=ChangeBase(C, Providers)
- R:=Projection(D, #Copies)

3)
- A:=Roll-up(borrows,Users.AllUsers,Sum)
- B:=Roll-up(A, Books.EditionYear,Sum)
- C:=ChangeBase(B, Books x Time)
- D:=Drill-across(C, buys)
- R:=Projection(D, Cost)

4)
- A:=Roll-up(buys, Providers.AllProviders, Sum)
- B:=ChangeBase(A, Books, Time)
- C:=Drill-across(B, borrows, Sum)
- D:=Projeccio(C, #Daysborrowed)
- R:=ChangeBase(D, Books, Time, Users)

5)
- A:=Roll-up(borrows, Time.Month, Sum)
- B:=Selection(A, Time.Month=’January’)
- R:=Drill-down(B, Time.Id)

6. One of the three necessary summarizability conditions is disjointness. How can you implement a dimension with a level whose elements are overlapping avoiding summarizability problems? Give a concrete example using a UML class diagram with only two levels (you do not need to draw the whole snowflake schema).

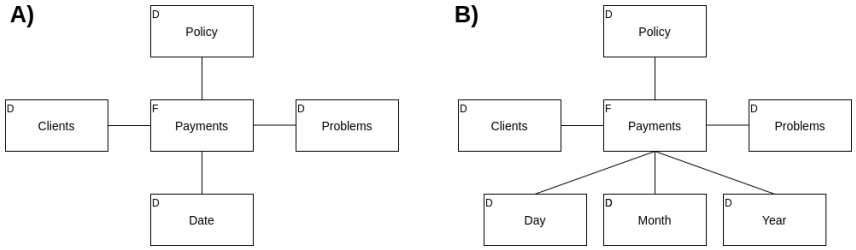
7. Name the three possible meanings of the NULL value in standard SQL.

   1)
   2)
   3)

8. Identify the problem in this sequence of algebraic multidimensional operations, and briefly explain how you would solve it.

   - A:=Roll-up(Sales, Providers.Province, Sum)
   - B:=ChangeBase(A, Books x Time)
   - C:=Drill-across(B, Loans, Sum)
   - D:=Projection(C, days)
   - R:=ChangeBase(D, Books x Time x Users)

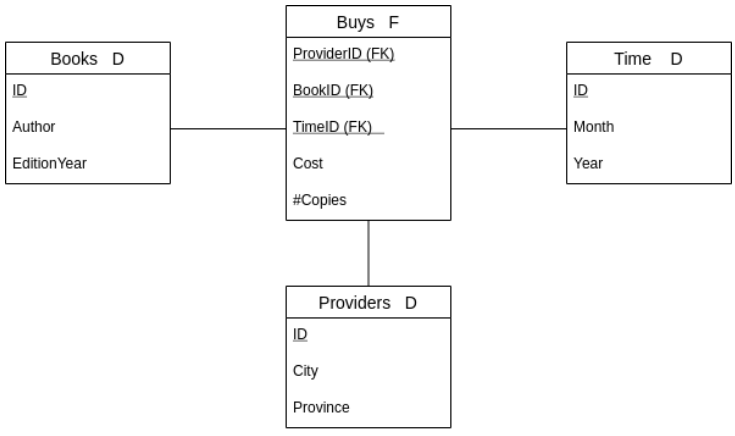
9. Which schema would you choose and **why**:



10. Briefly explain the difference between the different kinds of multidimensional schemas:

    a) Star vs. Snowflake
    b) Star vs. Galaxy

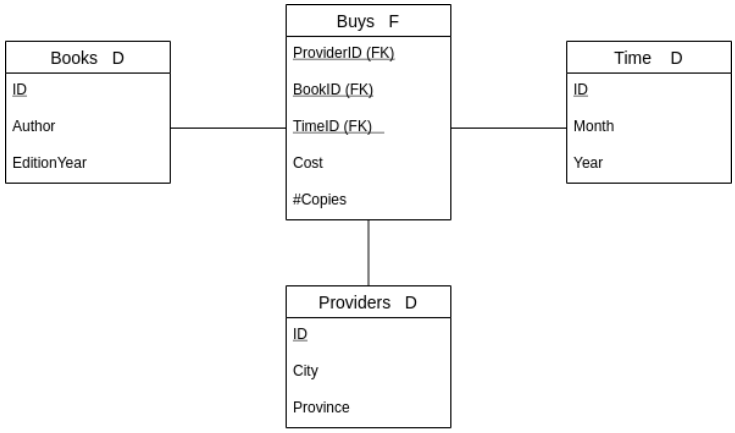
11. Given the multidimensional schema and the sequence of multidimensional operations below, give the equivalent SQL query (simplify it as much as possible).



- A:=Roll-up(Buys,Providers.All,Sum)
- B:=Roll-up(A, Books.AllBooks,Sum)
- C:=Roll-up(B, Time.AllDays,Sum)
- D:=ChangeBase(C, Providers)
- E:=Projection(D, #copies)
- R:=Drill-down(E, Providers.City)

12. Give the relational representation (i.e., tables and their corresponding integrity constraints) of the geographical **dimension** for a data mart in United Nations, so that it allows to **keep track of changes** in territories. Assume you only need to keep track of countries and their first administrative level (e.g., Autonomous communities in Spain). We would like to consider the split of countries (like in the case Catalonia would become independent), as well as annexations (like the case of Crimea peninsula by Russian Federation, assuming that peninsula was already a pre-existing component of Ukraine at the first administrative level). Briefly justify your answer and explicit any assumption you make.

13. Given the multidimensional schema and the sequence of multidimensional operations below, give the equivalent SQL query (simplify it as much as possible).



- A:=Projection(Buys, Cost)
- B:=Roll-up(A,Providers.Province,Sum)
- C:=Selection(B,Province=Barcelona)
- D:=Drill-down(C,Providers.City,Sum)
- E:=Selection(D,TimeID=20210115)
- R:=ChangeBase(E, Providers,Books)