Data Warehouse Modelling Workgroup

Business Intelligence & Data Warehousing

PROFESSOR JOSÉ CURTO DÍAZ

Group D

Adilet Gaparov
Amanda Marques
Benjamín Chumaceiro
David Facusse
Salim Aouzai
Tingting Sun

Data Set Analysis

The adventure of big data and analytics have industries going through process to try and optimize its decisions. The healthcare industry has also been hugely affected by those transformation as the aim is both to improve the patient well-being and to find efficiencies in the whole treatment and operational process.

Cognitive analysis refers to the process of generating insights from past unstructured data using technologies that find patterns and enable analysts to draw hypothesis. In the current paper, the analysis is of data associated with healthcare and patients with 3 different cognitive diseases, such as dementia, amnesia and delirium.

The structured data provided were recorded from 20 different patients on a daily basis, throughout the period of one year (2016). The data fields were (5 excel sheets):

- **Patients:** 20 people with its own number identification that were followed daily. Each of them was diagnosed with one disorder, being either DELIRIUM, DEMENTIA or AMNESIA. They were each also linked to a Spanish city where they were being treated and each city was related to an environment (RURAL, SEMIURBAN or URBAN).
- **Hours of sleep**: refers to number of hours each patient slept per day. Values are integer numbers, ranging from 4 to 6 hours included.
- **Activities:** related to what a patient did during the day. It reports only one activity per day per patient. The different activity values are EXERCISE, FAMILY, READ/STUDY, RADIO/TV, SLEEP/SOFA or NO ACTIVITY.
- **Episode value:** the table reports if the disease of the patient manifested during the day and its episode level. The different levels are NO EPISODE, LIGHT, MODERATE, SEVERE.
- **Mood values:** refers to reported mood the patient had this day. The measures are SAD, NORMAL and HAPPY.

Based on the data we have, we assumed the potential users might be medical researchers and doctors, who can follow how the behavior and wellness of a patient changes over time during year, and build hypotheses on the relationship between behavior and wellness state of patient and how environment might influence both.

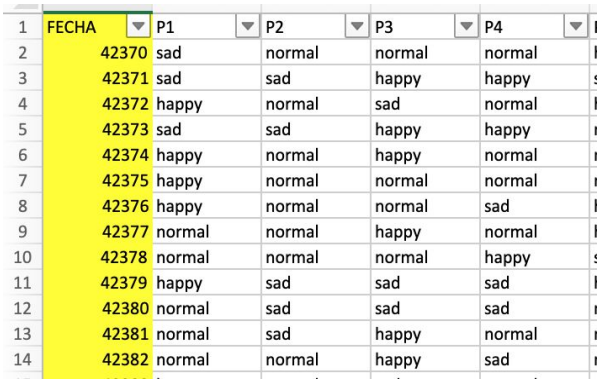
After a closer look at the data set we found some inconsistency and missing values :

- **Sheet "HOURS SLEEP VALUES":** first, there is an extra row (row 368) at the end with no date related to it. Based on the number of rows in other tables (367), we conclude that this row of values is not relevant and outside of time period. Second, the values of hours sleep are in text format, instead of integers.

366	30/12/2016	4	5	4	4	5	5	5	4	4	5	5	6	5	4	4	4	5	4	5	6		
367	31/12/2016	4	5	5	4	5	4	6	4	4	4	5	5	5	5	4	4	6	4	5	6		
368		4	4	5	5	5	4	6	4	4	5	5	6	5	5	4	4	5	5	5	5		

Figure 1: Invalid data in table "HOURS SLEEP VALUES"

- **Sheet “MOOD VALUES”:** first**,** the values in “Date” (“Fecha”) column are not in Datetime format, but in Text / General format, which results in incorrect display of dates. Second, mood values are written all in lowercase letters, whereas all other words in the Excel spreadsheet are written all in uppercase letters. We believe all tables should follow one written style, that is in our case being in all uppercase letters.



	FECHA	P1	P2	P3	P4
2	42370	sad	normal	normal	normal
3	42371	sad	sad	happy	happy
4	42372	happy	normal	sad	normal
5	42373	sad	sad	happy	happy
6	42374	happy	normal	happy	normal
7	42375	happy	normal	normal	normal
8	42376	happy	normal	normal	sad
9	42377	normal	normal	happy	normal
10	42378	normal	normal	normal	happy
11	42379	happy	sad	sad	sad
12	42380	normal	sad	sad	sad
13	42381	normal	sad	happy	normal
14	42382	normal	normal	happy	sad

Figure 2: incorrect date format in table “MOOD VALUES”

- **All Sheets, except Sheet “PATIENTS”:** the column name “FECHA” in all tables should be replaced by “DATE”, given that all other information in tables is written in English.

Data Warehouse Approach selection

Contrary to other industries such as retail, the healthcare industry is characterized by a series of events and processes that have different outcomes depending on the point of reference. Trying to focus on user understandability and due to relatively small amount of information, we decided to follow multidimensional architecture. Based on the Excel file provided with the data, we identified level of granularity, fact measures and conformed dimensions. In addition, multidimensional models are easily to adapt to unexpected changes in the behavior of the analyst - a common trait in the healthcare industry as doctors have different treatments approaches and process.

Moreover, the data set refers to past collected information that will not change over time and do not need that increased level of flexibility and agility that Data Vault approach would give.

Data Warehouse Design

Our process to establish the design was:

1. Define the most useful outcome of the research
2. Make some assumptions on how the data was collected and who would be the end users of the information
3. Define and choose the healthcare process to model;
4. Choose the dimensions, facts and attributes from the data set;
5. Define the relations and hierarchies between the facts and dimensions;
6. Define the facts (measures) that will populate each fact table, and design aggregation rules;

7. Implement the design of a constellation of fact tables.

Schema: The schema chosen to better model the data was the constellation of fact tables.

Flexibility of the chosen model:

The design permits some flexibility in case there is a need to update information in the future or the need to add new dimensions or attributes. As an example, it is possible to include new attributes to the dimension table patient if needed in the future.

On top of that, the existing fact tables are the symptoms or particular observations found in each patient, they are not modifiable because it is based on past documentation. However, the design allows for the addition of new fact tables in the constellation in case there is the need.

Identities:

- **Fact tables:**

It was the group choice to have two fact tables, instead of four (one for each Excel Sheet) . The first fact table contains information on both emotional (MOOD) and physical (EPISODES) wellness. The second table contains information on both day (ACTIVITY) and night (SLEEP) behaviour. We believe this grouping of tables simplifies and increases the understandability of the model.

  - ***Fact table "Wellness" :*** contains as attributes, the measure of the severity of the episode and the mood value a particular patient had in a specific date.
  - ***Fact table "Behavior"***: contains as attributes, type of activity a patient did and how many hours of sleep a patient had in a particular date.

- **Dimension tables :**

Since fact tables "Wellness" and "Behavior" shared same dimension of patient and date, and patient also has its own dimension of city and disorder status, we have four conformed dimensions in total.

  - ***Dimension table "Date"*** *:* this table stores every single date as a unique value. To give more granularity level to this dimension for further analysis, we expand date information with year, month, week and day.
  - ***Dimension table "Patient"****:* each patient is stored as a unique value (patient code). This table includes information on city and disorder of each patient. What's more, considering patients' information may change by time in the future, we added date of creating and updating for retrieving the change.
  - ***Dimension table "City"****:* linked to ***"Patient"*** table. This table stores information of cities, including city name and environment. We separated this table from the "Patient" table in case there would be the need to update this table by adding new attributes (population, air pollution level, noise pollution level, etc).
  - ***Dimension table "Disorder"****:* linked to ***"Patient"*** table. This table stores information of types of cognitive disorders of each patient. Currently it stores dementia, delirium and

amnesia. We separated this table from “Patient” table in case we would like to update the table by adding new attributes (disorder code, description of disorder, etc).

Structure of Data Warehouse design:

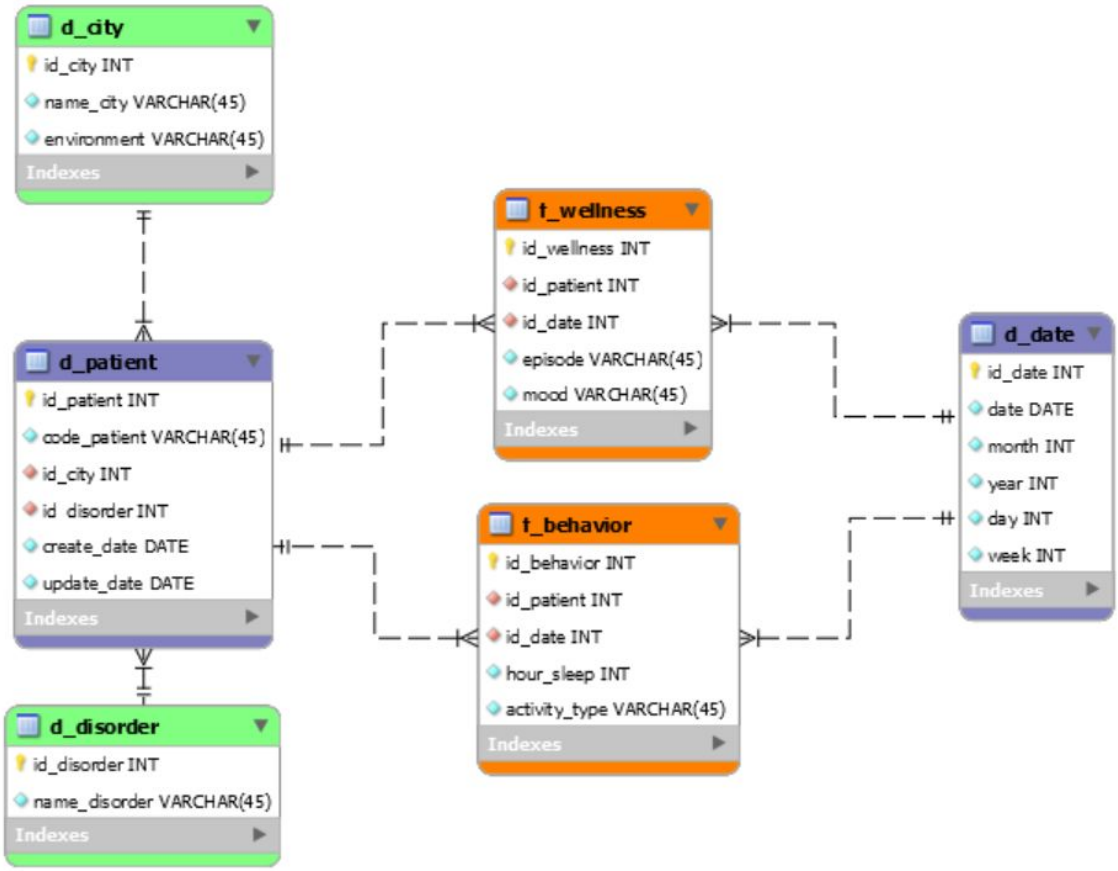


Figure 3: Structure of final database

References:

(1) *Kimball, Ralph, and Margy Ross. The Data Warehouse Toolkit : The Definitive Guide to Dimensional Modeling. Indianapolis, Wiley, Cop, 2013.*

(2) *Parmanto, Bambang et al. "A framework for designing a healthcare outcome data warehouse." Perspectives in health information management vol. 2 3. 6 Sep. 2005*

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2047311/