Data Warehouse ETL

Business Intelligence & Data Warehousing

PROFESSOR JOSÉ CURTO DÍAZ

Group D

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

Data Mapping document:

1. Business needs:

From an ETL designer's view, the business needs are the DW/BI system users' information requirements. Given our data set, we have created two separate fact tables since we have the possibility to analyze the information in a segregated way. This was done in order to satisfy the users' need of evaluating both behavior and wellness. Also the dataset was pivoted into a more simple structure to allow users to make comparisons between data easier.

In our case, the need is to provide the user (doctors, healthcare, therapists…) with a pluri-dimensional chronological view of the patient's health condition. For this we are creating two fact tables. The Wellness fact table will report the patient's wellness state based on two metrics (Episode values and Mood values). The Behavior fact table adds a new perspective as it reports patient behaviour based on two different metrics (Hours of Sleep and Activities performed during the day). This way the users can follow the reports and use the key performance indicators to assist on future decisions.

2. Compliance:

To provide proof that the reported numbers are accurate, complete, and have not been tampered we matched the results in Pentaho after finishing the transformation with the original Excel source. More than 10 samples were executed and metadata was included to control updates and creation.

After a comparative review of the reported numbers with our source dataset, we manually checked the compliance and reliability of our results. Our Dimension and Fact tables reported the correct values with no redundancy or errors.

3. Data Quality:

After carefully analyzing the source, some data elements were found inconsistent and some missing values were observed. Before extraction the following data issues were reported:

- **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.

4. Security:

Our dataset doesn’t have Personal Identifiable Information (PII) data and therefore we decided not to modify any information.

Besides, we created a backup of the database and origin source in each of our machines.

5.Data Integration

Data integration is a major matter as it aims to bring all dimensions and facts to work together and produce a business outcome.

In the context of data warehousing, dimensions are what describes the business environment with labels that connects to the business intelligence tool so that it can be properly used and understood by business analysts. Facts are the outcomes of a business process and it corresponds to a perceptible event rather than a specific demand of a report.

In order to identify the business processes and therefore the integration of data, we established the order the dimension tables should be populated in. From then on, the only dimension that needed a hierarchical rollup was “d_patient”. This particular dimension looks up information from “d_city” and “d_disorder” dimension tables in the database. Subsequently, the two fact tables are partly populated by the “d_date” and “d_patient” dimension tables. The final ETL process is as follows:

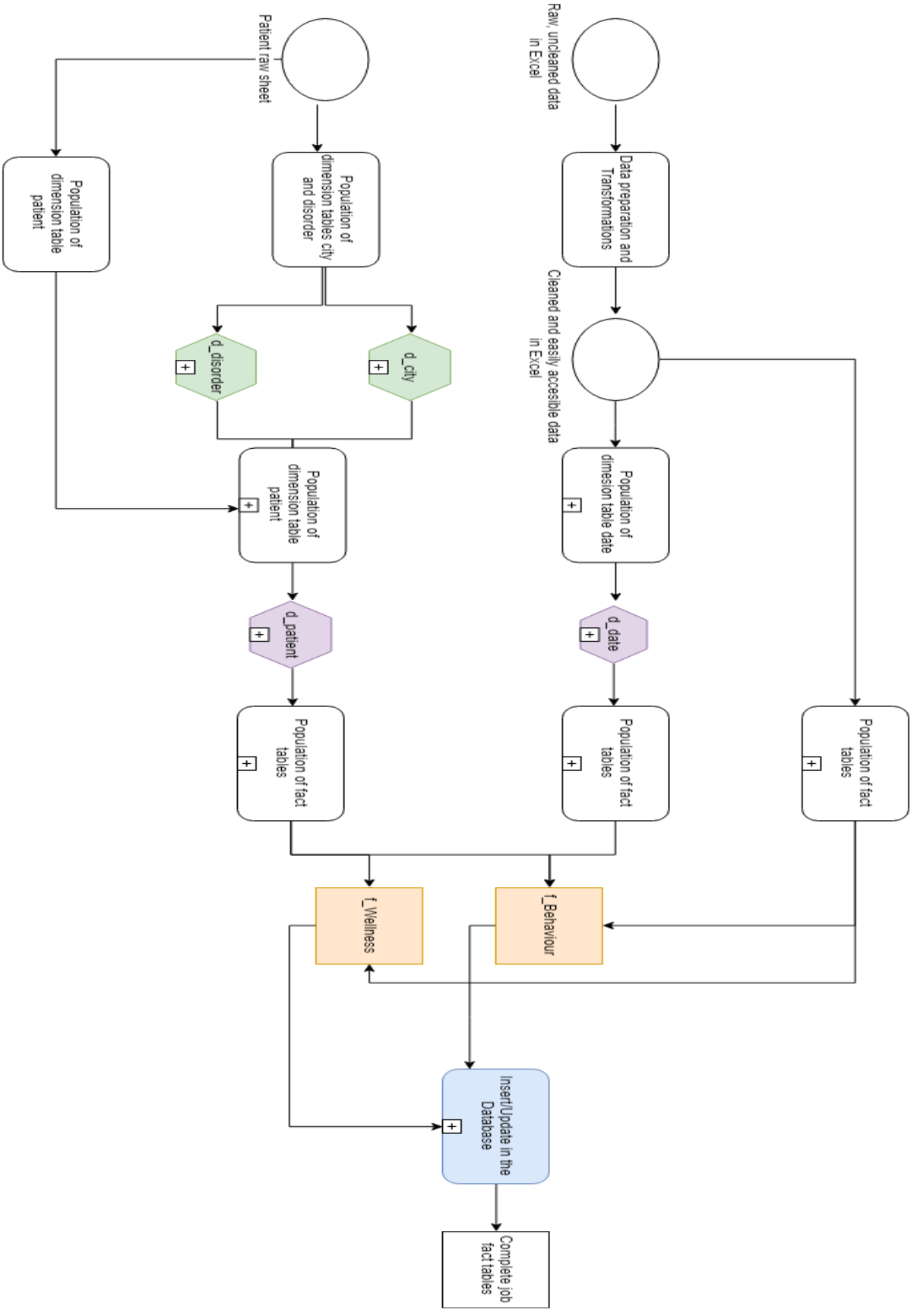

Figure 3: General ETL process in our case

6. Data Latency

Data latency describes how quickly the source system data must be delivered to the business users via the DW/BI system. It is safe to assume that the business users will be doctors, therapists or researchers. Having said this, latency should always be considered as an important aspect however we believe that in our case it will be sufficient. Given the structure of our DW and that our business users do not need to be updating the information in real time - as they are working with past data that can be updated every week or month.

7. Archiving and Lineage

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.

8. BI Delivery Interfaces

The fact tables that are going to be directly exposed to the BI tool are Wellness and Behaviour, already extensively described in the past paper.

9. Available Skills and Legacy Licenses

The ETL system used was Pentaho | Hitachi Ventara. On top of that, the database was created using MySQL Workbench.
We also made a backup of the database.

.

Data mapping process

In order to map the ETL process, the following steps were executed:

Step 1: Draw the high level plan:

We start the design process with a very simple schematic of the known pieces of the plan: sources and targets. See figure below.

Step 2: Choose ETL Tool:

Pentaho Data Integration was the tool used, as explained above.

Step 3 and 4: Develop default strategies and drill down by target table:

From the start, we have defined the need to drill down the multiple sheets from the Excel file where data were recorded into an unique and structured data base. There data should be clean and easily accessible in order to support the business users.

According to the database schema we chosen before and the data structure of Excel we have, we made the data mapping between target tables in database and sheets of Excel file.

Target					Source			Transformation
Table_name	Column_name	Data_type	Table_type	SCD_type	Sheet_name	Column_name	Data_type	
d_patient	id_patient	integer	dimension	Y				surrogate key
d_patient	code_patient	varchar(10)	dimension	Y	PATIENT	PATIENT	varchar	
d_patient	id_city	integer	dimension	Y				surrogate key
d_patient	id_disorder	integer	dimension	Y				surrogate key
d_patient	create_date	date	dimension	Y				define as the date of uploading the data at the first time
d_patient	update_date	date	dimension	Y				define as the date of updating the data
d_city	id_city	integer	dimension	N				surrogate key
d_city	name_city	varchar(45)	dimension	N	PATIENT	CITY	varchar	
d_city	environment	varchar(45)	dimension	N	PATIENT	ENVIRONMENT	varchar	
d_disorder	id_disorder	integer	dimension	N				surrogate key
d_disorder	name_disorder	varchar(45)	dimension	N	PATIENT	COGNITIVE DISORDER	varchar	
d_date	id_date	integer	dimension	N				surrogate key
d_date	date	date	dimension	N	every sheet, except PATIENT	FECHA	date	
d_date	month	integer	dimension	N				retrieved from date
d_date	year	integer	dimension	N				retrieved from date
d_date	day	integer	dimension	N				retrieved from date
d_date	week	integer	dimension	N				retrieved from date
f_wellness	id_wellness	integer	fact	N				surrogate key
f_wellness	id_patient	integer	fact	N				foreign key
f_wellness	id_date	integer	fact	N				foreign key
f_wellness	episode	varchar(45)	fact	N	EPISODE VALUES	cells	varchar	row normalizer to turn cells to values of the column
f_wellness	mood	varchar(45)	fact	N	MOOD VALUES	cells	varchar	row normalizer to turn cells to values of the column
f_behavior	id_behavior	integer	fact	N				surrogate key
f_behavior	id_patient	integer	fact	N				foreign key
f_behavior	id_date	integer	fact	N				foreign key
f_behavior	hour_sleep	integer	fact	N	HOURS SLEEP VALUES/ACTIVITY	cells	integer	row normalizer to turn cells to values of the column
f_behavior	activity_type	varchar(45)	fact	N	HOURS SLEEP VALUES/ACTIVITY	cells	varchar	row normalizer to turn cells to values of the column

Figure 4: Data mapping between target table and source of Excel

In our schema that dimensions for City (d_city) and Disorder (d_disorder) are separated from Patient dimension (d_patient). Since Patient dimension (d_patient) table looks up in the database information from City (d_city) and Disorder (d_disorder) dimension tables, we first needed to populate City (d_city) and Disorder (d_disorder) tables. After Patient dimension table (d_patient), we populate the last dimension table, which is Date dimension table (d_date). After dimension tables are ready, we start to populate 2 fact tables, Wellness (f_wellness) and Behavior (f_behavior). To sum up, we decided to input data in the following order: d_city→ d_disorder→d_patient →d_date → f_behavior→f_wellness.

Step 5: Populate Dimension Tables with Historic Data

1.Data Preparation:

The first step consisted on cleaning the source delivered in Excel format. A transformation called “TR_DATAPREP” which contains 19 steps was created for the main purpose of converting the four sheets in the Excel file (HOURS OF SLEEP VALUES, ACTIVITY VALUES, EPISODE VALUE and MOOD VALUE) into a single, clean data set. Other achievements in this stage were: the merging of

all dates and patients into two attributes, the normalization of the information from rows into columns, fixing empty values and empty rows, and fixing the date format. The final result is shown in the picture below:

OUTPUT: Clean fact data:

Logging | Execution History | Step Metrics | Performance Graph | Metrics | Preview data

First rows | Last rows | Off

#	DATE	PATIENT_CODE	MOOD_VALUES	EPISODE_VALUES	ACTIVITY_VALUES	SLEEP_VALUES
1	2016/01/...	P1	SAD	LIGHT	NO ACTIVITY	4.0
2	2016/01/...	P10	HAPPY	LIGHT	NO ACTIVITY	5.0
3	2016/01/...	P11	HAPPY	LIGHT	NO ACTIVITY	5.0
4	2016/01/...	P12	HAPPY	NO EPISODE	SLEEP/SOFA	6.0
5	2016/01/...	P13	HAPPY	LIGHT	SLEEP/SOFA	6.0
6	2016/01/...	P14	SAD	LIGHT	RADIO/TV	4.0
7	2016/01/...	P15	NORMAL	LIGHT	FAMILY	5.0
8	2016/01/...	P16	NORMAL	NO EPISODE	RADIO/TV	4.0
9	2016/01/...	P17	HAPPY	MODERATE	SLEEP/SOFA	6.0
10	2016/01/...	P18	NORMAL	SEVERE	SLEEP/SOFA	4.0
11	2016/01/...	P19	NORMAL	SEVERE	SLEEP/SOFA	4.0
12	2016/01/...	P2	NORMAL	SEVERE	RADIO/TV	5.0
13	2016/01/...	P20	NORMAL	SEVERE	READ/STUDY	5.0
14	2016/01/...	P3	NORMAL	NO EPISODE	RADIO/TV	5.0
15	2016/01/...	P4	NORMAL	NO EPISODE	NO ACTIVITY	4.0
16	2016/01/...	P5	HAPPY	NO EPISODE	FAMILY	6.0
17	2016/01/...	P6	SAD	LIGHT	READ/STUDY	5.0
18	2016/01/...	P7	HAPPY	NO EPISODE	READ/STUDY	6.0
19	2016/01/...	P8	HAPPY	MODERATE	RADIO/TV	5.0
20	2016/01/...	P9	SAD	LIGHT	NO ACTIVITY	5.0
21	2016/01/...	P1	SAD	NO EPISODE	EXERCISE	4.0
22	2016/01/...	P10	NORMAL	LIGHT	FAMILY	4.0
23	2016/01/...	P11	NORMAL	NO EPISODE	NO ACTIVITY	4.0
24	2016/01/...	P12	SAD	SEVERE	RADIO/TV	5.0
25	2016/01/...	P13	HAPPY	NO EPISODE	EXERCISE	6.0
26	2016/01/...	P14	NORMAL	LIGHT	RADIO/TV	4.0
27	2016/01/...	P15	HAPPY	NO EPISODE	READ/STUDY	4.0
28	2016/01/...	P16	SAD	NO EPISODE	EXERCISE	5.0
29	2016/01/...	P17	HAPPY	MODERATE	FAMILY	6.0
30	2016/01/...	P18	HAPPY	LIGHT	SLEEP/SOFA	5.0
31	2016/01/...	P19	HAPPY	LIGHT	RADIO/TV	5.0
32	2016/01/...	P2	SAD	SEVERE	NO ACTIVITY	4.0
33	2016/01/...	P20	HAPPY	LIGHT	READ/STUDY	5.0

Figure 5: Result of Data preparation process

The city, disorder, and patient dimension were populated directly from the original source but date dimension and fact tables were populated from the new clean file ***CleanFactData***.

2.Date dimension transformation and population process:

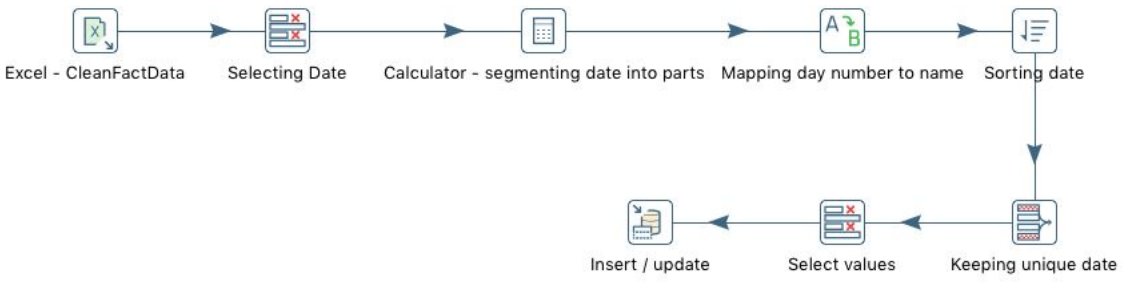

Figure 6: Process of Date dimension transformation

OUTPUT: Date Dimension Table

The table was reorganized and some attributes such as id, year, month, date, week of the year and day of the week were added.

First rows Last rows Off

#	DATE	YEAR	MONTH	DAY	WEEK OF YEAR	DAY_OF_WEEK_NAME
1	2016/01/01 00:00:0...	2016	1	1	1	SATURDAY
2	2016/01/02 00:00:0...	2016	1	2	1	SUNDAY
3	2016/01/03 00:00:0...	2016	1	3	2	MONDAY
4	2016/01/04 00:00:0...	2016	1	4	2	TUESDAY
5	2016/01/05 00:00:0...	2016	1	5	2	WEDNESDAY
6	2016/01/06 00:00:0...	2016	1	6	2	THURSDAY
7	2016/01/07 00:00:0...	2016	1	7	2	FRIDAY
8	2016/01/08 00:00:0...	2016	1	8	2	SATURDAY
9	2016/01/09 00:00:0...	2016	1	9	2	SUNDAY
10	2016/01/10 00:00:0...	2016	1	10	3	MONDAY
11	2016/01/11 00:00:0...	2016	1	11	3	TUESDAY
12	2016/01/12 00:00:0...	2016	1	12	3	WEDNESDAY
13	2016/01/13 00:00:0...	2016	1	13	3	THURSDAY
14	2016/01/14 00:00:0...	2016	1	14	3	FRIDAY
15	2016/01/15 00:00:0...	2016	1	15	3	SATURDAY
16	2016/01/16 00:00:0...	2016	1	16	3	SUNDAY
17	2016/01/17 00:00:0...	2016	1	17	4	MONDAY
18	2016/01/18 00:00:0...	2016	1	18	4	TUESDAY
19	2016/01/19 00:00:0...	2016	1	19	4	WEDNESDAY
20	2016/01/20 00:00:0...	2016	1	20	4	THURSDAY
21	2016/01/21 00:00:0...	2016	1	21	4	FRIDAY
22	2016/01/22 00:00:0...	2016	1	22	4	SATURDAY
23	2016/01/23 00:00:0...	2016	1	23	4	SUNDAY
24	2016/01/24 00:00:0...	2016	1	24	5	MONDAY
25	2016/01/25 00:00:0...	2016	1	25	5	TUESDAY
26	2016/01/26 00:00:0...	2016	1	26	5	WEDNESDAY
27	2016/01/27 00:00:0...	2016	1	27	5	THURSDAY
28	2016/01/28 00:00:0...	2016	1	28	5	FRIDAY
29	2016/01/29 00:00:0...	2016	1	29	5	SATURDAY
30	2016/01/30 00:00:0...	2016	1	30	5	SUNDAY
31	2016/01/31 00:00:0...	2016	1	31	6	MONDAY
32	2016/02/01 00:00:0...	2016	2	1	6	TUESDAY
33	2016/02/02 00:00:0...	2016	2	2	6	WEDNESDAY

Figure 7: Result of Date dimension transformation

3. Patient, city and disorder dimensions transformation and population process:

City and disorder dimension were created in two separate transformations. The creation process of each of them consisted in: the selection of the needed information, sorting the information in columns and unifying the rows into unique values. After this two dimensions were populated in the database, the transformation of the patient dimension was elaborated according to the diagram below:

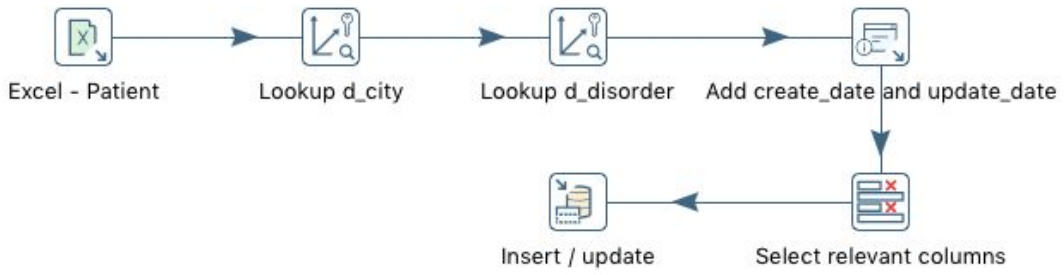

Figure 8: Process of Patient dimension transformation

This process begins with the extraction of patient information from source, then the city and disorder were lookup directly from the dimension d_disorder and d_city. Metadata was added to include “create_date” and “update_date”. It may be used later for traceability and extensive analysis.

OUTPUT: Patient dimension table

First rows Last rows Off

#	PATIENT	id_city	id_disorder	create_date	update_date
1	P1	2	2	2019/05/20 17:41:3...	2019/05/20 17:41:3...
2	P2	10	2	2019/05/20 17:41:3...	2019/05/20 17:41:3...
3	P3	13	2	2019/05/20 17:41:3...	2019/05/20 17:41:3...
4	P4	15	2	2019/05/20 17:41:3...	2019/05/20 17:41:3...
5	P5	5	2	2019/05/20 17:41:3...	2019/05/20 17:41:3...
6	P6	8	2	2019/05/20 17:41:3...	2019/05/20 17:41:3...
7	P7	14	2	2019/05/20 17:41:3...	2019/05/20 17:41:3...
8	P8	2	3	2019/05/20 17:41:3...	2019/05/20 17:41:3...
9	P9	3	3	2019/05/20 17:41:3...	2019/05/20 17:41:3...
10	P10	16	3	2019/05/20 17:41:3...	2019/05/20 17:41:3...
11	P11	8	3	2019/05/20 17:41:3...	2019/05/20 17:41:3...
12	P12	1	3	2019/05/20 17:41:3...	2019/05/20 17:41:3...
13	P13	7	3	2019/05/20 17:41:3...	2019/05/20 17:41:3...
14	P14	12	3	2019/05/20 17:41:3...	2019/05/20 17:41:3...
15	P15	2	1	2019/05/20 17:41:3...	2019/05/20 17:41:3...
16	P16	11	1	2019/05/20 17:41:3...	2019/05/20 17:41:3...
17	P17	4	1	2019/05/20 17:41:3...	2019/05/20 17:41:3...
18	P18	15	1	2019/05/20 17:41:3...	2019/05/20 17:41:3...
19	P19	8	1	2019/05/20 17:41:3...	2019/05/20 17:41:3...
20	P20	6	1	2019/05/20 17:41:3...	2019/05/20 17:41:3...

Figure 9: Result of Patient dimension transformation

Step 6: Perform the fact table historic load

Figure 10: Process of the transformation of Wellness table

Figure 11: Process of the transformation of Behavior table

In this step, we use the clean data set to load our Fact tables. Through the “Combined Lookup/Update” function, we are able to combine our attributes of interest (Patient id & Date id) from both patient and date dimensions tables. We then select the attributes for our fact tables and load the result to the database.

OUTPUT Fact tables

First rows Last rows Off

#	MOOD_VALUES	EPISODE_VALUES	id_patient	id_date
1	SAD	LIGHT	1	1
2	HAPPY	LIGHT	10	1
3	HAPPY	LIGHT	11	1
4	HAPPY	NO EPISODE	12	1
5	HAPPY	LIGHT	13	1
6	SAD	LIGHT	14	1
7	NORMAL	LIGHT	15	1
8	NORMAL	NO EPISODE	16	1
9	HAPPY	MODERATE	17	1
10	NORMAL	SEVERE	18	1
11	NORMAL	SEVERE	19	1
12	NORMAL	SEVERE	2	1
13	NORMAL	SEVERE	20	1
14	NORMAL	NO EPISODE	3	1
15	NORMAL	NO EPISODE	4	1
16	HAPPY	NO EPISODE	5	1
17	SAD	LIGHT	6	1
18	HAPPY	NO EPISODE	7	1
19	HAPPY	MODERATE	8	1
20	SAD	LIGHT	9	1
21	SAD	NO EPISODE	1	2
22	NORMAL	LIGHT	10	2
23	NORMAL	NO EPISODE	11	2
24	SAD	SEVERE	12	2
25	HAPPY	NO EPISODE	13	2
26	NORMAL	LIGHT	14	2
27	HAPPY	NO EPISODE	15	2
28	SAD	NO EPISODE	16	2
29	HAPPY	MODERATE	17	2
30	HAPPY	LIGHT	18	2
31	HAPPY	LIGHT	19	2
32	SAD	SEVERE	2	2

Figure 12: Result of the transformation of Wellness table

Step 7: ETL System Operation and Data quality check

At last, we combine all the transformations into a complete job and execute in Pentaho. After execution of all transformations, we verify total number of rows for all tables we created to make sure we can get right data in our database.

New changes in SQL model:

After a comprehensive review of our first model, we decided to optimise some aspects for better performance and granularity:

- Considering analyzing patients' status by day-of-week dimension, we added 'day_of_week' to table 'd_date'.
- In order to avoid misunderstanding of 'week' column, we changed 'week' in table 'd_date' to 'week_of_year';
- Considering the format for 'code_patient' (P plus number), we changed type of 'code_patient' in table 'd_patient' from varchar(45) to varchar(10).
- In order to avoid making mistakes when updating primary keys, we changed every primary key of all tables into auto-increment status.

References

1. Kimball, R. and Ross, M. (2013). *The data warehouse toolkit*. Indianapolis: Wiley.