Sistema de inteligencia de negocio entorno a las enfermedades cognitivas

María Barcia Calviño
Máster en Big Data y Business Inteligence
Business Intelligence

David Amorós Alcaraz
María Isabel Guitart Hormigo

03/04/2017

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Sistema de inteligencia de negocio entorno a las enfermedades cognitivas*
Nombre del autor:	*María Barcia Calviño*
Nombre del consultor/a:	*David Amorós*
Nombre del PRA:	*Isabel Guitart*
Fecha de entrega (mm/aaaa):	07/2017
Titulación:	*Máster en Big Data y Business Intelligence*
Área del Trabajo Final:	*Sistemas de Información*
Idioma del trabajo:	*Castellano*
Palabras clave	*Data Warehouse, Análisis, BI*

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados y conclusiones del trabajo.*

En el siglo 21 la calidad de vida ha aumentado lo que supone un aumento de personas de la tercera edad. Esto ha supuesto el incremento de personas con enfermedades cognitivas y es por ello que surge la necesidad de desarrollar técnicas y sistemas que nos permitan mejorar la vida de estos paciente.

En este trabajo fin de master se ha desarrollado un Data Warehouse para el almacenamiento de los datos obtenidos durante la exploración de los pacientes teniendo en cuenta sus episodios, actividades y horas de sueño a lo largo del año.

Para alimentar esta base de datos, se han desarrollado diferentes procesos ETL para la carga automática de datos y se han diseñado cubos OLAP que nos permiten analizar los datos y sacar conclusiones.

En general, los datos de los que disponemos para este TFM no son suficientes para sacar conclusiones determinantes, pero si se observa que el aumento de actividad física o social y el ambiente rural, mejora la calidad de vida de los pacientes o que dependiendo de que tipo de dolencia padezcan, los episodios y las horas de sueño se relacionan distintamente.

En un caso real sería necesario recopilar más datos y tener una población de prueba mayor que nos permitiese buscar métodos y soluciones que hagan la vida de estos enfermos más fácil.

Abstract (in English, 250 words or less):

In the 21st century, the life quality has increased potencially which means that Elder population has raised. This has created a need of understanding certain diseases, among them, the mental diseases.

In the Master thesis a Data Warehouse has been created and designed in order to understand better what is happening with this patient and try to find a way to improve their lifes.

To feed this data base, some ETLs processes have been developed in order to load the data automatically and some OLAP cubes have been created so that we can extract informaction from the Data Warehouse and we can rease some conclutions.

In fact, the data that it has been provided for this case is not enough but we can conclude that the patients that leaves in the country side and that do some exercise or they meet the famliy offenly, their life quality is better or that depending on the mental disorder, the episode and the actovioty the sleeping hours vary in a different way.

In a real case we would need some extra data and more patients in order to arrive to strong conclusions that would help in the research of ways to improve the parients life.

Índice

Lista de figuras

No se encuentran elementos de tabla de ilustraciones.

1. Introducción

1.1 Contexto y justificación del Trabajo

Es por todos sabido que actualmente la calidad de vida de los seres humano ha mejorado en comparación con los años anteriores, así como la esperanza de vida. Nuestros mayores cada vez son más longevos, pero con ello se hacen más números los casos de enfermedades cognitivas con los que se han de tratar. Si bien este tipo de enfermedades no sólo afectan a personas ancianas, un gran grupo de los afectados pertenecen a este sector de edad.

Se desconoce mucha información relacionada con este tipo de enfermedades y las herramientas de las que se disponen no son demasiadas. Es por ello que un sistema de inteligencia de negocio ayudaría no solo a entender un poco más estas enfermedades si no a analizar cuáles son los síntomas, los estados de ánimo y actividades relacionadas con ellas y qué actividades pueden ayudar a paliar sus consecuencias y por tanto mejorar la calidad de vida de los pacientes.

1.2 Objetivos del Trabajo

El objetivo general es el diseño e implementación de un sistema de Business Intelligence que facilite la adquisición, el almacenamiento y la explotación de datos asociados a pacientes con enfermedades cognitivas provenientes de diferentes centros médicos.

De manera más específica los objetivos son:

1. Diseñar un almacén de datos (Data Warehouse) que permita almacenar la información adquirida desde los diferentes orígenes de datos situados en cada centro médico, teniendo en cuenta que cada centro médico estará formado por un grupo de terapeutas con un cierto número de pacientes asignados.
2. Implementar este almacén de datos y programar los procesos ETL (extracción, transformación y carga) que permitan alimentar el DW a partir de los ficheros base facilitados.
3. Analizar las diferentes plataformas BI OS disponibles en el mercado que nos permitan explorar la información almacenada.
4. Elegir una de estas herramientas de tal forma que se disponga de una capa de software para el análisis de la información.

1.3 Enfoque y método seguido

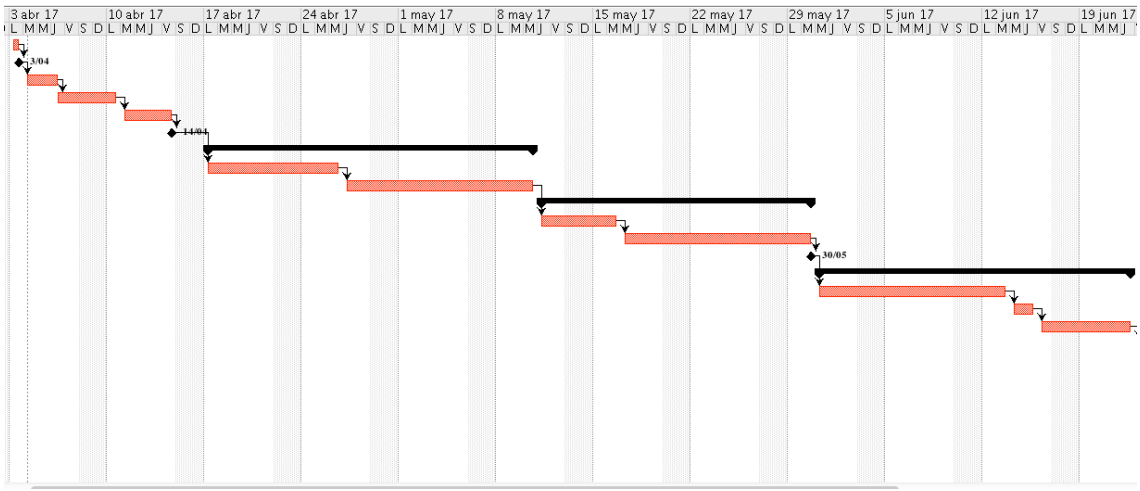
Debido a que el producto final no tiene fines comerciales y servirá únicamente como un trabajo de máster, se ha decidido desarrollar un producto nuevo de manera que se puedan explotar todas las fases de aprendizaje que se han llevado a cabo durante el

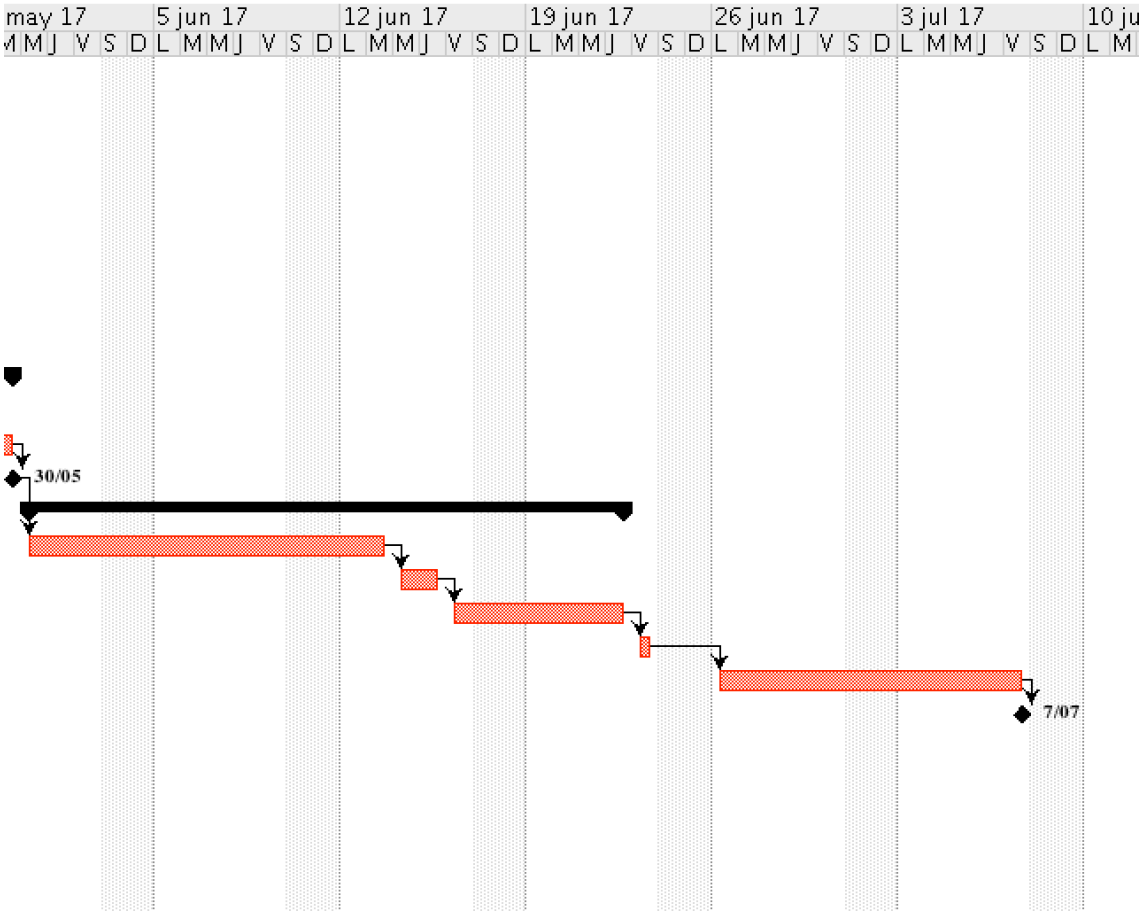
máster. De esta manera se diseñará e implementará un almacén de datos desde cero, así como las ETLs.

1.4 Planificación del Trabajo

A continuación se adjuntan el diagrama de Gantt y la planificación del proyecto.

	Nombre	Duracion	Inicio	Terminado	Predecesores
	Propuesta inicial	2 days	2/04/17 8:00	3/04/17 17:00	
	PEC 1	0 days	3/04/17 17:00	3/04/17 17:00	1
	Análisis de requisitos	5 days	4/04/17 8:00	6/04/17 13:00	2
	Elección entorno tecnoló	5 days	6/04/17 13:00	10/04/17 17:00	3
	Instalación y configuraci	8 days	11/04/17 8:00	14/04/17 17:00	4
	PEC 2	0 days	14/04/17 17:00	14/04/17 17:00	5
	DWH	36 days	**17/04/17 8:00**	**10/05/17 17:00**	
	Modelizar DWH	16 days	17/04/17 8:00	26/04/17 17:00	6
	Implementar DWH	20 days	27/04/17 8:00	10/05/17 17:00	8
	Datos	28 days	**11/05/17 8:00**	**30/05/17 17:00**	
	Análisis Datos origen	8 days	11/05/17 8:00	16/05/17 17:00	9
	Creacion de ETLS	20 days	17/05/17 8:00	30/05/17 17:00	11
	PEC 3	0 days	30/05/17 17:00	30/05/17 17:00	12
	Resultados	34 days	**31/05/17 8:00**	**22/06/17 17:00**	
	Creacion front end	20 days	31/05/17 8:00	13/06/17 17:00	13
	Generar resultados	4 days	14/06/17 8:00	15/06/17 17:00	15
	Interpretación resultac	10 days	16/06/17 8:00	22/06/17 17:00	16
	Conclusiones	2 days	23/06/17 8:00	23/06/17 17:00	17
	Informe final	20 days	26/06/17 8:00	7/07/17 17:00	18
	Memoria final	0 days	7/07/17 17:00	7/07/17 17:00	19





1.5 Breve sumario de productos obtenidos

Los productos obtenidos será el Data Warehouse y los ETLs para la carga y transformación de datos, así como una herramienta que permita analizar los datos almacenados y facilite la obtención de información. Además de un documento que recopile todo el proceso.

1.6 Breve descripción de los otros capítulos de la memoria

La memoria que sigue explicará el proceso de creación del sistema de negocio y su posterior explotación. Para ello habrá en primer lugar un análisis de los requisitos y tecnologías a utilizar. A continuación, se mostrará el proceso de creación del data warehouse y las ETL para la carga de datos. Finalmente, se explicará cómo se han mostrado los datos y se hará un análisis. Las conclusiones serán el punto y final del trabajo.

2. Análisis de requisitos

El proceso de análisis de requisitos consiste en definir y documentar las necesidades de los interesados a fin de cumplir con los objetivos del proyecto.

El proceso de recopilar requisitos es determinante para la correcta evolución del proyecto y está directamente relacionado con la implicación activa de los interesados pues se trata de conocer lo que ellos esperan del proyecto, del negocio y de los productos o servicios a obtener.

2.1 Catálogo de requisitos

Una vez definidos los requisitos, los detallaremos según la tabla que se adjunta a continuación.

Identificador	Identificador único del requisito. Comenzará por RF o RNF según el tipo de requisito, seguido de un número.
Título	Enunciado del contenido del requisito.
Descripción	Explicación del requisito.
Prioridad	Grado de importancia de cumplimiento del requisito.
Criterio de validación	Criterio según el cual, una vez entregado, se considerará que el requisito se ha cumplido.

2.1.1 Requisitos funcionales

Identificador	RF1
Título	Análisis de información.
Descripción	El sistema deberá ser capaz de permitir el análisis de la información almacenada .
Prioridad	Alta.
Criterio de validación	Se han podido extraer informaciones del sistema.

Identificador	RF2
Título	Carga de datos.
Descripción	El sistema deberá ser capaz de permitir la carga de datos.
Prioridad	Alta.
Criterio de validación	Se han podido cargar nuevos datos a partir de los documentos con datos de origen.

2.1.2 Requisitos no funcionales

Identificador	RNF1
Título	Ficheros de entrada Excel
Descripción	Los ficheros de entrada de los datos provistos han de ser formato Excel.
Prioridad	Alta
Criterio de validación	Los ficheros de datos son de formato Excel.

3. Elección entorno tecnológico

Se nos ha dado plena flexibilidad en lo referente a entorno tecnológico, donde la única limitación (como ya hemos dicho anteriormente) es que los ficheros de entrada sean en formato Excel.

Por un lado, se ha decidido utilizar una máquina con sistema macOS.

Se ha elegido como sistema gestor de base de datos PostgreSQL por ser open source, fácil de utilizar y poner en funcionamiento y por mis conocimientos previos con el sistema. Con este SGBD modelaremos el Data Warehouse para el almacenamiento de los datos que se hayan proveído.

En cuanto a qué herramienta elegiremos para los procesos ETL (extracción, transformación y carga) me he decantado por la plataforma Pentaho open source, que si bien no es completamente gratuita si se requiere soporto o ciertas funcionalidades extra, para el caso que nos ocupa es suficiente y nos provee toda la funcionalidad y potencia necesaria para tratar nuestros datos. Será con la herramienta Spoon con la que crearemos las ETLs.

Para los cubos OLAP se había decidido en un principio utilizar Schema Workbench de Pentaho pero tras utilizarlo los resultados no han sido los esperados ya que no facilita la exploración de los datos ni se integra fácilmente con otra aplicaciones como Excel.

Por ello, se ha decidido utilizar Analysis Service de Microsoft SQLServer que junto con Visual Studio nos proporciona una herramienta muy potente y versátil que permite explorar los datos de manera sencilla y que integra con Excel, de manera que la obtención de gráficos con los datos es muy sencilla.

Pero la elección de Analysis Service ha supuesto la utilización Windows 7. Como los datos ya se habían creado en la máquina local macOS en PostgreSQL, se ha creado una máquina virtual con VirtualBox en el que se ha instalado todo el entorno SQLServer y se ha conectado a la máquina local donde se encontraba el PostgreSQL con los datos ya insertados. Esto ha supuesto un esfuerzo importante.

4. Diseño de la base de datos

4.1 Datos de entrada

Se nos han proporcionado los datos en formato Excel, en los que tenemos 4 hojas con los diferentes datos proporcionados.

- PATIENTS: datos de los pacientes. Los campos son cognitive disorder, city, environment y el propio paciente (P1, P2, P3…)
- HOURS SLEEP VALUES: fecha de la toma de datos y las horas de sueño para cada paciente
- ACTIVITY VALUES: fecha de la toma de datos para cada paciente y la actividad realizada. Tenemos tipificadas las actividades: NO ACTIVITY, EXERCISE, FAMILY, SLEEP/SOFA, RADIO/TV (Nota: Cuando estaba escribiendo esta memoria, me he dado cuenta que me ha faltado la actividad READ/STUDY aunque no afecta al desarrollo del caso, si faltarían datos relacionados con esta actividad)
- EPISODE VALUES: fecha de la toma de datos para cada paciente y el episodio sufrido. Tendremos tipificados los episodios: LIGHT, MODERATE, SEVERE y NO EPISODE

4.2 Dimensiones

A partir de los datos anteriormente explicados se han diseñado las dimensiones. En este caso es fácil observar cuales son dichas dimensiones ya que prácticamente se corresponden con las hojas del fichero Excel.

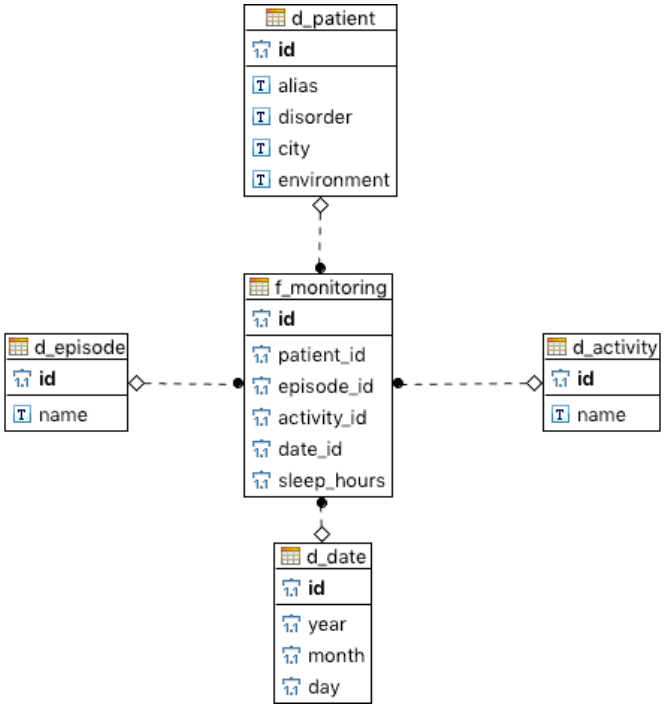
- d_patient: contendrá la información concerniente con los datos del paciente.
- d_episode: contendrá la información concerniente a los tipos de episodios que puede tener un paciente.
- d_activity: contendrá la información concerniente a las actividades que puede realizar un paciente.
- d_date: es la dimensión del tiempo, que almacenará las fechas con granularidad máxima de día

4.3 Tabla de hechos

En este caso tendremos una sola tabla de hechos, a la que llamaremos f_monitoring y que contendrá los id a las dimensiones anteriormente indicadas y las horas de sueño de cada paciente. Se le ha llamado monitoring a la tabla ya que al fin y al cabo, lo que estamos haciendo es un seguimiento = monitoreo de los pacientes.

4.4 Modelo ER del Data Warehouse

A continuación, se muestra el modelo entidad relación del Data Warehouse con sus dimensiones y tabla de hechos. Como se puede observar, tenemos un Data Warehouse con arquitectura en estrella, donde tenemos una única tabla de hechos que relaciona las diferentes dimensiones.


d_patient
id
alias
disorder
city
environment
f_monitoring
id
patient_id
episode_id
activity_id
date_id
sleep_hours
d_episode
id
name
d_activity
id
name
d_date
id
year
month
day

5. Transformaciones

Una vez tenemos la base de datos diseñada y creada en PostgreSQL, es el momento de insertar los datos. He de decir que se ha decidido insertar los datos de episodios y actividades directamente en la base de datos sin necesidad de transformaciones ya que éstos están tipificados y por tanto no pueden variar.

Lo primero que haremos será crear una transformación que inserte los pacientes en la base de datos. Para ello, leerá la hoja PATIENTS del fichero Excel, le añadirá la secuencia del id en la base de datos y finalmente guardará en paciente en la tabla correspondiente.

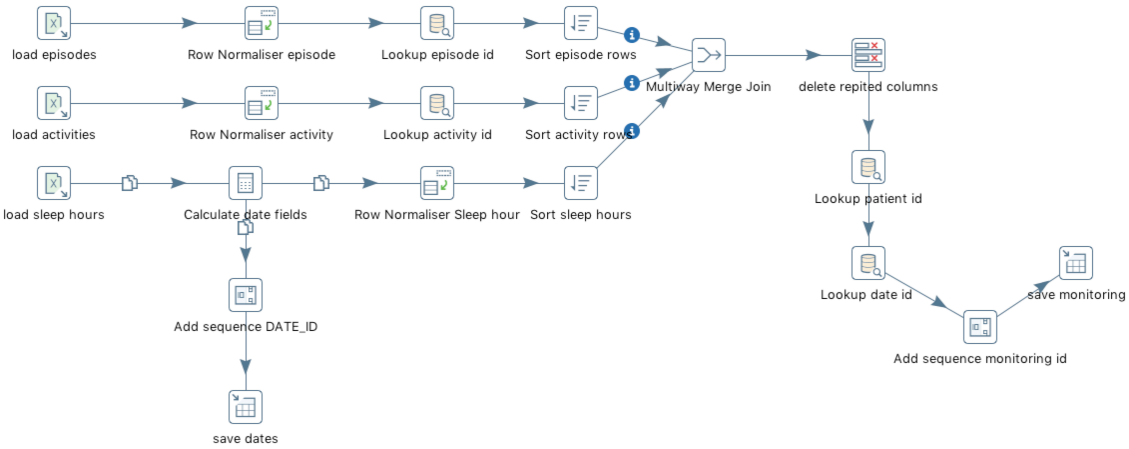
El siguiente paso es la inserción de la tabla de hechos f_monitoring. Lo primero que haremos será leer las restantes hojas del fichero Excel. En cuanto a la hoja de SLEEP HOURS es la que utilizaremos para recoger las fechas e insertarlas en la tabla d_date. Para ello, se ha hecho un cálculo de manera que dada una fecha obtengamos el año, mes y día de la misma y a continuación añadiremos la secuencia e insertaremos.

Una vez leidos los datos, normalizaremos la columna paciente de manera que cada paciente pase a ser una fila y cada valor (actividad, horas de sueño, episodio) sean una nueva columna. A continuación para los episodios y las actividades, buscaremos en la base de datos el id correspondiente(look up)

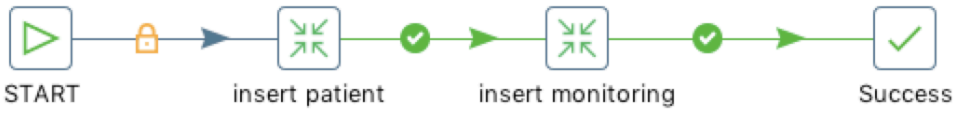
Ya tenemos los datos casi listos. El siguiente paso será hacer un join (previo ordenamiento) y la eliminación de las columnas repetidas.

Los datos están listos, pero necesitamos los ids para poder insertar en la tabla de hechos. Para ello, haremos un look up del paciente y la fecha para encontrar los ids, añadiremos la secuencia y finalmentye, insertaremos.

La transformación resultante se puede encontrar en la figura siguiente:



Finalmente, crearemos un trabajo que ejecute las 2 transformaciones anteriores.



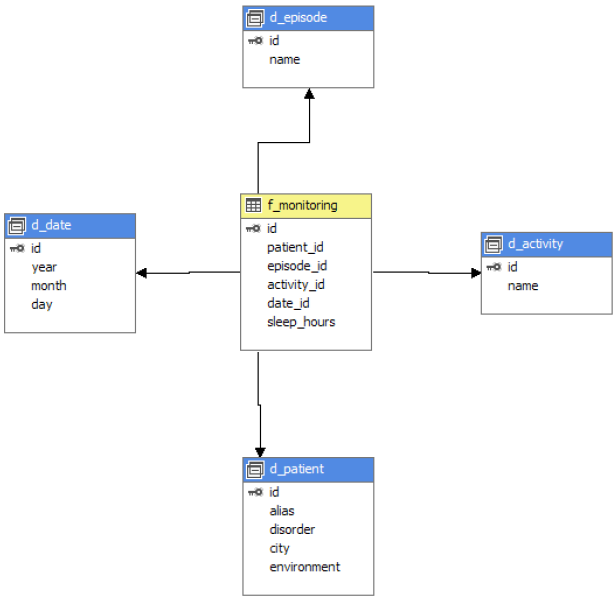
6. OLAP

Una vez insertados los datos en momento de crear el cubo OLAP. Como ya se ha indicado anteriormente, para ello utilizaremos el Analysis Service de Microsoft SQL Server con Visual Studio para modelar los cubos y analizar los datos. Además, nos serviremos de Microsoft Excel para obtener alguna gráfica.

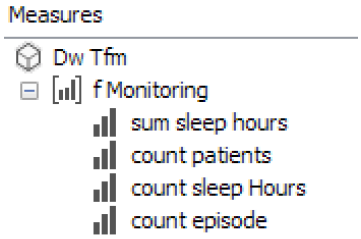
6.1 Cubo OLAP

El cubo OLAP que se ha diseñado y prácticamente igual que el modelo ER que se ha enseñado anteriormente.

Lo primero que se ha hecho es diseñar las dimensiones y sus jerarquías.
Para la dimensión fecha hemos creado una jerarquía año-mes-día, y otra jerarquía mes-día. Para las demás dimensiones no creo que haya jerarquías interesantes.



A continuación, se han creado las diferentes medidas. Tendremos la suma de las horas de sueño y la cuenta de pacientes, episodios y horas de sueño. A mayores se hará una medida calculada llamada horas media de sueño que se calculará a partir de la suma de las horas de sueño y el número de horas de sueño registradas.



6.2 Análisis de resultados

Una vez diseñado el cubo y las medidas, utilizaremos el browser para obtener resultados. Las conclusiones obtenidas son las siguientes:

1. Las actividades afectan significativamente a las horas de sueño en pacientes de episodios graves. Si bien en general la media de horas que duermen los pacientes no varía significativamente(solo hay media horas de diferencia de media entre los que menos duermen y los que más) aquellos que realizan actividades físicas descansan mejor.

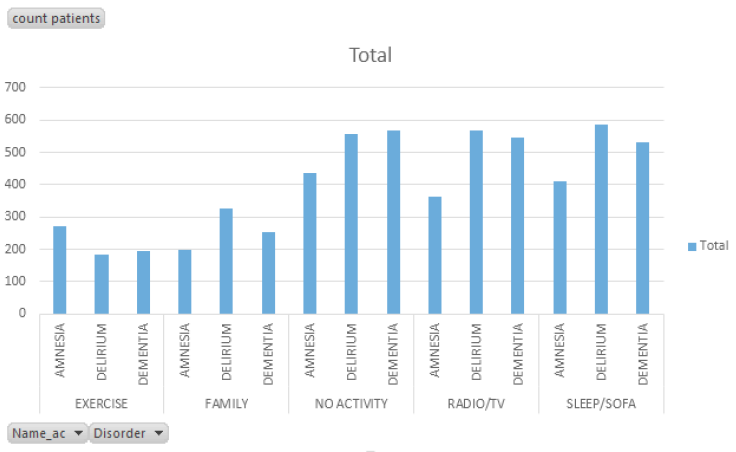
Name_ac	avg sleep hours
EXERCISE	5,4130434782...
FAMILY	5,109375
NO ACTIVITY	4,7749287749...
RADIO/TV	4,7164179104...
SLEEP/SOFA	4,6918604651...

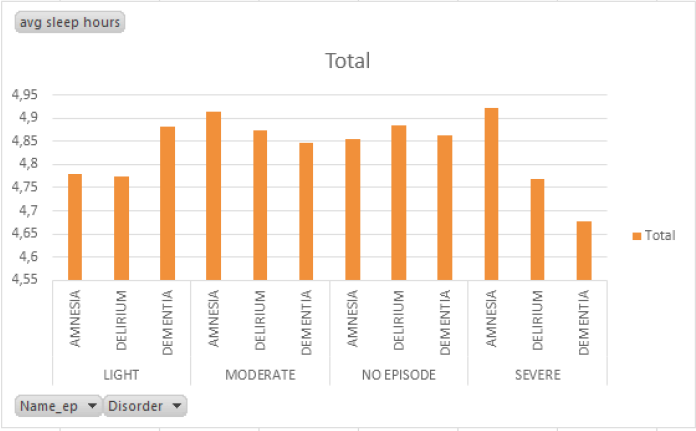
2. No hay una relación clara entre la gravedad del episodio y las horas de sueño. En este caso, a diferencia de las actividades, no parece que haya una gran diferencia en el descanso del paciente dependiendo del tipo de episodio que haya sufrido.

Name_ep	avg sleep hours
LIGHT	4,815296566077
MODERATE	4,8817317845...
NO EPISODE	4,8688193743...
SEVERE	4,7771929824...

3. Es mayor el número de pacientes con enfermedades cognitivas que no realizan ejercicios o no tienen visitas familiares. Además, mayoritariamente tiene delirium o demencia.

Name_ac	Disorder	count patients
EXERCISE	AMNESIA	271
EXERCISE	DELIRIUM	183
EXERCISE	DEMEN...	196
FAMILY	AMNESIA	197
FAMILY	DELIRIUM	325
FAMILY	DEMEN...	253
NO ACTIVITY	AMNESIA	437
NO ACTIVITY	DELIRIUM	558
NO ACTIVITY	DEMEN...	566
RADIO/TV	AMNESIA	363
RADIO/TV	DELIRIUM	568
RADIO/TV	DEMEN...	545
SLEEP/SOFA	AMNESIA	412
SLEEP/SOFA	DELIRIUM	585
SLEEP/SOFA	DEMEN...	532



4. No es inmediato sacar una relación directa entre las horas de sueño y el trastorno cognitivo, si bien los días que no se presentan episodios, los pacientes duermes más tiempo de media.
5. 

6. Los pacientes del ámbito rural tienen menos enfermedades cognitivas y duermen más. Podríamos decir que la calidad de vida en el ámbito rural aumenta.

Environment	sum patients	avg sleep hours
RURAL	1696	5,4958726415...
SEMIURBAN	1804	4,6629711751...
URBAN	2491	4,5126455238...

7. No se muestra una gran mejora ni se encuentra un patrón de los pacientes en el tiempo.
8. La mayor parte de los pacientes son sedentarios o se reúnen poco con sus familiares.
9. Los episodios más frecuentes son los leves y la ausencia de los mismos.
10. El mes con episodios más severos es agosto.
11. La actividad de los pacientes no varía a lo largo del año.