Universitat Oberta de Catalunya

uoc.edu

Sistema de inteligencia de negocio entorno a las enfermedades cognitivas

José Luis Martínez Pérez
Máster Business Inteligence 2015
Área del trabajo final

David Amorós Alcaraz
María Isabel Guitart Hormigo

07/07/2017

Licencias alternativas (elegir alguna de las siguientes y sustituir la de la página anterior)

A) Creative Commons:

Esta obra está sujeta a una licencia de Reconocimiento-NoComercial-SinObraDerivada 3.0 España de Creative Commons

Esta obra está sujeta a una licencia de Reconocimiento-NoComercial-CompartirIgual 3.0 España de Creative Commons

Esta obra está sujeta a una licencia de Reconocimiento-NoComercial 3.0 España de Creative Commons

Esta obra está sujeta a una licencia de Reconocimiento-SinObraDerivada 3.0 España de Creative Commons

Esta obra está sujeta a una licencia de Reconocimiento-CompartirIgual 3.0 España de Creative Commons

Esta obra está sujeta a una licencia de Reconocimiento 3.0 España de Creative Commons

B) GNU Free Documentation License (GNU FDL)

Copyright © AÑO TU-NOMBRE.

Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.3 or any later version published by the Free

C) Copyright

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Sistema de inteligencia de negocio entorno a enfermedades cognitivas*
Nombre del autor:	*José Luis Martínez Pérez*
Nombre del consultor/a:	*David Amorós Alcaraz*
Nombre del PRA:	
Fecha de entrega (mm/aaaa):	07/2017
Titulación::	*Máster Business Intelligence 2015*
Área del Trabajo Final:	*B2.343-MIB-SI*
Idioma del trabajo:	*Castellano*
Palabras clave	*Almacén de datos, inteligencia de negocio, machine learning*

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

Abstract (in English, 250 words or less):

Índice

1. Introducción.....1
1.1 Contexto y justificación del Trabajo.....1
1.2 Objetivos del Trabajo.....1
1.3 Enfoque y método seguido.....1
1.4 Planificación del Trabajo.....1
1.5 Breve sumario de productos obtenidos.....1
1.6 Breve descripción de los otros capítulos de la memoria.....1
2. Resto de capítulos.....2
3. Conclusiones.....3
4. Glosario.....4
5. Bibliografía.....5
6. Anexos.....6

Lista de figuras

No se encuentran elementos de tabla de ilustraciones.

1. Introducción

1.1 Contexto y justificación del Trabajo

Crear un entorno BI que posibilite el análisis de la información generada por pacientes y clínicos dentro del marco de las enfermedades llamadas trastornos cognitivos.

Antecedentes

Los trastornos cognitivos, como su nombre indica, alteran las funciones cognitivas de la persona que los padece como pueden ser la memoria, el lenguaje, la atención, la conducta, el aprendizaje o la orientación. Este tipo de trastornos suele darse en personas mayores, por lo que debemos trabajar para prevenir dicho deterioro cognitivo. Dentro de estos trastornos, podemos encontrarnos con el delirium, la demencia o los trastornos amnésicos que explicaremos a continuación:

Delirium: Se trata del deterioro agudo y global de las funciones superiores. Su dato más característico es el deterioro del nivel de conciencia. Al principio sólo se detectan dificultades de atención, concentración y desorientación (temporal al inicio, luego espacial). Conforme se agrava, se desestructura el pensamiento y la percepción. En el delirium se diferencian dos patrones según la alteración de la conducta: agitado y estuporoso.

Demencia: Es el síndrome caracterizado por el deterioro crónico y global de las denominadas funciones superiores. Lo normal en estos casos es un deterioro intelectual, acompañado de alteraciones de la conducta y del estado de ánimo. Su prevalencia aumenta con la edad (de 65 a 70 años, 2%; >80 años, 20%), siendo la principal causa de incapacidad a largo plazo en la tercera edad. Suele iniciarse con el deterioro de la memoria y cambios de personalidad, sin que el paciente tenga conciencia de sus cambios que con frecuencia niega o disimula. La conducta se vuelve inapropiada y se pierde el interés por las cosas debido en gran parte a fuertes déficits de atención.

Trastornos amnésicos: Es un deterioro específico en la memoria, normalmente de la memoria reciente. Los trastornos amnésicos más típicos son los siguientes:

- Psicosis de Korsakoff: Trastorno de la memoria provocado por la deficiencia de vitamina B1. Afecta sobre todo a la memoria a corto plazo. Los pacientes que presentan este síndrome manifiestan, por norma general dificultad al caminar y con el equilibrio, confusión, somnolencia, parálisis de algunos músculos oculares, neuropatía periférica, etc

- Traumatismos craneoencefálicos: Asociada a la amnesia retrógrada y anterógrada. Ambas se asocian con la intensidad del traumatismo. En él se asocian déficits cognitivos leves (deterioro de la atención o la memoria) con síntomas afectivos (ansiedad, labilidad emocional, tristeza), cambios de personalidad, cansancio, fatiga,cefalea, insomnio, o inestabilidad.

- Amnesia global transitoria: Caracterizada por una pérdida brusca de la memoria reciente, provocándole un estado de desorientación y perplejidad al no poder retener información; el resto de la exploración es normal. El paciente conserva recuerdos lejanos (nombre, lugar de nacimiento); pero es incapaz de recordar cosas recientes a pesar de mantener un buen nivel de atención; es característico que el paciente repita de forma insistente la misma pregunta.

En el marco de un proyecto que intenta desarrollar terapias que permitan entender y controlar estas enfermedades, se ha hecho un estudio sobre veinte pacientes afectados por estos trastornos. Este estudio intenta relacionar los estados de ánimo y las actividades realizadas con la aparición de crisis agudas o empeoramientos temporales de los síntomas asociados a estas enfermedades. Poder extraer conclusiones sobre esta relación podría ayudar en la mejora de las condiciones de vida de enfermos con estos trastornos.

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 este tipo de enfermedades provenientes de diferentes centros médicos.

Los objetivos específicos del trabajo 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.

5. Opcionalmente, crear un módulo de Machine Learning sobre el DW para poder prevenir cambios de tendencia o empeoramiento en la enfermedad. Este módulo podría implementar lo que se conoce como “early warnings”. También podría, a partir de los datos del paciente, poder intentar hacer una pre-clasificación de los síntomas en una de las enfermedades que se estudian en el juego de datos

Preguntas analíticas

Las preguntas que se quieren resolver son las siguientes:

- ¿Cuál es la relación entre las actividades realizadas y los episodios de crisis graves?
- ¿Se puede establecer algún tipo de relación entre los valores de los diferentes estados de ánimo y los episodios de crisis?
- Estas relaciones son iguales para cualquiera de las enfermedades o en cambio hay relaciones más acusadas por alguna de ellas.
- ¿Se puede establecer alguna relación en nivel geográfico, por ejemplo entorno urbano o rural?
- ¿Cuál sido la evolución de los diferentes pacientes a lo largo del tiempo?
- ¿Se puede establecer alguna relación entre los episodios de crisis y el momento del día o de la semana o del año?
- La realización de actividades físicas mejora o empeora el estado de ánimo de los pacientes.
- ¿Hay algún tipo de actividad que mejore el día a día de los pacientes?

1.3 Enfoque y método seguido

El objetivo es desarrollar un producto nuevo y para ello se va a seguir una metodología estandar de gestión de proyectos BI, basada en PMBoK. En esta metodología de gestión de proyectos, los actores e interesados principales, las fases del proyecto y las responsabilidades quedan muy bien definidas y así evitamos problemas de gestión en el desarrollo del proyecto, lo que añadiría aún más probabilidad de fracaso a un proyecto BI.
Siguiendo esta metodología planificaremos el proyecto en diversas fases que describiremos en los puntos siguientes y tales fases comprenderán tareas de estudio del estado del arte, análisis y de desarrollo e implementación.

1.4 Planificación del Trabajo

El proyecto se ha dividido en varias fases que comprenden en total 15 semanas de duración, desde la primera semana de abril de 2017 hasta la primera semana de julio de 2017 que es la fecha límite para entregar y finalizar el proyecto.
A continuación semuestra un diagrama de Gantt con la planificación por fases del proyecto:

Tarea	Abril				Mayo					Junio					Julio
	Semana 1	Semana 2	Semana 3	Semana 4	Semana 5	Semana 6	Semana 7	Semana 8	Semana 9	Semana 10	Semana 11	Semana 12	Semana 13	Semana 14	Semana 1
Estudio del proyecto	X														
Análisis de los ficheros fuentes de datos	X	X	X												
Análisis de soluciones BI para el análisis de datos			X	X	H1										
Diseño del almacén de datos				X	X	X	H2								
Implementación del almacén de datos				X	X	X	X	X	H3						
Diseño de los procedimientos de transformación de datos						X	X								
Implementación de los procedimientos de transformación de datos							X	X	X	H4					
Implementación de la capa de análisis de datos										X	X	H5			
Diseño e implementación del módulo de machine learning												X	X	H6	
Redacción de memoria final															H7

1.5 Breve sumario de productos obtenidos

Hito	Descripción
H1: Documento con el análisis de las fuentes de datos y de herramientas de análisis de datos.	Se presentará un informe con las fuentes de datos y el significado de cada uno de los campos del que se hará uso en el proyecto. Además el documento contendrá una sección donde se analizarán las herramientas existentes para el tratamiento de dichos datos.
H2: Diseño del almacén de datos.	Se ofrecerá un documento con el diseño de la arquitectura y los componentes princippales dle almacén de datos.
H3:Implementación del almacén de datos.	El resultado será un documento explicativo con las fases de la implementación del DW y los problemas más relevantes que han surgido durante la misma.
H4: Diseño e implementación de los procesos de transformación de datos.	Se obtendrá un informe con el resultado de la ejecución de los procesos de integración de datos y sus pruebas correspondientes.
H5: Implementación de la capa de análisis de datos.	El presentará un documento con la definicón de cada proceso de visualización de datos y la respuesta a la pregunta analítica correspondiente.
H6: Diseño e implementación del módulo de machine learning.	Se porporcionará un documento con el diseño y su implementación correspondiente del módulo de machine learning para previr cambios de tendencias en las enfermedades cognitivas.
H7: Entrega de la memoria final.	El resultado es el documento del proyecto de final de máster.

Además, en la tabla siguiente se hace una correspondencia entre hitos y entrega de PECs.

PEC	Hito	Descripción
PEC1	X	Documento inicial donde se define el proyecto.
PEC2 08/05/2017	H1	Entrega con el documento de análisis de los datos disposibles y un estudio

		de las herramientas para el tratamiento de dichos datos en etapas posteriores del proyecto.
PEC3 05/06/2017	H4	Implementación de los procesos de transformación de datos. Queda la parte de visualización y el modulo de machine learning para la última fase del proyecto.
PEC Final 07/07/2017	H7	Entrega de la memoria final junto con los ficheros de código fuente y ficheros relevantes en la ejecución del proyecto.

1.6 Breve descripción de los otros capítulos de la memoria

En el primer capítulo se presentará el proyecto de una forma general, definiendo la arquitectura general de la solución y sus principales componentes.

La definición y análisis de las fuentes de datos se expondrá en el segundo capítulo de la memoria junto con el estudio de las soluciones disponibles de BI para la capa de tratamiento, transformación y visualización de datos.

El tercer capítulo es un capítulo con contenido más técnico y abarca el diseño del almacén de datos, sus componentes, tablas maestras, tablas de hechos, etc.

El procesamiento de las fuentes de datos se verá en prfuncidad en el capítulo custro del proyecto. En él se desarrollarán las tareas de extracción y procesamiento de datos y se realizarán las pruebas pertinentes para su integración en el almacén de datos.

La implementación de la capa de análisis de datos se expondrá en el capítulo quinto del documento.

En el sexto capítulo se desarrollará el módulo de aprendizaje automático «machine learning» que será capaz de prevenircambios en las tendencias o empeoramientos de las enfermedades cognitivas.

Por último, el documento finalizará con un capítulo dedicado a conclusiones extraídas durante la ejecución del proyecto.

2. Resto de capítulos

2.1. Análisis de las fuentes de datos y selección de herramientas software.

En este capítulo, tal y como queda establecido en la planificación del proyecto se va a realizar por una parte un estudio de las fuentes de datos y cómo se relacionancon las preguntas analíticas que queremos resolver, y por otra parte, hay que seleccionar el software necesario para llevar a cabo su implementación. Por último se expondrá el diseño de la arquitectura de la solución.

2.1.1. Análisis de las fuentes de datos.

Para el desarrollo del proyecto es necesario obtener datos de los 20 pacentes que son estudiados y que reciben un seguimiento de su enfermedad. El seguimiento que se realiza tiene en cuenta factores como el ejercicio físico, las horas de sueño y cuando se producen los episodios de la enfermedad.
Los datos vienen dados en un fichero excel que contiene 4 hojas de datos. A continuación se describen los campos de datos de cada hoja proporcionada.

Hoja 1: Pacientes.

Campo	Descripción
ID	Identificador de paciente (P1,P2..Pn)
Desorden cognitivo	Nombre de la enfermedad que padece
Ciudad	Ciudad de origen del paciente
Entorno	Rural / Urbano / Semiurbano

Hoja 2: Horas de sueño.

Campo	Descripción
Fecha	Fecha de registro del dato
Horas / Paciente	Horas de sueño registradas por paciente.

Hoja 3: Actividades

Campo	Descripción
Fecha	Fecha de registro del dato
Actividad / Paciente	Actividad que realiza el paciente

	cuando sufrió síntomas de la enfermedad.

Hoja 4: Episodios

Campo	Descripción
Fecha	Fecha de registro del dato
Episodio / Paciente	Gravedad del episodio registrado. Toma los valores 'NO EPISODE, LIGHT, MODERATE, SEVERE'

2.1.3. Relación de los datos con las preguntas analíticas a resolver.

Respecto a las preguntas analíticas iniciales definidas en los objetivos del proyecto hay que hacer una rectificación sobre ellas y eliminar las que tienen que ver con el estado de ánimo del paciente, ya que no se disponen de datos que muestre esta información.
Una vez eliminadas estas preguntas, en la siguiente tabla veremos la información que tenemos en el excel y como podemos relacionarla para poder responder a las preguntas analíticas que propone el enunciado del proyecto.

Pregunta	Relación de datos del excel
¿Cuál es la relación entre las actividades realizadas y los episodios de crisis graves?	Cruzaremos los datos de actividades realizadas por pacientes con los registros de episodios graves de los pacientes.
¿Se puede establecer algún tipo de relación entre los valores de los diferentes estados de ánimo y los episodios de crisis?	De nuevo, cruzaremos los datos de actividades realizadas por pacientes con los registros de episodios de los pacientes. En este caso se tendrá en cuenta cualquier valor de la gravedad del episodio (LIGHT, MODERATE, SEVERE).
Estas relaciones son iguales para cualquiera de las enfermedades o en cambio hay relaciones más acusadas por alguna de ellas.	Con los datos obtenidos en las consultas anteriores se intentará responder a esta pregunta.
¿Se puede establecer alguna relación en nivel geográfico, por ejemplo entorno urbano o rural?	Habrá que cruzar los datos geográficos del paciente con los datos de los episodios registrados por cada paciente.
¿Cuál sido la evolución de los diferentes pacientes a lo largo del tiempo?	En este caso, se necesita realizar un análisis temporal de los registros de los episodios registrados de cada paciente.

¿Se puede establecer alguna relación entre los episodios de crisis y el momento del día o de la semana o del año?	Se analizarán los episodios registrados en función del día, semana o año.
¿Hay algún tipo de actividad que mejore el día a día de los pacientes?	Para resolver esta pregunta hay que agrupar por tipo de actividad y ver la evolución de cada paciente.

Estas relaciones se definirán completamente durante la fase de diseño del almacén de datos más adelante en el proyecto.

2.1.3. Análisis de las soluciones software para el tratamiento, transformación, análisis y visualización de los datos.

Para la creación del sistema completo de BI se va a necesitar software que cubra todas y cada una de las fases de un proyecto de construcción de un almacén de datos:

- Adquisición y transformación de datos.
- Almacenamiento de datos.
- Análisis de datos OLAP.
- Consulta y visualización de datos.

A continuación se expone un estudio de herramientas BI y se escoge una de ellas en función de las necesidades del proyecto y de las ventajas e inconvenientes de cada una de ellas.

Almacenamiento de datos.

Para la definición e implementación del almacén de datos se necesita un sistema gestor de bases de datos que guardará la información de las tablas de dimensiones y de las tablas de hechos.
En este apartado se da respuesta a las incognitas de si utilizar un sistema relacional SQL o NoSQL y en función de esta elección qué alternativa de sistema gestor de base de datos se selecciona. Como requisito indispensable para esta selección se tiene que es necesario que sea software libre y que se pueda instalar en Ubuntu.

Lo primero que se va a decidir es si se debe utilizar un sistema SQL o NoSQL. Entre las alternativas SQL de software libre destacan MySQL y PostgreSQL, mientras que si se opta por un sistema NoSQL el software que se postula como principal candidato es MongoDB.
Para decidir este apartado, se va tener en cuenta los datos disponibles que necesitamos almacenar y el tipo de esquema que requieren para su almecenamiento.
Los datos provienen de ficheros excel las columnas de cada uno de ellos están bien definidas, no dejando lugar a ambigüedades ni a posibilidad de flexibilidad en el esquema de datos. Los sistemas NoSQL permiten esquemas de datos mucho más flexibles que los relacionales SQL pero en este caso, al no necesitar de esta característica, se ha optado por un sistema relacional SQL tradicional.

Una vez decidio el tipo de sistema, se analizan las características de las dos principales alternativas, MySQL y PostgreSQL.
La elección entre estos sistemas es complicado, ya que los dos son utilizados ampliamente en el desarrollo de aplicaciones actualmente. A continuación se lista una comparativa a muy alto nivel entre ellos.

PostgresSQL:
- Cumple el standard SQL.
- Soporte completo para transacciones ACID.
- Menos popular que MySQL.
- Más difícil de comfigurar.
- Comunidad de soporte menos amplia.

MySQL:
- No cumple el standard SQL en su totalidad.
- Fácil de aprender y utilizar.
- Oracle es su desarrollador.
- El sistema más popular actualmente para el desarrollo de aplicaciones.

Para la realización del proyecto, se considera que cualquiera de las dos opciones serían más que válidas, pero para poder seleccionar una de ellas, se ha decidido que el cumplimiento del standard SQL es un argumento determinante para su elección, es por esto, que se selecciona PostgreSQL como sistema de gestión de bases de datos para la implementación del almacén de datos o data warehouse.

Herramientas BI.

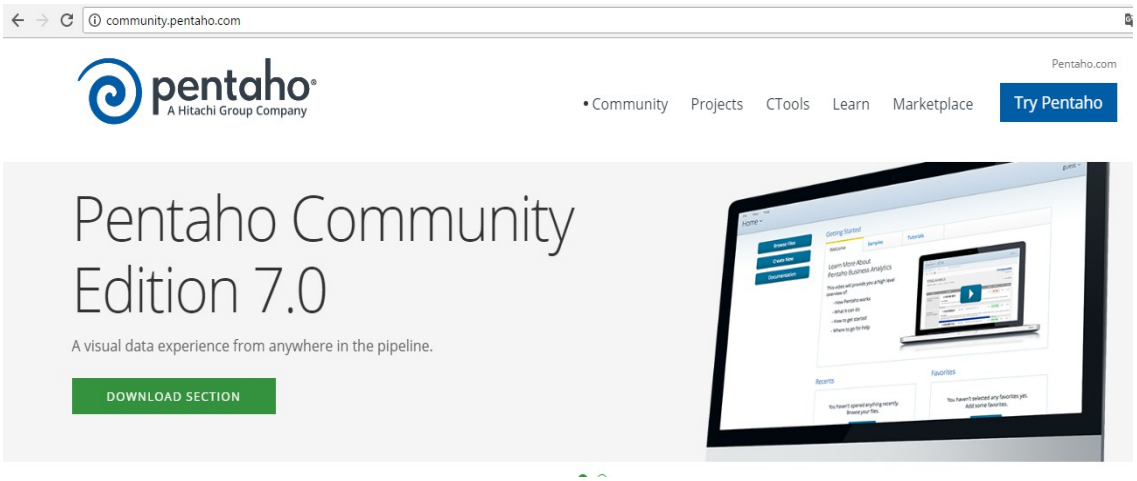
Actualmente existen muchas herramientas en el mercado de Business Intelligence. Entre las herramientas analizadas, de nuevo, el requisito primordial es que sean software libre y que se puedan instalar en Linux (concretamente en Ubuntu 16.04).

Los datos vienen dados en forma de ficheros excel, por lo tanto la carga, el acceso y el tratamiento de ficheros excel es requisito indispensable para la selección de la herramienta ETL. Además deben poder almacenar la información extraída del fichero en un SGBD PostgreSQL.

Las herramientas que vamos a analizar son:

- Microsoft SQL Server.
- IBM InfoSphere Information Server.
- Pentaho 7.0 community edition.
- SpagoBI

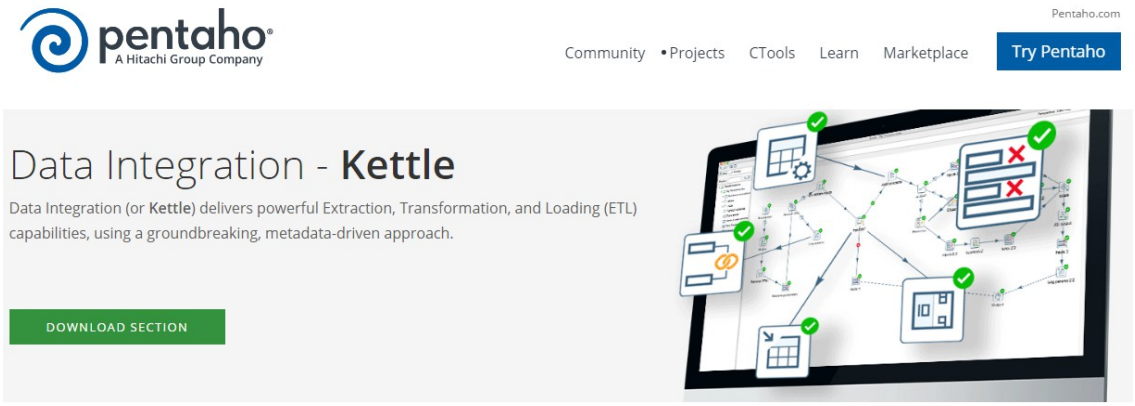
En una selección inicial, se descartan las herramientas de pago y las que funcionan únicamente en Windows. Por lo tanto, únicamente quedan en el análisis SpagoBI y Pentaho.



En ambos casos veremos el soporte que ofrecen ambas para los procesos de transformación de datos, para el análisis dinámico de datos (OLAP) y para la visualización de los datos.

Adquisición y transformación de datos.

Pentaho proporciona la herramienta Pentaho Data Integration para la definición de procesos de transformación de datos (ETL). SpagoBI, utiliza un módulo incluído en la instalación de la solución completa.

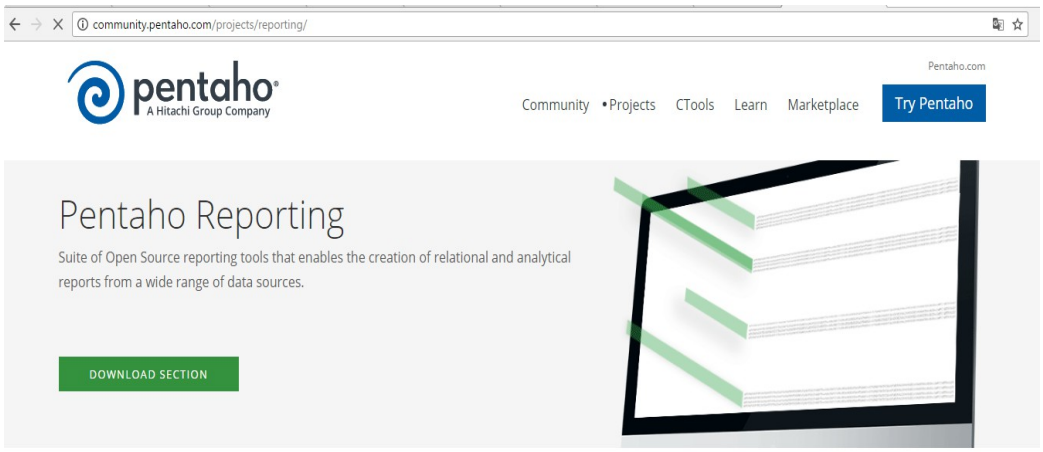


Análisis de datos OLAP.

Tanto Pentaho como SpagoBI cuentan con herramientas para soporte de análisis OLAP. En el caso de Pentaho, su producto Business Server incorpora el motor Mondrian, el cuál, además de realizar las tareas de análisis, también se puede instalar independientemente para usuarios avanzados o desarrolladores, mientras que en SpagoBi viene integrado como módulo del producto.

Consulta y visualización de datos.

Pentaho define y realiza los trabajos de visualización de datos a través de su herramienta Pentaho Report Designer, mientras que en SpagoBI utiliza su módulo integrado en la instalación completa del producto.



A la hora de seleccionar entre una de estas dos herramientas los criterios que han prevalecido han sido por una parte la facilidad en la instalación y el conocimiento previo de la herramienta y por otra parte la comunidad de soporte que tiene cada una de ellas. En cuanto al conocimiento de la herramienta, Pentaho se ha utlilizado en alguna asignatura del máster para la definición de procesos ETL, mientras que SpagoBI es una herramienta totalmente nueva. Por otra parte, en cuanto a soporte, Pentaho tiene un foro en su propia página (http://forums.pentaho.com/), SpagoBI, por su parte, no tiene foro en su propia web, aunque sí que existe alguna página extra oficial del producto con foros sobre su uso (https://www.spagoworld.org/jforum/forums/list.page).

Sin más motivos de selección que los expuestos anteriormente, la herramienta BI seleccionada es **Pentaho 7.0 Community Edition**.

2.1.4. Resumen de herramientas seleccionadas.

Después de analizar las diferentes herramientas expuestas en las secciones anteriores, se ha decidido optar por las siguientes:

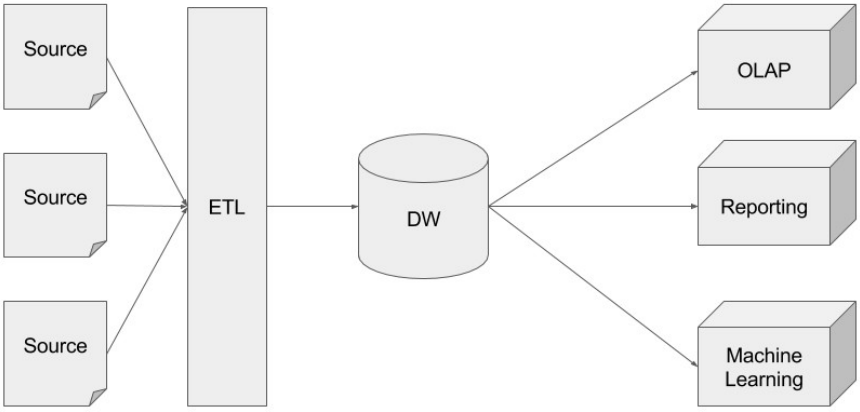
- PostgreSQL para la implementación del almacén de datos.
- Pentaho 7.0 Community Edition.

2.1.5. Arquitectura de la solución.

El sistema de DW se va a construir bajo los siguientes requisitos software:

- Sistema operativo base Windows 10.
- Virtual Box 5.1.6r
- Máquina Virtual 64 bits 4 GB de RAM.
- Sistema Operativo Ubuntu 16.04.2-LTS 64 bits.

La siguiente figura muestra el diseño general de la arquitectura del almacén de datos.



Referencias para la elección de las herramientas.
Se enumeran a continuación las referencias consultadas para la selección de las herramientas de los apartados anteriores.

- https://www.digitalocean.com/community/tutorials/sqlite-vs-mysql-vs-postgresql-a-comparison-of-relational-database-management-systems
- http://community.pentaho.com/projects/data-integration/
- http://wiki.pentaho.com/display/EAI/Excel+Input+Step
- http://forums.pentaho.com/showthread.php?131598-Using-Postgres-with-Pentaho-ETL
- https://en.wikipedia.org/wiki/SQL_Server_Integration_Services
- https://docs.microsoft.com/en-us/sql/integration-services/sql-server-integration-services
- https://docs.microsoft.com/en-us/sql/integration-services/connection-manager/connect-to-an-excel-workbook
- https://opensource.com/business/16/6/top-business-intelligence-reporting-tools
- http://www.datasciencecentral.com/profiles/blogs/10-open-source-etl-tools
- http://community.pentaho.com/projects/mondrian/
- http://mondrian.pentaho.com/documentation/olap.php
- http://forums.pentaho.com/showthread.php?77325-Absolute-begginer-s-Mondrian-tutorial
- http://forums.pentaho.com/showthread.php?83030-Create-OLAP-Cube-in-Pentaho
- http://mondrian.pentaho.com/documentation/workbench.php
- https://technet.microsoft.com/en-us/library/hh916543(v=sc.12).aspx
- https://technet.microsoft.com/en-us/library/ee677579.aspx
- https://www.youtube.com/watch?v=ctUiHZHr-5M
- http://www.spagobi.org/

- https://www.ibm.com/analytics/us/en/technology/information-server/
- https://www-01.ibm.com/software/

2.2. Diseño e implementación del almacén de datos.

En este capítulo, tal y como queda establecido en la planificación del proyecto se va a realizar por una parte el diseño del almacén de datos y por otra parte se implementarán los procesos para rellenar las tablas de dimensiones y la/s de hecho/s que se definirán partiendo de las fuentes de datos y las preguntas analíticas que queremos resolver.

2.2.1. Estrategia de construcción del almacén de datos.

La estrategia que se va a seguir para la implementación del almacén de datos es la de realizar un desarrollo incremental con un único DataMart y sin área de staging. Esto es así porque actualmente sólo existe la necesidad de acceder a los datos por parte de un departamento, el departamento médico, por lo que no hacen falta más data marts para implentar informes para otros tipos de usuarios. Además, si se dispusiera de varias fuentes de datos, el desarrollo de un área de staging sería aconsejable para consolidar datos, pero en este caso, no se considera necesario.

2.2.2. Diseño conceptual.

A continuación mostramos un ejemplo de cada una de las hojas de datos que se describieron y analizando en secciones anteriores del documento. Partiendo estos ejemplos se definirán las dimensiones y las tablas de hechos posteriormente en los siguientes párrafos.

Ejemplo de registro hoja 1: Pacientes

PATIENT	COGNITIVE DISORDER	CITY	ENVIRONME
P1	DELIRIUM	BARCELONA	URBAN
P2	DELIRIUM	MONTORO	SEMIURBAN

Ejemplo de registro hoja 2: Horas de sueño

FECHA	P1	P2	P3	P4	P5
01/01/2016	4	5	5	4	6

Ejemplo de registro hoja 3: Actividades

FECHA	P1	P2	P3	P4	P5
01/01/2016	NO ACTIVITY	RADIO/TV	RADIO/TV	NO ACTIVITY	FAMILY
02/01/2016	EXERCISE	NO ACTIVITY	READ/STUDY	NO ACTIVITY	NO ACTIVITY

Ejemplo de registro hoja 3: Episodios

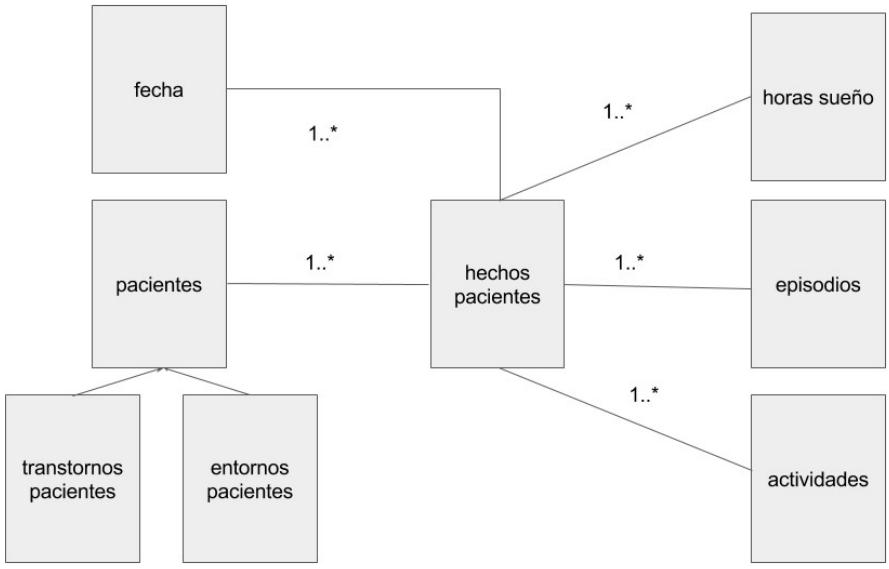
FECHA	P1	P2	P3	P4	P5
01/01/2016	LIGHT	SEVERE	NO EPISODE	NO EPISODE	NO EPISODE
02/01/2016	NO EPISODE	SEVERE	LIGHT	LIGHT	NO EPISODE

Con estos ejemplos obtenidos de las fuentes de datos y teniendo en cuenta las preguntas analíticas, se ha concluído que:

1. Se necesita una tabla de hechos que almacene la información necesaria para obtener informes que respodan a las preguntas analíticas definidas para este proyecto.
2. Se han identificado 5 tablas de dimensiones:
    1. dim_fecha, con información temporal para cada una de las entradas de los ficheros de datos.
    2. dim_paciente, que guardará los datos de identificación de cada paciente.
    3. dim_horas_sueño, que almacenará los valores de las horas de sueño realizadas por el paciente.
    4. dim_episodio, proporcionará datos sobre los episodios sufridos por cada paciente.
    5. dim_actividad, que guardará la información relativa a la actividad que realiza cada paciente.

Respecto a la dimensión paciente, se opta por definir una jerarquía para separar los transtornos del paciente del entorno en el que vive el paciente.

A continuación se muestra una imagen con el diseño conceptual definido en párrafos anteriores:



2.2.3. Diseño lógico.

Una vez definido el modelo conceptual, el siguiente paso en el modelado del almacén de datos es el modelo lógico. En este modelo definiremos, en un paso previo al diseño físico, los campos y tipos de datos que contendrá cada tabla de nuestro almacén de datos.

Primero definiremos las tablas de dimensiones.

dim_fecha

Nombre atributo	Tipo	nulo
id_fecha	Serial (autoincremental)	NO
Anyo	Int	NO
Mes	Int	NO
Dia	Int	NO
Fecha	String (dd/mm/yyyy)	NO

dim_paciente

Nombre atributo	Tipo	nulo
id_paciente	Serial (autoincremental)	NO
Nombre	String	NO
Ciudad	String	NO

dim_transtorno_paciente

Nombre atributo	Tipo	nulo
id_paciente	Serial (autoincremental)	NO
Transtorno	String	NO

dim_entorno_paciente

Nombre atributo	Tipo	nulo
id_paciente	Serial (autoincremental)	NO
Entorno	String	NO

dim_horas_sueño

Nombre atributo	Tipo	nulo
id_horas	Serial (autoincremental)	NO
num_horas	Int	NO

identificador_pac	Int	NO
id_fecha	Int	NO

dim_actividades

Nombre atributo	Tipo	nulo
id_actividad	Serial (autoincremental)	NO
Actividad	String	NO
identificador_pac	Int	NO
id_fecha	Int	NO

dim_episodios

Nombre atributo	Tipo	nulo
id_episodio	Serial (autoincremental)	NO
Episodio	String	NO
identificador_pac	Int	NO
id_fecha	Int	NO

La tabla de hechos tiene los siguientes atributos:

Nombre atributo	Tipo	nulo
id_hecho	Serial (autoincremental)	NO
id_fecha	Int	NO
id_paciente	Int	NO
Variable	String	NO
Valor	String	NO

2.2.4. Diseño físico.

En este punto de la memoria, se presenta el diseño físico del almacén de datos. El diseño físico es la implementación en un sistema gestor de bases de datos del modelo lógico expuesto anteriormente. El diseño físico se ha implementado bajo un sistema PostgreSQL 9.6 con la ayuda de PgAdmin 3, una herramienta gráfica para la gestión de postgresql.

Además el fichero dw_schema.sql se adjuntará como parte de este trabajo.

Las sentencias de creación de las tablas quedan descritas en las siguientes lineas.

dim_paciente

```
SQL pane
-- Table: public.dim_paciente

-- DROP TABLE public.dim_paciente;

CREATE TABLE public.dim_paciente
(
    id_paciente integer NOT NULL DEFAULT nextval('dim_paciente_id_paciente_seq'::regclass),
    ciudad character varying(50) NOT NULL,
    identificador character varying(10),
    CONSTRAINT id_paciente_pk PRIMARY KEY (id_paciente)
)
WITH (
    OIDS=FALSE
);
ALTER TABLE public.dim_paciente
    OWNER TO jlmartinez;
```

dim_entornos_paciente

```
SQL pane
-- Table: public.dim_entornos_paciente

-- DROP TABLE public.dim_entornos_paciente;

CREATE TABLE public.dim_entornos_paciente
(
-- Inherited from table dim_paciente:  id_paciente integer NOT NULL DEFAULT nextval('dim_paciente_id_paciente_seq'::regclass),
-- Inherited from table dim_paciente:  ciudad character varying(50) NOT NULL,
    entorno character varying(50),
-- Inherited from table dim_paciente:  identificador character varying(10),
    CONSTRAINT dim_entornos_paciente_pkey PRIMARY KEY (id_paciente)
)
INHERITS (public.dim_paciente)
WITH (
    OIDS=FALSE
);
ALTER TABLE public.dim_entornos_paciente
    OWNER TO jlmartinez;
```

dim_transtornos_paciente

SQL pane

```
-- Table: public.dim_transtornos_paciente

-- DROP TABLE public.dim_transtornos_paciente;

CREATE TABLE public.dim_transtornos_paciente
(
-- Inherited from table dim_paciente:  id_paciente integer NOT NULL DEFAULT nextval('dim_paciente_id_paciente_seq'::regclass),
-- Inherited from table dim_paciente:  ciudad character varying(50) NOT NULL,
  transtorno character varying(50),
-- Inherited from table dim_paciente:  identificador character varying(10),
  CONSTRAINT dim_transtornos_paciente_pkey PRIMARY KEY (id_paciente)
)
INHERITS (public.dim_paciente)
WITH (
  OIDS=FALSE
);
ALTER TABLE public.dim_transtornos_paciente
  OWNER TO jlmartinez;
```

dim_fecha

SQL pane

```
-- Table: public.dim_fecha

-- DROP TABLE public.dim_fecha;

CREATE TABLE public.dim_fecha
(
  id_fecha integer NOT NULL DEFAULT nextval('dim_fecha_id_fecha_seq'::regclass),
  anyo integer NOT NULL,
  mes integer,
  dia integer,
  fecha integer NOT NULL,
  fecha_orig character varying(20) NOT NULL,
  CONSTRAINT id_fecha_pk PRIMARY KEY (id_fecha)
)
WITH (
  OIDS=FALSE
);
ALTER TABLE public.dim_fecha
  OWNER TO jlmartinez;
```

dim_actividades

SQL pane

```sql
-- Table: public.dim_actividades

-- DROP TABLE public.dim_actividades;

CREATE TABLE public.dim_actividades
(
  identificador_pac character varying(50),
  id_actividad integer NOT NULL DEFAULT nextval('dim_actividades_id_actividad_seq'::regclass),
  actividad character varying(50) NOT NULL,
  id_fecha integer NOT NULL,
  CONSTRAINT dim_actividad PRIMARY KEY (id_actividad)
)
WITH (
  OIDS=FALSE
);
ALTER TABLE public.dim_actividades
  OWNER TO jlmartinez;
```

dim_horas_sueño

SQL pane

```sql
-- Table: public.dim_horas_suenyo

-- DROP TABLE public.dim_horas_suenyo;

CREATE TABLE public.dim_horas_suenyo
(
  id_horas_suenyo integer NOT NULL DEFAULT nextval('dim_horas_suenyo_id_horas_suenyo_seq'::regclass),
  horas integer,
  id_fecha integer NOT NULL,
  identificador_pac character varying(50) NOT NULL,
  CONSTRAINT id_horas_suenyo_pk PRIMARY KEY (id_horas_suenyo)
)
WITH (
  OIDS=FALSE
);
ALTER TABLE public.dim_horas_suenyo
  OWNER TO jlmartinez;
```

dim_episodios

SQL pane

```sql
-- Table: public.dim_episodios

-- DROP TABLE public.dim_episodios;

CREATE TABLE public.dim_episodios
(
  id_episodio integer NOT NULL DEFAULT nextval('dim_episodios_id_episodio_seq'::regclass),
  episodio character varying(25),
  identificador_pac character varying(50),
  id_fecha integer NOT NULL,
  CONSTRAINT dim_episodio_pk PRIMARY KEY (id_episodio)
)
WITH (
  OIDS=FALSE
);
ALTER TABLE public.dim_episodios
  OWNER TO jlmartinez;
```

hechos_paciente

SQL pane

```
-- Table: public.hechos_paciente

-- DROP TABLE public.hechos_paciente;

CREATE TABLE public.hechos_paciente
(
    id_hecho integer NOT NULL DEFAULT nextval('hechos_paciente_id_hecho_seq'::regclass),
    id_paciente integer,
    id_fecha integer,
    variable character varying(50),
    valor character varying,
    CONSTRAINT hechos_pk PRIMARY KEY (id_hecho)
)
WITH (
    OIDS=FALSE
);
ALTER TABLE public.hechos_paciente
    OWNER TO jlmartinez;
```

2.3. Diseño e implementación de los procesos de transformación de datos.

En este capítulo, tal y como queda establecido en la planificación del proyecto se va a realizar por una parte el diseño del almacén de datos y por otra parte se implementarán los procesos para rellenar las tablas de dimensiones y la/s de hecho/s.

2.3.1. Proceso general de transformación de datos.

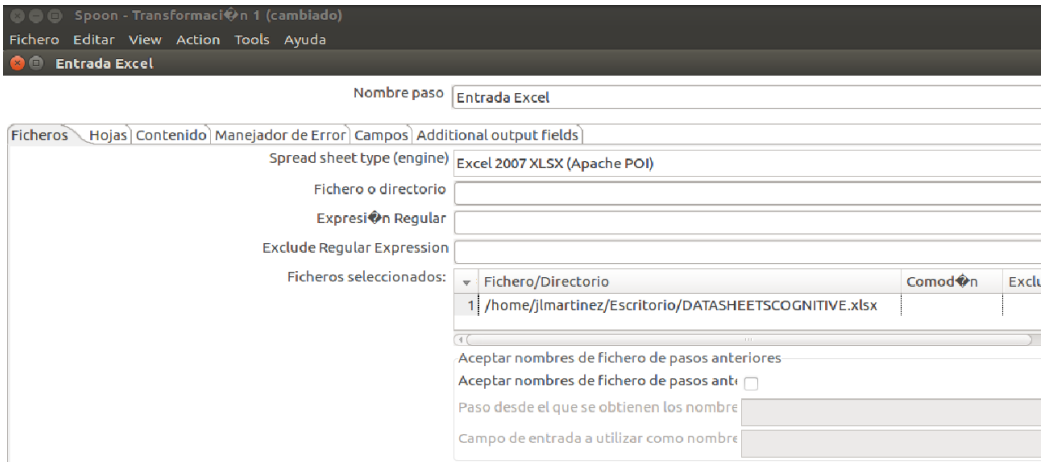
Como paso previo a la implementación de los procesos ETL se ha realizado un estudio de los ficheros excel que se han proporcionado como datos de entrada para la ejecución del proyecto. En este análisis previo se identifican tareas de renombramiento de columnas de los ficheros si fuera necesario, unificación o corrección de algunos registros, etc.
En este caso, no ha sido necesario tal modificación de los ficheros originales.

En las siguientes lineas se describen los procesos ETL utilizados en el proyecto.

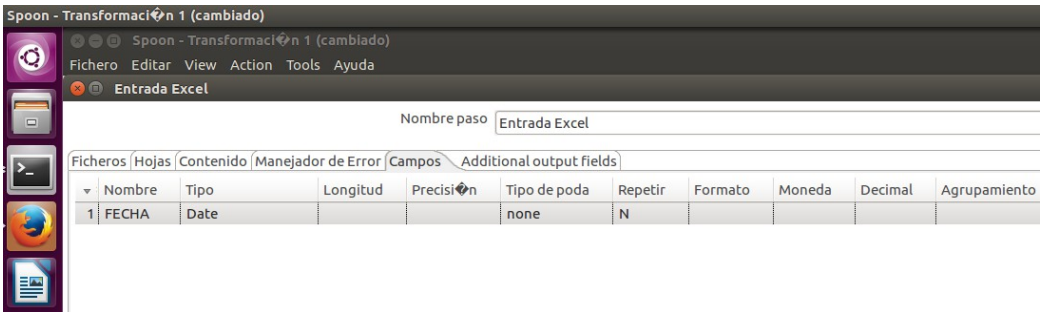
ETL fecha

Este proceso tomará como entrada el fichero excel, y cualquier hoja que contenga la columna fecha. Tal columna contiene un registro por cada día del año 2016, que será insertado en la dim_fecha.

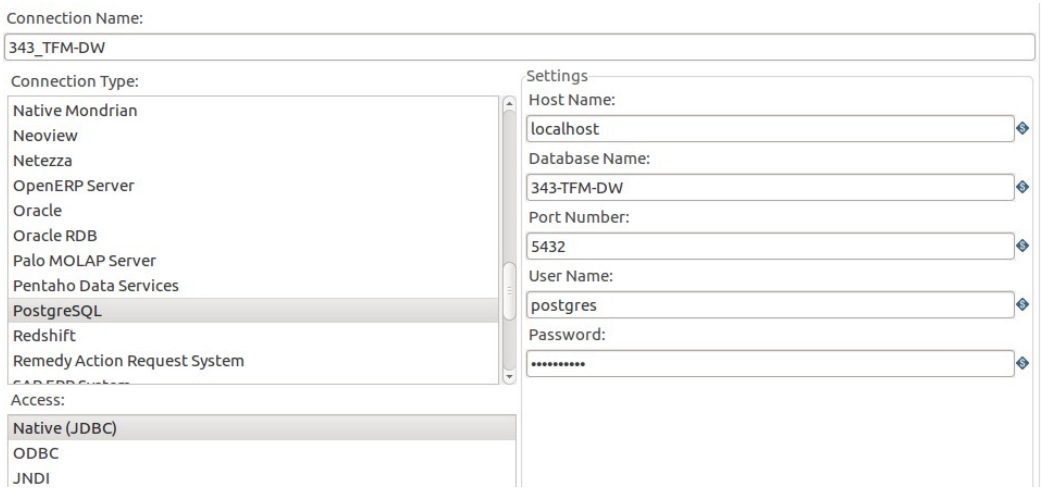
Configuramos la entrada de datos desde el fichero .xlsx

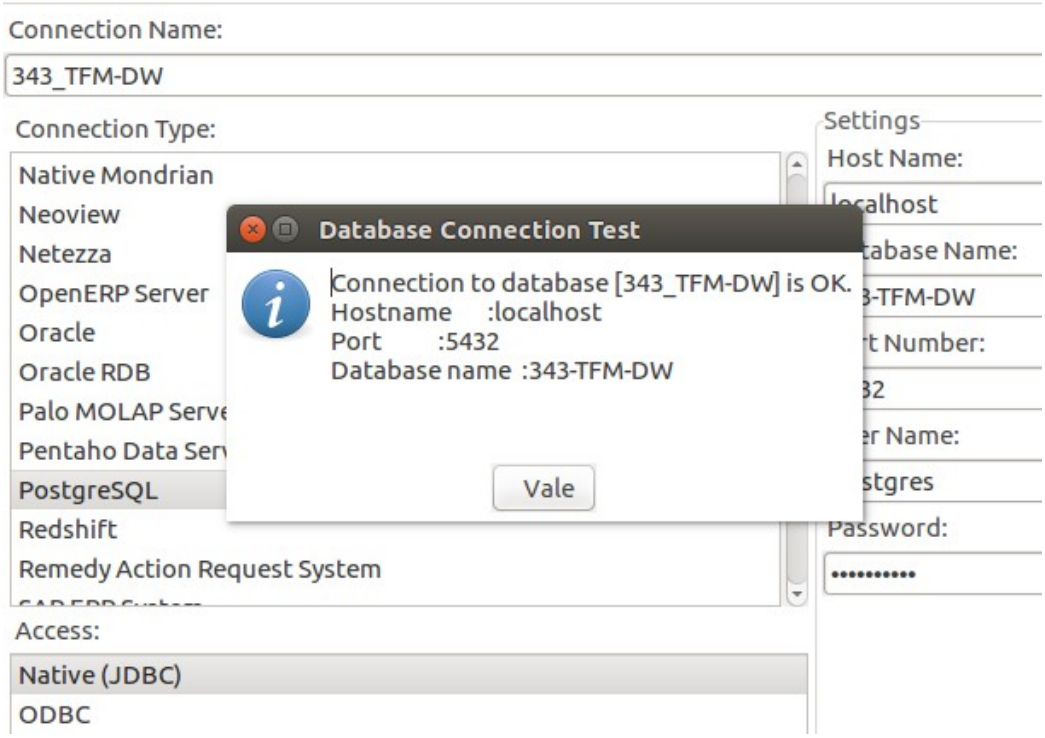


Y definimos las columnas que queremos obtener. En este caso únicamente la columna fecha.



Definimos un paso de salida para el volcado de datos de la transformación a la BBDD postgreSQL. Para ello hay que configurar un conector:

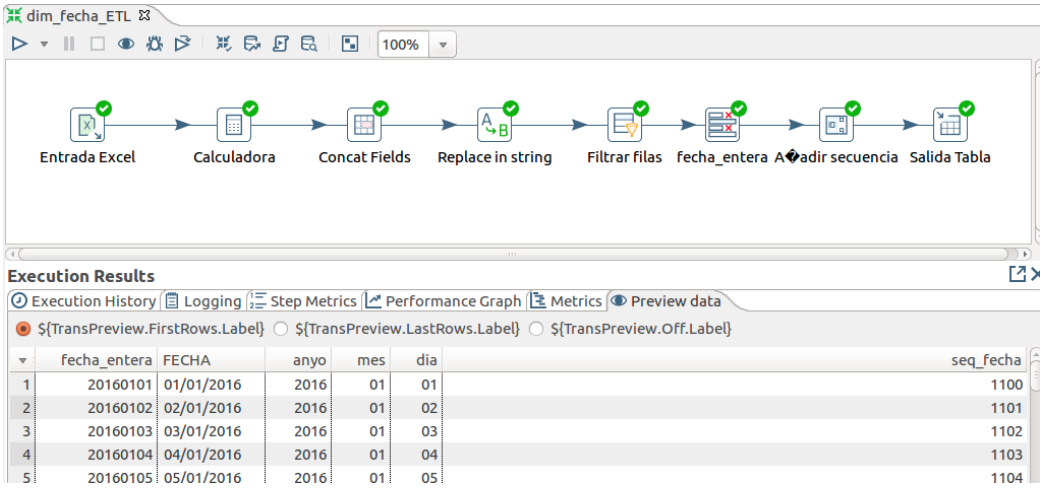




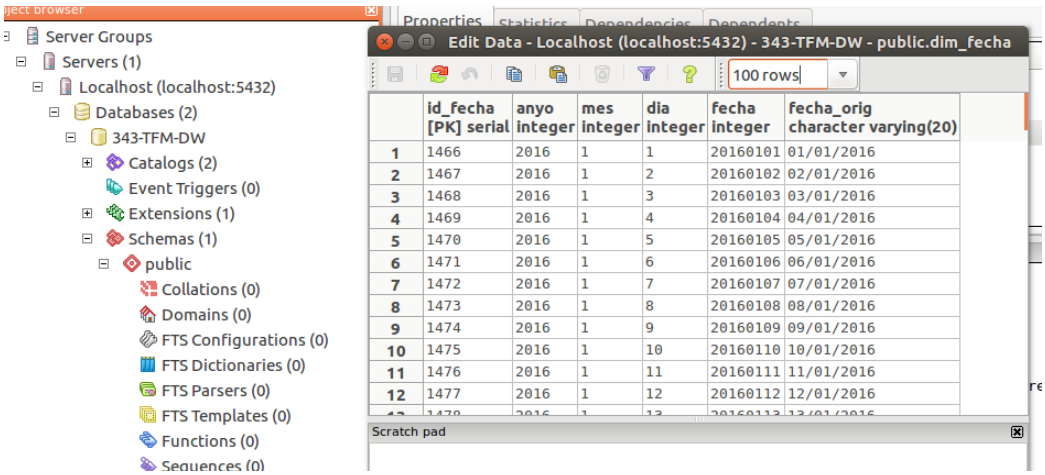
El resto de los pasos son:

- Utilizamos una transformación de tipo calculadora para obtener el día, mes y año de cada registro de fecha.
- Concatenamos los campos día, mes y año y además eliminamos de la cadena los espacios en blanco del resultado de esta concatenación. Generamos el campo fecha_entera.
- Realizamos un filtrado de filas con valor nulo.
- Utilizamos una transformación para cambiar el tipo de dato del campo fecha_entera de String a Number.

El resultado final del proceso es:



	fecha_entera	FECHA	anyo	mes	dia	seq_fecha
1	20160101	01/01/2016	2016	01	01	1100
2	20160102	02/01/2016	2016	01	02	1101
3	20160103	03/01/2016	2016	01	03	1102
4	20160104	04/01/2016	2016	01	04	1103
5	20160105	05/01/2016	2016	01	05	1104

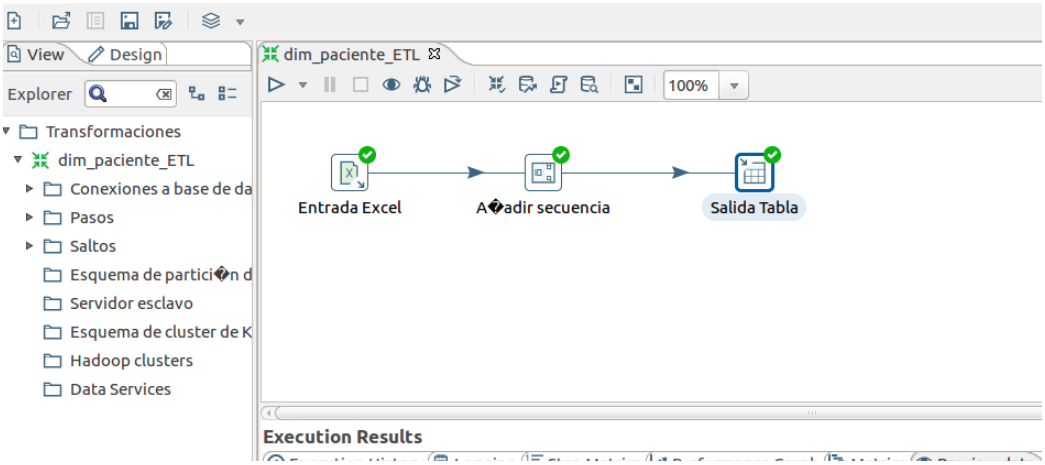


	id_fecha [PK] serial	anyo integer	mes integer	dia integer	fecha integer	fecha_orig character varying(20)
1	1466	2016	1	1	20160101	01/01/2016
2	1467	2016	1	2	20160102	02/01/2016
3	1468	2016	1	3	20160103	03/01/2016
4	1469	2016	1	4	20160104	04/01/2016
5	1470	2016	1	5	20160105	05/01/2016
6	1471	2016	1	6	20160106	06/01/2016
7	1472	2016	1	7	20160107	07/01/2016
8	1473	2016	1	8	20160108	08/01/2016
9	1474	2016	1	9	20160109	09/01/2016
10	1475	2016	1	10	20160110	10/01/2016
11	1476	2016	1	11	20160111	11/01/2016
12	1477	2016	1	12	20160112	12/01/2016

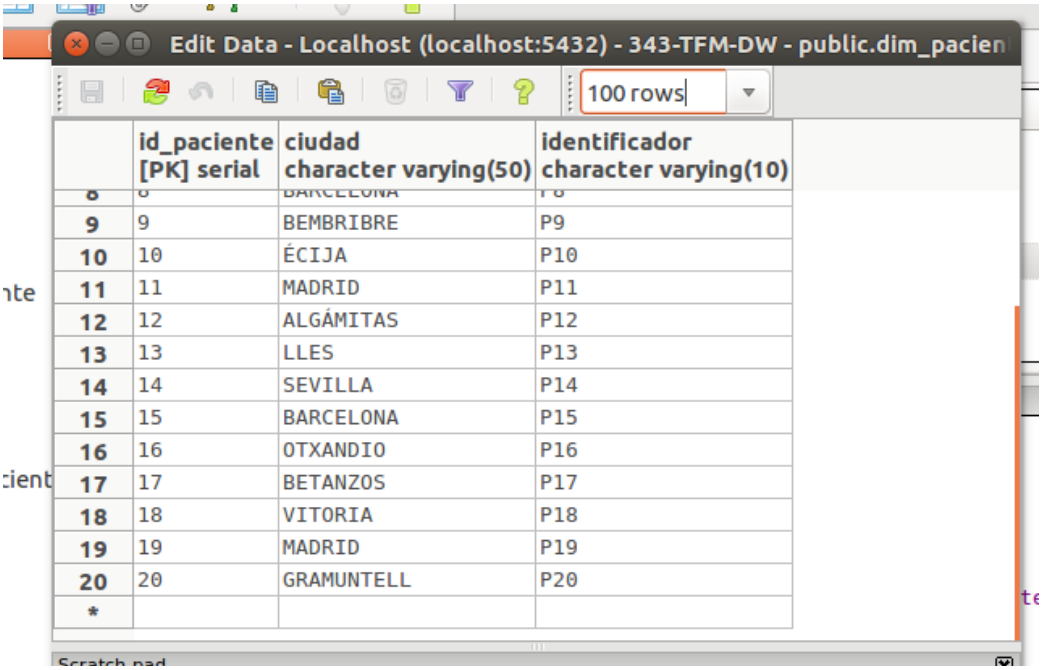
Y comprobamos la correcta inserción del proceso ETL en la BBDD:
ETL Pacientes

El proceso consta de los siguientes pasos:

- Carga de la hoja excel PATIENTS.
- Búsqueda de la secuencia definida para la tabla dim_pacientes.
- Inserción en la tabla dim_paciente de los campos id_paciente, identificador (P1,etc) y ciudad.



El resultado en la BBDD es:



	id_paciente [PK] serial	ciudad character varying(50)	identificador character varying(10)
9	9	BEMBRIBRE	P9
10	10	ÉCIJA	P10
11	11	MADRID	P11
12	12	ALGÁMITAS	P12
13	13	LLES	P13
14	14	SEVILLA	P14
15	15	BARCELONA	P15
16	16	OTXANDIO	P16
17	17	BETANZOS	P17
18	18	VITORIA	P18
19	19	MADRID	P19
20	20	GRAMUNTELL	P20
*			

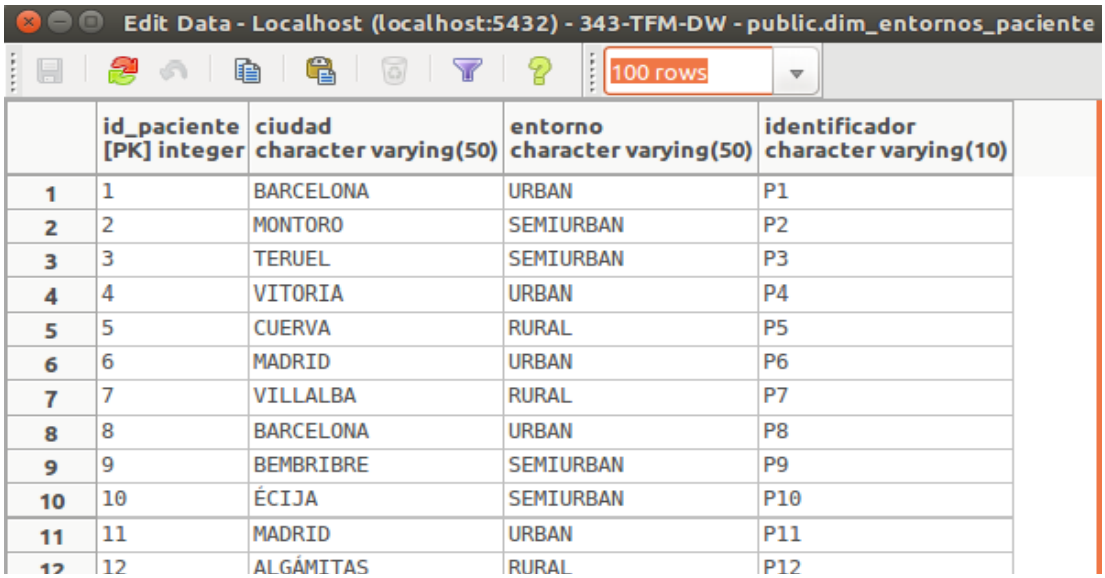
ETL Entornos paciente

El proceso carga el entorno de cada paicente en la tabla dim_entorno_paciente, y consta de los siguientes pasos:

- Carga de la hoja excel PATIENTS
- Búsqueda del id_paciente para cada identificador(P1) en la tabla dim_paciente creada anteriormente.
- Inserción en la tabla dim_entornos_ paciente de los campos identificador y entorno.

El proceso ETL es este:

El resultado de la BBDD es este:



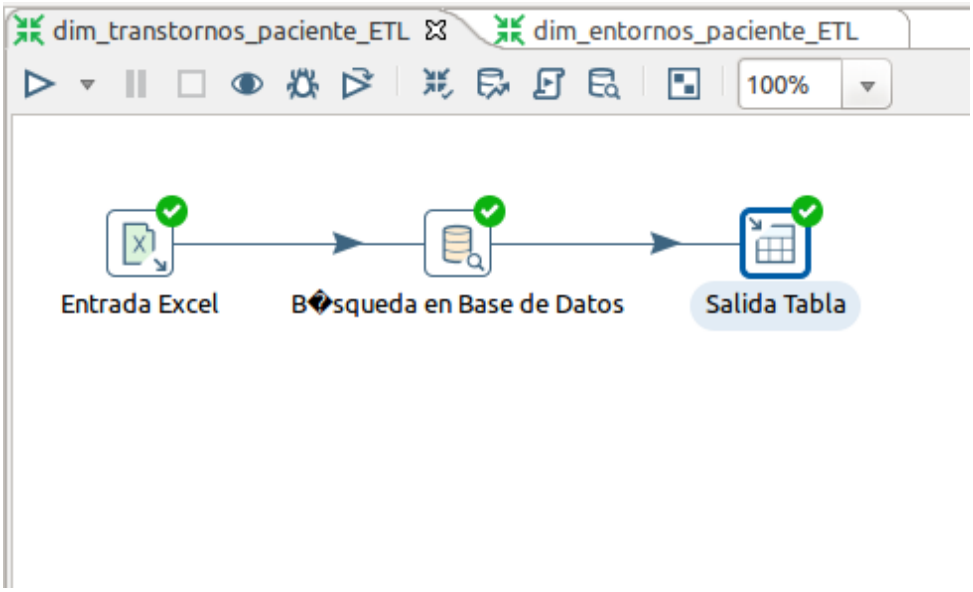
	id_paciente [PK] integer	ciudad character varying(50)	entorno character varying(50)	identificador character varying(10)
1	1	BARCELONA	URBAN	P1
2	2	MONTORO	SEMIURBAN	P2
3	3	TERUEL	SEMIURBAN	P3
4	4	VITORIA	URBAN	P4
5	5	CUERVA	RURAL	P5
6	6	MADRID	URBAN	P6
7	7	VILLALBA	RURAL	P7
8	8	BARCELONA	URBAN	P8
9	9	BEMBRIBRE	SEMIURBAN	P9
10	10	ÉCIJA	SEMIURBAN	P10
11	11	MADRID	URBAN	P11
12	12	ALGÁMITAS	RURAL	P12

ETL Transtornos paciente

El proceso carga el transtorno que sufre de cada paciente en la tabla dim_transtorno_paciente, y consta de los siguientes pasos:

- Carga de la hoja excel PATIENTS
- Búsqueda del id_paciente para cada identificador(P1) en la tabla dim_paciente creada anteriormente.
- Inserción en la tabla dim_transtorno_paciente de los campos identificador y transtorno.

El proceso ETL es muy similar al anterior, de hecho se podía haber fusionado en un único proceso con varios accesos a escritura en tablas de bbdd con variables diferentes:



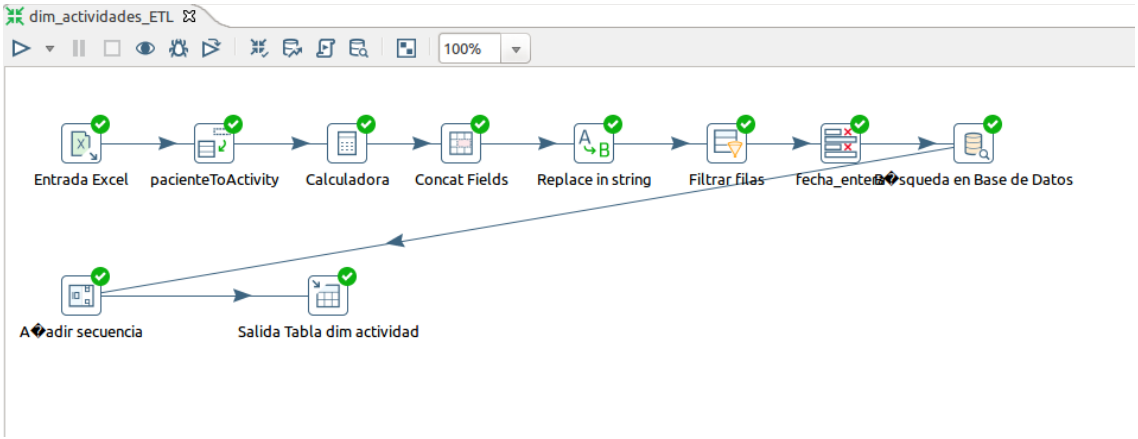
El resultado de ejecutar el proceso de inserción en la BBDD es este:

	id_paciente [PK] integer	ciudad character varying(50)	transtorno character varying(50)	identificador character varying(10)
1	1	BARCELONA	DELIRIUM	P1
2	2	MONTORO	DELIRIUM	P2
3	3	TERUEL	DELIRIUM	P3
4	4	VITORIA	DELIRIUM	P4
5	5	CUERVA	DELIRIUM	P5
6	6	MADRID	DELIRIUM	P6
7	7	VILLALBA	DELIRIUM	P7
8	8	BARCELONA	DEMENTIA	P8
9	9	BEMBRIBRE	DEMENTIA	P9

ETL Actividad paciente

El proceso almacena la actividad de cada paciente en la tabla de BBDD dim_actividad_paciente, y consta de los siguientes pasos:

- Carga de la hoja excel ACTIVITY VALUES
- Tratamiento de la fecha para obtener el id_fecha de la tabla dim_fecha.
- Búsqueda del id_actividad en la secuencia definida en la BBDD.
- Inserción en la tabla dim_actividades de los campos paciente, fecha y actividad.



El resultado en la base de datos es:

Edit Data - Localhost (localhost:5432) - 343-TFM-DW - public.dim_

100 rows

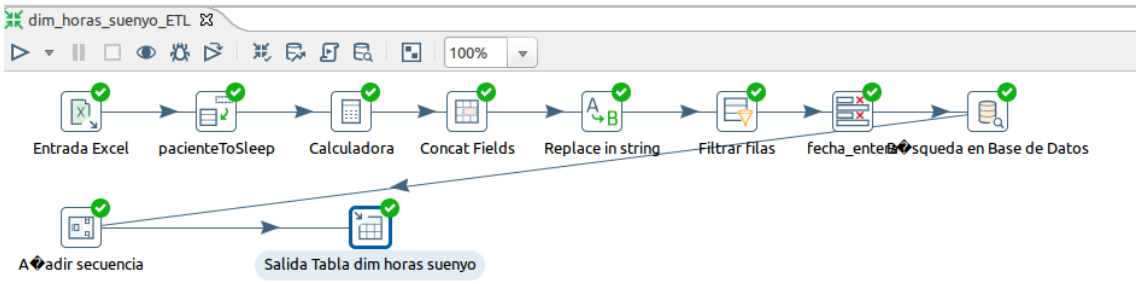
	identificador_pac character varying(50)	id_actividad [PK] serial	actividad character varying(50)	id_fecha integer
1	P1	7321	NO ACTIVITY	1466
2	P2	7322	RADIO/TV	1466
3	P3	7323	RADIO/TV	1466
4	P4	7324	NO ACTIVITY	1466
5	P5	7325	FAMILY	1466
6	P6	7326	READ/STUDY	1466
7	P7	7327	READ/STUDY	1466
8	P8	7328	RADIO/TV	1466
9	P9	7329	NO ACTIVITY	1466

ETL Horas de sueño paciente

El proceso almacena la actividad de cada paciente en la tabla de BBDD dim_actividad_paciente, y consta de los siguientes pasos:

- Carga de la hoja excel HOUR SLEEP VALUES
- Tratamiento de la fecha para obtener el id_fecha de la tabla dim_fecha.
- Búsqueda del id_sleep_hour en la secuencia definida en la BBDD.
- Inserción en la tabla dim_horas_suenyo de los campos paciente, fecha y horas_suenyo.

El proceso es este:



Y el volcado de datos en la tabla de dimensión es:

Edit Data - Localhost (localhost:5432) - 343-TFM-DW - p

100 rows

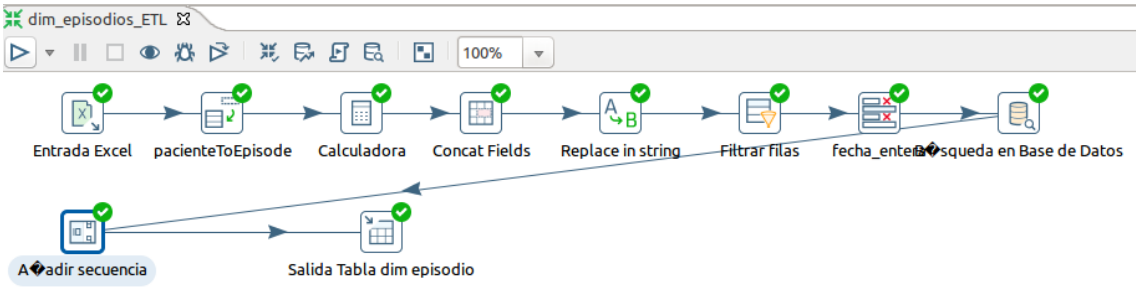
	id_horas_suenyo [PK] serial	horas integer	id_fecha integer	identificador_pac character varying(50)
1	16846	4	1466	P1
2	16847	5	1466	P2
3	16848	5	1466	P3
4	16849	4	1466	P4
5	16850	6	1466	P5
6	16851	5	1466	P6
7	16852	6	1466	P7
8	16853	5	1466	P8
9	16854	5	1466	P9
10	16855	5	1466	P10
11	16856	5	1466	P11

ETL Episodios paciente

El proceso almacena la actividad de cada paciente en la tabla de BBDD dim_actividad_paciente, y consta de los siguientes pasos:

- Carga de la hoja excel EPISODE VALUES
- Tratamiento de la fecha para obtener el id_fecha de la tabla dim_fecha en la que se produce el episodio.
- Búsqueda del id_episode en la secuencia definida en la BBDD..
- Inserción en la tabla dim_episodios_paciente de los campos identificador, fecha y episodio.

El proceso definido para la carga de los episodios de paciente es el siguiente:



La ejecución del mismo produce este resultado en la tabla dim_episodios de la base de datos:

Edit Data - Localhost (localhost:5432) - 343-TFM-DW - public.dim_e

100 rows

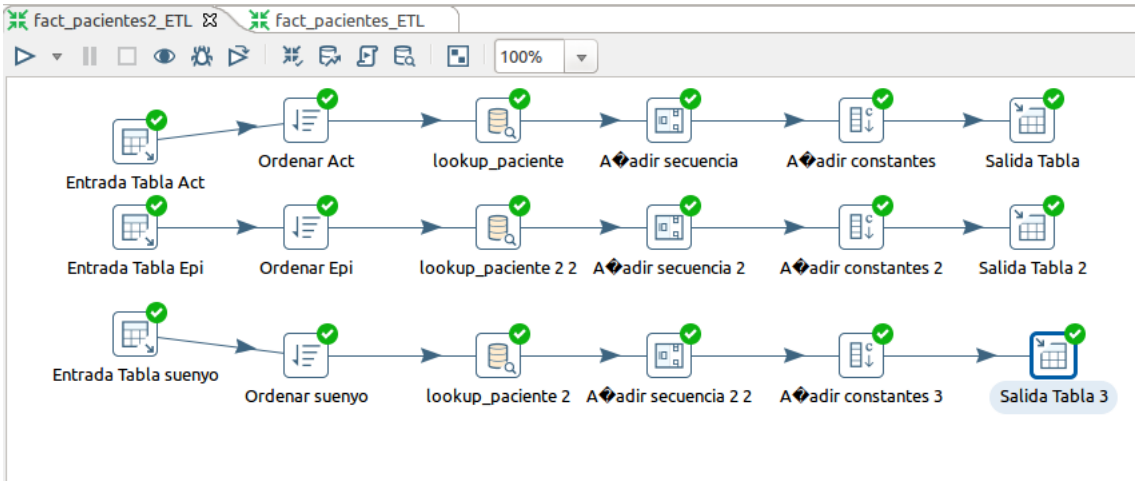
	id_episodio [PK] serial	episodio character varying(25)	identificador_pac character varying(50)	id_fecha integer
1	7321	LIGHT	P1	1466
2	7322	SEVERE	P2	1466
3	7323	NO EPISODE	P3	1466
4	7324	NO EPISODE	P4	1466
5	7325	NO EPISODE	P5	1466
6	7326	LIGHT	P6	1466
7	7327	NO EPISODE	P7	1466
8	7328	MODERATE	P8	1466
9	7329	LIGHT	P9	1466
10	7330	LIGHT	P10	1466
11	7331	LIGHT	P11	1466
12	7332	NO EPISODE	P12	1466
13	7333	LIGHT	P13	1466
14	7334	LIGHT	P14	1466
15	7335	LIGHT	P15	1466
16	7336	NO EPISODE	P16	1466
17	7337	MODERATE	P17	1466

ETL Hechos paciente

El proceso almacena la actividad de cada paciente en la tabla de BBDD hechos_paciente. Para llevar a cabo esta carga desde diferentes tablas de la base de datos se ha procedido de la siguiente manera:

- Se definen 3 procesos en paralelo que accederán a las tablas de dimensiones e insertarán concurrentemente en la tabla de hechos.
- Acceder a la tabla de dim_actividades, dim_episodio y dim_horas_suenyo y obtener todas las filas ordenadas por el campo identificador_pac.
- Búsqueda en cada tabla de dimensiones por id_paciente, necesario este dato para insertar en la tabla de hechos.
- Una vez tenemos todas las dimensiones cargadas, para cada una de ellas definimos una columna constante con la cabecera 'variable'. Esta columna almacenará el tipo dde variable que se almacena en el registro de la tabla de hechos, pudiento tomar los valores ACTIVIDAD, EPISODIO, HORAS_SUEÑO
- Inserción en la tabla hechos_paciente de cada uno de los tres procesos de carga en paralelo..

El proceso para rellenar la tabla de hechos se muestra a continuación:



Se han insertado un total de filas en la tabla de hechos. Una captura de los datos se puede ver a continuación:

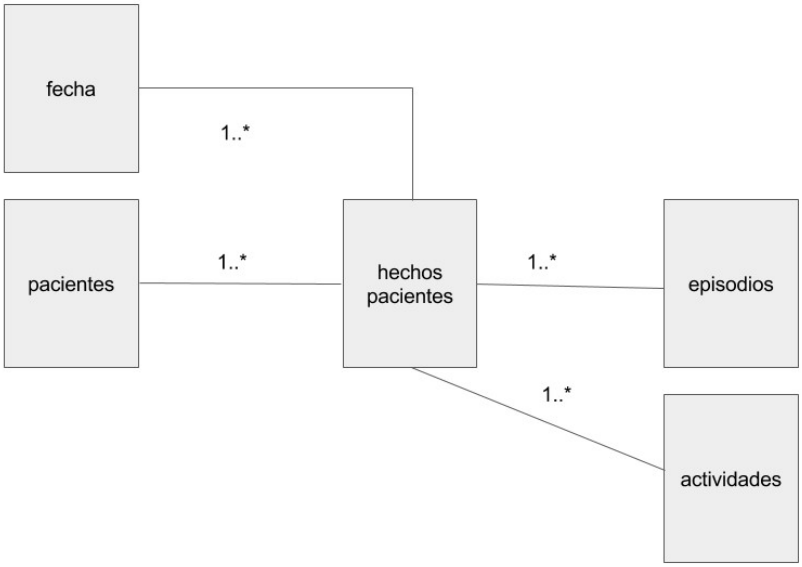
159	159	1	1492	HORAS_SUENYO	5
160	160	1	1541	ACTIVIDAD	READ/STUDY
161	161	1	1523	EPISODIO	NO EPISODE
162	162	1	1493	HORAS_SUENYO	5
163	163	1	1542	ACTIVIDAD	EXERCISE
164	164	1	1524	EPISODIO	LIGHT
165	165	1	1494	HORAS_SUENYO	5
166	166	1	1495	HORAS_SUENYO	5
167	167	1	1525	EPISODIO	NO EPISODE
168	168	1	1543	ACTIVIDAD	NO ACTIVITY
169	169	1	1526	EPISODIO	LIGHT
170	170	1	1544	ACTIVIDAD	EXERCISE
171	171	1	1527	EPISODIO	NO EPISODE
172	172	1	1528	EPISODIO	NO EPISODE
173	173	1	1545	ACTIVIDAD	FAMILY
174	174	1	1529	EPISODIO	LIGHT
175	175	1	1530	EPISODIO	LIGHT

21960 rows.

2.4. Correctivo sobre el diseño de la base de datos y los procesos ETL.

A raíz de comenzar a hacer las primeras pruebas en el diseño y visualización del cubo OLAP, se detectaron algunos problemas en el esquema de la BBDD inicial, así como en los procesos ETLs que rellenaban tal esquema.
A continuación se muestra el diagrama final de la base de datos y el proceso ETL que haciendo uso de tablas auxiliares inserta los datos en la tabla de hechos.

Diagrama del esquema final de la BBDD.



En él podemos ver que han quedado 4 dimensiones y pasamos el cálculo de las horas de sueño de cada paciente como una columna en la tabla de hechos.

Diagrama de tablas auxiliares para los procesos ETL.

Vemos que se han utilizado tres tablas auxiliares sobre las que insertan datos procesos ETL.
A continuación mostramos un extracto de cada tabla auxiliar.

Aux_episodios:

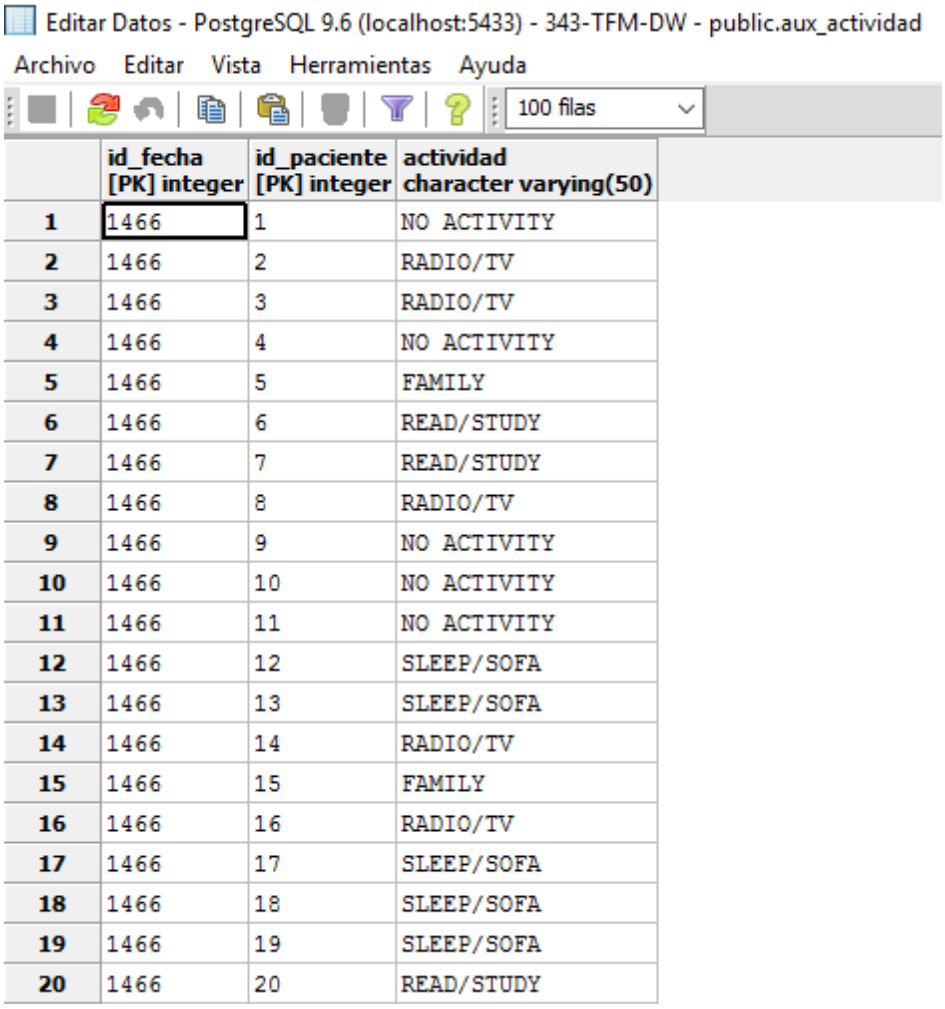
Editar Datos - PostgreSQL 9.6 (localhost:5433) - 343-TFM-DW - public.aux_episodio

Archivo Editar Vista Herramientas Ayuda

100 filas

	id_fecha [PK] integer	id_paciente [PK] integer	episodio character varying(50)
1	1466	1	LIGHT
2	1466	2	SEVERE
3	1466	3	NO EPISODE
4	1466	4	NO EPISODE
5	1466	5	NO EPISODE
6	1466	6	LIGHT
7	1466	7	NO EPISODE
8	1466	8	MODERATE
9	1466	9	LIGHT
10	1466	10	LIGHT
11	1466	11	LIGHT
12	1466	12	NO EPISODE
13	1466	13	LIGHT
14	1466	14	LIGHT
15	1466	15	LIGHT
16	1466	16	NO EPISODE
17	1466	17	MODERATE
18	1466	18	SEVERE
19	1466	19	SEVERE
20	1466	20	SEVERE

Aux_actividades:



	id_fecha [PK] integer	id_paciente [PK] integer	actividad character varying(50)
1	1466	1	NO ACTIVITY
2	1466	2	RADIO/TV
3	1466	3	RADIO/TV
4	1466	4	NO ACTIVITY
5	1466	5	FAMILY
6	1466	6	READ/STUDY
7	1466	7	READ/STUDY
8	1466	8	RADIO/TV
9	1466	9	NO ACTIVITY
10	1466	10	NO ACTIVITY
11	1466	11	NO ACTIVITY
12	1466	12	SLEEP/SOFA
13	1466	13	SLEEP/SOFA
14	1466	14	RADIO/TV
15	1466	15	FAMILY
16	1466	16	RADIO/TV
17	1466	17	SLEEP/SOFA
18	1466	18	SLEEP/SOFA
19	1466	19	SLEEP/SOFA
20	1466	20	READ/STUDY

Aux_horas_suenyo:

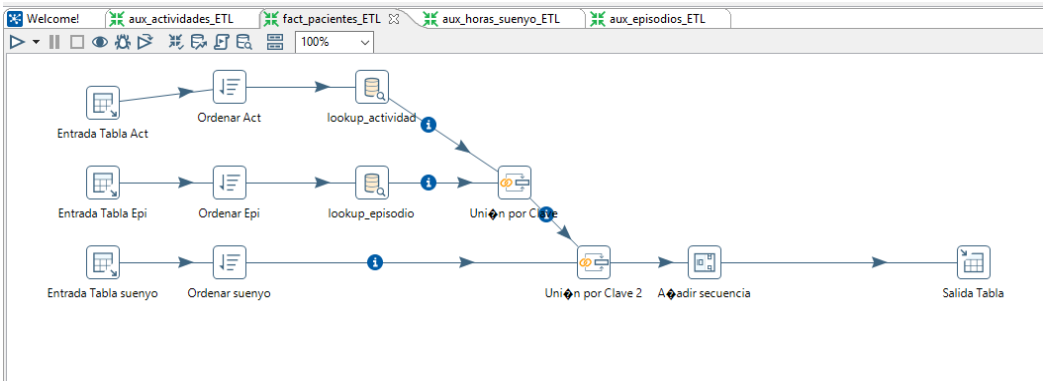
Editar Datos - PostgreSQL 9.6 (localhost:5433) - 343-TFM-DW - public.aux_horas

Archivo Editar Vista Herramientas Ayuda

100 filas

	id_fecha [PK] integer	id_paciente [PK] integer	horas_suenyo integer
1	1466	1	4
2	1466	2	5
3	1466	3	5
4	1466	4	4
5	1466	5	6
6	1466	6	5
7	1466	7	6
8	1466	8	5
9	1466	9	5
10	1466	10	5
11	1466	11	5
12	1466	12	6
13	1466	13	6
14	1466	14	4
15	1466	15	5
16	1466	16	4
17	1466	17	6
18	1466	18	4
19	1466	19	4
20	1466	20	5

Proceso ETL de carga de información en la tabla de hechos.



El proceso se carga se hace accediendo a las 3 tablas auxiliares y mediante la utilización de 2 transformacióes “Join” de Pentaho PDI unimos la salida de los tres “streams” de las tablas auxiliares e insertamos los datos correspondientes en la tabla de hechos.

Una muestra de la tabla de hechos después de estos ajustes es la siguiente:

Editar Datos - PostgreSQL 9.6 (localhost:5433) - 343-TFM-DW - public.hechos_paciente

Archivo Editar Vista Herramientas Ayuda

100 filas

	id_hecho [PK] serial	id_paciente integer	id_fecha integer	id_actividad integer	id_episodio integer	horas_suenyo integer
1	7321	1	1466	2	0	4
2	7322	2	1466	3	3	5
3	7323	3	1466	3	2	5
4	7324	4	1466	2	2	4
5	7325	5	1466	1	2	6
6	7326	6	1466	4	0	5
7	7327	7	1466	4	2	6
8	7328	8	1466	3	1	5
9	7329	9	1466	2	0	5
10	7330	10	1466	2	0	5
11	7331	11	1466	2	0	5
12	7332	12	1466	5	2	6
13	7333	13	1466	5	0	6
14	7334	14	1466	3	0	4
15	7335	15	1466	1	0	5
16	7336	16	1466	3	2	4
17	7337	17	1466	5	1	6
18	7338	18	1466	5	3	4
19	7339	19	1466	5	3	4
20	7340	20	1466	4	3	5

2.5. Implementación de la capa de análisis de datos.

En esta última sección, una vez que tenemos el modelo de datos definitivo, se va a implementar la capa de visualización de datos a través de la herramienta Pentaho Community Edition, que contiene módulos para análisis OLAP y visualización de datos.
Mediante esta visualización se dará respuesta a las preguntas analíticas definidas en los objetivos del proyecto.

Para ello, utilizando la herramienta Pentaho Schema Workbench, se definirá un cubo OLAP partiendo de la BBDD en postgreSQL donde tenemos las tablas con los datos.
La información oficial sobre mondrian y la creación de cubos OLAP, se puede encontrar en:
http://mondrian.pentaho.com/documentation/schema.php

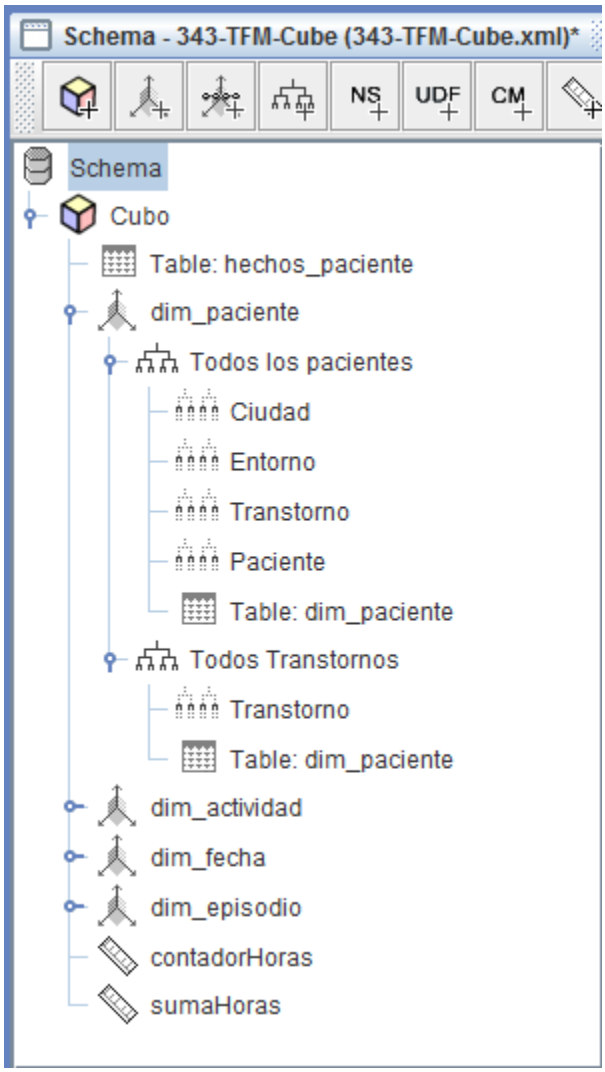
Lo primero que hay que hacer una vez iniciado el programa es crear un nuevo “Schema” y una nueva conexión a la base de datos para tener acceso a las tablas necesarias que se utilizarán en el cubo.

2.4.1. Creación del cubo

En esta sección se va a proceder a explicar la creación del cubo para realizar los análisis OLAP.
El cubo consta de los siguientes elementos:

- Dimensión Fecha
  - Nivel Año
  - Nivel Mes
  - Nivel Día
- Dimensión Episodio
- Dimensión Actividad
- Dimensión Paciente
  - Nivel Ciudad
  - Nivel Entorno
  - Nivel Transtorno
  - Nivel Paciente
- Medida contador de horas (cuenta las filas de la columna horas_sueño)
- Medida suma de horas (suma las filas de la columna horas_sueño)

El resultado final del cubo se puede ver en la siguiente imagen:

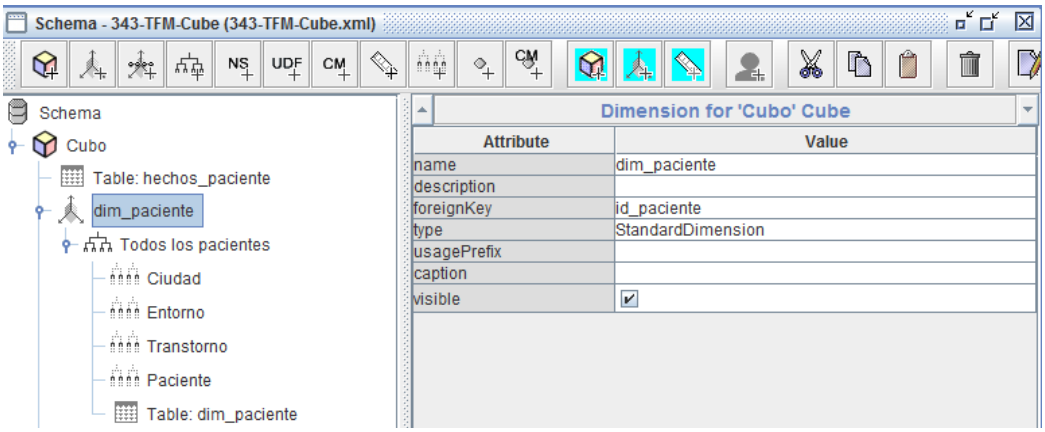


2.4.1.1. Definición de dimensiones

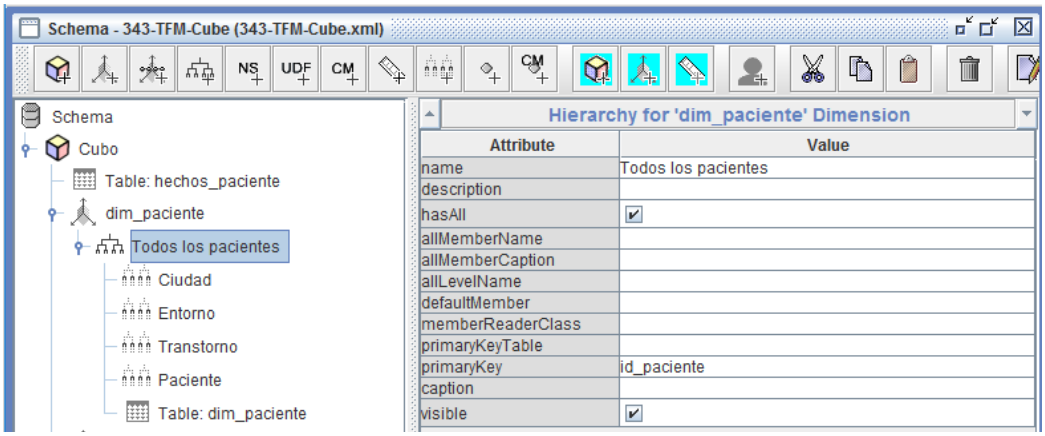
En este subapartado se creará una de las dimensiones necesarias para el análisis de datos. No se definirán todas, porque el proceso es el mismo para cada una de ellas. En concreto especificamos la dimesión paciente por ser las más completa en cuanto a niveles de jerarquía.

Dimensión Pacientes (dim_pacientes)

Creamos una nueva dimensión y asignamos la clave ajna de la tabla de hechos que relacionará la tabla de hechos con la dimensión.

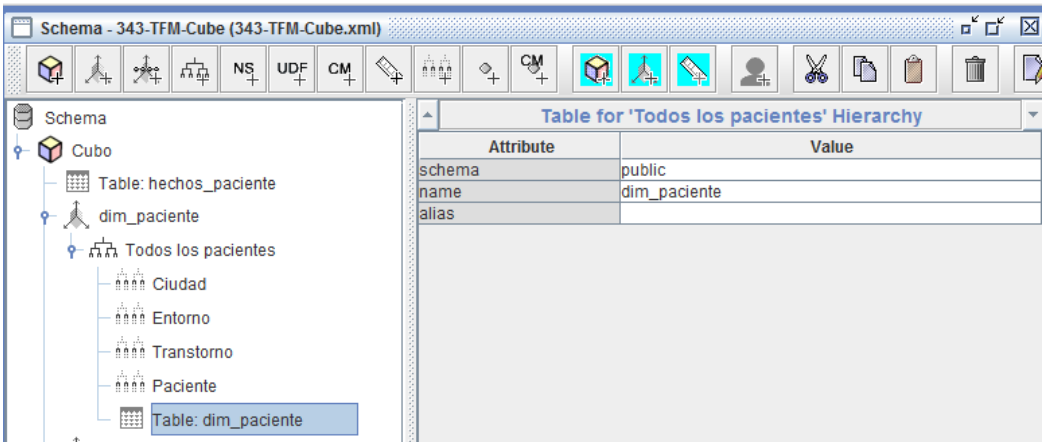


Una jerarquía:



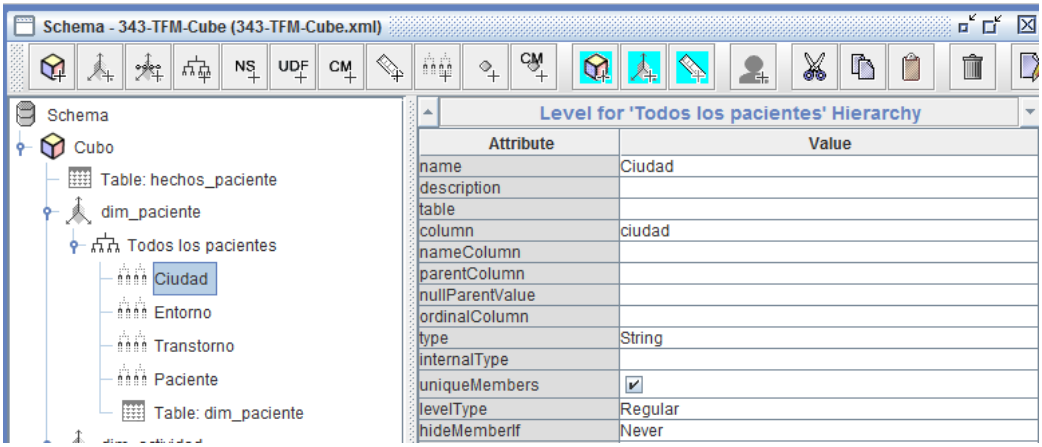
Vemos que en el campo PrimerayKey hemos puesto id_paciente, pero ese valor no se puede establacer hasta que no añadamos la tabla a la jerarquía, por lo que es un paso que debe realizarse después de añadir la tabla.

Dentro de la jerarquía añadimos una tabla, concretamente la tabla dim_paciente:

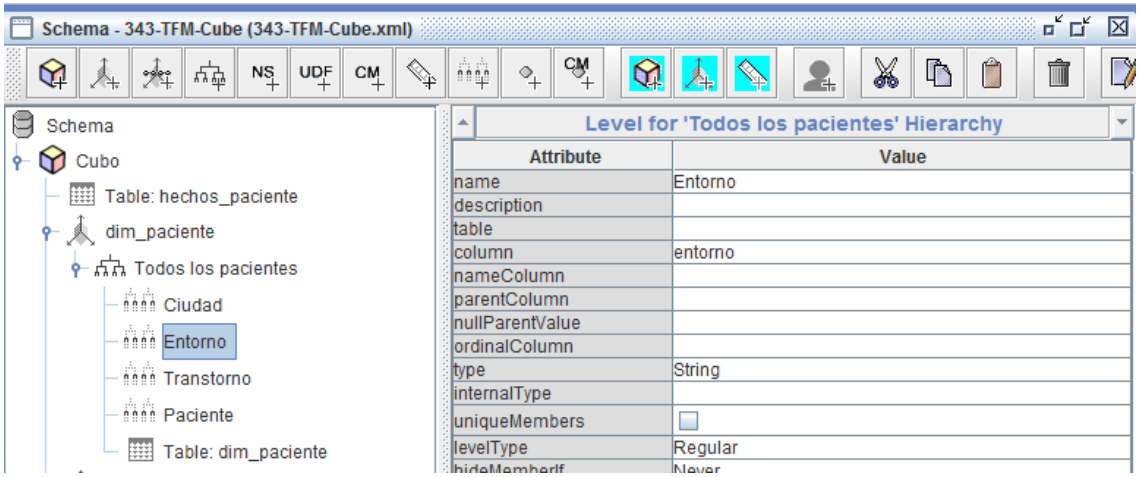


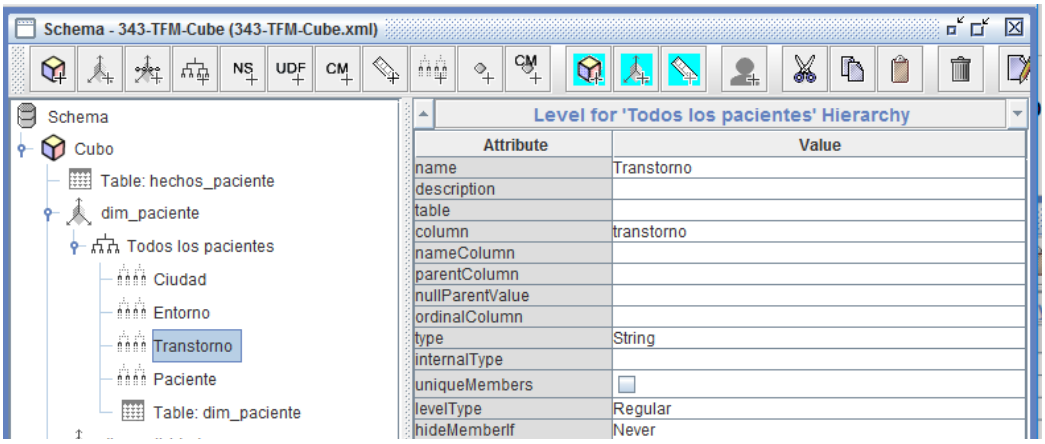
Ahora añadimos niveles en la jerarquía, asignando columnas de la dimensión de paciente por la que agruparemos los datos.

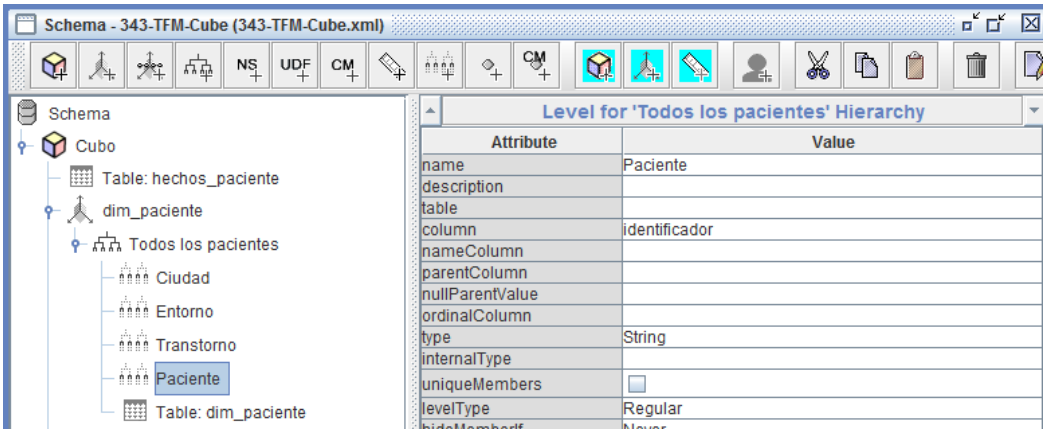
Comenzamos agrupando por el nivel ciudad:



Y a continuación definimos los demás niveles, entorno, transtorno y por último el nivel paciente.

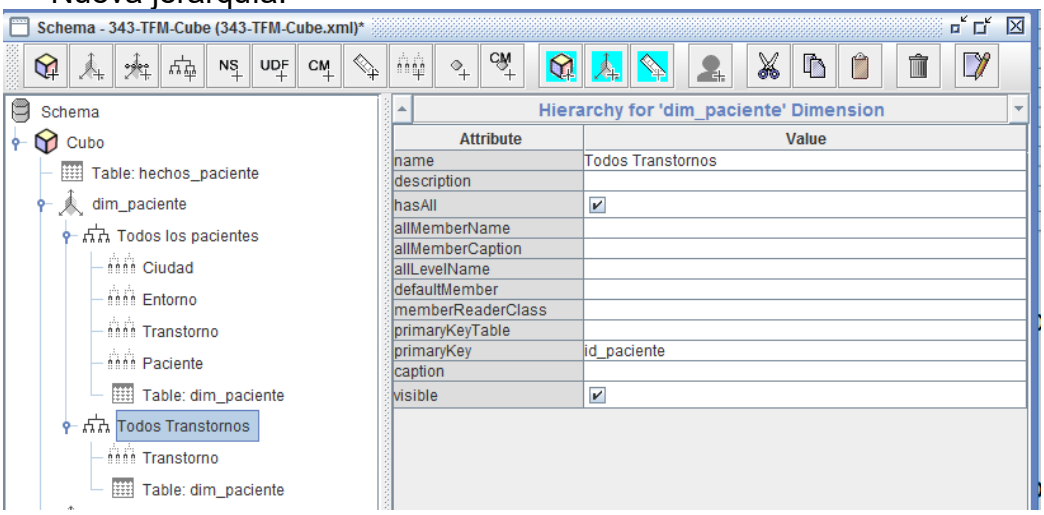




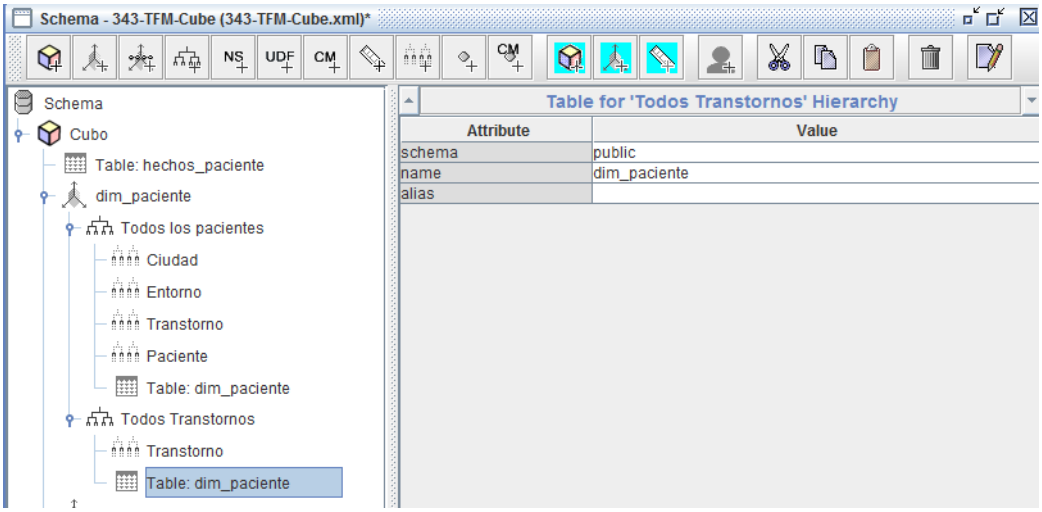


Hemos definido una jerarquía para agrupar datos por ciudades, ahora definimos otra jerarquía dentro de la misma dimensión para agrupar por tipo de transtorno. El procedimiento es el mismo que el anterior.

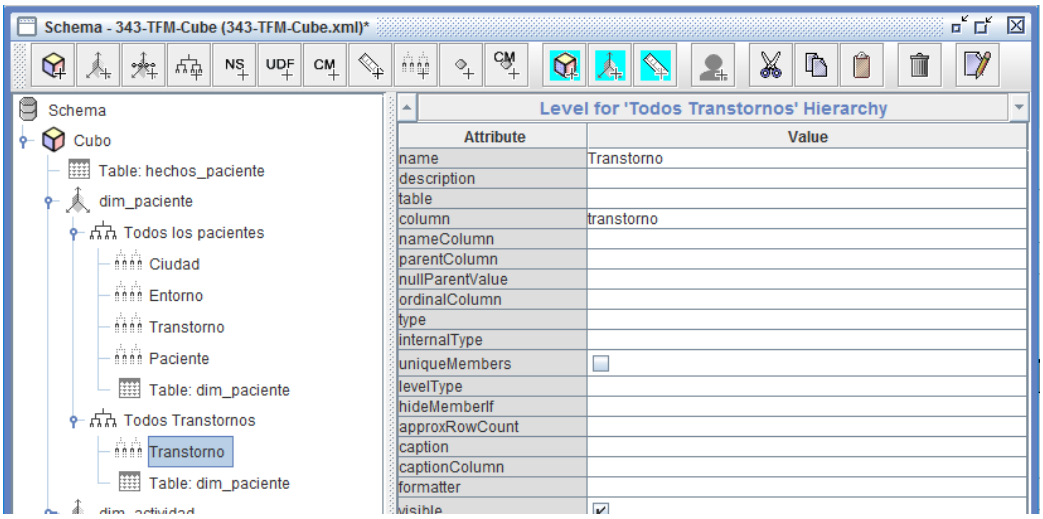
Nueva jerarquía:



Añadimos la tabla dim_paciente:



Y definimos el nivel Transtorno:

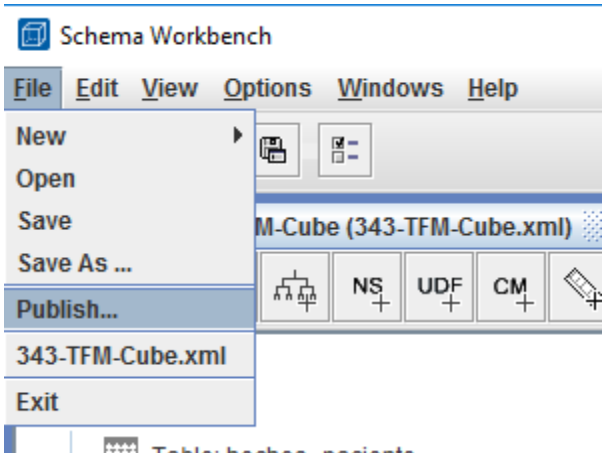


Cabe destacar, que en función de las pregunta analíticas, es posible que hayan que definir nuevas jerarquías para agrupar los datos de distintas formas.

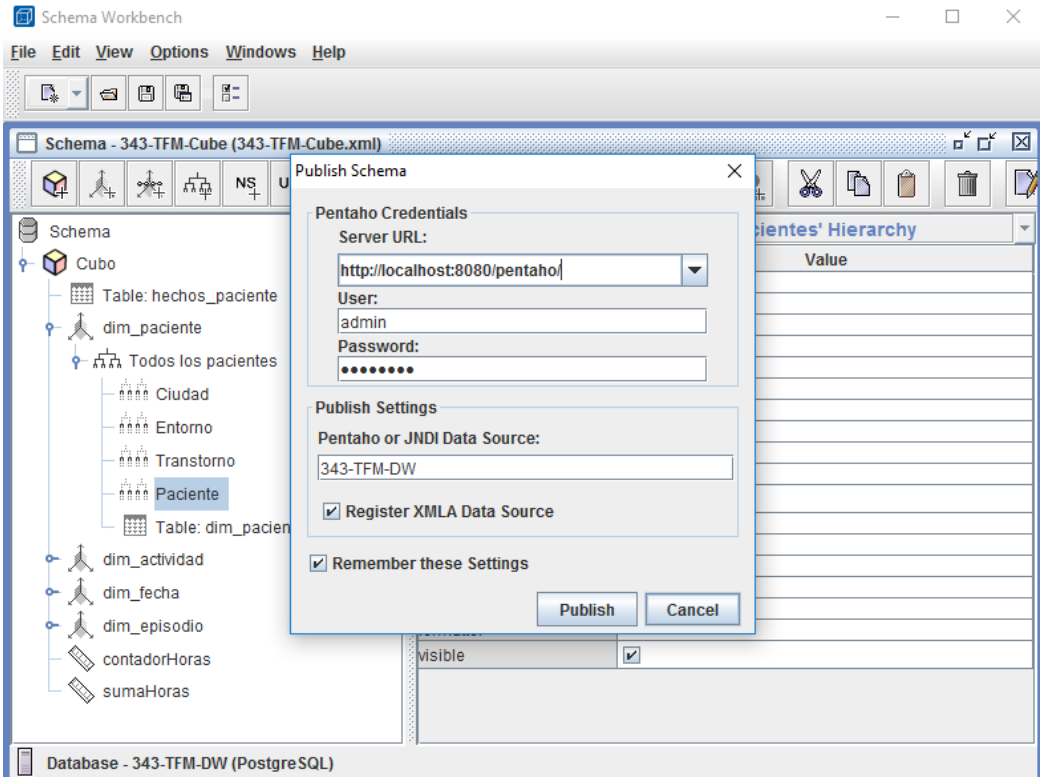
2.4.1.2. Publicación del cubo en Pentaho Server BI

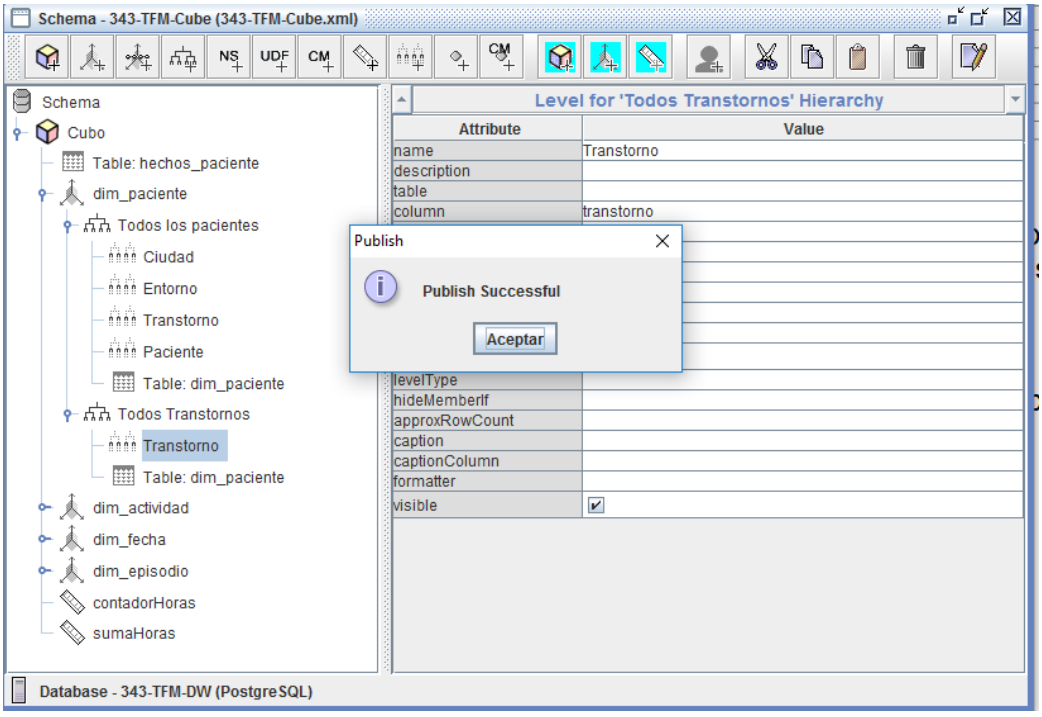
El último paso para poder tener acceso a los daots del cubo es publicarlo en el servidor de pentaho.

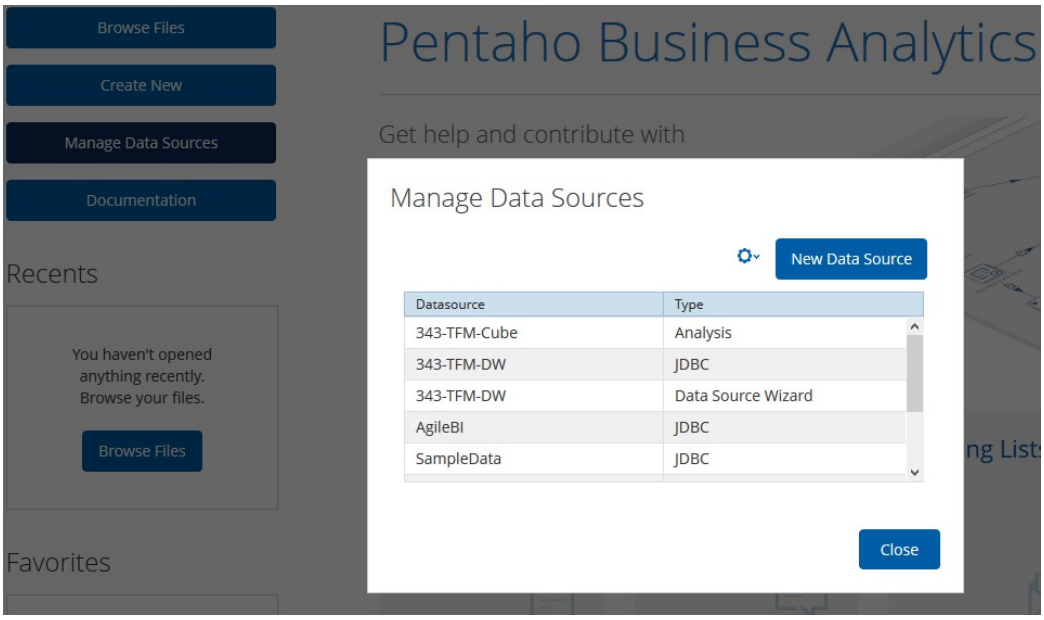
Para ello, desde Workbench, vamos a File->Publish...



A continuación, se nos preguntarán las credenciales de un usuario definido en el servidor y al aceptar, el cubo ya estará disponible desde pentaho server BI:


Schema Workbench
File Edit View Options Windows Help
Schema - 343-TFM-Cube (343-TFM-Cube.xml)
Publish Schema
Pentaho Credentials
Server URL:
http://localhost:8080/pentaho/
User:
admin
Password:
Publish Settings
Pentaho or JNDI Data Source:
343-TFM-DW
Register XMLA Data Source
Remember these Settings
Publish
Cancel
Schema
Cubo
Table: hechos_paciente
dim_paciente
Todos los pacientes
Ciudad
Entorno
Transtorno
Paciente
dim_actividad
dim_fecha
dim_episodio
contadorHoras
sumaHoras
Value
visible
Database - 343-TFM-DW (PostgreSQL)

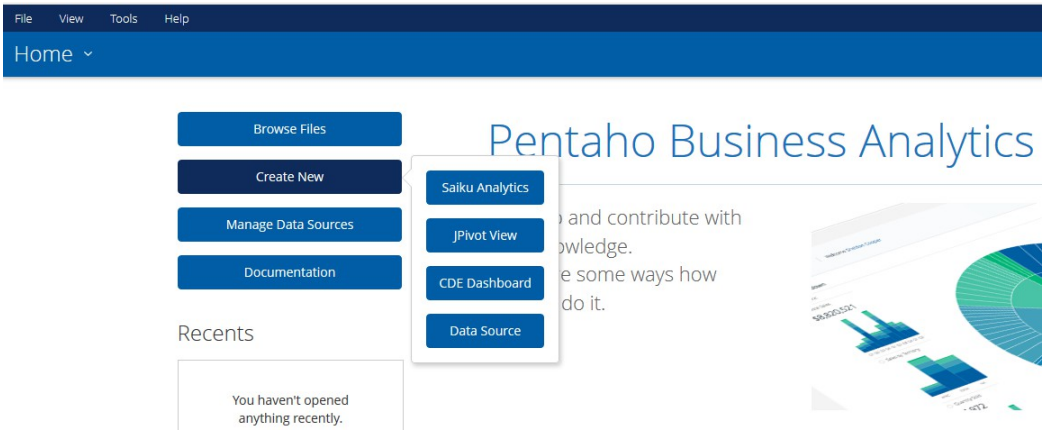

Schema - 343-TFM-Cube (343-TFM-Cube.xml)
Level for 'Todos Transtornos' Hierarchy
Attribute
Value
name
Transtorno
description
table
column
transtorno
Publish
Publish Successful
Aceptar
levelType
hideMemberIf
approxRowCount
caption
captionColumn
formatter
visible
Schema
Cubo
Table: hechos_paciente
dim_paciente
Todos los pacientes
Ciudad
Entorno
Transtorno
Paciente
Table: dim_paciente
Todos Transtornos
Transtorno
Table: dim_paciente
dim_actividad
dim_fecha
dim_episodio
contadorHoras
sumaHoras
Database - 343-TFM-DW (PostgreSQL)

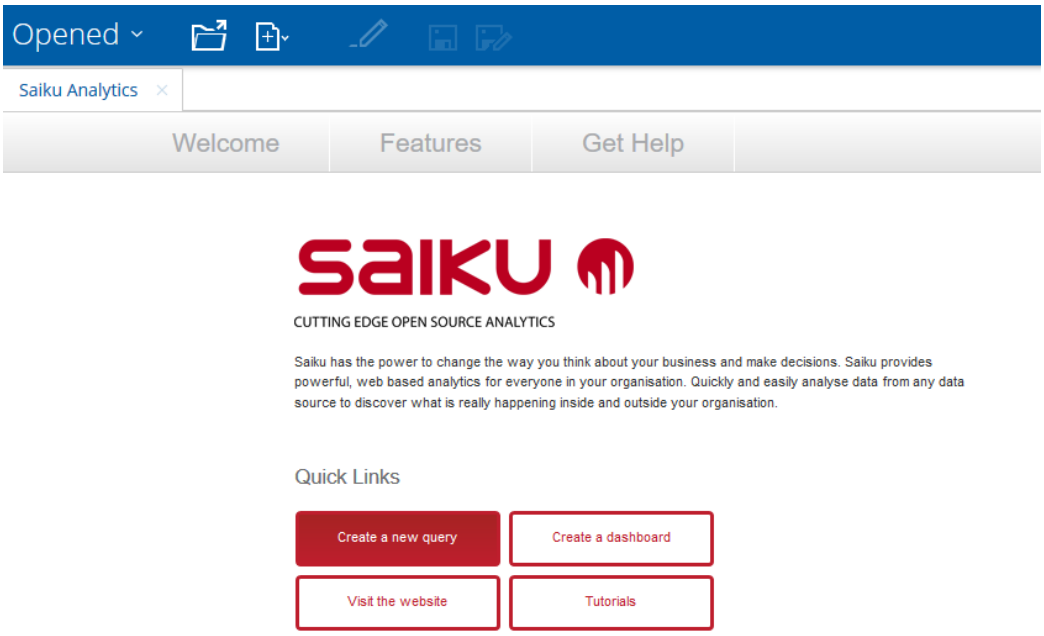


2.4.2. Preguntas analíticas

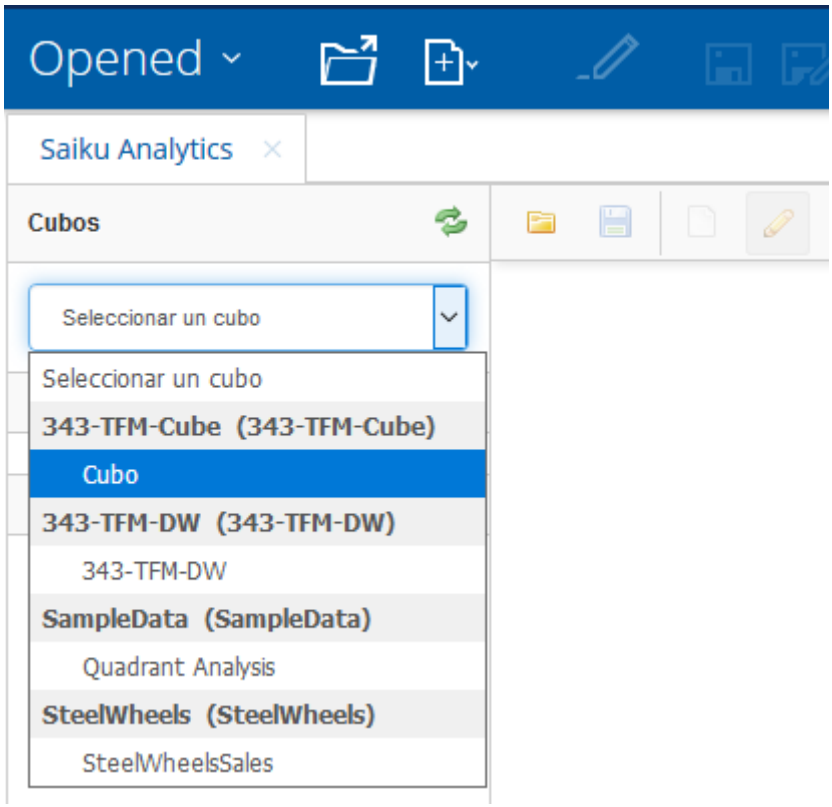
Una vez tenemos el cubo, se utilizará para visualizar los datos desde Pentaho Server BI, a través del plugin Saiku, que es bastante más flexible que el que viene por defecto en el servidor, Jpivot.

Para llegar a visualizar los datos, hay que entrar en el servidor de pentaho y si el plugin de Saiku está bien instalado, permitirá la creación de un “análisis Saiku” y crearemos una nueva “query”:





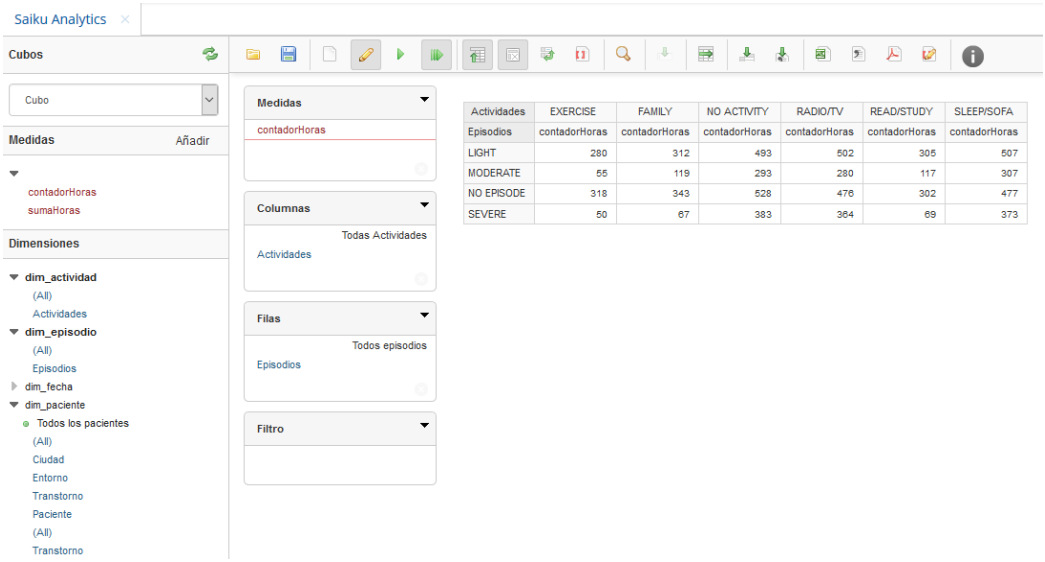
Por último seleccionamos el cubo de entre las fuentes de datos disponibles en el servidor:



2.4.2.1.¿Cuál es la relación entre las actividades realizadas y los episodios de crisis graves?

Para resolver esta pregunta, se ha cruzado la dimensión actividades con la dimesión episodios con la medida de contador de registros de horas

de sueño, que en realidad se ha hecho sobre esa columna pero es un contador de filas para una determinada condición:



Actividades	EXERCISE	FAMILY	NO ACTIVITY	RADIO/TV	READ/STUDY	SLEEP/SOFA
Episodios	contadorHoras	contadorHoras	contadorHoras	contadorHoras	contadorHoras	contadorHoras
LIGHT	280	312	493	502	305	507
MODERATE	55	119	293	280	117	307
NO EPISODE	318	343	528	476	302	477
SEVERE	50	67	383	364	69	373

Si ponemos un filtro sobre la dimensión Episodios para que nos muestre únicamente los severos:

Actividades	EXERCISE	FAMILY	NO ACTIVITY	RADIO/TV	READ/STUDY	SLEEP/SOFA
Episodios	contadorHoras	contadorHoras	contadorHoras	contadorHoras	contadorHoras	contadorHoras
SEVERE	50	67	383	364	69	373

Con estos datos, la conclusión es que la realización es la actividad que menos episodios severos ha registrado, seguido de estar con la familia y leer o esudiar. La realización de ninguna actividad o de dormir y/o estar en el sofá, son las actividades que más episodios severos han registrado, con 383 y 373 respectivamente en contraste con los excasos 50 de la realización de ejercicio.

2.4.2.2.¿Se puede establecer algún tipo de relación entre los valores de los diferentes estados de ánimo y los episodios de crisis?

No disponemos de datos con los estados de ánimo por lo que no podemos respondr a esta pregunta y queda fuera del alcance del proyecto.

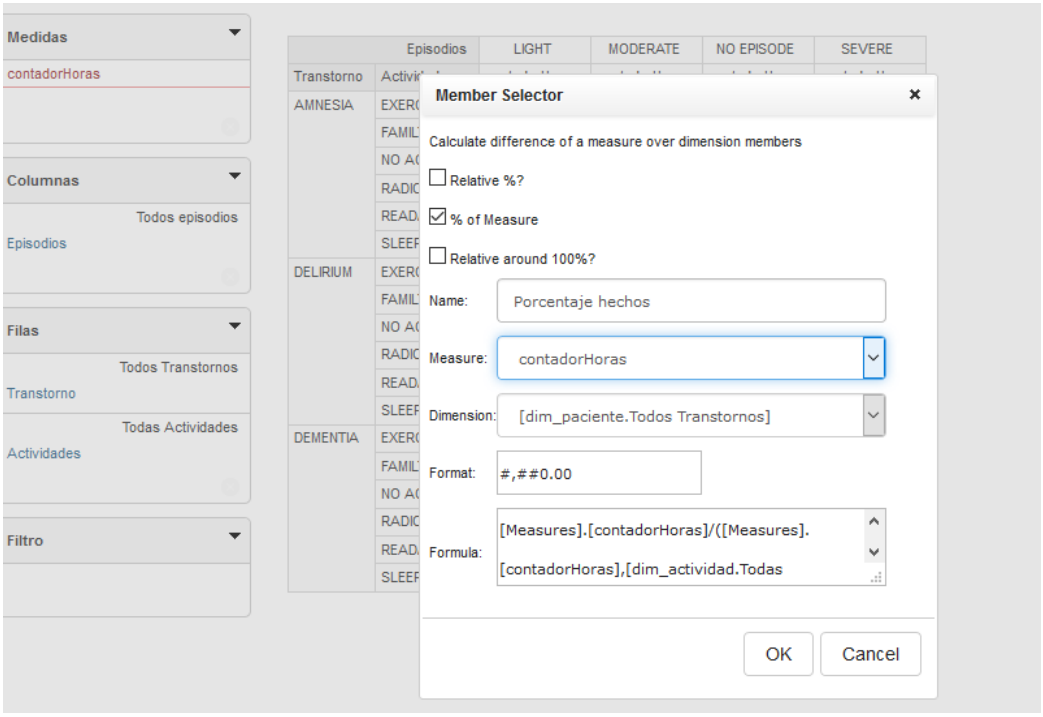
2.4.2.3.Estas relaciones son iguales para cualquiera de las enfermedades o en cambio hay relaciones más acusadas por alguna de ellas.

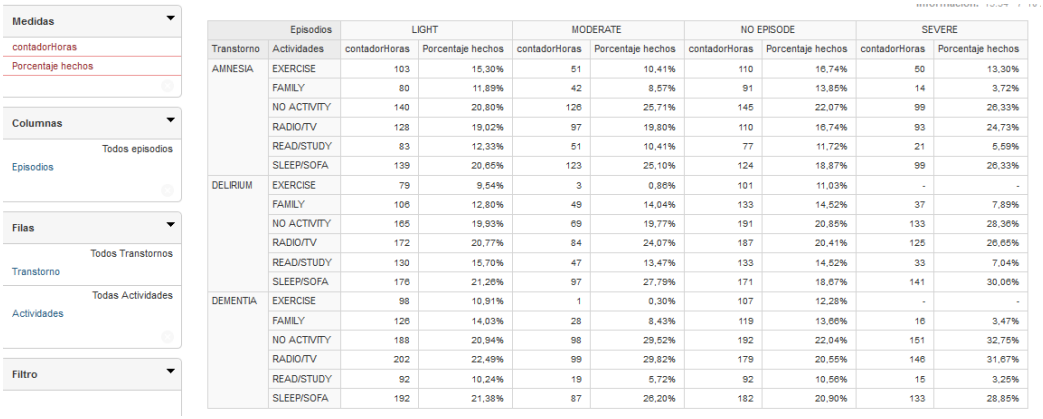
En este caso, se proporciona un informe con datos agrupados por Transtorno, de la dimensión paciente, cruzados con actividades y episodios.

Esto nos da el siguiente gráfico:

	Episodios	LIGHT	MODERATE	NO EPISODE	SEVERE
Transtorno	Actividades	contadorHoras	contadorHoras	contadorHoras	contadorHoras
AMNESIA	EXERCISE	103	51	110	50
	FAMILY	80	42	91	14
	NO ACTIVITY	140	126	145	99
	RADIO/TV	128	97	110	93
	READ/STUDY	83	51	77	21
	SLEEP/SOFA	139	123	124	99
DELIRIUM	EXERCISE	79	3	101	-
	FAMILY	106	49	133	37
	NO ACTIVITY	165	69	191	133
	RADIO/TV	172	84	187	125
	READ/STUDY	130	47	133	33
	SLEEP/SOFA	176	97	171	141
DEMENTIA	EXERCISE	98	1	107	-
	FAMILY	126	28	119	16
	NO ACTIVITY	188	98	192	151
	RADIO/TV	202	99	179	146
	READ/STUDY	92	19	92	15
	SLEEP/SOFA	192	87	182	133

De estos datos, la conclusión que extraigo es que los episodios de la enfermedad en relación a la actividad y el transtorno se distribuye en de la misma forma en las tres enfermedades. Los número más bajos para cada una de ellas siguen estando en el ejercicio, familia y lectura/estudio, mientras siguen siendo los valores más altos los correspondentes a no realizar actividades o permanecer en el sofá o durmiendo.
Podemos calcular una medida desde Saiku para contrastar esta hipótesis visual.





	Episodios	LIGHT		MODERATE		NO EPISODE		SEVERE	
Transtorno	Actividades	contadorHoras	Porcentaje hechos	contadorHoras	Porcentaje hechos	contadorHoras	Porcentaje hechos	contadorHoras	Porcentaje hechos
AMNESIA	EXERCISE	103	15,30%	51	10,41%	110	16,74%	50	13,30%
	FAMILY	80	11,89%	42	8,57%	91	13,85%	14	3,72%
	NO ACTIVITY	140	20,80%	126	25,71%	145	22,07%	99	26,33%
	RADIO/TV	128	19,02%	97	19,80%	110	16,74%	93	24,73%
	READ/STUDY	83	12,33%	51	10,41%	77	11,72%	21	5,59%
	SLEEP/SOFA	139	20,65%	123	25,10%	124	18,87%	99	26,33%
DELIRIUM	EXERCISE	79	9,54%	3	0,86%	101	11,03%	-	-
	FAMILY	106	12,80%	49	14,04%	133	14,52%	37	7,89%
	NO ACTIVITY	165	19,93%	69	19,77%	191	20,85%	133	28,36%
	RADIO/TV	172	20,77%	84	24,07%	187	20,41%	125	26,65%
	READ/STUDY	130	15,70%	47	13,47%	133	14,52%	33	7,04%
	SLEEP/SOFA	176	21,26%	97	27,79%	171	18,67%	141	30,06%
DEMENTIA	EXERCISE	98	10,91%	1	0,30%	107	12,28%	-	-
	FAMILY	126	14,03%	28	8,43%	119	13,66%	16	3,47%
	NO ACTIVITY	188	20,94%	98	29,52%	192	22,04%	151	32,75%
	RADIO/TV	202	22,49%	99	29,82%	179	20,55%	146	31,67%
	READ/STUDY	92	10,24%	19	5,72%	92	10,56%	15	3,25%
	SLEEP/SOFA	192	21,38%	87	26,20%	182	20,90%	133	28,85%

Con estos datos parece clara la hipótesis de que la no realización de actividades o permanecer en el sofá o durmiendo aumenta el riesgo de padecer un episodio severo.

También hay un dato que me llama la atención si miro la columna de “NO EPISODE”, y es que los datos son contrarios al “SEVERE”, es decir, la no realización de actividad o “SLEEP/SOFA” tienen más entradas en el caso de NO EPISODE.

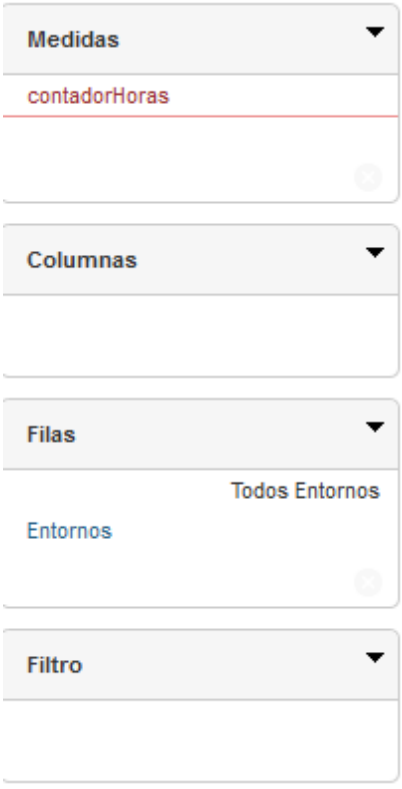
Este dato me hace replantearme si existe realmente una relación o no entre todos los tipos de episodio, pero lo que sí que parece claro es que la realización de ejercicio disminuye los casos severos.

2.4.2.4.¿Se puede establecer alguna relación en nivel geográfico, por ejemplo entorno urbano o rural?

La dimensión pacientes tiene definidad una jerarquía por ciudad, en la que podríamos ver la relación de episodios y/o actividades en diferentes ciudades y luego bajar al nivel de entorno como en la siguiente figura:

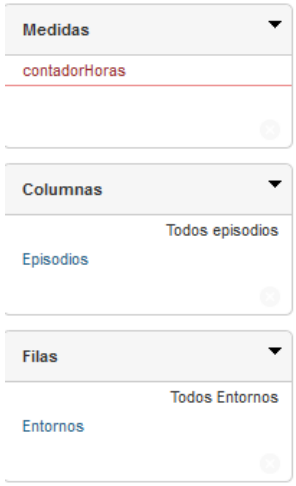
Ciudad	Entorno	LIGHT	MODERATE	NO EPISODE	SEVERE
ALGÁMITAS	RURAL	705	401	770	320
BARCELONA	URBAN	998	432	979	519
BEMBRIBRE	SEMIURBAN	696	338	695	467
BETANZOS	RURAL	705	401	770	320
CUERVA	RURAL	705	401	770	320
GRAMUNTELL	RURAL	705	401	770	320
LLES	RURAL	705	401	770	320
MADRID	SEMIURBAN	696	338	695	467
	URBAN	998	432	979	519
MONTORO	SEMIURBAN	696	338	695	467
OTXANDIO	SEMIURBAN	696	338	695	467
SEVILLA	URBAN	998	432	979	519
TERUEL	SEMIURBAN	696	338	695	467
VILLALBA	RURAL	705	401	770	320
VITORIA	URBAN	998	432	979	519
ÉCIJA	SEMIURBAN	696	338	695	467

Pero para agrupar primeramente por entorno, definiré en el cubo una nueva jerarquía en la dimensión paciente y la aplicaré en el informe.



Entornos	contadorHoras
RURAL	2.196
SEMIURBAN	2.196
URBAN	2.928

De aquí concluímos que en el entorno URBANO hay más registros que en los rurales y semiurbanos. Si afinamos un poco por episodios podemos ver su distribución entorno / episodio.



Episodios	LIGHT	MODERATE	NO EPISODE	SEVERE
Entornos	contadorHoras	contadorHoras	contadorHoras	contadorHoras
RURAL	705	401	770	320
SEMIURBAN	696	338	695	467
URBAN	998	432	979	519

2.4.2.5.¿Cuál sido la evolución de los diferentes pacientes a lo largo del tiempo?

Vamos a definir una relación entre la dimensión fecha, la dimensión episodio y la dimención paciente. Para cada paciente, veremos por cada mes el porcentaje de cada tipo de episodios que ha sufrido.

	Todos los pacientes - Paciente	P12	P15	P1	P8	P9	P17	P5	P20	P13	P19	P6	P11	P2
Mes	Episodios	Porcentaje	Porcentaje	Porcentaje	Porcentaje	Porcentaje	Porcentaje	Porcentaje	Porcentaje	Porcentaje	Porcentaje	Porcentaje	Porcentaje	Porcentaje
1	LIGHT	25,81%	41,94%	58,06%	12,90%	51,61%	45,16%	25,81%	12,90%	64,52%	12,90%	61,29%	22,58%	16,13%
	MODERATE	12,90%	25,81%	-	29,03%	-	25,81%	3,23%	19,35%	3,23%	22,58%	3,23%	19,35%	19,35%
	NO EPISODE	38,71%	32,26%	41,94%	19,35%	48,39%	29,03%	70,97%	25,81%	32,26%	22,58%	35,48%	32,26%	12,90%
	SEVERE	22,58%	-	-	38,71%	-	-	-	41,94%	-	41,94%	-	25,81%	51,61%
2	LIGHT	44,83%	37,93%	44,83%	27,59%	48,28%	48,28%	51,72%	31,03%	58,62%	20,69%	51,72%	10,34%	10,34%
	MODERATE	10,34%	34,48%	3,45%	13,79%	-	27,59%	-	20,69%	-	17,24%	3,45%	24,14%	20,69%
	NO EPISODE	17,24%	27,59%	51,72%	13,79%	51,72%	24,14%	48,28%	17,24%	41,38%	10,34%	44,83%	27,59%	41,38%
	SEVERE	27,59%	-	-	44,83%	-	-	-	31,03%	-	51,72%	-	37,93%	27,59%
3	LIGHT	29,03%	29,03%	48,39%	19,35%	45,16%	29,03%	29,03%	12,90%	54,84%	12,90%	51,61%	22,58%	22,58%
	MODERATE	16,13%	38,71%	3,23%	22,58%	-	41,94%	3,23%	16,13%	3,23%	25,81%	-	9,68%	12,90%
	NO EPISODE	32,26%	32,26%	48,39%	19,35%	54,84%	29,03%	67,74%	25,81%	41,94%	25,81%	48,39%	25,81%	22,58%
	SEVERE	22,58%	-	-	38,71%	-	-	-	45,16%	-	35,48%	-	41,94%	41,94%
4	LIGHT	10,00%	50,00%	56,67%	10,00%	43,33%	43,33%	40,00%	13,33%	50,00%	23,33%	46,67%	10,00%	16,67%
	MODERATE	36,67%	20,00%	3,33%	13,33%	-	23,33%	-	20,00%	-	20,00%	3,33%	13,33%	16,67%
	NO EPISODE	36,67%	30,00%	40,00%	23,33%	56,67%	33,33%	60,00%	23,33%	50,00%	20,00%	50,00%	40,00%	16,67%
	SEVERE	16,67%	-	-	53,33%	-	-	-	43,33%	-	36,67%	-	36,67%	50,00%
5	LIGHT	22,58%	35,48%	51,61%	38,71%	35,48%	35,48%	41,94%	22,58%	48,39%	32,26%	48,39%	6,45%	29,03%
	MODERATE	38,71%	32,26%	3,23%	6,45%	-	29,03%	3,23%	19,35%	3,23%	22,58%	3,23%	25,81%	16,13%
	NO EPISODE	16,13%	32,26%	45,16%	3,23%	64,52%	35,48%	54,84%	19,35%	48,39%	12,90%	48,39%	19,35%	12,90%
	SEVERE	22,58%	-	-	51,61%	-	-	-	38,71%	-	32,26%	-	48,39%	41,94%
6	LIGHT	33,33%	33,33%	40,00%	16,67%	56,67%	23,33%	26,67%	10,00%	56,67%	20,00%	50,00%	20,00%	13,33%
	MODERATE	20,00%	6,67%	-	16,67%	-	36,67%	-	33,33%	-	30,00%	-	23,33%	26,67%
	NO EPISODE	26,67%	60,00%	60,00%	33,33%	43,33%	40,00%	73,33%	20,00%	43,33%	13,33%	50,00%	13,33%	30,00%
	SEVERE	20,00%	-	-	33,33%	-	-	-	36,67%	-	36,67%	-	43,33%	30,00%

2.4.2.6.¿Se puede establecer alguna relación entre los episodios de crisis y el momento del día o de la semana o del año?

Ajustaremos la vista anterior para visualizar todos los episodios, excepto el valor “NO EPISODE” y los dias de cada mes, por si hubiera algún día en concreto que se producen más crisis.

Episodios		LIGHT		MODERATE		SEVERE	
Mes	Dia	Porcentaje	contadorHoras	Porcentaje	contadorHoras	Porcentaje	contadorHoras
1	1	0,33%	8	0,17%	2	0,31%	4
	2	0,38%	9	0,09%	1	0,23%	3
	3	0,46%	11	0,09%	1	0,38%	5
	4	0,29%	7	0,34%	4	0,31%	4
	5	0,33%	8	0,34%	4	0,38%	5
	6	0,29%	7	0,26%	3	0,15%	2
	7	0,29%	7	0,26%	3	0,23%	3
	8	0,17%	4	0,26%	3	0,31%	4
	9	0,17%	4	0,26%	3	0,15%	2
	10	0,08%	2	0,26%	3	0,38%	5
	11	0,54%	13	0,17%	2	0,15%	2
	12	0,25%	6	0,26%	3	0,15%	2
	13	0,29%	7	0,17%	2	0,31%	4
	14	0,29%	7	0,17%	2	0,31%	4
	15	0,38%	9	0,17%	2	0,23%	3
	16	0,13%	3	0,51%	6	0,31%	4
	17	0,21%	5	0,26%	3	0,38%	5
	18	0,33%	8	0,51%	6	0,08%	1
	19	0,17%	4	0,17%	2	0,38%	5

Después de ver los días de cada mes y los episodios de cada día, no parece que haya un día o un tipo de episodio que se produzca con más frecuencia que los demás. Los orcentajes son bastante uniformes para todos los pacientes.

2.4.2.7.La realización de actividades físicas mejora o empeora el estado de ánimo de los pacientes.

De nuevo, no disponemos de datos de estado de ánimo, por lo que esta pregunta queda fuera del alcance del proyecto.

2.4.2.8.¿Hay algún tipo de actividad que mejore el día a día de los pacientes?

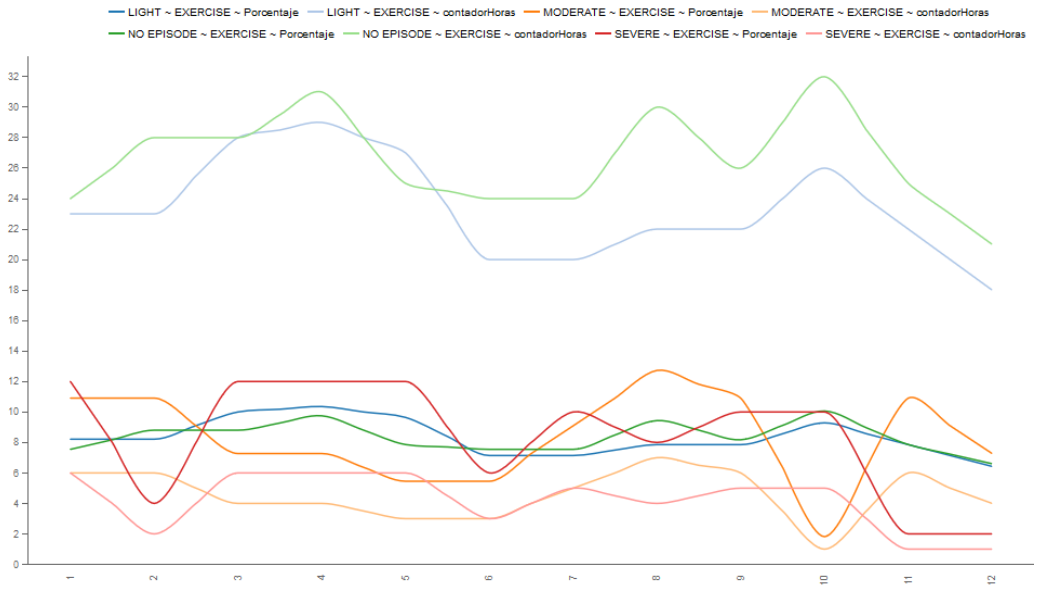
Vamos a cruzar los datos de episodios y actividades con fecha por mes, y veremos la evolusión de episodios “LIGHT”, “MODERATE”, “SEVERE” y “NO EPISODE” contra cada una de las actividades por separado para una mejor visualización.

Comenzamos con “EXERCISE”:

Episodios	LIGHT		MODERATE		NO EPISODE		SEVERE	
Actividades	EXERCISE		EXERCISE		EXERCISE		EXERCISE	
Mes	Porcentaje	contadorHoras	Porcentaje	contadorHoras	Porcentaje	contadorHoras	Porcentaje	contadorHoras
1	8,21%	23	10,91%	6	7,55%	24	12,00%	6
2	8,21%	23	10,91%	6	8,81%	28	4,00%	2
3	10,00%	28	7,27%	4	8,81%	28	12,00%	6
4	10,36%	29	7,27%	4	9,75%	31	12,00%	6
5	9,64%	27	5,45%	3	7,86%	25	12,00%	6
6	7,14%	20	5,45%	3	7,55%	24	6,00%	3
7	7,14%	20	9,09%	5	7,55%	24	10,00%	5
8	7,86%	22	12,73%	7	9,43%	30	8,00%	4
9	7,86%	22	10,91%	6	8,18%	26	10,00%	5
10	9,29%	26	1,82%	1	10,06%	32	10,00%	5
11	7,86%	22	10,91%	6	7,86%	25	2,00%	1
12	6,43%	18	7,27%	4	6,60%	21	2,00%	1

En ella parece que después de un año realizando ejercicio, los episodios se reducen, sobre todo los de tipo severo. Pero también es cierto que los registros de “NO EPISODE se mantienen, no aumentan, lo que indicaría que se producen menos episodios durante cada mes del año.”

La gráfica muestra la tendencia a la baja:



Actividad “FAMILY”:

Medidas

Porcentaje

contadorHoras

Columnas

Todos episodios

Episodios

Todas Actividades

Actividades

Filas

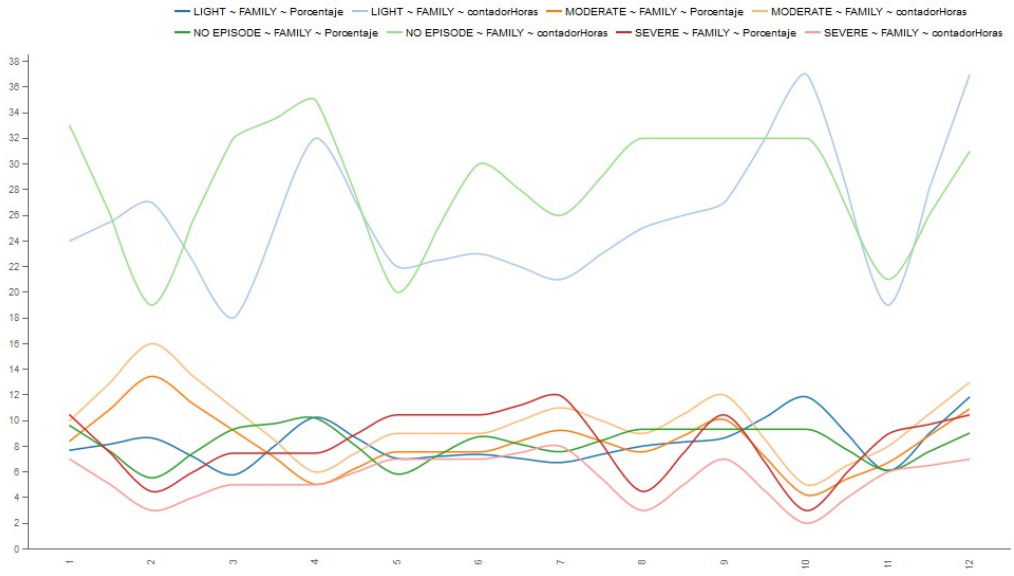
Fechas

Mes

Filtro

Episodios	LIGHT		MODERATE		NO EPISODE		SEVERE	
Actividades	FAMILY		FAMILY		FAMILY		FAMILY	
Mes	Porcentaje	contadorHoras	Porcentaje	contadorHoras	Porcentaje	contadorHoras	Porcentaje	contadorHoras
1	7,69%	24	8,40%	10	9,62%	33	10,45%	7
2	8,65%	27	13,45%	16	5,54%	19	4,48%	3
3	5,77%	18	9,24%	11	9,33%	32	7,46%	5
4	10,26%	32	5,04%	6	10,20%	35	7,46%	5
5	7,05%	22	7,56%	9	5,83%	20	10,45%	7
6	7,37%	23	7,56%	9	8,75%	30	10,45%	7
7	6,73%	21	9,24%	11	7,58%	26	11,94%	8
8	8,01%	25	7,56%	9	9,33%	32	4,48%	3
9	8,65%	27	10,08%	12	9,33%	32	10,45%	7
10	11,86%	37	4,20%	5	9,33%	32	2,99%	2
11	6,09%	19	6,72%	8	6,12%	21	8,96%	6
12	11,86%	37	10,92%	13	9,04%	31	10,45%	7

No parece que exista una evolución positiva al cabo de un año de esta actividad en ningún tipo de episodio. De hecho, los episodios tanto moderados como ligeros aumentan en número.



Actividad “Read/Study”:

Medidas

Porcentaje

contadorHoras

Columnas

Todos episodios

Episodios

Todas Actividades

Actividades

Filas

Fechas

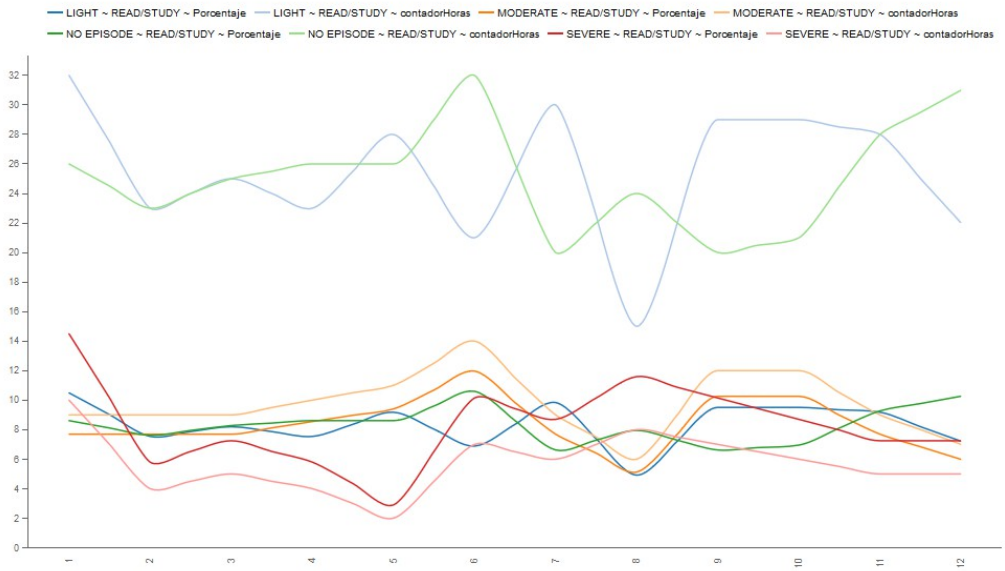
Mes

Filtro

Episodios	LIGHT		MODERATE		NO EPISODE		SEVERE	
Actividades	READ/STUDY		READ/STUDY		READ/STUDY		READ/STUDY	
Mes	Porcentaje	contadorHoras	Porcentaje	contadorHoras	Porcentaje	contadorHoras	Porcentaje	contadorHoras
1	10,49%	32	7,69%	9	8,61%	26	14,49%	10
2	7,54%	23	7,69%	9	7,62%	23	5,80%	4
3	8,20%	25	7,69%	9	8,28%	25	7,25%	5
4	7,54%	23	8,55%	10	8,61%	26	5,80%	4
5	9,18%	28	9,40%	11	8,61%	26	2,90%	2
6	6,89%	21	11,97%	14	10,60%	32	10,14%	7
7	9,84%	30	7,69%	9	6,62%	20	8,70%	6
8	4,92%	15	5,13%	6	7,95%	24	11,59%	8
9	9,51%	29	10,26%	12	6,62%	20	10,14%	7
10	9,51%	29	10,26%	12	6,95%	21	8,70%	6
11	9,18%	28	7,69%	9	9,27%	28	7,25%	5
12	7,21%	22	5,98%	7	10,26%	31	7,25%	5

En este caso, esta actividad sí que parece que ve reducido el número de casos ligeros, moderados y severos, sobre todo los ligeros. Además aumenta el caso de “NO EPISODE”, que se interpreta positivamente.

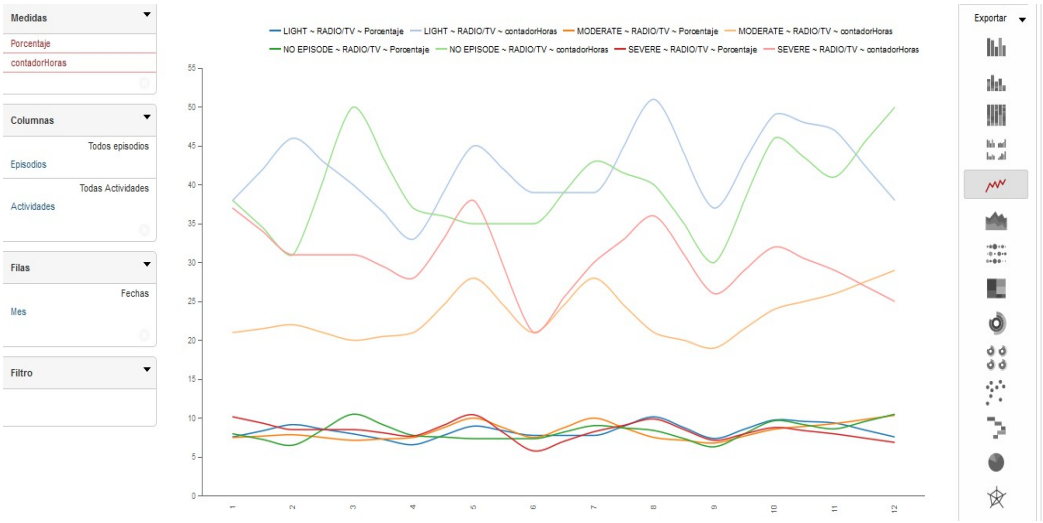
La gráfica muestra esta tendencia:



Actividad “RADIO/TV”:

Episodios	LIGHT		MODERATE		NO EPISODE		SEVERE		
Actividades	RADIO/TV		RADIO/TV		RADIO/TV		RADIO/TV		
Mes	Porcentaje	contadorHoras	Porcentaje	contadorHoras	Porcentaje	contadorHoras	Porcentaje	contadorHoras	
1	7,57%	38	7,50%	21	7,98%	38	10,16%	37	
2	9,16%	46	7,86%	22	6,51%	31	8,52%	31	
3	7,97%	40	7,14%	20	10,50%	50	8,52%	31	
4	6,57%	33	7,50%	21	7,77%	37	7,69%	28	
5	8,96%	45	10,00%	28	7,35%	35	10,44%	38	
6	7,77%	39	7,50%	21	7,35%	35	5,77%	21	
7	7,77%	39	10,00%	28	9,03%	43	8,24%	30	
8	10,16%	51	7,50%	21	8,40%	40	9,89%	36	
9	7,37%	37	6,79%	19	6,30%	30	7,14%	26	
10	9,76%	49	8,57%	24	9,66%	46	8,79%	32	
11	9,36%	47	9,29%	26	8,61%	41	7,97%	29	
12	7,57%	38	10,36%	29	10,50%	50	6,87%	25	

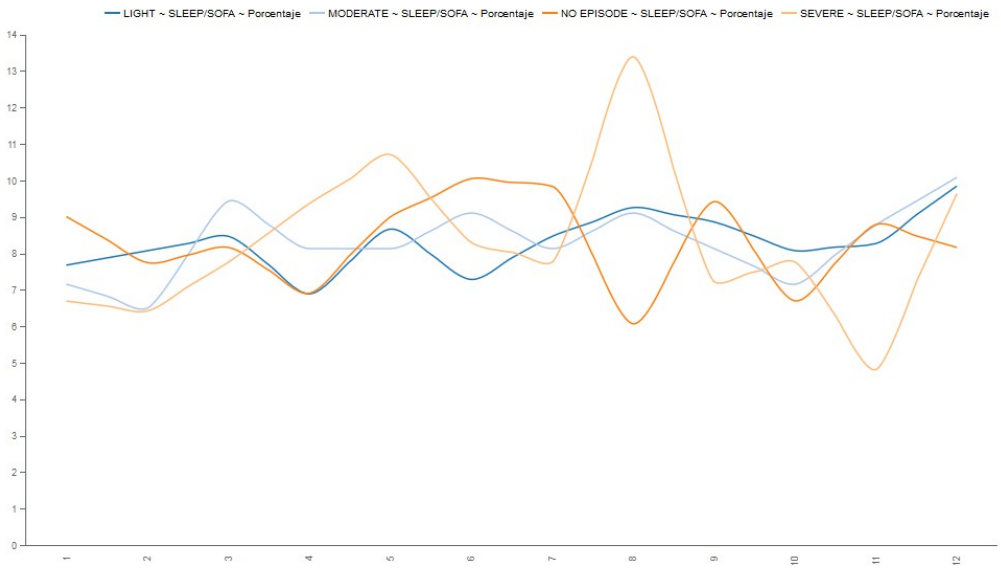
Si vemos la gráfica que proporciona Saiku, vemos que todas las lineas siguen el mismo patrón, por lo que no hayuna conclusión clara.



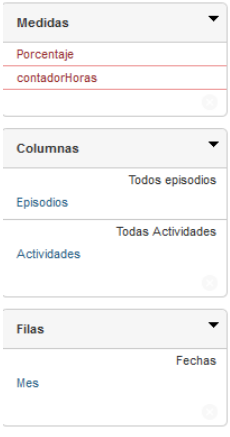
Actividad “SLEEP/SOFA”:

Episodios	LIGHT		MODERATE		NO EPISODE		SEVERE	
Actividades	SLEEP/SOFA		SLEEP/SOFA		SLEEP/SOFA		SLEEP/SOFA	
Mes	Porcentaje	contadorHoras	Porcentaje	contadorHoras	Porcentaje	contadorHoras	Porcentaje	contadorHoras
1	7,69%	39	7,17%	22	9,01%	43	6,70%	25
2	8,09%	41	6,51%	20	7,76%	37	6,43%	24
3	8,48%	43	9,45%	29	8,18%	39	7,77%	29
4	6,90%	35	8,14%	25	6,92%	33	9,38%	35
5	8,68%	44	8,14%	25	9,01%	43	10,72%	40
6	7,30%	37	9,12%	28	10,06%	48	8,31%	31
7	8,48%	43	8,14%	25	9,85%	47	7,77%	29
8	9,27%	47	9,12%	28	6,08%	29	13,40%	50
9	8,88%	45	8,14%	25	9,43%	45	7,24%	27
10	8,09%	41	7,17%	22	6,71%	32	7,77%	29
11	8,28%	42	8,79%	27	8,81%	42	4,83%	18
12	9,86%	50	10,10%	31	8,18%	39	9,65%	36

Con estos datos, parece que aumentan los episodios en todos los casos. Como complemento, los casos de “NO EPISODE” disminuyen.

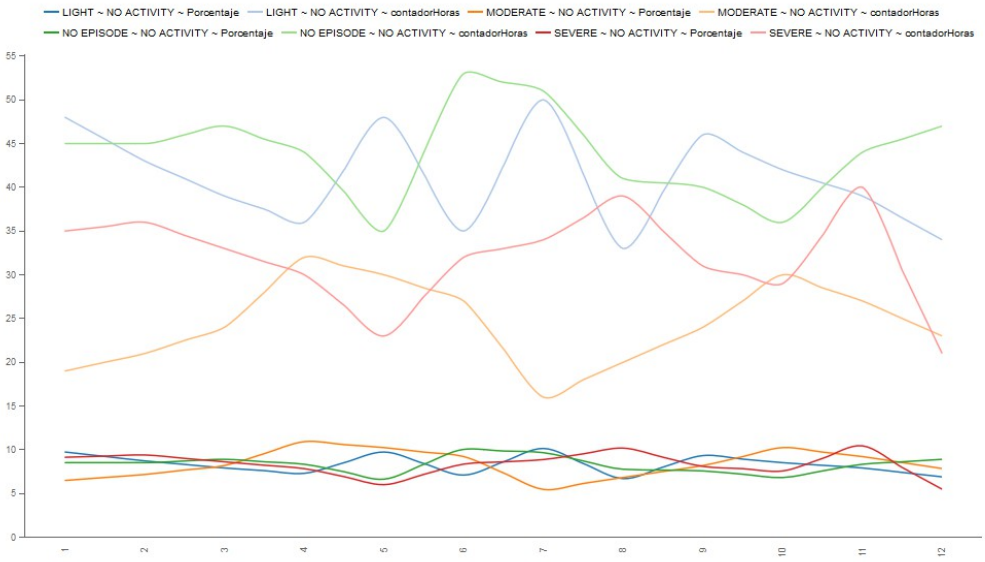


Actividad “NO ACTIVITY”:



Episodios	LIGHT		MODERATE		NO EPISODE		SEVERE	
Actividades	NO ACTIVITY		NO ACTIVITY		NO ACTIVITY		NO ACTIVITY	
Mes	Porcentaje	contadorHoras	Porcentaje	contadorHoras	Porcentaje	contadorHoras	Porcentaje	contadorHoras
1	9,74%	48	6,48%	19	8,52%	45	9,14%	35
2	8,72%	43	7,17%	21	8,52%	45	9,40%	36
3	7,91%	39	8,19%	24	8,90%	47	8,62%	33
4	7,30%	36	10,92%	32	8,33%	44	7,83%	30
5	9,74%	48	10,24%	30	6,63%	35	6,01%	23
6	7,10%	35	9,22%	27	10,04%	53	8,36%	32
7	10,14%	50	5,46%	16	9,66%	51	8,88%	34
8	6,69%	33	6,83%	20	7,77%	41	10,18%	39
9	9,33%	46	8,19%	24	7,58%	40	8,09%	31
10	8,52%	42	10,24%	30	6,82%	36	7,57%	29
11	7,91%	39	9,22%	27	8,33%	44	10,44%	40
12	6,90%	34	7,85%	23	8,90%	47	5,48%	21

Lano realización de actividad, parece que mejora los episodios ligeros y severos si se compara el mes 1 con el 12, pero existen altibajos en esta curva, por lo que no se puede concluir con rigurosidad que produzca mejora alguna en el paciente.



De las 6 pruebas que se han realizado, parece que la realizació de ejercicio es la que mejor resultado ofrece en la mejora de los pacientes.

3. Conclusiones

En este capítulo se exponen las conclusiones principales que se han extraído durante la realización de este proyecto.
La exposición se desglosa en conclusiones generales y conclusiones específicas de cada fase del proyecto.

En cuanto a las conclusiones generales, el proyecto ha sido satisfactorio, no es lo mismo realizar una práctica guiada que realizar un proyecto desde 0, desarrollando todas sus fases, desde la planificación hasta el cierre del proyecto. Los problemas a los que me he enfrentado durante la realización del proyecto, tales como gestión de cambio, malos diseños o decisiones incorrectas las valoro muy positivamente como experiencia adquirida en la realicación de proyectos BI.

Como conclusiones específicas de la parte de planificación del proyecto, en lineas generales creo que ha sido satisfactorio, teniendo en cuenta la dificultad que conlleva la planificación y desarrollo de un proyecto BI (aún siendo este a muy pequeña escala), se han cumplido las planificaciones de las fases especificadas en el diagrama de Gantt y los objetivos definidos en ellas.

En la parte de selección de software, la conclusión principal es que es una parte crítica que puede marcar la consecución satisfactoria del proyecto y/o facilitar su implementación, o puede ser uno de los factores que hagan fracasar el proyecto. Por ejemplo, en la selección del software de BI, los puntos principales que se han tenido en cuenta para su selección fueron:

- Software libre
- Ofrece funcionalidades necesarias para la realización del proyecto.
- Facilidad de instalación.
- Documentación disponible.
- Comunidad de soporte.

En este apartado, también es importante remarcar el tiempo que se requiere para la configuración del software a utilizar, definición de máquinas virtuales, instalación de software, comprobar que funciona correctamente, etc.

Otra de las conclusiones más importantes extraídas es que el diseño de la base de datos y la compresión de las fuentes de datos que se proporcionan es crucial para minimizar errores de implementación de capas superiores o procesos posteriores (ETLs, OLAP, etc). También es importante entender el objetivo de las tablas de hechos y de las tablas de dimensiones, así como de los esquemas de estrella y copo de nieve ya que en mi caso particular he confundido en ocasiones el diseño del DW con un diseño relacional tradicional.
Por ejemplo, para poder rellenar la tabla de hechos desde los procesos ETL, he tenido que apoyarme en tablas auxiliares en las que guardaba la información de cada fecha,paciente y actividad, para luego recorrerlas y poder insertar filas con todos los datos necesarios en la tabla de hechos. Esta parte en un

modelado inicial, la estaba realizando de forma incorrecta y en fases posteriores salían problemas a la hora de visualizar los datos en el cubo.

En cuanto al apartado de análisis de datos, la conclusión que saco es que he conseguido entender los conceptos principales de Cubo, Jerarquías, Niveles, etc y que he sido capaz de realizar consulas relativamente sencillas sobre el cubo cruzando dimensiones. Me hubiera gustado profundizar más en consultas más complicadas mediante el lenguaje MDX.
También, como complemento a futuro del análisis OLAP, se podría implementar un “Dashboard” utilizando la herramiento de pentaho CDE.

Hay un objetivo que no se ha cumplido, que es uno opcional sobre la implementación de un módulo de Machine Learning, tecnología en la que estoy bastante interesado. La principal razón por la que no se ha cumplido ha sido la ambigüedad a la hora de definir lo que habría que hacer, aunque esto ya se comenta en las fases iniciales del proyecto. También la falta de tiempo sirve un poco de excusa, por lo que puede caber como una linea de investigación futura.

Otra linea a futuro que creo que complementaría muy bien este proyecto es la introducción de frameworks de cálculo distribuído, ya que es una parte fundamental de entornos Big Data y en un desarrollo local, no se ha podido poner en marcha. Esta parte distribuída podría incluir Hadoop o Spark, y distribuir la base de datos en nodos, etc, lo que añade bastante complejidad de configuración de sistemas y entonos distribuídos.

4. Glosario

En la memoria no se utilizan acrónimos o se indica su significado inmediatamente después de su utilización para facilitar la comprensión al lector.

5. Bibliografía

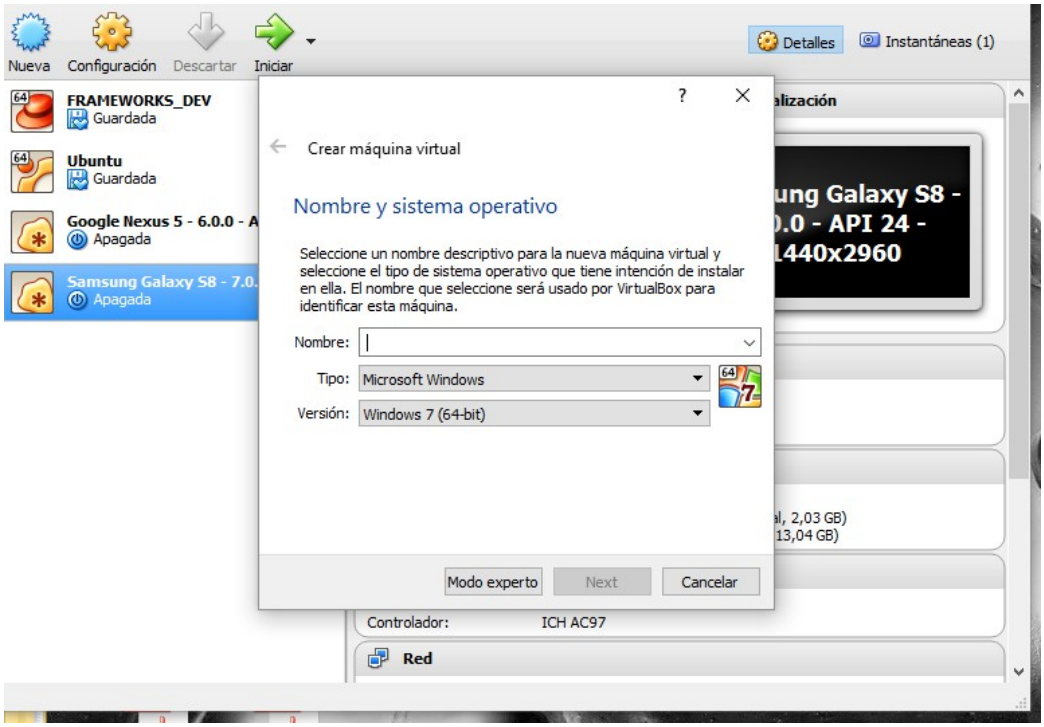
La mayoría de los recursos consultados son recuersos online utlizados en el anexo I, donde se explican los procesos para la instalación y configuración del software utlizado en el proyecto.
Además, en la sección de selección de software, se ofrece una lista de enlaces a sitios web de donde se ha extraído información relevante para la selección del software. Asímismo, tales enlaces se han dejado indicados en la propia sección del documento para un mejor seguimiento por parte del lector.

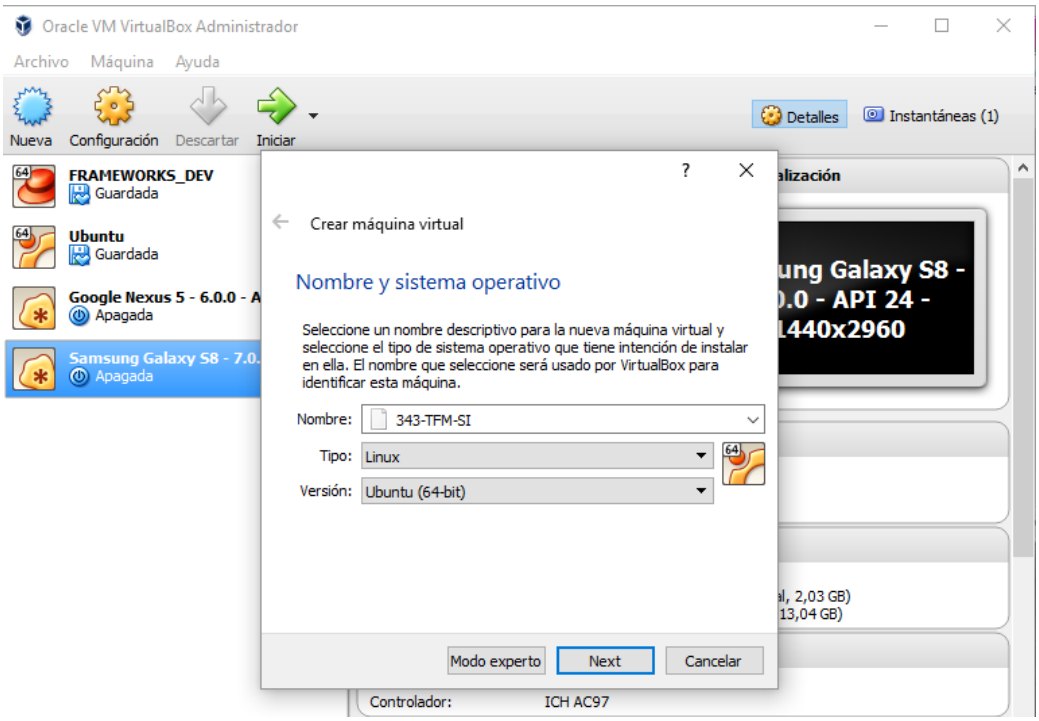
6. Anexos

En el siguiente anexo se va a proporcionar una serie de instrucciones para la instalación del entorno en el que se va a desarrollar el proyecto.
Se mostrarán los pasos necesarios para instalar una máquina virtual quecontenga todos los elementos software para la consecución del proyecto, PostgreSQL, Java, Pentaho, etc.

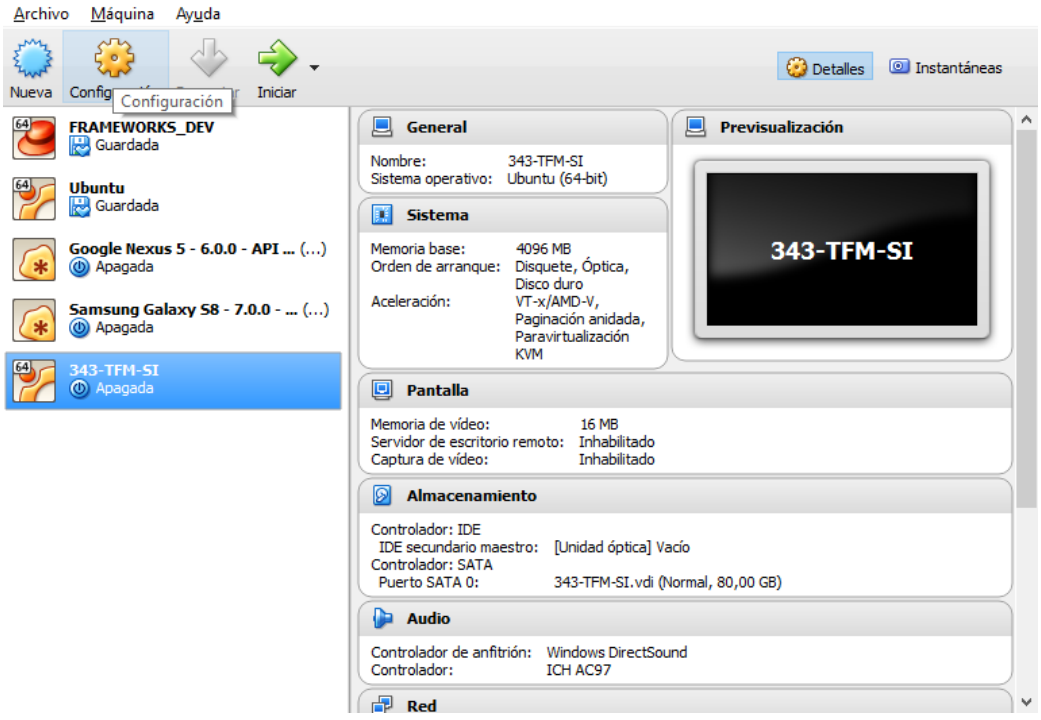
Instalación de la máquina virtual

Utilizando el software VirtualBox de Oracle, definimos una nueva máquina virtual con sistema operativo Linux.

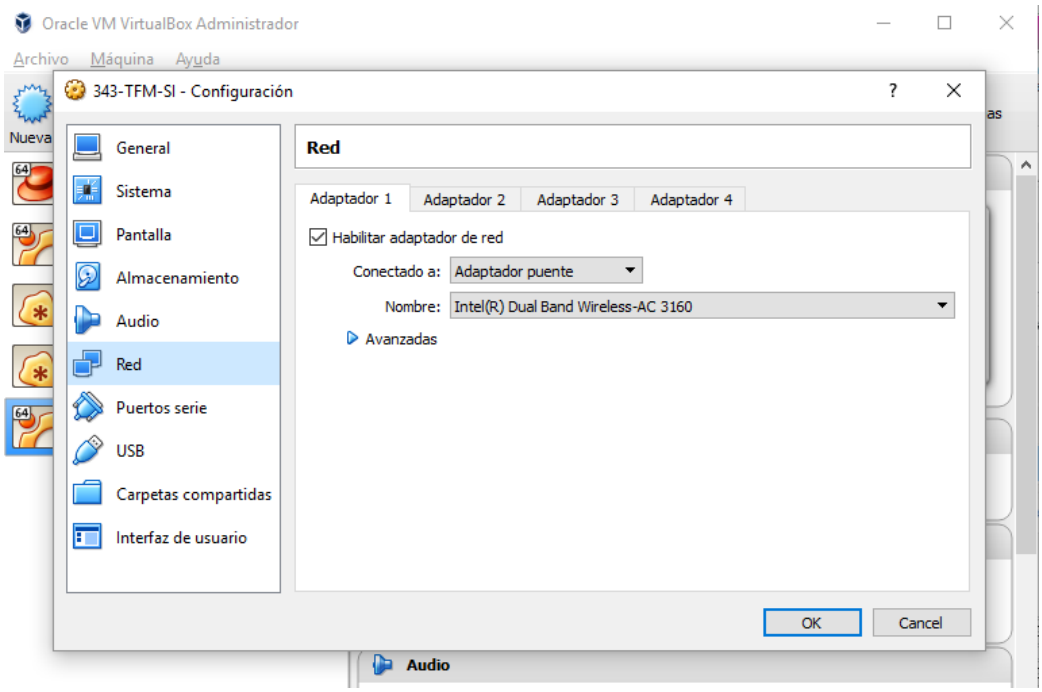




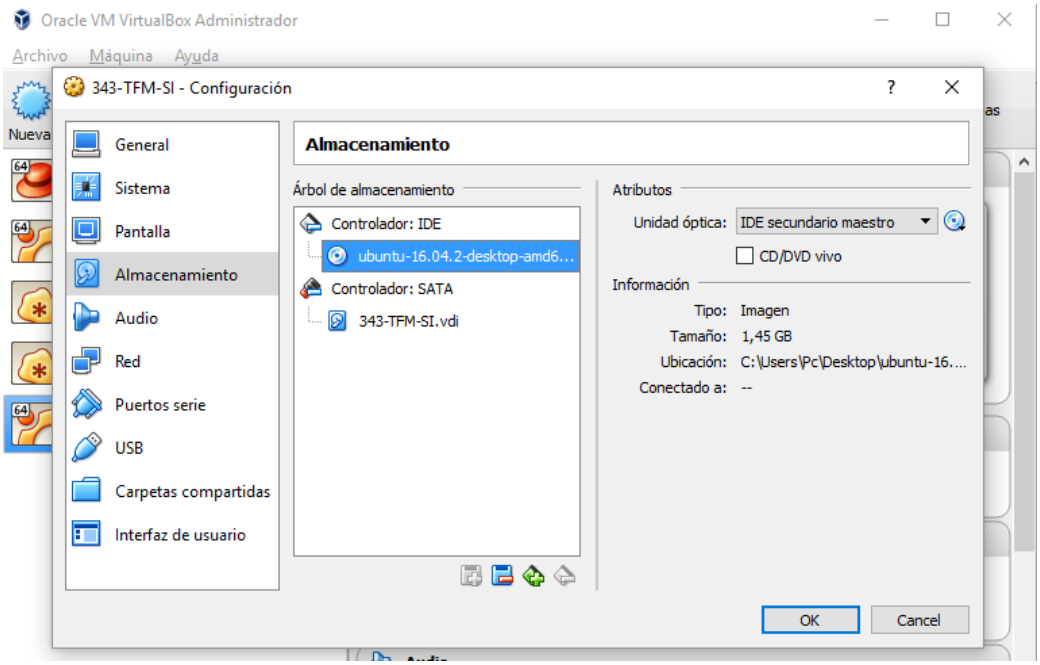
A continuación se procede a la asignación de memoria y disco, 4GB de RAM y 80GB de disco.



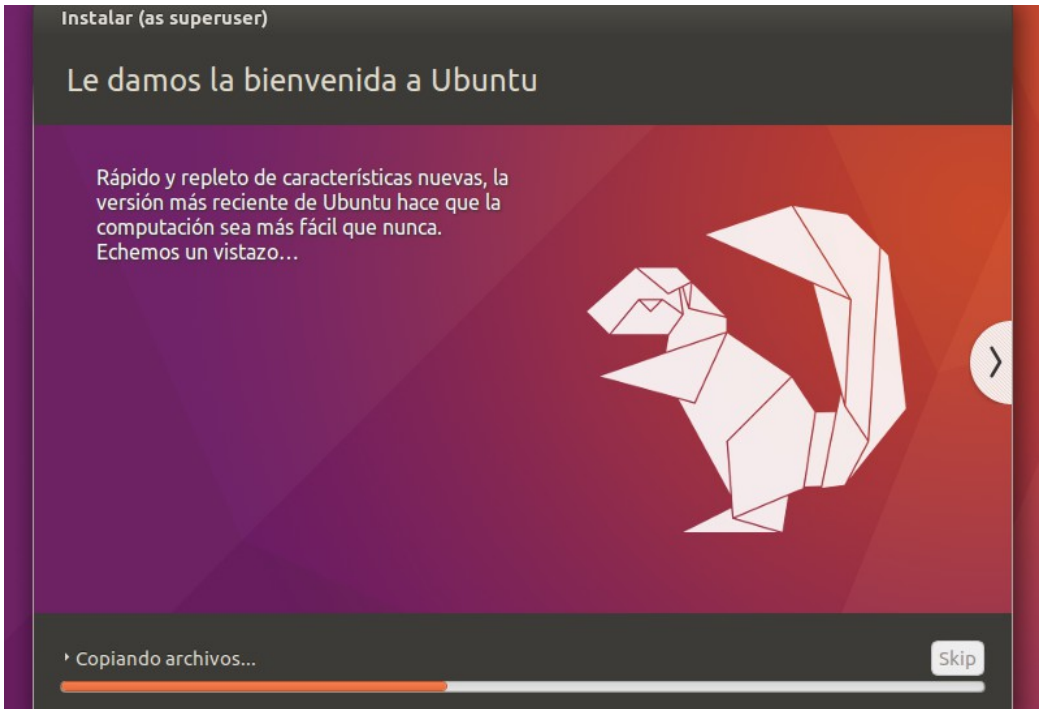
Configuración red como bridge para el acceso a internet.



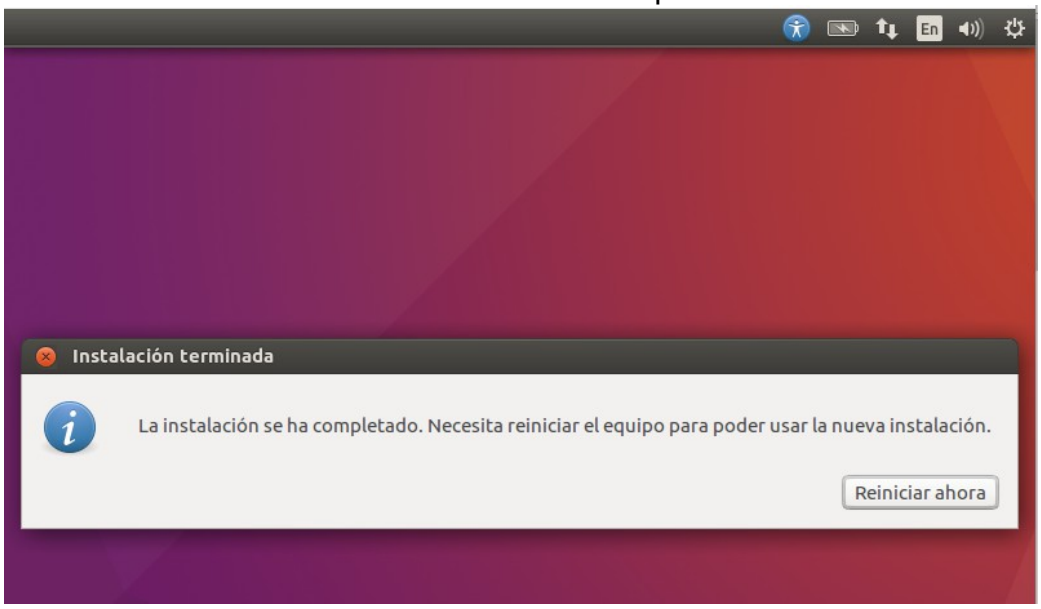
Para poder instalar el sistema operativo hay que añadirl a ISO de Ubuntu 16.04 TLS a la unidad óptica que simula un lector de DVD 'real'.



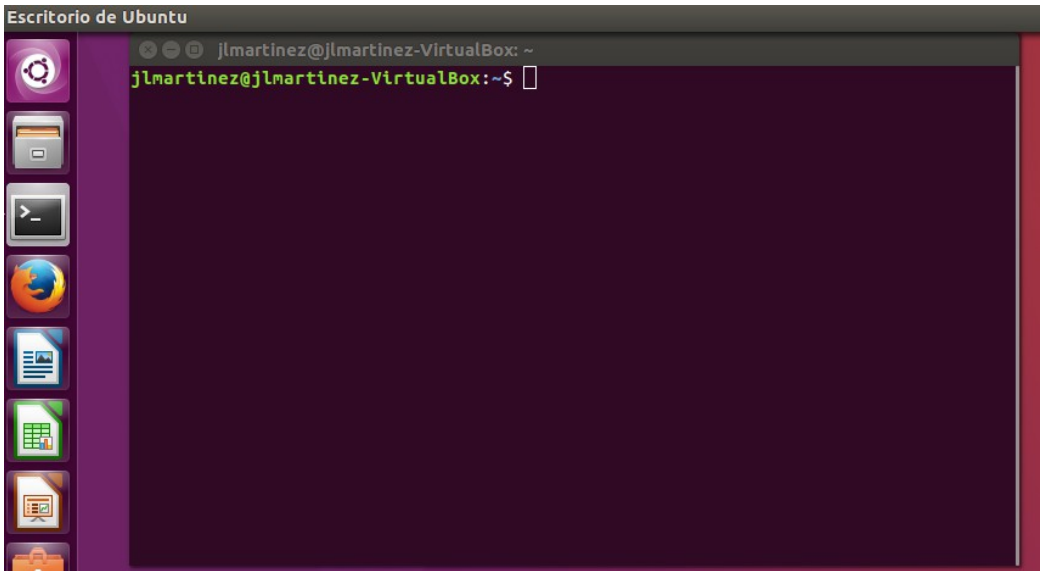
A continuación, se arranca la máquina virtual y se observa el proceso de instalación del SO.



Finalización de la instalación del sistema operativo.



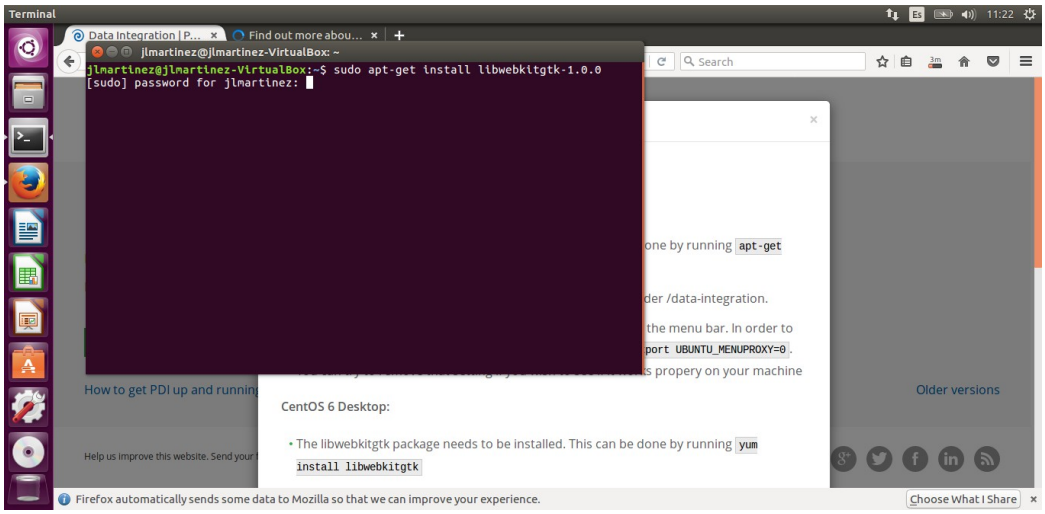
Comprobación de entrada al sistema.


Escritorio de Ubuntu
jlmartinez@jlmartinez-VirtualBox: ~
jlmartinez@jlmartinez-VirtualBox:~$

Intalación de Pentaho Data Integration

http://community.pentaho.com/projects/data-integration/
https://help.pentaho.com/Documentation/5.2/0F0/0G0/020

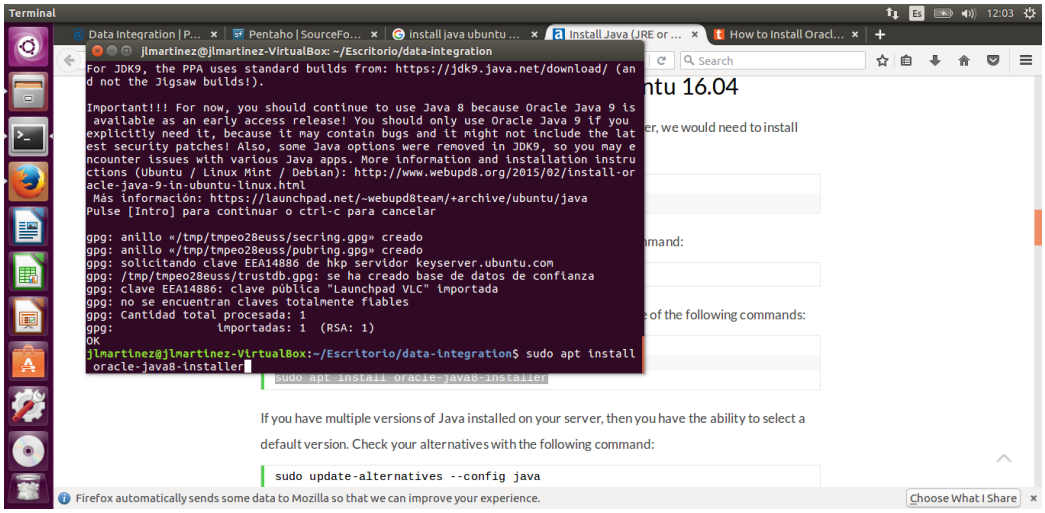
Instalar la librería requerida



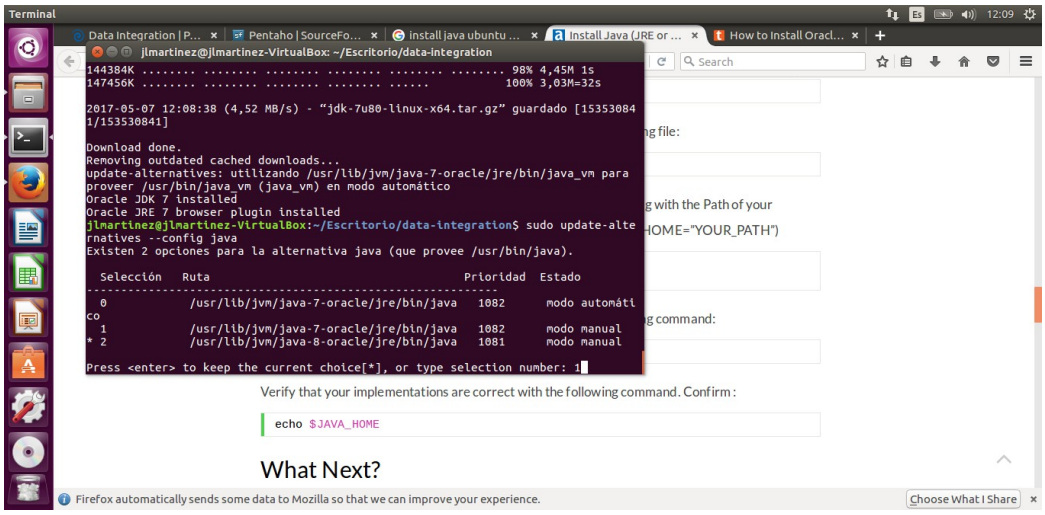
Instalar Java

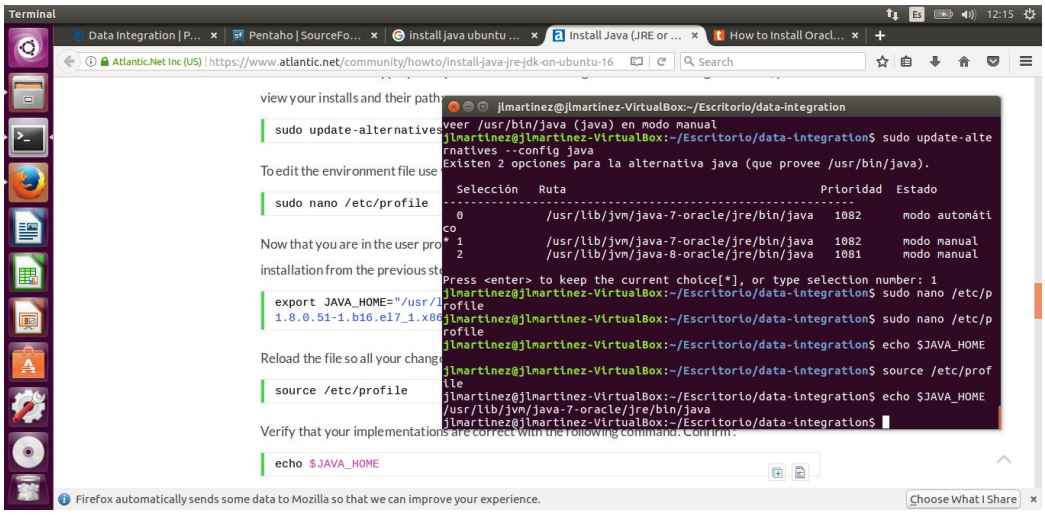
https://www.atlantic.net/community/howto/install-java-jre-jdk-on-ubuntu-16-04/

Siguiendo este tutorial, se ha optado por la instalación de Oracle JDK y se instalará Java 8.



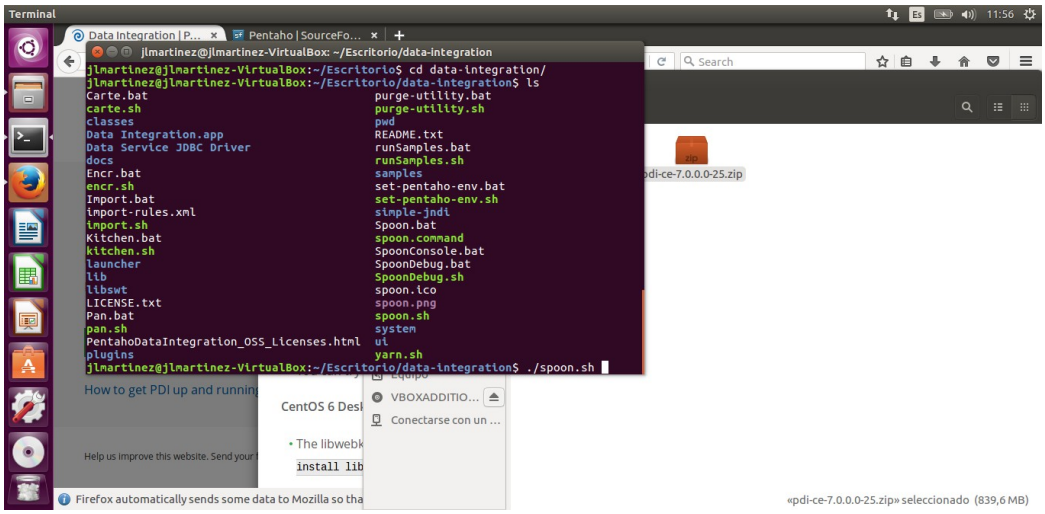
Seleccionamos Java 8, porque Pentaho 7.0 está compilado con Java 8 y si se selecciona Java 7 da un error de 'major minor versión' en la JVM de Java.

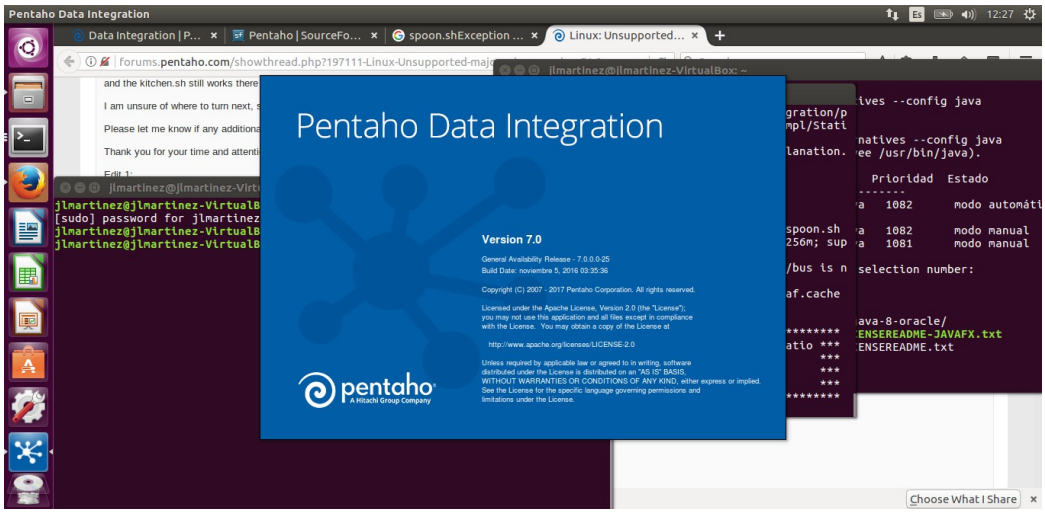

Terminal
12:09
Data Integration | P...
Pentaho | SourceFo...
install java ubuntu ...
Install Java (JRE or ...
How to Install Oracl...
jlmartinez@jlmartinez-VirtualBox: ~/Escritorio/data-integration
144384K 98% 4,45M 1s
147456K .... 100% 3,03M=32s
2017-05-07 12:08:38 (4,52 MB/s) - “jdk-7u80-linux-x64.tar.gz” guardado [153530841/153530841]
Download done.
Removing outdated cached downloads...
update-alternatives: utilizando /usr/lib/jvm/java-7-oracle/jre/bin/java_vm para proveer /usr/bin/java_vm (java_vm) en modo automático
Oracle JDK 7 installed
Oracle JRE 7 browser plugin installed
jlmartinez@jlmartinez-VirtualBox:~/Escritorio/data-integration$ sudo update-alternatives --config java
Existen 2 opciones para la alternativa java (que provee /usr/bin/java).
Selección Ruta Prioridad Estado
0 /usr/lib/jvm/java-7-oracle/jre/bin/java 1082 modo automático
1 /usr/lib/jvm/java-7-oracle/jre/bin/java 1082 modo manual
* 2 /usr/lib/jvm/java-8-oracle/jre/bin/java 1081 modo manual
Press <enter> to keep the current choice[*], or type selection number: 1
Verify that your implementations are correct with the following command. Confirm :
echo $JAVA_HOME
What Next?
Firefox automatically sends some data to Mozilla so that we can improve your experience.
Choose What I Share


Terminal
12:15
Data Integration | P...
Pentaho | SourceFo...
install java ubuntu ...
Install Java (JRE or ...
How to Install Oracl...
Atlantic.Net Inc (US) https://www.atlantic.net/community/howto/install-java-jre-jdk-on-ubuntu-16
Search
view your installs and their path
sudo update-alternatives
To edit the environment file use
sudo nano /etc/profile
Now that you are in the user pro
installation from the previous st
export JAVA_HOME="/usr/l
1.8.0.51-1.b16.el7_1.x86
Reload the file so all your chang
source /etc/profile
Verify that your implementations are correct with the following command. Confirm :
echo $JAVA_HOME
jlmartinez@jlmartinez-VirtualBox:~/Escritorio/data-integration
veer /usr/bin/java (java) en modo manual
jlmartinez@jlmartinez-VirtualBox:~/Escritorio/data-integration$ sudo update-alternatives --config java
Existen 2 opciones para la alternativa java (que provee /usr/bin/java).
Selección Ruta Prioridad Estado
0 /usr/lib/jvm/java-7-oracle/jre/bin/java 1082 modo automático
* 1 /usr/lib/jvm/java-7-oracle/jre/bin/java 1082 modo manual
2 /usr/lib/jvm/java-8-oracle/jre/bin/java 1081 modo manual
Press <enter> to keep the current choice[*], or type selection number: 1
jlmartinez@jlmartinez-VirtualBox:~/Escritorio/data-integration$ sudo nano /etc/profile
jlmartinez@jlmartinez-VirtualBox:~/Escritorio/data-integration$ sudo nano /etc/profile
jlmartinez@jlmartinez-VirtualBox:~/Escritorio/data-integration$ echo $JAVA_HOME
jlmartinez@jlmartinez-VirtualBox:~/Escritorio/data-integration$ source /etc/profile
jlmartinez@jlmartinez-VirtualBox:~/Escritorio/data-integration$ echo $JAVA_HOME
/usr/lib/jvm/java-7-oracle/jre/bin/java
jlmartinez@jlmartinez-VirtualBox:~/Escritorio/data-integration$
Firefox automatically sends some data to Mozilla so that we can improve your experience.
Choose What I Share

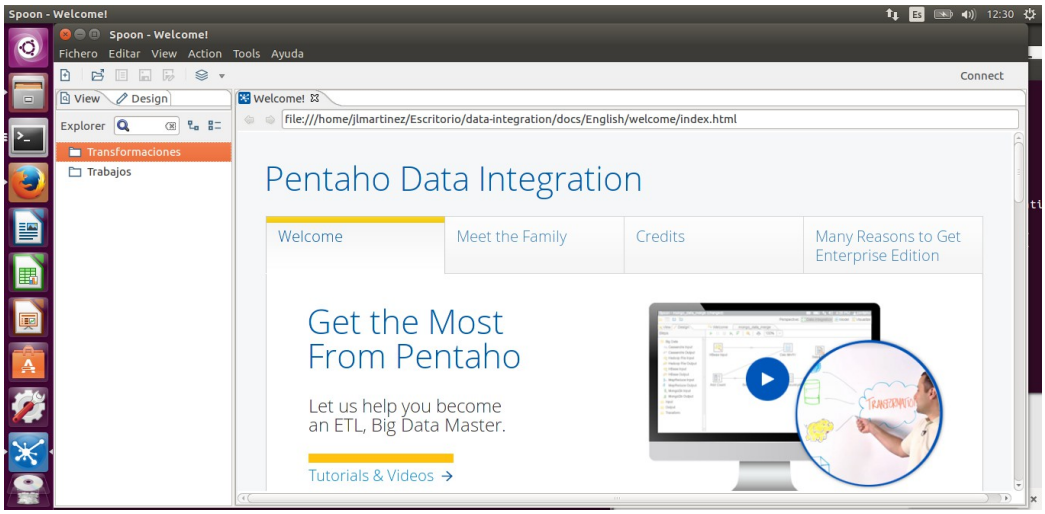
Instalar Pentaho Data Integration

Establecer variables de entorno en el fichero /etc/environment
export PENTAHO_JAVA_HOME=/usr/lib/jvm/java-8-oracle





El arranque del programa es correcto.


Spoon - Welcome!
12:30
Spoon - Welcome!
Fichero Editar View Action Tools Ayuda
Connect
View
Design
Explorer
Transformaciones
Trabajos
Welcome!
file:///home/jlmartinez/Escritorio/data-integration/docs/English/welcome/index.html
Pentaho Data Integration
Welcome
Meet the Family
Credits
Many Reasons to Get Enterprise Edition
Get the Most From Pentaho
Let us help you become an ETL, Big Data Master.
Tutorials & Videos

Instalar PostgreSQL

//http://ideafalaz.blogspot.com.es/2016/04/instalar-postgresql-y-pgadmin-en-linux.html
//https://www.digitalocean.com/community/tutorials/how-to-install-and-use-postgresql-on-ubuntu-16-04

Ahora se va a proceder a instalar tanto PostgreSQL como PgAdmin 3 para poder crear las tablas del almacén de datos y gestionar la consulta de datos.

```
jlmartinez@jlmartinez-VirtualBox:~$ sudo apt-get install postgresql postgresql-c
ontrib
```

```
Configurando postgresql (9.5+173) ...
Configurando postgresql-contrib-9.5 (9.5.6-0ubuntu0.16.04) ...
Configurando postgresql-contrib (9.5+173) ...
Configurando sysstat (11.2.0-1ubuntu0.1) ...

Creating config file /etc/default/sysstat with new version
update-alternatives: utilizando /usr/bin/sar.sysstat para proveer /usr/bin/sar (
sar) en modo automático
Procesando disparadores para libc-bin (2.23-0ubuntu5) ...
Procesando disparadores para systemd (229-4ubuntu16) ...
Procesando disparadores para ureadahead (0.100.0-19) ...
jlmartinez@jlmartinez-VirtualBox:~$
```

Modificamos el password del usuario postgres y lo asignamos con el valor 'jlmartinez'.

```
jlmartinez@jlmartinez-VirtualBox:~$ sudo -u postgres psql postgres
psql (9.5.6)
Type "help" for help.

postgres=# \password postgres
Enter new password:
Enter it again:
postgres=#
```

Salimos con el comando \q.

En el siguiente paso vamos a crear un usuario diferente a postgres, que es el usuario por defecto.

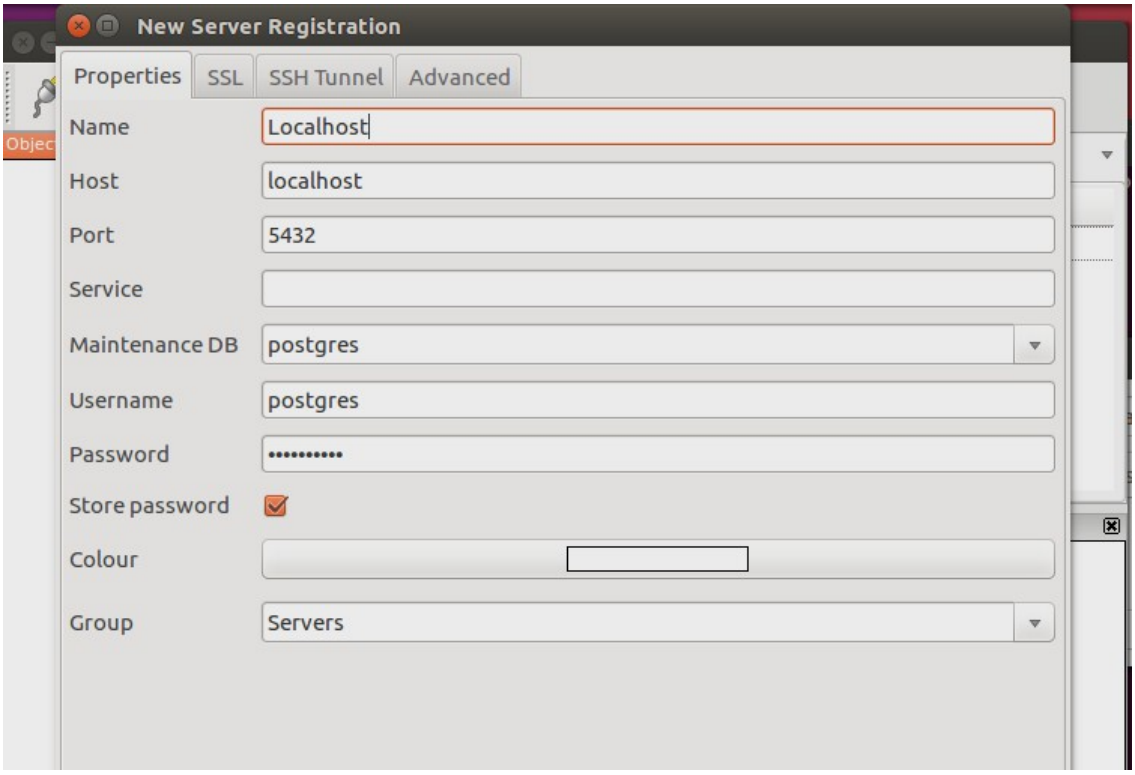
```
jlmartinez@jlmartinez-VirtualBox:~$ sudo -u postgres createuser --interactive
Enter name of role to add: jlmartinez
Shall the new role be a superuser? (y/n) y
jlmartinez@jlmartinez-VirtualBox:~$
```

Instalamos el gestor gráfico de la base de datos pgAdmin3:

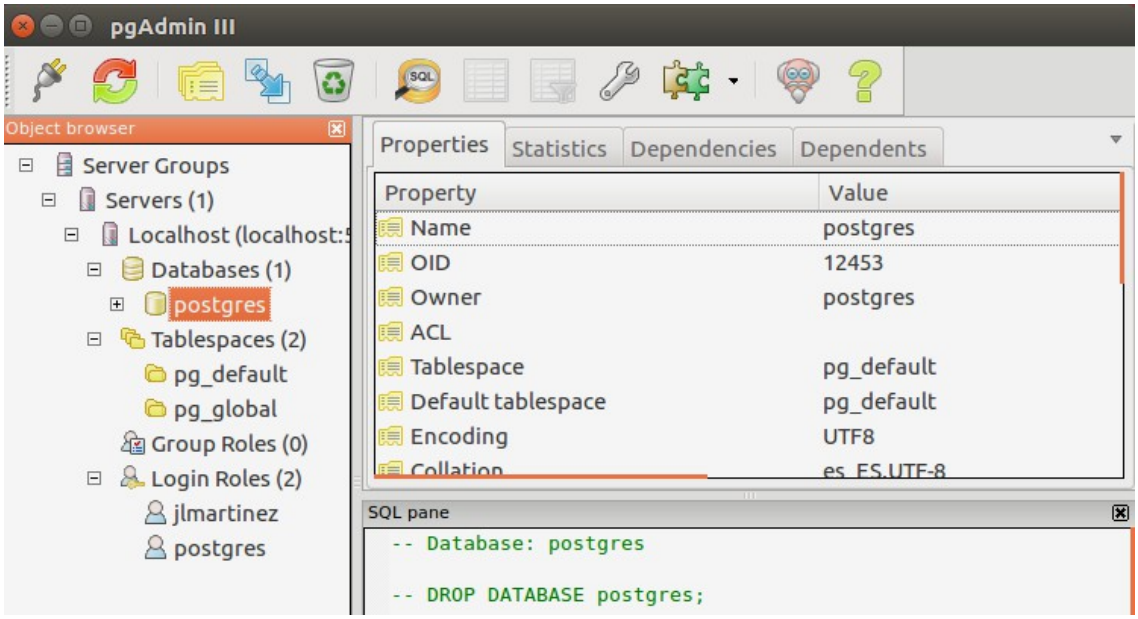
```
jlmartinez@jlmartinez-VirtualBox:~$ sudo apt-get install pgadmin3
Leyendo lista de paquetes... Hecho
Creando árbol de dependencias
Leyendo la información de estado... Hecho
```

Ahora, vamos a arrancar el programa y lo añadiremos a los accesos directos en la barra izquierda del escritorio de Ubuntu arrastrándolo con el ratón.

Arrancamos el programa y creamos una nueva conexión, a la que será nuestra base de datos de nuestro data warehouse.



Y podemos ver que tenemos una base de datos, y dos usuarios, postgres y el creado anteriormente.

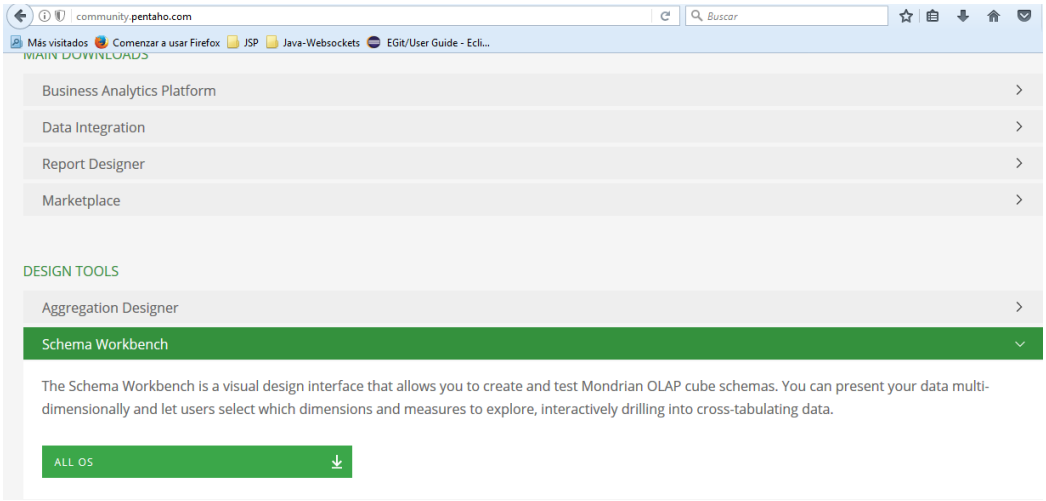


Con esto ya tenemos instalado tanto el gestor de la base de datos como una herramienta gráfica para su manejo y consulta.

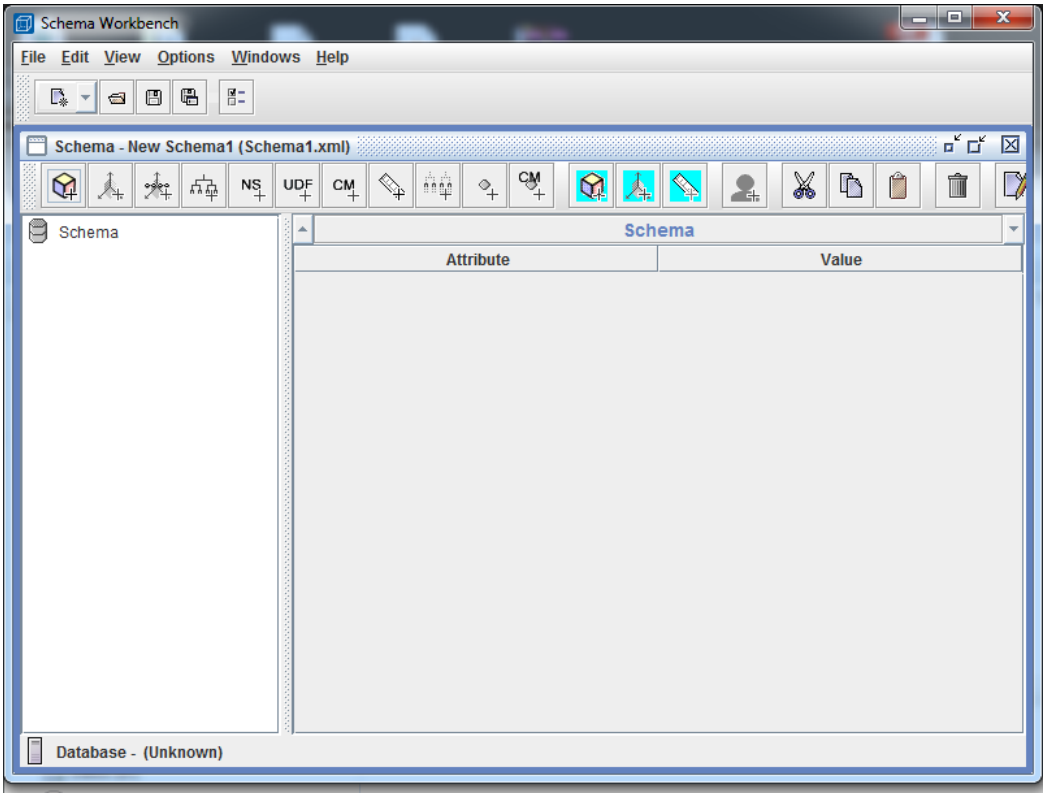
Instalar Pentaho Schema Workbench 7.1

Pentaho Schema Workbench es un software para confeccionar cubos OLAP que posteriormente serán analizados por el paquete Mondrian desde el servidor de Pentaho Community Edition.

Para su instalación hay que descargarse el paquete psw-ce-3.14.0.0-12.zip desde la web de Pentaho

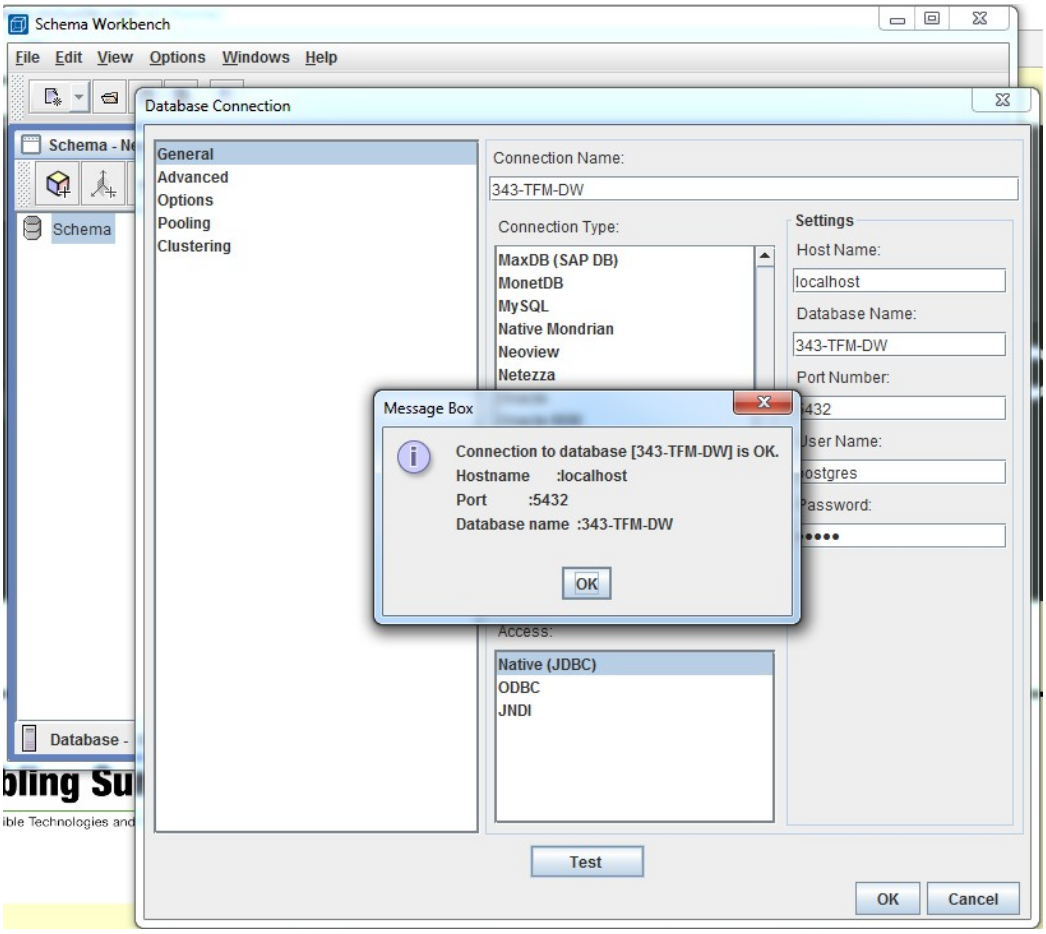


Una vez descargado el zip, se descomprime y se arranca el script workbench.sh para sistemas Linux ó workbench.bat para windows.



Lo siguiente que hay que hacer es definir una conexión a la base de datos.

Para ello vamos a 'Options-→ Connection…' y seleccionamos nuestra conexión a la BBDD de PostgreSQL con nuestro almacén de datos.

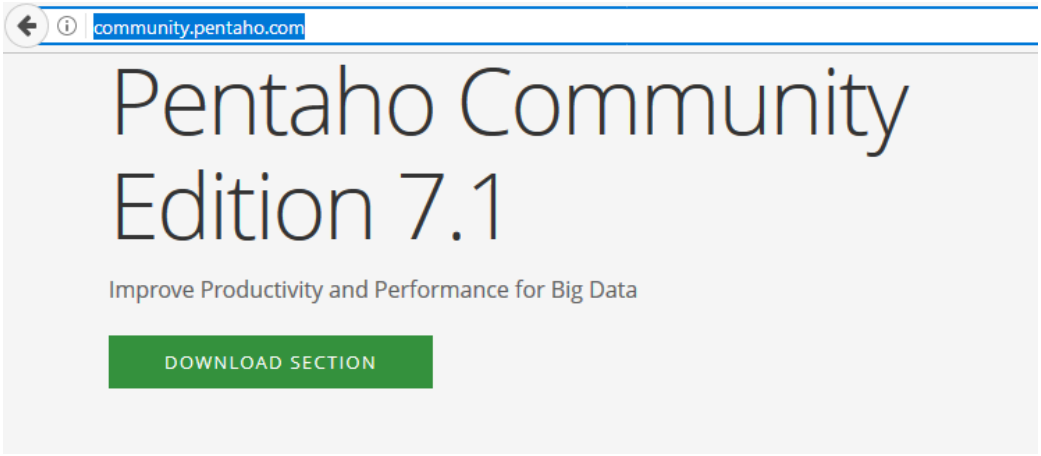


Con esto, ya tenemos instalado y configurado nuestro Pentaho Schema Workbench.

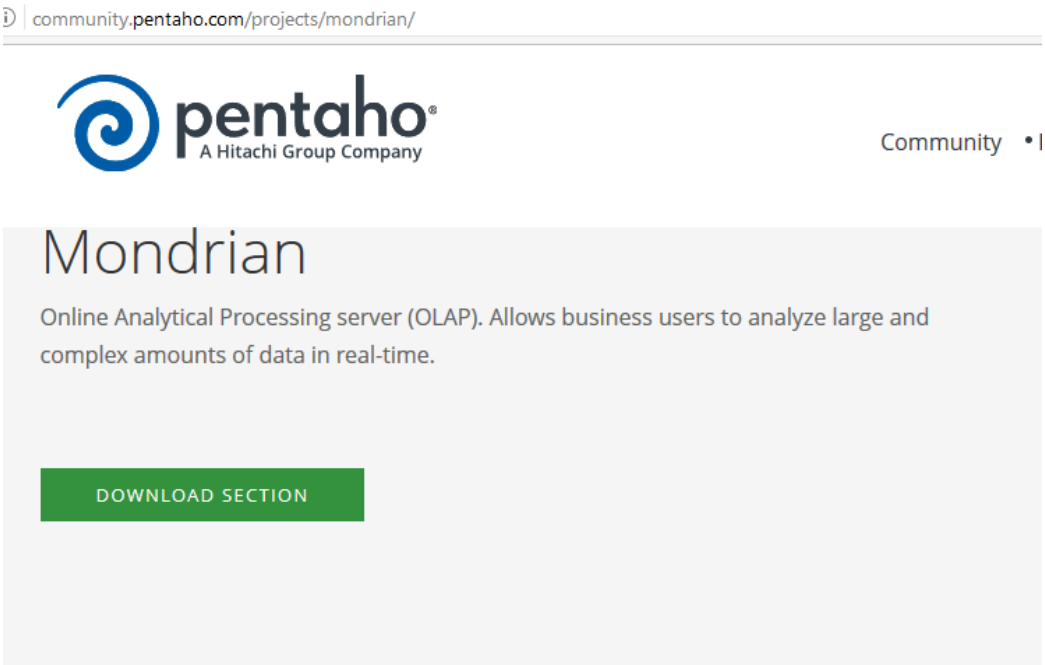
Instalar Pentaho Server Community Edition 7.1

Pentaho Server Community Edition es un software para realizar análisis de datos y análisis OLAP en un entorno web. El paquete contiene internamente el módulo mondrian que es el motor de análisis OLAP de pentaho.

http://community.pentaho.com/



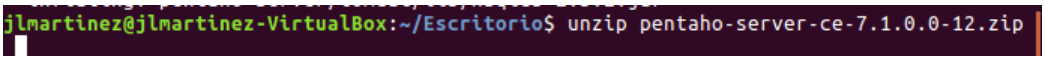
http://community.pentaho.com/projects/mondrian/

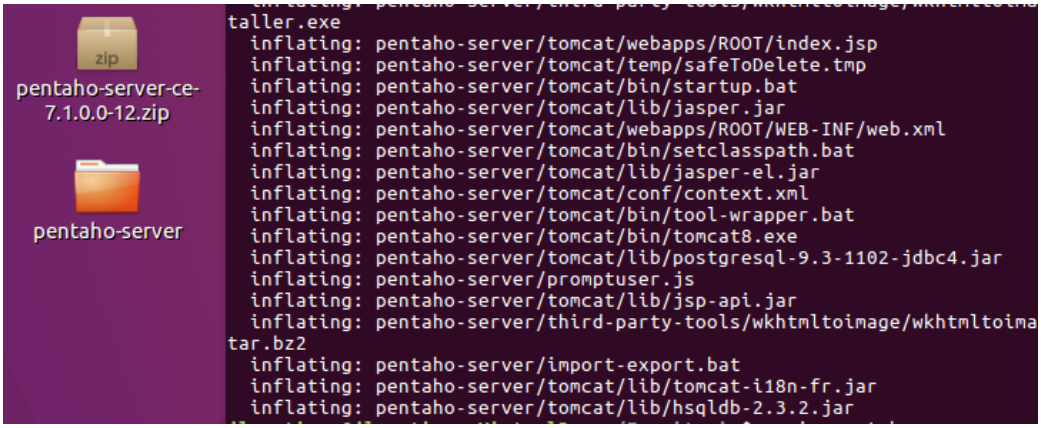


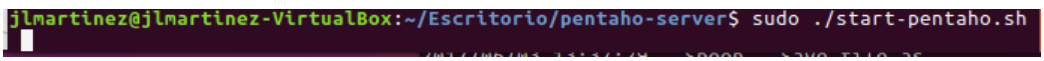
A continuación se explica el proceso de instalación de la herramienta para poder utilizarla en el contexto del proyecto.

Descargamos el software desde la sección downloads, seleccionamos el módulo Business Analytics Platform y esto redirecciona a SourceForge (https://sourceforge.net/projects/pentaho), donde hay que descargar un fichero ZIP.

http://fcorti.com/2016/12/05/install-pentaho-business-analytics-platform-7/

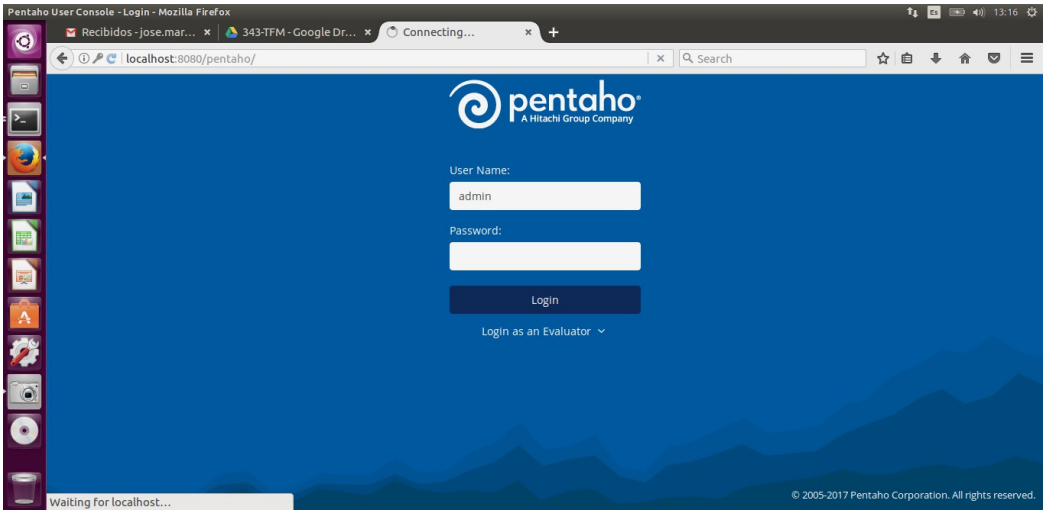


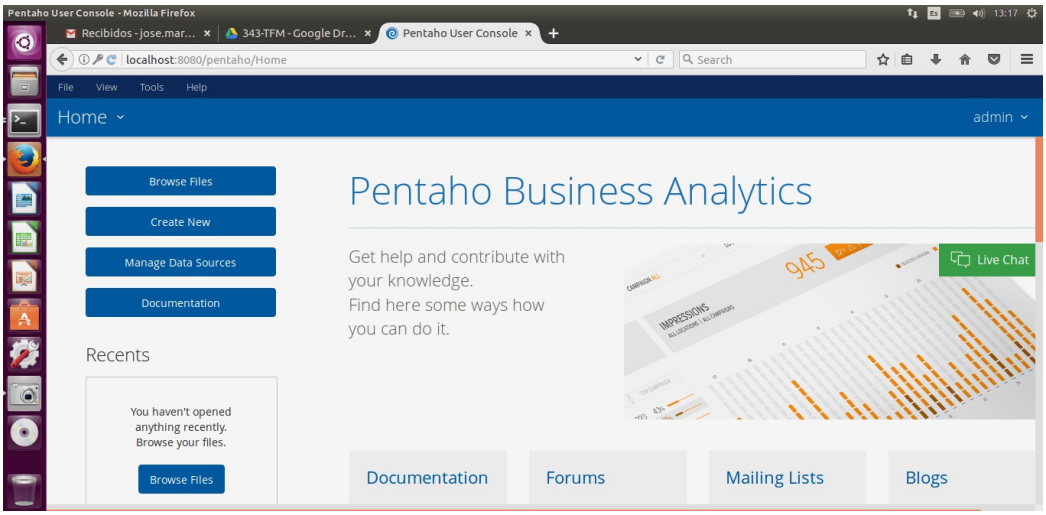




```
13:13:25,750 INFO  [PeriodicStatusLogger] Caution, the system is initializing. Do not
 shut down or restart the system at this time.
13:13:55,753 INFO  [PeriodicStatusLogger] Caution, the system is initializing. Do not
 shut down or restart the system at this time.
13:14:25,753 INFO  [PeriodicStatusLogger] Caution, the system is initializing. Do not
 shut down or restart the system at this time.
13:14:27,949 INFO  [PeriodicStatusLogger] The system has finished initializing.
Pentaho BI server listo.
```

User: admin
Pass: password



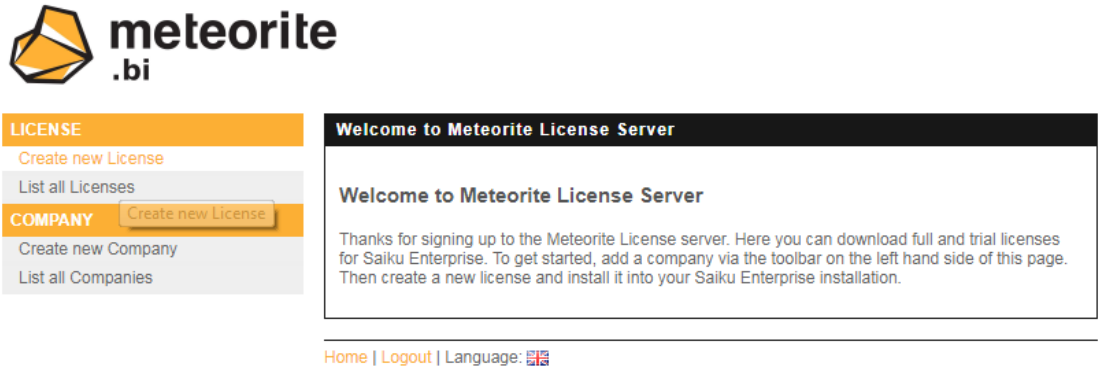


Instalar plugin Saiku

Se ha seguido este tutorial para su instalación:
http://edpflager.com/?p=3322

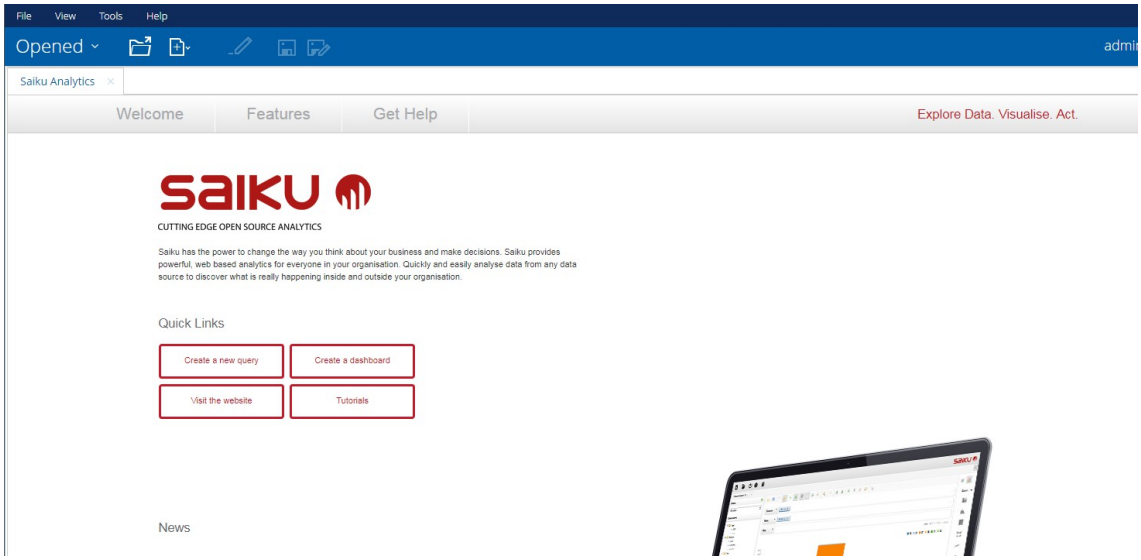
Los pasos que hay que realizar son sencillos y son los de la siguiente lista:

- Buscar en el marketplace de Pentaho el plugin Saiku Analytics y descargarlo.
- Descomprimirlo en /biserver-ce/pentaho-solutions/system.
- Acceder a http://licensing.meteorite.bi/signups?form y registrarse.
- Crear una licencia CE



- Copiar la licencia en la carpeta /biserver-ce/pentaho-solutions/system/saiku
- Renombrar la licencia a 'license.lic'
- Reiniciar el servidor Pentaho.

Tendremos disponible una nueva opción para crear un análisis Saiku:



Definimos un nuevo “Data Source” contra la base de datos de Postgresql y probamos la conexión.