Sistema de inteligencia de negocio para el análisis de los tratamientos de reducción del colesterol

Leidy Tatiana Vélez Londoño

Máster en Ingeniería Informática
TFM-Business Intelligence

David Amorós Alcaraz

Ferran Prados Carrasco

Enero de 2019

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Sistema de inteligencia de negocio para el análisis de los tratamientos de reducción del colesterol*
Nombre del autor:	*Leidy Tatiana Vélez Londoño*
Nombre del consultor/a:	*David Amorós Alcaraz*
Nombre del PRA:	*Ferran Prados Carrasco*
Fecha de entrega (mm/aaaa):	01/2019
Titulación:	*Máster en Ingeniería Informática*
Área del Trabajo Final:	TFM-Business Intelligence
Idioma del trabajo:	*Español*
Palabras clave	*Business Intelligence o Inteligencia de negocio, Colesterol LDL, pacientes.*

Resumen del Trabajo (máximo 250 palabras):

En este proyecto se aplican los conceptos relacionados con Business Intelligence con enfoque al sector salud. En particular, el diseño de un sistema que tiene como objetivo analizar la efectividad de los tratamientos para la reducción del colesterol. La metodología utilizada para la construcción de este es una adaptación al modelo de Tareas de la metodología de Kimball, denominada Business Dimensional Lifecycle, donde se desarrolla el análisis dimensional, el diseño físico, los procesos de extracción, transformación y carga y finalmente la implementación.

Se realiza el proceso de selección de la tecnología requerida para el análisis, dando como resultado a Pentaho y todas sus herramientas. Y se definen e implementan mecanismos para procesar en línea la información a través de la técnica OLAP, habilitando las siguientes conclusiones:

I. Los pacientes sometidos en un lapso determinado a un tratamiento homeopático presentan un incremento en los niveles de colesterol.
II. La terapia más eficaz para la diminución del colesterol, y a su vez unos valores de presión estables es la natural.
III. Los pacientes cuyo tratamiento es farmacológico y cuyos hábitos son sedentarios son más propensos a un alza del colesterol.
IV. Se presentan diferencias no representativas en el uso de un tratamiento en una misma ubicación geográfica del paciente.

Aunque son diversas las causas que llevan a que el colesterol se incremente más en unas personas que en otras, seguir una dieta sana (como la mediterránea) y hacer ejercicio físico a diario mejora los niveles de presión arterial y colesterol LDL.

Abstract (in English, 250 words or less):

The Business Intelligence (BI) concepts are applied in this project that focuses on healthcare area. Particularly, the design of a BI system in order to analyze the effectiveness for reduction of LDL cholesterol treatments. The methodology used to build the system is based on Kimball's Business Dimensional Lifecycle approach of Kimball, and It is mentioned about dimensional analysis, physical design, extraction, transformation and load process and finally the BI system implementation.

The open source BI technology selected it was Pentaho and all software needed and compatible with it. Online analytical processing (OLAP) can be used for analyzing information and in this project, it was development an OLAP cube that permitted to make the following conclusions:

I. Patients with a homeopathic treatment in a period have an increase in LDL cholesterol.
II. Natural therapy is extremely effective to decrease the LDL cholesterol and blood pressure.
III. The pharmacological treatment causes an increase in the LDL cholesterol in patients with sedentarism.
IV. There aren't relevant differences in the Cholesterol LDL level and blood pressure with the use of a treatment in the same city.

Although there are many different causes for high cholesterol, to have a healthy diet (like Mediterranean), and to do exercise every day can be reduce the LDL Cholesterol and be better the blood pressure indicators.

Índice

Lista de figuras

Lista de Abreviaturas

ACV	Accidente cerebro vascular.
BI	Inteligencia de negocio o Business Intelligence en inglés.
CEP	Complex Event Processing Engine.
CMD	Símbolo de sistema de Windows o Command prompt en inglés.
DW	Almacén de datos o Data Warehouse en inglés.
DSS	Sistema de soporte a las decisiones o Decision Support System en inglés.
EDT	Estructura de descomposición del trabajo.
ETL	Extraer, transformar y cargar o Extract, Transform and Load en inglés.
HDL	Lipoproteínas de alta densidad o high density lipoproteinsen inglés.
LDL	Lipoproteínas de baja densidad o Low density lipoproteins en inglés.
MDX	Multidimensional expressions.
MTC	Medicina Tradicional China.
OLAP	On-Line Analytical Processing.
OWL	Lenguaje de Ontologías para la Web o Web Ontology Language en inglés.
PDI	Pentaho Data Integration.
PSW	Pentaho Schema Workbench.
RDBMS	Relational Database Management System.

1. Introducción

En este capítulo se presentan los antecedentes del objeto de estudio, la justificación y el entendimiento del problema. Asimismo, se exponen los objetivos del proyecto.

1.1. Antecedentes del proyecto

¿Qué es el colesterol?

El colesterol es un lípido (grasa). Se forma en el hígado a partir de alimentos grasos y es necesario para el funcionamiento normal del organismo. El colesterol está presente en la membrana plasmática (capa exterior) de todas las células del organismo. [1]

El colesterol se desplaza por la sangre mediante unas moléculas denominadas lipoproteínas. Los tres tipos principales son:

a) Las lipoproteínas de baja densidad (LDL) o “colesterol malo”; se cree que causan enfermedades arteriales. Las LDL transportan el colesterol desde el hígado a las células y pueden causar una acumulación nociva si hay más del que las células pueden usar.

b) Las lipoproteínas de alta densidad (HDL) o “colesterol bueno”; se cree que previenen las enfermedades arteriales. Las HDL se traen el colesterol de las células y lo devuelven al hígado donde se descompone y se elimina como residuo corporal.

c) Los triglicéridos se forman en el hígado y están presentes en productos lácteos, carne y aceites culinarios. La obesidad y la alimentación rica en grasas aumentan el riesgo de tener niveles altos de triglicéridos.

Los síntomas.

El colesterol alto está relacionado con enfermedades graves, como las cardiopatías, la angina de pecho y los accidentes cerebrovasculares. La causa de la cardiopatía coronaria es el estrechamiento de las arterias (aterosclerosis) que suministran la aportación de sangre al corazón. Los depósitos grasos, como el colesterol o los productos residuales se acumulan en el interior de las arterias. Esta acumulación se llama placa e impide el flujo de sangre por las arterias.

Si tiene síntomas de aterosclerosis, también puede tener un nivel alto de colesterol. Los síntomas incluyen la angina de pecho (dolor en el pecho causado por una reducción de la aportación sanguínea al corazón), dolor de pierna (debido al estrechamiento de las arterias que traen sangre a las extremidades) y coágulos sanguíneos en las arterias que transportan sangre al corazón (trombosis coronaria). Los coágulos sanguíneos pueden traer a una deficiencia cardíaca.

Las manchas espesas de color amarillo (xantomas) alrededor de los ojos o en alguna otra zona de la piel se forman por los depósitos de colesterol. A menudo, se pueden apreciar en personas con colesterol alto hereditario, (Hipercolesterolemia familiar). [1]

Aunque el colesterol es esencial para la vida, cuando hay un exceso en el cuerpo humano, se generan problemas de salud. En el caso del hipercolesterolemia familiar, se ha evidenciado que afecta una de cada 250 personas. Si los afectados

por este trastorno no reciben tratamientos adecuados, son frecuentes las enfermedades cardiovasculares prematuras.

En países como España, por ejemplo, hay más de 10 millones de personas con el colesterol por encima de lo deseable. [2] Alrededor del mundo se reportan casi 48 mil fallecimientos por día a causa de padecimientos cardiovasculares, lo que representa 17.5 millones de personas que mueren por año y 56 por ciento, (56%) de estos problemas se le atribuye al colesterol alto. [3]

Los factores que influyen en el incremento del colesterol corresponden a dietas inadecuadas y sedentarismo. Por lo tanto, además de atacar estos malos hábitos, es importante recurrir a tratamientos de tipo: farmacológico, naturales u homeopáticos, según la efectividad que muestre cada uno de ellos en el cuerpo que los recibe.

1.2. Justificación

En la actualidad existen múltiples tratamientos diseñados para combatir el colesterol alto y sus enfermedades derivadas. Algunos estudios están demostrando la inefectividad de los tratamientos considerando rangos de edades y comienzo de suministro de estos. Otros afirman que, se encuentran efectos secundarios potencialmente dañinos como: la inflamación del hígado o dolores musculares e insomnio. [4] Es por esto por lo que, el interés principal en el desarrollo de este trabajo de grado se centrará en diseñar un sistema de Inteligencia de Negocios o Business Intelligence, en inglés. (En adelante, BI), que posibilite el análisis de la información generada durante un experimento que tiene como objetivo comprobar la eficacia de los diferentes tratamientos para la reducción de los niveles de colesterol.

1.3. Problema

Los efectos de un tratamiento en una persona con características “A” de tipo: corporal, hereditaria y de estilo de vida, pueden ser totalmente diferentes a los que puede tener una persona con características tipo “B”. Actualmente, la selección de los tratamientos puede depender mucho de las creencias de las personas, de su afinidad o cercanía con un tema. Por lo tanto, recomendar tratamientos eficaces acordes a las necesidades de las personas, puede resultar un problema interesante de resolver.

Con el fin de tomar decisiones acertadas al momento de recomendar tratamientos para la disminución de los niveles de colesterol, se plantea el diseño de un sistema de BI, capaz de generar análisis concretos y aplicables a la realidad.

1.4. Objetivos

1.4.1. Objetivo general

El objetivo de este trabajo es el diseño e implementación de un sistema de BI, que facilite la adquisición, el almacenamiento y la explotación de datos asociados a pacientes a los que se los ha diagnosticado los niveles de colesterol (LDL).

1.4.2. Objetivos específicos

i. Diseñar un almacén de datos (Data Warehouse) que permita almacenar la información adquirida de los diferentes orígenes de datos. Teniendo en cuenta que tendremos un conjunto de pacientes que han sido sometidos, por grupos, a diferentes tratamientos.

ii. Implementar este almacén de datos y programar los procesos ETL (siglas en inglés de extracción, transformación y carga) que permitan alimentar el Data Warehouse a partir de los ficheros base facilidades.
iii. Analizar las diferentes plataformas BI Open Source disponibles al mercado que nos permitirían explotar la información almacenada.
iv. Seleccionar e implantar una de estas herramientas Open Source de tal forma que se disponga de una capa de software dentro de la arquitectura del proyecto de BI, para el análisis de la información.
v. Extraer información relevante para determinar la eficacia de los diferentes tratamientos utilizados para el Colesterol alto.

Algunas de las preguntas analíticas por resolver son:

- ¿Cuál es la relación entre los diferentes tratamientos y la evolución de los pacientes?
- ¿Existen terapias más eficaces?
- ¿Ha influido en el resultado, los hábitos de los pacientes?
- ¿La evolución a lo largo del tiempo, por un mismo tratamiento, dependen de algún factor como los hábitos?
- ¿Hay diferencias en el resultado de un tratamiento según el lugar geográfico del paciente?
- ¿Hay algún periodo del año donde el tratamiento sea más o menos efectivo?

1.5. Enfoque y método seguido

Para llevar a cabo este proyecto tecnológico, se parte inicialmente de una planificación y unos requerimientos de negocio establecidos por el cliente. Además, se tiene claridad sobre los siguientes aspectos:

- Existencia de un problema claro y una necesidad previamente validada por el cliente.
- Delimitación de un alcance para el proyecto. (Establecido por la EDT, esquema de desglose de trabajo).
- Planeación base para llevar a cabo las actividades del proyecto.
- Definición del tiempo requerido para la finalización del proyecto y sus entregables.
- Existencia de un proceso para refinar y retroalimentar los entregables parciales con el cliente.

Posteriormente, y tomando como referencia uno de los modelos más utilizados para la implementación de un Data Warehouse, y en general proyectos de BI, diseñado por Kimball, se procede a establecer el diseño de la arquitectura, la selección del entorno tecnológico, el modelado del problema, las extracciones de datos y posterior implementación.

Dicho modelo como se ilustra a continuación incluye fases de crecimiento, mantenimiento y administración del proyecto, sin embargo, esto tiene más aplicabilidad en implementaciones comerciales con continuidad en la implementación de la solución. Para desarrollar el alcance fijado en este trabajo de máster, se ha realizado una “adaptación”, la cual puede ser visualizada en la Ilustración 2.

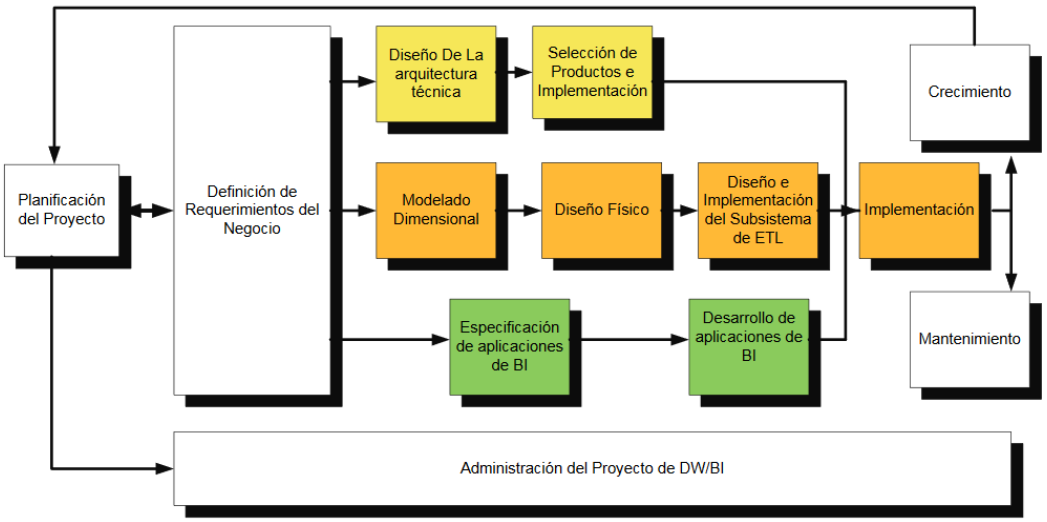


Ilustración 1, Modelo de Tareas de la metodología de Kimball, denominada Business Dimensional Lifecycle [5]

Las actividades enumeradas del uno (1) al ocho (8), permiten dar respuesta al enfoque seguido en el proyecto. Así mismo aquellas marcadas con un asterisco (*), son consideradas actividades trasversales, ya que las especificaciones y el desarrollo se darán a medida que se construya la solución.

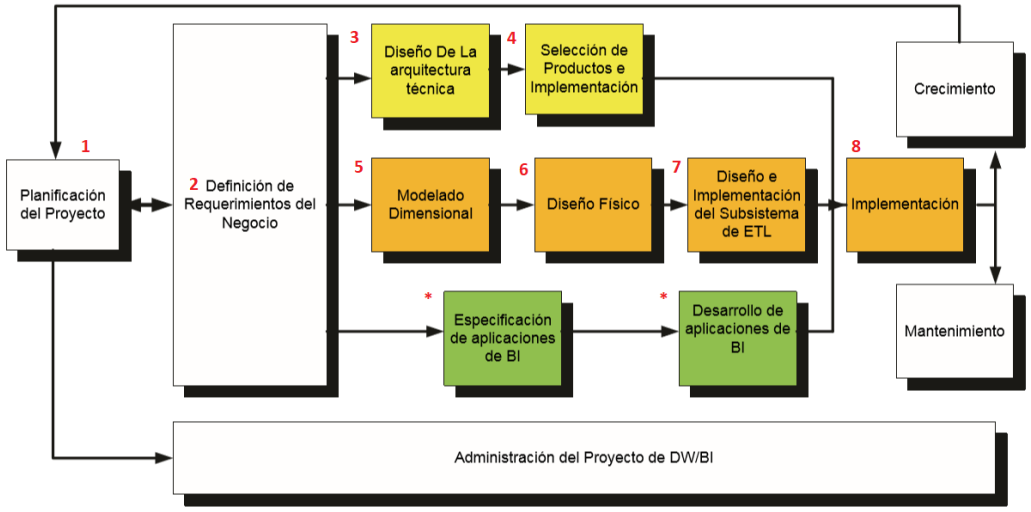


Ilustración 2, Adaptación de la Metodología de Kimball

Bajo las anteriores premisas y la metodología expuesta, en este trabajo de grado, se comenzó por presentar el estado del arte de la Inteligencia de negocios y su aplicabilidad en el ámbito de la salud. Además, de una arquitectura típica de BI, la cual da pie para la identificación de las diferentes herramientas que integran la arquitectura solución.

Se realizó entonces, una investigación exhaustiva sobre cada una de las soluciones libres de BI, previamente diseñando unos criterios base para determinar aquella(s), de mayor viabilidad para el proyecto.

El resultado de dicha investigación arrojó a Pentaho como la herramienta más viable a utilizar. Con esta se incluyeron múltiples componentes en la solución, los cuales además de permitir dar respuesta al almacenamiento de la información

debían suplir las necesidades relacionadas con la extracción, transformación y carga, el procesamiento analítico en línea y los reportes que le agregarían valor al cliente de la solución. (Hasta este aspecto se han completado los primeros cuatro (4) pasos de la metodología adaptada de Kimball).

Con la implementación del entorno tecnológico se realizó el modelamiento dimensional del Data Warehouse, lo que concluyó en una arquitectura tipo estrella para el DW. Este aspecto es declarado como uno de los más relevantes para garantizar la comprensión del problema y poder en adelante resolver cuestiones relacionadas con las preguntas analíticas planteadas.

Posteriormente, tomando como referencia el esquema del Data Warehouse (DW), se definieron las transformaciones necesarias sobre la base de datos origen de la información. Mediante el uso del conjunto de herramientas de Pentaho se logró diseñar escenarios clave para garantizar la consistencia y formato de los datos. Así mismo, escenarios enfocados a crear nuevos elementos sobre el DW que permitiesen dar respuesta al diseño lógico del mismo.

En adelante, se diseñó un cubo OLAP con la capacidad de visualizar y presentar en línea los datos que desde el negocio se consideran relevantes para correlacionar y agrupar información. Dicho cubo OLAP fue publicado en el servidor Mondrian de Pentaho y accedido mediante las herramientas de Jpivot y Saiku las cuales permitieron explorar su navegación y el reporte de información para el usuario final.

Finalmente, se logró dar respuesta a las preguntas analíticas presentadas en el proyecto, actividad con la que se dio cierre al abordaje de este.

1.6. Planificación del Trabajo

A continuación, se presenta de manera resumida los principales entregables y/o logros que se plantean a lo largo del proyecto. Lo anterior utilizando la técnica de EDT.

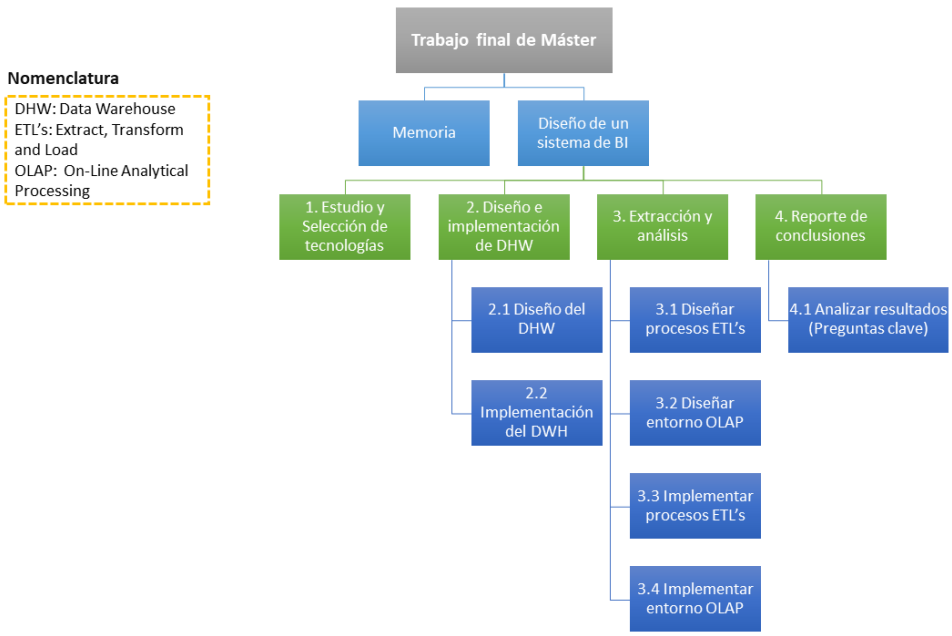


Ilustración 3, Estructura detallada del trabajo.
Realizada con la herramienta: Power Point 2016.

1.7. Breve sumario de productos obtenidos

En este proyecto de máster se generaron los siguientes entregables:

I. **Estudio y selección de tecnologías de BI libres:** consistió en una investigación relacionada con las tecnologías de BI libres en el mercado y sus principales funcionalidades. Lo anterior permitió seleccionar la plataforma de Pentaho que sería la base de la arquitectura técnica del sistema de BI.

II. **Diseño e implementación del data warehouse:** se realizó un análisis dimensional que generó un esquema tipo estrella para el data warehouse a ser implementado en el proyecto. La metodología utilizada para la realización del diseño fue la de Kimball. [5]

III. **Diseño e implementación de los procesos ETL:** contiene cada uno de los escenarios planteados desde la extracción y transformación de los datos de origen a ser tratados en el sistema de BI, hasta aquellos donde se cargó la información al DW. Estos escenarios fueron implementados a través de la herramienta PDI.

IV. **Diseño e implementación del entorno OLAP y cubos OLAP:** a partir del análisis dimensional y de la investigación realizada para entender el concepto de procesamiento analítico en línea, se construyó un entorno OLAP con la herramienta PSW, y se diseñó un cubo OLAP con las dimensiones y medidas necesarias del proyecto. EstE fue publicado en el servidor Mondrian de Pentaho donde finalmente se extrajeron los análisis relevantes del proyecto.

V. **Implementación de la arquitectura de un sistema de BI en funcionamiento:** este producto es el entregable final del proyecto de máster, consiste en el diseño técnico y operativo del sistema de BI. Con él se llevaron a cabo cada uno de los procesos de configuración, extracción, transformación, carga, reporte y otros del proyecto.

VI. **Conclusiones relacionadas con las preguntas analíticas del proyecto:** la construcción del sistema de BI planteaba el reto de analizar la eficacia de los tratamientos suministrados para la disminución de los niveles de colesterol, es por esto, que un entregable importante al final del proyecto era el de establecer las conclusiones relacionadas con la correlación y comprensión de la información. De no haber realizado un análisis claro del sistema de BI, el DW, los procesos ETL y demás, no se tendría completo este apartado.

1.8. Breve descripción de los otros capítulos de la memoria

Capítulo 2, Estado del arte: en este capítulo se brinda contexto sobre el concepto de BI, su evolución a lo largo del tiempo, los beneficios que representa en utilizar esta capacidad o herramienta en los entornos organizacionales y las arquitecturas tipo de BI.

Capítulo 3, Desarrollo del prototipo: el capítulo describe cómo se desarrolló el prototipo de arquitectura de BI del proyecto, diferenciando las fuentes de información y estableciendo criterios para la selección del entorno tecnológico.

Capítulo 4, Implementación del entorno tecnológico: se definen los requerimientos técnicos de la solución de BI a partir de la herramienta

seleccionada, (Pentaho). Así mismo, se implementan las herramientas para los procesos de:

I. **ETL:** Pentaho Data Integration.

II. **Procesamiento analítico en línea:** Mondrian y PSW.

III. **Reporte:** Saiku y Jpivot.

Capítulo 5, Data warehouse diseño lógico: define el tipo de DW a construir y el esquema que se utilizaría para técnicamente reflejar el diseño en las herramientas tecnológicas de la solución.

Capítulo 6, Diseño e implementación de ETL's: consiste en el diseño de los escenarios para las extracciones, transformación y carga al DW.

Capítulo 7, OLAP, On-Line Analytical Procesing: define el concepto de procesamiento analítico en línea y cubos OLAP. Adicionalmente, expone los componentes, las ventajas y operación de estos últimos. Comprendido el concepto se genera el diseño del Cubo OLAP del proyecto.

Capítulo 8, Reporte de información: consiste en la puesta en marcha de la herramienta de reporte y visualización de cubos OLAP, SAIKU. Bajo esta implementación se extraen diferentes análisis que son la fuente de información para la generación de conclusiones relacionadas con la efectividad de los tratamientos para el colesterol LDL.

Capítulo 9, Conclusiones: establece las conclusiones del trabajo final de máster a partir de la implementación del sistema BI.

2. Estado del arte

2.1. Introducción

Las organizaciones actuales cada vez más se ven retadas a tomar decisiones de manera rápida y frecuente, debido al creciente entorno competitivo y la oferta que existe en múltiples servicios.

Uno de los padecimientos que resulta ser más común para dichas organizaciones, es tener muchos datos, pero poca información; estos datos en la mayoría de los casos se pueden encontrar en múltiples fuentes o sistemas de información distribuidos, complejos arquitectónicamente y difíciles de integrar. Lo anterior, se presta para que en las organizaciones se implementen procedimientos manuales para la extracción y correlación de información. De la misma manera, se adquieren herramientas para la estructuración de reportes que en muchos casos son lentos, costosos, duplican esfuerzos y son propensos a errores o sujetos a la interpretación individual. [6]

Con el fin de habilitar el manejo y análisis eficiente y efectivo de los diferentes volúmenes de información de un negocio, adicionalmente, de entender mejor las necesidades de los clientes, encontrar formas de medir eficientemente los recursos, monitorear y tener visibilidad de la situación, se ha creado el concepto de Inteligencia de negocios.

Una de las definiciones más reconocidas o aceptadas sobre este concepto, es la emitida por el Data Warehouse Institute, donde se menciona que la Inteligencia de negocios es la combinación de tecnología, herramientas y procesos que

permiten convertir datos en información e información en reglas y planes que optimicen la toma de decisiones y las actividades de negocio. Las organizaciones actuales no se preocupan por los costos de almacenamiento de la información -lo cual ha disminuido notablemente-, sino por el obtener la mayor ventaja competitiva posible de la información que gestionan. [7]

2.2. Evolución de la Inteligencia de negocios.

De acuerdo con History of Business Intelligence, el término “Business Intelligence” apareció por primera vez en Cyclopedia of Commercial and Business de Richard Millar Devens en el año de 1865.

En un artículo publicado en 1958, el investigador de IBM Hans Peter Luhn definió y utilizó el concepto de inteligencia de negocios. Después de eso, la inteligencia empresarial se comprende y evoluciona a partir de los sistemas de soporte a las decisiones (Decision Support Systems en inglés o DSS) que comenzaron en la década de 1960 y se desarrollaron a mediados de la década de 1980, incluso en países como Holanda, Bélgica, Francia y Alemania.

En 1989, Howard Dresner, quien al año 2017, era analista de Gartner, brindó una definición sobre aquel ya mencionado concepto, y en la década de 1990 el uso de BI se extendió a todo el mundo.

Por su parte, en el siglo XX, a comienzos del año 2000, se da el punto de inflexión más grande en varias industrias, donde más y más compañías comenzaron a comprender el verdadero valor que la inteligencia de negocios podría agregarles a sus organizaciones. [8]

De manera resumida, como se ilustra a continuación, (Ver Ilustración 2), pasamos de tener reportes manuales a tener almacenes de datos completos y a extraer métricas e indicadores de valor para generar cada vez más casos de éxito.

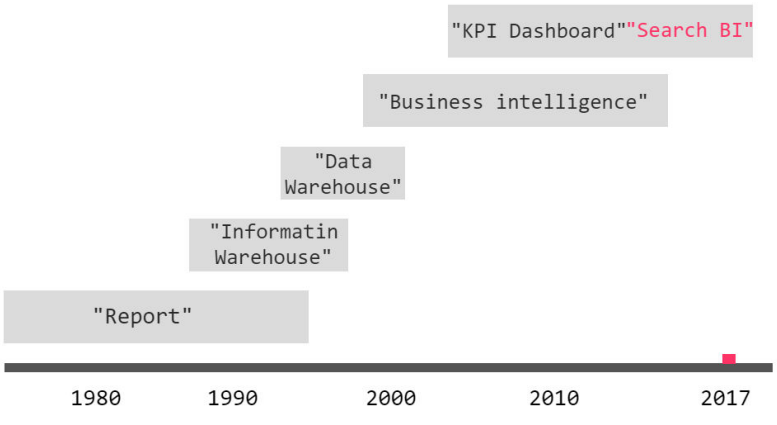


Ilustración 4, Evolución de la Inteligencia de negocios. [8]

2.3. Beneficios de la Inteligencia de negocios.

Los resultados de una encuesta realizada a 2,600 usuarios y publicada en 2017 en el portal de BI-Survey, evidencia los siguientes como los principales beneficios frente a la aplicación de la inteligencia de negocios en el entorno empresarial: [9]

- Lograr informes y análisis de manera más rápida y precisa.
- Toma de mejores decisiones en el negocio.
- Mejora en la calidad de los datos.
- Mejora en la satisfacción de los empleados.
- Mejora en la eficiencia operacional.
- Cliente más satisfecho.
- Mayor ventaja competitiva.
- Reducción de costos.
- Incremento de ingresos.

A partir de esta misma encuesta se logra evidenciar en la siguiente ilustración, cómo el software de BI realmente ayuda a las empresas.

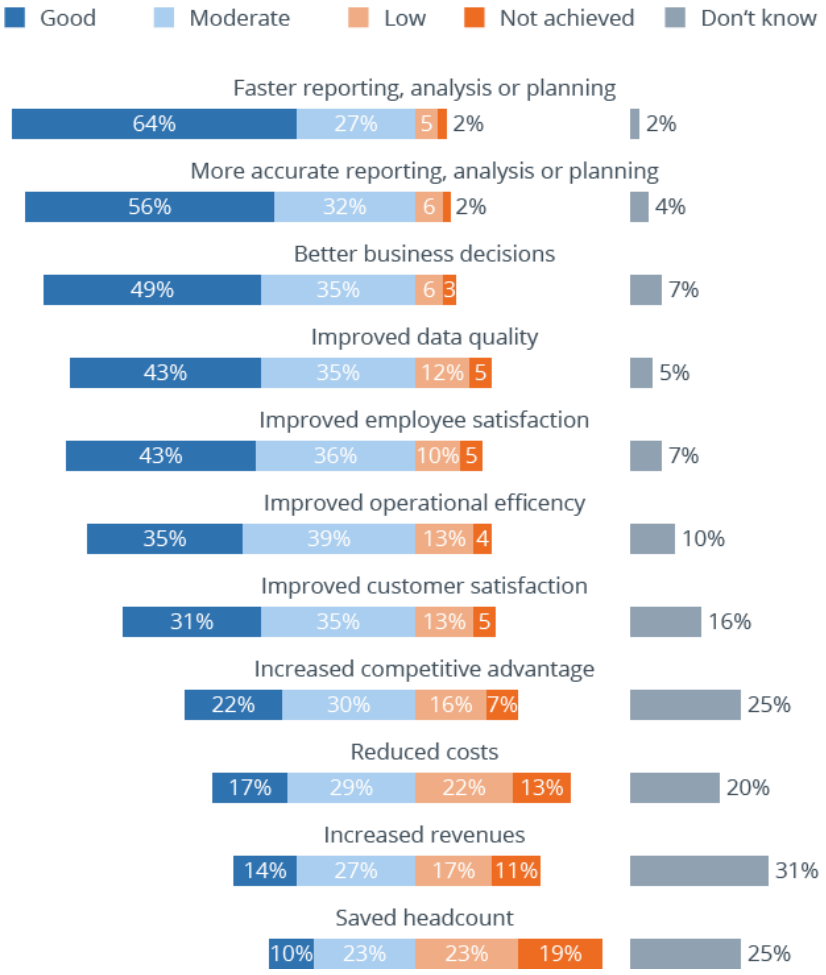


Ilustración 5, Beneficios de BI para las empresas. [9]

En el sector salud, por ejemplo, BI, permite habilitar las siguientes capacidades, algunas de estas no muy lejanas de los beneficios que en otros sectores se pueden evidenciar.

- Optimizar recursos (incluidos el espacio físico, equipos y dispositivos, personal y suministros).
- Desarrollar y monitorear indicadores clave de rendimiento e indicadores clínicos para mejorar el rendimiento y la calidad.
- Llevar a cabo la planificación, la elaboración de presupuestos y los pronósticos de manera más eficiente y precisa en grandes organizaciones.

- Comprender y gestionar eficazmente la cadena de suministro y la logística para contener los costos y garantizar un suministro constante.
- Garantizar una mejor seguridad del paciente a través de diagnósticos eficientes y la identificación y cumplimiento de los protocolos de tratamiento de mejores prácticas.
- Contener costos y mejorar el rendimiento y la calidad a través de la gestión de recursos humanos y la elaboración de perfiles médicos adecuados. [10]

Asimismo, a través de valorar a 2,220 encuestados se logra determinar algunos sectores o necesidades donde se está aplicando mayormente BI. (Ver Ilustración 5).

Es claro que en temas de mercadeo (Comportamiento integral de cliente), y ventas, el uso de BI es casi obligatorio en este contexto. Sin embargo, en actividades como la optimización de precios, los análisis geoespaciales, la simulación, la optimización de las cadenas de suministros y otros, se está aumentando considerablemente el uso de BI.

Lo que parece extraño, es que aún esta capacidad no se esté utilizando en gran medida para habilitar análisis de fraude y seguridad, ya que son aspectos fuertemente regulados en algunos países y adicionalmente, el no tener visibilidad sobre estos, genera pérdidas monetarias y reputacionales para las organizaciones. Esta misma conclusión podemos extraerla para aplicaciones relacionadas con el sector de la salud, ya que, aunque no es estrictamente regulado, su uso si brinda unos beneficios claves y una ventaja competitiva en el sector.

Ilustración 6, Sectores de inversión en BI. [9]

2.4. Aplicación de BI, en el sector salud.

En el año 2008, Xuezhong, en conjunto con otros estudiosos, sugirieron que los datos clínicos del proceso clínico diario, que se ajustan a las teorías y principios de la medicina tradicional china (MTC), son la fuente de conocimiento empírico fundamental para las investigaciones de la MTC. En este caso, introdujeron un sistema de almacenamiento de datos, que se basa en el sistema de registro médico electrónico estructurado y los datos clínicos diarios,

para las investigaciones clínicas de MTC y el descubrimiento de conocimiento médico. El sistema consta de varios componentes clave: esquema de datos clínicos, herramienta ETL, análisis analítico en línea (OLAP) basado en Business Objects (un software de inteligencia comercial) y funcionalidades integradas de extracción de datos. Su almacén de datos hasta el año 2012, almacenaba 20,000 datos de pacientes hospitalizados de diabetes, enfermedad coronaria y accidente cerebrovascular, y más de 20,000 datos de pacientes ambulatorios. En conclusión, sus aplicaciones de análisis mostraron que la plataforma de almacenamiento de datos clínicos desarrollada promete construir el puente para la práctica clínica y la investigación teórica de la medicina tradicional china que promoverá las investigaciones de la medicina tradicional china relacionadas. [10]

En el 2010, por su parte, Lihong, J., Hongming, C., & Boyi, en un intento por eliminar la heterogeneidad de los datos para construir un almacén de datos, introdujeron la ontología de dominio en el proceso ETL para encontrar las fuentes de datos y definir las reglas de transformación de los datos y a su vez eliminar la heterogeneidad. En esta tarea, incorporaron la ontología de dominio en los metadatos del almacén de datos, lo que llevó a que los registros de datos se asignaran de las bases de datos a las clases del Lenguaje de Ontología Web (OWL). Esto dio lugar a acceder a los recursos de información de manera más eficiente. Los autores probaron el método en un proyecto de almacenamiento de datos del hospital, y el resultado muestra que el método de ontología juega un papel importante en el proceso de integración de datos al proporcionar descripciones comunes de los conceptos y relaciones de los elementos de datos, y la ontología del dominio médico en el proceso ETL, es de viabilidad práctica. [10] [11]

En un artículo publicado en el mismo año, en la revista Journal of Health Informatics**,** se describe una implementación de un ambiente computacional utilizando tecnologías web para analizar datos de salud, a través de herramientas OLAP. Los resultados obtenidos a través del cruce de información permitieron la identificación de individuos enfermos o con predisposición para desarrollar una enfermedad arterial coronaria, y así aplicar programas preventivos. [12]

En la última década el desarrollo de soluciones de BI se ha incrementado y en el campo de la salud algunas de ellas ya se encuentran preconstruidas y son personalizables de acuerdo con las necesidades de cada cliente. Por ejemplo, una plataforma desarrollada por SAP, permite entre muchas cosas, que los médicos comprendan información clínica actualizada en los procesos de investigación, el acceso a los datos relevantes en tiempo real, la segmentación de pacientes, la creación de planes de salud alineados con las necesidades de cada persona y el monitoreo en línea de su avance, el seguimiento de la cadena logística de productos médicos y farmacéuticos desde la materia prima hasta el punto de consumo por parte del paciente, incluida la prevención eficaz de falsificaciones. [13]

En el 2018, por ejemplo, Orion Health anunció en Boston, (Massachusetts), la implementación de un Core en salud llamado Amadeus, compuesto de diferentes soluciones de gestión de datos, almacenamiento y uso compartido, para permitir a las empresas de atención médica dar un primer paso hacia la inteligencia de negocios. Lo que se busca con esta solución es la atención basada en el valor y la gestión de la salud de la población de manera predictiva. [14]

Por su parte, Martin Koehring, editor jefe de Salud en The Economist Intelligence Unit, (Para el año 2018), y quién tiene acceso a información proveniente de diferentes partes del mundo sobre los principales descubrimientos y hallazgos del sector salud, afirma que el gran aumento de los datos biométricos a través de aplicaciones móviles, pruebas genéticas, exámenes avanzados, etc. Cobrará cada vez más relevancia para las organizaciones y los estados gubernamentales, ya que se deberá analizar mejor los datos y aplicar mejores modelos predictivos. [15]

En conclusión, la implantación de un sistema de BI en el sector de la salud posibilita la búsqueda e interpretación de información almacenada para apoyar la toma de decisiones no sólo referentes a un negocio sino a la misma vida de las personas. [12]

2.5. Arquitectura de un sistema de BI.

Los datos sobre los cuales se realizan tareas de análisis y correlación de información a menudo provienen de diferentes fuentes, generalmente aplicaciones, bases de datos descentralizadas, información que envían proveedores de la organización y otras. En una arquitectura típica de BI, este es el primer componente. A continuación, encontramos además de este, los otros elementos importantes en dicha arquitectura:

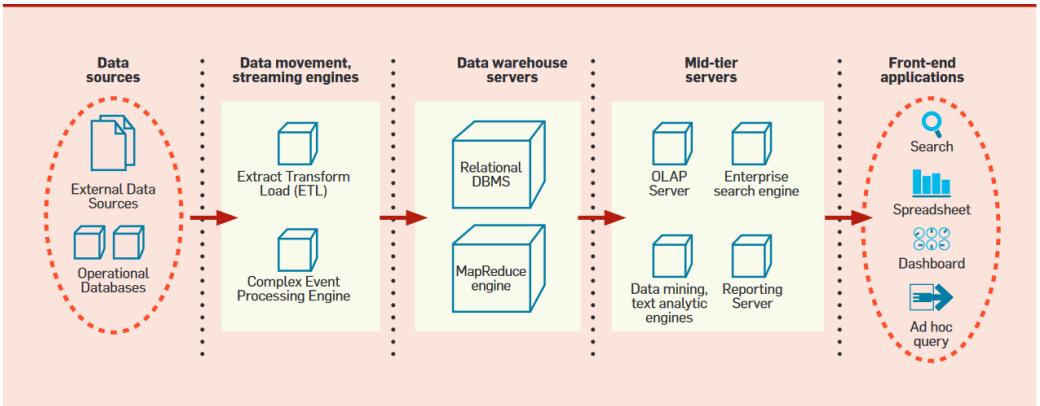


Ilustración 7, Arquitectura típica de un proyecto de BI. [7]

i. **Fuentes de información o Data sources:** fuentes de información, internas o externas a la organización.

ii. **Movimiento y carga de datos o Data Movement, Streaming engines:** una de las situaciones más desafiantes en BI, la cual se deriva de la heterogeneidad de las fuentes de información, es la limpieza, y estandarización de los datos para su respectivo uso. Por lo tanto, se debe definir a partir de toda la información obtenida de dichas fuentes, qué campos realmente se requieren, si necesitan algún tipo de modificación y/o transformación y donde se ubicarían estos, este proceso se le conoce como “mapping”. [16]

Las tecnologías de back-end se denominan colectivamente como herramientas de extracción-transformación-carga (ETL). Cada vez es más necesario respaldar las tareas de BI casi en tiempo real, es decir, tomar decisiones comerciales basadas en los datos operativos en sí. Los motores especializados denominados motores de Procesamiento de Eventos Complejos (CEP) han surgido para soportar tales escenarios.

iii. **Almacén de datos, Data warehouse servers:** los datos sobre los cuales se realizan las tareas de BI se cargan generalmente en un repositorio llamado el almacén de datos o Data Warehouse en inglés, el cual es administrado por uno o más servidores de almacenamiento de datos. Una opción popular de motores para almacenar y consultar datos de almacén son los sistemas de administración de bases de datos relacionales (RDBMS).

iv. **Servidores de nivel medio o intermedio, Mid-tiers servers:** los servidores de almacenamiento de datos se complementan con un conjunto de servidores de nivel medio que proporcionan una funcionalidad especializada para diferentes escenarios de BI. En esta capa de la arquitectura se encuentran los servidores de procesamiento analítico en línea (OLAP), los cuales exponen de manera eficiente la vista multidimensional de los datos a las aplicaciones o los usuarios y permiten las operaciones de BI más comunes, como el filtrado, la agregación, el desglose y la rotación. Además de los servidores OLAP tradicionales, están apareciendo nuevos motores de "BI en memoria" que explotan los grandes tamaños de memoria principal de hoy en día para mejorar drásticamente el rendimiento de las consultas multidimensionales. [7]

   Asimismo, los servidores de reportes, (*Reporting servers,* en inglés), permiten la definición, la ejecución eficiente y la representación de informes.

   Por su parte, los motores de búsqueda empresariales, (Enterprise Search Engine, en inglés), admiten la búsqueda de palabras clave sobre texto y datos estructurados en el Data Warehouse, y se han convertido en una herramienta valiosa para BI.

   Finalmente, en esta capa, encontramos un componente bastante importante y es el motor de minería de datos, (*Data Mining Text Analytics Engine, en inglés), permiten* un análisis en profundidad de los datos, el cual va mucho más allá de lo que ofrecen los servidores OLAP o de reportes, y es el proporcionar la capacidad de construir modelos predictivos para responder a preguntas como: ¿qué clientes actuales probablemente responderán a mi próximo envío de correos por catálogo?

v. Por último, encontramos la vista para el usuario final, (Front end Applications, en inglés). En esta capa podemos encontrar hojas de cálculo de Excel, portales empresariales para búsquedas, aplicaciones o herramientas de indicadores o métricas para hacer seguimiento de los aspectos claves de la empresa mediante paneles visuales, herramientas que permiten a los usuarios plantear consultas ad hoc, visores de modelos de minería de datos, etc. La visualización rápida y ad hoc de los datos puede permitir la exploración dinámica de patrones, valores atípicos y ayudar a descubrir hechos relevantes para BI. Es allí, donde evidencian el verdadero valor de la implementación de la solución de BI. [7]

La arquitectura anteriormente expuesta puede ser una de las diversas implementaciones que cada empresa puede desarrollar para dar respuesta a sus necesidades de negocio. En algunos casos, se pueden integrar herramientas de analítica de sitios web, comportamiento de los usuarios en nuestros sistemas empresariales, analítica sobre dispositivos móviles y otros. Lo anterior con el objetivo de integrar toda aquella información que de acuerdo con nuestros casos de negocio puede ser más relevante. Sin embargo, es claro que la arquitectura

debería ser lo más “limpia” posible para evitar dependencias que generen lentitudes y afectaciones en el servicio. La experiencia nos ha mostrado que estos sistemas no son realmente críticos en las empresas, sin embargo, el no tenerlos en un momento determinado puede significar no ver oportunidades o desaciertos a tiempo.

3. Desarrollo del prototipo.

Retomando los objetivos específicos del proyecto, en este apartado se dará respuesta al diseño e implementación de:

- Un almacén de datos o Data Warehouse.
- Procesos ETL.
- Reportes que permitan dar respuesta a las preguntas analíticas.

3.1. Arquitectura técnica

La arquitectura técnica describe cada uno de los aspectos de implementación requeridos en la solución tecnológica.

3.1.1. Fuentes de datos origen

Para nuestro proyecto, tenemos una base de datos compuesta por la siguiente información:

- **Treatment:** listado de los 3 tipos de tratamiento del estudio.
- **Patients:** datos de los pacientes (anonimizados), lugar de origen, sexo y rango de edad.
- **Habits:** dieta y actividad física a nivel semanal.
- **Indicators**: valor de los indicadores a nivel semanal.

En la siguiente ilustración, se presenta el diagrama que expone las entidades y sus relaciones en la base de datos. Este diagrama está fuertemente relacionado con el gráfico de burbujas ilustrado en anteriores apartados, el cual expone las dimensiones que son relevantes a considerar en el proyecto.

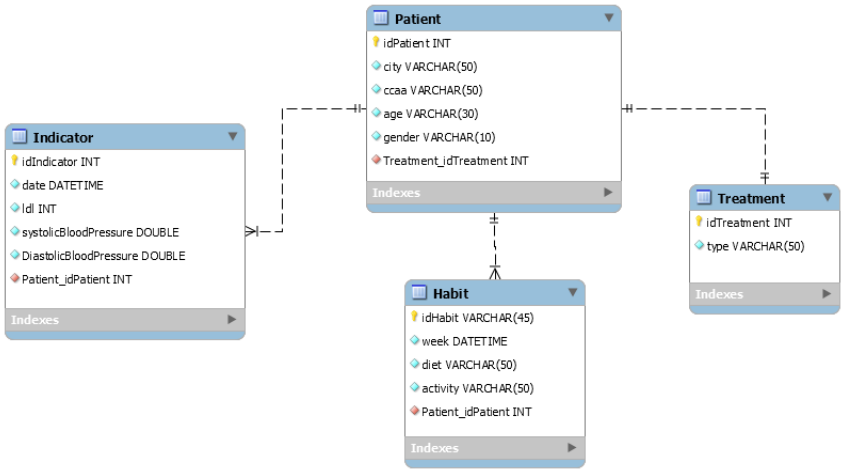


Ilustración 8, Diagrama entidad-relación.
Realizado con el software MySQL Workbench 8.0.

Para el uso de la información se asume que:

- Los identificadores de pacientes y tratamientos serán numéricos.

- Un paciente solo tiene un (1) tratamiento.
- Un paciente debe tener uno (1) o muchos indicadores.
- Un paciente debe tener uno (1) o muchos hábitos.

3.1.2. Selección del entorno tecnológico

Comúnmente se encuentra software para BI, con múltiples capacidades, pero con un factor bastante considerable en el momento de la implementación, y es, el costo. Por lo tanto, en este proyecto se hará uso de herramientas open source o de uso no comercial, previamente valoradas para el cumplimiento de los requisitos establecidos por el cliente.

Algunas de las herramientas BI open source más conocidas son: [5] [17] [18]

i. Jaspersoft BI Tools.
ii. Pentaho Community Edition.
iii. SpagoBI.
iv. KNIME Business Intelligence Tools.
v. Microsoft Power BI.
vi. Birt.
vii. QlikView Personal Edition.
viii. Tableau Public.
ix. ReportServer Community Edition.

Para seleccionar la herramienta se han definido los siguientes requisitos basados en Benchmarkings del mercado [19] [20]:

- **Generación de reportes (KPIS, Indicadores):** capacidad de generar reportes e informes para responder las preguntas analíticas del proyecto.
- **Procesos ETL:** capacidad de diseñar e implementar procesos ETL.
- **Nivel de esfuerzo para uso de la interfaz:** se refiere al nivel de esfuerzo (Bajo, medio o alto), requerido para aprender a manipular la herramienta. Lo anterior, considerando que mientras más intuitiva sea la misma, se puede dedicar mayor tiempo a estructurar lo que tome más esfuerzo.
- **OLAP Analysis:** se refiere a la capacidad de ofrecer un entorno para agilizar y establecer consultas especiales sobre la información. No aplica, si se permite conectar a cubos OLAP en otro entorno. Lo ideal es un ecosistema de BI, integrado con esta funcionalidad.
- **Data Mining:** se refiere a la capacidad de explorar datos de manera automática o semiautomática.
- **Reportes de Georreferenciación:** este criterio es deseable, y se quiere para a publicación de datos a través de una representación geográfica.
- **Flexibilidad para manejar diferentes orígenes de datos (CSV, XML, Excel, entre otros):** se refiere a la capacidad de integración con múltiples tipos de fuentes de información o base de datos.
- **Documentación:** se refiere a la cantidad y calidad en la documentación requerida para conocer y operar la herramienta.

En la evaluación de cada uno de estos requisitos para las plataformas listadas anteriormente, las herramientas de SpagoBI y Pentaho, obtuvieron el mismo puntaje. (12.1**Anexo 1, Selección entorno tecnológico**)

Por lo tanto, se tomó como criterio de selección de dicha tecnología, un benchmarking realizado sobre herramientas open source de BI, especializado en

aplicabilidades en el sector de la salud. [5] Para lo cual se tiene la siguiente información:

Features	BI Open-Source Tools						Group
	Jaspersoft BI	*Palo BI Suite*	*Pentaho BI Suite*	*QlikView*	*SpagoBI*	*Tableau Public*	
Performance	4	3	4	3	4	4	D
OLAP Ad hoc Queries	1	5	5	3	5	4	B
Architecture	4	4	5	4	5	4	D
Display of KPIs	1	1	5	4	4	4	A
Plug-ins	3	0	5	0	0	3	D
Interactive Visualization of Data	5	4	5	5	4	4	C
Documentation	4	4	2	2	2	3	F
Dashboards	1	1	4	4	5	4	B
Navigation Features	5	4	4	1	2	4	C
ETL	4	5	5	3	4	1	E
Connection to the Database	5	4	4	5	5	3	A
Integration of Dimensional Model	1	1	1	2	4	1	E
Open-source	5	5	5	1	5	5	D
Export	5	2	5	2	5	4	C
Pervasive	5	5	5	1	5	4	A
Online Help	4	2	3	4	3	4	F
Support for Mobile Devices	4	1	0	5	5	3	C
Data Mining	1	1	3	2	4	1	B
Ease of Use	4	4	4	4	4	5	F
Attractiveness	4	3	4	5	5	4	C
Customization of the Interface	4	0	5	5	5	5	F
User Profile	5	4	5	1	4	0	D
Real-time	5	4	5	1	5	1	A

Ilustración 9, Tabla comparativa de las herramientas de código abierto de BI seleccionadas por características y grupo.

***Nota:** 0 significa desconocido; 1 significa ausente; 2 significa insuficiente; 3 significa suficiente; 4 significa bueno y 5 significa excelente.*

Tal como se observa en la anterior ilustración, ambas herramientas presentan muy pocas disimilitudes. Sin embargo, revisando las características esenciales y deseables de la misma, entre ellas: OLAP, ETL, reportes, KPI's, data mining y personalización, la herramienta seleccionada es **Pentaho**.

4. Implementación del entorno tecnológico

En este apartado se detalla la implementación de Pentaho y su ecosistema de soluciones como herramienta para BI, en su versión 8.1.

4.1. Pentaho 8.1

4.1.1. Especificaciones técnicas

Los requerimientos técnicos por cubrir para el uso del software son: [21]

Tabla 1, Requerimientos para la instalación de Pentaho 8

Hardware	• Procesador: Intel EM64T o AMD64 Dual Core • RAM: 8 GB con 4GB dedicadas al servidor de Pentaho • Espacio en disco duro: 20 GB. Después de la instalación 2 GB.
Servidor de aplicaciones	• JBoss EAP 7.x with Oracle Java 8.x • Tomcat 8.0* & 8.5 (Default) con Oracle Java 8x.
Repositorios de base de datos	• MySQL 5.6 & 5.7* (SQL 92) • Oracle 11.2 & 12.1 (SQL 92) • PostgreSQL 9.5 & 9.6 (Por defecto). • MS SQL Server 2014, 2016.

**Requerimientos instalados para funcionamiento de Pentaho.*

4.1.2. Motor de base de datos MySQL

Para este proyecto se utilizó la versión 5.7.24 de MySQL. [22]

El modelo relacional presentado en el apartado 5.1 se creó sobre este motor de base de datos con una base de datos llamada tfm2018:

```
mysql> show databases;
+--------------------+
| Database           |
+--------------------+
| information_schema |
| mysql              |
| performance_schema |
| sys                |
| tfm2018            |
+--------------------+
5 rows in set (0.00 sec)
```

Ilustración 11, Base de datos MySQL

```
mysql> show tables;
+-------------------+
| Tables_in_tfm2018 |
+-------------------+
| habit             |
| indicator         |
| patient           |
| treatment         |
+-------------------+
4 rows in set (0.05 sec)
```

Ilustración 10, Tablas del modelo relacional

4.1.3. Servidor de aplicaciones Tomcat 8

Tal como se especificó anteriormente se hizo uso de la versión 8 de Tomcat. [23]

Ilustración 12, Apache Tomcat 8 [24]

Para su correcto funcionamiento en Windows 10, se configuraron las respectivas variables de entorno:

Variable	Valor
CATALINA_HOME	E:\Master UOC\Semestre 4\pentaho-server-ce-8.1.0.0-365\pentaho-server\tomcat
JAVA_HOME	C:\Program Files\Java\jdk1.8.0_191
JAVA_JRE	C:\Program Files\Java\jre1.8.0_191

Ilustración 13, Variables de entorno para el uso de Tomcat

Como se detalla a continuación, el servidor de aplicaciones se instala al momento de la instalación de Pentaho.

La versión utilizada del software Pentaho es la 8.1 y se encuentra disponible para descarga en [25]. Se procede entonces a descargar el archivo .zip llamado *pentaho-server-ce-8.1.0.365.zip*, el cual tiene el integrado el servidor de aplicaciones.

Inicialmente, se realiza la instalación de Tomcat 8, descomprimiendo el archivo y seleccionando el directorio \bin del servidor de aplicaciones. Allí a través de consola (CMD), se utiliza el comando .\service.bat install.

En adelante, se adiciona la variable de entorno para el correcto funcionamiento del servidor de aplicaciones, tal como se ilustró anteriormente, llamada CATALINA_HOME.

Y finalmente, se procede a ejecutar el servidor de aplicaciones con el comando .\startup.bat.

```
E:\Master UOC\Semestre 4\pentaho-server-ce-8.1.0.0-365\pentaho-server\tomcat\bin>.\service.bat install
Installing the service 'Tomcat8' ...
Using CATALINA_HOME:    "E:\Master UOC\Semestre 4\pentaho-server-ce-8.1.0.0-365\pentaho-server\tomcat"
Using CATALINA_BASE:    "E:\Master UOC\Semestre 4\pentaho-server-ce-8.1.0.0-365\pentaho-server\tomcat"
Using JAVA_HOME:        "C:\Program Files\Java\jdk1.8.0_191"
Using JRE_HOME:         "C:\Program Files\Java\jdk1.8.0_191\jre"
Using JVM:              "C:\Program Files\Java\jdk1.8.0_191\jre\bin\server\jvm.dll"
The service 'Tomcat8' has been installed.

E:\Master UOC\Semestre 4\pentaho-server-ce-8.1.0.0-365\pentaho-server\tomcat\bin>.\startup.bat
Using CATALINA_BASE:   "E:\Master UOC\Semestre 4\pentaho-server-ce-8.1.0.0-365\pentaho-server\tomcat"
Using CATALINA_HOME:   "E:\Master UOC\Semestre 4\pentaho-server-ce-8.1.0.0-365\pentaho-server\tomcat"
Using CATALINA_TMPDIR: "E:\Master UOC\Semestre 4\pentaho-server-ce-8.1.0.0-365\pentaho-server\tomcat\temp"
Using JRE_HOME:        "C:\Program Files\Java\jdk1.8.0_191"
Using CLASSPATH:       "E:\Master UOC\Semestre 4\pentaho-server-ce-8.1.0.0-365\pentaho-server\tomcat\bin\bootstrap.jar;E:\Master UOC\Semestre 4\pentaho-server-ce-8.1.0.0-36
5\pentaho-server\tomcat\bin\tomcat-juli.jar"
```

Ilustración 14, Instalación y ejecución de Tomcat 8.

En el momento de la instalación, los paquetes integrados en el archivo .zip, se despliegan para utilizar correctamente Pentaho.

Posterior a verificar la correcta instalación se puede visualizar en el localhost, puerto 8080, el entorno de Pentaho.

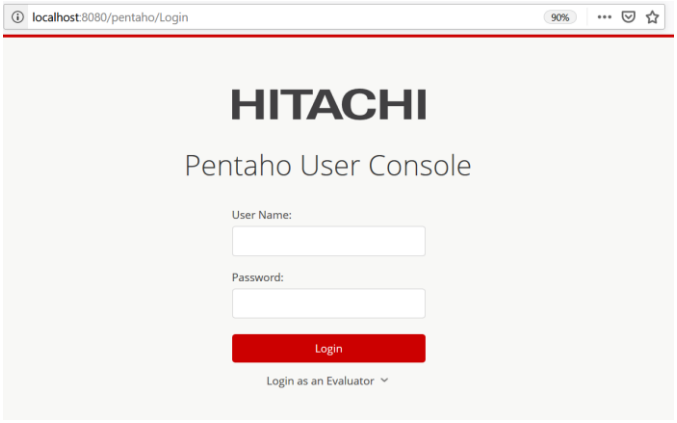


Ilustración 15, Pantalla de inicio de Pentaho

Aunque existen varios métodos de inicio de sesión en Pentaho, por ejemplo, como usuario evaluador, de negocio o administrador. Nuestro caso es este último.

Al ingresar los datos de inicio de sesión por defecto, se pueden visualizar las diferentes capacidades que habilita el software, por ejemplo, importar archivos para generar reportes, crear nuevos reportes, gestionar fuentes de datos, documentación y otros.

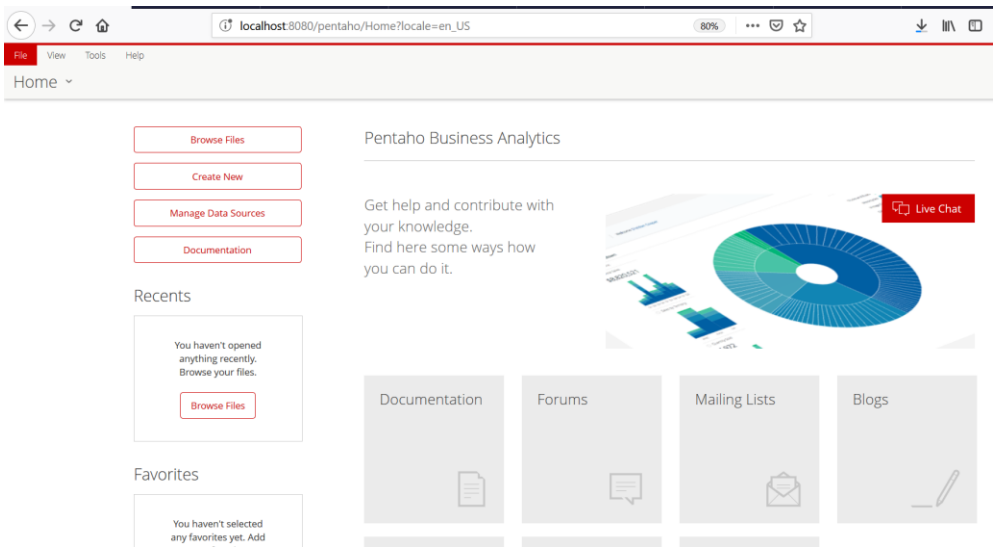


Ilustración 16, Funcionalidades de Pentaho 8.1

4.2. Data Warehouse

Para implementar adecuadamente el componente técnico del Data Warehouse que utilizaría Pentaho, se establece una conexión a la base de datos a través de la opción: File → New → Data Source:

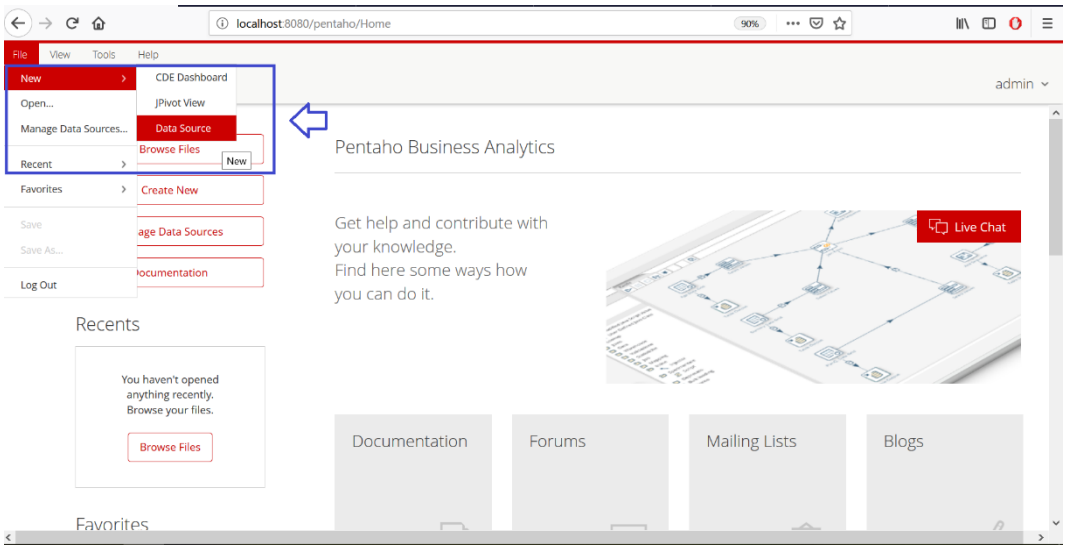


Ilustración 17, Nuevo fuente de datos para Pentaho

Para gestionar la nueva conexión se requiere indicar un nombre y un tipo de conexión, entre los cuales se tiene: archivos CSV, Query SQL, Tabla de base de datos.

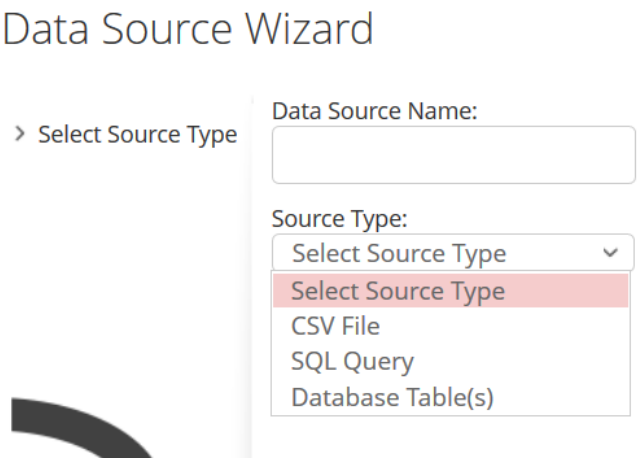


Ilustración 18, Ajustes a realizar para fuente de datos en Pentaho

Teniendo en cuenta que la base de datos ya está lista para usarse, se elige la última opción. Posteriormente se ingresan los datos solicitados y antes de confirmar la creación se puede realizar una prueba para verificar que la conexión sea exitosa.

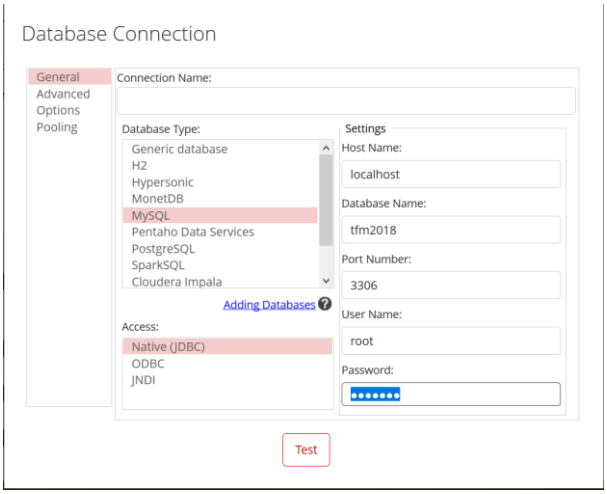


Ilustración 19, Creación de la conexión a la base de datos desde Pentaho

Ilustración 20, Test de la conexión a la base de datos.

Hasta este paso, el ecosistema de solución es el ilustrado a continuación:

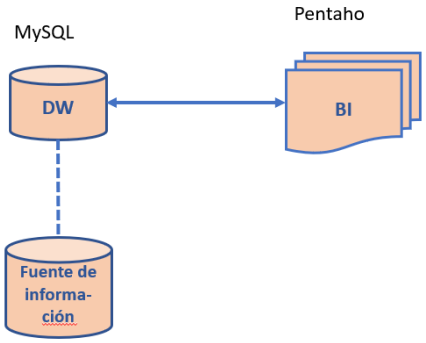


Ilustración 21, Ecosistema de la solución Versión 1.

Posterior a esta implementación, se hace necesario integrar a la solución una herramienta para realizar los procesos ETL.

4.3. Pentaho Data Integration, PDI.

Esta herramienta también es conocida como Kettle, (acrónimo recursivo: “Kettle Extraction, Transformation, Transportation, and Load Environment”).

La versión utilizada en este trabajo es la número 8.1, la cual es compatible con la versión previamente instalada de Pentaho. El archivo está disponible para descarga en esta ubicación [26].

Para su adecuado funcionamiento, se debe descomprimir el archivo pdi-ce-8.1.0.0-365.zip preferiblemente en la carpeta raíz de nuestro servidor Pentaho. Posterior a descomprimir esta carpeta, se procede a abrir el archivo Spoon.bat en una consola de Windows.

Si se tiene configurada adecuadamente la variable de entorno para el funcionamiento de Pentaho, no se tendrá ningún inconveniente en la apertura del ejecutable Java, tal como se ilustra a continuación:

```
DEBUG: Using PENTAHO_JAVA_HOME
DEBUG: _PENTAHO_JAVA_HOME=C:\Program Files\Java\jdk1.8.0_191
DEBUG: _PENTAHO_JAVA=C:\Program Files\Java\jdk1.8.0_191\bin\javaw.exe

E:\Master UOC\Semestre 4\pentaho-server-ce-8.1.0.0-365\pdi-ce-8.1.0.0-365\data-integration>start "Spoon" "C:\Program Fil
es\Java\jdk1.8.0_191\bin\javaw.exe"  "-Xms1024m" "-Xmx2048m" "-XX:MaxPermSize=256m" "-Dhttps.protocols=TLSv1,TLSv1.1,TLS
v1.2" "-Djava.library.path=libswt\win64" "-DKETTLE_HOME=" "-DKETTLE_REPOSITORY=" "-DKETTLE_USER=" "-DKETTLE_PASSWORD=" "
-DKETTLE_PLUGIN_PACKAGES=" "-DKETTLE_LOG_SIZE_LIMIT=" "-DKETTLE_JNDI_ROOT=" -jar launcher\launcher.jar -lib ..\libswt\wi
n64
```

Ilustración 22, Ejecución desde CMD de Pentaho Data Integration

Posterior a esta actividad, se inicia el .jar de Spoon, o la vista integral de Pentaho Data Integration.

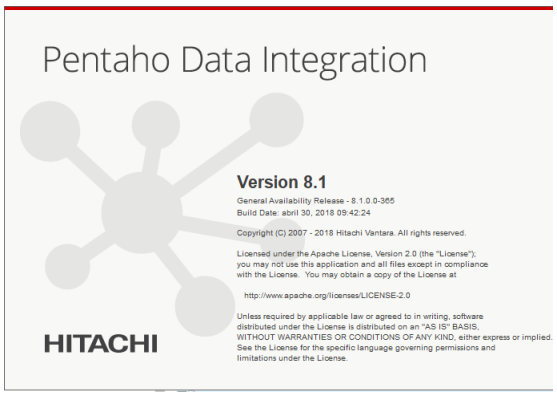


Ilustración 23, Inicio de Pentaho Data Integration

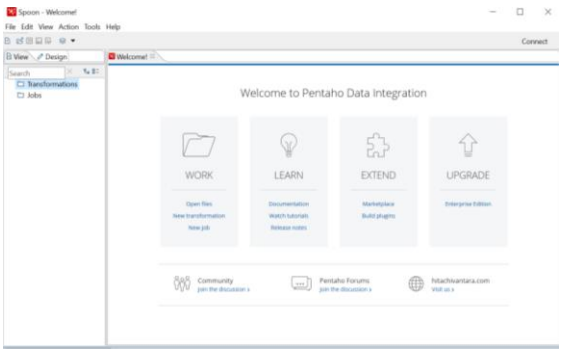


Ilustración 24, Pantalla principal de Pentaho Data Integration

Spoon, es el diseñador gráfico de transformaciones y trabajos del sistema de ETLs de PDI.

Nota: El uso de la plataforma requiere tener el conector de MySQL – Java, en la carpeta Lib, de Pentaho Data Integration.

Posterior a iniciar PDI, se crea una nueva conexión a la base de datos de nuestro sistema de BI temporal, a través de la opción del menú: File, New, Database connection.

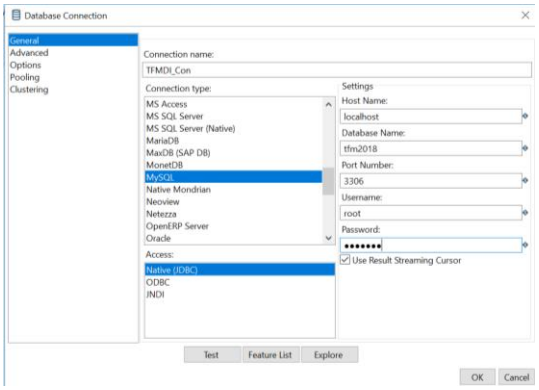


Ilustración 25, Creación de nueva conexión de base de datos en PDI.

Con el fin de verificar que la conexión es exitosa se utiliza la opción Explore en la ventana de creación de la nueva conexión, tal como se ilustra a continuación:

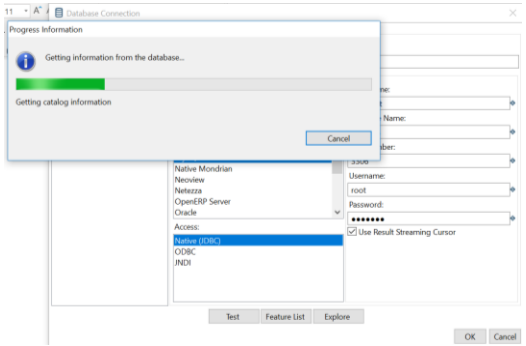


Ilustración 26, Test de la conexión a la Base de Datos en PDI.

Cuando la conexión está creada es posible visualizar la composición de nuestra base de datos en la interfaz de PDI.

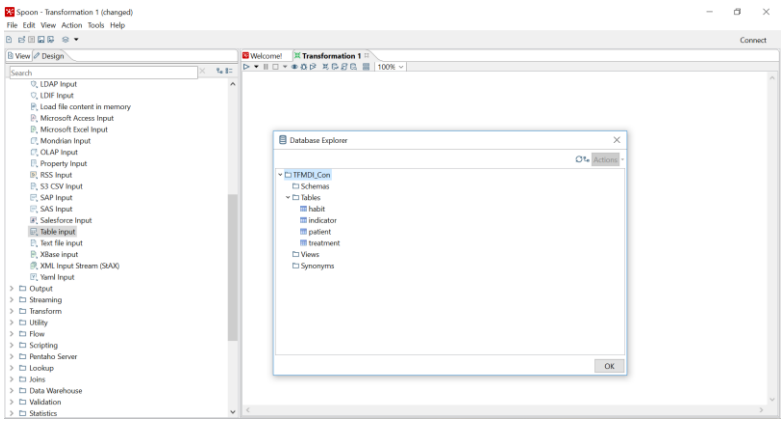


Ilustración 27, Esquema de la Base de Datos en PDI.

De manera general, se detallan a continuación los pasos a ejecutar para la creación de una transformación en PDI.

4.3.1. Creación de una transformación

En adelante, al mencionar la plataforma transformación en el contexto del software PDI, no sólo se refiere a la fase de transformación de los procesos ETL, sino a actividades relacionadas con cualquier de sus fases.

> **Nota:** el software PDI, con una wiki destinada a explicar en mayor profundidad su uso. [27]

Los pasos para crear una transformación en PDI, son:

- Seleccionar en la opción File, new, transformation.
- Sobre el menú del panel izquiero de PDI, se seleccionan los elementos requeridos para la transformación, así:

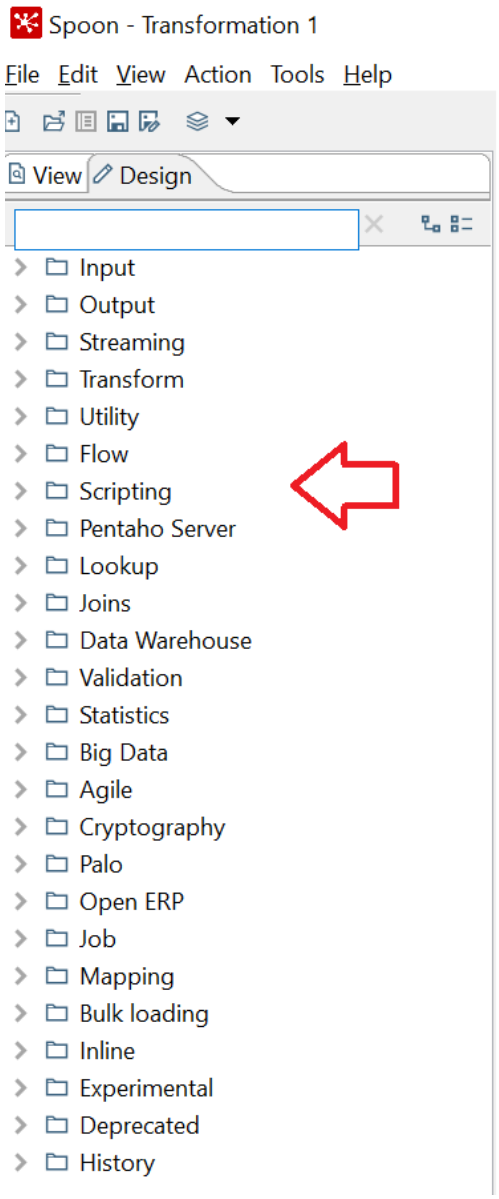


Ilustración 28, Elementos de una transformación

- Cuando se seleccione uno de los elementos y se arrastre hacia el entorno de PDI, se procede a realizar la configuración del componente, es decir, parametrizando la entrada de datos, los queries a realizar, las validaciones, el envío de correos o lo demás que se pueda ejecutar con caca uno de estos elementos.

4.3.2. Ejecución de una transformación

Para ejecutar una transformación en PDI, mínimamente se deben configurar tres elementos:

- **Elemento 1, fase Input:** fuente de la información o componente encargado de la extracción de la información requerida en la transformación.

- **Elemento 2, fase transform:** proceso de validación, cruce, modificación, concatenación, envío u otro, donde se toma como insumo la información resultante del punto 1.
- **Elemento 3, fase output:** componente de salida de la transformación realizada, es decir, donde se podrá visualizar el resultado del proceso ejecutado en el elemento 2.

En la siguiente ilustración, se evidencia el elemento 1, que corresponde a un archivo CSV, el elemento 2, correspondiente a una función de concatenación de datos extraídos desde el elemento 1. Posterior a la concatenación se establece la salida de la información en formato XML, la cual se representa en el elemento 3.

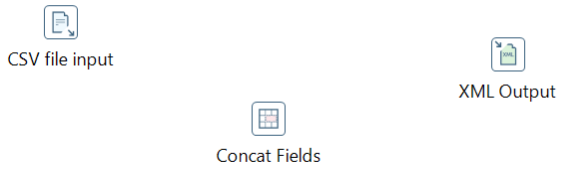


Ilustración 29, Ejemplo de componentes de una transformación

El tener definidos los elementos no garantiza un flujo u orden en la transformación, por lo tanto, se deben conectar los componentes. Para ello se selecciona la tecla ALT + clic sobre el componente a conectar y se lleva la flecha hacia el componente destino: (Esto se realiza en el orden definido).

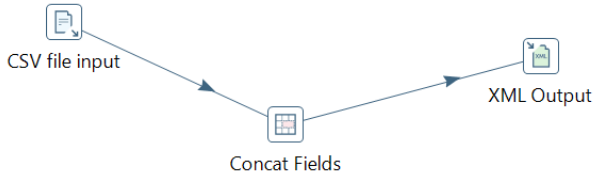


Ilustración 30, Conexión de los elementos de la transformación

Posteriormente, se debe hacer uso de la opción Preview this transformation, donde de manera previa se puede entender cuál será el resultado de la transformación. Esto ayudará a tener los componentes parametrizados para su funcionamiento.

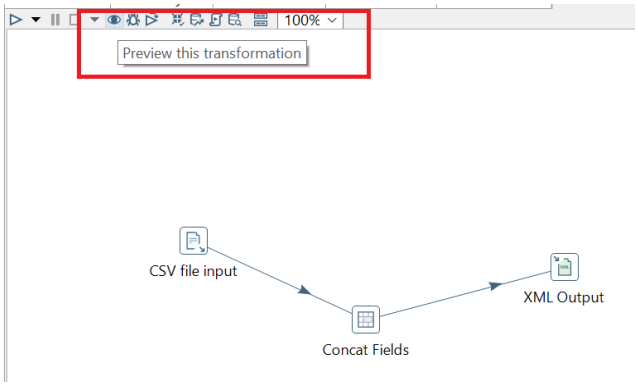


Ilustración 31, Visualización previa de la transformación

Cuando se han verificado los errores obtenidos se puede ejecutar la transformación adecuadamente y visualizar en la consola Execution Results, las métricas de la transformación, es decir, si se leyeron adecuadamente los datos, si se procesó y si se almacenó la información transformada en el elemento de salida.

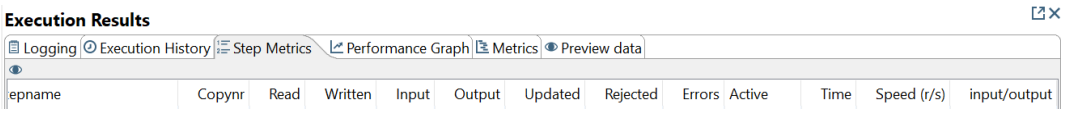


Ilustración 32, Revisión de métricas de la transformación para constatar ejecución

Al concluir el apartado de instalación de PDI, la arquitectura de nuestro sistema de BI se complementa como se ilustra a continuación:

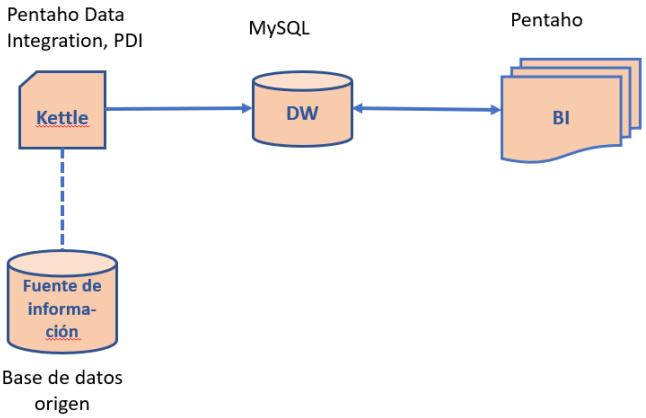


Ilustración 33, Ecosistema de la solución Versión 2

4.4. Mondrian para procesamiento analítico en línea, OLAP.

Pentaho ofrece un servidor OLAP llamado Mondrian, el cual está instalado por defecto.

La forma de corroborar esto es a través de la búsqueda del archivo de propiedades del servidor en la ruta: pentaho-solutions/System/Mondrian:

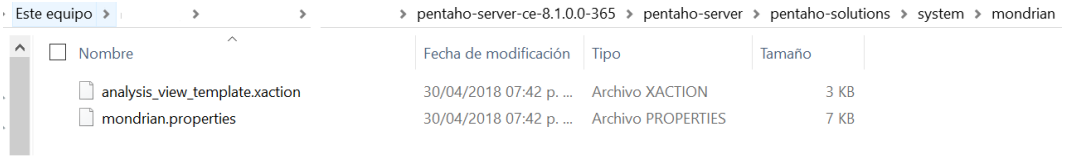


Ilustración 34, Instalación de Mondrian OLAP.

Esta herramienta es una de las más importantes de la plataforma Pentaho BI y en algunos casos también se le conoce con el nombre de: Pentaho Analysis Services. Mondrian es un servidor OLAP open source que gestiona comunicación entre una aplicación OLAP (escrita en Java) y la base de datos con los datos fuente. Es decir, Mondrian actúa como “JDBC para OLAP”. [28]

Para hacer uso de las funcionalidades provistas por dicha herramienta, se puede hacer uso de una versión gráfica que permite diseñar los cubos OLAP y está disponible en: [29] llamada **Schema WorkBench.** Para su instalación se descomprime el archivo: psw-ce-3.6.1. zip en la ruta donde se tienen las demás herramientas de Pentaho.

En la carpeta schema-workbench se encuentran los archivos necesarios para su uso, al iniciar el archivo workbench.bat, se logra acceder a esta herramienta:

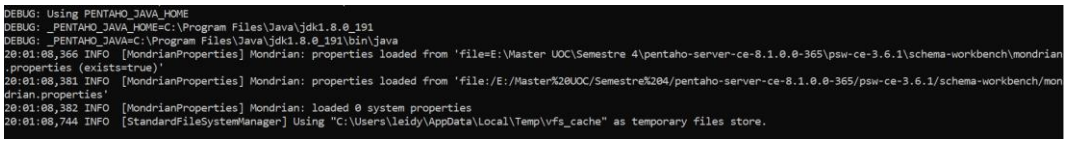


Ilustración 35, Ejecución de Schema Workbench

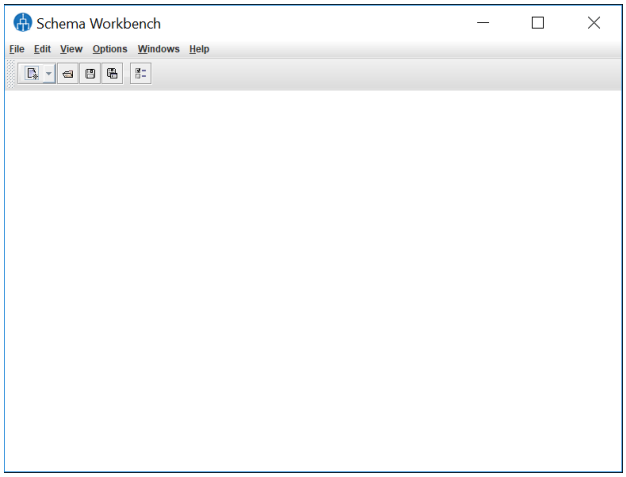


Ilustración 36, Vista inicial de Schema Workbench

Nota: para su correcto funcionamiento es necesario incluir en la carpeta drivers, el conector de MySQL en Java.

El elemento principal del sistema son los ficheros XML donde se representan los esquemas dimensionales. Para construir estos ficheros, es posible utilizar cualquier editor de texto o XML, o bien Schema Workbench. Posterior a crear los esquemas en esta herramienta, se podrá publicarlos al servidor BI para que puedan ser utilizados en los análisis por los usuarios de la plataforma.

En los ficheros de esquema XML, se describen las relaciones entre las dimensiones y medidas del cubo (modelo multidimensional), con las tablas y campos de la base de datos, a nivel relacional. Este mapeo se utiliza para ayudar la traducción de las queries MDX (que es el lenguaje con el que trabaja Mondrian), y para transformar los resultados recibidos de las consultas SQL a un formato dimensional. Vamos a ver a continuación como utilizar PSW para definir los esquemas del proyecto y publicar los resultados en el servidor BI. [30]

En adelante para el uso de la herramienta se debe configurar una conexión a la base de datos del DW:

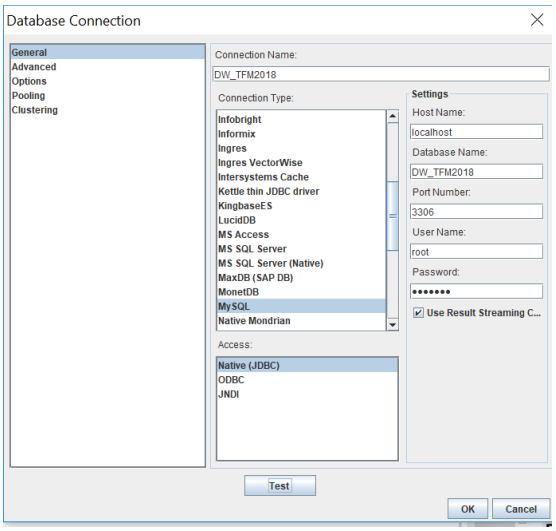


Ilustración 37, Creación de la conexión al DW para PSW.

En la conexión se ingresan los datos básicos de la base de datos donde se incluye el nombre, el servidor de la conexión, el puerto y el usuario y contraseña.

Al concluir la instalación de Mondrian dentro de la suite de herramientas de Pentaho, se actualiza el ecosistema de la solución:

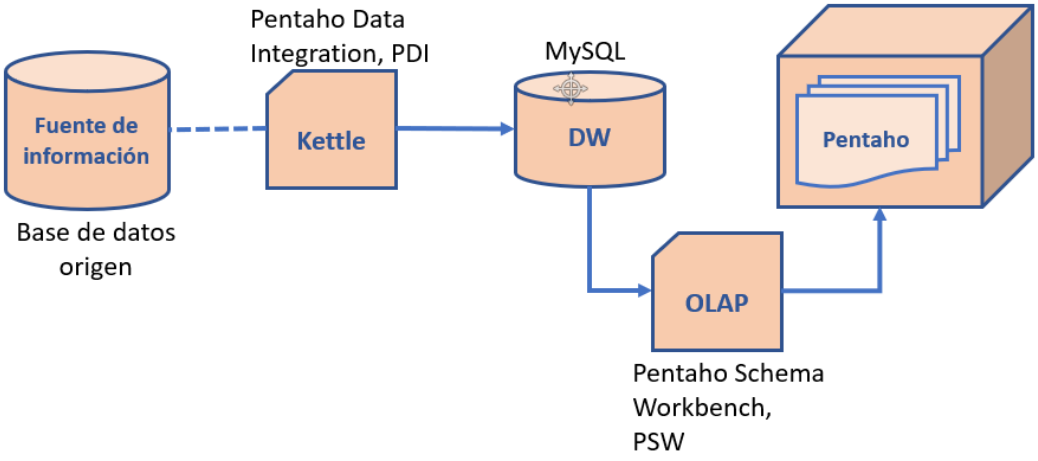


Ilustración 38, Ecosistema de la solución Versión 3.

4.5. SAIKU visualización de cubos OLAP

La herramienta Saiku provee un sitio web para la obtención de la licencia de tipo community o libre y se accede a través de [31]. Allí es posible crear una cuenta para que vía correo electrónico se pueda obtener el archivo con la licencia.

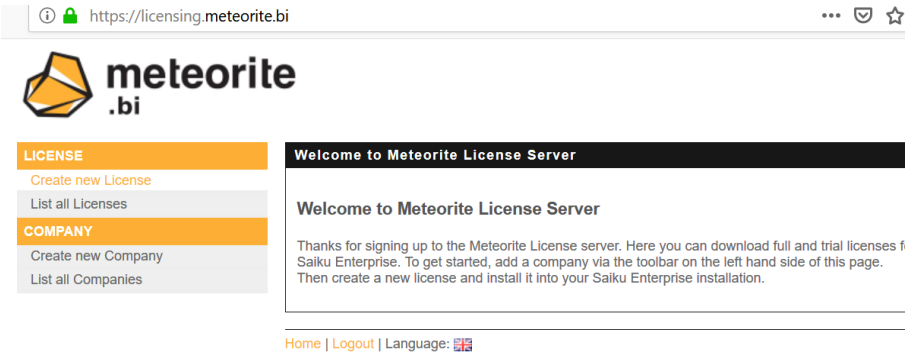


Ilustración 39, Obtención de la licencia de Saiku

Posterior a diligenciar la información solicitada se puede proceder a descargar la licencia:

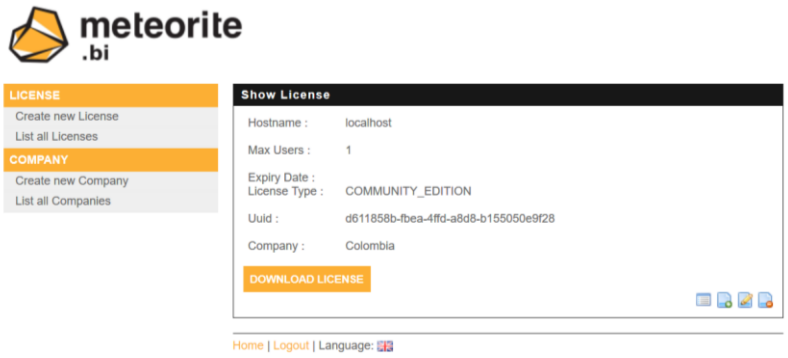


Ilustración 40, Licencia de Saiku

Al descargar la licencia esta debe ser almacenada sobre la raíz de la carpeta de Saiku con el nombre license.lic.

Aunque Saiku se puede usar como un cliente instalado en una máquina, también puede ser integrado a Pentaho siguiente los pasos descritos a continuación:

1. Se descarga el plugin de Saiku desde el Market Place de Pentaho. El cual debe ser compatible con la versión 8.1 [32]
2. Se descomprime el archivo: saiku-plugin-p7.1-3.90.zip en la ruta \pentaho-server\pentaho-solutions\system\saiku
3. Se accede al portal de Pentaho a través del Localhost si es el caso y en la opción File → New, se selecciona, Saiku Analytics.

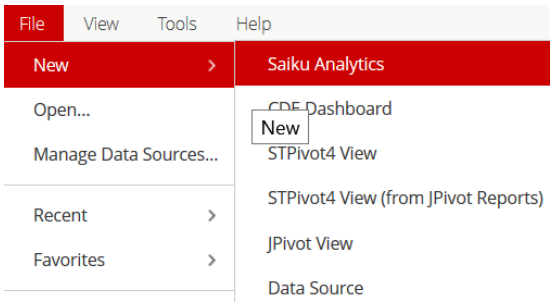


Ilustración 41, Opción de Saiku en Pentaho

Posteriormente se puede visualizar la siguiente pantalla:

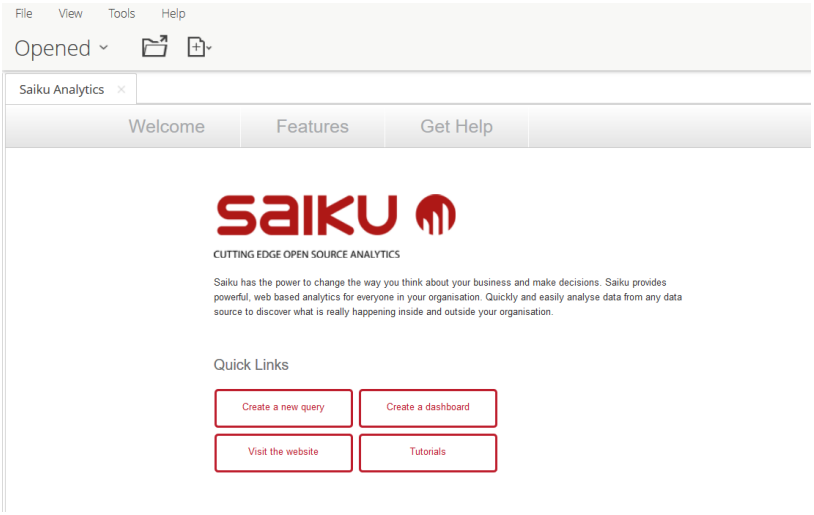


Ilustración 42, Pantalla principal de Saiku Analytics

En la opción *"Create a new query"* se puede navegar sobre el cubo OLAP requerido. Finalmente, se logra visualizar la siguiente pantalla donde arrastrando los elementos definidos en el cubo hacia los campos de medidas, filas y columnas, se logra visualizar la información.

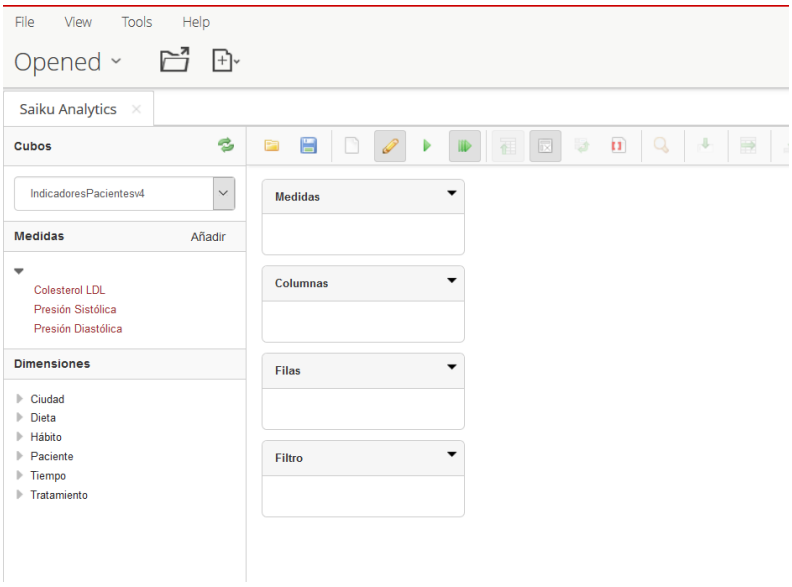


Ilustración 43, Entorno de navegación de SAIKU.

La arquitectura final implementada para el proyecto TFM es:

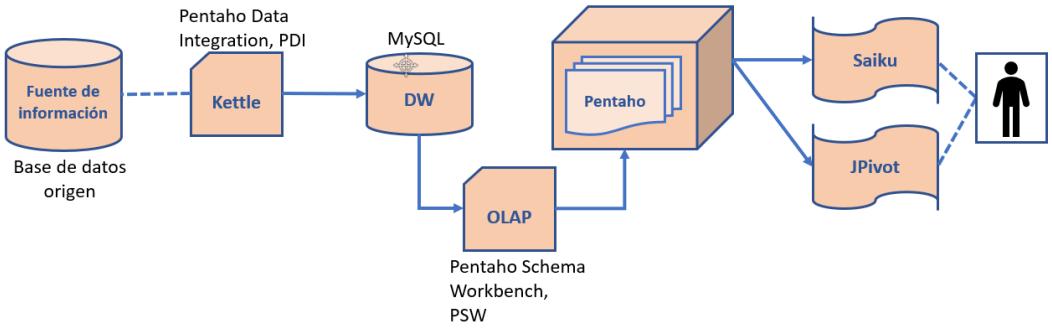


Ilustración 44, Arquitectura de BI completa

En este caso se implementó una arquitectura lineal ya que sólo se tiene una instancia del servidor local. Las implementaciones en alta disponibilidad o clúster pueden realizarse para otros proyectos basándose en este tipo de diseño.

5. Data Warehouse diseño lógico

Uno de los aspectos más relevantes en el momento de diseñar un sistema de BI, es pasar desde el diseño relacional al dimensional, esto lo sugiere Kimball en su metodología para la construcción de un data warehouse.

El diseño relacional por su parte se refiere a la estructura de la fuente de datos origen de nuestra información y el dimensional, a la estructura vista desde los procesos y la granularidad de los análisis requeridos para dar respuesta a las preguntas analíticas del cliente del sistema de BI.

5.1. Análisis dimensional

Para llevar a cabo el análisis dimensional del problema planteado en este proyecto, se pueden seguir los siguientes pasos descritos en la metodología de Kimball: [5]

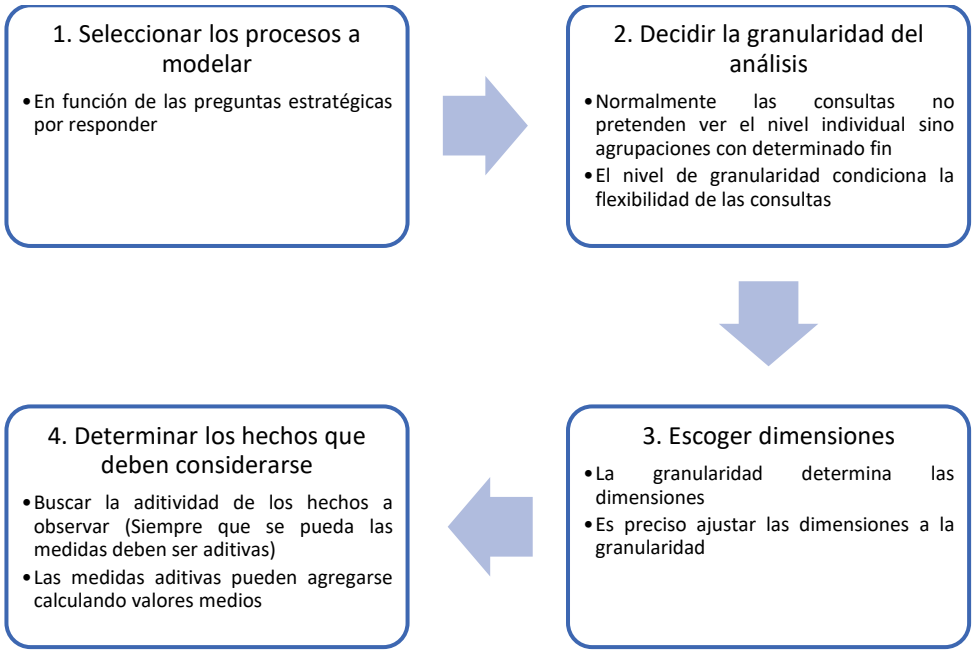


Ilustración 45, Metodología Kimball para el análisis dimensional [5]

Paso 1: en función de los preguntas analíticas o estratégicas que están planteadas en el proyecto, y considerando que el esquema de información no contiene datos que no están directamente ligados con el análisis de los tratamientos para la disminución de los niveles de colesterol, el único proceso a modelar es el ya mencionado en el enfoque del proyecto.

Paso 2: la granularidad del análisis será la máxima posible ya que las preguntas analíticas planteadas requieren abarcar los aspectos más específicos de cada paciente frente a cómo los tratamientos suministrados han mejorado o no su salud.

Paso 3: por su parte en la elección de las dimensiones es importante entender que una dimensión es una forma, vista o criterio por medio del cual se pueden sumariar, cruzar o cortar datos numéricos a analizar. Estos datos se denominan medidas (Measures en inglés).

Dichas dimensiones surgen naturalmente de las discusiones del equipo de trabajo y [5] los atributos de estas o campos textuales se pueden utilizar como campos de las tablas del DW. Lo anterior considerando que estos atributos se convierten muchos casos, como lo afirma Kimball, en fuente de reporte o monitoreo en el sistema de BI.

El siguiente gráfico de burbujas, (Bubble chart, en inglés), expone el conjunto de dimensiones objetivo del proyecto:

Ilustración 46, Gráfico de burbujas (Lenguaje Kimball), Análisis dimensional

En este gráfico se evidencian varias dimensiones relevantes: los tratamientos que se suministran a los pacientes, los datos de los pacientes, su dieta y actividad, el lugar donde están ubicados y el tiempo en que se desarrollan los hábitos e indicadores de estos. El centro del análisis dimensional son los indicadores de salud de los pacientes (Colesterol LDL, presión sistólica y diastólica).

5.2. Arquitectura del Data Warehouse

Actualmente, por ejemplo, para encontrar respuesta a cuál es la relación entre los diferentes tratamientos y la evolución de los pacientes, se tendría que realizar una consulta compleja sobre el modelo relacional que detalle cómo un paciente ha modificado sus indicadores de acuerdo con los hábitos y actividad realizados.

Para resolver esta complejidad que pueda darse al querer dar respuesta a los planteamientos analíticos del proyecto, se debe acudir a diseñar una arquitectura simple para el DW, la cual pueda centralizar adecuadamente la información. En este caso, se ha seleccionado la arquitectura tipo **Estrella** *(Star schema, en inglés).*

En este diseño del DW la tabla de variables (Hechos) está rodeada por dimensiones y juntos forman una estructura que permite implementar mecanismos básicos para poder utilizarla con una herramienta de consultas OLAP. [33]

Una tabla de hechos, (o *tabla fact en inglés*), es la tabla central de un esquema dimensional, y contiene los valores de las medidas de negocio o dicho de otra forma los indicadores de negocio. Cada medida se toma mediante la intersección de las dimensiones que la definen, dichas dimensiones estarán reflejadas en sus correspondientes tablas de dimensiones que rodearán la tabla de hechos y estarán relacionadas con ella. [33]

Para este proyecto, se ha decidido implementar un DW considerando la tabla **Indicator** como la **tabla de hechos**, la cual contiene los indicadores claves para el negocio: el nivel de colesterol (LDL), la presión sistólica y diastólica, la dieta, actividad y fecha de esta última. Lo anterior se da a partir de relaciones establecidas con las dimensiones.

Del modelo relacional expuesto en la ***Ilustración 8, Diagrama entidad-relación***. Pasamos al siguiente esquema dimensional en estrella:

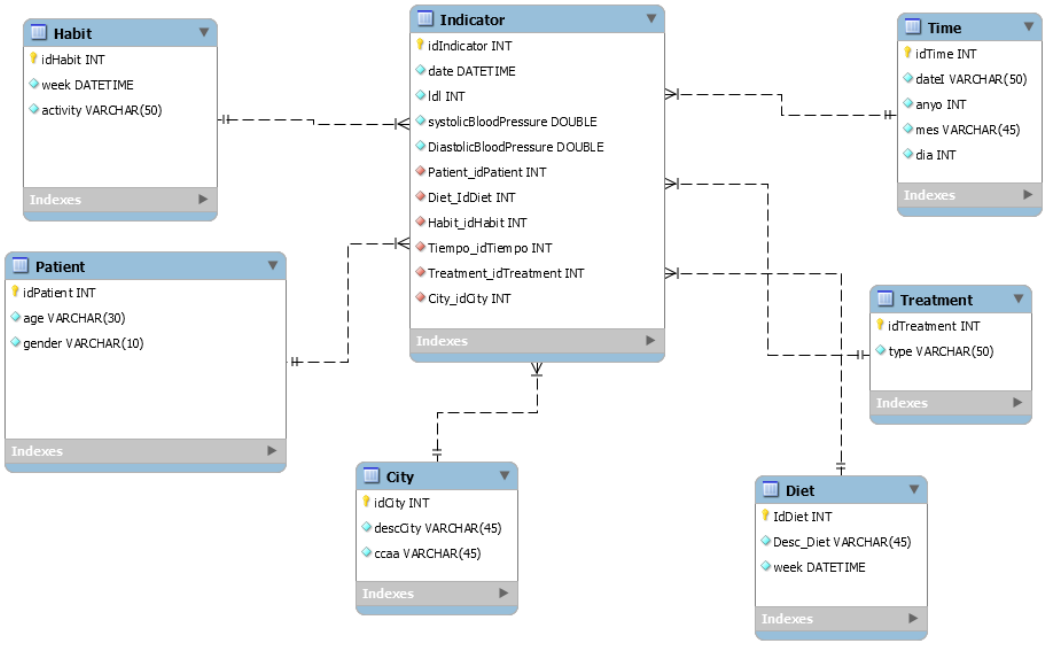


*Ilustración 47, Modelo dimensional en **Estrella**. Realizado con MySQL Workbench*

Este diseño es implementado en apartados posteriores a la extracción, transformación y carga de los datos.

6. Diseño e implementación de ETL's

Los procesos **ETL**, o de **E**xtracción, **T**ransformación y **C**arga, son el conjunto de procesos mediante el cual los datos se extraen de numerosas bases de datos, aplicaciones y sistemas, se transforman según corresponda y se cargan en los sistemas de destino, incluidos, entre otros, almacenes de datos, mercados de datos y análisis. aplicaciones, etc. [34] En resumen, son la base sobre la cual se alimenta el data warehouse.

A continuación, se detallan los componentes de los procesos ETL:

Tabla 2, Descripción de los componentes del proceso ETL. [35]

Componente	Elementos Objetivo (Entrada)	Operaciones realizadas (Proceso)	Resultado de la tarea (Salida)
Extracción	Fuentes de información como bases de datos de sistemas transaccionales, archivos de texto, hojas de cálculo.	Selección de los datos	Datos crudos (cargados en memoria)
Transformación	Datos crudos (cargados en memoria)	Limpieza, transformación, personalización, realización de cálculos y aplicación de	Datos formateados, estructurados y resumidos de acuerdo con las necesidades (aún en memoria).

		funciones de agregación.	
Carga	Datos formateados, estructurados y resumidos de acuerdo con las necesidades (aún en memoria).	Inserción en el data warehouse.	Datos formateados, estructurados y resumidos con persistencia en el data warehouse.

El principal objetivo de diseñar los procesos ETL, es extraer y consolidar los datos de diferentes fuentes de información, (Por ejemplo, para nuestro caso sólo tenemos la base de datos MySQL), aplicar reglas para incrementar la calidad de los datos y su consistencia y finalmente, guardar en nuestro data warehouse los datos en el formato acorde para que las demás herramientas que integran nuestra solución de BI puedan operar de manera adecuada.

El objetivo de implementar los escenarios ETL es identificar y mitigar los defectos de los datos en el momento de la extracción y verificar adecuadamente las transformaciones planteadas para estos.

Para materializar estos escenarios, se debe instalar una herramienta llamada: Pentaho Data Integration, PDI. El proceso de instalación se detalla en: **Pentaho Data Integration, PDI.**

6.1.1. Extracción

Existen diferentes metodologías para modelar de manera gráfica escenarios ETL [36], sin embargo, entendiendo que nuestros datos ya cumplen con unos estándares mínimos de consistencia y calidad, se han establecido algunos escenarios básicos en todo proceso de extracción de información. Los escenarios se listan y se prueban a continuación:

I. Calidad de los datos.

Se refiere a validar el correcto formato y/o correcta consistencia de los datos.

En el modelo de datos detallado en apartados anteriores, se cuenta con dos atributos cuyo formato corresponde a: dd/mm/aaaa, y son respectivamente: Date (Tabla Indicator), Week (Tabla Habit).

Una de las opciones más sencillas de validar la información, es a través de la consola SQL de PDI, y para esto se utilizan las siguientes sentencia SQL:

Tabla Habit:

```
SELECT * FROM Habit

WHERE week >= STR_TO_DATE('1/01/2017', '%d/%m/%Y')  AND

week NOT REGEXP '^[0-9\.]+$';
```

Tabla Indicator:

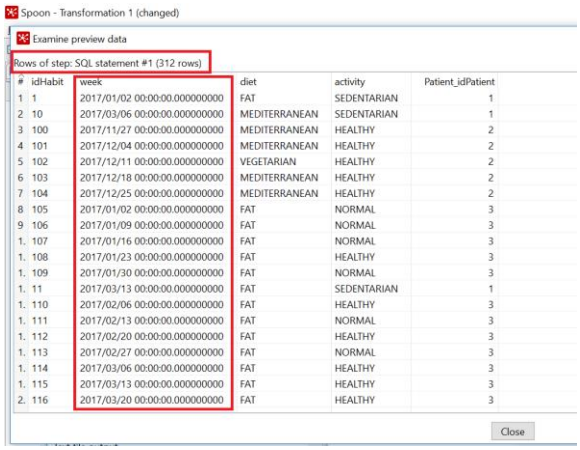
```
SELECT * FROM Indicator

WHERE dateI >= STR_TO_DATE('1/01/2017', '%d/%m/%Y')  AND
```

datel NOT REGEXP ‘^[0-9\.]+$’;

Las sentencias validan dos aspectos:

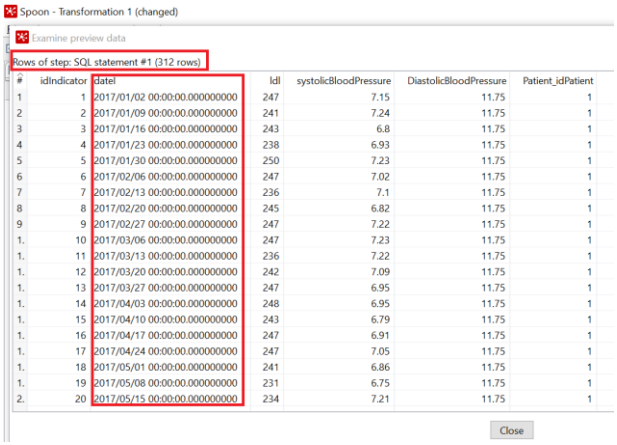
- Que el campo fecha presente una estructura dd/mm/yyyy, lo cual se hace a partir de la comparación >= del campo.
- Que el campo que contiene los valores en formato de fecha no tenga números, lo cual se valida con la expresión NOT REGEXP.

Las siguientes ilustraciones exponen el resultado de la verificación de este escenario mediante queries:



#	idHabit	week	diet	activity	Patient_idPatient
1	1	2017/01/02 00:00:00.000000000	FAT	SEDENTARIAN	1
2	10	2017/03/06 00:00:00.000000000	MEDITERRANEAN	SEDENTARIAN	1
3	100	2017/11/27 00:00:00.000000000	MEDITERRANEAN	HEALTHY	2
4	101	2017/12/04 00:00:00.000000000	MEDITERRANEAN	HEALTHY	2
5	102	2017/12/11 00:00:00.000000000	VEGETARIAN	HEALTHY	2
6	103	2017/12/18 00:00:00.000000000	MEDITERRANEAN	HEALTHY	2
7	104	2017/12/25 00:00:00.000000000	MEDITERRANEAN	HEALTHY	2
8	105	2017/01/02 00:00:00.000000000	FAT	NORMAL	3
9	106	2017/01/09 00:00:00.000000000	FAT	NORMAL	3
1.	107	2017/01/16 00:00:00.000000000	FAT	NORMAL	3
1.	108	2017/01/23 00:00:00.000000000	FAT	HEALTHY	3
1.	109	2017/01/30 00:00:00.000000000	FAT	NORMAL	3
1.	11	2017/03/13 00:00:00.000000000	FAT	SEDENTARIAN	1
1.	110	2017/02/06 00:00:00.000000000	FAT	HEALTHY	3
1.	111	2017/02/13 00:00:00.000000000	FAT	NORMAL	3
1.	112	2017/02/20 00:00:00.000000000	FAT	HEALTHY	3
1.	113	2017/02/27 00:00:00.000000000	FAT	NORMAL	3
1.	114	2017/03/06 00:00:00.000000000	FAT	HEALTHY	3
1.	115	2017/03/13 00:00:00.000000000	FAT	HEALTHY	3
2.	116	2017/03/20 00:00:00.000000000	FAT	HEALTHY	3

Ilustración 48, Resultado de verificación escenario de extracción 1.



#	idIndicator	datel	ldl	systolicBloodPressure	DiastolicBloodPressure	Patient_idPatient
1	1	2017/01/02 00:00:00.000000000	247	7.15	11.75	1
2	2	2017/01/09 00:00:00.000000000	241	7.24	11.75	1
3	3	2017/01/16 00:00:00.000000000	243	6.8	11.75	1
4	4	2017/01/23 00:00:00.000000000	238	6.93	11.75	1
5	5	2017/01/30 00:00:00.000000000	250	7.23	11.75	1
6	6	2017/02/06 00:00:00.000000000	247	7.02	11.75	1
7	7	2017/02/13 00:00:00.000000000	236	7.1	11.75	1
8	8	2017/02/20 00:00:00.000000000	245	6.82	11.75	1
9	9	2017/02/27 00:00:00.000000000	247	7.22	11.75	1
1.	10	2017/03/06 00:00:00.000000000	247	7.23	11.75	1
1.	11	2017/03/13 00:00:00.000000000	236	7.22	11.75	1
1.	12	2017/03/20 00:00:00.000000000	242	7.09	11.75	1
1.	13	2017/03/27 00:00:00.000000000	247	6.95	11.75	1
1.	14	2017/04/03 00:00:00.000000000	248	6.95	11.75	1
1.	15	2017/04/10 00:00:00.000000000	243	6.79	11.75	1
1.	16	2017/04/17 00:00:00.000000000	247	6.91	11.75	1
1.	17	2017/04/24 00:00:00.000000000	247	7.05	11.75	1
1.	18	2017/05/01 00:00:00.000000000	241	6.86	11.75	1
1.	19	2017/05/08 00:00:00.000000000	231	6.75	11.75	1
2.	20	2017/05/15 00:00:00.000000000	234	7.21	11.75	1

Ilustración 49, Resultado de verificación escenario de extracción 1.

Así mismo, se puede realizar también una automatización de este escenario mediante una transformación, la cual contiene los siguientes elementos:

- Un elemento de tipo input: **Table Input***, encargado de leer el valor del campo week en la tabla Indicator.
- Un elemento del tipo validation: **Data Validator***, encargado de ejecutar una regla de validación del formato del campo.

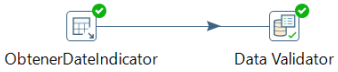


Ilustración 50, Elementos de la transformación

Nota**: los nombres de los elementos son tomados como aparecen en el menú del panel izquierdo de PDI y son totalmente personalizables. Tal como se evidenció en la **Ilustración 28, Elementos de una transformación.

Para configurar el elemento de tipo Validation, se utiliza la opción new Validation y se proporciona un nombre.

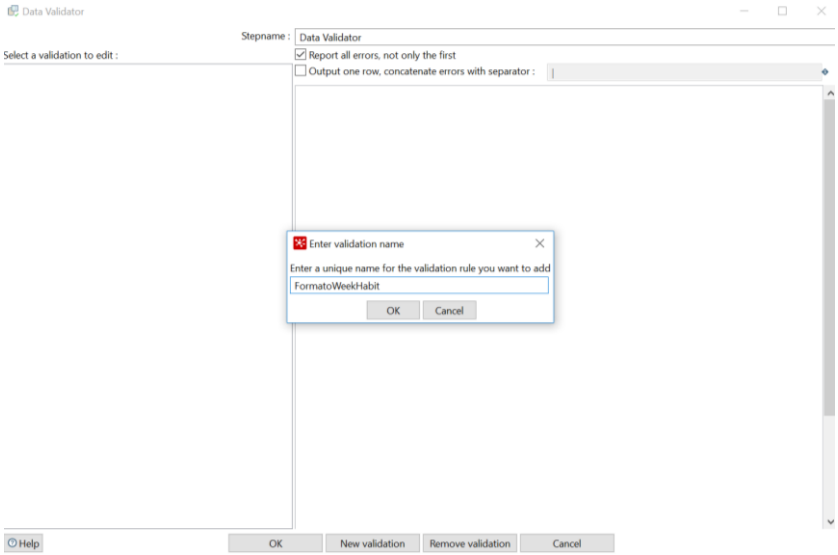


Ilustración 51, Creación de regla de validación formato fecha 1.0

La documentación básica de esta regla creada para la ejecución del escenario es:

- Descripción de la validación.
- Campo por validar.
- Descripción del error que arrojaría la validación cuando no coincide la información.
- El tipo de dato.
- Admisión de valores no nulos.

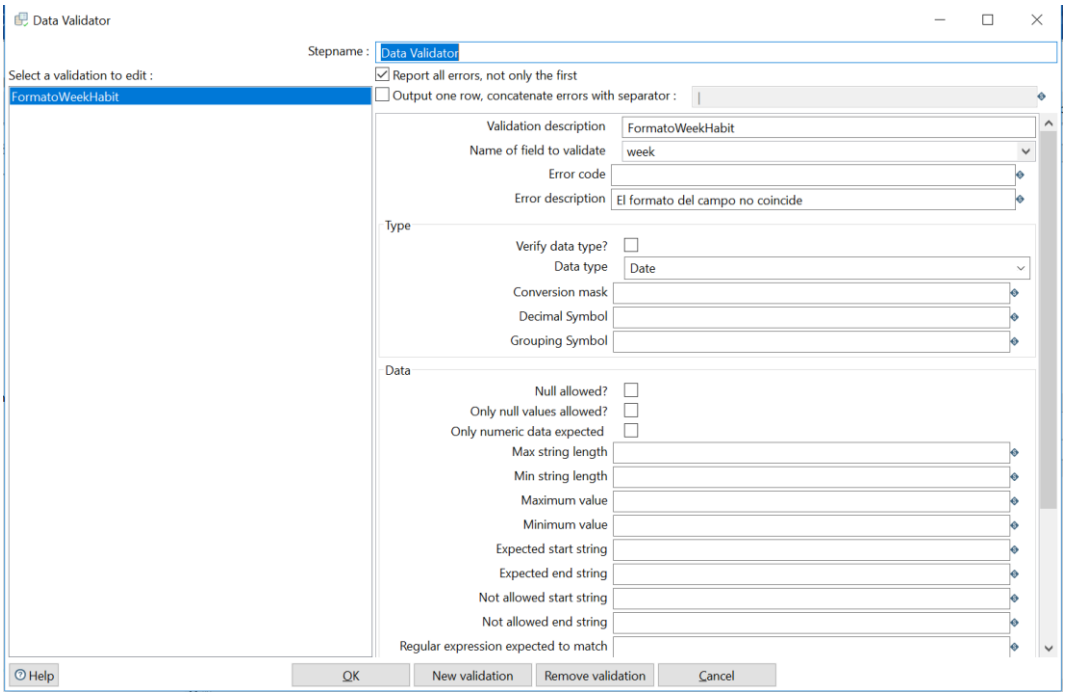
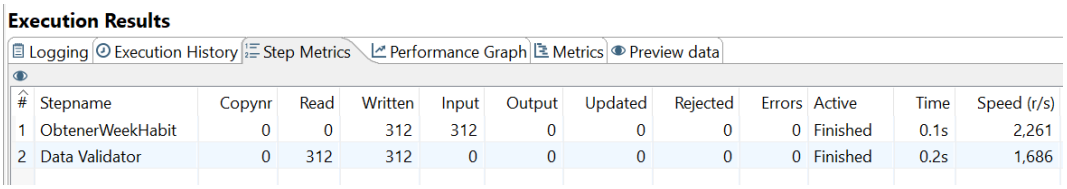


Ilustración 52, Parametrización de la regla de validación, escenario 1.0

Al ejecutar la transformación y visualizar las métricas, se evidencia que todos los datos presentan el formato adecuado.



#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)
1	ObtenerWeekHabit	0	0	312	312	0	0	0	0	Finished	0.1s	2,261
2	Data Validator	0	312	312	0	0	0	0	0	Finished	0.2s	1,686

Ilustración 53, Resultado de la validación, escenario 1.0

Este proceso se repite para el campo date de la tabla Indicator, donde el resultado es:

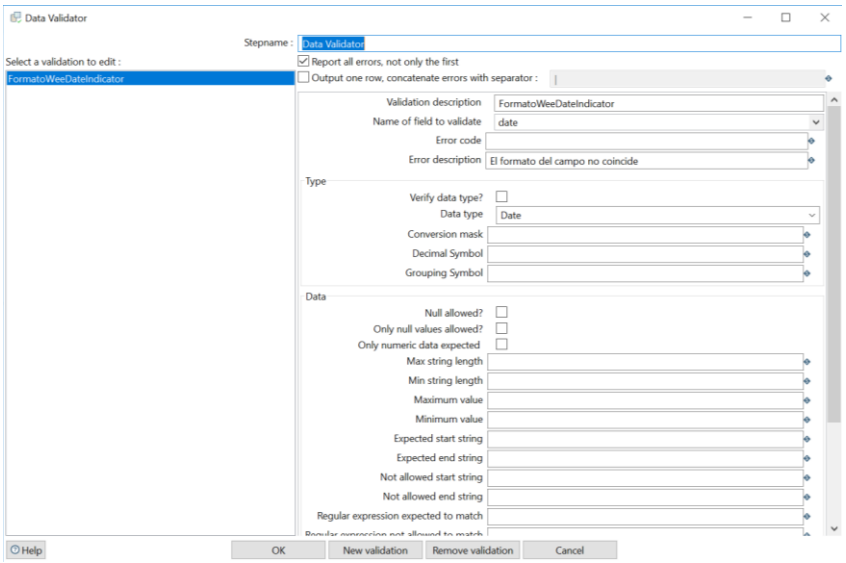
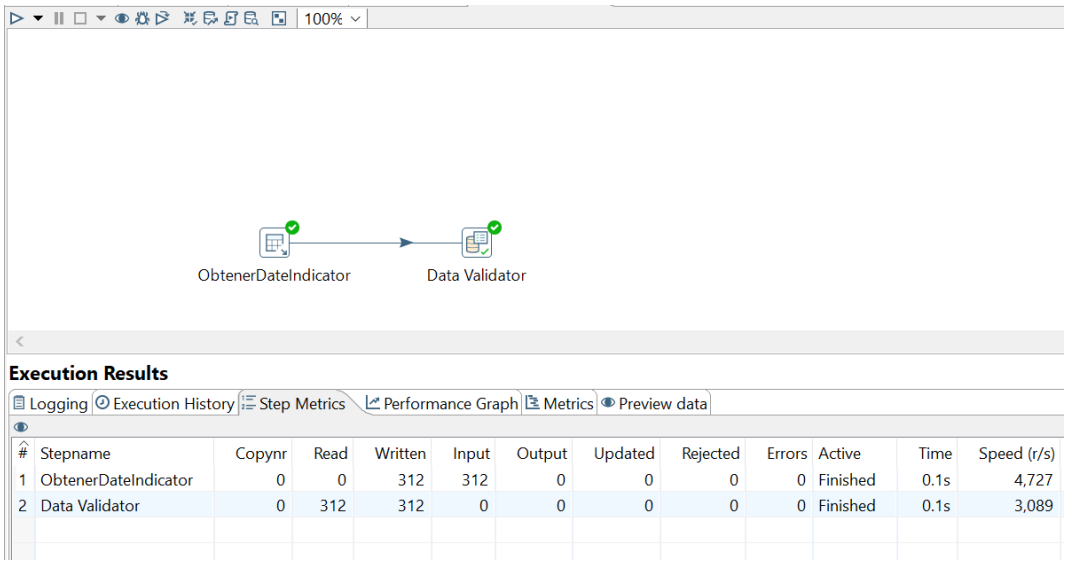


Ilustración 54, Parametrización de la regla de validación, escenario 1.1



#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)
1	ObtenerDateIndicator	0	0	312	312	0	0	0	0	Finished	0.1s	4,727
2	Data Validator	0	312	312	0	0	0	0	0	Finished	0.1s	3,089

Ilustración 55, Resultado de la validación, escenario 1.1

II. Valores no nulos.

Ningún campo del modelo debe ser nulo, por lo tanto, se debe constatar donde puede haber efectivamente un valor nulo cuando se había determinado que no lo debe estar.

Con ayuda de la siguiente sentencia se puede realizar una validación simple:

```
SELECT * FROM NombreTabla
WHERE Atributo=NULL;
```

Nota: esta sentencia se expuso de manera genérica puesto que la validación por realizar se debe efectuar de la misma manera para todos los atributos de nuestro interés. En este caso al validar no se han encontrado valores nulos en las tablas Habit e Indicator.

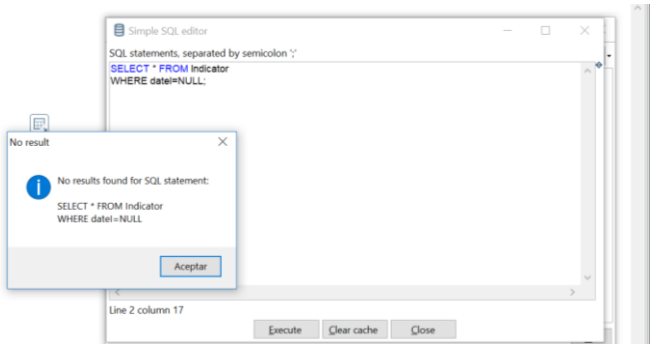


Ilustración 56, Resultado de verificación escenario de extracción 2.

Adicionalmente, esto puede automatizar mediante la creación de reglas en un componente del tipo Data Validation, las cuales contienen la siguiente información por cada atributo de la tabla Indicator que en este caso tomamos como ejemplo:

- Descripción de la validación.
- Descripción del error.

- En la sección Data, se debe retirar la marca que existe sobre los campos de admitir valores null.

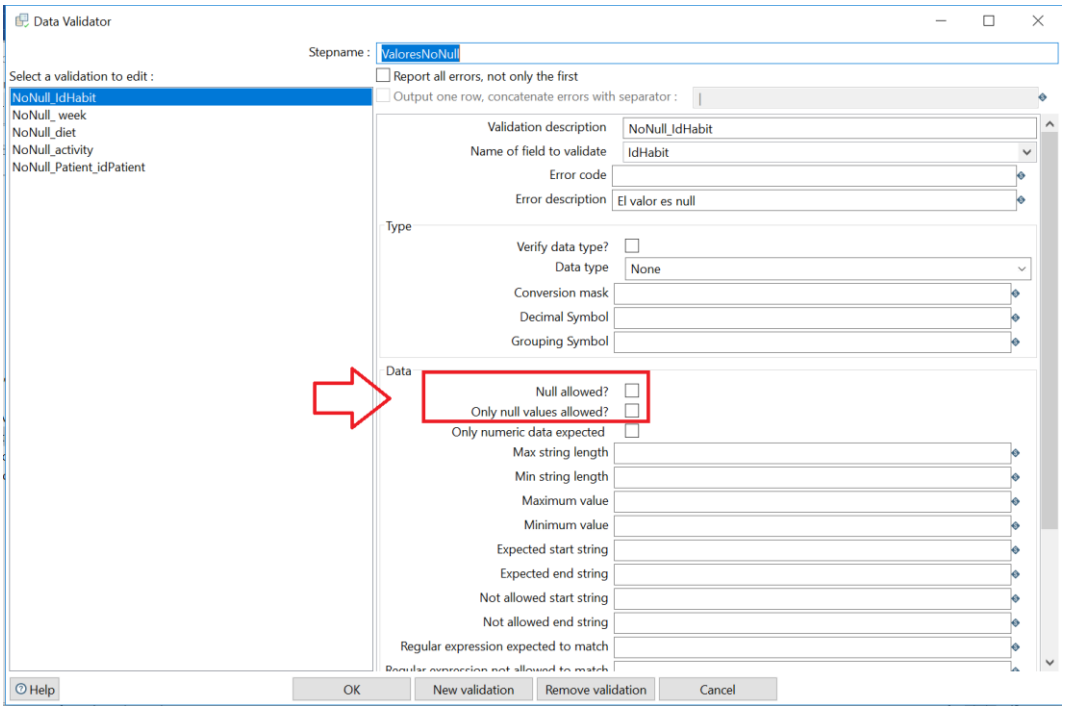


Ilustración 57, Regla para validación de valores no null.

Posterior a crear cada una de las reglas se procede a enlazar los componentes y ejecutar la transformación entendiendo que se debe realizar por cada atributo. Para este caso, se ha seleccionado el atributo diet, correspondiente a la dieta de cada paciente.

Lo anterior, porque en el momento de conectar los componentes PDI, valida la regla que se requiere ejecutar:

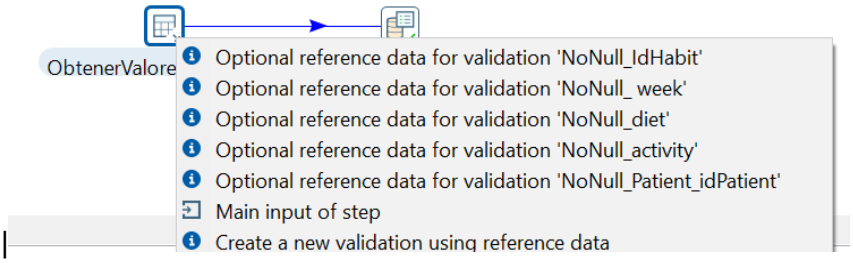


Ilustración 58, Reglas para la validación.

Los resultados son:

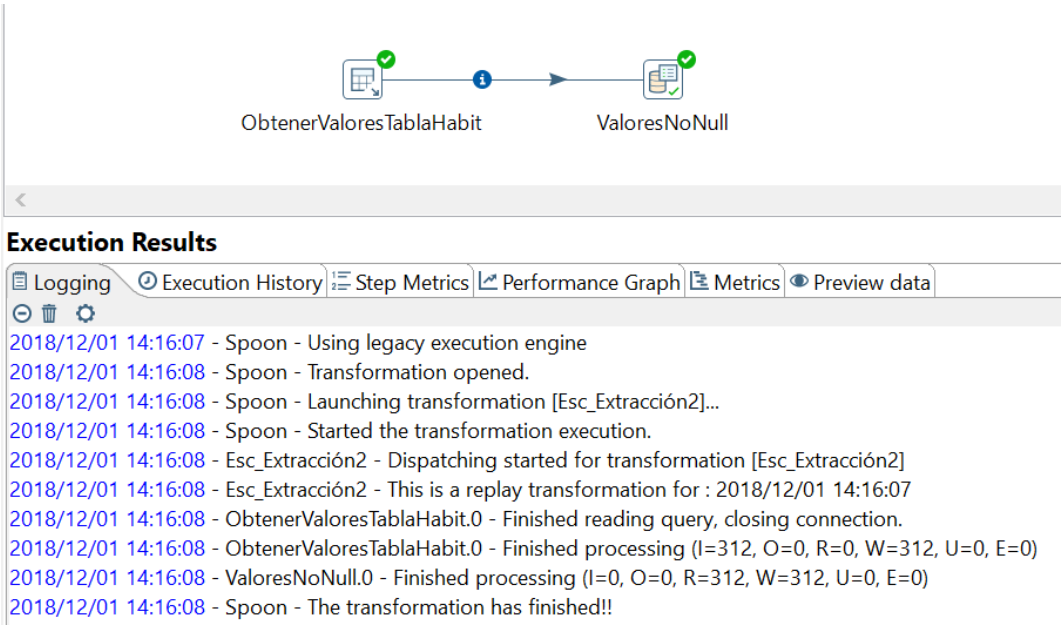


Ilustración 59, Resultado de verificación escenario de extracción 2.

De acuerdo con la anterior ilustración, ningún valor es nulo, de lo contrario, como resultado de la validación se mostraría un mensaje en color Rojo donde indique la situación y este mensaje contendría el texto especificado en la regla de validación creada.

III. Valores no duplicados.

Este escenario no se debe replicar para las claves primarias y foráneas de las tablas de la base de datos, ya que en la carga de información al sistema Data warehouse, no se repiten. Este es un control a nivel de motor de la base de datos. Sin embargo, sobre los demás atributos como, por ejemplo: treatment, diet, activity, LDL, Systolic Blood Pressure, Diastolic Blood Pressure, es normal de acuerdo con el contexto y conocimiento del proyecto, que estos datos se repitan.

Ahora bien, un dato que no debe aparecer duplicado es un hábito o indicador de un paciente para una misma fecha. Para validar esta situación se ejecuta la siguiente sentencia:

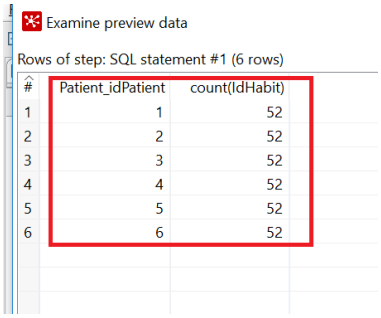
```
SELECT  Patient_idPatient, count(IdIndicator)
FROM Indicator
GROUP BY Patient_idPatient
HAVING COUNT(*) > 1
```

Esta sentencia permite obtener aquellos registros que tienen el mismo valor, en todos los campos especificados en el GROUP BY y que se repitan más de una vez (HAVING COUNT).

Así mismo, se ejecuta dicha sentencia para la tabla Habit:

```
SELECT  Patient_idPatient, count(IdHabit)
FROM Habit
GROUP BY Patient_idPatient
HAVING COUNT(*) > 1
```

El resultado se ilustra a continuación:



#	Patient_idPatient	count(IdHabit)
1	1	52
2	2	52
3	3	52
4	4	52
5	5	52
6	6	52

Ilustración 60, Resultado de verificación escenario de extracción 3.

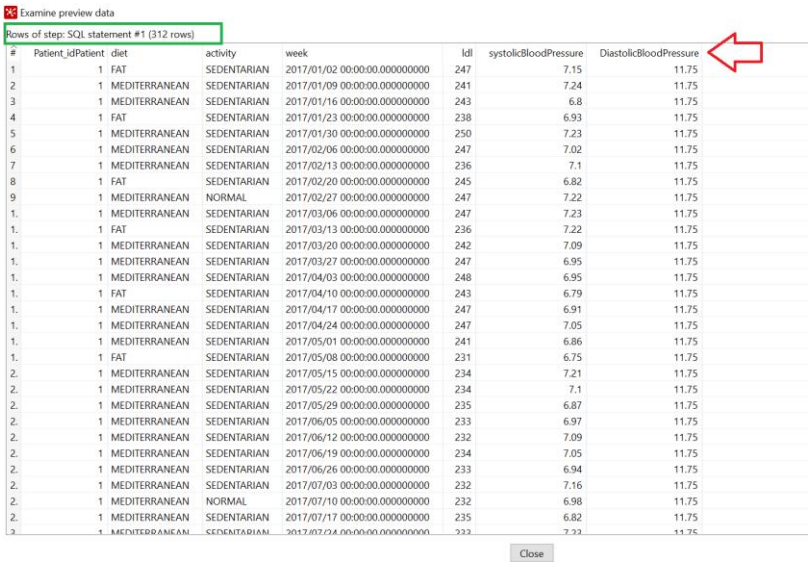
IV. Correspondencia de los datos de acuerdo con el caso de negocio.

En este proyecto, cada paciente tiene un indicador (Nivel de la presión y colesterol) que está asociado a una dieta y una actividad en un día específico. Para ello se debe garantizar que existe una correspondencia entre la fecha en la que se tuvo una dieta específica y el indicador medido.

Con este fin se utiliza el siguiente query:

```
SELECT H.Patient_idPatient, H.diet, H.activity, H.week, I.ldl,
I.systolicBloodPressure, I.DiastolicBloodPressure

FROM indicator I, habit H

where I.Patient_idPatient = H.Patient_idPatient and

H.week=I.dateI;
```

La sentencia corrobora la correspondencia de las fechas y de los pacientes. Posterior a su ejecución se evidencian los datos de sus hábitos e indicadores.



#	Patient_idPatient	diet	activity	week	ldl	systolicBloodPressure	DiastolicBloodPressure
1	1	FAT	SEDENTARIAN	2017/01/02 00:00:00.000000000	247	7.15	11.75
2	1	MEDITERRANEAN	SEDENTARIAN	2017/01/09 00:00:00.000000000	241	7.24	11.75
3	1	MEDITERRANEAN	SEDENTARIAN	2017/01/16 00:00:00.000000000	243	6.8	11.75
4	1	FAT	SEDENTARIAN	2017/01/23 00:00:00.000000000	238	6.93	11.75
5	1	MEDITERRANEAN	SEDENTARIAN	2017/01/30 00:00:00.000000000	250	7.23	11.75
6	1	MEDITERRANEAN	SEDENTARIAN	2017/02/06 00:00:00.000000000	247	7.02	11.75
7	1	MEDITERRANEAN	SEDENTARIAN	2017/02/13 00:00:00.000000000	236	7.1	11.75
8	1	FAT	SEDENTARIAN	2017/02/20 00:00:00.000000000	245	6.82	11.75
9	1	MEDITERRANEAN	NORMAL	2017/02/27 00:00:00.000000000	247	7.22	11.75
1.	1	MEDITERRANEAN	SEDENTARIAN	2017/03/06 00:00:00.000000000	247	7.23	11.75
1.	1	FAT	SEDENTARIAN	2017/03/13 00:00:00.000000000	236	7.22	11.75
1.	1	MEDITERRANEAN	SEDENTARIAN	2017/03/20 00:00:00.000000000	242	7.09	11.75
1.	1	MEDITERRANEAN	SEDENTARIAN	2017/03/27 00:00:00.000000000	247	6.95	11.75
1.	1	MEDITERRANEAN	SEDENTARIAN	2017/04/03 00:00:00.000000000	248	6.95	11.75
1.	1	FAT	SEDENTARIAN	2017/04/10 00:00:00.000000000	243	6.79	11.75
1.	1	MEDITERRANEAN	SEDENTARIAN	2017/04/17 00:00:00.000000000	247	6.91	11.75
1.	1	MEDITERRANEAN	SEDENTARIAN	2017/04/24 00:00:00.000000000	247	7.05	11.75
1.	1	MEDITERRANEAN	SEDENTARIAN	2017/05/01 00:00:00.000000000	241	6.86	11.75
1.	1	FAT	SEDENTARIAN	2017/05/08 00:00:00.000000000	231	6.75	11.75
2.	1	MEDITERRANEAN	SEDENTARIAN	2017/05/15 00:00:00.000000000	234	7.21	11.75
2.	1	MEDITERRANEAN	SEDENTARIAN	2017/05/22 00:00:00.000000000	234	7.1	11.75
2.	1	MEDITERRANEAN	SEDENTARIAN	2017/05/29 00:00:00.000000000	235	6.87	11.75
2.	1	MEDITERRANEAN	SEDENTARIAN	2017/06/05 00:00:00.000000000	233	6.97	11.75
2.	1	MEDITERRANEAN	SEDENTARIAN	2017/06/12 00:00:00.000000000	232	7.09	11.75
2.	1	MEDITERRANEAN	SEDENTARIAN	2017/06/19 00:00:00.000000000	234	7.05	11.75
2.	1	MEDITERRANEAN	SEDENTARIAN	2017/06/26 00:00:00.000000000	233	6.94	11.75
2.	1	MEDITERRANEAN	SEDENTARIAN	2017/07/03 00:00:00.000000000	232	7.16	11.75
2.	1	MEDITERRANEAN	NORMAL	2017/07/10 00:00:00.000000000	232	6.98	11.75
2.	1	MEDITERRANEAN	SEDENTARIAN	2017/07/17 00:00:00.000000000	235	6.82	11.75

Ilustración 61, Resultado de verificación escenario de extracción 4.

6.1.2. Transformación

De acuerdo con lo indicado anteriormente, los procesos de transformación se realizan para limpiar, resumir y/o formatear los datos que finalmente se almacenarán en nuestro data warehouse.

Con el fin de identificar si se requieren o no procesos de transformación en los datos del proyecto, se debe efectuar un proceso de ingeniería inversa donde se busca conocerlos y entenderlos en el contexto del problema planteado. Lo anterior, para determinar la necesidad de calcular nuevos campos, extraer métricas, aplicar reglas de formateo o criterios que se consideren relevantes para dar respuesta de manera asertiva a las preguntas analíticas planteadas en el proyecto.

Al ejecutar este proceso de ingeniería inversa se evidencia que es necesario, por ejemplo, adecuar los valores de referencia para la presión sistólica y diastólica de la tabla **Indicator**, donde contamos con valores que corresponden a una décima del valor real de la medición estándar.

Adicionalmente, se han planteado los siguientes escenarios.

I. Valores de referencia por ajustar

Como se mencionó anteriormente, los valores de referencia para la presión sistólica y presión diastólica de la tabla **Indicator**, corresponden a una décima del valor real en escalas de medición avaladas científicamente [37]. Por lo tanto, se debe multiplicar este valor por 10 y persistirlo en nuestro data warehouse.
La transformación se define así en PDI:

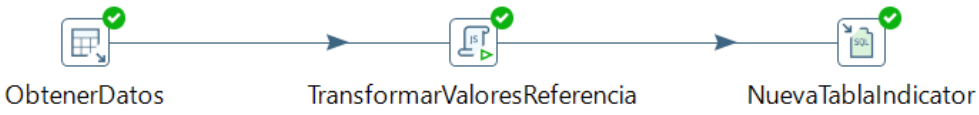


Ilustración 62, Transformación de valores de referencia de la presión

Elemento 1, Obtener datos: este se encarga de establecer la conexión a la base de datos y extraer toda la información de la tabla Indicator.

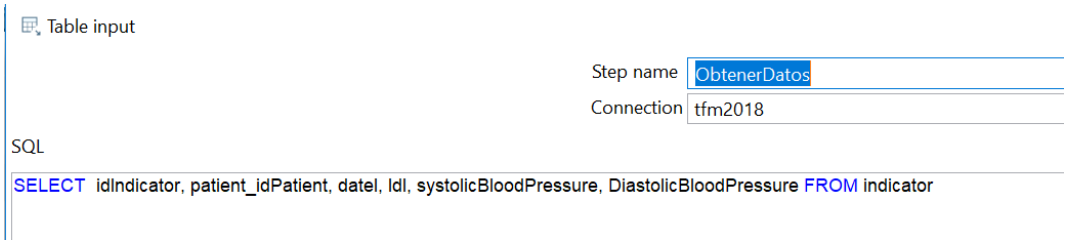


Ilustración 63, Acceso a la información de la tabla Indicator

Elemento 2, Transformar valores de referencia: se encarga de tomar los valores de la presión diastólica y sistólica y multiplicarlos por 10.

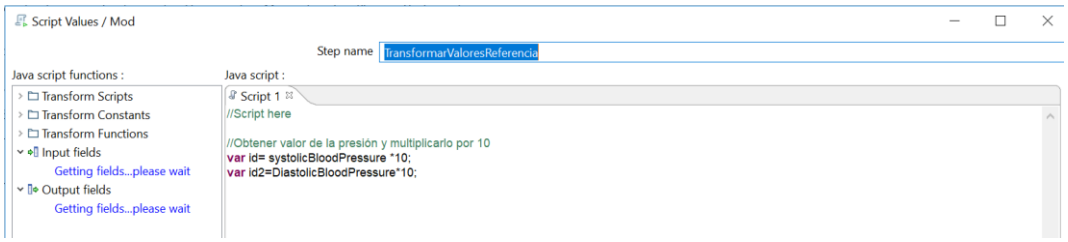


Ilustración 64, Multiplicación de los valores de referencia de la presión

Elemento 3, Nueva Tabla Indicator, en un principio se trató de actualizar la misma tabla Indicator con los valores de referencia modificados, sin embargo, no fue posible. Para solventar esto se genera un archivo SQL con la creación de la nueva tabla y los inserts requeridos para la tener la información de los indicadores de los pacientes.

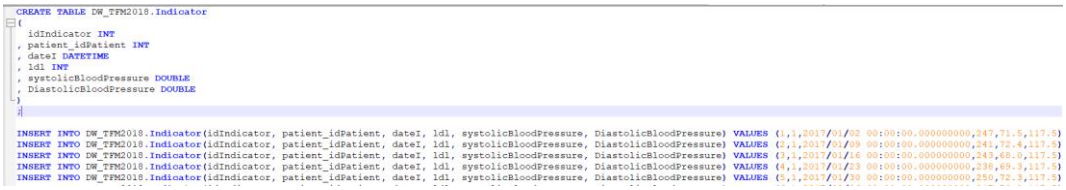


Ilustración 65, Resultado de la transformación

I. Tabla tiempo

Se plantea la creación de una nueva tabla, llamada: Tiempo, la cual contendrá el atributo Date o week, (Fecha del indicador o hábito del paciente), procedente de las tablas Indicator y Habit. Esta permitirá analizar desde una dimensión más granular la información.

Este atributo será dividido en tres partes:

- **Día**: día del indicador y hábito del paciente, valor numérico.
- **Mes**: mes del indicador y hábito del paciente, valor texto.
- **Año**: año del indicador y hábito del paciente, valor numérico.

La transformación se ve de la siguiente manera en PDI:

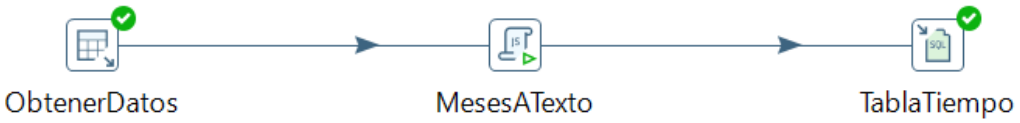


Ilustración 66, Transformación dimensión tiempo

Elemento 1, Obtener datos: se encarga de extraer la información de las tablas Indicator y Habit para la transformación:

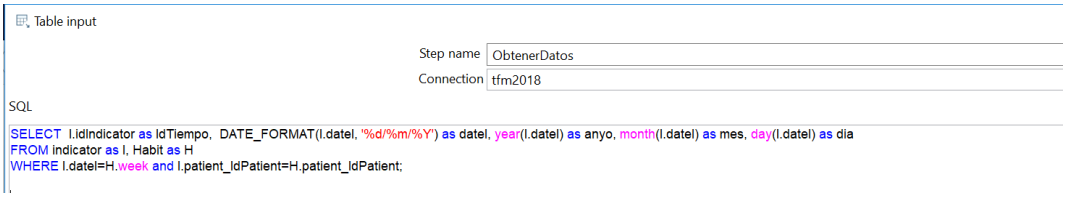


Ilustración 67, Obtención de datos a transformar para la tabla Tiempo

Elemento 2, MesesATexto: es el paso core de la transformación, el cual contempla la modificación de los meses en valores numéricos a valores en texto.

Ilustración 68, Script para transformar meses a texto.

En principio se trató de utilizar la sentencia switch case de Java, sin embargo, no se pudo operar. Para solventarlo se utilizó la sentencia elseif.

Elemento 3, Tabla Tiempo: consiste en la generación de un archivo SQL donde se define la creación de la tabla Tiempo y sus registros provenientes de las tablas Indicator y Habit.

Para la generación del archivo SQL, se define el esquema de la base de datos que se creará, el nombre de la tabla, la ruta donde se alojará el archivo y su extensión:

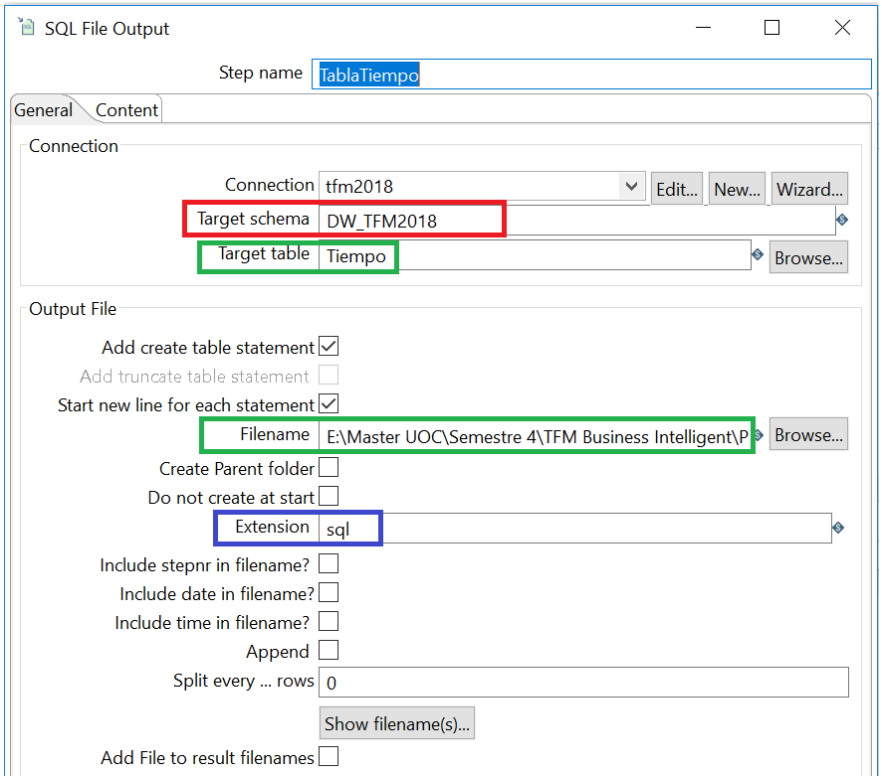


Ilustración 69, Configuración del archivo de salida SQL.

Posterior a ejecutar la transformación es posible visualizar en la ruta seleccionada la creación de la tabla y sus respectivos registros:

```
CREATE TABLE DW_TFM2018.Tiempo
(
  IdTiempo INT
, dateI TINYTEXT
, anyo BIGINT
, mes TINYTEXT
, dia BIGINT
)
;

INSERT INTO DW_TFM2018.Tiempo(IdTiempo, dateI, anyo, mes, dia) VALUES (1,'02/01/2017',2017,'Enero',2);
INSERT INTO DW_TFM2018.Tiempo(IdTiempo, dateI, anyo, mes, dia) VALUES (2,'09/01/2017',2017,'Enero',9);
INSERT INTO DW_TFM2018.Tiempo(IdTiempo, dateI, anyo, mes, dia) VALUES (3,'16/01/2017',2017,'Enero',16);
```

Ilustración 70, Creación básica de la tabla Tiempo.

Cabe aclarar que para hacer de esta una tabla válida en nuestro DW, fue necesario adicionar una clave primaria y verificar de manera previa la relación de la fecha ingresada con la fecha del indicador y hábito del paciente.

II. Tabla Dieta

De acuerdo con las dimensiones expuestas anteriormente en el diseño lógico del DW, se debe crear una tabla Dieta que contenga los valores de:

- La dieta del paciente.
- La fecha en la que llevaba la dieta en particular.

Para hacerlo se crea una transformación con los siguientes elementos:

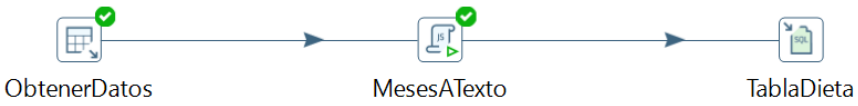


Ilustración 71, Transformación dimensión Dieta

- **Elemento 1, Obtener datos:** este se encarga de extraer desde la base de datos de Staging los datos correspondientes a la dieta del paciente:

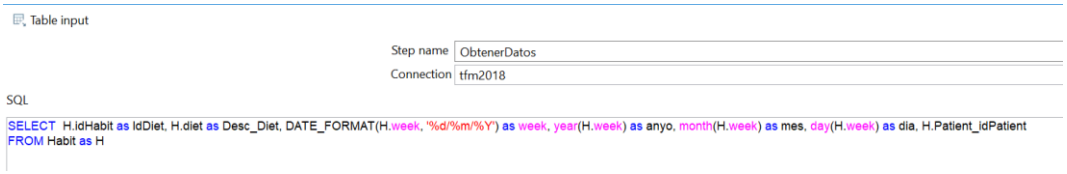


Ilustración 72, Obtención de datos a transformar para la tabla Dieta

- **Elemento 2, Meses a texto:** se transforman los valores de día, mes y año tal como se hizo con la tabla Tiempo.
- **Elemento 3, TablaTiempo:** se crea un script SQL para la tabla Dieta.

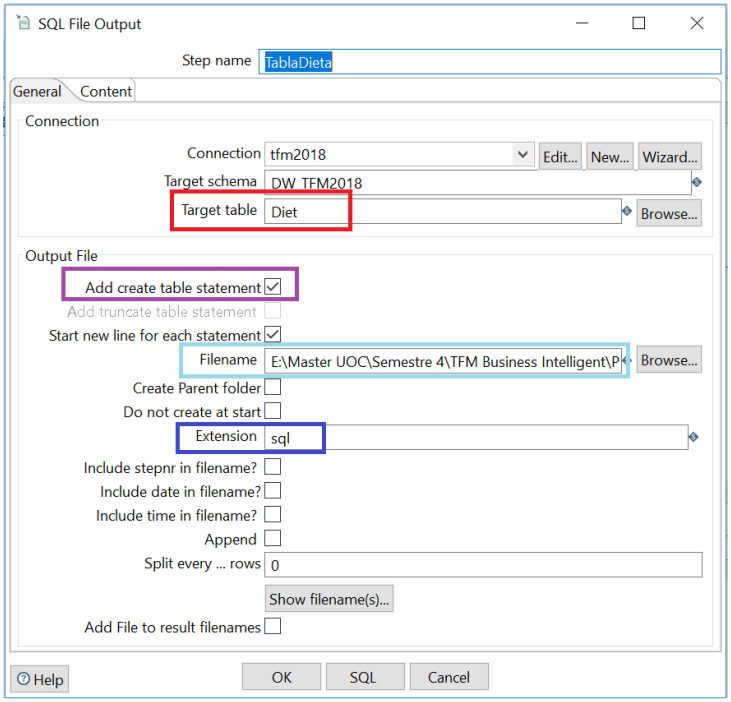


Ilustración 73, Configuración del archivo de salida SQL tabla Dieta.

Como se puede observar a continuación, el archivo de salida tiene la siguiente estructura:

```
CREATE TABLE DW_TFM2018.Diet
(
  IdDiet INT
, week DATETIME
, anyo BIGINT
, mes TINYTEXT
, dia BIGINT
, Patient_idPatient INT
)
;

INSERT INTO DW_TFM2018.Diet(IdDiet, week, anyo, mes, dia, Patient_idPatient) VALUES (1,2017/01/02 00:00:00.000000000,2017,'Enero',2,1);
INSERT INTO DW_TFM2018.Diet(IdDiet, week, anyo, mes, dia, Patient_idPatient) VALUES (2,2017/01/09 00:00:00.000000000,2017,'Enero',9,1);
INSERT INTO DW_TFM2018.Diet(IdDiet, week, anyo, mes, dia, Patient_idPatient) VALUES (3,2017/01/16 00:00:00.000000000,2017,'Enero',16,1);
```

Ilustración 74, Evidencia del archivo de salida SQL

III. Tabla Ciudad

Como se evidenció anteriormente es necesario desagregar de la base de datos origen el manejo de las ciudades o ubicaciones de los pacientes. Para ello se realizar una transformación básica donde se extraen los datos y se llevan a un SQL donde se define la nueva tabla a crear en el DW.

Los elementos de esta transformación son:

Elemento 1: Obtener Datos. Es el encargado de conectarse a la base de datos y extraer los datos de las ciudades directamente desde la tabla de Pacientes:

Ilustración 75, Transformación dimensión tiempo

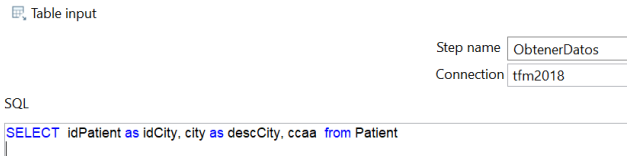


Ilustración 76, Obtención de datos a transformar para la tabla Ciudad

Elemento 2, TablaCityDW: es el elemento encargado de generar el archivo SQL para la creación de la nueva tabla.

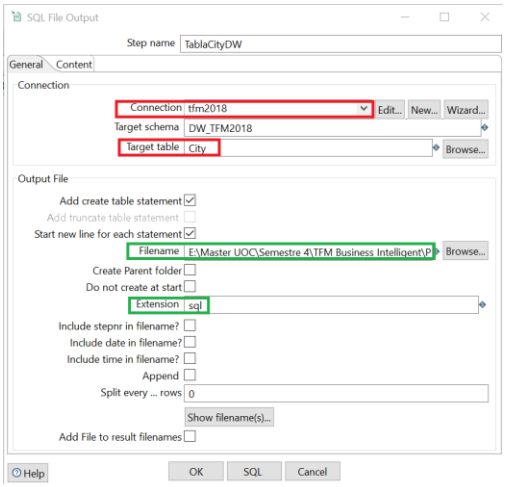


Ilustración 77, Configuración del archivo de salida SQL tabla Ciudad.

6.1.3. Carga

Considerando que la base de datos actual con la que se están desarrollando las transformaciones actúa como una base de datos de ***staging (Lugar de origen y pruebas de la información),*** posterior a ejecutar estos anteriores escenarios se creará una base de datos actualizada, la cual será el data warehouse de la solución.

Nota: previamente, se generó una copia de seguridad de la base de datos a través del uso de MySQL WorkBench. **Ver anexo**: BackupBDStaging.

Para completar el nuevo esquema se utilizará la copia de seguridad de la base de datos de Staging donde se tienen las sentencias requeridas para crear las tablas restantes de nuestro esquema, Patient y Treatment y Habit.

Aunque se exploró cómo realizar la creación de nuevo esquema desde PDI, no se logró encontrar una opción. Sólo se encontraron opciones para ejecutar el archivo que se obtuvo como resultado de la transformación en la consola de la base de datos.

En una consola de MySQL, se crea la respectiva bodega de datos llamada DW_TFM2018, y sus registros, así:

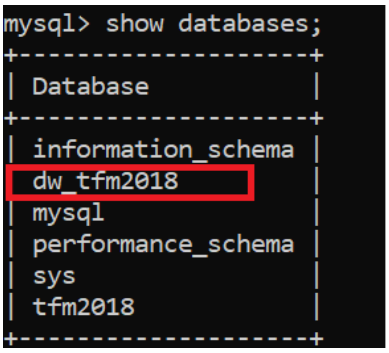


Ilustración 78, Resultado de la creación del esquema del DW

Tal como se mencionó en apartados anteriores, las tablas Tiempo y Dieta e Indicator fueron transformadas, es momento de reconstruir con esta información el DW.

Considerando que, en el apartado de transformaciones, la tabla Indicator mantiene actualmente una información básica para la integración con las dimensiones del modelo y con el fin de crear las respectivas relaciones con las dimensiones tiempo, paciente, tratamientos y hábitos, es necesario crear unas claves foráneas en dicha tabla. Para automatizar esta tarea se ha utilizado PDI, con una transformación básica como la expuesta a continuación:

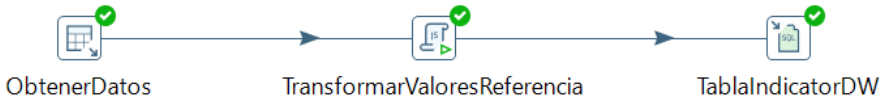


Ilustración 79, Nueva transformación para complementar tabla hechos

En esta segunda versión de la transformación ***Valores de referencia por ,*** se conservan los mismos elementos, sin embargo, el Elemento 1, Obtener Datos, complementará la información que requiere la tabla Indicator para relacionarse con las dimensiones del esquema lo modelo en estrella.

Ilustración 80, Modificación elemento 1 transformación tabla de hechos

Cómo se puede observar en la anterior ilustración, se adicionan las relaciones con las dimensiones tiempo, tratamiento, dieta, actividad o hábito y paciente.

Al ejecutar la transformación, se genera como resultado un completo archivo SQL con la creación de la tabla Indicator y sus registros:

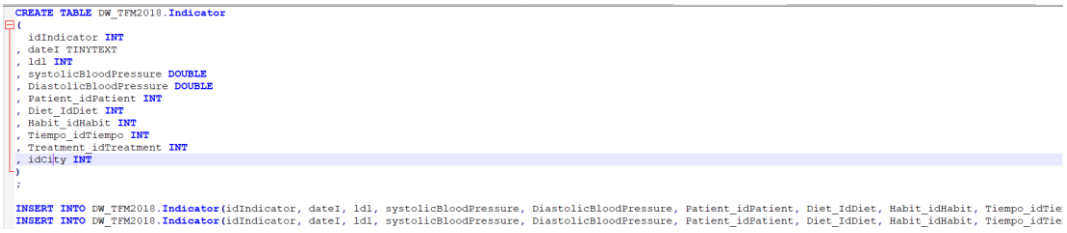


Ilustración 81, Salida de la transformación de la tabla de hechos.

Al ejecutar esto y los demás scripts de creación de la base de datos en una consola de MySQL se tiene:

```
mysql> show databases;
+--------------------+
| Database           |
+--------------------+
| information_schema |
| dw_tfm2018         |
| mysql              |
| performance_schema |
| sys                |
| tfm2018            |
+--------------------+
6 rows in set (0.00 sec)
```

Ilustración 82, Creación del esquema del DW

```
+----------------------+
| Tables_in_dw_tfm2018 |
+----------------------+
| city                 |
| diet                 |
| habit                |
| indicator            |
| patient              |
| time                 |
| treatment            |
+----------------------+
```

Ilustración 83, Listado de tablas del DW.

```
mysql> select count(*) from indicator;
+----------+
| count(*) |
+----------+
|      312 |
+----------+
1 row in set (0.00 sec)
```

Ilustración 84, Resultado de ejecución MySQL DW, número de registros insertados

Nota: en los anexos Data Warehouse y Transformaciones se encuentran los archivos utilizados en las transformaciones y cargas de la información al DW.

7. OLAP, On-Line Analytical Processing.

Es una técnica de análisis de multidimensional, utilizada para lograr un buen rendimiento de consultas ad-hoc. Permite tener una vista multidimensional de los

datos y es un punto intermedio entre el data warehouse y los reportes que finalmente se le muestran al usuario final.

Además, desde la usabilidad esta técnica permite esconder la complejidad de los datos y los análisis realizados sobre la información, a su vez que facilita el análisis, la visualización, el acceso y la generación de reportes. [38]

7.1. Cubos OLAP.

Un cubo OLAP es una estructura de datos que permite un análisis rápido de los datos de acuerdo con las múltiples dimensiones que definen un problema empresarial. [39] Uno de los casos típicos para explicar qué se entiende por dimensión en un cubo OLAP, es el caso de las ventas de productos, un cubo multidimensional para informar sobre ventas podría estar compuesto, por ejemplo, de 7 dimensiones: vendedor, monto de ventas, región, producto, región, mes, año. En nuestro caso podemos hablar de pacientes, tratamientos, tiempo y hábitos o actividades de los pacientes.

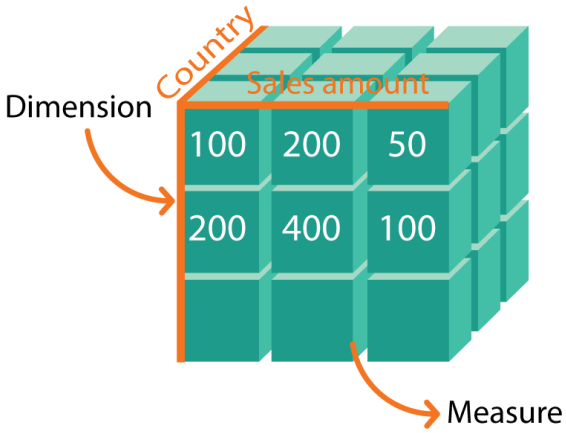


Ilustración 85, Ejemplo de cubo OLAP para ventas

7.1.1. Ventajas de los cubos OLAP.

I. Se logra tener acceso a grandes cantidades de datos de manera más ágil y flexible.
II. Analizan las relaciones entre muchos tipos de elementos empresariales.
III. Presentan los datos en diferentes perspectivas.
IV. Fácil uso para el usuario final.
V. Su uso genera ventajas competitivas como aprender más sobre el negocio, entender cómo generar nuevas estrategias de mercadeo o análisis para mejorar la productividad. [40]

7.1.2. Componentes de un cubo OLAP.

De manera general, un cubo OLAP puede componerse de los siguientes elementos: [41]

I. **Dimensiones:** los cubos contienen todas las dimensiones en las que los usuarios basan sus análisis de los datos de hechos. Una instancia de una dimensión de base de datos en un cubo se denomina dimensión de cubo y se relaciona con

uno o más grupos de medida en el cubo. Una dimensión de base de datos se puede utilizar varias veces en un cubo. [42]

II. **Medidas y grupos de medidas:** una medida es una agregación de valores de datos numéricos, como una suma, un recuento, un mínimo, un máximo o un promedio. Un *grupo de medida* es un contenedor para una o más medidas. Todas las medidas existen en un grupo de medida, incluso si solo hay una medida. Un cubo debe tener al menos una medida y un grupo de medida. [43]

III. **Particiones:** una partición proporciona el almacenamiento físico de los datos de hechos cargados en un grupo de medida. Se crea automáticamente una sola partición para cada grupo de medida, aunque es frecuente crear particiones adicionales que segmenten aún más los datos, lo que produce un procesamiento más eficiente y un rendimiento de las consultas más rápido. [44] En este caso no se tendrá en cuenta este elemento para el desarrollo de los cubos OLAP.

IV. **Perspectivas:** son un subconjunto de un cubo creado para una aplicación o grupo de usuarios específico. El cubo es la perspectiva predeterminada. Una perspectiva se muestra al cliente como un cubo. Cuando el usuario ve una perspectiva, se muestra como otro cubo. [45]

V. **Jerarquías:** una jerarquía es una colección de niveles basados en atributos. Por ejemplo, una jerarquía de tiempo puede contener los niveles año, trimestre, mes, semana y día. En algunas jerarquías, cada atributo de miembro implica únicamente al atributo de miembro que tiene por encima. Esto se conoce a veces como una jerarquía natural. Es común que los usuarios finales pueden utilizar una jerarquía para examinar los datos del cubo. [46]

VI. **Acciones:** una acción es una operación iniciada por el usuario final en un cubo seleccionado o en una parte de un cubo. La operación puede iniciar una aplicación con el elemento seleccionado como parámetro o recuperar información acerca del elemento seleccionado. [47]

7.1.3. Operaciones sobre un cubo OLAP.

Las siguientes son las operaciones que pueden realizarse sobre los cubos OLAP. [6]

I. **Roll up (Drill up):** capacidad de resumir los datos. Puede significar subir en la jerarquía lo reducir las dimensiones.

II. **Drill down (Roll down):** capacidad de aumentar el detalle de los datos. Puede significar bajar en la jerarquía o reducir las dimensiones.

III. **Slice and dice:** permite seleccionar y proyectar datos.

IV. **Pivot (Rotar):** se refiere a reorientar el cubo.

V. **Drill:** se utilizan las coordenadas dimensionales especificadas por un usuario para una celda en un cubo para moverse a otro cubo a ver información relacionada.

a. **Drill across:** implica utilizar más de una tabla de hechos.
b. **Drill through:** Ir desde el nivel de máximo detalle del cubo a sus tablas relacionales (utilizando SQL).

7.2. Diseño de cubo OLAP.

Para este proyecto solo se diseñará un cubo OLAP entendiendo que todas las dimensiones están al mismo nivel y que las medidas son comunes a todas estas.

De acuerdo con el análisis dimensional realizado en anteriores apartados, el cubo OLAP contendrá las siguientes dimensiones:

- Paciente.
- Tratamiento.
- Tiempo.
- Hábito: actividad del paciente.
- Dieta.
- Ciudad: ubicación geográfica del paciente.

Por su parte las medidas correspondientes para valorar los indicadores de salud del paciente son:

- Nivel de colesterol LDL.
- Presión sistólica.
- Presión diastólica.

Las jerarquías que se utilizan son principalmente para la dimensión tiempo, donde se tienen los días y se los meses como un nivel más alto para visualizar los datos.

Al iniciar la herramienta Pentaho Schema Workbench o PSW, y posterior a configurar la conexión al DW, se crea un nuevo esquema.

En el esquema lo primero que se crean son las dimensiones:

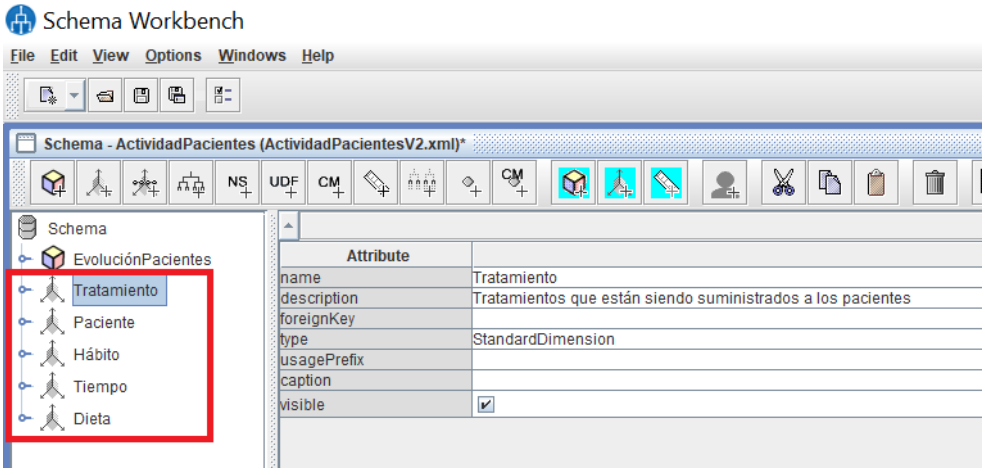


Ilustración 86, Definición de las dimensiones del cubo

Cada dimensión contiene como mínimo los siguientes atributos:

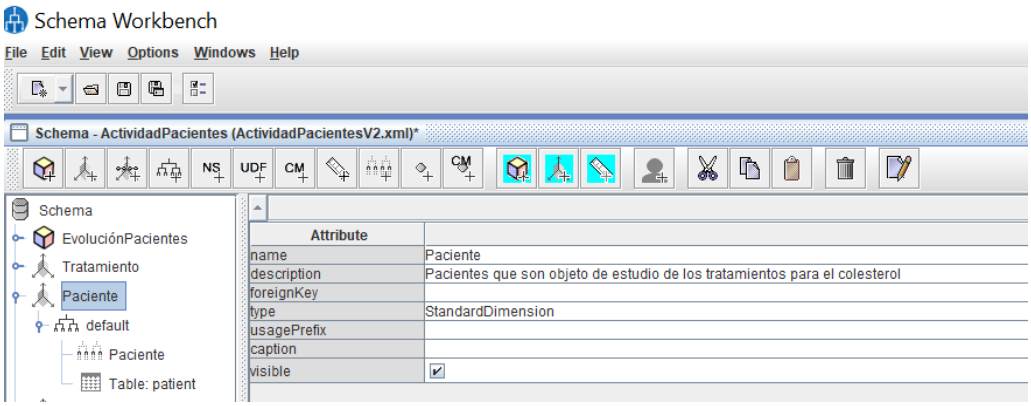


Ilustración 87, Atributos de una dimensión en PSW.

- **Name:** nombre de la dimensión
- **Description:** descripción de lo que significa la dimensión.
- **Foreing Key:** clave foránea de la tabla que representa la dimensión y la relación con la tabla de hechos, (Indicator).
- **Type:** se refiere al tipo de dimensión, existen dos tipos en PSW, la dimensión estándar y de tiempo. En este caso es Estándar.
- **UsagePrefix:** prefijo de uso de la dimensión.
- **Caption:** mensaje sobre la dimensión que aparece al usar el cubo.
- **Visible:** se refiere a si la dimensión es visible para los cubos que se vayan a crear en adelante sobre el mismo esquema.

Esta información se diligencia para todas las dimensiones. Así mismo, cada dimensión está compuesta por:

- Una jerarquía creada por defecto para la dimensión.
- Un nivel que contiene la descripción de las columnas que se utilizarían en el análisis del cubo y que además son la relación con la tabla de hechos. De igual manera identifica el tipo de datos de la dimensión, en este caso enteros y el tipo de nivel. (Para el caso regular), tal como se ilustra a continuación:

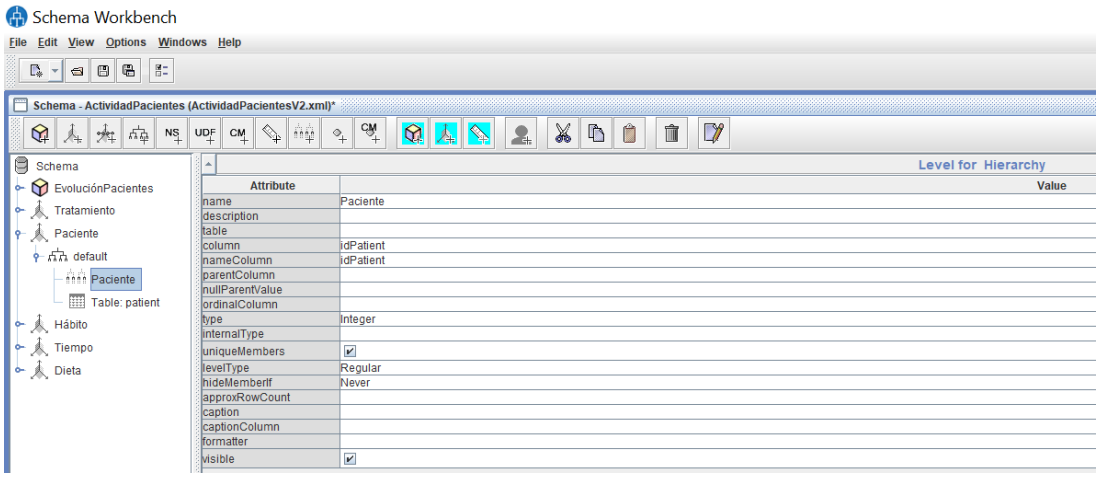


Ilustración 88, Datos básicos de un nivel en una dimensión en PSW.

- Una tabla en el Data Warehouse que contiene los datos básicos para la dimensión, en este caso es la tabla patient.

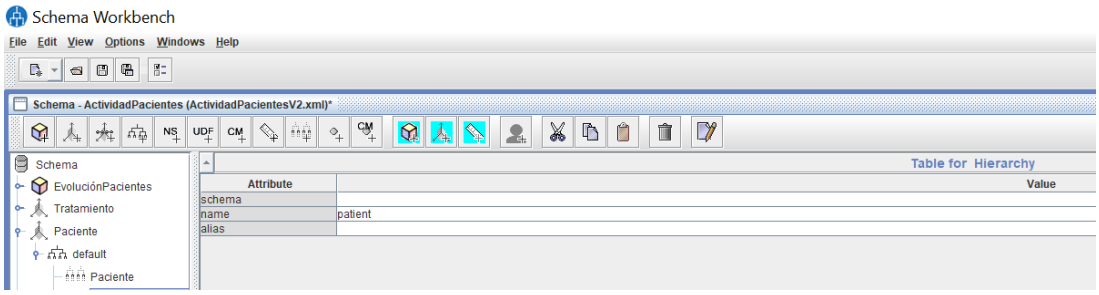


Ilustración 89, Tabla de la dimensión Paciente.

Tanto la definición de atributos como la definición de los componentes de cada dimensión se hace de manera idéntica para cada una de las dimensiones listadas.

Las dimensiones inicialmente se crean por fuera del cubo OLAP con el fin de que puedan ser usadas de manera trasversal o compartida en caso de que se deseara crear más cubos dentro del mismo esquema. Esto facilita el mantenimiento y orden de la información.

Posteriormente, se crea el cubo OLAP, a través de la opción Add Cube:

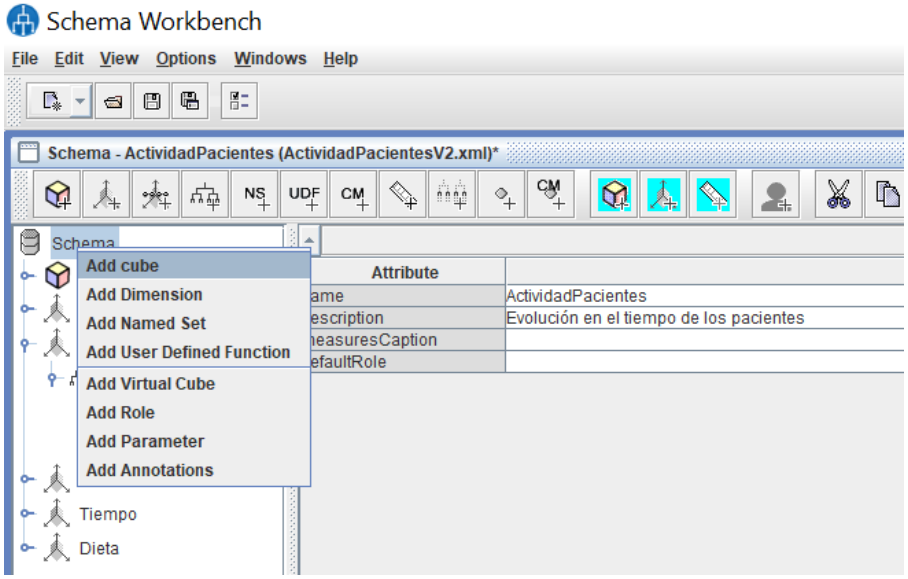


Ilustración 90, Creación de cubo OLAP en PSW.

Un cubo OLAP en PSW, tiene los siguientes atributos:

- **Name:** nombre que se le da al cubo OLAP.
- **Description:** texto que indica qué hace o a qué se refiere el cubo OLAP.
- **MeasuresCaption:** es el mensaje que aparece en las medidas del cubo.
- **DefaultRole:** indica el rol del cubo. (Este campo no es utilizado).

De manera general el cubo está compuesto por:

- La tabla de Hechos, en este caso Indicator.

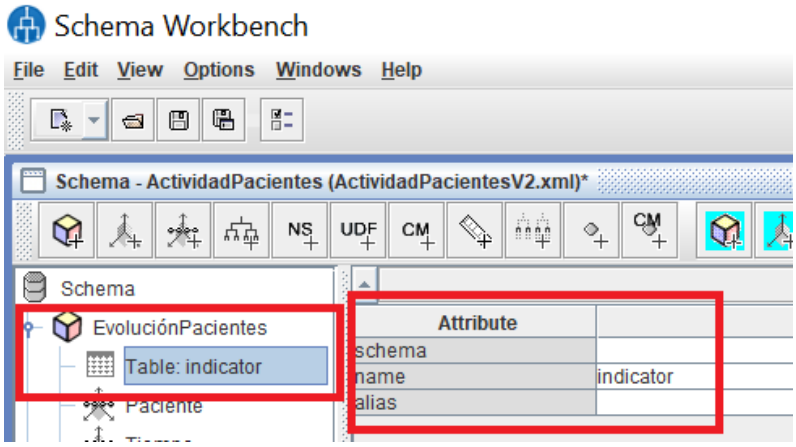


Ilustración 91, Tabla de hechos en el cubo OLAP.

Las dimensiones que se utilizarían en el cubo. Para el caso se hace uso de la opción add dimension usage, la cual permite reutilizar las dimensiones que de manera trasversal se crearon en el esquema:

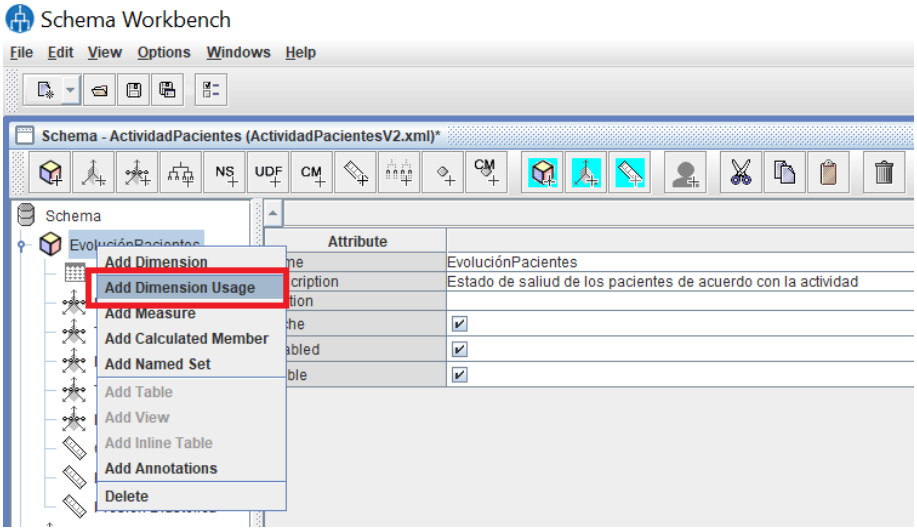


Ilustración 92, Agregar dimensiones de uso al cubo OLAP.

Al hacer esto las dimensiones son agregadas y se deben complementar los atributos básicos de la dimensión:

- **Name:** nombre de la dimensión.
- **Foregin Key:** clave foránea que identifica la relación con la tabla de hechos Indicator.
- **Source:** fuente de la dimensión que se va a reutilizar.

Los demás campos de la dimensión de uso no son utilizados.

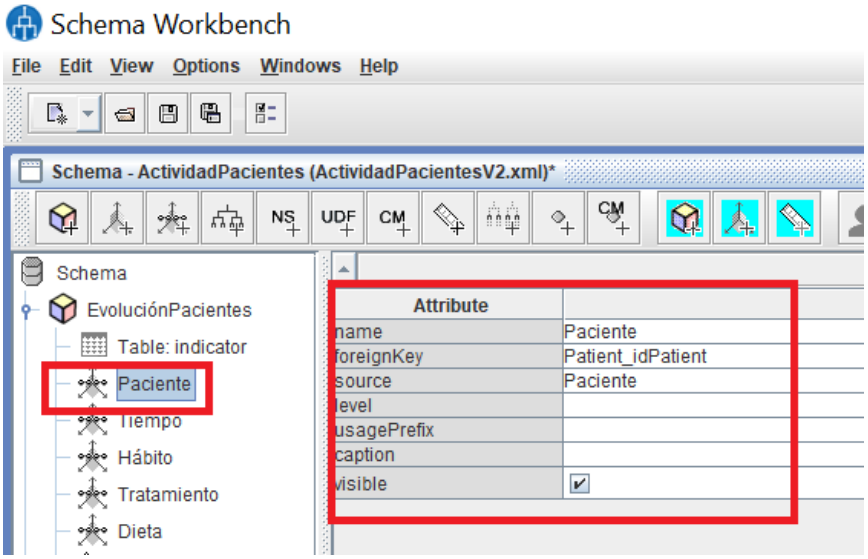


Ilustración 93, Datos básicos de una dimensión de uso

Al diligenciar esta información para todas las dimensiones del cubo, se procede entonces a crear las medidas bajo la opción add measure.

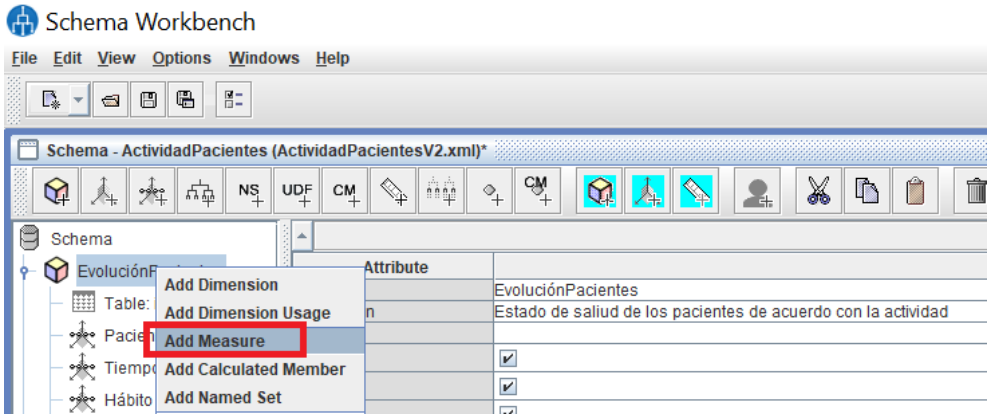


Ilustración 94, Datos básicos de una medida en PSW.

Los atributos básicos de una medida son:

- **Name:** nombre de la medida.
- **Description:** descripción de la medida.
- **Aggregator:** tipo de operación matemática que se realiza con los datos que son objeto de la medida, por ejemplo: avg, sum, count y otros.
- **Column:** columna en la tabla de hechos que identifica la medida.
- **DataType:** tipo de dato de la medida para este caso entero.

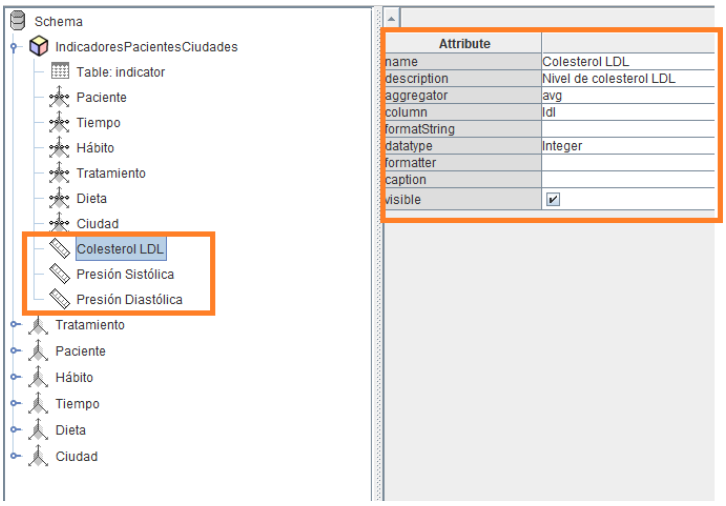


Ilustración 95, Datos básicos de una medida.

Posterior a diligenciar los datos básicos de cada una de las medidas: colesterol LDL, Presión sistólica y presión diastólica, se procede a guardar el cubo OLAP.

Para realizar la integración del PSW con el servidor Mondrian de PENTAHO, lo que se hace es seleccionar la opción de Publish o publicar:

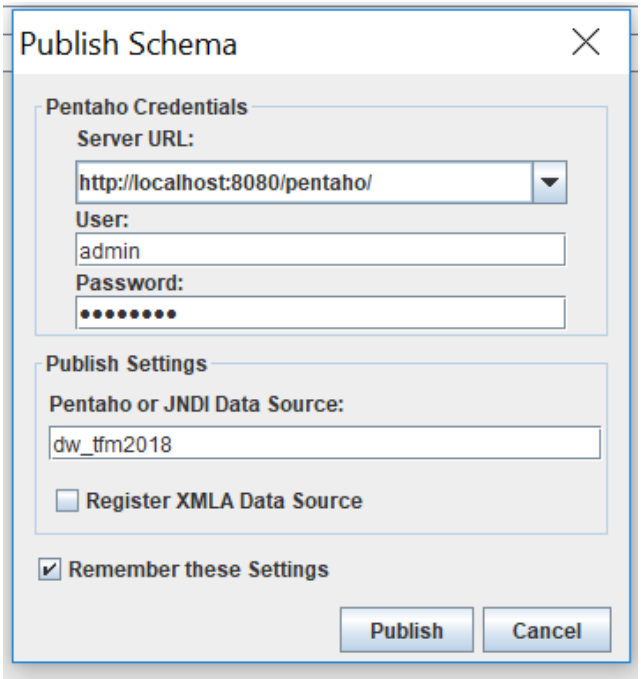


Ilustración 96, Datos requeridos para la publicación del esquema del Cubo OLAP en Mondrian.

Los datos requeridos para publicar el cubo OLAP son:

- **URL del servidor:** se refiere a la URL del servidor del donde está publicado PENTAHO. En este caso está sobre el locahost de la máquina.
- **User:** usuario requerido para iniciar sesión en Pentaho y sobre el servidor Mondrian.
- **Password:** contraseña requerida para iniciar sesión en Pentaho sobre el servidor Mondrian.

- **Pentaho o JNDI Data Source:** se refiere a la conexión al Data Warehouse. Para que funcione correctamente en Pentaho, se debe colocar el mismo nombre que tiene esta última herramienta para conectarse al Data Warehouse.

Si todo funciona correctamente entonces se tendrá el Cubo OLAP publicado correctamente en Pentaho.

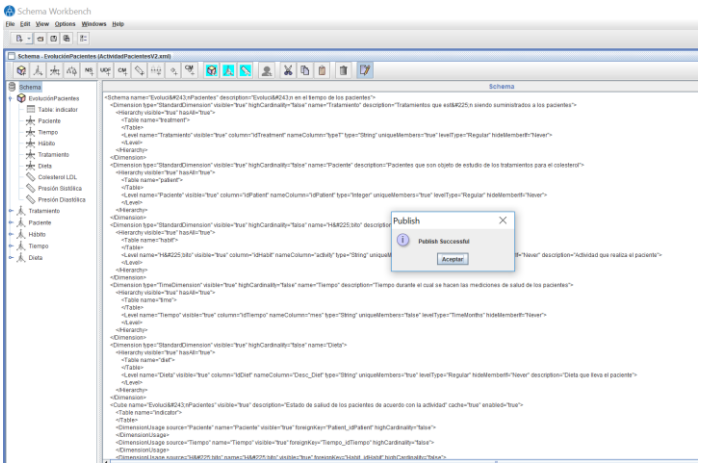


Ilustración 97, Publicación del cubo OLAP en PSW.

El paso siguiente es determinar cómo navegar sobre el cubo OLAP. Para esto JPivot es la herramienta que está integrada a Pentaho con fines a suplir dicha necesidad.

Para hacer uso de este componente en Pentaho, se utiliza la opción File, New, JPivot View.

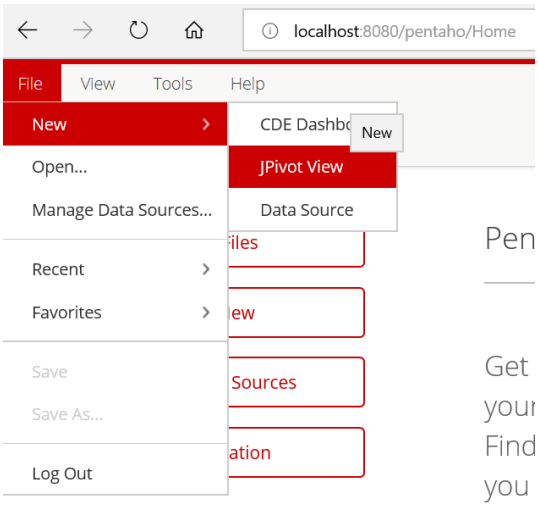


Ilustración 98, Opción de vista en JPivot, Pentaho.

Para completar la vista es necesario indicar el nombre del cubo y del esquema a utilizar:

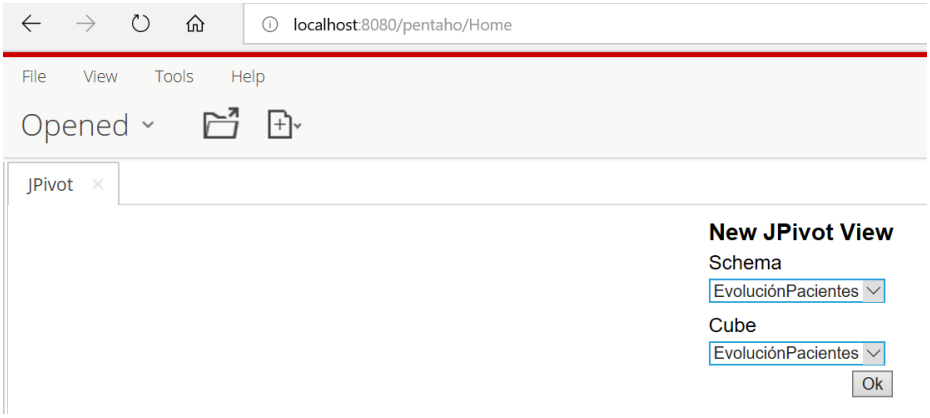


Ilustración 99, Creación de la vista en JPivot.

Si todo funciona correctamente se logrará visualizar lo siguiente:

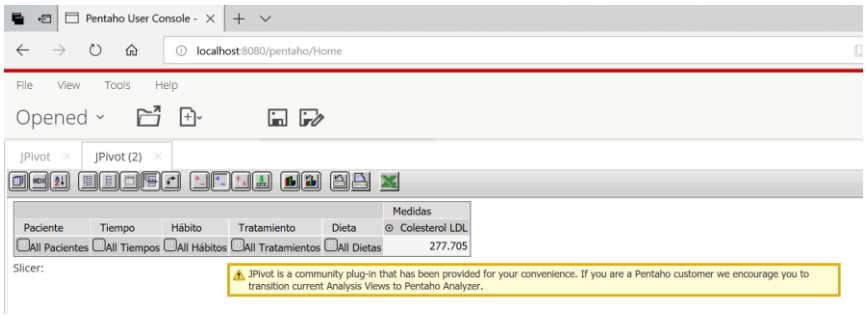


Ilustración 100, Vista del cubo OLAP en JPivot View.

Como se puede observar cada una de las dimensiones del cubo puede ser utilizada como en una tabla dinámica para filtrar y elegir los campos que sean requeridos para análisis. Adicionalmente, JPivot posee una herramienta la cual consiste en un editor MDX o de Lenguaje Mondrian que permite diseñar los queries sobre el cubo OLAP, tal como se ilustra a continuación:

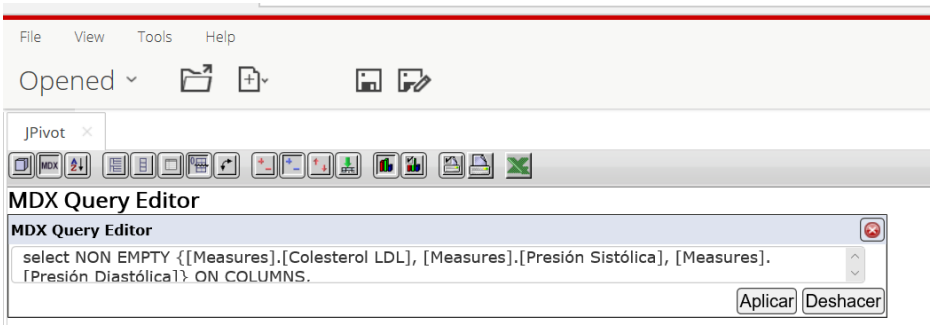


Ilustración 101, Mondrian Editor de Queries

Así mismo, es posible extraer gráficas que permitan interpretar mejor la información, por ejemplo:

A continuación, se ilustra cómo es posible graficar el nivel de colesterol de los pacientes por tipo de tratamiento.

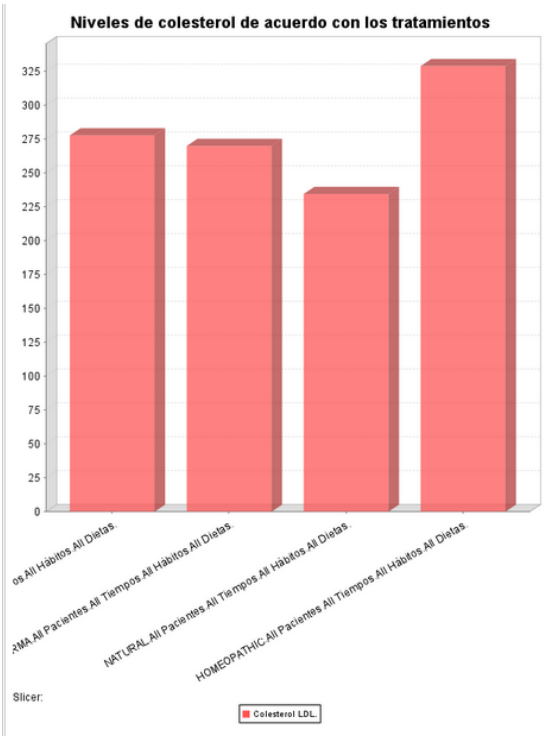


Ilustración 102, Ejemplo de gráfica obtenido con JPivot

Al evidenciar la poca usabilidad de la herramienta para un usuario final y lo difícil que es replicar un query o una consulta sobre el Cubo OLAP se analizan otras herramientas. En este punto aparece **SAIKU** como una de las más intuitivas y útiles para trabajar sobre dichos cubos.

8. Reporte de información

En este apartado se hará referencia al uso de la herramienta SAIKU como la habilitadora de la capacidad de reporte y navegación de cubos OLAP.

Al abrir el esquema del cubo OLAP creado, se realizan los siguientes análisis en términos de la disminución o el aumento de los siguientes indicadores de salud:

- Colesterol LDL
- Presión sistólica
- Presión diastólica

Nota: en adelante, cuando se mencione la palabra hábito se está haciendo referencia a la actividad del paciente.

Se toman los datos de los pacientes en el tiempo, dejando estos en las filas y en las columnas registrando los indicadores de salud o medidas del cubo OLAP. La vista de este análisis es:

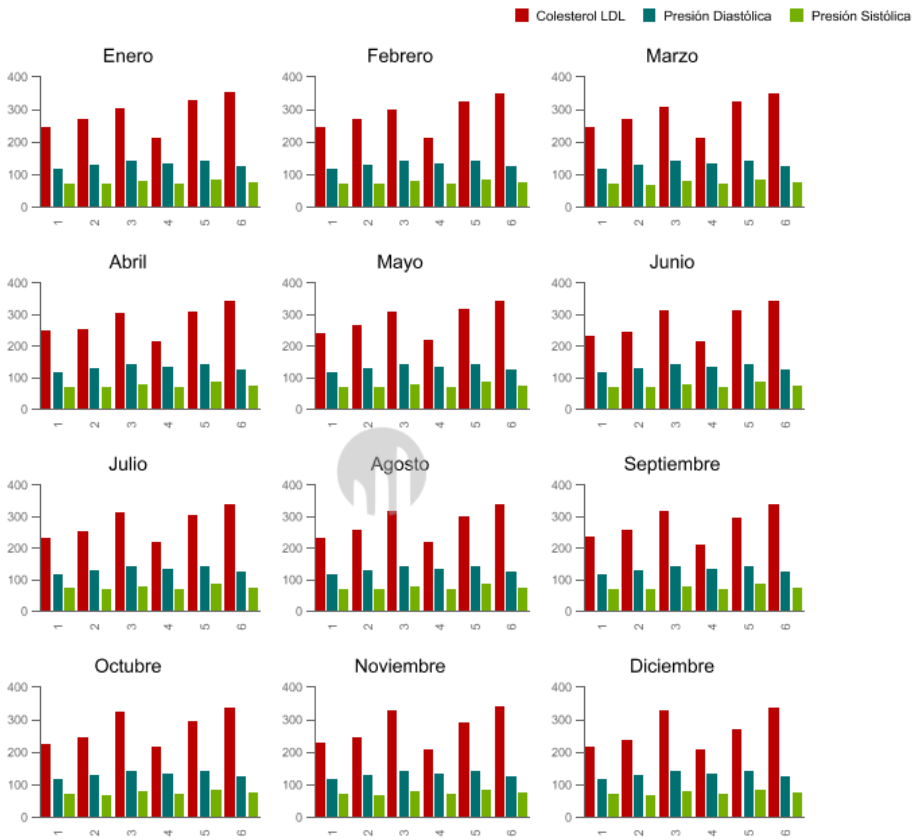


Ilustración 103, Evolución general de los pacientes en el tiempo

Como se evidencia anteriormente, en cada mes los pacientes tienen altas y bajas, si se observa este detalle por paciente se tiene:

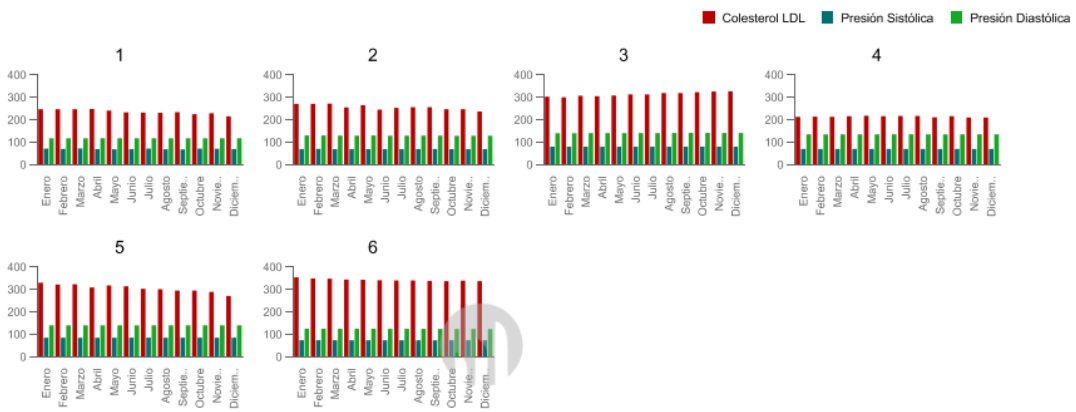


Ilustración 104, Evolución de cada paciente en el tiempo

Es importante destacar que la anterior gráfica visualmente es amigable porque son pocos pacientes, sin embargo, al tener miles de datos la mejor de manera de evidenciar la evolución sería acumulada:

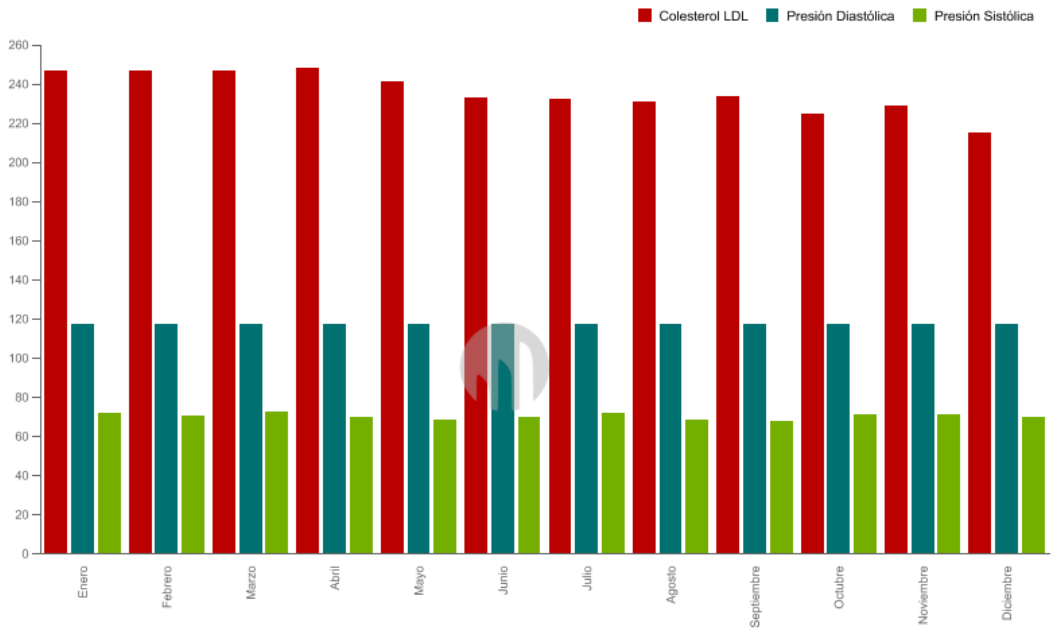


Ilustración 105, Evolución acumulada de todos los pacientes en el tiempo

Los indicadores de salud de los pacientes por tratamiento se visualizan así:

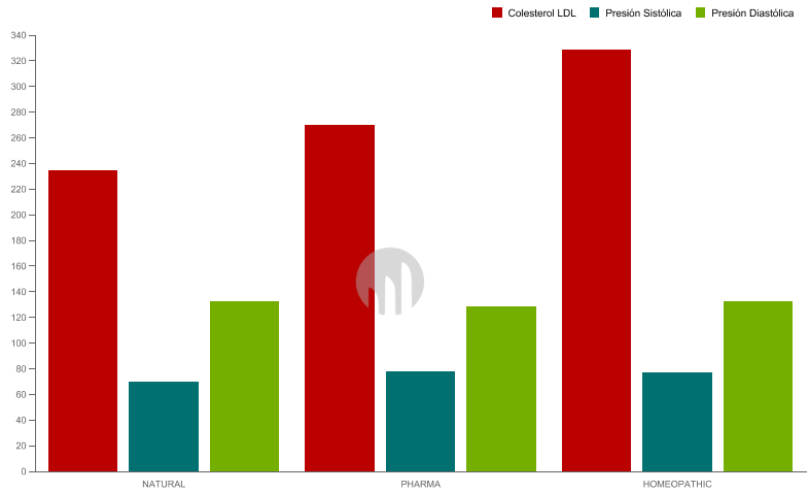


Ilustración 106, Indicadores de salud de los pacientes por tratamientos

Los datos de este análisis son:

Tratamiento	Colesterol LDL	Presión Sistólica	Presión Diastólica
NATURAL	234,481	69,69	132,121
PHARMA	269,865	77,656	128,752
HOMEOPATHIC	328,769	76,849	132,535
Grand Total	**278**	**75**	**131**

Ilustración 107, Datos de Indicadores de salud de los pacientes por tratamientos

En la siguiente ilustración se evidencian las **estadísticas** relacionadas con los indicadores de salud de todos los tratamientos.

Estadisticas	Colesterol LDL / PHARMA	Colesterol LDL / NATURAL	Colesterol LDL / HOMEOPATHIC
Mínimo	215.000	210.000	300.000
Máximo	330.000	274.000	354.000
Suma	28066.000	24386.000	34192.000
Promedio	269.865	234.481	328.769
Desviación Estandard	37.823	21.738	14.639

Ilustración 108, Análisis estadístico de indicadores de salud de los pacientes por tratamientos

Para visualizar cómo los hábitos de los pacientes influyen en sus indicadores de salud bajo un tratamiento específico se tiene:

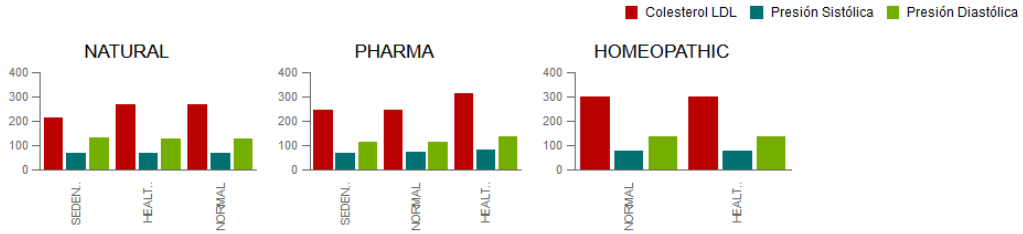


Ilustración 109, Relación tratamiento vs hábito de los pacientes

Si se analiza el comportamiento de los indicadores de los pacientes de acuerdo con el tratamiento y la dieta se tiene:

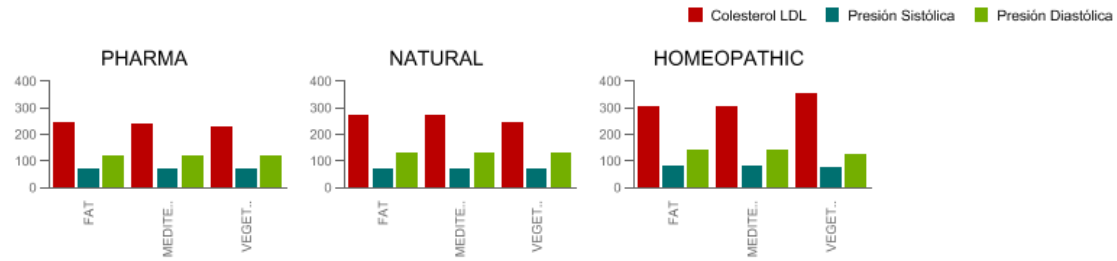


Ilustración 110, Relación tratamiento vs dieta de los pacientes

Sin determinar si el tratamiento tiene un efecto o no sobre los indicadores, la relación entre las dietas y los hábitos es:

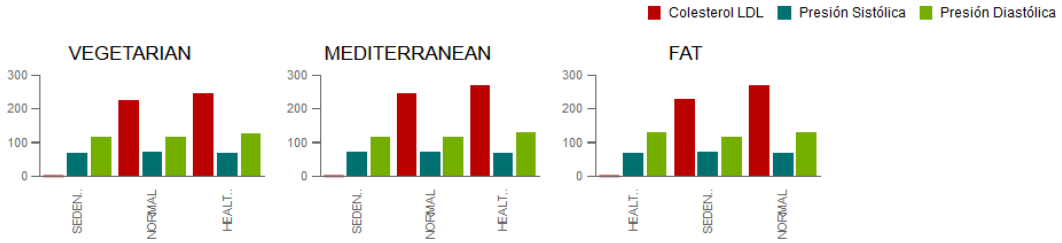


Ilustración 111, Relación entre las dietas y los hábitos de los pacientes

Cuando un paciente se ve sometido a un mismo tratamiento en el tiempo se tienen los siguientes indicadores:

I. Tratamiento homeopático

Un paciente bajo este tratamiento presenta la siguiente evolución en el tiempo:

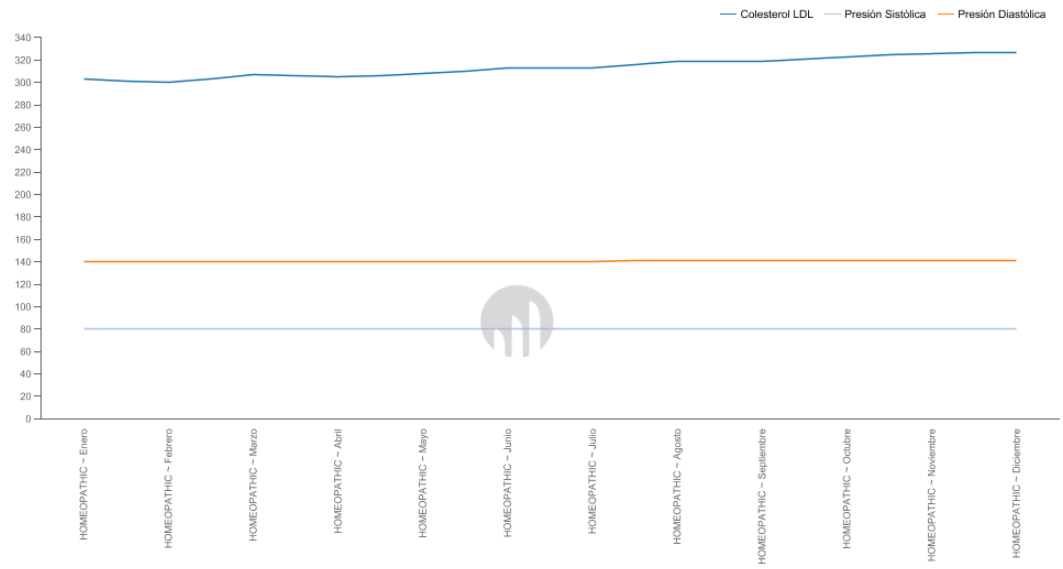


Ilustración 112, Evolución de los pacientes bajo el tratamiento homeopático

Al seguir diferentes **hábitos** en el tiempo estando sometidos a este tratamiento se puede evidenciar:

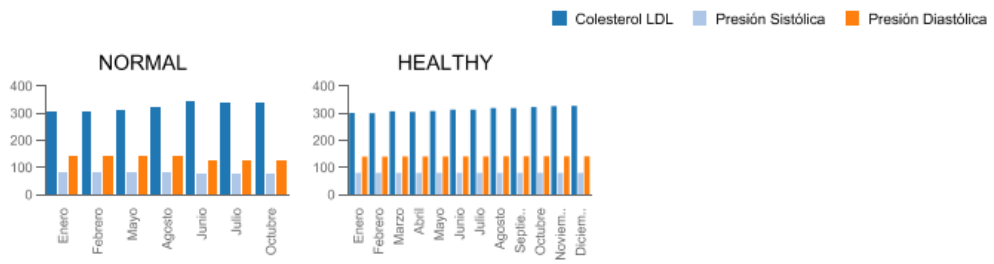


Ilustración 113, Evolución de los pacientes con sus y tratamiento homeopático

En este caso los pacientes con tratamientos homeopáticos tendieron a seguir siempre hábitos normales y saludables.

Agrupando esto por meses se puede visualizar así:

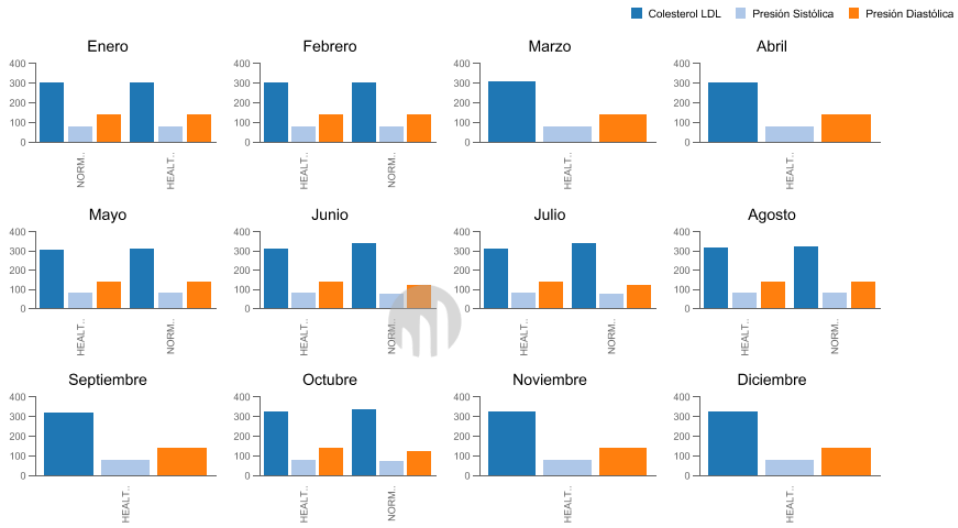


Ilustración 114, Evolución de los pacientes y sus hábitos por meses con el tratamiento homeopático

Según la dieta se tiene:

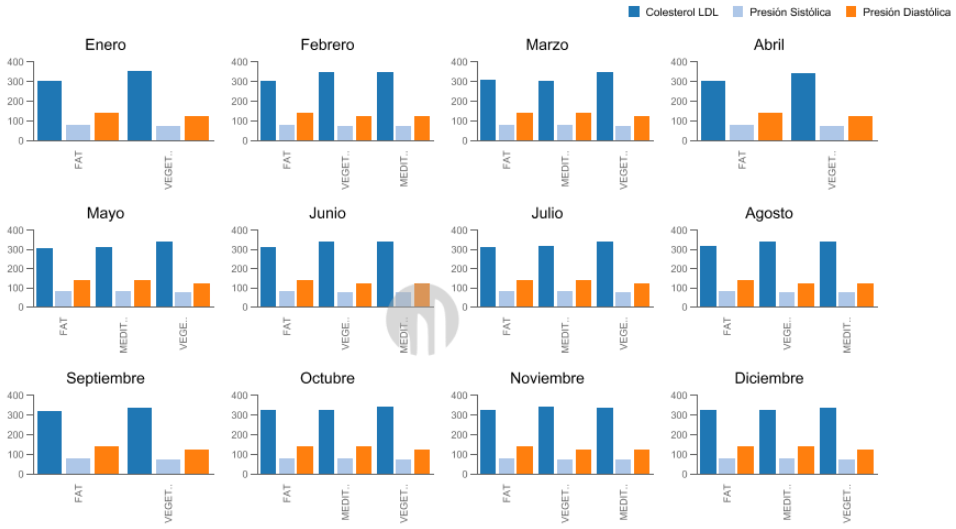


Ilustración 115, Evolución de los pacientes y sus dietas, por meses con el tratamiento homeopático

Visualizando esta información por dieta a cambio de meses se tiene:

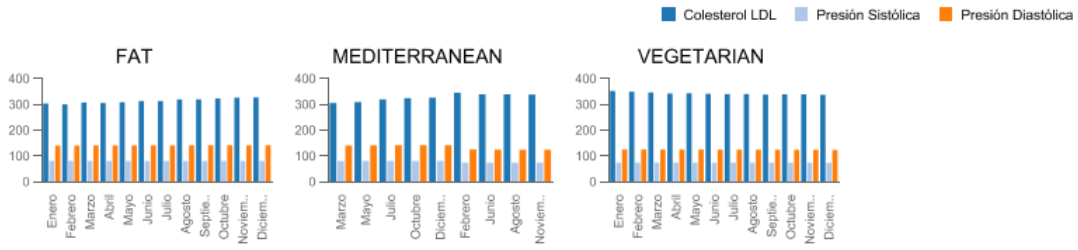


Ilustración 116, Evolución de los pacientes por meses con el tratamiento homeopático segregado por dietas

Existe una desventaja con el uso de estas gráficas y es que expone la información por pacientes. Por ejemplo, en el caso de la dieta mediterránea los pacientes cuyo tratamiento es farmacológico no siguieron esta dieta en los meses de enero, abril y septiembre. Por lo tanto, se debe saber interpretar asertivamente la información para no generar conclusiones erróneas.

Las estadísticas agrupadas de esta información en el tiempo son:

Estadisticas	Colesterol LDL	Presión Sistólica	Presión Diastólica
Mínimo	300.000	73.500	123.300
Máximo	354.000	80.100	141.500
Suma	34192.000	7992.300	13783.600
Promedio	328.769	76.849	132.535
Desviación Estandard	14.639	3.155	8.249

Ilustración 117, Indicadores resumidos para pacientes con tratamiento homeopático

Adicionalmente, se puede ver cuál es la dieta de mayor predominancia en un paciente:

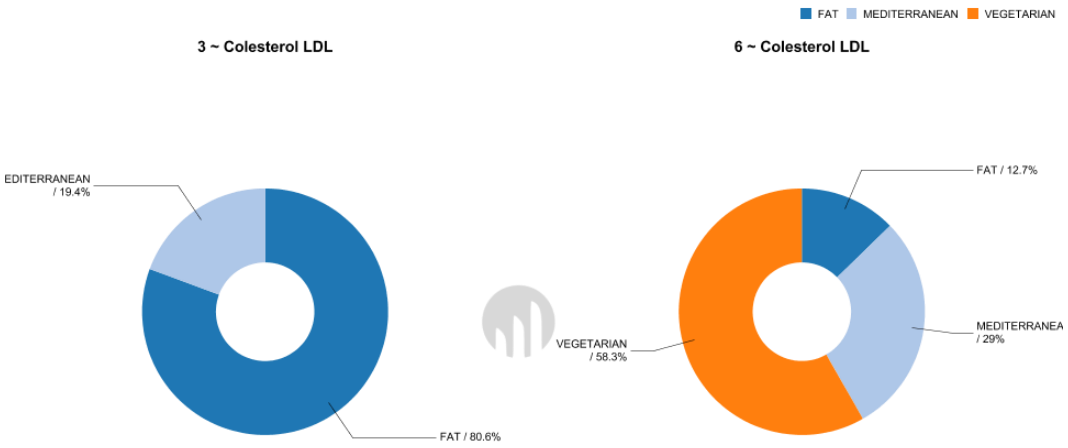
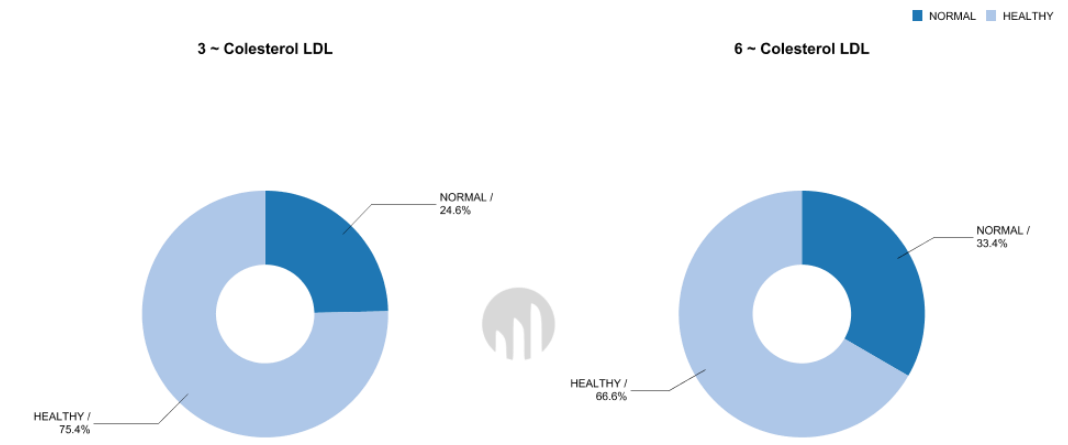


Ilustración 118, Dieta de mayor predominancia en pacientes con tratamiento homeopático

Así mismo se tiene dicha predominancia con relación al hábito del paciente:



II. Tratamiento natural

De manera general la evolución de un paciente en el tiempo con este tratamiento es:

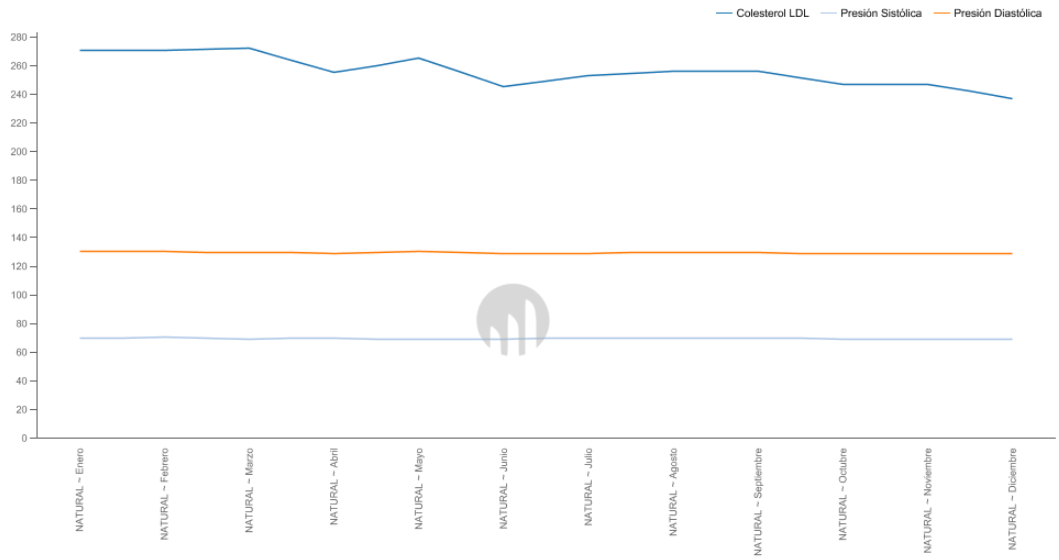


Ilustración 119, Evolución en el tiempo del tratamiento natural en pacientes

La relación que tiene el estar sometido a un tratamiento natural en el tiempo con diferentes hábitos es:

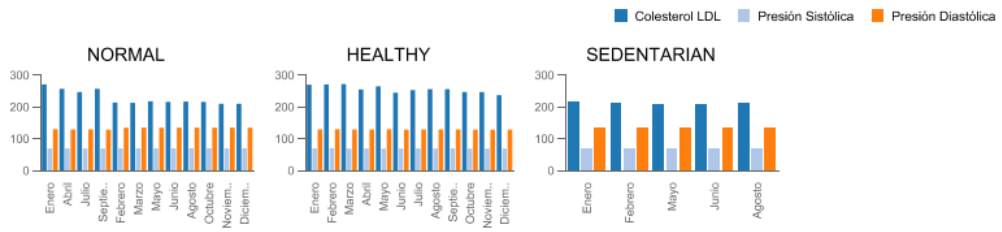


Ilustración 120, Evolución en el tiempo de los pacientes bajo un tratamiento natural por hábito

Otra forma de ver esta relación es la distribución por meses de manera que se pueda comparar más asertivamente cómo en un mismo mes dos o más hábitos pueden representar un aumento o disminución en los indicadores.

Ilustración 121, Relación por meses de los hábitos de los pacientes bajo un tratamiento natural

Según la dieta se tiene:

Ilustración 122, Relación por meses de las dietas de los pacientes bajo un tratamiento natural

Predominancia de dietas de los pacientes con tratamientos naturales:

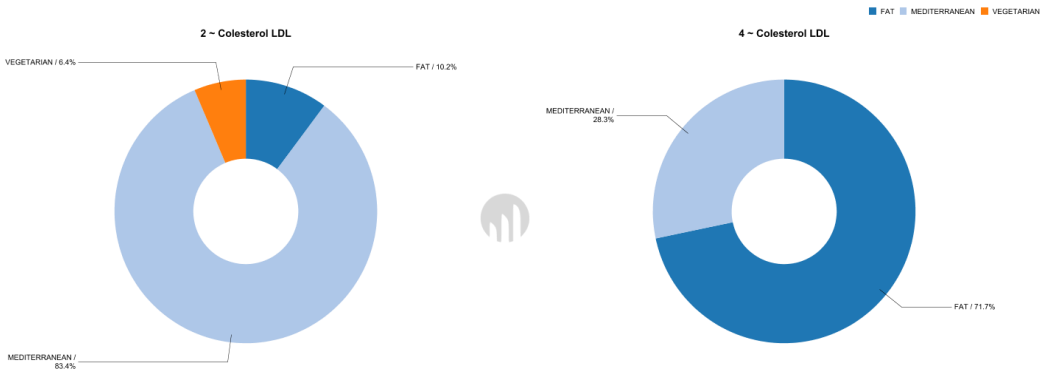


Ilustración 123, Predominancia de dietas en pacientes con tratamiento natural

De la misma manera se extrae la predominancia para los hábitos del paciente:

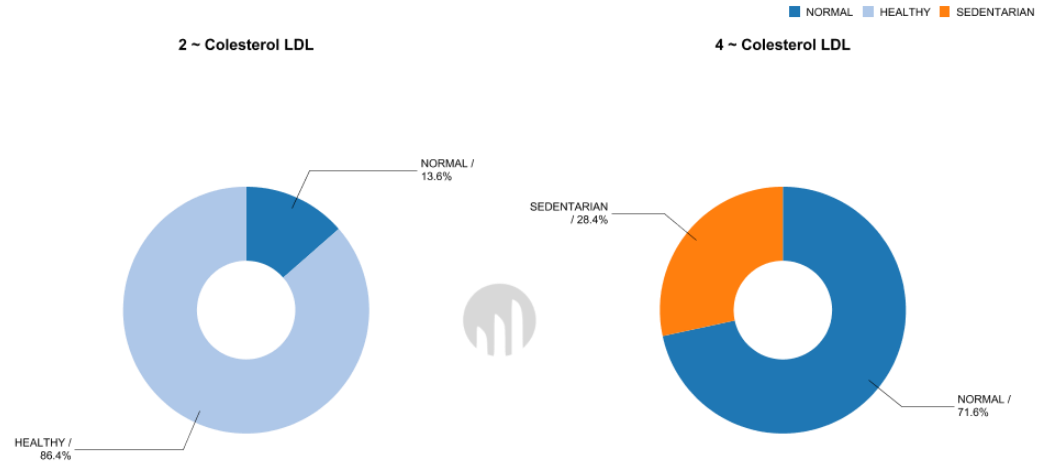


Ilustración 124, Predominancia de hábitos en pacientes con tratamiento natural

III. Tratamiento Farmacológico

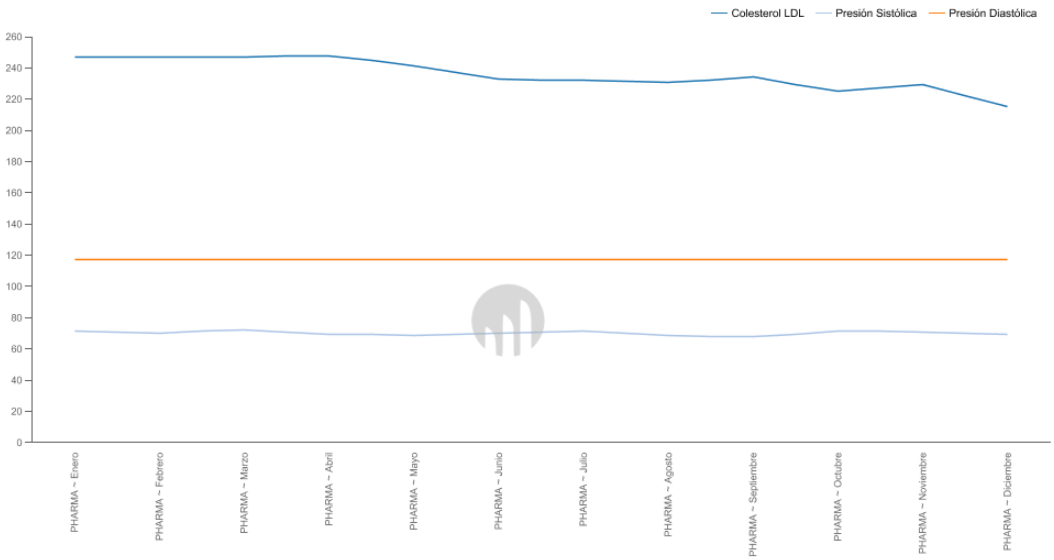


Ilustración 125, Evolución en el tiempo del tratamiento farmacológico en pacientes

La relación que tiene el estar sometido a un tratamiento farmacológico en el tiempo con diferentes hábitos es:

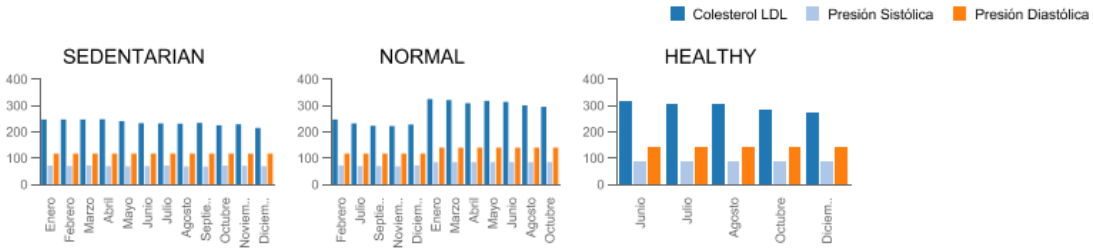


Ilustración 126, Evolución de los pacientes en el tiempo bajo un tratamiento farmacológico y sus hábitos

Otra forma de ver esto es identificando en cada mes los indicadores obtenidos al seguir varios hábitos:

Ilustración 127, Evolución bajo un tratamiento farmacológico y los hábitos de los pacientes por meses

De acuerdo con la dieta se tiene:

Ilustración 128, Evolución bajo un tratamiento farmacológico y las dietas de los pacientes por meses

A continuación, se detalla la predominancia de dietas de los pacientes con tratamientos farmacológicos:

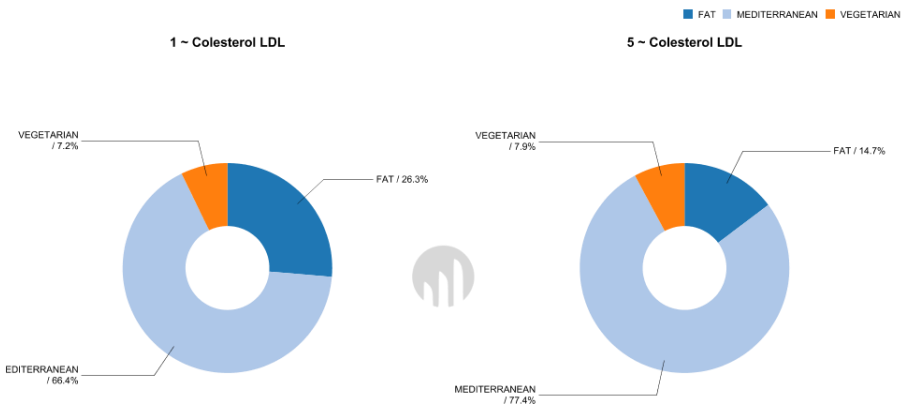


Ilustración 129, Predominancia de las dietas de los pacientes con tratamiento farmacológico

Predominancia del hábito en los pacientes con este tratamiento:

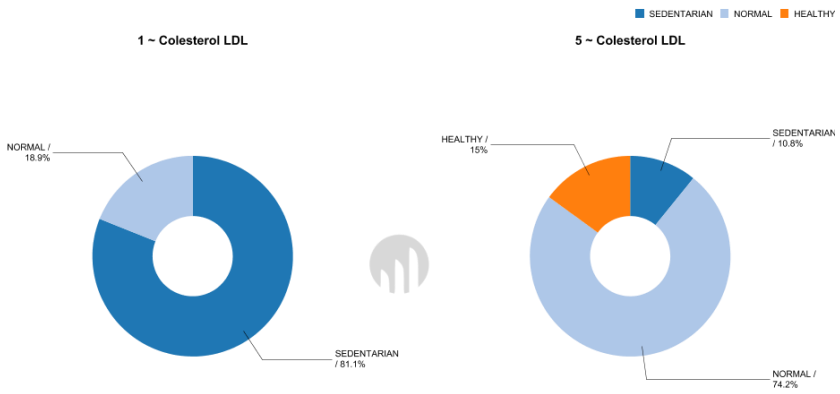


Ilustración 130, Predominancia de los hábitos de los pacientes con tratamiento farmacológico

Ahora bien, teniendo en cuenta la distribución geográfica de los pacientes es posible analizar los indicadores de salud de manera acumulada por tratamiento:

Tratamiento	Ciudad	Colesterol LDL	Presión Sistólica	Presión Diastólica
PHARMA	BARCELONA	234,115	70,317	117,5
	GETAFE	305,615	84,994	140,004
NATURAL	VILAFRANCA P.	254,788	69,388	129,24
	VITORIA	214,173	69,992	135,002
HOMEOPATHIC	ZUMAIA	315,788	80,002	140,767
	MADRID	341,75	73,696	124,302

Ilustración 131, Análisis acumulado de los indicadores por tratamiento y ciudad

Gráficamente la comparativa realizada entre tratamientos, ciudades e indicadores se puede visualizar de la siguiente manera:

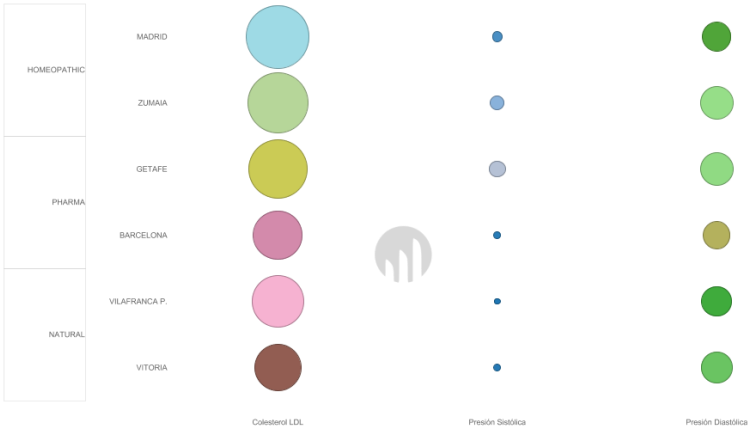


Ilustración 132, Análisis de los indicadores por tratamiento y ciudad

Si se desagrega por tratamiento la vista anterior se logra visualizar así:

I. Tratamiento Farmacológico

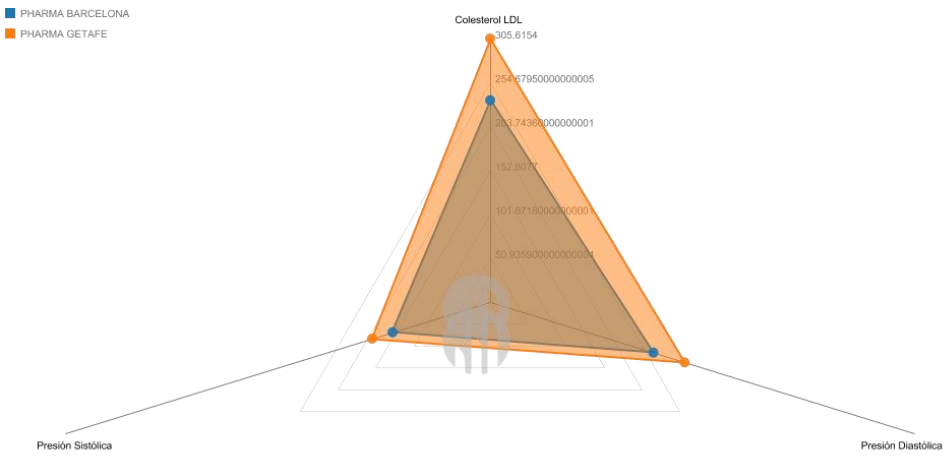


Ilustración 133, Indicadores de salud por ciudad con tratamiento farmacológico

II. Tratamiento natural

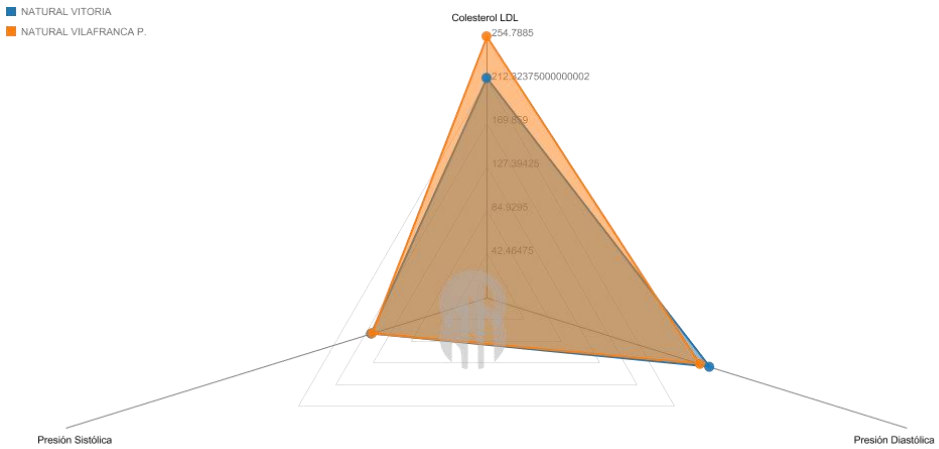


Ilustración 134, Indicadores de salud por ciudad con tratamiento natural

III. Tratamiento homeopático

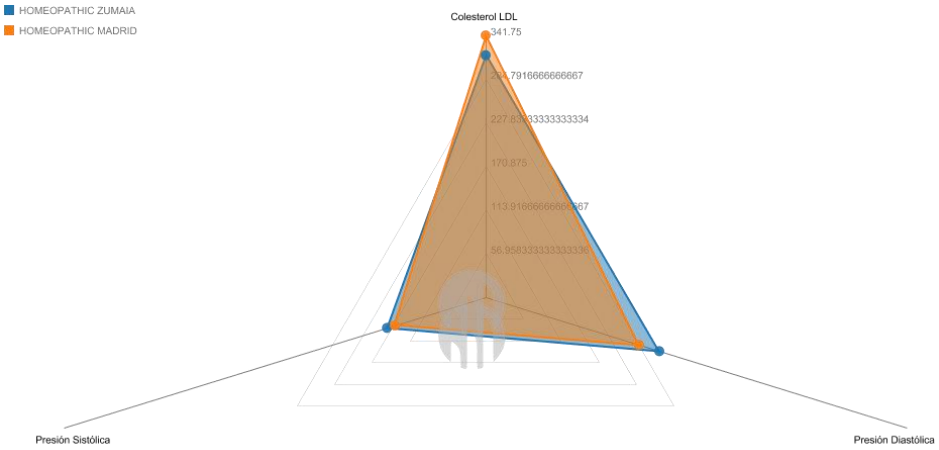


Ilustración 135, Indicadores de salud por ciudad con tratamiento homeopático

9. Conclusiones

Retomando las preguntas analíticas planteadas desde el inicio del proyecto, es posible determinar las siguientes conclusiones:

- **¿Cuál es la relación entre los diferentes tratamientos y la evolución de los pacientes?**

La evolución de los pacientes se mide en términos del aumento o disminución de los niveles de colesterol LDL, presión diastólica y presión sistólica en el tiempo. En este sentido, se ha evidenciado que en general los pacientes sometidos en un lapso determinado a un tratamiento homeopático presentan un incremento en los niveles de Colesterol LDL.

Un rango aceptable científicamente para la medición del colesterol LDL, indica que: [48]

V. 100 a 129 mg/dl es un valor de LDL ideal para la mayoría de las personas.
VI. 100 mg/dl o menos es un valor de LDL bueno para personas que tienen alto riesgo de infarto o ACV.
VII. 130 a 159 mg/dl está el valor límite superior de LDL.
VIII. 190 mg/dl o mayor se considera un valor de LDL muy alto.

En este caso, los pacientes presentan con los diferentes tratamientos un nivel de colesterol que oscila entre 215 y 330 bajo un tratamiento farmacológico, entre 210 y 274 para pacientes con tratamientos naturales y entre 300 y 354 para pacientes con tratamientos homeopáticos.
De acuerdo con los rangos aceptables para medición del Colesterol LDL, estos datos sobrepasan lo establecido en la escala más alta, lo que da entender que en general con cualquier tratamiento los pacientes tienen una alta probabilidad de sufrir un bloqueo en sus arterias que puede ser peligroso para el corazón.

Adicionalmente, teniendo en cuenta los rangos de presión arterial ilustrados a continuación:

Categoría	Sistólica (mmHg)		Diastólica (mmHg)
Hipotensión	menor de 80	o	menor de 60
Normal	80-120	y	60-80
Prehipertensión	120-139	o	80-89
Hipertensión grado 1 (HTA 1)	140-159	o	90-99
Hipertensión grado 2 (HTA 2)	160 o superior	o	100 o superior
Crisis hipertensiva (emergencia médica)	superior a 180	o	superior a 110

Fuente: American Heart Association

Ilustración 136, Gracias de la presión arterial. Tomado desde: Curiosoando

Y que los valores obtenidos para los diferentes tratamientos oscilan entre 75 y 131 (Presión sistólica y presión diastólica respectivamente), se puede inferir que estos pacientes presentan para los diferentes tratamientos un descontrol importante frente a este indicador. Según expertos, tener una presión sistólica elevada por un período prolongado puede aumentar el riesgo de padecer problemas cardiovasculares importantes, como un ataque cardíaco o un accidente cerebrovascular.

El objetivo para la presión sistólica recomendado para los adultos menores de 65 años con un riesgo de 10 por ciento o más de tener enfermedades cardiovasculares es de menos de 130 mm Hg. Para los adultos sanos de 65 años o más, el objetivo recomendado para tratamiento de la presión sistólica sigue siendo de menos de 130 mm Hg. [49]

Por lo tanto, entendiendo que el valor de 75 es menor que el indicado por expertos, las probabilidades de sufrir un accidente cerebrovascular son bajas.

La presión diastólica por su parte suele aumentar hasta los 50 años y a partir de esa edad tiende a disminuir. Esto provoca que en los pacientes mayores sea frecuente la presencia de hipertensión arterial sistólica aislada (tensión arterial sistólica mayor de 140 y una tensión arterial diastólica menor de 90), y que en los pacientes mayores de 40 años sea frecuente el encontrar una hipertensión arterial

diastólica aislada (tensión arterial sistólica menor de 140 y tensión arterial diastólica mayor de 90).

El tener solamente elevada la tensión arterial diastólica (la baja) suele ser habitual de gente joven (menor de 40 años) y aunque inicialmente se consideró una enfermedad benigna que no habría que tratar, actualmente se tiene claridad de que más del 80 % de estos pacientes terminan desarrollando también una elevación de la tensión arterial sistólica antes de 10 años y que es necesario tratar con alguna especificación médica. [50]

En consecuencia, los valores de presión para cada uno de los tratamientos oscilan entre rangos “normales” de acuerdo con los grupos etarios, para los cuales se puede suministrar un tratamiento de cualquier tipo, ***preferiblemente natural o farmacológico*** ya que el tratamiento homeopático es el que más representa un riesgo para el aumento del Colesterol LDL y la presión.

¿Existen terapias más eficaces?

La eficacia de los tratamientos se mide bajo dos criterios esenciales:

1. Disminución del colesterol LDL.
2. Normalización de la presión arterial o de acuerdo con los rangos de edades una presión sistólica baja y una presión diastólica en rangos “normales”.

Bajo las anteriores premisas, la terapia más eficaz para la diminución del colesterol LDL, y a su vez unos valores de presión estables es el tratamiento de tipo natural.

¿Ha influido en el resultado, los hábitos de los pacientes?

Para recoger todos aquellos análisis desagregados anteriormente donde se relacionaban los tratamientos, los hábitos y las dietas de los pacientes con sus indicadores de salud, se presenta la siguiente ilustración:

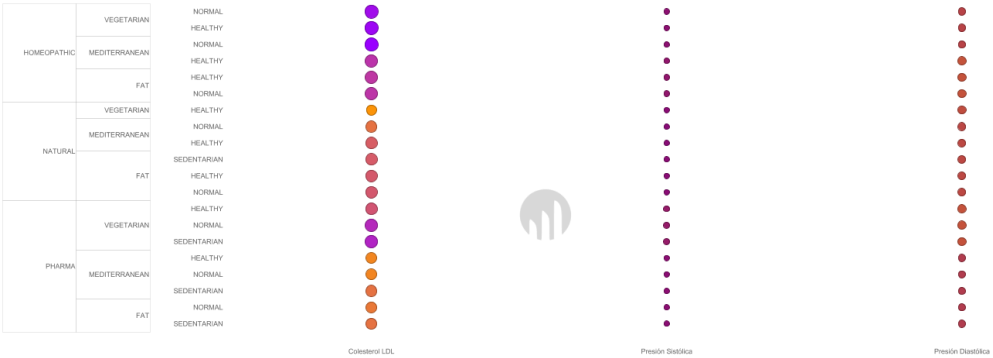


Ilustración 137, Relación entre tratamientos, dietas y hábitos

Los aspectos conclusivos son:

1. Los valores de presión sistólica no representan una variación “importante” dependiendo de la dieta, el hábito o el tratamiento. Su alza puede significar un alto riesgo de ACV.
2. Sostener en el tiempo un tratamiento homeopático siguiendo una dieta y hábito de cualquier tipo no representa una disminución importante en los niveles de colesterol LDL de los pacientes.

3. Los pacientes que llevan un tratamiento natural cuyos hábitos son saludables evidencian unos niveles de colesterol LDL más bajos.
4. Los pacientes cuyo tratamiento es farmacológico presentan menores niveles de colesterol LDL cuando llevan una dieta mediterránea o vegetariana y un hábito o actividad normal o saludable en su día a día.
5. El nivel de colesterol LDL más bajo es el resultante de combinar un tratamiento natural con una dieta vegetariana y un hábito saludable, entendiendo que el consumo de verduras y productos no grasos mejora notablemente los niveles de colesterol acompañado de actividad física constante.

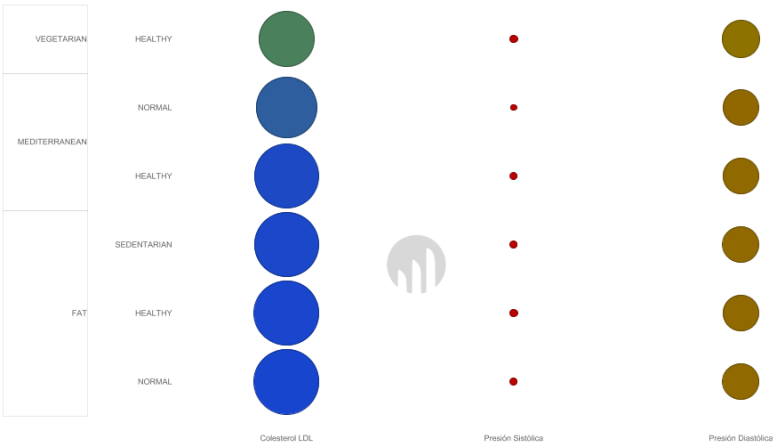
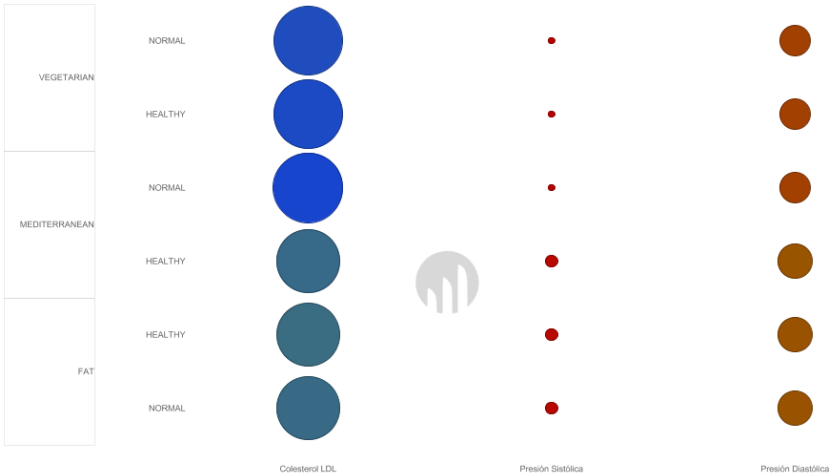


Ilustración 138, Relación tratamiento natural vs dieta vs hábito.

6. Los pacientes cuyo tratamiento es farmacológico y cuyos hábitos son sedentarios son más propensos a un alza del colesterol.

En la siguiente ilustración, por ejemplo, se relaciona el detalle del tratamiento homeopático, lo que permite concluir que los pacientes que lleven con este tratamiento, una dieta mediterránea y un hábito saludable evidencian disminución en sus niveles de colesterol.



- **¿La evolución a lo largo del tiempo, por un mismo tratamiento, dependen de algún factor como los hábitos?**

En general los expertos afirman que los resultados de un tratamiento dependen de la edad del paciente y de sus hábitos. Por lo tanto, considerando diferentes variables en el análisis, los hábitos o actividades del paciente siempre tendrán un efecto positivo o negativo sobre su salud.

Para el caso del tratamiento homeopático se ha podido retratar que seguir un hábito de actividad normal representa una variabilidad mayor en los niveles de presión diastólica. Adicionalmente, Los niveles de presión son más bajos cuando el hábito es saludable. Con respecto a la dieta se encuentra que una dieta vegetariana en el tiempo genera mayores niveles de colesterol.

El uso de un tratamiento natural en el tiempo evidencia que los pacientes cuyo hábito es normal encuentran una disminución notable de los niveles de colesterol. Igualmente, el seguir una dieta mediterránea representa mayor estabilidad en los niveles de colesterol.

Con respecto al tratamiento de tipo farmacológico, en el tiempo los niveles de presión representan una variabilidad. Un hábito sedentario representa una estabilidad en los niveles de colesterol y la dieta mediterránea contribuye a mejorar el estado de salud del paciente.

- **¿Hay diferencias en el resultado de un tratamiento según el lugar geográfico del paciente?**

Hablar de que el lugar geográfico del paciente puede implicar un cambio en los indicadores de salud de este, en principio puede ser erróneo. Lo anterior, porque los indicadores de salud como ya se ha mencionado, dependen de factores como los hábitos y la alimentación, sin embargo, las costumbres de un lugar u otro pueden influir en estos dos últimos parámetros. Probablemente, este sea un buen aspecto para relacionar la ubicación geográfica de un paciente con su estado de salud.

La siguiente ilustración resume los indicadores de salud por tratamiento y ciudad, en lo que respecta a Zumaia y Madrid los niveles de colesterol son los más altos. Adicionalmente, **los valores de presión y colesterol son similares bajo un mismo tratamiento en una misma ubicación, lo que quiere decir que hay diferencias, pero no son realmente notorias según la ubicación geográfica del paciente.**

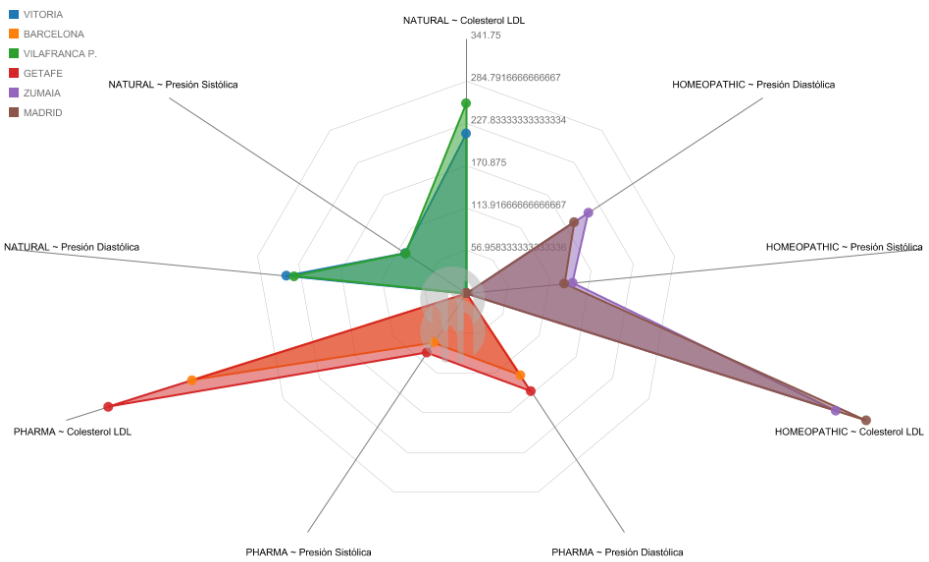


Ilustración 139, Análisis de los indicadores por tratamiento y ciudad

Investigando el estado de salud de la población madrileña, en 2018 **La Fundación Española del Corazón (FEC), promovida por la Sociedad Española de Cardiología,** urge al Ministerio de Sanidad y las comunidades autónomas que sitúen el hipercolesterolemia en sus agendas políticas y desarrollen un plan nacional para mejorar su abordaje, ya que se estima que más de la mitad de la población española lo tiene alto pero el 50 por ciento está sin diagnosticar. [51]
Adicionalmente, este mismo año se reportó que casi la mitad de los madrileños tiene sobrepeso o es obeso. [52]

En particular los tratamientos farmacológicos y homeopáticos son los que representan para el proyecto, los niveles más altos de presión y colesterol en las ciudades de Madrid, Zumaia, Barcelona y Getafe. No es la ubicación geográfica la que diferenciaría los indicadores, son los hábitos de las personas los que se reflejan en su estado de salud.

Por ejemplo, los remedios homeopáticos para reducir el colesterol son considerados medicamentos al igual que las estatinas. Estos remedios contienen sustancias naturales muy diluidas, y que complementan los tratamientos convencionales. Su objetivo es reducir la concentración de esta grasa en sangre, pero se debe indicar teniendo en cuenta la personalidad del paciente y buscar el remedio homeopático específico contra el colesterol alto. No sólo es asunto de tomar medicamentos.
Para la homeopatía es importante la personalidad del paciente, y aquellas problemáticas que estén sucediendo en el entorno de quien sufre de colesterol alto. Esto se debe a que existen muchas causas que podrían provocar la hipercolesterolemia, incluso angustia, stress y depresión. Estos factores pueden influir de dos formas, a través del aumento de consumo de alimentos, especialmente aquellos ricos en grasas y azucares, o a través del aumento de síntesis del colesterol endógeno sin causa aparente. [53]

Probablemente, estos pacientes no fueron tratados en otros aspectos durante el año en el que se hicieron las mediciones y sus hábitos en sus lugares de residencia tampoco fueron los mejores.

- **¿Hay algún periodo del año donde el tratamiento sea más o menos efectivo?**

El tratamiento homeopático es menos efectivo en el segundo semestre del año en particular el diciembre. El tratamiento natural, por su parte, es menos efectivo en el primer semestre del año en particular en enero y marzo. Y por último el farmacológico es menos efectivo entre los meses de enero y abril.

La tendencia entonces es que durante los tres primeros meses del año lo tratamientos son menos efectivos, esto puede explicarse a raíz de que durante este periodo la estación predominante en España es el invierno. Probablemente se tienen diferentes hábitos alimenticios pero poca actividad física. “La gente prefiere estar tranquilo y en casa”.

A raíz de todas las afirmaciones expuestas anteriormente, cabe destacar que, aunque son muchas las causas que llevan a que el colesterol se eleve más en unas personas que en otras, como por ejemplo heredar la hipercolesterolemia de los padres, aún en estas situaciones, seguir una dieta sana (como la mediterránea) hacer ejercicio físico a diario, y no fumar, mejora las cifras de colesterol”. [54]

Por último, alrededor de la realización de este proyecto es posible concluir que las herramientas de BI representan una utilidad importante en el momento de automatizar la correlación y el análisis de grandes volúmenes de información. La correcta definición de una arquitectura de BI puede llevar a casos de éxito o fracaso de acuerdo con el tamaño de la organización, el tipo de información a analizar, la descentralización de esta y demás.
Es importante reflejar en el diseño del sistema de BI las necesidades de negocio ya que pensar “lógicamente” desconociendo cómo se mueve una organización o el contexto de un problema planteado puede llevar a brindar soluciones con aspectos faltantes e inmaduros.

10. Glosario

Dieta fat: dieta basada en un alto consumo calórico.

Dieta Mediterránea: dieta basada en vegetales, carnes blancas (principalmente pollo, pescado), frutas y verduras.

Dieta vegetariana: Aunque hay diferentes tipos la más estricta excluye las carnes y productos animales.

Estatinas: son medicamentos para disminuir los niveles de colesterol.

Hábito normal: consiste en que el paciente realiza actividad física regularmente y de baja intensidad algunas veces a la semana.

Hábito saludable: consiste en que el paciente realiza actividad física de manera prologada varias veces por semana.

Hábito sedentario: consiste en que el paciente prefiere quedarse en total quietud y la actividad física es casi nula a la semana.

Hipercolesterolemia: aumento de la cantidad normal de colesterol en la sangre.

Presión arterial: la presión arterial es la fuerza que ejerce la sangre contra las paredes de las arterias. Cada vez que el corazón late, bombea sangre hacia las arterias. [55]

Presión diastólica: cuando el corazón está en reposo, entre un latido y otro, la presión sanguínea disminuye. A esto se le llama presión diastólica. [55]

Presión sistólica: cuando el corazón late para bombear la sangre se incrementa la presión arterial, a esto se le llama presión sistólica. [55]

Tratamiento natural: son tratamientos a base de plantas y suplementos naturales que permiten generar disminución en los niveles del colesterol.

Tratamiento farmacológico: es un tratamiento consistente en suministrar fármacos para disminuir el Colesterol LDL. Se cree que las estatinas son los mejores fármacos para usarlos en personas que necesitan medicamentos para bajar su colesterol. [56]

Tratamiento homeopático: el tratamiento homeopático es un elemento medicinal que no busca restablecer el orden o suprimir las manifestaciones de la enfermedad de manera directa, sino que tiene como objetivo estimular la reacción curativa del enfermo. Este se basa a partir de medicamentos homeopáticos elaborados a partir de extractos de plantas, como la Belladonna, la Pulsatilla o el Árnica; de sustancias minerales como el fósforo o el azufre; de sustancias de origen animal como la abeja o la cantárida. Con mucha menor frecuencia, de algunas sustancias químicas de síntesis. También se utilizan en homeopatía medicamentos llamados bioterápicos, hechos a partir de cultivos microbianos u otras sustancias de origen microbiológico. [57]

Saiku: es una herramienta OLAP destinada a usuarios finales de Pentaho, que permite visualizar y realizar análisis de datos de forma fácil e intuitiva. Es la interfaz gráfica del portal web mediante la cual pueden construirse vistas propias arrastrando y soltando campos. [58]

11. Bibliografía

[1] D. Amoróz Alcaraz, «M1.221- Trabajo final de máster Sistema de inteligencia de negocio para el análisis de los tratamientos de reducción del colesterol,» Universitat Oberta de Catalunya, España, 2018.

[2] P. B. y. F. Ruso, «Maldición genética en Sevilla: 400 miembros de una misma familia, enfermos por colesterol,» El Español, 22 Septiembre 2018. [En línea]. Available: https://www.elespanol.com/reportajes/20180922/maldicion-genetica-sevilla-miembros-familia-enfermos-colesterol/339717247_0.html. [Último acceso: 23 Septiembre 2018].

[3] Redacción Agencia Notimex, «Hoy Día Mundial del Colesterol, el asesino silencioso,» Agencia Notimex, 19 Septiembre 2018. [En línea]. Available: http://netnoticias.mx/2018-09-19-f86a83db/hoy-dia-mundial-del-colesterol-el-asesino-silencioso/. [Último acceso: 23 Septiembre 2018].

[4] E. Morán, «Un estudio cuestiona la eficacia de los fármacos contra el colesterol a edades avanzadas,» Huelva Información, 18 Abril 2018. [En línea]. Available: https://www.huelvainformacion.es/huelva/cuestiona-eficacia-farmacos-colesterol-avanzadas_0_1237376635.html. [Último acceso: 23 Septiembre 2018].

[5] Rivadera. Gustavo, «La metodología de Kimball para el diseño de almacenes de datos (Data warehouses),» S.A. [En línea]. Available: http://www1.ucasal.edu.ar/htm/ingenieria/cuadernos/archivos/5-p56-rivadera-formateado.pdf. [Último acceso: 29 Octubre 2018].

[6] Zorrilla. Marta, «Datawarehouse y OLAP, Universidad de Cantabria,» Noviembre 2010. [En línea]. Available: http://docshare01.docshare.tips/files/24773/247735327.pdf. [Último acceso: 16 Diciembre 2018].

[7] Chaudhuri. Surajit , Dayal. Umeshwar, Narasayy. Vivek , «An Overview of Business Intelligence Technology,» *Review Articles,* vol. 54, nº 8, p. 11, 2011.

[8] Roosboard.com, «History Of Business Intelligence To Be Evolved From Then & Now,» Roos Board, 2016. [En línea]. Available: https://roosboard.com/blog/history-of-business-intelligence-to-be-evolved-from-then-and-now.html. [Último acceso: 15 Octubre 2018].

[9] BARC – Business Application Research Center , «Benefits and Advantages of Business Intelligence Systems,» BI-Survey.com, 2017. [En línea]. Available: https://bi-survey.com/benefits-business-intelligence. [Último acceso: 21 Octubre 2018].

[10] Muraina, Ishola & Ahmad, Azizah, «Healthcare Business Intelligence: The Case of University's Health Center,» *International Conference on E-CASE & E-TECH,* p. 27, 2012.

[11] Jiang. Lihong, Cai. Hongming, Xu. Boyi, «A Domain Ontology Approach in the ETL Process of Data Warehousing,» *IEEE International Conference on e-Business Engineering,* vol. 3, pp. 30-35, 2010.

[12] Rodrigues. Eliana, Diniz Morais. Espelho, Sidnéia da Silva. Silvia, Carlos Caritá. Edilson, «Business Intelligence utilizando tecnologias Web para análise de fatores de risco na ocorrência de doença arterial coronariana,» *Journal of Health Informatics,* vol. 2, nº 1, p. 13, 2010.

[13] Mihaela IVAN, Manole VELICANU , «Healthcare Industry Improvement with Business Intelligence,» *Informatica Economică ,* vol. 19, nº 2, p. 9, 2015.

[14] Redacción Citizen Tribune, «Orion Health launches Amadeus CORE to help healthcare sector leverage data analytics and machine learning,» Citizen Tribune, 15 Octubre 2018. [En línea]. Available: https://www.citizentribune.com/news/business/orion-health-launches-amadeus-core-to-help-healthcare-sector-leverage/article_99e6b87d-e094-5f83-a0ba-705e373cb43f.html. [Último acceso: 20 Octubre 2018].

[15] P. S. S.A, «"Los sistemas de salud en el mundo empezarán a pagar basados en resultados",» Semana, 10 Agosto 2018. [En línea]. Available: https://www.semana.com/nacion/articulo/los-sistemas-de-salud-en-el-mundo-empezaran-a-pagar-basados-en-resultados/586205. [Último acceso: 20 Octubre 2018].

[16] Oracle, «¿Por qué Inteligencia de Negocios?,» S.A. [En línea]. Available: https://www.oracle.com/ocom/groups/public/@otn/documents/webcontent/317529_esa.pdf. [Último acceso: 15 Octubre 2018].

[17] Hoppe. Geoff, «The Top 15 Free and Open Source Business Intelligence Software,» Capterra, 12 Julio 2018. [En línea]. Available: https://blog.capterra.com/top-8-free-and-open-source-business-intelligence-software/. [Último acceso: 21 Octubre 2018].

[18] James. Harry, «Top 10 free and open source business intelligence software,» Crayon Data, 05 Enero 2018. [En línea]. Available: https://bigdata-madesimple.com/top-10-free-and-open-source-business-intelligence-software/. [Último acceso: 21 Octubre 2018].

[19] Redacción TodoBI, «Comparativa de herramientas Business Intelligence,» 12 Diciembre 2017. [En línea]. Available: http://www.todobi.com/2017/04/comparativa-de-herramientas-business.html. [Último acceso: 28 Octubre 2018].

[20] Brandão. Andreia, Pereira. Eliana, Esteves. Marisa, Portela.Filipe, Filipe Santos. Manuel, Abelha. António, Machado. José , «A Benchmarking Analysis of Open-Source Business Intelligence Tools in Healthcare Environments,» *MDPI,* vol. 7, nº 57, p. 16, 2016.

[21] Redacción Hitachi Vantara, «Components Reference,» Hitachi Inspire the Next, 12 Septiembre 2018. [En línea]. Available: https://help.pentaho.com/Documentation/8.1/Setup/Components_Reference. [Último acceso: 03 Noviembre 2018].

[22] Redacción MySQL Community Downloads , «MySQL Community Downloads,» S.I., 2018. [En línea]. Available: https://dev.mysql.com/downloads/. [Último acceso: 04 Noviembre 2018].

[23] Apache Tomcat, «Tomcat 8 Software Downloads,» S.I., 2018. [En línea]. Available: https://tomcat.apache.org/download-80.cgi. [Último acceso: 04 Noviembre 2018].

[24] Redacción Blog Noelonassis, «Blog Noelonassis,» Mayo 2017. [En línea]. Available: https://noelonassis.files.wordpress.com/2017/05/apache-tomcate-web-server-online-training.jpg. [Último acceso: 04 Noviembre 2018].

[25] Hitachi Vantara, «Pentaho 8.1,» S.I., 2017. [En línea]. Available: https://community.hitachivantara.com/community/products-and-solutions/pentaho. [Último acceso: 04 Noviembre 2018].

[26] Hitachi Vantara Corporation, «Download Data Integration,» Hitachi Vantara Inspire The Next, 2018. [En línea]. Available: http://events.pentaho.com/CE-Download_Data-Integration-ALL-OS.html. [Último acceso: 17 Noviembre 2018].

[27] Moran, Doug , «Pentaho Data Integration (Kettle) Tutorial,» Hitachi, Inspire The Next, 11 Diciembre 2015. [En línea]. Available: https://wiki.pentaho.com/display/EAI/Pentaho+Data+Integration+%28Kettle%29+Tutorial. [Último acceso: 30 Noviembre 2018].

[28] Redacción sitio web pentaho.almacen-datos, «Portada sobre la plataforma Pentaho Open Source Business Intelligence,» 2006-2011. [En línea]. Available: http://pentaho.almacen-datos.com/mondrian.html. [Último acceso: 02 Diciembre 2018].

[29] SourceForge, «Mondrian,» Slashdot Media, 2018. [En línea]. Available: https://sourceforge.net/projects/mondrian/files/schema%20workbench/. [Último acceso: 02 Diciembre 2018].

[30] Espinosa, Roberto, «El Rincón del BI, Descubriendo el Business Intelligence,» 04 Julio 2010. [En línea]. Available: https://churriwifi.wordpress.com/2010/07/04/17-3-preparando-el-analisis-

dimensional-definicion-de-cubos-utilizando-schema-workbench/. [Último acceso: 02 Diciembre 2018].

[31] Redacción Meteorite.bi, «Meteorite,» 2018. [En línea]. Available: https://licensing.meteorite.bi/login. [Último acceso: 03 Enero 2019].

[32] Pentaho Corporation, «Pentaho MarketPlace,» 2005-2019. [En línea]. Available: https://www.pentaho.com/marketplace/. [Último acceso: 01 Enero 2019].

[33] Redacción ETL-Tools.Info, «Business Intelligence - Almacenes de Datos - ETL,» ETL-Tools.Info, 2006-2018. [En línea]. Available: https://etl-tools.info/es/bi/almacendedatos_esquema-estrella.htm. [Último acceso: 15 Diciembre 2018].

[34] Hitachi Solutions Asia Pacific Pte. Ltd. , «BI and Data warehousing Services,» Hitachi Inspire The Next, 2018. [En línea]. Available: http://www.ignify.com/services_business_intelligence.asp. [Último acceso: 17 Noviembre 2018].

[35] Bustamante Martínez, A., & Galvis Lista, E., & Gómez Flórez, L., «Técnicas de modelado de procesos de ETL: una revisión de alternativas y su aplicación en un proyecto de desarrollo de una solución de BI,» *Redalyc.org,* vol. 18, nº 1, pp. 185-191, 2013.

[36] Bustamante Martínez, Alexander, Galvis Lista, Ernesto Amaru, Gómez Flórez, Luis Carlos, «Técnicas de modelado de procesos de ETL: una revisión de alternativas y su aplicación en un proyecto de desarrollo de una solución de BI,» *Redalyc.org,* vol. 18, nº 1, pp. 185-191, 2013.

[37] Blood Pressure Association, «Blood pressure chart,» 2008. [En línea]. Available: http://www.bloodpressureuk.org/BloodPressureandyou/Thebasics/Bloodpressurechart. [Último acceso: 02 Diciembre 2018].

[38] León Guzmán. Elizabeth , «Minería de Datos,» Universidad Nacional de Colombia, S.A.. [En línea]. Available: http://disi.unal.edu.co/~eleonguz/cursos/md/presentaciones/Sesion4-OLAP.pdf. [Último acceso: 14 Diciembre 2018].

[39] Redacción OLAP.com, «OLAP Cube,» OLAP.com, S.A.. [En línea]. Available: http://olap.com/learn-bi-olap/olap-bi-definitions/olap-cube/. [Último acceso: 14 Diciembre 2018].

[40] Ibarra. Maria de los Ángeles, «Procesamientos Analítico en Línea, OLAP,» Universidad Nacional del Nordeste , Corrientes, Argentina, 2006.

[41] Redacción Microsoft, «Cubos en modelos multidimensionales,» 01 Mayo 2018. [En línea]. Available: https://docs.microsoft.com/es-es/sql/analysis-services/multidimensional-models/cubes-in-multidimensional-models?view=sql-server-2017. [Último acceso: 16 Diciembre 2018].

[42] Redacción Microsoft, «Dimensiones en modelos multidimensionales,» 01 Mayo 2018. [En línea]. Available: https://docs.microsoft.com/es-es/sql/analysis-services/multidimensional-models/dimensions-in-multidimensional-models?view=sql-server-2017. [Último acceso: 16 Diciembre 2018].

[43] R. Microsoft, «Crear medidas y grupos de medida en modelos multidimensionales,» 01 Mayo 2018. [En línea]. Available: https://docs.microsoft.com/es-es/sql/analysis-services/multidimensional-models/create-measures-and-measure-groups-in-multidimensional-models?view=sql-server-2017. [Último acceso: 16 Diciembre 2018].

[44] R. Microsoft, «Particiones en modelos multidimensionales,» 05 Mayo 2018. [En línea]. Available: https://docs.microsoft.com/es-es/sql/analysis-services/multidimensional-models/partitions-in-multidimensional-models?view=sql-server-2017. [Último acceso: 16 Diciembre 2018].

[45] Redacción Microsoft, «Perspectivas de modelos multidimensionales,» 01 Mayo 2018. [En línea]. Available: https://docs.microsoft.com/es-es/sql/analysis-services/multidimensional-models/perspectives-in-multidimensional-models?view=sql-server-2017. [Último acceso: 16 Diciembre 2018].

[46] Redacción Microsoft, «Crear jerarquías definidas por el usuario,» 01 Mayo 2018. [En línea]. Available: https://docs.microsoft.com/es-es/sql/analysis-services/multidimensional-models/user-defined-hierarchies-create?view=sql-server-2017. [Último acceso: 16 Diciembre 2018].

[47] Redacción Microsoft, «Acciones en modelos multidimensionales,» 01 Mayo 2018. [En línea]. Available: https://docs.microsoft.com/es-es/sql/analysis-services/multidimensional-models/actions-in-multidimensional-models?view=sql-server-2017. [Último acceso: 16 Diciembre 2018].

[48] R. breastcancer.org, «Colesterol alto,» Breastcancer.org, 17 Septiembre 2012. [En línea]. Available: https://www.breastcancer.org/es/tratamiento/efectos_secundarios/colesterol_alto. [Último acceso: 05 Enero 2019].

[49] Sheldon G. Sheps, M.D., «Hipertensión sistólica aislada: ¿un problema de salud?,» 10 Abril 2018. [En línea]. Available: https://www.mayoclinic.org/es-es/diseases-conditions/high-blood-pressure/expert-answers/hypertension/faq-20058527. [Último acceso: 05 Enero 2019].

[50] N. Ortega, «¿Qué diferencia existe entre tensión arterial sistólica y diastólica?,» Febrero 2017. [En línea]. Available: https://salud-1.com/enfermedades/diferencia-entre-tension-arterial-sistolica-y-diastolica/. [Último acceso: 05 Enero 2019].

[51] R. tribunapalencia.com, «España tiene el colesterol alto... ¿por qué?,» 18 Enero 2018. [En línea]. Available: https://www.tribunapalencia.com/noticias/espana-tiene-el-colesterol-alto-dot-dot-dot-por-que. [Último acceso: 06 Enero 2018].

[52] Redacción Telemadrid.es, «Casi la mitad de los madrileños tiene sobrepeso o es obeso,» 2018 Febrero 12. [En línea]. Available: http://www.telemadrid.es/noticias/madrid/mitad-madrilenos-sobrepeso-obeso-0-1984901532--20180212041400.html. [Último acceso: 06 Enero 2019].

[53] R. Abajarcolesterol.com, «Homeopatía contra el colesterol alto,» 2012. [En línea]. Available: https://www.abajarcolesterol.com/homeopatia-para-reducir-el-colesterol/. [Último acceso: 06 Enero 2019].

[54] J. R. Vico, «Más de la mitad de los españoles tiene hipercolesterolemia,» 12 Septiembre 2017. [En línea]. Available: https://as.com/deporteyvida/2017/09/12/portada/1505231242_939055.html. [Último acceso: 06 Enero 2019].

[55] Redacción MedlinePlus, «Presión arterial baja,» 22 Agosto 2018. [En línea]. Available: https://medlineplus.gov/spanish/lowbloodpressure.html. [Último acceso: 05 Enero 2019].

[56] Redacción MedlinePlus, «Tratamiento farmacológico para el colesterol,» 22 Febrero 2018. [En línea]. Available: https://medlineplus.gov/spanish/ency/patientinstructions/000314.htm. [Último acceso: 05 Enero 2019].

[57] Eizayaga. José E, «Consultorio médico-homeopático Doctores Eizayaga,» Septiembre 2018. [En línea]. Available: https://consultorioeizayaga.com/en-que-consiste-la-homeopatia/#medicamentos. [Último acceso: 03 Enero 2019].

[58] Redacción Universidad Nacional de Córdoba, «Manual de usuario, Ucumari, Sistema de soporte a la toma de decisiones,» 2015. [En línea]. Available: http://ucumari.unc.edu.ar/repositorio/archivos/Tutorial_Ucumari_Pentaho5.pdf. [Último acceso: 03 Enero 2019].

[59] Vertex42.com, «Gantt Chart Template for Excel,» 2006 - 2018. [En línea]. Available: https://www.vertex42.com/ExcelTemplates/excel-gantt-chart.html. [Último acceso: 29 Septiembre 2018].

[60] Redacción Sitio Modrian.Pentaho.com, «Mondrian Schema Workbench,» S.A. [En línea]. Available: https://mondrian.pentaho.com/documentation/schema_workbench.pdf. [Último acceso: 02 Diciembre 2018].

12. Anexos

12.1. Anexo 1, Selección entorno tecnológico

ID	Herramienta	Descripción general	Requisito									
			Flexibilidad para manejar diferentes orígenes de datos (CSV, XML, Excel, entre otros).	Generación de reportes (KPIS, Indicadores)	Procesos ETL	Nivel de esfuerzo para uso de la interfaz	OLAP Analysis	Data Mining	Buena calidad y cantidad en la Documentación	Reportes de Georreferenciación	Enlace de consulta	Puntaje obtenido
1	Pentaho Community Edition.	El software Pentaho BI Suite fue desarrollado por Pentaho Corporation en 2001, en el lenguaje Java, siendo la primera plataforma de BI en ser lanzada como una alternativa de código abierto en el mercado. El proyecto Pentaho BI Suite comprende un conjunto de productos: plataforma de BI (servidor), informes, análisis OLAP, integración de datos (ETL), paneles y minería de datos.	x	x	x	Bajo	x	x	x	Sin información	https://goo.gl/QZTnS3 https://gravitar.biz/pentaho/ http://www.innoventsolutions.com/comparison-matrix.html	7
2	SpagoBI.	La herramienta Spago BI es un software de código abierto completo, y solo hay una versión única, es decir, la edición comunitaria, una versión completamente gratuita. Es una herramienta desarrollada por Spago World y respaldada por una comunidad de código abierto.	Sin información	x	x	Bajo	x	x	x	x	https://es.wikipedia.org/wiki/SpagoBI http://www.stratebi.com/spagobi	7
3	Jaspersoft BI Tools.	Jaspersoft BI es una herramienta desarrollada en 2001, en los lenguajes Java y Perl. Esta es una herramienta de código abierto disponible en dos versiones, la versión Community y la versión Enterprise. Jaspersoft OLAP es un entorno poderoso para el análisis de datos al que se puede acceder a través de una interfaz de usuario intuitiva, diseñada para el análisis de grandes volúmenes de conjuntos de datos y para realizar consultas analíticas complejas. Consistente en un motor OLAP, esta herramienta proporciona un entorno interactivo para que los usuarios realicen operaciones de división y dados, pivote y fi ltro y resuman los datos en tiempo real a través de una interfaz basada en la web o MS Excel.	x	x	x	Medio	x	x	x	Sin información	https://goo.gl/wzTPRL https://goo.gl/GJTBE7 http://www.innoventsolutions.com/comparison-matrix.html	6.5
4	Tableau Public.	Es una herramienta que está orientada a que usuarios comunes y corrientes sean capaces de interpretar y comprender los datos.	x	x	Poca información	Bajo	Sin información	x	x	x	https://blog.bi-geek.com/tableau-parte-i-introduccion/ https://goo.gl/QZTnS3	6
5	Microsoft Power BI.	Es una herramienta que transforma los datos de las empresas en objetivos visuales, que permiten que el usuario se centre en lo que realmente importa. Se compone de un conjunto de aplicaciones de análisis de negocios que permiten realizar análisis sobre los datos y compartir información.	x	x	x	Bajo	Conexión a cubos OLAP. Ambiente no integrado	Sin información	x	Sin información	https://goo.gl/QZTnS3 https://www.softeng.es/es-es/blog/power-bi-descubre-power-query.html	5
6	QlikView Personal Edition	QlikView (QV) es un software de BI desarrollado por la compañía sueca QlikTech. Aunque este software es un producto patentado, la compañía proporciona una versión de desarrollo completa del software de forma gratuita.	Sin información	x	x	Bajo	Conexión a cubos OLAP. Ambiente no integrado	x	x	Sin información	https://goo.gl/wzTPRL https://goo.gl/QZTnS3	5
7	KNIME Business Intelligence Tools.	KNIME, Konstanz Information Miner, es un software de código abierto de inteligencia empresarial que integra varios componentes para el aprendizaje automático y la extracción de datos. Una interfaz gráfica de usuario permite el ensamblaje de nodos para el preprocesamiento de datos mediante extracción, transformación y carga (ETL), para modelado, análisis de datos y visualización.	Sin información	x	x	Medio	Sin información	x	Poca documentación	Sin información	https://www.knime.com/knime-for-data-scientists https://blog.statsbot.co/open-source-business-intelligence-523ba185d530 https://www.lis-solutions.es/blog/knime-un-software-vivo/	3.5
8	ReportServer Community Edition.	ReportServer es la plataforma de BI de código abierto, que viene con una gran selección de herramientas poderosas como informes de píxeles perfectos, análisis ad-hoc, informes de Excel y Word, o análisis de OLAP multidimensional.	x	x	Sin información	Medio	x	Sin información	Poca documentación	Sin información	https://reviews.financesonline.com/p/reportserver/ https://blog.statsbot.co/open-source-business-intelligence-523ba185d530	3.5
9	Birt	BIRT (Business Intelligence and Reporting Tool) es un proyecto de la comunidad de Eclipse que incluye un generador de gráficos, un generador de informes y un entorno de diseño. El proyecto se lanzó en 2005. El motor BIRT es una biblioteca que genera informes configurados y no configurados en formatos HTML, PDF, XLS, DOC y PPT. Estos informes pueden ser complejos, ya que contienen varias tablas, gráficos avanzados e imágenes. BIRT también es capaz de crear tablas dinámicas dinámicas. Los datos mostrados se pueden recuperar de diferentes bases de datos relacionales y multidimensionales (HOLAP y MOLAP) y consultas.	x	x	Sin información	Alto	x	Sin información	Poca documentación	Sin información	http://www.innoventsolutions.com/comparison-matrix.html https://es.wikipedia.org/wiki/Business_Intelligence_and_Reporting_Tools	3.25