

Diseño de una Bodega de Datos en ambientes De Inteligencia De Negocios para el Análisis de Indicadores claves De enfermedades cardiacas.

Design of a Data Warehouse in Intelligence environments of Business for the Analysis of Key Indicators of Heart Disease.

T. Hernández *, E. Méndez*, D. Freiler & J. Freiler*

{Thalía. Hernández, Elvia. Méndez, Deyler. freile, Johnner. freile} @unisimon.edu.co

*Estudiante de Ingeniería industrial ** Juan Carlos Calabria Sarmiento **

Universidad Simón Bolívar, Barranquilla-Colombia.

Resumen | El objetivo de este artículo es presentar un análisis del número de riesgo cardiovascular al que padecen las personas teniendo en cuenta una base de dato "heart_2020_cleaned", en este sentido, este artículo propone la implementación de un análisis y medición de los datos, diseñando una bodega de datos, hacer un proceso de ETL (Extracción, Transformación y Carga) utilizando DataWereHouse para el almacenamiento de los datos y realizar análisis por medio de herramientas de la inteligencia de negocio, y con ayuda de las normas ISO 2512se revisará la calidad de los datos para llegar a una toma de decisión y lograr nuestro objetivo.

Palabras clave: | *Calidad de datos, DataWereHouse, enfermedades cardiacas, riesgos, base de Datos, inteligencia de Negocio, ETL (Extracción, Transformación y Carga.)*

Abstract | The objective of this article is to present an analysis of the number of cardiovascular risk that people suffer from, taking into account a "heart_2020_cleaned" database, in this sense, this article proposes the implementation of an analysis and measurement of the data, designing a data warehouse, carry out an ETL process (Extraction, Transformation and Loading) using DataWereHouse for data storage and analysis using business intelligence tools, and with the help of ISO 2512 standards, the quality of data will be reviewed. data to reach a decision making and achieve our goal.

Keywords: | *Data quality, Datawarehouse, heart disease, risks, Database, Business intelligence, ETL (Extraction, Transformation and Loading.).*

I. INTRODUCCIÓN

La demanda del crecimiento de los datos en estos tiempos ha traído controversia por muchos usuarios, pacientes en las clínicas y hospitales de Colombia, debido a que el volumen exagerado de datos genera un caos, de esta manera muchas organizaciones y empresas se ven obligados acudir a los administradores de sistemas de información actuales que desarrollan estrategias de análisis dirigida a la toma de decisiones que benefician a las organizaciones o empresas que lo requieran. [1]

En cualquier sector de clínicas u hospitales ya sea público o privado, las personas demandan entender y comprender su resultado que se les entregan cuando tienden a padecer enfermedades cardiacas, sin embargo, un pilar importante en estos pacientes son los desafíos que afrontan, cualquier persona de edad o sexo, Las enfermedades cardiovasculares son la principal causa de discapacidad y de muerte prematura en todo el mundo, y contribuyen sustancialmente al aumento de los costos

de la atención de salud. La lesión anatomopatológica fundamental es la aterosclerosis, que se presenta con el transcurso de los años y suele estar avanzada cuando aparecen los síntomas, en general en la

madurez. A menudo se producen repentinamente episodios de infartos y con frecuencia son mortales antes de que pueda prestarse atención médica. Se ha demostrado que la modificación de los factores de riesgo reduce la mortalidad y la morbilidad en personas con enfermedades cardiovasculares, diagnosticadas o no. [2]

En el campo de la inteligencia de negocios una de las soluciones más utilizadas son las estructuras OLAP, cuyo propósito es agilizar la consulta de grandes volúmenes de datos; Normalmente se utilizan estructuras multidimensionales o cubos OLAP, que contienen sistemas transaccionales para la generación de informes inteligentes en diversas áreas. [3]

En este artículo se mostrará la aplicación del proyecto de forma detallada las herramientas tecnológicas que se utilizarán y así mismo, el proceso de estructuración e implementación, aunque muchas de las investigaciones referentes a la enfermedad cardiovascular están orientadas a la utilización de I.A. y nuestro proyecto es cambiar la metodología utilizando la herramienta de inteligencia de negocios para comprender los datos mediante un Dashboard y de tal manera tomar una buena decisión.

II. ESTADOS DEL ARTE

El origen de las Plataformas de Business Intelligence basada en análisis Multidimensional tiene como función monitorear el comportamiento de los datos. El cual tiene como definición emprender estrategias, datos, aplicaciones, productos, procesos y todo aquel recurso digital que otorga una capacidad de recopilación, almacenamiento y análisis de toda la información que a diario genera una empresa con el objetivo de optimizar su rendimiento.

De Esta manera, este ayudara tener una amplia visión y conocimiento de lo que sucede de manera interna, por lo que cuando se aprovecha toda la información que arroja cada área y actividad empresarial, es posible mejorar la toma de decisiones, comprender el funcionamiento de la empresa e incluso, predecir posibles acontecimientos futuros. [4]

el business intelligence predomina desde épocas antiguas donde esto se hacía de forma manual, por lo que las empresas tenían que comprar mobiliario grande para tener un espacio donde almacenar todos los documentos. La invención de la computadora cambió ese panorama porque ya era posible tener esa información.

A finales de los años 60, se introdujo un análisis de datos empleando bases de datos que son herramientas que resultaron útiles al permitir la administración de la información y esto les ayudaba a los

directivos de la empresa a tomar decisiones pertinentes. Pero, para obtener informes confiables y de calidad, era necesario contar con una persona capacitada.

El estudio de la base de datos en inteligencia de negocios es importante porque proporciona tanto a las organizaciones como a las empresas, la funcionalidad que este aplica de manera general en todas las áreas que componen una empresa para obtener los datos relevantes de cada una y analizar su potencial para mejorar el rendimiento global.

Una base de datos multidimensional es un almacén de datos que soporta básicamente la ejecución eficiente de consultas complejas de decisión empresarial. Cada respuesta se puede mejorar significativamente almacenando un conjunto apropiado de vistas de Apariencia de los datos. Estas vistas se seleccionan a partir de las multidimensional cuyos elementos representan el espacio de solución del problema. Las técnicas discutidas en este documento pueden implementarse utilizando la tecnología de bases de datos relacionales. El enfoque también es capaz de explotar el procesamiento de consultas multidimensionales, técnicas como la agregación, esto produce una solución práctica con bajo coste computacional, En este artículo. Técnicas proponemos bases de datos multidimensionales. Las dimensiones, las jerarquías de dimensión y los cubos se introducen formalmente, además; ofrecemos un mapeo del modelo multidimensional al modelo relacional ya las matrices multidimensionales. Los resultados muestran una velocidad progresiva de inserción de datos, capacidad de decisión y reducción de costos, así como una comparación con la inserción tradicional de datos. [5]

III. HERRAMIENTAS DE EXTRACCIÓN-TRANSFORMACIÓN-CARGA (ETL)

A. Estructura de un ETL

Para ilustrar las características básicas de un modelo de herramientas de extracción, transformación y carga, consideraremos la información pertinente de la base de datos de las enfermedades cardiacas, el cual ETL hace referencia a piezas de software responsables de la extracción de datos de varias fuentes, su limpieza, personalización, reformato, integración e inserción en un almacén de datos. La construcción del proceso ETL es potencialmente una de las tareas más importantes de la construcción de un almacén; es complejo, requiere mucho tiempo y consume la mayor parte de los

esfuerzos, costos y recursos de implementación del proyecto de almacenamiento de datos.

La construcción de un almacén de datos requiere enfocar de cerca en comprender tres áreas principales: el área de origen, el área de destino y el área de mapeo (procesos ETL). [6]

B. Bodega de datos.

Se define bodega de datos como la integración de datos consolidados, almacenados en un dispositivo de memoria no volátil, proveniente de múltiples y posiblemente diferentes fuentes de datos. Con el propósito del análisis y a partir de este tomar decisiones en función de mejorar la gestión del negocio [7]. Es decir, que el objetivo principal de la data warehousing es construir un conjunto integrado de bases de datos, con orientación temática donde cada unidad de datos es relevante en algún momento del tiempo [8]

C. Número de datos de pacientes que padecen la enfermedad cardiovascular.

A medida que las enfermedades cardiovasculares van aumentando también aumenta el número de pacientes y por esta razón también aumenta el número de datos de los pacientes. Se realizó una previa investigación para conocer las posibles causas de las enfermedades para tener un promedio ponderado de los números de pacientes que padecen esta enfermedad, se encontró que algunos factores asociados para la hipertensión arterial pueden ser de la herencia, el peso, la raza, la edad, la genética y otras enfermedades asociadas. A partir de estos factores surgen las complicaciones cardiovasculares, entre las más comunes están: hipertensión arterial, enfermedad arterial coronaria, enfermedad valvular cardíaca, accidente cerebrovascular (trombosis o derrame cerebral), fiebre reumática o enfermedad cardíaca reumática. teniendo en cuenta lo anterior la problemática de las enfermedades cardiovasculares ha sido considerado un problema de salud pública a nivel mundial, la Organización Mundial de la Salud, especifica que las enfermedades cardiovasculares causan 17.5% millones de muertes en el mundo cada año en Colombia también se ha considerado que las enfermedades cardiovasculares por lo general están relacionadas con la “hipertensión arterial es el producto de muchos años de vida con hábitos y costumbres inadecuadas, hábitos como el comer salado, consumir en exceso grasas, licor o tinto, ser sedentario y fumar, el cual son considerados estilos de vida poco saludables, y hacen que la tensión arterial vaya aumentando sin

darse cuenta, generalmente, la hipertensión arterial no produce ningún síntoma y se le conoce como el asesino silencioso" [9] .

D. ¿Qué significa BI?

La definición del proceso de BI hace parte del Business intelligence, BI o inteligencia empresarial son tres términos que se emplean para referirse a la implementación de estrategias y herramientas que utilizan para sacar el máximo provecho de cada dato que se produce como consecuencia de las diversas actividades de una empresa con el objetivo de tener un impacto trascendental en su toma de decisiones.

Según la página de climatic, los negocios anteriormente usaban este concepto para explicar y comprender su rendimiento. Sin embargo, con el tiempo y la llegada de las nuevas tecnologías su definición se fue perfeccionando y perfilando hacia el uso de herramientas que ayudan a maximizar los resultados de las empresas, por eso en plena era digital, se dice que los datos son “oro puro”. [10]

IV. METODOLOGÍA DE LA INVESTIGACIÓN

La metodología de construcción de bodegas de datos propuesta se centra en una estructura tecnológica diseñada para que la organización disponga, en forma integrada y estandarizada, de la información correspondiente a la operación de la empresa, así como, proporciona a los usuarios, que tienen a su cargo la toma de decisiones, las herramientas adecuadas, para que, a través de consultas rápidas, ellos mismos accedan la información requerida [11]. En este orden de ideas la investigación está basada en el estudio de unos datos reales y enfocada en métodos de inteligencia de negocio, con el objetivo de lograr resultados capaces de esclarecer datos con el fin de tomar decisiones acertadas en los pacientes con enfermedades cardíacas.

El objetivo que encabeza esta propuesta se desarrollara mediante la evaluación de una base de datos, que será desglosada durante métodos de estadísticas productiva con el fin de procesar la información requerida, a continuación de lo anterior la información procesada será estructurada con un método llamado Data Warehouse que es un tipo base de datos orientada a constituir información de una o más fuentes distintas, para luego procesarla, de esta misma forma se le realizara un estudio para verificar la calidad de los datos para una toma de decisión. como consecuente de lo anterior tendremos un objetivo concreto, que será presentado como proyecto final y

realizado con una metodología de lenguaje que se caracteriza por ser entendible para todos los usuarios.

La metodología planteada para la elaboración del proyecto consta de 3 fases:

Fase 1. Planeación.

En esta primera fase se identificarán las fuentes de origen de datos y se definirán los presupuestos y el cronograma.

Figura 1: Sistema desarrollado bajo Data Warehouse.

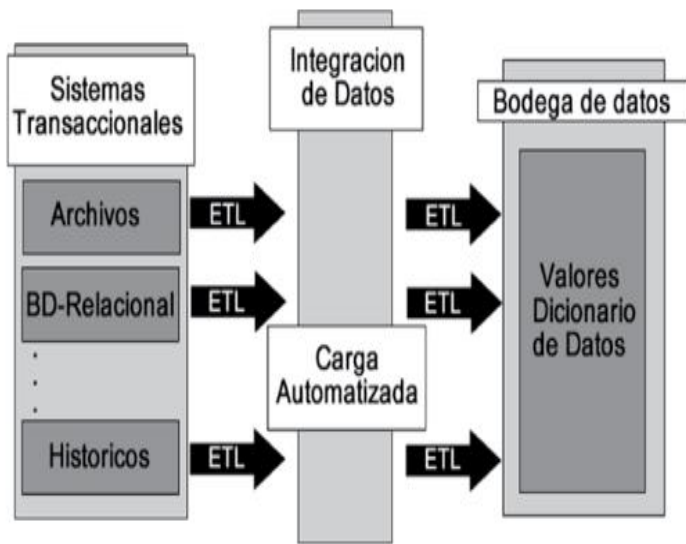
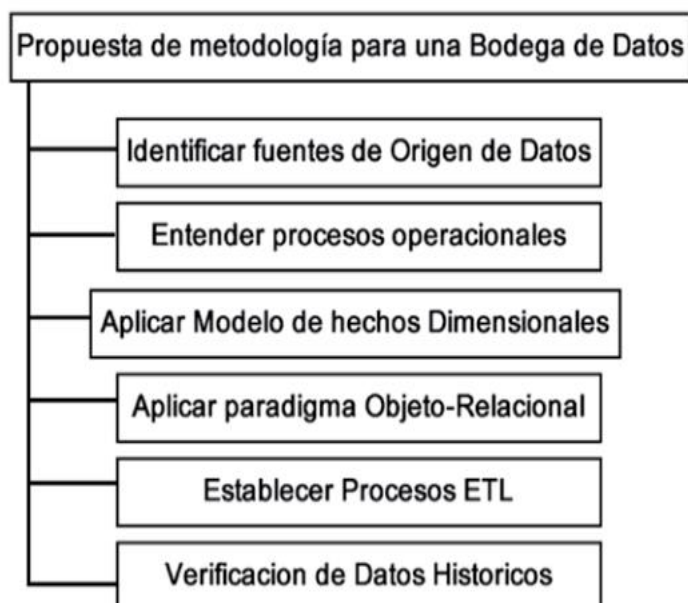


Figura 2: Propuesta de metodología para una Bodega de Datos.



En primera instancia, el objetivo fundamental es la identificación formal de las distintas fuentes de datos que servirán de origen para la alimentación y el suministro de la bodega. estas fuentes pueden ser de estructuras tan complejas como motores avanzados de administración de bases de datos, así como sistemas obsoletos y en cierto sentido reemplazados, limitados y de menor eficiencia como serian sistemas de archivos almacenados. Cada organización deberá ofrecer la información necesaria sobre los distintos sistemas de almacenamiento que hacen uso concerniente de las operaciones primordiales de sus procesos.

Para cuestión de comprensión del presente artículo se han identificado tres (3) fuentes de datos para el desarrollo de este caso guía citado, son las siguientes:

1. Access: es un programa, utilizado en los sistemas operativos Microsoft Windows, para la gestión de bases de datos creado y modificado por Microsoft y orientado a ser usado en entornos personales o en pequeñas organizaciones [12] por medio de sistemas de archivos.

2. PostgreSQL: es un sistema de gestión de bases de datos que trabaja bajo el paradigma Objeto-Relacional (ORDBMS) basado en el proyecto POSTGRES, de la universidad de Berkeley [13]

3. MySQL: es un sistema de gestión de bases de datos relacional, licenciado bajo la GPL de la GNU. Su diseño multihilo le permite soportar una gran carga de forma muy eficiente [13]

Fase 2. Desarrollo.

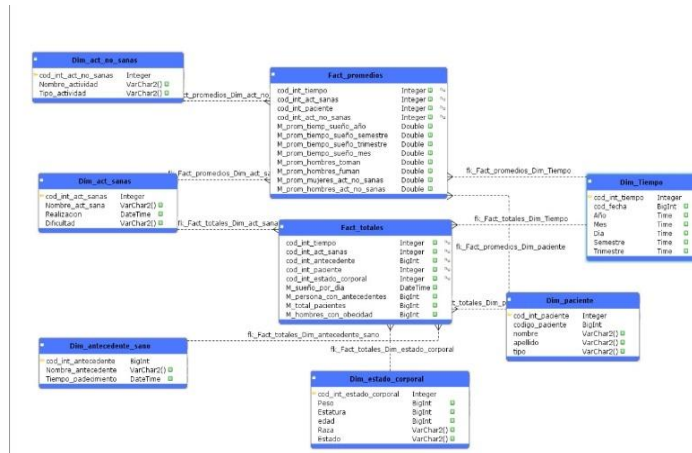
Se creará una base de datos, donde se importarán los datos recopilados, para luego implementar el proceso ETL, para la reconstrucción de los datos, y para que estos puedan ser almacenados en un DataWareHouse. Además de construir una plataforma con una interfaz gráfica visualmente agradable, donde se plasmarán los datos de la DataWareHouse.

Para la implementación de la fase 2 es necesario comprender y analizar algunos conceptos técnicos y se ha hecho necesario la utilización de dos orígenes de distintas procedencia y características incluyendo un limitado sistema de archivos, tales como:

A. Aplicar Modelo de hechos Dimensionales (Diagrama Estrella)

En este paso la importancia radica en construir a raíz de las bases de datos operacionales identificadas y comprendidas en los pasos anteriores de nuestra bodega de datos, este proceso de construcción presenta algunos altibajos que nos podrían perder más tiempo del dispuesto y podría ser un paso en falso si no se escoge un modelo adecuado para el diseño de nuestra data warehouse. El modelo de diseño escogido para la estructura de la bodega se denomina modelo de hechos dimensionales, a partir del cual implementamos la extracción y poblamiento de datos provenientes de las fuentes antes mencionadas.

Figura 3: Modelo Relacional de las fuentes de origen.



El uso del modelo de hechos dimensionales surge a partir del análisis de los datos en un tiempo finito, que ha traído consigo estudios sobre la mejor forma de almacenar y representar estos datos para que puedan ser consultados de una forma más rápida. El uso del modelo multidimensional es una de las aproximaciones más acertadas y seguidas por los especialistas en estos días [1]. Este se basa en el estudio de Los datos de pacientes con enfermedades cardiacas.

b. el uso y utilización de QlikSense.

QlikSense es una plataforma que permite explorar sus datos libremente, avanzando por su propia ruta de exploración hasta el conocimiento. [14] De este modo también permite democratizar los datos, es decir, nos permite recopilar, procesar y extraer el valor de los datos de una compañía para que estos datos sean accesibles por todo el personal de una empresa. [15]

Fase 3. Análisis.

A través de una herramienta de inteligencia de negocio y con ayuda de un proceso de calidad de datos se desarrolló un análisis con los datos reunidos, donde se plasmarán resultados, toma de decisión acertada y conclusiones sobre el comportamiento de los datos de los pacientes con enfermedades cardiacas. Para la implementación de la calidad de los datos se tomará en cuenta la norma ISO 25012, el cual se realizará una guía para para llegar a uno de nuestros objetivos específicos.

V. RESULTADOS



Figura 4: Aplicación de procesos ETL con Kettle (Spoon).

Fuente: elaboración propia.

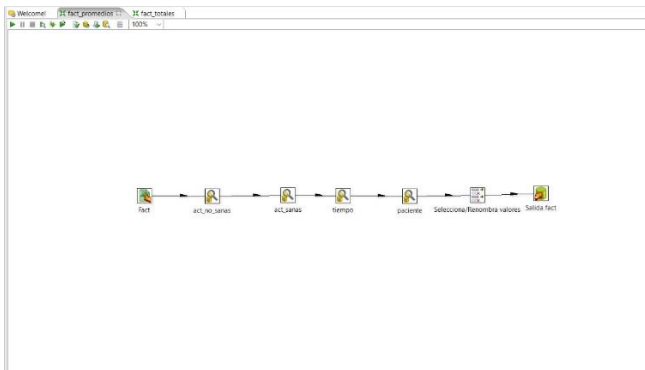


Figura 5: Aplicación de procesos ETL con Kettle (Spoon).

Fuente: elaboración propia.

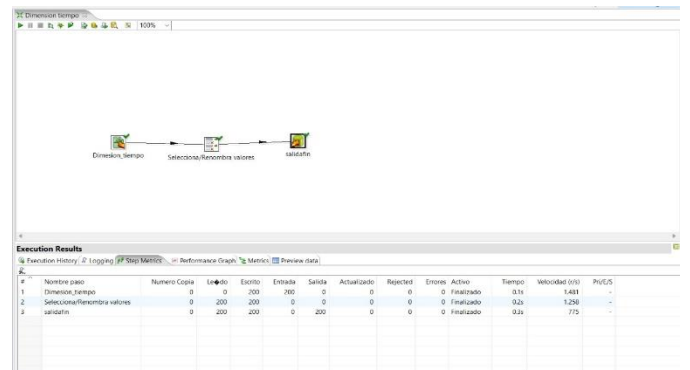


Figura 6: Aplicación de procesos ETL con Kettle (Spoon).

Fuente: elaboración propia.

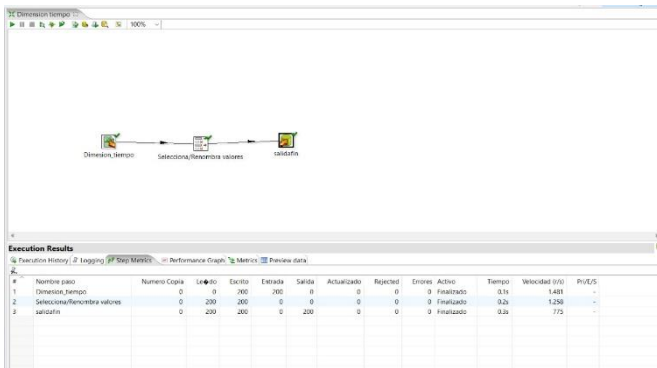


Figura 7: Aplicación de procesos ETL con Kettle (Spoon).

Fuente: elaboración propia.

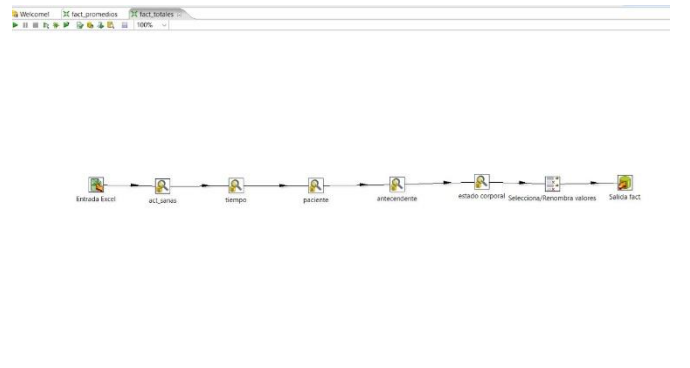


Figura 8: Aplicación de procesos ETL con Kettle (Spoon).

Fuente: elaboración propia.

cod_act_estado	tipo_actividad	dificultad
1	May buena	3
2	May buena	0
3	May buena	29
4	May buena	0
5	May buena	29
6	May buena	15
7	May buena	0
8	May buena	5
9	May buena	0
10	May buena	0
11	May buena	39

Figura 9: resultado de la bodega de datos.

Fuente: elaboración propia.

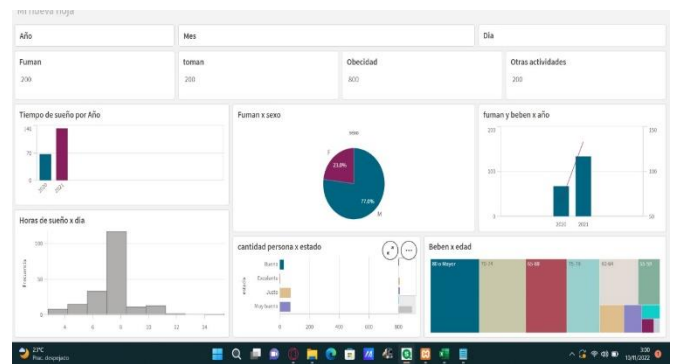


Figura 11: resultado del Dashboard

Fuente: elaboración propia.

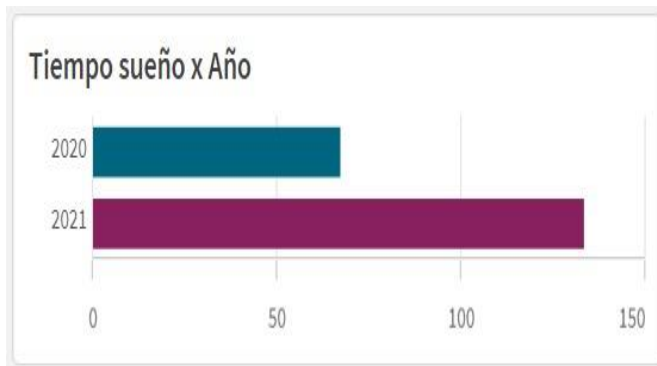


Figura 13: resultado del Dashboard

Fuente: elaboración propia.

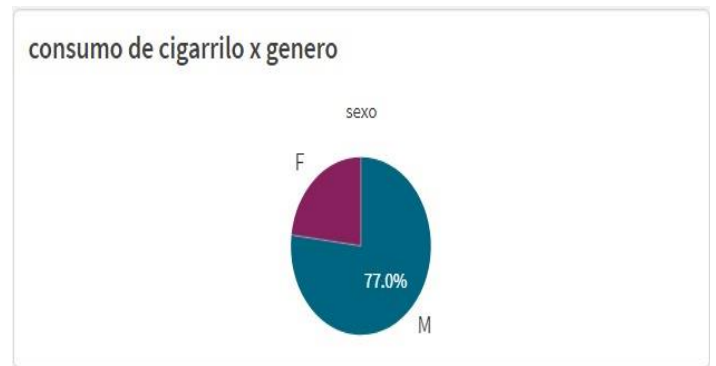


Figura 14: resultado del Dashboard

Fuente: elaboración propia.

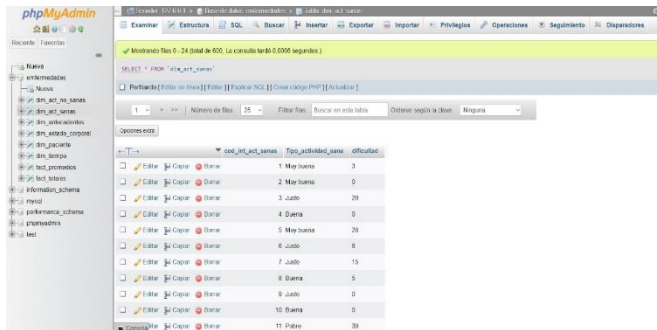


Figura 10: resultado de la bodega de datos.

Fuente: elaboración propia.

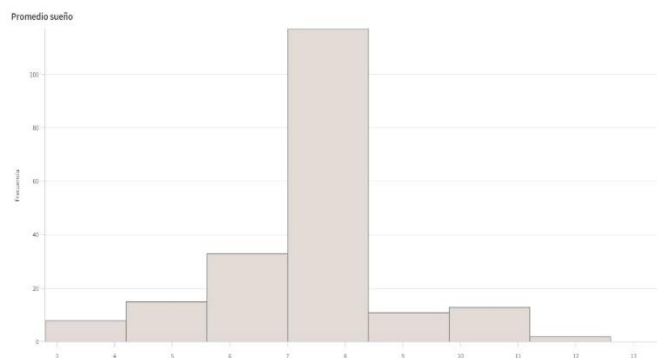


Figura 12: resultado del Dashboard

Fuente: elaboración propia.

VI. ANALISIS DE LOS RESULTADOS.

Los resultados obtenidos de llevar a cabo un proceso de extracción, transformación y carga de datos como está expuesto en pasos anteriores toman validez en cuanto a los atributos o dimensiones que se llevaron a cabo a los pacientes que sufren enfermedades cardiacas.

De este modo, por medio de consultas rápidas se le da una vista más amigable a los resultados, haciendo uso de la herramienta QlikSense, esto con el fin de permitir a las personas encargadas de la toma de decisiones, llevar el control en las organizaciones del hospital o la clínica y poder sacar conclusiones. Así pues, también se pueden fomentar estrategias para que los usuarios puedan entender con más facilidad los resultados arrojados, que encontramos representados por medio de tablas del caso citado que en el artículo se muestra.

VII. CONCLUSIONES.

- La implementación del modelo de hechos dimensionales facilitó la etapa de investigación y diseño; teniendo en cuenta que otros modelos podrían causar una complejidad mucho mayor en la estructura y administración de nuestra bodega de datos.
- Con la herramienta dataset de pentaho logramos llenar y limpiar la bodega de datos en su totalidad, lo cual nos facilitó la construcción del dashboard.
- La implementación de la bodega de datos en la herramienta Qlik Sense nos ayudó a diseñar un

dashboard que nos permite mostrar cada uno de los porcentajes de los pacientes de acuerdo con el año, mes, día, sexo, raza, antecedentes, actividades sanas y no sanas, padecimientos y masa corporal.

Referencias

- [1] J. C. C. SARMIENTO, «Construcción y poblamiento de un datawarehouse basado en el,» *Red de revistas científicas de Acceso Abierto no comercial propiedad de la academia.*, vol. 9 , nº 1, pp. 1-2, 2011.
- [2] W. H. ORGANIZATION, «Prevención de las enfermedades cardiovasculares,» *Organizacion panamericana de la salud* , pp. 6-8, 2007.
- [3] F. A. C. F. M. V. J. A. N. D. S. B. Daniel Sebastián Quintero Martínez, «diseño e implementación de un datawarehouse para la presentación de informes inteligentes sobre medicamentos entregados basados en herramientasmedicamentos entregados basados en herramientasmedicamentos entregados basados en herramientas OLAP,» *universidad simon bolivar* , 2018.
- [4] climatic, «BUSINESS INTELLIGENCE,» 12 agosto 2020. [En línea]. Available: <https://cimatic.com.mx/business-intelligence/#:~:text=Durante%20el%20a%C3%B1o%201989%20surge,apoyo%20basados%20en%20la%20evidencia..> [Último acceso: 3 noviembre 2022].
- [5] s. b. j. g. j. a. j. r. & j. c. l. ardila, «diseño y desarrollo de un dashboard utilizando una bodega de datos integrada con una palicacion web para realizar consultas especializadas,» *universidad simon bolivar* , 2013.
- [6] b. M. A. H. A. H. E. B. Shaker H. Ali El-Sappagh, «A proposed model for data warehouse ETL processes,Journal of King Saud University - Computer and Information Sciences,» *science direct*, vol. 23, pp. 91-104, juli 2011.
- [7] S. D. CHAUDHURI, «Umeshwar.An Overview of Data Warehousing and OLAP Technology,» *SIGMOD Record*,, 2007.
- [8] J. C. CALABRIA, «Construcción y poblamiento de un datawarehouse basado en el paradigma de base de datos,» *prospectiva en linea* , vol. 9, nº 1, pp. 2-4, 2011.
- [9] O. M. D. L. SALUD, 2 agosto 2004. [En línea]. Available: <http://www.geosalud.com/htm.indexz.htm>. [Último acceso: 2022 abril 15].
- [10] CLIMATIC, «Business intelligence, todo lo que debes saber,» 12 agosto 2020. [En línea]. Available: <https://cimatic.com.mx/business-intelligence/#:~:text=Durante%20el%20a%C3%B1o%201989%20surge,apoyo%20basados%20en%20la%20evidencia..> [Último acceso: 2 noviembre 2022].
- [11] I. F. WITTEN, «Eibe. Data Mining: Practical Machine Learning Tools and Techniques (Second Edition).,» *Morgan Series in Data Magnagement Systems*, 2005.
- [12] C. PEREZ, «Data Mining,» *soluciones con Enterprise.Alfaomega*, 2006.
- [13] R. STACKOWIAK, «Oracle Data Warehousing,» 2007.