UNIVERSIDAD POLITÉCNICA SALESIANA
SEDE QUITO

CARRERA:
INGENIERÍA DE SISTEMAS

Trabajo de titulación previo a la obtención del título de:
INGENIERA DE SISTEMAS E INGENIERO DE SISTEMAS

TEMA:
IMPLEMENTACIÓN DE UN BI PARA ANALIZAR LA INCIDENCIA DE LAS ENFERMEDADES Y PROBLEMAS RELACIONADAS CON LA SALUD (CIE-10) EN EL ECUADOR, DENTRO DE LOS ÚLTIMOS 4 AÑOS, UTILIZANDO POSTGRESQL, MATLAB Y PENTAHO.

AUTORES:
MAYRA ISABEL ALVAREZ JIMÉNEZ
VINICIO OBRAYANN CAÑAREJO PUENTE

TUTOR:
ALONSO RENÉ ARÉVALO CAMPOS

Quito, marzo de 2018

CESIÓN DE DERECHOS DE AUTOR

Nosotros MAYRA ISABEL ALVAREZ JIMÉNEZ y VINICIO OBRAYANN CAÑAREJO PUENTE con documento de identificación N° 172080404-4, y con documento de identificación N° 172257644-2, manifestamos que voluntariamente otorgamos a la Universidad Politécnica Salesiana la titularidad sobre los derechos patrimoniales en virtud de que somos autores del trabajo de titulación con el tema: "IMPLEMENTACIÓN DE UN BI PARA ANALIZAR LA INCIDENCIA DE LAS ENFERMEDADES Y PROBLEMAS RELACIONADAS CON LA SALUD (CIE-10) EN EL ECUADOR, DENTRO DE LOS ÚLTIMOS 4 AÑOS, UTILIZANDO POSTGRESQL, MATLAB Y PENTAHO", mismo que ha sido desarrollado para optar por el título de: INGENIERA DE SISTEMAS, INGENIERO DE SISTEMAS en la Universidad Politécnica Salesiana, quedando la Universidad facultada para ejercer plenamente los derechos cedidos anteriormente.

En aplicación a lo determinado en la Ley de Propiedad Intelectual, en nuestra condición de autores nos reservamos los derechos morales de la obra antes citada. En concordancia, suscribo este documento en el momento que hago entrega del trabajo final en formato impreso y digital a la Biblioteca de la Universidad Politécnica Salesiana.

MAYRA ISABEL	VINICIO OBRAYANN
ALVAREZ JIMÉNEZ	CAÑAREJO PUENTE
CI: 172080404-4	CI: 172257644-2

Quito, marzo del 2018

DECLARATORIA DE COAUTORÍA DEL TUTOR

Yo, ALONSO RENÉ ARÉVALO CAMPOS, declaro que bajo mi dirección y asesoría fue desarrollado el proyecto de titulación, con el tema: "IMPLEMENTACIÓN DE UN BI PARA ANALIZAR LA INCIDENCIA DE LAS ENFERMEDADES Y PROBLEMAS RELACIONADAS CON LA SALUD (CIE-10) EN EL ECUADOR, DENTRO DE LOS ÚLTIMOS 4 AÑOS, UTILIZANDO POSTGRESQL, MATLAB Y PENTAHO", realizado por los estudiantes MAYRA ISABEL ALVAREZ JIMÉNEZ y VINICIO OBRAYANN CAÑAREJO PUENTE, obteniendo un producto que cumple con todos los requisitos estipulados por la Universidad Politécnica Salesiana, para ser considerados como trabajo final de titulación.

Quito, marzo del 2018

ALONSO RENÉ ARÉVALO CAMPOS

CI: 140016489-1

Dedicatoria

El presente trabajo se lo dedico a mi familia que día tras día me ha dado su apoyo incondicional, para alcanzar cada uno de los objetivos planteados.

A mi madre mi pilar, mi apoyo incondicional.

Mi hijo mi inspiración, motivación para seguir adelante y no rendirme jamás.

Personas que siempre han estado a mi lado, brindándome su completo apoyo, dándome animo en los peores momentos.

Isabel

* * *

Dedico este proyecto, a mi familia, ya que ellos son los que me han ayudado a llegar a este momento de mi formación profesional.

A mis padres, por demostrarme siempre su cariño y apoyo, sin ellos no sería nada.

A mi hijo Josue, quien es mi logro más grande en esta vida, quien aspiro que llegue a ser el doble de lo que yo alguna vez podre ser, a pesar de su corta edad me llena de momentos felices con sus ocurrencias y travesuras.

A mi esposa Jessyca, porque te amo infinitamente y por siempre estar dispuesta ayudarme en cualquier momento, gracias por ayudarme a alcanzar esta meta.

Vinicio

Agradecimientos

Primero agradezco a Dios por haberme dado la hermosa familia que tengo, además de la fuerza, el coraje y la valentía para culminar las metas planteadas.

A mi madre e hijo por aguantar mi mal genio y falta de interés en ciertos temas, por brindarme su cariño, por estar siempre a mi lado y creer en mí.

A la familia que conforma la Universidad Politécnica Salesiana, docentes y colegas que formaron parte de esta gran aventura, que en todos estos años contribuyeron con su enseñanza y amistad.

Isabel

* * *

Gracias cerebro mío, por no fallarme pese al cansancio y frustración, además gracias a todas las personas que me ayudaron en el trayecto de esta carrera universitaria, familia, amigos, compañeros, profesores, etc., me han ayudado directa e indirectamente en la realización de este proyecto, con sus consejos y asesoramientos, a ustedes les debo la alegría que siento en este momento mientras redacto este texto, espero algún día poder devolverles el favor.

Vinicio

ÍNDICE

INTRODUCCIÓN 1
Objetivo General 4
Objetivos Específicos 5
Capítulo 1 6
Fundamentos Teóricos 6
1.1 Clasificación Internacional de Enfermedades (CIE-10) 6
1.2 Inteligencia de Negocio (BI) 9
1.3 Herramientas que involucran inteligencia de negocio 10
1.3.1 Almacén de datos. 10
1.3.2 Data Mart. 11
1.3.3 Procesamiento Analítico en Línea (OLAP). 11
1.3.4 Minería de datos. 12
1.3.5 Dimensiones. 12
1.3.6 Tabla de hechos. 13
1.4 Proceso de almacenamiento de datos 13
1.4.1 Metodologías para almacenamiento de datos. 13
1.4.2 Características del almacén de datos. 14
1.4.3 Metodología de Kimball. 15
1.4.4 Proceso de Extracción Transformación y Carga (ETL). 20
1.4.5 Arquitectura de almacenes de datos. 22
1.5 Proceso de minería de datos 23
1.5.1 Metodologías para minería de datos. 23
1.5.2 Requerimientos de un sistema de minería de datos. 24
1.5.3 Descripción del Proceso KDD. 25
1.5.4 Técnicas de minería de datos. 26
1.6 Metodología 28
1.6.1 Metodología de desarrollo. 28
1.6.2 Metodología ágil SCRUM. 29
1.6.3 Elementos de la metodología. 31
1.6.4 Implementación de la metodología. 32
Capítulo 2 36
Diseño y Construcción 36
2.1 Capa de Datos 36
2.1.1 Sprint 1. Identificación, validación y análisis de información. 36
2.2 Capa de Información 39
2.2.1 Sprint 2: Creación del modelo operacional y dimensional del DW. 39
2.2.2 Sprint 3. Definición de dimensiones para el almacén de datos. 51
2.2.3 Sprint 4. Análisis de datos para la creación de la tabla de hechos. 59
2.3 Diccionario de Datos 69
2.3.1 Diccionario de la base operacional (EgresoVCA). 69
2.3.2 Diccionario de la base AlmacenVCA. 78
2.3.3 Diccionario de la tabla de hechos. 80
Capítulo 3 81
Análisis 81
3.1 Adquisición de Conocimiento 81
3.1.1 Sprint 5. Definición de áreas de interés para adquirir conocimiento. 81
4.1.2 Sprint 6. Integración de la herramienta Pentaho 84
3.1.3 Sprint 7. Integración de la herramienta Matlab. 89
CONCLUSIONES 93

RECOMENDACIONES 94
LISTA DE REFERENCIAS 95
Anexos 97

ÍNDICE DE TABLAS

Tabla 1. Clasificación de Enfermedades y Lesiones (CIE-10). 6
Tabla 2. Capítulo XII – Enfermedades de la Piel y el Tejido Subcutáneo..... 7
Tabla 3. Comparación de metodologías para implementación de DW..... 13
Tabla 4. Formatos para el Análisis Dimensional. 17
Tabla 5. Comparación de tipos de esquemas dimensionales. 18
Tabla 6. Comparación de metodologías para minería de datos. 24
Tabla 7. Comparación de metodologías ágiles. 28
Tabla 8. Product Backlog del Proyecto..... 32
Tabla 9. Actividades sprint 1. 33
Tabla 10. Actividades sprint 2. 33
Tabla 11. Actividades sprint 3. 33
Tabla 12. Actividades sprint 4. 34
Tabla 13. Sprint con una actividad a realizar..... 34
Tabla 14. Matriz de Involucrados. 34
Tabla 15. Dimensiones y Jerarquías..... 52
Tabla 16. Diccionario de datos de la tabla anio 70
Tabla 17. Diccionario de datos de la tabla canton. 70
Tabla 18. Diccionario de datos de la tabla cap221rx. 71
Tabla 19. Diccionario de datos de la tabla cau221rx 71
Tabla 20. Diccionario de datos de la tabla cau298rx. 71
Tabla 21. Diccionario de datos de la tabla cau_cie10..... 72
Tabla 22. Diccionario de datos de la tabla clase. 72
Tabla 23. Diccionario de datos de la tabla cod_edad..... 72
Tabla 24. Diccionario de datos de la tabla cod_est..... 73
Tabla 25. Diccionario de datos de la tabla entidad. 73
Tabla 26. Diccionario de datos de la tabla mes..... 74
Tabla 27. Diccionario de datos de la tabla parroquia..... 74
Tabla 28. Diccionario de datos de la tabla provincia..... 75
Tabla 29. Diccionario de datos de la tabla sector..... 75
Tabla 30. Diccionario de datos de la tabla sexo..... 75
Tabla 31. Diccionario de datos de la tabla tipo..... 76
Tabla 32. Diccionario de datos de la tabla principal egreso_2011_a..... 76
Tabla 33. Diccionario de datos de la tabla dimensional d_entidad_salud. 78
Tabla 34. Diccionario de datos de la tabla dimensional d_geografia 78
Tabla 35. Diccionario de datos de la tabla dimensional d_sintoma..... 79
Tabla 36: Diccionario de datos de la tabla dimensional d_tiempo 79
Tabla 37. Tabla th_morbilidad..... 80

ÍNDICE DE FIGURAS

Figura 1. Fases de la Metodología de Kimball. 16
Figura 2.Esquema de componentes ETL. 20
Figura 3. Esquema del almacén de datos. 22
Figura 4. Esquema del proceso KDD (Knowledge Discovery in Database). 25
Figura 5. Ubicación de archivos de descarga. 38
Figura 6. Modelo relacional con información del año 2011. 40
Figura 7. Vista de datos año 2011. 42
Figura 8. Vista de variables año 2011. 42
Figura 9. Exportación de información a formato csv. 43
Figura 10. Archivo del año 2011 exportado a formato .csv. 43
Figura 11. Estandarización de formato tipo cadena a tipo Integer. 44
Figura 12. Clasificación de los datos (tabla cod_est). 46
Figura 13. Datos exportados en formato .txt. 46
Figura 14. Tablas secundarias de la base operacional. 47
Figura 15. Primera carga de información de la tabla cod_est. 47
Figura 16. Actualización de datos en cod_est. 48
Figura 17. Actualización de datos faltantes de la tabla cod_est. 48
Figura 18. Tabla principal y claves foráneas del año 2011. 49
Figura 19. Error por información faltante. 50
Figura 20. Esquema de BDD operacional. 50
Figura 21. Número total de registros cargados año 2011. 51
Figura 22. Esquema dimensional tipo estrella. 53
Figura 23. Tablas dimensionales. 53
Figura 24. Script para relacionar las BDD creadas. 54
Figura 25. Script de creación de claves foráneas de tablas dimensionales tiempo y entidad de salud. 54
Figura 26. Script de creación de claves foráneas de tablas dimensionales síntomas y geografía. 55
Figura 27. Secuenciales usadas en ambas bases. 55
Figura 28. Función de carga para la dimensión d_tiempo. 56
Figura 29. Número de datos cargados d_tiempo. 57
Figura 30. Script para contabilizar el número de registros en tablas. 57
Figura 31. Número de registros en la tabla d_entidad_salud 58
Figura 32. Número de registros en d_geografia. 58
Figura 33. Número de registros en d_sintoma. 59
Figura 34. Estructura de la tabla de hechos. 60
Figura 35. Script de mapeo de tablas. 61
Figura 36. Script para obtener el rango de edad entre 0 a 5años. 62
Figura 37. Índices creados en la tabla principal egreso_2011_a 62
Figura 38. Tiempo de ejecución de consultas. 63
Figura 39. Cabecera de la función cargar_hechos(). 64
Figura 40. Variables utilizadas en la función. 64
Figura 41. Cursores utilizados en la función. 65
Figura 42. Apertura de los cursores para cargar los datos. 65
Figura 43. Script para el cálculo de los rangos de edad. 66
Figura 44. Inserción de datos y cierre de cursores. 67
Figura 45. Cierre de cursores, finalización de la función. 67

Figura 46. Datos que forma la tabla de hechos. 68
Figura 47. Comparación de registros por rangos de edad. 69
Figura 48. Datos de Enfermedades por Sexo y Edad. 82
Figura 49. Datos de Enfermedades por Localidad. 82
Figura 50. Datos de Enfermedades por Entidad de Salud. 83
Figura 51. Datos para el método de regresión lineal. 83
Figura 52. Añade query para consultas de información. 84
Figura 53. Query de Enfermedades por entidad de salud. 85
Figura 54.Parámetros de búsqueda de información. 86
Figura 55. Barras de tipo sector de entidad de salud (Año-Sector). 86
Figura 56. Gráficos con porcentajes de enfermedad por sector de salud y año. 87
Figura 57. Barras de casos diagnosticados por clase de entidad de salud. 87
Figura 58. Gráfico con porcentajes de enfermedad por clase de entidad de salud. ... 88
Figura 59. Parte3-Reporte Enfermedad por entidad de salud. 89
Figura 60.Query de la consulta de regresión lineal. 90
Figura 61. Ocurrencia a lo largo de los años, por cada mes. 90
Figura 62. Vector de los meses de un determinado año. 91
Figura 63. Regresión lineal años 2014. 91

Resumen

El proyecto de titulación, planteado por la UPS, nace de la necesidad de tomar decisiones en el área de la salud y el análisis de grandes volúmenes de datos hospitalarios en relación con la incidencia de enfermedades reportadas en las diversas entidades de salud del Ecuador, la metodología implementada para crear el almacén de datos se basa en: 1) Descarga de datos desde el sitio web del INEC. 2) Análisis y migración de datos desde la aplicación estadística SPSS a la base de datos PostgreSQL. 3) Entrevistas a personal del INEC, área del ministerio de salud y médicos, para que aporten con ideas para el análisis de datos y definición de áreas de interés (dimensiones) y tablas de hechos. 4) Creación del proceso ETL (Extracción, Transformación y Carga). 5) Análisis de resultados y generación de vistas materializadas.

Para la creación del almacén de datos se utilizó la metodología de KIMBALL, para la gestión del proyecto se utilizó la metodología ágil SCRUM, que posibilita la estimación del tiempo de desarrollo del mismo, planificando entregas parciales del producto final, priorizando objetivos y adaptando los cambios en el trayecto de su desarrollo.

La plataforma utilizada consta de las siguientes herramientas: SPSS, PostgreSQL, Pentaho y Matlab

El presente trabajo aporta con una herramienta que permite visualizar los síntomas por criterios como son: Ubicación geográfica, rango de edad, síntoma (enfermedad), entidad de salud, por periodo, etc. Permitiendo visualizar la incidencia que tiene cada uno de estos criterios de análisis en la población ecuatoriana.

Abstract

The titling project, proposed by UPS, born from the need to make decisions in the area of health and perform analysis of a large volume of hospital data in connection with the incidence of diseases reported in the various entities of health of Ecuador, the methodology implemented to create the data warehouse is based on the following steps.1) Data download from the web site of the INEC. 2) Analysis and data migration from the statistics application SPSS to PostgreSQL database. 3) Interviews the staff of the INEC, area the ministry of health and doctors, to contribute with ideas for the analysis of the data and the definition of the areas of interest (dimensions) and fact tables. 4) Creation of ETL process (Extract, Transform, and Load). 5) Analysis of the results and generation of materialized views.

For the creation of the data warehouse was used the KIMBALL methodology, for the management of the project was used the agile methodology SCRUM, that allows the estimation of the development time of the same, planning partial deliveries of the final product, prioritizing objectives and adapting changes in the course of its development.

The platform used consists of the following tools: SPSS, PostgreSQL, Pentaho and Matlab.

The present work provides with a tool that allows visualizing the symptoms by criteria as: geographic location, age range, symptom (illness), health entity, period, etc. Allowing visualize the incidence that has each of these criteria of analysis in the Ecuadorian population.

INTRODUCCIÓN

Anualmente a nivel nacional, el INEC recopila gran cantidad de datos hospitalarios en su repositorio, esto a su vez involucra un gran crecimiento del volumen de la base, haciendo que resulte complejo el manejo de información.

La incertidumbre y el costo de tomar malas decisiones de nivel gerencial son factores que han dado origen a las aplicaciones basadas en información y conocimiento, el objetivo de estas es agilizar la consulta de grandes volúmenes de información, para lograr este objetivo utiliza estructuras multidimensionales, también llamadas cubos OLAP, los cuales contienen el resumen de bases de datos de volúmenes grandes. La minería de datos busca detectar patrones de comportamiento y tendencias, las mismas que generan conocimiento y permiten establecer estrategias para gestionar el negocio u organización.

Con el objeto de tener un mejor panorama de las condiciones existentes en áreas básicas de las entidades de salud a nivel nacional, se procesa la información levantada anualmente por el INEC, sobre las diferentes afecciones a la salud que son atendidas en las diferentes entidades de salud públicas o privadas, para implementar un almacén de datos.

La minería de datos se aplica sobre datos históricos de la organización, analizándola y transformándola en información útil para el usuario, convirtiéndose en un Sistema de Soporte a Decisiones (DSS), asegurando la entrega de información confiable, precisa, en formato adecuado y comprensible de modo oportuno.

El análisis de información permite conocer la incidencia de enfermedades diagnosticadas en la población ecuatoriana, de ahí la importancia de identificar variables y determinar sus patrones de comportamiento, relaciones e incidencias que

brinden conocimiento de la frecuencia con que se presenta una determinada enfermedad en la población ecuatoriana para poder tomar medidas pertinentes.

La centralización de información de egresos hospitalarios y aplicación de minería sobre ella permite la extracción de información sensible para la identificación de afecciones de salud Clasificación CIE-10 que afectan a la población, con procesos más eficientes que reducen el tiempo de respuesta.

Entre los beneficios de la creación del almacén de datos encontramos:

Apoyo para análisis, representando un ahorro de recursos para la entidad de salud.

Facilita el trabajo de médicos, seguimiento a pacientes que presentan una misma patología, de modo que permita agrupar datos y obtener conclusiones.

Proceso de consultas rápido en preguntas predeterminadas, aprovechando las dimensiones definidas en la base (tiempo, geográficas, entidad, enfermedad).

Transformación de datos en información, lo que permite adecuarse a cambios futuros, principalmente al crecimiento en cuanto al volumen de datos.

Información confiable, disponibilidad de información que aporte conocimiento de las enfermedades diagnosticadas en los diferentes establecimientos de salud.

Información recopilada en el mismo sitio, permite el acceso y consulta de datos de forma confiable, precisa y oportuna.

Información de fácil manejo para los usuarios incluso para aquellos que no están familiarizados con términos de salud y/o herramientas informáticas.

La fuente de información consultada para la realización de nuestro proyecto es la del banco de datos del INEC, la cual provee de información que se encuentra disponible de forma transparente en su sitio web.

El Instituto Nacional de Estadísticas y Censos es una entidad pública de función y administración autónoma, que brinda servicios especializados en fines estadísticos cuya disponibilidad está respaldada por la “LOTAIP” (cuyas siglas corresponden a la Ley Orgánica de Transparencia y Acceso a la Información Pública), aprobada en asamblea.

El sistema de salud en Ecuador esta segmentado en el sector público comprendido por el “MPS” (cuyas siglas corresponden al Ministerio de Salud Pública), “MIES” (cuyas siglas corresponden al Ministerio de Inclusión Económica y Social), “IESS” (cuyas siglas corresponden al Instituto Ecuatoriano de Seguridad Social), “ISSFA” (cuyas siglas corresponden al Instituto de Seguridad Social de las Fuerzas Armadas) e “ISSPOL” (cuyas siglas corresponden al Instituto de Seguridad Social de la Policía Nacional), además de los servicios de las municipalidades.

Mientras que el sector privado comprende entidades como hospitales, clínicas, consultorios, farmacias, etc., con fines de lucro, además de organizaciones no lucrativas de la sociedad civil y de servicio social.

Datos que son de utilidad a la hora del análisis de información resultado del proceso a realizar, para conocer el sector donde se reporta la mayor cantidad de casos de morbilidad, que permita la predicción de patrones e incidencia de modo que se pueda tomar medidas para su mitigación.

Sin embargo, la forma y el formato en que se almacena la información que provee el INEC, dificulta el acceso a ella, debido a que los datos se encuentran separados por años, se requiere descargar diferentes archivos para recopilar información histórica.

Además, el análisis de información requiere la generación de reportes, implicando un mayor tiempo de respuesta para la entrega de resultados, debido a que hay que acudir a diferentes fuentes para su formulación.

El almacenamiento de volúmenes grandes de datos trae problemas como:

Redundancia mínima, para el respaldo de información se realizan algunas copias de los datos y se almacenan en diferentes ficheros, implicando desperdicio de espacio de almacenamiento con información no relevante.

Consistencia de datos, es necesario eliminar y controlar las redundancias existentes en los datos. “La información debe estar disponible para el usuario de forma inmediata, por ente, si un dato está almacenado una sola vez, su actualización debe realizarse sólo una vez”. (Corrales, 2009).

Accesibilidad a los datos, una base, periódicamente almacena cientos de datos, que al pasar el tiempo se vuelven difíciles de acceder o consultar, si no se precisa de herramientas informáticas adecuadas para ello.

Mantenimiento, depuraciones que consumen mucho tiempo y recursos.

Algoritmos y patrones, su ejecución se dificulta por la cantidad de datos y el tiempo en los procesos.

Consulta y sub-consulta, una consulta simple de una tabla puede llevar demasiado tiempo en su proceso.

Objetivo General

Crear un almacén de datos con información del INEC, que permitan identificar las afecciones de clasificación CIE-10 en el Ecuador, dentro de los últimos 4 años, utilizando PostgreSQL, Matlab y Pentaho.

Objetivos Específicos

Construir un BI con información relevante, que permita analizar la incidencia de las enfermedades y problemas relacionados con la salud (CIE-10) en el Ecuador.

Aplicar minería de datos, para la obtención de patrones de comportamiento, relaciones, tendencias e incidencias que permitan la identificación de las enfermedades de la piel y sus diferentes afecciones, capítulo (L00 – L99).

Brindar información gerencial, derivada del análisis resultante de la minería de datos, aplicada sobre información de afecciones de salud CIE-10, para la generación de reportes.

Capítulo 1
Fundamentos Teóricos

1.1 Clasificación Internacional de Enfermedades (CIE-10)

Por disposición de la OMS, Organización Mundial de la Salud, las enfermedades son categorizadas mediante códigos que determinan el tipo de afección a la que pertenecen, basándose en múltiples síntomas, signos y características presentes en un sector afectado determinado.

La Clasificación Internacional de Enfermedades CIE-10, tiene fines estadísticos que contribuyen al análisis, interpretación y comparación de datos sobre morbilidad y mortalidad de la población en general, en diferentes áreas y épocas del año. (Salud, 2003).

Cada afección es asignada a una categoría que incluye un grupo de enfermedades similares y reciben un código, en la tabla Nº1, se muestra su clasificación.

Tabla 1. Clasificación de Enfermedades y Lesiones (CIE-10).

Cap.	Códigos	Título
I	A00-B99	Ciertas enfermedades infecciosas y parasitarias
II	C00-D48	Neoplasias
III	D50-D89	Enfermedades de la sangre y de los órganos hematopoyéticos y otros trastornos que afectan el mecanismo de la inmunidad
IV	E00-E90	Enfermedades endocrinas, nutricionales y metabólicas
V	F00-F99	Trastornos mentales y del comportamiento
VI	G00-G99	Enfermedades del sistema nervioso
VII	H00-H59	Enfermedades del ojo y sus anexos
VIII	H60-H95	Enfermedades del oído y de la apófisis mastoides
IX	I00-I99	Enfermedades del sistema circulatorio
X	J00-J99	Enfermedades del sistema respiratorio
XI	K00-K93	Enfermedades del aparato digestivo
XII	L00-L99	Enfermedades de la piel y el tejido subcutáneo

XIII	M00-M99	Enfermedades del sistema osteomuscular y del tejido conectivo
XIV	N00-N99	Enfermedades del aparato genitourinario
XV	O00-O99	Embarazo, parto y puerperio
XVI	P00-P96	Ciertas afecciones originadas en el periodo perinatal
XVII	Q00-Q99	Malformaciones congénitas, deformidades y anomalías cromosómicas
XVIII	R00-R99	Síntomas, signos y hallazgos anormales clínicos y de laboratorio, no clasificados en otra parte
XIX	S00-T98	Traumatismos, envenenamientos y algunas otras consecuencias de causa externa
XX	V01-Y98	Causas externas de morbilidad y de mortalidad
XXI	Z00-Z99	Factores que influyen en el estado de salud y contacto con los servicios de salud
XXII	U00-U99	Códigos para situaciones especiales

Nota: Detalle de la clasificación de enfermedades por capítulos.
Fuente: Dirección General de Salud Pública, Calidad e Innovación (Ministerio de Sanidad, 2016).

Tabla 2. Capítulo XII – Enfermedades de la Piel y el Tejido Subcutáneo.

Código	***Título***
	Infecciones de la piel y del tejido subcutáneo
L00	Síndrome estafilococo de la piel escaldada
L01	Impétigo
L02	Absceso cutáneo, furúnculo y carbunco
L03	Celulitis
L04	Linfadenitis aguda
L05	Quiste pilonidal
L08	Otras infecciones locales de la piel y del tejido subcutáneo
L10	Penfigo
L11	Otros trastornos acantoliticos
L12	Penfigoide
L13	Otros trastornos flictenulares
L14	Trastornos flictenulares en enfermedades clasificadas en otra parte
	Dermatitis y eczema
L20	Dermatitis atópica
L21	Dermatitis seborreica
L22	Dermatitis del pañal

L23	Dermatitis alérgica de contacto
L24	Dermatitis de contacto por irritantes
L25	Dermatitis de contacto, forma no especificada
L26	Dermatitis exfoliativa
L27	Dermatitis debida a sustancias ingeridas
L28	Liquen simple crónico y prurigo
L29	Prurito
L30	Otras dermatitis
	Trastornos papuloescamosos
L40	Psoriasis
L41	Parapsoriasis
L42	Pitiriasis rosada
L43	Liquen plano
L44	Otros trastornos papuloescamosos
L45	Trastornos papuloescamosos en enfermedades clasificadas en otra parte
L50	Urticaria
L51	Eritema multiforme
L52	Eritema nudoso
L53	Otras afecciones eritematosas
L54	Eritema en enfermedades clasificadas en otra parte
	Trastornos de la piel y del tejido subcutáneo relacionados con radiación
L55	Quemadura solar
L56	Otros cambios agudos de la piel debidos a radiación ultravioleta
L57	Cambios de la piel debidos a exposición crónica a radiación no ionizante
L58	Radiodermatitis
L59	Otros trastornos de la piel y del tejido subcutáneo relacionados con radiación
	Trastornos de las faneras
L60	Trastornos de las uñas
L62	Trastornos de las uñas en enfermedades clasificadas en otra parte
L63	Alopecia areata
L64	Alopecia androgena
L65	Otra perdida no cicatricial del pelo
L66	Alopecia cicatricial [perdida cicatricial del pelo]
L67	Anormalidades del tallo y del color del pelo
L68	Hipertricosis
L70	Canee

L71	Rosácea
L72	Quiste folicular de la piel y del tejido subcutáneo
L73	Otros trastornos foliculares
L74	Trastornos sudoríparos ecrinos
L75	Trastornos sudoríparos apocrinos
	Otros trastornos de la piel y del tejido subcutáneo
L80	Vitíligo
L81	Otros trastornos de la pigmentación
L82	Queratosis seborreica
L83	Acantosis nigricans
L84	Callos y callosidades
L85	Otros tipos de engrasamiento epidérmico
L86	Queratoderma en enfermedades clasificadas en otra parte
L87	Trastornos de la eliminación transepidermica
L88	Pioderma gangrenoso
L89	Ulcera de decúbito
L90	Trastornos atroficos de la piel
L91	Trastornos hipertróficos de la piel
L92	Trastornos granulomatosos de la piel y del tejido subcutáneo
L93	Lupus erítematoso
L94	Otros trastornos localizados del tejido conjuntivo
L95	Vasculitis limitada a la piel, no clasificada en otra parte
L97	Ulcera de miembro inferior, no clasificada en otra parte
L98	Otros trastornos de la piel y del tejido subcutáneo, no clasificados en otra parte
L99	Otros trastornos de la piel y del tejido subcutáneo en enfermedades clasificadas en otra parte

Nota: Detalle de la clasificación de enfermedades de la piel y tejido subcutáneo.
Fuente: CIE 10° Revisión (Naquista, 2013).

1.2 Inteligencia de Negocio (BI)

Inicialmente Howard Dresden señala la Inteligencia de Negocios como “conceptos y métodos para mejorar las decisiones gerenciales mediante el uso de sistemas de soporte basados en hechos” (2010, pág. 18), dicho concepto ha evolucionado a “conjunto de aplicaciones, metodologías, prácticas y capacidades enfocadas a la

administración de información, para apoyar al proceso de toma de decisiones en beneficio de la organización". (Caralt & Curto, 2010).

En una organización la Inteligencia de Negocios y las Tecnologías de la Información (TI) deben ir a la par, de modo que, en conjunto, hagan posible la recolección, depuración y transformación de datos en información que pueda ser analizada y convertida en conocimiento.

1.3 Herramientas que involucran inteligencia de negocio

La implementación de un BI requiere de herramientas tecnológicas que colaboren en conjunto y proporcionen facilidades para que el manejo de datos sea menos complejo, dichas herramientas tienen tres características comunes:

Proporcionar información confiable, clara y precisa para un mejor control en los procesos de negocio, independientemente del origen de los datos.

Dar soporte al proceso que implica el tomar decisiones, permitiendo determinar la prioridad de objetivos según requerimientos de la organización.

Capa semántica, es fundamental entender el lenguaje del negocio, la información debe ser útil, no trivial y entendible para todas las personas involucradas.

1.3.1 Almacén de datos.

Proceso donde la organización almacena datos que provee información trascendental para su desempeño, para favorecer el análisis y extracción de información de manera precisa y oportuna.

El contenido, distribución y estructura de los datos son destinados a cubrir la demanda de usuarios expertos en el tema. (2005, pág. 1). El objetivo de un almacén es brindar información clave para la empresa, de modo que represente una herramienta para el apoyo de procesos de gestión.

1.3.2 Data Mart.

Tiene un alcance de contenido de información limitado, se utiliza para el seguimiento de situaciones específicas que compete a un departamento parte de la organización.

Es un subconjunto del almacén de datos, su información es destinada al apoyo de toma decisiones administrativas, de un área determinada del negocio.

Sistema de Soporte de Decisiones (DSS). En ingles, Decision Support System, contiene servicios y procesos para la selección, manipulación y análisis información, haciendo transparente la presentación de resultado. Constituye la interfaz entre el usuario y la información que reposa en el almacén de datos.

Un DSS se orienta a proporcionar información oportuna al apoyo de áreas operacionales y administrativas para tomar mejores decisiones en beneficio de la organización. (Ayala, 2016).

1.3.3 Procesamiento Analítico en Línea (OLAP).

En ingles, Online Analytical Processing, su función principal es facilitar al usuario la consulta y acceso a volúmenes grandes de información, utilizando cubos multidimensionales.

Un sistema OLAP permite conocer la historia del negocio, realizar informes complejos de manera sencilla y la interacción del usuario con los datos que contiene el almacén, de modo que pueda especificar entradas para la búsqueda de información.

Características del sistema OLAP:

Optimización de tiempos de respuesta a consultas efectuadas por usuarios del sistema.

Los cubos de OLAP almacenan diferentes niveles de datos con estructuras optimizadas orientadas a los objetivos del negocio, reduciendo el tiempo de análisis de datos y recursos empleados para ello. (Martinez, 2010).

1.3.4 Minería de datos.

Extracción de información importante e implícita, en un inicio desconocida pero potencialmente valiosa, en análisis de relaciones, patrones y tendencias encontradas a partir de volúmenes grandes de datos. (Trujillo, 2013).

En Inteligencia de Negocio la minería de datos, constituye un proceso de transformación de información almacenada en conocimiento, para el apoyo de toma de decisiones gerenciales. La aplicación de minería de datos en campos de la medicina abarca los siguientes ámbitos:

Clínico, ayuda al reconocimiento y diagnóstico de enfermedades que afectan a la población.

Medicina preventiva, hace posible la detección e identificación de enfermedades y síntomas en pacientes que sufren alguna patología.

Gestión hospitalaria, para predicciones en períodos de tiempo, que permitan la optimización de recursos hospitalarios disponibles y la priorización de tratamientos a pacientes con una misma patología. (Zamarrón, García, Calvo, Pichel, & Rodríguez, 2006).

1.3.5 Dimensiones.

Las dimensiones son criterios o áreas de interés con información para presentar al usuario, es decir, contienen la información general de cada registro de la tabla de hechos, por ejemplo: ubicación, tiempo, enfermedades, entidades de salud, etc.

Además, permiten la selección, agrupación e identificación de datos que representan elementos fundamentales en un cubo multidimensional, son vistas para un proceso de negocio determinado.

1.3.6 Tabla de hechos.

Es el centro del almacén de datos, contiene los indicadores que representan los procedimientos que sigue la organización para alcanzar sus objetivos.

Los procesos de negocio están representados por un modelo dimensional, conformado por la tabla de hechos que contiene mediciones numéricas de un evento que no cambian en un futuro, pero puede ser actualizadas, rodeado por dimensiones que contienen características descriptivas asociadas al evento. (Kimball & Ross, 2013).

1.4 Proceso de almacenamiento de datos

1.4.1 Metodologías para almacenamiento de datos.

Existen varias metodologías para la implementación de un almacén de datos sin embargo las más conocidas son las propuestas por Bernabeu Ricardo Darío (Hefesto), Bill Inmon y Ralph Kimball. En la tabla Nº3, se compara las características de las metodologías, para seleccionar la más apropiada en la ejecución del proyecto.

Tabla 3. Comparación de metodologías para implementación de DW.

Bueno/Alto=3; Peq-Med/Medio=2; Complejo/Costoso/Bajo=1; No =0; Todos/ Si=1
Relacional =1; Estrella/Copo de nieve/ Empresarial =2; Dimensional =3

Factor de análisis	Inmon	Hefesto	Kimball
Flexibilidad	Alto	Alto	Alto
Adaptabilidad tecnológica	Si	Si	Si
Afinidad de desarrollo	Medio	Medio	Medio
Comunicación con el usuario	Alto	Alto	Alto

Tamaño del proyecto	Complejo	Todo	Peq-Med
Tiempo de análisis y diseño	Medio	Medio	Costoso
Tiempo en construcción	Medio	Medio	Bueno
Perspectiva	Relacional	Estrella/Copo	Estrella
Rápido acceso a reportes	Baja	Medio	Alto
Enfoque utilizado	Empresarial	Empresarial	Dimensional
Tipo de Perspectiva	Top-Down	Top-Down	Bottom-up
Total	**18**	**20**	**23**

Nota: Valor asignado a las diferentes características de la metodología, para seleccionar la más idónea para el proyecto en cuestión.

Como conclusión obtenemos que la metodología más compatible con las características de nuestro proyecto, es la propuesta por Ralph Kimball, ya que es un proyecto no muy grande que busca asegurar la usabilidad de los usuarios, con herramientas de fácil accesibilidad y rápido desarrollo.

1.4.2 Características del almacén de datos.

Según Kimball, un almacén de datos, está planteado para apoyar al proceso de toma de decisiones gerenciales, cuyas características son: datos alineados al negocio, cambiantes en el trascurso del tiempo, integrados y no volátiles.

Alineado al negocio, se enfocado en datos relevantes relacionados con áreas específicas del negocio.

Cambiante en el tiempo, los datos son recopilados cada cierto tiempo acorde a las necesidades de la organización, puede ser de forma diaria, mensual, trimestral, etc.

Un almacén contiene gran cantidad de información histórica, sobre actividades que realiza la organización, que son de importancia en el alcance de sus metas.

Integrados, los datos son almacenados en un único repositorio, si estos proceden de diferentes fuentes es necesario estandarizar su formato y significado.

No volátil, una vez añadidos los datos al almacén estos no cambian, por ente una misma consulta tendrá siempre un mismo resultado. (Kimball & Ross, 2013).

Generalmente un almacén de datos contiene los siguientes elementos:

Datos Antiguos, importante en el proceso inicial de recopilación de datos para el almacén, son los datos históricos de la organización.

Datos de Producción, datos que contiene la base operacional. Se selecciona datos relevantes que permitan conocer el estado de la organización, acorde al giro de negocio. Los datos pueden ser actualizados por aplicaciones OLTP.

Extractores de Datos, determina cada que tiempo se debe realizar la actualización de datos, además, del origen y destino de datos a copiar.

Herramientas de Consultas, interfaz que permite obtener información al usuario mediante la interacción con los datos del almacén.

1.4.3 Metodología de Kimball.

La metodología se basa en lo que Ralph Kimball denomina Ciclo de Vida Dimensional del Negocio, basado en cuatro principios básicos:

Centrarse en el negocio, identificación de requerimientos y el valor asociado que cada uno aporta al negocio, para determinar el orden de desarrollo del proyecto.

Infraestructura de información adecuada, diseño de la base con información relevante que refleje las necesidades del negocio.

Entregas en incrementos, se debe realizar la entrega de incrementos significativos del almacén de datos con plazos no mayores a un año.

Solución completa, el producto final debe satisfacer las necesidades de la organización, incluyendo herramientas de consulta y análisis de los datos. (Kimball & Ross, 2013).

1.4.3.1 *Fases de la Metodología Kimball.*

La metodología de Ralph Kimball se orienta al diseño del almacén de datos, resumido en cinco fases de enfoque iterativo que generan un conjunto de estructuras, informes y datos relacionados.

La planificación del proyecto tiene gran influencia sobre los requerimientos del negocio los mismos que constituyen el soporte inicial de la metodología, de donde parten tres caminos enfocados a áreas diferentes.

Tecnología, engloba tareas relacionadas con un software específico.

Datos, tratamiento de datos a utilizar.

Aplicaciones BI, tareas dedicadas a usuarios finales.

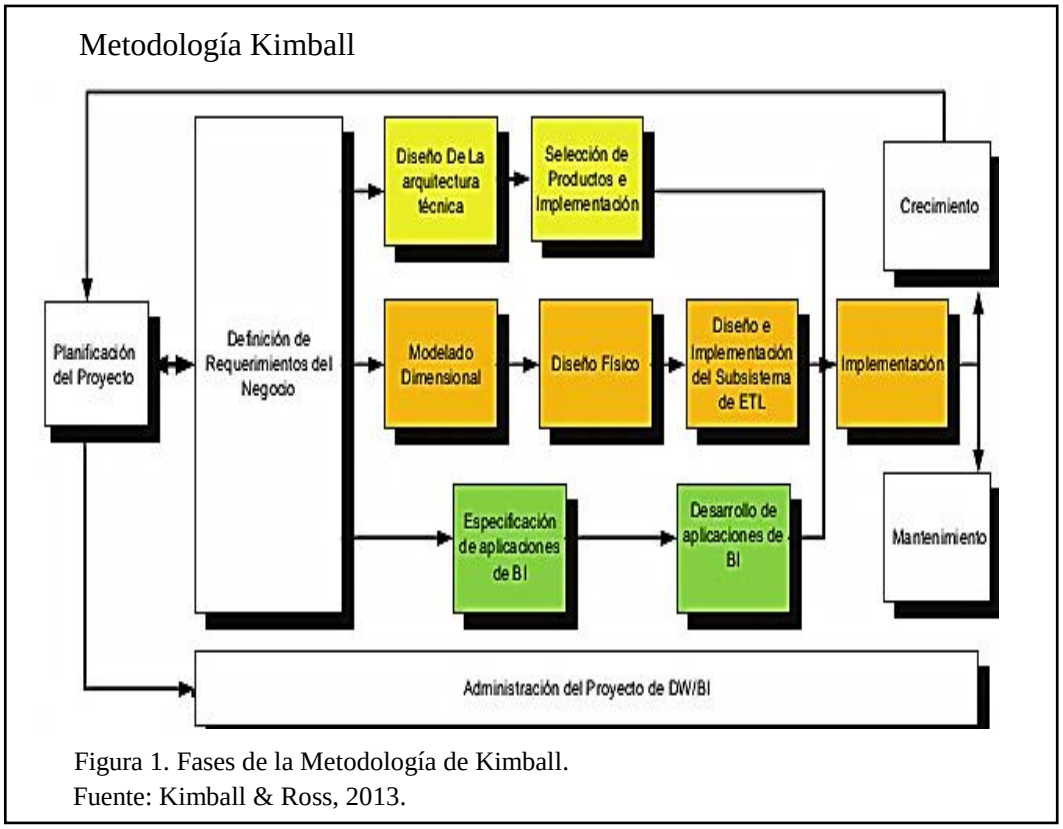


Figura 1. Fases de la Metodología de Kimball.
Fuente: Kimball & Ross, 2013.

Fase 1. Planeamiento.

Visión del Producto, definir el alcance del proyecto, los medios para ejecutarlo y el análisis de posibles riesgos en las etapas de desarrollo.

Equipo de Trabajo, cada integrante debe tener claro el rol y responsabilidades a cumplir, para evitar futuros contratiempos.

Cronograma, asignación de actividades a miembros del equipo, cada actividad tiene un tiempo de ejecución.

Fase 2. Requerimientos y análisis del negocio.

Requerimientos, las medidas y dimensiones que se aplican en el desarrollo del almacén se enfocan en los objetivos del negocio.

Análisis Dimensional, evalúan los recursos de los datos, determinando las necesidades de información encontradas en las fuentes de consulta.

Tabla 4. Formatos para el Análisis Dimensional.

Formato	**Descripción**
Gestión	Define los indicadores y las medidas respectivas.
Análisis	Define medidas y dimensiones.
Dimensiones	Jerarquía de datos, análisis de lo genérico a lo específico.
Medidas dimensionales	Define las medidas dimensiones a ser analizadas.

Nota: Hojas de formato para recolección y análisis de requerimientos del negocio.

Fase 3. Diseño dimensional.

En esta fase se define el esquema dimensional a utilizar en la creación del almacén de datos, constituye una técnica de diseño creada con el fin de brindar apoyo a las consultas realizadas al almacén por parte de los usuarios, basándose en hechos (valores numéricos) y dimensiones (jerarquías y contextos).

Existen varios esquemas dimensionales, sin embargo, los más utilizados son el esquema estrella y copo de nieve, cuyas principales características se describen en la tabla Nº5.

Tabla 5. Comparación de tipos de esquemas dimensionales.

Esquema Estrella	**Esquema Copo de Nieve**

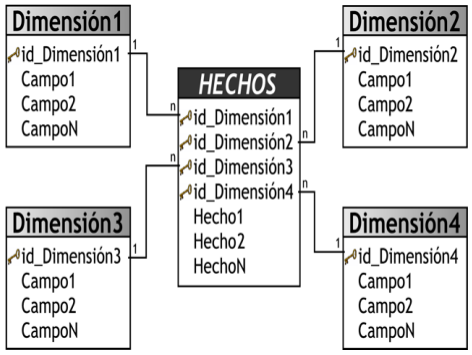
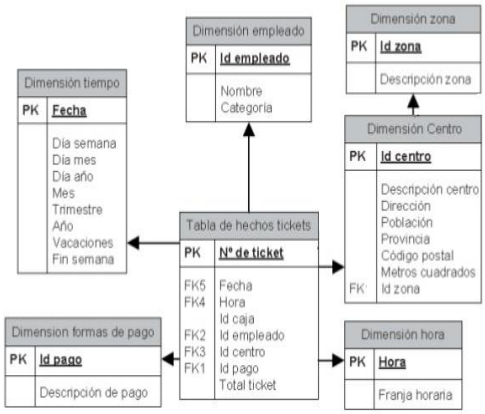

	

Tabla de hechos con los datos para el análisis, compuesta por las claves primarias de las dimensiones Tablas dimensionales que asemejan una estrella, con claves primarias simples. La extracción de datos es más rápida, debido a su simplicidad.	La tabla de hechos no es la única relacionada con otras tablas. Dimensiones relacionadas entre sí, sin tener relación directa con la tabla de hechos. La extracción de datos es más compleja, debido a las múltiples tablas y relaciones entre ellas.

Nota: Comparación de esquemas dimensionales y sus principales características.

Seleccionamos el esquema estrella ya que disponemos de una tabla de hechos central, alimentada por varias tablas dimensionales, además los script utilizados no son complejos y el número de tablas es pequeño, permitiendo la elaboración de reportes (cubos), disminuyendo el tiempo de respuesta en consultas.

Se implementan los datos resultantes de la fase 2, en las tablas dimensionales y las medidas en tablas de hecho. Donde es necesario:

Definir el nivel dimensional. Revisión de datos que pueden asociarse. Diseñar dimensiones y medidas, cada dimensión hallada debe ser convertida en tabla dimensional.

Creación de la tabla de hechos. Información importante para la organización, cada dimensión conforma una medida de esta tabla.

Documentación del modelo. Las dimensiones deben unirse con su tabla hecho correspondiente, de acuerdo a los análisis dimensionales.

Diseño Prototipo. Los prototipos deben tener características como:

Tiempo de respuesta adecuado, el prototipo debe reflejar la capacidad de proporcionar información de forma rápida.

Permitir la interpretación de roles y relaciones, para la visualización de información combinando las dimensiones.

Permitir la navegación a diferente nivel de detalle, el usuario debe poder seleccionar la información que desea consultar.

Permitir funcionalidades adicionales, como ordenamiento y agrupación de datos, filtros, rangos, totales, etc.

Fase 4. Construcción.

ETL, fase de eliminación o corrección de datos en los que se han detectado errores o faltantes.

Migración de registros necesarios a la base de datos que se utilizara en la construcción del Data Mart.

SGBD (Sistemas de Gestión de BDD), conjunto de software que sirva de interfaz entre usuarios del sistema y el almacén de datos.

Fase 5. Despliegue.

Al concretarse la planificación es necesario asegurar:

Resultados favorables en el uso de tecnología, consulta y análisis de datos.

Disponibilidad para capacitar y apoyar a los usuarios del sistema.

Documentación y validación de datos, de modo que el usuario pueda seguir determinados pasos basándose en manuales del sistema.

1.4.4 Proceso de Extracción Transformación y Carga (ETL).

Proceso que prepara los datos para conseguir variables o atributos de importancia para la organización, su objetivo es facilitar la aplicación de minería. El tipo de datos puede modificarse para favorecer el uso de métodos de estandarización de información. (Hernandez, Ramirez, & Ferri, 2014).

Procesos de Extracción, Transformación y Carga

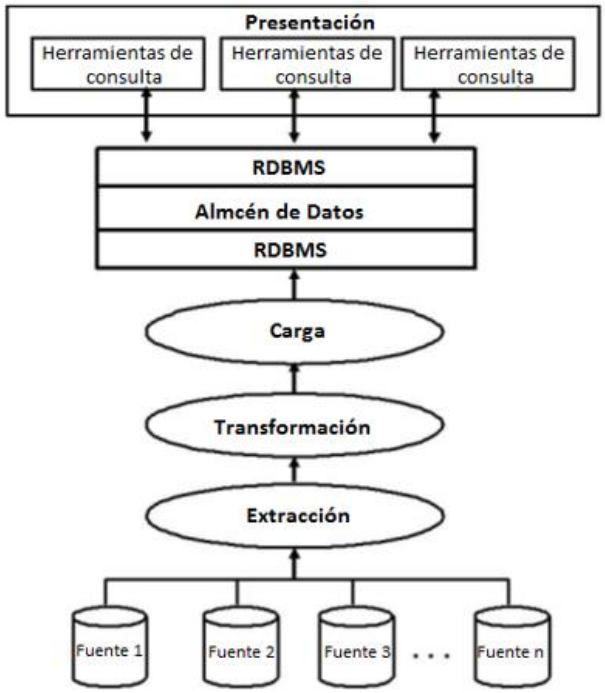


Figura 2.Esquema de componentes ETL.
Fuente: Hernandez.J & Varios, 2014.

Fuentes. Datos recopilados en bases operacionales, proveen los registros del almacén de datos.

Extracción. Extrae datos de fuentes operacionales, que luego de la revisión de sus registros y campos son recopilados en el almacén. No todos los datos de las fuentes son relevantes en el desarrollo del almacén.

Transformación. Prepara los datos para incorporarlos al almacén. Limpieza de datos: Debido a que los datos proceden de diferentes fuentes pueden contener errores y anomalías, que deben corregirse para añadir valor a los datos y establecer como atacar el problema, definiendo reglas para validar de los datos. (López Porrero, 2009). Es necesario:

Análisis de datos, detectar inconsistencias y tipo de errores a eliminar.

Determinar el flujo de trabajo, trasformaciones y reglas de correspondencia, que dependen del número de fuentes de datos.

Flujo inverso de datos limpios, después de eliminar los errores, el dato correcto debe reemplazar al incorrecto en las fuentes originales.

Estructuras internas, la estructura interna de datos operacionales pueden cambiar cuando pasan a ser parte del almacén de datos.

Agregación de datos, para colocar registros operacionales en el almacén, se precisa la creación de funciones de agregación.

Integración de datos, los registros del almacén deben ser correctos, ya que de ellos dependerá el resultado del análisis que se realice.

Integración de formato, las fuentes operacionales suelen estar soportadas por tecnología de diferente fabricante, esto puede provocar que un atributo sea de un tipo en un sistema y de otro tipo en otro sistema.

Integración semántica, cada registro ingresado en el almacén debe tener un mismo significado, que todos los usuarios deben conocer. (Diccionario de datos).

Carga. Recopilación de información extraída de fuentes operacionales para cargarla al almacén, una vez realizadas las transformaciones necesarias, la actualización de los datos debe ser periódica acorde al giro del negocio.

Presentación. Interfaz amigable para la interacción del usuario con información del almacén.

1.4.5 Arquitectura de almacenes de datos.

La arquitectura de un almacén de datos representa la estructura global de los datos, procesos, comunicaciones y presentación al usuario final.

Arquitectura de almacén de datos

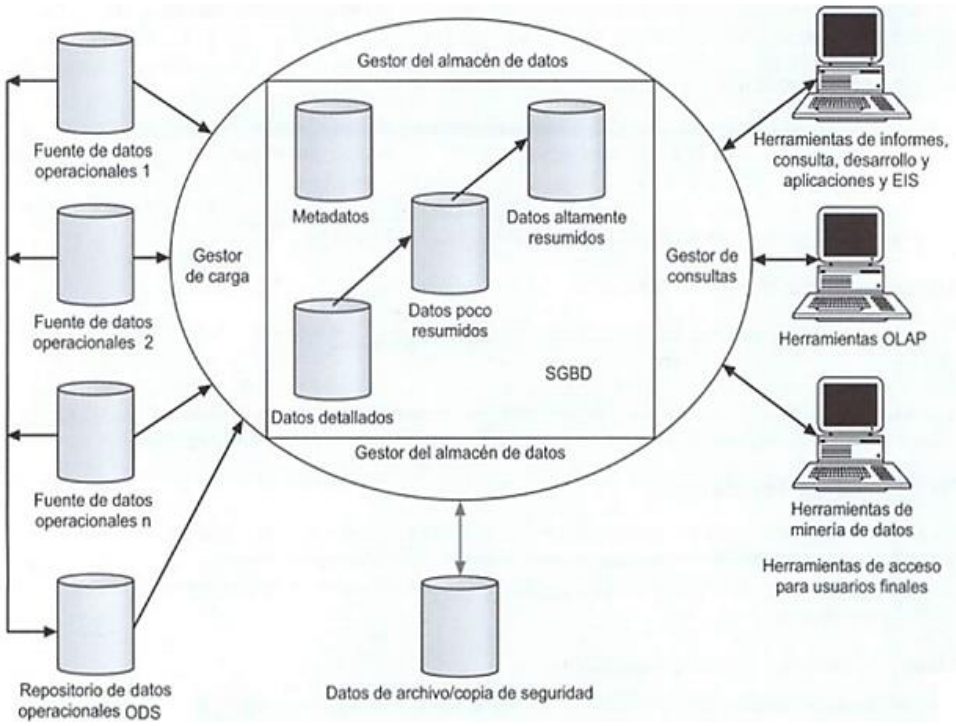


Figura 3. Esquema del almacén de datos.
Fuente: Thomas M. Cannolly & Carolyn E. Begg, 2005.

Nivel Operacional. Datos constituidos por diferentes fuentes operacionales de la organización. Utilizados para el análisis y generación de informes, a partir de estos se definen las dimensiones para la posterior creación de la tabla de hechos.

Almacén de datos. Es el núcleo del sistema, se almacenan los datos actualizados que se utilizan con fines informativos, además contiene copias de datos operacionales y externos que optimizan el acceso a las consultas.

Nivel departamental (Data Mart). Datos generados por cada departamento que integra la organización, ordenados y almacenados de modo que se puedan realizar consultas directas de información, que solicite el usuario acorde a sus roles y permisos.

Nivel individual. Contiene pocos datos, resultantes de la aplicación de procesos de conocimiento y estadísticos, a datos de nivel departamental. Los usuarios asignados a este nivel interactúan con los registros del almacén.

1.5 Proceso de minería de datos

1.5.1 Metodologías para minería de datos.

En el área de minería de datos las metodologías más dominantes son KDD (Knowledge Discovery in Database), SEMMA (Sample, Explore, Modify, Model, Assess) y CRISP-DM (Cross-Industry Standard Process for Data Mining).

Básicamente las metodologías mencionadas comparten la misma naturaleza y objetivo, que es el análisis de datos, de modo que constituya una herramienta de apoyo en procesos gerenciales.

Tabla 6. Comparación de metodologías para minería de datos.

Factor	KDD	CRISP-DM	SEMMA
Enfoque	Usuarios	Empresarial	Técnicos
Obtención de información	Análisis del problema	Análisis del problema	Muestreo de datos
Tareas a realizar	Guía general del trabajo a realizar en cada fase	Tareas y actividades concretas en cada fase del proyecto	Tareas y actividades concretas en cada fase del proyecto
Herramientas comerciales	Multiplataforma	Multiplataforma	Ligada a productos SAS

Nota: Tabla comparativa de características de las metodologías más populares en minería de datos.

El motivo de peso para la selección de la metodología KDD es el enfoque en el usuario, la metodología considera, la asistencia al usuario en la correcta selección de instrumentos y técnicas para aplicar en la organización, como parte de su implementación.

1.5.2 Requerimientos de un sistema de minería de datos.

Las organizaciones generalmente recopilan volúmenes grandes de datos en diferentes formatos, sin embargo, para que estos datos sean útiles, se precisa la identificación de patrones, relaciones y asociaciones entre los datos históricos proporcionados. De aquí, la importancia que el sistema de minería cumpla con ciertas características:

Manejar diferentes tipos de datos.

Eficiencia y escalabilidad, para la selección de algoritmos a utilizar en minería, debe tomarse en cuenta que el ingreso de registros es inevitable.

Utilidad, certeza y expresividad de resultados de la minería, se debe medir la calidad de la información descubierta.

Expresar de diferente forma las respuestas y resultados de la minería, las tareas de minería, pueden ser especificadas por expertos, pero el resultado debe ser sencillo de interpretar por los involucrados.

Interactividad con el conocimiento, se debe permitir que el usuario interactúe de manera flexible con la información que constituye el almacén, acorde al rol y función que desempeñe.

Protección, se debe eludir la mala utilización de información, implementando medidas para la seguridad y privacidad de la misma.

1.5.3 Descripción del Proceso KDD.

Proceso relacionado a la extracción de conocimiento, enfocado en descubrir información potencialmente útil que se encuentra oculta en los repositorios de las bases de datos. (Han & Kamber, 2011).

Fases del Proceso de Descubrimiento de Conocimiento en Bases de Datos

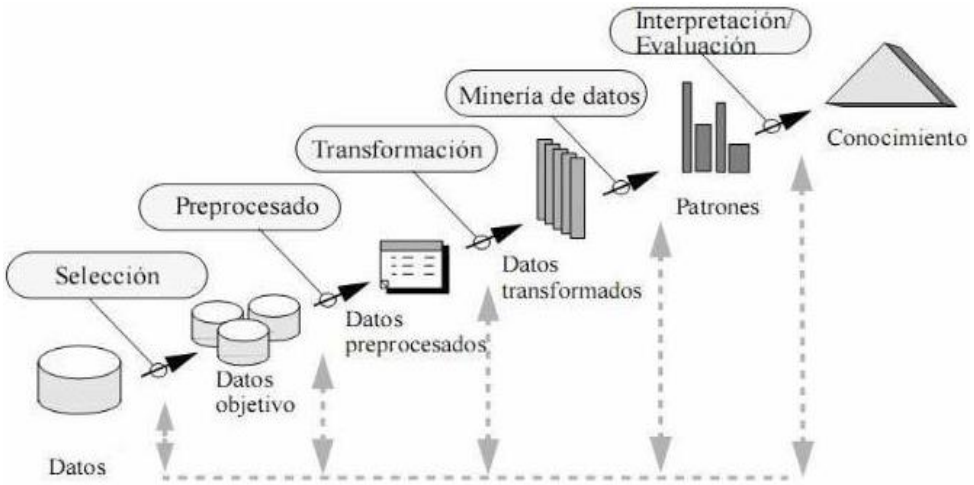


Figura 4. Esquema del proceso KDD (Knowledge Discovery in Database).
Fuente: Fayyad, Piatetsky, & Smyth, 1996.

Selección de datos.

Localización de datos, identificación de registros recopilados en el almacén.

Identificación de datos, requiere la ayuda de personas expertas en los datos a analizar, que estén familiarizadas con los propósitos del negocio.

Procesamiento de Datos.

Depuración de datos, proceso que asegura los propósitos de la minería, se debe identificar y corregir información perdida y limpiar los registros.

Enriquecimiento de datos, proceso en que se añade nuevos atributos a los datos ya existentes, por ejemplo: campos calculados, datos de fuentes externas, etc.

Transformación de datos. Proceso de cambiar la forma o estructura de los datos existentes.

Preparación de un conjunto de casos, elaboración del conjunto inicial de reglas y patrones que sirven de modelo para aplicar minería de datos.

Construcción del modelo de minería de datos.

Selección de algoritmos para minería, que se ajusten a los requerimientos del conjunto de casos, para su respectiva evaluación.

Interpretación y evaluación. Comprobación de conclusiones satisfactorias, de no cumplir las expectativas del cliente, debe revisarse la solución y realizar las respectivas modificaciones.

1.5.4 Técnicas de minería de datos.

Las técnicas de minería de datos son una especie de colador que al introducir los datos produce esa serie de patrones, para el caso del sector de la salud se utilizaremos la técnica predictiva de regresión

1.5.4.1 *Técnica predictiva de regresión.*

Cuando la variable de salida, así como las explicativas son cuantitativas, útiles para predecir la existencia de ciertas características en los datos. Estos modelos se utilizan por ejemplo para la predicción de demanda de insumos de salud en el futuro basándose en hechos pasados.

Además, permite conocer si una variable depende, o no de otras, es decir, expresan la probabilidad que un hecho ocurra en función de ciertas variables consideradas potencialmente influyentes, donde la función de regresión asigna un valor real a cada instancia, el objetivo es minimizar el error entre valores de predicción y los reales.

Regresión Lineal. Es un modelo matemático de ajuste lineal, usado para aproximar la relación entre variables que se adaptan una variedad de situaciones y predecir amplios rangos de fenómenos indiferentemente cual sea el aspecto a analizar.

Usando este modelo, se puede apreciar la relación de la variable dependiente, en nuestro caso los años (2011 - 2014) y sus meses (1-12), a lo cual definimos X, y las variables independientes Y, que vienen a ser el número de ocurrencias de una cierta enfermedad, en una determinada muestra, basado en la fórmula ‘Y = XB’.

1.5.4.2 *Modelo matemático para Minería de Datos.*

Los modelos matemáticos constituyen conjuntos heurísticos y de cálculos que se crean a partir de información, para resolver determinados problemas, para ello, primero debe analizarse los datos provistos en busca de patrones, relaciones o tendencias.

Una vez realizado el análisis respectivo, los resultados son comprobados aplicando la fórmula matemática, en repetido número de iteraciones para determinar parámetros adecuados en la generación de modelos de minería.

Fórmula de regresión lineal. La fórmula 'Y = XB' en la que se basa la regresión lineal, traza los valores reales, por otro lado, los valores ajustados, se los traza con el coeficiente de regresión llamada b1 donde b1 = x/y.

El valor calculado es un ajuste para establecer el mejor coeficiente de determinación, usando mínimos cuadrados, para comparación del ajuste de la recta con el ajuste del coeficiente de regresión. Los mínimos cuadrados representan la recta que mejor se ajusta a un conjunto de datos.

1.6 Metodología

1.6.1 Metodología de desarrollo.

Para el desarrollo del proyecto de software se consideró las metodologías ágiles, ya que aporta con aprendizaje continuo y adaptabilidad acorde a necesidades del entorno, para proyectos no muy bien definidos y alto grado de riesgo.

En la tabla Nº7, se realiza la comparación de las metodologías agiles SCRUM y XP que son las más utilizadas para proyectos pequeños y medianos.

Tabla 7. Comparación de metodologías ágiles.

	SCRUM	**XP (Extreme Programming)**
Enfoque	Iterativo	Iterativo e incremental
Ciclo	Cada sprint es un ciclo completo	- Pruebas unitarias continuas - Iteraciones cortas
Planificación	- El plan de iteraciones se define una vez terminado el anterior. - El usuario determina cuando se realiza el proyecto.	- Refactorización de código - Artículos en un orden de prioridad estricto
Alcance	Cartera de proyectos, que se re-evalúan al final de cada iteración.	Revisión continua de avances.
Artefactos	El único artefacto formal es el software operativo	Historias de usuarios
Producto	Mejoras rápidas, cambiante	Mejoras rápidas, proyectos pequeños
Equipo	Se asigna tareas a cada miembro del equipo.	Programación por parejas

Nota: Tabla comparativa de características entre metodologías agiles SCRUM y XP.

La metodología XP se centra más en la programación y creación del proyecto, mientras que SCRUM se basa su administración, motivo por el cual se optó por SCRUM para la gestión de nuestro proyecto, para entregas parciales y regulares del producto final, priorizando objetivos y controlando el tiempo de desarrollo.

1.6.2 Metodología ágil SCRUM.

Metodología ágil que tiene como prioridad minimizar los riesgos durante el desarrollo de proyectos, además de la colaboración de todos los involucrados.

Se concentra en definir las características que el producto final debe cumplir y en mantener unido al equipo de trabajo para lograr sus propósitos.

Entre sus ventajas están; la productividad del equipo, calidad del producto, seguimiento diario de avances, realización de cambios funcionales o prioritarios en el progreso del proyecto, indistintamente la fase en que este se encuentre.

Scrum utiliza iteraciones (Sprints) que representan una parte del producto entregable al cliente, en periodos de hasta 4 semanas, se realizan reuniones para la evaluación del avance de entregas realizadas y de las actividades por realizar.

1.6.2.1 *Características de la metodología ágil Scrum.*

Proceso que aplica buenas prácticas, que en conjunto buscan conseguir un buen trabajo en equipo y obtener resultados de calidad en menor tiempo.

Contempla los riesgos que conllevan desarrollar un proyecto donde los requerimientos son variables o no muy bien definidos.

Flexibilidad a cambios que usualmente se dan al presentar los entregables al cliente, el cual puede utilizar los módulos desarrollados de manera completa ya que las entregas son completamente funcionales.

1.6.2.2 *Beneficios de la metodología ágil Scrum.*

Cumplimiento de expectativas, el cliente identifica sus necesidades, definiendo el coste que contribuye al proyecto cada una, el equipo junto al Product Owner evalúan los requisitos y establecen su prioridad de desarrollo.

Flexibilidad a cambios, capacidad del equipo de desarrollo para acoplarse a cambios inadvertidos en requerimientos, sean generados por necesidades del cliente o cambios del mercado.

Reducción del Tiempo, el cliente puede realizar pruebas de la funcionalidad del producto acorde a las entregas realizadas, antes de que esté finalizado por completo, permitiendo mejoras oportunas en el mismo.

Mayor calidad del software, el objetivo es la entrega de un producto que cumpla las perspectivas del cliente, corrigiendo posibles errores o haciendo mejoras de las versiones funcionales después de cada entrega.

Mayor productividad, motivación constante al grupo de trabajo, eliminación de favoritismos entre los involucrados.

Predicción de tiempos, análisis acorde la funcionalidad del requerimiento, el tiempo promedio que tardará el equipo en la entrega del sprint, para un mejor control de avances del proyecto.

Reducción de riesgos, identificar procesos de negocio, para establecer prioridades al desarrollar el proyecto y anticiparse a posibles riegos, tomando las medidas pertinentes.

1.6.3 Elementos de la metodología.

1.6.3.1 *Artefactos.*

Product Backlog, (Pila de Producto), lista que contiene todos los requerimientos por realizar, del proyecto, ordenada acorde a la prioridad de sus elementos.

Sprint Backlog, (Pila del Sprint), es la planificación de actividades y/o tareas que conforman una entrega (sprint) del proyecto al cliente.

Sprint, (Iteración), tiempo estimado que le toma al equipo de trabajo, para entregar resultados completos de la iteración a la que se ha comprometido.

Iteración, ciclo repetitivo para alcanzar un determinado objetivo, donde se sigue una serie de instrucciones a realizar.

Incremento, resultado de cada sprint.

1.6.3.2 *Participantes.*

Product Owner, responsable de asegurar que se cumplan los objetivos del proyecto, es el intermediario entre el equipo responsable y el cliente.

Scrum Team, equipo de desarrollo que se encarga del cumplimiento de las tareas asignadas por el Product Owner.

Scrum Master, lidera las reuniones y apoya al equipo, para cumplimiento de los objetivos del sprint, guiándolo en la ejecución de procesos y reglas de la metodología, es un facilitador mas no un gestor.

Stakeholders, son las partes implicadas en el proyecto, ya sea de manera directa o indirecta. Su opinión interfiere en el desarrollo del mismo.

Cliente, dueño del producto final, influye en el proceso con la aportación de ideas y/o comentarios respecto al desarrollo.

1.6.4 Implementación de la metodología.

Para llevar a cabo la ejecución del proyecto en cuestión se optó por la metodología ágil Scrum, la cual fue seleccionada por las siguientes circunstancias:

Equipo de trabajo conformado por dos personas.

Obtención de resultados en corto plazo.

La posibilidad de cambios de forma y fondo fueron considerados como altamente probables.

Scrum cuenta con cuatro tipos de participantes, que en el contexto del presente proyecto se lo ha manejado como se detalla a continuación:

Scrum Master, responsabilidad asignada al docente tutor del proyecto, quien veló y asesoró a los miembros del equipo, para que el proyecto se desarrolle.

Scrum Team, formado por dos personas.

Stakeholders, personas que gestionan las entidades de salud que colaboraron con entrevistas para la identificación de variables y generación de reportes.

Clientes, los posibles beneficiarios de la información conseguida con la minería de datos, directores entidades de salud.

Acorde la metodología ágil SCRUM, se elaboró un Product Backlog para definir las tareas a realizar, segmentadas en 7 Sprints.

Tabla 8. Product Backlog del Proyecto.

Descripción	Sprint	Fecha	Observaciones
Identificación, validación y análisis de información.	Sprint 1	13/04/2017	Información obtenida de la página http://www.ecuadorencifras.gob.ec/banco-de-informacion/
Creación del modelo operacional y dimensional del DW.	Sprint 2	15/05/2017	Se tuvo en consideración los datos de los años 2011, 2012, 2013 y 2014.

Definición de dimensiones para el almacén de datos.	Sprint 3	14/06/2017	Se identificaron cuatro dimensiones tiempo, geografía, entidades y síntomas
Análisis de datos para la creación de tabla de hechos.	Sprint 4	16/07/2017	Se realizó entrevistas para identificar la información que contiene la tala de hechos.
Definición de áreas de interés para la adquirir conocimiento.	Sprint 5	14/08/2017	Identificación de tendencias, patrones y relaciones significativas entre los datos con los que se dispone.
Integración de la herramienta Pentaho.	Sprint 6	13/09/2017	Utilización de Pentaho para el análisis de información y generación de repostes.
Integración de la herramienta Matlab.	Sprint 7	14/11/2017	Utilización del método de regresión lineal en Matlab.

Nota: Tabla con la descripción general de las actividades a realizar en el transcurso de desarrollo del proyecto.

De acuerdo a la metodología ágil SCRUM se elaboró también el Sprint Backlog en el cual se definen las actividades o tareas a realizarse en cada iteración.

Tabla 9. Actividades sprint 1.

Descripción	Tiempo
Planeación del desarrollo del proyecto	7 días
Identificación de fuentes de información oficiales.	4 días
Validación de disponibilidad y descarga de datos.	7 días
Verificación de la calidad de datos obtenidos.	12 días

Nota: Tabla con la descripción de las actividades correspondientes al Sprint 1.

Tabla 10. Actividades sprint 2.

Descripción	Tiempo
Generación del modelo de la base operacional.	5 días
Proceso de extracción, transformación y carga (ETL).	10 días
Migración de información a la base operacional.	16 días

Nota: Tabla con la descripción de las actividades correspondientes al Sprint 2.

Tabla 11. Actividades sprint 3.

Descripción	Tiempo

Análisis e identificación de áreas de interés (Dimensiones)	11 días
Creación de la estructura dimensional para el almacén de datos	8 días
Carga de información a las tablas dimensiones del almacén de datos.	11 días

Nota: Tabla con la descripción de las actividades correspondientes al Sprint 3.

Tabla 12. Actividades sprint 4.

Descripción	Tiempo
Creación de la estructura para la tabla de hechos.	10 días
Carga de información para la Tabla de Hechos.	20 días

Nota: Tabla con la descripción de las actividades correspondientes al Sprint 4.

Tabla 13. Sprint con una actividad a realizar.

Descripción	Tiempo
Sprint 5. Definición de áreas de interés para la adquisición de conocimiento.	30 días
Sprint 6. Integración de la herramienta Pentaho	30 días
Sprint 7. Integración de la herramienta Matlab	30 días

Nota: Tabla con la descripción de las actividades correspondientes a los Sprint 5, 6 y 7.

La documentación respectiva se desarrolló en paralelo a las actividades de cada una de las iteraciones.

Las áreas consideradas de interés, se identifican según la matriz de involucrados, la misma que permite determinar el grado de influencia de los implicados y disponer de información de contacto.

Tabla 14. Matriz de Involucrados.

Nº	Nombre	Cargo/Rol	Contacto	Interés	Nivel
1	INEC	Institución Pública	Tel: (593-2) 2232303 *www.ecuadorencifras.gob.ec*	Verificación de información proporcionada, para el bienestar común.	10

2	MSP Coordinación Zonal 9 (Pichincha)	Institución Pública	Tel: 3931020 ext. 333 www.salud.gob.ec	Visualizar las afecciones de salud más frecuentes en la población ecuatoriana	10
4	Profesionales que gestionan establecimientos de salud.	Director, Asistente de dirección, Personal varios.	Tel: 0995789106 Director centro de salud tipo B. Tel:0992905103 Asistente de dirección entidad de salud privada.	Generación de reportes de su interés. Obtención de reportes en tiempos prudentes. Tener datos históricos de la situación de la población ecuatoriana.	10
5	Alvarez Mayra, Cañarejo Vinicio (Estudiantes de la Universidad Politécnica Salesiana)	Miembros del proyecto de titulación	0979350736 malvarezj@est.ups.edu.ec; 0987610918 vcanarejo@est.ups.edu.ec;	Crear una herramienta para el procesamiento de datos proporcionados por el INEC para obtención de información de las afecciones de la salud Clasificación CIE-10. Posteriormente su implementación para la obtención del título profesional.	10
4	Arévalo René (Docente de la Universidad Politécnica Salesiana)	Tutor	0995459598 *aarevalo@ups.edu.ec*	Verificar el cumplimiento de objetivos planteados en el presente proyecto de titulación.	10

Nota: Tabla con la descripción general de los involucrados directa o indirectamente en la elaboración del proyecto en cuestión.

Capítulo 2
Diseño y Construcción

Para lograr nuestro propósito en la implementación de un BI para analizar la incidencia de enfermedades CIE-10 en Ecuador, se debe tomar en cuenta tres conceptos clave:

Datos, en nuestro caso son provistos por el INEC, por sí solos no brindan información, sin embargo, a partir de ellos se la crea.

Información, lo constituyen las dimensiones tiempo, geografía, entidad y síntomas, para cagar dicha información se emplean funciones.

Conocimiento, resultado diferentes datos conectados para el correspondiente análisis, para su visualización utilizamos la herramienta Pentaho.

Una vez aclarado este punto, se procede a cumplir de forma ordenada con las siguientes fases, acorde a los ítems establecidos en el Product Backlog.

2.1 Capa de Datos

Como ya lo mencionamos anteriormente la capa de datos la constituyen las fuentes información, a la cual se debe realizar el análisis respectivo y validar la disponibilidad de la misma.

2.1.1 Sprint 1. Identificación, validación y análisis de información.

2.1.1.1 *Planeación de desarrollo del proyecto.*

En esta fase de implementación del almacén de datos se debe considerar:

Alcance del proyecto. En busca una solución útil que apoye al proceso de decisiones gerenciales orientadas a brindar un mejor servicio de salud, mejorando los sistemas tecnológicos enlazados a dichos procesos, se propone la implementación de BI,

fundamentado en almacenamiento y minería de datos, con información de egresos hospitalarios de los últimos cuatro años, obtenidos de la página oficial del INEC.

El resultado del análisis, es presentado acorde a requerimientos recopilados en entrevistas realizadas a personas con conocimiento del tema, mediante reportes, utilizando las herramientas de Pentaho y Matlab.

Plan de pruebas, si los scripts utilizados para comprobar la veracidad de los datos son incorrectos, no se podrá garantizar la confiabilidad de los datos finales y se deberá replantear el plan de pruebas y verificar la consistencia de datos acorde a los archivos con información original.

Atrasos en corrección de errores, en caso de haber muchas inconsistencias en datos obtenidos, para mitigar este riesgo se deberá verificar la consistencia de datos en cada avance que se entregue, comparando los resultados obtenidos con los originales.

Definición del equipo de trabajo y sus roles. Descrito en el literal 2.6.4 Implementación de la Metodología.

Asignación de actividades a desarrollar. Enlistadas en la Tabla Nº 7 y tiempo aproximado en que se llevaran a cabo cada una (Tablas de Sprints), lo cual ha sido definido como parte del plan del proyecto en cuestión.

2.1.1.2 *Identificación de fuentes de información oficiales.*

La fuente de información del INEC consultada para llevar a cabo el proyecto, tiene relación directa con el análisis de incidencias de enfermedades y problemas con la salud que afectan la población ecuatoriana, (Capítulo Sociales, Sección Egresos Hospitalarios), específicamente en el capítulo L del catálogo CIE-10, que pertenece a enfermedades de la piel.

2.1.1.3 *Validación de disponibilidad y descarga de datos.*

El Instituto Nacional de Estadísticas y Censos se especialista en brindar diversos servicios producto de investigaciones especializadas con fines estadísticos, permitiendo la consulta a diferentes bases de datos.

Los archivos que reposan en dichas bases poseen información original provista por las respectivas entidades, lo cual asegura la veracidad de información, mientras que "LOTAIP" (cuyas siglas corresponden a la Ley Orgánica de Transparencia y Acceso a la Información Pública,) asegura la disponibilidad de información a la ciudadanía en general, garantizando el acceso a la misma. Actualmente esta información está disponible en el sitio web http://www.ecuadorencifras.gob.ec/banco-de-informacion/.

Ubicación de datos de la fuente del INEC

Investigación	Producto	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Sociales	Divorcios										
Sociales	Egresos Hospitalarios										
Sociales	Entradas y Salidas Internacionales Migración										
Sociales	Matrimonios										
Sociales	Nacimientos										
Sociales	Recursos y Actividades de Salud										
Sociales	Victimización y Percepción de Inseguridad										
Sociodemográficas	Censo de Población y Vivienda										

Figura 5. Ubicación de archivos de descarga.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

Una vez identificada el área de interés (Egresos Hospitalarios), procedemos a descargar los archivos de los últimos 4 años (2011-2014), los mismos que se encuentran por separado en carpetas comprimidas, que contienen la información en

formato .sav (almacenamiento de archivos y datos) y en formato .xlxs las etiquetas empleadas en las tablas.

2.1.1.4 *Verificación de la calidad de datos obtenidos.*

Para establecer relación entre la información obtenida, comparamos la existencia de campos iguales en los archivos de cada año, para posteriormente identificar los campos de relevancia que serán utilizados en la creación del almacén de datos.

Para poder visualizar los datos descargada, es necesaria la instalación de la aplicación IBM SPSS Estatistics compatible con archivos. .sav que luego los exportaremos a formato .csv (texto delimitado por comas), para la carga de datos en la base operacional.

2.2 Capa de Información

La información la constituye los datos que han sido analizados y pasan a formar parte de las tablas de la base operacional y dimensiones del almacén de datos que alimentan la tabla de hechos.

2.2.1 Sprint 2: Creación del modelo operacional y dimensional del DW.

2.2.1.1 *Generación del modelo de BDD operacional.*

Al disponer de la información necesaria para nuestro propósito y teniendo claro los requerimientos a cumplir, nuestro siguiente paso es la elaboración del modelo relacional para la base operacional.

Para establecer la estructura de la base operacional, se identificó la relación entre la información obtenida y así procedimos al modelado de las tablas correspondientes, definiendo claves primarias y foráneas para asegurar la calidad de información que se almacenara en dicha base.

Esquema relacional de la base operacional

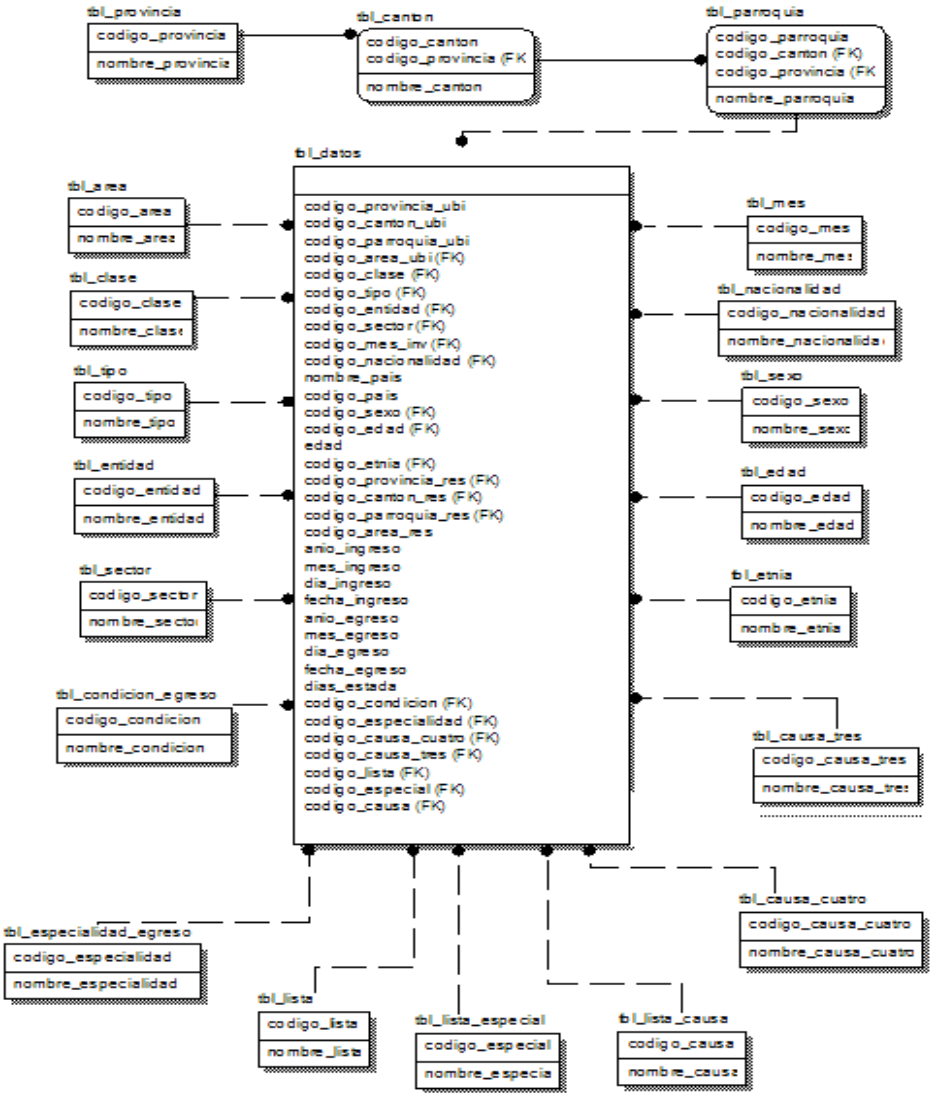


Figura 6. Modelo relacional con información del año 2011.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

Ya definido el modelo relacional de la base operacional, procedemos al análisis de información que contiene cada año, para así identificar los campos de interés que formaran parte del almacén.

La creación de la base operacional es una de las principales características del almacén, por ello, los datos deben pasar por el proceso de extracción, transformación y carga (ETL) para asegurar la calidad de información.

2.2.1.2 *Proceso de Extracción, transformación y caga (ETL).*

Fase 3. Construcción (Almacén de Datos).

Luego de la definición del modelo relacional y sus características, se procede a realizar la limpieza, preparación y carga de datos.

Proceso de Extracción.

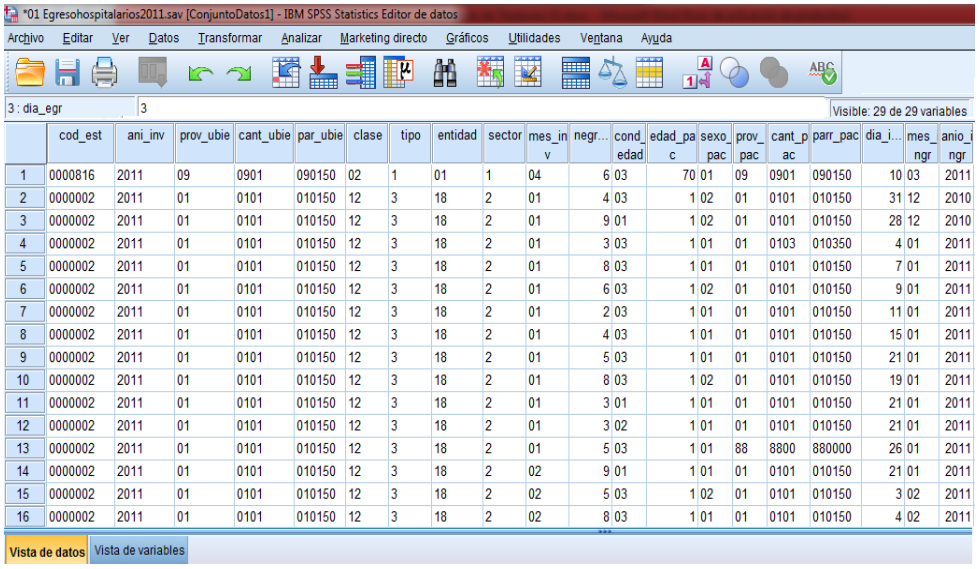
Para demostración ejemplo tomamos los datos del año 2011, al abrir el archivo obtenido, haciendo uso de la aplicación IBM SPSS, observamos que existen dos pestañas (Vista de datos y Vista de variables), la primera contiene información recolectada de diferentes entidades de salud a nivel nacional; mientras que la segunda contiene los parámetros que se muestra en los datos, además de características propias de cada atributo.

Los archivos que provee el INEC conservan todos los mismos formatos, la diferencia reside en los parámetros disponibles para cada año, por ejemplo, originalmente los archivos a analizar correspondían a los años desde el 2012 al 2015, sin embargo, luego del respectivo análisis se reemplazaron los archivos del año 2015 por los del año 2011.

La inconsistencia en el archivo del año 2015, fue que no existe información del campo cod_est, corresponde al Código de establecimiento de salud, motivo por el que se descartó para nuestro propósito.

Para dicha deducción nos basamos en la pestaña Vista de datos, que contiene la información a ser cargada en la tabla principal, una por año, mientras que la pestaña Vista de variables contiene los datos de las tablas secundarias que se usaran de manera global en la BDD operacional.

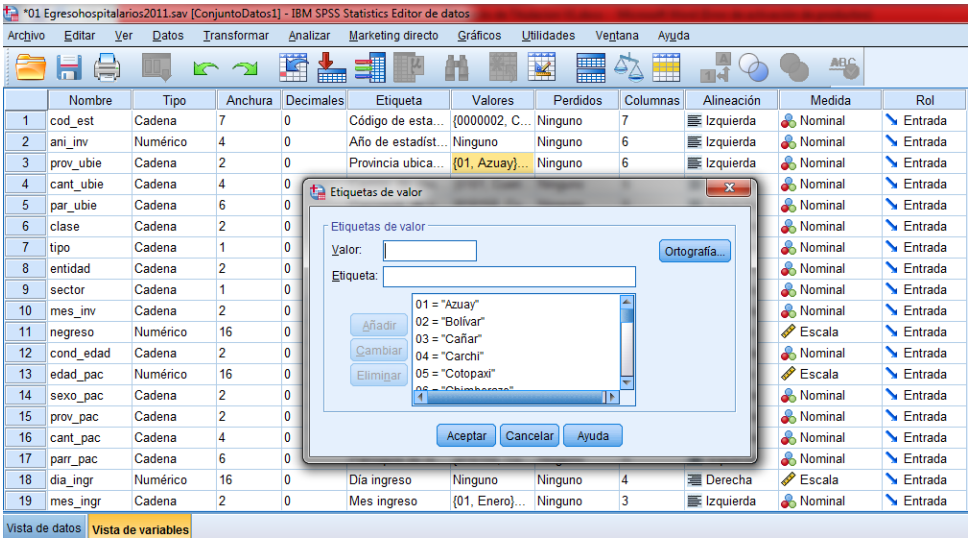
Información de recurso de egresos hospitalarios



	cod_est	ani_inv	prov_ubie	cant_ubie	par_ubie	clase	tipo	entidad	sector	mes_inv	negr...	cond_edad	edad_pac	sexo_pac	prov_pac	cant_pac	parr_pac	día_i...	mes_ngr	anio_ingr
1	0000816	2011	09	0901	090150	02	1	01	1	04	6	03	70	01	09	0901	090150	10	03	2011
2	0000002	2011	01	0101	010150	12	3	18	2	01	4	03	1	02	01	0101	010150	31	12	2010
3	0000002	2011	01	0101	010150	12	3	18	2	01	9	01	1	02	01	0101	010150	28	12	2010
4	0000002	2011	01	0101	010150	12	3	18	2	01	3	03	1	01	01	0103	010350	4	01	2011
5	0000002	2011	01	0101	010150	12	3	18	2	01	8	03	1	01	01	0101	010150	7	01	2011
6	0000002	2011	01	0101	010150	12	3	18	2	01	6	03	1	02	01	0101	010150	9	01	2011
7	0000002	2011	01	0101	010150	12	3	18	2	01	2	03	1	01	01	0101	010150	11	01	2011
8	0000002	2011	01	0101	010150	12	3	18	2	01	4	03	1	01	01	0101	010150	15	01	2011
9	0000002	2011	01	0101	010150	12	3	18	2	01	5	03	1	01	01	0101	010150	21	01	2011
10	0000002	2011	01	0101	010150	12	3	18	2	01	8	03	1	02	01	0101	010150	19	01	2011
11	0000002	2011	01	0101	010150	12	3	18	2	01	3	01	1	01	01	0101	010150	21	01	2011
12	0000002	2011	01	0101	010150	12	3	18	2	01	3	02	1	01	01	0101	010150	21	01	2011
13	0000002	2011	01	0101	010150	12	3	18	2	01	5	03	1	01	88	8800	880000	26	01	2011
14	0000002	2011	01	0101	010150	12	3	18	2	02	9	01	1	01	01	0101	010150	21	01	2011
15	0000002	2011	01	0101	010150	12	3	18	2	02	5	03	1	02	01	0101	010150	3	02	2011
16	0000002	2011	01	0101	010150	12	3	18	2	02	8	03	1	01	01	0101	010150	4	02	2011

Figura 7. Vista de datos año 2011.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

Información de recurso de Egreso Hospitalario



	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida	Rol
1	cod_est	Cadena	7	0	Código de esta...	{0000002, C...	Ninguno	7	Izquierda	Nominal	Entrada
2	ani_inv	Numérico	4	0	Año de estadíst...	Ninguno	Ninguno	6	Izquierda	Nominal	Entrada
3	prov_ubie	Cadena	2	0	Provincia ubica...	{01, Azuay}...	Ninguno	6	Izquierda	Nominal	Entrada
4	cant_ubie	Cadena	4	0						Nominal	Entrada
5	par_ubie	Cadena	6	0						Nominal	Entrada
6	clase	Cadena	2	0						Nominal	Entrada
7	tipo	Cadena	1	0						Nominal	Entrada
8	entidad	Cadena	2	0						Nominal	Entrada
9	sector	Cadena	1	0						Nominal	Entrada
10	mes_inv	Cadena	2	0						Nominal	Entrada
11	negreso	Numérico	16	0						Escala	Entrada
12	cond_edad	Cadena	2	0						Nominal	Entrada
13	edad_pac	Numérico	16	0						Escala	Entrada
14	sexo_pac	Cadena	2	0						Nominal	Entrada
15	prov_pac	Cadena	2	0						Nominal	Entrada
16	cant_pac	Cadena	4	0						Nominal	Entrada
17	parr_pac	Cadena	6	0						Nominal	Entrada
18	dia_ingr	Numérico	16	0	Día ingreso	Ninguno	Ninguno	4	Derecha	Escala	Entrada
19	mes_ingr	Cadena	2	0	Mes ingreso	{01, Enero}...	Ninguno	3	Izquierda	Nominal	Entrada

Figura 8. Vista de variables año 2011.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

Originalmente los datos están en formato .sav, motivo por el cual debemos exportarlos a formato .csv para poder manipular la información como archivo plano o haciendo uso de la herramienta ofimática Excel.

Abrimos los archivos obtenidos con la herramienta SPSS, en la pestaña Vista de datos, nos dirigimos a Archivo > Guardar Como, seleccionamos la opción Delimitado por comas (*.csv) y presionamos Almacenar archivo en Repositorio.

Pantalla para exportación de datos.

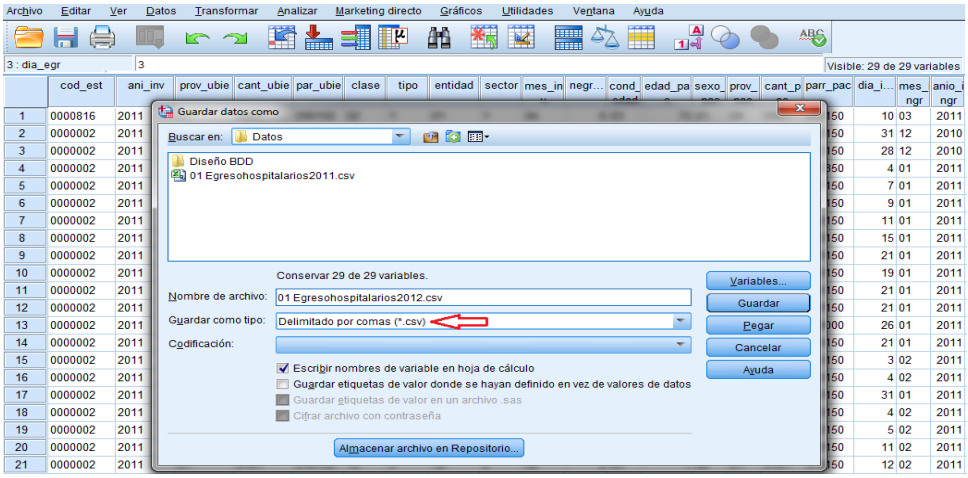
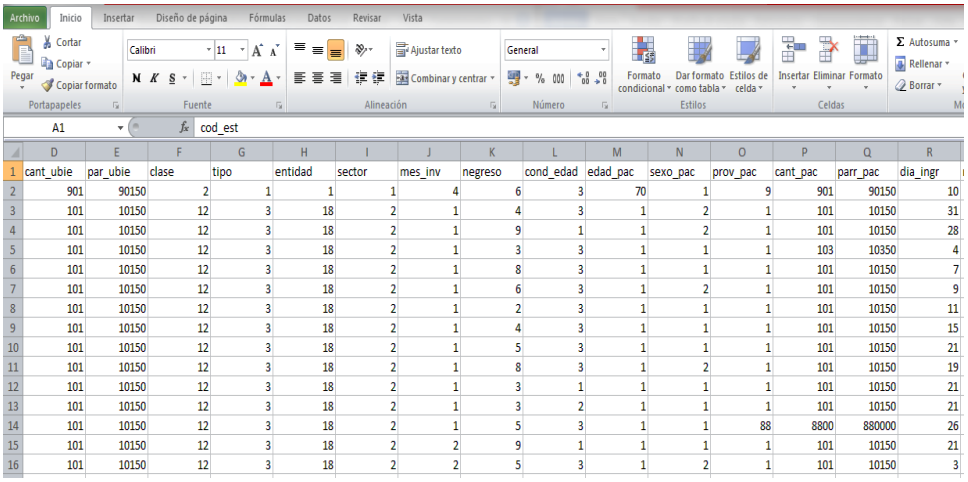


Figura 9. Exportación de información a formato csv.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

En la sección Buscar en, seleccionamos la carpeta que guardará el archivo que posteriormente usaremos para cargar los datos. Se realiza el mismo procedimiento por cada año.

Información visualizada en Excel



	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	cant_ubie	par_ubie	clase	tipo	entidad	sector	mes_inv	negreso	cond_edad	edad_pac	sexo_pac	prov_pac	cant_pac	parr_pac	dia_ingr
2	901	90150	2	1	1	1	4	6	3	70	1	9	901	90150	10
3	101	10150	12	3	18	2	1	4	3	1	2	1	101	10150	31
4	101	10150	12	3	18	2	1	9	1	1	2	1	101	10150	28
5	101	10150	12	3	18	2	1	3	3	1	1	1	103	10350	4
6	101	10150	12	3	18	2	1	8	3	1	1	1	101	10150	7
7	101	10150	12	3	18	2	1	6	3	1	2	1	101	10150	9
8	101	10150	12	3	18	2	1	2	3	1	1	1	101	10150	11
9	101	10150	12	3	18	2	1	4	3	1	1	1	101	10150	15
10	101	10150	12	3	18	2	1	5	3	1	1	1	101	10150	21
11	101	10150	12	3	18	2	1	8	3	1	2	1	101	10150	19
12	101	10150	12	3	18	2	1	3	1	1	1	1	101	10150	21
13	101	10150	12	3	18	2	1	3	2	1	1	1	101	10150	21
14	101	10150	12	3	18	2	1	5	3	1	1	88	8800	880000	26
15	101	10150	12	3	18	2	2	9	1	1	1	1	101	10150	21
16	101	10150	12	3	18	2	2	5	3	1	2	1	101	10150	3

Figura 10. Archivo del año 2011 exportado a formato .csv.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

Proceso de Transformación.

Limpieza de datos. Los datos que conforman la base operacional deben ser normalizados, estandarizando los campos que componen la tabla principal, de manera que las tablas de años a analizar contengan los mismos campos.

Por este motivo descartamos el archivo del año 2015, ya que posee características que otros años no, en cuanto a campos de información que componen la tabla principal, por ejemplo, no contiene los datos de establecimiento de salud correspondiente a la tabla cod_est, por ello en su lugar para el análisis respectivo se utilizó el archivo del año 2011.

Además, se debe estandarizar el formato de las variables, de lo contrario al realizar la carga de datos se generarán errores de compatibilidad, debido a que los parámetros pasan a constituir claves foráneas de la tabla principal.

En el archivo del año 2011, se modificó el formato de las variables: clase, tipo, entidad, sector y cod_est, de formato de tipo cadena a formato tipo Integer, de modo que coincida con el estándar del resto de archivos.

Estandarización campos.

A	B	D	E	F	G	H	I
cod_est	ani_inv	cant_u	par_ul	clase	tipo	entidad	sector
000816	2011	901	90150	02	01	01	01
000002	2011	101	10150	12	03	18	02
000002	2011	101	10150	12	03	18	02
000002	2011	101	10150	12	03	18	02
000002	2011	101	10150	12	03	18	02
000002	2011	101	10150	12	03	18	02
000002	2011	101	10150	12	03	18	02
000002	2011	101	10150	12	03	18	02
000002	2011	101	10150	12	03	18	02
000002	2011	101	10150	12	03	18	02
000002	2011	101	10150	12	03	18	02
000002	2011	101	10150	12	03	18	02
000002	2011	101	10150	12	03	18	02
000002	2011	101	10150	12	03	18	02
000002	2011	101	10150	12	03	18	02
000002	2011	101	10150	12	03	18	02
000002	2011	101	10150	12	03	18	02

A	B	D	E	F	G	H	I
cod_est	ani_inv	cant_u	par_ul	clase	tipo	entidad	sector
816	2011	901	90150	2	1	1	1
2	2011	101	10150	12	3	18	2
2	2011	101	10150	12	3	18	2
2	2011	101	10150	12	3	18	2
2	2011	101	10150	12	3	18	2
2	2011	101	10150	12	3	18	2
2	2011	101	10150	12	3	18	2
2	2011	101	10150	12	3	18	2
2	2011	101	10150	12	3	18	2
2	2011	101	10150	12	3	18	2
2	2011	101	10150	12	3	18	2
2	2011	101	10150	12	3	18	2
2	2011	101	10150	12	3	18	2
2	2011	101	10150	12	3	18	2
2	2011	101	10150	12	3	18	2
2	2011	101	10150	12	3	18	2

Figura 11. Estandarización de formato tipo cadena a tipo Integer.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

Para tener clara la información que representan los campos, realizamos un diccionario de datos con el significado de cada registro, el mismo que se puede consultar en la sección 3.4 dedicada a los diccionarios de datos.

2.2.1.3 *Migración de información a la base operacional.*

Proceso de Carga.

Al diseñar el esquema relacional tenemos un panorama más claro de los datos que formarán la base operacional, para la migración debemos definir las claves primarias y foráneas correspondientes, tomando en cuenta la información de la pestaña Vista de variables de la fuente de datos.

Los datos exportados a formato .csv deben ser almacenados en la base operacional, para lo cual deben crearse las tablas necesarias para dicho propósito basándonos en el modelo relacional de la Figura Nº6.

Para la creación de tablas deben considerarse aquellas cuyos campos sean comunes en los años a analizar, una vez identificadas las tablas que formaran parte del almacén procedemos a cargar la información, tomamos como ejemplo el año 2011, en el archivo descargado de la fuente se encuentra el documento .xlxs que contiene información de dichas tablas.

Clasificamos los datos en diferentes hojas, acorde a las tablas seleccionadas para el almacén, tomando en cuenta los campos de la tabla principal que cambiamos el formato, debiendo hacer lo mismo con el resto de datos.

Datos de la tabla cod_est (entidades de salud)

COD_EST	Código de establecimiento de salud	1
Value	Label	
0000002	Clínica Sta. Ana Centro Medico Quir. S.A.	
0000003	Hospital Clínica Latinoamericana	
0000004	Hospital Santa Inés	
0000006	Hospital Vicente Corral Moscoso	
0000007	Hospital Del I.E.S.S. Jose Carrasco Arteaga	
0000008	Hospital División N°. 3 D. E. Tarqui	
0000010	Hospital Dermatológico Mariano Estrella	
0000011	Centro De Reposo Y Adicciones Cra	
0000012	Hospital Aida León Rodríguez	
0000013	Hospital Moreno Vásquez Gualaceo	
0000014	Hospital Jose Félix Valdivieso	
0000015	Hospital San Sebastián	
0000016	Hospital Cantonal Paute	
0000017	Clínica Paucarbamba	
0000018	Clínica Bolívar	
0000019	Clínica España S.A.	
0000021	Clínica Santa Bárbara	
0000028	Centro Quirúrgico Metropolitano	
0000031	Maternidad San Martin De Porres	

Figura 12. Clasificación de los datos (tabla cod_est).
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

Guardamos los datos clasificados en formato de tipo Texto Unicode (*.txt), abrimos dicho archivo, borramos los datos marcados de color azul que se muestra en la Figura Nº13 y guardamos el archivo.

Preparación de datos a exportar

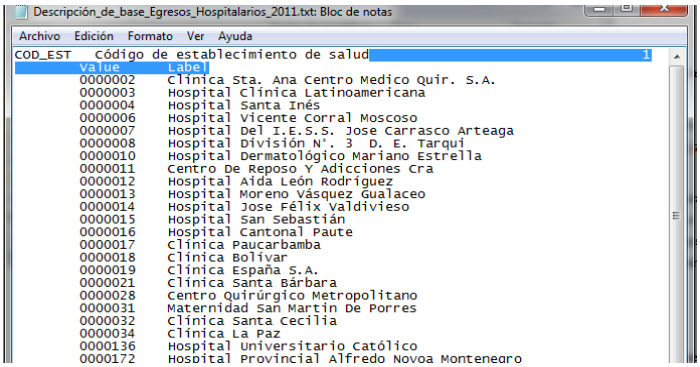


Figura 13. Datos exportados en formato .txt.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

Encontramos 16 tablas comunes, las mismas que se han creado en la base nombrada EgresoVCA, cargamos información haciendo uso del comando *copy nombre_tabla from 'ubicación /nombre_archivo .extension' using delimiters ' '* , cada una de estas tablas tiene su clave primaria.

Adicional creamos la tabla anio e ingresamos manualmente los años correspondientes a los datos a analizar, con su respectiva clave primaria.

Base de datos operacional y sus tablas.

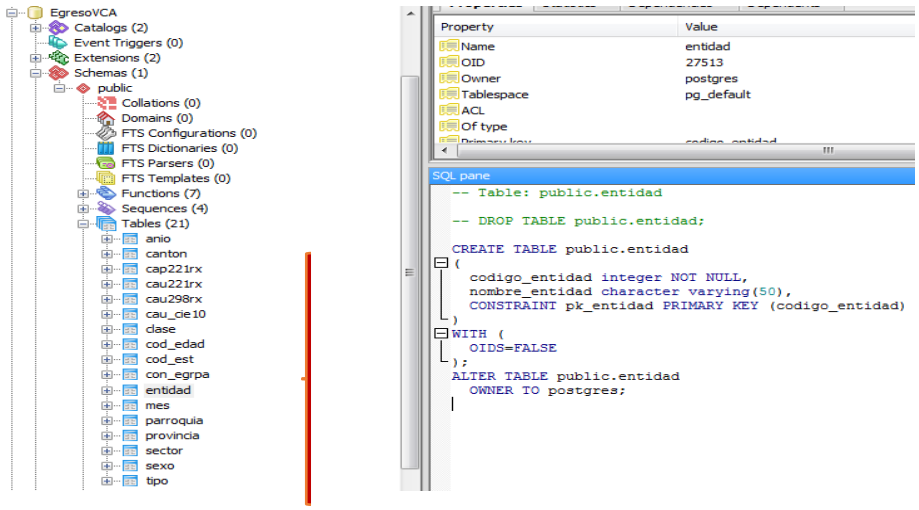


Figura 14. Tablas secundarias de la base operacional.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

Como caso especial encontramos la tabla cod_est, la cual se cargó en dos partes, la primera con datos proporcionados en el archivo fuente original.

Primera carga de la tabla cod_est

id_entidad [PK] integer	nombre_entidad character varying(100)	prov character	can charact	parr charact	sector integer	tipo integer	clase integer
2	CLÍNICA STA. ANA CENTRO MEDICO QUIR. S.A.						
3	HOSPITAL CLÍNICA LATINOAMERICANA						
4	HOSPITAL SANTA INÉS						
6	HOSPITAL VICENTE CORRAL MOSCOSO						
7	HOSPITAL DEL I.E.S.S. JOSE CARRASCO ARTEAGA						
8	HOSPITAL DIVISIÓN N°. 3 D. E. TARQUI						
10	HOSPITAL DERMATOLÓGICO MARIANO ESTRELLA						
11	CENTRO DE REPOSO Y ADICCIONES HUC						
12	HOSPITAL AIDA LEÓN RODRÍGUEZ						
13	HOSPITAL MORENO VÁSQUEZ GUALACEO						
14	HOSPITAL JOSE FELIX VALDIVIESO						
15	HOSPITAL SAN SEBASTIÁN						
16	HOSPITAL CANTONAL PAUTE						
17	CLÍNICA PAUCARBAMBA						
18	CLÍNICA BOLÍVAR						
19	CLÍNICA ESPAÑA S.A.						
21	CLÍNICA SANTA BARBARA						
28	CENTRO QUIRÚRGICO METROPOLITANO						
31	MATERNIDAD SAN MARTIN DE PORRES						
32	CLÍNICA SANTA CECILIA						

Figura 15. Primera carga de información de la tabla cod_est.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

Para completar la información de los demás campos, se realizó un script que al ejecutarse en PostgreSQL actualiza la información faltante en la tabla, cargando los datos en los campos de prov, can, parr, sector, tipo y clase respectivamente.

La generación del script de la Figura Nº16, debe realizarse por cada año de análisis, tomando como referencia la tabla principal acorde al año.

Script para actualizar los datos de cod_est

```
UPDATE cod_est SET
prov = egreso_2011_a.prov,
can = egreso_2011_a.can,
parr = egreso_2011_a.parr,
sector = egreso_2011_a.sector,
tipo = egreso_2011_a.tipo,
clase = egreso_2011_a.clase
FROM egreso_2011_a
WHERE egreso_2011_a.cod_est = id_entidad
```

Figura 16. Actualización de datos en cod_est.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

Luego de ejecutar el script para todos los años podremos constatar la actualización de datos en la tabla correspondiente.

Segunda carga de la tabla cod_est

id_entidad [PK] integer	nombre_entidad character varying(100)	prov character	can charact	parr charact	sector integer	tipo integer	clase integer
2	CLÍNICA STA. ANA CENTRO MEDICO QUIR. S.A.	01	01	50	2	3	12
3	HOSPITAL CLÍNICA LATINOAMERICANA	01	01	50	2	3	12
4	HOSPITAL SANTA INÉS	01	01	50	2	3	12
6	HOSPITAL VICENTE CORRAL MOSCOSO	01	01	50	1	1	2
7	HOSPITAL DEL I.E.S.S. JOSE CARRASCO ARTEAGA	01	01	50	1	1	11
8	HOSPITAL DIVISIÓN N°. 3 D. E. TARQUI	01	01	50	1	1	2
10	HOSPITAL DERMATOLÓGICO MARIANO ESTRELLA	01	01	50	1	5	1
11	CENTRO DE REPOSO Y ADICCIONES HUC	01	01	50	3	2	6
12	HOSPITAL AIDA LEÓN RODRÍGUEZ	01	02	50	1	5	1
13	HOSPITAL MORENO VÁSQUEZ GUALACEO	01	03	50	1	5	1
14	HOSPITAL JOSE FELIX VALDIVIESO	01	08	50	1	5	1
15	HOSPITAL SAN SEBASTIÁN	01	09	50	1	5	1
16	HOSPITAL CANTONAL PAUTE	01	05	50	1	5	1
17	CLÍNICA PAUCARBAMBA	01	01	50	2	3	12
18	CLÍNICA BOLÍVAR	01	01	50	2	3	12
19	CLÍNICA ESPAÑA S.A.	01	01	50	2	3	12
21	CLÍNICA SANTA BARBARA	01	03	50	2	3	12

Figura 17. Actualización de datos faltantes de la tabla cod_est.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

Creamos la tabla principal, para cargar la información de la pestaña Vista de datos, con los datos de todos los campos comunes, las claves principales de estas constituyen las claves foráneas de la tabla principal, de modo que, si los datos no se encuentran en la tabla secundaria dará un error en la carga.

Claves foráneas de la tabla principal del año 2011

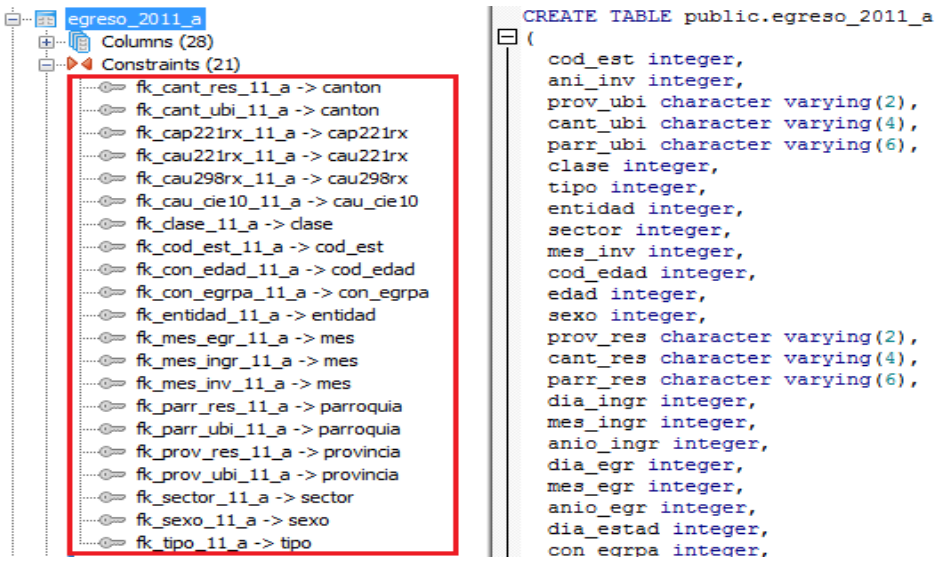


Figura 18. Tabla principal y claves foráneas del año 2011.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

Realizamos el mismo procedimiento de carga de las tablas secundarias

```
COPY egreso_2011 FROM 'E:/DatosAnios/subir2011.txt' USING DELIMITERS ' '
```

para todos los años en cuestión. Las tablas creadas deben ser del tipo que ya se definió en los datos cuyo formato fue modificado.

Inconvenientes, en este punto, al cargar los datos en las tablas principales de los años restantes (2012, 2013 y 2014), se encontraron inconvenientes en datos como cod_est, cantón y parroquia, ya que los datos de las tablas secundarias se obtuvieron basándonos en el año 2011, hubo registros de otros años que no constaban en nuestra base, generándose error en su carga.

Información faltante en la tabla secundaria

```
COPY egreso_2012 FROM 'E:/DatosAnios/subir2012.txt' USING DELIMITERS ';'
```

Output pane

Data Output | Explain | **Messages** | History

```
ERROR:  no existe la relación 1821,HOSPITAL PSIQUIÁTRICO SAN LÁZARO
********** Error **********

ERROR: no existe la relación «egreso_2012»
SQL state: 42P01
```

Figura 19. Error por información faltante.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

Revisamos los campos correspondientes a las tablas cod_est, cantón y parroquia de los demás años, realizando la inserción de cada registro faltante mediante un script. `INSERT INTO public.parroquia VALUES ('08','08','50','080850','LA CONDORDIA');`

Al completar la información faltante en las tablas secundarias, ejecutamos nuevamente el comando de carga. Al finalizar el proceso, podremos verificar la información en las respectivas tablas de la base operacional.

Base de datos operacional (EgresoVCA)

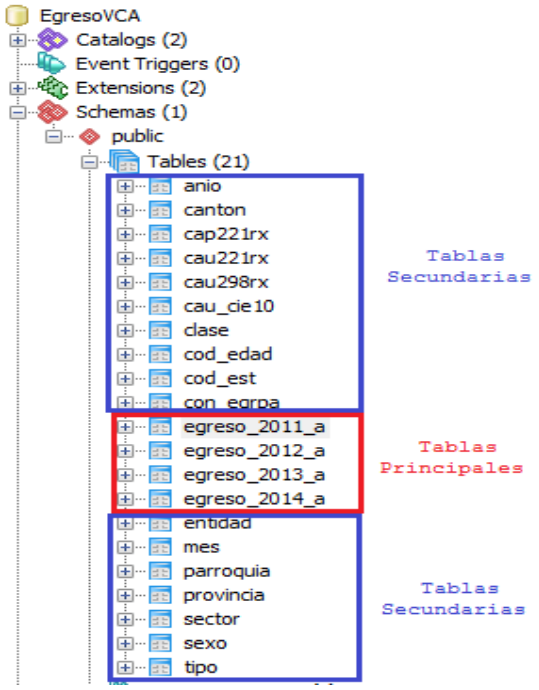


Figura 20. Esquema de BDD operacional.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

Prueba realizada. Llegado a este punto debemos verificar los que los datos que componen las tablas de nuestra base operacional hayan sido cargados correctamente, para ello realizamos un select COUNT(*) from egreso_2011_a, para ver el total de registros de la tabla y compararlos con los registros originales.

Comprobación de total de registros año 2011

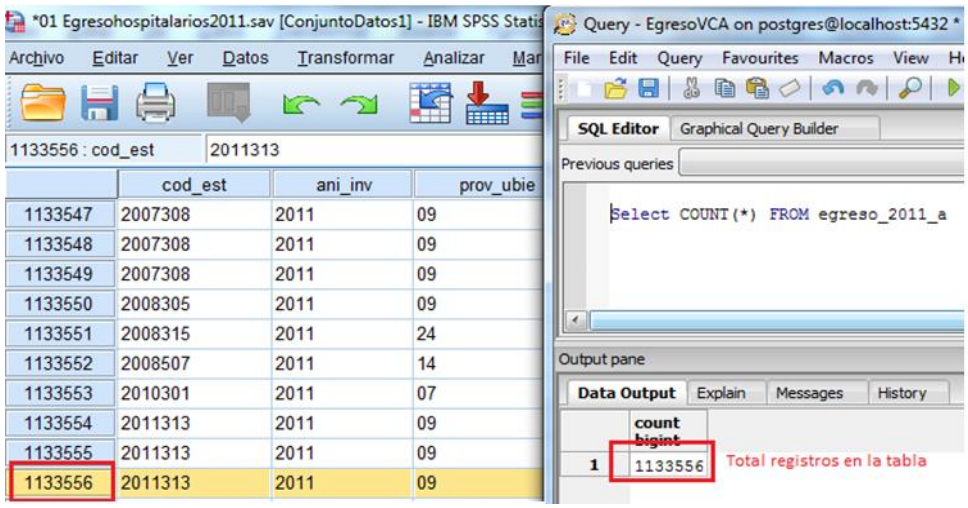


Figura 21. Número total de registros cargados año 2011.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

Esta comparación se realiza para cada tabla principal de los años a analizar. Si los datos no coincidieran en número, debemos reevaluar los procedimientos y corregir errores.

2.2.2 Sprint 3. Definición de dimensiones para el almacén de datos.

2.2.2.1 *Análisis e identificación de áreas de interés (Dimensiones).*

Fase 4. Diseño Dimensional.

La elaboración del modelo relacional de la base operacional permite identificar la relación existente entre los datos, de modo que podamos seleccionar los datos de mayor interés, para la elaboración del modelo dimensional basándonos el esquema estrella.

El análisis de las dimensiones a utilizar debe ser exhaustivo, revisando minuciosamente los datos que podrían asociarse y la jerarquía entre ellos, ya que en función de estas se cargará la información en la tabla de hechos. Para nuestro propósito definimos las siguientes dimensiones y la jerarquía entre los datos a utilizar.

Tabla 15. Dimensiones y Jerarquías.

Dimensiones	Nivel 1	Nivel 2	Nivel 3	Nivel 4
Tiempo	Año	Mes		
Geografía	Provincia	Cantón	Parroquia	
Entidad Salud	Sector	Tipo Entidad	Clase Entidad	Establecimiento
Síntoma	Capítulo	Lista	Causa	

Nota: Identificación de dimensiones y jerarquía de los datos

El modelo dimensional debe ser elaborado de tal manera que simplifique y facilite la comprensión del contenido de la BDD, permitiendo su análisis en menor tiempo, convirtiendo los datos en información útil al negocio.

2.2.2.2 *Creación de la estructura dimensional para el almacén.*

Tomando en cuenta el alcance del proyecto, la granulidad y jerarquía de datos se definieron 4 dimensiones: tiempo, geografía, entidades de salud y síntoma, que permitirán la adecuada clasificación de información.

La información que forma parte de las dimensiones debe ser analizada a detalle, ya que a partir de estas se creara la tabla de hechos, que a su vez dará origen a las vistas materializadas que se emplea para el análisis de información que se presenta al usuario final a manera de reportes.

Estructura dimensional tipo estrella.

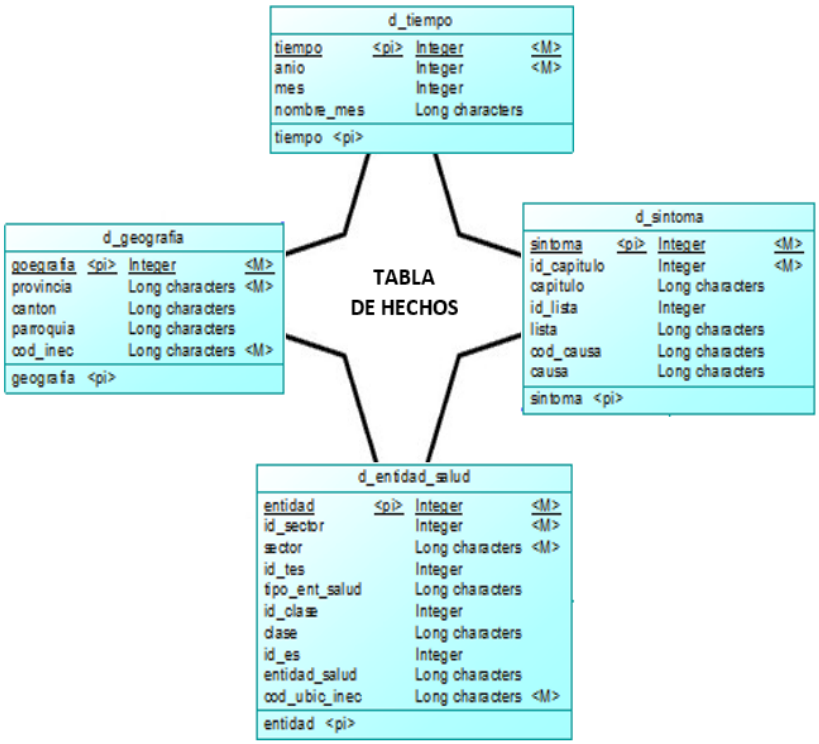


Figura 22. Esquema dimensional tipo estrella.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

2.2.2.3 *Carga de información a tablas dimensionales del almacén.*

Creamos la base de datos AlmacenVCA y las tablas dimensionales con los respectivos campos, acorde al modelo de la Figura Nº22.

Tablas dimensionales del almacén

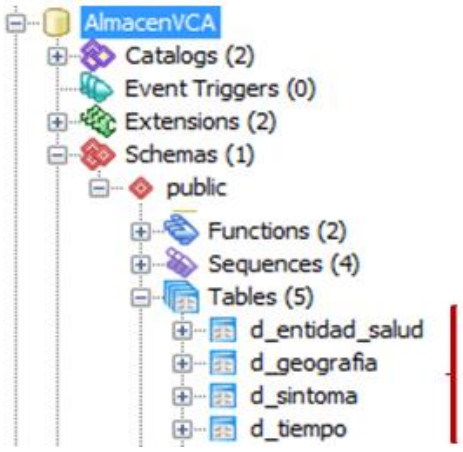


Figura 23. Tablas dimensionales.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

Para cargar los datos desde la base operacional a la base del almacén de datos, se debe crear la relación entre estas, para lo cual utilizamos el siguiente script, que deberemos ejecutar desde la base EgresosVCA.

Creación de la relación entre las bases de datos.

```
-- 1. Instancia de Conexion de Posgresql
CREATE EXTENSION postgres_fdw;

-- 2. Creación del servidor
CREATE SERVER almacenVCA_server
FOREIGN DATA WRAPPER postgres_fdw
OPTIONS (host 'localhost', port '5432',dbname 'AlmacenVCA');

-- 3. Usuarios con permiso para acceder a la BDD y
      direccionamiento del mapeo de datos
CREATE USER MAPPING FOR postgres
SERVER almacenVCA_server
OPTIONS (user 'postgres', password '123456');
```

Figura 24. Script para relacionar las BDD creadas.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

Además, se debe mapear cada una de las tablas dimensionales, acorde a los campos que contengan, como se muestra en la Figura Nº25.

Parte 1 de 2: Mapeo de las tablas dimensionales

```
CREATE FOREIGN TABLE d_entidad_salud (
id serial NOT NULL,
sector character varying OPTIONS (column_name 'sector'),
tipo_ent_salud character varying OPTIONS (column_name 'tipo_ent_salud'),
clase character varying OPTIONS (column_name 'clase'),
entidad_salud character varying OPTIONS (column_name 'entidad_salud'),
id_sector integer OPTIONS (column_name 'id_sector'),
id_tes integer OPTIONS (column_name 'id_tes'),
id_clase integer OPTIONS (column_name 'id_clase'),
id_es integer OPTIONS (column_name 'id_es'),
cod_ubic_inec character varying(6) OPTIONS (column_name 'cod_ubic_inec')
) SERVER almacenVCA_server
  OPTIONS (table_name 'd_entidad_salud');

CREATE FOREIGN TABLE d_tiempo (
id serial NOT NULL,
anio integer OPTIONS (column_name 'anio'),
mes integer OPTIONS (column_name 'mes'),
nombre_mes character varying OPTIONS (column_name 'nombre_mes')
) SERVER almacenVCA_server
  OPTIONS (table_name 'd_tiempo');
```

Figura 25. Script de creación de claves foráneas de tablas dimensionales de tiempo y entidad de salud.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

Parte 2 de 2: Mapeo de las tablas dimensionales

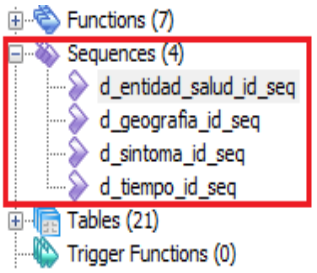
```
CREATE FOREIGN TABLE d_sintoma (
id serial NOT NULL,
id_capitulo integer OPTIONS (column_name 'id_capitulo'),
capitulo character varying OPTIONS (column_name 'capitulo'),
id_lista integer OPTIONS (column_name 'id_lista'),
lista character varying OPTIONS (column_name 'lista'),
cod_causa character varying OPTIONS (column_name 'cod_causa'),
causa character varying OPTIONS (column_name 'causa')
) SERVER almacenVCA_server
  OPTIONS (table_name 'd_sintoma');

CREATE FOREIGN TABLE d_geografia (
id serial NOT NULL,
provincia character varying OPTIONS (column_name 'provincia'),
canton character varying OPTIONS (column_name 'canton'),
parroquia character varying OPTIONS (column_name 'parroquia'),
cod_inec character varying(6) OPTIONS (column_name 'cod_inec')
) SERVER almacenVCA_server
  OPTIONS (table_name 'd_geografia');
```

Figura 26. Script de creación de claves foráneas de tablas dimensionales síntomas y geografía.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

Adicional se creó secuencias en ambas bases para cada dimensión, que son llamadas en los cursores para colocar los identificadores correspondientes a cada registro de forma automática.

Creación de secuenciales para tablas dimensionales




```
CREATE SEQUENCE public.d_entidad_salud_id_seq
  INCREMENT 1
  MINVALUE 1
  MAXVALUE 9223372036854775807
  START 1012
  CACHE 1;
ALTER TABLE public.d_entidad_salud_id_seq
  OWNER TO postgres;
```

Figura 27. Secuenciales usadas en ambas bases.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

Para nuestro propósito, creamos cuatro funciones que corresponden a las tablas dimensionales, a las cuales se realizó las respectivas pruebas antes de la carga definitiva. Dichas funciones deben ser ejecutadas desde la base operacional, el orden de ejecución es indistinto y que acorde a sus características deben o no, ser ejecutadas para cada año.

Función cargar_d_tiempo().

Carga los datos de las tablas anio y mes, realizando las respectivas combinaciones, que llenan los datos en la tabla dimensional d_tiempo, ejecutando el comando select cargar_d_tiempo(), por una sola vez.

Función cargar_d_tiempo

```
CREATE OR REPLACE FUNCTION public.cargar_d_tiempo() RETURNS integer AS
$BODY$ DECLARE
-- Definicion de los cursores (anio, mes)
cur_anio cursor for
        select anio from anio
        where (anio) not in (select anio from d_tiempo)
        order by anio asc;
cur_mes cursor for
        select anio, codigo_mes, nombre_mes
        from anio, mes;

 -- Variables de extraccion de datos al recorrido de los cursores
 v_anio anio.anio%type;
 v_codigo_mes mes.codigo_mes%type;
 v_nombre_mes mes.nombre_mes%type;
 v_existe integer := 3;
begin
---------------- CARGA ANIO ----------------
   open cur_anio;
   -- carga los registros de anio a la tabla
   fetch cur_anio into v_anio;
   while found loop
     insert into d_tiempo (id, anio, mes,nombre_mes)
            values (nextval('d_tiempo_id_seq'), v_anio, null,null);
   -- El siguiente registro del cursor 1...
     fetch cur_anio into v_anio;
   end loop;
   if not found then
      raise notice 'Los años estan incluidos (%)',found;
      close cur_anio;
      v_existe := 1;
   end if;
---------------- CARGA MES ----------------
   open cur_mes;
   -- carga los registros de mes a la tabla
   fetch cur_mes into v_anio, v_codigo_mes, v_nombre_mes;
   while found loop
     insert into d_tiempo (id, anio, mes, nombre_mes)
            values (nextval('d_tiempo_id_seq'), v_anio, v_codigo_mes, v_nombre_mes),
     -- El siguiente registro del cursor 1...
     fetch cur_mes into v_anio, v_codigo_mes, v_nombre_mes;
   end loop;
   if not found then
      raise notice 'Los meses estan incluidos (%)',found;
      close cur_mes;
      v_existe := 2;
   end if;

return v_existe;
end;
$BODY$
  LANGUAGE plpgsql VOLATILE
  COST 100;
ALTER FUNCTION public.cargar_d_tiempo()
  OWNER TO postgres;
```

Figura 28. Función de carga para la dimensión d_tiempo.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

Prueba realizada. Para la comprobación de que los datos son correctos realizamos una suma que involucra los meses del año por cada año analizar, como se indica en la Figura Nº29.

Comparación de datos, tabla dimensional d_tiempo

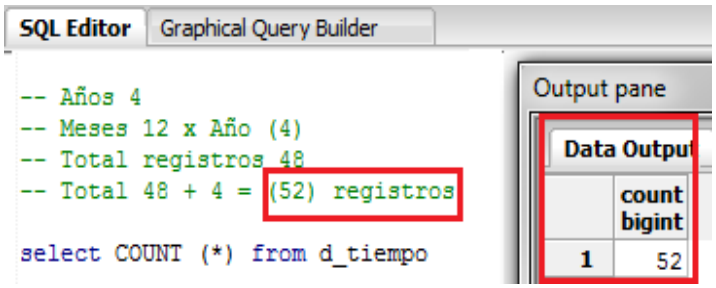


Figura 29. Número de datos cargados d_tiempo.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

Función cargar_d_entidad_salud().

Carga los datos combinados de las tablas sector, entidad, clase, cod_est de la base operacional, en la tabla dimensional d_entidad_salud ejecutando el comando *select cargar_d_entidad_salud(2011)* por cada año, el proceso llenara los datos correspondientes acorde al parámetro.

Prueba realizada. Para verificar los datos de la dimensión d_entidad_salud consultamos el total de registros de las tablas sector, tipo, entidad, clase y cod_est de la base operacional, el año 2011 suma un total de 831 registros.

Número de registros que conforman la dimensión d_entidad_salud

```
--    3 SECTORES
select count(*) from sector
 --  11 TIPO ESTABLECIMIENTO
select COUNT(*) from sector s, tipo tes,
(select distinct sector, tipo from egreso_2011_a ) egh
where cast(s.codigo_sector as numeric) =  egh.sector
and cast(tes.codigo_tipo as numeric) = egh.tipo
 --  28 CLASES ESTABLECIMIENTO
select COUNT(*) from sector s, tipo tes, clase ce,
(select distinct sector, tipo, clase from egreso_2011_a ) egh
where cast(s.codigo_sector as numeric) =  egh.sector
and cast(tes.codigo_tipo as numeric) = egh.tipo and ce.codigo_clase = egh.clase
--  789 ESTABLECIMIENTO DE SALUD
select count (*)from sector s, tipo tes, clase ce, cod_est esi,
(select distinct sector, tipo, clase, cod_est  from egreso_2011_a ) egh
where cast(s.codigo_sector as numeric) =  egh.sector
and cast(tes.codigo_tipo as numeric) = egh.tipo and ce.codigo_clase = egh.clase
and  esi.id_entidad = egh.cod_est
```

```
--TOTAL 3 + 11 + 28 + 789
831 Tablas BDD Operacional
```

Figura 30. Script para contabilizar el número de registros en tablas.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

Verificamos que el número de registros correspondientes al año 2011 de la tabla d_entidad_salud del almacén sea el mismo que de la base operacional, caso contrario se debe reevaluar la función y corregir errores.

Número de registros en la tabla dimensional d_entidad_salud.

```
-- TOTAL DIMENSION 831 registros
select COUNT(*) from d_entidad_salud
```

Output pane

Data Output | Explain | Messages | History

	count bigint
1	831

Figura 31. Número de registros en la tabla d_entidad_salud
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

Función cargar_d_geografia().

La función carga los datos de las tablas de provincia, cantón y parroquia, realizando las respectivas combinaciones, que llenan los datos en la tabla dimensional d_geografia. ejecutando select cargar_d_geografia().

Prueba realizada. Consultamos el número total de registros de las tablas provincia, cantón y parroquia de la tabla operacional su total lo comparamos con el número de registros de la tabla dimensiona d_geografia.

Verificación de registros de la tabla dimensional d_geografia

```
select count(*) from provincia --   26 provincias
select count(*) from canton    --  234 cantones
select count(*) from parroquia -- 1463 parroquias
     -- Total (26 + 234 + 1463) = 1723

select count(*) from d_geografia
```

Data Output | Explain | Messages | History

	count bigint
1	1723

registros de d_geografia

Figura 32. Número de registros en d_geografia.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

Función cargar_d_sintoma().

Esta función carga los datos combinando los datos de las tablas cap221rx, cau221rx y cau_cie10 de la base operacional, en la tabla dimensional d_sintoma, ejecutando para cada año select cargar_d_sintoma(2011).

Prueba realizada. Consultamos el número total de registros de las tablas cap221rx, cau221rx y cau_cie10 de la tabla operacional, su total lo comparamos con el número de registros de la tabla dimensional d_sintoma.

Verificación de registros de la tabla dimensional d_sintoma

```
-- Total cap221rx (21)
select COUNT(*) from cap221rx

-- Total cau221rx (221)
select COUNT(*)from cap221rx cap, cau221rx cau,
(select distinct cap221rx, cau221rx from egreso_2011_a) egh
where cast(cap.codigo_cap221rx as numeric) =  egh.cap221rx
and cast(cau.codigo_cau221rx as numeric) = egh.cau221rx

-- Total cau_cie10 (6571)
select COUNT(*)from cap221rx cap, cau221rx cau, cau_cie10 cie,
(select distinct cap221rx, cau221rx, cau_cie10 from egreso_2011_a) egh
where cast(cap.codigo_cap221rx as numeric) =  egh.cap221rx
and cast(cau.codigo_cau221rx as numeric) = egh.cau221rx
and cie.id =  egh.cau cie10

-- Total (21 + 221 + 6571) = 6813

select COUNT(*) from d_sintoma
```

Data Output | Explain | Messages | History

	count bigint
1	6813

Registros de la tabla d_sintoma

Figura 33. Número de registros en d_sintoma.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

2.2.3 Sprint 4. Análisis de datos para la creación de la tabla de hechos.

Debemos tomar en cuenta que cada dimensión representa una medida en la tabla de hechos, la cual contiene información relevante para el negocio, resultado de la recopilación de datos de las dimensiones, a la que se aplica la minería de datos.

2.2.3.1 *Creación de la estructura para la tabla de hechos.*

La tabla de hechos a más de medidas dimensiones puede incluir medidas totales que brinden un mayor nivel de detalle de los datos disponibles. Para identificar dichos campos adicionales de la tabla, se implementa el proceso de extracción de conocimiento (KDD) de minería de datos.

Selección de datos (Minería).

Diseño Dimensional y tabla de hechos

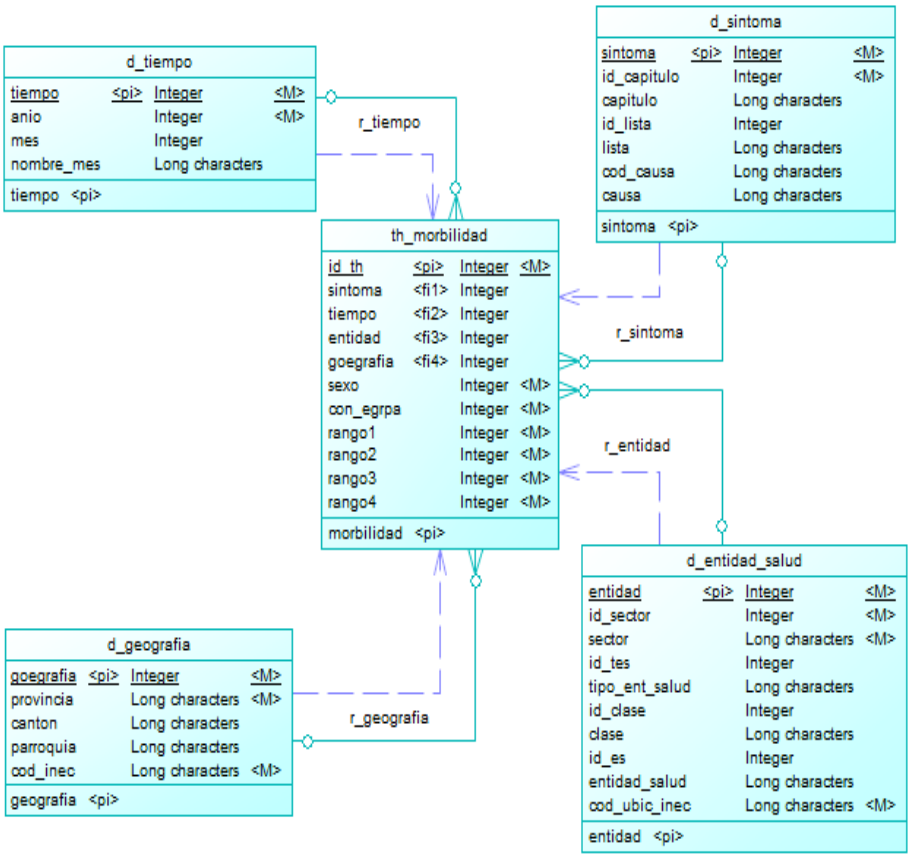


Figura 34. Estructura de la tabla de hechos.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

Se realizaron tres entrevistas con personas expertas en gestión de entidades de salud y seguimiento estadístico: Director del centro de salud tipo B en Latacunga; Asistente de dirección entidad de salud privada; Representante de la sección de estadística del MSP Coordinación Zonal Número 9.

Entrevistas donde concluimos que son necesarios los campos sexo, con_egrpa y rangos de edad de los pacientes comprendidos en rango 1 de 0 a 5 años, rango 2 de 6 a 17 años, rango 3 de 18 a 65 años y rango 4 de 66 años en adelante, para el análisis de datos y posterior elaboración de reportes, generando como resultado la estructura de la Figura Nº34.

Una vez definidos los campos de la tabla de hechos, procedemos crearla en la base AlmacenVCA, para la carga respectiva de los datos.

Transformación (Minería de datos).

Los campos extra sexo y con_egrpa que añadimos a la tabla de hechos son datos existentes en la base operacional que no necesitan de transformación, solo es necesario realizar la conexión entre dicha base y el almacén, así como, la creación de las tablas foráneas correspondientes.

Script de mapeo de tablas ejecutado en la base EgresoVCA

```
-- 1. Instancia de Conexion de Posgresql
CREATE EXTENSION postgres_fdw;

-- 2. Creación del servidor
CREATE SERVER egresoVCA_server
FOREIGN DATA WRAPPER postgres_fdw
OPTIONS (host 'localhost', port '5432',dbname 'EgresoVCA');

-- 3. Usuarios con permiso para acceder a la BDD y
--    direccionamiento del mapeo de datos
CREATE USER MAPPING FOR postgres
SERVER egresoVCA_server
OPTIONS (user 'postgres', password '123456');
```

Figura 35. Script de mapeo de tablas.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

El mapeo de las tablas, son necesarios para la elaboración de consultas, análisis y presentación de reportes que se realizan desde el almacén.

Los campos de rangos de edad son procesados de diferente forma, ya que para su obtención se realizó una comparación, contabilizando la cantidad de registros del campo edad de las tablas principales de la base operacional.

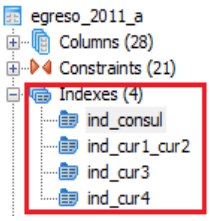
Obtención de datos del rango de edad (0 a 5 años).

```
-- RANGO EDAD 1
query:= 'SELECT count(*) FROM egreso_' || v_anio || '_a
        WHERE con_egrpa = ' || v_con_egrpa || '
        AND sexo = ' || v_sexo || ' AND edad >= 0 AND edad <=5
        AND anio_egr = ' || v_anio_dtiempo || '
        AND mes_egr = ' || v_mes_dtiempo || '
        AND cast(parr_ubi as numeric) = ' || v_parroquia || '
        AND cast(cod_est as numeric) = ' || v_id_es || '
        AND cau_cie10  = ' || chr(39) || v_cod_causa || chr(39);
raise NOTICE 'Rango 1 : %', query;
EXECUTE query into num_rango1;
```

Figura 36. Script para obtener el rango de edad entre 0 a 5años.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

Adicional se crearon índices para acelerar la carga de datos en la tabla de hechos, reduciendo considerablemente el tiempo de carga, estos índex corresponden a los datos que recorren los cursores. Se realiza un índex por cada cursor, y a su vez por cada año a analizar de la base operacional.

Índices creados en las tablas principales de la base operacional.




```
CREATE INDEX ind_consul ON egreso_2011_a (con_egrpa, sexo,
edad, anio_egr, mes_egr, parr_ubi, cod_est, cau_cie10);

CREATE INDEX ind_cur1_cur2 ON egreso_2011_a (mes_egr);

CREATE INDEX ind_cur3 ON egreso_2011_a (mes_egr, parr_ubi);

CREATE INDEX ind_cur4 ON egreso_2011_a (mes_egr,
parr_ubi,cod_est);
```

Figura 37. Índices creados en la tabla principal egreso_2011_a
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

Prueba realizada. Se realizó pruebas de carga de registros en tablas temporales por meses, al ejecutar las consultas el tiempo de respuesta fue de 4 a 8 minutos, con la creación de índices se redujo el tiempo de respuesta considerablemente.

Además, los índices aceleraron la carga de datos, ejecutamos cargar_hechos(2011) sin índices, en 3 días y medio el proceso no finalizaba así que lo detuvimos, ejecutamos el mismo proceso ya con índices y el proceso finalizo en 13 horas, 37 minutos y 28 segundos.

Tiempos de ejecución de consulta de información en tablas temporales

```
select COUNT(*) from egreso_2011_a where  anio_egr = 2011 AND mes_egr = 1 -- 92.570
select COUNT(*) from egreso_2011_a where  anio_egr = 2011 AND mes_egr = 2 -- 88.022
select COUNT(*) from egreso_2011_a where  anio_egr = 2011 AND mes_egr = 3 -- 97.244
-----------------------------------------------------------------------------------
-----------------------------------------------------------------------------------
-- 3:22 min (antes 8:12)
select SUM(rango1),SUM(rango2),SUM(rango3),SUM(rango4) from tmp_datos_2 -- 187.112

select COUNT(*) from egreso_2011_a where  anio_egr = 2011 AND mes_egr = 4 -- 92.428
select COUNT(*) from egreso_2011_a where  anio_egr = 2011 AND mes_egr = 5 -- 94.684

-----------------------------------------------------------------------------------
-----------------------------------------------------------------------------------
-- 3:17min (antes 7;14)
select SUM(rango1),SUM(rango2),SUM(rango3),SUM(rango4) from tmp_datos_2 -- 191.486

select COUNT(*) from egreso_2011_a where  anio_egr = 2011 AND mes_egr = 6 -- 94.819
select COUNT(*) from egreso_2011_a where  anio_egr = 2011 AND mes_egr = 7 -- 96.667

-----------------------------------------------------------------------------------
-----------------------------------------------------------------------------------
-- 3:22min (antes 5:14)
select SUM(rango1),SUM(rango2),SUM(rango3),SUM(rango4) from tmp_datos_2 -- 192.981

select COUNT(*) from egreso_2011_a where  anio_egr = 2011 AND mes_egr = 8 -- 96.367
select COUNT(*) from egreso_2011_a where  anio_egr = 2011 AND mes_egr = 9 -- 96.614
```

Figura 38. Tiempo de ejecución de consultas.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

2.2.3.2 *Carga de información en la Tabla de Hechos.*

La carga de datos a la tabla de hechos se realiza mediante procedimientos almacenados como en el caso anterior, la diferencia es que el query empleado es más complejo y exige más tiempo y recursos de la máquina. A continuación, se detalla la función creada.

Damos un nombre a la función e indicamos que va a recibir un parámetro de tipo integer que corresponde al año de carga de la información.

Parte 1 de 6: Cabecera de la función cargar_hechos().

```
CREATE OR REPLACE FUNCTION public.cargar_hechos(anio integer)
  RETURNS integer AS
$BODY$
 DECLARE
```

Figura 39. Cabecera de la función cargar_hechos().
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

Declaramos todas las variables a utilizar para extracción de datos y parametrización junto con el tipo de dato que almacenara cada registro de las tablas dimensionales.

Parte 2 de 6: Declaración de variables de la función cargar_hechos().

```
-- Variables de inicio y parametrizacion de nombre de la tabla
v_anio integer := $1;
x integer :=1;
query text; -- para rangos de edad
num_rango1 integer;
num_rango2 integer;
num_rango3 integer;
num_rango4 integer;

-- Variables de extraccion de datos al recorrido del cursor
-- Variable para el tiempo
v_id_dtiempo d_tiempo.id%type;
v_anio_dtiempo d_tiempo.anio%type;
v_mes_dtiempo d_tiempo.mes%type;

-- sexo, condicion egreso paciente
v_sexo integer;
v_con_egrpa integer;

-- Variable para la geografia
v_id_geografia d_geografia.id%type; -- cur_provincia
v_provincia d_geografia.cod_inec%type; -- cur_provincia
v_parroquia d_geografia.cod_inec%type; -- cur_provincia
v_canton d_geografia.cod_inec%type; -- cur_provincia canton

-- Variable para sector, entidad, clase, tipo etidad salud, entidad salud
v_id_entidad d_entidad_salud.id%type; -- cur_entidad
v_id_sector d_entidad_salud.id_sector%type; -- cur_entidad
v_id_tes d_entidad_salud.id_tes%type; -- cur_entidad
v_id_clase d_entidad_salud.id_clase%type; -- cur_entidad
v_id_es d_entidad_salud.id_es%type; -- cur_entidad

-- Variable para sintomas
v_id_sintoma d_sintoma.id%type; --cur_sintoma
v_id_capitulo d_sintoma.id_capitulo%type; --cur_sintoma
v_id_lista d_sintoma.id_lista%type; --cur_sintoma
v_cod_causa d_sintoma.cod_causa%type; --cur_sintoma
```

Figura 40. Variables utilizadas en la función.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

Parte 3 de 6: Declaración de cursores de la función cargar_hechos().

```
-- Definicion cursores
cur_sexo refcursor;
cur_codigoEgreso refcursor;
cur_anio_mes refcursor;
cur_provincia refcursor;
cur_entidad refcursor;
cur_sintoma refcursor;
```

Figura 41. Cursores utilizados en la función.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

Abrimos los cursores declarados, se arma la consulta considerando todas las variables que forman parte de las diferentes dimensiones, así como los parámetros de filtro de información que recorrerá el cursor.

Parte 4 de 6: Apertura de cursores declarados.

```
begin
-- LLENA TABLA POR AÑO-MES (SEXO, CODIGO_EGRESO Y RANGO 1-4)
open cur_anio_mes for execute 'select id, anio, mes from d_tiempo where anio = ' || v_anio ||'
            and mes IN (select distinct mes_egr from egreso_' || v_anio || '_a) order by id;';
fetch cur_anio_mes into v_id_dtiempo, v_anio_dtiempo, v_mes_dtiempo;
while found loop

   ------- Cursor dimencion Provincia (2), Canton (4), Parroquia (6)
   open cur_provincia for execute 'select id, cod_inec from d_geografia
            where cod_inec IN (select distinct parr_ubi from egreso_' || v_anio || '_a
            where mes_egr = ' || v_mes_dtiempo ||' ) order by id ';
   fetch cur_provincia into v_id_geografia, v_parroquia;
   while found loop

     -- Cursor dimencion Entidad_Salud (Sector,Tipo,Clase,Entidad y CodigoInec)
     open cur_entidad  for execute 'select id,id_es from d_entidad_salud
     where id_es IN (select distinct cod_est from egreso_' || v_anio || '_a
     where mes_egr = ' || v_mes_dtiempo ||'
     AND cast(parr_ubi as numeric) = '|| v_parroquia ||') order by id;';
     fetch cur_entidad into v_id_entidad, v_id_es;
     while found loop

       -- Cursor dimencion Sintoma
       open cur_sintoma  for execute 'select id,cod_causa from d_sintoma
       where cod_causa IN (select  distinct cau_cie10 from egreso_' || v_anio || '_a
       where mes_egr = ' || v_mes_dtiempo ||'
       and cast(parr_ubi as numeric) = '|| v_parroquia || '
       and cast(cod_est as numeric)= '|| v_id_es || ') order by id;';
       fetch cur_sintoma into v_id_sintoma, v_cod_causa;
       while found loop

             -- Cursor Sexo (1,2)
             open cur_sexo  for execute 'select codigo_sexo from sexo order by codigo_sexo;';
             fetch cur_sexo into v_sexo;
             while found loop

               -- Cursor condicion egreso paciente (1,2,3)
               open cur_codigoEgreso  for execute 'select codigo_con_egrpa
               from con_egrpa order by codigo_con_egrpa;';
               fetch cur_codigoEgreso into v_con_egrpa;
               while found loop
```

Figura 42. Apertura de los cursores para cargar los datos.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo

En la Figura Nº43, se indica la realización del cálculo de los respectivos rangos de edad, de igual forma considerando las variables que componen las dimensiones.

Parte 5 de 6: Cálculo de los rangos de edad.

```
-- RANGO EDAD 1
query:= 'SELECT count(*) FROM egreso_' || v_anio || '_a WHERE con_egrpa = ' || v_con_egrpa ||
        ' AND sexo = ' || v_sexo || ' AND edad >= 0 AND edad <=5
        AND anio_egr = ' || v_anio_dtiempo || ' AND mes_egr = ' || v_mes_dtiempo || '
        AND cast(parr_ubi as numeric) = ' || v_parroquia || '
        AND cast(cod_est as numeric) = ' || v_id_es || '
        AND cau_cie10  = ' || chr(39) || v_cod_causa || chr(39);
raise NOTICE 'Rango 1 : %', query;
EXECUTE query into num_rango1;

-- RANGO EDAD 2
query:= 'SELECT count(*) FROM egreso_' || v_anio || '_a WHERE con_egrpa = ' || v_con_egrpa ||
      ' AND sexo = ' || v_sexo || ' AND edad >= 6 AND edad <=17
        AND anio_egr = ' || v_anio_dtiempo || ' AND mes_egr = '|| v_mes_dtiempo || '
        AND cast(parr_ubi as numeric) = '|| v_parroquia || '
        AND cast(cod_est as numeric) = '|| v_id_es || '
        AND cau_cie10  = ' || chr(39) || v_cod_causa || chr(39);
--raise NOTICE 'Rango 2 : %', query;
EXECUTE query into num_rango2;

--RANGO EDAD 3
query:= 'SELECT count(*) FROM egreso_' || v_anio || '_a WHERE con_egrpa = ' || v_con_egrpa ||
        ' AND sexo = ' || v_sexo || ' AND edad >= 18 AND edad <=65
        AND anio_egr = ' || v_anio_dtiempo || ' AND mes_egr = '|| v_mes_dtiempo || '
        AND cast(parr_ubi as numeric) = '|| v_parroquia || '
        AND cast(cod_est as numeric) = '|| v_id_es || '
        AND cau_cie10  = ' || chr(39) || v_cod_causa || chr(39);
--raise NOTICE 'Rango 3 : %', query;
EXECUTE query into num_rango3;

--RANGO EDAD 4
query:= 'SELECT count(*) FROM egreso_' || v_anio || '_a WHERE con_egrpa = ' || v_con_egrpa ||
' AND sexo = ' || v_sexo || ' AND edad >= 66
        AND anio_egr = ' || v_anio_dtiempo || ' AND mes_egr = '|| v_mes_dtiempo || '
        AND cast(parr_ubi as numeric) = '|| v_parroquia || '
        AND cast(cod_est as numeric) = '|| v_id_es || '
        AND cau_cie10  = ' || chr(39) || v_cod_causa || chr(39);
--raise NOTICE 'Rango 4 : %', query;
EXECUTE query into num_rango4;
```

Figura 43. Script para el cálculo de los rangos de edad.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

Finalmente se procede a la inserción de datos en la tabla th_morbilidad, cierre de cursores, notificación de datos ingresados y cierre de la función.

La duración del proceso de carga de información en la taba de hechos fue entre 13 y 17 horas por cada año, en una computadora con características: Intel Core I7 con 3.860GHz y 8 GB en RAM.

Parte 6.1 de 6: Inserción de datos en la base y cierre de cursores.

```
            --Inserta los resultados
            INSERT INTO th_morbilidad VALUES (v_id_dtiempo,v_id_geografia,v_id_entidad,
            v_id_sintoma,v_sexo,v_con_egrpa,num_rango1,num_rango2,num_rango3,num_rango4);
         -- Cierre del cursor condicion egreso paciente
         fetch cur_codigoEgreso into v_con_egrpa;
         end loop;
         close cur_codigoEgreso;
      -- Cierre del cursor sexo
      fetch cur_sexo into  v_sexo;
      end loop;
      close cur_sexo;
   --Cierre del cursor sintoma
   fetch cur_sintoma into v_id_sintoma, v_cod_causa;
   end loop;
   close cur_sintoma;
-- Cierre del cursor entidad-salud
fetch cur_entidad into v_id_entidad, v_id_es;
end loop;
close cur_entidad;
```

Figura 44. Inserción de datos y cierre de cursores.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

Parte 6.2 de 6: Inserción de datos en la base y cierre de cursores.

```
      -- Cierre del cursor provincia
      fetch cur_provincia into v_id_geografia, v_parroquia;
      end loop;
      close cur_provincia;
  -- Cierre del cursor anio-mes
  fetch cur_anio_mes into  v_id_dtiempo, v_anio_dtiempo, v_mes_dtiempo;
  end loop;
  if not found then
  raise notice 'Todos los datos estan incluidos (%)',v_anio_dtiempo;
  close cur_anio_mes;
  end if;

return x;
end;
$BODY$
  LANGUAGE plpgsql VOLATILE
  COST 100;
ALTER FUNCTION public.cargar_hechos(integer)
  OWNER TO postgres;
```

Figura 45. Cierre de cursores, finalización de la función.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

La carga de información en la tabla de hechos estará completa una vez se ejecute ***select cargar_hechos(año)*** para cada año. La visualización de información debido a la pesada carga se tarda unos minutos.

Tabla de datos resultantes de la ejecución de la función.

Edit Data - PostgreSQL 9.5 (localhost:5432) - AlmacenVCA - public.th_morbilidad

File Edit View Tools Help

No limit

	tiempo integer	geografia integer	entidad integer	sintoma integer	sexo integer	con_egrpa integer	rango1 bigint	rango2 bigint	rango3 bigint	rango4 bigint
1	5	788	98	5528	2	2	0	0	0	0
2	5	788	98	5528	2	3	0	0	0	0
3	5	788	98	5532	1	1	0	0	2	0
4	5	788	98	5532	1	2	0	0	0	0
5	5	788	98	5532	1	3	0	0	0	0
6	5	788	98	5532	2	1	0	0	11	3
7	5	788	98	5532	2	2	0	0	0	0
8	5	788	98	5532	2	3	0	0	0	0
9	5	788	98	5535	1	1	0	0	0	0
10	5	788	98	5535	1	2	0	0	0	0
11	5	788	98	5535	1	3	0	0	0	0
12	5	788	98	5535	2	1	0	0	1	0
13	5	788	98	5535	2	2	0	0	0	0
14	5	788	98	5535	2	3	0	0	0	0
15	5	788	98	5536	1	1	0	0	1	0
16	5	788	98	5536	1	2	0	0	0	0
17	5	788	98	5536	1	3	0	0	0	0
18	5	788	98	5536	2	1	0	0	0	0
19	5	788	98	5536	2	2	0	0	0	0
20	5	788	98	5536	2	3	0	0	0	0
21	5	788	98	5537	1	1	0	0	5	0
22	5	788	98	5537	1	2	0	0	1	0
23	5	788	98	5537	1	3	0	0	0	1
24	5	788	98	5537	2	1	0	0	2	0
25	5	788	98	5537	2	2	0	0	0	0
26	5	788	98	5537	2	3	0	0	0	0
27	5	788	98	5539	1	1	0	0	0	0
28	5	788	98	5539	1	2	0	0	1	0
29	5	788	98	5539	1	3	0	0	0	1
30	5	788	98	5539	2	1	0	0	0	0

Figura 46. Datos que forma la tabla de hechos.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

Prueba realizada. Consultamos el número total de registros de la base operacional correspondientes al año 2011 (egresos_2011_a), lo comparamos con la sumatoria de los registros del almacén de datos (th_morbilidad), acorde a los rangos de edad. Si los resultados no coinciden debe reevaluarse la función de carga y corregir los errores.

Sumatoria de registros por rangos de edad.

```
-- Total de egresos para el año 2011 (1133325 registros)
select COUNT(*) from egreso_2011_a where anio_egr = 2011

-- Total de hechos por cada uno de los rangos
select * from d_tiempo where anio = 2011 and mes IS NOT NULL

select SUM(rango1),SUM(rango2),SUM(rango3),SUM(rango4) -- Sumatoria de los años
from th_morbilidad
where tiempo IN (select id from d_tiempo where anio = 2011 and mes IS NOT NULL)

-- Rango 1 = 130269
-- Rango 2 = 143500
-- Rango 3 = 730499
-- Rango 4 = 129057
```

--TOTAL = 1133325 registros

Total registros tabla th_morbilidad

Data Output | Explain | Messages | History

	count bigint
1	1133325

Total registros tabla egreso_2011_a

Figura 47. Comparación de registros por rangos de edad.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

La sección correspondiente Conocimiento, se tratará en el siguiente capítulo de Análisis, finalizando los procesos de elaboración del almacenamiento y minería de datos.

2.3 Diccionario de Datos

El objetivo del diccionario de datos es contener información precisa de los datos que se encuentran en las respectivas tablas de la base, indicando la información que cada campo representa, para evitar ambigüedades o malas interpretaciones entre los datos empleada.

2.3.1 Diccionario de la base operacional (EgresoVCA).

El diccionario de la base operacional se describe a continuación en orden alfabético, dejando para el final la tabla principal.

Tabla 16. Diccionario de datos de la tabla anio

Campo	Nombre	Tipo	Campo Obligatorio (S/N)	Restricción
anio	Identificador	Entero	S	Clave Primaria
Restricción				
Nombre	**Tipo**	**Definición**		
pk_anio	Primary key	(anio)		

Nota: Información de los años a analizar.

Tabla 17. Diccionario de datos de la tabla canton.

Campo	Nombre	Tipo	Campo Obligatorio (S/N)	Restricción
codigo_provincia	Código de provincia	Carácter(2)	S	Clave Primaria
codigo_canton	Código de cantón	Carácter(2)	S	Clave Primaria
codigo_canton_inec	Código INEC del cantón	Carácter(4)	S	Clave Primaria
Nombre_canton	Nombre del cantón	Carácter	N	
Restricción				
Nombre	**Tipo**	**Definición**		
pk canton	Clave Primaria	(codigo_provincia, codigo_canton, codigo_canton_inec)		
uk_canton	Clave Única	codigo_provincia, codigo_canton, codigo_canton_inec)		
uk canton_codigo _inec	Clave Única	(codigo_canton_inec)		
uk_canton_prov	Clave Única	(codigo_provincia, codigo_canton)		

Nota: Información del código y nombres de los cantones del Ecuador, conjuntamente con el código de la provincia a la que pertenecen.

Tabla 18. Diccionario de datos de la tabla cap221rx.

Campo	Nombre	Tipo	Campo Obligatorio (S/N)	Restricción
codigo_cap221rx	Código del capítulo de la lista 221	Entero	S	Clave Primaria
nombre_cap221rx	Nombre cap. de la lista 221	Caracter(150)	N	
Restricción				
Nombre	**Tipo**	**Definición**		
pk_cap221rx	Clave Primaria	(codigo_cap221rx)		

Nota: Información del identificador y nombres de los capítulos de la lista 221 de problemas relacionados con la salud.

Tabla 19. Diccionario de datos de la tabla cau221rx

Campo	Nombre	Tipo	Campo Obligatorio (S/N)	Restricción
codigo_cau221rx	Código de la lista especial de 221 grupos	Entero	S	Clave Primaria
nombre_ cau221rx	Nombre de la lista especial de 221 grupos	Caracter(150)	N	
Restricción				
Nombre	**Tipo**	**Definición**		
pk_cau221rx	Clave Primaria	(codigo_cau221rx)		

Nota: Información del identificador y nombres de la lista especial de 221 grupos de problemas relacionados con la salud.

Tabla 20. Diccionario de datos de la tabla cau298rx.

Campo	Nombre	Tipo	Campo Obligatorio (S/N)	Restricción
codigo_cau298rx	Código de lista de 298 causas	Entero	S	Clave Primaria
nombre_ cau298rx	Nombre de lista 298 causas	Caracter(150)	N	
Restricción				
Nombre	**Tipo**	**Definición**		
pk_cau298rx	Clave Primaria	(codigo_cau298rx)		

Nota: Información del identificador y nombres de la lista de 298 causas de problemas relacionados con la salud.

Tabla 21. Diccionario de datos de la tabla cau_cie10.

Campo	Nombre	Tipo	Campo Obligatorio (S/N)	Restricción
Id	Código causa morbilidad	Caracter(4)	S	Clave Primaria
descripcion	Descripción causa morbilidad	Caracter(300)	N	
Restricción				
Nombre	**Tipo**	**Definición**		
pk_cau_cie10	Clave Primaria	(id)		

Nota: Información del código y nombres de las causas de morbilidad de la 10ma revisión, en Ecuador.

Tabla 22. Diccionario de datos de la tabla clase.

Campo	Nombre	Tipo	Campo Obligatorio (S/N)	Restricción
codigo_clase	Código causa morbilidad	Entero	S	Clave Primaria
nombre_clase	Nombre causa morbilidad	Caracter(50)	N	
Restricción				
Nombre	**Tipo**	**Definición**		
pk_clase	Clave Primaria	(codigo_clase)		

Nota: Información del identificador y nombres de las clases de establecimientos de salud.

Tabla 23. Diccionario de datos de la tabla cod_edad.

Campo	Nombre	Tipo	Campo Obligatorio (S/N)	Restricción
codigo_edad	Código condición edad	Entero	S	Clave Primaria
nombre_edad	Nombre condición edad	Caracter(30)	N	
Restricción				
Nombre	**Tipo**	**Definición**		
pk_cd_edad	Clave Primaria	(codigo_edad)		

Nota: Información del identificador y nombres de la condición de edad del paciente.

Tabla 24. Diccionario de datos de la tabla cod_est.

Campo	Nombre	Tipo	Campo Obligatorio (S/N)	Restricción
id_entidad	Identificador del establecimiento de salud	Entero	S	Clave Primaria
nombre_entidad	Nombre del establecimiento de salud	Caracter(100)	N	
prov	Código de la provincia del Estab.Salud.	Caracter(2)	S	
can	Código del cantón del E.S.	Caracter(2)	S	
parr	Código de la parroquia del E.S.	Caracter(2)	S	
sector	Identificador del sector del establecimiento.	Entero	S	
tipo	Identificador del tipo de establecimiento.	Entero	S	
clase	Identificador de la clase del establecimiento.	Entero	S	
Restricción				
Nombre	**Tipo**	**Definición**		
pk_con_est	Clave Primaria	(id_entidad)		
uk_cod_est	Clave Unica	(id_entidad, prov, can, parr)		

Nota: Información del identificador y nombre del establecimiento de salud, además del código de la provincia, cantón y parroquia de su ubicación, así como el tipo, sector y clase de entidad la que pertenece.

Tabla 25. Diccionario de datos de la tabla entidad.

Campo	Nombre	Tipo	Campo Obligatorio (S/N)	Restricción
codigo_entidad	Código de la entidad del establecimiento	Entero	S	Clave Primaria
nombre_entidad	Nombre de la entidad del establecimiento	Caracter(50)	N	
Restricción				
Nombre	**Tipo**	**Definición**		
pk_entidad	Clave Primaria	(codigo_entidad)		

Nota: Información del identificador de la entidad del establecimiento de salud.

Tabla 26. Diccionario de datos de la tabla mes.

Campo	Nombre	Tipo	Campo Obligatorio (S/N)	Restricción
codigo_mes	Código del mes del año	Entero	S	Clave Primaria
nombre_mes	Nombre del mes del año	Caracter(10)	N	
Restricción				
Nombre	**Tipo**	**Definición**		
pk_mes	Clave Primaria	(codigo_mes)		

Nota: Información del identificador y nombre del mes del año.

Tabla 27. Diccionario de datos de la tabla parroquia.

Campo	Nombre	Tipo	Campo Obligatorio (S/N)	Restricción
codigo_provincia	Código de provincia	Caracter(2)	S	Clave Primaria
codido_canton	Código de cantón	Caracter(2)	S	Clave Primaria
codigo_parroquia	Código de parroquia	Caracter(2)	S	Clave Primaria
codigo_parroquia _inec	Código INEC	Caracter(6)	S	Clave Primaria
nombre_parroquia	Nombre de la parroquia	Caracter	N	
Restricción				
Nombre	**Tipo**	**Definición**		
pk_parroquia	Clave Primaria	(id)		
uk_parroquia	Clave Unica	(cod_provincia, cod_canton, cod_parroquia)		
uk_parroquia _codinec	Clave Unica	(cod_inec)		
uk_parroquia_pcp	Clave Unica	(cod_provincia, cod_canton, parroquia)		

Nota: Información del código y nombres de las parroquias del Ecuador, conjuntamente con el código de la provincia y cantón a las que pertenecen.

Tabla 28. Diccionario de datos de la tabla provincia.

Campo	Nombre	Tipo	Campo Obligatorio (S/N)	Restricción
codigo_provincia	Código de provincia	Caracter(2)	S	Clave Primaria
nombre_provincia	Nombre de la provincia	Caracter(2)	N	
Restricción				
Nombre	**Tipo**	**Definición**		
pk_codigo_provincia	Clave Primaria	(codigo_provincia)		
uk_ provincia	Clave Unica	(codigo_provincia)		

Nota: Información del código y nombres de las provincias del Ecuador.

Tabla 29. Diccionario de datos de la tabla sector.

Campo	Nombre	Tipo	Campo Obligatorio (S/N)	Restricción
codigo_sector	Código del sector establecimiento	Entero	S	Clave Primaria
nombre_sector	Nombre del sector establecimiento	Caracter(30)	N	
Restricción				
Nombre	**Tipo**	**Definición**		
pk_sector	Clave Primaria	(codigo_sector)		

Nota: Información del identificador y nombres de los sectores al que pertenecen los establecimientos.

Tabla 30. Diccionario de datos de la tabla sexo.

Campo	Nombre	Tipo	Campo Obligatorio (S/N)	Restricción
codigo_sexo	Código del sexo del paciente	Entero	S	Clave Primaria
nombre_sexo	Nombre del sexo del paciente	Caracter(10)	N	
Restricción				
Nombre	**Tipo**	**Definición**		
pk_sexo	Clave Primaria	(codigo_sexo)		

Nota: Información del identificador y nombre del sexo del paciente.

Tabla 31. Diccionario de datos de la tabla tipo.

Campo	Nombre	Tipo	Campo Obligatorio (S/N)	Restricción
codigo_tipo	Código del tipo de establecimiento	Entero	S	Clave Primaria
nombre_tipo	Nombre del tipo de establecimiento	Caracter(10)	N	
Restricción				
Nombre	**Tipo**	**Definición**		
pk_tipo	Clave Primaria	(codigo_tipo)		

Nota: Información del identificador y nombre del tipo de establecimiento de salud.

Tabla 32. Diccionario de datos de la tabla principal egreso_2011_a.

Campo	Nombre	Tipo	Campo Obligatorio (S/N)	Restricción
cod_est	Código del establecimiento de salud	Entero	S	
ani_inv	Año de registro de los datos.	Entero	S	
prov_ubi	Provincia de ubicación del establecimiento.	Caracter(2)	S	
cant_ubi	Cantón de ubicación del establecimiento.	Caracter(4)	S	
parr_ubi	Parroquia de ubicación del establecimiento.	Caracter(6)	S	
clase	Clase de establecimiento	Entero	S	
tipo	Tipo de establecimiento	Entero	S	
entidad	Entidad del establecimiento	Entero	S	
sector	Sector del establecimiento	Entero	S	
mes_inv	Mes de registro de los datos	Entero	S	
cod_edad	Condición de edad del paciente.	Entero	S	
edad	Edad del paciente	Entero	S	
sexo	Sexo del paciente	Entero	S	
prov_res	Provincia de residencia del paciente.	Caracter(2)	S	
cant_res	Cantón de residencia del paciente.	Caracter(4)	S	
parr_res	Parroquia de residencia del paciente.	Caracter(6)	S	
dia_ingr	Día de ingreso del paciente	Caracter(4)	S	
mes_ingr	Mes de ingreso del paciente	Caracter(6)	S	
anio_ingr	Año de ingreso del paciente	Entero	S	
dia_egr	Día de egreso del paciente	Entero	S	
mes_egr	Mes de egreso del paciente	Entero	S	
anio_egr	Año de egreso del paciente	Entero	S	
dia_estad	Indicador de servicios	Entero	S	
con_egrpa	Condición de egreso del paciente	Entero	S	
cau_cie10	Cusa morbilidad 10ma revisión.	Caracter(4)	S	
cau221rx	Lista especial de 221 grupos.	Entero	S	
cap221rx	Capítulo lista 221.	Entero	S	
Cau298rx	Lista de 298 causas.	Entero	S	

Restricción		
Nombre	**Tipo**	**Definición**
fk_cant_res_11_a	Clave Foranea	(codigo_canton_inec)
fk_cant_ubi_11_a	Clave Foranea	(codigo_canton_inec)
fk_cap221rx_11_a	Clave Foranea	(codigo_cap221rx)
fk_cau221rx_11_a	Clave Foranea	(codigo_cau221rx)
fk_cau298rx_11_a	Clave Foranea	(codigo_cau298rx)
fk_cau_cie10_11_a	Clave Foranea	(id)
fk_clase_11_a	Clave Foranea	(codigo_clase)
fk_cod_est_11_a	Clave Foranea	(id_entidad)
fk_con_edad_11_a	Clave Foranea	(codigo_edad)
fk_con_egrpa_11_a	Clave Foranea	(codigo_con_egrpa)
fk_entidad_11_a	Clave Foranea	(codigo_entidad)
fk_mes_egr_11_a	Clave Foranea	(codigo_mes)
fk_mes_ingr_11_a	Clave Foranea	(codigo_mes)
fk_mes_inv_11_a	Clave Foranea	(codigo_mes)
fk_parr_res_11_a	Clave Foranea	(codigo_parroquia_inec)
fk_parr_ubi_11_a	Clave Foranea	(codigo_parroquia_inec)
fk_prov_res_11_a	Clave Foranea	(codigo_provincia)
fk_prov_ubi_11_a	Clave Foranea	(codigo_provincia)
fk_sector_11_a	Clave Foranea	(codigo_sector)
fk_sexo_11_a	Clave Foranea	(codigo_sexo)
fk_tipo_11_a	Clave Foranea	(codigo_tipo)
ind_consul	Clave Unica	(con_egrpa, sexo, edad, anio_egr, mes_egr, parr_ubi pg_catalog. cod_est, cau_cie10
ind_cur1_cur2	Clave Unica	(mes_egr)
ind_cur3	Clave Unica	(mes_egr, parr_ubi)
ind_cur4	Clave Unica	(mes_egr, parr_ubi) ubi pg_catalog. cod_est

Nota: Información de año 2011, cuyos datos son resultado de las claves primarias de las tablas secundarias cuya definición se ha explicado.

La definición empleada en los campos que componen la tabla egreso_2011_a, se utiliza para las tablas correspondientes a los años 2012, 2013 y 2014, debido a que se estandarizo la información de modo que todas las tablas principales contienen los mismos campos, el mismo significado.

2.3.2 Diccionario de la base AlmacenVCA.

Tabla 33. Diccionario de datos de la tabla dimensional d_entidad_salud.

Campo	Nombre	Tipo	Campo Obligatorio (S/N)	Restricción
id	Identificador serial de tabla	Serial	S	Clave Primaria
sector	Tipo sector	Caracter	N	
ent_gest_salud	Entidad que gestiona	Caracter	N	
clase	Clase de entidad	Caracter	N	
entidad_salud	Nombre de la entidad de salud	Caracter	N	
id_sector	Identificador de tipo sector	Entero	N	
id_tes	Identificador de entidad que gestiona	Entero	N	
id_clase	Identificador de clase de entidad	Entero	N	
id_es	Identificador de la entidad de salud	Entero	N	
cod_ubic_inec	Código de ubicación INEC	Caracter (6)	N	
Restricción				
Nombre	**Tipo**	**Definición**		
pk_d_entidad _salud	Clave Primaria	(id)		
uk_ d_entidad	Clave Unica	(id_es)		

Nota: Información que relacionan el sector, tipo de entidad, clase y entidad d salud con su respectivo identificador y nombre, además del código de ubicación del INEC.

Tabla 34. Diccionario de datos de la tabla dimensional d_geografia

Campo	Nombre	Tipo	Campo Obligatorio (S/N)	Restricción
id	Identificador geográfico	Serial	S	Clave Primaria
provincia	Nombre de la provincia	Caracter	N	
cantón	Nombre del cantón	Caracter	N	
parroquia	Nombre de la parroquia	Caracter	N	
cod_inec	Código INEC	Caracter(6)	N	
Restricción				
Nombre	**Tipo**	**Definición**		
pk_id_geografia	Clave Primaria	(id)		
uk_d_geografia_codinec	Clave Unica	(cod_inec)		

Nota: Información que relaciona la provincia, el canto y parroquias del país, además del código de ubicación del INEC.

Tabla 35. Diccionario de datos de la tabla dimensional d_sintoma.

Campo	Nombre	Tipo	Campo Obligatorio (S/N)	Restricción
Id	Identificador serial de tabla	Serial	S	Clave Primaria
id_capitulo	Indicador del capítulo	Caracter	N	
capitulo	Descripción del capítulo	Caracter	N	
id_lista	Indicador de la lista	Caracter	N	
lista	Descripción de la lista	Caracter	N	
cod_causa	Código de la causas	Caracter	N	
causa	Descripción de la causas	Caracter	N	
Restricción				
Nombre	**Tipo**	**Definición**		
pk_d_sintoma	Clave Primaria	(id)		

Nota: Información que relaciona el capítulo de la lista 221, la lista especial de 221 grupos, y la causa de morbilidad de la 10ma revisión.

Tabla 36: Diccionario de datos de la tabla dimensional d_tiempo

Campo	Nombre	Tipo	Campo Obligatorio (S/N)	Restricción
id	Identificador serial de tabla	Serial	S	Clave Primaria
anio	Indicador del año	Entero	S	
mes	Indicador del mes	Entero	N	
nombre_mes	Descripción del mes	Caracter	N	
Restricción				
Nombre	**Tipo**	**Definición**		
pk_d_tiempo	Clave Primaria	(id)		

Nota: Información que relaciona el año y mes de ingreso de los registros.

2.3.3 Diccionario de la tabla de hechos.

Tabla 37. Tabla th_morbilidad.

Campo	Nombre	Tipo	Campo Obligatorio (S/N)	Restricción
tiempo	Identificador de la dimensión tiempo	Entero	S	
geografia	Identificador de la dimensión geografía	Entero	S	
entidad	Identificador de la dimensión entidad de salud	Entero	S	
sintoma	Identificador de la dimensión síntoma	Entero	S	
sexo	Número de especialistas de 8 horas	Entero	S	
con_egrpa	Identificador de la dimensión condición de egreso	Entero	S	
rango1	Rango de edad de 0 a 5 años.	Entero Largo	S	
rango2	Rango de edad de 6 a 17 años.	Entero Largo	S	
rango3	Rango de edad de 18 a 65 años.	Entero Largo	S	
rango4	Rango de edad mayo a 66 años.	Entero Largo	S	

Nota: Información que combina los datos de las dimensiones, el sexo y la condición de egreso del paciente, además contiene la cantidad de pacientes acorde a un rango de edad.

Capítulo 3
Análisis

3.1 Adquisición de Conocimiento

El proceso y construcción del almacén de datos está conformado por todos los datos que se recopilo de la fuente oficial del INEC de los últimos 4 años (2011, 2012, 2013 y 2014), relacionados con las enfermedades de la clasificación CIE-10, sin embargo el análisis realizado se centra en el capítulo XII correspondiente a enfermedades de la piel y el tejido subcutáneo, con código de identificación L.

Fase 5 Despliegue. Para la transformación de información en conocimiento, se realizó el análisis respectivo y presentación al usuario a manera de cifras y gráficos, utilizando la herramienta Pentaho, esto corresponde a la fase 5, despliegue.

3.1.1 Sprint 5. Definición de áreas de interés para adquirir conocimiento.

Identificación de la relación existente entre los datos, para la selección de criterios a analizar en función de los resultados de la minería, con el objeto de descubrir el conocimiento oculto en los grandes volúmenes de información.

3.1.1.1 *Vistas materializadas para el análisis de datos.*

Para el diseño y presentación del análisis de información a manera de reportes, que permitan realizar consultas dinámicas acorde al interés del usuario, utilizamos la herramienta Pentaho, para lo cual se crearon vistas materializadas, para los respectivos análisis, acorde a datos recolectados en entrevistas realizadas.

Una vista materializada es una tabla virtual que contiene el resultado de consultas a la base de datos, se actualiza periódicamente a partir de los datos que contienen las tablas de consulta, para la extracción de datos de forma más rápida, evitando la consulta de una tabla con millones de datos.

Vista materializada Enfermedades por Sexo y Edad. Esta vista contiene una parte de los datos que conforman la tabla de hechos (th_morbilidad), correspondientes a los campos de sexo, rangos de edad, causa por capítulo y año.

Vista materializada de enfermedad por sexo y edad

anio intege	capitulo character varying	causa character varying	sexo character	rang bigin	rang bigin	rang bigin	rango4 bigint
2013	TRAUMATISMOS, ENVENENAMIENTOS Y	TRAUMATISMO DE LA VE	MUJER	0	0	0	0
2011	TRAUMATISMOS, ENVENENAMIENTOS Y	TRAUMATISMO DE LA VE	MUJER	0	0	0	0
2013	TRAUMATISMOS, ENVENENAMIENTOS Y	TRAUMATISMO DE LA VE	HOMBRE	0	0	0	0
2011	TRAUMATISMOS, ENVENENAMIENTOS Y	TRAUMATISMO DE LA VE	HOMBRE	0	0	0	0
2013	TRAUMATISMOS, ENVENENAMIENTOS Y	TRAUMATISMO DE LA VE	HOMBRE	0	0	0	0
2011	TRAUMATISMOS, ENVENENAMIENTOS Y	TRAUMATISMO DE LA VE	HOMBRE	0	0	0	0
2013	TRAUMATISMOS, ENVENENAMIENTOS Y	TRAUMATISMO DE LA VE	MUJER	0	0	0	0
2011	TRAUMATISMOS, ENVENENAMIENTOS Y	TRAUMATISMO DE LA VE	HOMBRE	0	0	1	0
2013	TRAUMATISMOS, ENVENENAMIENTOS Y	TRAUMATISMO DE LA VE	HOMBRE	0	0	1	0
2011	TRAUMATISMOS, ENVENENAMIENTOS Y	TRAUMATISMO DE LA VE	MUJER	0	0	0	0
2013	TRAUMATISMOS, ENVENENAMIENTOS Y	TRAUMATISMO DE VENA	HOMBRE	0	0	0	0
2014	TRAUMATISMOS, ENVENENAMIENTOS Y	TRAUMATISMO DE VENA	HOMBRE	0	0	1	0
2011	TRAUMATISMOS, ENVENENAMIENTOS Y	TRAUMATISMO DE VENA	HOMBRE	0	0	0	0
2013	TRAUMATISMOS, ENVENENAMIENTOS Y	TRAUMATISMO DE VENA	HOMBRE	0	0	0	0
2013	TRAUMATISMOS, ENVENENAMIENTOS Y	TRAUMATISMO DE VENA	HOMBRE	0	0	1	0
2014	TRAUMATISMOS, ENVENENAMIENTOS Y	TRAUMATISMO DE VENA	HOMBRE	0	0	1	0

Figura 48. Datos de Enfermedades por Sexo y Edad.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

Vista materializada Enfermedades por Localidad. Vista que permite realizar consultas considerando la información de la tabla de hechos que corresponde a provincia, cantón, capítulo por año.

Vista materializada Enfermedades por Localidad

provincia character var	canton character va	parroquia character varying	causa character varying	sexo characte	rang bigin	rang bigin	rang bigin	rango4 bigint
GUAYAS	GUAYAQUIL	GUAYAQUIL, CABECI	CÓLERA DEBIDO A VIBRIO CH	MUJER	0	0	0	0
AZUAY	SIGSIG	SIGSIG, CABECERA	CÓLERA DEBIDO A VIBRIO CH	MUJER	0	0	0	0
GUAYAS	GUAYAQUIL	GUAYAQUIL, CABECI	CÓLERA DEBIDO A VIBRIO CH	MUJER	0	0	0	0
AZUAY	SIGSIG	SIGSIG, CABECERA	CÓLERA DEBIDO A VIBRIO CH	MUJER	0	0	0	0
GUAYAS	GUAYAQUIL	GUAYAQUIL, CABECI	CÓLERA DEBIDO A VIBRIO CH	HOMBRE	0	0	1	0
PICHINCHA	QUITO	QUITO , CABECERA	CÓLERA, NO ESPECIFICADO	MUJER	0	0	0	0
GUAYAS	GUAYAQUIL	GUAYAQUIL, CABECI	CÓLERA, NO ESPECIFICADO	HOMBRE	0	0	0	0
PICHINCHA	QUITO	QUITO , CABECERA	CÓLERA, NO ESPECIFICADO	MUJER	0	0	0	0
PICHINCHA	QUITO	QUITO , CABECERA	CÓLERA, NO ESPECIFICADO	MUJER	0	0	0	0
EL ORO	MACHALA	MACHALA , CABECEI	CÓLERA, NO ESPECIFICADO	MUJER	0	0	0	0
GUAYAS	GUAYAQUIL	GUAYAQUIL, CABECI	CÓLERA, NO ESPECIFICADO	MUJER	0	0	0	0
AZUAY	CUENCA	CUENCA , CABECERA	CÓLERA, NO ESPECIFICADO	MUJER	0	0	1	0
GUAYAS	MILAGRO	MILAGRO, CABECERA	CÓLERA, NO ESPECIFICADO	HOMBRE	0	0	0	0
LOJA	LOJA	LOJA , CABECERA (	CÓLERA, NO ESPECIFICADO	HOMBRE	0	0	0	0

Figura 49. Datos de Enfermedades por Localidad.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

Vista materializada Enfermedades por Entidad de Salud. En esta vista se han cargado datos correspondientes a la dimensión d_entidad, para la consulta de capítulos de enfermedades diagnosticadas por año en los diferentes sectores y clases de entidades de salud.

Vista materializada Enfermedades por Entidad de Salud

sector character varyi	tipo_entidad character varying	clase character varying	entidad character varying	causa character varying
PÚBLICO	AGUDO	HOSPITAL GENERAL	HOSPITAL LUIS VERNAZA	CÓLERA DEBIDO A VIBRIO
PÚBLICO	SIN TIPO HOSPITALES I	HOSPITAL BÁSICO	HOSPITAL SAN SEBASTIÁN	CÓLERA DEBIDO A VIBRIO
PÚBLICO	AGUDO	HOSPITAL GENERAL	HOSPITAL LUIS VERNAZA	CÓLERA DEBIDO A VIBRIO
PÚBLICO	SIN TIPO HOSPITALES I	HOSPITAL BÁSICO	HOSPITAL SAN SEBASTIÁN	CÓLERA DEBIDO A VIBRIO
PÚBLICO	AGUDO	HOSPITAL GENERAL	HOSPITAL LUIS VERNAZA	CÓLERA DEBIDO A VIBRIO
PÚBLICO	SIN TIPO HOSPITALES I	HOSPITAL BÁSICO	HOSPITAL SAN SEBASTIÁN	CÓLERA DEBIDO A VIBRIO
PÚBLICO	AGUDO	HOSPITAL GENERAL	HOSPITAL LUIS VERNAZA	CÓLERA DEBIDO A VIBRIO
PÚBLICO	SIN TIPO HOSPITALES I	HOSPITAL BÁSICO	HOSPITAL SAN SEBASTIÁN	CÓLERA DEBIDO A VIBRIO
PÚBLICO	AGUDO	HOSPITAL GENERAL	HOSPITAL LUIS VERNAZA	CÓLERA DEBIDO A VIBRIO
PÚBLICO	SIN TIPO HOSPITALES I	HOSPITAL BÁSICO	HOSPITAL SAN SEBASTIÁN	CÓLERA DEBIDO A VIBRIO
PÚBLICO	AGUDO	HOSPITAL GENERAL	HOSPITAL LUIS VERNAZA	CÓLERA DEBIDO A VIBRIO
PRIVADO CON F	CLÍNICAS GENERALES S	CLÍNICA GENERAL (SIN ES	CLÍNICA VIRGEN DE SCHOENTAH	CÓLERA, NO ESPECIFICAD
PÚBLICO	AGUDO	HOSPITAL GENERAL	HOSPITAL PROVINCIAL DEL I.E.S	CÓLERA, NO ESPECIFICAD
PÚBLICO	AGUDO	HOSPITAL GENERAL	HOSPITAL DEL I.E.S.S.	CÓLERA, NO ESPECIFICAD
PRIVADO CON F	AGUDO	HOSPITAL DE ESPECIALIDA	HOSPITAL METROPOLITANO	CÓLERA, NO ESPECIFICAD
PÚBLICO	AGUDO	HOSPITAL GENERAL	HOSPITAL LUIS VERNAZA	CÓLERA, NO ESPECIFICAD
PRIVADO CON F	CLÍNICAS GENERALES S	CLÍNICA GENERAL (SIN ES	CLÍNICA SANTA MARÍA	CÓLERA, NO ESPECIFICAD

Figura 50. Datos de Enfermedades por Entidad de Salud.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

Vista materializada de Enfermedades. Vista utilizada para el análisis de datos en Matlab con el modelo de regresión lineal, para la consulta total por año y mes de los casos registrados, el sexo de paciente y causa de la enfermedad.

Datos para el método de regresión lineal.

	causa character varying	id integer	anio integer	mes integer	sexo integer	total numeric
1	ÚLCERA DE MIEMBRO INFERIOR, NO CLASIFICADA EN OTRA PARTE	3535	2011	1	1	8
2	ÚLCERA DE MIEMBRO INFERIOR, NO CLASIFICADA EN OTRA PARTE	3535	2011	1	2	10
3	ÚLCERA DE MIEMBRO INFERIOR, NO CLASIFICADA EN OTRA PARTE	3535	2011	2	1	7
4	ÚLCERA DE MIEMBRO INFERIOR, NO CLASIFICADA EN OTRA PARTE	3535	2011	2	2	7
5	ÚLCERA DE MIEMBRO INFERIOR, NO CLASIFICADA EN OTRA PARTE	3535	2011	3	1	10
6	ÚLCERA DE MIEMBRO INFERIOR, NO CLASIFICADA EN OTRA PARTE	3535	2011	3	2	12
7	ÚLCERA DE MIEMBRO INFERIOR, NO CLASIFICADA EN OTRA PARTE	3535	2011	4	1	10
8	ÚLCERA DE MIEMBRO INFERIOR, NO CLASIFICADA EN OTRA PARTE	3535	2011	4	2	7
9	ÚLCERA DE MIEMBRO INFERIOR, NO CLASIFICADA EN OTRA PARTE	3535	2011	5	1	13
10	ÚLCERA DE MIEMBRO INFERIOR, NO CLASIFICADA EN OTRA PARTE	3535	2011	5	2	7
11	ÚLCERA DE MIEMBRO INFERIOR, NO CLASIFICADA EN OTRA PARTE	3535	2011	6	1	11
12	ÚLCERA DE MIEMBRO INFERIOR, NO CLASIFICADA EN OTRA PARTE	3535	2011	6	2	9
13	ÚLCERA DE MIEMBRO INFERIOR, NO CLASIFICADA EN OTRA PARTE	3535	2011	7	1	8
14	ÚLCERA DE MIEMBRO INFERIOR, NO CLASIFICADA EN OTRA PARTE	3535	2011	7	2	5
15	ÚLCERA DE MIEMBRO INFERIOR, NO CLASIFICADA EN OTRA PARTE	3535	2011	8	1	16
16	ÚLCERA DE MIEMBRO INFERIOR, NO CLASIFICADA EN OTRA PARTE	3535	2011	8	2	12
17	ÚLCERA DE MIEMBRO INFERIOR, NO CLASIFICADA EN OTRA PARTE	3535	2011	9	1	6

Figura 51. Datos para el método de regresión lineal.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

4.1.2 Sprint 6. Integración de la herramienta Pentaho

Pentaho pone a disposición diversas herramientas de inteligencia de negocios para la elaboración de análisis, reportes, minería e integración de datos.

Para la generación de conocimiento a través de esta herramienta hacemos uso de las vistas materializadas, para la búsqueda y visualización de información.

Se debe realizar la conexión entre Pentaho y PostgreSQL para tener acceso a la información del almacén de datos (Almacen_VCA) y poder generar consultas empleando elementos propios de la fuente de datos. (Para ver más detalles revisar el Manual Técnico de Pentaho adjunto).

Seleccionamos la conexión a utilizar (conexionAlmacen) vamos a la opción Add Query y colocamos los scripts para los parámetros y query de consultas de información a utilizar en el reporte.

Pantalla de ingreso de query para el reporte.

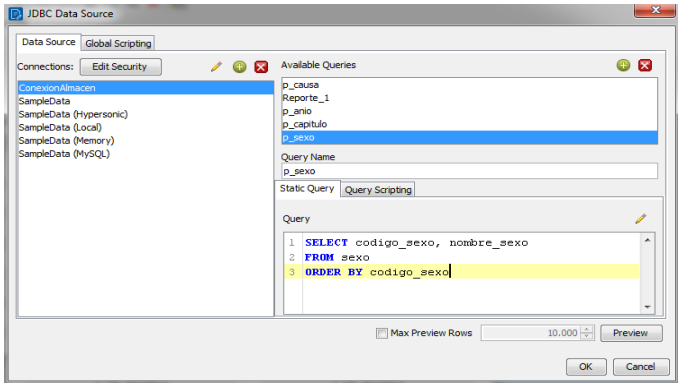


Figura 52. Añade query para consultas de información.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

Como ejemplo tomamos la vista materializada enfermedades por entidad de salud, que refleja información de enfermedades de la piel y el tejido subcutáneo diagnosticados por año acorde al sector y clase de entidad de salud, categorizando el sexo y la edad de los pacientes de los casos registrados.

Query de análisis de Enfermedades por entidad de salud

```
-- QUERY REPORTE ENFERMEDADES POR ENTIDAD DE SALLUD
select anio,sector,clase, capitulo, causa, nombre_sexo, nombre_con_egrpa,
SUM(rango1) AS "De 0 a 5 años",SUM(rango2) AS "De 6 a 17 años",
SUM(rango3) AS "De 18 a 65 años",SUM(rango4) AS "Más de 66 años",
(SUM(rango1) + SUM(rango2) + SUM(rango3) + SUM(rango4)) as total
from vm_entidad_causa
where anio in (${Anio})
and id_capitulo = ${Capitulo}
and id_sector in (${Sector})
and id_clase in (${Clase})
group by anio,sector,clase, capitulo, causa, nombre_sexo, nombre_con_egrpa
order by anio,causa,nombre_sexo,sector,clase,nombre_con_egrpa

-- PARÁMETROS
--Parámetro p_anio
SELECT anio FROM d_tiempo
WHERE mes IS NULL

--Parámetro p_capitulo
SELECT distinct id_capitulo, capitulo
FROM d_sintoma
order by id_capitulo;

--Parámetro p_sector
SELECT distinct id_sector, sector
FROM d_entidad_salud
order by id_sector;

--Parámetro p_clase
SELECT distinct id_clase, clase
FROM d_entidad_salud
where id_clase is not null
and id_sector in (${Sector})
order by id_clase;
```

Figura 53. Query de Enfermedades por entidad de salud.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

Ya creados los parámetros que permiten especificar la búsqueda a realizar (año, capítulo, sector y clase) identificados por la letra p al inicio y el query de consulta utilizando la vm_entidad_causa procedemos a diseñar el reporte. (Existen muchos tutoriales disponibles en la web para este propósito).

En la Figura Nº54, observamos los parámetros de búsqueda de información, a excepción de capítulo donde elige un solo parámetro, los demás (Año, Sector, Clase) pueden ser selecciones múltiples.

Parámetros para la búsqueda de información.

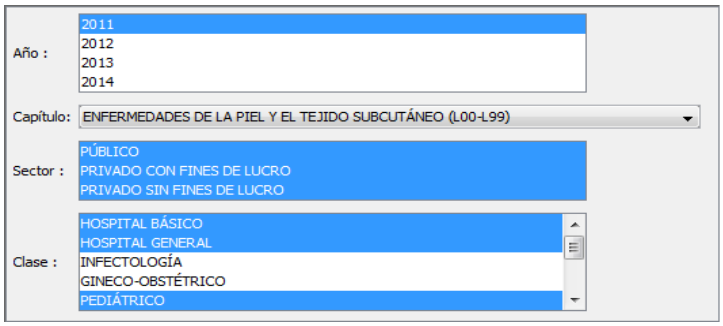


Figura 54.Parámetros de búsqueda de información.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

Los gráficos correspondientes a la primera sección del reporte muestran el total de casos diagnosticados por sector de entidad de salud al año representadas por barras con sus respectivas etiquetas y el porcentaje representadas por pasteles.

Parte 1 de 5. Reporte enfermedades por entidad de salud

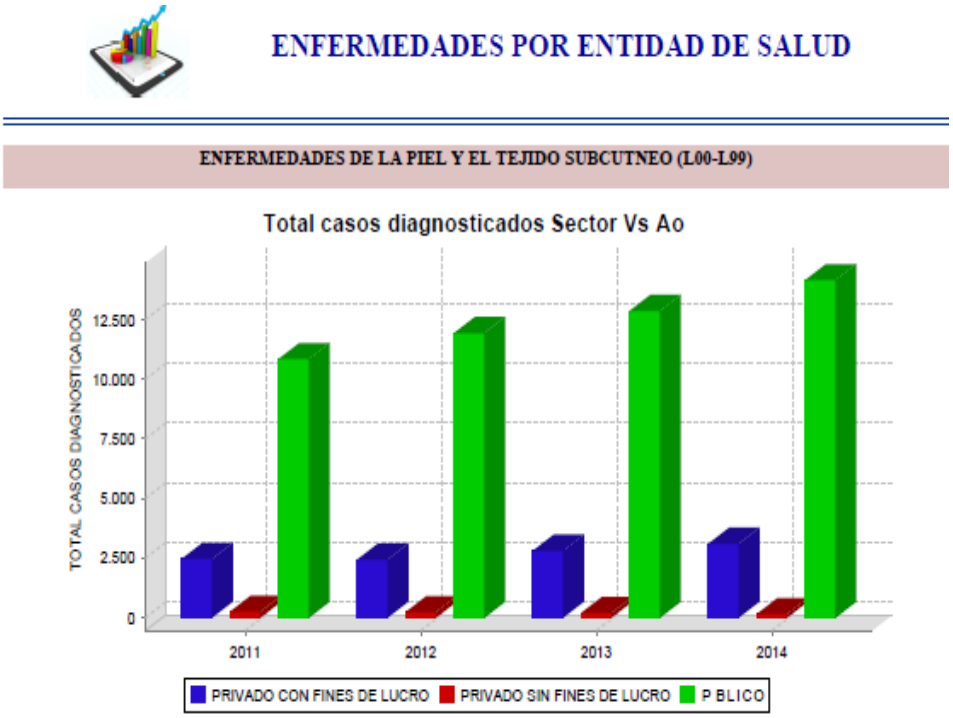


Figura 55. Barras de tipo sector de entidad de salud (Año-Sector).
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

Parte 2 de 5. Reporte enfermedades por entidad de salud

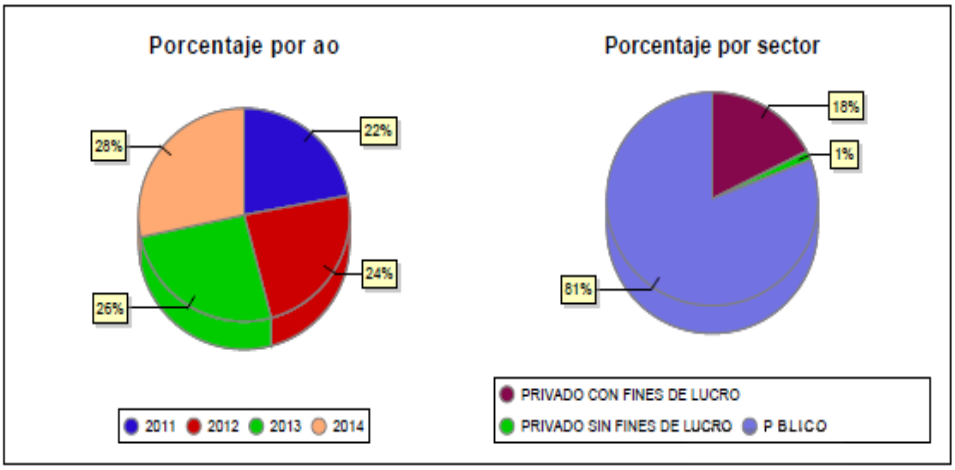


Figura 56. Gráficos con porcentajes de enfermedad por sector de salud y año.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

Se puede concluir el año donde se reportó el mayor número de casos de afecciones a la piel. Además, del sector con mayor porcentaje de casos.

La segunda sección analiza los casos diagnosticados por clase de entidad.

Parte 3 de 5. Reporte enfermedades por entidad de salud

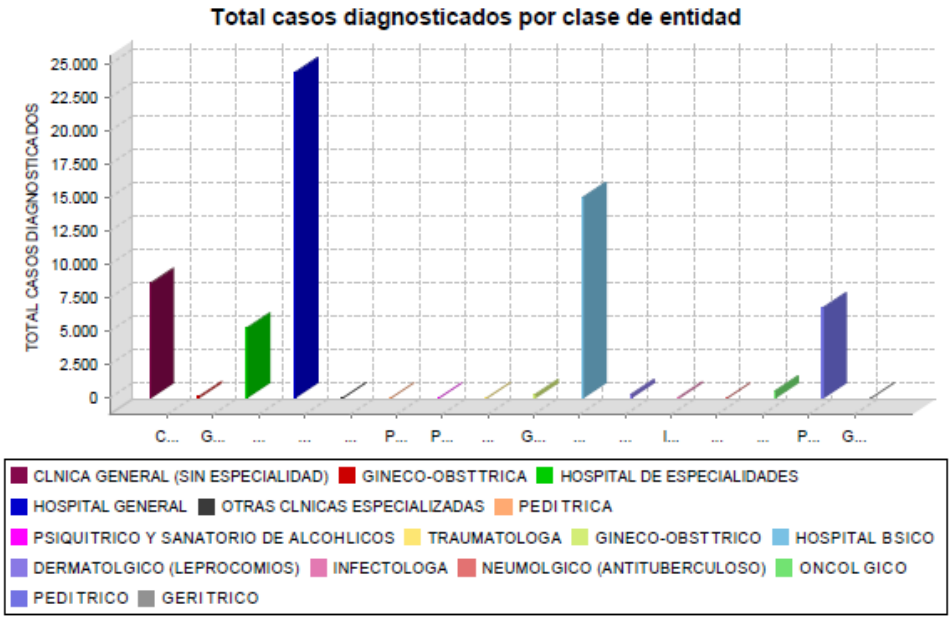


Figura 57. Barras de casos diagnosticados por clase de entidad de salud.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

Parte 4 de 5. Reporte enfermedades por entidad de salud

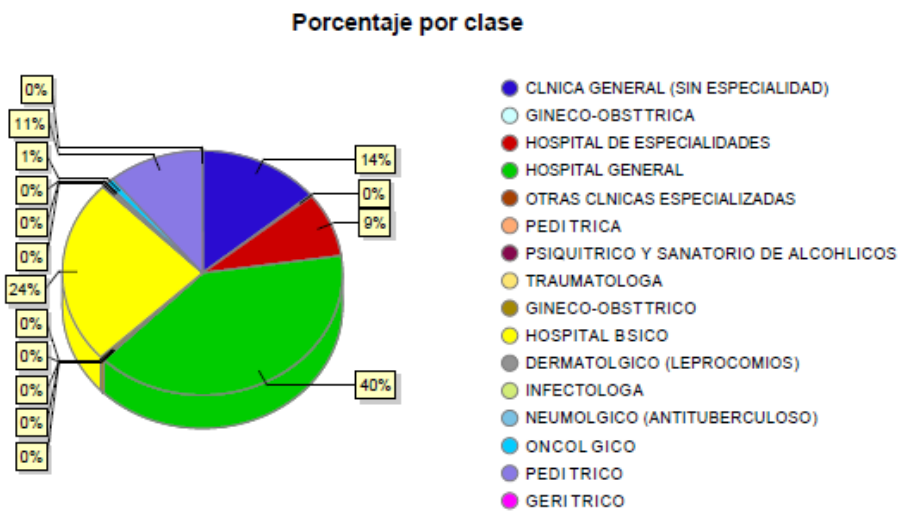


Figura 58. Gráfico con porcentajes de enfermedad por clase de entidad de salud.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

De este resultado podemos deducir que los casos diagnosticados del capítulo de enfermedades de la piel en mayor número se registran en el hospital general, seguido del hospital básico y la clínica general.

La Figura Nº59, corresponde a la tercera sección, aquí se muestra una visión más amplia de la información registrada por cada año, agrupada en el capítulo que estudia las enfermedades de la piel, sector y clase de entidad de salud, tomando en cuenta la causa y características como sexo y edad del paciente.

Parte 5 de 5. Reporte enfermedades por entidad de salud

Ao: 2.011

CAPTULO : ENFERMEDADES DE LA PIEL Y EL TEJIDO SUBCUTNEO (L00-L99)

SECTOR: PRIVADO CON FINES DE LUCRO

CLASE: CLNICA GENERAL (SIN ESPECIALIDAD)

>> ABSCESO CUTNEO, FURNCULO Y NTRAX DE GLTEOS

Sexo	0 a 5 aos	6 a 17 aos	18 a 65 aos	Ms de 66 aos	Total	%
HOMBRE	0	3	27	3	33	0,05
MUJER	0	2	27	6	35	0,06
				Total por causa	68	0,11

>> ABSCESO CUTNEO, FURNCULO Y NTRAX DE LA CARA

Sexo	0 a 5 aos	6 a 17 aos	18 a 65 aos	Ms de 66 aos	Total	%
HOMBRE	1	4	11	1	17	0,03
MUJER	7	3	8	1	19	0,03
				Total por causa	36	0,06

>> ABSCESO CUTNEO, FURNCULO Y NTRAX DE MIEMBRO

Sexo	0 a 5 aos	6 a 17 aos	18 a 65 aos	Ms de 66 aos	Total	%
HOMBRE	3	20	74	13	110	0,18
MUJER	10	7	58	16	91	0,15
				Total por causa	201	0,33

>> LCERA DE MIEMBRO INFERIOR, NO CLASIFICADA EN OTRA PARTE

Sexo	0 a 5 aos	6 a 17 aos	18 a 65 aos	Ms de 66 aos	Total	%
HOMBRE	0	1	0	0	1	0
MUJER	2	4	0	0	6	0,01
				Total por causa	7	0,01
				Total por clase	1.999	
				Total por sector	14.154	
				Total por ao	17.426	

TOTAL DIAGNOSTICADO 61.523

Figura 59. Parte3-Reporte Enfermedad por entidad de salud.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

3.1.3 Sprint 7. Integración de la herramienta Matlab.

Matlab, herramienta de software matemático de multiplataforma, orientado al análisis iterativo y procesos de diseño con un lenguaje de programación propio, con la manipulación de matrices pretende representar o aproximar la relación de dependencia entre las variables dependientes Y, y variables independientes Xi.

Para nuestro propósito, empleamos el método de regresión lineal, el cual se basa en la formula ‘Y = XB’, se creó un script, basado en un análisis previo del número de ocurrencias de una cierta enfermedad a lo largo del tiempo, en este caso la enfermedad "ÚLCERA DE MIEMBRO INFERIOR, NO CLASIFICADA EN OTRA PARTE", cuyo id es “3535” en la tabla de dimensión “d_sintoma”.

Query de la consulta para modelo de regresión lineal.

```
select
      b.causa,b.id, c.anio, c.mes, a.sexo, SUM(a.rango1 + a.rango2 + a.rango3 + a.rango4) as total
from th_morbilidad a, d_sintoma b, d_tiempo c where    a.sintoma = b.id
      and b.id_capitulo = 12 and b.id = 3535 and b.cod_causa is not null and
      a.tiempo = c.id and a.tiempo IN (select id from d_tiempo where anio =
      2011 and mes is not null)
group by b.causa,b.id,c.anio,c.mes,a.sexo order by causa,anio,mes,sexo
```

Figura 60.Query de la consulta de regresión lineal.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

El mismo script se utilizada para la extracción de las ocurrencias de dicha enfermedad al largo de los 4 años del análisis (2011, 2012, 2013 y 2014), estos valores obtenidos se convierten en el vector Y.

Totales de ocurrencias de casos por meses del año.

```
% Carga de datos Indices
total = [8 7 10 10 13 11 8 16 6 14 8 13]; %totales de los egresos 2011
total = [9 15 17 7 14 8 19 20 14 14 12 12]; %totales de los egresos 2012
total = [10 11 10 13 24 15 10 24 16 13 8 16]; %totales de los egresos 2013
total = [19 15 14 14 12 15 14 11 15 11 20 10]; %totales de los egresos 2014
```

Figura 61. Ocurrencia a lo largo de los años, por cada mes.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

Variables independientes son los meses de cada respectivo año.

```
% Vector x
mes = [1 2 3 4 5 6 7 8 9 10 11 12]; %meses del año 2011
x = mes;
x = x';
```

Figura 62. Vector de los meses de un determinado año.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

Obteniendo el coeficiente de regresión, con la aplicación del método de mínimos cuadrados obtenemos el ajuste de los datos.

Grafica generada aplicando el modelo de regresión lineal.

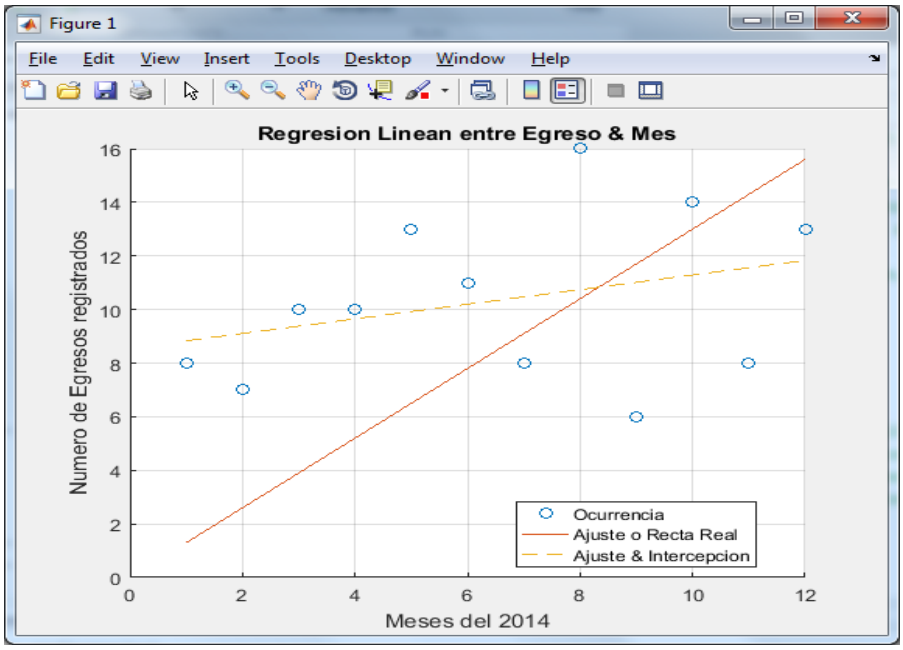


Figura 63. Regresión lineal años 2014.
Elaborado por: Mayra Alvarez y Vinicio Cañarejo.

En la gráfica representa las ocurrencias (circulo), son los datos obtenidos de la query, es decir el número de egresos por mes que se obtuvo del Almacén de Datos, de la enfermedad antes mencionada.

Ajuste o Recta Real, es una pendiente con ajuste promedio, que atraviesa las ocurrencias en una media.

Ajuste e Intercepción, es el ajuste de regresión lineal, verificado el mejor ajuste, la línea punteada, nos muestra un mejor promedio, más cerca de la realidad, en donde se aprecia que el número de egresos va en aumento entre 8 a 12 ocurrencias, por lo tanto, se puede predecir un amplio rango de ocurrencias basado en el gráfico.

CONCLUSIONES

- Se logró el propósito de la implementación de un BI para analizar la incidencia de enfermedades y problemas relacionados con la salud (CIE-10) en el Ecuador, con información provista por el INEC, con los datos de los últimos cuatro años.
- Se han presentado varias opciones para la implementación del almacén y minería de datos, de manera que, basados en la revisión de las fases a seguir, se facilitó la elección de una de las metodologías que más se ajustó a nuestras necesidades.
- Para la construcción del Almacén de Datos, es importante conocer el negocio, más que nada los datos que posee y cuales se podría obtener, para reducir el tiempo de análisis de las dimensiones evitandoo la creación de tablas y vistas innecesarias.
- Mediante la adecuada aplicación de los procesos en ETL sobre los datos, se logró extraer información depurada, descartando los problemas que existieron al momento de realizar el análisis de información, para posteriormente proceder a la caga de las respectivas tablas, tanto en la base operacional como en el almacén de datos.
- El proceso realizado y detallado en el presente documento, permite la consulta directa acorde a criterios de interés del usuario de enfermedades categorizadas en 10ma clasificación de la OMS, facilitando la búsqueda y tiempos de respuesta en la obtención de conocimiento.
- El almacén de datos es una herramienta de gran utilidad, con resultados en mano se puede visualizar de forma más clara lo que en realidad quiere decir la información, un reporte filtrado por provincia, por mes, por entidad etc., presentado en un gráfico, es lo que el gerente de una empresa desea, para de esa forma crear estrategias y toma de decisiones de una forma más rápida.

RECOMENDACIONES

- Antes de realizar la carga definitiva de datos en la base operacional, debemos asegurarnos que todas las tablas secundarias tengan la información necesaria, que abarque los registros de las tablas principales de todos los años a analizar.
- En la creación de dimensiones, buscar la forma de optimizar los procesos y sus tiempos de ejecución, en tablas de poca información o procesamiento, el tiempo es corto y se lo desprecia, pero al momento de utilizar millones de datos, esas optimizaciones suponen un gran beneficio y un dolor de cabeza menos.
- Tablas normalizadas, no solo que posean clave principal, sino también un buen performance de base y relaciones, seriales e indexaciones, son realmente útiles para agilitar las consultas en una base tan amplia y poblada.
- Para el análisis de datos y presentación de resultados, utilizar vistas materializadas u otra forma de extraer los datos de una forma más rápida, evitar la consulta de una tabla con millones de datos.
- Para presentación de resultado, usar una herramienta amigable e intuitiva de manejar, Pentaho es una buena opción, con un conocimiento básico de base de datos y consultas SQL, se puede explotar al máximo los datos del almacén.

LISTA DE REFERENCIAS

Ayala, A. (2016). *Inteligencia de Negocios: Una Propuesta para su Desarrollo en las organizaciones.* México: Instituto Politécnico Nacional.

Cannolly, T. M., & Wesley, C. E. (2005). *Sistemas de bases de datos: un enfoque práctico para diseño, implementación y gestión.* Pearson Educación.

Caralt, J., & Curto, J. (2010). *Introducción al Business Intelligence.* Barcelona: Editorial UOC.

Corrales, J. (2009). Gestión de la Función Administrativa Del Servicio Gallego de Salud. Partes Especifica VI: Sistemas de Información. Madrid: Eduformu.

Fayyad, U., Piatetsky, G., & Smyth, P. (1996). *Advances in knowledge and data mining.* Recuperado el 09 de Octure de 2017, de http://sedici.unlp.edu.ar/bitstream/handle/10915/21220/Documento_completo.pdf?sequence=1

Han, J., & Kamber, M. (2011). *Data Mining: Concepts and Techniques.* USA: Morgan Faufmann.

Hernandez, J., Ramirez, J., & Ferri, C. (2014). *Introducción a la Mineria de Datos.* Madrid: Departamento de Sistemas Informáticos y Computación. Universidad Politécnica de Valencia.

Hernandez, R. (2005). *Herramientas del Marketing: Data Warehousing,Tecnologia necesaria para el Comercio Internacional.* Recuperado el 19 de Septiembre de 2017, de http://revista.cnic.edu.cu/revistaCB/sites/default/files/articulos/CB-2005-4-CB-061.pdf

Hernandez.J, & Varios. (2014). *Introducción a la Mineria de Datos.* Madrid: Departamento de Sistemas Informáticos y Computación. Universidad Politécnica de Valencia.

Kimball, R., & Ross, M. (2013). *The Data Warehouse Tollkit: The Definitive Guide to Dimensional Modeling.* Indianapolis: Kimball Group.

López Porrero, B. E. (2009). *Limpieza de datos.* Editorial Feijóo.

Martinez, N. (05 de Mayo de 2010). *Business Intelligence*. Recuperado el 23 de Julio de 2017, de http://normanmartinez.blogspot.com/2010/05/modelos-olap.html

Ministerio de Sanidad, S. S. (Enero de 2016). *eCIE10ES.* Recuperado el 23 de Julio de 2017, de Edición electrónica de la CIE-10-ES Diagnósticos: http://eciemaps.msssi.gob.es/ecieMaps/browser/index_10_mc.html

Naquista. (Marzo de 2013). *Anexo:CIE-10*. Recuperado el 23 de Julio de 2017, de Capítulo XII: Enfermedades de la piel y el tejido subcutáneo: https://wiki.itcsoluciones.com/index.php/Anexo:CIE-10_Cap%C3%ADtulo_XII:_Enfermedades_de_la_piel_y_el_tejido_subcut%C3%A1neo

Salud, O. P. (2003). *lasificación Estadística Internacional de Enfermedades y Problemas Relacionados con la Salud.* Recuperado el 07 de 10 de 2017, de Volumen 2. Manualñ de Instruciones: http://ais.paho.org/classifications/Chapters/pdf/Volume2.pdf

Trujillo, J. (2013). *Diseño y explotación de almacenes de datos: conceptos básicos de modelado multidimensional.* ECU Editorial.

Zamarrón, C., García, V., Calvo, U., Pichel, F., & Rodríguez, J. (2006). *Alteración de la Minería de Datos al estudio de las alteraciones respiratorias.* Recuperado el 23 de Julio de 2017, de Hospital Clínico Universitario de Santiago de Compostela, Servicio de Neumología.: http://www.sogapar.org/pneuma/pneuma6/pneuma-n-6-5c.pdf

Anexos

Para revisar los anexos de este trabajo, por favor diríjase al CD.