UNIVERSIDAD POLITÉCNICA SALESIANA

SEDE QUITO

CARRERA:

INGENIERÍA DE SISTEMAS

Trabajo de titulación previo a la obtención del título de:

INGENIEROS DE SISTEMAS

TEMA:

IMPLEMENTACIÓN DE UN BI PARA ANALIZAR LA PROVISIÓN DE SERVICIOS DE SALUD EN LOS ÚLTIMOS 8 AÑOS, UTILIZANDO POSTGRESQL Y PENTAHO

AUTORES:

JORGE DAVID QUIROZ QUIROZ

JAVIER ANDRÉS REYES ZÚÑIGA

TUTOR:

ALONSO RENÉ ARÉVALO CAMPOS

Quito, julio de 2017

CERTIFICADO DE CESIÓN DE DERECHOS DE AUTOR DEL TRABAJO DE TITULACIÓN

Nosotros JORGE DAVID QUIROZ QUIROZ y JAVIER ANDRÉS REYES ZÚÑIGA, con documentos de identificación N° 1723560270 y 1719623215 respectivamente, manifestamos nuestra voluntad y cedemos 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 PROVISIÓN DE SERVICIOS DE SALUD EN LOS ÚLTIMOS 8 AÑOS, UTILIZANDO POSTGRESQL Y PENTAHO", mismo que ha sido desarrollado para optar por el título de: INGENIEROS 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.

Jorge David Quiroz Quiroz

CI: 1723560270

Javier Andrés Reyes Zúñiga

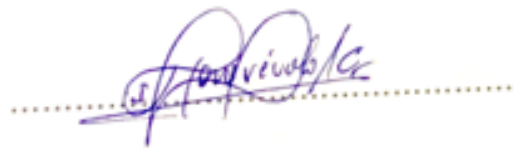
CI: 1719623215

Quito, julio de 2017

DECLARATORIA DE COAUTORÍA DEL TUTOR

Yo declaro que bajo mi dirección y asesoría fue desarrollado el trabajo de titulación, "IMPLEMENTACIÓN DE UN BI PARA ANALIZAR LA PROVISIÓN DE SERVICIOS DE SALUD EN LOS ÚLTIMOS 8 AÑOS, UTILIZANDO POSTGRESQL Y PENTAHO", realizado por los estudiantes JORGE DAVID QUIROZ QUIROZ y JAVIER ANDRÉS REYES ZÚÑIGA, obteniendo un producto que cumple con todos los requisitos estipulados por la Universidad Politécnica Salesiana, para ser considerados como trabajo final de titulación.

Quito, julio de 2017



Ing. Alonso René Arévalo Campos

CI: 1400164891

Dedicatoria

A todos los *argnonáutas*, que caminantes, sin descanso, emprenden y continúan su travesía en pos del *vellocino de oro*, en el sendero de la formación personal, académica e intelectual.

Jorge

* * *

A:

Dios, por haberme dado la oportunidad de vivir e iluminarme a lo largo de mi existencia, brindándome esa paz y alegría que me da aún más fuerza de seguir adelante.

A mis padres, por haberme brindado su amor incondicional, a la vez que me han enseñado como valores fundamentales la humildad, el respeto y la responsabilidad.

A mis hermanos, por haberme dado su apoyo y cariño incondicional en todos estos años, que, aunque no estén cerca los llevo siempre presentes en mi corazón.

A mis sobrinos, que puedan verme como ejemplo a seguir y sepan que con humildad y constancia se pueden cumplir cualquier meta o propósito en la vida.

Javier

Agradecimiento

Este proyecto va dirigido con expresión de gratitud a la Universidad Politécnica Salesiana, especialmente a la carrera de Ingeniería de Sistemas, a sus autoridades y personal docente por haber contribuido en nuestra formación profesional y personal, principalmente a nuestro tutor, el Ing. René Arévalo que con entusiasmo y compromiso nos ha brindado su apoyo incondicional, orientación y motivación en este proceso académico.

Gratitud y reconocimiento también a nuestros padres, familiares, amigos y a todas las personas implicadas, que han contribuido a cristalizar nuestras metas y objetivos.

Jorge

Javier

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

Resumen

El presente proyecto de titulación tiene como objetivo la construcción de un almacén de datos con información obtenida del portal web del INEC, referente a la provisión de servicios de salud en los ocho últimos años, para analizar la dotación de servicios de salud en función de cuatro criterios de análisis: tiempo, especialidad, entidad de salud y ubicación geográfica, mediante herramientas OpenSource de inteligencia de negocios.

Para la elaboración del presente proyecto, se emplearon técnicas y procesos de recopilación de datos como ETL (extracción, transformación y carga) para la construcción del almacén empleando PostgreSQL, para llevar a cabo tareas propias de inteligencia de negocios (BI), como la elaboración de reportes en Pentaho Report Designer y la aplicación de algoritmos de minería de datos en Weka.

La metodología SCRUM permitió gestionar el proyecto en función de entregas parciales funcionales, facilitando la gestión de cambios, prioridades y control del tiempo.

Abstract

The goal of this Titulation project is the datawarehouse construction with information obtained from the web portal of the INEC, referring to the provision of health services in the last eight years, to analyze the provision of health services according to four analysis criteria: time, specialty, health entity and geographical location, through OpenSource business intelligence tools.

For the elaboration of the present project, techniques and processes of data collection like ETL (extraction, transformation and load) were used for the construction of the warehouse using PostgreSQL were used, to carry out own tasks of intelligence of businesses (BI), like the elaboration of reports in Pentaho Report Designer and the application of algorithms of data mining in Weka.

The SCRUM methodology allowed to manage the project based on functional partial deliveries, facilitating change management, priorities and time control.

Introducción

Una de las áreas más sensibles de un estado es la de la salud, por lo tanto, es de vital importancia gestionarla apropiadamente. El presente proyecto de titulación aborda esta temática con el objeto de mostrar los resultados de la gestión de los diferentes gobiernos de turno en esta área, en relación con la provisión de servicios de salud, y a su vez también contrastarla con la que proveen las entidades de salud privada.

El disponer de información de la provisión de servicios de salud, basado en un análisis de datos reales obtenidos el portal del INEC, en la que se pone particular atención en la ubicación geográfica, entidades gestoras y establecimientos de salud en el mencionado periodo de tiempo, facilitará la detección de grupos humanos con mayores necesidades, especialidades médicas con mayor demanda y menor cobertura, etc. Este conocimiento facilitará que la autoridad de salud pertinente visualice los problemas detectados y planifique las actividades necesarias para corregir o solventar las mismas.

Los beneficiarios serán todos los interesados en conocer el desempeño de los servicios de salud en el país y su impacto en la población, con la finalidad de cubrir las necesidades de manera oportuna.

Para aportar a la solución de esta problemática este proyecto de titulación, propone la construcción de un almacén de datos en PostgreSQL, en base a la generación de información clasificada por geografía, entidades de salud, tipo de servicio, en el periodo comprendido entre 2007 y 2014, tomando como base la información obtenida del Banco de Datos del INEC sobre provisión de servicios de salud y posteriormente efectuar un análisis sobre los resultados obtenidos.

Problema

El INEC proporciona en su banco de información, disponible en su portal web, datos sobre recursos y actividades de salud, con periodicidad anual que constituye información histórica sobre la que se efectuará minería de datos, para realizar análisis que permita extraer información relacionada con la provisión de servicios de salud, que ofrezca patrones de comportamiento y tendencias con el fin de conocer la cobertura de servicios, en contraste con las necesidades de la población, en cualquier zona geográfica del país.

De esta manera, es posible determinar la necesidad de especialistas en el área de la salud en función de su ubicación geográfica y el tipo de entidad de salud.

Justificación

Debido a la información disponible en el banco de datos del INEC en lo referente a recursos y actividades de salud, se determina que es viable la construcción de un almacén de datos con información de los últimos ocho años, en relación con la provisión de servicios de salud, con la finalidad de efectuar un análisis que permita obtener conclusiones importantes en este contexto en beneficio de la población.

Objetivos

Objetivo general

Crear un almacén de datos para analizar la provisión de los servicios de salud en el Ecuador en los últimos ocho años, utilizando la información disponible en el portal del INEC, correspondiente al sector público y privado.

Objetivos específicos

Establecer un modelo multidimensional para alojar en PostgreSQL la información obtenida del banco de información del INEC.

Ofrecer información relevante referente al resultado del análisis efectuado, utilizando criterios de clasificación para la creación del cubo de información contenido en las tablas de hechos en base a cuatro dimensiones: periodo de

tiempo, tipo de servicio de salud, localización geográfica y entidades (establecimientos) de salud.

Analizar la data del almacén, con la finalidad de obtener patrones de comportamiento, tendencias y disponibilidad de los servicios de salud tanto en el sector público como privado.

Marco Referencial

La fuente de información oficial para recabar datos en nuestro análisis ha sido el banco de información del INEC, ubicado en su portal web, cuya información está transparentada y disponible para todos los ciudadanos.

El Instituto Nacional de Estadística y Censos (INEC) es el órgano rector de la estadística nacional y el encargado de generar las estadísticas oficiales del Ecuador para la toma de decisiones en la política pública.

> El sistema de salud de Ecuador, por su parte, está compuesto por dos sectores: público y privado. El sector público comprende al Ministerio de Salud Pública (MSP), el Ministerio de Inclusión Económica y Social (MIES), los servicios de salud de las municipalidades y las instituciones de seguridad social (Instituto Ecuatoriano de Seguridad Social, Instituto de Seguridad Social de las Fuerzas Armadas e Instituto de Seguridad Social de la Policía Nacional). El MSP ofrece servicios de atención de salud a toda la población. El MIES y las municipalidades cuentan con programas y establecimientos de salud en los que también brindan atención a la población no asegurada. Las instituciones de seguridad social cubren a la población asalariada afiliada. El sector privado comprende entidades con fines de lucro (hospitales, clínicas, dispensarios, consultorios, farmacias y empresas de medicina prepagada) y organizaciones no lucrativas de la sociedad civil y de servicio social. Los seguros privados y empresas de medicina prepagada cubren otros grupos poblacionales pertenecientes a estratos de ingresos medios y altos. Además, existen consultorios médicos particulares, en general dotados de infraestructura y tecnología elementales, ubicados en las principales ciudades y en los que la población

suele hacer pagos directos de bolsillo en el momento de recibir la atención. (Lucio, R., Villacrés, N., 2011)

2. Fundamentos teóricos

2.1 Servicios de salud

Según la definición de la Organización Mundial de la Salud (OMS) “son servicios entregados por personal de salud en forma directa o por otras personas, bajo la supervisión de éstas, con los propósitos de, en primer lugar, promover, mantener y/o recuperar la salud, y, en segundo lugar, de minimizar las disparidades tanto en el acceso a los servicios de salud, como en el nivel de salud de la población”. (OMS, 2003)

2.2 Inteligencia de negocios (BI)

Es la utilización de los datos históricos y fiables de las actividades empresariales de una entidad, mediante procesos y mecanismos de refinación, tratamiento y análisis; entre otros, para extraer patrones de conducta, tendencias e información que dé soporte a la toma de decisiones de la organización identificando las actividades que generan valor.

Integra metodologías, prácticas, y conceptos entre los cuales se tiene: análisis predictivo, reportería, minería de datos, integración de datos, etc. (Curto Díaz, 2010)

2.3 Minería de datos

Debido a la existencia de grandes volúmenes de información y al uso extendido de herramientas ofrecidas por las tecnologías de la información, el análisis de datos hoy en día, está orientado por un conjunto de técnicas especializadas, pertenecientes a lo que se conoce como minería de datos *(data mining).*

Estas técnicas tienen como fin descubrir el conocimiento existente en grandes almacenes de información, descubriendo tendencias, patrones y relaciones significativas entre datos, procesos o eventos que suceden dentro de una organización, aportando a éstas en la toma de decisiones, actividades gerenciales y ejecutivas, brindando mayores prestaciones que la de simplemente dar soporte a los procesos organizacionales más elementales. (Pérez López & Santín González, 2007)

2.4 Almacén de datos

"Es una colección de datos orientados por tema, integrados, variables en el tiempo, y no volátiles", almacenan información histórica para ser empleada en procesos estratégicos, difieren de las bases de datos convencionales en el hecho de que están orientados a la extracción de conocimientos de datos históricos y efectuar consultas a nivel gerencial y ejecutivo. Su objetivo no es permitir llevar a cabo transacciones (inserción, eliminación o actualización) desde una aplicación externa como sucede con una base de datos convencional.

Son orientados, porque pueden estar conformados por información proveniente de diversas bases de datos o fuentes externas en respuesta al contexto en el que se quiera efectuar el análisis conforme a los intereses de la empresa.

Son integrados porque contiene los datos de forma uniforme, resolviendo problemas de heterogeneidad entre formatos de distintas fuentes de almacenamiento.

Son variables en el tiempo ya que los datos están delimitados bajo una referencia temporal bajo la cual son válidos y no son volátiles ya que los datos almacenados no presentan tendencia a ser actualizados, sino que son históricos. (Trujillo, Mazón, & Pardillo, 2010)

2.5 Sistemas de soporte a las decisiones (SSD)

"Es un ambiente de trabajo compuesto por el usuario, procedimientos para el tratamiento de información y el equipo de cómputo, orientado a proveer información que apoye las operaciones, la administración y la función de toma de decisiones en una organización." (Ayala, 2006) en donde el personal es el encargado de alimentar los datos, procesarlos y explotarlos conforme los requerimientos de la organización para generar así valiosa información que aporte la toma de decisiones a nivel empresarial, los mismos que se pueden lograr mediante informes, reportes análisis, etc.

Estos sistemas se estructuran de acuerdo al giro de negocio de la organización, soportando sus funciones y atendiendo las tareas realizadas, contribuyendo a su adecuado funcionamiento. Se encarga de atender los niveles jerárquicos de la empresa a través de tres aspectos:

- Sistema de proceso de transacciones (SPT). Se enfoca en las operaciones cotidianas, reportes de estado, consultas, procurando rapidez y precisión de respuesta.
- Sistema de administración de la información (SAI). Arroja información solicitada por actividades de planeación y control, mediante consultas flexibles, no complejas.
- Sistema de información gerencial (SIG). Destinado a análisis de situaciones complejas, para lo cual dispone de herramientas e información pertinente. (Ayala, 2006)

Sistemas de soporte a las decisiones

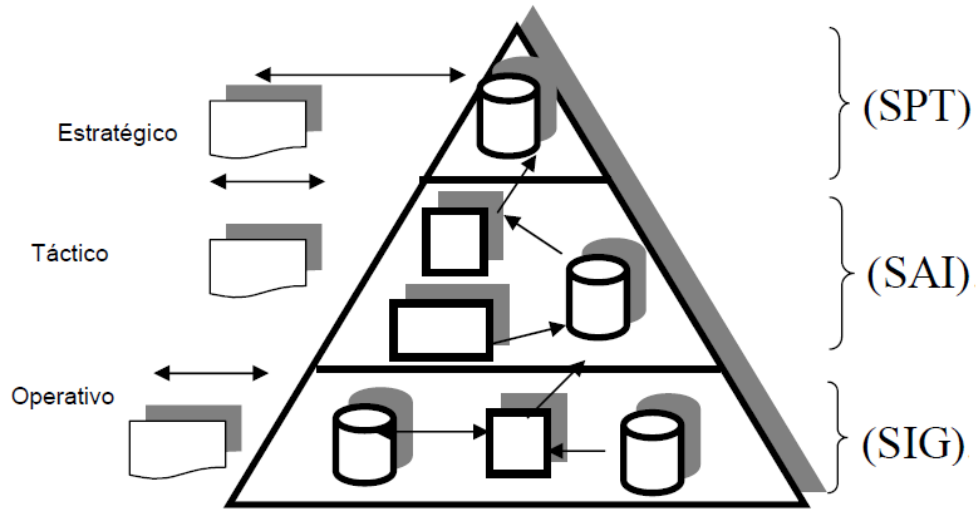


Figura 1: Estructura de un SSD basado en soporte a niveles.
Fuente: Ayala, 2006

2.6 Almacenamiento de Datos

2.6.1 Introducción

Desde décadas atrás, las empresas han optado por utilizar sistemas informáticos que puedan ayudar a automatizar los procesos de negocio, por lo que, al lograr esto generaban una ventaja competitiva ante las demás organizaciones. Anteriormente, dichos sistemas operacionales no eran lo bastante consistente para soportar tales requerimientos en la automatización de ciertos procesos y ayuda a la toma de decisiones de la empresa. Este hecho se daba, ya que cada sistema operacional que utilizaba una empresa no se maneja, por ejemplo los mismos tipos de datos e incluso eran programados en un lenguaje único para

dicho sistema. Es aquí, donde se vio en la necesidad de diseñar una solución que permita al usuario disponer de un sistema que le pueda ayudar al proceso de toma de decisiones, recibiendo datos de múltiples fuentes de información dados por estos sistemas operacionales.

2.6.2 Definición

Gracias a Bill Inmon, conocido como el "padre de los almacenes de datos", el concepto de los almacenes de datos logró verse como una solución valiosa y fiable.

Por su parte, el concepto pudo verse de la siguiente manera "Una colección de datos clasificada por temas, integrada, variable en el tiempo y no volátil que se utiliza como ayuda al proceso de toma de decisiones por parte de quienes dirigen una organización." (Thomas M. Connolly, 2005).

Dentro de esta definición, se puede rescatar y analizar lo siguiente con respecto a los datos:

- Están organizados de acuerdo con los temas de mayor importancia que tiene la cada organización.
- Se encuentran integradas, es decir, que los datos provienen de diferentes sistemas.
- Son variables en el tiempo, ya que los datos son válidos en algún instante temporal de tiempo.
- Son no volátiles, es decir, que no se actualizan en tiempo real sino gracias a través de la manipulación de sistemas operacionales.

Según (Thomas M. Connolly, 2005), el objetivo de los almacenes de datos es la de integrar todos los datos corporativos y convertirlos en un solo repositorio en el cual todos los usuarios puedan ejecutar consultas de manera rápida y fácil, además de generar reportes y realizar análisis.

Dentro de los beneficios de implementar de manera adecuada un almacén de datos se tienen los siguientes:

- Debido a la diversidad de soluciones que proporciona un almacén de datos existe un alto retorno de la inversión. Según (Thomas M. Connolly, 2005), un estudio realizado por International Data Corporation (IDC) se informa

que el ROI (Returns on Investiment) promedio a lo largo de tres años en sistemas de almacenes de datos alcanzaban el 401%, consiguiendo un 90% de las empresas un ROI de un 40%, la mitad de las empresas un ROI del 160% y la cuarta parte de ellas un ROI del 600%.

- Se obtienen ventajas competitivas ya que, se puede tomar una decisión que previamente se puede conocer el impacto que podría tener ya sea en los clientes, tendencias, etc. Esto anteriormente no se lo podía hacer por la falta de información necesaria para ello.
- Un almacén de datos permite tener un único repositorio de información compacta y coherente que ayuda a la toma de decisiones mejorando a su vez la productividad de la organización.

2.6.3 Problemas de los almacenes de datos

A continuación, se presenta una lista de los problemas asociadas en el desarrollo y gestión de un almacén de datos más comunes:

- *Subestimación de recursos,* esto se da debido a la subestimación de tiempo que los desarrolladores otorgan para poder extraer, limpiar y cargar los datos al almacén.
- *Problemas ocultos* que vienen de los sistemas origen de donde se captura la información, por ejemplo, en ciertos campos de un sistema se permite la introducción de valores nulos.
- *Incremento a la demanda,* esto se debe a que después de la implementación del almacén los usuarios hagan varias solicitudes de información y si no se tiene una adecuada herramienta que soporte este tipo de solicitudes esto creará varios inconvenientes como demora en ciertas consultas, etc.
- *Alta demanda de recursos,* al realizar las respectivas dimensiones de las tablas requeridas y a su vez añadiendo índices relacionados con las tablas de hechos, la realización de esto produciría que se pueda utilizar una gran cantidad de espacio en disco que los propios datos obtenidos inicialmente.
- *Altos costes de mantenimiento,* los sistemas de almacenes de datos requieren de una labor intensiva de mantenimiento, ya que, al realizar la reorganización de los procesos de la empresa, se debe velar por la consistencia y coherencia de los datos obtenidos por el almacén.

- *Proyectos de larga duración:* la construcción que tomaría en realizar un almacén de datos tiene un promedio de tres años, por tal razón, es que varias organizaciones hayan optado por otras alternativas en la gestión de sus datos.
- *Complejidad de la integración*, este tema es de mucha importancia y complejidad ya que habrá en algún punto donde debemos ver hasta donde nuestro almacén podrá anexar todo el contenido de los diferentes sistemas que maneja la organización y verificar que dicha información sea coherente a lo que se está buscando (Thomas M. Connolly, 2005)

2.6.4 Metodología Kimball

Para la construcción de un almacén de datos, existen muchas metodologías, entre las más utilizadas se encuentran Inmon y Kimball. La primera sigue un enfoque descendente o *top-down*, comenzando desde un inicio con la construcción del almacén o minería de datos (*datawarehouse*), se basa en conceptos de bases relacionales; mientras que la segunda, tiene un paradigma *bottom-up* o ascendente ya que inicia con la construcción de mercados de datos (*data marts*), que al igual que un minería de datos (*data mining*) son repositorios de datos, pero orientados a un contexto o área específica, de menor alcance, para luego integrarlos de manera ascendente en un solo almacén, tiene un enfoque de modelado dimensional. (Rivadera, 2010)

Para esto se siguen las consideraciones de granularidad mencionadas en el modelo dimensional y sus respectivas tablas de hechos

2.6.5 Proceso de extracción, transformación y carga (ETL)

Este proceso es la base sobre la que se alimenta el almacén de datos. Si este sistema tiene un diseño adecuado, puede extraer los datos de los sistemas de origen de datos, aplicar diversas reglas para aumentar la calidad y consistencia de los mismos, consolidar la información proveniente de distintas fuentes, y finalmente cargar (grabar) la información en el almacén de datos, siguiendo un formato estandarizado y acorde para su utilización por parte de herramientas de análisis. (Rivadera, 2010)

Este proceso según (Gómez, 2012), es clave en el desarrollo de un BI y consume alrededor de un 60 a 80 por ciento del tiempo del proyecto.

Extracción: Tiene como objetivo obtener los datos de diferentes fuentes, en esta fase se tiene los datos en bruto.

Transformación: En esta fase se recuperan los datos en bruto y se asegura su calidad, mediante depuración, corrigiendo valores duplicados, inconsistentes, etc. Es decir, se transforman reduciendo errores de carga dando como resultado datos limpios, consistentes y útiles.

Carga: Una vez conseguido un nivel de consistencia de formato y definición de datos mediante la fase de transformación, se procede a almacenarlos en un almacén o *datawarehouse*, en este caso en la base *almacen*, con el objetivo de analizarlos y apoyar un proceso de negocio. Los procedimientos empleados en esta parte suelen ser complejos. (Gómez, 2012)

2.6.5 Arquitectura de un almacén de datos.

A continuación, se presenta los componentes principales y la arquitectura de un almacén de datos (Thomas M. Connolly, 2005)

Arquitectura de un almacén de datos

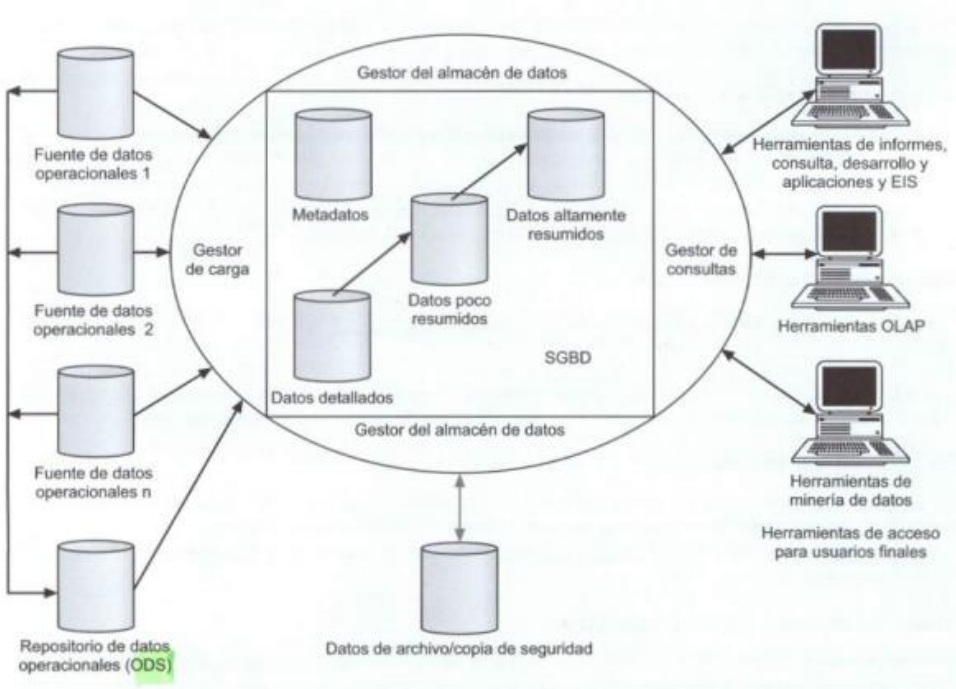


Figura 2: Esquema del gestor de almacén de datos
Fuente: Thomas M. Connolly, 2005

2.6.5.1 Datos operacionales

Dentro de los datos operacionales, se encuentran los siguientes tipos de fuentes para obtener los datos:

- Datos procedentes de sistemas mainframe (Sistemas de datos procedentes de bases de datos comerciales, servidores de transacciones y aplicaciones).
- Datos departamentales almacenados en archivos propietarios.
- Datos privados de estaciones de trabajo y servidores.
- Sistemas externos como la Internet, bases de datos disponibles asociados a clientes y proveedores.

2.6.5.2 Repositorio de datos operacionales

Los repositorios de datos operacionales *Operational Data Store* (ODS), son repositorios utilizados para realizar análisis y generación de informes, pero la diferencia con un almacén de datos es que no llega a ofrecer ayuda en la toma de decisiones, pero si en la integración de datos para el almacén.

2.6.5.3 Gestor de carga

El gestor de carga es la operación asociada a la extracción y carga de los datos en el almacén, esto se da utilizando herramientas de carga de datos y programas personalizados.

2.6.5.4 Gestor de almacén de datos

Este gestor permite, realizar diversas operaciones para incluir los datos en el almacén como:

- Análisis de datos para garantizar la coherencia de los mismos.
- Transformación y combinación de los datos origen, extrayéndoles del espacio de almacenamiento temporal y almacenándolos en las tablas del almacén de datos.
- Creación de vistas e índices.
- Copia de seguridad de datos.

2.6.5.5 Gestor de consultas

El gestor de consultas, utiliza herramientas de acceso a datos para que los usuarios finales puedan obtener sus reportes. Por lo que, las operaciones

realizados por este gestor incluye dirigirse a tablas apropiadas y la planificación en la ejecución de consultas.

2.6.5.6 Herramientas de acceso para usuarios finales

Como se ha dicho anteriormente, el objetivo de un almacén de datos es proporcionar información a los usuarios finales para la toma de decisiones estratégicas de cada organización, para ello se deben optar por la utilización de alguna herramienta que permita obtener dichos reportes o informes.

Para facilitar la clasificación de este tipo de herramientas, se han dividido en cinco grupos principales:

- Herramientas de consulta y generación de informes.
- Herramientas de desarrollo de aplicaciones.
- Sistemas de información ejecutiva (EIS, *Executive Information System*).
- Herramientas de procesamiento analítico en línea (OLAP, *Online Analytical Processing*).
- Herramientas de minería de datos.

En este caso, la herramienta a utilizar es la minería de datos, cuya definición indica que "es el proceso de descubrir nuevas correlaciones, patrones y tendencias significativas procesando grandes cantidades de datos mediante técnicas estadísticas, matemáticas y de inteligencia artificial." (Thomas M. Connolly, 2005). Otro de los aspectos importantes de una minería de datos es que, permite construir modelos predictivos y retrospectivos.

2.7 Sistemas basados en conocimiento (SBC)

Estos sistemas "se orientan a la adquisición, representación y empleo de conocimiento de un dominio de aplicación para identificar y resolver problemas a través de la toma de decisiones" (Ayala, 2006). Se pude decir que tienen un alcance mayor que los sistemas de información al permitir llevar a cabo interpretaciones y dotar de un significado a un problema en particular.

Sus objetivos fundamentales son la solución de problemas y automatización del proceso de toma de decisiones, la conexión a bases de datos alimenta los SSD y es, en última instancia, la persona responsable, quien luego de analizar e interpretar toma una decisión.

2.7.1 Arquitectura de un SBC

Los SBC están integrados por una interfaz, que se encarga de comunicarse con el usuario, permitiendo que este genere instrucciones; una base de conocimientos, que almacena y permite acceder al conocimiento; el mecanismo de adquisición del conocimiento, alimenta la base de conocimientos con conceptos ingresados manualmente; el motor de inferencias, se encarga de resolver problemas en base a procedimientos de búsqueda, heurística y lógica difusa, haciendo uso de la base de conocimientos y la información del problema proporcionada, que constituye la entrada para la obtención de la solución y, finalmente, el área de trabajo, que es donde se representa el problema y se presenta el desarrollo de la solución.

Sistemas basados en conocimiento

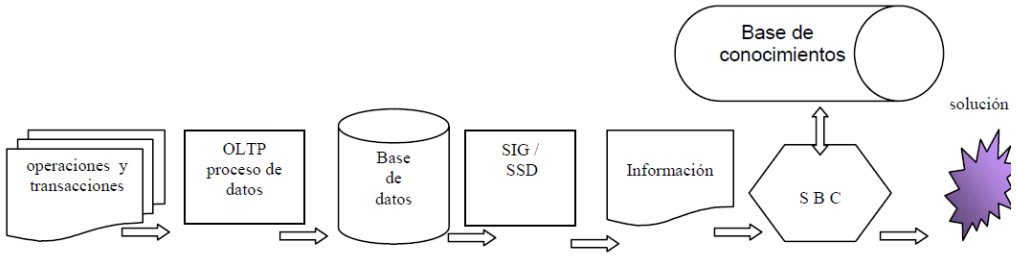


Figura 3: Modelo de un SBC, presenta el logro de la solución partiendo de transacciones cotidianas.
Fuente: Ayala, 2006

2.7.2 Involucrados en un SBC

Los participantes en este tipo de sistemas son el usuario, que es el responsable de plantear problemas e interpretar las soluciones arrojadas, el especialista que constituye la fuente del conocimiento, aporta con razonamientos, experiencias, métodos para resolver problemas y el ingeniero del conocimiento, quien se encarga de la modelar, desarrollar y construir el SBC en base a los requerimientos del usuario y el conocimiento del especialista. (Ayala, 2006)

Se pueden establecer diferentes categorías de SBC, de acuerdo a los requerimientos del usuario y posibilidades identificadas por el especialista y el ingeniero del conocimiento, entre ellas tenemos: SBC dedicados a la interpretación, que comparan un conjunto de datos con otros (de referencia), dando una calificación o asociación como respuesta, los SBC de predicción de consecuencias derivadas de una interpretación, SBC que arrojan diagnósticos en

base a una predicción, usados en medicina, SBC de diseño que establecen configuraciones acorde a requerimientos planteados, SBC de monitoreo y rastreo, control y reparación. (Ayala, 2006)

2.8 Ingeniería del Conocimiento

La Ingeniería del conocimiento (IC) nace tras la idea del mundo contemporáneo de que el conocimiento es el activo de mayor valor que posee cada persona. Es ahí donde varios autores la definen como "una disciplina emergente, que establece los mecanismos necesarios para poder almacenar, procesar y gestionar todo el conocimiento de un experto" (Garzón N. M., 2015).

Los sistemas basados en conocimiento (SBC), son aquellos sistemas que permiten realizar una completa gestión acerca del conocimiento obtenido de un experto. Para ello, es importante delimitar el problema o causa específica a solucionar, a partir de ello se puede aplicar lo que se conoce como lógica matemática a los diferentes escenarios que se presentan en la vida real.

Sistemas basados en conocimiento

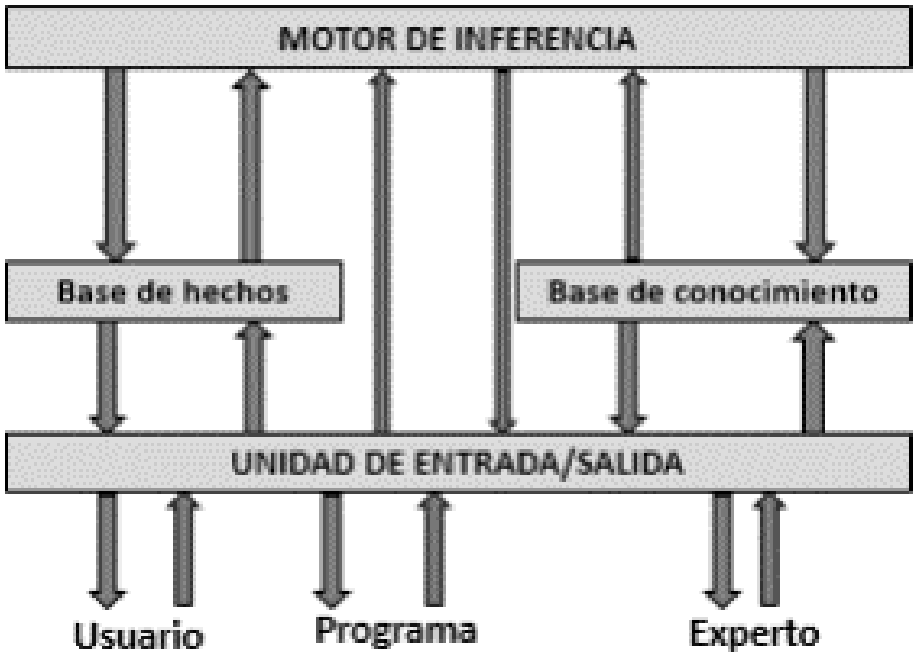


Figura 4: Componentes genéricos de un sistema basado en conocimiento.
Fuente: Garzón N. M., 2015

Por lo tanto, para la creación de estos sistemas se aplican ciertos pasos como los señalados a continuación.

2.8.1 Adquisición del Conocimiento

La adquisición del conocimiento, es la parte primordial para la construcción de este tipo de sistemas, ya que es la parte donde se provee de información al mismo y este se basa en la experiencia y conocimiento del experto en la materia.

2.8.2 Estrategias

Para la adquisición de la información, se deben considerar varias estrategias las más comunes son entrevistas formales o informales a los expertos del conocimiento, otra podría ser la observación de cómo realizan cada uno su trabajo o finalmente a través de encuestas. Por ejemplo, las causas de alguna enfermedad, se las puede diagnosticar a través de un sinnúmero de síntomas.

2.8.3 Representación de Conocimiento

La representación del conocimiento, se basa en la representación de los objetos, identificación de eventos o escenarios, condiciones y reglas que intervienen para la realización o toma de una decisión por parte del sistema.

2.8.4 Búsqueda

En esta parte se aplican lo que se conocen como métodos de búsqueda, que en cierto modo deberían ser aplicados de acuerdo al escenario que se presente en cada caso, entre los más importantes se tienen:

- Criterios de control. Este método se basa, en ramificar o extender la solución del conflicto o viceversa a través de niveles como un árbol.
- Métodos heurísticos. Se basa en criterios cualitativos relacionados al mundo real como experiencias, evaluaciones, costos, estimados de aproximación, etc.

2.8.5 Inferencia de Conocimiento

Este es uno de las etapas más críticas, ya que en esta parte lo que realiza es un diagnóstico o toma de decisión final al problema o situación planteada al inicio. Por eso, debe realizarse un proceso anterior de pruebas, en la que el experto del conocimiento valide cada uno de los procesos o reglas que realiza el sistema para tomar un diagnóstico o conclusión final. (Garzón N. M., 2015).

2.9 Matriz de confusión y valores estadísticos

2.9.1 Matriz de confusión

La matriz de confusión, contiene información real y de predicción de diferentes clases, a través de la aplicación de algún sistema de clasificación. Su representación, viene dada a través de una matriz de clases al cuadrado. (Oliver A., 2008)

2.9.2 Kappa statistic

Kappa mide el porcentaje de valores de datos en la diagonal principal de la matriz de confusión y ajusta estos valores para la cantidad de acuerdo que se podría esperar debido a la casualidad.

2.9.3 Mean absolute error

Es la diferencia entre el valor medido y el valor real o verdadero.

2.9.4 Root mean squared error

Es el valor calculado entre la raíz cuadrada de la media dividido para la media del cuadrado del error.

3. Marco metodológico

Para el desarrollo del presente proyecto se ha seleccionado la metodología SCRUM, debido a la adaptabilidad y el manejo de iteraciones, de vital importancia para poder llevar a cabo dicho proyecto, según lo menciona la guía SBOK (SCRUMstudy, 2016).

3.1 Metodología SCRUM

SCRUM es la metodología ágil y flexible para gestionar el desarrollo de un proyecto, permite segmentar, dar una frontera de alcance al proyecto y en cualquier momento introducir cambios funcionales o de prioridad, sin inconveniente alguno. Promueve, además, la motivación y compromiso del equipo que forma parte del proyecto, y así poder desarrollar al máximo cada una de sus capacidades.

3.1.1 Beneficios

- Calidad de Software. El sistema de trabajo empleado junto con la obtención de una versión funcional en cada iteración, hace que se consiga una mayor calidad.
- Tolerancia a cambios. Se tiene una capacidad de reacción y adaptación alta ante los requerimientos generados por parte de los interesados.
- Pronosticación de tiempos. Con esta metodología es posible estimar fácilmente cuando se puede entregar una versión funcional de una iteración del proyecto.
- Cumplimiento de expectativas. En esta parte los usuarios indican su expectativa con respecto a los requerimientos, el cual es evaluado por el dueño del producto quién se encarga de dar prioridad a cada uno de estos.
- Reducir Riesgos. La acción de conocer las funcionalidades de mayor prioridad y como el equipo de trabajo lo está ejecutando, permite que se pueda despejar o anticipar a ciertos riesgos que afecten el transcurso o viabilidad del proyecto.

3.1.2 Elementos de la metodología

- Interesados. Son aquellas individuos o partes interesadas que pueden afectar al proyecto de alguna manera.
- Pila del producto. Es aquella que contiene todos los requerimientos funcionales y no funcionales dados por los interesados.

- Dueño del producto. Es el que conoce el giro del negocio y se encarga de que se trabaje de la manera más adecuada según la perspectiva del mismo.
- Entrega parcial (*Sprint*). Es un subconjunto de requerimientos, extraídas de la pila del producto para ser ejecutadas en un periodo de 1 a 4 semanas de trabajo, entregando en cada una de ellas un producto funcional, es un ciclo repetitivo, en donde se sigue una serie de instrucciones a realizar
- Maestro SCRUM. Es aquel que se encarga de guiar y encaminar al equipo de desarrollo para el logro de los objetivos planteados para la culminación del proyecto.
- Equipo de desarrollo. Responsables de entregar el producto y participar en el proyecto (SCRUMstudy, 2016).

3.2 Planificación de la metodología utilizando SCRUM

Tabla 1: Segmentación del proyecto (iteraciones)

NÚMERO DE ITERACIÓN	DESCRIPCIÓN	ESTIMACIÓN DE TIEMPO (días)
1	Identificación y validación de fuentes de información	
	1.1 Identificación de áreas de interés nacional y fuentes oficiales	15
	1.2 Validación de disponibilidad y calidad de la información obtenida	15
2	Análisis y obtención de datos	
	2.1 Análisis de la información disponible y visualización de las posibles áreas de interés.	7
	2.2 Obtención de las bases de datos con la información existente	23
3	Definición de estructura relacional de la base de datos operacional en PostgreSQL	30
4	Migración de la información disponible a la base operacional	30

5	Análisis y creación de la estructura dimensional para el almacén de datos	30
6	Creación de procedimientos almacenados	30
7	Análisis de la provisión de servicios de salud e implementación de tablas de hechos	30
8	Definición de áreas para la aplicación de herramientas de inteligencia de negocios	30
9	Integración de las herramientas Pentaho Report Designer y Weka	30
10	Análisis de resultados	30

Nota: El proyecto tiene diez iteraciones, las cuales tienen un tiempo máximo de duración de 30 días, cabe recalcar que la documentación se desarrolló en paralelo a las otras actividades, en todas las iteraciones.

3.2.1 Primera iteración: Identificación y validación de fuentes de información

3.2.1.1 Identificación de áreas de interés nacional y fuentes oficiales

Las áreas de interés nacional e institucional que consideramos son partes importantes para los análisis de nuestro proyecto se identifican según la matriz de interesados.

Tabla 2: Matriz de interesados

#	NOMBRE	CARGO/ROL	CONTACTO	INTERESES	INFLUEN-CIA
1	INEC	Institución Pública	0 22232-303 *www.ecuadorencifras.gob.ec*	Visualizar la utilización de la información levantada por el INEC, orientada al bienestar de la sociedad.	10

2	MSP Coordinación Zonal 9 (Pichincha)	Institución Pública	Teléfono: 2 3931020 Dirección: Juan León Mera N26-38 y Santa María MSP Teléfono: 2381-4400 *www.salud.gob.ec*	Visualizar los servicios de salud brindados por un establecimiento sea público o privado en cualquier región geográfica.	10
3	Javier Reyes, Jorge Quiroz (Estudiantes de la Universidad Politécnica Salesiana)	Miembros del proyecto de titulación	0989568063 *jreyesz1@est.ups.edu.ec;* 0999989465 *jquirozq@est.ups.edu.ec;*	Crear una herramienta que permita procesar la información publicada por el INEC para obtener información básica sobre los servicios de salud relacionados con la ubicación geográfica. Implementar proyecto de titulación para la	10

				obtención del título de ingeniero de sistemas.	
4	René Arévalo (Investigador de la Universidad Politécnica Salesiana)	Tutor / Docente	0995459598 *aarevalo@ups.edu.ec*	Evaluar la calidad del proyecto respecto a los parámetros indicados.	10

Nota: La matriz de interesados, permite identificar la influencia de los involucrados en el proyecto y disponer de información de contacto, además de sus intereses particulares

3.2.1.2 Validación de disponibilidad y calidad de la información obtenida

La disponibilidad de la información a través de las entidades públicas como es el caso del INEC, nos permite asegurarnos de la fiabilidad y calidad de esta información. Que, a través, de una ley aprobada por parte de la asamblea dice lo siguiente:

"La Ley Orgánica de Transparencia y Acceso a la Información Pública (LOTAIP) plantea la participación ciudadana y el derecho de acceso a la información relacionada con asuntos públicos, para ejercer un efectivo control y exigir la rendición de cuentas a las instituciones gubernamentales o aquellas que perciben recursos estatales." (Censos, 2016)

Mientras que, en la propia página web de la institución, dicha entidad pública transmite lo siguiente:

"Hoy el INEC se encuentra en un proceso de transparencia y de liberalización de bases de datos a través de nuestro compromiso con el país de entregarle cifras de calidad, de manera adecuada y oportuna. Para esto el Instituto ha implementado portales y servidores proveedores de información de las encuestas. (www.inec.gob.ec,www.ecuadorencifras.com, el Banco de Información y el Visualizador ESPAC) para garantizar la disponibilidad de este servicio. " (Censos, 2016)

Con el contexto anterior, se puede decir que existe un compromiso de calidad en cuanto a la información brindada por la institución, es por ello que existe un grado de fiabilidad de dicha información, la cual se usará para los respectivos análisis.

3.2.2 Segunda iteración: Análisis y obtención de datos

3.2.2.1 Análisis de la información disponible y visualización de las posibles áreas de interés.

Esto permite establecer las relaciones existentes entre la información provista por las fuentes de información (INEC).

Una vez estructurado el almacén de datos en PostgreSQL y cargada la información obtenida del INEC, comparando con información de otras fuentes, en anuarios, publicaciones o reportes se identifican áreas de interés y relaciones entre la información disponible para posteriormente efectuar los análisis correspondientes sobre estos puntos relevantes, que permite corroborar conclusiones como abstraer nueva información.

3.2.2.2 Obtención de las bases de datos con la información existente

Se procede a descargar la información concerniente a los años 2007 a 2014 de la página web del INEC. Como se ha mencionado anteriormente esta información se encuentra en formato *.SAV* y es la base para realizar el almacén de datos y efectuar los análisis.

3.2.3 Tercera iteración: Definición de estructura relacional de la base de datos operacional en PostgreSQL

Una vez que se dispone de toda la información obtenida y que se ha seleccionado los datos de interés, es importante establecer la estructura de la base operacional, a través de la creación de tablas y relaciones mediante claves primarias, foráneas y únicas. Con esto, se obtiene un escenario preparado para albergar los datos de manera ordenada y sistemática en el gestor de base de datos mencionado.

3.2.4 Cuarta iteración: Migración de la información disponible a la base operacional

En el Banco de Información del INEC, estan disponibles bases de datos correspondientes a Recursos y Actividades de Salud desde el año 2000 al 2014, de donde se toman los 8 últimos años. Estos datos están codificados en archivos con formato .*SA*V, propio de IBM SPSS (*Statistical Package for the Social Sciences*).

Para llevar a cabo este proceso se realiza un proceso de formateo mediante depuración de campos de las bases de datos disponibles, en los que se evidencian diferentes formatos, como para el campo Codenc, en unos años tiene anchura de 3 y en otros años de 7. Esto se logra analizando el formato de los campos y determinando un formato final, que integre, sin conflictos, los datos de todos los años considerados, permitiendo almacenar toda la información pertinente.

La depuración permite cumplir con las características que debe tener un almacén de datos de ser integrados al contener datos de forma uniforme. (Trujillo, Mazón, & Pardillo, 2010)

En esta fase, se hace el levantamiento de un diccionario de datos para poder conocer qué atributos vamos a tomar en cuenta para nuestro análisis y el tipo de dato que van a tomar cada uno de estos.

Un diccionario de datos ofrece una descripción de los datos la misma que se conoce como metadatos; es decir, datos acerca de los datos. (Connolly, 2005)

Después de esto, se procede a transformar los archivos SAV obtenidos del INEC a un formato conocido como CSV para poder migrar todos estos datos al gestor de base de datos PostgreSQL. El formato SAV es una extensión propia del programa informático estadístico SPSS, permite manipular grandes bases de datos. El formato CSV por su parte, permite la representación de datos a manera de una tabla, separando los datos por un punto y coma.

Entonces, se utiliza la herramienta SPSS para poder manipular los archivos *.SAV* y visualizar la información. En esta herramienta se elige la opción *guardar como* y se selecciona el formato para guardar los datos, en este caso, se elige la opción *formato delimitado por comas (CSV)* y con esto el archivo obtenido se lo puede utilizar en Microsoft Excel.

3.2.5 Quinta iteración: Análisis y creación de la estructura dimensional para el almacén de datos

Una vez que se dispone de la información estructurada en la base *operacional* con la selección de los datos de interés, se realiza la definición del modelo dimensional, en este caso modelo estrella, que alberga los datos con una orientación hacia su análisis empleando herramientas de inteligencia de negocio considerando las áreas críticas que se quieren analizar definidas en las tablas de hechos mediante el cruce de información de las dimensiones consideradas.

3.2.6 Sexta iteración: Creación de procedimientos almacenados

Mediante la elaboración de procedimientos almacenados que son programas, que ejecuta un conjunto de acciones específicas, dentro un gestor de base de datos, en este caso, orientadas a extraer, procesar, clasificar y cargar datos desde la base operacional hacia el almacén de datos.

3.2.7 Séptima iteración: Análisis de la provisión de servicios de salud e implementación de tablas de hechos

Una vez que se encuentra almacenada toda la información depurada y consistente en el almacén, se procede a determinar aquellos indicadores que se relacionen con entidades de salud y especialidades, en los sectores públicos o privados. Con esto se procede a elaborar procedimientos almacenados orientados a construir las tablas de hechos con información del cruce de datos dimensional.

3.2.8 Octava iteración: Definición de áreas para la aplicación de herramientas de inteligencia de negocios

Se identifican las áreas de interés y relaciones puntuales de datos, mediante varios criterios, a ser analizados en función de los resultados obtenidos de la minería de datos, que es un conjunto de técnicas que tienen como fin descubrir el conocimiento existente en grandes almacenes de información, tendencias, patrones y relaciones significativas entre datos, procesos o eventos que suceden dentro de una organización, aportando a éstas en la toma de decisiones, actividades gerenciales y ejecutivas, brindando mayores prestaciones que la de simplemente dar soporte a los procesos organizacionales más elementales. (Pérez López & Santín González, 2007).

3.2.9 Novena iteración: Integración de las herramientas Pentaho Report Designer y Weka

Una vez determinadas las áreas de interés, se procede a mostrar la información obtenida a través de la generación de reportes gerenciales empleando Pentaho y a aplicar algoritmos de inteligencia de negocios con Weka.

3.2.10 Décima iteración: Análisis de resultados

Una vez aplicadas las herramientas de inteligencia de negocios, se pueden establecer conclusiones, tendencias, patrones de comportamiento e información relevante sobre determinados puntos críticos en la provisión de servicios de salud en el país, mediante reportes, gráficas y representación tangibles, resultantes de la aplicación de análisis.

Esto se logra debido a que, los almacenes de datos contienen información en un modelo de datos común (estandarizado) con los que, una entidad pueda emplear los datos históricos para apoyar los procesos de tomar decisiones estratégicas.

4. Diseño y construcción

Un sistema de inteligencia de negocios está compuesto por tres capas: de datos, información y conocimiento. La primera está formada por los datos provistos por el INEC, los mismos que han pasado por un proceso de depuración y migración, y están contenidos como se ha expuesto anteriormente en la base operacional en PostgreSQL.

La capa de información está en la base almacén, para la cual se emplean procedimientos almacenados que permiten la creación del almacén de datos, el mismo que contiene las dimensiones geografía, entidad salud, especialidades y tiempo.

Finalmente, la capa de conocimiento corresponde al análisis aplicado a la capa de información con Pentaho.

4.1 Desarrollo de la capa de información

Para el desarrollo de la capa de información, se establecen ciertos pasos en detalle de lo que se quiere para el análisis del proyecto y como lo íbamos a resolver. Conforme a esto, nos enfocamos en lo siguiente (Calabria, 2011):

- Fuente de datos
- Planificación
- Modelo Relacional
- Procedimientos
- Modelo Dimensional
- Tabla Hechos
- Verificación de datos
- Resultados

4.1.1 Fuente de datos

Un paso primordial e importante es tener las fuentes de datos con la que se va a trabajar, en este caso los datos obtenidos fueron tomados principalmente del INEC, con las consideraciones detalladas en el numeral 3.2.4; así como otras fuentes secundarias como Ecuador en cifras, entre otros. Posteriormente, se deben elegir las herramientas o programas que ayudan en el proceso de carga y análisis de los datos.

Para el presente proyecto, se resolvió trabajar con herramientas Open Source, como sistema gestor de base de datos, pgAdmin (PostgreSQL) y para el análisis

de la información, Pentaho Report Designer. Ambas herramientas, son muy destacadas en cuanto al rendimiento y versatilidad de sus componentes y funcionalidades.

4.1.2 Planificación

Una vez, que se ha obtenido las fuentes de datos e instalado las herramientas a utilizar, se procede a asignar tareas y actividades a seguir, de forma secuencial, para construir el almacén de datos. A continuación, se describen brevemente (Rivadera, 2010)

- Creación de la Base de datos Operacional.
- Normalización de datos.
- Modelado y relacionamiento de tablas.
- Construcción de la base de datos Almacén.
- Creación de procedimientos de carga.
- Modelo Dimensional.
- Creación tablas de hechos.
- Verificación datos.
- Análisis datos y resultados.

4.1.3 Modelado Relacional

Para el modelado relacional, se debe explorar cada uno de los campos o atributos en común de cada archivo o fuente de datos de los años considerados. En este caso, se encuentran campos que hacen referencia a ciertos códigos, los cuáles se han convertido en tablas, ubicando su identificador y la descripción de cada una de ellas, para que los datos sean más comprensibles. Luego se procede, a relacionar las tablas con claves foráneas y también agregar restricciones (*constraints*) para lograr así, disponer de un conjunto de datos compacto, gestionando cualquier tipo de inconsistencia.

4.1.4 Procedimientos

Luego de haber modelado relacionalmente los datos, se procede a elaborar los procedimientos de carga a las tablas dimensionales y tablas de hechos. En este caso, hubo ciertos conflictos con algunos datos de indicadores que se procedió a descartar en los procedimientos para evitar inconsistencia de información.

4.1.5 Modelo Dimensional

Se considera el alcance de este proyecto definido en un inicio, que es crear un almacén de datos con información provista por el INEC sobre servicios de salud en el Ecuador, en los últimos ocho años.

El nivel de granularidad, dada la cantidad y calidad de datos disponibles en el banco de información del INEC, está dada por cuatro perspectivas que se definen como tablas dimensionales, las mismas que permiten clasificar la información de acuerdo a indicadores y atributos y cifras respectivamente.

El diseño dimensional, se lo efectúa según entidades y campos con mayor relevancia para el análisis del proyecto. Con la creación de los procedimientos estas dimensiones fueron cargadas desde la base operacional en la base almacén, y fueron los siguientes:

- Dimensión geografía. (Provincias, cantones y parroquias)
- Dimensión tiempo. (Año)
- Dimensión Entidades de salud. (Entidad, clase y tipo)
- Dimensión Especialidades. (Especialidades de Salud)

Modelo *Bubble chart* de dimensiones

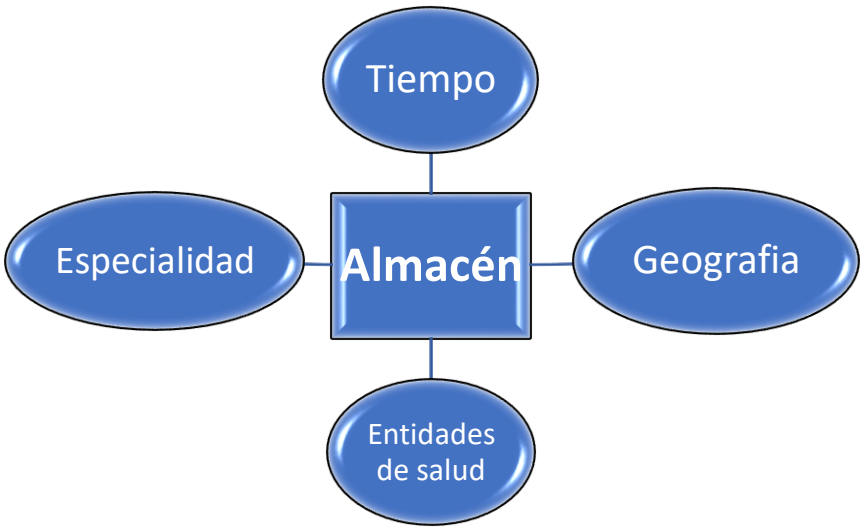


Figura 5: Modelado gráfico de alto nivel, gráfico de burbujas dimensional
Elaborado por: Jorge Quiroz y Javier Reyes

4.1.6 Tabla Hechos del almacén de datos

Las tablas de hechos, son una recopilación de información de las tablas dimensionales creadas previamente, este proyecto se obtiene dos tablas una referente a especialidades y otra a entidades de salud. Las mencionadas tablas de hechos, son combinaciones de cada uno de los registros de las tablas dimensionales y poseen cifras que permiten obtener información e indicadores, dotándose así de una característica especial e importante que es el nivel de detalle de toda la información disponible.

El presente almacén de datos (*datawarehouse*), que se encuentra en la base *almacen*, sigue un modelo estrella, que consta de dos tablas de hechos, y cuatro dimensiones, donde las tablas de hechos son las únicas en relacionarse con otras, es decir, con las tablas dimensionales. Este tipo de modelos, a diferencia del modelo copo de nieve, es sencillo de mantener y provee eficacia en la extracción de información. (Gálvez, 2015)

Directorio base de datos almacén

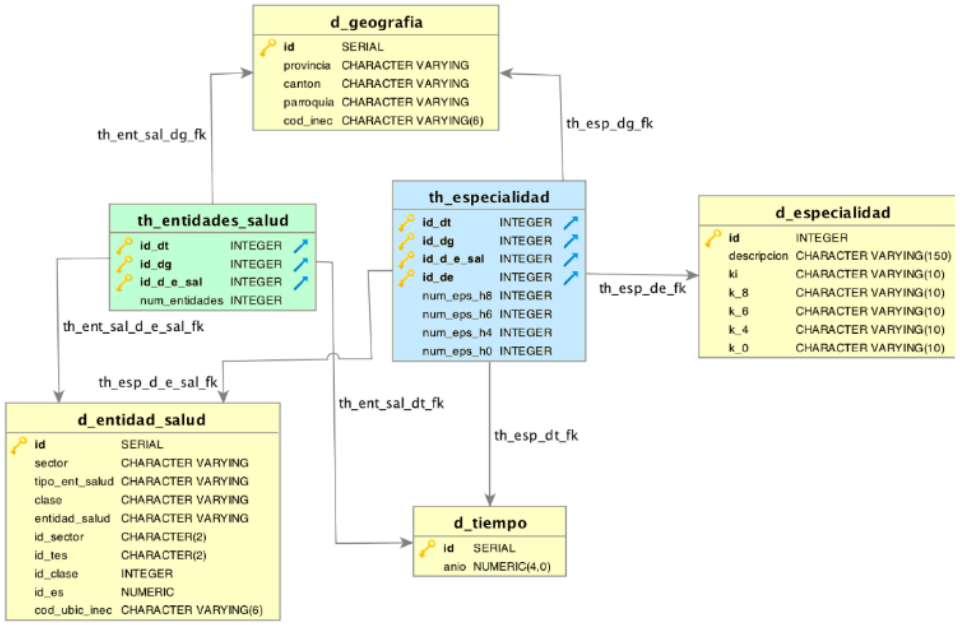


Figura 6: Modelo estrella tablas de hechos entidad salud y especialidad
Elaborado por: Jorge Quiroz y Javier Reyes

4.1.7 Verificación

Una vez que se ha realizado la carga de datos a las tablas de hechos, se procede a realizar pruebas manuales y a verificar que efectivamente estos resultados sean consistentes.

4.1.8 Resultados

Para una mejor optimización de la muestra de los resultados, se ha optado por crear vistas materializadas, ya que estas nos proporcionan el acceso a la información de manera más eficiente. Con esto, se puede levantar reportes utilizando la herramienta Pentaho y mostrar, según los criterios que se determinen, los resultados obtenidos.

Directorio base de datos almacén

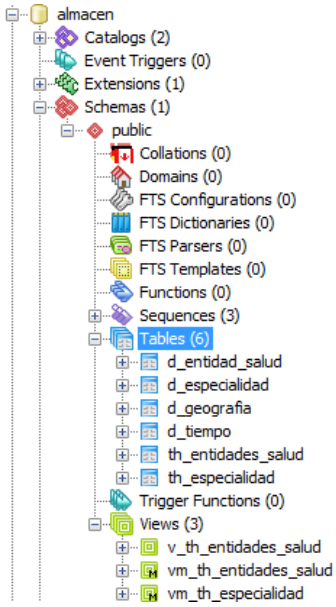


Figura 7: Vistas materializadas y tablas dimensionales
Elaborado por: Jorge Quiroz y Javier Reyes

4.2 Implementación del proceso ETL

Para la aplicación del proceso de extracción, transformación y carga, se considera la Metodología Kimball.

4.2.1 Extracción

En este caso puntual, la fuente de información oficial para recabar datos en este proyecto es el banco de información estadística del INEC, sección Recursos y Actividades de Salud, ubicado en su portal web, cuya información está

transparentada y disponible para todos los ciudadanos como puede corroborase en la figura 8. Aquí se encuentran disponibles bases de datos en formato *.SAV*, propio de IBM SPSS (*Statistical Package for the Social Sciences*). de los años 2007 a 2014, de donde se toman los 8 últimos años.

Lista de descargas del banco de información del INEC

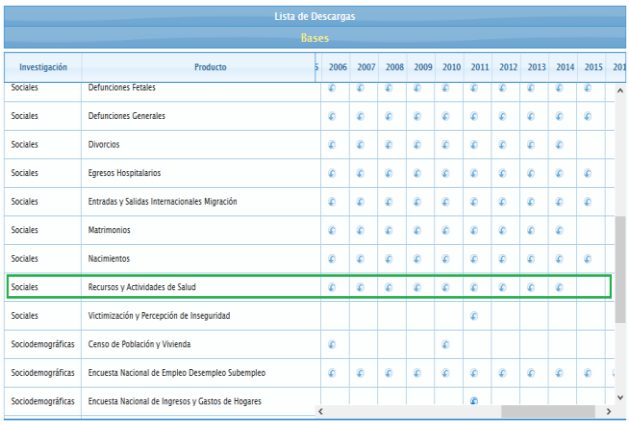


Figura 8: Recursos actividades de salud en la lista de descargas del banco de datos del INEC
Elaborado por: Jorge Quiroz y Javier Reyes

Tomando como ejemplo el año 2013 se puede observar que los archivos se componen de dos pestañas, *Vista de datos* y *Vista de variables*. Aquí se encuentran respectivamente, los datos correspondientes ordenados de acuerdo a las tablas *Codenc, Año, Establecimiento*, etc. Y los atributos de cada una de dichas variables como tipo, anchura, etiqueta, valores que pueden tomar, etc. (Figura 9).

Cabe mencionar que el INEC, sigue este mismo formato para todos los años.

Base de Recursos y Actividades de Salud, año 2013 (.sav)

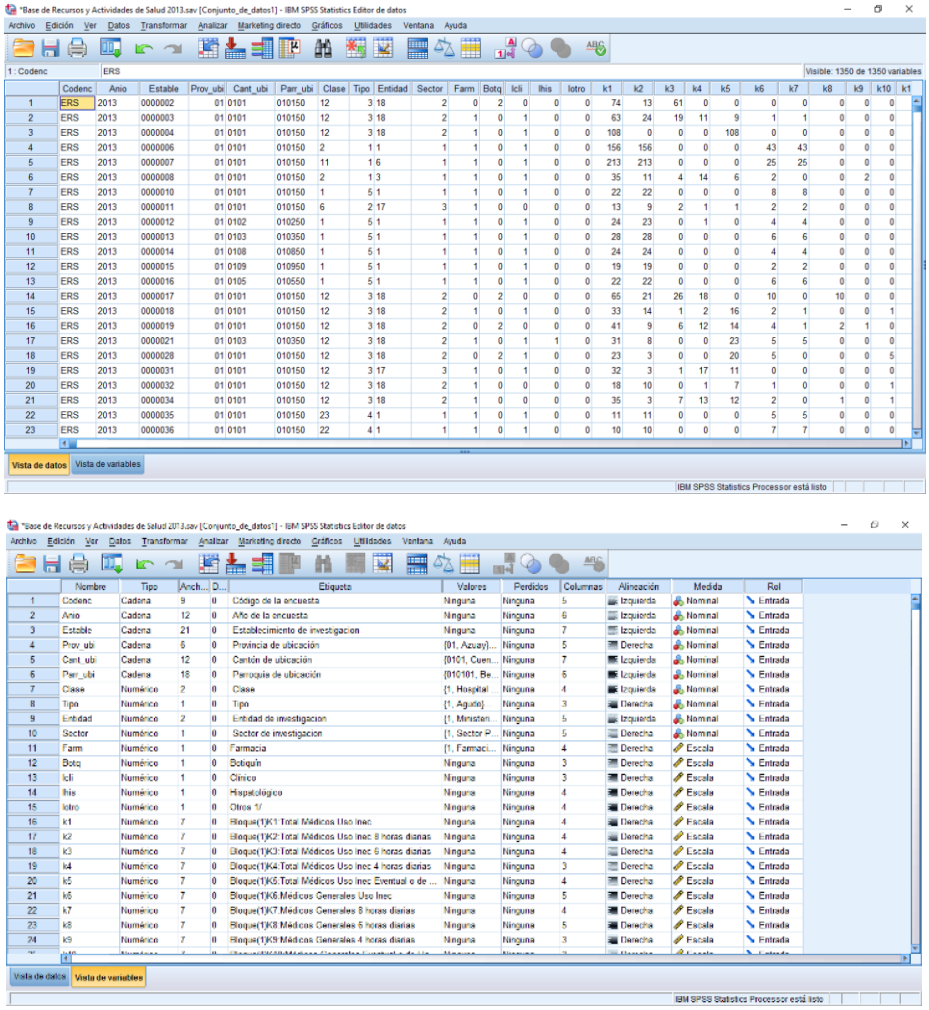


Figura 9: Pestañas “Vista de datos” y “Vista de variables” correspondientes al año 2011
Elaborado por: Jorge Quiroz y Javier Reyes

4.2.1 Transformación

Los datos recuperados en esta fase se encuentran alojados en la base operacional, que corresponde a la normalización de las bases de datos descargadas del INEC y sirve para calcular datos acumulados por áreas de interés.

Como puede observarse en la figura posterior, para esto se han creado un modelo de base de datos con datos que corresponden a los años seleccionados para el análisis (2007 a 2014), y otras tablas relacionadas con los datos de los ocho años como cantón, parroquia, etc.

Base de datos operacional

Figura 10: Estructura de la base de datos operacional
Elaborado por: Jorge Quiroz y Javier Reyes

4.2.1.1 Migración de SPSS a PostgreSQL

Se toma como referencia la tabla serv_salud_2008, este proceso se realiza siguiendo los mismos principios para los demás años, salvo en los casos en los que se ha visto necesario depurar los datos, estandarizarlos, y manipularlos a fin de poder realizar su integración en la base de datos operacional migrando desde SPSS hacia PostgreSQL, estos detalles son tratados más adelante, en depuración.

Lo primero que se realiza es crear la tabla serv_salud_año, correspondiente a uno de los años, en este caso el 2008, considerando todas las columnas que lo forman, tomando como referencia el modelo del archivo SPSS correspondiente, como el indicado en la Figura 9, donde cada variable constituye una columna de la tabla.

Para la definición del tipo de dato de cada columna, se considera un estándar que permita alojar todos los campos de la base de recursos por año en cada tabla.

Exportación de datos .CSV

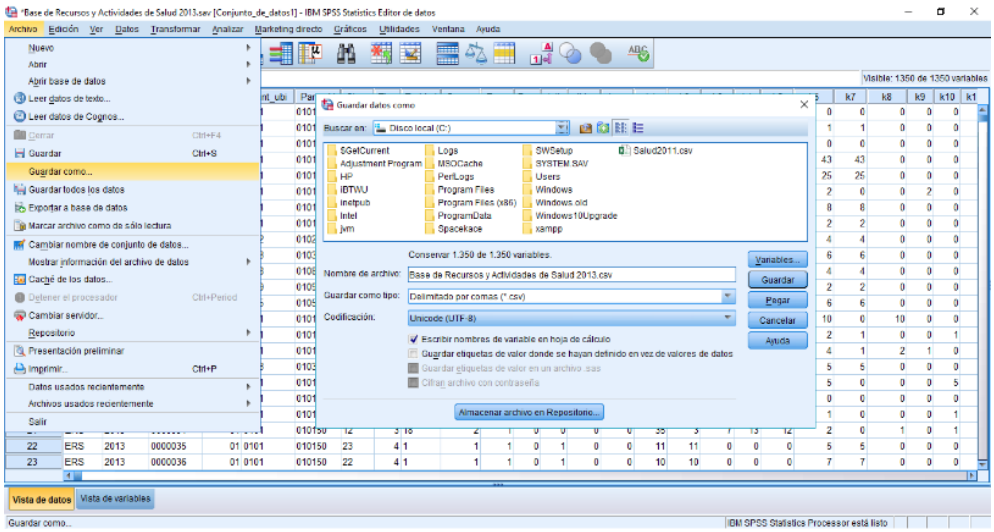


Figura 11: Procedimiento de exportación a formato delimitado por comas en SPSS
Elaborado por: Jorge Quiroz y Javier Reyes

Esto se realiza mediante un script por cada año, en este caso el que se encuentra en la Figura 13. Luego se procede a la carga de datos provenientes del archivo *.SAV* para esto, en SPSS en el base de recursos de salud del año en cuestión, en la pestaña *Vista de datos* se selecciona la opción *Archivo*, *guardar como*, se elige el directorio de ubicación, el tipo: delimitado por comas (*.*csv*), la codificación *UTF8* y se procede a guardar (Figura 11).

En el directorio seleccionado se puede verificar el archivo *.CSV* en el que se procede a eliminar las cabeceras (fila que contiene *Codenc*, *Anio*, *Estable*, etc.) y dejar solo los datos para evitar conflictos en la migración.

Archivo SPSS

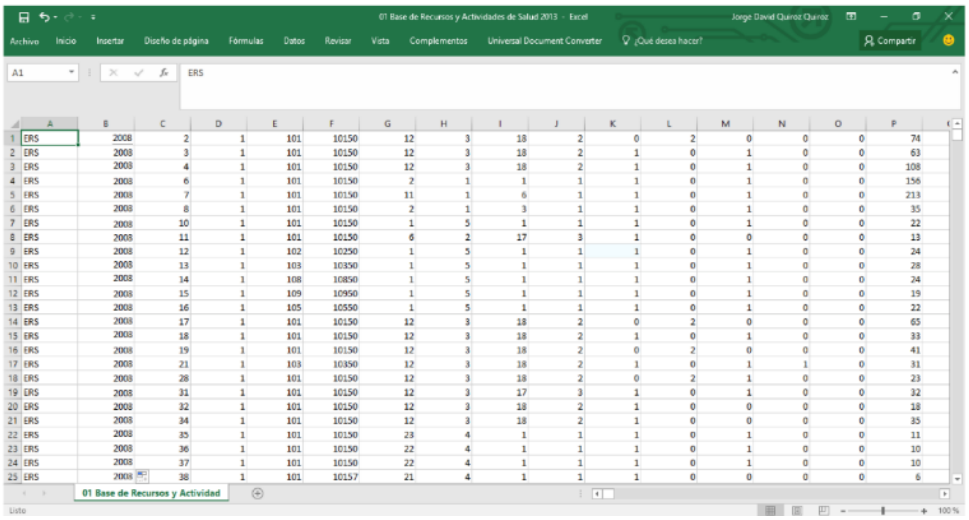


Figura 12: Resultado de exportación desde SPSS, de formato. SAV a .CSV y eliminación de cabeceras.
Elaborado por: Jorge Quiroz y Javier Reyes

A continuación, se procede a migrar los datos desde al archivo .*CVS* a la tabla del año correspondiente, esto se realiza ejecutando el siguiente *query* en PostgreSQL en el cual se indica el directorio en el que se encuentra el archivo *.csv* luego del *from*, indicando que es de tipo delimitado por comas, el mismo que se debe copiar o importar a la tabla correspondiente, en este caso: *serv_salud_2008.*

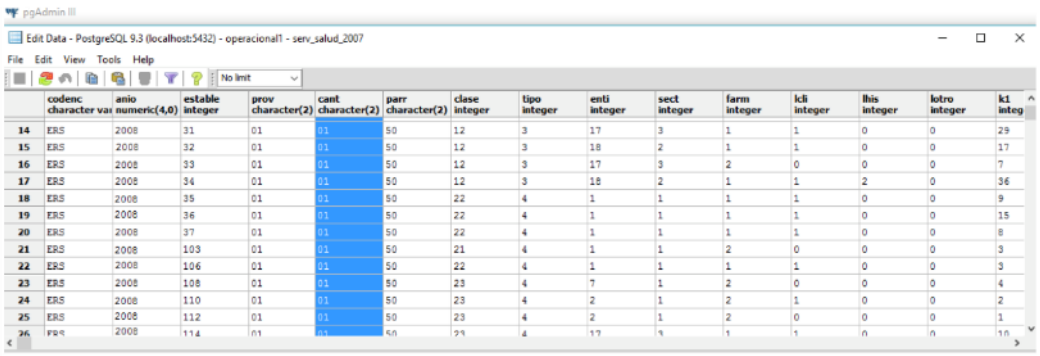
Instrucción para importar csv

```
copy  serv_salud_2008  from  'C:\Bases\recursos_salud_2008.csv'
with delimiter ';'
```

Figura 13: Query para importar datos desde formato .csv hacia la tabla serv_salud_2008
Elaborado por: Jorge Quiroz y Javier Reyes

Una vez efectuada esta instrucción podemos verificar que los datos se encuentran disponibles en la tabla *serv_salud_2008* en PostgreSQL.

Datos de la tabla 2008



	codenc character var	anio numeric(4,0)	estable integer	prov character(2)	cant character(2)	parr character(2)	clase integer	tipo integer	enti integer	sect integer	farm integer	lcli integer	lhis integer	lotro integer	k1 integ
14	ERS	2008	31	01	01	50	12	3	17	3	1	1	0	0	29
15	ERS	2008	32	01	01	50	12	3	18	2	1	1	0	0	17
16	ERS	2008	33	01	01	50	12	3	17	3	2	0	0	0	7
17	ERS	2008	34	01	01	50	12	3	18	2	1	1	2	0	36
18	ERS	2008	35	01	01	50	22	4	1	1	1	1	0	0	9
19	ERS	2008	36	01	01	50	22	4	1	1	1	1	0	0	15
20	ERS	2008	37	01	01	50	22	4	1	1	1	1	0	0	8
21	ERS	2008	103	01	01	50	21	4	1	1	2	0	0	0	3
22	ERS	2008	106	01	01	50	22	4	1	1	1	1	0	0	3
23	ERS	2008	108	01	01	50	23	4	7	1	2	0	0	0	4
24	ERS	2008	110	01	01	50	23	4	2	1	2	1	0	0	2
25	ERS	2008	112	01	01	50	23	4	2	1	2	0	0	0	1
26	ERS	2008	114	01	01	50	23	4	17	3	1	1	0	0	10

Figura 14: Consulta de datos en la tabla serv_salud_2007
Elaborado por: Jorge Quiroz y Javier Reyes

Tabla entidad_salud_inec

Esta tabla está construida con la estructura que puede observarse en la Figura 15.

Tabla entidad_salud_inec

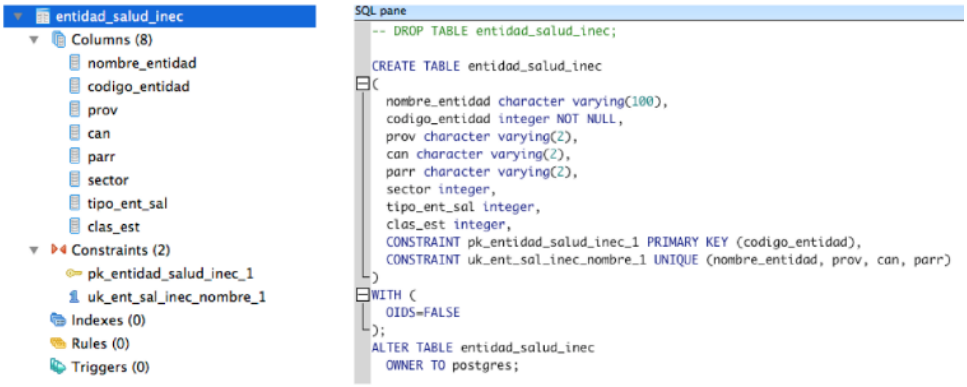


Figura 15: Campos y constraints de la tabla entidad_salud_inec
Elaborado por: Jorge Quiroz y Javier Reyes

Se utiliza la información provista por el INEC, la misma que está disponible en el Archivo Nacional de Datos y Metadatos Estadísticos (ANDA), en donde se encuentra la documentación de las bases de datos que utiliza el INEC (Anda.inec.gob.ec, 2017) correspondiente a Estadísticas Hospitalarias Camas y Egresos 2012, la cual ha sido creada en febrero de 2015, en la que se encuentra la información sobre las entidades de salud, su código y su nombre o categoría, como se muestra en la Figura 16. Esta información, se la guarda en un documento separado por comas (*CSV*) en Excel, como puede observarse en la Figura 17, lo que permite importar esta información en la tabla anteriormente creada.

Estadísticas Hospitalarias Camas y Egresos

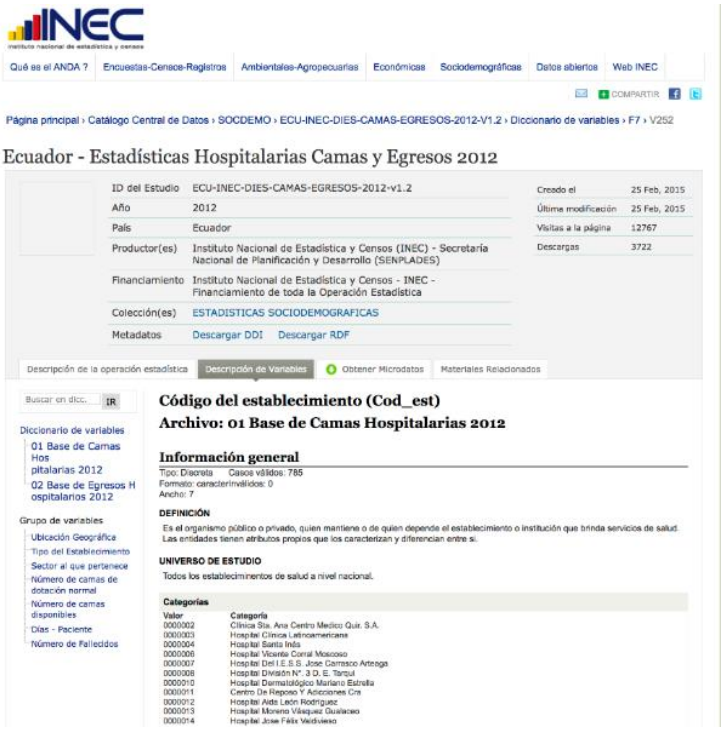


Figura 16: Listado de establecimiento y códigos
Elaborado por: Jorge Quiroz y Javier Reyes

Entidades de salud

	A	B	C	D
1	2	Clínica Sta. Ana Centro Medico Quir. S.A.		
2	3	Hospital Clínica Latinoamericana		
3	4	Hospital Santa Inés		
4	6	Hospital Vicente Corral Moscoso		
5	7	Hospital Del I.E.S.S. Jose Carrasco Arteaga		
6	8	Hospital División N°. 3 D. E. Tarqui		
7	10	Hospital Dermatológico Mariano Estrella		
8	11	Centro De Reposo Y Adicciones Cra		
9	12	Hospital Aida León Rodríguez		
10	13	Hospital Moreno Vásquez Gualaceo		
11	14	Hospital Jose Félix Valdivieso		
12	15	Hospital San Sebastián		
13	16	Hospital Cantonal Paute		
14	17	Clínica Paucarbamba		
15	18	Clínica Bolívar		
16	19	Clínica España S.A.		
17	21	Clínica Santa Bárbara		
18	28	Centro Quirúrgico Metropolitano		
19	31	Maternidad San Martin De Porres		
20	32	Clínica Santa Cecilia		
21	34	Clínica La Paz		

Figura 17: Listado de entidades de salud en formato CSV
Elaborado por: Jorge Quiroz y Javier Reyes

Con esto, se tiene como resultado que en la tabla se carga la información correspondiente a las dos primeras columnas, código y nombre de la entidad, como se evidencia en la Figura 18.

Para completar las demás columnas, se procede a crear un script en PostgreSQL (Figura 19) que compare, para cada una de las tablas *serv_salud_año*, correspondientes a todos los años, el código de la entidad (*estable*) con el de la tabla *entidad_salud_inec* y actualice los campos faltantes desde provincia hasta clase de establecimiento con la información respectiva.

Campos de la tabla entidad_salud_inec

	codigo_entid integer	nombre_entidad character varying(100)	prov character var	can character var	parr character var	sector integer	tipo_ent_sal integer	clas_est integer
1	2	Clínica Sta. Ana Centro Medico Quir. S.A.						
2	3	Hospital Clínica Latinoamericana						
3	4	Hospital Santa Inés						
4	6	Hospital Vicente Corral Moscoso						
5	7	Hospital Del I.E.S.S. Jose Carrasco Arteaga						
6	8	Hospital División N°. 3 D. E. Tarqui						
7	10	Hospital Dermatológico Mariano Estrella						
8	11	Centro De Reposo Y Adicciones Cra						
9	12	Hospital Aida León Rodríguez						
10	13	Hospital Moreno Vásquez Gualaceo						
11	14	Hospital Jose Félix Valdivieso						
12	15	Hospital San Sebastián						
13	16	Hospital Cantonal Paute						
14	17	Clínica Paucarbamba						
15	18	Clínica Bolívar						
16	19	Clínica España S.A.						
17	21	Clínica Santa Bárbara						
18	28	Centro Quirúrgico Metropolitano						
19	31	Maternidad San Martín De Porres						
20	32	Clínica Santa Cecilia						

Figura 18: Tabla entidad_salud_inec con los campos código y nombre de entidad
Elaborado por: Jorge Quiroz y Javier Reyes

Carga de datos en tabla entidad_salud_inec

```
UPDATE establecimientos_salud_t
SET prov = serv_salud_2007.prov,
  can= serv_salud_2007.cant,
  parr= serv_salud_2007.parr,
  sector=serv_salud_2007.sect,
  tipo_ent_sal=serv_salud_2007.enti,
  clas_est=serv_salud_2007.clase
FROM serv_salud_2007
WHERE serv_salud_2007.estable = codigo_entidad
```

Figura 19: Script para actualizar información de campos de la tabla
Elaborado por: Jorge Quiroz y Javier Reyes

Este procedimiento se lo repite para cada año, consecutivamente, con lo que se va completando la información de la tabla. De esta manera, con el año 2007 se completan 668 campos; con el año 2008,708: para el 2009, 754 y así sucesivamente hasta completar un total de 864 registros con el último año, el 2014 (Figura 20)

Datos de la tabla entidad_salud_inec

	codigo_entid integer	nombre_entidad character varying(100)	prov character var	can character var	parr character var	sector integer	tipo_ent_sal integer	clas_est integer
853	841	Hospital Cantonal La Libertad	24	03	52	1	1	1
854	866	Clínica Cadena (Climaco)_	24	02	50	2	18	12
855	840	Clínica De La Base Naval De Salinas	24	03	50	1	3	23
856	843	Hospital Cantonal Manglar alto	24	01	54	1	1	1
857	824	Policlínico Mater. La CigÜeña (Granados)	24	03	52	2	18	12
858	2008316	Clínica Baste	24	02	50	2	18	12
859	2008315	Clínica Maternidad Garcia	24	03	50	2	18	12
860	842	Hospital Del I.E.S.S. Ancón	24	01	56	1	6	2
861	2009332	Clínica Santa Martha	24	02	50	2	18	12
862	848	Hospital Cantonal De Salinas-Dr. Jose Gar.	24	03	50	1	1	1
863	2003410	Hospital Alcívar "La Península"	24	02	50	2	18	12
864	2012307	Hospital Liborio Panchana	24	01	50	1	1	2

Figura 20: Tabla entidad_salud_inec con datos en todos los campos.
Elaborado por: Jorge Quiroz y Javier Reyes

Debido a que a partir del año 2011 los campos cantón (cant) y parroquia (parr) disponen de una longitud de cuatro y seis caracteres respectivamente, para que la tabla que se muestra en la figura anterior conserve su formato de dos caracteres para cada campo, se emplea el script que se indica en la figura 21.

Corrección de datos en tabla entidad_salud_inec

```
--Estandarización de campos canton y parroquia
update establecimientos_salud_t
SET can = substring(can,3,2)
where char_length(can)=4


update establecimientos_salud_t
SET parr = substring(parr,5,2)
where char_length(parr)=6
```

Figura 21: Script para ajustar tamaño de determinados campos de campos de la tabla
Elaborado por: Jorge Quiroz y Javier Reyes

4.2.1.2 Depuración de datos

Como se ha señalado en un comienzo consiste en realizar correcciones en las inconsistencias encontradas con el fin de obtener datos sin inconsistencias que sean útiles y sigan un formato determinado

A continuación, se menciona algunas de las tareas que se han llevado a cabo en el proceso de depuración

En el proceso de migración de datos a PostgreSQL, que se ha señalado anteriormente, al momento de importar los datos en formato CSV, un cuadro de diálogo mostró un error de sintaxis Figura 22.

En el archivo *.SAV* correspondiente al año 2012 en las tablas *sininter* (sin intervención) y *coninter* (con intervención), correspondientes a actividades con intervención y sin intervención, respectivamente, se encuentran celdas en blanco (Figura 23)**,** lo que ocasiona el mencionado conflicto, al momento de migrar los datos contenidos en estos campos a PostgreSQL, en donde las columna están definida como tipo entero (*integer*) en ambos casos.

Mensaje de inconsistencia de valores

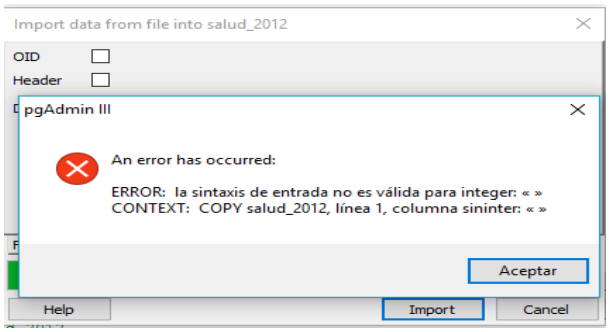


Figura 22: Cuadro de diálogo error (PgAdmin, Postgresql)
Elaborado por: Jorge Quiroz y Javier Reyes

Datos de la tabla 2008

region	coninter	sininter	area
2,00	1,00	.	1,00
2,00	1,00	.	1,00
2,00	1,00	.	1,00
1,00	1,00	.	1,00
1,00	1,00	.	1,00
1,00	1,00	.	1,00
2,00	.	1,00	1,00
1,00	1,00	.	1,00
1,00	1,00	.	1,00
1,00	1,00	.	1,00
1,00	1,00	.	1,00
1,00	1,00	.	2,00
1,00	1,00	.	1,00
1,00	1,00	.	1,00

Figura 23: Vista de datos 2012 (SPSS)
Elaborado por: Jorge Quiroz y Javier Reyes

El centro de conocimiento de IBM, define este particular como “valores perdidos” en SPSS, lo cual se debe a que no se presenta información en estos campos o sin campos en blanco (IBM, 2016)

Según el Departamento de Matemáticas Aplicadas de la Universidad Complutense de Madrid, es menester reflexionar acerca del riesgo que conlleva la sustitución de estos datos perdidos, no se puede determinar que los datos faltantes sean por puro azar, o si se debe a otro factor desconocido, sin embargo no se puede reemplazar estos valores por otros en base a algún razonamiento o criterio ya que esto provocaría una “diferencia estadísticamente significativa” que afectaría además la dispersión (varianza) de los datos en cuestión (Sánchez, Vargas, & otros, 2016)

Dada esta situación de ausencia de datos, se procede a reemplazar por “0” estos valores para que no produzcan problemas de carga en el gestor de base de datos, y posteriormente se establece estos campos como valores nulos para no alterar la data, para lo cual se emplearon consultas como la que se puede observar en a fin de no alterar la información disponible.

Mensaje de inconsistencia de valores

```
update serv_salud_20xx
set noninter=null
where coninter=0

update serv_salud_20xx
set sininter=null
where sininter=0
```

Figura 24: Cuadro de diálogo error (PgAdmin, PostgreSQL)
Elaborado por: Jorge Quiroz y Javier Reyes

Con el objetivo de determinar el número de incidencias de este tipo, se procede a realizar un análisis estadístico descriptivo de frecuencias (E-stadistica.bio.ucm.es, 2017) para las variables en cuestión, lo cual arroja como resultado el número de valores perdidos respectivamente.

Para ello, en SPSS de IBM se selecciona la pestaña *Analizar*, la opción *Estadístico-Descriptivo*, se selecciona frecuencias y se añade la variable que se necesita (*sininter* y *coninter*).

En la figura 25 se puede visualizar que *sininter* tenía 735 valores nulos o perdidos; mientras que *coninter*, 735, ambos de un total de 3280 valores.

Mensaje de inconsistencia de valores

```
FREQUENCIES VARIABLES=sininter
  /ORDER=ANALYSIS.
```

Frecuencias

[Conjunto_de_datos1] C:\Users\Jorge Quiroz\Documents\Mis Documentos\

Estadísticos

sininter

N	Válidos	3280
	Perdidos	735

sininter

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	1,00	3280	81,7	100,0	100,0
Perdidos	Sistema	735	18,3		
Total		4015	100,0		

```
FREQUENCIES VARIABLES=coninter
  /ORDER=ANALYSIS.
```

Frecuencias

[Conjunto_de_datos1] C:\Users\Jorge Quiroz\Documents\Mis Documentos\

Estadísticos

coninter

N	Válidos	735
	Perdidos	3280

coninter

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	1,00	735	18,3	100,0	100,0
Perdidos	Sistema	3280	81,7		
Total		4015	100,0		

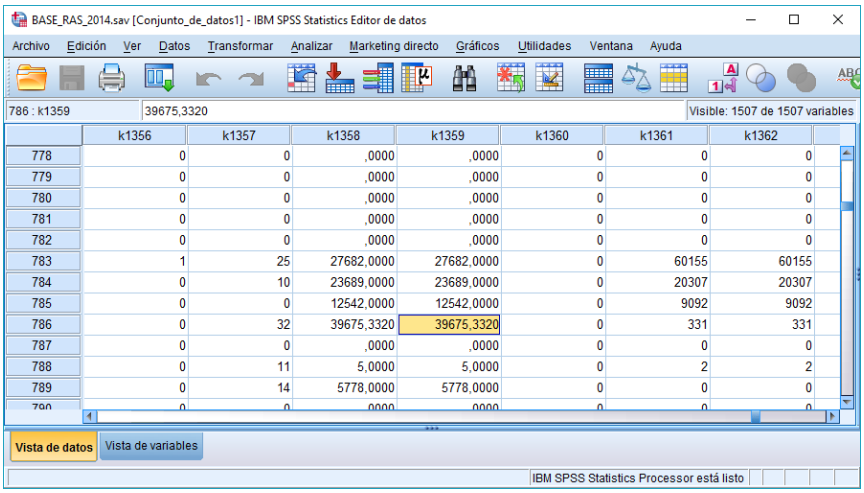
Figura 25: Tabla estadística descriptiva de frecuencias para variables sininter y cointer
Elaborado por: Jorge Quiroz y Javier Reyes

Otro inconveniente, se presenta al intentar migrar los datos correspondientes al año 2014, al encontrarse valores en los indicadores *K1358, K1359* y otros, valores con decimales, cuando estos se han definido como tipo *integer* (Figura 26)

Para solucionar este error, se procede a eliminar los decimales, ya que representa una inconsistencia en los campos correspondientes a *No.tratam.fisioterapia* de electroterapia uso inec y *No.tratam.fisioterapia* de electroterapia *cons.externa* respectivamente ya que estos valores no deben ser enteros, ya que indican el número de tratamientos de fisioterapia.

Para ello se procede a eliminar la parte decimal del archivo *.CSV* en Microsft Excel, como paso previo la importación de información desde PostgreSQL (Figura 28)

Valores inconsistentes



	k1356	k1357	k1358	k1359	k1360	k1361	k1362
778	0	0	,0000	,0000	0	0	0
779	0	0	,0000	,0000	0	0	0
780	0	0	,0000	,0000	0	0	0
781	0	0	,0000	,0000	0	0	0
782	0	0	,0000	,0000	0	0	0
783	1	25	27682,0000	27682,0000	0	60155	60155
784	0	10	23689,0000	23689,0000	0	20307	20307
785	0	0	12542,0000	12542,0000	0	9092	9092
786	0	32	39675,3320	39675,3320	0	331	331
787	0	0	,0000	,0000	0	0	0
788	0	11	5,0000	5,0000	0	2	2
789	0	14	5778,0000	5778,0000	0	0	0

Figura 26: Visualización de valores decimales en campo de datos enteros (SPSS)
Elaborado por: Jorge Quiroz y Javier Reyes

Corrección de inconsistencia de datos enteros

AZT	AZU	AZV
0	0,00	0,00
0	0,00	0,00
0	0,00	0,00
0	0,00	0,00
0	0,00	0,00
0	0,00	0,00
0	0,00	0,00
25	27682,00	27682,00
10	23689,00	23689,00
0	12542,00	12542,00
32	39675,33	39675,33
0	0,00	0,00
11	5,00	5,00
14	5778,00	5778,00
0	0,00	0,00
0	0,00	0,00
6	3240,00	3240,00

	AZU	AZV
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
25	27682	27682
10	23689	23689
0	12542	12542
32	39675	39675
0	0	0
11	5	5
14	5778	5778
0	0	0
0	0	0
6	3240	3240

Figura 27: Eliminación de la parte decimal de valores, con el fin de garantizar la calidad de los datos y disponer de datos consistentes.
Elaborado por: Jorge Quiroz y Javier Reyes

4.2.3 Carga

Una vez que se ha llevado a cabo el proceso de transformación, se procede a transcribir la base operacional al almacén de datos, realizando un análisis en función de las áreas de interés en el tiempo, expresadas en cuatro tablas dimensionales y

dos tablas de hechos, siguiendo el modelo estrella como se lo puede verificar en la figura 6 mediante procedimientos almacenados.

4.2.3.1 Procedimientos almacenados

Para llevar a cabo la carga del almacén de datos, como primer punto, se crea un apuntador desde la base operacional, que permite almacenar los datos transformados y limpios en el almacén de datos.

Una vez realizado lo mencionado anteriormente, se procede con la construcción de procedimientos almacenados (funciones) en la base operacional, que, apuntando al almacén, se encargan de la carga de registros correspondientes.

Dichas funciones están conformadas mediante cursores que permiten manipular los datos arrojados de una consulta que son alojados temporalmente en una variable de tipo cursor.

Esto permite seleccionar la información requerida, proveniente de la consulta, se almacena en una tabla virtual (Silberschatz, K, Korth, H, 2002), para cargar los registros en las tablas dimensionales del almacén.

Las funciones declaradas se explican a continuación.

Función carga_d_geografia

Este procedimiento tiene por objetivo cargar los datos respectivos en la dimensión geografía del almacén, en el que consideran un identificador de tipo serial, los nombres de la parroquia, el código INEC correspondiente y la relación con el cantón al que pertenece y la provincia correspondiente a este último.

Función carga_dgeografia 1 de 2

```
CREATE OR REPLACE FUNCTION carga_dgeografia()
  RETURNS integer AS
$BODY$
 DECLARE

cur_geografia_p cursor for select cod_provincia,provincia
        from provincia
        where (cod_provincia) not in (select cod_inec from d_geografia)
                order by cod_provincia;

 cur_geografia_c cursor for select c.cod_provincia,provincia,c.cod_canton,canton
        from provincia p, canton c
        where p.cod_provincia = c.cod_provincia
        and (c.cod_provincia||c.cod_canton) not in (select cod_inec from d_geografia)
                order by cod_provincia,cod_canton;

cur_geografia_pa cursor for select p.cod_provincia,p.provincia,c.cod_canton,c.canton,pa.cod_parroquia, parroquia
        from provincia p, canton c, parroquia pa
        where p.cod_provincia = pa.cod_provincia
        and c.cod_canton = pa.cod_canton
        and c.cod_provincia = p.cod_provincia
        and cod_inec not in (select cod_inec from d_geografia)
                order by p.cod_provincia,c.cod_canton,pa.cod_parroquia;
```

Figura 28: Declaración de la función y cursores
Elaborado por: Jorge Quiroz y Javier Reyes

En primera instancia, se declara la función, en cuyo cuerpo se incluyen tres cursores geografia_p, geografia_c y geografia_pa.

El primero contiene la consulta del código de provincia y su nombre de la tabla provincia, el segundo, además incluye el código de cantón y su respectivo nombre y el tercero adiciona la información pertinente a parroquia.

Función carga_dgeografia 2 de 2

```
v_cod_prov  parroquia.cod_provincia%type;
v_provincia provincia.provincia%type;
v_cod_cant  parroquia.cod_canton%type;
v_canton    canton.canton%type;
v_cod_parr parroquia.cod_parroquia%type;
v_parroquia parroquia.parroquia%type;
v_existe integer := 3;

BEGIN
   open cur_geografia_p;

   fetch cur_geografia_p into  v_cod_prov, v_provincia;
   while found loop
     insert into d_geografia (id,provincia,canton,parroquia,cod_inec)
                      values (nextval('d_geografia_id_seq'), v_provincia, null, null, v_cod_pr
   fetch cur_geografia_p into  v_cod_prov, v_provincia;
   end loop;
   if not found then
      raise notice 'Todas las provincias estan incluidas (%)',found;
      close cur_geografia_p;
      v_existe := 0;
   end if;
```

Figura 29: Apertura y recorrido en los cursores
Elaborado por: Jorge Quiroz y Javier Reyes

A continuación, se abre el cursor geografía y se realiza el recorrido de la consulta con el comando *fetch* el cual se encarga de la inserción de los campos id, provincia, canton, parroquia y cod_inec mediante un bucle while que realiza iteraciones hasta barrer con todos los registros mediante la instrucción found.

En caso de haber completado todos los registros se desplegará el mensaje: “Todas las provincias están incluidas”.

Se siguen los mismos pasos con los cursores correspondientes a cantón y parroquia.

Función cargar_entidad_salud

Tiene por objetivo hacer el cruce de los datos de las tablas sector, tipo entidad y clase de establecimientos de salud que se encuentran en la base de datos operacional hacia la dimensión entidad salud de la base de datos almacén.

Función cargar_entidad_salud parte 1 de 4

```
-- Function: cargar_entidad_salud()

-- DROP FUNCTION cargar_entidad_salud();

CREATE OR REPLACE FUNCTION cargar_entidad_salud()
  RETURNS integer AS
$BODY$
 DECLARE
x integer :=1;

cur_sector cursor
    for select codigo, descripcion
        from sector
        --where (descripcion,null,null,null) not in (select sector, tipo_ent_salud, clase, entidad_salud  from d_entidad_salud
        where (descripcion) not in (select sector  from d_entidad_salud)
        order by codigo;

cur_sector_tipoentidad cursor
    for select s.codigo, s.descripcion,
               tes.codigo, tes.descripcion
                from sector s, tipo_entidad_salud tes,
                     (select distinct sect, tipo from serv_salud_2013 order by 1,2) egh
        where s.codigo =  egh.sect and tes.codigo = egh.tipo
        and (tes.descripcion ) not in (select tipo_ent_salud from d_entidad_salud where tipo_ent_salud is not null )
        order by s.codigo,tes.codigo;
```

Figura 30: Creación de procedimiento cargar_entidad_salud()
Elaborado por: Jorge Quiroz y Javier Reyes

Al igual que en el caso anterior, se procede a declarar cursores y realizar los cruces de los datos de las tablas sector, tipo entidad y clase de establecimientos de salud.

Tabla de hechos entidad salud (th_entidad_salud)

En primer lugar, se crean los siguientes cursores para la dimensión entidad salud respectivos:

- Cur_ent_salud_sector. Cursor de carga de los tipos de sectores salud.
- Cur_ent_salud_sec_t_e_s. Cursor de carga de los tipos de entidad salud.
- Cur_ent_salud_sec_t_e_s_c. Cursor de carga de las clases de salud.

De igual manera, se crean los cursos para la dimensión geografía salud respectivos:

- Cur_geo_prov. Cursor de carga para las provincias.
- Cur_geo_prov_can. Cursor de carga para los cantones.
- Cur_geo_prov_can_parr. Cursor de carga para las parroquias.

Procedimiento th_entidades_salud parte 2 de 4

```
CREATE OR REPLACE FUNCTION cargar_entidad_salud_hechos(tabla character varying)
  RETURNS integer AS
$BODY$
 DECLARE
query text;
v_anio integer:=substring(tabla,12,4);  --variable anio
id_t integer; --id de tiempo


--Dimension Entidad salud
cur_ent_salud_sector cursor --sector: publico, priv con y sin fines de lucro
    for select id,id_sector,id_tes,id_clase  from d_entidad_salud where id_es is null and id_clase is null and id_tes is null
        order by id;

cur_ent_salud_sec_t_e_s cursor -- t_e_s:tipo_entidad_salud, Clinicas particulares, hospital general, sin tipo, etc.
    for select id,id_sector,id_tes,id_clase  from d_entidad_salud where id_es is null and id_sector is not null and id_tes is
        and id_clase is null
        order by id;

cur_ent_salud_sec_t_e_s_c cursor --c: clase, hospital basico, clinica privada, dispensario, etc.
    for select id,id_sector,id_tes,id_clase  from d_entidad_salud where id_es is null and id_sector is not null and id_tes is
        and id_clase is not null
        order by id;


--Dimension geografia
cur_geo_prov cursor
    for select id, substr(cod_inec,1,2) prov from d_geografia
        where length(cod_inec)=2
    order by id;

cur_geo_prov_can cursor
    for select id, substr(cod_inec,1,2) prov, substr(cod_inec,3,2) can  from d_geografia
        where length(cod_inec)=4
    order by id;

cur_geo_prov_can_parr cursor
    for select id, substr(cod_inec,1,2) prov, substr(cod_inec,3,2) can, substr(cod_inec,5,2) parr  from d_geograf
        where length(cod_inec)=6
    order by id;
```

Figura 31: Cursores de la dimensión entidad salud
Elaborado por: Jorge Quiroz y Javier Reyes

Luego, se procede a declarar las variables que van a ser utilizadas para la carga

Procedimiento th_entidades_salud parte 3 de 4

```
v_id_e_s    d_entidad_salud.id%type;
v_id_sector   d_entidad_salud.id_sector%type;
v_id_egs  integer;--d_entidad_salud.id_tes%type;
v_id_clase d_entidad_salud.id_clase%type;

v_id_g    d_geografia.id%type;
v_id_prov  d_geografia.cod_inec%type;
v_id_can  d_geografia.cod_inec%type;
v_id_parr  d_geografia.cod_inec%type;

v_num_entidades integer;

v_existe integer := 2;
```

Figura 32: Variables para la carga de la tabla de hechos entidad salud
Elaborado por: Jorge Quiroz y Javier Reyes

A continuación, se procede a abrir los cursores, almacenar la información necesaria de manera temporal en las variables creadas anteriormente para finalmente con la sentencia *insert* proceder a cargar la información.

Procedimiento th_entidades_salud parte 4 de 4

```
  open cur_ent_salud_sector;

   -- inicia captura reg
   fetch cur_ent_salud_sector into  v_id_e_s, v_id_sector;

   while found loop -- Este found tiene true o false dependiendo si el fetch esta en el eof del archiv
      select id into v_id_g
        from d_geografia
        where provincia like 'Todas'; --Se agrega el campo 3367 para Todas en d_geografia
-----------
 query :=  'SELECT
        count(estable) from  '
         || quote_ident(tabla)
         || ' WHERE sect  = ' || v_id_sector::integer;

EXECUTE query into v_num_entidades ;
-----------------

select id into id_t
 from d_tiempo
where anio=v_anio;

     insert into th_entidades_salud(id_dt, id_dg, id_d_e_sal,num_entidades)
                            values (id_t, v_id_g, v_id_e_s,v_num_entidades);

   -- El siguiente registro del cursor 1...
     fetch cur_ent_salud_sector into  v_id_e_s, v_id_sector;
   end loop;

if not found then
      raise notice 'Todas los sectores estan incluidos (%)',found;
      close cur_ent_salud_sector;
      v_existe := 0;
   end if;
```

Figura 33: Carga en la tabla de hechos entidad salud y manejo de cursores
Elaborado por: Jorge Quiroz y Javier Reyes

Tabla de hechos especialidad (th_especialidad_hechos)

De manera similar, al procedimiento empleado para el almacenamiento de datos de la tabla de hechos de entidades de salud, en el presente caso, se emplean cursores por cada una de las dimensiones consideradas: entidad salud, geografía y especialidad (Figura 34)

Luego, se declara un conjunto de variables destinadas a almacenar la información previa al proceso de carga de datos, empleadas para almacenar temporalmente la información y efectuar las comparaciones correspondientes de acuerdo a los criterios establecidos como sector, parroquia, tipo de entidad, etc.

A continuación, se procede a trabajar con los cursores, primero se declaran las sentencias de apertura, el *sql* dinámico contenido en la instrucción *query*, la sentencia de inserción de los campos correspondientes en *th_especialidad* y finalmente el cierre de cada uno de los cursores, relacionando las dimensiones

mencionadas anteriormente con cada una de las posibles combinaciones con el objetivo de obtener el número de especialistas que trabajan 8, 6, 4 y 0 (ocasional) horas a nivel de país, provincia, cantón, parroquia, por cada sector, clase de establecimiento y entidad de salud.

Función cargar_especialidades_hechos (tabla) parte 1 de 2

```
-- Function: cargar_especialidades_hechos(character varying)

-- DROP FUNCTION cargar_especialidades_hechos(character varying);

CREATE OR REPLACE FUNCTION cargar_especialidades_hechos(tabla character varying)
  RETURNS integer AS
$BODY$
 DECLARE
query text;
v_anio integer:=substring(tabla,12,4);  --variable anio
id_t integer; --id de tiempo


--Dimension Entidad salud
cur_ent_salud_sector cursor --sector: publico, priv con y sin fines de lucro
    for select ........
    ......

cur_ent_salud_sec_t_e_s cursor -- t_e_s:tipo_entidad_salud,
                        --Clinicas particulares, hospital general, sin tipo, etc.
    for select ....
    .....

cur_ent_salud_sec_t_e_s_c cursor --c: clase, hospital basico, clinica privada,
                              --dispensario, etc.
    for select  ....
    .....

--Entidad salud
cur_ent_salud_sec_t_e_s_c_ent_sal cursor --entidad salud, establecimientos de salu
    for select ....
    .....

--Dimension geografia
cur_geo_prov cursor
    for select id, substr(cod_inec,1,2) prov from d_geografia
        ....


cur_geo_prov_can cursor
    for select  ....

cur_geo_prov_can_parr cursor
    for select ....
```

Figura 34: Declaración de cursores
Elaborado por: Jorge Quiroz y Javier Reyes

Función cargar_especialidades_hechos (tabla) parte 2 de 2

```
--Variables
 v_id_e_s    d_entidad_salud.id%type;
 v_id_sector   d_entidad_salud.id_sector%type;
 v_id_egs  integer;--d_entidad_salud.id_tes%type;
 v_id_clase d_entidad_salud.id_clase%type;
 v_id_es d_entidad_salud.id_es%type; --establecimiento salud

 v_id_g    d_geografia.id%type;
 v_id_prov  d_geografia.cod_inec%type;
 v_id_can  d_geografia.cod_inec%type;
 v_id_parr  d_geografia.cod_inec%type;

 v_id_espec integer; --variables cursor
 v_eps_h8 character varying (10);
....
....

------------------
-- Inicio de carga sector
------------------
  open cur_ent_salud_sector;

   -- inicia captura reg
   fetch cur_ent_salud_sector into  v_id_e_s, v_id_sector;

   while found loop -- Este found tiene true o false dependiendo si el fetch esta en el eof del archivo;
      ....
  ------------
  --Inicio de Carga especialidad
  ------------
        open cur_especialidad;
       ....

        query := .....

      insert into th_especialidad....
                          values ....
       -----------------
          fetch cur_especialidad  into v_id_espec, v_eps_h8, v_eps_h6, v_eps_h4 ,v_eps_h0 ;
       end loop;
  ------------
  --Fin de Carga especialidad
  ------------
   -- El siguiente registro del cursor sector: cur_ent_salud_sector
     fetch cur_ent_salud_sector into  v_id_e_s, v_id_sector;
   end loop;

if not found then
      raise notice 'Todas los sectores estan incluidos (%)',found;
        ....
   end if;
------------------
--Inicio carga tipo entidad salud --e_gestiona
```

Figura 35: Manejo de cursores y consultas dinámicas para almacenamiento de datos.
Elaborado por: Jorge Quiroz y Javier Reyes

La ejecución de procedimientos almacenados, consume gran cantidad de recursos de hardware, este procedimiento, por ejemplo, ha tardado 95,37 horas (343342932 ms) en los tres primeros años, debido a la cantidad de datos que se almacenan en memoria caché para posteriormente ingresar como atributos en los diferentes cursores, realizar comparaciones y consultas de tipo SQL, para finalmente llevar a cabo la inserción de los datos deseados, en cada caso.

Tiempo de carga para la tabla de hechos especialidad

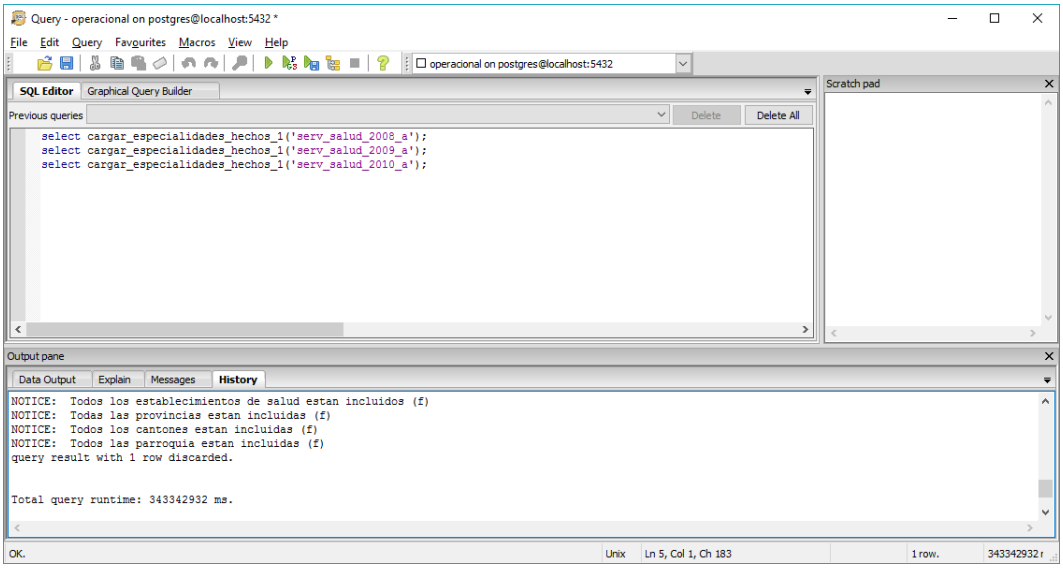


Figura 36: Tiempo de carga tabla de hechos especialidad
Elaborado por: Jorge Quiroz y Javier Reyes

4.3 Diccionario de datos

4.3.1 Base Operacional

Tabla 3: Tabla provincia

Nombre del Campo	Descripción/ Comentario	Tipo de dato	Obligatorio S/N	Restricción /Constraint	Valor por defecto
cod_provincia	Código de provincia	character varying(2)	S	Primary key	
provincia	Nombre de la provincia	character varying	N		
CONSTRAINTS					
Nombre del constraint	Tipo de constraint	Definición			
provincias_pk	Primary key	(cod_provincia)			
provincia_uk	Unique key	(provincia)			

Nota: La presente tabla, contiene la información de la descripción y código de las provincias del país.

Tabla 4: Tabla cantón

Nombre del Campo	Descripción/ Comentario	Tipo de dato	Obligatorio S/N	Restricción /Constraint	Valor por defecto
cod_provincia	Código de provincia	character varying(4)	S	Primary key	
cod_canton	Código de cantón	character varying(4)	S	Primary key	
canton	Nombre del cantón	character varying	N		
CONSTRAINTS					
Nombre del constraint	Tipo de constraint	Definición			
canton_pk	Primary key	(cod_provincia, cod_canton)			
canton_prov_fk	Foreign key	(cod_provincia) MATCH SIMPLE ON UPDATE NO ACTION ON DELETE NO ACTION			
canton_uk	Unique key	(canton)			

Nota: La presente tabla, contiene la información de la descripción y código de los cantones, además del código de la provincia a la que pertenece cada uno de los mismos.

Tabla 5: Tabla parroquia

Nombre del Campo	Descripción/ Comentario	Tipo de dato	Obligatorio S/N	Restricción /Constraint	Valor por defecto
cod_provincia	Código de provincia	character varying(4)	S		
cod_canton	Código de cantón	character varying(4)	S		
cod_inec	Código INEC	character varying(6)	S		
parroquia	Nombre de la parroquia	character varying	N		
id	Identificador de parroquia	serial	S	Primary key	
cod_parroquia	Código de parroquia	character varying(2)	N		
CONSTRAINTS					
Nombre del constraint	Tipo de constraint	Definición			
pk_parroquia	Primary key	(id)			
uk_parroquia	Unique key	(cod_provincia, cod_canton, cod_parroquia)			
uk_parroquia_codinec	Unique key	(cod_inec)			

uk_parroquia_pcp	Unique key	(cod_provincia, cod_canton, parroquia)

Nota: La presente tabla, contiene la información de la descripción y código de las parroquias, además del código de la provincia y cantón a la que pertenece cada uno de los mismos. Por otra parte, el código INEC y un id para facilitar la carga de la dimensión geografía.

Tabla 6: Tabla sector

Nombre del Campo	Descripción/ Comentario	Tipo de dato	Obligatorio S/N	Restricción /Constraint	Valor por defecto
codigo	Código del sector	character (2)	S	Primary key	
descripcion	Nombre del sector	character varying	N		
CONSTRAINTS					
Nombre del constraint	Tipo de constraint	Definición			
pk_sector	Primary key	(codigo)			

Nota: La presente tabla, contiene la información de los códigos y descripciones de todos los sectores de salud que se encuentran en las tablas salud de cada año respectivo.

Tabla 7: Tabla clase_establecimiento_salud

Nombre del Campo	Descripción / Comentario	Tipo de dato	Obligatorio S/N	Restricción /Constraint	Valor por defecto
codigo	Código de establecimiento	integer	S	Primary key	
descripcion	Nombre de establecimiento	character varying	N		
CONSTRAINTS					
Nombre del constraint	Tipo de constraint	Definición			
pk_clase_establecimiento_salud	Primary key	(codigo)			

Nota: La presente tabla, contiene la información de los códigos y descripciones de todas las clases de establecimientos de salud que existen y que se encuentran en las tablas salud de cada año respectivamente.

Tabla 8: Tabla entidad_gestiona_estab_sal

Nombre del Campo	Descripción/ Comentario	Tipo de dato	Obligatorio S/N	Restricción /Constraint	Valor por defecto
id	Código de establecimiento	numeric	S	Primary key	
descripcion	Nombre de establecimiento	text	N		
CONSTRAINTS					
Nombre del constraint	Tipo de constraint	Definición			
id_entidad_establecimiento	Primary key	(id)			

Nota: La presente tabla, contiene la información de las descripciones con su identificador para todos los establecimientos que existen a nivel nacional y se encuentran en las tablas salud de cada año respectivamente.

Tabla 9: Tabla tipo_entidad_salud

Nombre del Campo	Descripción/ Comentario	Tipo de dato	Obligatorio S/N	Restricción /Constraint	Valor por defecto
codigo	Código del Tipo Entidad Salud	character (2)	S	Primary key	
descripcion	Nombre del tipo de entidad	character varying	N		
CONSTRAINTS					
Nombre del constraint	Tipo de constraint	Definición			
pk_tipo	Primary key	(codigo)			

Nota: La presente tabla, contiene la información de las descripciones con su identificador para todos los tipos de entidades de salud a nivel nacional y se encuentran en las tablas salud de cada año respectivamente.

Tabla 10: Tabla serv_salud_año

Nombre del Campo	Descripción/ Comentario	Tipo de dato	Obligatorio S/N	Restricción /Constraint	Valor por defecto
codenc	Código de la encuesta	numeric			
descripcion	Año de la encuesta	text	N		
estable	Establecimiento de investigación	integer	N		
prov	Provincia de investigación	Carácter (2)	N		
cant	Cantón de investigación	Carácter (4)	N		
parr	Parroquia de investigación	Carácter (6)	N		
clase	Clase	integer	N		
tipo	Tipo	integer	N		
enti	Entidad de investigación	integer	N		
sect	Sector de investigación	integer	N		
farm	Farmacia	integer	N		
botq		integer	N		
lcli	Laboratorio clínico	integer	N		
lhis	Laboratorio histopatológico	integer	N		
lotro	Otros laboratorios	integer	N		
K1	Indicador de servicios	integer	N		
:	:	:	:		
:	:	:	:		
K1492	Indicador de servicios	integer	N		
id_serv_salud	Id de tabla	serial	S	Primary key	
CONSTRAINTS					
Nombre del constraint	Tipo de constraint	Definición			
pk_serv_salud_2014	Primary key	(id_serv_salud)			

Nota: Se sigue la definición de esta tabla para todos los años considerados, debido a su extensión, no se han considerado todos los indicadores (K1, K2, …K1400, Kn), está conformada por descripción, ubicación geográfica correspondiente (cantón, parroquia), el tipo de entidad y demás datos informativos que se relacionan con las otras tablas para obtener información.

4.3.2 Almacén de datos

La base *almacen* corresponde a la capa de información y se la realiza una vez concluida la construcción de la base operacional utilizando todos los datos contenidos en la misma, para su desarrollo se emplean procedimientos almacenados, que se encargan de cargar la información, esto se realiza mediante un apuntador hacia la base operacional que permite trabajar con estos datos.

Tabla 11: Tabla d_geografia

Nombre del Campo	Descripción/ Comentario	Tipo de dato	Obligatorio S/N	Restricción /Constraint	Valor por defecto
id	Identificador geográfico	Serial	S	Primary key	
provincia	Nombre de la provincia	character varying	N		
cantón	Nombre del cantón	character varying	N		
parroquia	Nombre de la parroquia	character varying	N		
cod_inec	Código INEC	character varying(6)	N		
CONSTRAINTS					
Nombre del constraint	Tipo de constraint	Definición			
pk_d_geografia	Primary key	(id)			
uk_ d_geografia	Unique key	(cod_inec)			

Nota: La presente tabla, contiene la información geográfica, clasificada por provincia, cantón y parroquia con su respectivo código (cod_inec)

Tabla 12: Tabla d_entidad_salud

Nombre del Campo	Descripción/ Comentario	Tipo de dato	Obligatorio S/N	Restricción Constraint	Valor por defecto
id	Identificador serial de tabla	Serial	S	Primary key	
sector	Tipo sector	character varying	N		
ent_gest_salud	Entidad que gestiona	character varying	N		
clase	Clase de entidad	character varying	N		
entidad_salud	Nombre de la entidad de salud	character varying	N		
id_sector	Identificador de tipo sector	character(2)	N		
id_egs	Identificador de entidad que gestiona	character(2)	N		
id_clase	Identificador de clase de entidad	integer	N		
id_es	Identificador de la entidad de salud	numeric	N		
cod_ubic_inec	Código de ubicación INEC	character varying (6)	N		
CONSTRAINTS					
Nombre del constraint	Tipo de constraint	Definición			
pk_d_entidad_salud	Primary key	(id)			
uk_ d_entidad_salud_des	Unique key	(sector, tipo_ent_salud, clase, entidad_salud, cod_ubic_inec)			

Nota: La presente tabla, refleja el cruce y carga realizada por el procedimiento cargar_entidad_salud() de las tablas sector, clase_establecimiento_salud y entidad_gestiona_estab_sal con sus respectivas descripciones y códigos pertenecientes a cada uno. Además, del código de ubicación geográfica del INEC, para poder identificar el lugar al que pertenece cualquier entidad de salud.

Tabla 13: Tabla d_tiempo

Nombre del Campo	Descripción/ Comentario	Tipo de dato	Obligatorio S/N	Restricción /Constraint	Valor por defecto
id	Código de año	serial	S	Primary key	
anio	Año correspondiente	numeric	N		
CONSTRAINTS					
Nombre del constraint	Tipo de constraint	Definición			
pk_d_tiempo	Primary key	(id)			

Nota: La presente tabla, contiene la dimensión tiempo, determinada por id y año correspondiente.

Tabla 14: Tabla d_especialidad

Nombre del Campo	Descripción/ Comentario	Tipo de dato	Obligatorio S/N	Restricción /Constraint	Valor por defecto
id	Identificador serial de tabla	Serial	S	Primary key	
descripción	Descripción indicador	character varying	N		
ki	Indicador Kn	character varying	N		
k_8	Indicador de especialista por 8 horas	Character varying	N		
k_6	Indicador de especialista por 6 horas	Character varying	N		
k_4	Indicador de especialista por 4 horas	Character varying	N		
k_0	Indicador de especialista por 0 horas	Character varying	N		
CONSTRAINTS					
Nombre del constraint	Tipo de constraint	Definición			
pk_d_especialidad	Primary key	(id)			
fk_kp_ki_especialidad	Foreign key	(kp)			

uk_ d_entidad_salud_des	Unique key	(ki)

Nota: La presente tabla, refleja la información de los indicadores que se encuentran en cada una de las tablas de salud de cada año de la base operacional. Estos indicadores reflejan la composición o relación que tienen, ya que unos son considerados padres porque contienen los códigos totales de los indicadores hijos. Además, se encuentran agregados los campos correspondientes al número de horas, que por lo general se encuentran clasificadas de 8, 6, 4 y 0 horas respectivamente.

Tabla 15: Tabla th_entidades_salud (Tabla de hechos)

Nombre del Campo	Descripción/ Comentario	Tipo de dato	Obligatorio S/N	Restricción /Constraint	Valor por defecto
id_dt	Identificador de la dimensión tiempo	Integer	S	Primary key	
id_dg	Identificador de la dimensión geografía	Integer	S	Primary key	
id_d_e_sal	Identificador de la dimensión entidad de salud	Integer	S	Primary key	
num_entidades	Número (total) de entidades de salud	Integer	N		
CONSTRAINTS					
Nombre del constraint	Tipo de constraint	Definición			
th_ent_sal_pk	Primary key	(id_dt, id_dg, id_d_e_sal)			
th_ent_sal_d_e_sal_fk	Foreign key	(id_d_e_sal)			
th_ent_sal_dg_fk	Foreign key	(id_dg)			
th_ent_sal_dt_fk	Foreign key	(id_dt)			

Nota: La presente tabla, refleja el cruce y carga de información de los identificadores de las tablas dimensionales d_tiempo, d_geografia, y d_entidad_salud, además del cálculo del número de entidades, esto a través del procedimiento carga_entidad_salud_hechos()

Tabla 16: Tabla th_especialidad (Tabla de hechos)

Nombre del Campo	Descripción/ Comentario	Tipo de dato	Obligatorio S/N	Restricción /Constraint	Valor por defecto
id_dt	Identificador de la dimensión tiempo	Integer	S	Primary key	
id_dg	Identificador de la dimensión geografía	Integer	S	Primary key	
id_d_e_sal	Identificador de la dimensión entidad de salud	Integer	S	Primary key	
id_de	Identificador de especialidad	Integer	S	Primary key	
num_eps_h8	Número de especialistas de 8 horas	Integer	N		
num_eps_h6	Número de especialistas de 6 horas	Integer	N		
num_eps_h4	Número de especialistas de 4 horas	Integer	N		
num_eps_h0	Número de especialistas de 0 horas	Integer	N		
CONSTRAINTS					
Nombre del constraint	Tipo de constraint	Definición			
th_esp_pk	Primary key	(id_dt, id_dg, id_d_e_sal, id_de)			
th_ent_sal_d_e_sal_fk	Foreign key	(id_d_e_sal)			
th_esp_de_fk	Foreign key	(id_de)			
th_esp_dg_fk	Foreign key	(id_dg)			
th_esp_dt_fk	Foreign key	(id_dt)			

Nota: La presente tabla, refleja el cruce y carga de información de los identificadores de las tablas dimensionales d_tiempo, d_geografia, y d_entidad_salud y d_especialidad, además del cálculo del número de especialistas, esto a través del procedimiento carga_especialidad_hechos(), a nivel nacional, por provincia, cantón y parroquia.

4.3.3 Vistas materializadas

Tabla 17: Vista vm_th_entidades_salud_prov_cant

Nombre del Campo	Descripción/ Comentario	Tipo de dato	Obligatorio S/N	Restricción /Constraint	Valor por defecto
anio	Año	Numeric(4,0)	N		
ubicación_geografica	Ubicación geográfica	Text	N		
sector	Tipo de sector	Character varying	N		
tipo_ent_salud	Tipo entidad de salud	Character varying	N		
clase	Clase de establecimiento	Character varying	N		
entidad_salud	Nombre entidad de salud	Character varying	N		
tot_num_ent	Número total de entidades de salud	Integer	N		

Nota: La presente vista materializada, refleja la misma información de la tabla de hechos entidades de salud, pero añadiendo las descripciones de cada uno de los códigos que existe en la misma, con el fin de poder entender y mostrar dicha información.

Tabla 18: Vista vm_th_especialidad_cant / Vista vm_th_especialidad_prov

Nombre del Campo	Descripción/ Comentario	Tipo de dato	Obligatorio S/N	Restricción /Constraint	Valor por defecto
anio	Año	Numeric(4,0)	N		
ubicación_geografica	Ubicación geográfica	Text	N		
sector	Tipo de sector	Character varying	N		
tipo_ent_salud	Tipo entidad de salud	Character varying	N		
clase	Clase de establecimiento	Character varying	N		
entidad_salud	Nombre entidad de salud	Character varying	N		
num_eps_h8	Número total de especialistas de 8 horas	Integer	N		

num_eps_h6	Número total de especialistas de 6 horas	Integer	N		
num_eps_h4	Número total de especialistas de 4 horas	Integer	N		
num_eps_h0	Número total de especialistas de 0 horas	Integer	N		
tot_esp	Número total de especialistas	Integer	N		

Nota: La presente vista materializada, refleja la misma información de la tabla de hechos especialidad, pero añadiendo las descripciones de cada uno de los códigos que existe en la misma, con el fin de poder entender y mostrar dicha información.

Tabla 19: Vista vm_th_especialidad_sect_provcant

Nombre del Campo	Descripción/ Comentario	Tipo de dato	Obligatorio S/N	Restricción /Constraint	Valor por defecto
anio	Año	Numeric (4,0)	N		
ubicación_geografica	Ubicación geográfica	Text	N		
sector	Tipo de sector	Character varying	N		
num_eps_h8	Número total de especialistas de 8 horas	Integer	N		
num_eps_h6	Número total de especialistas de 6 horas	Integer	N		
num_eps_h4	Número total de especialistas de 4 horas	Integer	N		
num_eps_h0	Número total de especialistas de 0 horas	Integer	N		
tot_esp	Número total de especialistas	Integer	N		

Nota: La presente vista materializada, refleja la misma información de la tabla de hechos especialidad, considerando solamente la clasificación por sector (público o privado), sin el nivel de detalle que la vista anterior (sector, tipo, clase), pero añadiendo las descripciones de cada uno de los códigos que existe en la misma.

5. Análisis

5.1 Vistas materializadas

Se obtuvieron vistas materializadas de las tablas de hechos por motivo de restricciones físicas en el equipo de trabajo (computador de uso personal).

5.1.1 Vista materializada entidades de salud.

Para esta vista, se han cargado los datos correspondientes, desde el año 2008 hasta el 2013, incluyendo como ubicación geográfica a todo el Ecuador, adicional cinco provincias principales y diez cantones de las mismas. Debido a las inconsistencias encontradas como la falta del identificador de establecimiento(estable) en el año 2014 y la discrepancia en los indicadores Kn en el año 2007, estos años no son considerados para el análisis.

Vista materializada de entidades de salud de la tabla de hechos.

	anio numeric(4,0)	ubicacion_ge text	sector character varying	tipo_ent_salud character varying	clase character varying	entidad_salud character varying	tot_num_ent integer
1	2008	Todas	Sector Publico				3103
2	2008	Todas	Sector Privado con fines de lucro				542
3	2008	Todas	Sector Privado sin fines de lucro				168
4	2008	Todas	Sector Publico	Ministerio de Salud Pública			1799
5	2008	Todas	Sector Publico	Ministerio de Gobierno y Policía			34
6	2008	Todas	Sector Publico	Ministerio de Defensa Nacional			68
7	2008	Todas	Sector Publico	Ministerio de Educación			95
8	2008	Todas	Sector Publico	Otros Ministerios			20
9	2008	Todas	Sector Publico	Instituto Ecuatoriano de Seguridad Social			75
10	2008	Todas	Sector Publico	Anexos al Seguro Social			277
11	2008	Todas	Sector Publico	Seguro Social Campesino			592
12	2008	Todas	Sector Publico	Otros Públicos			20
13	2008	Todas	Sector Publico	Consejos Provinciales			6
14	2008	Todas	Sector Publico	Municipios			74
15	2008	Todas	Sector Publico	Universidades y Politécnicas			22
16	2008	Todas	Sector Publico	Junta de Beneficiencia de Guayaquil			6
17	2008	Todas	Sector Publico	Cruz Roja Ecuatoriana			3
18	2008	Todas	Sector Publico	Sociedad de Lucha contra el Cáncer			12
19	2008	Todas	Sector Publico	Fisco Misionales			4
20	2008	Todas	Sector Privado con fines de lucro	Universidades y Politécnicas			22
21	2008	Todas	Sector Privado con fines de lucro	Privados con Fines de Lucro			539
22	2008	Todas	Sector Privado sin fines de lucro	Universidades y Politécnicas			22
23	2008	Todas	Sector Privado sin fines de lucro	Privados Sin Fines de Lucro			167
24	2008	Todas	Sector Publico	Ministerio de Salud Pública	Hospital Basico		90

Figura 37: Vista materializada entidades de salud
Elaborado por: Jorge Quiroz y Javier Reyes

5.1.2 Vista materializada especialidades de salud.

En esta vista, se incluyen de igual manera que la anterior vista, datos correspondientes desde el año 2008 al 2013, como ubicación geográfica a nivel nacional considerando las áreas más pobladas: seis provincias y doce cantones.

Vista materializada de especialidades de salud de la tabla de hechos.

	anio numeric(4,0)	ubicacion_ge text	especialidad character varying(150)	sector text	num_eps_h8 integer	num_eps_h6 integer	num_eps_h4 integer	num_eps_h0 integer	tot_e intege
1	2008	Todas	Total Medicos	Sector Publico	5651	1185	5726	939	13501
2	2008	Todas	Medicos Generales	Sector Publico	1541	168	1779	405	3893
3	2008	Todas	Cirujanos Generales	Sector Publico	96	45	336	32	509
4	2008	Todas	Cirujanos Plasticos	Sector Publico	8	12	47	3	70
5	2008	Todas	Medicina Interna (Internistas)	Sector Publico	33	26	151	2	212
6	2008	Todas	Anestesiologos	Sector Publico	98	96	310	50	554
7	2008	Todas	Cardiologos	Sector Publico	12	33	147	7	199
8	2008	Todas	Neurologos	Sector Publico	7	18	70	3	98
9	2008	Todas	Traumatologos	Sector Publico	30	45	196	28	299
10	2008	Todas	Psiquiatras	Sector Publico	12	13	101	1	127
11	2008	Todas	Oftalmologos	Sector Publico	9	24	109	13	155
12	2008	Todas	Otorrinolaringologos	Sector Publico	10	19	89	15	133
13	2008	Todas	Hematologos	Sector Publico	4	13	23	4	44
14	2008	Todas	Intensivistas	Sector Publico	6	37	70	2	115
15	2008	Todas	Nefrologos	Sector Publico	2	8	36	3	49
16	2008	Todas	Neumologos	Sector Publico	4	12	47	5	68
17	2008	Todas	Gastroenterologos	Sector Publico	7	19	84	13	123
18	2008	Todas	Geriatras	Sector Publico	0	3	9	2	14
19	2008	Todas	Oncologos	Sector Publico	5	24	31	2	62
20	2008	Todas	Urologos	Sector Publico	4	28	94	6	132

Figura 38: Vista materializada entidades de salud
Elaborado por: Jorge Quiroz y Javier Reyes

5.2 Pentaho

Pentaho es una suite que "proporciona un espectro completo de herramientas de inteligencia de negocio, reportes, análisis, dashboards, minería de datos e integración de datos. Ofrece una serie de servicios críticos entre los que están la autentificación, programación de tareas, seguridad y servicios Web. Este conjunto de herramientas y servicios forman una plataforma integral de inteligencia de negocio, convirtiendo a Pentaho en el proveedor líder de soluciones BI de código abierto." (Revista telemática, 2016)

Para llevar a cabo el análisis de datos, se utiliza la herramienta descrita anteriormente, Pentaho Report Designer, utilizando la información según los resultados del Censo de Población y Vivienda 2010, disponible en la página de Ecuador en Cifras (Censos, 2017).

Para el mismo, se consideran las seis provincias y los doce cantones más poblados, en las diferentes regiones del país, como puede observarse en las tablas 20 y 21, luego, se generan, vistas materializadas, que contengan la información necesaria que será visualizada en los respectivos reportes.

Tabla 20: Provincias con mayor población en el país

Región	Provincia	Población
Costa	Guayas	3.645.483
Costa	Manabí	1.369.780
Sierra	Pichincha	2.576.287
Sierra	Azuay	712.127
Oriente	Sucumbíos	176.472
Insular	Galápagos	25.124

Nota: Provincias con mayor número de habitantes en el país, según los datos oficiales del INEC

Tabla 21: Cantones con mayor población en el país

Región	Provincia	Cantón	Población
Costa	Guayas	Guayaquil	2.350.915
Sierra	Pichincha	Quito	2.239.191
Sierra	Azuay	Cuenca	505.585
Costa	Santo Domingo de los Tsáchilas	Santo Domingo	368.013
Sierra	Tungurahua	Ambato	329.856
Costa	Manabí	Portoviejo	280.029
Costa	El Oro	Machala	245.972
Costa	Guayas	Durán	235.769
Costa	Manabí	Manta	226.477
Sierra	Loja	Loja	214.855
Oriente	Sucumbíos	Shushufindi	44.328
Insular	Galápagos	Santa Cruz	15.393

Nota: De conformidad a las cifras del INEC se muestran las ciudades con mayor número de habitantes

Una vez establecida la conexión entre Pentaho Report Designer (PRD) y PostgreSQL, en este caso con la base de datos *almacen*, se generan consultas utilizando los elementos propios de la fuente de datos, en este caso, se emplean las vistas materializadas, ya que contienen la información procedente de las tablas de hechos y dimensiones de manera entendible, que permite visualizar la información.

Datasource de PRD

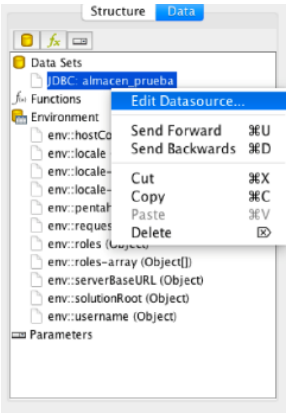


Figura 39: Configuración de Datasource (Fuente de datos)
Elaborado por: Jorge Quiroz y Javier Reyes

Para esto, se procede a editar la conexión, en donde se agregan las consultas respectivas de acuerdo a la necesidad de información que se quiere visualizar, como se puede ver en la figura 40, se puede obtener una vista preliminar de las mismas con *preview*.

Gestión de consultas SQL en PRD

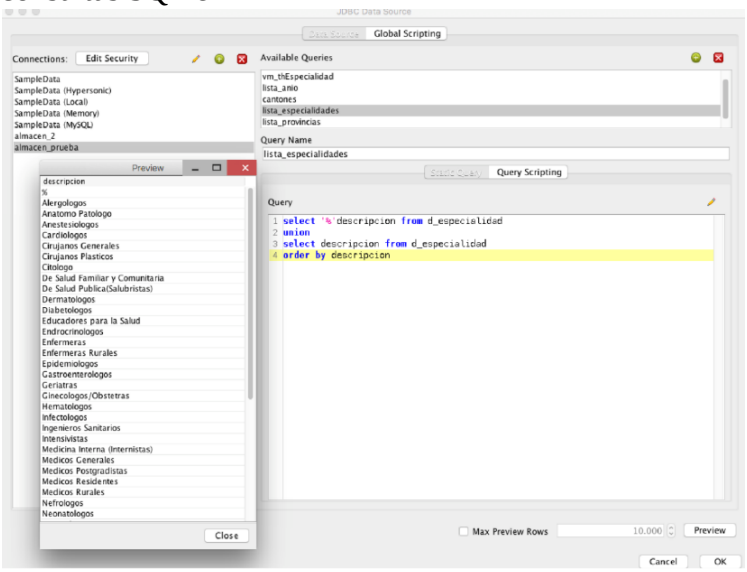


Figura 40: Creación y gestión de consultas en el Datasource almacen_prueba
Elaborado por: Jorge Quiroz y Javier Reyes

En el caso del reporte de especialidades por provincia, que es del cual se muestra el procedimiento de creación, se tienen consultas que permiten obtener información en forma de variables con las que se puede mostrar atributos o gráficos estadísticos en el reporte como se puede observar en la figura 41.

La primera consulta hace referencia a la vista de especialidades, con la que se obtiene el año, sector tipo de entidad y ubicación geográfica, lista de años, provincias y sector que permite obtener estos datos en un componente tipo combo box denominado *drop down*, en PRD.

Queries del reporte, PRD

```
-----------
--Query name: vm_thEspecialidad
select * from vm_th_especialidad_prov
where especialidad like ${especialidades}
AND ubicacion_geografica like ( ${ubicacion}::text || '%'::text )
AND anio = ${anio}
AND sector like ${sect}
AND tipo_ent_salud is not null
-----------
--Query name: lista_anio
select anio from d_tiempo where anio not in (2014)
-----------
--Query name: lista_especialidades
select '%'descripcion from d_especialidad
union
select descripcion from d_especialidad
order by descripcion
-----------
--Query name:lista_provincias
select provincia from d_geografia g
where g.cod_inec is null or g.cod_inec
         in ('01','09','13','17','20','21')
-----------
--Query name:lista_sector
select distinct sector from d_entidad_salud
```

Figura 41: Consultas (queries) empleados en el reporte de especialistas por provincia
Elaborado por: Jorge Quiroz y Javier Reyes

A continuación, se procede a crear parámetros que permitan seleccionar, de las listas respectivas, el cantón o parroquia, sector y año según sea el caso, para mostrar información estadística clasificada bajo estos filtros.

En la figura 42, se muestra la creación del parámetro *ubicación*, en donde se establece el texto que se muestra, el tipo de dato, en este caso *String*, valores por defecto, el elemento de selección de la GUI, en este caso, *DropDown* y la consulta que alimenta la lista de elementos a ser seleccionados, en este caso, corresponde a la consulta con nombre *lista_sector*.

Gestión de parámetros en PRD

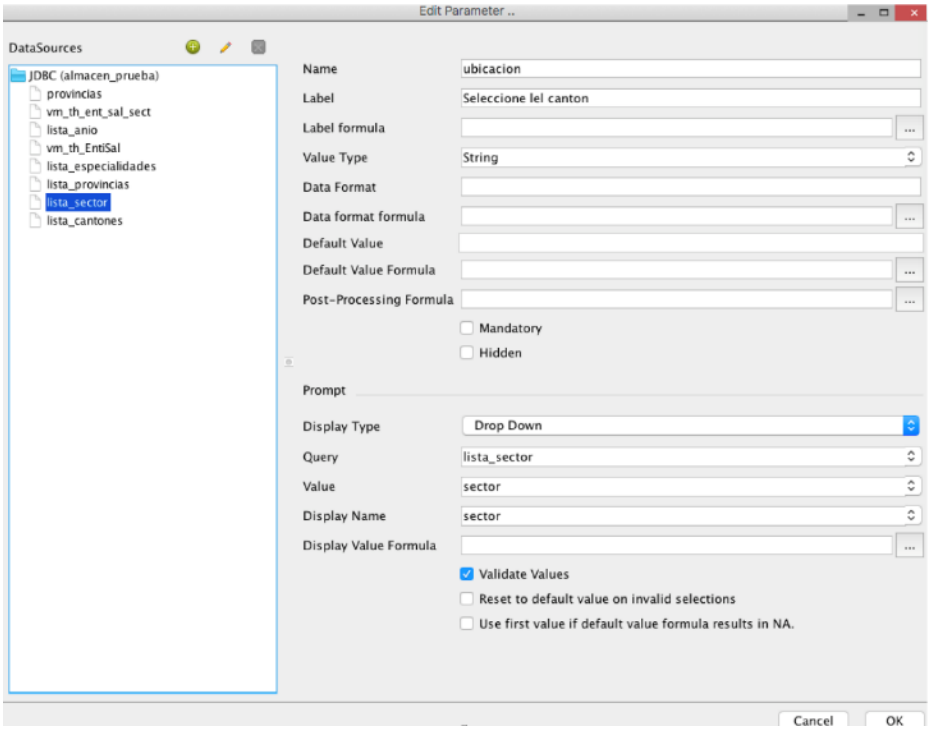


Figura 42: Configuración del parámetro ubicación, asignación de query y leyendas.
Elaborado por: Jorge Quiroz y Javier Reyes

Como se puede observar en la figura 43, PRD presenta una plantilla inicial personalizable que contiene varias secciones en las cuales se puede ubicar la información que se requiera desplegar. Estas secciones se pueden aumentar o disminuir dependiendo de los criterios de información que se quieran presentar.

Para mostrar los datos, resultado de las consultas realizadas, se arrastra cada variable disponible en la carpeta contenedora de la consulta (*query*), en este caso *vm_thEspecialidad*, a la sección correspondiente donde se visualizará el conjunto de datos resultantes.

En este caso, la sección *Details Header*, contiene la cabecera de los campos mostrados, mientras que la sección *Details*, muestra a detalle el resultado del *query*.

Personalización de plantilla PRD

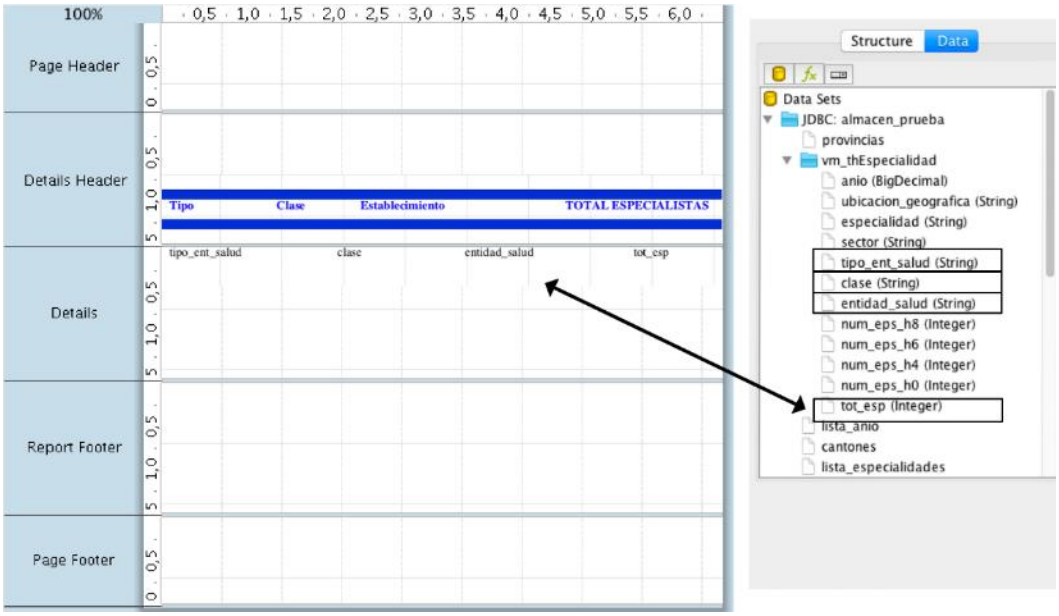


Figura 43: Configuración de secciones de PRD, para desplegar información y títulos respectivos provenientes de consultas sql.
Elaborado por: Jorge Quiroz y Javier Reyes

Para gráficos estadísticos, se emplea la opción *chart*, en la cual se selecciona el tipo de gráfico deseado, y se procede realizar las configuraciones necesarias, como se muestra en la figura 45; *char-title*, contiene el título de la gráfica, el apartado *CategorySet Data Collector*, permite establecer la categoría y valores por columnas y la organización se las series y campos, entre otros.

La siguiente imagen muestra la estructura final del reporte.

Plantilla personalizada PRD

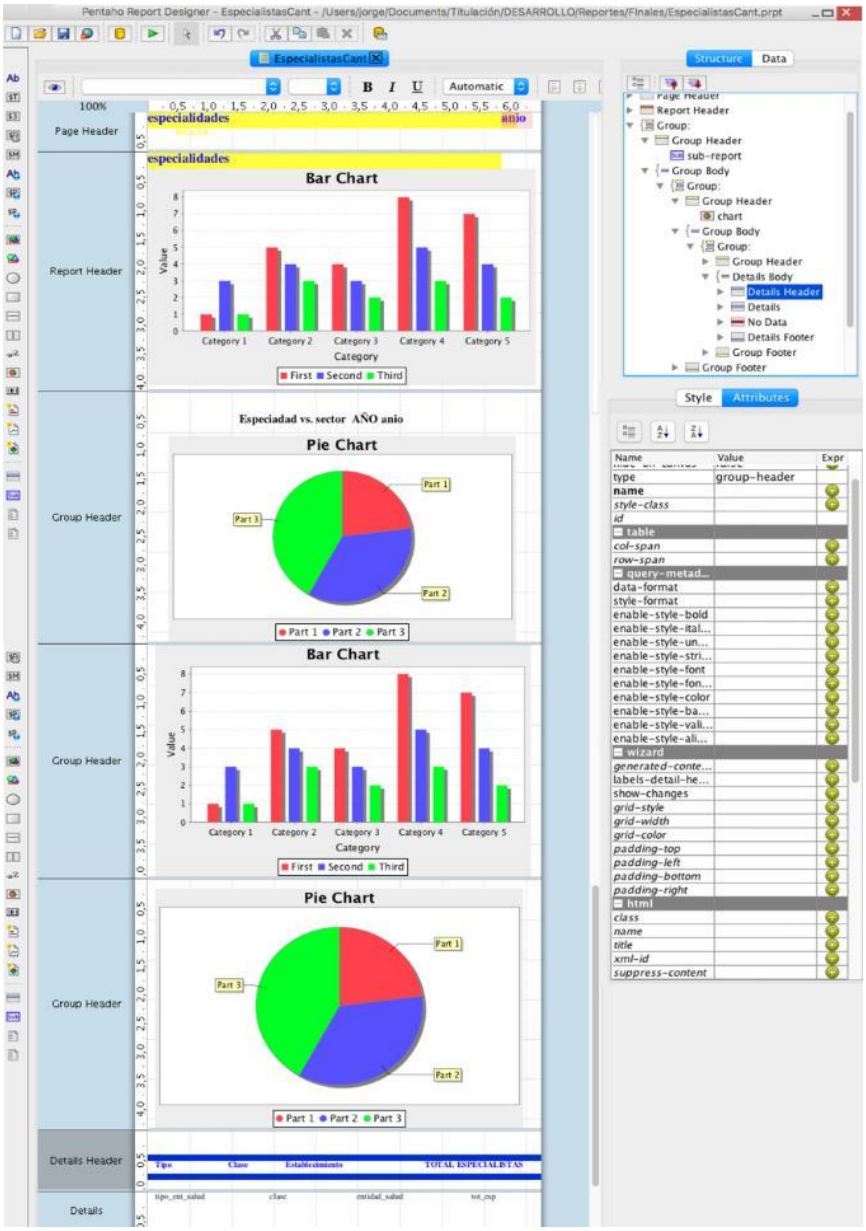


Figura 44: Visualización de la estructura del reporte, mediante la personalización de elementos y secciones establecidas para despliegue de datos y gráficas.
Elaborado por: Jorge Quiroz y Javier Reyes

Considerando todas estas indicaciones, se han llevado a cabo cuatro reportes. Dos correspondientes a cada tabla de hechos, uno por filtro parroquia y otro por cantón.

Gráficos estadísticos en PRD

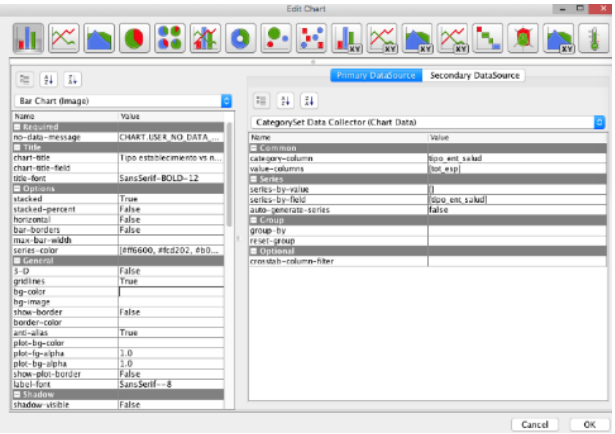


Figura 45: Configuración de elementos para generar gráficos estadísticos, utilizando campos provenientes de las consultas (query) establecidas.
Elaborado por: Jorge Quiroz y Javier Reyes

Entidades de salud

Reporte Entidades de salud por cantón 1 de 3

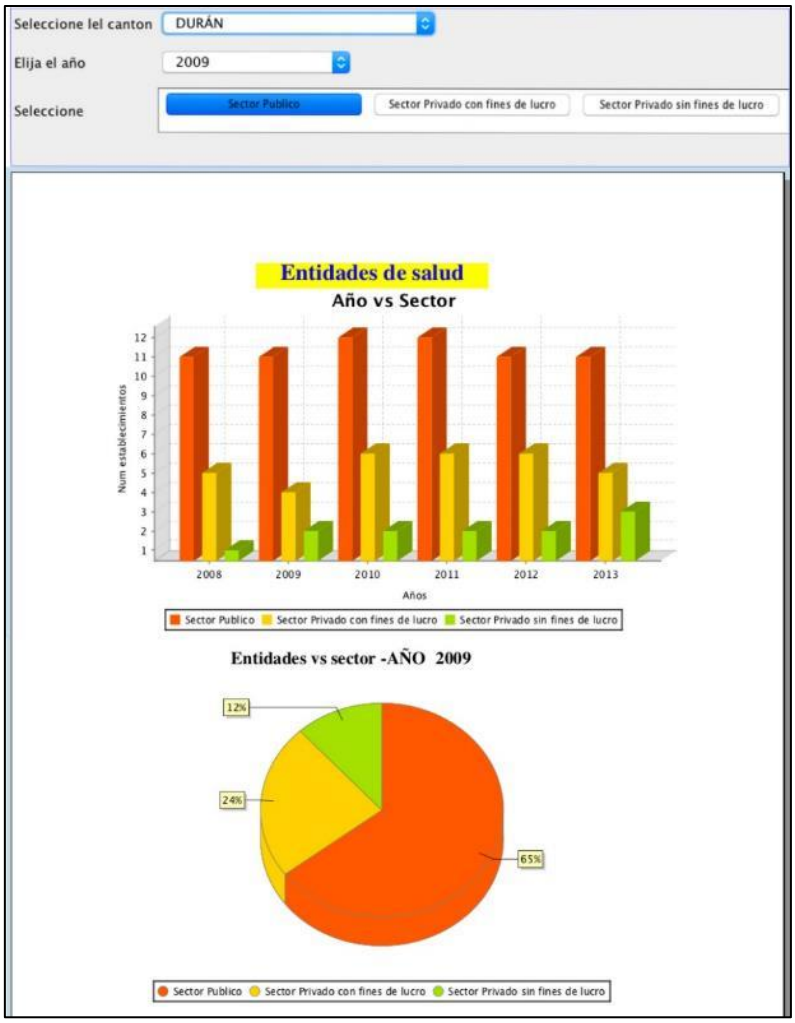


Figura 46: Interfaz de selección de filtros: ubicación, año y sector. Gráficos con relación a sector por todos los años y en porcentajes de acuerdo al año seleccionado.
Elaborado por: Jorge Quiroz y Javier Reyes

Reporte Entidades de salud por cantón 2 de 3

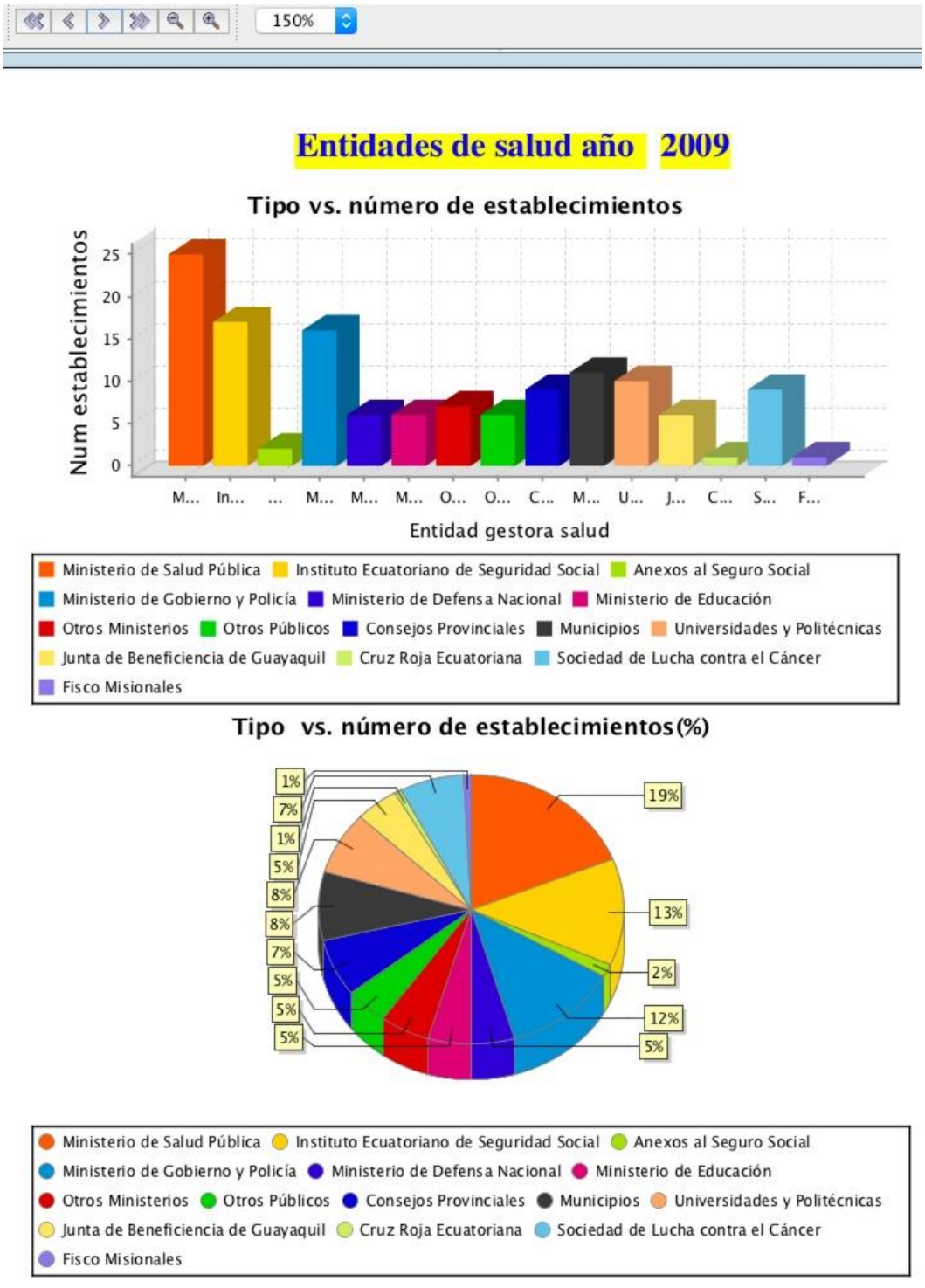
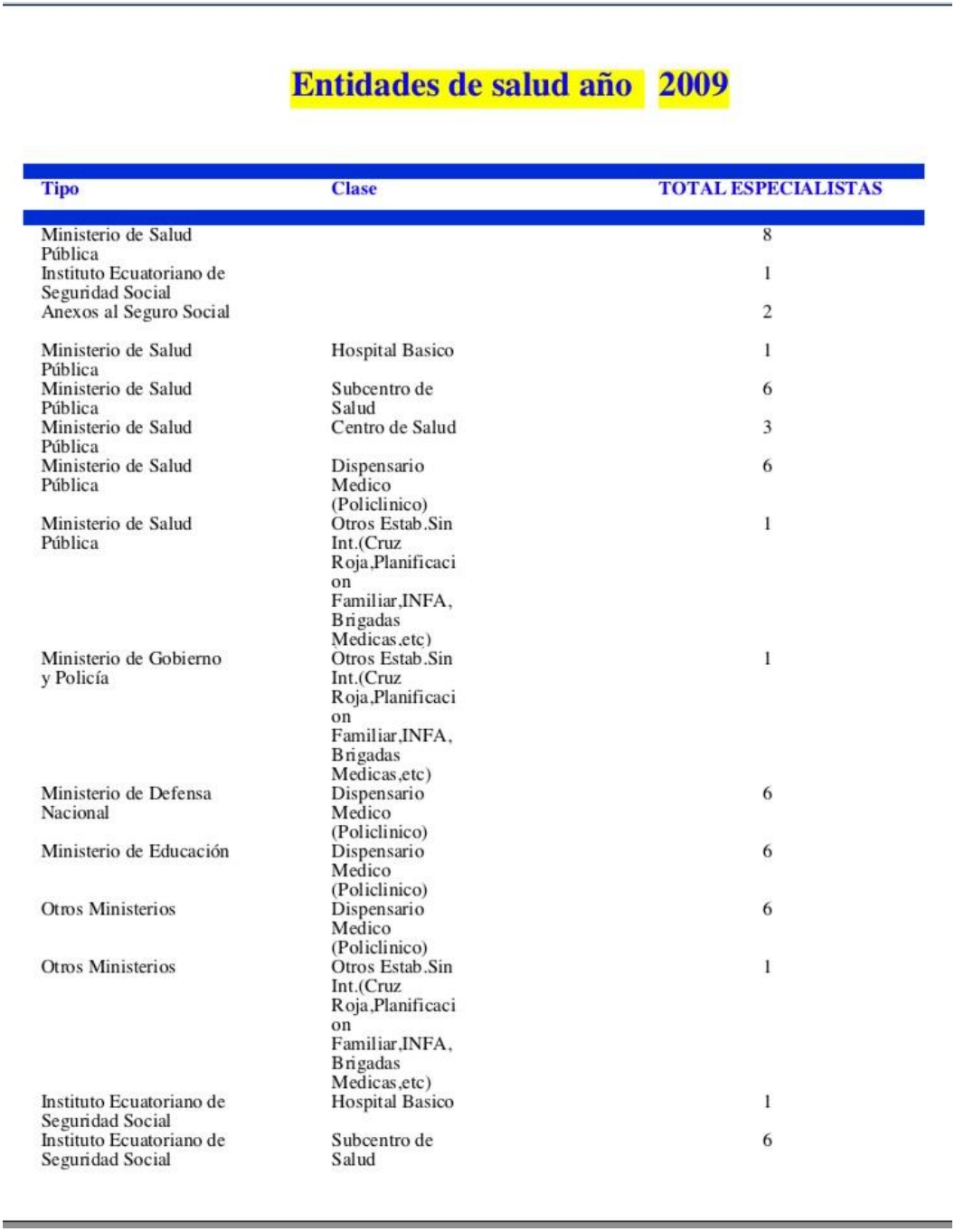


Figura 47: Gráficas de tipo de entidad de salud por número de establecimientos en cifras y porcentajes
Elaborado por: Jorge Quiroz y Javier Reyes

Reporte Entidades de salud por cantón 3 de 3



Entidades de salud año 2009

Tipo	Clase	TOTAL ESPECIALISTAS
Ministerio de Salud Pública		8
Instituto Ecuatoriano de Seguridad Social		1
Anexos al Seguro Social		2
Ministerio de Salud Pública	Hospital Basico	1
Ministerio de Salud Pública	Subcentro de Salud	6
Ministerio de Salud Pública	Centro de Salud	3
Ministerio de Salud Pública	Dispensario Medico (Policlinico)	6
Ministerio de Salud Pública	Otros Estab.Sin Int.(Cruz Roja,Planificacion Familiar,INFA, Brigadas Medicas,etc)	1
Ministerio de Gobierno y Policía	Otros Estab.Sin Int.(Cruz Roja,Planificacion Familiar,INFA, Brigadas Medicas,etc)	1
Ministerio de Defensa Nacional	Dispensario Medico (Policlinico)	6
Ministerio de Educación	Dispensario Medico (Policlinico)	6
Otros Ministerios	Dispensario Medico (Policlinico)	6
Otros Ministerios	Otros Estab.Sin Int.(Cruz Roja,Planificacion Familiar,INFA, Brigadas Medicas,etc)	1
Instituto Ecuatoriano de Seguridad Social	Hospital Basico	1
Instituto Ecuatoriano de Seguridad Social	Subcentro de Salud	6

Figura 48: Reporte en texto plano de acuerdo a variables consideradas y sus cifras respectivas
Elaborado por: Jorge Quiroz y Javier Reyes

Especialidades

El objetivo de este reporte es mostrar información relativa a las entidades de salud del país considerando provincias y cantones con mayor población.

Reporte Especialidades por provincia 1 de 3

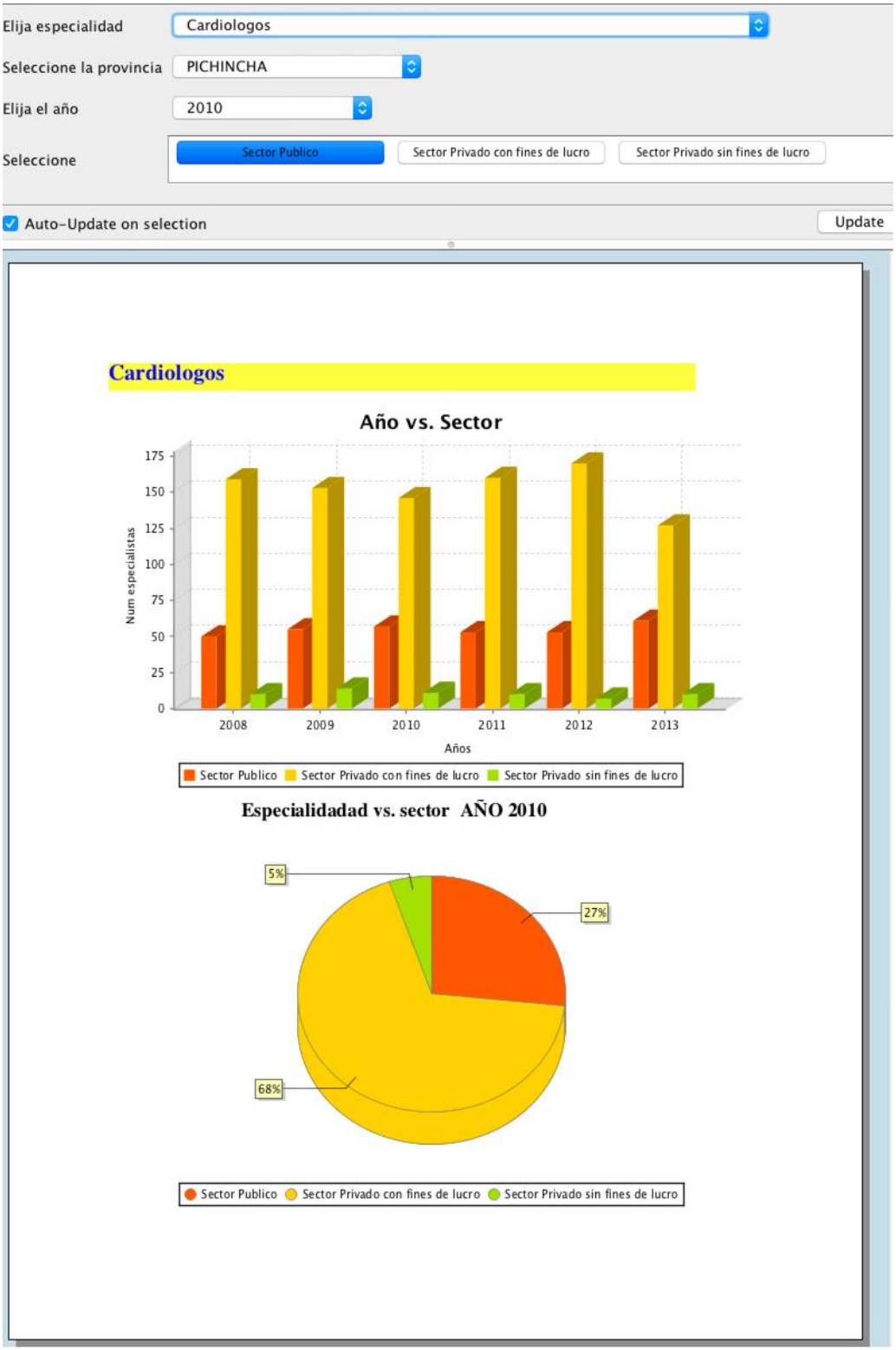


Figura 49: Interfaz de selección de filtros: ubicación, año y sector. Gráficos con relación a sector por todos los años y en porcentajes de acuerdo al año seleccionado.
Elaborado por: Jorge Quiroz y Javier Reyes

Reporte Especialidades por provincia 2 de 3

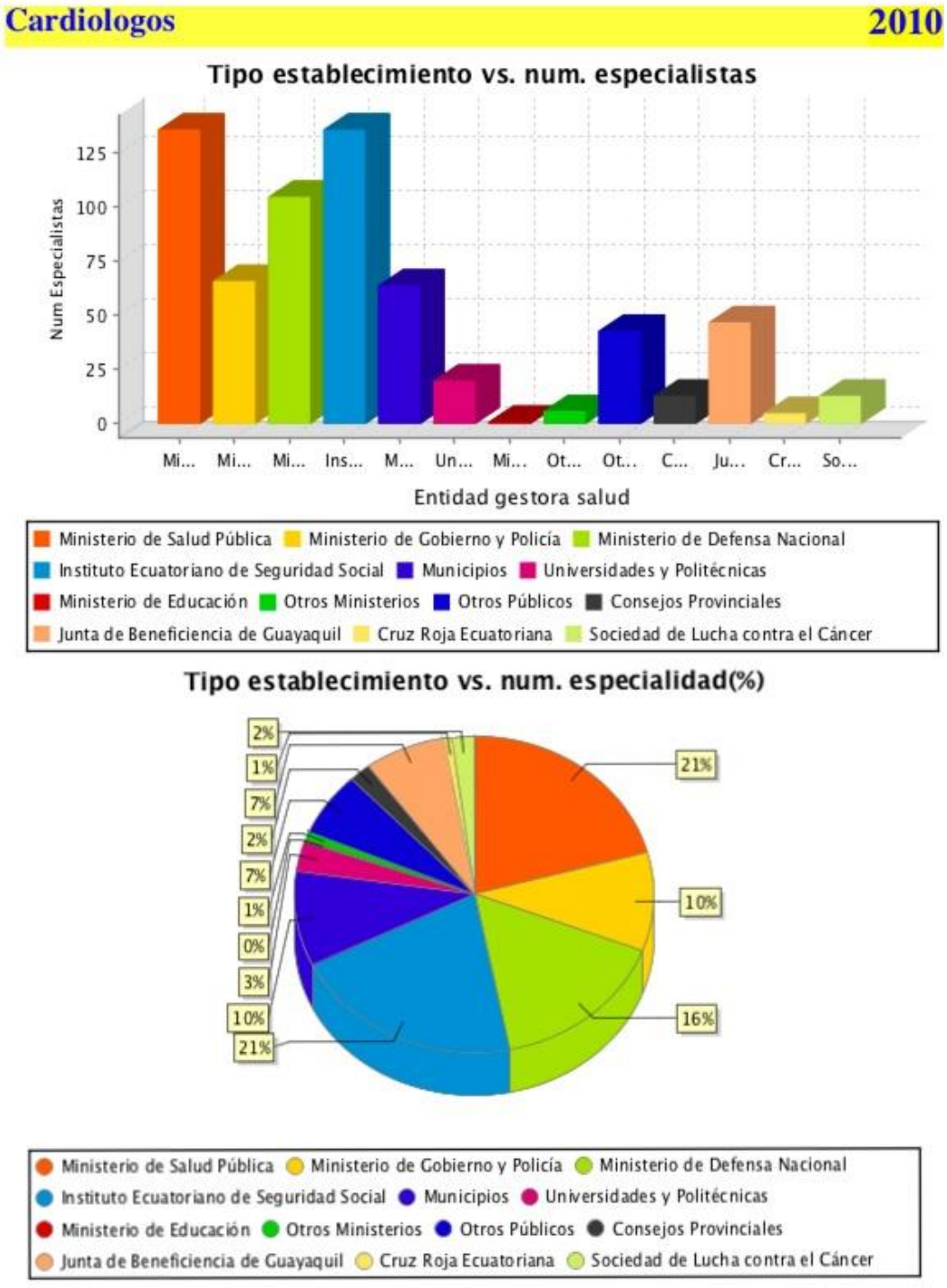


Figura 50: Gráficas de tipo de entidad de salud por número de especialistas en cifras y porcentajes
Elaborado por: Jorge Quiroz y Javier Reyes

Reporte Especialidades por provincia 3 de 3

Figura 51: Resultado de tasa de médicos y reporte en texto plano de acuerdo a variables consideradas
Elaborado por: Jorge Quiroz y Javier Reyes

5.3 Weka

Weka es un software de código abierto desarrollado en lenguaje de programación JAVA, contiene un conjunto de algoritmos que son implementados para tareas o actividades de minería de datos. Además, Weka contiene herramientas para el pre-procesamiento de datos, así como también de clasificación, regresión, agrupación y visualización de los mismos. (Waikato, 2017)

Weka Data Mining Task

Figura 52: Interfaz de presentación Weka Data Mining Task
Elaborado por: Jorge Quiroz y Javier Reyes

5.3.1 Conexión Weka con base de datos PostgreSQL.

Para la implementación de los algoritmos en minería de datos, es necesario configurar la conexión a la base de datos u otra opción es tener la información en un archivo con formato .csv, .mat, etc.

Conexión a Base de datos Postgres desde Weka

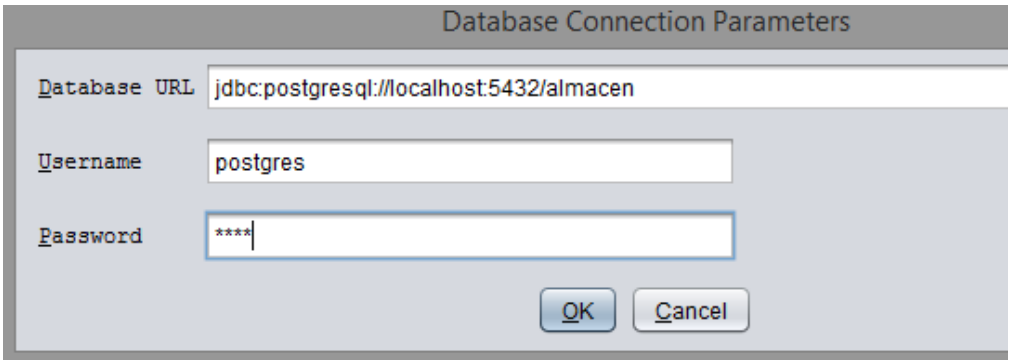


Figura 53: Conexión base de datos desde Weka
Elaborado por: Jorge Quiroz y Javier Reyes

Posteriormente, para la conexión a una base de datos en Weka, se insertan los datos, esto se lo puede observar en la siguiente imagen.

Verificación de conexión

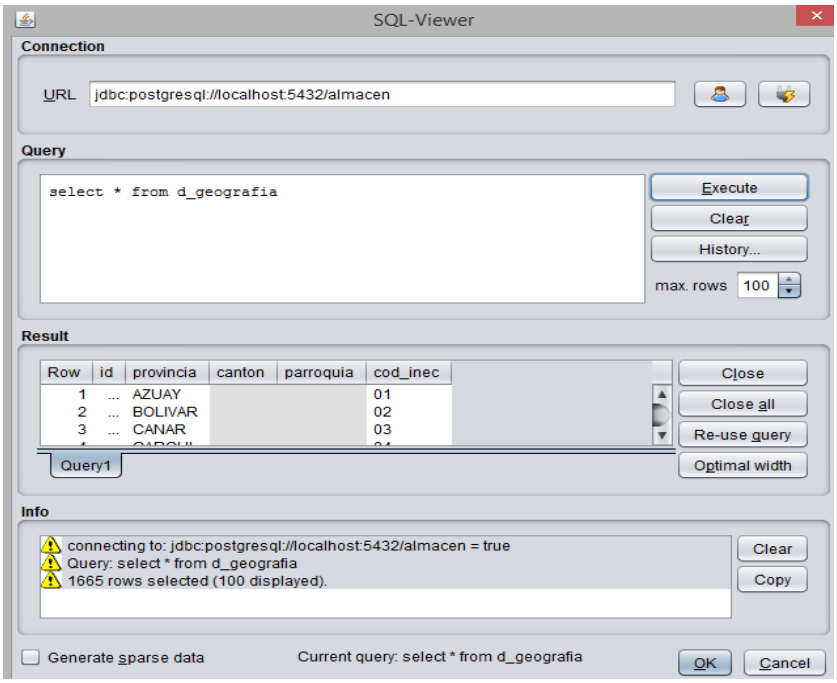


Figura 54: Verificación de conexión y consulta para obtener datos de la base de datos.
Elaborado por: Jorge Quiroz y Javier Reyes

Para el análisis de la información, se emplean las vistas materializadas creadas desde PostgreSQL sobre las tablas de hechos, tanto de especialidad como entidad de salud, para que de esta manera la carga de datos en Weka sea rápida.

5.3.3 Pre procesamiento de datos en Weka

Una vez, que se ha verificado la conexión con la base de datos PostgresSQL y que se han obtenido los datos de la vista materializada, Weka presenta todos los datos de la consulta establecida para el análisis, inclusive se puede visualizar una estadística de los datos extraídos como se puede constatar en la Figura 55.

Visualización de atributos y estadísticas de datos.

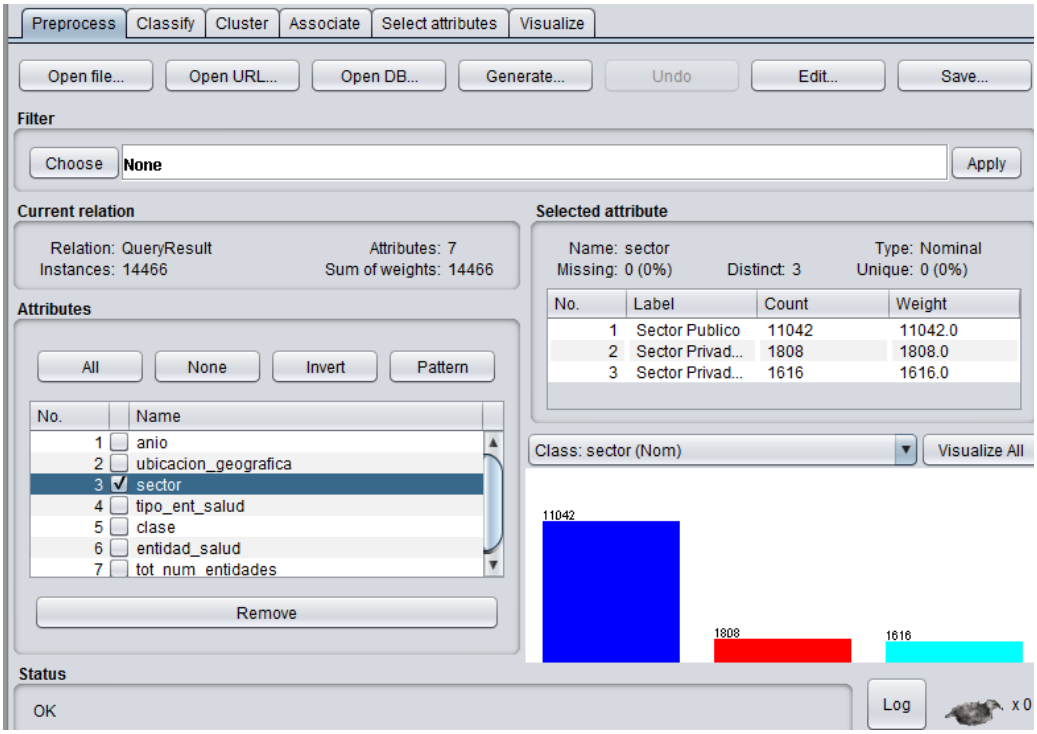


Figura 55 Visualización y pre procesado de datos
Elaborado por: Jorge Quiroz y Javier Reyes

5.3.4 Algoritmos de clasificación y regresión

Los algoritmos de clasificación son aquellos que permiten conocer a que clase pertenece un conjunto de datos, mientras que los algoritmos de regresión tienen como objetivo visualizar o predecir los valores futuros y tendencias de un mismo conjunto de datos. (Bouckaert R., 2015)

Weka ofrece en sus herramientas el uso de estos tipos de algoritmos para el análisis y criterio de la información.

Como punto de referencia para la aplicación de estos algoritmos, el atributo sector será la base para todas estas consultas y análisis.

Clasificador Weka para algoritmos de regresión, árboles y clasificación.

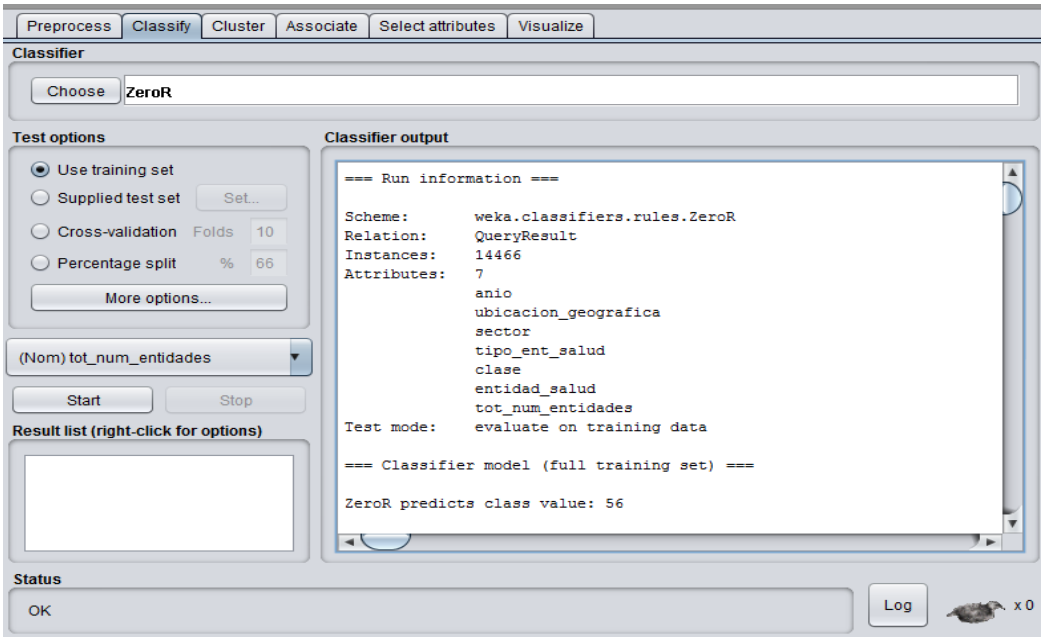


Figura 56: Herramienta clasificador WEKA
Elaborado por: Jorge Quiroz y Javier Reyes

5.3.4.1 Algoritmo ZeroR

El algoritmo ZeroR, tiene como objetivo predecir la clase mayoritaria de los datos, es usado como base para comparaciones con otros algoritmos. (Bouckaert R., 2015). Si se observa la siguiente imagen, el número mayoritario de entidades pertenecen al sector público. Por lo que, ZeroR clasificará a todos los demás sectores como sector público:

Número de datos por Sector de salud

Name: sector
Missing: 0 (0%)
Distinct: 3
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	Sector Publico	11042	11042.0
2	Sector Privado con fines de lucro	1808	1808.0
3	Sector Privado sin fines de lucro	1616	1616.0

Figura 57: Número de datos por sector de salud
Elaborado por: Jorge Quiroz y Javier Reyes

A continuación, se presentan los datos arrojados por el algoritmo ZeroR:

Algoritmo ZeroR – Resumen estadístico

```
=== Summary ===

Correctly Classified Instances       11042               76.3307 %
Incorrectly Classified Instances      3424               23.6693 %
Kappa statistic                          0
Mean absolute error                      0.2595
Root mean squared error                  0.3602
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances            14466
```

Figura 58: Resumen de algoritmo ZeroR
Elaborado por: Jorge Quiroz y Javier Reyes

A Continuación, se observa la siguiente información estadística de ZeroR:

- Número de datos clasificados correctamente: 76%, es decir 11042 registros.
- Número de datos clasificados incorrectamente: 24%, es decir un total de 3424 registros.
- Error absoluto medio: 0.2595
- Error cuadrático medio: 0.3602
- Error absoluto relativo: 100%
- Error cuadrático relativo: 100%
- Número de datos: 14466

Mientras tanto, la matriz de confusión arrojó los siguiente:

Algoritmo ZeroR - Matriz de confusión algoritmo

```
=== Confusion Matrix ===

     a     b     c   <-- classified as
 11042     0     0 |     a = Sector Publico
  1808     0     0 |     b = Sector Privado con fines de lucro
  1616     0     0 |     c = Sector Privado sin fines de lucro
```

Figura 59: Matriz confusión ZeroR
Elaborado por: Jorge Quiroz y Javier Reyes

Este algoritmo permitió conocer el sector con mayor provisión de entidades de salud, corresponde en este caso, al sector público, con los errores estadísticos respectivos y los datos reflejados en la matriz de confusión.

5.3.4.2 Algoritmo OneR

El algoritmo OneR, se encuentra basado por reglas con una única variable. Este algoritmo, genera todas las reglas del tipo "IF variable=valor THEN clase=categoría", además es otro algoritmo usado para comparaciones. (Bouckaert R., 2015)

Algoritmo OneR - Reglas de tipo

```
=== Classifier model (full training set) ===

tipo_ent_salud:
        Ministerio de Salud Pública     -> Sector Publico
        Ministerio de Gobierno y Policía        -> Sector Publico
        Ministerio de Defensa Nacional  -> Sector Publico
        Ministerio de Educación -> Sector Publico
        Otros Ministerios       -> Sector Publico
        Instituto Ecuatoriano de Seguridad Social       -> Sector Publico
        Anexos al Seguro Social -> Sector Publico
        Seguro Social Campesino -> Sector Publico
        Otros Públicos  -> Sector Publico
        Consejos Provinciales   -> Sector Publico
        Municipios      -> Sector Publico
        Universidades y Politécnicas    -> Sector Publico
        Junta de Beneficiencia de Guayaquil     -> Sector Publico
        Cruz Roja Ecuatoriana   -> Sector Publico
        Sociedad de Lucha contra el Cáncer      -> Sector Publico
        Fisco Misionales        -> Sector Publico
        Privados con Fines de Lucro     -> Sector Privado con fines de lucro
        Privados Sin Fines de Lucro     -> Sector Privado sin fines de lucro
        ?       -> Sector Publico
(13794/14466 instances correct)
```

Figura 60: Reglas de tipo Algoritmo OneR
Elaborado por: Jorge Quiroz y Javier Reyes

Una de las reglas, como se ve en la imagen anterior por ejemplo es:

Ministerio de Salud Pública → Sector Público, es decir OneR asoció en sus reglas a esta variable con la clase Sector Público.

Algoritmo OneR - Resumen estadístico

```
=== Summary ===

Correctly Classified Instances       13794               95.3546 %
Incorrectly Classified Instances       672                4.6454 %
Kappa statistic                          0.8707
Mean absolute error                      0.031
Root mean squared error                  0.176
Relative absolute error                 11.932  %
Root relative squared error             48.8545 %
Total Number of Instances            14466
```

Figura 61: Detalle Algoritmo OneR
Elaborado por: Jorge Quiroz y Javier Reyes

A Continuación, se observa la siguiente información estadística de OneR:

- Número de datos clasificados correctamente: 95%, es decir 13794 registros.
- Número de datos clasificados incorrectamente: 5%, es decir un total de 672 registros.
- Kappa statistic(Concordancia): 0.8707
- Error absoluto medio: 0.031
- Error cuadrático medio: 0.176
- Error absoluto relativo: 11.932%
- Error cuadrático relativo: 48.85%
- Número de datos: 14466

Mientras tanto, la matriz de confusión arrojó los siguiente:

Algoritmo ZeroR - Matriz de confusión algoritmo

```
=== Confusion Matrix ===

     a     b     c   <-- classified as
 11042     0     0 |     a = Sector Publico
   384  1424     0 |     b = Sector Privado con fines de lucro
   288     0  1328 |     c = Sector Privado sin fines de lucro
```

Figura 62: Matriz confusión ZeroR
Elaborado por: Jorge Quiroz y Javier Reyes

Este algoritmo hizo posible obtener las reglas de una única variable, en este caso referente a tipo de entidad de salud que nos permite visualizar su pertenencia al sector público o privado (con o sin fines de lucro).

5.3.4.3 Red Bayesiana

La Red Bayesiana es una representación gráfica que relaciona un conjunto de variables que ayuda a la toma de decisiones. El modelo resultante de una red bayesiana permite realizar una inferencia probabilística de manera eficiente. (Todd A., 2000)

Red Bayesiana – Detalles estadísticos

```
=== Summary ===

Correctly Classified Instances       13194               95.4427 %
Incorrectly Classified Instances       630                4.5573 %
Kappa statistic                          0.8723
Mean absolute error                      0.0394
Root mean squared error                  0.1454
Relative absolute error                 15.2931 %
Root relative squared error             40.5089 %
Total Number of Instances            13824
```

Figura 63: Detalles estadísticos Red Bayesiana
Elaborado por: Jorge Quiroz y Javier Reyes

En cuanto, a la matriz de confusión se dieron los siguientes resultados:

Red Bayesiana – Matriz de confusión de las clases de sectores

```
=== Confusion Matrix ===

     a     b     c   <-- classified as
 10584     0     0 |     a = Sector Publico
   360  1350     0 |     b = Sector Privado con fines de lucro
   270     0  1260 |     c = Sector Privado sin fines de lucro
```

Figura 64: Matriz de confusión Red Bayesiana
Elaborado por: Jorge Quiroz y Javier Reyes

Para la visualización de la red bayesiana, la selección de la lista de resultados presenta la siguiente gráfica:

Red Bayesiana – Visualización gráfico de las probabilidades

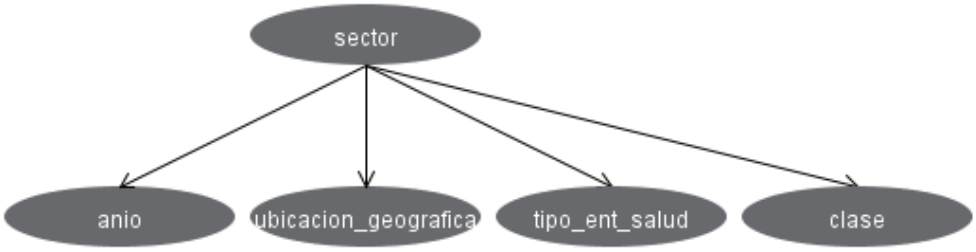


Figura 65: Visualización de gráfico Red bayesiana
Elaborado por: Jorge Quiroz y Javier Reyes

Red Bayesiana – Visualización gráfico

'Sector Publico'	'Sector Privado con fines de lucro'	'Sector Privado sin fines de lucro'
0,766	0,124	0,111

(a)

sector	2008	2009	2010	2011	2012	2013
'Sector Publico'	0,167	0,167	0,167	0,167	0,167	0,167
'Sector Privado con fines de lucro'	0,167	0,167	0,167	0,167	0,167	0,167
'Sector Privado sin fines de lucro'	0,167	0,167	0,167	0,167	0,167	0,167

(b)

Probability Distribution Table For tipo_ent_salud

sector	'Ministerio de Salud Pública'	'Ministerio de Gobierno y Policía'	...	'Ministerio de Educación'	...	'Instituto Ecuatoriano de Seguridad Social'	'Anexos al Seguro Social'
'Sector Publico'	0,214	0,071	...	0,041	...	0,082	0,031
'Sector Privado con fines de lucro'	0,053	0	...	0	...	0	0
'Sector Privado sin fines de lucro'	0,059	0	...	0	...	0	0

(c)

Probability Distribution Table For clase

sector	'Hospital Basico'	'Hospital General'	'Hospital de Infectologia (Agudo)'	...	...	'Hospital Psiquiatrico y Sanatorio ...	'Hospital Dermatologico (le...
'Sector Publico'	0,041	0,071	0,01	...	...	0,02	0,01
'Sector Privado con fines de lucro'	0	0,105	0	...	...	0,052	0
'Sector Privado sin fines de lucro'	0,059	0,117	0	...	...	0,059	0

(d)

Figura 66: Probabilidad de distribución para (a) sector, (b) año, (a) tipo entidad salud y (a) clase establecimiento.
Elaborado por: Jorge Quiroz y Javier Reyes

El presente algoritmo permite descubrir las relaciones existentes entre sector, tipo de entidad, clase y año, mostrando resultados, entre los que se puede referir que el Hospital de Infectología, tiene una presencia de 0,01 en el sector público y no está provisto en los demás sectores.

CONCLUSIONES

- Se logró cumplir con los objetivos propuestos en el presente proyecto, es decir: crear un almacén de datos para analizar la provisión de los servicios de salud en el país, utilizando la información disponible en el portal del INEC, analizar la información, utilizando criterios de clasificación para determinar la disponibilidad de los servicios de salud tanto en el sector público como privado.
- El resultado y análisis del presente proyecto, permite conocer a detalle cómo se encuentra la provisión de servicios de salud en el Ecuador, de acuerdo a los criterios de análisis establecidos (geografía, sector de salud, clase de establecimiento y entidad), permitiendo tomar decisiones para rectificar deficiencias detectadas en la provisión de servicios de salud, en relación con el número de especialidades o número de entidades de salud reportados.
- Se puede visualizar, la deficiencia en la dotación de servicios de salud relacionadas con las especialidades y entidades de salud y determinar en qué localidad geográfica exist menor provisión de especialidades y que tipo de entidades tienen menor presencia en el país.
- La suite Pentaho Report Designer permitió explotar de mejor manera la información disponible en el almacén, a través de reportes gráficos que muestran de manera estadística la situación de la provisión de servicios de salud en el país, mediante la aplicación de filtros de búsqueda, que segmentan la información que se quiere visualizar. Esto permite obtener una mejor interpretación de la información.
- El uso de la herramienta Weka, ha hecho posible la aplicación de varios algoritmos orientados a la minería de datos, gracias versatilidad y fácil manejo con grandes volúmenes de información, lo que permitió interpretar de mejor manera los resultados.
- A través de la aplicación del algoritmo bayesiano se puede constatar que la mayor concentración de tipos de entidades de salud se encuentra bajo el poder del gobierno central por parte del ministerio de salud pública y que existen ciertas clases de establecimiento inexistentes en los sectores privados, como es el caso de hospitales de infectología, dermatológicos (leprocomios) y geriátricos.

RECOMENDACIONES

- Para la construcción de un almacén de datos es importante definir, en primera instancia, de acuerdo a un previo análisis, la metodología que se aplicará para su creación. Esto permite optimizar tiempo y recursos y lograr resultados de alta calidad que permitan obtener, información del mismo.
- Antes de implementar un almacén de datos, es importante definir la viabilidad y su alcance. Esto se logra analizando las fuentes de información y la calidad de los datos disponibles y definiendo la granularidad de abstracción de la información.
- Como paso previo a trabajar con la información que pone a disposición el INEC, considerando que existe una seria de inconsistencias, es necesario tomar medidas de control como la normalización y depuración para poder sacar provecho de estos datos.
- Para sacar mayor provecho a una solución de inteligencia de negocios (*Bussiness Intelligence, BI*), sería importante llevar a cabo la toma de contacto con los usuarios finales y una capacitación, con el objetivo de cubrir todas sus expectativas, que permitan el uso y explotación del almacén de datos y herramientas de análisis implementadas, de manera que apoye a procesos de toma de decisiones, logrando un cambio cultural; considerando que los encargados de la toma de decisiones enfocan su actividad en construir información, más no en analizarla.

LISTA DE REFERENCIAS

Curto Díaz, J. (2010). Introducción al Business Intelligence. Barcelona: Editorial UOC.

Connolly, T (2005). Sistemas de Bases de Datos. Madrid: Pearson Educación S.A.

Silberschatz, A., Korth, H. and Sudarshan, S. (2002). Fundamentos de bases de datos. Cuarta Edición. McGraw-Hill Inc.

Pérez López, C., & Santín González, D. (2007). Minería de datos: técnicas y herramientas. Madrid: Thomson Ediciones Paraninfo, S.A.

Trujillo, J., Mazón, J., & Pardillo, J. (2010). Diseño y explotación de almacenes de datos. Alicante: Editorial Club Universitario.

OMS. (2003). Informe sobre la salud en el mundo. Francia: Organización Mundial de la Salud, World Health Report, 1211 Ginebra 27, Suiza.

Torres, E.; Arzuza, E. & Becerra, O. (2012) Aplicación de la metodología SCRUM para la optimización de procesos académicos. Universidad de San Buenaventura, Colombia. Recuperado de: http://bibliotecadigital.usbcali.edu.co/jspui/bitstream/10819/2353/1/Aplicaci%C3%B3n%20de%20la%20metodolog%C3%ADa%20SCRUM_Elkin%20Jos%C3%A9%20Torres%20Mart%C3%ADnez_USBCTG_2012.pdf

Lucio R., MSc., Villacrés N., MD., Henríquez R., MD. (2011). Sistema de salud de Ecuador. Recuperado de: http://www.scielosp.org

Instituto Superior Politécnico José Antonio Echeverría (CUJAE). (2016). Pentaho: software líder de Inteligencia de Negocio de código abierto. Recuperado de: http://revistatelematica.cujae.edu.cu/index.php/tele/article/viewFile/44/43

Ecuador en cifras. (2016) Censos I. Historia Instituto Nacional de Estadística y Censos. Recuperado de: http://www.ecuadorencifras.gob.ec/historia/

Instituto Nacional de Estadística y Censos (2016). Transparencia. Recuperado de: http://www.ecuadorencifras.gob.ec/transparencia/

Anda.inec.gob.ec. (2017). Ecuador - Estadísticas Hospitalarias Camas y Egresos 2012 - variable - V252. Recuperado de: http://anda.inec.gob.ec/anda/index.php/catalog/394/datafile/F7/V252

SCRUMstudy.(2016). Una guía para el cuerpo de conocimiento de SCRUM (SBOK Guide™). EEUU, Arizona: VMEdu, Inc. Recuperado de:

https://www.SCRUMstudy.com/SBOK/SCRUMstudy-SBOK-Guide-2016-spanish.pdf

E-stadistica.bio.ucm.es. (2017). Reemplazamiento de valores perdidos. [online] Recuperado de: http://e-stadistica.bio.ucm.es/web_spss/reemplazar_missing.html

Censos, I. (2017). Población y Demografía. [online] Instituto Nacional de Estadística y Censos. Recuperado de: http://www.ecuadorencifras.gob.ec/censo-de-poblacion-y-vivienda/

Rivadera, G. (2010). La metodología de Kimball para el diseño de almacenes de datos. 1st ed. [ebook] Argentina: Universidad Católica de Salta. Recuperado de: http://s3.amazonaws.com/academia.edu.documents/37011190/Metodologia-Kimball_DWH.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1496713400&Signature=sF%2FdZIndppcM6W%2FVmj9gQbrwhAk%3D&response-content-disposition=inline%3B%20filename%3DLa_metodologia_de_Kimball_para_el_diseno.pdf

Ayala, A. P. (2006). *Inteligencia de Negocios: Una Propuesta para su Desarrollo en las organizaciones.* México: Instituto Politécnico Nacional.

Bouckaert R., F. E. (2015). *WEKA Manual.* Hamilton, New Zealand: GNU.

Calabria, J. (26 de Junio de 2011). *Repositorio Digital UNIAUTONOMA.* Obtenido de Repositorio Digital UNIAUTONOMA: http://repositorio.uac.edu.co/bitstream/handle/11619/1298/Construcci%C3%B3n%20y%20poblamiento%20de%20un%20datawarehouse%20basado%20en%20el.pdf?sequence=1&isAllowed=y

Cárdenas M. (01 de Dicembre de 2014). *CIEMAT*. Obtenido de CIEMAT: http://wwwae.ciemat.es/~cardenas/docs/lessons/ArbolesDeDecision.pdf

Gálvez, A. P. (2015). *Business Intelligence Y La Tecnología de la Información.* (C. I. Platform, Ed.) USA: United States.

Garzón N. M., T. L. (2015). *Ingeniería del conocimiento.* Obtenido de Facultad de Ciencias básicas e ingeniería: http://fcbi.unillanos.edu.co/cici/Articulos/CICI_2016_paper_14.pdf

Gómez, A. J. (2012). Inteligencia de negocios, una ventaja competitiva para las organizaciones. *Revista "Ciencia y Tecnología", Escuela de Postgrado - UNT, 8*. Perú.

IBM. (noviembre de 2016). *IBM Knowledge Center*. Obtenido de http://www.ibm.com/support/knowledgecenter/es/SSLVMB_22.0.0/com.ibm.spss.statistics.help/spss/mva/idh_miss_pat.htm

Oliver A. (17 de 06 de 2008). *Universidad de Girona.* Obtenido de Universidad de Girona: http://eia.udg.edu/~aoliver/publications/tesi/node143.html

Rivadera, G. (2010). *Universidad Católica de Salta.* Obtenido de Universidad Católica de Salta: http://www.ucasal.edu.ar/htm/ingenieria/cuadernos/archivos/5-p56-rivadera-formateado.pdf

Sánchez, J. L., Vargas, A. P., & otros. (noviembre de 2016). *Introducción a SPSS*. Obtenido de Aula Virtual de Bioestadística: http://e-stadistica.bio.ucm.es/web_spss/reemplazar_missing.html

The MathWorks, Inc. (03 de 01 de 2016). *Mathworks*. Obtenido de https://es.mathworks.com

Thomas M. Connolly, C. E. (2005). Sistemas de bases de datos: un enfoque práctico para diseño, implementación y gestión. En C. E. Thomas M. Connolly, *Sistemas de bases de datos: un enfoque práctico para diseño, implementación y gestión.* Pearson Educación.

Todd A. (2000). *An Introduction to Bayesian Networks.* Martigny.

Waikato, M. L. (2017). *Machine Learning Group at the University of Waikato*. Obtenido de Machine Learning Group at the University of Waikato: http://www.cs.waikato.ac.nz/ml/weka/