



**Universidad de las Ciencias Informáticas
Facultad 6**

Título: Sistema de Información de Gobierno. Mercado de Datos Salud Pública y Asistencia Social.

Trabajo de Diploma para optar por el título de
Ingeniero en Ciencias Informáticas.

Autores:

Erllys de la Caridad Quintero Méndez.

Orelmis Aguilar Álvarez.

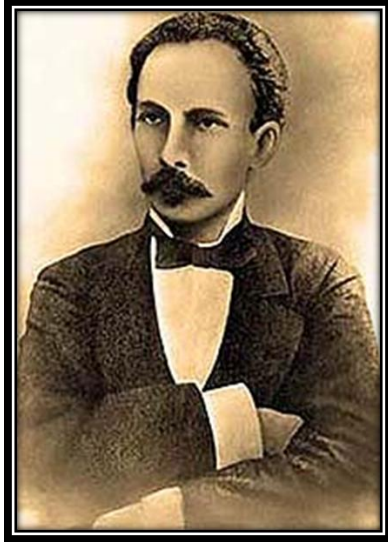
Tutores:

Ing. Yurima Ibañez Alfonso

Ing. Yoendris Lacoste Ricardo

La Habana, Junio de 2011

“Año 53 de la Revolución”



A nuevas ciencias que todo lo invaden, reforman y minan, nuevas cátedras.

José Martí

DECLARACIÓN DE AUTORÍA

Declaramos ser autores de la presente tesis y reconocemos a la Universidad de las Ciencias Informáticas los derechos patrimoniales de la misma, con carácter exclusivo.

Para que así conste firmo la presente a los ____ días del mes de _____ del año _____.

Erllys de la Caridad Quintero Méndez

Firma del Autor

Orelmis Aguilar Álvarez

Firma del Autor

Yoendris Lacoste Ricardo

Firma del Tutor

Yurima Ibañez Alfonso

Firma del Tutor

AGRADECIMIENTOS

A mis profesores, caudal del que bebí el conocimiento

A Carlos por su ayuda incondicional

A Teresa por estar siempre ahí, cuando la necesitaba,

A mis tutores, sin los que no hubiese sido posible desarrollar, este trabajo.

A Todos lo que no he mencionado pero que se sienten parte de este proyecto, mi agradecimiento.

Erllys

AGRADECIMIENTOS

A mis padres, por el apoyo incondicional, por el infinito amor que a diario cultivan, por ayudarme a trazar el camino correcto, por confiar en mí en cada momento a pesar de las dificultades.

A mis familiares por ser la mejor familia del mundo.

A todos los que compartieron esta etapa conmigo y especialmente a Leonel, a Liniuska, a Manresa, a Yoendy, a Fabian por dejarme compartir con ellos estos años y por ayudarme cuando lo he necesitado.

A mis colegas de la música por los buenos consejos y por estar ahí en todo momento.

A todos los que de una forma u otra se hicieron parte de mi vida y me ayudaron a construir este sueño.

A los tutores, por la ayuda invaluable y por la guía que han sido para mí porque sin ella no lo hubiera logrado, son el mejor equipo de trabajo en verdad.

A todos los que no menciono porque serían muchos nombres, pero de quienes guardo muchos recuerdos e historias y pueden estar seguros que de todos he aprendido algo. Gracias por la colaboración, y por aun seguir aquí.

Orelmis

DEDICATORIA

A mi mami por ser el ángel hacedor de mis triunfos, por enseñarme a escalar en la vida, por el sacrificio que sin pedirle siempre ha hecho por mí.

A la Universidad de las Ciencias Informáticas, hazaña de la Revolución que me convierte en una profesional al servicio de la patria.

A mi familia en general que ha estado siempre orgulloso de mis adelantos en la carrera, y ha confiado en mí.

Erllys

Dedico este trabajo a mis padres, responsables de mi existencia y de mis logros, por su apoyo incondicional, y a mi familia por su ayuda, en el largo camino que he transitado hasta aquí.

Orelmis

RESUMEN

La solución del presente trabajo de diploma está enmarcada en los sistemas de soporte a la toma de decisión (DSS), el cual a su vez engloba un conjunto de acciones y procesos que hay que llevar a cabo para poder implementar el mismo. El trabajo de diploma abarca una investigación de los elementos que componen un DSS, los mismos son los mercados de datos o almacenes de datos, los procesos de ETL (extracción, transformación y carga) y BI (inteligencia de negocio); también se detallan las herramientas y metodología para el desarrollo de este tipo de soluciones. Se efectúa un levantamiento de requisitos, para conocer las necesidades del usuario y pasar al proceso de definición del diseño de la solución. Como resultado se obtiene la estructura del modelo dimensional que comprende: las dimensiones, las jerarquías, las tablas de hechos y las medidas necesarias para proceder a la visualización de los datos. Se precisan las reglas del negocio utilizadas y se detalla el proceso de carga de los datos de la fuente de Salud Pública y Asistencia Social al mercado de datos en cuestión, al igual que se detalla el proceso de la capa de visualización de los reportes. De igual manera, la solución incluye las estrategias de seguridad, respaldo y recuperación de los datos, así como el proceso de validación completo. Una de las características especiales de este sistema, se basa en la utilización de herramientas libres en su implementación.

CONTENIDO

AGRADECIMIENTOS	I
AGRADECIMIENTOS	II
DEDICATORIA	III
RESUMEN	IV
INTRODUCCIÓN	1
CAPÍTULO 1: FUNDAMENTACIÓN TEÓRICA	4
Introducción	4
1.2 Tecnologías de almacenamiento de datos.....	4
1.2.1 Bases de Datos	4
1.2.3 Mercado de Datos (Data Mart).....	5
1.2.4 Almacenes de Datos (Data Warehouse)	5
1.3 Integración de datos	6
1.3.1 Ventajas de la integración de datos.....	8
1.4 Inteligencia de Negocios (BI).....	9
1.5 Metodologías para el desarrollo del mercado de datos.....	11
1.6 Herramientas para el proceso de ETL	12
1.7 Herramientas para el proceso Inteligencia de Negocios (BI).....	13
1.8 Herramientas de modelado	15
1.9 Sistema Gestor de Bases de Datos.....	16
1.9.1 Herramienta para gestionar Bases de Datos. PgAdmin III	17
1.10 Modos de almacenamiento de datos	18
1.10.1 Proceso Analítico Relacional (ROLAP).....	18
1.10.2 Proceso Analítico Multidimensional (MOLAP).....	18
1.10.3 Proceso Analítico Híbrido (HOLAP)	19
1.10.4 Comparación entre los modos de almacenamiento de los datos ROLAP y MOLAP	19
1.10.5 Justificación del modo de almacenamiento a utilizar.....	20
Conclusiones	21
CAPITULO 2: ANÁLISIS Y DISEÑO DEL MERCADO DE DATOS	22
Introducción.....	22
2.1 Estudio preliminar del negocio	22
2.2 Necesidades de información.....	24
2.3 Requisitos funcionales	24

2.4 Requisitos no funcionales	24
2.5 Requisitos de información	25
2.6 Reglas del negocio.....	28
2.7 Modelo de Casos de Uso del Sistema	29
2.7.1 Actores del sistema	29
2.7.2 Casos de Uso del Sistema.....	30
2.7.3 Diagrama de CUS.....	31
2.8 Desarrollo de la matriz BUS	33
2.9 Identificación de las dimensiones, hechos y medidas.	34
2.10 Arquitectura de la información.....	41
2.11 Política de seguridad	41
2.12 Política de respaldo y recuperación.....	43
Conclusiones	44
CAPÍTULO 3: IMPLEMENTACIÓN DEL MERCADO DE DATOS.....	45
Introducción.....	45
3.1 Estructura de datos.....	45
3.1.2 Esquemas y tablas.....	45
3.1.2 Restricciones y secuencias	46
3.1.3 Índices	49
3.2 Usuarios y privilegios	50
3.3 Implementación de los procesos ETL.....	51
3.3.1 Implementación de los flujos de transformación	52
3.4 Implementación de los procesos de BI	52
3.4.1 Implementación de los cubos OLAP.....	53
3.4.2 Reportes candidatos	53
Conclusiones	55
CAPÍTULO 4: VALIDACIÓN Y PRUEBA DEL MERCADO DE DATOS.....	56
Introducción.....	56
4.1 Pruebas de software	56
4.2 Elaboración y aplicación de las Listas de Chequeo	57
4.3 Casos de prueba.....	60
4.4 Prueba de aceptación	61
Conclusiones	61
CONCLUSIONES GENERALES.....	62
RECOMENDACIONES.....	63

TRABAJOS CITADOS	64
BIBLIOGRAFÍA	65
ANEXOS	67
GLOSARIO DE TÉRMINOS	69

INTRODUCCIÓN

En los últimos años ha sido notable el desarrollo mundial de la industria del software. También es visible la necesidad de muchas naciones de aprovechar estos avances para controlar el elevado volumen de información que se maneja en cada uno de estos países, con el propósito de extraer estadísticas acertadas y estratégicas que permitan la detección de amenazas y oportunidades con anticipación y de esta forma poder tomar decisiones correctas.

Cuba no está exenta a los avances tecnológicos existentes en el mundo. En nuestro país se está llevando a cabo un proceso de informatización que comprende todas las esferas de la sociedad y la economía. Como resultado de este proceso es fundada la Universidad de las Ciencias Informáticas (UCI). Este centro de altos estudios vincula la docencia con la producción para formar profesionales competentes y fortalecer la industria del software. La línea productiva de la UCI está conformada por varios centros de desarrollo, uno de estos es DATEC, centro que tiene bajo su responsabilidad varios proyectos. Uno de los proyectos más importantes del centro es el relacionado con la Oficina Nacional de Estadísticas (ONE). Institución creada para proponer, organizar y ejecutar, según corresponda, la aplicación de la política estatal en materia de estadística del país. Para ello cuenta con el Sistema Estadístico Nacional (SEN), el cual está conformado por órganos, organismos, entidades e instituciones que coordinadamente generan información estadística importante para el país; llevando un control confiable de los datos estadísticos, a través de una red de transmisión de datos por toda Cuba, desde las oficinas que radican en cada municipio. Por lo que se guardan datos históricos de distintos esferas de la sociedad cubana tales como los de la Salud, Medio Ambiente, Turismo, Educación. Por lo que se generan anualmente grandes volúmenes de información que deben ser analizados manualmente para la obtención de información, impidiéndose de esta forma la inmediatez en la obtención de reportes.

Actualmente, en aras de satisfacer las necesidades de información existentes, parte del análisis estadístico se realiza a través de mecanismos no automatizados, poco confiables y tediosos debido a que la información se encuentra en formato Excel y solo puede ser accedida por un especialista de la informática que tenga conocimiento del negocio. Existen múltiples versiones de los datos, lo que provoca que existan datos no integrados y por otra parte mala calidad de los mismos. Todo esto dificultando la toma de decisiones. Esto afecta la calidad de la información, medida bajo los parámetros de seguridad, disponibilidad, rapidez y facilidad de accesos necesarios para este centro.

A partir de lo antes expuesto se define como **Problema científico**: ¿Cómo contribuir a la toma de decisiones en el área de Salud Pública y Asistencia Social?

Para dar solución al Problema científico se define como **Objeto de estudio**: Almacenes de datos. Enmarcado en el **Campo de acción**: Mercado de datos para el área de Salud Pública y Asistencia Social del Sistema de Información de Gobierno.

Para dar solución al Problema científico se plantea como **Objetivo general**: Desarrollar el mercado de datos Salud Pública y Asistencia Social del Sistema de Información de Gobierno que contribuya a la toma de decisiones.

A partir de un análisis del Objetivo general se derivaron los siguientes **Objetivos específicos**:

- Realizar el análisis y diseño del mercado de datos del área Salud Pública y Asistencia Social.
- Implementar el mercado de datos del área Salud Pública y Asistencia Social.
- Validar el mercado de datos del área Salud Pública y Asistencia Social.

Para dar cumplimiento a los objetivos planteados se definen las siguientes **Tareas de la Investigación**:

- Caracterización de las metodologías, herramientas y tecnologías a utilizar en el desarrollo de almacenes de datos.
- Levantamiento de los requisitos del mercado de datos.
- Descripción de los casos de uso del mercado de datos.
- Diseño de los casos de prueba.
- Definición de la arquitectura del mercado de datos.
- Diseño del modelo de datos.
- Diseño del subsistema de integración.
- Diseño del subsistema de visualización.
- Implementación del subsistema de integración.
- Implementación del subsistema de visualización.

- Aplicación de las listas de chequeo.
- Aplicación de los casos de pruebas.

Para garantizar el cumplimiento de todos los elementos planteados anteriormente, la presente investigación estará compuesta por cuatro capítulos que serán resumidos a continuación.

El capítulo 1 **"Fundamentación Teórica"**: Está referido al estudio sobre los sistemas de almacenes de datos y los mercados de datos, sus principales metas, características, y los elementos fundamentales que los componen. Se describen las metodologías existentes y las principales herramientas para el desarrollo de los almacenes de datos a nivel mundial. También está referido al estudio de las técnicas y herramientas para el desarrollo de la capa de visualización.

El capítulo 2 **"Análisis y diseño del mercado de datos Salud Pública y Asistencia Social"**: Aborda aspectos referentes al análisis de los datos, el modelo dimensional propuesto, tablas de hechos, tablas de dimensiones y medidas identificadas. Se describen los casos de uso identificados al igual que las relaciones de los actores con los casos de uso.

El capítulo 3 **"Implementación del mercado de datos Salud Pública y Asistencia Social"**: Aborda aspectos referentes a la descripción e implementación de la solución.

El capítulo 4 **"Validación del mercado de datos Salud Pública y Asistencia Social"**: Se valida la solución a través del empleo de las listas de chequeo, entre otras.

CAPÍTULO 1: FUNDAMENTACIÓN TEÓRICA

Introducción

El primer capítulo de esta investigación aborda los temas referentes a los almacenes de datos y mercados de datos, sus principales características y sus componentes. También se realiza un estudio de las metodologías existentes además de las herramientas fundamentales para el desarrollo de los almacenes de datos y de la capa de visualización.

1.2 Tecnologías de almacenamiento de datos

Desde el surgimiento de las tecnologías de almacenamiento de datos, estas se han convertido en una herramienta fundamental para el control y manejo de operaciones. Estas no solo son útiles para el manejo de operaciones y para almacenar información, sino que también nos ayuda en los procesos de toma de decisiones. A continuación se abordará sobre varias tecnologías de almacenamiento de datos que han surgido a través de los años y que forman hoy día un elemento importante para el almacenamiento de información en disímiles empresas.

1.2.1 Bases de Datos

Una base de datos es un conjunto de datos que pertenecen al mismo contexto almacenados sistemáticamente para su posterior uso. En este sentido, una biblioteca puede considerarse una base de datos compuesta en su mayoría por documentos y textos impresos en papel e indexados para su consulta. En la actualidad, y debido al desarrollo tecnológico de campos como la informática y la electrónica, la mayoría de las bases de datos tienen formato electrónico, lo que ofrece un amplio rango de soluciones al problema de almacenar datos. [1]

Existen varios tipos de Bases de datos:

- Bases de datos estáticas: Estas son bases de datos de solo lectura, utilizadas primordialmente para almacenar datos históricos que posteriormente se pueden utilizar para estudiar el comportamiento de un conjunto de datos a través del tiempo, realizar proyecciones y tomar decisiones. [1]
- Bases de datos dinámicas: Estas son bases de datos donde la información almacenada se modifica con el tiempo, permitiendo operaciones como actualización y adición de datos, además de las operaciones fundamentales de consulta. Un ejemplo de esto puede ser la base de datos utilizada en un sistema de información de una tienda, una farmacia, un

videoclub, etc. [1]

1.2.3 Mercado de Datos (Data Mart)

El data mart es una base de datos departamental o un subconjunto de una bodega de datos, especializada en el almacenamiento de los datos de un área de negocio específica, para un propósito específico. Un data mart puede ser alimentado desde los datos de un data warehouse, o integrar por sí mismo un compendio de distintas fuentes de información. Su función es apoyar a otros sistemas para la toma de decisiones.

El motivo por el cual se crean mercados de datos es el crecimiento que tiene el almacén de datos y así facilitar su construcción y utilización. Las características de los mercados de datos son:

- Se centran en los requisitos de los usuarios asociados a un departamento o área de negocio concretos. [2]
- Como diferencia con los almacenes de datos, los mercados no contienen datos operacionales detallados. [2]
- Son más sencillos a la hora de utilizarlos y comprender sus datos, debido a que la cantidad de información que contienen es mucho menor que en los almacenes de datos. [2]

1.2.4 Almacenes de Datos (Data Warehouse)

A partir de los años noventa surgen los almacenes de datos (data warehouse) debido a que las empresas necesitaban guardar grandes volúmenes de información, conservarla y consultarla en algún momento para poder llevar a cabo un análisis de la información almacenada, además de que los Data Warehouse sirven de base para la toma de decisiones.

En el libro “**Mastering Data Warehouse Design Relational and Dimensional Techniques**” se precisa que los Data Warehouse producen una fuente estable de la información histórica que es constante, consistente y fiable para cualquier consumidor, además de que está configurado para facilitar los datos de cualquier forma de tecnología de análisis dentro de la comunidad de negocios. La definición universalmente aceptada fue desarrollada por Bill Inmon en los años noventa, y dice que los Data Warehouse “Son un conjunto de datos orientados a un tema, integrados, de tiempo variante y no volátiles usados en la estrategia de toma de decisiones administrativas.” Las características que tienen los Data Warehouse según Bill Inmon son las siguientes:

Integrado: Los datos almacenados en Data Warehouse deben integrarse en una estructura consistente.

Temático: Los datos se organizan por temas para facilitar su acceso y entendimiento por parte de los usuarios finales.

Histórico: El almacén de datos se carga con los distintos valores que toma una variable en el tiempo para permitir comparaciones. Se utiliza para realizar análisis de tendencias.

No volátil: El almacén de datos puede ser leído pero no modificado. Es decir, la incorporación de los últimos valores que tomarán las distintas variables contenidas en él sin ningún tipo de acción sobre lo que ya existía.

1.3 Integración de datos

El proceso de integración de datos está basado en la necesidad de unificar los datos pertenecientes a base de datos fuentes, archivos u otro sistema de almacenamiento de datos. Este proceso se lleva a cabo con el fin de poder tener una mejor visión del problema a resolver, así como para poder sincronizar los datos de las diversas aplicaciones. Hoy día existen varias tecnologías de integración de datos, entre ellas se encuentra Enterprise Application Integration (EAI); Enterprise Information Integration (EII); Extract, Transform, Load (ETL); las mismas han avanzado lo suficiente como para rendir significativos retornos al negocio, pero han sido desarrolladas durante mucho tiempo de forma independiente unas de las otras. EAI ha evolucionado para ser una tecnología destinada a la mensajería en tiempo real. EII nos ha ofrecido un querying federado y on-the-fly. ETL ha sido tradicionalmente muy bueno en el manejo de transformaciones de datos complejos y BI se beneficia de los datos procedentes de todos estos métodos de integración para comprender el estado en que se encuentra una empresa.

Son muchos libros que abordan al respecto, un ejemplo de eso es el libro “Wiley Publishing The Data Warehouse ETL Toolkit”, en el cual en su primer capítulo, se hace una referencia en cuanto a la integración de datos y se precisa que “En el sistema de ETL, la integración de datos es un paso separado identificado en nuestro hilo de flujo de datos que conforman el paso. Físicamente, este paso implica hacer cumplir nombres comunes de los atributos de dimensión conformada y los hechos, así como la aplicación de contenidos comunes de dominio y unidades comunes de medición”.

La tecnología ETL (extracción, transformación y carga) está enfocada a la integración de datos, tanto por lotes como a tiempo real hacia almacenes de datos [3]. Esta integración de datos requiere un enfoque constante e interactivo ya que una buena integración ayuda a evitar/reducir las redundancias y las inconsistencias. El problema de la redundancia aparece al integrar los datos de las diferentes fuentes de datos.

El proceso ETL se divide a su vez en tres subprocesos los cuales se explican a continuación:

Extracción

La primera parte del proceso ETL consiste en la extracción de los datos desde los sistemas de origen. Es necesario fusionar los datos provenientes de las diferentes fuentes de almacenamiento. Los formatos de las fuentes normalmente se encuentran en bases de datos relacionales o ficheros planos, pero pueden incluir bases de datos no relacionales u otras estructuras diferentes. La extracción convierte los datos a un formato preparado para iniciar el proceso de transformación. Una parte intrínseca del proceso de extracción es la de analizar los datos extraídos, de lo que resulta un chequeo que verifica si los datos cumplen la pauta o estructura que se esperaba. De no ser así los datos son rechazados. Un requerimiento importante que debe exigir la tarea de extracción es que esta cause un mínimo impacto en el sistema origen, pues si los datos a extraer son muchos, el sistema de origen se podría colapsar, provocando que no pueda utilizarse con normalidad para su uso cotidiano. Por esta razón, en sistemas grandes las operaciones de extracción suelen programarse en horarios o días donde este impacto sea nulo o mínimo.

Limpieza y Transformación

La transformación es la encargada de convertir aquellos datos inconsistentes en un conjunto de datos compatibles y congruentes, para que puedan ser cargados en el almacén de datos. Estas acciones se llevan a cabo, debido a que pueden existir diferentes fuentes de información, y es vital conciliar un formato y forma única, definiendo estándares, para que todos los datos que ingresarán al almacén de datos estén integrados.

Pero no solo es necesario hacer transformaciones a los datos, a estos también se les realiza un proceso de Limpieza de Datos (Data Cleansing). Para esta etapa medular, los datos deben ser limpiados, pues en el mundo real son sucios; se muestran incompletos, donde faltan valores de los atributos, o contienen solo agregados de datos; se presentan con ruido, conteniendo errores o valores fuera de límites, manteniendo discrepancias en nombre o códigos. Por ello se realiza un

proceso de limpieza que elimina errores e inconsistencias en los datos y resuelve el problema de identidad de los objetos. Este proceso tiene como tareas fundamentales: llenar valores ausentes, identificar valores fuera de límite, eliminar el ruido en los datos, corregir las inconsistencias de los datos, e integrarlos.

Luego que los datos se encuentran limpios, homogeneizados, y se cuenta con metadatos fiable. Se generan las reglas de transformación que va desde realizar tratamientos a valores nulos, aplicar reglas del negocio, combinar los datos de las distintas fuentes o realizar búsquedas de valores en diversas tablas, hasta implementar agregaciones que aceleran los tiempos de análisis, ocultan complejidad de los datos, y proveen múltiples vistas del mismo conjunto de datos. Culminando esta fase con la retroalimentación del flujo inverso de datos limpios.

Carga

La fase de carga es el momento en el cual los datos de la fase anterior (transformación) son cargados en el sistema de destino. Dependiendo de los requerimientos de la organización, este proceso puede abarcar una amplia variedad de acciones diferentes. En algunas bases de datos se sobrescribe la información antigua con nuevos datos. Los almacenes de datos mantienen un historial de los registros de manera que se pueda hacer una auditoría de los mismos y disponer de un rastro de toda la historia de un valor a lo largo del tiempo.

La fase de carga interactúa directamente con la base de datos de destino. Al realizar esta operación se aplicarán todas las restricciones y triggers¹ (disparadores) que se hayan definido en esta (por ejemplo, valores únicos, integridad referencial, campos obligatorios, rangos de valores). Estas restricciones y triggers (si están bien definidos) contribuyen a que se garantice la calidad de los datos en el proceso ETL, y deben ser tenidos en cuenta.

En el siguiente epígrafe se abordará sobre las ventajas que ofrece la integración de datos.

1.3.1 Ventajas de la integración de datos

La integración de datos tiene varias ventajas, entre ellas están las siguientes:

¹ Llamados también disparadores, es un procedimiento que se ejecuta cuando se cumple una condición establecida al realizar una operación de inserción (INSERT), actualización (UPDATE) o borrado (DELETE).

- Control sobre la redundancia de datos. Los sistemas de ficheros almacenan varias copias de los mismos datos en ficheros distintos. Esto hace que se desperdicie espacio de almacenamiento, además de provocar la falta de consistencia de datos. En los sistemas de bases de datos todos estos ficheros están integrados, por lo que no se almacenan varias copias de los mismos datos. Sin embargo, en una base de datos no se puede eliminar la redundancia completamente, ya que en ocasiones es necesaria para modelar las relaciones entre los datos, o bien es necesaria para mejorar las prestaciones.
- Consistencia de datos. Eliminando o controlando las redundancias de datos se reduce en gran medida el riesgo de que haya inconsistencias. Si un dato está almacenado una sola vez, cualquier actualización se debe realizar solo una vez, y está disponible para todos los usuarios inmediatamente. Si un dato está duplicado y el sistema conoce esta redundancia, el propio sistema puede encargarse de garantizar que todas las copias se mantienen consistentes. Desgraciadamente, no todos los Sistemas Gestores de Base de Datos (SGBD) de hoy día se encargan de mantener automáticamente la consistencia.
- Compartición de datos. En los sistemas de ficheros, los ficheros pertenecen a las personas o a los departamentos que los utilizan. Pero en los sistemas de bases de datos, la base de datos pertenece a la empresa y puede ser compartida por todos los usuarios que estén autorizados. Además, las nuevas aplicaciones que se vayan creando pueden utilizar los datos de la base de datos existente.
- Mantenimiento de estándares. Gracias a la integración es más fácil respetar los estándares necesarios, tanto los establecidos a nivel de la empresa como los nacionales e internacionales. Estos estándares pueden establecerse sobre el formato de los datos para facilitar su intercambio, pueden ser estándares de documentación, procedimientos de actualización y también reglas de acceso.

1.4 Inteligencia de Negocios (BI)

Al igual que otros conceptos o términos, el de Inteligencia de Negocios no escapa a la variedad de interpretaciones. Sin embargo queda esencialmente claro que: “no es una metodología, software, sistema o herramienta específica, es más bien una colección de tecnologías que van desde arquitecturas para almacenar datos, metodologías, técnicas para analizar información y software, entre otros, con un fin común para el apoyo a la toma de decisiones”.

De acuerdo al **Data Warehousing Institute**, la definición de Inteligencia de Negocios es la siguiente:

“...Son los procesos, tecnologías, y herramientas que se necesitan para convertir los datos en información, la información en conocimiento, y el conocimiento en planes que impulsan acciones rentables para el negocio. La Inteligencia de Negocios abarca el almacenamiento de datos, herramientas analíticas, y contenido y gestión del conocimiento...”

Después de analizar el concepto anterior se puede decir que la Inteligencia de Negocio es considerada un término "agrupador". El que sea considerado como una colección de conceptos le da un poder enorme, pues pueden integrarse funciones que tradicionalmente estaban separadas, tales como el acceso de datos, reporte, explotación, pronóstico y análisis.

De este modo, dentro de las empresas grandes, las soluciones de Inteligencia de Negocios se han convertido en un apoyo indispensable para la toma de decisiones, en cualquier nivel de la organización y mucha gente está explotando el potencial estratégico de los datos operativos. Bien utilizada, BI puede ser un arma estratégica para los cargos directivos, sustentada en tecnología de sistemas informáticos.

Las soluciones de Inteligencia de Negocios proporcionan amplias ventajas ya sean competitivas para las empresas o de índole cognoscitivo en temas no empresariales, al permitir que los datos se conviertan en un centro de beneficios, facilitando el análisis de información útil a las organizaciones para la toma de decisiones.

Entre las principales ventajas que ofrece una solución de Inteligencia de Negocio se tienen:

- Control de costes, al tener un solo sistema que permite manejar fácilmente los distintos programas que se encuentran en los diferentes departamentos de su compañía.
- Mejora de la colaboración y la calidad de las decisiones, facilitando el acceso a la información en todos los niveles de la organización.
- Orienta las soluciones tecnológicas hacia el usuario, porque reduce los tiempos de aprendizaje mediante el uso de herramientas de uso cotidiano.

1.5 Metodologías para el desarrollo del mercado de datos.

Para la creación de un producto informático es necesario realizar varios pasos, es ahí donde juega un papel muy importante la metodología, esta es el conjunto de procedimientos, técnicas y herramientas que ayuda a los desarrolladores a realizar un software. Estas indican quien debe hacer qué, cuándo y cómo. Las metodologías de desarrollo de software surgieron a raíz de la necesidad de documentar proyectos muy complejos, impulsadas principalmente por instituciones económicas muy importantes. Para diseñar un mercado de datos también existen varias metodologías, estas son las que definen y guían todo el ciclo de vida del desarrollo completo.

Debido a la necesidad de una metodología robusta que garantice eficazmente la integración de la información que se dispone en el área de Salud Pública y Asistencia Social de la ONE, se tomó como base la metodología de Kimball, adecuándola a las necesidades de la UCI. Esta metodología fue la seleccionada por los siguientes elementos:

- Crea los conceptos de Hechos y Dimensiones, lo que indudablemente es muy eficaz en el proceso de la toma de decisiones y proporciona mayor agilidad en el proceso de desarrollo.
- Propone ir construyendo el Almacén de Datos a través de la construcción de los Mercados de Datos departamentales, lo que constituye una estrategia buena y coincide con la división lógica de las empresas, entidades, organismos, etcétera.
- La técnica de Kimball posee una abundante documentación, la respuesta a todas las dudas y preguntas que puedan surgir se pueden encontrar en la Web, a través de los servicios que brindan el grupo creador de la metodología.
- Es una metodología madura y reconocida por el resto de la comunidad dedicada al tema. Tiene bien definidas las etapas, actividades, artefactos y roles.

Como complemento a la misma y fortaleciendo la etapa del levantamiento de requisitos; se tomó lo planteado por Leopoldo Zenaido Zepeda Sánchez en su Tesis de Doctorado, orientando así el trabajo a los Casos de Uso y se logra estar más alineado con las tendencias y normas de la Universidad.

Siguiendo lo planteado en las metodologías seleccionadas como base y teniendo en cuenta las características de la UCI y DATEC se mencionan las fases y los flujos de trabajo del modelo de desarrollo con el cual se trabajará, el mismo fue realizado por especialistas del centro DATEC.

- Requerimientos y Gestión de Proyectos

- Arquitectura Técnica
- Diseño e Implementación
- Implementación y Crecimiento

Durante el ciclo de vida se tienen los siguientes flujos de trabajo:

- Estudio Preliminar o Planeación
- Requerimientos
- Arquitectura y Diseño
- Implementación
- Prueba
- Despliegue
- Soporte y Mantenimiento
- Gestión y Administración del Proyecto

1.6 Herramientas para el proceso de ETL

En el mundo actual se han desarrollado varias herramientas para los procesos de ETL con el fin de dar un acercamiento a la automatización, construcción, implementación y mantenimiento de los almacenes de datos; existen una amplia variedad de las mismas tanto comerciales como de código abierto.

A continuación se menciona la herramienta a utilizar en la implementación del mercado de datos de Salud Pública y Asistencia Social, y las características de la misma:

La herramienta a utilizar para llevar a cabo la integración de datos en el mercado de datos es el **Pentaho Data Integration (Kettle)** ya que esta abre, limpia e integra la información y la pone en manos del usuario. Provee una consistencia, una sola versión de todos los recursos de información, que es uno de los más grandes desafíos para las organizaciones hoy día. Pentaho Data Integration proporciona una extracción de gran alcance, transformación y carga (ETL) utilizando un enfoque innovador, orientado a los metadatos. El Kettle permite evitar grandes cargas de trabajo manual frecuentemente difícil de mantener y de desplegar. Algunas de las propiedades de esta herramienta son mencionadas a continuación.

Aparte de ser open source y sin costes de licencia, las características básicas de esta herramienta son:

- Entorno gráfico de desarrollo.
- Uso de tecnologías estándar: Java, XML, JavaScript
- Fácil de instalar y configurar.
- Multiplataforma: Windows, Linux.
- Admite una amplia gama de formatos de entrada y salida, incluyendo archivos de texto, hojas de datos, archivos XML, propiedades de Java y los motores de base de datos open source y propietarios.
- Cada proceso es creado con una herramienta gráfica donde se especifica qué se va hacer sin necesidad de escribir código que indique cómo hacerlo.
- Basado en dos tipos de objetos: Transformaciones (colección de pasos en un proceso ETL) y trabajos (colección de transformaciones).
- Ofrece una licencia pública GPL.
- Soporta Oracle, DB2, SQL Server y Sybase así como MySQL, Postgres, entre otras.

Esta herramienta fue seleccionada por las características antes mencionadas además por tener entre sus ventajas, que es una de las más antiguas herramientas ETL de código abierto, cuenta con una gran comunidad de usuarios y su interfaz gráfica permite un aumento de la productividad. Después de elegirse la herramienta para la implementación del proceso de ETL se da paso a la elección de la herramienta para los procesos de inteligencia de negocios. En el siguiente epígrafe se abordará al respecto.

1.7 Herramientas para el proceso Inteligencia de Negocios (BI)

Las herramientas de Inteligencia de Negocios están diseñadas para apoyar la toma decisiones dentro de las empresas o instituciones del estado, mostrando una visión general de todos los procesos de la entidad a sus directivos, facilitando el análisis y la presentación de los datos. A continuación se presentarán las herramientas a utilizar.

De las herramientas a utilizar para realizar el proceso de BI, están en primer lugar el *Mondrian Schema Workbench* la cual se utilizará para la creación de los cubos dimensionales. En segundo lugar será útil el *Pentaho BI Server* ya que es el encargado de los servicios básicos además de incluir autenticación, registro, auditoría, servicios web y motor de reglas, incorporando un motor

de solución que integra reportes, análisis, tableros de comandos y componentes de minería de datos.

A continuación se mencionan las principales características de estas herramientas.

Mondrian Schema Workbench

Mondrian Schema Workbench es un entorno visual para el desarrollo y prueba de cubos OLAP Mondrian. Si bien la definición del XML para esquemas Mondrian no es extremadamente compleja, en la práctica resulta engorroso recordar cada uno de los elementos junto a sus atributos y sub-elementos. Con esta aplicación, se puede configurar una conexión JDBC como el modelo físico, para luego elaborar el esquema lógico de manera simple y efectiva. Permite crear y probar los cubos OLAP visualmente para que luego el motor de Mondrian procese las solicitudes MDX con los esquemas creados. Esta herramienta ofrece las siguientes funcionalidades:

- Editor de esquemas integrados con un origen de datos subyacente para su validación.
- Prueba de consultas MDX sobre el esquema y la base de datos.
- Examinar la estructura subyacente de bases de datos.

Pentaho BI Server

El B.I. Server de Pentaho es una aplicación 100% Java2EE que permite gestionar todos los recursos de BI. Cuenta con una Interfaz de Usuario de BI donde se encuentran disponibles todos los informes y vistas OLAP. También tiene acceso a una consola de administración que permitirá gestionar y supervisar tanto la aplicación como los usuarios. Qué informes consulta cada usuario, cuando se han consultado, el rendimiento de la aplicación, etc.

Con esta plataforma se suministra soporte e infraestructura para crear soluciones de inteligencia de negocio. Proporciona servicios básicos además de incluir autenticación, registro, auditoría, servicios web y motor de reglas. Incorpora un motor de solución que integra reportes, análisis, tableros de comandos y componentes de minería de datos. Funciona como un sistema basado en administración web de informes, el servidor de integración de aplicaciones y un motor de flujo de trabajo ligero (secuencias de acciones). Además, está diseñada para integrarse fácilmente en cualquier proceso de negocio.

Tres de sus principales ventajas son:

1. Administra y programa reportes.
2. Administra seguridad de usuarios.

3. Brinda la posibilidad de guardar la consulta que se ejecute.

Luego de haber sido elegidas las herramientas para la implementación de la capa de visualización se da paso al siguiente epígrafe, donde se abordará sobre la herramienta de modelado a utilizar.

1.8 Herramientas de modelado

Existen un sin número de herramientas para el modelado, a estas también se le conocen como herramientas CASE (Computer Aided Software Engineering). Se puede definir a las herramientas CASE como un conjunto de programas y ayudas que dan asistencia a los analistas, ingenieros de software y desarrolladores, durante todos los pasos del ciclo de vida de desarrollo de un software. CASE es también definido como el conjunto de métodos, utilidades y técnicas que facilitan el mejoramiento del ciclo de vida del desarrollo de sistemas de información, completamente o en alguna de sus fases. Entre las herramientas de diseño más utilizadas se encuentran:

- Rational Rose
- ER Studio
- Visual Paradigm

A continuación se hace referencia a la herramienta de modelado a utilizar.

Se selecciona al Visual Paradigm como herramienta CASE de modelado por su integración a UML, además por ser multiplataforma, ser amigable su uso y poseer interoperabilidad con otras aplicaciones e integración con distintos Ambientes de Desarrollo Integrado (IDE). Esta herramienta permite dibujar todos los tipos de diagramas de clases, código inverso, generar código desde los diagramas y generar documentación, proporcionando abundantes tutoriales de UML. Cabe destacar su robustez, usabilidad y portabilidad. Además el Visual Paradigm, proporciona el diseño de ingeniería inversa, código a modelo, código a diagrama, permitiendo la realización de diagramas de flujo de datos, generación de bases de datos y la ingeniería inversa de bases de datos, siendo un potente generador de informes.

Es una herramienta visual de ingeniería de software para el modelado, brinda una colección de menús, barras de herramientas y ventanas que forman el área de trabajo, lo cual permite crear diferentes tipos de diagramas en un ambiente completamente visual. Acelera el desarrollo de aplicaciones, ya que sirve de puente visual entre arquitectos, analistas y diseñadores de sistemas de información, haciendo el trabajo más fácil y dinámico.

Algunas de las particularidades que ofrece el Visual Paradigm son:

- Entorno de creación de diagramas para UML 2.1.

- Diseño centrado en casos de uso y enfocado al negocio que generan un software de mayor calidad.
- Uso de un lenguaje estándar común a todo el equipo de desarrollo que facilita la comunicación.
- Disponibilidad de múltiples versiones, para cada necesidad.
- Disponibilidad en múltiples plataformas.
- Sincronización entre Diagramas de Entidad Relación y Diagramas de Clases.
- Provee soporte para la generación de código y la ingeniería inversa para Java. Se integra con algunas herramientas de este lenguaje, como: Eclipse, Netbeans, Jbuilder, Oracle, entre otras.

Al ser seleccionada la herramienta de modelado se da paso a la elección del sistema gestor de base de datos a utilizar en el desarrollo de esta solución.

1.9 Sistema Gestor de Bases de Datos

Se define como un Sistema Gestor de Bases de Datos (SGBD), también llamado DBMS (Data Base Management System) a la colección de datos relacionados entre sí, estructurados y organizados, y un conjunto de programas que acceden y gestionan esos datos. También cabe destacar que son múltiples las compañías que han marcado hitos en este sentido. Entre los SGBD más conocidos se encuentran:

- Oracle
- MySQL
- PostgreSQL

Debido a que la Oficina Nacional de Estadísticas es una de las entidades del país que está migrando hacia la independencia tecnológica, se seleccionó como mecanismo de almacenamiento el Sistema Gestor de Base de Datos PostgreSQL. Esta decisión ha sido previamente colegiada y aceptada por parte del cliente final debido a que dentro de sus políticas de migración se encuentran las de llevar a todas sus bases de datos hacia dicha plataforma. La versión que se utilizará es la 8.4 por ser lo suficientemente estable y segura. Este SGBD posee las siguientes características:

- La cantidad máxima de BD que permite es ilimitada.
- El máximo de registros por tablas es ilimitado.
- El máximo de índices por tablas es ilimitado.

- Se permite programar funciones en lenguajes como: Pl/pgsql, Pl/java, Pl/Perl, Pl/pyton, TCL, Pl/PHP, C, C++, Ruby, entre otros.
- Alta capacidad de almacenar información, con una alta velocidad de respuesta ante consultas complejas y/o extensas.
- Incorpora una estructura de datos y arreglos.
- Permite la declaración de funciones propias, así como la definición de disparadores.
- Soporta el uso de índices, reglas y vistas.
- Incluye herencia entre tablas (aunque no entre objetos, ya que no existen), por lo que a este gestor de bases de datos se le incluye entre los gestores objeto-relacionales.
- Permite la gestión de diferentes usuarios, como también los permisos asignados a cada uno de ellos.

Al ser seleccionado el PostgreSQL como el SGBD es necesario utilizar una herramienta que gestione el SGBD PostgreSQL 8.4, por sus características se ha seleccionado el PgAdmin III 1.10, siendo la más completa y popular con licencia Open Source. En el siguiente epígrafe se abordará al respecto.

1.9.1 Herramienta para gestionar Bases de Datos. PgAdmin III

El PgAdmin III es una interfaz de administración para gestionar bases de datos PostgreSQL. Es multiplataforma y puede funcionar bajo GNU / Linux, FreeBSD, Mac, Windows y Solaris. Es capaz de gestionar versiones a partir de PostgreSQL 7.3 ejecutándose en cualquier plataforma, así como versiones comerciales de PostgreSQL como Pervasive Postgres, EnterpriseDB, Mammoth Replicator y SRA PowerGres. Incluye funcionalidades para responder a las necesidades de todos los usuarios, desde escribir consultas SQL simples hasta desarrollar bases de datos complejas. La interfaz gráfica soporta todas las características de PostgreSQL y facilita enormemente la administración. Otras de sus características son:

- Con este gestor de bases de datos es posible añadir nuevos servidores, acceder a un completo explorador de objetos (tablas, usuarios, dominios, funciones, grupos, esquemas etc.), ver estadísticas de tablas, índices, iniciar y detener los servidores, crear reportes, scripts, opciones de mantenimiento, respaldo y backup de los datos, ver el estado de un servidor, también contiene un asistente para configurar permisos, completa documentación y manual del programa, y muchas otras opciones para gestionar las BD.

- Utiliza procedimientos almacenados.
- Muy rápido para la visualización y entrada de datos.
- Incluye una interfaz gráfica de administración, una herramienta para el trabajo con SQL, un editor de código de procedimientos y funciones y un agente para lanzar scripts programados.

Al ser definido el SGBD que se utilizará en el proceso de desarrollo de la tecnología de almacenamiento de datos seleccionada, se da paso a la selección del modo de almacenamiento de datos, el cual se abordará a continuación.

1.10 Modos de almacenamiento de datos

Existen tres modelos para el proceso analítico en línea (OLAP, por sus siglas en inglés) de la información ROLAP, MOLAP y HOLAP. El proceso de análisis se realiza de igual forma lo que varía en uno y otro caso es la metodología de almacenamiento. La forma de almacenamiento es crítica para garantizar la velocidad de recuperación de la información, las zonas de ubicación de las agregaciones y el procesamiento de los datos en general.

1.10.1 Proceso Analítico Relacional (ROLAP)

La arquitectura ROLAP, accede a los datos almacenados en un Data Warehouse para proporcionar los análisis OLAP. La premisa de los sistemas ROLAP es que las capacidades OLAP se soportan mejor contra las bases de datos relacionales.

El sistema ROLAP utiliza una arquitectura de tres niveles. La base de datos relacional maneja los requerimientos de almacenamiento de datos, y el motor ROLAP proporciona la funcionalidad analítica. El nivel de base de datos usa bases de datos relacionales para el manejo, acceso y obtención del dato. El nivel de aplicación es el motor que ejecuta las consultas multidimensionales de los usuarios. El motor R-OLAP se integra con niveles de presentación, a través de los cuales los usuarios realizan los análisis OLAP. [4]

1.10.2 Proceso Analítico Multidimensional (MOLAP)

La arquitectura MOLAP usa bases de datos multidimensionales para proporcionar el análisis, su principal premisa es que el OLAP está mejor implantado almacenando los datos multidimensionalmente.

Un sistema MOLAP usa una base de datos propietaria multidimensional, en la que la información se almacena multidimensionalmente, para ser visualizada en varias dimensiones de análisis, a diferencia del ROLAP. El sistema MOLAP utiliza una arquitectura de dos niveles: las bases de datos multidimensionales y el motor analítico. La base de datos multidimensional es la encargada del manejo, acceso y obtención del dato.

Las estructuras de almacenamiento son grandes arreglos dimensionales que son una copia de la fuente de datos y persisten físicamente en la misma estación de trabajo donde está instalada la herramienta Data Warehousing. Esto provoca que el acceso a la información almacenada se realice de forma más rápida y efectiva utilizándose en depósito donde el tiempo en la velocidad de respuesta es crítico. [4]

1.10.3 Proceso Analítico Híbrido (HOLAP)

El modo de almacenamiento HOLAP (Hybrid Online Analytical Process, por sus siglas en inglés), como su nombre lo indica es un híbrido el cual combina las arquitecturas ROLAP y MOLAP para brindar una solución con las mejores características de ambas. Permite almacenar una parte de los datos como en un sistema MOLAP y el resto como en uno ROLAP.

Cada uno de los tipos de OLAP tiene beneficios en dependencia del problema en que se aplique. MOLAP requiere de menor espacio de almacenamiento y es más rápido calculando las agregaciones y devolviendo las respuestas, aunque se recomienda emplear para pequeños volúmenes de datos. ROLAP es considerado el más escalable, pero es más lento en el pre procesamiento y rendimiento de las consultas. HOLAP es rápido en el pre procesamiento y rendimiento de las consultas, aunque más lento que MOLAP y es escalable. HOLAP es ideal para grandes fuentes de datos.

1.10.4 Comparación entre los modos de almacenamiento de los datos ROLAP y MOLAP

ROLAP	MOLAP
Soportan análisis OLAP contra grandes volúmenes de datos elementales.	Se comportan razonablemente en volúmenes más reducidos (menos de 5 Gb).
Pueden crecer hasta un gran número de dimensiones.	Generalmente son adecuados para diez o menos dimensiones.

Las herramientas R-OLAP tienen menor rendimiento que las herramientas M-OLAP.	Las herramientas M-OLAP tradicionalmente tienen dificultades para consultar con modelos con dimensiones muy altas (del orden de millones de miembros).
Todos los datos de acceso están en la bodega almacenados.	Resúmenes de acceso a datos detallados en BD multidimensionales.
Vistas multidimensionales en la capa de presentación.	Tecnología propietaria para almacenar las vistas multidimensionales en arreglos no en tablas. Matriz de alta velocidad para la recuperación de los datos.
Las implementaciones ROLAP son más escalables y son frecuentemente atractivas a los clientes debido a que aprovechan las inversiones en tecnologías de bases de datos relacionales ya existentes en la organización.	En las implementaciones MOLAP el acceso a la información almacenada se realiza de forma más rápida y efectiva utilizándose un depósito donde el tiempo en la velocidad de respuesta es crítico. Normalmente se desempeñan mejor que la tecnología ROLAP, pero tienen problemas de escalabilidad.

Tabla1: Comparación ROLAP y MOLAP

La selección de uno u otro modelo depende de cuán importante sea el rendimiento de las consultas para los usuarios y de la tecnología disponible a utilizar. En el modelo ROLAP la respuesta a las consultas y el tiempo de procesamiento suelen ser más lentos que con los modos de almacenamiento MOLAP o HOLAP. No obstante, ROLAP permite a los usuarios ver los datos en tiempo real y ahorrar espacio de almacenamiento al trabajar con conjuntos de datos grandes a los que no se suele consultar con frecuencia, como datos puramente históricos.

1.10.5 Justificación del modo de almacenamiento a utilizar

Para el desarrollo del mercado de datos se escogió como modo de almacenamiento de datos al Procesamiento Analítico Relacional (ROLAP) por las ventajas que el mismo significa en este tipo de entornos, donde el detalle y la consolidación de los distintos conceptos relacionados condicionan el funcionamiento del sistema. Además porque posee un ambiente conocido y

disponibilidad de herramientas para su desarrollo lo cual facilita su puesta en práctica para desarrollar soluciones. También el tipo de gestor de BD a utilizar exige que sea este el modelo de almacenamiento a utilizar.

Conclusiones

En este capítulo se realizó una descripción acerca de las tecnologías de almacenamiento de datos, abordando los conceptos esenciales para la realización del mercado de datos en el área de Salud Pública y Asistencia Social. También se investigó acerca del desarrollo de las tecnologías de almacenamiento de datos existentes a nivel mundial.

Después de realizado el estudio del arte se seleccionó como SGBD el PostgreSQL versión 8.4 para la construcción del mercado de datos, utilizando para su diseño y modelado dimensional el Visual Paradigm for UML versión 6.4 por las características antes mencionadas. Para el proceso de ETL se seleccionó el Pentaho Data Integration versión 4.1, más conocido como Kettle, para el proceso de BI, el Pentaho BI Server versión 3.6 y el Mondrian Schema Workbench versión 3.2. Como metodología se eligió la Metodología de Kimball, pero ajustándola a las necesidades de la UCI.

CAPITULO 2: ANÁLISIS Y DISEÑO DEL MERCADO DE DATOS.

Introducción

Para la elaboración del mercado de datos Salud Pública y Asistencia Social son muy necesarias las fases de análisis y diseño para tener un control acerca de las necesidades de los usuarios, para lo cual se realiza la captura de los requerimientos, funcionales, no funcionales y de información; creando una guía para los desarrolladores en la fase de implementación con el objetivo de lograr un producto con la calidad requerida. En este capítulo se fijan las bases para la correcta realización del proyecto en cuestión.

2.1 Estudio preliminar del negocio

La Oficina Nacional de Estadísticas es una entidad creada para gestionar grandes volúmenes de información. Su misión fundamental es garantizar la producción de estadísticas de calidad a través del Sistema Estadístico Nacional (SEN) ejerciendo una eficaz captación de las cifras económicas y sociales así como su adecuada difusión de acuerdo con las necesidades de la economía y las demás esferas del país en información estadística. En el Sistema Estadístico Nacional participan coordinadamente los órganos, organismos, instituciones y entidades que elaboran estadísticas en Cuba y está integrado por 3 subsistemas:

- ✓ Sistema de Información Estadística Nacional (SIEN): Está dirigido a la elaboración de estadísticas destinadas a satisfacer los requerimientos informativos de los más altos niveles del Estado y el Gobierno además de informar a los organismos internacionales. La ONE controla la información relacionada con la población y las entidades, siendo la productora de estadísticas a través de su Red.
- ✓ Sistema de Información Estadística Complementaria (SIEC): Está dirigido a la elaboración de estadísticas destinadas a satisfacer los requerimientos informativos de los organismos a los efectos del control administrativo de sus entidades. La ONE revisa y ejerce su control metodológico, además de que controla la información de las entidades subordinadas y no subordinadas.
- ✓ Sistema de Información Estadística Territorial (SIET): Dirigido a la elaboración de estadísticas destinadas a satisfacer los requerimientos informativos del Gobierno en el territorio y otros actores locales.

Las funciones y atribuciones de la ONE son:

1. Organizar y aprobar la producción de las estadísticas centralizadas y territoriales.
2. Dirigir metodológicamente la actividad estadística de los órganos, organismos, instituciones y entidades estatales y velar y dictaminar acerca del funcionamiento de las estadísticas complementarias.
3. Definir las atribuciones y responsabilidades de otros productores de estadísticas de interés nacional y su lugar en el sistema estadístico del país.
4. Aprobar las normas metodológicas y de clasificación que se utilizan en las estadísticas centralizadas, territoriales y complementarias.
5. Identificar las unidades de observación estadística y captar, a través de la red territorial, la información correspondiente a las estadísticas centralizadas.
6. Dirigir los procesos y ejecutar, según corresponda, los censos económicos y de población y encuestas económicas o sociales de carácter nacional. Aprobar la realización de este tipo de investigaciones estadísticas en el país.
7. Llevar los registros estatales de empresas, unidades presupuestadas y otras entidades.
8. Supervisar el trabajo estadístico de organismos y entidades, organizar la auditoría y comprobación estadísticas velando por la autenticidad de la información.
9. Centralizar y emitir la estadística oficial del país.

El flujo de información de la ONE es a través de una red de transmisión de datos por todo el país. Se transmiten más de 484 millones de datos al año a través de las oficinas en los municipios donde se digitan y revisan las informaciones. En los territorios (provincias) se revisan e informan a la oficina nacional donde se procesan y emiten las informaciones.

La información estadística de la salud en la ONE proviene del Ministerio de Salud Pública este al igual que otros organismos le brinda la información de todos los servicios de salud y asistencia social desde el triunfo de la revolución hasta la actualidad, la ONE procesa esta información a través del SIEC, cuyo funcionamiento fue explicado con anterioridad.

2.2 Necesidades de información

Para el desarrollo del análisis en el proceso del negocio, es necesario conocer que es lo que buscan y necesitan los clientes o usuarios, por esta razón en esta etapa se realizan reuniones con los especialistas de la ONE para analizar la información referente al área de Salud Pública y Asistencia Social, identificando cuáles son las necesidades que tiene la organización. Las necesidades de información se recogen realizando un estudio donde se especifica cuáles son los objetivos de la organización, redactándose en forma de requisitos para luego analizar los indicadores que serán tomados para el desarrollo del mercado de datos.

2.3 Requisitos funcionales

Los requisitos funcionales son capacidades o condiciones que el sistema debe cumplir para dar respuesta a los requisitos de información. Un requisito funcional define el comportamiento interno del software: cálculos, detalles técnicos, manipulación de datos y otras funcionalidades específicas que muestran cómo los casos de uso serán llevados a la práctica. Los mismos son identificados a partir de las necesidades de los usuarios y las reglas del negocio. A continuación se mostrarán los requisitos funcionales que fueron identificados.

RF1. Realizar la extracción, transformación y carga de la información, de los ficheros Excel.

Administrar usuario

RF2. Insertar usuario.

RF3. Eliminar usuario.

Administrar rol.

RF4. Insertar rol.

RF5. Eliminar rol.

Gestionar reportes OLAP.

RF6. Crear reportes OLAP.

RF7. Modificar reportes OLAP.

RF8. Eliminar reporte OLAP

RF9. Autenticar usuario.

2.4 Requisitos no funcionales

Los requisitos no funcionales son propiedades o cualidades que el producto debe tener; son aquellos que tienen que ver con características que de una u otra forma pueden limitar el sistema,

como por ejemplo la seguridad, el rendimiento, la fiabilidad entre otros. A continuación se mostrarán una selección de los requisitos no funcionales identificados, basados en las características del sistema que se va a realizar. Para ver detalladamente cada uno de estos requisitos se puede hacer referencia al artefacto **Especificación de requisitos**.

Hardware

- Computadora/Procesador: con un procesador Pentium.
- Memoria: 512 MB de RAM ó más (1 GB ideal).
- Disco duro: 10 GB (mínimo).
- Hardware (Conectividad) de Red: para el acceso a los datos en el Datawarehouse, para la emisión de los datos y los servicios a las estaciones clientes.

Software

Instalar en las estaciones de trabajo el software necesario para el correcto funcionamiento del sistema. Las configuraciones de software de las máquinas clientes deben contar al menos con los siguientes elementos:

- Firefox 2.0 o superior.
- Schema Workbench 3.2.1 en caso de que un usuario capacitado requiera la construcción de esquemas multidimensionales para el diseño de nuevos reportes.
- Pentaho Data Integration para los procesos de ETL.

2.5 Requisitos de información

Los requisitos de información describen qué información debe almacenar el sistema para satisfacer las necesidades de clientes y usuarios; es la información que debe estar disponible en el almacén para su consulta. Identifican los conceptos relevantes sobre los que se debe almacenar información y los datos específicos que son de interés. Además constituyen la entrada fundamental para el proceso de Inteligencia de Negocio y para futuros reportes.

A continuación se mostrarán los requisitos de información que fueron identificados en el proceso de análisis, agrupándolos por dominios informativos.

Causas de muerte

RI1. Obtener anualmente la cantidad de las principales causas de muerte de todas las edades.

RI2. Obtener anualmente las tasas de las principales causas de muerte.

RI3. Obtener anualmente la cantidad de las principales causas de muerte en niños de 1 a 4 años de edad.

RI4. Obtener anualmente la cantidad de las principales causas de muerte en niños menores de un año de edad.

Consultas

RI5. Obtener anualmente la cantidad de consultas por el tipo de consultas.

RI6. Obtener anualmente la cantidad de consultas por habitantes.

RI7. Obtener anualmente la cantidad de consultas externas de asistencia médica por el tipo de unidad de servicio.

RI8. Obtener anualmente la cantidad de consultas en cuerpos de guardia de asistencia médica por el tipo de unidad de servicio.

Camas

RI9. Obtener anualmente el promedio de camas reales de asistencia médica por el tipo de unidad de servicio.

RI10. Obtener anualmente la cantidad de dotación normal de camas por el tipo de unidad de servicio.

RI11. Obtener anualmente la cantidad de dotación normal de camas por el tipo de unidad de servicio y por provincias.

Donantes de sangre

RI12. Obtener anualmente la cantidad de donantes de sangre por provincias.

RI13. Obtener anualmente la cantidad de donantes de sangre.

RI14. Obtener anualmente la cantidad de sangre útil.

RI15. Obtener anualmente la cantidad de sangre útil por provincias.

Enfermedades

RI16. Obtener anualmente la cantidad de incidencias por enfermedades de declaración obligatoria.

RI17. Obtener anualmente la tasa de las enfermedades de declaración obligatoria por habitantes.

Inmunizaciones

RI18. Obtener anualmente la cantidad de inmunizaciones por tipo de vacunas.

Ingresos

RI19. Obtener anualmente la cantidad de ingresos en hogares maternos por provincias.

RI20. Obtener anualmente el total de ingresos en hogares maternos.

RI21. Obtener anualmente la cantidad de ingresos por unidad de servicio.

RI22. Obtener anualmente la tasa de gestantes que ingresaron en hogares maternos por cada 100 nacidos vivos.

RI23. Obtener anualmente la cantidad de gestantes (mujeres en estado de gestación o mujeres embarazadas) que ingresaron en hogares maternos.

Índices

RI24. Obtener anualmente el índice de embarazadas que reciben atención en hogares maternos por provincias.

RI25. Obtener anualmente el total del índice de embarazadas que reciben atención en hogares maternos.

RI26. Obtener anualmente el índice de bajo peso al nacer por provincia.

RI27. Obtener anualmente el total del índice de bajo peso al nacer.

Personal facultativo

RI28. Obtener anualmente la cantidad de personal facultativo del Ministerio de Salud Pública por el tipo de personal.

RI29. Obtener anualmente la cantidad de personal facultativo del Ministerio de Salud Pública por provincias.

RI30. Obtener anualmente la cantidad de habitantes que existen por médicos y estomatólogos.

Tasa de mortalidad

RI31. Obtener anualmente la tasa de mortalidad materna según las causas.

RI32. Obtener anualmente el total de la tasa de mortalidad materna por 100 mil nacidos vivos.

RI33. Obtener anualmente la tasa de mortalidad infantil por provincias.

RI34. Obtener anualmente el total de la tasa de mortalidad infantil por mil nacidos vivos.

RI35. Obtener anualmente la tasa de mortalidad de los niños menores de 5 años de edad por provincias.

RI36. Obtener anualmente el total de la tasa de mortalidad de los niños menores de 5 años de edad por mil nacidos vivos.

RI37. Obtener anualmente la tasa de mortalidad materna por provincia.

Unidades de servicios

RI38. Obtener anualmente la cantidad de unidades de servicio del Ministerio de Salud Pública.

RI39. Obtener anualmente la cantidad de unidades de servicio del Ministerio de Salud Pública por provincias.

RI40. Obtener anualmente la cantidad de hospitales y clínicas según el rango de camas.

2.6 Reglas del negocio

En la fase de análisis una de las tareas más importante son las reglas de negocio, las mismas describen políticas que deben cumplirse o condiciones que deben satisfacerse, por lo que regulan algún aspecto del negocio. También dan la oportunidad de definir cómo es que se van a calcular las medidas del almacén de datos.

A continuación se mostrarán las reglas de negocios identificadas en el mercado de datos Salud Pública y Asistencia Social.

En el fichero Excel **Salud Series 1958-2009** en la hoja:

- **19.1** El total del personal facultativo es la suma de todas sus temáticas.
- **19.4** El total de unidades de servicio es la suma de todas sus temáticas y el total de cada una de las temáticas es la suma de sus subtemáticas.
- **19.6** El total de dotación normal de camas en unidades de servicio es la suma de cada una de sus temáticas.
- **19.8** El total de camas en hospitales y clínicas es la suma de las cantidades que existen en los diferentes rangos.
- **19.9** El total del promedio de camas reales de asistencia médica es la suma de cada una de sus temáticas.

- **19.10** El total de ingresos en unidades de servicio es la suma de cada una de sus temáticas.
- **19.11** El total de consultas médicas es la suma de las consultas externas más las consultas en cuerpo de guardia.
- **19.12** El total de consultas externas de asistencia médica es la suma de cada una de sus temáticas.
- **19.13** El total de consultas en cuerpo de guardia de asistencia médica es la suma de cada una de sus temáticas.
- **19.26** En la tabla ingresos en hogares maternos el total es la suma de la cantidad de los ingresos (anualmente).
- Los identificadores de las tablas no pueden tomar valores repetidos.
- Las medidas de los veintidós hechos deben poseer valores mayores o iguales a cero, nunca un número negativo.

2.7 Modelo de Casos de Uso del Sistema

Los Caso de Uso del Sistema (CUS) se utilizan para capturar los requisitos del sistema, los mismos proporcionan la interacción del o los usuarios con el sistema para lograr un objetivo específico. Estos se describen como un conjunto de secuencias.

En el desarrollo de la investigación se identificaron 11 casos de uso informativos y 6 casos de uso funcionales los cuales representarán los requisitos de información y los requisitos funcionales respectivamente. En el siguiente epígrafe se mostrarán cuáles fueron los actores identificados en el desarrollo del análisis.

Para ver más detallada la información de los casos de uso identificados y sus respectivas descripciones, consultar el expediente de proyecto, en el artefacto **Modelo de Casos de Uso**.

2.7.1 Actores del sistema

Actor	Descripción
Analista	Es el usuario que analiza y consulta la información de los diferentes indicadores de Salud Pública y Asistencia Social.
Administrador del Sistema	Es el responsable de la administración de los usuarios,

	roles y de los reportes OLAP.
Administrador ETL	Es el responsable de la extracción, transformación y carga de los datos.

Tabla 2: Descripciones de los actores.

2.7.2 Casos de Uso del Sistema

Los CUS son los encargados de que los usuarios puedan obtener la información que desean mediante la representación visual de la misma. A continuación se dará una breve descripción de los CUS identificados, tanto los de información como los funcionales.

Los Casos de Uso Informativos (CUI) responden a los requisitos de información, los cuales dan respuesta a las necesidades de los usuarios de obtener o consultar información. A continuación se describirán brevemente los CUI identificados.

Casos de Uso de Información	Descripción
Mostrar información de las causas de muerte	Visualiza los reportes de los indicadores referentes a las causas de muerte.
Mostrar información de las consultas	Visualiza los reportes de los indicadores referentes a las consultas.
Mostrar información de las enfermedades de declaración obligatoria	Visualiza los reportes de los indicadores referentes a las enfermedades de declaración obligatoria.
Mostrar información de las inmunizaciones	Visualiza los reportes de los indicadores referentes a las inmunizaciones.
Mostrar información de las unidades de servicio	Visualiza los reportes de los indicadores referentes a las unidades de servicio.
Mostrar información del personal facultativo	Visualiza los reportes de los indicadores referentes al personal facultativo.
Mostrar información de las camas	Visualiza los reportes de los indicadores referentes a las camas.
Mostrar información de los ingresos	Visualiza los reportes de los indicadores referentes a los ingresos.
Mostrar información de los índices	Visualiza los reportes de los indicadores referentes a los índices.

Mostrar información de las donaciones de sangre	Visualiza los reportes de los indicadores referentes a las donaciones de sangre.
Mostrar información de los tasas de mortalidad	Visualiza los reportes de los indicadores referentes a las tasas de mortalidad.

Tabla 3: Descripción de los CUI.

Los Casos de Uso Funcionales (CUF) describen los requerimientos funcionales del sistema. A continuación se describirán brevemente los CUF identificados.

Casos de Uso Funcionales	Descripción
Extraer, Transformar y Cargar datos Salud Pública y Asistencia Social	Realiza la extracción, transformación y carga de los datos necesarios para la construcción del mercado de datos del fichero fuente.
Administrar usuario	Inserta y elimina los usuarios que interactúan con el sistema.
Administrar reportes OLAP	Crea, modifica y elimina los reportes OLAP que se visualizan.
Administrar rol	Inserta y elimina los roles que existan en el sistema.
Autenticar usuario	Realiza la autenticación de los usuarios en el sistema.
Visualizar reporte	Modifica la visualización de los reportes, en dependencia de cómo el usuario desee analizar el mismo.

Tabla 4: Descripción de los CUF.

2.7.3 Diagrama de CUS

Un diagrama de caso de uso del sistema representa gráficamente a los procesos y su interacción con los actores.



Fig1: Diagrama de CUS

2.8 Desarrollo de la matriz BUS

En la matriz BUS o matriz dimensional se representa la información como matrices multidimensionales o cuadros de múltiples entradas denominados cubos. En la misma se refleja la relación que existe entre las tablas hechos y sus tablas dimensiones asociadas. A los ejes de la matriz se les llama dimensiones y representan los criterios de análisis, y a los datos almacenados en la matriz se los llama medidas y representan los indicadores o valores a analizar. A continuación se muestra como quedó la matriz BUS realizada al mercado de datos a partir de las dimensiones y hechos identificados, los mismos se muestran en el siguiente epígrafe.

Matriz Dimensional													
	Dimensiones												
Hechos	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13
H1	x									x			x
H2							x			x		x	x
H3							x		x	x		x	
H4						x				x			x
H5							x			x		x	x
H6									x	x			
H7			x							x			x
H8					x					x		x	x
H9									x	x			
H10							x			x		x	x
H11									x	x			
H12										x	x		
H13		x								x			x
H14								x	x	x		x	
H15								x		x		x	x
H16									x	x			
H17										x			x
H18										x			x
H19										x			x

H20							X			X		X	X
H21							X		X	X		X	
H22				X						X		X	X

Tabla 2: Matriz Dimensional

2.9 Identificación de las dimensiones, hechos y medidas.

Luego de realizar el análisis al fichero Excel en el área de Salud Pública y Asistencia Social se procede a identificar las dimensiones, hechos y medidas que tendrá el mercado de datos para a partir de estos poder realizar el modelo de datos.

Medidas

La medida es un término del modelado dimensional que constituyen el que analizar, también se refiere a los valores o datos cuantificables, normalmente numéricos que miden algunos aspectos del negocio. Una medida se asigna a una columna de una tabla de hechos.

Las medidas numéricas definidas se almacenarán en las tablas de hechos, los extraídos de cada temática aparecen a continuación en sus respectivas tablas de hechos.

Hechos

Llamamos evento o **Hecho** a una operación que se realiza en el negocio en un tiempo determinado. Los hechos son objeto de análisis para la toma de decisiones y están caracterizados por medidas numéricas. A continuación se mostrarán cada uno de los hechos identificados.

H1. hech_causas_muerte: En esta tabla se encuentra toda la información referente a las causas de muerte.

hech_causas_muerte		
+#dim temporal_anno_id	int4	Nullable = false
+#dim causas_defuncion_id	int4	Nullable = false
+#dim dpa_id	int4	Nullable = false
tasa_prin_cm_por_100_mil_hab	float4	Nullable = true
cant_muertes_todas_edades_unidades	int4	Nullable = true
cant_muertes_ninnos_menos_1año_unidades	int4	Nullable = true
cant_muertes_ninnos_1a4años_unidades	int4	Nullable = true

H2. hech_camaz: En esta tabla se encuentra toda la información referente a la cantidad de camas a nivel nacional.

hech_camas		
+#dim_dpa_id	int4	Nullable = false
+#dim_temporal_anno_id	int4	Nullable = false
+#dim_unidades_servicio_id	int4	Nullable = false
+#dim_um_id	int4	Nullable = false
cant_dotacion_normal_camas	int4	Nullable = true
promedio_camas_reales	int4	Nullable = true

H3. hech_camas_prov: En esta tabla se encuentra toda la información referente a la cantidad de camas por provincia.

hech_camas_prov		
+#dim_provincia_id	int4	Nullable = false
+#dim_um_id	int4	Nullable = false
+#dim_unidades_servicio_id	int4	Nullable = false
+#dim_temporal_anno_id	int4	Nullable = false
cant_dotacion_normal_camas	int4	Nullable = true

H4. hech_consultas: En esta tabla se encuentra toda la información referente a las consultas.

hech_consultas		
+#dim_consultas_id	int4	Nullable = false
+#dim_temporal_anno_id	int4	Nullable = false
+#dim_dpa_id	int4	Nullable = false
cant_cons_miles	int4	Nullable = true
cant_cons_por_hab	float4	Nullable = true

H5. hech_consultas_us: En esta tabla se encuentra toda la información referente a la cantidad de consultas a nivel nacional.

hech_consultas_us		
+#dim_unidades_servicio_id	int4	Nullable = false
+#dim_um_id	int4	Nullable = false
+#dim_temporal_anno_id	int4	Nullable = false
+#dim_dpa_id	int4	Nullable = false
cant_cons_externas	int4	Nullable = true
cant_cons_cuerpo_guardia	int4	Nullable = true

H6. hech_donaciones_sangre: En esta tabla se encuentra toda la información referente a las donaciones de sangre.

hech_donaciones_sangre		
+#dim_temporal_anno_id	int4	Nullable = false
+#dim_provincia_id	int4	Nullable = false
cant_donantes_sangre_unidades	int4	Nullable = true
cant_sangre_util_unidades	int4	Nullable = true

H7. hech_enfermedades: En esta tabla se encuentra toda la información referente a las enfermedades que son de declaración obligatoria.

hech_enfermedades		
+#dim_enf_dec_obligatoria_id	int4	Nullable = false
+#dim_temporal_anno_id	int4	Nullable = false
+#dim_dpa_id	int4	Nullable = false
cant_incidentes_enfermedades_unidades	int4	Nullable = true
tasa_enfermedades_por_100_mil_hab	float4	Nullable = true

H8. hech_hosp_clinicas: En esta tabla se encuentra toda la información referente a los hospitales y clínicas según el rango de camas

hech_hosp_clinicas		
+#dim_rango_cama_id	int4	Nullable = false
+#dim_temporal_anno_id	int4	Nullable = false
+#dim_dpa_id	int4	Nullable = false
+#dim_um_id	int4	Nullable = false
cant_hyc_según_rango_camás	int4	Nullable = true

H9. hech_indices: En esta tabla se encuentra toda la información referente a los índices tanto de embarazadas como de bajo peso al nacer.

hech_indices		
+#dim_provincia_id	int4	Nullable = false
+#dim_temporal_anno_id	int4	Nullable = false
indice_embarazadas_reciben_atencion_hm_porcentaje	float4	Nullable = true
indice_bajo_peso_nacer_por_100_nv	float4	Nullable = true

H10. hech_ingresos: En esta tabla se encuentra toda la información referente a los ingresos en las diferentes unidades de servicio.

hech_ingresos		
+#dim_unidades_servicio_id	int4	Nullable = false
+#dim_temporal_anno_id	int4	Nullable = false
+#dim_dpa_id	int4	Nullable = false
+#dim_um_id	int4	Nullable = false
cant_ingresos_us	int4	Nullable = true

H11. hech_ingresos_hm: En esta tabla se encuentra toda la información referente a los ingresos en hogares maternos a nivel provincial.

hech_ingresos_hm		
+#dim_provincia_id	int4	Nullable = false
+#dim_temporal_anno_id	int4	Nullable = false
cant_ingresos_hogares_maternos	int4	Nullable = true

H12. hech_ingresos_hogares_maternos: En esta tabla se encuentra toda la información referente a los ingresos en hogares maternos y la tasa.

hech_ingresos_hogares_maternos		
+#dim_temporal_anno_id	int4	Nullable = false
+#dim_pais_id	int4	Nullable = false
cant_ingresos_mujeres_estado_gestacion_unidades	int4	Nullable = true
tasa_ingresos_por_100_nv	float4	Nullable = true

H13. hech_mortalidad_materna: En esta tabla se encuentra toda la información referente a la mortalidad materna.

hech_mortalidad_materna		
+#dim_mortalidad_materna_id	int4	Nullable = false
+#dim_temporal_anno_id	int4	Nullable = false
+#dim_dpa_id	int4	Nullable = false
tasa_mm_por_100_mil_nv	float4	Nullable = true
total_tasa_mm_por_100_mil_nv	float4	Nullable = true

H14. hech_personal_facultativo: En esta tabla se encuentra toda la información referente al tipo de personal facultativo del Ministerio de Salud Pública.

hech_personal_facultativo		
+#dim_temporal_anno_id	int4	Nullable = false
+#dim_provincia_id	int4	Nullable = false
+#dim_personal_facultativo_id	int4	Nullable = false
+#dim_um_id	int4	Nullable = false
cant_personal_facultativo	int4	Nullable = true

H15. hech_otro_personal_facultativo: En esta tabla se encuentra toda la información referente a al personal facultativo a nivel nacional.

hech_otro_personal_facultativo		
+#dim_temporal_anno_id	int4	Nullable = false
+#dim_personal_facultativo_id	int4	Nullable = false
+#dim_dpa_id	int4	Nullable = false
+#dim_um_id	int4	Nullable = false
cant_personal_facultativo	int4	Nullable = true
cant_hab_por_personal	int4	Nullable = true

H16. hech_tasa_mortalidad: En esta tabla se encuentra toda la información referente a las tasas de mortalidad.

hech_tasa_mortalidad		
+#dim_provincia_id	int4	Nullable = false
+#dim_temporal_anno_id	int4	Nullable = false
tasa_m_ninnos_menosres_5annos_edad_por_1000_nv	float4	Nullable = true
tasa_mortalidad_materna_por_100_mil_nv	float4	Nullable = true
tasa_m_infantil_por_1000_nv	float4	Nullable = true

H17. hech_totales_indices: En esta tabla se encuentra toda la información referente a los índices totales.

hech_totales_indices		
+#dim_dpa_id	int4	Nullable = false
+#dim_temporal_anno_id	int4	Nullable = false
total_indice_embarazadas_reciben_atencion_hm_porcentaje	float4	Nullable = true
total_indice_bajo_peso_nacer_por_100_nv	float4	Nullable = true

H18. hech_totales_ingresos_hm: En esta tabla se encuentra toda la información referente a los totales de los ingresos en hogares maternos.

hech_totales_ingresos_hm		
+#dim_dpa_id	int4	Nullable = false
+#dim_temporal_anno_id	int4	Nullable = false
total_ingresos_hogares_maternos	int4	Nullable = true

H19. hech_totales_tasa_mortalidad: En esta tabla se encuentra toda la información referente a las donaciones de sangre.

hech_totales_tasa_mortalidad		
+#dim_dpa_id	int4	Nullable = false
+#dim_temporal_anno_id	int4	Nullable = false
total_tasa_m_infantil_por_1000_nv	float4	Nullable = true
total_tasa_m_ninnos_menores_5años_edad_por_1000_nv	float4	Nullable = true
total_tasa_mortalidad_materna_por_100_mil_nv	int4	Nullable = true

H20. hech_unidades_servicio: En esta tabla se encuentra toda la información referente a los tipos de unidades de servicio.

hech_unidades_servicio		
+#dim_unidades_servicio_id	int4	Nullable = false
+#dim_um_id	int4	Nullable = false
+#dim_temporal_anno_id	int4	Nullable = false
+#dim_dpa_id	int4	Nullable = false
cant_unidades_servicio	int4	Nullable = true

H21. hech_us_prov: En esta tabla se encuentra toda la información referente a las unidades de servicio por provincias.

hech_unidades_servicio_prov		
+#dim_unidades_servicio_id	int4	Nullable = false
+#dim_provincia_id	int4	Nullable = false
+#dim_temporal_anno_id	int4	Nullable = false
+#dim_um_id	int4	Nullable = false
cant_unidades_servicio	int4	Nullable = true

H22. hech_vacuna: En esta tabla se encuentra toda la información referente a los tipos de vacunas.

hech_vacuna		
+#dim_tipo_vacuna_id	int4	Nullable = false
+#dim_temporal_anno_id	int4	Nullable = false
+#dim_dpa_id	int4	Nullable = false
+#dim_um_id	int4	Nullable = false
cant_inmunizaciones	int4	Nullable = true

Dimensiones

Una **Dimensión** es una característica de un hecho que permite su análisis posterior, en el proceso de toma de decisiones. A continuación se describirán cada una de las dimensiones que se obtuvieron.

D1. Dimensión Causas de defunción: Dimensión que maneja la información de las causas de muerte de manera general para todas las edades.

dim_causas_defuncion		
+dim_causas_defuncion_id	int4	Nullable = false
clasificacion	varchar(50)	Nullable = true
tipo_causa	varchar(80)	Nullable = true

D2. Dimensión Mortalidad materna: Dimensión que maneja información de la mortalidad materna de forma general independientemente de su tipo de causa.

dim_mortalidad_materna		
+dim_mortalidad_materna_id	int4	Nullable = false
tipo_mortalidad	varchar(50)	Nullable = true
clasificacion	varchar(50)	Nullable = true

D3. Dimensión Enfermedades de declaración obligatoria: Dimensión que maneja información de las enfermedades de declaración obligatoria.

dim_enf_dec_obligatoria		
+dim_enf_dec_obligatoria_id	int4	Nullable = false
tipo_enfermedad	varchar(50)	Nullable = true

D4. Dimensión Tipo vacuna: Dimensión que maneja información de las inmunizaciones por tipo de vacuna.

dim_tipo_vacuna		
+dim_tipo_vacuna_id	int4	Nullable = false
tipo_vacuna	varchar(50)	Nullable = true

D5. Dimensión Rango de cama: Dimensión que maneja información del rango de camas.

dim_rango_cama		
+dim_rango_cama_id	int4	Nullable = false
rango_camas	varchar(50)	Nullable = true

D6. Dimensión Consultas: Dimensión que maneja información de los tipos de consultas de forma general, teniendo en cuenta que pueden ser médicas o estomatológicas.

dim_consultas		
+dim_consultas_id	int4	Nullable = false
clasificacion	varchar(50)	Nullable = true
tipo_consulta	varchar(50)	Nullable = true

D7. Dimensión Unidades de Servicio: Dimensión que maneja información de las unidades de servicio del MINSAP.

dim_unidades_servicio		
+dim_unidades_servicio_id	int4	Nullable = false
temático	varchar(50)	Nullable = true
subtemático	varchar(50)	Nullable = true
nombre	varchar(50)	Nullable = true

D8. Dimensión Personal Facultativo: Dimensión que maneja información del personal facultativo según su tipo.

dim_personal_facultativo		
+dim_personal_facultativo_id	int4	Nullable = false
temático	varchar(50)	Nullable = true
tipo_personal_facultativo	varchar(50)	Nullable = true

D9. Dimensión Provincia: Dimensión que maneja información de las provincias de la antigua División Político-Administrativa (DPA).

dim_provincia		
+dim_provincia_id	int4	Nullable = false
provincia_codigo	varchar(50)	Nullable = true
provincia_nombre	varchar(50)	Nullable = true
provincia_descripcion	varchar(50)	Nullable = true

D10. Dimensión Temporal año: Dimensión que maneja información de los años de los cuales se tiene información del MINSAP.

dim_temporal_anno		
+dim_temporal_anno_id	int4	Nullable = false
anno_codigo	varchar(50)	Nullable = true
anno_nombre	varchar(50)	Nullable = true
anno_numero	int4	Nullable = true

D11. Dimensión País: Dimensión que maneja información de los países.

dim_pais		
+dim_pais_id	int4	Nullable = false
nombre	varchar(50)	Nullable = true
descripcion	varchar(10)	Nullable = true
codigo_aladi	varchar(25)	Nullable = true
area_geog_nombre	varchar(10)	Nullable = true
iso_a1_codigo	varchar(25)	Nullable = true
iso_a2_codigo	varchar(25)	Nullable = true
iso_a3_codigo	varchar(25)	Nullable = true

D12. Dimensión unidad de medida: Dimensión que maneja información relacionada con las unidades de medida.

dim_um		
+dim_um_id	int4	Nullable = false
codigo	varchar(25)	Nullable = true
descripcion	varchar(100)	Nullable = true
nombre	varchar(20)	Nullable = true

D13. Dimensión DPA: Dimensión que maneja información de la división política administrativa.

dim_dpa		
+dim_dpa_id	int4	Nullable = false
provincia_codigo	char(2)	Nullable = false
provincia_nombre	varchar(25)	Nullable = true
provincia_descripcion	varchar(45)	Nullable = true
municipio_nombre	varchar(30)	Nullable = true
municipio_codigo	varchar(30)	Nullable = true
municipio_descripcion	varchar(45)	Nullable = true
municipio_extension	int4	Nullable = true
municipio_ext_cayosady	int4	Nullable = true
municipio_ext_tierrafirme	int4	Nullable = true
dpa_anno_nombre	varchar(4)	Nullable = true

2.10 Arquitectura de la información

La arquitectura de la información del mercado de datos para el área de Salud Pública y Asistencia Social debido al volumen y la variedad de indicadores estará conformada por un área de análisis la cual no es más que la agrupación de la información brindada por el MINSAP. Contará con 11 libros de trabajo los que se encargan de agrupar los reportes generados dentro del área de análisis teniendo en cuenta cada indicador. También tendrá los reportes que se generan en cada uno de estos libros de trabajo y que dan respuesta a las necesidades de información de los clientes. A continuación se relaciona el área de análisis Salud Pública y Asistencia Social con cada uno de sus libros de trabajo y los mismos con sus reportes y se muestra una imagen con un diseño del mapa de navegación. Para un mayor entendimiento de esta estructura consultar el artefacto **Arquitectura de la información**.

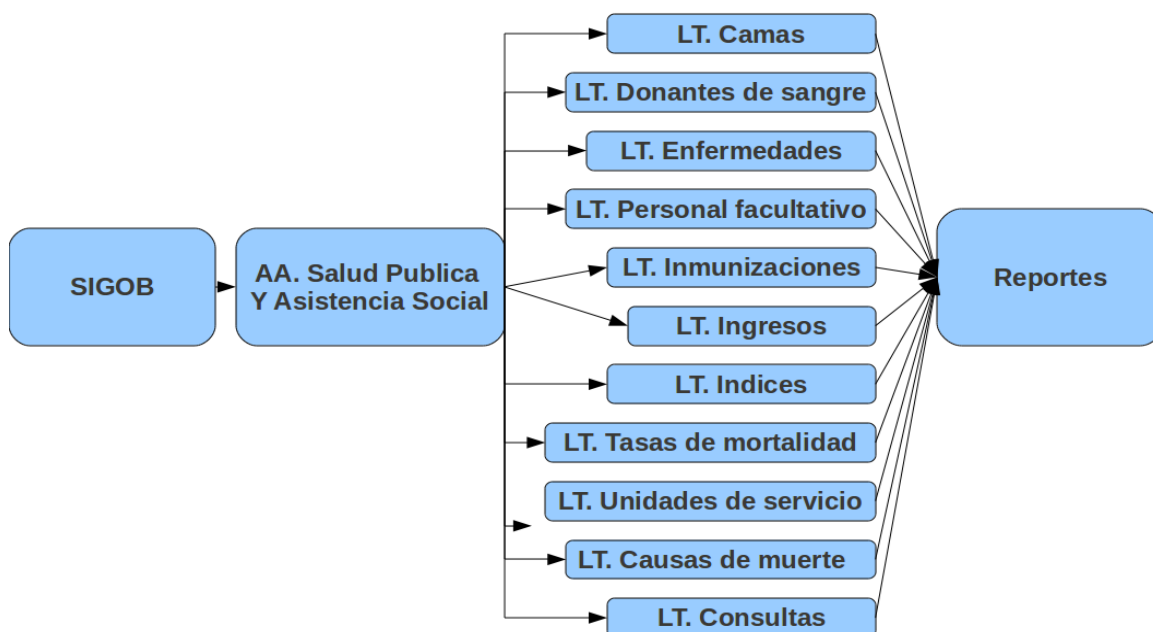


Fig3: Mapa de navegación

2.11 Política de seguridad

En el mercado de datos de Salud Pública y Asistencia Social las políticas de seguridad que se emplean están basadas mayormente en los niveles de acceso al sistema. Esta seguridad es

manejada fundamentalmente por los roles y permisos que los usuarios poseen en su interacción con la base de datos y la aplicación.

Seguridad en la base de datos

Para la interacción de los usuarios con la base de datos se definió el rol:

Actor	Permiso
Administrador ETL	Tiene total acceso a la base de dato del sistema.

Tabla3: Actores y permisos

Seguridad en la aplicación

Las aplicaciones desplegadas en el Servidor de Inteligencia de Negocios de Pentaho muestran un continuo incremento, así como los usuarios que tiene acceso a estas. Lo que trae como resultado que se especifiquen los siguientes roles:

Roles	Permisos
Administrador	Tiene acceso total a todas las áreas de análisis general.
Analista	Tiene acceso de solo lectura al área de análisis Salud Pública y asistencia Social. Visualiza los reportes.

Tabla4: Roles y permisos

Elemento de la aplicación	Roles con acceso
AA General	1. Administrador 2. Analista
Carpeta raíz: AA Salud Pública y Asistencia Social	1. Administrador 2. Analista

Tabla4: Elemento de aplicación y roles con acceso

Conjuntamente con los niveles de accesos y roles antes determinados, para interactuar con la aplicación, la Plataforma Pentaho BI tiene incluida su propia seguridad, la cual está basada en la

infraestructura proporcionada por el Sistema de Seguridad Acegi (es un framework de seguridad existente en Java, creado por Ben Alex). El mismo se divide en cuatro áreas fundamentales, estas son:

Seguridad de acceso a datos de objetos: Incluye usuarios, contraseñas, autorizaciones permitidas, recursos web y protección a datos.

Autenticación: Tiene que ver con el procesamiento de información interactiva de inicio de sesión (por ejemplo nombre de usuario y contraseña) comparándola con la información recuperada del almacén de datos de seguridad.

Autorización de recursos web (URL): Brinda protección a las URL para responder a cada usuario si pueden o no acceder a una determinada página. Esto es decidido por el administrador de recursos web, el cual le brinda a cada usuario autenticado un permiso de seguridad, delimitando las páginas a las que tiene acceso y a las que no.

Autorización a objetos del dominio: En el sistema los únicos objetos del dominio protegidos por la plataforma son los objetos de repositorio otorgados al usuario autenticado. Es responsabilidad de los objetos del dominio autorizar las operaciones solicitadas por este.

2.12 Política de respaldo y recuperación

La política de respaldo y recuperación son las medidas y precauciones que se toman en caso de que ocurra algún fallo en el sistema; las que se emplearán en el mercado de datos consta de 3 puntos esenciales:

- **Periodicidad de las salvadas del sistema:** Estas se harán mensualmente a toda la información contenida en el mercado de datos, verificando siempre la existencia de una copia de toda la información almacenada.
- **Tablas involucradas:** Las tablas que se involucran en la realización son: hech_causas_muerte, hech_consultas, hech_consultas_us, hech_enfermedades, hech_camas, hech_camas_prov, hech_vacuna, hech_mortalidad_materna, hech_unidades_servicio, hech_us_prov, hech_índices, hech_personal_facultativo, hech_otro_personal_facultativo, hech_ingresos_hogares_maternos,

hech_donaciones_sangre, hech_tasa_mortalidad, hech_ingresos, hech_ingresos_hm,
hech_hosp_clinicas, hech_totales_indices, hech_totales_ingresos_hm,
hech_totales_tasa_mortalidad, dim_causas_defuncion, dim_consultas,
dim_enf_dec_obligatoria, dim_mortalidad_materna, dim_personal_facultativo,
dim_rango_cama, dim_tipo_vacuna y dim_unidades_servicio.

- **Salvas existentes:** A pesar de que actualmente no existen salvadas en esta área se prevé la realización de reemplazos de estas cada un año, así como también el chequeo de su estado mensualmente, mediante pruebas de integridad.

Conclusiones

En este capítulo se realizó un estudio del negocio, logrando tener un mejor conocimiento del flujo de la información en el área de Salud Pública y Asistencia Social. Se hizo un análisis de las series de salud para extraer la información que ayudará al correcto desarrollo del mercado de datos, también se definieron los requisitos funcionales y no funcionales. Para entender mejor el negocio se realizó el diagrama de casos de uso y el modelo dimensional y se generaron todos los artefactos correspondientes a esta etapa. Mediante la culminación de este capítulo quedan creadas las bases para la próxima etapa en la cual se implementará el mercado de datos.

CAPÍTULO 3: IMPLEMENTACIÓN DEL MERCADO DE DATOS

Introducción

En este capítulo se aborda todo lo referente a la implementación de la solución, tanto de los procesos ETL como la capa de inteligencia del negocio para el área de Salud Pública y Asistencia Social de la ONE, teniendo en cuenta las necesidades del negocio. También se muestra cómo está estructurado el modelo de datos físico y los usuarios y privilegios que están definidos. Se describe además todo el proceso de carga de los datos al almacén de datos así como todo el proceso de BI, desde la confección de los cubos OLAP hasta la realización de los reportes.

3.1 Estructura de datos

Las estructuras de datos son una colección de datos cuya organización se caracteriza por las funciones de acceso que se usan para almacenar y acceder a elementos individuales de datos; se pueden definir también como una forma de organizar un conjunto de datos elementales con el objetivo de facilitar su manipulación.

La base de datos se organiza en dos marcadas secciones; el esquema y los datos (o instancia). El esquema es la definición de la estructura de la base de datos y principalmente almacena los siguientes datos: el nombre de cada tabla, el nombre de cada columna, el tipo de dato de cada columna y la tabla a la que pertenece cada columna.

3.1.2 Esquemas y tablas

El Esquema de una Base de datos (en Inglés Database Schema) describe la estructura de una Base de datos, en un lenguaje formal soportado por un Sistema administrador de Base de Datos (DBMS). En una Base de datos relacional, el esquema define sus tablas, sus campos en cada tabla y las relaciones entre cada campo y cada tabla.

El esquema es generalmente almacenado en un Diccionario de Datos. Aunque generalmente el esquema es definido en un lenguaje de Base de datos, el término se usa a menudo para referirse a una representación gráfica de la estructura de base de datos.

Esquemas

Los esquemas que fueron creados se describen a continuación:

Esquema dimensiones: Contiene todas las dimensiones estáticas, en este caso se encuentran las tablas dimensiones *temporal_anno*, *provincia* y *país*.

Esquema mart_salud: Contiene todas las tablas hechos que controlan toda la información referente al área de Salud Pública y Asistencia Social, así como las dimensiones que son específicas de dicha área.

Tablas

Algunas de las tablas identificadas fueron las siguientes:

- **Tabla causas de muerte**

PK	FK	Nombre del atributo	Tipo de dato	No null	Descripción
x		dim_causas_defuncion_id	serial	x	Llave primaria.
		clasificación	varchar(50)		Clasificación de las causas de muerte.
		tipo_causa	varchar(80)		Tipos de causas de muerte.

- **Tabla hecho vacuna**

PK	FK	Nombre del atributo	Tipo de dato	No null	Descripción
x	x	dim_vacuna_id	serial	x	Llave primaria.
x	x	dim_temporal_anno_id	serial	x	Llave primaria.
		cant_inmunizaciones	double		Indicador para los cálculos de la cantidad de inmunizaciones.

3.1.2 Restricciones y secuencias

Restricciones

Una restricción es una condición que obliga el cumplimiento de ciertas condiciones en la base de datos. En el mundo real existen ciertas restricciones que deben cumplir los elementos que en el existen; por ejemplo, una persona sólo puede tener un número de carné de identidad y una única dirección oficial. Cuando se diseña una base de datos se debe reflejar fielmente el universo del discurso que se trata, lo que es lo mismo, reflejar las restricciones existentes en el mundo real. Las restricciones proveen un método de implementar reglas en la base de datos. Las restricciones restringen los datos que pueden ser almacenados en las tablas.

Secuencias

Son los atributos que se van a ir incrementando secuencialmente durante la entrada de los datos a la BD, en este caso las llave primarias. Las secuencias identificadas son las siguientes:

Secuencias	Esquemas	Propietario	Incrementado	Valor mínimo	Valor máximo
dim_pais_dim_pais_id_seq	dimensiones	postgres	1	1	922337203 685477580 7
dim_provincia_dim_provincia_id_seq	dimensiones	postgres	1	1	922337203 685477580 7
dim_temporal_anno_dim_temporal_anno_id_seq	dimensiones	postgres	1	1	922337203 685477580 7
dim_causas_defuncion_dim_causas_defuncion_id_seq	mart_salud	postgres	1	1	922337203 685477580 7
dim_consultas_dim_consultas_id_seq	mart_salud	postgres	1	1	922337203 685477580 7
dim_enf_dec_obligatoria_dim_enf_dec_obligatoria_id_seq	mart_salud	postgres	1	1	922337203 685477580 7
dim_mortalidad_materna_dim_mortalidad_materna_id_seq	mart_salud	postgres	1	1	922337203 685477580 7
dim_personal_facultativo_dim_personal_facultativo_id_seq	mart_salud	postgres	1	1	922337203 685477580 7
dim_rango_cama_dim_rango_cama_id_seq	mart_salud	postgres	1	1	922337203 685477580 7
dim_tipo_vacuna_dim_tipo_vacuna_id_seq	mart_salud	postgres	1	1	922337203 685477580 7
dim_unidades_servicio_dim_unidades_servicio_id_seq	mart_salud	postgres	1	1	922337203 685477580 7

3.1.3 Índices

Un índice es una estructura de disco asociada con una tabla o una vista que acelera la recuperación de filas de la tabla o de la vista. Un índice contiene claves generadas a partir de una o varias columnas de la tabla o la vista. Dichas claves están almacenadas en una estructura (árbol b) que permite que se busque de forma rápida y eficiente la fila o filas asociadas a los valores de cada clave. [5]

Dos filas no pueden tener el mismo valor para la clave de índice, de lo contrario, el índice no es único y varias filas pueden compartir el mismo valor de clave. Los índices se mantienen automáticamente para una tabla o vista cuando se modifican los datos de la tabla.

Los índices se crean automáticamente cuando las restricciones PRIMARY KEY y UNIQUE se definen en las columnas de tabla. Por ejemplo, cuando crea una tabla e identifica una determinada columna como la clave primaria, el motor de base de datos crea automáticamente una restricción PRIMARY KEY y un índice en esa columna. Los índices generados son los siguientes:

Índice	Tabla	Esquema	Tipo	Campo	PK	Único
PK35	dim_pais	dimensiones	btree	dim_pais_id	x	x
PK16	dim_provincia	dimensiones	btree	dim_pais_id	x	x
PK15	dim_temporal_anno	dimensiones	btree	dim_temporal_anno_id	x	x
PK1	dim_causas_defuncion	mart_salud	btree	dim_causas_defuncion_id	x	x
PK7	dim_consultas	mart_salud	btree	dim_consultas_id	x	x
PK4	dim_enf_dec_obligatoria	mart_salud	btree	dim_enf_dec_obligatoria_id	x	x
PK3	dim_mortalidad_materna	mart_salud	btree	dim_mortalidad_materna_id	x	x
PK18	dim_personal_facultativo	mart_salud	btree	dim_personal_facultativo_id	x	x
PK6	dim_rango_cama	mart_salud	btree	dim_rango_cama_id	x	x
PK5	dim_tipo_vacuna	mart_salud	btree	dim_tipo_vacuna_id	x	x
PK17	dim_unidades_servicio	mart_salud	btree	dim_unidades_servicio_id	x	x
PK13	hech_camas	mart_salud	btree	dim_rango_cama_id, dim_temporal_anno_id	x	x
PK2	hech_causas_muerte	mart_salud	btree	dim_temporal_anno_id, dim_causas_muerte_id	x	x
PK14	hech_consultas	mart_salud	btree	dim_consultas_id, dim_temporal_anno_id	x	x

PK11	hech_enfermedades	mart_salud	btree	dim_enf_dec_obligatoria_id, dim_temporal_anno_id	x	x
PK34	hech_ingresos_hogares_ maternos	mart_salud	btree	dim_temporal_anno_id, dim_pais_id	x	x
PK10	hech_mortalidad_matern a	mart_salud	btree	dim_mortalidad_materna_id , dim_temporal_anno_id	x	x
PK29	hech_otras_us	mart_salud	btree	dim_unidades_servicio_id, dim_temporal_anno_id	x	x
PK33	hech_otro_personal_facu ltativo	mart_salud	btree	dim_temporal_anno_id, dim_personal_facultativo_id	x	x
PK22	hech_personal_facultativ o	mart_salud	btree	dim_temporal_anno_id, dim_provincia_id, dim_personal_facultativo_id	x	x
PK21	hech_unidades_servicio	mart_salud	btree	dim_unidades_servicio_id, dim_provincia_id, dim_temporal_anno_id	x	x
PK12	hech_vacuna	mart_salud	btree	dim_tipo_vacuna_id, dim_temporal_anno_id	x	x
PK25	hech_varios_indicadores	mart_salud	btree	dim_temporal_anno_id, dim_provincia_id	x	x

3.2 Usuarios y privilegios

Usuarios

Los usuarios de la base de datos serán:

Analista: Es el usuario que analiza y consulta la información de los diferentes indicadores de Salud Pública y Asistencia Social.

Administrador del sistema: Es el que administra toda la base de datos, dígase administrar los roles y los privilegios de cada uno, también se encarga del rendimiento y buen funcionamiento de la base de datos.

Administrador de ETL: Es el responsable de la extracción, transformación y carga de los datos.

Privilegios

Los privilegios serán otorgados a los usuarios del sistema, según el rol que representen. En este caso:

- El Analista tendrá el privilegio de actuar con los esquemas y las tablas, poseerá permisos de lectura, en este caso, la acción llamada SELECT en la base de datos.
- El Administrador del sistema tendrá el privilegio de controlar la base de datos, su configuración y administración, disfrutará de todos los permisos permitidos en la base de datos.
- El Administrador de ETL tendrá el privilegio de visualizar y modificar los datos de la base de datos, por lo que poseerá permisos de lectura y escritura, las acciones nombradas SELECT, INSERT, UPDATE, DELETE en la base de datos.

3.3 Implementación de los procesos ETL

El proceso de ETL es el más importante, pues es quien organiza el flujo de los datos y aporta los métodos y herramientas necesarios para mover datos desde múltiples fuentes a un almacén de datos, reformatearlos, limpiarlos y cargarlos en otra base de datos o mercado de datos.

La primera parte del proceso consiste en extraer los datos desde el sistema origen, una parte importante de este proceso es la de analizar los datos extraídos, de lo que resulta un chequeo que verifica si los datos cumplen la pauta o estructura que se esperaba, de no ser así, los datos son rechazados. En el caso del área de salud los datos son recogidos automáticamente del Excel proporcionado por la ONE por lo que los datos se encuentran listos para pasar al proceso de transformación. En esta etapa del proceso no fue necesario realizar una limpieza de los datos, pues contamos con series históricas las cuales ya están limpias y no contienen ninguna inconsistencia.

El proceso de transformación de los datos se realizó de acuerdo a las reglas del negocio que se definieron, como es el caso de realizar tratamientos a valores nulos. Para su realización se ejecutaron 30 transformaciones. Finalmente está el proceso de la carga, este es el momento cuando los datos de la fase anterior, transformación, son cargados al sistema destino. En algunas bases de datos se sobrescribe la información antigua con nuevos datos, pero en el caso del mercado de datos Salud Pública y Asistencia Social la información es cargada por primera vez en la base de datos. Cuando se realiza la carga de los datos se aplican todas las restricciones que se definieron, ejemplo de esto es que los identificadores de las tablas no pueden tomar valores repetidos, deben ser únicos.

Para la implementación de los procesos de ETL nos apoyamos en el diseño del subsistema de integración, ver **anexo 1**. A continuación se dará una breve explicación de cómo se implementaron los flujos de transformación.

3.3.1 Implementación de los flujos de transformación

La transformación es el elemento básico de diseño de los procesos ETL, la misma no es ningún programa ni un ejecutable, simplemente es un conjunto de metadatos en XML que le indican al motor de la herramienta *Pentaho Data Integration* las acciones a realizar. Una transformación se compone de pasos, que están enlazados entre sí a través de saltos. Cada paso tiene una ventana de configuración específica, donde se determina los elementos a tratar y su forma de comportamiento.

Para la realización del proceso de ETL se ejecutaron 30 transformaciones, a continuación se mostrará un ejemplo de las transformaciones realizadas, quedando de la siguiente manera:

hech_vacunat: En la transformación del *hecho vacunas* se realiza la carga correspondiente a la serie histórica **Inmunizaciones por tipo de vacunas**.

En esta transformación se garantiza que el registro fuente (series históricas) no contengan campos nulos y que los id de las dimensiones sean válidos. Se realizan las búsquedas en la BD de las dimensiones relacionadas con el hecho y finalmente se carga el hecho con la transformación correspondiente.

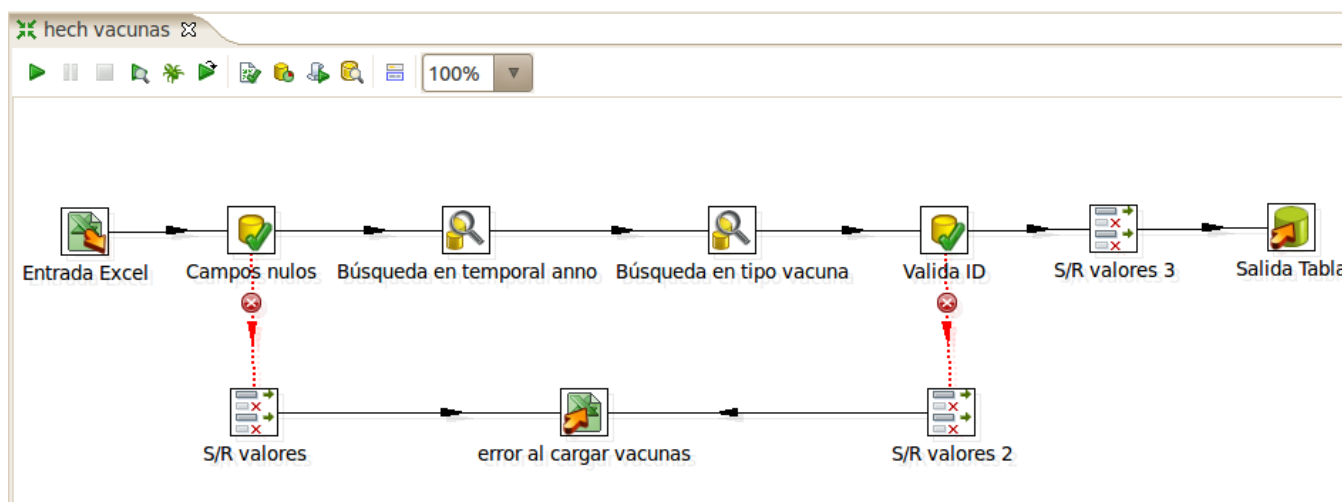


Fig5: Transformación del hecho vacunas

3.4 Implementación de los procesos de BI

La Inteligencia de Negocios o BI es una forma de manejar la información histórica a través de la construcción de almacenes de datos o mercados de datos, con el fin de explotarla y de esta forma poder hacer respectivos análisis de la misma para llevar a cabo una mejor toma de decisiones.

Los procesos ETL son los componentes más importantes y de valor añadido de una infraestructura de BI, por lo que el éxito de la presentación de la capa de visualización depende

fundamentalmente de la forma en que se organiza la información estructural, estos son los cubos OLAP, los cuales pueden contener una o más dimensiones.

3.4.1 Implementación de los cubos OLAP

Uno de los elementos fundamentales donde se organiza la información estructuralmente son los cubos OLAP los cuales pueden poseer más de tres dimensiones por lo que son llamados hipercubos, específicamente en el área de Salud Pública y Asistencia Social, los cubos OLAP poseen hasta cuatro dimensiones solamente. En el caso concreto de la solución se desarrollaron 22 cubos multidimensionales principales, apoyándose en la herramienta Pentaho Workbench; en el capítulo 1 fueron abordadas las facilidades y características que esa herramienta facilita a la hora de crear los cubos OLAP. A continuación se mostrará uno de estos cubos multidimensionales creados.

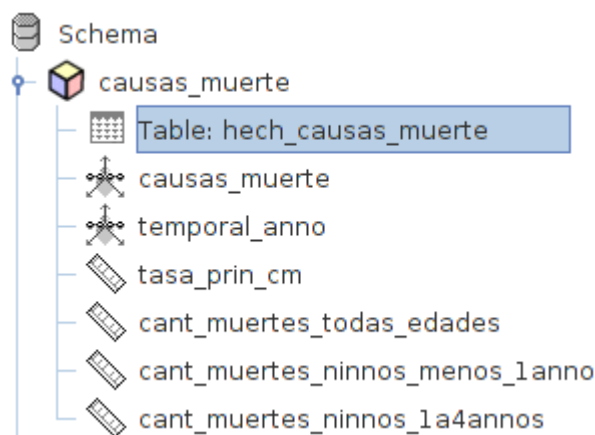


Fig7: Cubo Causas de muerte

Los cubos OLAP se definen uno por cada tabla hecho, debido a que estas tablas son las fuentes de información principal de la organización estructural de los datos, además se observan las tablas dimensiones relacionadas con este hecho así como las medidas del mismo.

3.4.2 Reportes candidatos

Los reportes candidatos son la información que el cliente desea que se muestre como finalidad del producto. Los mismos fueron identificados luego de realizar un análisis de los ficheros Excel, donde se recoge toda la información referente al área de Salud Pública y Asistencia Social de la ONE. Los reportes son agrupados por libros de trabajo, en el epígrafe 2.10, se explica al respecto y para más información detallada se puede consultar los artefactos **Reportes candidatos** y

Arquitectura de la información. A continuación se describirá un ejemplo de los 40 reportes candidatos que fueron identificados.

- **Obtener anualmente la cantidad de las principales causas de muerte de todas las edades:** En este reporte se muestra anualmente la cantidad de defunciones teniendo en cuenta los tipos de causa de defunción.

Nueva vista de aná..		Medidas											
		Cantidad de muertes de todas las edades											
		tiempo											
causa		Todos	1968	1969	1970	1971	1972	1973	1974	1975	1976	1977	1999
Todas		2.258.061	33.939	34.687	34.868	34.105	33.864	36.088	37.949	37.136	39.661	41.994	42.1
De todas las edades		2.258.061	33.939	34.687	34.868	34.105	33.864	36.088	37.949	37.136	39.661	41.994	42.1
Accidentes		180.355	2.644	2.807	3.089	2.754	2.822	2.766	2.939	3.110	3.167	3.787	3.7
Cirrosis y otras enfermedades crónicas del hígado		23.748	0	0	0	0	0	0	0	0	0	0	
Diabetes mellitus		68.192	1.042	889	847	897	890	929	905	949	1.055	1.120	1.0
Enfermedades cerebrovasculares		281.492	5.484	5.258	5.155	4.956	4.694	4.736	4.873	4.746	4.913	5.479	5.1
Enfermedades crónicas de las vías respiratorias inferiores		43.198	0	0	0	0	0	0	0	0	0	0	
Enfermedades de las arterias, artereolas y vasos capilares		64.766	0	0	0	0	0	0	0	0	0	0	
Enfermedades del corazón		775.571	12.352	12.537	12.704	12.253	12.003	13.251	14.305	13.840	15.053	15.615	16.3
Influenza y neumonía		181.736	3.373	4.204	3.602	3.421	3.261	3.949	4.188	3.618	3.965	4.276	4.3
Suicidios y lesiones autoinfligidas		76.010	1.041	936	1.011	1.151	1.265	1.449	1.617	1.608	1.667	1.695	1.8
Tumores malignos		562.993	8.003	8.056	8.460	8.673	8.929	9.008	9.122	9.265	9.841	10.022	9.6
En niños de 1 a 4 años de edad		0	0	0	0	0	0	0	0	0	0	0	
En niños menores de 1 año de edad		0	0	0	0	0	0	0	0	0	0	0	

Fig10: Reporte causas de muerte de todas las edades

El mismo se implementó mediante consultas mdx. El lenguaje MDX es en los sistemas OLAP el equivalente al SQL en los sistemas gestores de bases de datos relacionales. Eso significa que es el lenguaje a través del cual podemos explotar la información que reside en los motores OLAP y satisfacer las consultas analíticas. La consulta que se realizó para este reporte es la siguiente:

```

select                                NON                                EMPTY
Hierarchize(Union(Crossjoin({[Measures].[cant_muertes_todas_edades]},
{[temporal_anno.tiempo].[Todos]}), Crossjoin({[Measures].[cant_muertes_todas_edades]},
[temporal_anno.tiempo].[Todos].Children))) ON COLUMNS, NON EMPTY
Hierarchize(Union(Union({[causas_defuncion.causa].[Todas]},

```


[causas_defuncion.causa].[Todas].Children), [causas_defuncion.causa].[De todas las edades].Children)) ON ROWS from [causas_difuncion]

Conclusiones

En este capítulo se abordaron los elementos fundamentales de la implementación del mercado de datos, dándose a conocer como quedó físicamente el mercado de datos, logrando una estructura robusta entre las relaciones de las tablas multidimensionales, las cuales se encuentran agrupadas por esquemas para una mejor organización en la BD. Por lo que se puede arribar a las siguientes conclusiones: Se definieron 2 esquemas en la BD: *mart_salud* y *dimensiones*, se definieron 35 tablas: 13 de dimensiones y 22 hechos, se realizaron 30 transformaciones, se diseñaron los cubos OLAP, se identificó un área de análisis, 11 libros de trabajo y se crearon un total de 40 reportes.

CAPÍTULO 4: VALIDACIÓN Y PRUEBA DEL MERCADO DE DATOS

Introducción

En este capítulo se aborda todo lo referente a la validación y pruebas realizadas al mercado de datos de Salud Pública y Asistencia Social. Se procede a validar la solución propuesta a través de la aplicación de las listas de chequeo, los casos de prueba y la carta de aceptación del cliente.

4.1 Pruebas de software

La prueba de software es un elemento crítico para la garantía de la calidad del software y representa una revisión de las especificaciones del diseño y de la implementación. Se le realizan pruebas a la documentación, a un módulo de la aplicación, a toda la aplicación, etc., en dependencia de lo que se quiera probar.

También es importante destacar que la calidad de un producto software debe ser considerada en todos sus estados de evolución a medida que avanza su desarrollo de acuerdo al ciclo de vida seleccionado para su construcción (especificaciones, diseño, código, etc.). La calidad siempre va a depender de los requisitos o necesidades que se desee satisfacer. Por eso, la evaluación de la calidad de un producto siempre va a implicar una comparación entre los requisitos preestablecidos y el producto realmente desarrollado.

Para lograr un producto con calidad y que cumpliera con las expectativas del cliente se realizaron las pruebas al mercado de datos de acuerdo a los casos de prueba y las listas de chequeo. En los siguientes epígrafes se abordará sobre las listas de chequeo y los casos de prueba. Las pruebas realizadas fueron las siguientes:

- ❖ **Pruebas de Unidad:** Se buscan errores en los componentes más pequeños del programa, se realizan sobre cada módulo del software de manera independiente.
- ❖ **Pruebas de Integración:** Se comprueba la compatibilidad y funcionalidad de las interfaces entre las distintas 'partes' que componen un sistema, estas 'partes' pueden ser módulos, aplicaciones individuales, aplicaciones cliente/servidor, etc. Este tipo de pruebas es especialmente relevante en aplicaciones distribuidas.
- ❖ **Pruebas de Sistema:** El software ya validado se integra con el resto del sistema donde algunos tipos de pruebas a considerar son:

- ✓ Seguridad: Se determinan los niveles de permiso de usuarios, las operaciones de acceso al sistema y acceso a datos.
- ✓ Usabilidad: Se determina la calidad de la experiencia de un usuario en la forma en la que éste interactúa con el sistema, se considera la facilidad de uso y el grado de satisfacción del usuario.
- ❖ **Pruebas de Aceptación:** Son las que hará el cliente, se determina que el sistema cumple con lo deseado y se obtiene la conformidad del cliente.

4.2 Elaboración y aplicación de las Listas de Chequeo

Los requisitos una vez definidos necesitan ser validados. La validación de los mismos tiene como misión demostrar que la definición de los requisitos describe realmente el sistema que el usuario necesita o el cliente desea. La técnica empleada para validar los requisitos identificados en la etapa de análisis es las listas de chequeos. Esta técnica se realiza con el objetivo de evaluar la calidad de los artefactos que se generan en el análisis de la solución. Mediante estas listas se pueden cubrir todos los aspectos aplicables en los temas de análisis, particularmente para el tema de Salud Pública y Asistencia Social. La lista de chequeo contiene diferentes indicadores a evaluar los cuales se encuentran distribuidos en tres secciones fundamentales:

- **Estructura del documento:** Abarca todos los aspectos definidos por el expediente de proyecto o el formato establecido por el proyecto.
- **Indicadores definidos:** Abarca todos los indicadores a evaluar durante la etapa de desarrollo del mercado.
- **Semántica del documento:** Contempla todos los indicadores a evaluar respecto a la ortografía, redacción y demás.

Al ser aplicados estos chequeos a los artefactos que se generaron en la elaboración del mercado de datos, se detectaron 18 no conformidades.

Elementos que forman parte de la estructura de la lista de chequeo:

- **Peso:** Define si el indicador a evaluar es crítico o no.
- **Indicadores a evaluar:** Son los indicadores a evaluar en las secciones *Estructura del documento*, *Indicadores definidos* y *Semántica del documento*.

- **Evaluación (Eval):** Es la forma de evaluar el indicador en cuestión. El mismo se evalúa de 1 en caso de que exista alguna dificultad sobre el indicador y 0 en caso de que el indicador revisado no presente problemas.
- **N.P. (No Procede):** Se usa para especificar que el indicador no es necesario evaluarlo en ese caso.
- **Cantidad de elementos afectados:** Especifica la cantidad de errores encontrados sobre el mismo indicador.
- **Comentario:** Especifica los señalamientos o sugerencias que quiera incluir la persona que aplica la lista de chequeo. Pueden o no existir señalamientos o sugerencias.

Una vez aplicada la lista de chequeo se detectan los indicadores evaluados de mal y con el objetivo de darles solución se especifican en una tabla de no conformidades (NC), la cual presenta la siguiente estructura:

- **No.:** Es un número consecutivo e indica la cantidad de no conformidades identificadas.
- **Elemento de evaluación:** Se refiere a un número que identifica al elemento de evaluación para el cual se corresponden los indicadores identificados.
- **No conformidad:** Especifica la no conformidad a la que se refiere.
- **Fase correspondiente:** Especifica la fase del procedimiento a la que corresponde la no conformidad encontrada.
- **Significación:** Especifica si la no conformidad es o no significativa, dependiendo si el indicador es o no crítico.
- **Recomendación:** Especifica si la no conformidad es una recomendación, es decir que no es de obligatorio cumplimiento que se solucione por parte de los diseñadores.
- **Estado NC:** Especifica el estado de solución en que se encuentra la no conformidad, puede ser: pendiente o solucionada.
- **Respuesta del equipo de desarrollo:** Si es necesario se especifica la respuesta que le da el equipo de desarrollo a la no conformidad.

Estructura del documento					
Peso	Indicadores a evaluar	Evaluación	No procede	Cantidad de elementos afectados	Comentarios

Crítico	1. ¿Los entregables contienen las secciones obligatorias de la plantilla estándar definidas para un expediente de proyecto? (Portada, Control de Versiones, Reglas de Confidencialidad, Tabla de Contenidos y Contenido) (Ver Expediente de Proyecto)				
Indicadores definidos por la etapa					
Peso	Indicadores a evaluar	Evaluación	No procede	Cantidad de elementos afectados	Comentarios
	1. ¿Se utilizó un lenguaje cuyas sentencias son expresables mediante una sintaxis bien definida?				
Crítico	2. ¿Los reportes son configurables a través de la interfaz del sistema?				
	3. ¿La interfaz está orientada a facilitar el uso de las funciones del sistema por parte de los usuarios?				
Crítico	4. ¿No existen restricciones para construir cubos OLAP con dimensiones y niveles de agregación ilimitados?				
Crítico	5. ¿Los usuarios son capaces de manipular los resultados de manera que se ajusten a sus necesidades, conformando nuevos reportes?				
	6. ¿El sistema responde de una forma rápida a la información que le sea solicitada por el usuario?				
Crítico	7. ¿El sistema refleja cualquier lógica del negocio para poder responder a preguntas específicas?				
Crítico	8. ¿El sistema garantiza la confidencialidad y seguridad de acceso a los datos por rol de los usuarios?				
	9. ¿Los datos e información derivados del proceso de análisis realizado mediante la aplicación, apoyan la toma de decisiones en la institución?				

Crítico	10. ¿Los cambios en los datos se reflejan automáticamente en los reportes de forma instantánea?				
Semántica del documento					
Peso	Indicadores a evaluar	Evaluación	No procede	Cantidad de elementos	Comentarios
Crítico	1. ¿Se han identificado errores ortográficos en los entregables?				
Crítico	2. ¿Se entiende claramente lo que se ha especificado en el documento?				
	3. ¿El número de página que aparece en el índice coincide con el contenido que se refleja realmente en dicha página?				

Tabla8: Lista de Chequeo

No. de evaluación	Elemento	No conformidad	Fase correspondiente	Significación	Recomendación	Estado NC	Respuesta del equipo de desarrollo
--------------------------	-----------------	-----------------------	-----------------------------	----------------------	----------------------	------------------	---

Tabla9: No conformidad

4.3 Casos de prueba

A partir del momento en que se ha elaborado los casos de uso del sistema se pueden elaborar los casos de prueba; el propósito de un caso de prueba es especificar una forma de probar el sistema, incluyendo las entradas con las que se ha de validar, los resultados esperados y las condiciones bajo las que ha de probarse. Los casos de prueba son necesarios para verificar que la aplicación cumple con los requisitos identificados en la etapa de análisis. Los casos de pruebas deben chequear:

- Si el producto satisface los requerimientos del usuario, tal y como se describe en las especificación de los requerimientos.
- Si el producto se comporta como se desea, tal y como se describe en las especificaciones funcionales del diseño.

Se elaboraron en la etapa de análisis 11 casos de prueba, los cuales fueron creados uno por cada caso de uso, para posteriormente en la etapa de pruebas proceder a probar cada uno arrojando los siguientes resultados en la primera iteración:

Fueron encontradas 18 no conformidades.

4.4 Prueba de aceptación

La prueba de aceptación del usuario es la última que se realiza antes del despliegue de la aplicación. El objetivo de esta prueba es verificar que el sistema está listo para ser usado por los usuarios finales para ejecutar aquellas funciones y tareas para las cuales el sistema fue construido. En el caso del mercado de datos para el área de Salud Pública y Asistencia Social para verificar la aceptación del producto estuvo presente la especialista de la ONE: Elena Leonila Fernández García. La compañera estuvo de acuerdo con la solución desarrollada. Durante el proceso de análisis y diseño la compañera estuvo supervisando cada avance, quedando satisfecha con el diseño lógico y la arquitectura de la solución que se generó al finalizar esta fase. Es importante resaltar que durante la implementación de la solución se obtuvieron resultados tangibles, como el modelo físico de la BD y la aplicación Mondrian para la visualización de los reportes en el área de Salud Pública y Asistencia Social, con los que la especialista de la ONE estuvo satisfecha. (Ver anexo 2)

Conclusiones

En este capítulo se abordaron los temas referente a las validaciones y pruebas realizadas al mercado de datos de Salud Pública y Asistencia Social, también se elaboró y aplicó una lista de chequeo con el objetivo de evaluar el mercado de datos, donde se obtuvieron los siguientes resultados. Se identificaron de forma general 14 indicadores necesarios e imprescindibles para la evaluación final de esta etapa, la evaluación final del mercado de datos fue de Bien, lo que demuestra la calidad del producto desarrollado.

CONCLUSIONES GENERALES

Al concluir esta solución, se puede plantear que fueron cumplidos los objetivos trazados y las tareas de la investigación propuestas. Por tanto, se llega a las siguientes conclusiones:

- ❖ Se elaboró el marco teórico de la investigación acerca de las principales herramientas y metodologías para el desarrollo de mercado de datos, procesos de ETL y procesos de BI, donde se definieron y utilizaron las herramientas necesarias para la implementación del mercado de datos y de la capa de visualización.
- ❖ Se diseñó e implemento el subsistema de integración del Mercado de Datos de Salud Pública y Asistencia Social.
- ❖ Se diseñó e implemento el subsistema de visualización del Mercado de Datos de Salud Pública y Asistencia Social.
- ❖ Se validó la solución desarrollada mediante las listas de chequeo, los casos de prueba y la carta de aceptación del cliente, obteniéndose resultados satisfactorios.
- ❖ Se desarrolló el expediente de proyecto con los documentos necesarios para su análisis con el objetivo de dejar en archivos todo el proceso de desarrollo del producto final.

RECOMENDACIONES

Con el propósito de enriquecer lo plasmado en este trabajo se sugiere:

- Que se garantice la continuidad de la integración de los datos en esta área.
- Que se implemente minería de datos sobre los datos del almacén debido a que este tipo de análisis prepara y explora los datos para sacar la información oculta en ellos.

TRABAJOS CITADOS

1. **Escobar Domínguez René Raydel** Desarrollo de un Almacén de Datos para el Control del Consumo de Portadores Energéticos en la Oficina Nacional de Estadística [Libro]. - Ciudad Habana : [s.n.], 2010.
2. **Casale Cabrera María Evelia** Almacenes de Datos [En línea]. - 2009. - <http://hp.fciencias.unam.mx/~alg/bd/dwh.pdf..> - sn.
3. **Kimball Ralph y Caserta Joe** Wiley Publishing The Data Warehouse ETL Toolkit [Libro]. - [s.l.] : Wiley Publishing, 2004.
4. **Nader Javier** Sistema de Apoyo Gerencial Universitario [Libro]. - 2003. - sn.
5. Microsoft [En línea] // Microsoft. - 2011. - <http://msdn.microsoft.com/es-es/library/ms190457.aspx>.

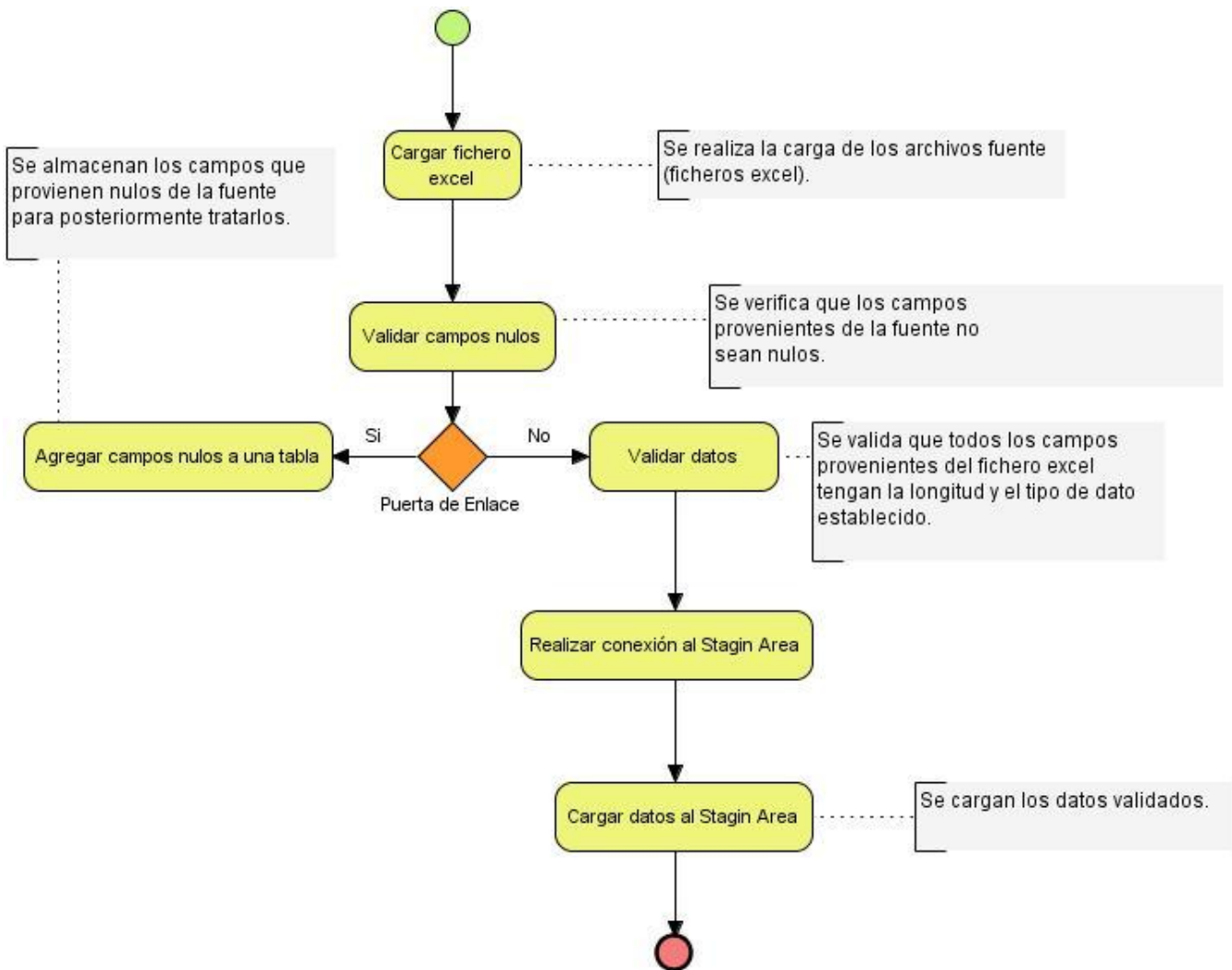
BIBLIOGRAFÍA

1. **Casale Cabrera María Evelia** Almacenes de Datos [En línea]. - 2009. - <http://hp.fciencias.unam.mx/~alg/bd/dwh.pdf..> - sn.
2. **Correa Cristián** Cognus [En línea] // Cognus. - 2008. - <http://www.cognus.cl/content/view/794163/Pentaho-Data-Integration-4-0.html#content-top>.
3. **Escobar Domínguez René Raydel** Desarrollo de un Almacén de Datos para el Control del Consumo de Portadores Energéticos en la Oficina Nacional de Estadística [Libro]. - Ciudad Habana : [s.n.], 2010.
4. **Espinosa Roberto** El Rincon del BI [En línea] // El Rincon del BI. - 10 de mayo de 2010. - <http://churriwifi.wordpress.com/2010/05/10/16-3-construccion-procesos-etl-utilizando-kettle-pentaho-data-integration/>.
5. Free Download Manager [En línea] // Free Download Manager. - 5 de marzo de 2007. - http://www.freedownloadmanager.org/es/downloads/Paradigma_Visual_para_UML_%28M%C3%8D%29_14720_p/.
6. Gravitar [En línea] // Gravitar. - 2011. - <http://www.gravitar.biz/index.php/herramientas-bi/pentaho/caracteristicas-pentaho/>.
7. **Imhoff Claudia y Glemmo Nicholas** Mastering Data Warehouse Design - Relational and Dimensional Techniques [Libro]. - [s.l.] : Wiley Publishing, 2003. - sn.
8. **Juristo Natalia** Técnicas de evaluación de Software [Libro]. - 2006. - sn.
9. **Kimball Ralph y Caserta Joe** Wiley Publishing The Data Warehouse ETL Toolkit [Libro]. - [s.l.] : Wiley Publishing, 2004.
10. Microsoft [En línea] // Microsoft. - 2011. - <http://msdn.microsoft.com/es-es/library/ms190457.aspx>.
11. **Microsoft Corporation** [En línea]. - 2011. - 3 de 3 de 2011. - <http://msdn.microsoft.com/es-es/library/ms190457.aspx>.
12. **Moné Roque Diana** Aplicación de las pruebas de liberación al sistema Informático de Genética Médica [Libro]. - Ciudad Habana : [s.n.], 2009.

13. **Nader Javier** Sistema de Apoyo Gerencial Universitario [Libro]. - 2003. - sn.
14. Oficina Nacional de Estadísticas [En línea] // Oficina Nacional de Estadísticas. - 2011. - <http://www.one.cu>.
15. **Ortiz Luis** Webquest [En línea] // Webquest. - 25 de octubre de 2010. - <http://www.webquest.es/wq/data-warehouse-y-datamart>.
16. **Pérez Lamancha Beatriz** Proceso de Testing Funcional Independiente [Libro]. - Uruguay : [s.n.], 2006.
17. pgAdmin PostgreSQL Tools [En línea] // pgAdmin PostgreSQL Tools. - 2011. - <http://www.pgadmin.org/>.
18. **Ponniah P** Data Warehousing Fundamentals [Libro]. - EUA : Wiley Publishing Inc., 2001.
19. **Rosillo Solano Marco Vinicio** Scribd [En línea] // Scribd. - 25 de marzo de 2011. - <http://es.scribd.com/doc/51548013/modulo-base-de-datos>.
20. Sinnexus [En línea] // Sinnexus. - 2011. - http://www.sinnexus.com/business_intelligence/datamart.aspx.
21. **sn** Fedora [En línea] // Fedora. - 2011. - <http://proyectofedora.org/colombia/?p=771>.
22. **Soft Jasper** Pentaho Mondrian Documentation [En línea] // Pentaho Mondrian Documentation. - 2007. - <http://mondrian.pentaho.com/documentation/workbench.php>.
23. **Universitat Jaume I** Universitat Jaume I [En línea] // Universitat Jaume I / ed. Andrés María M. - 02 de 12 de 2001. - 03 de 03 de 2011. - <http://www3.uji.es/~mmarques/f47/apun/apun.html>.
24. Visual Paradigm [En línea] // Visual Paradigm. - 2011. - <http://www.visual-paradigm.com/product/vpuml/editions/community.jsp>.

ANEXOS

Anexo1: Diseño del subsistema de integración



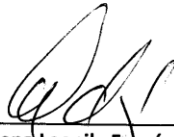
Anexo2: Carta de aceptación del cliente

Carta de aceptación del cliente.

Fecha: 8/6/2011

Yo: **Elena Leonila Fernández García**, representante de la Oficina Nacional de Estadísticas en la Universidad de las Ciencias Informáticas para el desarrollo del Sistema de Información de Gobierno. Apruebo:

1- El mercado de datos para el área de SALUD cumple con los requisitos especificados por el cliente.



Elena Leonila Fernández García
Firma del cliente.

GLOSARIO DE TÉRMINOS

ONE: Oficina Nacional de Estadísticas.

ETL: Extracción, Transformación y Carga

BI: Inteligencia de Negocio.

SGBD: Sistema Gestor de Base de Datos.

BD: Base de datos.

CUS: Casos de Uso del Sistema.