Universidad de las Ciencias Informáticas

Facultad 6

Título: Mercado de datos Series históricas de población para el Sistema de Información de Gobierno

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

Autora:

Katy Medina García

Tutores:

Ing. Themis Patricia Díaz Morales

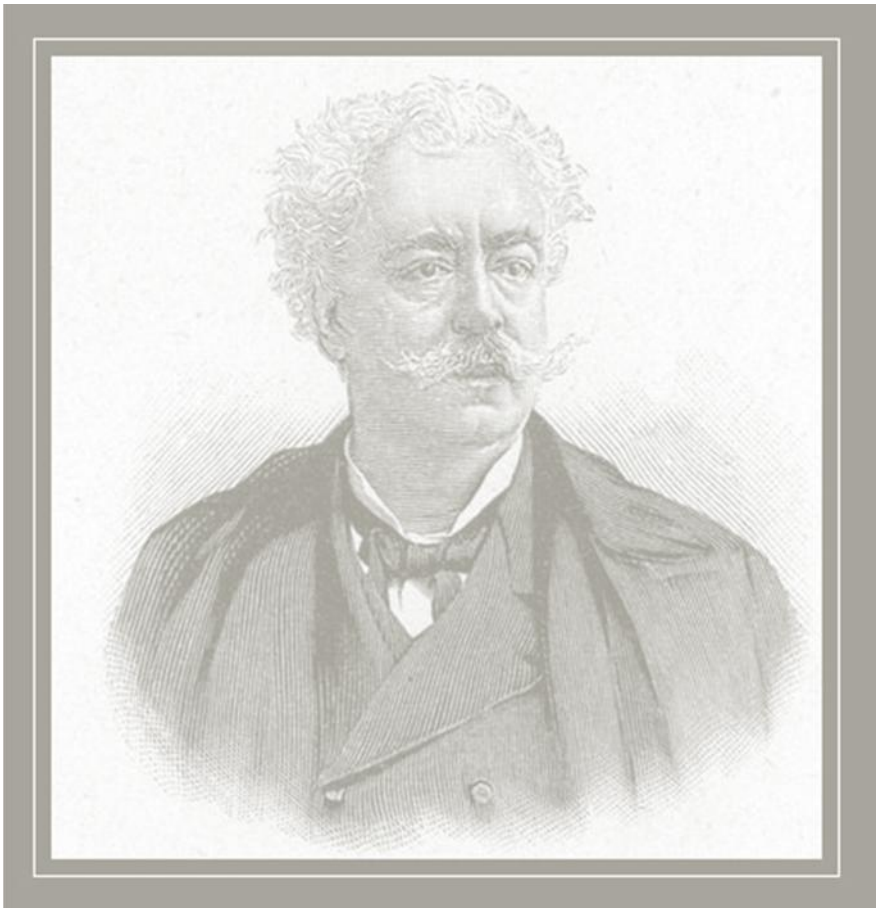
Ing. José Salvador Bermúdez Rodríguez

La Habana, junio de 2011

"Año 53 de la Revolución"

"Por este mundo pasaré solamente una vez, si hay una buena obra que pueda hacer, si hay una buena palabra que pueda decir; haré esa buena obra y diré esa buena palabra, pues ya nunca volveré a pasar por aquí"

Edmundo De Amicis



Declaración de autoría

Declaro ser autora del presente Trabajo de Diploma “Mercado de datos Series históricas de población para el Sistema de Información de Gobierno” y reconozco 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 _____.

Katy Medina García

_____

Firma del Autor

Ing. Themis Patricia Díaz Morales

_____

Firma del Tutor

Ing. José Salvador Bermúdez Rodríguez

_____

Firma del Tutor

Tutores:

Tutor: Ing. Themis Patricia Díaz Morales

Especialidad de graduación: Ingeniería en Ciencias Informáticas

Años de experiencia en el tema: 2

Años de graduado: 1

Correo Electrónico: tpdiaz@uci.cu

Tutor: Ing. José Salvador Bermúdez Rodríguez

Especialidad de graduación: Ingeniería en Ciencias Informáticas

Años de experiencia en el tema: 2

Años de graduado: 1

Correo Electrónico: jsbermudez@uci.cu

Muchas gracias:

A mi mamita linda por ser mí mejor amiga, por apoyarme siempre en las buenas y en las malas, por ser la persona más noble y luchadora que conozco, por abrigarme en su pecho cuando me dejo vencer por el llanto y darme fuerzas para continuar. Pero sobre todo por su comprensión y su amor infinito.

A mi papito por la confianza que ha depositado en mí, por el esfuerzo realizado para que pudiera llegar hasta aquí, por servirme de guía y por quererme tanto.

A mis abuelitos por todo lo que me enseñaron en los primeros pasos de mi vida y por el amor y el cariño que sienten por mí.

A mi hermanito Rone por preocuparse tanto por mí y porque a pesar de no saber mucho de la vida, en momentos de tristezas me ha ayudado a sonreír.

A mi novio Javi por enseñarme todo lo que sabe, por estar siempre conmigo y apoyarme en momentos difíciles como los que he vivido este curso lejos de él, donde los segundos han sido horas. Por ser parte de este sueño, gracias mi Amor.

A todos mis amigos que durante estos 5 años me ayudaron de una forma u otra, a los del departamento de almacenes entre ellos Cecilia, Leonel y Yoendy.

A mi amiga y compañera de cuarto Raiza que a pesar de tener un fuerte carácter, en momentos difíciles me tendió su mano y sobre todo por estar siempre ahí para escuchar cada charla nocturna que le daba sobre mi novio…aunque algunas veces se quedaba dormida.

A mis tutores Themis y Salvi por el cariño y el apoyo brindado, por el tiempo dedicado y el esmero con el que me han guiado por el mejor camino.

A las profesoras del tribunal por todo lo que he aprendido con ellas en cada corte, por hacer sus señalamientos de una forma tan sutil que parecían consejos de un amigo.

A mi oponente Yoamel por su cooperación durante la realización de este trabajo y por las asignaturas impartidas en años anteriores.

Dedico esta investigación:

A toda mi familia por ser muy unida y porque en parte todos me han ayudado a conseguir este sueño.

A mi tía Negra por ser para mí como una madre.

A mi abuelito José que siempre me aconseja y me trata como a su nieta más chiquita.

A mi abuelita Violeta que siempre ha estado muy orgullosa de mí y por lo mucho que deseaba que llegara este día.

Por último quisiera mencionar a las 4 personas más especiales e importantes en mi vida: mi madre, mi padre, mi hermanito y mi novio Javi, a ustedes por llenar mis días de felicidad y estar siempre conmigo dedico especialmente este éxito.

Resumen

La presente investigación surgió como parte de la colaboración que existe entre la Universidad de las Ciencias Informáticas y la Oficina Nacional de Estadísticas. La principal tarea de esta última es gestionar y analizar los datos estadísticos de Cuba, con el objetivo de contribuir a la toma de decisiones en los principales sectores socioeconómicos del país. Con el paso de los años la información que se recoge en este centro se ha incrementado, siendo necesario transformarla en conocimiento útil. Al Centro de Tecnologías de Gestión de Datos, perteneciente a dicha universidad se le dio la tarea de construir un almacén de datos que integre toda esta información, dicho almacén recibió el nombre de Sistema de Información de Gobierno. En el presente Trabajo de Diploma se realizó el mercado de datos para las series históricas de población, el cual formará parte del Sistema de Información de Gobierno, y apoyará la toma de decisiones sobre el control de la población. Para esto se hizo un estudio y se caracterizaron las metodologías, herramientas y tecnologías utilizadas en el desarrollo del mercado de datos. Se realizó el análisis y el diseño, donde se obtuvieron entre otros artefactos la especificación de requerimientos y se diseñaron los subsistemas de integración y visualización. Luego se implementaron dichos subsistemas y se hicieron pruebas para validar lo implementado, aplicando lista de chequeo y casos de pruebas.

PALABRAS CLAVES:

Almacén de datos, mercado de datos, Oficina Nacional de Estadísticas, series históricas, Sistema de Información de Gobierno.

Tabla de contenido

Introducción

Actualmente las Tecnologías de la Información y las Comunicaciones (TICs) tienen un desarrollo vertiginoso. La mayoría de los campos de la sociedad poseen una nueva forma de solucionar los problemas que se presentan, la estadística no está exenta de estos avances y transformaciones. El uso de las tecnologías se manifiesta progresivamente como una necesidad en este sector, ya que los datos estadísticos dentro de la infraestructura de un país constituyen el eslabón fundamental para la toma de decisiones en los principales sectores socioeconómicos.

En Cuba se manifiesta también este auge tecnológico, siendo la informática un factor importante para su desarrollo económico. Existen numerosos centros investigativos, dedicados a la producción de aplicaciones informáticas para contribuir con la informatización de todo el país. Una de estas instituciones es la Universidad de las Ciencias Informáticas (UCI), la cual cuenta con pequeños centros de desarrollo enfocados a soluciones informáticas en una rama determinada. El Centro de Tecnologías de Gestión de Datos (DATEC) es uno de los centros de desarrollo de la UCI, en el cual el departamento: Almacenes de Datos, se encuentra trabajando en conjunto con la Oficina Nacional de Estadísticas (ONE).

La ONE es el órgano rector de la estadística en Cuba y la responsable de gestionar los principales indicadores de los diferentes sectores del país. Tiene una estructura institucional distribuida territorialmente en las provincias y municipios de la isla. Existen 16 oficinas provinciales: una en cada provincia, incluyendo otra en el municipio especial Isla de la Juventud y otra adicional en Ciudad de La Habana. A su vez, 169 oficinas municipales, subordinadas a las provinciales, las cuales son las encargadas de interactuar directamente con los Centros Informantes (CI), siendo estos el último eslabón de la cadena de la actividad estadística [1]. Todos los centros asociados reciben atención administrativa y metodológica por la ONE, la cual mediante su Sistema Estadístico Nacional (SEN) ejerce una adecuada dirección, ejecución y control de la captación de las cifras económicas y sociales, así como su difusión de acuerdo con los requerimientos de la economía y las demás necesidades del país en cuanto a información estadística.

Producto de la gran cantidad de información que recibe la ONE diariamente, la cual se ha acumulado con el paso de los años, se torna engorroso el proceso de análisis por parte de los especialistas de la institución. Las principales deficiencias que existen en la gestión de estos datos son:

- La información se tiene en varios formatos: el departamento de estadísticas de la ONE, almacena la información en excel, archivos de texto, archivos DBF, formato duro (papeles), documentos Word, entre otros.

- No pueden ser consultadas a no ser por un especialista de la informática y de la información con alto conocimiento del negocio.
- Se generan ficheros anuales con los cuales se hace muy difícil la obtención de información.
- Los datos no están integrados, lo que atenta contra la calidad de estos: la integración se refiere al hecho de existir referencias a la misma información que usan diferente codificación o diferente cantidad de caracteres, y ocurre porque existen múltiples versiones de los mismos datos.

Debido a estas deficiencias, en el centro DATEC se crea un proyecto encargado de construir un almacén de datos con el objetivo de integrar la información, dicho almacén recibió el nombre de Sistema de Información de Gobierno (SIGOB). Este proyecto liberó, como parte de un Trabajo de Diploma precedente, el mercado de datos correspondiente al área de demografía. Las series históricas de población pertenecientes a esta área no fueron incluidas en dicho mercado, surgiendo dificultades para garantizar un mejor análisis de estos datos, que contienen información censal de Cuba en un período de tiempo y se registran en formato excel.

Por a la necesidad de dar solución a la problemática planteada, surge el siguiente **problema de la investigación:** La diversidad de información de los datos de las series históricas de población dificulta su adecuado análisis por parte de los especialistas.

La presente investigación tiene como **objeto de estudio** los almacenes de datos, enmarcado en el **campo de acción** mercado de datos para las series históricas de población del Sistema de Información de Gobierno.

El **objetivo general** de este trabajo es desarrollar el mercado de datos Series históricas de población para el Sistema de Información de Gobierno que contribuya al almacenamiento homogéneo de los datos, para un adecuado análisis de la información por parte de los especialistas.

En correspondencia con ello se plantean como **objetivos específicos:**

- Realizar el análisis y diseño del mercado de datos Series históricas de población.
- Implementar el mercado de datos Series históricas de población.
- Validar el mercado de datos Series históricas de población.

Para dar cumplimiento a los objetivos específicos se definen las siguientes **tareas de la investigación:**

- Caracterización de las metodologías, herramientas y tecnologías a utilizar para el desarrollo de almacenes de datos.

- Levantamiento de requisitos para definir las necesidades del cliente.
- Descripción de los casos de uso del mercado de datos para un mejor entendimiento de los casos de uso informativos, funcionales y no funcionales identificados.
- Definición de los hechos, las medidas y las dimensiones del mercado de datos para estructurar el modelo de datos.
- Diseño del modelo de datos para comprender la estructura de la base de datos.
- Definición de la arquitectura del sistema para establecer las bases del desarrollo del mercado de datos.
- Diseño del subsistema de integración para definir el flujo de datos desde los sistemas fuentes hacia el mercado de datos.
- Diseño del subsistema de visualización para definir el flujo de datos entre el sistema y el cliente.
- Diseño de los casos de pruebas para aplicarlos posteriormente durante la liberación el sistema.
- Implementación del modelo de datos para cumplir con el diseño de la estructura de la base de datos.
- Implementación del subsistema de integración para poblar el mercado de datos.
- Implementación del subsistema de visualización para gestionar los reportes candidatos necesarios y de esta forma satisfacer las necesidades del cliente.
- Aplicación de la lista de chequeo para comprobar el correcto funcionamiento de los subsistemas de integración y visualización.
- Aplicación de los casos de pruebas para comprobar el correcto funcionamiento de los reportes candidatos definidos en el sistema.

El Trabajo de Diploma está estructurado de la siguiente manera: introducción, cuatro capítulos, conclusiones, recomendaciones, referencias bibliográficas, bibliografía, anexos y glosario de términos.

Capítulo 1: Fundamentación teórica sobre almacenes de datos

Se realiza un análisis del estado del arte del objeto de estudio. Se caracterizan las metodologías, herramientas y tecnologías a utilizar en el desarrollo del mercado de datos.

Capítulo 2: Análisis y diseño del mercado de datos Series históricas de población

Se hace un estudio preliminar del negocio para realizar el levantamiento de requisitos, describir los casos de uso del mercado de datos e identificar las reglas del negocio. Se define la arquitectura, así como los hechos, las medidas y las dimensiones de dicho mercado. Además se confecciona la matriz bus que sirve de guía durante el diseño del modelo de datos.

Capítulo 3: Implementación del mercado de datos Series históricas de población

Se implementan los subsistemas de integración y visualización para los datos de las series históricas de población, teniendo en cuenta el análisis y el diseño del capítulo anterior.

Capítulo 4: Validación del mercado de datos Series históricas de población

Se realizan pruebas para validar la implementación de los subsistemas de integración y de visualización, en este caso se aplican la lista de chequeo y los casos de pruebas diseñados.

CAPÍTULO 1: FUNDAMENTACIÓN TEÓRICA SOBRE ALMACENES DE DATOS

1.1 Introducción

En este capítulo se hace un análisis del estado del arte del objeto de estudio; se presentan elementos conceptuales que permitirán comprender el problema de la investigación. Se realiza un análisis de la gestión de las series históricas de población en la ONE, así como la dificultad que trae para la toma de decisiones sobre el control de la población. Además, se caracterizan las metodologías, herramientas y tecnologías a utilizar en el desarrollo de almacenes de datos.

1.2 Proceso de digitalización de los datos de las series históricas de población que se almacenan en la Oficina Nacional de Estadísticas

Las series históricas de población se almacenan en tablas, las cuales se encuentran en formato excel. Estos datos dependen de información censal de un período de tiempo. Para analizarlos es necesaria la participación de varios especialistas demográficos, que realizan el análisis de los datos de forma manual con la ayuda de herramientas informáticas. Este es un trabajo minucioso donde se debe revisar tabla por tabla para detectar incongruencias en los datos calculados, en un intervalo de tiempo entre un censo y otro. Todo este proceso descrito anteriormente hace que se dificulte la manera de realizar los análisis estadísticos sobre la población cubana; corriéndose el riesgo de que se pierda información útil al no contar con una herramienta informática que contribuya a mejorar la eficiencia del tratamiento de la información. Por todas estas razones es objetivo del presente trabajo de investigación, desarrollar un mercado de datos con el objetivo de permitirle a los especialistas una mejor visualización de los datos y apoyar la toma de decisiones.

1.3 Almacenes de datos

Un almacén de datos (datawarehouse, DWH por sus siglas en inglés) es una base de datos (BD) corporativa que se caracteriza por integrar y depurar información desde una o varias fuentes de datos. De esta manera se facilita el análisis de la información desde infinidad de perspectivas y con grandes velocidades de respuesta. Por tales razones se podría ver también a un almacén de datos como una enciclopedia especializada en los temas que afectan el quehacer de la empresa [2].

Características de un almacén de datos:

- Orientado a temas: los datos en la BD están organizados de manera que todos los elementos de datos relativos al mismo evento u objeto del mundo real queden unidos entre sí.
- Variante en el tiempo: los cambios producidos en los datos a lo largo del tiempo quedan registrados para que los informes que se puedan generar reflejen esas variaciones.

- No volátil: la información no se modifica ni se elimina, una vez almacenado un dato, éste se convierte en información de sólo lectura y se mantiene para futuras consultas.
- Integrado: la BD contiene los datos de todos los sistemas operacionales de la organización y dichos datos deben ser consistentes [3].

1.3.1 Mercado de datos

El concepto de mercado de datos o datamart en inglés como también es conocido, se utiliza para definir un almacén de datos "a la medida" de las necesidades de un departamento o área particular en una institución, es decir, sólo contiene el subconjunto de datos del almacén de datos corporativo que son de interés para análisis en esa área.

Un concepto más formal sería: *"es una BD departamental, especializada en el almacenamiento de los datos de un área de negocio específica. Se caracteriza por disponer la estructura óptima de datos para analizar la información al detalle desde todas las perspectivas que afecten a los procesos de dicho departamento. Un datamart puede ser alimentado desde los datos de un almacén, o integrar por sí mismo un compendio de distintas fuentes de información"* [4].

De forma sencilla, se puede decir que un mercado de datos es como un almacén de datos que se puede consultar rápidamente, pero a un nivel más pequeño (áreas), mientras que el almacén es a nivel de toda la empresa.

OnLine Transaction Processing

Para crear el mercado de datos de un área funcional de la empresa es preciso encontrar la estructura óptima para el análisis de su información, que puede estar montada sobre una BD OLTP (OnLine Transaction Processing, OLTP por sus siglas en inglés), o sobre una BD OLAP (OnLine Analytical Processing, OLAP por sus siglas en inglés). La designación de una u otra dependerá de los datos, los requisitos y las características específicas de cada departamento.

Características de un mercado de datos:

- ✓ Se centran en los requisitos de los usuarios asociados a un departamento o área de negocio concretos.
- ✓ A diferencia de los almacenes de datos, los mercados no contienen datos operacionales detallados.
- ✓ 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 [5].

1.4 Experiencias de uso de los almacenes de datos

Hoy día los almacenes de datos son muy usados en todo el mundo, ya que constituyen uno de los soportes fundamentales para el proceso de toma de decisiones. Numerosos son los almacenes de datos que han tenido experiencias exitosas. Algunos ejemplos de estos se mencionan a continuación:

La compañía Wall-Mart, considerada la empresa más grande a nivel mundial cuenta con el almacén de datos más voluminoso y poblado del mundo; el cual usa para tomar decisiones acerca de todos los procesos que realizan en el mercado internacional, elevar su economía y mantenerse en competencia respecto a otras compañías [6].

Igualmente, Twentieth Century Fox utiliza la información relacionada con las películas que se proyectan en distintos lugares de los Estados Unidos para predecir qué actores, argumentos y filmes serán populares, con el objetivo de ganar audiencia en sus producciones [7].

En Cuba también se hace uso de esta herramienta para la toma de decisiones. El grupo empresarial CIMEX, caracterizado desde su creación por el crecimiento constante y la estabilidad financiera, tanto dentro como fuera del país, utiliza un almacén de datos para la gestión de inventarios. En la UCI se ha desarrollado un almacén de datos para la toma de decisiones en cuanto al consumo energético. También el Centro de Inmunología Molecular utiliza un almacén de datos, desarrollado por la UCI para analizar los ensayos clínicos que se gestionan en dicho centro.

1.5 Etapas de desarrollo de un almacén de datos

1.5.1 Etapa de diseño

La etapa de diseño comprende el análisis del negocio en cuestión, con el objetivo de identificar los requerimientos del negocio. De esta forma especificar el modelo de datos dimensional, las dimensiones, las medidas y las tablas de hechos del mercado de datos, y a partir de esto se debe especificar además el modelo físico de los datos.

El modelado multidimensional es una técnica de diseño lógico que busca presentar los datos en un estándar y facilitar una recuperación adecuada de estos. Los datos son almacenados como hechos y dimensiones en un modelo de datos relacional.

Se le llama **hecho** a una operación que se realiza en el negocio la cual está estrechamente relacionada con el tiempo y es objeto de análisis para la toma de decisiones. También puede verse como un valor numérico que representa una actividad específica casi siempre con cifras que se suman entre sí.

Se conoce como **dimensión** a la característica de un hecho que permite su análisis posterior en el proceso de toma de decisiones y brinda una perspectiva adicional a un hecho dado. Son agrupaciones lógicas de atributos con un significado común y atómico. Por lo general son estables.

Una **medida** es un atributo numérico de un hecho que representa el rendimiento o comportamiento del negocio relativo a la dimensión. Representan los valores que son analizados. Deben ser numéricas ya que estos valores son las bases sobre las cuales el usuario puede realizar cálculos.

Las **tablas hechos** son las que contendrán los hechos a través de los cuales se construirán los indicadores de estudio. Las **tablas dimensiones** contienen la descripción de atributos y características asociadas con medidas de eventos tangibles y específicos. El **modelo físico de los datos** se obtiene a partir del modelo dimensional, en él se especifican los tipos de datos de las variables que fueron definidas anteriormente y la cardinalidad entre las tablas.

Tipología de esquema:

La tipología de esquema no es más que la forma en la cual se va a estructurar el depósito de datos. Es muy importante definir que tipología se empleará, ya que esta decisión afecta considerablemente la elaboración de los modelos dimesional y físico. Generalmente se utiliza la que se adapte mejor a los requerimientos y necesidades del cliente.

Tipos de esquemas:

- Esquemas en estrella: la base de datos relacional consiste en una tabla simple de hecho relacionada con las tablas de dimensiones, las cuales no se relacionan entre sí. Cada tupla de la tabla de hechos incluye las medidas consideradas y una referencia a cada dimensión.

- Esquema en bola de nieve o copo de nieve: es una variedad más compleja del esquema estrella. Lo que distingue a la arquitectura en copo de nieve del esquema estrella, es que las tablas de dimensiones pueden estar relacionadas entre sí, o sea existen caminos alternativos en ellas.

- Constelaciones de hechos: constituye una generalización de los esquemas en estrella y bola de nieve, que puede obtenerse con la inclusión de distintas tablas de hechos que comparten todas o algunas de las dimensiones.

1.5.2 Extracción, transformación y carga

La etapa de diseño sirve como punto de partida para la implementación del proceso de extracción, transformación y carga (ETL por sus siglas en inglés). Este proceso consiste en extraer los datos

desde las diversas fuentes que sean necesarias, transformarlos para resolver posibles problemas de inconsistencias entre ellos y finalmente, después de haberlos depurado se procede a cargarlos en un depósito de datos. En síntesis, las funciones específicas de la etapa de ETL son tres: la extracción, transformación y carga de los datos.

Extracción

La primera parte del proceso ETL consiste en extraer los datos desde los sistemas de origen. La mayoría de los proyectos de almacenamiento de datos fusionan datos provenientes de diferentes sistemas de origen. Cada sistema separado puede usar una organización diferente de los datos o formatos distintos. Los formatos de las fuentes normalmente se encuentran en BD relacionales o ficheros planos, pero pueden incluir BD no relacionales u otras estructuras diferentes. La extracción convierte los datos a un formato preparado para iniciar el proceso de transformación.

Transformación

Esta fase 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.

Carga

Este proceso es el responsable de cargar la estructura de datos del DWH con aquellos datos que han sido transformados y que residen en el almacenamiento intermedio y aquellos datos de los OLTP que tienen correspondencia directa con el depósito de datos. Se debe tener en cuenta, que los datos antes de cargarse en el almacén de datos, deben ser analizados con el propósito de asegurar su calidad, ya que este es un factor clave, que no debe dejarse de lado [2].

1.5.3 Inteligencia de negocios

La inteligencia de negocios conocida también como Business Intelligence (BI, por sus siglas en inglés) debe ser parte de la estrategia empresarial, esta le permite optimizar la utilización de recursos, monitorear el cumplimiento de los objetivos de la empresa y la capacidad de tomar buenas decisiones para así obtener mejores resultados.

La base de la inteligencia de negocios es el **análisis de datos,** proceso en el que a través de las distintas técnicas del análisis, como: OLAP, la minería de datos y los reportes y consultas, se le da

valor real de los datos al extraer de ellos la información requerida para auxiliar la toma de decisiones y mostrarla de manera entendible [8].

Minería de datos

Es *"…el proceso de descubrir conocimientos interesantes, como patrones, asociaciones, cambios, anomalías y estructuras significativas a partir de grandes cantidades de datos, almacenadas en bases de datos, DWH, o cualquier otro medio de almacenamiento de información*" [9].

Reportes y consultas

Esta técnica no es más que el uso de herramientas destinadas a la producción de consultas y reportes. Estas ofrecen a los usuarios la posibilidad de generar informes del área de interés del negocio que se esté analizando, a través de pantallas gráficas intuitivas. El usuario únicamente debe seleccionar las opciones que le brindan para especificar los elementos de datos, sus condiciones, criterios de agrupación y demás atributos que se consideren significativos [10].

OLAP

Es la técnica que permite que la información sea vista multidimensionalmente, a través de cubos con categorías descriptivas (dimensiones) y valores cuantitativos (medidas). Se pueden definir varios sistemas OLAP, dependiendo de las técnicas que se utilicen a la hora de obtener los datos y la forma en la que están estructurados y almacenados, entre los principales se encuentran:

- **ROLAP:** almacena los datos en un motor relacional. La arquitectura está compuesta por un servidor de banco de datos relacional y el motor OLAP, que se encuentra en un servidor dedicado. La principal ventaja de esta arquitectura es que permite el análisis de una enorme cantidad de datos.
- **MOLAP:** esta implementación OLAP almacena los datos en una BD multidimensional.

Se decidió usar la tecnología OLAP porque a pesar de ser similar a la minería de datos, cuenta con la capacidad para analizar los datos en múltiples dimensiones de análisis, de tendencias y pronósticos (cubos). Además, este modelo de datos libra a los especialistas de la ONE de formular consultas complejas, por el hecho de concebir una única consulta MDX (acrónimo de MultiDimensional eXpression) sobre el cubo OLAP de datos previamente diseñado, permitiéndoles obtener reportes donde la sencillez y el alto nivel de detalles sean elevados. Otra ventaja que ofrece el modelo dimensional, es poder cambiar de datos resumidos a datos detallados y filtrar o rebanar los datos en subconjuntos significativos; además permiten explicar la causa y los efectos de las ocurrencias y las tendencias. Específicamente se decide usar la arquitectura ROLAP, ya que accede directamente a los

datos del almacén, soporta técnicas de optimización de accesos, tales como particionado de los datos a nivel de aplicación, soporte a la desnormalización y uniones múltiples, para acelerar las consultas.

1.6 Metodologías para el desarrollo de almacenes de datos

Para desarrollar un almacén de datos, existen dos estrategias fundamentales dadas por quienes se conocen como los padres de los almacenes de datos: Bill Inmon y Ralph Kimball.

Bill Inmon define una estrategia descendente (top-down), donde los mercados de datos se crean después de desarrollar el almacén de datos de la empresa.

Ralph Kimball plantea el desarrollo ascendente (bottom-up), donde se crean los mercados de datos de cada departamento para luego unirlos y conformar el almacén de datos de toda la empresa.

De estas dos estrategias, la de Kimball es la más aceptada en todo el mundo como la más efectiva para desarrollar una solución de construcción de almacenes de datos. Además es de fácil comprensión y rápida de implementar por etapas.

La de Inmon por el contrario puede tener una implementación mucho más tardada, y es recomendada cuando se hace demasiado difícil representar el modelo a través de dimensiones y la complejidad de la solución se hace demasiado grande [7].

Además de estas tendencias existen algunas metodologías para el desarrollo de almacenes de datos, como por ejemplo Hefesto, Desarrollo de almacenes de datos dirigidos por modelos, Data Warehouse Engineering Process (DWEP), Rapid Warehousing Methodology (RWM) y el Ciclo de vida Kimball, entre otras.

1.6.1 Ciclo de vida Kimball

A continuación se presenta una figura donde aparece el ciclo de vida Kimball (ver figura 1):

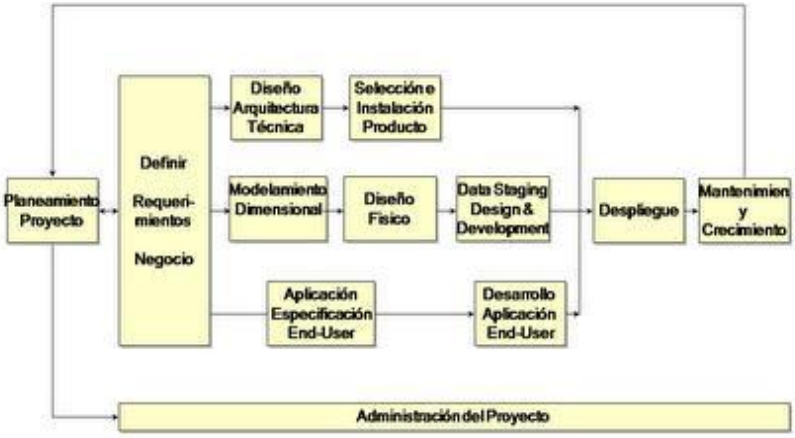


Figura 1: Ciclo de vida Kimball

El ciclo de vida Kimball comienza con una planificación de proyecto, en la cual se define el alcance, se identifican y programan las tareas, se planifica el uso de los recursos, conformando con todo esto el plan de proyecto. En la segunda etapa de este ciclo se definen los requerimientos del negocio. Luego de definir los requerimientos del negocio, el proyecto se enfoca en tres líneas concurrentes: tecnología, datos y aplicaciones de la inteligencia de negocios (ver figura 1).

Tomando como base esta metodología de Kimball, DATEC propone un modelo para este tipo de soluciones, adaptada a su proceso de desarrollo, el cual está basado en Líneas de Productos de Software, y en los lineamientos de calidad exigidos por la UCI, denominándose éste: Modelo para el Desarrollo de Soluciones de Almacenes de Datos e Inteligencia de Negocio en DATEC.

1.6.2 Modelo de DATEC

El modelo de DATEC denominado Modelo para el Desarrollo de Soluciones de Almacenes de Datos e Inteligencia de Negocio en DATEC, cubre todas las fases por las que pasa la construcción de un almacén de datos, desde el levantamiento de información inicial hasta la capa de visualización. Este modelo reúne elementos de varias metodologías de desarrollo de proyectos de integración de datos, toma como base la Metodología de Kimball. En una primera fase contempla el levantamiento de información a nivel de negocio para identificar los posibles indicadores y aspectos a medir en los análisis, que luego de algunas transformaciones se convierten en los requerimientos de información de entrada y de salida para la solución de integración.

De forma paralela a esta actividad se lleva a cabo un estudio de las fuentes de datos que soportan los datos a cargar. Finalizadas estas dos tareas, se corrobora que la información levantada sobre las necesidades de los clientes esté realmente almacenada en las fuentes correspondientes, para posteriormente, teniendo los requerimientos informativos correctamente definidos, proceder a diseñar la solución de almacén de datos. Una vez diseñada la estructura del almacén, se realiza la carga de los datos desde las fuentes y posteriormente se implementan los requerimientos de inteligencia de negocio identificados en el levantamiento de información inicial. Las actividades y artefactos de la solución son realizados por cuatro grupos que conforman la línea, especializados en componentes específicos de la solución.

Ventajas del Modelo para el Desarrollo de Soluciones de Almacenes de Datos e Inteligencia de Negocio en DATEC:

- La solución completa se puede implementar en poco tiempo.
- Cuenta con mayor velocidad de respuesta al cliente.
- Los productos son más comprensibles para los usuarios.

- Es resistente y tolerante ante los cambios [11].

En el presente Trabajo de Diploma se decide utilizar el modelo propuesto por DATEC, ya que toma como base la metodología de Ralph Kimball, cubriendo todas las fases de construcción de un almacén de datos, además de ser considerada la más adecuada para la situación que se plantea en esta investigación.

1.7 Herramientas de modelado

Al aplicar la Ingeniería de Software durante el proceso de desarrollo de un producto informático es necesario modelar artefactos, para esto existen varias herramientas automatizadas que utilizan técnicas de diseño y metodologías bien definidas. Ejemplo de ello, son las herramientas CASE (Computer Aided Software Engineering), que constituyen un conjunto de ayudas para el desarrollo de programas informáticos.

Entre las herramientas CASE orientadas a UML se encuentran:

- Rational Rose
- ArgoUML
- Poseidón
- Visual Paradigm
- MagicDraw UML
- Borland Together

En la presente investigación se decidió utilizar Visual Paradigm en su versión 6.4, ya que es multiplataforma y permite su uso en cualquier sistema operativo. Está disponible en varios idiomas, en conjunto con esto, es fácil de instalar y fácil de actualizar. Además admite compatibilidad con las demás versiones.

1.7.1 Visual Paradigm

Visual Paradigm es una herramienta de modelado que utiliza UML como lenguaje de modelado y posibilita una rápida construcción de las aplicaciones con alta calidad. Es factible a la hora de dibujar diagramas de clases, y generar script para diferentes sistemas gestores de base de datos. Además, permite una integración con sistemas de control de versiones que almacenan centralmente los artefactos y realizan un seguimiento de los cambios realizados sobre un proyecto. Los desarrolladores lo utilizan para facilitar el modelado simultáneo, almacenar los archivos de proyectos y hacer un seguimiento de los cambios.

1.8 Sistema gestor de base de datos

Los Sistemas Gestores de Bases de Datos (SGBD) son un elemento clave dentro del mundo de la información ya que contienen las rutinas necesarias para el manejo de los datos: dígase definición, construcción y manipulación. Permiten la eliminación y actualización de registros, la combinación con otras bases de datos y la generación de informes impresos.

En la presente investigación se decide usar como SGBD PostgreSQL en su versión 8.4, debido a que es una herramienta libre muy potente. Posee numerosas ventajas, tales como gran soporte profesional tanto de la comunidad de usuarios como de empresas especializadas. Además es multiplataforma, por lo que puede ser usado en cualquier sistema operativo.

1.8.1 PostgreSQL

PostgreSQL es un potente SGBD objeto-relacional libre, liberado bajo la licencia BSD (del inglés Berkeley Software Distribution). Como muchos otros proyectos Open Source, el desarrollo de PostgreSQL no es manejado por una sola compañía sino que es dirigido por una comunidad de desarrolladores y organizaciones comerciales las cuales trabajan en su desarrollo, dicha comunidad es denominada el PostgreSQL Grupo Global de Desarrollo (PGDG), sus siglas en inglés se definen como: PostgreSQL Global Development Group [12].

En cuanto a la arquitectura de la herramienta, PostgreSQL utiliza un modelo cliente/servidor. El servidor y los clientes pueden estar ejecutándose en diferentes equipos, por lo tanto PostgreSQL permite la comunicación entre estos procesos a través del protocolo TCP/IP. El servidor soporta múltiples conexiones concurrentes de clientes. Con este propósito ejecuta nuevos procesos para cada conexión. Este proceso es transparente para los usuarios.

Ventajas encontradas en PostgreSQL:

- Soporta lenguajes: PHP, C, C++, Perl y Python.
- Drivers: ODBC (del inglés Open Database Connectivity), JDBC (del inglés Java Database Connectivity) y.Net.
- Soporta: *triggers*, procedimientos almacenados, funciones, secuencias, relaciones, reglas, tipos de datos definidos por el usuario, vistas y vistas materializadas.
- Soporte de tipos de datos de SQL92, SQL99 y SQL2003.
- Soporte de protocolo de comunicación encriptado por SSL.
- Máximo de bases de datos: ilimitado.
- Máximo de tamaño de tabla: 32 TB.
- Máximo de tamaño de registro: 1.6 TB.

- Máximo de tamaño de campo: 1GB.
- Máximo de registros por tabla: ilimitado.
- Máximo de campos por tabla: 250 a 1600 (depende de los tipos usados).
- Máximo de índices por tabla: ilimitado.
- Número de lenguajes en los que se puede programar funciones: aproximadamente 10 (pl/pgsql, pl/java, pl/perl, pl/python, tcl, pl/php, C, C++ y Ruby [2].

PostgreSQL requiere de una herramienta para administrar y desarrollar las BD, en la presente investigación se decide utilizar con tal fin PgAdmin. Este es un motor de bases de datos de código abierto muy avanzado. Es multiplataforma y también funciona con otros motores comerciales basados en PostgreSQL. Se diseña para responder a las necesidades de la mayoría de los usuarios, desde escribir simples consultas SQL hasta desarrollar BD complejas. La interfaz gráfica soporta todas las características de PostgreSQL y facilita la administración. Está disponible en más de una docena de lenguajes y para varios sistemas operativos, incluyendo Microsoft Windows, Linux, FreeBSD, Mac OSX y Solaris.

1.9 Técnicas de captura de requisitos

Desde el inicio del desarrollo de sistemas, los ingenieros han presentado dificultad con la identificación de los requisitos. Esto no es un proceso que se pueda determinar matemáticamente, sino un proceso en el cual los datos son extraídos de las personas y pueden variar; dependiendo de la persona a la que se esté consultando. Se le ha dado solución a esto a través del desarrollo de técnicas que permitan hacer este proceso de una forma más eficiente y segura. Algunas de las técnicas más utilizadas para la captura de requisitos son las entrevistas, tormenta de ideas, cuestionarios, observaciones, discusiones, análisis de protocolo y casos de uso.

En la presente investigación se utilizaron las entrevistas, ya que permiten comprender mejor el problema y los objetivos de la solución buscada, de esta forma se obtiene una amplia visión del trabajo y de las necesidades del usuario. En conjunto con las entrevistas se utilizaron las tormentas de ideas, esta consiste en reuniones de grupos, con la ventaja de que los participantes muestran sus ideas de forma libre. Otras técnicas combinadas con las mencionadas anteriormente fueron las observaciones y discusiones con los especialistas, para aclarar dudas y realizar una correcta captura de los requisitos.

1.10 Herramientas informáticas para el proceso de extracción, transformación y carga

Por la gran importancia que tiene la reducción de costos y la independencia con respecto a los proveedores, se decidió utilizar herramientas de software libre para desarrollar el proceso de ETL.

Dentro de las herramientas informáticas que existen, se encuentran Kettle, Scriptella, Octopus y Talend. En la presente investigación se decidió utilizar la versión 4.0 de Kettle para el desarrollo del proceso de ETL, ya que es una herramienta libre muy potente, así como una de las más antiguas y utilizadas por los usuarios. Producto a esto tiene gran soporte técnico y los usuarios comparten muchos consejos y trucos en los foros. Además, Kettle es la más completa de las mencionadas por la gran cantidad de conectores que posee y la posibilidad de crear flujos de trabajo integrados con transformaciones de datos de manera muy sencilla y funcional.

1.10.1 Kettle

Conocido actualmente como Pentaho Data Integration, es un proyecto belga de código abierto, ahora adoptado por Pentaho BI, que incluye un grupo de herramientas para realizar el proceso de ETL. Uno de sus objetivos es que dicho proceso sea más fácil de generar, mantener y desplegar. Kettle está compuesto por cuatro herramientas:

- **SPOON**: permite diseñar de forma gráfica las transformaciones ETL.
- **PAN**: ejecuta un conjunto de transformaciones diseñadas con SPOON.
- **CHEF**: permite diseñar la carga de datos incluyendo un control de estado de los trabajos.
- **KITCHEN**: permite ejecutar los trabajos *batch* diseñados con CHEF.

El uso de esta herramienta permite evitar grandes cargas de trabajo manual frecuentemente difícil de mantener y de desplegar. Además, es una herramienta que permite definir transformaciones de forma gráfica, interconectando bloques que tienen diversas funciones. Es extremadamente versátil, ya que se tienen bloques que permiten leer y escribir de cualquier BD, fichero excel, Access y otros que permiten operar con los campos renombrando, normalizando, calculando campos en función de otros, mapeando valores, realizando búsquedas auxiliares en BD y normalizando los datos de distintas filas en una sola. Las transformaciones que se hacen con el Kettle se guardan en un fichero *ktr* que luego puede ser ejecutado mediante líneas de comandos o un fichero *batch* [13].

Ventajas de Kettle:

- Funciona en Windows, Unix y Linux.
- Tiene una interfaz gráfica con indicadores de las trasformaciones.
- Es una aplicación implementada en Java con algunas características avanzadas en JavaScript.
- Ofrece una licencia púbica GPL (del inglés General Public License).
- Basada en metadatos.
- Como soporte se encuentran los foros de Pentaho y la comunidad Pentaho.

- Soporta Oracle, DB2.SQL Server, Sybase así como MySQL y Postgres. También soporta la conectividad con SAP.
- Con respecto a las escalabilidad, soporta la arquitectura de procesamiento en paralelo para distribuir las tares de ETL a través de múltiples servidores.
- Basado en dos tipos de objetos: transformaciones (colección de pasos en un proceso ETL) y trabajos (colección de transformaciones).

1.10.2 DataCleaner

Antes de iniciar el proceso de ETL, es necesario realizar la limpieza y estandarización de los datos, de esta forma se pueden detectar y corregir errores en la información a integrar, con tal fin se decide usar DataCleaner. Esta herramienta es una aplicación de código abierto para el perfilado, validación y comparación de los datos. Es muy fácil de usar, genera sofisticados informes y gráficos que permiten a los usuarios determinar el nivel de calidad de los datos, identificar y analizar la estructura del origen de datos. Es una alternativa libre para la metodología de administración de datos, para proyectos de almacenes de datos, búsquedas estadísticas, actividades de preparación de ETL y más.

1.11 Herramientas informáticas para la inteligencia de negocios, aplicando técnicas OLAP

Pentaho BI Server

Pentaho BI Server ofrece soporte y una infraestructura para solucionar los problemas de negocios con soluciones de inteligencia empresarial. Incluye servicios básicos como autenticación, registro, auditoria, servicios web y motor de reglas. También un motor de soluciones unido a generador de reportes, un tablero de comandos, análisis y componentes de minería de datos. La arquitectura con plug-in permite la integración de aplicaciones de terceros por los usuarios finales, así como de los equipos originales. Funciona como un sistema de administración web de informes, un servidor de integración de aplicaciones y un motor de flujo de trabajo ligero. Su diseño permite la integración de forma muy fácil a cualquier proceso del negocio [13].

Pentaho Schema Workbench

Shema Workbench es un entorno visual para el desarrollo y prueba de cubos OLAP. Con esta aplicación, se puede configurar una conexión JDBC como el modelo físico, para luego elaborar el esquema lógico de manera simple y efectiva; para ello el entorno de trabajo de la herramienta ofrece un editor de esquemas con la fuente de datos subyacente para su validación; permite la ejecución de consultas MDX contra el esquema, la BD y la navegación por la BD subyacente [13].

Servidor Mondrian OLAP

Para obtener la funcionalidad de procesamiento OLAP la suite utiliza otras dos aplicaciones: el servidor OLAP Mondrian, que combinado con Jpivot, permiten realizar consultas a almacenes de datos, que los resultados sean presentados mediante un navegador de modo que el usuario pueda realizar las actividades típicas de navegación. Mondrian utiliza MDX como lenguaje de consulta, que fue un lenguaje propuesto por Microsoft. Funciona sobre las bases de datos estándar del mercado: Oracle, DB2, SQL-Server, MySQL, PostgreSQL, lo cual habilita y facilita el desarrollo del negocio basado en la plataforma Pentaho.

Servidor web Apache Jakarta Tomcat

La suite de Pentaho no incluye esta herramienta, pero si la utiliza como servidor web. Tomcat es un servidor de código abierto, un contenedor de aplicaciones web basadas en Java que fue creado para ejecutar Servlets y Java Server Page, (JSP por sus siglas en inglés), de aplicaciones web. Existe en el marco del subproyecto Apache Jakarta, donde es apoyado y reforzado por un grupo de voluntarios de la comunidad de código abierto de Java.

En la presente investigación se decide utilizar el servidor Mondrian OLAP (versión 3.0.4), ya que además de las características mencionadas anteriormente, tiene un alto desempeño, reflejado en el análisis interactivo de grandes o pequeños volúmenes de información. Este debe ser ejecutado sobre un servidor web que soporte el lenguaje JSP, por lo cual se decidió utilizar el servidor web Apache Jakarta Tomcat en su versión 5.5. Para el diseño de los cubos que se cargarán en el servidor Mondrian se propone utilizar la herramienta Pentaho Shema Workbench (versión 3.2.0). Esta posee un entorno visual muy cómodo para el desarrollo y prueba de los cubos OLAP, que soporta además las consultas MDX, permitiendo la realización de pruebas y corrección de consultas sobre los cubos.

1.12 Conclusiones del capítulo

En este capítulo se mostró una panorámica general del proceso de desarrollo de un almacén de datos, así como una caracterización de las metodologías, herramientas y tecnologías que se utilizarán en este Trabajo de Diploma. Luego de esta investigación se arribaron a las siguientes conclusiones:

- Se decidió adoptar el Modelo para el Desarrollo de Soluciones de Almacenes de Datos e Inteligencia de Negocio propuesta por DATEC la cual toma como base la metodología de Ralph Kimball.
- Se decidió utilizar Visual Paradigm para el Lenguaje de Modelado Unificado en su versión 6.4 como herramienta de modelado, en el caso del lenguaje de modelado se hará uso de su versión 2.0.

- Se decidió utilizar PostgreSQL como SGBD en su versión 8.4.
- Se decidió utilizar herramienta informática Kettle en su versión 4.0 de la suite Pentaho, para la implementación del proceso de ETL.
- Para el análisis de la información se decidió utilizar el Pentaho BI Server en su versión 3.6.0.
- Se decidió hacer uso del servidor Mondrian OLAP en su versión 3.0.4, que permitirá analizar el contenido del almacén de datos sin tener que definir consultas SQL.
- Para el diseño de los cubos OLAP se decidió utilizar la herramienta Pentaho Shema Workbench en su versión 3.2.0.
- Como servidor web se propone utilizar el Apache Jakarta Tomcat en su versión 5.5.

CAPÍTULO 2: ANÁLISIS Y DISEÑO DEL MERCADO DE DATOS SERIES HISTÓRICAS DE POBLACIÓN

2.1 Introducción

En este capítulo se hace un estudio preliminar del negocio y de la organización, con el objetivo de identificar las reglas del negocio, los requerimientos y los casos de uso del mercado de datos. Se describen brevemente los actores que van a interactuar con el sistema, así como sus funcionalidades. También se definen los hechos, las medidas y las dimensiones del mercado de datos y se diseña el modelo de datos a partir de la confección de la matriz bus.

2.2 Caracterización de las áreas de la organización

La principal tarea de la ONE es almacenar y analizar los datos estadísticos de Cuba, con el objetivo de contribuir a la toma de decisiones en los principales sectores socioeconómicos del país. Por la gran cantidad de información que se procesa en esta entidad, la información se agrupa en distintos departamentos o áreas, cada uno relacionado con un sector del país. En el área de demografía se analizan entre otras cosas las series históricas de población, donde los datos se encuentran almacenados en archivos excel y dependen de información censal de un período de tiempo. Para analizar las series históricas de población es necesaria la participación de varios especialistas, ya que es muy difícil consultar estos datos por sus características.

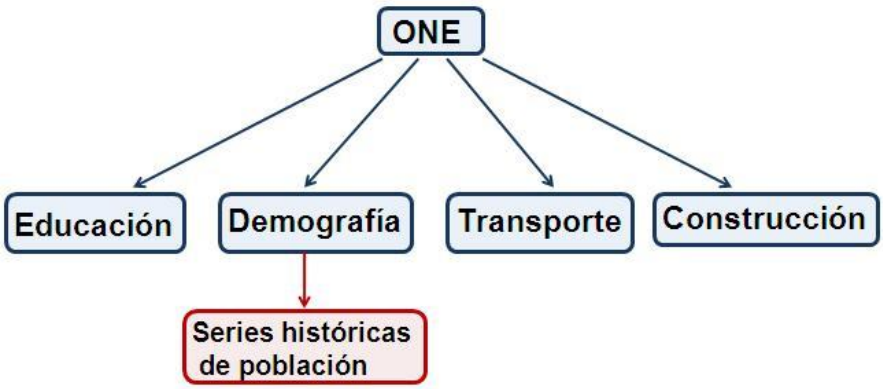


Figura 2: Áreas de la organización

2.3 Reglas del negocio

La función fundamental de las reglas del negocio es definir políticas o condiciones que se deben cumplir, por esta razón regulan aspectos del negocio. A continuación se muestran las reglas del negocio identificadas en los datos de las series históricas de población:

RN 1

Las medidas cant_nacimientos, cant_defunciones, pr_total_ambas_zonas, pr_varones_ambas_zonas, pr_hembras_ambas_zonas, pr_total_urbano, pr_varones_urbano, pr_hembras_urbano, pr_total_rural, pr_varones_rural, pr_hembras_rural, pm_total_ambas_zonas, pm_varones_ambas_zonas, pm_hembras_ambas_zonas, pm_total_urbano, pm_varones_urbano, pm_hembras_urbano, pm_total_rural, pm_varones_rural, pm_hembras_rural, pob_resid_capit_prov y pob_media_capit_prov son valores enteros, por lo que no pueden existir números con coma.

RN 2

Todas las medidas identificadas son valores positivos, por lo que solamente pueden tomar valores mayores o iguales que cero.

RN 3

No se admiten valores nulos.

2.4 Necesidades de información

Después de realizar un análisis minucioso del negocio en la ONE, específicamente de las series históricas de población se definen las siguientes necesidades de información:

- Obtener la cantidad de nacimientos según provincia de residencia y año.
- Obtener la tasa de natalidad según provincia de residencia y año.
- Obtener la cantidad de defunciones según provincia de residencia y año.
- Obtener la tasa de mortalidad según provincia de residencia y año.
- Obtener la tasa de crecimiento natural según provincia de residencia y año.
- Obtener la población residente según provincia, municipio y año, por sexo y zona de residencia.
- Obtener la población media según provincia, municipio y año, por sexo y zona de residencia.
- Obtener el total de población residente según capital provincial y año.
- Obtener el total de población media según capital provincial y año.

2.5 Especificación de requerimientos

2.5.1 Requisitos de información

RI1-Obtener información del total de nacimientos por provincias, en el tiempo.

RI2-Obtener información del total de la tasa de natalidad por provincias, en el tiempo.

RI3-Obtener información del total de defunciones por provincias, en el tiempo.

RI4-Obtener información del total de la tasa de mortalidad por provincias, en el tiempo.

RI5-Obtener información del total de tasa de crecimiento natural por provincias, en el tiempo.

RI6-Obtener información del total de población residente en ambas zonas por DPA, en el tiempo.

RI7-Obtener información del total de población residente varones en ambas zonas por DPA, en el tiempo.

RI8-Obtener información del total de población residente hembras en ambas zonas por DPA, en el tiempo.

RI9-Obtener información del total de población residente en zona urbana por DPA, en el tiempo.

RI10-Obtener información del total de población residente varones en zona urbana por DPA, en el tiempo.

RI11-Obtener información del total de población residente hembras en zona urbana por DPA, en el tiempo.

RI12-Obtener información del total de población residente en zona rural por DPA, en el tiempo.

RI13-Obtener información del total de población residente varones en zona rural por DPA, en el tiempo.

RI14-Obtener información del total de población residente hembras en zona rural por DPA, en el tiempo.

RI15-Obtener información del total de población media en ambas zonas por DPA, en el tiempo.

RI16-Obtener información del total de población media varones en ambas zonas por DPA, en el tiempo.

RI17-Obtener información del total de población media de hembras en ambas zonas por DPA, en el tiempo.

RI18-Obtener información del total de población media en zona urbana por DPA, en el tiempo.

RI19-Obtener información del total de población media varones en zona urbana por DPA, en el tiempo.

RI20-Obtener información del total de población media de hembras en zona urbana por DPA, en el tiempo.

RI21-Obtener información del total de población media en zona rural por DPA, en el tiempo.

RI22-Obtener información del total de población media varones en zona rural por DPA, en el tiempo.

RI23-Obtener información del total de población media de hembras en zona rural por DPA, en el tiempo.

RI24-Obtener información del total de población residente por capitales provinciales, en el tiempo.

RI25-Obtener información del total de población media por capitales provinciales, en el tiempo.

2.5.2 Requisitos funcionales

RF1- Autenticar usuario.
RF2 - Adicionar roles.
RF3 - Eliminar roles.
RF4 - Adicionar usuarios.
RF5 - Eliminar usuarios.
RF6 - Insertar reportes.
RF7 - Modificar reportes.
RF8 - Eliminar reportes.
RF9 - Realizar extracción de información de archivos excel de series históricas de población.
RF10 - Realizar transformación y carga de información de archivos excel de series históricas de población.
RF11- Abrir navegador OLAP.
RF12 - Mostrar editor MDX.
RF13 - Mostrar Padres.
RF14 - Ocultar repeticiones.
RF15 - Intercambiar ejes.
RF16 - Mostrar gráfico.
RF17 - Configurar gráfico.
RF18 - Configurar impresión.
RF19 - Exportar a PDF.
RF20 - Exportar a excel.
RF21 - Mostrar propiedades.
RF22 - Suprimir filas.
RF23 - Detallar miembros.
RF24 - Entrar en detalles.
RF25 - Mostrar datos de origen.

2.5.3 Requerimientos no funcionales

Requerimientos de usabilidad

RNF 1. Cumplir con las pautas de diseño de las interfaces.
El sistema debe tener una interfaz gráfica uniforme que incluya pantallas, menús y opciones. Las pautas de diseño se realizarán siguiendo la arquitectura de información definida durante el diseño del subsistema de visualización.

RNF 2. Mostrar los mensajes, títulos y demás textos que aparezcan en la interfaz del sistema en idioma español.
Los títulos de los componentes de la interfaz, los mensajes para interactuar con los usuarios y los mensajes de error, deben ser en idioma español y tener una apariencia uniforme en todo el sistema. Los mensajes de error deberán ser lo suficientemente informativos para dar a conocer la severidad del error.

RNF 3. Agilizar el acceso a los reportes del almacén de datos mediante la distribución de la información por áreas de análisis.
El usuario podrá acceder de manera rápida a la información que solicita en el área correspondiente de acuerdo al objetivo de su solicitud.

Requerimientos de fiabilidad

RNF 4. Asegurar la disponibilidad del sistema.
El sistema debe estar disponible durante el horario de trabajo. En caso de fallo debe ser capaz de recuperarse, teniendo en cuenta la complejidad y naturaleza de éste. El tiempo para su correcta recuperación fluctúa entre 10 minutos y 72 horas. Este tiempo comprende la solución al problema, así como su validación y prueba.

RNF 5. Garantizar la persistencia de la información.
Se debe realizar un respaldo total de los datos del almacén de datos con una frecuencia anual. Esta información se almacenará en el edificio correspondiente a la oficina de estadísticas de La Habana y será responsabilidad del grupo de administración de redes de la ONE.

Requerimientos de soporte

RNF 6. Lograr la homogeneidad de la estructura de los elementos definidos en el almacén.
Las estructuras del almacén de datos deben tener un nombre estándar teniendo en cuenta el tipo de estructura que sea. A continuación se definen convenciones de nombrado con el objetivo de manejar un vocabulario común en todo el almacén de datos, permitiendo un entendimiento claro y conciso por parte de los desarrolladores:

- Tablas de hechos: todas las tablas de hechos tendrán una cadena que demuestra que son hechos y el concepto que describen. Ejemplo hech_<concepto>.
- Tablas de dimensiones: todas las tablas de dimensiones tendrán una cadena que demuestra que son dimensiones y el concepto que describen. Ejemplo dim_<concepto>.
- Llaves primarias: todas las llaves primarias tendrán el nombre de la tabla a la que pertenecen y una cadena que demuestra que son llaves primarias. Ejemplo <nombre_tabla>_id.

Requerimientos de restricciones de diseño

RNF 7. Utilizar los lenguajes de programación definidos durante la investigación.

Como lenguaje dentro del sistema gestor de base de datos para la programación en el almacén de datos se utilizará PL/pgSQL. En la implementación de los procesos de integración de datos se utilizará el lenguaje JavaScript. También se hará uso del lenguaje MDX para realizar las consultas.

RNF 8. Utilizar el Sistema Gestor de Base de Datos definido durante la investigación.

El gestor de base de datos que se utilizará es PostgreSQL y como interfaz de administración de dicho gestor PgAdmin.

RNF 9. Utilizar la herramienta de integración de datos definida durante la investigación.

Para el proceso de integración de datos se usará la herramienta Pentaho Data Integrator.

RNF 10. Utilizar las herramientas para la implementación de la capa de inteligencia de negocios definidas durante la investigación.

De la suite Pentaho, se usarán los siguientes componentes:

- Schema Workbench: herramienta gráfica que se utiliza para construir el esquema multidimensional que soportará la creación de los reportes multidimensionales.
- Pentaho BI Server: servidor que se encarga de visualizar los reportes, tableros de control digital, controlar el acceso a la información y unificar en una solución de inteligencia de negocios el uso de las demás herramientas que componen la suite.
- Pentaho Administrator Console: herramienta para administrar el Pentaho BI Server, que permite la administración de las conexiones a las bases de datos, tareas programadas así como los roles y usuarios.

Para el uso de las herramientas anteriores se requiere la instalación de la máquina virtual de java (Java Virtual Machine 6.0).

Requerimientos para la documentación de usuarios y ayuda del sistema

RNF 11. Confección de un manual de usuario.

El sistema debe estar acompañado de un documento que guiará la ejecución del usuario teniendo en cuenta cada funcionalidad.

Interfaz

RNF 12. Acceso al sistema.

El usuario deberá acceder a la aplicación mediante el protocolo HTTP, usando preferiblemente el navegador web Firefox 2.0 en adelante.

Interfaces de usuario

RNF 13. Garantizar una interfaz amigable al usuario.

El sistema debe tener una interfaz amigable y sencilla de utilizar, teniendo en cuenta que los usuarios finales no son personas adiestradas en el campo de la informática.

Interfaces de hardware

RNF 14. Definir las interfaces de hardware que soportará el sistema.

El sistema podrá interactuar solamente con una interfaz de hardware: la impresora. Esta interacción se ocasionará cuando se necesite imprimir un reporte en formato físico. El acceso a la impresora será mediante el protocolo TCP/IP a través de la interfaz que ofrece el hardware.

RNF 15. Proporcionar características mínimas de hardware a las estaciones de trabajo.

Características de un cliente ligero.

RNF 16. Proporcionar características mínimas de hardware a los servidores.

Para lograr una explotación aceptable del sistema los servidores deben contar con los siguientes requerimientos de hardware:

- Windows server 2003.
- 1 GB RAM.
- 1 Microprocesador Core2Duo.

Interfaces de software

RNF 17. 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:

- Firefox 2.0 o superior.

- Java Virtual Machine 6.0 y Schema Workbench 3.2.0 en caso de que un usuario capacitado requiera la construcción de esquemas multidimensionales para el diseño de nuevos reportes.

Requerimientos legales de derechos de autor y otros

RNF 18. Entregar el sistema a la ONE.

El sistema debe ser transferido a la ONE mediante un proceso de transferencia una vez que esté en explotación, incluyendo el código fuente y la documentación correspondiente.

RNF 19. Requerimientos legales, de derecho de autor y otros.

No se hace solicitud de derecho de autor, patentes, marca comercial o complacencia con logotipo para el software, debido a que se usan soluciones con Licencia Pública General (GNU GPL por sus siglas en inglés), bajo el principio de software libre.

2.6 Casos de uso del sistema

2.6.1 Actores del sistema

Actor	Descripción
Analista	Es el encargado de analizar los datos.
Administrador	Administra los roles, los usuarios y los reportes, además de analizar los datos.
Administrador ETL	Realiza los procesos ETL.

Tabla 1: Actores del sistema

2.6.2 Diagrama de casos de uso del sistema

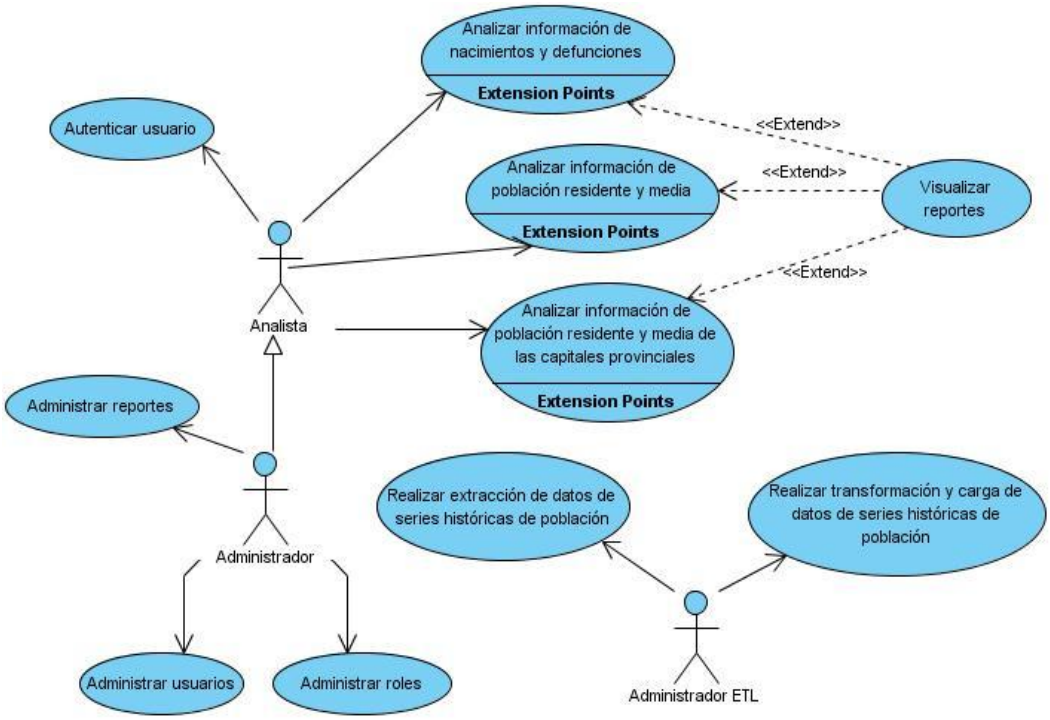


Figura 3: Diagrama de casos de uso del sistema

2.6.3 Especificación de casos de uso del sistema

Autenticar usuario:

Caso de Uso:	Autenticar usuario.
Tipo:	Funcional
Actores:	Analista, Administrador.
Resumen:	El CU inicia cuando el actor accede al sistema para autenticarse, introduce los datos, el sistema verifica que sean correctos y le permite acceder a la información según sus privilegios. Una vez que el actor queda autenticado en el sistema termina el CU.
Precondiciones:	-
Referencia	RF1
Prioridad	Crítico
Flujo Normal de Eventos	
Acción del Actor	**Respuesta del Sistema**
1. El actor ejecuta la aplicación.	2. El sistema muestra una interfaz para que

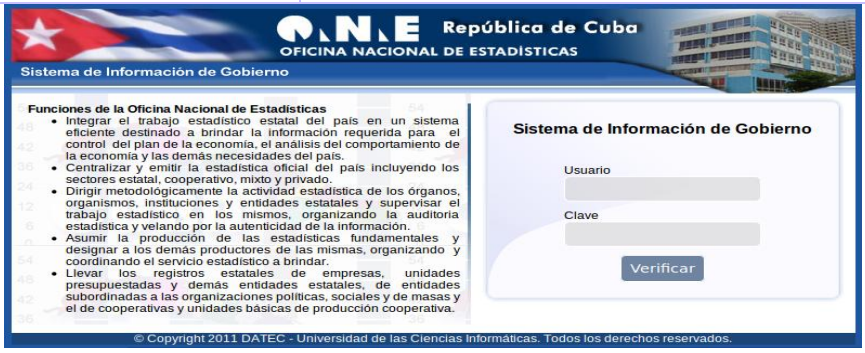
	el actor introduzca los datos.
3. El actor introduce los datos requeridos.	4. El sistema verifica que los datos sean correctos.
	5. Si los datos son correctos, el sistema permite al actor acceder a la aplicación. Finaliza el CU.
Flujo Alterno	
Acción del Actor	**Respuesta del Sistema**
	5.1 Si los datos no son correctos, muestra un mensaje de error y regresa al punto 2 del flujo normal de eventos.
Prototipo de interfaz	
Poscondiciones	El actor (Analista o Administrador) queda autenticado.

Tabla 2: Descripción del CU Autenticar usuario

Analizar información de nacimientos y defunciones:

Caso de Uso:	Analizar información de nacimientos y defunciones
Tipo:	Información
Actores:	Analista, Administrador.
Resumen	El CU inicia cuando el actor desea hacer un análisis de los nacimientos y defunciones desde diferentes perspectivas. El actor selecciona el reporte que desea ver; el sistema muestra la información contenida en él y las opciones de los posibles cambios que le puede hacer al reporte. Cuando el actor termina el análisis de la información de nacimientos y defunciones finaliza el CU.
Precondiciones:	- El actor se autenticó correctamente. - Los datos correspondientes fueron cargados en el mercado de datos. - Los reportes relacionados con los nacimientos y defunciones fueron creados.

Referencias	RI1, RI2, RI3, RI4, RI5, CU Visualizar Reportes.
Prioridad	Crítico.

Flujo Normal de Eventos	
Acción del Actor	**Respuesta del Sistema**
1. El actor se autentica en el sistema.	2. Muestra la interfaz principal con las áreas de análisis existentes.
3. El actor selecciona el área de análisis general A.A.G SIGOB.	4. Muestra las áreas de análisis contenidas en el A.A.G SIGOB.
5. El actor selecciona el área de análisis A.A Series históricas de población.	6. Muestra los libros de trabajo contenidas en el A.A Series históricas de población.
7. El actor selecciona el libro de trabajo L.T Nacimientos y defunciones.	8. Muestra los reportes contendidos en el L.T Nacimientos y defunciones.
9. El actor selecciona el reporte que desea analizar.	10. Muestra la información contenida en el reporte seleccionado y brinda opciones al actor para hacer cambios al reporte durante su análisis. Ir al CU Visualizar reportes. Finaliza el CU.

Opciones de reportes de Analizar información de nacimientos y defunciones.		
Perspectivas de análisis	**Posibles resultados**	
	Medidas	**Periodicidad**
Variables de entrada relacionadas con el CU Analizar información de nacimientos y defunciones: • Año. • Provincia.	Variables de salida disponibles en el hecho Analizar información de nacimientos y defunciones**:** • Cantidad de defunciones. • Tasa de mortalidad (por mil habitantes). • Cantidad de nacimientos. • Tasa de natalidad (por mil habitantes). • Tasa de crecimiento natural (por mil habitantes).	Rango de tiempo en que se solicitan las variables de salida: • Anual.

Prototipo de interfaz	 Medidas Año | Provincia | Cantidad de defunciones | Tasa de mortalidad (por mil habitantes) 1970 | Pinar del Río | 2.787 | 5,1 La Habana | 3.126 | 6 Ciudad de La Habana | 14.097 | 7,9 Matanzas | 3.749 | 6,7 Villa Clara | 4.285 | 6,3 Cienfuegos | 1.913 | 7,2
Poscondiciones	Los reportes correspondientes al libro de trabajo L.T Nacimientos y defunciones han sido consultados por el actor (Analista, Administrador).

Tabla 3: Descripción del CU Analizar información de nacimientos y defunciones

Realizar extracción de datos de series históricas de población.

Caso de Uso:	Realizar extracción de datos de series históricas de población
Tipo:	Funcional
Actores:	Administrador ETL
Resumen	Inicia cuando el Administrador ETL desea hacer la extracción de los datos. Luego selecciona la fuente de información correspondiente y extrae los datos contenidos en ella. El CU finaliza cuando todos los datos de la fuente son extraídos.
Precondiciones:	Fuentes disponibles.
Referencias	RF9
Prioridad	Crítico.

Flujo Normal de Eventos	
Acción del Actor	**Respuesta del sistema**
1. El administrador de ETL realiza la conexión al 'excel' correspondiente.	2. Responde a la solicitud de conexión.
3. El administrador de ETL selecciona estructuras o archivos a extraer.	
4. El administrador de ETL realiza la extracción de	5. Ejecuta la extracción de los datos. Finaliza

los datos.	el CU.
Flujos Alternos	
Acción del Actor	**Respuesta del sistema**
	2.1 No responde a la solicitud de conexión.
	2.2 Notifica el error al administrador de ETL. Vuelve al paso 1 del Flujo Normal de Eventos.
3.1 Si hay control de cambios, verifica si hay modificaciones. • En caso positivo, va al paso 3 del flujo normal. • En caso negativo, va al paso 4 del flujo normal.	
Poscondiciones	Los datos del 'excel' correspondiente están preparados para ser transformados y cargados.

Tabla 4: Descripción del CU Realizar extracción de datos de series históricas de población

Realizar transformación y carga de datos de series históricas de población.

Caso de Uso:	Realizar transformación y carga de datos de series históricas de población
Tipo:	Funcional.
Actores:	Administrador ETL
Resumen	El CU inicia cuando el administrador de ETL desea transformar y cargar los datos. El actor selecciona la fuente de información correspondiente, realiza las transformaciones necesarias y carga la información en el mercado de datos, de esta forma finaliza el CU.
Precondiciones:	Se realizó la extracción completada de los datos en el área temporal y las estructuras del mercado de datos se encontraron disponibles para su uso.
Referencias	RF10
Prioridad	Crítico.
Flujo Normal de Eventos	
Acción del Actor.	**Respuesta del sistema.**
1. El administrador de ETL selecciona las estructuras del área temporal a transformar.	
2. El administrador de ETL carga los datos seleccionados en memoria.	
3. El administrador de ETL aplica transformaciones	

pertinentes y genera los datos.	
4. El administrador de ETL carga los datos en el mercado de datos.	5. Ejecuta la consulta. Finaliza el CU.
Poscondiciones	Han sido transformados y cargados en el mercado de datos los datos del fichero 'excel' correspondiente.

Tabla 5: Descripción del CU Realizar transformación y carga de datos de series históricas de población

Para un mejor entendimiento de estos CU se encuentra la descripción correspondiente a cada uno en el documento *Modelo de Casos de Uso del Sistema* del expediente de proyecto.

2.7 Especificación del modelo dimensional

2.7.1 Matriz BUS

La siguiente tabla representa la matriz BUS, esta contiene las relaciones que existen entre los hechos y las dimensiones del mercado de datos Series históricas de población.

Hecho/Dimensión	provincia	temporal_anno	dpa	capitales_prov
nacimientos_defunciones	X	X		
pob_resid_y_media		X	X	
pob_resid_y_media_capit_prov		X		X

Tabla 6: Matriz BUS

2.7.2 Tablas de dimensiones

Dimensión provincia: la dimensión describe los valores bajo los cuáles puede clasificarse la información teniendo en cuenta solamente la provincia donde ocurre un determinado hecho.

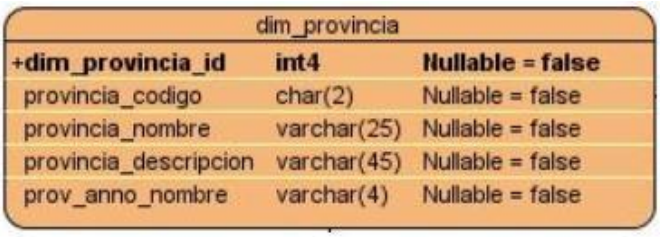


Figura 4: Dimensión provincia

Dimensión temporal año: la dimensión describe los valores bajo los cuáles puede clasificarse la información atendiendo al tiempo donde ocurre un determinado hecho.

dim_temporal_anno		
+dim_temporal_anno_id	**int4**	**Nullable = false**
anno_codigo	varchar(4)	Nullable = true
anno_nombre	varchar(4)	Nullable = true
anno_numero	int4	Nullable = true

Figura 5: Dimensión temporal año

Dimensión capitales provinciales: la dimensión describe los valores bajo los cuáles puede clasificarse la información en cuanto al lugar de residencia de una persona o el lugar de ocurrencia de un hecho determinado según la capital provincial.

dim_capitales_prov		
+dim_capitales_prov_id	**int4**	**Nullable = false**
#dim_provincia_id	*int4*	*Nullable = true*
capit_prov_codigo	varchar(4)	Nullable = true
capit_prov_nombre	varchar(25)	Nullable = true

Figura 6: Dimensión capitales provinciales

Dimensión dpa: la dimensión describe los valores bajo los cuáles puede clasificarse la información en cuanto al lugar de residencia de una persona o el lugar de ocurrencia de un hecho determinado según la división política administrativa.

dim_dpa		
+dim_dpa_id	**int4**	**Nullable = false**
provincia_codigo	char(2)	Nullable = false
provincia_nombre	varchar(25)	Nullable = false
municipio_codigo	varchar(4)	Nullable = false
municipio_nombre	varchar(30)	Nullable = false
provincia_descripcion	varchar(45)	Nullable = false
municipio_descripcion	varchar(45)	Nullable = false
municipio_extension	float4	Nullable = true
municipio_ext_cayosady	float4	Nullable = true
municipio_ext_tierrafirme	float4	Nullable = true
dpa_anno_nombre	varchar(4)	Nullable = false

Figura 7: Dimensión dpa

2.7.3 Tablas de hechos

Hecho nacimientos y defunciones: en este hecho se almacena toda la información relacionada con la cantidad de nacimientos, el valor de la tasa de crecimiento natural, la cantidad de defunciones, el valor de la tasa de mortalidad y el valor de la tasa de crecimiento natural, específicamente en las provincias del país.

hech_nacimientos_defunciones		
+#dim_provincia_id	***int4***	***Nullable = false***
+#dim_temporal_anno_id	***int4***	***Nullable = false***
cant_nacimientos	int4	Nullable = true
tasa_natalidad	float4	Nullable = true
cant_defunciones	int4	Nullable = true
tasa_mortalidad	float4	Nullable = true
tasa_crecim_natural	float4	Nullable = true

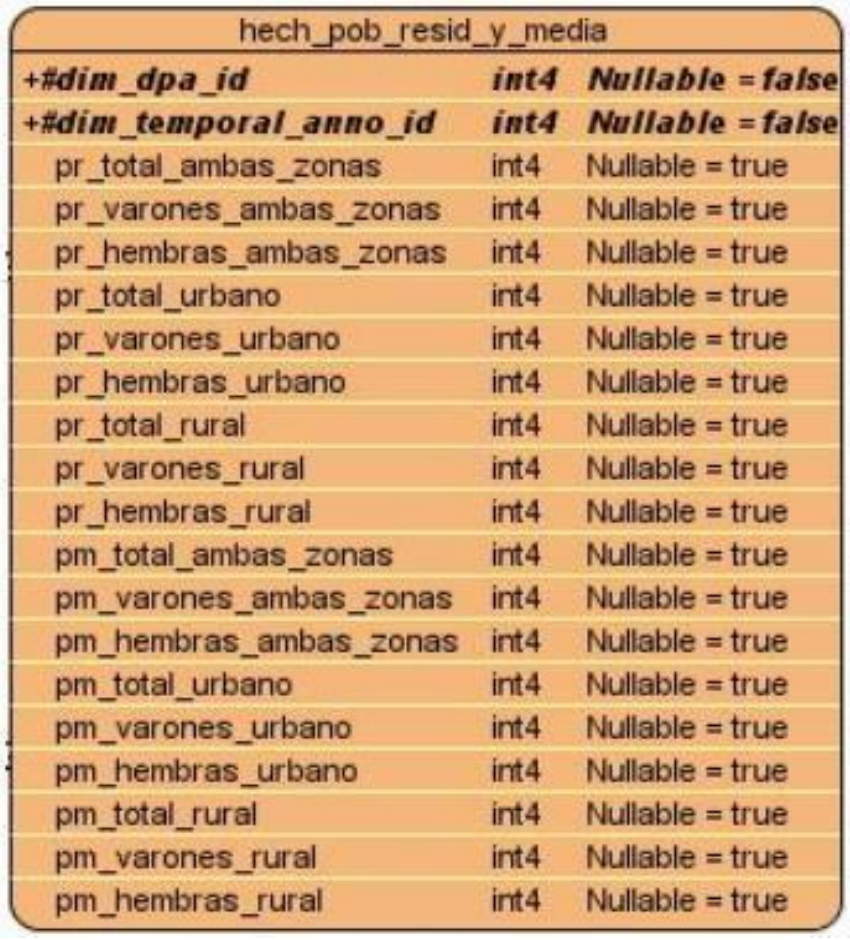
Figura 8: Hecho nacimientos y defunciones

Hecho población residente y media de las capitales provinciales: en este hecho se almacena toda la información relacionada con la cantidad de población residente y media, específicamente de las capitales provinciales del país.

hech_pob_resid_y_media_capit_prov		
+#dim_capitales_prov_id	***int4***	***Nullable = false***
+#dim_temporal_anno_id	***int4***	***Nullable = false***
pob_resid_capit_prov	int4	Nullable = true
pob_media_capit_prov	int4	Nullable = true

Figura 9: Hecho población residente y media de las capitales provinciales

Hecho población residente y media: en este hecho se almacena toda la información relacionada con la cantidad de población residente y media, específicamente en las provincias y municipios del país.



hech_pob_resid_y_media		
+#dim_dpa_id	***int4***	***Nullable = false***
+#dim_temporal_anno_id	***int4***	***Nullable = false***
pr_total_ambas_zonas	int4	Nullable = true
pr_varones_ambas_zonas	int4	Nullable = true
pr_hembras_ambas_zonas	int4	Nullable = true
pr_total_urbano	int4	Nullable = true
pr_varones_urbano	int4	Nullable = true
pr_hembras_urbano	int4	Nullable = true
pr_total_rural	int4	Nullable = true
pr_varones_rural	int4	Nullable = true
pr_hembras_rural	int4	Nullable = true
pm_total_ambas_zonas	int4	Nullable = true
pm_varones_ambas_zonas	int4	Nullable = true
pm_hembras_ambas_zonas	int4	Nullable = true
pm_total_urbano	int4	Nullable = true
pm_varones_urbano	int4	Nullable = true
pm_hembras_urbano	int4	Nullable = true
pm_total_rural	int4	Nullable = true
pm_varones_rural	int4	Nullable = true
pm_hembras_rural	int4	Nullable = true

Figura 10: Hecho población residente y media

2.7.4 Modelo dimensional

Una vez identificadas las tablas de hechos y de dimensiones, y luego de confeccionar la matriz bus para definir las relaciones entre estas tablas, se procede a diseñar el modelo dimensional. Este modelo en cuanto a tipología de esquema es un copo de nieve parcial, debido a que contiene una relación de uno-uno entre dos dimensiones, y a la vez es constelación de hechos ya que presenta tres hechos que comparten al menos una dimensión. Se definió dicha estructura para el depósito de datos, debido a que se adapta mejor según las necesidades del cliente.

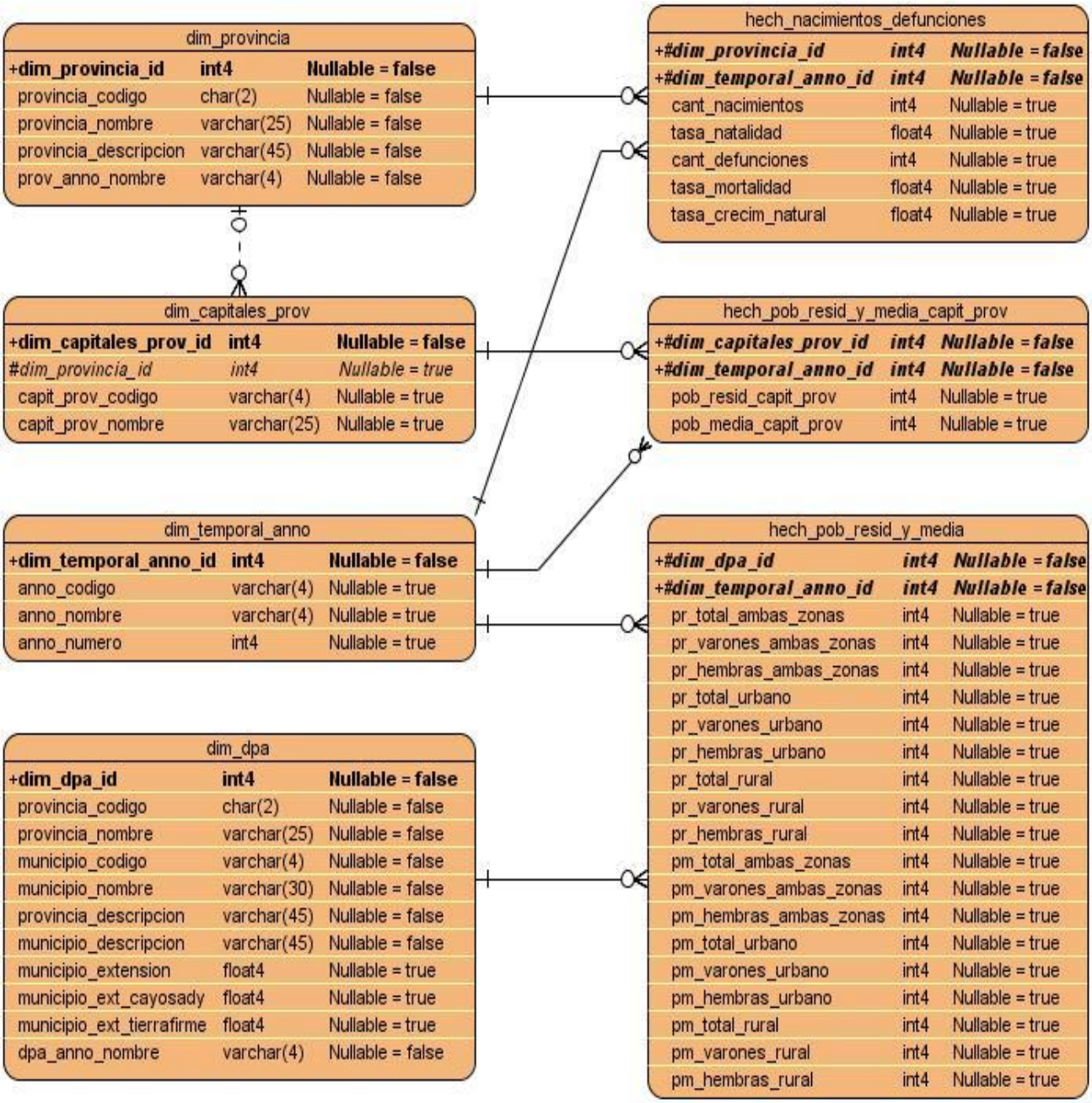


Figura 11: Modelo de datos dimensional

2.7.5 Especificación del modelo físico

El modelo físico se obtuvo a partir del modelo de datos dimensional, ambos modelos cuentan con las mismas tablas (tres hechos y cuatro dimensiones), ya que no existes relaciones de mucho-mucho, por lo cual no se generaron nuevas tablas. El tipo de datos que tendrá la llave primaria de cada tabla será entero.

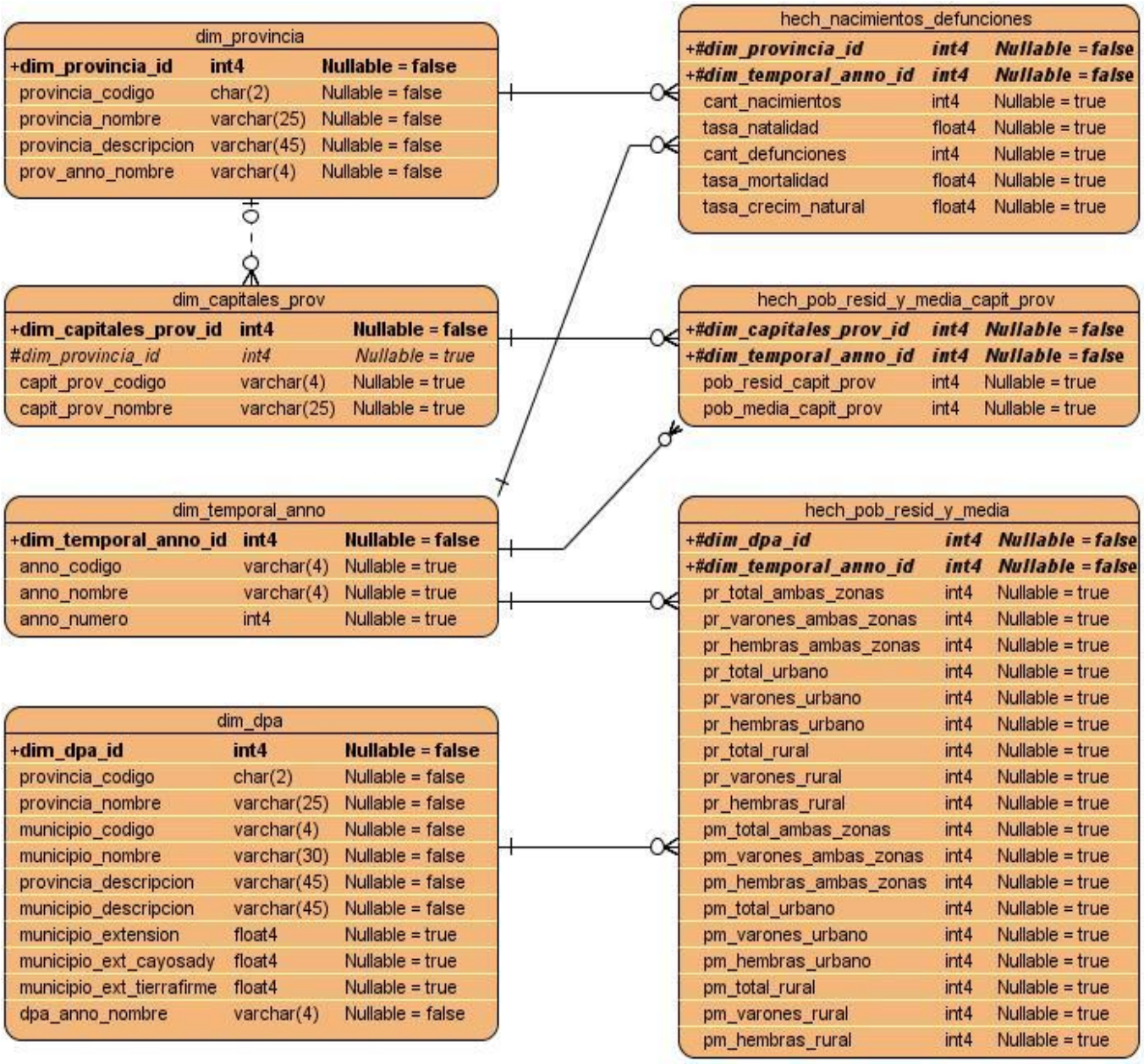


Figura 12: Modelo físico

2.8 Políticas de respaldo y recuperación

La solución utiliza una política de respaldo y recuperación sencilla pero sólida, para ello se miden tres puntos fundamentales:

- Periodicidad de las salvas: las salvas de toda la información contenida en la BD se realizan anualmente, así lo define la organización, certificando en todo momento que exista una copia escrita de la información presente en el servidor.
- Tablas involucradas: las tablas que se involucran en la realización son las tres tablas de hechos

identificadas en el proceso de análisis.

- Backups existentes: En esta área no existen backups actualmente.

2.9 Conclusiones del capítulo

Luego de haber realizado el análisis y diseño del mercado de datos Series históricas de población se obtuvieron los siguientes resultados:

- Se identificaron nueve necesidades de información.
- Se identificaron tres reglas del negocio.
- Se realizó la especificación de requisitos, donde se definieron 25 requisitos de información, 25 funcionales y 19 no funcionales.
- Se identificaron los actores del sistema.
- Se realizó una especificación de los casos de uso del sistema.
- Se identificaron cuatro dimensiones y tres hechos, con lo cual se satisfacen las necesidades del cliente.
- Se hizo una especificación del modelo físico.

CAPÍTULO 3: IMPLEMENTACIÓN DEL MERCADO DE DATOS SERIES HISTÓRICAS DE POBLACIÓN

3.1 Introducción

En este capítulo se realiza la implementación del modelo de datos previamente diseñado, a partir de esto se procede diseñar e implementar el subsistema de integración con el objetivo de poblar el mercado de datos. Después de tener el mercado de datos poblado se diseña y desarrolla el subsistema de visualización para gestionar los reportes candidatos necesarios que satisfacen las necesidades del cliente.

3.2 Diseño del subsistema de integración

Para el subsistema de integración se realizan diferentes transformaciones, que al ser ejecutadas permiten que los datos se encuentren disponibles en una tabla de salida. A continuación se muestra el diseño general que se siguió para implementar cada una de las transformaciones por vez primera:

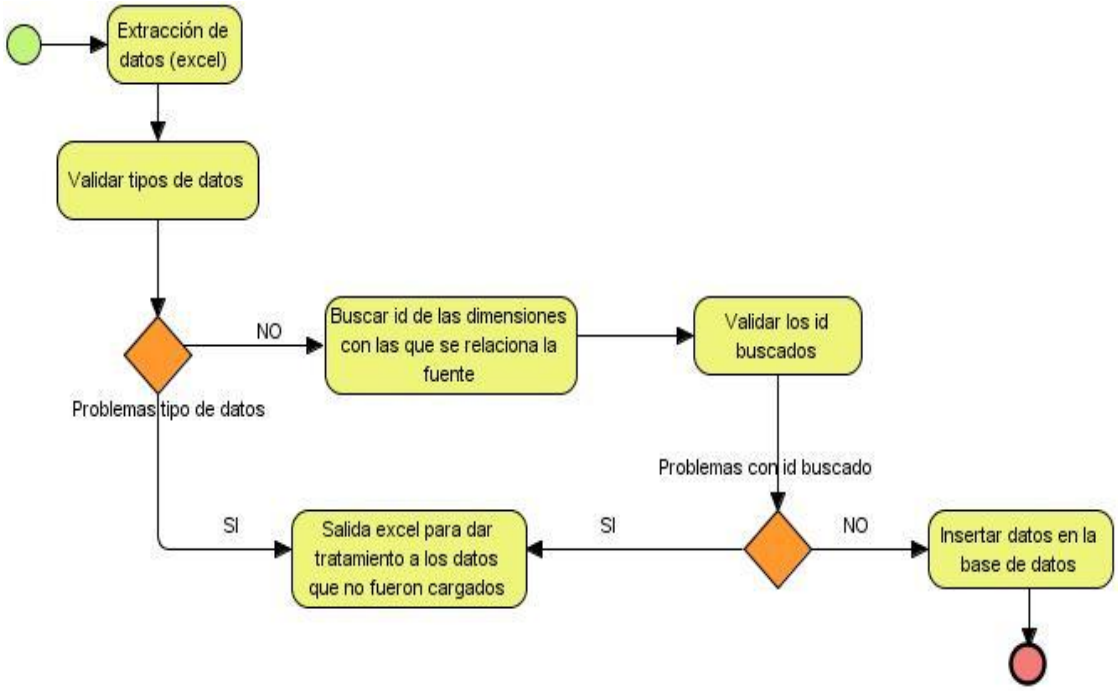


Figura 13: Diseño de una transformación

Registro de sistemas fuentes

Es necesario realizar un análisis del estado en que se encuentran los sistemas fuentes de las series históricas de población, para esto se generó un artefacto en el expediente de proyecto llamado Registro de sistemas fuentes. Este documento tiene como objetivo fundamental realizar una correcta

documentación de los sistemas fuentes, específicamente del estado físico en el que se encuentran los datos, así como los responsables de los diferentes sistemas fuentes.

Diccionario de datos

Para dar paso al proceso de ETL es también necesario hacer un correcto análisis de las variables de la fuente de datos, y de esta forma comprender mejor su contenido, estructura y dependencia. Esto se realizó en el artefacto Diccionario de datos del expediente de proyecto, en el cual se especifica de cada variable su significado en el negocio y los posibles valores que puede tomar.

Mapa lógico de datos

Es fundamental antes de abordar la implementación del subsistema de integración, la correcta identificación del flujo entre los datos fuentes y los datos destino. Para esto existe en el expediente de proyecto el artefacto: Mapa lógico de datos, el cual servirá como guía durante el proceso de ETL.

3.3 Implementación del subsistema de integración

3.3.1 Implementación del modelo de datos

Durante la implementación del modelo de datos se tuvo en cuenta todo lo especificado anteriormente en el modelo físico, como por ejemplo los tipos de datos de las variables, la cardinalidad entre las tablas, así como los esquemas. Luego de obtener el script a partir de este modelo, se desarrolló la BD en el PgAdmin (ver figura 14).

En el mercado de datos Series históricas de población se definieron para la organización de las tablas de la BD dos esquemas: dimensiones, que agrupa todas las dimensiones comunes en el almacén de datos, y el esquema mart_demografia_series, que contiene todos los hechos y dimensiones propias de dicho mercado. La solución cuenta con siete tablas en total, cuatro dimensiones y tres hechos, distribuidas por los dos esquemas de la siguiente forma:

Esquemas	Tablas
dimensiones	dim_provincia
dimensiones	dim_dpa
dimensiones	dim_temp_anno
mart_demografia_series	dim_capitales_prov

mart_demografia_series	hech_nacimientos_defunciones
mart_demografia_series	hech_pob_resid_y_media
mart_demografia_series	hech_pob_resid_y_media_capit_prov

Tabla 7: Esquemas y tablas

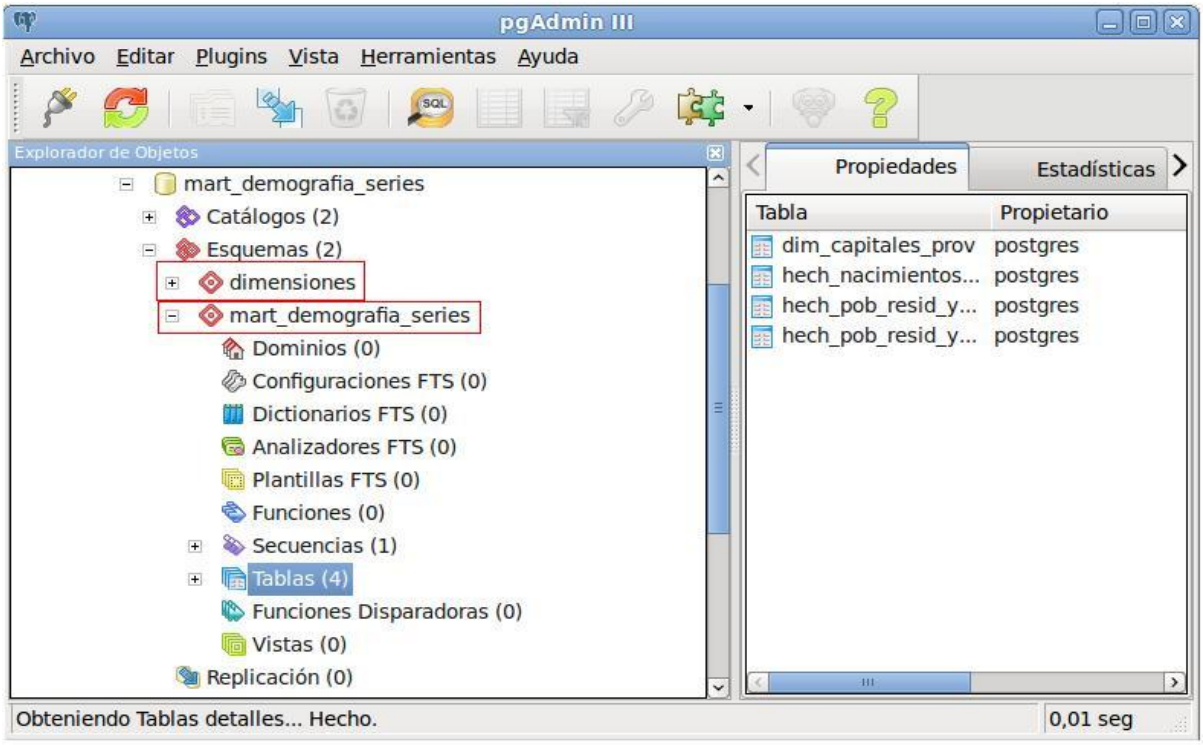


Figura 14: Implementación del modelo de datos

3.3.2 Implementación de los flujos de transformación

El elemento básico de diseño de los procesos ETL es la transformación, la cual está compuesta por pasos enlazados entre sí a través de los saltos. Los pasos constituyen el elemento más pequeño dentro de las transformaciones, y a través de los saltos fluye la información entre los diferentes pasos. En el presente trabajo se realizó un flujo de transformación para la carga de cada una de las tablas pertenecientes al esquema mart_demografia_series, a continuación se muestra la carga del hecho población residente y media de las capitales provinciales:

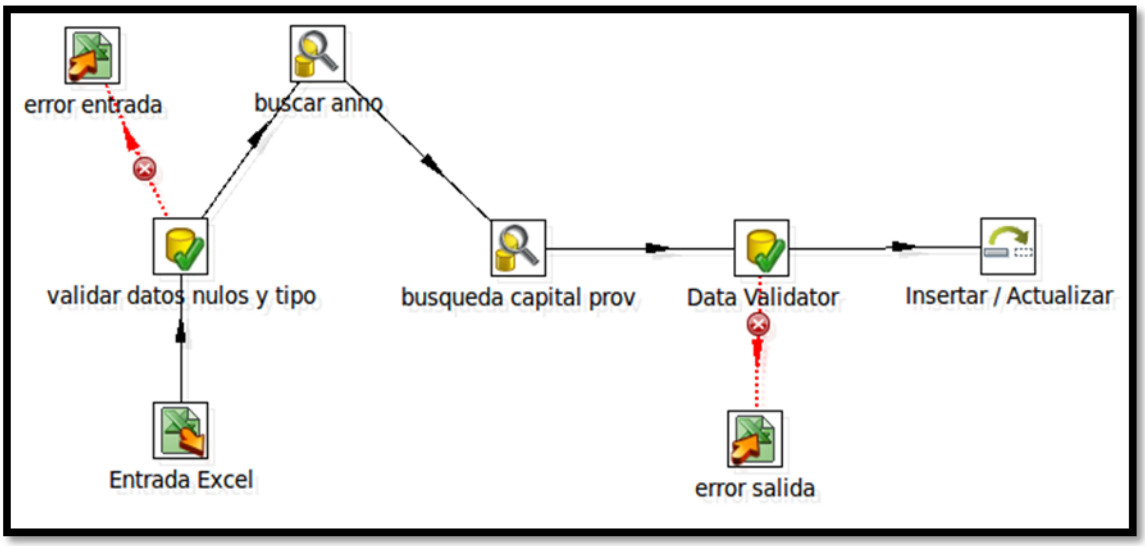


Figura 15: Carga del hecho Población residente y media de las capitales provinciales

Como se puede apreciar en la figura 15, el primer paso de la transformación es una entrada excel, debido a que la información perteneciente al hecho población residente y media de las capitales provinciales, se encuentra en un fichero de este tipo. El segundo está destinado a la validación de los datos, donde se valida que las medidas sean valores enteros, no nulos (ver figura 16) y que sean positivos (ver figura 17); en caso de no cumplirse esto los valores pasarían a un excel para su posterior tratamiento. En el tercer y cuarto paso se realiza una búsqueda para obtener el id de las dimensiones con las que se relaciona el hecho. El penúltimo paso es el encargado de verificar que se hayan encontrado todos los id, es decir que no hayan campos de id vacío, en caso contrario se le da salida a los valores hacia un fichero excel para solucionar el error. El último paso es un componente que permite insertar nuevos valores en la base de datos, o actualizar los que ya se encuentran insertados.

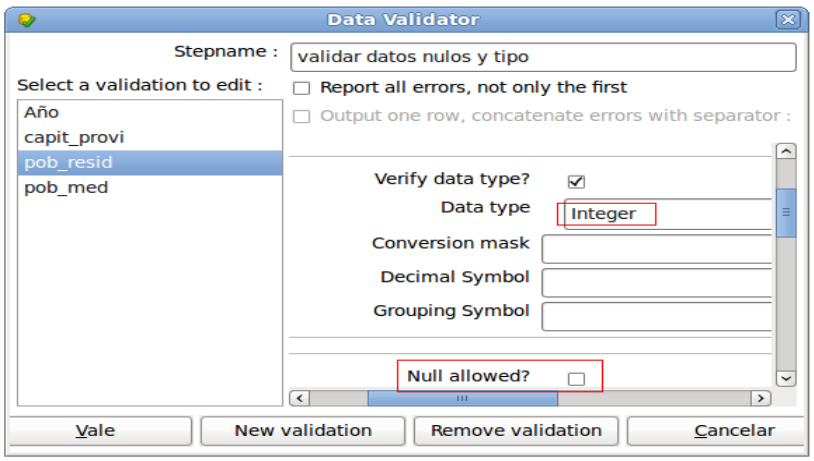


Figura 16: Validar tipo de dato y no permitir datos nulos

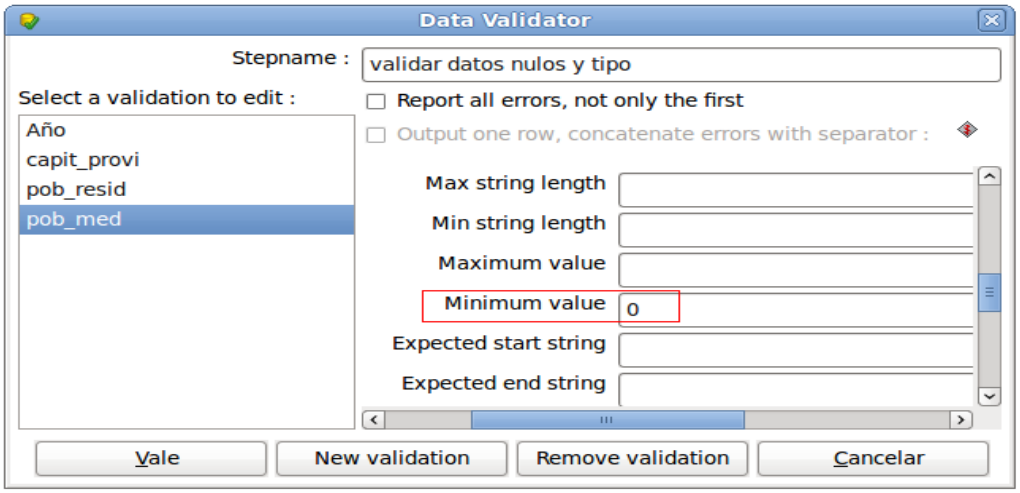


Figura 17: Validar que los valores sean positivos

3.3.3 Implementación de los trabajos

Un trabajo o job es similar a un proceso: conjunto de tareas con el objetivo de realizar una acción determinada. Estos permiten ejecutar varias transformaciones o trabajos previamente diseñados y organizar una secuencia de ejecución de éstos. Los trabajos se encuentran en un nivel superior de las transformaciones. En la presente investigación se realizó un trabajo general, donde se ejecutan primeramente la transformación correspondiente a la carga de la dimensión particular del mercado de datos, y posteriormente las cargas de los tres hechos identificados. A continuación se muestra el trabajo realizado:

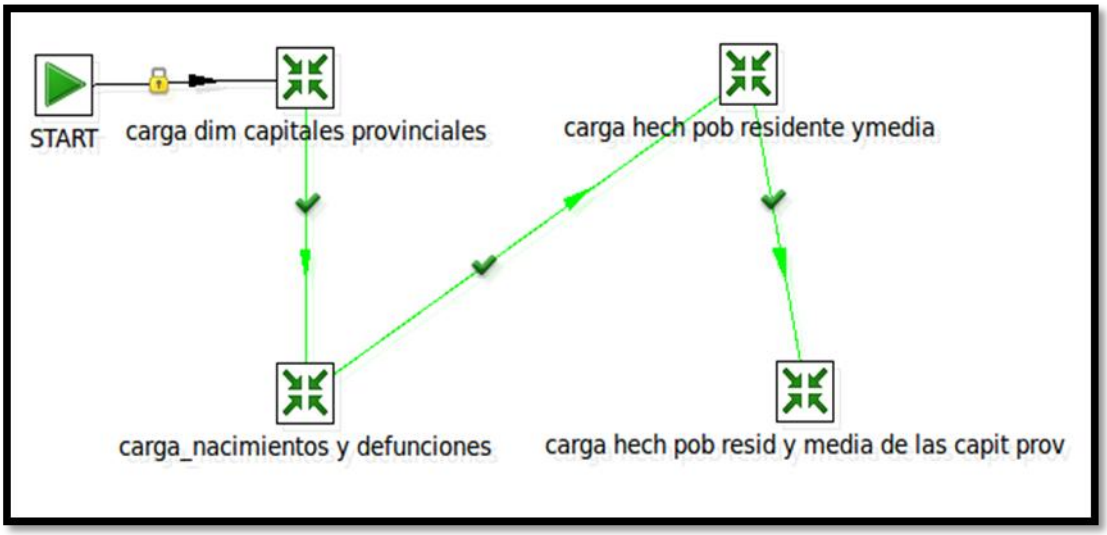


Figura 18: Implementación del trabajo

3.4 Diseño del subsistema de visualización

3.4.1 Arquitectura de información

El artefacto Arquitectura de información del expediente de proyecto, contribuye a definir el entorno de análisis, monitoreo y control de las series históricas de población. En este se identificó un Área de Análisis (A.A), que contiene tres Libros de Trabajo (LT) y diez reportes agrupados en los libros de trabajo. A continuación se detallan los elementos que componen las estructuras de navegación de la información presentada en la capa de visualización:

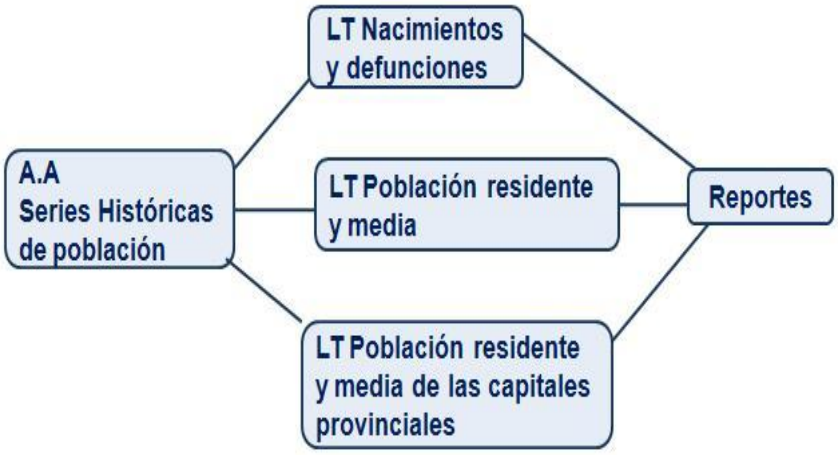


Figura 19: Mapa de navegación

3.4.2 Diseño de los reportes candidatos

En el expediente de proyecto existe un artefacto denominado Reportes candidatos, en este se describen los reportes candidatos identificados para el desarrollo del mercado de datos Series históricas de población. A continuación se muestra la descripción de uno de los reportes:

Área de análisis (AA)	**Series históricas de población**
Libro de Trabajo (LT)	LT1: Nacimientos y defunciones
Reporte (Tabla de Salida – TS)	TS4: Nacimientos y tasas de natalidad según provincia de residencia
Descripción	El reporte muestra los nacimientos y tasas de natalidad según la provincia de residencia, en el

	tiempo.
Elementos del reporte	• Año • Provincia • Cantidad de nacimientos • Tasa de natalidad (por mil habitantes)

Tabla 8: Descripción del reporte Nacimientos y tasas de natalidad según provincia de residencia

3.5 Implementación del subsistema de visualización

3.5.1 Implementación de los cubos OLAP

La implementación de los cubos OLAP se hizo en la herramienta Pentaho Schema Workbench. Esta herramienta permite generar un fichero de configuración XML, en el cual se definen los hechos, las dimensiones con sus niveles de jerarquía, así como la conexión con la BD que contiene los datos para el cubo multidimensional.

En el presente trabajo se modelaron tres cubos y cuatro dimensiones, con las características correspondientes a cada una de las tablas de hechos y dimensiones respectivamente, esto se muestra a continuación:

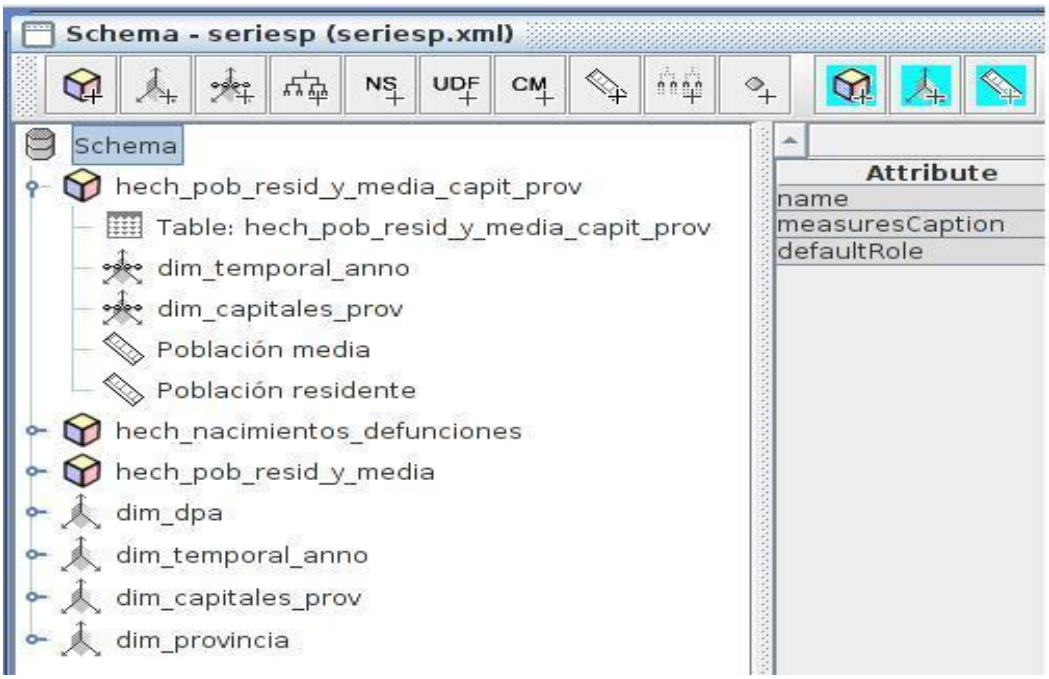


Figura 20: Implementación de los cubos OLAP

3.5.2 Implementación de los reportes candidatos

Los reportes candidatos necesarios para satisfacer las necesidades del cliente fueron implementados en la herramienta Pentaho BI Server, mediante consultas MDX (ver figura 20), las cuales varían su complejidad en dependencia de los tipos de reportes.

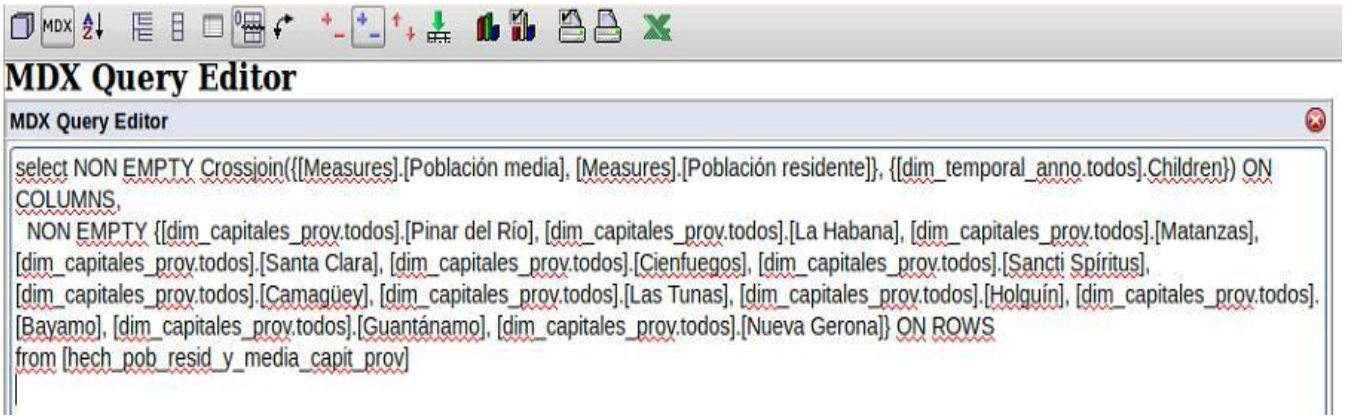
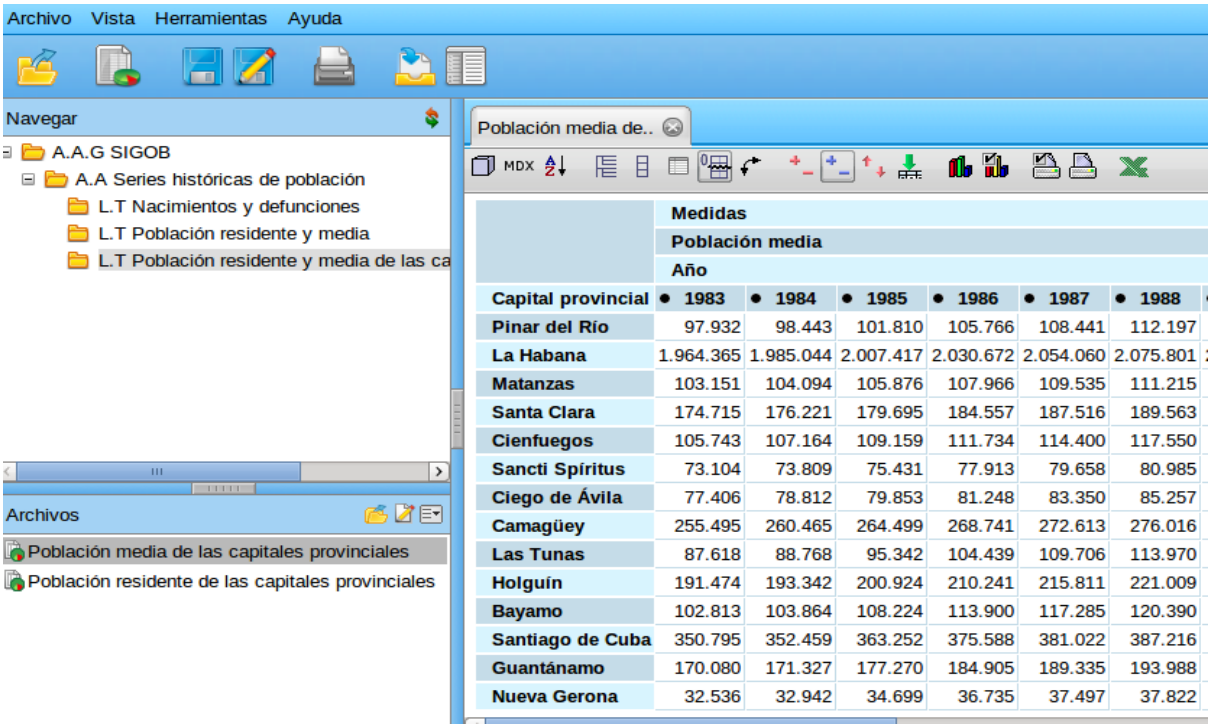


Figura 21: Consulta MDX

A continuación se muestra la vista de uno de los reportes o tablas de salidas luego de su implementación:



	Medidas					
	Población media					
	Año					
Capital provincial	1983	1984	1985	1986	1987	1988
Pinar del Río	97.932	98.443	101.810	105.766	108.441	112.197
La Habana	1.964.365	1.985.044	2.007.417	2.030.672	2.054.060	2.075.801
Matanzas	103.151	104.094	105.876	107.966	109.535	111.215
Santa Clara	174.715	176.221	179.695	184.557	187.516	189.563
Cienfuegos	105.743	107.164	109.159	111.734	114.400	117.550
Sancti Spíritus	73.104	73.809	75.431	77.913	79.658	80.985
Ciego de Ávila	77.406	78.812	79.853	81.248	83.350	85.257
Camagüey	255.495	260.465	264.499	268.741	272.613	276.016
Las Tunas	87.618	88.768	95.342	104.439	109.706	113.970
Holguín	191.474	193.342	200.924	210.241	215.811	221.009
Bayamo	102.813	103.864	108.224	113.900	117.285	120.390
Santiago de Cuba	350.795	352.459	363.252	375.588	381.022	387.216
Guantánamo	170.080	171.327	177.270	184.905	189.335	193.988
Nueva Gerona	32.536	32.942	34.699	36.735	37.497	37.822

Figura 22: Vista del reporte Población media de las capitales provinciales

3.5.3 Configuración de la seguridad de los usuarios

La seguridad de los usuarios es un elemento muy importante, ya que se deben definir los usuarios y roles que pueden acceder a la capa de visualización del mercado de datos, así como las acciones que estos pueden realizar sobre los datos. En el presente trabajo se definieron dos usuarios: administrador que cuenta con todos los permisos de administración y analista que solamente tiene permisos de lectura. A continuación se muestran imágenes sobre la administración de los roles y usuarios para el mercado de datos Series históricas de población, así como la asignación de los permisos que incluye Pentaho BI Server:

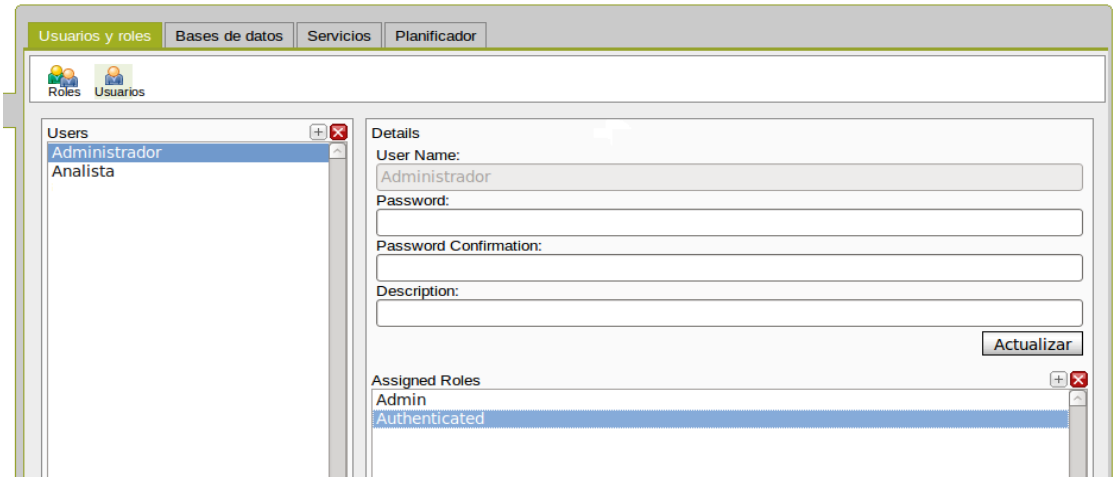


Figura 23: Administración de roles y usuarios

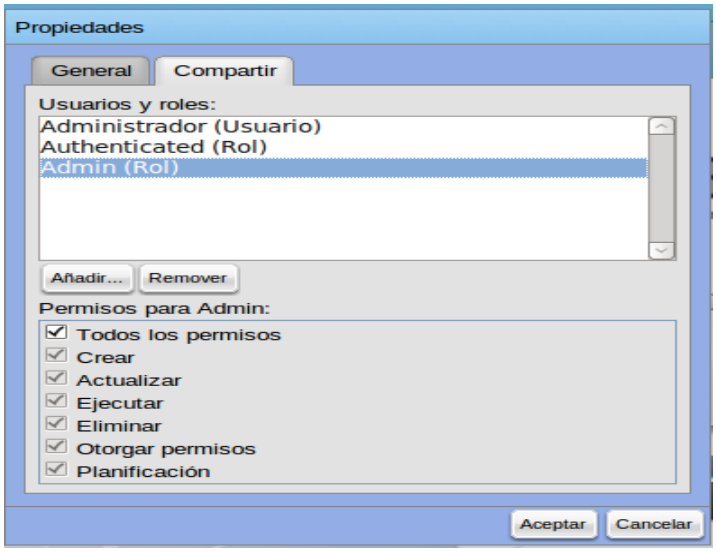


Figura 24: Asignación de permisos

3.6 Conclusiones del capítulo

Después de realizar la implementación del mercado de datos Series históricas de población se obtuvieron las siguientes conclusiones:

- Se realizó el diseño de los subsistemas de integración y visualización.
- Se realizaron cuatro transformaciones y un trabajo, quedando implementado el subsistema de integración.
- Se implementaron los cubos OLAP en correspondencia con las tablas de hechos y dimensiones identificadas anteriormente.
- Se implementó el subsistema de visualización, este cuenta con un área de análisis, tres libros de trabajos y diez reportes.

CAPÍTULO 4: VALIDACIÓN DEL MERCADO DE DATOS SERIES HISTÓRICAS DE POBLACIÓN

4.1 Introducción

En este capítulo se realizan pruebas al mercado de datos Series históricas de población, con el objetivo de verificar su correcto funcionamiento y que cuente con la calidad requerida por el cliente. Esta validación se realiza a partir de la aplicación de una lista de chequeo y los casos de pruebas diseñados. También se recibe una carta de aceptación por parte del cliente.

4.2 Casos de pruebas

En la presente investigación se utilizan los casos de pruebas para validar la implementación del subsistema de visualización, ya que estos constituyen un tipo de prueba que se realiza a la interfaz de una aplicación informática. La siguiente tabla muestra el diseño de uno de los casos de prueba que fueron aplicados:

Escenario	Descripción	Perfiles de análisis	Indicadores a medir	Respuesta del sistema	Flujo central
EC 1.1: Población media de las capitales provinciales.	Permite visualizar el reporte con las variables presentes en el mismo.	• Año. • Capital provincial.	• Población media.	Se muestra la tabla con los valores correspondientes a cada escenario.	Se abre la aplicación. Se autentifica. Se entra al sistema. Se despliega hacia la derecha el componente ubicado en el lateral izquierdo que contiene el navegador. Se selecciona el área de análisis **AA. Series históricas de población.** Se selecciona el libro de trabajo **LT. Población residente y media de las capitales provinciales.** En la parte inferior izquierda se selecciona el reporte deseado. En el área de trabajo se visualiza la tabla correspondiente al reporte.
EC 1.2: Población residente de las capitales provinciales.		• Año. • Capital provincial.	• Población residente.		

Tabla 9: Caso de prueba Analizar información de población residente y media de las capitales provinciales

Además del caso de prueba mencionado anteriormente, se aplicaron: Caso de prueba Analizar información de nacimientos y defunciones y Caso de prueba Analizar información de población residente y media (ver anexo 1).

Aplicación de los casos de prueba

Los casos de prueba previamente diseñados fueron aplicados primeramente por especialistas del departamento, donde se identificaron cuatro no conformidades (NC), las cuales fueron solucionadas. Luego especialistas del centro DATEC aplicaron nuevamente los casos de pruebas, surgiendo una NC, a la cual se le dio tratamiento. Por último el mercado de datos pasó a ser revisado por especialistas de la empresa especializada en temas de calidad: Calisoft, identificando solamente una NC.

Una vez aplicados los casos de pruebas, se obtuvieron resultados satisfactorios. De forma general se verificó la disponibilidad de los perfiles de análisis (dimensiones) y de los indicadores a medir (medidas) para cada reporte, y se solucionaron todas las NC surgidas en cada una de las revisiones.

4.3 Lista de chequeo

Una lista de chequeo contiene un listado de preguntas en forma de cuestionario que sirven para verificar el grado de cumplimiento de determinadas reglas establecidas. Es un instrumento que contiene criterios o indicadores a partir de los cuales se miden y evalúan las características de un objeto, comprobando si cumple con los atributos establecidos.

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 diferentes secciones.
- **Eval. (Evaluación):** es la forma de evaluar el indicador en cuestión, éste se evalúa de uno en caso de que exista alguna dificultad o de cero en caso contrario.
- **NP (No Procede):** se usa para especificar que no es necesario evaluar el indicador en ese caso.
- **Cantidad de elementos afectados:** especifica la cantidad de errores encontrados sobre un indicador.
- **Comentario:** especifica los señalamientos o sugerencias que se quieran incluir. Pueden o no existir señalamientos o sugerencias.

La lista de chequeo definida para la evaluación del mercado de datos Series históricas de población, analiza 14 indicadores, de los cuales ocho son críticos. Estos indicadores están distribuidos en tres

secciones fundamentales: Estructura del documento, Indicadores definidos en el desarrollo y Semántica del documento. A continuación se muestra la lista de chequeo aplicada en esta investigación:

Estructura del documento					
Peso	**Indicadores a evaluar**	**Eval**	**(NP)**	**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)	0		0	
Indicadores definidos en el desarrollo					
Peso	**Indicadores a evaluar**	**Eval**	**(NP)**	**Cantidad de elementos afectados**	**Comentarios**
	1. ¿Se utilizó un lenguaje cuyas sentencias son expresables mediante una sintaxis bien definida?	0		0	
crítico	2. ¿Los reportes pueden configurarse a través de la interfaz del sistema?	0		0	
	3. ¿La interfaz está orientada a facilitar el uso de las funciones del sistema por parte de los usuarios?	0		0	
crítico	4. ¿No existen restricciones para construir cubos OLAP con dimensiones y niveles de agregación ilimitados?	0		0	

crítico	5. ¿Los usuarios son capaces de manipular los resultados de manera que se ajusten a sus necesidades, conformando nuevos reportes?	0		0	
	6. ¿El sistema responde de una forma rápida a la información solicitada por el usuario?	0		0	
	7. ¿El sistema refleja cualquier lógica del negocio para poder responder a preguntas específicas?	0		0	
crítico	8. ¿El sistema garantiza la confidencialidad y seguridad de acceso a los datos por roles de usuarios?	0		0	
	9. ¿Los datos e información derivados del proceso de análisis realizado mediante la aplicación, apoyan la toma de decisiones en la ONE?	0		0	
crítico	10. ¿Los cambios en los datos se reflejan automáticamente en los reportes de forma instantánea?	0		0	
Semántica del documento					
Peso	**Indicadores a evaluar**	**Eval**	**(NP)**	**Cantidad de elementos afectados**	**Comentarios**
crítico	1. ¿Se han identificado errores ortográficos en los entregables?	0		0	
crítico	2. ¿Se entiende claramente lo que se ha especificado en el documento?	0		0	

	3. ¿El número de página que aparece en el índice coincide con el contenido que se refleja realmente en dicha página?	1		2	

Tabla 10: Lista de chequeo aplicada

Aplicación de la lista de chequeo

Durante la revisión del desarrollo del mercado de datos, a través de la aplicación de la lista de chequeo, se identificó un indicador con dificultades del cual se generó una NC. Esta fue tratada al corregir los dos elementos afectados, haciendo coincidir el número de página que aparecía en el índice con el contenido que se reflejaba en dicha página. De forma general la implementación fue evaluada de **Bien**, ya que no hubo ningún indicador crítico afectado, no existieron problemas con los formatos de las plantillas, no se encontraron errores ortográficos en los documentos revisados y fue solucionada la NC generada. La siguiente figura, representa el comportamiento de los indicadores en las diferentes secciones de la lista de chequeo, luego de evaluar la implementación del mercado de datos:

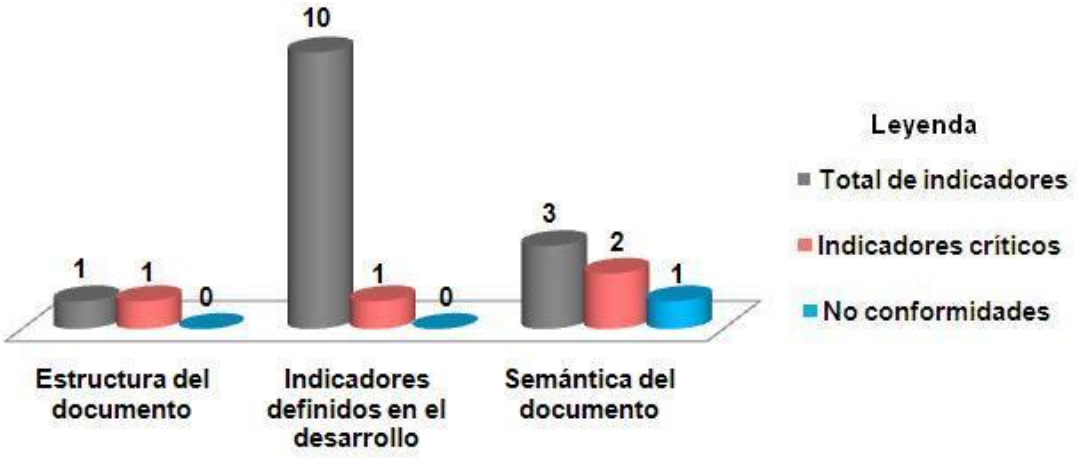


Figura 25: Comportamiento de los indicadores por secciones

4.4 Carta de aceptación de la solución

Luego de concluir el desarrollo del mercado de datos Series históricas de población, así como las validaciones expuestas anteriormente, se realizó un encuentro con la especialista Elena Leonila Fernández García, representante de la ONE en la UCI, en el cual la compañera hizo entrega de una carta de aceptación de todo el desarrollo de dicho mercado.

4.5 Conclusiones del capítulo

Luego de validar el mercado de datos Series históricas de población se obtienen como conclusiones:

- Se aplicaron tres casos de pruebas para validar la capa de visualización del mercado de datos.
- Se aplicó una lista de chequeo con el fin de evaluar de forma general el mercado de datos.
- Se solucionaron las no conformidades identificadas durante la aplicación de las pruebas definidas.
- Se obtuvo una carta de aceptación del mercado de datos firmada por el cliente.

CONCLUSIONES

Luego de finalizada la investigación se arriba a las siguientes conclusiones:

- Se realizó el análisis y diseño del mercado de datos, identificándose tres tablas hecho y cuatro dimensiones.
- Se realizaron cuatro transformaciones y un trabajo, quedando implementado el subsistema de integración.
- Se implementó el subsistema de visualización, este cuenta con un área de análisis, tres libros de trabajos y diez reportes.
- Se validó satisfactoriamente el mercado de datos a través de la aplicación de la lista de chequeo y los casos de prueba.

RECOMENDACIONES

- Utilizar el contenido de la investigación como base de referencia para el desarrollo de futuros mercados de datos.
- Ampliar el nivel de detalle de la información que se recoge de las series históricas de población, como por ejemplo por edades.
- Realizar la integración del almacén de datos con el Sistema de Gestión Estadística (SIGE) que se está desarrollando en el departamento de Soluciones Integrales.

Referencias bibliográficas

1. **Martín Bravo, Ivan y Díaz Morales, Yoel.** *Almacén de datos estadísticos de la ONE: Desarrollo de la capa de visualización del mercado de datos demografía.* 2010. Tesis (Ingeniero en Ciencias Informáticas).

2. **Rodríguez Sotolongo, Javier y Peralta Góngora, Yohan Orlando.** *Implementación del proceso de extracción, transformación y carga de un Datawarehouse para los Ensayos Clínicos del Centro de Inmunología Molecular.* Habana : s.n., 2010. Tesis (Ingeniero en Ciencias Informáticas).

3. **Galindo González, Lic. Carlos y Pérez Vázquez, Dr. Ramiro.** Gestiopolis. [En línea] 27 de Agosto de 2009. [Citado el: 6 de Noviembre de 2010.] http://www.gestiopolis.com/administracion-estrategia/almacenes-de-datos-y-microsoft-sql-server.htm.

4. Sinnexus. [En línea] [Citado el: 8 de Octubre de 2010.] http://www.sinnexus.com/business_intelligence/datamart.aspx.

5. **Casales Cabrera, María Evelia.** *Data Warehouse (Almacenes de Datos).* 2009. Maestría en Ciencias e Ingeniería de la Computación.

6. **Ohlinger, Patrick.** *Wal-Mart's Data Warehouse.* 2006.

7. **Díaz Morales, Themis Patricia y Bermúdez Rodríguez, José Salvador.** *Diseño de un Datawarehouse para los Ensayos Clínicos que se gestionan en el Centro de Inmunología Molecular.* Habana : s.n., 2010. Tesis (Ingeniero en Ciencias Informáticas).

8. **Simón Mir, Yailin y Iñiguez Bermúdez, Yoander.** *Análisis de datos a un Datawarehouse para los Ensayos Clínicos del Centro de Inmunología Molecular, aplicando técnicas OLAP.* Habana : s.n., 2010. Tesis (Ingeniero en Ciencias Informáticas).

9. **Servente, Magdalena y García Martínez, Dr. Ramón.** *Algoritmos TDIDT aplicados a la Minería de Datos Inteligente.* Facultad de Ingeniería, Universidad de Buenos Aires. 2002. Tesis de Grado en Ingeniería Informática. (http://laboratorios.fi.uba.ar/lsi/servente-tesisingenieriainformatica.pdf).

10. **Ricardo Dario, Ing. Bernabeu.** *Data Warehousing: Investigación y sistematización de conceptos. Hefesto: Metodología propia para la construcción de un Datawarehouse.* Córdoba : s.n., 2009.

11. **DATEC, Especialistas de.** *Metodología para el Desarrollo de Soluciones de Almacenes de Datos e Inteligencia de Negocios en DATEC.* Habana : s.n., 2010.

12. **The PostgreSQL Global Development Group.** PostgreSQL. [En línea] 2005. [Citado el: 2 de Noviembre de 2010.] http://www.postgresql.org/docs/8.0/interactive/index.html.

13. Pentaho Open Source Business Intelligence. [En línea] [Citado el: 28 de Octubre de 2010.] http://www.pentaho.com/.

Bibliografía

Biosca, Eric. *Tutorial MDX.* 2005. (http://www.dataprix.com/files/Tutorial_MDX.pdf).

Casales Cabrera, María Evelia. *Data Warehouse (Almacenes de Datos).* 2009. Maestría en Ciencias e Ingeniería de la Computación.

DATEC, Especialistas de. *Metodología para el Desarrollo de Soluciones de Almacenes de Datos e Inteligencia de Negocios en DATEC.* Habana : s.n., 2010.

Díaz Morales, Themis Patricia y Bermúdez Rodríguez, José Salvador. *Diseño de un Datawarehouse para los Ensayos Clínicos que se gestionan en el Centro de Inmunología Molecular.* Habana : s.n., 2010. Tesis (Ingeniero en Ciencias Informáticas).

Galindo González, Lic. Carlos y Pérez Vázquez, Dr. Ramiro. Gestiopolis. [En línea] 27 de Agosto de 2009. [Consultado el: 6 de Noviembre de 2010.] http://www.gestiopolis.com/administracion-estrategia/almacenes-de-datos-y-microsoft-sql-server.htm.

KLE. *Transforming Knowledge into action! BI en la práctica. Artículos de BI en la práctica.* 2010. http://www.siskle.com/spanish/articulo04.html.

Lanzillotta, Analia. *Definición de OLAP. Tecnología OLAP.* 2004. http://www.mastermagazine.info/termino/6841.php.

Martín Bravo, Ivan y Díaz Morales, Yoel. *Almacén de datos estadísticos de la ONE: Desarrollo de la capa de visualización del mercado de datos demografía.* 2010. Tesis (Ingeniero en Ciencias Informáticas).

Nader, Javier. *Sistema de Apoyo Gerencial Universitario.* Buenos Aires : s.n., 2004. pág. 416 . http://www.itba.edu.ar/archivos/secciones/nader-tesisdemagister.pdf.

Ohlinger, Patrick. *Wal-Mart's Data Warehouse.* 2006.

Pentaho Open Source Business Intelligence. [En línea] [Consultado el: 28 de Octubre de 2010.] http://www.pentaho.com/.

Ricardo Dario, Ing. Bernabeu. *Data Warehousing: Investigación y sistematización de conceptos. Hefesto: Metodología propia para la construcción de un Datawarehouse.* Córdoba : s.n., 2009.

Rodríguez Sotolongo, Javier y Peralta Góngora, Yohan Orlando. *Implementación del proceso de extracción, transformación y carga de un Datawarehouse para los Ensayos Clínicos del Centro de Inmunología Molecular.* Habana : s.n., 2010. Tesis (Ingeniero en Ciencias Informáticas).

Simón Mir, Yailin y Iñiguez Bermúdez, Yoander. *Análisis de datos a un Datawarehouse para los Ensayos Clínicos del Centro de Inmunología Molecular, aplicando técnicas OLAP.* Habana : s.n., 2010. Tesis (Ingeniero en Ciencias Informáticas).

Sinnexus. [En línea] [Consultado el: 8 de Octubre de 2010.] http://www.sinnexus.com/business_intelligence/datamart.aspx.

Servente, Magdalena y García Martínez, Dr. Ramón. *Algoritmos TDIDT aplicados a la Minería de Datos Inteligente.* Facultad de Ingeniería, Universidad de Buenos Aires. 2002. Tesis de Grado en Ingeniería Informática. (http://laboratorios.fi.uba.ar/lsi/servente-tesisingenieriainformatica.pdf).

The PostgreSQL Global Development Group. PostgreSQL. [En línea] 2005. [Consultado el: 2 de Noviembre de 2010.] http://www.postgresql.org/docs/8.0/interactive/index.html.

Anexos

Anexo 1: Diseño de casos de prueba

- **Caso de prueba Analizar información de nacimientos y defunciones:**

Escenario	Descripción	Perfiles de análisis	Indicadores a medir	Respuesta del sistema	Flujo central
EC 1.1: Defunciones y tasas de mortalidad de Cuba.	Permite visualizar el reporte con las variables presentes en el mismo.	• Año.	• Cantidad de defunciones. • Tasa de mortalidad (por mil habitantes).	Se muestra la tabla con los valores correspondientes a cada escenario.	Se abre la aplicación. Se autentifica. Se entra al sistema. Se despliega hacia la derecha el componente ubicado en el lateral izquierdo que contiene el navegador. Se selecciona el área de análisis **AA. Series históricas de población.** Se selecciona el libro de trabajo **LT. Nacimientos y defunciones.** En la parte inferior izquierda se selecciona el reporte deseado. En el área de trabajo se visualiza la tabla correspondiente al reporte.
EC 1.2: Defunciones y tasas de mortalidad según provincia de residencia.		• Año. • Provincia.	• Cantidad de defunciones. • Tasa de mortalidad (por mil habitantes).		
EC 1.3: Nacimientos y tasas de natalidad de Cuba.		• Año.	• Cantidad de nacimientos. • Tasa de natalidad (por mil habitantes).		
EC 1.4: Nacimientos y tasas de natalidad según provincia de residencia.		• Año. • Provincia.	• Cantidad de nacimientos. • Tasa de natalidad (por mil habitantes).		

EC 1.5: Tasa de crecimiento natural de Cuba.		• Año.	• Tasa de crecimiento natural (por mil habitantes).		
EC 1.6: Tasa de crecimiento natural según provincia de residencia.		• Año. • Provincia.	• Tasa de crecimiento natural (por mil habitantes).		

Tabla 11: Caso de prueba Analizar información de nacimientos y defunciones

- **Caso de prueba Analizar información de población residente y media:**

Escenario	Descripción	Perfiles de análisis	Indicadores a medir	Respuesta del sistema	Flujo central
EC 1.1: Población media según provincia y municipio.	Permite visualizar el reporte con las variables presentes en el mismo.	• Año. • División Política Administrativa.	• Población media: total ambas zonas. • Población media: varones ambas zonas. • Población media: hembras ambas zonas. • Población media: total zona urbana. • Población media: varones zona urbana. • Población media: hembra zona urbana. • Población media: total zona rural. • Población media: varones zona rural.	Se muestra la tabla con los valores correspondientes a cada escenario.	Se abre la aplicación. Se autentifica. Se entra al sistema. Se despliega hacia la derecha el componente ubicado en el lateral izquierdo que contiene el navegador. Se selecciona el área de análisis **AA. Series históricas de población.** Se selecciona el libro de trabajo **LT. Población residente y media.** En la parte inferior izquierda se selecciona el reporte deseado. En el área de trabajo se visualiza la tabla correspondiente al

			• Población media: hembras zona rural.		reporte.
EC 1.2: Población residente según provincia y municipio.		• Año. • División Política Administrativa.	• Población residente: total ambas zonas. • Población residente: varones ambas zonas. • Población residente: hembras ambas zonas. • Población residente: total zona urbana. • Población residente: varones zona urbana. • Población residente: hembras zona urbana. • Población residente: total zona rural. • Población residente: varones zona rural. • Población residente: hembras zona rural.		

Tabla 12: Caso de prueba Analizar información de población residente y media

Glosario de términos

El objetivo de esta sección es facilitar la comprensión del contenido que se expone en este documento. A continuación se presentan los términos que podrían resultar nuevos al lector, de difícil comprensión, o de diversos significados dependiendo del contexto que se analiza:

DPA: División Política Administrativa.

Dimensión: característica de un hecho que permite su análisis posterior en el proceso de toma de decisiones y brinda una perspectiva adicional a un hecho dado.

ETL: Proceso que permite extraer, transformar y cargar los datos de un almacén de datos.

Hecho: operación que se realiza en el negocio la cual está estrechamente relacionada con el tiempo y es objeto de análisis para la toma de decisiones.

JDBC: Protocolo de conexión de Java a base de datos (del inglés Java Data Base Connectivity).

NC: No Conformidad.

ONE: Oficina Nacional de Estadísticas.

Perspectiva: se refiere a un objeto mediante el cual se quiere examinar un indicador, con el fin de responder a una pregunta planteada.

PENTAHO: Plataforma Open Source Pentaho Business Intelligence que cubre amplias necesidades de análisis de los datos y de los Informes empresariales.

RN: Regla del negocio.