Universidad de las Ciencias Informáticas

Facultad 6

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

Trabajo de Diploma para optar por el título de

Ingeniero en Ciencias Informáticas

Autora: Evelyn Yanez Clark

Tutor: Ing. Marisel Santana Rodríguez

Ing. Vladimir Urquía Cordero

Julio de 2012

Declaración de autoría

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

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

Datos de contacto

Tutora: Ing. Marisel Santana Rodríguez

Especialidad de graduación: Ingeniería en Ciencias Informáticas

Categoría docente: -

Categoría Científica: -

Años de experiencia en el tema: 1

Años de graduado: 1

Correo Electrónico: msantana@uci.cu

Tutor: Ing. Vladimir Urquia Cordero

Especialidad de graduación: Ingeniería en Ciencias Informáticas

Categoría docente: -

Categoría Científica: -

Años de experiencia en el tema: 1

Años de graduado: 1

Correo Electrónico: vurquia@uci.cu

Agradecimientos

Esta tesis no hubiera podido salir a la luz sin la contribución de muchas personas. Me es imposible referirlas a todas, sin embargo, hay quienes han aportado mucho más de lo que ellos imaginan: en primer lugar, mi tutora Ing. Marisel Santana Rodríguez, quiero agradecerle, por enseñarme con su ejemplo como profesional y su comprensión, a Liván, Leisy, compañeros, por su constancia incondicional de siempre. Me regalaron sus ideas y júbilo.

Sería imposible olvidar al grupo de compañeros que me acompañaron durante todo el tiempo en este bregar. Ellos fueron protagonistas y jueces en todo lo que aquí se expone.

No puedo dejar de mencionar a mi abuela, mi padre, mi hermano, mi esposo, mi suegra, que unido a su constante preocupación por los resultados, sin ellos me hubiera sido muy difícil la materialización de esta investigación.

Por último, voy a mencionar a mi madre, Irma, por su apoyo afectivo, su comprensión y su constante preocupación porque culminara con calidad esta tesis.

A ellos y a todos los que confiaron en mí, muchas gracias.

Dedicatoria

A mi mamá por saber comprenderme y darme su apoyo siempre que lo he necesitado. Gracias mami por el cariño y cuidados que me das todo el tiempo.

A mi familia por toda su preocupación y cariño que me han dado y porque sin ella no hubiera podido llegar a ser quien soy.

Tabla de contenido

Resumen

El incremento impetuoso de las Tecnologías de la Información y las Comunicaciones a nivel mundial, ha generado el procesamiento de datos en todos las esferas de la sociedad, lo que trae aparejado la posibilidad de almacenar un gran número de informaciones que permiten, a través de los Mercados de Datos (MD), utilizar toda la información operacional y convertirla en información estratégica, para su ulterior empleo y potenciar la toma de decisiones.

En Cuba, la necesidad de contar con datos precisos de diferentes áreas de una institución, utilizar una forma de trabajo ordenada y economizar los tiempos en el momento de brindar u obtener una búsqueda estipulada, son factores de importancia en el desarrollo e implantación de los MD, los cuales están enfocados a satisfacer los requerimientos de información internos de determinado organismo con una mayor facilidad de acceso.

La Oficina Nacional de Estadísticas e Información (ONEI) requiere analizar los datos referentes al área Tecnologías de la información para operar acertadamente en cuanto a la toma de decisiones en el país, estos análisis se manejan con herramientas poco factibles para la agilidad y veracidad que se requiere con el monto de información. El departamento de Almacenes de Datos de la Universidad de Ciencias Informáticas (UCI) en contrato con el Sistema de Gobierno (SIGOB) contribuirá con la elaboración de un mercado de datos para la ONEI que le permita obtener la información centralizada acerca de las Tecnologías de la información. Todo esto permitirá integrar datos y realizar análisis estadísticos a nivel nacional.

Palabras claves: Tecnologías de la información, desarrollo, integrar.

Introducción

En la actualidad las Tecnologías de la Información y las Comunicaciones (TIC) ocupan un lugar esencial en el desarrollo de la sociedad y la economía. El concepto de las TIC nace con la convergencia tecnológica de la electrónica, el software y las infraestructuras de las telecomunicaciones y proveen herramientas que ofrecen la posibilidad de encontrar soluciones novedosas ante los desafíos sociales de hoy. (2)

Debido al auge que ha tenido la implantación y utilización de las TIC en todo el mundo, se presentan como una necesidad para el desarrollo económico y social de cualquier país. Esto ha traído como resultado que las empresas informáticas enfrenten cada día un reto mayor para brindar una respuesta rápida y con calidad a sus clientes.

En los últimos años, en Cuba se ha emprendido el reto de lograr la informatización de la sociedad, este proyecto se ha desarrollado de forma acelerada, alcanzando resultados satisfactorios en áreas tan esenciales como la Educación, la Salud y la Investigación. La aplicación de la política de autonomía tecnológica en la que se encuentra inmerso el país, y la selección del software libre como alternativa viable en el desarrollo de sistemas informáticos para la explotación por parte de la infraestructura informática, resulta actualmente una necesidad de primer orden, a la que se ha volcado un enorme interés político y económico.

Uno de los centros de investigación de la Universidad de Ciencias Informáticas (UCI) es el Centro de Tecnología de Gestión de Datos (DATEC). Actualmente en la línea Almacenes de Datos (AD) se realiza el proyecto Sistema de Información de Gobierno (SIGOB) que tiene entre sus objetivos la integración de los datos estadísticos.

La Oficina Nacional de Estadística e Información (ONEI), constituye el órgano rector que obtiene, analiza y difunde, entre otros, los datos estadísticos del país. El área de Tecnologías de la información procesa grandes cantidades de datos de manera ineficiente e inadecuada, cuestión que se torna muy compleja y lenta debido a que los datos se encuentran disgregados en múltiples ficheros. Esta empeora a medida que transcurre el tiempo. Asimismo, como los datos provienen de distintos orígenes se produce la aparición de inconsistencias en estos, influyendo negativamente en la calidad de la información deseada. Por lo anteriormente expuesto, se identifica la siguiente **situación problemática**:

¿Cómo contribuir a la toma de decisiones sobre las Series históricas del área Tecnologías de la información del Sistema de Información de Gobierno? Dicho problema, nos conlleva a

definir como **objeto de estudio**: los Almacenes de Datos, delimitando el **campo de acción** al Mercado de Datos para el área de Tecnologías de la Información.

Para dar solución a la situación problemática, se establece como **objetivo general** Desarrollar el mercado de datos Series históricas Tecnologías de la información para el Sistema de Información de Gobierno, que contribuya a la toma de decisiones.

Para dar cumplimiento al **objetivo general,** los **objetivos específicos** formulados son:

1. Realizar el análisis y diseño del mercado de datos Series históricas Tecnologías de la información.

2. Implementar el mercado de datos Series históricas Tecnologías de la información.

3. Validar el mercado de datos Series históricas Tecnologías de la información.

Para darle cumplimiento a cada uno de los objetivos específicos se proyectan las siguientes **tareas de investigación:**

1. Caracterización de las metodologías, herramientas y tecnologías a utilizar en el desarrollo de Almacenes de Datos.
2. Levantamiento de requisitos.
3. Descripción de los casos de uso del mercado de datos Series históricas Tecnologías de la información para el Sistema de información de Gobierno.
4. Definición de los hechos, las medidas y las dimensiones del mercado de datos Series históricas Tecnologías de la información para el Sistema de información de Gobierno.
5. Diseño del modelo de datos.
6. Definición de la arquitectura del mercado de datos Series históricas Tecnologías de la información para el Sistema de información de Gobierno.
7. Diseño del subsistema de integración.
8. Diseño del subsistema de visualización.
9. Diseño de los casos de pruebas.
10. Implementación del modelo de datos.
11. Implementación del subsistema de integración.
12. Implementación del subsistema de visualización.
13. Aplicación de las listas de chequeo.
14. Aplicación de los casos de prueba.

Estructura de la investigación

El presente trabajo de diploma tiene la siguiente estructura: introducción, cuatro capítulos, conclusiones, recomendaciones, referencias bibliográficas, bibliografía, anexos y glosario de términos.

Capítulo 1: Fundamentos teóricos sobre el desarrollo de un mercado de datos

En el capítulo, se abordan los principales conceptos relacionados con los Almacenes de Datos, sus principales características y las ventajas que proporcionan, su uso en el mundo empresarial. Se exponen además, aspectos fundamentales tales como la metodología de desarrollo a utilizar y las herramientas más empleadas en los procesos de integración de datos e inteligencia de negocio a nivel internacional.

Capítulo 2: Análisis y diseño de un mercado de datos

En el capítulo, se realiza un análisis del negocio con el propósito de comprender mejor los aspectos de mayor relevancia y se propone un diseño de la solución. Se abordan aspectos concernientes a la descripción de las fuentes a integrar, se definen los requisitos de información, los funcionales y no funcionales que debe cumplir el sistema así como el modelo dimensional para el desarrollo del Mercado de Datos a partir de los indicadores que se seleccionaron con las características necesarias para satisfacer las necesidades manifestadas por el cliente. Se definen las reglas del negocio, además, se realiza la matriz BUS, el modelo de datos y el esquema de seguridad.

Capítulo 3: Implementación del mercado de datos

El capítulo está dirigido a la descripción de la implementación de los disímiles aspectos relacionados con los procesos de integración de datos, con el propósito de brindar una mayor comprensión de las estrategias y procedimientos utilizados. Además, se abordan elementos relacionados con la implementación de la capa de inteligencia de negocio, incluyendo la creación de las estructuras necesarias para la navegación y el análisis de los datos.

Capítulo 4: Pruebas

En el capítulo se exponen las pruebas realizadas al mercado de datos, así como los resultados obtenidos en cada una de ellas luego de su aplicación. Dichas pruebas son realizadas para garantizar el cumplimiento de las exigencias del cliente y la calidad del producto final.

Capítulo1: Fundamentos teóricos sobre el desarrollo de un mercado de datos

1. Introducción

En el capítulo, se abordan los principales conceptos relacionados con los Almacenes de Datos, sus principales características y las ventajas que proporcionan, su uso en el mundo empresarial. Se exponen además, aspectos fundamentales como la metodología de desarrollo a utilizar y las herramientas más empleadas en los procesos de integración de datos e inteligencia de negocio a nivel internacional.

1.1 Almacenes de Datos

En la actualidad, y debido al desarrollo tecnológico, la información se hace cada vez más indispensable en la toma de decisiones por parte de los directivos de las disímiles empresas. Un Almacén de Datos (AD), en inglés Data Warehouse (DWH), ofrece la obtención de datos históricamente guardados ya que contiene una copia de los datos de los sistemas operacionales, brinda la posibilidad a los directivos de las organizaciones formular preguntas, realizar consultas y analizar los datos en el momento, forma y cantidad que precisen sin necesidad de tener que acudir al personal informático de la empresa. (1)

La finalidad de los AD, consiste en convertir los datos contenidos en las bases de datos corporativas de las organizaciones en información y ésta, a su vez, en conocimiento útil en el proceso de toma de decisiones estratégicas. (1)

Según Ralph Kimball, conocido autor en el tema de los Almacenes de Datos, lo precisa como “una copia de las transacciones de datos específicamente estructurada para la consulta y el análisis”. William H. Inmon, uno de los primeros autores en escribir sobre esta temática, define un AD como: Una colección de datos orientada al negocio, integrada, no volátil y variante en el tiempo, para el apoyo a la toma de decisiones administrativas.” (2)
Basándonos en la definición que nos proporciona Bill Inmon, podemos definir un AD como un repositorio de datos con las siguientes propiedades:

Orientado a temas: los datos en la base de datos están organizados de manera que todos los elementos de datos relativos al mismo evento u objeto del mundo real queden unidos entre sí.

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 base de datos contiene los datos de todos los sistemas operacionales de la organización, y dichos datos deben ser consistentes.
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.

1.2 Componentes de un Almacén de Datos

Un AD está compuesto por varios elementos necesarios para lograr el cumplimiento de sus objetivos. A continuación, se brinda una explicación de cada uno de ellos:

Sistemas fuentes operacionales: son los sistemas utilizados en las empresas para gestionar sus transacciones, información que es almacenada en diferentes formatos de acuerdo a las necesidades del negocio. Estos sistemas, conservan pocos datos históricos, pues generalmente realizan salvas de la información para trabajar con los datos generados en un corto período de tiempo y de esta forma hacer las recuperaciones más fácilmente. Las prioridades principales que poseen son el procesamiento del rendimiento y la disponibilidad. (3)

Área de procesamiento (staging area): es un área de almacenamiento donde se realizan un conjunto de procesos comúnmente conocidos como extracción, transformación y carga (ETL), en los cuales se invierte la mayor cantidad de tiempo y esfuerzo durante la construcción de un AD. Primeramente, se realiza la extracción de los datos necesarios para el almacén de las diferentes fuentes, para luego pasar por un proceso de transformación donde se eliminan errores e inconsistencias que dificulten su posterior análisis. Finalmente, una vez que los datos están listos para ser almacenados, son cargados en el área de presentación del AD. (3)

Área de presentación: en esta área los datos son almacenados, organizados y puestos a disposición de los usuarios para ser consultados, analizados o realizar reportes sobre ellos. En ella, se almacena toda la información que puede ser de utilidad para el proceso de toma de decisiones en la empresa, diseñada mediante esquemas dimensionales. Generalmente, es referenciada como un conjunto de MD integrados, donde cada uno representa a un proceso específico del negocio. (3)

Herramientas de acceso a los datos: en este componente, se utiliza la palabra herramienta para referirse a la variedad de habilidades que pueden ser provistas a los usuarios del negocio, para soportar el proceso de toma de decisiones. Por definición, su actividad fundamental consiste en consultar la información que se encuentran en el área de presentación. (3)

1.2.1 Ventajas y desventajas de un Almacén de Datos

Un almacén es una base de datos, orientada al análisis de la información histórica contenida en ella, característica que lo ha convertido en una potente herramienta para apoyar el proceso de toma de decisiones. Provee numerosos beneficios en cuanto al proceso de toma de decisiones administrativas, las cuales influyen fuertemente en su desarrollo, pero también presenta sus desventajas, las cuales pueden ser un impedimento para la utilización de este tipo de sistema en una entidad determinada. Seguidamente se mostrarán algunas de las ventajas y desventajas de los AD.

Ventajas

- Soporta diferentes tipos de procesamiento, sin afectar el rendimiento.
- Capaz de actualizarse con las distintas bases de datos para poder almacenar esta nueva información, resumirla, y así poderle dar las características que esta tecnología brinda.
- Hacen más fácil el acceso a una gran variedad de datos a los usuarios finales.
- Se obtiene una base de datos histórica y clasificada por temas.
- Integra información procedente de múltiples sistemas externos.
- Permite un análisis inmediato de las actividades de la empresa.
- Transforma datos orientados a las aplicaciones, en información orientada a la toma de decisiones.
- Capacidad de analizar y explorar las diferentes áreas de trabajo.
- Facilidades en la gestión y análisis de recursos.
- Permite establecer una conexión entre los departamentos empresariales que son independientes y el resto.

Desventajas

- No es muy útil para la toma de decisiones en tiempo real debido al largo tiempo de procesamiento que puede requerir. En cualquier caso la tendencia de los productos actuales (junto con los avances del hardware) es la de solventar este problema convirtiendo la desventaja en una ventaja.
- Requiere de continua limpieza, transformación e integración de datos. Mantenimiento.
- En un proceso de implantación puede encontrarse dificultades ante los diferentes objetivos que pretende una organización.
- Una vez implementado puede ser complicado añadir nuevas fuentes de datos.
- Se necesita de un constante y costoso soporte técnico.

- Debido a su complejidad, AD es muy “frágil” en cuanto a su funcionamiento se refiere, por lo que se hace necesaria una continua revisión.
- Los propietarios de algunos datos (como es el caso de las fuentes) pierden el control de sus datos. Es decir, el AD trabaja con la información recabada indiscriminadamente, enfocándose solamente en el tratamiento de dicha información. Así pues, vemos que es el sistema el que los controla y no el propietario, lo que puede llevar a un plano de inseguridad. (4)

1.3 Mercado de datos

Un MD es una base de datos 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 MD puede ser alimentado desde los datos de un AD, o integrar por sí mismo un compendio de distintas fuentes de información. Los MD son generalmente subconjuntos del AD, diseñados para satisfacer las necesidades específicas de grupos comunes de usuarios.

Dispone de una estructura óptima de datos para analizar la información al detalle desde todas las perspectivas que afecten a los procesos de dicho departamento. En general, un mercado de datos es un subconjunto del almacén de datos enfocado a un departamento o área específica. (5)

Seguidamente se muestran algunas de las características de los MD:

- Se centran en los requisitos de los usuarios asociados a un departamento o área de negocio.
- Como diferencia con los almacenes, los MD no contienen información operacional detallada. Son más sencillos a la hora de utilizar y comprender sus datos, debido a que la cantidad de información que contienen es mucho menor que en los del almacén

Ventajas que proporciona el uso de los MD:

- Son simples de implementar
- Conllevan poco tiempo de construcción y puesta en marcha.
- Permiten manejar información confidencial.
- Reflejan rápidamente sus beneficios y cualidades.
- Reducen la demanda del depósito de datos.

1.3.1 Clasificación de los sistemas OLAP

Para crear el MD de un área funcional de la empresa es preciso encontrar la estructura óptima para el análisis de su información, estructura que puede estar montada sobre una base de datos OLTP (Procesamiento de Transacciones en Línea), como el propio AD, o sobre una base de datos OLAP (El

Procesamiento Analítico en Línea). La designación de una u otra dependerá de los datos, los requisitos y las características específicas de cada departamento.

OLAP

Se basan en los populares cubos OLAP, que se construyen agregando, según los requisitos de cada área o departamento, las dimensiones y los indicadores necesarios de cada cubo relacional. El modo de creación, explotación y mantenimiento de los cubos OLAP es muy heterogéneo, en función de la herramienta final que se utilice.

Existen tres modelos para realizar este proceso: ROLAP, MOLAP y HOLAP. Seguidamente, se explican brevemente las características principales de cada uno de ellos.

ROLAP

En el Procesamiento Analítico Relacional en Línea, en inglés Relational Online Analytical Process (ROLAP), los datos son guardados en filas y columnas de forma racional. Este modelo presenta los datos a los usuarios en forma de dimensiones del negocio, con el objetivo de ocultar las estructuras de almacenamiento a los usuarios y presentar los datos multidimensionalmente.

El modelo ROLAP, es usado fundamentalmente sobre la información que no se consulta frecuentemente, debido a que resulta muy útil cuando se desea consultar información que se almacena durante muchos años. En la siguiente figura se muestra la forma de almacenamiento de ROLAP:

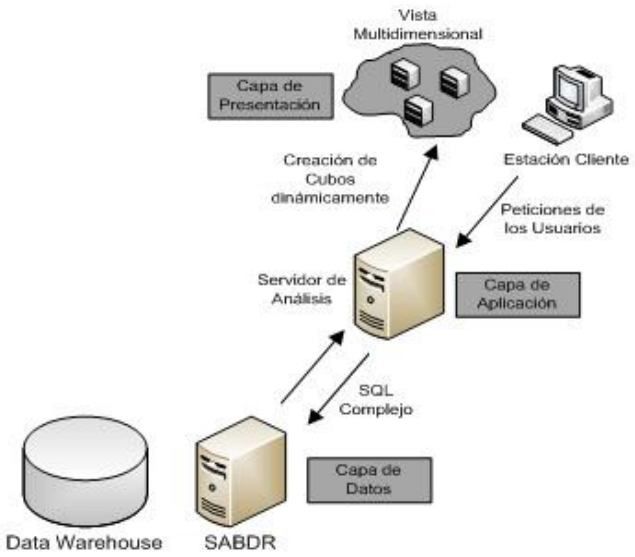

Figura 1: Modelo de almacenamiento ROLAP

MOLAP

El Procesamiento Analítico Multidimensional en Línea, en inglés Multidimensional Online Analytical Process (MOLAP), almacena los datos multidimensionalmente a diferencia del modelo ROLAP. La estructura de los datos en este modelo, es estática para que la lógica al procesar el análisis multidimensional pueda ser basada en métodos bien definidos, con el propósito de establecer las coordenadas del almacenamiento de los datos.

Para realizar el acceso a la información almacenada de forma más rápida y efectiva, las estructuras de almacenamiento se organizan en grandes arreglos dimensionales, que son una copia de la fuente de datos y persisten físicamente en la misma estación de trabajo donde está instalado el DWH. En la siguiente figura se muestra la forma de almacenamiento de MOLAP.

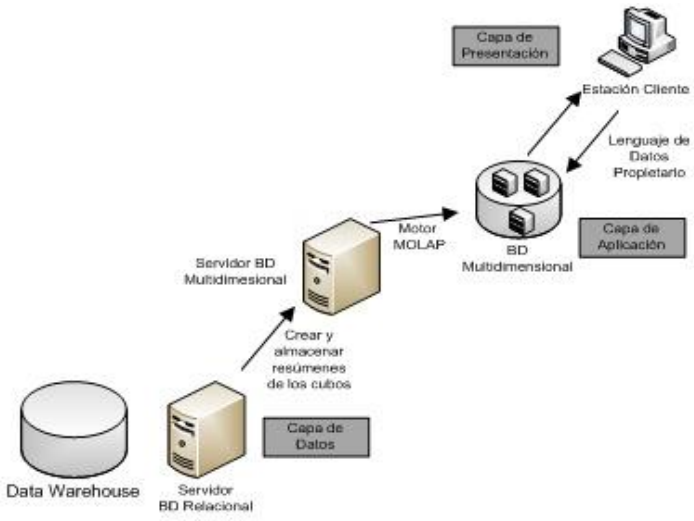

Figura 2: Modelo de almacenamiento MOLAP

HOLAP

El modo de almacenamiento Procesamiento Analítico Híbrido en Línea, en inglés (HOLAP), como su nombre lo indica, es un híbrido entre los métodos ROLAP y MOLAP, que permite almacenar una parte de los datos como en un sistema MOLAP y el resto como en uno ROLAP. Este modelo ofrece 2 tipos de partición de los datos:

Partición vertical: almacena los datos agregados según MOLAP para mejorar la velocidad de las consultas, y los datos en detalle según ROLAP para optimizar el tiempo de procesa-miento del cubo.

Partición horizontal: en este tipo de división una parte de los datos, normalmente los más recientes, es decir divididos según la dimensión Tiempo, se almacenan según el modelo MOLAP, para mejorar la

velocidad de las consultas mientras que los datos más antiguos se guardan según el modelo ROLAP. Aún más se pueden almacenar fragmentos del cubo según MOLAP y otros según ROLAP facilitando el hecho de que en un cubo grande existan subregiones densas en datos y subregiones escasas de datos.

Para la realización del trabajo se decide utilizar como modelo de almacenamiento el Procesamiento Analítico Relacional en Línea (ROLAP). La principal razón por la que se selecciona esta arquitectura es porque se utilizará el Sistema Gestor de Bases de Datos (SGBD) PostgreSQL, el cual solamente permite utilizar un modelo relacional. Adicionalmente, por sus características, este modelo es más escalable para grandes volúmenes e igualmente posee una estructura más dinámica que lo hace atractivo para el cliente.

1.4 Etapas de desarrollo de un Almacén de Datos

Análisis y diseño

Para la elaboración de un mercado de datos es necesaria la etapa de análisis y diseño, con el objetivo de tener un control acerca de las necesidades de los usuarios y poder obtener finalmente un sistema que responda a los intereses del negocio. El análisis es la base fundamental para el desarrollo del mercado de datos, pues a partir de él se sientan las bases para los posteriores procesos de diseño e implementación.

Realizar el proceso de análisis es una tarea compleja, partiendo de que se necesita un estudio del proceso del negocio que se pretende informatizar para entender de manera clara y transparente lo que el usuario necesita. En la fase de análisis se generan un conjunto de artefactos que facilitan el desarrollo del sistema. La realización de dichos artefactos orienta el avance del cumplimiento de las tareas planteadas y garantizan la utilidad y el éxito del diseño de las estructuras.

Es necesario tener una descripción acerca de la organización que se desempeña como cliente, haciendo una definición del negocio en cuestión. Durante el período de análisis se tiene en cuenta el levantamiento de los requerimientos, creando una guía para los desarrolladores en la fase de implementación. Se elabora además el diagrama de diseño de la base de datos, donde se definen las relaciones entre los hechos y dimensiones, así como los diagramas de casos de uso. Se especifican los actores del negocio y del sistema, y su relación con los diferentes casos de uso.

En el diseño es donde se transforman los modelos lógicos conseguidos en la fase de análisis a modelos físicos. En esta fase se realiza el modelo de datos. Se construye también la matriz dimensional y se muestra el modelo de diseño realizado.

ETL, es un proceso de tres etapas en el uso de base de datos y almacenamiento de datos. Excepto para el almacenamiento de datos e inteligencia empresarial, las herramientas ETL se pueden utilizar para mover datos de un sistema operativo a otro. (6)

Para el desarrollo del sistema que es objeto de estudio, se realizará la integración y el análisis de los datos almacenados por el cliente en distintos ficheros "Excel". Para llevar a cabo lo anterior, los datos se recopilarán de diversas fuentes. Una vez obtenidos, se limpiarán y modificarán con la finalidad de satisfacer las necesidades operacionales. Finalmente se cargarán en una base de datos de destino para ser analizados.

Extracción, Transformación y Carga (ETL)

Extracción

La primera parte de un proceso de ETL es extraer los datos de los sistemas de la fuente. Un elemento esencial de la extracción es el análisis de los datos extraídos, dando por resultado una evaluación donde se valida si los datos resuelven un patrón o una estructura prevista. De no cumplirse esta condición, los datos se rechazan o se transforman. La extracción convierte los datos en un formato común para el proceso de la transformación.

Transformación

La etapa de transformación de un proceso ETL consiste en la aplicación de una serie de reglas o funciones a los datos extraídos. Incluye la validación de los registros y su rechazo si no son aceptables. El importe de la manipulación necesaria para este proceso depende de los datos. Existen fuentes de datos que requieren poca transformación, mientras que otras pueden requerir una o más técnicas para cumplir con los requisitos empresariales y técnicos de la base de datos de destino o el almacén de datos. Los procesos más comunes utilizados son la conversión, la limpieza, la normalización, el perfilado y selección.

Carga

La carga es la última etapa del proceso de ETL y se realiza en un repositorio de destino. Existen varias maneras en las que se pueden cargar los datos. Los datos son cargados al almacén durante esta fase. Dependiendo de los requisitos de la organización, este proceso se puede extender.

Inteligencia de Negocio

Hoy en día, las organizaciones están comprendiendo la importancia de la gestión de la información y las ventajas competitivas que implica su uso. Este proceso de gestión consiste en lograr de una manera eficiente el análisis de distintos tipos de datos de la empresa y su entorno, a través de la

explotación de la información por medio de las tecnologías de la información (TI), facilitando la adaptación de aplicaciones para la inteligencia de negocios (Business Intelligence) (7).

La inteligencia de negocios (BI) es un enfoque estratégico para orientar sistemáticamente el seguimiento, la comunicación y la transformación relacionada al débil conocimiento de la información procesable en la cual se basa la toma de decisiones. Una de las actividades más significativas en el ámbito del BI lo constituye el diseño y construcción de los Almacenes de Datos, conocidos como "una colección de datos orientados a un ámbito (empresa, organización), integrada, no volátil y variante en el tiempo, que ayuda al proceso de los sistemas de soporte de decisiones (DSS)". Los AD están ganando cada vez mayor popularidad en las organizaciones. Ellas se están dando cuenta de las ventajas que involucra el análisis de los datos históricos de forma multidimensional para apoyar el proceso de toma de decisiones. (7)

Entre las principales ventajas que ofrece una solución de BI se encuentran las siguientes:

- Rentabilidad, reducción de costos e incremento de la competitividad.
- Proporcionan la capacidad de extraer, depurar y agregar datos de múltiples sistemas de información en un MD o AD independiente.
- Almacenan datos en esquemas multidimensionales para permitir la entrega de información resumida y examinada al detalle, rápidamente.
- Entregan vistas personalizadas y capacidades de consulta, reporte y análisis relevantes que van más allá de las capacidades de informe estándar de los sistemas, permitiendo obtener una mejor comprensión del negocio y tomar mejores decisiones rápidamente. (8)

1.5 Metodología de desarrollo de Almacenes de Datos

Una metodología es el conjunto de métodos o procedimientos de investigación que se siguen para alcanzar una gama de objetivos en una ciencia y rigen una investigación científica o una exposición doctrinal, a la hora de abordar un AD no hay una única metodología en la que basar el diseño, sino que dependiendo del contexto en el que se encuentre la empresa y los objetivos que persiga se puede emplear una u otra metodología. Estas diferentes metodologías se pueden englobar dentro de dos grandes bloques: descendente (top-down) y ascendente (bottom-up) que se corresponden con las metodologías propuestas por Bill Inmon y Ralph Kimball respectivamente. Estos autores merecen una especial atención porque, en muchos aspectos, se consideran los precursores del AD y sus opiniones son muy valoradas en la industria.

El enfoque descendente se adapta a la visión de Bill Inmon, quien considera que el almacén de datos debe responder a las necesidades de todos los usuarios en la organización, y no sólo de un determinado grupo.

La metodología de Inmon tiene un enfoque a modo de explosión en el sentido de que en cierto modo no viene acompañada del ciclo de vida normal de las aplicaciones, sino que los requisitos irán acompañando al proyecto según vaya comprobándose su necesidad. Esta visión de Inmon puede traer consigo mucho riesgo a la compañía, ya que invierte grandes esfuerzos en el desarrollo del AD y no es hasta la aparición de los MD cuando se empieza a explotar la inversión y obtener beneficios.

Esta estrategia se contempla en el marco de que es imposible conocer cuáles son las necesidades concretas de información de una empresa, el ambiente dinámico en que se mueve la organización, el cambio de estructura que conlleva el desarrollo de la nueva plataforma y los consiguientes cambios a los sistemas transaccionales que su introducción implica. Esto hace muy probable que después de la gran inversión en tiempo y recursos en el desarrollo del AD, se haga patente la necesidad de cambios fundamentales que traen consigo altos costos de desarrollo para la organización, poniendo en evidente peligro el éxito de todo el proyecto en sí y que podían ser evitados. Esta metodología para la construcción de un sistema de este tipo es frecuente a la hora de diseñar un sistema de información, utilizando las herramientas habituales como el esquema Entidad/Relación pero al tener un enfoque global, es más difícil de desarrollar en un proyecto sencillo, pues estamos intentando abordar el "todo", a partir del cual luego iremos al "detalle". Esta es otra de las restricciones que trabajan en contra de la metodología de Inmon ya que implica un consumo de tiempo mayor, teniendo como consecuencia que muchas empresas se inclinen por usar metodologías con las que obtengan resultados tangibles en un espacio menor de tiempo. Sin embargo, esta metodología, se contrapone con la metodología ascendente que defienden otros autores como Ralph Kimball, el cual define un AD como: "Una copia de las transacciones de datos específicamente estructurada para la consulta y el análisis". Según el mismo Kimball, un AD no es más que: "la unión de todos los MD de una entidad". Ahora bien, una vez almacenados los datos de la empresa, se pueden emplear aplicaciones para la obtención estructurada de lo que se quiera consultar en cada momento.

Por otro lado el enfoque ascendente es una metodología rápida que se basa en experimentos y prototipos. Es un método flexible que permite a la organización ir más lejos con menores costos. La idea es construir MD independientes para evaluar las ventajas del nuevo sistema a medida que avanzamos. En él, las partes individuales se diseñan con detalle y luego se enlazan para formar componentes más grandes, que a su vez se enlazan hasta que se forma el sistema completo. Las estrategias basadas en el flujo de información ascendente se antojan potencialmente necesarias y suficientes porque se basan en el conocimiento de todas las variables que pueden afectar a los elementos del sistema. (4)

Para el desarrollo del sistema que es objeto de estudio se decidió utilizar como metodología El modelo para el Desarrollo de Soluciones de AD e Inteligencia de Negocio en DATEC.

1.6 Justificación de la metodología a utilizar

El modelo para el Desarrollo de Soluciones de AD e Inteligencia de Negocio en DATEC, se basa en el ciclo de vida Kimball y en la propuesta realizada por Leopoldo Zenaido Zepeda en su tesis de doctorado, adaptada a las necesidades de la UCI, y cubre las fases por las que pasa la construcción de un almacén de datos.Los elementos que influyeron en esta decisión, se exponen a continuación:

- 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.

Se decidió aplicar este modelo ya que cumplía con las cuatro fases por las que pasa la construcción de un almacén de datos:

- Requerimientos y gestión de proyectos.
- Arquitectura técnica.
- Implementación.
- Implantación y crecimiento.

Otra de las metodologías para el desarrollo de los AD es la propuesta por Leopoldo Zenaido Zepeda en su tesis de doctorado, en la cual plantea incluir los casos de uso para guiar el proceso de desarrollo. Los flujos de trabajo que presenta esta metodología son:

- Estudio preliminar o planeación: se realiza el estudio de la entidad cliente para determinar lo que se desea construir y las condiciones que existen para el desarrollo de la misma, la planeación del proyecto, se definen los objetivos, el alcance preliminar, los costos estimados y otras series de actividades.
- Requerimientos: se realiza en dos direcciones, una, identificando las necesidades de información y reglas del negocio; y la otra con un levantamiento detallado de las fuentes de datos a integrar. Es aquí donde se definen los requerimientos a través de la comparación de las necesidades y las reglas del negocio.
- Arquitectura y diseño: aquí se definen las estructuras de almacenamiento, se diseñan las reglas de extracción, transformación y carga, así como la arquitectura de información que regirá el desarrollo de la solución.
- Implementación: se lleva a cabo el diseño físico del repositorio de datos, se crean las estructuras de almacenamiento, el área temporal de almacenamiento, se ejecutan las reglas de ETL y se configuran e implementan las herramientas de BI para la obtención de los requerimientos acordados con el cliente.

- Prueba: se realizan las pruebas de unidad, luego las pruebas de integración y sistema, hasta las pruebas de aceptación con el cliente final.
- Despliegue: consta de dos etapas, despliegue piloto en el cual se configuran los servidores y se instalan las herramientas según la arquitectura definida y se carga una muestra de los datos para demostrar al cliente que el sistema funciona. Posterior a la aceptación del cliente se realiza la carga histórica de los datos y la capacitación y transferencia tecnológica.
- Soporte y mantenimiento: después de haber implantado la solución, se brindan los servicios de soporte en línea, vía telefónica, web u otros, hasta el acompañamiento junto al cliente según el contrato firmado y las condiciones de soporte establecidas.
- Gestión y administración del proyecto: se lleva durante todo el ciclo de vida, es aquí donde se controla, gestiona y chequea todo el desarrollo, los gastos, las utilidades, los recursos, las adquisiciones y demás actividades relacionadas con la gestión del proyecto.

1.7 Herramientas para el desarrollo de un Almacén de Datos

1.7.1 Herramientas de modelado

Para el modelado de la presente investigación, se decide utilizar Visual Paradigm for UML en su versión 6.4, debido a que es una de las herramientas UML más utilizadas para realizar el ciclo de vida completo del desarrollo de un software: análisis, diseño, implementación y despliegue. Es muy fácil de usar, ya que posee una interfaz gráfica amigable y posibilita diseñar todos los tipos de diagramas de clases necesarios, generar código desde cualquiera de estos diagramas y generar documentación. Además, es fácil de instalar, actualizar y sus diferentes ediciones son compatibles entre ellas.

1.7.2 Sistema Gestor de Bases de Datos

Se define como un conjunto de programas que administran y gestionan la información contenida en una base de datos, contribuyendo en la realización de las siguientes acciones: (3)

- Definición de los datos
- Mantenimiento de la integridad de los datos dentro de la base de datos
- Control de la seguridad y privacidad de los datos
- Manipulación de los datos

PostgreSQL v8.4.0: es un potente motor de bases de datos, que tiene prestaciones y funcionalidades equivalentes a muchos gestores de bases de datos comerciales. Está considerado como el gestor de bases de datos de código abierto más avanzado del mundo, ya que proporciona un gran número de características que normalmente sólo se encontraban en gestores comerciales como DB2 u Oracle. (3)

A continuación se listan las características que más lo identifican y que justifican su utilización en el proyecto SIGOB:

- PostgreSQL aproxima los datos a un modelo objeto-relacional, siendo capaz de manejar complejas rutinas y reglas, como por ejemplo: consultas SQL declarativas, control de concurrencia multi-versión, soporte multi-usuario, optimización de consultas y herencia.
- Es altamente extensible pues soporta operadores, funciones, métodos de acceso y tipos de datos definidos por el usuario.
- Soporta la integridad referencial, la cual es utilizada para garantizar la validez de los datos en una base de datos.
- Posee un API flexible.

PgAdmin 3

PgAdmin III es una aplicación gráfica para gestionar el gestor de bases de datos PostgreSQL, siendo la más completa y popular con licencia *Open Source*. Está escrita en C++ usando la librería gráfica multiplataforma wxWidgets, lo que permite que se pueda usar en Linux, FreeBSD, Solaris, Mac OS X y Windows.

1.8 Herramienta de ETL

Data Cleaner 1.5.3: el DataCleaner es una aplicación Open Source para el perfilado, la validación y comparación de datos, ayuda a administrar y supervisar la calidad de los datos con el fin de garantizar que la información sea útil y aplicable a su situación de negocio. Es una aplicación muy fácil de usar, genera sofisticados informes y gráficos que permiten a los usuarios determinar de un vistazo el nivel de calidad de los datos, identificar y analizar la estructura del origen de datos y combinar resultados y gráficos, creando vistas fáciles de interpretar para evaluar la calidad de los datos.

Principales características de DataCleaner:

- Los perfiles de datos se utilizan para calcular y analizar diversas medidas importantes basadas en los valores de los datos.
- Validación de datos: el validador le dará un resultado que puede ser interpretado como bueno o malo.
- Soporta acceso de lectura a muchos tipos de Almacenes de Datos.

Pentaho Data Integration 4.0.1: fue concebida para apoyar el desarrollo de soluciones de (BI) mediante metodologías ágiles, reduciendo y optimizando el ciclo de vida de aplicaciones BI al permitir avanzar de forma paralela en el diseño de las ETL, modelamiento y visualización de datos, que a su vez ayuda a reducir costos, mejorar la productividad y acortar el tiempo necesario para obtener resultados concretos.

1.9 Herramientas para la inteligencia de negocios

Las herramientas de inteligencia de negocio, han sido creadas para ayudar a la toma de decisiones entre las diferentes empresas e instituciones de un estado. Además, permiten mostrar una visión general de todos los procesos de la entidad a sus directivos, facilitando un mejor entendimiento en el análisis y en la presentación de los datos. A continuación se presentarán algunas de estas herramientas y sus principales características.

Pentaho Schema Workbench v3.2.0: es un entorno visual para el desarrollo y prueba de cubos OLAP, los cuales proveen un mecanismo para buscar datos con rapidez y tiempo de respuesta uniforme, independientemente de la cantidad de datos en el cubo o la complejidad del procedimiento de búsqueda. 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 ofrece un editor de esquemas con la fuente de datos subyacente para su validación, además de permitir la ejecución de consultas ´MDX´ contra el esquema y la base de datos.

Mondrian OLAP Server v3.0.4: es un servidor OLAP de código abierto muy popular, que gestiona la comunicación entre una aplicación OLAP escrita en Java y la base de datos con los datos fuentes. El núcleo del servidor Mondrian es similar a ´JDBC´, pero exclusivo para OLAP. Este proporciona la conexión a la base de datos y ejecuta las sentencias ´SQL´. Entre sus principales características, se encuentra las facilidades que brinda para el análisis de grandes volúmenes de información, que se encuentren almacenados en bases de datos que soporten ´JDBC´.

Para el desarrollo de la capa de inteligencia de negocio, se seleccionan estas herramientas ya que además de ser las utilizadas en el centro, poseen las características necesarias para realizar esta última etapa en el desarrollo de un almacén de datos.

Apache Tomcat v5.5: fue creado bajo el proyecto Jakarta de la Fundación Apache, siendo mantenido por una comunidad de desarrollo, logrando destacar como un producto robusto y altamente eficiente. Esta herramienta es gratis, fácil de instalar y se puede ejecutar en máquinas con pocos recursos, así como también es compatible con las API más recientes de Java. Debido a que su código binario posee un tamaño total de un poco más de un megabyte, no ocupa mucho espacio de modo que no resulta extraño que se ejecute tan rápidamente. Además, su remarcada estabilidad y la capacidad de ejecución multiplataforma, lo han colocado como uno de los servidores más utilizados para el despliegue de aplicaciones web basadas en la tecnología Java.

Pentaho BI Server 3.8.0: la plataforma Pentaho BI Server provee el soporte y la infraestructura necesarios para crear soluciones de inteligencia empresarial a problemas de negocios. El marco proporciona los servicios básicos, incluidos autenticación, registro, auditoría, servicios web y motor de reglas. La plataforma también incluye un motor de solución que integra reportes, análisis, tableros de comandos y componentes de minería de datos. El diseño modular y arquitectura basada en plug-in permite a todos o parte de la plataforma estar inmersa en aplicaciones de terceros por los usuarios finales, así como fabricantes de equipos originales.

La aplicación Pentaho BI Server funciona como un sistema basado en administración web de informes, el servidor de integración de aplicaciones y un motor de flujo de trabajo ligero. Está diseñado para integrarse fácilmente en cualquier proceso de negocio.

Algunas de sus ventajas son:

- Integración con procesos de negocio
- Administra y programa reportes
- Administra seguridad de usuarios

1.10 Conclusiones parciales del capítulo

En el presente capítulo se abordaron los principales conceptos relacionados con los AD, así como sus principales características, ventajas y desventajas, ofreciendo una mejor comprensión del tema.
Se seleccionó como metodología de desarrollo el Modelo para el Desarrollo de Soluciones de Datos e Inteligencia de Negocios en DATEC, que tiene como base la propuesta por Ralph Kimball y las necesidades del centro. Como herramienta de modelado el Visual Paradigm 6.4 para UML, como como gestor de base de datos PostgreSQL y como administrador de este gestor el PgAdmin 3, para la integración de los datos se escogió el Pentaho Data Integration (PDI) y el DataCleaner para realizar el perfilado de datos. Para el proceso de inteligencia de negocio, se decidió utilizar el Pentaho Schema Workbench, el Pentaho BI Server y como motor OLAP de esta suite el Mondrian OLAP Server, además se seleccionó como servidor web el Apache Tomcat.

Capítulo 2: Análisis y diseño del mercado de datos series históricas tecnologías de la información para el sistema de información de gobierno.

2.1 Introducción

En el capítulo, se realiza un análisis del negocio con el propósito de ganar en claridad en los aspectos de mayor relevancia y se propone un diseño de la solución. Se abordan aspectos concernientes a la descripción de las fuentes a integrar, se definen los requisitos de información, los funcionales y no funcionales que debe cumplir el sistema así como el modelo dimensional para el desarrollo del MD a partir de los indicadores que se seleccionaron con las características necesarias para satisfacer las necesidades manifestadas por el cliente. Se definen las reglas del negocio, además, se realiza la matriz BUS, el modelo de datos y el esquema de seguridad.

2.2 Caracterización de las áreas del negocio

La ONEI es el órgano rector encargada de recopilar toda la información proveniente de las diferentes provincias del país. La entidad cuenta con 23 áreas de trabajo, una de ellas es la de Tecnologías de la Información donde todos sus datos son almacenados en ficheros “Excel”. La información que se guarda en estos ficheros es recogida por la Oficina Municipal de Estadísticas de las provincias donde la misma proviene de los centros informantes que pertenecen a un organismo y luego se envía a la Oficina Provincial de Estadísticas en la cual se hace un resumen por municipios de la información. Posteriormente a esto los especialistas de la ONEI son los responsables de limpiar los datos y crear un fichero general, el cual es publicado en el sitio de la misma para así ponerla a disposición de todo el que desee consultarla.

2.3 Necesidades de los usuarios

El Mercado de Datos Tecnologías de la Información persigue como objetivo mejorar el trabajo en el área con que cuenta la ONEI de esta temática, para ello se gestionan las necesidades informáticas de los usuarios del departamento. A continuación se detallan las necesidades de usuarios identificadas:
Análisis y difusión de los diferentes indicadores referentes a Tecnologías de la Información que es recogida en los “Excel” de las series históricas, estos indicadores están dados por el uso social de las TIC en el sector de la salud, la educación y la población, así como recoger indicadores de inversiones e ingresos de las TIC de infraestructura, tráfico internacional de telefonía, y físicos. Toda la información es recogida anualmente.

2.4 Reglas del negocio

1. Las cifras deben estar representadas por valores numéricos.
2. Todas las medidas identificadas deben ser valores positivos.

2.5 Especificación de requerimientos

2.5.1 Requerimientos de información

Para lograr satisfacer las necesidades del cliente, se hace necesario identificar los requisitos de información que son los requerimientos disponibles para el usuario final a la hora de realizar las consultas para analizar los datos. A continuación se definen las siguientes necesidades existentes en la ONEI en la temática de Tecnologías de la información:

- RI 1 - Obtener cantidad de indicadores por infraestructura y por año.
- RI 2 - Obtener cantidad de indicadores físicos por servicio internacional de teléfonos y por año.
- RI 3 - Obtener cantidad de indicadores por los servicios de correo y telégrafo por año.
- RI 4 - Obtener cantidad de indicadores físicos de las TICs por indicadores por año.
- RI 5 - Obtener cantidad de indicadores TICs por entidades y por año.
- RI 6 - Obtener uso social de las TICs en el sector educacional por tipo de enseñanza, por indicadores sector educación por año.
- RI 7 - Obtener uso social de las TICs en el sector salud por indicadores del sector salud, unidades de servicio por año.
- RI 8 - Obtener uso social de las TICs en el sector población por indicadores del sector población, por localización por año.
- RI 9 - Obtener ingresos del Comercio de TICs por tipos de ingresos, por indicadores del comercio por año.
- RI 10 - Obtener inversiones en las Tecnologías de la Información por indicadores de ejecución física, por entidad por año.

2.5.2 Requerimientos funcionales

Los requerimientos funcionales representan las capacidades y condiciones que el sistema debe cumplir para dar respuesta a los requerimientos de información. A continuación se muestran los requerimientos funcionales que fueron identificados:

- RF 1 - Autenticar usuario.
- RF 2 - Adicionar rol.
- RF 3 - Eliminar rol.
- RF 4 - Adicionar usuario.

- RF 5 - Eliminar usuario.
- RF 6 - Adicionar reporte.
- RF 7 - Modificar reporte.
- RF 8 - Eliminar reporte.
- RF 9 - Realizar extracción de archivos “Excel” de Series históricas de transporte.
- RF 10 - Realizar transformación y carga de información de archivos “Excel” de Series históricas de transporte.
- RF 11 - Abrir navegador OLAP.
- RF 12 - Mostrar editor MDX.
- RF 13 - Mostrar padres.
- RF 14 - Ocultar repeticiones.
- RF 15 - Intercambiar ejes.
- RF 16 - Mostrar gráfico.
- RF 17 - Configurar gráfico.
- RF 18 - Configurar impresión.
- RF 19 - Exportar a “PDF”.
- RF 20 - Exportar a “Excel”.
- RF 21 - Mostrar propiedades.
- RF 22 - Suprimir filas.
- RF 23 - Detallar miembros.
- RF 24 - Entrar en detalles.
- RF 25 - Mostrar datos de origen.

2.5.3 Requerimientos no funcionales

Los requerimientos no funcionales son las propiedades o cualidades que el sistema debe cumplir. Estos determinan como debe comportarse el producto. En la presente investigación fueron definidos 26 requerimientos no funcionales los cuales se encuentran contemplados en el artefacto Especificación de Requisitos de Software, a continuación se muestra un ejemplo de ellos:

- RNF 1 - Usabilidad: Cumplir con las pautas de diseño de las interfaces.
- RNF 2 - Usabilidad: Mostrar los mensajes, títulos y demás textos que aparezcan en la interfaz del sistema en idioma español.
- RNF 3 - Usabilidad: Establecer tiempo de entrenamiento requerido para que usuarios normales sean productivos operando el sistema.
- RNF 4 - Usabilidad: Diseñar un reporte del Almacén de Datos de manera sencilla y ágil.
- RNF 5 - Usabilidad: Navegar en los reportes del Almacén de Datos de manera ágil.

- RNF 6 - Confiabilidad: Garantizar la persistencia de la información.
- RNF 7 - Confiabilidad: Garantizar el cumplimiento de actualización de los datos en el almacén.
- RNF 8 - Eficiencia: El sistema debe permitir la abundancia de usuarios conectados simultáneamente sin que se afecte el tiempo de respuesta, al menos 700.
- RNF 9 - Eficiencia: El tiempo de respuesta debe ser en tiempo real.
- RNF 10 - Eficiencia: En el proceso de integración solo tendrá conectado un usuario que tendrá la tarea de vigilar el proceso de integración de datos.
- RNF 11 - Soporte: Lograr que los elementos definidos en el almacén tengan una nomenclatura homogénea.
- RNF 12 - Restricciones del Diseño: Utilizar el lenguaje de programación definido durante la investigación.
- RNF 13 - Requisitos para la documentación de usuarios en línea y ayuda del sistema: Confeccionar manual de usuario.
- RNF 14 - Interfaz: Acceder al sistema.
- RNF 15 - Interfaces de usuario: Los requerimientos de interfaz de usuario se centran en la presentación de la información de cara al cliente.
- RNF 16 - Interfaz: Los reportes estadísticos deben contar con una interfaz simple que facilite la interacción usuario-aplicación.
- RNF 17 - Interfaz: Las interfaces de salida no serán cargadas con información innecesaria.
- RNF 18 - Interfaz: Los gráficos serán con los colores establecidos por la ONEI ajustándose a los estándares establecidos de un buen diseño.
- RNF 19 - Interfaces Hardware: Proporcionar características mínimas de hardware a las estaciones de trabajo.
- RNF 20 - Interfaces Hardware: Proporcionar características mínimas de hardware a los servidores.
- RNF 21 - Interfaces Software: Instalar en las estaciones de trabajo el software necesario para el correcto funcionamiento del sistema.
- RNF 22 - Interfaces Software: Las configuraciones de software de las máquinas clientes deben contar al menos con los siguientes elementos: Firefox 2.0 o superior, Java Virtual Machine, Schema Workbench en caso de que un usuario capacitado requiera la construcción de esquemas multidimensionales para el diseño de nuevos reportes y Pentaho Data Integrator.
- RNF 23 - Interfaces de Comunicación: La comunicación entre la base de datos de integración y el almacén de datos es a través del protocolo TCP/IP.

- RNF 24 - Interfaces de Comunicación: El sistema necesita estar conectado directamente a un dispositivo de red.
- RNF 25 - Requisitos de Licencia: las herramientas a utilizar se encuentran bajo licencia GPL.
- RNF 26 - Requisitos 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

Durante la fase de análisis y diseño de un almacén de datos, se definen los casos de uso del sistema, que tienen como objetivo describir la relación entre la aplicación y los usuarios. En ellos, se agrupan todos los requisitos funcionales y de información que hayan sido identificados con anterioridad, proporcionando una mejor comprensión del sistema. En la figura 3 se muestra el Diagrama de Casos de Uso del Sistema (DCUS) para la solución.

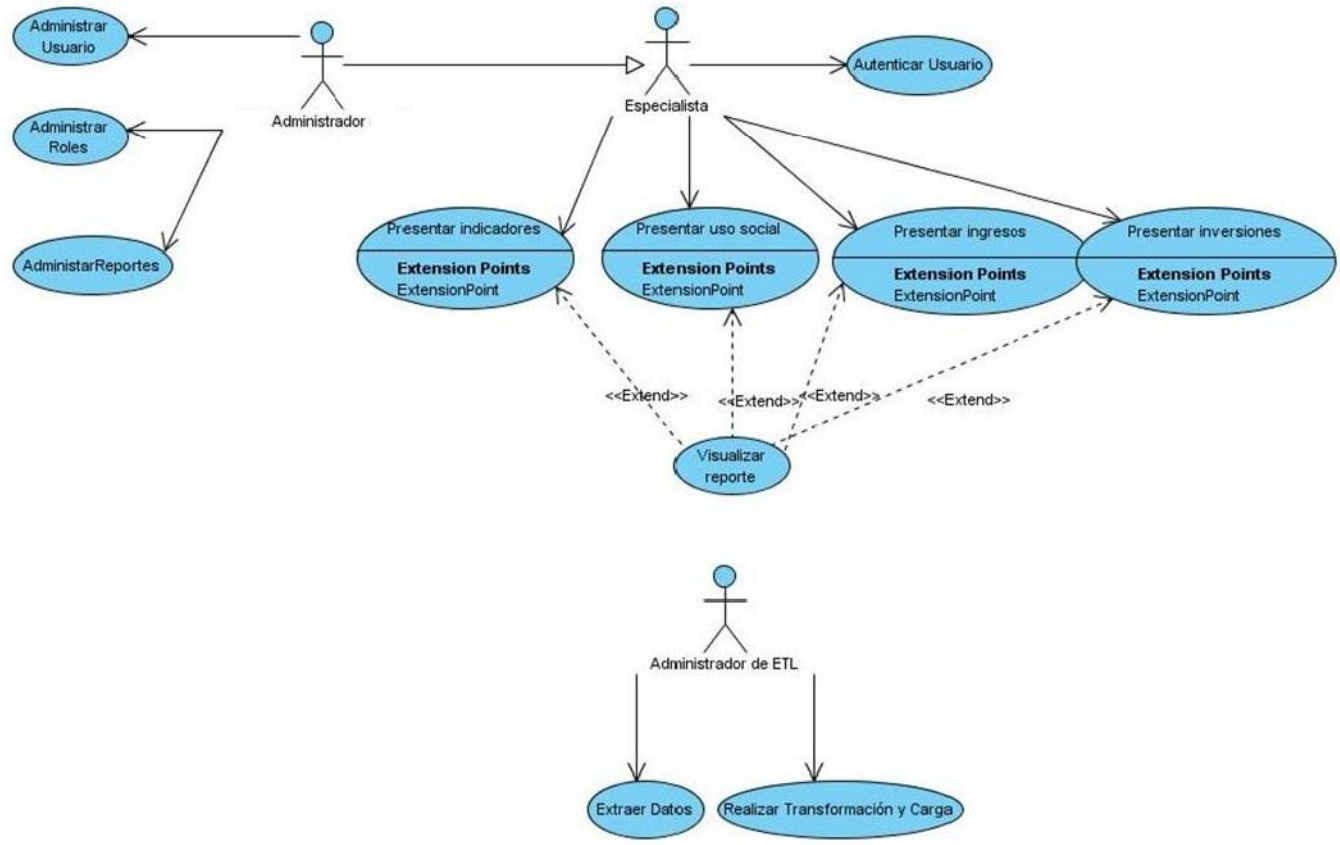

Figura 3: Diagrama de Casos de Uso del Sistema

2.7 Diseño de la solución

El diseño de un mercado de datos es el sostén de la solución a las necesidades planteadas por el cliente. En esta fase se construye la matriz BUS y el modelo de datos.

2.7.1 Matriz bus o matriz dimensional

La Matriz Bus representa las relaciones existentes entre los hechos y las dimensiones del modelo de datos.

	dimensiones								
Hechos	tipo _sector	Tipo _ingreso	criterio_entidad	tráfico_telefonia	indicador _tic	nivel_ escolaridad	Inversion _tic	temporal _anno	dpa
servicio_telefonia				X				X	X
tic_entidades			X					X	X
sector_poblacion	X							X	X
Ingresos		X						X	X
tic_sector_educacion						X		X	X
general_tic					X			X	X
sector_salud	X							X	X
Inversiones_tic							X		

Tabla1: Matriz dimensional o matriz bus

2.7.2 Modelo de datos

El modelado dimensional, es una técnica para lograr que las bases de datos fueran más simples y entendibles, pero ha llegado a ser ampliamente aceptada como la técnica dominante para la presentación de los almacenes de datos, también resulta beneficioso en cuanto a diseño, pues posibilita una mejor comprensión de la aplicación por parte de los usuarios, un mayor rendimiento en las consultas y flexibilidad ante los cambios. [9]

En el modelo de datos, se definen las dimensiones y hechos que serán las futuras tablas de la base de datos de la solución. Debido a esto, se hace necesario conocer cada uno de sus componentes, los cuales se explican a continuación con el propósito de facilitar su entendimiento:

Hechos

En un modelo dimensional, una tabla de hechos representa una transacción o un evento del negocio y en ella se almacenan un conjunto de medidas o atributos, que permiten cuantificar o medir el rendimiento en los diferentes procesos del mismo. Generalmente, estas tablas poseen su propia llave

primaria que se forma por la unión de las llaves pertenecientes a las dimensiones que se relacionan con ella, por lo que también se conoce como llave compuesta. [9]

Dimensiones

Por su parte, las tablas de dimensiones tienden a ser poco profundas en cuanto a la cantidad de filas, pero suelen ser bastante amplias en cuanto al número de columnas o atributos. Estos últimos, son la fuente principal de las restricciones de las consultas y los agrupamientos, por lo que en una consulta o una solicitud de reporte, son los diferentes aspectos por los que se pueden analizar las medidas de los hechos. Además, cada dimensión debe tener una llave primaria que sirva como base para la integridad referencial con cualquier tabla de hechos con la que se relacione. [9]

El diseño del modelo de datos de la presente investigación presenta ocho hechos, nueve dimensiones, de ellas tres compartidas, seis propias y 29 medidas, evidenciándose una topología constelación de hechos, debido a que está compuesto por una serie de esquemas de estrella, es decir, una tabla de hechos central con otras auxiliares y sus respectivas tablas de dimensiones.

Luego del análisis realizado a la fuente de datos recogida en la ONEI se identificaron los siguientes hechos y dimensiones.

Hechos

- hech_tic_entidades
- hech_sector_poblacion
- hech_sector_salud
- hech_ingresos
- hech_tic_sector_educacion
- hech_servicio _telefonia
- hech_general_tic
- hech_inversiones_tic

Dimensiones

- dim_temporal_anno
- dim_dpa
- dim_tipo_sector
- dim_tipo_ingreso
- dim_nivel_escolaridad
- dim_ criterio_entidad
- dim_trafico_telefonia
- dim_indicador
- dim_inversion_tic

A continuación, se muestran las figuras en las que se representan los hechos con sus correspondientes dimensiones.

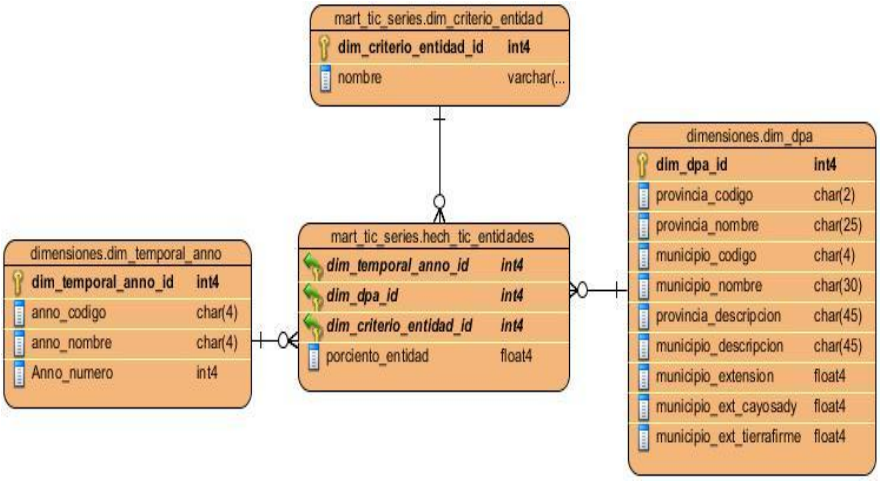

Figura 4: Modelo de datos dimensional. hech_tic_entidades

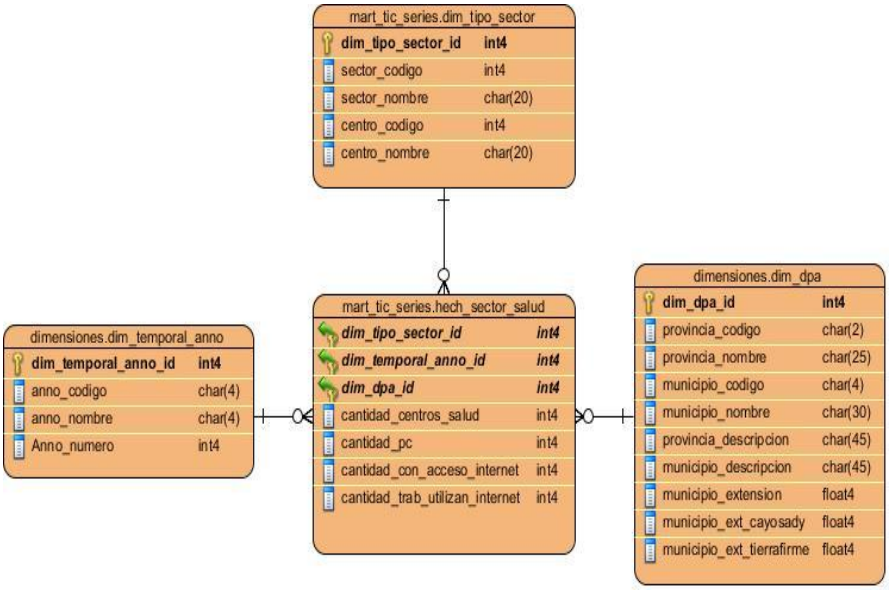

Figura 5: Modelo de datos dimensional. hech_sector_salud

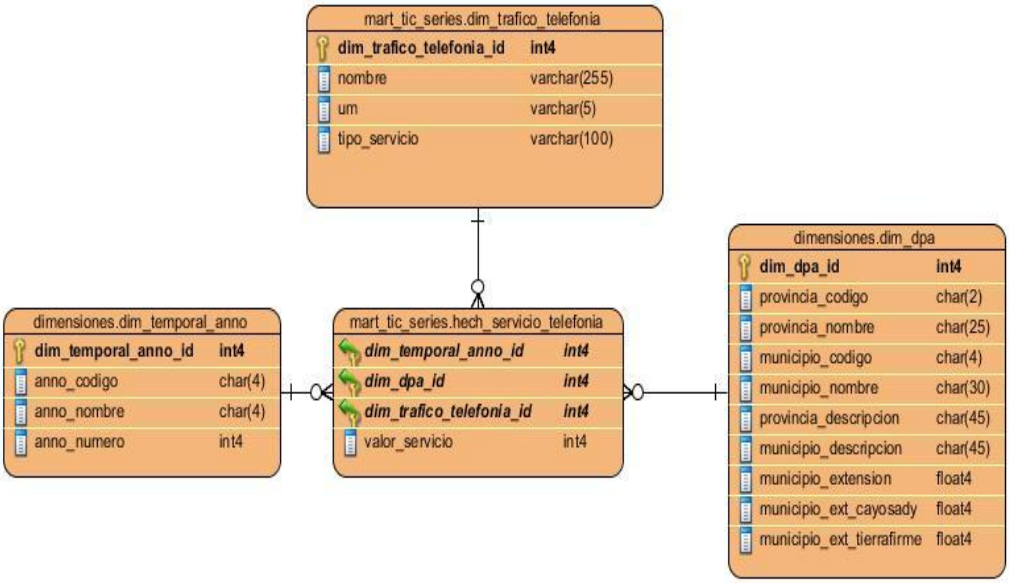

Figura 6: Modelo de datos dimensional. hech_servicio_telefonia

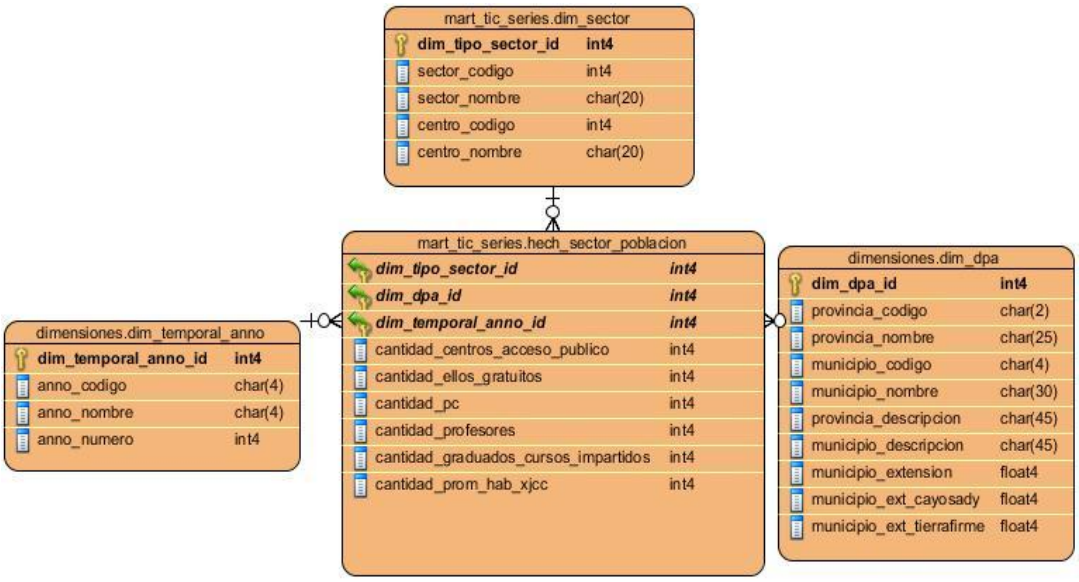

Figura 7: Modelo de datos dimensional. hech_sector_poblacion

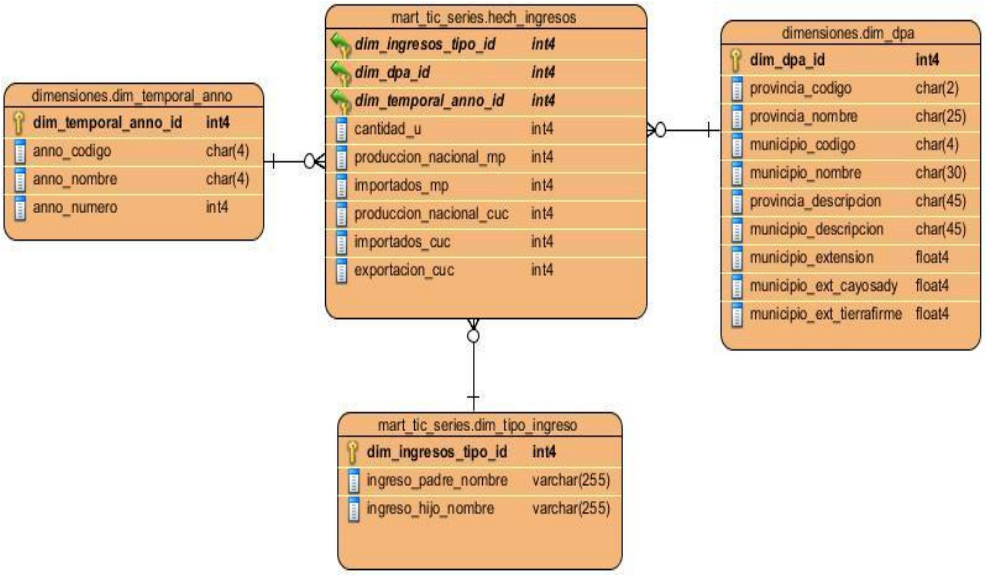

Figura 8: Modelo de datos dimensional. hech_ingresos

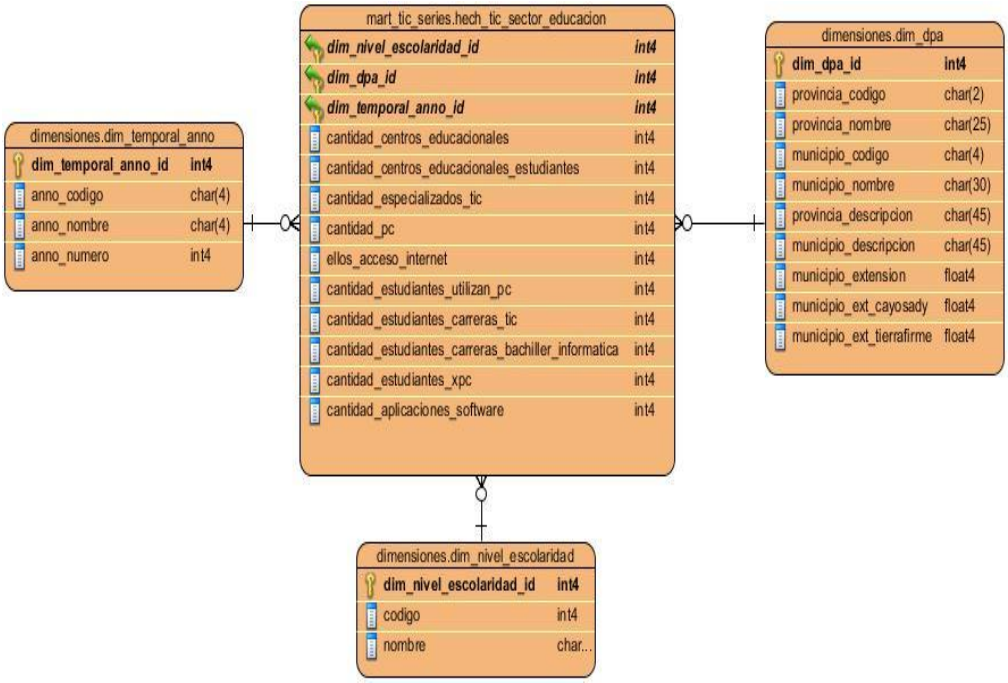

Figura 9: Modelo de datos dimensional. hech_educacion

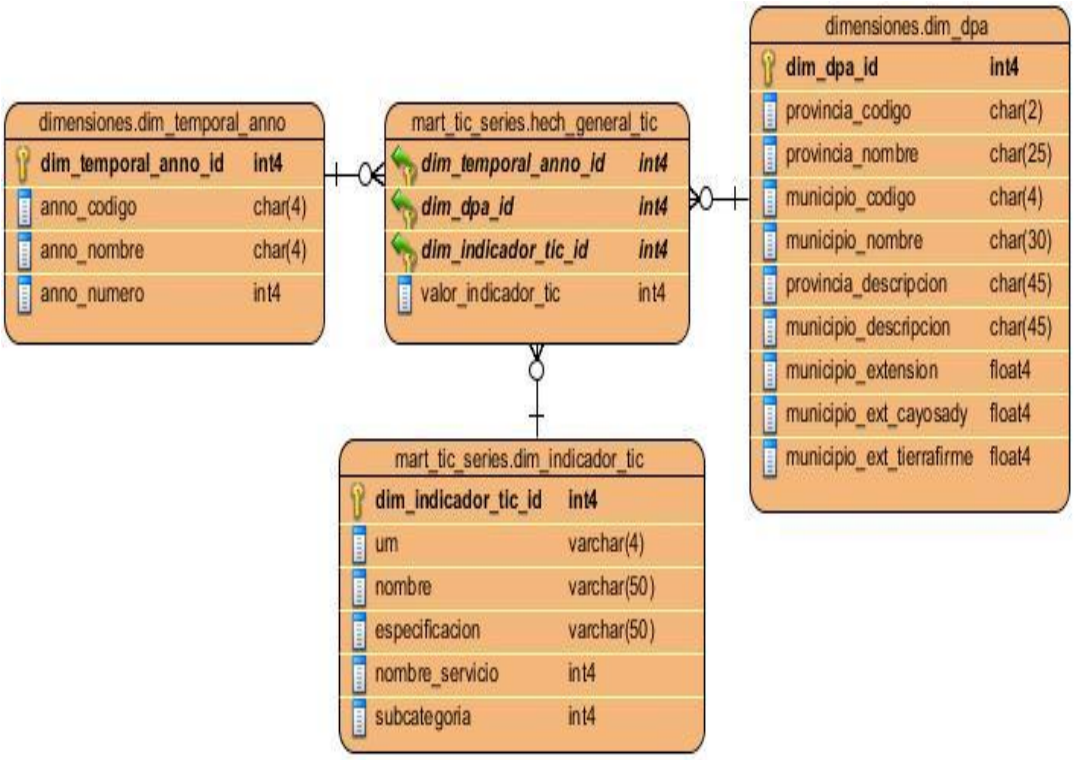

Figura 10: Modelo de datos dimensional. hech_general_tic

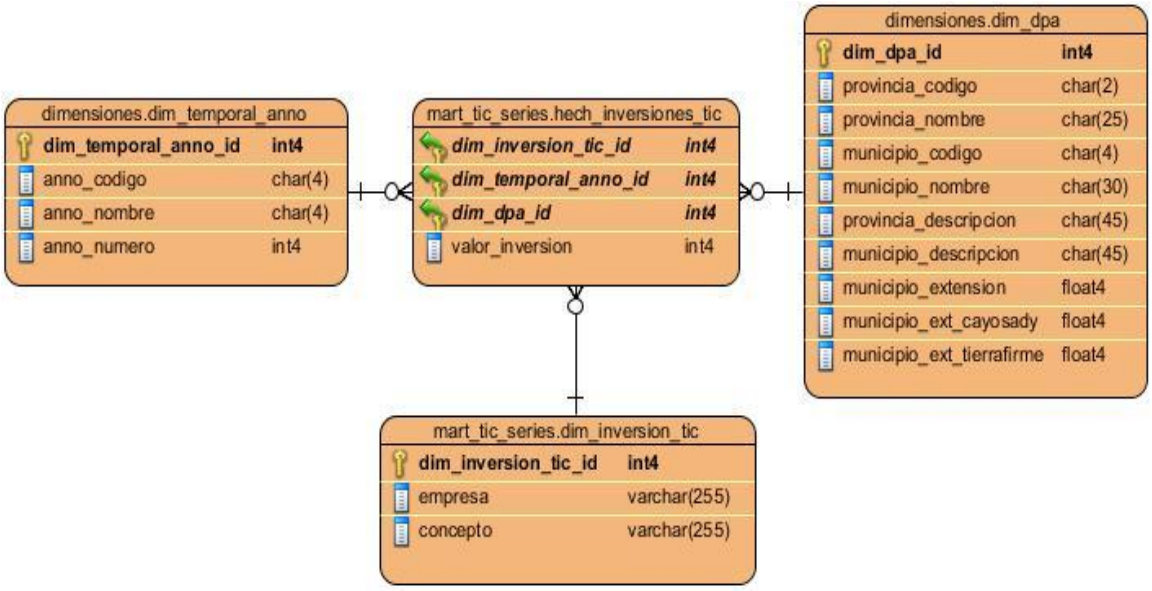

Figura 11: Modelo de datos dimensional. hech_inversiones_tic

2.8 Política de respaldo y recuperación

En el MD Series históricas Tecnologías de la información se utiliza una política de respaldo y recuperación sólida, midiéndose en tres aspectos fundamentales:

- Periodicidad de las salvas: las salvas de toda la información contenida en la BD se realizan anualmente por ser una serie histórica que contiene datos anuales, verificando en todo momento que exista una copia escrita de la información almacenada en el servidor.
- Tablas involucradas: las tablas que se involucran en la realización son las ocho tablas de hechos identificadas en el proceso de análisis.
- Salvas existentes: a pesar de que actualmente no existen salvas en esta área se prevé la realización de reemplazos de estas cada un año, así como también el chequeo de su estado mensualmente, mediante pruebas de rendimiento y flexibilidad.

2.9 Esquema de seguridad

La seguridad en el MD Series históricas Tecnologías de la Información está dada por los niveles de acceso al sistema, regida fundamentalmente por los permisos y roles que los usuarios tienen a la hora de interactuar con la base de datos y la aplicación. Para la seguridad en la base de datos se decidió definir un rol de Administrador el cual posee total acceso a la base de datos del sistema; y para la seguridad en la aplicación de definen los roles que se muestran a continuación:

Roles	Permisos
Administrador	Tiene acceso total a todas las Áreas de Análisis General (AAG). Gestiona el Sistema de información de gobierno.
Analista	Tiene acceso de solo lectura al AA Series históricas Tecnologías de la Información.

Tabla 2: Roles y permisos

Elementos de aplicación	Roles con acceso
AA General	Administrador
Carpeta raíz: AA Series históricas Tecnologías de la Información	Administrador y Analista

Tabla 3: Nivel de acceso a los elementos de aplicación.

2.10 Definición de la arquitectura base del mercado de datos

La arquitectura base del MD Series históricas Tecnologías de la Información se define a partir de los tres subsistemas fundamentales que lo componen: el subsistema de almacenamiento, el subsistema de integración y el subsistema de visualización (como se muestra en la figura 6).

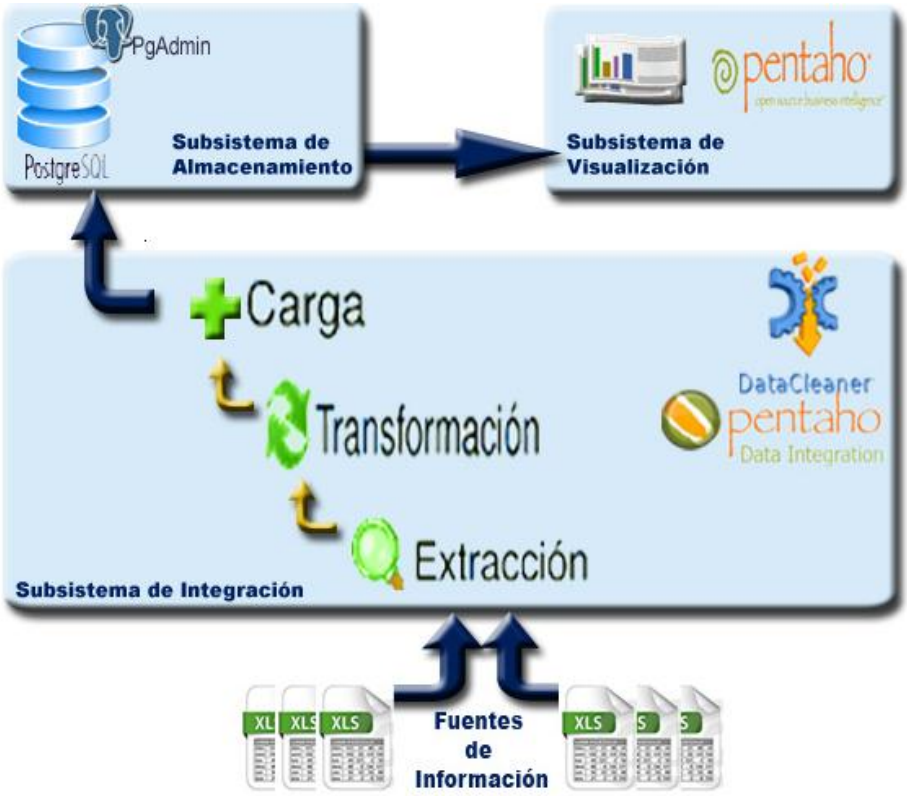

Figura 12: Arquitectura del mercado de datos Tecnologías de la información

A continuación se explicarán cada uno de los subsistemas que definen la arquitectura base del MD:

- **Subsistema de almacenamiento**: es el encargado de contener toda la información correspondiente al área Tecnologías de la información. El almacén estará compuesto por nueve dimensiones y ocho tablas de hechos que a su vez contendrán los datos que describirán un hecho, así como las medidas.
- **Subsistema de integración**: se encarga de integrar, estandarizar y limpiar la información recogida de la serie Tecnologías de la información para cargarla al almacén, luego de la realización de un total de 14 transformaciones, seis dimensiones propias y ocho hechos.
- **Subsistema de visualización**: tiene como objetivo principal consultar la información contenida en el AD para así mostrar los reportes donde se visualizan los datos de Tecnologías de la información que el cliente desea analizar a través del diseño de ocho cubos y 10 reportes que responden a las necesidades del cliente.

2.11 Conclusiones parciales del capítulo

En este capítulo se identificaron 10 requisitos de información, 25 requisitos funcionales y 26 requisitos no funcionales, ocho tablas de dimensiones y siete tablas de hechos, se confeccionó la matriz BUS y el modelo de datos, además se detectaron dos reglas de negocio. Además, se definió el DCUS y el modelo de datos dimensional donde se reflejan las nueve dimensiones y ocho hechos identificados. Finalmente, se explicó cómo se manejará la seguridad de los datos en la aplicación y las políticas a seguir para la recuperación de los datos, en caso de ocurrir algún incidente que atente contra la integridad de los mismos.

Capítulo 3: Implementación del mercado de datos series históricas tecnologías de la información para el sistema de información de gobierno.

3.1 Introducción

El capítulo está dirigido a la exposición del modo en que se realiza la implementación de los diferentes aspectos relacionados con los procesos de integración de datos, con el propósito de brindar una mejor comprensión de las estrategias y procedimientos utilizados. También, se abordan elementos relacionados con la implementación de la capa de inteligencia de negocio, entre los que se incluye la creación de las estructuras necesarias para la navegación y el análisis de los datos.

3.2 Modelo de datos físico

El modelo de datos es una representación de un diseño de datos, que tiene en cuenta las facilidades y restricciones de los sistemas de base de datos. Permite describir las estructuras de datos de la base de datos y la forma en que estos se relacionan. Para esta investigación el modelo de datos físicos queda conformado como se muestra en la figura 13.

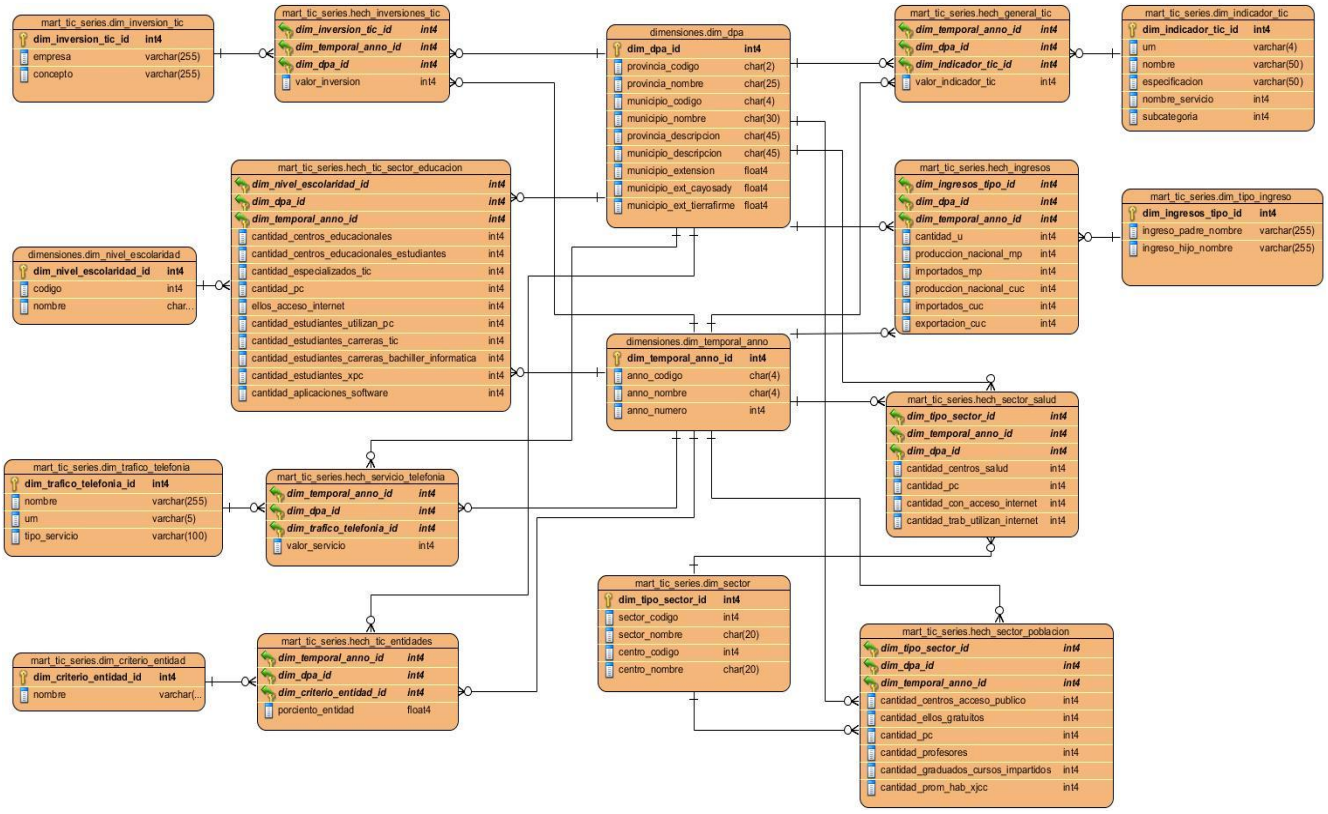

Figura 13: Modelo de datos ffísico.

El modelo de datos físico se puede clasificar dependiendo de los tipos de conceptos que ofrece para describir la estructura de la base de datos, asimismo proporciona conceptos que describen los detalles de cómo se almacenan los datos en el ordenador.

3.3 Esquemas y tablas

Los esquemas son formas de organización de la base de datos que pueden contener tablas, tipos de datos, funciones y operadores. Permiten organizar los objetos de la base de datos en grupos lógicos, posibilitando así utilizar el mismo nombre para un objeto en esquemas diferentes sin ocasionar un conflicto. Para el desarrollo del sistema propuesto en esta investigación se definieron dos esquemas: el esquema dimensiones, que contiene las dimensiones comunes del almacén de datos, y el esquema mart_tic_series, que agrupa todos los hechos de dicho mercado y las dimensiones propias del mercado de datos. La solución cuenta con nueve tablas dimensiones y ocho tablas hechos, distribuidas por los dos esquemas propuestos con anterioridad quedando de la siguiente manera:

Esquemas	Tablas
dimensiones	dim_nivel_escolaridad
dimensiones	dim_dpa
dimensiones	dim_temporal_anno
mart_tic_series	dim_criterio_entidad
mart_tic_series	dim_indicador_tic
mart_tic_series	dim_inversion_tic
mart_tic_series	dim_sector
mart_tic_series	dim_tipo_ingreso
mart_tic_series	dim_trafico_telefonia
mart_tic_series	hech_inversiones_tic
mart_tic_series	hech_sector_poblacion
mart_tic_series	hech_sector_salud
mart_tic_series	hech_sector_educacion
mart_tic_series	hech_servicio_telefonia
mart_tic_series	hech_tic_entidades

mart_tic_series	hech_general_tic
mart_tic_series	hech_ingresos_ti

Tabla 4: Esquemas y tablas identificadas.

3.4 Proceso de integración de datos

3.4.1 Perfilado de datos

El perfilado de datos consiste en realizar un primer análisis sobre los datos de origen, con el objetivo de empezar a conocer su estructura, formato y nivel de calidad. A partir de este proceso se establecen reglas para corregir los defectos de los datos y así garantizar la disponibilidad de los mismos. Para el análisis de los datos de Tecnologías de la información se utilizó la herramienta DataCleaner 1.5.4, la cual nos permitió obtener los resultados que se muestran en la figura 13, los que se emplearán posteriormente al realizar el proceso ETL.

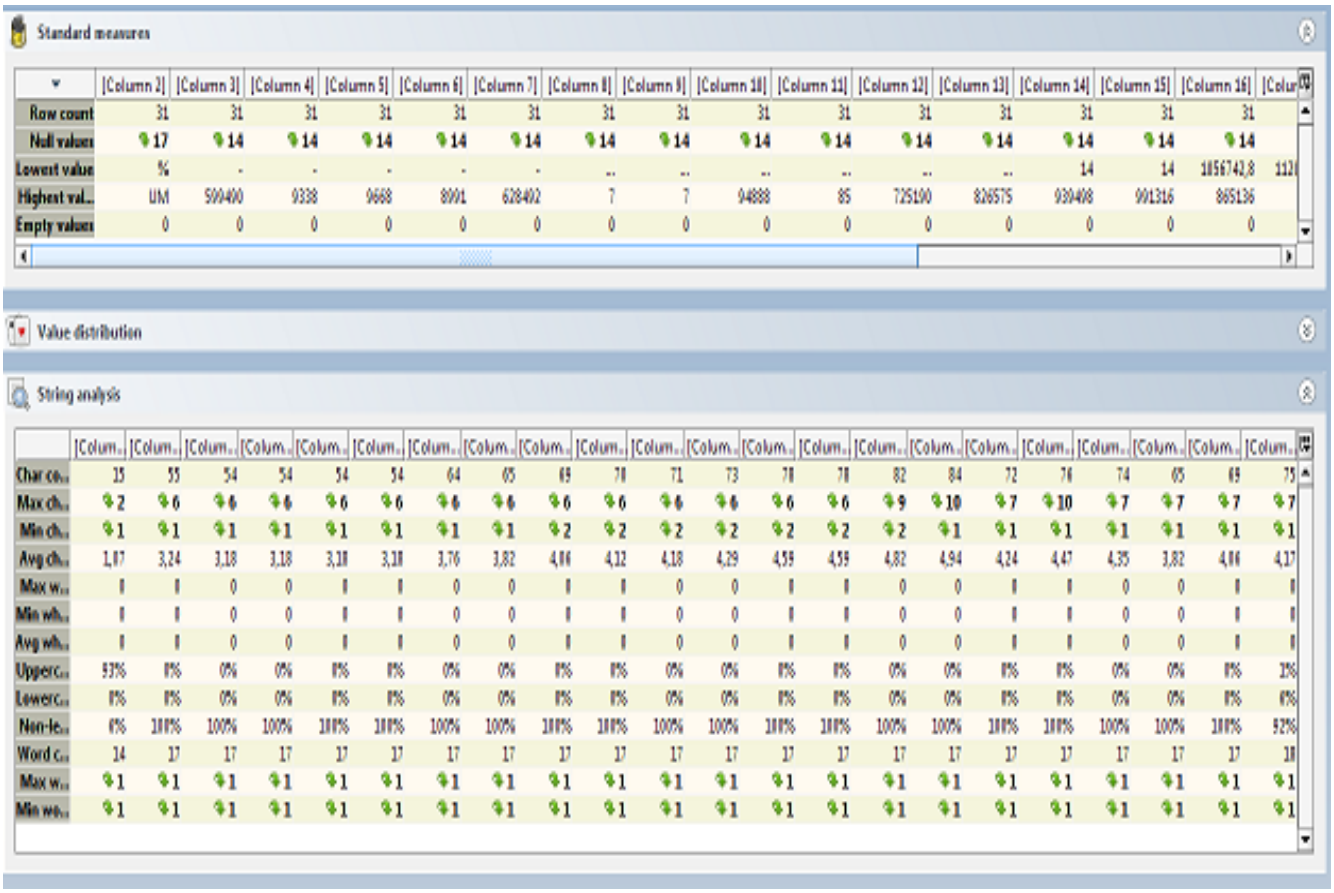

Figura 14: Perfilado para Indicadores de Infraestructura.

3.4.2 Diseño general de las transformaciones

Una vez efectuada la extracción de los datos, se realizan las validaciones necesarias teniendo en cuenta las dos reglas del negocio identificadas y en caso de encontrarse valores incorrectos, el flujo se desvía hacia un fichero donde es almacenado con la correspondiente descripción del error. Si se comprueba que los datos poseen la calidad requerida, se les realiza un conjunto de transformaciones y se procede a su inserción en el mercado de datos, en la figura 15 se muestra el diseño general de una transformación realizada al mercado de datos que es objeto de estudio.

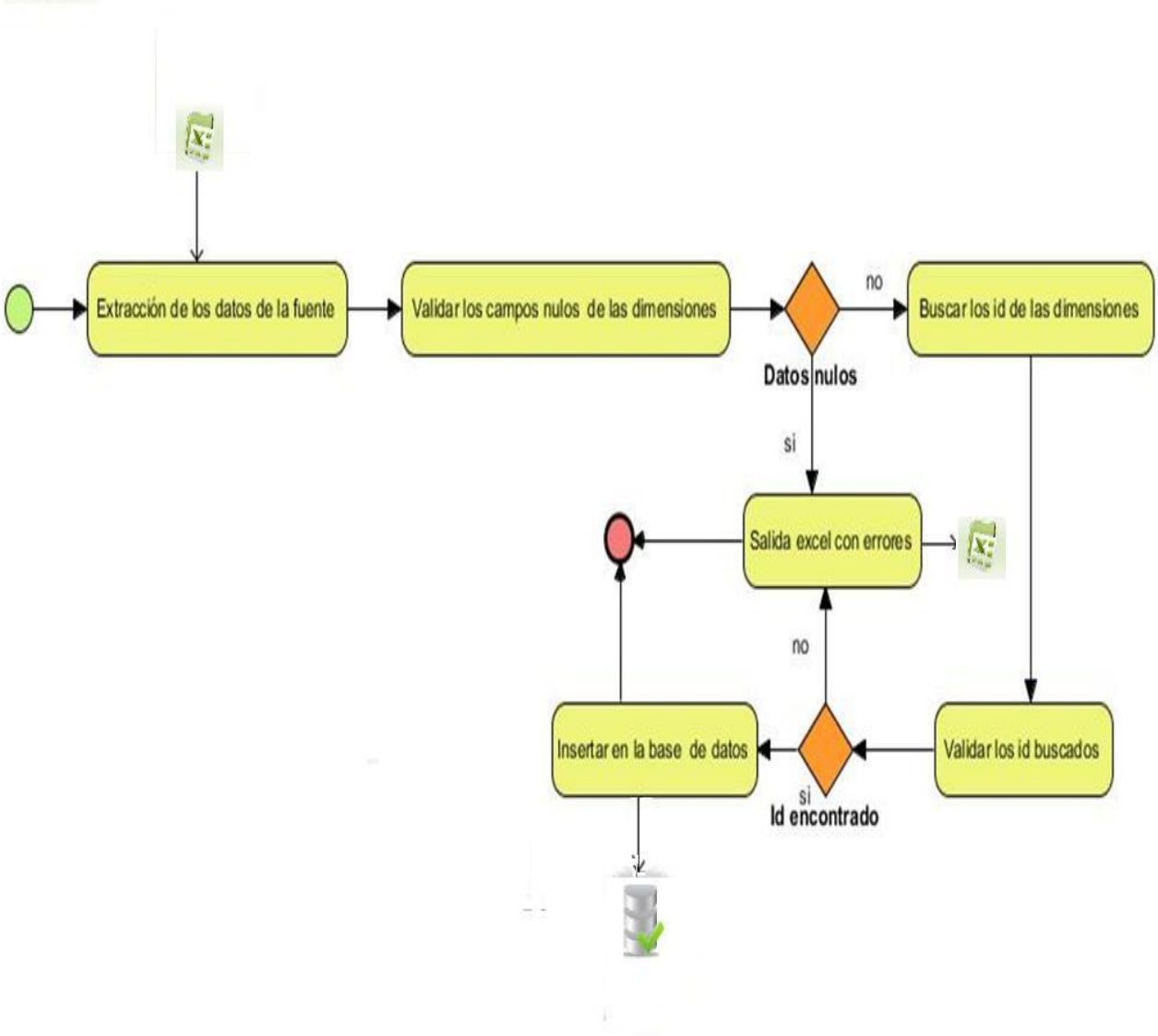

Figura 15: Diseño general de las transformaciones.

3.4.3 Extracción, transformación y carga de los datos

La información almacenada en los sistemas fuente (“Excel”) es extraída para realizarle un proceso de transformación y limpieza, ajustándola a las necesidades de información del cliente. En la presente investigación se realizaron 14 transformaciones, de las cuales ocho pertenecen a las transformaciones de los hechos y seis a las transformaciones realizadas a las dimensiones.

Transformaciones para las dimensiones

Figura 16: Transformación de la dimensión indicador_tic.

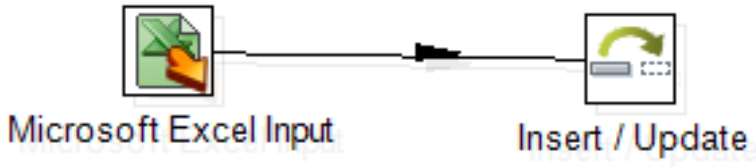

Figura 17: Transformación de la dimensión dim_criterio_entidad.

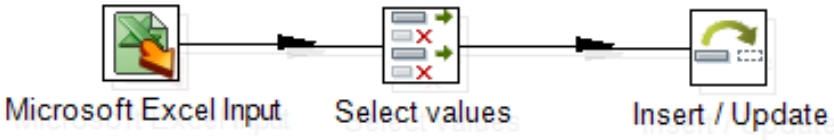

Figura 18: Transformación de la dimensión dim_sector.

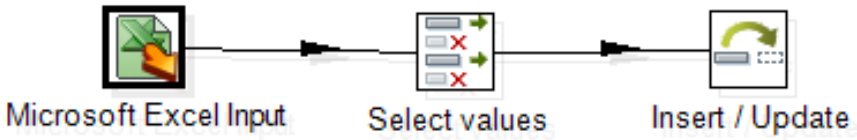

Figura 19: Transformación de la dimensión dim_ tipo_ingreso.

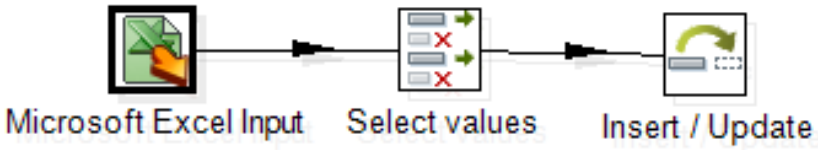

Figura 20: Transformación de la dimensión dim_inversion.

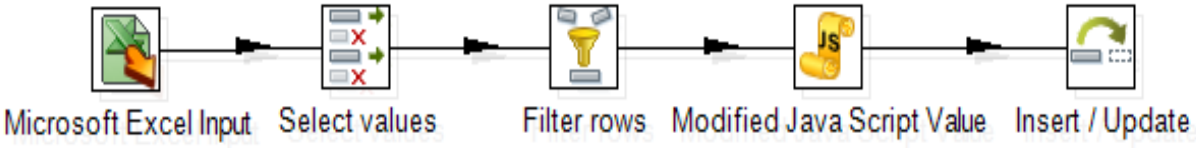

Figura 21: Transformación de la dimensión dim_trafico.

Transformaciones para los hechos

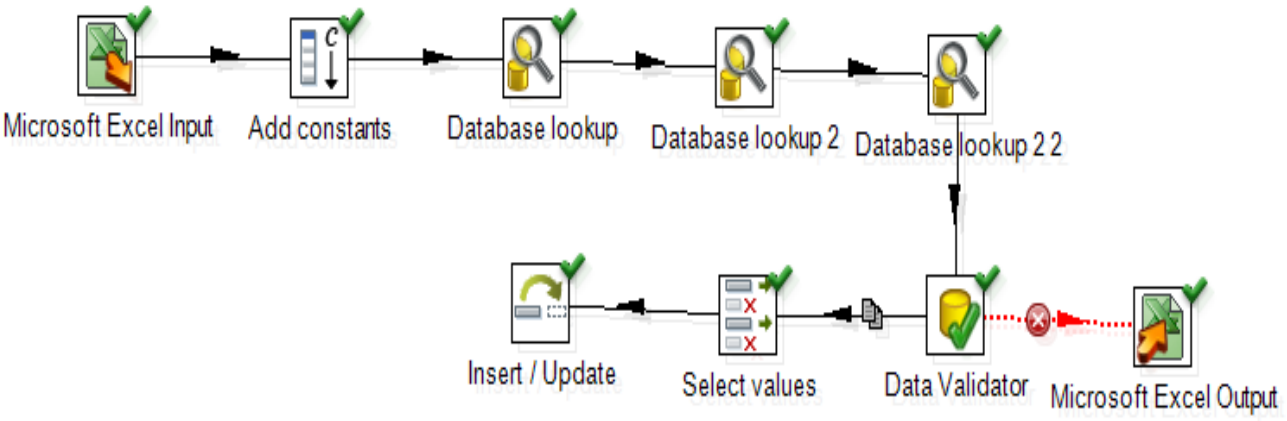

Figura 22: Carga del hecho hech_criterio_entidad.

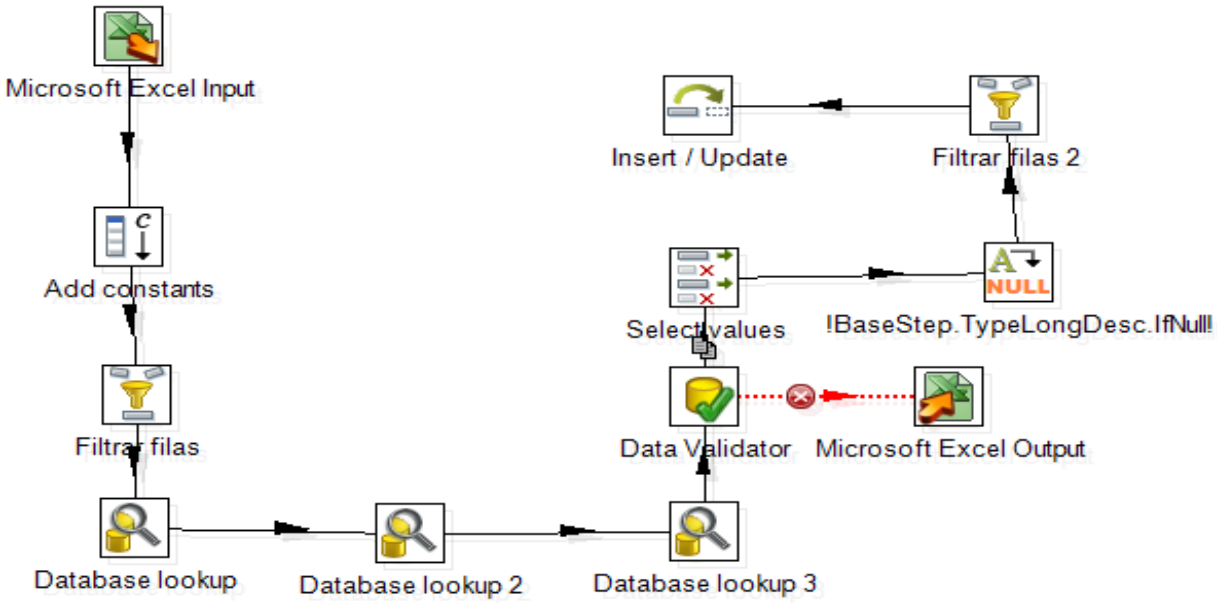

Figura 23: Carga del hecho hech_inversiones.

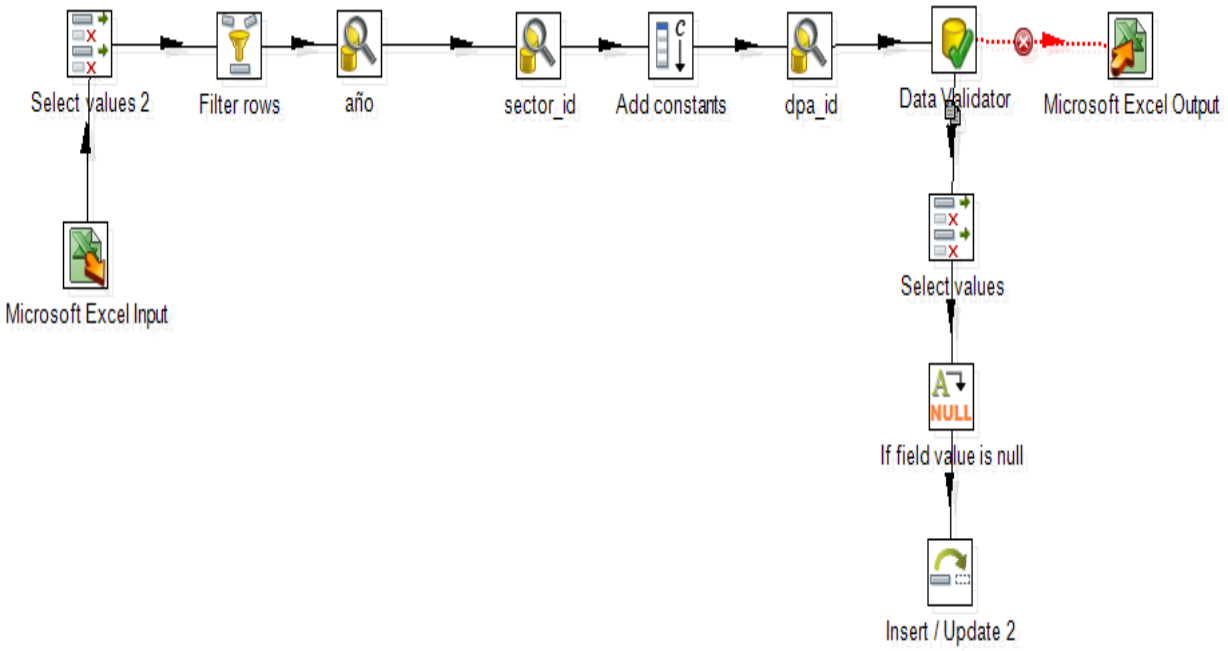

Figura 24: Carga del hecho hech_sector_poblacion.

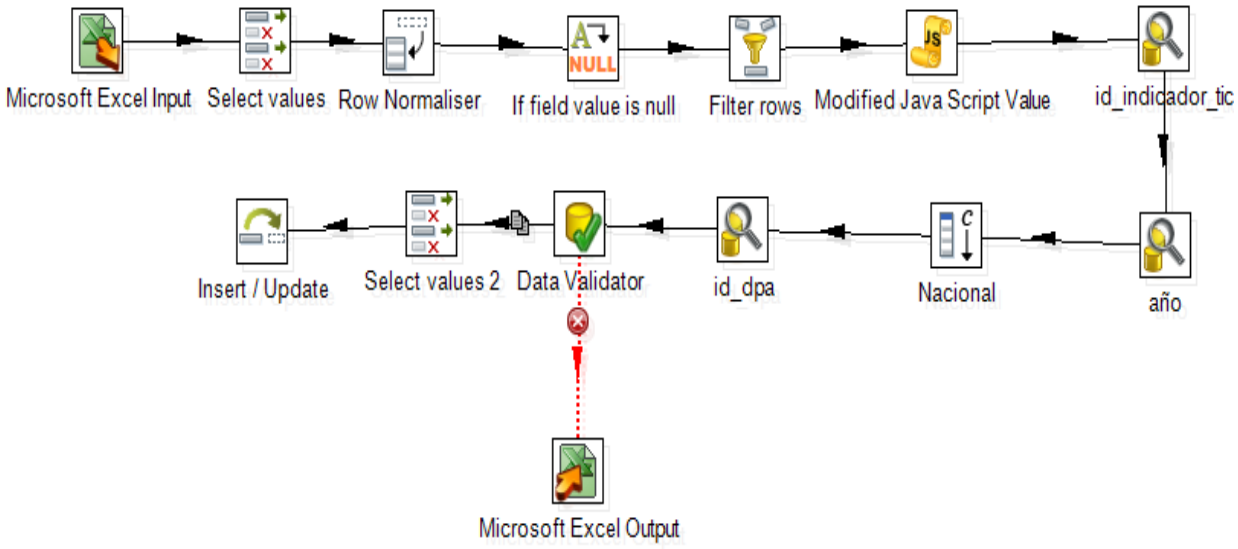

Figura 25: Carga del hecho hech_general_tic

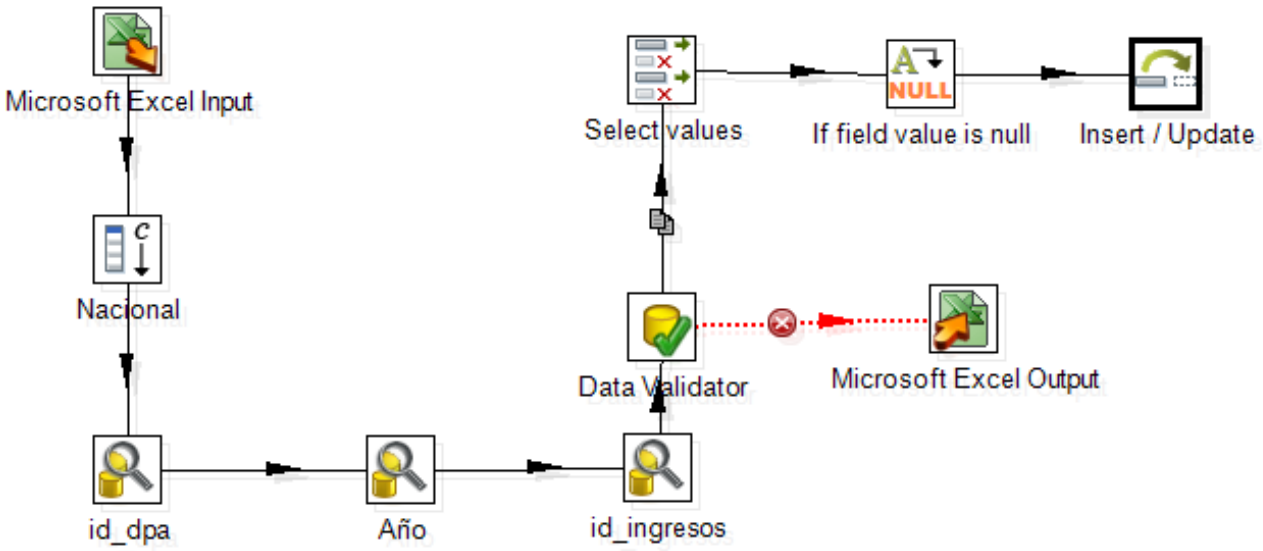

Figura 26: Carga del hecho hech_ingresos.

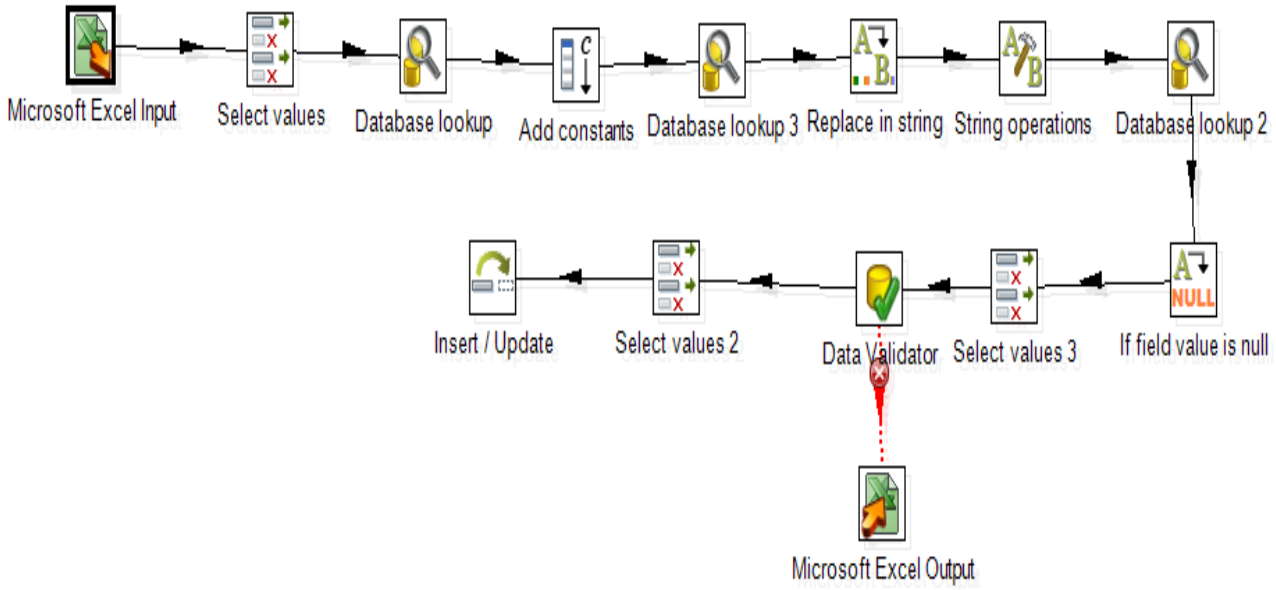

Figura 27: Carga del hecho hech_educación.

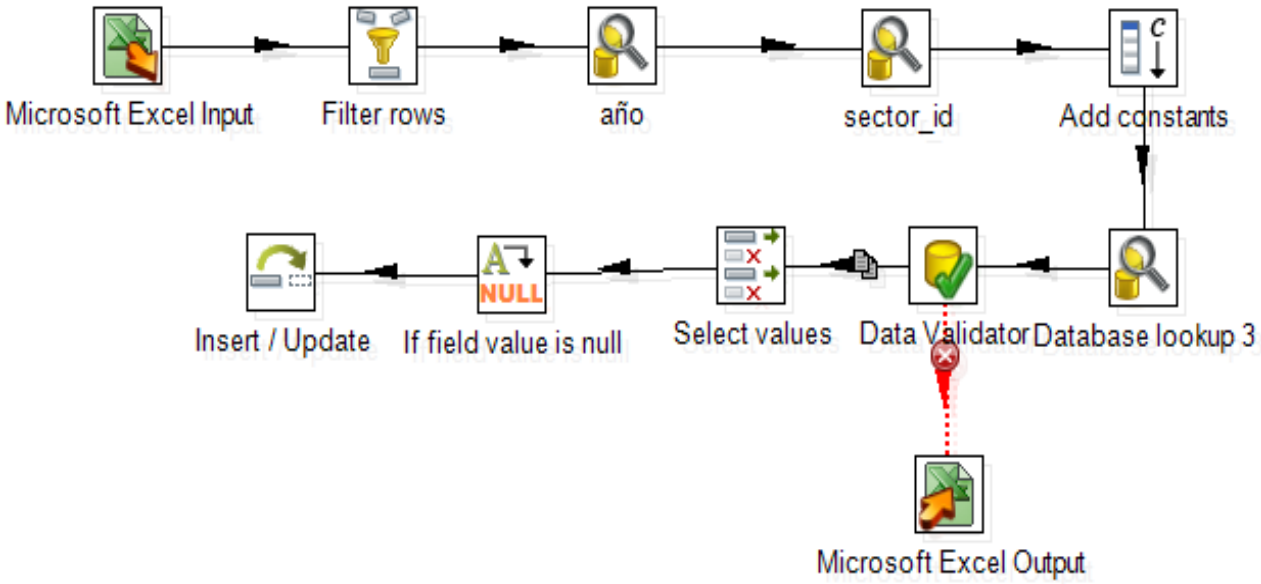

Figura 28: Carga del hecho hech_salud.

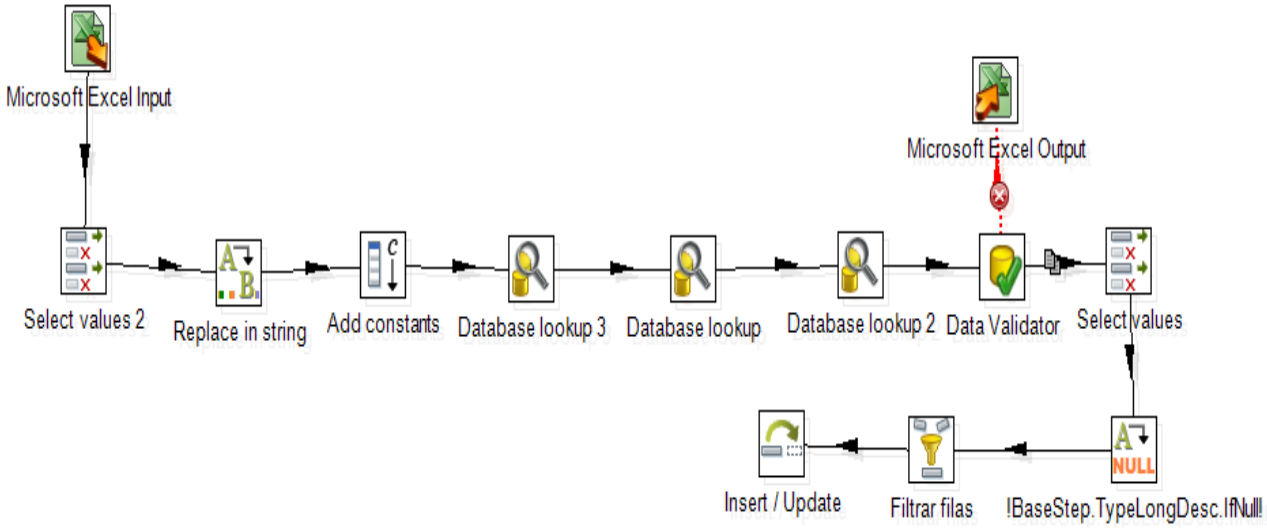

Figura 29: Carga del hecho hech_trafico.

3.5 Trabajo para organizar el orden de la carga

Luego de haber realizado las transformaciones a los datos, es necesario organizar el orden de las cargas de las tablas. El trabajo o job define el orden en que se van a ejecutar las transformaciones, el horario y frecuencia de las cargas.

En este proceso el trabajo se organizó de manera que se ejecutaran cada una de las dimensiones y luego los hechos como se muestra en la figura 30.

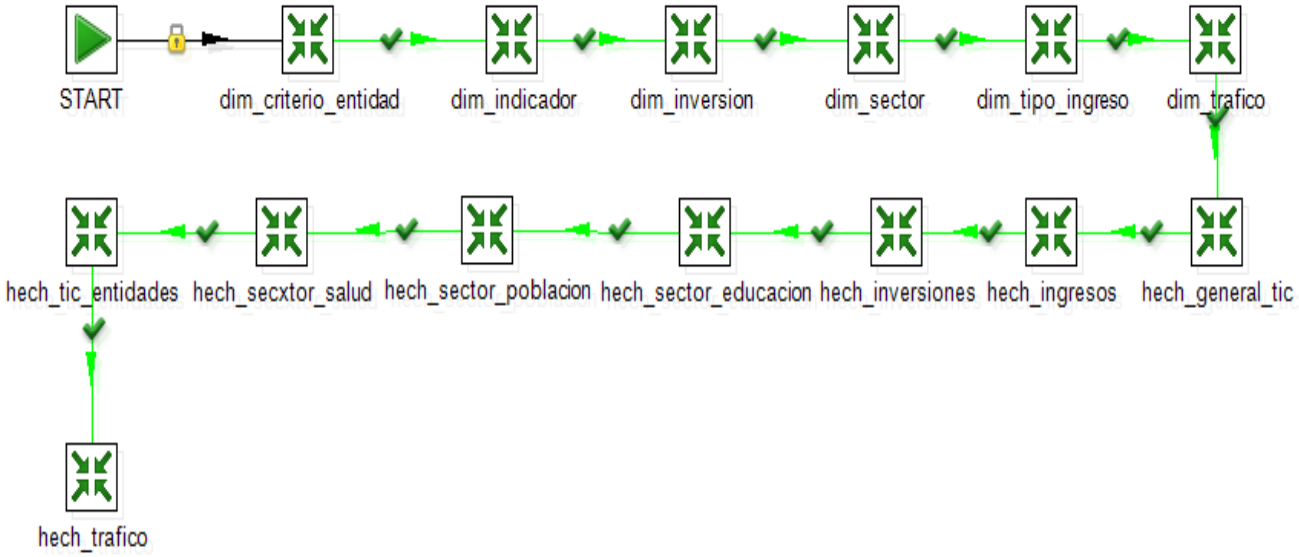

Figura 30: Trabajo de las cargas de las dimensiones y hechos

3.6 Proceso de Inteligencia de Negocios

La Inteligencia de Negocio consiste en transformar los datos en información, y la información en conocimiento, dependiendo en gran medida en la forma que quede organizada la información estructural, de forma que se pueda optimizar el proceso de toma de decisiones en las instituciones.

3.6.1 Diseño del subsistema de visualización de datos

El diseño del subsistema de visualización viene seguido de la implementación del subsistema de visualización de datos. A continuación se detallan en la figura 31 los elementos que componen las estructuras de navegación de la información presentada en la capa de visualización:

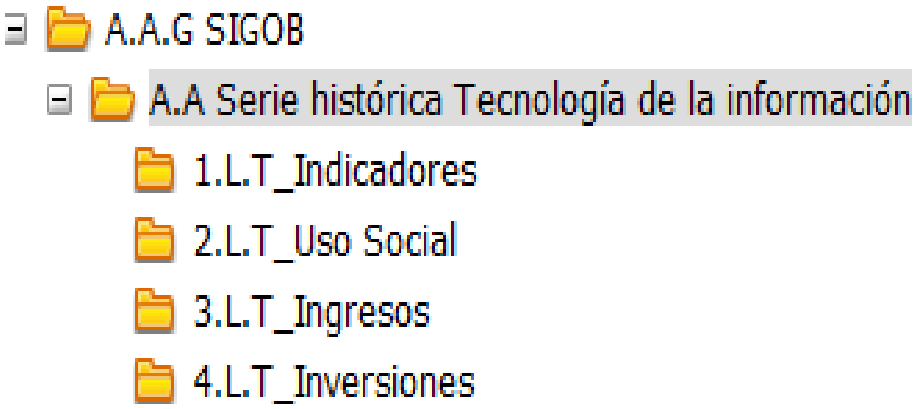

Figura 31: Arquitectura de la información.

3.6.2 Implementación del subsistema de visualización de datos

En correspondencia con los requisitos de información identificados para el área tecnología de la información, se definieron 10 reportes agrupados en ocho Libros de Trabajo (LT) ubicados dentro del Área de Análisis (AA) Series históricas Tecnologías de la información, abarcando información estadística de las diferentes aristas tales como, la educación, la salud, la población, las entidades, los ingresos e inversiones de las TICs a nivel nacional. Dicha área se corresponde con una sección del almacén de datos del proyecto SIGOB, mientras que los LT representan las diferentes categorías a las que pueden pertenecer los reportes, conteniendo cada LT diferentes indicadores a medir, por ejemplo, el total de tráfico internacional de telefonía, porciento de entidades que poseen computadoras, el total de centrales telefónicas, entre otros aspectos que necesitan ser analizados por los especialistas de la información, ver figura 32.

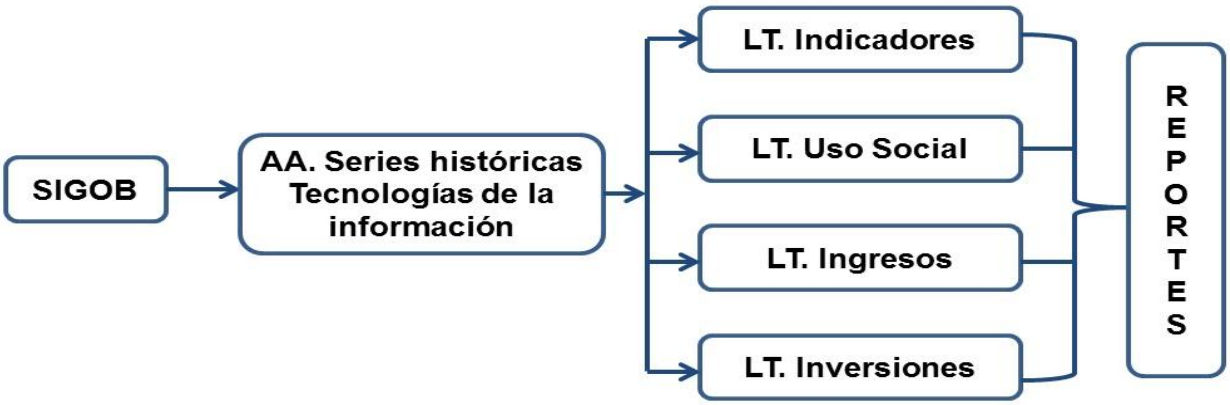

Figura 32: Mapa de navegación de la solución.

3.6.3 Diseño de los cubos OLAP

Uno de los factores claves en el procesamiento analítico en línea son los cubos OLAP, estos proveen rápido acceso a los datos almacenados independientemente de la cantidad de datos en el cubo, al mismo tiempo son un subconjunto de datos del almacén de datos. Luego de haber realizado los procesos de ETL se realizó la creación de los cubos correspondientes al mercado de datos. En la presente investigación se diseñaron ocho cubos multidimensionales como se muestra en la tabla cinco, uno para cada tabla de hecho, especificando en los cubos las dimensiones y características que se corresponden con estas tablas, además el esquema cuenta también con nueve de dimensiones y 29 medidas, todas ellas enfocadas a las necesidades del cliente para darle solución al problema planteado.

Cubos	Hechos correspondientes
Inversiones	hech_inversiones_tic
Sector_poblacion	hech_sector_poblacion
Sector_salud	hech_sector_salud
Sector_educacion	hech_sector_educacion
Servicio_telefonia	hech_servicio_telefonia
Entidades	hech_tic_entidades
General_tic	hech_general_tic
Ingresos	hech_ingresos_ti

Tabla 5: Cubos y hechos correspondientes.

A continuación se muestran algunos ejemplos de diseño de los cubos en las figuras 33 y 34.

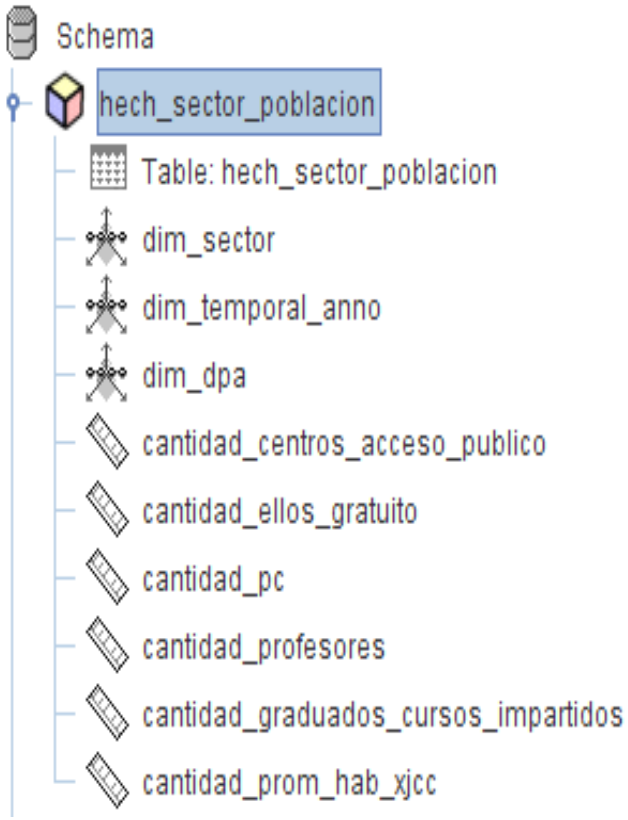

Figura 33: Cubo OLAP correspondiente al hecho hech_sector_poblacion.

Figura 34: Cubo OLAP correspondiente al hecho hech_inversiones_tic.

Después de diseñado cada uno de los cubos, se procede a la creación de los reportes, a continuación se listan todos los implementados en la solución, seguido de la visualización en la aplicación de algunos de ellos:

- ✓ Indicadores de infraestructura.
- ✓ Indicadores físicos del servicio internacional de telefonía.

- ✓ Indicadores de los servicios de correo y telégrafo.
- ✓ Indicadores físicos de las TIC.
- ✓ Indicadores TIC en Entidades.
- ✓ Uso social, sector población.
- ✓ Uso social, sector educación.
- ✓ Uso social, sector salud.
- ✓ Ingresos por Comercio de TIC.
- ✓ Inversiones en las TIC.

	valores						
	valor indicador						
	Años						
Indicadores	2004	2005	2006	2007	2008	2009	2010
Indicadores de infraestructura	3.654.672,38	3.965.709,6	4.397.321,66	3.259.504,4	3.493.002,2	3.917.834,59	4.612.150,01
Centrales telefónicas	502	505	505	506	506	549	617
Teléfonos instalados de todo tipo	1.120.147,58	1.167.441	1.354.231,56	13.777,3	50.530	53.414	54.087
Estaciones públicas totales instaladas	28.605	34.571	40.358	44.126			
Total de centrales telegráficas	8	4	4	4	4	3	1
Líneas telegráficas telex en servicio	1.673	486	486	486	486	850	607
Líneas telefónicas instaladas nacionalmente	897.122	934.999	982.801	999.490	1.033.565	1.076.587	1.147.358
Cantidad de líneas digitales existentes	765.387	839.590	905.516	947.639	987.978	1.045.377	1.117.907
Por ciento de digitalización nacional	85,3	89,8	92,2	94,9	95,6	97,1	98,71
Líneas telefónicas en servicio nacionalmente	841.135	988.015	1.113.318	1.253.370	1.419.825	1.740.942	2.291.455
Densidad telefónica por 1000 habitantes	7,5	8,8	9,9	11,2	12,6	15,49	19,3

Slicer:

Figura 35: Visualización del reporte Indicadores de Infraestructura.

	valores								
	valor indicador								
	Años								
Indicadores	2002	2003	2004	2005	2006	2007	2008	2009	2010
Indicadores físicos de las TICs	752,199	2.067,156	2.695,92	3.157,942	3.444,819	3.690,9	4.838,8	5.533,7	5.041,4
Cantidad de computadoras existentes	250	270	300	377	430	509	630	700	724
Cantidad de usuarios de servicios de Internet	420	584,967	940	1.090	1.250	1.310	1.450	1.600	1.790
Cantidad de sitios de Internet	0,1	0,7	1,5	2,5	2,9	3,4			
Computadoras Personales por 1 000 habitantes	22	24	27	34	38	45	56	62	64,4
Usuarios de Internet por 1 000 habitantes	37,499	52,089	83,62	96,942	111,219	117	129	142	159
Dominios registrados bajo .cu		1.100	1.209	1.351	1.389	1.431	2.168	2.331	2.225
Abonados a teléfonos móviles celulares	22,6	35,4	73,8	135,5	152,7	198,3	330	621,2	1
Cobertura de la Población de celular móvil			61	71	71	77,2	75,8	77,5	78

Slicer:

Figura 36: Visualización del reporte Indicadores físicos de las TICs.

	valores
	Porciento
	Años
Concepto	2010
Poseen computadoras	95,12
Presencia en la web	30,23
Poseen intranet	55,75
Cuentan con red de área local (LAN)	72,48
Poseen red externa	25,86

Slicer:

Figura 37: Visualización del reporte Indicadores TIC en Entidades.

3.7 Conclusiones parciales del capítulo

En este capítulo se abordaron los elementos que permitieron el desarrollo de la implementación del mercado de datos. Se realizó la implementación del modelo de datos físico definiéndose dos esquemas: dimensiones y mart_tic_series para lograr una mejor estructura de la información. Se realizó el proceso de ETL, extrayendo los datos de las series históricas de Tecnologías de la Información para luego realizar la carga de los datos a la base de datos. Luego de tener los datos cargados se transforman en información valiosa para el cliente a través de la implementación de los reportes y la visualización de los datos logrando una mejor perspectiva de análisis.

Capítulo 4: Pruebas

4.1 Introducción

En el capítulo se exponen las pruebas realizadas al mercado de datos, así como los resultados obtenidos en cada una de ellas luego de su aplicación. Dichas pruebas son realizadas para garantizar el cumplimiento de las exigencias del cliente y la calidad del producto final.

4.2 Diseño de casos de prueba

Para los ingenieros del software desarrollar un producto o servicio de buena calidad y aceptación por el cliente, constituye un requisito indispensable para que el propio ingeniero pueda ganar determinada reputación dentro de la entidad donde se desempeña como profesional y también la empresa u organización a la que pertenecen.

A nivel mundial se trata de lograr que las empresas que producen software, lo hagan de acuerdo a criterios comunes, para lograr uniformidad, muchos son los esfuerzos que se realizan para lograr que los productos de software tengan una alta calidad al salir al mercado y por lo tanto que la satisfacción del cliente sea elevada.

Las pruebas son una medida de la calidad del software. Estas verifican el desarrollo que va alcanzando el producto durante todas sus etapas, identificando posibles fallos de implementación, calidad o usabilidad de un programa. Para determinar el nivel de calidad se deben efectuar pruebas que permitan comprobar el grado de cumplimiento de las especificaciones iniciales del sistema. (11)

Para lograr obtener un producto con calidad que cumpliera con los requerimientos establecidos y la satisfacción del cliente, se realizaron las pruebas de software de acuerdo al Modelo V definido en el centro DATEC para evaluar y determinar la calidad del producto.

4.3 Pruebas de software

Para el desarrollo de cualquier producto de software se realizan diferentes actividades desde que surge la idea inicial hasta la obtención del producto final. Para establecer un orden de ejecución de estas actividades se utilizó el modelo en V (13), este modelo considera las actividades de prueba como un proceso que corre en paralelo con las actividades de análisis y diseño, en lugar de establecer una fase independiente al final del proyecto, el cual es empleado por el centro DATEC para garantizar la calidad del producto final. En la figura 38, se aprecia una representación gráfica del ciclo de vida del software propuesta en el Modelo en V, donde a la izquierda se muestran las diferentes etapas de desarrollo, mientras que a la derecha se muestran las pruebas correspondientes a cada una de ellas.
Para la validación del MD Tecnologías de la información se puede aplicar diferentes tipos de pruebas, a continuación se muestran algunas de las pruebas que pueden ser utilizadas para la validación de un

software. (12).

Para la presente investigación se realizarán las siguientes pruebas:

Prueba unitaria: es el proceso de probar los componentes individuales de la solución. El propósito es identificar diferencias entre la especificación de los artefactos y el comportamiento real de cada módulo.

Prueba de integración: es el proceso en el cual los componentes son agregados para crear componentes más grandes. Es la prueba realizada para mostrar que aunque los componentes hayan pasado satisfactoriamente las pruebas de unidad, la integración de los componentes es incorrecta.

Prueba de sistema: se refiere al comportamiento del sistema integrado. Durante la etapa de las pruebas unitarias y de integración deben haberse identificado la mayoría de las no conformidades. La prueba de sistema se aplica generalmente para probar los requerimientos no funcionales de la solución.

Pruebas de aceptación: se realizan para probar que el sistema cumpla con los requerimientos especificados por el cliente.

Además de realizará la prueba de calidad del dato para conocer si el producto satisface las necesidades del cliente.

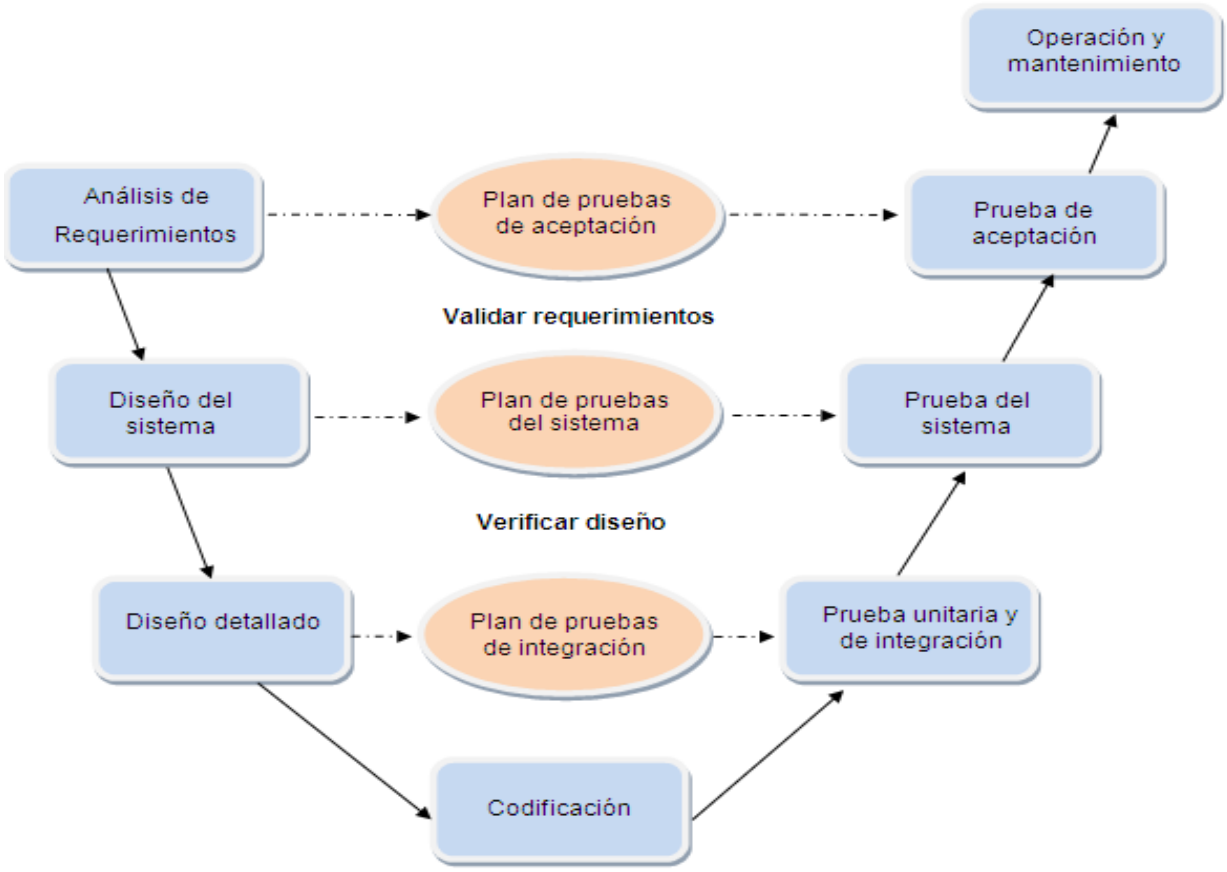

Figura 38: Modelo V

4.4 Listas de chequeo

Las listas de chequeo constituyen un mecanismo para el control de los riesgos y su función básica es la de detectar condiciones peligrosas que puedan generar incidentes al producto de software.

Para elaborar la lista de chequeo aplicadas al mercado de datos Tecnologías de la información, se tuvieron en cuenta elementos de evaluación que son importantes una vez realizado el proceso de ETL y BI, permitiendo identificar las deficiencias que posean los artefactos generados de estos procesos. La lista de chequeo contiene diferentes indicadores a evaluar los cuales se encuentran distribuidos en tres secciones fundamentales:

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

4.5 Calidad del dato

La prueba de Calidad del dato posibilita conocer si los datos cargados y visualizados satisfacen las necesidades del cliente. Con el objetivo de conocer la calidad de los datos almacenados en el MD se realizó el perfilado de los mismos a través de la herramienta Data Cleaner. Una vez analizado el estado de estos se definieron la cantidad de valores únicos, duplicados, distintos, mínimos y máximos de cada medida de las tablas de hechos. La especialista del área revisó la serie con los datos cargados en la que se probaron diez “Excel”, de ellos fueron cargados diez correctamente, aunque se detectó un dato inconsistente desde la fuente original, que se decidió mantener así y se validó de esta forma el MD.

4.6. Resultados de las pruebas

Al mercado de datos Tecnologías de la información se le aplicaron dos casos de pruebas basados en casos de uso para las pruebas del sistema, un caso de prueba y cuatro listas de chequeo para los proceso de extracción, transformación y carga. Además de aplicarse las pruebas unitarias por cada uno de los subsistemas del mercado, las de integración por parte de los especialistas del departamento y finalmente las pruebas de aceptación del cliente.

Pruebas unitarias y de integración

Una vez culminada la implementación se realizaron pruebas unitarias a los flujos de integración de datos y a los diferentes componentes relacionados con la capa de visualización, siendo detectadas las siguientes no conformidades: en ETL, exceso de selecciona/renombra, los nombres de los

componentes deben ser más sugerentes, error en el diseño de la transformación hech_generales, de estas no conformidades solamente la tercera presentó complejidad alta. Cada una de ellas se resolvió satisfactoriamente.

Las no conformidades detectadas en BI fueron las siguientes: creación del fichero ¨.Properties ¨ en cada una de los LT y organización de los indicadores por orden de aparición en la fuente, siendo todas de complejidad baja y resueltas positivamente.

Pruebas de sistema

Se realizaron pruebas por parte del grupo de especialistas de calidad del centro DATEC, donde fueron identificadas tres no conformidades: uso excesivo de mayúsculas en la aplicación, un error ortográfico y el orden de los reportes no se corresponden con los casos de prueba, siendo todas de complejidades bajas y resueltas positivamente.

Pruebas de aceptación

Una vez finalizado el desarrollo del mercado de datos, el cliente probó el sistema para verificar el cumplimiento de los requerimientos definidos, que los valores cargados en el almacén se corresponden con la fuente de datos y las funcionalidades de la aplicación. Estas pruebas demostraron que el producto satisface realmente las necesidades de los especialistas de la ONEI, ya que el personal que las ejecuta forma parte del equipo de trabajo que utilizará el mercado de datos en dicho centro. La especialista Elena Leonila Fernández García, quien es la representante de la ONEI en la UCI emitió la carta de aceptación, la cual valida que el mercado de datos Tecnologías de la información se encuentra listo para ser desplegado y utilizado en la ONEI.

4.7 Conclusiones parciales del capítulo

En este capítulo se realizó la validación y pruebas al mercado de datos, aplicándose los casos de prueba correspondientes a los casos de uso de información del sistema y se explicó brevemente la estrategia seguida para la validación del mercado de datos tomando como base la propuesta del modelo en V.

Después de realizadas estas pruebas el Mercado de datos Tecnologías de la información fue validado por CALISOFT certificando de esta manera que está listo para su uso, posteriormente se realizaron las pruebas de aceptación del cliente, el cual entregó la carta que evidencia que el sistema está listo.

Conclusiones generales

Al concluir el presente trabajo de diploma se arriba a las siguientes conclusiones:

- La fundamentación de la metodología, herramientas y tecnologías utilizadas en la investigación, permitió organizar de manera estructurada el proceso de desarrollo del mercado de datos así como un correcto diseño e implementación de la solución presentada.
- El análisis y diseño realizado posibilitó la identificación de las necesidades del cliente. De igual manera, contribuyó a la obtención del diagrama de casos de uso del sistema, la matriz bus, así como el modelo lógico y físico, logrando una mejor comprensión del proceso de negocio del área Tecnología de la información.
- La implementación de los subsistemas que componen la solución contribuyó a la creación de dos esquemas en el gestor de bases de datos PostgreSQL que organizan nueve tablas dimensiones y ocho tablas de hechos, lo que admitió separar las tablas de dimensiones comunes de las específicas del Mercado de Datos. Se implementaron un total de 14 transformaciones que permitieron la carga de los datos hacia el Mercado de Datos Tecnología de la información. La implementación de los ocho cubos OLAP y los diez reportes garantizan la disponibilidad de la información de manera organizada facilitando el proceso de toma de decisiones.
- Las pruebas unitarias, integración, sistema, liberación y aceptación permitieron validar el mercado de datos Tecnologías de la información garantizando que cumple con las necesidades del cliente.

Recomendaciones

- Se recomienda la integración al Sistema de Información de Gobierno y su despliegue en la ONEI.
- Implementar una estrategia para realizar la carga de los ficheros de error obtenidos como resultado de las transformaciones, luego de ser analizados y corregidos por los especialistas del área de tecnología de la ONEI.
- Gestionar los metadatos que registren información acerca de cada uno de los procesos de integración, permitiendo de esta forma facilitar al equipo de ETL, depurar los errores que puedan ocurrir de una forma más acertada, así como la generación de informes detallados sobre el estado de los datos.

Referencias bibliográficas

1. ARAQUE CUENCA, FRANCISCO.*Definicion del Modelo y Esquemas del Almacén de Datos en función de las características temporales de los sistemas operacionales componentes.*. [En línea]. Granada: Universidad de Granada, Noviembre 2005. [Citado el: 20 octubre de 2011]Tesis Doctoral.
   Disponible en Web: <http://digibug.ugr.es/bitstream/10481/844/1/15837087.pdf>.
2. CURTO, JOSEP.*DW:Definiciones de Inmon y Kimball.*[En línea].Information Management.Reflexiones sobre las tecnologías de la información,28 de noviembre del 2006..[Citado el: 21 de octubre de 2011].Blog de WordPress.com. Disponible en Web: <http://informationmanagement.wordpress.com/2006/11/28/dw-definiciones-de-inmon-y-kimball/>.
3. KIMBALL, RALPHyROSS, MARGY.*The Data Warehouse Toolkit:The complete Guide to Dimensional Modeling. 2nd edition.* Hoboken (N.J.): Wiley Publishing Inc., 2002. 464 p.ISBN 0-471-20024-6
4. RFV. *Data WareHouse.* [En línea]. Renziton, 8 de septiembre de 2008. [Citado el: 3 de diciembre de 2011].Disponible en Web: http://renziton.blogspot.com/2008/09/data-warehouse.html.
5. *Business Intelligence.Datamart.* [En línea].A Coruña (Galicia): Sinnexus,2007. [Citado el: 20 de octubre de 2011].
   Disponible en Web: <http://www.sinnexus.com/business_intelligence/datamart.aspx>.
6. SANZ RODRÍGUEZ, MIGUEL.*Análisis y Diseño de un Data Mart para el seguimiento académico de alumnos en un entorno universitario.*[En línea]. Madrid: Universidad Carlos III, 22 de julio de 2010.[Citado el: 3 de diciembre de 2011]. Proyecto Fin de Carrera.Disponible en Web:
   http://e-archivo.uc3m.es/bitstream/10016/9856/6/PFC_Miguel_Rodriguez_Sanz.pdf.
7. TAPIA FUENTES, LUIS yVALDIVIA PINTO, RICARDO.*Incorporación de elementos de inteligencia de negocios en el proceso de admisión y matrícula de una universidad chilena.* [En línea]. Arica (Chile): Universidad de Tarapacá, 18 de noviembre de 2010. Ingeniare. Revista chilena de ingeniería, vol. 18 Nº 3, 2010, pp. 383-394.[Citado el: 5 de diciembre de 2011]. Disponible en Web: www.scielo.cl/scielo.php?pid=S0718-33052010000300012...sci_arttext.
   ISSN 0718-3305.

8. BOUMAN, ROLAND y VAN DONGEN, JOS. *Pentaho Solutions, Business Intelligence and Data Warehousing with Pentaho and MySQL.* Hoboken (N.J.): Wiley Publishing Inc., 2009. 648 p. ISBN: 978-0-470-48432-6

9. *Conveniencia del "Business Intelligence".* [En línea]. Buenos Aires (Argentina): Técnicas Cuantitativas Srl., 2007. [Citado el: 21 de octubre de 2011].
Disponible en Web: http://www.tecnicas.com/business-intelligence/conveniencia-de-business-intelligence.aspx.

10. KIMBALL, RALPH y CASERTA, JOE.*The Data Warehouse ETL Toolkit Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data.* Hoboken (N.J.): Wiley Publishing Inc., 2004. 528 p. ISBN: 978-0-7645-6757-5.

11. PRESSMAN, ROGER S.*Ingeniería del software.* s.l.: McGraw-Hill, 2001. 640 p. ISBN: 84-481-3214-9.

12. SCALONE, FERNANDA.*"Estudio comparativo de los modelos y estándares de calidad del software".*[En línea]. Buenos Aires: Universidad tecnológica nacional, Facultad regional de Buenos Aires, 2006.[Citado el: 5 de diciembre de 2011].Tesis de Maestría. Disponible en Web: <http*://laboratorios.fi.uba.ar/lsi/scalone-tesis-maestria-ingenieria-en-calidad.pdf>.*

.

Bibliografía

1. ARAQUE CUENCA, FRANCISCO.*Definicion del Modelo y Esquemas del Almacén de Datos en función de las características temporales de los sistemas operacionales componentes*.. [En línea]. Granada: Universidad de Granada, Noviembre 2005. [Citado el: 20 de octubre de 2011]Tesis Doctoral. Disponible en Web: http://digibug.ugr.es/bitstream/10481/844/1/15837087.pdf.

2. *Business Intelligence. Datamart.* [En línea].A Coruña (Galicia): Sinnexus,2007. [Citado el: 20 de octubre de 2011].Disponible en Web: http://www.sinnexus.com/business_intelligence/datamart.aspx.

3. CARAMAZANA CÁRCAMO, ALBERTO. *Tecnologías y Metodologías para la Construcción de Sistemas de Gestión del Conocimiento. "Business Intelligence".* [En línea].Madrid: Universidad Pontificia de Salamanca, Septiembre 2002.[Citado el:21 de octubre de 2011].Tesis de Doctorado. Disponible en Web: http://www.willydev.net/descargas/articulos/general/bi.pdf.

4. *Conveniencia del "Business Intelligence".* [En línea]. Buenos Aires (Argentina): Técnicas Cuantitativas Srl. [Citado el: 21 de octubre de 2011]. Disponible en Web: http://www.tecnicas.com/businessintelligence/convenienciadebusinessintelligence.aspx.

5. CURTO, JOSEP.*DW: Definiciones de Inmon y Kimball.*[En línea].Information Management. Reflexiones sobre las tecnologías de la información, 28 de noviembre del 2006.[Citado el: 21 de octubre de 2011]. Blog de WordPress.com. Disponible en Web: http://informationmanagement.wordpress.com/2006/11/28/dw-definiciones-de-inmon-y-kimball/.

6. *DataCleaner.* [En línea].Holanda:Human Inference, abril 2008. [Citado el: 30 de Noviembre de 2010].Descripción de una aplicación informática. Disponible en Web: http://datacleaner.eobjects.org.

7. IMHOFF, CLAUDIA, GALEMMO, NICHOLAS y GEIGER, JONATHAN G. *Mastering Data Warehouse Design, Relational and Dimensional Techniques.*Hoboken(N.J.): Wiley Publishing, 2003. 456 p. ISBN: 0-471-32421-3.

8. KIMBALL, RALPH y CASERTA, JOE.*The Data Warehouse ETL Toolkit Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data.*Hoboken(N.J.): Wiley Publishing, 2004. 528 p. ISBN: 978-0-7645-6757-5.

9. KIMBALL, RALPH, REEVES, LAURA, ROSS, MARGY y THORNTWHAITE, WARREN.*The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing, and Deploying Data Warehouses.*Hoboken(N.J.): Wiley Publishing, 1998.800 p. ISBN: 0-471-25547-5.

10. KIMBALL, RALPH, ROSS MARGY.*The Data Warehouse Toolkit:The complete Guide to Dimensional Modeling. 2nd edition.* Hoboken (N.J.): Wiley Publishing, 2002.464 p. ISBN 978-0-471-20024-6

11. *Mondrian Documentation.Mondrian Schema Workbench.*[En línea]. Orlando (Florida): Pentaho Corporation,2009. [Citado el: 5 de diciembre de 2011].Disponible en Web: http://mondrian.pentaho.com/documentation/workbench.php.

12. *Pentaho Mondrian Proyect.* [En línea]. Orlando (Florida): Pentaho Corporation, 28 de octubre de 2011. [Citado el: 5 de diciembre de 2011]. Disponible en Web:http://mondrian.pentaho.com.

13. *Perfilado de datos.* [En línea]. La Florida (Barcelona): Dataprix, 2007-2012. [Citado el: 25 de noviembre de 2011]Portal de referencia sobre Tecnologías de la Información. Disponible en Web: http://www.dataprix.com/category/integracion-datos/perfilado-datos.

14. RFV. *Data WareHouse.* [En línea] s.l.: Renziton, 8 de septiembre de 2008.[Citado el:5 diciembre de 2011]Disponible en Web: http://renziton.blogspot.com/2008/09/data-warehouse.html.

15. ROLDÁN, MARÍA CARINA.*Pentaho 3.2 data integration: beginner's guide.* Birmingham (Inglaterra): Packt Publishing, 2010. 492 p. ISBN: 978-1-847-19954-6.

16. SANZ RODRÍGUEZ, MIGUEL.*Análisis y Diseño de un Data Mart para el seguimiento académico de alumnos en un entorno universitario.*[En linea]. Madrid: Universidad Carlos III, 22 de julio de 2010.Proyecto Fin de Carrera. [Citado el: 3 de diciembre de 2011]. Disponible en Web: http://e-archivo.uc3m.es/bitstream/10016/9856/6/PFC_Miguel_Rodriguez_Sanz.pdf

17. TAPIA FUENTES, LUISyVALDIVIA PINTO, RICARDO.*Incorporación de elementos de inteligencia de negocios en el proceso de admisión y matrícula de una universidad chilena.*[En línea]. Arica (Chile): Universidad de Tarapacá, 18 de noviembre de 2010.Ingeniare. Revista chilena de ingeniería, vol. 18 Nº 3, 2010, pp. 383-394.[Citado el: 3 de diciembre de 2011]. Disponible

en Web: www.scielo.cl/scielo.php?pid=S0718-33052010000300012...sci_arttext.
ISSN 0718-3305.

Glosario de términos

ONEI: Oficina Nacional de Estadísticas e Información.

TIC: Tecnologías de la Información y las Comunicaciones.

SIGOB: Sistema de Información de Gobierno.

DATEC: Centro de Tecnologías de Gestión de Datos.

API: una API o Interfaz de Programación de Aplicaciones, es el conjunto de funciones y procedimientos (o métodos si se refiere a programación orientada a objetos) que ofrece cierta biblioteca para ser utilizado por otro software como una capa de abstracción.

JDBC: es el acrónimo de Java Database Connectivity, una API que permite la ejecución de operaciones sobre bases de datos desde el lenguaje de programación Java, independientemente del sistema operativo donde se ejecute o de la base de datos a la cual se accede utilizando el dialecto SQL del modelo de base de datos que se utilice.

ETL: proceso de extracción, transformación y carga.

UML: lenguaje visual para especificar, construir y documentar un sistema de software. Sus siglas vienen dadas por su nombre en inglés Unified Modeling Language.

SQL: lenguaje de consulta estructurado o SQL (por sus siglas en inglés Structured Query Language) es un lenguaje declarativo de acceso a bases de datos relacionales que permite especificar diversos tipos de operaciones en ellas.

Data Warehouse: almacén de datos.

Data Mart: mercado de datos.

BI: Inteligencia del negocio.

Lista de chequeo: instrumento de medición y evaluación que consiste básicamente en un formulario de preguntas referentes al atributo de calidad que se está probando y de las características del documento en el caso de la documentación.

No conformidad: defecto, error o sugerencia que se le hace al equipo de desarrollo una vez encontrada alguna dificultad en lo que se está evaluando.