Universidad de las Ciencias Informáticas

Facultad 6

"Sistema de Información de Gobierno. Mercado de datos Contabilidad, el trabajo y los salarios."

Trabajo de Diploma para optar por el título de

Ingeniero en Ciencias Informáticas

Autores:

Yoendy Martínez Contreras

Daniel Ben Durán

Tutores:

Ing. Yosbel Rodríguez Rodríguez

Ing. Yunier Santana Aldana

Ciudad de La Habana, junio de 2011

"Año 53 de la Revolución"

"Quien no se resuelve a cultivar el hábito de pensar, se pierde el mayor placer de la vida."

Thomas Alva Edison

DECLARACIÓN DE AUTORÍA

Declaramos ser autores del presente trabajo “Sistema de Información de Gobierno. Mercado de datos Contabilidad, el trabajo y los salarios.” y reconocemos a la Universidad de las Ciencias Informáticas (UCI) los derechos patrimoniales de la misma, con carácter exclusivo.

Para que así conste firmamos la presente a los ____ días del mes de_____ del año _____.

Autores: _____ _____

Yoendy Martínez Contreras Daniel Ben Durán

Tutores: _____ _____

Ing. Yosbel Rodríguez Rodríguez Ing. Yunier Santana Aldana

Tutores:

Tutor: Ing. Yunier Santana Aldana

Especialidad de graduación: Ingeniería en Ciencias Informáticas

Categoría docente: Instructor en Adiestramiento

Categoría Científica: Ingeniero

Correo Electrónico: ysaldana@uci.cu

Tutor: Ing. Yosbel Rodríguez Rodríguez

Especialidad de graduación: Ingeniería en Ciencias Informáticas

Categoría docente: Instructor en Adiestramiento

Categoría Científica: Ingeniero

Correo Electrónico: yrdguezro@uci.cu

Agradecimientos

Agradecer primeramente a mis padres por darme todo el cariño y el apoyo necesario para cumplir mis metas, a mi hermano Yojander por aconsejarme en los momentos más difíciles, a mi tía Mercedes mi segunda madre, por todo su apoyo y comprensión, a mis primos y primas por todo su cariño.

A todos mis amigos que me han estado presentes en el transcurso de estos cinco años, a Ariel, Orelmis, Leonel, Fabián, Liniuska por ayudarme a superar todos los obstáculos para cumplir mi meta.

Al colectivo de profesores y estudiantes del departamento, Laritza, Susel, Patricia, Yulier, David, Aylenis, Yelena, Yurislaine que me ayudaron en el desarrollo de la tesis, a mis tutores por su guía durante todo el proceso investigativo. Agradecer también a todos los profesores que contribuyeron en mi formación como profesional durante estos cinco años.

Yoendy Martínez Contreras

En primer lugar les agradezco a mis padres por el apoyo que me brindaron en todos estos años de estudio y que fueron mi principal inspiración para convertirme en un profesional. También a mi hermano que siempre me apoyó y me sirvió de ejemplo, en general a toda mi familia que siempre confió en mí. Además mencionar a los profesores que en algún momento supieron guiarme por el buen camino a base de regaños y consejos. Por último y no por eso menos importante a mis amigos que estuvieran en las buenas y en las malas, me aconsejaron y me ayudaron en los momentos difíciles.

Daniel Ben Durán

Dedicatoria

Dedicada a las personas más importantes de mi vida a mis padres y a mi hermano.

Yoendy Martínez Contreras

Dedicada a mis padres que siempre estuvieron pendientes y dieron todo lo que tenían para que yo pudiera ser ingeniero. Además se la quisiera dedicar a mi abuela Blanca que también estuvo pendiente y se preocupó por mis estudios.

Daniel Ben Durán

Resumen

El presente trabajo de diploma surge como parte de la colaboración existente entre la Universidad de las Ciencias Informáticas y la Oficina Nacional de Estadísticas. Esta última se encarga de recopilar información estadística de todos los sectores de la economía y la sociedad cubana, entre los que se encuentran datos referentes al área de la Contabilidad, el trabajo y los salarios. Las herramientas utilizadas para la recolección y gestión de la información en la Oficina Nacional de Estadísticas dificulta el análisis de los datos almacenados en la entidad. El presente trabajo de diploma tiene como objetivo desarrollar un Mercado de Datos para el área de Contabilidad, el trabajo y los salarios del Sistema de información de Gobierno que apoye a los especialistas en el proceso de toma de decisiones. Para el desarrollo de la solución se realizó una caracterización de las metodologías, herramientas y tecnologías a utilizar en el desarrollo de los Almacenes de Datos. Posteriormente se refinó el análisis y diseñó del Mercado de Datos que dio paso a la implementación, obteniéndose como resultado un Mercado de Datos poblado y funcional, que cumple con los requerimientos planteados por el cliente. Para validar la solución se realizaron pruebas mediante casos de pruebas y listas de chequeo.

Palabras clave:

Mercado de Datos (MD), Almacén de Datos (AD), toma de decisiones, Inteligencia de Negocio (BI).

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

INTRODUCCIÓN

En la actualidad, manejar y analizar correctamente la información, es una necesidad primordial para la sociedad. Los avances tecnológicos ocurridos durante las últimas décadas, han facilitado la manipulación y almacenamiento de modo eficiente de grandes volúmenes de datos. Sin embargo, las crecientes necesidades de los usuarios, en su afán de obtener información cada vez más confiable a partir de los datos acumulados, evidenciaron la incapacidad de realizar en los entornos transaccionales, procesos analíticos de gran envergadura. Esto provocó que surgiera una división en la línea del manejo de la información: por una parte quedaban los ambientes transaccionales, encargados de la entrada de datos; por otra, los destinados a análisis, especializados en la obtención y buen aprovechamiento de los mismos. Uno de los resultados más notorios en esta separación ha sido la concepción de una nueva arquitectura, destinada a apoyar los procesos de toma de decisiones: los Almacenes de Datos (AD). Con ellos se pretende concentrar la información para brindar una visión global del comportamiento del negocio.

En Cuba se realizan múltiples esfuerzos por informatizar gran parte de los sectores estatales, por tal motivo se han convocado a un grupo de instituciones del Ministerio de la Informática y las Comunicaciones (MIC) a colaborar en dicho propósito. La Universidad de las Ciencias Informáticas (UCI), ha intervenido de manera directa en el proceso de informatización de los principales renglones de la sociedad cubana como son la educación, la salud, la seguridad social y la cultura. La UCI como pilar fundamental del proceso tecnológico que se desarrolla en la nación, elabora productos de *software* en cada uno de los centros productivos que la conforman. Uno de ellos es el Centro de Tecnologías y Gestión de Datos (DATEC), que se encarga de desarrollar productos y brindar servicios relacionados con bases de datos y el análisis de información. Una de las instituciones que recibe apoyo del centro es la Oficina Nacional de Estadísticas (ONE).

La ONE, órgano rector de la estadística en Cuba, tiene como misión garantizar la obtención de estadísticas de calidad a través del Sistema Estadístico Nacional (SEN). Incluye también dentro de sus misiones ejercer una adecuada dirección, ejecución y control de la captación de las cifras económicas y sociales. La entidad recopila mediante modelos estadísticos, información de todos los sectores de la economía y la sociedad cubana, entre los que se encuentran datos referentes al ámbito de la Contabilidad, el trabajo y los salarios. Estos modelos contienen información sobre las entidades estatales, las sociedades mercantiles cubanas, las empresas mixtas y de capital totalmente extranjero, las organizaciones políticas y de masas, entre otras.

Para la renovación de esta entidad se realizó un estudio de su situación general, identificándose un conjunto de problemáticas en la organización y difusión de la información que se maneja. En la ONE controlar todo el proceso de organización de los datos es complejo, debido a que la información digital se encuentra almacenada en diferentes formatos, como ".x*ls"*, ficheros de texto, archivos ".*dbf"*, formato duro (papeles), archivos *".doc"* y otros. Se necesita conocimiento del negocio para entender los ficheros que se generan de manera mensual, trimestral y semestral, con el fin de ser procesados para obtener los principales reportes, cruces de variables, indicadores, tasas y porcentajes. Esto trae consigo que existan datos no integrados, múltiples versiones de los mismos, carencia de reportes flexibles y dificultad en el análisis de la información acumulada en el tiempo, obstaculizando así el proceso de la toma de decisiones.

Por lo anteriormente planteado surge como **problema de la investigación** ¿cómo apoyar la toma de decisiones en el área de Contabilidad, el trabajo y los salarios de la Oficina Nacional de Estadísticas?

Teniendo como **objeto de estudio**: los Almacenes de Datos, enmarcado en el **campo de acción** el Mercado de Datos para los indicadores de la Contabilidad, el trabajo y los salarios.

Para dar solución al problema planteado se ha trazado como **objetivo general**: desarrollar el mercado de datos Contabilidad, el trabajo y los salarios de la Oficina Nacional de Estadísticas que apoye a la toma de decisiones.

En correspondencia con el objetivo general propuesto se definieron los siguientes **objetivos específicos**:

1. Refinar el análisis y diseño del mercado de datos del área Contabilidad, el trabajo y los salarios.
2. Implementar el mercado de datos del área Contabilidad, el trabajo y los salarios.
3. Validar el mercado de datos del área Contabilidad, el trabajo y los salarios.

Para dar cumplimiento a los objetivos específicos y lograr una solución adecuada a la problemática especificada se plantean las siguientes **tareas de la investigación**:

Refinar el análisis y diseño del mercado de datos del área Contabilidad el trabajo y los salarios.

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.
4. Definición de los hechos, las medidas y las dimensiones del mercado de datos.
5. Diseño del modelo de datos.
6. Definición de la arquitectura del mercado de datos.
7. Diseño del subsistema de integración.
8. Diseño del subsistema de visualización.
9. Diseño de los casos de pruebas.

Implementar el mercado de datos del área Contabilidad el trabajo y los salarios.

1. Implementación del subsistema de integración.
2. Implementación del subsistema de visualización.

Validar el mercado de datos del área Contabilidad el trabajo y los salarios.

1. Aplicación de las listas de chequeo.
2. Aplicación de los casos de pruebas.

El presente documento está estructurado en cuatro capítulos. En ellos se describen los métodos y procedimientos a seguir para dar cumplimiento a los objetivos trazados. A continuación se expone una breve descripción de cada uno de ellos.

Capítulo 1: **Fundamentos teóricos**

En este capítulo se realiza un estudio del estado del arte acerca de los Almacenes de Datos (AD) enmarcándose en los Mercados de Datos (MD), así como conceptos esenciales relacionados al tema. Se estudian las metodologías y herramientas más utilizadas, analizando sus características, ventajas y desventajas, con el propósito de proponer la adecuada para el desarrollo de la investigación.

Capítulo 2: **Análisis y diseño del mercado de datos**

En este capítulo se realiza una descripción de los pasos a seguir durante el análisis y el diseño de la presente investigación. Se definen los requisitos que debe cumplir el sistema, así como el modelo dimensional propuesto para el desarrollo del Mercado de Datos (MD) a partir de los indicadores seleccionados.

Capítulo 3: **Implementación del mercado de datos**

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

Capítulo 4: **Validación del mercado de datos**

Se aplican las pruebas de validación al Mercado de Datos (MD) a través de las listas de chequeo, carta de aceptación del cliente y casos de pruebas.

CAPÍTULO 1: Fundamentos teóricos

1.1 Introducción

En este capítulo se realiza un estudio del estado del arte acerca de los Almacenes de Datos (AD) enmarcándose en los Mercados de Datos (MD), así como conceptos esenciales relacionados al tema. Se estudian las metodologías y herramientas más utilizadas, analizando sus características, ventajas y desventajas, con el propósito de proponer la adecuada para el desarrollo de la investigación.

1.2 Herramientas existentes para llevar estadísticas sobre indicadores económicos

En la actualidad el término Almacén de Datos (AD) o también conocido como *Data Warehouse* (DWH) por sus siglas en inglés, ofrece la solución como ubicación central para que todos puedan acceder a la información con los reportes necesarios, dando respuestas a necesidades de diferentes tipos de usuarios. Los AD surgieron con el objetivo de hacer consultable la información que se tiene de una empresa tanto de meses como de años anteriores.

En el AD se organiza y orienta la información desde la perspectiva del usuario final, mientras que en los sistemas operacionales se organiza desde la perspectiva de la aplicación, para lograr eficiencia en el acceso a datos [1].

1.2.1 Algunas herramientas existentes en el mundo

En el mundo los AD han servido de gran utilidad, al constituir una tecnología que les permite a los empresarios analizar con mayor rapidez las decisiones importantes de la empresa. Existen varias instituciones que emplean tecnologías de este tipo, por ejemplo: Visa, Telefónica de Argentina, Wallmart, Tv Azteca y muchos que han ido incorporando el uso de estas tecnologías para la toma de decisiones importantes.

En Europa las empresas BonPreu, WH Smith Books, Eroski, además de las grandes transaccionales como Coca Cola, Adidas, Bosh Siemens se han ido incorporando al empleo de los AD para la realización de estudios de mercado y de Inteligencia de Negocio (BI).

En el ámbito estadístico, en México, específicamente en el Instituto Nacional de Estadísticas e Informática (INEGI), se maneja la información mediante AD, de esta forma, mejoran el proceso de toma de decisiones a nivel gubernamental.

1.2.2 Algunas herramientas existentes en Cuba

En Cuba la aplicación de estas nuevas tecnologías aún se encuentra reducida y faltan muchos aspectos por mejorar, se han dedicado esfuerzos en este sentido, pero en general se puede decir que estos constituyen los primeros pasos para el futuro desarrollo de las organizaciones en este campo. Entre las organizaciones que han dado pasos firmes en este sentido se encuentra el AD comercial de la corporación CIMEX, que se dedica fundamentalmente a la exportación e importación de mercancías. El organismo centra su atención en la gestión de inventario, permitiendo una gestión de compra–venta eficiente, con la finalidad de disminuir los costos, sin afectar al cliente, permitiendo prestaciones eficientes y con la calidad requerida.

En el XIII Concurso Nacional de Computación (CNC) y en la Feria de Informática del 2002 se presentó un AD para CUBACEL desarrollado sobre la plataforma Oracle, con grandes resultados obtenidos a partir de su implantación. La UCI, se encuentra desarrollando este tipo de tecnología, aplicada en uno de sus casos al Sistema de Información de Gobierno (SIGOB), que se encuentra en proceso de construcción.

1.3 Almacenes de Datos (AD)

La utilización de bases de datos como plataforma para el desarrollo de aplicaciones informáticas en las organizaciones se ha incrementado notablemente en los últimos años, debido a la necesidad de las empresas de disponer de gran cúmulo de información almacenada. Con la necesidad de unir las distintas fuentes de información en un lugar único para la futura introducción de la documentación relevante, y como respuesta a la misma, es que surgen los AD [2].

Muchos son los criterios que se pueden encontrar para definir un AD. A continuación se muestran algunas de las definiciones dadas por algunos autores.

Los AD según Inmon[1], que fue uno de los primeros en hablar del tema, son un conjunto de datos que deben tener las siguientes características [3]:

Orientado por temas: los datos en el almacén están organizados de manera que todos los elementos de datos relativos al mismo evento u objetivo queden unidos entre sí.

[1] William Harvey Inmon (1945), experto reconocido mundialmente, es el creador de la llamada Corporate Information Factory.

Variables en el tiempo: los cambios producidos en los datos a lo largo del tiempo quedan registrados para que los informes que se puedan generar reflejen esas variaciones.

No volátil: la información no se modifica ni se elimina, una vez almacenado un dato, éste se convierte en información de solo lectura, y se mantiene para futuras consultas.

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

Por otra parte Ralph Kimball[2] otro conocido autor del tema lo define como "una copia de las transacciones de datos específicamente estructurada para la consulta y el análisis" [4].

El uso de almacenes de datos provee las siguientes ventajas:

- Integrar datos históricos sobre la actividad de la organización (o negocio) en un repositorio.
- Analizar los datos del negocio desde la perspectiva de su evolución en el tiempo.
- Prever tendencias de evolución del negocio.
- Identificar nuevas oportunidades de negocio y tomar decisiones estratégicas.
- Reducir los costes materiales y humanos en la toma de decisiones [5].

A continuación se muestran algunas de las desventajas de los AD:

- Riesgo de fracaso en la construcción del sistema, al subestimar los costes de captura y preparación de los datos.
- Riesgo de fracaso en la construcción del sistema por cambios continuos en los requisitos de los usuarios.
- Problemas con la privacidad de los datos [6].

Seguidamente se explican algunos de los elementos por los que están compuestos los AD:

Tablas de hechos: es donde las mediciones numéricas del negocio son almacenadas, cada una de las mediciones es tomada como la intersección de todas las dimensiones. Los mejores y más útiles hechos son numéricos, valorados continuamente y aditivos. La razón de esto es que, virtualmente cada consulta hecha contra la tabla de hechos pregunta por cientos, miles o aún millones de registros a ser usados por el sistema gestor de base de datos para construir el conjunto respuesta [7].

[2] Ralph Kimball, Doctor en Filosofía, ha sido uno de los mayores visionarios en la industria del Almacén de Datos desde 1982, actualmente, reconocido conferencista, consultante y profesor.

Tablas dimensionales: son aquellas donde las descripciones textuales de las dimensiones del negocio son almacenadas. Cada una de ellas ayuda a describir un miembro de la dimensión respectiva [8].

Vistas materializadas: a diferencia de las vistas "normales" una vista materializada almacena físicamente los datos resultantes al ejecutar la consulta definida en la vista. Este tipo realiza una carga inicial de los datos cuando se definen en la tabla de hechos y posteriormente se actualizan con una frecuencia establecida [8].

1.3.1 Componentes de un Almacén de Datos (AD)

Un AD está compuesto por un conjunto de elementos, necesarios para lograr el cumplimiento de sus objetivos. A continuación, se ofrece 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 [9].

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 [9].

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 [9].

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 encuentra en el área de presentación, lo que constituye el objetivo principal de los AD [9].

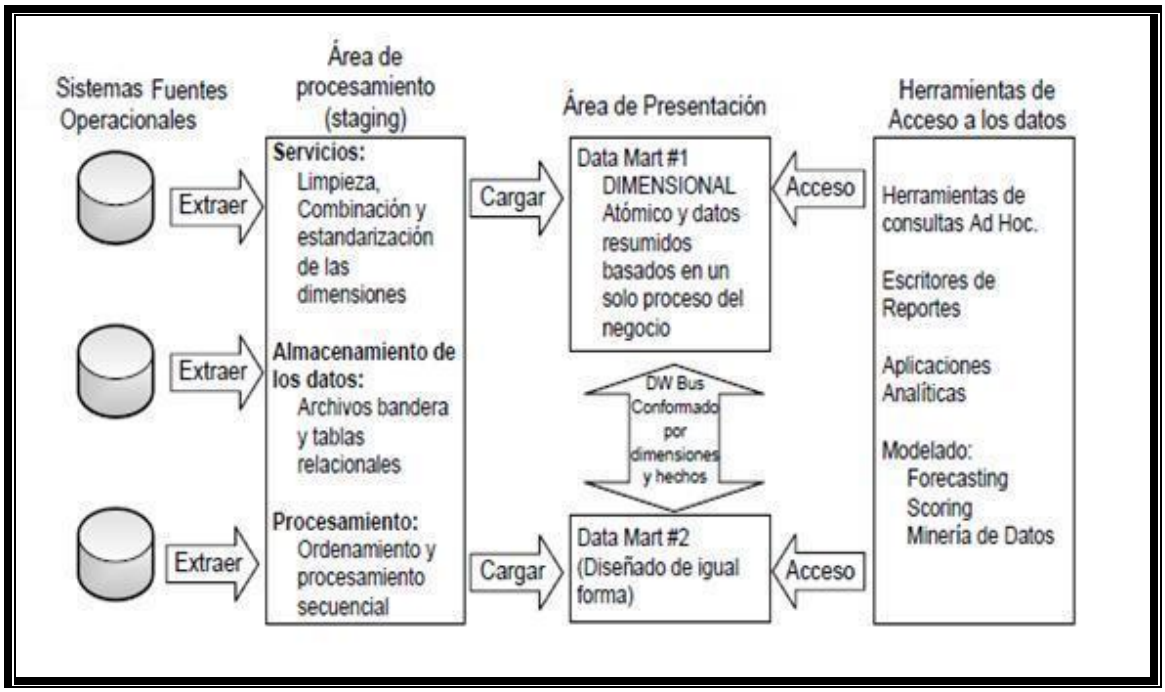

Figura 1: Componentes de un Almacén de Datos (AD)

1.3.2 Tipos de esquemas de un Almacén de Datos (AD)

El diseño multidimensional es un método de diseño de base de datos basados en el modelo relacional, está compuesto por las tablas de hechos y las dimensionales [9], a continuación se describen los esquemas multidimensionales estudiados para el desarrollo del MD (ver anexo uno):
Esquema estrella: consta de una tabla de hechos central y de varias tablas dimensionales relacionadas a esta, a través de sus respectivas claves [10].

Esquema copo de nieve: es una variante del esquema estrella en el que las tablas dimensionales de este último se organizan jerárquicamente mediante su normalización [10].

Esquema constelación de hechos: es un conjunto de tablas de hechos que comparten algunas tablas dimensionales [10]. Por sus características es la seleccionada para la realización del Mercado de Datos (MD).

1.3.3 Mercado de Datos (MD)

Los Mercados de Datos (MD) o *Data Mart* como se conoce en inglés, son un subconjunto de los Almacenes de Datos (AD) relativos a los requisitos de un departamento o área de negocio. Este subconjunto puede funcionar de manera autónoma, o bien enlazado al AD. El motivo por el cual se crean MD es el crecimiento que tiene el almacén y así facilitar su construcción y utilización [11].

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 [12].

Ventajas que proporciona el uso de los MD:

- Son un subconjunto de los AD.
- Permiten satisfacer las necesidades específicas de grupos de usuarios.
- Una misma organización puede tener varios MD.
- Son más fáciles de manejar [13].

Desventajas de los MD:

- Se pierde capacidad de procesamiento debido al crecimiento de los datos.
- Los usuarios necesitan acceder a varios MD.
- Complejidad a la hora de realizar el proceso de administración [13].

1.4 Integración de Datos

1.4.1 Procesos de Integración de datos

Los procesos de integración de datos están basados en la necesidad de unir los datos pertenecientes a múltiples fuentes con el fin de tener de forma centralizada, una mirada única e integrada al problema en cuestión.

En principio, y como principal problema los datos son heterogéneos y se encuentran distribuidos, dispersos y en la mayoría de los casos, no estandarizados. Existe de esta forma numerosa cantidad de información inconsistente que imposibilita una comprensión unificada en cuanto a los términos, cantidades, unidades de medida, de todas las entidades generadoras de datos, provocando que al unificar estos datos sea una tarea sumamente compleja y costosa, de forma tal que la integración tenga sentido, y puedan obtenerse resultados comparables y compatibles [14].

Existen diferentes procesos de integración dependiendo de la concepción que se tenga, algunos de ellos son:

- Replicación de datos.
- Integración de Información Empresarial (EII).
- Extracción, Transformación y Carga (ETL).

1.4.2 Extracción, Transformación y Carga (ETL)

El proceso ETL proporciona consolidación de datos para la construcción de bases de datos permanentes, utilizadas para el análisis o la generación de informes. Estas funciones se combinan en una herramienta para extraer datos fuentes y colocarlas en una base de datos destino. ETL se utiliza para migrar información de una o más bases de datos a terceros, para formar repositorios de datos, MD, AD y también para convertir bases de datos de un tipo o formato a otro. Es la tecnología que se utiliza actualmente en el centro DATEC para desarrollar los procesos de integración de datos [15].

Los subprocesos que componen el proceso de ETL son:

Extracción: proceso de lectura de datos desde los sistemas fuentes.

Transformación: proceso de conversión de los datos extraídos de su forma actual, en el formato que debe ser, en la que se pueden colocar en otros sistemas o bases de datos.

Carga: proceso de creación y ejecución de flujos de trabajo, para escribir los datos en los sistemas destinos. La carga de datos puede provocar el refrescado completo de un AD o puede hacerse mediante la actualización de la base de datos destino [15].

Características del proceso ETL:

- Es un mecanismo de carga muy eficiente y efectivo orientado a los AD.
- Enfocado a migrar y mezclar datos.
- Necesita pocos servicios de administración y mantenimiento.
- Gran capacidad para llevar a cabo transformaciones.
- Tecnología enfocada a la integración de datos en bases de datos versátiles hacia los AD [16].

1.5 Modos de almacenamiento de datos

OLAP (Procesamiento Analítico en Línea): es una tecnología que se basa en el análisis multidimensional de los datos y que le permite al usuario tener una visión más rápida e interactiva de

los mismos. Posee una gran capacidad para realizar cálculos de múltiples dimensiones, lo que permite gran variedad de informes y análisis de grandes volúmenes de datos [17].

Las principales características de OLAP son:

- ✓ **Rápido:** la primera regla se refiere a que el sistema debe ser capaz de responder de una forma rápida y ágil a la información que le sea solicitada por el usuario, el cual no deberá esperar más de cinco segundos a la hora de resolver peticiones sencillas y no más de veinte segundos en las peticiones complejas [18].
- ✓ **Análisis:** significa que el sistema debe poder reflejar cualquier lógica del negocio para poder responder a las preguntas específicas y necesidades empresariales [18].
- ✓ **Compartido**: el sistema deberá proporcionar herramientas que garanticen la confidencialidad de los datos y la seguridad de acceso por perfiles de los usuarios [18].
- ✓ **Multidimensional*:*** la herramienta deberá proporcionar soporte a cada una de las múltiples jerarquías que puedan existir dentro de la organización de información [18].
- ✓ **Información:** son todos los datos e información derivada de este proceso de análisis, la cual permitirá la toma de decisiones [18].

1.5.1 Procesamiento Analítico Relacional (ROLAP)

ROLAP: es organización física que se implementa sobre tecnología relacional. Para proporcionar los análisis OLAP esta arquitectura accede directamente a los datos almacenados en un AD. Los datos son acumulados en filas y columnas de forma relacional, y se les presentan a los usuarios en forma de dimensiones del negocio. ROLAP guarda la información en bases de datos relacionales, aprovechando así la tecnología relacional, permitiendo usar la integridad y seguridad de los sistemas gestores de bases de datos relacionales, además es capaz de manejar grandes volúmenes de datos [19].

Ventajas de ROLAP:

- ➢ Seguridad e integridad en la base de datos.
- ➢ Escalable para grandes volúmenes.
- ➢ Los datos pueden ser compartidos con aplicaciones SQL.
- ➢ Estructura más dinámica [19].

1.5.2 Procesamiento Analítico Multidimensional (MOLAP)

MOLAP: usa bases de datos multidimensionales para almacenar los datos, presentando un mejor rendimiento que la tecnología relacional en el procesamiento de las consultas, ofrece una mayor flexibilidad y rapidez de acceso para realizar el análisis de los datos [20].

Ventajas de MOLAP:

- Mayor rendimiento en el procesamiento de consultas.
- Posibilita hacer cálculos más complicados [20].

1.5.3 Procesamiento Analítico Híbrido (HOLAP)

HOLAP: es una combinación de ambas arquitecturas, que recoge las mejores características de cada una de ellas. Este modelo posee dos tipos de particionamiento: el vertical y el horizontal [21].

Ventajas de HOLAP:

- Reduce los costes de *hardware* ya que se necesita menos espacio en disco que en las bases de datos relacionales.
- Las respuestas de las consultas sobre las bases de datos multidimensionales son más rápidas que sobre las relacionales [21].

1.5.4 Justificación del modo de almacenamiento seleccionado

Después de realizarse un análisis de todos los modos de almacenamiento de datos mencionados anteriormente, se llegó a la conclusión de que el más adecuado a la solución es ROLAP pues soporta PostgreSQL, siendo este el sistema gestor de base datos seleccionado para la solución del Mercado de Datos (MD) Contabilidad, el trabajo y los salarios.

1.6 Metodologías

1.6.1 Metodología para el desarrollo de un Almacén de Datos (AD)

En múltiples disciplinas existen diferentes enfoques para abordar un mismo concepto o problema. El diseño de un AD, como disciplina que ha alcanzado ya un grado de madurez considerable a lo largo de estos años, también presenta diferentes vertientes. Existen dos criterios bien identificados y que han marcado claramente su tendencia sirviéndole de guía a la comunidad mundial en cuanto a este tema.

Los dos enfoques son la propuesta de Ralph Kimball (considerado por todos como el padre de la disciplina) y la de Bill H. Inmon. La principal diferencia que existe entre ambas tendencias está basada en la forma de enfrentar el problema.

La visión de Inmon se basa principalmente en un enfoque descendente (*top-down*), además plantea la creación de un repositorio de datos corporativo como fuente de información consolidada, persistente, histórica y de calidad. Al ser construido descendentemente, los MD se nutren del AD corporativo, convirtiéndose en un complejo empresarial de bases de datos relacionales [22].

La propuesta de Kimball se basa principalmente en un enfoque ascendente (*bottom-up*) debido a que plantea que se debe crear por cada departamento un conjunto de MD independientes, orientados a los temas que estén relacionados con él. Y "El Almacén de Datos es la unión de todos los Mercados de Datos de una entidad" [22].

La visión de Kimball de dividir el mundo de Inteligencia de Negocios (BI) entre el hecho y las dimensiones es muy eficaz y conduce a una solución completa en una cantidad muy pequeña de tiempo. Además, la propuesta de Kimball tiene una gran cantidad de documentación y se puede encontrar una respuesta a casi todas las preguntas que usted puede tener [23].

1.6.2 Metodología seleccionada

Se definió como metodología de desarrollo a utilizar, el modelo para el desarrollo de soluciones de Almacenes Datos (AD) e Inteligencia de Negocios (BI) de DATEC, que toma como base la metodología planteada por Kimball para el desarrollo de la solución, donde crea los conceptos de hechos y dimensiones, lo que indudablemente es muy eficaz en el proceso de la toma de decisiones y proporciona mayor agilidad en el proceso de desarrollo. Además, propone ir construyendo el AD a través de la construcción de los MD departamentales, lo que constituye una buena estrategia y coincide con la división lógica de las empresas, entidades y organismos. Es una metodología madura y reconocida por el resto de la comunidad dedicada al tema. Tiene bien definidas las etapas, actividades, artefactos y roles. Una característica importante para la selección, es que existe abundante documentación sobre la misma, a través de los servicios que brindan el grupo creador de la metodología. Como complemento y fortaleciendo la etapa del levantamiento de requisitos, se tomó lo planteado por Leopoldo Zenaido Zepeda Sánchez en su tesis de doctorado, orientando así el trabajo a los Casos de Uso (CU) y se logra estar más alineado con las tendencias y normas de la universidad.

1.7 Herramientas de modelado

Visual Paradigm

Visual Paradigm es una herramienta *CASE*[3] que utiliza *UML*[4] como lenguaje de modelado, con el uso del acercamiento orientado al objeto. Esta herramienta apoya los estándares más altos de las notaciones de *Java* y de *UML*. Está dotada de una buena cantidad de productos o módulos para facilitar el trabajo durante la confección de un *software*, lo cual garantiza la calidad del producto final [24].

Entre sus características están:

- Diseño centrado en Casos de Uso (CU) y enfocado al negocio que genera un *software* de mayor calidad.
- Uso de un lenguaje estándar común a todo el equipo de desarrollo que facilita la comunicación.
- Disponibilidad en múltiples plataformas.
- Contiene facilidades para redactar especificaciones de Casos de Uso del Sistema (CUS).
- Sincronización entre Diagramas de Entidad Relación (DER) y Diagramas de Clases (DC).
- Generación de documentos.
- Integración con distintos Ambientes de Desarrollo Integrados (IDE) [24].

1.8 Herramientas para el proceso de extracción, transformación y carga (ETL)

Pentaho Data Integration (PDI)

El Pentaho Data Integration (PDI) es el componente de Pentaho responsable de la extracción, transformación y carga (ETL) de los datos en el proceso de integración de los AD [25].

Como principales características se tiene:

[3] Las herramientas CASE (Computer Aided Software Engineering, Ingeniería de Software Asistida por Computadora) son diversas aplicaciones informáticas destinadas a aumentar la productividad en el desarrollo de software reduciendo el coste de las mismas en términos de tiempo y de dinero.

[4] Lenguaje Unificado de Modelado (UML, por sus siglas en inglés, Unified Modeling Language) es el lenguaje de modelado de sistemas de software más conocido y utilizado en la actualidad.

- Cada proceso es creado con una herramienta gráfica donde se especifica qué se va hacer sin necesidad de escribir un código que indique cómo hacerlo.
- Admite una amplia gama de formatos de entrada y salida, incluyendo archivos de texto, hojas de datos, archivos *XML*[5].
- Basado en repositorio, facilita la reutilización de componentes como transformación, colaboración y administración de modelos, conexiones, *logs*.
- *Debugger* integrado [26].

DataCleaner

El perfilado de datos es una de las principales tareas que se realizan en el proceso de calidad de datos, consiste en realizar un análisis sobre los datos provenientes de las fuentes, normalmente, sobre tablas, con el objetivo de empezar a conocer su estructura, formato y nivel de calidad [14].

Dentro de sus características se incluyen:

- Validación de los datos: el validador le dará un resultado que puede ser interpretado como bueno o malo, ya que valida los datos.
- Compatibles con diferentes tipos de base de datos: Oracle, MySQL, PostgreSQL, Firebird, SQLite [27].

La herramienta para el perfilado de datos que se decide utilizar en este trabajo es DataCleaner en su versión 1.5.3. Según su creador Kasper Sorense el sistema requiere *Java Runtime Enviroment* 5.0 o una versión superior y *Drivers* de JDBC[6].

1.9 Administrador de Base de datos

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

[5] Siglas en inglés de eXtensible Markup Language (lenguaje de marcas extensible).

[6] Java Database Connectivity (JDBC), permite la ejecución de operaciones sobre bases de datos desde el lenguaje de programación Java.

y Windows. Es capaz de gestionar versiones a partir de PostgreSQL 7.3, ejecutándose en cualquier plataforma [28].

1.10 Herramientas para el proceso de Inteligencia de Negocio (BI)

Mondrian Schema Workbench

El esquema Mondrian Workbench es una interfaz de diseño que le permite crear y probar esquemas de cubos OLAP visualmente. Los modelos de esquemas *XML* de metadatos se crean en una estructura específica utilizada por el motor de Mondrian. La estructura de estos modelos se pueden considerar de forma de cubos, que utilizan hechos existentes y tablas de dimensiones que se encuentran en su gestor de base de datos. Ofrece las siguientes funcionalidades:

- ✓ Editor de esquema integrado con el origen de datos subyacente para su validación.
- ✓ Probar consultas *MDX*[7] en contra del esquema de base de datos en pantalla.
- ✓ Examinar bases de datos de estructura subyacente en pantalla [27].

Mondrian OLAP Server

Para obtener la funcionalidad de Procesamiento Analítico en Línea (OLAP) se utiliza el servidor OLAP Mondrian, que permiten realizar consultas al AD y posibilitando que los resultados sean presentados mediante un navegador, de modo que el usuario pueda realizar las actividades típicas de navegación. Mondrian utiliza *MDX* como lenguaje de consulta, que fue un lenguaje propuesto por Microsoft. Funciona sobre las bases de datos estándar del mercado: Oracle, DB2, SQL-Server, PostgreSQL, MySQL, lo cual habilita y facilita el desarrollo del negocio basado en la plataforma Pentaho. Es un servidor OLAP *open source* que gestiona comunicación entre una aplicación OLAP (escrita en Java) y la base de datos con los datos fuente [27].

Pentaho BI Server

La plataforma Pentaho BI Server provee el soporte y la infraestructura necesaria 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

[7] MDX es el acrónimo de MultiDimensional eXpressions, es un lenguaje de consulta para bases de datos multidimensionales sobre cubos OLAP.

también incluye un motor de solución que integra reportes, análisis y componentes de minería de datos. Su diseño modular y la arquitectura basada en *plug-in* permiten a todos o parte de la plataforma estar inmersa en aplicaciones de terceros por los usuarios finales, así como fabricantes de equipos originales.

Algunas de sus ventajas son:

- ✓ Integración con procesos de negocio
- ✓ Administra y programa reportes
- ✓ Administra seguridad de usuarios [27].

Apache Tomcat

Apache Tomcat es una implementación de *software* de código abierto de *Java servlet*[8] y tecnologías *Java Server Pages* (JSP)[9]. Es desarrollado en un entorno abierto y participativo y publicado bajo la licencia Apache versión 2. Es la intención de ser una colaboración de los mejores desarrolladores de su clase de todo el mundo [29].

1.11 Sistema gestor de bases de datos

PostgreSQL

Es un sistema de gestión de base de datos relacional orientado a objetos y libre, publicado bajo la licencia BSD[10].

A continuación se enumeran las principales características de este gestor de bases de datos:

- ✓ Soporta distintos tipos de datos: además del soporte para los tipos base, también soporta datos de tipo fecha, monetarios, elementos gráficos, datos sobre redes, cadenas de bits, etc. También permite la creación de tipos propios.
- ✓ Incorpora *array* como una estructura de datos.

[8] Pequeño programa que corre en un servidor. Por lo general son aplicaciones Java que corren en un entorno de servidor web.

[9] JavaServer Pages (JSP) es una tecnología Java que permite generar contenido dinámico para web, en forma de documentos HTML, XML o de otro tipo.

[10] La licencia BSD es la licencia de software otorgada principalmente para los sistemas BSD (Berkeley Software Distribution).

- ✓ Soporta el uso de índices, reglas y vistas.
- ✓ Incluye herencia entre tablas.
- ✓ Permite la gestión de diferentes usuarios, como también los permisos asignados a cada uno de ellos [30].

1.12 Conclusiones del capítulo

En el presente capítulo se abordaron conceptos elementales relacionados con el tema donde:

- Se realizó una investigación sobre el estado del arte de las distintas bases de datos existentes para la confección de aplicaciones con Almacenes de Datos (AD).
- Se observaron las metodologías asentadas en el tema de Almacenes de Datos (AD) y basada en los fundamentos de Kimball se estableció como metodología el modelo para el desarrollo de soluciones de Almacenes de Datos (AD) e Inteligencia de Negocio (BI).
- Se decidió utilizar como gestor de base de datos PostgreSQL, por ser una herramienta liberada y brindar una serie de funcionalidades.
- Se definió como herramienta de administración PgAdmin3 y para el modelado de los datos se propone Visual Paradigm apoyado en UML.
- Para realizar el perfilado de los datos se seleccionó la herramienta DataCleaner y para el desarrollo de las transformaciones el Pentaho Data Integration (PDI).

CAPÍTULO 2: Análisis y Diseño del mercado de datos

Introducción

En este capítulo se realiza una descripción de los pasos a seguir durante el análisis y el diseño de la presente investigación. Se definen los requisitos que debe cumplir el sistema, así como el modelo dimensional propuesto para el desarrollo del Mercado de Datos (MD) a partir de los indicadores seleccionados.

2.1 Análisis de la solución

2.1.1 Definición del negocio

La ONE es el centro encargado de garantizar la producción de estadísticas de calidad a través del SEN, ejerciendo una adecuada dirección, ejecución y control de la captación de las cifras económicas y sociales, así como su adecuada difusión de acuerdo con las necesidades de la economía y las demás necesidades del país en información estadística.

La entidad cuenta con 22 áreas de trabajo, una de ellas es la dirección de Cuentas nacionales y dentro de esta el departamento de Sectores institucionales que abarca el área de Contabilidad, el trabajo y los salarios. En dicho departamento se definen los formularios o modelos con los que se trabaja y los indicadores y columnas que lo conforman; además se encarga de establecer la periodicidad con la que se recoge la información en los modelos estadísticos. Los datos recopilados son usados por áreas como: la dirección de Industria, Medio Ambiente, Agropecuaria, Comercio, Turismo, Servicios y la de Estadísticas sociales.

Ejemplos de cómo pueden intervenir en el tema de Contabilidad, el trabajo y los salarios en estas direcciones: en la parte de industria se necesita saber los gastos de la fuerza de trabajo en un período de tiempo dado; en la Agropecuaria se necesita saber la productividad o la cantidad de trabajadores disponibles; en Comercio, Turismo y Servicios, se quiere saber el total de ingresos en divisas, en un período de tiempo dado.

2.1.2 Tema de análisis identificado

Dada el área de trabajo se definió el tema de análisis que permite agrupar las principales necesidades de información. Esto posibilita determinar una organización global de los datos y enfocar la investigación en dominios informativos. Para la construcción de la propuesta se definió como tema de análisis: controles estadísticos de indicadores generales de la contabilidad, trabajo y los salarios.

2.1.3 Roles y permisos

Para garantizar la seguridad de la solución se definen los roles y los permisos correspondientes, en dependencia de la actividad que se desarrolle en el MD, ya sea lectura o escritura (ver tabla uno).

Roles	**Permisos**	
	Lectura	**Escritura**
Sobre la base de datos		
Administrador ETL	X	X
Administrador	X	
Analista	X	
Sobre la aplicación		
Administrador	X	X
Analista	X	

Tabla 1: Roles y permisos

Analista: analiza y consulta los reportes relacionados con los temas de análisis correspondientes a los modelos del área Contabilidad, el trabajo y los salarios.

Administrador de ETL: es la persona encargada de realizar el proceso de ETL del Mercado de Datos (MD).

Administrador: es la persona encargada de administrar permisos para el acceso a reportes y áreas de análisis, gestiona los roles y usuarios.

2.1.4 Reglas del negocio

Las reglas del negocio definen las políticas, normas, operaciones, definiciones y restricciones que la propuesta debe contemplar para cumplir con las necesidades del cliente. Para el desarrollo del MD se

definieron 15 reglas del negocio en el artefacto reglas del negocio y transformación del expediente de proyecto, a continuación se especifican dos ejemplos:

RN-1. La utilidad o pérdida del período, se obtiene mediante la diferencia entre el total de ingresos y el total de gastos.

RN-2. Para obtener el fondo de salario es necesario realizar la adición de los siguientes indicadores: fondo de salario escala, fondo de salario del pago adicional del perfeccionamiento empresarial, fondo de salario de otros pagos adicionales legalmente aprobados, fondo de salario por resultados y vacaciones acumuladas.

2.1.5 Necesidades de los usuarios

Es necesario y de gran importancia conocer que es lo que necesitan los usuarios. Pues de este conocimiento proviene la posibilidad de que los resultados estén en correspondencia de las necesidades, es decir, de que el producto sea satisfactorio o no. A continuación se describen las necesidades de los usuarios planteadas:

- ✓ El usuario necesita analizar la información del modelo 0005-11 “Indicadores generales” por división política administrativa, organismos, nomenclador de actividades económicas, forma de financiamiento, entidad y por el tiempo en que se recopilan la información, de todos los datos que se solicitan en el modelo: plan y real del año actual y plan del año anterior.
- ✓ El usuario necesita analizar la información del modelo 5901-08 “Indicadores seleccionados de la contabilidad” por división política administrativa, organismos, nomenclador de actividades económicas, forma de financiamiento, entidad y por el tiempo en que se recopilan la información, de todos los datos que se solicitan en el modelo: saldo inicial y final del año anterior, además del saldo inicial y final del año actual.
- ✓ El usuario necesita analizar la información del modelo 5903-04 “Cumplimiento del plan económico” por división política administrativa, organismos, nomenclador de actividades económicas, forma de financiamiento, entidad y por el tiempo en que se recopilan la información, de todos los datos que se solicitan en el modelo: real del año anterior al cierre 31/12, real acumulado del período del año anterior, plan del año actual al cierre 31/12, plan acumulado del período del año actual, real acumulado del período del año actual y el plan del año próximo.

2.1.6 Requisitos de información

Los requisitos de información son especificaciones que los clientes precisan para darle cumplimiento a las tareas internas del área Contabilidad, el trabajo y los salarios. Con el objetivo de controlar las necesidades establecidas, se han definido una serie de medidas e indicadores que muestran su comportamiento. Para el desarrollo del MD se identificaron 13 requisitos de información los cuales se encuentran descritos en el artefacto especificación de requerimientos del expediente de proyecto, a continuación se especifica un ejemplo.

RI1 Obtener el plan del año actual para todos los indicadores del modelo 0005-11, por las diferentes áreas: entidad, organismos, división político administrativa, forma de financiamiento, nomenclador de actividades económicas, por indicadores según el tiempo para la que se tiene la información.

2.1.7 Requisitos funcionales

Los requisitos funcionales están orientados a las necesidades de los usuarios finales. Son capacidades o condiciones que el sistema debe cumplir, lo que es muy importante para satisfacer las expectativas del cliente. En el desarrollo del presente trabajo se identificaron 16 requisitos funcionales.

RF1 Realizar extracción de los datos fuentes.

RF2 Realizar transformación y carga de los datos fuentes.

RF3 Autenticar usuario.

RF4 Adicionar usuario.

RF5 Eliminar usuario.

RF6 Adicionar rol.

RF7 Eliminar rol.

RF8 Adicionar reporte.

RF9 Eliminar reporte.

RF10 Modificar reporte.

RF11 Mostrar consulta *MDX*.

RF12 Suprimir filas y columnas vacías.

RF13 Imprimir reporte.

RF14 Visualizar reporte.

RF15 Exportar reporte como *Excel*.

RF16 Exportar reporte como *PDF.*

2.1.8 Requisitos no funcionales

Los requisitos no funcionales son características que de una u otra forma pueden limitar el correcto funcionamiento del sistema, como por ejemplo la seguridad, el rendimiento, la fiabilidad, entre otros. También hacen posible que el producto pueda desplegarse al ser terminado y funcionar de manera eficiente. Para el desarrollo del MD Contabilidad, el trabajo y los salarios se identificaron 22 requisitos funcionales los cuales se encuentran descritos en el artefacto especificación de requerimientos del expediente de proyecto, basados en las características del sistema que se va a realizar, a continuación se muestran algunos ejemplos:

Usabilidad

RNF 1. El sistema debe contar con un diseño del modelo físico sencillo, con una estructura y distribución que permita trabajar con rapidez y eficiencia.

Fiabilidad

RNF 2. El acceso a la información debe estar disponible el tiempo especificado y se tendrán en cuenta los permisos establecidos.

RNF 3. Se establecerán estrategias que permitan detectar posibles errores y darles solución en caso de ser posible.

Soporte

RNF 4. Lograr que los elementos definidos en el almacén tengan una estructura homogénea.

Las estructuras del AD se nombrarán de una manera estándar teniendo en cuenta el tipo de estructura que se maneje. Se definen convenciones de nombrado (ver anexo tres) con el objetivo de manejar un vocabulario común en el MD que permita un entendimiento claro y conciso de las estructuras por parte de los desarrolladores que interactúen con el AD.

Requisitos de seguridad

RNF 5. Sesión de usuarios.

Los permisos correspondientes al usuario autenticado se activarán una vez que éste se autentique y en caso de cambiar, tendrá acceso sólo a la información que le compete de acuerdo con sus privilegios.

Requisitos de *software*

RNF 6. Gestor de base de datos PosgreSQL 8.4.

RNF 7. Navegador *web* preferentemente Firefox 2.0 o una versión superior.

RNF 8. *Java Virtual Machine* 6.0 o una versión superior.

RNF 9. Schema Workbench 3.2.1 para el diseño de los cubos multidimensionales.

Requisitos de *hardware*

RNF 10. Se debe garantizar al menos una impresora para imprimir los reportes de salida.

2.1.10 Casos de Uso del Sistema

Para el diseño del diagrama de CUS (ver figura dos), se agruparon los 16 requisitos funcionales y los 13 de información en CU y se definieron las relaciones existentes entre ellos y los actores del sistema (ver anexo dos).

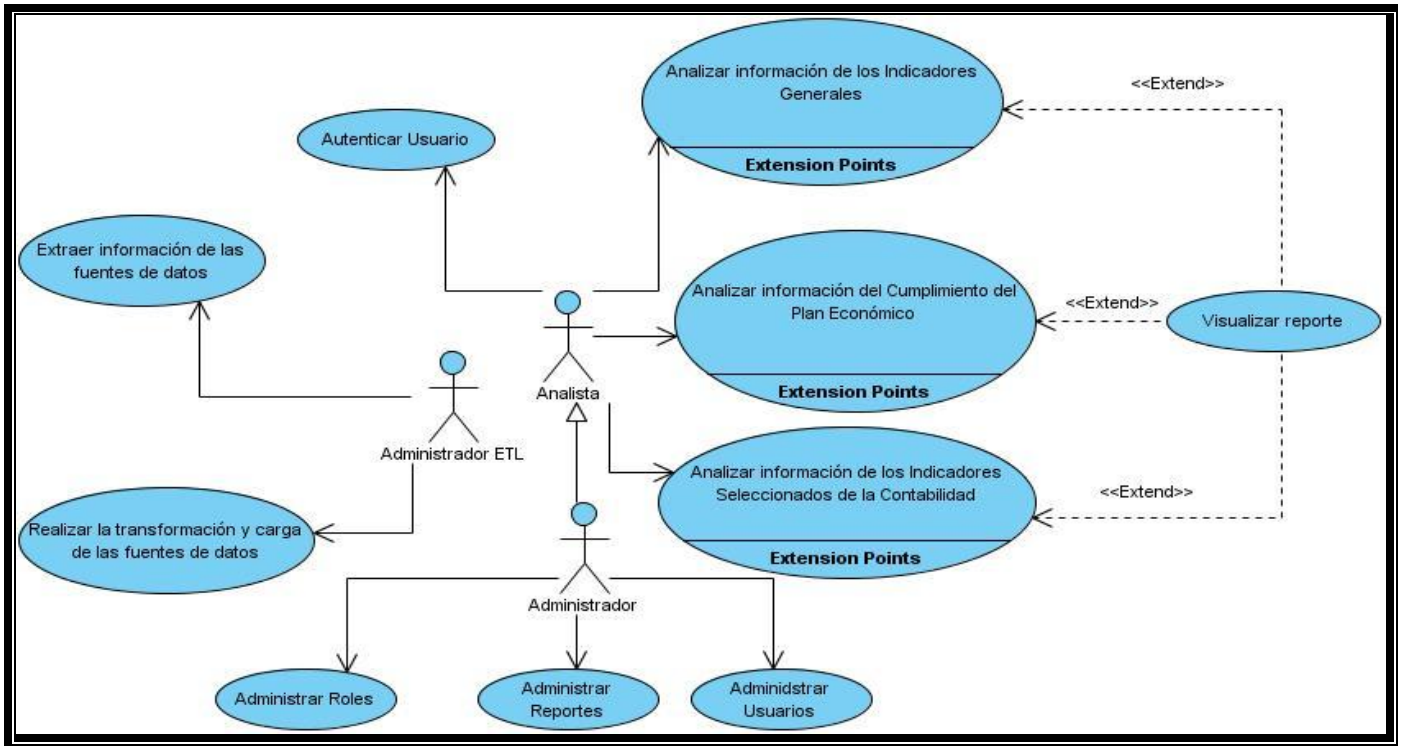

Figura 2: Diagrama de Casos de Usos del Sistema (CUS).

2.1.11 Descripción de los casos de usos críticos

Caso de Uso:	**Extraer información de las fuentes de datos.**
Tipo:	Funcional
Actor:	Administrador ETL
Resumen:	El caso de uso inicia cuando el administrador de ETL desea realizar la extracción de datos. Se selecciona la fuente de información correspondiente y extrae los datos contenidos en ella. El CU finaliza cuando todos los datos de la fuente son extraídos.
Precondiciones:	Disponibilidad de la fuente de datos.
Referencias	RF1

Prioridad	Crítico
Flujo Normal de Eventos	
Acción del Actor	**Respuesta del Sistema**
1. El administrador de ETL interactúa con la herramienta PDI para realizar la extracción de los datos	**2.** Muestra los repositorios disponibles para acceder a las transformaciones.
3. El administrador de ETL selecciona el repositorio con el cual va a trabajar.	**4.** Muestra el área de trabajo de la herramienta.
5. El administrador de ETL carga la transformación a ejecutar.	**6.** Muestra la transformación seleccionada por el usuario.
7. El administrador de ETL configura los parámetros de entrada de la transformación y presiona el botón pre visualizar.	**8.** Muestra los datos de los ficheros fuentes.
	9. Ejecuta la transformación seleccionada.
Flujos Alternos	
Acción del Actor	**Respuesta del Sistema**
	9.1. No muestra los datos de los ficheros
	9.2. Notifica el error al administrador de ETL.

Poscondiciones	Los datos de los ficheros "*dbf*" correspondientes son extraídos de la fuente y almacenados en un área temporal.

Tabla 2: Descripción del CU extraer información de las fuentes de datos.

Caso de Uso:	**Realizar la transformación y carga de los fuentes de datos**
Tipo:	Funcional.
Actores:	Administrador ETL.
Resumen:	El CU se inicia cuando el administrador de ETL selecciona los datos que van a ser transformados y cargados previamente extraídos. El actor realiza las transformaciones pertinentes carga la información hacia el MD finalizando así el CU.
Precondiciones:	La información debe ser extraída correctamente hacia el área temporal y las estructuras del almacén deben estar disponibles para ser usadas.
Referencias	RF 2
Prioridad	Crítico.

Flujo Normal de Eventos	
Acción del Actor.	**Respuesta del almacén.**
1. El administrador de ETL selecciona las estructuras del área temporal a transformar.	
2. El administrador de ETL carga los datos seleccionados en memoria.	
3. El administrador de ETL aplica las transformaciones pertinentes y genera datos de auditoría.	

4. El administrador de ETL carga datos en el almacén.	5. Ejecuta la consulta para insertar los datos en el almacén.
Prototipo de interfaz:	
Pos condiciones	Datos de la fuente transformados y cargados en el MD.

Tabla 3: Descripción del CU realizar la transformación y carga de las fuentes de datos.

Caso de Uso:	**Analizar información de los Indicadores generales.**
Tipo:	Información
Actores:	Analista, Administrador
Resumen:	El CU inicia cuando el actor desea consultar la información relacionada con el modelo Indicadores generales desde diferentes perspectivas de análisis. Luego de seleccionar el reporte deseado, el sistema muestra la información contenida en él y las opciones de los posibles cambios que le puede hacer al reporte. El CU finaliza cuando el actor termina el análisis de la información proveniente del modelo.
Precondiciones:	-El MD debe estar poblado correctamente. -Los reportes relacionados con el modelo deben estar disponibles.
Referencias	RI1, RI2, RI3
Prioridad	Crítico
Flujo Normal de Eventos	
Sección “”	

Acción del Actor	Respuesta del Sistema
1. El administrador se autentica en el sistema.	2. Muestra las áreas de análisis existente.
3. El administrador selecciona el AA Contabilidad, el trabajo y los salarios.	4. Muestra los Libros de Trabajos (LT) que contiene el AA Contabilidad, el trabajo y lo salarios.
5. El administrador selecciona el LT Indicadores generales.	6. Muestra los reportes pertenecientes al LT Indicadores generales.
7. El administrador selecciona el reporte deseado.	8. Muestra la información contenida en el reporte seleccionado y brinda la posibilidad de analizar el reporte desde diferentes perspectivas. Ir al CU: **Visualizar reporte.**

Opciones de reportes de Indicadores generales		
Entradas	**Posibles resultados**	
	Salidas	**Periodicidad**
Variables de entrada relacionadas con el CU: Analizar información de los Indicadores Generales. ➢ Entidad ➢ NAE ➢ DPA ➢ Empresas en perfeccionamiento ➢ Formas de financiamiento	Variables de salida disponibles en el CU: Analizar información de los Indicadores Generales. • Plan año actual. • Real año actual. • Año anterior.	Rango de tiempo en que se solicitan las variables de salidas. - Mensual

➢ Organismo ➢ Subordinación ➢ Indicador general ➢ Temporal semestre		
Pos condiciones	Los reportes correspondientes al libro de trabajo LT Indicadores generales se encuentran disponibles en el sistema.	

Tabla 4: Descripción del CU analizar información de los indicadores generales.

Caso de Uso:	**Analizar información de los Indicadores Seleccionados de la Contabilidad.**
Tipo:	Información
Actores:	Analista, Administrador
Resumen:	El CU inicia cuando el actor desea consultar la información relacionada con el modelo Indicadores Seleccionados de la Contabilidad desde diferentes perspectivas de análisis. Luego de seleccionar el reporte deseado, el sistema muestra la información contenida en él y las opciones de los posibles cambios que le puede hacer al reporte. El CU finaliza cuando el actor termina el análisis de la información proveniente del modelo.
Precondiciones:	-El mercado de datos debe estar poblado correctamente. -Los reportes relacionados con el modelo deben estar disponibles.
Referencias	RI4, RI5, RI6, RI17
Prioridad	Crítico
Flujo Normal de Eventos	

<table>
<tr><th colspan="3">Sección “”</th></tr>
<tr><th>Acción del Actor</th><th colspan="2">Respuesta del Sistema</th></tr>
<tr><td>1. El administrador se autentica en el sistema.</td><td colspan="2">2. Muestra las áreas de análisis existente.</td></tr>
<tr><td>3. El administrador selecciona el AA Contabilidad, el trabajo y los salarios.</td><td colspan="2">4. Muestra los Libros de Trabajos (LT) que contiene el AA Contabilidad, el trabajo y lo salarios.</td></tr>
<tr><td>5. El administrador selecciona el LT Indicadores seleccionados de la contabilidad.</td><td colspan="2">6. Muestra los reportes pertenecientes al LT Indicadores seleccionados de la contabilidad.</td></tr>
<tr><td>7. El administrador selecciona el reporte deseado.</td><td colspan="2">8. Muestra la información contenida en el reporte seleccionado y brinda la posibilidad de analizar el reporte desde diferentes perspectivas. Ir al CU: Visualizar reporte.</td></tr>
<tr><td colspan="3"></td></tr>
<tr><th colspan="3">Opciones de reportes de Indicadores seleccionados de la Contabilidad</th></tr>
<tr><th rowspan="2">Entradas</th><th colspan="2">Posibles resultados</th></tr>
<tr><th>Salidas</th><th>Periodicidad</th></tr>
<tr><td>Variables de entrada relacionadas con el CU: Analizar información de los Indicadores Seleccionados de la Contabilidad.
➢ Entidad
➢ NAE
➢ DPA</td><td>Variables de salida disponibles en el CU: Analizar información de los Indicadores Seleccionados de la Contabilidad.
• Saldo inicial del año anterior.
• Saldo final del año</td><td>Rango de tiempo en que se solicitan las variables de salidas.
- Semestral.</td></tr>
</table>

➢ Empresas en perfeccionamiento ➢ Formas de financiamiento ➢ Organismo ➢ Subordinación ➢ Indicador general ➢ Temporal semestre	anterior. • Saldo inicial del año actual. • Saldo final del año actual.	
Pos condiciones	Los reportes correspondientes al libro de trabajo LT Indicadores seleccionados de la contabilidad se encuentran disponibles en el sistema.	

Tabla 5: Descripción del CU analizar información de los indicadores seleccionados de la contabilidad

Caso de Uso:	**Analizar información del Cumplimiento del Plan Económico.**
Tipo:	Información
Actores:	Analista, Administrador
Resumen:	El CU inicia cuando el actor desea consultar la información relacionada con el modelo Cumplimiento del plan económico desde diferentes perspectivas de análisis. Selecciona el reporte que desea consultar, se muestra la información del reporte y brinda la posibilidad de realizarle nuevos cambios al reporte. El CU finaliza cuando el actor termina el análisis de la información proveniente del modelo.
Precondiciones:	-El mercado de datos debe estar poblado correctamente. -Los reportes relacionados con el modelo deben estar disponibles.
Referencias	RI8, RI9, RI10, RI11, RI12, RI13
Prioridad	Crítico

<table>
<tr><th colspan="3">Flujo Normal de Eventos</th></tr>
<tr><th colspan="3">Sección “”</th></tr>
<tr><th>Acción del Actor</th><th colspan="2">Respuesta del Sistema</th></tr>
<tr><td>1. El administrador se autentica en el sistema.</td><td colspan="2">2. Muestra las áreas de análisis existente.</td></tr>
<tr><td>3. El administrador selecciona el AA Contabilidad, el trabajo y los salarios.</td><td colspan="2">4. Muestra los LT que contiene el AA Contabilidad, el trabajo y lo salarios.</td></tr>
<tr><td>5. El administrador selecciona el LT Cumplimiento del plan económico.</td><td colspan="2">6. Muestra los reportes pertenecientes al LT Cumplimiento del plan económico.</td></tr>
<tr><td>7. El administrador selecciona el reporte deseado.</td><td colspan="2">8. Muestra la información contenida en el reporte seleccionado y brinda la posibilidad de analizar el reporte desde diferentes perspectivas. Ir al CU: Visualizar reporte.</td></tr>
<tr><td colspan="3"></td></tr>
<tr><th colspan="3">Opciones de reportes de Cumplimiento del Plan Económico.</th></tr>
<tr><th rowspan="2">Entradas</th><th colspan="2">Posibles resultados</th></tr>
<tr><th>Salidas</th><th>Periodicidad</th></tr>
<tr><td>Variables de entrada relacionadas con el CU: Analizar información del Cumplimiento del Plan Económico.
➢ Entidad
➢ NAE
➢ DPA
➢ Empresas en perfeccionamiento</td><td>Variables de salida disponibles en el CU: Analizar información de los Indicadores Seleccionados de la Contabilidad.
• Real año anterior al cierre
• Real acumulado del</td><td>Rango de tiempo en que se solicitan las variables de salidas.
- Trimestral.</td></tr>
</table>

➢ Formas de financiamiento ➢ Organismo ➢ Subordinación ➢ Indicador general ➢ Temporal semestre	período año anterior. • Plan año actual al cierre. • Plan acumulado del periodo año actual. • Real acumulado del período año actual. • Plan año próximo.	
Pos condiciones	Los reportes correspondientes al libro de trabajo LT Cumplimiento del plan económico se encuentran disponibles en el sistema.	

Tabla 6: Descripción del CU analizar información del cumplimiento del plan económico

2.2 Diseño de la Solución

2.2.1 Matriz BUS

La Matriz Bus o dimensional representa las relaciones existentes entre los hechos y las dimensiones del Mercado de Datos (MD).

TH/DIM	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12
hech_indic_generales_0005_11	X	X	X	X	X	X	X	X	X			X
hech_saldos_dl_contabilidad_5901_08	X	X	X	X	X	X	X	X		X		X
hech_cumplim_plan_econ_5903_04	X	X	X	X	X	X	X	X			X	X
vm_calculo_d_indicadores	X	X	X	X	X	X	X			X		X

Tabla 7: Matriz Bus.

Lista de dimensiones

D1 dim_entidad

D2 dim_nae

D3 dim_dpa

D4 dim_empresa_perfeccionamiento

D5 dim_ffinanciamiento

D6 dim_organismo

D7 dim_subordinacion

D8 dim_indicador_general

D9 dim_temporal_mes

D10 dim_temporal_trimestre

D11 dim_temporal_semestre

D12 dim_esfera

2.2.2 Modelo de Datos

El modelo de datos representa una descripción de la estructura de los datos, se identifican las relaciones entre las dimensiones y los hechos así como las variables que los componen. Para el desarrollo del MD se identificaron tres hechos, una vista materializada y 12 dimensiones (ver anexo tres). La figura tres, muestra una porción del modelo de datos perteneciente al área de la Contabilidad, el trabajo y los salarios.

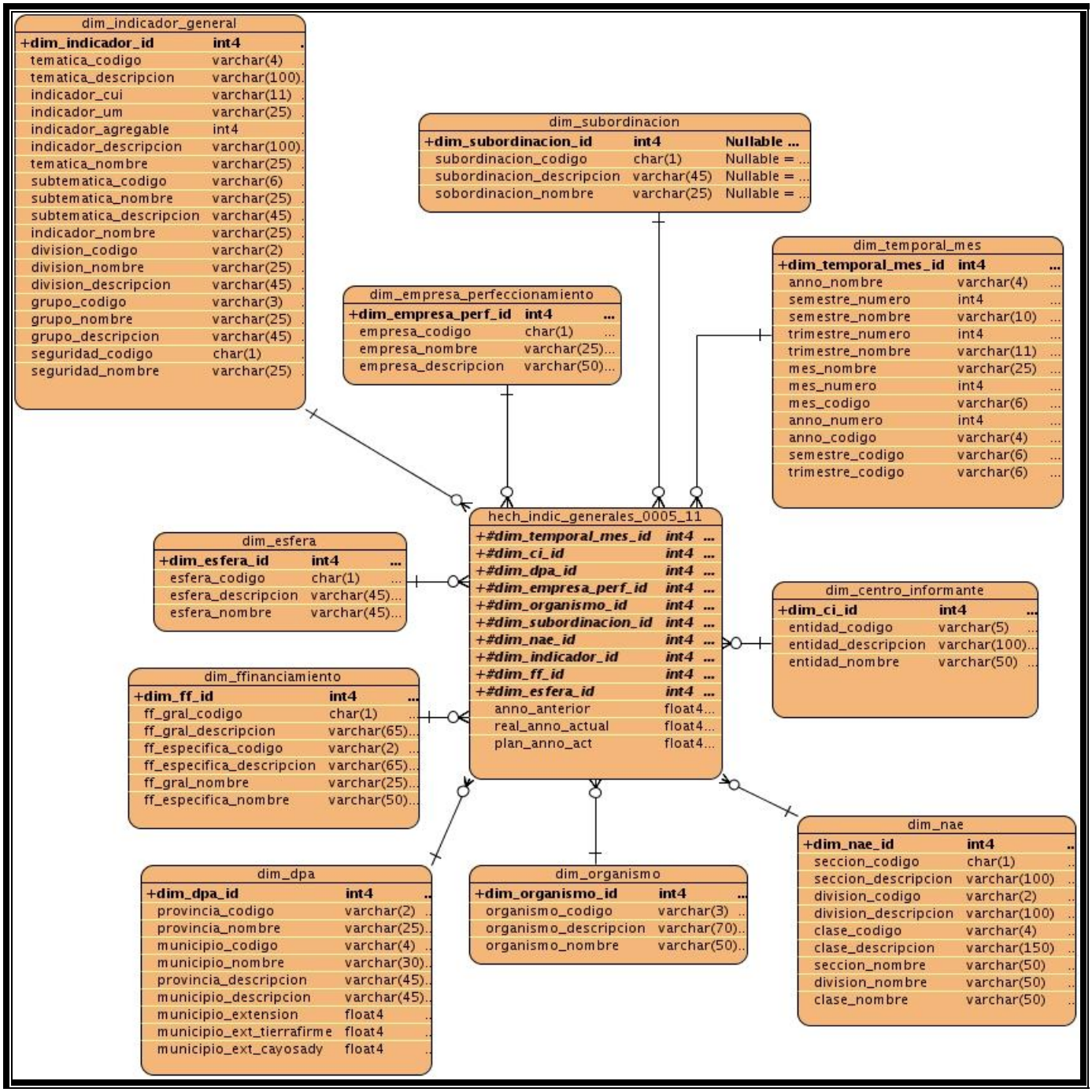

Figura 3: Modelo físico de datos.

2.3 Conclusiones del capítulo

En el presente capítulo se realizó un análisis y diseño del Mercado de Datos (MD) Contabilidad, el trabajo y los salarios donde:

- Se definieron las reglas del negocio, roles y permisos.
- Se definieron los Casos de Uso del Sistema (CUS).
- El modelo de datos fue refinado, quedando aprobado por el cliente.
- Se refinaron las dimensiones y los hechos asociados.
- Se realizó el modelado tanto lógico como físico de los datos, generándose el *script* necesario para llevar a cabo la propuesta de solución.

CAPÍTULO 3: Implementación del mercado de datos

Introducción

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

3.1 Implementación del modelo de datos

Durante la implementación del modelo de datos se tuvo en cuenta todo lo especificado anteriormente en el modelo físico, los tipos de datos de las variables, los esquemas que agrupan las dimensiones y los hechos del MD. El modelo de datos cuenta con 12 tablas dimensionales, tres tablas de hechos y una vista materializada (ver anexo cuatro).

3.2 Implementación del subsistema de integración

3.2.1 Arquitectura del subsistema de integración

A partir del estudio de las características del proceso ETL, se puede deducir que es un proceso complejo, debido a su alto nivel de detalle, por lo cual se debe regir por una arquitectura para así lograr un buen diseño, la cual consta de los siguientes componentes:

Fuente de datos: contiene la información que abarca el área de análisis, en el caso del MD Contabilidad, el trabajo y los salarios, los ficheros están en formato "*dbf"*.

Área temporal (*Staging Area*): es la línea divisoria entre la fuente de datos y el MD, donde se realizan los subprocesos de transformación y limpieza de los datos.

Almacén de datos: constituye el área de destino a donde se almacenaran los datos [13].

Figura 4: Arquitectura del subsistema de integración.

3.2.2 Perfilado de datos

El perfilado de datos consiste en realizar una descripción detallada de los sistemas fuentes, con el objetivo de conocer su estructura, formato y calidad de la información. A partir del desarrollo de este proceso, se identifican las principales reglas de transformación, útiles para lograr una perfecta disponibilidad del MD.

Para el análisis de los datos correspondiente al desarrollo de la solución se empleó la herramienta DataCleaner 1.5.4, con el cual se identificaron los posibles valores de las variables de cada campo proveniente de la fuente, así como los valores máximos y mínimos de las medidas. En el caso de la información proveniente de los modelos relacionados con la Contabilidad, el trabajo y los salarios, no se obtuvieron valores nulos (ver anexo cinco).

3.2.3 Extracción, transformación y carga (ETL) de los datos

El proceso de ETL se inicia al extraer los datos desde los sistemas de origen. El formato de los datos fuentes correspondientes al MD se encuentran en ficheros de tipo "*dbf*" separados por modelos en diferentes directorios.

Después de realizada la extracción de los datos se procede a realizar las transformaciones pertinentes al MD, las cuales constituyen un elemento básico dentro de la implementación del proceso ETL. En este paso la información se valida y se adapta al modelo de datos desarrollado anteriormente.

El último de los subprocesos de ETL válidos para el desarrollo de la solución, es la carga de la información. Consiste en migrar los datos que han sido transformados anteriormente al AD, para después realizar la implementación de la capa de visualización. La figura cinco, muestra un ejemplo de una de las transformaciones desarrolladas para poblar el MD Contabilidad, el trabajo y los salarios. En ella se visualiza una entrada de fichero "*.dbf"* donde se recoge toda la información de la fuente perteneciente al modelo Indicadores generales, se valida la entrada previendo la existencia de valores nulos en los campos, en caso de encontrarlos, se envían a un fichero "*.xls"* en el área temporal para darle un posterior tratamiento, aplicando la estrategia de llaves nulas. Posteriormente se obtienen las llaves primarias de cada dimensión perteneciente al hecho en el cual se van a insertar los datos, una vez realizadas las búsquedas pertinentes, se valida la información obtenida dándole el mismo tratamiento antes descrito.

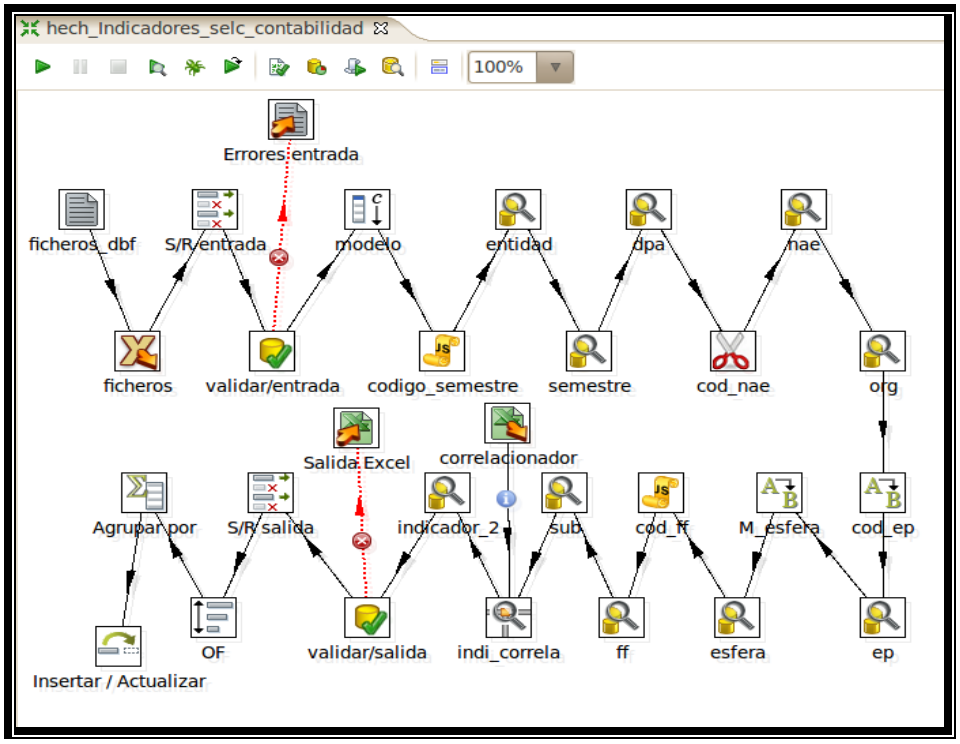

Figura 5: Transformación del modelo 0005 indicadores generales.

3.2.4 Implementación de los trabajos

Los trabajos constituyen un proceso sencillo de tareas con el objetivo de realizar una operación específica. En ellos podemos utilizar diferentes pasos como por ejemplo mandar un correo electrónico o emitir un mensaje de texto. Además, se puede ejecutar una o varias transformaciones de las que se diseñaron y elaborar una secuencia de ejecución lógica de ellas. Los trabajos normalmente se planifican para ser ejecutados automáticamente en intervalos regulares.

Para la ejecución de las transformaciones desarrolladas para la solución se implementó un trabajo (ver figura seis), el cual está compuesto por un paso inicial que da inicio a la ejecución, después se verifica la conexión a la base de datos donde serán insertados los datos, en caso de no estar funcionando se muestra un mensaje donde se detalla el problema ocurrido. Posteriormente se muestran los siete pasos para ejecutar las transformaciones realizadas para el MD, por último, se muestra un mensaje donde se especifica que el flujo ha concluido. Durante todo este proceso se guardan los *logs* de cada transformación en el área temporal, para tener una información detallada de cada ejecución del trabajo.

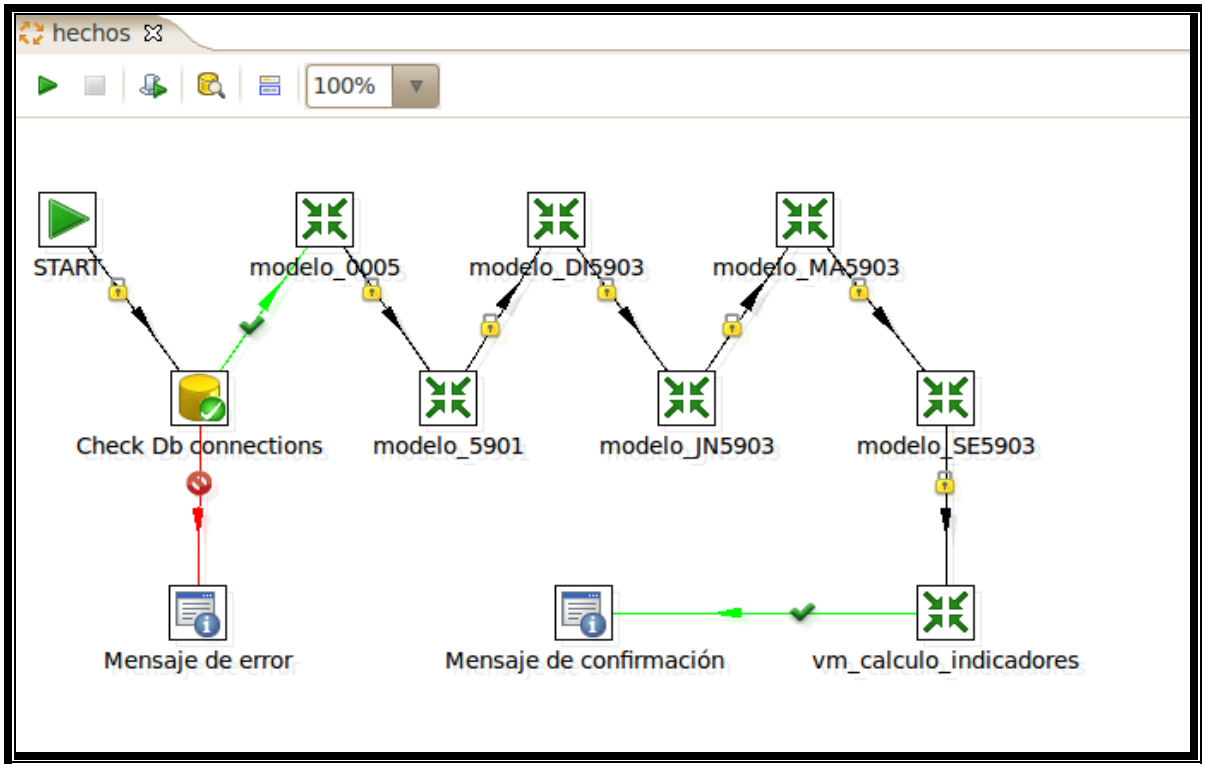

Figura 6: Trabajo de los hechos del Mercado de Datos (MD).

3.3 Implementación del subsistema de visualización

3.3.1 Cubos OLAP

La implementación de los cubos OLAP se realizó en la herramienta Pentaho Schema Workbench. Esta permite generar un fichero de configuración "*.xml"*, en el cual se definen los cubos y las dimensiones con sus niveles de jerarquía, así como la conexión con el mercado que contiene los datos para el cubo multidimensional. En el presente trabajo se modelaron cinco cubos, con las características correspondientes a cada una de las tablas de hechos y dimensiones (ver figura siete).

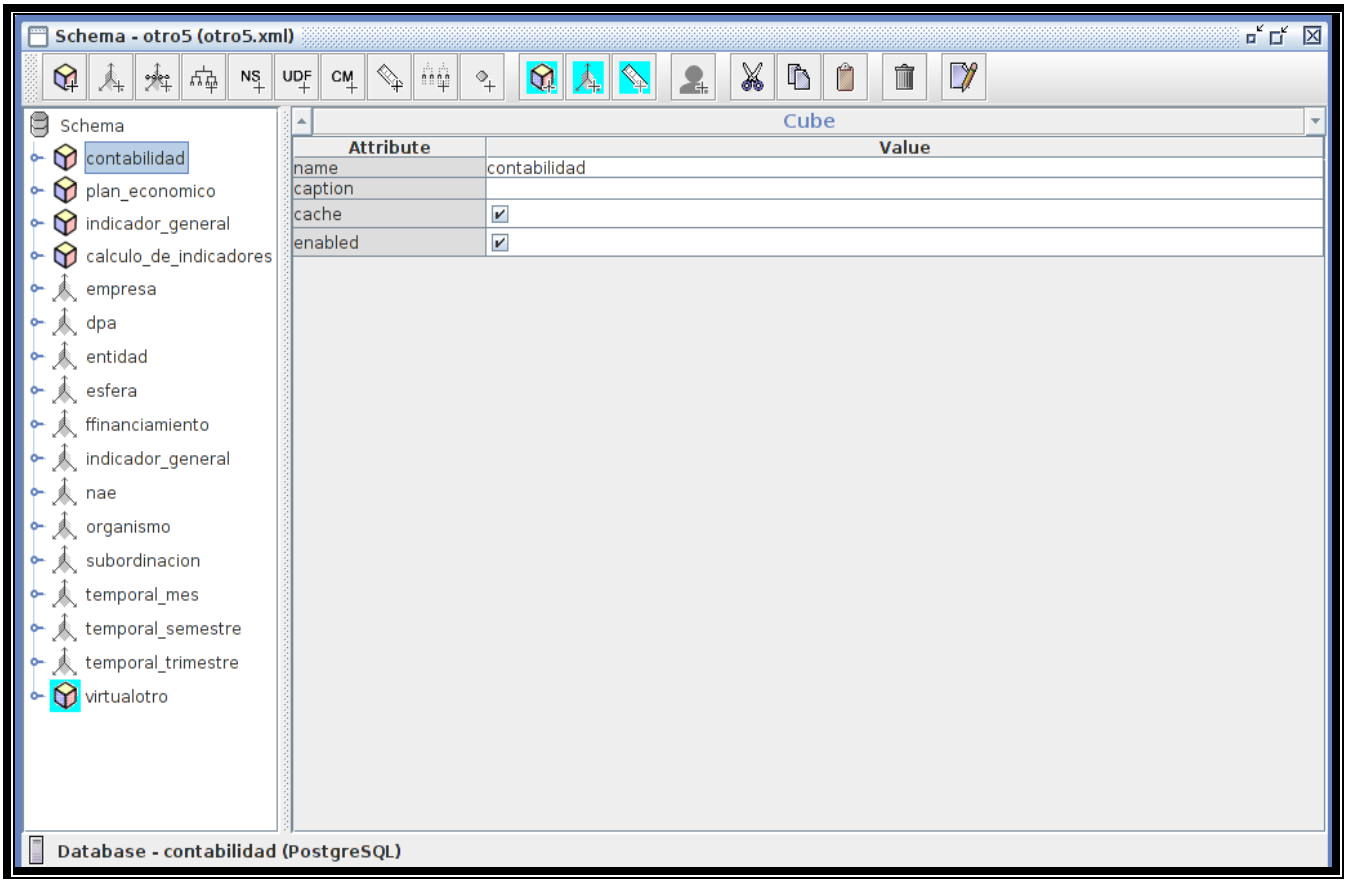

Figura 7: Diseño de los cubos utilizando Pentaho Schema Workbench.

3.3.2 Navegación de la capa de visualización

El artefacto Arquitectura de información del expediente de proyecto, contribuye a definir el entorno de análisis, monitoreo y control de los modelos estadísticos de la Contabilidad, el trabajo y los salarios. En este se identificó un Área de Análisis (A.A), tres Libros de Trabajos (LT) y 16 vistas de análisis. A continuación se detallan los elementos que componen las estructuras de navegación de la información presentada en la capa de visualización:

Descripción de los Libros de Trabajo (LT):

L.T Contabilidad: libro de trabajo contenido dentro del área de Contabilidad, el trabajo y los salarios, que agrupa las vistas de análisis correspondientes al modelo Saldos de la contabilidad.

L.T Indicadores generales: libro de trabajo contenido dentro del área de Contabilidad, el trabajo y los salarios, que contiene las vistas de análisis correspondiente al modelo Indicadores generales.

L.T Plan económico: libro de trabajo contenido dentro del área de Contabilidad, el trabajo y los salarios, que agrupa las vistas de análisis del modelo Cumplimiento del plan económico.

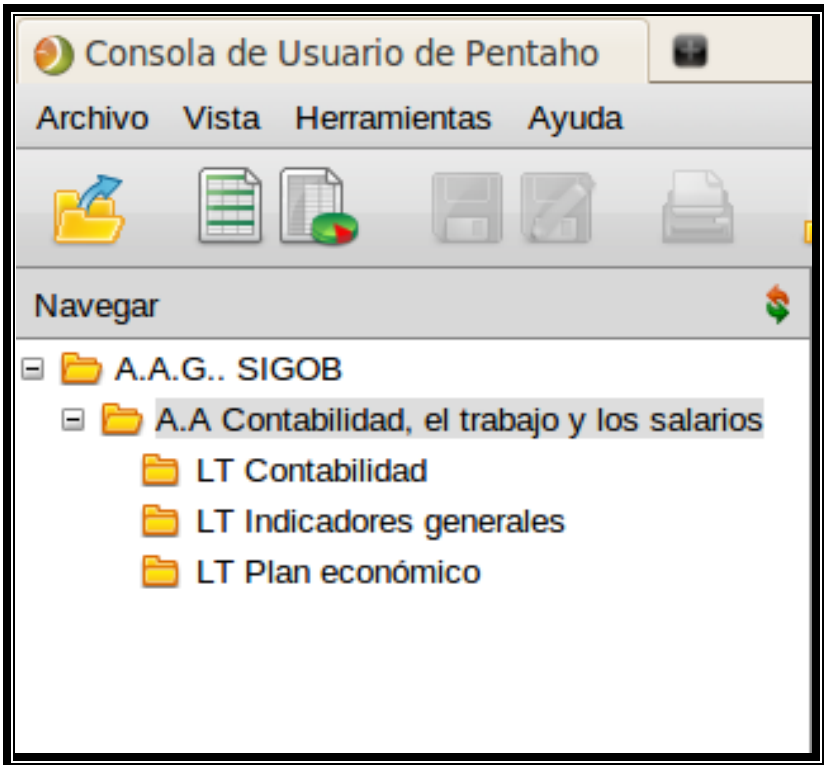

Figura 8: Arquitectura de información.

3.3.3 Implementación de los reportes candidatos

Los reportes candidatos representan la información que el cliente desea que se muestre como finalidad del producto. Fueron seleccionados luego de realizar un análisis de los modelos antes descritos, donde se recoge toda la información referente al área de Contabilidad, el trabajo y los salarios de la ONE. A continuación se representa un reporte del hecho saldos de la contabilidad, en el cual se identifican los indicadores y los valores de las medidas.

	Temporal semestre			
	Segundo Semestre			
	Medidas			
Indicador general	● Saldo final del año actual	● Saldo final del año anterior	● Saldo inicial del año actual	● Saldo inicial del año anterior
Gasto Material	359,8	407,1	0,0	0,0
Otros gastos de la fuerza de trab.	969,6	1099,7	0,0	0,0
Impuesto utiliz.defuerza de trabajo	623,3	705,5	0,0	0,0
Depreciación y amortización	317,6	409,3	0,0	0,0
Deprec.de activos fijos tangibles	1890,8	1865,5	1521,7	1446,9
Otros gastos monetarios	21803,6	26834,3	0,0	0,0
Servicios comprados	1515,8	2016,0	0,0	0,0
Comisión por servicios	105,2	50,9	0,0	0,0
Otros gastos de personal	17,4	10,5	0,0	0,0
Primas de seguros	11,7	15,7	0,0	0,0
Gastos financieros	206,3	214,7	0,0	0,0
Activos fijos tangibles	3240,3	3188,7	3188,7	3200,0
Act. fijos tangibles en ejecución	0,0	0,0	0,0	15,1
Activos fijos intangibles	6,8	6,8	6,8	6,8
Amortiz. activos fijos intangibles	6,8	3,9	3,9	0,6

Figura 9: Reporte "total de indicadores" del modelo Saldos de la contabilidad.

3.4 Conclusiones del capítulo

Después de realizar la implementación del Mercado de Datos (MD) Contabilidad, el trabajo y los salarios se obtuvieron las siguientes conclusiones:

- Se implementó el trabajo y las transformaciones necesarias para extraer, transformar y cargar los datos hacia el Mercado de Datos (MD).
- Quedó definida la estructura de los datos a partir del modelo de datos físico, contando con dos esquemas: dimensiones compuesto por 12 tablas dimensionales y mart_contabilidad que contiene tres hechos y una vista materializada.
- Se diseñaron e implementaron los cubos OLAP, quedando definidos cinco cubos, 12 dimensiones y 63 medidas calculables.
- Se desarrolló el subsistema de visualización, determinándose tres libros de trabajo, con un total de 16 vistas de análisis.

CAPÍTULO 4: Validación del mercado datos

Introducción

Se aplican las pruebas de validación al Mercado de Datos (MD) a través de las listas de chequeo, carta de aceptación del cliente y casos de pruebas.

4.1 Pruebas

Las pruebas de *software* son los procesos que permiten verificar y revelar la calidad de un producto. Son utilizadas para identificar posibles fallos de implementación, calidad y usabilidad en la solución. Para determinar el nivel de calidad se deben efectuar pruebas que permitan comprobar el grado de cumplimiento de las especificaciones iniciales del sistema.

Para la validación del MD Contabilidad, el trabajo y los salarios se pueden aplicar diferentes tipos de pruebas a continuación se especifican algunas de ellas:

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.

Para realizar las pruebas al MD Contabilidad, el trabajo y los salarios se diseñó una lista de chequeo y tres casos de prueba (ver anexo dos).

Listas de chequeo: constituyen un mecanismo para el control de los riesgos, tienen como función básica detectar condiciones peligrosas que puedan generar incidentes al producto de *software*.

Para elaborar la lista de chequeo se tuvieron en cuenta elementos de evaluación que son importantes una vez realizado el proceso de ETL y BI, permitiendo recoger los puntos eficientes e ineficientes que posean dichos procesos. La lista de chequeo contiene diferentes indicadores a evaluar los cuales se encuentran distribuidos en tres secciones fundamentales (ver anexo siete):

- 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.

Casos de prueba: tienen como propósito validar cada Caso de Uso de Información (CUI) definido en el DCUS, aunque pueden existir más casos de prueba en dependencia de la complejidad de los CUI que hayan sido identificados. La aplicación de estas pruebas, permite demostrar que las vistas de análisis definidas en el MD satisfacen los requisitos de información identificados, garantizando así el cumplimiento de uno de los principales objetivos del sistema.

Escenario	**Descripción**	**Perfiles de análisis**	**Indicadores a medir**	**Respuesta del sistema**	**Flujo central**
Indicador por DPA	Se obtiene las medidas del modelo 0005 para los indicadores generales por la división política administrativa	Indicador general DPA Temporal mes	Año anterior Real del año actual Plan año actual Cumplimiento Crecimiento Variación con respecto al plan	El sistema muestra todas las variables disponibles para los análisis, ubicados en las filas y las columnas que pueden ser visualizadas para cada reporte.	Se abre la aplicación. Se autentifica. Se entra al sistema. Se selecciona la opción Nueva Vista de Análisis. Se selecciona el esquema “otro5”. Se selecciona el cubo

			Variación respecto al año anterior		indicador_general. Se selecciona la opción Abrir navegador OLAP en la parte superior izquierda. Se verifica que estén disponibles en las filas y las columnas todos los perfiles de análisis y los indicadores a medir.
Indicadores por NAE.	Se obtiene las medidas del modelo 0005 para los indicadores generales por el nomenclador de actividades económicas	Indicador general NAE Temporal mes	Año anterior Real del año actual Plan año actual Cumplimiento Crecimiento Variación con respecto al plan Variación respecto al año anterior		
Indicador por forma de financiamiento.	Se obtiene las medidas del modelo 0005 para los indicadores generales por la forma de financiamiento	Forma de financiamiento Indicador general Temporal mes	Año anterior Real del año actual Plan año actual Cumplimiento Crecimiento Variación con respecto al plan		

			Variación respecto al año anterior		
Indicadores por organismos	Se obtiene las medidas del modelo 0005 para los indicadores generales por organismos	Indicador general Organismo Temporal mes	Año anterior Real del año actual Plan año actual Cumplimiento Crecimiento Variación con respecto al plan Variación respecto al año anterior		
Total de indicadores	Se obtiene las medidas del modelo 0005 para los indicadores generales	Indicador general Temporal mes	Año anterior Real del año actual Plan año actual Cumplimiento Crecimiento Variación con respecto al plan		

			Variación respecto al año anterior

Tabla 8: Escenario "reportes candidatos" del caso de prueba indicadores generales.

4.1.1 Pruebas aplicadas

Casos de prueba

Para la validación de la solución se aplicaron pruebas de interfaz mediante los casos de prueba diseñados (ver anexo cinco). Se aplicaron a nivel departamental, a nivel de centro y por parte de calidad UCI, donde se identificaron 28 no conformidades en total, las que posteriormente se corrigieron para lograr una correcta disponibilidad de la información.

Pruebas de aceptación

Para aprobar la propuesta de solución se realizó un encuentro con la representante de la ONE en la UCI Elena Leonila Fernández García, quedando conforme con la propuesta mostrada y satisfecha con el cumplimiento de los requisitos anteriormente planteados.

4.2 Conclusiones del capítulo

En el capítulo se realizó la validación de la propuesta del Mercado de Datos (MD) para el área de Contabilidad, el trabajo y los salarios donde:

- Se diseñaron y aplicaron los casos de prueba para validar la solución.
- Se diseñó una lista de chequeo para posteriormente aplicarla al Mercado de Datos (MD).
- Se realizaron pruebas de validación al Mercado de Datos (MD) con el cliente para confirmar que el producto cumple con sus necesidades.

Conclusiones

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

- Se refinó el análisis y diseño del Mercado de Datos (MD) Contabilidad, el trabajo y los salarios, donde se seleccionó la metodología y las herramientas útiles para el desarrollo de la solución.
- Se implementó el MD Contabilidad, el trabajo y los salarios, quedando desarrollado el modelo de datos, así como la implementación del subsistema de integración y visualización.
- Se validó el Mercado de Datos (MD) Contabilidad, el trabajo y los salarios, donde se aplicaron por parte de los especialistas del departamento los tres casos de prueba diseñados.

RECOMENDACIONES

- Implementar un mecanismo más efectivo para el tratamiento de errores de los ficheros fuentes, permitiéndole a los usuarios insertar los datos ya arreglados, sin necesidad de cargar toda la información nuevamente.

- Crear un repositorio para los metadatos, donde se guarde toda la información de la ejecución de las transformaciones de una manera más detallada y entendible para el administrador de ETL.

BIBLIOGRAFÍA

1. Adriana Collaguazo, Grace Cornejo, Joffre Pesantez, Galo Solis. 2009. *MANUAL DEL ETL DE PENTAHO PDI PENTAHO DATA INTEGRATION PREVIOUS KETTLE.* 2009.
2. Abad Grau, M. Mar, Hornos Barranco, Miguel J. y Montes Soldado, Rosana. *BASES DE DATOS Y DATA WAREHOUSE: HERRAMIENTAS ESTRATÉGICAS PARA LA EFICACIA COMERCIAL.*
3. Adamson, Christopher. 2006. *Mastering Data Warehouse Aggregates Solutions for Star Schema Performance.* Indianapolis : Wiley Publishing, Inc., 2006.
4. Alonso, Roberto Abajo. 2006. *DATA WAREHOUSE.* Madrid : s.n., 2006.
5. 2010. Apache Tomcat - Welcome! [En línea] The Apache Software Foundation, 2010. [Citado el: 29 de Octubre de 2010.] http://tomcat.apache.org/.
6. Cabrera, María Evelia Casales. 2009. *Data Warehouse (Almacenes de Datos).* 2009.
7. CNyS S.C Web Site. [En línea] [Citado el: 4 de Diciembre de 2010.] http://www.cnys.com.mx/databases/pdfs/sesion12_1.pdf.
8. Collaguazo, Adriana, Cornejo, Grace y Pesantez, Joffre. 2009. *MANUAL DEL ETL DE PENTAHO PDI PENTAHO DATA INTEGRATION PREVIOUS KETTLE.* 2009.
9. Gerardo, Clemente Garcia. 2008. *Un Sistema para el Mantenimiento de Almacenes de Datos.* Valencia : s.n., 2008.
10. Gerardo, Clemente García. 2008. *Un Sistema para el Mantenimiento de Almacenesde Datos.* Valencia : s.n., 2008.
11. Imohoff, Claudia, Galemmo, Nicholas y Geiger, Jonathan G. 2003. *Mastering Data Warehouse Desing Relational and Dimensional Techniques.* s.l. : Wiley Publishing, Inc, 2003.
12. Inmon, William H. 2005. *Building the Data Warehouse.* s.l. : Wiley Publishing, Ing., 2005.
13. 2010. Introducción a Pentaho Business Intelligence. [En línea] 2010. http://pentaho.almacen-datos.com/.
14. Keyla Ferreira, Jose Schmidt. 2009. *Sistemas de Información.* 2009.
15. Kimball, Ralph. 1996. *El Juego de Herramientas del Almacén de Datos.* s.l. : John Wiley & Sons, 1996.

16. kimball, Ralph y Caserta, Joe. 2004. *The Data Warehouse ETL Tolkit Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data.* s.l. : Wiley Publishing, Inc., 2004.

17. Kimball, Ralph y Joe, Caserta. 2004. *The Data Warehouse ETL Toolkit.* s.l. : Wiley Publishing, Inc, 2004.

18. Kimball, Ralph y Ross, Margy. 2002. *The Data Warehouse Toolkit.* s.l. : John Wiley & Sons, Inc., 2002.

19. Kimball, Ralph, y otros. *The Data Warehouse Lifecycle Toolkit.* s.l. : Wiley Publishing, Inc.

20. Límia Navarro, Alberto, y otros. 2008. *METODOLOGÍA PARA EL DESARROLLO DE SOLUCIONES DE ALMACENES DE DATOS E INTELIGENCIA DE NEGOCIO EN CENTALAD.* 2008.

21. Lucas-Torres Torrilla, Francisco José, y otros. Almacenes de Datos y Bases de Datos XML. [En línea]

22. Lujan-Mora, Sergio. 2004. *Diseño de Almacenes de Datos con UML.* 2004.

23. Main Page - Postgres SQL Wiki. [En línea] [Citado el: 16 de Noviembre de 2010.] http://wiki.postgresql.org/wiki/Main_Page.

24. Microsoft Download Center. [En línea] [Citado el: 16 de Noviembre de 2010.] http://download.microsoft.com/download/C/8/F/C8FF6EEE-C33F-428E-93B2-E8C0002C782D/binextel.pdf.

25. Open Source Business Intelligence - Open Source Reporting, ETL & Data Integration and Olap | Pentaho. [En línea] [Citado el: 3 de Noviembre de 2010.] http://www.pentaho.com/.

26. *Pentho Open Source business intelligence Creación de soluciones Pentaho.*

27. pgAdmin: PostgreSQL administration and management tools. [En línea] [Citado el: 20 de Noviembre de 2010.] http://www.pgadmin.org/.

28. Ponniah, Paulraj. 2001. *DATA WAREHOUSING FUNDAMENTALS.* New York : JOHN WILEY & SONS, INC., 2001.

29. Sumathi, S. y Sivanandam, S. N. 2006. *Introduction to Data Mining and its Applications.* New York : s.n., 2006. 978-3-540-34350-9/1860-9503.

30. Tandrón, Iván Maykel Cárdenas. 2008. *TÉCNICAS Y HERRAMIENTAS DE EXTRACCIÓN, TRANSFORMACIÓN Y CARGA DE DATOS APLICADAS A LA SEGURIDAD CIUDADANA.* Las Villas : s.n., 2008.

31. Trujillo, Juan C. 2006. *Diseño y explotación de Almacenes de Datos.* 2006.

32. UML, BPMN and Database Tool for Software Development. [En línea] [Citado el: 27 de Octubre de 2010.] http://www.visual-paradigm.com/.

33. Wang, John. 2006. *Encyclopedia of Data Warehousing and Mining.* 2006.

REFERENCIAS BIBLIOGRÁFICAS

1. Inmon, William H. 2005. *Building the Data Warehouse.* s.l. : Wiley Publishing, Ing., 2005.
2. Hurtado Torres,. 2005. *BASES DE DATOS Y DATA WAREHOUSE: HERRAMIENTAS ESTRATÉGICAS PARA LA EFICACIA COMERCIAL.* 2005.
3. Imohoff, Claudia, Galemmo, Nicholas y Geiger, Jonathan G. 2003. *Mastering Data Warehouse Desing Relational and Dimensional Techniques.* s.l. : Wiley Publishing, Inc, 2003.
4. Kimball, Ralph y Ross, Margy. 2002. *The Data Warehouse Toolkit.* s.l. : John Wiley & Sons, Inc., 2002.
5. Alonso, Roberto Abajo. 2006. *DATA WAREHOUSE.* Madrid : s.n., 2006.
6. Gerardo, Clemente Garcia. 2008. *Un Sistema para el Mantenimiento de Almacenes de Datos.* Valencia : s.n., 2008.
7. Wang, John. 2006. *Encyclopedia of Warehousing and Mining.* EUA: Idea Group reference, 2006.
8. Kimball, Ralph. 1996. *El Juego de Herramientas del Almacén de Datos.* s.l. : John Wiley & Sons, 1996.
9. Chuc-Durán, Diana Graciela. 2007. *Introducción a los Datawarehouses.* México : s.n., 2007.
10. Cabrera, María Evelia Casales. 2009. *Data Warehouse (Almacenes de Datos).* 2009.
11. PONNIAH, PAULRAJ. 2001. *DATA WAREHOUSING FUNDAMENTALS.* New York : John Wiley & Sons, Inc., 2001.
12. Lujan-Mora, Sergio. 2004. *Diseño de Almacenes de Datos con UML.* 2004.
13. Lucas-Torres Torrillas, Francisco José, y otros. Modelos Avanzados de Base de Datos Almacenes de Datos y Base de Datos XML. *Grupo Alarcos-Universidad de Castilla-La Mancha.* [En línea] [Citado el: 17 de Octubre de 2010.] http://alarcos.inf-cr.uclm.es/.
14. Orallo, Hernández, Quintana, Ramírez y César, Ferri Ramírez. 2004. *Introducción a la Minería de Datos.* Madrid : PEARSON EDUCACIÓN, S.A, 2004.

15. Kimball, Ralph y Joe, Caserta. 2004. *The Data Warehouse ETL Toolkit.* s.l. : Wiley Publishing, Inc, 2004.

16. Tandrón, Iván Maykel Cárdenas. 2008. *TÉCNICAS Y HERRAMIENTAS DE EXTRACCIÓN, TRANSFORMACIÓN Y CARGA DE DATOS APLICADAS A LA SEGURIDAD CIUDADANA.* Las Villas : s.n., 2008.

17. kimball, Ralph y Caserta, Joe. 2004. *The Data Warehouse ETL Tolkit Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data.* s.l. : Wiley Publishing, Inc., 2004.

18. 2010. Introducción a Pentaho Business Intelligence. [En línea] 2010. http://pentaho.almacen-datos.com/.

19. Rochnik, Nikolai. 2006. Oracle Warehouse Builder 10gR2: Transforming Data into Quality Information. U.S.A : Redwood Shores, 2006.

20. Trujillo, Juan y Manuel, C.Palomar. 2002. *Uso y Diseño de Bases de Datos Multidimensionales y Almacenes de Datos.* 2002.

21. Trujillo, Juna C. 2006. *Diseño y explotación de Almacenes de Datos.* 2006.

22. Chuc-Durán, Diana Graciela. 2007. *Introducción a los Datawarehouses.* México : s.n., 2007.

23. Límia Navarro, Alberto, y otros. 2008. *METODOLOGÍA PARA EL DESARROLLO DE SOLUCIONES DE ALMACENES DE DATOS E INTELIGENCIA DE NEGOCIO EN CENTALAD.* 2008.

24. UML, BPMN and Database Tool for Software Development. [En línea] [Citado el: 27 de Octubre de 2010.] http://www.visual-paradigm.com/.

25. 2008. *Pentaho Data Integration.* 2008.

26. Collaguazo, Adriana, Cornejo, Grace y Pesantez, Joffre. 2009. *MANUAL DEL ETL DE PENTAHO PDI PENTAHO DATA INTEGRATION PREVIOUS KETTLE.* 2009.

27. Open Source Business Intelligence - Open Source Reporting, ETL & Data Integration and Olap | Pentaho. [En línea] [Citado el: 3 de Noviembre de 2010.] http://www.pentaho.com/.

28. pgAdmin: PostgreSQL administration and management tools. [En línea] [Citado el: 20 de Noviembre de 2010.] http://www.pgadmin.org/.

29. 2010. Apache Tomcat - Welcome! [En línea] The Apache Software Foundation, 2010. [Citado el: 29 de Octubre de 2010.] http://tomcat.apache.org/.

30. Main Page - Postgres SQL Wiki. [En línea] [Citado el: 16 de Noviembre de 2010.] http://wiki.postgresql.org/wiki/Main_Page.

GLOSARIO DE TÉRMINOS

Área de análisis: no es más que la agrupación de información según su propósito, aunque el criterio depende de las necesidades de la institución o empresa donde se aplica el sistema. Permite restringir el número de usuarios que acceden a los datos.

Cubo: colección de dimensiones y medidas en un área temática particular.

Libro de trabajo: estructura organizativa que agrupa las vistas de. Puede ser creado teniendo en cuenta criterio que permitan organizar la información: Emisor de los reportes, receptor del reporte, contenido, entre otros.

MDX: es el acrónimo de MultiDimensional eXpressions, es un lenguaje de consulta para bases de datos multidimensionales sobre cubos OLAP.

MOLAP: Procesamiento Analítico Híbrido.

OLAP: Procesamiento Analítico en Línea.

Open Source: el *software Open Source* se define por la licencia que lo acompaña, que garantiza a cualquier persona el derecho de usar, modificar y redistribuir el código libremente

ROLAP: Procesamiento Analítico Relacional.

UML: Lenguaje Unificado de Modelado (*UML*, por sus siglas en inglés, *Unified Modeling Language*) es el lenguaje de modelado de sistemas de *software* más conocido y utilizado en la actualidad.

Vistas de análisis: muestran la información almacenada en la base de datos de acuerdo a las necesidades del usuario.

XML: *extensible markup language* ó lenguaje de anotación extensible.