Universidad de las Ciencias Informáticas

Facultad 6

Título: "Sistema de Información de Gobierno. Mercado de datos Tecnologías de la información"

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

Autora: *Laritza Rodríguez La Rosa*

Tutores: *Ing. Yamila Mateu Romero*

Ing. Yosvany Arrastia Machin

La Habana, Junio de 2011

"Año 53 de la Revolución"

"No todo lo bello es fácil, quien aspira lo perfecto, debe escoger lo difícil"

Che

Declaración de autoría

Declaro ser la autora de la presente tesis y reconozco a la Universidad de las Ciencias Informáticas los derechos patrimoniales de la misma, con carácter exclusivo.

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

_____

Laritza Rodríguez La Rosa

Autor

_____

Ing. Yamila Mateu Romero

Tutor

_____

Ing. Yosvany Arrastia Machin

Tutor

Datos de contacto

Ing. Yamila Mateu Romero.

Centro de Tecnología y Gestión de Datos, Universidad de las Ciencias Informáticas, La Habana, Cuba.

ymateu@uci.cu

Ing. Yosvany Arrastia Machin

Centro de Tecnología y Gestión de Datos, Universidad de las Ciencias Informáticas, La Habana, Cuba.

yarrastia@uci.cu

Agradecimientos

A mis padres por ser la razón de todo mi esfuerzo, por guiarme, por ser mi estrella, por darlo todo por mí, por no dudar ni un segundo que lo lograría, por hacer de mí una persona sencilla. Sé que están muy orgullosos de mí. Todo lo que soy es por ustedes. Los Amo.

A mi papá por darme fuerzas cuando lo necesité, por decirme “tú si puedes”.

A mi abuela (mamá) y mi abuelo (papá), por darme tanto cariño y amor, por ayudar a mis padres a darme este tamaño, por ayudar a mi educación y por aconsejarme tanto. Los quiero muchísimo y no quisiera que se apartaran nunca de mí.

A mi tío Luis, por quererme tanto, por confiar en mí, por creerme capaz de sobrepasar los obstáculos y por ser tan cariñoso. Te quiere mucho, tu sobri.

A mi prima Yamelis por ser amiga y hermana, gracias por tus consejos, por quererme tanto y por ser siempre esa personita tan bella.

A mis sobrinos, aunque no he tenido mucho tiempo de compartir con ustedes, me han llenado de alegría. Dariel sé que eres muy inteligente y puedes lograr cosas mejores, por eso hoy te exhorto a que continúes así, a que escuches a mamá y a papá cuando te dan un consejo, que continúes siendo el primero en todo, que sigas así de cariñoso, que quieras mucho a tus papis y a tus abuelos que te adoran.

A mi amiga del IPI Delvis; sabía que lo lograrías, ya eres una Ingeniera. Gracias por no olvidarme nunca.

A todos los de mi antiguo grupo 6104 y 6204.

A mis amigas y amigos de siempre: Yudelkis, Lisandra, Aleida, Sailín, Yulier, Alejo, que aunque nos hemos separado la amistad ha perdurado.

A Alejo por no decir nunca no cuando necesitaba de él, gracias por todo.

A los del antiguo grupo 6303, por ser tan divertidos y sobre todo por ser buenos compañeros.

A Daynelis y Patricia mis compañeras de estudio de la PNP porque siempre estuvimos muy unidas hasta el final, por ser perseverantes en el estudio y por ser excelentes compañeras de cuarto.

A Patry por ser una gran amiga, por escucharme, porque cuando tuvimos una caída fuimos capaces de levantarnos y salir adelante. A su tía Delkis por acogerme en su casa, por preocuparse por mí.

A mi compañera de tesis, que mientras estuvo aquí su apoyo fue incondicional.

A Yoendy, Leandro, Ariel, Yani, Susel, David, Leonel y Yuslier por brindarme su ayuda. Gracias a ti Yoendy por ser tan buena persona.

A mi otra compañera de tesis: Laidy por nunca dejarme a tras cuando necesitaba ayuda, por ser tan buena persona, te deseo de todo corazón que tengas muchos éxitos en la vida.

A mis compañeras de cuarto con las que compartí un feliz quinto año.

Al profe Tellez por darme la mano cuando más lo necesité.

A los todos los maestros y profesores que han puesto un granito en la formación de la persona que soy.

A los tutores, gracias por su apoyo.

Al tribunal y al oponente gracias por sus críticas constructivas.

Dedicatoria

A mis padres y a mis abuelos maternos por ser las personas más importante de mi vida.

Resumen

La Oficina Nacional de Estadística de Información (ONEI) es la encargada de garantizar estadísticas de alta calidad al Sistema Estadístico Nacional. Para una mejor organización de la información esta se encuentra dividida en diferentes departamentos. Uno de ellos es el de Tecnologías de la información, encargado de recoger los datos referentes a insumos informáticos y acceso a internet. Para este departamento controlar todo el proceso de organización de los datos es complejo, pues los mismos se encuentran en formato excel y dbf y sólo pueden ser consultados por un especialista informático y de la información. El presente trabajo tiene como objetivo desarrollar un mercado de datos para el área de Tecnologías de la información del Sistema de Información de Gobierno. Para el desarrollo de la solución se utilizaron algunas herramientas de la Suite de Pentaho, además del PostgreSQL como Sistema Gestor de Base de datos. Como resultado se obtuvo el mercado de datos poblado y con una capa de inteligencia de negocio, permitiéndole a los especialistas un mejor análisis de la información.

Palabras Clave: Almacén de datos, Mercado de datos, Tecnologías de la información, Inteligencia de Negocio.

ÍNDICE

Agradecimientos I

Dedicatoria III

Resumen IV

Introducción 1

Capítulo 1: Fundamento Teórico de los almacenes de datos 5

1.1 Almacenes de datos (Data Warehouse) 5

1.1.1 Mercado de datos (Data Mart) 7

1.2 Tendencias actuales de los almacenes de datos 7

1.2.1 Sistemas existentes a nivel mundial 7

1.2.2 Sistemas existentes en Cuba 8

1.3 Integración de Datos 9

1.4 Modelos de diseño de datos 9

1.4.1 Tipos de modelamiento de un almacén de datos 10

1.5 Inteligencia de negocios 11

1.6 Modos de almacenamiento de datos 12

1.7 Metodología de desarrollo 13

1.8 Herramientas de Modelado 14

1.8.1 Visual Paradigm 6.4 15

1.9 Sistema Gestor de Base de Datos 15

1.9.1 PostgreSQL 8.4 15

1.10 Herramienta de Perfilado de Datos 16

1.10.1 DataCleaner 1.5.3 16

1.11 Herramienta de ETL 16

1.11.1 Pentaho Data Integration 4.0.1 16

1.12 Herramientas para la Inteligencia de Negocios 17
1.12.1 Pentaho BI Server 3.6.0 17
1.12.2 Pentaho Analysis Services 3.0.4 17
1.12.3 Pentaho Schema Workbench 3.2.0 18
1.12.4 Apache Tomcat 5.5 18
1.13 Conclusiones del capítulo 18
Capítulo 2: Análisis y Diseño del mercado de datos Tecnologías de la información 20
2.1 Análisis 20
2.1.1 Definición del Negocio 20
2.1.2 Tema de análisis 20
2.1.3 Necesidades de información 21
2.1.4 Requisitos de información 21
2.1.5 Requisitos multidimensionales 25
2.1.6 Requisitos funcionales 26
2.1.7 Requisitos no funcionales 27
2.1.8 Reglas del Negocio 28
2.1.9 Especificación de Casos de Uso 30
2.2 Diseño 35
2.2.1 Dimensiones identificadas 35
2.2.2 Hechos identificados 36
2.2.3 Medidas identificadas 37
2.2.4 Matriz Bus 38
2.2.5 Modelo de datos 38
2.2.6 Roles y permisos 40
2.3 Conclusiones del capítulo 40

Capítulo 3: Implementación del mercado de datos Tecnologías de la información 41
3.1 Modelo de datos físico 41
3.1.1 Esquemas y tablas 41
3.2 Arquitectura de integración 42
3.2 Proceso de integración de datos 42
3.2.1 Perfilado de datos 42
3.2.2 Extracción, transformación y carga de los datos 43
3.3 Trabajo para organizar el orden de la carga 43
3.4 Proceso de Inteligencia de Negocios 44
3.4.1 Implementación de los cubos OLAP 44
3.4.2 Arquitectura de información 45
3.4.3 Reportes Candidatos 45
3.5 Conclusiones del capítulo 47
Capítulo 4: Validación del mercado de datos Tecnologías de la información 48
4.1 Pruebas 48
4.2 Herramientas para aplicar las pruebas 48
4.2.1 Listas de chequeo 48
4.2.2 Casos de prueba 51
4.3 Pruebas aplicadas al mercado de datos 53
4.3.1 Prueba integración 53
4.3.2 Prueba de aceptación 53
4.4 Conclusiones del capítulo 54
Conclusiones 55
Recomendaciones 56
Referencias Bibliográficas 57

Bibliografía 58

Glosario de Términos 60

ÍNDICE DE FIGURAS

Figura 1: Esquema en Estrella. 10

Figura 2: Esquema Copo de Nieve. 11

Figura 3: Esquema Constelación. 11

Figura 4: Diagrama de Casos de Uso del sistema. 31

Figura 5: Modelo de datos. 39

Figura 6: Arquitectura de integración. 42

Figura 7: Carga del hecho Ingresos de las TI. 43

Figura 8: Trabajo del mercado de datos. 44

Figura 9: Cubo Indicadores generales. 45

Figura 10: Arquitectura de información. 45

ÍNDICE DE TABLA

Tabla 1: Casos de Uso de información ¨Analizar los Ingresos de las TI¨. 34

Tabla 2: Descripción del caso de uso funcional ¨Autenticar Usuario¨. 35

Tabla 3: Matriz Bus. 38

Tabla 4: Roles identificados y permisos. 40

Tabla 5: Esquemas y tablas identificados. 41

Tabla 6: Lista de chequeo aplicada al mercado de datos Tecnologías de la información. 51

Tabla 7: Caso de prueba "Ingresos de las Tecnologías de la información". 53

Introducción

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

Debido al auge que ha tenido la implantación y utilización de las TIC en todo el mundo, se presentan como una necesidad para el desarrollo económico y social de cualquier país. Esto ha traído como consecuencia que las empresas informáticas enfrenten cada día un reto mayor para brindar una respuesta rápida y con calidad a sus clientes, que cada vez adquieren más experiencia y se vuelven más exigentes en cuanto a la confiabilidad que deben brindar los productos de software en la actualidad.

En los últimos años, en Cuba se ha emprendido el reto de la informatización de la sociedad, este proyecto se ha realizado de manera acelerada, alcanzando resultados satisfactorios en áreas tan esenciales como la Educación, la Salud y la Investigación.

Como resultado de esto surge en el 2002 la Universidad de las Ciencias Informáticas (UCI), que se plantea entre sus objetivos formar excelentes profesionales altamente comprometidos con la Revolución y producir software y servicios informáticos, a partir de la vinculación estudio-trabajo como modelo de formación.

Uno de los Centros de Investigación de la UCI que pertenece a la Facultad 6 es el Centro de Tecnología de Gestión de Datos (DATEC). Actualmente en la línea Almacenes de Datos se realiza el proyecto ¨Sistema de Información de Gobierno¨ que tiene entre sus objetivos la integración de los datos recopilados a través del Sistema Estadístico Nacional en un almacén de datos, el cual se desglosa en diferentes áreas de análisis que se hacen corresponder con cada uno de los departamentos de la ONEI.

El Sistema de Información de Gobierno (SIGOB) está formado por un conjunto de organismos y entidades que interactúan entre sí para procesar información y distribuirla de manera adecuada en función de los objetivos del gobierno. La facultada de centralizar, emitir, organizar y aprobar las estadísticas destinadas a satisfacer los requerimientos informativos y los compromisos de los órganos de dirección del gobierno en los territorios es la ONEI.

La ONEI es la entidad encargada de garantizar la producción de estadísticas de calidad a través del Sistema Estadístico Nacional, ejerciendo un adecuado control de la captación de las cifras económicas y sociales, así como su correcta difusión de acuerdo con las necesidades de la economía y las demás necesidades del país en información estadística. Para ello resulta imprescindible tener un alto nivel de organización en los datos que maneja y difunde para contribuir a la correcta toma de decisiones.

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 ONEI controlar todo el proceso de organización de los datos es complejo, debido a que la información digital se encuentra en formato .dbf y solo puede ser consultada por un especialista informático y de la información. Se necesita conocimiento del negocio para entender los ficheros, los cuales son generados anualmente y deben ser procesados para obtener información. Actualmente la ONEI no cuenta con un sistema informático que le permita analizar la información, lo que trae consigo que haya 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.

Luego de un análisis en el área de Tecnologías de la información en el Sistema de Información de Gobierno referente a los mecanismos de procesamiento y almacenamiento de la información se identifica como **problema científico**: ¿Cómo contribuir a la toma de decisiones en el área Tecnologías de la información del Sistema de Información de Gobierno?

Este problema se enmarca en el **objeto de estudio**: Almacenes de datos e Inteligencia de Negocios. El objeto de estudio delimita el **campo de acción**: Mercado de datos y capa de visualización para el área Tecnologías de la información del Sistema de Información de Gobierno.

Para dar solución al problema científico se ha definido como **objetivo general** de la investigación: Desarrollar el mercado de datos Tecnologías de la información del Sistema de Información de Gobierno que contribuya a la toma de decisiones.

Del objetivo general se derivan los siguientes **objetivos específicos**:

- Realizar el análisis y diseño del mercado de datos del área Tecnologías de la información.
- Implementar el mercado de datos del área Tecnologías de la información.
- Implementar la capa de visualización de datos.
- Validar el mercado de datos del área Tecnologías de la información.

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

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

El presente trabajo está desglosado en 4 capítulos:

En el capítulo 1 ¨***Fundamento Teórico de los almacenes de datos***¨ se definen los temas relacionados con el desarrollo de los almacenes de datos, mercados de datos e inteligencia de negocio, se describen los conceptos fundamentales que serán tratados a lo largo de la investigación, así como la selección de la metodología y las herramientas que se van a utilizar.

En el capítulo 2 ¨***Análisis y diseño del mercado de datos Tecnologías de la información***¨ se definen los temas de análisis y se realiza el diseño del almacén, del subsistema de integración y de visualización, teniendo como principales resultados la descripción de los casos de uso, el modelo de datos, los reportes candidatos, entre otros.

En el capítulo 3 ¨***Implementación del mercado de datos Tecnologías de la información***¨ se implementa el modelo de datos, los cubos OLAP, los reportes candidatos, los subsistemas de integración y de visualización.

En el capítulo 4 ¨***Validación del mercado de datos Tecnologías de la información***¨ se valida la solución mediante las listas de chequeo, los casos de pruebas y la carta de aceptación del cliente.

Capítulo 1: Fundamento Teórico de los almacenes de datos

Introducción

Durante el desarrollo de este capítulo se hace un estudio sobre los almacenes de datos y los mercados de datos, así como sus características y los elementos que lo integran. Se realiza un análisis de las metodologías y herramientas existentes sobre los almacenes de datos.

1.1 Almacenes de datos (Data Warehouse)

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, convirtiéndose en una herramienta esencial para el control y el manejo de las operaciones del comercio, 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 almacenes de datos.

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

Un almacén de datos es una gran colección de datos históricos, que recoge información de múltiples sistemas fuentes u operacionales dispersos, y cuya actividad se centra en la toma de decisiones. (VELASCO, 2010)

Un almacén de datos es una colección de datos orientada a temas o materias, integrada, variable en el tiempo y no volátil que será utilizada fundamentalmente en el proceso de toma de decisiones. (INMON, 1992)

Ralph Kimball[1] define un Data Warehouse como: “una copia de las transacciones de datos específicamente estructurada para la consulta y el análisis.” (LÖFBERG and MOLIN, 2005)

Teniendo en cuenta las definiciones anteriores se puede concluir que un almacén de datos es una colección de datos orientados a temas, integrados, no volátiles y variante en el tiempo, que apoya la toma de decisiones. Contiene información no solo de bases de datos relacionales, sino de otras fuentes relacionadas con la actividad de la empresa y cuya finalidad es la obtención de información estructurada y útil para la toma de decisiones.

[1] Ralph Kimball, autor ampliamente publicado sobre el tema de almacén de datos y de inteligencia de negocios. Es considerado el ¨Gurú del Data Warehousing¨.

Entre las características de los almacenes de datos se pueden mencionar las siguientes:(EVELIA, 2009)

- ✓ ***Organizado en torno a temas***: la información se clasifica en base a los aspectos que son de interés para la empresa.
- ✓ ***Integrado***: integra datos recogidos de diferentes sistemas operacionales de la organización.
- ✓ ***Dependiente del tiempo***: los datos son relativos a un período de tiempo y deben ser incrementados periódicamente.
- ✓ ***No volátil***: el almacén de datos solo permite cargar nuevos datos y acceder a los ya almacenados; pero no permite ni borrar ni modificar los datos.

El almacén de datos consta de 2 elementos principales: (FERREIRA and SCHMIDT, 2009)

Tablas de dimensiones: describen los objetos relevantes para la organización. Cada tupla de la tabla se compone del identificador de un objeto y de sus atributos descriptores. Los datos de las tablas de dimensiones son generalmente de tipo alfanumérico.

Tablas de hechos: contienen datos sobre las actividades básicas de la organización. Cada tupla de la tabla se compone de los datos observables de la actividad y de las referencias a las dimensiones que los caracterizan. Los datos de las tablas de hechos son generalmente de tipo numérico.

Los principales aportes que brinda un almacén de datos son:(GALICIA, 2007)

- ✓ Proporciona una herramienta para la toma de decisiones en cualquier área funcional, basándose en información integrada y global del negocio.
- ✓ Facilita la aplicación de técnicas estadísticas de análisis para encontrar relaciones ocultas entre los datos del almacén; obteniendo un valor añadido para el negocio de dicha información.
- ✓ Proporciona la capacidad de aprender de los datos del pasado y de predecir situaciones futuras en diversos escenarios.

Estos sistemas poseen las siguientes ventajas:(ORALLO, 2003)

- ✓ Rentabilidad de las inversiones realizadas para su creación.
- ✓ Aumento de la competitividad en el mercado.
- ✓ Aumento de la productividad de los técnicos de dirección.

Algunas de las desventajas de estos almacenes son:(ORALLO, 2003)

- ✓ Incremento continuo de los requisitos de los usuarios.
- ✓ Privacidad de los datos.

1.1.1 Mercado de datos (Data Mart)

Los mercados de datos son generalmente subconjuntos del almacén de datos, diseñados para satisfacer las necesidades específicas de grupos comunes de usuarios. (TORRES, 2007)

Es una base de datos departamental, especializada en el almacenamiento de los datos de un área de negocio específica. Dispone de una estructura óptima de datos para analizar la información al detalle desde todas las perspectivas que afecten a los procesos de dicho departamento. Un mercado de datos puede ser alimentado desde los datos de un almacén de datos, o integrar por sí mismo un compendio de distintas fuentes de información. (DÍAZ, 2009)

En general, un mercado de datos es un subconjunto del almacén de datos enfocado a un departamento o área específica. Permite examinar los datos desde diferentes perspectivas, logrando así un mejor control de la información que se está abarcando.

Dentro de las ventajas de aplicar un mercado de datos a un negocio, se han seleccionado las siguientes:(BERNABEU, 2009)

- ✓ Son simples de implementar.
- ✓ Conllevan poco tiempo de construcción y puesta en marcha.
- ✓ Permiten manejar información confidencial.
- ✓ Reflejan rápidamente sus beneficios y cualidades.

1.2 Tendencias actuales de los almacenes de datos

La información estadística es de gran importancia para el desarrollo de todo país, pues permite el análisis de datos acumulados y contribuye a la toma de decisiones de los altos directivos. Por ello resulta imprescindible tener un elevado nivel de organización de datos que se manejan y difunden, y esto puede lograrse aplicando técnicas de inteligencia de negocio.

1.2.1 Sistemas existentes a nivel mundial

- ✓ El 11 de febrero del 2010 se puso en línea el Almacén Central de Datos de la ONE para el Sistema Estadístico Nacional de la República Dominicana. Este almacén de datos sirve de

apoyo a la recolección, procesamiento, análisis y difusión de la información estadística del país. Además ofrecerá servicios de información basándose en modelos analíticos y de minería de datos. Actualmente el almacén de datos recoge la información del Censo Nacional de Población y Vivienda del 2002, del Directorio de Establecimientos Económicos 2009, de los nacimientos, defunciones, matrimonios y divorcios registrados entre los años 2001-2008, de las encuestas de ingresos y gastos de los hogares del 2007, entre otras.

- ✓ El almacén de datos para el Análisis y Difusión de la Información Estadística del Turismo en España o DATATUR se encarga de la recolección de todos los datos concernientes al área del turismo español. Sus principales objetivos son conservar y mantener al día los datos estadísticos de esta área, facilitar la actualización de manera ágil y sin errores, cubrir las demandas externas de información, entre otros aspectos.
- ✓ El INEGI (Instituto Nacional de Estadísticas e Informática) de México tiene su información almacenada en un almacén de datos cuyo propósito es integrar dicha información en un repositorio para su consulta y análisis, para permitir a sus usuarios tomar decisiones en el contexto del Sistema Nacional de Estadísticas e Información Geográfica.

1.2.2 Sistemas 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 encuentran:

- ✓ El almacén de datos comercial de la corporación CIMEX, que se dedica fundamentalmente a la exportación e importación de mercancías. Forman parte de ella un conjunto de empresas que se encuentran enfocadas en diversos negocios, aquí se puede citar la red de comercio minorista y la dirección de logística, esta última dedicada al comercio mayorista. El mismo centra su atención en la gestión de inventario, permitiendo una gestión de compra–venta eficiente, con una finalidad de disminuir los costos, sin afectar al cliente, permitiendo prestaciones eficientes y con la calidad requerida, aumentando las utilidades de las mismas.
- ✓ En el XIII Concurso Nacional de Computación y en la Feria de Informática del 2002 se presentó un almacén de datos para CUBACEL desarrollado sobre la plataforma Oracle, con grandes resultados obtenidos a partir de su implantación.

1.3 Integración de Datos

El término “integración de datos” se entiende normalmente como el proceso de combinación de datos procedentes de distintas fuentes heterogéneas con el objetivo de proporcionar una visión única y global de los datos combinados.

Extracción, Transformación y Carga de Datos (ETL): es la tecnología enfocada a la integración de datos, tanto por lote, como a tiempo real hacia almacenes de datos, logrando alto grado de transformaciones para la migración y consolidación de datos. Este sincroniza datos desde diversas aplicaciones e involucra procesos de manipulación de datos.

Actividades generales que realizan dentro del proceso de ETL:

- ✓ Extracción: proceso consistente en la adquisición de datos desde una o varias fuentes.
- ✓ Transformación: proceso de acondicionamiento de la forma y/o contenido de los datos para encajarlos en la estructura del almacén de datos destino.
- ✓ Carga: proceso de carga de los datos en el almacén destino.

Características más significativas de este proceso:(KIMBALL and CASERTA, 2004)

- ✓ Es un mecanismo de carga muy eficiente orientado a los almacenes de datos.
- ✓ 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 almacenes de datos.

1.4 Modelos de diseño de datos

Modelo relacional: el diagrama de entidad-relación es un lenguaje para realizar el modelado de los datos de un sistema de información, se basa en la separación de los datos en entidades para formar parte del diseño físico. (CASTILLO, 2008)

Modelo dimensional: es uno de los más reconocidos en el mundo de los almacenes de datos, contiene la misma información que un modelo de entidad-relación, pero la forma de empaquetar los datos en un formato simétrico tiene como objetivo garantizar una ejecución eficiente y rápida de las consultas, así como lograr una mayor comprensión del usuario. Dicho modelo separa sus datos en dos

grandes tipos: las medidas, que generalmente son valores numéricos que se almacenan en las tablas de hechos, las cuales son las tablas primarias de este modelo; y las descripciones de los entornos, que son textuales y se almacenan en las tablas de dimensiones.(VERÁSTEGUI, 2007)

La estructura básica de un almacén de datos para el modelo multidimensional está definida por dos elementos: esquemas y tablas.

Existen dos tipos básicos de tablas en el modelo multidimensional:
Tablas de hechos: es la tabla central en un esquema dimensional. Contienen los valores de las medidas de negocios.

Tablas dimensionales: contienen el detalle de los valores que se encuentran asociados a la tabla de hechos.

1.4.1 Tipos de modelamiento de un almacén de datos

Esquema Estrella: consta de una tabla de hechos central y de varias tablas de dimensiones relacionadas a esta, a través de sus respectivas claves. En la siguiente figura se puede apreciar un esquema en estrella estándar:

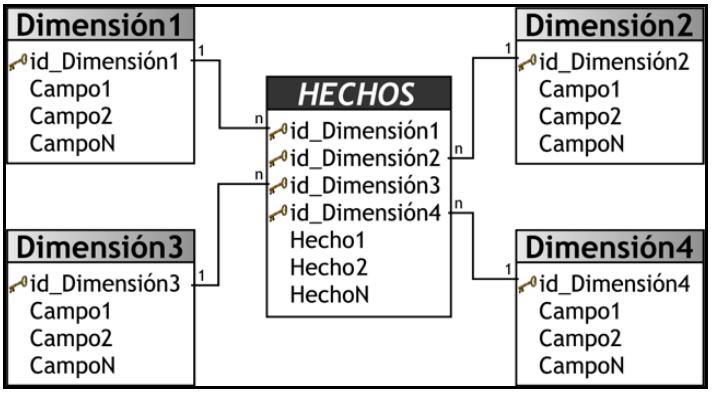


Figura 1: Esquema en Estrella.

Esquema Copo de Nieve: existe una tabla de hechos central que está relacionada con una o más tablas de dimensiones, quienes a su vez pueden estar relacionadas o no con una o más tablas de dimensiones.

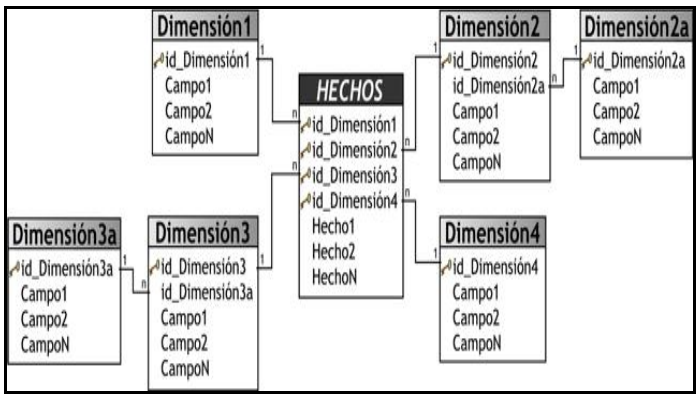


Figura 2: Esquema Copo de Nieve.

Esquema Constelación: está compuesto por una serie de esquemas de estrella, es decir, una tabla de hechos central con otras auxiliares y sus respectivas tablas de dimensiones. Por sus características es la seleccionada para la realización del mercado de datos.

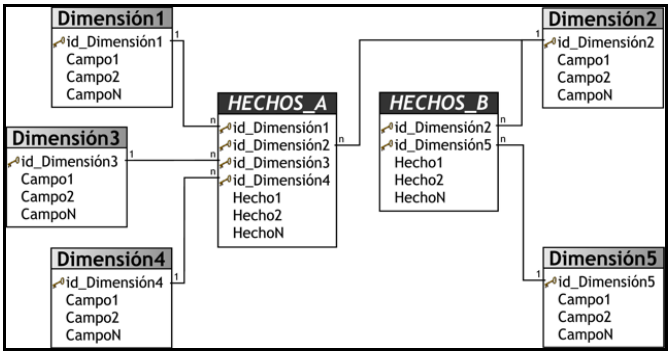


Figura 3: Esquema Constelación.

1.5 Inteligencia de negocios

La Inteligencia de Negocios (BI) se utiliza con el fin de analizar los datos acumulados en una empresa y extraer una cierta inteligencia o conocimiento de ellos. Se puede definir como el conjunto de metodologías, aplicaciones y tecnologías que permiten reunir, limpiar y transformar datos de los sistemas transaccionales en información estructurada, para su análisis. Su característica principal radica en estar dirigida generalmente a usuarios no expertos en el manejo de computadoras.

La necesidad principal que cubre la BI es permitirles a los usuarios el poder acceder a la información, extraerlas a través de consultas SQL sencillas, para luego dar formato y presentación a la misma en la cual van a estar basando su toma de decisiones.

Beneficios

- Genera reportes
- Apoya a la toma de decisiones
- Aprovecha datos históricos
- Mejora el servicio al cliente

1.6 Modos de almacenamiento de datos

OLAP (Procesamiento Analítico en Línea[2]) 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.

Las principales características de OLAP son:

- ✓ ***Rápido:*** el sistema debe ser capaz de responder de una forma rápida y ágil a la información que le sea solicitada por el usuario.
- ✓ ***Análisis:*** significa que el sistema debe poder reflejar cualquier lógica del negocio para poder responder a las preguntas específicas y necesidades empresariales.
- ✓ ***Compartido***: el sistema deberá proporcionar herramientas que garanticen la confidencialidad de los datos y la seguridad de acceso por perfiles de los usuarios.
- ✓ ***Multidimensional:*** la herramienta deberá proporcionar soporte a cada una de las múltiples jerarquías que puedan existir dentro de la organización de información.
- ✓ ***Información:*** son todos los datos e información derivada de este proceso de análisis, la cual permitirá la toma de decisiones.

OLAP es un componente clave en las soluciones de Inteligencia de Negocios. Permite agilizar las consultas de grandes cantidades de datos, para ello utiliza estructuras multidimensionales o cubos OLAP que contienen datos resumidos de grandes bases de datos.

Tipos de sistemas OLAP.

ROLAP: organización física que se implementa sobre tecnología relacional. Para proporcionar los análisis OLAP esta arquitectura accede directamente a los datos almacenados en un almacén. Los datos son acumulados en filas y columnas de forma relacional, y se les presentan a los usuarios en

[2] Del inglés Online Analytical Processing, conocido por sus siglas OLAP.

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 y gratuitos, y es capaz de manejar grandes volúmenes de datos.

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.

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.

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 Tecnologías de la información.

1.7 Metodología de desarrollo

Una metodología es el conjunto de métodos o procedimientos de investigación que se siguen para alcanzar una gama de objetivos en una ciencia y rigen una investigación científica o una exposición doctrinal.

Las metodologías de desarrollo para soluciones de inteligencia de negocios están marcadas por dos tendencias muy bien definidas, la Metodología Kimball, en honor de su creador Ralph Kimball y la Metodología de Inmon dada a su creador Bill H. Inmon, reconocido como el padre del data warehosing.

Kimball basa su desarrollo dividiendo el mundo de la Inteligencia de Negocios entre los hechos y las dimensiones. Tiene una gran cantidad de documentación y se puede encontrar una respuesta a casi todas las preguntas que se puedan tener. Este método es iterativo, en el cual el almacén de datos es construido pieza por pieza, mediante el desarrollo de todos los mercados de datos. Esta tendencia es conocida como “Bottom-Up”.

La metodología de Inmon plantea la creación de un repositorio de datos corporativo como fuente de información consolidada, persistente, histórica y de calidad. Se basa en un enfoque descendente el cual propone construir primero el almacén de datos para luego a partir de este, los mercados de datos, por lo que es necesario la creación de un repositorio de datos corporativo como fuente de información consolidada. Esta tendencia es conocida como “Top-Down”.

DATEC propone una modelo de metodología para este tipo de soluciones, adaptada a su proceso de desarrollo, el cual está basado en Líneas de Productos de Software, y a los lineamientos de calidad exigidos por la Universidad de Ciencias Informáticas.

La misma cubre todas las fases por las que pasa la construcción de un almacén de datos, desde el levantamiento de información inicial hasta la capa de visualización. Es una metodología mixta que reúne elementos de varias metodologías de desarrollo de proyectos de integración de datos, toma como base la Metodología de Kimball. En una primera fase contempla el levantamiento de información a nivel de negocio para identificar los posibles indicadores y aspectos a medir en los análisis, que luego de algunas transformaciones se convierten en los requerimientos de información de entrada y de salida para la solución de integración.

De forma paralela a esta actividad se lleva a cabo un estudio de las fuentes de datos que soportan los datos a cargar. Finalizadas estas dos tareas, se corrobora que la información levantada sobre las necesidades de los clientes esté realmente almacenada en las fuentes correspondientes, para posteriormente, teniendo los requerimientos informativos correctamente definidos, proceder a diseñar la solución de almacén de datos. Una vez diseñada la estructura del almacén, se realiza la carga de los datos desde las fuentes y posteriormente se implementan los requerimientos de Inteligencia de Negocios identificados en el levantamiento de información inicial. Las actividades y artefactos de la solución son realizados por cuatro grupos que conforman la línea, especializados en componentes específicos de la solución. (DATEC, 2008)

1.8 Herramientas de Modelado

Hoy día, muchas empresas se han extendido a la adquisición de herramientas de Ingeniería de Sistemas Asistida por Computadora (CASE), con el fin de automatizar los aspectos clave de todo el proceso de desarrollo de un sistema, desde el principio hasta el final e incrementar su posición en el mercado competitivo.

La Ingeniería de Sistemas Asistida por Computadora es la aplicación de tecnología informática a las actividades, las técnicas y las metodologías propias de desarrollo, su objetivo es acelerar el proceso para el que han sido diseñadas, en el caso de CASE para automatizar o apoyar una o más fases del ciclo de vida del desarrollo de sistemas.

1.8.1 Visual Paradigm 6.4

Visual Paradigm es una herramienta CASE que utiliza UML[3] como lenguaje de modelado, proporciona a los desarrolladores una plataforma que les permite diseñar un producto con calidad de forma rápida.

Entre sus características están:

- ✓ Diseño centrado en casos de uso 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.

1.9 Sistema Gestor de Base de Datos

Un Sistema de Gestión de Bases de Datos es un tipo de software muy específico que permite entre otras cosas, la administración de todos los datos almacenados en una base de datos y el acceso a los mismos por medio de la gestión de usuarios y diversos permisos de acceso. De acuerdo con estos permisos, permite a cada usuario realizar tareas como definir y crear la base de datos; agregar, modificar y mantener los datos almacenados y llevar a cabo el control de acceso.

1.9.1 PostgreSQL 8.4

Es un sistema de gestión de base de datos relacional orientado a objetos y libre, publicado bajo la licencia BSD[4].

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

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

[3] UML, del ingles Unified Modeling Language

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

Ventajas

- ✓ Es capaz de ajustarse al número de CPU y a la cantidad de memoria que posee el sistema de forma óptima, haciéndole capaz de soportar una mayor cantidad de peticiones simultáneas de manera correcta.
- ✓ Tiene la capacidad de comprobar la integridad referencial, así como también la de almacenar procedimientos en la propia base de datos.

1.10 Herramienta de Perfilado de Datos

El perfilado de los datos forma parte del ciclo de la integración de los datos, y permite localizar, medir, monitorizar y reportar los problemas de calidad de datos de las fuentes de origen.

1.10.1 DataCleaner 1.5.3

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

Principales características de DataCleaner:

- ✓ Los perfiles de datos se utilizan para calcular y analizar diversas medidas importantes basadas en los valores de los datos.
- ✓ Validación de datos: el validador le dará un resultado que puede ser interpretado como bueno o malo.

1.11 Herramienta de ETL

1.11.1 Pentaho Data Integration 4.0.1

Pentaho Data Integration, también conocido como Kettle, es el componente de Pentaho responsable de la extracción, transformación y carga (ETL) de procesos. Este se encarga de abrir, limpiar e integrar la información disponible en diferentes fuentes de datos y ponerla en manos del usuario.

A continuación se muestran las características de esta herramienta:

- ✓ **Plataforma**: Windows, Unix y Linux
- ✓ **Código fuente**: el código fuente está disponible.

- ✓ **Soporte**: existe un foro, un buscador de problemas y la comunidad Pentaho, con profundos artículos técnicos que son mejores que algunos de los vendedores oficiales de productos para ETL.

Kettle admite una amplia gama de formatos de entrada y salida, incluyendo archivos de texto, hojas de datos, archivos XML, facilita la reutilización de componentes de transformación, colaboración y administración de modelos. Provee un ambiente de diseño intuitivo y soporta conexión a bases de datos PostgreSQL.

Desventajas del Kettle:

- ✓ No cuenta con un componente de calidad de datos.
- ✓ Las búsquedas de mayores volúmenes de información dificultan el rendimiento de ETL.

1.12 Herramientas para la Inteligencia de Negocios

1.12.1 Pentaho BI Server 3.6.0

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 también incluye un motor de solución que integra reportes, análisis, tableros de comandos y componentes de minería de datos. Está diseñado para integrarse fácilmente en cualquier proceso de negocio.

Algunas de sus ventajas son:

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

1.12.2 Pentaho Analysis Services 3.0.4

Pentaho Analysis Services, también conocido como Mondrian, 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. Permite a los usuarios analizar grandes cantidades de datos en tiempo real.

Algunas de sus ventajas son:

- ✓ Da la posibilidad de utilizar siempre los datos residentes en la base de datos, de forma que se trabaje con datos actualizados.
- ✓ Agiliza la consulta de grandes cantidades de datos.
- ✓ Posee alta velocidad de respuesta.
- ✓ Permite realizar querys a Data Marts.

1.12.3 Pentaho Schema Workbench 3.2.0

Pentaho Schema Workbench (PSW) es un entorno visual para el desarrollo y prueba de cubos OLAP Mondrian. Permite crear y probar los cubos OLAP visualmente para que luego el motor de Mondrian procese las solicitudes MDX con los esquemas creados.

Ofrece las siguientes funcionalidades:

- ✓ Prueba de consultas MDX contra el esquema y la base de datos.
- ✓ Consultar la base de datos que sirve de origen para el esquema de Mondrian.
- ✓ Publicar directamente el esquema en el servidor de Pentaho.

1.12.4 Apache Tomcat 5.5

Apache Tomcat es un servidor web y de aplicaciones que gestiona solicitudes y respuestas Http.

Esta herramienta posee las siguientes características:

- ✓ Es de código abierto.
- ✓ Es implementado con tecnología Java bajo la licencia de Apache 2.
- ✓ Permite funcionar en cualquier Sistema Operativo que se encuentre la máquina virtual de Java.

1.13 Conclusiones del capítulo

En este capítulo se realizó un análisis de los principales conceptos relacionados con los almacenes de datos, permitiendo una mejor comprensión del tema. Para el desarrollo del mercado de datos se tomó como base la Metodología de Ralph Kimball y como complemento a la misma y fortaleciendo la etapa de levantamiento de requisitos; se tomó lo planteado por Leopoldo Zenaido Zepeda Sánchez en su Tesis de Doctorado, orientado así el trabajo a los casos de uso. Se utilizarán herramientas pertenecientes a Pentaho Suite debido a las posibilidades de análisis y presentación que ofrece la

misma. Se definieron además como sistema gestor de bases de datos PostgreSQL 8.4, por ser un sistema robusto y poseer licencia libre, y como herramienta de modelado el Visual Paradigm 6.4.

Capítulo 2: Análisis y Diseño del mercado de datos Tecnologías de la información

Introducción

Muchos de los software que se desarrollan automatizan algunos o todos los procesos existentes en un negocio, por lo que es necesario realizar un análisis y diseño de la solución lo más adecuado posible, para tener garantía de que el software a desarrollar va a cumplir su propósito.

En el presente capítulo se realiza una descripción detallada del negocio y los temas de análisis. Se definen las reglas del negocio, los requisitos de información, los funcionales, no funcionales, multidimensionales y los casos de uso del sistema. Además, se realiza la matriz BUS, el modelo de datos y el esquema de seguridad.

2.1 Análisis

El análisis es el soporte principal para la construcción de un mercado de datos, ya que a partir del mismo se crean las bases para los siguientes procesos de diseño e implementación. Para lograr que un mercado de datos tenga la calidad requerida se necesita realizar un complejo proceso de análisis partiendo del estudio del negocio que se desea automatizar para comprender lo que el cliente necesita.

2.1.1 Definición del Negocio

La ONEI es el centro encargado de garantizar la producción de estadísticas de calidad a través del Sistema Estadístico Nacional, 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 misma está delimitada por varias áreas que cumplen con diversas funciones. Una de estas áreas es Tecnologías de la información, encargada de recoger todo lo relacionado con las tecnologías que permiten el procesamiento, transmisión, manipulación, almacenamiento y recuperación de la información. Viabilizan obtener información sobre un grupo de indicadores que posibilitan controlar y evaluar el consumo del país en Tecnologías de la informática.

2.1.2 Tema de análisis

Los temas de análisis permiten agrupar las principales necesidades en áreas de información. Esto posibilita determinar una organización global de la información y enfocar la investigación en dominios informativos. En el análisis de la investigación se definió como tema de análisis las Tecnologías de la información.

2.1.3 Necesidades de información

Con la solución actual se gestionan las necesidades informáticas de los usuarios del departamento de Tecnologías de la información de la ONEI. A continuación se detallan las necesidades de usuarios identificadas:

El usuario necesita analizar la información del modelo 0009 “Indicadores específicos de los servicios” por División Político Administrativa (DPA), por Nomenclador de Actividad Económica (NAE), por entidad y por organismos, de todos los datos que se solicitan en el modelo: real del año actual y real del año anterior.

El usuario necesita analizar la información del modelo 0764 “Ingresos por Tecnologías de la informática” por DPA y por organismos, de todos los datos que se solicitan en el modelo: cantidad de producción nacional en MP, cantidad de producción nacional en MCUC, cantidad de productos importados en MP, cantidad de productos importados en MCUC, cantidad de productos exportados y cantidad de donaciones.

2.1.4 Requisitos de información

Después de realizar un análisis del negocio se definen las siguientes necesidades existentes en la ONEI en la temática de Tecnologías de la información:

RI1: Obtener cantidad de transacciones realizadas por comercio electrónico por NAE, DPA, organismo y entidad en un tiempo dado.

RI2: Obtener facturación total de ventas efectuadas por comercio electrónico por NAE, DPA, organismo y entidad en un tiempo dado.

RI3: Obtener facturación en divisa de ventas efectuadas por comercio electrónico por NAE, DPA, organismo y entidad en un tiempo dado.

RI4: Obtener gasto total en adquisiciones realizadas por comercio electrónico por NAE, DPA, organismo y entidad en un tiempo dado.

RI5: Obtener gastos en divisa en adquisiciones realizadas por comercio electrónico por NAE, DPA, organismo y entidad en un tiempo dado.

RI6: Obtener total de nodos de transmisión de datos por NAE, DPA, organismo y entidad en un tiempo dado.

RI7: Obtener cantidad de enlaces conmutados para acceder a internet por NAE, DPA, organismo y entidad en un tiempo dado.

RI8: Obtener cantidad de enlaces arrendados para acceder a internet por NAE, DPA, organismo y entidad en un tiempo dado.

RI9: Obtener ancho de banda de salida para el servicio de internet internacional por NAE, DPA, organismo y entidad en un tiempo dado.

RI10: Obtener ancho de banda de entrada para el servicio de internet internacional por NAE, DPA, organismo y entidad en un tiempo dado.

RI11: Obtener ancho de banda disponible en la red troncal para el servicio internet nacional por NAE, DPA, organismo y entidad en un tiempo dado.

RI12: Obtener tráfico de internet conmutado por NAE, DPA, organismo y entidad en un tiempo dado.

RI13: Obtener cantidad de computadoras existentes por NAE, DPA, organismo y entidad en un tiempo dado.

RI14: Obtener cantidad de computadoras en red por NAE, DPA, organismo y entidad en un tiempo dado.

RI15: Obtener cantidad de computadoras en red con acceso a intranet por NAE, DPA, organismo y entidad en un tiempo dado.

RI16: Obtener cantidad de computadoras en red con acceso a internet por NAE, DPA, organismo y entidad en un tiempo dado.

RI17: Obtener cantidad de computadoras servidores de internet por NAE, DPA, organismo y entidad en un tiempo dado.

RI18: Obtener cantidad de computadoras usadas fundamentalmente en la actividad de la empresa por NAE, DPA, organismo y entidad en un tiempo dado.

RI19: Obtener cantidad de computadoras usadas directamente en actividades científicas por NAE, DPA, organismo y entidad en un tiempo dado.

RI20: Obtener cantidad de computadoras usadas directamente en la gestión administrativa y/o económica por NAE, DPA, organismo y entidad en un tiempo dado.

RI21: Obtener cantidad de computadoras destinadas a la automatización de procesos industriales por NAE, DPA, organismo y entidad en un tiempo dado.

RI22: Obtener cantidad de computadoras con sistema operativo Windows por NAE, DPA, organismo y entidad en un tiempo dado.

RI23: Obtener cantidad de computadoras con sistema operativo Unix por NAE, DPA, organismo y entidad en un tiempo dado.

RI24: Obtener cantidad de computadoras con sistema operativo Linux por NAE, DPA, organismo y entidad en un tiempo dado.

RI25: Obtener cantidad de computadoras con otros sistemas operativos por NAE, DPA, organismo y entidad en un tiempo dado.

RI26: Obtener cantidad de usuarios con correo electrónico por NAE, DPA, organismo y entidad en un tiempo dado.

RI27: Obtener cantidad de usuarios con correo electrónico con acceso internacional por NAE, DPA, organismo y entidad en un tiempo dado.

RI28: Obtener cantidad de usuarios con correo electrónico en el hogar por NAE, DPA, organismo y entidad en un tiempo dado.

RI29: Obtener cantidad de usuarios con correo electrónico en el hogar con acceso internacional por NAE, DPA, organismo y entidad en un tiempo dado.

RI30: Obtener cantidad de usuarios de internet por NAE, DPA, organismo y entidad en un tiempo dado.

RI31: Obtener cantidad de usuarios con internet en el hogar por NAE, DPA, organismo y entidad en un tiempo dado.

RI32: Obtener cantidad de sitios web por NAE, DPA, organismo y entidad en un tiempo dado.

RI33: Obtener cantidad de sitios web bajo dominio .cu por NAE, DPA, organismo y entidad en un tiempo dado.

RI34: Obtener cantidad de autómatas instalados por NAE, DPA, organismo y entidad en un tiempo dado.

RI35: Obtener cantidad de procesos industriales instrumentados por NAE, DPA, organismo y entidad en un tiempo dado.

RI36: Obtener cantidad de procesos industriales instrumentados automatizados por NAE, DPA, organismo y entidad en un tiempo dado.

RI37: Obtener cantidad de máquinas herramientas de control numérico por NAE, DPA, organismo y entidad en un tiempo dado.

RI38: Obtener total de gastos en divisa en equipos de computación por NAE, DPA, organismo y entidad en un tiempo dado.

RI39: Obtener gastos en divisa en computadoras por NAE, DPA, organismo y entidad en un tiempo dado.

RI40: Obtener gastos en divisas en otros equipos y accesorios por NAE, DPA, organismo y entidad en un tiempo dado.

RI41: Obtener gastos en divisa en insumos informáticos por NAE, DPA, organismo y entidad en un tiempo dado.

RI42: Obtener gastos en divisa en servicios técnicos por NAE, DPA, organismo y entidad en un tiempo dado.

RI43: Obtener total de gastos en divisa en software por NAE, DPA, organismo y entidad en un tiempo dado.

RI44: Obtener gastos en divisa en paquetes y aplicaciones por NAE, DPA, organismo y entidad en un tiempo dado.

RI45: Obtener gastos en divisa en servicios informáticos por NAE, DPA, organismo y entidad en un tiempo dado.

RI46: Obtener total de gastos en moneda nacional en equipos de computación por NAE, DPA, organismo y entidad en un tiempo dado.

RI47: Obtener gastos en moneda nacional en computadoras por NAE, DPA, organismo y entidad en un tiempo dado.

RI48: Obtener gastos en moneda nacional en otros equipos y accesorios por NAE, DPA, organismo y entidad en un tiempo dado.

RI49: Obtener gastos en moneda nacional en insumos informáticos por NAE, DPA, organismo y entidad en un tiempo dado.

RI50: Obtener gastos en moneda nacional en Servicios técnicos por NAE, DPA, organismo y entidad en un tiempo dado.

RI51: Obtener total de gastos en moneda nacional en software por NAE, DPA, organismo y entidad un tiempo dado.

RI52: Obtener gastos en moneda nacional en paquetes y aplicaciones por NAE, DPA, organismo y entidad en un tiempo dado.

RI53: Obtener gastos en moneda nacional servicios informáticos por NAE, DPA, organismo y entidad en un tiempo dado.

RI54: Obtener ingresos por equipos de computación por DPA y organismo en un tiempo dado.

RI55: Obtener ingresos por computadoras por DPA y organismo en un tiempo dado.

RI56: Obtener ingresos por otros equipos y accesorios por DPA y organismo en un tiempo dado.

RI57: Obtener ingresos por insumos informáticos por DPA y organismo en un tiempo dado.

RI58: Obtener ingresos por servicios técnicos por DPA y organismo en un tiempo dado.

RI59: Obtener ingresos por software por DPA y organismo en un tiempo dado.

RI60: Obtener ingresos por paquetes y aplicaciones por DPA y organismo en un tiempo dado.

RI61: Obtener ingresos por servicios informáticos por DPA y organismo en un tiempo dado.

RI62: Obtener total de ingresos por DPA y organismo en un tiempo dado.

2.1.5 Requisitos multidimensionales

Los requisitos multidimensionales se caracterizan por poseer variables de entrada y salida para la obtención de la información. Estos se definen a partir de los requisitos de información anteriormente descritos.

Variables de entrada y salida correspondientes a los requisitos de información del modelo “Indicadores específicos de los servicios”:

Variables de entrada:

- NAE
- DPA
- Organismo
- Entidad
- Tiempo

Variables de salida:

- Real del año actual
- Real del año anterior
- Por ciento de crecimiento
- Variación

Variables de entrada y salida correspondientes a los requisitos de información del modelo “Ingresos por Tecnologías de la informática”.

Variables de entrada

- Organismo
- DPA

Variables de salida

- Cantidad de producción nacional en MP
- Cantidad de producción nacional en MCUC
- Cantidad de importación en MP
- Cantidad de importación en MCUC
- Cantidad de exportación
- Cantidad de donaciones

2.1.6 Requisitos funcionales

Los requerimientos funcionales son capacidades o condiciones que el sistema debe cumplir de acuerdo con las necesidades y especificaciones del cliente. A continuación se enumeran las funcionalidades que debe poseer el sistema:

RF1: Realizar la extracción de los datos de TI.

RF2: Realizar la transformación y carga de los datos de TI.

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: Realizar cruce de variables.

RF12: Mostrar consulta MDX.

RF13: Suprimir filas y columnas vacías.

RF14: Imprimir reporte.

RF15: Visualizar reporte.

RF16: Exportar reporte como Excel.

2.1.7 Requisitos no funcionales

Los requisitos no funcionales son propiedades o cualidades que el producto debe tener. Estas propiedades se ven como las características que hacen al producto atractivo, usable, rápido o confiable.

Requisito de Usabilidad

RNF1: El sistema debe contar con una estructura y distribución que permita trabajar con rapidez y eficiencia.

Requisitos de fiabilidad

Son los que brindan una solución final exitosa del análisis y almacenamiento de los datos.

RNF2: El acceso a la información debe estar disponible el tiempo especificado y se tendrán en cuenta los permisos establecidos.

RNF3: El mantenimiento del sistema se realizará 1 vez por mes y en horario establecido que no afecte la productividad.

RNF4: El tiempo de reparación del sistema dependerá de lo crítico que sea el fallo.

Requisito de eficiencia

La eficiencia constituye un elemento clave para la calidad requerida de las funciones que deberá realizar el sistema.

RNF5: El sistema debe tener un tiempo de respuesta aproximado de 5 segundos, no debe excederse de 1 minuto.

RNF6: El sistema debe permitir de 5 a 10 usuarios conectados simultáneamente sin que se afecte el tiempo de respuesta.

Requisito de diseño

RNF7: Para realizar las consultas a la Base de datos se utilizará PostgreSQL en su versión 8.4.

RNF8: Visual Paradigm 6.4 es la herramienta CASE utilizada por el centro para el modelado conceptual de la información, no es un software libre pero se cuenta con la licencia de este programa.

Requisitos de Interfaz

Los requerimientos de interfaz son aquellos que estarán en contacto directo con el usuario, por lo tanto, incluyen el aspecto del sistema y el contenido que verá el usuario.

RNF9: Los reportes mostrarán una interfaz sencilla que permita la interacción usuario-aplicación.

RNF10: Para el proceso de transformación es necesaria una memoria RAM de 512 MB como mínimo.

RNF11: Para el proceso de visualización e inteligencia de negocio se necesita una memoria RAM de 1 GB como mínimo, para garantizar el correcto funcionamiento del sistema cuando es accedido por varios usuarios simultáneamente.

RNF12: Se debe garantizar al menos una impresora para imprimir los reportes de salida.

2.1.8 Reglas del Negocio

A continuación se definen las reglas del negocio, o sea, las transformaciones que se le deben realizar a ciertos datos para obtener otros datos o los segmentos de un dato que pueden generar otros datos sin necesidad de ser introducidos por una persona.

- Crecimiento: Representa el por ciento del crecimiento del año actual con respecto al anterior. El resultado se debe devolver con una cifra decimal; en caso de devolver un entero se deja un cero después de la coma. Ej. 77,8% ó 50,0%.

  Crecimiento= (real_anno_actual / real_anno_anterior) * 100

- Variación contra año actual: Representa la diferencia entre el real existente y el real del año anterior para el período que se está analizando. El resultado se debe devolver con una cifra decimal; en caso de devolver un entero se deja un cero después de la coma. Ej. 75,6 ó 97,0.

  Variación = real_anno_actual - real_anno_anterior

- La cantidad de computadoras existentes tiene que ser mayor e igual que la cantidad de computadoras en red.
- La cantidad de computadoras existentes tiene que ser mayor e igual que las usadas fundamentalmente en la actividad de la empresa.
- La cantidad de computadoras existentes tiene que ser mayor e igual que las usadas directamente en actividades científicas.
- La cantidad de computadoras existentes tiene que ser mayor e igual que las usadas directamente en la gestión administrativa y/o económica.
- La cantidad de computadoras existentes tiene que ser mayor e igual que las destinadas a la automatización de procesos industriales.
- La cantidad de computadoras en red tienen que ser mayor e igual que las computadoras con acceso a intranet.
- La cantidad de computadoras con acceso a internet tiene que ser mayor e igual que la cantidad de servidores de internet.
- La cantidad de usuarios con correo electrónico tiene que ser mayor e igual que la cantidad de usuarios con correo electrónico con acceso internacional.
- La cantidad de usuarios con correo electrónico en el hogar tiene que ser mayor e igual que la cantidad de usuarios con correo electrónico en el hogar con acceso internacional.
- La cantidad de sitios web tiene que ser mayor e igual que la cantidad de sitios web bajo dominio cu.
- La cantidad de procesos industriales instrumentados tiene que ser mayor e igual que la cantidad de procesos industriales instrumentados automatizados.
- El total de gastos en equipos de computación es igual a la suma de computadoras, otros equipos y accesorios, insumos informáticos y servicios técnicos.

- El total de gastos en software es igual a la suma de paquetes y aplicaciones y servicios informáticos.
- La facturación total de ventas efectuadas por comercio electrónico es mayor que la facturación total de ventas efectuadas por comercio electrónico en divisas.
- El gasto total en adquisiciones realizadas por comercio electrónico es mayor que el gasto total en adquisiciones realizadas por comercio electrónico en divisas.
- El total de ingresos es igual a la suma de los ingresos por equipos de computación y los ingresos por software.
- Los ingresos por equipos de computación son igual a la suma de computadoras, otros equipos y accesorios, insumos informáticos y servicios técnicos.
- Los ingresos por software son igual a la suma de paquetes y aplicaciones y servicios informáticos.

2.1.9 Especificación de Casos de Uso

- Casos de uso de información.
  - Analizar indicadores generales de las Tecnologías de la información.
  - Analizar los ingresos de las Tecnologías de la información.
- Casos de uso funcionales
  - Realizar transformación y carga de los datos de Tecnologías de la información.
  - Realizar la extracción de los datos de Tecnologías de la información.
  - Autenticar usuario.
  - Administrar reporte.
  - Administrar rol
  - Administrar usuario.

Diagrama de Casos de Uso del sistema

Un diagrama de casos de uso del sistema es un modelo que contiene actores, casos de uso y sus relaciones.

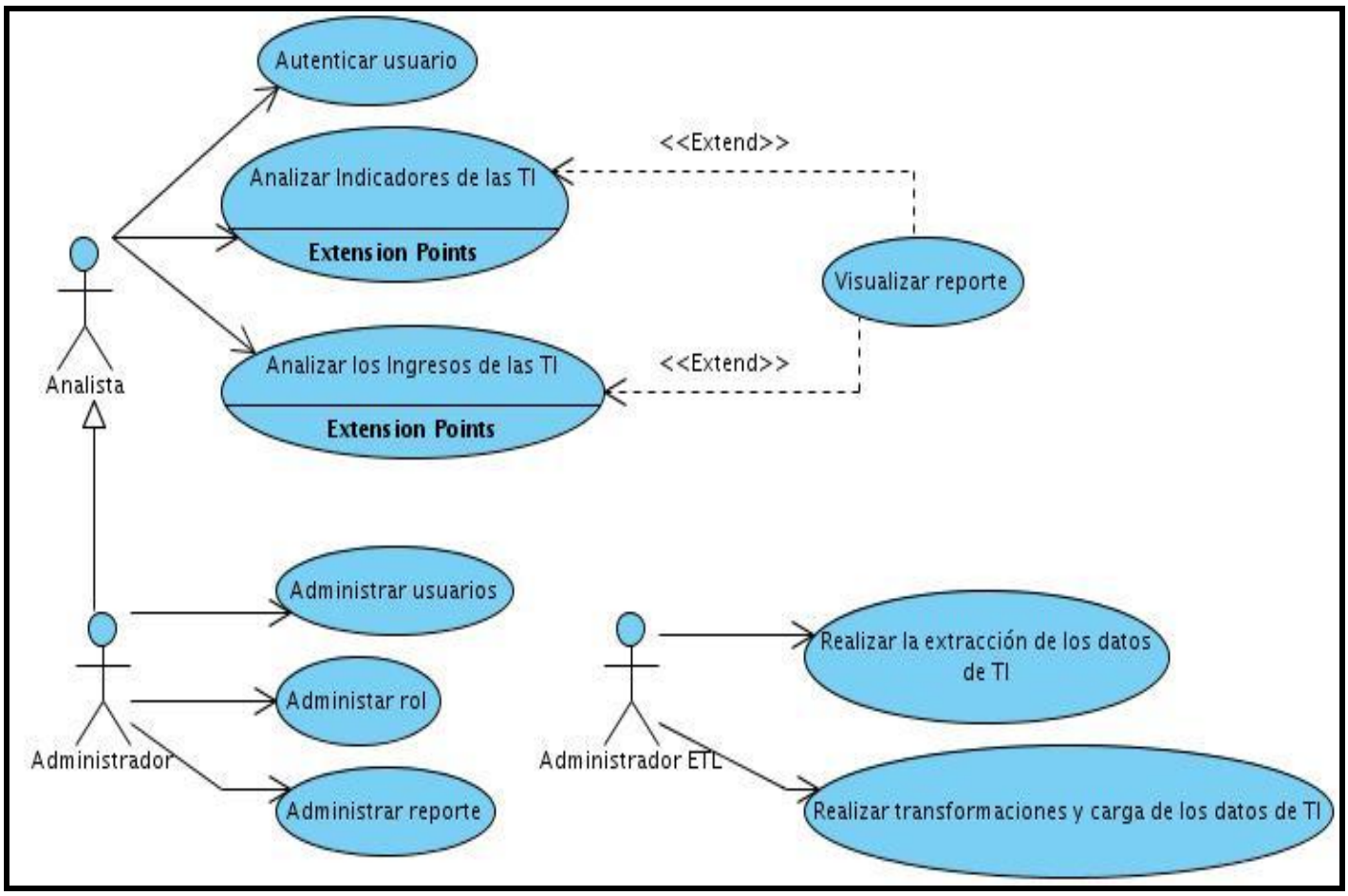


Figura 4: Diagrama de Casos de Uso del sistema.

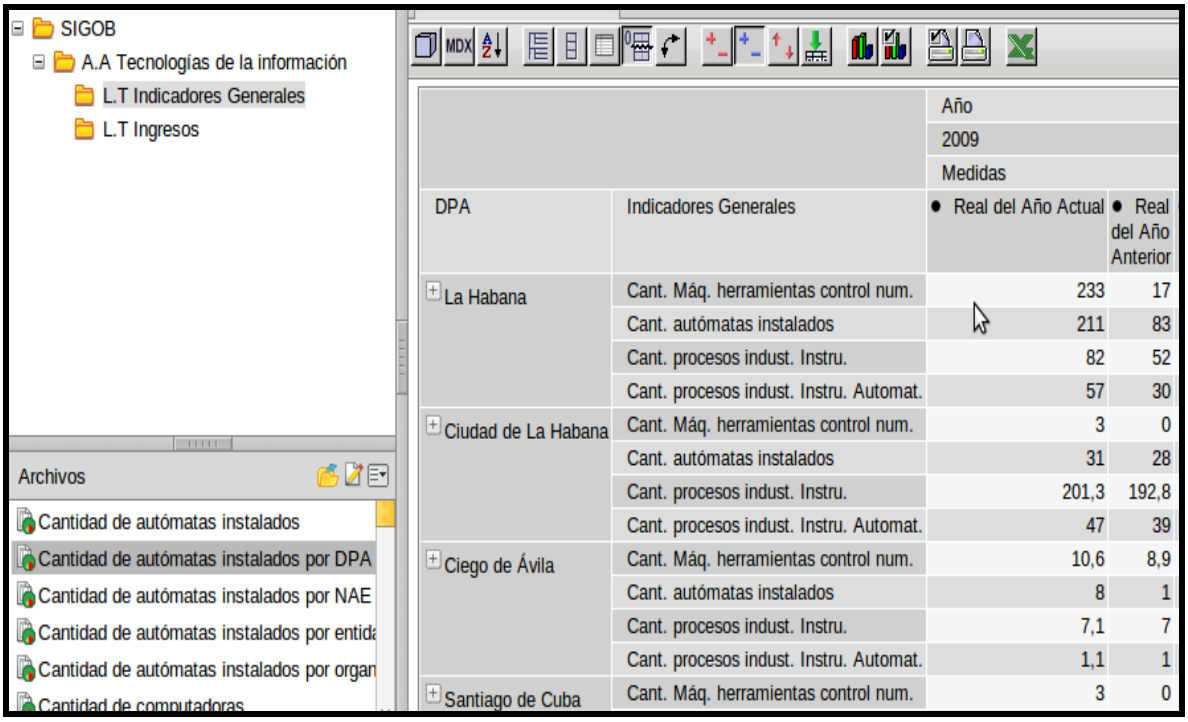
Descripción de los Casos de Uso.

A continuación se muestra la descripción del caso de uso de información Analizar los Ingresos de las TI y del caso de uso funcional Autenticar usuario. Para mayor información, consultar el expediente de proyecto.

Caso de Uso:	Analizar los Ingresos de las TI
Tipo:	Información
Actores:	Analista, Administrador
Resumen:	El CU inicia cuando el actor entra al sistema, selecciona el reporte Ingresos de las TI. El CU finaliza una vez que el especialista termina de analizar el reporte.
Precondiciones	Carga de los datos. El usuario tiene que estar autenticado.
Referencias	RI54, RI55, RI56, RI57, RI58, RI59, RI60, RI1, RI62: Caso de Uso

	Visualizar reporte.
Prioridad	Crítico
Flujo Normal de Eventos	
Acción del Actor	**Respuesta del Negocio**
1. El actor se autentica en el sistema.	2. Muestra la interfaz principal con las áreas de análisis existentes.
3. El actor selecciona el área de análisis A.A Tecnologías de la información.	4. Muestra los libros de trabajo que están contenidos dentro del A.A Tecnologías de la información.
5. El especialista selecciona el libro de trabajo L.T Ingresos.	6. Muestra los reportes contendidos dentro del L.T Ingresos.
7. El actor selecciona el reporte deseado.	8. Muestra la información contenida en el reporte seleccionado y brinda la posibilidad al actor de hacerle cambios al reporte para su análisis. Ir al Caso de Uso **Visualizar reporte**. Finaliza el caso de uso.

Prototipo de Interfaz



Opciones de reportes

Entradas	Posibles resultados	
	Salidas	Periodicidad
Variables de entrada disponibles relacionadas con el caso de uso Analizar los Ingresos de las TI. ✓ DPA ✓ Organismo ✓ Temporal trimestre ✓ Ingresos	Variables de salida disponibles el caso de uso Analizar los Ingresos de las TI. ✓ Cantidad de producción nacional en MP ✓ Cantidad de producción nacional en MCUC ✓ Cantidad de importación en MP ✓ Cantidad de importación en MCUC ✓ Cantidad de exportación	Rango de tiempo en que se solicitan las variables de salida: ✓ Trimestral

	✓ Cantidad de donaciones	

Tabla 1: Casos de Uso de información ¨Analizar los Ingresos de las TI¨.

Caso de Uso:	Autenticar usuario
Tipo:	Funcional
Actores:	Analista, Administrador
Resumen:	El CU inicia cuando el actor selecciona la opción “Autenticar usuario”, que le permite entrar al sistema a través de un usuario y contraseña. El CU termina una vez validado el usuario y la contraseña cuando el actor logra acceder al sistema.
Precondiciones:	• El sistema debe estar disponible. • El usuario debe existir en la BD.
Referencias	RF3

Flujo Normal de Eventos	
Acción del Actor	**Respuesta del Sistema**
1. El actor accede al sistema.	1.1. El sistema muestra un formulario para introducir los datos.
2. El actor introduce los datos.	2.1. El sistema valida los datos.
	2.2. Le da los permisos para entrar a la aplicación.

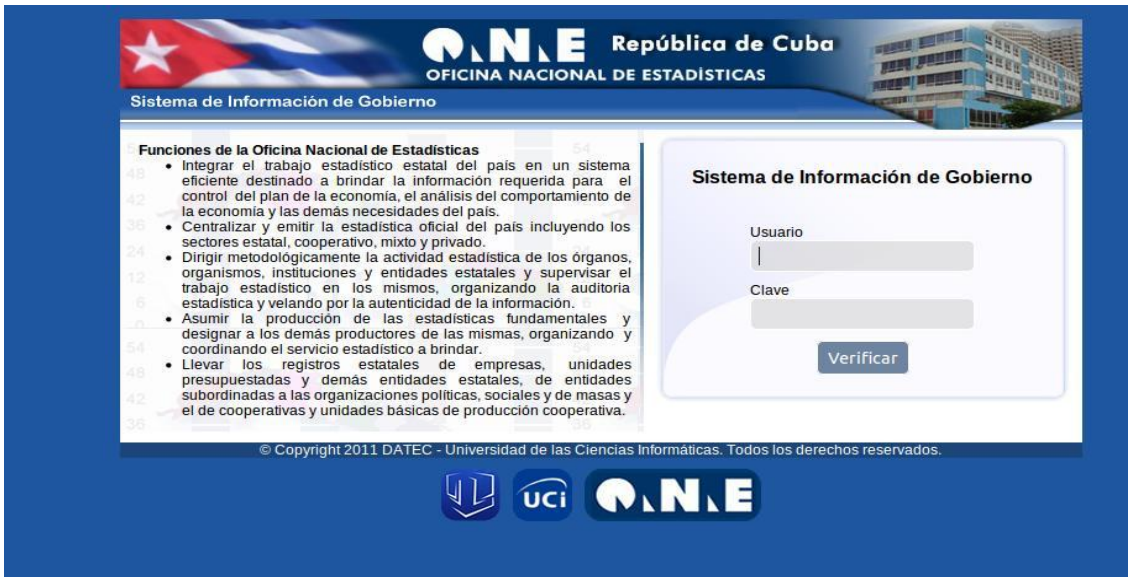
Prototipo de Interfaz	
	
Flujos Alternos	
Acción del Actor	**Respuesta del Sistema**
	2.2. El sistema muestra un mensaje de error y regresa al punto 1.1.

Tabla 2: Descripción del caso de uso funcional ¨Autenticar Usuario¨.

2.2 Diseño

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

2.2.1 Dimensiones identificadas

Luego del análisis de los Modelos 0009 y 0764, así como de los reportes que se realizan de estos, se definieron las dimensiones que intervendrán en el sistema a diseñar.

- dim_dpa

Esta dimensión describe el universo de valores bajo los cuales puede clasificarse la información atendiendo a la situación geográfica. Presenta un nivel “provincia” y dentro de este se encuentra “municipios”.

Provincia → Municipios

- dim_nae

Esta dimensión describe el universo de valores bajo los cuales puede clasificarse la información atendiendo a los Nomencladores de Actividad Económica. Presenta los niveles sección, división y clase.

Sección → División → Clase

- dim_entidad

Esta dimensión describe el universo de valores bajo los cuales puede clasificarse la información de acuerdo con el centro que emite la misma.

- dim_temporal_anno

Esta dimensión describe el universo de valores bajo los cuales puede clasificarse la información atendiendo a los valores anuales.

- dim_temporal_trimestre

Esta dimensión describe el universo de valores bajo los cuales puede clasificarse la información atendiendo a los valores anuales y trimestrales que posee el modelo. Presenta un nivel denominado “anno” y dentro la categoría “trimestre”.

Anno → Trimestre

- dim_indicador_general

Esta dimensión describe el universo de valores bajo los cuales puede clasificarse la información atendiendo a los indicadores que se recogen del Modelo 0009.

- dim_organismo

Esta dimensión describe el universo de valores bajo los cuales puede clasificarse la información atendiendo al organismo que la emite.

- dim_ingresos

Esta dimensión describe el universo de valores bajo los cuales puede clasificarse la información atendiendo a los indicadores que se recogen del Modelo 0764.

2.2.2 Hechos identificados

La tabla de hechos es la tabla central en un esquema dimensional y contienen los valores de las medidas.

El modelo consta de 2 tablas de hechos, por la necesidad de dividir la información según los modelos que genera el departamento de Tecnologías de la información.

- Tabla hech_indicadores_ti

Se utiliza para el trabajo con los datos de los indicadores generales de las Tecnologías de la información recogidos en el modelo 0009. Se relaciona con las dimensiones dim_indicador_general, dim_dpa, dim_nae, dim_entidad, dim_temporal_anno y dim_organismo.

- Tabla hech_ingresos_ti

Esta tabla recoge los datos de los ingresos por Tecnologías de la información generados en el modelo 0764. Se relaciona con las dimensiones dim_ingresos, dim_dpa, dim_temporal_trimestre y dim_organismo.

2.2.3 Medidas identificadas

Las medidas son indicadores de negocio, representan datos cuantificables que permiten medir el proceso de negocio.

- Tabla hech_indicadores_ti
  - Real del año actual
  - Real del año anterior
  - Variación
  - Por ciento de crecimiento
- Tabla hech_ingresos_ti
  - Cantidad de producción nacional en MP
  - Cantidad de producción nacional en MCUC
  - Cantidad de importación en MP
  - Cantidad de importación en MCUC
  - Cantidad de exportación
  - Cantidad de donaciones

2.2.4 Matriz Bus

La Matriz Bus representa las relaciones existentes entre los hechos y las dimensiones del modelo de datos. Define parte de la arquitectura del mercado de datos y se aplica en las siguientes fases del modelado dimensional y el desarrollo de la solución.

DIMENSIONES	HECHOS	
	hech_indicadores_ti	hech_ingresos_ti
dim_dpa	X	X
dim_nae	X	
dim_organismo	X	X
dim_entidad	X	
dim_temporal_trimestre		X
dim_temporal_anno	X	
dim_indicador_general	X	
dim_ingresos		X

Tabla 3: Matriz Bus.

2.2.5 Modelo de datos

A continuación se expone el modelo de datos compuesto por los hechos, dimensiones y medidas seleccionados según los modelos estudiados en la investigación.

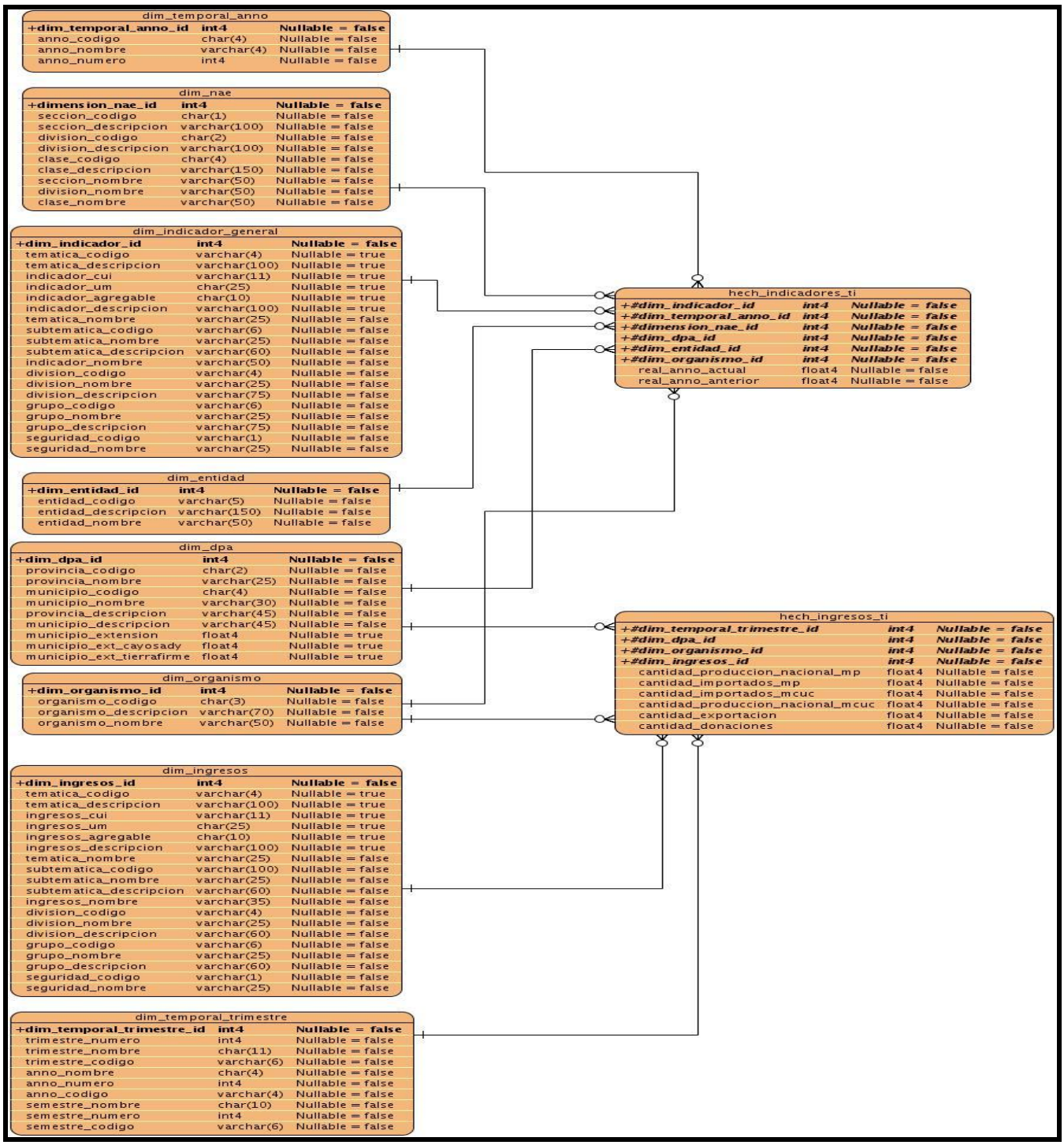


Figura 5: Modelo de datos.

2.2.6 Roles y permisos

Se definen los roles y permisos de cada rol en dependencia de la actividad que vaya a desarrollar en el mercado de datos, ya sea lectura o escritura y depende además del tipo de información que vaya a manejar.

ROLES	PERMISOS	
	Lectura	Escritura
Analista	X	
Administrador	X	X
Administrador de ETL	X	X

Tabla 4: Roles identificados y permisos.

Analista: Analiza y consulta los reportes relacionados con los temas de análisis correspondientes a los modelos de Tecnologías de la información.

Administrador de ETL: Es la persona encargada de realizar el proceso de ETL.

Administrador: Es la persona encargada de administrar la base de datos, tiene acceso a todas las áreas de análisis en general y gestiona el Sistema de Información de Gobierno.

2.3 Conclusiones del capítulo

En este capítulo se definió el tema de análisis Tecnologías de la información, además de los roles y permisos en el mercado de datos para garantizar la seguridad del mismo. Se identificaron 62 requisitos de información, 16 requisitos funcionales y 12 requisitos no funcionales, los cuales fueron agrupados en 9 Casos de Usos. Se realizó el diseño del modelo de datos identificándose 8 tablas de dimensiones y 2 tablas de hechos, se confeccionó la matriz Bus y fueron detectadas 21 reglas de negocio que fueron importantes a la hora de realizar el proceso de ETL.

Capítulo 3: Implementación del mercado de datos Tecnologías de la información

Introducción

En el presente capítulo se muestra cómo está estructurado el modelo físico y se realiza la implementación de los subsistemas de integración y visualización. Se implementan los reportes candidatos y se describen los procesos de perfilado, extracción, transformación y carga de los datos.

3.1 Modelo de datos físico

El modelo de datos es una representación de un diseño de datos, que tiene en cuenta las facilidades y restricciones de los sistemas de base datos. Permite describir las estructuras de datos de la base datos y la forma en que estos se relacionan.

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

3.1.1 Esquemas y tablas

Los esquemas son formas de organización de la base de datos que pueden contener tablas, tipos de datos, funciones y operadores. Permiten organizar los objetos de la base de datos en grupos lógicos, posibilitando así utilizar el mismo nombre para un objeto en esquemas diferentes sin ocasionar un conflicto. Para el desarrollo del sistema propuesto en esta investigación se definieron 2 esquemas: el esquema dimensiones, que contiene todas las dimensiones comunes del almacén de datos, y el esquema mart_tecnologia_inf, que agrupa todos los hechos de dicho mercado y las dimensiones propias del mercado de datos. La solución cuenta con 8 tablas dimensiones y 2 tablas hechos, distribuidas por los dos esquemas propuestos con anterioridad quedando de la siguiente manera:

Esquemas	Tablas
dimensiones	dim_nae
dimensiones	dim_organismo
dimensiones	dim_entidad
dimensiones	dim_temporal_trimestre
dimensiones	dim_temporal_anno
dimensiones	dim_indicador_general
dimensiones	dim_dpa
mart_tecnologia_inf	dim_ingresos
mart_tecnologia_inf	hech_indicadores_ti
mart_tecnologia_inf	hech_ingresos_ti

Tabla 5: Esquemas y tablas identificados.

3.2 Arquitectura de integración

La arquitectura en el proceso de desarrollo de software es el diseño de más alto nivel de la estructura de un sistema, consiste en un conjunto de patrones que proporcionan el marco de referencia necesario para guiar la construcción del software.

En el proceso de integración de datos no es recomendable comenzar el desarrollo de una solución sin haberla premeditado. La arquitectura de este proceso está formada dos elementos necesarios en la implementación del sistema:

- Fuente de datos.
- Mercado de datos.

Fuente de datos: son los ficheros que guardan información histórica de los sistemas, los cuales se encuentran en formato Excel y DBF.

Mercado de datos: constituye el destino a donde se integrarán los datos a cargar.

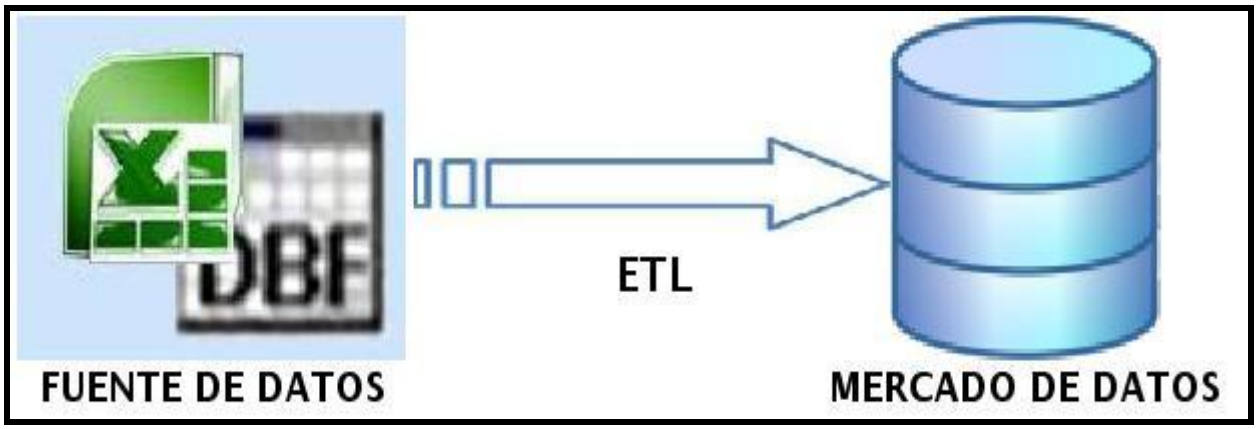


Figura 6: Arquitectura de integración.

3.2 Proceso de integración de datos

3.2.1 Perfilado de datos

El perfilado de datos consiste en realizar un primer análisis sobre los datos de origen, con el objetivo de empezar a conocer su estructura, formato y nivel de calidad. A partir de este proceso se establecen reglas para corregir los defectos de los datos y así garantizar la disponibilidad de los mismos.

Para el análisis de los datos de Tecnologías de la información se utilizó la herramienta DataCleaner 1.5.3, la cual arrojó resultados importantes para posteriormente realizar el proceso ETL.

3.2.2 Extracción, transformación y carga de los datos

Los ficheros Excel y DBF acumulan información relacionada con los ingresos e indicadores generales de Tecnologías de la información de los modelos 0009 y 0764 que serán almacenados en las tablas dim_indicador_general, dim_ingresos, hech_indicadores_ti y hech_ingresos_ti respectivamente.

A continuación se muestra un ejemplo de la carga de la tabla hech_ingresos_ti.

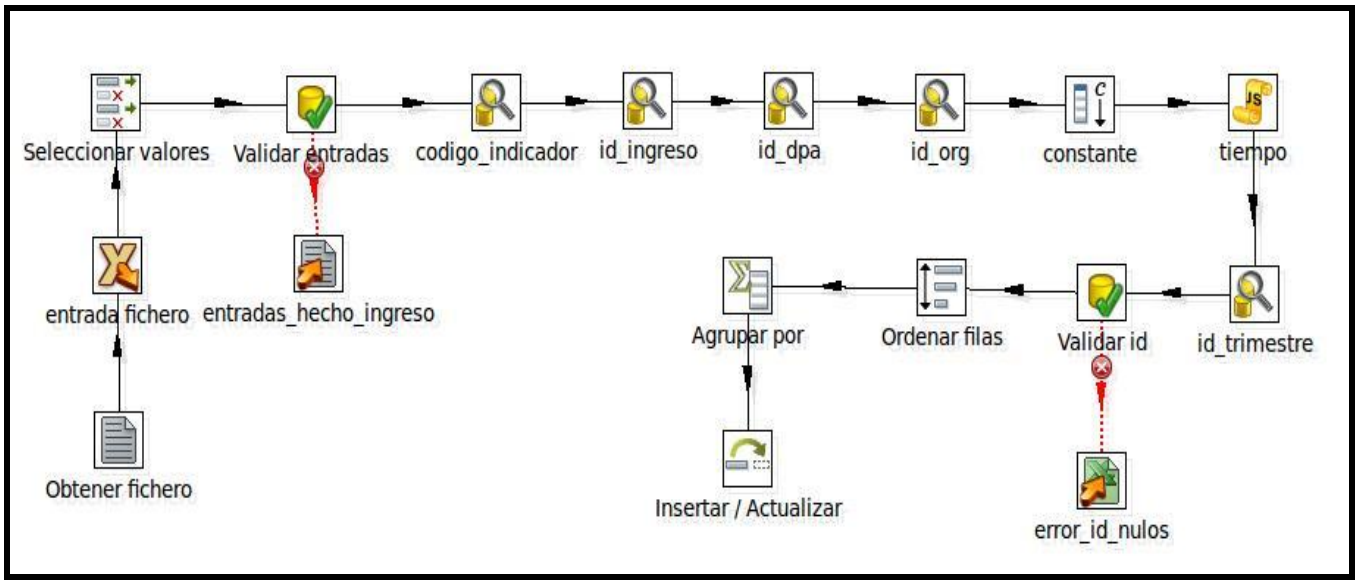


Figura 7: Carga del hecho Ingresos de las TI.

Como se muestra en la Figura la información almacenada en los ficheros fuentes correspondientes al Modelo 0764 es extraída para realizarle un proceso de transformación y limpieza, ajustándola a las necesidades de información del cliente. Posteriormente se realiza un proceso de búsqueda de los id de las dimensiones relacionadas con el hecho, para luego realizar la carga de los datos en la tabla destino.

3.3 Trabajo para organizar el orden de la carga

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

La siguiente figura muestra el trabajo para ejecutar este proceso.

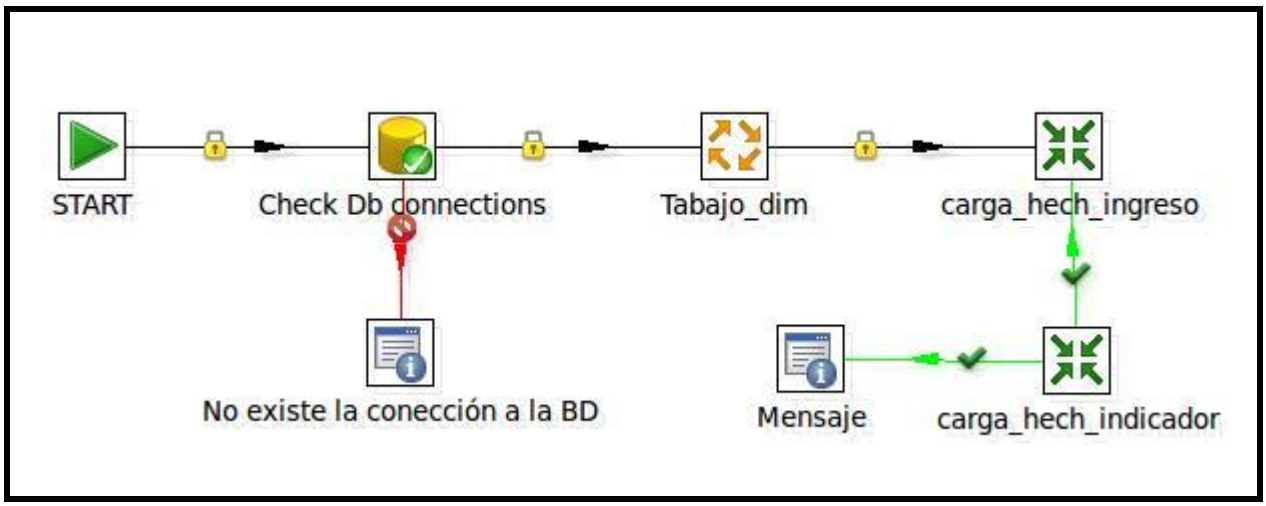


Figura 8: Trabajo del mercado de datos.

En este proceso se realizó primero la carga de la tabla de dimensión y luego las tablas de hecho. Destacar que no fue necesario seguir un orden entre las cargas de las tablas hechos, ya que no existe dependencia entre ellas.

3.4 Proceso de Inteligencia de Negocios

El éxito final del almacén de datos depende en gran medida de la presentación de la capa de visualización y de la forma en que se organiza la información estructural.

3.4.1 Implementación de los cubos OLAP

Uno de los elementos fundamentales donde se organiza la información estructuralmente son los cubos multidimensionales. Estos se realizan utilizando la herramienta Pentaho Schema Workbench que genera un fichero de configuración XML.

En este fichero de esquema se pueden definir las dimensiones, los niveles de jerarquía de dimensiones y los hechos. Se modelaron 2 cubos multidimensionales compuestos por 8 dimensiones y 10 medidas, de ellas 2 calculables que dan solución al problema planteado.

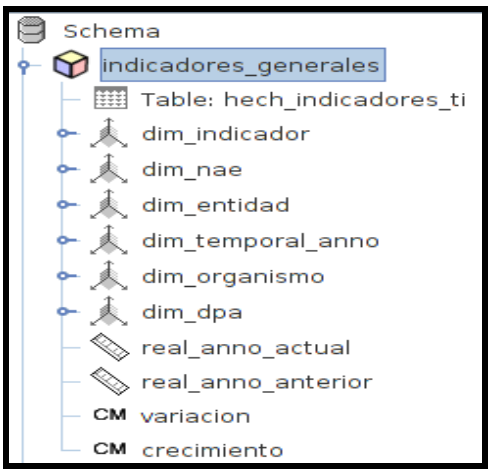


Figura 9: Cubo Indicadores generales.

3.4.2 Arquitectura de información

La Arquitectura de información define la organización de la información con el objetivo de permitir al usuario encontrar su vía de navegación hacia el conocimiento y la comprensión de la información.

La capa de visualización del mercado de datos está formada por el área de análisis Tecnologías de la información, compuesta por los libros de trabajo Indicadores Generales e Ingresos, los cuales contienen 38 reportes multidimensionales.

Figura 10: Arquitectura de información.

3.4.3 Reportes Candidatos

Los reportes candidatos representan la información que el cliente desea que se muestre como finalidad del producto. Los mismos fueron seleccionados luego de realizar un análisis de los modelos 0009 y 0764, donde se recoge toda la información referente al área de Tecnologías de la información de la ONEI. A continuación se muestran los reportes candidatos que fueron identificados.

Reportes candidatos del libro de trabajo Indicadores Generales de las TI:

- ✓ Gastos Generales.

- ✓ Gastos por DPA.
- ✓ Gastos por organismos.
- ✓ Gastos por NAE.
- ✓ Gastos por entidades.
- ✓ Comercio electrónico.
- ✓ Comercio electrónico por DPA.
- ✓ Comercio electrónico por organismos.
- ✓ Comercio electrónico por NAE.
- ✓ Comercio electrónico por entidades.
- ✓ Transmisión de datos.
- ✓ Transmisión de datos por DPA.
- ✓ Transmisión de datos por organismos.
- ✓ Transmisión de datos por NAE.
- ✓ Transmisión de datos por entidades.
- ✓ Cantidad de computadoras.
- ✓ Cantidad de computadoras por DPA.
- ✓ Cantidad de computadoras por organismos.
- ✓ Cantidad de computadoras por NAE.
- ✓ Cantidad de computadoras por entidades.
- ✓ Cantidad de usuarios.
- ✓ Cantidad de usuarios por DPA.
- ✓ Cantidad de usuarios por organismos.
- ✓ Cantidad de usuarios por NAE.
- ✓ Cantidad de usuarios por entidades.
- ✓ Cantidad de sitios web.
- ✓ Cantidad de sitios web por DPA.
- ✓ Cantidad de sitios web por organismos.
- ✓ Cantidad de sitios web por NAE.
- ✓ Cantidad de sitios web por entidades.
- ✓ Cantidad de autómatas.
- ✓ Cantidad de autómatas por DPA.
- ✓ Cantidad de autómatas por organismos.
- ✓ Cantidad de autómatas por NAE.

- ✓ Cantidad de autómatas por entidades.

Reportes candidatos del libro de trabajo Ingresos de las TI:

- ✓ Total de ingresos.
- ✓ Total de ingresos por organismos.
- ✓ Total de ingresos por DPA.

3.5 Conclusiones del capítulo

En el presente capítulo se realizó el modelo físico y perfilado de los datos. Se definió la arquitectura de integración del sistema dando paso a la implementación del proceso de integración, lográndose la carga exitosa de los datos. Como parte de la estructura OLAP del sistema se diseñaron dos cubos, tributando al área de análisis Tecnologías de la información, desglosada en dos libros de trabajo, para un total de 38 vistas de análisis identificadas en la etapa de levantamiento de requisitos, las cuales finalmente quedaron implementadas.

Capítulo 4: Validación del mercado de datos Tecnologías de la información

Introducción

Una vez culminada la implementación de la solución, se da paso a la validación de la misma, verificando que se cumplan todos los requisitos que hacen confiable la información almacenada en el mercado de datos. En este capítulo se realizan las validaciones y pruebas a la solución.

4.1 Pruebas

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

A continuación se mencionan algunos tipos de pruebas que demuestran cómo se relacionan las actividades de prueba con las de análisis y diseño:

Prueba de integración: es la fase de prueba en la cual los componentes son agregados para crear componentes más grandes.

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

4.2 Herramientas para aplicar las pruebas

Para realizar las pruebas al mercado de datos del área Tecnologías de la información se utilizaron las siguientes herramientas:

4.2.1 Listas de chequeo

Las listas de chequeo constituyen un mecanismo para el control de los riesgos, la cual tiene 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:

- 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.

A continuación se muestra la lista de chequeo para evaluar el análisis y diseño realizado en la presente investigación:

Estructura del documento					
Peso	**Indicadores a evaluar**	**Eval**	**(NP)**	**Cantidad de elementos afectados**	**Comentarios**
crítico	✓ ¿Los entregables contienen las secciones obligatorias de la plantilla estándar definidas para un expediente de proyecto? (portada, control de versiones, reglas de confidencialidad, tabla de contenidos y contenido)				
Indicadores definidos en el desarrollo					
Peso	**Indicadores a evaluar**	**Eval**	**(NP)**	**Cantidad de elementos afectados**	**Comentarios**
	✓ ¿Se utilizó un lenguaje cuyas sentencias son expresables mediante una sintaxis bien definida?				
crítico	✓ ¿Los reportes son configurables a través de la interfaz del sistema?				
	✓ ¿La interfaz está orientada a facilitar el uso de las funciones del sistema por parte de los usuarios?				
crítico	✓ ¿No existen restricciones para				

	construir cubos OLAP con dimensiones y niveles de agregación ilimitados?				
crítico	✓ ¿Los usuarios son capaces de manipular los resultados de manera que se ajusten a sus necesidades, conformando nuevos reportes?				
	✓ ¿El sistema responde de una forma rápida a la información que le sea solicitada por el usuario?				
	✓ ¿El sistema refleja cualquier lógica del negocio para poder responder a preguntas específicas?				
crítico	✓ ¿El sistema garantiza la confidencialidad y seguridad de acceso a los datos por rol de los usuarios?				
	✓ ¿Los datos e información derivados del proceso de análisis realizado mediante la aplicación, apoyan la toma de decisiones en la Institución?				
crítico	✓ ¿Los cambios en los datos se reflejan automáticamente en los reportes de forma instantánea?				
Semántica del documento					
Peso	**Indicadores a evaluar**	**Eval**	**(NP)**	**Cantidad de elementos afectados**	**Comentarios**
crítico	✓ ¿Se han identificado errores ortográficos en los entregables?				

crítico	✓ ¿Se entiende claramente lo que se ha especificado en el documento?				
	✓ ¿El número de página que aparece en el índice coincide con el contenido que se refleja realmente en dicha página?				

Tabla 6: Lista de chequeo aplicada al mercado de datos Tecnologías de la información.

4.2.2 Casos de prueba

Los casos de prueba tienen como propósito comprobar que un requisito es completamente satisfactorio, con este propósito, debe diseñarse al menos un caso de prueba por caso de uso de información.

A continuación se muestra el caso de prueba diseñado para evaluar el caso de uso de información "Analizar los Ingresos de las TI":

Escenario	Descripción	Perfiles de análisis	Indicadores a medir	Respuesta del sistema	Flujo central
EC 1.1: Ingresos por DPA	Este reporte muestra los ingresos por DPA de Tecnologías de la información.	DPA ingresos Trimestre	Cantidad de Importados en MCUC Cantidad de Importados en MP Cantidad de Producción Nacional en MCUC Cantidad de Producción Nacional en MP Cantidad de Donaciones Cantidad de Exportaciones	Se muestra la tabla con los valores correspondientes a cada escenario.	Se abre la aplicación. Se autentifica. Se entra al sistema. Se despliega hacia la derecha el componente ubicado en el lateral izquierdo que contiene el navegador.

					Se selecciona el área de análisis de **AA. Tecnologías de la información.** Se selecciona el libro de trabajo **LT. Ingresos.** En la parte inferior izquierda se selecciona el reporte deseado. En el área de trabajo se visualiza la tabla correspondiente al reporte.
EC 1.2: Ingresos por organismo	Este reporte muestra los ingresos por organismo de Tecnologías de la información.	ingresos Organismo Trimestre	Cantidad de Importados en MCUC Cantidad de Importados en MP Cantidad de Producción Nacional en MCUC Cantidad de Producción Nacional en MP		

			Cantidad de Donaciones Cantidad de Exportaciones		
EC 1.3: Ingresos totales	Este reporte muestra los ingresos totales de Tecnologías de la información.	ingresos Trimestre	Cantidad de Importados en MCUC Cantidad de Importados en MP Cantidad de Producción Nacional en MCUC Cantidad de Producción Nacional en MP Cantidad de Donaciones Cantidad de Exportaciones		

Tabla 7: Caso de prueba "Ingresos de las Tecnologías de la información".

4.3 Pruebas aplicadas al mercado de datos

4.3.1 Prueba integración

Para la validación del mercado de datos y la capa de inteligencia de negocios se aplicaron los casos de prueba diseñados y la lista de chequeo elaborada. En esta prueba se integran los módulos, y se verifica que los datos cargados en el mercado sean los que se muestran finalmente al usuario por medio de la capa de visualización.

En una primera iteración se realizaron las pruebas en el centro, durante el proceso los especialistas detectaron doce no conformidades, que no afectan el buen funcionamiento del producto pero si la calidad de la información presentada. Las no conformidades detectadas fueron resueltas de inmediato y en estos momentos el producto se encuentra en otra iteración de la etapa de prueba, en la que está siendo evaluado por los especialistas de Calisoft.

4.3.2 Prueba de aceptación

La validación es la comprobación de que la solución está acorde a las necesidades y exigencias de los clientes. Para aprobar la propuesta de solución se realizó un encuentro con la representante de la ONEI en la universidad: Elena Leonila Fernández García, detectándose algunas diferencias en los

datos. Al realizar una revisión en las transformaciones para las cargas de las tablas de hecho se detectó que había indicadores que se estaban desviando para el fichero de error. Luego de solucionar el error la especialista reevaluó el producto y aprobó la carga de los datos, su presentación visual y las funcionalidades del sistema.

4.4 Conclusiones del capítulo

En este capítulo se realizó una breve descripción de los tipos de pruebas. Para evaluar el mercado de datos Tecnologías de la información se diseñó un caso de prueba por caso de uso de información y se elaboró una lista de chequeo. Luego de ser evaluados por un especialista del centro, fueron detectadas algunas no conformidades relacionadas con el uso excesivo de mayúsculas en el nombre de las vistas de análisis, la descripción y el orden de las mismas, las cuales no se correspondían con lo descrito en el artefacto de ¨Reportes Candidatos¨ y ¨Casos de Prueba¨, las deficiencias señaladas fueron resueltas de inmediato. Se realizó la prueba de aceptación, verificando que los datos fueron cargados correctamente y que la presentación visual de los mismos coincida con los datos almacenados, como resultado de esta prueba se obtiene la carta de aceptación del cliente.

Conclusiones

Al finalizar el trabajo de diploma “Sistema de Información de Gobierno. Mercado de datos Tecnologías de la información” se arribaron a las siguientes conclusiones:

- Las herramientas seleccionadas garantizaron el buen funcionamiento del mercado de datos y la capa de visualización.
- Los 62 requisitos de información y los 16 requisitos funcionales definidos dan solución a las solicitudes del cliente.
- Las estructuras dimensionales modeladas engloban la forma de organización de la información del Modelo 0009 y el Modelo 0764 y permiten a los especialistas realizar sin dificultades el proceso de toma de decisiones.
- El Sistema Gestor de Base de datos PostgreSQL permitió la administración de los volúmenes de información que necesita este tipo de solución.
- Al realizar la carga de los datos en el almacén se observó que no hubo pérdida de información y la presentación visual de la misma permite a los usuarios analizar los principales reportes y cruce de variables.
- Durante las etapas de validación se demostró que el sistema cumple con las expectativas del cliente.
- Se efectuaron las pruebas internas al mercado de datos.

Recomendaciones

- Integrar el mercado de datos Tecnologías de la información con el Sistema Informático de Gestión Estadística (SIGE).

Referencias Bibliográficas

1. VELASCO, R. H. *Almacenes de datos*, 2009. [2010]. Disponible en: http://www.rhernando.net/modules/tutorials/doc/bd/dw.html
2. INMON, W. H. *Using the Data Warehouse* 1992. p.
3. LÖFBERG, M. and P. MOLIN. *Web vs. Standalone Application: A maintenance application for Business Intelligence*, Blekinge Institute of Technology, 2005. 40. p.
4. EVELIA, C. C. M. *Data Warehouse (Almacenes de Datos)*, 2009. Disponible en: http://hp.fciencias.unam.mx/~alg/bd/dwh.pdf
5. FERREIRA, K. and J. SCHMIDT Sistemas de Información. Pentaho, 2009.
6. GALICIA, B. I. *Data Warehouse*, 2007. [2010]. Disponible en: http://www.sinnexus.com/business_intelligence
7. ORALLO, J. H. *Análisis y Extracción de Conocimiento en Sistemas de Información: Data Warehouse y Datamining*, 2003.
8. TORRES, L. *BI – Terminología Básica 1a Parte*, 2007. Disponible en: http://www.gravitar.biz/index.php/bi/bi-terminologia-1/
9. DÍAZ, S. O. *Inteligencia de Negocios*, 2009.
10. BERNABEU, D. *Data Mart*, 2009. Disponible en: http://www.dataprix.com/data-warehousing-y-metodologia-hefesto/i-data-warehousing-investigacion-y-sistematizacion-concepto-13
11. KIMBALL, R. and J. CASERTA. *The Data Warehouse ETL Tolkit Practical for Extracting,Cleaning,Conforming, and Delivered Data*. 2004. p.
12. CASTILLO, C. *Sistemas de información. Modelo relacional.*, 2008. Disponible en: http://www.tejedoresdelweb.com/wiki/images/a/a5/Basesdatos_teo5_modelo_relacional.pdf
13. VERÁSTEGUI, H. C. *Modelo Dimensional de Datos*, 2007. Disponible en: http://api.ning.com/files/ei6*2AatbAfAFDwyLdYnMfJ7ZXvd3xwXakwY*gIIpDQk4Ykcb3K9af3G*OvF06uLSTXQwIltUimXRHhSRU8enhjUX2QDs-kR/MODELADODIMENSIONALDEDATOS_V2.pdf
14. DATEC, METODOLOGÍA PARA EL DESARROLLO DE SOLUCIONES DE ALMACENES DE DATOS E INTELIGENCIA DE NEGOCIO EN DATEC, 2008.

Bibliografía

BERNABEU, D. *Data Mart*, 2009. Disponible en: http://www.dataprix.com/data-warehousing-y-metodologia-hefesto/i-data-warehousing-investigacion-y-sistematizacion-concepto-13

BERNABEU, D. *Data Warehouse manager*, 2009. Disponible en: http://www.dataprix.com/data-warehousing-y-metodologia-hefesto/arquitectura-del-data-warehouse/34-Data Warehouse-manager

CASTILLO, C. *Sistemas de información. Modelo relacional.*, 2008. Disponible en: http://www.tejedoresdelweb.com/wiki/images/a/a5/Basesdatos_teo5_modelo_relacional.pdf

DARWIN; I. F.*, et al. TOMCAT 6.0. LA GUÍA DEFINITIVA*. Disponible en: http://www.libros.universia.es/fichalibro.aspx?id=63371

DATEC, METODOLOGÍA PARA EL DESARROLLO DE SOLUCIONES DE ALMACENES DE DATOS E INTELIGENCIA DE NEGOCIO EN DATEC, 2008.

DÍAZ, S. O. *Inteligencia de Negocios*, 2009.

ENGINEERING, N. *La Inteligencia de Negocios*, 2010. Disponible en: http://www.nexteleng.es

EVELIA, C. C. M. *Data Warehouse (Almacenes de Datos)*, 2009. Disponible en: http://hp.fciencias.unam.mx/~alg/bd/dwh.pdf

FERREIRA, K. and J. SCHMIDT Sistemas de Información. Pentaho, 2009.

FOUNDATION, T. A. S. *Apache Tomcat*, 1999-2010. Disponible en: http://tomcat.apache.org/

GALICIA, B. I. *Data Warehouse*, 2007. [2010]. Disponible en: http://www.sinnexus.com/business_intelligence

GROUP, K. Disponible en: http://www.kimballgroup.com

INMON, W. H. Using the Data Warehouse 1992. p.

KIMBALL, R. and J. CASERTA. The Data Warehouse ETL Tolkit Practical for Extracting,Cleaning,Conforming, and Delivered Data. 2004. p.

LÖFBERG, M. and P. MOLIN. Web vs. Standalone Application: A maintenance application for Business Intelligence, Blekinge Institute of Technology, 2005. 40. p.

MARTINEZ, R. Sobre PostgreSQL, 2009. Disponible en :http://www.PostgreSQL-es.org/sobre_PostgreSQL

ORALLO, J. H. Análisis y Extracción de Conocimiento en Sistemas de Información: Data Warehouse y Datamining, 2003.

PENTAHO. Pentaho Analysis Services (Mondrian) 2005-2010. Disponible en: http://mondrian.pentaho.com/

SÁNCHEZ, L. Z. Z. Metodología para el diseño conceptual de almacenes de datos, 2008. p.

S.L., D. I. Soluciones Profesionales Danysoft: Soluciones empresariales de Embarcadero Technologies.

SUMMAN. Pentaho BI Platform Server 2006-2008. Disponible en: http://www.summan.com/index.php/productos/software/pentaho-.html

SYNTEL EAI vs. ETL: Drawing Boundaries for Data Integration, 2007.

TORRES, L. BI – Terminología Básica 1a Parte, 2007. Disponible en: http://www.gravitar.biz/index.php/bi/bi-terminologia-1/

VELASCO, R. H. Almacenes de datos, 2009. [2010]. Disponible en: http://www.rhernando.net/modules/tutorials/doc/bd/dw.html

VERÁSTEGUI, H. C. Modelo Dimensional de Datos, 2007. Disponible en: http://api.ning.com/files/ei6*2AatbAfAFDwyLdYnMfJ7ZXvd3xwXakwY*gIIpDQk4Ykcb3K9af3G*OvF06uLSTXQwIltUimXRHhSRU8enhjUX2QDs-kR/MODELADODIMENSIONALDEDATOS_V2.pdf

WOLFF, C. G. Modelamiento multidimensional.

WOOD, S. and JASPERSOFT. Mondrian Schema Workbench, 2007. Disponible en: http://mondrian.pentaho.com/documentation/workbench.php

Glosario de Términos

UCI: Universidad de las Ciencias Informáticas.

DATEC: Centro de Tecnología de Gestión de Datos.

ONEI: Oficina Nacional de Estadísticas e Información.

SIGOB: Sistema de Información de Gobierno.

DPA: División Político Administrativa

NAE: Nomenclador de Actividad Económica.

TIC: Tecnologías de la Información y las Comunicaciones.

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

Código abierto: término con el que se conoce al software distribuido y desarrollado libremente.

Data Mart: mercado de datos.

Data Warehouse: almacén de datos.

HOLAP: Procesamiento Analítico en Línea Híbrido.

MOLAP: Procesamiento Analítico en Línea Multidimensional.

Open source: Código abierto.

ROLAP: Procesamiento Analítico en Línea Relacional.

BI: Inteligencia de Negocios.

CASE: Ingeniería de Sistemas Asistida por Computadora.

UML: Lenguaje Unificado de Modelado.

ETL: Extracción, Transformación y Carga.

MDX: Lenguaje de consulta a estructuras multidimensionales.