Universidad de las Ciencias Informáticas

Facultad 8

Título: Análisis, Diseño e Implementación del mercado de datos Indicadores relacionados con el Comercio Exterior para la Oficina Nacional de Estadísticas.

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

Autores: Reinier Cárdenas Ramírez.

Daimel Rubén González Alarcón.

Tutor: Ing. Yunesti Pérez La Rosa.

DATEC

Ciudad de la Habana, Junio 2010

"Año 52 de la Revolución Cubana"

DECLARACIÓN DE AUTORÍA

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

Para que así conste firmo la presente a los ____ días del mes de _____ del año 2010.

_____	_____	_____
Firma del Autor	Firma del Autor	Firma del Tutor
Daimel Rubén González Alarcón	Reinier Cárdenas Ramírez	Ing. Yunesti Pérez La Rosa

DATOS DE CONTACTO

Síntesis del Tutor: Graduado en el 2007 de Ingeniero en Ciencias Informáticas de la universidad de Ciencias Informáticas (UCI). Está categorizado docentemente como Instructor. Ha tutorado las tesis: “Sistema de Gestión y Control de Métricas”. 2008. UCI. “Multimedia para el entrenamiento de las acciones ofensivas del Fútbol Sala Femenino”. 2009. UCI.´ Es el Arquitecto principal del Proyecto Productivo: “Deportes de Combate”, 2008-2009. Dos artículos publicados: Propuesta de un Software para la planificación del entrenamiento en el deporte de boxeo. UCIENCIA 2008, Sistema Informatizado de Estudio de Adversarios en el Boxeo. UCIENCIA 2008.

Nombre y Apellidos: Ing. Yunesti Pérez La Rosa

Especialidad: Ingeniero en Ciencias Informáticas.

Ocupación actual: Profesor de Sistemas Operativos y Seguridad Informática, Facultad 8, UCI.

Correo electrónico: yperezla@uci.cu

"Es mejor saber después de haber pensado y discutido que aceptar los saberes que nadie discute para no tener que pensar."

Fernando Savater

AGRADECIMIENTOS

A mi abuela muy especialmente, por enseñarme el camino a seguir durante estos años, te llevo siempre presente.

A mis padres que siempre me han dado todo lo que ha estado en sus manos.

A todo el resto de la familia, en especial a mi hermana, mis tíos, mi primo y mi bisabuela.

A Eric Rey, Raúl García, Juan Enrique Pérez, Héctor René Sánchez, Rayner Lauries y Henry Hechevarría mis hermanos del alma.

A mi tutor, amigo y hermano Yunesti Pérez La Rosa.

A mi compañero de tesis, sin él este resultado no hubiese sido posible.

A mis compañeros de cuarto y a todo el que de una u otra forma a tenido que ver con el presente trabajo, les agradezco de todo corazón.

AGRADECIMIENTOS

Les agradezco en primer lugar a mi familia, mi hermano, mi abuela, mi tía y mi primita; y en especial a mi madre que a pesar de la distancia siempre ha estado cerca, apoyándome en cada paso que doy.

A esa otra familia que es la que se nos permite elegir, mis amigos, con los cuales he disfrutado y sufrido cada etapa.

A mi novia que me ha acompañado y ayudado en las buenas y en las malas. A su familia que me ha acogido como si fuera uno de ellos.

A nuestro tutor por haber jugado perfectamente su papel.

Por último pero no menos importante a mi compañero de tesis porque sin él de una forma u otra este trabajo no se hubiera completado.

DEDICATORIA

A la memoria de mi bisabuelo Otto Pírez Gómez, siempre estarás conmigo.

DEDICATORIA

Dedico este trabajo a mi madre por su apoyo incondicional, su confianza y comprensión.

RESUMEN

El presente trabajo de diploma abordará los temas relacionados con los almacenes de datos y los mercados de datos, así como las principales técnicas para el análisis de la información de carácter estadístico. Plasmará a su vez un estudio detallado de las principales tecnologías y metodologías encargadas del desarrollo de soluciones referentes a los sistemas de almacenamiento y gestión de datos anteriormente mencionados.

Con el objetivo de optimizar y aumentar la velocidad de respuesta a los pedidos que se realicen sobre la información que se necesite consultar, surge como tarea la creación de un almacén de datos. Dicho almacén de datos contendrá varios mercados de datos, y más específicamente el mercado de datos “Indicadores relacionados con el Comercio Exterior”, para la recogida de información estadística relacionada con el comercio exterior en el país. En aras de garantizar un mejor proceso de toma de decisiones para la ONE, se realizará el análisis, diseño e implementación del mercado de datos anteriormente referenciado. Como parte de la solución se hará alusión a los procedimientos para agregar datos con el objetivo de garantizar un rendimiento aceptable a las peticiones que requieran el procesamiento de un elevado número de tuplas. También se expondrán las estructuras dimensionales y la estrategia de indexado y particionamiento a utilizar para un correcto funcionamiento del sistema. Todo lo expuesto con anterioridad será el principal contenido a tratar en el siguiente material.

PALABRAS CLAVE: Almacén de Datos, Mercado de Datos, indicadores.

TABLA DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

INTRODUCCIÓN

Desde hace ya casi medio siglo la humanidad ha sido partícipe del desarrollo vertiginoso de la informática y su vinculación con casi todos los sectores de la sociedad. Muchas son las personas que han invertido su tiempo en hacerla prosperar desde los últimos años del siglo pasado hasta la actualidad. La informática es una ciencia que propone el tratamiento automatizado de la información utilizando dispositivos electrónicos y sistemas computacionales. En sus inicios procesar información mediante el uso de la informática consistía solamente en facilitar trabajos repetitivos y monótonos del área administrativa, lo cual trajo consigo una notable disminución de los costes y un importante incremento de la producción. Ya en la actualidad convergen los fundamentos de las ciencias de la computación, la programación y las metodologías para el desarrollo de software, la arquitectura de computadores, las redes de datos como Internet, la inteligencia artificial, así como determinados temas de electrónica. De ahí su aplicación en numerosas y variadas áreas, como por ejemplo: gestión de negocios, almacenamiento y consulta de información, monitorización y control de procesos, comunicaciones, control de transportes, investigación, desarrollo de juegos, diseño computarizado, aplicaciones y herramientas multimedia. También se ha diversificado en los distintos sectores de la actividad humana como: medicina, biología, física, química, meteorología, ingeniería, industria, investigación científica, comunicaciones y las estadísticas. Hoy día es difícil concebir un área que no use, de alguna forma, el apoyo de la informática; en un enorme abanico que cubre desde las más simples cuestiones hogareñas hasta los más complejos cálculos científicos. Entre las utilidades más importantes de la informática está facilitar información en forma oportuna y veraz, lo cual, por ejemplo, puede tanto facilitar la toma de decisiones a nivel gerencial como permitir el control de procesos críticos. Actualmente es un renglón importante en cualquier economía e incluso el principal en algunos países.

Cuba ha experimentado un importante auge en este sector durante el transcurso de la última década, al punto que se quiere una informatización masiva de toda la sociedad. Siguiendo una estrategia del estado cubano en función de lo antes mencionado y en medio de un sin número de vicisitudes nace la Universidad de las Ciencias Informáticas (UCI), institución que se caracteriza por ser docente y productiva a la vez. En ella se forman profesionales comprometidos con la revolución y altamente calificados en la rama informática. Recientemente la universidad ha realizado algunos cambios organizacionales cuyo destino apunta a lograr sus objetivos de la manera más eficiente posible. En virtud de ello se han creado los Centros de

Desarrollo, es válido mencionar como ejemplo al Centro de Tecnologías de Gestión de Datos (DATEC), el cual brinda su apoyo para el desarrollo de software.

Dicho centro tiene la misión de proveer soluciones integrales y de consultoría relacionadas con las tecnologías de bases de datos y análisis de información. Además, organiza la producción en líneas de trabajo como por ejemplo el Desarrollo de Almacenes de Datos e Inteligencia de Negocio (BI), la cual es responsable de brindarle apoyo a la Oficina Nacional de Estadísticas (ONE) en busca de soporte para organizar y mantener el control de gran cantidad de información que continuamente se manipula en esa entidad.

En el año 1976 se crea el Comité Estatal de Estadísticas, el cual contaba para el procesamiento de la información, con una red encargada del procesamiento de datos a nivel nacional. Sin embargo, no es hasta el 21 de abril de 1994, con la reorganización de los organismos de la administración central del Estado, que aparece lo que actualmente se conoce como la Oficina Nacional de Estadísticas (ONE). La ONE es el organismo rector de las estadísticas en Cuba, dicha oficina es la encargada de captar, analizar y difundir los datos acumulados en todo el territorio cubano. Para ello tiene una serie de modelos estadísticos en los cuales se recoge la información de todos los sectores de la economía y la sociedad. Estos datos son almacenados en formatos de difícil acceso para su consulta, por lo que se hace muy complejo el proceso de acceder y difundir dicha información.

La Oficina Nacional de Estadísticas lleva el registro de todas las instituciones del país. Para almacenar, recuperar y presentar la información proveniente de un modelo estadístico se utiliza un especialista en informática que la introduce en un fichero, utilizando herramientas que dificultan la integración de los principales reportes y cruces de variables, indicadores, tasas, porcentajes y demás aspectos de interés para los órganos del Estado, afectando así el proceso de toma de decisiones. La información contenida por la ONE de los indicadores relacionados con el comercio exterior es almacenada en formato dbf, es por ello que el acceso y propagación de la misma no se realiza de la manera más eficiente. Si un usuario con permisos desea transformar la información, no posee la herramienta más indicada para ello.

En correspondencia con lo planteado anteriormente se plantea como **problema a resolver:** ¿Cómo mejorar el proceso de almacenamiento de datos para optimizar el análisis y la difusión de la información de los indicadores del comercio exterior almacenada en la ONE?

Se define como **objeto de estudio** los Almacenes de Datos y como **campo de acción** los Mercados de Datos estadísticos.

El **objetivo general** de esta investigación es desarrollar el análisis, diseño e implementación del Mercado de Datos para los indicadores de Comercio Exterior para la Oficina Nacional de Estadística.

Como **objetivos específicos** se tienen:

1. Elaborar el marco teórico de la investigación acerca de las principales tendencias de implementación de los Almacenes de Datos, los Mercados de Datos y estudio del arte del tema referente a los indicadores del Comercio Exterior.
2. Realizar el análisis del modelo de control de Comercio Exterior de la ONE.
3. Diseñar el Mercado de Datos para los indicadores del Comercio Exterior para el Almacén de Datos de la Oficina Nacional de Estadísticas.
4. Implementar y cargar los clasificadores para el Mercado de Datos para los indicadores del Comercio Exterior para el Almacén de Datos de la ONE.
5. Validar la solución desarrollada mediante la realización de pruebas.

Para el cumplimiento de los objetivos de esta investigación se plantean las **tareas de investigación** que a continuación se enumeran:

1. Estudio de los temas relacionados con el desarrollo de Mercados de Datos.
2. Estudio de la metodología a utilizar en el desarrollo.
3. Planificación y realización de entrevistas.
4. Identificación de la estructura de usuarios y permisos.
5. Definición de los temas de análisis.
6. Identificación de las necesidades de información, requisitos funcionales y no funcionales.
7. Modelación de los requerimientos.
8. Validación de los requerimientos.
9. Definición de requisitos de entradas y salidas.

10. Elección de la granularidad del proceso del negocio.
11. Definición de las dimensiones del Mercado de Datos.
12. Definición de los hechos asociados a las dimensiones definidas.
13. Estructuración del modelo dimensional.
14. Transformación del modelo dimensional al diseño físico.
15. Implementación del Mercado de Datos.
16. Montaje de los clasificadores para el Mercado de Datos: “Comercio Exterior” para el almacén de datos de la Oficina Nacional de Estadísticas.
17. Realización de las pruebas al Mercado de Datos.

Con el conocimiento de lo anteriormente planteado se ha arribado a la siguiente **idea a defender:** Si se realizara el análisis, diseño e implementación del Mercado de Datos para los indicadores del Comercio Exterior, se lograría una integración más completa de la información y una mejor disponibilidad de la misma en beneficio del proceso de toma de decisiones mediante la informatización de los datos.

La investigación usa como parte de todo trabajo **métodos científicos** que servirán para arribar a los resultados esperados. Dichos métodos no son más que los procedimientos que se encargan de estudiar la realidad, la sociedad, la naturaleza y el pensamiento, para poder describir su esencia y sus principales relaciones. Implica una combinación de inducción y deducción que se retroalimentan.

Como **métodos teóricos** se utilizarán:

- **Análisis histórico-lógico:** Mediante el estudio del estado del arte de los Almacenes de Datos y los Mercados de Datos, se podrá conocer la evolución histórica y el desarrollo actual de los mismos, enfocado en los Modelos Estadísticos para los indicadores del Comercio Exterior en la Oficina Nacional de Estadística.
- **Analítico-Sintético:** Se utilizará para distinguir, extraer y unificar los elementos que forman parte del proceso de construcción del mercado de datos Indicadores relacionados con el Comercio Exterior para la Oficina Nacional de Estadísticas.

Como **método empírico** se utilizará:

- **Entrevista:** Se recurrirá a dicho método con el propósito de conocer las funcionalidades que debe tener el sistema a desarrollar teniendo en cuenta las necesidades del cliente.

El presente trabajo estará compuesto por 3 capítulos:

Capítulo 1: Fundamentación teórica

Se realizará el estudio de Sistemas de Almacenes de Datos y Mercados de Datos, sus principales características así como sus funcionalidades. Se llevará a cabo un estudio del arte de estas tecnologías a nivel mundial y nacional respectivamente. También se definirá la metodología a utilizar para el desarrollo de la solución y las principales herramientas que se utilizarán.

Capítulo 2: Análisis y Diseño

Se plantearán las principales características del sistema a desarrollar y se hará una descripción de la solución definiendo áreas del negocio, la arquitectura, el diseño, la relación entre los principales componentes y por último se dará una propuesta de solución a la problemática.

Capítulo 3: Implementación y Prueba

Se llevará a cabo la implementación del Mercado de Datos y el modelo multidimensional propuesto. Además, se realizará la normalización, el análisis del rendimiento, las pruebas y la validación del sistema, así como el análisis de los resultados obtenidos.

CAPÍTULO 1: FUNDAMENTACIÓN TEÓRICA

1.1. Introducción

Durante el desarrollo del presente capítulo se detallarán diferentes elementos teóricos sobre las tecnologías de Almacenes de Datos y Mercados de Datos. También se analizarán las diferentes metodologías existentes a nivel mundial y se realizará una comparación de estas con el objetivo de justificar la seleccionada. Se seleccionarán las herramientas que se utilizarán para todo el proceso de desarrollo del producto a partir de un análisis profundo, dejando plasmado los elementos relacionados con la elección de cada una de ellas. De igual forma se expondrán las principales características del gestor de bases de datos escogido.

1.2. Almacenes de Datos (AD).

Debido a la gran cantidad de información que se maneja actualmente en los diferentes organismos y empresas del mundo, ha surgido la necesidad de digitalizarla en función de la manipulación eficiente de la misma. A pesar del esfuerzo realizado por los especialistas en la materia el volumen de datos sigue siendo cada vez mayor. Con el objetivo de darle solución a este problema ha surgido el término Almacén de Datos o *DataWareHouse*.

Un Almacén de Datos es una gran colección de datos orientados a temas, integrados, no volátiles e históricos, que recoge información de múltiples sistemas fuentes u operacionales dispersos, y cuya actividad se centra en la “Toma de Decisiones”. Es un sistema computarizado para guardar registros cuya principal finalidad es almacenar información de manera que los usuarios puedan actualizarla y recuperarla.(1; 2)

Una vez reunidos los datos de los sistemas fuentes se guardan durante mucho tiempo, permitiendo el acceso a datos históricos y proporcionando una mayor facilidad a la hora de realizar consultas en función de la toma de decisiones.(3)

Michael J. Corey y Michael Abbey en su libro “*Oracle Data Warehousing*” definen a un Almacén de Datos como:”…una colección de información corporativa derivada directamente de los sistemas operacionales y de algunos orígenes de datos externos. Su propósito específico es soportar la toma de decisiones en un negocio, no las operaciones de un negocio”.(4)

En su libro “*Using the DataWareHouse*” William H. Inmon (1997) define un Almacén de Datos como "una colección de datos, orientados a hechos relevantes del negocio, integrados, que

incluyen el tiempo como característica importante de referencia y no volátiles para el proceso de toma de decisiones".

También se puede citar lo planteado por otro reconocido autor como Ralph Kimball (2002), quien define a un Almacén de Datos como: "…los *DataWareHouse* son una copia de los datos de la transacción estructurados específicamente para la pregunta y el análisis".

Como se ha podido ver, existen disímiles autores y estudiosos del tema que definen de forma muy particular lo que para ellos representa un Almacén de Datos. Sin embargo, no deja de ser un hecho que un AD no es más que un sistema para la colección de datos con respecto a temas específicos, en los que la información va a prevalecer de una manera efectiva a lo largo del tiempo, y su principal razón de existir es la de brindar ayuda a una organización o empresa con respecto a la toma de decisiones.

Características del almacén de datos.

El Almacén de Datos (AD) posee un grupo de características que los distinguen y que están estrechamente relacionadas con su estructura y funcionamiento. Acerca de las mismas se puede plantear que un Almacén de Datos es:(5)

Temático

Porque los datos están almacenados por materias o temas. Estos se organizan desde la perspectiva del usuario final, mientras que en las Bases de Datos operacionales se organizan desde la perspectiva de la aplicación, con vistas a lograr una mayor eficiencia en el acceso a los datos.

Integrado

Porque todos los datos almacenados en el Almacén de Datos están integrados. Las bases de datos operacionales orientadas hacia las aplicaciones fueron creadas sin pensar en su integración, por lo que un mismo tipo de dato puede ser expresado de distintas maneras en dos bases de datos operacionales distintas (Por ejemplo, para representar el sexo: 'Femenino' y 'Masculino', 'F' y 'M' o '0' y '1').

No volátil

Porque únicamente hay dos tipos de operaciones en el Almacén de Datos: la carga de los datos procedentes de los entornos operacionales (carga inicial y carga periódica) y la consulta de los mismos. La actualización de datos no forma parte de la operativa normal de un AD, ya que no

son muy frecuentes los cambios en ellos y por ende se puede mantener por largos períodos de tiempo la información.

Histórico

El tiempo debe estar presente en todos los registros contenidos en un Almacén de Datos. Las bases de datos operacionales contienen los valores actuales de los datos. Un AD no es más que una serie de instantáneas en el tiempo tomadas periódicamente, que permiten mantener y referenciar información.

Diferencia entre Almacenes de Datos y Bases de Datos.

A partir de lo anterior se pueden establecer diferencias sustanciales entre las Bases de Datos Operacionales y los Almacenes de Datos, las cuales apuntan hacia la utilización de los AD como solución a muchos de los problemas que no se resuelven con las Bases de Datos. La siguiente tabla muestra de manera sintetizada estas diferencias.(2)

Base de Datos Operacional	Almacén de Datos
Datos Operacionales	Datos del Negocio para la Información
Orientado a la aplicación	Orientado a sujeto
Actual	Actual + Histórico
Detallada	Detallada + Resumida
Cambia continuamente	Estable

Tabla 1 Diferencia entre Almacenes de Datos y Bases de Datos

Ventajas del uso de Almacenes de Datos.

1. La inversión que realiza una organización para la implantación de un sistema de Almacén de Datos conlleva un coste muy elevado, sin embargo, el retorno de la inversión es garantizado en gran medida.
2. Facilitan el proceso de funcionamiento de las aplicaciones de los sistemas de apoyo a la toma de decisiones de una organización, al poseer una mejor ventaja competitiva con respecto a los demás sistemas de almacenamiento de datos.
3. Los Almacenes de Datos hacen más fácil el acceso a una gran variedad de datos.

4. Se obtiene una Base de Datos clasificada por temas e histórica.
5. Con la utilización de los Almacenes de Datos se logra una mejor integración de la información procedente de múltiples sistemas externos.

Desventajas del uso de Almacenes de Datos.

1. La subestimación del tiempo requerido para extraer, limpiar y cargar los datos en el Almacén.
2. Problemas con los sistemas de origen de los datos.
3. Los datos obtenidos no son suficientes.

1.3. Mercado de Datos (MD).

A través del análisis de diversas fuentes existentes sobre la tecnología de almacenes de datos, muchos autores hacen referencia al término de Mercado de Datos (MD) o *DataMart* cuando se refieren al proceso de almacenar información.

Josep Curto un estudioso en el ámbito de los Almacenes de Datos y Mercados de Datos, define a un MD como: "…un subconjunto de datos del *DataWareHouse* con el objetivo de responder a un determinado análisis, función o necesidad y con una población de usuarios específica".(6)

En su libro Oracle Data Warehousing Michael J. Corey y Michael Abbey (1997) definen a un Mercado de Datos como:…." bases de datos multidimensionales orientadas a una materia específica"(4)

Según Claudia Imhoff en su libro "*Mastering DataWareHouse Design, Relational and Dimensional Techniques*", los Mercados de Datos son un subconjunto de datos de un AD donde se almacenan la mayoría de las actividades de análisis que en el entorno de Inteligencia de Negocio se llevará a cabo.(7)

Los Mercados de Datos (MD) son un subconjunto de datos que pueden funcionar de forma autónoma, o bien enlazado al Almacén de Datos. El motivo por el cual se crean MD es el crecimiento que tiene el almacén y así facilitar su construcción y utilización.(8)

Por lo anteriormente mencionado se puede llegar a la conclusión de que un Mercado de Datos es un componente de un Almacén de Datos pero dedicado a un área más específica, que

contienen información de datos operacionales y ayudan a decidir sobre estrategias del negocio teniendo en cuenta experiencias pasadas.

Las características de los Mercados de Datos son:

- Se centran en los requisitos de los usuarios asociados a un departamento o área de negocios concretos.
- A diferencia de los Almacenes de Datos, los mercados contienen datos operacionales detallados.
- Son más sencillos a la hora de utilizarlos y comprender sus datos, debido a que la cantidad de información que contienen es mucho menor que en los Almacenes de Datos.

Ventajas de utilizar Mercados de Datos: (8)

- Poco volumen de datos.
- Mayor rapidez de consulta.
- Consultas SQL y/o MDX sencillas.
- Facilidad para conservar los datos a través del tiempo.
- Fácil acceso a los datos que se necesitan frecuentemente.
- Crea vista colectiva para grupo de usuarios.
- Facilidad de creación.
- Costo inferior de instalación que la de un almacén de datos.

1.4. OLAP (Proceso Analítico en Línea)

El sistema Proceso Analítico en Línea (OLAP) es una proyección multidimensional redundante de una relación. Al computar todos los *group by* realiza una agregación de sus resultados en un espacio N-dimensional para responder consultas. (9; 10)

OLAP fue presentado en 1993 en el artículo titulado "Providing OLAP to user-analysts: An IT mandate" publicado por Codd y asociados. Según la definición que le dio Codd," es un tipo de

procesamiento de datos que se caracteriza, entre otras cosas, por permitir el análisis multidimensional de datos".

Según Michael J. Corey y Michael Abbey (1997) en su libro"Oracle Data Warehousing", el proceso analítico en línea no es más que:…un tipo de tecnología que permite a los usuarios mejorar la visión que tienen de sus datos de una manera rápida, interactiva y fácil de usar.

Dicho análisis se basa en modelar la información en medidas, dimensiones y hechos. Las medidas son los valores de un dato en particular, las dimensiones son las descripciones de las características que definen dicho dato y los hechos son la definición de una o más medidas para una combinación particular de dimensiones.(3; 10)

En otras palabras una aplicación OLAP permite ver los datos en función de muchas dimensiones. Su importancia viene dada por darles la posibilidad a los usuarios de expresar los datos de la misma forma en que ellos lo piensan.

Entre las funcionalidades que puede ofrecer OLAP y que incluyen declaración de dimensiones y jerarquías, óptima indexación de los datos y definición de operaciones predefinidas de navegación en las dimensiones y de agrupación de medidas, se encuentran las siguientes:(3; 10)

- *Slice-and-dice*: selecciona solo la información de un miembro en particular de una dimensión, o lo que es lo mismo, se trabaja con un subconjunto del total de los datos.
- *Roll-up (drill-up)*: pasa la información del nivel anterior de la dimensión actual a una jerarquía definida, consolidando los datos del nivel actual y mostrando el valor consolidado, correspondiente al nivel inmediatamente superior.
- *Drill-down*: es la operación inversa del Roll-up. Permite ver la información del nivel siguiente de la dimensión actual en una jerarquía definida.
- *Pívot (swap)*: cambia la dimensión que está caracterizando los datos actualmente considerados.
- *Drill-across*: visualiza la información de otro miembro del mismo nivel de la dimensión que se está evaluando. No detalla ni consolida la información, sino que cambia el miembro para el cual se están presentando los datos.

- *Drill-through*: permite consultar información del nivel inferior en la dimensión actual y la navegación por fuera del modelo multidimensional. La ejecución de esta operación depende de poder establecer el acceso al sistema fuente desde el OLAP.

Características de OLAP:(11)

- El acceso a los datos suele ser de sólo lectura. La acción más común es la consulta, con muy pocas inserciones, actualizaciones o eliminaciones.
- Los datos se estructuran según las áreas de negocio, y los formatos de los datos están integrados de manera uniforme en toda la organización.
- El historial de datos es a largo plazo, normalmente, de dos a cinco años.
- Provee análisis multidimensional dinámico, permitiendo a los usuarios finales realizar actividades analíticas y navegacionales, que incluyen cálculo de dimensiones, análisis en períodos de tiempo, visualización de subconjuntos de datos, subir o bajar niveles, comparaciones de varias dimensiones en el área de visualización, entre otros.
- Está basado en una modalidad cliente/servidor multiusuario, que ofrece respuestas rápidas, independientemente del tamaño y la complejidad de la base de datos.
- Solución de manipulación de datos multiusuario de alta capacidad diseñada para soportar y operar en una estructura de datos multidimensional.
- Preparado físicamente para responder rápida y consistentemente a los usuarios finales y/o cargar datos en tiempo real desde las bases de datos.

1.4.1. ROLAP (Procesamiento Analítico en Línea Relacional)

En ROLAP se utiliza una arquitectura de tres niveles. La base de datos relacional maneja el almacenamiento de datos, el motor OLAP proporciona la funcionalidad analítica, y alguna herramienta especializada es empleada para el nivel de presentación. El nivel de aplicación es el motor OLAP, que ejecuta las consultas de los usuarios. La arquitectura ROLAP es capaz de usar datos precalculados (si estos están disponibles), o de generar dinámicamente los resultados desde la información elemental (menos resumida).

Esta arquitectura accede directamente a los datos del Almacén de Datos y soporta técnicas de optimización para acelerar las consultas como tablas particionadas, soporte a la

desnormalización, precalculado de datos, índices, entre otros. Se pueden mencionar como algunos productos que basan sus implementaciones en ROLAP a:(9)

- BusinessObjects [BOS]
- Microstrategy's DSS Agent [MIC]
- Redbrick [RED]
- Oracle Warehouse [ORA]
- DB2 Data Warehouse [DB2]

1.4.2. MOLAP (Procesamiento Analítico en Línea Multidimensional)

Un sistema MOLAP usa una Base de Datos Multidimensional (BDMD), en la que la información se almacena multidimensionalmente. MOLAP utiliza una arquitectura de dos niveles: la BDMD y el motor analítico. La BDMD es la encargada del manejo, acceso y obtención de los datos.

La información procedente de los sistemas transaccionales se carga en el sistema MOLAP. Una vez cargados los datos en la BDMD, se realiza una serie de cálculos para obtener datos agregados a través de las dimensiones del negocio, poblando la estructura de la BDMD; luego de llenar esta estructura, se generan índices y se emplean algoritmos de tablas hash para mejorar los tiempos de accesos de las consultas.(9)

1.4.3. HOLAP (Procesamiento Analítico en Línea Híbrido)

Solución OLAP híbrida que combina el uso de las arquitecturas ROLAP y MOLAP. En una solución con HOLAP, los registros detallados (los volúmenes más grandes) se mantienen en la base de datos relacional, mientras que los agregados lo hacen en un almacén MOLAP independiente (Ibarzábal, 2003). Un sistema HOLAP resuelve el problema de dispersión, dejando los datos más granulares (menos agregados) en la base de datos relacional, pero almacena los agregados en un formato multidimensional, minimizando así la presencia de celdas vacías.(9)

1.5. Componentes de un Almacén de Datos

La composición de un Almacén de Datos está definida por una serie de elementos que muestran de manera general el ambiente de estos. Es irrefutable que la construcción de un AD está justificada por la necesidad en específico de cada empresa u organización que lo requiera.

Generalmente ellos cumplen con un estándar en específico. A continuación se hace referencia a los componentes de un AD.

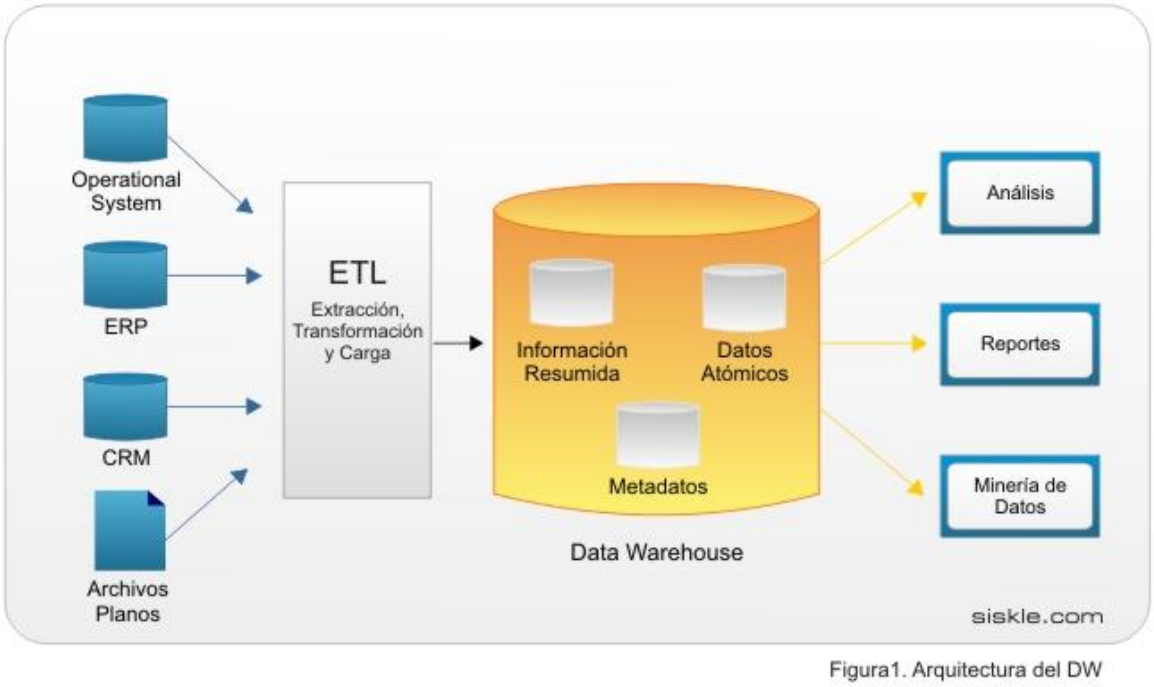

Figura1. Arquitectura del DW

Figura 1 Arquitectura de un Almacén de Datos

1.5.1. Sistema de fuentes operacionales

Con el objetivo de gestionar las transacciones que se hacen diariamente en las empresas aparecen los sistemas de fuentes operacionales. Dichas transacciones pueden ser almacenadas con distintos formatos que suelen ser vistos de distintas maneras, desde una base de datos relacional hasta ficheros como Excel, xml, dbf, texto plano y muchos más.(12)

Por no tenerse control o casi ningún control sobre el volumen y formato de los datos de estas fuentes, estos sistemas están localizados fuera del depósito central. Estos componentes tienen una importancia vital a la hora de evaluar el rendimiento y la disponibilidad de la información y funcionan con el objetivo de ir realizando salvas a la información que se encargan de gestionar.

1.5.2. Área de procesamiento (Staging Área)

El sistema que se encuentra entre las fuentes de datos y el Almacén de Datos es llamado Área de Procesamiento (Staging Área). Tiene entre sus objetivos facilitar la extracción de los datos desde las fuentes orígenes. El área de procesamiento realiza lo que se conoce como data cleansing (limpieza de datos) con el fin de mejorarlos. Se usa también para acceder con un nivel detallado a la información que no está contenida en el AD.(12)

Robert Wrembel y Christian Koncilia en su libro "*DATAWAREHOUSES AND OLAP*" hacen referencia a este componente como: "…área donde los datos extraídos por los procesos ETL son transformados y limpiados antes de ser cargados dentro del Almacén de Datos".

Finalmente, se puede concluir definiendo que el Staging Area es el componente de un Almacén de Datos donde los datos son almacenados temporalmente y que realiza un conjunto de procesos como los de Extracción, Transformación y Carga (ETL).

Extracción, Transformación y Carga de los Datos

Al proceso de analizar la información consolidada en un Almacén de Datos y que implica actividades de extracción de diversas fuentes de datos, transformación de la información necesaria y finalmente su carga se le denomina proceso ETL (Extracción, Transformación y Carga). Otro aspecto de mucha importancia y que provee los procesos ETL es la limpieza de los datos, que es la habilidad de chequear filtrar y corregir los errores que puedan ser encontrados en los datos. Entre los principales productores actuales de herramientas ETL destacan Oracle con *OracleWarehouseBuilder*, Microsoft con *MicrosoftwithDataTransformation Services* e IBM con *DataWareHouse* Center. A continuación se describe más detalladamente dicho proceso.(10; 13)

- Extracción.- Acción de obtener la información deseada a partir de los datos almacenados en fuentes externas.
- Transformación.- Cualquier operación realizada sobre los datos para que puedan ser cargados en el Almacén de Datos o se puedan migrar de este a otra base de datos.
- Carga.- Consiste en almacenar los datos en la base de datos final.

El flujo de trabajo que representa a los procesos de extracción, transformación y carga (ETL) está compuesto por varias funcionalidades que a continuación se mencionan:(10)

- Identificación de la información relevante a los recursos externos.

- Extracción de la información identificada.
- Transportación de la información a la Staging Área.
- Transformación de la información proveniente de múltiples recursos a un formato común.
- Limpieza de los datos entrados a la base de datos.
- Propagación de los datos al Almacén de Datos y actualización del Mercado de Datos.

1.5.3. Área de Presentación

Se ha definido como Área de Presentación al componente de un Almacén de Datos donde los datos significativamente importantes para la toma de decisiones están organizados, almacenados y disponibles para la consulta o reporte que deseen hacer los usuarios. Según Kimball en esta área las estructuras aparecen como esquemas dimensionales llamados esquemas en estrella.

1.5.4. Herramientas de acceso a datos

Son vistas como la variedad de capacidades que se les proveen a los usuarios del negocio para la toma de decisiones. Básicamente son herramientas que permiten la consulta del área de presentación de un Almacén de Datos. Puede ser herramientas de consultas muy simples o tan complejas y sofisticadas como una aplicación de modelado o de Minería de Datos.

1.6. Modelo Entidad-Relación y Modelo Dimensional

1.6.1. Modelo Entidad-Relación

Introducido por Peter Chen en 1976 los modelos Entidad-Relación (E/R, por sus siglas más populares) es un lenguaje para el modelado de los datos dentro de un sistema de información. Es muy útil para la creación de esquemas conceptuales de bases de datos. En ellos los datos son fraccionados en muchas entidades donde estas posteriormente se convierten en una tabla física dentro de la base de datos.

Dentro de este modelo hay una gran simetría por lo parecido de sus tablas. Esto da al traste con que no se pueda identificar que tabla tiene mayor importancia que otra. Otra desventaja

que poseen es que no existe una manera eficiente de diferenciar cuales son las tablas que contienen medidas numéricas con respecto a las que incluyen información estática.

De manera general los modelos E/R no son recomendables para el diseño de un Almacén de Datos por no garantizar una óptima recuperación de la información almacenada. Las consultas muy complejas que abarcan una gran cantidad de registros y tablas hacen de estos modelos, algo muy difícil de entender para los usuarios.

1.6.2. Modelo Dimensional

Cuando se hace referencia a los Almacenes de Datos es imposible obviar el Modelo Dimensional. A diferencia de los sistemas de bases de datos más convencionales, un Almacén de Datos requiere de un Modelo Dimensional para su diseño. Estos modelos poseen en su estructura la misma información que uno Entidad-Relación pero la diferencia entre ellos es la manera de organizarla.

El modelado dimensional se puede adaptar a dos entornos principales: el relacional y el dimensional (ROLAP y MOLAP) y es una técnica que hace más entendible para el usuario la base de datos. Existen dos tipos de tablas fundamentales en este modelo:(14; 15)

- Tablas Dimensionales (Lock table): son las que se conectan a las tablas de hechos (Fac table). Almacenan valores que están relacionados con una dimensión en particular y están compuestas por una clave primaria.
- Tabla de Hechos (Fac table): tabla central en el esquema dimensional, contiene valores de las medidas de los negocios.

Tiene distintas formas de representación basado en esquemas como son:(14; 15)

- Esquema estrella: compuesto por una tabla central y un conjunto de tablas que la atienden respectivamente.
- Esquema copo de nieve (snowflake): en este esquema las tablas de dimensiones están normalizadas, con respecto al esquema estrella.

1.6.2.1 Tabla de Dimensiones

Estas acompañan a la tabla de hechos y determinan los parámetros (dimensiones) de los que dependen los hechos registrados en la tabla de hechos. Son elementos que contienen atributos

(o campos) que se utilizan para restringir y agrupar los datos almacenados en una tabla de hechos cuando se realizan consultas sobre dichos datos en un entorno de Almacén de Datos o Mercado de Datos. Los atributos de las dimensiones sirven como fuente primaria de las restricciones de las consultas, agrupaciones y las etiquetas de los reportes.(1; 10)

1.6.2.2. Tabla de hechos

Todo Almacén de Datos tiene incluido una o varias tablas de hechos, eso depende en gran medida de su complejidad. Estas tablas básicamente capturan los datos encargados de medir las operaciones de un equipo. Contienen datos numéricos (hechos) que proporcionan información sobre el historial de la organización.

Están compuestas por un gran número de filas en dependencia de los años de trabajo de un proyecto, y su capacidad de almacenar información a un nivel elevadísimo de detalle sobresale como una de sus principales características. No deben incluir datos que no procedan de los campos numéricos y los campos de índice que guardan relación entre los hechos con las entradas en las tablas de dimensiones correspondientes.(10; 16)

1.7. Estado actual de los Almacenes de Datos y Mercados de Datos

Mundialmente hoy día la competencia sigue siendo uno de los temas más tangibles entre las principales empresas y compañías del mundo. La sed de poder y el deseo de ser cada vez más exitosas en el ámbito de las tecnologías y las comunicaciones, trae consigo un aumento de la necesidad de incorporar lo más reciente y avanzado en el campo informático. Debido a que la economía actual se centra específicamente en las necesidades más importantes del cliente, el avance tecnológico ha dado un vuelco total y se ha enfocado más seriamente en estas exigencias.

Cambiar la concepción del almacenamiento de datos se hace cada vez más importante. El uso del modelo relacional no es ya la solución más factible; pues los sistemas OLTP (Procesamiento de Transacciones En Línea) trabajan con transacciones diarias de diferentes departamentos pero no es común el trabajo con los datos históricos. Estos sistemas no han sido construidos para resolver los principales problemas de las funciones de análisis, síntesis y consolidación de datos y de ahí su actual desuso. Basado en esto, ha comenzado sin lugar a dudas una nueva era para enfocar los esfuerzos hacía una mayor utilización de los Almacenes de Datos.

1.7.1. En el mundo

En los orígenes del Almacén de Datos y de las herramientas de Inteligencia del Negocio, la implantación de estos era muy esporádica. Al hacer alusión a los costos de inversión y mantenimiento, y de diversas licencias de software como el Sistema Gestor de Base de Datos, las empresas dudaban constantemente en enfocarse a estas técnicas tan novedosas. Todo esto, unido al cambio de filosofía de trabajo de los usuarios tan necesaria para el funcionamiento de estos sistemas, dio al traste con la utilización de estas técnicas. Solo empresas con una fuerte orientación del marketing fueron capaces en un principio de dar los primeros pasos hacia el tema.

No obstante, con el transcurso del tiempo y con el crecimiento de la información en distintas fuentes de datos la respuesta no se hizo esperar, y los directivos de las empresas se dieron cuenta de que era insoslayable la utilización de los Almacenes de Datos y las herramientas de Business Intelligence. Actualmente existen muchas corporaciones que manejan grandes cantidades de datos como son el caso de la Adida, Hewlett Packard y Sun, dichas empresas construyen sus sistemas operacionales con herramientas propietarias.

En la industria minorista son utilizados también hoy día los Almacenes de Datos. American Stores (E.U.A), Canadian Tyre (Canadá), WH Smith Books (Gran Bretaña), Great Universal (Gran Bretaña), Supermercados Casino (Francia), son algunos ejemplos de industrias de este tipo que hacen uso de esta tecnología. (17)

Algunas empresas muy conocidas como Coca Cola, Nike, Hallmark, Maybelline, Karsten Ping Golf Clubs y Walt Disney también utilizan Almacenes de Datos. (17)

No solo en países desarrollados se dan avances con respecto al tema de los Almacenes de Datos. En América se pueden mencionar muchos ejemplos de empresas como Telefónica de Argentina, Visa, Arcor, Wallmart, Procter & Gamble, TV Azteca, Baxter, entre otras que han incorporado el uso de los AD para el proceso de toma de decisiones. (18)

El uso de los Almacenes de Datos se ha extendido a muchas esferas, ya no están concebidos solo en la economía. La salud, las telecomunicaciones y los sistemas bancarios son ejemplos fehacientes que demuestran que los AD han ido tomando cada día más fuerza y ya no pasarán desapercibidos como antes.

En la telecomunicación se usan para la monitorización de los clientes, los cobros y pagos, el marketing y los servicios. Otro ejemplo a mencionar lo constituye Bouygues Telecom,

considerada como la tercera compañía inalámbrica más grande de Francia, Jazztel, Vodafone, France Telecom y la compañía de Radio y Televisión de Galicia (CRTVG), como algunos exponentes en este campo tan moderno e importante.

La salud se ha visto representada también y ese es el caso de La Congregación de Hermanas Hospitalarias una organización internacional que se dedica a la asistencia médica en 24 países de Europa, Asia, América y África. Ellos utilizan un Almacén de Datos para el control de patrones de comportamiento y para darles el adecuado seguimiento a pacientes enfermos.

Prestigiosos bancos del mundo utilizan Almacenes de Datos para gestionar su información. Entidades como Banco París de Francia, European Central Bank de la Union Europea, BBVA considerado el segundo banco más grande de España, Caja Madrid, Caja Extremadura también españoles hacen uso de ellos.

1.7.2. En Cuba

En Cuba el uso de estos sistemas de almacenamiento de información aún se consideran muy incipientes. No obstante, el estado cubano y el Ministerio de la Informática y las Comunicaciones siguen haciendo esfuerzos y demostrando su interés en dar pasos sólidos hacia una futura migración a esta tecnología por sus demostradas ventajas.

La corporación CIMEX es un ejemplo de uso de Almacén de Datos. Dedicada a la Exportación e Importación de productos centra básicamente su atención a la actividad del comercio. En la feria Informática 2002 se mostró entre sus productos un AD para Cubacel desarrollado con plataforma Oracle obteniendo resultados satisfactorios. Otras empresas que actualmente se inclinan al uso de estos sistemas son UNION y CUPET, las cuales se encuentran en pleno desarrollo de sus AD respectivamente.

1.8. Herramientas a utilizar en la investigación

1.8.1. Sistemas Gestores de Bases de Datos

Se considera un Sistema Gestor de Base de Datos a un conjunto de programas que ayudan a administrar la información que está contenida en una base de datos. Estos programas se encargan de la integridad y la seguridad de los datos; y garantizan la interacción con el sistema operativo. Entre los tipos de gestores de bases de datos que existen actualmente se pueden mencionar a los relacionales, jerárquico y de red; pero el más utilizado es el relacional (SGBDR). (19)

Los Sistemas Gestores de Bases de Datos definen además lenguajes para permitir a los administradores de la base de datos especificar los datos que componen la base de datos. Estos lenguajes se pueden clasificar en Lenguaje de definición de datos (LDD o DDL), y en Lenguaje de manipulación de datos (MDL o DML). Algunas características que se pueden mencionar de los Sistemas Gestores de Bases de Datos son las siguientes:

- Son capaces de controlar la concurrencia y las operaciones que implican la recuperación de fallos.
- Definen usuarios y sus restricciones de acceso.
- Respetan la integridad y seguridad de los datos.
- Toleran definiciones de esquemas y vistas.

1.8.1.1. Sistema Gestor de Bases de Datos MySQL

Es un Sistema Gestor de Base de Datos Relacional muy conocido y usado por su rendimiento, rapidez, facilidad para ser aprendido y sencillez al ser utilizado, siendo frecuente su utilización en aplicaciones web y plataformas. MySQL es un programa de intercambio que permite conectarse a servidores MySQL, para la realización de consultas y obtención de los resultados. Suele ser combinado para trabajar con PHP, siendo esta mezcla bastante segura.(20)

Entre las principales características de este SGDB se encuentran:

- Es una aplicación sencilla de instalar y configurar en servidores tanto Windows como Linux.
- Dispone de varias funciones exigidas por desarrolladores, como compatibilidad para la mayoría de las partes de SQL ANSI21, absoluta compatibilidad con ACID (Atomicidad, Consistencia, Aislamiento y Durabilidad), duplicación, volcados online e integración con gran parte de los entornos de programación y funciones SSL.
- Tiene gran facilidad de portabilidad pues se ejecuta prácticamente en casi todos los sistemas operativos y los datos pueden ser transferidos de un sistema a otro sin dificultad.
- Es fácil su utilización y administración.

1.8.1.2. Sistema Gestor de Bases de Datos Oracle

Es un Sistema Gestor de Base de Datos Relacional que utiliza los recursos de los sistemas informáticos en las arquitecturas de hardware, garantizando así un gran aprovechamiento en entornos con grandes volúmenes de información, ya que posee la capacidad de almacenar y acudir a los datos de forma recurrente. Este SGDBR es una herramienta cliente-servidor considerándose como unos de los Sistemas de Base de Datos más completos, orientada al acceso remoto, redes y multiusuario. Oracle es un sistema multiplataforma y se encuentra entre los más usados.(21)

Entre sus principales características se encuentra su seguridad, garantizando la autenticidad apropiada de los usuarios y la privacidad e integridad de la información. Posee lectura de multiversión, proporcionándoles a los usuarios respuestas consistentes. Oracle es configurable en ambientes OLTP (Procesamiento de Transacciones en línea), paralelos, clúster, también es una buena solución a nivel de Almacén de Datos. Este está desarrollado para la plataforma Unix, pues soporta gran parte de la carga de los sistemas a nivel mundial, así como un sistema abierto y configurable.(21)

1.8.1.3. Sistema Gestor de Bases de Datos SQLite

Este Sistema Gestor de Bases de Datos relacional creado por el D. Richard Hipp, es compatible con ACID (Atomicidad, Consistencia, Aislamiento y Durabilidad) y está contenido en una pequeña librería de aproximadamente 500kb. Entre sus principales ventajas es válido mencionar que puede trabajar tanto cargado en memoria como en disco, brindando la posibilidad de pasar la base de una modalidad a otra.(22)

Tiene como principal diferencia con los demás gestores que su motor no es un proceso independiente con el que se comunica el programa principal. Esta característica le permite a la librería SQLite enlazarse con el programa para formar parte del mismo. A través de llamadas simples a subrutinas el programa utiliza la funcionalidad de SQLite, reduciendo considerablemente la latencia en el acceso a la base de datos, principalmente porque las llamadas a funciones son más eficientes que la comunicación entre procesos. La versión tres de SQLite llega a permitir Bases de Datos con un tamaño que puede llegar hasta dos terabytes.(22)

1.8.1.4. Sistema Gestor de Bases de Datos PostgreSQL

Creado por el proyecto POSTGRES de la universidad de Berkeley, PostgreSQL es un Sistema Objeto-Relacional, ya que incluye características de la orientación a objetos, como puede ser la herencia, tipos de datos, funciones, restricciones, disparadores, reglas e integridad transaccional. (23)

Ventajas:

- DBMS (Sistemas Manejadores de Bases de Datos) Objeto-Relacional. Aproxima los datos a un modelo Objeto-Relacional, y es capaz de manejar complejas rutinas y reglas. Ejemplos de su avanzada funcionalidad son consultas SQL declarativas, control de concurrencia multi-versión, soporte multiusuario, transacciones, optimización de consultas, herencia, y arreglos.(23)
- Cliente/Servidor. Usa una arquitectura proceso por usuario cliente/servidor. Hay un proceso maestro que se ramifica para proporcionar conexiones adicionales para cada cliente que intente conectarse a PostgreSQL.(23)
- Altamente Extensible. Soporta los tipos de datos base, así como: tipo, fecha, monetarios, elementos gráficos, datos sobre redes (MAC, IP…), cadenas de bits, etc. Además, operadores, funciones, métodos de acceso y tipos de datos definidos por el usuario.(23)
- Soporte SQL Compresivo. Soporta la especificación SQL99 e incluye características avanzadas tales como las uniones (joins) SQL92.(23)
- Integridad Referencial. Es utilizada para garantizar la validez de los datos de la base de datos.(23)
- Lenguajes Procedurales. Tiene soporte para lenguajes procedurales internos, incluyendo un lenguaje nativo denominado PL/pgSQL. Este lenguaje es comparable al lenguaje procedural de Oracle, PL/SQL. Además, tiene la habilidad para usar Perl, Python, o TLC como lenguaje procedural embebido.(23)
- MVCC (Multi-Version Concurrency Control) Control de Concurrencia Multi-Versión. Es la tecnología que PostgreSQL usa para evitar bloqueos innecesarios, es decir, permite la lectura sin que sea bloqueada por los que escriben que están actualizando registros.(23)
- Write Ahead Loging (WAL). Esta característica incrementa la dependencia de la base de datos al registro de cambios antes de que estos sean escritos en ella. Esto garantiza que en caso de que la base de datos se caiga, existirá un registro de las transacciones a partir del cual se podrá restaurar la base de datos desde el punto en que se quedó.(23)

- Es un gestor bajo licencia Berkeley Software Distribution (BSD), que posee una gran escalabilidad, haciéndolo idóneo para su uso en sitios web. Además, por su arquitectura de diseño, escala muy bien al aumentar el número de CPUs y la cantidad de RAM.(23)
- Sus tablas pueden llegar a 32 TB, sus tuplas 1.6 TB y los campos a 1GB de tamaño respectivamente.(23)
- El tamaño de la base de datos es ilimitada.(23)

Desventajas

- Consume bastantes recursos y carga con mucha facilidad el sistema.
- Velocidad de respuesta un poco deficiente al gestionar bases de datos relativamente pequeñas, aunque esta misma velocidad la mantiene al gestionar bases de datos realmente grandes.

1.8.2. Justificación del Sistema Gestor de Bases de Datos a utilizar

Actualmente la Oficina Nacional de Estadísticas se encuentra en pleno proceso de migración hacia una independencia tecnológica. Sus dispositivos de almacenamiento están siendo enfocados hacia la plataforma PostgreSQL. Por ser una propuesta Código Abierto (Open Source) que sobrepasa a muchas propietarias y por su gran potencialidad y adaptabilidad al problema en cuestión, se ha elegido como Sistema Gestor de Base de Datos PostgreSQL.

1.8.3. Herramientas de modelado

Las herramientas CASE (Ingeniería de Software Asistida por Ordenador) son de gran ayuda para el desarrollo de un software. Estas son un grupo de programas que utilizan las personas que intervienen en el desarrollo de un software, como es el caso de los diseñadores, desarrolladores, analistas, entre otros, durante las fases del desarrollo del software (análisis, diseño, implementación, etc.), para agilizar y facilitar el trabajo, ya que dichas herramientas proveen de métodos, técnicas y utilidades que ayudan al perfeccionamiento del desarrollo de sistemas de información, de forma total o parcial.(24)

1.8.3.1. Herramienta de modelado ERWIN

ERWIN es una CASE empleada en el diseño de base de datos, con la cual se puede hacer un trabajo productivo por las facilidades que brinda para la generación y mantenimiento de aplicaciones de forma sencilla para el diseñador. ERWIN permite realizar el diseño del modelo

lógico de los requerimientos de información, así como el diseño del modelo físico, ya con nivel mayor, refinando las características de la base de datos diseñada.

Con esta herramienta CASE es posible visualizar la estructura de la base de datos diseñada, lo cual tiene como ventaja que el diseñador pueda observar en su totalidad el trabajo realizado, para realizar un análisis y si fuera necesario hacer cambios en busca de optimizar el diseño final. Esta herramienta permite generar, de forma automática, las tablas y el código referente a los procedimientos almacenados (stored procedure) y disparadores (triggers) para los principales tipos de bases de datos.

1.8.3.2. Herramienta de modelado Rational Rose Enterprise Edition

Es una de las herramientas más utilizadas y eficaces de modelado visual para el análisis y diseño de sistemas orientados a objetos. Es utilizada para el modelamiento de los sistemas antes de comenzar a desarrollarlos. Soporta todo el ciclo de vida de un proyecto. Esta permite el completamiento de varias partes de los flujos de trabajos principales de RUP (Proceso Unificado de Desarrollo Software).(25)

En cuanto a las tecnologías de almacenamiento de datos esta herramienta acelera el diseño de bases de datos por su entorno de modelado tan sofisticado y una gran flexibilidad entre los modelos lógicos y físicos. Le facilita una panorámica a los desarrolladores de bases de datos de cómo accederá la aplicación a la base de datos.(26)
Entre las principales características de esta aplicación se encuentran:(25)

- Permite analizar y diseñar el sistema antes de iniciar con el código.
- Mantiene la solidez de los modelos del sistema de software.
- Contiene chequeo de la sintaxis UML.
- Permite la generación y documentación automática.
- Permite generación de código a partir de los modelos.

1.8.3.3. Herramienta de modelado ER/Estudio

Es una herramienta para el modelado de las bases de datos fácil de usar, ofrece capacidades de diseño lógico y la construcción automática de las bases de datos, posee abundante documentación y fácil creación de los reportes. La realización de los diagramas es rápida y clara, permitiendo la documentación del mismo. ER/Studio permite realizar cambios en los

datos del modelo de diseño de forma directa a la base de datos. En su versión 8.0, ER/Studio aborda temas importantes y posee un repositorio para el modelado.(27)

Con esta herramienta se puede realizar reingeniería inversa de la base de datos. Para un Almacén de Datos ofrece la documentación visual del linaje de datos y el modelado dimensional de los modelos lógico y físico. Una de sus versiones más recientes lanzada al mercado el 18 de febrero del 2009 llamada Embarcadero All-Access soporta una amplia variedad de plataformas y lenguajes como Oracle, IBM® DB2, Sybase®, Microsoft® SQL Server, InterBase®, MySQL®, Java, Microsoft .NET, VCL, Windows®, Linux®, Mac OS®, SQL, UML®, XML, HTML, C++, Delphi® y PHP. Su precio de licencia está fijado en $2,250.00 para un usuario.(28)

1.8.3.4. Herramienta de modelado Visual Paradigm

Es una herramienta CASE que utiliza lenguaje de modelado UML y soporta las etapas del ciclo de vida completo de desarrollo de software. Permite la ingeniería inversa, generación de código, importación desde Rational Rose, exportación/importación XMI, generador de informes, editor de figuras y otros elementos particulares.(29)

Visual Paradigm es un entorno de creación de diagramas para UML. El mismo utiliza el diseño centrado en casos de uso y enfocado al negocio que genera un software de mayor calidad. Usa un lenguaje estándar común a todo el equipo de desarrollo que facilita la comunicación. También contiene un modelo y código que permanece sincronizado en todo el ciclo de desarrollo. Presenta disponibilidad de múltiples versiones, disponibilidad para integrarse en los principales IDEs que existen en la actualidad y disponibilidad en múltiples plataformas.(29)

1.8.4. Justificación de la herramienta de modelado a utilizar

A pesar de no ser Visual Paradigm una herramienta de plataforma libre como el Gestor de Base de Datos PostgreSQL, se ha tenido en cuenta a la hora de su selección el hecho de que la Universidad de las Ciencias Informáticas (UCI) paga la licencia de este producto. Proporciona un entorno amigable e interactivo al usuario y permite la exportación de script de una manera íntegra y eficiente.

1.9. Metodologías existentes para el desarrollo de un Mercado de Datos

En la producción de software es primordial el uso de las metodologías, con el fin de organizar, planificar y controlar el proceso de desarrollo de estos, sirviendo estas de guía para mejorar la productividad en el desarrollo y calidad del software. Para el desarrollo e implementación de las soluciones de Almacenes de Datos e Inteligencia de Negocio se han creado distintas metodologías.

Existen dos juicios que marcan su tendencia de desarrollo conocidas como; Metodología de Kimball y Metodología de Inmon, en honor de sus creadores Ralph Kimball y William H. Inmon respectivamente. La diferencia existente entre ambas está basada en la forma de enfrentar el problema.

Inmon se basa en un enfoque descendente (top-down), propone construir primero el Almacén de Datos y a partir de este los Mercados de Datos (MD). Afirma, que la creación de una base de datos relacional con una ligera normalización precisa ser las bases para los MD.(30)

Kimball se basa en dividir el mundo de Inteligencia de Negocio entre los hechos y las dimensiones. Defiende por tanto una metodología ascendente (bottom-up) a la hora de diseñar un almacén de datos. Plantea que se debe crear por cada departamento un conjunto de MD independientes orientados a los temas que estén relacionados con él.(30)

Sin embargo, estas dos metodologías no son las únicas existentes en el ámbito de Almacenes de Datos y Mercados de Datos.

Otra metodología es DM2 que se enfoca en las necesidades de información a nivel gerencial (atiende las necesidades a nivel ejecutivo y gerencial). Semejante a la forma top down que propone Inmon acorta considerablemente el tiempo entre el inicio y la implantación del análisis.

En 1996 se presenta como una herramienta industrial y de aplicación neutral CRISP-DM. Basada en un modelo de proceso jerárquico, consiste en un grupo de tareas marcadas en cuatro niveles de abstracción. Estos niveles son: fase, tarea genérica, tarea especializada e instancia de procesos.

En su tesis de doctorado Metodología para el Diseño Conceptual de Almacenes de Datos, Leopoldo Zenaido Zepeda Sánchez explica su punto de vista y realiza aportes importantes y novedosos en el tema de desarrollo de estructuras de almacenamiento de datos. Incorpora los casos de uso para guiar el proceso de desarrollo del software y define las principales

transformaciones que se deben hacer para llevar un diagrama relacional a uno dimensional y de esta forma obtener las estructuras que conformarán el repositorio de datos.(31)

1.9.1. Justificación de la metodología a utilizar

Por ser la Oficina Nacional de Estadísticas el órgano rector de las estadísticas en Cuba y por la importancia de la información que maneja, se hace necesario la elección de una metodología lo más robusta y adaptable posible.

Teniendo en cuenta la problemática que se plantea para la realización de este trabajo, y después de un minucioso estudio fue seleccionada como base para definir la metodología de desarrollo a utilizar, una metodología híbrida pasada fundamentalmente en una adaptación de la Metodología de Ralph Kimball, decisión tomada por las razones siguientes:

- Elabora los conceptos de Hechos y Dimensiones, lo que es efectivo en la toma de decisiones y suministra superior rapidez en el proceso de desarrollo.
- Expone ir montando el Almacén de Datos a través de la construcción de los Mercados de Datos departamentales, siendo esto una buena estrategia y coincide con la división lógica de las empresas.
- Existe una amplia documentación sobre la misma, la respuesta a todas las dudas y preguntas que puedan surgir se pueden encontrar en la web.
- Es una metodología madura y reconocida por la comunidad dedicada al tema; tiene bien definidas las etapas, actividades, artefactos y roles.

Para la perfección de la misma se tomó lo esbozado por Leopoldo Zenaido Zepeda Sánchez en su Tesis de Doctorado, orientando el trabajo a los Casos de Uso. (31)

Conclusiones del Capítulo

A lo largo de este capítulo se ha explicado y detallado la base sobre la que se trabajará en aras de confeccionar un Mercado de Datos que sea capaz de darle solución a la problemática de la Oficina Nacional de Estadísticas. Después de haber realizado el estado del arte concerniente a ellos se concluye lo siguiente:

- Se decide como tecnología más apropiada, llevar a cabo el desarrollo de un Mercado de Datos para la ONE en el área del Comercio Exterior.
- Con la adaptación de la metodología de Kimball propuesta por el grupo de Almacenes de Datos e Inteligencia del Negocio se facilitará y agilizará el trabajo durante el desarrollo de la solución propuesta.
- Las herramientas de desarrollo que se utilizarán serán el Sistema Gestor de Base de Datos PostgreSQL en su versión 8.4, y para desarrollar el diseño de la solución se ha seleccionado al Visual Paradigm.

CAPÍTULO 2: ANÁLISIS Y DISEÑO

2.1. Introducción

Los Almacenes de Datos como ha sido referenciado con anterioridad en el capítulo 1 están formados por varios elementos; las fuentes de datos externas, el área de procesamiento (Staging Area), el área de presentación y las herramientas de acceso a datos son algunos de ellos. Cada uno de estos conllevan una serie de responsabilidades específicas propias y trabajan de manera semi-independiente aportando así en su conjunto resultados tangibles dentro de la solución. Además, permiten lograr un diseño robusto y que sea adaptable a las condiciones de los usuarios.

Existen 3 grupos fundamentales en los que se pueden dividir dichos elementos, ellos son:

- Diseño dimensional de estructuras.
- Extracción, Transformación y Carga (ETL) de los datos de las fuentes.
- Inteligencia de Negocio.

Se hace imprescindible también la definición de las distintas áreas de análisis existentes, así como el proceso del negocio a modelar en la solución, por lo que estos aspectos unidos a un correcto diseño de la solución serán los principales temas a tratar dentro del presente capítulo.

2.2. Análisis.

2.2.1. Definición del Negocio.

La ONE es el organismo rector de la estadística en Cuba. Comprende la elaboración de estadísticas y análisis del Estado y del Gobierno a los efectos de conocer el comportamiento de los procesos económicos, demográficos y sociales y, especialmente para el control del plan de economía nacional y del presupuesto, los compromisos estadísticos internacionales, la población y otras instituciones.

2.2.2. Temas de análisis.

La identificación de los temas de análisis es de suma importancia para el desarrollo del Mercado de Datos, estos permiten la factibilidad, utilidad, y éxito de las estructuras que se están diseñando. Básicamente la ONE maneja toda la información relacionada con las exportaciones y las importaciones, es por ello que se ha identificado a los indicadores para el

Comercio Exterior como el principal tema de información ó análisis. Dentro de este tema y como anteriormente se ha descrito surgen dos tipos de información, ellos son:

- Análisis de las exportaciones.
- Análisis de las importaciones.

2.2.3. Roles y permisos.

La ONE publica la gran mayoría de la información que maneja. Sin embargo, hay ciertos datos que no se publican y para estos se han definido dos roles fundamentales. El analista encargado de verificar y analizar la información, es el de menor privilegio, a este rol no le es permitido la transformación de los datos almacenados solo puede ser capaz de consultarlos. Para los derechos de transformación de la información contenida en el Mercado de Datos se ha definido al Administrador, quien lleva a cabo los procesos ETL y a su vez goza de los permisos definidos anteriormente para el analista.

Roles	Escritura	Lectura	Administración
Analista		x	
Administrador ETL	x	x	x

2.2.4. Reglas del negocio.

En el proceso de almacenamiento de los indicadores del comercio exterior se guarda la información de las exportaciones y las importaciones en forma de clasificadores para un mejor control de estos parámetros. Varios de estos clasificadores tienen significados propios que se especifican a continuación.

Subpartida: Este clasificador tiene un código de ocho dígitos que describe a los productos que están siendo comercializados ya sea en una exportación o en una importación. De estos ocho dígitos los dos primeros pertenecen al capítulo, el tercer y cuarto número pertenecen a la partida y los últimos cuatro a la subpartida.

Grandes Categorías Económicas (GCE): Clasificador que divide a los productos en grandes categorías económicas, este posee un código que puede ser desde uno hasta tres dígitos. EL primer número de este código representa a las secciones de dichas categorías.

Clasificador Uniforme para el Comercio Internacional (CUCI): Este es el clasificador por el cual se rigen todas las entidades a nivel mundial que llevan algún tipo de control sobre el comercio internacional. Posee un código de cuatro ó cinco cifras del cual el primero significa la sección.

2.2.5. Necesidades de los usuarios.

En el análisis del comercio exterior, los especialistas se enfocan en las exportaciones e importaciones. De estas dos operaciones se almacenan varios parámetros en los cuales ellos enfocan sus estudios. Todo este análisis está destinado a satisfacer necesidades gubernamentales, como último y no menos importante para su publicación. Para más información referirse al anexo número 4.

2.2.6. Requisitos de información.

Los requisitos de información son las principales informaciones que deben estar disponibles al realizar los análisis sobre los datos. Constituyen una entrada fundamental para el proceso de inteligencia del negocio y para futuros reportes. Los siguientes son los requisitos que se han identificado hasta el momento:

- Obtener las exportaciones por tiempo.
- Obtener las importaciones por tiempo.
- Obtener el intercambio comercial con países, áreas geográficas por tiempo.
- Obtener las exportaciones según país del último destino de las mercancías por países seleccionados y áreas geográficas en el tiempo.
- Obtener las importaciones según país de origen de las mercancías por países seleccionados y áreas geográficas en el tiempo.
- Obtener las importaciones de mercancías agrupadas en grandes categorías en el tiempo.
- Obtener las exportaciones de mercancías según secciones de la Clasificación Uniforme para el Comercio Internacional (CUCI) por tiempo.
- Obtener las importaciones de mercancías según secciones de la CUCI por tiempo.

- Obtener las exportaciones de productos seleccionados según secciones y capítulos de la CUCI por tiempo.
- Obtener las importaciones de productos seleccionados según secciones y capítulos de la CUCI por tiempo.

2.2.7. Requisitos Multidimensionales (entradas y salidas-dirigidos al diseño del Almacén).

Los requisitos multidimensionales son las variables de entrada (VE) y las de salida (VS). Constituyen el acceso fundamental para el diseño de las estructuras del Mercado de Datos. Se derivan principalmente de los requisitos de información. A continuación se representan en forma de VE y VS.

VE	VS
temporal	peso_bru
países	peso_neto
moneda	can_com
subpartidas	val_flet
empresas	val_seg
con_pag	val_otrgas
regimen_exp	base_imp
regimen_imp	aran_min
	aran_usd
	serv_min
	preciounit
	val_fact

2.2.8. Requisitos funcionales.

Las funcionalidades del producto deben estar orientadas a las necesidades de información de los usuarios finales. Para darle cumplimiento a la afirmación anterior se han identificado los siguientes requisitos funcionales:

- Obtener el comportamiento de las importaciones y las exportaciones.

- Elaborar una copia de seguridad mensual en el repositorio.
- Permitir cargar, extraer y transformar los datos de cada fuente.

2.2.9. Requisitos no funcionales.

A continuación se presentan los requisitos no funcionales mediante los cuales se definen las propiedades o cualidades que el software debe tener.

Fiabilidad:

1. Disponibilidad:
    - El sistema será accedido para su mantenimiento una vez al mes.
    - Debe estar disponible el 100% del tiempo mientras no se le estén aplicando cambios y mientras esté fuera de mantenimiento.
2. Máximo de errores:
    - La cantidad de errores en el proceso de integración define la calidad de los datos que se están almacenando, es por ello que es crítico, definiéndose así 0 errores/puntos de función.
3. Tiempo medio entre fallos:
    - El tiempo medio entre fallos es de aproximadamente 6 días.
4. Tiempo medio de reparación:
    - El tiempo medio de reparación depende fundamentalmente de la magnitud del fallo pero se estima que como promedio sea de 24 horas.

Eficiencia:

1. Utilización de recursos:
    - El tiempo de respuesta deberá estar comprendido entre 1 y 10 segundos.
    - El sistema deberá permitir al menos 50 usuarios conectados sincrónicamente sin que se afecte el tiempo de respuesta.

2. Capacidad:

- En el proceso de integración solo tendrá conectado un usuario que tendrá la tarea de vigilar el proceso de integración de datos.

Restricciones del diseño:

El lenguaje de programación del proceso de integración de la base de datos será SQL, desarrollado en PostgreSQL 8.4.

2.2.10. Casos de uso del sistema

Con el fin de poder definir los requisitos de entrada y salida de datos al almacén, se han identificado los siguientes casos de uso del sistema.

Casos de uso informativos:

- Solicitar información de importaciones.
- Solicitar información de exportaciones.

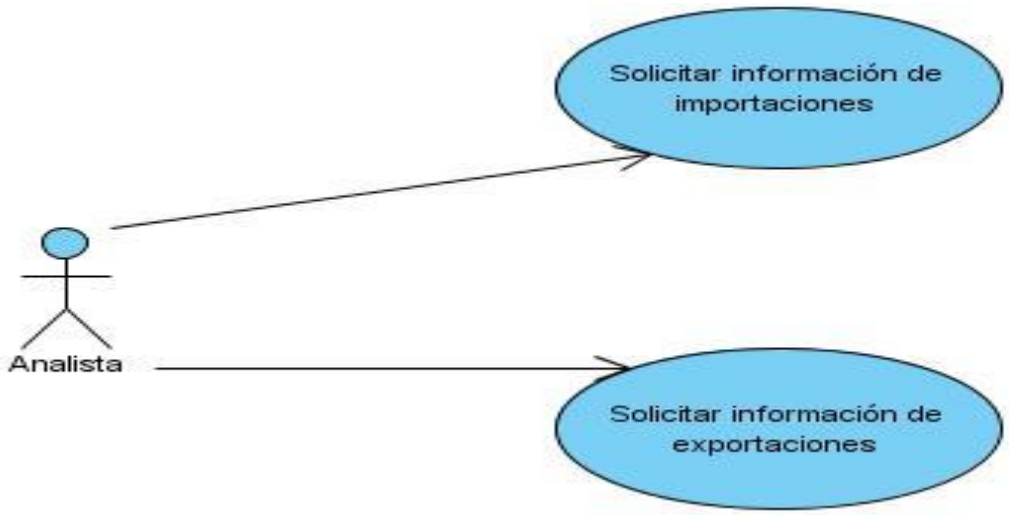

Figura 2 Diagrama de Casos de uso informativos

Casos de uso funcionales:

- CUS-1 Extraer CLA_AGEO.
- CUS-2 Transformar y cargar CLA_AGEO.
- CUS-3 Extraer CLA_GCE.

- CUS-4 Transformar y cargar CLA_GCE.
- CUS-5 Extraer CLA_GCEC.
- CUS-6 Transformar y cargar CLA_GCEC.
- CUS-7 Extraer CLA_SCUC.
- CUS-8 Transformar y cargar CLA_SCUC.
- CUS-9 Extraer CLA_SELE.
- CUS-10 Transformar y cargar CLA_SELE.
- CUS-11 Extraer CLA_SELI.
- CUS-12 Transformar y cargar CLA_SELI.
- CUS-13 Extraer CLA_SGCE.
- CUS-14 Transformar y cargar CLA_SGCE.
- CUS-15 Extraer CLA_CLASIF.
- CUS-16 Transformar y cargar CLA_CLASIF.
- CUS-17 Extraer CLA_CPAISES.
- CUS-18 Transformar y cargar CLA_CPAISES.
- CUS-19 Extraer CLA_EMPRESAS.
- CUS-20 Transformar y cargar CLA_EMPRESAS.
- CUS-21 Extraer CLA_EXP.
- CUS-22 Transformar, y cargar CLA_EXP.
- CUS-23 Extraer CLA_IMP.
- CUS-24 Transformar y cargar CLA_IMP.
- CUS-25 Extraer CLA_REGEXP.
- CUS-26 Transformar y cargar CLA_REGEXP.
- CUS-27 Extraer CLA_REGIMP.

- CUS-28 Transformar, y cargar CLA_REGIMP.
- CUS-29 Extraer CLA_UNIDADES.
- CUS-30 Transformar y cargar CLA_UNIDADES.
- CUS-31 Extraer CLA_RGS.
- CUS-32 Transformar y cargar CLA_RGS.

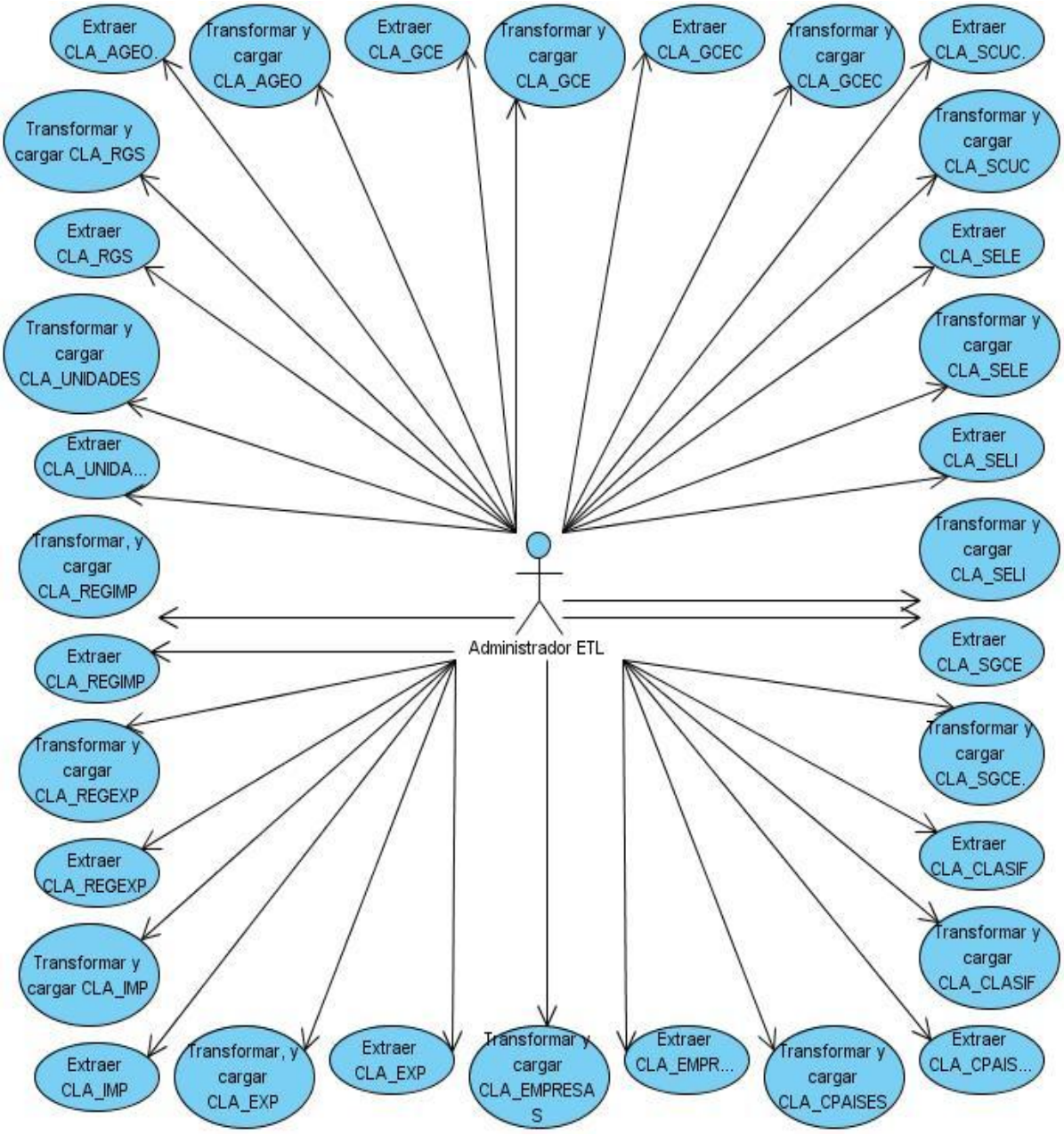

Figura 3 Diagrama de Casos de Uso del Sistema

2.3. Diseño.

2.3.1. Matriz BUS.

El propósito de la matriz bus es obtener el modelo lógico inicial, donde queda identificado el mercado de datos y sus dimensiones identificadas.

Dimensiones
1. Dimensión dim_regimen_exp
2. Dimensión dim_regimen_imp
3. Dimensión con_pag.
4. Dimensión subpartida
5. Dimensión empresa
6. Dimensión países
7. Dimensión temporal
8. Dimensión moneda.

Tabla 2 Dimensiones

Hechos	
AA1. Hech_Exp.	**AA2. Hech_Imp.**

Tabla 3 Hechos

AA/dim	1	2	3	4	5	6	7	8
AA1	x		x	x	x	x	x	x
AA2		x	x	x	x	x	x	x

Tabla 4 Matriz Bus

2.3.2. Modelo de datos.

A continuación se muestra el modelo de datos realizado con la herramienta de modelado Visual Paradigm.

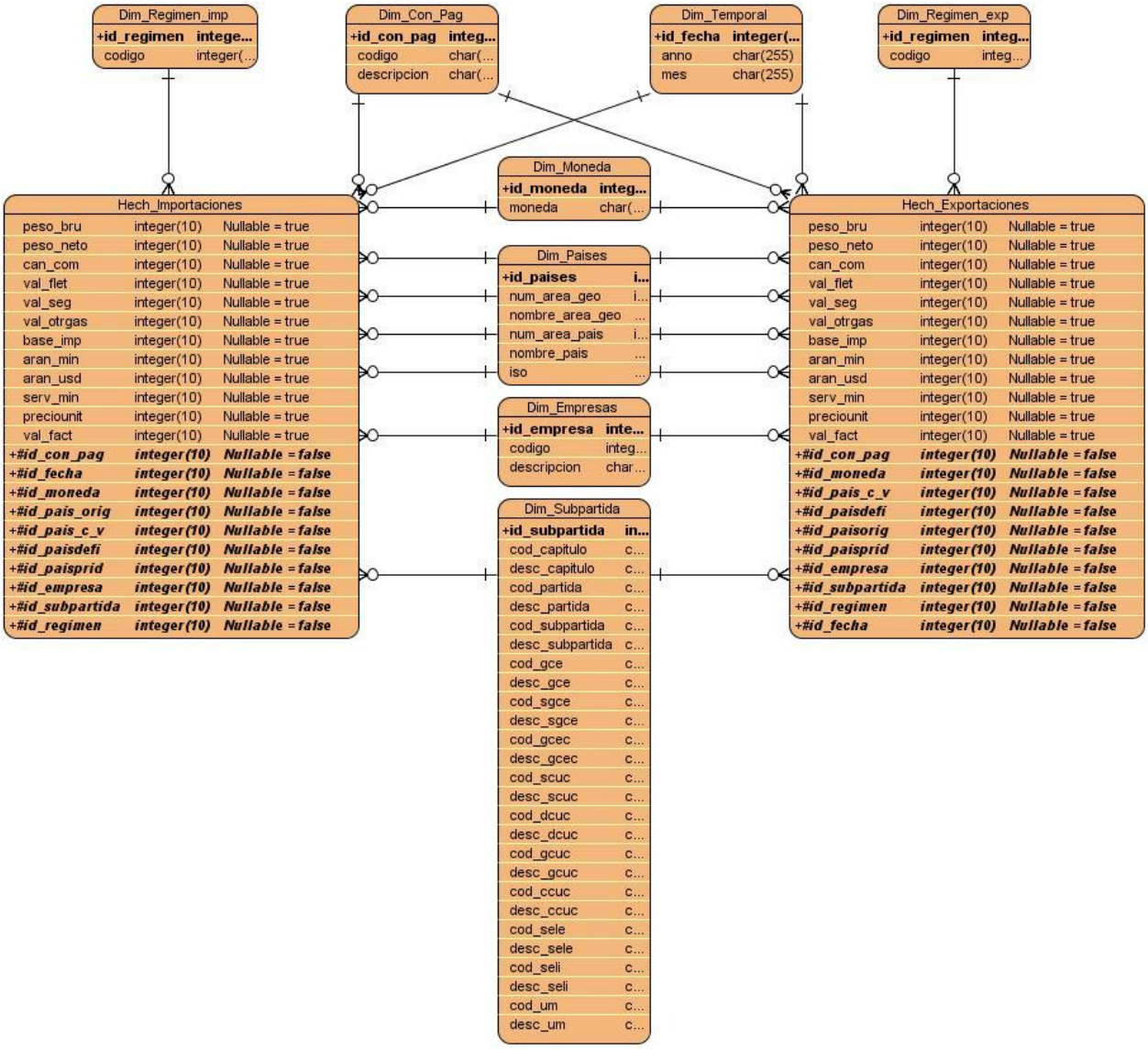

Figura 4 Propuesta de solución

2.3.2.1. Dimensiones y Jerarquías.

Dimensión Temporal (dim_temporal). Jerarquía anno → mes
Dimensión Países (dim_paises). Jerarquía nombre_area_geo → nombre_pais
Dimensión Moneda (dim_moneda). Jerarquía moneda
Dimensión Empresas (dim_empresa). Jerarquía codigo
Dimensión subpartidas (Dim_Subpartidas). Jerarquía capitulo → partida → subpartida gce → sgce gcec cuci → scuc → dcuc → gcuc → ccuc
Dimensión Condición de Pago (dim_con_pag).

Jerarquía codigo
Dimensión Régimen de Exportación (dim_regimen_exp). Jerarquía codigo
Dimensión Régimen de Importación (dim_regimen_imp). Jerarquía codigo

Tabla 5 Tabla de Dimensiones y Jerarquías

2.3.2.2. Tablas de hechos.

Las dos tablas de hechos identificadas contienen los mismos atributos por lo que se muestra una sola tabla representando los dos hechos.

Exportaciones/Importaciones:

Atributos	Tipo de dato	Exportaciones	Importaciones
id_empresa	intiger	x	x
id_pais_orig	Intiger	x	x
Id_pais_c_v	Intiger	x	x
Id_paisdefi	Intiger	x	x
Id_pais_prid	Intiger	x	x
Id_subpartidas	intiger	x	x
Id_fecha	Intiger	x	x

Id_moneda	Intiger	x	x
Id_con_pag	intiger	x	x
Id_reg_exp	intiger	x	-
Id_reg_imp	intiger	-	x

Tabla 6 Tabla de Hecho Exportaciones/Importaciones

Medidas:

Atributos	Tipo de dato	Exportaciones	Importaciones
Peso_bru	float	x	x
Peso_neto	float	x	x
Can_com	float	x	x
Val_fet	float	x	x
Val_seg	float	x	x
Val_otrogas	float	x	x
Base_imp	float	x	x
Aran_min	float	x	x
Aran_usd	float	x	x
Serv_min	float	x	x
Preciounit	float	x	x
Val_fact	float	x	x

Tabla 7 Tabla de Hecho de medidas

2.3.3. Esquema de Seguridad.

El esquema de seguridad estará apoyado en gran medida por los niveles de acceso al sistema y fundamentalmente por los roles definidos con anterioridad.

Se ha definido una arquitectura en 3 capas para lograr una alta disponibilidad y rendimiento del sistema de seguridad, ellas son:

- Capa de Funcionamiento
- Capa de Servidores
- Capa de Presentación

En la consola de administración radicará una Interfaz de Administración Gráfica como parte de la solución y con la utilización de un editor de políticas de seguridad se logrará hacer una mejor visualización de la topología que tiene la red implicada. Para lograr una correcta representación de la topología se representará de manera que se pueda observar la relación entre los distintos objetos definidos con el editor de políticas y la red.

Para el control del tráfico existente en toda la red se utilizará un registro Log File con el fin de monitorear los posibles ataques a la aplicación. Este registro tiene un formato, que se define en las entradas de información detectadas por la red con una regla hecha para ello.

Con vistas a lograr un mejor rendimiento del esquema previsto las pruebas deberán realizarse en las instalaciones de la Oficina Nacional de Estadísticas y en conjunto con el personal calificado para la situación. En aras de prevenir posibles ataques a la aplicación se realizarán visitas semestrales mientras dure el período de garantía técnica; y el mantenimiento correctivo se realizará anualmente. La actualización de todo el software se hará mediante parches que se le hagan al mismo y la utilización si es debida de sistemas operativos, más compatibles con él.

2.3.4. Política de respaldo y recuperación.

La política de respaldo y recuperación que utiliza el software es robusta y está medida por 3 puntos fundamentales.

- Constante realización de salvas al sistema: estas se realizarán en un período aproximado de 30 días y se le harán a toda la información que contenga la base de datos, revisando siempre que exista una copia de toda la información almacenada.
- Tablas de hechos involucradas para el registro de toda la información entrante y saliente.
- Salvas existentes: aunque actualmente no existan backups en esta área, se prevé que una vez estén se le realizarán reemplazos por períodos anuales y los mismos se chequearán mensualmente para monitorear su comportamiento.

Conclusiones del Capítulo

Al finalizar el anterior capítulo, se ha fundamentado con claridad el ciclo del proceso del negocio y los métodos de solución que se seguirán para el desarrollo del mercado de datos concerniente al comercio exterior para la Oficina Nacional de Estadísticas. Los puntos que a continuación se muestran evidencian de manera sintetizada y clara el trabajo realizado.

- Se definió el tema de análisis basado en un criterio fundamental:

  Indicadores para el Comercio Exterior.

- Se establecieron las principales reglas del negocio, roles y permisos.

- Se realizó el modelado del diseño lógico definiendo las dimensiones y tablas de hechos involucradas en la solución.

CAPÍTULO 3. IMPLEMENTACIÓN Y PRUEBAS

3.1. Introducción.

Al finalizar el proceso de análisis y diseño del Mercado de Datos, resulta importante adentrarse en los temas de implementación y validación. Después de identificadas con anterioridad las medidas, hechos y dimensiones correctamente, se pasará a realizar el proceso de ETL, teniendo en cuenta que para la solución en cuestión se realizará solo una parte del mismo.

Con la interrelación entre los usuarios finales y el Mercado de Datos, el sistema pasará por una prueba de rendimiento con el comienzo del almacenamiento de los datos a través de los años. Resulta un hecho irrefutable que al transcurrir el tiempo, el MD se hará cada vez más grande y complejo de manejar. Es por ello que realizar las correspondientes pruebas de calidad al producto resulta un paso importante a realizar.

El Mercado de Datos debe ser capaz de brindarle dinamismo a los usuarios a la hora de realizar los principales reportes y cruces de variables y de esta manera ayudar al proceso de toma de decisiones de la Oficina Nacional de Estadísticas. Para un uso más adecuado del MD y en aras de facilitar el proceso de extracción de la información almacenada, la preparación y capacitación de los usuarios finales resulta vital para garantizar el éxito en la utilización del sistema.

3.2. Implementación

3.2.1. Modelos de Datos Físico.

Para el desarrollo del Modelo de Datos Físico es necesario tener en cuenta al Modelo Lógico lo más integralmente posible. No obstante, si bien uno depende en gran medida del otro es realmente un hecho que esta relación no puede ser 100% íntegra. Por tanto, para la elaboración de un Modelo de Datos Físico partiendo de su Modelo Lógico es necesario realizar ciertos cambios dentro de la estructura de sus tablas y columnas en miras de garantizar un correcto trabajo con el Sistema Gestor de Base de Datos Relacional. Por otra parte, el Modelo Físico presenta tablas de mantenimiento que no pueden ser incluidas dentro del Modelo Lógico.

El tamaño estimado de la base de datos inicial es algo complicado, pero constituye el paso final del proceso de construcción del modelo por su importancia a la hora de evaluar el rendimiento del sistema. Las longitudes de los valores almacenados en las columnas pueden llevar a un

crecimiento desmedido de la base de datos. Tal es el caso de las cadenas VARCHAR, que suelen no ser explotadas del todo lo que trae como consecuencia un desmedido almacenamiento de espacio en el Sistema Gestor de Base de Datos cuando puede que realmente no estén llenos los campos.

3.2.2. Estructuras de Datos.

Representan la colección de datos organizados por funciones de acceso utilizadas para almacenar y acceder a elementos individuales de la base de datos. Es una forma de organización para los conjuntos de datos fundamentales, y de esta forma facilitar su uso. En un nivel inferior a estos elementos se encuentran los archivos, discos, particiones y espacios de tabla; el correcto uso de dichos elementos y su dominio facilita en gran medida el trabajo, por lo que constituyen una práctica a realizar para garantizar el éxito de la aplicación.

3.2.3. Esquemas y Tablas

El elemento que realiza una descripción de la estructuración de la base de datos mediante un lenguaje formal y que es soportado por el Sistema Gestor de Base de Datos es denominado esquema de la base de datos. En una base de datos relacional, el esquema es el encargado de definir sus tablas, los campos en cada tabla y las relaciones entre cada campo y tabla. Este elemento es almacenado en un diccionario de datos y su definición de manera general se hace en un lenguaje de base de datos. Es utilizado para hacer referencia a una representación gráfica de la estructura de la base de datos. A continuación se muestran los esquemas definidos para la solución.

- Esquema Dimensiones: contiene todas las tablas dimensiones que manejan información relacionada con los indicadores del comercio exterior.
- Esquema Hecho: contiene todas las tablas hechos que manejan información relacionada con los indicadores del comercio exterior.

Por otra parte, las tablas manejan el tipo de modelado de datos donde se guardan los datos recogidos por un programa. Tiene dos estructuras fundamentales, ellas son:

- Campo: corresponde al nombre de la columna, debe ser único y tener un tipo de dato asociado.

- Registro: corresponde a la fila que compone la tabla y en ellos se guardan los datos y registros, los cuales pueden tomar valores nulos.

Ejemplos de esquemas y tablas utilizados en la solución

Tablas	Esquemas	Usuarios
dim_con_pag	Dimensiones	Analista
dim_empresas	Dimensiones	Analista
dim_moneda	Dimensiones	Analista
dim_paises	Dimensiones	Analista
dim_regimen_exp	Dimensiones	Analista
dim_regimen_imp	Dimensiones	Analista
dim_subpartidas	Dimensiones	Analista
dim_temporal	Dimensiones	Analista
hech_exp	Hechos	Analista
hech_imp	Hechos	Analista

Ejemplo 1. Tabla dim_empresas

Tabla	Esquema	Campo	PK	Tipo de Datos	No nulo	Único
dim_empresas	Dimensiones	id_empresa	✓	serial	✓	✓
		codigo		varchar		
		descripcion		varchar		

Ejemplo 2. Tabla dim_regimen_imp

Tabla	Esquema	Campo	PK	Tipo de Datos	No nulo	Único
dim_regimen_imp	Dimensiones	id_regimen	✓	serial	✓	✓
		codigo		varchar		✓

Ejemplo 3. Tabla hech_exp

Tabla	Esquema	Campo	FK	Tipo de Datos	No nulo	Único
Hech_Exp	Hechos	id_fecha	✓	integer	✓	
		id_empresa	✓	integer	✓	
		id_PAIS_ORIG	✓	integer	✓	
		id_PAISDEFI	✓	integer	✓	
		peso_bru		real		

Ejemplo 4. Tabla Hech_Imp

Tabla	Esquema	Campo	FK	Tipo de Datos	No nulo	Único
Hech_Imp	Hechos	id_empresa	✓	integer	✓	
		id_PAIS_ORIG	✓	integer	✓	
		id_PAIS_C_V	✓	integer	✓	
		id_PAISDEFI	✓	integer	✓	
		peso_bru	✓	real	✓	

3.2.4. Restricciones y Secuencias

El uso de las restricciones facilita un correcto control de determinadas reglas establecidas para el uso de la base de datos. La mayoría de estas son definidas por el Gestor de Base de Datos, en este caso por PostgreSQL, pero existen otras que son creadas por los usuarios. Entre las restricciones que pueden definir los usuarios se encuentran la del uso de rangos para registrar valores en determinados campos, en dependencia de lo que se desee guardar. La correcta utilización de las restricciones permite definir distintos métodos de implementación de reglas dentro de la base de datos, y a su vez controlan los datos que puedan ser almacenados restringiendo el acceso a ellos.

Por otra parte, se hace imprescindible la utilización de las secuencias, que no son más que los atributos que se van a ir incrementando de una manera secuencial durante todo el proceso de almacenamiento de los datos en la base de datos. El estudio y definición de las llaves, tanto primarias como foráneas también será objeto de análisis en el presente epígrafe. A continuación se muestran algunos ejemplos de secuencias y llaves foráneas identificadas en la solución:

Ejemplo 1 Llaves Foráneas

Foreign Key	Table
RefDim_Subpartidas50	Hech_Exp
RefDim_Temporal43	Hech_Exp
RefDim_Con_Pag74	Hech_Imp
RefDim_Empresas60	Hech_Imp

Secuencias: Son atributos que se van a ir incrementando secuencialmente durante la entrada de datos a la BD, un ejemplo de estos tipos de atributos son las llaves primarias.

Ejemplo 2 Secuencias

Secuencia	Esquema	Prox. Valor	Incremental	Min. Valor	Máx. Valor
dim_con_pag_id_con_pag_seq	Dimensiones	1	1	1	2147483647
dim_empresas_id_empresa_seq	Dimensiones	17697	1	1	2147483647
dim_moneda_id_moneda_seq	Dimensiones	2	1	1	2147483647
dim_paises_id_paises_seq	Dimensiones	292	1	1	2147483647
dim_regimen_exp_id_regimen_seq	Dimensiones	85	1	1	2147483647
dim_regimen_imp_id_regimen_seq	Dimensiones	85	1	1	2147483647
dim_subparti	Dimensiones	11797	1	1	2147483647

das_id_subpartidas_seq					
dim_temporal_id_fecha_seq	Dimensiones	1	1	1	2147483647

3.2.5. Índices

Las políticas de indexado son restricciones que el programador de BD define a la hora de indexar o crear un índice. Un índice es una estructura de disco asociada con una tabla o una vista que acelera la recuperación de filas de la tabla o de la vista. Un índice contiene claves generadas a partir de una o varias columnas de la tabla o la vista. Dichas claves están almacenadas en una estructura (árbol b) que permite que SQL Server busque de forma rápida y eficiente la fila o filas asociadas a los valores de cada clave.

A continuación se mostrarán aquellos índices que genera automáticamente el gestor, por ejemplo a la hora de crear una llave primaria, cuando se crea una tabla o cuando se identifica una columna según la llave primaria coincidiendo con las propiedades o características que poseen los índices agrupados. Es significativo resaltar que los usuarios también pueden definir índices según las necesidades específicas de cada uno de estos. También se mostrarán los atributos UNIQUE que enriquecen el indexado de las tablas.

Tabla de Índices

Índice	Tabla	Esquema	Tipo	Campo	PK	Único
PKcon_pag	dim_con_pag	dimensiones	btree	id_con_pag	x	x
dim_con_pag_codigo_key	dim_con_pag	dimensiones	btree	codigo		x
PKempresas	dim_empresas	dimensiones	btree	id_empresa	x	x

PKmoneda	dim_moneda	dimensiones	btree	id_moneda	x	x
PKpaises	dim_paises	dimensiones	btree	id_paises	x	x
PKregimen_exp	dim_regimen_exp	dimensiones	btree	id_regimen	x	x
dim_regimen_exp_codigo_key	dim_regimen_exp	dimensiones	btree	codigo		x
PKregimen_imp	dim_regimen_imp	dimensiones	btree	id_regimen	x	x
dim_regimen_imp_codigo_key	dim_regimen_imp	dimensiones	btree	codigo		x
PKsubpartidas	dim_subpartidas	dimensiones	btree	id_subpartidas	x	x
dim_subpartidas_cod_subpartida_key	dim_subpartidas	dimensiones	btree	cod_subpartida		x
PKtemporal	dim_temporal	dimensiones	btree	id_fecha	x	x

3.2.6. Usuarios y Privilegios

La definición de los usuarios y sus privilegios resulta una tarea imposible de pasar por alto a la hora de construir un Mercado de Datos. Es importante que las personas involucradas con el sistema tengan bien definido los permisos para realizar solo las operaciones que se le estén permitidas. Asignar las acciones más complejas al personal debidamente capacitado constituye un método a poner en práctica. A continuación se representará de una manera más detallada los roles identificados y sus respectivos privilegios.

3.2.6.1 Usuarios y Roles

- Consultor: consulta la información contenida en la base de datos.
- Administrador: derechos de lectura y escritura de los datos almacenados en el Mercado de Datos

3.2.6.2. Privilegios

Para asignar los privilegios a los usuarios del sistema se ha tomado en cuenta el rol que desempeñan y el uso que hagan de la base de datos, quedando definido de la forma que se muestra a continuación en la tabla:

Usuario	Permisos	Descripción
Consultor	Derechos de lectura de los datos almacenados en el Mercado de Datos	
Administrador	Derechos de lectura y escritura de los datos almacenados en el Mercado de Datos	
Postgres	Como un súper usuario hace uso completo de la base de datos	

3.2.7. Carga de nomencladores

El proceso de verificación para la inclusión al sistema de los datos provenientes de una o varias fuentes recibe el nombre de fase de carga. Con la creación, distribución y carga de una manera homogénea de los nomencladores se hace factible la estandarización de todos los criterios de datos para la construcción de un Mercado de Datos. Realizar el proceso de carga de nomencladores recogidos en los archivos dbf, constituye una operación de vital importancia a la hora de completar la solución. Dicha carga se realiza desde el propio gestor de bases de datos, el cual brinda esta posibilidad de una manera fácil y efectiva. A continuación se verá la

correspondencia de la tabla a cargar con respecto al fichero que contiene los datos. Para la efectiva carga de las dimensiones países y sus subpartidas se hizo necesario la agrupación de sus respectivos clasificadores (dbf), en un Excel logrando una importación de los datos de forma más sencilla.

Tabla de Ficheros

Dimensiones	Ficheros
Dim_regimen_imp	REGIMP.DBF
Dim_regimen_exp	REGEXP.DBF
Dim_empresas	EMPRESAS.DBF
Dim_con_pag	-
Dim_paises	CLA_AGEO.DBF, PAISES.DBF
Dim_moneda	-
Dim_temporal	-
Dim_subpartidas	CLA_GCE.DBF, CLA_GCEC.DBF, CLA_SCUC.DBF, CLA_SELE.DBF, CLA_SELI.DBF, CLA_SGCE.DBF, CLASIF.DBF, RGS.DBF, UNIDADES.DBF, CLA_DCUC.DBF

3.2.8. Guía de Implantación

3.2.8.1. Requerimientos

Los requerimientos son un aspecto muy importante a tener en cuenta para la implantación de un Mercado de Datos. Para el funcionamiento del producto es vital analizar ciertas condiciones que garanticen el uso óptimo del mismo. Estos son los principales requerimientos que se deben tener en cuenta agrupados por dos criterios fundamentales:

Software:

- Se debe utilizar un navegador común, asociado principalmente a un sistema operativo, garantizando de esta forma una única visualización de la interfaz web para los reportes y consultas que se hagan.
- El montaje de la base de datos se hará con el Gestor de Base de Datos PostgreSQL en su versión 8.4.
- El lenguaje a utilizar para realizar las consultas dentro de la base de datos será SQL.

Hardware:

- Para garantizar un rápido funcionamiento del sistema a las peticiones hechas cuando sea accedido por varios usuarios a la vez, se requiere un mínimo de 1 GB de memoria RAM.
- La capacidad mínima requerida para el almacenamiento de los datos debe ser de al menos 60 GB de disco duro.
- Memoria mínima de 512 Mb para realizar el proceso de transformación.
- El sistema debe de estar correctamente conectado a la red en aras de garantizar el acceso a los usuarios.
- Para la impresión de los reportes es necesario el uso de una impresora.

3.2.8.2. Pasos para la instalación de la base de datos.

Para realizar la instalación de la base de datos se deben tener en cuenta los siguientes pasos:

1. Una vez instalado el Gestor de Base de Datos PostgreSQL, se procede a la creación de una nueva base de datos de la siguiente forma:
    a. Seleccionar Bases de Datos/Editar/Nuevo objeto/Nueva Base de Datos.
    b. Introducir los datos que demande el gestor y crear la base de datos.
2. Seleccionar la base de datos creada y hacer uso de la herramienta de consulta del gestor PostgreSQL.
3. Cargar el script denominado *DDL Dimensiones PT.sql* y ejecutarlo. Con este paso se creará el esquema “dimensión” con sus tablas correspondientes.

4. Volver al paso 2 y cargar el script denominado *DDL Hechos PT.sql* y ejecutarlo. Con este paso se creará el esquema “hechos” con sus tablas correspondientes.
5. Volver al paso 2 nuevamente y cargar el script denominado *DML Dimensiones PT.sql* y ejecutarlo. Con este paso se cargará los nomencladores de las tablas “dimensión”.
6. Volver al paso 2 una vez más y cargar el script denominado *DSL PT.sql* y ejecutarlo. Con este paso se cargarán los usuarios creados y se establecerán los privilegios de ellos sobre los objetos de la base de datos.

Después de realizado los pasos descritos anteriormente se puede afirmar que la base de datos se encuentra lista para su explotación.

3.3. Validación y pruebas

Las pruebas son los procesos que se le realizan al producto para verificar su correcto funcionamiento y calidad. Estas se hacen con el objetivo principal de identificar dificultades en temas de implementación, o de manera general en la usabilidad del sistema. La definición de cuándo y cómo hacerlas varía en dependencia de la metodología de desarrollo utilizada. En el presente trabajo se ha definido una estrategia para garantizar el mejor uso posible del sistema apoyado en la “Metodología para el desarrollo de Soluciones de Almacenes de Datos e Inteligencia de Negocio (DW&BI) de DATEC”. Los epígrafes siguientes tratarán sobre aspectos fundamentales dentro de este vital flujo de la metodología; las Listas de Chequeo Análisis, la Validación de requisitos por el cliente, la Lista de Chequeo de Diseño y por último las Pruebas de Implantación al sistema serán los temas a abordar en ellos.

3.3.1. Listas de Chequeo de Análisis

Como parte de la fase de análisis de la solución, se encuentran los pasos a seguir para el desarrollo de la solución en los documentos encargados de detallar las actividades a realizar para la implementación del sistema. De esta forma, son generadas una serie de listas de chequeo que cubren los aspectos más importantes en los temas de análisis, en particular de los indicadores del Comercio Exterior. Seguidamente se muestran las listas generadas en el proceso de análisis realizado:

- Lista de Chequeo de Especificación de Requisitos
- Lista de Chequeo de la Herramienta para la colección y análisis de los datos.

- Lista de Chequeo de las Áreas de la organización.

3.3.2. Validación de requisitos por el cliente

En el ámbito de aprobación de un producto el cliente es el aspecto más importante a tener en cuenta. Es por ello que la realización de entrevistas con él se torna una necesidad de primer orden en vista a satisfacer sus exigencias. Para la confección del sistema se realizó un encuentro con el cliente donde se abordaron los requisitos definidos y se mostró el diseño del Modelo de Datos Lógico, manifestando su acuerdo con los aspectos tratados.

3.3.3. Lista de Chequeo de Diseño

Con la finalidad de lograr un buen diseño de la solución se hace imprescindible el uso de la Lista de Chequeo de Diseño. Estas surgen en dependencia del sistema a desarrollar y pretende abarcar todos los aspectos de interés en los temas de diseño e implementación, en el presente caso para los indicadores del comercio exterior. La lista es generada y es aplicable en la fase de diseño. A continuación se muestra la lista generada en dicha fase:

- Lista de Chequeo del Modelo de Datos.

3.3.4. Pruebas de Implantación

En aras de garantizar una adecuada calidad en el proceso de implementación se ha procedido a definir un modelo de casos de prueba para apoyar la identificación de los resultados y de esta manera hacerlos más verídicos y eficientes. La tabla que se muestra a continuación representa los pasos seguidos para dicho proceso.

Casos de Prueba	Pre-Condiciones	Resultados esperados	Pos-Condiciones
Caso de Prueba 1	Antes de implantar el sistema, es importante tener instalado el gestor PostgreSQL y para la administración del mismo el PgAdminlll. Para lograr	Al contar con las pre-condiciones requeridas se puede obtener las condiciones necesarias para	Crear la base de datos donde se montará el Mercado de Datos y poder copiar el sitio del servidor Apache que visualizará los datos

	una mayor eficiencia en la aplicación web se debe tener instalado el servidor Apache Tomcat.	proceder a la implantación del sistema.	gráficamente.
Caso de Prueba 2	Tener creada correctamente la base de datos para el proceso de carga de los DCL.	Contar con los usuarios y permisos definidos cargados en el Mercado de Datos.	Contar con las condiciones requeridas para una adecuada estructura de la base de datos.
Caso de Prueba 3	Tener cargado el DCL, para de esta forma poder correr el script DDL.	Contar con una correcta estructura de la base de datos con sus usuarios y permisos bien definidos.	Contar con las condiciones mínimas necesarias para poder cargar los nomencladores.

Conclusiones del Capítulo

En este capítulo se ha procedido a la implementación del sistema. Se detalló el Modelo de Datos Físico, la extracción y carga de los datos y por último se le han realizado las pruebas definidas para validar un satisfactorio uso del sistema. Algunos aspectos importantes a destacar dentro del mismo son:

- La correcta elaboración del Modelo de Datos Físicos utilizando las herramientas adecuadas permite una óptima implementación del sistema.
- El almacenamiento organizado y eficiente de los datos, así como la utilización de técnicas de indexado asegura un mejor rendimiento del sistema en las respuestas a los pedidos de información.
- La realización de pruebas de validación al sistema garantiza su uso de una manera adecuada y permite su implantación con un alto nivel de éxito.

CONCLUSIONES GENERALES

Al finalizar el desarrollo del presente trabajo se puede arribar a las siguientes conclusiones generales:

- Los Mercados de Datos constituyen una robusta solución para el almacenamiento de los datos relacionados con los indicadores del comercio exterior.
- Con la utilización del Gestor de Base de Datos PostgreSQL es posible manejar el volumen de información que presenta la solución en cuestión.
- Se logró implementar las estructuras dimensionales utilizadas para la solución, y con el uso de las mismas se agilizará el proceso de toma de decisiones para la ONE.
- Las pruebas efectuadas al sistema permitieron la validación del mismo, obteniendo resultados positivos y la aceptación del cliente.

RECOMENDACIONES

- La continuación y terminación de la carga de datos que no han estado disponibles hasta el momento, para alguna de sus dimensiones.

- Desarrollar la capa de visualización de inteligencia de negocio, para el mercado de datos “Indicadores relacionados con el Comercio Exterior para la Oficina Nacional de Estadísticas”.

BIBLIOGRAFÍA

1. MENÉNDEZ, D. E. S. *Diseño y Optimización de Bases de Datos* [Consultado el: 24 de febrero de 2010]. Disponible en: http://www.oei.eui.upm.es/Asignaturas/BD/DYOBD/Ejemplo_DW.pdf.

2. VELASCO, R. H. *Almacenes de datos (Datawarehouse)* [Consultado el: 25 de febrero de 2010]. Disponible en: http://www.rhernando.net/modules/tutorials/doc/bd/dw.pdf.

3. CASTILLO, D. O. A. F. Y. J. N. P. *Estado actual de las tecnologías data warehousing y OLAP aplicadas a bases de datos espaciales* Luís Joyanes Aguilar, [Consultado el: 25 de febrero de 2010]. Disponible en: http://novella.mhhe.com/sites/dl/free/8448118952/540197/ActasVol2SISOFT2006.pdf#page=113.

4. ABBEY, M. J. C. Y. M. *Oracle Data Warehousing.* 2002.

5. CASARES, C. *Data Warehousing* [Consultado el: 26 de febrero de 2010]. Disponible en: http://www.programacion.com/bbdd/tutorial/warehouse/.

6. CURTO, J. *Data Warehousing, Data Warehouse y Datamart* [Consultado el: 26 de febrero de 2010]. Disponible en: http://informationmanagement.wordpress.com/tag/data-warehousing/.

7. IMHOFF, C. *Mastering Data Warehouse Design, Relational and Dimensional Techniques.* 2003.

8. SINNEXUS. *Datamart* [Consultado el: 22 de febrero de 2010]. Disponible en: http://www.sinnexus.com/business_intelligence/datamart.aspx.

9. VÁZQUEZ, F. P. Y. G. *Relevamiento: Diseño Físico de Sistemas OLAP* [Consultado el: 25 de febrero de 2010]. Disponible en: http://www.fing.edu.uy/~fpiedrab/downloads/Physical_OLAP_Design.pdf.

10. KONCILIA, R. W. Y. C. *DATA WAREHOUSES AND OLAP.* 2007.

11. V., P. F. P. *Sistemas de Soporte a la toma de Decisiones* [Consultado el: 27 de febrero de 2010]. Disponible en: http://palomo.usach.cl/bdnc/2005-02/Presentaciones/U3-1-OLAP.pdf.

12. CURTO, J. *¿Qué es una Staging Area?* [Consultado el: 28 de febrero de 2010].

Disponible en: http://informationmanagement.wordpress.com/2007/10/15/%C2%BFque-es-una-staging-area/.

13. LORENA ETCHEVERRY, P. G., SALVADOR TERCIA. *Análisis del proceso de carga del Sistema de Data Warehousing de Enseñanza de la Facultad de Ingeniería* [Consultado el: 27 de febrero de 2010]. Disponible en: http://www.info.univ-tours.fr/~veronika/publications/tr0606-le.pdf.

14. WOLFF, C. G. *Modelamiento Multidimensional* [Consultado el: 25 de febrero de 2010]. Disponible en: http://www.inf.udec.cl/~revista/ediciones/edicion4/modmulti.PDF.

15. VERÁSTEGUI, H. C. *Modelado Dimensional de Datos* [Consultado el: 24 de febrero de 2010]. Disponible en: http://www.db-system.com/pls/portal/docs/PAGE/SITIOWWWDB/ARTICULOS/MODELADO%20DIMENSIONAL%20DE%20DATOS_V2.PDF.

16. CURTO, J. *Diseño de un data warehouse: tabla de hecho* [Consultado el: 26 de febrero de 2010]. Disponible en: http://informationmanagement.wordpress.com/tag/data-warehousing/.

17. GONZÁLEZ, Y. I. *Diseño e Implementación de un Data Warehouse para el Sistema de Gestión Estadística en Cuba.* Universidad de las Ciencias Informáticas, 2008.

18. SIERRA, J. E. O. *Diseño e Implementación de un Mercado de Datos para la Oficina Nacional de Estadísticas.* Universidad de las Ciencias Informáticas, 2009.

19. ALVAREZ, S. *Sistemas gestores de bases de datos* [Consultado el: 20 de febrero de 2010]. Disponible en: http://www.desarrolloweb.com/articulos/sistemas-gestores-bases-datos.html.

20. GONZÁLEZ, M. H. Á. *DISEÑO DE LA BASE DE DATOS DEL SISTEMA DE INFORMACIÓN DE PERFORACIÓN DE POZOS.* Universidad de las Ciencias Informáticas, 2009.

21. SANTOS, D. F. A. Y. Y. F. *Administración, configuración y optimización de un Sistema de Bases de Datos Descentralizado en Oracle Database 10g release 2* [Consultado el: 28 de febrero de 2010]. Disponible en: http://bibliodoc.uci.cu/TD/TD_0289_07.pdf.

22. MALDONADO, D. M. *SQLite, el motor de base de datos ágil y robusto* [Consultado el: 28 de febrero de 2010]. Disponible en: http://www.aplicacionesempresariales.com/sqlite-el-motor-de-base-de-datos-agil-y-robusto.html.

23. OCHOA, D. G. *Diseño e Implementación de un Almacén de Datos Operacionales para la Corporación CIMEX.* [Consultado el: 28 de febrero de 2010]. Disponible en: http://bibliodoc.uci.cu/TD/TD_2188_09.pdf.

24. ALFARO, F. M. *Herramientas Case* [Consultado el: 27 de febrero de 2010]. Disponible en: http://www.innovavirtual.org/campus/file.php/178/archivos_curso/CAP_12_2006_I_SI905/CAP_12_2006_I_SI905_VA8_M.pdf.

25. GONZÁLEZ, R. T. C. Y. L. G. *Análisis, diseño e implementación del Sistema de Gestión de Flotas por GSM/GPRS.* Universidad de las Ciencias Informáticas, 2009.

26. GSINNOVA, G. D. S. *Rational Rose Data Modeler* [Consultado el: 4 de marzo de 2010]. Disponible en: http://www.rational.com.ar/herramientas/rosedatamodeler.html.

27. COMPANY, E. *ER/Studio Enterprise* [Consultado el: 4 de marzo de 2010]. Disponible en: http://www.embarcadero.com/products/er-studio-enterprise.

28. CHAVEZ, M. P. Y. J. *Embarcadero Presenta a la Industria el Primer Kit de Herramientas para Desarrollo de Software Multiplataforma Bajo Demanda: Embarcadero® All-Access™* [Consultado el: 3 de marzo de 2010]. Disponible en: http://etnaweb04.embarcadero.com/news/press_releases/all-access-press-release-latam.pdf.

29. MARTÍNEZ, J. P. *DISEÑO DEL SUBSISTEMA PARA EL TRATAMIENTO Y MODELADO DE DIAGRAMAS DE FLUJO DE INFORMACIÓN* [Consultado el: 27 de febrero de 2010]. Disponible en: http://bibliodoc.uci.cu/TD/TD_2783_09.pdf.

30. CURTO, J. *CIF vs MD : Dos enfoques clásicos en el diseño de la arquitectura de un Data Warehouse* [Consultado el: 25 de febrero de 2010]. Disponible en: http://bi-businessintelligence.blogspot.com/2009/01/cif-vs-md-dos-enfoques-clsicos-en-el.html.

31. ZEPEDA SÁNCHEZ, L. Z. *Metodología para el Diseño Conceptual de Almacenes de Datos.* Universidad Politécnica de Valencia, 2008.

REFERENCIAS BIBLIOGRÁFICAS

1. MENÉNDEZ, D. E. S. *Diseño y Optimización de Bases de Datos* [Consultado el: 24 de febrero de 2010]. Disponible en: http://www.oei.eui.upm.es/Asignaturas/BD/DYOBD/Ejemplo_DW.pdf.
2. VELASCO, R. H. *Almacenes de datos (Datawarehouse)* [Consultado el: 25 de febrero de 2010]. Disponible en: http://www.rhernando.net/modules/tutorials/doc/bd/dw.pdf.
3. CASTILLO, D. O. A. F. Y. J. N. P. *Estado actual de las tecnologías data warehousing y OLAP aplicadas a bases de datos espaciales* Luís Joyanes Aguilar, [Consultado el: 25 de febrero de 2010]. Disponible en: http://novella.mhhe.com/sites/dl/free/8448118952/540197/ActasVol2SISOFT2006.pdf#page=113.
4. ABBEY, M. J. C. Y. M. *Oracle Data Warehousing.* 2002.
5. CASARES, C. *Data Warehousing* [Consultado el: 26 de febrero de 2010]. Disponible en: http://www.programacion.com/bbdd/tutorial/warehouse/.
6. CURTO, J. *Data Warehousing, Data Warehouse y Datamart* [Consultado el: 26 de febrero de 2010]. Disponible en: http://informationmanagement.wordpress.com/tag/data-warehousing/.
7. IMHOFF, C. *Mastering Data Warehouse Design, Relational and Dimensional Techniques.* 2003.
8. SINNEXUS. *Datamart* [Consultado el: 22 de febrero de 2010]. Disponible en: http://www.sinnexus.com/business_intelligence/datamart.aspx.
9. VÁZQUEZ, F. P. Y. G. *Relevamiento: Diseño Físico de Sistemas OLAP* [Consultado el: 25 de febrero de 2010]. Disponible en: http://www.fing.edu.uy/~fpiedrab/downloads/Physical_OLAP_Design.pdf.
10. KONCILIA, R. W. Y. C. *DATA WAREHOUSES AND OLAP.* 2007.
11. V., P. F. P. *Sistemas de Soporte a la toma de Decisiones* [Consultado el: 27 de febrero de 2010]. Disponible en: http://palomo.usach.cl/bdnc/2005-02/Presentaciones/U3-1-OLAP.pdf.

12. CURTO, J. *¿Qué es una Staging Area?* [Consultado el: 28 de febrero de 2010]. Disponible en: http://informationmanagement.wordpress.com/2007/10/15/%C2%BFque-es-una-staging-area/.

13. LORENA ETCHEVERRY, P. G., SALVADOR TERCIA. *Análisis del proceso de carga del Sistema de Data Warehousing de Enseñanza de la Facultad de Ingeniería* [Consultado el: 27 de febrero de 2010]. Disponible en: http://www.info.univ-tours.fr/~veronika/publications/tr0606-le.pdf.

14. WOLFF, C. G. *Modelamiento Multidimensional* [Consultado el: 25 de febrero de 2010]. Disponible en: http://www.inf.udec.cl/~revista/ediciones/edicion4/modmulti.PDF.

15. VERÁSTEGUI, H. C. *Modelado Dimensional de Datos* [Consultado el: 24 de febrero de 2010]. Disponible en: http://www.db-system.com/pls/portal/docs/PAGE/SITIOWWWDB/ARTICULOS/MODELADO%20DIMENSIONAL%20DE%20DATOS_V2.PDF.

16. CURTO, J. *Diseño de un data warehouse: tabla de hecho* [Consultado el: 26 de febrero de 2010]. Disponible en: http://informationmanagement.wordpress.com/tag/data-warehousing/.

17. GONZÁLEZ, Y. I. *Diseño e Implementación de un Data Warehouse para el Sistema de Gestión Estadística en Cuba.* Universidad de las Ciencias Informáticas, 2008.

18. SIERRA, J. E. O. *Diseño e Implementación de un Mercado de Datos para la Oficina Nacional de Estadísticas.* Universidad de las Ciencias Informáticas, 2009.

19. ALVAREZ, S. *Sistemas gestores de bases de datos* [Consultado el: 20 de febrero de 2010]. Disponible en: http://www.desarrolloweb.com/articulos/sistemas-gestores-bases-datos.html.

20. GONZÁLEZ, M. H. Á. *DISEÑO DE LA BASE DE DATOS DEL SISTEMA DE INFORMACIÓN DE PERFORACIÓN DE POZOS.* Universidad de las Ciencias Informáticas, 2009.

21. SANTOS, D. F. A. Y. Y. F. *Administración, configuración y optimización de un Sistema de Bases de Datos Descentralizado en Oracle Database 10g release 2* [Consultado el: 28 de febrero de 2010]. Disponible en: http://bibliodoc.uci.cu/TD/TD_0289_07.pdf.

22. MALDONADO, D. M. *SQLite, el motor de base de datos ágil y robusto* [Consultado el: 28 de febrero de 2010]. Disponible en: http://www.aplicacionesempresariales.com/sqlite-el-motor-de-base-de-datos-agil-y-robusto.html.

23. OCHOA, D. G. *Diseño e Implementación de un Almacén de Datos Operacionales para la Corporación CIMEX.* [Consultado el: 28 de febrero de 2010]. Disponible en: http://bibliodoc.uci.cu/TD/TD_2188_09.pdf.

24. ALFARO, F. M. *Herramientas Case* [Consultado el: 27 de febrero de 2010]. Disponible en: http://www.innovavirtual.org/campus/file.php/178/archivos_curso/CAP_12_2006_I_SI905/CAP_12_2006_I_SI905_VA8_M.pdf.

25. GONZÁLEZ, R. T. C. Y. L. G. *Análisis, diseño e implementación del Sistema de Gestión de Flotas por GSM/GPRS.* Universidad de las Ciencias Informáticas, 2009.

26. GSINNOVA, G. D. S. *Rational Rose Data Modeler* [Consultado el: 4 de marzo de 2010]. Disponible en: http://www.rational.com.ar/herramientas/rosedatamodeler.html.

27. COMPANY, E. *ER/Studio Enterprise* [Consultado el: 4 de marzo de 2010]. Disponible en: http://www.embarcadero.com/products/er-studio-enterprise.

28. CHAVEZ, M. P. Y. J. *Embarcadero Presenta a la Industria el Primer Kit de Herramientas para Desarrollo de Software Multiplataforma Bajo Demanda: Embarcadero® All-Access™* [Consultado el: 3 de marzo de 2010]. Disponible en: http://etnaweb04.embarcadero.com/news/press_releases/all-access-press-release-latam.pdf.

29. MARTÍNEZ, J. P. *DISEÑO DEL SUBSISTEMA PARA EL TRATAMIENTO Y MODELADO DE DIAGRAMAS DE FLUJO DE INFORMACIÓN* [Consultado el: 27 de febrero de 2010]. Disponible en: http://bibliodoc.uci.cu/TD/TD_2783_09.pdf.

30. CURTO, J. *CIF vs MD : Dos enfoques clásicos en el diseño de la arquitectura de un Data Warehouse* [Consultado el: 25 de febrero de 2010]. Disponible en: http://bi-businessintelligence.blogspot.com/2009/01/cif-vs-md-dos-enfoques-clsicos-en-el.html.

31. ZEPEDA SÁNCHEZ, L. Z. *Metodología para el Diseño Conceptual de Almacenes de Datos.* Universidad Politécnica de Valencia, 2008.

GLOSARIO DE TÉRMINOS

Indicadores: puntos de referencia, que brindan información cualitativa o cuantitativa, conformada por uno o varios datos, constituidos por percepciones, números, hechos, opiniones o medidas, que permiten seguir el desenvolvimiento de un proceso y su evaluación, y que deben guardar relación con el mismo.

Referencia: relación entre ciertas expresiones y aquello de lo cual se habla cuando se usan dichas expresiones.

Vertiginoso: que es muy rápido o intenso: un éxito vertiginoso; un ritmo vertiginoso.

Informática: ciencia aplicada que abarca el estudio y aplicación del tratamiento automático de la información, utilizando dispositivos electrónicos y sistemas computacionales. También está definida como el procesamiento automático de la información.

Automatizar: aplicar procedimientos automáticos a un aparato, proceso o sistema. Ej.

Han automatizado la biblioteca universitaria.

Inteligencia artificial: rama de las Ciencias de la Computación dedicada al desarrollo de agentes racionales no vivos.

Monitorización: control de las constantes vitales de un paciente a través de monitores.

Estadísticas: ciencia con base matemática referente a la recolección, análisis e interpretación de datos, que busca explicar condiciones regulares en fenómenos de tipo aleatorio.

Bases de datos: conjunto de datos pertenecientes a un mismo contexto y almacenados sistemáticamente para su posterior uso.

Implementación: poner en funcionamiento, aplicar los métodos y medidas necesarios para llevar algo a cabo: implementar un algoritmo.

Metodología: ciencia que estudia los métodos de conocimiento.

Funcionalidad: conjunto de características que hacen que algo sea práctico y utilitario: en el diseño de este vehículo se ha buscado la funcionalidad.

Multidimensional: que tiene varias dimensiones: espacio multidimensional.

Gestor de bases de datos: tipo de software muy específico, dedicado a servir de interfaz entre la base de datos, el usuario y las aplicaciones que la utilizan.

Digitalizar: transformar una información a un sistema de dígitos para su tratamiento informático: digitalizar una fotografía.

Volátil: mudable, inconstante: es de carácter volátil.

Requisito: circunstancia o condición necesaria para una cosa.

Modelo: arquetipo digno de ser imitado que se toma como pauta a seguir.

Procesamiento: tratamiento de la información: procesamiento de ficheros.

Indexar: indizar.

Sistema: conjunto de elementos dinámicamente relacionados formando una actividad para alcanzar un objetivo operando sobre datos, energía o materia para proveer información.

Funciones SSL: Secure Sockets Layer -Protocolo de Capa de Conexión Segura.

Cupet: Cuba Petróleo.

COBOL (COmmon Business - Oriented Language): Lenguaje Común Orientado a Negocios.

FORTRAN: lenguaje de programación alto nivel de propósito general, procedurimental e imperativo.

Tablas hash: estructura de datos que asocia llaves o claves con valores.

Consultas MDX: permiten consultar objetos multidimensionales, como los cubos, y devolver conjuntos de celdas multidimensionales que contengan los datos del cubo.

IDEs (Entorno de Desarrollo Integrado): programa informático compuesto por un conjunto de herramientas de programación.

Lenguaje DDL (Lenguaje de Definición de Datos): lenguaje proporcionado por el sistema de gestión de base de datos que permite a los usuarios de la misma llevar a cabo las tareas de definición de las estructuras que almacenarán los datos así como de los procedimientos o funciones que permitan consultarlos.

ANEXOS

Anexo 1 Herramienta para la recolección y análisis de la información.

Anexo 2 Especificación de Requisitos.

Anexo 3 Modelo de Casos de Usos del Sistema.

Anexo 4 Evaluación de áreas de la organización.