Universidad de las Ciencias Informáticas

Oficina Nacional de Estadísticas

Título: Diseño e Implementación de un Data Warehouse para el Sistema de Gestión Estadística en Cuba.

Trabajo de Diploma para optar por el título de

Ingeniero Informático

Autor: Yonelbys Iznaga González

Tutor: Ing. Carlos Yasmany Hidalgo

Ciudad de La Habana, Junio de 2008
"Año 50 de la Revolución"

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 _____.

_____ _____

Firma del Autor Firma del Tutor

"Sólo en el momento justo, en el lugar preciso y frente a oídos deseosos de nuevas melodías, habrás de pronunciar lo exacto para alcanzar la gloria."

John Elvis

AGRADECIMIENTOS

Ante todo agradecer a todos los hombres de ciencia y de bien que forjaron nuestro pasado y que acondicionaron nuestro presente para alcanzar lo que tenemos y conocemos hoy.

A mis queridos padres, Digna y Humbe, que tanto me han apoyado en esta lucha, siempre estuvieron y estarán presentes en mi corazón, donde quiera que esté.

A Yonelkis y a Yorelbis, mis dos hermanos, que siempre están conmigo, sepan que nada nos podrá separar.

A Mayra, por tu ayuda y apoyo en todos estos años de carrera, sin ti, todo habría sido más difícil.

A Elio Veidis, gracias por todo lo que hiciste por mí, tú formas parte de este resultado.

A mi abuela Nené, por ser tan buena y tan especial conmigo, te llevo siempre presente.

A todo el resto de la familia, en especial a mis tíos y tías que tanto apoyo y ayuda me dieron, a todos gracias.

A Ridosbey, por ser quien me recibió en un proyecto para integrarme a la producción de software y mostrarme este fascinante mundo de los Sistemas Data Warehouse.

A todos mis amigos, que me brindaron su apoyo incondicional, gracias de verdad.

A todos los que contribuyeron de una forma u otra en la elaboración de este trabajo, les agradezco de todo corazón.

DEDICATORIA

Dedico este trabajo a todos mis familiares y amigos.

RESUMEN

Entre los principales proyecto de nuestra Universidad se destaca el Sistema Integrado para la Gestión Estadística (SIGE), el cual desarrolló un complejo producto de software para la Oficina Nacional de Estadísticas (ONE), que se encargará del control, manejo, colección y publicación de la información estadística de una manera rápida y eficiente. En medio del desarrollo de este software y debido a la necesidad por parte de la ONE de manejar un inmenso cúmulo de información, además de su análisis e investigación, entre otras causas, se propone diseñar e implementar un Data Warehouse para el logro de tales objetivos. Constantemente los distintos directivos de los organismos y de la administración central del Estado y del Gobierno, así como los máximos líderes de nuestro país indagan y analizan distintas informaciones estadísticas, necesitan informes y reportes pormenorizados tanto productivos como administrativos, sociales, de ámbito económico, educacional, de salud, etc., para lograr una eficiente y efectiva dirección de nuestro país, y es la Oficina Nacional de Estadísticas la encargada de ofrecer estos servicios, mediante el uso y explotación de un Data Warehouse que posibilitará el acceso rápido y eficiente a un inmenso cúmulo de información histórica, ordenada e integrada, mediante complejas consultas o simples pedidos estadísticos, así como el estudio y análisis de los datos y el descubrimiento de patrones, sobre la base de la información almacenada.

ÍNDICE

INTRODUCCIÓN:

El impetuoso avance y desarrollo indiscutible de las ciencias informáticas y de las comunicaciones han marcado pautas y diferencias notables hoy en día con respecto a épocas anteriores y más aún cuando se trata de elevar las potencialidades tecnológicas de la sociedad, luego es una necesidad imperiosa tanto para los países en desarrollo como para los desarrollados la adquisición y uso ético de las mismas para el logro de un desarrollo sostenible e incremental.

Sin embargo nuestro país no queda exento del esfuerzo por un desarrollo científico, tomando como principal fuerza de trabajo su capital humano. En consecuencia ha puesto en práctica un conjunto de estrategias y vías de solución, concretamente se puede hablar de la idea del surgimiento de una universidad enfocada al campo informático, y objetivamente sobre la línea de producción de software, solución económicamente viable para un país en desarrollo. Es así que surge la Universidad de las Ciencias Informáticas, un centro de altos estudios encaminado a la formación de futuros jóvenes profesionales en este campo y paralelamente dedicado a la producción de productos informáticos.

Entre las disímiles tareas encargas a esta Universidad, se encuentra la informatización de las oficinas de estadísticas en todo el país, llevando a cabo el desarrollo del Sistema Integrado de Gestión Estadística (SIGE) el cual constituye un paquete con diversos módulos que automatizan y gestionan todo lo relacionado con la información estadística a nivel nacional, y regido por su principal institución, la Oficina Nacional de Estadísticas (ONE), encargada fundamentalmente de la dirección, control, organización y regulación de toda la información estadística en Cuba, la cual sirve de análisis para el proceso de apoyo en la toma de decisiones por parte de la dirección del país.

Centralmente en el nivel superior se encuentra la ONE, la cual rige y controla la actividad estadística de todas las provincias a través de las Oficina Territorial de Estadísticas (OTE) ubicadas cada una en dieciséis territorios de nuestro país, una en cada provincia, otra en el municipio especial Isla de la Juventud y otra adicional en Ciudad de La Habana. En un tercer nivel y regidos por las OTE se encuentran las Oficinas Municipales de Estadísticas (OME) en cada uno de los ciento sesenta y nueve municipios, los cuales a su vez controlan los centros informantes, último eslabón en esta cadena de la actividad estadística, tan importante para nuestro país.

Las empresas, en sus actividades de negocios han coleccionado los datos operacionales durante años, y estas cantidades continúan aumentando a medida que pasa el tiempo, en paralelo con el crecimiento en complejidad de las redes de comunicación y el flujo de comercio.

Durante varios años los Data Warehouse o Almacenes de datos (en el habla hispana) se han usado para apoyar a los directores, gerentes, especialistas y hombres de negocios en las decisiones comerciales. Una de sus características especiales es el modelo dimensional, lo cual no es un arte o una ciencia; sino que constituye una metodología madura que organiza los datos en una representación espacial simple e intuitiva de manera que los especialistas puedan observar y analizar sus datos.

MARCO TEÓRICO

Situación problémica:

En la Oficina Nacional de Estadísticas debe existir un sistema eficiente y dinámico de acceso a la información, con la rapidez requerida en los reportes y consultas, donde la información estadística se encuentre centralizada, ordenada e integrada, sin embargo, hoy esta entidad cuenta con grandes bancos de datos, es decir, carpetas que guardan los ficheros en varios formatos con la información histórica la cual no se encuentra integrada, lo que hace muy complejo y lento el acceso a la información, los reportes requeridos y las consultas necesarias.

Problema científico:

Después de un análisis de la situación problémica, queda conformado el problema científico mediante el cuestionamiento de:

¿Cómo lograr la integración de todos los datos almacenados históricamente en la Oficina Nacional de Estadísticas de manera que su acceso sea eficiente?

Objeto de estudio: Sistemas Informacionales

Campo de acción: Sistemas Data Warehouse.

Objetivo general: Desarrollar un Data Warehouse para el Sistema Integrado de Gestión Estadística.

Hipótesis:

Si se concreta el desarrollo de un Sistema Data Warehouse para la Oficina Nacional de Estadísticas, se contribuirá al proceso de integración de sus datos históricos así como la optimización dinámica de las consultas y los reportes que debe brindar dicha entidad.

Tareas de la investigación:

- Estudiar sistemáticamente los temas relacionados
- Seleccionar el proceso de negocio a modelar.
- Elegir el grano del proceso de negocio
- Definir las dimensiones.
- Definir los hechos mensurables.
- Obtener las copias de los datos operacionales.
- Confeccionar el esquema de diseño físico.
- Configurar la extracción y transformación de datos.
- Construir el sistema de almacenamiento de datos.
- Evaluar la validez de los datos y la velocidad de respuesta de los reportes.

Estructuración del trabajo:

Este trabajo se compone de 3 capítulos, de los cuales el primero aborda los distintos temas relacionados con la fundamentación teórica, el segundo trata sobre la descripción de la solución y el tercero está orientado al análisis de los resultados.

En el Capítulo 1 los puntos implicados están referidos a un estudio sobre los Sistemas Informacionales, los Sistemas Data Warehouse, un estudio del arte tanto a nivel mundial como nacional, de los Sistemas Data Warehouse, de las herramientas en usabilidad, así como también los objetivos y características de dichos sistemas, y los modelos que se utilizan para su diseño.

En el Capítulo 2 se abordan aspectos concernientes a la arquitectura, la interacción entre los componentes, el diseño y la implementación del Data Warehouse, los tipos de almacenamiento que existen, el cubo de datos y el modelo multidimensional propuesto.

Finalmente en el Capítulo 3 se detallan las temáticas referidas a la normalización, el tamaño y crecimiento del Sistema, el análisis del tiempo de respuesta de las consultas y reportes así como la validación general del Sistema y el análisis de los resultados.

CAPITULO 1: FUNDAMENTACIÓN TEÓRICA.

El siglo XXI parece heredar del XX todo un espectro amplio de tecnologías y conocimientos para ser implantados y esparcidos por el mundo en un intento por el logro de un desarrollo superlativo e inigualable, jamás visto por generaciones de épocas pasadas. Es así que entre sus grandes avances se encuentran los cada vez más famosos Data Warehouse, denominados también en el habla hispana Almacenes de Datos, los cuales constituyen una alternativa de gestión para el negocio empresarial, convirtiéndose para quienes lo usan y explotan en un sistema casi inteligente para entender y analizar la más compleja consulta o para ofrecer el más grande y engorroso reporte, además de constituir un motor de dirección y guía en un mundo de marketing cada vez más complejo, principalmente para aquellos directivos de empresas que necesitan de sistemas o herramientas que los ayuden a tomar decisiones importantes, a dirigir con eficacia una importante entidad económica y a plantearse metas y objetivos de mediano y largo alcance.

1.1 LOS SISTEMAS INFORMACIONALES:

Para comprender mejor todo lo relacionado con el inmenso mundo de los Sistemas Informacionales, se comenzará diferenciando los sistemas en cuanto a sus finalidades; las cuales pueden ir desde el apoyo para realizar las funciones específicas de la empresa y manipular el negocio, hasta el estudio y el comportamiento de la organización para proyectar nuevas estrategias.

Se tienen así, por una parte los Sistemas Operacionales o Sistemas de Producción, los cuales se utilizan para garantizar el desempeño del negocio en tiempo real; es decir, se encargan de registrar y controlar los procesos que constituyen el núcleo del comportamiento diario de la empresa. Entre los objetivos de estos sistemas operacionales, además de apoyar las funciones diarias de la empresa, se encuentran brindar servicios de oficina y entregar la información de manera automatizada, asegurando la calidad y la protección de la misma.

Por otro lado, los Sistemas Informacionales se utilizan para administrar y controlar la empresa, es decir, se basan en puntos estables en el tiempo o datos históricos y se diseñan principalmente, para ejecutar consultas complejas, no planas o dimensionales y de sólo lectura. El objetivo fundamental de los sistemas informacionales es posibilitar el hecho de mantener disponible un compendio de la

información histórica garantizando una fuente única, contribuyendo a realizar análisis y toma de decisiones estratégicas a largo plazo.

Uno de los aspectos esenciales que distingue a los sistemas operacionales de los informacionales, es el tipo de procesamiento de la información que realizan, a saber, transaccional o analítico. El Procesamiento Transaccional en Línea, OLTP,(On Line Transaction Processing), conocido también como Procesamiento Operacional, soporta las operaciones diarias del negocio y responde a los requerimientos del comportamiento diario de una organización. Mientras que el Procesamiento Analítico en Línea, OLAP (On Line Analitical Processing), conocido también como Procesamiento para la Toma de Decisiones, soporta las actividades de investigación y navegación del usuario terminal, sustenta el estudio del comportamiento del negocio y su proyección, se caracteriza por un análisis dimensional y dinámico -desde diferentes puntos de vista- de los datos consolidados de la empresa, ayudando al usuario a sintetizar la información de la empresa a través de vistas personalizadas, análisis históricos y pronósticos.

Básicamente se utilizan para administrar y controlar la empresa. Se basan en datos estables en un momento en el tiempo o periódicos, llamados datos históricos, y se diseñan, principalmente, para ejecutar consultas que involucran perspectivas de los datos, complejas y de solo lectura.

Una vez detallada algunas de las características de los sistemas informacionales, se puede decir que constituye la herramienta esencial para dirigir el procesamiento de la información hacia el análisis y la toma de decisiones.

Evidentemente y formando parte de este gran grupo se encuentran los Sistemas Data Warehouse que pasaría a ser nuestro centro de atención en el presente trabajo.

1.2 LOS SISTEMAS DATA WAREHOUSE:

Para comenzar a hablar sobre los orígenes del Data Warehouse se debe partir explicando que hacia los años sesenta, con el surgimiento de los primeros sistemas computarizados, los procesos de negocios se fueron haciendo aunque lentamente, más complejos debido a que las corporaciones fueron adquiriendo un significativo aumento de tamaño en cuanto a tecnología, capital humano y aplicaciones informáticas, las cuales durante bastante tiempo lograron manejar satisfactoriamente todo el volumen de información y soportar todos los procesos a los cuales estaban sujetos, pero ya en las puertas de la década del noventa, y con la consecuente competencia mundial de mercado, la inevitable

globalización de las tecnologías y la aglomeración descomunal de información, no era suficiente con que los sistemas operacionales hasta ese momento existentes realizarán múltiples cálculos diarios o computaran millones de procesos continuos, sino que se abría paso a la necesidad de tomar decisiones estratégicas, adoptar certeros criterios de negocios y ofrecerle a los usuarios rápidos, concretos y resumidos informes sobre la base de grandes montañas de información y conocimiento. (Ponniah, P. 2001).

Es por eso que en la década de los ochenta y noventa los Data Warehouse vienen a relucir como paradigma en el desarrollo estratégico de las empresas y en el proceso de ofrecer información y conocimiento, valiosos para conducir por el camino correcto en el abstracto mundo de las decisiones estratégicas logrando de forma factible hacer consultable la información que se tiene de una empresa o institución determinada tanto de meses como de años anteriores.

De esta manera sobre los Data Warehouse existen varias definiciones ofrecidas por grandes especialistas y profesionales del tema, que aunque difieren en algunos aspectos, giran todas sobre el mismo eje central, por ejemplo:

W.H. Inmon, considerado el padre del Data Warehouse lo define como “un conjunto de datos integrados, orientados a una materia, que varía con el tiempo y que no son transitorios, los cuales soportan el proceso de toma de decisiones de la administración. “ (Inmon, W. H., 2001)

Ralph Kimball, otro gran especialista sobre estas colecciones de datos, reconocido mundialmente y escritor del famoso libro “The Data Warehouse Toolkit” afirma que es "una copia de los datos de la transacción estructurados específicamente para la pregunta y el análisis." (Kimball, R. y Margy, R. 2002).

Otra de las definiciones plantea que constituye una base de datos integrada para el soporte de decisiones cuyo contenido es derivado de varias bases de datos operacionales. (Hoffer, Prescott, & McFadden, 2005; Sen & Jacob, 1998).

Se debe ver más bien como un proceso en espiral donde la implantación inicial conducirá a los usuarios finales al planteamiento de nuevas interrogantes y necesidades que a su vez llevarán al grupo de desarrolladores a la concesión de nuevas estructuras y entidades que satisfagan las actuales necesidades. Dicho proceso se repetirá continuamente creando un ciclo que fomente el desarrollo del Data Warehouse en cuestión. Constituye pues una nueva manera de pensar en los datos. Lo primero que se necesita es realizar un estudio de la estructura y funcionamiento de la empresa. Esta etapa se denomina modelación de la empresa y en ella también es imprescindible definir las necesidades de los

usuarios finales del Data Warehouse. Estos se orientan a entidades, es decir, como resultado de la modelación previa de los procesos y las entidades de una empresa se genera un modelo empresarial que luego se utiliza para generar las estructuras y el flujo informativo específico de la organización. Es importante destacar entonces que solo se puede contar con los componentes y las herramientas, mientras que la estructura y el desarrollo del sistema dependen enteramente de las habilidades de los diseñadores e implementadores. (Veliz Monteagudo, Mijaíl. 2003).

En fin, conceptualmente se trata de una colección de datos, muy conocido también como Almacén de datos, no igualmente diseñado y estructurado como los sistemas transaccionales (las bases de datos comunes), que además de sus tantas características, bien sintetizadas en el concepto ofrecido por Inmon se constituye como banco central de datos el cual se nutre de varias bases de datos de tal manera que todos puedan acceder a su información mediante las consultas y reportes, dando respuestas a las necesidades de los distintos tipos de usuarios.
Quizá mucho más sintetizado, se puede decir que se trata de un sistema integrado y multidimensional de almacenamiento histórico de datos estables, funcionalmente dirigido a consultas, reportes y análisis de información.

1.3 OBJETIVOS GENERALES DE LOS DATA WAREHOUSE

Como parte de las metas a mediano y largo plazo por las cuales se rigen estos sistemas de especial almacenamiento de información, se pueden diferenciar cuatro objetivos fundamentales y de mayor importancia para el diseño, implementación y mantenimiento de los Data Warehouse, estos son: (Kimball, R. y Margy, R. 2002).

1- **Lograr que la información sea fácilmente accesible.**

De cualquier manera el contenido del Data Warehouse debe ser entendible y legible tanto para los que la diseñan y construyen como para los que la usan finalmente. Los reportes devolverán resultados que servirán para el análisis de la información, en tanto que las herramientas que acceden al almacén de datos, así como los modelos multidimensionales y diseños realizados deben ser de fácil uso, mejorando así el proceso para devolver consultas con grandes volúmenes de datos en el mínimo tiempo posible.

2- **Lograr que la información sea consistente.**

Teniendo una información constante se contribuye a su alta calidad. Además de entendibles, los datos deben ser creíbles y palpables. Los datos, los cuales provienen de varias fuentes deben estar cuidadosamente ensamblados, depurados y con buena calidad. Si por ejemplo, dos medidas de funcionamiento tienen el mismo nombre entonces deben significar lo mismo. En caso contrario, si dos medidas no significan lo mismo, entonces deben ser calificadas de forma diferente. La consistencia también implica que las definiciones comunes para el contenido del almacén de los datos están disponibles para los usuarios.

3- Lograr que los la información almacenada sea adaptable y resistente al cambio.

No se resuelve nada con evitar los posibles cambios que puedan ocurrir. Tanto las necesidades del usuario como las condiciones de negocios, los datos y la tecnología están sujetas a cambios posibles que puedan ocurrir por lo que el Data Warehouse debe ser diseñado para manejar cualquier transformación inevitable. Inclusive, pueden ocurrir cambios que no afecten o que no invalidan los datos o usos existentes, lo cuales no deben ser cambiados o ser interrumpidos cuando la comunidad de negocio pide nuevas preguntas o se agreguen nuevos datos al almacén. Si los datos descriptivos en el almacén de datos se modifican, se deben explicar los cambios apropiadamente.

4- Lograr que la información almacenada esté totalmente segura y protegida

Puesto que generalmente se trabaja con información de empresas, centros de prestigio o entidades gubernamentales, los datos almacenados deben gozar de la más absoluta seguridad. El data Warehouse debe así mismo controlar el acceso a la información que la empresa considera es confidencial o de limitado acceso.

1.4 CARACTERÍSTICAS DE LOS DATA WAREHOUSE

Existen según W. H. Inmon cuatro características fundamentales que describen a los almacenes de datos: (Inmon, W. H., 2001).

1. Orientado al sujeto:

Una primera característica del Data Warehouse es que la información se clasifica en base a los aspectos que son de interés para la empresa. Siendo así, los datos tomados están en contraste con

los clásicos procesos orientados a las aplicaciones. El ambiente operacional se diseña alrededor de las aplicaciones y funciones tales como préstamos, ahorros, tarjeta bancaria y depósitos para una institución financiera. Por ejemplo, una aplicación de ingreso de órdenes puede acceder a los datos sobre clientes, productos y cuentas. La base de datos combina estos elementos en una estructura que acomoda las necesidades de la aplicación.

En el ambiente Data Warehouse se organiza alrededor de sujetos tales como cliente, vendedor, producto y actividad. Por ejemplo, para un fabricante, éstos pueden ser clientes, productos, proveedores y vendedores. Para una universidad pueden ser estudiantes, clases y profesores. Para una oficina de estadística pueden ser los indicadores de la producción a medir, los modelos estadísticos a llenar y los clasificadores de empresas. Para un hospital pueden ser pacientes, personal médico, medicamentos, etc.

La alineación alrededor de las áreas de los temas afecta el diseño y la implementación de los datos encontrados en el Data Warehouse. Las principales áreas de los temas influyen en la parte más importante de la estructura clave.

2. Integrado:

El aspecto más importante del ambiente Data Warehouse es que la información encontrada al interior está siempre integrada.

La integración de datos se muestra de muchas maneras: en convenciones de nombres consistentes, en la medida uniforme de variables, en la codificación de estructuras consistentes, en atributos físicos de los datos consistentes, fuentes múltiples y otros.

Cuando los datos residen en muchas aplicaciones separados por los distintos entornos operacionales, la descodificación de los datos es a menudo inconsistente. Por ejemplo, en una aplicación, la palabra “género” podría codificarse como "m" y "f" en otra como "0" y "1". Cuando los datos fluyen de un entorno operacional a un entorno de Data Warehouse, ellos asumen una codificación consistente, por ejemplo “género” siempre se transformaría a "m" y "f".

3. De tiempo variante

Toda la información del almacén de datos es requerida en algún momento. Esta característica básica de los datos en un depósito, es muy diferente de la información encontrada en el ambiente

operacional. En éstos, la información se requiere al momento de acceder. En otras palabras, en el ambiente operacional, cuando usted accede a una unidad de información, usted espera que los valores requeridos se obtengan a partir del momento de acceso.

Como la información en el Data Warehouse es solicitada en cualquier momento (es decir, no "ahora mismo"), los datos encontrados en el depósito se llaman de "tiempo variante". Los datos históricos son de poco uso en el procesamiento operacional. En cambio la información del almacén de datos, debe incluir los datos históricos para usarse en la identificación y evaluación de tendencias así como en comparaciones y previsiones.

4. No volátil:

La información es útil sólo cuando es estable, es por eso que los datos no serán modificados o cambiados de ninguna manera una vez ellos han sido introducidos en el almacén de datos, solamente podrán ser cargados, leídos o accedidos. La actualización de información que implican procesos de inserción, eliminación y actualización, se hace regularmente en el ambiente operacional sobre una base de registro por registro. Pero la manipulación básica de los datos que ocurre en el Data Warehouse es mucho más simple. Hay dos únicos tipos de operaciones: la carga inicial de datos y el acceso a los mismos. No hay actualización de datos (en el sentido general de actualización), como una parte normal de procesamiento.

1.5 MODELO DE DATOS ENTIDAD- RELACIÓN Y MODELO DIMENSIONAL

1.5.1 El Modelo de Datos Entidad –Relación

La diferencia más importante y final entre los sistemas OLTP, propios de las bases de datos comunes y el almacén de datos es la organización de los datos en los sistemas, o más simple, el modelo de datos. Para entender por qué el dato es organizado tan distantemente, se necesita retornar a la emisión del desempeño de transacciones. Muchos de los milagrosos beneficios en la ejecución de transacciones son debido a la técnica denominada modelación entidad – relación. Este modelo busca dirigir la eliminación de toda redundancia de los datos. Si no existe redundancia en los datos, entonces una transacción que cambia cualquier dato (o adiciona o elimina datos) solo necesita tocar la base de datos en un lugar. Este es el secreto detrás del perfeccionamiento fenomenal en la velocidad del procesamiento de transacciones a partir de los inicios de los 80.

La modelación entidad – relación trabaja dividiendo los datos en muchas entidades discretas, cada una de las cuales se convierte en una tabla en la base de datos OLTP.

Existe un número importante de observaciones que hacer sobre estos diagramas entidad – relación, orientadas a OLTP. Primero, en un extraño orden de formas, este diagrama es muy simétrico. Todas las tablas parecen iguales. No existe manera para decir que tabla es más importante o la mayor. No existe forma de decir que tablas contienen medidas numéricas de los negocios y que tablas incluyen descriptores estáticos o cuasi – estáticos de los objetos. Esa simetría significa que uno puede revolver el diagrama medianamente más arbitrariamente y este parecerá ser el mismo. Los diagramas similares a este son muy difíciles para las personas (usuarios finales o diseñadores) visualizar y conservar en sus cabezas. (Ponniah, P. 2001).

Para consultas, que alcanzan muchos registros y muchas tablas, los diagramas entidad – relación, son muy complejos de comprender por los usuarios y muy complejos de recorrer por el software.

1.5.2 El Modelo Dimensional.

El Data Warehouse se soporta sobre el modelo dimensional a diferencia de los sistemas de bases de datos que están basados en el modelo Entidad-Relación. Este modelo contiene la misma información que el modelo E/R pero empaqueta los datos en un formato simétrico cuyo objetivo es ganar una mayor comprensión del usuario y garantizar la ejecución rápida y eficiente de las consultas. A diferencia del modelo E/R, el modelo dimensional no necesita anticipar las consultas que se van a realizar y es muy elástico a los cambios que se produzcan en los patrones de los usuarios.

Este tipo de modelo es muy comúnmente llamado “esquema de enlace estrella” debido a que los esquemas se parecen a una estrella, con una gran tabla central y un conjunto de pequeñas tablas acompañantes presentadas en un modelo radial alrededor de la tabla central.

El modelo dimensional divide el mundo de los datos en dos grandes tipos: las medidas y las descripciones del entorno de estas medidas. Las medidas, que generalmente son numéricas, se almacenan en las tablas de hechos y las descripciones de los entornos que son textuales se almacenan en las tablas de dimensiones. Las tablas de hechos son las tablas primarias en el modelo dimensional y contiene los valores del negocio. Los hechos más comunes son valores numéricos. Cada tabla representa una interrelación muchos – muchos y contiene dos o más llaves extranjeras que acoplan con sus respectivas tablas de dimensiones. (Ponniah, P. 2001).

Las tablas de dimensiones son las compañeras de las tablas de hechos. Cada dimensión se define por su llave primaria que sirve para mantener la integridad referencial en la tabla de hechos a la que se acopla. Los atributos de estas tablas sirven de base a las solicitudes que se hacen al Data Warehouse.

A diferencia del modelo entidad – relación, el modelo dimensional es muy asimétrico. Existe aquí una gran tabla dominante en el centro del esquema. Ella es la única tabla en el esquema con múltiples enlaces conectándola a otras tablas. Las otras tablas tienen un enlace simple que las enlaza con la tabla central. Se le llama a la tabla central la tabla de hechos y a las otras tablas las tablas dimensionales.

Las tablas de dimensiones contienen información jerárquica que permitirán la realización de las agregaciones o las profundizaciones. En la siguiente figura se visualiza una representación de este tipo de modelo con las tablas dimensionales Producto, Cliente y Tiempo, y en el centro la tabla de hechos, con la cual se relacionan las restantes tablas mediante sus respectivos identificadores, es decir, que se encontrarían en la tabla de hechos las llaves primarias de las tablas dimensionales, más los atributos que representan las medidas o hechos mesurables, como ventas, compras, transacciones, etc. (Kimball, R. y Margy, R. 2002)..

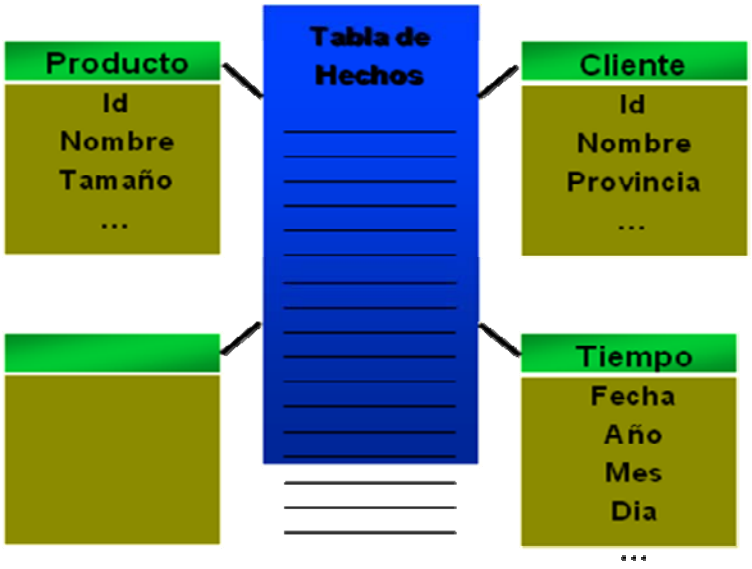


Figura 1.1: Representación del modelo multidimensional.

1.6 ESTADO ACTUAL DE LOS DATA WAREHOUSE:

En un principio en la mayor parte de las empresas, esta necesaria capitalización de la información comercial ha venido de la mano de la incorporación de bases de datos relacionales. El modelo relacional tiene entre sus objetivos guardar la integridad de los datos obtenidos en los procesos transaccionales automatizados (OLTP: Procesamiento Transaccional en Línea). Sin embargo, este modelo no se corresponde con la forma en la que el usuario percibe la gestión del conocimiento de un negocio, en general, y la gestión del conocimiento comercial, en particular. De hecho, Codd afirmó que aunque los sistemas de gestión de bases de datos relacionales, han sido muy beneficiosos para los usuarios, nunca han sido diseñados para proporcionar funciones potentes de síntesis, análisis y consolidación de los datos.
Teniendo presente ese hecho y el hecho de que la economía actual está centrada en el cliente, las corporaciones deben impulsar diversos esfuerzos técnicos y metodológicos para intentar acometer el objetivo de reenfocar su atención a la relación con el cliente.

1.6.1 SISTEMAS DATA WAREHOUSE EN EL MUNDO:

Hoy en día en el mundo son utilizados los Data Warehouse de forma significativa por miles de empresas y entidades, por ejemplo, en la industria minorista se pude mencionar de American Stores (Estados Unidos), Canadian Tyre (Canadá), WH Smith Books (Gran Bretaña), Great Universal (Gran Bretaña), Supermercados Casino (Francia), Migrosgenossenschaftsbund (Suiza), Otto Versand (Alemania). Las mismas necesitan hacer un manejo cada vez más ágil de la información para mantenerse competitivas en la industria. Los Data Warehouse aquí se utilizan para predecir la cantidad de producto que se venderá a un determinado precio y, por consiguiente, producir la cantidad adecuada para una entrega "justo a tiempo". A su vez se coordina el suministro a las grandes cadenas minoristas con inmensas cantidades de productos "en consignación", que no son pagados hasta que estos productos son vendidos al consumidor final.

Otras empresas que cuentan con Almacenes de datos de importancia son: Coca Cola, Nike, Procter & Gamble, Hallmark, Maybelline, Helene Curtis, 3M, Owens Corning Glass, Karsten Ping Golf Clubs y Walt Disney.

En el área del transporte de cargas y pasajeros se utilizan entre otras cosas para almacenar y acceder a meses o años de datos de clientes y sistemas de reservas para realizar actividades de marketing,

planeamiento de capacidad, monitoreo de ganancias, proyecciones y análisis de ventas y costos, programas de calidad y servicio a clientes.

Las empresas de transporte de cargas llevan datos históricos de años, de millones de cargamentos, capacidades, tiempos de entrega, costos, ventas, márgenes, equipamiento, etc. En este campo se pueden mencionar empresas de magnitud como: Cornrail, Union Pacific, Norfolk Southern, American President Lines, Delta, Lufthansa, QANTAS, British Airways, American Airlines, Canadian Airlines y SNFC.

En las telecomunicaciones se están utilizando fundamentalmente para operar en un mercado crecientemente competitivo, desregulado y global que, a su vez, atraviesa profundos cambios tecnológicos. Se almacenan datos de millones de clientes: sus circuitos, facturas mensuales, volúmenes de llamados, servicios utilizados, equipamiento vendido, configuraciones de redes, etc. así como también información de facturación, utilidades, y costos son utilizadas con propósitos de marketing, contabilidad, reportes gubernamentales, inventarios, compras y administración de redes.

Muchas otras industrias y actividades utilizan actualmente, o están comenzando a instalar, Data Warehouse como por ejemplo entidades gubernamentales, especialmente para el control impositivo, empresas de servicios públicos, de entretenimiento, editoriales, fabricantes de automóviles, empresas de petróleo y gas, laboratorios farmacéuticos, droguerías, etc. En la industria informática NCR se cuenta con los Data Warehouse de mayor magnitud y antigüedad. Sus mayores instalaciones se encuentran en distintos centros de la compañía en Estados Unidos. Por ejemplo en la NCR San Diego, California, se encuentra el centro de desarrollo de los computadores WorldMark. Sobre los mismos se realizó la demostración del Data Warehouse más grande del mundo: 10 Terabytes de información (=10.000 Gigabytes=10.000.000 Megabytes), para poner esto en términos manejables se debe considerar que toda la información escrita de la Biblioteca del Congreso de los Estados Unidos se podría almacenar en unos 20 Terabytes.

Otras organizaciones como *Bacardí Martini* (distribución de bebidas) utiliza la información de ventas existente en el Data Warehouse para optimizar la utilización de recursos con el fin de lograr el máximo de ventas con un coste preestablecido de antemano.
Pierre Fabré Ibérica (laboratorio multinacional cosmético y farmacéutico) utiliza un Data Warehouse comercial para el seguimiento de ventas por zona geográfica, organización comercial, por producto, cliente, cadena y campaña etc., integrado en la aplicación de red de ventas, produce también un extenso informe mensual requerido por la casa matriz francesa.

Pastas *La Familia* (producción y distribución de alimentos) cuenta con un Data Warehouse comercial que se destaca por la integración de la información presupuestaria en el ámbito de familia de producto y cadena, genera hojas electrónicas con información real del año en curso, sobre las cuales el departamento correspondiente calcula los presupuestos del próximo año.

SEUR (empresa de mensajería y transporte de paquetes) posee un DW de más de 80 millones de registros para seguimiento estadístico de los movimientos operativos, que permite realizar unos análisis mucho más detallados y precisos de envíos por ejemplo por origen y destino, por volumen, peso o precios de envío.

El diario *El Mundo* cuenta con un Data Warehouse cuyo objetivo es obtener información completa sobre la contratación de publicidad en sus medios.

1.6.2 SISTEMAS DATA WAREHOUSE EN CUBA:

Pero nuestro país no ha quedado exento de este desarrollo sobre Almacenes de Datos. A pesar de los muchos esfuerzos que se están realizando hoy en día, ya se cuenta con el desarrollo de varios de estos, destacando sin dudas el más sobresaliente, perteneciente a CIMEX, corporación dedicada fundamentalmente a la Exportación e importación de mercancías, cuyo Data Warehouse centra su atención en la actividad del comercio, principalmente en la gestión de inventario, permitiendo una gestión de compra–venta eficiente con el objetivo fundamental de disminuir los costos sin afectar al cliente, permitiendo prestaciones eficientes y con la calidad requerida, aumentando las ganancias o utilidades de las empresas.

En la Feria Informática 2002 se presentó un Data Warehouse por y para Cubacel el cual basado en Oracle brinda amplias posibilidades para el diseño, la implementación y la administración de un sistema de este tipo.

Otras entidades como UNION CUPET y Copextel se encuentran en franco proceso de diseño y desarrollo al respecto. (Veliz Monteagudo, Mijaíl. 2003).

1.7 HERRAMIENTAS EN USABILIDAD:

Actualmente se encuentran en uso numerosas herramientas destinadas al diseño, construcción, implementación y mantenimiento de almacenes de datos. Definitivamente, uno de los gigantes informáticos , la compañía de software Oracle lleva la delantera sobre dichas herramientas, tanto para optimizar el tiempo y fortalecer las formas y métodos de consultas complejas y reportes, como para el desarrollo de sistemas estratégicos de negocios empresariales, con amplia aplicabilidad en los llamados sistemas inteligentes de mercado. Con su producto Oracle Warehouse Builder 10g han

logrado afianzarse, y mucho más con la salida de su segunda versión, debido a que constituye una herramienta multiplataforma que ayuda a los clientes a administrar el ciclo de vida de almacenes de datos desde el diseño hasta la implementación y mantenimiento. Esta nueva versión incorpora importantes funciones de administración, integración y calidad de datos, para los usuarios que buscan una herramienta fácil de usar a fin de diseñar, implementar y administrar rápidamente proyectos de integración de datos y sistemas de Business Intelligence (BI). Brinda calidad de datos, auditoría de datos, modelado dimensional y relacional totalmente integrado y gestión de todo el ciclo de vida de datos y metadatos de Oracle Database.

El mismo incluye la habilidad de diseñar estructuras de base de datos OLAP y relacionales, facilitando la tarea de almacenar datos en un repositorio común de Oracle Database y de ofrecer a los usuarios una elección de herramientas de BI, tales como Oracle Business Intelligence Suite, planillas de cálculo, etc.

En sus opciones se soportan implementaciones de múltiples entornos típicos de los proyectos de Data Warehouse para empresas, permitiendo un mejor desempeño y escalabilidad de los procesos de ETL. En esta opción se incluye una línea nueva y altamente interactiva, un analizador de impacto, objetos y asociaciones definidas por los usuarios, y un propagador de cambios que facilita la rápida respuesta a los metadatos en constante cambio.
Sus conexiones permiten a los clientes extraer datos rápida y fácilmente, y en algunos casos, dirigir los datos hacia sus aplicaciones centrales de CRM y ERP, incluidas Oracle E-Business Suite y PeopleSoft Enterprise de Oracle. Incluye soporte para apuntar a bases de datos que no pertenecen a Oracle, característica que permite a los usuarios elegir el lugar donde sus datos serán almacenados finalmente.

“Experts” es una característica nueva y exclusiva que posibilita a los usuarios ahorrar tiempo y costos. Esto permite a las empresas encapsular sus propios estándares de desarrollo y las mejores prácticas como wizards; además que una amplia gama de usuarios, incluidos los usuarios finales, accedan a la funcionalidad declarativa de Oracle Warehouse Builder.

Otra de las herramientas destacadas en este ámbito es la implementada por la compañía Microsoft nombrada Microsoft SQL Server 2000 Analysis Services el cual resulta una extensión del anterior paquete de componentes OLAP Services, y que incluye la tecnología OLAP (Procesamiento Analítico en Línea) y la minería de datos especialmente usado para el descubrimiento de información en los

cubos OLAP y las bases de datos relacionales. Dichos cubos pueden ser creados con múltiples configuraciones y ser actualizados en tiempo real al ocurrir cambios en las fuentes operacionales.

Otra de sus mejorías es la forma de crear las dimensiones, con nuevos tipos y características. Es el caso del tipo de dimensión Padre- Hijo, donde se representa la información dependiendo del negocio, en una estructura jerárquica.

También esta herramienta incluye características que proveen de mayor flexibilidad y control de acceso al cubo de datos, con métodos adicionales de autentificación de usuarios y roles.

En el plano de las herramientas Open Source, se destacan el Mondrian, el cual es un servidor OLAP escrito en Java/Servlets/JSPs que se puede instalar en servidores de aplicaciones como JBoss. Mondrian permite interactivamente analizar grandes cantidades de información almacenada en cualquier Base de Datos que soporte JDBC.

Mondrian soporta el lenguaje Microsoft's Multidimensional Expressions (MDX). También soporta los APIs: Java OLAP (JOLAP) y XML for Analysis application programming.
Sin embargo existen otras herramientas Open Source, orientadas a reportar la información almacenada multidimensionalmente en servidores, como el OpenI, interfaz Web para publicar reportes interactivos y gráficos, así como BIRT, perteneciente al grupo Eclipse.

BIRT es un sistema para la generación de reportes basado en Java/J2EE que tiene 2 componentes: un diseñador de reportes basado en Eclipse y un ambiente de ejecución que puede ser adicionado en un servidor de aplicaciones. Ofrece un motor de gráficos como: barras, pies, etc. que le adicionan el componente gráfico a los reportes.

Además cuenta con soporte interactivo para la visualización de los reportes, a través de JavaScript e incluso con el nuevo modelo propuesto por AJAX (Asynchronous JavaScript and XML).

Realmente en software libre se pueden encontrar todas las herramientas necesarias para la implementación de Data Warehouse, cuyos proyectos se orientan en la línea de Business Intelligence (BI). La única diferencia con los productos privativos es que sobre esta plataforma no se encuentra un producto que lo haga todo. Esa es la mayor ventaja, pues en las experiencias de implementación en el desarrollo de proyectos de este tipo, rara vez se requiere todo; en algunos casos solo se necesita una herramienta de reportes, en otro el servicio OLAP. En muchos casos no se necesita la herramienta de

ETL (Extracción, Transformación y Carga, por sus siglas en inglés), en otros por el contrario son tan complejas las fuentes de datos, que se requiere programar todo el proceso de ETL.

Como en todo proyecto de software, incluidos los proyectos de BI, el éxito depende de qué tan claros están definidos los requerimientos y no en qué tan poderosa es la herramienta.

En cuanto a las herramientas de consulta y reporte, existe una gran cantidad de estas en el mercado. Algunos proveedores ofrecen productos que permiten tener más control sobre qué procesamiento de consulta es hecho en el cliente y qué procesamiento en el servidor.

La herramienta de consulta genera entonces un llamado a una base de datos, extrae los datos pertinentes, efectúa cálculos adicionales, manipula los datos si es necesario y presenta los resultados en un formato claro.

Se puede almacenar las consultas y los pedidos de reporte para trabajos subsiguientes, como está o con modificaciones. El procesamiento estadístico se limita comúnmente a promedios, sumas, desviaciones estándar y otras funciones de análisis básicas.

Para hacer consultas más accesibles a usuarios no-técnicos, los productos tales como Crystal Reports de Seagate, Impromptu de Cognos, Reportsmith de Borland, Intelligent Query de IQ Software, Esperant de Software AG y GQL de Andyne, ofrecen interfases gráficas para seleccionar, arrastrar y pegar.

En general, los administradores de data warehouses que usen estos tipos de productos, deben estar dispuestos a ocupar su tiempo para resolver las tareas de estructuración, como administrar bibliotecas y directorios, instalar software de conectividad, establecer nombres similares en Inglés y precalcular "campos de datos virtuales".

Una vez que se han creado las pantallas SQL, puede necesitar desarrollar un conjunto de consultas y reportes estándares, aunque algunos productos ofrecen librerías de plantillas prediseñadas y reportes predefinidos que se pueden modificar rápidamente.

1.8 JUSTIFICACIÓN DE LAS HERRAMIENTAS UTILIZADAS:

La Oficina Nacional de Estadísticas, en calidad de ser el centro nacional de almacenamiento histórico de todos los datos estadísticos en Cuba, asume la necesidad del uso de un Data Warehouse para la

prestación de servicios de consultas y reportes de manera optimizada, eficiente y con la calidad requerida, para lo cual hace uso de la herramienta Microsoft SQL Server Analysis Services 2000 por varias razones:

- Alta compatibilidad con el gestor de base de datos utilizado en el proyecto Sistema Integrado de Gestión Estadística (SIGE), para el cual se diseña e implementa el Data Warehouse. SIGE es concebido en los inicios de su creación (agosto de 2006) para ser desarrollado sobre la base de software propietario, teniendo como gestor de Base de datos SQL Server 2000, con la consecuente aceptación del cliente y previa justificación de su utilización en el proyecto, y por el hecho de que el Analysis Services constituye un sub- módulo adjunto al paquete de instalación del SQL Server 2000 pues es perfectamente compatible y relacionable con el gestor donde se montara la base de datos del proyecto. La razón fundamental del uso de este gestor, además de su alto potencial, es que Cuba no tiene que pagar licencia para utilizar este producto, ya que el software que se va a desarrollar no se va a comercializar, sino que será un producto nacional, aunque ya se planea la migración en una próxima iteración hacia el Software de Código Abierto haciendo un fuerte estudio del estado del arte de las herramientas de Código Abierto, hoy usadas a nivel mundial.
- Posibilidad de almacenar y manipular cubos multidimensionales de información, unidad fundamental de almacenamiento de los Data Warehouse.
- Alta compatibilidad con las herramientas de Extracción, Transformación y Carga (ETL por sus siglas en inglés) del gestor SQL Server 2000 en los llamados Paquetes, donde se ofrecen servicios de importación, exportación, transporte, transformación de datos; tan importantes y necesarios para la implementación del Data Warehouse.
- Alta capacidad de almacenar un inmenso volumen de información, con una alta velocidad de respuesta en las consultas por muy complejas y/o extensas que sean.
- Alta capacidad de gestión de dimensiones, en sus distintas formas y estructuras de concepción.
- Soporte de disímiles funcionalidades matemáticas y estadísticas para mejorar y dinamizar el servicio de reportes y consultas, así como el soporte de miembros pre calculados y agregaciones, así como la posibilidad de crear particiones de los cubos.
- Soporte de minería de datos, mediante el uso de clúster y árboles de decisión, para llevar a cabo el análisis de información y patrones de descubrimiento, algo que no se concibe en este trabajo por su alta complejidad, pero que queda en un segundo plano para futuras iteraciones, en el transcurso del desarrollo de este Data Warehouse para la Estadística en Cuba.
- Eficiente gestión de la seguridad. El SQL Server Analysis Manager posee entre sus características el manejo eficiente de usuarios, permisos y roles, que tendrán acceso a distintas partes de los

cubos multidimensionales, en cuyo contenido también estará almacenado información harto sensible por lo que se requiere de una factible gestión de la seguridad.

- Rendimiento eficiente: Altísima velocidad de respuesta de las consultas Se tarda mucho menos en acceder a los datos del repositorio del *Data Warehouse* que en hacer una consulta a una base de datos relacional con un inmenso volumen de información. Además hacer consultas complicadas a las bases de datos de los sistemas operacionales puede empeorar el tiempo de respuesta de estos sistemas para otros usuarios.
- Posibilidad de integración de múltiples orígenes de datos: Nos brinda la capacidad de combinar los datos de distintas fuentes, lo cual suele ser una tarea bastante complicada para las personas encargadas de tomar decisiones con esa información. Normalmente hay que homogenizar los datos de una forma u otra. Por ejemplo, la información estadística histórica se encuentra almacenada en ficheros, con formatos específicos y fragmentados de distintas formas. Sin embargo, en este sistema, mediante la ayuda de los paquetes habilitados en Microsoft SQL Server 2000, los datos se homogenizan durante el proceso de extracción, transformación y carga.
- Soporte de agregados: Con esta herramienta se tiene la posibilidad de no solo almacenar los datos al más atómico de los detalles, sino que se pueden guardar los agregados necesarios (por ejemplo, la producción total de azúcar de la provincia Granma en el primer trimestre del año 2007, etc.).

1.9 CONCLUSIONES DEL CAPÍTULO:

Luego de este primer capítulo se puede concluir que en el mismo se han abordado de manera muy descriptiva y enfocada las principales características, definiciones y aspectos relacionados tanto con los Sistemas Informacionales como con los Sistemas Data Warehouse. Se ha realizado un estudio muy actual del estado del arte de los principales Data Warehouse en uso, tanto a nivel mundial como en nuestro país. Consecuentemente ha sido abordado el por qué de usar las herramientas que han sido utilizadas para el desarrollo de nuestro Sistema, así como los objetivos y características fundamentales de los Data Warehouse.

Finalmente se explican los aspectos principales, semejanzas y diferencias ente los Modelos Entidad Relación y los Modelos Multidimensionales por lo que se puede plantear que los distintos puntos y

temas descritos en este capítulo se encuentran ampliamente detallados y documentados así como claros y transparentes para permitir su fácil entendimiento.

CAPÍTULO 2: DESCRIPCIÓN DE LA SOLUCIÓN

Llevar a cabo un proceso de diseño e implementación de un Data Warehouse constituye una tarea bastante difícil y engorrosa puesto que en ella están implícitas un conjunto de elementos arquitectónicos, los cuales se deben conjugar satisfactoriamente para lograr un ensamblado adaptable a las necesidades de la empresa, así como también constituye un hito importante la efectiva definición de los procesos de negocio, dimensiones, medidas, cubos de datos, granularidad y modelos dimensionales, para entonces, para entonces efectuar eficientes procesos de extracción, transformación y carga de los datos, hasta la presentación al cliente de la información requerida. Es por eso que entran a jugar un conjunto de factores claves para el logro de tales metas, como un dominio amplio y extenso de lo que necesita saber el cliente, el cómo debe ser mostrada la información para que se puedan realizar a partir de esta los análisis precisos y las deducción correctas de los datos almacenados, en tanto que estos adquieren la connotación de información una vez que se muestra en forma de reporte, y dicha información se convierte en conocimiento, pues se aprovecha en el proceso de toma de decisiones gerenciales.

2.1 ARQUITECTURA E INTERACCIÓN DE LOS COMPONENTES DEL SISTEMA

Para comenzar, puede entenderse una arquitectura, en el ámbito computacional, como un conjunto de reglas o estructuras que proveen un esqueleto para el diseño general de un producto o sistema. En particular, una Arquitectura de Data Warehouse (Data Warehouse Architecture) es una forma de representar todas las estructuras de datos, comunicaciones, procesos, presentaciones y frentes usuarios que existen para los que hacen uso de un sistema computacional dentro de la empresa. Entre los objetivos fundamentales que debe satisfacer la arquitectura de un Data Warehouse se encuentra soportar diferentes configuraciones de los datos empresariales y al mismo tiempo, brindar facilidades al usuario para la ejecución y la administración de sus tareas en ambientes complejos. El procedimiento de reconciliación de los datos consiste en ejecutar reglas de concertación apropiadas que establecen relaciones de correspondencia entre los datos de tiempo real de diferentes sistemas operacionales y los datos que se generan, llamados datos reconciliados, los cuales mantienen la información detallada e histórica de manera consistente, integrada y veraz. El proceso de derivación de los datos aplica transformaciones adecuadas a los datos reconciliados o a los datos de tiempo real para generar los datos derivados, los cuales se refieren a períodos o puntos específicos en el tiempo y

son utilizados para administrar la empresa y servir de apoyo en los sistemas para la toma de decisiones. Proponer una arquitectura robusta y eficiente para un Sistema Data Warehouse constituye junto a la implementación otra de las tareas más difíciles e importantes en su proceso de desarrollo, puesto que en ella entran a interactuar un conjunto de elementos, componentes y factores, donde todos interrelacionados como un todo y cada uno con su funcionalidad y características específicas, le imprimen al Data Warehouse lo necesario para que los resultados esperados a mediano y largo plazo sean visibles y palpables, además con consecuencias positivas y beneficiosas. (Humphries, M, Hawkins, M. y Michelle, C. D. 1999).

Para comenzar a describir nuestra arquitectura, es necesario recordar algunos elementos referentes a la implementación de nuestro Sistema como son:

- Origen o Fuente de datos
- Base de datos Copia.
- Base de datos Operacional
- Base de datos Multidimensional
- Data Warehouse. Cubo de Datos
- Aplicación de Reporte

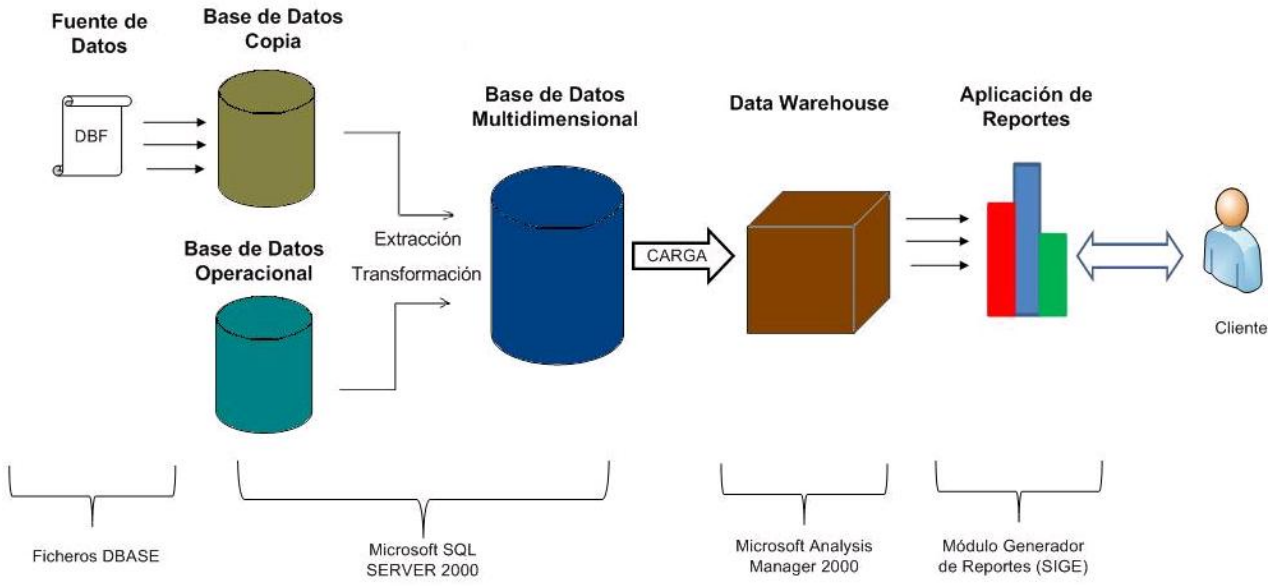


Figura 2.1: Arquitectura propuesta del Sistema Data Warehouse.

Origen de datos:

Son los ficheros o datos en bruto, que se encuentran almacenados en los bancos de datos o carpetas que guardan la información histórica de la ONE. Dichos ficheros se encuentran con extensión DBF, es decir que son de tipo DBase. Estos sufrirán un proceso de extracción hacia un Base de datos Copia para facilita en trabajo con la transformación y la homogenización de los tipos de datos, la información y los campos de las tablas en estos ficheros.

Base de datos Copia:

Constituirá una copia fiel de la información almacenada en los Orígenes de datos, solo que será una Base de datos montada en el SQL Server 2000. La misma mantiene sus tablas, campos, y tipos de datos, e información, respetando la estructura de los ficheros, para que pueda facilitar los procesos de Extracción, Transformación y Carga (ETL por sus siglas en inglés) hacia la Base de datos Multidimensional, que será descrita más adelante.

Base de datos Operacional:

Es aquella que interactúa constantemente con el SIGE, software desarrollado por nuestro Proyecto, que como se ha explicado se encarga de automatizar la actividad estadística en Cuba. Esta Base de datos intercambia permanentemente información a través de los distintos Módulos del software como son el MED, el MRC y el MGM, y contendrá la información concerniente a un año estadístico. La misma también está implicada en otro proceso de extracción y transformación mediante el uso de paquetes hacia la hacia la Base de Datos Multidimensional que contendría toda la información estadística de todos los años lo cual se realizará con una frecuencia de 12 horas, es decir 2 veces al día, y que a su vez, su información es cargada hacia el Data Warehouse.

Base de datos Multidimensional:

Los datos provenientes de las bases de datos Copia y Operacional abastecerán de información a la Multidimensional, la cual con un diseño orientado a dimensiones y medidas, contendría entonces la información limpia, integrada, válida, y lista para ser cargada hacia el Cubo de datos preparado en el Analysis Manager. Esta carga se realizará diariamente, en horarios nocturnos para no afectar el tiempo laboral, y será automático mediante los paquetes SQL, alertando siempre al Administrador del Data Warehouse mediante correo electrónico el día que no se realice la carga de los datos para que sea efectuado manualmente. De tal forma esta Base de datos es cargada la información proveniente de la

Operacional cada 12 horas, y la Base datos con Copia de la información histórica lo haría mediante los procesos de Extracción y Transformación solo una vez y año por año para evitar errores y duplicaciones en la información a reportar.

Data Warehouse. Cubo de datos:

Se refiere al cubo de datos definido en el Analysis Manager, el cual integra todas las dimensiones definidas en el diseño, tales como Aspecto, Indicador, Modelo, Temática, Tiempo, Ubicación. Dicho cubo se llamará CUBO_SIGE y contendrá la medida valor, como indicador para regir el hecho numérico en esta estructura. Algunas de sus características son que será procesado diariamente en horario nocturno, mediante configuraciones realizadas con ayuda de los paquetes de Microsoft SQL Server 2000. El tipo de almacenamiento definido para este cubo es el MOLAP. Como tal quedaría representado el Sistema Data Warehouse, con sus dimensiones, medidas, y cubos que integra. Solo nos queda la presentación de la información almacenada.

Aplicación de Reportes:

La Aplicación MGR desarrollada por el Módulo Generador de Reportes de nuestro Proyecto SIGE, contiene un ambiente amigable y entendible mediante la utilización de varios asistentes para no obligar a que el usuario final conozca el lenguaje SQL para estructurar las consultas, además de utilizar el mundialmente conocido asistente y Generador dinámico de consultas Active Report. El mismo se encarga de preparar la información proveniente del Data Warehouse, y presentársela adecuadamente al cliente en la forma de reportes. Dicha aplicación puede estar instalada en cualquier máquina, siempre que exista una conexión al Data Warehouse para poder acceder a los datos del mismo.

2.1.1 FLUJO DE DATOS:

El flujo de los datos desde sus orígenes hasta su presentación al cliente en forma de reportes tabulados, se realiza de manera unidireccional, es decir, en una sola dirección, sufriendo los cambios y las transformaciones necesarias. Esto se realiza después que los datos son copiados hacia una Base de datos inicial, montada y preparada en el gestor Microsoft SQL Server 2000, y una vez allí, efectuar mediante el uso de los paquetes de dicho gestor los procesos de extracción y transformación hacia otra Base de datos que contendrá un diseño multidimensional, ubicada en otro servidor, para entonces, una vez almacenada allí toda la información histórica de la ONE, efectuar la carga de dicha información hacia el cubo de datos preparado en el servidor destinado para el Data Warehouse. Una

vez que la toda información esté almacenada, preparada e integrada en el Data Warehouse, la aplicación de Reportes MGR accederá a la misma mediante las potencialidades que le brinda el Active Report en su compatibilidad con el OLAP. De igual forma la Base de Datos Operacional actualizará diariamente sus datos en la Base de datos Multidimensional utilizando los paquetes de SQL Server, y los datos fluirán de manera igual que el proceso descrito anteriormente.

2.1.2 NECESIDADES TECNOLÓGICAS:

La Arquitectura Data Warehouse a implantarse en la Oficina Nacional de Estadísticas tendrá como requisitos mínimos de hardware, los siguientes:

- 1 Servidor con 1 GB de memoria RAM, 160 GB de capacidad de disco duro, procesador a 3,1 GHz de velocidad.
- 1 Servidor con 2 GB de memoria RAM, 500 GB de capacidad de disco duro, con procesador a 3,1 GHz de velocidad.

Tanto la Base de datos operacional como la Copia serán administradas en un mismo servidor, en este caso el primero que aparece descrito, debido en primer lugar a que la copia de información desde los ficheros a la Base de datos Copia se realiza una sola vez, y automáticamente, al terminar de realizar los procesos de extracción y transformación hacia la Multidimensional, se procederá a realizar una salva en formato duro de esta Base de datos y a eliminarla de este servidor para liberar la carga de procesamiento del mismo y habilitar a la Base de datos operacional en sus funcionalidades diarias con el software que automatiza la actividad estadística.(SIGE).

En el caso del segundo servidor, este estará destinado a la base de datos multidimensional sobre el gestor Microsoft SQL Server 2000, con toda la información histórica, y al mismo tiempo, sobre el Analysis Manager 2000 tendrá el Data Warehouse, al cual se accede para realizar los pedidos de reportes desde la aplicación MGR, que se encontraría en un PC cliente.

2.2 PASOS PARA EL DISEÑO DE UN DATA WAREHOUSE

1-Seleccionar el proceso de negocio a modelar.

Un proceso de negocio es el mayor proceso operacional en la organización que es soportado por alguna clase de sistema (o sistemas) heredado desde el cual los datos pueden ser recolectados con el

propósito del almacén de datos. Ejemplos de procesos comerciales son los pedidos, inventarios, facturas, embarques, contabilidad, ventas, etc.

2- Elegir el grano del proceso de negocio

El grano es el nivel atómico fundamental del dato a ser representado en la tabla de hechos para este proceso. Granos típicos son las transacciones individuales, las instantáneas diarias individuales o las instantáneas mensuales individuales. Es imposible proceder al paso 3 sin definir el grano.

3- Elegir las dimensiones que se aplicaran a cada registro de la tabla de hechos.

Dimensiones típicas son el tiempo, el producto, el cliente, la promoción, el depósito, el tipo de transacción y el estado. Con la elección de cada dimensión, se describen todos los atributos dimensionales de tipo texto discreto, que llenarán cada tabla dimensional.

4- Elegir los hechos mensurables que existirán en cada registro de la tabla de hechos.

Cuando se habla de hechos mesurables se refiere a hechos contables, es decir de cuantitativa existencia. Los típicos hechos mensurables son cantidades numéricas aditivas similares a Cantidades Vendidas y Ganancia en Dinero.

2.3 MODOS DE ALMACENAMIENTO DE DATOS

De la forma en que se almacenen los datos en un Data Warehouse, dependerán los requerimientos de almacenamiento, las velocidades necesarias, los lugares de almacenamiento de las agregaciones y las características del almacenamiento en sí. Recordar que las agregaciones son resúmenes pre calculados de datos que hacen más rápidas las consultas al cubo.

Existen 3 modos o formas de almacenar los datos, estas son:

- Multidimensional OLAP (MOLAP)
- Relational OLAP (ROLAP)
- Hybrid OLAP (HOLAP)

El modo de almacenamiento MOLAP guarda la información de las agregaciones y una copia de la fuente de datos en una estructura multidimensional en la máquina donde se encuentra instalada la

herramienta Data Warehousing. Debido a esto, el acceso tanto a los datos en detalle como a los resúmenes, es mucho más rápido, por lo que este tipo de almacenamiento es apropiado para aquellos almacenes de datos donde sea fundamental la velocidad de respuesta sin importar el espacio físico de almacenamiento.

El modo ROLAP almacena las agregaciones en la base de datos relacional de la fuente de datos especificada. El mismo, a diferencia de MOLAP, no guarda una copia de la fuente de datos, y de los resúmenes de estos, y entre otras cosas las consultas las consultas son lentas. Lo típico de ROLAP es que se usa generalmente para inmensos volúmenes de datos con poca frecuencia de consultas, ejemplo la información histórica de muchos años de antigüedad.

En cuanto al modo HOLAP, como su nombre lo indica, es un híbrido, es decir, que adquiere características de los otros restantes. Como el MOLAP, es capaz de almacenar las agregaciones en una estructura multidimensional. En cambio no lo hace para almacenar copias de las informaciones provenientes de la fuente de datos. Para consultas que acceden solamente a resúmenes de dato, sobre las propias bases de datos relacionales, estas serán más rápidas que en el modo ROLAP, pero al acceder a datos en detalle, no lo hace en el tiempo requerido, por lo que en ambos casos suele ser más lento que el MOLAP.

Después de hacer un análisis detallado de estos tipos de almacenamiento de datos en estructuras multidimensionales, se considera como el más indicado el MOLAP (Multidimensional Online Analytical Processing) porque:

- Se requiere una alta velocidad de respuestas en las consultas.
- Se requiere almacenar los agregados o pre calculados en una estructura multidimensional para facilitar su rápido acceso.
- El espacio en disco de la información almacenada no afectará el desempeño del sistema, es decir, no es problema la capacidad de almacenamiento de datos, puesto que la Oficina Nacional de Estadísticas cuenta con los recursos necesarios.
- El modo HOLAP, aunque plantea una mayor velocidad de respuesta que el ROLAP, no satisface el tiempo de consulta requerido por el negocio y en ocasiones genera conflictos y errores en los reportes de datos.
- Se accede frecuentemente a datos en detalle y con necesidades de respuestas rápidas.

2.4 DISEÑO DEL SISTEMA DATA WAREHOUSE

A partir de lo planteado por Kimball en su obra The Data Warehouse Toolkit sobre los pasos que deben seguir para diseñar un Data Warehouse y además para darle cumplimiento a las tareas de la investigación relacionadas con esta etapa de construcción, se describen cuatro pasos adaptados a nuestras propias necesidades, que definen como queda estructurado el diseño de este Sistema.

2.4.1 DESCRIPCIÓN DEL NEGOCIO A MODELAR.

La Oficina Nacional de Estadísticas, en calidad de ser la rectora de toda la actividad estadística en nuestro país, constituye la entidad en la cual se aglutinan y confluyen un conjunto de procesos que contribuyen a que la gestión de la información estadística en Cuba se desarrolle favorablemente. En tal sentido en esta actividad intervienen los procesos con los Modelos de Estadísticas Continuas, los cuales a diferencia de los Modelos de Estadísticas Periódicas se llenan regularmente en períodos fijos de tiempo (ejemplo, existen modelos mensuales, trimestrales, anuales, etc.), no son de tipo encuesta, y se encargan de captar la información con respecto a los indicadores relacionados con los centros informantes (término para nombrar las empresas) en dependencia de si sean entidades económicas, instituciones sociales u organismos estatales y cuya captación se realiza en todos los niveles, es decir, en los municipios siendo controlados por la Oficina Municipal de Estadísticas correspondiente, en las provincias, siendo estas controladas por la Oficina Territorial de Estadísticas, y a nivel nacional por la entidad que controla la nación en este sentido.

De tal forma que el dato estadístico es captado o recogido en estos Modelos, en cada centro informante de cada municipio, en una fecha determinada y donde cada uno de ellos representará un indicador ya sea económico o no.

Proceso de negocio

Precisamente y después de haber descrito brevemente el negocio de la Oficina Nacional de Estadísticas, sale a relucir que su proceso de negocio constituye la captación de la información estadística, en todos los niveles y en todos los sectores, registrando los datos desde el mayor nivel de detalle hasta los resúmenes más densos, para permitir después que esta información pueda ser consultada y reportada sin contratiempos y con la validez requerida.

2.4.2 GRANO DEL PROCESO DE NEGOCIO A MODELAR

Una vez que ya se han definido el proceso de negocio a modelar, se debe seleccionar cuál será el máximo nivel de detalles de nuestro diseño, es decir el grano del proceso de negocio. Recordar que el grano es importante porque determina el grado dimensional de nuestro diseño, y por supuesto tiene también un profundo impacto en el tamaño del sistema. Por tal motivo y en dependencia de las necesidades de nuestro negocio, el grano queda definido como la captación diaria de datos estadísticos, en todos los municipios, de cualquier indicador estadístico y por cualquier aspecto y variante de algún Modelo estadístico (Documento encargado de recoger los datos estadísticos en los centros informantes)

2.4.3 DIMENSIONES

Una vez que se ha seleccionado el grano del proceso de negocio, de inmediato se deja entrever cuales posibles dimensiones podrán existir para este diseño. De momento sale a relucir una dimensión tiempo, muy necesaria para controlar la captación de los datos estadísticos en la línea temporal, otra dimensión lugar, para el control de la localización donde se realiza el proceso de negocio, una dimensión modelo que identifica en cuál Modelo de los tantos que tiene la Gestión Estadística es que se ha captado el dato, y una dimensión indicador, que define a qué indicador corresponde la información captada. Dependiendo de las necesidades y las características del negocio en general, se hace necesaria la definición de otras cuatro dimensiones adicionales para facilitar el buen desempeño de este sistema, por lo que salen a relucir la dimensión Temática y Aspecto, las cuales se explicarán detalladamente a continuación para su mejor entendimiento.

Dimensión Tiempo

La dimensión tiempo es una dimensión virtualmente garantizada a estar presente en cada oficina de estadística, porque virtualmente cada oficina es una serie de tiempo. El tiempo es usualmente la primera dimensión en el orden subyacente de organización en la base de datos, debido a que cuando está primero en el orden de organización, las cargas sucesivas de intervalos de tiempos de datos, cargarán los datos en un territorio virgen en el disco. En tal sentido se define una dimensión tiempo con varias categorías o niveles para su mejor organización, estos son los niveles Año, Semestre, Trimestre, Mes y Día, pues precisamente las oficinas se rigen por una captación en series periódicas que pueden ser mensuales, trimestrales, semestrales, anuales, bianuales, etc., y los cuales estarían representados en una correspondiente tabla dimensional llamada Tiempo, que se encargaría de registrar el dato estadístico, en coincidencia por supuesto con las demás dimensiones ocurrentes.

Figura 2.2: Tabla dimensional Tiempo.

Dimensión Ubicación

La dimensión Ubicación describe la organización en el espacio que tendrá la captación de los datos estadísticos en cada una de las Oficinas. Se considera entonces como una dimensión geográfica para localizar espacialmente desde el más detallado nivel que es el centro informante hasta el más resumido, en este caso la provincia, los proceso de captación. Por tal motivo existen varias categorías o niveles en esta dimensión, como son Provincia, Municipio, y Centro Informante.

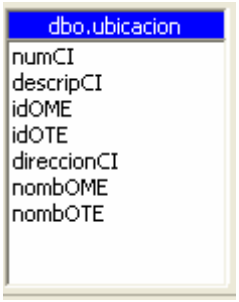


Figura 2.3: Tabla dimensional Ubicación

Dimensión Indicador

En el proceso de gestión estadística se manejan un conjunto de indicadores de distintos sectores como la economía, la educación, la producción, etc., y ejemplos de indicadores son el maíz, el azúcar, ventas netas de níquel, ingresos netos en divisas del turismo, etc. Cada indicador especifica una fila del modelo donde se captan los datos, por lo que la intersección de un indicador, que en este caso sería una fila, con una columna, llamada Aspecto en las terminologías estadísticas, es lo que define una celda o escaque del modelo. También se puede dar el caso de que un indicador pueda incluir varios sub indicadores como es el caso del indicador leche, y sus sub indicadores lecha condensada y

lecha en polvo, por lo que se modela aquí una dimensión con herencia, donde se tiene un miembro padre del cual dependen varios hijos, y cada hijo pueden contener otros hijos, y así sucesivamente. Este tipo de dimensión también es llamada Parent- Child (Padre- Hijo).

Figura 2.4: Tabla dimensional Indicador

Dimensión Aspecto

El aspecto en términos estadísticos se refiere a las columnas en los modelos, y la intersección de estos con los indicadores, que serían las filas, es lo que determina una celda que guarda el dato que ha sido captado en un centro informante determinado, en una fecha determinada. Ejemplo de aspectos se tienen Plan, Real, Año Anterior, entre otros, y al igual que los indicadores, cada aspecto puede contener sub aspectos dentro de él, por lo que se percibe una dimensión con herencia o de tipo Parent- Child.

Figura 2.5: Tabla dimensional Aspecto

Dimensión Variante

La variante en los modelos estadísticos es otra forma de clasificación de la información, pero no clasificando las celdas sino el modelo como tal, en el cual se incluyen variantes por ejemplo como

Actividad Hotelera, Actividad de apoyo al turismo, Perteneciente a empresa mixta, perteneciente a campismo, entre otras, que realmente lo que determinan es ya sea la actividad, el sector, o el organismo al que pertenece el centro informante cuando llena un modelo determinado. Aunque parezca complicado, realmente no lo es, simplemente se trata de una clasificación de los modelos en dependencia del centro informante que lo llene en un momento determinado.

Figura 2.6: Tabla dimensional Variante.

Dimensión Temática

Dentro de la dimensión indicador se podrán registrar todos los indicadores que estarán almacenados pero debido a que se trata de una relación de tipo Parent- Child (Padre - Hijo), solo estarán los indicadores y los subgrupos de estos que tiene asociado. Sin embargo existe otra clasificación para estos. Es la temática, la cual agrupa en distintas ramas los distintos indicadores que existen, de tal manera que puedan estar clasificados de una forma más genérica y su organización y estructura jerárquica sea más desglosada. De tal manera que lo que describe esta dimensión sería la descripción de la temática, como elemento que agrupa los indicadores y cuyo campo en el Data Warehouse se identifica como descripTemática, y como elemento subyacente e interno de este nivel, sería la descripción del indicador, descrito en este caso como descripInd. De tal forma se tiene otra clasificación para los indicadores, elemento tan importante en nuestro sistema constituyendo además uno de centros fundamentales del negocio del Data Warehouse.

Figura 2.7: Tabla dimensional Indicador

Dimensión Modelo

Ya por último se tiene la dimensión modelo, la cual se encarga de registrar en la Tabla de Hechos que se verá a continuación, el número del modelo en el cual se capta la información estadística en cada uno de los centros informantes, en una fecha determinada. Este es quizá el más simple de las dimensiones, pero no por eso deja de ser importante, pues se convierte en imprescindible en el momento de realizar un reporte estadístico donde se involucren datos del modelo como su número, su sub número, su descripción, etc. En el Sistema Estadístico en Cuba se encuentran modelos como el 0005, el 0333, 0467, el 0760, entre otros.

Figura 2.8: Tabla dimensional Modelo.

2.4.4 LA TABLA DE HECHOS.

La tabla de hechos es donde las mediciones numéricas del negocio son almacenadas. Cada una de las mediciones es tomada como la intersección de todas las dimensiones. Esta tabla tiene que ver mucho con la granularidad de nuestro modelo multidimensional, es decir, el significado conceptual de que en una fecha determinada, en un centro informante determinado, en un modelo determinado, con una variante determinada, en un indicador (fila del modelo) determinado, y en un aspecto (columna del modelo) determinado, se encuentra almacenado el dato estadístico, porque es precisamente este dato el que responde a todas estas coincidencias de los atributos identificadores de cada una de las

dimensiones. Se puede ver a la Tabla de Hechos como una tabla formada por cada uno de estos atributos de cada dimensión, (por ejemplo el código de Tiempo, el código de la ubicación, el código del modelo, etc.) y el valor estadístico como tal, que en este caso es un valor medible o mesurable, y que corresponde al hecho de ser captado coincidiendo con todas estas características. Las llaves provenientes de las tablas dimensionales serían llaves también en esta tabla, siendo también además atributos indexados, ofreciéndole al sistema un acceso a la información almacenada en la tabla de hechos, mucho más rápida y eficiente.

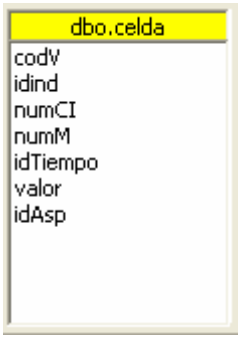


Figura 2.10: Tabla de hechos celda.

2.5 CUBO DE DATOS

Un cubo es una estructura multidimensional que contiene dimensiones y medidas. Las dimensiones definen su estructura mientras que las medidas le proporcionan los valores numéricos de interés para el usuario final. Cada celda del cubo es definida por la intersección de las dimensiones, y el valor numérico almacenado constituye un agregado que se almacena en cada celda. En la siguiente figura se muestra una representación espacial de cómo se vería un cubo de sólo tres dimensiones, en este caso, la dimensión Tiempo (1998, 1999, 2000), geografía (Matanzas, Holguín, Cienfuegos) y la dimensión Producto (Producto 1, Producto 2), en el cual la celda de color rojo representa la coincidencia de estas dimensiones, almacenando la cantidad de Productos 2 en el año 2000 en Holguín.

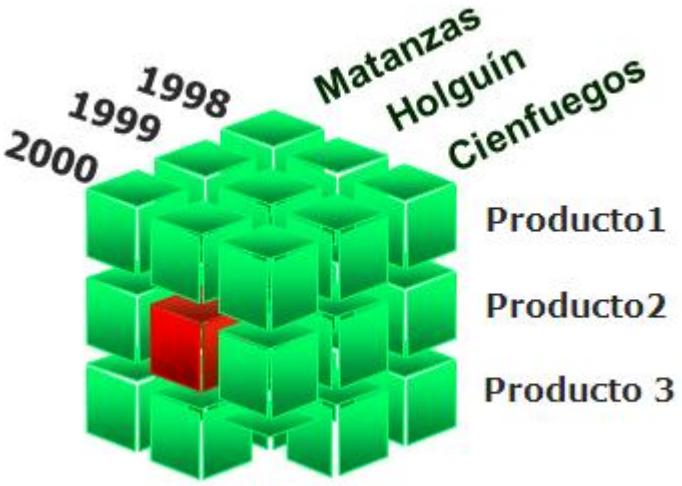


Figura 2.11: Representación espacial de un cubo de tres dimensiones.

En nuestro sistema se ha definido hasta el momento un solo cubo que englobará todas las dimensiones descritas en el diseño, tales como Aspecto, Indicador, Modelo, Temática, Tiempo, Ubicación. Dicho cubo se llamará CUBOSIGE y contendrá la medida valor, como indicador para regir el hecho numérico en esta estructura.

Algunas de sus características son que será procesado diariamente en horario nocturno, mediante configuraciones realizadas con ayuda de los paquetes de Microsoft SQL Server 2000. El tipo de almacenamiento definido para este cubo es el MOLAP.

En otras iteraciones y a medida que el sistema se haga más grande y complejo, donde se incluirán más reglas de negocios de la Oficina Nacional de Estadísticas, será necesario crear otros cubos que utilicen estas y otras nuevas dimensiones que deberán crearse, pero por el momento, como se está en una primera versión del Sistema Data Warehouse, donde no se han integrado aún todos los procesos de negocios de la Entidad, se mantiene un solo cubo.

2.6 LA GRANULARIDAD

La granularidad en la tabla de hechos se determina después de identificadas las columnas que existirán en dicha tabla. La granularidad es una medida del nivel de detalle enfocada a cada ocurrencia que exista en la tabla de hechos. Ejemplos de granos se pueden mencionar "Las ventas a cada cliente, por producto, por mes." Por lo que es evidente percibir la gran relación que existe entre la granularidad y las dimensiones asociadas.

Se recomienda siempre no mezclar varias granularidades en una misma tabla de hechos, ni almacenar en dicha tabla, sumas o resúmenes, pues van en contra de la filosofía de llegar al detalle, más bien en estos casos se deben almacenar dichos resúmenes o agregados en tablas separadas con sus respectivos niveles de granularidad.

Independientemente de la importancia de mantener un mínimo de granularidad en el Sistema Data Warehouse, también es muy importante contener la información en bajos niveles de detalles, es decir, con alta granularidad puesto que podría ser muy beneficioso para aquellas empresas que no requieren altos nivele de detalles, para su futuro en el análisis de sus negocios.

Después de este análisis se puede decir que nuestra granularidad está concentrada en registrar el dato estadístico en un centro informante (empresa), en un día, con un indicador asociado, con un aspecto, con una variante, en un modelo determinado. De esta manera se puede determinar el alcance dimensional de nuestro negocio.

2.7 MODELO MULTIDIMENSIONAL

Una vez definido en nuestro negocio las dimensiones, las medidas, el o los cubos de datos y la granularidad, se procede a la estructuración del modelo o los modelos multidimensionales que existirán. En tal sentido se puede destacar que por las necesidades actuales de nuestro negocio solo existe un modelo que unifica las dimensiones definidas y la medida que se ha especificado hasta el momento.

El modelo dimensional divide el mundo de los datos en dos grandes tipos: las medidas y las descripciones del entorno de estas medidas. Las medidas, que generalmente son numéricas, se almacenan en las tablas de hechos y las descripciones de los entornos que son textuales se almacenan en las tablas de dimensiones. Las tablas de hechos son las tablas primarias en el modelo dimensional y contiene los valores del negocio. Los hechos más comunes son valores numéricos. Cada tabla representa una interrelación muchos – muchos y contiene dos o más llaves extranjeras que acoplan con sus respectivas tablas de dimensiones. (Kimball, R. y Margy, R. 2002).

El tipo de modelo ha utilizar en este caso es el llamado modelo tipo estrella debido con la tabla de hechos en el centro y las tablas dimensionales relacionadas con la central en un modelo radial alrededor de la tabla central.

En este caso cada dimensión se define por su llave primaria que sirve para mantener la integridad referencial en la tabla de hechos a la que se acopla. Los atributos de estas tablas sirven de base a las solicitudes que se hacen al DWH.

Las tablas de dimensiones contienen información jerárquica que permitirán la realización de las agregaciones o las profundizaciones.

La tabla del centro es la única tabla en el esquema con múltiples enlaces conectándola a otras tablas. Las otras tablas tienen un enlace simple que las enlaza con la tabla central.

En la siguiente figura se muestra el modelo multidimensional definido en nuestro Sistema:

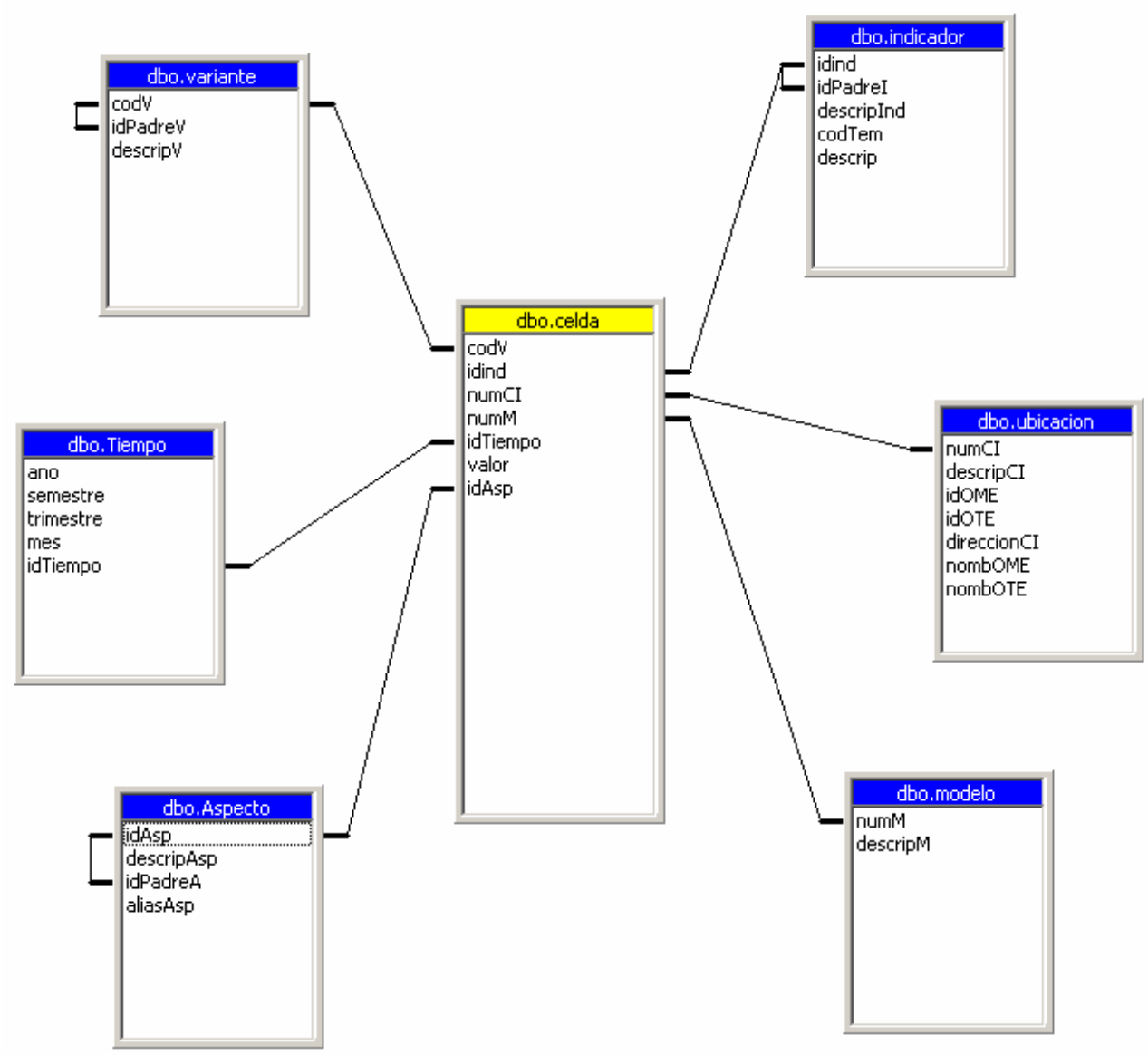


Figura 2.12. Modelo multidimensional con esquema de estrella.

2.8 IMPLEMENTACIÓN DEL SISTEMA DATA WAREHOUSE

La tarea de implementar el data Warehouse una vez definido su diseño multidimensional es quizá la más difícil dentro de la construcción del sistema en su conjunto, por lo que al utilizar como guía los pasos propuestos anteriormente para la implementación y mediante la adaptación como es lógico, a las características del negocio del Sistema de Estadísticas en Cuba, se comienza con esta compleja etapa.

Varios autores han propuesto ingeniosas metodologías para la construcción de Data Warehouse, entre ellos se han destacado los prestigiosos investigadores y empresarios informáticos, Mark Humphries, Michael W. Hawkins y Michelle C. Dy, quienes en su libro Data Warehousing Arquitecture e Implementation abordan un conjunto de detallados pasos relacionados con la implementación. Guiándonos por ese proceso de construcción, a continuación se definen un conjunto de pasos que logran satisfactoriamente la implementación de nuestro Sistema en una primera iteración.

2.8.1 PREPARACIÓN DEL AMBIENTE DE DESARROLLO:

Este primer paso está dedicado a la adquisición y montaje de todo el ambiente de desarrollo necesario para la implementación. Es aquí donde se instala el hardware, el sistema operativo con el que se trabajará, todas las herramientas relacionadas con el Data Warehouse, las conexiones de redes necesarias así como la creación de los usuarios de acceso requeridos.

Para empezar, se deberá configurar todo lo relacionado con el montaje adecuado del software y el hardware y todas las conexiones y habilitaciones necesarias para el buen funcionamiento del Data Warehouse por lo que esta etapa tiene gran relación con la arquitectura propuesta para el Sistema de Gestión Estadística. Como se ha planteado para la justificación de las herramientas utilizadas para el desarrollo del Data Warehouse, estas son el SQL Server 2000, y el Analysis Services 2000, aprovechando sus potencialidades y sus comodidades, principalmente en los procesos de almacenamiento de la información y en los de extracción, transformación y carga de los datos.

Se dispondría entonces de un servidor de Base de datos con el gestor SQL Server 2000, al cual llegan todas las informaciones provenientes de los distintos orígenes de datos que utiliza la Estadística, como son ficheros DBF, y que por supuesto estos orígenes de datos estarían en otra máquina para no sobrecargar el servidor de Base de Datos, en el cual los datos extraídos sufrirán procesos de

transformación, limpieza y validación de los datos, antes de que sean utilizados para ser consultados o reportados.

2.8.2 OBTENCIÓN DE LAS COPIAS DE LOS DATOS OPERACIONALES.

Almacenar copias de las tablas de la base de datos operacional (fuente de datos original) en el servidor de Data Warehouse, o quizá en algún dispositivo de almacenamiento, por si en algún momento no se puede acceder a la base de datos operacional, y además para realizar las primeras pruebas de extracción y transformación de datos con copias de la información original.

La Oficina Nacional de Estadísticas, entidad para la cual precisamente se construye este sistema, presenta su información histórica en ficheros de extensión DBF, organizados por meses, y agrupados en carpetas por años, es decir, que existen por ejemplo las carpetas 2000, 2001 y 2002, y cada carpeta contiene los ficheros enero.dbf, febrero.dbf, marzo.dbf, etc., los cuales contienen las información estadística de cada modelo, en cada empresa o centro informante, de cada indicadores, y así sucesivamente, hasta completar el resto de los atributos de los ficheros. Para realizar las copias desde estos ficheros hacia el gestor de base de datos SQL Server 2000, que es donde los datos sufrirán los procesos de extracción, transformación y carga, se han utilizado los paquete que provee este gestor, con los cuales se ha configurado para que se replique la información que existe en las carpetas a una base de datos preparada para este fin, que tendrá doce tablas con los doce meses del año, y en donde se guardará en cada tabla, la información correspondiente a cada mes que existe en las carpetas.

La figura siguiente muestra una parte del paquete configurado para este fin:

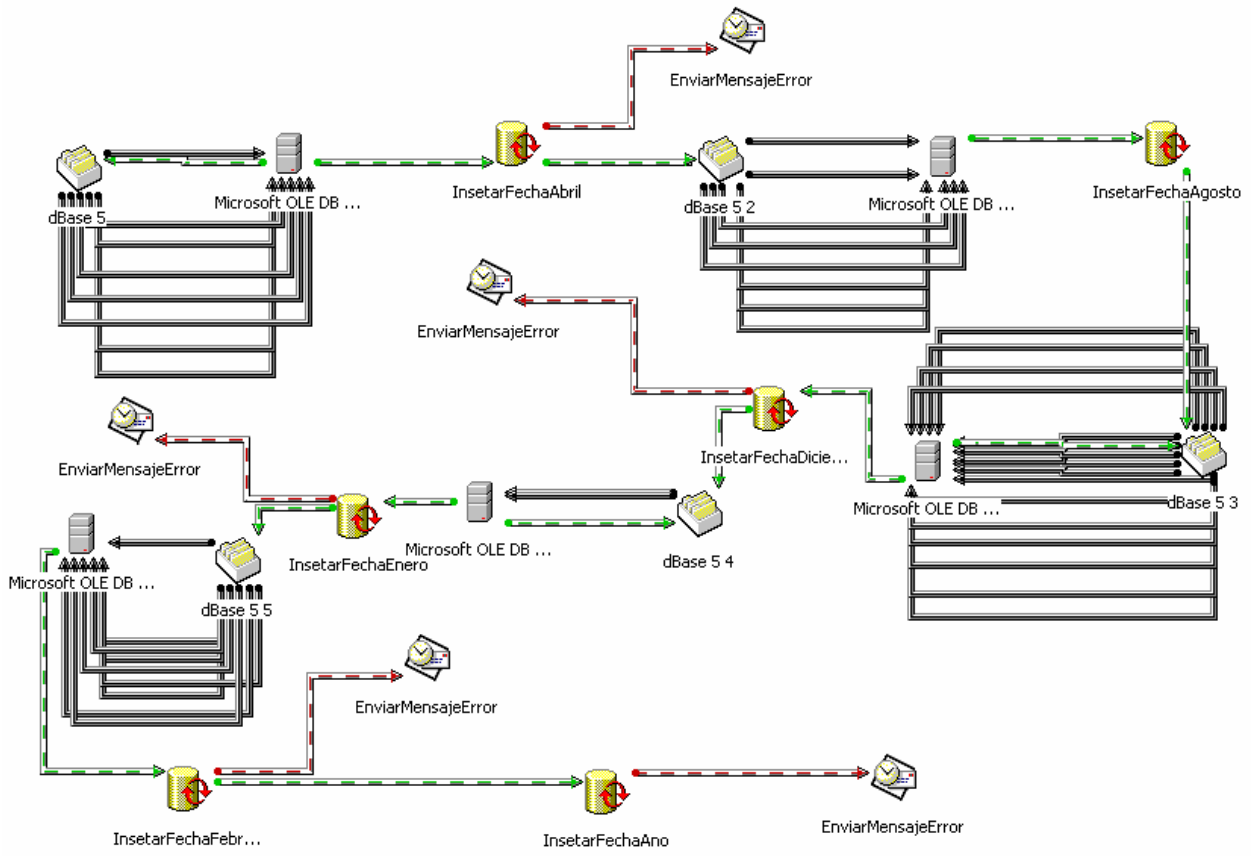


Figura: 2:13: Proceso de extracción, transformación y carga en paquete de SQL Server 2000.

2.8.3 COMPLETAMIENTO DEL ESQUEMA DE DISEÑO FÍSICO

Completar con el diseño físico multidimensional es un paso fundamental donde se debe tener en cuenta tres puntos, el esquema de diseño, donde se debe concluir con el diseño físico de las tablas de hechos y dimensionales y sus respectivos atributos. Como segundo punto, los índices, donde se debe identificar el método de indexación apropiado para usar en las tablas y sus atributos basado en el volumen de datos esperado y en las consultas que se puedan realizar. Como tercer punto, la fragmentación, donde el administrador o encargado del Data Warehouse puede crear particiones de las tablas de hechos y dimensionales, en dependencia del tamaño y complejidad de las mismas, teniendo en cuenta siempre las potencialidades del software donde esta montado el almacén de datos.

En este paso corresponde completar el diseño, para que la información pueda ser validada, limpia, debidamente transformada y lista para ser cargada a la herramienta Analysis Services 2000. Se debe empezar diciendo que en nuestro gestor ya se tiene definida y montada una base de datos con un

diseño perfectamente adaptable a nuestra problemática, la cual actuaría como almacén de datos históricos, pero no sería la propia base de datos en sí, sino que tendría las tablas necesarias para adaptar nuestro diseño físico con nuestras propias necesidades, en lo que a dimensiones y medidas se refiere, es decir que se tendrían las tablas Tiempo, Indicador, Aspecto, Variante, Modelo correspondientes a las tablas dimensionales, y la tabla celda que sería la de hechos, la cual contendría todos los atributos identificadores de cada tabla dimensional, y además los atributos medibles, que en nuestro caso es uno solo, es decir el atributo valor.

Se puede concluir que nuestro esquema de diseño quedaría de la siguiente manera:

Tablas dimensionales:

Tiempo (idTiempo, Año, Semestre, Trimestre, Mes)

Indicador (idInd, descripInd, idPadreI, codTem, descrip)

Aspecto (idAsp, idPadreAsp , descripAsp, aliasAsp)

Variante (codV, descripV, idPadreV)

Modelo (idModelo, descripM)

Ubicación (numCI, descripCI, idOME, idOTE, direccionCI, nombOME, nombOTE)

Tabla de hecho:

Celda (numM, codInfAcum, idAsp, codV, idind, numCI, valor)

2.8.4 CONFIGURACIÓN DE LA EXTRACCIÓN Y TRANSFORMACIÓN DE DATOS.

Debido a que los datos deberán ser extraídos, transformados, depurados y cargados del sistema operacional al Data Warehouse, es imprescindible conocer como se realizarán cada una de estas actividades. En este sentido, controlar desde donde y en qué tiempo se realizará la extracción de los datos, que herramienta se utilizará así como dominar sus potencialidades, de qué tablas y cuáles datos se extraerán. En cuanto a la transformación de los datos, esta se hará de acuerdo a las reglas que se hayan definido en el negocio. Transformaciones tales como cambios de formato, la cual asegura la unicidad y estandarización de los tipos de datos, por ejemplo la fecha, deberá ser tratada como un solo tipo de dato y siempre de la misma forma.

La duplicación de datos es otro tipo de transformación, donde se debe evitar que la información proveniente de una o varias fuentes de datos se multiplique a la hora de su unión o juntura. Otros tipos de transformaciones son los agregados y los campos integrados.

Para los procesos de extracción y transformación se utilizarán las ventajas del Gestor SQL Server 2000, en el cual se definen un conjunto de paquetes con distintas características que posibilitarán este proceso. En este caso ya se encontraría en nuestro Gestor las tablas de cada uno de los meses, pobladas de datos provenientes de los ficheros de extensión dbf que se encontraban organizadas en carpetas. Estas tablas, cuyos meses contendrían la información de todos los años tendrían los siguientes atributos:

Enero (mod, emp, var, ind, c01, c02, c03, c04, c05,…, c20)

Febrero (mod, emp, var, ind, c01, c02, c03, c04, c05,…, c20)

Marzo (mod, emp, var, ind, c01, c02, c03, c04, c05,…, c20)

Abril (mod, emp, var, ind, c01, c02, c03, c04, c05,…, c20)

Mayo (mod, emp, var, ind, c01, c02, c03, c04, c05,…, c20)

Junio (mod, emp, var, ind, c01, c02, c03, c04, c05,…, c20)

Julio (mod, emp, var, ind, c01, c02, c03, c04, c05,…, c20)

Agosto (mod, emp, var, ind, c01, c02, c03, c04, c05,…, c20)

Septiembre (mod, emp, var, ind, c01, c02, c03, c04, c05,…, c20)

Octubre (mod, emp, var, ind, c01, c02, c03, c04, c05,…, c20)

Noviembre (mod, emp, var, ind, c01, c02, c03, c04, c05,…, c20)

Diciembre (mod, emp, var, ind, c01, c02, c03, c04, c05,…, c20)

Donde el atributo mod se refiere al número del modelo, emp es el código de la empresa o centro informante, var es el código de la variante asociada, ind es el código del indicador y los atributos desde c1 hasta c20 contendrán el dato estadístico que concuerda con los atributos anteriores. Por supuesto que habrá modelos donde en cada fila de estas tablas tendrán datos nulos o que no existan en los atributos columnas (c01, c02, c03,…, c20).

Teniendo ya estas tablas en el gestor, se realiza la carga de los datos hacia el modelo multidimensional preparado previamente. Este modelo multidimensional, físicamente concluido en pasos anteriores constituirá una base de datos separada, donde los datos estarán poblados sin inconsistencia, con limpieza, sin duplicaciones ni anulaciones. Esta base de datos que contiene el modelo multidimensional, presenta entre sus tablas una llamada celda, la cual contiene atributos con significados similares a las tablas mensuales, estos son:

Celda (numM, idTiempo, idAsp, codV, idind, numCI, valor)

En este caso se realiza la extracción, transformación y carga de los datos almacenados en las tablas mensuales hacia esta tabla, mediante el uso eficiente de los paquetes de SQL Server 2000. La correspondencia de estos atributos queda como se muestra en la siguiente tabla:

Atributos de las tablas Mensuales	**Atributos de la tabla celda**
mod	numM
var	codV
ind	idind
emp	numCI
FECHA	idTiempo
COLUMNA	idAsp
c01, c02, c03, c04, c05,…, c20	valor

Tabla 1: Correspondencia de los atributos de las tablas mensuales con los atributos de la tabla celda.

En el caso de la extracción, transformación y carga de la FECHA, se realizaría en dependencia del mes y el año; igualmente ocurre con el atributo COLUMNA; y en el caso del atributo valor, se realizará primero comenzando siempre que la información no sea nula, por el atributo c01 de la tabla mensual, y así sucesivamente hasta el atributo c20. El siguiente fragmento de código en el lenguaje SQL describe el proceso descrito anteriormente

```
--- LA COLUMNA 1

INSERT INTO datawarehouse.dbo.celda( variante, indicador,ci, modelo, tiempo, valor, aspecto)  select distinct [2000].dbo.enero.[v01]+[2000].dbo.enero.[v02],  left ([2000].dbo.enero.mod,4)+right [2000].dbo.enero.fil,5) ,  [2000].dbo.enero.emp, [2000].dbo.enero.mod, [2000].dbo.enero.esf+ [2000].dbo.enero.[sin],[c01],  'c01'  from [2000].dbo.enero where  [2000].dbo.enero.[v01] is not null and [2000].dbo.enero.[V02] is not null and  [2000].dbo.enero.emp is not null and  [2000].dbo.enero.mod is not null and  [2000].dbo.enero.[sin] is not null and  [2000].dbo.enero.fil is not null and [2000].dbo.enero.[c01] is not null

--- LA COLUMNA 2

INSERT INTO datawarehouse.dbo.celda  ( variante, indicador,ci, modelo, tiempo, valor, aspecto) select distinct [2000].dbo.enero.[v01]+[2000].dbo.enero.[v02],  left ([2000].dbo.enero.mod,4)+right ([2000].dbo.enero.fil,5) ,  [2000].dbo.enero.emp, [2000].dbo.enero.mod,  [2000].dbo.enero.esf+ [2000].dbo.enero.[sin],[c02],  'c02'  from [2000].dbo.enero where  [2000].dbo.enero.[v01] is not null and [2000].dbo.enero.[V02] is not null and  [2000].dbo.enero.emp is not null and  [2000].dbo.enero.mod is not null and  [2000].dbo.enero.[sin] is not null and  [2000].dbo.enero.fil is not null and [2000].dbo.enero.[c02] is not null

--- Y así sucesivamente se realiza la extracción de los datos, desde la columna 1 hasta la columna 16.

--- LA COLUMNA 16

INSERT INTO datawarehouse.dbo.celda ( variante, indicador,ci, modelo, tiempo, valor, aspecto) select distinct [2000].dbo.enero.[v01]+[2000].dbo.enero.[v02],  left ([2000].dbo.enero.mod,4)+right ([2000].dbo.enero.fil,5) ,  [2000].dbo.enero.emp, [2000].dbo.enero.mod,  [2000].dbo.enero.esf+ [2000].dbo.enero.[sin],[c16],  'c16'  from [2000].dbo.enero where  [2000].dbo.enero.[v01] is not null and [2000].dbo.enero.[V02] is not null and  [2000].dbo.enero.emp is not null and  [2000].dbo.enero.mod is
```

not null and [2000].dbo.enero.[sin] is not null and [2000].dbo.enero.fil is not null and [2000].dbo.enero.[c16] is not null

Por último, en este primer proceso que se ejecutará una única vez, se procede a la actualización de la tabla celda, donde el atributo idAsp adquiere el número de la columna correspondiente al modelo en el cual tiene lugar ese dato o valor estadístico. Corresponde entonces a una segunda etapa la cual enmarca la configuración de la extracción, transformación y carga hacia la base de datos del modelo multidimensional desde la base de datos operacional, es decir, la que constantemente está en procesos transaccionales de gestión estadística, y que al igual que la primera, contiene una tabla celda que incluye, además de otros, los mismos atributos que la tabla de hechos. De tal manera que este proceso no es complejo de realizar, y estaría configurado para efectuarse diariamente, puesto que con esta misma frecuencia, se insertan datos estadísticos, se actualiza o se puede actualizar la información de las empresas, de los indicadores, o de los clasificadores, y es necesario que esta información permanezca constantemente actualizada en la base de datos de diseño multidimensional.

La estrategia para los procesos ETL para los datos almacenados en la tabla celda que pertenece a la base de datos operacional mediante los paquetes de SQL Server 2000 quedaría de la siguiente manera:

Base de datos operacional	**Base de datos Multidimensional**
numM	numM
codV	codV
idind	idind
numCI	numCI
codInfAcum	idTiempo
idAsp	idAsp
valor	valor

Tabla 2: Correspondencia de los atributos de la tabla celda de la Base de datos operacional, con los atributos de dicha tabla en la Base de datos Multidimensional.

Así mismo, se debe actualizar la información referente a las tablas dimensionales, es decir las tablas Tiempo, Ubicación, Aspecto, Modelo, Variante, Indicador, donde por supuesto se especifican otros detalles, pero trabajando con el código que se describe a continuación:

Tabla Tiempo

SELECT codInfAcum, ano, semestre, trimestre, mes FROM dbo.InformeAcum

Tabla Indicador

SELECT dbo.Indicador.idind, dbo.Indicador.descripInd , dbo.Indicador.idPadreI , dbo.Indicador.codTem , dbo.Tematica.descrip FROM dbo.Indicador INNER JOIN

dbo.Tematica ON dbo.Indicador.codTem = dbo.Tematica.codTem

Tabla Aspecto

SELECT idAsp, idPadreA, descripAsp, aliasAsp FROM dbo.Aspecto

Tabla Variante

SELECT codV, idPadreV, descripV FROM dbo.Variante

Tabla Modelo

SELECT numM, descripM FROM dbo.Modelo

Tabla Ubicación

SELECT dbo.CI.numCI, dbo.CI.descripCI, dbo.CI.idOME, dbo.CI.idOTE, dbo.CI.direccionCI, dbo.OME.nombOME, dbo.OTE.nombOTE FROM dbo.CI INNER JOIN dbo.OME ON dbo.CI.idOME = dbo.OME.idOME AND dbo.CI.idOTE = dbo.OME.idOTE INNER JOIN dbo.OTE ON dbo.OME.idOTE = dbo.OTE.idOTE

El paquete que describe e implementa el proceso descrito anteriormente se muestra a continuación:

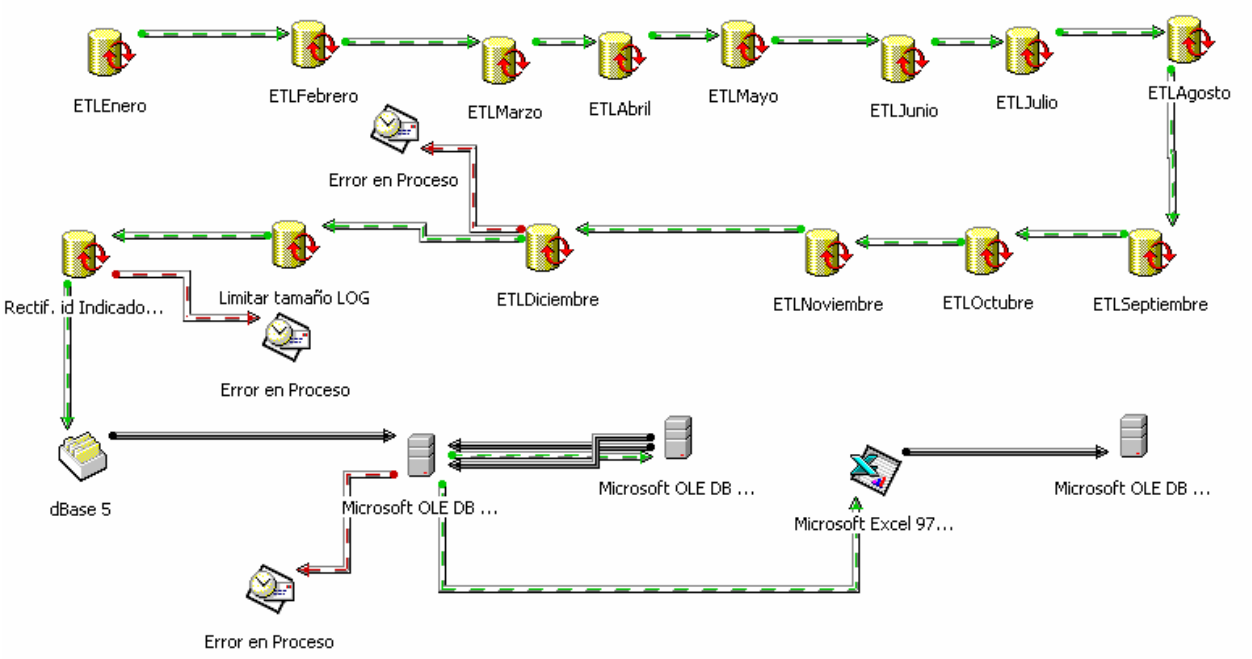


Figura: 2:14: Proceso de extracción, transformación y carga en paquete de SQL Server 2000.

2.8.5 CONFIGURACIÓN DEL ASEGURAMIENTO DE LA CALIDAD DE LOS DATOS.

Asegurar la calidad de los datos es una tarea fundamental. Se debe monitorear la forma en que se almacenarán los datos, que no haya duplicados, que exista compatibilidad en cuanto a unidades de medidas y tipos de datos, que no hayan múltiples jerarquías de dimensiones, es decir por ejemplo, que existan dos dimensiones temporales relacionadas jerárquicamente con elementos comunes, tampoco que haya información errónea o ausente, o algún conflicto o inconsistencia en cuanto a las reglas y términos definidos en el negocio de la empresa. Sería bueno también elaborar mecanismos para posibles errores o problemas que puedan surgir. Se recuerda que una casi perfecta validez de los datos y su idónea explotación, contribuye con la total confianza de los usuarios por el sistema.

Debido a que se realiza una limpieza exquisita en los procesos ETL en los pasos anteriores, evitando el almacenamiento de los valores nulos, los datos erróneos, la inconsistencia de tipos de datos o de términos propios del negocio; esta configuración se considera implícita dentro de la etapa de los procesos ETL. Lo que sí se debe priorizar es el monitoreo constante por parte de quien o quienes administren el Data Warehouse, efectuando periódicamente revisiones a la configuración de las dimensiones, las tablas implicadas y principalmente chequeos de consultas a la misma información

tanto a la base de datos multidimensional como al Data Warehouse, para comprobar la exactitud y similitud de los reportes.

2.8.6 CONSTRUCCIÓN DEL SISTEMA DE ALMACENAMIENTO DE DATOS.

Corresponde entonces al montaje en el Gestor de Base de datos SQL Server 2000 del diseño concebido y completado anteriormente, por lo que se crean las tablas con sus atributos y las relaciones entre ellas, dando paso a un sistema de almacenamiento adecuado para los procesos posteriores. Esta base de datos multidimensional contendría la información histórica de las actividades estadísticas en Cuba, desde el año 1994 hasta la actualidad, por lo cual, una vez montada en el gestor de base de datos, en conjunto con las copias operacionales de la información histórica a ser cargada una única vez y la base de datos operacional, que es la que sufre constantemente procesos de selección, inserción, actualización y eliminación de datos, a ser cargada diariamente hacia la base de datos multidimensional, entonces se procedería al movimiento de información hacia el Data Warehouse, el cual conserva el mismo diseño multidimensional que esta última base de datos.

El Módulo Generador de Reportes es uno de los módulos de nuestro Proyecto que se encarga de preparar, construir y ofrecer reportes de la información estadística almacenada en las base de datos en cada una de las Oficinas Provinciales y Municipales de Estadísticas así como en la Oficina Nacional, y al mismo tiempo integra la posibilidad de reportar la información estadística almacenada en el Data Warehouse debido a la situación problémica descrita en el capítulo anterior, donde la velocidad de respuesta a las consultas y reportes a decenas de Giga Bytes de datos almacenados llegaba a ser bastante lenta sin el uso de este sistema. La oficina crítica en este sentido es la Nacional, la cual recoge toda la información de Cuba, y es aquí donde se montaría el Data Warehouse, donde el Módulo Generador de Reportes haciendo uso del Active Report, herramienta especializada en este servicio, ha implementado una aplicación con la cual se accede perfectamente y sin demora alguna a la información del Data Warehouse.

Mediante el Analysis Services 2000, se definen entonces la conexión a la base de datos multidimensional, las dimensiones que existirán con sus propiedades, ya definidas anteriormente, y el cubo al cual se asociarán estas dimensiones, el que constituye la unidad estructural y funcional del Data Warehouse, pues además de contener toda la información estadística almacenada en distintas dimensiones, preserva el dato estadístico desde el detalle más específico hasta el resumen más general, y es al que la aplicación le realiza los reportes necesarios.

2.8.7 CARGA DE LOS DATOS AL DATA WAREHOUSE.

Es aquí donde se procede a almacenar los datos después que hayan sido extraídos y transformados previamente, para ello es necesario seguir algunos puntos, tales como: evitar la indexación en tablas con grandes volúmenes de datos, pues insertar miles o millones de tuplas en tablas indexadas hace más lento el procesamiento. Definir e identificar cuales serán las llaves o identificadores de las tablas dimensionales y de hechos, donde la llave de esta última es la concatenación o coincidencia de las demás llaves dimensionales. Al concluir este paso ya deben estar creados las tablas, índices, llaves y dimensiones necesarias.

Una vez definidos, configurados y ejecutados todos los anteriores pasos, se procede entonces a la carga de los datos al Data Warehouse. Se aconseja siempre empezar a cargar los datos por los períodos críticos de explotación, es decir, si el tiempo sobrepasa varios años, realizar primero el almacenamiento del año en curso, y luego, en retroceso cronológico, los años anteriores. Es aquí donde se define el tiempo requerido para ejecutar esta tarea, pues se requiere de un tiempo en el cual no se interactúe con el sistema, es decir fuera del horario laboral de los usuarios que intercambian con el mismo, generalmente se realiza una vez todos los días, en horarios nocturnos o los fines de semana.

En tanto que cuando los datos se encuentren almacenados previamente en la base de datos multidimensional y después de haber configurado el Analysis Manager 2000, se procede entonces a procesar el cubo, llamado CuboSIGE, para así transportar toda la información estadística que se encuentra almacenada en aquella base de datos desde el SQL Server 2000 hacia el cubo multidimensional.

Este procesamiento se debe ejecutar diariamente, puesto que con esta misma frecuencia se almacena información en la base de datos. Se ha de definir en el gestor SQL Server 2000 un paquete específico para realizar automáticamente, y en horario no laborable, este procesamiento, que puede tardar entre 1 y 3 horas aproximadamente.

2.8.8 PRESENTACIÓN DE LA INFORMACIÓN

Este es otra de las partes más importantes de la implementación del Data Warehouse, la construcción del Front End. Se le llama Front End puesto que se trata de la interfaz que interactúa con el usuario para mostrarle los datos de una forma fácil y con el formato adecuado. Sería entonces la aplicación MGR la que se encargaría de acceder a los datos almacenados en el Data Warehouse, de tal manera que los clientes que se encargarán de elaborar las consultas y reportes no tengan que poseer conocimientos de SQL ni OLAP para construir las consultas, sino que mediante configuraciones fáciles

podrá elaborarlas sin ningún problema. A continuación se muestra una de las interfaces de la aplicación donde se construye el reporte, y en los anexos aparecen otras imágenes referentes a dicha aplicación, con ejemplos de consultas realizadas. Otra de las ventajas en este sentido es la posibilidad de ofrecer los reportes tanto en forma tabular como gráfica, siendo esta ultima mucho más descriptiva, ya sea en la manera de barras, de pastel, etc., aportándole mucha más información al cliente. En el anexo 3 se puede apreciar la Interfaz de configuración del Reporte en el MGR.

Figura 2.15: Interfaz de configuración del reporte para la posterior presentación de la información en forma gráfica o tabular.

2.9 CONCLUSIONES DEL CAPÍTULO

En el presente capítulo se han descrito detalladamente cómo interactúan los componentes del Data Warehouse y la arquitectura que se ha propuesto para su desarrollo. En el mismo ha quedado de

manifiesto la interacción de los distintos factores claves para el logro de un buen diseño así como los elementos relacionados con el almacenamiento de la información, los cubos de datos, la granularidad y el modelo multidimensional. Por último y muy importante se abordan los diferentes pasos para la implementación de este Sistema, estando estos muy bien adaptados a las condiciones y características del negocio en cuestión, por lo que se han cumplido las metas propuestas para este capítulo de diseño e implementación.

CAPÍTULO 3: ANÁLISIS DE LOS RESULTADOS

Al concluir el proceso de construcción del Sistema Data Warehouse, el análisis y ejecución de algunos aspectos tales como la normalización, la validación del sistema, el testeo por parte de los clientes y el análisis de los tiempos de respuesta, resultan tan importantes como el diseño y la implementación misma. El Data Warehouse al entrar en contacto con los usuarios finales, entra en un ciclo iterativo e incremental, de lo simple a lo complejo, donde el sistema nunca descansará puesto que a él son adheridos, con el transcurso del tiempo, nuevas necesidades, procesos de negocios de la empresa o insatisfacciones del cliente. En el momento en que se implanta en la empresa y al entrar en plena explotación, el Data Warehouse crece ilimitadamente, al ser alimentado con los datos históricos, a la vez que se vuelve complejo, y es cuando comienzan a observarse los beneficios de los tiempos de respuesta, el dinamismo en la elaboración de los reportes, los conocimientos que puedan ser extraídos de la información almacenada, la efectiva preparación de los usuarios finales y la validación satisfactoria del Sistema.

3.1 NORMALIZACIÓN

En nuestro Sistema, el diseño de la tabla de hechos llega a ser tan compacto en las clases y en los datos que no hay forma aquí para normalizar en el futuro las extremadamente complejas interrelaciones muchos a muchos entre todas las llaves en la tabla de hechos que en nuestro caso vendría siendo la tabla celda con sus atributos numM, idTiempo, idAsp, codV, idind, numCI, y valor, porque nuestras seis dimensiones esencialmente no están correlacionadas unas con otras en nuestro negocio.

Ralph Kimball, toda una personalidad mundial en el diseño y construcción de Data Warehouse, en su famoso libro The Data Warehouse Toolkit plantea claramente que las tablas dimensionales no tienen que estar normalizadas sino deben permanecer como tablas planas puesto que las tablas dimensionales normalizadas destruyen la habilidad de la presentación tabulada. Los espacios en disco salvados por la normalización de las tablas dimensionales, son típicamente menores que un porciento del espacio total de disco necesario para el esquema completo. Los esfuerzos para normalizar cualquiera de las tablas en una base de datos dimensional solamente con el objetivo de salvar espacio en disco, son una perdida de tiempo.

3.2 TAMAÑO Y CRECIMIENTO.

A partir de un estimado razonable que se hará en cuanto al tamaño de la base de datos, se tendrá una concepción aproximada su la dimensión espacial. En tal sentido, se realizará un análisis de las dimensiones para calcular la cantidad de unidades en cada una, la cantidad de filas implicadas en cada una de las tablas hasta llegar al número de Bytes que serán ocupados por concepto de tamaño. Es necesario destacar que este análisis se hará con toda la información almacenada en la Oficina Nacional de Estadísticas que data desde 1994, año a partir del cual comenzaron a almacenar los datos estadísticos en las nuevas tecnologías de la información de aquella época.

Dimensión Tiempo: 14 años X 12 meses = 168 meses.

Dimensión Ubicación: 6833 Empresas o Centros Informantes distribuidas en todos los Municipios que llenan cada mes los modelos estadísticos.

Dimensión Indicador: 3056 Indicadores (aparecen como filas del modelo) de los cuales 40 como promedio se encuentran implicados en algún modelo estadístico que captan mensualmente los datos en los centros informantes.

Dimensión Aspecto: 32 Aspectos (aparecen como columnas del modelo) de los cuales 8 como promedio se encuentran implicados en cada modelo estadístico que captan mensualmente los datos en los centros informantes.

Dimensión Modelo: 52 Modelos estadísticos de los cuales 30 captan cada mes los datos en los centros informantes.

Número de registros en base de hechos: 168 X 6833 X 40 X 8 X 30 = 11020262400

Número de campos claves: 6

Número de campos hechos: 1

Total de campos: 7

Tamaño de la tabla de hechos base: 11020262400 X 7 campos X 4 bytes = 308 GB

Crecimiento anual: 308 GB / 12 meses del año = 27 GB

De tal forma se puede concluir que nuestro Data Warehouse tendría un crecimiento anual de aproximadamente 27 Giga Bytes de información, y todo el histórico almacenaría un cúmulo de aproximadamente 308 Giga Bytes. La Oficina Nacional de Estadísticas cuenta con la tecnología y la infraestructura necesaria para almacenar dicha información, por lo las limitaciones de almacenamiento no existen.

3.3 ANÁLISIS DEL RENDIMIENTO DEL SISTEMA

En este punto se analizan los rendimientos del Sistema data Warehouse que se ha construido, al dar respuesta a distintos pedidos de información, comparándolos con el rendimiento en cuanto a tiempo de respuestas a estos mismos pedidos accediendo a la base de datos que se encuentra en el gestor Microsoft SQL Server 2000.
En este análisis se propone determinar los tiempos que demora en reportar la información usando la aplicación MGR desarrollada en el Proyecto ONE para devolver los resultados para consultas de distintos grados de complejidad, sobre una cantidad determinada de filas.
El objetivo de este análisis es demostrar cuán óptimo, fácil y rápido se realizan las consultas OLAP sobre el servidor del Data Warehouse, en comparación con los mismos pedidos hechos sobre la misma información almacenada en un servidor de base de datos en SQL Server 2000.

De tal forma se da comienzo a nuestro análisis ofreciendo algunos datos de nuestro caso de estudio:

Datos:

- Fuente de datos a reportar: Cubo de datos en Analysis Manager 2000 y Base de datos con diseño multidimensional, en SQL Server 2000, ambos con información que ha sido facilitada por la ONE referente los datos estadísticos del año 2000 de todo el país.
- Tipos de consulta a realizar: Consultas SQL y OLAP.
- Dimensiones implicadas: Modelo, Tiempo, Ubicación, Aspecto, Indicador, Temática y variante. (Total: 7 dimensiones).
- Cantidad de filas en la tabla de hechos: 1 832 596
- Requerimientos de hardware del Servidor:

- Características del Servidor:

  a) Hardware: 768 MB de memoria RAM, 70 GB de capacidad de disco duro, procesador a 2.4 GHz de velocidad.
  b) Software: SO Windows server 2003, SQL Server 2000, Analysis Services 2000 y aplicación de reportes MGR.

- Consultas a realizar:

  a) Total de Salarios y Sueldos Devengados, correspondiente al mes de Enero, en la provincia de Ciudad de la Habana.

  b) Producción mercantil del último trimestre en el municipio de Bayamo

  c) Ventas Netas de Bienes y Servicios del país en el primer semestre del ano.

  d) La exportación de servicios del año en el país.

  e) El resultado anual de los indicadores asociados al modelo 500 en la provincia Camagüey.

  f) La sumatoria numérica de todos los valores asociados a todos los indicadores en Cuba en todo el año.

Las consultas aparecen enumeradas, y cada una es diferente a las demás en cuanto a complejidad y a la cantidad de tuplas que devolverán. Estas primero se estructurarán utilizando la aplicación MGR, el cual hace uso del Active Report para realizar el reporte a los orígenes de datos, y el cual, en dependencia del tipo de consulta que sea, elaborará el código pertinente para poder ejecutarla, ya sea de tipo OLAP o SQL. Luego se ejecutarán dichas consultas extrayendo los datos pertinentes, y el tiempo de respuesta será medido utilizando el propio reloj del sistema, lo cual daría una aproximación de dicho tiempo, con el cual se basó para comparar ambos tipos de consultas.

Las consultas que se realizarán sobre la base de datos montada en el SQL Server 2000 con toda la información histórica del año 2000, que además constituye la Base de datos multidimensional, se realizarán en sentencias Transact SQL, y tendrán como algo común el FROM, es decir, desde donde se realiza el pedido, el cual será desde la juntura o unión de varias tablas puesto que estas son las

implicadas con el almacenamiento de datos, además que debe coincidir con los datos almacenados en los cubos del Sistema Data Warehouse.

A continuación las consultas en Sentencias SQL:

Consulta 1:

```
select   sum(valor)as VENTAS_NETAS_TOTALES
FROM        dbo.Aspecto INNER JOIN
                  dbo.celda ON dbo.Aspecto.idAsp = dbo.celda.idAsp INNER JOIN
                  dbo.ubicacion ON dbo.celda.numCI = dbo.ubicacion.numCI INNER JOIN
                  dbo.Tiempo ON dbo.celda.idTiempo = dbo.Tiempo.idTiempo INNER JOIN
                  dbo.modelo ON dbo.celda.numM = dbo.modelo.numM INNER JOIN
                  dbo.variante ON dbo.celda.codV = dbo.variante.codV INNER JOIN
                  dbo.indicador ON dbo.celda.idind = dbo.indicador.idind
where  (ano=2006 )  and (Semestre='Semestre1' )
and  (descripInd='Ventas Netas de Bienes y Servicios' )
```

Consulta 2:

```
select   sum(valor) as ExportacionesServicios
FROM        dbo.Aspecto INNER JOIN
                  dbo.celda ON dbo.Aspecto.idAsp = dbo.celda.idAsp INNER JOIN
                  dbo.ubicacion ON dbo.celda.numCI = dbo.ubicacion.numCI INNER JOIN
dbo.Tiempo ON dbo.celda.idTiempo = dbo.Tiempo.idTiempo INNER JOIN        dbo.modelo ON
dbo.celda.numM = dbo.modelo.numM INNER JOIN
                  dbo.variante ON dbo.celda.codV = dbo.variante.codV INNER JOIN
```

```
                dbo.indicador ON dbo.celda.idind = dbo.indicador.idind
 where  (ano=2006 )  and  (dbo.Indicador.descripInd='Exportaciones de Servicios' )
```

Consulta 3:

```
select  dbo.Indicador.descripInd , sum(valor) as Valor
FROM        dbo.Aspecto INNER JOIN
                dbo.celda ON dbo.Aspecto.idAsp = dbo.celda.idAsp INNER JOIN
                dbo.ubicacion ON dbo.celda.numCI = dbo.ubicacion.numCI INNER JOIN
                dbo.Tiempo ON dbo.celda.idTiempo = dbo.Tiempo.idTiempo INNER JOIN
                dbo.modelo ON dbo.celda.numM = dbo.modelo.numM INNER JOIN
                dbo.variante ON dbo.celda.codV = dbo.variante.codV INNER JOIN
                dbo.indicador ON dbo.celda.idind = dbo.indicador.idind
 where   (ano=2006 )  and  (nombOTE='Camaguey' ) and (celda.numm = 500)
group by( descripInd )
```

Consulta 4:

```
select nombome, sum(valor) ProduccionMercantilProvincial
FROM        dbo.Aspecto INNER JOIN
                dbo.celda ON dbo.Aspecto.idAsp = dbo.celda.idAsp INNER JOIN
                dbo.ubicacion ON dbo.celda.numCI = dbo.ubicacion.numCI INNER JOIN
                dbo.Tiempo ON dbo.celda.idTiempo = dbo.Tiempo.idTiempo INNER JOIN
                dbo.modelo ON dbo.celda.numM = dbo.modelo.numM INNER JOIN
                dbo.variante ON dbo.celda.codV = dbo.variante.codV INNER JOIN
                dbo.indicador ON dbo.celda.idind = dbo.indicador.idind
```

where (ano=2006) and (Trimestre= 'Trimestre4') and (nombOmE='bayamo') and (dbo.Indicador.descripInd='Produccion Mercantil') group by nombome

Consulta 5:

```
select  nombOTE as Provincia, sum(valor) as Total
FROM        dbo.Aspecto INNER JOIN
                  dbo.celda ON dbo.Aspecto.idAsp = dbo.celda.idAsp INNER JOIN
                  dbo.ubicacion ON dbo.celda.numCI = dbo.ubicacion.numCI INNER JOIN
                  dbo.Tiempo ON dbo.celda.idTiempo = dbo.Tiempo.idTiempo INNER JOIN
                  dbo.modelo ON dbo.celda.numM = dbo.modelo.numM INNER JOIN
                  dbo.variante ON dbo.celda.codV = dbo.variante.codV INNER JOIN
                  dbo.indicador ON dbo.celda.idind = dbo.indicador.idind
where   (dbo.tiempo.ano=2006 ) and (dbo.tiempo.Mes='Enero' ) and   (nombOTE='Ciudad de la Habana' ) and (dbo.Indicador.descripInd='Salarios y Sueldos Devengados' ) group by( nombOTE )
```

Consulta 6:

```
select sum(valor) as Valor
FROM        dbo.Aspecto INNER JOIN
                  dbo.celda ON dbo.Aspecto.idAsp = dbo.celda.idAsp INNER JOIN
                  dbo.ubicacion ON dbo.celda.numCI = dbo.ubicacion.numCI INNER JOIN
                  dbo.Tiempo ON dbo.celda.idTiempo = dbo.Tiempo.idTiempo INNER JOIN
                  dbo.modelo ON dbo.celda.numM = dbo.modelo.numM INNER JOIN
                  dbo.variante ON dbo.celda.codV = dbo.variante.codV INNER JOIN
                  dbo.indicador ON dbo.celda.idind = dbo.indicador.idind
```

Luego de la elaboración de las 6 consultas en la aplicación MGR y de su ejecución en el Active Report, del cual hace uso dicha aplicación, se concluyó con los siguientes resultados:

Número de la consulta	Tiempo de respuesta (segundos)
1	8
2	8
3	10
4	10
5	14
6	21

Tabla 1: Tiempo de respuesta de las consultas SQL.

Igualmente se realizan las consultas, pero en la opción OLAP que nos brinda la aplicación MGR, quedando estructuradas los códigos de las siguientes consultas:

Consulta 1:

```
with  member [dimubicacion].[Total] as   'sum(  [dimubicacion].[nomb ote].members,measures.valor)'

  select { [dimindicador].[Level 02].[ventas netas de bienes y servicios]} on rows, {
[dimubicacion].[Total] } on columns from cubo_sige  where ( [dimtiempo].[2006].[Semestre1])
```

Consulta 2:

```
with  member [dimubicacion].[Total] as   'sum(  [dimubicacion].[nomb ote].members,measures.valor)'
select {  [dimubicacion].[Total] } on rows, {  [dimtiempo].[ano].members } on columns from cubo_sige
where ( [dimIndicador].[Exportaciones de Servicios])
```

Consulta 3:

select non empty{ [dimindicador].[Level 02].members} on rows, { [dimtiempo].[ano].[2006]} on columns

from cubo_sige where ([dimUbicacion].[nomb ote].[Camaguey] ,

dimmodelo.[num m].[500])

Consulta 4:

select { [dimindicador].[Level 02].[Produccion Mercantil]} on rows,

{ [dimubicacion].[nomb ome].[bayamo] } on columns from cubo_sige where ([dimtiempo].[2006].[Semestre2].[Trimestre4])

Consulta 5:

select { [dimindicador].[Level 02].[Salarios y sueldos devengados]} on rows, { [dimubicacion].[nomb ote].[Ciudad de la habana] } on columns from cubo_sige

where ([dimtiempo].[2006].[Semestre1].[Trimestre1].[Enero])

Consulta 6:

select { [dimtiempo].[ano].[2006] } on columns from cubo_sige

Número de la consulta	Tiempo de respuesta (segundos)
1	0.5
2	0.5
3	0.5
4	0.5
5	0.5
6	0.8
7	0.1

Tabla 2: Tiempos de respuesta de las consultas OLAP.

De tal forma se puede realizar una comparación entre ambos tipos de consultas, accediendo a la misma información, pero por supuesto, almacenada en distintas localizaciones y de diferentes formas. Como auxiliar se utilizará siguiente gráfico donde el color azul aparecen las consultas SQL y en rojo las OLAP.

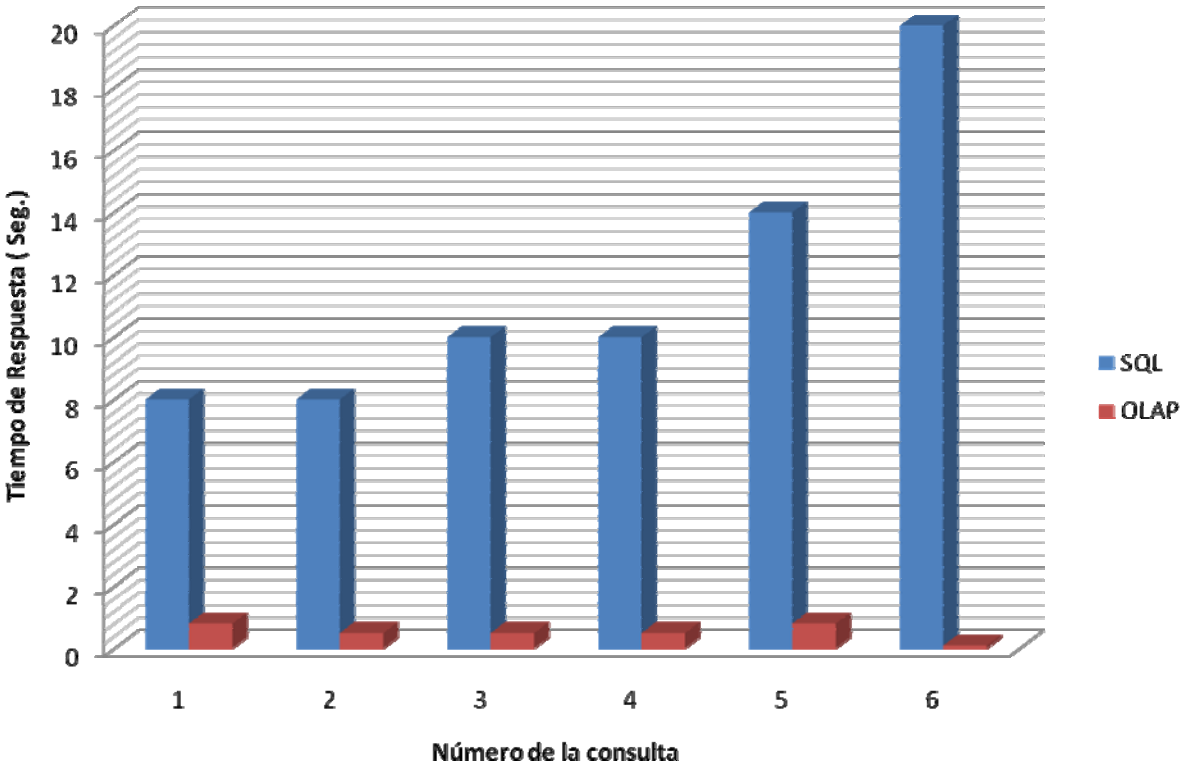


Figura 3.1: Comparación entre consultas SQL y OLAP en cuanto a tiempo de respuesta.

Una vez analizada la gráfica se percibe la abismal diferencia en tiempos de repuesta entre ambos tipos de consultas, demostrando así cuán óptimo y eficiente resulta para el acceso a la información, la utilización de Sistemas Data Warehouse con almacenamiento de grandes volúmenes de datos.

3.4 VALIDACIÓN DEL SISTEMA. RESULTADOS GENERALES

Tras haber transitado por una primera iteración de construcción del sistema, en el cual ya se ha definido un conjunto de elementos importantes referentes tanto al diseño como a la implementación del

mismo, corresponde a evaluar, validar y probar el Data Warehouse, en el cual la participación de los clientes marca un punto estratégico en esta etapa.

Realmente resulta de gran importancia que los clientes estén inmersos en la etapa de prueba del sistema data Warehouse, pues:

- Pueden ser encontradas incompatibilidades o discrepancias
- Los clientes podrán familiarizarse con el sistema.
- Pueden realizarse algunas refinaciones en el sistema.

En nuestro proyecto esta etapa duro aproximadamente 3 meses, dentro de la cual 3 especialistas de la ONE estuvieron implicados directamente en este proceso, y de los cuales uno permaneció a tiempo completo , en colaboración con el sistema en aras de pulir o refinar aquellas incongruencias o insatisfacciones que pudieran aparecer.

Uno de los puntos fundamentales en este proceso de validación fueron los tipos consultas y reportes que la ONE debía realizar, por lo que los especialistas se encargaron de recopilar y proveernos de una variado y amplio conjunto de consultas, que nos ayudaron sobremanera para enfocarnos en la forma y estructura de cómo debían realizarse los pedidos a la información almacenada.

Algunos de los detalles detectados por nuestros clientes, y que han sido solucionados satisfactoriamente, han sido los referentes a los procesos de extracción, transformación, y carga, como son:

- Conversión de los meses, trimestres y semestres de números a términos semánticos.
- Especificación de los nombres de los centros informantes, y no de sus números de identificación.
- Conversión de los identificadores de los indicadores, a partir de los identificadores de los modelos y de las filas.
- Especificación de los nombres de las Oficinas Territoriales y Municipales de Estadísticas, y no de sus números de identificación

Y otras referentes al Sistema, como:

- Posibilidad de obtener totales, promedios, porcientos, máximos, y mínimos, en las consultas OLAP.

- Existencia de sub Aspectos, que podían existir dentro de algunos Aspectos.
- Eliminación de miembros pres calculados, definidos en un inicio en el cubo de datos, por resultar innecesarios en el mismo.
- Efectuar comparaciones de resultados estadísticos, entre distintos periodos de tiempo.
- Necesidad de extender el Sistema Data Warehouse, a las restantes áreas estadísticas, de la ONE.

Esta última comenzaría a realizarse en las próximas iteraciones ya que se trata de un proceso de desarrollo iterativo e incremental, donde a medida que se integran más procesos de negocio, departamentos de la empresa y necesidades de los usuarios, el sistema va creciendo en volumen y complejidad.

Realmente esta etapa de validación y prueba, ha tenido un impacto significativo en la eficiencia y buen desempeño del Data Warehouse, puesto que se ha contado desde el primer momento de su concepción, con la colaboración, la asistencia y el apoyo de los especialistas de la ONE, especialmente de su Directora de Informática, Elena Fernández.

El hecho de que el Sistema haya cumplido los objetivos propuestos inicialmente y satisfecho total o mayoritariamente los requisitos definidos, no significa que ya esté apto para ser montado e instalado en la entidad, sino que es importante también la preparación que deben tener los especialistas que interactuarán con el Sistema, lo cual en nuestro caso se ha llevado a cabo en esta etapa, donde los clientes han adquirido los conocimientos tanto teóricos como prácticos relacionados con este tema. Es válido además mencionar que se efectuaron dos talleres de capacitación a los principales especialistas en estadísticas del país, sobre el software SIGE, y especialmente y como algo novedoso dentro del mismo el Data Warehouse, el cual tuvo una gran aceptación y apoyo por parte de estos, principalmente los que laboran en la ONE y en las Oficinas Provinciales de Estadísticas.

Resulta meritorio destacar que este trabajo se ha presentado en la Jornada Científica Estudiantil donde a nivel de Facultad obtuvo el premio Relevante y a nivel UCI el premio Destacado, dando fe de ello las imágenes de dichos premios en los Anexos.

Finalmente se puede decir que el proceso de validación del Sistema se ha realizado satisfactoriamente, a pesar que el Data Warehouse no abarca todas las áreas de la Oficina Nacional de Estadísticas, detalle que será solucionado en próximas iteraciones, pero en general los clientes se encuentran satisfechos con el trabajo realizado, hecho que ha quedado plasmado en las cartas de

aceptación por parte de ellos, las cuales se pueden encontrar en los anexos, así como también distintos premios y reconocimientos en los cuales este Sistema ha estado involucrado, por tales motivos se puede decir que los objetivos propuestos han sido cumplidos y las expectativas que se tenían con el Data Warehouse han sido superadas.

3.5 CONCLUSIONES DEL CAPÍTULO

Al concluir este capítulo se puede plantear que se han cumplido los objetivos propuestos para el mismo, superando al mismo tiempo las expectativas esperadas. Se ha demostrado fehacientemente cuán optimo resulta ser el acceso a los datos en un Data Warehouse, el dinamismo y la rapidez de los reportes y consultas, y todo sobre la base de un estudio profundo del tamaño y crecimiento del Sistema, de su rendimiento, y de la validación del mismo, tanto por sus especialistas como por los usuarios finales.

CONCLUSIONES GENERALES

Al concluir este trabajo se puede plantear que se ha cumplido con los objetivos del mismo así como con las tareas de la investigación. Se debe empezar planteando que se trata de un proceso de desarrollo bastante complejo y tedioso, el cual integra mucha información tanto de la entidad como de las herramientas con las que se trabajan y que ha llevado varios meses de trabajo intenso, en consecuencia, los resultados son alentadores.

A partir de la fundamentación teórica se enriqueció y abundó en el conocimiento sobre todo lo que respecta y se relaciona con el Data Warehouse, así como toda una investigación sobre los principales Sistemas de este tipo utilizados en el mundo y en Cuba, incluido entrevistas a especialistas en el tema, todo lo cual sirvió de base para una futura construcción del mismo. Mediante el diseño y la implementación de este Data Warehouse se podrá proveer a la Oficina Nacional de Estadísticas de un sistema eficiente y dinámico de acceso a la información con la rapidez requerida en los reportes y consultas, donde la información histórica de dicha entidad estará centralizada, ordenada e integrada, posibilitando su acceso simultáneo por varias personas. Finalmente y luego de constatar las grandes potencialidades y beneficios que brinda este Data Warehouse, nuestro Clientes se encuentran ampliamente satisfechos con lo realizado hasta hoy, en tanto que al tratarse de proceso con un desarrollo iterativo e incremental, se podrán incluir en el futuro nuevas necesidades, procesos de negocios y cambios que surjan en la Oficina Nacional de Estadísticas.

RECOMENDACIONES

- Incentivar el estudio sobre los Sistemas Data Warehouse, que tanto auge tienen hoy en día.

- Incluir dentro de este Sistema el resto de los procesos de negocios, áreas y departamentos de la Oficina Nacional de Estadísticas.

- Continuar con el proceso de desarrollo en sus próximas iteraciones para lograr que el Sistema sea robusto, flexible y que satisfaga todas las necesidades de los clientes.

BIBLIOGRAFÍA

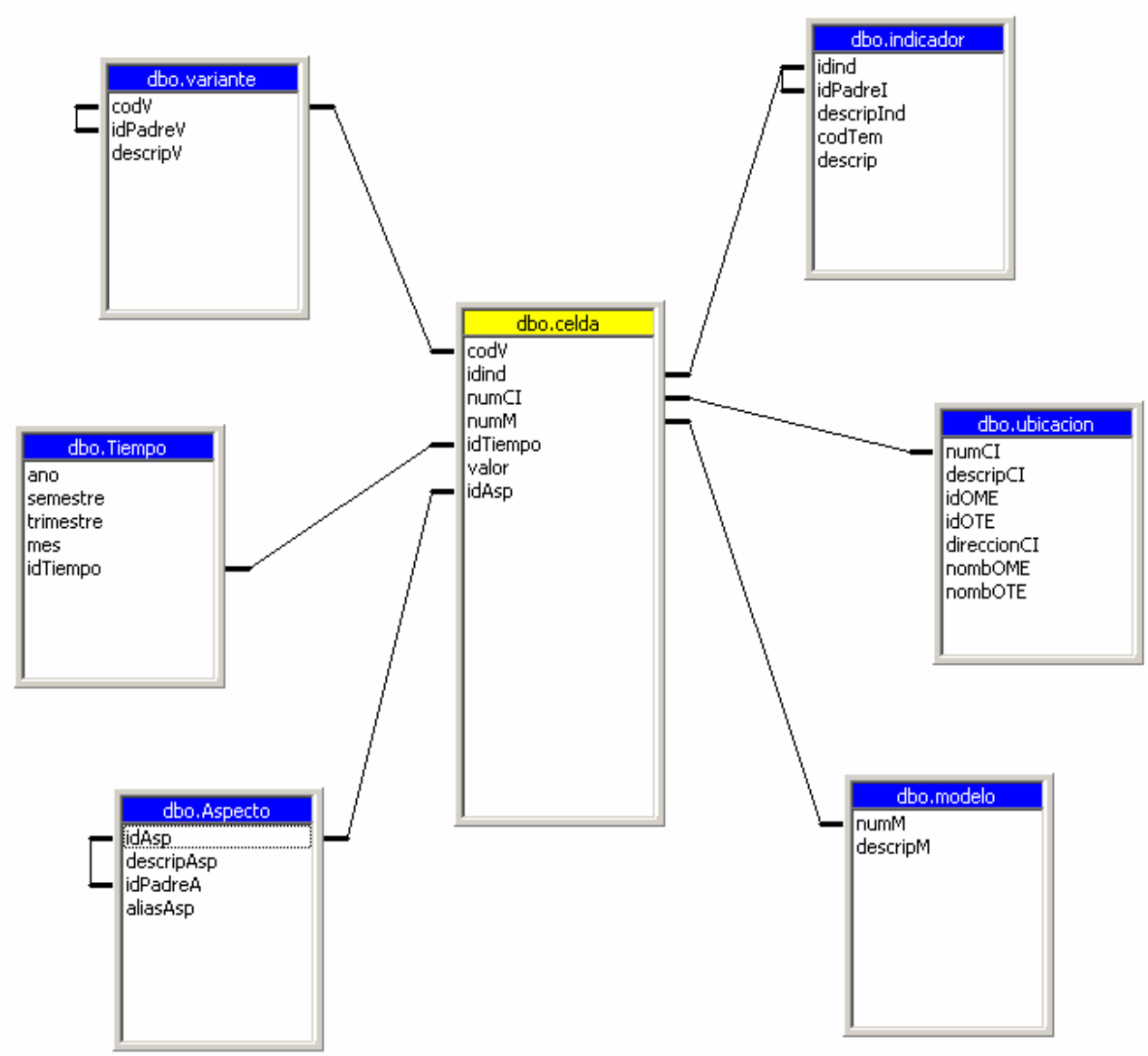
1. *Más allá de un Almacen de datos. GIGA.* Veliz, M. 2003**.** Ciudad de la Habana : n 1, 2003.
2. Greenfield, Larry. 1995. The Data warehouse information center. [En línea] 1995. [Citado el: 23 de Enero de 2008.] http://www.dwinfocenter.org.
3. Humphries, M, Hawkins, M. y Michelle, C. D. 1999. *Data Warehousing Architecture and Implementation. .* 1999.
4. Imhoff, C, Galemmo, N. y Jonathan, G. 2003. *Mastering Data Warehouse Design.* Indianapolis : Publicaciones Wiley, 2003.
5. Inmon, W. H. 2002. *Building the Data Warehouse.* New York : John Wiley & Sons, Inc., 2002.
6. Kimball, R. y Margy, R. 2002. *The Data Warehouse Toolkit.* 2002.
7. Ponniah, P. 2001. *DATA WAREHOUSING FUNDAMENTALS.* New York : Publicaciones Wiley-Interscience, 2001.
8. Poole, J., Chang, D., Tolbert, D. y David, M. 2002. *Common Warehouse Metamodel.* New York : John Wiley & Sons, Inc., 2002.
9. Wang, J. 2006. *Encyclopedia of Data Warehousing and Mining.* Hershey, PA. : s.n., 2006.
10. Corey, Michael J y Michael, Abbey. 1997**.** *Oracle Data Warehousing.* s.l. : McGraw-Hill, Inc., 1997.
11. Devlin, Barry. 1997. *Data Warehouse from Architecture to Implementaton*, 1997.
12. Singh, Harry S. *Data Warehousing: Concepts, Technologies, Implementations and Management*, 1998.
13. Inmon, William H. *Data Mart does not equal Data Warehouse*, DM Review magazine, May 1998.
14. Carter G.M, *Building Organizational Decision Support Systems*, Academic Press, 1992.
15. Anónimo. Decision Support Systems (DSS). [Citado el 5 de Mayo de 2008]. http://www.informationbuilders.com/decision-support-systems-dss.html
16. Turban, E. Aronson, J. 2001. *Decision Support Systems and Intelligent Systems*. (6a ed.). United States of America: Prentice Halll, 2001.
17. Craig, Robert S.; Vivona, Joseph A.; Bercovitch, David: Microsoft Data Warehousing: Building Distributed Decision Support Systems. Wiley Computer Publishing. ISBN: 0-471-32761-1, USA, 1999.

18. García, Lucina: Apuntes sobre Data Warehousing. Italian Cuban Seminar on Informatics Methods and Applications. Universidad de La Habana, Cuba, Diciembre, 2000.

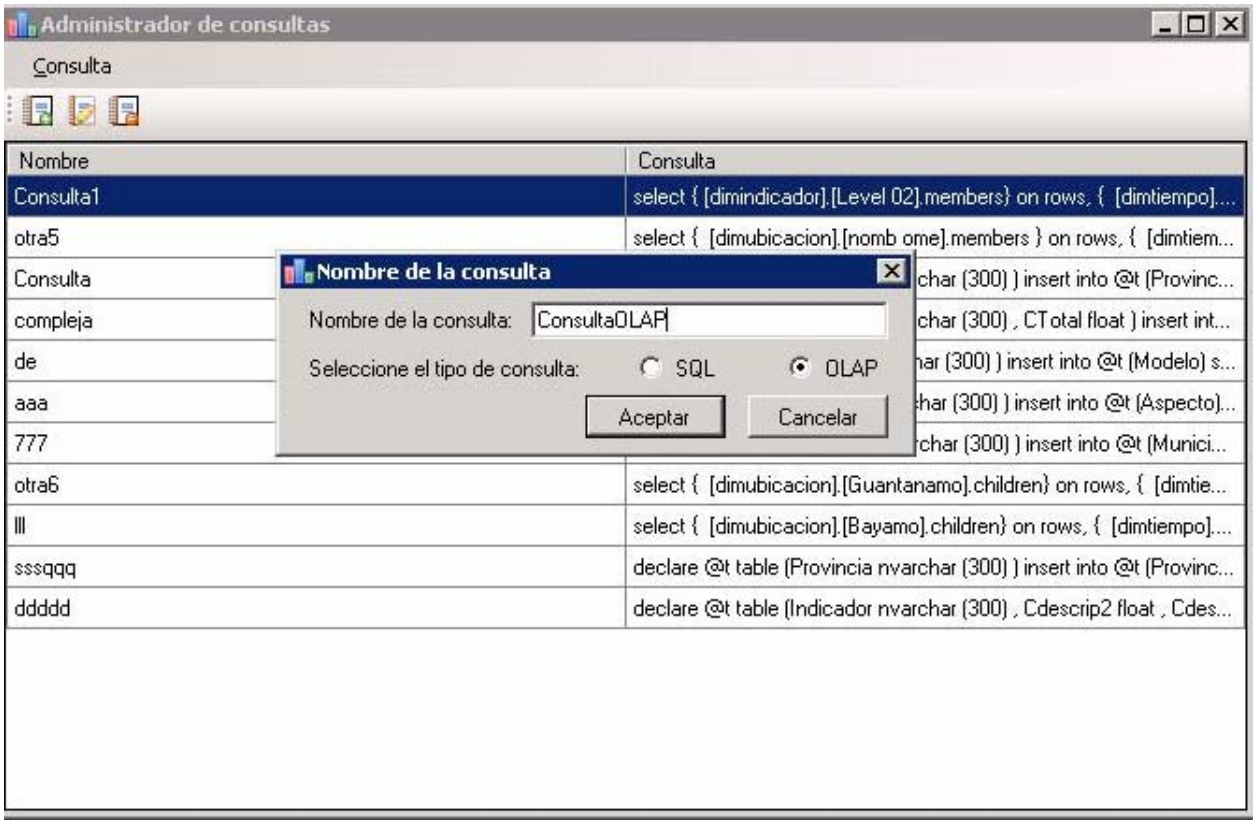
19. Giovinazzo, William A.; Foreword by Bill Inmon: Object-Oriented Data Warehouse Desing: Building a Star Schema. ISBN: 0-13-085081-0, 2001.

ANEXOS

Anexo 1: Modelo multidimensional con esquema de estrella.

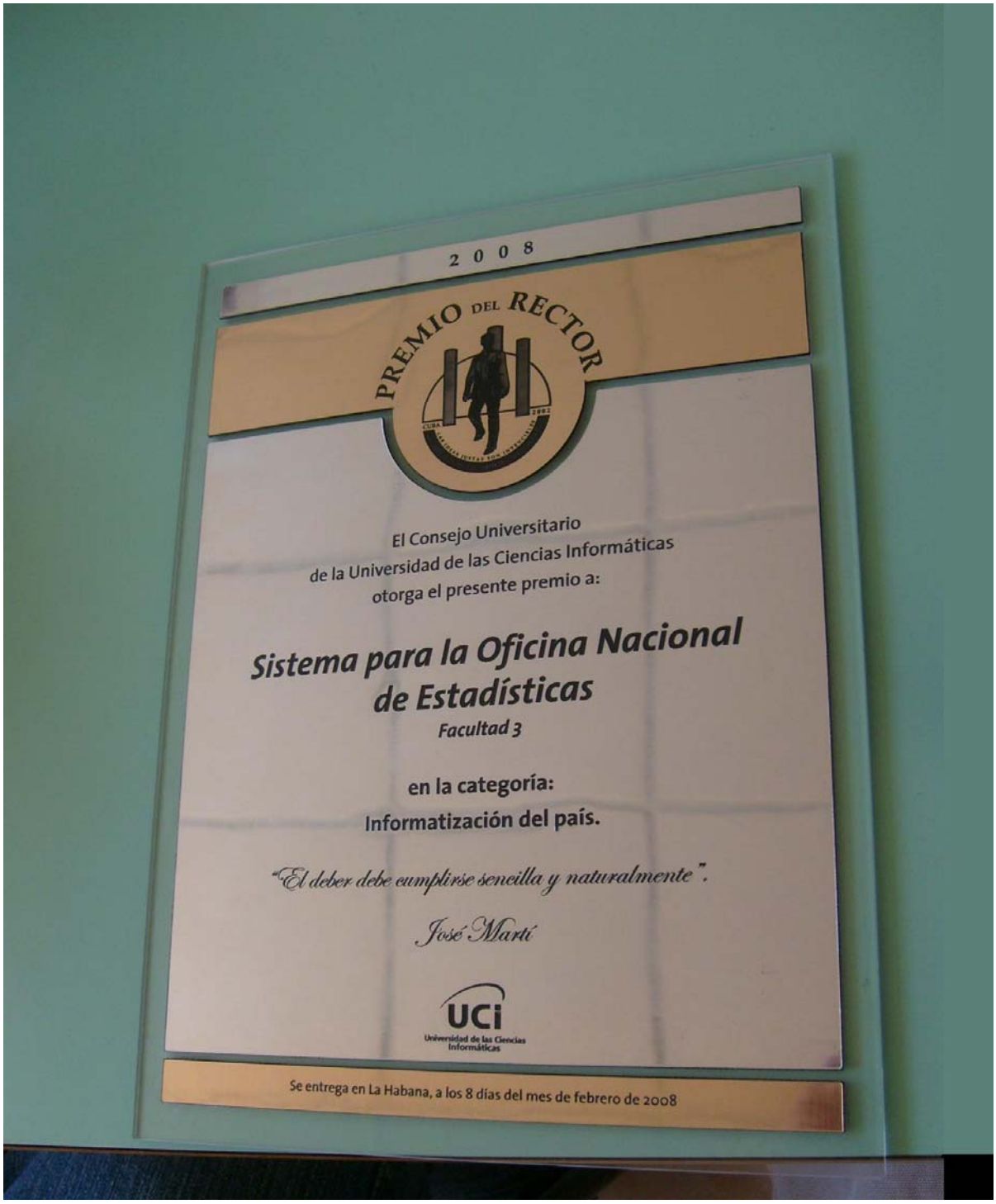


Anexo 2: Interfaz de creación de una consulta OLAP en el MGR.



Anexo 3: Interfaz de configuración del Reporte en el MGR.

Anexo 4: Premio del Rector como Mejor Software para la informatización de la Sociedad Cubana, otorgado al Proyecto SIGE, dentro del cual se encuentra este Sistema Data Warehouse.



Anexo 5: Premio Relevante otorgado a este Sistema Data Warehouse, en la pasada JCE efectuada en la Facultad 3.

UCI

Facultad 3

VI Jornada Científica

El comité organizador de la VI JCE
otorga el presente:

Reconocimiento

a: Yonelbys Iznaga González

por haber alcanzado la condición de

"Relevante"

con el trabajo: Sistema Data Warehouse para
la la Oficina Nacional de Estadística

Dado a los 15 días del mes de mayo del 2008.
"Año 50 de la Revolución"

P.S.R.
Yvonne Caridad Callado Peña
Decana Fac. 3

Yenier Figueroa Machado
Sec. Cte. UJC

Carlos R. Rodríguez Rodríguez
Pdte Fac. 3

Anexo 6: Documento avalado por la Gerente General del Proyecto y Directora de Informática de la ONE, que expresa la conformidad con los requisitos de este Sistema.

ONE
OFICINA NACIONAL DE ESTADÍSTICAS

Ciudad de la Habana, 16 de mayo del 2008.
"Año 50 de la Revolución"

Mediante el presente documento se expresa la conformidad con los requisitos del Sistema Data Warehouse en su integración con el Módulo Generador de Reportes, los cuales aparecen especificados en los siguientes puntos:

- ✓ Construir consultas OLAP de forma fácil mediante la aplicación MGR para acceder a la información histórica almacenada en el Data Warehouse.
- ✓ Almacenar los reportes como consultas tipo OLAP, para que puedan ser utilizadas cuando el cliente lo estime necesario.
- ✓ Modificar las consultas tipo OLAP una vez creadas con el asistente de la aplicación MGR, para que pueda ser adaptada a sus propias necesidades.
- ✓ Mostrar los reportes tanto en forma gráfica como tabular, de una manera rápida y sencilla.

De manera general los requisitos constituyen una propuesta que integra muy bien las funcionalidades de un Sistema Data Warehouse.

Avala la presente.

UCI Universidad de las Ciencias Informaticas
Decanato Facultad 3

Elena Leonila Fernández García
Gerente General Proyecto SIGE y Directora de Informática de la Oficina Nacional de Estadísticas.

Anexo 7: Paquete de Extracción y Transformación (Año 2000. Meses: Enero, Abril, Agosto)

```
'************************************************************************
'Microsoft SQL Server 2000
'Visual Basic file generated for DTS Package
'File Name: ETLsige.bas
'Package Name: ETLsige
'Package Description:
'Generated Date: 20/04/2008
'Generated Time: 05:49:18 p.m.
'************************************************************************

Option Explicit
Public goPackageOld As New DTS.Package
Public goPackage As DTS.Package2
Private Sub Main()
        set goPackage = goPackageOld

        goPackage.Name = "ETLsige"
        goPackage.WriteCompletionStatusToNTEventLog = False
        goPackage.FailOnError = False
        goPackage.PackagePriorityClass = 2
        goPackage.MaxConcurrentSteps = 4
        goPackage.LineageOptions = 0
        goPackage.UseTransaction = True
        goPackage.TransactionIsolationLevel = 4096
        goPackage.AutoCommitTransaction = True
        goPackage.RepositoryMetadataOptions = 0
        goPackage.UseOLEDBServiceComponents = True
        goPackage.LogToSQLServer = False
        goPackage.LogServerFlags = 0
        goPackage.FailPackageOnLogFailure = False
        goPackage.ExplicitGlobalVariables = False
        goPackage.PackageType = 0
```

```
'-------------------------------------------------------------------------
' create package connection information
'-------------------------------------------------------------------------

Dim oConnection as DTS.Connection2

'------------ a new connection defined below.
'For security purposes, the password is never scripted

Set oConnection = goPackage.Connections.New("Microsoft.Jet.OLEDB.4.0")

        oConnection.ConnectionProperties("User ID") = "Admin"
        oConnection.ConnectionProperties("Data Source") = "D:\yonelbys\documentacion\TODO
SOBRE PROYECTO ONE\La informacion de un año de ONE\Informacion\Informacion"
        oConnection.ConnectionProperties("Extended Properties") = "dBase 5.0"

        oConnection.Name = "dBase 5"
        oConnection.ID = 3
        oConnection.Reusable = True
        oConnection.ConnectImmediate = False
        oConnection.DataSource = "D:\yonelbys\documentacion\TODO SOBRE PROYECTO ONE\La
informacion de un año de ONE\Informacion\Informacion"
        oConnection.UserID = "Admin"
        oConnection.ConnectionTimeout = 60
        oConnection.UseTrustedConnection = False
        oConnection.UseDSL = False

        'If you have a password for this connection, please uncomment and add your password below.
        'oConnection.Password = "<put the password here>"

goPackage.Connections.Add oConnection
Set oConnection = Nothing
```

```
'------------ a new connection defined below.
'For security purposes, the password is never scripted

Set oConnection = goPackage.Connections.New("SQLOLEDB")

        oConnection.ConnectionProperties("Integrated Security") = "SSPI"
        oConnection.ConnectionProperties("Persist Security Info") = True
        oConnection.ConnectionProperties("Initial Catalog") = "datawarehouse"
        oConnection.ConnectionProperties("Data Source") = "(local)"
        oConnection.ConnectionProperties("Application Name") = "DTS Designer"

        oConnection.Name = "Microsoft OLE DB Provider for SQL Server 2"
        oConnection.ID = 2
        oConnection.Reusable = True
        oConnection.ConnectImmediate = False
        oConnection.DataSource = "(local)"
        oConnection.ConnectionTimeout = 60
        oConnection.Catalog = "datawarehouse"
        oConnection.UseTrustedConnection = True
        oConnection.UseDSL = False

        'If you have a password for this connection, please uncomment and add your password below.
        'oConnection.Password = "<put the password here>"

goPackage.Connections.Add oConnection
Set oConnection = Nothing

 '*************************************************************
```

'Debido al gran tamaño del código se ha omitido en este espacio fragmentos referentes a la conexión con los orígenes de datos.

```
'*************************************************************
'------------ a new connection defined below.
'For security purposes, the password is never scripted
```

```
Set oConnection = goPackage.Connections.New("Microsoft.Jet.OLEDB.4.0")

        oConnection.ConnectionProperties("User ID") = "Admin"
        oConnection.ConnectionProperties("Data Source") = "C:\Documents and
Settings\yiznaga\Escritorio\LOS INDICADORES.....xls"
        oConnection.ConnectionProperties("Extended Properties") = "Excel 8.0;HDR=YES;"

        oConnection.Name = "Microsoft Excel 97-2000"
        oConnection.ID = 5
        oConnection.Reusable = True
        oConnection.ConnectImmediate = False
        oConnection.DataSource = "C:\Documents and Settings\yiznaga\Escritorio\LOS
INDICADORES.....xls"
        oConnection.UserID = "Admin"
        oConnection.ConnectionTimeout = 60
        oConnection.UseTrustedConnection = False
        oConnection.UseDSL = False

        'If you have a password for this connection, please uncomment and add your password below.
        'oConnection.Password = "<put the password here>"

goPackage.Connections.Add oConnection
Set oConnection = Nothing
'-----------------------------------------------------------------------
' create package steps information
'-----------------------------------------------------------------------
Dim oStep as DTS.Step2
Dim oPrecConstraint as DTS.PrecedenceConstraint
'------------- a new step defined below
Set oStep = goPackage.Steps.New

        oStep.Name = "DTSStep_DTSExecuteSQLTask_1"
        oStep.Description = "ETLEnero"
        oStep.ExecutionStatus = 1
```

```
        oStep.TaskName = "DTSTask_DTSExecuteSQLTask_1"
        oStep.CommitSuccess = False
        oStep.RollbackFailure = False
        oStep.ScriptLanguage = "VBScript"
        oStep.AddGlobalVariables = True
        oStep.RelativePriority = 3
        oStep.CloseConnection = False
        oStep.ExecuteInMainThread = False
        oStep.IsPackageDSORowset = False
        oStep.JoinTransactionIfPresent = False
        oStep.DisableStep = False
        oStep.FailPackageOnError = False

'------------ a new step defined below

Set oStep = goPackage.Steps.New

        oStep.Name = "DTSStep_DTSExecuteSQLTask_4"
        oStep.Description = "Limitar le tamaño del LOG"
        oStep.ExecutionStatus = 1
        oStep.TaskName = "DTSTask_DTSExecuteSQLTask_4"
        oStep.CommitSuccess = False
        oStep.RollbackFailure = False
        oStep.ScriptLanguage = "VBScript"
        oStep.AddGlobalVariables = True
        oStep.RelativePriority = 3
        oStep.CloseConnection = False
        oStep.ExecuteInMainThread = False
        oStep.IsPackageDSORowset = False
        oStep.JoinTransactionIfPresent = False
        oStep.DisableStep = False
        oStep.FailPackageOnError = False

goPackage.Steps.Add oStep
```

```
Set oStep = Nothing

'------------- a precedence constraint for steps defined below

Set oStep = goPackage.Steps("DTSStep_DTSExecuteSQLTask_2")
Set oPrecConstraint = oStep.PrecedenceConstraints.New("DTSStep_DTSExecuteSQLTask_1")
        oPrecConstraint.StepName = "DTSStep_DTSExecuteSQLTask_1"
        oPrecConstraint.PrecedenceBasis = 1
        oPrecConstraint.Value = 0
oStep.precedenceConstraints.Add oPrecConstraint
Set oPrecConstraint = Nothing

'------------- a precedence constraint for steps defined below

Set oStep = goPackage.Steps("DTSStep_DTSExecuteSQLTask_3")
Set oPrecConstraint = oStep.PrecedenceConstraints.New("DTSStep_DTSExecuteSQLTask_2")
        oPrecConstraint.StepName = "DTSStep_DTSExecuteSQLTask_2"
        oPrecConstraint.PrecedenceBasis = 1
        oPrecConstraint.Value = 0

oStep.precedenceConstraints.Add oPrecConstraint
Set oPrecConstraint = Nothing
'----------------------------------------------------------------------
' create package tasks information
'----------------------------------------------------------------------

'------------- call Task_Sub1 for task DTSTask_DTSExecuteSQLTask_1 (ETLEnero)
Call Task_Sub1( goPackage )

'------------- call Task_Sub2 for task DTSTask_DTSExecuteSQLTask_2 (ETLAbril)
Call Task_Sub2( goPackage )

'------------- call Task_Sub3 for task DTSTask_DTSExecuteSQLTask_3 (ETLAgosto)
Call Task_Sub3( goPackage )
```

```
'------------- call Task_Sub4 for task DTSTask_DTSDataPumpTask_1 (Transform Data Task: undefined)
Call Task_Sub4( goPackage )

'------------- call Task_Sub5 for task DTSTask_DTSExecuteSQLTask_4 (Limitar le tamaño del LOG)
Call Task_Sub5( goPackage )
'**************************************************************
'Debido al gran tamaño del código se ha omitido en este espacio fragmentos referentes a la validación de valores no nulos en los campos
'**************************************************************
        oCustomTask1.SQLStatement = oCustomTask1.SQLStatement & " [2000].dbo.enero.emp is not null and" & vbCrLf
        oCustomTask1.SQLStatement = oCustomTask1.SQLStatement & "   [2000].dbo.enero.mod is not null and " & vbCrLf
        oCustomTask1.SQLStatement = oCustomTask1.SQLStatement & " [2000].dbo.enero.[sin] is not null and " & vbCrLf
        oCustomTask1.SQLStatement = oCustomTask1.SQLStatement & " [2000].dbo.enero.fil is not null and" & vbCrLf
        oCustomTask1.SQLStatement = oCustomTask1.SQLStatement & "   [2000].dbo.enero.[c16] is not null "
        oCustomTask1.ConnectionID = 1
        oCustomTask1.CommandTimeout = 0
        oCustomTask1.OutputAsRecordset = False
goPackage.Tasks.Add oTask
Set oCustomTask1 = Nothing
Set oTask = Nothing
End Sub
```

GLOSARIO DE TÉRMINOS:

ONE: Oficina Nacional de Estadísticas. Oficina encargada de las estadísticas a nivel nacional.

OTE: Oficina Territorial de Estadísticas. Oficina encargada de las estadísticas a nivel territorial. Las oficinas a nivel territorial son generalmente oficinas provinciales aunque en algunos casos existen OTE 's que realizan las estadísticas de un territorio que no se considera una provincia, como por ejemplo: La Isla de la Juventud.

OME: Oficina Municipal de Estadísticas. Oficina encargada de las estadísticas a nivel municipal.

MicroSet NT: Software utilizado actualmente por la ONE para procesar los modelos.

Centros Informantes (CI): Los Centros Informantes son las empresas u organismos que deben dar parte a las oficinas de estadísticas. En algunos casos existen establecimientos que pueden ser considerados Centros Informantes.

Aspecto: Definen los datos que van en las columnas de los modelos.

Modelo: Especie de planilla compuesta por tablas, cuadres e indicadores.

Indicador: se dice de una variable que puede tomar un valor de una determinada unidad de medida y de un determinado tipo de datos (generalmente numérico). Los indicadores de la ONE están bien definidos y tienen un código que los identifica.

Variante: Un modelo contiene variantes que especifican si el mismo pertenece a una empresa estatal o privada y a que organización. Aunque en otros caso permite especificar si el modelo tiene frecuencia mensual, trimestral o anual.

SIGE: Sistema Integrado de Gestión Estadística.

MGR: Módulo Generador de Reportes.