Universidad de las Ciencias Informáticas
Facultad 4

Título: Diseño de un Data Mart para el módulo Control de Personas del Sistema Gestión Integral Aduanera.

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

Autor: Oscar Luis Garcell Martínez.

Tutores: Ing. Pedro Manuel Alás Verdecia.
Ing. Rafael Andrés Céspedes Basterio.

Ciudad de La Habana, Junio del 2010.

"Año del 51 Aniversario de la Revolución."

Declaración De Autoría:

Declaro que soy el único autor de este trabajo y autorizo a la Facultad 4 de la Universidad de las Ciencias Informáticas; así como a dicho centro para que hagan el uso que estimen pertinente con este trabajo.

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

_____

Oscar Luis Garcell Martínez

Firma del Autor

_____ _____

Pedro Manuel Alás Verdecia

Firma del Tutor

Rafael Andrés Céspedes Basteiro

Firma del Tutor

Datos De Contacto:

Ing. Rafael Andrés Céspedes Basteiro.
Fac 15. Dpto Soluciones Financieras.
E-mail: racespesdes@uci.cu

Ing. Pedro Manuel Alás Verdecia.
Fac 15. Dpto Ingenieria De Software.
E-mail: pmalas@uci.cu

Resumen:

Entre los principales proyecto de nuestra Universidad se destaca el Sistema para la Gestión Integral Aduanera (GINA), el cual desarrolló un complejo producto de software para la Aduana General de la República (AGR). Dentro de el GINA se encuentra el subsistema Lucha contra el Fraude(LCF) y dentro de este el módulo Control de Personas, que se encarga de control, manejo, colección y gestión de la información referente a los cruces de personas naturales por frontera.

En medio del desarrollo de este software y debido a la necesidad por parte de Control de Personas de manejar un inmenso cúmulo de información, además de su análisis e investigación, entre otras causas, se propone diseñar un Data Mart para el logro de tales objetivos.

Constantemente los distintos analistas de LCF indagan y analizan distintas informaciones estadísticas, necesitan informes y reportes pormenorizados de el control de personas, para lograr un eficiente y efectivo cuidado de la soberanía económica y patrimonial de nuestro país, y es Control de Personas el encargado de ofrecer estos servicios, mediante el uso y explotación de un Data Mart 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, toma de decisiones y el descubrimiento de patrones, sobre la base de la información almacenada.

PALABRAS CLAVE

Aduana, Control de Personas, Data Mart, gestión del conocimiento, toma de decisiones.

Índice De Contenido:

Introducción:

Si innegable es el imparable desarrollo de las ciencias informáticas y las comunicaciones, aún mayor es el impacto que ha tenido en la sociedad del presente catapultándola años adelante en muchas esferas de su desarrollo. Una sociedad que aplique de forma eficiente, eficaz y competitiva la informatización en todas sus esferas, logrará un aumento en su riqueza y un aumento en la calidad de vida de todos sus ciudadanos. Cuba no está ajena a este fenómeno. El impulso de una cultura informática y una cultura digital, constituyen los pilares principales dentro de los planes de desarrollo de la Revolución.

Como decisión del Consejo de Estado de la República de Cuba de informatizar todas las empresas y entidades del país, la Aduana General de la República de Cuba (AGR) no se ha quedado rezagada en este aspecto. La AGR es una organización creada el 5 de febrero de 1963 con el objetivo fundamental de garantizar la seguridad nacional. Con este propósito se crea el área de Lucha Contra el Fraude (LCF) encargada de enfrentar las acciones terroristas, de narcotráfico, y las que ponen en riesgo el patrimonio cultural y natural del país.

Dentro del área LCF se encuentra Control de Personas, operación aduanal de gran importancia ya que controla y regula el paso de personas naturales por frontera, así como recopila la información de los hechos, indicios e ilícitos en que incurren personas naturales, procedentes del tráfico mercantil, viajero o postal.

Las decisiones que se toman en el seno de la AGR provienen de todas las operaciones aduanales y de los fenómenos que se producen en su entorno organizacional. Estas operaciones y fenómenos se expresan en términos de información. Para gestionar esta información se está desarrollando el Sistema para la Gestión Integral Aduanera (GINA), el cual constituye un sistema con diversos módulos que se encargan de automatizar y gestionar todo lo relacionado con los procesos y fenómenos aduanales.

El GINA tiene un subsistema dedicado enteramente al área LCF, y dentro de este subsistema se encuentra el módulo Control de Personas, encargado de la actividad de mismo nombre. En cada aeropuerto de nuestro país se encuentra al menos un analista de LCF, encargado de gestionar todo el flujo de pasajeros, así como monitorizar y generar reportes sobre personas que se consideren de interés para la aduana o para los órganos de seguridad del país.

Los especialistas de LCF manejan de forma eficiente pero no eficaz el gran volumen de información que se genera. A diario están entrando y saliendo personas por los puertos y aeropuertos del país. La información de los pasajeros es gestionada en sistemas computacionales con bases de datos relacionales dado su gran volumen operacional. Esto se debe a que los

especialistas cuentan con la información necesaria pero la herramienta que poseen aún no les permite acceder a ella de forma óptima y rápida. Muchas veces existe información sugerente a la que podrían tener acceso los especialistas, sin embargo se pierde, en otras ocasiones esta información es redundante, donde el reconocimiento de patrones de comportamiento se torna algo trabajoso y se emplea un tiempo muy extenso para estas tareas. Aquí el factor crítico es la integración de grandes volúmenes de información.

En estos momentos la AGR está usando los datos en sus sistemas operacionales para atender las necesidades de información contenida dentro del propio sistema operacional, o desde la Base de datos operacional. Estos datos se han usado de forma no limpia, inconsistente y no estructurada, datos sobre los cuales a diario se toman decisiones importantes. Una manera de elevar la eficiencia es hacer mejor uso de la información, pero el logro de esto se debe en gran medida a la estructura de datos reunidos en en ambiente integral centralizado.

Dentro del Control de Personas la consolidación, resumen, clasificación y reconciliación de la información se torna vital para la toma de decisiones, dado que una mala decisión podría repercutir en la economía y/o seguridad del país. Durante años los procesos aduanales dentro de LCF han generado gran cantidad de información histórica, y a diario esta cantidad crece más y más, siguiendo un desarrollo paralelo al de las tecnologías de comunicación. Sobre esta base se hace necesaria la implantación de un sistema que centralice, clasifique y de soporte a la toma de decisiones.

Los especialistas de LCF encargados de Control de Personas necesitan un sistema que les de apoyo y soporte a la toma de decisiones. El objetivo principal es tener disponible la información necesaria para los agentes durante el proceso de toma de decisiones. Al tener en cuenta el carácter cambiante de los datos actuales es necesario contar con un solo depósito integrado desde el cual los especialistas puedan fácilmente ejecutar consultas, confeccionar reportes y efectuar análisis, además que les sirva en un ambiente en que se toman las decisiones. Este ambiente de ver ser capaz de poner a completa disposición los datos almacenados en diferentes fuentes, así como reforzar la información procedente del análisis de dichos datos.

Durante los últimos años los Data Warehouse se han usado ampliamente como un apoyo para la toma de decisiones a gran nivel. Esta técnica es utilizada para la recuperación y la integración de los datos a partir de fuentes distribuidas, autónomas y posiblemente heterogéneas. Los datos son almacenados en un gran depósito llamado Data Warehouse (almacén de datos en español), que resume los datos que son organizados en dimensiones, disponiéndolos para consultas y análisis a través de aplicaciones OLAP: (proceso analítico en línea en español) y sistemas de soporte de decisiones.

Los Data Warehouse siguen un modelo dimensional, siendo esta una de sus características más sobresalientes, constituyendo a la vez una Metodología eficiente para organizar los datos de forma espacial simple, óptima y de fácil acceso para que los analistas y especialistas puedan observar y catalogar sus datos.

Un Data Warehouse resultaría una herramienta poderosa en manos de los analistas de LCF para la identificación de patrones o perfiles de riesgo. De esta forma se da un primer paso para crear una herramienta que permita a la Aduana tomar decisiones en el menor tiempo posible con respecto al flujo de viajeros.

Por lo anteriormente expuesto se infiere la necesidad de un sistema que muestre los casos de interés para los analistas de LCF, para un tratamiento y acciones operativas oportunas por parte de los inspectores y autoridades de aduana. El diseño de un Data Warehouse, constituirá el primer paso para agrupar y acceder a información que se ha almacenado, y crea una base necesaria para su implementación. Además proveerá un ambiente integral centralizado donde se simplificará el acceso a la información y se acelerará el proceso de análisis y consulta de los datos.

Ante esta situación se tiene como **problema científico:** la falta de una arquitectura de datos para la identificación de patrones y perfiles de riesgo en el flujo de viajeros.

Debido a lo anteriormente planteado se tiene como **objeto de estudio:** Los procesos relacionados con los Sistemas Informacionales.

Como **campo de acción** están los procesos de Control de Personas en la Aduana.

Para dar solución a la problemática planteada se establece como **objetivo general** realizar el diseño de un Data Warehouse para la identificación de patrones y perfiles de riesgos en el Control de Personas en la Aduana General de la República de Cuba.

Para cumplir con el objetivo y lograr una solución adecuada al problema especificado se plantean las siguientes **tareas investigativas:**

- Asimilar las herramientas, tecnologías y tendencias actuales, propuestas para el desarrollo del Data Warehouse.
- Seleccionar una metodología para el diseño e implementación del Data Warehouse.
- Analizar el negocio de Control de Personas en frontera así como la base de datos relacional actualmente en uso.
- Identificar Perspectivas e Indicadores que relacionen los temas.

- Implementar la base de datos utilizando el gestor de base de datos Oracle 11g sobre Linux.
- Realizar el proceso de ETL (Extracción, Transformación y Carga) utilizando una herramienta libre sobre Linux.
- Realizar el proceso OLAP (Procesamiento Analítico en Línea) utilizando una herramienta libre sobre Linux.
- Obtener una propuesta de diseño de un Data Mart que permita luego de su implementación la identificación de patrones y perfiles de riesgo en el flujo de viajeros.

Hipotéticamente si se concreta el diseño de un Data Mart para el Módulo Control de Personas del Sistema de Gestión Integral Aduanera, se asegura el primer paso para al proceso de integración y eliminación de información redundante de sus datos históricos, así como la agilización de la generación de reportes y el proceso de toma de decisiones.

Capítulo 1. Fundamentación teórica.

Este capítulo brinda un acercamiento a la descripción de los Sistemas informacionales, enfatizando sus principales características, así como en la descripción de su arquitectura. Se ofrecerá la descripción de los sistemas Data Warehouse, su estado actual de desarrollo, así como la caracterización de las metodologías actuales de diseño de los Data Warehouse.

El área de Lucha contra el Fraude (LCF) de la Aduana General de la República hoy en día cuenta con gran información dinámica, donde los especialistas deben tomar decisiones de forma rápida y efectiva basándose en la última información disponible. Por otra parte, las bases de datos operativas de LCF están aumentando de tamaño día a día, donde es necesario consultar gran parte de estos datos, para el trabajo diario de los especialistas. Para los especialistas es necesario un motor de dirección y una herramienta que ayude al proceso de Toma de decisiones y al reconocimiento de patrones de riesgo con eficacia. Resulta significativa la cantidad de información que puede ser extraída y analizada de la base de datos relacionales de LCF, necesaria para la toma de decisiones, por lo que se puede llegar al concepto de la necesidad de el uso de un sistema informacional.

1.1 Los Sistemas Informacionales:

1.1.1 Características de los sistemas operacionales.

En la actualidad el comportamiento de las funciones diarias de una empresa u organismo en general se monitorean a través de Sistemas operacionales o sistemas de producción, asociados a bases de datos relacionales[1]. Las bases de datos relacionales constituyen un apoyo vital para el quehacer diario de una empresa, ya que el tipo de procesamiento de los datos que están realizan garantizan el flujo de información y la automatización de procesos dentro de la misma.

El diseño e implementación de los sistemas operacionales están dirigidos a cumplir sus objetivos, que consisten en apoyar las funciones diarias de la entidad u organización, brindar servicios de oficina, entregar los datos de manera automatizada y asegurar la calidad y la protección de los datos. En general un sistema operacional se utiliza para el funcionamiento de los negocios en tiempo real, ya que las operaciones que realiza sobre los datos son la de entrada y producción. Estas operaciones son óptimas para datos que están cambiando constantemente y que necesitan de un amplio espectro por el número simultáneo de Transacciones realizadas por los usuarios.

Los sistemas operacionales realizan los que se denomina OLTP (procesamiento transaccional en línea, por sus siglas en inglés), conocido también como procesamiento operacional, que sustenta

las operaciones diarias de la empresa y describe los requerimientos operacionales del sistema. Este procesamiento se refiere a un tipo de cómputo en el cual el énfasis está en el procesamiento de las transacciones tal y como son recibidas por las aplicaciones. Este procesamiento se caracteriza además por su verticalidad, o sea los datos y el procesamiento de los mismos están orientados a la aplicación que maneja dichos datos, por lo que para un usuario común estos datos no le brindarán información, al carecer de una forma de presentación para los mismos.

Los sistemas operacionales se diseñan para proveer un análisis táctico del negocio de la empresa, por lo que debe tener una buena colocación de los datos, transacciones que minimicen los bloqueos del sistema, así como una alta Normalización de los datos y poco o ningún dato histórico o agregados. De esta manera se garantiza que se oriente correctamente el funcionamiento de el sistema operacional, que es la gestión de los datos.

1.1.2 Características de los sistemas informacionales.

Los sistemas que se utilizan para administrar y controlar la empresa son los llamados sistemas informacionales[2]. Ellos se apoyan en los datos que sustentan el proceso de toma de decisiones en una organización y en datos estables en el tiempo (datos periódicos o históricos). Se diseñan principalmente para ejecutar consultas complejas y de sólo lectura, que involucran perspectivas a partir de dichos datos.

Los sistemas informacionales realizan un procesamiento analítico en línea (OLAP, por sus siglas en inglés), conocido también como procesamiento informacional o procesamiento para la toma de decisiones. Los sistemas informacionales realizan un análisis dinámico y desde diferentes perspectivas de los datos históricos de la empresa. Estas actividades consolidan la navegación e investigación terminal de los usuarios, a través de diferentes puntos de vista.

Los sistemas informacionales explotan información estratégica relativa a la empresa, obteniendo más conocimiento que información. La implantación de un sistema informacional en una empresa evita duplicidades e incoherencias a la hora de identificar hábitos, tendencias o simulaciones de carácter histórico, enfocados en el espacio-tiempo del negocio empresarial. Los sistemas informacionales para la toma de decisiones deben ser diseñados para promover un indexado fuerte que permita agilizar la ejecución de las consultas. Además, la Denormalización de la base de datos satisface aquellos requerimientos de consultas más comunes y agiliza el tiempo de respuesta de las mismas; introduciendo datos pre-agregados o sumarizados.

Los sistemas para el proceso informacional se caracterizan además por la horizontalidad de los datos, o sea, la información se encuentra disponible y orientada a nivel de usuario, facilitando el análisis del negocio. El mejoramiento organizativo y la gestión empresarial son las principales

funciones de un sistema informacional, así como su razón de ser brindar soporte para la toma de decisiones en el ámbito empresarial.

1.1.3 Clasificación de los sistemas informacionales.

Acorde a la finalidad que persigue el sistema informacional estos se han dividido en 4 categorías o niveles[3]. Cuando un sistema informacional se usa de apoyo a la toma de decisiones y para brindar un soporte básico a el proceso de dirección se denomina sistema estratégico. Estos sistemas se caracterizan por no ser de uso periódico, o sea, su utilización no es predecible. Un ejemplo ilustrativo de este tipo de sistemas son los Sistemas de Información Georeferencial(GIS por sus siglas en inglés). Los sistemas que se ocupan de gestión documental y coordinación de actividades, para su posterior consulta, resumen o análisis se nombran sistemas tácticos. Dentro de los sistemas tácticos se encuentran los Sistemas Ofimáticos y los Sistemas de Control y Coordinación de Tareas(Workflow).

Los sistemas que operan entradas masivas de datos, así como las operaciones básicas de tratamiento de estos se denominan sistemas técnico-operativos. Estos sistemas tienen tareas predefinidas para gestionar los datos de forma administrativa, como contabilidad, facturación, almacén, personal. Estos sistemas actualmente están evolucionando con el surgimiento de autómatas, sistemas multimedia, bases de datos relacionales avanzadas y sistemas Data Warehouse.

Actualmente están surgiendo nuevos formas de categorizar a un sistema informacional, debido al aumento del uso de la Internet, el nuevo enfoque global del mercado y el propio desarrollo institucional. Estos nuevos sistemas basan su funcionamiento orientado al desarrollo organizacional del mercado, lo cual obliga a que los canales de comunicación sean mas estrechos entre la organización y su entorno circundante. En estos casos los conceptos de Intranet, Extranet y Red Global adquieren nuevos matices, dado que su uso frecuente los convierte en la columna vertebral de estos nuevos sistemas, denominados interinstitucionales Figura 1.

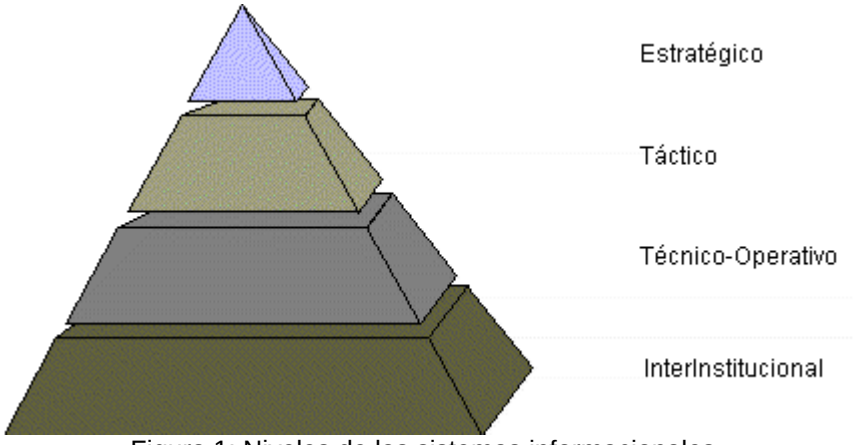

Figura 1: Niveles de los sistemas informacionales.

1.1.4 Arquitectura de los sistemas informacionales.

La complejidad de los nuevos sistemas de información, es un problema que corre a cargo de la arquitectura de los sistemas informacionales. Se entiende como arquitectura las actividades de clasificar, describir, estructurar y etiquetar los contenidos de la información. La arquitectura de los sistemas informacionales tienen un gran impacto en la usabilidad el propio sistema, aunque esta no constituye un objeto tangible para los usuarios del sistema, dado que no pertenece al universo palpable de los sistemas informacionales. La arquitectura de la información organiza los patrones inherentes de los datos, crea la estructura o mapa de la información y enfoca el conocimiento y la ciencia de la información.

La arquitectura de los sistemas informacionales deben proveer una misión y visión clara del entorno informativo, así como definir puntos de equilibrio entre las necesidades de los usuarios del sistema y las necesidades informativas del entorno. Esto se refiere a encontrar los medios por los cuales los usuarios del sistema informacional recibirán la información sin entorpecer el flujo de datos operacionales existente. En los escenarios naturales, donde la información es abundante y accesible por varias esferas se hace necesario la estructuración de forma lógica y orientada, para la posterior transformación de la información en conocimiento.

Generalmente, para la definición de la arquitectura de un sistema informacional, se definen tres niveles[4]. El primer nivel es el nivel organizacional, donde se define el ámbito en el cual se va a establecer el nuevo sistema informacional. El segundo nivel es el de definición, aquí se establece el universo conjunto de datos sobre los cuales se va a operar. El tercer y último nivel es el nivel de procesamiento, donde ya se opera directamente sobre los datos definidos en el nivel anterior.

El estado del arte actual define que un sistema informacional debe estar conformado por una fuente de datos, almacén de datos y terminales de usuarios. La fuente de datos sera conformada

por los datos operacionales, fuentes de datos externas y otras formas de manejar los datos inherentes al entorno. El almacén de datos se encargará de todo el procesamiento de los datos, los cuales serán gestionados a través de las terminales de usuarios Figura 2.

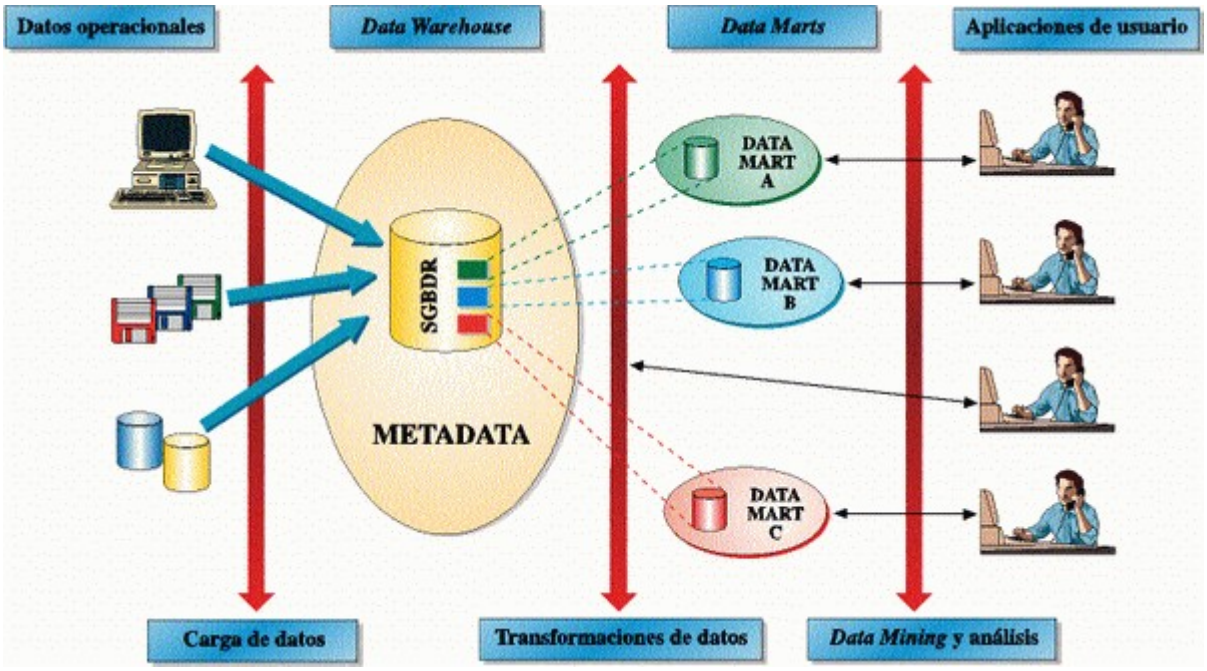

Figura 2: Arquitectura de datos

1.1.5 Objetivos generales de los sistemas informacionales.

El objetivo fundamental de un sistema informacional y las tecnologías asociadas es la mejora del rendimiento organizativo general de una entidad específica[5]. Un sistema informacional debe servir, en primer lugar, como soporte para la toma de decisiones basado en en informaciones clave. El diseño de un plan preciso de acción asociado a decisiones, de tipo jerárquicas y organizativas dentro de el ámbito organizacional corre a cargo también de este tipo de sistemas. Además el sistema informacional debe brindar monitoreos en tiempo real de las acciones inherentes a las decisiones que se toman, así como una visión global o detallada de los niveles jerárquicos del entorno del sistema informacional.

En resumen, un sistema informacional debe cumplir con el poder de síntesis de la información, o sea, el usuario debe acceder a la información globalmente o con el nivel de resumen que este desee. La integración y autonomía son otros factores que influyen sensiblemente en el desempeño de un sistema informacional. Los datos deben ser integrados correctamente y las herramientas para acceder a estos deben ser de fácil uso para los usuarios, además de que el grado de complejidad debe ser mínimo para reducir riesgos en el manejo de el sistema. Además la información debe ser accedida de forma jerárquica, para que cada usuario tenga acceso a el

conocimiento que realmente le compete.

1.2 Inteligencia de negocio.

La inteligencia de negocio es el concepto que integra por un lado el almacenamiento y por otro el procesamiento de grandes cantidades de datos, con el objetivo de transformarlos en conocimientos y en decisiones en tiempo real. La inteligencia de negocios hace hincapié en recolectar y utilizar efectivamente la información, con el fin de mejorar las operaciones dentro de una organización.

Cuanto más útil y relevante sea la inteligencia a la que tiene acceso una organización sobre su negocio, clientes, usuarios y entorno en general, podrá tomar mejores decisiones. Esto se debe a que cuando una organización conoce mejor su negocio, podrá anticiparse a comportamientos y satisfacer sus necesidades mas rápido y eficientemente. La inteligencia de negocio tiene sus raíces en los sistemas informacionales de carácter estratégico, y se han transformado en u conjunto de tecnologías capaces de satisfacer necesidades específicas en cuanto al análisis de la información.

1.2.1 Proceso de inteligencia de negocio.

A fin de comprender como una organización puede dirigir la inteligencia de sus datos, se divide el proceso de inteligencia de negocios en 5 fases[6] Figura 3:

- Fase 1 — Dirigir y Planear. En esta fase inicial es donde se deberán recolectar los requerimientos de información específicos de los diferentes usuarios, así como entender sus diversas necesidades, para que luego en conjunto con ellos se generen las preguntas que les ayudarán a alcanzar sus objetivos.
- Fase 2 — Recolección de Información. Es aquí en donde se realiza el proceso de extraer desde las diferentes fuentes de información de la empresa, tanto internas como externas, los datos que serán necesarios para encontrar las respuestas a las preguntas planteadas en el paso anterior.
- Fase 3 — Procesamiento de Datos. En esta fase es donde se integran y cargan los datos en crudo en un formato utilizable para el análisis. Esta actividad puede realizarse mediante la creación de una nueva base de datos, agregando datos a una base de datos ya existente o bien consolidando la información.
- Fase 4 — Análisis y Producción. Ahora, se procederá a trabajar sobre los datos extraídos e integrados, utilizando herramientas y técnicas propias de la tecnología BI, para crear inteligencia. Como resultado final de esta fase se obtendrán las respuestas a las

preguntas, mediante la creación de reportes, indicadores de rendimiento, cuadros de mando, gráficos estadísticos, etc.

- Fase 5 — Difusión. Finalmente, se les entregará a los usuarios que lo requieran las herramientas necesarias, que les permitirán explorar los datos de manera sencilla e intuitiva.

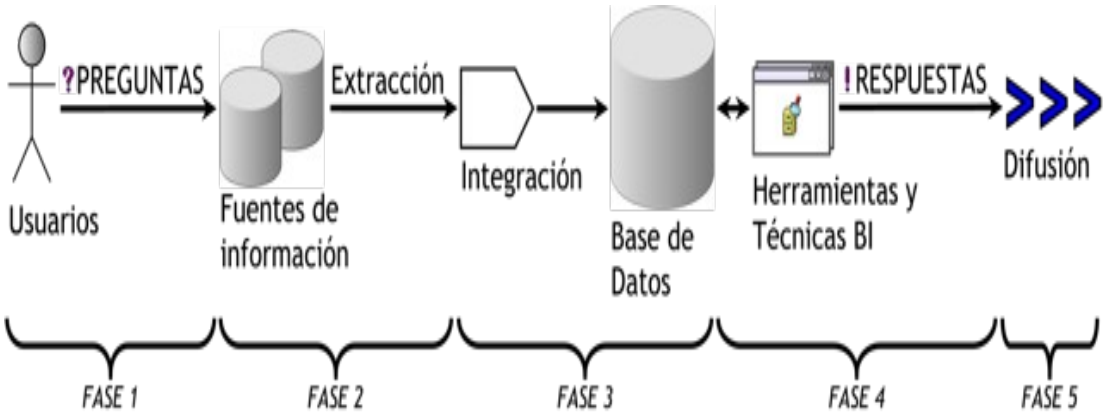

Figura 3: Procesos de inteligencia de negocio

1.2.2 Beneficios de la inteligencia de negocio.

- Reduce el tiempo mínimo que se requiere para recoger toda la información relevante de un tema en particular, ya que la misma se encontrará integrada en una fuente única de fácil acceso.
- Automatiza la asimilación de la información, debido a que la extracción y carga de los datos necesarios se realizará a través de procesos predefinidos.
- Proporciona herramientas de análisis para establecer comparaciones y tomar decisiones.
- Cierra el círculo que hace pasar de la decisión a la acción.
- Permite a los usuarios no depender de reportes o informes programados, porque los mismos serán generados de manera dinámica.
- Posibilita la formulación y respuesta de preguntas que son claves para el desempeño de la empresa.
- Permite acceder y analizar directamente los indicadores de éxito.
- Se pueden identificar cuáles son los factores que inciden en el buen o mal funcionamiento de la empresa.

- Se podrán detectar situaciones fuera de lo normal.
- Permitirá predecir el comportamiento futuro con un alto porcentaje de certeza, basado en el entendimiento del pasado.
- El usuario podrá consultar y analizar los datos de manera sencilla e intuitiva.

1.3 Los sistemas Data Warehouse.

1.3.1 Objetivos de los Data Warehouse.

Con la implementación de un Data Warehouse se deben cumplir los siguientes objetivos:

- Proveer una única visión de los clientes a través de toda la compañía.
- Proveer la mayor cantidad de información a la mayor cantidad de personas dentro de la organización.
- Mejorar el tiempo de emisión de informes.
- Monitorizar el comportamiento de los clientes.
- Mejorar la capacidad de respuesta a las cuestiones del negocio.
- Mejorar la productividad.

1.3.2 Características de los Data Warehouse.

Los sistemas Data Warehouse (DWH) o almacenes de datos constituyen la herramienta fundamental para el manejo y control de conocimiento generado en una entidad, así como soporte principal para el proceso de toma de decisiones. Los DWH constituyen fuentes organizadas, integradas, lógicas y dinámicas de los datos históricos de un ambiente organizativo. Existen según W. H. Inmon[7] cuatro características fundamentales que describen a los almacenes de datos: orientado al tema, integrado, de tiempo variante y no volátilFigura 4.

- **Orientado a temas:** Un DWH debe clasificar su información por aspectos que sean de interés para el entorno sobre el cual se está desarrollando. Una diferencia notable existente entre los sistemas operativos y los sistemas DWH es que los sistemas operativos centran su atención en aplicaciones y funciones, y los sistemas DWH se centran en sujetos. Las aplicaciones en ambientes organizacionales serían todos los artefactos sobre las cuales la entidad organizacional opera, ejemplo libros, historias clínicas, cheques. Los sistemas DWH centran su atención en el sujeto del ambiente organizacional, ejemplo de estos serian estudiantes y profesores, para una universidad o doctores y pacientes para un

hospital. La alineación alrededor de las áreas de los temas afecta el diseño y la implementación de los datos encontrados en el DWH. Las principales áreas de los temas influyen en la parte más importante de la estructura clave. El data warehousing se enfoca en el diseño de la base de datos yen el proceso. El Data Warehouse excluye aquellos datos que no son necesarios para el proceso de toma de decisiones, y despliega estos en un espectro de tiempo como cantidades cuantificables.

- **Integrado:** El objetivo principal que se persigue con un almacén de datos es que la información esté siempre integrada. La integración de datos se puede lograr de muchas maneras: convenciones de nombres, medida uniforme de variables. Un problema puntual constituye las fuentes múltiples de datos. No importa a que fuente pertenezcan los datos, ni las diferencias estructurales o de convenciones que existan entre estas: los datos deben llegar al Data Warehouse siguiendo una misma medida. Cualquiera que sea el diseño de la base de datos, el resultado debe ser el mismo, la información debe ser almacenada en el Data Warehouse en un modelo globalmente aceptable y singular, aún cuando las fuentes de datos almacenen los suyos de manera diferente.
- **De tiempo variante:** Toda la información contenida dentro de un DWH es requerida en algún momento. En ambientes operacionales, la información se requiere en el momento de acceder, en los sistemas Data Warehouse, la información se solicita en cualquier instante de tiempo, por eso los datos reciben la denominación de tiempo variante. Otra característica es que los datos históricos en ambientes operacionales son de poco o nulo uso, en los sistemas DWH los datos históricos son de uso continuo, ya que estos se utilizan para la identificación y evaluación de tendencias. Las aplicaciones DWH tiene un largo horizonte de tiempo, y dado que los DWH son aplicaciones diseñadas los suficientemente flexibles, pueden procesar gran cantidad de datos sin afectar su rendimiento.
- **No volátil:** Es un hecho que la información solo es útil cuando es estable. En los sistemas operacionales, los datos sufren, registro a registro, constantes cambios que influyen en su actualización, entiéndase las operaciones de inserción, modificación y borrado de datos. El los sistemas DWH estos es mucho mas fácil, dado que las únicas operación que se realizan es la de carga y acceso de los datos. Las bases de datos operacionales necesitan de backup y recuperación, transacciones e integridad de los datos y la detección y solución al estancamiento. En los sistemas DWH no es necesario el procesamiento de los datos.

Figura 4: Características de los Data Warehouse

Debe considerarse que los datos son filtrados y resumidos cuando pasan de las fuentes externas al DWH, por lo que el DWH solo tendrá acceso a los datos que realmente son necesarios. Los rangos de tiempo entre las fuentes y el DWH también varían, las fuentes de datos tienen datos recientes, sin embargo el DWH tiene acceso a datos con rangos de tiempo grandes. Al efectuarse el resumen de los datos que van a ser pasados al DWH, estos sufren una transformación, se alteran física y radicalmente cuando son llevados al almacén de datos. Sin embargo, no existirá redundancia entre los ambientes operacionales e informacionales teniendo en cuenta la integración de los datos.

1.3.3 Costos versus Valor.

Antes de iniciar cualquier proyecto de software, y más si se trata de sistemas informacionales de gran envergadura como los Data Marts, es necesario tener en cuenta si el valor agregado compensa la inversión. Como se señalo, para un proyecto general de data warehousing es necesario tener en cuenta los costos principales, así como los beneficios, los cuales quedan resumidos en la siguiente tabla[8].

Costos	Beneficios
Construcción	Entrega de información.
Mantenimiento	Mejora de el proceso de toma de decisiones.
Operación	Valor agregado sobre procesos empresariales.

En los costos de construcción se pueden señalar como valores a medir **los recursos humanos**, **la tecnología** y **el tiempo**[8]. De los recursos humanos se requiere la colaboración de la empresa que domine las características de el negocio, así como de los especialistas tecnológicos. De la tecnología se debe tener en cuenta que en un futuro el Data Warehouse puede incluir nuevas tecnologías que incurra en determinados gastos.

En los costos de operación se debe tener en cuanta que cuando el Data Warehouse esté construido y entregado este debe tener asociado un valor empresarial, tanto de el tipo evolutivo, de crecimiento como versatilidad ante cambios.

El valor de un Data Warehouse se puede medir en tres dimensiones:[8]

- Mejora en la entrega de la información: información completa, correcta, consistente, oportuna y accesible.
- Mejora en el proceso de toma de decisiones: con un mayor soporte a las decisiones estas se toma mucho más rápido, además de tener una mayor visión de el impacto de las mismas.
- Impacto positivo sobre los procesos de negocio: se eliminan los retardos producidos por informaciones incorrectas o inconsistente, se optimiza los procesos empresariales a través de el uso integrado de las fuentes de información y se elimina el uso de los datos que no son necesarios para ciertos procesos empresariales.

1.3.4 Estructura de los Data Warehouse.

Los Data Warehouse tienen una estructura completamente distinta la los demás sistemas informacionales, dado que los DWH persiguen niveles de esquematización y detalles delimitadores. En general un Data Warehouse se compone de detalles de datos actuales, detalle de datos antiguos, datos ligeramente resumidos, datos completamente resumidos y los meta-datos. Figura 5

Los detalles de datos actuales son aquellos que reflejan las ocurrencias más recientes, son de gran volumen y generalmente se almacenan en disco duro. Los detalles de datos actuales son de fácil acceso, aunque su costo de administración se torne costoso y complejo. Los detalles de datos antiguos se almacenan en formas de almacenamiento masivo externas. Los datos antiguos no son accedidos con frecuencia, y se almacenan con un nivel de detalle consistente con los datos actuales.

Los datos que proviene desde un bajo nivel de detalle encontrado al nivel de detalle actual son los datos ligeramente resumidos. Este tipo de datos se obtienen sobre unidades de tiempo de

esquematización hechas y definiendo los atributos que tendrán los datos a resumir. Los datos completamente resumidos se encuentran dentro del DWH de forma compacta y de fácil acceso. Los datos completamente resumidos son parte esencial del Data Warehouse sin considerar donde se alojan los datos físicamente.

Los meta-datos son el componente final de un Data Warehouse. Los meta-datos son usado como directorios para ayudar a los analistas a ubicar contenidos dentro del DWH. Constituyen una guía para el Mapeo y algoritmización de datos. Los meta-datos son los responsables de guiar los procesos de extracción, carga y limpieza de los datos, además de ayudar a las herramientas de consulta y los generadores de informe funcionen correctamente [9]. Los meta-datos se refieren a información estructural, de contenido e interdependencia existente entre los componentes del DWH. Generalizando, los meta-datos definen objetos dentro del Data Warehouse.

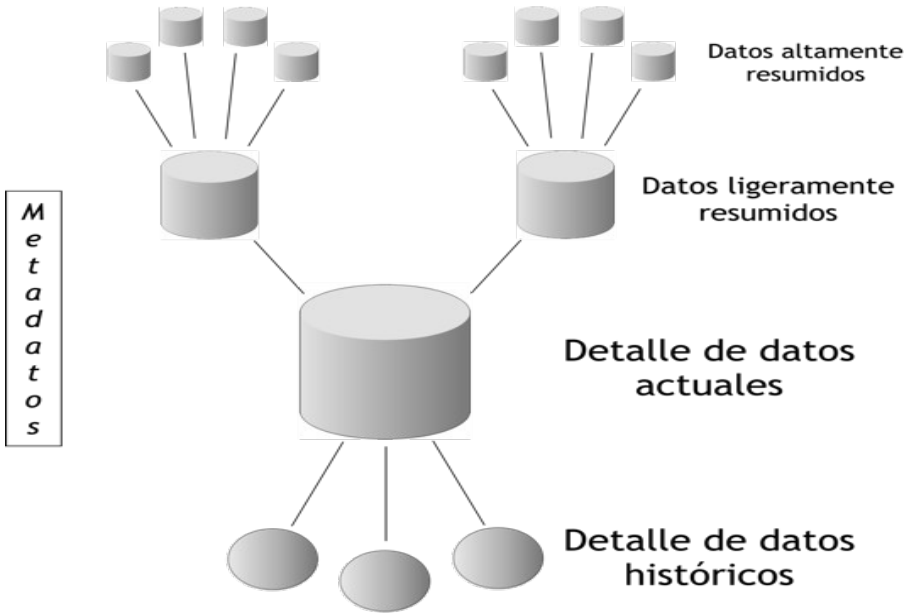

Figura 5: Estructura de los Data Warehouse

1.3.5 Métodos más usados en la construcción de Data Warehouses.

Los modelos propuestos por William H. Inmon y Ralph Kimball para llevar a cabo el diseño de un DWH son los más aplicados en la actualidad, coincidiendo en que un Data Mart o un Data Warehouse independiente no satisface las necesidades que tienen las compañías a escala corporativa de acceder inmediatamente y con facilidad a sus datos, pero sus criterios difieren en cuanto al modelo de datos y a las arquitecturas.

El término Data Mart es usado para designar a los almacenes de datos cuyo ámbito es más reducido, normalmente un departamento o área específica dentro de la empresa, es definido por

Ralph Kimball como bodegas de datos con información de interés particular para un determinado sector de la empresa y aunque su enfoque sea para una sola perspectiva departamental, no lo exime de tener que seguir los lineamientos generales de implementación que posee el Data Warehouse[9].

Kimball propone como modelo de datos el modelo dimensional. Este modelo se caracteriza por ser sencillo de implementar, muy estable en cuanto a cambios además de ser comprensible para los usuarios. Kimball sugiere este modelo para la creación de los Data Warehouse y los Data Marts[10].

Inmon reconoce el modelo dimensional como el mejor para el desarrollo de los Data Mart, por las ventajas que se refieren, pero propone para el diseño de el Data Warehouse el modelo Entidad-Relación, por ser mas adaptable[7].

Para la arquitectura Inmon refiere que no se debe sustituir la construcción de el Data Warehouse con el diseño de varios Data Marts. Pone como respuesta que los Data Marts son para áreas especificas de la empresa, y en algunos casos puede traer conflictos de datos, y que estos al final no sean lo suficientemente flexibles, reusables o útiles. Inmon propone que se diseñe primeramente el Data Warehouse, y de ahí se parte a darle atención a los diferentes departamentos, para definir y crear los Data Marts. Figura 6

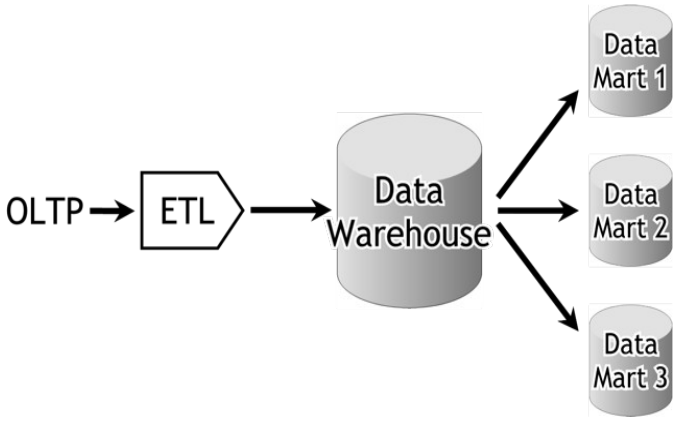

Figura 6: Modelo de William Inmon

Kimball, en desacuerdo con el modelo de Inmon, refiere a que los Data Marts están indiscutiblemente ligados a los datos de la fuente, y no constituyen una visión departamental. Kimball refiere que los Data Marts constituyen solo una parte del producto general. Kimball se pronuncia además por que los Data Mart se centran en estrategias adaptables e incrementales al tener una visión más particular sobre porciones del negocio. Por esta razón manifiesta que el proceso de construcción de un almacén de datos parte de los sistemas operacionales existentes, creando los diferentes Data Marts basados en la información de dichas fuentes, para luego de tenerlos desarrollados y funcionales se comience con la construcción del Data Warehouse basado

en la información que éstos contienen. Figura 7

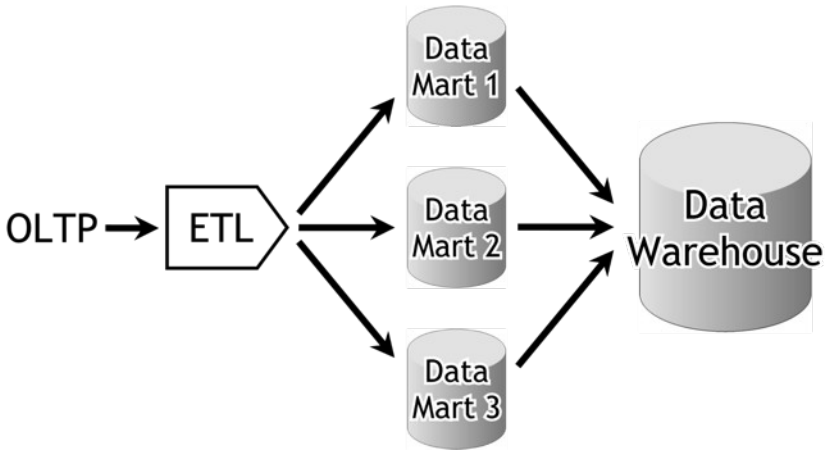

Figura 7: Modelo de Ralph Kimball

En la actualidad este es el método mas usado, gracias a las ventajas que proporciona: provee a las empresas y entidades llevar a cabo los proyectos de forma separada y de esta forma reducir los efectos fallidos que podría llevar a cabo la implementación de un Data Warehouse general.

1.3.6 Arquitectura conceptual de los datos.

Para realizar el diseño de un Data Warehouse es necesario definir una estructura que reúna todos los componentes del Data Warehouse. Este estructura es conocida como arquitectura , y es la forma de representar la organización de los datos, comunicación, procesamiento y presentación. El estado del arte actual define el uso de diferentes arquitecturas conceptuales de datos a nivel lógico. El los Data Warehouse se representa a través de capas los niveles por los cuales circulan los datos, siendo el numero de capas la forma de nombrar la arquitectura conceptual de los datos.

La arquitectura de una sola capa [11] se caracteriza por la información se guarda una sola vez en el Data Warehouse, almacenándose solamente los datos de tiempo real. Sobre estos datos actúan las bases de datos informacionales y operacionales, lo que en ocasiones suele traer bloqueos a la hora de realizar operaciones sobre un mismo conjunto de datos. Figura 8

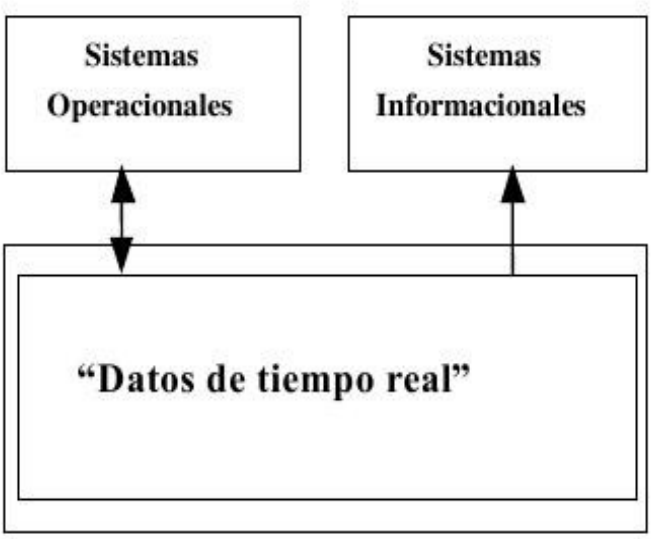

Figura 8: Arquitectura de una capa

La arquitectura de dos capas, se perfecciona el modelo de una capa, ya que se añade otra capa que contendrá los datos en tiempo real utilizados por las aplicaciones operacionales en modo lectura/escritura. Además los datos usados por las aplicaciones informacionales se almacenaran en una capa habilitada para el efecto, los cuales serán una copia directa de los datos en tiempo real o el resultado de procesos sobre los mismos. Este modelo tiene como defecto la capacidad de almacenamiento ya que se duplican los datos, pero se mejora la accesibilidad y disponibilidad de los mismos. Figura 9

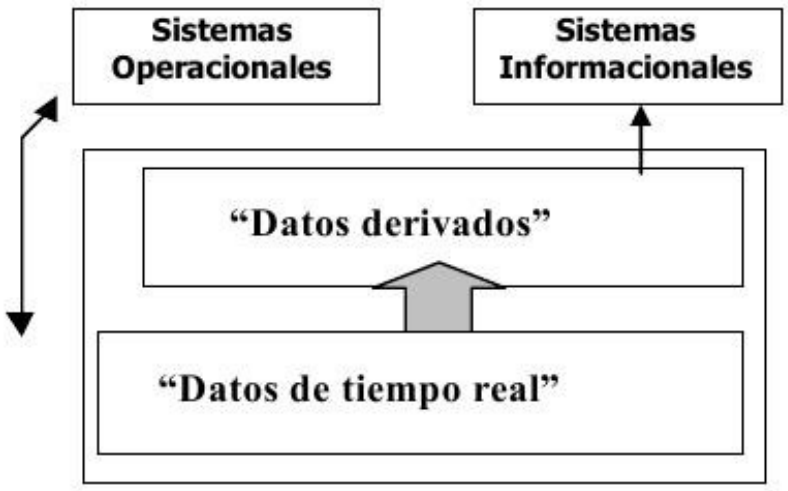

Figura 9: Arquitectura de dos capas

La arquitectura de tres capas añade una capa intermedia para la transformación de los datos de tiempo real a datos derivados y para evitar problemas de inconsistencias. A esta capa se le

denomina capa de datos reconciliados. Figura 10

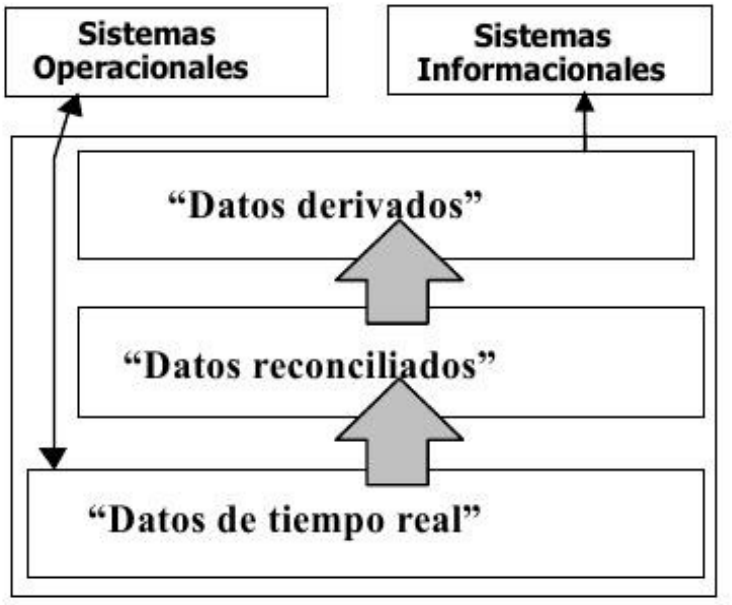

Figura 10: Arquitectura de tres capas

1.3.7 Arquitectura de los Data Warehouse.

Los Data Warehouse son una tecnología de fácil comprensión, por lo que tienen un crecimiento acelerado. La forma más fiel de representar la estructura amplia de administración de información dentro de un ambiente organizacional es usando un Data Warehouse. Para definir la estrategia de un Data Warehouse, es necesario comprender como se relacionan los componentes involucrados en la arquitectura de un Data Warehouse. Una Arquitectura Data Warehouse representa la estructura total de datos, comunicación, procesamiento y presentación, que existe para los usuarios finales del Data Warehouse. Un Data Warehouse se compone de ocho niveles arquitectónicos[12]:

- **Nivel de base de datos externo:** El objetivo de un Data Warehouse es combinar las información proveniente de bases de datos operacionales con fuentes de datos externas. Los ambientes organizacionales necesitan datos de fuentes externas, generalmente combinadas con reportes de otras organizaciones o a través de la Internet. El nivel de base de datos externo cumple la función de definir las bases de datos operacionales y fuentes externas que serán usadas.
- **Nivel de acceso a la información:** Este nivel es completa responsabilidad del usuario final del Data Warehouse, ya que define las herramientas que serán usadas para el acceso a la información. En este nivel también se define el Hardware y el Software

involucrados en las formas de presentación de la información.

- **Nivel de acceso a datos:** Este nivel sirve de puente entre el nivel de base de datos y el nivel de acceso a la información. En el estado del arte actual, el Lenguaje Estructurado de Consultas (SQL, por sus siglas en inglés), es el lenguaje utilizado para la comunicación entre los mismos.
- **Nivel de directorio de datos:** Este nivel actúa como repositorio para almacenar y gestionar los meta-datos con el fin de proveer acceso a los datos universales. Para que este repositorio de datos sea completamente funcional es necesario tener una variedad de meta-datos disponible, información sobre las vistas de datos e información sobre las bases de datos operacionales. La idea general del repositorio de datos es que los usuarios accederán a la información desde el Data Warehouse sin necesidad de conocer donde residen los datos ni de que forma están siendo gestionados.
- **Nivel de gestión de procesos:** En este nivel se definen las tareas que deben realizarse para la construcción y mantenimiento del Data Warehouse y del directorio de datos. Además es el encargado de controlar varios procesos que deben ocurrir para mantener el Data Warehouse actualizado.
- **Nivel de mensaje de la aplicación:** Este nivel regula el flujo de información a través de toda la red organizacional. Puede usarse además como herramienta para recolectar las transacciones y los mensajes, así como su almacenamiento en un lugar seguro.
- **Nivel de Data Warehouse (físico):** En este nivel es donde ocurre el almacenamiento físico de datos, generalmente para usos estratégicos. Los datos almacenados deben ser flexibles y fáciles de acceder. Generalmente el Data Warehouse se ve como una vista lógica o virtual de datos, pues en muchas ocasiones los Data Warehouse pueden no involucrar almacenamiento de datos.
- **Nivel de organización de datos:** Conocido también como gestión de copia o réplica, incluye procesos para combinar, cargar datos para el depósito, resumir, acceder a la información desde bases de datos operacionales y/o externas, permite además el análisis de calidad de datos y filtros que identifican modelos y estructura de datos dentro de la data operacional existente. Este nivel es el componente final de la arquitectura de un Data Warehouse.

1.3.8 Procesos de los Data Warehouse.

Detrás de la arquitectura de un Data Warehouse existen un conjunto de procesos de gran

importancia para el mismo[13]. El ciclo de uso de un Data Warehouse comienza con la **extracción** de los datos, que consiste en el estudio de los datos fuente, filtrando aquellos que son de utilidad. Estos datos después de extraídos, sufren un proceso de **transformación** para llegar de forma presentable y valuable para los usuarios finales. El proceso de transformación incluye además corrección de errores, filtrado de datos, generación de claves, gregación de información, etc.

Terminada la transformación de los datos, se realiza la **carga** de los mismos en el Data Warehouse, paralelamente con los **controles de calidad** para asegurar que dichos datos sean correctos. Al encontrarse los datos completamente cargados, y disponiendo de las herramientas de consulta adecuados, ya se puede comenzar la explotación de el Data Warehouse. El Data Warehouse por su versatilidad permite también que se realice el proceso inverso a la carga de datos, o sea, envió de información desde el Data Warehouse a las bases de datos operacionales, este proceso es conocido como **retro-alimentación de datos** o **feedback**.

Durante todo el proceso de creación del Data Warehouse se deben realizar **auditorías** al mismo, para conocer de donde proviene la información, así como que cálculos la generaron. Ya construido el Data Warehouse, es de interés que la información llegue a el mayor numero de usuarios y a la vez que estos usuarios sean los correctos para el acceso a dicha información, por lo que debe implementarse un **sistema de seguridad.** Además, se deben realizar actividades de **backup** y **restauración de la información**, tanto de la almacenada en el Data Warehouse como de la que circula desde los sistemas fuente al almacén, para garantizar la integridad y perdurabilidad de la misma.

1.3.9 Bases de datos multidimensional.

Los Data Warehouse gestionan el depósito de datos y lo organizan en torno a una base de datos multidimensional[14]. Como su nombre indica, la base de datos multidimensional almacena los datos en dimensiones, que conforman un cubo multidimensional, donde el cruce de los valores de los atributos de cada Dimensión determinan un hecho específico. Los cálculos que se aplican son matriciales, dando a lugar a reportes tabulares. Las bases de datos multidimensionales implican tres variantes posibles de modelamiento, que permiten realizar consultas de soporte de decisión:

- Esquema copo de nieve (Snowflake Scheme).
- Esquema constelación o copo de estrellas (Starflake Scheme).
- Esquema en estrella (Star Scheme).

Esquema copo de nieve:

Este esquema representa una extensión del Esquema Estrella cuando las dimensiones se

organizan en Jerarquía de dimensiones. Figura 11

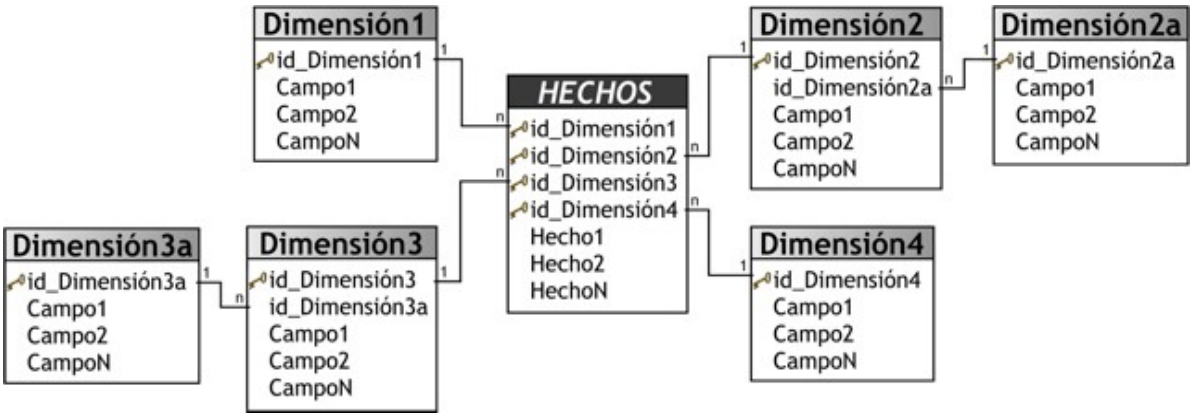

Figura 11: Esquema copo de nieve

Esquema constelación: Figura 12

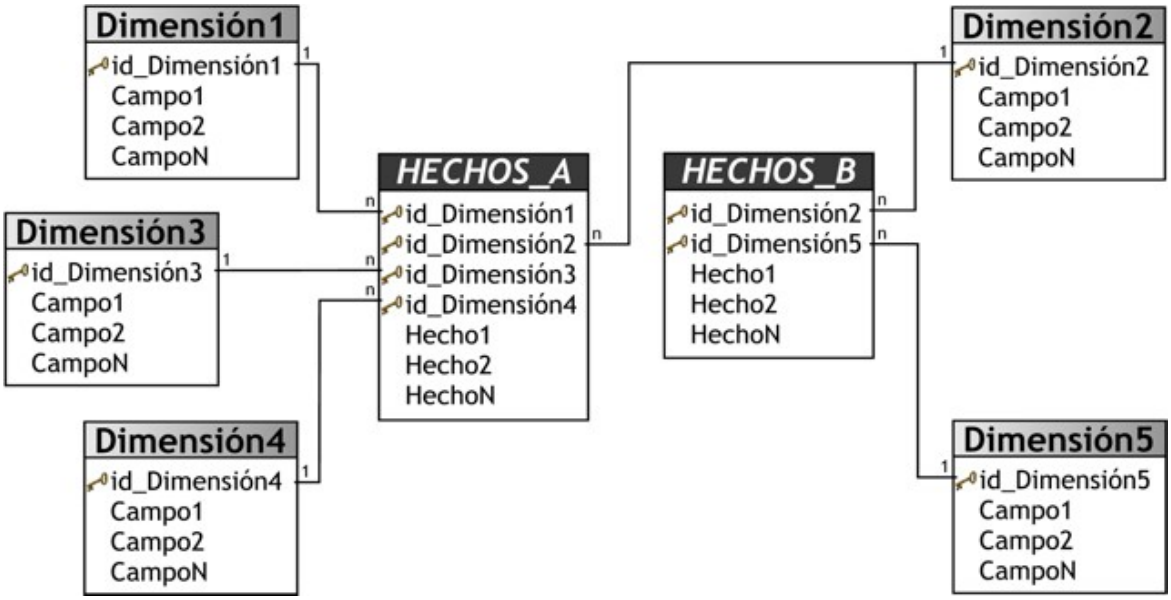

Figura 12: Esquema constelación

Este modelo se compone por un conjunto de de esquemas en estrella. Se compone de una tabla de hechos principal y una serie de tablas de hechos auxiliares. Dichas tablas están en el centro del modelo y están relacionadas con sus respectivas tablas de dimensiones.

Esquema estrella:

Este esquema esta conformado por una tabla de hechos central y de varias tablas dimensiones relacionadas a esta a través de sus respectivas claves. Figura 13

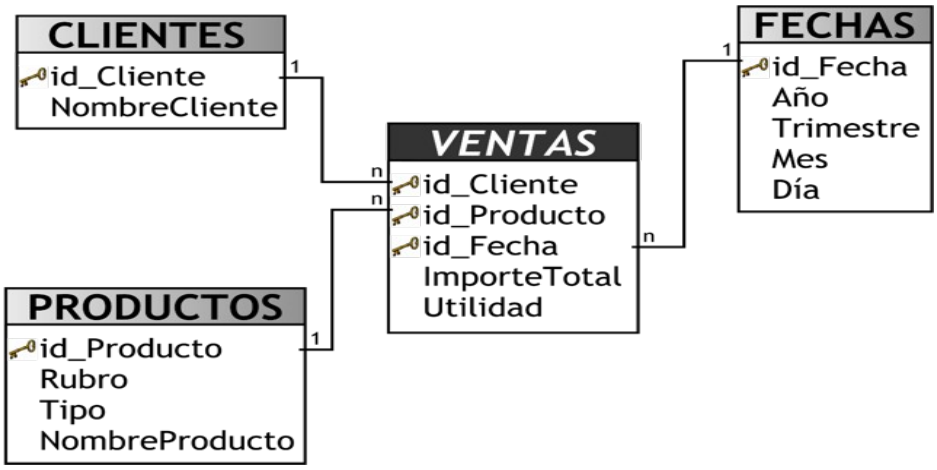

Figura 13: Esquema estrella

La **tabla de hechos** contiene los hechos o medidas que serán utilizados por los analistas de negocio para apoyar el proceso de toma de decisiones, donde los hechos son datos instantáneos en el tiempo. El esquema en estrella, consta de una tabla de hechos central y de varias tablas de dimensiones relacionadas a esta, a través de sus respectivas claves. El esquema en estrella es el más simple de interpretar y optimiza los tiempos de respuesta ante las consultas de los usuarios. Este modelo es soportado por casi todas las herramientas de consulta y análisis, y los meta-datos son fáciles de documentar y mantener. Figura 14

Figura 14: Tabla de hechos

1.3.10 Procesos OLAP, ROLAP, MOLAP, HOLAP.

OLAP:

El procesamiento analítico en línea (OLAP por sus siglas en inglés) es el motor de consultas especializado de los data warehouse. Las herramientas OLAP, son la tecnología de software para el análisis en línea, administración y ejecución de consultas que permiten predecir información del comportamiento del negocio. Su principal objetivo es brindar respuesta a preguntas complejas, para interpretar el comportamiento del negocio y tomar decisiones. Cabe destacar que lo que es realmente interesante en OLAP, no es la ejecución de simples consultas tradicionales, sino la posibilidad de utilizar operadores tales como **Drill Up**, **Drill Down**, para explotar profundamente la

información. Este tipo de herramientas, se puede analizar el negocio desde diferentes escenarios históricos, y proyectar como se ha venido comportando y evolucionando en un ambiente multidimensional, o sea, mediante la combinación de diferentes perspectivas, temas de interés o dimensiones. Esto permite deducir tendencias, por medio del descubrimiento de relaciones entre las perspectivas que a simple vista no se podrían encontrar sencillamente.

Drill-down: es ir de lo general a lo específico. Permite apreciar los datos en un mayor detalle, bajando por la jerarquía de una dimensión. Esto brinda la posibilidad de introducir un nuevo nivel o criterio de agregación en el análisis, disgregando los grupos actuales. Figura 15

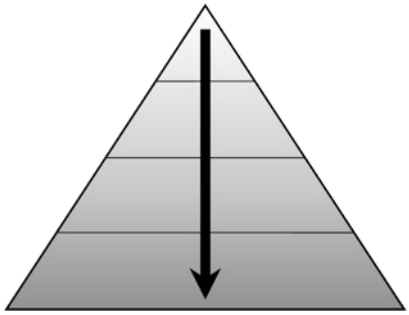

Figura 15: Drill-down

Drill-up: es ir de lo específico a lo general. Permite apreciar los datos en menor nivel de detalle, subiendo por la jerarquía de una dimensión. Esto brinda la posibilidad de quitar un nivel o criterio de agregación en el análisis, agregando los grupos actuales. Figura 16

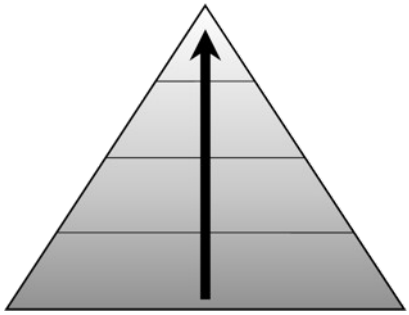

Figura 16: Drill-up

Las herramientas OLAP requieren que los datos estén organizados dentro del depósito en forma multidimensional, por lo cual es que utilizan los cubos multidimensionales. Además de las características ya descritas, se pueden enumerar las siguientes:

- Permite recolectar y organizar la información analítica necesaria para los
- usuarios y disponer de ella en diversos formatos, tales como tablas, gráficos, reportes, etc.
- Soporta análisis complejos de grandes volúmenes de datos.

- Complementa las actividades de otras herramientas que requieran procesamiento analítico en línea.
- Presenta al usuario una visión multidimensional de los datos (matricial).
- Es transparente al tipo de tecnología que soporta el DWH, ya sea ROLAP, MOLAP o HOLAP.
- Permite definir de forma flexible las dimensiones que se quieren analizar, sus restricciones, jerarquías y combinaciones.
- No tiene limitaciones con respecto al número máximo de dimensiones permitidas.
- Permite a los usuarios, analizar la información basándose en más criterios que un análisis de forma tradicional.
- Al contar con muestras grandes, se pueden explorar mejor los datos en busca de respuestas.
- Permiten realizar agregaciones y combinaciones de los datos de maneras complejas y específicas, con el fin de realizar análisis más estratégicos.

ROLAP:

El procesamiento relacional analítico en línea(ROLAP por sus siglas en inglés) es un tipo de organización física que se implementa sobre tecnología relacional. ROLAP cuenta con todos los beneficios de un Sistema Gestor de Bases de Datos Relacional al cual se le provee con extensiones y herramientas para poder ser utilizado como un Sistema Gestor de Data Warehouse. Esto tiene como inconveniente que es mucho más lenta que las demás estrategias de almacenaje.

MOLAP:

Esta estrategia, procesamiento multidimensional en línea(MOLAP por sus siglas en inglés), usa las bases de datos multidimensionales para proporcionar el análisis. Su principal precepto es que OLAP está mejor implementado almacenando los datos multidimensionalmente. El almacenaje de MOLAP provee un excelente rendimiento y compresión de los datos, así como el mejor tiempo de respuesta de todas las estrategias.

HOLAP:

La estrategia OLAP híbrida, ha sido desarrollada recientemente. Esta estrategia combina ROLAP y MOLAP para brindar una solución aprovechando las ventajas de cada una: mejor desempeño y gran escalabilidad. HOLAP mantiene los volúmenes de datos más grandes en la base de datos relacional y las agregaciones en un almacén MOLAP separado. Un cubo HOLAP es mas pequeño

que los almacenados con MOLAP y responden más rápido que los ROLAP. Generalmente esta estrategia se utiliza cuando es necesario una respuesta rápida de los datos pero usando un gran volumen de estos.

1.4 Modelo de datos.

Existe una diferencia notable entre los procesos que se realizan para la gestión de los datos en los sistemas operacionales y los sistemas Data Warehouse. Los sistemas operacionales realizan el procesamiento de transacciones en línea(OLTP por sus siglas en inglés) y los sistemas Data Warehouse siguen el modelo dimensional(ADD por sus siglas en inglés). Estas diferencias viene marcadas por la forma en que estos manipulan a los usuarios, contenido, estructura de datos, hardware, software, administración, etc. Pese a estas diferencias, se sigue utilizando su filosofía para el diseño de Data Warehouses. Ralph Kimball[9], expone que las ideas de diseño para el procesamiento de transacciones son inapropiados y en ocasiones destructivos con la información, por lo que propone una serie de técnicas denominadas modelo dimensional.

1.4.1 Consistencia y dimensión tiempo.

La consistencia de datos, ya sea en sistemas OLTP como ADD, es sumamente necesaria, pero se enfoca de manera diferente en ambos sistemas. En los sistemas OLTP la consistencia se garantiza a nivel transaccional, centrándose en no perder nada de estas. El modelo dimensional la consistencia se gestiona a nivel global, cuidando que la carga de datos nuevos sea un conjunto integrado y consistente.

Los sistemas OLTP y los sistemas ADD gestionan en tiempo de forma diferente. Los sistemas OLTP reciben la denominación de parpadeantes, ya que la información contenida dentro de los mismos cambia constantemente, imposibilitando que se pueda hacer una instantánea o Snapshot de los datos. Los sistemas ADD los datos sufren pocas transformaciones cíclicas, por lo que es fácil representar un extracto de datos productivos en diferentes instantes de tiempo.

1.4.2 El modelo entidad-relación.

Kimball[9] plantea que el modelo entidad-relación constituyen un desastre para las consultas, dado que este modelo no puede ser comprendido por los usuarios y no puede ser útilmente recorridos por los sistemas gestores de bases de datos. El modelo entidad-relación persigue la meta de la eliminación de redundancia de los datos, logrando así que las transacciones solo alteren los datos en un punto específico, mejorando en rendimiento de la base de datos. El modelos entidad-relación tiene una gran simetría, por lo que no permite diferenciar que tablas son mas importantes o cual contiene medidas numéricas de los objetos del negocio.

1.4.3 El modelo dimensional.

El modelo dimensional es el nombre adoptado por una vieja técnica que permite hacer bases de datos comprensibles. Este modelo puede ser visualizado como un cubo con 3 o mas dimensiones, donde cualquier punto interior es la intersección de las coordenadas definidas por los ejes del cubo[9]. el cubo seria la unidad encargada del almacenamiento de la información, de forma equivalente a las tablas de las bases de datos relacionales. Estos cubos representan la información como matrices, las que reciben el nombre de dimensiones y representan criterios de análisis. A los datos almacenados en la matriz, se les llama medidas y representan los indicadores o valores a analizar.

Este modelo se caracteriza por su sencillez, permitiéndole a los usuarios finales una fácil comprensión de la base de datos y a los sistemas gestores de bases de datos un recorrido eficiente de sus estructuras. Este modelo es conocido como esquema estrella, por su semejanza con la misma, dado que se compone de una tabla principal nombrada hechos, y varias tablas alrededor de esta, nombradas dimensiones.

1.4.4 Modelado de datos.

Los datos, tanto en los sistemas OLTP como en los sistemas ADD tiene que pasar por 3 fases de modelado: conceptual, lógico y físico. Las diferencias entre los Data Warehouse con las bases de datos operacionales en cuanto al tipo de consultas y al rendimiento esperado, hacen que las estrategias de diseño y modelo de datos sean diferentes. El **modelo conceptual** captura la esencia del problema, refleja la información fundamental acerca de las entidades del dominio del problema y sus relaciones. El **modelo lógico** describe los datos detalladamente, incluye todas las entidades, sus relaciones, atributos, llaves, tipos de datos. El modelo lógico sirve de puente entre el modelo conceptual y el modelo conceptual y el modelo físico. El modelo físico describe las estructuras de almacenamiento y los métodos usados para el acceso efectivo a los datos.

1.6 Estado actual de los Data Warehouse.

La forma de los usuarios de percibir y gestionar el conocimiento no siempre se corresponde con el modelo que proponen los sistemas operacionales , conocido como procesamiento transaccional en línea. Este modelo genera una serie de datos, los cuales son gestionados por bases de datos operacionales. Es un hecho innegable la importancia para la vida diaria de las bases de datos operacionales, pero estas nunca son diseñadas para proporcionar funciones de síntesis, análisis consolidación de datos. El estado del arte actual centra los procesos del negocio en el cliente, por lo que se realizan esfuerzos tecnológicos para reenfocar esto.

1.6.1 Sistemas Data Warehouse en el mundo.

Las empresas y entidades comerciales a nivel mundial necesitan cada vez más un manejo ágil de la información para mantenerse competitivas en el mercado. En el plano comercial minorista los Data Warehouse se usan para predecir la cantidad de un determinado producto, que se venderá a determinado precio. 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.

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, Delta, Lufthansa, QANTAS, British Airways, American Airlines, Canadian Airlines.

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.

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.

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 Data Warehouse en Cuba.

Nuestro país no ha quedado rezagado en el desarrollo de almacenes de datos. Ya se cuenta con el desarrollo de varios de estos, el 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.

Existencia en CUBAENERGIA de un Datawarehouse para coleccionar toda la documentación concerniente al sistema de información de CUBAENERGIA soportado básicamente en sus sistemas operacionales y aplicaciones en explotación. Los hechos modelados fueron: Ventas, Compras y Gestión de Recursos Humanos por Competencias y por Tiempo.

1.7 Metodologías de diseño de los sistemas Data Warehouse.

En la actualidad existen varias metodologías para llevar a cabo el diseño y construcción de un Data Warehouse. La mayoría propuestas no han sido aceptadas como modelos estándares para el modelado dimensional, ya que no cubren todas las etapas o transformaciones necesarias.

Dos de las metodologías más completas para el diseño de un Data Warehouse son la metedología Hefestos[14] y la DWEP(Proceso de Ingeniería de un Data Warehouse, por sus siglas e inglés), esta última propuesta en la tesis de Sergio Luján-Mora[15].

Hefestos propone la construcción de un Data Warehouse de forma metódica y sencilla, guiada por las necesidades de los usuarios. Utiliza modelos conceptuales lógicos, los cuales son sencillos de interpretar y analizar. Esta metología es independiente del ciclo de vida, de las herramientas y de las estructuras físicas que contengan el Data Warehouse.

DWEP es una metodología orientada a objetos, independiente de cualquier implementación especifica, ya sea multidimensional, relacional, etc. Esta metodología esta basada en UML, que es el Lenguaje Unificado de Modelado[16] y RUP[17], que es el proceso unificado de desarrollo de software.

1.7.1 Justificación de la metodología escogida.

Para el diseño del Data Mart para el Modulo Control de Personas del Sistema para la Gestión Integral Aduanera, se utilizará DWEP[15], siguiendo el modelo dimensional propuesto por Kimball. Teniendo como principal ventaja el empleo de la misma notación (basada en UML) para el diseño de los diferentes diagramas y las correspondientes transformaciones de una manera integrada. Como UML es un lenguaje de modelado general se utilizan sus mecanismos de extensión para adaptarlo al dominio específico de los almacenes de datos.

El método DWEP propone la estructuración del almacén de datos en cinco etapas y tres niveles[15].

Etapas:

- Origen: Define los orígenes de datos del almacén de datos, como los sistemas OLTP,

fuentes de datos externas, etc.

- Integración: Define el mapeo entre los orígenes de datos y el propio almacén de datos.
- Almacén de Datos: Define la estructura del almacén de Datos.
- Adaptación: Define el mapeo entre el almacén de datos y las estructuras empleadas por el cliente.
- Cliente: Define las estructuras concretas que son empleadas por los clientes para acceder al almacén de datos, como Data Marts o aplicaciones OLAP.

Niveles:

- Conceptual: Define el almacén de datos desde un punto de vista conceptual, es decir, desde el mayor nivel de abstracción y contiene únicamente los objetos y relaciones mas importantes.
- Lógico: Abarca aspectos lógicos del diseño del almacén de datos, como la definición de las tablas y claves, la definición de los procesos ETL (Extraction, Transformation and Loading), etc.
- Físico: Define los aspectos físicos del almacén de datos, como el almacenamiento de las estructuras lógicas en diferentes discos o la configuración de los servidores de bases de datos que mantienen el almacén de datos.

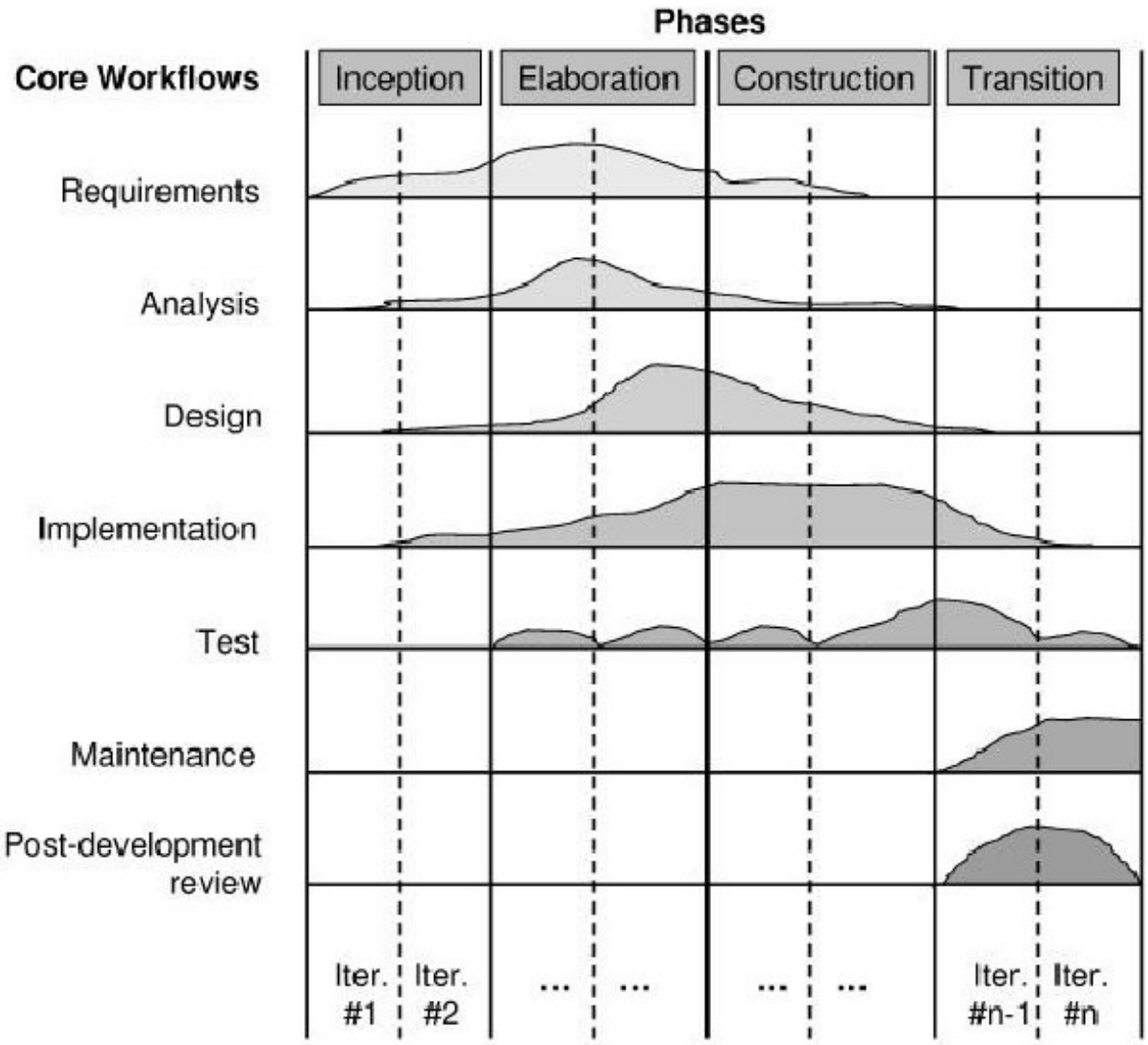

Figura 17: Flujos de la metodología DWEP

Como el DWEP es una instanciación de RUP para el desarrollo de almacenes de datos establece al igual que este, el ciclo de vida de un proyecto en cuatro fases: Inicio, Elaboración, Construcción y Transición, así como cinco flujos de trabajo fundamentales: Requerimientos, Análisis y Diseño, Implementación y Prueba [18], además adiciona dos nuevas actividades: Mantenimiento y Revisión Pos-Desarrollo.Figura 17

En cada flujo de trabajo se utilizan diferentes diagramas UML, para modelar y documentar el proceso de desarrollo.

1.8 Herramientas en usabilidad.

Actualmente se encuentran en uso numerosas herramientas dedicadas al diseño, implementación y mantenimiento de almacenes de datos. Oracle, el gigante de las compañías informáticas, lleva la

delantera en dichas herramientas. Con el producto Oracle Warehouse Builder, que constituye una herramienta multiplataforma, incorporan herramientas de administración, integración y calidad de datos, además de cumplir con todo el ciclo de vida de un Data Warehouse. 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.

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).

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 interfaces gráficas para seleccionar, arrastrar y pegar.

1.8.1 Justificación de las herramientas seleccionadas.

Control de Personas, de el área Lucha contra el Fraude, de la Aduana General de la República, asume la necesidad de la implementación de un Data Mart para la prestación de servicios de consultas y reportes de manera optimizada. Estas consultas y reportes tributarán a el reconocimiento de patrones de riesgo en el flujo de viajeros por frontera. Le herramienta seleccionada debe cumplir con los siguientes requisitos:

- Alta compatibilidad con el gestor de base de datos utilizado por el Sistema de Gestión Integral Aduanera(GINA), para el cual es diseñado el Data Mart. La Aduana General de la República utiliza como gestor de base de datos Oracle.
- Posibilidad de almacenar y manipular cubos multidimensionales de información.
- Alta compatibilidad con las herramientas ETL(Extracción, Transformación y Carga, por sus siglas en inglés).
- Capacidad de almacenar y gestionar un gran volumen de información.
- Alta capacidad de gestión de dimensiones, en sus distintas formas y estructuras de

concepción.

- Soporte de funciones matemáticas y estadísticas para mejorar y dinamizar el servicio de reportes y consultas.
- Eficiente gestión de seguridad.
- Rendimiento eficiente.
- Posibilidad de integración con disímiles fuentes de datos.

Atendiendo a todas estos requerimientos, y acorde con la política de la Aduana General de la República de Cuba y el Consejo de Estado y de Ministros, de promover una cultura libre, extendiendo el uso de herramientas de código abierto. Teniendo esto como principal premisa se hace uso de los siguientes herramientas:

- Visual Paradigm for UML v6.1: para modelado de clases y objetos[19].
- SQL Power Architect: diseño lógico de bases de datos[20].
- Umbrello: diseño de diagramas UML[21].
- PostgreSQL Server v8.4: Sistema Gestor de Bases de Datos[22].
- PgAdmin3: herramienta utilizada como Sistema de Administración de Bases de Datos[23].
- Schema Workbech: herramienta para el diseño lógico de el Data Mart[24].
- Pentaho Mondrian: herramienta utilizada para el procesamiento analítico en linea(OLAP) [24].

1.9 Conclusiones.

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. Situación actual de Control de Personas.

En el presente capítulo se explican detalladamente las necesidades de información adicional que se requieren para realizar el Control de Personas en el área Lucha Contra el Fraude. Se describen los procesos fundamentales, ya que constituyen la entrada principal para el diseño de el Data Mart. Se hace una panorámica además de los puntos de contacto que tiene el área LCF con otros negocios como API(Información Adelantada de Pasajeros) y con Inmigración, ya que constituyen fuentes externas de datos.

Se realiza un estudio además sobre el mapa conceptual y el modelo de objetos de el módulo Control de Personas. Se realiza un breve resumen sobre los aspectos generales y la estructura informativa-computacional del módulo Control de Personas del GINA.

2.1 Aspectos generales

2.1.1 Descripción

El módulo Control de Personas de el subsistema LCF del GINA se constituye como el área principal para el diagnóstico y vigilancia de las personas de interés aduanal(PIA). Desde el año 2007 la Aduana recibe información adelantada de pasajeros(API), además de nutrirse con información proveniente de otros órganos de seguridad de el Estado, como el MININT. El volumen de informaciones a procesar sugiere la necesidad de automatizar estos procesos.

Esta área cuenta actualmente con un software que reúne tres funcionalidades principales:

- Monitoriza las personas al pasar por frontera.
- Gestionar los PIA.
- Realizar estudios del API.

Además este software permite a las jefaturas de la Aduana, monitorizar aeropuertos determinados y realizar estudios API definidos por el usuario. Este sistema permite configurar a los PIA y acceder a los datos históricos de estas personas, así como establecer posibles redes de vínculos entre los viajeros.

2.1.2 Objetivos

Control de Personas es la encargada de llevar a cabo de forma efectiva el control de personas naturales por frontera. Esta es una de las responsabilidades de mayor peso de el sistema GINA, específicamente de el subsistema LCF.

Actualmente, el subsistema LCF tiene asociado un módulo dedicado a la gestión de los procesos

relacionados al control de personas, nombrado Control de Personas. Este módulo fue diseñado e implementado centrándose en las deficiencias que existían en la Aduana para efectuar este tipo de control. El software se encuentra actualmente en explotación, realizando las tres funcionalidades mencionadas anteriormente.

Sin embargo, este sistema no cumple con todas las funcionalidades que se requieren en el área Control de Personas: análisis estratégicos, algoritmos de minería de datos, inteligencia de negocio; requieren de rendimiento y de sistemas de información que no están implementados aún.

2.2 Estructura informativa-computacional

Logísticamente se cuenta con tecnología para realizar un cluster en PostgreSQL, que funcionaría como el sistema de almacenamiento externo. Este cluster estaría conformado por dos nodos, capaces de acceder simultáneamente a la base de datos y a los servicios del cluster. Los nodos estarían conectados entre sí por Ethernet a 1Gb/s. Estos nodos tendrán una interfaz de red dedicada solamente a la comunicación entre ellos y otra para dar servicio de base de datos y para acceder a los sistemas de comunicación.

Los nodos poseerán un mínimo de 1Gb de RAM y el almacenamiento suficiente para el Sistema Operativo y el software de PostgreSQL. El cluster estará implementado sobre Ubuntu Server 10.04 LTS, usando la versión PostgreSQL 8.4. Accederán un mínimo de 5 computadoras cliente, también con Ubuntu como sistema operativo, aunque se cuenta con recursos para que accedan más computadoras cliente.

2.3 Características de el sistema.

2.3.1 Análisis de los procesos de Control de Personas.

En el Control de Personas se identificó un proceso, Controlar Persona, compuesto por tres subprocesos: Controlar Viajeros, Estudiar API y Gestionar Persona. De manera general, el proceso Controlar Persona permite realizar análisis de vuelos, para conocer el número de taquilla por donde hará salida la persona de interés aduanal, también se insertan los resultados de las personas que cometieron alguna infracción o indicio, y se gestionan las personas que son de interés aduanal.

El subproceso Controlar Viajeros permite realizar actividades de control sobre las personas que entran y salen del país, además de registrar toda la información recogida de las personas infractoras, posibilitando el estudio y análisis.

El subproceso Estudiar API analiza y selecciona las personas de un vuelo que se considere

importante aplicarle actividades de control para descubrir un indicio o una infracción. Además se seleccionan las categorías que tienen los pasajeros ante la aduana.

El subproceso Gestionar Persona gestiona las personas de interés aduanal, para realizar estudios y control a los viajeros cuando pasen por frontera.

2.4 Definición de el alcance.

Control de Personas, a pesar de que cuenta con un software que le ayuda en la automatización de todo su proceso de negocio tiene necesidades de información, por lo que se infiere la necesidad de la explotación de un sistema informacional. Antes de comenzar a diseñar un sistema informacional, se necesita un plan para su desarrollo. Las entradas criticas para este plan de desarrollo son los requerimientos de información y las prioridades de el negocio donde se vaya a desplegar el sistema informacional.

2.4.1 Alcance.

Definir el alcance de el sistema informacional es la primera tarea, ya que centra las expectativas, guía y prioriza el desarrollo incremental, identifica riesgos y ayuda en el proceso de estimación de costos.

Los alcances de el sistema informacional para Control de Personas serán los siguientes:

- Procesos de control de personas naturales por frontera.
- Proceso de identificación de infracciones, Modus operandis recurrentes, productos ilícitos, tanto en la entrada o salida de personas naturales de el país.
- Proceso de identificación de aduanas recurrentes en la identificación de hechos ilícitos.
- Proceso de identificación de las categorías recurrentes de las personas naturales.
- Proceso de identificación de los países de las personas naturales de los cuales y hacia los cuales ocurren mas hechos ilícitos.
- Proceso de identificación de las líneas de enfrentamiento y las medidas mas efectivas.

2.4.2 Necesidades de información.

Para realizar un control de personas efectivo es de vital interés definir las necesidades de información reales y potenciales, para afrontar la tarea de lucha contra los hechos ilícitos. Destacar que en los estudios de necesidades de información en ocasiones pueden proporcionar mas datos de los que se vayan a utilizar.

Se realizó un análisis de las necesidades de información para controlar las personas, centrado en los datos que son necesarios a los analistas de LCF y en los datos que actualmente les provee el software que tienen en explotación. Como resultado de dicho análisis se arribo a la conclusión de que el módulo Control de Personas necesita algún mecanismo automatizado y normalizado que le de respuesta a los siguientes objetivos:

- Identificación unívoca de los patrones de riesgo.
- Interoperabilidad entre todos los sistemas de información.
- Integración con los sistemas actuales de información, manteniendo sus peculiaridades.
- Conocimiento global y particular de el negocio de Control de Personas.
- Consistencia de la información.
- Establecimiento de estándares de intercambio de información.
- Poseer un núcleo o repositorio central de información estratégica.
- Inclusión de nuevos servicios de manejo de los datos, tanto tácticos como estratégicos.
- Minimizar la carga de conexiones de datos.

2.4.3 Requisitos.

El sistema informacional a implantar en el negocio de Control de Personas deberá cumplir con ciertos requisitos o prioridades, que deben ser respetados durante las fases de desarrollo del mismo.

1) Orientación al proceso: El sistema informacional deberá centrarse principalmente en el área Control de Personas, de forma general, y no a partes o elementos de la misma. La solución será de carácter global y no parcial.
2) Los datos se cargarán una sola vez: La captura de los datos a procesar dentro de el sistema informacional se harán en el origen, sin intermediarios ni cambio de soportes que haga peligrar la integridad de los datos.
3) Descentralización de la información: Permitirá la carga, búsqueda y difusión de la información en cualquier momento y a cualquier usuario que esté acreditado y registrado en el sistema.
4) Centralización de la definición de la información: Definición de criterios unificados de los diferentes elementos informacionales (entidades, tablas, funciones, entradas y salidas de datos, etc).

5) Peculiaridades organizativas: Definición de una herramienta que permita configurar niveles de compartimentación de la información y de usuario.
6) Seguridad: Protección de accesos no autorizados. Autenticidad, confidencialidad, integridad y protección de la calidad de los datos.
7) Arquitectura: La arquitectura de la información debe adaptarse a la naturaleza táctica o estratégica de la información.

2.5 Aproximación de la solución.

Control de Personas necesita a diario mas servicios en línea y más accesibilidad a la información, con el objetivo de obtener el mayor rendimiento en los procesos de negocio. No se puede concebir un trabajo eficiente en el área Control de Personas sin un sistema de información eficiente que lo soporte, y que implique directamente en el modelo de gestión, constituyendo una estrategia fundamental.

El sistema informacional a poner en práctica en Control de Personas permitirá realizar proyecciones futuras y servirá de soporte para el proceso de toma de decisiones, ya que:

- Soportará el manejo de las funciones claves de el negocio de Control de Personas.
- Estará orientado a aplicaciones con un patrón de uso predefinido.
- Influirá de manera directa con los cambios constantes de los datos en información, y de la información en conocimiento.
- Abarcará desde los datos globales hasta los datos detallados.

Para dar respuesta a las necesidades informativas de Control de Personas y teniendo en cuenta los requisitos de el sistema informacional que se necesita; se determinó el diseño de un Data Mart. Para lograr la integración de este tipo de sistema, se contará con un repositorio de datos preparado para tal fin. Este repositorio se creará siguiendo las características de un Data Warehouse. Para llevar adelante el diseño de el sistema, se utilizará la metodología propuesta DWEP(Data Warehouse Engineering Process, por sus siglas en inglés).

2.6 Validando la solución.

Para validar la solución propuesta se utilizó la metodología propuesta en el trabajo de tesis de Erika Muñoz Leiva[25]. Esta metodología de validación se proponen 10 criterios por los cuales regirse para revisar la propuesta metodológica de diseño y hacer un estudio de factibilidad de implantación de el Data Mart. Estos criterios de evalúan de forma cuantitativa, asignándole valores

entre 0 y 100, atendiendo a el nivel alcanzado en el desarrollo de el proyecto.

- Organización
- Planificación.
- Grado de entendimiento con el usuario.
- Gestión de limpieza de datos.
- Consistencia lógica en la metodología
- Gestión de modelado.
- Gestión de diseño.
- Gestión de pruebas.
- Gestión de uso.
- Análisis de valor agregado.

Siguiendo esta metodología se concluye que es factible el diseño de el Data Mart, centrando como actividades priorizadas el Grado de entendimiento con el usuario, Gestión de limpieza de datos y Gestión de pruebas.

2.7 Análisis de los requerimientos e indicadores.

El primer paso consiste en obtener los datos que son objetos de análisis en cuanto al tema de control de personas, lo cual se puede llevar a cabo por diferentes vías: entrevistas, cuestionarios u observaciones,etc. Aquí se persigue el objetivo de identificar la información clave que guiará el proceso de análisis, dicho proceso es esencial para la identificación de patrones de riesgo y para trazar metas estratégicas relativas a el cruce de personas naturales por frontera.

A partir de observaciones y teniendo en cuenta el análisis realizado se determinó que se desea conocer los siguiente:

- Infracciones en las aduanas por línea de enfrentamiento y su ocurrencia en el tiempo, desglosadas por las modalidades de enfrentamiento en el tiempo.
- Productos de las infracciones y su ocurrencia en el tiempo, desglosados por los modus operandis y las técnicas de detección.
- Países y su ocurrencia en el tiempo, desglosados por las líneas de enfrentamiento y las modalidades de enfrentamiento.

- Países de los cuales entran más personas por frontera, desglosados por las aduanas, además una relación porcentual de las personas que han presentado infracciones y cuales han sido controles en rangos de tiempo.
- Controles realizados a las personas desglosadas en aduanas, rango de tiempo, categorías, líneas de enfrentamiento y modalidades.
- Controles realizados a las personas desglosadas en países, rango de tiempo, categorías, líneas de enfrentamiento y operaciones.

Estos requerimientos identificados permitirán a los analistas de LCF conocer la cantidad de cruces por frontera que presentan infracciones en determinadas fechas y por cuales Aduanas, para así canalizar la fuerza de enfrentamiento hacia esas Aduanas. Además permitirá conocer que infracciones ocurren con más frecuencia en determinadas Aduanas para trazar estrategias sobre las técnicas de detección de infracciones, así como de potenciar y capacitar el personal más en unas regiones del país que en otras según el histórico de los hechos delictivos. Permitirá conocer desde y hacia cuales países y en que época de el año ocurren más infracciones, así como los modos operandis de los infractores y los productos de las infracciones.

2.8 Análisis de los OLTP.

Otro de los aspectos a tener en cuenta para el diseño de el Data Mart es el análisis de los sistemas de procesamiento transaccional en línea(OLTP). En el caso de Control de Personas, toda la información se gestiona en un schema de la base de datos implementada en Oracle, sobre el cual se sustenta GINA. El schema perteneciente a LCF, recibe el mismo nombre(ver Anexo # X). A continuación se representa un fragmento de dicho schema, donde se representan las tablas puntuales para el Data Mart.

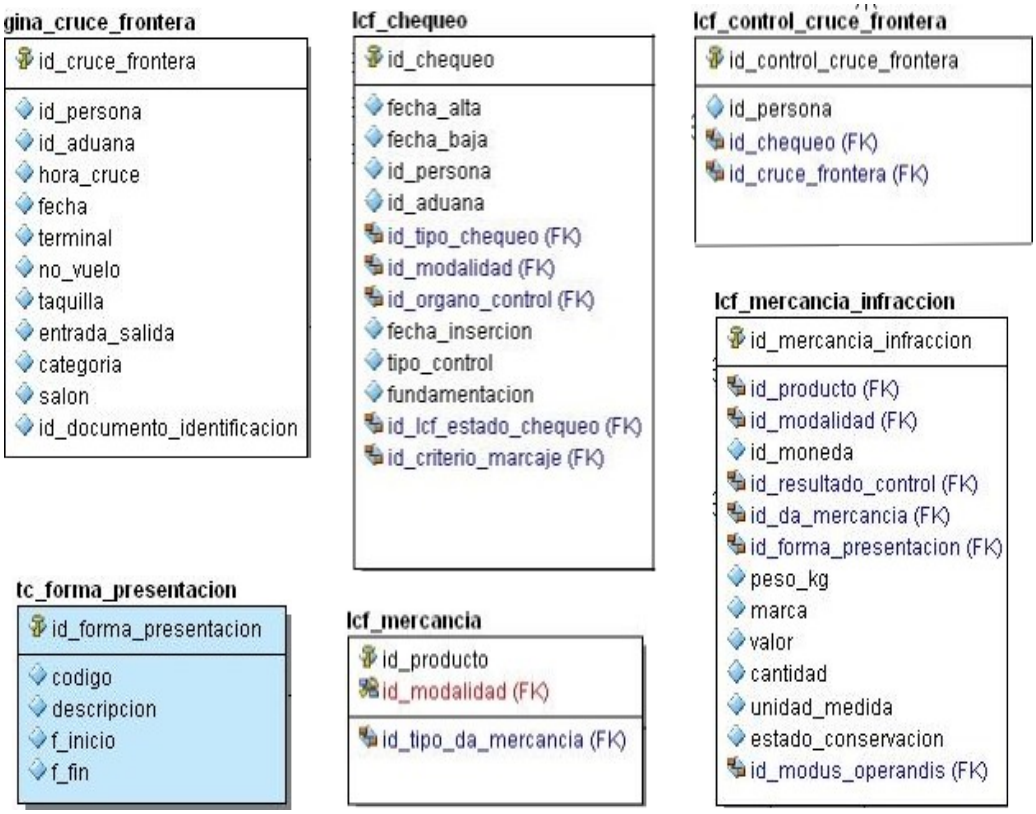

Figura 18: Fragmento de el schema LCF(1)

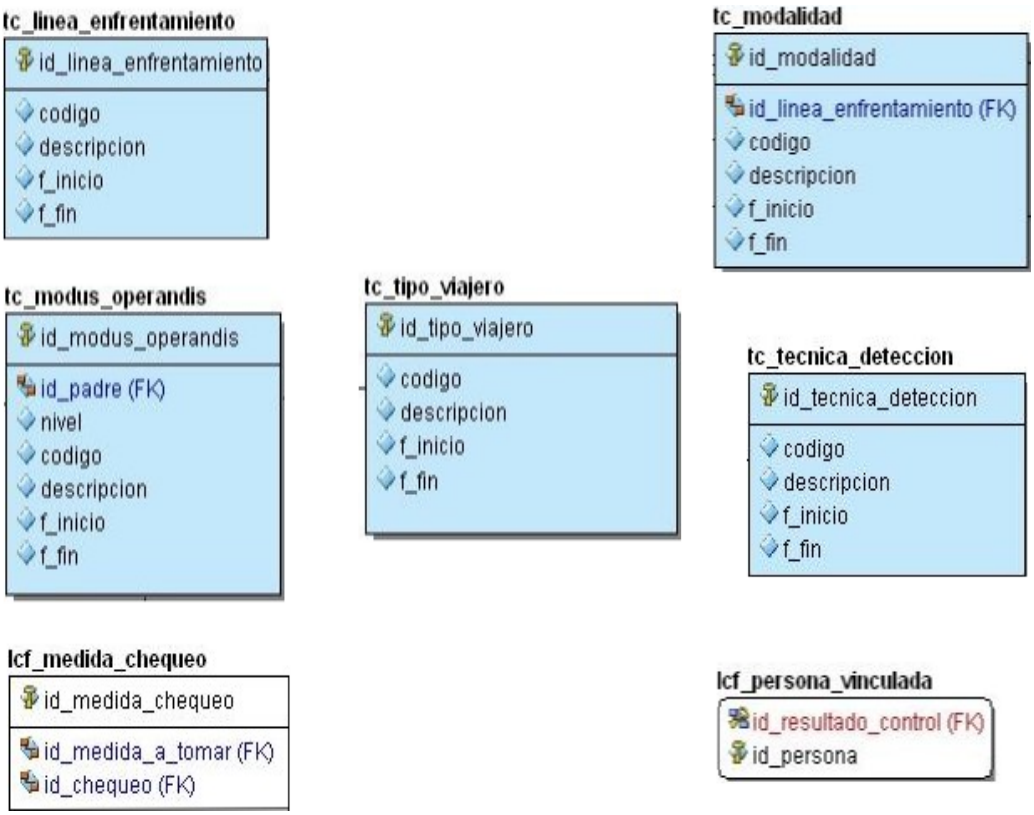

Figura 19: Fragmento de el schema LCF(2)

2.9 Conclusiones.

Capítulo 3 Diseño de el Data Mart.

En el presente capítulo se realiza en diseño de el Data Mart, así como la descripción de el ciclo de vida del mismo representado en cinco fases, como establece la metodología DWEP. Se realizan además los diagramas de los tres niveles: conceptual, lógico y físico.

3.1 Descripción de el Data Mart.

Se desea construir un Data Mart, adjunto al Módulo Control de Personas del subsistema Lucha Contra el Fraude(LCF), perteneciente a el Sistema para la Gestión Integral Aduanera(GINA). Este Data Mart posibilitará el análisis y reconocimiento de patrones de riesgo en el cruce de personas naturales por frontera.

El Data Mart tendrá tres enfoques de exploración de los datos: Infracciones, Cruces por Frontera y Controles.

De las Infracciones se almacenará las Técnicas de Detección, los Productos, Modus Operandis, País de procedencia de la Persona, así como la Línea de Enfrentamiento. De los Cruce por Frontera se almacenará el País de la persona y las Operaciones. De los Controles se almacenará las Medidas, las Líneas de Enfrentamiento y sus Modalidades, el País de la persona, la Categoría y la Operación.

Todos estos datos estarán acotados por espacios de tiempo, así como por las Aduanas, para garantizar el máximo nivel de granularidad.

3.2 Proceso de construcción de un Data Mart.

De manera general, los procesos de construcción de un Data Mart son diseñar el esquema, construir el almacenamiento físico, popular el Data Mart con datos provenientes de las fuentes externas, comenzar su explotación para obtener informes y la administración temporal de el mismo[15]. A continuación se desglosan los procesos.

- Diseño: este proceso cubre todas las tareas de iniciar la solicitud de creación de un Data Mart a través de recopilación de información sobre los requisitos. En este paso se modelan además los modelos lógico y físico de el Data Mart. En este paso se deben cumplir las siguientes tareas:
    1. Obtención de los requerimientos funcionales y no funcionales.
    2. Identificar las fuentes de datos.
    3. Selección de el subconjunto apropiado de datos.

4. Diseñar la estructura lógica y física de el Data Mart.

- Construcción: este proceso incluye la creación de la base de datos física y las estructuras lógicas asociadas con el Data Mart, para proveer un acceso rápido y eficiente a los datos. En este proceso se deben cumplir las siguientes tareas:
    1. Creación de la base de datos física y las estructuras de almacenamiento.
    2. Creación de los objetos de el esquema, como tablas e índices, definidos en el proceso de Diseño.
    3. Determinación de la mejor vía para establecer las tablas y las estructuras de acceso.
- Llenado: el proceso de llenado cubre todas las tareas relacionadas a obtener los datos de las fuentes, el proceso de limpieza de dichos datos, modificación de los datos hasta el nivel de detalle deseado. En este proceso se deben cumplir las siguientes tareas:
    1. Mapeo de las fuentes de datos hasta las estructuras de datos de el Data Mart.
    2. Extracción de los datos.
    3. Limpieza y transformación de los datos.
    4. Carga de los datos dentro de el Data Mart.
    5. Creación y almacenamiento de los meta-datos.
- Acceso: el proceso de acceso cubre las tareas de poner los datos aptos para su uso, consultarlos, analizarlos, crear reportes, crear gráficos, así como la publicación de estos. De manera general se usa una herramienta gráfica para hacer las consultas a la base de datos y visualizarlas. En este proceso se deben cumplir las siguientes tareas:
    1. Establecer una capa intermedia para la herramienta visual a utilizarse.
    2. Mantener y administrar estas interfaces de negocio.
    3. Establecer y administrar estructuras de bases de datos, como tablas sumarizadas.
- Administración: este paso cubre todas las tareas de administración de el Data Mart a través de su ciclo de vida. En este proceso se deben cumplir las siguientes tareas:
    1. Proveer acceso seguro a los datos.
    2. Administración de el crecimiento de los datos.
    3. Optimización de el sistema para un mejor desempeño.

4. Aseguramiento de la disponibilidad y seguridad de los datos en caso de fallos de el sistema.

3.3 Descripción de la solución.

El Data Mart a diseñar para el módulo Control de Personas cumplirá con las siguientes características:

- Modelo dimensional propuesto por Ralph Kimball, con la variante de modelamiento Esquema estrella.
- Para la arquitectura conceptual de los datos se utilizará la de tres capas.
- Para el proceso analítico en linea se utilizará la variante MOLAP.
- Metodología de desarrollo DWEP.

3.4 Aplicación de la metodología DWEP.

La metodología DWEP es un método global para el diseño e implementación de todas las fases de un almacén de datos, incluyendo las fuentes de datos operacionales, los procesos ETL y el propio esquema de el almacén de datos. Se representarán cada fase de el ciclo de vida con sus respectivos diagramas.

Para el diseño de los diagramas se hace uso de un Perfil UML para Modelación Dimensional, que permite modelar las principales propiedades multidimensionales de un almacén de datos. Este perfil es creado por Sergio Luján Mora y Juan Trujillo, el cual permite llevar a cabo el modelado del DWH utilizando la metodología DWEP[26].

3.4.1 Requerimientos.

En este flujo de trabajo se define el alcance de el Data Mart. Los requerimientos son modelados usando Casos de Uso. Una vez que los requerimientos son definidos, el proyecto Data Mart es establecido y asignado los diferentes roles.

3.4.1.1 Requerimientos funcionales.

- El sistema permitirá a los usuarios la presentación de la información por los criterios pre-establecidos.
- El sistema permitirá alternar las filas y las columnas, correspondientes a las dimensiones en la matriz de datos.
- El sistema permitirá activar o desactivar dimensiones en las consultas.

- El sistema tendrá que extraer, transformar y cargar datos de los sistemas operacionales.

3.4.1.2 Requerimientos no funcionales.

- Rendimiento:

1. El tiempo medio de respuesta a consultas simples echas al Data Mart deberá ser de no más de 1 minuto y para las consultas complejas (que son las que implican más de 30 millones de registros), podría ser de 1 a 3 minutos.
2. El sistema deberá soportar hasta 100 usuarios activos.
3. El sistema deberá soportar hasta 200 accesos a la misma vez.

- Seguridad.

1. El sistema está definido para operar conjuntamente con el sistema del control del acceso y garantizar de esta forma el acceso a los datos solamente a las personas debidamente autorizadas.

- Accesibilidad.

1. El Data Mart debe estar disponible para todos los usuarios autorizados las 24 horas del día durante los 7 días de la semana.

- Precisión de los datos.

1. El sistema tendrá que proceder a la actualización periódica de los datos de los proveedores de las fuentes, dentro de los períodos establecidos entre el cliente y equipo de desarrollo.

- Usabilidad.

1. El Data Mart deberá ser sencillo, flexible y de fácil uso.

3.4.1.3 Actores de el sistema.

Actores de el sistema	Justificación
Analista LCF	Encargado de consultar los datos de el Data Mart para llevar a cabo el proceso de toma de decisiones.
Desarrollador ETL	Encargado de extraer los datos de los sistemas OLTP, transformarlos y cargarlos en el Data

	Mart.
Desarrollador OLAP	Responsable de desarrollar los cubos OLAP.
Grupo de calidad de datos.	Encargado de asegurar la exactitud y calidad de los datos.
Sistema LCF.	Sistema externo que contiene la base de datos operacional LCF, la cual provee de datos a el Data Mart. Además contendrá una función automática para la carga de los datos.

3.4.1.4 Diagramas de Casos de Uso.

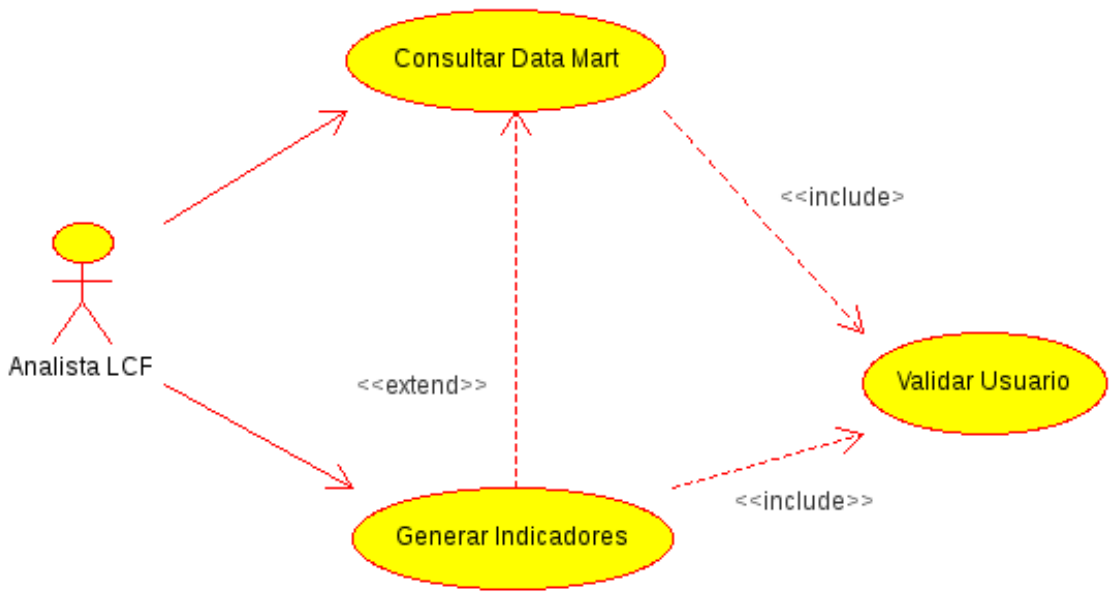

Figura 20: Caso de Uso Consultar Data Mart

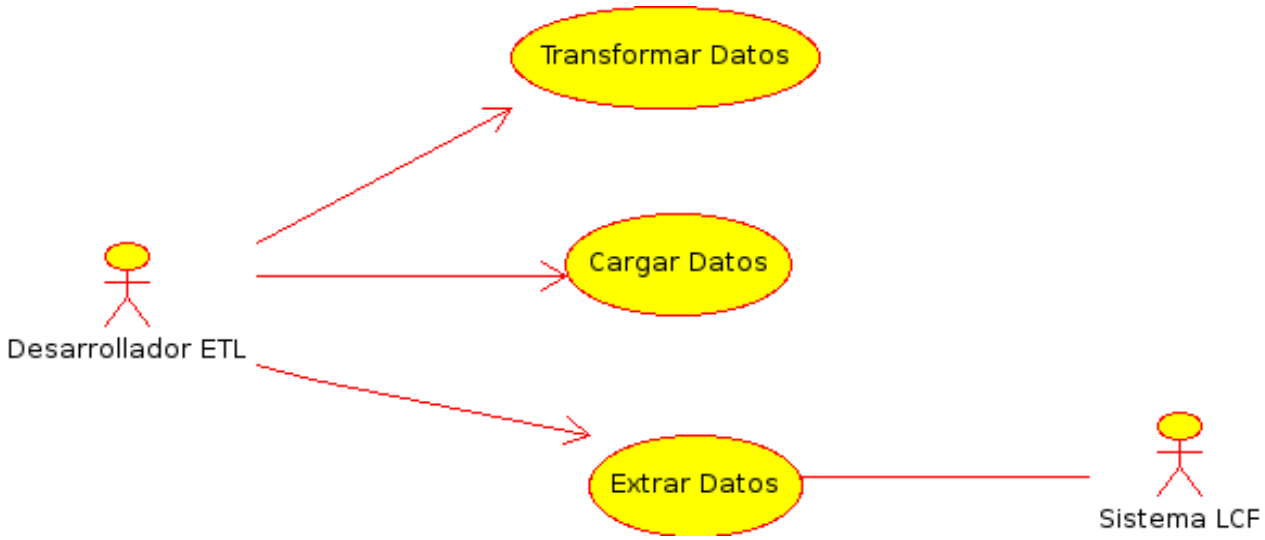

Figura 21: Caso de Uso Proceso ETL

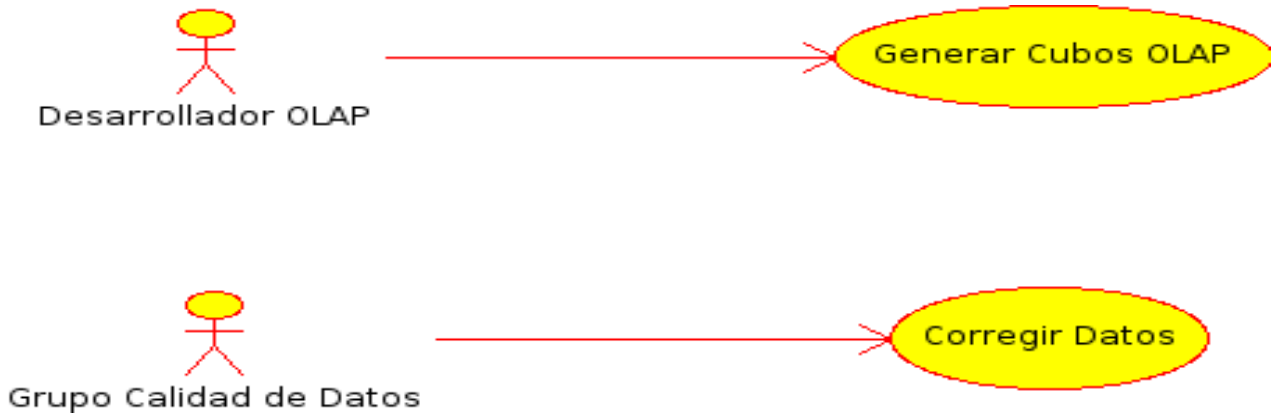

Figura 22: Casos de Uso Procesos OLAP y Proceso Corrección de Datos

3.4.2 Análisis.

El objetivo de este flujo es refinar y estructurar los requerimientos obtenidos en el flujo anterior, así como definir las fuentes de datos operacionales y externas que nutrirán los datos de el Data Mart. En este flujo de trabajo se generan los artefactos esquema conceptual de la fuente y esquema lógico de la fuente.[15]

Para realizar el llenado de el Data Mart de Control de Personas se parte de una base de datos operacional, el esquema LCF, perteneciente al sistema GINA, la cual se encuentra normalizada y almacena datos sobre las personas de interés aduanal(PIA), las técnicas de detección de hechos delictivos, los productos de dichos hechos delictivos, los modus operandis utilizados en los hechos delictivos, las fechas en que estos se han realizado, la aduana donde se han detectado, y si se

han cometido en la entrada o salida de el país.

3.4.2.1 Esquema conceptual de la fuente.

El esquema conceptual de la fuente representa la base de datos operacional que proveerá de datos a el Data Mart. Para ellos se representan las entidades mas importantes y las relaciones entre ellas. Ver Figura 23.

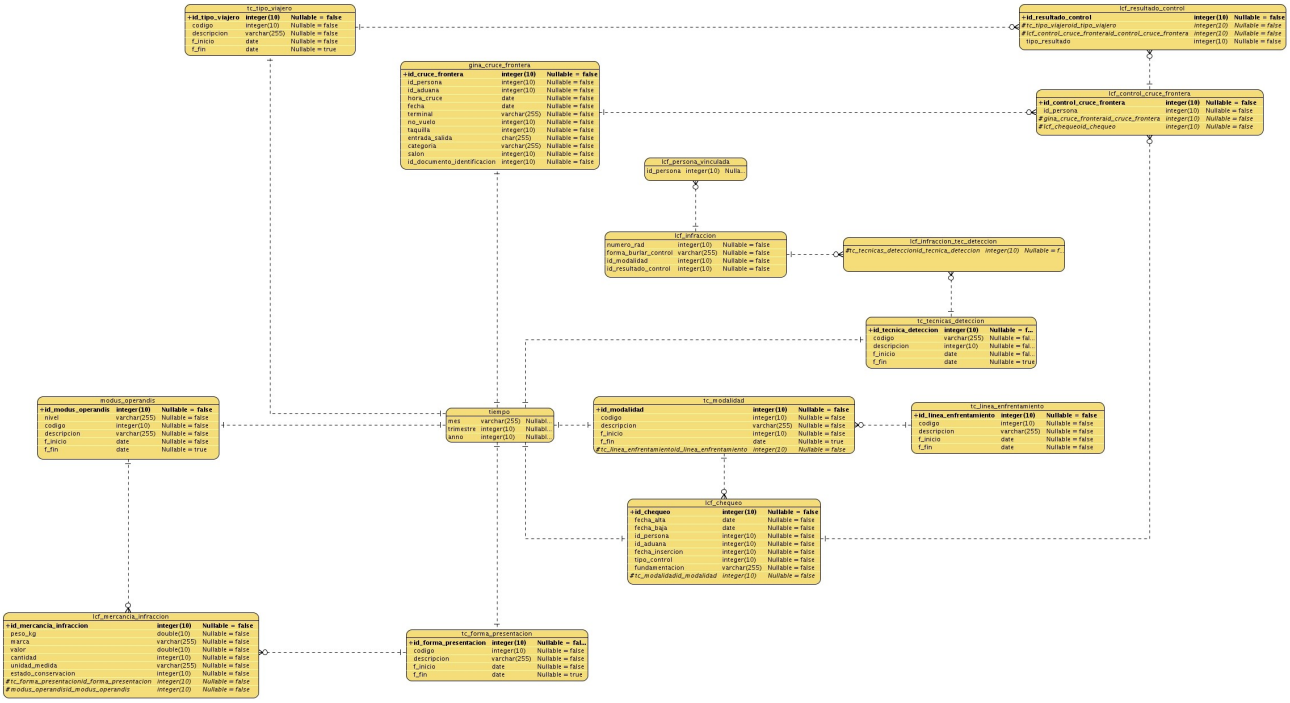

Figura 23: Esquema conceptual de la fuente(SCS)

3.4.2.2 Esquema lógico de la fuente.

Para realizar el esquema lógico de la fuente se toma como entrada en esquema conceptual de la fuente, para ello las clases son convertidas a tablas, los atributos a columnas y las asociaciones a relaciones.

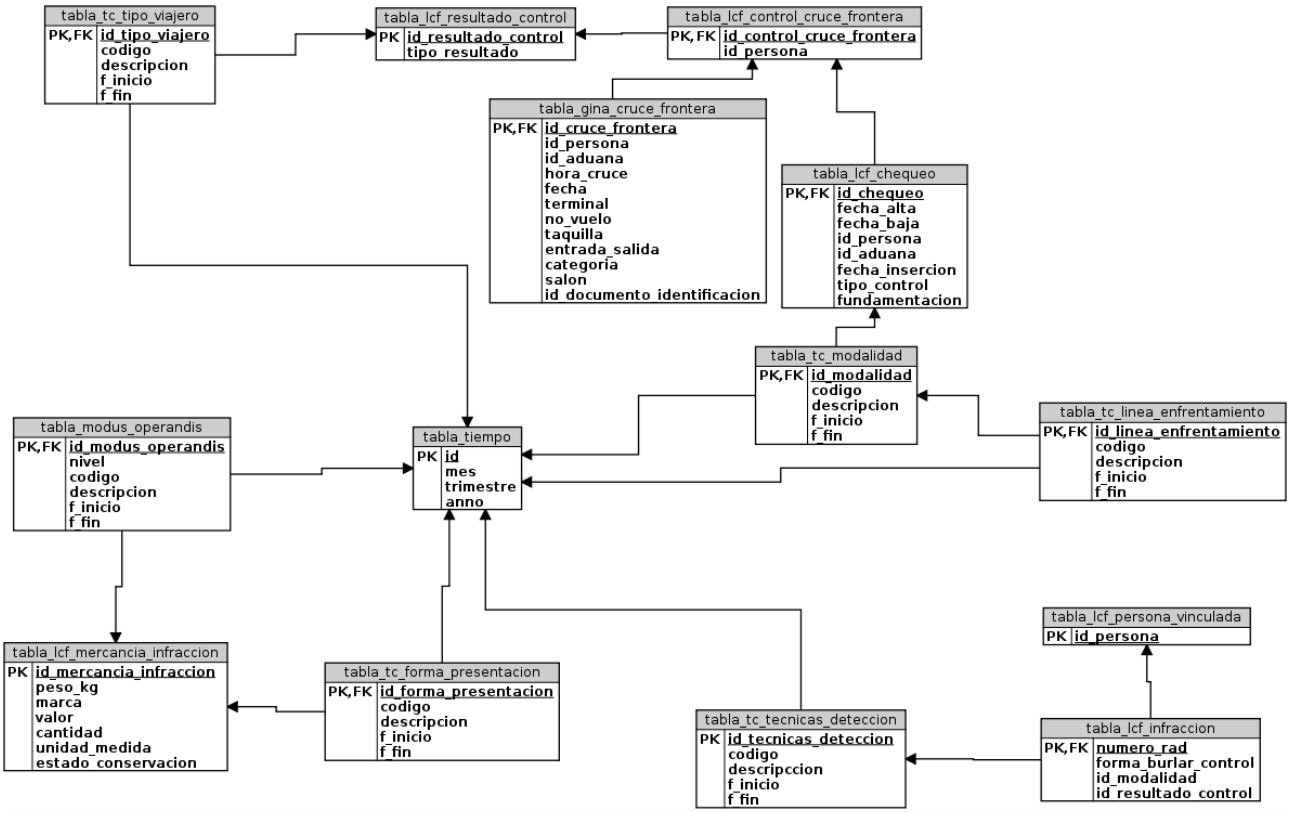

Figura 24: Esquema lógico de la fuente(SLS)

3.4.3 Diseño.

En este flujo de trabajo se construye el esquema conceptual de el Data Mart(DWCS). Para realizar el esquema conceptual existen dos estrategias: *top-down*, que recomienda la construcción de el Data Warehouse primero y después la de los Data Marts y la estrategia *bottom-up*, que recomienda la creación de los Data Mart y luego la unión de estos en un Data Warehouse. En el caso particular de Control de Personas se seguirá la estrategia *buttom-up*, ya que, acorde con la definición de Kimball, provee un mayor índice de éxito y se pueden apreciar los resultados con mayor rapidez.

También en este flujo se determinan los procesos ETL, los cuales se definen como mapeo de datos entre las fuentes(esquema conceptual de la fuente) y el almacén de datos(esquema conceptual de el Data Warehouse), definiéndose el esquema de mapeo de datos(Data Mapping).

Como se expuso en el Capítulo 1, se utilizará el enfoque de Ralph Kimball, que propone el modelado dimensional de datos, donde se estructura la información en hechos y dimensiones.

3.4.3.1 Esquema conceptual de el Data Mart.

El diseño conceptual tiene como objetivo la construcción abstracta y completa de el problema, se

seleccionan los objetos y criterios relevantes para el proceso de toma de decisiones, y se modelan estos como dimensiones y/o medidas.

En la metodología propuesta por Luján-Mora 2005[15], se divide este proceso en tres niveles:

- Nivel 1: Definición del Modelo: Un Paquete representa un esquema estrella de un modelo multidimensional. En este nivel, una dependencia entre dos paquetes indica que los esquemas estrellas comparten al menos una dimensión.
- Nivel 2: Definición de un esquema estrella. Un paquete representa un hecho o una dimensión de un esquema estrella. En este nivel, una dependencia entre dos paquetes de dimensión indica que las dimensiones comparten al menos un nivel en sus correspondientes jerarquías.
- Nivel 3: Definición de un hecho o una dimensión. Se compone de un conjunto de clases que representan los niveles jerárquicos en un paquete de dimensión o el esquema estrella completo en el caso de un paquete de hecho.

El Data Mart para Control de Personas estará compuesto por tres cubos de datos:

- **Cubo Infracciones:**

  Tabla de hechos: Tfact_Infracciones.

  Tablas de dimensiones: Tdim_TecnicasDeteccion, Tdim_Producto, Tdim_ModusOperandis, Tdim_Paises, Tdim_Tiempo, Tdim_LineaEnfrentamiento, Tdim_Modalidad y Tdim_Aduana.

- **Cubo Cruces Frontera:**

  Tabla de hechos: Tfact_CrucesFrontera.

  Tablas de dimensiones: Tdim_Aduana, Tdim_Tiempo, Tdim_Paises y Tdim_Operaciones.

- **Cubo Controles:**

  Tabla de hechos: Tfact_Controles.

  Tablas de dimensiones: Tdim_Aduana, Tdim_Operacion, Tdim_Categoria, Tdim_Paises, Tdim_LineaEnfrentamiento, Tdim_Modalidad, Tdim_Medidas y Tdim_Tiempo.

En la siguiente figura se representa el Nivel 1, representando los cubos de datos que conformarán el Data Mart. Las flechas denotan dimensiones comunes entre los cubos.Figura 25

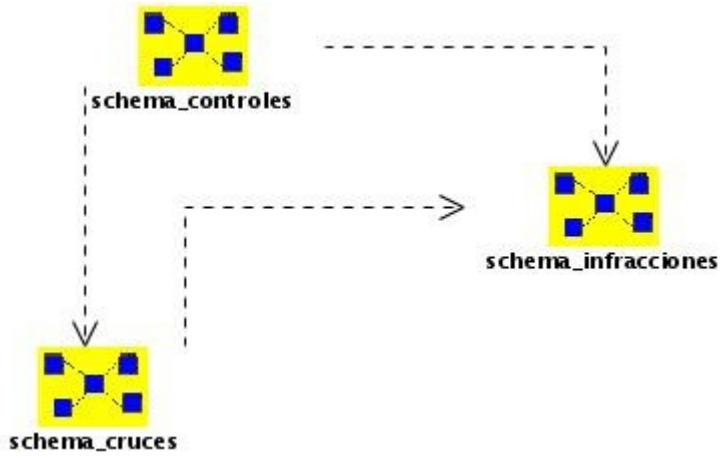

Figura 25: Esquema conceptual de el Data Mart (DWCS)

3.4.3.2 Modelo estrella para Infracciones.

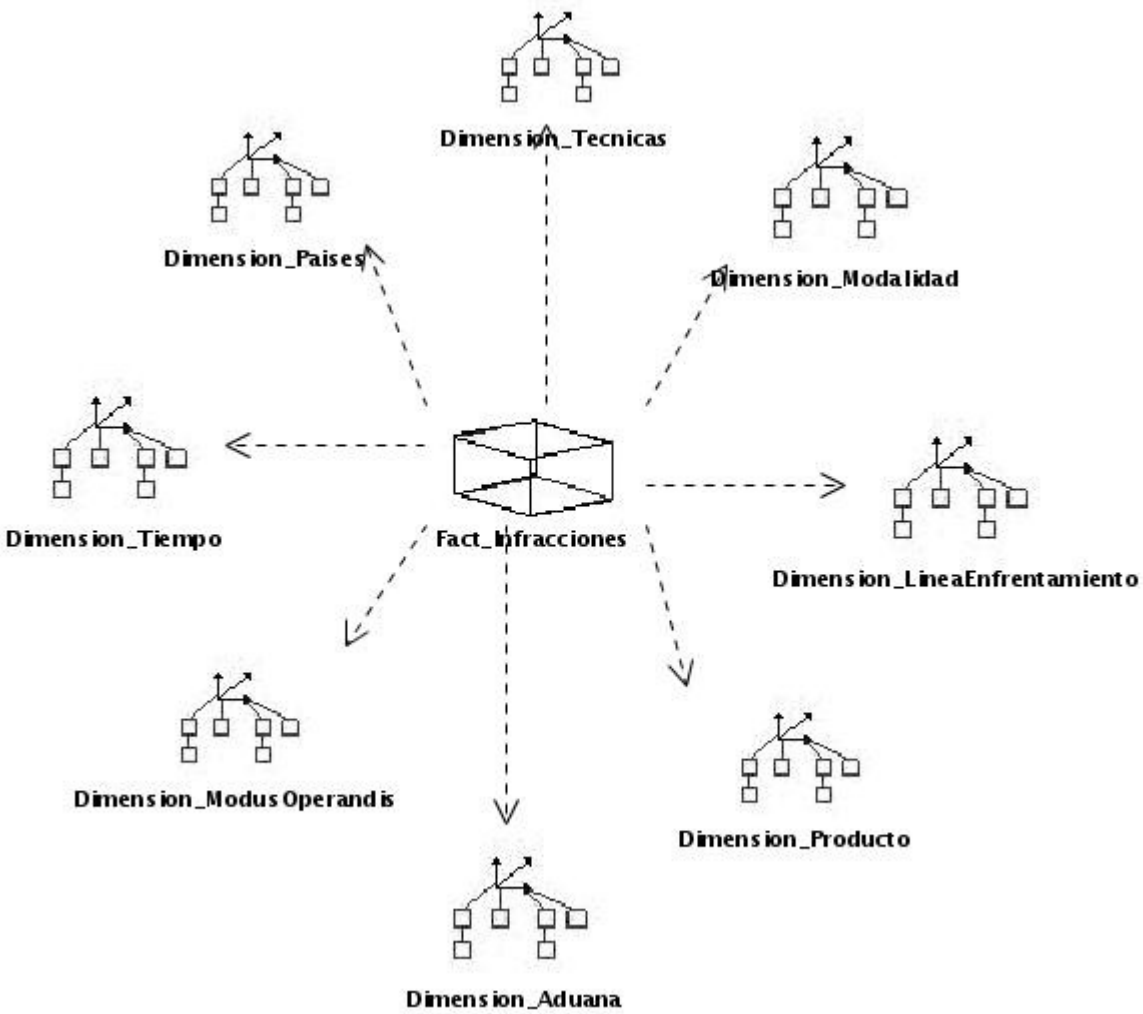

Figura 26: Esquema conceptual de el DWH. Schema Infracciones (Nivel 2)

Este esquema en estrella permite obtener información sobre cada una de las infracciones, así como los productos y las técnicas de detección asociadas a dichas infracciones; los países de donde provienen las personas que incurren en las infracciones y la aduana donde fueron detectadas. Además brinda un enfoque de las técnicas de detección y las modalidades de enfrentamiento.

3.4.3.3 Modelo estrella para Cruces por Frontera.

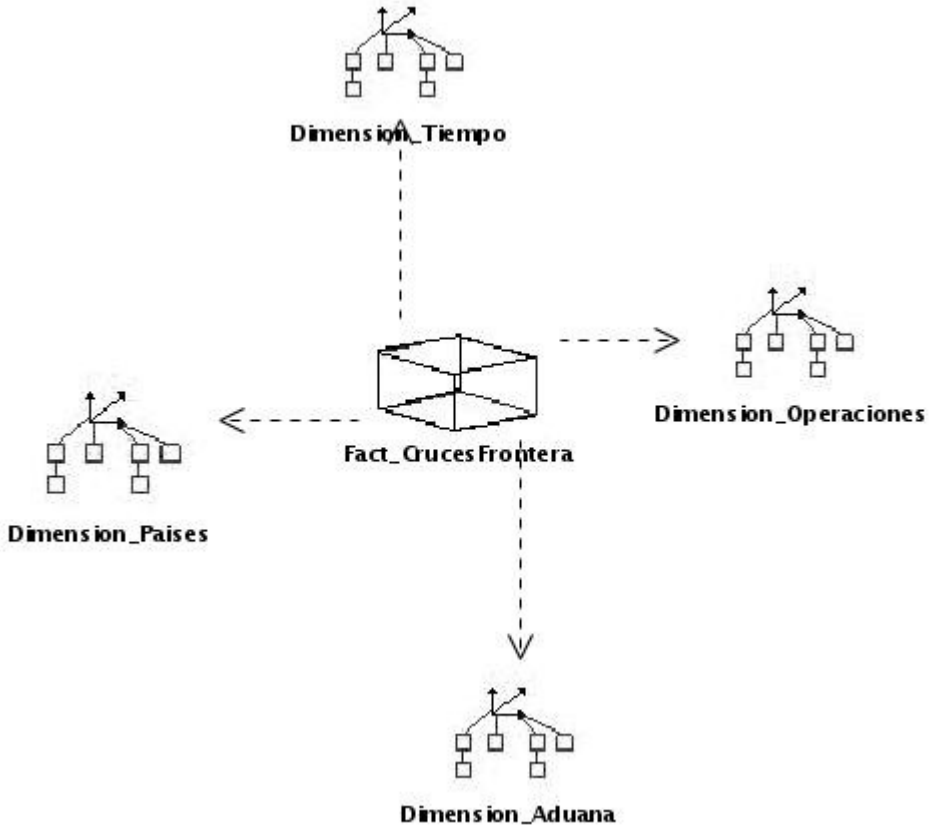

Figura 27: Esquema conceptual de el DWH. Schema CrucesFrontera(Nivel 2)

Este esquema en estrella permite obtener información sobre cada uno de las cruces por frontera, así como las operaciones(entrada o salida de el país) y las aduanas por donde se registran los cruces por frontera; acotados en rangos de tiempo.

3.4.3.4 Modelo estrella para Controles.

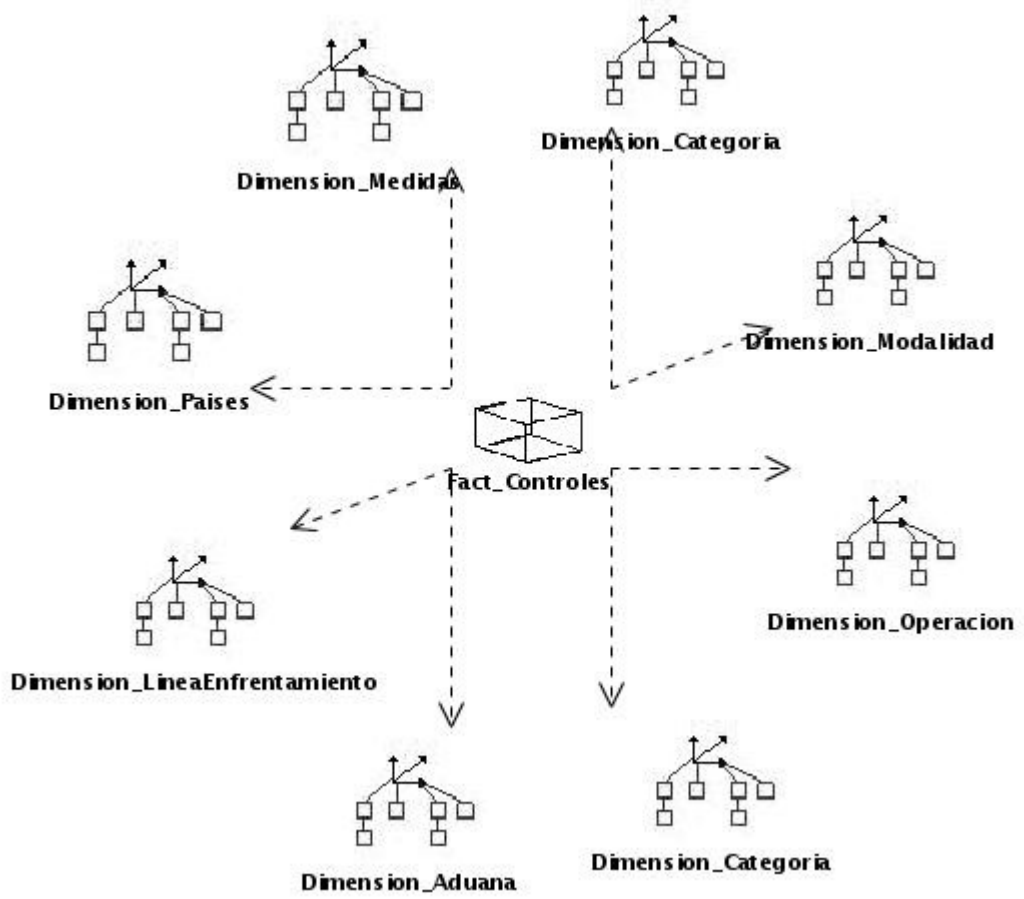

Figura 28: Esquema conceptual de el DWH. Schema Controles(Nivel 2)

Este esquema en estrella permite obtener información sobre cada una de los controles, así como las medidas, modalidades, líneas de enfrentamiento, operaciones(entrada o salida de el país) asociadas a los controles. Todos estos datos se presentarán acotados en rangos de tiempo.

3.4.4 Hechos de el modelo de datos lógico.

3.4.4.1 Hecho Cruces por Frontera.

En la Figura 29 se muestra el contenido de el paquete hecho Fact_CrucesFrontera, perteneciente al Nivel 3, en el cual la clase hecho es definida, y se representan además las dimensiones con sus respectivas jerarquías.

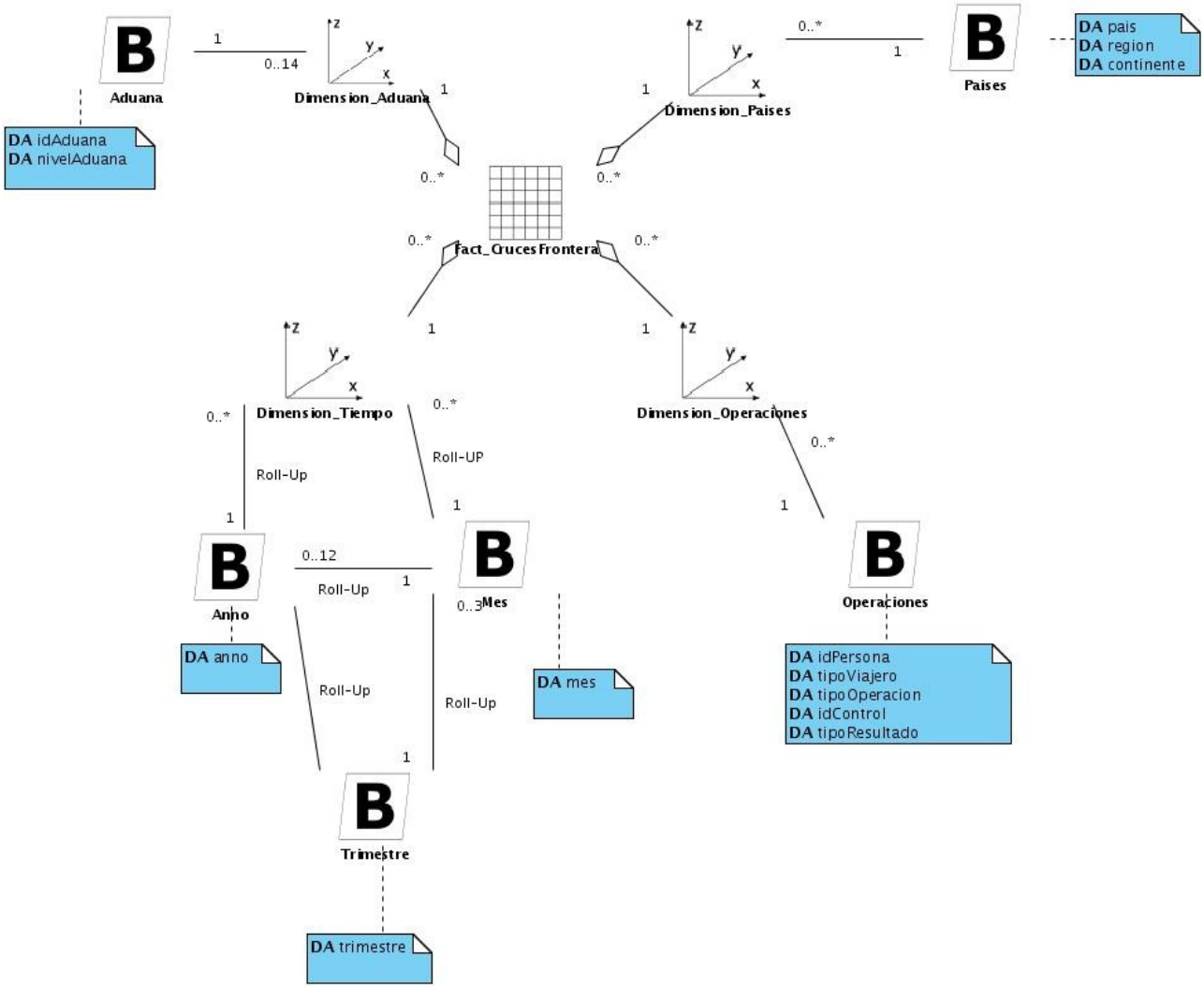

Figura 29: Esquema conceptual de el DWH. Fact_CrucesFrontera(Nivel 2)

En los Anexos se exponen el contenido de los restantes hechos:

1. Fact_Infracciones.
2. Fact_Controles.

3.4.5 Dimensiones de el modelo de datos lógico.

3.4.5.1 Dimensión Producto.

En la Figura 30 se muestra el contenido de el paquete Dim_Producto, perteneciente al Nivel 3, que contiene la definición de la clase <<Dimensión>> Producto y los diferentes niveles de jerarquías que son representadas por clases <<Base>>, estos niveles de jerarquías definen como las diferentes operaciones OLAP (roll-up, drill-down) pueden ser aplicadas.

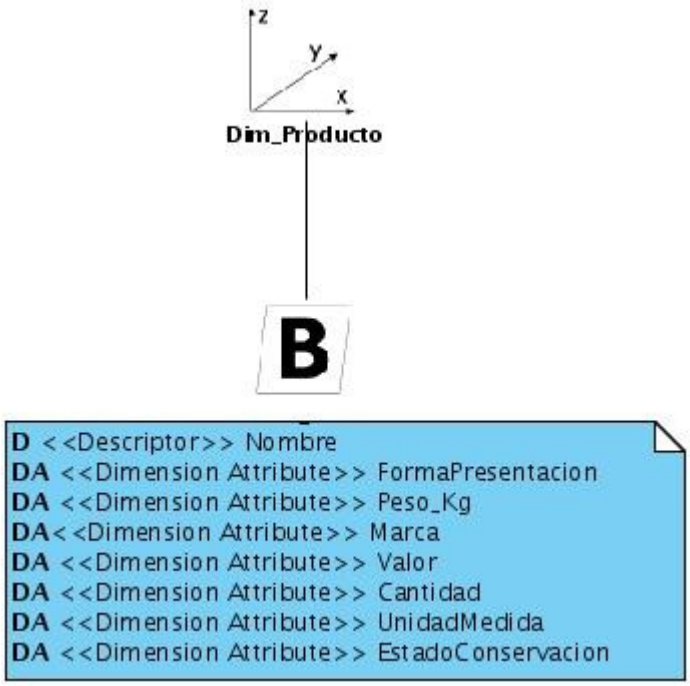

Figura 30: Esquema conceptual de el DWH.
Dimension_Producto(Nivel 3)

En los Anexos se exponen el contenido de las restantes dimensiones:

1. Dimension_Paises
2. Dimension_Tecnicas
3. Dimension_Modalidad
4. Dimension_LineaEnfrentamiento
5. Dimension_Aduana.
6. Dimension_ModusOperandis
7. Dimension_Tiempo
8. Dimension_Medida
9. Dimension_Operacion
10. Dimension_Categoria

3.4.6 Mapeo de datos.

El proceso de mapeo de datos constituye la consolidación y limpieza de la información a utilizarse, e incluye las actividades de extracción, transformación y carga de los datos, denominado proceso

ETL.

Las transformaciones aplicadas a los datos provenientes de las fuentes externas son transformaciones básicas de limpieza y estructuración. Estas transformaciones son necesarias para garantizar la calidad de los datos de el Data Mart e incluye corrección de errores, eliminación de redundancia, cambios de formato, agregaciones, etc.

Para representar el flujo de datos desde las fuentes externas hacia el Data Mart se utiliza el diagrama de mapeo de datos(Data Mapping), propuesto por Lujan-Mora [15]. Estos diagramas de mapeos, como en ocasiones pueden tornarse muy complejos, el autor propone dividirlos en cuatro niveles:

- Nivel de base de datos o Nivel 0: En este nivel cada esquema del almacén de datos se representa mediante un paquete. Los mapeos entre los diferentes esquemas se modelan en un único paquete de mapeo, que encapsula todos los detalles.
- Nivel de flujo de datos o Nivel1: Este nivel describe las relaciones de datos a nivel individual entre las fuentes de datos hacia los respectivos destinos en el almacén de datos.
- Nivel de tabla o Nivel2: Mientras que el diagrama de mapeo en el nivel 1 describe las relaciones entre las fuentes y los destinos de datos mediante un único paquete, el diagrama de mapeo de datos en el nivel de tabla detalla todas las transformaciones intermedias que tienen lugar durante ese flujo.
- Nivel de atributo o Nivel 3: En este nivel, el diagrama de mapeo de datos captura los mapeos existentes a nivel de atributo.

En la Figura 31 se representa el Nivel 0 de el mapeo de datos, representado por un paquete llamado CPMapping, relacionado con el Esquema Conceptual de la Fuente(SCS por sus siglas en inglés) y el Esquema Conceptual del Data Mart(DWCS por sus siglas en inglés).

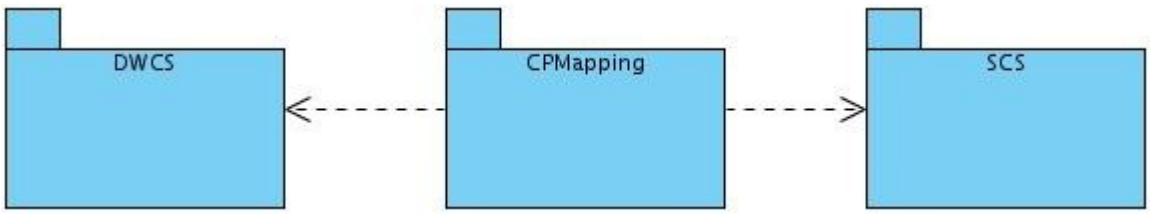

Figura 31: Mapeo de datos(Nivel 0)

El Data Mart para Control de Personas posee catorce tablas que se quieren llenar, once de dimensiones y tres de hechos, por lo que existen nueve escenarios dentro de el paquete CPMapping, una para cada uno de las tablas de las dimensiones; a excepción de las tablas

dimensión Producto y la tabla dimensión Modus Operandis que se llenan de forma simultánea, al igual que las tablas LineaEnfrentamiento y Modalidad. El nivel 1 representa el flujo de datos entre la fuente de datos y el destino de los datos, en el contexto de cada escenario, representado en la Figura 32.

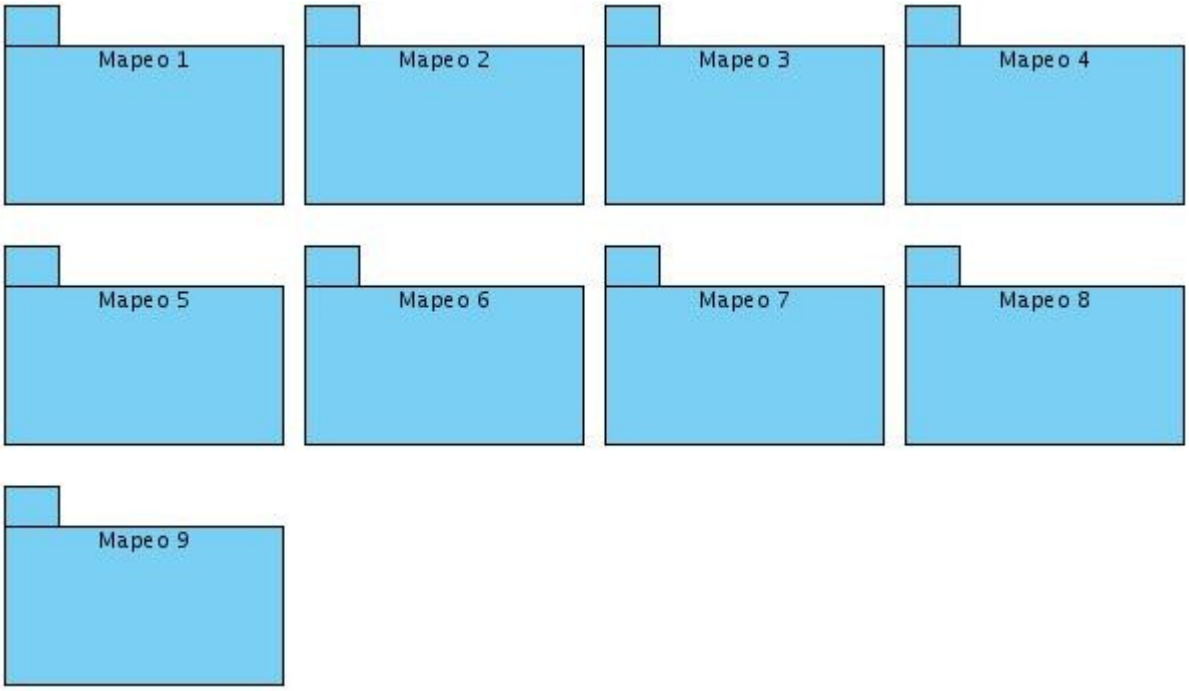

Figura 32: Mapeo de datos(Nivel 0)

En la Figura 33 se muestran las 3 transformaciones que sufren los datos provenientes de la SCS, correspondiente al Nivel 2, de las dimensiones ModusOperandis y Producto. Asimismo, en la Figura 34 se presenta el mapeo a nivel de atributo, correspondiente al Nivel 3, de la dimensión ModusOperandis. En los Anexos se muestran los restantes mapeos

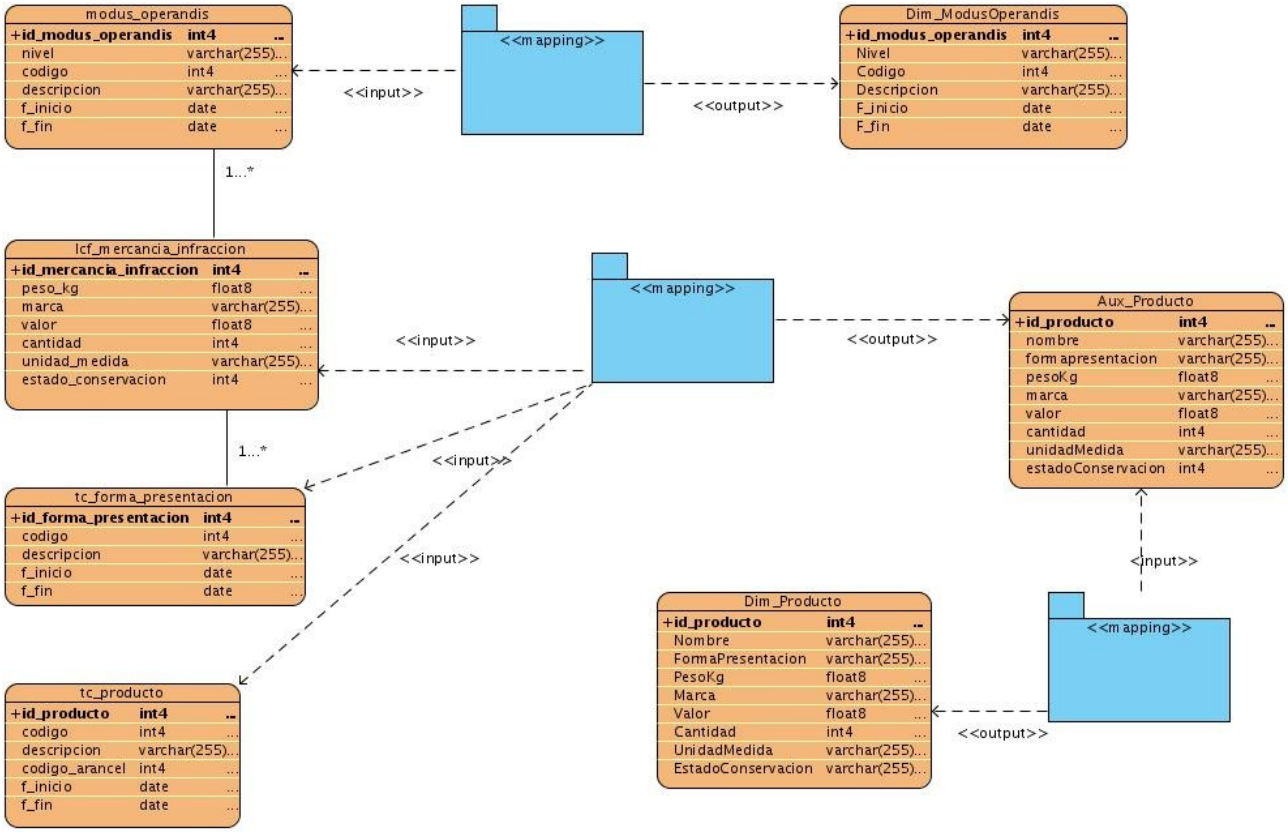

Figura 33: Mapeo de datos Dim_Producto y Dim_ModusOperandis(Nivel 2)

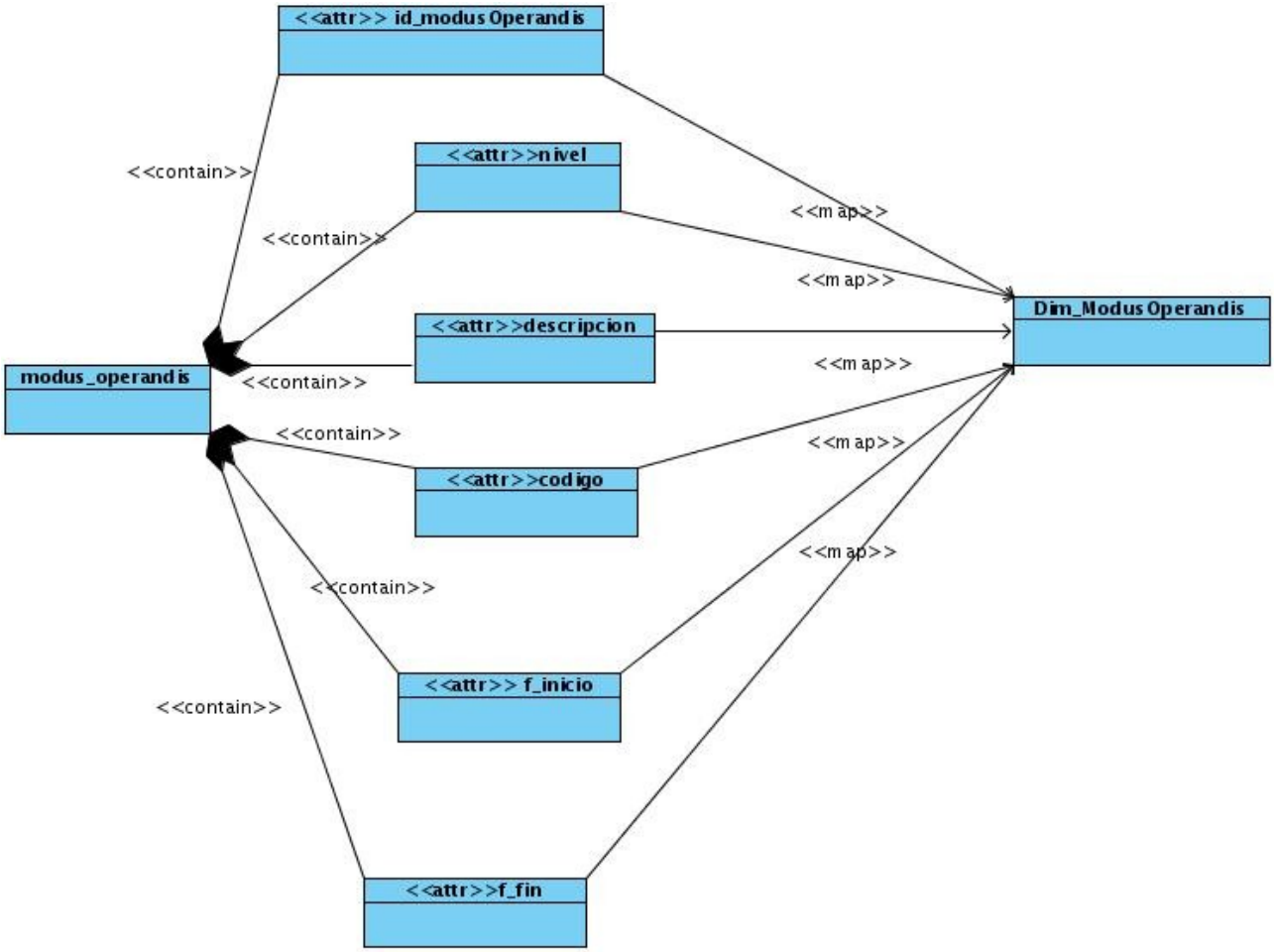

Figura 34: Mapeo de datos Dim_ModusOperandis(Nivel 3)

3.5 Conclusiones.

En este capítulo se realizó la descripción de el proceso de diseño de el Data Mart de el módulo Control de Personas. Se expuso la información más relevante que se va a almacenar, además de los enfoques de exploración de datos definidos. Se expuso además la metodología de diseño del mismo, aplicando el método DWEP, además se obtuvieron los diagramas definidos para los tres niveles, conceptual, lógico y físico.

Conclusiones Generales.

Al culminar el diseño del Data Mart del módulo Control de Personas se dio cumplimiento al objetivo general del trabajo de diploma. El Data Mart será capaz de dar soporte a una base de experiencia para la gestión de el cruce de personas naturales por frontera, así las tareas asignadas a los analistas se desarrollaran de una manera eficaz y en la menor brevedad posible.

El análisis del sistema operacional de Lucha contra el Fraude, junto a la guía de la metodología escogida para el proceso de diseño contribuyó a la selección apropiada de la información a modelar para el diseño del Data Mart y a un mejor entendimiento de los datos que se obtienen de este sistema.

Al mismo tiempo es capaz de servir de repositorio de datos para la aplicación de técnicas de minería de datos e inteligencia artificial que posibilitarían extraer el conocimiento tácito almacenado en los datos y convertirlo en conocimiento explícito con valor de uso para la organización.

Con el resultado obtenido en el análisis de los datos de las fuentes de datos y la guía de la metodología DWEP se diseñó el Data Mart del módulo Control de Personas. Como esta metodología es lo suficientemente extensible y versátil se deja el camino abierto para que en un futuro sea facilmente extensible.

Recomendaciones.

Culminar la construcción del Data Mart del módulo de Control de Personas y ponerlo en explotación en la Aduana General de la República.

Que este diseño sirva como paso impulsor para el diseño de futuros Data Marts en la Aduana General de la República.

Abrir líneas de investigación cuyos objetivos sean la aplicación de algoritmos de minería de datos a la base existente de manera que contribuyan a la transformación del conocimiento tácito a conocimiento explícito.

Citas Bibliográficas

1: Cerezal Tamargo, L; Gutiérrez García, R, Los sistemas operacionales: pilar fundamental..., 2002

2: Cerezal Tamargo,Lourdes, Primeros pasos para adentrarnos en el mundo de los Data Warehouse, 2002

3: Chavez Monzon, Carlos, Metodologia Integradora De Procesos Empresariales a Nivel Estrategico, 2008, http://www.gestiopolis.com/administracion-estrateg

4: Castillo R; Morata J; del Arbol J., Operational Data Store, 2005

5: Artiles Visual S; García González F, Dimension Informacional De La Gestion Del Conocimiento, 2005, http://www.alide.org.pe/download/Foro/Dimension%20

6: Barreto Véliz, Bernard Pavel, Inteligencia de Negocios,

7: Inmon, William, Building the Data Warehouse, 2002

8: Wolff, Carmen Gloria, La tecnología Data Warehouse,

9: Kimbal, Ralph, The Data Warehouse Toolkit, 1996

10: Kimball R, J Casserta, The Data Warehouse ETL Toolkit: Practical Techniques., 2004

11: VIDAL, L. V;MONTEAGUDO M. V., Estudio Teórico-Conceptual sobre Data Warehouse, 2000

12: Curto, Josep, Arquitectura de un datawarehouse, 2007, http://informationmanagement.wordpress.com/2007/06

13: Gutierrez D R, Data WareHouse, 2004, http://www.monografias.com/trabajos17/data-warehou

14: Ricardo Dario, Bernabeu, Hefesto: Metodología Propia Para La Construcción de un Data Warehouse, 2009

15: Lujan-Mora, S, Data Warehouse Desing with UML, 2005

16: , UML Lenguaje Unificado de Modelado, 2010, http://es.wikipedia.org/wiki/Lenguaje_Unificado_de

17: , RUP Proceso Unificado de Desarrollo de Software, 2010, http://es.wikipedia.org/wiki/RUP

18: Jacobson I; Booch G, El Proceso Unificado de Desarrollo de Software, 2000

19: , Visual Paradigm Sitio Oficial, 2010, http://www.visual-paradigm.com/

20: , SQL Power Architect Sitio Oficial, , http://www.sqlpower.ca/page/architect

21: , Umbrello Sitio Oficial, , http://uml.sourceforge.net/

22: , PostgreSQL Sitio Oficial, 2010, http://www.postgresql.org/

23: , pgAdmin3 Sitio oficial, , http://www.pgadminorg

24: , Pentaho Sitio Oficial, , http://www.pentaho.com/

25: Muñoz Leiva, Erika, Validación de Metodología Data Warehouse, 2008, http://validacionmetodologiadwh.blogspot.com/

26: Lujan-Mora S; Trujillo J, A UML Based Approach for Modeling ETL Process In Data Warehouses, 2003

Glosario de términos.

A

Aduana: Oficina pública, establecida generalmente en las costas y fronteras, para registrar, en el tráfico internacional, los géneros y mercaderías que se importan o exportan, y cobrar los derechos que adeudan.

Agregación: Actividad de combinar datos desde múltiples tablas para formar una unidad de información más compleja, necesitada frecuentemente para responder consultas del Data Warehouse en forma más rápida y fácil.

B

Base de datos: Colección de datos -estructurada y organizada- para permitir el rápido acceso a la información de interés.

D

Data Mart : Conjunto de hechos y datos organizados para soporte decisional basados en la necesidad de un área o departamento específico. Los datos son orientados a satisfacer las necesidades particulares de un departamento dado teniendo sólo sentido para el personal de ese departamento y sus datos no tienen porque tener las mismas fuentes que los de otro DataMart.

Data Warehouse : Base de datos que almacena una gran cantidad de datos transaccionales integrados para ser usados para análisis gestionales por usuarios especializados (tomadores de decisión de la empresa).

Denormalización(base de datos): proceso de procurar optimizar el desempeño de una base de datos por medio de agregar datos redundantes.

Dimensión : Entidad independiente dentro del modelo multidimensional de una organización, que sirve como llave de búsqueda (actuando como índice), o como mecanismo de selección de datos.

Drill Down : Exponer progresivamente más detalle (dentro de un reporte o consulta), mediante selecciones de ítemes sucesivamente.

Drill Up : Es el efecto contrario a drill-down. Significa ver menos nivel de detalle, sobre la jerarquía significa generalizar o sumarizar, es decir, subir en el árbol jerárquico.

H

Hardware: corresponde a todas las partes físicas y tangibles1 de una computadora: sus componentes eléctricos, electrónicos, electromecánicos y mecánicos;sus cables, gabinetes o

cajas, periféricos de todo tipo y cualquier otro elemento físico involucrado; contrariamente al soporte lógico e intangible que es llamado software.

J

Jerarquía : Es un conjunto de atributos descriptivos que permite que a medida que se tenga una relación de muchos a uno se ascienda en la jerarquía.

M

Mapeo: Proceso de convertir los datos que son transmitidos en un formato por el remitente, al formato de datos que puede ser aceptado por el receptor.

Metodología: Es un proceso de software detallado que define con precisión los artefactos, roles y actividades involucradas.

Minería de datos: Proceso de extracción de información y patrones de comportamiento que permanecen ocultos entre grandes cantidades de información.

Modus operandis: Del latín, modo de obrar (literalmente: “modo de operar”) o de hacer las cosas cuando es característica y reiterada. Manera especial de actuar o trabajar para alcanzar el fin propuesto.

N

Normalización(base de datos): proceso que consiste en aplicar una serie de reglas a las relaciones obtenidas tras el paso del modelo entidad-relación al modelo relacional.

O

OLAP: OLAP es el acrónimo en inglés de procesamiento analítico en línea (On-Line Analytical Processing). Es una solución utilizada en el campo de la llamada Inteligencia empresarial (o Business Intelligence) cuyo objetivo es agilizar la consulta de grandes cantidades de datos. Para ello utiliza estructuras multidimensionales (o Cubos OLAP) que contienen datos resumidos de grandes Bases de datos o Sistemas Transaccionales (OLTP). Se usa en informes de negocios de ventas, marketing, informes de dirección, minería de datos y áreas similares.

OLTP: OLTP es la sigla en inglés de Procesamiento de Transacciones En Línea (OnLine Transaction Processing) es un tipo de sistemas que facilitan y administran aplicaciones transaccionales, usualmente para entrada de datos y recuperación y procesamiento de transacciones (gestor transaccional). Los paquetes de software para OLTP se basan en la arquitectura cliente-servidor ya que suelen ser utilizados por empresas con una red informática distribuida.

R

RUP: Proceso Unificado del Rational, es un proceso de desarrollo de software, constituye la metodología estándar más utilizada para el análisis, implementación y documentación de sistemas orientados a objetos.

S

Sistema Gestor de Bases de Datos (SGBD): Es el software que permite la utilización y/o la actualización de los datos almacenados en una (o varias) base(s) de datos por uno o varios usuarios desde diferentes puntos de vista y a la vez.

Sistemas informacionales: Sistemas que permiten estudiar el comportamiento de la empresa, se utilizan para administrar y controlar la empresa.

Sistemas operacionales: Sistemas que garantiza la automatización de los procesos y el flujo de la información a través de la organización, se utilizan en el funcionamiento de los negocio en tiempo real, basándose en datos actuales.

Software: Se refiere a los programas y datos almacenados en un ordenador.

Snapshot : Imagen instantánea de los datos en un tiempo dado.

T

Transacciones: Suceso externo que involucra el traslado de algo de valor entre dos o más entidades.

Toma de decisiones: Conjunto de actividades intelectuales o cibernéticas que utilizando una cierta información disponible dan como resultado una ciertas acciones, todo ello en un contexto real y concreto.

U

UML: Es el Lenguaje de Modelado Unificado para detallar, construir, visualizar y documentar las partes o artefactos de un software.