Universidad de las Ciencias Informáticas

Título: Arquitectura y Componente de Almacenamiento del ODS para SIIPOL

Trabajo de Diploma para optar por el título de

Ingeniero en Ciencias Informáticas

Autores: Vladimir Urquia Cordero

Elieyis Gómez Padrón

Tutores: Ing. Alberto Limia Navarro

Ing. Yurima Ibañez Alfonso

Ciudad de la Habana

Junio de 2010

A veces sentimos que lo que hacemos es tan sólo una gota en el mar, pero el mar sería menos si le faltara esa gota.

Madre Teresa de Calcuta.

DECLARACIÓN DE AUTORÍA

Declaramos que somos los únicos autores de este trabajo y autorizamos a la Universidad de las Ciencias Informáticas a hacer uso del mismo en su beneficio.

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

_____	_____
Vladimir Urquia Cordero	Elieyis Gómez Padrón
Autor	Autora
_____	_____
Ing. Alberto Limia Navarro	Ing. Yurima Ibañez Alfonso
Tutor	Tutora

RESUMEN

El presente trabajo comprende las ideas fundamentales para el desarrollo del Almacén de Datos Operacionales (ODS) para el Sistema de Investigación e Información Policial (SIIPOL). Recoge una valoración crítica de las principales tendencias, herramientas y metodologías existentes para el desarrollo de este tipo de Soluciones de Almacenes de Datos e Inteligencia de Negocio. Se describe la metodología seleccionada y se presenta como resultado fundamental de la investigación, la arquitectura y el componente de almacenamiento del ODS para SIIPOL.

PALABRAS CLAVE

ODS, almacén operacional, integración de datos

ÍNDICE

ÍNDICE DE FIGURAS Y TABLAS

INTRODUCCIÓN

En las últimas décadas, se ha logrado un amplio desarrollo científico-técnico en las Tecnologías de la Información y las Comunicaciones (TIC)[1]; lo que provoca que todos los sectores de la sociedad deseen aprovechar las oportunidades que estas ofrecen para incrementar su productividad. Los sistemas judiciales no quedan exentos de este vertiginoso proceso de transformaciones. Las TIC les ofrecen todo un universo de posibilidades para mejorar los servicios, aumentar la eficiencia y eficacia en la investigación de actos delictivos y proveer un marco más transparente en las acciones de la administración pública.

En estos momentos, la República Bolivariana de Venezuela apuesta por la modernización y transformación de la Administración Pública. El Cuerpo de Investigaciones Científicas Penales y Criminalísticas (CICPC) de Venezuela es la organización con la misión de garantizar la eficiencia en la investigación del delito mediante su determinación científica, asegurando el ejercicio de la acción penal que conduzca a una sana administración de justicia.[1] Para ello resulta imprescindible tener un alto nivel de credibilidad nacional e internacional en la investigación del fenómeno delictivo organizado.

Para la renovación del CICPC se realizó un estudio de su situación general, identificándose un conjunto de problemáticas en los procesos que en él se desarrollan. Entre estas problemáticas se destacan las relacionadas con el Sistema de Investigación e Información Policial (SIIPOL).

Uno de los principales objetivos estratégicos de esta aplicación informática es gestionar la información policial. Entre las operaciones que controla se encuentra la confección y seguimiento de los expedientes policiales. Para este proceso se requieren datos que resultan claves en la mayoría de los actos delictivos, como son las personas, armas, vehículos y otros objetos.

Estos datos se encuentran dispersos en varios sistemas pertenecientes a diferentes instituciones y empresas privadas o estatales tales como: la Oficina Nacional de Identificación y Extranjería, el Registro y Notarías, el Instituto Nacional del Tránsito. Así como la Superintendencia Tributaria, el Centro de Tratamiento y Análisis de la Información, el Parque de Armas Nacional, la Dirección de Prisiones y Compañías de servicios de telefonía celular.

[1] Toda forma de tecnología usada para crear, almacenar, procesar e intercambiar información.

Para el trabajo de SIIPOL se hace necesario intercambiar información con los sistemas informáticos de dichas organizaciones. Entre ellos se encuentran: SAIME que almacena la información relativa a las personas incluyendo datos personales, filiatorios y movimientos migratorios; INTTT que recoge por otro lado lo referente a los vehículos con sus propietarios; y DARFA que por su parte controla las armas ya sea que estén en manos de funcionarios adscritos al cuerpo policial o de personas con licencia para portarlas.

SIIPOL no solo necesita datos de otros sistemas, sino que también los provee con información propia de su negocio, por lo que requiere un nivel adecuado de interoperabilidad[1] con los mismos. Sin embargo, existen problemas en la infraestructura de información en cuanto a la colaboración eficiente entre los distintos sistemas para el intercambio de datos. Su integración constituye un tema complicado, debido a la heterogeneidad de la base tecnológica y la calidad de la comunicación con cada uno, que va desde fiable hasta inexistente. Actualmente, parte de la gestión de información se realiza a través de mecanismos no automatizados y poco confiables; estos retardan la obtención de información oportuna y veraz sobre los hechos delictivos, recargando de trabajo administrativo a los funcionarios.

La situación descrita afecta la calidad de la información gestionada por el CICPC, teniendo en cuenta parámetros tales como: cantidad, seguridad y disponibilidad. Así como la rapidez y facilidad de accesos necesarios teniendo en cuenta la premura y alcance de las funciones de dicha entidad.

Partiendo de lo anteriormente expuesto se identifica el siguiente **problema científico**: ¿Cómo lograr la integración a nivel de datos entre el sistema de investigación e información policial del CICPC y los sistemas externos a dicha organización? Con el propósito de solucionar el mismo se traza como **objetivo general**: Definir la arquitectura base y el componente de almacenamiento del Almacén de Datos Operacionales (ODS, del inglés Operational Data Store) para SIIPOL.

Para dar cumplimiento al **objetivo general,** los **objetivos específicos** propuestos son:

- Definir lineamientos base de arquitectura ODS para SIIPOL.
- Definir las estructuras de almacenamiento.
- Instalar el ODS en un entorno de pruebas.

[1] Es la habilidad de los sistemas TIC, y de los procesos de negocios que ellas soportan, de intercambiar datos y posibilitar compartir información y conocimiento.

- Configurar el ODS en un entorno de pruebas.
- Validar la solución.

Por tanto el **objeto de estudio** lo constituyen los ODS y el **campo de acción** la arquitectura e implementación de ODS en entornos de integración e investigación policial.

Como resultado de la investigación se obtiene la arquitectura y el componente de almacenamiento del ODS para SIIPOL.

Para darle cumplimiento a los **objetivos específicos** trazados, se definieron las siguientes **tareas**:

1. Estudio de los fundamentos teóricos de los ODS haciendo énfasis en su diseño e implementación.
2. Caracterización de las tendencias actuales de diseño e implementación de ODS.
3. Documentación de la metodología a utilizar en el desarrollo.
4. Selección de las herramientas que brinden los servicios y las funcionalidades necesarias para el diseño y la implementación del ODS.
5. Instalación de las herramientas de trabajo.
6. Caracterización de los sistemas relacionados con SIIPOL.
7. Definición de la arquitectura del ODS para SIIPOL.
8. Caracterización de cada componente definido en la arquitectura.
9. Identificación de las entidades, atributos y relaciones.
10. Estructuración del modelo lógico.
11. Transformación del modelo relacional al diseño físico.
12. Definición de políticas de seguridad y control de acceso.
13. Implementación del modelo de datos del ODS.

14. Montaje del ODS en un entorno de pruebas.
15. Configuración del ODS en un entorno de pruebas.
16. Definición de las pruebas de validación del sistema.
17. Realización de pruebas al ODS.

Para dar cumplimiento al objetivo planteado se aplicaron los **Métodos Teóricos** siguientes:

Analítico Sintético: Se utiliza en la revisión de documentos y artículos, de donde se extrajeron ideas y elementos importantes vinculados con la investigación. Permitió ampliar más sobre el tema, estudiando sus particularidades, obteniendo ideas centrales y relacionándolas como un todo.

Histórico Lógico: Permite estudiar los antecedentes y evolución de los ODS, documentar los temas relacionados con el desarrollo del ODS para SIIPOL y la metodología a utilizar.

Inductivo Deductivo: Se utilizó para el planteamiento del objetivo y para realizar una propuesta que se adapte a las necesidades del SIIPOL; teniendo en cuenta los conceptos esenciales, aceptados como válidos, relacionados con la arquitectura e implementación de los ODS.

Modelación: Permite descubrir y estudiar nuevas relaciones y cualidades mediante la reproducción simplificada de la realidad. Es necesario para realizar los modelos de datos de las diferentes fuentes con los datos de interés para SIIPOL y escoger la alternativa que responda a las necesidades identificadas.

Como **Método Empírico** es empleada la **Entrevista**, se entrevistó sistemáticamente al personal especializado del proyecto CICPC en la UCI.

El documento está estructurado en tres capítulos que se describen a continuación:

Capítulo 1: Fundamentación Teórica.

En este capítulo se caracterizan las tecnologías de almacenamiento de datos, los Almacenes de Datos (DW, del inglés Data Warehouse) y ODS. Se profundiza en los ODS, sus principales características, los elementos que los componen, el estado del arte de sus desarrollos a nivel mundial y nacional, las metodologías y herramientas existentes para su desarrollo, así como la justificación de su uso.

Capítulo 2: Diseño e Implementación del ODS.

En este capítulo se establecen los lineamientos base de la arquitectura y la descripción de la solución en general. Se especifica la caracterización de las fuentes a integrar, la definición de las áreas de análisis, la arquitectura, el diseño, la implementación del modelo de datos propuesto, las políticas de seguridad y los procedimientos de recuperación ante desastres.

Capítulo 3: Análisis de los Resultados.

Este capítulo está orientado al análisis y validación de los resultados. Se detallan las temáticas referidas a la normalización, calibrado de la base de datos, análisis del rendimiento, pruebas y validación general del sistema.

CAPÍTULO 1: FUNDAMENTACIÓN TEÓRICA

En el presente capítulo se describen aspectos teóricos relacionados con la evolución de las tecnologías de almacenamiento de datos, particularmente las conocidas como DW y ODS. Se muestra un análisis del estado del arte de sus desarrollos a nivel mundial y nacional. Se presentan las metodologías y herramientas existentes para el desarrollo de este tipo de soluciones. Además, se exponen conceptos importantes que permiten una comprensión significativa de la investigación y justifican la solución escogida.

1.1 Almacenes de Datos

La aparición de la computación en los años 70 y su vertiginoso desarrollo provocó la necesidad inmediata de preservar y gestionar la creciente información generada. En este contexto surgen los Sistemas de Ficheros Relacionados, pioneros de una continua búsqueda de la eficiencia y calidad en el almacenamiento y gestión de datos, quienes rápidamente evolucionaron hacia los Sistemas de Bases de Datos, los cuales prestaban un servicio con mayor rapidez y calidad.

Con el transcurso de los años y el aumento de la complejidad de las tecnologías, las bases de datos se convirtieron en una herramienta fundamental de control y manejo de las operaciones dentro de las instituciones y la información que se almacenaba comenzó a aumentar exponencialmente en cantidad e importancia.

Los sistemas creados hasta ese momento fueron decayendo ante la creciente necesidad de realizar análisis exhaustivos de los datos, estas operaciones se tornaban altamente costosas y atentaban contra su funcionamiento. Se fueron creando otros sistemas y programas que realizaban estas funciones, pero lo hacían de forma independiente, extraían y analizaban parte de la información, que posteriormente era sometida a criterios de especialistas en determinada área de la organización.

Estos sistemas tomaron auge por un tiempo pues separaban la información y cada usuario analizaba solo la parte de los datos conveniente a su departamento. Si se consideran las valoraciones de William H. Inmon se puede constatar que dicha forma de concebir la extracción y análisis de la información se conoció como “Evolución Natural de la Arquitectura” y solo se hizo efectiva para resolver problemas específicos y puntuales, ver **Figura 1**. [2]

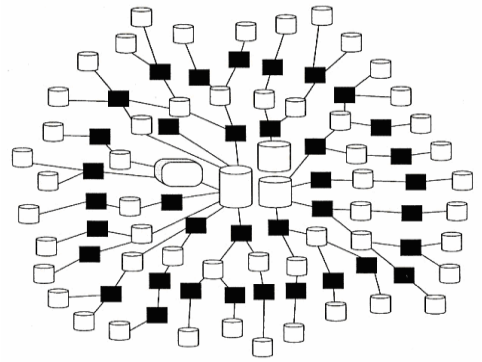

Figura 1: Evolución Natural de la Arquitectura

Con el tiempo la ventaja de ver los datos por separados se convirtió en un gran problema, se extraía información de datos ya analizados y se formaba una especie de "*tela de araña*" de la cuál resultaba difícil escapar. Para resolver los problemas identificados, se concluyó que lo ideal sería unificar las diferentes fuentes de información de las cuales se disponía, en un único lugar, al que sólo se le incorporaría información relevante, sobre la base de una estructura organizada, integrada, lógica, dinámica y de fácil explotación. Surgen entonces los DW. [3],[4]

Para adentrarse en el tema de los DW se considera necesario hacer referencia a las principales personalidades que se desenvuelven en este campo. Existen diversas tendencias y formas de conceptualizar esta terminología que aunque difieren en algunos aspectos, giran sobre el mismo eje central.

Al referirse a este particular; Imhoff, Galemmo y Geiger prestan especial atención al concepto desarrollado por Inmon en los años 90, en el que señala que los DW "*son un conjunto de datos orientados a un tema, integrados, de tiempo variante y no volátiles usados en la estrategia de toma de decisiones administrativas*". Aseveran además que "*Los Almacenes de Datos se han venido reconociendo cada vez más como una herramienta efectiva de las organizaciones para transformar los datos en información útil y estratégica para la toma de decisiones*". [5]

Según las apreciaciones de Wang se debe analizar la siguiente definición esbozada en 1996: "*Los Almacenes de Datos entraron en existencia para satisfacer las necesidades, consolidando e integrando la información de fuentes internas o externas en función de organizarla en un formato útil, para soportar a las decisiones empresariales*".[6]

Ralph Kimball, mundialmente reconocido por sus libros sobre el tema, ha expuesto que "*los Almacenes de Datos son una copia de los datos de la transacción estructurados específicamente para la pregunta y el análisis*".[3] Criterio que ha sido reafirmado por determinados estudiosos del tema, un ejemplo de ello lo constituye las reflexiones de Adamson en el 2006 donde argumenta que "*Si el objetivo de los sistemas operacionales es la ejecución del proceso del negocio, los DW soportan la evaluación del proceso del negocio*."[7]

Aunque existe diversidad de puntos de vista y percepciones sobre el tema, se pueden identificar claramente algunos elementos cruciales en todos los casos concluyendo en que los DW son estructuras que se definen en función de temas específicos donde la información histórica debe estar integrada, robusta ante los cambios que puedan afectar a la organización y que su objetivo principal, y es lo que define su razón de ser, es servir de soporte a la toma de decisiones.

Con esta nueva manera de concebir el almacenamiento de los datos se persigue, a pesar de las incompatibilidades que puedan existir entre diferentes sistemas y sus contenidos, unificar la información y transformarla, haciéndola claramente legible y que pueda llegar al usuario un conocimiento general de los procesos que ocurren en su entorno de trabajo.

1.2 Almacenes de Datos Operacionales

Los DW están enfocados fundamentalmente a ofrecer una panorámica integral de los datos, que sostenga las funciones generales de la gerencia a la hora de tomar decisiones estratégicas. A estos niveles generalmente las solicitudes de reportes van encaminadas a ver la información agrupada por diferentes criterios, donde se pueda observar el comportamiento general que siguen determinados datos, pero solo de forma global, no hechos detallados.

En diversas ocasiones, cuando se está realizando el proceso de carga hacia los DW, se puede manejar de una forma u otra, cierto nivel de atomicidad de los mismos, pero a medida que este proceso avanza, y se incrementa el nivel de almacenamiento, se pierden los detalles de los datos que están siendo cargados; pero la práctica ha confirmado que en determinadas circunstancias resulta necesario realizar un análisis exhaustivo a un nivel altamente detallado como es el caso de las transacciones.

Los sistemas transaccionales no deben afectarse con un alto consumo de recursos, esto pudiera provocar problemas de rendimiento, inestabilidad o incluso caída parcial o total de los mismos en análisis de gran

alcance; sin contar la necesidad de utilizar datos que se encuentran en constante actualización, cuyo proceso de cambio constituye el principal objetivo del análisis.

Tampoco sería una alternativa viable alterar la arquitectura del DW para que contuviese datos operacionales, esto significaría alterar por completo la arquitectura general de los Sistemas de Información. Surge entonces la necesidad de contar con un sistema integrador de datos que brinde la información con un alto nivel de detalle operacional, los ODS.

Dos de las personalidades reconocidas mundialmente por sus conocimientos en esta teoría han elaborado sus propias definiciones de ODS, se considera oportuno citar las mismas:

Según Ralph Kimball un ODS es un almacén de información detallada, orientado a temas, integrado, aumentado con frecuencia, dentro del DW de una empresa.[8]

Por su parte, Inmon considera que un ODS es una colección de datos orientada a temas, integrada, volátil, actualizada, sólo detallada, que sustenta las necesidades de información reciente, operacional, integrada y colectiva de la organización.[9]

Al examinar dichos conceptos se puede apreciar que discrepan en cuanto a la permanencia de los datos en el sistema; es decir, en cuanto a lo que a almacenamiento histórico se refiere. Inmon ofrece un carácter de mayor volatilidad a la información almacenada, previendo el incremento de datos con ese alto nivel de detalle, que es básicamente lo que propone Kimball. Pero es relevante destacar la necesidad actual de realizar análisis de información cada vez más detallada y de un período histórico específico, según lo requiera el negocio de la empresa en cuestión.

Solo cabría preguntarse si las transacciones pueden analizarse estando al nivel de detalle que se describe en el ODS. Quizás, hace algún tiempo atrás, esto no sería posible; pues el soporte tecnológico no era suficiente, todo indicaba a la tendencia de ir disminuyendo en cuantía de detalles y atomicidad de datos. Hoy sucede lo contrario, potencialmente existen los soportes de hardware y software que permiten la incorporación y almacenamiento por períodos históricos de datos altamente granulados.

Tampoco puede obviarse completamente lo planteado por Inmon. ¿Con qué período de tiempo es factible el almacenamiento de estos datos?, ¿es necesario conservarlos todos? En los ambientes operacionales, por lo general no hace falta retener los datos por períodos históricos prolongados, y es aquí una de las

interpretaciones fundamentales que se le pudiera dar a la definición de Bill Inmon cuando hablaba de volatilidad de datos. Se puede afirmar entonces que ambas variantes, tanto la de Kimball como la de Inmon, son favorables a la hora de concebir el ODS; en cuanto a la conservación de los datos, es responsabilidad de la institución definir los recursos disponibles y necesidades objetivas a la hora de implementar su sistema.

Atendiendo a los elementos analizados se entiende por ODS una colección de datos integrados, orientados a un tema, que brindan información actualizada con un alto nivel de detalle operacional. Su principal función es constituir la fuente de datos para herramientas que ayudan a la toma de decisiones tácticas.

1.2.1 Estado Actual de los ODS

Dadas las características de un sistema de almacenamiento de datos puede exponerse que su aplicación más rica corresponde a entornos de empresas que como parte de su negocio manejan o generan grandes volúmenes de datos asociados a sus clientes, compras, marketing y de manera general a sus transacciones u operaciones diarias.
Se pueden citar algunos ejemplos de sectores y empresas que han implantado DW, entre ellos las Empresas de telecomunicaciones Telefónica móviles, Jazztel, Vodafone y France Telecom. Empresas de transporte (Aerolíneas, Transporte de Cargas y Transporte de Pasajeros) entre ellas British Airways, Union Pacific y Air France. Empresas de fabricación de bienes de consumo masivo como Coca-Cola, Adidas, Nike, 3M, Bosh Siemens, prácticamente todas las empresas de fabricación de automóviles, etc. Entidades Financieras: BBVA, Caja Madrid, Caja Extremadura.[10]

Algunas de ellas han implementado ODS como complemento de sus propios DW. Entre las más importantes de acuerdo a la gran cantidad de recursos que manejan están: la gigante de circuitos electrónicos dedicados a la telecomunicación AT&T, quién utiliza el ODS como soporte a la toma de decisiones de situaciones operacionales que se puedan presentar como parte de su negocio y Telefónica, operador de telecomunicaciones líder en el mundo de habla hispana y portuguesa, con más de 82 millones de clientes, quién ya tiene operativo su nuevo Sistema de Información Unificado bajo el concepto de ODS.[11]

En Cuba, producto a factores económicos que influyen directamente sobre las cuestiones de desarrollo de las tecnologías no se ha visto un sustancial avance en este tema específico. Muchas entidades que

manejan volúmenes cuantiosos de datos han implementado su propio DW, sin comportarse de la misma manera con los ODS, cuyo desarrollo es aún incipiente. Se conoce por fuentes documentables que la Corporación Cimex es la única que trabaja con ODS en su negocio interno. Resultando la utilización de los mismos de una utilidad notable a la hora de realizar el proceso de toma de decisiones[12].

1.2.2 Actualización de los ODS. Clasificación

El ODS puede asumir diferentes frecuencias para actualizar su contenido, esto puede determinar los métodos que se utilicen para el tránsito de la información desde los sistemas transaccionales, así como el nivel de integración y transformación de los datos.

En sus inicios, los sistemas ODS que fueron desarrollados eran utilizados como una herramienta de reportes con propósitos administrativos. Se actualizaban diariamente y ofrecían resúmenes sobre las transacciones empresariales del día. Este tipo de sistemas recibe hoy el nombre de **ODS de Clase III**.

En esta categoría el movimiento de datos hacia el sistema se realiza con periodicidad diaria, de preferencia durante la noche. El método más utilizado para el traspaso es el de almacenar los datos y enviarlos más tarde (técnica conocida como Store and Forward). Los cambios, por lo general, se escriben a un fichero y luego son cargados por lotes al sistema. Son posibles una mayor integración y transformaciones complejas sobre los datos; esto implica el uso considerable de herramientas de Extracción, Transformación y Carga (ETL, del inglés Extraction Transformation and Load). Las sumarizaciones para reportes son realizadas una vez al día, esta clase de ODS es la más fácil de desarrollar y mantener. [13]

Con el aumento de las necesidades empresariales, el ODS evolucionó para convertirse en lo que hoy se conoce como **ODS de Clase II**, siendo capaz de manejar información más compleja y renovarse frecuentemente. Los intervalos entre una carga y otra pueden ir desde 15 minutos hasta varias horas. Para el traspaso es común utilizar el método Store and Forward descrito. Se puede realizar algo de integración y transformación mientras los datos son cargados, para este proceso se sugiere el uso de herramientas ETL. Además, se pueden hacer sumarizaciones instantáneas, aunque se recomienda hacerlas sólo una vez al día, por ejemplo, en la madrugada. [13]

Los **ODS de Clase I** surgieron con la llegada de los Sistemas de Administración de Relaciones con el Cliente[1] (CRM). Los CRM requerían la creación de un ODS enfocado a los clientes, cuyas actualizaciones fueran sincrónicas o casi sincrónicas con los sistemas transaccionales, de tal manera que se pudiera ofrecer información consistente y organizada inmediatamente después de la ocurrencia de cambios en estos últimos.

En el ODS de Clase I los cambios aparecen 2 ó 3 segundos después de ocurridos en los sistemas fuentes. Por lo general el intercambio de información se realiza a través de la transmisión de mensajes por capas intermedias[2]. Raras veces se utilizan herramientas ETL debido a que las transformaciones sobre los datos tienden a ser escasas. El modelo de datos usualmente es parecido en algunos aspectos al de los sistemas fuentes. En esta clase de ODS, las sumarizaciones instantáneas pueden ser difíciles de realizar, por lo que se recomienda efectuarlas en diferentes intervalos durante el transcurso del día, por ejemplo, cada una hora. [13]

En conformidad con las investigaciones de Imhoff, existe otra categoría más reciente para los ODS, la **Clase IV**, donde la información se recibe desde el DW, realizando un proceso de retroalimentación para chequear el estado de la información actual y completar análisis tácticos. El movimiento de datos, puede realizarse en intervalos regulares o irregulares. Cualquiera de las clases I, II, ó III pueden convertirse en una Clase IV. El requerimiento es, por supuesto, que debe existir un DW antes de asumir esta categoría. El proceso de traspaso de información de un entorno a otro se puede realizar de manera sencilla, haciendo uso, por ejemplo, de herramientas ETL.[13]

De las categorías vistas, la I es un caso que se encuentra muy pocas veces, utiliza muchos recursos, son considerables los gastos en su mantenimiento y en sus inicios es muy difícil su sincronización. La implementación de un sistema de este tipo, con actualización instantánea, resulta complicada por lo que sólo debe ser creado cuando la tecnología disponible lo permita.

Las clases II y III son comúnmente usadas pues pueden mantenerse con tecnología estándar siendo menos costosas y más fáciles de desarrollar que la clase I. Estas estructuras son relativamente sencillas y requieren de menos recursos en línea para la carga. [13]

---

[1] *El término Administración de Relaciones con el Cliente (Customer Relationship Management, CRM) es usado para describir herramientas de personalización sofisticadas, desarrolladas por algunos distribuidores para definir grupos de clientes y clasificarlos correctamente de acuerdo a los productos y servicios que se le ofrecen.*

[2] *Conocido también como messaging middleware*

Puede concluirse que además de las estrategias y necesidades de renovación de los datos que tenga la empresa, de igual manera otros factores van a determinar asumir una frecuencia de actualización u otra. Entre ellos, las posibilidades económicas de la organización para brindar el respaldo tecnológico que implica cada clase de ODS; así como el tiempo que demoran los sistemas operacionales en generar los datos a cargar y la complejidad del procesamiento que el ODS realice sobre la información durante la carga.

Debe tenerse en cuenta que a cada clase le corresponden distintas tecnologías y tienen propósitos diferentes. Es aconsejable comenzar desarrollando un ODS de Clase III e ir incrementando la frecuencia de actualización a medida que el sistema evolucione. En el desarrollo de la presente solución se ha seleccionado la Clase III, la misma cumple con los requerimientos de actualización para el tipo de negocio manejado por CICPC.

1.3 Almacenes de Datos vs Almacenes de Datos Operacionales

Comparando los conceptos ODS y DW desde un punto de vista superficial puede parecer que no existen diferencias notables entre ellos, pues coinciden en la estructura y los datos almacenados. El ODS es temático e integrado como el DW pero volátil con valores actuales y detallados. Sin embargo, la principal característica que los distingue es el tipo de consulta que sobre ellos se realice.

En un DW, por lo general, se efectúan análisis de tendencias para comprender mejor el comportamiento del negocio de manera global e histórica; una tendencia no se puede definir con la recopilación de los datos que se brindan de unos días, es un proceso más abarcador, en el cuál se necesita información incluso de años. En un ODS el análisis que se realiza es operacional, de forma más detallada, de último momento, y como se señaló anteriormente, esta información posee un período de vigencia que resulta interesante para el analista. Esto determina la diferencia de los datos almacenados en estos entornos, en uno datos operacionales, en otro, datos informativos.

La diferencia de objetivos entre un entorno y otro, ocasiona que se cree una divergencia de agregación en cada sistema, en el ODS se encontrará información atómica e indivisible con un alto nivel de detalle, mientras que en el DW se mantendrán los datos agregados. Además, los distintos tipos de análisis que se realicen sobre el ODS y el DW como sistemas contenedores de información determinan que la frecuencia de actualización de los datos en el ODS debe ser mayor, o en el peor de los casos igual a la del DW, ya que el primero debe contener las últimas modificaciones ocurridas en los sistemas transaccionales.

Otra diferencia sustancial es la probabilidad de acceso de los diferentes usuarios al ODS y el DW. En cualquiera de los dos sistemas pueden hacer consultas tanto personas como otros sistemas propiamente dicho; las personas con el objetivo de intentar descubrir, a simple vista, comportamientos en el negocio que los conlleve a tomar determinada decisión; los sistemas por su parte accederán al usar un subconjunto de la información, reestructurarla a conveniencia y servir de soporte a los analistas en el proceso de toma de decisiones. Debido a que la información contenida en el ODS es de gran riqueza y de fácil restructuración es más frecuente el acceso de otros sistemas, mientras que los analistas querrán acceder al DW, pues les hace falta tener una visión global y utilizar datos totalizados, y es en este ambiente donde podrán lograr esto, recorriendo de lo más general hasta el nivel de detalle deseado.

Las características y objetivos de cada entorno están bien delimitados, y no es intención de uno sustituir las funcionalidades del otro, sino que cada cual tiene su finalidad bien definida dentro de la arquitectura de los Sistemas de Información. En el **Anexo 1** se detallan las principales diferencias entre los ODS y los DW.

1.4 Metodología de Desarrollo

Existen numerosas metodologías para controlar y documentar el desarrollo de soluciones de almacenamiento de datos. Entre ellas se distinguen dos enfoques fundamentales con los cuales se implementaron DW en un principio y aún hasta el día de hoy, pues han sido los más influyentes y mejor conceptualizados por sus creadores, Ralph Kimball y William H. Inmon respectivamente. Inmon es el creador del término Data Warehouse así como de la Fábrica de Información Corporativa (CIF, del inglés Corporate Information Factory), conjuntamente con Claudia Imhoff; es considerado por todos el padre de la disciplina. Por su parte, Ralph Kimball es un gurú del diseño de almacenes de datos y creador del enfoque Arquitectura Multidimensional (MD de sus siglas en inglés, Multidimensional Architecture).

Inmon trata la construcción de los almacenes con un enfoque descendente (top-down) donde los pequeños almacenes departamentales (DM, del inglés Data Mart) se nutrirán del DW. Esboza la creación de un repositorio de datos corporativo como fuente de información consolidada, consistente e histórica. Al ser construido descendentemente los DM se nutren del DW Corporativo, convirtiéndose en un complejo empresarial de bases de datos relacionales.[14]

Por su parte, Ralph Kimball expone que el DW se compone por el conglomerado de todos los DM generados en una empresa y que la información siempre se almacena en un modelo dimensional.

Respecto a esto, los razonamientos expuestos por **Curto**[1] confirman que una de las características principales de la metodología respaldada por Kimball es que su arquitectura es ascendente (bottom-up), consigna que se debe crear por cada departamento un conjunto de DM independientes orientados a los temas que estén relacionados con él. Y "*El DW es la unión de todos los DM de una entidad*".

Basados en estas propuestas se han desarrollado un conjunto de metodologías que no siguen obligatoriamente una específica sino que realizan una selección de lo mejor de cada una y definen su propia metodología. Ejemplo de esto se tiene la avalada por Microsoft llamada **Metodología SQLBI** orientada totalmente a las herramientas que propone el gigante del software como son Microsoft SQL Server, SQL Server Analysis Services y su oferta más completa en este campo que es Microsoft Suite for Business Intelligence.[15]

Otra metodología para este tema es la denominada **HEFESTOS**, de acuerdo a las indagaciones de Bernabeu; entre sus principales directrices plantea que la construcción e implementación de un DW puede adaptarse muy bien a cualquier ciclo de desarrollo de software. Lo que se busca, es entregar una primera implementación que satisfaga una parte de las necesidades, para demostrar las ventajas del DW y motivar a los usuarios. [16]

Cada una de estas metodologías pretende dar un acercamiento a una propuesta ideal para el desarrollo de DW. Cada autor la orienta a la optimización del rendimiento y a su visión de los principales procesos que se deben tener en cuenta para construir un DW flexible y dinámico.

El análisis realizado por Curto en el 2008 esclarece que existen situaciones en las que una de las arquitecturas clásicas proporciona ventajas competitivas sobre la otra. Hecho que influye en la selección de una u otra. Normalmente la realidad es que ambas se combinan para proporcionar la mejor respuesta a las necesidades del cliente.[14]

La construcción del ODS para SIIPOL amerita la utilización de una metodología robusta y madura que garantice el éxito de la integración de los datos provenientes de los sistemas externos. Por estas razones se escoge la adaptación de la **metodología de Kimball** denominada Metodología para el Desarrollo de Soluciones de Almacenes de Datos e Inteligencia de Negocio en el Centro de Tecnologías de Gestión de

[1] Josep Curto: Master en *Business Intelligence* y en Dirección y Gestión de Sistemas y Tecnologías de la Información, colaborador puntual en revista Gestión del Rendimiento, autor del libro "Introducción al Business Intelligence".

Datos (DATEC). Esta adaptación es guiada por casos de uso e incluye una etapa de pruebas en aras de estar más alineados con las tendencias y normas de la Universidad de Ciencias Informáticas (UCI).[15] La selección de esta metodología se sustenta en las siguientes razones:

- La técnica de Kimball posee una gran cantidad de documentación y generalmente se puede encontrar una respuesta a casi todas las problemáticas que se puedan presentar.
- Su creador Ralph Kimball es una figura emblemática en el mundo de warehousing teniendo publicados alrededor de 100 artículos científicos proponiendo mejoras al proceso, además de innumerables libros que se han posicionado como guías de obligatoria consulta para el desarrollo.
- Propone con claridad cada actividad, en cada uno de los períodos de construcción, que deben realizar los roles involucrados en el proyecto.
- Esta metodología de dividir el mundo de la Inteligencia de Negocio (BI, del inglés Business Intelligence) entre el hecho y las dimensiones es muy eficaz y conduce a una solución completa en un tiempo razonable.
- Es iterativo e incremental, donde el almacén de datos es construido pieza por pieza garantizando mayor velocidad de respuesta a los clientes.
- La forma de almacenar la información es de fácil entendimiento para los usuarios finales lo que permite mayor comprensión para el análisis de los datos que se encuentran integrados y detallados.
- Es una metodología dúctil, es decir, resistente y adaptable a los cambios.

1.5 Modelo Conceptual de Datos

En 1976 Peter Chen publicó el Modelo Entidad Relación (MER) original, que proveía un enfoque visual fácil de usar del diseño lógico de la base de datos. Este modelo elude las complicaciones de almacenamiento y consideraciones de eficiencia, las cuales son reservadas para el diseño físico de la base de datos.

Los principales elementos de dicho modelo son las entidades con sus atributos y relaciones. En consonancia con los razonamientos de Wolff se puede afirmar que el MER se caracteriza por dividir los datos en entidades discretas, las que son representadas como tablas físicas en una base de datos.

Realizar un cambio en la base significa tocarla en un solo lugar pues la redundancia se trata fuera de los datos. Bajo estas directrices se realizan los sistemas operacionales, cuya función es reflejar el estado y funcionamiento de las empresas mediante el registro de sus operaciones diarias.[17]

El modelo relacional se basa en el concepto de relación, siendo esta un conjunto de n-tuplas que pueden representar tanto entidades como relaciones entre estas. Su aceptación no se debe a que permite de forma implícita operaciones conceptualmente abstractas sobre los datos sino a los altos niveles de fiabilidad e integridad que aporta en el manejo de grandes cantidades de datos.

Este modelo tiene un diseño flexible, permite el descubrimiento de nuevas relaciones, patrones e información que pueden ser relevantes para la organización. Proporciona el sustento para la naturaleza transaccional del ODS pues facilita el análisis exploratorio, abarcando visiones agregadas y predefinidas e igualmente el análisis a nivel de detalle.

1.6 Modos de Almacenamiento de Datos

La forma de almacenamiento es crítica para garantizar el rendimiento de las consultas, las zonas de ubicación de las agregaciones y el procesamiento en general. Existen tres modelos para el proceso analítico en línea (OLAP) de la información: ROLAP, MOLAP y HOLAP. En dichos modelos el proceso de análisis se realiza de igual forma, lo que varía en uno y otro caso es el modo en que son almacenados los datos.

1.6.1 Sistemas ROLAP

Los sistemas ROLAP usan bases de datos relacionales para el manejo, acceso y obtención de los datos; aunque presentan los datos a los usuarios en forma de dimensiones de negocio. Esto es posible gracias a la creación de la semántica de las etiquetas de los metadatos que soportan el mapeo de las dimensiones a las tablas relacionales, permitiendo ocultar las estructuras de almacenamiento y presentar los datos dimensionalmente. Estos metadatos también son almacenados en tablas relacionales. La **Figura 2** muestra la arquitectura que presenta el modelo ROLAP.[18]

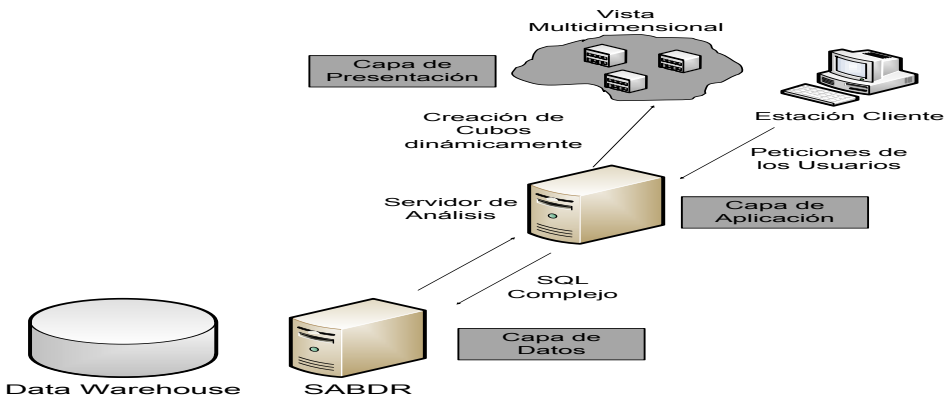

Figura 2: Modelo de almacenamiento ROLAP

Los usuarios finales ejecutan sus análisis multidimensionales, el motor ROLAP transforma dinámicamente sus consultas a consultas SQL que se ejecutan en las bases de datos relacionales, y cuyos resultados se relacionan mediante tablas cruzadas y conjuntos multidimensionales para devolver los resultados.

La arquitectura ROLAP accede directamente a los datos del almacén, y soporta técnicas de optimización de accesos, tales como particionado de los datos a nivel de aplicación, soporte a la desnormalización y joins múltiples, para acelerar las consultas.

1.6.2 Sistemas MOLAP

Por su parte el modelo MOLAP almacena los datos dimensionalmente. Es decir, las estructuras de almacenamiento son grandes arreglos dimensionales que son una copia de la fuente de datos y persisten físicamente en la misma estación de trabajo donde está instalada la herramienta de almacenamiento de datos. Aquí las estructuras de los datos están fijas para que la lógica, al procesar la información, pueda estar basada en métodos bien definidos para establecer las coordenadas del almacenamiento de los datos. La **Figura 3** muestra la arquitectura que presenta el modelo ROLAP.[18]

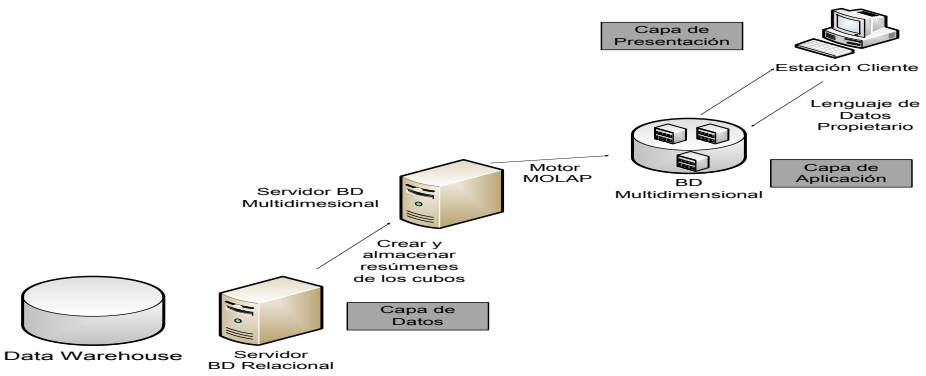

Figura 3: Modelo de almacenamiento MOLAP

1.6.3 MOLAP versus ROLAP

La selección de uno u otro modelo depende de cuán importante sea el rendimiento de las consultas para los usuarios y de la tecnología disponible a utilizar. En el modelo ROLAP la respuesta a las consultas y el tiempo de procesamiento suelen ser más lentos que con los modos de almacenamiento MOLAP ó HOLAP. No obstante, ROLAP permite a los usuarios ver los datos en tiempo real y ahorrar espacio de almacenamiento al trabajar con conjuntos de datos grandes a los que no se suele consultar con frecuencia, por ejemplo, datos puramente históricos.

Por otro lado, las implementaciones ROLAP son más escalables y son frecuentemente atractivas a los clientes debido a que aprovechan las inversiones en tecnologías de bases de datos relacionales ya existentes en la organización. En las implementaciones MOLAP el acceso a la información almacenada se realiza de forma más rápida y efectiva utilizándose un depósito donde el tiempo en la velocidad de respuesta es crítico. Normalmente se desempeñan mejor que la tecnología ROLAP, pero tienen problemas de escalabilidad En el **Anexo 2** se muestra una tabla comparativa entre ambos modelos basándose en Almacenamiento de los Datos, Tecnologías Subyacentes, Funciones y Características.[18] Para el desarrollo del ODS se utiliza el sistema ROLAP por las ventajas que representa en este tipo de entornos, donde el detalle y la consolidación de los distintos conceptos relacionados condicionan el funcionamiento del sistema.

1.7 Metadatos

El término metadatos no tiene una definición única. Según la definición más difundida, metadatos son "*datos sobre datos*". Otra clase de definiciones trata de precisar el término como "*descripciones*

estructuradas y opcionales que están disponibles de forma pública para ayudar a localizar objetos"[19] o "*datos estructurados y codificados que describen características de instancias conteniendo informaciones para ayudar a identificar, descubrir, valorar y administrar las instancias descritas*"[20].

Los metadatos pueden describir colecciones de objetos y también los procesos en los que están involucrados, describiendo cada uno de los eventos, sus componentes y las restricciones que se les aplican. Definen las relaciones entre los objetos, como las tuplas en una base de datos o clases en orientación a objetos, generando estructuras.

Los beneficios derivados de la utilización de metadatos son diversos y dependen del área en que se utilicen. Los metadatos adhieren contenido, contexto y estructura a los objetos de información, asistiendo de esta forma al proceso de recuperación de conocimiento desde colecciones de objetos. [21] Entre sus principales potencialidades se encuentran:

- Permiten generar distintos puntos de vista conceptuales para usuarios o sistemas, y liberan a estos últimos de tener conocimientos avanzados sobre la existencia o características del objeto que describen.

- Permiten el intercambio de información sin la necesidad de que implique el intercambio de los propios recursos.

- Facilitan un acceso a los recursos en forma controlada ya que se conoce con precisión el objeto descrito.

- Preservan los objetos de información, permitiendo migrar sucesivamente éstos, para su posible uso por parte de las futuras generaciones.

- Coordinan búsquedas, integración y recuperación del conocimiento desde un mayor número de fuentes heterogéneas.

La clasificación de los diferentes modelos de metadatos, que se utilizan para codificar la información y permiten su recuperación, se muestran en la **Tabla 1**.

Tipo	Objetivo
Metadatos descriptivos	Descripción e identificación de recursos de información. En el nivel (sistema) local facilitan la búsqueda y recuperación. En el nivel web, permite a los usuarios descubrir recursos.
Metadatos estructurales	Facilitan la navegación y presentación de recursos electrónicos. Proporcionan información sobre la estructura interna de los recursos, incluyendo página, sección, capítulo, numeración, índices, y tabla de contenidos. Describen la relación entre los materiales. Unen los archivos y los textos relacionados.
Metadatos administrativos	Facilitan la gestión y procesamiento de las colecciones digitales tanto a corto como a largo plazo. Incluyen datos técnicos sobre la creación y el control de calidad. Incluyen gestión de derechos y requisitos de control de acceso y utilización. Información sobre acción de preservación.

Tabla 1: Tipos de Metadatos

1.8 Sistemas Gestores de Bases de Datos

Los Sistemas Gestores de Bases de Datos (SGBD) [1]son un tipo de software muy específico, dedicado a servir de interfaz entre la base de datos, el usuario y las aplicaciones que la utilizan. El propósito general de los sistemas de gestión de base de datos es el de manejar de manera clara, sencilla y ordenada un conjunto de datos que posteriormente se convierten en información relevante.

Existen tres grandes agrupaciones de sistemas gestores de base de datos: Los SGBD considerados productos **libres**. Dentro de este grupo se pueden encontrar 6 principales gestores. Constituyendo el más relevante PostgreSQL. La segunda agrupación son el conjunto de los gestores **no libres** donde resaltan como principales ORACLE, Microsoft SQLServer y MySQL. La tercera agrupación y más pequeña, los gestores considerados productos no libres y gratuitos con 2 principales gestores: Microsoft SQLServer Compact Edition y Sybase ASE Express Edition para Linux.[22]

[1] *En numerosas literaturas se conocen por las siglas DBMS, del inglés Data Base Management System.*

MySQL es un sistema de gestión de bases de datos relacional, fue creado por la empresa sueca MySQL AB, la cual tiene el copyright del código fuente del servidor SQL, así como también de la marca. Fue un software de código abierto, licenciado bajo la GPL de la GNU, comprado a principios de 2009 por la compañía productora de software propietario SUN MicroSystem, lo cual hace que el precio por la obtención de este producto sea casi inaccesible.[23]

Oracle es un sistema de gestión de base de datos relacional, desarrollado por Oracle Corporation surge a finales de los años 70 bajo el nombre de Relational Software a partir de un estudio sobre SGBD de George Koch. Computer World definió este estudio como uno de los más completos jamás escritos sobre bases de datos. Este artículo incluía una comparativa de productos que erigía a Relational Software como el más completo desde el punto de vista técnico. Esto se debía a que usaba la filosofía de las bases de datos relacionales, algo que por aquella época era todavía desconocido.[23]

PostgreSQL es un potente SGBD relacional (Open Source, su código fuente está disponible) liberado bajo licencia Berkeley software Distribución (BSD). Desarrollado en la Universidad de California, en el departamento de ciencias de la computación de Berkeley. Posee más de 15 años de activo desarrollo y arquitectura probada que se ha ganado una muy buena reputación por su confiabilidad e integridad de datos.[22] Es el gestor de código abierto más avanzado del mundo; lo que se refleja en que lo utilicen empresas como Yahoo, Greenplum, MyYearbook, Hi.5 y Facebook. [24]

1.9 Justificación de las herramientas a utilizar

Siguiendo la política nacional de migración hacia la independencia tecnológica, DATEC utiliza como **SGBD PostgreSQL**. Esta decisión ha sido previamente colegiada y aceptada por parte del cliente final debido a que dentro de sus políticas de migración se encuentran las de llevar todas sus bases de datos hacia dicha plataforma. En este sentido la versión seleccionada es la **8.4** por ser lo suficientemente estable y segura.

Entre las principales características que avalan la decisión de tomar PostgreSQL como SGBD figuran las siguientes:[25]

- Corre en más de 30 plataformas diferentes.
- Excelente documentación.

- Es sumamente adaptable a las necesidades propias.
- Soporta casi toda la sintaxis SQL. Soporte para los tipos de datos de SQL92, SQL99, SQL2003 y parte del SQL2008.
- Soporta llaves foráneas, tipos de datos definidos por el usuario, secuencias, relaciones, uniones, vistas, reglas, triggers, y procedimientos almacenados en múltiples lenguajes.
- Usa una arquitectura proceso-por-usuario cliente/servidor. Hay un proceso maestro que se ramifica para proporcionar conexiones adicionales para cada cliente que intente conectar a PostgreSQL.
- Tiene soporte para varios lenguajes procedurales internos incluyendo un lenguaje nativo denominado PL/pgSQL. Este lenguaje es comparable al lenguaje procedural de Oracle, PL/SQL. Posee habilidad para usar Perl, Python, Ruby o TCL como lenguaje procedural embebido, además de C, C++ y, Java.
- Puntos de recuperación a un momento dado, tablespaces, replicación asincrónica, transacciones jerarquizadas (savepoints), copia de seguridad en línea.
- Un sofisticado analizador/optimizador de consultas.
- Soporta juegos de caracteres internacionales, codificación de caracteres multibyte.

Incorporando en su versión más reciente:

- Alta capacidad de almacenar información, con una alta velocidad de respuesta ante consultas complejas y/o extensas.
- Restauración de base de datos en procesos paralelos, que acelera la recuperación de un respaldo hasta 8 veces con respecto a la versión 8.3
- Privilegios por Columna, que permiten un control más granular de datos confidenciales.
- Configuración de ordenamiento configurable por base de datos, lo cual hace a PostgreSQL más útil en entornos con múltiples idiomas.
- Nuevas Herramientas para Monitoreo de Consultas que le otorgan a los administradores mayor información sobre la actividad del sistema.

SE-PostgreSql

Además de utilizar las facilidades inherentes del SGBD PostgreSQL 8.4, se incluyen las funcionalidades brindadas por la herramienta SE-PostgreSQL elevando considerablemente el nivel de seguridad y control de acceso a la información.

SE-PostgreSQL proporciona un mecanismo para controlar el acceso a la información en una base de datos de manera detallada integrando las políticas de seguridad del SGBD al sistema operativo, logra no solo potenciar la seguridad y control a nivel de presentación de los datos, gestiona el acceso a los mismos por parte de otros usuarios o administradores dentro del mismo gestor. Esto posibilita proteger los datos incluso de otros administradores de la misma base de datos.

Además se puede configurar el control de acceso basado en filas, columnas, tablas y bases de datos. Si bien este nivel de personalización puede ser algo complejo en cuanto a su configuración para múltiples usuarios, el mismo es altamente recomendable debido a la flexibilidad que ofrece. Requiere para su utilización una correcta implementación del soporte SE-Linux en el servidor.

Visual Paradigm for UML

Es una herramienta CASE profesional para el desarrollo de aplicaciones, que integra el diseño visual, la generación de código fuente, bases de datos y documentación, abarcando todo el ciclo de vida del software. Entre las facilidades de uso que ofrece es necesario mencionar la ayuda para asegurar la consistencia en el nombrado, las definiciones de tablas y columnas y la generación de los objetos físicos mediante el lenguaje DDL. Además, provee mecanismos para estimar las consecuencias de los cambios y su impacto en los diagramas de análisis, así como la integración con múltiples herramientas de desarrollo. Se utiliza como herramienta de modelado en la solución del ODS.

Conclusiones

En concordancia con el estudio del estado del arte realizado puede concluirse que la tecnología adecuada para dar respuesta a la problemática en cuestión es el ODS, por constituir una solución definida, conocida y validada teóricamente. La metodología de desarrollo adoptada es la adaptación de la metodología de Kimball. Se determinó además que el modelo relacional cumple con los requerimientos necesarios para el diseño y estructuración del ODS; para cuyo desarrollo se seleccionan las siguientes herramientas: el Sistema Gestor de Base de Datos PostgreSQL en su versión 8.4 y Visual Paradigm for UML versión 6.4.

CAPÍTULO 2: DISEÑO E IMPLEMENTACIÓN DEL ODS

En el presente capítulo se abordan aspectos concernientes a la descripción de la solución. Se caracterizan los sistemas fuentes y los lineamientos base de la arquitectura del sistema; así como los componentes y otros elementos que la conforman. También se definen las áreas de análisis para la concepción de los reportes y objetos de estudio del CICPC. Además, se especifican el diseño e implementación de las estructuras de datos del ODS, las políticas de seguridad y la estrategia de copias de respaldo.

2.1 Descripción de las Fuentes de Datos

Las Fuentes de Datos son el punto de partida para la construcción de cualquier sistema de bases de datos, su selección depende fundamentalmente de las necesidades de los usuarios y son de especial importancia para el diseño y desarrollo de un ODS. Estas pueden agruparse en cuatro categorías fundamentales: Datos Internos, Datos Externos, Datos de Producción y Datos Archivados.[18]

Datos Internos

Son los datos que cada departamento, dentro de la organización, posee almacenados en archivos o bases de datos internas para auxiliarse en sus actividades.

Datos Externos

Son los datos que provienen de fuentes externas a la organización. Generalmente son informaciones compartidas entre competidores o entre proveedores y clientes.

Datos de Producción

Son los datos generados dentro de la organización en sus funciones diarias que resultan de interés para el ODS y se encuentran almacenados en los diferentes sistemas operacionales.

Datos Archivados

Son los datos provenientes de sistemas operacionales que se almacenan con el objetivo de llevar un histórico de la información en la organización.

Las fuentes de datos identificadas en el negocio de CICPC son los sistemas operacionales de SAIME, SAREN, DARFA, SENIAT, INTTT, SIGEP y SIGEPOL que entran en la clasificación de Datos Externos y el SIIPOL que se clasifica como una fuente de Datos de Producción.

2.1.1 Fuente 1: SAIME

SAIME es el sistema perteneciente a la Oficina Nacional de Inmigración y Extranjería de Venezuela, esta aplicación es la encargada de gestionar la información referente a las personas ceduladas en el país, así como a los extranjeros registrados en este sistema. Incluye datos personales, filiatorios, fotografía, registro dactilar, movimientos migratorios, entre otros.

El esquema para el Servicio de Notificaciones de datos que propone SAIME, permite que diferentes entidades externas puedan obtener información del sistema a partir de los esquemas o estructura de la información que se definan previamente. Cualquier entidad podrá tener un "buzón", sobre el cual se le permitirá consumir las notificaciones que el SAIME va generando como parte de los procesos de identificación que ocurren en los trámites definidos.

Actualmente la comunicación con este sistema se realiza mediante un sistema de FTP sobre SSL donde cada Entidad Externa, que solicita determinadas notificaciones, debe definir el esquema o estructura de la información que necesita obtener. Se crea un "buzón" por el protocolo FTPS (Secure FTP) con el nombre de la entidad, y dentro varias carpetas o directorios con el nombre de cada notificación que haya solicitado. En cada una de estas carpetas se escribirán los archivos compactados en formato "ZIP", que contienen archivos en formato "XML", con los datos relativos a cada notificación. Ejemplo de "buzón": "**ftps://ftp.saime.gob.ve/entidad**".

Los archivos compactados tienen su nombre en el formato: "Nombre del tipo de notificación _Fecha" (en el formato: yyyyMMddHHmmss) si se repiten archivos con la primera parte del nombre se agrega:"_" más un consecutivo de hasta 4 dígitos.zip", ejemplo: "**Cedulado_20080101123010.zip**".

Para la conexión externa se debe acordar previamente, entre la entidad y la institución, la dirección IP con la que se establecerá la comunicación para transmitir la información.

A la entidad externa se le crea un usuario que tendrá como nombre el propio nombre de la entidad, para hacer la conexión a su "buzón" FTPS, y se le emite un certificado digital que permita establecer una comunicación segura.

A medida que la entidad vaya consumiendo los archivos compactados con los datos de las notificaciones, se debe encargar de ir borrándolos con cierta frecuencia o en el mismo proceso en que los consume, para evitar que se acumulen los archivos en su "buzón".

2.1.2 Fuente 2: SAREN

Registro y Notarías de Venezuela cuenta con un sistema, conocido como SAREN, que maneja la información relativa a los inmuebles, compañías u oficinas así como lo concerniente a las operaciones y trámites que se realicen con dichas propiedades. Almacena además datos identificativos de sus propietarios, considerando que estos pueden ser personas naturales o jurídicas.

SAREN posee un ambiente de intercambio con otros entes vía FTP que es actualizado constantemente; sin embargo la comunicación con este sistema se materializará a través de servicios web [1]y copias locales de las Bases de Datos, para lo cual se definen los datos a ser intercambiados en documentos a entregar en cada una de las instituciones, exigiéndose además, en los caso que lo lleve, mecanismos de autenticación y restricción de uso de los mismos de manera dedicada entre las partes, que garanticen la seguridad y fiabilidad del intercambio.

2.1.3 Fuente 3: INTTT

INTTT es el sistema, perteneciente al Instituto Nacional de Tránsito venezolano, que gestiona la información relacionada con los vehículos incluyendo su certificado de origen y un historial de las placas, colores, tipos, usos, propietarios y seriales que ha tenido desde su fabricación y puesta en el mercado. Los datos almacenados incluyen además cualquier trámite que se realice con dichos vehículos así como el registro de las licencias de conducción adquiridas por cualquier persona. La comunicación con este sistema se posibilitará a través de los servicios web.

[1](en inglés *Web service*): es una colección de protocolos y estándares que sirven para intercambiar datos entre aplicaciones. Distintas aplicaciones de software desarrolladas en lenguajes de programación diferentes, y ejecutadas sobre cualquier plataforma, pueden utilizar los servicios web para intercambiar datos en redes de ordenadores como Internet.

2.1.4 Fuente 4: SIGEP

La Humanización Penitenciaria venezolana maneja la información de los individuos durante su tránsito por el Sistema Penitenciario, así como los datos recogidos en su expediente de Seguridad y Custodia, las sanciones disciplinarias que le han sido aplicadas, el historial de ubicaciones dentro del establecimiento penitenciario, su situación jurídica, sus salidas transitorias y traslados interpenales y otros datos registrados en aras de apoyar el cumplimiento de la legalidad durante los procesos penales y la ejecución de la sentencia. Todos estos datos son gestionados por la aplicación informática denominada SIGEP. La comunicación con dicho sistema se realizará a través de los servicios web.

2.1.5 Fuente 5: SENIAT

La Superintendencia Tributaria de Venezuela es otra de las organizaciones que genera, en sus operaciones diarias, datos que resultan de interés para las investigaciones policiales. Esta información se almacena en el sistema SENIAT y es referente a las empresas incluyendo ubicación geográfica, registro de información fiscal, relación de pagos efectuados, las declaraciones impuestos sobre la renta (ISLR) de la persona jurídica, si esta es un agente de retención del impuesto al valor agregado (IVA) y si ha realizado alguna declaración de IVA. La comunicación con este sistema se materializa a través de los servicios web.

2.1.6 Fuente 6: SIGEPOL

SIGEPOL contiene información acerca de los elementos de investigación, de los funcionarios de los cuerpos policiales con su respectiva ubicación policial y área de cobertura, las denuncias, las reseñas asociadas a un caso policial, las armas, vehículos y objetos denunciados.

La tecnología utilizada en la comunicación con este sistema son los servicios web, para ello se definen los datos a ser intercambiados exigiéndose mecanismos de autenticación y restricción de uso de los mismos de manera dedicada entre las partes, que garanticen la seguridad y fiabilidad del intercambio. Se definió el protocolo estándar SOAP en todas las aplicaciones que intercambiarán información mediante Servicios Web. Se utiliza el protocolo SSL para asegurar toda la información que se intercambie.

2.1.7 Fuente 7: DARFA

La dirección de Armamento de las Fuerzas Armadas venezolanas almacena información de las armas, dígase seriales, propietario, número del porte, marca, modelo, calibre y los trámites que con ellas se

realizan. A esto se le adicionan los datos de aquellas personas que han adquirido licencias para portarlas. Para la comunicación con el sistema DARFA tendrán que utilizarse, de igual modo que en la mayoría de las fuentes antes descritas, los servicios web.

2.2 Arquitectura del Sistema

De manera general, la arquitectura, dentro del desarrollo de software, es el diseño de más alto nivel de la estructura de un sistema o producto basado en reglas, objetivos y restricciones. Más específicamente, en la Tecnología Warehousing, es una forma de representar la estructura total de datos, comunicación, procesamiento y presentación, en función de los usuarios finales.

Ponniah la define como la estructura que unifica los componentes del DW, donde provee un marco general para su desarrollo y despliegue. Además define los estándares, mediciones, diseño general y técnicas de soporte.[18]

La propuesta de arquitectura del sistema describe todo el flujo de datos desde su extracción en los sistemas fuentes, hasta su preparación para la utilización por parte de los clientes del negocio. Para una explicación más detallada se describe atendiendo a cada subsistema que la conforma. Ver **Figura 4**.

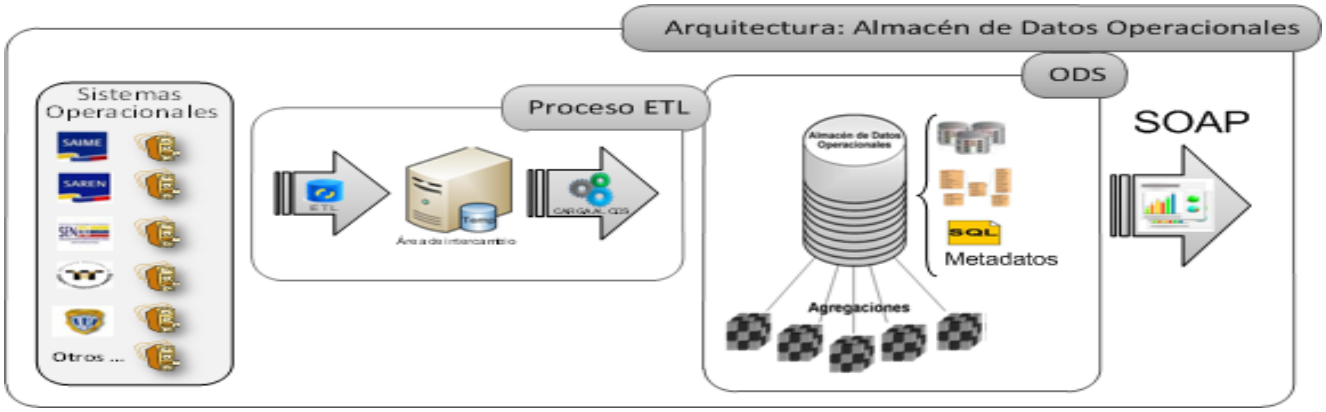

Figura 4: Arquitectura del Almacén de Datos Operacionales

2.2.1 Sistemas Fuentes Operacionales

Estos son los sistemas OLTP que poseen las compañías o empresas para la gestión de sus transacciones diarias. Estas transacciones son almacenadas en los más diversos formatos, desde una base de datos relacional hasta cualquier tipo de ficheros, ya sea Excel, XML, DBF, texto plano, entre otros.

Se encuentran localizados fuera del repositorio debido a que se tiene poco o ningún control sobre el volumen y formato de los datos de estas fuentes. Las prioridades principales de este componente son el procesamiento, el rendimiento y la disponibilidad. Generalmente realizan salvas de la información que gestionan y sólo trabajan con los datos generados en un período corto de tiempo para hacer las recuperaciones de forma más óptima. Para el ODS de SIIPOL los sistemas fuentes son los sistemas operacionales externos a CICPC.

2.2.2 Subsistema de Integración

Un ODS se alimenta de un conjunto de sistemas en los cuales existe información no integrada, cada uno con estándares diferentes y en plataformas diversas, razón por la cual cobra especial importancia el subsistema de integración. Este tiene la responsabilidad de recopilar los datos necesarios consumiendo los servidores FTP o servicios web implementados en los sistemas fuentes. Posterior a este proceso de extractar los datos, estos se cargan al área de intercambio donde se almacenan antes de ingresar al ODS. Una vez los datos estén almacenados en bases de datos temporales se procede a su depuración y limpieza donde se detectan inconsistencias, duplicaciones, errores de formato e inexistencias, estandarizándose la información almacenada en diferentes fuentes. Se realizan además procesos de transformación de acuerdo a las reglas definidas en el negocio, todo esto previo a la alimentación de las estructuras de nivel detallado.

2.2.3 Subsistema de Almacenamiento

Este subsistema es responsable de la centralización, seguridad y respaldo de los datos provenientes de todos los sistemas operacionales. Realiza las operaciones relacionadas con la gestión de los datos dentro del almacén utilizando herramientas específicas que realizan operaciones como la transformación de datos para su incorporación en las tablas, la creación de índices y vistas de las tablas base, creación de copias de seguridad y archivado de datos, además del análisis de los datos para garantizar la coherencia de los mismos.

2.3 Arquitectura de Componentes del Sistema

La arquitectura general consta básicamente con tres escenarios bien definidos, un componente para la administración, el componente ETL y el componente de almacenamiento, ver **Figura 5**.

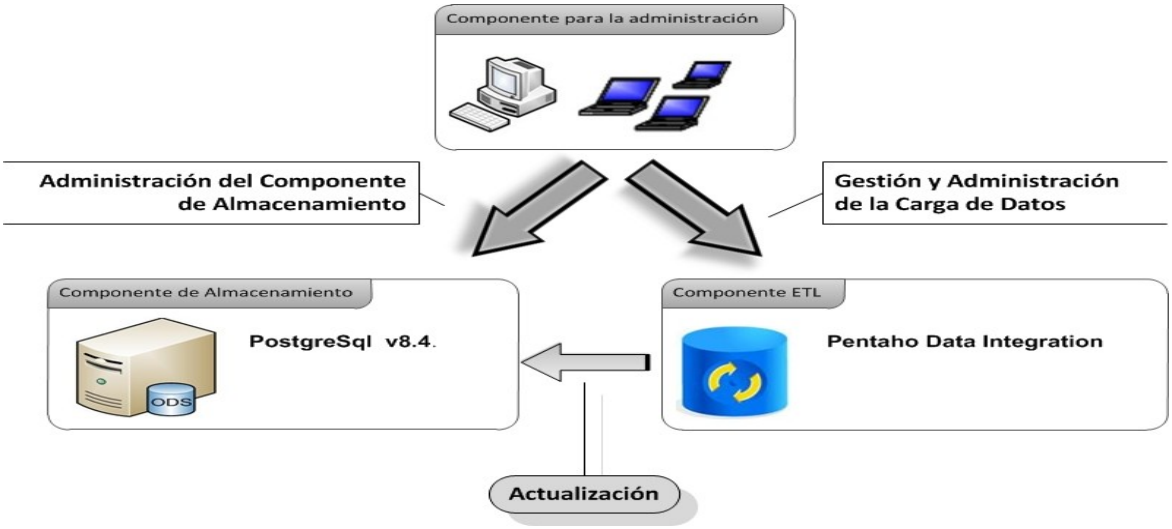

Figura 5: Arquitectura de Componentes

Para describir el ambiente de desarrollo de la solución propuesta se va a dividir en las siguientes secciones:

- Servidor Proceso ETL.
- Repositorio Central.
- Clientes para la Administración.

En cada una de las secciones existen un conjunto de herramientas que soportan cada proceso.

2.3.1 Componente ETL

El proceso ETL es el encargado de impulsar el flujo de datos haciendo transformaciones intermedias y permitiendo una integración exitosa. Es por esto que cada paso, desde su diseño hasta la puesta en marcha, precisa de la mayor atención y esfuerzo posible. Este componente permite mover datos desde múltiples fuentes, reformatearlos, limpiarlos, y cargarlos en otra base de datos, protegiendo su linaje. También realiza la actualización del almacén (operación periódica que propaga los cambios de las fuentes externas al ODS).

ETL se encuentra enfocado a la integración de datos, tanto por lote, como en tiempo real, logrando un alto grado de transformaciones para la consolidación de la misma. El procedimiento fortalece los datos para su

utilización en bases de datos permanentes, dedicadas para el análisis o la generación de informes, como es el caso de repositorios de datos, data marts y almacenes de datos, entre otras funcionalidades, que hacen de este proceso imprescindible a la hora de integrar datos. [21]

Entre las características que describen las tecnologías de ETL se tienen:[26]

- Es un mecanismo de carga muy eficiente y efectivo orientado a los DW.
- Enfocado a migrar y mezclar datos.
- Reduce la exposición a desarrollos manuales (codificación) producto de la existencia en el mercado de herramientas potenciales para la implementación visual, con manejo de excepciones, gestión y planificación de tareas.
- Necesita pocos servicios de administración y mantenimiento.
- Gran capacidad para llevar a cabo transformaciones.
- Tecnología enfocada a la integración de datos desde fuentes versátiles hacia los DW.

Este proceso resulta complejo, pues precisa de un alto nivel de detalle, y en caso de ser mal diseñado puede provocar serios problemas operativos. En la **Figura 6** se presenta la arquitectura definida para el tipo de solución en cuestión.

Figura 6: Arquitectura de la solución ETL

Para los nodos del componente ETL se propone el sistema operativo Debian 5 con la herramienta Pentaho Data Integration 3.2 y se requiere la máquina virtual de java 6, actualización 2.

2.3.2 Componente de almacenamiento

El componente más importante y es sobre el cual se basa el sistema es el repositorio central, la estructura del mismo está compuesta por el Gestor de Base de Datos PostgreSQL 8.4.1 que es donde va a estar desplegado el sistema sobre el Sistema Operativo Debian 5.0. Además se utilizan algunas herramientas de PostgreSQL como PgPool, para el control de altas concurrencias sobre el sistema, y Slony-I en el proceso de réplica de datos orientado a la alta disponibilidad.

2.3.3 Componente Administración

El componente de administración es responsable de dar mantenimiento, respaldo y actualización al ODS. Está conformado por una o varias estaciones para la administración soportadas sobre cualquier sistema operativo que cuente con las herramientas necesarias para este proceso. Se propone la utilización de la herramienta Pentaho Data Integration 3.2.2 para la gestión y administración de todo el proceso de extracción, transformación y carga de los datos y el PgAdmin III v1.10 para la administración y mantenimiento del componente de almacenamiento. Se requiere además la máquina virtual de java en su versión 6, segunda actualización.

2.4 Arquitectura de la Base de Datos

Las estructuras diseñadas están definidas en dos niveles de agrupación y concentradas en una misma instancia de la base de datos. Por un lado están las estructuras con los datos detallados, metadatos y base de datos intermedia y en un segundo nivel se encuentran las agregaciones diseñadas en función de los reportes más comunes. Vale destacar que la carga de los datos hacia estas estructuras se realiza mediante funciones definidas dentro del mismo gestor. En la **Figura 7** se especifica la arquitectura interna dentro del repositorio.

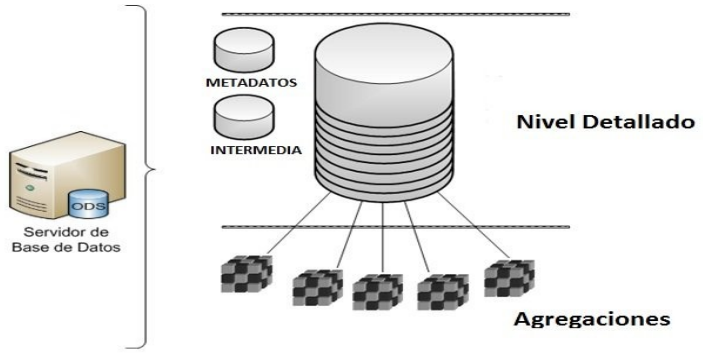

Figura 7: Arquitectura interna de la BD

El nivel detallado consta de 3 particiones específicamente. Una partición de Metadatos con la información de las tablas del repositorio central, todo sobre los datos necesarios para la administración y uso del ODS. Una partición Intermedia que sirve como puente para la integración de los datos hacia el ODS. Y un conjunto de esquemas de datos detallados que tienen como función principal actualizar las agregaciones definidas en función de los principales reportes.

El nivel de las agregaciones contiene la información pública del sistema, sobre este nivel se implementan vistas, funciones y otros mecanismos orientados a la publicación de la información.

2.5 Definición de las áreas de análisis

Las áreas de análisis (AA) se definen con el objetivo de garantizar la factibilidad, utilidad y de manera general el éxito de las estructuras que se están diseñando. En la solución propuesta se han definido 2 AA en concordancia con las necesidades de información de los clientes que a continuación se señalan:

- Obtener información sobre operaciones con inmuebles, vehículos, armas u otras propiedades llevadas a cabo por una persona o un conjunto de ellas involucradas en algún proceso judicial.
- Obtener información sobre los vínculos familiares y afectivos de personas involucradas en un caso policial.
- Obtener información sobre las entradas y salidas del país de personas involucradas en un caso policial, así como cualquier restricción que tenga de entrar a un país.

- Obtener información sobre el comportamiento de los funcionarios del CICPC y SIGEPOL así como de cualquier miembro del cuerpo policial.

Las AA identificadas se denominaron:

- Incidencias Delictivas.
- Movimientos Migratorios.
- Compra y Venta de Propiedades.
- Captura de armas.
- Compra y Venta de Vehículos.
- Actuar Policial.

2.6 Diseño del ODS

El diseño de bases de datos consta de tres etapas: diseño conceptual, lógico y físico. Los modelos de datos conceptuales o de alto nivel son los modelos orientados a la descripción de estructuras de datos y restricciones de integridad, disponen de conceptos muy cercanos al modo en que la mayoría de los usuarios percibe los datos; los modelos de datos lógicos, son orientados a las operaciones más que a la descripción de una realidad y los modelos de datos físicos son estructuras de datos a bajo nivel implementadas dentro del propio manejador y proporcionan conceptos que describen los detalles de cómo se almacenan los datos en el ordenador.

A continuación se describen los pasos llevados a cabo para el desarrollo del diseño lógico del ODS, en esta etapa, partiendo del análisis e identificación de los principales conceptos asociados a las necesidades informacionales de los clientes, se obtienen un conjunto de relaciones o tablas que representan los datos de interés. Mediante este proceso se construye un esquema que representa la información que manejan los distintos sistemas y que resulta interesante para SIIPOL, basándose en un modelo lógico relacional, pero independiente del SGBD concreto que se vaya a utilizar para implementar la base de datos y de cualquier otra consideración física.

2.6.1 Conceptos identificados

Al analizar los requerimientos de información para establecer los conceptos fundamentales representados en el almacén, partiendo de las solicitudes de los clientes y considerando la disponibilidad informacional en las fuentes de datos, se ha llegado a la identificación de los siguientes elementos como entidades del sistema.

Concepto Persona: La persona, puede ser natural o jurídica, constituye el eje fundamental para el análisis y gestión de información dentro del sistema a modelar, de la misma se conocen elementos descriptivos tales como: nombre, apellido, cédula, dirección, relaciones personales, licencias entre otras.

Concepto Propiedad: La propiedad sobre algún inmueble resulta de interés en determinadas investigaciones, especialmente los trámites que con ella se realicen, la dirección en la que esté ubicada y sus propietarios.

Concepto Expediente: El expediente está fuertemente vinculado a una persona, almacena todo su historial delictivo, los procesos penales en los que se ha visto involucrada, las condenas que ha cumplido o cumple por estos, las denuncias que se le han implicado, diagnósticos mentales e informes periciales asociados, si se encuentra o se ha encontrado recluida en algún centro penitenciario, en cuál o cuáles, fechas, etc.

Concepto Vehículo: Del vehículo resultan interesantes algunas características como su placa, seriales, colores, los trámites efectuados, propietarios, el histórico, sus características principales, seriales, colores, marca, el modelo, el tipo, el uso, etc.

Concepto Arma: Las armas tienen especial importancia en las investigaciones policiales, estas pueden pertenecer a funcionarios de CICPC ó SIGEPOL, a cualquier miembro del cuerpo policial e incluso a personas civiles que posean una licencia para portarlas. Pueden ser identificadas por sus 3 seriales (primario, secundario y terciario), marca, modelo y calibre. Los trámites asociados a ellas pueden ser claves para el esclarecimiento de determinados hechos delictivos o su localización.

Concepto Objeto: Se denomina objeto a cualquier elemento asociado a un delito, encontrado en una escena del crimen, decomisado, es decir, por alguna razón se encuentra bajo la custodia de las entidades policiales y por supuesto resulta necesario investigar ciertos puntos como por ejemplo: su dueño si se

conoce, la marca, modelo, tipo, tamaño, peso, color, valor desde el punto de vista del dinero que se necesita para adquirirlo, su descripción, etc.

Concepto Trámite: El trámite como concepto se refiere a las operaciones de compra, venta, traspaso, u otras que pueden hacerse con los inmuebles, vehículos y armas, cada trámite tiene sus particularidades y sus características generales como la fecha en que se hizo, en qué oficina, la dirección de dicha oficina, quién fue el beneficiario, etc.

Concepto Viaje Internacional: En una investigación policial pueden ser objeto de análisis y seguimiento los movimientos migratorios asociados a una o varias personas, conocer en qué fechas viajó, de qué lugar partió, hacia donde se dirigió, cualquier observación referente al viaje: si el pasajero tiene restricciones de acceso a determinado país ó si pasó algo relevante que merezca ser estudiado.

Derivados a estos conceptos existen un conjunto de elementos que los describen o complementan, los cuales constituyen también objetos de interés y son representados como entidades persistentes.

2.6.2 Diseño del modelo de datos del sistema

El siguiente paso es convertir los esquemas conceptuales locales en un esquema lógico global. El objetivo del diseño lógico es obtener una representación que use, del modo más eficiente posible, los recursos que el modelo de SGBD posee para estructurar los datos y modelar las restricciones.

El modelo relacional carece de ciertos rasgos de abstracción que se usan en los modelos conceptuales. Por lo tanto, el primer paso en la fase del diseño lógico consiste en la conversión de esos mecanismos de representación de alto nivel en términos de las estructuras de bajo nivel disponibles en el modelo relacional.

Tablas existentes de la BD

La el modelo de datos de la solución para el nivel detallado está conformado por 34 tablas como resultado de la integración de 7 fuentes. Estas tablas fueron definidas a partir de los conceptos previamente identificados y de las relaciones que entre ellos se establecen. Entre las más importantes figuran:

Tabla Persona Natural (***"tb_personanatural"***): Esta tabla describe a la persona natural mediante el conjunto de atributos que la identifican como son nombres, apellidos, letra y número de cédula, pasaporte,

nacionalidad, etc. Constituye una de las principales fuentes de información para la realización de las estructuras que soporten los reportes más comunes.

Tabla Persona Jurídica (***"tb_personajuridica"***): Al igual que la tabla persona natural, esta tabla surge a partir del concepto persona previamente identificado. Recoge todas las propiedades de este tipo de persona, por ejemplo, su registro de información fiscal.

Tabla Viajes Internacionales (***"tb_viajesinternacionales"***): Esta tabla agrupa todas las características de los viajes internacionales registrados.

Tabla Vehículo (***"tb_vehiculo"***)***:*** Describe los vehículos con todas sus propiedades, dígase marca, modelo, seriales, certificado de origen, entre otras. Se relaciona de manera directa con la tabla persona.

Tabla Arma (***"tb_arma"***)***:*** Reúne todas las características de las armas, su calibre, seriales, marca, modelo, etc. Se relaciona de manera directa con la tabla persona.

Tabla Objeto (***"tb_objeto"***): Describe los objetos, que por una u otra causa están bajo la custodia de las fuerzas policiales, mediante características cómo su color, tamaño, forma, clasificación y demás.

Tabla Propiedad (***"tb_inmueble"***)***:*** Agrupa todas las cualidades que describen al concepto propiedad. Se relaciona de manera directa con la tabla persona.

Tabla Expediente (***"tb_expediente"***)***:*** Posee todas las propiedades o atributos descriptivos del expediente policial. Se relaciona de manera directa con la tabla persona.

Tabla Trámite Persona Vehículos (***"tb_tramitepersonavehiculos"***)***:*** Se define a partir del concepto Trámite y describe los trámites que las personas realizan con los vehículos.

Tabla Trámite Persona Armas (***"tb_tramitepersonaarmas"***): También se deriva del concepto Trámite relacionando aquellos trámites realizados con las armas por parte de las personas.

Tabla Trámite Persona Propiedades (***"tb_tramitepersonapropiedades"***)***:*** Surge a partir del concepto Trámite y reúne las características de las operaciones que las personas realizan con sus propiedades.

Una vez definidas las tablas con sus atributos y relaciones se procede a la estructuración del modelo relacional. Con este fin cada tabla posee una llave primaria encargada de mantener la integridad referencial entre ella y otras tablas. Ver **Figura 8**.

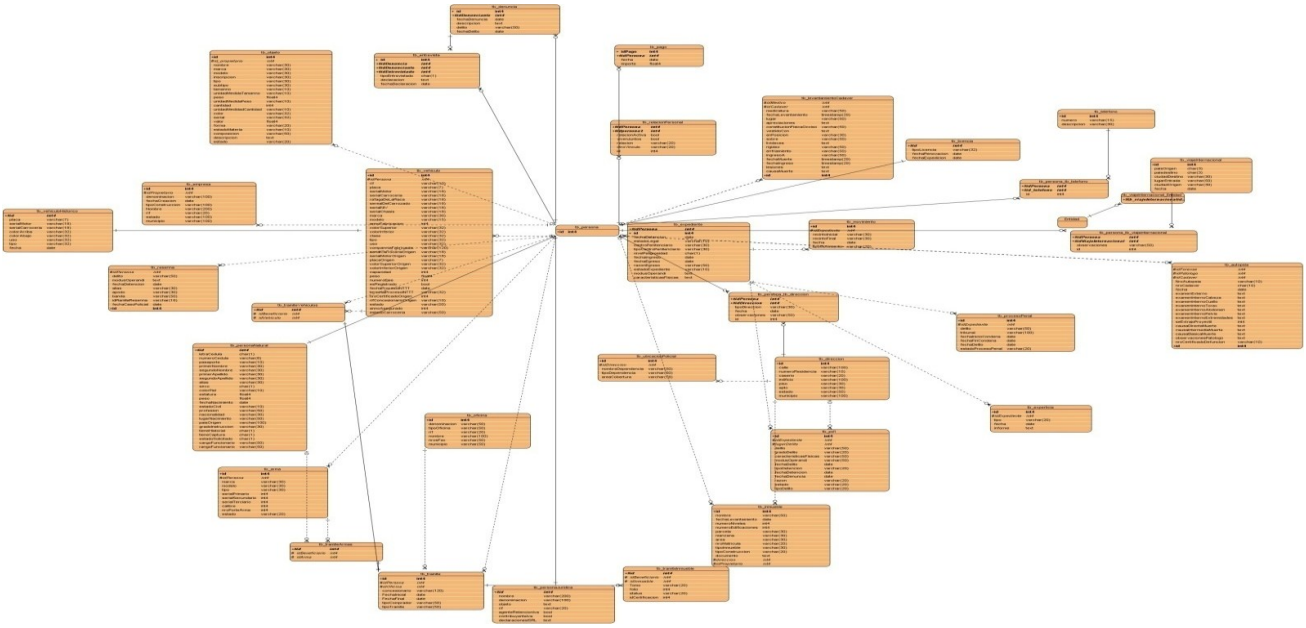

Figura 8: Modelo de Datos ODS-SIIPOL

2.7 Implementación del ODS

Luego de haber realizado el diseño del modelo de datos relacional que conforma el ODS se procede a la implementación física del mismo. En este punto se habla de la forma de llevar a cabo todo lo diseñado en el modelo lógico y plantearlo en la realidad, para lo cual es necesario el desarrollo de los estándares y del modelo de datos físico, la definición del plan inicial de índices, el diseño y construcción de la instancia de la base de datos y el diseño de la estructura física de almacenamiento.

2.7.1 Estandarización de nombres

Esencialmente existen 3 componentes básicos de Nombres de Objetos de Bases de Datos: palabra primaria (*prime word*), clases de palabras (*class words*) y calificadores (*qualifiers*). Definiciones de elementos de datos, nombres de elementos lógicos de datos y nombres de elementos físicos de datos están todos compuestos por los tres componentes básicos

Palabra primaria: Describe el elemento de datos de la materia. Algunos ejemplos de esta categoría son: clientes, productos, cuentas, ciudades, regiones, esquema, regiones, etc. Cada palabra primaria debe ser clara, definición inequívoca que va la derecha en la información de catálogo.

Clases de palabras: Describe la mayor clasificación de los datos asociados con cada elemento. Algunos ejemplos son: total, cantidad, fecha, bandera, nombre, descripción y número.

Clasificadores: Los clasificadores son elementos opcionales que pueden describir o definir más las Clases o Palabras Primarias. Existen algunos ejemplos como: inicio, final, histórico, actual, etc.

En la solución propuesta, a nivel global, se mantuvo la misma estructura en cuanto a la Clasificación, específicamente en lo referente a si la estructura es una tabla, un esquema o una base de datos. Si es una tabla al nombre le precede las letras ***"tb"*** ejemplo ***tb_persona***, en caso de ser un esquema donde se organizan tablas, secuencias y funciones se le antepone las siglas ***"ods"*** seguido del nombre de la fuente, ejemplo, ***ods_saime*** y cuando es una agregación se le especifica con las letras ***"agg"*** ejemplo ***agg_viajes_internacionales***.

En el caso de los atributos de las tablas se siguió la misma política en todas. Cuando se refiere a la llave primaria se le denominó ***"pk_id"*** seguido del nombre de la tabla, en caso que fuera llave foránea se especificó ***"fk_id"*** seguido del nombre de la tabla fuente. Así mismo con respecto a las funciones en el caso de aquellas que se implementan para el control de cambios la nomenclatura definida es ***"fc_tgr_control_cambios"*** ejemplo ***fc_tgr_.***

Igualmente se nombraron los scripts ya sea para la creación como para el llenado de cada una de las tablas. Estos nombres están ordenados por las siglas "DDL" (*data definition language*, en inglés) y el nombre de la acción ejemplo ***DDL_metadatos_ODS***, al referirse a la estructura del ODS.

Al finalizar este paso queda completamente estructurado la nomenclatura utilizada para la denominación de las tablas, atributos y scripts dentro de la base de datos. Ya después de tenerlos definidos se comienza con la implementación de las estructuras físicas. Ver **Anexo 3**.

2.7.2 Desarrollo del modelo físico

En esta etapa, se parte del esquema lógico global obtenido durante el diseño lógico y se obtiene una descripción de la implementación de la base de datos. Esta descripción es completamente dependiente

del SGBD que se vaya a utilizar, en este caso PostgreSQL 8.4. El modelo físico incluye algunos cambios en las estructuras de las tablas para ajustarlas al SGBD, contiene además las tablas de mantenimiento que usualmente no se incluyen en el modelo lógico y detalla las características físicas de la base de datos, desde los tipos de datos hasta la segmentación, los parámetros de almacenamiento de tablas y bandas de discos.

Una vez que se ha concluido el proceso de carga del modelo físico en la herramienta, se da paso a las tareas de personalización de las estructuras físicas en función de la estandarización de formatos, nombres de objetos, corrección de relaciones, se corrige la utilización de algunos datos como son: los nombres de los atributos de las tablas, los tipos de datos, la cantidad de caracteres que soporta, y la documentación de cada una de las estructuras y atributos, con el objetivo de concretar aún más el modelo de datos.

La actividad final de este paso es la estimación inicial del tamaño de la base de datos. Para los desarrolladores de almacenes de datos es realmente crítico el saber cuánto va a almacenar con el fin de utilizar el impacto de esta variable en el rendimiento del sistema.

Las longitudes de las filas afectan significativamente el tamaño de una base de datos debido a que ejemplos como las cadenas VARCHAR no siempre son explotadas en su totalidad y el SGBD si les almacena espacio como si estuvieran totalmente llenas. Igualmente pasa con los campos vacíos o null, la inclusión de estos campos en la base de datos aumenta su tamaño y generalmente no es información útil para el usuario final. En el sistema propuesto se redujo a cero la cantidad de campos null dentro de la BD no siendo así con los campos VARCHAR debido a que la información de las tablas, en ocasiones, son oraciones completas y se hace muy difícil acotar el tamaño, como alternativa se establecieron los tamaños lo más pequeño posible para reducir al máximo la cantidad de espacios sin utilizar dentro del campo.

2.7.3 Estrategia inicial de indexado

Sobre el ODS se realizan, muchas veces, consultas de gran complejidad que solicitan información que cumpla determinados criterios, es decir, los usuarios frecuentemente querrán especificar los valores con los cuales se filtrarán los datos que deben ser retornados. La mayoría de estas consultas incluyen operaciones de join entre tablas muy grandes, lo cual puede resultar extremadamente costoso. Para ganar en eficiencia a la hora de realizar estas operaciones se han investigado y creado técnicas especializadas que hoy ofrecen varios gestores, como los índices.

Un índice es una estructura física que permite un tipo de acceso alternativo al secuencial. Es creado a partir de una o varias columnas de una tabla, y, por lo general, es construido en forma de árbol balanceado (B-Tree). Al ser estructuras físicas, los índices van a tener un fichero asociado, en cuyas páginas se pueden almacenar uno o varios nodos del árbol. Cada uno de ellos apunta hacia otros nodos o hace referencia a las filas de la tabla. En cada nodo, los valores están ordenados, y los que se encuentran en un nodo hijo son menores o iguales que el valor en el nodo padre que le hace referencia.

Existe un tipo de índice con el cual se impone que los datos de la tabla estén ordenados en el nivel físico, reciben el nombre de índices clusterizados. Para cada tabla sólo se puede especificar un índice clusterizado, pues este afecta la forma en que son almacenadas las filas. Aquellos que no influyen en la organización física se denominan índices no clusterizados y varios pueden ser creados para una misma tabla. [26]

Las ventajas que tiene el uso de los índices están dadas, precisamente, por su estructura. Por ejemplo, las búsquedas de filas en las que un valor en particular aparezca no implican recorrer toda la tabla, sino que se utiliza la estructura arbórea del índice que se haya definido. Con esto se consume menos tiempo en hallar el resultado y es menor la cantidad de veces que se accede al disco para leer.

Sin embargo, no deben crearse estas estructuras innecesariamente pues aunque permiten especificar caminos de acceso adicionales para las relaciones base, hay que tener en cuenta que conllevan un coste de mantenimiento que hay que medir frente a la ganancia en prestaciones.

Un detalle a tener en cuenta es que los índices se almacenan en ficheros al igual que los datos de una tabla, por tanto, van a ocupar espacio de almacenamiento físico. Mientras más grande sea una tabla, mayores serán los índices asociados a ella, de manera que resulta necesario analizar la capacidad de almacenamiento disponible.

Además, la creación de demasiados índices puede traer consecuencias no deseadas, por ejemplo, si se modifican valores asociados a columnas sobre las que se hayan creado índices, o se insertan o eliminan filas, la estructura del índice se actualiza, pues el árbol asociado debe ser consistente con respecto a la información de la tabla. Esto va a influir en el comportamiento del gestor, pudiendo reducir la velocidad de procesamiento a la hora de realizar dichas operaciones. Este detalle debe tenerse en cuenta a la hora de realizar las cargas hacia el sistema, donde las operaciones de inserción y modificación son abundantes.

Una alternativa que se podría analizar es eliminar los índices antes de comenzar la carga y volverlos a crear después.

Una solución apropiada consiste en definir solo aquellos índices que impliquen una mejora significativa en el rendimiento del sistema ante consultas. Algunas instrucciones que se pueden seguir son:

- Construir un índice sobre la llave primaria de cada relación base.
- No crear índices sobre relaciones pequeñas.
- Añadir un índice sobre los atributos que se utilizan para acceder con mucha frecuencia.
- Añadir un índice sobre las llaves foráneas que se utilicen con frecuencia para hacer joins.
- Evitar los índices sobre atributos que se modifican a menudo.
- Evitar los índices sobre atributos poco selectivos (aquellos en los que la consulta selecciona una porción significativa de la relación).
- Evitar los índices sobre atributos formados por tiras de caracteres largas.

La mayoría de los Sistemas Gestores de Bases de Datos proporcionan herramientas de prueba y evaluación para determinar la efectividad de un índice, con las cuales, luego de creado, se puede comprobar si traerá mejoras significativas en el sistema.

Para agilizar consultas complejas que involucran operaciones de *joins* entre tablas grandes, han sido desarrolladas formas especiales de índices. Algunos gestores los han incorporado, permitiendo lograr mayor eficiencia en los tiempos de respuesta ante solicitudes con propósitos analíticos.

Los índices multitabla o índices *join*, por ejemplo, permiten definir índices sobre columnas de dos o más tablas. Desde el punto de vista físico, la modificación con respecto a los índices antes explicados es que las referencias de las páginas hojas apuntan a varias filas en tablas diferentes. Esto mejora notoriamente las operaciones de unión donde participen dichas columnas.

Otros índices son los de columnas virtuales, también denominados índices basados en funciones, que dan la posibilidad de definir índices sobre una expresión más allá que sobre columnas. La principal ventaja que

ofrecen es la mejoría de la velocidad de procesamiento en consultas donde se utilice la expresión para filtrar. Existen también otras formas especiales de índices que se basan en estructuras diferentes del **B-Tree**, como el *índice Hash* y *el Bitmap*, este último utilizado generalmente cuando la cardinalidad de la columna es baja.

El estudio de los índices ha sido y continúa siendo un campo en desarrollo. A medida que surgen nuevas necesidades informativas y las consultas van ganando en complejidad, se hacen necesarias estas técnicas de optimización, con el fin de mejorar el comportamiento de los sistemas ante las solicitudes.

Cada tipo de índice generalmente está enfocado a hacer eficientes las consultas, pero teniendo en cuenta los datos almacenados, su cantidad y variabilidad, factores que influyen a la hora de tomar la decisión de qué índices definir.

La solución posee implementado como indexado el que trae por defecto el gestor PostgreSQL, para la búsqueda de datos utilizando las llaves primarias, foráneas y campos únicos. Todas las llaves primarias, que son llaves subrogadas además, poseen índices de tipo “b-tree” (Árboles-B), de esta manera, cualquier búsqueda que se realice utilizando las llaves se optimiza mediante este método.

2.7.4 Diseño y construcción de la instancia de la Base de Datos

El diseño y construcción de la instancia de la Base de Datos tiene como objetivo principal garantizar la existencia de los requerimientos físicos necesarios para el buen funcionamiento del ODS. Uno de los parámetros más importantes es la disponibilidad de memoria, ya que en el ODS se realizan numerosas y complejas consultas dónde se tienen que unir un número considerable de tablas físicas, almacenándose en memoria para ser ofrecidas al usuario final. Esto implica que la solución propuesta necesite de al menos 2Gb de memoria para garantizar un rendimiento óptimo a las peticiones de información que se le soliciten. Otro parámetro a tener en cuenta es el procesador que tendrá el servidor físico, para la solución en cuestión con un procesador Pentium IV cumpliría con los requerimientos básicos.

2.7.5 Desarrollo de la estructura física de almacenamiento

Los SGBD almacenan las tablas en ficheros, cada fichero está dividido en páginas y en cada una de ellas se puede almacenar un número fijo de filas. Este número lo van a determinar dos factores: el tamaño de la

página[1] y la longitud de las filas. Las páginas no tienen por qué estar completamente llenas, sino que pueden contener espacios en blanco, originados, posiblemente, por eliminaciones de filas previamente almacenadas. Para la inserción de nuevos datos son usadas diferentes técnicas. Algunos gestores asumen que siempre una fila nueva es añadida detrás de la última fila en la página final. En caso de que dicha página esté llena, una página vacía es agregada al fichero y en ella se colocan los nuevos datos. Otros gestores llenan de manera automática los espacios intermedios, haciendo uso de algoritmos de manejo de páginas más sofisticados.

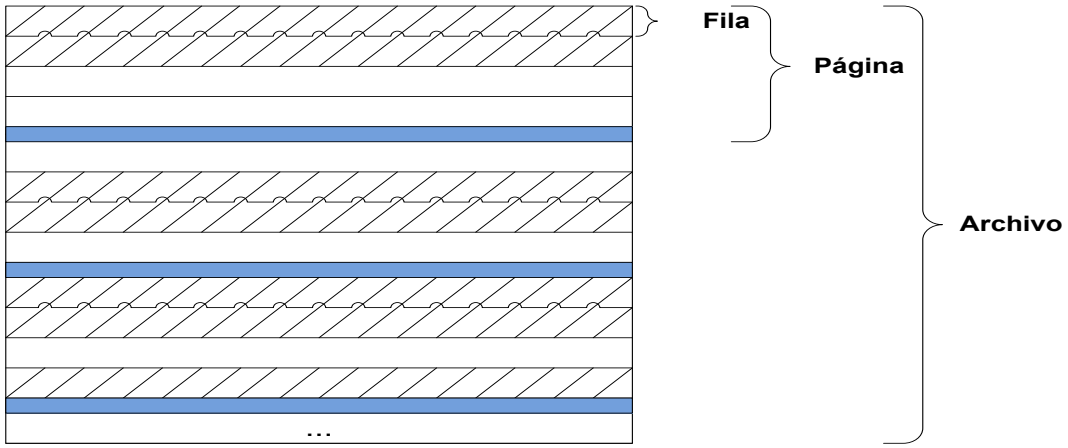

Figura 9: Posible estructura de un fichero donde se almacenan datos

Es importante aclarar que las páginas forman la unidad de Entrada/Salida: si el Sistema Operativo recupera datos en el disco duro, esto es hecho página por página. Por lo tanto, un gestor de bases de datos puede solicitar al Sistema Operativo una página dentro del fichero, pero sólo una fila. Para lograr recuperar una fila, son necesarios dos pasos: primero, debe recuperarse del disco la página donde se encuentra dicha fila. La manera de identificar una página es usando el identificador del fichero donde se encuentra, más el número que tiene esa página dentro del fichero. Luego, se debe buscar la fila dentro de la página.

Este último paso puede hacerse de diferentes maneras. Una vía es examinar toda la página hasta encontrar la fila deseada. Como este proceso se realiza completamente en memoria, es llevado a cabo

[1] El tamaño de la página depende del Sistema Operativo y del Gestor de Bases de Datos. Tamaños como 2K, 4K, 8K y 32K son muy comunes.

relativamente rápido. Otro método más directo es el siguiente: cada página puede contener una lista de entidades enumeradas en las que se puede encontrar la ubicación de todas las filas almacenadas en dicha página. Cada fila tendrá, entonces, un identificador único, formado por el identificador de la página y el número que tiene esa fila en la lista. Por lo tanto, el procedimiento para encontrar una fila sería: primero seleccionar la página correcta y luego acceder a la ubicación especificada en la lista para el identificador de dicha fila.

Como se ha visto, manejar la información almacenada implica efectuar distintas operaciones en el nivel físico, en las que se combinan las funciones a realizar en disco, con las capacidades en cuanto a memoria y CPU.

En este punto se deben tener en cuenta elementos como el particionamiento de las tablas, en función de lograr una mayor organización de la información y velocidad en su recuperación, y estructuras de control de cambios. En el repositorio central se definió un Sistema de Control de Cambio que, mediante tablas de metadatos, almacena los cambios realizados con el fin de minimizar la utilización de recursos físicos cuando se refresquen las agregaciones. La tabla definida para este objetivo se puede ver en el **Anexo 4**

2.7.6 Esquemas definidos

A la descripción de la estructura de una base de datos, en un lenguaje formal soportado por un SGBD, se le denomina esquema de la base de datos. Estos se especifican durante el diseño, y su modificación es poco frecuente. En el esquema de una base de datos relacional, se definen las tablas, sus campos y relaciones. Los esquemas definidos para la solución propuesta se describen a continuación:

Un esquema por cada fuente denominado “ods_*nombre de la fuente*” por ejemplo *ods_saime* donde se ubican las tablas cuyos datos provienen fundamentalmente de la fuente en cuestión, en este caso SAIME. El esquema “ods_agg” donde aparecen las tablas de las agregaciones y sus funciones de refrescamiento. El esquema “ods_cc_auditoria_part” donde se encuentra la tabla relacionada con el control de cambios y mecanismos de auditoría y control de datos.

Las estructuras diseñadas para el control de cambios están basadas específicamente en una tabla de metadatos cuya finalidad es registrar cada acción de inserción o actualización realizada sobre las tablas del repositorio central de interés para el seguimiento. Se excluye el proceso de eliminación debido a que el mismo no estará disponible según lo establecido en las políticas y reglas del negocio. Se garantiza

mediante este mecanismo que cuando se ejecuten las funciones de refrescamiento sólo se procesen los valores nuevos o modificados minimizándose de esta forma el tiempo y la utilización de recursos del servidor.

Las agregaciones constituyen consolidados de información, orientados a los reportes más comunes que relacionan la información en dependencia de las tablas que interactúan en ellas. Debido al gran volumen de información que se manipula en un ODS se le deposita gran importancia a la utilización de las mismas.

A continuación se relacionan las agregaciones que se conformaron sobre la base a las necesidades de información detectadas en SIIPOL. Vale aclarar que el llenado de dichas tablas no se realiza desde las fuentes sino desde la información almacenada en el repositorio central mediante funciones de refrescamiento implementadas con este objetivo.

Agregación Viajes Internacionales (agg_viajesInternacionales):

La agregación Viajes Internacionales tendrá como objetivo fundamental presentar toda la información de interés a analizar en cuanto a algunas características importantes sobre las personas y los viajes, tanto nacionales como internacionales, en los que se ven involucradas. La misma ofrece la posibilidad de identificar posibles indicadores o comportamientos sospechosos de carácter migratorio, así como facilitar el seguimiento y localización de una persona bajo investigación.

Agregación Trámites sobre Propiedades (agg_tramitePersonaPropiedad):

La agregación Trámites sobre propiedades suple las necesidades de análisis del proceso de compra, venta, traspaso o cualquier otro tipo de trámite realizados sobre las propiedades. En la misma se recoge toda la información necesaria para caracterizar compradores, vendedores y ambiente de las transacciones, elementos que pueden resultar de vital importancia para iniciar e incluso sustentar mecanismos de investigación en esta esfera. Uno de los principales elementos a reconocer los constituyen la compra y venta excesiva de propiedades.

Agregación Trámites sobre Armas (agg_tramitePersonaArmas):

La caracterización y seguimiento de las armas y los trámites que con ellas se realizan es el principal objetivo de esta agregación. En la misma se vinculan los elementos fundamentales a tener en cuenta tanto de las armas como de sus propietarios. Se orienta principalmente a tener un estricto control de las

armas en circulación, reconocer la posesión de armas por parte de elementos potencialmente hostiles e identificación de las armas vinculada a un proceso penal así como el análisis de los principales indicadores de comportamientos de interés en esta esfera.

Agregación Trámites sobre los Vehículos (agg_tramitePersonaVehiculo):

A nivel mundial el robo y contrabando de vehículos es uno de los delitos más proliferados. Para garantizar un mecanismo eficiente de identificación y control de los vehículos para CICPC se lleva a cabo la construcción de esta agregación, en ella se recogen las características identificativas de los vehículos y sus propietarios, así como los trámites legales en los que los primeros se ven involucrados. Esta agregación está orientada al reporte, el análisis y a la identificación de vehículos relacionados con algún proceso dentro del sistema.

Además de los esquemas definidos se utilizaron un conjunto de **tablespace** para organizar y optimizar el uso de los recursos físicos en el almacenamiento de los datos. Ver **Tabla 2**.

Definición de los **tablespace** del ODS para SIIPOL		
Nombre	**Dirección**	**Esquema**
ts_saime	/mnt/saime	ods_saime
ts_saren	/mnt/saren	ods_saren
ts_inttt	/mnt/inttt	ods_inttt
ts_darfa	/mnt/darfa	ods_darfa
ts_seniat	/mnt/seniat	ods_seniat
ts_siipol	/mnt/siipol	ods_siipol
ts_sigep	/mnt/sigep	ods_sigep
ts_sigepol	/mnt/sigepol	ods_sigepol
ts_agg	/mnt/agregaciones	ods_agg
ts_cc_auditoria	/mnt/cc_auditoria	ods_cc_auditoria_part

Tabla 2: Definición de los tablespace

2.8 Seguridad en el ODS

El rigor que en el manejo de la información se exige a una institución como el CICPC requiere el diseño de estrategias de seguridad para el tratamiento de la información que se gestione, ya que se habla de datos personales de la población e información que tiene trascendencia legal, por lo que el acceso de terceros, tanto para conocimiento como para alteración de la misma, puede traer consecuencias negativas importantes.

Por estas razones, ninguna información que se haya ingresado en el sistema será eliminada físicamente de la BD, independientemente de que para el sistema, este elemento ya no exista, se ha implementado una estrategia para seguridad de acceso y administración de usuarios a través del otorgamiento de privilegios y roles, el acceso al sistema deberá ser a partir de contraseñas y la información será clasificada en cuanto a importancia y nivel de confidencialidad.

Se materializó la implementación de un control de cambios a determinados campos de información, de forma tal que sea posible determinar cuáles han sido las actualizaciones que se le han realizado. Además se ha diseñado una estrategia de respaldo mediante copias de seguridad de la base de datos hacia otro dispositivo de almacenamiento externo de manera que sea posible la recuperación de la base de datos a partir de los respaldos realizados.

2.8.1 Usuarios y Roles

Para el trabajo de actualización, administración y configuración del ODS se han definido un grupo de usuarios y establecido roles de trabajo como se muestra en la **Figura 10**.

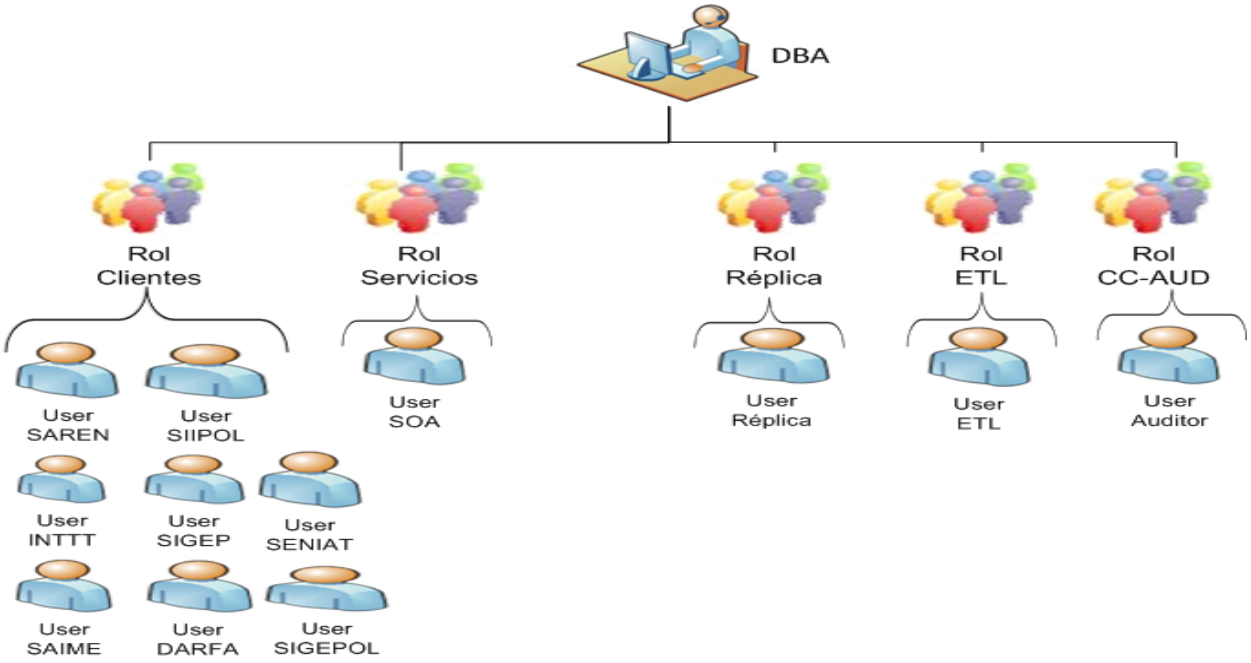

Figura 10: Usuarios y Roles

En primer lugar se definió un usuario para cada sistema externo, este usuario tiene como objetivo principal la protección y control de acceso sobre la información perteneciente a cada sistema. Premisa establecida debido al alto grado de sensibilidad de la información. Esto significa que aunque la Capa de Servicios será la encargada de la autenticación y filtrado de la información presentada para cada usuario, a nivel de datos sólo se podrá acceder a aquellos campos permitidos para cada proceso, usando el rol y los permisos previamente acordados, de forma tal que se lleva la seguridad no a nivel de tabla sino a cada campo contenido en la misma. Este mecanismo unido a elementos de encriptación, garantizan la total confidencialidad y presentación de la información almacenada en el sistema solamente a aquellas personas con los permisos suficientes para acceder a la misma.

Usuarios:

1. MPPRIJ (Ministerio del Poder Popular para las Relaciones Interiores y Justicia)
2. SIIPOL (Sistema de Investigación e Información Policial)
3. CICPC (Cuerpo de Investigaciones Científicas, Penales y Criminalísticas de Venezuela)
4. SAIME (Oficina Nacional de Identificación y Extranjería)

5. SAREN(Registro y Notarías)
6. INTTT(Instituto Nacional del Transito)
7. SIGEP (Humanización Penitenciaria)
8. SIGEPOL
9. DARFA (Dirección de Armamento de la Fuerza Armada)
10. SENIAT(Superintendencia Tributaria)

Roles y privilegios:

Administrador: Los usuarios con este rol tendrán un acceso total al sistema, con privilegios de lectura y actualización; es por esto que el mismo constituye el principal objetivo ante ataques. Se definen como máximo 3 usuarios pertenecientes a este rol y su distribución será discreta y por mecanismos seguros a las personas encargadas de la administración del almacén.

Servicios: Dentro del rol Servicios los usuarios tendrán acceso a las funciones, vistas, agregaciones, u otros mecanismo presentados por los administradores para la visualización de los datos. Este rol es el encargado de garantizar la presentación de la información requerida por cada cliente al sistema.

Replica: Rol interno del SGBD, contiene permisos de lectura sobre las bases de datos centrales y escritura sobre los nodos esclavos o de replicación. Posee cierto grado de sensibilidad debido al hecho que mediante este elemento se podrá visualizar toda la información contenida en el almacén, elemento imprescindible para garantizar una réplica total de la información.

CC-AD: Control de Cambios y Auditoría de datos, rol dedicado fundamentalmente a gestionar metadatos y tablas de control de cambios, posibilitando la realización de análisis sobre el comportamiento de los mismos.

ETL: Rol altamente sensible, posee acceso de lectura y escritura sobre toda la información contenida en el gestor. Es el único rol que podrá actualizar datos sobre el almacén, de ahí que su sensibilidad sea crítica, se estima que a este rol pertenezcan un máximo de 3 personas. Por estos motivos la acreditación del mismo en el Repositorio Central sólo se establece por canales seguros.

CLIENTES: Este rol tiene como objetivo garantizar el acceso a la información pública dentro del sistema.

2.8.2 Encriptación

Los mecanismos de encriptación resultan imprescindibles cuando se trabaja con información altamente sensible donde se necesiten mecanismos especiales para su protección, tal es el caso de las contraseñas y otros atributos que ostenten esta categoría. El proceso se realiza usando el algoritmo md5 soportado por el gestor de bases de datos. La encriptación unida a las políticas de seguridad y acceso garantizadas por Se-PostgreSQL posibilita que la alteración o visualización de la información protegida esté fuera del alcance por numerosos mecanismos de ataques tanto internos como externos, permitiendo proteger la información incluso de otros administradores del mismo sistema.

2.8.3 Control de Acceso a nivel de columnas

Debido al objeto integracional de la información contenida en las tablas del ODS existen elementos que aunque pertenecen a un mega concepto recogido en una entidad, no son propiedad del mismo, por tal motivo se han definido un conjunto de políticas de acceso que se llevan a cabo a nivel de atributo dentro de las tablas. Esto significa que aunque el usuario A sea propietario de la tabla1 pueden existir atributos de dicha tabla a los cuales no tenga acceso.

2.9 Estrategia de Copias de Respaldo

Para garantizar la persistencia de la información y la contribución a no tener almacenada información que no sea útil a los sistemas externos y SIIPOL se definieron las siguientes directrices.

Se realizarán backups completos anuales y mensuales, constituyendo cada uno un consolidado de los backups que los preceden en jerarquía, por ejemplo se establece debido al dinamismo de la información gestionada un backup completo semanal y posteriormente de tipo incremental, con periodicidad diaria de la información actualizada sobre la BD, garantizando en todo momento que exista una copia exacta de la información que está vigente en el servidor. La estructura de carpetas definidas con este sentido será con la jerarquía ***ODS -> Año -> Mes -> Semana*** y el nombre del scripts será de la misma forma ***anno_mes_dia*** logrando dejar plasmado claramente la fecha de la copia realizada.

Debido a que no es necesario que el tamaño del histórico crezca indefinidamente se establece, un máximo de años a almacenar de 2 años y que a partir de su aumento la información sea almacenada

anualmente en una estructura similar que las copias de respaldo pero llegando solamente hasta el nivel de año ***ODS -> Año*** en cintas de seguridad. El nombre de los scripts para este caso sería solamente ***ods_anno.***

Conclusiones

En este capítulo se han caracterizado los sistemas externos relacionados con SIIPOL, se definieron los lineamientos base de la arquitectura y las estructuras de datos del ODS. Se han descrito paso a paso: la arquitectura propuesta para el desarrollo de la solución y el diseño e implementación de las estructuras relacionales que soportan la integración de los datos en el sistema. La estrategia de desarrollo se centra en un proceso iterativo-incremental donde se deja la arquitectura preparada para la integración de las restantes fuentes en dependencia de las necesidades de información que posea SIIPOL.

CAPÍTULO 3: ANÁLISIS DE LOS RESULTADOS

El presente capítulo está orientado al análisis y validación de los resultados. Se detallan las temáticas referidas a la normalización, calibrado de la base de datos, análisis del rendimiento, pruebas y validación general del sistema.

3.1 Validación del Sistema

Al concluir el proceso de construcción del ODS, el análisis de algunos aspectos tales como la normalización, las pruebas de carga y stress, la validación del rendimiento ante la concurrencia de usuarios, los tiempos de respuesta, en fin, la validación del sistema; resulta tan importante como el diseño y la implementación misma.

El ODS al entrar en contacto con los usuarios finales, inicia un ciclo iterativo e incremental, de lo simple a lo complejo, donde son incluidos, con el transcurso del tiempo, años de información, requerimientos, funcionalidades, incluso nuevos principios y actualizaciones en el funcionamiento de las instituciones que lo conforman. Es entonces cuando comienzan a observarse los beneficios de los tiempos de respuesta, el dinamismo en la elaboración de los reportes, los conocimientos que puedan ser extraídos de la información almacenada y la efectiva preparación de los usuarios finales, garantizando así el éxito del sistema.

En la concepción de este trabajo se han seguido las pautas del Ciclo de Vida del Software [1], el que se ha particularizado añadiendo los elementos ya discutidos, una vez que fue identificada la necesidad de crear un ODS. Para sustentar la decisión asumida, se ha tomado como base la metodología propuesta por Kimball del Ciclo de Vida Dimensional del Negocio durante el desarrollo de un DW[2], pero adaptado al entorno actual. Aunque a lo largo de este trabajo se han seguido los pasos del ciclo de desarrollo formulado, no es objetivo detallar cada uno de ellos, sino solamente resaltar aquellos aspectos en los que se avala la propuesta teórica – práctica realizada en capítulos anteriores.

[1] El Ciclo de Vida del Software consta de varias etapas: análisis de requerimientos, diseño, implementación, pruebas, mantenimiento, y actualización.
[2] Kimball brinda esta propuesta en la Segunda Edición de su libro “The Data Warehouse Toolkit: the complete guide to dimensional modeling”.

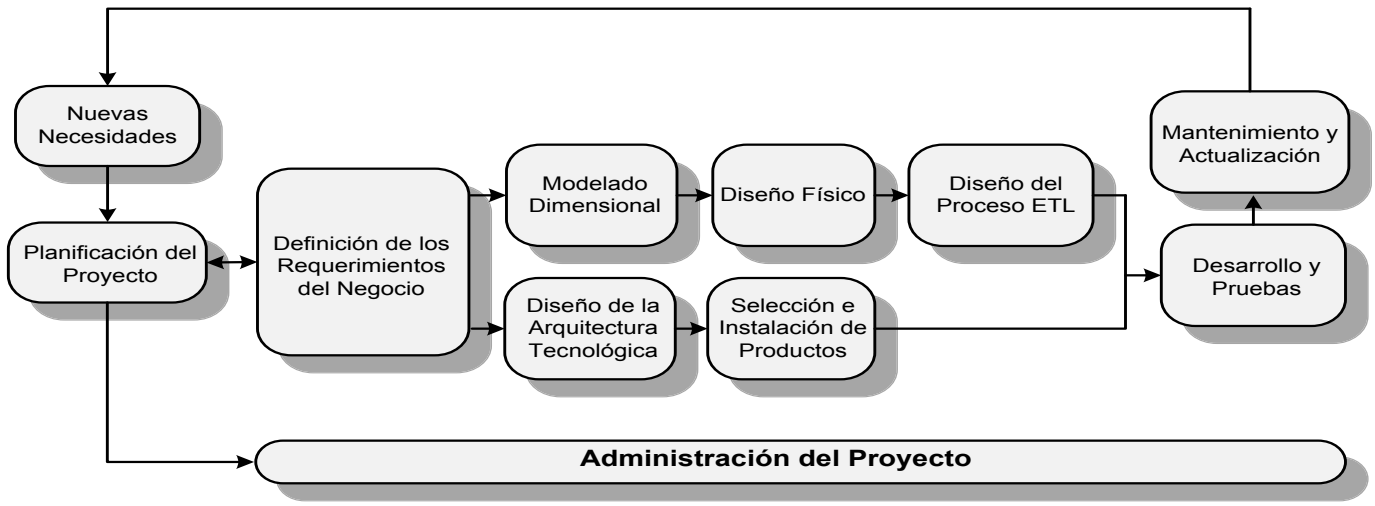

Figura 11: Diagrama del Ciclo de desarrollo del ODS

Después de haber concluido una primera iteración de ciclo de desarrollo del ODS, en el cual quedan definidos aspectos importantes referentes tanto al diseño como a la implementación del mismo, corresponde evaluar y validar el Sistema.

Resulta de gran importancia que los clientes estén inmersos en esta etapa de evaluación y validación del sistema, pues:

Pueden ser encontradas discrepancias con los requerimientos identificados en la etapa de análisis.

Familiarización con el ambiente de explotación de la información.

Para refinar el sistema en función de que quede lo más completo posible.

3.2 Normalización

La normalización, dentro del universo de los diseños relacionales se justifica y adquiere un valor incalculable debido a que garantizan el éxito conceptual y lógico de la base de datos. Es el proceso mediante el cual se transforman datos complejos a un conjunto de estructuras de datos más pequeñas, que además de ser más simples y más estables, son más fáciles de mantener.[27] Las bases de datos relacionales se normalizan para evitar la redundancia de los datos, y proteger la integridad de los mismos. Otra ventaja es el consumo de espacio. Una base de datos normalizada ocupa menos espacio en disco

que una no normalizada. Quizás la principal interrogante es: ¿qué tan lejos debe llegar la normalización en el sistema?

La normalización es una ciencia subjetiva. Determinar las necesidades de simplificación depende de los desarrolladores. Las reglas de normalización existen como guías para crear tablas que sean fáciles de manejar, así como flexibles y eficientes. A veces puede ocurrir que normalizar los datos hasta el nivel más alto no tenga sentido

3.3 Análisis del tamaño, crecimiento y calibrado del ODS

A partir de un estimado razonable en cuanto al tamaño de la base de datos, se tendrá una concepción aproximada de la dimensión espacial total que alcanzaría el ODS. Por tal razón, se realiza un análisis de cada una de las tablas propuestas para calcular la cantidad de unidades, la cantidad de filas implicadas en cada una de las tablas hasta llegar al número de Bytes que serán ocupados por concepto de tamaño.

3.3.1 Pruebas de Volumen y Carga

Existen un conjunto de posibles pruebas que se le pueden realizar a un sistema informático para validar su uso, ejemplo de ellas se pueden mencionar: pruebas de unidad, integración y sistema. Las que más impactan en el desarrollo de almacenes de datos son las pruebas que tengan relación con el rendimiento, capacidad y concurrencia. En este sentido las pruebas realizadas al ODS fueron las de volumen y carga.

Las **pruebas de volumen** son pruebas típicas de entornos que utilicen bases de datos. Las mismas se realizan para analizar el comportamiento del sistema o base de datos con volúmenes de datos almacenados lo más similar posible a los esperados en la explotación real del sistema. Para el sistema en cuestión la BD se pobló con los datos ficticios en un ambiente de pruebas. La información real resulta imposible de acceder debido a cuestiones elementales de seguridad de los sistemas relacionados.

Al introducir los datos en el almacén no se presentaron problemas de límite de capacidad, ni de volumen de datos. Tampoco se detectaron desbordamientos de matrices, columnas, atributos, tipos de datos, ni peticiones excesivas de memoria. Las llaves autogeneradas no se salieron del rango especificado, ni se detectaron problemas con los tipos de datos definidos en el paso de diseño. Lo anteriormente planteado garantiza que el gestor utilizado y el diseño de las estructuras de la base de datos implementadas soportan completamente el almacenamiento de los niveles de información requeridos.

Por otro lado, las **pruebas de carga** consisten en someter a una aplicación y/o base de datos a un régimen de carga de trabajo (habitualmente por simulación de concurrencia) similar al esperado en la explotación real del sistema. El objetivo de estas pruebas es buscar consultas mal diseñadas, consultas candidatas a optimización, la necesidad de índices adicionales, código mal diseñado, tiempo de demora de respuesta de magnitudes inaceptables, hardware insuficiente, problemas de control de concurrencia, etc.

Para la realización de las Pruebas de Carga existen diversos mecanismos y herramientas que automatizan dicho proceso. Se pueden utilizar desde navegadores ordinarios, trazas del servidor de base de datos, una aplicación simplificada con consultas de la aplicación real con un mínimo de código y sin complejidad algorítmica ni iteraciones, la utilización de herramientas diseñadas con este fin, entre otras.

Para realizar las pruebas se utilizan las bondades que brinda la herramienta **Jmeter** por la facilidad de su uso y funcionalidades. Apache-Jakarta Jmeter es un generador de carga diseñado para la realización de pruebas de carga y stress. Corre sobre la máquina virtual de java por lo que es multiplataforma. Genera carga por diversos protocolos, ya sea, FTP, HTTP, HTTPS, SQL, etc. Maneja cookies y autenticación. Realiza carga variable, en niveles de concurrencia, número de veces, tiempo, etc.; y su característica principal radica en que pertenece a la familia de software libre.

La herramienta posee dos tipos de generación de carga: la **indirecta**, es decir, a través de una aplicación y la **directa** que basa fundamentalmente su utilización en consultas grabadas en la traza o log del servidor de base de datos. La que se va a utilizar para las pruebas del sistema es la **directa** configurada específicamente para la realización de consultas sobre el servidor de base de datos.

La arquitectura general que se utilizó para la realización de las pruebas fue una estación cliente conectada directamente con el servidor de BD mediante la configuración del Jmeter. Para el análisis de los resultados se realizaron pruebas con distintas cantidades de usuarios concurrentes por sistema. El número de conexiones establecidas por sistema tuvo 10 como valor mínimo y 20 en su proyección máxima, una cantidad superior a esta no está prevista atendiendo a las características del negocio. Las consultas se ejecutaron sobre las agregaciones definidas y se limitó la cantidad de tuplas a 10 000 argumentando que un reporte que se acerque a esa cifra se convierte, prácticamente, en inmanejable.

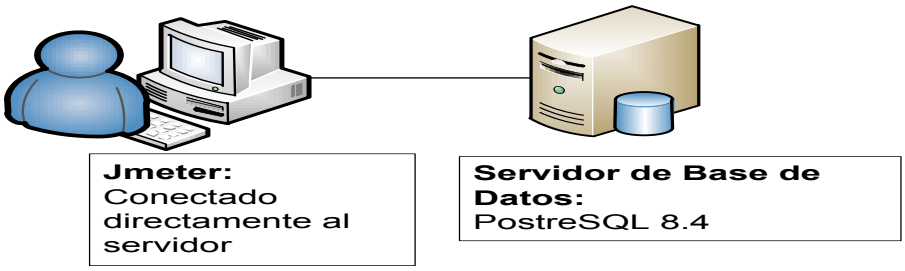

Figura 12: Configuración para las pruebas de carga

3.4 Pruebas de Rendimiento

En este punto se analizan los tiempos de rendimientos del Sistema en la respuesta a distintos pedidos de información accediendo a la base de datos que se encuentra en el servidor PostgreSQL.

En este análisis se propone determinar los tiempos que demora en recuperar la información almacenada mediante consultas de distintos grados de complejidad, sobre una cantidad determinada de filas. El objetivo es demostrar cuán óptimo, fácil y rápido se recupera la información del almacén.

De tal forma se da comienzo al análisis ofreciendo algunos datos del caso de estudio:

Fuente de datos a reportar: Comportamiento de las agregaciones: Viajes Internacionales, Trámites de vehículos, Trámites de propiedades, Trámites de armas.

Características del Hardware del Servidor:

Hardware: 1 Gb de memoria RAM, 160 Gb de capacidad de disco duro SATA, procesador Intel Pentium IV a 3.0 GHz

Software: SO Debian 5.0, PostgreSQL 8.4

Prueba No.1 Comportamiento de la agregación Viajes internacionales

- Tabla involucrada: agg_viajesInternacionales
- Cantidad total de filas: 12594 filas
- Cantidad de filas recuperadas: 10000 filas
- Consulta SELECT * FROM ods_agg."agg_viajesInternacionales";

- Cantidad de Usuarios:

  - 5 usuarios concurrentes en 4 sistemas de prueba: Total 20 usuarios

	Mín. (seg)	Media (seg)	Mediana (seg)	Max (seg)
Resultados	**1,437**	**0,64**	**1,392**	**2,172**

  - 10 usuarios concurrentes en 4 sistemas de prueba: Total 20 usuarios

	Mín. (seg)	Media (seg)	Mediana (seg)	Max (seg)
Resultados	**3,5**	**0,61**	**2,771**	**3,938**

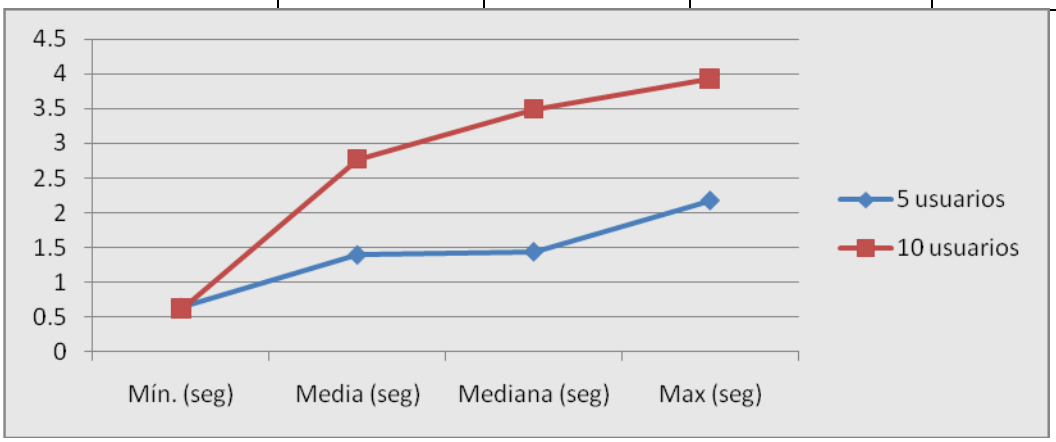

Gráfico 1: Representación de la prueba 1

Prueba No.2 Comportamiento de la agregación Trámites de vehículos

- Tabla involucrada: agg_tramitePersonaVehiculo
- Cantidad total de filas: 16500 filas
- Cantidad de filas recuperadas: 10000 filas
- Consulta SELECT * FROM ods_agg."agg_tramitePersonaVehiculo";
- Cantidad de Usuarios:

  - 5 usuarios concurrentes en 4 sistemas de prueba: Total 20 usuarios

	Mín. (seg)	Media (seg)	Mediana (seg)	Max (seg)
Resultados	**0,44**	**1,42**	**1,32**	**2,45**

- 10 usuarios concurrentes en 4 sistemas de prueba: Total 20 usuarios

	Mín. (seg)	Media (seg)	Mediana (seg)	Max (seg)
Resultados	**0,67**	**2,91**	**2,83**	**3,59**

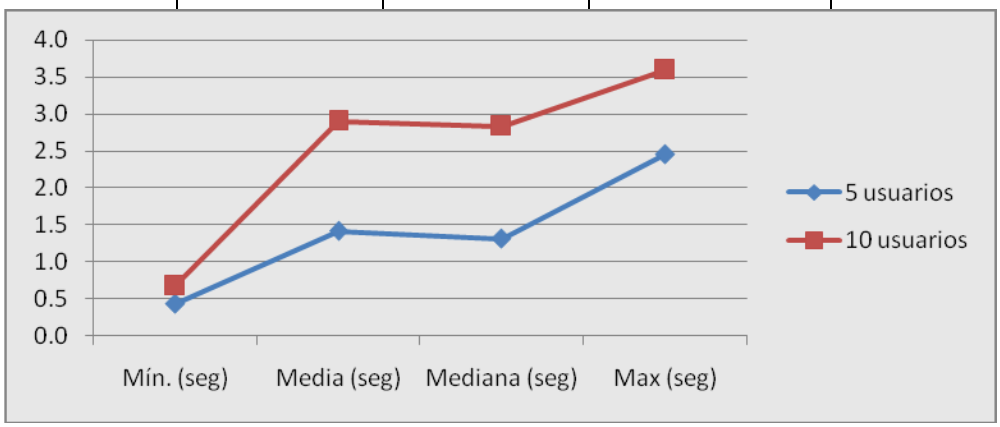

Gráfico 2: Representación de la prueba 2

Prueba No.3 Comportamiento de la agregación Trámites de propiedades

- Tabla involucrada: agg_tramitePersonaVehiculo
- Cantidad total de filas: 12700 filas
- Cantidad de filas recuperadas: 10000 filas
- Consulta SELECT * FROM ods_agg."agg_tramitePersonaInmueble";
- Cantidad de Usuarios:
  - 5 usuarios concurrentes en 4 sistemas de prueba: Total 20 usuarios

	Mín. (seg)	Media (seg)	Mediana (seg)	Max (seg)
Resultados	**0,50**	**1,59**	**1,63**	**2,39**

- 10 usuarios concurrentes en 4 sistemas de prueba: Total 20 usuarios

	Mín. (seg)	Media (seg)	Mediana (seg)	Max (seg)
Resultados	**2,20**	**3,92**	**3,52**	**4,58**

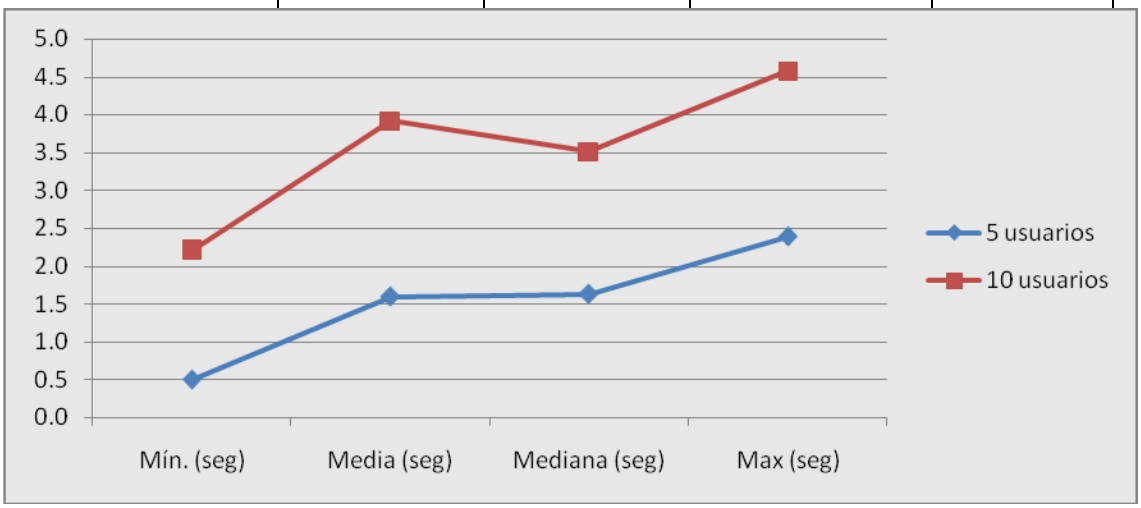

Gráfico 3: Representación de la prueba 3

Prueba No.4 Comportamiento de la agregación Trámites sobre armas

- Tabla involucrada: agg_tramitePersonaArmas
- Cantidad total de filas: 1500 filas
- Cantidad de filas recuperadas: 1500 filas
- Consulta SELECT * FROM ods_agg."agg_tramitePersonaArmas";
- Cantidad de Usuarios:
  - 5 usuarios concurrentes en 4 sistemas de prueba: Total 20 usuarios

	Mín. (seg)	Media (seg)	Mediana (seg)	Max (seg)
Resultados	**0,13**	**0,34**	**0,30**	**0,66**

  - 10 usuarios concurrentes en 4 sistemas de prueba: Total 20 usuarios

	Mín. (seg)	Media (seg)	Mediana (seg)	Max (seg)
Resultados	**0,16**	**0,77**	**0,72**	**1,63**

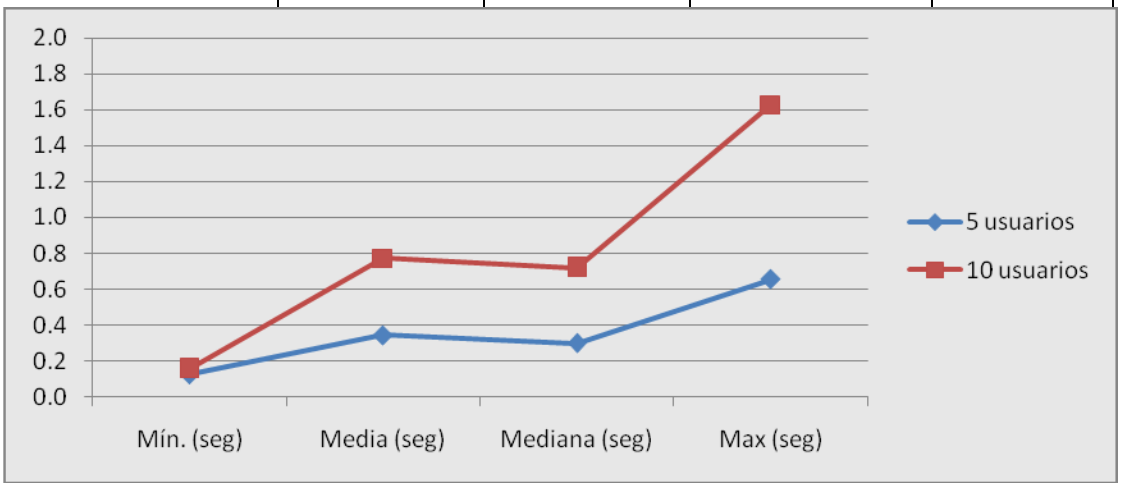

Gráfico 4: Representación de la prueba 4

Conclusiones del Capítulo

Al concluir este capítulo, se puede plantear que los objetivos propuestos para el mismo se han cumplido satisfactoriamente. No solo se ha probado la estabilidad y eficiencia en las prestaciones del SGBD escogido, PostgreSQL v8.4, para cumplir los requerimientos necesarios para la estabilidad y eficiencia del ODS, además se ha realizado un proceso de validación y perfeccionamiento de las sentencias SQL y estructuras físicas del almacén de datos, en aras de lograr una óptima funcionalidad bajo explotación.

CONCLUSIONES

Al finalizar la investigación, se exponen las siguientes conclusiones:

- El ODS constituye una solución viable para la integración a nivel de datos entre el SIIPOL de CICPC y los sistemas externos.
- Las estructuras de almacenamiento de datos diseñadas e implementadas soportan las necesidades informacionales analizadas.
- La arquitectura definida garantiza la flexibilidad y escalabilidad del sistema.
- Se identificaron las agregaciones cuya implementación y carga de datos mejora el rendimiento y potencia el análisis para la toma de decisiones tácticas.
- Los esquemas de seguridad implementados gestionan el control de acceso a varios niveles del sistema elevando consistentemente la seguridad del mismo.
- Las pruebas realizadas confirmaron que el SGBD PostgreSQL 8.4 garantiza el manejo de los volúmenes de datos previstos.
- Las pruebas realizadas permitieron validar la solución propuesta obteniendo resultados satisfactorios en cada una de ellas.

RECOMENDACIONES

Con el propósito de enriquecer la propuesta realizada en este trabajo, se sugiere:

- Incorporar otros sistemas según las necesidades informacionales de CICPC, teniendo en cuenta las condiciones tecnológicas para la integración.
- Definir e implementar nuevas agregaciones en vista a futuros requerimientos de análisis.
- Implementar un Almacén de Datos, que podría nutrirse del ODS, para potenciar el análisis estratégico.

REFERENCIAS BIBLIOGRÁFICAS

[1] Portal Web Institucional del Cuerpo de Investigaciones Científicas, Penales y Criminálisticas. Disponible en: http://www.cicpc.gov.ve/

[2] INMON, W. H. Building the Datawarehouse. EUA, 2002. p.

[3] RALPH KIMBALL, M. R. The Data Warehouse Toolkit. EUA y Canadá, Wiley Publishing Inc, 2002. p.

[4] INMON, W. H. Building the Data Warehouse. 4. EUA y Canadá, Wiley Publishing Inc, 2005. p.

[5] CLAUDIA IMHOFF, N. G., JONATHAN G. GEIGER. Mastering Data Warehouse Desing, Relational and Dimentional Techniques. EUA y Canadá, Wiley Publishing Inc, 2003. 9-13 p.

[6] WANG, J. Encyclopedia of Warehousing and Mining. E.U.A, Idea Group Reference, 2006. 17 p.

[7] ADAMSON, C. Mastering Data Warehouse Aggregates. EUA, Wiley Publishing Inc, 2006.

[8] INMON, W. H. The Operational Data Store InfoDB, 1995.

[9] KIMBALL, R. A Dimensional Modeling Manifesto. Revista DBMS Online, 1997.

[10] Empresas que utilizan Data Ware House. 2005.

[11] Historias de éxito de SYBASE. Telefónica., 2006.

[12] RIVERO, D. V. Sistema de Gestión de las Compras Mayoristas para la Corporación Cimex. Ciudad de La Habana, Universidad de La Habana, 2009. p.

[13] IMHOFF, C. A New Class of Operational Data Store.: Revista Information Management 2000.

[14] CURTO, J. CIF vs MD Dos enfoques clásicos en el diseño de la arquitectura de un Data Warehouse. La revista de la Gestión de Rendimiento, 2008.

[15] METODOLOGÍA PARA EL DESARROLLO DE SOLUCIONES DE ALMACENES DE DATOS E INTELIGENCIA DE NEGOCIO EN CENTALAD. UCIENCIA 2010, 2010.

[16] BERNABEU, R. D. Hefesto: Metodología propia para la construcción de un Data Warehouse. Argentina, 2007. p.

[17] WOLFF, C. G. Modelamiento Multidimensional, 2002.

[18] PONNIAH, P. Data Warehousing Fundamentals. EUA, Wiley Publishing Inc, 2001. p.

[19] BULTERMANN, D. C. A. Is It Time for a Moratorium on Metadata? . IEEE Multimedia, IEEE Computer Society Press., 2004.

[20] W. R. DURRELL, M.-H. Data Administration.A Practical Guide to Data Administration, 1985.

[21] KIMBALL, R. The Data Warehouse ETL Toolkit. Practical Techniques for extracting, cleaning, conforming, and delivering data. WILEY PUBLICHING, INC., 2004. p.

[22] Sitio Oficial de Oracle. 2010]. Disponible en: http://www.oracle.com/solutions/business_intelligence/feature_dw_leadership.html

[23] Sitio Oficial de Pentaho. 2010]. Disponible en: http://pentaho.almacen-datos.com/

[24] G. The PostgreSQL Global Development. PostgreSQL Conference. East 09, Philadelphia, Pennsylvania, United States., 2009.

[25] NAKAMURA, Y. Sistemas Gestores de Bases de Datos. Revista de Posgrado. Universidad Autónoma de México., 2007.

[26] ING. YUDISNEY VAZQUEZ ORTÍZ, I. M. L. O. V., ALDO CRISTIÁ ÁLVAREZ, Y YENI MORGADO SÁNCHEZ. Propuesta del PostgreSQL Empresarial Cubano. Ciudad de la Habana, 2009.

[27] ALARCÓN, J. M. Administración de SGBD PostgreSQL. 2006.

BIBLIOGRAFÍA

ADAMSON, C. Mastering Data Warehouse Aggregates. EUA, Wiley Publishing Inc, 2006.

ALARCÓN, J. M. Administración de SGBD PostgreSQL. 2006.

BERNABEU, R. D. Hefesto: Metodología propia para la construcción de un Data Warehouse. Argentina, 2007.

BULTERMANN, D. C. A. Is It Time for a Moratorium on Metadata? IEEE Multimedia, IEEE Computer Society Press. 2004.

CLAUDIA IMHOFF, N. G., JONATHAN G. GEIGER. Mastering Data Warehouse Desing, Relational and Dimentional Techniques. EUA y Canadá, Wiley Publishing Inc, 2003. 9-13 p.

CURTO, J. CIF vs MD Dos enfoques clásicos en el diseño de la arquitectura de un Data Warehouse. La revista de la Gestión de Rendimiento, 2008.

IMHOFF, C. A New Class of Operational Data Store.: Revista Information Management 2000.

INMON, W. H. Building the Data Warehouse. 4. EUA y Canadá, Wiley Publishing Inc, 2005.

INMON, W. H. The Operational Data Store InfoDB, 1995.

KEN ENGLAND, G. P. Performance, Optimization and Tunning handbook. EUA, Elsevier Inc, 2007.

KIMBALL, R. A Dimensional Modeling Manifesto. Revista DBMS Online, 1997.

NAKAMURA, Y. Sistemas Gestores de Bases de Datos. Revista de Posgrado. Universidad Autónoma de México., 2007.

PONNIAH, P. Data Warehousing Fundamentals. EUA, Wiley Publishing Inc, 2001.

ANEXOS

Anexo 1: Tabla comparativa entre los ODS y los DW

CRITERIO	ODS	DW
Orientación	Temas	Temas
Contenido	Información	Información
Actualización	Frecuente	No Frecuente
Intereses del Usuario	Operacionales	Gerenciales
Tipo de Análisis	Operacional	De Tendencias
Modelamiento	Multidimensional	Multidimensional
Historia de los Datos	Limitada, enfocada a períodos vigentes	Larga
Volatilidad de los Datos	Volátil	No Volátil
Nivel de agregación	Bajo. Muy detallado	Alto

Tabla 3: Comparación ODS - DW

Anexo 2: Tabla comparativa entre los modos de almacenamiento ROLAP y MOLAP

	Almacenamiento de Datos	Tecnologías Subyacentes	Funciones y Características
ROALP	Almacenamiento como tablas relacionales. Resumen detallado de los datos disponibles Volúmenes altos de datos Todos los datos de acceso están en la bodega almacenados.	Uso de SQL complejo para obtener los datos del depósito. Motor ROLAP en el servidor de análisis crea los cubos de datos sobre la marcha. Vistas multidimensionales en la capa de presentación.	Ambiente conocido y disponibilidad de herramientas Limitación en funciones de análisis complejas Realización de agregaciones no siempre son fáciles.
MOLAP	Almacenamiento como tablas relacionales. Diversos resúmenes de datos se mantienen en las BD propietarias Volúmenes de datos moderados. Resúmenes de acceso a datos detallados en BD multidimensionales.	Creación de cubos de datos prefabricados por el motor MOLAP. Tecnología propietaria para almacenar las vistas multidimensionales en arreglos no en tablas. Matriz de alta velocidad para la recuperación de los datos. Escasa tecnología de matriz de datos para gestionar la escasez de los resúmenes.	Acceso rápido. Fuerte librería de funciones para cálculos complejos Facilita el análisis independientemente de la cantidad de dimensiones. Amplitud en la capacidad de acción en el Drill-Down, Roll-Up y Slicing-Dicing.

Tabla 4: Comparación ROLAP - MOLAP

Anexo 3: Estandarización de los nombres

Tipo de Objeto	Función	Nomenclatura	Descripción
BD	ODS	ods_[nombre]	Almacén de datos Operacional
	catálogo	meta_[nombre]	Repositorio de Metadatos
	temporal	middle_[nombre]	Base de Datos temporal
Esquemas	Objetos de Configuración organizados por sistema fuente.	ods_[nombre fuente]	Esquemas donde se organizan los objetos de uso interno del ODS (Tablas, secuencias y funciones)
	Esquemas de Presentación de datos	ods_agg	Esquemas donde se almacenan las vistas materializadas definidas para gestionar los datos asociados a cada área de análisis.
Tablas	Data	tb_[nombre]	Tablas para almacenar la data.
	Configuración	cf_[nombre]	Tablas internas para almacenar datos de configuración asociados a las rutinas de control de cambio y sincronización de las vistas materializadas.
	Vistas Materializadas	vm_[nombre]	Tablas que materializan determinados cortes de información de interés para algunos análisis específicos.
Secuencias	Secuencias Tablas	seq_ tb_[nombre]	Secuencias de Tablas.
	Secuencias configuración	seqConf_ tb_[nombre]	Secuencias configuración
Funciones	ETL	fc_etl_[nombre]	Funciones para las tareas de Integración de Datos al DWH.
	Sincronización	fc_scn_[nombre]	Funciones internas para la sincronización de Vistas materializadas.
	Trigger	fc_tgr_[nombre]	Funciones para el control de cambios.
	Particionamiento	fc_prt_[nombre]	Funciones para particionar tablas.
	Replicación	fc_rp_[nombre]	Funciones de replicación.
Vistas	Vistas	v_[nombre]	Vistas de datos definidas.

Índices	Índices	Idx_tb_[nombre]	Índices definidos
Constraints	Claves Primarias	pk_id [tabla]	Clave primaria
	Clave Foránea	Fk_id [tablaFuente]	Clave foránea
	Unicidad	unq_[nombre]	Unicidad de Campo(s)
Trigger	Control de Cambio	tgr_cdc_[nombre]	Trigger de control de cambio
	Sincronización	tgr_agg_[nombre]	Trigger de agragación.
	Réplica	tgr_rp_[nombre]	Trigger de Replicación.

Tabla 5: Estandarización de nombres

Anexo 4: Diseño de la tabla para el Control de Cambio

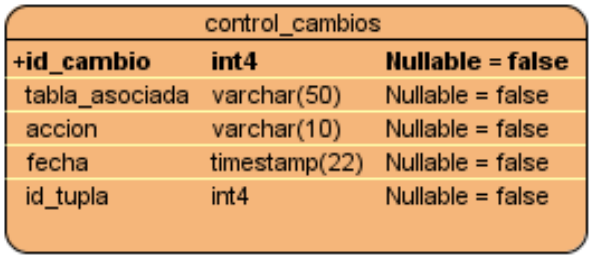

Figura 13: Tabla definida para el control de cambios

GLOSARIO DE TÉRMINOS

Agregación: Es la agrupación dos o más conjuntos de entidades relacionados para conformar un solo conjunto lógico. El objetivo primordial en la agregación es establecer relaciones entre conjuntos de entidades agrupadas.

Atomicidad: Se dice que una operación es atómica cuando es imposible para otra parte de un sistema encontrar pasos intermedios. Si esta operación consiste en una serie de pasos, todos ellos ocurren o ninguno. Por ejemplo en el caso de una transacción bancaria o se ejecuta tanto el depósito y la deducción o ninguna acción es realizada. Es una característica de los sistemas transaccionales.

OLAP: El término OLAP proviene de Online Analitical Processing (Procesamiento Analítico en Línea), define a una tecnología que se basa en el análisis multidimensional de los datos y que le permite al usuario tener una visión más rápida e interactiva de los mismos.

OLTP: (On-line Transaction Processing). Es un tipo de proceso especialmente rápido en el que las solicitudes de los usuarios son resueltas de inmediato; naturalmente, ello implica la concurrencia de un «mecanismo» que permite el procesamiento de varias transacciones a la vez.

Ralph Kimball: Conocido innovador, escritor, educador y consultor en el campo de Almacenes de Datos. En la actualidad posee más de 100 artículos sobre inteligencia empresarial. Es Vicepresidente de Metaphor Cumputer Systems, pionera en software para ayuda a la toma de decisiones y proveedora de servicios de esta índole. La asociación Ralph Kimball fue creada en 1992 para proveer consultoría y educación sobre la tecnología de warehousing.

Sumarizaciones: Actividad de incremento de la granularidad (nivel de detalle de los datos) de la información en una base de datos. La sumarización reduce el nivel de detalle, y es muy útil para presentar los datos para apoyar al proceso de Toma de Decisiones.

Transacciones: Una transacción es una secuencia de operaciones realizadas como una sola unidad lógica de trabajo. Una unidad lógica de trabajo debe exhibir cuatro propiedades, conocidas como atomicidad, coherencia, aislamiento y durabilidad para ser calificada como transacción.

William H. Inmon: Conocido como “El padre de la tecnología de Almacenes de Datos”, es el creador de la metodología CIF (Corporate Information Factory) y más recientemente GIF (Government Information Factory). Posee más de 35 años de experiencia en tecnología de administración de base de datos y diseño de almacenes de datos. Ha escrito más de 650 artículos sobre construcción, uso y mantenimiento de almacenes de datos. Posee la autoría de más de 46 libros de temas relacionados a tecnologías de base de datos.