Universidad de las Ciencias Informáticas

Facultad 6

TÍTULO: “Subsistemas de almacenamiento e integración N Glicolil GM3 para el almacén de datos Ensayos Clínicos del Centro de Inmunología Molecular.”

Trabajo de Diploma para optar por el título de

Ingeniero en Ciencias Informáticas

Autores: Yudith Acosta Anglés
Dainiel Viltre Guillén

Tutores: Ing. Esley León Valdés
Ing. Yulién Figueredo Guzmán

La Habana, junio de 2013
“Año 55 de la Revolución”

"No se puede dirigir si no se sabe analizar, y no se puede analizar si no hay datos verídicos; y si no hay todo un sistema de recolección de datos confiables, sin mentiras y globos, si no hay toda una representación de un sistema estadístico y de hombres habituados a recoger los datos y transformarlos en números.
Esta es una tarea esencial"
Ernesto Guevara

DECLARACIÓN DE AUTORÍA

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

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

Yudith Acosta Anglés

_____

Firma del autor

Dainiel Viltre Guillén

_____

Firma del autor

Ing. Esley León Valdés

_____

Firma del tutor

Ing. Yulién Figueredo Guzmán

_____

Firma del tutor

DATOS DE CONTACTO

Tutores:

Ing. Esley León Valdés.

Universidad de las Ciencias Informáticas, Habana, Cuba.

Email: elvaldes@uci.cu

Ing. Yulién Figueredo Guzmán

Universidad de las Ciencias Informáticas, Habana, Cuba.

Email: yfguzman@uci.cu

Agradecimientos

A mi abuela por mimarme y quererme como una hija, por estar presente en cada momento de mi vida.

A mi mejor amiga y cómplice, mi soporte y refugio seguro, compañera de conciertos en noches de apagón. A la mujer de cuyas entrañas nunca saldré.

A mi compañero eterno de juegos y aventuras, a mi guía y mi inspiración, a mi héroe: al hombre que me parió.

A mi hermanita por su cariño, te adoro siempre estaré ahí para ti.

A mi hermana Elizabeth por permitirme ser su amiga, por siempre estar apoyándome y demostrarme lo que valgo.

A mis tías Elaine, Tania y Bertica por el amor que me han dado y tantos consejos que me han convertido en una mejor persona.

A mis tíos Betico, Tony y Gilberto por todo el cariño y el apoyo durante estos difíciles años.

A las mejores primas del mundo Alina y Evelyn por estar a mi lado y darme todo su apoyo incondicional.

A los primos mas lindos del mundo Carlitos, Alejandro, Pavel y Pedrito por demostrarme que con primos como ustedes no hace falta hermanos.

La amistad no se conquista, no se impone; se cultiva como una flor. No importan las distancias o los niveles sociales. A mis eternos compañeros de noches de desvelo. A ustedes que me han soportado: Yaima, Gleivis, Liena, Lili, Titi, Jorge, Keke, Ronniel, Kique, Oscar y Alfredo.

A Marianela y Ariannis por tantos consejos, por su apoyo y el gran cariño que siempre me han brindado.

Agradezco también a mis suegros, Alicia y Leonel, que siempre me aconsejaron y animaron a seguir para adelante, en especial a Alicia que te has portado como una madre para mí.

A mi dúo de tesis por su ayuda y apoyo durante todo este tiempo, por ser el mejor compañero de trabajo del mundo.

A mis tutores Esley y Yuya por ayudarnos y guiarnos durante este proceso investigativo.

A todos los profesores que contribuyeron en mi formación como profesional.

A todas las personas que contribuyeron que este sueño se hiciera realidad.

A Fidel y a la Revolución por darme la oportunidad de estudiar en esta universidad y hacer este sueño realidad.

No por último es el menos importante, agradezco a Leonel, por ser ante todo mi amigo, por aguantarme durante estos 4 años, por hacerme sentir la mujer más feliz del mundo, pero sobre todo por ser mi cosha, te amo.

Yudith

Agradecimientos

Quiero agradecer primeramente a dios por permitirme compartir este momento tan importante junto a mi familia y amistades más cercanas. A mingui por apoyarme en todo momento y confiar en mi desde un principio, simplemente eres todo para mi, sin ti nunca hubiera llegado a ningún lado. A mi papá que me ayudó en todo, pues siempre estuvo pendiente de mí para que me sintiera de la mejor forma posible y pudiera alcanzar así mis sueños. Les debo la vida a los dos y por eso los quiero mucho.

A mis abuelos y al resto de la familia, a mi hermano, que me causó muchos problemas pero al final ayudándolo me ayudé a mí mismo.

A mi mamita la más pequeña, la más gordita, por ayudarme a crecer, a ser mejor persona, a estar junto a mí en las buenas y en las malas, por ser una persona increíble que existe en pocos lados y que tuve la suerte de encontrar.

A mi dúo de tesis por ayudarme a obtener buenos resultados y apoyarme en todo momento. A Dayatni por brindarme su amistad y ayudarme con el trabajo, por ser una gran amiga de las que llegan para quedarse por siempre.

A mis compañeros de clases, que me permitieron tener días mágicos junto a ellos. A la gorda por quererme y apoyarme tanto. A mis compañeros del cuarto que siempre confiaron en mí y me apoyaron en todo momento: Flaco, Fito, Frank, Fuki y Arián.

A todos los profesores que de una forma contribuyeron a mi formación como ingeniero: Irina, Esley, Yulién, Fabián, Adalennis. Por cada duda aclarada y cada información brindada.

A mis amigas por enseñarme a enfrentar la vida y ser un profesional: Yisel, Lisandra, Yadira, Sheyla.

A mis amigos más cercanos por apoyarme en la vocacional y aquí en la UCI, por creer en mí, por estar siempre ahí cuando los necesitaba, por compartir momentos felices juntos: Aramis, Asiel y Joel.

A Fidel, Raúl y a la Revolución por brindarme la oportunidad de estudiar en esta universidad.

Dainiel

Dedicatoria

A mi madre y mi padre.

A mi ángel guardián, mi abuelo Beto, por guiarme, cuidarme y estar a mi lado en estos años que tanto lo he necesitado, pues aunque no estés físicamente junto a nosotros estarás por siempre en mi corazón.

Yudith

A mi mamá por entregarme su amor incondicional, por estar siempre ahí en las buenas y en las malas, por creer en mí en todo momento y por cuidarme con todo su esfuerzo y dedicación a pesar de su salud.

A mi papá que me quiere a morir y siempre ha confiado en mí, a él gracias por todo el apoyo y dedicación durante toda mi carrera, espero que siempre este orgulloso de mí y de todo lo que logre en la vida.

A mis abuelos que me lo dan todo y me quieren con el alma, a ellos gracias por toda su ayuda, por toda su humildad y por todos los momentos felices que me han regalado.

A mi familia en general por hacerme sentir parte de ella y admirarme siempre.

A mi hermano Osmani por apoyarme siempre y por el cariño y amor que le tengo.

Dainiel

RESUMEN

La presente investigación surge como parte de la colaboración existente entre la Universidad de las Ciencias Informáticas y el Centro de Inmunología Molecular. Este último, tiene como principal misión, la búsqueda y obtención de nuevos biofármacos destinados al tratamiento del cáncer y otras enfermedades crónicas no transmisibles relacionadas con el sistema inmune, introducirlos en la salud pública cubana, así como realizar la actividad científica y productiva económicamente sostenible y lograr importantes aportes a la economía del país. A dicha entidad se le hace necesario almacenar toda la información generada relacionada con los ensayos clínicos que en ella se realizan. Por este motivo la presente investigación tiene como objetivo general desarrollar los subsistemas de almacenamiento e integración N Glicolil GM3 para el almacén de datos de los ensayos clínicos del Centro de Inmunología Molecular que permita el almacenamiento homogéneo de la información. Para dar cumplimiento al objetivo general definido se plantearon tres objetivos específicos relacionados con la fundamentación de la selección de las metodologías, herramientas y tecnologías a utilizar en el desarrollo de los almacenes de datos, el análisis y diseño de los subsistemas de almacenamiento e integración N Glicolil GM3, así como la implementación y validación de los subsistemas de almacenamiento e integración N Glicolil GM3. Todo esto contribuyó a proporcionarle al Centro de Inmunología Molecular el subsistema de almacenamiento e integración que se presenta, el cual satisface todas las necesidades planteadas por los especialistas del área de ensayos clínicos de dicha entidad.

Palabras claves: Almacén de datos, Centro de Inmunología Molecular, Ensayos clínicos, Mercado de datos, Toma de decisiones.

ÍNDICE

INTRODUCCIÓN

En el mundo actualmente, ha cobrado gran importancia la búsqueda de soluciones para diversas enfermedades que afectan el sistema inmunológico, como es el caso del cáncer, que constituye un serio problema de salud para la humanidad debido a las altas tasas de incidencia y mortalidad que se presenta en todo el planeta, y a los problemas que genera en el orden psicológico, familiar, laboral y económico. A pesar del gran desarrollo tecnológico alcanzado, no se ha logrado encontrar una solución para la cura de dicha enfermedad, manteniéndose este flagelo no solo como un peligro para la vida del enfermo, sino también influyendo negativamente en la calidad de vida de sus familiares.

La Organización Mundial de la Salud (OMS) promueve el control del cáncer en el marco de los programas nacionales de lucha contra el mismo, integrándolo en los programas de prevención y el control de las enfermedades no transmisibles. Su cometido fundamental es promover políticas, planes y programas nacionales que estén incorporados a estas iniciativas, así como generar y divulgar conocimientos para facilitar la aplicación de métodos de control basados en datos científicos.

A nivel mundial se han creado varios centros de investigaciones científicas, con el objetivo de fabricar medicamentos para el tratamiento de enfermedades incurables. La gran mayoría de estos centros se encuentran en países desarrollados tales como Estados Unidos y Gran Bretaña. En este ámbito, Cuba no se encuentra exenta de estas investigaciones ya que se han creado, desde el triunfo de la Revolución, varios centros biotecnológicos y científicos dentro de los que se encuentra el Centro de Inmunología Molecular (CIM), fundado el 5 de diciembre de 1994 al oeste de La Habana.

El CIM tiene como principal misión, la búsqueda y obtención de nuevos biofármacos destinados al tratamiento del cáncer y otras enfermedades crónicas no transmisibles relacionadas con el sistema inmune, introducirlos en la salud pública cubana, así como realizar la actividad científica y productiva económicamente sostenible y lograr importantes aportes a la economía del país.

Las líneas de investigación básicas están concentradas en la inmunoterapia del cáncer, especialmente en el desarrollo de vacunas moleculares, ingeniería de anticuerpos, ingeniería celular, bioinformática y regulación de la respuesta inmune. El CIM realiza en hospitales altamente especializados, ensayos clínicos (EC) para el diagnóstico de tumores por métodos de imagenología y tratamiento de diferentes tipos de cáncer. Un EC es una evaluación experimental de un producto, sustancia, medicamento, técnica diagnóstica o terapéutica que, en su aplicación a seres humanos, pretende valorar su eficacia y seguridad.

Actualmente el CIM trabaja en varios productos como es el caso del N-Glicolil GM3, gangliósido que se encuentra sobre expresado en las células tumorales del cáncer de mama y se ha convertido en blanco importante para el tratamiento de esta enfermedad. Los gangliósidos son carbohidratos que pertenecen al grupo de los glicoesfingolípidos y se localizan en las células epiteliales actuando como receptores de virus, bacterias y toxinas. Estos se encuentran expresados en mayor concentración en los tejidos cancerosos, lo que posibilita el uso de anticuerpos contra él. Por ello se ha convertido en uno de los componentes fundamentales de una vacuna contra el cáncer de mama que se encuentra en fase de EC.

Con el transcurso de los años, el volumen de información almacenada en el área de EC del CIM se ha incrementado considerablemente, debido a la gran cantidad de datos que se genera en cada uno de los EC que son aplicados tanto fuera como dentro del país. Dicha entidad presenta dificultades a la hora de confeccionar los reportes, analizar y consultar la información recopilada y presentar los indicadores relacionados con dichos ensayos.

Los EC que se gestionan llevan asociados un gran volumen de documentación. La institución cuenta con una aplicación informática llamada EpiData, donde se encuentra almacenada toda la información, la cual no ha sido implementada con tecnologías avanzadas, además que presenta algunos problemas con su funcionamiento debido a errores en su confección. La aplicación genera ficheros Excel (¨xls¨) donde se recoge la información relacionada con los EC.

Para gestionar la información de cada proceso, los especialistas lo realizan manualmente. Dicha situación trae como consecuencia que resulte difícil para los especialistas realizar análisis certeros que contribuyan con las decisiones que la entidad debe tomar.

Por la situación anteriormente descrita, se plantea como **problema de la investigación**: ¿Cómo lograr la estandarización de los datos del producto N Glicolil GM3 para su almacenamiento de forma homogénea?

La investigación tiene como **objeto de estudio** los almacenes de datos, enmarcado en el **campo de acción** subsistema de almacenamiento e integración del producto N Glicolil GM3.

Para dar solución a la situación planteada anteriormente la investigación tiene como **objetivo general:** desarrollar los subsistemas de almacenamiento e integración N Glicolil GM3 para el almacén de datos Ensayos Clínicos del CIM que permita el almacenamiento homogéneo de la información. En correspondencia con el objetivo general, se plantean como **objetivos específicos**

1. Fundamentar la selección de la metodología, herramientas y tecnologías a utilizar en el desarrollo de los almacenes de datos.
2. Realizar el análisis y diseño de los subsistemas de almacenamiento e integración N Glicolil GM3.
3. Realizar la implementación y validación de los subsistemas de almacenamiento e integración N Glicolil GM3.

Resultados esperados:

1. Mercado de datos poblado.

Para el cumplimiento de estos objetivos se realizaron esencialmente las siguientes **tareas investigativas:**

1. Caracterización de las metodologías, herramientas y tecnologías a utilizar en el desarrollo de los almacenes de datos.
2. Levantamiento de los requisitos para definir las necesidades del cliente.
3. Descripción de los casos de uso de los subsistemas de almacenamiento e integración N Glicolil GM3.
4. Definición de la arquitectura de los subsistemas de almacenamiento e integración N Glicolil GM3 para identificar los subsistemas fundamentales que componen la solución.
5. Diseño del modelo lógico de datos de los subsistemas de almacenamiento e integración N Glicolil GM3.
6. Diseño del subsistema de integración para definir cómo se realizará la carga de los datos desde la fuente hacia los subsistemas de almacenamiento e integración N Glicolil GM3.
7. Implementación del modelo físico de datos de los subsistemas de almacenamiento e integración N Glicolil GM3.
8. Implementación del subsistema de integración que cargará los datos desde la fuente hacia los subsistemas de almacenamiento e integración N Glicolil GM3.
9. Aplicación de las listas de chequeo para garantizar la correcta implementación de los subsistemas definidos en la arquitectura definida.
10. Aplicación de los casos de prueba para validar la calidad de cada uno de los elementos de los

subsistemas de almacenamiento e integración N Glicolil GM3.

Estructura del documento:

El documento está estructurado como se muestra a continuación: resumen, introducción, tres capítulos, conclusiones, recomendaciones, referencias bibliográficas, bibliografía, Anexos y glosario de términos.

Capítulo 1: Fundamentación teórica.

En este capítulo se abordan definiciones y conceptos de los Almacenes de Datos (AD) y Mercados de Datos (MD) así como ventajas y desventajas, y sus características. También se relacionan las metodologías, tecnologías y herramientas para el desarrollo de un AD.

Capítulo 2: Análisis y diseño de los subsistemas de almacenamiento e integración N Glicolil GM3.

En este capítulo se realiza un estudio y análisis del negocio. Se abordan aspectos referentes al levantamiento de requisitos. Se especifican las necesidades de información, las reglas del negocio (RN), requisitos funcionales (RF), los requisitos no funcionales (RNF), y los requisitos de información (RI), así como los casos de uso del sistema (CUS). Se realiza además, el diseño de los subsistemas de almacenamiento e integración del MD.

Capítulo 3: Implementación y validación de los subsistemas de almacenamiento e integración N Glicolil GM3.

En este capítulo se implementan los procesos definidos en el diseño. Se realiza la implementación de cada uno de los subsistemas de almacenamiento e integración, que conforman el MD. También se valida la solución mediante listas de chequeos y los casos de prueba, para garantizar la calidad del producto.

CAPÍTULO 1: FUNDAMENTACIÓN TEÓRICA

1.1 Introducción

En este capítulo se abordan definiciones y conceptos de los AD y MD así como ventajas y desventajas, sus principales características, metas y elementos que lo componen. También se documentan las metodologías, tecnologías y herramientas para el desarrollo de un AD.

1.2 Almacenes de datos

El cada vez mayor poder de procesamiento y sofisticación de las herramientas y técnicas analíticas ha dado como resultado la creación de los AD, los cuales proporcionan almacenamiento, funcionalidad y receptividad a las consultas que van más allá de las posibilidades de las bases de datos (BD) destinadas a transacciones. Dado que se han creado AD para satisfacer las necesidades particulares de las empresas, no existe una sola definición canónica del término AD. Algunas de las motivaciones que conllevaron a la creación del concepto de AD fueron:

- ✓ La mayoría de decisiones de empresas, organizaciones e instituciones se basan en información de experiencias pasadas.
- ✓ Generalmente, la información que es necesario investigar sobre un cierto dominio de la organización se encuentra en:
  - BD tanto internas como externas.
  - Otras fuentes muy diversas, no necesariamente BD.
- ✓ Muchas de estas fuentes son las que se utilizan para el trabajo diario.
- ✓ Tradicionalmente el análisis para la toma de decisiones se realizaba sobre estas mismas BD de trabajo o BD transaccionales.
- ✓ La BD está diseñada para el trabajo transaccional y no para el análisis de los datos, por lo que el análisis es lento.

William H. Inmon definió un AD como: “un conjunto de datos orientado a temas, integrado, no volátil, variante en el tiempo, como soporte para la toma de decisiones” [1].

Ralph Kimball lo define como: "una copia de las transacciones de datos específicamente estructurada para la consulta y el análisis". También fue Kimball quien determinó que un AD no era más que: "la unión de todos los MD de una entidad". Defiende por tanto una metodología ascendente (bottom-up) a la hora de

diseñar un AD [2].

Son diversos los criterios y definiciones de los AD, en la presente investigación se define como AD a una recopilación de datos integrados, no volátil y variante en el tiempo donde se pueden acceder a los mismos, revelar conocimientos y tomar decisiones.

1.2.1 Mercados de datos

Un MD es una versión especial de los AD. Como los AD, los MD contienen una visión de datos operacionales que ayudan a decidir sobre estrategias de negocio basadas en el análisis de tendencias y experiencias pasadas. La diferencia principal es que la creación de un MD es específica para una necesidad de datos seleccionados, enfatizando el fácil acceso a una información relevante. [3].

1.2.2 Características de los almacenes de datos

- ✓ Orientado a tema: Los datos están almacenados por materiales o temas, estos se organizan desde la perspectiva del usuario final, mientras que en las BD se organizan desde la perspectiva de la aplicación.
- ✓ Integrado: Todos los datos en el almacén están integrados, facilitando una descripción global y un análisis comprensivo de los datos en el almacén.
- ✓ No volátil: El AD solo permite cargar nuevos datos y consultar los datos ya almacenados, no permite ni borrar, ni modificar los mismos.
- ✓ Variante en el tiempo: El tiempo está implícito en los registros de los datos en el almacén, posibilitando el análisis de las tendencias [4].

Para examinar los AD y distinguirlos de las BD transaccionales es necesario contar con un modelo de datos que sea apropiado.

- ✓ El modelo de datos multidimensional es una buena opción para las tecnologías OLAP (Procesamiento analítico en línea) y de soporte a la toma de decisión.
- ✓ Los AD suelen mantener series de tiempo y análisis de tendencia, que necesitan más datos históricos de los que contienen generalmente las BD transaccionales.

- ✓ Los AD son no volátiles. Esto significa que la información contenida en el almacén de datos cambia con menos frecuencia. Aunque los AD solo permiten consultar la información, dan la posibilidad de actualizar datos que en un futuro puedan sufrir cambios.
- ✓ La información del AD es menos precisa y se actualiza de acuerdo a una política de actualización, elegida con cuidado, y que es generalmente incremental.
- ✓ Las actualizaciones del AD las realiza el componente de adquisición del almacén, que proporciona todo el procesamiento previo necesario.

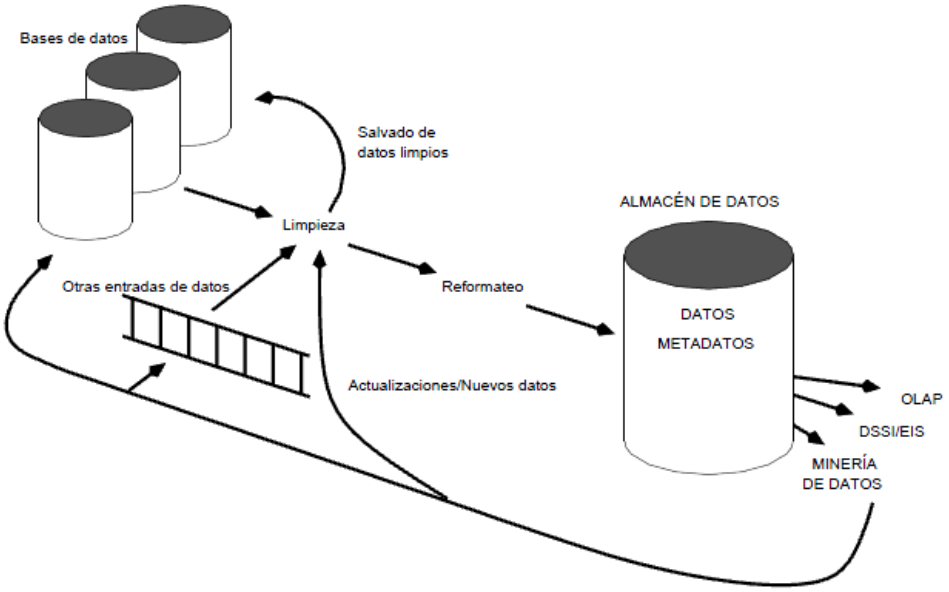


Figura 1. *Perspectiva general de la estructura conceptual de un almacén de datos.*

1.2.3 Componentes de los almacenes de datos

Características distintivas de un AD:

- ✓ Visión conceptual multidimensional.
- ✓ Dimensionalidad genérica.
- ✓ Dimensiones ilimitadas y niveles de agregación.
- ✓ Operaciones de dimensiones cruzadas sin restricciones.
- ✓ Arquitectura cliente-servidor.
- ✓ Soporte multiusuario.
- ✓ Accesibilidad.

- ✓ Transparencia.
- ✓ Manipulación de datos intuitiva.
- ✓ Buen rendimiento al crear informes consistentes.
- ✓ Creación de informes flexibles.

Bases de datos operacionales	AD
Datos operacionales	Datos del negocio para información
Orientado a aplicación	Orientado al sujeto
Actual	Actual + histórico
Detallada	Detallada+ más resumida
Cambia Continuamente	Estable

Tabla 1. *Comparación entre bases de datos operacionales y AD.*

1.2.4 Arquitectura de los almacenes de datos

La arquitectura de un AD viene determinada por su situación central como fuente de información para las herramientas de análisis.

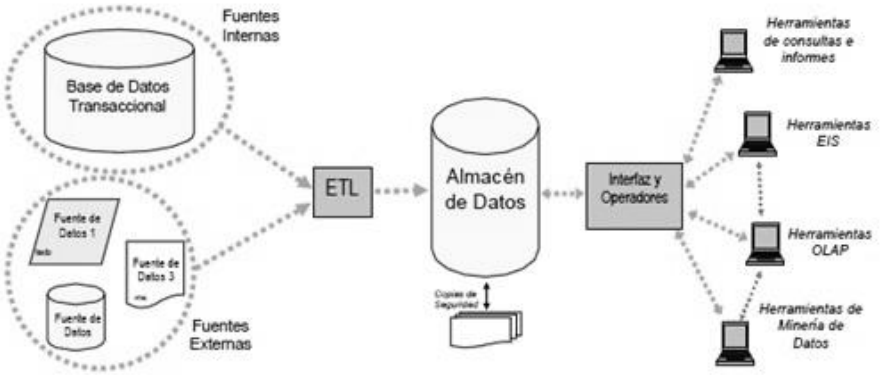


Figura 2. *Arquitectura de un almacén de datos.*

Componentes de los MD:

- ✓ Sistema de extracción, transformación y carga (ETL): Son los componentes más críticos de la infraestructura de Integración de Datos para la Inteligencia de Negocio (BI). Mientras que estos pasan desapercibidos para la mayoría de los usuarios de BI los procesos ETL recuperan los datos de las distintas fuentes y los pre-procesan para las herramientas de análisis e informes.

Estos procesos que tienen gran importancia en la precisión de la BI se detallan en:

- Extracción: Normalmente el AD integra diferentes sistemas de fuentes de datos. Cada uno de estos sistemas por separado puede utilizar una organización diferente de los datos o formatos distintos. Es aquí donde entra a jugar un papel importante el proceso de extracción, ya que convierte los datos de los diferentes sistemas a un formato preparado para iniciar el proceso de transformación. Este mecanismo para especificar las correspondencias entre el esquema fuente y un esquema intermedio para cargar la información en el almacén de datos se le denomina *mapping* o mapeo, que no son más que cálculos y funciones.
- Transformación: Esta fase aplica una serie de RN o funciones sobre los datos extraídos para convertirlos en datos que puedan ser cargados en el AD. Algunas fuentes de datos requieren de una pequeña manipulación de los datos para poder ser cargados. Entre las transformaciones más sencillas se encuentran:
  - Seleccionar solo ciertas columnas para ser cargadas, esto da la posibilidad de dejar columnas con valores nulos fuera de la carga.
  - Traducir códigos.
  - Codificar valores libres.
  - Derivar nuevos valores calculados.
  - Unir datos de múltiples fuentes.
  - Sumar múltiples filas de datos.
  - Generación de campos claves en el destino.
  - Trasponer o pivotar, dando la posibilidad de girar una o múltiples filas en columnas o viceversa.
- Carga: Esta fase transcurre en el momento en el que los datos de la fase anterior son cargados en el almacén de datos de destino. Este proceso abarca una amplia variedad de procesos diferentes dependiendo de los requerimientos de la organización. Algunos AD sobrescriben información antigua con datos nuevos, los sistemas más complejos pueden mantener un historial de los registros de manera que se pueda hacer una auditoría de los mismos y disponer de un rastro de toda la historia de un dato; este proceso se le denomina seguimiento de datos.

✓ Repositorio propio de datos: Información relevante, metadatos.

- ✓ Interfaces y gestores de consulta: Permiten acceder a los datos y sobre ellos se conectan herramientas más sofisticadas (OLAP, EIS, minería de datos).
- ✓ Sistemas de integridad y seguridad: Se encargan de un mantenimiento global, copias de seguridad [5].

1.2.5 Ventajas y Desventajas de los almacenes de datos

Ventajas de los almacenes de datos:

1. Proporciona información clave para la toma de decisiones empresariales.
2. Especialmente útil para el medio y largo plazo.
3. Son sistemas relativamente sencillos de instalar si las fuentes de datos y los objetivos están claros.
4. Muy útiles para el almacenamiento de análisis y consultas de históricos.
5. Proporciona un gran poder de procesamiento de información.
6. Permite una mayor flexibilidad y rapidez en el acceso a la información.
7. Facilita la toma de decisiones en los negocios.
8. Las empresas obtienen un aumento de la productividad.
9. Proporciona una comunicación fiable entre todos los departamentos de la empresa.
10. Mejora las relaciones con los proveedores y los clientes.
11. Permite conocer qué está pasando en el negocio, es decir, estar siempre enterado de los buenos y malos resultados.
12. Transforma los datos en información y la información en conocimiento.
13. Permite hacer planes de forma más efectiva.
14. Reduce los tiempos de respuesta y los costes de operación.
15. Facilita la aplicación de técnicas estadísticas de análisis y modelización para encontrar relaciones ocultas entre los datos del almacén; obteniendo un valor añadido para el negocio de dicha información.
16. Supone una optimización tecnológica y económica en entornos de centro de información, estadística o de generación de informes con retornos de la inversión espectaculares.

Desventajas de los almacenes de datos:

1. No es muy útil para la toma de decisiones en tiempo real debido al largo tiempo de procesamiento

que puede requerir. En cualquier caso la tendencia de los productos actuales (junto con los avances del hardware) es la de solventar este problema convirtiendo la desventaja en una ventaja.

2. Requiere de continua limpieza, transformación e integración de datos.
3. En un proceso de implantación puede encontrarse dificultades ante los diferentes objetivos que pretende una organización.
4. Una vez implementado puede ser complicado añadir nuevas fuentes de datos.
5. Requieren una revisión del modelo de datos, objetos, transacciones y además del almacenamiento.
6. Tienen un diseño complejo y multidisciplinar.
7. Tienen un alto coste.
8. Requieren sistemas, aplicaciones y almacenamiento específico [6].

1.3 Modelo Multidimensional

El modelo multidimensional (MMD) dentro del entorno de las BD, es una disciplina de diseño que se sustenta en el modelo entidad-relación y en las realidades de la ingeniería de texto y datos numéricos [7].

Dadas las características de los AD es ideal la utilización en su diseño de un MMD. Este tipo de diseño tiene como ventajas sobre el Modelo Entidad-Relación (MER), que es muy flexible, está desnormalizado y orientado a los intereses de un usuario final, aunque esto no significa que existan inconsistencias en los datos. Mediante la utilización de un MMD se disminuye la cantidad de tablas y relaciones entre ellas, lo que agiliza el acceso a los datos [8].

El MMD se representa a través de la definición de las tablas de hechos y dimensiones.

Tablas de Hechos: Representan la ocurrencia de un determinado proceso dentro de la organización y no tienen relación entre sí. Generalmente, almacenan medidas numéricas, las que representan valores de las dimensiones, aunque en ocasiones estas no están presentes y se les denominan “tablas de hechos sin hechos”. La llave de la tabla de hecho, es una llave compuesta, debido a que se forma de la composición de las llaves primarias de las tablas dimensionales a las que está unida.

Tablas de Dimensiones: Contienen, generalmente, una llave simple y atributos que la describen. En

dependencia del esquema de diseño que se asuma pueden contener llaves foráneas de otras tablas de dimensión. Existe una dimensión fundamental en todo el AD, la dimensión tiempo. Esto ocurre porque

todo registro que se incluya constituye la ocurrencia de un fenómeno en un instante de tiempo definido.

Dicha dimensión es la que establece uno de los objetivos fundamentales de la construcción de un AD, la conservación de un "histórico". Los atributos dimensionales son fundamentalmente textos descriptivos, estos juegan un papel determinante porque son la fuente de gran parte de todas las necesidades que deben cubrirse, además, sirven de restricciones en la mayoría de las consultas que realizan los usuarios. Esto significa, que la calidad del MMD, dependerá en gran parte de cuan descriptivos y manejables, sean los atributos dimensionales escogidos [9].

Existen tres esquemas para representar un MMD, ellos son:

- ✓ **Esquema en estrella:** Formado por una tabla de hechos con una única tabla para cada dimensión (Figura 3).
- ✓ **Esquema en copos:** Es una variante del esquema de estrella en el que las tablas dimensionales de este último se organizan jerárquicamente mediante su normalización (Figura 4).
- ✓ **Constelación de hechos:** Es un conjunto de tablas de hechos que comparten algunas tablas de dimensiones (Figura 5).

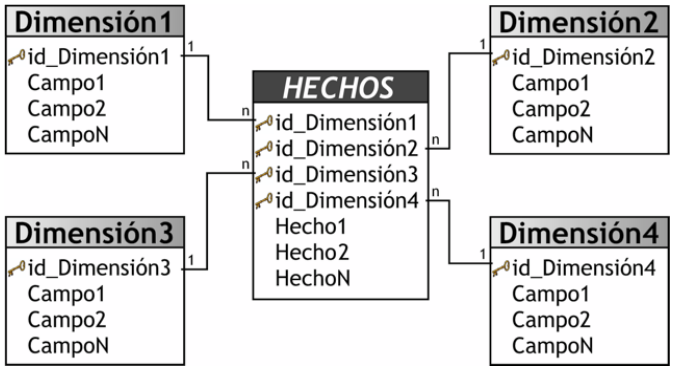


Figura 3. *Esquema en estrella.*

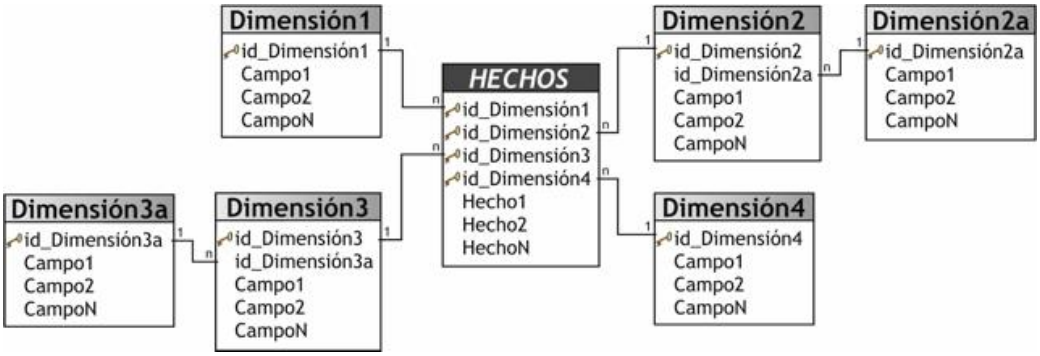


Figura 4. *Esquema en copo de nieve.*

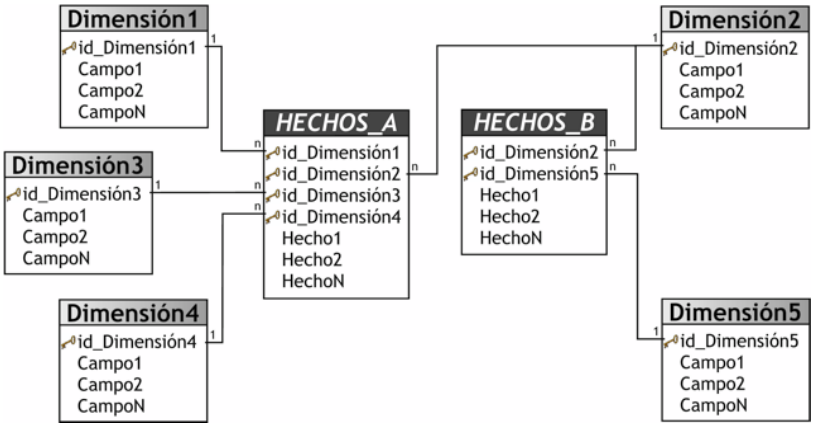


Figura 5. *Constelación de hechos.*

1.4 Estado actual de los almacenes de datos

En el marco actual de la empresa, la confluencia de nuevas infraestructuras de comunicación con potentes y flexibles herramientas de tratamiento de información (BD, AD, minería de datos,) mejoran la calidad, cantidad y eficiencia de los datos comerciales, así como el análisis, procesamiento y comunicación de los mismos.

En otras palabras, pueden aportar a las corporaciones la base tecnológica necesaria para afrontar los nuevos retos de la situación actual y las perspectivas de futuro de la gestión comercial.

Las BD y el AD permiten en primera instancia el almacenamiento adecuado de los datos obtenidos de las actividades habituales de organización, producción, control de gestión, marketing y planificación estratégica [12].

Hoy día, existen en el mundo numerosas empresas que han incorporado el uso de AD debido a su

importancia para la toma de decisiones y que brinda la facilidad de conocer a fondo el comportamiento de su negocio como para sacarle el máximo provecho [13]. Un ejemplo de estas empresas son: Wallmart, Whirpol, Coca Cola,Tv Azteca, Banco de México, pertenecientes a México, MonerisSolution, Canadian Tyre de Canadá, American Stores, Owens Corning Glass de los Estados Unidos, entre otras como Walt Disney, Nike, Adidas, 3M, Bosh Siemens, las cuales emplean AD para la realización de estudios de mercado principalmente.

Cuba no se ha quedado atrás en cuanto al desarrollo de los AD, pues existen organismos como la Oficina Nacional de Estadísticas e Información (ONEI) y Copextel que se encuentran en fases de despliegue de sus respectivos almacenes además de que en el XIII Concurso Nacional de Computación y en la Feria de Informática del 2002, se presentó un AD para Cubacel con grandes resultados obtenidos a partir de su implantación. También, en la Universidad de las Ciencias Informáticas (UCI) se han desarrollado diversos AD para la toma de decisiones.

1.5 Metodologías para el desarrollo de los almacenes de datos

Las metodologías para el desarrollo de los AD no son más que un conjunto de procedimientos, técnicas y ayuda a la documentación para el desarrollo de software. Las mismas indican detalladamente todas aquellas actividades que se deben realizar para lograr finalmente un producto con la calidad requerida, también describen el personal que debe participar en el desarrollo de las actividades así como el papel que estos deben desempeñar. Estas se basan en distintos modelos a la hora de construir un almacén. La empresa decide qué metodología utilizar dependiendo del modelo a seguir, del contexto en que se encuentre la misma y del objetivo que persiga. Estos modelos son:

- ✓ Modelo Top Down (de arriba hacia abajo o descendente).
- ✓ Modelo Bottom Up (de abajo hacia arriba o ascendente).
- ✓ Modelo Paralelo.
- ✓ Modelo Top Down con Retroalimentación.
- ✓ Modelo Bottom Up con Retroalimentación.
- ✓ Modelo Paralelo con Retroalimentación.

Cada fabricante de software de BI busca imponer una metodología con sus productos. Sin embargo, se imponen entre la mayoría dos metodologías, la de Ralph Kimball y la de Bill Inmon. La metodología de Inmon plantea un enfoque descendente donde crea primeramente el almacén y posteriormente los MD,

donde estos últimos se nutren del almacén. Kimball plantea un enfoque ascendente donde crea los MD y luego el almacén.

La metodología a utilizar para el presente subsistema de almacenamiento e integración es la Propuesta de Metodología para el Desarrollo de AD, realizada por el departamento de AD del Centro de Tecnologías de Gestión de Datos (DATEC) perteneciente a la UCI. Esta metodología se basa en el ciclo de vida de Kimball y en la propuesta de Leopoldo Zenaido Zepeda Sánchez en su tesis de doctorado, donde plantea incluir los casos de uso para guiar el proceso de desarrollo de un AD. Esta metodología incluye además las fases por las cuales tiene que pasar la construcción de un AD.

Entre las peculiaridades de esta metodología se encuentra la identificación de los requisitos de información y su trazabilidad a lo largo del ciclo de desarrollo de un MD, además de la inclusión de la etapa de prueba para validar que el producto se haya realizado con la calidad requerida. Ajusta las fases de actividades y artefactos a la propuesta de mejora llevada a cabo por el centro CALISOF. El ciclo de vida de esta metodología cuenta con ocho fases de las cuales para el desarrollo del presente subsistema de almacenamiento e integración solo se desarrollaron cinco fases (Estudio preliminar y planeación, Requisitos, Arquitectura, Diseño e Implementación y por último la fase de Prueba):

1. Estudio preliminar y planeación: Se realiza un estudio minucioso en la entidad cliente. Esto incluye un diagnóstico integral de la organización, con el fin de determinar qué se desea construir y qué condiciones existen para el desarrollo y montaje de la misma. Además se llevan a cabo las tareas de planeación del proyecto.
2. Requisitos: Se realiza el proceso entrevistas al cliente para determinar los requisitos de información. Se hace levantamiento detallado de las fuentes de datos para validar la disponibilidad de la información. Además se definen los RF y no funcionales de la solución y se hace el análisis de los requisitos que dan paso al diseño e implementación.
3. Arquitectura: Se definen las vistas arquitectónicas de la solución, aspectos como, los subsistemas y componentes, la seguridad, la comunicación y la tecnología a utilizar.
4. Diseño e Implementación: Se define el diseño de las estructuras de almacenamiento de datos, se diseñan los procesos de integración de datos como, el mapa lógico de datos, los cubos OLAP para la presentación de la información, así como el diseño gráfico de la aplicación definido por el cliente. Después se implementan cada uno de los subsistemas (repositorio de datos, integración de datos, presentación de datos).

5. Prueba: Se realizan las pruebas que validan la calidad del producto, comenzando por las Pruebas de Unidad llevadas, las Pruebas de Integración y Sistema, hasta llegar a las
   Pruebas de Aceptación con el cliente final. Esta fase no es la única en la que se realizan pruebas durante el desarrollo del proyecto, en todas las fases hay actividades de aseguramiento de la calidad.
6. Despliegue: Consta de dos etapas, despliegue piloto, donde se configuran los servidores necesarios y se instalan las herramientas según la arquitectura definida, se cargan una muestra de los datos en un ambiente controlado, con el fin de mostrarle al cliente final el sistema en funcionamiento. Una vez aceptada la solución por el cliente, se realiza la carga histórica de los datos, puede ser en el mismo entorno que el despliegue piloto u otro, todo depende de las condiciones que establezca el cliente. Además se realiza la capacitación y transferencia tecnológica de la solución a los clientes. El resultado fundamental es la solución desplegada en el entorno real y en correcto funcionamiento.
7. Soporte y Mantenimiento: Comienza cuando la solución está implantada y en explotación, y se ejecuta según el contrato firmado y las condiciones de soporte establecidas. Puede realizarse a través de variados servicios, que pueden ser soporte en línea, vía telefónica, web, correo u otros y el acompañamiento al cliente. Además se realizan las tareas de mantenimiento de la aplicación tan necesarias para este tipo de desarrollo y que garantiza el adecuado funcionamiento y crecimiento del almacén de datos.
8. Gestión del proyecto: Esta fase se ejecuta a lo largo de todo el ciclo de vida del proyecto. Es aquí donde se controla, gestiona y chequea todo el desarrollo, los gastos, las utilidades, los recursos, las adquisiciones, los planes y cronogramas entre otras actividades relacionadas con la gestión de proyectos. Esta fase es la columna vertebrar del proyecto y si no se ejecuta de forma continua y correcta el proyecto puede fracasar.

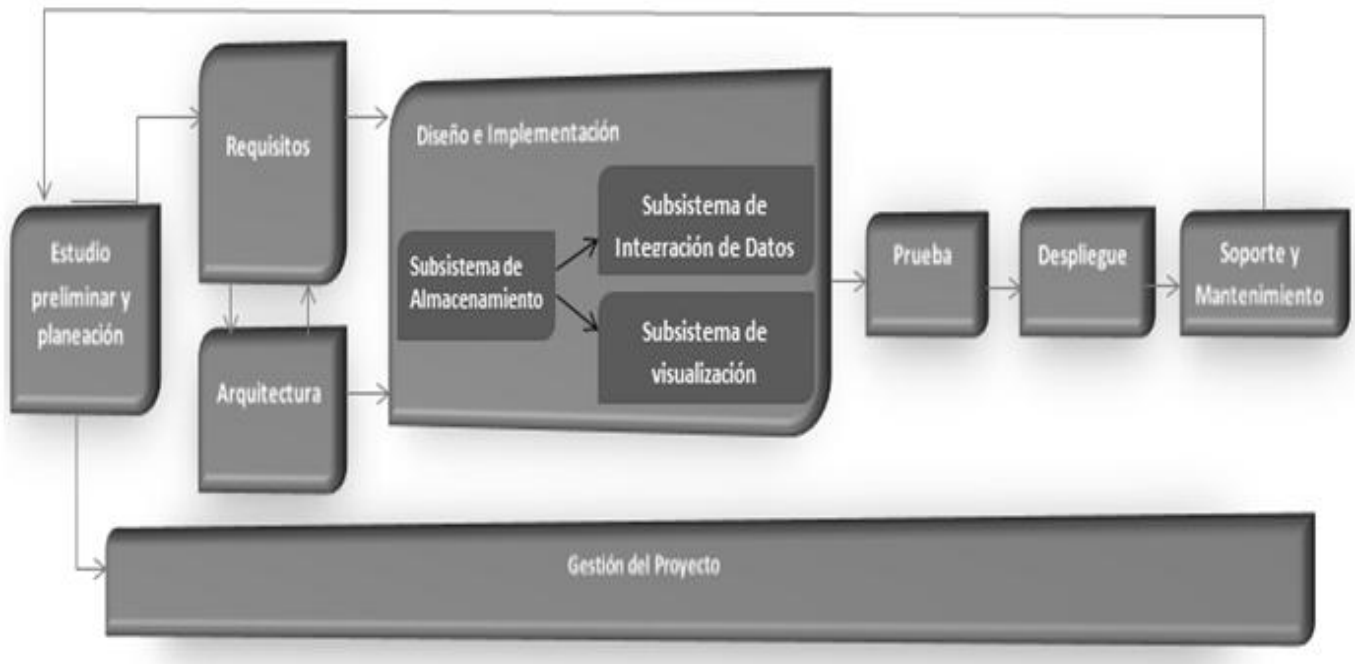


Figura 8. *Ciclo de vida de la* Propuesta de Metodología para el Desarrollo de Almacenes de Datos.

1.6 Herramientas de almacenes de datos

1.6.1 Herramienta para el modelado de los datos

Las herramientas de modelado se pueden definir como un conjunto de programas y ayudas que son utilizadas por los analistas, ingenieros de software y desarrolladores durante todo el ciclo de vida de desarrollo de un software. Estas herramientas también son definidas como un conjunto de métodos, utilidades y técnicas que facilitan la automatización del ciclo de vida del desarrollo de sistemas de información, completamente o en alguna de sus fases.

Visual Paradigm

Visual Paradigm 8.0 es una herramienta UML profesional que soporta el ciclo de vida completo del desarrollo de software: análisis y diseño orientados a objetos, construcción, pruebas y despliegue. La misma permite dibujar todos los tipos de diagramas de clases, código inverso, generar códigos desde diagramas y generar documentación. También proporciona abundantes tutoriales de UML, demostraciones interactivas de UML y proyectos UML. Presenta licencia gratuita y comercial. Es fácil de instalar y actualizar y compatible entre ediciones. Es multiplataforma y permite su uso en cualquier sistema

operativo.

Principales características:

- ✓ Diagramas de Procesos de Negocio - Proceso, Decisión, Actor de negocio, Documento.
- ✓ Diagramas de flujo de datos.
- ✓ Generación de BD - Transformación de diagramas de Entidad-Relación en tablas de base de datos.
- ✓ Ingeniería inversa de BD - Desde Sistemas Gestores de Bases de Datos existentes a diagramas de Entidad-Relación.
- ✓ Distribución automática de diagramas - Reorganización de las figuras y conectores de los diagramas UML.
- ✓ Importación y exportación de ficheros XMI. [14]

Se decidió utilizar Visual Paradigm 8.0 para modelar porque es una herramienta CASE profesional que soporta todo el ciclo de vida de desarrollo de software. Brinda además la posibilidad de modelar todo tipo de diagramas de clases. Los desarrolladores lo utilizan para facilitar el modelado simultáneo, almacenar los archivos de proyectos y hacer un seguimiento de los cambios. Una de las razones fundamentales por la que se utiliza la misma, es que nuestra universidad cuenta con la licencia de uso que presenta dicha herramienta.

1.6.2 Sistema Gestor de Base de Datos

Un Sistema Gestor de bases de datos (SGBD), en inglés Databases Management System (DBMS), es un sistema de software que permite la definición de BD; así como la elección de las estructuras de datos necesarios para el almacenamiento y búsqueda de los datos, ya sea de forma interactiva o a través de un lenguaje de programación. Un SGBD relacional es un modelo de datos que facilita a los usuarios describir los datos que serán almacenados en la base de datos junto con un grupo de operaciones para manejar los datos [15].

Los SGBD relacionales son una herramienta efectiva que permite a varios usuarios acceder a los datos al mismo tiempo. Brindan facilidades eficientes y un grupo de funciones con el objetivo de garantizar la confidencialidad, la calidad, la seguridad y la integridad de los datos que contienen, así como un acceso fácil y eficiente a los mismos.

Existe una gran variedad de SGBD ejemplos de estos son: PostgreSQl, SQLite, DB2 Express-C, Apache

Derby, Microsoft SQL, Sybase ASE Express Edition, Oracle Express Edition 10.

PostgreSQL

El SGBD que se utilizará en el desarrollo del presente trabajo es PostgreSQL 9.1: Es un sistema de gestión de BD objeto-relacional, distribuido bajo licencia BSD y con su código fuente disponible libremente. Es el sistema de gestión de BD de código abierto más potente del mercado y en sus últimas versiones presentan la misma calidad que otras BD comerciales [16].

Presenta las siguientes características:

- ✓ Soporte nativo para los lenguajes más populares del medio: PHP, C, C++, Perl, Python, etc.
- ✓ Soporte de todas las características de una base de datos profesional (triggers, store procedures, funciones, secuencias, relaciones, reglas, tipos de datos definidos por usuarios, vistas, vistas materializadas).
- ✓ Altamente adaptable a las necesidades del cliente.
- ✓ Se ejecuta en casi todos los principales sistemas operativos: Linux, Unix, BSDs, Mac OS, Beos, Windows.
- ✓ Documentación muy bien organizada, pública y libre, con comentarios de los propios usuarios.
- ✓ Comunidades muy activas, varias comunidades en castellano.
- ✓ Máximo número de filas por tabla ilimitado.
- ✓ Máximo número de columnas por tabla es de 250-1600 dependiendo del tipo.
- ✓ Máximo número de índices por tabla es ilimitado.
- ✓ Máximo tamaño de base de datos es ilimitado.
- ✓ Máximo tamaño de tabla es de 32 TB.
- ✓ Máximo tamaño de campo es de 1GB.

Ventajas que presenta:

- ✓ Diseñado para ambientes de alto volumen.
- ✓ Multiplataforma.
- ✓ Extensible.
- ✓ Ahorros considerables en costos de operación.
- ✓ Mejor soporte que los proveedores comerciales.
- ✓ Tamaño de base de dato ilimitado.

- ✓ Es una herramienta libre.

PgAdmin3

PgAdmin3 1.14 es una herramienta de código abierto y multiplataforma para la administración de BD PostgreSQL. Está diseñado para responder a las necesidades de todos los usuarios, desde la escritura simple de consultas SQL hasta desarrollar BD complejas. Su interfaz gráfica soporta todas las características de PostgreSQL y hace simple la administración. Está disponible en más de una docena de lenguajes y para varios sistemas operativos, como Microsoft Windows, Linux. PgAdmin3 soporta versiones de servidores 7.3 y superiores.

1.6.3 Herramientas para los procesos de Extracción, Transformación y Carga

DataCleaner

DataCleaner 1.5.3 es una herramienta cuyo objetivo principal es detectar errores e inconsistencias en los datos, para proveer una mejor calidad de los mismos. Por ejemplo, permite detectar entradas duplicadas, incompletas y establecer reglas para corregirlas. El objetivo no es borrar información sino mejorar la calidad de los datos construyendo un proceso de mejora continua. Es una herramienta de código abierto y multiplataforma. Genera sofisticados informes y gráficos que permiten a los usuarios identificar y analizar la estructura del origen de los datos y determinar el nivel de calidad de los mismos

Pentaho Data Integration

Pentaho Data Integration (PDI, también conocida como Kettle) es una de las soluciones más extendidas y mejor valoradas en el mercado, que reúne un conjunto de componentes que permiten modelar y ejecutar transformaciones sobre flujos de datos. Posee capacidades de integración de datos, entorno de diseño gráfico intuitivo y rico y una arquitectura altamente escalable, proporciona la solución ideal para cualquier tipo de integración de datos, análisis de negocio o proyectos con grandes capacidades de datos.

Debido a las características que esta valiosa herramienta presenta, se utilizó la misma en su versión 4.2.1 para el desarrollo del proceso de ETL, algunas de sus características se exponen a continuación:

- ✓ Multiplataforma: Windows, Macintosh, Linux.
- ✓ Entorno gráfico de desarrollo.
- ✓ Basado en dos tipos de objetos: Transformaciones (colección de pasos en un proceso ETL) y trabajos (colección de transformaciones).

- ✓ Uso de tecnologías estándar: Java, XML, JavaScript.
- ✓ Fácil de instalar y configurar.
- ✓ Incluye cuatro herramientas:
  - Spoon: para diseñar transformaciones ETL usando el entorno gráfico.
  - PAN: para ejecutar transformaciones diseñadas con spoon.
  - CHEF: para crear trabajos.
  - Kitchen: para ejecutar trabajos. [17]

1.7 Conclusiones del capítulo

Las definiciones y conceptos abordados en este capítulo permiten tener un mayor entendimiento en el área de los AD y su desarrollo, teniendo en cuenta las principales ventajas y desventajas del uso de los mismos. El estudio de las metodologías, tecnologías y herramientas permitió definir cuáles son las que serán empleadas en el desarrollo de la solución, sin descartar las ventajas y desventajas analizadas durante el estudio de las mismas. Este estudio y selección hace posible:

- ✓ Cubrir las etapas por las que transita el desarrollo de la solución con la metodología seleccionada.
- ✓ La confección de los diagramas fundamentales que formarán parte de la solución.
- ✓ Crear y mantener disponibles las estructuras físicas para un correcto almacenamiento de la información.
- ✓ Implementar los procesos de integración que posibilitarán que los datos tengan la calidad requerida al ser cargados al MD.

CAPÍTULO 2: ANÁLISIS Y DISEÑO DE LOS SUBSISTEMAS DE ALMACENAMIENTO E INTEGRACIÓN

2.1 Introducción

En este capítulo se realiza un estudio y análisis del negocio. Se abordan aspectos referentes al levantamiento de requisitos. Se especifican las necesidades de información, las RN, RF, RNF, y los RI, así como los CUS. Se construye la matriz bus, el modelo de datos donde se determinan las dimensiones, los hechos y las medidas, también se elabora el diseño de los subsistemas de almacenamiento e integración del MD.

2.2 Necesidades del negocio

Para desarrollar un AD que cumpla con las necesidades del cliente, en el presente caso el CIM, es muy importante realizar un análisis previo para que los especialistas de este centro queden satisfechos con el producto.

Varias son las técnicas que pueden llevarse a cabo para identificar las necesidades de la organización, tales como la entrevista, cuestionarios y observaciones. Obtener del usuario los requisitos es una tarea compleja, por lo que se optó por la técnica de la entrevista, ya que esta permitió interactuar directamente con el cliente. A partir de estas entrevistas se identificaron diversos problemas con la gestión de los datos relacionados con los ensayos clínicos realizados con los pacientes que presentan cáncer y de estos los que han hecho tratamiento con el compuesto vacunal N Glicolil GM3.

Como parte de las necesidades de información se acordó almacenar toda la información contenida de los diferentes EC, dichos ensayos se encuentran agrupados en tres grupos:

- ✓ **EC 058:** Ensayo clínico realizado a pacientes que presentan cáncer de mama metastásico, contiene los datos de inclusión, eventos adversos, tipo de evento adverso, laboratorio clínico, cumplimiento de inmunización presentados por estos.

- ✓ **EC 066:** Ensayo clínico realizado a pacientes que presentan melanoma cutáneo metastásico, contiene los datos de inclusión, eventos adversos, tipo de evento adverso, laboratorio clínico, cumplimiento de inmunización presentados por estos.

- ✓ **EC 067:** Ensayo clínico realizado a pacientes que presentan cáncer de mama metastásico, contiene los datos de inclusión, eventos adversos, tipo de evento adverso, laboratorio clínico, cumplimiento de inmunización presentados por estos.

En el ámbito del diseño de un MD, los grupos de información se corresponden con los temas de análisis que agrupan toda la información solicitada por los especialistas.

2.3 Reglas del negocio

Las RN son el conjunto de normas y políticas que deben cumplirse para alcanzar el correcto cumplimiento de los objetivos del sistema. Tomando como base el estudio preliminar del negocio se definieron las siguientes reglas del negocio:

RN1. La variable que identifica a un paciente en las BD es una concatenación entre el hospital, sus iniciales y el número de inclusión en el ensayo.

RN2. En el caso de las perspectivas que no se les encuentren relación con los datos fuentes en algunas de las BD porque no se recogieron en ese ensayo (Ejemplo: dosis, tratamientos previos) se registra en el AD como "No procede".

RN3. En el caso de las perspectivas que no se les encuentren relación con los datos fuentes en algunas de las BD porque se recogieron en ese ensayo, pero no está disponible la información en la fuente de datos (Ejemplo: estadio, raza, signos vitales) se registra en el AD como "No disponible".

RN4. Para los tipos de respuestas se determinó que se selecciona la respuesta global independientemente de la respuesta a las lesiones dianas y no dianas que pueda tener un paciente.

RN5. Los únicos tratamientos previos que se tendrán en cuenta para el análisis serán quimioterapia, radioterapia y cirugía. Se determinó que en los ensayos donde no se hayan recogido, ese campo aparecerá como "No procede".

RN6. En el caso de que un paciente no cumpla con el modelo de inclusión se determina que no forma parte del ensayo.

RN7. Se determinó crear rangos para la edad (Ejemplo: 01 a 10, 11 a 20, 21 a 30).

RN8. El hospital Celestino Hernández se recoge en los ensayos como CH. Se determinó cambiar ese valor por CHR.

RN9. La raza se recoge en los ensayos como 1(blanca), 2(negra), 3(mestiza) y 4(amarilla). Se determinó cambiar esos valores por 2(blanca), 5(negra), 4(mestiza) y 6(amarilla).

RN10. El nivel de dosis se recoge en los ensayos como un valor numérico. Se determinó cambiar esos valores por rangos: 1 a 300(Baja), 301 a 900(Media), 901 a 1500(Alta).

RN11. En algunos ensayos el nivel de dosis está representado en gramos. Se determinó convertir dicho valor a miligramos.

RN12. Los identificadores de los indicadores y de las dimensiones no pueden tomar valores nulos, ni repetidos.

2.4 Especificación de requisitos

La especificación de requisitos es una descripción completa del comportamiento del sistema que se va a desarrollar. Consiste en identificar las necesidades de información de la organización, las características y cualidades que debe poseer el sistema. A continuación se describen los RI, RF y RNF.

2.4.1 Requisitos de información

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

RI1. Obtener la cantidad de pacientes incluidos y evaluados inicialmente del ensayo clínico 058 de mama metastásico por hospital, provincia, edad, raza, estadio, tnm, tratamientos previos y tiempo.

RI2. Obtener la cantidad de pacientes incluidos y evaluados inicialmente del ensayo clínico 066 de melanoma cutáneo metastásico por hospital, edad, estadio, tnm, tratamientos previos y tiempo.

RI3. Obtener la cantidad de pacientes incluidos y evaluados inicialmente del ensayo clínico 067 de mama metastásico por TNM, estadio.

RI4. Obtener la cantidad de pacientes que presentaron eventos adversos de los ensayos clínicos del producto NGCGM3 por hospital, edad, tiempo, tipo de eventos adverso, grado de severidad del evento adverso y causalidad del evento adverso.

RI5. Obtener la cantidad de pacientes que abandonaron los ensayos clínicos del producto NGCGM3 por hospital, necropsia, tiempo_sobrevida, interrupción tratamiento.

RI6. Obtener la cantidad de pacientes que sufrieron lesiones durante los ensayos clínicos del producto NGCGM3 por hospital, edad, tiempo, respuesta global y evaluación de la lesión.

RI7. Obtener la cantidad de pacientes que se le realizaron exámenes físicos durante los ensayos clínicos del producto NGCGM3 por el tiempo, edad, hospital y exámenes físicos.

RI8. Obtener la cantidad de pacientes inmunizados durante los ensayos clínicos del producto NGCGM3 por hospital, edad, tiempo, dosificación y signos vitales.

RI9. Obtener la cantidad de pacientes que se le realizaron imagenología durante los ensayos clínicos del producto NGCGM3 por hospital, edad, tiempo y método de diagnóstico.

RI10. Obtener la cantidad de pacientes que se le realizaron exámenes de laboratorio clínico durante los ensayos clínicos del producto NGCGM3 por hospital, edad, tiempo y examen de laboratorio.

2.4.2 Requisitos funcionales

Los RF son las capacidades o condiciones que el sistema debe hacer para que sea exitoso y cumpla con las especificaciones del cliente. Seguidamente se presentan los RF de la solución:

RF1. Realizar la extracción de los datos.

RF2. Realizar la transformación y carga de los datos.

2.4.3 Requisitos no funcionales

Los RNF son las cualidades que hacen al sistema usable, confiable y atractivo para los usuarios. A continuación se muestran un conjunto de requisitos que forman parte de los RNF definidos para el subsistema de almacenamiento e integración:

Confiabilidad:

RNF1. Garantizar la persistencia de la información.

Soporte:

RNF2. Lograr la homogeneidad de la estructura de los elementos definidos en el almacén.

Restricciones de diseño:

RNF3. Utilizar las tecnologías definidas durante la investigación.

RNF4. Utilizar el SGBD definido durante la investigación.

RNF5. Utilizar la herramienta de integración de datos definida durante la investigación.

2.5 Diagrama de Caso Uso del Sistema

El diagrama de CUS es una representación de todos los actores del sistema, los casos de uso y las relaciones que existen entre ellos. Para la confección del diagrama, se cuenta con 12 CUS (diez casos de uso de información (CUI) y dos casos de uso funcionales (CUF)), como se muestra en la Figura siguiente.

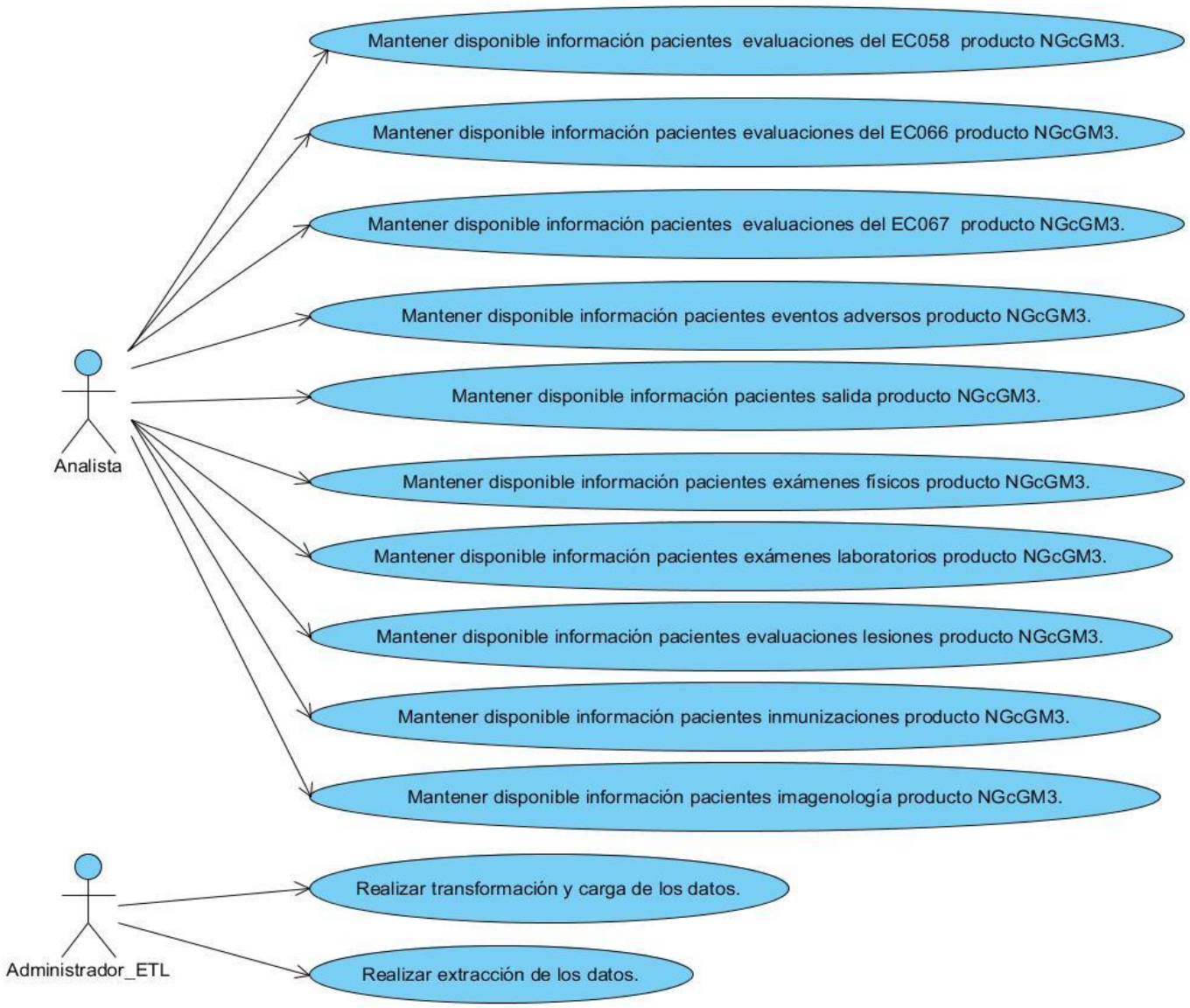


Figura 9. *Diagrama de Casos de uso de sistema*.

2.6 Especificación de los Casos de Uso

Para realizar la especificación de los CUS se sigue una secuencia de transacciones que desarrolla el sistema en respuesta a un evento o acción que inicia un actor sobre el sistema. A continuación se presenta la especificación del CU Realizar la extracción de los datos, las demás especificaciones puede consultarlo en el artefacto “Especificación de Casos de Uso.doc” dentro del Expediente de Proyecto de los Subsistemas de almacenamiento e integración N Glicolil GM3.

Objetivo:	Realizar la extracción de los datos.
Actores:	Administrador ETL.
Resumen	El caso de uso inicia cuando el actor selecciona los datos a extraer. Se extraen los datos de la fuente. Finaliza cuando los datos se encuentran en el área temporal.
Complejidad:	Media
Prioridad:	Media
Precondiciones:	Disponibilidad de las fuentes.
Poscondiciones:	Los datos de la fuente correspondiente han sido extraídos y almacenados en un área temporal.

Flujo Normal de Eventos	
Acción del Actor	**Respuesta del sistema**
1. El administrador de ETL realiza la conexión a la fuente correspondiente.	2. Responde a la solicitud de conexión.
3. El administrador de ETL selecciona los datos a extraer.	
4. El administrador de ETL realiza la extracción de los datos.	5. Ejecuta la extracción de los datos. Finaliza el caso de uso.

Flujos Alternos	
Acción del Actor	**Respuesta del sistema**
	2.1. No responde a solicitud de conexión.
	2.2. Notifica el error al administrador de ETL. Vuelve al paso 1 del Flujo Normal de Eventos.
3.1. Si hay control de cambios, el administrador de ETL verifica si hay modificaciones. En caso afirmativo ir al paso 3 del flujo normal.	

En caso negativo ir al paso 2 del flujo normal.	

Tabla 2. *Descripción del Caso de Uso: Realizar la extracción de los datos.*

2.7 Definición de la arquitectura base de los mercados de datos

El MD está estructurado mediante una arquitectura compuesta por la fuente de datos y tres subsistemas bases, estos subsistemas se comunican mediante el protocolo TCP/IP:

- ✓ Subsistema de integración: Es el encargado de extraer la información, así como limpiarla, estandarizarla e integrarla, preparándola para la carga al MD.
- ✓ Subsistema de almacenamiento: Es el encargado de almacenar todos los datos del mercado en diferentes tablas de hechos y dimensiones.
- ✓ Subsistema de visualización: Se encarga de consultar los datos existentes en el MD para luego presentarlos de disímiles maneras mediante gráficos y esquemas facilitando así la toma de decisiones.

La arquitectura base del presente sistema no incluye el subsistema de visualización por lo que cuenta con solo dos subsistemas, el de almacenamiento y el de integración:

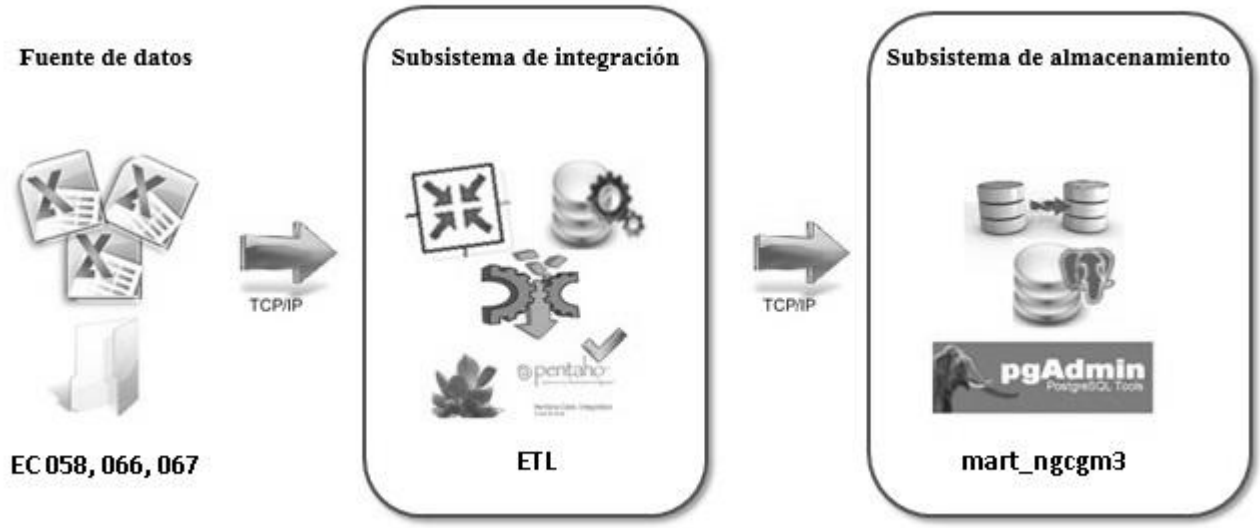


Figura 10. *Arquitectura del mercado de datos.*

2.8 Diseño del mercado de datos

2.8.1 Subsistema de almacenamiento

Para la realización y correcto funcionamiento del MD se realiza el modelo dimensional de los datos, el cual contiene las tablas de hechos identificadas en el negocio, las dimensiones y las relaciones que existen entre estas. Se debe definir además, una política de respaldo y recuperación que garantice la integridad de los datos almacenados.

Dimensiones

Las dimensiones representan las características de un hecho y permiten el análisis de los datos desde varias perspectivas. Pueden estar constituidas por jerarquías o niveles, que permiten detallar o agregar la información de forma dinámica. Las dimensiones presentan una perspectiva de los datos y son usadas para seleccionar y agregar datos a un cierto nivel de detalle [18].

Se identificaron 20 dimensiones, de las cuales nueve son dimensiones propias del MD N Glicolil GM3 y 11 dimensiones compartidas, estas son dimensiones comunes de todos los MD del CIM, a continuación se muestra un ejemplo de una de las dimensiones de la solución, las demás puede consultarlo en el artefacto “Especificación del modelo de datos.doc” dentro del Expediente de Proyecto de los Subsistemas de almacenamiento e integración N Glicolil GM3.

- ✓ Dimensión tratamientos_previos: Dimensión que describe los tratamientos previos realizados a los pacientes que se encuentran incluidos en cada uno de los EC del CIM.

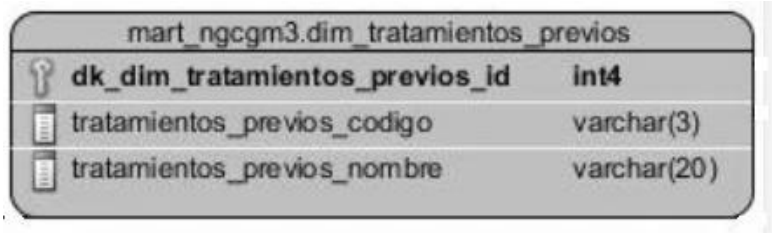


Figura 11. *Dimensión tratamientos previos.*

Hechos

Los hechos no son más que operaciones que se producen en el negocio, que los analistas utilizan para encontrar y examinar las tendencias. Las tablas de hechos contienen datos numéricos (hechos) que se

pueden resumir para proporcionar información sobre el historial de las operaciones a los usuarios. Cada tabla de hechos incluye un índice de varias partes que contiene los atributos de los registros de hechos y como llaves foráneas las llaves primarias de las tablas de dimensiones. Estas no deben contener información descriptiva ni datos que no procedan de los campos de medición numéricos y los campos de índice que relacionan los hechos con las correspondientes entradas en la tabla de dimensiones [19].

Se identificaron diez hechos, a continuación se muestra un ejemplo de uno de los hechos que presenta la solución, los restantes hechos puede consultarlo en el artefacto “Especificación del modelo de datos.doc” dentro del Expediente de Proyecto de los Subsistemas de almacenamiento e integración N Glicolil GM3.

- ✓ Hecho hech_eventos_adversos: Este hecho recoge la información relacionada con los eventos adversos presentados por los pacientes de los EC.

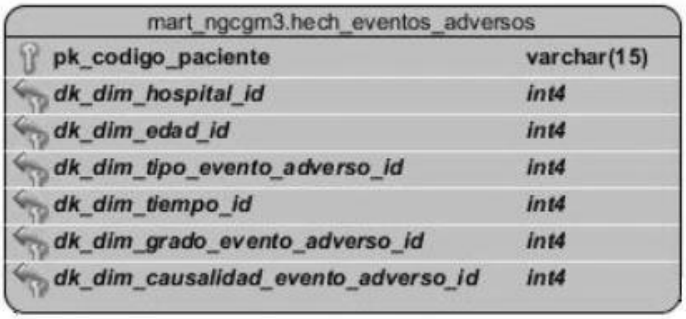


Figura 12. *Hecho eventos adversos.*

Matriz bus o matriz dimensional

La matriz bus describe las relaciones entre los hechos y las dimensiones del MD, permitiendo determinar el impacto que provocaría un cambio en la solución durante el desarrollo del sistema. Las celdas marcadas indican que la columna del hecho está relacionada con la fila de la dimensión.

A continuación se muestra la Matriz bus para el MD N Glicolil GM3.

Dimensiones/Hechos	hech_ evaluaciones _ec058	hech_ evaluaciones _ec066	hech_ evaluaciones _ec067	hech_ eventos _adversos	hech_ salida	hech_ examenes _fisicos	hech_ examenes _laboratorio	hech_ evaluacion _lesiones	hech_ imagenologia	hech_ inmunizacion
dim_provincia	X		X							
dim_tiempo	X	X	X	X	X	X	X	X	X	X
dim_raza	X									
dim_edad	X	X	X	X		X	X	X	X	X
dim_hospital	X	X	X	X	X	X	X	X	X	X
dim_signos_vitales										X
dim_tratamientos_previos	X	X	X							
dim_dosificacion										X
dim_examen_fisico						X				
dim_examen_laboratorio							X			
dim_tipo_eventos_adversos				X						
dim_interrupcion_tratamiento					X					
dim_estadio	X	X	X							
dim_tiempo_sobrevida					X					
dim_metodos_diagnostico									X	
dim_evaluacion_lesion								X		
dim_respuesta_global								X		
dim_clasificacion_tnm	X	X	X							
dim_grado_evento_adverso				X						
dim_causalidad_evento_adverso				X						

Tabla 3. *Matriz bus.*

Modelo de datos

Un modelo de datos es un lenguaje utilizado para la descripción de una BD. Estos permiten describir las estructuras de datos, o sea, el tipo de datos que se incluyen en la base de datos y la forma en que se relacionan los mismos; las restricciones de integridad, que no son más que las condiciones que los datos deben cumplir para reflejar correctamente la realidad deseada.

En el capitulo anterior se realizó un estudio de las tipologías de esquemas existentes para diseñar los modelos dimensionales, en el modelo de datos diseñado para el presente subsistema de almacenamiento e integración se evidencia el uso de la topología constelación de hechos, ya que en dicho modelo existen varias tablas de hechos que comparten dimensiones entre ellas. En la siguiente figura se muestra un fragmento del modelo de datos diseñado, para una mejor comprensión del mismo puede consultar el artefacto “Especificación del modelo de datos.doc” dentro del Expediente de Proyecto de los Subsistemas de almacenamiento e integración N Glicolil GM3.

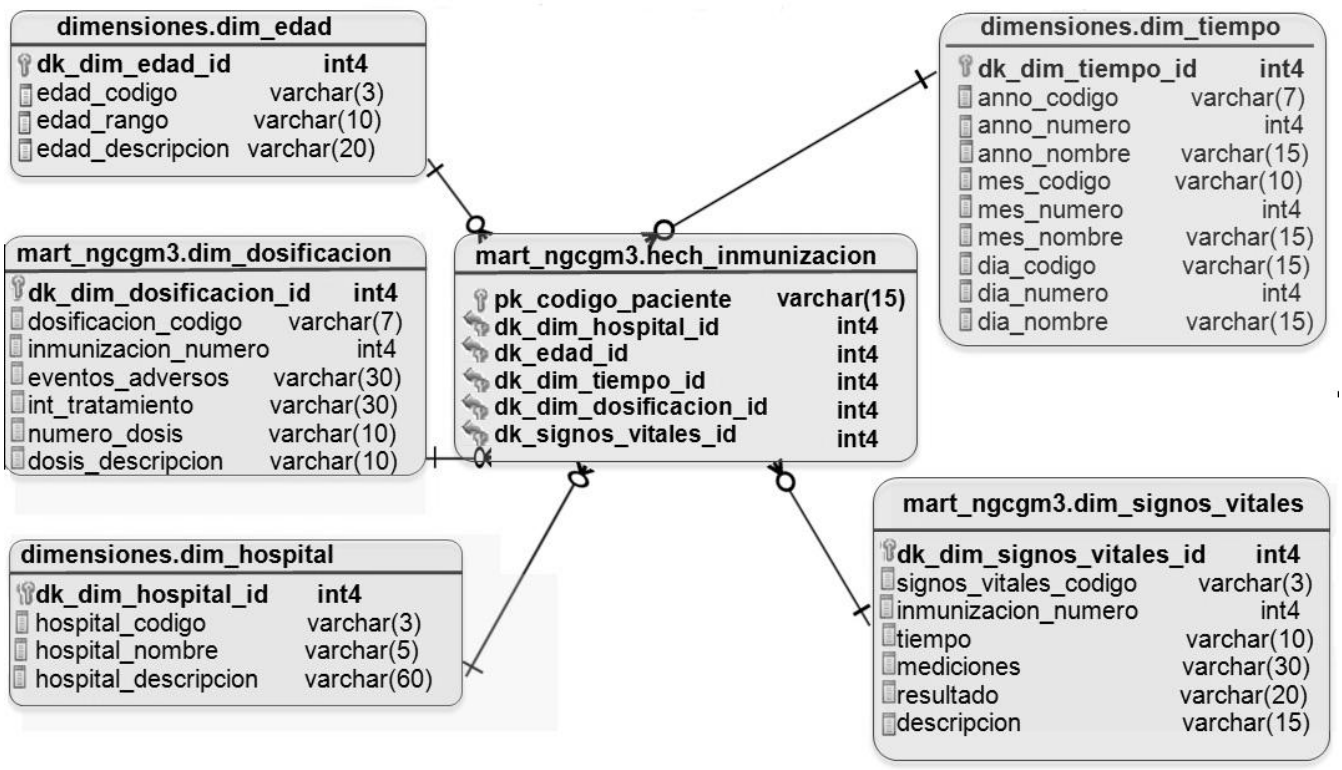


Figura 13. *Modelo de datos.*

2.8.2 Subsistema de integración

El perfilado de los datos y la extracción de los mismos desde los sistemas fuentes constituyen elementos esenciales para lograr el diseño del subsistema de integración, estos datos son sometidos a un conjunto de transformaciones, de manera que fuentes separadas puedan ser aprovechadas conjuntamente.

Perfilado de datos

El perfilado de datos es un precursor necesario para diseñar cualquier tipo de sistema para utilizar información. El mismo emplea métodos analíticos para examinar estos, con el fin de desarrollar una comprensión profunda del contenido, estructura y calidad de los mismos. Un buen sistema de perfilado puede procesar grandes cantidades de información, y con las habilidades del analista, descubre todo tipo de cuestiones que deben abordarse. El perfilado es un examen sistemático de la calidad, el alcance y el contexto de una fuente de datos, el cual permite que se construya un sistema de ETL [20].

Se realizó un análisis de todos los campos de la fuente, detectando diversos errores descritos en el

artefacto “Perfil de los Datos.doc” en el Expediente de Proyecto de los Subsistemas de almacenamiento e integración N Glicolil GM3. Conjuntamente se encuentra reflejado de los diferentes campos la longitud de la cadena, la cantidad de valores nulos, los valores mínimos y máximos. El proceso de perfilado de datos permitió identificar los tipos de datos contenidos en los diferentes campos como se muestra en el siguiente gráfico:

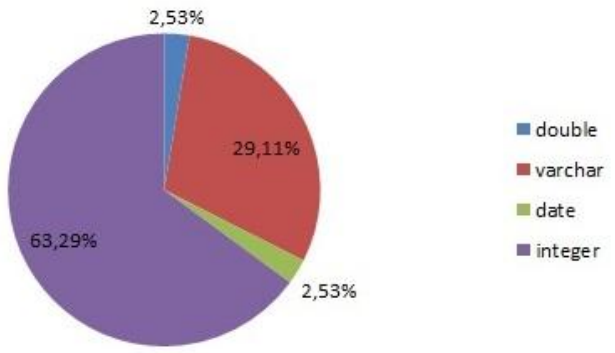


Figura 14. *Porcentaje de los tipos de datos.*

Diccionario de datos

El diccionario de datos contiene la información necesaria para el correcto entendimiento del sistema fuente, enfocado principalmente en las variables que interactúan con la fuente de datos, también es utilizado para definir una correspondencia entre los valores de la fuente y los que manejan el MD. Este es recogido en el artefacto “Diccionario de Datos.xls” dentro de Expediente de Proyecto de los Subsistemas de almacenamiento e integración N Glicolil GM3, este artefacto recoge cada una de las variables descritas con los posibles valores que pueden tomar.

Nombre de las variables	Descripción de las variables
tipo_respuesta	Se refiere a los tipos de respuestas que presenta paciente ante el tratamiento.
raza	Se refiere a la raza de los pacientes.
sexo	Se refiere al sexo de los pacientes.
provincia	Se refiere a las provincias.
tiempo	Se refiere al año, al mes y al día.
edad	Se refiere a la edad de los pacientes.
t	Se refiere al tumor primario.
n	Se refiere a los ganglios linfáticos regionales.

m	Se refiere a la metástasis distante.
estadio	Se refiere a la etapa en que se encuentra el cáncer.
localización	Se refiere al lugar donde el paciente presenta el cáncer.
nivel_dosis	Se refiere a la cantidad de medicamento (ml) que se le suministra al paciente.
tratamiento	Se refiere al tipo de tratamiento que se le está dando al paciente.
numero_dosis	Se refiere al número de dosis del compuesto vacunal que se le suministra al paciente
tipo_evento_adverso	Se refiere al tipo de evento adverso que presenta el paciente.
intensidad_evento_adverso	Se refiere a la intensidad del evento adverso que presenta el paciente.
causalidad_evento_adverso	Se refiere a la causalidad del evento adverso que presenta el paciente.
grado_evento_adverso	Se refiere al grado del evento adverso que presenta el paciente.
tratamientos previos	Se refiere al grado de diferenciación que tiene el paciente.
resp_event_adv	Se refiere a la respuesta del evento adverso del paciente.
examenes_lab	Se refiere al tipo de examen de laboratorio realizado al paciente.
examenes_fisicos	Se refiere a al tipo de examen físico realizado al paciente.

Tabla 4. *Diccionario de datos.*

Diseño general de las transformaciones

Las transformaciones son de gran importancia en el proceso ETL, estas permiten cargar correctamente los datos al mercado en las tablas correspondientes. Una vez acabado el perfilado de los datos fuente, se procede a realizar el diseño de las transformaciones. Estos diseños son flexibles en muchas ocasiones a la hora de implementar las transformaciones, debido al surgimiento de algunas situaciones con los datos extraídos de las fuentes y sería necesario establecer un conjunto de estrategias para resolverlas.

En el diseño general para implementar las transformaciones se establecen primeramente las variables de entorno que se utilizarán en los cambios, luego se verifica la conexión a la BD con el objetivo de verificar si esta se encuentra disponible. Seguidamente se realiza la extracción de los datos de la fuente obteniendo la información existente. Posteriormente se verifica la existencia de datos nulos, en caso de que existan se les realiza las estrategias ETL correspondientes. Se procede a buscar las llaves dimensionales, una vez obtenidas estas, se valida que no estén huérfanas, en caso de que existan se almacenan en el esquema metadatos en la tabla correspondiente a las mismas, luego se insertan los datos en la tabla del hecho correspondiente, una vez concluida la inserción se obtiene la información del sistema acerca del nombre de la transformación ejecutada, dirección IP donde se ejecutó y la fecha, esta será almacenada en la tabla gestión de carga histórica del esquema metadatos, todos los datos

correspondiente a los procesos internos de los cambios se almacenará en la tabla gestión de procesos del esquema metadatos. En la siguiente figura se muestra el diseño general de las transformaciones del Subsistema de almacenamiento e integración N Glicolil GM3.

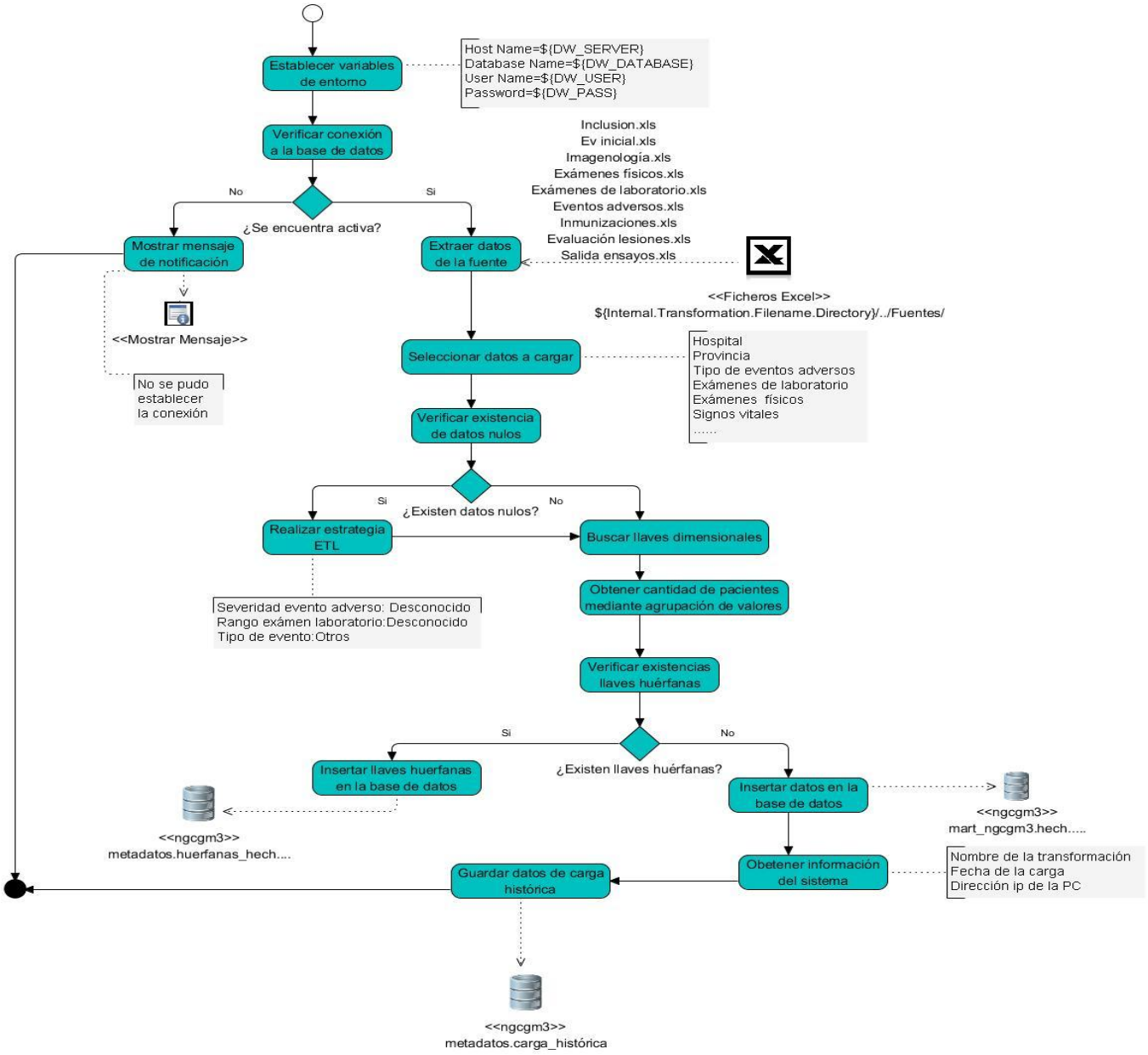


Figura 15. *Diseño general de las trasformaciones de los hechos.*

2.9 Política de respaldo y recuperación

Garantizar la persistencia y seguridad de la información es un objetivo primordial para cualquier entidad, la pérdida de la misma trae consigo consecuencias embarazosas ya sea una gran pérdida monetaria o la pérdida de prestigio de dicha entidad. Con el objetivo de lograr la persistencia y seguridad de la información que se maneja en el CIM y por la importancia que la misma provee a los especialistas de esa organización, se establece una política de respaldo y recuperación la cual comprende dos elementos esenciales:

- ✓ Verificar la existencia de las salvas: Verificar la existencia de una copia de toda la información almacenada. Las tablas involucradas en el proceso son las diez tablas de hechos identificadas con las veinte tablas de dimensiones asociadas.
- ✓ Backups existentes: En esta área actualmente no existen backups, se debe realizar una copia de los datos en un dispositivo de almacenamiento.

2.10 Esquema de seguridad

Es de gran importancia para un sistema de información implementar un mecanismo de protección contra aquellas acciones que puedan afectar la integridad, confidencialidad y disponibilidad de los datos almacenados. Por tal motivo, para el acceso al MD es necesario definir los roles que tendrán acceso a la DB.

- ✓ Administrador ETL: Realiza los procesos de ETL sobre los datos y tiene permisos de lectura y escritura sobre los esquemas pertenecientes al subsistema de almacenamiento e integración de los EC del producto N Glicolil GM3.

2.11 Conclusiones

Luego de haber realizado el análisis y diseño del subsistema de almacenamiento e integración N Glicolil GM3 posibilitó arribar a las siguientes conclusiones:

- ✓ Las entrevistas realizadas al cliente, el estudio de los protocolos y las fuentes de datos proporcionadas, permitió la identificación de diez RI, dos RF, cinco RNF y 12 RN.
- ✓ Se identificaron 12 CUS, dos CUF y diez CUI.
- ✓ Mediante el diseño del modelo de datos se identificaron 20 dimensiones y diez hechos con una medida asociada a cada una de ellos, que garantizan el correcto funcionamiento del sistema.

- ✓ El diseño de la arquitectura para la solución contará con dos subsistemas: el de almacenamiento y el subsistema de integración.
- ✓ El perfilado de datos realizado permitió conocer el estado de la fuente, permitió además establecer nuevas RN aplicables durante el proceso de transformación.
- ✓ El diseño general de las transformaciones proporcionó una aproximación a los pasos que se deben realizar para lograr la estandarización de la información y su almacenamiento.
- ✓ Las políticas de recuperación y respaldo establecidas contribuyen a mantener la integridad y seguridad de los datos almacenados.

CAPÍTULO 3: IMPLEMENTACIÓN Y VALIDACIÓN DE LOS SUBSISTEMAS DE ALMACENAMIENTO E INTEGRACIÓN

3.1 Introducción

En este capítulo se implementan los procesos definidos en el diseño. Se realiza la implementación de cada uno de los subsistemas de almacenamiento e integración, que conforman el MD. También se valida la solución mediante listas de chequeos y los casos de prueba, para garantizar la calidad del producto.

3.2 Implementación del subsistema de almacenamiento

Una vez diseñado el modelo dimensional, prosigue la implementación del subsistema de almacenamiento. Este proceso incluye estándares de codificación de las estructuras, para facilitar la comprensión de los nombres definidos en cada uno de los esquemas.

3.2.1 Estándares de codificación

Los estándares de codificación se utilizan para lograr un entendimiento entre las partes implicadas en un proyecto. Tienen como objetivo estandarizar la forma de las estructuras del subsistema de almacenamiento e integración del producto y lograr utilizar un patrón que conduzca a la correcta normalización de los términos utilizados. De esa forma permite a los desarrolladores un mejor entendimiento de las estructuras utilizadas.

En la solución la nomenclatura se mantiene atendiendo a la clasificación de las diferentes estructuras, teniendo en cuenta si es una tabla de hechos o una dimensión. Si la tabla es una dimensión, al nombre de la misma le preceden las letras “dim” separadas del nombre de la dimensión por el caracter “_”, ejemplo “dim_tratamientos_previos”. En caso de ser una tabla de hechos, como prefijo se ubican las letras “hech”, igualmente separadas del nombre de la tabla, por el caracter “_”, ejemplo “hech_eventos_adversos”.
Para los atributos de las dimensiones se siguió la misma política para cada una de ellas. En el caso de las llaves de las dimensiones se les denominó “dk_dim_nombre-dimension_id”. Para el caso de que el atributo de la misma sea un código del negocio se le especificó como “nombre-dimension_codigo”, igualmente para los nombres, descripciones u otros atributos: “nombre-dimension_resultado” y “nombre-dimension_descripcion” respectivamente. De manera general los atributos fueron nombrados como “nombre-dimension_atributo”. Las medidas fueron definidas de la forma “cant_medida”, por ejemplo “cant_pacientes”.

Luego de finalizar el proceso de estandarización de las nomenclaturas a utilizar para cada una de las tablas del negocio, atributos y medidas dentro de la BD, seguidamente se procede a la implementación del modelo de las estructuras físicas.

3.2.2 Implementación del modelo de datos físico

El modelo de datos físico constituye una colección integrada de entidades que describe las estructuras de los datos. Dicho modelo se genera a partir del modelo lógico dimensional mostrado en el capítulo anterior. Para la solución se cuenta con 30 tablas, estas se encuentran divididas en 20 tablas de dimensiones y diez tablas de hechos. Se define la utilización de tres esquemas:

- ✓ El esquema “dimensiones” contiene las dimensiones compartidas con los demás mercados del CIM, del almacén central.
- ✓ El esquema “mart_ngcgm3” contiene las tablas de hechos y las dimensiones propias del subsistema de almacenamiento e integración del producto N Glicolil GM3.
- ✓ El esquema “metadatos” contiene los metadatos, llaves huérfanas y carga histórica que son metadatos técnicos y los metadatos de procesos del subsistema de almacenamiento e integración del producto N Glicolil GM3.

Figura 16. *Esquemas y tablas del modelo físico.*

3.3 Implementación del subsistema de integración

La implementación del subsistema de integración encierra los procesos de ETL. Para llevar dichos procesos es recomendable haber realizado un análisis de las fuentes de datos, para identificar los principales problemas que esta presenta.

Inicialmente se extraen los datos de las fuentes, seleccionando los que aportan información significativa al negocio. Luego se realiza la transformación y limpieza de los datos para cargar la información hacia la BD. En el proceso de carga de las dimensiones y hechos los mismos son transmitidos del modelo temporal hacia el subsistema de almacenamiento e integración, teniendo como salida las tablas correspondientes en la BD relacional.

Es primordial tener en cuenta el uso de los subsistemas propuestos por Kimball para lograr el correcto funcionamiento del mismo. A continuación se mencionan y describen los subsistemas utilizados en la implementación de la aplicación.

- ✓ Perfilado de datos: Permitió examinar los datos para verificar su calidad y el cumplimiento de los estándares conforme a los requisitos especificados por el cliente. A través de este subsistema fueron definidas nuevas reglas de transformación.
- ✓ Sistema de extracción: Permitió la extracción de los datos desde la fuente origen para su transformación y posterior carga.
- ✓ Subsistema de transformación: Posibilitó realizar transformaciones como el mapeo de valores, el cambio de datos en algunos campos, el filtrado de valores, entre otras.
- ✓ Subsistema de carga: Permitió realizar la carga de los datos a las tablas de dimensiones y hechos del Subsistemas de almacenamiento e integración N Glicolil GM3.
- ✓ Llave subrogada: Permite crear llaves subrogadas independientes para cada tabla.
- ✓ Repositorio de metadatos: Captura los metadatos de los procesos de ETL, de los datos de negocio y de los aspectos técnicos.
- ✓ Rastreo de eventos de errores: Permite capturar errores que proporcionan información valiosa sobre la calidad de los datos y posibilita mejorarlos.
- ✓ Control de versiones: Permite hacer control de versiones del proyecto ETL y de los metadatos asociados, permitiendo conocer cuando se realizó la última carga.

3.3.1 Implementación de las transformaciones

En el presente trabajo se realizaron transformaciones, estas están compuestas por pasos enlazados entre sí a través de saltos. Estos simbolizan el componente más pequeño dentro de las transformaciones, y por

ellos fluye la información que se obtiene de la fuente. De los saltos emana la información entre los distintos pasos (siempre es la salida de un paso y la entrada de otro). Estos pasos se encuentran agrupados por categorías y cada uno de ellos ha sido diseñado para cumplir una función determinada. Cada paso tiene una ventana de configuración específica, donde se determina los elementos a tratar y su forma de comportamiento.

Las transformaciones se realizaron teniendo en cuentas las RN identificadas. Se ejecutaron 19 transformaciones, nueve para cargar las dimensiones propias del MD y diez para los hechos. Todas estas transformaciones tienen el objetivo de poblar los Subsistemas de almacenamiento e integración N Glicolil GM3.

Para realizar la carga de las dimensiones (Figura 17) primeramente se generan los datos debido a que los EC se encuentran cerrados, luego realiza una unión por una clave generada, con el objetivo de obtener un producto cartesiano entre los campos. Una vez concluido este paso se genera el código de los datos y finalmente se insertan en la BD.

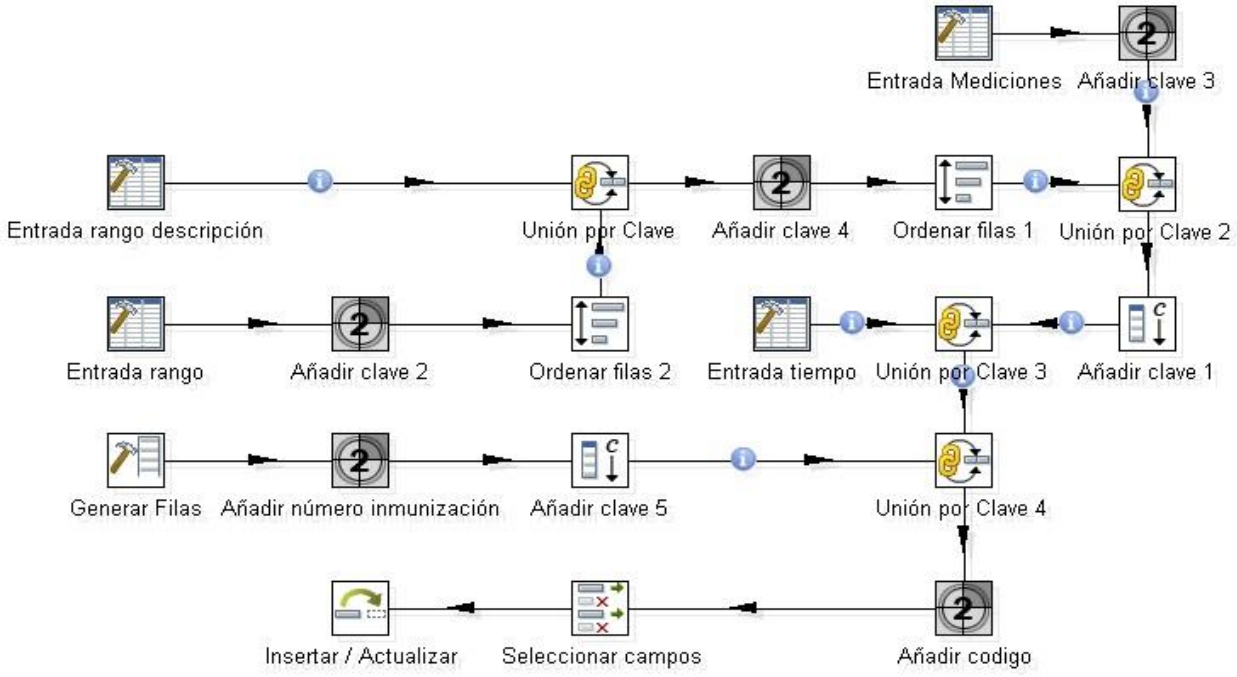


Figura 17. *Transformación para cargar la dimensión signos vitales.*

En la Figura 18 se muestra el flujo de la transformación que contiene los pasos que se siguieron para la carga del hecho hech_inmunización. En este flujo se hace uso del esquema metadatos, el cual posibilita la

gestión de la carga histórica de los datos y tratamiento de llaves huérfanas, que permite conocer el momento en que ocurrió la última carga realizada y la gestión de errores respectivamente. Inicialmente se extraen los datos, luego se unen y normalizan los valores extraídos. Una vez normalizado los datos se realiza el tratamiento de llaves nulas y el mapeo de valores. Luego se hace una búsqueda de llaves dimensionales en la BD, las cuales serán validadas, permitiendo así insertar las mismas en la BD. En caso de que estas llaves contengan valores nulos se almacenarán en la tabla de llaves huérfanas correspondiente a este hecho.

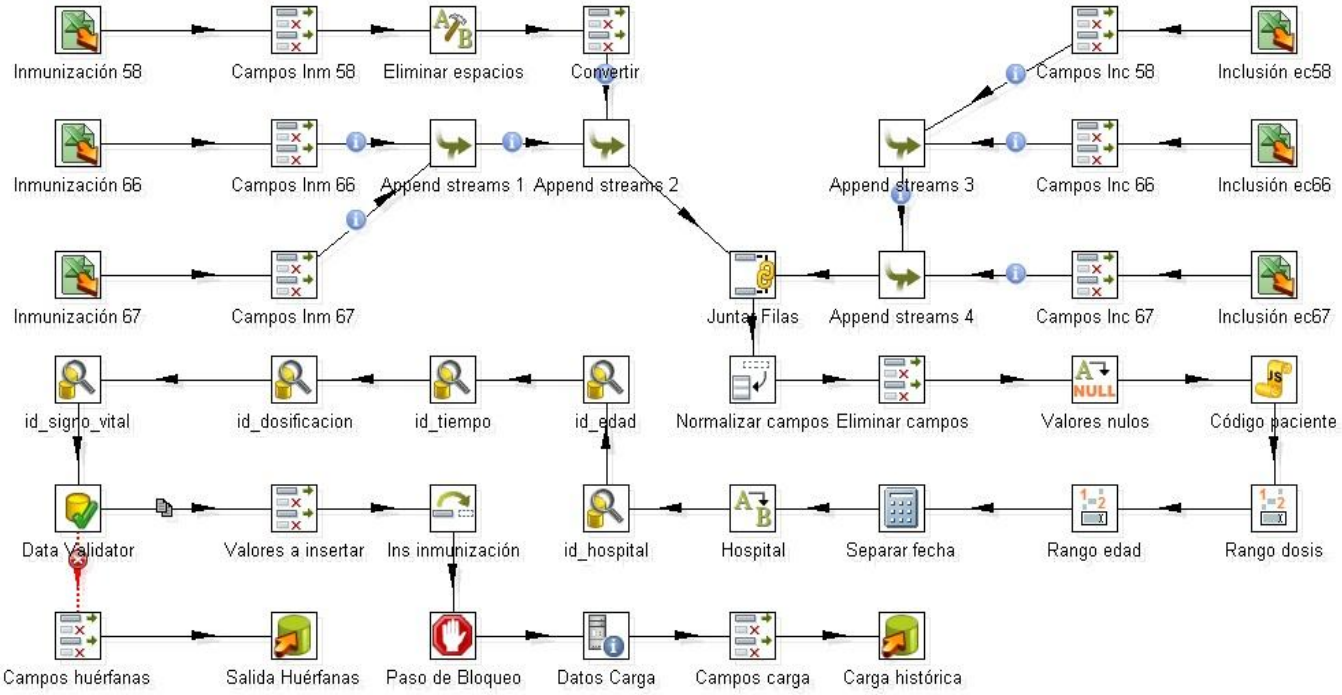


Figura 18. *Transformación para cargar el hecho inmunización.*

3.3.2 Implementación de los trabajos

En el contexto de integración de datos, el término trabajo o “Job”, no es más que un conjunto de tareas con el objetivo de realizar una acción determinada. Estos permiten la ejecución de una o varias transformaciones, las cuales siguen una secuencia lógica de pasos. Los trabajos se encuentran en un nivel superior a las transformaciones. Los saltos o “hops” entre los elementos de un “Job” indican el orden de ejecución de cada uno de ellos (teniendo en cuenta que no comienza la ejecución del componente siguiente hasta que el anterior no haya concluido su ejecución).

Se realizaron en la investigación cuatro trabajos, de los cuales tres se realizaron para cargar los hechos evaluaciones_exámenes, inmunización_imagenología y lesiones_salida y finalmente se realizó un trabajo general que se encarga de ejecutar los trabajos mencionados anteriormente. En la Figura 19 se muestra la estructura de las transformaciones y los trabajos realizados y en la Figura 20 se muestra la implementación de los trabajos.

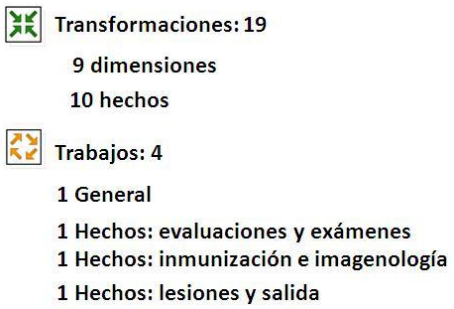


Figura 19. *Estructura de las transformaciones y de los trabajos.*

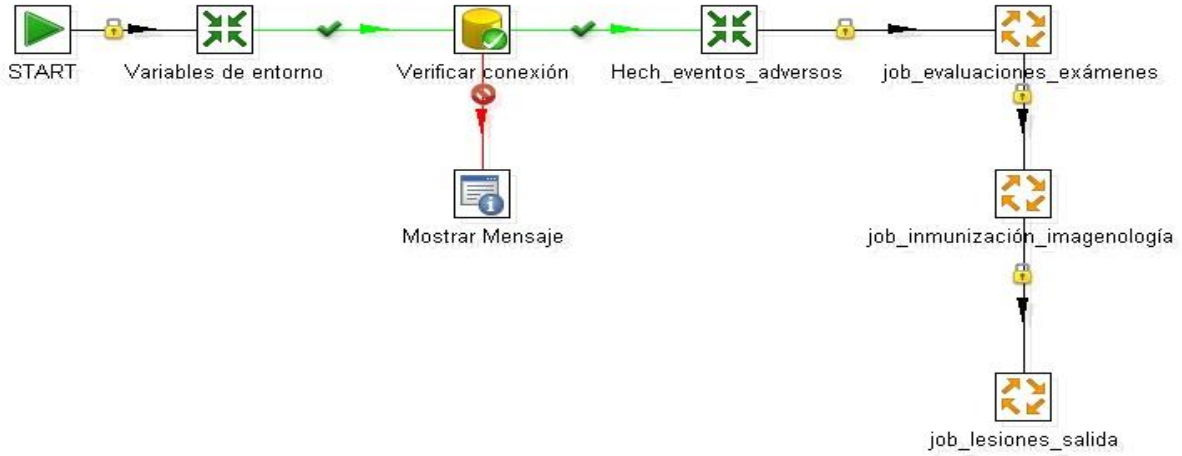


Figura 20. *Implementación de los trabajos.*

3.3.3 Gestión del cambio lento en las dimensiones

Las dimensiones lentamente cambiantes o SCD (*Slowly Changing Dimensions*) determinan cómo se manejan los cambios históricos en las tablas de dimensiones. Estos cambios pueden ocurrir de forma ocasional o constantemente e implicar a un solo registro o a la tabla completa. Cuando ocurren estos cambios, se puede optar por seguir alguna de estas dos opciones:

- ✓ Registrar el historial de cambios.

- ✓ Reemplazar los valores que sean necesarios.

Ralph Kimball planteó tres estrategias a seguir para las SCD, estas son las SCD tipo 1, 2 y 3. Posteriormente se ha profundizado en estas estrategias y creado nuevas como los tipos 0, 4 y 6. A continuación se describen las mismas:

- ✓ Tipo 0: No se tiene en cuenta la gestión histórica y no se realiza esfuerzo alguno para lidiar con los problemas del cambio de la dimensión. De este modo alguna información es sobrescrita mientras que otra queda sin cambios.
- ✓ Tipo 1 (sobrescribir): Es utilizado cuando la información histórica no es importante. Este tipo sobrescribe los datos antiguos con nuevos, es utilizado mayormente para corregir errores de datos en las dimensiones. A pesar de ser fácil de implementar presenta como desventaja principal que no permanece ningún registro histórico en la dimensión.
- ✓ Tipo 2 (añadir fila): Cuando hay un cambio se crea una nueva entrada en la tabla. Al nuevo registro se le asigna una nueva llave subrogada y a partir de este momento será el valor usado para futuras entradas, las antiguas usarán el valor anterior. En este modo se gestiona un versionado que puede incluir fechas para indicar los períodos de validez, así como numeradores de registros o indicadores de registros activos o no. Este tipo permite guardar toda la información histórica en el almacén de datos.
- ✓ Tipo 3 (añadir columna): Esta estrategia requiere que se agregue una nueva columna a la tabla por cada columna cuyos valores se desea mantener un historial de cambios. De este modo en la nueva columna se coloca el valor antiguo antes de sobrescribir el valor actual con el nuevo. Este tipo presenta como principal desventaja que solo permite guardar un historial limitado de los datos, dependiendo del número de columnas que se cree.
- ✓ Tipo 4 (tabla de historia separada): Su función es almacenar en una tabla adicional los detalles de cambios históricos realizados a la tabla de dimensión. La tabla con la información histórica indicará el tipo de operación que se ha realizado, sobre qué campo se realizó el cambio y la fecha del mismo. Esta tabla tiene como objetivo mantener un detalle de los cambios realizados.
- ✓ Tipo 6 (híbrido): Este método combina los tipos anteriores 1, 2 y 3; y se le denomina tipo 6 debido a la suma de los tres tipos que integra (1+2+3=6). Esta estrategia utiliza el Tipo 1 (sobrescribir) junto con el Tipo 2 (añadir filas) y el Tipo 3 (añadir columnas), añadiendo además una pareja adicional de columnas para indicar el rango de fechas al cual aplica cada fila en particular [21].

Para la solución se implementó la gestión de cambio lento para las dimensiones mediante el tipo 0, ya que la fuente no se le agregarán más datos, por lo que no se realizará una carga incremental debido a que los EC del producto N Glicolil GM3 se encuentran cerrados.

3.3.4 Gestión de los metadatos

Los metadatos son datos que ayudan a identificar, describir y localizar recursos digitales, son información estructurada que describe y/o permite encontrar, gestionar, controlar y entender o preservar otra información; o sea, que no son más que datos sobre los propios datos. Estos se utilizan para describir recursos y no se limitan a un tipo de formato, sino que cubren una amplia gama de recursos, además pueden describir una colección en general, un recurso en particular o un solo elemento.

Los metadatos presentan entre sus funciones básicas proporcionar una descripción de una entidad u objeto de la información a través de otra información necesaria para su manejo y preservación, además de proporcionar puntos de acceso a esa descripción y codificar la misma. Estos pueden ser agrupados en tres categorías:

- ✓ Metadatos técnicos: Están relacionados con la función de un sistema o el modo en que se interrelacionan sus componentes.
- ✓ Metadatos del negocio: Posibilita obtener los datos y la información referente a los aspectos del negocio, como son los datos provenientes de la fuente.
- ✓ Metadatos de proceso: Permiten obtener información de los procesos que se ejecuten. Es la presentación de las estadísticas sobre los resultados de la ejecución del proceso de ETL, incluyendo medidas tales como filas cargadas con éxito, filas rechazadas y la cantidad de tiempo de carga; es muy importante en el proceso de limpieza de metadatos [21].

En la investigación se hizo uso fundamentalmente de los metadatos técnicos para el tratamiento de llaves huérfanas y la gestión de carga histórica y los metadatos de proceso para obtener la información correspondiente a los procesos de las transformaciones.

3.4 Pruebas

Todo proceso de creación de un software está sujeto a fallos, es por esto que las pruebas de software constituyen una fase importante en el desarrollo de este, ya que permiten comprobar que no existan fallos en la implementación del mismo, proporcionándole la calidad requerida al software y elevar de esta forma

la reputación del equipo de desarrollo. El proceso de pruebas comienza con la planificación de las mismas, seguidamente la ejecución, control y como paso final la evaluación.

Diversos son los tipos de pruebas que pueden ser aplicadas a un MD, haciendo uso de herramientas que validan estos procesos. La metodología adoptada por el departamento define las siguientes pruebas a utilizar.

3.4.1 Modelo V

El Modelo V fue definido por CALISOFT y es utilizado en el centro DATEC con el propósito de crear un estándar para comprobar que el producto cumpla con las especificaciones del negocio. El Modelo V demuestra cómo se relacionan las actividades de prueba con las de análisis y diseño. En la siguiente figura se muestra el Modelo V representando las fases del ciclo de vida del modelo:

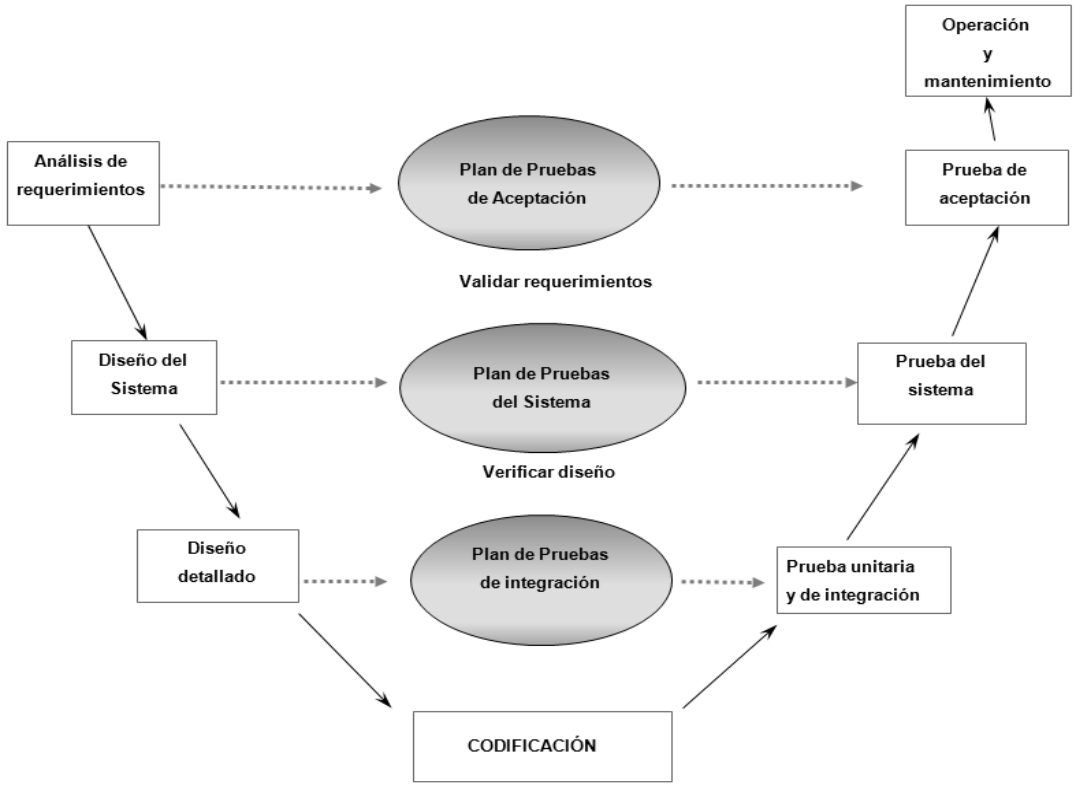


Figura 21. *Modelo V.*

En la Figura 21 se puede observar la punta de la V que es la codificación, a la izquierda del mismo se puede detallar las etapas de desarrollo del software y a la derecha de este las pruebas correspondientes a cada etapa.

3.4.2 Pruebas unitarias

Permiten probar el correcto funcionamiento de un componente o subsistema específico y son desarrollados por los propios desarrolladores durante la implementación [23]. Luego de haber concluido la etapa de implementación se realizaron las pruebas unitarias a los subsistemas de almacenamiento e integración, detectando diez no conformidades (NC) cuatro pertenecientes al subsistema de almacenamiento y seis al subsistema de integración, las cuales fueron resueltas inmediatamente luego de su detección:

Subsistema de almacenamiento:

NC1. Realizar control de versiones en todo el expediente de proyecto.

NC2. Revisar la aditividad de las variables de salida.

NC3. Revisar los RF y los RNF.

NC4. Los artefactos RN y Especificación de requisitos de software se encuentran desactualizados, por lo que la revisión no pudo completarse correctamente.

Subsistema de integración:

NC5. Modificar los nombres de las actividades y objetos en los diseños de los proceso de integración de datos, para que reflejen su verdadero objetivo.

NC6. Eliminar el perfilado de datos como una actividad en el diseño de los procesos de integración de datos.

NC7. Incluir las actividades que faltan en el diseño de los procesos de integración de datos.

NC8. Revisar el orden lógico de las actividades y su relación con otras actividades.

NC9. Revisar la estrategia de gestión de metadatos tanto en el diseño como en la implementación.

NC10. Revisar los nombres de las transformaciones y sus componentes.

3.4.3 Pruebas de integración

Permite verificar la correcta integración de los componentes y subsistemas que conforman la solución. Pone a prueba la vista arquitectónica del sistema de definida en una infraestructura de desarrollo. Estas pruebas son ejecutadas por los arquitectos de software. No son verdaderamente pruebas de sistema

debido a que los componentes no se encuentran implementados en el ambiente operativo [23].

En la presente investigación se validó que los datos ubicados en la fuente fueran cargados completamente. Se realizaron consultas a la BD donde se obtuvo un resultado satisfactorio, demostrando de esta manera que los datos de fuente fueron cargados en su totalidad. La estrategia definida para realizar estas pruebas incluye la confección de casos de prueba que resultan de gran importancia para demostrar la funcionalidad del mismo. A continuación se muestra un ejemplo de una de las consultas realizadas al CU Mantener_disponible_información_pacientes_eventos_adversos_producto_NGCGM3. Las restantes consultas se encuentran en el artefacto Casos de pruebas de integración ubicado en el expediente de proyecto.

SELECT count (distinct hech_eventos_adversos.pk_codigo_paciente)

FROM mart_ngcgm3.hech_eventos_adversos, dimensiones.dim_causalidad_evento_adverso

WHERE hech_eventos_adversos.dk_dim_causalidad_evento_adverso_id =

dim_causalidad_evento_adverso.dk_dim_causalidad_evento_adverso_id AND

dim_causalidad_evento_adverso.causalidad_evento_adverso = 'Definitiva';

Mediante esta consulta se obtuvo la cantidad de pacientes con causalidad de evento adverso definitiva.

3.5 Herramientas para la aplicación de las pruebas

Dentro de las herramientas utilizadas para aplicar los distintos tipos de pruebas se tienen los casos de prueba y las listas de chequeo.

3.5.1 Casos de pruebas

Mediante los casos de prueba el analista podrá determinar si el grado de cumplimiento de los requisitos de una aplicación es parcial o completamente satisfactorio. En los subsistemas de almacenamiento e integración N Glicolil GM3 se diseñaron diez casos de prueba correspondientes a los casos de uso de información descritos en la etapa de análisis.

Los casos de prueba son esenciales para todas las actividades de pruebas porque son la base para diseñar y ejecutar los procedimientos de pruebas. Reflejan trazabilidad con los CU, ya que estos muestran una secuencia ordenada de eventos, al describir flujos básicos, flujos alternos, precondiciones y postcondiciones. Si los casos de prueba no son correctos, la calidad del sistema se pone en duda y las

pruebas dejan de ser confiables.

En el Anexo 1 se puede encontrar un ejemplo correspondiente al Caso de pruebas de integración, perteneciente al CU Mantener_disponible_información_pacientes_eventos_adversos_producto_

NGCGM3. Los restantes casos de pruebas se pueden consultar en el expediente de proyecto.

3.5.2 Listas de chequeos

La lista de chequeo cuenta con una serie de preguntas, mediante el cual se verifica el grado de cumplimiento de determinadas reglas establecidas para los procesos de desarrollo del sistema, además de medir la calidad de los artefactos generados durante la realización del producto. Para evaluar los Subsistemas de almacenamiento e integración N Glicolil GM3 fueron definidas cuatro listas de chequeo (Lista de chequeo del Mapa Lógico de Datos, Lista de chequeo de Registro de Sistema Fuente, Lista de chequeo del Perfilado de Datos y Lista de chequeo del Diccionario de Datos).
Esta evaluación se desarrolla a través del análisis de un grupo de indicadores, distribuidos en tres secciones fundamentales:

- ✓ Estructura del documento: Abarca todos los aspectos definidos por el expediente de proyecto o el formato establecido por el proyecto.
- ✓ Indicadores definidos: Abarca todos los indicadores a evaluar durante la etapa de desarrollo.
- ✓ Semántica del documento: Contempla todos los indicadores a evaluar respecto a la ortografía, redacción y demás.

Las listas de chequeo están estructuradas por los siguientes elementos:

- ✓ Peso: Define si el indicador a evaluar es crítico o no. El mismo se describe con una C si es crítico.
- ✓ Indicadores a evaluar: Constituyen los indicadores a evaluar en las secciones Estructura del documento, Semántica del documento e Indicadores definidos para el artefacto a evaluar.
- ✓ Evaluación: Es la forma de evaluar el indicador en cuestión. El mismo se evalúa de uno en caso de que exista alguna dificultad sobre el indicador y de cero, en caso de que el indicador revisado no presente problemas.
- ✓ N.P. (No Procede): Se usa para especificar que no es necesario evaluar el indicador en ese caso.

- ✓ Cantidad de elementos afectados (CEA): Especifica la cantidad de errores encontrados sobre el mismo indicador.
- ✓ Comentario: Especifica los señalamientos o sugerencias que quiera incluir la persona que aplica la lista de chequeo. Pueden o no existir señalamientos o sugerencias.

En el Anexo 2 se muestra la lista de chequeo correspondiente al Mapa Lógico de Datos. Las restantes listas de chequeo se pueden consultar en el expediente de proyecto. La siguiente tabla muestra un resumen de los resultados obtenidos luego de haber aplicado las listas de chequeo a los artefactos de ETL.

Artefactos	Estructura	Indicadores	Semántica	Total indicadores	Indicadores críticos	NC
Registro sistema fuente	9	1	3	13	5	0
Perfilado de datos	8	1	3	12	5	0
Diccionario de datos	9	1	3	13	5	1
Mapa lógico de datos	5	1	3	9	5	1

Tabla 5. *Aplicación de las listas de chequeo a los artefactos de ETL.*

En la siguiente gráfica se muestra el comportamiento de los indicadores para la lista de chequeo realizada los artefactos de ETL.

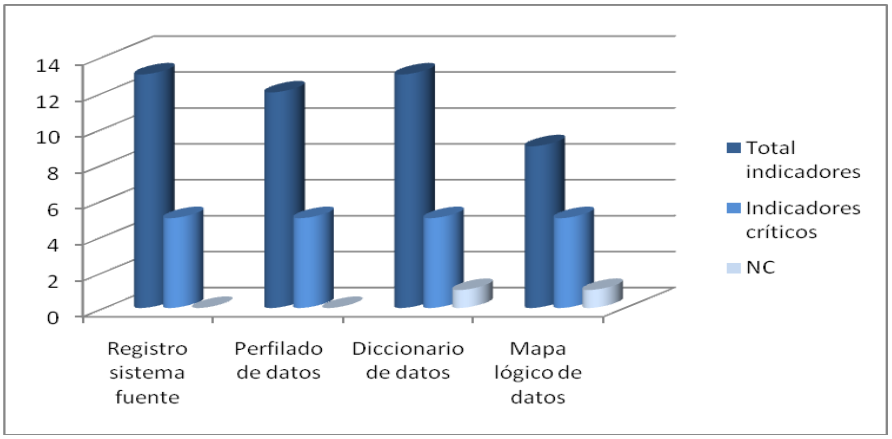


Figura 22. *Resultados de la aplicación de las listas de chequeos a los artefactos de ETL.*

3.6 Calidad de datos

El proceso de calidad de datos es de vital importancia para el desarrollo del MD, ya que permite comprobar que los datos cargados no posean errores. Una vez culminado el proceso de integración y

carga de los datos, se realizó el perfilado de los datos que fueron cargados a los subsistemas de almacenamiento e integración, con el objetivo que estos adquiriesen la calidad requerida. Este proceso de perfilado permite obtener estadísticas e información sobre los datos, lo cual posibilita corregir problemas que pudiesen existir, como son valores escritos incorrectamente, duplicados o nulos.

3.6.1 Perfilado de datos

A través del uso de la herramienta DataCleaner se obtuvieron reportes que evidenciaron un resultado positivo respecto a los datos cargados al MD. Mediante el análisis de los reportes arrojados por este proceso se pudo concluir que la carga de los datos correspondientes a cada uno de los hechos se realizó correctamente. No se almacenaron valores nulos ni vacíos, y en las tablas de hechos contienen únicamente valores numéricos.

3.7 Conclusiones del capítulo

Luego de haber realizado la implantación de los subsistemas de almacenamiento e integración N Glicolil GM3 y realizadas las pruebas al mismo con el objetivo de comprobar la calidad de los datos integrados, permitiendo concluir el proceso de construcción y validación, se arrojan las siguientes conclusiones:

- ✓ Fueron implementados completamente los dos subsistemas que componen la aplicación: subsistema de almacenamiento y subsistema de integración, teniendo como resultado la disponibilidad de la información.
- ✓ Con la implementación del subsistema de almacenamiento e integración N Glicolil GM3 quedaron definidos tres esquemas: dimensiones, mart_ngcgm3 y metadatos, posibilitando la correcta integración de los datos a la BD.
- ✓ Se realizaron diez transformaciones para la carga de los hechos, nueve para las dimensiones y se realizaron cuatro trabajos, los cuales posibilitaron la ejecución de las transformaciones.
- ✓ Con la aplicación de las listas de chequeo a los artefactos de ETL y los casos de prueba basados en los casos de uso y reglas de transformación, se logró probar que los subsistemas de almacenamiento e integración cumplen con los requisitos identificados y que los artefactos cuentan con la estructura correspondiente, obteniendo de esta forma la carta de aceptación por parte del cliente.

CONCLUSIONES GENERALES

El estudio de los principales conceptos relacionados con el desarrollo de los AD, proporcionó la elaboración del presente trabajo, cuyo resultado fue el "Subsistemas de almacenamiento e integración N Glicolil GM3 para el almacén de datos Ensayos Clínicos del CIM". Dicho resultado contribuyó a mitigar las deficiencias encontradas con respecto al manejo de la información en el departamento de EC del CIM, lo cual dificultaba el proceso de toma de decisiones. Los siguientes resultados demuestran el cumplimiento de los objetivos propuestos en la investigación:

- ✓ El estudio realizado de las metodologías, herramientas y tecnologías a utilizar en el desarrollo del subsistema de almacenamiento e integración, garantizó que la metodología seleccionada guiara el proceso de desarrollo a través de cada etapa del ciclo de vida, permitiendo la documentación de cada una de ellas. Facilitó la selección de las herramientas y tecnologías para la construcción de la solución propuesta, las cuales cumplen con la política de migración a software libre, dando soporte a las necesidades del equipo de desarrollo.
- ✓ El análisis y diseño de los subsistemas de almacenamiento e integración N Glicolil GM3, permitió arrojar como resultados los artefactos necesario para la posterior etapa de implementación.
- ✓ La implementación de los subsistemas de almacenamiento e integración N Glicolil GM3 permitió la integración de los datos históricos y su almacenamiento.
- ✓ Las pruebas realizadas durante las distintas etapas del ciclo de vida, permitieron demostrar la funcionalidad del sistema a partir de los requisitos establecidos por el cliente. Los resultados obtenidos durante las últimas pruebas fueron satisfactorios, por lo cual queda validado el cumplimiento de los objetivos propuestos.

RECOMENDACIONES

Con el propósito de mejorar la propuesta plasmada en este trabajo se sugiere:

- ✓ Utilizar la presente investigación como base de referencia para la construcción de dos subsistemas de un MD: subsistema almacenamiento y subsistema de integración.
- ✓ Aplicarle una técnica de minería de datos o realizarle BI a la BD de los EC del producto N Glicolil GM3, que permitan detectar patrones de comportamiento sobre la información almacenada.

REFERENCIAS BIBLIOGRÁFICAS

[1] Díaz Villanueva, Wladimiro (Wladimiro.Diaz@uv.es).Universidad de Valencia. Almacenes de datos (DataWarehouses).

[2] Ralph Kimball. The Data Warehouse Lifecycle Toolkit. New York :s.n., 1998.

[3] Computer Audio Video System Integrator. [En línea]. Disponible en: http://www.cavsi.com/preguntasrespuestas/que-es-data-mart/

[4] González Hidalgo-Gato, G. "Data Mart para la Gestión Contable de la Empresa de Proyectos de Arquitectura e Ingeniería". Tesis de Maestría de Informática Aplicada.

[5] Espinosa, Itziar Angoitia. Data Warehouse para la Gestión de Lista de Espera Sanitaria. Facultad de Informática, Universidad Politécnica de Madrid. Madrid : s.n. pág. 148. http://oa.upm.es/1095/1/PFC_ITZIAR_ANGOITIA_ESPINOSA.pdf.

[6] Vega Torres, Ing. Lilliam y Rojas Díaz, Ing. Luis. La inteligencia de negocio. Su implementación mediante la plataforma Pentaho. La Habana: s.n., 2008.

[7]Kimball Ralph: "FactTables and Dimension Tables". [En línea]. Disponible en:

http://www.intelligententerprise.com/030101/602warehouse1_1.jhtml.

[8] TODO BI, "Informe Business Intelligence". Recopilación de los mejores artículos de inteligencia de negocio del 2006.

[9] Universidad de Murcia. [En línea] http://www.um.es/docencia/barzana/IAGP/IAGP2-Metodologias-de-desarrollo.html.

[10] Arley, Ricardo Chinchilla. Mercado de datos: conceptos y metodologías de desarrollo.3, julio-septiembre de 2011, Tecnología en Marcha, Vol. 24, págs. 55-66.

[11] CAVSI. CAVSI Computer Audio Video Systems Integrator. [En línea] [Citado el: 4 de diciembre de 2012.] http://www.cavsi.com/preguntasrespuestas/que-es-procesamiento-analitico-en-liea-olap/.

[12] Hurtado Torres, M. Visitación; Abad Grau, M. Mar; Hornos Barranco, Miguel J. y Montes Soldado, Rosana. Bases de datos y data warehouse: Herramientas estratégicas para la eficacia comercial. Departamento de Lenguajes y Sistemas Informáticos. Facultad de Ciencias Económicas y Empresariales. Universidad de Granada.

[13] Rizo Rizo, MSc. Emma R.; Tápanes Mora, Ing. Mayté; Pedro Febles, Dr. Juan; Estrada Senti, Dra. Vivian y Sánchez Pérez, Dr. Efraín. Importancia de la utilización de un Data Warehouse (DW) en las empresas.

http://www.tec.ac.cr/sitios/Vicerrectoria/vie/editorial_tecnologica/Revista_Tecnologia_Marcha/pdf/tecnologia_marcha_24-3/TM%2024-3%20art%206.pdf.

[14] PostgreSql Cuba. [En línea] http://postgresql.uci.cu/node/63.

[15] Sistema Gestor de Base de Datos. [En Línea]. Disponible en: http://www.ecured.cu/index.php/Sistema_Gestor_de_Base_de_Datos. Fecha de consulta: día 9 de 01 2013.

[16] PostgreSQL. [En línea] http://archives.postgresql.org/pgsql-es-fomento/2009-07/msg00000.php.

[17] Gravitar Información sin Límites. [En línea] http://www.gravitar.biz/index.php/herramientas-bi/pentaho/caracteristicas-pentaho/.

[18] autores, Colectivo de. SlideShare. DataWarehouse. [En línea] [Citado el: 6 de marzo de 2012.] http://www.slideshare.net/g2ticstelesup/data-warehouse-7285737.

[19] MSDN. Tablas de hechos. [En línea] [Citado el: 2 de abril de 2012.] http://msdn.microsoft.com/es-es/library/ms244679%28v=vs.80%29.aspx.

[20] Ralph Kimball, Joe Caserta. The Data Warehouse ETL Toolkit. s.l. : Wiley. pág. 526.

[21] autores, Colectivo de. Auditoría de sistemas. s.l., Manizales : Universidad de Caldas, 2009.

[22] Medina, Doris Mustelier. Técnicas de Extracción, Transformación y Carga de Datos del Sistema de Información Nacional de Seguridad Ciudadana en la República Bolivariana de Venezuela, Marzo 2009.

[23] Gonzáles Hernández, Yanisbel. La Metodología de Desarrollo de Almacenes de Datos. noviembre del 2011.

BIBLIOGRAFÍA

1. **Arley, Ricardo Chinchilla.** *Mercado de datos: conceptos y metodologías de desarrollo.* 3, julio-septiembre de 2011, Tecnología en Marcha, Vol. 24, págs. 55-66.
2. **autores, Colectivo de.** *Auditoría de sistemas.* s.l., Manizales : Universidad de Caldas, 2009.
3. **CAVSI.** CAVSI Computer Audio Video Systems Integrator. [En línea] [Citado el: 4 de diciembre de 2012.] http://www.cavsi.com/preguntasrespuestas/que-es-procesamiento-analitico-en-liea-olap/.
4. **Computer Audio Video System Integrator.** [En línea]. Disponible en:
5. **Díaz Villanueva, Wladimiro** (Wladimiro.Diaz@uv.es).Universidad de Valencia. Almacenes de datos (DataWarehouses).
6. Empresa de Proyectos de Arquitectura e Ingeniería". Tesis de Maestría de
7. **Espinosa, Itziar Angoitia.** *Data Warehouse para la Gestión de Lista de Espera Sanitaria.* Facultad de Informática, Universidad Politécnica de Madrid. Madrid : s.n. pág. 148. http://oa.upm.es/1095/1/PFC_ITZIAR_ANGOITIA_ESPINOSA.pdf.
8. **Gonzáles Hernández, Yanisbel.** *La Metodología de Desarrollo de Almacenes de Datos.* noviembre del 2011.
9. **González Hidalgo-Gato, G.** "Data Mart para la Gestión Contable de la
10. **Gravitar Información sin Límites.** [En línea] http://www.gravitar.biz/index.php/herramientas-bi/pentaho/caracteristicas-pentaho/.
11. http://www.cavsi.com/preguntasrespuestas/que-es-data-mart/
12. http://www.intelligententerprise.com/030101/602warehouse1_1.jhtml.
13. http://www.tec.ac.cr/sitios/Vicerrectoria/vie/editorial_tecnologica/Revista_Tecnologia_Marcha/pdf/tecnologia_marcha_24-3/TM%2024-3%20art%206.pdf.
14. **Hurtado Torres, M. Visitación; Abad Grau, M. Mar; Hornos Barranco, Miguel J. y Montes Soldado, Rosana.** Bases de datos y data warehouse: Herramientas estratégicas para la eficacia comercial. Departamento de Lenguajes y Sistemas Informáticos. Facultad de Ciencias Económicas y Empresariales. Universidad de Granada.
15. Informática Aplicada.
16. **Kimball Ralph**: "FactTables and Dimension Tables". [En línea]. Disponible en: http://www.intelligententerprise.com/030101/602warehouse1_1.jhtml.

17. **Medina, Doris Mustelier.** *Técnicas de Extracción, Transformación y Carga de Datos del Sistema de Información Nacional de Seguridad Ciudadana en la República Bolivariana de Venezuela*, Marzo 2009.
18. **MSDN.** Tablas de hechos. [En línea] [Citado el: 2 de abril de 2012.] http://msdn.microsoft.com/es-es/library/ms244679%28v=vs.80%29.aspx.
19. **PostgreSql Cuba.** [En línea] http://postgresql.uci.cu/node/63.
20. **PostgreSQL.** [En línea] http://archives.postgresql.org/pgsql-es-fomento/2009-07/msg00000.php.
21. **Ralph Kimball, Joe Caserta.** *The Data Warehouse ETL Toolkit.* s.l. : Wiley. pág. 526.
22. **Ralph Kimball.** The Data Warehouse Lifecycle Toolkit. New York :s.n., 1998.
23. **Rizo Rizo, MSc. Emma R.; Tápanes Mora, Ing. Mayté; Pedro Febles, Dr. Juan; Estrada Senti, Dra. Vivian y Sánchez Pérez, Dr. Efraín.** Importancia de la utilización de un Data Warehouse (DW) en las empresas.
24. **Sistema Gestor de Base de Datos.** [En Línea]. Disponible en: http://www.ecured.cu/index.php/Sistema_Gestor_de_Base_de_Datos. Fecha de consulta: día 9 de 01 2013.
25. **TODO BI,** "Informe Business Intelligence". Recopilación de los mejores artículos de inteligencia de negocio del 2006. Universidad de Murcia. [En línea] http://www.um.es/docencia/barzana/IAGP/IAGP2-Metodologias-de-desarrollo.html.
26. **Vega Torres, Ing. Lilliam y Rojas Díaz, Ing. Luis**. La inteligencia de negocio. Su implementación mediante la plataforma Pentaho. La Habana: s.n., 2008.

ANEXOS

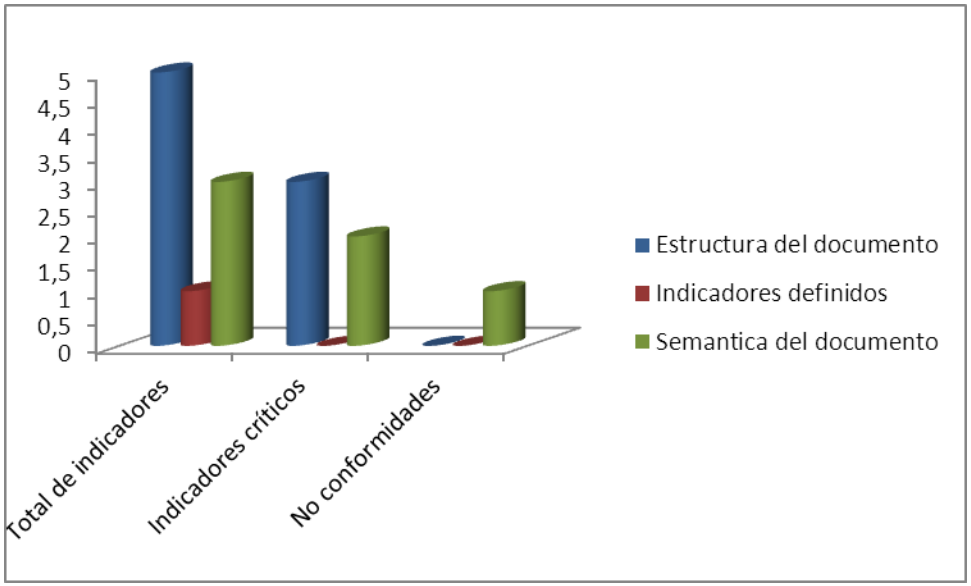
Caso de uso de información	Requisito de información	Tablas implicadas	Variables de entrada	Variables de salida	Consulta SQL realizada	Datos obtenidos	Fuente de datos	Variables de la fuente implicadas	Datos almacenados en la fuente	Resultados de la prueba
Mantener_disponible_informacion_pacientes_eventos_adversos_producto_NGCGM3	Obtener la cantidad que presentaron eventos adversos de los ensayos clínicos del producto NGCGM3 por hospital,edad, tiempo, tipo de eventos adverso, grado de severidad del eventos adverso y causalidad del eventos adverso.	dim_tiempo, dim_hospital, dim_tipo_evento_adverso, dim_grado_severidad_evento_adverso, causalidad_evento_adverso y hech_eventos_adversos.	grado de severidad del eventos adverso	cant_pacientes	SELECT count(distinct hech_eventos_adversos.pk_codigo_paciente) FROM mart_ngcgm3.hech_eventos_adversos, dimensiones.dim_grado_evento_adverso WHERE hech_eventos_adversos.dk_dim_grado_evento_adverso_id = dim_grado_evento_adverso.dk_dim_grado_evento_adverso_id AND dim_grado_evento_adverso.grado_evento_adverso = 'Moderado';	La cantidad de pacientes obtenidos con grado de evento adverso moderado fue 62.	Repositorio/Fuentes/EC058 NGc GM3 Mama FII/05.1 Eventos Adversos.xls Repositorio/Fuentes/EC066 NGc Melanoma SC FI/05.1 Eventos Adversos.xls Repositorio/Fuentes/EC067 NGc Mama SC FI/05.1 Eventos Adversos.xls	**Eventos_adversos058** var3, incl, var6,var7, var11,var12. **Eventos_adversos066** inic, incl,var3,var4,var8, var9. **Eventos_adversos067** INIC,INCL, VAR3, VAR4,VAR8,VAR9.	Los datos existentes en la fuente de datos arrojaron el mismo resultado que la consulta a la base de datos.	Satisfactorio

Anexo 1. *Caso de prueba realizado al CU*

Mantener_disponible_informacion_pacientes_eventos_adversos_producto_NGCGM3.

Estructura del documento					
Peso	**Indicadores a evaluar**	**Eval**	**(NP)**	**Cantidad de elementos afectados**	**Comentarios**
crítico	1. ¿El alcance del proyecto describe correctamente los datos de las dimensiones y hechos del MD?	0			
crítico	2. ¿El objetivo expresa correctamente el propósito del documento?	0			
crítico	3. ¿Existe una adecuada correspondencia entre el origen del los datos y los atributos del mercado?	0			
	4. ¿Se hace un uso adecuado del control del documento?	0			
	5. ¿En la sección de acrónimos se definen todos los acrónimos utilizados en el documento?	0			
Indicadores definidos en el desarrollo					
Peso	**Indicadores a evaluar**	**Eval**	**(NP)**	**Cantidad de elementos afectados**	**Comentarios**
	1. ¿Se utilizó un lenguaje cuyas sentencias son expresables mediante una sintaxis	0			

	bien definida?				
Semántica del documento					
Peso	**Indicadores a evaluar**	**Eval**	**(NP)**	**Cantidad de elementos afectados**	**Comentarios**
crítico	1. ¿Se han identificado errores ortográficos en el documento?	0	1	1	
crítico	2. ¿Se entiende claramente lo que se ha especificado en el documento?	0			
	3. ¿El número de página que aparece en el índice coincide con el contenido que se refleja realmente en dicha página?		NP		



Anexo 2. *Lista de chequeo realizada al artefacto Mapa lógico.*

GLOSARIO DE TÉRMINOS

AD: Almacén de datos.

BD: Bases de Datos.

BI: Inteligencia de Negocio.

CIM: Centro de Inmunología Molecular.

CUS: Casos de Uso del Sistema.

DATEC: Centro de Tecnologías de Gestión de Datos.

EC: Ensayos Clínicos.

ETL: Extracción, Transformación y Carga.

MD: Mercado de datos.

MER: Modelo Entidad-Relación.

MMD: Modelo Multidimensional.

OLAP: Procesamiento analítico en línea.

OMS: Organización Mundial de la Salud.

ONEI: Oficina Nacional de Estadísticas e información.

RI: Requisitos de Información.

RF: Requisitos Funcionales.

RN: Reglas del Negocio.

RNF: Requisitos No Funcionales.

SGBD: Sistema Gestor de Base de Datos.

UCI: Universidad de las Ciencias Informáticas.