# Universidad de las Ciencias Informáticas

## Facultad 6


# Título: Subsistemas de almacenamiento e integración EPOCIM para el almacén de datos de los ensayos clínicos del Centro de Inmunología Molecular

Trabajo de Diploma para optar por el título de

Ingeniero en Ciencias Informáticas

**Autores**

Henrry García Ortega

Yanelis Collado Fajardo

**Tutor**

Ing. Lázaro José Estupiñan Cutiño

**La Habana, junio 2013.**

**"Año 55 de la Revolución"**

*"EL SENTIDO DE LAS COSAS NO ESTÁ EN LAS COSAS MISMAS, SINO EN NUESTRA ACTITUD HACIA ELLAS"*

## 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 año _____.

Henrry García Ortega

_____

Firma del Autor

Yanelis Collado Fajardo

_____

Firma del Autor

Ing. Lázaro José Estupiñan Cutiño

_____

Firma del Tutor

## DATOS DE CONTACTO

**Tutor:** Ing. Lázaro José Estupiñan Cutiño

**Especialidad de graduación:** Ingeniería en Ciencias Informáticas

**Correo Electrónico:** ljestupinan@uci.cu

**DEDICATORIA**

*Dedico este trabajo a mi mama por ser cada día más que una madre y ser lo más importante que tengo en mi vida, a mi papá por guiar mis pasos y tratar de llevarme siempre por el buen camino.*

*Yanelis*

*A mi madre, mi única flor en el universo*

*Henrry*

**AGRADECIMIENTOS**

*Primeramente agradecerle a la Universidad por darme la oportunidad de hacer mis sueños realidad. A mi mamá por siempre brindarme su apoyo incondicional y darme ánimo cuando siempre lo he necesitado. A mi papa porque a pesar de todos los momentos malos por los que hemos pasado siempre ha estado atento a todo lo que correspondía con mi carrera y porque este también es su sueño. A mi hermana Yelena por apoyarme, ayudarme y siempre estar a mi lado. A mi hermana Narumis porque a pesar de los problemas que han ocurrido no nos hemos podido separar. A mi prima Irene por ser una hermana más y siempre apoyarme. A toda mi familia en especial a mi abuela Fe, mi tía Virgen, mis primos, mis tíos, a Thaimi por ser parte también de mi familia. A mi amiga Albis por apoyarme y ayudarme con mis estudios .A mi dúo de tesis por el tiempo dedicado y sacrificado para que esta tesis salga de la mejor manera. A varias personas que no nombraré pero que siempre los tengo en mi mente y mi corazón. A todas las personas de las cuales estoy muy agradecida por la influencia positiva que tuvieron sobre mí para que este gran sueño se lograra.*

*Yanelis*

**RESUMEN**

La presente investigación surge como parte de la colaboración que existe entre la Universidad de las Ciencias Informáticas (UCI) y el Centro de Inmunología Molecular (CIM). El CIM tiene la necesidad de gestionar, almacenar y analizar toda la información que se recoge en los ensayos clínicos cuando son aplicados los medicamentos allí realizados. En el presente Trabajo de Diploma, se pretende crear un repositorio donde la información, sobre el producto Epocim, se encuentre centralizada, estandarizada y accesible para su consulta. Se realiza un estudio de las metodologías para el desarrollo de almacenes de datos y se definen las herramientas que se utilizan durante la construcción de la solución. Se abarca el análisis, diseño y todo el conjunto de transformaciones al que son sometidos los datos para ser estandarizados. Como resultado de estos procesos se obtiene los subsistemas de almacenamiento e integración, que contienen la información del producto Epocim, permitiendo mantener disponible la información histórica para su análisis.

**Palabras claves:** Centro de Inmunología Molecular, subsistemas de almacenamiento e integración, Epocim, almacenes de datos, estandarizados.

## ÍNDICE

## ÍNDICE FIGURAS

## ÍNDICE TABLAS

# INTRODUCCIÓN

El mundo en los últimos años ha experimentado un vertiginoso crecimiento de la tecnología, impulsado por el desarrollo de las Tecnologías de la Información y las Comunicaciones (TICs). Las TICs se encuentran en constante evolución, lo que le ha permitido convertirse en un eslabón importante de la sociedad fortaleciendo el avance tecnológico que estaban experimentando algunas ramas. Es por ello que su uso ha dado paso a la obtención de mejores resultados y desarrollo de la Salud Pública, aquí se encuentran los avances médicos tanto logísticos como tecnológicos, la administración de paciente, la gestión médica, bibliográfica y de pruebas.

Cuba ha incorporado el uso de estas tecnologías para así incrementar su desarrollo económico. El uso de las TICs cobra gran importancia en el uso social y colectivo, su aplicación se potencia en la ciencia, la cultura, la economía, servicios públicos, educación y la salud pública, siendo esta última rama de vital importancia ya que la gestión médica ha sido uno de los objetivos priorizados en el país.

Cuba ha obtenido múltiples logros en el sector de la salud, se han creado varios centros biotecnológicos y de investigación con el objetivo de mejorar la calidad de vida de los pacientes. El Centro de Inmunología Molecular (CIM) creado el 5 de diciembre de 1994 tiene como misión obtener y producir nuevos biofármacos destinados al tratamiento de enfermedades crónicas no transmisibles e introducirlos en la salud pública cubana (1). En dicho centro, para llevar el control de estos estudios, los especialistas diseñan los Cuadernos de Recogida de Datos (CRD), donde se reúne toda la información relacionada con el paciente durante su tratamiento. Una vez culminado dicho estudio se envían los cuadernos para el CIM, donde se realiza el proceso de digitalización de la información almacenada en los cuadernos, mediante el sistema EpiData.

El EpiData cuando genera los reportes, crea varios ficheros en formato de hojas de cálculo Excel, por cada ensayo clínico con gran volumen de información, debido al alto número de personas que se involucran. Vale destacar que estos reportes no son uniformes, la herramienta ni sus resultados son capaces de interactuar con Sistemas Gestores de Bases de Datos (SGBD). Concluido este proceso, debido a la gran cantidad de información, se dificulta el análisis estadístico de los distintos ensayos clínicos en el CIM y se convierte en un problema a la hora de satisfacer la demanda de información

biomédica para la toma de decisiones en la práctica clínica. Esto trae como consecuencia que una vez terminado el procesamiento de la información y la generación de los reportes, la toma de decisiones demora más tiempo del que debería y afectando a los pacientes. Teniendo en cuenta la situación anterior y la importancia que tiene obtener mejores resultados de los ensayos clínicos en el CIM, así como el avance de la medicina en Cuba, se identifica el siguiente **problema a resolver:** ¿Cómo lograr la estandarización de los datos del producto Epocim para su almacenamiento de forma homogénea?

Se define como **objeto de estudio,** los almacenes de datos enmarcados en el **campo de acción** subsistemas de almacenamiento e integración Epocim para los ensayos clínicos que se gestionan en el Centro de Inmunología Molecular.

El **objetivo general** de este trabajo es desarrollar los subsistemas de almacenamiento e integración Epocim para el almacén de datos de los ensayos clínicos del Centro de Inmunología Molecular que contribuya al almacenamiento homogéneo de la información.

En correspondencia con el objetivo general se definen los siguientes **objetivos específicos**:

- Fundamentar la selección de la metodología, herramientas y tecnologías a utilizar en el desarrollo de los almacenes de datos.
- Realizar el análisis y diseño de los subsistemas de almacenamiento e integración Epocim.
- Realizar la implementación y validación de los subsistemas de almacenamiento e integración Epocim.

Para dar cumplimiento a estos objetivos se definieron las siguientes **tareas de la investigación**:

- Caracterización de las metodologías, herramientas y tecnologías a utilizar en el desarrollo de almacenes de datos para profundizar en el nivel de comprensión y dominio sobre las mismas.
- Levantamiento de los requisitos del sistema.
- Descripción de los casos de uso de los subsistemas de almacenamiento e integración del producto Epocim.
- Definición de la arquitectura de los subsistemas de almacenamiento e integración del producto Epocim mediante la identificación de los subsistemas fundamentales que componen la solución.

- Diseño del subsistema de almacenamiento.
- Diseño del subsistema de integración.
- Implementación del subsistema de almacenamiento.
- Implementación del subsistema de integración.
- Aplicación de las listas de chequeo.

El documento está estructurado de la siguiente forma: introducción, tres capítulos, conclusiones, recomendaciones, referencias bibliográficas, bibliografía, glosarios de términos y anexos.

**Capítulo 1: Fundamentos teóricos de los almacenes de datos.**

En este capítulo se abordan los principales conceptos relacionados con el producto Epocim, los mercados de datos y almacén de datos, sus principales características y las ventajas que proporciona su uso. Se fundamenta el uso de la metodología de desarrollo a utilizar y las herramientas empleadas.

**Capítulo 2: Análisis y diseño de los subsistemas de almacenamiento e integración Epocim.**

En este capítulo se realiza el análisis y diseño de los subsistemas de almacenamiento e integración Epocim para los ensayos clínicos que se gestionan en el Centro de Inmunología Molecular. Además, se aplica el procedimiento propuesto para el trabajo con el Centro de Inmunología Molecular y se obtiene un diseño de la solución.

**Capítulo 3: Implementación y Pruebas.**

Se realiza la implementación física del sistema, que costa de la implementación de los subsistemas de almacenamientos e integración Epocim y se exponen las pruebas realizadas a la solución así como sus resultados.

# CAPÍTULO 1: Fundamentos teóricos de los almacenes de datos.

**Introducción**

En este capítulo se realiza un estudio sobre las características del producto Epocim, los ensayos clínicos y su gestión en el CIM. También se abordan los principales conceptos relacionados con los Mercados de Datos (MD) y Almacenes de Datos (AD), sus principales características y las ventajas que proporciona su uso. Se fundamenta el uso de la metodología de desarrollo a utilizar y las herramientas empleadas.

## 1.1. Ensayos clínicos

Un Ensayo Clínico (EC) es un estudio clínico en el que se evalúan la eficacia o seguridad de nuevos fármacos o tratamientos médicos que son aplicados a seres humanos con un protocolo de investigación estrictamente controlado. Permiten determinar si un nuevo tratamiento o medicamento contribuirá a prevenir, detectar o tratar una enfermedad. Cuando se habla de un ensayo clínico, es importante conocer el término localización, lugar del cuerpo donde está concentrada la enfermedad que padece el paciente y sobre el cual se aplican los fármacos en cuestión (2).

### 1.1.1. Epocim

Epocim es una solución para inyección fabricada por el CIM en Cuba. El producto está compuesto por Eritropoyetina humana recombinante alfa, Albúmina humana, Citrato de sodio, Cloruro de sodio, Ácido cítrico, Polisorbato 20 y Agua para inyección. Sus efectos están siendo utilizados para el tratamiento de la anemia renal crónica, el tratamiento de la anemia de pacientes con virus de inmunodeficiencia humana (VIH) en régimen terapéutico con Zidovudina y en pacientes oncológicos con tratamiento de Quimioterapia. Resaltar que no tiene como indicación fundamental resolver los casos de anemias severas que requieren corrección inmediata, el producto sustituye la necesidad del tratamiento con transfusión pero no la transfusión de emergencia en cualquiera que sea el caso (3).

## 1.2. Almacén de Datos

En una empresa la toma de decisiones por la parte administrativa es fundamental e indispensable. Por tanto es de gran importancia digitalizar toda la información, en función de una eficiente manipulación.

Existen diversas revisiones conceptuales acerca de AD; lo que demuestra que los mismos pueden ser de mucha utilidad en el entorno empresarial para tomar mejores decisiones. William H. Inmon, uno de los pioneros en el tema de los almacenes, es el actor de una de las definiciones más conocidas, él plantea: "…un AD consiste en una colección de datos orientada al negocio, integrada, no volátil y variante en el tiempo, para el apoyo a la toma de decisiones administrativas" (4).

Ralph Kimball, otro prestigioso autor en el área de los AD, propone otra definición al catalogarlo como "una copia de las transacciones de datos específicamente estructurada para la consulta y el análisis" (21).

A pesar de existir algunas diferencias entre los conceptos expuestos, se puede apreciar que giran sobre la misma idea y es que los AD, tiene como tarea fundamental, organizar la información recogida de diferentes fuentes, para facilitar la extracción de conocimientos y la posterior toma de decisiones.

### 1.2.1. Características principales

Las principales características que definen un AD son:

**Orientados por temas**: los datos de toda la información se encuentran almacenados por materias o temas (clientes, campañas, productos) según los intereses o los resultados que desee obtener cada empresa. Esta característica se organiza de la perspectiva del usuario final.

**Integrados:** los datos con los que se trabajan provienen de diferentes fuentes y deben pasar por un proceso de integración donde se corrigen, validan y estandarizan, para lograr llevarlos a un formato único que facilite su consulta.

**No volátiles:** únicamente hay dos tipos de operaciones en el almacén de datos: la carga de los datos procedentes de los entornos operacionales (carga inicial y carga periódica) y la consulta de los mismos. La actualización de datos no forma parte de la operativa normal de un almacén de datos (5).

**Variables en el tiempo:** permite realizar análisis en el comportamiento de una variable a lo largo del tiempo pues la información nunca es modificada ni eliminada del sistema. El tiempo debe estar presente en todos los registros contenidos de un AD.

En la siguiente tabla se encuentran algunas de las características que diferencian a los AD con las bases de datos operacionales (Ver tabla1) (6).

**Tabla 1. Diferencia entre almacén de datos y bases de datos operacionales.**

| Bases de Datos Operacional | Almacén de datos |
|---|---|
| Datos Operacionales | Datos del negocio para información |
| Orientado a la aplicación | Orientado al sujeto |
| Actual | Actual + Histórico |
| Detallada | Detallada + Resumida |
| Cambia continuamente | Estable |

### 1.2.2. Ventajas del uso de un almacén de datos

La utilización de un AD trae consigo numerosas ventajas que proveen beneficios en cuanto al proceso de toma de decisiones administrativas. El mundo de la investigación biomédica se encuentra constantemente en diversos cambios debidos al alto nivel de desarrollo en cuanto a la investigación de medicamentos y nuevos fármacos para el tratamiento de muchas enfermedades que afectan a las personas, por esta razón los datos de información han sufrido cambios y el uso de un almacén de datos es muy importante para la toma de decisiones en la práctica clínica.

A continuación se muestran otros argumentos necesarios que especifican de forma más clara lo explicado anteriormente (16):

- Proporciona acceso a datos para análisis complejos, revelación de conocimientos y toma de decisiones.
- Integra y consolida diferentes fuentes de datos en una única plataforma sólida y centralizada.
- Provee la capacidad de analizar y explotar toda la información que posee.
- Aumenta la competitividad en el mercado.
- Mejora la entrega de información, o sea, información completa, correcta, consistente, oportuna y accesible.

- Aprovecha el enorme valor potencial de los recursos de información y los transforma en valor verdadero.
- Permite al usuario adquirir mayor confianza acerca de sus propias decisiones y de las del resto y lograr así, un mayor entendimiento de los impactos ocasionados.

Los sistemas de almacenes de datos tienen también entre sus características (5):

- Integración de bases de datos heterogéneas (relacionales, documentales, geográficas, archivos, entre otros).
- Ejecución de consultas complejas no predefinidas visualizando el resultado en forma gráfica y en diferentes niveles de agrupamiento y totalización de datos.
- Agrupar y desagrupar datos en forma interactiva.
- Análisis del problema en términos de dimensiones.

### 1.3. Mercado de Datos

Los MD son creados con el propósito de facilitar la construcción y utilización de un almacén de datos, constituyen una tecnología de bases de datos multidimensionales orientadas a una materia específica, que ha tomado gran auge debido al crecimiento de datos históricos almacenados en grandes organizaciones.

Básicamente, los mercados son una versión reducida de los almacenes, que se especializan en el almacenamiento de los datos relativos a un área específica del negocio, a diferencia de un AD que gestiona la información a nivel de empresa. A continuación se encuentran características que definen aun más a los mercados (22):

- Se centran en los requisitos de los usuarios asociados a un departamento o área de negocio específico.
- Presentan un mayor nivel de detalle.
- Reducen la demanda del depósito de datos.
- Son más sencillo a la hora de utilizarlos y comprender sus datos, debido a que la cantidad de información que contiene es mucho menor que en los almacenes de datos (7).

Según las características y definiciones expuestas se puede concluir que el mercado de datos está definido por la forma en que el usuario necesita ver la información y como desea que se le presente, guarda la información a un nivel más específico, permite el procesamiento independiente del resto de los departamentos de la organización y un costo de almacenamiento inferior.

### 1.4. Modelo multidimensional

El modelo multidimensional es la base de los almacenes de datos. Representa la estructura de sus tablas y relaciones mediante un moldeo multidimensional, o sea, se almacena la información como en el Diagrama Entidad Relación (DER). El modelo multidimensional es una técnica de diseño lógico que busca facilitar una recuperación adecuada de los datos puesto que la supervivencia de cada organización depende de la correcta gestión, seguridad y confidencialidad de la información. Los datos son almacenados como hechos y dimensiones en un modelo de datos relacional.

El hecho son los datos que brindan información cuantitativa sobre las características del negocio que se quieren analizar; el objeto de análisis para la toma de decisiones. Presenta valores cuantitativos que almacenan las métricas del negocio y se denominan medidas (15).

Se conoce como dimensión a la característica de un hecho que permite su análisis posterior en el proceso de toma de decisiones y brinda una perspectiva adicional a un hecho dado. Son agrupaciones lógicas de atributos con un significado común, atómico y por lo general son estables (15).

Los atributos por su parte, se utilizan en las tablas de dimensiones, para búsquedas, filtrado o clasificación de los hechos (5).

#### 1.4.1. Esquemas del modelo multidimensional

El modelo dimensional brinda toda la información de manera sencilla y estándar para administradores de base de datos y los usuarios finales (23). Existen tres tipos de esquemas multidimensionales que están determinados por la complejidad del sistema como se muestra a continuación en la Figura 1:

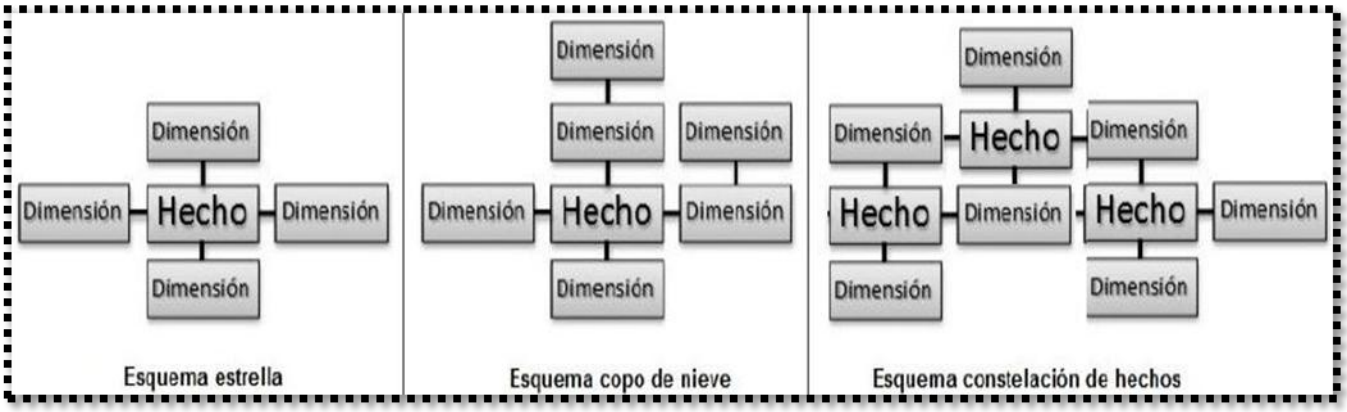


**Figura 1. Esquemas del modelo multidimensional.**

**Esquema en estrella:** presenta el modelo básico de sobre el cual las aplicaciones pueden extenderse hacia los otros dos modelos. A nivel de diseño, consiste en una tabla de hechos en el centro para el hecho objeto de análisis y una o varias tablas de dimensión por cada dimensión de análisis que participa de la descripción de ese hecho.

**Esquema copo de nieve:** es un refinamiento del esquema en estrella en el que las tablas de dimensión se normalizan en múltiples tablas.

**Constelación de hechos:** son varios esquemas en estrella o copo de nieve que comparten dimensiones.

Para poder diseñar las tablas de dimensiones se debe tener en cuenta que en los tres tipos de esquemas todas las perspectivas definidas en el modelo conceptual constituirá una tabla de dimensión. Por esto se deberá tomar cada perspectiva con sus campos relacionados.

Es de suma importancia agregar que para los tres tipos de esquemas, se realizarán las uniones correspondientes entre sus tablas de dimensiones y sus tablas de hechos.

A los grupos de atributos que siguen un orden preestablecido se definen **jerarquías.** Implican una organización de niveles dentro de una dimensión, o sea, cada nivel representa el total agregado de los datos del nivel inferior. Una dimensión típica soporta una o más jerarquías naturales. Una jerarquía puede, pero no exige contener todos los valores existentes en la dimensión (5).

### 1.5. Metodologías para diseñar mercados de datos

La metodología de desarrollo provee una guía para facilitar el trabajo del equipo de trabajo en el proceso de construcción del producto y definir los objetivos del negocio. Esta permite estructurar, planear y controlar todo el proceso de desarrollo de todo tipo de sistemas informáticos.

Para realizar el diseño de la solución es preciso analizar las opciones que se tienen en cuanto a las principales metodologías de desarrollo existentes, Hefesto propuesta por Dario Bernabeu, la de Ralph Kimball y además Bill Inmon:

**Hefesto:** es una metodología ágil en desarrollo que propone cómo guiar la construcción de los AD, analizar los requerimientos de la empresa, identificando las carencias de información que se tienen, y los indicadores y perspectivas de su negocio. Basada en los requerimientos de los usuarios creando una estructura adaptable y rápida a los cambios en el negocio. Utiliza modelos conceptuales y lógicos sencillos de interpretar y analizar; además de incluir al usuario final en cada etapa para que participe en la toma de decisiones.

**Ciclo de vida Kimball:** (principal promotor del enfoque dimensional para el diseño de almacenes de datos), considera que un mercado de datos es una copia de los datos transaccionales específicamente estructurada para la consulta y el análisis. En su contenido define cuatro fases en el ciclo de vida iniciando con la selección del proceso de negocio, definir la granularidad de la información, elegir las dimensiones del análisis y finalizar con la identificación de las métricas (hechos).

**Bill Inmon:** la metodología puede tener una implementación mucho más tardada, y es recomendada cuando se hace demasiado difícil representar el modelo a través de dimensiones y la complejidad de la solución es demasiado grande. Si es necesario utilizar esta metodología, se recomienda hacerlo en iteraciones ya avanzadas y siempre empezar con Kimball pues ambas metodologías pueden implementarse en un mismo mercado de datos (8).

**Propuesta de Metodología para el Desarrollo de Almacenes de Datos**

Para desarrollar la solución se selecciona la Propuesta de Metodología para el Desarrollo de Almacenes de Datos definida por el Centro de Tecnologías de Gestión de Datos (DATEC), la cual toma como base la

metodología propuesta por Ralph Kimball, que abarca las cuatro fases de construcción de un almacén de datos, análisis, diseño, integración de datos e inteligencia de negocio, y como complemento, por las características del trabajo en la UCI, se incluye el empleo de casos de uso para guiar el proceso de desarrollo y una etapa de prueba para comprobar, la calidad de los productos desarrollados (3). A continuación la Figura 2 recrea el ciclo de vida de dicha metodología:

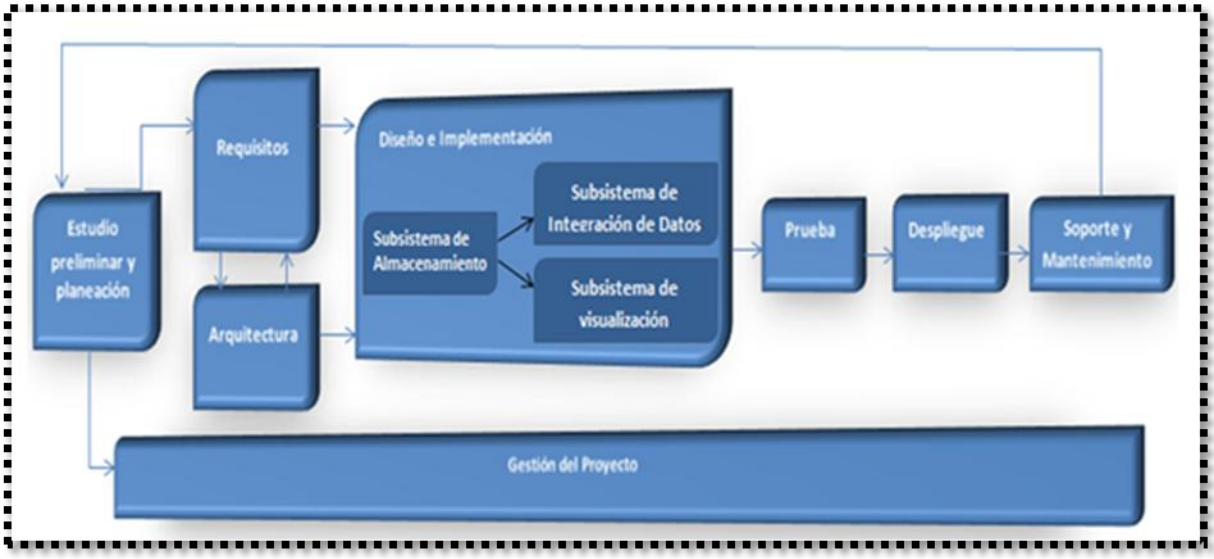


**Figura 2. Ciclo de vida de la Propuesta de Metodología para el Desarrollo de Almacenes de Datos.**

En la solución se llevaron a cabo cinco de las ocho fases que presenta la metodología, las cuales son: estudio preliminar y planeación, requisitos, arquitectura, diseño e implementación de los subsistemas de almacenamiento e integración y por último la etapa de prueba.

### 1.6. Integración de datos

El principal objetivo de esta etapa es integrar los datos que se encuentran almacenados en fuentes heterogéneas, pero que en la mayoría de los casos no están estandarizados, haciendo prácticamente imposible, un correcto análisis de la información, y la toma de decisiones.

Para llevar a cabo la integración se realiza la Extracción, Transformación y Carga (ETL, por sus siglas en inglés), que consiste en la recuperación de los datos de diversas fuentes ajenas al AD, con el fin de limpiarlos e insertarlos en el almacén.

**Extracción**: se obtienen y analizan los datos de fuentes externas, para los subsistemas de almacenamiento e integración Epocim los datos se encuentran en documentos con formato .DBF y .XLS
**Transformación**: una vez analizados, se transforman los datos extraídos, a un formato homogéneo para el almacén de datos. Este paso se lleva a cabo mediante las reglas de transformación, algún ejemplo de estas reglas serian, los tratamiento de valores nulos, o la combinación de datos de diversas fuentes, entre otras.
**Carga:** se insertan los datos preparados en el AD, destacando, que nunca se modifican, o se actualizan los datos.

### 1.7. Herramientas para el desarrollo de los subsistemas de almacenamiento e integración

Durante el desarrollo de los subsistemas de almacenamiento e integración se utilizan varias herramientas que facilitan el trabajo en las distintas fases que abarca el proceso. A continuación se exponen las herramientas definidas por el departamento de almacenes de datos de DATEC, para la implementación de la solución.

#### 1.7.1. Herramienta y lenguaje de modelado

En la actualidad se han desarrollado diversas herramientas con el propósito de un acercamiento a la automatización del diseño, construcción, implementación y mantenimiento de los almacenes de datos, debido a que con el transcurso de los años el avance de la tecnología aumenta para dar seguimiento a la obtención de resultados positivos. Existen varias herramientas creadas para el desarrollo de la Ingeniería de Software, con el fin de desarrollar programas. En el presente trabajo de investigación se utiliza Visual Paradigm para UML en su versión 8.0, herramienta CASE (Ingeniería de software asistida por computadoras), multiplataforma, factible a la hora de dibujar diagramas de clases, y generar script para diferentes SGBD. Permite una integración con sistemas de control de versiones que almacenan centralmente los artefactos y realizan un seguimiento de los cambios realizados sobre un proyecto. Los desarrolladores lo utilizan para facilitar el modelado simultáneo, almacenar los archivos de proyectos y hacer un seguimiento de los cambios, además de posibilitar una rápida construcción de las aplicaciones con alta calidad (9).

Esta herramienta utiliza el **Lenguaje Unificado de Modelado** (UML, por sus siglas en inglés) que en la actualidad es el lenguaje de sistemas de software más conocido y utilizado. UML ofrece un estándar para describir de la manera más sencilla y entendible para cualquier desarrollador los aspectos conceptuales tales como procesos de negocios y funciones del sistema. Un detalle de gran importancia resulta ser que UML es un lenguaje para especificar y no para describir métodos o procesos, por tanto su definición se expone de la siguiente manera: (…) lenguaje gráfico para visualizar, especificar, construir y documentar un sistema de software (9).

### 1.7.2 Herramientas para la gestión y administración de la base de datos

**PostgreSQL 9.1**: es el Sistema Gestor de Base de Datos (SGBD) de código abierto más potente del mercado, es relacional, orientado a objetos y multiplataforma. Su código es estable y sólido, pues utiliza multiprocesos, lo que permite que si un proceso falla, no afecte al sistema, minimizando los errores.

Cuenta con una documentación bien organizada y amplia, permite la creación de funciones personalizadas, el manejo, la configuración de disparadores y es capaz de ajustarse a la cantidad de memoria que ofrece el sistema de forma óptima (10).

**PgAdmin III 1.14**: es un sistema de administración de bases de datos multiplataforma, diseñado para versiones iguales o superiores a PostgreSQL 7.3, la interfaz gráfica es compatible con las características de PostgreSQL y facilita la administración. Cuenta con una amplia documentación, y es una herramienta de consulta de gran alcance con resaltado de sintaxis. Muestran la dependencia entre objetos con su definición SQL (11).

### 1.7.3 Herramientas para el proceso de integración de datos

**Pentaho Data Integration 4.2.1:** es una herramienta libre, sin costes de licencia, con una interfaz de usuario sencilla que permite realizar el proceso de ETL. Otra de las facilidades que brinda su uso es la potente de extracción, transformación y carga que permite la integración de ambientes y datos para soportar las áreas de negocios. Proporciona la solución ideal para cualquier tipo de integración de datos, análisis de negocio o proyectos con grandes capacidades de datos (12). Es una herramienta

multiplataforma lo que permite ejecutarla en cualquier sistema operativo. Cuenta con una gran comunidad de usuarios.

**Datacleaner 1.5.3:** es una aplicación de código abierto para el perfilado, análisis, transformación y limpieza de datos. Estas actividades ayudan a administrar y controlar la calidad de los datos pues genera reportes y gráficos sofisticados. (13)

### 1.8. Conclusiones parciales

En este capítulo se abarca una panorámica general del proceso de desarrollo de los almacenes de datos, y más específicamente los mercados de datos, evidenciando la relación existente con los subsistemas de almacenamiento e integración, así como el estudio de la metodología y herramientas a utilizar.

Para la implementación de la solución se seleccionaron:

- Como metodología para el desarrollo del subsistema de almacenamiento e integración Epocim la propuesta de Metodología para el Desarrollo de Almacenes de Datos en DATEC, siendo una adaptación de la metodología propuesta por Kimball, que se ajusta con las tendencias y normas de la UCI.
- Para lograr la integración de los datos de EC que se gestiona en el CIM se seleccionó las herramientas propuestas por DATEC, Pentaho Data Integration 4.2.1, el Datacleaner 1.5.3 para el perfilado de los datos.
- Como herramienta de modelado se utiliza el Visual Paradigm en su versión (8.0).
- Para la gestión y administración de la base de datos, PostgreSQL en su versión 9.1 y PgAdminIII 1.14.0.

# CAPÍTULO 2: Análisis y Diseño de los subsistemas de almacenamiento e integración Epocim.

**Introducción**

En este capítulo se aborda la etapa de análisis y diseño de los subsistemas de almacenamiento e integración Epocim para los EC que se gestionan en el CIM. Se realiza un análisis profundo del negocio, y se obtiene un diseño de la solución que responde a las necesidades manifestadas por el cliente.

## 2.1 Análisis del negocio

Para el desarrollo de un sistema, el levantamiento de los requisitos es una actividad necesaria, ya que muestran las necesidades de los clientes, y a través de las mismas se podrán definir las funcionalidades que tendrá la aplicación.

Para identificar las necesidades de la organización, en este caso el CIM, se entrevistaron a los especialistas del producto Epocim y la estadística del centro. Acordándose almacenar toda la información sobre las variables requeridas de los siguientes ensayos:

- EpoAdultos.
- EpoNiños.

**Variables**

**Enfermedad de base:** recoge la enfermedad de base que presenta el paciente.

**Esquema de tratamiento con quimio y/o radioterapia:** recoge el esquema de tratamiento con quimio y/o radioterapia que recibe el paciente para su enfermedad de base.

**Ciclo de tratamiento con quimio y/o radioterapia:** recoge el ciclo de tratamiento con quimio y/o radioterapia que recibe el paciente para su enfermedad de base.

**Número de transfusiones recibidas en los 2 meses previos:** recoge la cantidad de transfusiones administradas al paciente en los 2 meses previos al estudio.

**Dosis de EPOCIM indicada:** recoge la dosis de EPOCIM indicada en unidades totales de acuerdo al peso corporal del paciente.

**Persistencia de dosis inicial:** registra si hubo cambios en la dosis indicada para el paciente durante el tratamiento.

**Tratamiento anterior con Epocim en esta etapa de quimio y/o radioterapia:** recoge si el paciente recibió tratamiento anterior con EPOCIM en esta misma etapa de quimio y/o radioterapia y en caso afirmativo el número de tratamientos.

**Estado del paciente según la Organización Mundial de la Salud (OMS):** recoge el estado general del paciente según la OMS.

**Paciente ambulatorio u hospitalizado:** recoge si el paciente recibe tratamiento en forma ambulatoria u hospitalizada.

**Tratamiento concomitante para la anemia:** recoge el tratamiento concomitante antianémico que ha recibido el paciente durante el estudio.

**Número de dosis de EPOCIM recibidas:** recoge el número de dosis de EPOCIM que ha recibido el paciente.

**Tratamiento concomitante con Leukocim:** se evaluará mediante la respuesta dicotómica sí/no.

**Transfusiones sanguíneas:** se evaluará mediante la respuesta dicotómica sí/no.

**Edad:** recoge la edad del paciente en años.

**Peso:** recoge el peso del paciente en Kg.

**Sexo:** recoge el sexo del paciente: femenino o masculino.

**Color de la piel:** recoge el color de la piel del paciente: blanca, negra, mestiza o amarilla.

**Ocurrencia de algún evento adverso (EA) en el sujeto:** se registra el evento adverso que sufre el paciente.

**Exámenes:** se registrarán los rangos de bajo, normal y alto, en correspondencia con los valores establecidos para cada uno de ellos.

**Interrupción de tratamiento:** recoge si al paciente se le interrumpió el tratamiento y la causa.

**Fallecimiento:** registra si el paciente falleció durante el tratamiento, y su relación con el tratamiento.

**Hospital:** registra el hospital donde se atiende el paciente.

**Fecha de la evaluación:** registra la fecha en que le realizan la evaluación al paciente.

### 2.2 Especificación de los requisitos

En el levantamiento de los requisitos es donde se identifican todos los requerimientos de la solución. El siguiente paso se materializa con la entrevista al cliente para determinar los requisitos de información. Seguidamente se efectúa un levantamiento detallado de las fuentes de datos para validar la disponibilidad de la información, se definen los requisitos funcionales, los no funcionales y se hace el análisis de los requisitos que dan paso al diseño e implementación de la solución. Todos los requisitos se pueden encontrar de forma íntegra dentro del Expediente de Proyecto de los Subsistemas de almacenamiento e integración Epocim, en el artefacto “Especificación de Requisitos de Software.doc”.

#### 2.2.1. Requisitos de Información (RI)

Son los que describen la información y los datos que son almacenados en el sistema. Se definen a partir de las necesidades del negocio para poder satisfacer a los clientes, pues permiten el análisis del producto según los objetivos y metas de la organización.

El presente trabajo cuenta con 26 requisitos de información:

**RI1:** Obtener la cantidad de pacientes niños que se incluyeron en el ensayo atendiendo a los exámenes de laboratorio.

**RI2:** Obtener la cantidad de pacientes niños que se incluyeron en el ensayo atendiendo a los datos demográficos.

**RI3:** Obtener la cantidad de pacientes niños que se incluyeron en el ensayo atendiendo a la enfermedad base presentada y esquema de tratamiento.

**RI4:** Obtener la cantidad de pacientes niños que se incluyeron en el ensayo atendiendo a la fecha de las evaluaciones y el hospital donde se atendieron.

**RI5:** Obtener la cantidad de pacientes niños que se incluyeron en el ensayo atendiendo a su estado según la OMS y la dosis de Epocim indicada.

**RI6:** Obtener la cantidad de pacientes niños que se incluyeron en el ensayo atendiendo a su situación y el número de transfusiones recibidas en los 2 meses previos.

**RI7:** Obtener la cantidad de pacientes niños que se incluyeron en el ensayo atendiendo al tratamiento concomitante y con Leukocim recibidos.

**RI8:** Obtener la cantidad de pacientes niños que se incluyeron en el ensayo atendiendo al ciclo de tratamiento, y el número de quimioterapia o radioterapia recibidas.

**RI9:** Obtener la cantidad de pacientes niños que se incluyeron en el ensayo atendiendo al número de dosis Epocim y de transfusiones recibidas.

**RI10:** Obtener la cantidad de pacientes niños que fallecieron durante el tratamiento y su causa

**RI11:** Obtener la cantidad de pacientes niños que interrumpieron el tratamiento y su causa.

**RI12:** Obtener la cantidad de pacientes niños que sufrieron eventos adversos atendiendo a su tipo.

**RI13:** Obtener la cantidad de eventos adversos en pacientes niños atendiendo a su tipo.

**RI14:** Obtener la cantidad de eventos adversos en pacientes adultos atendiendo a su tipo.

**RI15:** Obtener la cantidad de pacientes adultos que sufrieron eventos adversos atendiendo a su tipo.

**RI16:** Obtener la cantidad de pacientes adultos que interrumpieron el tratamiento y su causa.

**RI17:** Obtener la cantidad de pacientes adultos que fallecieron durante el tratamiento y su causa.

**RI18:** Obtener la cantidad de pacientes adultos que se incluyeron en el ensayo atendiendo al número de dosis Epocim y de transfusiones recibidas.

**RI19:** Obtener la cantidad de pacientes adultos que se incluyeron en el ensayo atendiendo al ciclo de tratamiento, y el número de quimioterapia o radioterapia recibidas.

**RI20:** Obtener la cantidad de pacientes adultos que se incluyeron en el ensayo atendiendo a al tratamiento concomitante y con Leukocim recibidos.

**RI21:** Obtener la cantidad de pacientes adultos que se incluyeron en el ensayo atendiendo a su situación y el número de transfusiones recibidas en los 2 meses previos.

**RI22:** Obtener la cantidad de pacientes adultos que se incluyeron en el ensayo atendiendo a su estado según la OMS y la dosis de Epocim indicada.

**RI23:** Obtener la cantidad de pacientes adultos que se incluyeron en el ensayo atendiendo a la fecha de las evaluaciones y el hospital donde se atendieron.

**RI24:** Obtener la cantidad de pacientes adultos que se incluyeron en el ensayo atendiendo a la enfermedad base presentada y esquema de tratamiento.

**RI25:** Obtener la cantidad de pacientes adultos que se incluyeron en el ensayo atendiendo a los datos demográficos.

**RI26:** Obtener la cantidad de pacientes adultos que se incluyeron en el ensayo atendiendo a los exámenes de laboratorio.

### 2.2.2 Requisitos Funcionales (RF)

Los requisitos funcionales expresan las condiciones o capacidades que el sistema en cuestión debe cumplir. En esta investigación solo se incluye el análisis de los subsistemas de almacenamiento y de integración de datos por lo que se especificarán las funcionalidades respecto a estos según las características del negocio.

Atendiendo a que los ensayos clínicos del producto Epocim son cerrados, no se especifica la persistencia de la información ni las vistas integradas, por tanto el subsistema de almacenamiento no tiene requisitos asociados. Los requisitos referentes al subsistema de integración de datos se describen a continuación:

**RF1:** Realizar la extracción de los datos.

**RF2:** Realizar la transformación y carga de los datos.

### 2.2.3. Requisitos No Funcionales (RNF)

Los requisitos no funcionales son propiedades o cualidades que debe tener la solución. Se agrupan por categorías, y estas vienen dadas en dependencia de las características del negocio. Se definen 4 RNF:

**RNF1:** Garantizar la disponibilidad del sistema en el tiempo requerido.

**RNF2:** Garantizar la persistencia de la información.

**RNF3:** Lograr la homogeneidad de la estructura de los elementos definidos en el almacén.

**RNF4:** Proporcionar el software necesario para el correcto funcionamiento del sistema, en las estaciones de trabajo.

### 2.3 Reglas del Negocio (RN)

Las reglas del negocio son identificadas a partir del análisis del negocio y de los resultados del perfilado de datos. Para cumplir con los aspectos del negocio se describen políticas que deben cumplirse y condiciones que deben satisfacer al cliente (14). Algunas de estas reglas no suelen definirse en etapas tan tempranas del desarrollo por lo que pueden ser actualizadas a medida que van surgiendo. La metodología llevada a cabo clasifica las de reglas del negocio e categorías, que se pueden encontrar en su totalidad en el Expediente de Proyecto de los Subsistemas de almacenamiento e integración Epocim, en el artefacto “Reglas del Negocio.doc”. Estas se presentan a continuación:

**Reglas de variables**

**RN1.** Los identificadores de las dimensiones no pueden tomar valores nulos, ni repetidos.

**Reglas de almacenamiento**

**RN2.** La enfermedad base será de tipo cadena con una longitud máxima de 60 caracteres.

**RN3.** Las fechas contenidas en fecha de la evaluación serán de tipo (dd/mm/aaaa).

**Reglas de transformación**

**RN4.** Los valores del sexo definidos en las variables informacionales, serán 1 y 2 para Femenino y Masculino respectivamente.

**RN5.** Las fechas con formato aaaammdd, serán convertidas al formato dd/mm/aaaa

**RN6.** Los valores de la raza serán: 1, 2, 3, 4 para Blanca, Negra, Mestiza y Amarilla respectivamente

**RN7.** En el modelo evaluación de los ensayos clínicos EpoAdultos y EpoNiños en la variable tratamiento estandarizar el valor “FUMARATO FERROSO”, que presenta inconsistencias en su redacción.

**RN8.** En el modelo evaluación de los ensayos clínicos EpoAdultos y EpoNiños en la variable motivo estandarizar el valor “SUPLEMENTO”, que presenta inconsistencias en su redacción.

**RN9.** En el modelo evaluación del ensayo clínico EpoAdultos la columna tratamiento concomitante almacena los valores 1 y 2. En los casos en que dicho campo contiene el valor ¨2¨, los campos tratamiento y motivo almacenarán como valor la cadena “NO TIENE”.

**RN10.** Luego de ser aplicada la regla de transformación número 9, en la variable motivo almacenará la cadena “SEGÚN PROTOCOLO” en los valores nulos.

**RN11.** En el modelo evtto 4 de los ensayos clínicos EpoAdultos y EpoNiños en la variable transfusiones, almacenar el valor “2” en los campos nulos.

**RN12.** En el modelo evtto 4 de los ensayos clínicos EpoAdultos y EpoNiños en la variable número de transfusiones, almacenar el valor “0” en los campos con valor nulos.

**RN13.** En el modelo evtto 4 de los ensayos clínicos EpoAdultos y EpoNiños en las variables 14, 18, 22, 26 y 30, almacenar el valor “2” en los campos con valores nulos.

**RN14.** En el modelo evtto 8 de los ensayos clínicos EpoAdultos y EpoNiños en las variables 18, 22, 26, 30 y 34 almacenar el valor “2” en los campos con valores nulos.

**RN15.** En todos los campos de exámenes cuando exista valores nulos dicho campo tomara el valor “-1”.

**RN16.** En el modelo evtto 4 de los ensayos clínicos EpoAdultos y EpoNiños en la variable interrupción de tratamiento los valores nulos tomarán el valor “2”.

**RN17.** En el modelo evtto 8 de los ensayos clínicos EpoAdultos y EpoNiños en la variable interrupción de tratamiento mientras el valor sea “2”, almacenar en la variable causa la cadena “NO TIENE”.

**RN18.** En el modelo evtto 8 de los ensayos clínicos EpoAdultos y EpoNiños en la variable fallecimiento cuando el valor sea “2”, asignar a causa de fallecimiento “NO PROCEDE”

**RN19.** Luego de ser aplicada la regla de transformación número 18, en la variable causa de fallecimiento si el campo es nulo, almacenar “EPISODIO DE ANEMIA”.

**RN20.** En el modelo inclusión de los ensayos clínicos EpoAdultos y EpoNiños en la variable esquema de tratamiento almacenar en los campos nulos la cadena “NO DISPONIBLE”

**RN21.** En el modelo Evtto4 de los ensayos clínicos EpoAdultos y EpoNiños en la variable fecha de la evaluación, cuando el valor es nulo; buscar en el modelo evini, fecha de evaluación por el id del paciente, y sumarle 30 días.

**RN22.** En todos los campos de exámenes donde el valor sea “-1”, se va almacenar “NO DISPONIBLE”.

**RN23.** El valor de situación del paciente será: 1 Ambulatorio y 2 Ingresado.

**RN24.** Para la variable peso existen los siguientes rangos

- 1 - 25 kg
- 26 – 50 kg
- 51 – 75 kg
- 76 – 100 kg
- 101 – 125 kg
- 126 – 150 kg
- 151 – 175 kg
- 176 – 200 kg

**RN25.** Para la variable edad existen los siguientes rangos

- 0 - 10
- 11 -20
- 21 - 30
- 31 - 40
- 41 - 50
- 51 - 60

- 61 -70
- 71 - 80
- 81 – 90
- 91 - 100

**RN26.** Para la variable conteo de hemoglobina existen los siguientes rangos

- 0 - 116 bajo
- 117 - 179 normal
- Más de 180 alto

**RN27.** Para la variable hematocrito existen los siguientes rangos

- 0 – 36.1 bajo
- 36.2 – 50.3 normal
- Más de 50.4 alto

**RN28.** Para la variable conteo de plaquetas existen los siguientes rangos

- 0 - 150 bajo
- 151 - 400 normal
- Más de 401 alto

**RN29.** Para la variable reticulocitos existen los siguientes rangos

- 0 - 20 baja
- 21 - 70 normal
- Más de 71 alta

**RN30.** Los exámenes de laboratorio que sean nulos se van a almacenar “Sin información” y cuando no se realiza se almacena “No realizado”.

**RN31.** En la dimensión persiste dosis inicial, cuando el atributo dosis_futura toma valor “-1”, significa que la dosis no cambió.

## 2.4 Definición de la arquitectura base de los subsistemas de almacenamiento e integración

Para el desarrollo de la solución se define la siguiente arquitectura, ver **Figura 3**, consta de dos subsistemas y la fuente de información. El subsistema de integración obtiene toda la información de las fuentes de datos Epocim y su trabajo es llevar a cabo los procesos que integran y transforman la información para su almacenamiento. Mientras que el subsistema de almacenamiento obtiene todos los datos ya procesados durante la extracción, transformación y carga, la cual es almacenada, en una base de datos soportada por el gestor PostgreSQL y administrada por los usuarios autorizados en la herramienta PgAdminIII.

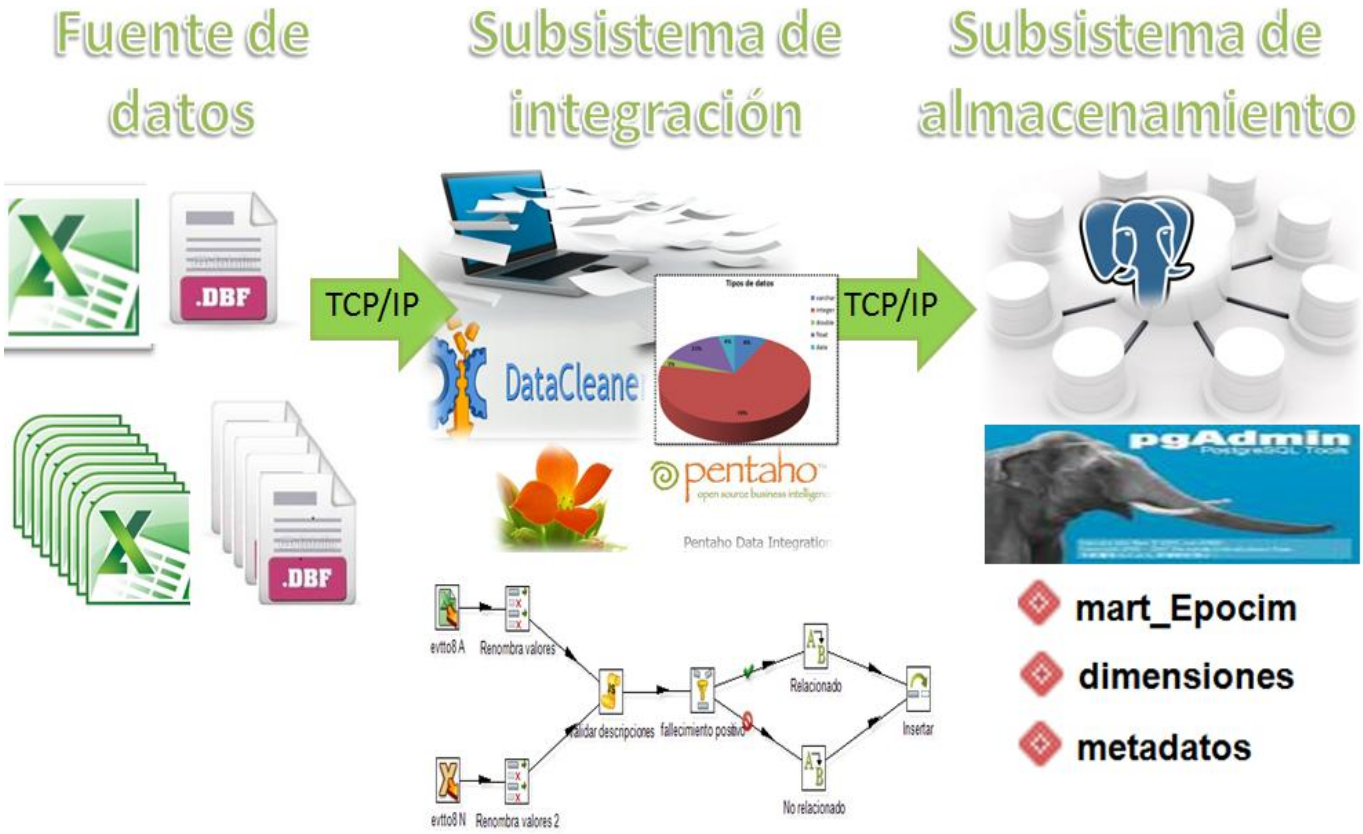


**Figura 3. Arquitectura del sistema.**

## 2.5 Casos de uso del sistema

Los casos de uso son una representación lógica y visual para que el usuario obtenga la información que necesita de la relación existente entre los actores y el sistema. Muestran una secuencia de pasos que son desarrolladas por un sistema en respuesta a un evento que inicia un actor sobre el propio sistema.

### Actores del sistema

**Analista:** es el encargado de analizar y consultar la información de los diferentes indicadores.

**Administrador ETL:** es el encargado de realizar la extracción, transformación y carga de los datos de la fuente.

### Diagrama de casos de uso del sistema

Los diagramas de casos de uso representan como un Cliente (Actor) interactúa con el sistema en desarrollo. Esta secuencia de actividades gráficas hay que automatizarlas.

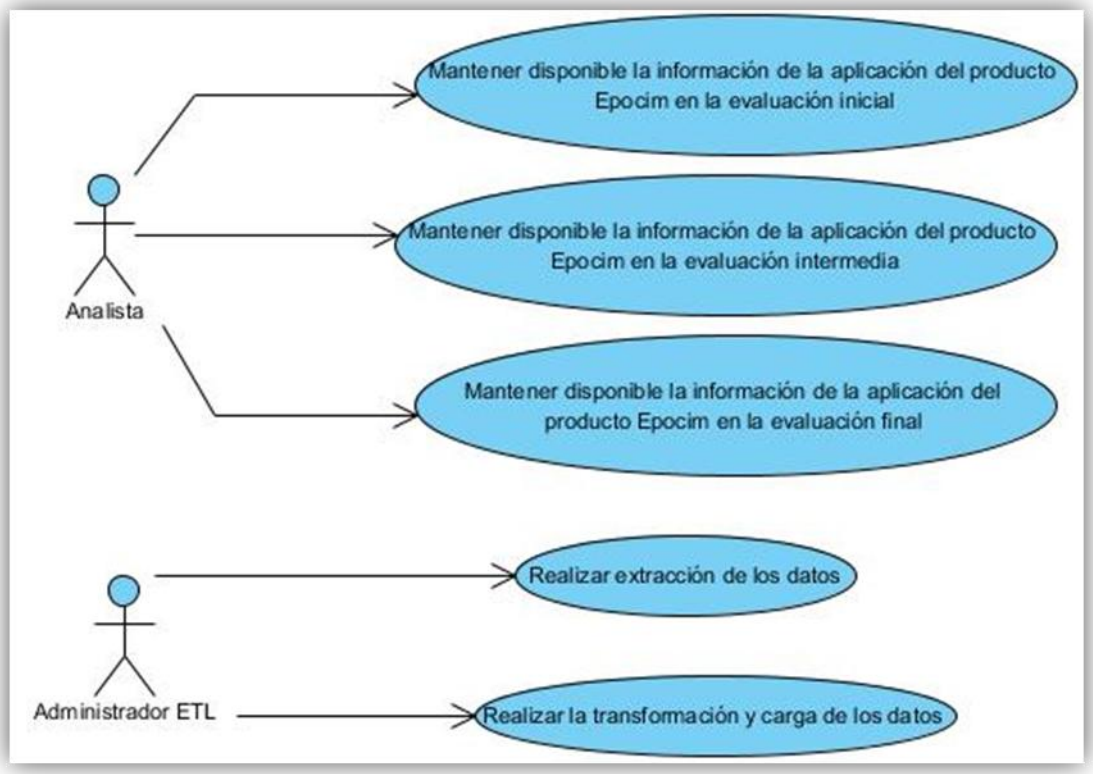


**Figura 4. Diagrama de casos de uso del sistema.**

### 2.4.2 Casos de uso de información

Los casos de uso de información (CUI), responden a los requisitos de información en dependencia de los conceptos de negocio que manejan, principalmente por temas de análisis.

**CUI1:** mantener disponible la información de la aplicación del producto Epocim en la evaluación inicial.
**CUI2:** mantener disponible la información de la aplicación del producto Epocim en la evaluación intermedia.
**CUI3:** mantener disponible la información de la aplicación del producto Epocim en la evaluación final.

### 2.4.3 Casos de uso funcionales

Los Casos de Uso Funcionales (CUF) agrupan los requisitos funcionales definidos para los subsistemas que componen la solución.

**CUF1:** realizar la extracción de los datos.
**CUF2:** realizar la transformación y carga de los datos.

### 2.4.4 Especificación de los casos de uso

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

| | |
|---|---|
| **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 mismos de la fuente. Finaliza cuando los datos se encuentran en la base de datos. |
| **Complejidad:** | Media |
| **Prioridad:** | Media |
| **Precondiciones:** | Disponibilidad de las fuentes. |

| | |
|---|---|
| **Poscondiciones:** | Los datos de la fuente correspondiente han sido extraídos y almacenados en la base de datos. |
| **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 la estructura o archivo 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. |

**Tabla 3. Descripción del Caso de Uso Funcional: Realizar la transformación y carga de los datos.**

| | |
|---|---|
| **Objetivo:** | Realizar la transformación y carga de los datos. |
| **Actores:** | Administrador ETL. |
| **Resumen** | El caso de uso inicia cuando el actor desea realizar la transformación y carga de los datos. Finaliza cuando los datos son insertados satisfactoriamente en la base de datos. |
| **Complejidad:** | Media |

| | |
|---|---|
| **Prioridad:** | Media |
| **Precondiciones:** | Los datos se encontraron correctamente extraídos de la fuente y las estructuras del mercado de datos se encontraron disponibles para su uso. En la base de datos debe existir una estructura para almacenar la información. |
| **Poscondiciones** | Los datos han sido transformados y cargados en el mercado de datos. |

| **Flujo Normal de Eventos** | |
|---|---|
| **Acción del Actor** | **Respuesta del sistema** |
| 1. El administrador de ETL selecciona las estructuras de la fuente que desea transformar. | |
| 2. El administrador de ETL carga los datos seleccionados en memoria. | |
| 3. El administrador de ETL aplica las transformaciones pertinentes y genera datos de auditoría. | |
| 4. El administrador de ETL carga los datos en el mercado de datos. | 5. Ejecuta la consulta. Finaliza el caso de uso. |

| **Flujos Alternos** | |
|---|---|
| **Acción del Actor** | **Respuesta del Sistema** |
| | 3.3 El sistema muestra un mensaje de error y regresa al paso 3. |

**2.6 Diseño del subsistema de almacenamiento de datos.**

Como parte del diseño del subsistema de almacenamiento se identifican las dimensiones y las tablas de hechos con sus medidas asociadas. También debe definir, una política de respaldo y recuperación que garantice la integridad de los datos almacenados.

**Hechos.**

Como se plantea anteriormente el hecho es el objeto a analizar, en este caso sobre los temas de análisis identificados.

- **Evaluación inicial** (hech_evaluacion_inicial): Muestra los resultados de los pacientes de la primera evaluación. (Figura 5)

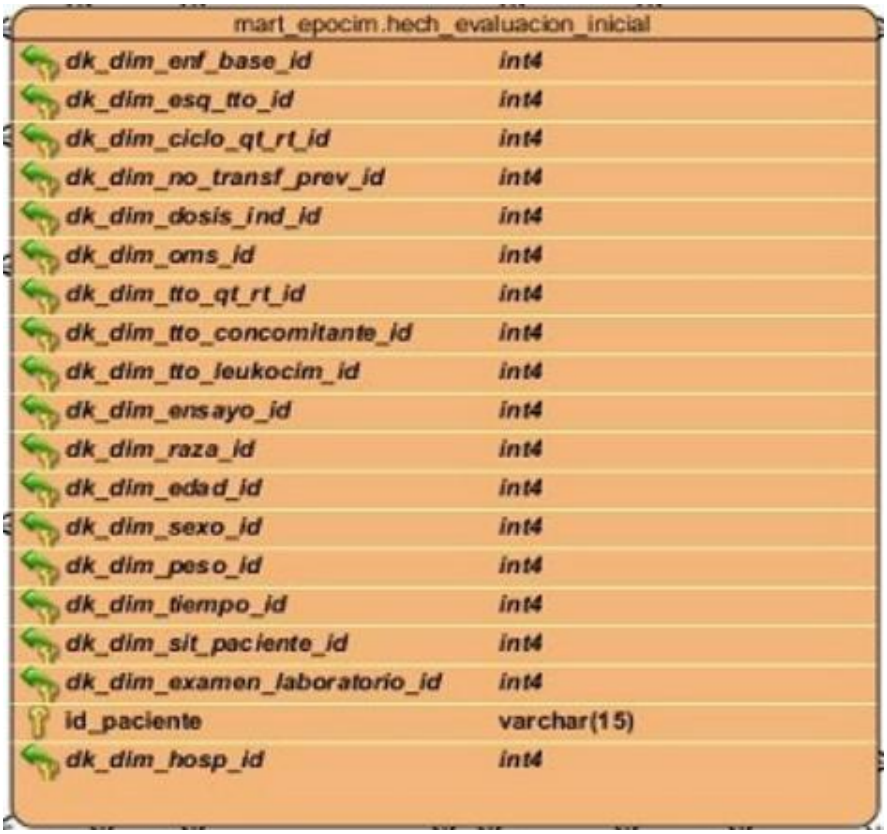


**Figura 5. Hecho hech_evaluacion_inicial.**

- **Evaluación intermedia** (hech_evaluacion_intermedia): Muestra los resultados de los pacientes de la segunda evaluación realizada a la cuarta semana. (Figura 6)


**Figura 6. Hecho hech_evaluacion_intermedia.**

- **Evaluación final** (hech_evaluacion_final): Muestra los resultados de los pacientes de la tercera evaluación realizada a la octava semana. (Figura 7)

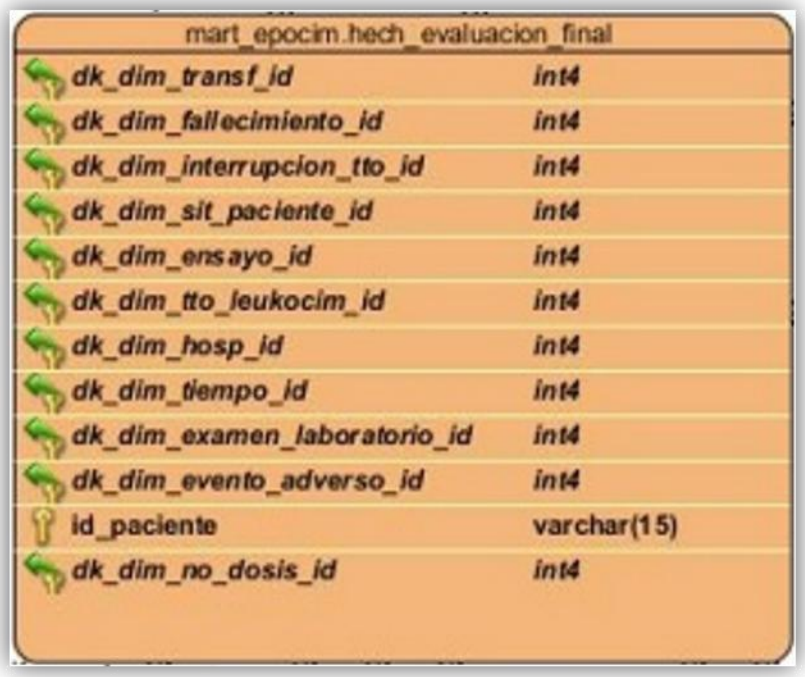


**Figura 7. Hecho hech_evaluacion_final.**

**Dimensiones.**

Las dimensiones constituyen las perspectivas de análisis de la información, se utilizan para almacenar los datos en un nivel de detalle deseado; se pueden dividir en niveles, donde en cada nivel la información es más resumida que en el anterior (15). Para el desarrollo de la solución se identificaron 24 dimensiones; una dimensión por cada variable requerida por el cliente. A continuación se presentan algunos ejemplos:

- Para la variable **tratamiento concomitante para la anemia** se define la dimensión **dim_no_dosis.** (Figura 8)
- Para la variable **número de dosis de EPOCIM recibidas** se define la dimensión **dim_tto_concomitante.** (Figura 9)


**Figura 8. Dimensión dim_no_dosis.**


**Figura 9. Dimensión dim_tto_concomitante.**

**Matriz bus.**

La Matriz Bus representa las relaciones existentes entre los hechos y las dimensiones, aportando una visión del impacto que provocaría un cambio durante el desarrollo del sistema.

**Tabla 4. Matriz bus.**

| Dimensiones | Hechos | | |
|---|---|---|---|
| | Evaluación inicial | Evaluación intermedia | Evaluación final |
| Edad | x | | |
| Peso | x | x | |

| | | | |
|---|---|---|---|
| Sexo | x | | |
| Color de la piel | x | | |
| Enfermedad base | x | | |
| Esquema de tratamiento | x | | |
| Hospital | x | x | x |
| Dosis de Epocim indicada | x | | |
| Fecha de evaluación | x | x | x |
| Ciclo de tratamiento QT y/o RT | x | | |
| No de transfusiones recibidas en los 2 meses previos | x | | |
| Quimioterapia y/o radioterapia | x | | |
| Estado clínico del paciente según OMS | x | | |
| Situación del paciente | x | | x |
| Tratamiento concomitante | x | | |
| Tratamiento con Leukocim | x | x | x |
| No. de dosis Epocim recibidas | | x | x |
| Transfusiones durante tratamiento | | x | x |
| Eventos adversos | | x | x |
| Mantiene dosis Epocim inicial | | x | |
| Interrupción de tratamiento | | x | x |
| Falleció durante tratamiento | | | x |
| Exámenes de laboratorio | x | x | x |
| Ensayo | x | x | x |

La realización de la matriz bus permitió conocer que aunque los tres hechos definidos comparten relación con algunas dimensiones, no existen hechos que se relacionen exactamente con las mismas dimensiones. Esto indica que se confirmó la inexistencia de solapamiento de hechos.

**Modelo de datos**

El Modelo de datos representa una descripción de la estructura de los datos, donde se definen las relaciones que se establecen entre las dimensiones y los hechos que lo componen. Para el desarrollo de los subsistemas de almacenamiento e integración Epocim, atendiendo a que existen tres tablas de hechos que comparten algunas de las dimensiones existentes, ver **figura 10,** se utiliza como topología la constelación de hechos.

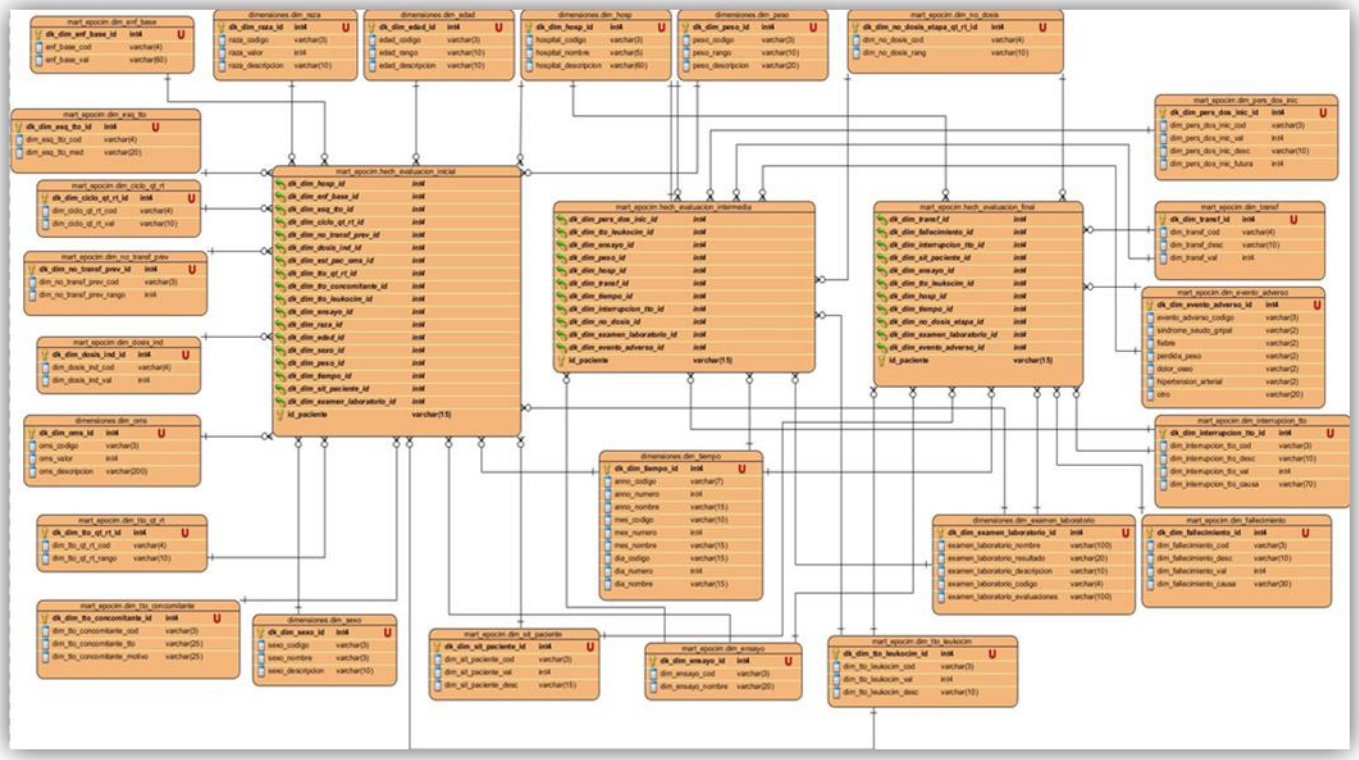

**Figura 10. Modelo de datos de los subsistemas de almacenamiento e integración Epocim.**

### 2.6.2 Subsistema de integración de datos

En el diseño del subsistema de integración de datos se realizaron tres importantes actividades para su posterior implementación; el perfilado de los datos, el diccionario de datos y el diseño de las transformaciones. Las mismas serán abordadas a continuación.

- Perfilado de datos.

El perfilado de los datos permite lograr una mejor comprensión de los mismos y verificar la existencia de valores nulos, distintos, duplicados, entre otros; permitiendo así, definir nuevas reglas del negocio que posteriormente pasan a ser las reglas de trasformación aplicadas en la implementación de la solución. Para mayor detalle consultar el artefacto “Perfil de Datos”. En la **Figura 11** se muestra un gráfico con los tipos de datos existentes en la fuente.

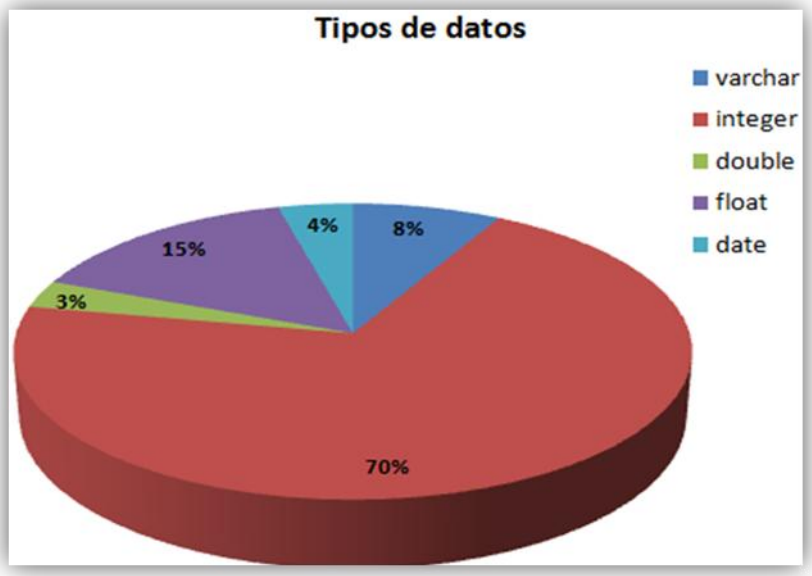


**Figura 11. Tipos de datos.**

- Diccionarios de datos.

El diccionario de datos es una lista organizada alfabéticamente de todos los elementos que forman parte del flujo de datos de todo el sistema, permitiendo guardar los detalles y la descripción de todos estos. La

elaboración de los diccionarios se desarrolla durante el análisis de flujo de datos facilitando el trabajo al evitar malas interpretaciones o ambigüedades. En la tabla que se muestra a continuación se listan las variables del sistema a desarrollar:

**Tabla 5. Diccionario de datos.**

| Nombre de las variables | Descripción de las variables |
|---|---|
| Ciclo de tratamiento QT y/o RT | Ciclo de quimio y/o radioterapia en la que se encuentra el paciente. |
| Dosis de Epocim indicada | Es la dosis del producto indicada al paciente. |
| Edad | Se refiere a la cantidad de años del paciente. |
| Enfermedad de base | Se refiere a la enfermedad que presenta el paciente al comienzo del ensayo. |
| Ensayo | Se refiere a si el producto se aplica en pacientes niños o adultos. |
| Esquema de tratamiento | Esquema de tratamiento con quimio y/o radioterapia que recibe el paciente para su enfermedad de base. |
| Estado clínico del paciente según OMS | Estado del paciente según la clasificación de la organización mundial de la salud. |
| Exámenes | Recoge valores cualitativos de los resultados de los exámenes. |
| Falleció durante tratamiento | Dice si el paciente falleció durante su tratamiento. |
| Fecha de la evaluación | Es la fecha en la que se realiza la evaluación del paciente. |
| Hospital | Se refiere al hospital donde se atiende el paciente. |
| Interrupción de tratamiento | Recoge si el paciente interrumpe tratamiento. |
| Mantiene dosis Epocim inicial | se refiere a si persiste la dosis inicial indicada. |
| No. de transfusiones recibidas en los 2 meses previos | Se refiere al número de transfusiones recibidas antes del tratamiento. |
| Número de dosis recibidas | Se refiere al número de dosis que ha recibido el paciente. |
| Peso | Se refiere al peso de los pacientes en kg. |

| | |
|---|---|
| Quimio y/o radioterapia | Se refiere a si ha recibido tratamiento con quimioterapia o radioterapia. |
| Raza | Se refiere al color de la piel. |
| Sexo | se refiere al sexo de los pacientes |
| Situación del paciente | Recoge si el paciente se atiende hospitalizado o ambulatorio. |
| Transfusiones durante tratamiento | Recoge si el paciente recibe transfusiones durante el tratamiento. |
| Tratamiento con Leukocim | Es el tratamiento que llevara el paciente. |
| Tratamiento concomitante | Dice si el paciente se trata con algún medicamento junto con Epocim. |

- Diseño general de las transformaciones.

El proceso ETL es el encargado de impulsar el flujo de datos haciendo transformaciones para lograr una integración exitosa. Este componente permite recopilar datos desde múltiples fuentes, formatearlos, limpiarlos y cargarlos en la base de datos, protegiendo su integridad. Este proceso resulta complejo, pues precisa de un alto nivel de detalle, por tanto se hace necesario un diseño general de las transformaciones que describa los pasos para realizar la carga de los hechos y las dimensiones a la base de datos.

Este diseño puede variar a la hora de ejecutar la implementación de las transformaciones, pues durante su desarrollo, pueden surgir situaciones con los datos, que requieren de diferentes estrategias para su solución. La Figura 12 muestra el diseño general para la carga de los hechos de los subsistemas de almacenamiento e integración Epocim.

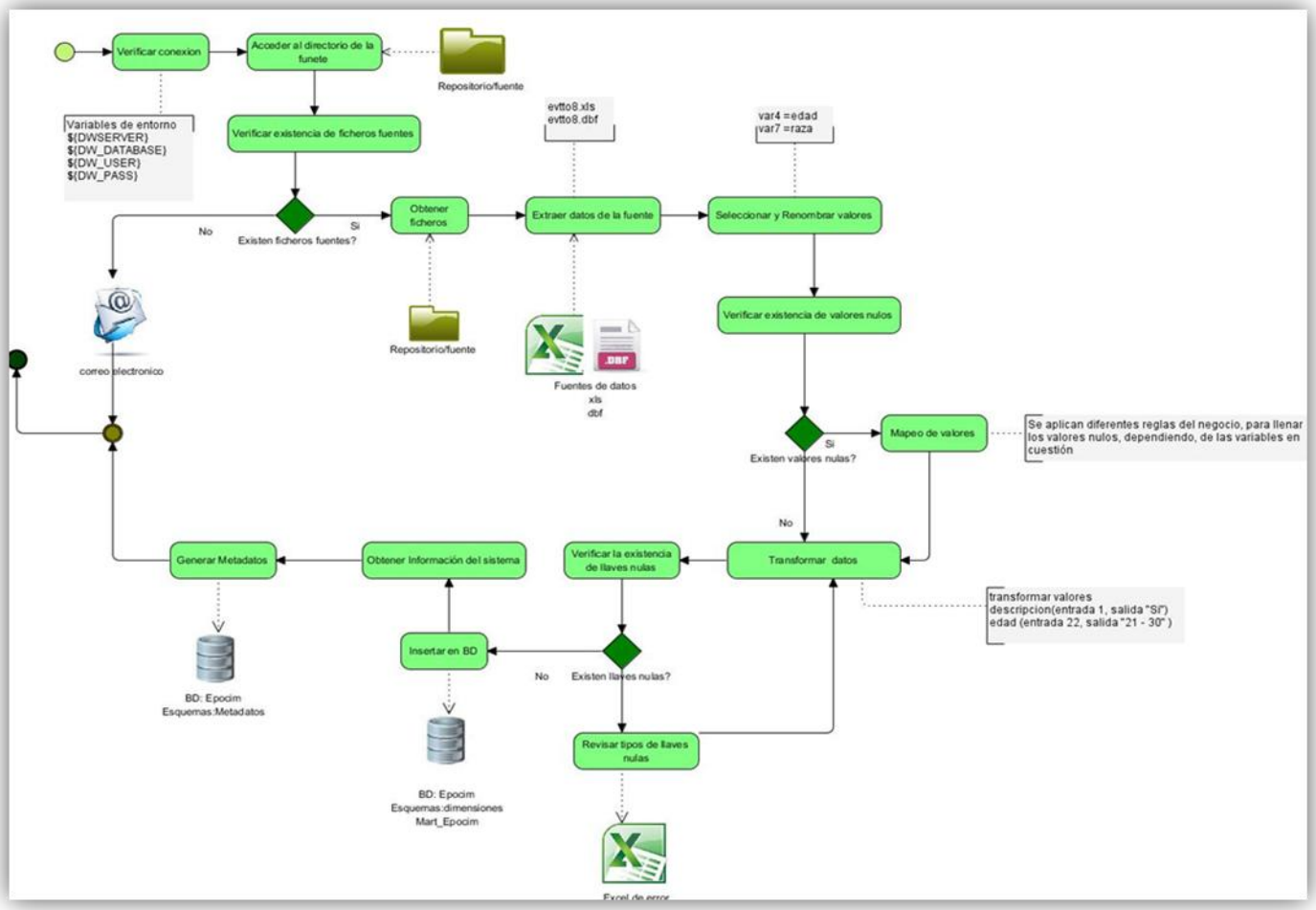


**Figura 12. Diseño general.**

## 2.7 Política de respaldo y recuperación

La política de respaldo y recuperación se establece con el objetivo de garantizar la persistencia de la información. En la presente investigación está basada fundamentalmente en las copia de seguridad, pues el sistema solo efectúa una carga histórica y no realizará carga incremental. Basándose en las características anteriores se efectuará una salva de toda la información presente en la base de datos, y se debe verificar la existencia de dos copias en ubicaciones diferentes, previniendo la ocurrencia de fallos en el sistema o de otra índole que incida en la perdida de la información. Igualmente ejecutando nuevamente las transformaciones se vuelve a obtener la información en el subsistema de almacenamiento.

## 2.8 Esquema de seguridad

La seguridad de los datos durante el proceso de ETL constituye una tarea de gran importancia, se hace necesario garantizar la misma pues de esta depende la confidencialidad e integridad de los datos. El esquema de seguridad representa el respaldado por los niveles de acceso, específicamente por los roles definidos:

- **Administrador BD:** administra la base de datos relacional que contiene todos los esquemas del almacén. Posee todos los permisos de administración y otorga los permisos a los diferentes usuarios.
- **Administrador ETL:** realiza los procesos de ETL, y tiene permiso de lectura y escritura sobre los esquemas.

En la realización de las ETL se garantiza la seguridad mediante las opciones brindadas por el sistema operativo, donde se asignan permisos a los diferentes usuarios sobre la carpeta Repositorio (nombre de la carpeta donde su ubican todos los datos de las fuentes y transformaciones); control total para el administrador, y solo lectura a los restantes usuarios que accedan a la información.

## 2.9 Conclusiones parciales

Durante el presente capítulo, en conjunto con los clientes, se hizo un estudio de las necesidades de información a contener en los subsistemas de almacenamiento e integración Epocim, arrojando como resultados:

- La realización de un estudio de las necesidades de información permitieron identificar 26 requisitos de información, 6 requisitos funcionales y 4 no funcionales.
- Identificación de 26 reglas del negocio, que se aplicarán durante la implementación de las ETL.
- Identificación de 3 casos de uso de información y 2 casos de uso de funcionales.
- Identificación de 3 tablas de hechos y 24 tablas de dimensiones, garantizando el cumplimiento de las necesidades de información del cliente.
- Obtención de la matriz bus, donde se aprecia la relación existente, entre cada una de las tablas de dimensiones y las tablas de hechos.
- El perfilado de datos realizado a la fuente arrojó el estado en que se encuentran los datos y la solución a las inconsistencias existentes.

- Obtención del diseño de las transformaciones, que permite guiarse a la hora de la implementación de las ETL.
- El esquema de seguridad y la política de respaldo y recuperación contribuyen a mantener la seguridad e integridad de los datos.

# CAPÍTULO 3: Implementación y prueba de los subsistemas de almacenamiento e integración Epocim.

**Introducción**

Luego de realizado el análisis y diseño de la solución, se procede a la realización de la implementación física del sistema. Esta etapa comprende la implementación de los subsistemas de almacenamientos e integración Epocim y se realizan las pruebas a la solución para comprobar su funcionalidad y calidad.

## 3.1 Implementación del subsistema de almacenamiento

La esencia de la implementación del subsistema de almacenamiento, es dotar al sistema de una correcta estructura y organización de los datos. Para ello se encauzó el trabajo del equipo en la estandarización de los nombres y la implementación de la estructura física de almacenamiento.

### 3.1.1 Estandarización de los nombres

Con la estandarización de los nombres se organiza la forma de denominar las estructuras, para lograr un patrón que contribuya a la correcta normalización de los términos utilizados y permita a los desarrolladores un mejor entendimiento de las estructuras.

En el presente trabajo, de forma general, las diferentes estructuras mantienen una nomenclatura similar. En los casos de que la tabla sea un hecho, el nombre va a estar compuesto por “hech” + “_” + ”clasificación_de_la_evaluación”, ejemplo hech_evaluación_intermedia. En el caso de las dimensiones estas se componen de la abreviatura “dim” + “_” + “nombre_de_la_dimensión”, ejemplo dim_tto_leukocim.

La estructura que presentan los atributos de las dimensiones están dadas por: “*nombre_referente_a_la_dimension*” + “**_**” + “**referencia al atributo**”, ejemplo “*dim_tto_leukocim*_**val**” o “*evento_adverso*_**codigo**”. En el caso de las llaves primarias de las dimensiones están denominadas de la forma “dk” + “_” + “nombre_de_la_dimension” + “id”, ejemplo “dk_dim_tto_leukocim_id”.

Las transformaciones de los hechos y las dimensiones no tienen variación en la nomenclatura, cada transformación presenta el mismo nombre de la estructura que se está ejecutando. En el caso de los trabajos los nombres vienen dados por el grupo al que se refieren, ejemplo “dimensiones iniciales”,

“dimensiones intermedias y finales”. Por su parte los metadatos están conformados por las letras “mp” o “mt” en dependencia del tipo metadatos al que hace alusión, seguido del carácter “_” y el nombre ejemplo mp_trabajos.

Realizada la estandarización de los atributos y tablas dentro del subsistema de almacenamiento, se puede dar paso a la implementación de la estructura física.

### 3.1.3 Implementación del modelo de datos físico

Con el objetivo de lograr brindar una estructura y organización a la información almacenada se definieron 3 esquemas que contienen las 30 tablas que presenta la solución; dimensiones con 5 tablas, mart_epocim contiene 22 y metadatos que presenta las 3 tablas restantes.

- **Dimensiones:** este esquema contiene todas las dimensiones compartidas con el resto de los mercados de datos pertenecientes al almacén del CIM.
- **Mart_epocim:** recoge las tablas de dimensiones y hechos propias de los Subsistemas de almacenamiento e integración Epocim.

- **Metadatos:** su objetivo está en realizar el control de cambios, donde se ha diseñado una estructura basada en el uso de 3 tablas de metadatos “mp_trabajos”, “mp_transformaciones” y “mt_gestion_carga_historica”. Las dos primeras tablas recogen los **Metadatos de proceso,** que permite obtener información de los procesos que se ejecuten, son una presentación de las estadísticas sobre los resultados de la ejecución del proceso de ETL, incluyendo medidas tales como filas cargadas con éxito, filas rechazadas, la cantidad de tiempo de carga. Mientras que la table “mt_gestion_carga_historica” recoge los **Metadatos técnicos,** los cuales están relacionados con la función del sistema o el modo en que se interrelacionan sus componentes (24).

En la **Tabla 6** se muestran las tablas asociadas a cada uno de los esquemas:

**Tabla 6. Esquemas y tablas de la aplicación.**

| Esquemas | Tablas | |
|---|---|---|
| **dimensiones** | dim_edad | dim_oms |

| | | |
|---|---|---|
| | dim_peso | dim_sexo |
| | dim_tiempo | |
| **mart_epocim** | dim_ciclo_qt_rt | dim_dosis_ind |
| | dim_enf_base | dim_ensayo |
| | dim_esq_tto | dim_evento_adverso |
| | dim_examen_laboratorio | dim_fallecimiento |
| | dim_hosp | dim_interrupcion_tto |
| | dim_no_dosis | dim_no_transf_prev |
| | dim_pers_dos_inic | dim_raza |
| | dim_sit_paciente | dim_transf |
| | dim_tto_concomitante | dim_tto_leukocim |
| | dim_tto_qt_rt | hech_evaluacion_final |
| | hech_evaluacion_inicial | hech_evaluacion_intermedia |
| **metadatos** | mp_trabajos | mp_transformaciones |
| | mt_gestion_carga_historica | |

### 3.2 Implementación del subsistema de integración

Para la implementación del subsistema de integración se realizan los procesos de extracción, transformación y carga. Antes de iniciar el desarrollo de estos procesos para la integración es de vital importancia realizar un análisis previo de cada una de las fuentes, para ello se realizó el perfilado de datos. El perfilado consiste en analizar la información a extraer y conocer el estado en que se encuentran los datos, formato, estructura y calidad de los mismos. La solución posee como fuente de datos archivos de extensión ***.dbf** y ***.xls**, los que almacenan la información histórica relacionada con el producto Epocim.

El proceso se inicia al extraer la información necesaria a utilizar para el producto Epocim desde las fuentes. Luego de obtener los datos se realizan la limpieza y transformaciones pertinentes, donde se valida y adapta la información a la estructura definida en la base de datos. Para finalizar el proceso se inicia el subproceso de carga de la información logrando así migrar los datos ya transformados con anterioridad al subsistema de almacenamiento.

### 3.2.1 Implementación de las transformaciones.

Las transformaciones están compuestas por pasos, que se encuentran unidos a través de saltos, en ellos se definen las reglas que serán establecidas en las rutinas de transformación. La Figura 13 muestra la transformación correspondiente al hecho hech_evaluación_final, donde se puede apreciar el flujo a través de todos los componentes, que limpian y modifican los datos hasta poder ser insertados en la BD. El tratamiento de errores existentes se implementó con el objetivo de obtener los campos que no se insertaban en la base de datos, durante las primeras iteraciones de la transformación, para corregir los detalles que lo impedían, y así lograr que se insertaran todos los datos provenientes de las fuentes.

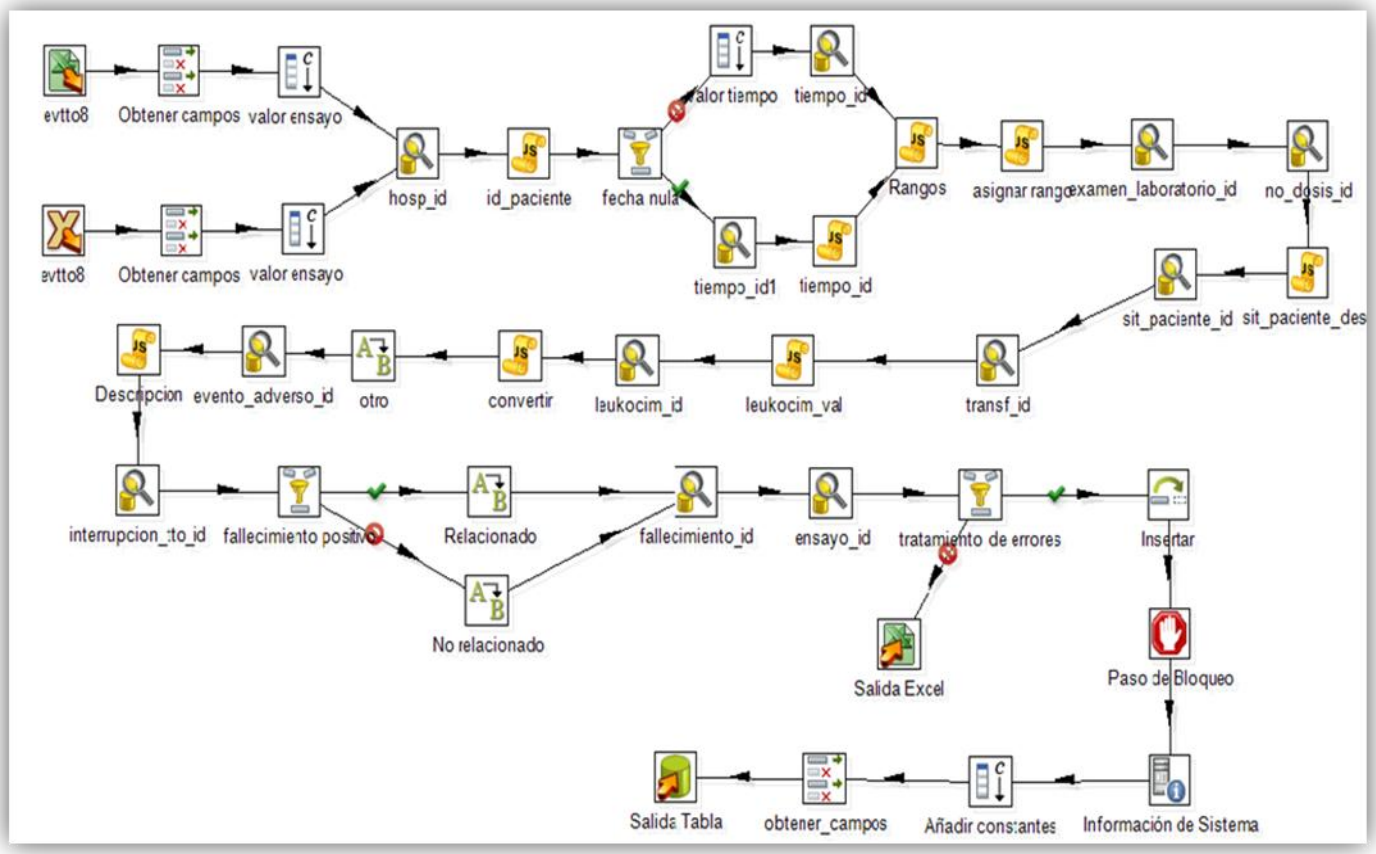


**Figura 13. Transformación del hecho hech_evaluación_final.**

### 3.2.2 Implementación de los trabajos

Los trabajos constituyen un proceso de tareas con el objetivo de realizar una operación específica. Los pasos disponibles en los trabajos, difieren de los existentes en las transformaciones, permiten la ejecución de una o varias de estas que se encuentren diseñadas y elaborar una secuencia lógica de ellas. En la Figura 14 se aprecia el trabajo principal que ejecuta todas las transformaciones diseñadas, con el objetivo de cargar los datos al subsistema de almacenamiento. El trabajo presenta la peculiaridad que toda la información fluye por un solo flujo, pues los hechos tienen dependencia entre sí para su ejecución, pero se divide la carga de las dimensiones en tres etapas, según la dependencia de estas y los hechos, logrando así, que en caso de error en alguna de las ejecuciones de las dimensiones, no impidan la carga de los hechos que le anteceden.

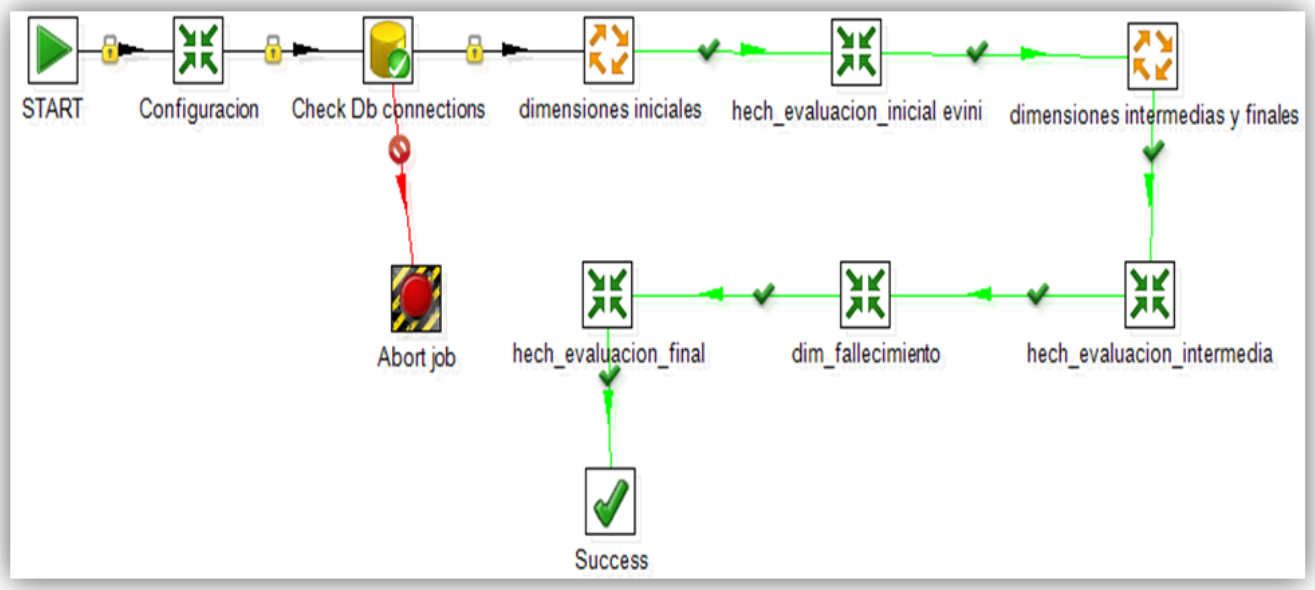


**Figura 14. Trabajo general.**

### 3.2.3 Gestión del cambio lento en las dimensiones

Las dimensiones lentamente cambiantes o SCD (Slowly Changing Dimensions) son dimensiones las cuales sus datos tienden a modificarse a través del tiempo, ya sea de forma ocasional o constante. Cuando ocurre este tipo de cambios, se puede optar por realizar el registro del historial de cambios o reemplazar los valores que sean necesarios.

Ralph Kimball inicialmente planteó en su libro “The Data Warehouse ETL Toolkit” tres estrategias a seguir cuando se tratan las SCD: tipo 1, tipo 2 y tipo 3. A través de los años, se ha profundizado en el estudio de las definiciones iniciales y han surgido los tipos 0, 4 y 6. (21)

- **Tipo 0 (no tiene en cuenta la gestión histórica):** no se realiza ningún esfuerzo para lidiar con los problemas del cambio de la dimensión.

- **Tipo 1 (sobrescribir):** es utilizado cuando la información histórica no es relevante. Este tipo de SCD sobrescribe los datos antiguos con nuevos y es usado por lo general con el objetivo de

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 ocurre algún cambio en la dimensión se crea una nueva entrada en la tabla. Al nuevo registro es asignada una nueva llave subrogada, valor que será usado para futuras entradas, mientras que las antiguas utilizarán el valor anterior.

- **Tipo 3 (añadir columna):** esta estrategia requiere que se agregue una nueva columna a la tabla de dimensión por cada campo cuyos valores deben incluir 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.

En los Subsistemas de almacenamiento e integración Epocim no se tiene en cuenta el cambio histórico, pues son ensayos cerrados, por tanto se usa el tipo 0 de los SDC.

### 3.3 Pruebas

Para lograr que un producto cumpla con las expectativas y la calidad que el mercado merita es necesario pasar por la etapa de pruebas. Es de gran importancia para los desarrolladores llevar este proceso durante la etapa de construcción hasta finalizar el mismo con la entrega del producto, pues permite encontrar los errores en etapas tempranas para darle la solución adecuada y entregar dicha solución en el tiempo pactado con el cliente. De esta manera se obtienen resultados satisfactorios para todos los interesados incluyendo el beneficio que se obtiene moralmente para el equipo de trabajo o empresa. El

Instituto de Ingenieros Eléctricos y Electrónicos (IEEE según sus siglas en inglés) define como pruebas de software a la actividad en la cual un sistema o componente es ejecutado bajo condiciones específicas, se observan o almacenan los resultados y se realiza una evaluación de algún aspecto del sistema o componente (20).

Las pruebas se realizan con objetivos específicos en diferentes escenarios, utilizando diversas estrategias y herramientas. En el presente trabajo diploma para validar los resultados de la solución Subsistema de almacenamiento e integración Epocim se realizaron pruebas Unitarias, de Integración y Listas de chequeo.

### 3.3.1 Pruebas unitarias

Las pruebas **Unitarias** consisten en probar diferentes componentes de la solución que tengan funcionalidades específicas del producto; su objetivo es asegurar el correcto funcionamiento de todas las partes del subsistema de manera independiente (19). Las pruebas se desarrollaron con la ayuda de los profesores del departamento Almacenes de Datos y otros integrantes del centro DATEC lo que permitió que el producto estuviera listo para su uso. A continuación se muestran 5 no conformidades encontradas durante todo el proceso de construcción de la solución por el equipo de probadores:

1) Los casos de uso no están acorde con la representación del modelo.
2) Revisar las variables de salida.
3) Revisar el diseño de las bases de datos y reglas de transformación.
4) No especifica las variables de entorno que utiliza en la implementación a través del diseño de ETL.
5) No define en los diseños de los proceso de integración de datos la estructura del repositorio definido por la dirección del proyecto.

### 3.3.2 Pruebas de integración

Dentro de las soluciones informáticos algunos componentes son combinados con otros. Para asegurarse que en los datos compartidos ellos, no son introducidos errores, se efectúan las pruebas de integración. En la presente investigación se realizan estas pruebas para determinar que la unión del subsistema de almacenamiento y el subsistema de integración funcione correctamente y no se hayan introducido datos diferentes a los existentes en la fuente.

Se diseñaron seis casos de prueba, con el propósito de verificar los requisitos de información, agrupados en tres casos de uso de información que fueron definidos previamente durante la etapa de análisis.

En la tabla 7 se muestra un fragmento del caso de prueba correspondiente al caso de uso mantener disponible la información de la aplicación del producto Epocim en la evaluación inicial, donde se aprecia una consulta realizada al sistema, de la cantidad de pacientes niños de sexo femenino que se incluyeron en el ensayo, obteniendo igual resultado, al efectuar la misma búsqueda en la fuente de datos.

**Tabla 7. Caso de Prueba: Mantener disponible la información de la aplicación del producto Epocim en la evaluación final.**

| Caso de uso de información | Requisito de información | Tablas implicadas | Variables de entrada | Variables de salida | Consulta SQL realizada | Datos obtenidos | Fuente de datos | Columnas de la fuente | Datos almacenados en la fuente | Resultados de la prueba |
|---|---|---|---|---|---|---|---|---|---|---|
| Mantener disponible la información de la aplicación del producto Epocim en la evaluacion inicial | Obtener la cantidad de pacientes niños que se incluyeron en el ensayo atendiendo a los datos demograficos | dim_sexo<br>hech_evaluacion_ini | dim_sexo<br>id_paciente<br>hech_evaluacion_ir | cant_pacientes | select sum(p.cantidad) as cantidadpacientes from (SELECT count(evaluacioninicial.id_paciente) as cantidad FROM mart_epocim.hech_evaluacion_inicial evaluacioninicial inner join dimensiones.dim_sexo sexo on evaluacioninicial.dk_dim_sexo_id = sexo.dk_dim_sexo_id inner join mart_epocim.dim_ensayo ensayo on evaluacioninicial.dk_dim_ensayo_id = | Se obtuvo 105 pacientes niños del sexo femenino | Archivo en formato ,dbf evini.dbf | var 7<br>CODID | Se obtuvo 105 pacientes niños del sexo femenino | Satisfactorio |

Las pruebas fueron diseñadas y ejecutadas por los desarrolladores cuando la solución estaba completa. Las mismas demostraron la existencia de dos No Conformidades (NC) que fueron resueltas con inmediatez para la segunda iteración como se muestra en la Figura 15:

I. Los nombres de los atributos y dimensiones poseen nombres complicados debido a la cantidad de abreviaturas que se emplean.
II. El reporte “Obtener la cantidad de pacientes adultos que fallecieron durante el tratamiento y su causa” no muestra resultados correctos.

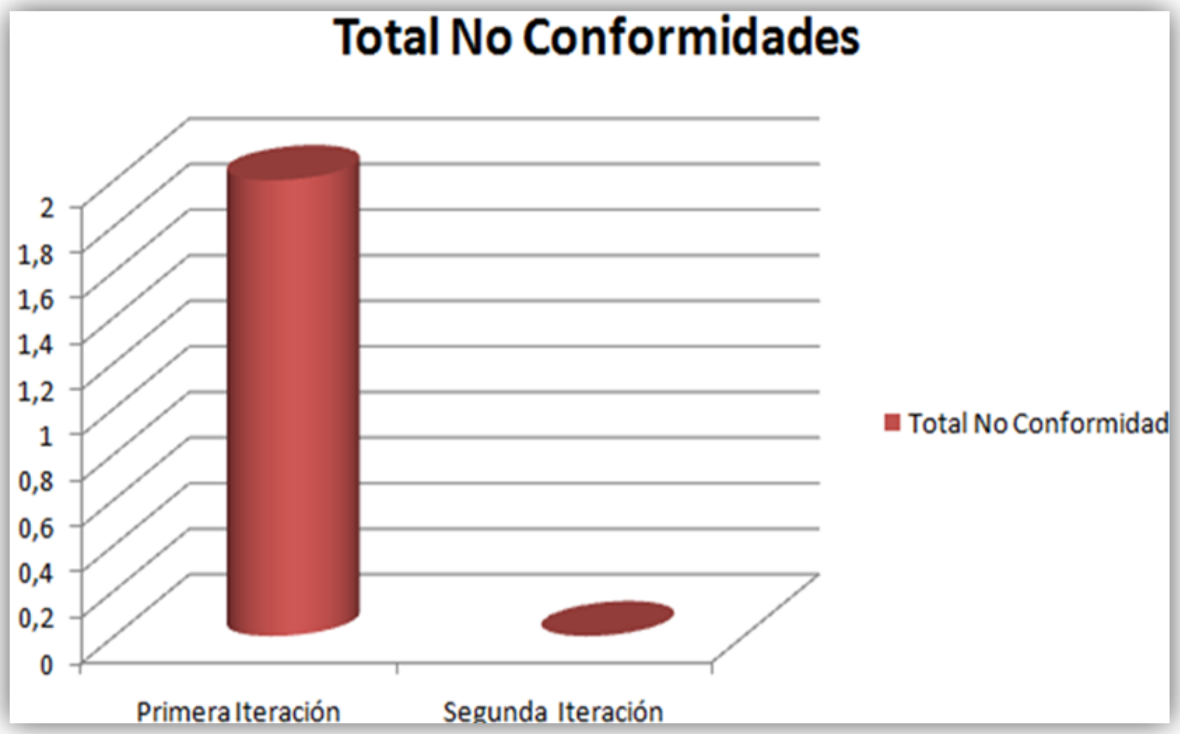


**Figura 15. Cantidad de No Conformidades encontradas en las pruebas de sistema.**

### 3.3.3 Listas de chequeo

Las listas de chequeo son las encargadas de guiar el proceso de evaluación en función de la información que se tiene de las características y calidad del producto, estos datos se recogen de un cuestionario donde las respuestas son concretas SI (1) o NO (0). Después de tener estos resultados se realiza una evaluación que arroja resultados cualitativos.

Las listas de chequeo contienen diferentes indicadores a evaluar los cuales se encuentran distribuidos en tres secciones fundamentales (17):

- **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.
- **Semántica del documento:** contempla todos los indicadores a evaluar respecto a la ortografía, redacción y demás.

Elementos que forman parte de la estructura de la lista de chequeo:

- **Peso:** define si el indicador a evaluar es crítico o no.
- **Indicadores a evaluar:** son los indicadores a evaluar en las secciones Estructura del documento, Semántica del documento e Indicadores definidos por la etapa.
- **Evaluación:** es la forma de evaluar el indicador en cuestión. El mismo se evalúa de 1 en caso de que exista alguna dificultad sobre el indicador y 0 en caso de que el indicador revisado no presente problemas.
- **N.P. (No Procede):** se usa para especificar que el indicador no es necesario evaluarlo en ese caso.
- **Cantidad de elementos afectados:** 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.

**Evaluación a través de la Lista de Chequeo:**

Se elaboraron listas de chequeo para evaluar el diccionario de datos, mapa lógico de los datos, perfilado de datos y el registro del sistema fuente con un total de 13 indicadores distribuidos en 3 secciones: Estructura del documento, Indicadores definidos y Semántica del documento; de los que 4 fueron definidos como críticos para así obtener el producto la evaluación de bien, en la Figura 16 se revelan los resultados obtenidos.

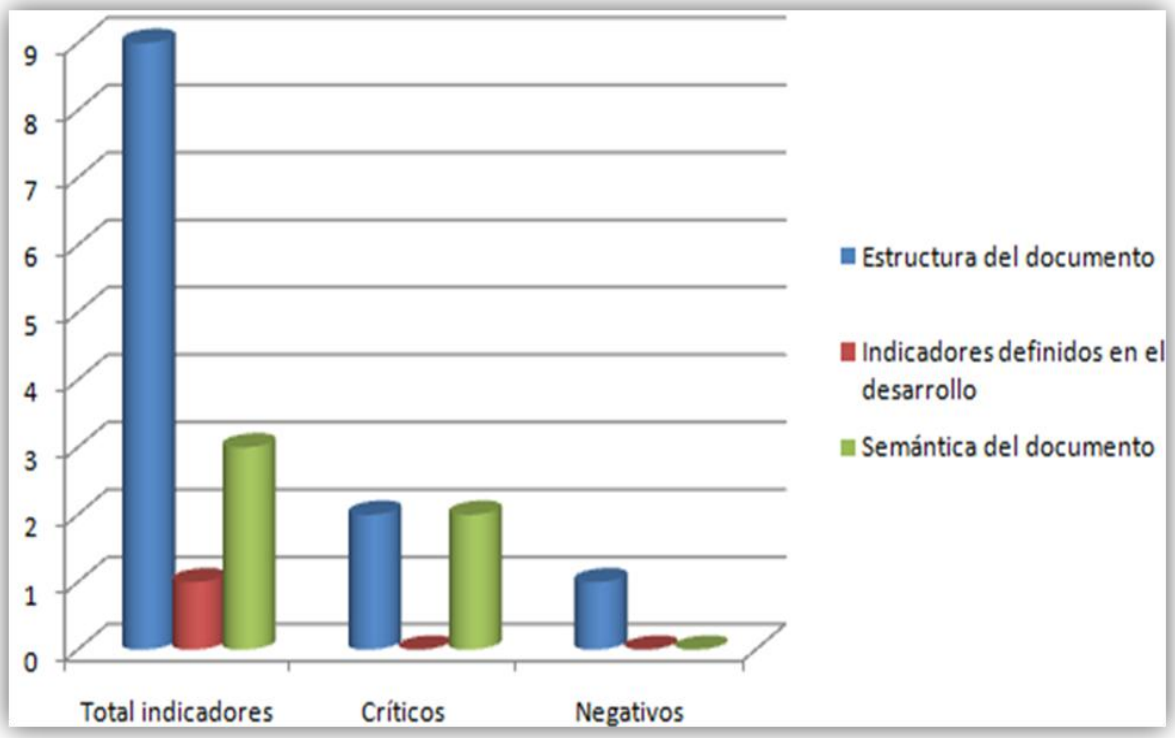


**Figura 16. Aplicación de los indicadores de las Listas de Chequeo al subsistema de almacenamiento e integración Epocim.**

### 3.4 Conclusiones parciales

Durante el capítulo se abordó sobre la implementación y pruebas realizadas a los Subsistemas de almacenamiento e integración Epocim, concluyendo así el proceso de construcción y validación, obteniéndose los siguientes resultados:

- Quedó definida la estructura de los Subsistemas de almacenamiento e integración Epocim, de forma conjunta con los demás subsistemas de productos del CIM, con tres esquemas, permitiendo que al integrarlos al AD del CIM, exista una estructura homogénea.
- Se realizaron 3 transformaciones para la carga de los hechos, 3 para los trabajos y 20 para las dimensiones que permitieron eliminar las inconsistencias existentes y el almacenamiento de toda la información necesaria de la fuente.

- Con la realización de pruebas unitarias, de integración y de las listas de chequeo a los artefactos de ETL, se logró probar que la solución propuesta cumple con los requisitos definidos por el cliente, de esta manera quedo validado el subsistema de almacenamiento e integración.

# CONCLUSIONES

Una vez terminada la investigación y desarrollo de los Subsistemas de almacenamiento e integración Epocim se arriba a las siguientes conclusiones:

- El estudio de las metodologías y herramientas permitió la selección de las más acorde para que el proceso de desarrollo de la solución transitara por las fases del ciclo de vida y la construcción de la solución, fuera coherente con las normas y tendencias de la UCI.
- Mediante el análisis y diseño de los Subsistemas de almacenamiento e integración Epocim se generaron los artefactos necesarios, que guiaron la posterior etapa de implementación.
- La implementación de las estructuras definidas, permitieron la integración de los datos históricos del producto Epocim y su almacenamiento
- Las validaciones realizadas permitieron comprobar la calidad del producto a partir de los requisitos establecidos. Certificando de tal forma que la solución cumple con las necesidades del cliente.

De esta manera se logra el cumplimiento del objetivo planteado de contribuir a homogenizar la información almacenada de los ensayos clínicos del Centro de Inmunología Molecular.

## RECOMENDACIONES

Con el objetivo de mejorar la solución propuesta realizada en el presente trabajo diploma los autores se sugieren:

- Realizarle técnicas de minería de datos al subsistema de almacenamiento del producto Epocim, permitiendo detectar patrones de comportamiento sobre la información almacenada.
- Adicionar el subsistema de visualización de la información que permitan al usuario realizar análisis de la información más detallado; y aplicar técnicas de inteligencia de negocios para contribuir con la toma de decisiones.

## REFERENCIAS BIBLIOGRAFICAS

**1.** CENTRO DE INMUNOLOGÍA MOLECULAR. Información General [en línea] [Consulta: 25 de noviembre 2009]. Disponible en: http://www.cim.sld.cu/.

**2**. Ministerio de Salud Pública. Centro para el control estatal de la calidad de los medicamentos. Regulación No. 45-2007. [En línea] [Citado el: 27 de Noviembre de 2010] Disponible en: http://www.bvv.sld.cu./

**3.** Almacén de datos Sala situacional /Yaneisy Pedraza González, Edgar Rojas Ricardo. Disponible en: http://publicaciones.uci.cu/index.php/SC/search/advancedResults 15/12/2012.

**4.** Inmon, W. H. Building the Data Warehouse. sl: Wiley Publishing. ISBN: 0-471-08130-2.

**5.** Cabrera Casales, Ms. María Evelia. [En línea] [Citado el: 27 de Noviembre de 2010] Disponible en: http://hp.fciencias.unam.mx/~alg/bd/Almacen de Datos.pdf.

**6.** Velasco, Roberto Hernando. [En línea] [Citado el: 20 de Marzo de 2011.] Disponible en: http://www.rhernando.net/modules/tutorials/doc/bd/dw.html.

**7**. Bello Mesa, Viviana Extracción, transformación y carga del mercado de datos Racotumumab para el almacén de datos del Centro de Inmunología Molecular: s.n., 2011.

**8.** Díaz Morales, Themis Patricia y Bermúdez Rodríguez, José Salvador. Diseño de un Datawarehouse para los Ensayos Clínicos que se gestionan en el Centro de Inmunología Molecular. Habana: s.n., 2010.

**9.** RICARDO DARIO, Ing. Bernabeu. Data Warehousing: Investigación y sistematización de conceptos. Hefesto: Metodología propia para la construcción de un Datawarehouse. Córdoba, Argentina, 2009.

**10.** POSTGRESQL, O. PostgreSQL. Global DevelopmentGroup, 2011 [22/11/2011] Disponible en: http://www.postgresql.org.

**11.** PGADMIN, O. PgAdmin for PostgreSQL Tools, 2011. [22/11/2011]. Disponible en: http://www.pgadmin.org.

**12.** HUMAN INFERENCE, O.Data Cleaner, 2011. [22/11/2011] Disponible en:http://datacleaner.eobjects.org.

**14.** [En línea]. [Citado el: 21 de abril de 2013.] Disponible en: http://www.msdn.microsoft.com/es-es/library/aa577691(v=bts.10).aspx

**15.** Tamayo, Marysol y Moreno, Francisco J. Análisis del modelo de almacenamiento MOLAP frente al modelo de almacenamiento ROLAP. Ingeniería de investigación, vol. 26, no. 3. [En línea] Septiembre-Diciembre de 2006. [Citado el: 11 de Noviembre de 2012.] Disponible en: http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0120-56092006000300016&lng=pt&nrm=iso. .

**16.** JUAN MANUEL CUEVA LOVELLE, Calidad de software, Universidad de Oviedo, España, 1999[En línea][Disponible en:http://gidis.ing.unlpam.edu.ar/downloads/pdfs/Calidad_software.PDF]

**17.** CALISOFT. [En línea] [10/5/2012]. [Disponible en:http://calisoft.uci.cu/tmp/documentos/normas/iso/NC-ISO-IEC%209126-1.pdf.

18. Técnicas de prueba; [En línea][Citado el: 16 de febrero de 2013.] Disponible en: http://indalog.ual.es/mtorres/LP/Prueba.pdf

**19.** PRESSMAN, R. S. ¨Ingeniería del Software. Un enfoque práctico¨. Madrid: s.n.,[Citado el: 16 de febrero de 2013.]

**20.** IEEE, (1990). "IEEE-610-12-Standard Glossary of Software Engineering Terminology Institute of Electrical and Electronics Engineers". ISBN: 155937067X. Disponible en: http://www.ieee.org/sitemap.html. Fecha de consulta: enero de 2011.

**21.** Kimball, R. y Ross, M. The Data Warehouse Toolkit: the Complete Guide to Dimensional Modelling. New York, EE.UU: John Wiley & Sons, 2002.

**22.** Peñaloza, Lucía Victoria Hernández. Tesis para lograr el título de Magíster: Diseño y Construcción de un DataMart para la mantención de Indicadores de Sostenibilidad de la Industria del Salmón. . Chile: s.n., 2008.

**23.** Data Warehouse [En línea][Citado el: 15 de noviembre de 2012.] Disponible en: http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/MonoAdsDiseno.pdf.

**24. Medi**na Mustelier, Doris**.** *Técnicas de Extracción, Transformación y Carga de Datos del Sistema de Información Nacional de Seguridad Ciudadana en la República Bolivariana de Venezuela*, Marzo 2009.

## BIBLIOGRAFIA CONSULTADA


1. CENTRO DE INMUNOLOGÍA MOLECULAR. Información General [en línea] .[Consulta: 25 de noviembre 2009]. Disponible en: http://www.cim.sld.cu/.
2. Inmon, W. H. Building the Data Warehouse. sl: Wiley Publishing. ISBN: 0-471-08130-2.
3. Metodología para el desarrollo de soluciones de Almacenes de Datos e Inteligencia de Negocio en CENTALAD. Edtion ed., 2009.
4. KIMBALL, R. *The Data Warehouse Toolkit*. 2. Canada, Robert Ipsen, 2002. 447 p.
5. Almacén de datos Sala situacional /Yaneisy Pedraza González, Edgar Rojas Ricardo (http://publicaciones.uci.cu/index.php/SC/search/advancedResults) 15/12/2012.
6. Inmon, W. H. Building the Data Warehouse. sl: Wiley Publishing. ISBN: 0-471-08130-2.
7. Cabrera Casales, Ms. María Evelia. [En línea] [Citado el: 27 de Noviembre de 2010] http://hp.fciencias.unam.mx/~alg/bd/Almacen de Datos.pdf.
8. Velasco, Roberto Hernando. [En línea] [Citado el: 20 de Marzo de 2011.] http://www.rhernando.net/modules/tutorials/doc/bd/dw.html.
9. Bello Mesa, Viviana Extracción, transformación y carga del mercado de datos Racotumumab para el almacén de datos del Centro de Inmunología Molecular: s.n., 2011.
10. Díaz Morales, Themis Patricia y Bermúdez Rodríguez, José Salvador. Diseño de unDatawarehouse para los Ensayos Clínicos que se gestionan en el Centro de Inmunología Molecular.Habana: s.n., 2010.
11. RICARDO DARIO, Ing. Bernabeu. Data Warehousing: Investigación y sistematización de conceptos.Hefesto: Metodología propia para la construcción de un Datawarehouse. Córdoba, Argentina, 2009.
12. POSTGRESQL, O. PostgreSQL. Global DevelopmentGroup, 2011 [22/11/2011] Disponible en: http://www.postgresql.org.
13. .PGADMIN, O. PgAdmin for PostgreSQL Tools, 2011. [22/11/2011]. Disponible en:http://www.pgadmin.org.
14. HUMAN INFERENCE, O.Data Cleaner, 2011. [22/11/2011] Disponible en:http://datacleaner.eobjects.org.