Universidad de las Ciencias Informáticas

Facultad 6

Título: *Subsistemas de almacenamiento e integración N Acetil GM3 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

Autor:

Luis Ángel Fis Aldana

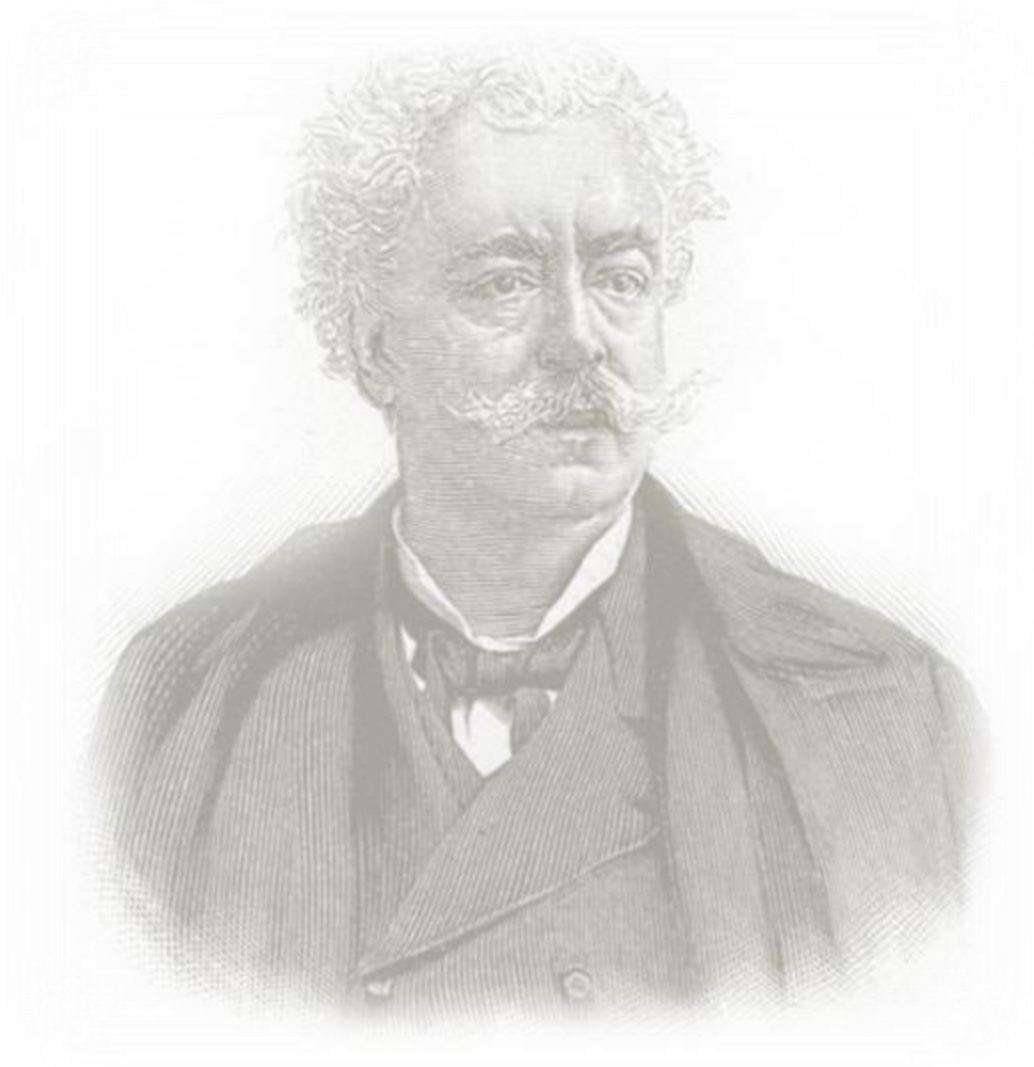
Tutoras:

Ing. Adalennis Buchillón Soris

Ing. Marisel Santana Rodríguez

La Habana

"Año 55 de la Revolución"



"Por este mundo pasaré solamente una vez, si hay una buena obra que pueda hacer, si hay una buena palabra que pueda decir; haré esa buena obra y diré esa buena palabra, pues ya nunca volveré a pasar por aquí"

Edmundo De Amicis

Declaración de autoría

Yo: Luis Ángel Fis Aldana, declaro ser autor de la presente tesis y reconozco a la Universidad de las Ciencias Informáticas los derechos patrimoniales de la misma, con carácter exclusivo.

Para que así conste firmo la presente a los _____ días del mes de _____ del año 2013.

_____

Autor: *Luis Ángel Fis Aldana.*

_____ _____

Tutora: *Ing. Adalennis Buchillón Soris.* ***Tutora:*** *Ing. Marisel Santana Rodríguez.*

Contacto

Ing. Adalennis Buchillón Soris

Graduado en Ingeniería en Ciencias Informáticas en el 2012 en la UCI.

La Habana, Cuba

Correo Electrónico: abuchillon@uci.cu

Ing. Marisel Santana Rodríguez

Graduado en Ingeniería en Ciencias Informáticas en el 2010 en la UCI.

La Habana, Cuba

Correo Electrónico: msantana@uci.cu

Agradecimientos

Una vez Neruda quiso escribir las cosas más tristes esa noche, pues yo quisiera, aunque se me hace muy difícil, plasmar en tan solo pocas páginas a todas esas personas que de una forma u otra han hecho de estos 5 años los mejores de mi vida.

Primeramente gracias a dios por darme la fuerza para seguir en los momentos que más la necesitaba y por siempre estar a mi lado, pues nunca temí porque él siempre estuvo a mi lado y principalmente por poner en mi vida a todos estos ángeles que han hecho de mi vida el paraíso en la tierra.

Gracias a mi familia que es lo más grande que dios me ha dado, por apoyarme en los momentos que más lo necesité y por confiar en que nunca los defraudaría.

A mi mamá que me dio la vida hace 24 años y que hoy me toca pagarle ese regalo con este título que tanto deseaba.

A mi papá por siempre estar dándome apoyo, ánimo y recordarme que estaba estudiando para beneficio mío y alegrías de mi familia y amigos, que ya lo decía la biblia que el camino del triunfo es largo y difícil y solo los valientes lograrán cruzarlo.

A mi hermanito, la sangre de mi sangre, le agradezco por haberme regalado la alegría de haber aprobado las pruebas de ingreso, por estar ahí siempre dándome el ánimo y los consejos de jóvenes que a veces mis padres no podrían entender porque les tocó vivir otra época.

A mi abuela Ramona, o mejor dicho mi otra madre, porque nunca dudó de mí y ser una de mis más grandes inspiraciones en los retos grandes que me ha dado la vida.

A mi prima Elizabeth que ha sabido levantarme los ánimos y ganarse el título de ser mi primita predilecta y dedicarme en sus correos las frases más alentadoras y lindas que pueda dedicarle una prima a su primo.

A mi abuela Mima, a mi abuelo, a mis tías Nena, María Elsa, Marianita, Miladis y Bárbara que siempre estuvieron pendiente de mí, a Carlitos, Carmen, Edelmys, Susana y Mercedes.

A las personas más importantes en mi paso por esta escuela, los que fueron mi familia durante cinco años y otros que fueron durante menos pero han marcado una pauta en mi vida.

Elisabeth, por ser aquí la hermana hembra que no tengo, por haberme aguantado en este año y medio que nos conocemos, por compartir cada momento de alegría y tristeza, por haber hecho de mí una mejor persona, por haberme enseñado que hay personas que definitivamente llegan y te marcan la vida, y no es porque ellas quieran, es simplemente porque ellas saben dar cambios en tu vida y pintar tus días de los colores que incitan la felicidad, aunque sea el más oscuro de ellos.

Agradecimientos

Yoanna, por haber aguantado todas mis pesadeces durante cuatro años, por ser mi amiga, por ser esa persona que no vacila para decirte las verdades y las cosas que tienes que hacer sin darle vueltas al asunto, mi consejera, mi guía, gracias Yoa.

Irela, mi gordita del alma, por haber compartido cinco años a mi lado en mi misma aula, haber sido como yo le decía la mujer de muchos de los varones del aula, gracias Iry, te quiero mucho.

Lianet, por ser mi tercera tutora, mi amiga y la persona que un día me pidió que ella quería que nos graduáramos juntos. El destino no nos quiso complacer poniéndonos de dúo de tesis pero nos complació hoy en el momento en que los dos compartimos el título de ingenieros.

A Chávez y Tomás, por haber estado en las buenas y en las malas, por compartir las notas buenas y las malas, por ser partícipe de muchos de mis más grandes momentos en esta escuela, esos momentos que nunca olvidaré y por haber entablado conmigo esta amistad que tanto aprecio.

A René por haberme aguantado durante todo este tiempo y aun así conservar nuestra amistad y dar el paso al frente siempre que lo necesité.

A Elizita, por ser esa mujer que cualquier hombre quisiera en su vida, la detallista, la cariñosa, por ser la mujer más completa que conocí en esta escuela, por regalarme su te quiero cada día, su sonrisa, sus palabras de aliento, su apoyo incondicional, Ely, te quiero mi vida.

A Ana del Carmen por enseñarme que el esfuerzo no lo es todo, hay veces que las cosas están escritas en las estrellas y cuando no está para ti ni aunque te pongas.

A mis compañeros de aula durante estos cinco años, a esas personas que formaron parte de mi club de amistades: Masacre, el choco, Sandy, Darlon, Marcos, Osvaldo, Alejandro, Félix, Osman, Yaily, Natalie, Lijandy, Martha Maricela, Diana, Laritza y Carlos, Yanet, Yuliet, Pablo, Yahima, Dayatni, Rebeca, Jineth, Yadir, Reanna, Anita, a la gente de Moa que compartieron estos años conmigo. A Ruberlando y Lissette que aunque no están presentes en este momento conmigo, fueron muy importantes para mí en el tiempo que lo estuvieron y los recordaré siempre por lo que representaron para mí. En general le agradezco a todas esas amistades que hice en estos años.

A Katherine, la que bauticé como mi ángel por aparecer en mi vida en los momentos que más necesitaba y ser mi ángel de la suerte y enseñarme que lo que eres, es el regalo de dios para ti, lo que haces de ti mismo, es tu regalo para él y yo quisiera que este regalo llegara en estos momentos a él, como recompensa a lo que ha hecho por mí.

Agradecimientos

Para los profes que de una forma u otra me ayudaron a lo largo de toda mi carrera y que contribuyeron a que hoy yo fuera un profesional, Arodys, Yania, Sonia, Liván, el profe Raymundo Bermúdez, la profe Milena, Lienner, Manolo, Yanelis, Adilen, Yordanka, Esley, Ernesto, Elio, Carlos Luis, a Fabián, a Leonel, a Themis, a Miguel, a las personas que me tutoraron durante el trabajo de diploma, la profe Marisel y Adalennis y en especial al profesor que más le debo en el transcurso de mi tesis al profe Monchy por estar siempre cuando lo necesité, por no apartarse de mi lado hasta que todo funcionara, por su entrega a mi causa, gracias profe.

Dedicatoria

Se la dedico a mi familia en general que son mi razón de ser, a Dios por estar siempre conmigo, a esta revolución que me dio la oportunidad de estudiar y graduarme en esta escuela, y a todas las personas que de una forma u otra han formado parte mi vida.

La investigación surge como parte de una colaboración entre el Centro de Inmunología Molecular (CIM) y la Universidad de las Ciencias Informáticas (UCI). La solución propuesta consiste en el desarrollo de los subsistemas de almacenamiento e integración de un mercado de datos para el CIM. El objetivo fundamental de la investigación radica en la estandarización de los datos asociados al ensayo clínico realizado en el CIM mediante la aplicación del producto N Acetil GM3 para su almacenamiento de forma homogénea. De esta manera se contribuye a la toma de decisiones en la institución.

Con el fin de dar cumplimiento al objetivo propuesto se integra la información en un espacio único de almacenamiento donde se encuentra centralizada y estandarizada. Los procesos de integración de datos fueron realizados a partir de la técnica Extracción, Transformación y Carga (ETL), haciendo uso de la herramienta Pentaho Data Integration 4.2.1. Se utilizó PostgreSQL 9.1 como Sistema Gestor de Bases de Datos (SGBD), administrado por pgAdminIII en su versión 1.14.0. Dichas herramientas son multiplataforma y cumplen con las políticas actuales de migración a software libre establecidas en la UCI. Una vez concluido el proceso de desarrollo y realizadas las pruebas unitarias y de integración, se obtiene como resultado el subsistema de integración que permite poblar el subsistema de almacenamiento.

Palabras claves

Centro de Inmunología Molecular (CIM), ensayo clínico, mercado de datos, procesos de integración de datos, subsistema de integración, subsistema de almacenamiento.

Introducción

Uno de los principales problemas del hombre a lo largo de la historia ha sido la lucha contra las enfermedades; este hecho lo ha impulsado a ser perseverante en su estudio, llevando a cabo diversas investigaciones acerca del tema. A pesar del desarrollo científico y los grandes avances tecnológicos que se poseen actualmente, aún existen enfermedades que causan la muerte a millones de personas y para las cuales no se tiene una cura definitiva. Dentro de este grupo se encuentra el cáncer, el cual es uno de los padecimientos que más afecta a la población mundial y constituye una de las principales causas de muerte; por esta razón cada día se intensifican las investigaciones en busca de la sobrevivencia de los pacientes afectados.

"El término cáncer es asociado a enfermedades en las que células anormales se dividen sin control y pueden invadir otros tejidos. Las células cancerosas pueden diseminarse a otras partes del cuerpo por el sistema sanguíneo y por el sistema linfático" (1). Existen más de 100 tipos de cáncer, que cuanto más agresivos y malignos se manifiestan, menos recuerdan a la estructura del tejido del que proceden.

Miles de especialistas de todo el mundo han hecho grandes esfuerzos en la búsqueda de una solución para esta enfermedad. Cuba no se ha quedado atrás y hoy cuenta con grandes avances en el desarrollo de vacunas contra el cáncer, SIDA, hepatitis y dengue, que no revierten dichas enfermedades pero mejoran la calidad de vida del paciente. El CIM es una de las instituciones que trabaja para alcanzar estos avances, fue inaugurado el 5 de diciembre de 1994, en medio del denominado "período especial" declarado a inicios de la pasada década de los 90. *"Su principal misión es obtener y producir nuevos biofármacos destinados al tratamiento del cáncer y otras enfermedades crónicas no transmisibles e introducirlos en la salud pública cubana. Sus investigaciones están concentradas en la inmunoterapia del cáncer, especialmente en el desarrollo de vacunas moleculares, ingeniería de anticuerpos, ingeniería celular, bioinformática y regulación de la respuesta inmune. Otro de sus objetivos se traduce en convertir el cáncer, que es una enfermedad mortal, en crónica".* (2)

El CIM posee laboratorios equipados para inmunoquímica, radioquímica, biología molecular y cultivo celular, además de instalaciones para la experimentación con modelos animales y una planta piloto que suministra los productos para ensayos clínicos. En este centro se prueban diferentes productos en grupos de personas que presentan la enfermedad y todas las pruebas son archivadas. (2)

Entre los ensayos clínicos que ha desarrollado el CIM se encuentran los del producto N Acetil GM3[1], el cual se obtuvo por la conjugación del antígeno Neu-acetil GM3 con un transportador el VSSP[2] y es administrado junto con un potenciador inespecífico de la respuesta inmune: el Montanide ISA 51. El producto es desarrollado para combatir el cáncer de mama.

La información asociada a cada ensayo clínico se encuentra almacenada en un sistema llamado *EpiData*, el cual arroja ficheros en varios formatos (Text, dBase III, Excel, Stata, SPSS y SAS). Algunos ensayos clínicos basados en la aplicación de los productos elaborados en el CIM son desarrollados durante varios años. La información arrojada por cada etapa es almacenada hasta que se realiza el cierre del ensayo. El volumen de datos se ha incrementado considerablemente en los últimos tiempos, debido al gran cúmulo de información que se genera en cada uno de los ensayos clínicos aplicados tanto fuera como dentro del país. Por tal motivo se torna complicado el proceso para el manejo de la información por parte de los directivos de la institución, teniendo en cuenta que los datos son gestionados manualmente, por lo que se presentan dificultades en el momento de realizar los reportes, análisis estadísticos complejos, consultas y el estudio de los indicadores relacionados con los ensayos clínicos.

La situación anteriormente descrita trae consigo el siguiente **problema a resolver**: ¿Cómo estandarizar los datos del producto N Acetil GM3 para su almacenamiento de forma homogénea?

Se define como **objeto de estudio** los almacenes de datos, enmarcado en el **campo de acción** subsistemas de almacenamiento e integración N Acetil GM3 para el almacén de datos de los ensayos clínicos del Centro de Inmunología Molecular.

El **objetivo general** de la investigación es desarrollar los subsistemas de almacenamiento e integración N Acetil GM3 para el almacén de datos de los Ensayos Clínicos del Centro de Inmunología Molecular, que permita el almacenamiento homogéneo de la información.

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

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

[1] **N Acetil GM3:** variante acetilada del gangliósido GM3.

[2] **VSSP:** proteínas de pequeño tamaño obtenidas de la membrana externa de una bacteria llamada Neisseria Meningitidis.

3. Realizar la implementación de los subsistemas de almacenamiento e integración N Acetil GM3.
4. Realizar pruebas a los subsistemas de almacenamiento e integración N Acetil GM3.

Para dar cumplimiento a los objetivos de la investigación y obtener una solución al problema planteado se trazan las siguientes **tareas de la investigación**:

1. Caracterización de las metodologías, herramientas y tecnologías a utilizar en el desarrollo de almacenes de datos.
2. Levantamiento de los requisitos.
3. Descripción de los casos de uso.
4. Definición de la arquitectura de la solución.
5. Diseño del subsistema de almacenamiento.
6. Diseño del modelo de datos.
7. Diseño del subsistema de integración.
8. Implementación del subsistema de almacenamiento.
9. Implementación del subsistema de integración.
10. Diseño de las listas de chequeo.
11. Diseño de los casos de prueba.
12. Aplicación de las listas de chequeo.
13. Aplicación de los casos de prueba.

El trabajo de diploma está estructurado en: Introducción, tres Capítulos, Conclusiones, Recomendaciones, Referencias bibliográficas, Bibliografía consultada, Anexos y Glosario de términos.

Capítulo 1: Fundamentos teóricos de los almacenes de datos.

El Capítulo 1 constituye el análisis teórico de los temas relacionados con la tecnología de los almacenes de datos. Se abordan los conceptos relacionados con el estado del arte de la investigación. Luego se expone y documenta la selección de la metodología, herramientas y tecnologías a utilizar en la implementación de la solución propuesta.

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

En el Capítulo 2 se realiza un análisis preliminar del negocio a partir del cual se especifican los requisitos de información que permiten dar cumplimiento a las necesidades del cliente, los cuales son agrupados en casos de uso de información. Una vez definidas las reglas del negocio, se identifican los hechos, dimensiones y medidas, elementos básicos para la obtención del modelo de datos. En el capítulo se estructura el diseño de los procesos de integración de datos.

Capítulo 3: Implementación y prueba de los subsistemas de almacenamiento e integración.

El Capítulo 3 está dirigido a la implementación y prueba de los subsistemas de almacenamiento e integración. Luego de elaborar el diseño de ambos subsistemas y asumiendo como pauta la metodología de desarrollo utilizada, se procede a la implementación. Además se exponen los resultados de la estrategia de pruebas aplicada, que permite avalar la disponibilidad de cada uno de los elementos incluidos en la solución.

Capítulo 1: Fundamentos teóricos de los Almacenes de Datos.

1.1. Introducción

El Capítulo 1 constituye el análisis teórico de los temas relacionados con la tecnología de los almacenes de datos. Se abordan los conceptos relacionados con el estado del arte de la investigación. Luego se expone y documenta la selección de la metodología, herramientas y tecnologías a utilizar en la implementación de la solución.

1.2. ¿Qué es un ensayo clínico?

"Un ensayo clínico es cualquier estudio de investigación que asigna de manera prospectiva participantes humanos o grupos de humanos a una o más intervenciones sanitarias a fin de evaluar los efectos en los resultados sanitarios" (3). Este término también hace referencia a un ensayo clínico de intervención. Las intervenciones incluyen, pero no se limitan a fármacos, células y otros productos biológicos, procedimientos quirúrgicos, procedimientos radiológicos, dispositivos, tratamientos conductuales, cambios en el proceso de atención y atención preventiva.

El CIM ha desarrollado hasta el momento tres vacunas que han sido incluidas en ensayos clínicos. Dos de ellas surgen a partir de la conjugación de dos gangliósidos[3] distintos a un transportador y asociándolo a una sustancia adyuvante potenciadora de la respuesta inmune. Uno de esos dos gangliósidos seleccionados es el N Acetil GM3, a partir del cual se llevan a cabo ensayos clínicos en la actualidad.

En base a sus propiedades inmunológicas, el desarrollo clínico de NAcGM3 / VSSP involucra tanto estudios en pacientes oncológicos como en sujetos infectados con el VIH. Varios ensayos clínicos con pacientes portadores de tumores positivos para N Acetil GM3 demostraron que el tratamiento con NAcGM3 / VSSP tiene una tolerabilidad aceptable. La aplicación del producto trajo consigo un gran volumen de información asociado a los procesos clínicos realizados.

1.3. Almacenes de datos

Durante la investigación se estudiaron varios conceptos de almacenes de datos, los cuales se relacionan a continuación:

[3] **Gangliósidos:** Lípido formado por dos ácidos grasos, una molécula de esfingosina, una cabeza de polisacáridos y uno o más grupos de ácido siálico.

Ralph Kimball, conocido autor en el tema de los almacenes de datos, define un almacén de datos como: *"una copia de las transacciones de datos específicamente estructurada para la consulta y el análisis"* (4). Además, define que un almacén de datos es: *"la unión de todos los data marts de una entidad"* (4).

Para Fernando Bocigas, ingeniero en telecomunicaciones y actual Director de Marketing de Grandes Cuentas y Partners en Microsoft Ibérica, un almacén de datos "*es una técnica que consolida y administra datos de diferentes fuentes para dar respuesta y servir de guía a distintos negocios, de una forma que no era posible hasta ahora. Consiste, básicamente, en la elaboración de un expediente empresarial más allá de la información transaccional y operacional; para después, almacenarlo en una base de datos que facilitará su posterior análisis y divulgación*" (5).

Bill Inmon fue uno de los primeros autores en escribir sobre el tema de los almacenes de datos. Este importante autor define un almacén de datos como *"una colección de datos orientada a temas, integrado, no volátil y variante en el tiempo que ayuda a la toma de decisiones de la empresa u organización"* (4). Después de un análisis de los conceptos se decide utilizar el de Inmon para realizar el trabajo de diploma. A continuación se describen las características que menciona Inmon en su definición de almacén de datos:

- **Orientado a temas:** los datos en la base de datos están organizados de manera que todos los elementos relativos al mismo evento u objeto del mundo real queden unidos entre sí.
- **Variante en el tiempo:** los cambios producidos en los datos a lo largo del tiempo quedan registrados para que los informes que se puedan generar reflejen esas variaciones.
- **No volátil:** la información no puede ser modificada ni eliminada por el usuario final, una vez almacenado un dato, éste se convierte en información de sólo lectura y se mantiene para futuras consultas.
- **Integrado:** la base de datos contiene los datos de todos los sistemas operacionales de la organización y dicha información debe ser consistente (4).

1.3.1. Mercado de datos

Kimball define un mercado de datos como *"una base de datos departamental, especializada en el almacenamiento de los datos de un área de negocio específica"* (4).

Los mercados de datos constituyen una pequeña porción del almacén de datos que contiene la información específica de un área del negocio de la organización, a fin de mantener consistencia de datos corporativos y mantener la seguridad e integridad de la información que se está usando.

Según Kimball, los mercados de datos están conectados con la arquitectura de los almacenes de datos en su forma más simple y que representan los datos de un sólo proceso del negocio a la vez. La mayor diferencia entre un mercado de datos y un almacén de datos está en el ámbito de la información que contienen debido a que en los mercados de datos son más pequeños y la información se obtiene de un menor número de fuentes y comúnmente el tiempo de desarrollo es menor.

1.3.2. Modelo de datos

"Un modelo de datos es un conjunto de conceptos que permiten describir los datos, las relaciones que existen entre ellos, la semántica y las restricciones de consistencia" (5).

De acuerdo con el Instituto Nacional Estadounidense de Estándares por sus siglas en ingles ANSI, un modelo de datos puede ser:

- **Conceptual:** especifica las expresiones permitidas por el modelo mismo, comunica las reglas y definiciones esenciales de los datos a los usuarios.
- **Lógico:** describe la semántica de tablas y columnas, clases orientadas a objetos y otros elementos representados por una tecnología de manipulación en particular (como es el lenguaje SQL).
- **Físico:** describe las estructuras de almacenamiento y métodos usados para tener acceso a los datos.

1.3.2.1. Modelo multidimensional

Un modelo dimensional constituye el diseño de las estructuras de almacenamiento de la información asociadas a la tecnología de almacenes de datos. Su uso facilita el acceso a los datos almacenados, los cuales se localizan en tablas de dimensiones y hechos. De esta manera es posible realizar análisis de la información mediante su visualización como cubos multidimensionales, donde las variables asociadas existen a lo largo de varios ejes o dimensiones y la intersección de las mismas representa la medida, indicador o el hecho que se está evaluando. (6)

Las medidas consisten en propiedades de un hecho (casi siempre numéricas), que son usadas para su análisis. Sin embargo, en algunos casos puede que los hechos no posean ninguna medida (6). En el caso

de que el hecho no posea las medidas se dice que el hecho es vacío y solo se usa para contar su ocurrencia en el tiempo.

Las dimensiones son un concepto esencial de las bases de datos multidimensionales. Constituyen perspectivas de análisis de la información utilizadas para seleccionar y agregar datos a un cierto nivel deseado de detalle y cada instancia o valor corresponde a un nivel particular. Constituyen características de un hecho que permiten su análisis orientado al proceso de toma de decisiones. Un hecho debe estar relacionado al menos con una dimensión: el tiempo. Las tablas de hechos contienen datos temporales, que son filtrados, agrupados y examinados mediante los contextos delimitados en las tablas de dimensiones. (6)

"Se le llama evento o hecho a una operación que se realiza en el negocio en un tiempo determinado. Son el objeto de análisis para la toma de decisiones y se representan en una caja con su nombre y las medidas que lo caracterizan" (6). Los hechos representan un patrón de interés o un evento dentro de una empresa u organización que necesita ser analizado. En la Figura 1 se muestra el ejemplo de un hecho llamado venta con sus medidas y las dimensiones que tiene asociada:

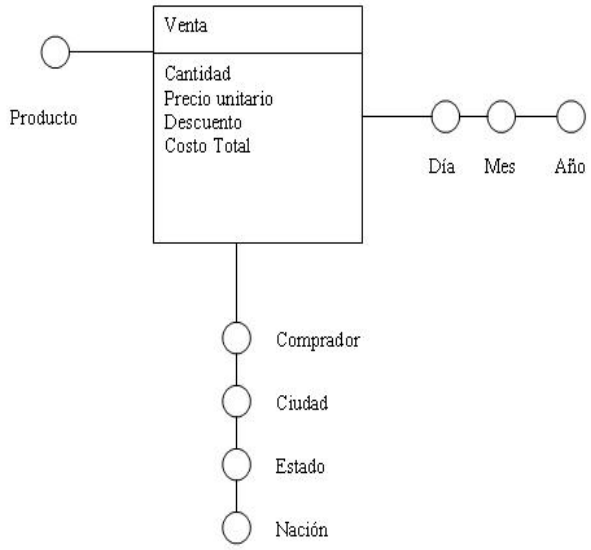


Figura 1: *Representación gráfica de un hecho y sus dimensiones*

1.3.3. Topologías de almacenes de datos

Una topología de almacenes de datos consiste en un conjunto de conceptos, reglas y convenciones que permiten describir y manipular los datos que se pretenden almacenar en una base de datos. (7) En el modelado de almacenes de datos existen tres tipos de topologías a utilizar para describir y estructurar la información, los cuales se mencionan a continuación:

1.3.3.1. Topología en estrella

La topología en estrella es un tipo de esquema de base de datos relacional que consta de una sola tabla de hechos central relacionada con varias tablas de dimensiones a través de sus respectivas llaves (7). El modelo no se encuentra normalizado, es decir, no se encuentra en tercera forma normal, lo que trae consigo la ventaja obviar uniones (join) entre tablas cuando se realizan consultas. De esta manera se genera un cierto grado de redundancia, sin embargo, el ahorro de espacio no es significativo en este tipo de modelos. En la Figura 2 se muestra un ejemplo de un esquema en estrella:

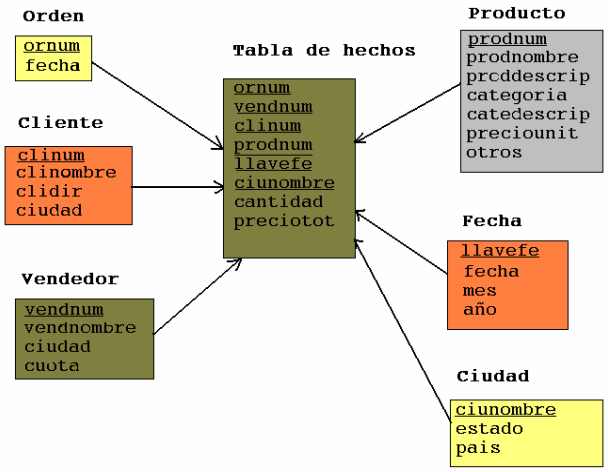


Figura 2: *Representación gráfica de un esquema en estrella*

Como se aprecia en la Figura 2, las tablas de dimensiones están relacionadas con las tablas de hechos, mediante relaciones de uno a mucho. La integridad referencial se garantiza mediante la creación de llaves foráneas en la tabla hecho, que a su vez forman parte de la llave principal de dicha tabla. Nótese, además, que un hecho puede tener asociadas una o varias dimensiones.

1.3.3.2. Esquema copo de nieve

El esquema copo de nieve *"es una variante de la topología en estrella que presenta las tablas de dimensión normalizadas"* (7). Al estar normalizadas se simplifican las operaciones de selección de datos, con lo que se logra presentar la información sin redundancia. Por ejemplo si una tabla de productos contiene 1000 tuplas, sería mejor crear una tabla con tipo de productos y mover los datos comunes para cada tipo de producto a esta tabla. El tamaño de estas dos tablas será menor que el de la tabla no normalizada que contenía los datos de los productos en general. Sin embargo este esquema presenta inconvenientes, al poseer múltiples dimensiones y cada una de ellas con jerarquías lo que provoca la existencia de múltiples uniones para enlazar la información requerida por lo que ralentiza el proceso. En la Figura 3 se muestra un ejemplo de este tipo de topología:

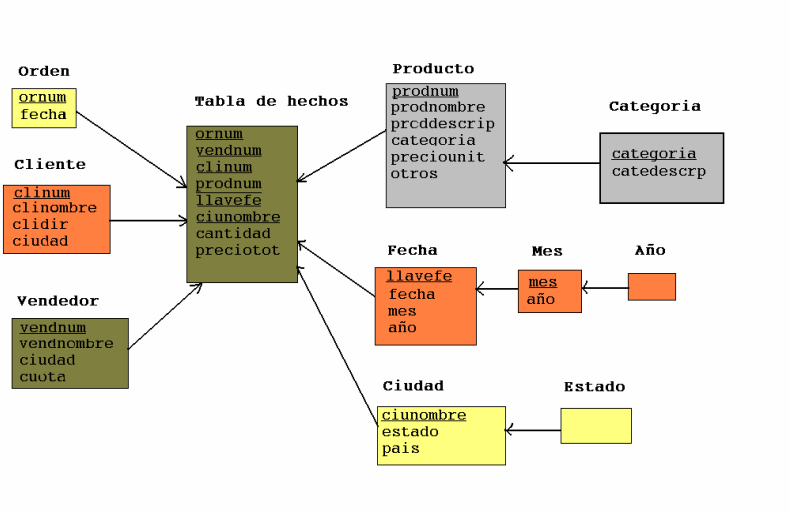


Figura 3: *Representación gráfica de un esquema copo de nieve.*

1.3.3.3. Esquema constelación de hechos

El esquema constelación de hechos son varios esquemas en estrella o copo de nieve que comparten algunas dimensiones (7). Es la más compleja de las topologías porque contiene varias tablas de hechos y contribuye a la reutilización de las tablas de dimensiones. De esta manera, una misma tabla de dimensión puede ser relacionada a varias tablas de hechos. La principal ventaja de esta topología radica en que al garantizar que las dimensiones sean compartidas, se logra un mejor uso del espacio de almacenamiento, evitando así la redundancia de la información. Además, teniendo en cuenta que permite la existencia de

varias tablas de hechos en el modelo, es posible analizar más aspectos del negocio con un menor esfuerzo adicional de diseño.

1.3.4. Procesos de integración de datos

La integración de datos puede ser enfocada de diferentes estrategias en dependencia, fundamentalmente, de los objetivos del sistema a desarrollar y las características de las fuentes de datos. A continuación se describen cuatro de las técnicas de integración existentes:

- **Replicación de datos:** *"es una técnica de integración que se basa en la creación y mantenimiento de múltiples copias de una misma base de datos"* (8). En la mayoría de las implementaciones de replicación, la copia primaria de la base de datos se mantiene en un servidor y servidores adicionales mantienen las copias esclavas de la misma.
- **Integración de Información Empresarial (EII):** *"es un mecanismo de transformación y acceso a datos transparente y optimizado para suministrar una única interfaz a lo largo de los datos de las organizaciones"* (8). La información es capturada en tiempo real lo que implica que las fuentes de datos tengan una estructura tecnológica sólida y bien establecida. EII protege a las aplicaciones de la complejidad de recuperar datos de múltiples localizaciones, donde los datos pueden diferir en semántica y formato y emplear diferentes interfaces de datos. Teniendo en cuenta estos aspectos, para la integración de datos a tiempo real la técnica EII constituye una buena alternativa, sin embargo no es factible para la integración de aplicaciones.
- **Integración de Aplicaciones Empresariales (EAI):** *"es el proceso de integrar múltiples aplicaciones desarrolladas independientemente, que utilizan tecnología incompatible y que son gestionadas de forma independiente, permitiendo que se comuniquen e intercambien transacciones de negocio, mensajes y datos entre sí"* (8). Uno de los principales objetivos de esta técnica es proporcionar acceso transparente a la amplia gama de aplicaciones que existen en una organización o empresa.
- **Extracción, Transformación y Carga de Datos (ETL):** *"como su nombre lo indica extrae información de un sistema fuente, transforma esos datos para satisfacer los requerimientos del negocio y carga el resultado en el sistema destino"* (8).

Se decide utilizar la técnica ETL en la investigación, atendiendo a que la información será extraída desde ficheros de tipo Excel, por lo que no es preciso trabajar desde un nivel alto de transparencia con

respecto al nivel de los datos de la fuente. La técnica seleccionada permite cumplir los objetivos propuestos mediante la extracción de los datos de las fuentes, su transformación en dependencia de las características del negocio y finalmente la carga en una base de datos. En la Figura 4 se representan dichos procesos:

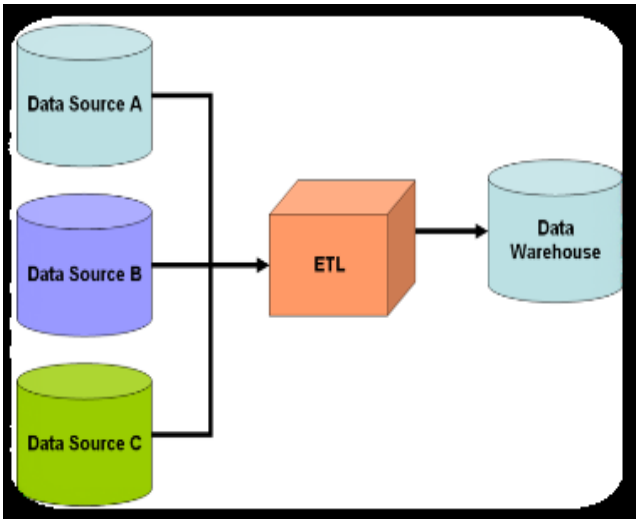


Figura 4: *Procesos de integración en la arquitectura de un almacén de datos*

- ✓ **Extraer:** proceso que consiste en extraer la información desde las fuentes de datos origen. Constituye el primer paso a desarrollar, el cual permite obtener los datos aunque procedan de diversas fuentes con formatos desiguales. Este proceso es el encargado de dejar listo los datos para la transformación.
- ✓ **Transformar:** en este proceso se modifica la información obtenida en el proceso de extracción. Tiene como entrada una o varias fuentes de datos origen y como salida una fuente de datos única. Requiere la aplicación de funcionalidades específicas a partir de las reglas de negocio y transformación definidas. Luego de extraída la información se convierte en un conjunto de datos homogéneos y congruentes, mediante la estandarización y limpieza de datos.
- ✓ **Cargar:** una vez realizados los procesos de extracción y transformación y asegurada la calidad de los datos, se procede a la carga de la estructura de datos al repositorio. En esta fase se interactúa directamente con la base de datos destino.

1.3.5. Dimensiones Lentamente Cambiantes (SCD[4])

Las dimensiones pueden ser estáticas o cambiantes. Las estáticas son aquellas cuya información no está propensa a cambios con el tiempo. A diferencia de las estáticas, las SCD son aquellas cuya información está propensa a cambiar según pasa el tiempo y que en dependencia de su tipo, permiten o no mantener en el almacén un registro histórico de los valores asociados a un identificador del sistema operacional. Inicialmente Ralph Kimball planteó tres estrategias a seguir cuando se tratan las SCD: tipo 1, tipo 2 y tipo 3; pero a través de los años la comunidad de personas que se encargaba de modelar bases de datos profundizó las definiciones iniciales e incluyó 3 tipos de SCD: tipo 0, tipo 4 y tipo 6. Para cada situación se puede aplicar una de las estrategias descritas a continuación:

- **Tipo 0 (no tiene en cuenta la gestión histórica):** no se realiza ningún esfuerzo para lidiar con los problemas del cambio de la dimensión. De esta manera se sobrescribe parte de la información, mientras que otra queda intacta.
- **Tipo 1 (sobrescribir):** se utiliza cuando la información histórica no es importante. Sobrescribe los datos antiguos con nuevos y es usado por lo general con el propósito de corregir errores en los datos almacenados en las tablas de dimensiones. Presenta como desventaja principal que no persiste ningún registro histórico en la dimensión.
- **Tipo 2 (añadir fila):** al ocurrir algún cambio en la dimensión se crea una nueva entrada en la tabla. Al nuevo registro es asignada una nueva llave subrogada, las cuales son de tipo serial y no representan información significativa del negocio, valor que será usado para futuras entradas, mientras que las antiguas utilizarán el valor anterior. De esta manera se gestionan versiones que pueden incluir fechas para indicar los períodos de validez, así como numeradores de registros o indicadores de registros activos o no.
- **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. Presenta como principal desventaja que solo permite guardar un historial limitado de los datos, dependiendo del número de columnas que se añadan.

[4] **SCD:** Slowly Changing Dimensions.

- **Tipo 4 (tabla de historia separada):** se almacenan en una tabla adicional los detalles de cambios históricos realizados a la tabla de dimensión. La tabla que contempla 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):** el método combina los tipos 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.

1.3.6. Metodología de desarrollo de almacenes de datos

Lograr la construcción de un sistema informático, que cumpla con los requisitos planteados, es una tarea realmente intensa y sobre todo difícil de cumplir. A la hora de abordar el tema del desarrollo de almacenes de datos se tienen en cuenta dos enfoques, el planteado por Inmon y el propuesto por Kimball. Inmon defiende un enfoque descendente (*top-down*) a la hora de diseñar un almacén de datos, puesto que de esta manera se consideraran mejor todos los datos corporativos mientras que Ralph Kimball defiende por tanto un enfoque ascendente (*bottom-up*) para el diseño de un almacén de datos.

En la investigación se seleccionó como metodología la Propuesta de Metodología para el Desarrollo de Almacenes de Datos en DATEC[5] que constituye una adaptación de la metodología Ciclo de vida de Kimball, la cual sigue un enfoque ascendente e introduce los conceptos de hechos y dimensiones que permiten el modelado multidimensional. Se complementa, además, con lo planteado por Leopoldo Zenaido Zepeda Sánchez de incluir los casos de uso para guiar el proceso de desarrollo y de esta manera estar alineados con las tendencias y normas de la UCI. Se introduce una etapa de pruebas que permite comprobar la calidad del producto.

1.4. Herramientas y tecnologías

A partir del análisis de las tendencias y tecnologías actuales posibles a emplear, considerando las características de la solución propuesta, se decide utilizar las siguientes herramientas y tecnologías en función de dar cumplimiento a los objetivos de la investigación:

[5] Centro de Tecnologías de Gestión de Datos.

1.4.1. Sistema Gestor de Base de Datos (SGBD)

PostgreSQL es un SGBD orientadas a objetos y usado en entornos de software libre. Se distribuye bajo licencia BSD[6], lo que permite su uso, redistribución y modificación con la única restricción de mantener el *copyright* del *software* a sus autores, el *PostgreSQL Global Development Group* y la Universidad de California. Funciona en múltiples plataformas como: Linux, Unix, Windows. Se destaca por su lista de prestaciones, dentro de las que se encuentran:

- Su administración se basa en usuarios y privilegios.
- Sus opciones de conectividad abarcan TCP/IP[7].
- Soporte para vistas, claves foráneas, integridad referencial, disparadores, procedimientos almacenados, subconsultas y casi todos los tipos y operadores soportados por *SQL*[8].(13)

1.4.2. Administrador de base de datos

La herramienta pgAdminIII *"es un programa de diseño y manejo de bases de datos para su uso con PostgreSQL. La aplicación se puede utilizar para manejar PostgreSQL 7.3 y superiores y funciona sobre casi todas las plataformas"* (14). Este *software* fue diseñado para responder a las necesidades de todos los usuarios, desde la escritura de simples consultas SQL a la elaboración de bases de datos complejas. La interfaz gráfica es compatible y facilita la administración. En el pgAdminIII se puede trabajar con casi todos los objetos de la base de datos, examinar sus propiedades y realizar tareas administrativas.

1.4.3. Modos de almacenamiento OLAP[9]

La tecnología OLAP facilita el análisis de datos en línea de un almacén de datos, proporcionando respuestas rápidas a consultas analíticas complejas. Es utilizada generalmente para contribuir en la toma de decisiones y presentar los datos a los usuarios a través de un modelo de datos intuitivo. Con este estilo

---

[6] **BSD:** Berkeley System Distribution

[7] **TCP/IP:** son las siglas de Protocolo de Control de Transmisión/Protocolo de Internet (en inglés *Transmission Control Protocol/Internet Protocol*), un sistema de protocolos que hacen posibles servicios Telnet, FTP, E-mail, y otros entre ordenadores que no pertenecen a la misma red.

[8] **lenguaje de consulta estructurado o SQL (por sus siglas en inglés Structured Query Language):** es un lenguaje declarativo de acceso a bases de datos relacionales que permite especificar diversos tipos de operaciones en ellas.

[9] **OLAP (Online Analytical Processing o procesamiento Analítico en Línea):** se trata de una forma de almacenar la información en una base de datos que permita realizar las consultas de forma más efectiva.

de presentación los usuarios finales pueden ver y comprender con mayor facilidad la información de sus bases de datos, lo que le permite a las organizaciones reconocer el valor de sus datos. Existen tres modos para el proceso analítico en línea de la información: ROLAP, MOLAP y HOLAP.

- ***MOLAP (Multidimensional OLAP):*** el objetivo de los sistemas MOLAP es almacenar físicamente los datos en estructuras multidimensionales de manera que la representación externa y la interna coincidan.
- ***ROLAP (Relational OLAP):*** este tipo de organización física se implementa sobre la tecnología relacional y dispone de algunas facilidades para mejorar el rendimiento.
- ***HOLAP (Hybrid OLAP):*** constituye un sistema híbrido entre MOLAP y ROLAP, que combina estas dos implementaciones para almacenar algunos datos en un motor relacional y otros en una base de datos multidimensional. (11)

A través de esta tecnología se analiza el negocio desde diferentes escenarios históricos y se proyecta cómo se ha venido comportando y evolucionando en un ambiente multidimensional, mediante la combinación de diferentes perspectivas, temas de interés o dimensiones. Esto permite deducir tendencias, por medio del descubrimiento de relaciones entre las perspectivas que a simple vista no se podrían encontrar sencillamente.

A pesar de que el sistema propuesto no incluye la etapa de visualización de los datos se tiene en cuenta la posibilidad futura de realizar análisis visual de la información histórica almacenada. Por tal motivo se selecciona el modo de almacenamiento ROLAP, teniendo en cuenta que el SGBD a utilizar es PostgreSQL, el cual no soporta el almacenamiento multidimensional, solamente el relacional. PostgreSQL es un SGBD multiplataforma de código abierto, a diferencia de los SGBD que dan soporte al almacenamiento multidimensional existentes en la actualidad.

1.4.4. Visual Paradigm 8.0

Visual Paradigm es una herramienta *CASE*[10] que utiliza UML[11] como lenguaje de modelado y propicia un conjunto de ayudas para el desarrollo de programas informáticos, desde la planificación, pasando por el análisis y el diseño, hasta la generación del código fuente de los programas y la documentación.

[10] **CASE:** (Ingeniería de Software Asistida por Computación)

[11] **UML:** Unified Modeling Language (Lenguaje de Modelado Unificado).

Esta herramienta ofrece:

- Entorno de creación de modelos conformes a UML.
- Diseño centrado en casos de uso y enfocado al negocio que generan un software de mayor calidad.
- Capacidades de ingeniería directa e inversa.
- Modelo y código que permanece sincronizado en todo el ciclo de desarrollo.
- Disponibilidad en múltiples plataformas.
- Extensible mediante desarrollo de nuevos *plug-ins*[12]. (12)

Visual Paradigm 8.0 para UML soporta todo el ciclo de vida del desarrollo de software. Permite modelar distintos tipos de diagramas, así como la generación de documentación asociada a cada etapa del proceso de desarrollo. Presenta compatibilidad entre ediciones, licencia gratuita y comercial. La UCI cuenta con la licencia para su uso.

1.4.5. Herramientas para los procesos de integración de datos

1.4.5.1. Data Cleaner 1.5.4

Antes de iniciar los procesos de ETL, es necesario realizar un análisis de los datos contenidos en las fuentes, con el propósito de conocer su estructura, formato y calidad. Dicho proceso es conocido como perfilado de datos y permite reconocer las características de la información, además de las necesidades de transformación de la misma.

Para ello son utilizadas algunas herramientas como *Microsoft Excel*, *Open Office*, así como consultas *SQL*. Se decide usar *Data Cleaner* 1.5.4 que resulta eficaz en el tratamiento de cadenas. Esta herramienta es una aplicación de código abierto para el perfilado, validación y comparación de los datos. Genera informes estadísticos que permiten determinar el nivel de calidad de los datos e identificar y analizar la distribución de los valores asociados a cada campo.

[12] Un ***plugin*** es un programa o aplicación que añade funcionalidad al programa principal donde está hospedado.

1.4.5.2. Pentaho Data Integration 4.2.1

El nombre *Kettle* viene de KDE (*Extraction, Transportation, Transformation and Loading Environment)*, puesto que originariamente la herramienta sería desarrollada para KDE, el famoso escritorio de Linux. El producto ha sido renombrado como *Pentaho Data Integration* (PDI).

El PDI está formado por un conjunto de cuatro herramientas cada una un propósito específico:

- **Spoon:** es la herramienta gráfica que permite el diseño de las transformaciones y trabajos. Incluye opciones para pre visualizar y probar los elementos desarrollados. Es la principal herramienta de trabajo de PDI y con la que construyen y validan los procesos ETL.
- **Pan:** es la herramienta que permite la ejecución de las transformaciones diseñadas en Spoon, ya sea desde un fichero o desde el repositorio. Permite desde la línea de comandos preparar la ejecución mediante scripts.
- **Kitchen:** similar a Pan, pero para ejecutar los trabajos o jobs.
- **Carte:** es un pequeño servidor web que permite la ejecución remota de transformaciones y jobs. (15)

Esta herramienta tiene como ventajas que es multiplataforma, está basada en metadatos, soporta múltiples gestores de base de dato entre ellos PostgreSQL que es el utilizado para realizar la investigación y posee una interfaz con indicadores de las transformaciones. Además, se basa en dos tipos de objetos: transformaciones (pasos en el proceso ETL) y trabajos (pasos para la ejecución de las transformaciones).

1.5. Conclusiones del capítulo

Una vez estudiados los fundamentos teóricos del objeto de estudio de la investigación se decidió utilizar la “Propuesta de Metodología para el Desarrollo de Almacenes de Datos en DATEC”, la que permite guiar el proceso de desarrollo de la solución a través de cada etapa del ciclo de vida. Como herramienta CASE se escogió el Visual Paradigm en su versión 8.0 pues posibilita la generación de los diagramas necesarios para modelar el funcionamiento del sistema propuesto. Se decide emplear el modo de almacenamiento ROLAP, teniendo en cuenta que el SGBD a utilizar es PostgreSQL en su versión 9.1 debido a las políticas de migración a software libre que se llevan en la universidad. Este es potencialmente adaptable al problema en cuestión, sin embargo no permite el almacenamiento multidimensional. Además, se decide utilizar el pgAdmin III en su versión 1.14.0 como administrador de base de datos. Las herramientas seleccionadas para el desarrollo de los procesos de ETL son:

- ✓ Para el proceso de perfilado de los datos el Data Cleaner en su versión 1.5.4 pues resulta eficaz en el tratamiento de cadenas y además genera sofisticados informes que permiten a los usuarios determinar el nivel de calidad de los datos, identificar y analizar la estructura del origen de datos.
- ✓ Para el desarrollo de los procesos de ETL fue seleccionada la herramienta Pentaho Data Integration en su versión 4.2.1, que permite su conducción a través de una interfaz gráfica. Además, posibilita la realización de transformaciones de datos en lotes desde y hacia un amplio rango de fuentes.

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

2.1. Introducción

En el Capítulo 2 se realiza un análisis preliminar del negocio a partir del cual se especifican los requisitos de información que permiten dar cumplimiento a las necesidades del cliente, los cuales son agrupados en casos de uso de información. Una vez definidas las reglas del negocio y los requisitos, se procede a descripción de los casos y su representación en el Diagrama de Caso de Uso del Sistema (DCUS) y para el diseño del subsistema de almacenamiento se identifican los hechos, dimensiones y medidas, elementos básicos para la obtención del modelo de datos. El capítulo se estructura el diseño de los procesos de integración de datos.

2.2. Análisis del negocio

En la construcción de un sistema informático, la etapa de análisis constituye el punto de partida para comprender los requisitos de la organización, a través de los cuales se definen las estructuras de almacenamiento, se diseñan las reglas de extracción, transformación y carga de los datos. Con el objetivo de lograr un diseño adaptable a las necesidades reales de los usuarios finales, se hace necesario realizar un estudio preliminar del negocio para identificar las necesidades de información de los especialistas del área.

En el CIM se recoge información de los ensayos clínicos de los productos que ahí se realizan. Entre estos ensayos se encuentra el del producto N Acetil GM3. Para identificar las necesidades de la organización, pueden llevarse a cabo diferentes técnicas, las que poseen características inherentes y específicas, como son las entrevistas, cuestionarios y observaciones. En el análisis del negocio de los subsistemas de almacenamiento e integración N Acetil GM3 para el CIM se escogieron las técnicas de entrevista con el cliente y observación. A partir de estas técnicas fueron identificados los principales requisitos de información para el cumplimiento de las necesidades del cliente y luego fueron definidas las reglas del negocio. Se diseñó el diagrama de casos de uso del sistema, así como los subsistemas de integración y almacenamiento. Además, el método de observación de la fuente de datos fue primordial en la realización del perfilado de datos. A partir del estudio del negocio realizado se decide clasificar la información en ocho grupos:

- Inclusión
- Evaluación inicial
- Examen físico
- Cumplimiento de las inmunizaciones
- Laboratorio clínico
- Imagenología
- Evaluación de las lesiones
- Interrupción del tratamiento
- Salida del ensayo

2.3. Reglas del negocio

Las reglas del negocio (RN) definen y controlan la estructura, funcionamiento y estrategia de una organización mediante medidas, políticas y restricciones de vital importancia. A continuación se definen las reglas del negocio identificadas:

RN1. La provincia se recoge en la fuente de datos como 1, 2, 3 ,4 y 5. Se determinó cambiar estos valores por 1: PR, 2: CH, 3: VC, 4: CM y 5: SC.

RN2. El sexo se recoge en la fuente de datos como 1 y 2. Se determinó cambiar estos valores por 1: M y 2: F.

RN3. La raza se recoge en la fuente de datos como 1, 2, 3 y 4. Se determinó cambiar estos valores por Blanca, Negra, Mestiza y Amarilla respectivamente.

RN4. El hospital se recoge en la fuente de datos como 1, 2, 3, 4, 5, 6, 7 y 8. Se determinó cambiar estos valores por 1: IIIC, 2: IN, 3: CQ, 4: HA, 5: SC, 6: MC, 7: AS y 8: HO.

RN5. El estadío se recoge en la fuente de datos como 1, 2, 3 y 4. Se determinó cambiar estos valores por Estadío I, Estadío II, Estadío III y Estadío IV.

RN6. Los receptores de estrógeno se recogen en la fuente de datos como 1, 2, 3. Se determinó cambiar estos valores por 1: Desconocido, 2: Positivo y 3: Negativo.

RN7. La evaluación de la respuesta se recoge en la fuente de datos como 1, 2, 3. Se determinó cambiar estos valores por 1: Remisión completa, 2: Remisión Parcial y 3: Estabilización.

RN8. El método diagnóstico se recoge en la fuente de datos como 1, 2, 3 y 4. Se determinó cambiar estos valores por 1: Rx, 2: TAC, 3: US y 4: Otro.

RN9. El método diagnóstico se recoge en la fuente de datos con los números del 1 al 8. Se determinó cambiar estos valores por 1: hígado, 2: pulmón, 3: ganglios linfáticos, 4: piel, 5: TCS, 6: huesos, 7: cerebro y 8: otro.

RN10. Las etapas se recoge en la fuente de datos con valor verdadero o falso Se determinó cambiar estos valores por el nombre de la etapa en que se encuentra el paciente ya sea etapa 1, etapa 2 o etapa 3.

RN11. La edad se recoge en la fuente de datos con valor entero. Se determinó cambiar estos valores por rangos de 10 hasta 100.

RN12. El peso se recoge en la fuente de datos con valor real. Se determinó cambiar estos valores por rangos de 10 kg hasta 120.

RN13. La respuesta lesiones dianas se recoge con valores enteros del 1 al 4. Se determinó cambiar estos valores por 1: Respuesta completa, 2: Respuesta parcial, 3: Enfermedad estable y 4: Progresión.

RN14. La respuesta lesiones no dianas se recoge con valores enteros del 1 al 4. Se determinó cambiar estos valores por 1: Respuesta completa, 2: Respuesta parcial, 3: Enfermedad estable y 4: Progresión.

RN15. La respuesta global se recoge con valores enteros del 1 al 3. Se determinó cambiar estos valores por 1: Respuesta completa, 2: Respuesta imparcial, 3: Progresión.

Las reglas de transformación que se definieron son:

RN16. Los valores nulos que se encuentran en los campos de tipo entero se decidió sustituirlos por (-1).

RN17. Los valores nulos que se encuentran en los campos de tipo date se decidió sustituirlos por una fecha fuera de rango de la aplicación del producto por eso se escogió 01/01/1999.

RN18. Los valores nulos que se encuentran en los campos de tipo cadena se decidió sustituirlos por ND (no disponible).

RN19. Los valores nulos del campo fecha_salida, del modelo salida del ensayo, se decidió cambiarlos por la fecha de fallecimiento.

RN20. Las fechas se encontrarán entre el 3/11/2000 y el 25/12/2007.

RN21. La cantidad de pacientes a los que se les aplico el ensayo es de 28, de los cuales sobrevivieron 27.

RN22. La cantidad de pacientes de raza blanca supera a los de raza negra en 4.

RN23. La edad de los pacientes incluidos en el ensayo está en los 35 y los 74 años.

RN24. El peso de los pacientes incluidos en el ensayo se encuentra entre 37 y 104 kg.

2.4. Especificación de requisitos

El análisis de requisitos constituye una de las fases más importantes en la construcción de un mercado de datos. Su objetivo general es guiar el proceso de desarrollo hacia el sistema correcto. En dicha fase se definen los requisitos de información, funcionales y no funcionales del sistema, partiendo de las necesidades del cliente. Se trabaja en conjunto con los futuros usuarios en función de que las especificaciones de la aplicación sean descritas por quien tiene una visión más precisa de los procesos que se llevan a cabo en la organización.

2.4.1. Requisitos de información

Los requisitos de información (RI) representan toda la información que debe estar disponible en el sistema para satisfacer las necesidades de los clientes. A continuación se especifican los identificados durante el proceso clasificados según el tema de análisis de Ensayo Clínico N Acetil GM3 Cáncer de mama. Los requisitos de información identificados garantizan que se encuentre disponible la información de manera que sea posible:

RI_1: Obtener la cantidad de pacientes por provincia, raza, etapa, edad, hospital, esquema tratamiento, tnm, estadío, receptores de estrógeno, método diagnóstico, exámenes físicos, exámenes de laboratorios, enfermedades y tratamientos previos en los modelos de inclusión y evaluación inicial de la aplicación del producto N Acetil GM3 en los pacientes con cáncer de mama metastásico.

RI_2: Obtener la cantidad de número de ciclos por esquema de tratamiento en los modelos de inclusión y evaluación inicial de la aplicación del producto N Acetil GM3 en los pacientes con cáncer de mama metastásico.

RI_3: Obtener la cantidad de pacientes con carcinoma de mama estadio IV o enfermedad metástasis evolutiva en los modelos de inclusión y evaluación de la aplicación del producto N Acetil GM3 en los pacientes con cáncer de mama metastásico.

RI_4: Obtener la cantidad de pacientes embarazadas o en periodo de lactancia en los modelos de inclusión y evaluación de la aplicación del producto N Acetil GM3 en los pacientes con cáncer de mama metastásico.

RI_5: Obtener la cantidad de pacientes con consentimiento informado para participar en la investigación en los modelos de inclusión y evaluación de la aplicación del producto N Acetil GM3 en los pacientes con cáncer de mama metastásico.

RI_6: Obtener la cantidad de pacientes con evaluación de la capacidad funcional de grado 0 a 2 en los modelos de inclusión y evaluación de la aplicación del producto N Acetil GM3 en los pacientes con cáncer de mama metastásico.

RI_7: Obtener la cantidad de pacientes con metástasis contra lateral del producto N Acetil GM3 en los pacientes con cáncer de mama metastásico en los modelos de inclusión y evaluación inicial de la aplicación del producto N Acetil GM3 en los pacientes con cáncer de mama metastásico.

RI_8: Obtener la cantidad de pacientes con metástasis cerebral del producto N Acetil GM3 en los pacientes con cáncer de mama metastásico en los modelos de inclusión y evaluación inicial de la aplicación del producto N acetil GM3 en los pacientes con cáncer de mama metastásico.

RI_9: Obtener la cantidad de pacientes con enfermedades malignas del producto N Acetil GM3 en los pacientes con cáncer de mama metastásico en los modelos de inclusión y evaluación inicial de la aplicación del producto N acetil GM3 en los pacientes con cáncer de mama metastásico.

RI_10: Obtener la cantidad de pacientes con parámetros de laboratorio dentro de los límites establecidos del producto N Acetil GM3 en los pacientes con cáncer de mama metastásico en los modelos de inclusión y evaluación inicial de la aplicación del producto N Acetil GM3 en los pacientes con cáncer de mama metastásico.

RI_11: Obtener la cantidad de pacientes que hayan recibido terapéuticas oncoespecíficas de primera línea para su enfermedad metastásica y hayan transcurrido 6 meses de aplicado el producto en los modelos de

inclusión y evaluación inicial de la aplicación del producto N Acetil GM3 en los pacientes con cáncer de mama metastásico.

RI_12: Obtener la cantidad de pacientes con edades entre 18 y 80 en los modelos de inclusión y evaluación inicial de la aplicación del producto N Acetil GM3 en los pacientes con cáncer de mama metastásico.

RI_13: Obtener la cantidad de pacientes con parámetros de laboratorio entre los límites establecidos en los modelos de inclusión y evaluación inicial de la aplicación del producto N acetil GM3 en los pacientes con cáncer de mama metastásico.

RI_14: Obtener la cantidad de pacientes con citología en los modelos de inclusión y evaluación inicial de la aplicación del producto N Acetil GM3 en los pacientes con cáncer de mama metastásico.

RI_15: Obtener la cantidad de pacientes con biopsia en los modelos de inclusión y evaluación inicial de la aplicación del producto N Acetil GM3 en los pacientes con cáncer de mama metastásico.

RI_16: Obtener la cantidad de pacientes por provincia, etapa, peso, exámenes físicos, estado funcional según la OMS[13] y hospital en el modelo de examen físico de la aplicación del producto N Acetil GM3 en los pacientes con cáncer de mama metastásico.

RI_17: Obtener la cantidad de pacientes por provincia, hospital, modificación del tratamiento, retraso de la administración, tratamiento médico, esquema de tratamiento y signos vitales en el modelo de cumplimiento de la inmunización de la aplicación del producto N Acetil GM3 en los pacientes con cáncer de mama metastásico.

RI_18: Obtener la cantidad de pacientes por provincia, hospital, etapa, grupo de tratamiento y examen de laboratorio en el modelo de laboratorio clínico de la aplicación del producto N Acetil GM3 en los pacientes con cáncer de mama metastásico.

RI_19: Obtener la cantidad de pacientes por provincia, hospital, etapa, grupo de tratamiento y respuestas en los modelos de imagenología y evaluación de las lesiones de la aplicación del producto N Acetil GM3 en los pacientes con cáncer de mama metastásico.

[13] **OMS:** Organización Mundial de la Salud.

RI_20: Obtener la cantidad de pacientes por provincia, hospital, causas del fallecimiento y causas de la interrupción en el modelo de interrupción del tratamiento de la aplicación del producto N Acetil GM3 en los pacientes con cáncer de mama metastásico.

RI_21: Obtener la cantidad de pacientes por provincia, hospital, causas del fallecimiento y causas de la salida en el modelo de salida del ensayo de la aplicación del producto N Acetil GM3 en los pacientes con cáncer de mama metastásico.

2.4.2. Requisitos funcionales

Los requisitos funcionales (RF) definen las capacidades o condiciones que el sistema debe cumplir. Permiten expresar específicamente las responsabilidades del sistema que se propone con la intención de lograr la satisfacción plena del cliente. A continuación se mencionan:

RF_1: Extraer los datos de la fuente.

RF_2: Transformar y cargar los datos de la fuente.

2.4.3. Requisitos no funcionales

Los requisitos no funcionales (RNF), son requisitos que imponen restricciones en el diseño o la implementación como estándares de calidad. Son propiedades o cualidades que el producto debe tener.

Confiabilidad

RNF_1: Asegurar la disponibilidad del sistema: el sistema debe estar disponible 10 horas, 6 días de la semana.

RNF_2: Asegurar la recuperación ante un fallo: el sistema debe ser capaz de recuperarse ante un fallo, teniendo en cuenta la complejidad y naturaleza de éste. El tiempo para su correcta recuperación debe ser entre 24 y 48 horas. Este tiempo comprende la solución al problema, así como su validación y prueba.

Fiabilidad

RNF_3: Garantizar la persistencia de la información: para garantizar la persistencia de la información se realizará una copia de respaldo de ella, teniendo en cuenta que lo que se realiza en la solución es una carga histórica. Toda esta información se almacenará en un área segura.

Restricciones de diseño

RNF_4: Utilizar los lenguajes de programación definidos durante la investigación: como lenguaje dentro del SGBD se utilizará PL/pgSQL. En la implementación de los procesos de integración de datos se utilizará el lenguaje Java Script.

RNF_5: Utilizar el Sistema Gestor de Base de Datos definido durante la investigación: el gestor de base de datos que se utilizará es PostgreSQL y como interfaz de administración de dicho gestor, el PgAdminIII.

RNF_6: Utilizar la herramienta de integración de datos definida durante la investigación: para los procesos de integración de los datos se usará la herramienta PDI.

2.5. Casos de uso del sistema

La fase análisis y diseño incluye la definición de los casos de uso del sistema (CUS). Los casos de usos constituyen la interacción entre los actores del sistema y sus respectivas actividades, en respuesta a un evento que inicia un actor sobre el propio sistema. Los casos de usos del sistema se dividen en casos de usos de información y casos de usos funcionales. Los requisitos de información y los requisitos funcionales son agrupados en casos de uso de información y casos de uso funcionales, respectivamente.

2.5.1. Actores del sistema

Actores	Objetivo
Administrador ETL	Es el encargado de los procesos de ETL de los datos.
Analista	Es el responsable de mantener disponible la información de los diferentes modelos.

Tabla 1: *Actores del sistema*

2.5.2. Casos de uso de información

Los casos de uso de información (CUI) representan el criterio por el cual se agrupan los requisitos de información. A continuación se describe cada CUI definidos en la solución:

1. **CUI_1. Mantener disponible la información de los modelos de inclusión y evaluación inicial:** Mantiene disponible la información de los modelos de inclusión y evaluación inicial.

2. **CUI_2. Mantener disponible la información del modelo de examen físico:** Mantiene disponible la información del modelo de examen físico.
3. **CUI_3. Mantener disponible la información del modelo de cumplimiento de las inmunizaciones:** Mantiene disponible la información del modelo de cumplimiento de las inmunizaciones.
4. **CUI_4. Mantener disponible la información del modelo de laboratorio clínico:** Mantiene disponible la información del modelo de laboratorio clínico.
5. **CUI_5. Mantener disponible la información de los modelos de imagenologia y evaluación de las lesiones:** Mantiene disponible la información de los modelos de imagenologia y evaluación de las lesiones.
6. **CUI_6. Mantener disponible la información del modelo de interrupción del tratamiento:** Mantiene disponible la información del modelo de interrupción del tratamiento.
7. **CUI_7. Mantener disponible la información del modelo de salida del ensayo:** Mantiene disponible la información del modelo de salida del ensayo.

2.5.3. Casos de uso funcionales

Los casos de uso funcionales (CUF) representan el criterio por el cual se agrupan los requisitos funcionales. A continuación se describe cada CUF definido en la solución:

1. **CUF_1.Extraer datos de la fuente:** se extraen los datos de la fuente para su posterior transformación y carga.
2. **CUF_2.Transformar y cargar los datos de la fuente:** se transforman y cargan los datos extraídos.

A continuación se presenta la descripción textual del CUF_1. Extraer datos de la fuente (ver Tabla 2) y del CUF_2. Transformar y cargar los datos de la fuente (ver Tabla 3).

Objetivo	Extraer datos de la fuente
Actores	Administrador de ETL
Resumen	El caso de uso (CU) inicia cuando el actor desea realizar la extracción de los datos correspondientes a las fuentes de datos. Se extraen los datos de la fuente y el CU finaliza una vez que los datos

	seleccionados por el actor son extraídos.
Complejidad	Media
Prioridad	Crítica
Precondiciones	Disponibilidad de las fuentes
Poscondiciones	Los datos seleccionados de las fuentes de datos quedan extraídos y disponibles para transformar.

Flujo de eventos		
Flujo básico Extraer datos de la fuente		
	Actor	**Sistema**
1	Ejecuta la transformación.	
2		Realiza la conexión a la fuente de información correspondiente.
3		Verifica el control de las extracciones.
4		Si no se extrajeron esos datos, procede a realizar la extracción. Finaliza el CU.
Prototipo de interfaz		
Flujos alternos		
2ª. No responde a la solicitud de conexión.		
	Actor	**Sistema**
		Notifica el error al Administrador de ETL a través de un mensaje. Vuelve al paso 1 del flujo normal.
5ª. Se extrajeron los datos.		
		Aborta la ejecución del proceso. Finaliza el CU.

Relaciones	**CU Incluidos**	No aplica.
	CU	No aplica.

	Extendidos	
Requisitos no funcionales	Sección: “3.2 Requisitos no funcionales” del documento: “0113_ERS”.	
Asuntos pendientes	[Posibles mejoras al CU	

Tabla 2: *Descripción del CUF_1. Extraer datos de la fuente.*

Objetivo	Transformar y cargar los datos de la fuente	
Actores	El CU inicia cuando el actor desea realizar la transformación y carga de los datos correspondientes a las fuentes de información ya extraídas. Se transforman y cargan los datos extraídos y el CU finaliza una vez que los datos han sido transformados y cargados en la base de dato.	
Resumen	Media	
Complejidad		
Prioridad	Crítica	
Precondiciones	Extracción de los datos completada.	
Poscondiciones	Los datos son cargados en la base de dato.	
Flujo de eventos		
Flujo básico Transformar y cargar los datos de la fuente		
	Actor	**Sistema**
1	Ejecuta la transformación	
2		Carga los datos extraídos de la fuente.
3		Aplica las transformaciones diseñadas a partir de las reglas del negocio.
4		Carga los datos en la base de datos. Finaliza el CU.
Flujos alternos		
4ª. No responde a la solicitud de conexión		
	Actor	**Sistema**

		Notifica el error al Administrador de ETL a través de un mensaje. Finaliza el CU.
Relaciones	**CU Incluidos**	No aplica.
	CU Extendidos	No aplica.
Requisitos no funcionales	Sección: "3.2 Requisitos no funcionales" del documento: "0113_ERS".	
Asuntos pendientes	Posibles mejoras al CU.	

Tabla 3: *Descripción del CUF_2.Transformar y cargar los datos de la fuente.*

2.6. Diagrama de casos de uso

Con el objetivo de especificar la comunicación, el comportamiento de un sistema mediante su interacción con los usuarios y las actividades y entender mejor el funcionamiento del sistema se realiza una representación gráfica con la relación entre los actores y sus casos de uso. En la Figura 5: Representación del DCUS. se muestra el DCUS donde se evidencia las relaciones entre los actores y sus respectivos casos de uso:

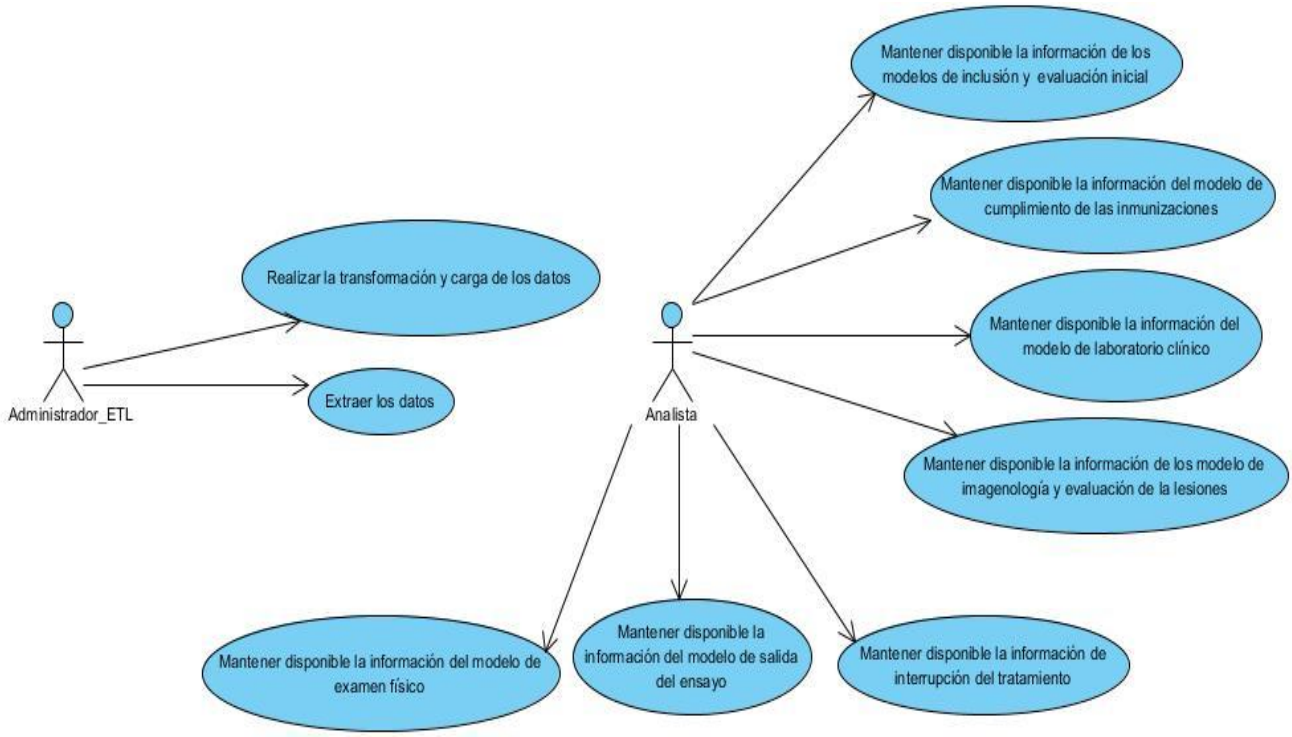


Figura 5: *Representación del DCUS.*

En la Figura 5 se muestra el diagrama de CU de la investigación compuesto por dos actores que inicializan los casos de uso que le corresponden. El administrador de ETL es el encargado de extraer los datos de la fuente y de realizar la transformación y carga de los datos de la fuente, los que constituyen CUF. El analista es el encargado de mantener disponible la información de los modelos. Los CU inicializados por el analista constituyen los CUI por lo que se agruparán los requisitos de información y estos a su vez se agruparán en el tema de análisis "Ensayo Clínico 048 para el tratamiento de paciente con cáncer de mama metastásico". Este diagrama no presenta ningún patrón de CU.

2.7. Definición de la arquitectura base de los subsistemas de almacenamiento e integración N Acetil GM3.

Al abordar el tema de la arquitectura de los almacenes de datos se parte del concepto de arquitectura de software que *"es un conjunto de patrones que proporcionan un marco de referencia necesario para guiar la construcción de un software, permitiendo a los programadores, analistas y todo el conjunto de desarrolladores del software compartir una misma línea de trabajo y cubrir todos los objetivos y*

restricciones de la aplicación. Es considerada el nivel más alto en el diseño de la arquitectura de un sistema puesto que establece la estructura, funcionamiento e interacción entre las partes del software" (7).

En el momento de abordar la arquitectura de un mercado de datos se debe tener en cuenta la forma de representar el origen y estructura global de los datos, la comunicación, los procesos y la presentación al usuario final. La arquitectura lógica de este tipo de sistemas consta de cuatro niveles (fuente de datos, subsistema de integración, subsistema de almacenamiento y subsistema de visualización) de los cuales el trabajo de diploma incluye tres solamente, la fuente de datos, el subsistema de integración y el de almacenamiento, los cuales se describen en la Figura 6

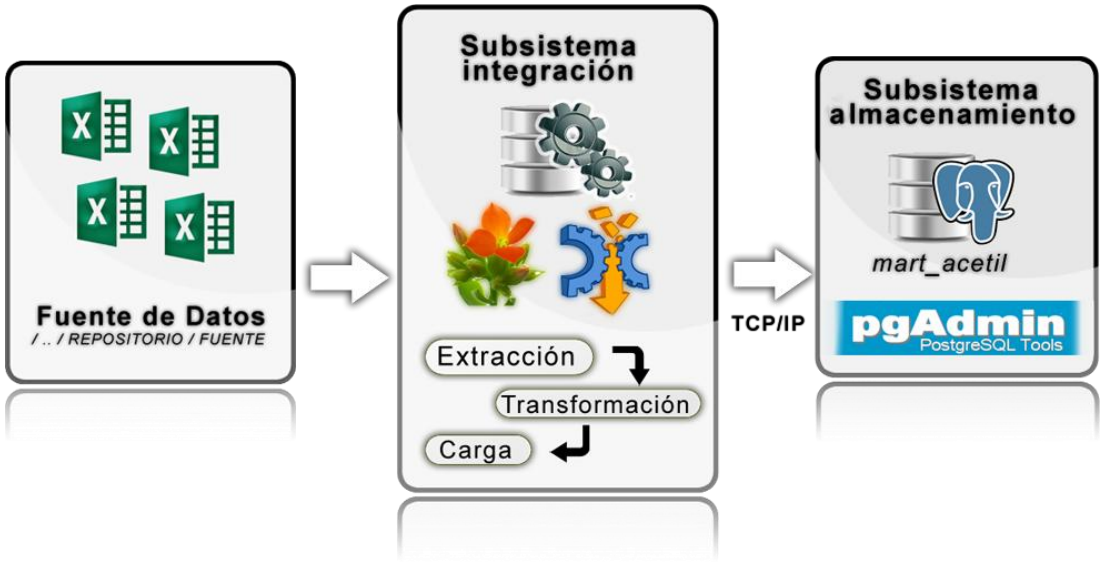


Figura 6: *Relación entre subsistema de almacenamiento e integración.*

La fuente de datos está compuesta por ficheros Excel los que son extraídos desde el subsistema de integración a través del PDI, el cual se encarga de realizar los procesos que integran y transforman la información para posteriormente almacenarla en la base de dato final. Por su parte, el subsistema de almacenamiento recibe la información manipulada durante los procesos de extracción, transformación y carga. Los datos se almacenan en una base de datos llamada mart_acetil soportada por el SGBD PostgreSQL y administrada solo por los usuarios autorizados mediante la herramienta PgAdminIII.

2.8. Diseño de la solución

La solución consiste en desarrollar los subsistemas de integración y almacenamiento N Acetil para el almacén de datos del CIM. Para el diseño de la solución se procede al diseño de los subsistemas de

integración y almacenamiento, realizar el perfilado de los datos, así como a establecer los esquemas de seguridad y las políticas de seguridad.

2.8.1. Diseño del subsistema de almacenamiento

A la hora de diseñar el subsistema de almacenamiento se deben tener en cuenta aspectos como la identificación de los hechos, dimensión y medidas para la realización del modelo de datos usando la topología definida. A continuación en los siguientes epígrafes se procede al desarrollo de los anteriores elementos:

2.8.1.1. Dimensiones

En el subsistema de almacenamiento se identifican las dimensiones que se incluirán en la solución. Se identificaron 41 dimensiones, de ellas 15 son degeneradas, las cuales son dimensiones que toman pocos valores y estos no cambian en el tiempo. Ejemplo de estas son los números de asiento de un teatro, los números de ticket y en este caso son preguntas que toman valores verdadero o falso. Este tipo de dimensiones se usa para para reducir la duplicación de datos y la simplificación las consultas. Estos campos se podrían incluir en la tabla de dimensiones pero en este caso ese estaría manteniendo una fila de esta dimensión por cada fila en la tabla de hechos, por tanto se obtendría la duplicación de información y complejidad, que precisamente se pretende evitar.

A continuación se describen las dimensiones incluidas en la solución:

1. **Dimensión raza (dim_raza):** describe el sexo del paciente.
2. **Dimensión edad (dim_edad):** describe la edad del paciente.
3. **Dimensión dpa (dim_dpa):** especifica la provincia donde se encuentra el hospital en que se aplicó el ensayo al paciente.
4. **Dimensión hospital (dim_hospital):** especifica el hospital donde se le aplicó el ensayo.
5. **Dimensión grupo de tratamiento (dim_grupo_tratamiento):** describe el grupo de tratamiento al que pertenece el paciente.
6. **Dimensión estadío (dim_estadio):** especifica la gravedad de la enfermedad del paciente al diagnosticar la enfermedad.

7. **Dimensión tnm (dim_tnm):** especifica el tamaño del tumor, número de ganglios y si el paciente tiene metástasis o no (ver Anexo 1).
8. **Dimensión tiempo (dim_tiempo):** describe todos los espacios de tiempo en que se aplicó el ensayo (fechas, tiempo que demora).
9. **Dimensión receptores de estrógeno (dim_receptores_estrogeno):** Especifica la cantidad de receptores de estrógeno.
10. **Dimensión exámenes de laboratorio (dim_examen_lab):** especifica los resultados de los exámenes de laboratorio.
11. **Dimensión esquema de tratamiento (dim_esquema_tratamiento):** Se refiere al esquema de tratamiento utilizado.
12. **Dimensión método diagnóstico (dim_metodo_diagnostico):** Se refiere al método de diagnóstico utilizado en el ensayo ya sea rayos x, ultrasonido, TAC o cualquier otro.
13. **Dimensión peso (dim_peso):** Se refiere al peso corporal del paciente.
14. **Dimensión estado funcional según la Organización Mundial de la salud (OMS) (dim_estado_oms):** se refiere a estado funcional según la OMS del paciente (ver Anexo 2)**.**
15. **Dimensión etapa (dim_etapa):** Se refiere a la etapa de la enfermedad en que se encuentra el paciente.
16. **Dimensión tratamiento asociado (dim_tratamiento):** Se refiere al tratamiento o los tratamientos aplicados al paciente.
17. **Dimensión respuesta (dim_respuesta):** Se refiere a las respuestas a los tratamientos aplicados al paciente.
18. **Dimensión exámenes físicos (dim_exam_fis):** Describe los exámenes físicos que se le realizan al paciente.
19. **Dimensión causas de la salida (dim_causas_salida):** Describe las causas de salida del ensayo.
20. **Dimensión causas del fallecimiento (dim_causas_fallecimiento):** Describe las causas de fallecimiento del paciente.

21. **Dimensión signos vitales (dim_signos_vitales):** Describe el comportamiento de los signos vitales del paciente por inmunización.
22. **Dimensión retraso de la administración (dim_restraso_admin):** se refiere a las causas del retraso de la administración.
23. **Dimensión tratamientos previos (dim_trat_previos):** se refiere a los tratamientos previos que se le habían aplicado al paciente ya sea cirugía u otro cualquiera.
24. **Dimensión modificación del tratamiento (dim_mod_trat):** describe el proceso de modificación de los tratamientos que se le aplican a un paciente.
25. **Dimensión causa de la interrupción (dim_causa_interrup):** describe la causa de interrupción por las que el paciente abandonó el tratamiento.
26. **Dimensión enfermedades (dim_enfermedades):** describe las enfermedades que presentaba el paciente antes de aplicarle el producto.

Dimensiones degeneradas:

27. **Dimensión paciente con carcinoma o metástasis estadio IV (pac_carcinoma_metastasis):** especifica si el paciente tiene carcinoma o metástasis con estadio IV.
28. **Dimensión paciente con consentimiento informado (pac_consent_inf):** especifica si el paciente ha sido informado para participar en la investigación.
29. **Dimensión paciente con edad entre 18 y 80 (pac_entre_18_y_80):** especifica si el paciente tiene edad entre 18 y 80.
30. **Dimensión paciente parámetros entre límites establecidos (param_entre_lim_est):** Especifica si el paciente tiene los parámetros entre los límites establecidos.
31. **Dimensión paciente enfermedad presenta enfermedad no establecidas en el protocolo (param_entre_lim_est):** Especifica si el paciente presenta carcinoma o metástasis estadio IV no establecidas en el protocolo.
32. **Dimensión paciente con terapia antes de entrar al ensayo (pac_con_terap):** Especifica si el paciente recibió alguna terapia con anterioridad.

33. **Dimensión paciente con embarazadas o en periodo de lactancia (embaraz_o_lact):** Especifica si el paciente está embarazada o en periodo de lactancia.
34. **Dimensión paciente con enfermedades malignas (pac_enferm_malig):** especifica si el paciente presenta alguna enfermedad maligna.
35. **Dimensión paciente metástasis cerebral (pac_metast_cerebral):** Especifica si el paciente presenta metástasis cerebral.
36. **Dimensión paciente con enfermedades crónicas o agudas (pac_enfer_cron_agud):** describe si el paciente presenta enfermedades crónicas o enfermedades crónicas agudas.
37. **Dimensión paciente con enfermedades autoinmunes o enfermedades crónicas descompensadas (enferm_autoinm_cron_desc):** se refiere a si el paciente presenta enfermedades autoinmumnes o crónicas descompensadas.
38. **Dimensión paciente estadíos alérgicos o reacciones alérgicas severas (alerg_reacc_adv):** especifica si el paciente presenta estadíos agudos alérgico o presentó alguna historia de reacciones alérgicas severas.
39. **Dimensión paciente con mama contralateral (mama_contralateral):** especifica si el paciente tiene o no metástasis de mama contralateral.
40. **Dimensión biopsia (biopsia):** especifica si se le realizó biopsia.
41. **Dimensión citología (citología):** especifica si se le realizó citología.

2.8.1.2. Hechos y medidas

Las tablas de hechos diseñadas almacenan en algunos casi todos los casos, el código del paciente, que se utilizará para contar los pacientes y además cuenta con las llaves primarias de las dimensiones. Existen otros casos que cuentan con dimensiones degeneradas y otros tienen una medida cantidad de pacientes. Para el desarrollo de la solución se identificaron los siguientes siete hechos:

1. **Modelo de inclusión, evaluación inicial:** En este hecho se almacena la información de los modelos inclusión y evaluación inicial Tiene como atributos propios el código del paciente a través del cual se cuentan los pacientes y 16 atributos que representan dimensiones degeneradas con valor booleano.

2. **Modelo examen físico:** En este hecho se almacena la información del modelo examen físico el cual tiene como único atributo propio el código del paciente a través del que se cuenta los pacientes.
3. **Cumplimiento de las inmunizaciones:** En este hecho se almacena la información del modelo de cumplimiento de las inmunizaciones el cual tiene como único atributo propio el código del paciente a través del que se cuentan los pacientes.
4. **Laboratorio clínico:** En este hecho se almacena la información del modelo de laboratorio clínico el cual tiene como único atributo propio el código del paciente a través del cual se cuentan los pacientes.
5. **Evaluación de las lesiones e Imagenología:** En este hecho se almacena la información del modelo de evaluación de las lesiones y de imagenología. Tiene como único atributo propio el código del paciente a través del que se cuentan los pacientes.
6. **Interrupción del tratamiento:** En este hecho se almacena la información del modelo de interrupción del tratamiento cuya medida es cantidad de pacientes.
7. **Modelo de salida:** En este hecho se almacena la información del modelo de salida del ensayo cuya medida es cantidad de pacientes.

2.8.1.3. Matriz bus

La matriz bus es esencialmente la arquitectura dimensional de los datos de la organización para cada proceso del negocio identificado. Especifica la relación entre los hechos y las dimensiones, donde en las columnas se encuentran todas las dimensiones y en las filas las tablas de hechos, la intersección de una fila con una columna especifica si hay relación entre una tabla de hechos y una dimensión. En la Tabla 4, se muestra la matriz bus del trabajo de diploma donde se reflejan las relaciones entre hechos y dimensiones donde:

H1: hech_inclusion_evinicial.

H2: hech_ex_fisico.

H3: hech_cump_inmunizacion.

H4: hech_labclco.

H5: hech_imagenologia_evallesiones.

H6: hech_interrup_tto.

H7: hech_salida_ensayo.

Dimensiones/Hechos	H1	H2	H3	H4	H5	H6	H7
dpa	x	x	x	x	x	x	x
hospital	x	x	x	x	x	x	x
edad	x						
raza	x						
tiempo	x	x	x	x	x	x	x
peso		x					
etapa		x		x	x		
esquema_tratamiento	x		x				
estado_oms	x	x					
enfermedades	x						
tnm	x						
estadio	x						
receptores_estrogeno	x						
pac_carcinoma_metastasis	x						
pac_consent_inf	x						
pac_entre_18_y_80	x						
evaluacion_funcional	x						
param_entre_lim_est	x						
no_est_protocolo	x						
pac_con_terap	x						
pac_enferm_malig	x						
embaraz_o_lact	x						
pac_metast_cerebral	x						
pac_enferm_autoinm_cron_desc	x						

pac_alerg_reacc_adv	x						
mama_contralateral	x						
biopsia	x						
citologia	x						
mod_trat			x				
exam_fis	x	x					
grupo_tratamiento				x	x		
metodo_diagnostico	x						
respuestas					x		
causa_fallecimiento						x	x
causa_salida						x	
examen_lab	x			x			
tratamiento			x				
retraso_admin			x				
signos vitales			x				
trat_previos	x						
causa_interrupcion						x	

Tabla 4: *Representación de la matriz bus.*

2.8.1.4. Modelo de datos

Una vez identificados los hechos, dimensiones y medidas se procede a realizar el modelo de datos. En este modelo se evidencia la relación entre los hechos y las dimensiones, una relación de uno a mucho donde un hecho puede tener una o varias dimensiones asociadas. Cada dimensión es identificada una llave primaria, que sirve para mantener la referencia en las tablas de hechos con las que se relaciona. En la Figura 7 se muestra un fragmento del modelo de datos de la solución, donde se muestran los hechos hech_salida_ens y hech_int_tto y las dimensiones asociadas a ellos, que encuentra íntegramente en el artefacto “DATEC_CIM_Especificación del modelo de datos” del Expediente de Proyecto:

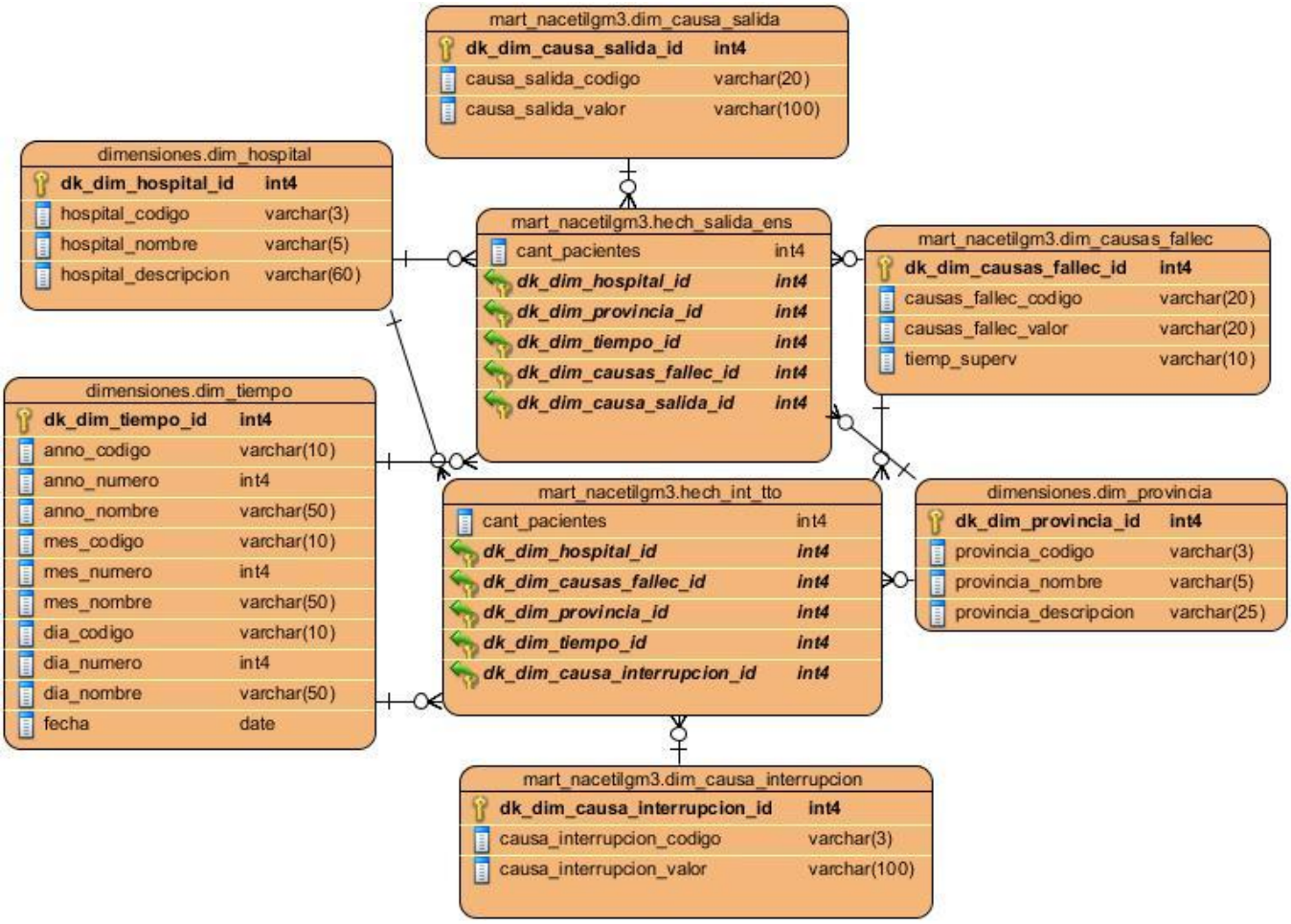


Figura 7: *Modelo de datos*

El modelo de datos diseñado presenta una topología constelación de hechos teniendo en cuenta dos elementos fundamentales:

- ✓ El modelo realizado presenta siete tablas de hechos, es decir, se incluyen varios aspectos del negocio en la estructura del depósito de datos, en función de los requisitos identificados.
- ✓ Las tablas de hechos comparten algunas dimensiones, cuya reutilización optimiza el diseño del modelo de datos.

2.8.2. Diseño general del subsistema de integración

El subsistema de integración contiene el perfilado de los datos y el diseño de las transformaciones, los que constituyen elementos esenciales a realizar. Este subsistema comprende el perfilado y la extracción

de los datos desde la fuente de datos. A estos datos se les aplican un conjunto de transformaciones y luego se almacenan en una base de datos. Para lograr una mejor organización de estos procesos, se cuenta con una carpeta de repositorio de datos organizada por transformaciones, trabajos, fuente de datos y variables de configuración.

2.8.2.1. Perfilado de los datos

El perfilado de datos es el análisis de los datos de las fuentes de datos para entender su contenido, estructura, calidad y dependencias. Este paso es muy importante ya que a la hora de plantear un análisis de los datos de origen es posible encontrarse en muchas ocasiones que realmente no se sabe que preguntar, ni donde pueden residir algunos problemas. Las reglas de transformación definidas se encuentran en el 21 de reglas del negocio. En la Figura 8, un gráfico que revela el comportamiento de los tipos de datos en la fuente de datos:

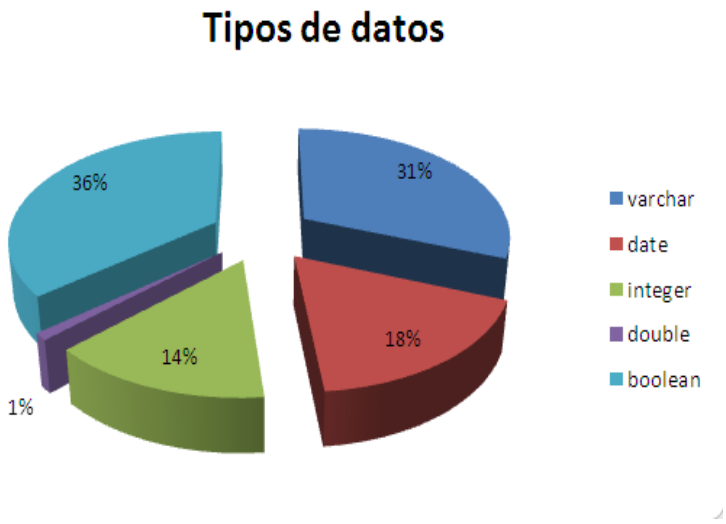


Figura 8: *Gráfico de comportamiento de los tipos de datos en la fuente.*

También se analizó la calidad de los datos de la fuente, donde se observa que de los 16513 tuplas a cargar, 844 eran nulas, representando el 5 por ciento de los datos a cargar. A continuación en la Figura 9 se muestra el análisis:

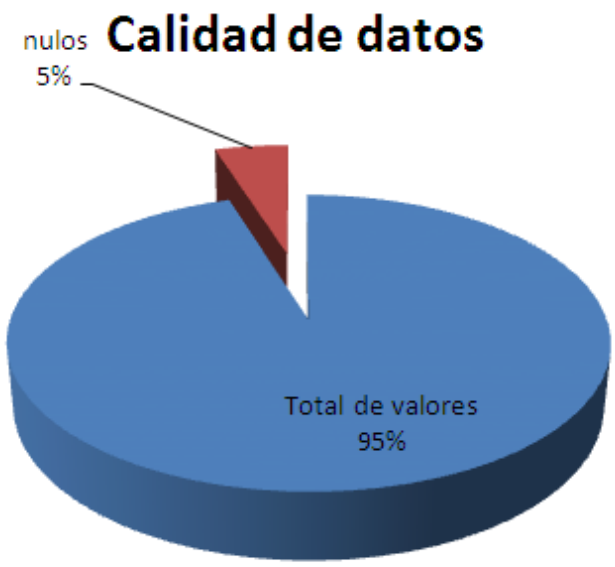


Figura 9: *Calidad de los datos.*

Después de haber analizado la fuente completamente, se definieron nuevas reglas del negocio entre las que se encuentran:

- Las fechas se encontrarán entre el 3/11/2000 y el 25/12/2007.
- La cantidad de pacientes a los que se les aplico el ensayo es de 28, de los cuales sobrevivieron 27.
- La cantidad de pacientes de raza blanca supera a los de raza negra en 4.
- La edad de los pacientes incluidos en el ensayo está en los 35 y los 74 años.
- El peso de los pacientes incluidos en el ensayo se encuentra entre 37 y 104 kg.

Además se declararon nuevas reglas de transformación entre las que están:

- Cambiar los valores vacíos de tipo entero por el valor (-1).
- Cambiar las fechas nulas por la fecha 01/01/1999 ya que es una fecha fuera del rango en que se aplicó el ensayo.
- Cambiar los valores cadenas por ND (no disponible) en los casos que sean vacíos.

2.8.2.2. Diseño de las transformaciones

A continuación se muestra en las Figura 10 y Figura 11, el diseño de las transformaciones para los hechos y para las dimensiones, que permite dar una visión de los pasos a seguir en la implementación:

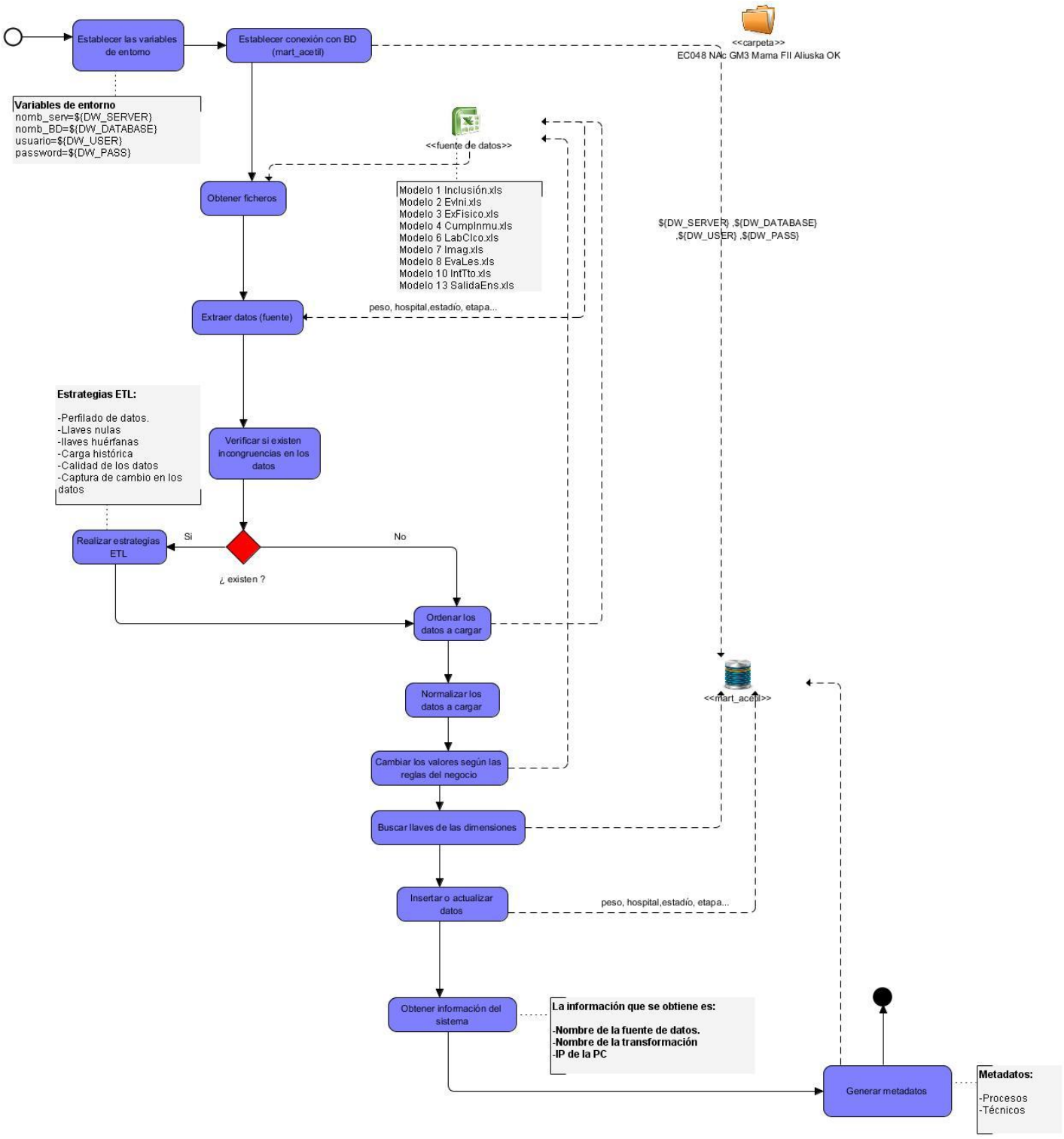


Figura 10: *Representación del diseño del subsistema de integración para los hechos.*

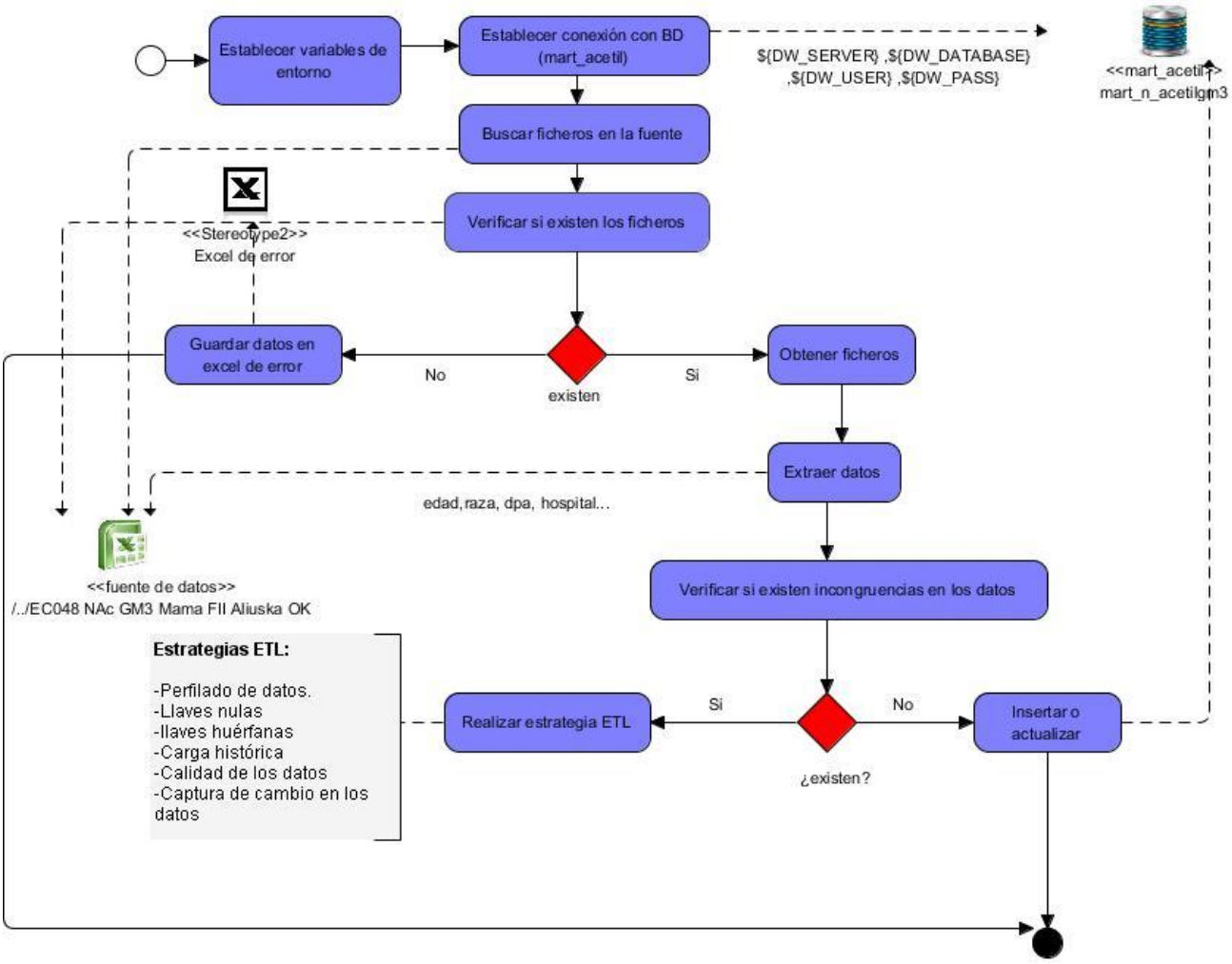


Figura 11: *Representación del diseño del subsistema de integración para las dimensiones.*

2.9. Política de respaldo y recuperación

Con el objetivo de garantizar la persistencia de la información, se establece una política de respaldo y recuperación que comprende:

Hacer una copia de respaldo de la información: se realizará una copia de respaldo de toda la información almacenada en la base de datos, de las transformaciones y trabajos realizados con el objetivo de contar con un método de recuperación ante contingencias.

2.10. Esquema de seguridad

La seguridad de los subsistemas está dada por el nivel de acceso de diferentes roles a la información. Además se pueden tomar medidas tales como la restricción por usuario, de modo que estos roles solo tengan acceso a la información correspondiente, lo cual garantiza la integridad de los datos y permite controlar el acceso a la información. A continuación se describe el esquema de seguridad para los diferentes subsistemas de la solución:

Subsistema de almacenamiento:

- **Administrador ETL:** realiza los procesos de ETL sobre los datos y tiene permisos de lectura y escritura sobre los cuatro esquemas de la base datos pertenecientes a la solución.
- **Analista:** tiene acceso de solo lectura sobre los esquemas de la base de datos pertenecientes a la solución.

Subsistema de integración:

La seguridad del subsistema de integración se garantiza a nivel de sistema operativo. El SGBD, donde se almacenará la información, estará instalado en el sistema operativo *Linux*, en su versión 12.04 para la distribución *Ubuntu*. Este sistema operativo permite asignar permisos (o derechos de acceso) a los archivos para determinados usuarios y grupos de usuarios y a través de esto se puede restringir o permitir el acceso a los archivos que contienen las transformaciones y los trabajos, así como su ejecución y modificación.

2.11. Conclusiones del capítulo

En este capítulo se diseñaron los subsistemas de almacenamiento e integración, se describió la propuesta y se aplicó a uno de los productos sobre el cual se realizan los ensayos clínicos en el CIM. Una vez conocido el negocio se identificaron los requisitos de información, funcionales y no funcionales, así como las reglas del negocio. Se definieron las tablas de hechos y dimensiones obteniéndose el modelo de datos, a partir del cual es posible obtener la estructura física de almacenamiento. Se realizó el perfilado de datos a las fuentes de información para garantizar el estado y calidad de los datos y para verificar la existencia de posibles reglas de negocio y transformación. Se diseñó los procesos de integración de datos relacionados a la carga de las dimensiones y los hechos, quedando establecido el flujo de actividades para la implementación del subsistema de integración.

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

3.1. Introducción

Una vez realizado el diseño de la solución y teniendo como guía la metodología utilizada, se procede a realizar la implementación de los subsistemas de integración y almacenamiento. También se procede a la realización de las pruebas que se seleccionaron para comprobar la calidad de la solución.

3.2. Implementación del subsistema de almacenamiento

En la implementación del subsistema de almacenamiento se debe tener en cuenta varios pasos entre los que se incluye la estandarización de los nombres de los hechos, dimensiones, medidas y atributos. Además se debe desarrollar la estructura física de almacenamiento.

3.2.1. Estandarización de los nombres

En las tablas de dimensiones, al nombre de las mismas le precederán las letras “dim” separadas del nombre de la dimensión por el caracter “_”, ejemplo dim_provincia. En el caso de ser una tabla de hecho, le precederán las letras “hech” e igualmente se separa del nombre de la tabla de hechos por el caracter “_”, ejemplo hech_int_tto, el que hace referencia a los datos almacenados en el modelo interrupción del tratamiento. Las llaves primarias de las dimensiones se nombran de la forma “dk_dim_dimension_id”. En el caso de que este atributo fuera el código de la dimensión se le especificó “dimension_codigo”. También, con respecto a los atributos de los nombres y las descripciones de las dimensiones se decidió nombrarlos: “dimensión_nombre” y “dimension_descripcion”, respectivamente. Las medidas fueron nombradas de la manera “cant_medida”, donde medida es lo que se quiere contar, por ejemplo cant_pacientes.

Para lograr una nomenclatura estándar de las estructuras del subsistema de integración, se definió que las transformaciones correspondientes a las dimensiones fueran nombradas de la manera dim” separadas del nombre de la dimensión por el caracter “_”, ejemplo dim_peso. En el caso de las transformaciones correspondientes a los hechos, se definió la estructura “hech” y el caracter “_” lo separa del nombre de la tabla de hechos, por ejemplo hech_int_tto, Si la transformación es un job o trabajo, entonces su nomenclatura será “job_nombre del trabajo”, ejemplo job_trabajo_general. Al finalizar el proceso de estandarización de los nombres, se encuentra organizada la nomenclatura utilizada para las tablas, atributos y medidas dentro de la base de datos y de las estructuras del subsistema de integración.

3.2.2. Indexado de la base de datos

El índice de una base de datos es una estructura de datos que mejora la velocidad de las operaciones, permitiendo un rápido acceso a los registros de una tabla en una base de datos. Al aumentar la velocidad de acceso, se suelen usar sobre aquellos campos sobre los cuales se hacen búsquedas frecuentemente. Tiene un funcionamiento similar al índice de un libro, guardando el elemento que se desea indexar y su posición en la base de datos. Para la búsqueda de un elemento que esté indexado, sólo hay que buscar dicho elemento en el índice, para una vez encontrado, devolver el registro que se encuentre en la posición marcada por el índice. Los índices son creados a partir de una o varias columnas de una tabla y evitan se escanee completamente la misma. Las columnas de las tablas cuyo indexado es necesario, se determinan al conocer los datos y los tipos de consultas que se van a realizar. La cláusula más frecuente a tener en cuenta es WHERE. En un mercado de datos se recomienda usar una estrategia de indexado que consiste en crear índices para las llaves primarias y foráneas. Esto se debe a que para la mayoría de las uniones entre tablas que se realizan, las columnas implicadas son precisamente las llaves y este tipo de operaciones son las que más tiempo de consulta consumen. Para esto se utiliza los indexados que posibilita el gestor PostgreSQL, que son los árboles B y los índices tipo hash:

Arboles B: es aquel en el que se mantiene incluso la cantidad de datos en el lado izquierdo y derecho de cada división, de modo que la cantidad de niveles que tiene que descender para llegar a cualquier fila individual es aproximadamente igual. Los árboles B se puede utilizar para encontrar un solo valor o para buscar un rango de ellos, buscando por los valores de la llave que son mayor que, menor que, y / o igual a algún valor. También pueden encontrar (o evitar) los valores NULL en una tabla.

Hash: los índices de tipo hash pueden ser útiles en casos en los que sólo se está haciendo la igualdad buscando en un índice, y no se quiere permitir valores NULL en ella. Sin embargo, este tipo de índices pueden corromperse fácilmente después de un accidente de base de datos, y por lo tanto es ineficaz para las consultas hasta su reconstrucción manual. Las ventajas de utilizar un índice hash en lugar de un árbol B son pequeñas en comparación con el riesgo. Normalmente nunca se debe utilizar el tipo de índice hash, pero si se está dispuesto a gastar una cantidad considerable de tiempo a analizar su utilidad y la garantía de que el índice se reconstruirá si se corrompe, es posible que se pueda encontrar un propósito para ellos (17).

De los tipos de indexado anteriormente explicados en la solución se utilizó el indexado hash, debido a que todas las consultas y operaciones que se hicieron entre las tablas de hechos y dimensiones fueron de igualdad.

3.2.3. Implementación del modelo de datos físico

El esquema de una base de datos define todas sus tablas y cada campo en cada una de ellas. La base de datos mart_acetil cuenta con 42 tablas, divididas en 26 tablas de dimensiones, 7 tablas de hechos y 9 tablas de metadatos. A continuación se define los esquemas:

- El esquema “dimensiones” contiene las dimensiones compartidas de la solución.
- El esquema “mart_nacetilgm3” recoge las tablas de hechos y las dimensiones propias de la solución.
- El esquema “metadatos” se define con el objetivo de llevar la gestión de errores y de la carga histórica, para ello se ha diseñado una estructura basada en el uso de nueve tablas de metadatos.

3.3. Implementación del subsistema de integración

En la implementación del subsistema de integración es donde se realizan los procesos de ETL. Es recomendable iniciar los procesos de integración de datos después de haber realizado un análisis previo de las fuentes de datos, la que está integrada por ficheros “.xls”.

Se extraen los datos de las fuentes, seleccionando los campos relevantes, a partir del modelo de datos realizado. Con el perfilado de los datos se detectan las incongruencias en los ellos tales como campos nulos, entradas duplicadas, errores en los tipos de datos y con las transformaciones se combinan y ordenan los datos. Para la gestión de errores y la carga histórica se crearon nueve tablas de metadatos, de ellas 7 de gestión de carga histórica, dos de gestionar la información del sistema y nueve ficheros Excel que contendrán los errores de existencia de los ficheros en la fuente A continuación se describen algunos de los subsistemas que se utilizaron en el desarrollo de la solución:

- **Perfilado de datos:** permitió verificar la calidad de los datos y el cumplimiento de los estándares conforme a los requisitos especificados por el cliente. A través este subsistema fueron definidas nuevas reglas de transformación.
- **Sistema de extracción:** posibilitó la extracción de los datos desde la fuente de origen para su transformación y posterior carga. Para ello se tuvo en cuenta la información relacionada con cada uno de los hechos y dimensiones.

- **Subsistema de transformación:** este subsistema contribuyó a realizar el mapeo de valores, el cambio de tipo de dato en algunos campos, la búsqueda de información en flujos de datos y el filtrado de valores.
- **Subsistema de carga:** permitió realizar la cargar de los datos a las tablas de dimensiones y hechos de la solución.
- **Llave subrogada:** permite crear claves subrogadas independientes para cada tabla.
- **Dimensiones Lentamente Cambiantes:** implementa la lógica para crear atributos de variabilidad lenta a lo largo del tiempo. En el caso de la presente solución se utilizó el tipo 1 de SCD pues solo se sobrescribirán valores.
- **Sistema de back up:** realiza copias de respaldo de los procesos ETL.
- **Seguridad:** gestiona el acceso a ETL y metadatos.
- **Repositorio de metadatos:** captura los metadatos de los procesos de ETL, de los datos de negocio y de los aspectos técnicos.

3.3.1. Implementación de las transformaciones

Para realizar la extracción de los datos correspondientes a cada una de las tablas de hechos de la solución, se accede a la fuente de datos, se extraen los campos necesarios atendiendo cada hecho con qué dimensión se relaciona y se procede a realizar las transformaciones correspondientes. A continuación se muestran dos ejemplos que corresponden a la dimensión enfermedades (dim_enfermedades) y al hecho interrupción del tratamiento (hech_intto).

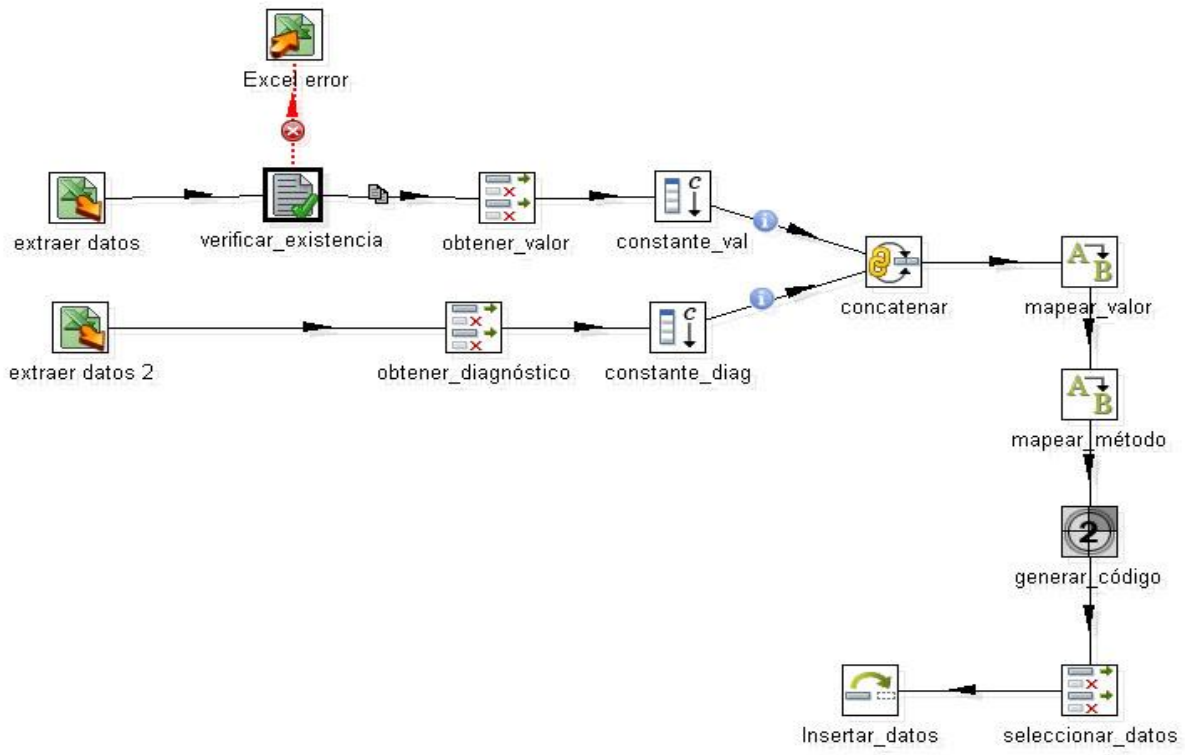


Figura 12: *Ejemplo de la carga de dimensiones (dim_enfermedades).*

En la Figura 12**Figura 12:** Ejemplo de la carga de dimensiones (dim_enfermedades). se hace referencia a la carga de una dimensión donde se extraen los datos a cargar, luego se seleccionan los datos a insertar y se mapean los valores según las reglas de la transformación antes definidas. Para terminar se generan los códigos de las dimensiones y se insertan los datos en las base de datos. Otro ejemplo de las transformaciones es el de los hechos, en el ejemplo de la Figura 13**Figura 13:** Ejemplo de la carga de un hecho (hech_intto)., donde se muestra el proceso de extracción, transformación y carga del hecho interrupción del tratamiento en el que después de extraer los datos se pasa al proceso de transformación según el negocio. Luego de obtener las llaves de las dimensiones a cargar, las que se validan y si existe alguna llave nula se agrega a la tabla de llaves huérfanas del esquema de metadatos. En el caso de que no existan incongruencias de ningún tipo se procede a insertar los datos y luego a generar la información del sistema.

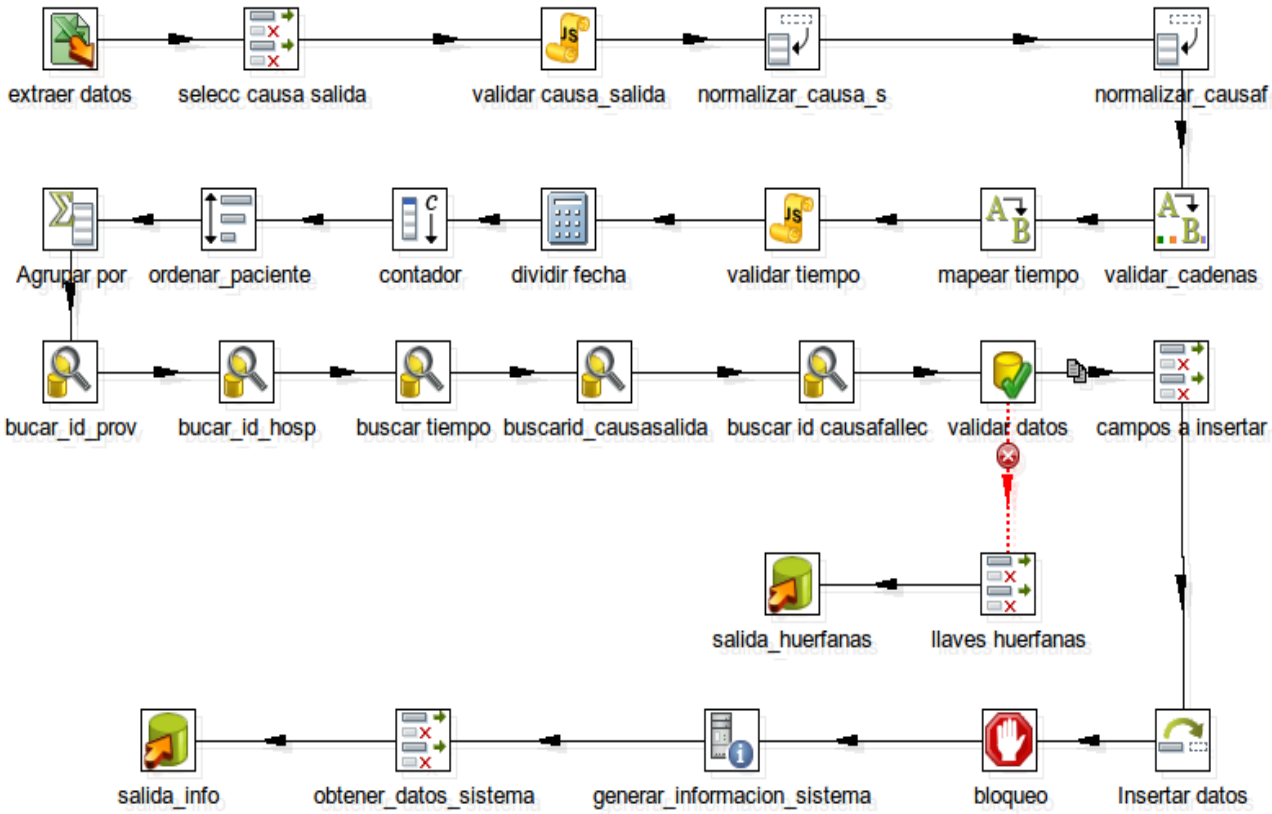


Figura 13: *Ejemplo de la carga de un hecho (hech_intto).*

3.3.2. Implementación de los trabajos

El término trabajo o job, en el contexto de los procesos ETL, es un conjunto de tareas cuyo objetivo consiste en realizar una acción determinada (16). Permite ejecutar varias transformaciones o trabajos previamente diseñados y organizar una secuencia de ejecución. Los trabajos se encuentran en un nivel superior de las transformaciones. Además, es posible ejecutar una o varias transformaciones de las que se han diseñado y definir una secuencia de ejecución para éstas.

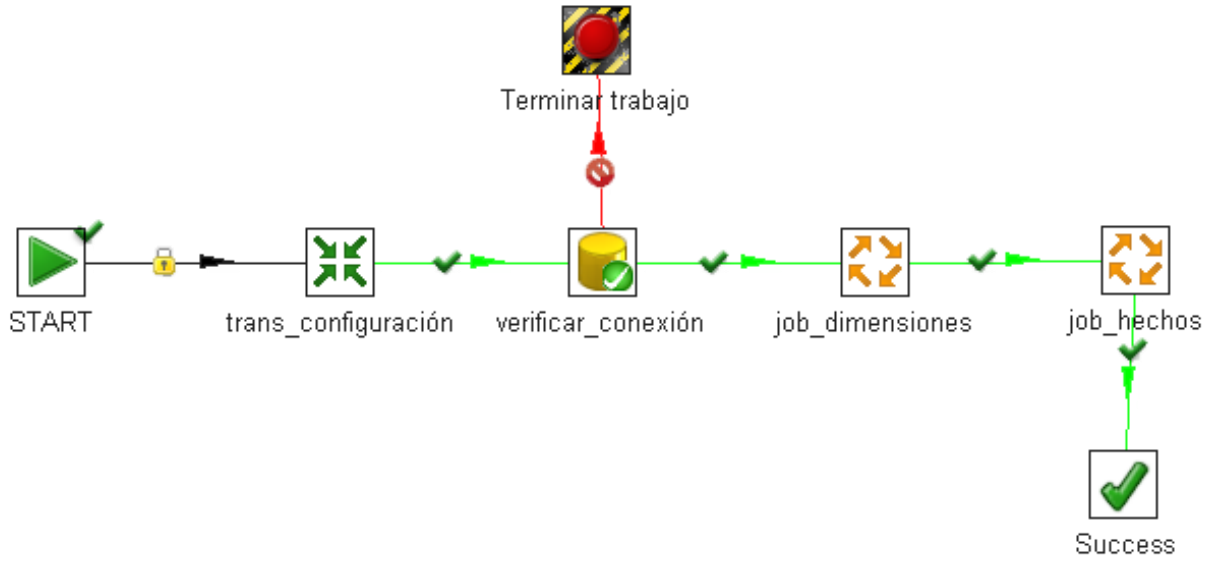


Figura 14: *Ejemplo de un trabajo (job_general).*

En la Figura 14***Figura 14:*** *Ejemplo de un trabajo (job_general).* se muestra la carga de todos los jobs a través de un job general. El proceso empieza con la obtención de la conexión y su validación. En el caso de que existan fallos en la conexión se termina el job, sino se procede a la carga de las dimensiones y luego a la de los hechos.

3.4. Aplicación de las pruebas.

Para determinar la calidad del producto se realizan distintas pruebas de software. Este proceso comienza con la planificación de las mismas, posteriormente se procede a la ejecución, el control y como paso final a la evaluación de éstas. A continuación se describen las pruebas que se le realizaron a la solución:

3.4.1. Aplicación de las pruebas de integración.

Se aplicaron las pruebas de integración para comprobar la correcta integración entre los subsistemas de la solución. Para ello se realizaron 14 casos de pruebas, donde se escogieron 14 requisitos de información (dos por cada CUI) y se comprobó que el valor de las medidas del hecho entrantes al flujo coincidiera con el resultado esperado.

3.4.1.1. Casos de prueba

Los casos de prueba son un conjunto de condiciones o variables bajo las cuales el analista podrá determinar si el requisito de una aplicación es parcial o completamente satisfactorio. Con el objetivo de

verificar los RI definidos durante la etapa de análisis y para validar la solución, se diseñaron catorce casos de pruebas basándose en los siete CUI, los cuales están recogidos en el Expediente de Proyecto. En la Figura 15**Figura 15:** Caso de prueba basado en el CU mantener disponible la información de los modelos inclusión y evaluación inicial. se muestra el caso de prueba correspondiente un caso de prueba basado en el CUI Mantener disponible la información de los modelos inclusión y evaluación inicial, refiriéndose a la cantidad de pacientes de color blanco que existen en ese modelo:

Caso de uso de información	Requisito de información	Tablas implicadas	Variables de entrada	Variables de salida	Consulta SQL realizada	Datos obtenidos	Fuente de datos	Variables de la fuente implicadas	Datos almacenados en la fuente	Resultados de la prueba
Mantener disponible la información de los modelos de inclusión y evaluación inicial	Obtener la cantidad de pacientes de color blanco en el modelo inclusión y evaluación inicial a los que se le aplicó el producto.	dim_raza y hech_inclusion_evinicial	cod_paciente, raza_id	distinct(cod_paciente)	SELECT distinct (hech_inclusion_evinicial.cod_paciente) FROM mart_nacetilgm3.hech_inclusion_evinicial, dimensiones.dim_raza WHERE dim_raza.dk_dim_raza_id = hech_inclusion_evinicial.dk_dim_raza_id and dim_raza.dk_dim_raza_id =2;	24	Modelo 1 Inclusión .xls	v7	Se obtuvieron 24 pacientes de raza blanca	Resultado satisfactorio

Figura 15: *Caso de prueba basado en el CU mantener disponible la información de los modelos inclusión y evaluación inicial.*

3.4.2. Aplicación de las pruebas unitarias.

Las pruebas unitarias son las pruebas que se le realizan a cada uno de los subsistemas por separado de la solución, para su verificar correcto funcionamiento. Estas pruebas fueron realizadas por los especialistas del centro de DATEC a cada uno de los subsistemas en específico. En el subsistema de almacenamiento se detectaron las siguientes no conformidades (NC) las que fueron arregladas luego de detectadas:

NC_1: Los casos de usos no se encuentran por criterio de agrupación.

NC_2: Existen problemas en la identificación de los niveles y las jerarquías de las dimensiones.

NC_3: Faltan algunas unidades de medidas y no existe aditividad en las variables de salida.

NC_4: Las reglas del negocio no se encuentran clasificadas.

El otro subsistema que se evaluó fue el de integración, en el cual se detectaron algunas no conformidades que fueron arregladas también luego de detectadas y a continuación se describen:

NC_5: Se deben modificar los nombres de las actividades en los diseños de los proceso de integración de datos que reflejen su verdadero objetivo.

NC_6: Existen errores de ortografía en el diseño de los proceso de integración de datos.

NC_7: Se deben especificar los nombres de los ficheros a cargar en los diseños del proceso de integración de datos.

NC_8: Se deben arreglar los resultados del perfilado de datos que corresponda con la realidad del negocio.

Además se comprobó el correcto funcionamiento de las transformaciones y que los datos cargados en la base de datos fueran correctos, comprobando así la calidad de los procesos realizados.

3.4.3. Aplicación de las listas de chequeo.

La lista de chequeo contiene un listado de preguntas en forma de cuestionario que sirven para verificar el grado de cumplimiento de determinadas reglas establecidas. Esta herramienta sirve para posibles soluciones o la detección de oportunidades de mejora y es uno de los mecanismos más utilizados para la evaluación y control de la calidad de los productos que se desarrollan en la universidad. Además se decidió hacer pruebas de integración, las que son ejecutadas para asegurar que los componentes en el modelo de implementación operen correctamente cuando son combinados para ejecutar un CU. Se prueba un paquete o un conjunto de paquetes del modelo de implementación y éstas descubren errores o incompletitud en las especificaciones de las interfaces de los paquetes.

3.4.3.1. Elaboración y evaluación de la lista de chequeo

Las listas de chequeo, aplicadas a cada uno de los artefactos de ETL, contienen diferentes indicadores a evaluar los cuales se encuentran distribuidos en tres secciones fundamentales:

- **Estructura del documento:** abarca todos los aspectos definidos por el expediente de proyecto o el formato establecido por el proyecto.
- **Indicadores definidos:** abarca todos los indicadores a evaluar durante la etapa.

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

3.4.3.2. 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 (Eval):** 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.

Una vez aplicada la lista de chequeo se detectan los indicadores evaluados de mal y con el objetivo de darles solución se especifican en una tabla de no conformidades, la cual presenta la siguiente estructura:

- **No.**: es un número consecutivo e indica la cantidad de no conformidades identificadas.
- **Elemento de evaluación**: se refiere a un número que identifica al elemento de evaluación para el cual se corresponden los indicadores identificados.
- **No conformidad**: especifica la no conformidad a la que se refiere.
- **Fase correspondiente**: especifica la fase del procedimiento a la que corresponde la no conformidad encontrada.
- **Significación**: especifica si la no conformidad es o no significativa, dependiendo si el indicador es o no crítico.
- **Recomendación**: especifica si la no conformidad es una recomendación, es decir, que no es de obligatorio cumplimiento que se solucione por parte de los diseñadores.

- **Estado NC**: especifica el estado de solución en que se encuentra la no conformidad, puede ser: **pendiente** o **solucionada**.
- **Respuesta del equipo de desarrollo**: si es necesario se especifica la respuesta que le da el equipo de desarrollo a la no conformidad.

3.4.3.3. Evaluación de las listas de chequeo

Se aborta la revisión si:

- Existen al menos dos indicadores críticos evaluados de mal en la sección **Indicadores** que posee la lista de chequeo.
- Más del 50 % de los indicadores a evaluar están evaluados de mal.
- Se mantienen las no conformidades de una revisión a otra.

Se evalúa de regular la calidad del diseño revisado si no cumple los criterios para ser abortado y además:

- Incumple con los indicadores críticos a evaluar de las secciones **Estructura del documento** y **Semántica del documento** que posee la lista de chequeo.
- Existe al menos un indicador crítico evaluado de mal.
- Existen al menos cinco indicadores no críticos evaluados de mal de la sección **Indicadores evaluados por la etapa** que posee la lista de chequeo.

El diseño es evaluado de bien si no cumple ninguno de los criterios anteriores y:

- No existe ningún indicador crítico evaluado de mal.
- Si la cantidad de indicadores no críticos evaluados de mal de la sección **Indicadores** que posee la lista de chequeo no es mayor que cuatro.

Después de elaborada y aplicada las listas de chequeo en la presente investigación se obtuvieron los siguientes resultados:

- Se identificaron, en total, 50 indicadores indispensables en general para así poder evaluar la etapa de diseño presente. Los mismos fueron distribuidos en tres secciones: Estructura del documento, Indicadores definidos y Semántica del documento y de ellos 20 fueron definidos como críticos y se

detectaron en una primera iteración 5 no conformidades, las cuales fueron resueltas inmediatamente.

- Después de analizar los resultados el producto fue evaluado de Bien.

En la Figura 16 se muestra una gráfica con el comportamiento de los indicadores definidos en la aplicación de la lista de chequeo a los diferentes artefactos de los procesos ETL.

Figura 16: *Comportamiento de los indicadores en general.*

3.5. Conclusiones del capítulo

En este capítulo se realizó todo el proceso de implementación de los subsistemas de almacenamiento e integración, desarrollando las transformaciones pertinentes en el proceso ETL, a través de las cuales se realizaron las transformaciones según las reglas definidas en el perfilado de los datos. Además se cargó toda la información en una sola base de datos y la aplicación de las pruebas seleccionadas y de las listas de chequeo contribuyó a la verificación de la calidad del producto final y las deficiencias encontradas fueron corregidas exitosamente.

Conclusiones generales

Luego de finalizada la investigación se puede afirmar que se le dio cumplimiento al objetivo general, cumpliéndose además con las tareas de la investigación planteadas para su cumplimiento. Por tanto se arriba a las siguientes conclusiones :

- La metodología seleccionada guió el proceso de desarrollo de la solución durante cada etapa del ciclo de vida. Las herramientas y tecnologías seleccionadas guiaron el proceso de desarrollo de software.
- El análisis y diseño de la solución generó los artefactos necesarios para la posterior etapa de implementación.
- La implementación de las estructuras modeladas permitió la integración de los datos y su almacenamiento.
- Las pruebas efectuadas permitieron comprobar la calidad del producto a partir de los requisitos establecidos. Los resultados obtenidos durante las pruebas realizadas fueron satisfactorios, validando el cumplimiento de los objetivos propuestos.

Recomendaciones

- Integrar la solución al almacén de datos de los ensayos clínicos del Centro de Inmunología Molecular.
- Se recomienda la aplicación de una técnica de minería de datos a la solución para descubrir patrones y tendencias en los datos.

Referencias bibliográficas

1. Sitio del Instituto Nacional del Cáncer. [En línea] [Citado el: 26 de Noviembre de 2012.] http://www.cancer.gov/espanol/cancer/que-es.

2. Tomado del sitio del CIMAB. [En línea] [Citado el: 26 de Noviembre de 2012.] http://www.cimab-sa.com/index.php?action=cim.

3. Plataforma de registros internacionales de ensayos clínicos de la OMS. [En línea] [Citado el: 29 de Noviembre de 2012.] http://www.who.int/ictrp/es/.

4. *Mercado de datos para la dirección de cuadros de la Administración Provincial de Artemisa.* **Rodríguez, Yisel Valdés.** 2013.

5. Definicion.de. *Definición de modelo de datos.* [En línea] [Citado el: 13 de Enero de 2013.] http://definicion.de/modelo-de-datos/.

6. **Berzal, Fernando.** Modelo multidimensional. [En línea] [Citado el: 12 de Diciembre de 2012.] http://elvex.ugr.es/.

7. **Méndez, Ana Laura Alba.** *Arquitectura, Diseño, Mantenimiento y Consulta de un Almacén de Datos.*

8. **Basallo, Yasser Azán.** *Una experiencia sobre integración de aplicaciones informáticas.*

9. **Sanz, Miguel Rodríguez.** *ANÁLISIS Y DISEÑO DE UN DATAMART PARA EL SEGUIMIENTO ACADÉMICO DE ALUMNOS EN UN ENTORNO UNIVERSITARIO.*

10. Tomado del sitio Dataprix. *Dimensiones Lentamente Cambiantes.* [En línea] [Citado el: 12 de Diciembre de 2012.] http://www.dataprix.com/blogs/bernabeu-dario/dimensiones-lentamente-cambiantes.

11. **Darío, Bernabeu Ricardo.** *DATA WAREHOUSING: Investigación y Sistematización de Conceptos.*

12. **López, Patricia y Sierra, María.** *Herramienta CASE Visual Paradigm.*

13. **Gibert Ginesta, Marc y Pérez Mora, Oscar.** *Bases de datos en PostgreSQL.*

14. Tomado del sitio Debian. *Herramienta para la administración gráfica de PostgreSQL. .* [En línea] [Citado el: 11 de Diciembre de 2012.] http://packages.debian.org/es/squeeze/pgadmin3.

15. **Expósito Martín, Miguel.** *Arquitectura de Integración de Datos del ICANE.*

16. **González Cornejo, José Enrique.** *¿Qué es UML? .* [En línea] [Citado el: 11 de diciembre de 2012.] http://www.docirs.cl/uml.htm.

17. **Smith, Gregory.** *PostgreSQL 9.0 High Performance.*

Bibliografía consultada

1. Sitio del Instituto Nacional del Cáncer. [En línea] [Citado el: 26 de Noviembre de 2012.] http://www.cancer.gov/espanol/cancer/que-es.

2. Tomado del sitio del CIMAB. [En línea] [Citado el: 26 de Noviembre de 2012.] http://www.cimab-sa.com/index.php?action=cim.

3. Plataforma de registros internacionales de ensayos clínicos de la OMS. [En línea] [Citado el: 29 de Noviembre de 2012.] http://www.who.int/ictrp/es/.

4. *Mercado de datos para la dirección de cuadros de la Administración Provincial de Artemisa.* **Rodríguez, Yisel Valdés.** 2013.

5. Definicion.de. *Definición de modelo de datos.* [En línea] [Citado el: 13 de Enero de 2013.] http://definicion.de/modelo-de-datos/.

6. **Berzal, Fernando.** Modelo multidimensional. [En línea] [Citado el: 12 de Diciembre de 2012.] http://elvex.ugr.es/.

7. **Méndez, Ana Laura Alba.** *Arquitectura, Diseño, Mantenimiento y Consulta de un Almacén de Datos.*

8. **Basallo, Yasser Azán.** *Una experiencia sobre integración de aplicaciones informáticas.*

9. **Sanz, Miguel Rodríguez.** *ANÁLISIS Y DISEÑO DE UN DATAMART PARA EL SEGUIMIENTO ACADÉMICO DE ALUMNOS EN UN ENTORNO UNIVERSITARIO.*

10. Tomado del sitio Dataprix. *Dimensiones Lentamente Cambiantes.* [En línea] [Citado el: 12 de Diciembre de 2012.] http://www.dataprix.com/blogs/bernabeu-dario/dimensiones-lentamente-cambiantes.

11. **Darío, Bernabeu Ricardo.** *DATA WAREHOUSING: Investigación y Sistematización de Conceptos.*

12. **López, Patricia y Sierra, María.** *Herramienta CASE Visual Paradigm.*

13. **Gibert Ginesta, Marc y Pérez Mora, Oscar.** *Bases de datos en PostgreSQL.*

14. Tomado del sitio Debian. *Herramienta para la administración gráfica de PostgreSQL.* [En línea] [Citado el: 11 de Diciembre de 2012.] http://packages.debian.org/es/squeeze/pgadmin3.

15. **Expósito Martín, Miguel.** *Arquitectura de Integración de Datos del ICANE.*

16. **González Cornejo, José Enrique.** *¿Qué es UML? .* [En línea] [Citado el: 11 de diciembre de 2012.] http://www.docirs.cl/uml.htm.

17. **Cárdenas, Clara Pérez.** Revista Cubana Medicina General Integral. [En línea] 2001. [Citado el: 11 de enero de 2013.] http://www.bvs.sld.cu/revistas/mgi/vol17_3_01/mgi10301.htm.

18. Programa de Desarrollo Clínico Laboratorios Elea SACIFyA. *Vacunas basadas en Gangliósidos.* [En línea] [Citado el: 27 de noviembre de 2012.] http://www.elea.com/vacunas-oncologicas-investigacion.htm.

19. Recombio. *Investigación de vacunas para HIV/SIDA.* [En línea] [Citado el: 1 de diciembre de 2013.] http://www.recombio.com/es/investigacion_hiv.php.

20. **Gallardo, Erith Eduardo Perez.** *Data Warehouse, Conceptos Fundamentales.*

21. **Herrera, Cristhian.** Adictos al trabajo. *Planificación Tecnológica.* [En línea] [Citado el: 5 de enero de 2013.] http://www.adictosaltrabajo.com/tutoriales/tutoriales.php?pagina=datawarehouse3#2.9.3.3.Herramientas%20de%20Procesamiento%20Anal%C3%ADtico%20en%20L%C3%ADne.

22. Tomado de Informacion Tecnica. [En línea] [Citado el: 5 de diciembre de 2012.] http://www.synerplus.es/Informacion-Tecnica/Data-Mart/309.html.

23. Lerrot.com. *Gestión de Contenido Especializado.* [En línea] [Citado el: 9 de enero de 2013.] http://www.leroot.com/site/index.php?option=com_content&view=article&id=63&Itemid=69.

24. **Hernández, Yanisbel González.** *PROPUESTA DE METODOLOGIA DE DASARROLLO DE ALMACENES DE DATOS. .* La Habana : s.n., Octubre, 2012.

25. Scribd. *Scribd.* [En línea] 14 de abril de 2010. [Citado el: 9 de enero de 2013.] http://es.scribd.com/doc/75043572/PostgreSQL.

26. Tomado del Portal en español sobre PostgreSQL. [En línea] [Citado el: 28 de noviembre de 2012.] http://www.postgresql.org.es/sobre_postgresql.

27. **Duque Méndez, Néstor Darío.** AUDITORÍA A SISTEMAS DE BODEGAS DE DATOS, DATA WAREHOUSE. [En línea] [Citado el: 27 de noviembre de 2012.] http://www.virtual.unal.edu.co/cursos/sedes/manizales/4060035/lecciones/Capitulo5.html.

28. **Pibaque Pillasagua, Flor Maricela.** *Tesis Desarrollo de un prototipo de inteligencia de negocio para Pymes usando herramientas Open Souce(Pentaho).* Guayaquil : s.n., 2011.

29. **Pressman, Roger S.** *Ingeniería del Software. Un enfoque Práctico. Cuarta edición.*

30. **Pressman, Roger S.** *Ingeniería del Software. Un enfoque Práctico. Quinta edición.*

31. **Smith, Gregory.** *PostgreSQL 9.0 High Performance.*

Anexos

Anexo 1

Sistema de clasificación del cáncer de mama

TNM:

Tumor primario (T)

Tx - El tumor primario no puede ser valorado.

T0 - Sin signos de tumor primario.

Tis - Carcinoma *in situ* (Carcinoma interductal, Carcinoma lobular *in situ*, enfermedad de Paget del pezón sin tumor).

T1 - Tumor con una dimensión máxima igual o inferior a 2 cm.

T2 - Tumor mayor de 2 cm pero con una dimensión máxima igual o inferior a 5 cm.

T3 - Tumor mayor de 5 cm

T4 - Tumor de cualquier tamaño con las siguientes características:

- extensión a la pared torácica,
- carcinoma inflamatorio.

Ganglios linfáticos regionales (N)

- Nx - Los ganglios linfáticos que no pueden ser valorados (pueden haber sido extirpados).
- N0 - Sin metástasis en ganglios linfáticos regionales.
- N1 - Metástasis a ganglios linfáticos axilar(es) homolaterales móvil(es) con las siguientes características:
  - ✓ Solo micro metástasis (no mayor de 0.2 cm),
  - ✓ Metástasis a ganglios regionales (menor de 0.2 cm):
  - ✓ 1-3 ganglios regionales (menores de 0.2 cm) y todos menores de 2 cm de dimensión máxima.
  - ✓ 4 ó más ganglios regionales, ninguno mayor que 0.2 cm y todos menor de 2 cm.

- ✓ extensión del tumor más allá de la cápsula de una metástasis de ganglio linfático, que sea menor de 2 cm.
- ✓ un ganglio linfático de 2 cm o mayor.
- ✓ N2 - Metástasis en ganglios linfáticos homolaterales adheridos a otros ganglios o a alguna otra estructura.
- ✓ N3 - Metástasis en ganglios linfáticos mamarios internos homolaterales.

Metástasis a distancia (M)

- Mx - No puede valorarse la metástasis.
- M0 - Sin metástasis a distancias.
- M1 - Metástasis a distancia.

Anexo 2

CAPACIDAD FUNCIONAL DEL PACIENTE, SEGÚN GRADOS DE LA OMS

GRADO	**CARACTERÍSTICAS**
0	Capaz de llevar a cabo una actividad física normal sin restricciones.
1	Paciente ambulatorio capaz de llevar a cabo un trabajo ligero.
2	Paciente ambulatorio incapaz de realizar ningún trabajo, pero capaz de realizar sus cuidados personales; más del 50 % del tiempo vigil, fuera de la cama.
3	Capaz de realizar sus cuidados personales, pero más del 50 % del tiempo confinado a la cama o la silla.
4	Completamente incapaz de realizar ningún esfuerzo, confinado totalmente a la cama.

Glosario de términos

Antígeno: Sustancia que induce la formación de anticuerpos, debido a que el sistema inmune la reconoce como una amenaza.

CIM: Centro de Inmunología Molecular.

Data Mart: mercado de datos.

Data Warehouse: almacén de datos.

DATEC: Centro de Tecnologías de Gestión de Datos.

ETL: proceso de extracción, transformación y carga.

Lista de chequeo: instrumento de medición y evaluación que consiste básicamente en un formulario de preguntas referentes al atributo de calidad que se está probando y de las características del documento en el caso de la documentación.

No conformidad: defecto, error o sugerencia que se le hace al equipo de desarrollo una vez encontrada alguna dificultad en lo que se está evaluando.

OMS: Organización Mundial de la Salud.

Transportador: El transporte celular es el mecanismo mediante el cual entran a la célula los materiales que se necesitan mientras salen los materiales de desecho y las secreciones celulares.