Universidad de las Ciencias Informáticas
Facultad 6

Título: Subsistemas de almacenamiento e integración del mercado de datos Racotumumab para el almacén de datos de los ensayos clínicos del Centro de Inmunología Molecular.

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

Autores: Ailin Guerra Fernández
Yasmany Pérez Torres

Tutores: Ing. Yaneisy Pedraza González
Ing. Adilen Guerra Sanabria

La Habana, junio 2013
"Año 55 de la Revolución"

"Nuestra recompensa se encuentra en el esfuerzo y no en el resultado. Un esfuerzo total es una victoria completa"

Mahatma Gandhi

DECLARACIÓN DE AUTORÍA

Declaramos que somos los únicos autores de este trabajo y autorizamos a la Facultad 6 de la Universidad de las Ciencias Informáticas a hacer uso del mismo en su beneficio.

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

Ailin Guerra Fernández

_____
Firma del autor

Yasmany Pérez Torres

_____
Firma del autor

Ing. Yaneisy Pedraza González

_____
Firma del tutor

Ing. Adilen Guerra Sanabria

_____
Firma del tutor

Tutora: Ing. Yaneisy Pedraza González

Especialidad de graduación: Ingeniería en Ciencias Informáticas

Años de experiencia en el tema: 3

Años de graduado: 4

Correo electrónico: ypedraza@uci.cu

Tutor: Ing. Adilen Guerra Sanabria

Especialidad de graduación: Ingeniería en Ciencias Informáticas

Años de experiencia en el tema: 2

Años de graduado: 1

Correo Electrónico: agsanabria@uci.cu

A nuestros padres, por todo el amor, apoyo, dedicación que nos han ofrecido y por haberse mantenido a nuestro lado todos estos años, brindándonos fuerzas y sabiduría para poder alcanzar nuestras metas.

A nuestros hermanos, por estar siempre a nuestro lado y apoyarnos tanto en los malos como en los buenos momentos.

A nuestras familias, por confiar en nuestras decisiones aunque no siempre fueran las mejores.

A Yaneisy, Adilen, Adalennis, Esley, Fabian, Monchi y Doris por la dudas aclaradas, los consejos brindados y por su extraordinaria paciencia y comprensión.

A Rosa, Adrián, Yenisleidi, Liena, Yari, Nelson, Sureny, Bañobre, Yuned, Yailema y a todo el grupo 6507, por apoyarnos, ayudarnos y por los tantos recuerdos imborrables que tenemos juntos.

A los profesores que nos dieron clases en estos cinco años, a los del departamento de Almacenes de Datos, al tribunal y al oponente, por todas las horas dedicadas que contribuyeron en nuestra formación como ingenieros informáticos.

A todos los que de una forma u otra han aportado en nuestra preparación, como jóvenes comprometidos con la Revolución para hacer este sueño realidad.

Ailin y Yasmany

RESUMEN

La presente investigación surge como parte de la colaboración que existe entre la Universidad de las Ciencias Informáticas y el Centro de Inmunología Molecular (CIM). Este último presenta una serie de problemas que dificultan el manejo de los datos por parte de los especialistas, lo que propicia la pérdida de información útil y valiosa. Como solución a la problemática existente en el CIM, se propone desarrollar los subsistemas de almacenamiento e integración del mercado de datos Racotumumab cuyo objetivo fundamental es lograr la estandarización de los datos para su almacenamiento de forma homogénea. Se utilizó para la implementación del producto la Propuesta de metodología para el desarrollo de almacenes de datos del Centro de Tecnologías y Gestión de Datos y las herramientas Visual Paradigm, PostgreSQL, PgAdmin, DataCleaner, Microsoft Excel y Pentaho Data Integration. Se llevó a cabo las etapas de análisis, diseño, implementación y prueba una vez conocida las peculiaridades del negocio. Finalmente el producto que se obtuvo almacena los datos del fármaco Racotumumab y permite realizar análisis sobre la información histórica, facilitando el proceso de toma de decisiones.

Palabras Claves: almacén de datos, CIM, extracción, transformación, carga, toma de decisiones.

ÍNDICE GENERAL

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

INTRODUCCIÓN

Desde las primeras investigaciones realizadas sobre el cáncer hasta la actualidad se ha tratado de buscar una cura contra esta enfermedad. A pesar de los esfuerzos realizados, todavía se hace difícil encontrarla, no obstante se han obtenido varios tratamientos que mejoran la calidad de vida de las personas que enfrentan este padecimiento. Tanto es así, que en los últimos años se ha visto una disminución significativa de la mortalidad por cáncer en algunos países desarrollados.

La necesidad de prever y comprender la notabilidad de las enfermedades crónicas y de intervenir contra ellas es una cuestión cada vez más imprescindible, por ello en el mundo existen numerosas instituciones dedicadas a su estudio, encontrándose la mayoría en países desarrollados. Los productos biológicos desarrollados en estos centros, no son solamente nuevos medicamentos, sino también son las herramientas de una transición más fundamental: la transformación del cáncer de una enfermedad rápidamente fatal en una condición crónica.

Cuba no ha estado exenta de los avances que existen en el mundo, ya que a partir del triunfo revolucionario se crearon numerosos centros científicos e investigativos que trabajan en programas de producción de alto valor agregado, especialmente en el campo biotecnológico y médico-farmacéutico. También, integró los conocimientos científicos al desarrollo tecnológico, un ejemplo lo constituye la Universidad de las Ciencias Informáticas (UCI)

La UCI es un centro de estudios que tiene como propósito formar profesionales comprometidos y altamente calificados en su trabajo. Está conformada por centros de desarrollo que facilitan la producción de software y servicios informáticos, siendo uno de estos el Centro de Tecnologías de Gestión de Datos (DATEC).

DATEC tiene como misión crear bienes y servicios informáticos relacionados con la gestión de datos, área del conocimiento que agrupa tanto a los sistemas de información, como a los denominados sistemas de inteligencia empresarial o de negocios, cuyo propósito fundamental es apoyar el proceso de toma de decisiones [1]. Dentro del centro se encuentra el departamento de Almacenes de Datos, el cual actualmente se encuentra realizando la integración de la información de un producto médico desarrollado por el CIM.

El CIM se encarga de obtener y producir aportes importantes a la economía del país. Para ello, se realizan los análisis pertinentes con el objetivo de probar el funcionamiento de los fármacos y luego se aplican en pacientes que padecen diferentes enfermedades. Para la prueba, aprobación e introducción de dichos

productos en el mercado, se realizan ensayos clínicos[1] (EC) con el objetivo de comprobar la seguridad y eficacia de los productos médicos que se desarrollan en dicho centro [2]. Uno de los fármacos que se desarrollan en este centro y sobre el cual se realizan los EC es Racotumumab, anticuerpo monoclonal anti-idiotípico conocido también como 1E10, que se utiliza para el tratamiento de cáncer de colon, mama y pulmón.

Para realizar el proceso de digitalización de la información de los productos del CIM, los especialistas elaboran varios modelos mediante el sistema informático EpiData, el cual genera ficheros con extensión "xls", estos datos no se encuentran integrados, lo que entorpece el manejo de la información por parte de los especialistas del CIM. Además, no permite realizar análisis estadísticos complejos entre uno o diferentes EC, lo que conlleva a la pérdida de datos útiles y valiosos.

Con el transcurso del tiempo se incrementa continuamente el volumen de información que se almacena en el CIM; debido al gran cúmulo de datos que se genera en cada uno de los EC que se gestionan. La entidad presenta problemas en los análisis y consultas que se realizan sobre la información recopilada, así como en la exposición de los indicadores relacionados con los ensayos.

Con el objetivo de almacenar y consultar la información de los EC en el CIM; se hace necesario integrar toda la información que se maneja en este centro, permitiendo realizar análisis certeros, tener acceso directo a todos los datos y apoyar el proceso de toma de decisiones.

Por todo lo anteriormente planteado surge como **problema de la investigación**: ¿Cómo estandarizar los datos del producto Racotumumab para su almacenamiento de forma homogénea?

La investigación tiene como **objeto de estudio** los almacenes de datos, enmarcado en el **campo de acción** los subsistemas de almacenamiento e integración del mercado de datos Racotumumab.

Para dar solución a la problemática que dio surgimiento al presente trabajo, se propone como **objetivo general**: implementar los subsistemas de almacenamiento e integración del mercado de datos Racotumumab para el almacén de datos de los EC del CIM que permita el almacenamiento homogéneo de la información.

Una vez realizado el análisis general, los **objetivos específicos** quedan desglosados en:

- Fundamentar la selección de la metodología y herramientas a utilizar en el desarrollo de los almacenes de datos.

[1] Ensayos clínicos: es un estudio experimental o investigación en seres humanos dirigida a descubrir o verificar los efectos clínicos, farmacológicos u otros efectos farmacodinámicos de un producto. Permiten determinar si un nuevo tratamiento o medicamento ayudará a prevenir, diagnosticar o tratar una enfermedad.

- Realizar el análisis y diseño de los subsistemas de almacenamiento e integración del mercado de datos Racotumumab.
- Realizar la implementación y pruebas de los subsistemas de almacenamiento e integración del mercado de datos Racotumumab.

Para alcanzar los objetivos propuestos se plantean las siguientes **tareas**:

1. Caracterización de la metodología, herramientas y tecnologías a utilizar en el desarrollo de almacenes de datos, permitiendo determinar cuáles se utilizarán durante la investigación.
2. Levantamiento de requerimientos para definir las necesidades del cliente.
3. Realización del perfilado de datos para garantizar la limpieza y calidad de los mismos.
4. Descripción de los casos de uso del mercado de datos para determinar cada una de las funcionalidades del sistema.
5. Definición de la arquitectura del mercado de datos identificando los subsistemas fundamentales que componen la solución.
6. Definición de los hechos, las medidas y las dimensiones del mercado de datos para identificar los elementos que forman parte del modelo lógico de datos.
7. Diseño del modelo lógico de datos para determinar los elementos que conforman el modelo físico.
8. Diseño del subsistema de integración para definir cómo se realizará la carga de las dimensiones y los hechos al mercado de datos.
9. Diseño de los casos de prueba para apoyar la realización de las pruebas.
10. Implementación del modelo físico de datos para garantizar la disponibilidad de las estructuras de la base de datos.
11. Implementación del subsistema de integración para poblar el mercado de datos.
12. Aplicación de las listas de chequeo para garantizar la correcta implementación de los subsistemas de almacenamiento e integración.
13. Aplicación de los casos de prueba para avalar la disponibilidad de cada uno de los elementos del mercado de datos.

El presente trabajo está estructurado en: introducción, tres capítulos, conclusiones, recomendaciones, referencias bibliográficas, bibliografía, anexos y glosario de términos.

Capítulo 1: Fundamentos teóricos de los almacenes de datos

En este capítulo se realiza un estudio de las tendencias actuales, así como la metodología y herramientas empleadas para la creación de un mercado de datos. Además, se recogen los conceptos, características, ventajas y desventajas relacionados con el tema para obtener un mayor entendimiento del negocio.

Capítulo 2: Análisis y diseño de los subsistemas de almacenamiento e integración del mercado de datos Racotumumab

En este capítulo se aborda la etapa de análisis, la cual constituye el punto de partida para comprender los requisitos de la organización, mediante los cuales se definen las estructuras de almacenamiento, se diseñan las reglas de extracción, transformación y carga de los datos, así como la arquitectura de información que regirá el desarrollo de la solución propuesta.

Capítulo 3: Implementación y prueba de los subsistemas de almacenamiento e integración del mercado de datos Racotumumab

En este capítulo se aborda todo lo referente a la implementación de la estructura física de la solución y la realización de los procesos de extracción, transformación y carga de los subsistemas de almacenamiento e integración del mercado de datos Racotumumab. Una vez concluida la implementación se da paso a una de las etapas más importantes en el ciclo de desarrollo de un software: las pruebas, las cuales permitirán encontrar y corregir no conformidades existentes, obteniéndose como resultado una aplicación con mayor calidad.

CAPÍTULO 1: FUNDAMENTOS TEÓRICOS DE LOS ALMACENES DE DATOS

Introducción

Hoy en día, la mayoría de las empresas requieren de un sistema capaz de realizar un buen análisis a partir de grandes volúmenes de información, con el objetivo de apoyar el proceso de toma de decisiones en dichas empresas. Esta necesidad conlleva al surgimiento de los conceptos de almacén de datos y mercado de datos.

En este capítulo se realiza un estudio de las tendencias actuales, así como la metodología y herramientas empleadas para la creación de un mercado de datos. Además, se recogen los conceptos, características, ventajas y desventajas relacionados con el tema para obtener un mayor entendimiento del negocio.

1.1 Almacenes de datos

El término almacén de datos (Data Warehouse, AD) fue introducido por Bill Inmon a principios de la década de los 90`, quien lo definió como: *"…una colección de datos orientado a temas, integrado, variable en el tiempo y no volátil para ayudar al proceso de toma de decisiones gerenciales"* [3].

Según Ralph Kimball, quien es una de las personalidades más influyentes en el área, propone otra definición al catalogarlo como *"…una copia de datos transaccionales, específicamente estructurados para la consulta y el análisis"* [4].

A partir de estos dos conceptos se ha creado una gran polémica sobre cual enfoque es el mejor, tratando de definir la alternativa adecuada para la correcta creación de un AD. Ninguna de las dos ideas está mal, simplemente cada una tiene un punto de vista diferente sobre lo que debe prevalecer a la hora de diseñar un AD. Durante la presente investigación se decidió ajustarse a la definición dada por Inmon por las características que este señala sobre los AD.

1.1.1 Características principales de los almacenes de datos

Los AD son la solución para obtener un sistema capaz de dar soporte al proceso de toma de decisiones estratégicas y tácticas, recibiendo datos de múltiples fuentes. Inmon plantea que existen cuatro características fundamentales que debe cumplir un almacén, estas son [3]:

- **Orientado a temas**: los datos presentes en la base de datos están organizados por materias o temas (paciente, personal médico, medicamentos) según los intereses de la empresa, de manera que todos los elementos de datos respectivos al mismo evento u objeto del mundo real queden unidos entre sí.

- **Integrado**: la integración implica que todos los datos provenientes de diversas fuentes deben ser consolidados en una instancia antes de ser agregados al AD, y deben por lo tanto ser analizados para asegurar su calidad y limpieza.
- **No volátil**: sólo existen dos tipos de operaciones que se llevan a cabo en un AD, cuando se cargan inicialmente los datos y cuando se acceden a él. Los datos almacenados no se modifican, ni actualizan, ni se eliminan nunca, una vez almacenado un dato, éste se convierte en información de sólo lectura, y se mantiene para futuras consultas.
- **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. El almacenamiento de los datos históricos, permitirá desarrollar pronósticos y análisis de tendencias y patrones, a partir de una base estadística de información.

1.1.2 Ventajas y desventajas de los almacenes de datos

Utilizar un AD implica numerosas ventajas tales como [5]:

- Permite el acceso a los usuarios finales de una gran variedad de datos.
- Permite obtener una base de datos histórica y clasificada por temas.
- Facilita la integración de información procedente de múltiples sistemas.

Aunque son muchas las ventajas que posee también presenta algunas desventajas como [5]:

- A lo largo de su vida puede suponer altos costos en cuanto a recursos humanos y tiempo de desarrollo.
- Corren el riesgo de quedar obsoletos debido a que en un futuro cambien los requisitos de los usuarios o surjan nuevas tecnologías que posibiliten la obtención de mejores resultados.

1.2 Mercado de datos

La eficiente manipulación de la información se ha convertido en un hecho indispensable y fundamental para la toma de decisiones de cualquier organización. Para lograr la misma, es necesario integrar los datos que han sido tomados de distintas fuentes para posteriormente almacenarlos de una forma globalmente aceptable. El problema está dado, en que al intentar integrar toda la información de la organización se requiere tratar con mucho más (fuentes de datos, plataformas, requisitos de usuarios, presupuesto, tiempo de desarrollo y personal). Una solución sería enfocarse en un área de negocio y crear un mercado de datos que satisfaga sus requisitos de información [6].

Los términos AD y mercado de datos (Data Mart, MD) se utilizan muchas veces indistintamente, según Kimball un MD es *"…una solución que, compartiendo tecnología con el almacén de datos (pero con contenidos específicos, volumen de datos más limitado y un alcance histórico menor), permita dar soporte a una empresa pequeña, un departamento o área de negocio de una empresa grande"*. Entre algunas de sus características se encuentran [3,4]:

- Se centran en los requisitos de los usuarios asociados a un departamento o área de negocio específico.
- El nivel de detalle de los AD no necesariamente debe coincidir con el de los MD asociados.
- Son más sencillos a la hora de utilizarlos y comprender sus datos, debido a que la cantidad de información que contienen es mucho menor que en los AD.

1.3 Tendencias actuales

En la actualidad con el desarrollo tecnológico de la informática, Cuba se ha centrado en la manera de usar los datos para dar soporte a la toma de decisiones, con el objetivo de crear sistemas capaces de aplicarse en el país y en el mundo. Aunque todavía no se han alcanzado los niveles de perfeccionamiento y apropiación que tienen otros países, se han realizado notables avances en el desarrollo y utilización de los MD. Algunos de los MD creados para el campo de la salud son:

- Para la Dirección de Salud en Cuba [7].
- Diseño del MD CIMAVAX EGF[2] para el CIM [8].
- Para el módulo visor de historias clínicas del Sistema Integral de Atención Primaria de la Salud [9].
- Para la Unidad Central de Cooperación Médica [10].

Algunos de estos mercados fueron desarrollados con el propósito de resolver una problemática similar a la de la presente investigación, pero ninguno responde exactamente a las necesidades y características de la misma. Por lo cual se decide desarrollar, los subsistemas de almacenamiento e integración del MD Racotumumab para el AD de los EC del CIM.

1.4 Metodologías para el desarrollo de un mercado de datos

Las metodologías son un conjunto de pasos definidos para lograr uno o varios objetivos. En el desarrollo de software esta surge ante la necesidad de utilizar una serie de procedimientos, técnicas, herramientas y soporte documental a la hora de desarrollar un producto [11]. Pretenden guiar a los desarrolladores al

[2] CIMAVAX EGF: vacuna para el factor del crecimiento epidérmico.

crear un nuevo software. Existen dos grandes enfoques para enfrentar la construcción de un AD: descendente (top-down) y ascendente (bottom-up), que se corresponden con las metodologías propuestas por Bill Inmon y Ralph Kimball respectivamente.

La principal diferencia que existe entre ambas tendencias está basada en la forma de enfrentar el problema. El enfoque descendente tiene como base un sistema de AD para toda la empresa y a partir de este se desarrollan los mercados de datos para cada departamento. A diferencia de la anterior, el enfoque ascendente tiene como base los distintos MD de los departamentos y a partir de estos se construye el almacén principal para toda la empresa. Este enfoque conduce a la obtención de la solución en un corto período de tiempo [12].

En determinadas circunstancias una metodología puede ofrecer ventajas sobre otras, sin embargo es posible que una combinación proporcione una mejor respuesta en dependencia de las necesidades del cliente. Algunos ejemplos de metodologías híbridas son:

- **Hefesto**: plantea que la construcción e implementación de un AD puede adaptarse muy bien a cualquier ciclo de vida de desarrollo de software, excepto para algunas fases en particular, donde las acciones que se han de realizar serán muy diferentes. Lo que busca, es entregar una primera implementación que satisfaga una parte de las necesidades, para demostrar las ventajas del AD y motivar a los usuarios [13].
- **Metodología para el diseño conceptual de almacenes de datos**: es presentada en la tesis de doctorado de Leopoldo Zenaido Zepeda Sánchez. Aporta como aspecto novedoso la incorporación de una serie de transformaciones para llevar un diagrama relacional a uno dimensional y así obtener las estructuras que conformarán el repositorio de datos [14].
- **Propuesta de metodología para el desarrollo de almacenes de datos en DATEC**: es presentada para estar más alineados a las tendencias y normas de la UCI, vincula el enfoque ascendente de la metodología Ciclo de vida de Kimball e incluye los casos de uso para guiar el proceso de desarrollo, planteado por el doctor Leopoldo Zenaido Zepeda en su tesis de doctorado. Además, a esta metodología se le agrega una etapa de prueba que permite comprobar la calidad de los productos que se desarrollan [12].

1.4.1 Metodología utilizada

En la presente investigación se definió como metodología a utilizar para el desarrollo de la solución la Propuesta de metodología para el desarrollo de almacenes de datos en DATEC, basado en las características que presenta. Se divide en ocho fases, las que se explicarán a continuación [12]:

Estudio preliminar y planeación: se realiza un estudio minucioso a la entidad cliente, el cual incluye un diagnóstico integral de la organización con el objetivo de determinar lo que se desea construir, y las condiciones que existen para su desarrollo y montaje.

Requisitos: se realizan entrevistas con el cliente para hacer el levantamiento de requisitos de información, requisitos funcionales y no funcionales de la solución.

Arquitectura: se definen las vistas arquitectónicas de la solución, los subsistemas y componentes, la seguridad, la comunicación y la tecnología a utilizar.

Diseño e implementación: se define el diseño de las estructuras de almacenamiento de datos, se diseñan los procesos de integración de datos como, el mapa lógico de datos, los cubos de procesamiento analítico en línea (On-Line Analytical Processing, OLAP) para la presentación de la información, así como el diseño gráfico de la aplicación definido por el cliente.

Prueba: se realizan las pruebas que validan la calidad del producto. Esta fase no es la única en la que se realizan pruebas durante el desarrollo del proyecto, en todas las fases hay actividades de aseguramiento de la calidad.

Despliegue: se realiza el despliegue de la aplicación en el entorno real y en correcto funcionamiento. Además se realiza la capacitación y transferencia tecnológica de la solución a los clientes.

Soporte y mantenimiento: puede realizarse a través de variados servicios, que pueden ser soporte en línea, vía telefónica, correo u otros. Se realizan las tareas de mantenimiento de la aplicación tan necesarias para este tipo de desarrollo y que garantizan el adecuado funcionamiento y crecimiento del AD.

Gestión del proyecto: esta fase se ejecuta a lo largo de todo el ciclo de vida del proyecto, es donde se controla, gestiona y chequea todo el desarrollo, los gastos, los recursos, las adquisiciones, los planes y cronogramas entre otras actividades relacionadas con la gestión de proyectos.

A pesar de que la metodología tiene ocho fases, en el presente trabajo solo se aplicaron las cinco primeras, es decir, hasta la fase de prueba. Las tres fases restantes (despliegue, gestión de proyecto y soporte y mantenimiento) no se encuentran dentro del alcance de la investigación debido a que estas son aplicadas por el personal del departamento de Almacenes de Datos.

1.5 Herramientas informáticas

Para realizar el modelado del negocio, diseño e implementación del sistema se hace necesario el uso de herramientas que permitan el desarrollo de la aplicación. A continuación se muestra un análisis de las herramientas que se utilizan para la realización de un MD.

1.5.1 Herramientas de modelado

Con el creciente desarrollo de la industria del software se hace cada vez más necesario la creación de herramientas para su modelado, que permitan mejorar y facilitar el trabajo de los desarrolladores. Existen varias herramientas de Ingeniería de Software Asistida por Computadora (Computer Aided Software Engineering, CASE) orientadas al Lenguaje Unificado de Modelado (UML), las cuales constituyen una ayuda para el desarrollo de programas informáticos. Algunas de ellas son:

- Rational Rose
- Visual Paradigm
- MagicDraw UML

Para establecer una especie de comparativa entre las herramientas de modelado anteriormente mencionadas con el objetivo de escoger la mejor opción para llevar a cabo el presente trabajo, se realizó un estudio del cual se obtuvieron los siguientes resultados [15,16]:

Tabla 1. Comparación de herramientas de modelado

	Rational Rose	**Visual Paradigm**	**MagicDraw UML**
Costo	Software de alto costo (licencia de €5,199.00).	Software de precio moderado ($99.00).	Software de precio moderado.
Facilidad de uso	Entorno gráfico no amigable para el usuario.	Herramienta potente, fácil de utilizar y con un entorno gráfico amigable.	Interfaz elegante e intuitiva con la mayor parte de las opciones accesibles con un solo clic.
Plataforma	No se encuentra disponible para Linux.	Multiplataforma.	Multiplataforma.

Rational Rose es una herramienta propietaria que tiene un elevado precio y presenta algunos problemas en cuanto a facilidad de uso y plataforma que soporta. Visual Paradigm y MagicDraw UML por otra parte son herramientas aceptadas por los usuarios, y según las características que se midieron están parejas en cuanto a resultados obtenidos. Sin embargo, desde hace ya unos años la UCI ha estado utilizando la herramienta de Visual Paradigm con fines académicos (no requiere de ningún costo), permitiendo así que

los usuarios se familiaricen con el sistema aportando mayor agilidad en el diseño, con MagicDraw a pesar de los beneficios que aporta, se tendría que brindar capacitación para poder trabajar con ella.

Finalmente se llegó a la conclusión de que para el desarrollo de este trabajo se utilizará la herramienta Visual Paradigm para UML en su versión 8.0 Enterprise Edition.

1.5.1.1 Visual Paradigm

El Visual Paradigm es una herramienta CASE de diseño UML que tiene como propósito apoyar el desarrollo del software. Alguna de sus características son [17]:

- Diseño centrado en casos de uso y enfocado al negocio que genera un software de mayor calidad.
- Uso de un lenguaje estándar común a todo el equipo de desarrollo facilitando la comunicación.
- Multiplataforma.
- Exportación de imágenes en formato jpg, png y svg.

1.5.2 Sistemas Gestores de Base de Datos

Un Sistema Gestor de Base de Datos (SGBD) es un conjunto de programas no visibles al usuario final que proporcionan una interfaz entre el usuario, las aplicaciones y la base de datos. Además permiten manipular los datos garantizando la privacidad, integridad y seguridad de los datos [18]. Algunos SGBD son:

- MySQL
- Oracle
- PostgreSQL

MySQL, Oracle y PostgreSQL son SGBD muy potentes y proporcionan muchas de las funciones estándar que el programador necesita para lograr grandes avances. Con el objetivo de escoger uno para el desarrollo del producto se realizó una comparación entre ellos, de la cual se obtuvieron los siguientes resultados [19, 20, 21]:

Tabla 2. Comparación de sistemas gestores de base de datos

	Oracle	**PostgreSQL**	**MySQL**
Costo	Software de alto costo (según formación, versiones y licencias de personal).	Su código fuente está disponible, sin costo alguno. Licencia liberada, que permite usarlo, modificarlo y distribuirlo de forma gratuita para cualquier fin.	Software de precio moderado.

Despliegue	Requiere de un procesador Pentium 166 MHz o superior, mínimo 128 Mb RAM y 1 Gb de disco duro.	Requiere como mínimo para instalarlo de 8 Mb de RAM, 30 Mb de espacio en disco para el código fuente, 5 Mb de espacio en disco.	Posee un bajo consumo, por lo que puede ser ejecutado en una máquina con escasos recursos sin ningún problema.
Velocidad	Muy veloz (puede llegar a insertar miles de filas por segundo).	Por lo general es lento en cuanto a inserciones y actualizaciones de datos.	Muy veloz (para 1.000.000 de inserciones con datos aleatorios en el mismo equipo, lo hace en un 25% menos de tiempo que PostgreSQL).
Estabilidad	Su estabilidad se puede ver a través del grado de satisfacción de la funcionalidad del producto respecto a versiones anteriores y la agilidad del soporte técnico Oracle para solucionar problemas.	Usa multiprocesos en vez de multihilos para garantizar la estabilidad del sistema.	Todos los problemas reportados y conocidos se arreglan en la última versión, con las excepciones listadas en las secciones de problemas y que están relacionados con dificultades de diseño.

La UCI se encuentra inmersa en un proceso de migración, vinculada principalmente a las nuevas tendencias tecnológicas, que consiste en la utilización de herramientas libres, por lo que claramente esa restricción dejaría completamente fuera a Oracle, debido a que es una herramienta propietaria altamente costosa. Por otra parte, MySQL es una herramienta que supera a PostgreSQL en todo lo referente a velocidad, en cambio, PostgreSQL gana en otros campos como a la hora de mantener la integridad referencial, en estabilidad y facilidad de uso.

Finalmente se llegó a la conclusión de que para el desarrollo de este trabajo se utilizará como SGBD la herramienta PostgreSQL en su versión 9.1.2.

1.5.2.1 PostgreSQL

PostgreSQL es un SGBD objeto-relacional, distribuido bajo licencia Berkeley Software Distribution (BSD) y con su código fuente disponible libremente. Este sistema utiliza un modelo cliente/servidor y usa multiprocesos en vez de multihilos para garantizar la estabilidad del sistema. Por lo cual un fallo en uno de los procesos no afecta al resto y el sistema continúa funcionando [22].

Entre las características generales que posee PostgreSQL se encuentran:

- Posibilita la realización de copias de seguridad.
- Es multiplataforma, está disponible para Linux, UNIX en todas sus variantes y Windows.

De las características de desarrollo más importantes soportadas por PostgreSQL se encuentran:

- Posee funciones y procedimientos almacenados en numerosos lenguajes de programación.
- Soporta almacenamiento de objetos binarios grandes (gráficos, videos, sonido).
- Posee diversas interfaces de programación de aplicaciones (Application Programming Interfaces, APIs) que facilitan el uso de la programación en C/C++, java, .net, python, php y ruby.

Las herramientas de administración de bases de datos actúan de interfaz entre la base de datos, el usuario y las aplicaciones que las utilizan, controlan la creación, el mantenimiento y el uso de las bases de datos de una organización y de sus usuarios finales.

Algunas de las herramientas más conocidas de administración para el SGBD PostgreSQL son:

- EMS SQL Manager
- PgAdmin
- PgAccess

Con el objetivo de seleccionar la mejor herramienta para la administración de PostgreSQL y poder desarrollar la solución requerida, se realizó un estudio de las herramientas antes mencionadas, obteniéndose como resultado la siguiente tabla comparativa [23, 24]:

Tabla 3. Comparación de administradores para el SGBD PostgreSQL

	EMS SQL Manager	**PgAccess**	**PgAdmin**
Costo	Software de alto costo (puede estar dado en $135.00).	Software libre (su código fuente se encuentra disponible para todos sin costo alguno).	Software libre (su código fuente se encuentra disponible para todos sin costo alguno).
Plataforma	Solo puede ser ejecutado en Windows.	Utiliza las librerías tcl/tk que le permite correr en cualquier plataforma a la que haya sido portado tcl/tk como Linux y Unix.	Usa la librería gráfica wxWidgets, que permite ejecutarla en múltiples plataformas como: Linux, FreeBSD, Solaris y Windows.
Facilidad de uso	Interfaz visual que brinda facilidades a los usuarios que interactúan con la misma (las barras de herramientas).	Ambiente visual pobre y poco amigable.	Interfaz gráfica que soporta todas las características de PostgreSQL, permitiendo que se administre la base PostgreSQL con mayor facilidad.

Seguridad	Las etiquetas de seguridad se aplican a los datos con el fin de protegerlo. Se conceden a los usuarios para que puedan acceder a los datos protegidos. Cuando un usuario intenta acceder a datos protegidos, su etiqueta de seguridad se compara con la etiqueta de seguridad que protege los datos. Si la etiqueta de seguridad del usuario está bloqueada, el usuario no puede acceder a los datos.	No se le brinda soporte constantemente, por lo que carece de seguridad.	Seguro (permite encriptar mediante SSL para mayor seguridad).

EMS SQL Manager es una herramienta propietaria que tiene un elevado precio y solo puede ser ejecutada en una única plataforma. PgAcces el mayor inconveniente que tiene es que posee una interfaz gráfica compleja, lo que podría ser un problema a la hora de interactuar con esta. Por otra parte, la robustez de la combinación PostgreSQL-PgAdmin permite obtener resultados con marcadas características de rapidez y flexibilidad, al permitir al usuario la modificación de código de acuerdo a sus necesidades.

Finalmente se llegó a la conclusión de que para el desarrollo de este trabajo se utilizará como herramienta de administración PgAdmin en su versión 1.14.1.

1.5.2.2 PgAdmin

Es una herramienta de código abierto para la administración de bases de datos PostgreSQL. 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 con todas las características de PostgreSQL y facilita la administración. Algunas de sus principales características son [25]:

- Multiplataforma.
- Diseñado para múltiples versiones de PostgreSQL y derivados.
- Amplia documentación.
- Acceso a todos los objetos de PostgreSQL.
- Interfaz multilingüe.

1.5.3 Herramientas para el perfilado de datos

El perfilado de los datos es una de las primeras tareas a realizar en el proceso de calidad de datos y consiste en realizar un análisis inicial sobre los datos de las fuentes, con el propósito de empezar a conocer su estructura, formato y nivel de calidad. Para llevar a cabo este proceso se utilizaron el Microsoft Excel 2010 y DataCleaner en su versión 1.5.4.

1.5.3.1 DataCleaner

DataCleaner es una aplicación de código abierto para el perfilado, la validación y comparación de datos. Estas actividades ayudan a administrar y supervisar la calidad de los datos con el fin de garantizar que la información sea útil y aplicable a la situación del negocio. Es una aplicación fácil de usar que genera sofisticados informes y gráficos que permiten a los usuarios determinar el nivel de calidad de los datos. Es utilizada, además, para identificar y analizar la estructura del origen de datos y combinar resultados y gráficos, creando vistas fáciles de interpretar para evaluar la calidad de los mismos. Algunas de las características que presenta son [26, 27]:

- Presenta una interfaz gráfica amigable y fácil de usar.
- Genera informes y gráficos que permiten determinar el nivel de calidad de los datos.
- Soporta tipos de ficheros como: hojas de cálculo y archivos xml.

1.5.4 Herramientas para los procesos de extracción, transformación y carga

Los procesos de extracción, transformación y carga (ETL), son una etapa importante en el desarrollo de un MD ya que organizan el flujo de datos entre diferentes sistemas de una organización y aportan los métodos y herramientas necesarias para mover datos desde múltiples fuentes a un AD, reformatearlos, limpiarlos y cargarlos. A continuación se explican brevemente los procesos de ETL [14]:

- Extracción: en este paso se obtiene la información de las diferentes fuentes. Los datos generalmente se encuentran en formatos distintos, la extracción deja los datos en un formato listo para transformarlos.
- Transformación: luego de extraer la información, se prepararan los datos para integrarlos en el AD. Por lo cual se realizan una serie de actividades como: limpieza de datos, estandarización de formatos e integración de datos.
- Carga: después de ser transformados los datos, se realiza la carga en el AD.

Algunas de las herramientas de integración de datos son:

- Talend

- Pentaho Data Integration
- Informática PowerCenter

La informatización como parte del avance tecnológico trae consigo a gran escala la utilización de modernos medios de integración de la información, entre los que se encuentran Talend, Pentaho Data Integration e Informática PowerCenter. Para escoger una de estas herramientas para el desarrollo de la solución se realizó una especie de comparativa entre las mismas, donde se obtuvieron los siguientes resultados [28]:

Tabla 4. Comparación de herramientas que realizan el proceso de integración de datos

	Informática PowerCenter	**Pentaho**	**Talend**
Costo	Herramienta propietaria con precio muy elevado debido al alto costo de las licencias de mantenimiento, soporte, capacitación y consultoría.	Herramienta de código abierto con un buen precio (los productos comerciales de software libre suelen ser de uso gratuito, y solo el soporte, la formación y la consultoría son los servicios que se tiene que pagar).	Herramienta de código abierto con un buen precio (ya que los productos comerciales de software libre suelen ser de uso gratuito, y solo el soporte, la formación y la consultoría son los servicios que se tiene que pagar).
Facilidad de uso	Posee una interfaz gráfica potente, aunque requiere de cierto entrenamiento para hacer uso de todas sus capacidades.	Dispone de entrenamiento que puede ser proporcionado en línea o dentro de la comunidad Pentaho.	Dispone de una interfaz gráfica de tipo add-on dentro del entorno Eclipse.
Despliegue	Requiere de dos CPU's con 1 Gb de RAM para la Standard Edition.	Requiere de 1Ghz de CPU con 512 Mb de RAM.	Requiere de 1GHz CPU con 512 Mb RAM.
Conectividad	Permite conectarse a múltiples bases de datos, ficheros excel y servicios web.	Permite conectarse a múltiples bases de datos, ficheros excel y servicios web.	Permite conectarse a múltiples bases de datos, ficheros excel y servicios web (requiere y depende de controladores java para dichas conexiones).

Las herramientas Informática y Pentaho son consideradas productos muy completos. A pesar de que la Informática tiene una gama mucho más amplia de productos en comparación con Pentaho, es muy cara. Talend posee una interfaz muy compleja, lo que dificulta la interacción entre los usuarios y el sistema.

Pentaho por otra parte, es una herramienta que ha demostrado que puede manejar pequeñas y grandes soluciones de desarrollo de integración, además de tener una interfaz gráfica fácil de comprender. Finalmente se llegó a la conclusión de que para el desarrollo de este trabajo se utilizará como herramienta para el proceso de integración de datos Pentaho Data Integration en su versión 4.2.1.

1.5.4.1 Pentaho Data Integration

Pentaho Data Integration (PDI) permite extraer la información de las diferentes fuentes, transformar la información a través de un modelo dimensional y cargar los resultados de la transformación en una base de datos tipo AD, para que luego pueda ser consultada y analizada a través de herramientas para desarrollar reportes especializados las cuales Pentaho también posee [29].

PDI, es una herramienta libre muy poderosa y una de las más antiguas y utilizadas por los usuarios. Sus principales ventajas son [30]:

- Multiplataforma.
- Uso de tecnologías estándar: java, XML, javascript.
- Incluye cuatro herramientas:
    - Spoon: para diseñar transformaciones de ETL usando el entorno gráfico.
    - PAN: para ejecutar transformaciones diseñadas con spoon.
    - CHEF: para crear trabajos.
    - Kitchen: para ejecutar trabajos.
- Basada en dos tipos de objetos: transformaciones (colección de pasos en los procesos de ETL) y trabajos (colección de transformaciones).
- Soporta diferentes lenguajes de base de datos como por ejemplo: MySQL y Postgres.

1.6 Modelo multidimensional

Para el diseño de un MD, se emplea la representación de un modelo multidimensional, que es una técnica de diseño lógico que busca presentar los datos en un estándar [31]. Proporciona dos conceptos fundamentales: hecho y dimensión.

Las tablas de dimensiones constituyen las características de un hecho y permiten el análisis de los datos desde varias perspectivas. Representan los elementos de análisis, proporcionándole al usuario el filtrado y manipulación de la información almacenada en la tabla de hechos [31].

Los hechos por otra parte, son operaciones que se producen en el negocio. Al insertar un hecho se define su llave primaria como la composición de las llaves primarias de las tablas de dimensiones relacionadas a este y poseen medidas de procesos del negocio [31].

Las medidas son atributos numéricos que representan el comportamiento del negocio relativo a la dimensión. Constituyen los valores que son analizados.

El modelo multidimensional incluye tres variantes de modelación, las cuales son [32]:

- Esquema en estrella: está formado por una tabla de hecho y una o más tablas de dimensiones. La tabla de hecho es la única que tiene múltiples uniones que la conectan con las tablas de dimensiones a través de la llave foránea.
- Esquema en copo de nieve: es derivado del esquema en estrella. La tabla de hecho deja de ser la única que se relaciona con otras tablas del esquema y las dimensiones de análisis se representan entre las tablas de dimensiones normalizadas. En la estructura dimensional normalizada, la tabla que representa el nivel base de la dimensión, es la que realiza la unión directa con la tabla de hecho. Existen dos tipos de representación de este esquema; el copo de nieve completo, en el que todas las tablas de dimensión están normalizadas, y el copo de nieve parcial, en el que sólo se lleva a cabo la normalización de algunas tablas de dimensión.
- Esquema constelación de hechos: es la generalización de los esquemas en estrella o copo de nieve con la inclusión de distintas tablas de hechos que comparten todas o algunas tablas de dimensiones.

El tipo de esquema que se utiliza en la presenta investigación es constelación de hechos, debido a que existen varios hechos que comparten las mismas dimensiones. Un ejemplo es: hech_inclusion_evaluacion_inicial_sclc y hech_inclusion_inicial_NSCLC_int, los cuales comparten la dimensión dim_terapia_empleada.

1.7 Tecnologías para el desarrollo de un mercado de datos

Los sistemas OLAP son bases de datos orientadas al procesamiento analítico. Este análisis suele implicar generalmente, estudios complejos de grandes volúmenes de datos para llegar a extraer algún tipo de información útil: tendencias de ventas y patrones de comportamiento de los consumidores. Hoy en día para lograr el almacenamiento de los datos en las grandes bases de datos y especialmente en los AD, se utilizan varias tecnologías OLAP, dentro de las cuales se destacan [33]:

- **Procesamiento Analítico Relacional en Línea** (ROLAP): almacena los datos en un motor relacional. Utiliza una arquitectura de tres niveles que permite el análisis de una enorme cantidad de datos:
    - El nivel de base de datos usa bases de datos relacionales para el manejo, acceso y obtención de los datos.
    - El nivel de aplicación es el motor que ejecuta las consultas multidimensionales de los usuarios.
    - El motor ROLAP se integra con niveles de presentación, a través de los cuales los usuarios realizan los análisis OLAP.
- **Procesamiento Analítico Multidimensional en Línea** (MOLAP): esta implementación OLAP almacena los datos en una base de datos multidimensional. Provee excelente rendimiento, compresión de los datos y tiene un buen tiempo de respuesta.
- **Procesamiento Analítico Híbrido en Línea** (HOLAP): es una solución que incluye las implementaciones anteriores (MOLAP y ROLAP). Las agregaciones de los datos son almacenadas en una estructura multidimensional usada por MOLAP y la base de datos fuente en una base de datos relacional. Los cubos almacenados como HOLAP son más pequeños que los MOLAP y responden más rápido que los ROLAP.

En la presente investigación se seleccionó el modo de almacenamiento ROLAP, teniendo en cuenta que el SGBD PostgreSQL no soporta el almacenamiento multidimensional, solamente el relacional, ya que PostgreSQL es un SGBD de código abierto y multiplataforma, a diferencia de los SGBD que dan soporte al almacenamiento multidimensional existentes en la actualidad, los cuales no están en correspondencia con las políticas de desarrollo de la UCI y el país, puesto que son software propietario.

Conclusiones del capítulo

Una vez expuestos los principales conceptos asociados a la tecnología de AD, se propone como solución a la necesidad de estandarizar y centralizar la información, el desarrollo de un MD que permita el almacenamiento, procesamiento y obtención de los datos históricos asociados a los EC del producto Racotumumab. La metodología de desarrollo seleccionada permite guiar el proceso de construcción del sistema durante cada una de las etapas, orientando el trabajo hacia el logro de los resultados esperados. Las herramientas definidas para el desarrollo de la solución propuesta cumplen con las políticas de migración a software libre.

CAPÍTULO 2: ANÁLISIS Y DISEÑO DE LOS SUBSISTEMAS DE ALMACENAMIENTO E INTEGRACIÓN DEL MERCADO DE DATOS RACOTUMUMAB

Introducción

En el proceso de construcción de un MD la etapa de análisis constituye el punto de partida para comprender los requisitos de la organización, mediante los cuales se definen las estructuras de almacenamiento, se diseñan las reglas de extracción, transformación y carga de los datos, así como la arquitectura de información que regirá el desarrollo de la solución propuesta. 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 que permita identificar las necesidades de información de los especialistas del CIM.

2.1 Análisis del negocio

Las necesidades de información constituyen especificaciones que los especialistas precisan para darle cumplimiento a sus tareas internas. Su conocimiento por parte del equipo de desarrollo tributa a un correcto análisis y diseño de los procesos de negocio. La participación de los usuarios durante todo el ciclo de vida del producto permite conocer si los resultados alcanzados son satisfactorios o no, según sus necesidades.

Actualmente en el CIM, la gestión de los EC se realiza mediante el sistema informático EpiData, el cual genera ficheros en diferentes formatos, como por ejemplo excel. La presente investigación incluye tres EC cerrados, donde la información es cargada una sola vez. Estos son facilitados por los especialistas del producto Racotumumab en el CIM junto con sus respectivos protocolos, a partir de los cuales es posible comprender varios aspectos que pueden llegar a ser significativos para la investigación. Los ensayos son: EC de cáncer de pulmón de células pequeñas (SCLC), EC de cáncer de pulmón de células no pequeñas internacional (NSCLC Internacional) y EC de cáncer de pulmón de células no pequeñas compacional (NSCLC Compacional). Cada EC tiene un aproximado de nueve ficheros, donde se tiene almacenada toda la información personal y pertinente a los resultados del tratamiento en los pacientes.

Luego de varias entrevistas realizadas a los especialistas del CIM, se procede a realizar la agrupación de sus necesidades en áreas de información, con el objetivo de garantizar un nivel organizativo que orienta la realización de las actividades hacia el cumplimiento de las metas establecidas. La información fue agrupada por tipo de información, quedando de la siguiente manera:

- EC SCLC
- EC NSCLC Compacional

- EC NSCLC Internacional

2.2 Especificación de requerimientos

El análisis de requisitos constituye una de las fases más importantes en la construcción de un MD. Una correcta descripción de los requisitos del sistema permite llegar a un acuerdo entre los clientes y desarrolladores sobre qué debe y qué no debe hacer el sistema. En dicha fase se definen los requisitos de información, funcionales y no funcionales del sistema, partiendo de las necesidades reales de los usuarios.

2.2.1 Requisitos de información

Los requisitos de información describen la información y los datos que el sistema debe proveer. Estos se definen a partir de las necesidades de información identificadas en el negocio, que permiten el análisis del comportamiento de los indicadores a medir según los objetivos y metas de la organización [34]. Luego de analizado el proceso que se lleva a cabo en la obtención de los datos del producto 1E10 para el CIM se capturaron los requisitos para mantener disponible la información de la cantidad de pacientes:

1. Incluidos y evaluados inicialmente en el EC SCLC por tiempo, localización, sexo, edad, provincia, hospital, etapa enfermedad, clasificación tnm, examen de laboratorio, respuesta clínica, examen físico, preguntas, técnica de imagenología y radiología, análisis anatomopatológico y terapia empleada.
2. Evaluados durante el tratamiento 1[3] en el EC SCLC por tiempo, localización, provincia, hospital, signos vitales, examen físico y preguntas.
3. Evaluados durante el tratamiento 2[4] en el EC SCLC por tiempo, localización, provincia, hospital, examen de laboratorio, signos vitales, examen físico y preguntas.
4. Evaluados durante el tratamiento 3[5] en el EC SCLC por tiempo, localización, provincia, hospital, signos vitales, examen físico, preguntas, técnica de imagenología y radiología y respuesta al tratamiento.
5. Evaluados durante el tratamiento 4[6] en el EC SCLC por tiempo, localización, provincia, hospital, examen de laboratorio, signos vitales, examen físico, preguntas, técnica de imagenología y radiología, respuesta al tratamiento y tipo de evaluación.

[3] Tratamiento 1: abarca las inmunizaciones 1, 2, 3, 5, 8, 10, 11 y 13 del EC SCLC.
[4] Tratamiento 2: abarca las inmunizaciones 4, 7 y 14 del EC SCLC.
[5] Tratamiento 3: abarca las inmunizaciones 6 y 15 del EC SCLC.

6. De acuerdo a la seguridad en el EC SCLC por tiempo, localización, provincia, hospital y tipo, grado y causalidad del evento adverso.
7. Que interrumpieron el tratamiento en el EC SCLC por tiempo, localización, provincia, hospital, preguntas, sobrevida, causa de interrupción y causa de muerte.
8. Incluidos y evaluados inicialmente en el EC NSCLC Internacional por tiempo, localización, hospital, edad, sexo, clasificación tnm, tipo histológico, examen de laboratorio, respuesta clínica, examen físico, preguntas, técnica de imagenología y radiología, país, raza, estadío, ecog, karnofsky, hábito tóxico, análisis anatomopatológico y terapia empleada.
9. Evaluados durante el tratamiento 1[7] en el EC NSCLC Internacional por tiempo, localización, signos vitales, examen físico y preguntas.
10. Evaluados durante el tratamiento 2[8] en el EC NSCLC Internacional por tiempo, localización, examen de laboratorio, signos vitales, examen físico, preguntas, técnica de imagenología y radiología, respuesta al tratamiento y tipo de evaluación.
11. De acuerdo a la seguridad en el EC NSCLC Internacional por tiempo, localización, tipo de evaluación, tipo, grado y causalidad del evento adverso.
12. Que interrumpieron el tratamiento en el EC NSCLC Internacional por tiempo, localización, hospital, preguntas, causa de interrupción y causa de muerte.
13. Que terminaron con el EC NSCLC Internacional por tiempo, localización, hospital, preguntas y sobrevida.
14. Evaluados de acuerdo a los datos generales en el EC NSCLC Compacional por tiempo, localización, hospital, edad, sexo, clasificación tnm, tipo histológico, preguntas, estadío, karnofsky y estado por la OMS.
15. Evaluados durante el tratamiento en el EC NSCLC Compacional por tiempo, localización, preguntas y dosis.
16. De acuerdo a la seguridad en el EC NSCLC Compacional por localización, tipo y grado del evento adverso.

[6] Tratamiento 4: abarca las inmunizaciones 9 y 12 del EC SCLC.
[7] Tratamiento 1: abarca las inmunizaciones 1, 2, 3, 4, 5, 7, 8, 10, 11, 13 y 14 del EC NSCLC Internacional.
[8] Tratamiento 2: abarca las inmunizaciones 6, 9, 12, 15 y el seguimiento durante los meses 3, 6, 9 y 12 del EC NSCLC Internacional.

17. Evaluados clínicamente en el EC NSCLC Compacional por tiempo, localización, respuesta clínica y técnica de imagenología y radiología.
18. Evaluados de acuerdo al laboratorio clínico en el EC NSCLC Compacional por localización y examen de laboratorio.
19. Evaluados según resultados de imagenología en el EC NSCLC Compacional por tiempo, localización y técnica de imagenología y radiología.
20. Evaluados finalmente en el EC NSCLC Compacional por tiempo, localización, respuesta clínica, preguntas, causa interrupción y estado al concluir tratamiento.

2.2.2 Requisitos funcionales

Para lograr satisfacer las necesidades del cliente definidas con anterioridad, se hace necesario identificar las funcionalidades que debe poseer el sistema. Estas funcionalidades son conocidas como requisitos funcionales de la aplicación [12].

Los requisitos funcionales identificados fueron agrupados por subsistemas, quedando de la siguiente manera:

Subsistema de almacenamiento:

1. Conservar los datos del EC del producto Racotumumab accesibles en el almacén por 5 años.
2. Crear un servidor de copias de respaldo, el cual debe encontrarse en el departamento de investigaciones clínicas del CIM. Se realizará una copia de respaldo luego de tener almacenada la información en el almacén.

Subsistema de integración:

3. La información del producto recogida se corresponderá con la de los ficheros entregados al equipo de desarrollo.
4. La información cargada tendrá un peso en disco de 7.71 MB.
5. La información recogida acerca del producto será obtenida una sola vez a través de un dispositivo de almacenamiento electrónico o por correo electrónico y será proporcionada por el especialista del CIM.
6. Los datos de los EC del producto Racotumumab que serán cargados en el almacén, están comprendidos entre los años 1922 y 2011.
7. Se cargarán 25 ficheros en formato .xls, proporcionados por el especialista del CIM.

8. Extraer datos de la fuente.
9. Realizar la transformación y carga de los datos.

2.2.3 Requisitos no funcionales

Describen las propiedades y cualidades que debe tener el producto, estos determinan además como debe comportarse el sistema o la aplicación una vez finalizado. Representan las características del producto [12]. Luego de un profundo análisis, los requisitos no funcionales detectados se agruparon por categorías según las características del negocio. A continuación se muestra un ejemplo de requisitos no funcionales de restricciones del diseño (el resto de los requisitos no funcionales se encuentran en el artefacto "Especificación de requisitos de software"):

Restricciones del diseño

1. Estandarizar la estructura de los elementos definidos en el almacén. Las estructuras del AD se nombrarán de una manera estándar teniendo en cuenta el tipo de estructura que se maneje. (En el anexo 5 se muestra una tabla con la estructura de los datos en el almacén para lograr un mejor entendimiento por parte de los desarrolladores).
2. Utilizar el SGBD definido durante la investigación. El gestor de base de datos que se utilizará es PostgreSQL en su versión 9.1.2 y como interfaz de administración de dicho gestor PgAdmin III en su versión 1.14.1.
3. Utilizar la herramienta de integración de datos definida durante la investigación. Para el proceso de integración de datos se usará la herramienta Pentaho Data Integration en su versión 4.2.1.

2.3 Reglas del negocio

Son declaraciones que definen o restringen algún aspecto del negocio. Estas pretenden hacer valer la estructura del negocio, controlar o influir en el comportamiento de la empresa [35]. A continuación se muestran algunas de las reglas del negocio identificadas (el resto de las reglas se encuentran en el artefacto "Reglas del Negocio").

Según las reglas de variables:

- En caso de no tener la variable edad, se calcula restando la fecha de inclusión menos la fecha de nacimiento.

Según las reglas de almacenamiento:

- El tipo de dato que toma la variable nombre del paciente es varchar con tamaño cinco.

- El tipo de dato que toma la variable número de inclusión del paciente es integer con tamaño dos.

Según las reglas de transformación:

- Las variables que son fechas se toman con el siguiente formato dd/mm/aaaa.
- Los campos que son fechas y aparezcan vacíos o tengan un valor fuera de los rangos establecidos para día, mes y año se remplazan por la fecha 31/12/2015.
- En la variable fecha, el día solo puede tomar un valor comprendido entre 1 y 31, por otra parte el mes solo puede tomar un valor comprendido entre 1 y 12, donde 1 corresponde al mes de enero, 2 a febrero, 3 a marzo, 4 a abril, 5 a mayo, 6 a junio, 7 a julio, 8 a agosto, 9 a septiembre, 10 a octubre, 11 a noviembre y 12 a diciembre. En caso del año la fecha tiene que estar entre los años 1922 y 2011 para ser válida.
- En caso de que la variable localización en la fuente de datos sea “nula” o “HIL D” o “pulmón”, el valor que le corresponde es Pulmón.
- En caso de que la variable provincia proveniente de la fuente sea “MZ” el valor que le corresponde es MT, el cual representa a la provincia de Matanzas.

2.4 Diagrama de casos de uso del sistema

Durante la fase de análisis y diseño de un MD se definen los casos de uso del sistema. Un caso de uso constituye una secuencia de interacciones que se desarrollan entre un sistema y sus actores, en respuesta a un evento que inicia un actor sobre el propio sistema [36]. Los requisitos de información y requisitos funcionales identificados son agrupados en casos de uso de información y funcionales respectivamente, los cuales aportan una apreciación más acertada del funcionamiento del sistema. A continuación se presenta un fragmento del diagrama obtenido (el diagrama completo se encuentra en el artefacto “Especificación de casos de uso” y en el anexo 1 del presente trabajo).

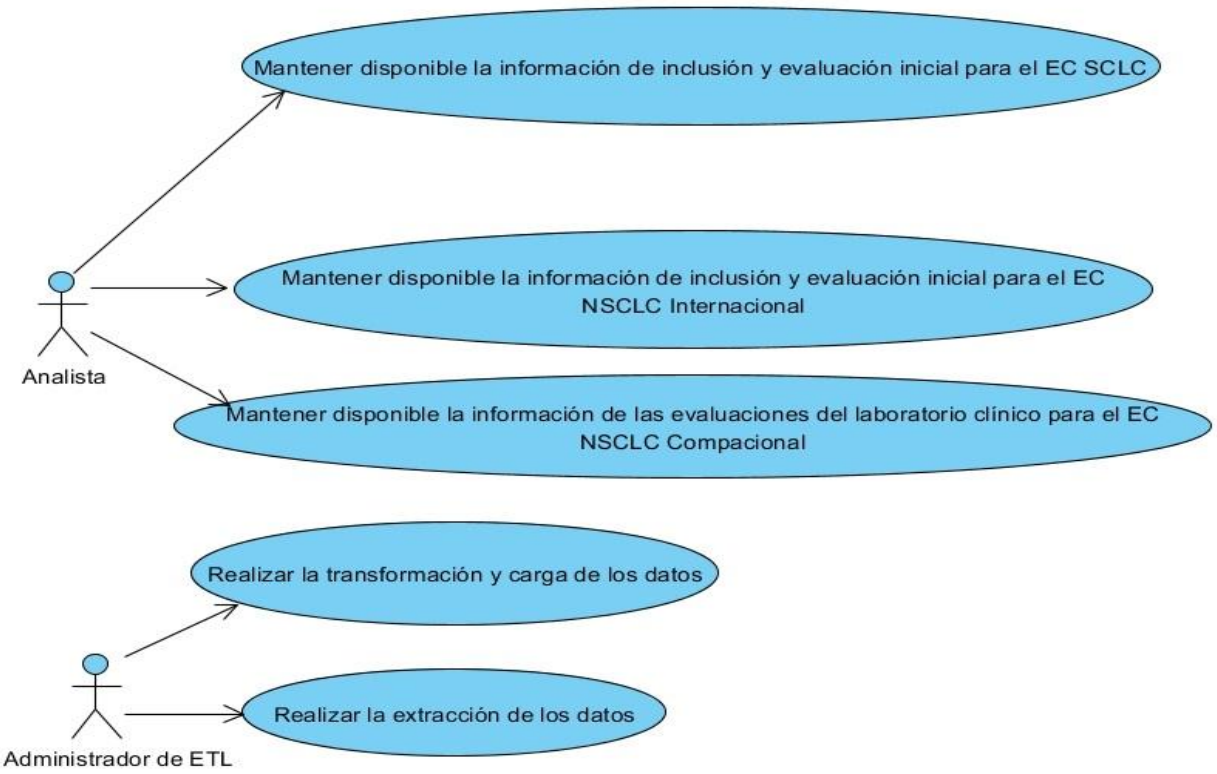


Figura 1. Fragmento del diagrama de casos de uso del sistema

Dependiendo de las necesidades del cliente, los datos que se manejan se agruparon mediante el criterio tipo de información, en casos de uso informativos, donde se mantiene disponible la información de:

- Inclusión y evaluación inicial en el EC SCLC.
- Evaluación durante el tratamiento 1 en el EC SCLC.
- Evaluación durante el tratamiento 2 en el EC SCLC.
- Evaluación durante el tratamiento 3 en el EC SCLC.
- Seguimiento y evaluación durante el tratamiento 4 en el EC SCLC.
- Seguridad en el EC SCLC.
- Interrupción del tratamiento en el EC SCLC.
- Inclusión y evaluación inicial en el EC NSCLC Internacional.
- Evaluación durante el tratamiento 1 en el EC NSCLC Internacional.
- Evaluación durante el tratamiento 2 en el EC NSCLC Internacional.
- Seguridad en el EC NSCLC Internacional.

- Interrupción en el EC NSCLC Internacional.
- Terminación del ensayo para NSCLC Internacional.
- Datos generales en el EC NSCLC Compacional.
- Evaluación durante el tratamiento en el EC NSCLC Compacional.
- Seguridad en el EC NSCLC Compacional.
- Evaluaciones clínicas en el ensayo NSCLC Compacional.
- Evaluaciones del laboratorio clínico en el ensayo NSCLC Compacional.
- Evaluación de los resultados de imagenología en el EC NSCLC Compacional.
- Evaluación final en el EC NSCLC Compacional.

Por otra parte los CU funcionales identificados agrupan los requisitos funcionales relacionados con la ejecución de las operaciones de ETL que serán realizadas a las bases del producto 1E10 perteneciente al CIM, en este caso son:

- Realizar la extracción de los datos.
- Realizar la transformación y carga de los datos.

2.5 Arquitectura de la solución

La arquitectura de un sistema informático es la organización de los componentes y las relaciones entre ellos. En la cual, se deben tener en cuenta los requisitos del sistema y las restricciones a las que está sujeto.

A la hora de abordar la arquitectura de un MD se debe tener en cuenta la forma de representar el origen y estructura global de los datos, la comunicación y los procesos. La arquitectura de la presente investigación consta de dos niveles:

- Subsistema de integración: incluye los procesos que permiten que los datos sean extraídos de las fuentes, transformarlos e integrarlos.
- Subsistema de almacenamiento: base de datos que contiene las tablas de dimensiones y hechos cargadas a través de los procesos de ETL.

En la presente investigación la arquitectura base queda compuesta de la siguiente manera:

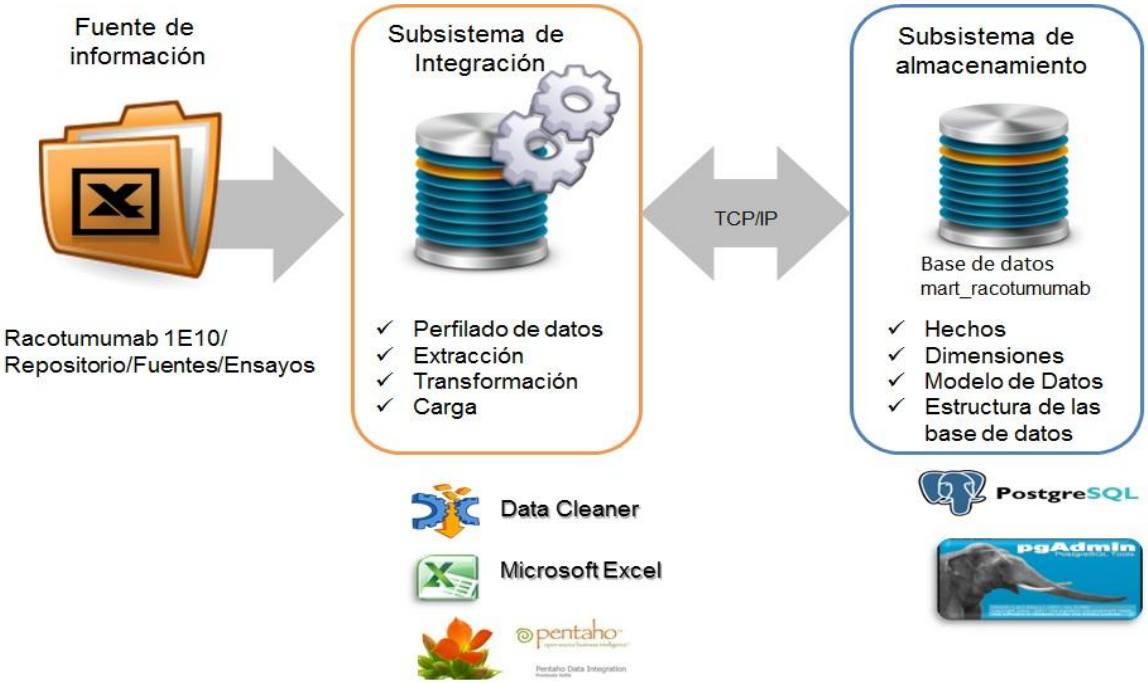


Figura 2. Arquitectura del mercado

En el primer nivel se encuentra el subsistema de integración, el cual se abastece de 25 ficheros excel pertenecientes a los tres EC del producto Racotumumab. Luego, se lleva a cabo los procesos de integración y transformación de la información obtenida de las fuentes de datos para su posterior almacenamiento. Los usuarios que acceden a este subsistema son los encargados de la administración de dichos procesos.

Por último, el subsistema de almacenamiento recibe la información tratada durante la extracción, transformación y carga, la cual se almacena en una base de datos soportada por el SGBD PostgreSQL y administrada por los usuarios autorizados mediante la herramienta PgAdmin.

2.6 Matriz bus

La matriz describe las relaciones entre los hechos y las dimensiones en un MD. Además, permite verificar que no haya solapamiento de hechos, o sea, que no existan dos hechos que compartan exactamente las mismas dimensiones en el almacén. A continuación se muestra la matriz bus, donde la celda marcada con una x indica la relación de una columna (hecho) con una fila (dimensión).

dimensiones/hechos	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11	v12	v13	v14	v15	v16	v17	v18	v19	v20
dim_tiempo	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		X		X	X
dim_localizacion	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
dim_provincia	X	X	X	X	X	X	X													
dim_hospital	X	X	X	X	X	X	X	X				X	X	X						
dim_edad	X							X						X						
dim_sexo	X							X						X						
dim_etapa_enfermedad	X																			
dim_tnm	X							X						X						
dim_tipo_histologico								X						X						
dim_examen_laboratorio	X		X		X			X		X								X		
dim_respuesta_clinica	X							X									X			X
dim_signos_vitales		X	X	X	X				X	X										
dim_examen_fisico	X	X	X	X	X			X	X	X										
dim_preguntas	X	X	X	X	X		X	X	X	X		X	X	X	X					X
dim_tecnica_imagenologia_radiologia	X			X	X			X		X							X		X	
dim_respuesta_tratamiento				X	X					X										
dim_sobrevida							X						X							
dim_tipo_evento_adverso						X					X					X				
dim_grado_evento_adverso						X					X					X				
dim_causalidad_evento_adverso						X					X									
dim_pais								X												
dim_raza								X												
dim_estadio								X						X						
dim_ecog								X												
dim_karnofsky								X						X						
dim_estado_oms														X						
dim_habito_toxico								X												
dim_dosis															X					
dim_causa_interrupcion							X					X								X
dim_causa_muerte							X					X								
dim_estado_concluir_tratamiento																				X
dim_tipo_evaluacion					X					X	X									
dim_analisis_anatomopatologicos	X							X												
dim_terapia_empleada	X							X												

Figura 3. Matriz bus de la presente investigación

Hechos:

V1: hech_inclusion_y_evaluacion_inicial_sclc

V2: hech_evaluacion_tratamiento_sclc_1

V3: hech_evaluacion_tratamiento_sclc_2

V4: hech_evaluacion_tratamiento_sclc_3

V5: hech_evaluacion_tratamiento_sclc_4

V6: hech_seguridad_sclc

V7: hech_interrupcion_sclc

V8: hech_inclusion_y_evaluacion_inicial_nsclc_internacional

V9: hech_evaluacion_tratamiento_nsclc_internacional_1

V10: hech_evaluacion_tratamiento_nsclc_internacional_2

V11: hech_seguridad_nsclc_internacional

V12: hech_interrupcion_nsclc_internacional

V13: hech_terminacion_ensayo_nsclc_internacional

V14: hech_datos_generales_nsclc

V15: hech_tratamiento_nsclc

V16: hech_seguridad_nsclc

V17: hech_evaluaciones_clinicas_nsclc

V18: hech_evaluaciones_laboratorio_clinico_nsclc

V19: hech_evaluacion_imagenologica_nsclc

V20: hech_evaluacion_final_nsclc

2.7 Estándares de codificación

En la presente investigación el diseño de los estándares de codificación es un aspecto importante a tener en cuenta, ya que permite organizar la forma en que se denominan las estructuras, con el objetivo de lograr un patrón que tribute a la correcta normalización de los términos utilizados y un entendimiento entre las partes implicadas en el proyecto.

En la presente investigación la nomenclatura fue clasificada teniendo en cuenta si la estructura es una dimensión o una tabla de hecho.

Tabla 5. Nomenclatura de las tablas de dimensiones y hechos

Tablas	Codificación	Ejemplo de tablas
dimensiones	dim_<nombre dimensión>	dim_localizacion
		dim_raza
hechos	hech_<nombre hecho>	hech_seguridad_sclc
		hech_interrupcion_sclc

En el caso de los atributos de las dimensiones se siguió la siguiente política:

Tabla 6. Nomenclatura de los campos de la tabla dimensión

Campos de la tabla dimensión	Codificación	Ejemplo de tablas
Llave primaria	dk_<nombre dimensión>_id	dk_dim_etapa_enfermedad_id
Código de la dimensión	<nombre dimensión>_codigo	etapa_enfermedad_codigo
Valor de la dimensión	<nombre dimensión>_valor	etapa_enfermedad_valor
Descripción de la dimensión	<nombre dimensión>_descripcion	etapa_enfermedad_descripcion

En el caso de las tablas hechos queda de la siguiente forma:

Tabla 7.Campos de la tabla hecho

Campos de la tabla hecho	Codificación	Ejemplo de tablas
Llave primaria	pk_<código>	pk_codigo_paciente

Al finalizar el proceso de estandarización de los nombres, quedó organizada la nomenclatura propuesta para la denominación de las tablas y atributos, dando paso al modelo de datos.

2.8 Modelo de datos de la solución

El modelo de datos es un lenguaje utilizado para describir y caracterizar los tipos de datos que se incluyen en la base de datos. Además, describe las relaciones entre las distintas entidades que lo componen.

A continuación se muestra un fragmento del modelo de datos obtenido (el diagrama completo se encuentra en el artefacto “Especificación del Modelo de Datos” y en el anexo 2).

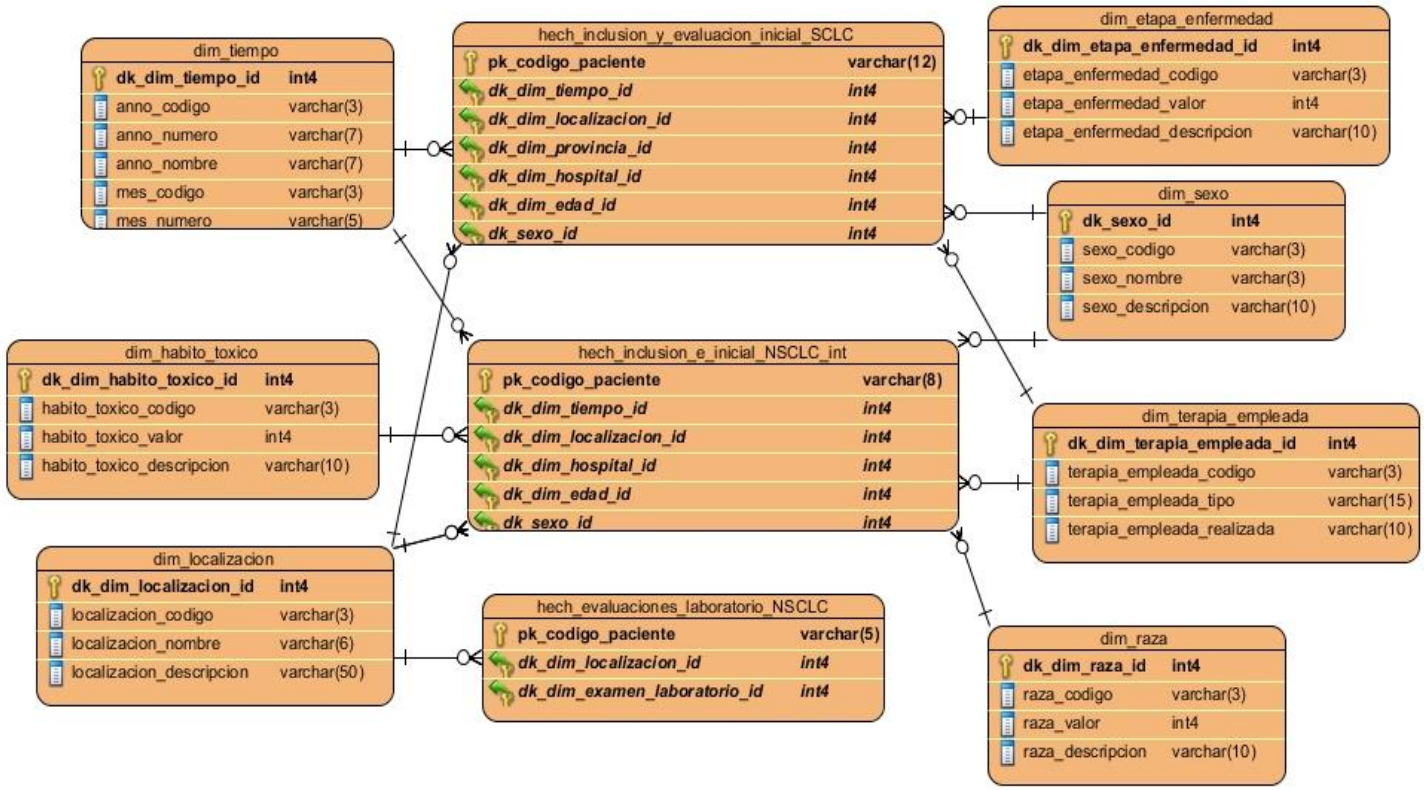


Figura 4. Muestra del modelo de datos de la solución

Inicialmente se definieron varias relaciones de mucho a mucho entre algunos hechos y dimensiones, debido a las características presentadas en el negocio a la hora de agrupar la información de cada paciente. Por ejemplo, durante las diferentes inmunizaciones son realizadas las siguientes mediciones a los pacientes: tensión arterial máxima, tensión arterial mínima, frecuencia cardiaca y temperatura. Para almacenar los valores asociados a las mediciones fue definida la tabla de dimensión dim_signos_vitales que se relaciona con varios hechos como hech_evaluacion_tratamiento_sclc_1. Teniendo en cuenta que a cada paciente se realizan todas las mediciones mencionadas anteriormente, se relacionaron las tablas dim_signos_vitales y hech_evaluacion_tratamiento_sclc_1 mediante una relación mucho a mucho, a partir de la cual se genera una nueva tabla (puente) con las llaves primarias de ambas. Sin embargo, atendiendo a que solamente se va a realizar una carga histórica de los datos, es decir, no se tiene en cuenta la carga incremental puesto que la fuente no va a aumentar, se decide establecer solamente relaciones uno a mucho entre los hechos y las dimensiones. Las tablas de hechos van a almacenar las llaves asociadas de cada una de las dimensiones y una llave para representar el código de cada paciente,

a partir de la cual es posible obtener la cantidad de pacientes. De esta manera se evita la utilización de tablas puentes cuyos datos no van aumentar.

2.9 Calidad de datos

Para garantizar que los datos tengan la calidad requerida para su posterior análisis deben pasar por un proceso de perfilado, permitiendo verificar la limpieza de los datos, la cantidad de valores nulos y cantidad de valores repetidos. Los problemas detectados, serán recogidos como reglas de transformación que después se convertirán en reglas del negocio y así quedan planteadas las anomalías que suponen pérdida de información para los clientes para ser solucionadas posteriormente.

Para llevar a cabo el perfilado de datos se utilizaron las herramientas Microsoft Excel y el DataCleaner, las cuales se aplicaron a los ficheros con formato excel que contienen la información de las fuentes de datos: EC de 1E10 NSCLC Compacional, EC 071 SCLC Fase II y EC NSCLC Internacional. Luego de haber sometido los datos de las fuentes a un proceso de perfilado se obtuvieron los siguientes resultados:

- Las fuentes de datos contienen 5 tipos de datos diferentes, lo cual permitió realizar el análisis individual de cada uno, para determinar el tratamiento que recibirán.

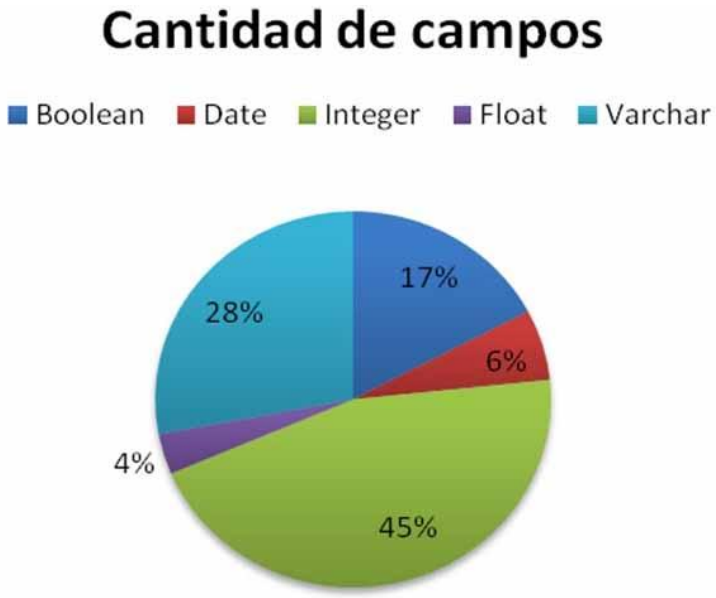


Figura 4. Cantidad de datos que contiene la fuente por tipos de datos

Se determinaron las fechas mínimas y máximas que se pueden encontrar en las fuentes de datos, lo cual permitió definir los rangos de fechas con los que se va a trabajar en cada una de las variables de fecha. Los valores de los números se refieren a:

1- Fecha de inclusión
2- Fecha de nacimiento
3- Fecha diagnóstico de la enfermedad
4- Fecha inicio del evento adverso
5- Fecha interrupción el tratamiento
6- Fecha inmunización
7- Fecha fallecimiento

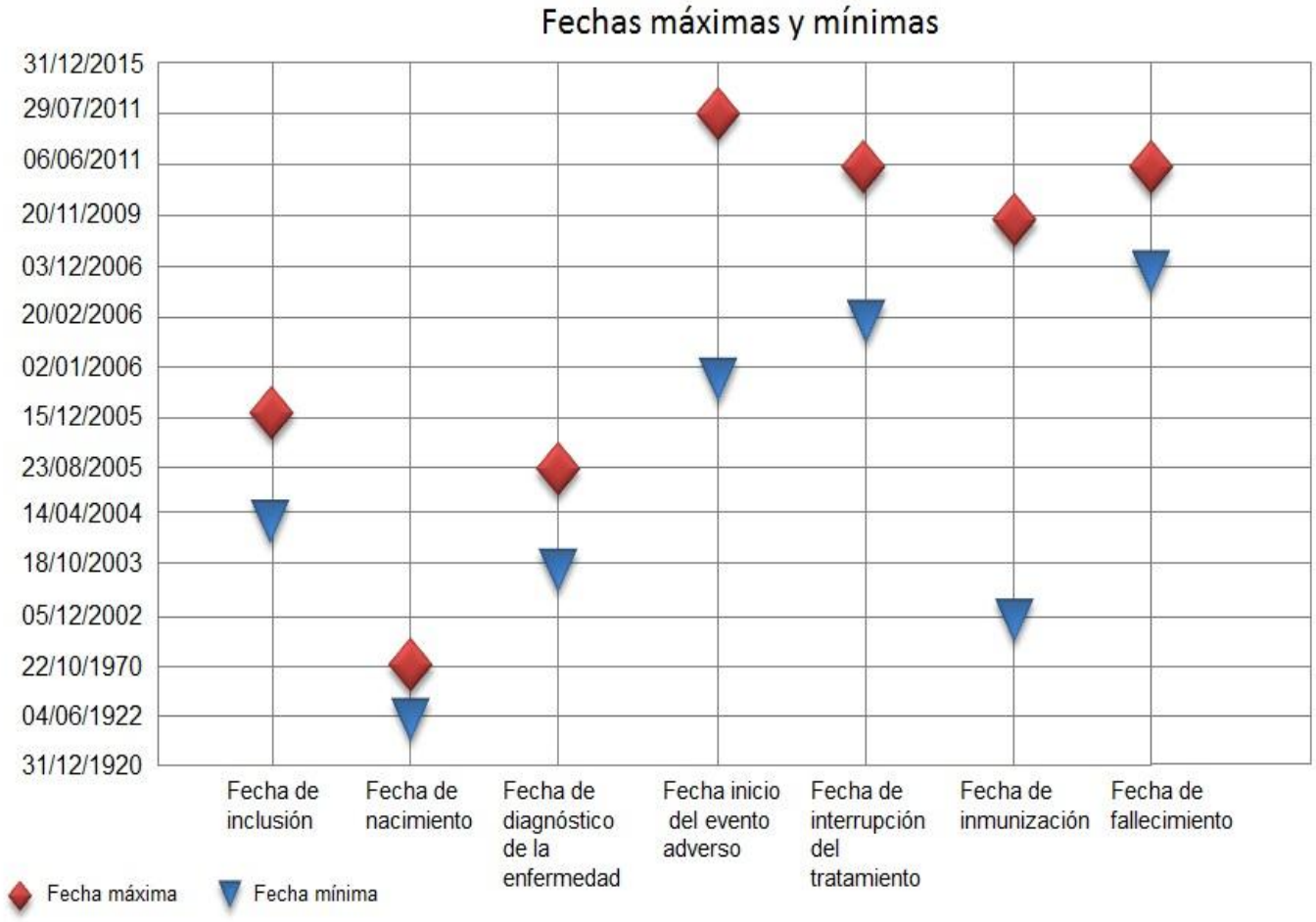


Figura 5. Fechas mínimas y máximas registradas en la fuente

2.10 Diseño del proceso de integración de datos

Las transformaciones en los procesos de integración de datos son una serie de pasos, que luego de ser ejecutadas permiten que los datos sean cargados en las tablas que se encuentran en la base de datos de

un MD. En el diseño general para realizar la carga de los hechos del MD lo primero es obtener los parámetros de conexión a la base de datos, luego se verifica si es válida la conexión a la base de datos, en caso de no ser válida se registran los parámetros de conexión en un excel. Se realiza la extracción de los datos desde la fuente (excel). Posteriormente se verifica la existencia de llaves nulas y se les da tratamiento. Luego, se procede a la limpieza y transformación de los datos aplicando las reglas de transformación necesarias. Además, se busca el identificador correspondiente en las tablas de dimensiones y luego se verifica la existencia de llaves huérfanas; si existen problemas se envían hacia un excel de errores. Se cargan los datos en la tabla de hecho correspondiente y se obtiene información del sistema necesaria para generar los metadatos: nombre del fichero fuente, nombre de la transformación, fecha de ejecución de la transformación y estado de la misma. Por último se inserta dicha información en la tabla respectiva de la base de datos perteneciente a la solución (el diseño del proceso de integración de datos se encuentra en el anexo 4).

2.11 Seguridad en el mercado de datos

La seguridad de un sistema es un proceso fundamental a tratar en cualquier mercado, ya que es un mecanismo de protección contra aquellas acciones que puedan afectar la integridad de los datos almacenados. Por tal motivo, para el acceso al producto se ha establecido un usuario por cada uno de los roles existentes en el sistema, con el objetivo de garantizar un control de acceso basado en roles, de esta manera cada usuario opera en el sistema según los permisos que se le definan al rol correspondiente.

2.11.1 Seguridad en el subsistema de almacenamiento

Los roles que se identificaron fueron el analista y el administrador de ETL. El analista sólo puede hacer consultas de tipo SELECT a las tablas, debido a que tiene permiso de lectura únicamente. Por otra parte el administrador de ETL realiza los procesos de ETL sobre los datos y tiene permiso de lectura y escritura sobre los esquemas pertenecientes al presente trabajo.

2.11.2 Seguridad en el subsistema de integración

Se garantiza a nivel de sistema operativo, el cual permite asignar permisos a los archivos para determinados usuarios y grupos de usuarios. De esta manera, se puede restringir el acceso de un archivo a determinados usuarios. Esta ventaja es utilizada para restringir el acceso por parte de usuarios no autorizados a los archivos que contienen las transformaciones y trabajos que permiten el desarrollo de los procesos de integración del producto.

Conclusiones del capítulo

El análisis del negocio realizado permitió definir los requisitos de información, funcionales y no funcionales. El establecimiento de las reglas del negocio y la identificación de las tablas de hechos y dimensiones tributaron al diseño del modelo de datos, a partir del cual es posible obtener la estructura física de almacenamiento. El perfilado de datos realizado a las fuentes de información, posibilitó verificar el estado y calidad de los datos, así como la creación de nuevas reglas del negocio y transformación. A partir del diseño de los procesos de integración de datos asociados a la carga de las dimensiones y los hechos, quedó definido el flujo de acciones o actividades a realizar durante la implementación del subsistema de integración.

CAPÍTULO 3: IMPLEMENTACIÓN Y PRUEBA DE LOS SUBSISTEMAS DE ALMACENAMIENTO E INTEGRACIÓN DEL MERCADO DE DATOS RACOTUMUMAB

Introducción

En este capítulo se aborda todo lo referente a la implementación de la estructura física de la solución y la realización de los procesos de ETL de los subsistemas de almacenamiento e integración del MD Racotumumab. Una vez concluida la implementación se da paso a una de las etapas más importantes en el ciclo de desarrollo de un software: las pruebas, las cuales permitirán encontrar y corregir no conformidades existentes, obteniéndose como resultado una aplicación con mayor calidad.

3.1 Implementación del subsistema de almacenamiento

Luego de haber diseñado el modelo dimensional se dio lugar al modelo físico, que permite describir el almacenamiento de los datos y la relación de las tablas. Fueron creados los esquemas así como las tablas correspondientes a cada uno.

Para organizar las tablas de la base de datos en la presente investigación se definieron tres esquemas, que cuentan con un total de 54 tablas, divididas en 34 tablas de dimensiones, 20 tablas de hechos y cuatro tablas de metadatos. Los esquemas son:

- dimensiones: contiene las dimensiones compartidas con el resto de los MD del almacén.
- mart_racotumumab: recoge las tablas de dimensiones y hechos propias del MD.
- metadatos: este esquema cuenta con dos tablas destinadas a los metadatos técnicos de las dimensiones y los hechos, y otras dos, para los metadatos de procesos en las transformaciones y trabajos.

Algunas de las tablas contenidas en los tres esquemas utilizados para la solución se mencionan a continuación:

Tabla 8.Estructura de datos

Esquemas	Ejemplo de tablas
dimensiones	dim_edad
	dim_estadio
	dim_hospital
	dim_tiempo
mart_racotumumab	dim_causa_muerte
	dim_dosis
	hech_datos_generales_nsclc

	hech_interrupcion_sclc
metadatos	md_procesos_trabajos
	md_procesos_transformaciones
	md_tecnicos_dimensiones
	md_tecnicos_hechos

3.2 Implementación del subsistema de integración de los datos

El proceso de integración de los datos consta de tres etapas fundamentales relacionadas entre sí: la extracción, transformación y carga de los datos. Una vez extraída la información se realiza la limpieza de los datos para identificar y corregir los problemas que se identifiquen. Luego los datos son transformados y cargados, poblando las dimensiones y los hechos que conforman la estructura de la solución.

Los procesos de ETL cuentan con varios subsistemas que posibilitan una correcta implementación de cada uno de los subprocesos que lo componen. Kimball propuso 34 subsistemas de ETL separados en 4 grupos, los cuales son [4]:

- Extracción
- Limpieza y conformación
- Entrega
- Gestión

A continuación se describen los subsistemas de ETL identificados que se utilizan en el desarrollo de la solución propuesta:

- Sistema de extracción: se encuentra en el grupo de “Extracción” y permite la permite la extracción de datos desde la fuente de origen a la fuente destino.
- Rastreo de eventos de errores: se encuentra en el grupo de “Limpieza y conformación”. Este subsistema captura todos los errores que proporcionan información valiosa sobre la calidad de datos y permiten la mejora de los mismos.
- Dimensiones Lentamente Cambiantes (SCD): se encuentra en el grupo de “Entrega” e implementa la lógica para crear atributos de variabilidad lenta a lo largo del tiempo.
- Llave subrogada: se encuentra en el grupo de “Entrega”. Este subsistema permite crear llaves subrogadas independientes para cada tabla.
- Tablas de hecho: se encuentra en el grupo de “Entrega” y permite crear tablas de hecho.

- Repositorio de metadatos: se encuentra en el grupo de “Gestión”. Este subsistema captura los metadatos de los procesos ETL y de los aspectos técnicos.

3.2.1 Implementación de las transformaciones

El elemento básico de diseño de los procesos de ETL es la transformación, la cual está compuesta por pasos que se encuentran unidos a través de saltos. Una vez identificadas las reglas del negocio, deben crearse e incluirse las definiciones en las rutinas de transformación. Se realizaron un total de 54 transformaciones, 34 para las dimensiones y 20 para los hechos. A continuación se describen las transformaciones del hecho hech_evaluacion_tratamiento_sclc_3 y de la dimensión dim_raza.

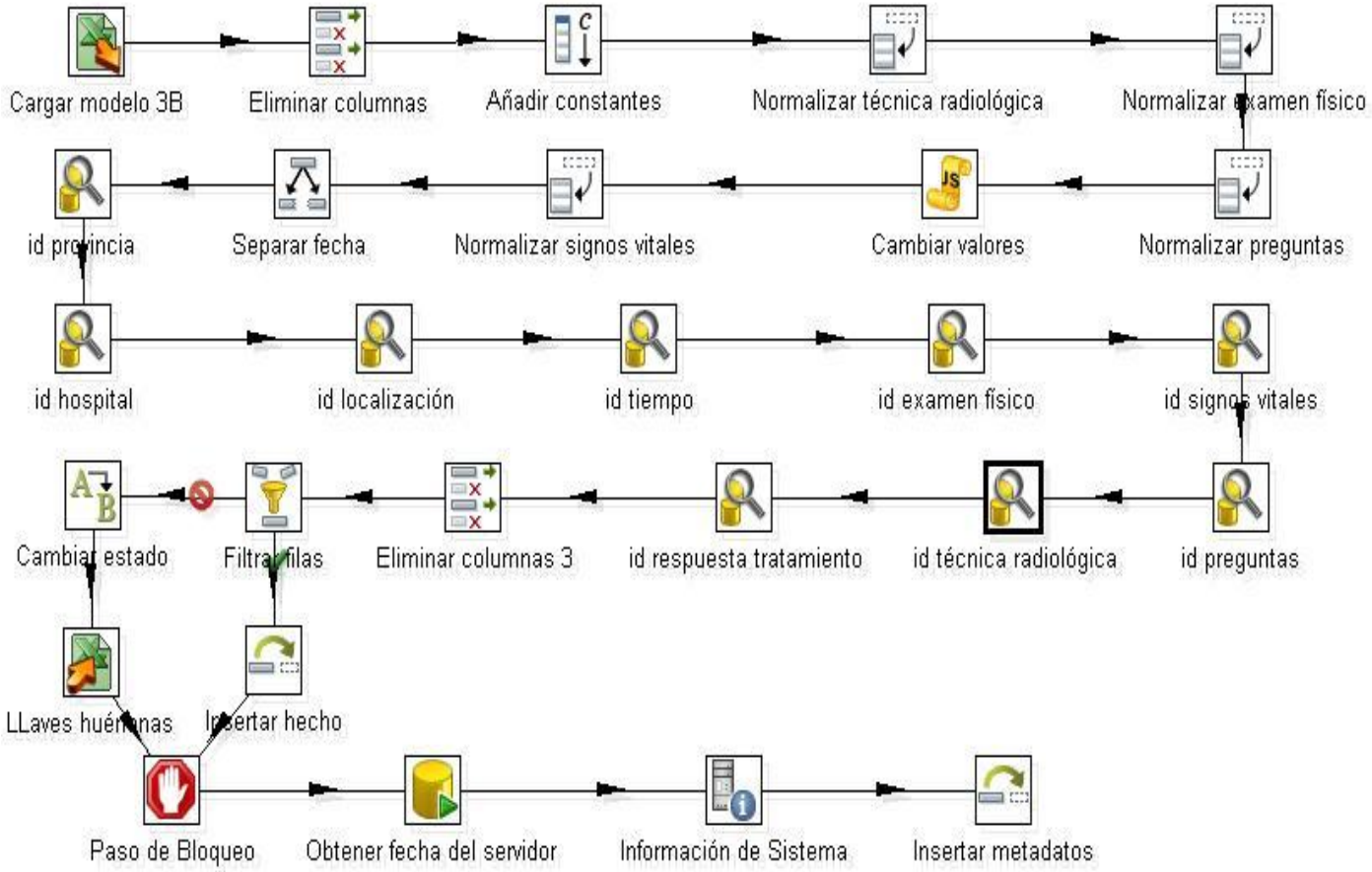


Figura 5. Transformación hech_evaluacion_tratamiento_sclc_3

Para realizar la extracción de los datos correspondientes a la transformación del hecho: evaluación del tratamiento SCLC 3, se accede a la fuente de datos de donde se extraen los campos necesarios en dependencia de las dimensiones con que se relaciona el hecho. Se crean un grupo de constantes, se

realizan una serie de transformaciones como, normalizar algunos campos y cambiar algunos valores que están mal escritos. Se buscan los identificadores de cada dimensión asociada, se eliminan los campos innecesarios. Se verifica la existencia de llaves huérfanas, en caso de existir alguna, se cambia el valor de la constante estado de la transformación y se almacena en un excel el código del paciente y los demás identificadores. Se inserta el hecho y se obtienen varios parámetros para insertar los metadatos, como la hora actual del servidor y el nombre de la trasformación.

Figura 6. Transformación dim_raza

Para realizar la extracción de los datos correspondientes a la transformación de la dimensión raza, se accede a la fuente de datos y se extraen los campos. Luego se inserta la dimensión y se obtienen varios parámetros para insertar los metadatos, como la hora actual del servidor y el nombre de la trasformación.

3.2.2 Implementación de los trabajos

En el contexto de integración de datos, el término trabajo se entiende como un conjunto sencillo o complejo de tareas cuyo objetivo consiste en realizar una acción determinada. La implementación de un trabajo define una secuencia lógica para la ejecución de las transformaciones, mediante el uso de pasos definidos, los cuales son diferentes a los disponibles en las transformaciones. Además, es posible ejecutar una o varias transformaciones de las que se hayan diseñado y orquestar una secuencia de ejecución para ellas. Se realizaron un total de dos trabajos. Para realizar el trabajo general primeramente se obtienen los parámetros de conexión de la base de datos y se verifica que se pueda establecer la conexión, en caso de no poderse se guarda en un excel los parámetros de conexión para su análisis. Luego se cargan las dimensiones que más relaciones tienen con los hechos y se verifican los hechos que se puedan ejecutar.

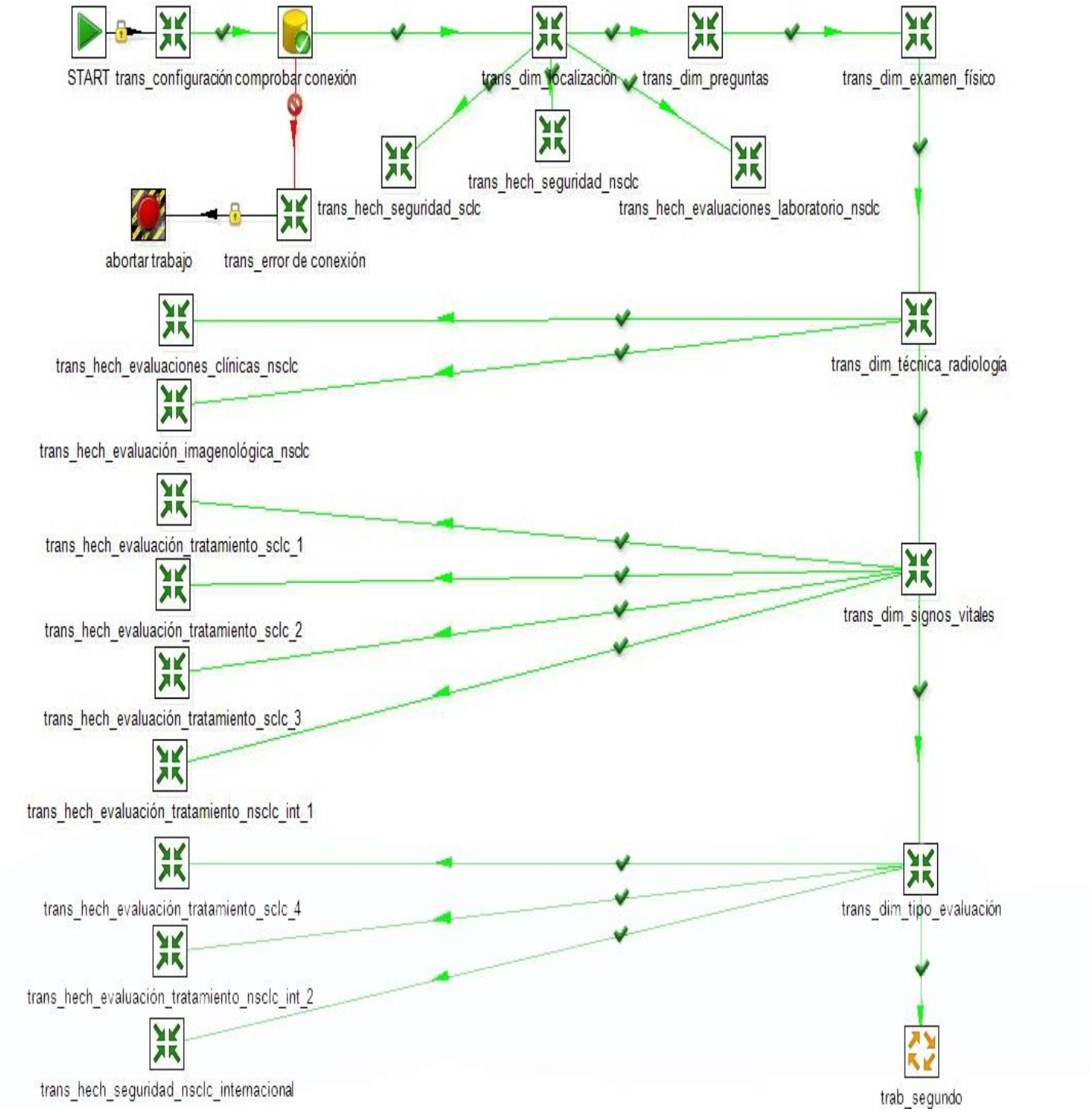


Figura 7. Trabajo general

3.2.3 Estrategias de carga de dimensiones y hechos

Una vez realizadas todas las transformaciones correspondientes se procedió a cargar en la solución propuesta las tablas de hechos y dimensiones. Para llevar a cabo este proceso fue necesario seguir varias estrategias de carga tanto en las dimensiones como en los hechos. En el presente trabajo se utilizó para la carga de las dimensiones las llaves sustitutas o llaves subrogadas y las dimensiones lentamente cambiantes (Slowly Changing Dimensions, SCD). Por otra parte para la carga de los hechos se utilizaron las estrategias de llaves nulas y llaves huérfanas.

3.2.3.1 Dimensiones lentamente cambiantes

Las SCD determinan cómo se manejan los cambios históricos en las tablas de dimensiones. Ralph Kimball propone seis estrategias a seguir para las SCD. A continuación se describen las mismas [4]:

- SCD Tipo 0: no se tiene en cuenta la gestión histórica y no se realiza esfuerzo alguno para lidiar con los problemas del cambio de la dimensión. De este modo alguna información es sobrescrita mientras que otra queda sin cambios.
- SCD Tipo 1 (sobrescribir): es utilizado cuando la información histórica no es importante. Este tipo sobrescribe los datos antiguos con nuevos, es utilizado mayormente para corregir errores de datos en las dimensiones. A pesar de ser fácil de implementar presenta como desventaja principal que no permanece ningún registro histórico en la dimensión.
- SCD Tipo 2 (añadir fila): cuando hay un cambio se crea una nueva entrada en la tabla. Al nuevo registro se le asigna una nueva llave subrogada y a partir de este momento será el valor usado para futuras entradas, las antiguas usarán el valor anterior. En este modo se gestiona un versionado que puede incluir fechas para indicar los períodos de validez, así como numeradores de registros o indicadores de registros activos o no. Este tipo permite guardar toda la información histórica en el almacén de datos.
- SCD Tipo 3 (añadir columna): esta estrategia requiere que se agregue una nueva columna a la tabla por cada tupla cuyos valores se desea mantener un historial de cambios. De este modo en la nueva columna se coloca el valor antiguo antes de sobrescribir el valor actual con el nuevo.
- SCD Tipo 4 (tabla de historia separada): su función es almacenar en una tabla adicional los detalles de cambios históricos realizados a la tabla de dimensión. La tabla con la información histórica indicará el tipo de operación que se ha realizado, sobre qué campo se realizó el cambio y la fecha del mismo. Esta tabla tiene como objetivo mantener un detalle de los cambios realizados.

- SCD Tipo 6 (híbrido): este método combina los tipos anteriores 1, 2 y 3; y se le denomina tipo 6 debido a la suma de los tres tipos que integra. Esta estrategia utiliza los 3 tipos antes mencionados y añade además una pareja adicional de columnas para indicar el rango de fechas al cual aplica cada fila en particular.

Para la solución se escogió la opción de sobrescribir los datos utilizando la estrategia de tipo 1, en caso de que se cometiera algún error ortográfico no sería necesario guardar los cambios efectuados. Esta estrategia fue aplicada a todas las dimensiones del presente trabajo.

3.2.3.2 Llaves subrogadas

Una llave subrogada es un identificador único que se asigna a cada registro de una tabla de dimensión. Esta clave, generalmente, no tiene ningún sentido específico de negocio y son siempre de tipo numérico. La aplicación de las mismas va a ser muy necesaria por diferentes motivos [37]:

- Fuentes heterogéneas: el MD se alimenta de diferentes fuentes, cada una de ellas con sus propias llaves, por lo que es arriesgado asumir un código de alguna aplicación en particular.
- Rendimiento: en la base de datos, ocupa menos espacio un entero que una cadena. De hecho, el espacio que ocupa no es tan preocupante como lo es el tiempo que se pierde en leerlo, ya que las llaves subrogadas se llevan a las tablas de hechos, por lo que cada código es susceptible de repetirse cientos de millones de veces. Conviene optimizarlo al máximo.

Esta estrategia brinda varias ventajas como [38]:

- Ocupan menos espacio.
- Son de tipo numérico entero. Al ser de tipo numérico permite realizar la búsqueda en base de datos de manera más eficiente, debido a que es un número lo que se va a comparar, de lo contrario si fuera una cadena, se tiene que comparar carácter a carácter, por lo que tardaría más tiempo.

En el presente MD todas las dimensiones aplican la estrategia de la llave subrogada.

3.2.3.3 Llaves nulas

En el presente trabajo se dio tratamiento a los valores nulos sustituyéndolos por valores preestablecidos. En las dimensiones correspondientes se añadió una fila con el valor por el que fue sustituido el campo nulo. De esta manera, al encontrarse con las llaves nulas cambia su destino hacia la fila indicada en la dimensión para esos casos y es la clave de ese campo la que se añade a la tabla de hechos.

Algunos hechos en los cuales se aplicó la estrategia son: hec_datos_generales_nsclc y hech_evaluacion_imagenologica_nsclc. Algunas de las dimensiones que están asociadas a los mismos son: dim_tiempo, dim_clasificacion_tnm y dim_tecnica_imagenologia_radiologia.

3.2.3.4 Llaves huérfanas

Esta estrategia se aplicó ya que en algunos campos se encuentran clasificaciones que no son las establecidas en las fuentes, es decir, el valor que se encuentra en el campo es huérfano porque no tiene ningún padre. Un ejemplo ocurre con la variable tipo histológico, la cual tiene cuatro clasificaciones: Epidermoide, Adenocarcinoma, Carcinoma de células grandes y Carcinoma, pero en el flujo de datos aparece el tipo histológico "Carcinoma broncogénico de pulmón derecho" el cual no pertenece a ninguna de las clasificaciones anteriores. Para dar solución a este problema se relacionó la información a "Otros" o algún tipo de clasificación de las anteriores (ejemplo: Carcinoma). Algunos hechos en los que se aplicó esta estrategia son: hec_datos_generales_nsclc y hech_evaluacion_imagenologica_nsclc y algunas de las dimensiones que están asociadas a los mismos son: dim_tiempo y dim_hospital.

3.2.4 Gestión de los metadatos del proceso de integración

Los metadatos son datos que ayudan a identificar, describir y localizar recursos digitales, son información estructurada que describe y/o permite encontrar, gestionar, controlar y entender o preservar otra información; o sea, que no son más que datos sobre los propios datos. A continuación se presentan algunos de los grupos o categorías de los metadatos [39, 40]:

- Metadatos administrativos: son utilizados para el manejo y administración de los recursos de información.
- Metadatos descriptivos y de descubrimiento: utilizados para describir, descubrir o identificar los recursos de información.
- Metadatos técnicos o modelos: están relacionados con la función de un sistema o el modo en que interrelacionan sus componentes.
- Metadatos de proceso: permiten obtener información de los procesos en que se ejecutan.
- Metadatos de negocio: posibilita obtener los datos y la información referente a los aspectos del negocio.

En la investigación se usaron los metadatos técnicos y los de procesos para obtener la información correspondiente a los procesos de las transformaciones y los trabajos referentes a los subprocesos de ETL. Esta estrategia se aplicó con el objetivo de poder presentar los resultados de la ejecución del propio

proceso de integración de datos, tales como: nombre del mercado, nombre de la tabla de la base de datos en la que se va a trabajar, nombre de la transformación, fecha en que se ejecutó la transformación y el estado de la misma, filas leídas, filas escritas, filas actualizadas, entre otras más. Las tablas que gestionan los metadatos son: md_técnicos_dimensiones, md_técnicos_hechos, md_procesos_trabajos y md_procesos_transformaciones.

3.2.5 Gestión de errores

Durante el proceso de carga, los errores que se encontraron relacionados con la existencia de llaves huérfanas, fueron guardados en un excel. Luego se notificó a los especialistas de la dirección de Investigaciones Clínicas del CIM de forma personal, en los casos donde se debió consultar al cliente para aplicar transformaciones no definidas en las reglas identificadas a partir del perfilado de datos. Esta estrategia se aplicó en todas las tablas de hechos.

3.2.6 Estructura de la información

Para asegurar el acceso a las transformaciones y trabajos implementados se definió una carpeta o repositorio de datos la cual contiene los ficheros utilizados y generados en el proceso de integración. Las carpetas son:

- Configuración: se guarda un archivo .xml donde se encuentran las variables para la conexión a la base de datos y una transformación que permite asignar estas a variables en el PDI.
- Fuentes: contiene otras cuatro carpetas, una para las dimensiones donde se guardan las fuentes de donde se cargaron las mismas y otras tres para guardar las fuentes de cada ensayo (NSCLC Compacional, NSCLC Internacional y SCLC).
- Log: contiene dos carpetas, una para guardar los errores de conexión y la otra para guardar todas las llaves huérfanas que se encontraron.
- Trabajos: contiene todos los trabajos que se realizaron.
- Transformaciones: contiene todas las transformaciones que se realizaron.

Los ficheros están ubicados en la siguiente estructura de carpetas:

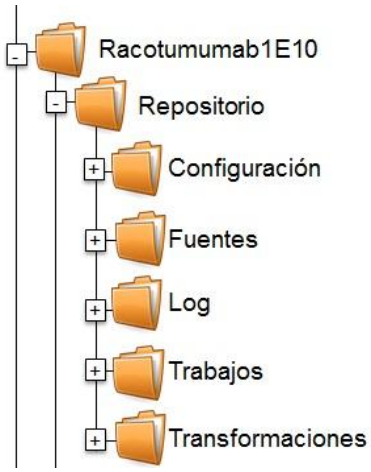


Figura 9. Estructura del repositorio

3.3 Pruebas realizadas al mercado de datos

Todo proceso de creación de software está sujeto a fallos, por lo que las pruebas que se realicen constituyen una fase importante en el desarrollo del producto, ya que permiten comprobar que no existan imperfecciones en la implementación del mismo, proporcionándole calidad al software.

3.3.1 Pruebas unitarias

Las pruebas unitarias permiten probar el correcto funcionamiento de un componente o subsistema específico y son realizadas por los propios desarrolladores durante la implementación [12]. Los programadores siempre prueban el código durante el desarrollo, por lo que las pruebas unitarias son realizadas solamente después de que el programador considera que el componente se encuentra libre de errores.

A medida que se desarrollaba la solución del problema que dio surgimiento a la presente investigación se realizaron pruebas unitarias a los subsistemas de almacenamiento e integración, donde se detectaron 4 no conformidades, las cuales fueron corregidas rápidamente. Además, se realizaron una serie de iteraciones de prueba por parte de los especialistas del departamento de Almacenes de Datos, con el objetivo de identificar no conformidades en distintas etapas de la solución. Las no conformidades encontradas fueron:

1. Debido a que no se realizará una carga incremental se pueden eliminar las tablas puentes existentes y poner a cada uno de los hechos una llave primaria para evitar las llaves duplicadas.
2. Nombrar de manera sugerente las actividades en los diseños de los proceso de integración de datos, de manera que reflejen su verdadero objetivo.

3. Establecer un orden lógico para la ejecución de las actividades.
4. Incluir el estado de cada carga en los metadatos capturados para saber cómo ocurrió el proceso, tanto en los diseños de los proceso de integración de datos como en la implementación de las transformaciones.
5. Intercambiar el orden de las llaves nulas con el de llaves huérfanas en el diseño de los procesos de integración de datos.
6. Realizar ajustes en algunos componentes de manera que permitan optimizar las transformaciones.

Cada una de las no conformidades encontradas en la revisión fue tratada y corregida en su totalidad.

3.3.2 Pruebas de integración

Las pruebas de integración permiten verificar la correcta integración de los componentes y subsistemas que conforman la solución. Estas pruebas son ejecutadas por los arquitectos de software [12]. La estrategia definida para realizar las pruebas de integración incluye la confección de casos de prueba que resultan de gran importancia para demostrar la funcionalidad del mismo.

3.3.2.1 Casos de prueba

Los casos de prueba son utilizados para identificar posibles fallos de implementación y comprobar el grado de cumplimiento de los requisitos especificados para el sistema. En la presente investigación se diseñaron 20 casos de prueba asociados a los casos de uso de información identificados en la etapa de análisis. La siguiente tabla muestra el caso de prueba asociado al caso de uso mantener disponible la información de la evaluación durante el tratamiento tres en el EC SCLC (los restantes casos de prueba asociados a casos de uso se encuentran en el artefacto “Casos de prueba de integración”).

Tabla 9. Caso de prueba asociado a un caso de uso (1)

Caso de uso de información	Requisito de información	Tablas implicadas	Variables de entrada
Mantener disponible la información de la evaluación durante el tratamiento 3 en el EC SCLC.	Mantener disponible la información de la cantidad de pacientes evaluados durante el tratamiento 3 en el ensayo clínico SCLC por tiempo, localización, provincia, hospital, signos vitales, examen físico, preguntas, técnica de imagenología y radiología y respuesta al tratamiento.	dim_provincia dim_preguntas hech_evaluacion_tratamiento_sclc_3	Provincia Preguntas

Tabla 10. Caso de prueba asociado a un caso de uso (2)

Variables de salida	Consulta SQL realizada	Datos obtenidos
Cantidad de pacientes	SELECT count (distinct pk_codigo_paciente) FROM mart_racotumumab.hech_evaluacion_tratamiento_sclc_3 where dk_dim_pregunta_id=79 and dk_dim_provincia_id=11;	Se obtuvieron 8 pacientes que respondieron falso a la pregunta "la administración se retrasó" y que son de la provincia Pinar del Río.

Tabla 11. Caso de prueba asociado a un caso de uso (3)

Fuente de datos	Variables de la fuente implicadas	Datos almacenados en la fuente	Resultados de la prueba
02_1E10 (4 cerrados)/ EC071 1E10 SCLC Fase II /Modelo3B.Evaluacion durante Tto(6,15).xls	v1 v20	Se obtuvieron 8 pacientes que respondieron falso a la pregunta "la administración se retrasó" y que son de la provincia Pinar del Río.	Resultados satisfactorios

3.3.3 Listas de chequeo

La lista de chequeo consta de una serie de preguntas en las cuales se verificará el grado de cumplimiento de determinadas reglas establecidas en el proceso en que se desarrollaba el sistema. Además, mide la calidad de los artefactos generados durante la realización del producto. Esta evaluación se desarrolla a través del análisis de un grupo de indicadores, distribuidos en tres secciones fundamentales:

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

La estructura de la lista de chequeo está formada por los siguientes elementos:

- Peso: define si el indicador a evaluar es crítico o no.
- Indicadores a evaluar: constituyen los indicadores a evaluar en las secciones estructura del documento y semántica del documento e indicadores definidos para el artefacto a evaluar.
- Evaluación: es la forma de evaluar el indicador en cuestión. El mismo se evalúa de uno en caso de que exista alguna dificultad sobre el indicador y de cero, en caso de que el indicador revisado no presente problemas.
- N.P. (No Procede): se usa para especificar que no es necesario evaluar el indicador en ese caso.
- Cantidad de elementos afectados: 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 definida la estructura de la lista de chequeo esta fue aplicada a varios artefactos del MD, donde se midieron 13 indicadores en el diccionario de datos, nueve en el mapa lógico, 12 en el perfilado y 13 en el registro del sistema fuente. Luego de aplicar la lista de chequeo fueron corregidas todas las no conformidades encontradas en dichos artefactos. (Las gráficas con los resultados obtenidos se pueden encontrar en la carpeta de artefacto de prueba “Lista de chequeo”).

3.3.4 Auditoría de datos

La auditoría de datos es el proceso de gestionar cómo los datos se ajustan a los propósitos definidos por la organización. A través de la realización de auditoría a los datos se obtiene conocimiento relacionado con la confiabilidad de los mismos, así como la información asociada a la ejecución de las transformaciones, como son: nombre de la transformación, fecha y hora de ejecución, cantidad de elementos de entrada, cantidad de elementos de salida, número de errores, entre otros elementos. La estrategia definida para auditar los datos almacenados se basa fundamentalmente en el uso de las tablas

de metadatos implementadas, que proveen la información necesaria para comprobar y garantizar que los datos cargados sean confiables.

Conclusiones del capítulo

En este capítulo quedan expuestos los detalles de la implementación de las tablas físicas del MD en PostgreSQL, ajustándose al modelo físico diseñado previamente y a las convenciones de nombrado establecidas. De igual forma quedan definidos los esquemas, secuencias y restricciones. Se obtuvo con éxito la información necesaria para proceder a la realización de los procesos de ETL, por medio del cual se transformó y cargó hacia el mercado una muestra representativa de cada categoría de la información que se manipula en el MD. Luego se expusieron los detalles de las pruebas del MD aplicando pruebas unitarias y de integración. Durante esta fase se diseñaron y aplicaron los casos de pruebas, las deficiencias encontradas fueron corregidas de forma exitosa arrojando como resultado un producto con mayor calidad y más completo en funcionalidad.

CONCLUSIONES GENERALES

Al finalizar el proceso de desarrollo del MD se puede afirmar que se le ha dado cumplimiento de forma satisfactoria al objetivo general del presente trabajo de diploma, obteniéndose los resultados que se esperaban. De igual forma durante el ciclo completo se cumplieron cada una de las tareas de la investigación planteadas.

- Una vez analizados los fundamentos teóricos de la investigación, quedaron establecidas las bases que sustentan el desarrollo del MD subsistemas de almacenamiento e integración del producto Racotumumab como solución al problema planteado. La metodología utilizada permitió guiar la construcción del sistema propuesto a través de los casos de uso. Las herramientas utilizadas posibilitaron cumplir con las políticas de migración a software libre.
- El análisis y diseño del MD subsistemas de almacenamiento e integración del producto Racotumumab permitió identificar las tablas de dimensiones y hechos, garantizando una adecuada estructura física de almacenamiento. Se diseñaron los procesos de integración de datos para la carga de las dimensiones y los hechos, quedando definido el flujo de acciones o actividades a realizar durante la implementación del subsistema de integración.
- A partir de la implementación de las estructuras modeladas se obtuvo un repositorio único de datos estandarizados, que contiene la información histórica extraída desde los archivos *Excel* asociados a los EC del producto Racotumumab.
- Con el objetivo de comprobar el cumplimiento de los requisitos definidos inicialmente, se realizaron seis pruebas unitarias y se utilizó una estrategia basada en la aplicación de cuatro listas de chequeo y 20 casos de prueba de integración. Los resultados satisfactorios obtenidos permitieron corroborar las funcionalidades del sistema a partir de los requisitos establecidos.

Al concluir la investigación se obtuvo un sistema capaz de resolver el problema existente en el CIM y apoyar el proceso de toma de decisiones en dicho centro.

RECOMENDACIONES

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

- Aplicar al MD subsistemas de almacenamiento e integración del producto Racotumumab algunas técnicas de minerías de datos que permitan detectar patrones de comportamiento sobre la información almacenada.

REFERENCIAS BIBLIOGRÁFICAS

[1] Centro Tecnologías de Gestión de Datos. [Consultado: el 9 de noviembre del 2012]. Disponible en: <http://gespro.datec.prod.uci.cu/>

[2] LAPORTE, Joan Ramón. *Principios Básicos de Investigación Clínica*. Fundación Instituto Catalán de Farmacología, 2007. [Consultado: el 5 de enero del 2012]. Disponible en: <http://www.icf.uab.es/llibre/llibre.htm>

[3] H.INMON, William. *Building the Data Warehouse*. Published by Wiley Publishing, 2000, Fourth edition. [Consultado: el 31 de enero del 2012]. Disponible en: <http://inmoncif.com/inmoncif-old/www/library/whiteprs/ttbuild.pdf>

[4] KIMBALL, R. y ROSS, M. *The Data Warehouse Toolkit: the Complete Guide to Dimensional Modelling*. New York, EE.UU: John Wiley & Sons, 2002. [Consultado: el 31 de enero del 2012].

[5] BUSTOS, Jorge. *Business Intelligence y Data Warehousing en Windows*, 2005.

[6] CHICIAZA, Janneth Alexandra. *Guía didáctica: Modelamiento de Datos*. Universidad Técnica Particular de Loja, 2011, Ecuador. [Consultado: el 15 de noviembre del 2012]. Disponible en: <http://rsa.utpl.edu.ec/material/208/G181003.1.pdf>

[7] ACOSTA MÉNDEZ, Geidy. *Mercado de datos para una Dirección de Salud en Cuba*. XV Convención y Feria Internacional Informática 2013. [Consultado: el 10 de mayo del 2013].

[8] SÁNCHEZ GALLARDO, Yisel de Lisy, PÉREZ REBOLLO, Rolando y WILKINSON BRITO, Bárbara. *Diseño del mercado de datos CIMAVAX EGF para el almacén de datos del Centro de Inmunología Molecular*. XV Convención y Feria Internacional Informática 2013. [Consultado: el 10 de mayo del 2013].

[9] RUANO VALDÉS, Osmar y QUERALTA POZO, Alejandro. *Mercado de Datos para el módulo visor de historias clínicas del Sistema Integral de Atención Primaria de la Salud*. Centro de Tecnologías de Gestión de Datos (DATEC), Facultad 6. Universidad de las Ciencias Informáticas, La Habana, 2012. [Consultado: el 20 de mayo del 2013].

[10] HIDALGO LÓPEZ, Leydis. *Mercado de datos para la Unidad Central de Cooperación Médica*. Universidad de las Ciencias Informáticas, La Habana, 2012. [Consultado: el 20 de mayo del 2013].

[11] EcuRed. *Metodologías de desarrollo de software*, 2010. [Consulta: 15 de noviembre del 2012]. Disponible en: <http://www.ecured.cu/index.php/Metodolog%C3%ADas_de_desarrollo_de_software>

[12] GONZÁLEZ HERNÁNDEZ, Yanisbel. *Propuesta de metodología para el desarrollo de almacenes de datos en DATEC*. La Habana, 2011. [Consultado: el 9 de febrero del 2013].

[13] BERNABEU, R.D. *Hefesto: Metodología propia para la construcción de un Data Warehouse*, 2007. [Consultado: el 9 de febrero del 2013]. Disponible en: <http://www.dataprix.com/es/hefesto-metodologia-propia-para-la-construccion-un-data-warehouse>

[14] ZEPEDA SÁNCHEZ, Leopoldo Zenaido. Universidad Politécnica de Valencia. Departamento de Sistemas Informáticos y Computación, 2008. [Consultado: el 22 de febrero del 2013]. Disponible en: <http://riunet.upv.es/bitstream/handle/10251/2506/tesisUPV2841.pdf>

[15] FUSTER, G. *Evaluación comparativa de herramientas CASE para UML desde el punto de vista notacional*. Departamento de Informática, Universidad Carlos III de Madrid, 2006. [Consultado: el 24 de febrero del 2013]. Disponible en: <http://www.ie.inf.uc3m.es/ggenova/pub-novatica2006b.html>

[16] *Comparativa de Herramientas UML de libre distribución*. Universidad politécnica de Madrid. [Consultado: el 24 de febrero del 2013]. Disponible en: <www.diatel.upm.es/malvarez/UML/comparativa.html>

[17] VIZCAÍNO, Aurora, GARCÍA, Félix y CABALLERO, Ismel. *Trabajando con Visual Paradigm for UML*. Universidad Cantabria, María Sierra, facultad de Ciencias. [Consultado: el 12 de febrero del 2013]. Disponible en: <http://personales.unican.es/ruizfr/is1/doc/lab/01/is1-p01-trans.pdf>

[18] EcuRed. *Sistema Gestor de Base de Datos*. [Consultado: el 15 de noviembre del 2012]. Disponible en: <http://www.ecured.cu/index.php/Sistema_Gestor_de_Base_de_Datos>

[19] SILVA VIERA, Diego. *Diferencias entre los SGBD´s*, 2012. [Consultado: el 24 de febrero del 2013]. Disponible en: <http://www.slideshare.net/diegosilvaviera1/diferencias-entre-los-sgbds>

[20] AGRAMONTE REY, Joel. *Comparación de SGBD*, 2012. [Consultado: el 24 de febrero del 2013]. Disponible en: <http://www.slideshare.net/wagramonter/sgbd-comparacion-13333341>

[21] TORRES GASTELÚ, Carlos Arturo y otros. *Comparación entre Oracle y MySQL*. Universidad Veracruzana, 2012. [Consultado: el 24 de febrero del 2013]. Disponible en: <http://www.slideshare.net/KARYBALK/bd-eq-3-actividad-extra-comparacion-oracle-y-mysql>

[22] PostgreSQL. [Consultado: el 12 de febrero del 2013]. Disponible en: <http://www.postgresql.org.es/sobre-postgresql>

[23] PostgreSQL. Universidad Mariano Gálvez, Guatemala, 2011. [Consultado: el 24 de febrero del 2013]. Disponible en: <http://www.slideshare.net/etumax/postgresql-9649848>

[24] SQLManager.net. *EMS SQL Manager for PostgreSQL*. [Consultado: el 24 de febrero del 2013]. Disponible en: <http://www.sqlmanager.net/en/products/postgresql/manager>

[25] PostgreSQL tools. [Consultado: el 15 de noviembre del 2012]. Disponible en: <http://www.pgadmin.org>

[26] Human Inference Enterprise. *DataCleaner: Powerful environment for Data Quality Analysis (DQA)*, 2011. [Consultado: el 24 de febrero del 2013]. Disponible en: <http://www.humaninference.com/media/66098/factsheet_datacleaner.uk.update2.pdf>

[27] Pentaho. *Pentaho Data Integrator*. [Consultado: el 3 de diciembre de 2012.] Disponible en: <http://www.pentaho.com/explore/pentaho-data-integration/>

[28] MEDINA MUSTELIER, Doris, IZNAGA GONZÁLEZ, Yonelbys, TÉLLEZ PÉREZ, Yuneimy y otros. *Herramientas de integración de datos*. Octubre, 2012. [Consultado: el 24 de febrero del 2013].

[29] CORNEJO, Grace, PESANTEZ, Joffre y SOLIS, Galo. *Herramientas PDI*. Pentaho Data Integration previous Kettle, 2009. Manual del ETL de Pentaho. [Consultado: el 23 de noviembre del 2012].

[30] *Ventajas de PENTAHO*. [Consultado: el 15 de noviembre del 2012]. Disponible en: < https://sites.google.com/site/pentahounicah/ventajas-de-pentaho>

[31] DAPENA BOSQUET, Isabel, MUÑOZ SAN ROQUE, Antonio y SÁNCHEZ MIRALLES, Álvaro. *Sistemas de Información Orientados a la Toma de Decisiones: el enfoque multidimensional*. Madrid, España, 2005. [Consultado: el 23 de mayo del 2013]

[32] WOLFF, Carmen Gloria. Modelamiento multidimensional. [Consultado: el 20 de mayo del 2013].

[33] RIVAS, Antonio. OLAP, MOLAP, ROLAP. Aprendiendo Business Intelligence, 2011. [Consultado: el 20 de mayo del 2013]. Disponible en: <http://www.bi.dev42.es/2011/02/23/olap-molap-rolap>

[34] SCHIEFER, J. y otros. *A holistic approach for managing requirements of Data Warehouse Systems*. Eighth Americas Conference on Information Systems, 2002. [Consultado: el 20 de mayo del 2013].

[35] Business Rules Group. [Consultado: el 12 de febrero del 2013]. Disponible en: <http://www.businessrulesgroup.org/defnbrg.shtml>

[36] JACOBSON, I. y otros. Ingeniería de Software Orientada a Objetos. *Un acercamiento a través de los casos de uso*. N.J., 1992.

[37] URQUIZU, Pau. *Claves subrogadas*. Business Intelligence fácil, 2009. [Consultado: el 26 de abril del 2013]. Disponible en: <http://www.businessintelligence.info/serie-dwh/claves-subrogadas.html>

[38] BERNABEU R., Dario. *Claves subrogadas*. DataPrix, Business Intelligence, 2010. [Consultado: el 26 de abril del 2013. Disponible en: <http://www.dataprix.com/blogs/bernabeu-dario/claves-subrogadas>

[39] Librería Digital. *Metadata resources*. [Consultado: el 22 de mayo del 2013]. Disponible en: <http://www.ifla.org/II/metadata.htm>

[40] MURTHA BACA, ed. *Introduction to Metadata: Pathways to Digital Information*. Los Angeles: Getty Research Institute, 2001. [Consultado: el 22 de mayo del 2013]. Disponible en: <http://www.getty.edu/research/institute/standards/intrometadata/>

BIBLIOGRAFÍA

1. Cubadebate. *OMS alerta el aumento de muertes por enfermedades no transmissible.* Noticias, 2011. [Consultado: el 1 de noviembre del 2012]. Disponible en: <http://mesaredonda.cubadebate.cu/noticias/2011/04/27/oms-alerta-aumento-muertes-por-enfermedades-no-transmisibles/>
2. University of MaryLand Medical Center. *Cáncer – Overview*, 2011. [Consultado: el 9 de noviembre del 2012]. Disponible en: <http://www.umm.edu/>
3. Organización Mundial de la Salud. *Cáncer*, 2013. Disponible en: <http://www.who.int/mediacentre/factsheets/fs297/es/>
4. Organización Mundial de la Salud. *Datos y cifras del cáncer*, 2011. Disponible en: <http://www.who.int/cancer/about/facts/es/>
5. Centro de Inmunología Molecular. *Productos Comerciales*, 2011. Disponible en: <http://www.cim.co.cu/productos_comerciales.php>
6. SAROKA, Raúl Horacio. *Sistemas de información en la era digital*. Programa Avanzado de Perfeccionamiento en Management de la fundación OSDE, con la supervisión académica y certificación de la Universidad Nacional de San Martín. Argentina, 2002. [Consultado: el 1 de febrero del 2012]. Disponible en: <http://www.fundacionosde.com.ar/pdf/biblioteca/Sistemas_de_informacion_en_la_era_digital-Modulo_I.pdf>
7. BERNABEU, Ricardo Dario. *Data warehousing: Investigación y Sistematización de Conceptos-Hefesto: Metodología propia para la Construcción de un Data Warehouse*. Córdoba, Argentina, 2009.
8. CURTO, Josep. *Diseño de un Data Warehouse: estrella y copo de nieve*. Information Management, 19 de noviembre de 2007. [Consultado: el 13 de febrero del 2013]. Disponible en: <http://informationmanagement.wordpress.com/2007/11/19/diseno-de-un-data-warehouse-estrella-y-copo-de-nieve/>
9. Visual Paradigm. *Visual Paradigm*. [Consultado: el 12 de febrero del 2013]. Disponible en: <http://www.visual-paradigm.com/product/vpuml/>
10. Colectivo de autores. *Auditoría de sistemas*. Universidad de Caldas, 2009.

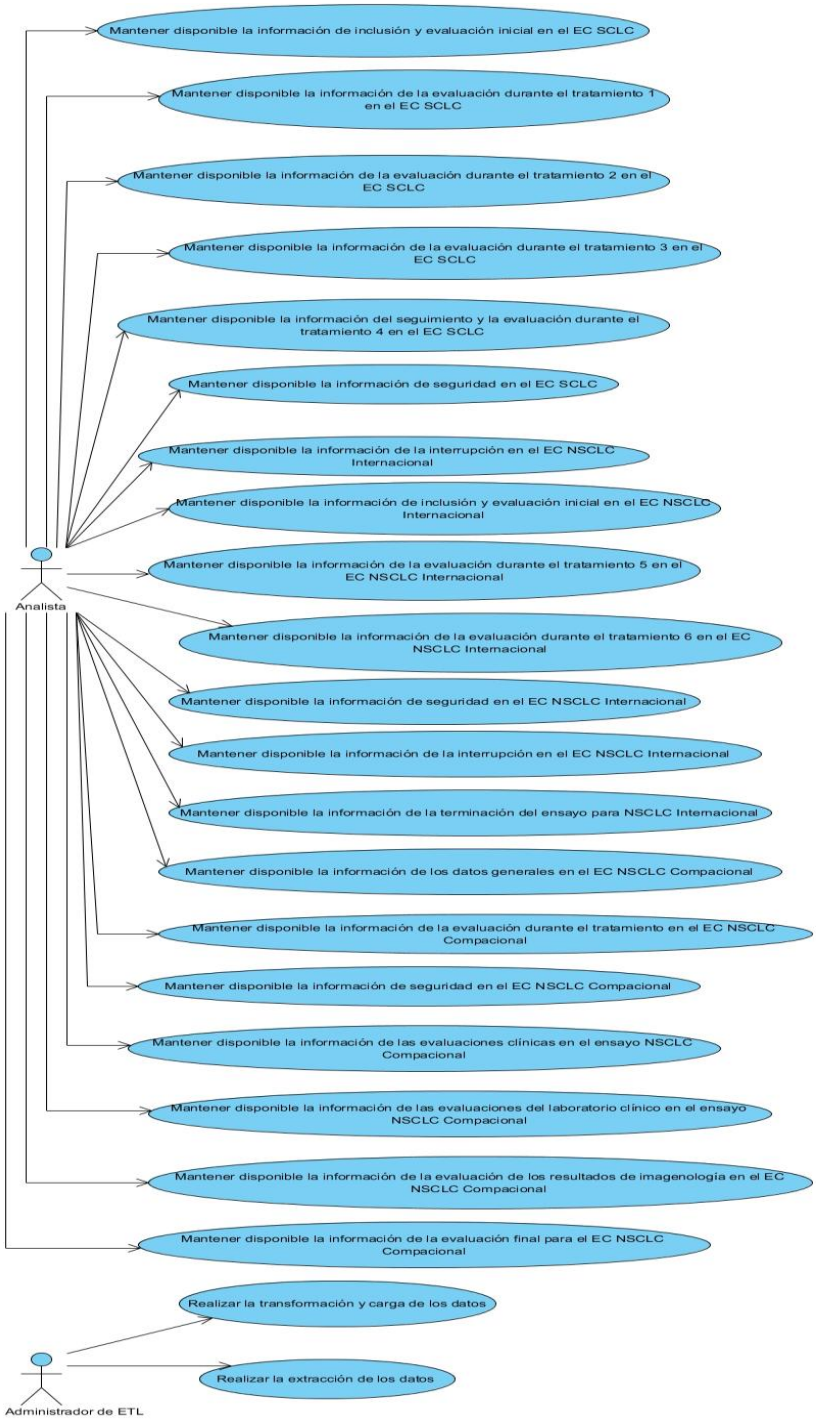
11. ZEPEDA SÁNCHEZ, Leopoldo Zenaido. Universidad Politécnica de Valencia. Departamento de Sistemas Informáticos y Computación, 2008. [Consultado: el 22 de febrero del 2013]. Disponible en: <http://riunet.upv.es/bitstream/handle/10251/2506/tesisUPV2841.pdf>
12. MEDINA MUSTELIER, Doris, ORTIZ, Julio Ernesto, GONZALES HERNÁNDEZ, Yanisbel y otros. *Metodología para el desarrollo de soluciones de Almacenes de Datos e Inteligencia de Negocios en DATEC*. [Consultado: el 15 de febrero del 2013].Ciudad de La Habana, 2010.
13. *Requerimientos para la notificación y el reporte de eventos adversos graves e inesperados en los ensayos clínicos*. Regulación no. 45, 2007. [Consultado: el 9 de noviembre del 2012].
14. EcuRed. [Consultado: el 9 de noviembre de 2012]. Disponible en: <http://www.ecured.cu/index.php/Almac%C3%A9n_de_Datos>
15. ETL-Tools. *Info Bussines Intelligence-Almacenes de datos-ETL*. [Consultado: el 3 de enero de 2013]. Disponible en: <http://etl-tools.info/es/bi/almacendedatos_arquitectura.htm>
16. H. INMON, William. *Building the Data Warehouse*. s.l.: Wiley Publishing. ISBN: 0-471-08130-2.
17. KIMBALL, Ralph, REEVES, Laura, ROSS, Margy y THORNTWHAITE, Warren. *The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing, and Deploying Data Warehouses*. s.l.: Wiley Publishing. ISBN: 978-0-471-25547-5.
18. KIMBALL, Ralph y ROSS, Margy. *The Data Warehouse Toolkit*. s.l.: Wiley Computer Publishing. ISBN: 0-471-20024-7.
19. KIMBALL, Ralph y CASERTA, Joe. *The Data Warehouse ETL Toolkit Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data*. s.l.: Wiley Publishing. ISBN: 0-764-57923-1.
20. PENTAHO. [Consultado: el 12 de enero de 2013]. Disponible en: <http://www.gravitar.biz/index.php/herramientas-bi/pentaho/caracteristicas-pentaho/>
21. PRESSMAN, R.S. *Ingeniería del software. Un enfoque práctico.* s.l.:McGraw-Hill. ISBN: 84-481-3214-9.
22. PgAdmin. [Consultado: el 3 de enero de 2013]. Disponible en: <http://www.pgadmin.org/>
23. PostgreSQL. [Consultado: el 3 de enero de 2013]. Disponible en: <http://www.postgresql.org/>
24. *Centro de Tecnologías de Gestión de Datos*. [Consultado: el 9 de noviembre del 2012]. Disponible en: <http://gespro.datec.prod.uci.cu/>
25. LAPORTE, Joan Ramón. *Principios Básicos de Investigación Clínica.* Fundación Instituto Catalán de Farmacología, 2007. [Consultado: el 5 de enero del 2012]. Disponible en: <http://www.icf.uab.es/llibre/llibre.htm>

26. HERNÁNDEZ RIVAS, Jesús M., GARCÍA ORTIZ, Luis. *Metodología en investigación clínica. Fases del estudio de investigación (II)*. [Consultado: el 30 de octubre del 2012]. Disponible en: <http://campus.usal.es/~dermed/Modulo%203%202%20Fases%20del%20estudio%20de%20investigaci%F3n%20II%202%20df.pdf>
27. CHICIAZA, Janneth Alexandra. *Guía didáctica: Modelamiento de Datos*. (Ingeniería en Informática). Ecuador. Universidad Técnica Particular de Loja, 2011. [Consultado: el 15 de noviembre del 2012]. Disponible en: <http://rsa.utpl.edu.ec/material/208/G181003.1.pdf>
28. SAROKA, Raúl Horacio. *Sistemas de información en la era digital*. Programa Avanzado de Perfeccionamiento en Management de la fundación OSDE, con la supervisión académica y certificación de la Universidad Nacional de San Martín, Argentina, 2002. [Consultado: el 1 de febrero del 2012]. Disponible en: <http://www.fundacionosde.com.ar/pdf/biblioteca/Sistemas_de_informacion_en_la_era_digital-Modulo_I.pdf>
29. FUSTER, G. *Evaluación comparativa de herramientas CASE para UML desde el punto de vista notacional*. Departamento de Informática, Universidad Carlos III de Madrid, 2006. [Consultado: el 24 de febrero del 2013]. Disponible en: <http://www.ie.inf.uc3m.es/ggenova/pub-novatica2006b.html>
30. *Comparativa de Herramientas UML de libre distribución*. Universidad politécnica de Madrid. [Consultado: el 24 de febrero del 2013]. Disponible en: <www.diatel.upm.es/malvarez/UML/comparativa.html>
31. Visual Paradigm. *Visual Paradigm*. [Consultado: el 12 de febrero del 2013]. Disponible en: <http://www.visual-paradigm.com/product/vpuml/>
32. BERNABEU, Ricardo Dario. *Data Warehousing: Investigación y Sistematización de Conceptos* - Hefesto: Metodología propia para la Construcción de un Data Warehouse. Córdoba, Argentina, 2009.
33. CLEMENTE, Gerardo. *Un Sistema para el Mantenimiento de Almacenes de Datos*, 2008.
34. CURTO, Josep*. Data Warehouse y Datamart*. Information Management Data Warehousing, 2007. Disponible en: <http://informationmanagement.wordpress.com/2007/10/07/data-warehousing-data-warehouse-y-datamart/>
35. GONZÁLEZ HERNÁNDEZ, Yanisbel. *Propuesta de metodología para el desarrollo de Almacenes de Datos en DATEC*. La Habana, 2011.

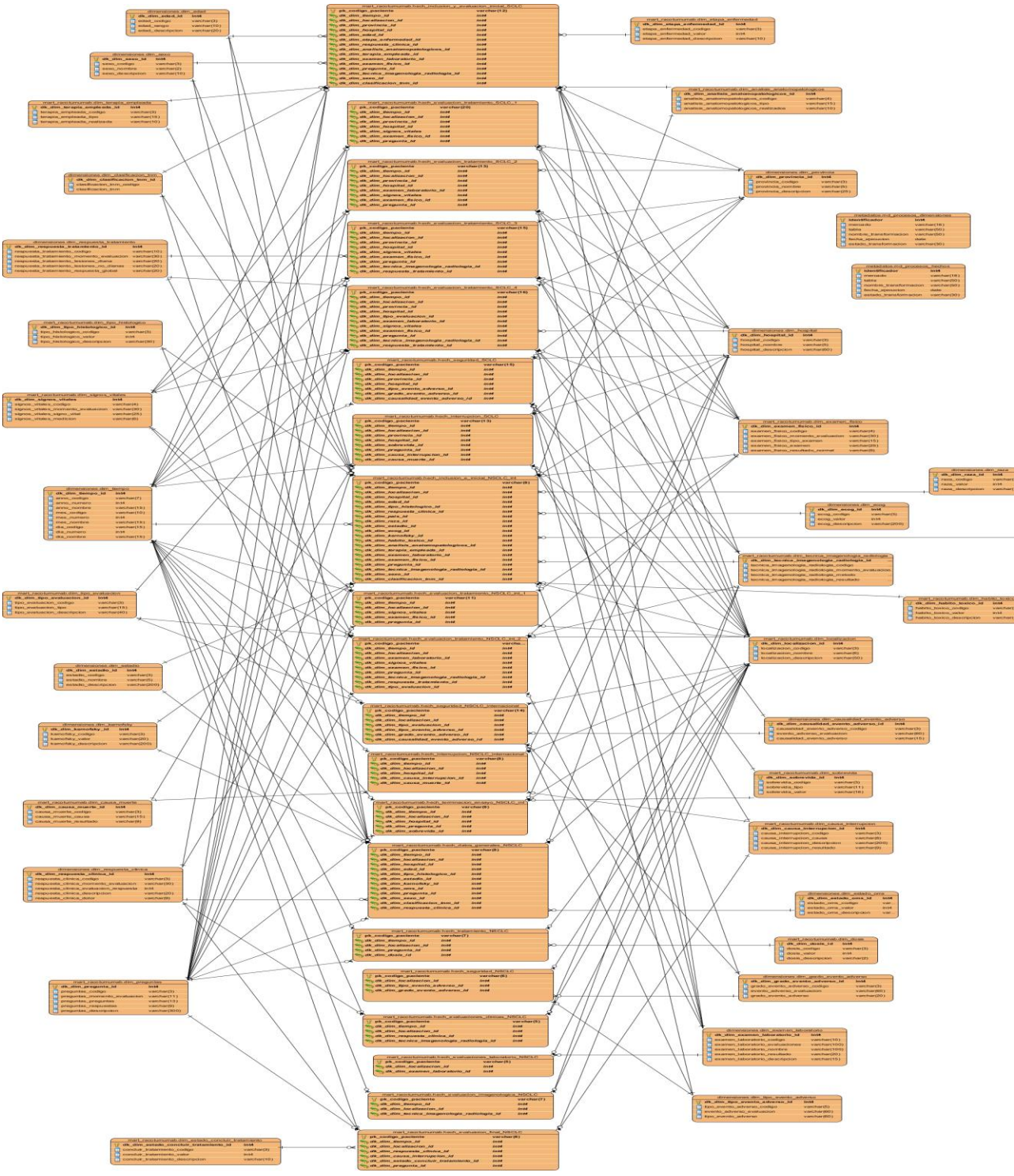
36. SCRIVE, Michael. *The Logic and Methodology of Checklists*, 2000.
37. DataCleaner, 2008. Disponible en: <http://datacleaner.eobjects.org>
38. *La plataforma Pentaho Open Source Business Intelligence*. Portada sobre la plataforma Pentaho Open Source Business Intelligence. [Consultado: el 30 de marzo del 2013]. Disponible en: <http://pentaho.almacen-datos.com>
39. UCI. [Consultado: el 2 de octubre del 2013]. Disponible en: < http://www.uci.cu>

ANEXOS

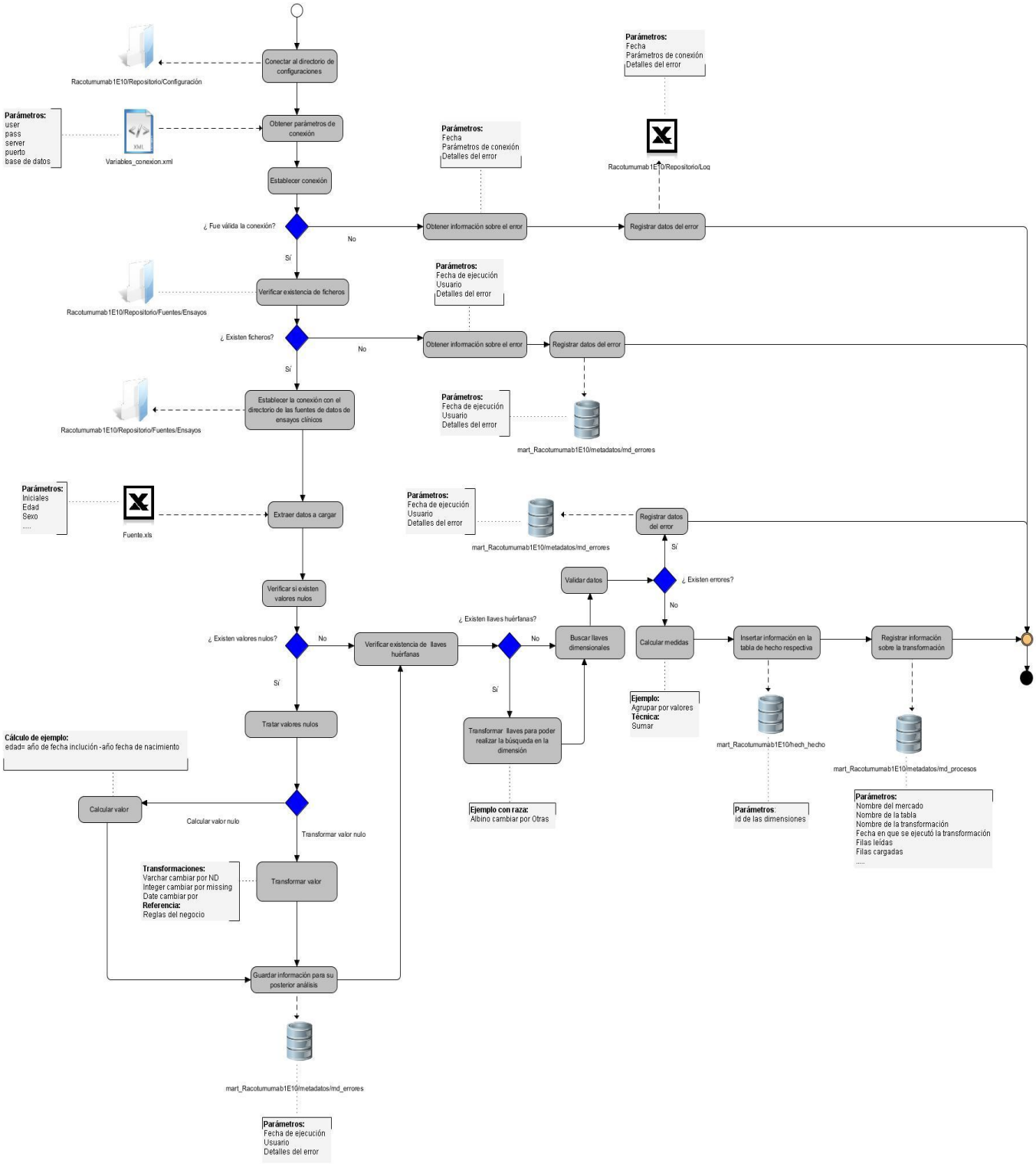
Anexo 1. Diagrama de CU del Sistema



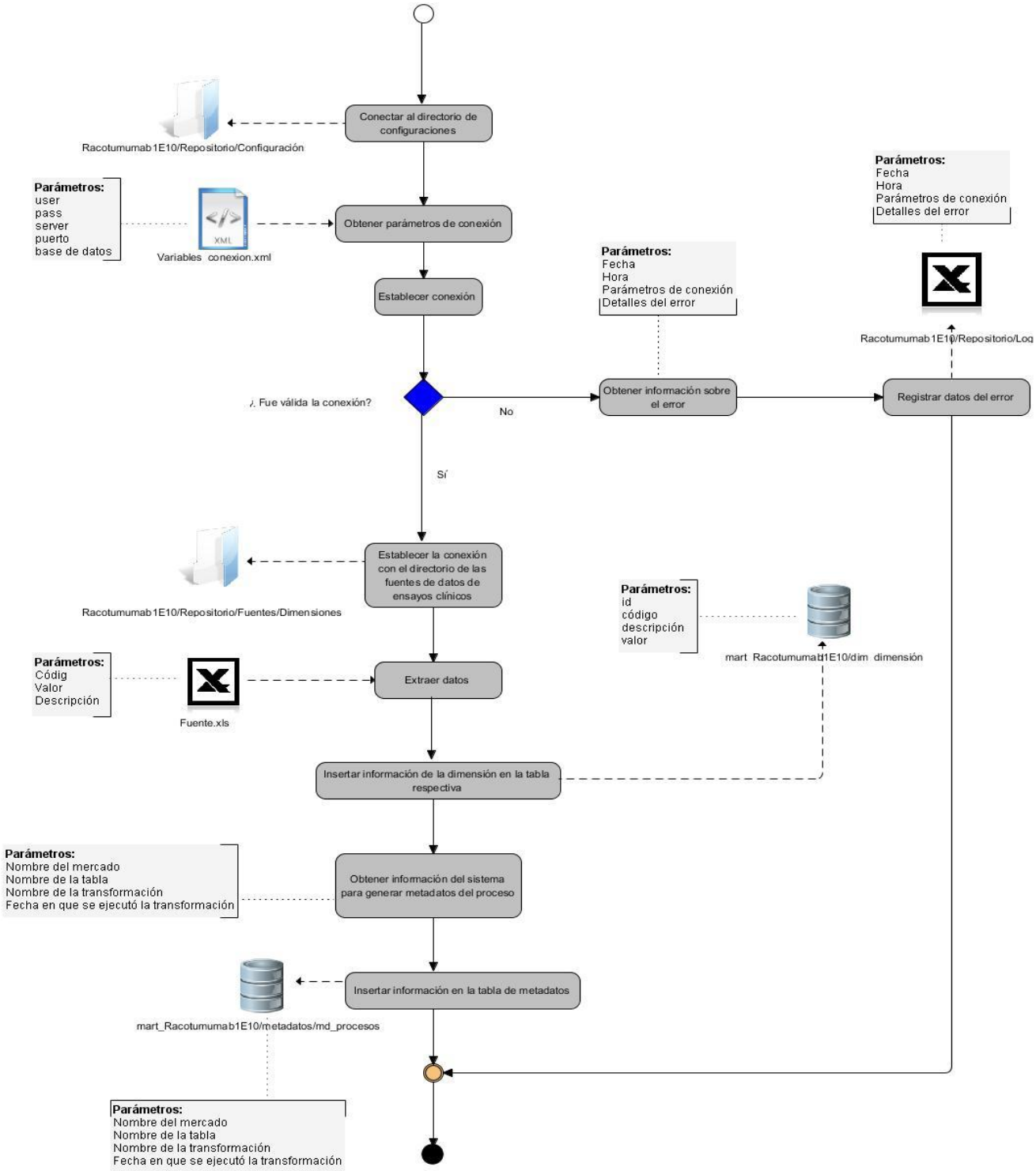
Anexo 2. Modelo de Datos



Anexo 3. Diseño de los procesos de integración para los hechos



Anexo 4. Diseño de los procesos de integración para las dimensiones



Anexo 5. Estructura de los datos en el almacén

Estructura	Descripción	Ejemplo
Tablas de hechos	Todas las tablas de hechos tendrán una cadena que demuestra que son hechos y el nombre que se corresponde con el mismo.	hech_<nombre del hecho>
Tablas de dimensiones	Todas las tablas de dimensiones tendrán una cadena que demuestra que son dimensiones y el nombre que se corresponde con las mismas.	dim_<nombre de la dimensión>
Llaves primarias	Todas las llaves primarias tendrán una cadena que demuestra que son llaves primarias y el nombre de la tabla a la que pertenecen.	dk_dim_<nombre de la tabla>_id
Atributos compuestos	En los atributos donde el nombre es compuesto se debe especificar el primer componente del atributo separado del segundo será por un carácter de _.	<Nombre de la tabla>_<Nombre del tributo>

GLOSARIO DE TÉRMINOS

Add-on: se entiende, del inglés, como una extensión o añadidura puede referirse a una mejora instalable para los proyectos de la Fundación Mozilla. También conocidos como extensiones, plugins, snap-ins, etc, son programas que sólo funcionan anexados a otro y que sirven para incrementar o complementar sus funcionalidades.

API: es el conjunto de funciones y procedimientos (o métodos, en la programación orientada a objetos) que ofrece cierta biblioteca para ser utilizado por otro software como una capa de abstracción. Son usadas generalmente en las bibliotecas.

CASE: son diversas aplicaciones informáticas destinadas a aumentar la productividad en el desarrollo de software reduciendo el costo de las mismas en términos de tiempo y de dinero. Estas herramientas pueden ayudar en todos los aspectos del ciclo de vida de desarrollo del software en tareas como el proceso de realizar un diseño del proyecto, cálculo de costos, implementación de parte del código automáticamente con el diseño dado, compilación automática, documentación o detección de errores entre otras.

Control de Concurrencia Multiversión: es una técnica de concurrencia optimista en donde ninguna tarea o hilo es bloqueado mientras se realiza una operación en la tabla, porque el otro hilo usa su propia copia del objeto dentro de una transacción.

Indicadores: son valores numéricos y representan lo que se desea analizar concretamente, por ejemplo: promedios, cantidades, sumatorias y fórmulas.

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.

PDF: (Portable Document Format, formato de documento portátil) es un formato de almacenamiento de documentos digitales independiente de plataformas de software o hardware. Este formato es de tipo compuesto (imagen vectorial, mapa de bits y texto).

Perspectivas: no son más que los objetos mediante los cuales se quiere examinar los indicadores, con el fin de responder a las preguntas planteadas. Teniendo en cuenta las características de los MD de ser variables en el tiempo, se sugiere incluir siempre el tiempo como una perspectiva más.

SSL: es un protocolo desarrollado por Netscape Communications Corporation para dar seguridad a la transmisión de datos en transacciones comerciales en Internet. Utilizando la criptografía de llave pública, SSL provee autentificación del servidor, encriptar de datos, e integridad de los datos en las comunicaciones cliente/servidor.

wxWidgets: son bibliotecas multiplataforma y software libre, que se utilizan para el desarrollo de interfaces gráficas programadas en lenguaje C++. Están publicadas bajo licencia pública general limitada de GNU (LGPL), similar a la Licencia pública general de GNU (GPL) con la excepción de que el código binario producido por el usuario a partir de ellas, puede ser propietario, permitiendo desarrollar aplicaciones empresariales sin coste. Proporcionan una interfaz gráfica basada en las bibliotecas ya existentes en el sistema, con lo que se integran de forma óptima y resultan muy portables entre distintos sistemas operativos.