Universidad de las Ciencias Informáticas

Facultad 6

Trabajo de Diploma para optar por el título de

Ingeniero en Ciencias Informáticas

Título: *Subsistemas de almacenamiento e integración del producto hr3 para los ensayos clínicos del Centro de Inmunología Molecular.*

Autores:

Dunia Espinosa López

Dianeyis Rivero Chacón

Tutores:

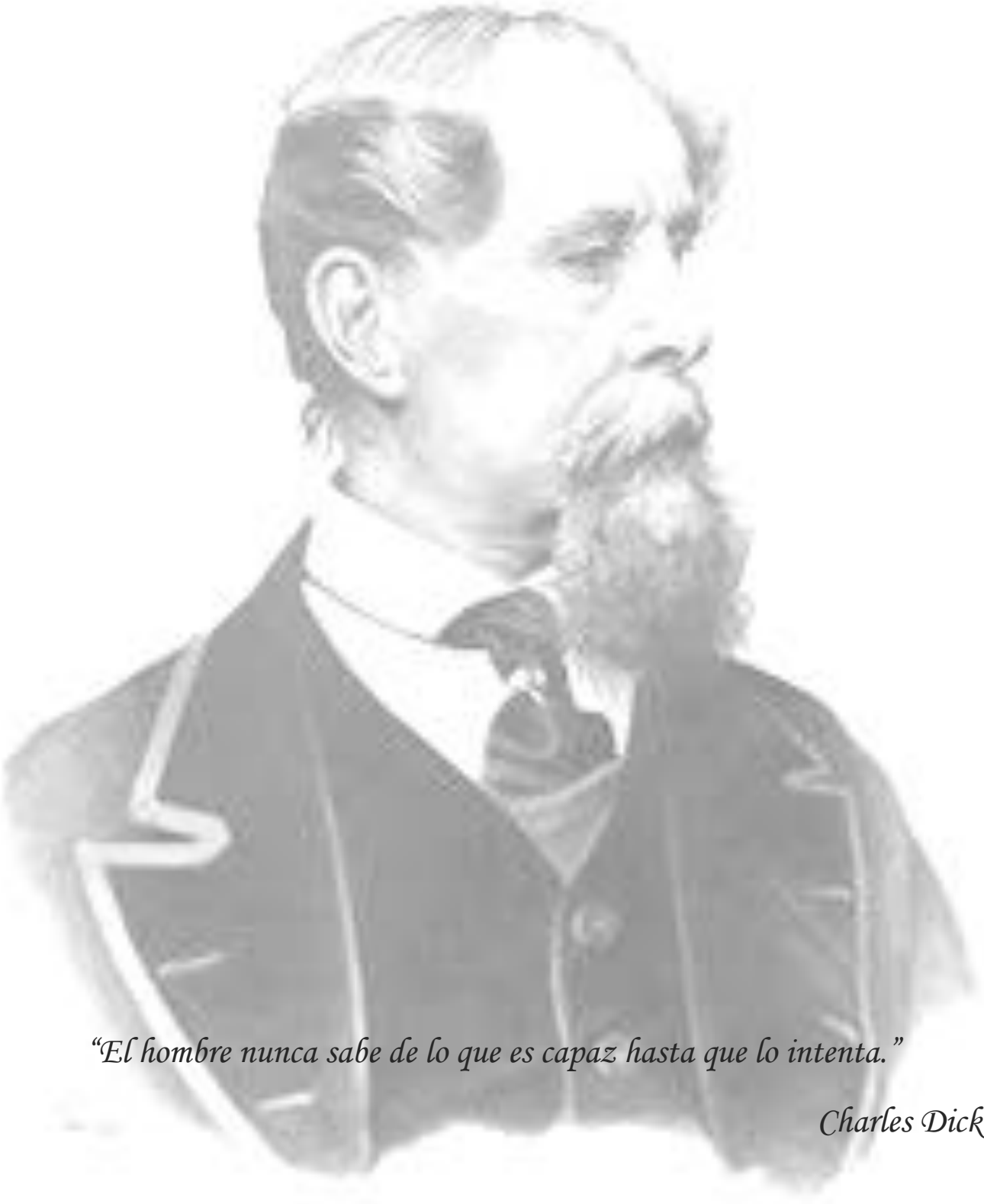
Msc. Yadira Barroso Rodríguez

Ing. Adilen Guerra Sanabria

Ing. Yanet Cardoso García

La Habana, Junio 2014

"Año 56 de la Revolución"



"El hombre nunca sabe de lo que es capaz hasta que lo intenta."

Charles Dickens

Declaración de Autoría.

Declaramos, Dianeyis Rivero Chacón y Dunia Espinosa López, ser autoras del presente trabajo de diploma y concedemos a la Universidad de las Ciencias Informáticas los derechos patrimoniales del mismo con carácter exclusivo.

Para que así conste firmamos a los ___días del mes de _____ del año_____.

_____	_____
Firma de la autora	Firma de la autora
Dianeyis Rivero Chacón	Dunia Espinosa López

_____	_____	_____
Firma del Tutor	Firma del Tutor	Firma del Tutor
MsC. Yadira Barroso Rodríguez	Ing. Adilen Guerra Sanabria	Ing. Yanet Cardoso García

Tutor: MsC. Yadira Barroso Rodríguez.

Especialidad de graduación: Máster en Informática Aplicada.

Correo Electrónico: ybarroso@uci.cu.

Tutor: Ing. Adilen Guerra Sanabria.

Especialidad de graduación: Ingeniería en Ciencias Informáticas.

Correo Electrónico: agsanabria@uci.cu.

Tutor: Ing. Yanet Cardoso García.

Especialidad de graduación: Ingeniería en Ciencias Informáticas.

Correo Electrónico: ycardosog@uci.cu.

Agradecimientos.

Quiero agradecer en primer lugar a mi mamá por guiarme siempre por el buen camino y estar conmigo en las buenas y las malas, por ayudarme a conseguir mi sueño, por alegrarme y darme fuerzas cuando más lo necesitaba. Muchas gracias mamita. Te quiero.

A mi abuela y hermanita, que aunque no pudieron estar hoy aquí, les quiero dar muchas gracias por estar conmigo en todos los momentos de mi vida y apoyarme siempre. Las quiero mucho.

A mi novio o mejor dicho a mi chitito por soportarme en estos cuatro años, por aguantarme en mis momentos de estrés y malos genios, por estar conmigo en las buenas y las malas, por permitirme ser parte de su vida y apoyarme en mis decisiones. Te quiero mucho mi amor.

A mi familia en general y a mi suegra por ser aquellas personas que permitieron mis malas crianzas. Los quiero.

A mi familia de la UCI, que estuvieron conmigo en mis desmayos, fiestas y estudios, ellos son la Choty, Javico, Yadian, Dunia, Katy y su bebito, Dami y a mi hermanito Alfredo. Gracias.

A mis dos grandes amigas o mis dos hermanas, que me vienen soportando desde la secundaria e IPI por estar siempre en el momento que más lo necesitaba, gracias Dami y Haileen.

A mi dúo de tesis y amiga Duni, por soportarme cuando tenía deseos de llorar por la tesis y por permitirme cumplir mi sueño. Gracias amiga.

A mis otras amistades que cuando no me decían enana café, me decían pelua, enana sácame la rana, esas que nunca voy a olvidar gracias Yuni, Alberto, Lili, Idalberto, Jeandy (Bicho) , a roxy y a mi amigo Jaciel. Gracias a todos.

A Nelsy por soportarme desde la secundaria hasta ahora. Al guajiro, la china, y sus bebés, por permitirme ser parte de su familia. A mis tutoras por ayudarme y permitirme molestarlas cuando estaban de descansos. A el profe Omar que siempre va a ser un ejemplo para mí. Gracias a todos.

A mi grupo por compartir conmigo estos cinco años de universidad y a todos aquellos que de una forma u otra me apoyaron en mi carrera. ***Dianeyis.***

Agradecimientos.

Primero que todo quiero agradecer a todas las personas que han sido mi apoyo y guía en los momentos más difíciles, en especial a mi mamá, por ser lo más lindo de mi vida y mi fuerza. Gracias por estar siempre a mi lado.

A mi papá que aunque sigue peleonero sé que se siente muy orgulloso de mi. A mi hermano y abuelos que tanto los quiero, por su apoyo incondicional.

A mi novio, gracias por regalarme 4 años de tu vida, por ser tan paciente conmigo, por estar junto a mí en las buenas y en las malas, gracias por ser mi amigo, mi novio, mi todo.

A mis suegros por dejarme formar parte de su familia y por su amor y apoyo les mando muchos besitos. A mi cuñada a Yoelkys, Yoelito, Yudi, Elvira y a toda la familia en general, le doy gracias por todo el cariño que me han dado en este tiempo.

A mi dúo de tesis Dianeyis por todos los momentos buenos y malos, por su preocupación, por ayudarme a realizar esta meta que nos hemos propuesto.

A mis amigos, por ser parte de mi vida, de mis momentos tristes y alegres, por apoyarme, por siempre estar ahí, en especial a la Choti, Kati, Roxana, Yanislet, Leydis, Alian, Yuni, Dami, Alfredo, Liliana, Eduar, Arianne, Anneris, a mis compañeros de aula y a todas esas bellas personas con las que he compartidos durante estos 5 años, por el apoyo ofrecido y por haber tenido la dicha de haberlos conocido.

A la UCI en general y a todos los profesores que fueron parte de mi formación como profesional, a mis tutoras, gracias por su paciencia y dedicación, al oponente por sus certeras recomendaciones, al tribunal en general gracias por su imparcialidad.

Dunia.

Dedicatoria.

Dedico esta tesis a cuatro personitas muy importantes en mi vida, a mi abuelita, mi hermana, mi novio y a la persona que más quiero en este mundo a mi mamita linda.

Dianeyis.

Les dedico el presente trabajo a toda mi familia y a mi novio por todo el amor y apoyo que me han brindado, en especial a mi mamá que es mi razón de ser.

Dunia.

Resumen.

La presente investigación surge como parte de la relación que existe entre la Universidad de las Ciencias Informáticas y el Centro de Inmunología Molecular. 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. Para resolver la problemática existente se propuso desarrollar los subsistemas de almacenamiento e integración del producto hr3, permitiendo la estandarización de los datos y el almacenamiento homogéneo de la información que se maneja en el centro. Luego de realizar un estudio preliminar se seleccionó la Metodología de Desarrollo para Proyectos de Almacenes de Datos del Centro de Tecnologías y Gestión de Datos y las herramientas Visual Paradigm, PostgreSQL, PgAdmin, DataCleaner, y Pentaho Data Integration. Una vez conocidas las peculiaridades del negocio se llevaron a cabo las etapas de análisis, diseño, implementación y prueba, obteniéndose una base de datos poblada en su totalidad. La solución permite realizar análisis estadísticos e históricos de los principales indicadores en el área de ensayos clínicos, facilitando el proceso de toma de decisiones.

Palabras claves: almacén de datos, Centro de Inmunología Molecular, ensayos clínicos, estandarización.

Abstract.

This research began as part of the relationship between the University of Informatics Sciences and the Center of Molecular Immunology. This center presents a number of challenges to the management of data by specialists, which encourages the loss of valuable and useful information. To solve the existing problems the storage and integration subsystems developing of hr3 product were proposed, allowing standardization of data and information homogeneous storage handled at the center. After conducting a preliminary study for developing data warehouses Data Management and Technologies Center a methodology was selected and Visual Paradigm, PostgreSQL, PgAdmin, DataCleaner and Pentaho Data Integration tools were selected. Once known the peculiarities of business stages of analysis, design, implementation and testing were conducted, yielding an entirely populated database. The solution allows to execute statistical and historical analysis of the main indicators in the area of clinical trials, facilitating the process of decision making.

Keywords: data warehouse, Center of Molecular Immunology, clinical trials, standardization.

Índice de contenido.

Índice de Figuras.

Introducción.

En la sociedad actual, el gran avance de las Tecnologías de la Información y las Comunicaciones (TICs) ha sido aprovechado ampliamente dentro del campo de la medicina, destacando el papel que juegan las ciencias informáticas, apoyando significativamente la toma de decisiones.

En Cuba se han creado varios centros biotecnológicos y científicos - investigativos asociados a la producción de fármacos; dentro de los cuales se encuentran: el Instituto de Investigación Carlos J. Finlay, el Centro de Ingeniería Genética y Biotecnología (CIGB) y el Centro de Inmunología Molecular (CIM).

La Universidad de las Ciencias Informáticas (UCI) juega un papel protagónico en el avance científico-técnico de Cuba, conjuntamente con el Ministerio de Salud Pública (MINSAP) y algunas empresas de *software* y *hardware*. La UCI está estructurada por facultades y en estas existen distintos centros que desarrollan diversas soluciones informáticas. Entre ellos se encuentra el Centro de Tecnologías de Gestión de Datos (DATEC), que cuenta con un departamento especializado en el desarrollo de soluciones de Inteligencia de Negocios y Almacenes de Datos (AD).

La UCI actualmente le brinda servicios al CIM, el cual tiene como principal objetivo la búsqueda de nuevos productos para el diagnóstico y tratamiento del cáncer y enfermedades relacionadas con el sistema inmune. Los proyectos de investigación básicos están concentrados en la inmunoterapia del cáncer, especialmente en el desarrollo de vacunas moleculares, ingeniería de anticuerpos, ingeniería celular, bioinformática y regulación de la respuesta inmune. Se encarga además de gestionar, almacenar y analizar toda la información recogida en los Ensayos Clínicos (EC), siempre que se aplique un producto determinado.

Uno de los fármacos que se está administrando hoy en día es el hr3, Anticuerpo Monoclonal Humanizado (ACM) que genera cantidad de información y con el paso del tiempo el volumen de datos se ha incrementado considerablemente, los mismos se encuentran almacenados en el sistema informático EpiData, el cual exporta ficheros con extensión *xls*.

Los datos que genera cada uno de los EC no están estandarizados, es decir, presentan diferentes formatos y estructuras. Esto implica que se convierta en un problema el análisis de la documentación para la toma de decisiones en la práctica clínica; así como la necesidad de involucrar una gran cantidad de personas y tiempo para este estudio. Trayendo consigo que una vez terminada de examinar toda la

información muchas de las decisiones tomadas sean obsoletas, además, debido al gran cúmulo de los datos, la institución presenta problemas con el manejo de la información para confeccionar reportes, realizar análisis y consultas que se llevan a cabo sobre los datos recopilados y presentar los indicadores relacionados con dichos ensayos. Por otra parte, no permite realizar análisis estadísticos entre diferentes EC, lo que conlleva a la pérdida de información útil y valiosa.

Por lo anteriormente descrito se plantea como **problema de la investigación:** ¿Cómo lograr el almacenamiento de forma homogénea de los datos asociados al producto hr3 para los Ensayos Clínicos del Centro de Inmunología Molecular?

La investigación tiene como **objeto de estudio:** los mercados de datos, enmarcado en el **campo de acción:** subsistemas de almacenamiento e integración para los Ensayos Clínicos del Centro de Inmunología Molecular.

Para solucionar el problema de la investigación planteado, se identifica como **objetivo general:** Desarrollar los subsistemas de almacenamiento e integración del producto hr3, para lograr su almacenamiento de forma homogénea.

Para orientar la investigación hacia el cumplimiento del objetivo general se plantean las siguientes **preguntas científicas**:

- ✓ ¿Qué fundamentos teóricos sustentan a los subsistemas de almacenamiento e integración para los Ensayos Clínicos del Centro de Inmunología Molecular?
- ✓ ¿Cuáles son las peculiaridades que deben tener los subsistemas de almacenamiento e integración para lograr el almacenamiento de forma homogénea de los datos asociados al producto hr3 para los Ensayos Clínicos del Centro de Inmunología Molecular?
- ✓ ¿Cómo organizar el proceso de desarrollo de los subsistemas de almacenamiento e integración de los datos asociados al producto hr3 para los Ensayos Clínicos del Centro de Inmunología Molecular?
- ✓ ¿La solución desarrollada de los subsistemas de almacenamiento e integración garantiza el almacenamiento homogéneo de los datos asociados al producto hr3 para los Ensayos Clínicos del Centro de Inmunología Molecular?

Para dar cumplimiento al objetivo general se proponen las siguientes **tareas de la investigación:**

1. Caracterización de la metodología, tecnología y herramientas a utilizar en el desarrollo de los subsistemas de almacenamiento e integración del producto hr3 permitiendo determinar cuáles se utilizarán.
2. Relación de los requisitos para definir las necesidades del cliente.
3. Definición de la arquitectura del mercado de datos identificando los subsistemas fundamentales que formarán parte de la solución.
4. Descripción de los casos de uso de los subsistemas de almacenamiento e integración del producto hr3 para determinar cada una de las funcionalidades del sistema.
5. Definición de los hechos, las medidas y las dimensiones de los subsistemas de almacenamiento e integración del producto hr3 para identificar los elementos que forman parte del modelo lógico de datos.
6. Realización del diseño del modelo lógico para determinar los elementos que conforman el modelo físico.
7. Realización del perfilado de datos para garantizar la limpieza y calidad de los datos.
8. Realización del diseño del subsistema de integración para definir cómo se realizará la carga de las dimensiones y los hechos.
9. Implementación del modelo físico para los subsistemas de almacenamiento e integración del producto hr3 permitiendo tener una base de datos bien estructurada.
10. Implementación del subsistema de integración para poblar la base de datos.
11. Realización del diseño de los casos de prueba para los subsistemas de almacenamiento e integración del producto hr3.
12. Aplicación de las listas de chequeo para comprobar la calidad de los artefactos de Extracción, Transformación y Carga (ETL).
13. Aplicación de los casos de prueba para probar que los datos extraídos de la fuente fueron cargados en su totalidad a los subsistemas de almacenamiento e integración del producto hr3.

Declaración de los métodos de investigación.

En la investigación fueron utilizados métodos teóricos y empíricos con el propósito de entender el negocio, sus particularidades, prioridades y elementos esenciales para poder darle solución al problema planteado en la investigación.

Como métodos teóricos se utilizaron:

Histórico-lógico: se utilizó en el estudio realizado sobre los AD y los Mercados de Datos (MD) con el fin de analizar su evolución y desarrollo en la actualidad, así como los logros y limitaciones de los mismos en Cuba y en el mundo.

Analítico-sintético: este se utilizó en el estudio realizado para analizar y caracterizar los elementos más importantes relacionados con el objeto de estudio, además ayudó en el establecimiento de los métodos, herramientas y procedimientos más factibles para la implementación de los subsistemas de almacenamiento e integración del producto hr3.

Método de modelación: este método fue utilizado para darle cumplimiento a las tareas de la investigación, asociadas al diseño e implementación de los subsistemas de almacenamiento e integración del producto hr3. Se utilizó para el diseño de los diferentes modelos de la solución, con el objetivo de tener una representación gráfica para poder implementar los procesos del negocio.

La presente investigación se encuentra desglosada en tres capítulos estructurados de la siguiente manera:

Capítulo 1. Fundamentos teóricos de los mercados de datos.

En este capítulo se abordan definiciones y conceptos sobre los AD, MD y la gestión de los EC del producto hr3 en el CIM. Incluyendo además, características, ventajas y desventajas de los mismos. Se realiza un estudio bibliográfico de la metodología, herramientas y tecnología que se utilizarán para el desarrollo de los subsistemas de almacenamiento e integración del producto hr3.

Capítulo 2. Análisis y diseño de los subsistemas de almacenamiento e integración del producto hr3.

En este capítulo se realiza el análisis para comprender mejor el negocio, a partir del cual son identificados los requisitos del sistema, los mismos son agrupados según los diferentes criterios existentes. Se identifican las reglas del negocio, tablas de dimensiones, de hechos y las medidas que tributan para

obtener el modelo de datos. Se definen los requisitos de información, los funcionales y no funcionales, así como la conformación del diagrama de casos de uso del sistema.

Capítulo 3. Implementación y pruebas de los subsistemas de almacenamiento e integración del producto hr3.

En este capítulo se realiza la implementación y prueba de los subsistemas de almacenamiento e integración del producto hr3. Se aborda todo lo referente a la implementación de la estructura física de la solución y la realización de los procesos de ETL. Se definen los estándares de codificación y la construcción del modelo físico. Se exponen las pruebas realizadas a los subsistemas, así como los resultados obtenidos en cada una de ellas. Dichas pruebas permiten encontrar y corregir no conformidades existentes, obteniéndose como resultado una aplicación con mayor calidad.

Capítulo 1. Fundamentos teóricos de los mercados de datos.

Introducción.

En este capítulo se abordan definiciones y conceptos importantes sobre los AD, MD e información necesaria relacionada con los EC del producto hr3. Asimismo, se hace un estudio bibliográfico de las metodologías existentes a nivel mundial y en Cuba para el desarrollo de un MD. Se realiza una explicación detallada de todos los aspectos a tener en cuenta en la etapa de diseño, además se recogen las características, ventajas y desventajas de las distintas herramientas a utilizar en el desarrollo de la investigación.

1.1 Ensayo clínico.

Un EC es un estudio que permite a los médicos determinar si un nuevo tratamiento, medicamento o dispositivo contribuirá a prevenir, detectar o tratar una enfermedad. Los ensayos clínicos también ayudan a los médicos a descubrir si estos nuevos tratamientos son inocuos y si son mejores que los tratamientos actuales. (1)

1.2 Almacén de datos.

Los AD son temáticos, orientados a cubrir las necesidades a las organizaciones con el fin de obtener un sistema de soporte para la gestión, control y apoyo a la toma de decisiones y así obtener una mejor ventaja comercial. (2)

Un AD según Inmon “es una colección de datos orientado a un determinado ámbito (empresa, organización, etc.), integrado, no volátil y variable en el tiempo, que ayuda a la toma de decisiones en la entidad en la que se utiliza”. (2)

Ralph Kimball propone otra definición al catalogarlo como “…una copia de datos transaccionales, específicamente estructurados para la consulta y el análisis”. (2)

Basándose en el concepto de Inmon se concluye que un AD es la unión de varios MD (subconjunto del almacén) que contienen información de las principales áreas de la empresa. Obteniéndose de una o varias fuentes para su posterior análisis y que persistirá en el tiempo, permitiendo a la empresa el apoyo a la toma de decisiones.

Características de los AD.

- ✓ **Integrado:** los datos almacenados deben integrarse en una estructura consistente, por lo que deben ser eliminadas en su totalidad las inconsistencias provenientes de los sistemas operacionales.

- ✓ **Orientado al tema:** el AD está orientado por las principales áreas de temáticas de la empresa. Los datos se organizan por temas para facilitar su acceso y entendimiento por parte de los usuarios finales.

- ✓ **Histórico:** en los sistemas operacionales, los datos siempre reflejan el estado de la actividad del negocio en el presente. Por el contrario, la información en el almacén sirve, entre otras cosas, para realizar análisis de tendencias. Por tanto, el almacén se carga con los distintos valores que toma una variable en el tiempo para permitir comparaciones.

- ✓ **No volátil:** los datos almacenados no se modifican ni se actualizan, solo se añaden nuevos datos. Los datos de un almacén existen para ser leídos, y no modificados, por lo tanto se carga una sola vez y siguen igual en lo adelante. (3)

1.3 Mercado de datos.

El MD constituye una tecnología de base de datos que ha tomado gran auge debido al crecimiento y muchas veces subutilización de los grandes bancos de datos históricos almacenados en las organizaciones.

Un MD es un conjunto de datos de una base de datos especializada, departamental, orientada a satisfacer las necesidades específicas de un grupo de usuarios. Normalmente es un subconjunto de un AD con transformaciones específicas para el área al que está dirigido. (4)

Se caracteriza por disponer de una estructura de datos que analiza la información al detalle desde las perspectivas que afectan los procesos del departamento. Este puede obtener los datos desde un almacén o puede integrar un compendio de distintas fuentes de información. (7)

1.3.1 Características de los MD.

Teniendo en cuenta que los MD son una parte de los AD, estos comparten las mismas características de ser integrado, orientado al tema, no volátil e histórico, además incluye otras características en las que se pueden citar:

- ✓ Según las necesidades de los usuarios el diseño del MD se realiza siguiendo una estructura consistente.
- ✓ Contiene el grado de granularidad necesaria.
- ✓ La cantidad de información que contienen es mucho menor que en los AD.
- ✓ Debido a que hay un grupo de usuarios que solo acceden a un subconjunto preciso de datos, se tiene mejor acceso a las herramientas de consulta y dividiendo los datos se tiene un control de estos. (2)

1.3.2 Ventajas y desventajas de los MD.

Los MD presentan una gran variedad de ventajas, de las que se pueden citar:

- ✓ Poco volumen de datos, mayor rapidez de consulta.
- ✓ Son simples de implementar.
- ✓ Conllevan poco tiempo de construcción y puesta en marcha.
- ✓ Permiten manejar información confidencial, de un área específica de una empresa.
- ✓ Reflejan rápidamente sus beneficios y cualidades.

Como desventaja presenta que al crecer el MD, el rendimiento de las consultas decae y deja de ser óptimo. Por lo que no permite el manejo de grandes volúmenes de datos.

1.3.3 Subsistemas de los MD.

Los MD están compuestos por tres subsistemas:

Subsistema de almacenamiento: en este subsistema se realiza un estudio del negocio y se elaboran los requisitos de información, cumpliéndose así las necesidades del cliente. Se crea el modelo de datos que

contiene las tablas de hechos y dimensiones, además de las relaciones que existen entre ellas, almacenándose la información en dichas tablas.

Subsistema de integración: se encargará de limpiar los datos una vez creado el modelo de datos, se realiza la extracción de los datos, se detectan los incorrectos y las entradas duplicadas. Luego se aplican una serie de transformaciones que combinan y ordenan los datos para estandarizarlos, integrarlos y por último cargarlos en la base de datos.

Subsistema de visualización: este subsistema tiene como objetivo principal presentar la información al cliente y organizar los reportes por áreas de análisis, facilitando al usuario una búsqueda rápida de la información.

Para dar paso a la solución se ha determinado por las características de la presente investigación, desarrollar los subsistemas de almacenamiento e integración, debido a que el cliente aplicará técnicas de minería de datos, que permiten descubrir patrones de comportamiento en los datos.

1.4 Tendencias actuales.

Los sistemas de AD constituyen un recurso corporativo primario y parte importante del patrimonio de la institución, manejando gran cantidad de datos en forma centralizada y manteniendo subsistemas en línea.

Ejemplos de la utilización de estos sistemas tanto en Cuba como en el mundo son:

- ✓ Almacén de datos de Meditech: brinda a la institución de salud un entorno abierto y propicio para la creación de informes robustos y para el apoyo a la toma de decisiones. (5)

- ✓ Procedimiento para el desarrollo de un sistema de inteligencia de negocios en la gestión de ensayos clínicos en el Centro de Inmunología Molecular: el objetivo fue desarrollar un procedimiento que contribuyera al almacenamiento y análisis de los EC y que facilitara la aplicación integral de la inteligencia de negocios en esta actividad. (6)

El objetivo de los AD expuestos es resolver una problemática similar a la presentada en la investigación, pero ninguno responde a las necesidades y características de la misma. Por lo que se decide desarrollar los subsistemas de almacenamiento e integración del producto hr3.

1.5 Metodologías para el desarrollo de un AD.

Con el fin de mejorar la calidad del *software* han surgido diferentes metodologías de desarrollo, las cuales definen un conjunto de pasos y procesos a seguir, que permiten planificar, estructurar y controlar el proceso de desarrollo de *software.*

Existen dos grandes enfoques arquitectónicos para enfrentar la construcción de un AD: el enfoque propuesto por Bill Inmon (descendente o *Top-down*) y por Ralph Kimball (ascendente o *Bottom-up*).

El enfoque *Botton-up* se centra en el detalle y construye la solución desde lo específico a lo genérico (crea los MD y luego el AD) y en cambio *Top-down*, parte de construir la solución desde lo genérico para a partir de ahí propagar la solución detallada (crea el AD y después los MD). Este último logra reducir el tiempo de creación de la solución.

En diferentes perspectivas de análisis, una metodología puede ofrecer ventajas con respecto a otras, e incluso para tener mejores soluciones se pueden realizar algunas combinaciones de las mismas.

Algunos ejemplos de metodologías híbridas son:

Metodología Hefesto: es una metodología ágil en desarrollo que propone cómo guiar la construcción de los AD. Se basa en los requerimientos del usuario, por lo cual su estructura es capaz de adaptarse con facilidad y rapidez ante los cambios en el negocio. En cada fase el usuario final toma decisiones respecto al comportamiento y funciones de los AD, distinguiendo fácilmente los objetivos y resultados esperados. Estos resultados se convierten en el punto para llevar a cabo el paso siguiente. La metodología utiliza modelos conceptuales y lógicos. (7)

Metodología para el diseño conceptual de almacenes de datos: es presentada en la tesis de doctorado de Leopoldo Zenaido Zepeda Sánchez, la misma define un conjunto de transformaciones que son utilizadas para llevar un diagrama relacional a uno dimensional para así obtener los posibles esquemas multidimensionales candidatos para el AD. (8)

Metodología de Desarrollo para Proyectos de Almacenes de Datos: esta metodología se centra el enfoque ascendente del Ciclo de vida de Kimball e incorpora los casos de uso para guiar el proceso de

desarrollo según lo planteado por el doctor Leopoldo Zenaido Zepeda en su tesis de doctorado; y así lograr estar alineados a las tendencias y normas de la UCI. Se agrega, además, una etapa de prueba para así comprobar la calidad de los productos que se desarrollen. (9)

1.5.1 Metodología utilizada.

Para guiar el proceso de desarrollo de los subsistemas de almacenamiento e integración del producto hr3, se escoge la *Metodología de Desarrollo para Proyectos de Almacenes de Datos*, que es una adaptación del Ciclo de vida de la metodología de Kimball por los siguientes elementos: (9)

1. Identifica las tablas de hechos y dimensiones, lo cual agiliza el proceso de desarrollo y con ello la toma de decisiones.
2. Es una metodología madura y reconocida por los usuarios dedicados al tema, además de tener bien definidas sus etapas, actividades, roles y artefactos.
3. Propone la construcción de mercados de datos departamentales y después el AD. Esto trae como ventaja que la creación y la puesta en marcha de los mercados de datos se producen en un lapso de tiempo corto y después se valora si se construye o no el AD.
4. Existe amplia documentación de la misma, así cualquier duda que exista puede ser atendida rápidamente.

Esta metodología cuenta con un flujo de trabajo y siete fases, las cuales se describen a continuación: (9)

Gestión del proyecto: constituye un flujo de trabajo que se ejecuta a lo largo de todo el ciclo de vida del proyecto. Está compuesto por un grupo de procesos que se encargan de mantener la adecuada gestión del proyecto a partir de la aplicación de conocimientos, habilidades, herramientas y técnicas.

1. **Estudio preliminar y planeación:** la fase se compone por dos procesos: el estudio preliminar y la planeación inicial del proyecto. El estudio preliminar consiste en hacer un diagnóstico integral de la organización dividido en tres áreas: diagnóstico del negocio, diagnóstico de los datos y diagnóstico de la infraestructura tecnológica. Con los resultados del diagnóstico se hace un estudio de factibilidad que permita estimar los costos de desarrollo, con el fin de establecer el monto del presupuesto que se necesita para desarrollar el proyecto. Durante esta fase también se realizan

las tareas de planeación inicial del proyecto, para ello se definen un grupo de aspectos importantes relacionados con la gestión de proyectos como son: alcance del proyecto, riesgos, calidad del producto, recursos humanos, adquisiciones, cronograma, entregables, costos y presupuesto.

2. **Requisitos:** se realiza en dos direcciones, una, identificando las necesidades de información y reglas del negocio; y la otra con un levantamiento detallado de las fuentes de datos a integrar. Es aquí donde se definen los requisitos a través de la comparación de las necesidades y las reglas del negocio.

3. **Arquitectura:** aquí se definen las estructuras de almacenamiento, se diseñan las reglas de extracción, transformación y carga, así como la arquitectura de información que regirá el desarrollo de la solución.

4. **Diseño e implementación:** en esta etapa se obtiene el producto de *software*, se diseñan e implementan los tres subsistemas que conforman el AD. Cada subsistema puede verse como un componente de *software* que se desarrolla de forma independiente, para luego ser integrados conformando el producto final.

5. **Prueba:** en esta fase se realizan las pruebas necesarias para validar la calidad del *software,* una vez implementado el mismo. Aquí se realizan las pruebas de unidad, las pruebas de integración y pruebas al sistema, hasta las pruebas de aceptación con el cliente final.

6. **Despliegue:** primeramente se realiza un despliegue piloto en el cual se configuran los servidores, se instalan las herramientas según la arquitectura definida y se carga una muestra de los datos para demostrar al cliente que el sistema funciona. Posterior a la aceptación del cliente se realiza la carga histórica de los datos, la capacitación y por último, la transferencia tecnológica.

7. **Soporte y mantenimiento:** después de haber implantado la solución, se brindan los servicios de soporte en línea, vía telefónica, web u otros, hasta el acompañamiento junto al cliente según el contrato firmado y las condiciones de soporte establecidas.

De las siete etapas con que cuenta esta metodología, se aplicarán las cinco primeras (Estudio preliminar o planeación, Requisitos, Arquitectura, Diseño e implementación y Prueba). Las dos fases restantes

(Despliegue, Soporte y mantenimiento) no se desarrollarán porque no se encuentran dentro del alcance de la investigación, debido a que estas son aplicadas por el Departamento de AD.

1.6 Procesos de integración.

El término integración de datos es usualmente entendido como el proceso que combina datos de diferentes fuentes para proveer una visión simple y comprensible de toda la información combinada. Existen varias técnicas de integración de datos como: Integración de Aplicaciones Empresariales (EAI), Integración de Informaciones empresariales (EII), Extracción, Transformación y Carga (ETL), entre otras. (10)

Por las características de la presente investigación se decide utilizar la técnica de ETL. Como su nombre lo indica, extrae información de un sistema fuente, transforma esos datos para satisfacer los requisitos del negocio y carga el resultado en el sistema destino. (11)

Procesos de ETL

Los procesos de ETL se encargan de extraer los datos desde las diversas fuentes, en este caso desde ficheros con formato *xls*, los transforman para resolver posibles problemas de inconsistencias entre los mismos y finalmente se procede a su carga en la base de datos destino.

Extracción: es el primer paso para la obtención de los datos de las diferentes fuentes hacia el AD. Como los datos pueden provenir de distintas fuentes, generalmente se encuentran en formatos distintos y organizados de acuerdo a los procesos de cada organización. Este proceso extrae los datos de las fuentes para luego aplicarles las transformaciones necesarias.

Transformación: una vez que los datos están extraídos, el proceso de transformación se encarga de preparar los datos de la manera adecuada para integrarlos en el AD. Para ello este proceso se compone de algunas actividades como: limpieza de datos, estandarización de formato e integración de datos.

Carga: una vez que los datos han sido extraídos de las diferentes fuentes y transformados, se realiza la carga hacia el AD y luego de la carga inicial se procede a mantener el almacén actualizado periódicamente en el caso que lo necesiten. (12)

1.7 Herramientas

Durante el desarrollo de los subsistemas de almacenamiento e integración del producto hr3 se utilizan varias herramientas que facilitan el trabajo en las distintas fases que abarca el proceso. A continuación se exponen las herramientas definidas por el Departamento de AD de DATEC para la implementación de la soluciones de AD.

1.7.1 Herramienta de modelado.

Las herramientas de modelado se utilizan para representar los elementos claves del proceso de manera que sea posible alcanzar una mejor comprensión del mismo. Las herramientas de Ingeniería Asistida por Computadora (*Computer Aided Software Engineering,* CASE por sus siglas en inglés), tienen como objetivo incrementar la productividad y calidad de los productos de *software*, mejorar la planificación del proyecto, así como reducir el tiempo y costo de su desarrollo.

Se puede definir a las herramientas CASE como un conjunto de herramientas y métodos asociados que proporcionan asistencia automatizada en el proceso de desarrollo del *software* a lo largo de su ciclo de vida. Fueron desarrolladas para automatizar esos procesos y facilitar las tareas de coordinación de los eventos que necesitan ser mejorados en el ciclo de desarrollo de *software*. (13) Existen varias herramientas CASE, ejemplo de estas es el Visual Paradigm.

Visual paradigm (10.1)

Es una herramienta para desarrollo de aplicaciones utilizando Lenguaje Unificado de Modelado (*Unified Modeling Language*, UML por sus siglas en inglés), útil para ingenieros de *software*, analistas y arquitectos de sistemas que están interesados en la construcción de estos, a gran escala y necesitan confiabilidad y estabilidad en el desarrollo orientado a objetos. Es fácil de instalar y actualizar, además de que puede ser utilizado durante todo el ciclo de vida del desarrollo de *software*. Esta herramienta acelera el desarrollo de aplicaciones, ya que sirve de puente visual entre arquitectos, analistas y diseñadores de sistemas de información, haciendo el trabajo fácil y dinámico. (14)

Se decidió utilizar Visual Paradigm (10.1) para UML porque es multiplataforma y permite su uso en cualquier sistema operativo. Además, soporta todo el ciclo de vida del desarrollo de *software*. Es factible a la hora de dibujar diagramas de clases, generar *script* para diferentes Sistemas Gestores de Bases de Datos (SGBD), permite una integración con sistemas de control de versiones que almacenan centralmente

los artefactos y realizan un seguimiento de los cambios realizados sobre un proyecto. Los desarrolladores los utilizan para facilitar el modelado simultáneo, almacenar los archivos de proyectos y hacer un seguimiento de los cambios. Además, permite exportar imágenes en formato *jpg, png y svg (w3g estándar)*.

1.7.2 Sistema Gestor de Base de Datos (SGBD).

Un SGBD es uno o varios *software* cuyo objetivo es servir de interfaz entre la base de datos, los usuarios y las aplicaciones. Estos sistemas permiten definir los datos a distintos niveles de abstracción y manipular los mismos, garantizando la seguridad e integridad de estos. (15) Es una herramienta efectiva que permite a varios usuarios acceder a la información al mismo tiempo. Dentro de los SGBD se pueden encontrar Oracle, Sybase, PostgreSQL, MySQL, entre otros.

Un SGBD debe permitir:

- ✓ Definir una base de datos: especificar tipos, estructuras y restricciones de datos.
- ✓ Construir una base de datos: guardar los datos en algún medio controlado por el mismo SGBD.
- ✓ Manipular la base de datos: realizar consultas, actualizar la base de datos, generar informes.

PostgreSQL (9.1)

El departamento propone utilizar como SGBD PostgreSQL por ser lo suficientemente estable y seguro, se utilizará con la versión 9.1. PostgreSQL funciona bien con grandes cantidades de datos y una alta concurrencia de usuarios accediendo a la vez al sistema. El código fuente está disponible bajo los más liberales términos de licencia de código abierto: la licencia *Berkeley Software Distribution* (BSD), por tanto pueden hacerse todas las modificaciones, mejoras o cambios que se estimen convenientes. (16)

Se caracteriza por:

- ✓ Multiplataforma: corre en diferentes sistemas operativos, incluyendo GNU/Linux, UNIX (AIX, BSD, HP-UX, SGI IRIX, Mac OS X, Solaris, Tru64) y Windows.
- ✓ Altamente escalable tanto en la cantidad de datos que puede administrar como en el número de usuarios concurrentes que puede manejar.

- ✓ Soporta llaves foráneas, tipos de datos definidos por el usuario, secuencias, relaciones, uniones, vistas, reglas, *triggers* y procedimientos almacenados en múltiples lenguajes.
- ✓ Puntos de recuperación a un momento dado, *table spaces*, replicación asincrónica, transacciones jerárquicas (*save points*), copia de seguridad en línea.
- ✓ Usa una arquitectura Cliente/Servidor. (16)

1.7.3 Administrador de base de datos.

Un administrador de base de datos (*Data Base Administrator*, DBA por sus siglas en inglés) es fundamental para el desarrollo de una investigación de este tipo, ya que son los responsables de la integridad y disponibilidad de los datos. El DBA se encarga también de garantizar el funcionamiento adecuado del sistema y de proporcionar otros servicios de índole técnica relacionados.

PgAdmin III (1.14.0.)

PgAdmin es una plataforma de desarrollo y administración de código abierto para la administración de bases de datos PostgreSQL. Es multiplataforma, puede ser usada en Linux, FreeBSD, Solaris, Mac OSX y Windows para gestionar PostgreSQL, así como las versiones derivadas y comerciales como Postgres Plus Advanced Server y Greenplumdatabase. (17)

La aplicación es desarrollada por una comunidad de expertos de PostgreSQL de todo el mundo, por lo que es *software* libre lanzado bajo la licencia *Berkeley Software Distribution* (BSD) de PostgreSQL y está disponible en más de una docena de idiomas. (17)

El departamento propone utilizar PgAdmin como la herramienta de administración de bases de datos, ya que está diseñado para responder a las necesidades de todos los usuarios, desde simples consultas SQL (por sus siglas en inglés *Structured Query Language*) hasta el desarrollo de complejas bases de datos. La interfaz gráfica que posee es compatible con todas las características de PostgreSQL y facilita su administración. Entre algunas de las funcionalidades que incluye la aplicación se encuentra un editor de sobresaltado de sintaxis SQL (*syntax highlighting SQL editor*) y un editor de código seguro del lado del servidor (*server*-sidecode editor). Puede realizar la conexión mediante el Protocolo de Control de Transmisión/Protocolo de Internet (TCP/IP) o Unix Domain Sockets (variante de los sockets que tienen como propósito la intercomunicación entre programas dentro de la misma computadora), además, puede

ser encriptado mediante la capa de conexión segura (*Secure Sockets Layer,* SSL por sus siglas en inglés) para mayor seguridad.

1.7.4 Herramientas para los procesos de ETL.

ETL es el proceso que organiza el flujo de datos entre diferentes sistemas de una organización y aporta los métodos y herramientas necesarias para mover datos desde múltiples fuentes a un AD, reformatearlos, limpiarlos y cargarlos en otra base de datos, MD o AD.

Data Cleaner (1.5.3)

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. Data Cleaner es una aplicación de código abierto para el perfilado, la validación y comparación de los mismos. 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 su situación de negocio.

El uso del Data Cleaner permite la evaluación del nivel de calidad de los datos contenidos en el sistema de información. 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. (18)

Pentaho Data Integration (4.2.1)

Pentaho Data Integration (PDI), también conocido como Kettle, es una herramienta de código abierto que se utiliza para implementar los procesos de ETL. Reúne un conjunto de componentes que permiten modelar y ejecutar transformaciones sobre flujos de datos. Posee una arquitectura altamente escalable y un entorno de diseño gráfico intuitivo y rico, proporciona la solución para cualquier tipo de integración de datos, análisis de negocio o proyectos con grandes capacidades de datos.

PDI está integrado por cuatro componentes:

- ✓ SPOON: para el diseño gráfico de las transformaciones.

- ✓ PAN: para la ejecución de los trabajos y las transformaciones.
- ✓ CHEF: para el diseño y la carga de datos.
- ✓ KITCHEN: para la ejecución de los trabajos *batch* diseñados con CHEF. (19)

Entre las características que Pentaho Data Integration incluye y por las cuales el departamento lo propone se encuentran:

- ✓ Escalabilidad y rendimiento empresarial, incluyendo el almacenamiento en caché en la memoria.
- ✓ Integración de datos, análisis y presentación de informes, incluyendo, NoSQL, OLTP tradicional y bases de datos analíticas.
- ✓ Interfaz de diseño ETL fácil de usar: el diseñador gráfico intuitivo que presenta PDI permite hacer exactamente lo que expertos desarrolladores pueden lograr, en solo una fracción del tiempo que emplearían estos y sin utilizar código manual.
- ✓ Perfilado y calidad de datos: con los pasos de transformación de los principales proveedores de calidad de datos integrados en el diseñador gráfico de ETL, PDI ofrece una plataforma de datos de mejor calidad que cualquier otro proveedor de inteligencia de negocios. (20)

1.8 Modos de almacenamiento de datos.

Procesamiento Analítico en Línea (On-Line Analytical Processing OLAP).

Es una solución utilizada en el campo de la inteligencia empresarial (*Business Intelligence*) cuyo objetivo es agilizar la consulta de grandes cantidades de datos. Para ello utiliza estructuras multidimensionales (o cubos OLAP) que contienen datos resumidos de bases de datos o sistemas transaccionales. Se usa en informes de negocios de ventas, *marketing*, informes de dirección, minería de datos y áreas similares. (21)

Con OLAP se puede ver un conjunto de datos de su negocio de diversas formas sin mucho esfuerzo. Los archivos OLAP o cubos modelan los datos en dimensiones. (21)

Características:

- ✓ Permite recolectar y organizar la información analítica necesaria para los usuarios y disponer de ella en diversos formatos, tales como: tablas, gráficos, reportes, tableros de control, entre otros.
- ✓ Soporta análisis complejos de grandes volúmenes de datos.

- ✓ Presenta al usuario una visión multidimensional de los datos para cada tema de interés del negocio. (22)

Procesamiento Analítico en Línea Relacional (ROLAP).

La arquitectura ROLAP accede a los datos almacenados en un AD para proporcionar los análisis OLAP. La premisa de los sistemas es que las capacidades OLAP se soportan mejor contra las bases de datos relacionales. La base de datos relacional maneja los requisitos de almacenamiento de datos, y el motor ROLAP proporciona la funcionalidad analítica. El nivel de aplicación es el motor que ejecuta las consultas multidimensionales de los usuarios. Este motor se integra con niveles de presentación, a través de los cuales los usuarios realizan los análisis OLAP. Una vez que el modelo de datos para el AD se ha definido, la información se carga desde el sistema operacional. Se ejecutan rutinas de base de datos para agregar el dato, si así es requerido por el modelo de datos. (22)

Procesamiento Analítico en Línea Multidimensional (MOLAP).

La arquitectura MOLAP usa una base de datos propietaria multidimensional, donde la información se almacena multidimensionalmente para ser visualizada en varias dimensiones de análisis. Utiliza una arquitectura de dos niveles: las bases de datos multidimensionales y el motor analítico. La base de datos multidimensional es la encargada del manejo, acceso y obtención del dato. (22) MOLAP consigue consultas rápidas a costa de mayores necesidades de almacenamiento.

Procesamiento Analítico en Línea Híbrido (HOLAP).

Un desarrollo un poco más reciente ha sido la solución HOLAP, la cual combina las arquitecturas ROLAP y MOLAP para brindar una solución con las mejores características de ambas: desempeño superior y gran escalabilidad. HOLAP mantiene los registros de detalle (los volúmenes más grandes) en la base de datos relacional, mientras que mantiene las agregaciones en un almacén MOLAP separado. (22)

En la presente investigación se selecciona como 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.

Modelo multidimensional.

En un modelo de datos multidimensional la información se organiza alrededor de los temas de la organización. La estructura de datos manejada en este modelo es la matriz multidimensional o hipercubo. Se puede decir que un hipercubo es un conjunto de celdas, cada una se identifica por la combinación de los miembros de las diferentes dimensiones y contiene el valor de la medida analizada para dicha combinación de dimensiones. Una de sus características principales es que no necesita que los reportes se hayan predefinido, debido a que se diseñan de forma tal que cubra el universo de variantes que los usuarios necesiten consultar en la información almacenada. (23)

El modelo multidimensional incluye tres variantes de modelación, que están determinadas por la complejidad del sistema:

- ✓ Esquema en estrella: es un tipo de esquema de base de datos relacional que consta de una sola tabla de hechos central rodeada de tablas de dimensiones.
- ✓ Esquema copo de nieve: el esquema de copo de nieve consta de una tabla de hechos que está conectada a varias tablas de dimensiones, que pueden estar conectadas a otras tablas de dimensiones a través de una relación de muchos a uno.
- ✓ Constelación de hechos: es una combinación de un esquema de estrella y un esquema de copo de nieve. Donde varias tablas de hechos pueden compartir algunas dimensiones.(19)

Para una adecuada comprensión de los distintos tipos de modelado existentes es necesario dominar algunos conceptos básicos referentes al tema, que a continuación se exponen.

El hecho es el objeto a analizar, posee atributos llamados de hechos o de síntesis y son de tipo cuantitativo. Sus valores (medidas) se obtienen generalmente por la aplicación de una función estadística que resume un conjunto de valores en un único valor. (23)

Las dimensiones representan cada uno de los ejes en un espacio multidimensional. Suministran el contexto en el que se obtienen las medidas de un hecho. Las dimensiones se utilizan para seleccionar y agrupar los datos en un nivel de detalle deseado. Los componentes de una dimensión se denominan niveles y se organizan en jerarquías. (23)

La jerarquía se define como el orden determinado dentro de los campos de una dimensión. Estas jerarquías son utilizadas a la hora de agregar o desagregar la información. Las tablas de hechos y dimensiones contienen datos de interés, que presentan un nivel de granularidad. La granularidad es el

nivel más bajo de información que será almacenado en estas tablas. El primer paso al diseñar una tabla de dimensión es determinar la granularidad. (23)

Conclusiones del capítulo.

Después de analizado el estado del arte de los AD y los MD con sus características, ventajas y desventajas, se caracterizaron la metodología y herramientas a utilizar en el desarrollo de la solución. Lo que permitió arrojar las siguientes conclusiones:

- ✓ La creación del marco teórico apoyó el estudio de los diferente conceptos, características, ventajas y desventajas de los AD y MD.
- ✓ La metodología seleccionada guiará el proceso de desarrollo de los subsistemas de almacenamiento e integración del producto hr3, permitiendo transitar por todas sus fases.
- ✓ Las herramientas seleccionadas posibilitarán un correcto diseño e implementación de los subsistemas de almacenamiento e integración del producto hr3.
- ✓ La técnica de integración seleccionada permitirá realizar una correcta implementación de los subsistemas de almacenamiento e integración del producto hr3.

Capítulo 2: Análisis y diseño de los subsistemas de almacenamiento e integración del producto hr3.

Introducción.

En este capítulo se realiza el análisis y diseño del negocio. Se abordan aspectos referentes al levantamiento de requisitos (información, funcionales, no funcionales), a la definición de las RN y al modelado de los datos con sus respectivos elementos tales como las dimensiones, los hechos y las medidas. Se define la arquitectura y se diseñan los subsistemas de almacenamiento e integración del producto hr3. Se representa en la matriz bus las relaciones existentes entre los hechos y dimensiones. Por último se crea el esquema de seguridad y se describe la política de respaldo y recuperación, definiendo los roles y permisos.

2.1 Análisis del negocio.

Para el desarrollo de un sistema el levantamiento de los requisitos es una actividad necesaria, ya que constituyen especificaciones que los especialistas precisan para darle cumplimiento a sus tareas internas. Las necesidades del negocio están en correspondencia con lo que el cliente necesita, por lo que es muy importante realizar un análisis previo, para que los especialistas del CIM queden satisfechos con el producto.

Para identificar las necesidades de la organización existen varias técnicas como la entrevista, los cuestionarios y observaciones. En la presente investigación se utilizó la entrevista, ya que posibilitó interactuar directamente con el especialista en desarrollo de los subsistemas de almacenamiento e integración de los EC, determinando de esta manera los requisitos del cliente. A partir de este estudio se identificaron varios problemas con el análisis de los datos relacionados con los EC realizados a los pacientes que presentan cáncer y de estos los que se han tratado con el ACM hr3.

Actualmente en el CIM, la información que han generado los EC del producto hr3 es almacenada en ficheros de Excel, exportados mediante el sistema EpiData. A medida que se realizan dichos estudios, se suministra a los pacientes el medicamento y comprueban mediante exámenes de laboratorio y físicos, las reacciones ante y durante el tratamiento. Se registran las respuestas, las lesiones y los eventos adversos presentados por los pacientes durante el proceso.

Como parte de las necesidades de información se decidió integrar toda la información contenida de los diferentes EC formando estos los temas de análisis, los que se encuentran agrupados de la siguiente manera:

- ✓ EC079 hr3 Meta Cerebral FII: EC realizados a pacientes con cáncer de pulmón de células no pequeñas (NSCLC) portadores de metástasis cerebral.
- ✓ EC070 hr3 Mama FI: EC realizados a pacientes con cáncer de mama.
- ✓ EC040 hr3 C y C: EC realizados a pacientes con tumores epiteliales de cabeza y cuello en estadios avanzados.
- ✓ EC035 hr3 C y C Farmacodinamia: EC realizados a pacientes con cáncer avanzado de cabeza y cuello.

2.2 Especificación de requisitos.

El análisis de requisitos constituye una de las fases más importantes en el desarrollo de aplicaciones informáticas. Pudiéndose identificar las necesidades de información de la empresa, las características y cualidades que debe poseer el sistema. En dicha fase se definen los requisitos de información, funcionales y no funcionales, partiendo de las necesidades de los clientes.

2.2.1 Requisitos de información (RI).

Los requisitos de información describen la información y los datos que el cliente necesita almacenar en el sistema. Se definen a partir de las necesidades del negocio, pues permiten el análisis del producto con el objetivo de almacenar la información de los EC realizados en el CIM.

Los requisitos de información identificados durante el proceso de análisis fueron clasificados según los temas de análisis definidos. A continuación se presentan algunos de ellos. Para más detalles diríjase a los artefactos “DATEC_CIM_hr3-0113_ERS” y “HerramientaRecolAna” dentro del Expediente de Proyecto de los subsistemas de almacenamiento e integración del producto hr3.

RI1: Almacenar la información de la cantidad de pacientes que tuvieron una evaluación inicial NSCLC en el EC079 hR3 Meta Cerebral FII por examen físico, peso, karnofsky, examen clínico, método diagnóstico y localización de la lesión.

RI2: Almacenar la información de la suma de diámetros que tuvieron una evaluación inicial NSCLC en el EC079 hR3 Meta Cerebral FII por examen físico, peso, karnofsky, examen clínico, método diagnóstico y localización de la lesión.

RI3: Almacenar la información de la cantidad de pacientes que le administraron el producto hr3, del EC070 hr3 Mama FI por ecog, examen físico, peso, dosis administrada de hr3, tiempo, tensión mínima, tensión máxima, frecuencia cardíaca, temperatura y si asistió a la administración.

2.2.2 Requisitos funcionales (RF).

Para lograr satisfacer las necesidades del cliente definidas con anterioridad, se hace necesario identificar los requisitos funcionales, los cuales expresan las condiciones o capacidades que el sistema debe cumplir, para que cumpla con las especificaciones del cliente y sea exitoso. Los requisitos funcionales identificados fueron agrupados por subsistemas.

Atendiendo a que los EC del producto hr3 son cerrados, no se especifica la persistencia de la información ni las vistas integradas, por tanto no hay requisitos relacionados con el subsistema de almacenamiento. Los requisitos asociados al subsistema de integración de datos se describen a continuación:

Subsistema de integración:

RF1: Realizar la extracción de los datos.

Entrada: Poseer los datos de la fuente.

Salida: Los datos estén extraídos de la fuente y listos para ser transformados.

Descripción: Se extraen los datos de la fuente y se tienen listos para ser transformados.

RF2: Realizar la transformación y carga de los datos.

Entrada: Los datos deben estar extraídos de la fuente.

Salida: Los datos transformados y cargados en la base de datos.

Descripción: Después de tener los datos extraídos de la fuente, serán transformados y cargados en la base de datos.

2.2.3 Requisitos No Funcionales (RNF).

Los requisitos no funcionales determinan cómo debe estar el sistema o la aplicación una vez terminada. Describen las propiedades y cualidades que debe tener la solución, es decir, representan las características del producto. Los requisitos no funcionales detectados se agruparon por categorías según las características del negocio. A continuación se describen algunos RNF.

Restricciones de diseño:

RNF1: Para el desarrollo de los subsistemas de almacenamiento e integración del producto hr3 se utilizarán diferentes herramientas, como SGBD PostgreSQL 9.1 con su interfaz de administración PgAdmin III 1.14.0, para el proceso de integración de datos se usará Pentaho Data Integration en su versión 4.2.1, para el modelado Visual Paradigm 10.1 y para la limpieza de los datos Data Cleaner en la versión 1.5.3.

Confiabilidad:

RNF2: Para garantizar la persistencia de la información se crearán copias de respaldo que incluirán todas las transformaciones, trabajos y *backup* de los subsistemas de almacenamiento e integración del producto hr3, las mismas se harán cada vez que se realice algún cambio en la implementación.

Soporte:

RNF3: La estación de trabajo donde estarán montados los subsistemas de almacenamiento e integración del producto hr3, debe contar con características de *hardware* como:

- ✓ 1 GB RAM o superior.
- ✓ 1 Microprocesador Core2Duo.
- ✓ Pentium 4.
- ✓ Almacenamiento en disco al menos 60GB.

2.3 Reglas del Negocio.

Las Reglas del Negocio (RN) son el conjunto de normas y políticas, en este caso agrupadas por categorías, que deben cumplirse para que el sistema funcione correctamente y cumpla con su objetivo, es decir, son declaraciones que definen o restringen algún aspecto del negocio. Tomando como base el

estudio preliminar del negocio se identificaron las siguientes RN. Para más detalles de las mismas puede consultar el artefacto “DATEC_CIM_hr3_Reglas de negocio y transformación” dentro del Expediente de Proyecto de los subsistemas de almacenamiento e integración del producto hr3.

Reglas de almacenamiento.

RN1: Las fechas serán de tipo date en formato (yyyy/mm/dd).

Reglas de transformación

RN2: Los valores del estadio tomarán valor 1 para I, 2 para II, 3 para III, 4 para IV, 5 para IIIA, 6 para IIIB y 7 para no disponible.

RN3: Los valores del sexo tomarán el valor 1 para femenino y 2 para masculino.

RN4: Los campos que son fechas y aparezcan vacíos o tengan un valor fuera de los rangos establecidos para día, mes y año se reemplazan por la fecha 31/12/2031.

RN5: Para la variable peso existen los siguientes rangos: 1 - 25, 26 – 50, 51 – 75, 76 – 100, 101 –125,

126 – 150, 151 – 175, 176 – 200. (Kg)

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 (CUS). El diagrama de CUS es una representación de todos los casos de uso, los actores y cómo interactúan con el sistema en desarrollo. Los casos de uso son una representación lógica y visual de una secuencia de pasos en respuesta a un evento que inicia un actor sobre el propio sistema y los actores del sistema. Para la presente investigación se cuenta con el analista que es el encargado de analizar y consultar la información de los diferentes indicadores y el administrador de ETL es el encargado de realizar los procesos de ETL.

Para la confección del diagrama se cuenta con 20 CUS (18 Casos de Uso de Información (CUI) y 2 Casos de Uso Funcionales (CUF)). En la Fig. 1 se muestra un fragmento del diagrama de CUS, para más detalles consultar en el artefacto “DATEC_CIM_hr3-0114_ECU” dentro del Expediente de Proyecto de los subsistemas de almacenamiento e integración del producto hr3.

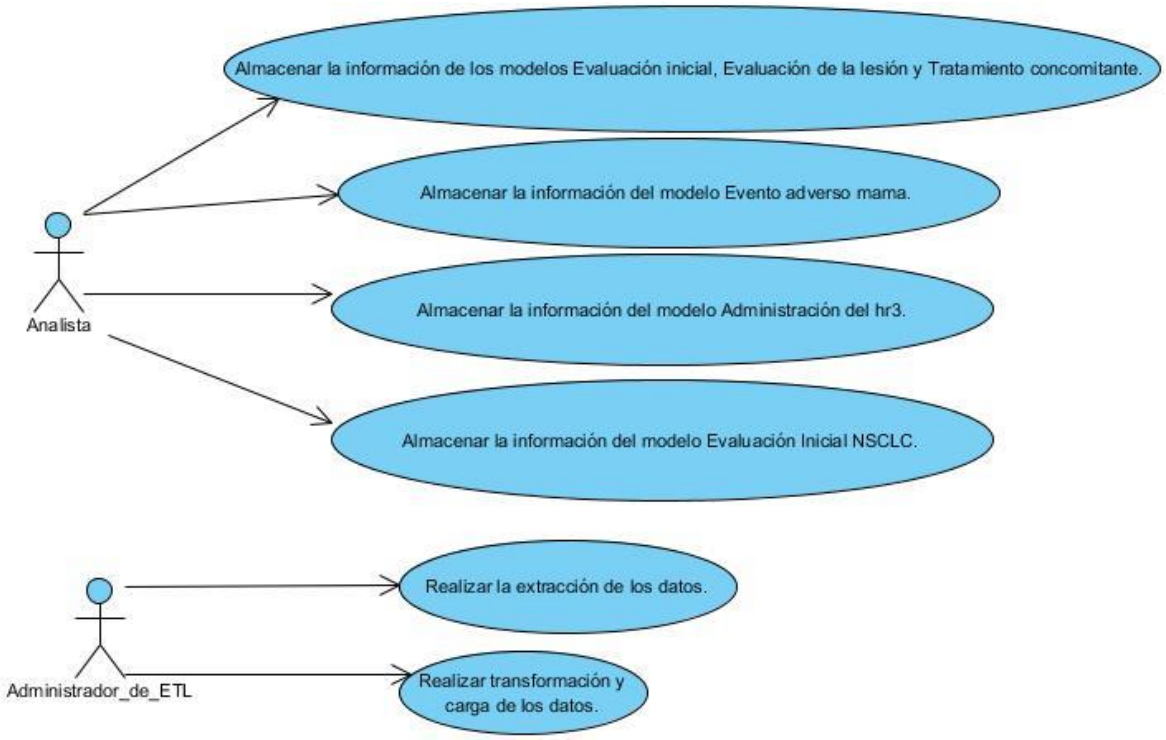


Fig. 1: Fragmento del diagrama de Casos de Uso del Sistema.

Los datos manejados partiendo de los requisitos del cliente, se agruparon por el criterio tipo de información en CUI, los cuales responden a los requisitos de información, mencionándose a continuación algunos ellos. Para más detalles consultar el artefacto “DATEC_CIM_hr3-0114_ECU” dentro del Expediente de Proyecto de los subsistemas de almacenamiento e integración del producto hr3.

CUI1: Almacenar la información de los modelos Evaluación inicial, Evaluación de la lesión y Tratamiento concomitante.

CUI2: Almacenar la información del modelo Evento adverso mama.

CUI3: Almacenar la información del modelo Administración del hr3.

CUI4: Almacenar la información del modelo Evaluación inicial NSCLC.

Por otra parte se agrupan los requisitos funcionales que se basan en la ejecución de las operaciones de ETL que serían aplicadas a las fuentes de datos relacionadas con los EC. A continuación se mencionan los CUF definidos para los subsistemas de almacenamiento e integración del producto hr3.

CUF1: Realizar la extracción de los datos.

CUF2: Realizar la transformación y carga de los datos.

2.5 Definición de la arquitectura base de los subsistemas de almacenamiento e integración del producto hr3.

La arquitectura general para el desarrollo de un MD consta de tres subsistemas (almacenamiento, integración, visualización) y la fuente de datos. Como se explicaba en el Capítulo 1, por las características de la investigación no se realizará el subsistema de visualización, quedando conformada la arquitectura para los subsistemas de almacenamiento e integración del producto hr3 con los siguientes elementos (Véase Fig. 2):

- ✓ **Fuente de datos:** está conformada por todos los ficheros en formato Excel de los EC del producto hr3.
- ✓ **Subsistema de integración:** es el encargado de extraer los datos, y llevar a cabo los procesos que integran y transforman los mismos para su almacenamiento, es decir, es donde se prepara los datos para su posterior carga.
- ✓ **Subsistema de almacenamiento:** es una base de datos soportada por el SGBD PostgreSQL y administrada por los usuarios autorizados en la herramienta PgAdminIII, que contiene las tablas de dimensiones y hechos cargadas a través de los procesos de ETL.

Fig. 2: Diseño de la arquitectura.

2.6 Diseño de la solución de la investigación.

2.6.1 Diseño del subsistema de almacenamiento.

Para realizar el diseño del subsistema de almacenamiento es de vital importancia confeccionar el modelo dimensional de los datos, el cual contiene las dimensiones y los hechos con sus medidas asociadas, siendo estos el punto de partida para el diseño. Se debe definir una política de respaldo y recuperación que garantice la integridad de los datos almacenados.

Se identificaron 42 dimensiones propias de los subsistemas de almacenamiento e integración del producto hr3. A continuación se muestra un ejemplo (Véase Fig. 3) de una de las dimensiones de la solución, para más información consultar el artefacto “DATEC_CIM_siglas mercado_hr3_Especificación del modelo de datos” dentro del Expediente de Proyecto de los subsistemas de almacenamiento e integración del producto hr3.

1. **Dimensión peso (dim_peso):** describe el peso del paciente en kilogramos (Kg).

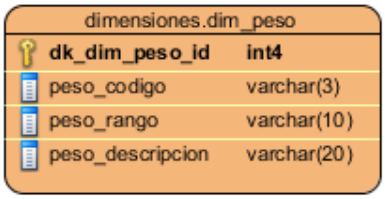


Fig. 3: Dimensión peso.

Se identificaron 10 dimensiones degeneradas. Se le denomina así al campo que será utilizado como criterio de análisis y que es almacenado en una tabla de hecho, en vez de ser definido como una dimensión. La inclusión de estos campos en las tablas de hechos se lleva a cabo con el objetivo de evitar que se dupliquen los datos y simplificar las consultas. Un ejemplo de dimensión degenerada en la investigación es *presenta metástasis*, donde la única información que se obtiene es sí o no.

La dimensión degenerada se utiliza cuando un campo posee el mismo nivel de granularidad que los datos almacenados en una tabla de hecho y no se pueden realizar agrupaciones o sumarizaciones a través del mismo.

Se identificaron 18 tablas de hechos de los subsistemas de almacenamiento e integración del producto hr3, mostrándose en la Fig. 4 un ejemplo. Para más información consultar el artefacto

“DATEC_CIM_siglas mercado_hr3_Especificación del modelo de datos” dentro del Expediente de Proyecto de los subsistemas de almacenamiento e integración del producto hr3.

1. **Administración del hr3 (hech_administracion_hr3):** muestra los resultados de los pacientes que se le administraron el producto hr3.

Fig. 4: Hecho Administración del hr3.

2.6.2 Matriz bus.

La matriz bus representa las relaciones entre los hechos y las dimensiones en un MD, además, se puede determinar el impacto que provocaría un cambio en la solución durante el desarrollo del sistema y permite verificar que no haya solapamiento de hechos, o sea, que no existan hechos que compartan exactamente las mismas dimensiones en un MD. A continuación se muestra un fragmento de la matriz bus (Véase la Fig. 5), donde la celda marcada con una x indica la relación de una columna (hecho) con una fila (dimensión). Para más información ir al documento “DATEC_CIM_siglas mercado_hr3_Especificación del modelo de datos” dentro del Expediente de Proyecto de los subsistemas de almacenamiento e integración del producto hr3.

dimensiones/hechos	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11	H12	H13	H14	H15	H16	H17	H18
dim_estadio	X									X				X				
dim_tnm	X									X				X				
dim_ecog	X		X											X				X
dim_examen_fisico	X		X	X	X					X				X	X			X
dim_examen_clinico	X			X	X								X	X				X
dim_peso			X	X	X					X								X
dim_droga																		X
dim_metodo_diagnostico	X					X								X				

Fig. 5: Fragmento de la matriz bus.

Leyenda:

H1: hech_evaluacion_inicial_evaluacion_lesion_tto_concomitante.

H2: hech_eventos_adversos_ec_mama.

H3: hech_administracion_hr3.

H4: hech_evaluacion_inicial_nsclc.

H5: hech_evaluacion_durante_tratamiento_nsclc.

H6: hech_evaluacion_de_respuesta_nsclc.

H7: hech_tratamiento_concomitante_nsclc.

H8: hech_evento_adverso_nsclc.

H9: hech_evaluacion_respuesta_clinica.

H10: hech_datos_demograficos_examen_fisico.

H11: hech_salida_del_ensayo_cyc.

H12: hech_interrupcion_del_tratamiento.

H13: hech_laboratorio_clinico.

H14: hech_evaluacion_inicial_farmacodinamia.

H15: hech_evaluacion_durante_tto_farmacodinamia.

H16: hech_evento_adverso_farmacodinamia.

H17: hech_salida_ensayo_farmacodinamia.

H18: hech_tratamiento_quimioterapia.

Al realizar la matriz bus se pudo conocer que de los 18 hechos definidos para el modelo de datos, algunos comparten dimensiones, por ejemplo, el hecho administración del hr3 y el de evaluación inicial NSCLC. Sin embargo, no existen dos o más hechos que se relacionen con exactamente las mismas dimensiones. Esto indica que pudo verificarse la inexistencia de solapamiento entre hechos.

2.6.3 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, representa una descripción de la estructura de los datos, donde se definen las relaciones que se establecen entre las dimensiones y los hechos que lo componen. En el capítulo anterior se realizó un estudio de las topologías de esquemas existentes para diseñar los modelos dimensionales, en el modelo de datos diseñado para los subsistemas de almacenamiento e integración del producto hr3 se evidencia el uso de la topología constelación de hechos, puesto que en dicho modelo existen 18 hechos y algunos comparten dimensiones. En la Fig. 6 se muestra un fragmento del modelo de datos diseñado que representa el tipo de topología seleccionada. Para más detalles ir al artefacto “DATEC_CIM_ mercado_hr3_Especificación del modelo de datos” dentro del Expediente de Proyecto de los subsistemas de almacenamiento e integración del producto hr3.

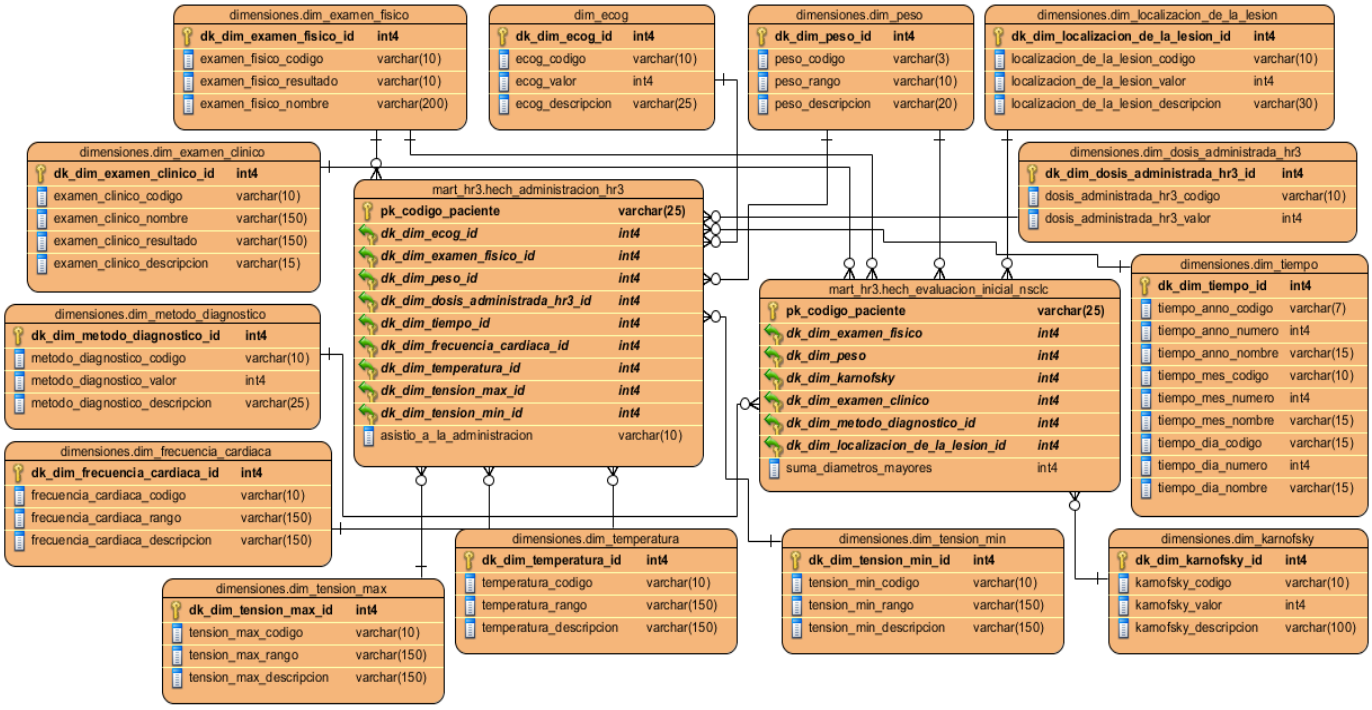


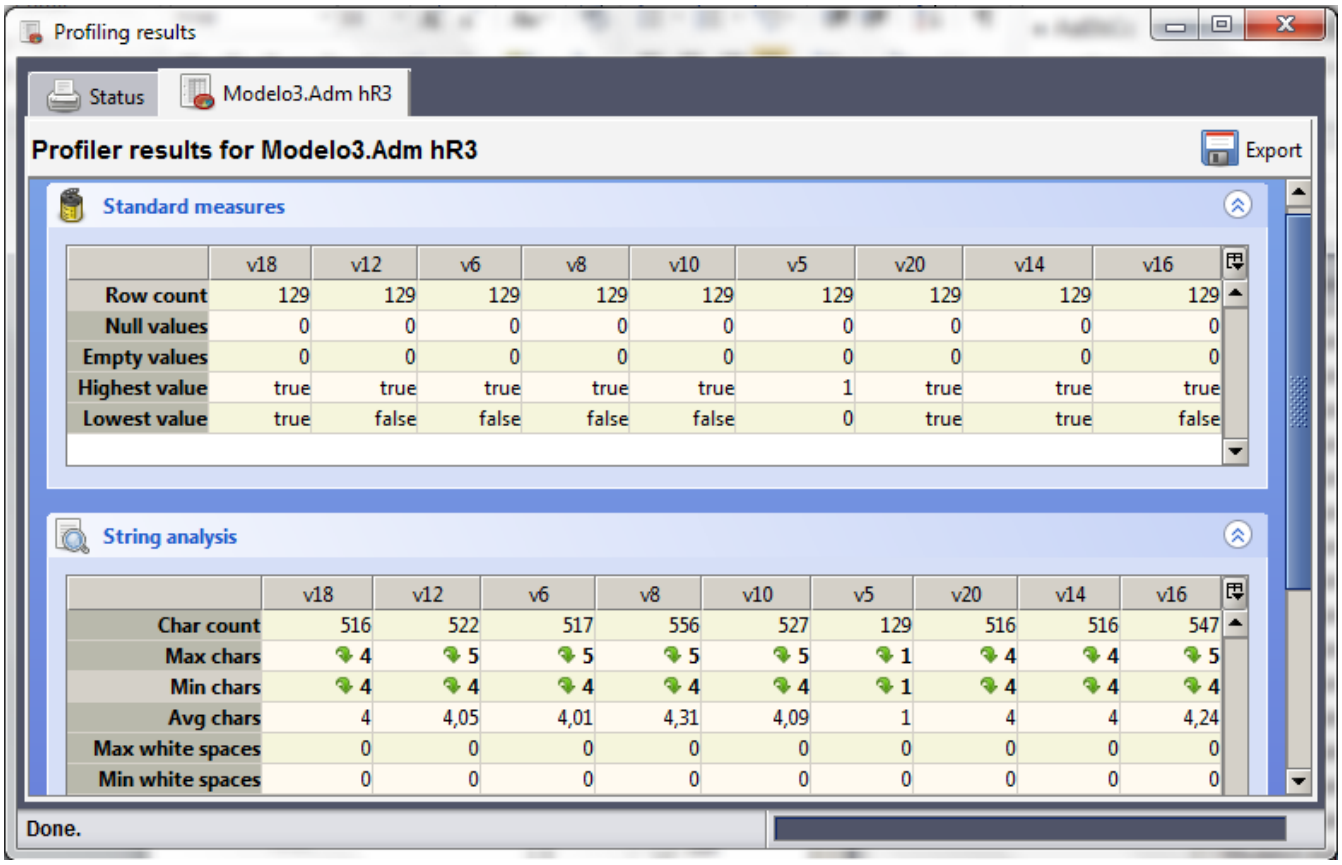
Fig. 6: Fragmento del modelo de datos de los subsistemas de almacenamiento e integración del producto hr3.

2.6.4 Diseño del subsistema de integración.

El diseño del subsistema de integración abarca el perfilado de los datos y los procesos de extracción de las fuentes de información. Dichas fuentes sufren un proceso de transformaciones con el objetivo de homogenizar los datos y finalmente estar listos para su almacenamiento.

Perfilado de datos.

La herramienta Data Cleaner se utiliza para realizar el perfilado de los datos, la cual permite identificar la cantidad y distribución de los valores nulos y desconocidos, así como la cantidad de filas de las variables. En la Fig. 7 se muestra un ejemplo de su utilización.



	v18	v12	v6	v8	v10	v5	v20	v14	v16
Row count	129	129	129	129	129	129	129	129	129
Null values	0	0	0	0	0	0	0	0	0
Empty values	0	0	0	0	0	0	0	0	0
Highest value	true	true	true	true	true	1	true	true	true
Lowest value	true	false	false	false	false	0	true	true	false

	v18	v12	v6	v8	v10	v5	v20	v14	v16
Char count	516	522	517	556	527	129	516	516	547
Max chars	4	5	5	5	5	1	4	4	5
Min chars	4	4	4	4	4	1	4	4	4
Avg chars	4	4,05	4,01	4,31	4,09	1	4	4	4,24
Max white spaces	0	0	0	0	0	0	0	0	0
Min white spaces	0	0	0	0	0	0	0	0	0

Fig. 7: Ejemplo del uso de la herramienta Data Cleaner.

El perfilado de los datos permite lograr una mejor comprensión de los datos y verificar la existencia de valores nulos, distintos, duplicados, entre otros; permitiendo así, definir nuevas RN que posteriormente pasan a ser las reglas de trasformación aplicadas en la implementación de la solución. Para más información consultar el artefacto “DATEC_CIM_hr3 - Perfilado de datos” dentro del Expediente de Proyecto de los subsistemas de almacenamiento e integración del producto hr3.

Después de realizarse el proceso de perfilado de datos a los EC del producto hr3 se obtuvieron los siguientes resultados:

- ✓ Se pudo identificar que los tipos de datos existentes eran *varchar, integer, date* y *double,* de los cuales en su mayoría eran enteros y cadenas como se muestra en la Fig. 8.

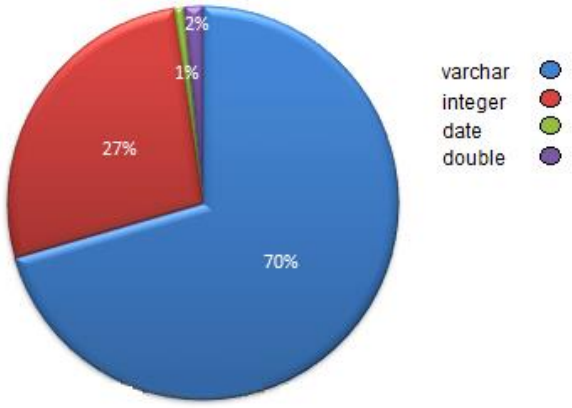


Fig. 8: Distribución de los tipos de datos en la fuente.

- ✓ Una regla de transformación que se pudo definir después de realizar el perfilado fue: las variables numéricas de tipo entero que se encuentren vacías tomarán valor -1 y 'No disponible' para las de tipo cadena.

2.6.5 Diseño general de las transformaciones.

Los procesos de ETL son los encargados de realizar las transformaciones para lograr una integración exitosa, permitiendo cargar correctamente los datos al mercado en las tablas correspondientes. Una vez acabado el perfilado de los datos fuente, se procede a realizar el diseño de las transformaciones. Estos procesos requieren de un alto nivel de detalle, por lo que es necesario realizar un diseño general de las transformaciones que describa los pasos para realizar la carga de los hechos y las dimensiones a la base de datos.

En la Fig. 9 se muestra el diseño general de las transformaciones para la carga de los hechos de los subsistemas de almacenamiento e integración del producto hr3. El diseño comienza estableciendo las variables de entorno que permitirán verificar la conexión con la base de datos *mercado_hr3*. Después se accede al directorio donde se encuentran los ficheros fuente y se obtienen los mismos. Luego se realiza la extracción de los datos que serán cargados en cada una de las tablas de los hechos del esquema *mart_hr3* dependiendo de las dimensiones con que se relaciona el hecho, y se realizan las transformaciones a los datos pertinentes. Se buscan las llaves dimensionales, verificándose la existencia

de llaves nulas, si existen se transforman los datos para realizar nuevamente la búsqueda de las llaves de las dimensiones. Si no hay llaves nulas se verifica la existencia de llaves huérfanas, si se cumple se pone un *id* por defecto según las dimensiones asociadas y se vuelve a verificar las llaves huérfanas. De no haber se inserta en la base de datos. Luego se obtiene la información del sistema para generar los dos tipos de metadatos: técnicos, donde serán almacenadas la fecha de inicio y fin de la carga, la dirección ip de la computadora donde se ejecutó la transformación y el nombre de la misma, y el de procesos contendrá las líneas leídas, líneas escritas, los errores, entre otros datos que brindan información acerca de la transformación que se cargó.

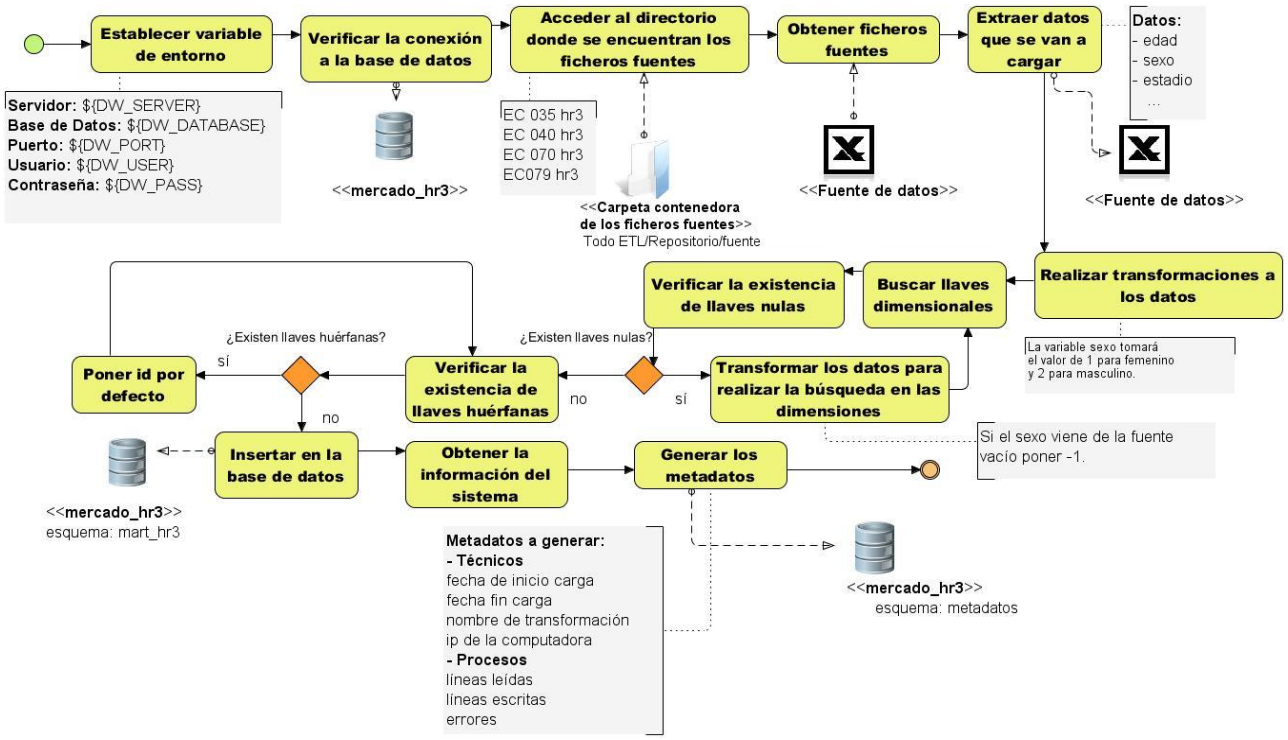


Fig. 9: Diseño general de las transformaciones para la carga de los hechos.

2.7 Política de respaldo y recuperación.

La pérdida de información trae consigo consecuencias embarazosas para la institución. Para evitar este tipo de problema se establece una política de respaldo y recuperación, con el objetivo de garantizar que sea duradera la misma. Debido a que el sistema posee una carga histórica, es decir, que los datos serán cargados una vez, se realizarán copias de seguridad a los subsistemas de almacenamiento e integración

del producto hr3, de manera que, al ocurrir una pérdida de la información se puede volver a cargar en la base de datos la copia de los mismos, quedando nuevamente poblada la base de datos.

También se propone realizar otras copias de la información en ubicaciones diferentes, previniendo la ocurrencia de fallos en el sistema o de otra índole, ya que es de gran importancia mantener segura y disponible la información.

2.7.1 Esquema de seguridad.

Es de gran importancia para un sistema de información implementar un mecanismo de protección contra aquellas acciones que puedan afectar la integridad, confidencialidad y disponibilidad de los datos almacenados. Por tal motivo, para el acceso al subsistema de almacenamiento del producto hr3 es necesario definir los roles que serán autorizados a acceder a la base de datos.

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

La 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, al marcar el atributo de solo lectura en las propiedades de la carpeta donde se almacenan todos los datos de las fuentes, las transformaciones y los trabajos. 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 hr3.

Conclusiones del capítulo.

Luego de haber realizado el análisis y diseño de los subsistemas de almacenamiento e integración del producto hr3 se pudo arribar a las siguientes conclusiones:

- ✓ Las necesidades de información identificadas fueron la base para definir los 20 requisitos de información y los 2 requisitos funcionales, agrupados en 18 casos de uso de información y 2

funcionales respectivamente. Además, se identificaron 3 requisitos no funcionales y se establecieron las RN que servirán de apoyo para realizar las transformaciones.

- ✓ La arquitectura diseñada contribuyó a definir los 3 componentes fundamentales que formarán parte de los subsistemas de almacenamiento e integración del producto hr3.
- ✓ A través del diseño del subsistema de almacenamiento se realizó el modelo dimensional de los datos, que cuenta con una topología de constelación de hechos, donde se identificaron 18 tablas para los hechos y 42 para las dimensiones.
- ✓ El perfilado de los datos realizado a la fuente de datos de los EC del producto hr3, se pudo identificar el estado en que estaban los datos y se definieron nuevas RN que luego pasaron a ser reglas de transformación.
- ✓ El diseño de las transformaciones permitió guiar la implementación de los procesos de ETL.

Capítulo 3: Implementación y pruebas de los subsistemas de almacenamiento e integración del producto hr3.

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 producto hr3. Se definen los estándares de codificación y la construcción del modelo físico. Se exponen las pruebas realizadas a los subsistemas, así como los resultados obtenidos en cada una de ellas. Dichas pruebas permitirán encontrar y corregir no conformidades existentes, obteniéndose como resultado una aplicación con mayor calidad.

3.1 Implementación del subsistema de almacenamiento.

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

3.1.1 Estándares de codificación.

Los estándares de codificación se utilizan para lograr un entendimiento entre las partes implicadas en un proyecto. Tienen como objetivo lograr un patrón que conduzca a la correcta normalización de los términos utilizados, es decir, estandarizar la forma de las estructuras de los subsistemas de almacenamiento e integración del producto hr3. Resulta conveniente conservar una nomenclatura estándar en el nombrado que permita un mejor entendimiento de las estructuras por parte de los desarrolladores.

En la solución, la nomenclatura se mantiene atendiendo a la clasificación de las diferentes estructuras, teniendo en cuenta el tipo de tabla. Si la tabla es una dimensión, al nombre de la misma le preceden las letras *dim* separadas del nombre de la dimensión por el caracter '_', ejemplo *dim_peso*. En caso de ser una tabla de hechos, como prefijo se ubican las letras *hech*, igualmente separadas del nombre de la tabla, por el caracter '_', ejemplo *hech_administracion_hr3*.

Para los atributos de las dimensiones se siguió la misma política para cada una de ellas. En el caso de las llaves de las dimensiones se les denominó *dk_dim_nombre_dimension_id*, ejemplo *dk_dim_peso_id*. Para

el caso de que el atributo de la misma sea un código del negocio se le especificó como *nombre_dimension_codigo*, ejemplo *dim_peso_codigo*; igualmente para los nombres, descripciones u otros atributos: ejemplo *peso_resultado* y *peso_descripcion,* respectivamente. De manera general los atributos fueron nombrados como *nombre_dimension_atributo*. Las medidas fueron definidas de la forma *cantidad_medida*, por ejemplo *cantidad_pacientes*.

El nombre de las transformaciones comienza con las letras *trans,* luego el caracter especial '_' y finalmente el nombre de la misma, ejemplo *trans_dim_sexo*. De la misma forma sucede con los trabajos, donde se antepone el nombre a las letras *trab* seguido del caracter '_', ejemplo *trab_general*. Por su parte los metadatos están conformados por las letras *md* y el nombre del mismo seguido del caracter '_', ejemplo *md_carga_historica*.

Luego de finalizar el proceso de estandarización de los nombres, queda organizada la nomenclatura utilizada para la denominación de las tablas, atributos y medidas dentro de la base de datos *mercado_hr3*, así como de las transformaciones y trabajos. Se procede entonces a la implementación del modelo de datos físico.

3.2 Implementación del modelo de datos físico.

El esquema de una base de datos define todas sus tablas, cada campo en cada una de ellas, incluyendo las relaciones entre ellos y las tablas. Para la solución se cuenta con 62 tablas, estas se encuentran divididas en 42 tablas de dimensiones, 18 tablas de hechos y 2 tablas de metadatos (Véase Fig. 10). Se define la utilización de tres esquemas:

- ✓ El esquema *dimensiones* contiene las dimensiones propias de los subsistemas de almacenamiento e integración del producto hr3.
- ✓ El esquema *mart_hr3* contiene las tablas de hechos de los subsistemas de almacenamiento e integración del producto hr3.
- ✓ El esquema *metadatos* contiene los metadatos técnicos y de procesos de los subsistemas de almacenamiento e integración del producto hr3.

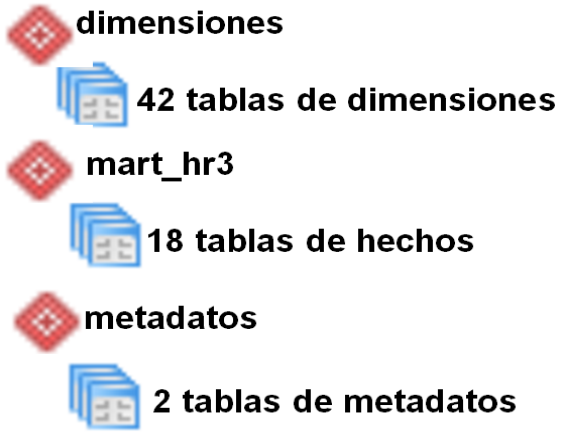


Fig. 10: Estructura física de la base de datos.

3.3 Implementación del subsistema de integración.

El proceso de integración de los datos consta de tres etapas fundamentales relacionadas entre sí: extracción, transformación y carga de los datos. La primera etapa de los procesos de ETL consiste en extraer la información desde los sistemas fuentes, en este caso ficheros en formato *xls*, donde se seleccionan los campos relevantes teniendo en cuenta el modelo de datos realizado. La segunda etapa es de transformación y limpieza, donde se detectan los datos incorrectos, las entradas duplicadas y se utilizan las transformaciones para corregir cualquier error existente en los datos. Por último, se procede a la carga, donde se toman los datos de la fase de transformación para cargarlos en el sistema destino, que consiste en la base de datos *mercado_hr3*.

Para lograr el correcto funcionamiento de los procesos de integración se tuvo en cuenta los cuatro grupos en que se dividen los subsistemas identificados por Kimball: extracción, limpieza y conformación, entrega y gestión. A continuación se mencionan los subsistemas identificados en el desarrollo de la solución propuesta.

Extracción

- ✓ **Sistema de extracción:** permite extraer los datos desde la fuente de origen, las cuales se encuentran en formato *xls*, para luego realizar las transformaciones de los datos y posteriormente

cargarlos. Para ello se tuvo en cuenta la información relacionada con cada uno de los hechos y dimensiones.

Limpieza y conformación

- ✓ **Perfilado de datos:** mediante este subsistema fueron definidas nuevas reglas de transformación. Permitió explorar los datos para verificar el cumplimiento de los estándares conforme a los requisitos especificados por el cliente y la calidad de los mismos.
- ✓ **Subsistema de transformación:** con este subsistema se realizaron transformaciones utilizando diferentes componentes para su creación, como los de mapeo de valores, creación de constantes, el filtrado de valores, entre otros.

Entrega

- ✓ **Dimensiones Lentamente Cambiantes (SCD):** son dimensiones en las cuales sus datos tienden a modificarse a través del tiempo. Fue utilizado el tipo 1 de SCD para dar tratamiento al cambio de la información asociada a las dimensiones.
- ✓ **Llave subrogada:** permite crear llaves subrogadas independientes para cada tabla. Para crearlas se utilizó el componente *Añadir secuencia* agregándole un *código* que pasaría a ser la llave subrogada de la dimensión.

Gestión

- ✓ **Repositorio de metadatos:** este subsistema permitió la captura de los metadatos de los procesos de ETL y de los aspectos técnicos. Con el primero se obtuvieron las líneas leídas, líneas escritas, errores, entre otros y para realizar el segundo se utilizaron los componentes *Información del sistema* e *Insertar Actualizar*.
- ✓ **Programador de trabajos:** permitió gestionar todos los trabajos de la solución, un ejemplo es el trabajo general, que se encarga de ejecutar en un orden específico las transformaciones creadas en los procesos de ETL. Se utilizaron algunos componentes para su creación como son el de crear trabajos y transformaciones, finalizar el trabajo, verificar conexión, entre otros.
- ✓ **Subsistema de carga:** permite realizar la carga de los datos a las tablas de dimensiones y hechos del subsistema de almacenamiento e integración del producto hr3.

3.4 Implementación de las transformaciones.

Las transformaciones están compuestas por pasos, que se encuentran unidos a través de saltos, en ellos se definen las reglas que serán establecidas en las transformaciones. A través de los saltos se distribuye la información entre los diferentes pasos.

Para el esquema *dimensiones* se confeccionó un flujo de transformación para la inserción de los datos en cada una de sus dimensiones. La transformación se realizó a partir de la carga de los datos de cada uno de los ficheros de la fuente, estos contienen todos los campos que son necesarios para poblar la base de datos, para luego poder elaborar las transformaciones de los hechos.

Para realizar la carga de la dimensión examen físico (Véase Fig. 11), el primer paso que se realiza es la extracción de los datos de la fuente (Excel de los EC mama modelo 2, modelo 3, modelo 4, meta cerebral modelo 4 y modelo 5a). Luego se seleccionaron los valores específicos a cargar, para después ser normalizados. Utilizando el componente *Append streams* se realizaron las uniones de todos los datos y con el de *Mapeo de Valores* realizado a la variable *valor*, donde se especificaron el nombre y los valores que tomará el campo origen y destino, se cambió el valor que venía de la fuente al que deseaba el cliente. Se realizó el ordenamiento con el componente *Ordenar Filas,* para posteriormente utilizar el de *Filas únicas* para seleccionar uno de los valores que se encuentran repetidos en el campo. Luego se añadió un *código*, que constituirá la llave subrogada de la dimensión y se le cambió el tipo de dato. Después se inserta en la base de datos *mercado_hr3*.

Inmediatamente se obtiene la información necesaria del sistema para generar los metadatos técnicos, que guarda el nombre del fichero fuente, el nombre de la transformación, del destino y la fecha de ejecución de la transformación. Por último se inserta dicha información en la tabla *md_carga_histórica*. Se crean metadatos de procesos, para obtener información de la transformación, haciendo *clic* derecho sobre la misma y seleccionando *Transformation Settings*, y en este caso se obtuvieron las líneas de entrada, leídas, actualizadas, de salida y los errores encontrados durante la ejecución, además de contar con otros elementos.

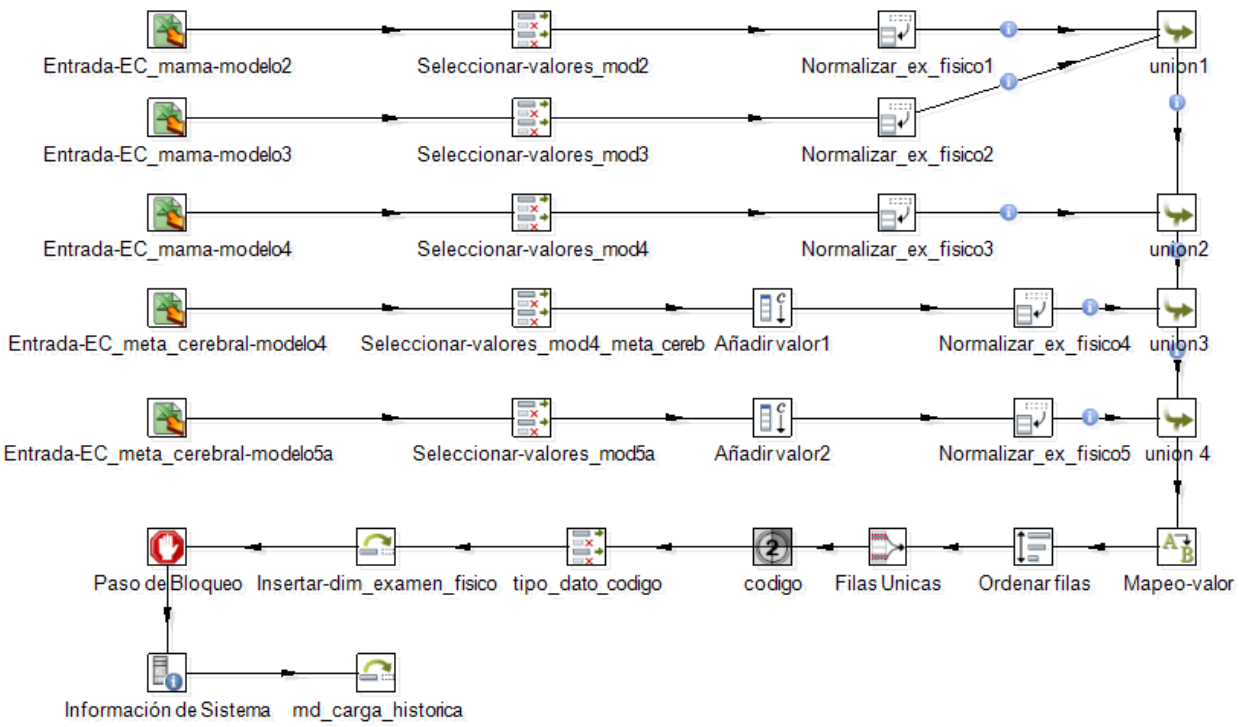


Fig. 11: Transformación de la dimensión examen físico.

Para realizar la extracción de los datos correspondientes a cada una de las tablas de hechos de la solución, primeramente se accede al fichero de la fuente (modelo 3, EC mama), donde son extraídos los campos necesarios dependiendo de las dimensiones con que está relaciona el hecho. Luego se procede a realizar las transformaciones pertinentes, donde se buscan las llaves dimensionales a partir de los datos que vienen de la fuente, para finalmente insertarlos en el esquema *mart_hr3* en las tablas de hechos.

Posteriormente, de igual forma que en las dimensiones, se realizan los metadatos técnicos y de procesos que son cargados en el esquema *metadatos* correspondiente. En la Fig. 12 se muestra la transformación para el hecho administración hr3 (*hech_adm_hr3*).

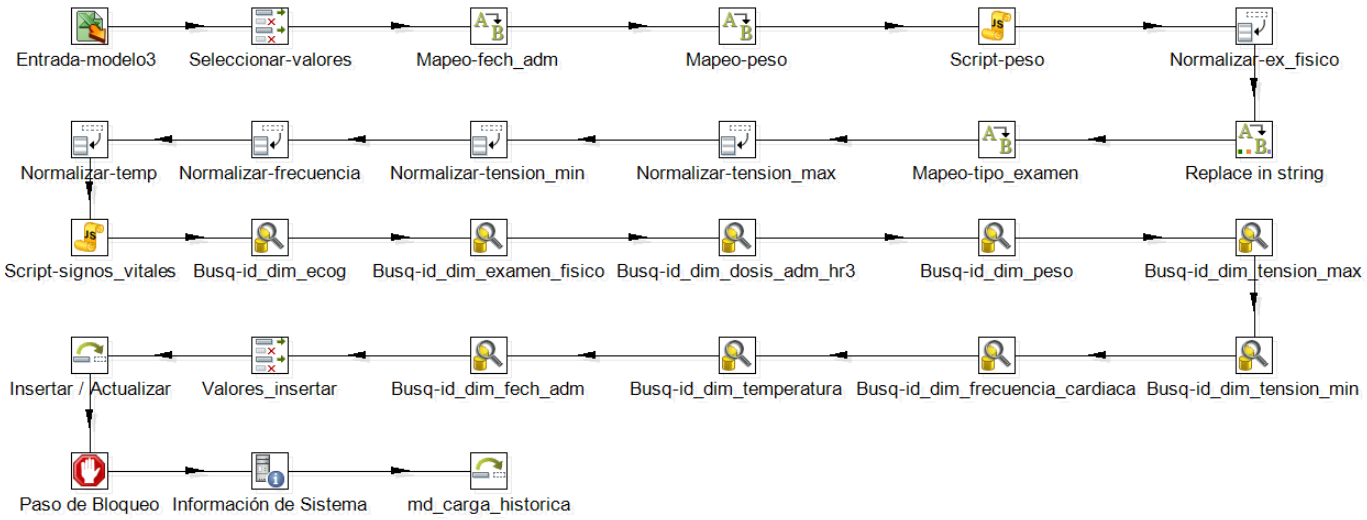


Fig. 12: Transformación del hecho administración hr3.

3.5 Implementación de los trabajos.

Una vez finalizadas las transformaciones necesarias para la carga de los datos, se llevó a cabo la implementación de los trabajos, los cuales representan un conjunto de tareas con el objetivo de realizar una acción determinada, siguiendo una secuencia lógica de pasos que permiten la ejecución de una o varias transformaciones. Con la utilización de los saltos o *hops* se indica el orden de ejecución de cada uno de los trabajos (no empezando la ejecución del elemento siguiente hasta que el anterior no haya concluido).

En la Fig. 13 se evidencia el trabajo general para la carga de cada una las dimensiones y hechos del producto hr3, de los esquemas *dimensiones* y *mart_hr3,* respectivamente. En primer lugar se conecta a la base de datos *mercado_hr3* y verifica la conexión, si no está conectada termina su ejecución, de lo contrario comienza la carga ejecutando los trabajos de las transformaciones de las dimensiones compartidas y de los hechos que solo tienen las que lo identifican, estos trabajos son: *trab_salida_ensayo* y el *trab_hech_evento_adverso*.

Luego se cargan los trabajos *trab_hechos1*, *trab_hech2* y *trab_hech3*, los cuales contienen las dimensiones compartidas, las que los identifican y la transformación del hecho en sí. Quedando conformado el trabajo general con los siguientes trabajos:

- ✓ ***trab_hechos1*** contiene los hechos *hech_tratamiento_quimioterapia, hech_adm_hr3, hech_ev_durante_ttmiento_nsclc, hech_evaluacion_inicial_nsclc, hech_evaluacion_de_respuesta_nsclc, hech_ev_inicial_ev_lesion_tto_concomitante* y *hech_laboratorio_clinico*.
- ✓ ***trab_hech2*** contiene los hechos *hech_datos_demografico_ex_fisico, hech_ev_inicial_farmacodinamia* y *hech_ttmiento_concomitante_nsclc.*
- ✓ ***trab_hech3*** contiene el hecho *hech_evaluacion_respuesta_clinica*.
- ✓ ***trab_salida_ensayo*** contiene los hechos *hech_interrupcion_del_tratamiento_nsclc, hech_salida_del_ensayo_cyc* y *hech_salida_ensayo_farmac*.
- ✓ ***trab_hech_evento_adverso*** contiene los hechos *hech_evento_adverso_ec_mama, hech_evento_adverso_farmac* y *hech_evento_adverso_nsclc.*
- ✓ ***trab_dim_compartidas*** contiene las dimensiones *dim_examen_fisico, dim_ecog, dim_signos_vitales, dim_examen_clinico, dim_peso, dim_estadio, dim_tnm* y *dim_karnosfky*.

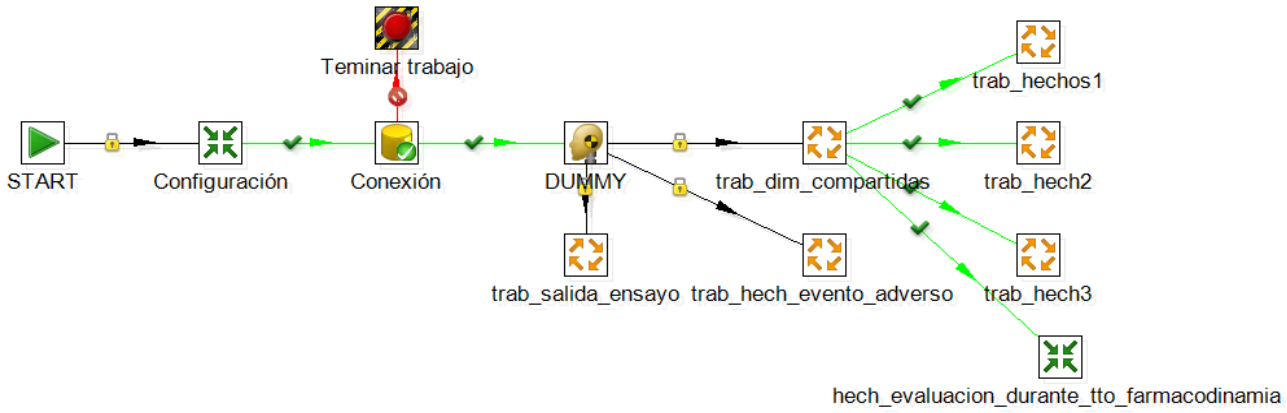


Fig. 13: Trabajo general para la carga de las dimensiones y hechos.

3.5.1 Gestión del cambio lento en las dimensiones.

Las dimensiones lentamente cambiantes o SCD (*Slowly Changing Dimensions*) determinan cómo se manejan los cambios históricos en las tablas de dimensiones. Sus datos tienden a modificarse a través del tiempo, ya sea de forma ocasional o constante e implicar a un solo registro o a la tabla completa. Cuando ocurren estos cambios se puede optar por seguir algunas de estas dos grandes opciones:

- ✓ Registrar el historial de cambios.
- ✓ Reemplazar los valores que sean necesarios.

Ralph Kimball planteó algunas estrategias a seguir para las SCD, las más usadas son las SCD tipo 1, 2 y 3 describiéndose a continuación: (24)

Tipo 1 (sobrescribir): es utilizado cuando la información histórica no es importante. Este tipo sobrescribe los datos antiguos por nuevos y es utilizado mayormente para corregir errores de datos en las dimensiones.

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 AD.

Tipo 3 (añadir columna): esta estrategia requiere que se agregue una nueva columna a la tabla por cada columna, cuyos valores se desea mantener en un historial de cambios. De este modo en la nueva columna se coloca el valor antiguo antes de sobrescribir el valor actual con el nuevo. Este tipo presenta como principal desventaja que solo permite guardar un historial limitado de los datos, dependiendo del número de columnas que se cree.

En la presente investigación la estrategia de SCD empleada es de tipo 1. La cual permite una vez cargados los datos corregir errores que sean identificados en los procesos de integración en caso de que fuese necesario. Un ejemplo encontrado en la base de datos *mercado_hr3* después de estar poblada, fue el error ortográfico en la dimensión *examen físico* en la columna *examen_fisico_nombre* en la tupla que contenía el valor *Corazon* que se cambió por *Corazón*.

3.5.2 Gestión de los metadatos.

Los metadatos son datos que ayudan a identificar, describir y localizar recursos digitales, son información estructurada que describe y/o permite encontrar, gestionar, controlar y entender o preservar otra información; o sea que no son más que datos sobre los propios datos. (24)

El almacenamiento y el uso de los metadatos permiten y facilitan el manejo de los datos, el uso consistente de los datos, el entendimiento de los datos, y la explotación de volúmenes de información que son accesibles en línea.

A continuación se presentan algunas de las clasificaciones de los metadatos en el contexto de AD: (24)

- ✓ 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, como son los datos provenientes de la fuente.

En la presente investigación se utilizan los metadatos de proceso para obtener la información correspondiente a los procesos de las transformaciones y los técnicos para la gestión de carga histórica, mostrándose en la Fig. 14 un ejemplo del último.

id serial	fecha_inicial date	fecha_fin date	nombre character varying(50)	ip character varying(15)
15	2014-04-23	2014-04-23	hech_evaluacion_inicialnsclc	10.8.107.4
16	2014-04-23	2014-04-23	hech_ev_durante_ttonsclc	10.8.107.4
17	2014-04-24	2014-04-24	hech_tto_concomitante_nsclc	10.8.107.4
18	2014-04-24	2014-04-24	hech_evaluacion_respuesta_nsclc	10.8.107.4
19	2014-04-24	2014-04-24	hech_evaluacion_resp_clinica	10.8.107.4

Fig. 14: Metadatos técnicos para la gestión de la carga histórica.

3.6 Pruebas.

Todo proceso de creación de un *software* está sujeto a fallos, es por esto que las pruebas constituyen una fase importante en su desarrollo y de esta manera lograr que un producto cumpla con la calidad requerida. Una prueba se considera exitosa si encuentra alguna deficiencia en el *software*. Para obtener diferentes tipos de errores en el sistema se hace necesario aplicar un amplio conjunto de pruebas. En el presente trabajo de diploma para probar los resultados de la solución se realizaron pruebas unitarias y de integración en las cuales se utilizaron los casos de prueba, además se aplicaron las listas de chequeo a los artefactos de ETL.

3.7 Pruebas unitarias.

Las pruebas unitarias permiten probar el correcto funcionamiento de un componente o subsistema en específico. Estas pruebas son desarrolladas por los propios desarrolladores durante la implementación de la solución. (9) Luego de concluida la etapa de implementación fueron aplicadas las pruebas unitarias a los subsistemas de almacenamiento e integración del producto hr3, donde se realizaron dos iteraciones. Para una primera iteración se detectaron 5 No Conformidades (NC) a los dos subsistemas, las cuales se mencionan a continuación. En una segunda iteración fueron tratadas y corregidas en su totalidad (Véase Fig. 15).

NC_1. Revisar los RF y RNF, los cuales estaban mal redactados.

NC_2. El artefacto de “DATEC_CIM_hr3-0113_ERS” se encontraba desactualizado, por lo que la revisión no pudo completarse correctamente.

NC_3. Revisar el modelo de datos, atendiendo a posibles cambios en los atributos de los hechos y dimensiones.

NC_4. Revisar el diseño general de las transformaciones.

NC_5. Definir la estructura física de los datos.

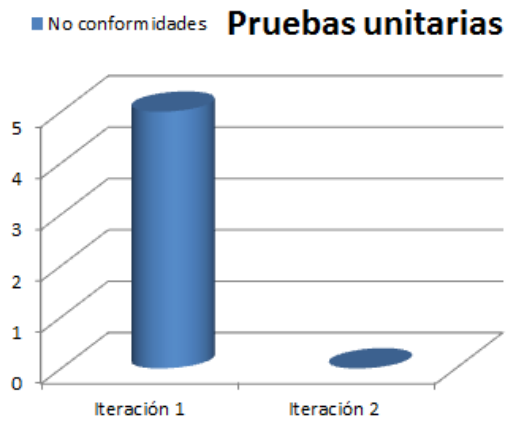


Fig. 15: Resultado de aplicar las pruebas unitarias.

3.8 Pruebas de integración.

Las pruebas de integración permiten verificar la correcta integración de los componentes y subsistemas que conforman la solución. Ponen a prueba la vista arquitectónica del sistema definida en una infraestructura de desarrollo. Estas pruebas son ejecutadas por los arquitectos de *software*. (9) Las herramientas utilizadas para aplicar las pruebas de integración son las listas de chequeo y los casos de prueba.

3.8.1 Casos de prueba.

Los casos de prueba son utilizados para identificar posibles errores en la implementación y comprobar que los requisitos especificados por el cliente se cumplan correctamente en el sistema.

En la presente investigación se realizaron consultas a la base de datos donde se obtuvieron resultados satisfactorios, demostrándose de esta manera que los datos de la fuente fueron cargados en su totalidad. Se diseñaron 18 casos de prueba, con el propósito de verificar los requisitos de información que fueron definidos previamente durante la etapa de análisis. Estos se encuentran en el artefacto “DATEC_CIM_hr3” ubicado en el Expediente de Proyecto de los subsistemas de almacenamiento e integración del producto hr3.

En la Fig. 16 se muestra un ejemplo de la consulta realizada al CU Almacenar la información del modelo Administración del hr3. Luego de ejecutarse esta consulta en lenguaje SQL en la base de datos

mercado_hr3, se obtuvo como resultado que 21 pacientes cumplían con el conjunto de condiciones de dicha consulta.

```
SELECT
  count (distinct (hech_administracion_hr3.pk_codigo_paciente,
  hech_administracion_hr3.asistio_a_la_administracion )) as resultado
FROM
  mart_hr3.hech_administracion_hr3,
  dimensiones.dim_dosis_administrada_hr3,
  dimensiones.dim_examen_fisico
WHERE
  dim_dosis_administrada_hr3.dk_dim_dosis_administrada_hr3_id =
  hech_administracion_hr3.dk_dim_dosis_administrada_hr3_id AND
  dim_examen_fisico.dk_dim_examen_fisico_id =
  hech_administracion_hr3.dk_dim_examen_fisico_id AND
  hech_administracion_hr3.asistio_a_la_administracion LIKE 'verdadero' AND
  dim_dosis_administrada_hr3.dosis_administrada_hr3_valor = 400 AND
  dim_examen_fisico.examen_fisico_nombre LIKE 'Piel' AND
  dim_examen_fisico.examen_fisico_valor LIKE 'verdadero';
```

Fig. 16: Consulta realizada al CU Almacenar la información del modelo Administración del hr3.

Luego se aplicaron los filtros necesarios al *Modelo3.Adm hr3* del EC Mama FI, donde se obtuvo que de los 129 pacientes contenidos en el modelo, solo 21 de ellos fueron evaluados por la *dosis administrada de hr3* donde su valor fue 400, *si asistió a la administración*, además, fueron evaluados por el *examen físico* donde el nombre fue *Piel* y su valor fue *verdadero*.

Al realizarse las comparaciones entre el resultado arrojado por la consulta en el lenguaje SQL y los filtros al *Modelo3.Adm hr3* se pudo comprobar que ambos eran iguales, por lo que la prueba fue satisfactoria.

3.9 Listas de chequeo.

Las listas de chequeos están conformadas por una serie de preguntas, a partir de las cuales se verifica el grado de cumplimiento de las reglas establecidas. Se utilizaron con el fin de medir una serie de indicadores implicados en la creación de la capa de integración, además de medir la calidad de los artefactos y documentos generados durante la realización del producto. Los indicadores a evaluar se encuentran distribuidos en tres secciones fundamentales:

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

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

En la Fig. 17 se muestra el resultado después de haber aplicado las listas de chequeo a los artefactos DATEC_CIM_hr3 - Mapa lógico de datos, DATEC_CIM_hr3 - Perfilado de datos, DATEC_CIM_hr3 - Registro del sistema fuente y DATEC_CIM_hr3 - Diccionario de Datos.

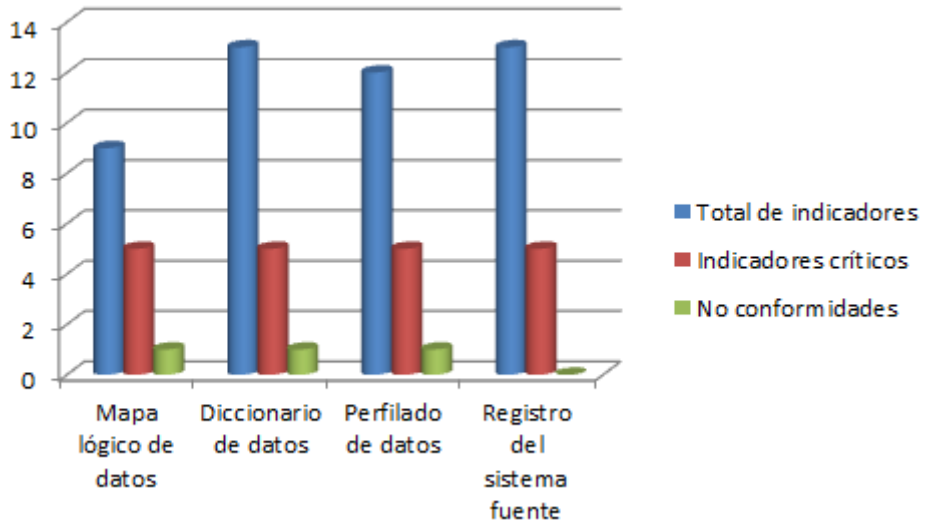


Fig. 17: Resultado después de aplicar la lista de chequeo a los artefactos.

3.10 Calidad de datos.

En el desarrollo de los subsistemas de almacenamiento e integración del producto hr3 el proceso de calidad de datos es de gran importancia, debido a que de esta forma se puede comprobar que no posean errores los datos cargados.

Perfilado de datos

A través del uso de la herramienta Data Cleaner se obtuvieron resultados positivos respecto a los datos cargados del producto hr3 y se pudo apreciar con el análisis de este proceso que se realizó correctamente la carga de los datos correspondientes a cada uno de los hechos, permitiendo comprobar que no fueron almacenados valores vacíos ni nulos y los hechos contienen únicamente valores enteros, exceptuando el código del paciente que es cargado directamente de la fuente. El siguiente gráfico muestra los resultados

correspondientes al perfilado de datos realizado a la base de datos *mercado_hr3*, que contiene todos los datos cargados de la fuente. (Véase Fig. 18).

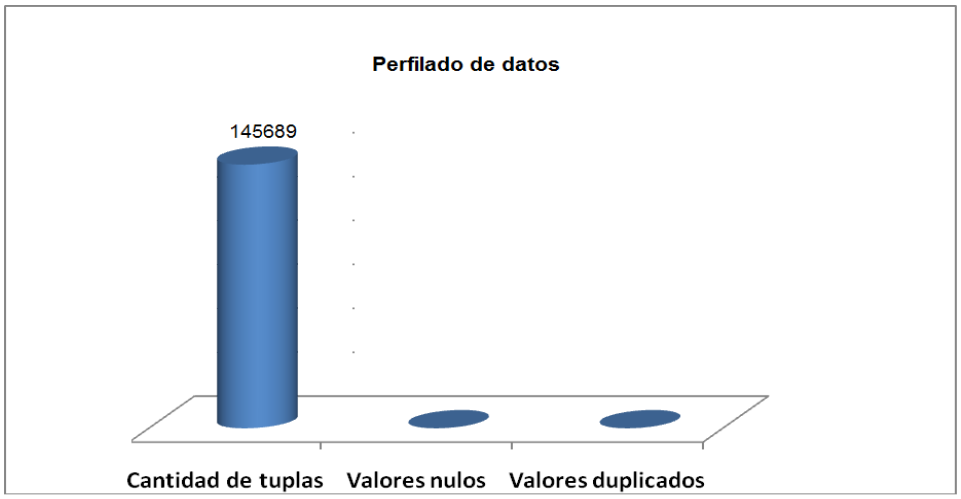


Fig. 18: Resultado realizado a la base de datos mercado_hr3.

Conclusiones parciales.

Durante el capítulo se abordó acerca de la implementación y pruebas realizadas a los subsistemas de almacenamiento e integración del producto hr3, concluyendo así, el proceso de construcción y prueba, obteniéndose los siguientes resultados:

- ✓ Se definió la estructura de los subsistemas de almacenamiento e integración del producto hr3, quedando conformada con los esquemas *dimensiones, mart_hr3* y *metadatos*.
- ✓ Se implementaron los subsistemas de almacenamiento e integración del producto hr3, obteniéndose la base de datos *mercado_hr3* poblada.
- ✓ Se realizaron 18 transformaciones para la carga de los hechos y 42 para las dimensiones que permitieron eliminar la existencia de valores duplicados, errores ortográficos e inconsistencia de los datos, además, se implementaron 7 trabajos, de los cuales 1 fue para la carga de dimensiones comunes, 5 para la carga de los hechos y 1 trabajo general que permitió la ejecución ordenada de las transformaciones.

- ✓ Se realizaron las pruebas necesarias durante las distintas etapas de desarrollo, permitiendo comprobar la funcionalidad del sistema a partir de los requisitos establecidos.

Conclusiones Generales.

Una vez concluida la investigación y desarrollo de los subsistemas de almacenamiento e integración del producto hr3 se arriba a las siguientes conclusiones:

- ✓ La elaboración del marco teórico facilitó la selección de herramientas y tecnología para la construcción de los subsistemas de almacenamiento e integración del producto hr3 y el estudio de la metodología permitió guiar el proceso de desarrollo.
- ✓ El análisis de los subsistemas de almacenamiento e integración del producto hr3, permitió identificar los requisitos de información y funcionales, sirviendo estos de guía para la elaboración del diagrama de casos de uso del sistema, además de identificarse los requisitos no funcionales y las reglas del negocio.
- ✓ Se realizó el diseño de los subsistemas de almacenamiento e integración del producto hr3 arrojando como elementos principales el modelo dimensional, el perfilado de los datos, el diseño general de las transformaciones y la arquitectura general para los subsistemas de almacenamiento e integración del producto hr3.
- ✓ La implementación de los procesos de ETL contribuyó a la carga de los datos de los subsistemas de almacenamiento e integración del producto hr3, donde su estructura física quedó compuesta por esquemas y tablas.
- ✓ Utilizando los casos de pruebas, las listas de chequeo y las pruebas unitarias se comprobó la calidad y funcionalidad de los subsistemas de almacenamiento e integración del producto hr3, a partir de los requisitos establecidos.

Recomendaciones.

Aplicar proceso de descubrimiento de conocimiento sobre el subsistema de almacenamiento del producto hr3, que permitan detectar patrones de comportamiento sobre la información almacenada.

Bibliografías Referenciadas.

1. Centro de Información Cardiovascular. [En línea] [Citado el: 18 de enero de enero.] http://www.texasheartinstitute.org/HIC/Topics_Esp/FAQ/clinical_trials_span.cfm.

2. **Ross, R. Kimball y M.** The Data Warehouse Toolkit: the Complete Guide to Dimensional Modelling. New York, Estados Unidos: s.n., 2002.

3. **NUÑEZ CÁRDENAS, Felipe de Jesús.** Introducción a Almacén de Datos. 2011.

4. **HURTADO TORRES, M. Visitación, et al.** Bases de Datos y Data Warehouse: Herramientas estratégicas para la eficacia comercial. Facultad de Ciencias Económicas y Empresariales: Universidad de Granada, 2002.

5. MEDITECH. Almacén de Datos para MEDITECH. [En línea] [Citado el: 6 de enero de 2014.] https://www.meditech.com/MeditechEnEspanol/Pages/es_data_repository.pdf,citado.

6. **Ramirez, Ms. C Martha Denia Hernandez.** Procedimiento para el desarrollo de un sistema de inteligencia de negocios en la gestión de ensayos clínicos en el Centro de Inmunología Molecular. s.l.: Hernández Ramírez, Revista Cubana de Información en Ciencias.

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

8. **Sanchez, Leopoldo Zenaido Zepeda.** Departamento de Sistemas Informáticos y Computación. [En línea] [Citado el: 28 de febrero de 2014.] http://personales.unican.es/ruizfr/is1/doc/lab/01/is1-p01-trans.pdf.

9. **Hernández, Yanisbel González.** Metodología de Desarrollo para proyectos de almacenes de datos. La Habana : s.n.: s.n., 2013.

10. **BASALLO, Yasser Azán, ESTRADA, Anay Díaz y GÓMEZ, Salvador González.** Una experiencia en integración de aplicaciones empresariales. s.l.: Revista Cubana de Ciencias Informáticas, 2009. vol. 3, no 3-4.

11. **MUFIOZ, L., MAZON, Jose-Norberto y TRUJILLO, Juan.** ETL process modeling conceptual for data warehouses: a systematic mapping study. Latin America Transactions: IEEE (Revista IEEE America Latina), 2011. vol. 9, no 3, p. 358-363.

12. **WOLFF, C.** La Tecnología Data Ware Housing. [En línea] [Citado el: 3 de marzo de 2014.] http://www. utpl. edu. ec/eva/descargas/material/140/INFAII21 G, vol. 4181003.

13. **Meza, Mirna.** Herramientas Case . [En línea] [Citado el: 2 de diciembre de 2014.] http://fds-herramientascase.blogspot.com.

14. **Sierra, Maria.** Herramientas CASE. [En línea] [Citado el: 20 de diciembre de 2014.] http://personales.unican.es/ruizfr/is1/doc/lab/01/is1-p01-trans.pdf.

15. **CAVSI.** cavsi. ¿Qué es un Sistema Gestor de Bases de Datos o SGBD? . [En línea] [Citado el: 27 de diciembre de 2014.] http://www.cavsi.com/preguntasrespuestas/que-es-un-sistema-gestor-de-bases-de-datos-o-sgbd.

16. **kinderman, Hans.** PostGreSQL. [En línea] [Citado el: 20 de enero de 2014.] http://postgressql-adsi.blogspot.com.

17. **pgAdmin.** PostgreSQL administration and management tools. [En línea] [Citado el: 20 de enero de 2014.] http://www.pgadmin.org/index.php.

18. **SOLÍS VELASCO, Margarita Isabel.** Propuesta Metodológica para la Gestión de la Calidad de Datos en Proyectos de Integración, Caso Práctico. s.l.: SII-ESPOCH, 2012.

19. **HERNÁNDEZ, Esther Naranjo y MOSQUERA, C. Ing Inty Sáez.** Pentaho: software líder de Inteligencia de Negocio de código abierto. s.l.: Revista Telem@ tica, 2012. vol. 10, no 2.

20. **DÍAZ, Josep Curto.** Introducción al Business Intelligence. . s.l.: Editorial UOC, 2012.

21. **Urquisu, Pau.** ¿Qué es OLAP? [En línea] [Citado el: 11 de enero de 2014.] http://www.businessintelligence.info/definiciones/que-es-olap.html.

22. **TAMAYO, Marysol y MORENO, Francisco Javier.** Análisis del modelo de almacenamiento MOLAP frente al modelo de almacenamiento ROLAP. s.l.: Ingeniería e Investigación, 2006. vol. 26, no 3, p. 135-142.

23. **CEDEÑO TRUJILLO, Alexis.** MODELO MULTIDIMENSIONAL. s.l.: Ingeniería Industrial, 2010. vol. 27, no 1, p. 4 pág.

24. **Autor corporativo.** Auditoría de sistemas. s.l., Manizales : Universidad de Caldas, 2009.

Bibliografía Consultada.

BASALLO Yasser Azán, ESTRADA Anay Díaz y GÓMEZ Salvador González Una experiencia en integración de aplicaciones empresariales [Libro].- [s.l.]: Revista Cubana de Ciencias Informáticas, 2009.- vol. 3, no 3-4.

BONIFATI Angela, et al. Designing data marts for data warehouses. ACM transactions on software engineering and methodology [Libro].- 2001.- vol. 10, no 4, p. 452-483.

CAVALLERI M., et al A set of tools for building PostgreSQL distributed databases in biomedical environment. En Engineering in Medicine and Biology Society, 2000. Proceedings of the 22nd Annual International Conference of the IEEE [Libro].- [s.l.]: IEEE, 2000.- p. 540-544.

CAVSI. cavsi [En línea]// ¿Qué es un Sistema Gestor de Bases de Datos o SGBD? .- 27 de diciembre de 2014 - http://www.cavsi.com/preguntasrespuestas/que-es-un-sistema-gestor-de-bases-de-datos-o-sgbd.

CEDEÑO TRUJILLO Alexis. MODELO MULTIDIMENSIONAL [Libro].- [s.l.]: Ingeniería Industrial, 2010.- vol. 27, no 1, p. 4 pág.

Centro de Información Cardiovascular [En línea].- 18 de enero de enero.- http://www.texasheartinstitute.org/HIC/Topics_Esp/FAQ/clinical_trials_span.cfm.

CHAUDHURI Surajit y DAYAL Umeshwar An overview of data warehousing and OLAP technology [Libro].- [s.l.]: ACM Sigmod record, 1997.- vol. 26, no 1, p. 65-74.

DÍAZ Josep Curto. Introducción al Business Intelligence. [Libro].- [s.l.]: Editorial UOC, 2012.

HERNÁNDEZ Esther Naranjo y MOSQUERA C. Ing Inty Sáez. Pentaho: software líder de Inteligencia de Negocio de código abierto [Libro].- [s.l.]: Revista Telem@ tica, 2012.- vol. 10, no 2.

Hernández Yanisbel González Metodología de Desarrollo para proyectos de almacenes de datos [Libro].- La Habana : s.n.: [s.n.], 2013.

HURTADO TORRES M. Visitación, et al Bases de Datos y Data Warehouse: Herramientas estratégicas para la eficacia comercial [Libro].- Facultad de Ciencias Económicas y Empresariales: Universidad de Granada, 2002.

IEEE-SA [En línea]// IEEE-SA.- 19 de mayo de 2014- http://standards.ieee.org/develop/project/29119-1.html.

INMON William H.e. Building the data warehous [Libro].- [s.l.]: John wiley & sons, 2005.

kinderman Hans. [En línea]// PostGreSQL.- 20 de enero de 2014.- http://postgressql-adsi.blogspot.com.

LEE Mong Li, LING Tok Wang y LOW Wai Lup IntelliClean: a knowledge-based intelligent data cleaner. En Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining [Libro].- [s.l.]: ACM, 2000.- p. 290-294.

MEDITECH. [En línea]// Almacén de Datos para MEDITECH.- 6 de enero de 2014.- https://www.meditech.com/MeditechEnEspanol/Pages/es_data_repository.pdf,citado.

Meza Mirna. [En línea]// Herramientas Case .- 2 de diciembre de 2014.- http://fds-herramientascase.blogspot.com.

MUFIOZ L., MAZON Jose-Norberto y TRUJILLO Juan ETL process modeling conceptual for data warehouses: a systematic mapping study. [Libro].- Latin America Transactions: IEEE (Revista IEEE America Latina), 2011.- vol. 9, no 3, p. 358-363.

NUÑEZ CÁRDENAS Felipe de Jesús Introducción a Almacén de Datos. [Libro].- 2011.

pgAdmin [En línea]// PostgreSQL administration and management tools.- 20 de enero de 2014.- http://www.pgadmin.org/index.php.

POMPA Yisel de los Ángeles González y GONZÁLEZ María Teresa Rosales Mercados de datos para el análisis estadístico de la información. [Libro].- [s.l.]: 3C Tecnología, 2014.- vol. 3, no 1.

Ramirez Ms. C Martha Denia Hernandez Procedimiento para el desarrollo de un sistema de inteligencia de negocios en la gestión de ensayos clínicos en el Centro de Inmunología Molecular [Libro].- [s.l.]: Hernández Ramírez, Revista Cubana de Información en Ciencias.

RICARDO DARIO Ing. Bernabeu Data Warehousing: Investigación y sistematización de conceptos. Hefesto: Metodología propia para la construcción de un Datawarehouse [Libro].- Córdoba, Argentina: [s.n.], 2009.

Ross R. Kimball y M The Data Warehouse Toolkit: the Complete Guide to Dimensional Modelling. [Libro].- New York, Estados Unidos: [s.n.], 2002.

Sanchez Leopoldo Zenaido Zepeda. [En línea]// Departamento de Sistemas Informáticos y Computación.- 28 de febrero de 2014.- http://personales.unican.es/ruizfr/is1/doc/lab/01/is1-p01-trans.pdf.

Sierra Maria [En línea]// Herramientas CASE.- 20 de diciembre de 2014.- http://personales.unican.es/ruizfr/is1/doc/lab/01/is1-p01-trans.pdf.

SOLÍS VELASCO Margarita Isabel Propuesta Metodológica para la Gestión de la Calidad de Datos en Proyectos de Integración, Caso Práctico [Libro].- [s.l.]: SII-ESPOCH, 2012.

TAMAYO Marysol y MORENO Francisco Javier. Análisis del modelo de almacenamiento MOLAP frente al modelo de almacenamiento ROLAP. [Libro].- [s.l.]: Ingeniería e Investigación, , 2006.- vol. 26, no 3, p. 135-142.

Urquisu Pau. [En línea]// ¿Qué es OLAP?.- 11 de enero de 2014.- http://www.businessintelligence.info/definiciones/que-es-olap.html.

WOLFF C. [En línea]// La Tecnología Data Ware Housing.- 3 de marzo de 2014.- http://www. utpl. edu. ec/eva/descargas/material/140/INFAII21 G, vol. 4181003.

Glosario de Términos.

TCP/IP: siglas de Protocolo de Control de Transmisión/Protocolo de Internet (del inglés Transmisión Control Protocol/Internet Protocol), sistema de protocolos que posibilita diversos servicios de red.

SSL: Secure Sockets Layer (SSL; en español capa de conexión segura) es un protocolo criptográfico que proporciona comunicaciones seguras por una red, comúnmente Internet.

NoSQL: en informática es llamado “no sólo SQL”es una amplia clase de sistemas de gestión de bases de datos que difieren del modelo clásico del sistema de gestión de bases de datos relacionales (RDBMS) en aspectos importantes, el más destacado que no usan SQL como el principal lenguaje de consultas.

OMS: siglas de la Organización Mundial de la Salud.

Ecog: escala para calidad de vida del paciente**.**

Karnofsky: escala para evaluar la capacidad de un paciente para sobrevivir a la quimioterapia para el cáncer.

Trigger (o disparador): es un procedimiento que se ejecuta cuando se cumple una condición establecida al realizar una operación.

SQL: es un lenguaje declarativo de acceso a bases de datos relacionales que permite especificar diversos tipos de operaciones en ellas.

Anexos.

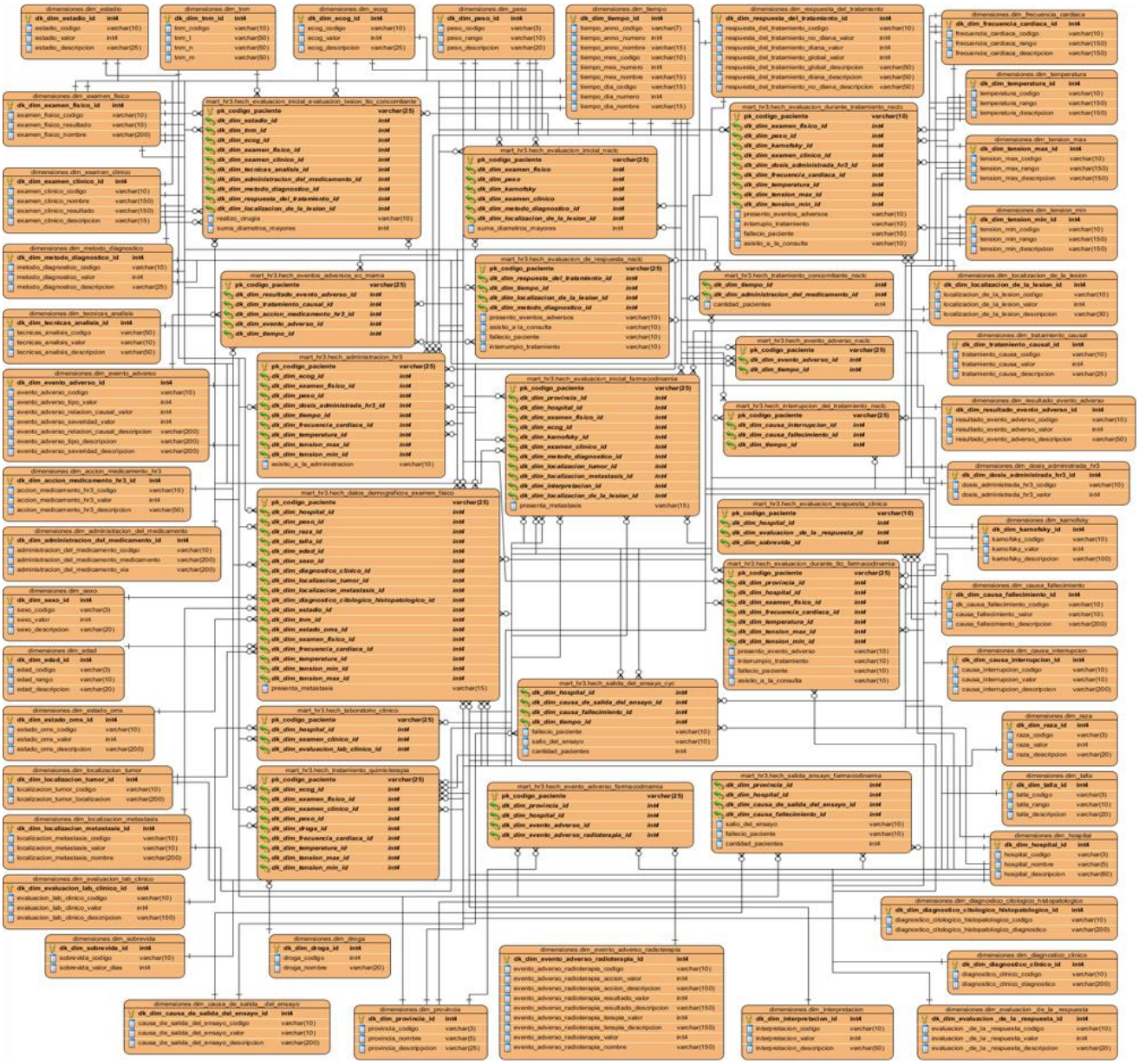


Fig. 19: Modelo de datos.

Fig. 20: Diagrama de casos de uso del sistema.