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 ACM 14F7 para los ensayos clínicos del Centro de Inmunología Molecular.*

Autora:

Yanislet Lorenzo Paz

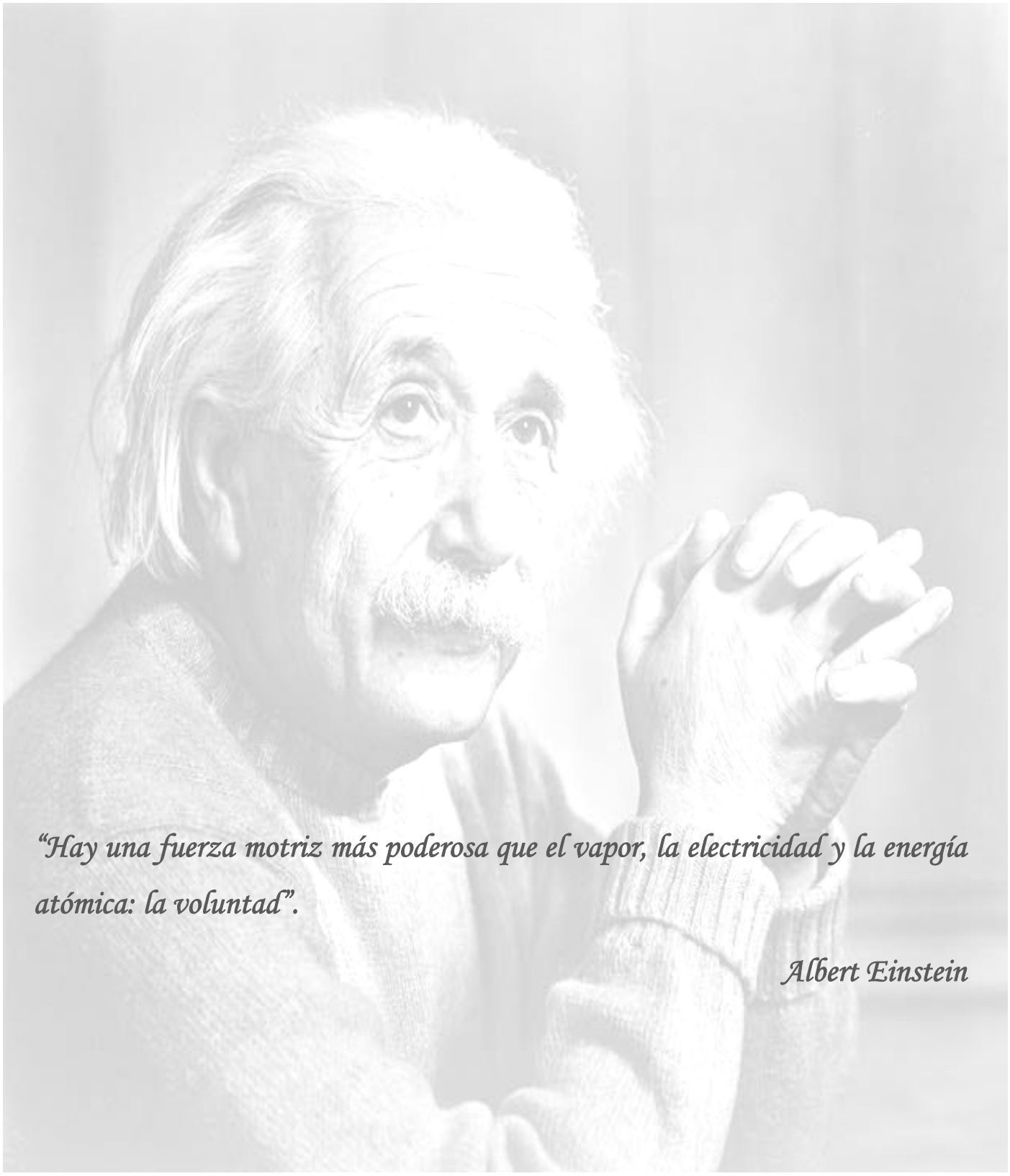
Tutores:

Msc. Yadira Barroso Rodríguez

Ing. Lazaro José Estupiñan Cutiño

La Habana, 2014

"Año del 56 Aniversario de la Revolución"



"Hay una fuerza motriz más poderosa que el vapor, la electricidad y la energía atómica: la voluntad".

Albert Einstein

Yo: Yanislet Lorenzo Paz, declaro ser autora de la presente tesis y reconozco a la Universidad de las Ciencias Informáticas los derechos patrimoniales de la misma, con carácter exclusivo.

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

Yanislet Lorenzo Paz

_____

Firma del Autor

Lazaro José Estupiñan Cutiño

_____

Firma del Tutor

Yadira Barroso Rodríguez

_____

Firma del Tutor

Msc. Yadira Barroso Rodríguez

Universidad de las Ciencias Informáticas, La Habana, Cuba.

Correo Electrónico: ybarroso@uci.cu

Ing. Lazaro José Estupiñan Cutiño

Universidad de las Ciencias Informáticas, La Habana, Cuba.

Correo Electrónico: ljestupinan@uci.cu

Yanislet Lorenzo Paz

Universidad de las Ciencias Informáticas, La Habana, Cuba.

Correo Electrónico: ylorenzo@estudiantes.uci.cu

Resumen

La presente investigación se desarrolla en el marco de la colaboración entre la Universidad de las Ciencias Informáticas y el Centro de Inmunología Molecular, debido a la necesidad de este último de gestionar, almacenar y analizar toda la información que se recoge en los Ensayos Clínicos (EC) asociados al producto ACM 14F7. Los mismos recogen información de pacientes que padecen cáncer de mama, incluyendo datos relacionados con los eventos adversos, la interrupción del tratamiento, la inclusión y evaluación inicial de los pacientes que participaron en dichos ensayos.

Ante la necesidad expuesta por este centro se pretende desarrollar los subsistemas de almacenamiento e integración del producto ACM 14F7 que contribuyan a mantener la información referente a este producto centralizada, estandarizada y accesible para su consulta. Con este fin se realiza un estudio de las metodologías, herramientas y tecnologías que permitirán la construcción del sistema y el correcto almacenamiento de los datos. Además se efectúa el análisis, diseño e implementación de los subsistemas de almacenamiento e integración y se aplican un conjunto de pruebas, con el objetivo de verificar la calidad del producto.

Palabras claves: Centro de Inmunología Molecular, Ensayos Clínicos, cáncer de mama, subsistemas de almacenamiento e integración.

Abstract

The present research is developed by the collaboration between the University of Informatics Sciences and the Center for Molecular Immunology, because this center has the need to manage, store and analyze all information contained in Clinical Trials (EC) of the product ACM 14F7. These clinical trials collect information about patients with breast cancer; also include data as adverse events, discontinuation of treatment, inclusion and initial evaluation of the patients.

Given the need of this center is to create a storage and integration subsystems where the information is centralized, standardized and accessible for consultation. A study of methodologies, tools and technologies that will enable the construction of the system and correct storage of the data is carry out. It also analyzes, design and implement the storage and integration subsystems. In addition, checklists and test cases are applied to obtain an efficient product.

Keywords: Center of Molecular Immunology, Clinical Trials, Breast Cancer, storage and integration subsystems.

Introducción

El vertiginoso avance de las Tecnologías de la Información y las Comunicaciones (TICs) a nivel mundial ha permitido el incremento de los logros científicos, investigativos y tecnológicos aplicados en diferentes esferas como la salud, la educación y la industria. Basado en los resultados alcanzados muchas empresas han optado por automatizar sus procesos, ya que se vive una época donde la información adquiere una importancia vital, además su adecuado manejo y preservación podría representar una ventaja competitiva en los negocios.

Cuba es uno de los tantos países del mundo donde se han experimentado estos avances tecnológicos y se ha propuesto extender la informática hacia todos los sectores de la sociedad. Tras el triunfo de la Revolución en aras de lograr la informatización del país se crearon numerosos centros científicos e investigativos que trabajan en programas de producción de alto valor agregado, especialmente en el campo biotecnológico y médico-farmacéutico. Como prueba de ello puede destacarse la creación del Centro de Inmunología Molecular (CIM) el 5 de diciembre de 1994 y de la Universidad de las Ciencias Informáticas (UCI) fundada el 23 de septiembre de 2002. Esta última ha sido vanguardia en el desarrollo de sistemas informáticos que fortalecen indudablemente la industria del software en Cuba (1).

El CIM se creó para analizar el comportamiento de enfermedades que aquejan la sociedad actual y causan la muerte de millones de personas a nivel global. En esta institución se crean biofármacos capaces de combatir peligrosas afecciones y contrarrestar sus síntomas, garantizando el mejoramiento de la calidad de vida del pueblo. Además son realizados numerosos estudios o Ensayos Clínicos (EC) en varios hospitales cubanos, por lo que uno de sus objetivos es: gestionar, almacenar y analizar toda la información que se recoge en los mismos una vez aplicado un producto a pacientes enfermos. Las investigaciones llevadas a cabo en este centro se enfocan en la inmunoterapia del cáncer, específicamente en vacunas moleculares, bioinformática e ingeniería celular (1).

Una de las enfermedades que se investigan en el CIM es el cáncer, afección que provoca una generación rápida de células anormales que pueden invadir zonas adyacentes del organismo o diseminarse a otros órganos dando lugar a la formación de las llamadas metástasis. Cada año aparecen en el planeta unos 10 millones de casos nuevos y mueren por esta causa 7 millones de personas (2).

Entre los EC desarrollados en esta institución se encuentran los del producto ACM 14F7, el cual ha sido utilizado para combatir el cáncer de mama. En la actualidad la cantidad de EC asociados al mencionado producto ha generado un volumen considerable de información. La misma se encuentra almacenada en bases de datos dispersas y en diferentes formatos, donde predominan las hojas de cálculo del programa Excel (formato xls). La dispersión de los datos y la falta de estandarización de los mismos ha condicionando su manejo inadecuado por parte de los especialistas del centro.

Para gestionar los EC relacionados a este producto debe consultarse un gran volumen de documentación. Actualmente el CIM no cuenta con la cantidad de especialistas necesarios para realizar esta ardua labor, por lo que se torna engorroso el análisis de la misma y se ven afectadas las decisiones que se deben tomar respecto a la efectividad del producto. Por este motivo debe invertirse mucho más tiempo para realizar el trabajo y es común que una vez terminada de examinar toda la documentación muchas de las decisiones tomadas sean obsoletas.

Otro de los problemas presentados por el CIM es la carencia de una plataforma de trabajo que permita dar seguimiento a los datos desde su captación, procesamiento, consulta y estandarización. Esto ha propiciado que los especialistas deban realizar manualmente la gestión de cada uno de los procesos lo que resulta sumamente difícil y dificulta la realización de análisis certeros que contribuyan con las decisiones estadísticas que deben ser tomadas.

Además existen inconvenientes a la hora de confeccionar los reportes, analizar y consultar la información recopilada y presentar los indicadores relacionados con dichos ensayos. A esto se le añade el riesgo de que se pierda información útil y valiosa con el transcurso del tiempo impidiendo un adecuado tratamiento de la misma.

Por la situación anteriormente descrita, se plantea como **Problema de la Investigación**: ¿Cómo lograr la estandarización de los datos del producto ACM 14F7 para el Centro de Inmunología Molecular que contribuya a su almacenamiento homogéneo?

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 expuesto anteriormente se plantea como **Objetivo General**: Desarrollar los subsistemas de almacenamiento e integración del producto ACM 14F7 para los ensayos clínicos del Centro de Inmunología Molecular que contribuya al almacenamiento homogéneo de la información.

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

1. Selección de la metodología, herramientas y tecnologías a utilizar en el desarrollo de los subsistemas de almacenamiento e integración del producto ACM 14F7 para los ensayos clínicos del Centro de Inmunología Molecular.
2. Levantamiento de los requisitos del sistema para definir las necesidades del cliente.
3. Perfilar los datos para garantizar la limpieza y calidad de los mismos.
4. Descripción de los subsistemas de almacenamiento e integración del producto ACM 14F7 para definir las funcionalidades del sistema.
5. Definición de la arquitectura de los mercados de datos mediante la identificación de los subsistemas fundamentales que los componen.
6. Diseño del subsistema de almacenamiento para definir las estructuras necesarias que permitan almacenar la información referente al producto ACM 14F7.
7. Diseño del subsistema de integración para definir cómo se realizará la carga de las dimensiones y los hechos.
8. Implementación del subsistema de almacenamiento para crear las estructuras necesarias que permitan almacenar la información referente al producto ACM 14F7.
9. Implementación del subsistema de integración para estandarizar los datos del producto ACM 14F7.
10. Aplicación de las listas de chequeo para garantizar la correcta implementación de los subsistemas de almacenamiento e integración del producto ACM 14F7.
11. Aplicación de los casos de prueba para avalar la disponibilidad de cada uno de los elementos de los subsistemas de almacenamiento e integración del producto ACM 14F7.

A lo largo de la investigación fueron utilizados métodos del nivel teórico y empírico como métodos científicos, estos permitieron comprender las características y prioridades del negocio, contribuyendo así a la resolución del problema general.

Como métodos del nivel teórico fueron utilizados los métodos **históricos** y **lógicos** ya que a través de estos fueron analizados los antecedentes históricos y racionales de los almacenes de datos, su evolución y principales tendencias a nivel mundial. Además fue utilizado el método **analítico-sintético**, que permitió el establecimiento de un conjunto de tareas que posibilitaron el cumplimiento del objetivo general de la investigación. También se utilizó el método de **modelación**, el cual contribuyó a diseñar e implementar los subsistemas de almacenamiento e integración del producto ACM 14F7 a través de modelos que permitieron representar gráficamente la solución.

Como métodos empíricos fueron seleccionados la **observación** ya que a través de la misma se pudo analizar y comprender de forma directa la información relacionada a la investigación y la **entrevista** la cual permitió la definición de las necesidades del cliente y posteriormente las reglas del negocio.

Una vez definidos los métodos científicos utilizados en la investigación se hace necesario plantear una estrategia a seguir basada en las características de la investigación, la trayectoria del problema, el conocimiento acumulado sobre el mismo y los objetivos que se persiguen. Existen tres tipos fundamentales de estrategias o estudios; exploratoria, descriptiva y experimental o explicativa. En la investigación se utiliza la **estrategia exploratoria**, la misma tiene como objetivo la familiarización del investigador con un tema específico, en este caso se evidencia con el estudio y análisis de los almacenes y mercados de datos, sus características y tendencias.

Las investigaciones que utilizan la estrategia exploratoria carecen de hipótesis por lo que es imprescindible el planteamiento de un conjunto de preguntas de investigación condicionadas por la descomposición del problema en subproblemas más pequeños. Estas orientarán la investigación hacia el cumplimiento del objetivo general. Las **preguntas de investigación** elaboradas se muestran a continuación:

1. ¿Cuáles son los fundamentos teóricos que proporcionan la base para el desarrollo de los subsistemas de almacenamiento e integración del producto ACM 14F7 para los ensayos clínicos del Centro de Inmunología Molecular?
2. ¿Cómo organizar el proceso de desarrollo de los subsistemas de almacenamiento e integración que permita la estandarización de los datos del producto ACM 14F7 para los ensayos clínicos del Centro de Inmunología Molecular?

3. ¿Cuáles pruebas deben realizarse para verificar que los subsistemas de almacenamiento e integración desarrollados permiten la estandarización y el almacenamiento homogéneo de los datos del producto ACM 14F7?

Estructura del documento

El presente documento está estructurado de la siguiente manera: resumen, introducción, tres capítulos, conclusiones, recomendaciones, referencias bibliográficas, bibliografía, anexos y glosario de términos.

Capítulo 1: Fundamentación teórica de los mercados de datos.

En el capítulo se abordan los conceptos y definiciones relacionadas con los almacenes y los mercados de datos, sus principales características y las ventajas que proporciona su uso en el mundo empresarial. Se exponen además, aspectos importantes tales como: la metodología de desarrollo a utilizar y las herramientas empleadas en los procesos de integración de datos.

Capítulo 2: Análisis y diseño de los subsistemas de almacenamiento e integración del producto ACM 14F7.

En el capítulo se aborda la etapa de análisis que constituye la base para comprender los requisitos establecidos por el cliente y posteriormente obtener las reglas del negocio. Se realiza el diagrama de casos de uso del sistema y se define además la arquitectura en la que se basará la solución proponiéndose un diseño de la misma, con las características necesarias para satisfacer las necesidades manifestadas por el cliente.

Capítulo 3: Implementación y prueba de los subsistemas de almacenamiento e integración del producto ACM 14F7.

El capítulo está dirigido a la implementación de los diferentes aspectos relacionados con los procesos de integración y almacenamiento de datos, con el propósito de brindar una mayor comprensión de las estrategias y procedimientos utilizados. Además se realiza la validación de los subsistemas de almacenamiento e integración del producto ACM 14F7 a través de varias pruebas que garantizan el cumplimiento de las exigencias del cliente y la calidad del producto.

Capítulo 1: Fundamentación teórica de los mercados de datos

En el capítulo se abordan los conceptos y definiciones relacionadas con los almacenes y los mercados de datos, sus principales características y las ventajas que proporciona su uso en el mundo empresarial. Se exponen además, aspectos importantes tales como: la metodología de desarrollo a utilizar y las herramientas empleadas en los procesos de integración de datos.

1.1 Gestión de ensayos clínicos en el Centro de Inmunología Molecular

Un EC es un tipo de estudio clínico en el que se evalúa la eficacia o seguridad de nuevos fármacos o tratamientos médicos a través de la aplicación a seres humanos con un protocolo de investigación estrictamente controlado. Estos permiten a los médicos determinar si un nuevo tratamiento, medicamento o dispositivo contribuirá a prevenir, detectar o tratar una enfermedad (3).

Los EC son conducidos por agencias del gobierno, instituciones educativas, organizaciones sin ánimo de lucro, o empresas comerciales que prueban el funcionamiento de los nuevos enfoques clínicos en las personas. Cada EC tiene un protocolo o plan de acción para llevarlo a cabo. El plan describe lo que se hará en el estudio, cómo se hará y por qué cada parte del estudio es necesaria (4).

La información recogida en cada estudio se almacena en Cuadernos de Recogida de Datos (CRD), en los cuales se reúne toda la información relacionada con el paciente durante su tratamiento. Una vez culminado dicho estudio se envían los CRD para el CIM, lugar donde se realiza el proceso de digitalización de la información almacenada mediante el sistema EpiData, generándose reportes en diferentes formatos (Text, dBaseIII, Excel, Stata, SPSS y SAS) (1).

En la investigación se realizará específicamente el almacenamiento y la integración a los datos referentes al ensayo 14F7 Mama FII 073, el cual incluye pacientes con cáncer de mama.

1.2 Almacenes de datos

Para cualquier organización la información es vital, por lo que su digitalización es un punto clave para manipularla de forma adecuada. Los almacenes de datos (AD) son la respuesta a la necesidad de muchas empresas de analizar grandes volúmenes de información apoyando el proceso de toma de decisiones en las mismas. La definición más difundida de un AD es la expresada por William H. Inmon, quien plantea

que: ..."Un almacén de datos consiste en una colección de datos orientada al negocio, integrada, no volátil y variante en el tiempo, para el apoyo a la toma de decisiones administrativas" (5).

Una de las personalidades más influyentes en el área, Ralph Kimball propone otra definición al catalogarlo como: ..."una copia de datos transaccionales, específicamente estructurados para la consulta y el análisis" (6). También existen otras propuestas de definiciones, por ejemplo la de Ricardo Chinchilla: un almacén de datos es una base de datos donde la información extraída de los sistemas operacionales corrientes de la empresa es transformada, integrada y resumida para luego ser usada con efectividad en el soporte de decisiones (7).

En la actualidad existe divergencia entre los diferentes conceptos expuestos por Inmon y Kimball. En realidad ninguno de los dos criterios es incorrecto, simplemente cada autor optó por un enfoque diferente de cómo diseñar un AD. En la presente investigación se concluye que un AD es la unión de varios mercados de datos los cuales constituyen subconjuntos del almacén. Estos contienen información de las principales áreas temáticas de la empresa, obteniéndola de una o varias fuentes para su posterior análisis a través del tiempo.

Los almacenes de datos presentan cuatro características fundamentales (5):

- **Temático:** los datos almacenados están organizados de manera que todos los elementos relativos al mismo evento u objeto del mundo real queden unidos entre sí.
- **Integrado:** la base de datos contiene información de todos los sistemas operacionales de la organización, y dichos datos deben ser consistentes.
- **No volátil:** la información no se modifica ni se elimina, una vez almacenado un dato, éste se convierte en información de sólo lectura, y se mantiene para futuras consultas.
- **Histórico:** los cambios producidos en los datos a lo largo del tiempo quedan registrados para que los informes que se puedan generar reflejen esas variaciones.

1.3 Mercado de datos

Diversos autores han planteado su criterio acerca del concepto Mercado de Datos (MD). Una de las definiciones más abarcadoras es la propuesta por Ralph Kimball: "...un conjunto flexible de datos, idealmente basado en el dato más atómico posible (granular) para ser extraído de las fuentes operacionales y presentado en un modelo simétrico (dimensional), que es más resistente cuando se

enfrentan con las más inesperadas consultas de los usuarios...". Estos están conectados con la arquitectura de los AD en su forma más simple y representan los datos de un proceso del negocio (6).

Los MD facilitan la construcción de los AD ya que representan un subconjunto del almacén que agrupa requisitos de un área temática de la empresa. Además ofrecen una mejor manejabilidad de la información almacenada ya que la misma es mucho menor que la contenida en un AD (8).

1.3.1 Características de los mercados de datos

Por la gran similitud entre los almacenes y los MD estos comparten las mismas características de integración, variación histórica, no volatilidad y orientación temática. Además estos últimos se caracterizan por (9):

- Proporcionar un esquema de almacenamiento que permita realizar consultas complejas, manteniendo solo un repositorio con la información agregada y ordenada.
- Permitir al usuario la posibilidad de análisis de la información para la toma de decisiones.
- Solucionar problemas derivados de la dispersión de los datos.

1.3.2 Ventajas y desventajas de los mercados de datos

Los mercados de datos presentan un gran número de ventajas, entre ellas se pueden señalar (10):

- Se centran en los requisitos de los usuarios asociados a un departamento o área de negocio específica.
- Contienen metadatos (datos sobre los datos), los cuales permiten conocer la procedencia de la información, su periodicidad de actualización, fiabilidad y forma de cálculo.
- Son más fáciles de utilizar y de comprender, debido a que la cantidad de información que contienen es mucho menor que la de un AD.

La principal desventaja de los MD es que no permiten el manejo de grandes volúmenes de datos por lo que muchas veces deben utilizarse varios de estos para cubrir todas las necesidades de información de una empresa.

1.3.3 Tendencias actuales de los mercados de datos

Cuba ha alcanzado un gran desarrollo tecnológico en las últimas décadas con el auge de la informática. En la actualidad se han desarrollado en el país varios mercados de datos, por ejemplo:

- Mercado de datos para la Dirección de Salud en Cuba (11).
- Diseño del MD CIMAVAX EGF2[1] para el CIM (8).
- Mercado de datos para el módulo visor de historias clínicas del Sistema Integral de Atención Primaria de la Salud (12).

1.4 Subsistemas de almacenamiento e integración del producto ACM 14F7

Para el desarrollo de los subsistemas de almacenamiento e integración del producto ACM 14F7 es necesario conocer las etapas por las cuales transitarían estos para así dar cumplimiento al objetivo general de la investigación. Un MD se divide en tres etapas, las cuales se corresponden a cada uno de los subsistemas que lo componen (almacenamiento, integración y visualización). La primera es el análisis y diseño donde se realiza un estudio del negocio y se diseñan las estructuras que contendrán los datos, la segunda etapa es la extracción, transformación y carga donde se extraen los datos de la fuente, se estandarizan, integran y se cargan en la base de datos. La última etapa es la inteligencia de negocio donde se analiza el comportamiento de los datos.

Se ha determinado en la investigación realizar dos de los subsistemas de un MD que darán cumplimiento al objetivo de la misma. El primero es el subsistema de almacenamiento que estará en correspondencia con la primera etapa y el segundo el de integración, el cual está relacionado a la segunda etapa. Se desarrollarán solamente los subsistemas de almacenamiento e integración ya que serán aplicadas posteriormente técnicas de minería de datos a los mismos.

1.5 Modelo multidimensional

El modelo multidimensional permite la modelación y el análisis de los datos, los cuales se almacenan como hechos, medidas y dimensiones.

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

Se define como hecho una operación que se realiza en el negocio la cual está estrechamente relacionada con el tiempo y es objeto de análisis para la toma de decisiones dentro del área del negocio que se está estudiando. También puede verse como un valor numérico que representa una actividad específica casi siempre con cifras que se suman entre sí (13).

Las medidas pueden ser denominadas como atributos numéricos asociados a los hechos (lo que realmente se mide) por ejemplo: volumen de las ventas, coste asociado a un producto, número de transacciones efectuadas y porcentaje de beneficios (17). Las mismas se clasifican según su tipo y su aditividad (6).

Tipo: se clasifican en físicas y calculables; la primera es cuando la medida es un valor físico que debe estar en el mercado y la segunda es cuando se calcula de alguna manera teniendo en cuenta medidas físicas del mercado.

Aditividad: la medida puede ser aditiva cuando es combinada y no pierde sentido su valor; no aditiva cuando no puede ser combinada y semi-aditivas cuando puede ser combinada en algunas dimensiones pero en otras no.

Las tablas de dimensiones especifican la organización lógica de los datos y proporcionan el medio para analizar el contexto del negocio. Representan los elementos del análisis, proporcionándole al usuario el filtrado y manipulación de la información almacenada en la tabla de hechos (14). También pueden definirse como perspectivas o entidades respecto a las cuales una organización quiere mantener sus datos organizados (por ejemplo: tiempo, localización, clientes, proveedores) (15) (16).

Contenida en estas tablas, en muchos casos aparece la relación lógica entre dos o más atributos, es decir las jerarquías de datos, conjuntamente, de acuerdo a las dimensiones del negocio, se refleja la granularidad, que representa el nivel de detalle en que se desea almacenar la información en el modelo de datos (17).

1.6 Metodología de desarrollo

Las metodologías son un conjunto de pasos definidos para lograr uno o varios objetivos. En el desarrollo de software esta surge ante la necesidad de utilizar una serie de procedimientos, técnicas, herramientas y soporte documental a la hora de desarrollar un producto (18).

En la actualidad existen varias propuestas de metodologías pero las más relevantes son las propuestas por Inmon y Kimball. Los conceptos de estos dos autores se contraponen de forma tal que Inmon propone una metodología *top-down* donde se plantea que el AD debe responder a las necesidades de todos los usuarios en la organización por lo que debe construirse primero y luego dividirse en MD, mientras que la metodología *bottom-up* defiende que el AD debe desarrollarse por pequeños MD los cuales deben ser integrados al final (19).

Existen también otras metodologías que integran los dos enfoques especificados anteriormente utilizando los aspectos más importantes de cada uno y conformando así una nueva metodología derivada hasta cierto punto de las otras dos, a continuación se muestran ejemplos de estas metodologías (19):

Metodología SQLBI, avalada por Microsoft y orientada totalmente a sus herramientas *Microsoft SQL Server*, *SQL Server Analysis Services* y su oferta más completa en este campo que es *Microsoft Suite for Business Intelligence* (19).

Metodología DM2, se basa en las necesidades de información a nivel gerencial, donde la información debe ser encarada como patrimonio de la empresa, accesible a quien la necesite (19).

Hefesto, plantea que la construcción e implementación de un AD puede adaptarse muy bien a cualquier ciclo de vida de desarrollo de software, excepto para algunas fases en particular, donde las acciones que se han de realizar serán muy diferentes (21).

Metodología para el Diseño Conceptual de Almacenes de Datos, esta última es presentada en la Tesis de Doctorado de Leopoldo Zenaido Zepeda Sánchez. Aporta como aspecto novedoso con respecto a las anteriores la incorporación de una serie de transformaciones para llevar un diagrama relacional a uno dimensional y así obtener las estructuras candidatas que conformarán el repositorio de datos. Además plantea incluir los casos de uso para guiar el proceso de desarrollo de un AD (20).

Ninguna de las metodologías anteriormente expuestas abarcan todas las fases de desarrollo de los subsistemas de almacenamiento e integración que serán realizados en la investigación, por lo que se hace necesario la utilización de una metodología específica que ha sido definida por el departamento de Almacenes de Datos del Centro de Tecnologías de Gestión de Datos (DATEC) perteneciente a la UCI, la misma se expone a continuación.

1.6.1 Metodología a utilizar

En la investigación será utilizada la **Metodología para el Desarrollo de Almacenes de Datos en DATEC** ya que reúne las características para el cumplimiento del objetivo y se ajusta al marco de trabajo de DATEC. Esta metodología está alineada a las tendencias y normas de la UCI. La misma toma como base la metodología Ciclo de Vida de Kimball y la complementa con algunos aspectos de la Metodología para el Diseño Conceptual de Almacenes de Datos. Otra de sus características es que permite la identificación de los requisitos de información y su trazabilidad a lo largo del ciclo de desarrollo de un MD, además incluye una etapa de pruebas que permite determinar la calidad del producto (19).

La metodología a utilizar se divide en ocho fases, las cuales se explicarán a continuación (19):

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

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

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

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

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

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

Soporte y mantenimiento: puede realizarse a través de varios servicios, que pueden ser: soporte en línea, vía telefónica, correo u otros. Se realizan las tareas de mantenimiento de la aplicación tan

necesarias para este tipo de desarrollo y que garantizan el adecuado funcionamiento y crecimiento del AD.

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

En la investigación se llevaron a cabo solo las primeras cinco fases; estudio preliminar y planeación, requisitos, arquitectura, implementación y prueba, ya que las restantes son implementadas por personal capacitado del departamento Almacenes de Datos del centro DATEC y no se encuentran dentro del alcance de la investigación.

1.7 Herramientas

Para llevar a cabo el modelado, diseño e implementación de los subsistemas de almacenamiento e integración se requieren una gran variedad de herramientas que harán posible el desarrollo de la aplicación. Las características y definiciones de las mismas serán especificadas a continuación:

1.7.1 Herramientas de modelado

Con el incremento de la información manejada por las empresas se hace imprescindible la utilización de una herramienta de modelado que permita realizar un nuevo diseño de datos o el estudio de un modelo existente.

Se podrían ejemplificar muchas herramientas de modelado como *MagicDraw UML*, *Rational Rose*, *open System Architect 4.0.0*, *Squirrel SQL 3.2.1* pero dependiendo de sus características, costo, ventajas y desventajas se utilizará para el modelado de los subsistemas de almacenamiento e integración la herramienta *Visual Paradigm for UML 8.*0.

Visual Paradigm for UML 8.0

Visual Paradigm es una herramienta para el desarrollo de aplicaciones utilizando un lenguaje unificado de modelado (UML). Es ideal para ingenieros de software, analistas y arquitectos de sistemas que están interesados en construcción de sistemas a gran escala y necesitan confiabilidad y estabilidad en el desarrollo orientado a objetos (22).

Entre las funcionalidades que ofrece Visual Paradigm se encuentran:

- Navegación intuitiva entre la escritura del código y su visualización.

- Potente generador de informes en formato PDF/HTML.
- Documentación automática *Ad-hoc*.
- Ambiente visualmente superior de modelado.
- Sofisticado diagramador automáticamente de *layout.*
- Sincronización de código fuente en tiempo real.
- Exportación de imágenes en formato jpg, png y svg.

1.7.2 Sistemas Gestores de Base de Datos

Se define un Sistema Gestor de Bases de Datos (SGBD) como un conjunto de programas no visibles al usuario final que proporcionan una interfaz entre el usuario, las aplicaciones y la base de datos. Además permiten manipular la información garantizando la privacidad, integridad y seguridad de los datos (23).

Entre los sistemas gestores de bases de datos más utilizados se encuentran Oracle, PostgreSQL y MySQL. La UCI se encuentra actualmente inmersa en un proceso de migración a software libre por lo que se descarta el SGBD Oracle, debido a que es una herramienta propietaria altamente costosa. La herramienta MySQL es mucho más veloz que PostgreSQL, pero no lo supera en cuanto a estabilidad y facilidad de uso.

Luego de analizados los SGBD existentes y sus ventajas, se optó por la utilización de la herramienta PostgreSQL en el desarrollo de los subsistemas de almacenamiento e integración del producto ACM 14F7.

PostgreSQL 9.1

PostgreSQL es un SGBD relacional, orientado a objetos y libre, publicado bajo la licencia BSD[2]. Presenta funcionalidades que permiten la existencia de herencia, tipos de datos, funciones, restricciones y disparadores (24).

Esta herramienta presenta varias ventajas como son (25):

- Puede soportar distintos tipos de datos como los de tipo fecha, monetarios, elementos gráficos, datos sobre redes, cadenas de bits, entre otros. También permite la creación de tipos propios.

[2] BSD: distribución de software *Berkeley*

- Incluye herencia entre tablas.
- Es multiplataforma, está disponible para Windows, Linux y Unix en todas sus variantes.
- Es extensible, su código fuente está disponible para todos los usuarios sin costo. Se puede extender o personalizar con un mínimo esfuerzo y sin costos adicionales.

1.7.3 Administrador de Base de Datos

Las herramientas que permiten la administración de base de datos son fundamentales para el desarrollo de una investigación de este tipo ya que se encargan de mantener la integridad y disponibilidad de los datos. Estas herramientas crean y configuran bases de datos relacionales, además se encargan de llevar a cabo el diseño de la distribución de la información, las soluciones de almacenamiento, el despliegue y monitorización de servidores de bases de datos. En la presente investigación se seleccionó para la gestión del SGBD PostgreSQL el administrador de bases de datos PgAdmin III en su versión 1.14.1.

PgAdmin III 1.14.1

PgAdmin III es una herramienta de código abierto para la administración de bases de datos PostgreSQL. Está diseñada para responder a las necesidades de la mayoría de los usuarios, desde escribir simples consultas SQL hasta desarrollar bases de datos complejas. La interfaz gráfica soporta todas las características de PostgreSQL y hace simple la administración. Está disponible en más de una docena de lenguajes y para varios sistemas operativos, incluyendo Microsoft Windows, Linux, FreeBSD, Mac OSX y Solaris (26).

Entre las características que posee PgAdmin III se encuentran (27):

- Es multiplataforma.
- Posee entradas SQL aleatorias.
- Posee pantallas de información y “ayudas" para bases de datos, tablas, índices, secuencias, vistas, programas de arranque, funciones y lenguajes.
- Ofrece preguntas y respuestas para configurar usuarios, grupos y privilegios.
- Ofrece control de revisión con mejora de la generación de script.
- Ofrece configuración de las tablas de *Microsoft MSysConf.*
- Posee “ayudas” para importar y exportar datos.

- Posee “ayudas” para migrar bases de datos.
- Ofrece informes predefinidos en bases de datos, tablas, índices, secuencias, lenguajes y vistas.

1.7.4 Herramientas para la integración de datos

Para la construcción de los subsistemas de almacenamiento e integración el proceso de ETL[3] es uno de los más importantes ya que garantiza la disponibilidad y confiabilidad de la información. Este proceso organiza el flujo de los datos entre diferentes sistemas en una organización y aporta los métodos y las herramientas necesarias para mover datos desde múltiples fuentes hacia un AD, reformatearlos, limpiarlos y cargarlos en otra base de datos o MD (28). Los procesos de ETL se explican a continuación (11):

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

Entre las herramientas de Extracción, Transformación y Carga más conocidas en el mercado mundial se destacan (29):

- *IBM Websphere DataStage*
- *Pentaho Data Integration*
- *SAS ETL Studio*
- *Oracle Warehouse Builder*
- *Informatica PowerCenter*
- *Ab Initio*
- *Business Objects Data Integrator (BODI)*
- *Microsoft SQL Server Integration Services (SSIS)*

[3] ETL: extracción, transformación y carga.

Después de realizar un estudio de las herramientas para la integración de los datos se determinó utilizar *Pentaho Data Integration* en su versión 4.4.0 por las potencialidades que presenta, y *DataCleaner* en su versión 1.5.3, para el perfilado de los datos.

Pentaho Data Integration 4.4.0 (PDI)

Esta herramienta permite extraer la información de las diferentes fuentes, transformar la información a través de un modelo dimensional y cargar los resultados de la transformación en una base de datos tipo AD, para que luego pueda ser consultada y analizada a través de herramientas para desarrollar reportes especializados las cuales Pentaho también posee (25). Entre sus ventajas se encuentran (31):

- Es multiplataforma.
- Permite el uso de tecnologías estándar: Java, XML[4], *javaScript*.
- Está basada en dos tipos de objetos: transformaciones (colección de pasos en los procesos de ETL) y trabajos (colección de transformaciones).
- Soporta diferentes sistemas gestores de bases de datos, por ejemplo: MySQL y PostgreSQL.

DataCleaner 1.5.4

Una de las tareas más importantes que se deben realizar en el proceso de análisis de la información es el perfilado de datos, ya que a través de este se puede ver claramente la estructura y el formato de las fuentes de datos. La herramienta *DataCleaner* está orientada a preparar los datos para cualquier proyecto en el que se deban aplicar técnicas de calidad de datos, es multiplataforma e incluye varias funcionalidades tales como (30):

- ***Data Profiling****,* para determinar la calidad de los datos.
- ***Data Validator***, para validar datos contra reglas que deben verificarse bajo la política de calidad establecida.
- ***Comparator***, para comparar la información de diferentes fuentes de origen.
- ***Monitor***, para establecer un seguimiento de la calidad de los datos.

[4] XML: lenguaje de marcas extensible

- ***Dictionary***, permite crear un repositorio de datos correctos para comparar los datos a los cuales se les realizará el perfilado.

1.8 Tipos de almacenamiento de datos

El Procesamiento Analítico en Línea (OLAP, por sus siglas en inglés), es una herramienta muy utilizada a nivel mundial para el desarrollo de soluciones de inteligencia de negocios. Procesan las transacciones de tiempo real de un negocio, además de contener una estructura de datos optimizados para la introducción y la adición de datos (32). En la actualidad se usan principalmente tres tecnologías derivadas de los sistemas OLAP, estas son (33):

Procesamiento Analítico Relacional en Línea (ROLAP): almacena los datos en un motor relacional. Utiliza una arquitectura de tres niveles que permite el análisis de una enorme cantidad de datos:

- El nivel de base de datos usa bases de datos relacionales para el manejo, acceso y obtención de los datos.
- El nivel de aplicación es el motor que ejecuta las consultas multidimensionales de los usuarios.
- El nivel de presentación se integra con el motor relacional y permite a los usuarios realizar los análisis OLAP.

Procesamiento Analítico Multidimensional en Línea (MOLAP): Esta implementación OLAP almacena los datos de forma lógica en una base de datos multidimensional para optimizar los tiempos de respuesta. Provee excelente rendimiento, compresión de los datos, tiene un buen tiempo de respuesta, puede escribir sobre la base de datos y además permite realizar cálculos mucho más complicados.

Procesamiento Analítico Híbrido en Línea (HOLAP): es una solución que incluye las implementaciones anteriores (MOLAP y ROLAP). Las agregaciones de los datos son almacenadas en una estructura multidimensional usada por MOLAP y la base de datos fuente en una base de datos relacional. Los cubos almacenados como HOLAP son más pequeños que los MOLAP y responden más rápido que los ROLAP.

En la presente investigación se seleccionó para el desarrollo de los subsistemas de almacenamiento e integración la tecnología ROLAP, ya que se tuvo en cuenta que el SGBD PostgreSQL seleccionado solo soporta el almacenamiento relacional y no el multidimensional. En la actualidad los SGBD que dan soporte al almacenamiento multidimensional son propietarios, por lo que no están en correspondencia con las políticas de desarrollo de la UCI y el país.

Conclusiones

Luego de haber profundizado en los conceptos principales de los AD se puede determinar que estos constituyen una alternativa formidable en la gestión de grandes volúmenes de información haciendo uso de los MD para distribuir la información según las áreas temáticas de la empresa. Luego de analizados todos estos elementos se establecen varias pautas con el fin de desarrollar los subsistemas de almacenamiento e integración del producto ACM 14F7:

- La metodología a utilizar será la Metodología para el Desarrollo de Almacenes de Datos en DATEC ya que guía el proceso de construcción del sistema durante cada una de las etapas.
- Se seleccionará el Visual Paradigm for UML en su versión 8.0 como herramienta de modelado, ya que posibilita la generación de los diagramas necesarios para modelar el funcionamiento del sistema propuesto. También para el diseño e implementación del subsistema de almacenamiento se decide utilizar PostgreSQL 9.1 como SGBD y PgAdmin III 1.14.1 como administrador de base de datos. Para realizar los procesos de integración se utilizará el DataCleaner 1.5.4 para el perfilado y limpieza de los datos y el PDI 4.4.0 para la implementación de los procesos de ETL.
- Además se elige para el desarrollo de los subsistemas de almacenamiento e integración del producto ACM 14F7 la tecnología ROLAP como modo de almacenamiento ya que el SGBD PostgreSQL 9.1 soporta la administración de los grandes volúmenes de datos que utiliza el sistema.

Capítulo 2: Análisis y diseño de los subsistemas de almacenamiento e integración del producto ACM 14F7

En el capítulo se aborda la etapa de análisis que constituye la base para comprender los requisitos establecidos por el cliente y posteriormente obtener las reglas del negocio. Se realiza el diagrama de casos de uso del sistema y se define además la arquitectura en la que se basará la solución proponiéndose un diseño de la misma, con las características necesarias para satisfacer las necesidades manifestadas por el cliente. Se diseñan el subsistema de almacenamiento y el subsistema de integración a través de la definición de las dimensiones, hechos y medidas identificadas. También se construye la matriz buz y posteriormente re realiza el modelo de datos. Conjuntamente se establece el esquema de seguridad y se describe la política de respaldo y recuperación a utilizar, definiendo roles y permisos.

2.1 Análisis del negocio

Las necesidades de información son las especificaciones definidas por los especialistas del negocio para dar cumplimiento a determinadas tareas. Para lograr un diseño que se ajuste a las necesidades del cliente en este caso el CIM, es necesario realizar un estudio preliminar del negocio para identificarlas y a través de las mismas se podrán definir las funcionalidades que tendrá el sistema.

Para la identificación de las necesidades planteadas por el cliente pueden ser utilizadas diversas técnicas, entre ellas las entrevistas, cuestionarios y encuestas. En el análisis del negocio de los subsistemas de almacenamiento e integración del producto ACM 14F7 se seleccionó la técnica de entrevista, la misma fue utilizada para identificar las necesidades del cliente y además definir las reglas del negocio.

A partir del estudio del negocio se decide clasificar la información en nueve grupos:

- **Inclusión:** se almacena la información de inclusión de los pacientes al ensayo.
- **Evaluación inicial:** se almacena la información de la evaluación inicial realizada a los pacientes del ensayo.
- **Administración:** se almacena información referente a la administración del producto a los pacientes del ensayo.
- **Control de signos vitales:** se almacena la información de los signos vitales de los pacientes del ensayo.

- **Control de exámenes de laboratorio:** se almacena la información de los exámenes de laboratorio realizados a los pacientes del ensayo.
- **Interrupción del tratamiento:** se almacena la información de la interrupción del tratamiento de los pacientes del ensayo.
- **Eventos adversos:** se almacena la información de los eventos adversos presentados por los pacientes del ensayo.
- **Evaluación durante el estudio diagnóstico:** se almacena la información de la evaluación durante el estudio diagnóstico de los pacientes del ensayo.

2.2 Especificación de requisitos

Al proceso de identificación de las necesidades del cliente se le denomina levantamiento de requisitos. En la investigación se definirán tres tipos de requisitos: Requisitos de Información (RI), Requisitos Funcionales (RF) y Requisitos No Funcionales (RNF). El análisis de requisitos guía el proceso de desarrollo del sistema hacia la dirección correcta a partir de las necesidades definidas por el cliente.

2.2.1 Requisitos de Información

Después de realizar un análisis minucioso del negocio se definen los RI asociados al producto ACM 14F7. Estos fueron agrupados según el tipo de información. La especificación de estos requisitos puede encontrarse en el artefacto “DATEC_CIM_ACM 14F7_Especificacion_requisitos_SW.doc”.

Inclusión y evaluación inicial

RI_1: Almacenar la información de los pacientes por edad, fecha de inclusión, peso, tnm[5], estadio[6], estado según OMS[7], exámenes de laboratorio, localización del tumor y técnica de análisis empleada en los modelos de inclusión y evaluación inicial de la aplicación del producto ACM 14F7 en los pacientes con cáncer de mama.

[5] tnm: tamaño del tumor, número de ganglios y si el paciente tiene metástasis o no

[6] estadio: en términos oncológicos se utiliza para designar la etapa o fase de la enfermedad

[7] OMS: Organización Mundial de Salud

RI_2: Almacenar la información de los pacientes con consentimiento informado para participar en la investigación en los modelos de inclusión y evaluación de la aplicación del producto ACM 14F7 en los pacientes con cáncer de mama.

RI_3: Almacenar la información de los pacientes con carcinoma de mama estadio IV o enfermedad metástasis evolutiva en los modelos de inclusión y evaluación inicial de la aplicación del producto ACM 14F7 en los pacientes con cáncer de mama.

RI_4: Almacenar la información de los pacientes con edades entre 18 y 80 años en los modelos de inclusión y evaluación inicial de la aplicación del producto ACM 14F7 en los pacientes con cáncer de mama.

RI_5 Almacenar la información de los pacientes con estado general menor o igual a 2 según los criterios de la OMS en los modelos de inclusión y evaluación inicial de la aplicación del producto ACM 14F7 en los pacientes con cáncer de mama.

RI_6: Almacenar la información de los pacientes con función renal conservada con parámetros de creatinina entre los límites normales entre 35-132 mmol/l en los modelos de inclusión y evaluación inicial de la aplicación del producto ACM 14F7 en los pacientes con cáncer de mama.

RI_7: Almacenar la información de los pacientes con valores de hemoglobina a partir 10g/l, leucocitos, mayor que 4000 /mm3, plaquetas 100 x 109, transaminasa y fosfatasa alcalina hasta dos veces y media por encima de los valores de referencias normales en los modelos de inclusión y evaluación inicial de la aplicación del producto ACM 14F7 en los pacientes con cáncer de mama.

Administración 14F7

RI_8 Almacenar la información de los pacientes por fecha de administración y localización del tumor en el modelo administración 14F7 del producto ACM 14F7 en los pacientes con cáncer de mama.

RI_9: Almacenar la información de los pacientes que presentaron eventos adversos en el modelo administración 14F7 del producto ACM 14F7 en los pacientes con cáncer de mama.

RI_10 Almacenar la información de los pacientes que interrumpieron el estudio en el modelo administración 14F7 del producto ACM 14F7 en los pacientes con cáncer de mama.

Control de signos vitales

RI_11 Almacenar la información de los pacientes por tipo de signos vitales y hora en que fueron tomados en el modelo control de signos vitales del producto ACM 14F7 en los pacientes con cáncer de mama.

Control de exámenes de laboratorio

RI_12: Almacenar la información de los pacientes por exámenes de laboratorio y día en que se realizaron en el modelo control de exámenes de laboratorio del producto ACM 14F7 en los pacientes con cáncer de mama.

Interrupción del tratamiento

RI_13: Almacenar la información de los pacientes por fecha de interrupción y causas de interrupción del tratamiento en el modelo interrupción del tratamiento del producto ACM 14F7 en los pacientes con cáncer de mama.

Eventos adversos

RI_14: Almacenar la información de los pacientes por tipo, grado, causalidad, resultado y fecha en que ocurrió el evento adverso en el modelo reacciones adversas con el producto ACM 14F7 en los pacientes con cáncer de mama.

Evaluación durante el estudio diagnóstico

RI_15: Almacenar la información de los pacientes por fecha de realización de la evaluación y localización del tumor en el modelo evaluación durante el estudio diagnóstico del producto ACM 14F7 en los pacientes con cáncer de mama.

RI_16: Almacenar la información de los pacientes que presentaron eventos adversos en el modelo evaluación durante el estudio diagnóstico del producto ACM 14F7 en los pacientes con cáncer de mama.

RI_17: Almacenar la información de los pacientes que interrumpieron el tratamiento en el modelo evaluación durante el estudio diagnóstico del producto ACM 14F7 en los pacientes con cáncer de mama.

2.2.2 Requisitos funcionales

Los requisitos funcionales describen las funcionalidades o capacidades que debe presentar el sistema para dar cumplimiento a los RI anteriormente descritos. En la investigación se realiza solo el análisis de los subsistemas de almacenamiento e integración de datos por lo cual se determinarán específicamente las funcionalidades asociadas a los mismos.

A continuación se mencionan los RF identificados, la especificación de estos requisitos se encuentra en el artefacto“DATEC_CIM_ACM14F7_Especificacion_requisitos_SW.doc”.

RF1: Realizar la extracción de los datos.

RF2: Realizar la transformación y carga de los datos.

2.2.3 Requisitos no funcionales

Los requisitos no funcionales representan las propiedades o cualidades de la solución. Estos son agrupados por categorías en dependencia de las características del negocio. A continuación se exponen algunos RNF definidos, los demás pueden encontrarse en el artefacto “DATEC_CIM_ACM 14F7_Especificacion_requisitos_SW.doc”.

Confiabilidad

RNF1: Garantizar la disponibilidad del sistema en el tiempo requerido. El sistema debe estar en un servidor disponible durante el horario de trabajo, de 8:00 am a 5:00 pm.

RNF2: Garantizar la persistencia de la información. Para garantizar la persistencia de la información se realizará un respaldo total de los datos del almacén de datos y además se guardará la información en algún dispositivo de almacenamiento.

Eficiencia

RNF3: Lograr la uniformidad de la estructura de los elementos definidos en los subsistemas de almacenamiento e integración del producto ACM14F7. Las estructuras definidas deben estar estandarizadas para permitir un mejor entendimiento entre los desarrolladores del sistema.

Soporte

RNF4: Proporcionar características mínimas de hardware a los servidores.

Para cumplir con este RNF se deben contar con los siguientes requerimientos de hardware:

- 512 MB RAM o superior.
- Pentium 4.
- Almacenamiento en disco 1GB.

Restricciones de diseño

RNF5: Instalar en las estaciones de trabajo el software necesario para el correcto funcionamiento del sistema.

Para un correcto funcionamiento del sistema la aplicación debe contar con:

- *Java Virtual Machine (JVM)*[8] versión 6.6 o superior.
- *Pentaho Data Integration* versión 4.4.0 o superior, en caso de que un usuario capacitado requiera realizar la carga de los datos nuevamente.

RNF6: Utilizar todas las tecnologías definidas en la investigación. Visual Paradigm para UML en su versión 8.0, PostgreSQL 9.1 como SGBD, PgAdmin III 1.14.1 como administrador de base de datos, DataCleaner 1.5.4 para el perfilado y limpieza de los datos y el PDI 4.4.0 para la implementación del proceso de ETL.

2.3 Reglas del negocio

Las reglas del negocio (RN) describen políticas que deben cumplirse o condiciones que deben satisfacer al cliente. Estas son identificadas a través del análisis del negocio y los resultados del perfilado de datos. La metodología utilizada clasifica según categorías las reglas del negocio, las mismas se presentan a continuación:

Reglas de variables

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

RN2. Los valores que indiquen cantidad tienen que ser mayores o igual que cero.

Reglas de almacenamiento

RN3. El evento adverso será de tipo cadena con una longitud máxima de 60 caracteres y se almacenará como: 1 para Fiebre, 2 para Náuseas, 3 para Vómitos, 4 para Disnea, 5 para Dolor precordial ,6 para Otro.

RN4. Los valores del grado del evento adverso según la OMS serán: 0 para Normal, 1 para Ligero, 2 para Moderado, 3 para Severo y 4 para Muy severo.

RN5. Los valores de causalidad del evento adverso serán: 1 Definitiva, 2 Muy probable, 3 Probable, 4 Posible, 5 No relacionado y 6 Desconocido.

[8] Java Virtual Machine (JVM): Máquina virtual de java (MVJ).

RN6. Los resultados del evento adverso serán: 1 para Efecto reversible, 2 para Efecto irreversible y 3 para Muerte.

RN7. Las fechas serán de tipo date en formato (aaaa/mm/dd).

Reglas de transformación

RN8. Cuando el resultado de los exámenes de laboratorio aparezca nulo se pondrá No Especificado.

RN9. Cuando el valor del estado según la OMS aparezca nulo se pondrá Desconocido.

RN10. Cuando el valor del estadio aparezca nulo se pondrá Desconocido.

RN11. Cuando el valor del grado, causalidad y resultado del evento adverso esté vacío se pondrá Desconocido.

2.4 Casos de Uso del Sistema

Los casos de uso representan las diversas interacciones entre el actor y el sistema, en respuesta a un evento que inicia el actor sobre el propio sistema. Los RI y RF identificados son agrupados en Casos de Usos de Información (CUI) y Casos de Usos Funcionales (CUF) respectivamente, los cuales aportan una especificación detallada del funcionamiento del sistema en cuestión.

2.4.1 Actores del sistema

Administrador de ETL: se encarga de realizar los procesos de extracción, transformación y carga de los datos del sistema fuente.

Analista: es el responsable de analizar, consultar y mantener disponible la información contenida en los diferentes modelos.

2.4.2 Casos de uso de información

Los CUI permiten agrupar varios RI a través de criterios predefinidos. En la presente investigación se seleccionaron 7 CUI que agrupan los RI identificados por el tipo de información.

1. **CUI_1. Almacenar información de los modelos inclusión y evaluación inicial:** agrupa los RI pertenecientes a los modelos de inclusión y evaluación inicial.
2. **CUI_2. Almacenar información del modelo administración 14F7:** agrupa los RI pertenecientes al modelo administración 14F7.

3. **CUI_3.Almacenar información del modelo control de signos vitales:** agrupa los RI pertenecientes al modelo control de signos vitales.
4. **CUI_4. Almacenar información del modelo control de exámenes de laboratorio:** agrupa los RI pertenecientes al modelo control de exámenes de laboratorio.
5. **CUI_5. Almacenar información del modelo interrupción del tratamiento:** agrupa los RI pertenecientes al modelo interrupción del tratamiento.
6. **CUI_6. Almacenar información del modelo eventos adversos:** agrupa los RI pertenecientes al modelo de interrupción del tratamiento.
7. **CUI_7. Almacenar información del modelo evaluación del estudio diagnóstico:** agrupa los RI pertenecientes al modelo evaluación del estudio diagnóstico.

La especificación de los CUI puede encontrarse en el artefacto "DATEC_CIM_ACM14F7_Especificacion_CU.doc" dentro de Expediente de Proyecto de los Subsistemas de almacenamiento e integración del producto ACM 14F7.

2.4.3 Casos de uso funcionales

Los CUF agrupan los RF definidos para cada uno de los subsistemas que componen la solución. A continuación se describen los 2 CUF identificados en la investigación:

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

En la siguiente tabla se muestra la especificación de los CUF:

Tabla 1: Especificación del CUF_1: Extraer los datos de la fuente

Objetivo	Realizar la extracción de los datos.
Actores	Administrador ETL

Resumen	El CU inicia cuando el actor selecciona los datos a extraer. Se extraen los mismos de la fuente de datos. Finaliza cuando los datos seleccionados por el actor son extraídos.
Complejidad	Media
Prioridad	Media
Precondiciones	Disponibilidad de las fuentes.
Pos condiciones	Los datos de la fuente correspondiente han sido extraídos y almacenados en la base de datos.

Flujo de eventos		
	Actor	**Sistema**
1.	El administrador de ETL realiza la conexión a la fuente correspondiente.	2. Responde a la solicitud de conexión a la fuente.
3.	El administrador de ETL selecciona el archivo a extraer.	
4.	El administrador de ETL realiza la extracción de los datos.	5. Ejecuta la extracción de los datos. Finaliza el caso de uso.

Flujos alternos		
	Acción del Actor	**Respuesta del Sistema**
		2.1. No responde a solicitud de conexión. 2.2. Notifica el error al administrador de ETL. Vuelve al paso 1 del Flujo Normal de Eventos.
	3.1. Si hay control de cambios, el administrador de ETL verifica si hay modificaciones. • En caso afirmativo ir al paso 3 del flujo normal. • En caso negativo ir al paso 2 del flujo normal.	

Tabla 2: Especificación del CUF_2: Transformar y cargar los datos de la fuente

Objetivo	Realizar la transformación y carga de los datos
Actores	Administrador ETL

Resumen	El CU inicia cuando el actor desea realizar la transformación y carga de los datos. Finaliza cuando los datos son insertados satisfactoriamente en la base de datos.
Complejidad	Media
Prioridad	Media
Precondiciones	Los datos se encontraron correctamente extraídos de la fuente y las estructuras de los subsistemas de almacenamiento e integración se encontraron disponibles para su uso. En la base de datos debe existir una estructura para almacenar la información.
Pos condiciones	Los datos han sido transformados y cargados satisfactoriamente.

Flujo de eventos		
	Acción del Actor	**Respuesta del Sistema**
1.	El administrador de ETL selecciona las estructuras de la fuente que desea transformar.	
2.	El administrador de ETL carga los datos seleccionados en memoria.	
3.	El administrador de ETL aplica las transformaciones pertinentes y genera datos de auditoría.	
4.	El administrador de ETL carga los datos en la base de datos	5. Ejecuta la consulta. Finaliza el caso de uso.
Flujos Alternos		
	Acción del Actor	**Respuesta del Sistema**
		3.3 El sistema muestra un mensaje de error y regresa al paso 3.

2.4.4 Diagrama de casos de uso del sistema

Los diagramas de Casos de Uso del Sistema (CUS) describen el comportamiento del sistema a través de la representación de actores, CU y sus relaciones. El diagrama de CUS de la presente investigación está constituido por 2 CUF y 7 CUI, representándose en total 9 CUS. A continuación se muestra el diagrama de CUS obtenido:

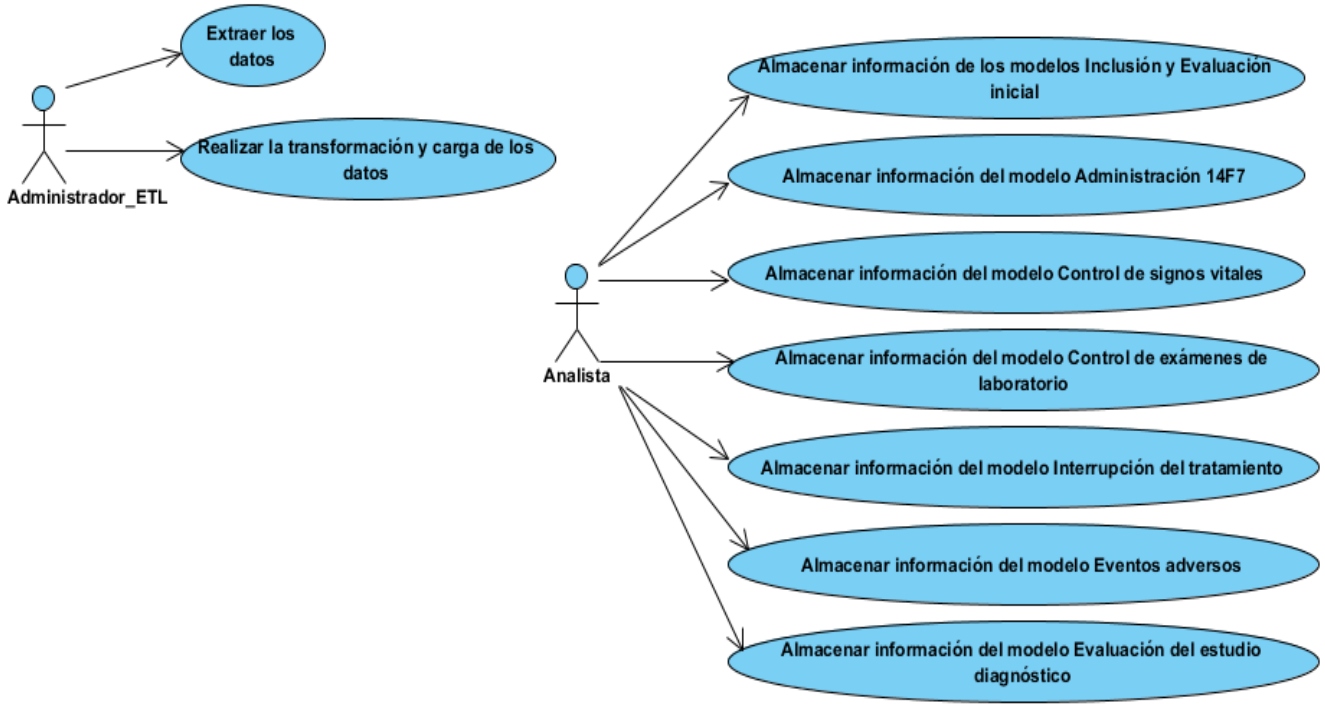


Fig 1. Diagrama de casos de uso del sistema

2.5 Arquitectura base de los subsistemas de almacenamiento e integración del producto ACM 14F7

Los MD presentan una arquitectura compuesta por la fuente de datos y tres subsistemas (almacenamiento, integración y visualización). Como se había mencionado en el Capítulo 1 el subsistema de visualización no formará parte de la solución, por lo que la misma solo contendrá en su arquitectura base la fuente de datos, el subsistema de almacenamiento y el subsistema de integración.

El subsistema de integración se encarga de extraer los datos de la fuente de datos en este caso ficheros Excel; limpiarla, estandarizarla e integrarla, dejándola lista para su posterior almacenamiento. Por su parte el subsistema de almacenamiento guarda todos los datos distribuyéndolos en tablas de hechos y dimensiones.

A continuación se muestra la arquitectura definida para el desarrollo de la solución:

Fig 2. Arquitectura de los subsistemas de almacenamiento e integración del producto ACM 14F7

La fuente de datos está compuesta por ficheros Excel, de los cuales se extraen los datos del subsistema de integración a través de la herramienta PDI, que es la encargada de estandarizarlos y transformarlos dejándolos listos para luego ser almacenados en la base de datos. Por su parte, el subsistema de almacenamiento recibe los datos integrados durante los procesos de ETL y los almacena en un esquema representado por hechos y dimensiones soportadas por el SGBD PostgreSQL y administradas solo por los usuarios autorizados mediante la herramienta PgAdminIII.

2.6 Diseño de la solución

Para realizar el diseño de la solución debe incluirse una arquitectura que soporte los requisitos funcionales y no funcionales. Es necesario también realizar el perfilado de los datos y posteriormente diseñar los subsistemas de almacenamiento e integración estableciendo esquemas y políticas de seguridad.

2.6.1 Diseño del subsistema de almacenamiento

Para realizar el diseño del subsistema de almacenamiento es fundamental la identificación de los hechos, medidas y dimensiones que permitan crear el modelo de datos. Es importante también definir una política de respaldo y recuperación que garantice la integridad de los datos almacenados.

Hechos

Los hechos como se había mencionado anteriormente serán los objetos de análisis de un negocio, por lo que los mismos serán definidos a través de los CUI identificados.

H1. Hecho inclusión y evaluación inicial (**hecho_inclus_y_eval_inicial**): contiene la información de los pacientes incluidos y evaluados inicialmente en el EC. Está relacionado al CUI_1. Almacenar información de los modelos inclusión y evaluación inicial

H2. Hecho administración 14F7 (**hecho_administración**): contiene la información de los pacientes a los cuales fue suministrado el producto ACM 14F7. Está relacionado al CUI_2. Almacenar información del modelo administración 14F7.

H3. Hecho control de signos vitales (**hecho_control_signos_vitales**): contiene la información de los signos vitales tomados a los pacientes del ensayo. Está relacionado al CUI_3. Almacenar información del modelo control de signos vitales.

H4. Hecho control de exámenes de laboratorio (**hecho_examenes_laboratorio**): contiene la información de los exámenes de laboratorio realizados a los pacientes del ensayo. Está relacionado al CUI_4. Almacenar información del modelo control de exámenes de laboratorio.

H5. Hecho interrupción del tratamiento (**hecho_interrupcion_tratamiento**): contiene la información de los pacientes del ensayo que interrumpieron su tratamiento. Está relacionado al CUI_5. Almacenar información del modelo interrupción del tratamiento.

H6. Hecho eventos adversos (**hecho_eventos_adversos**): contiene la información de los pacientes del ensayo que sufrieron eventos adversos. Está relacionado al CUI_6. Almacenar información del modelo eventos adversos.

H7. Hecho evaluación del estudio diagnóstico (**hecho_evaluacion_diagnostico**): contiene la información de los pacientes del ensayo a los cuales se les realizó la evaluación del estudio

diagnóstico. Está relacionado al CUI_7. Almacenar información del modelo evaluación del estudio diagnóstico.

Dimensiones

Una vez identificados los hechos deben determinarse las dimensiones y medidas asociadas a estos. Para ello se analizan las variables de entrada y de salida de cada uno de los RI que han sido agrupados en un CUI específico, de forma tal que las variables de entrada de estos requisitos pasan a ser las dimensiones y las variables de salida las medidas asociadas al hecho.

En la investigación se identificaron 25 dimensiones, de estas 8 son dimensiones degeneradas ya que toman pocos valores y estos no cambian con el tiempo. Este término se refiere al campo que será utilizado como criterio de análisis y que es almacenado en una tabla de hechos, en vez de ser definido como una dimensión. Un ejemplo de este tipo de dimensión sería: “ocurrió evento adverso”, donde la única información que se obtiene es sí o no. Por tanto se podría plantear la opción de simplemente incluir estos campos en una tabla de dimensión, pero en este caso se estaría manteniendo una fila de esta dimensión por cada fila en la tabla de hechos, por consiguiente se tendría la duplicación de información y complejidad, que precisamente es lo que se pretende evitar.

A continuación se describen las dimensiones identificadas:

1. **Dimensión edad (dim_edad):** describe la edad del paciente.
2. **Dimensión peso (dim_peso):** se refiere al peso corporal del paciente.
3. **Dimensión estadio (dim_estadio):** especifica la gravedad de la enfermedad del paciente al ser diagnosticado.
4. **Dimensión tnm (dim_tnm):** especifica el tamaño del tumor, número de ganglios y si el paciente presenta metástasis o no.
5. **Dimensión estado funcional según la Organización Mundial de la Salud (OMS) (dim_estado_oms):** se refiere al estado funcional según la OMS del paciente (Ver Anexo 1).
6. **Dimensión técnica de análisis (dim_tecnica_analisis):** se refiere a la técnica de análisis realizada al paciente, la misma puede ser biopsia o citología.

7. **Dimensión tiempo (dim_tiempo):** describe todos los espacios de tiempo en que se aplicó el ensayo.
8. **Dimensión causas de la interrupción del tratamiento (dim_causas_interrupcion):** describe las causas por las cuales el paciente interrumpió el tratamiento.
9. **Dimensión exámenes de laboratorio (dim_examen_lab):** especifica los resultados de los exámenes de laboratorio realizados a los pacientes del ensayo.
10. **Dimensión localización del tumor (dim_localizacion_tumor):** especifica la localización de la metástasis del paciente.
11. **Dimensión tad (dim_sv_tad):** especifica la tensión arterial mínima del paciente.
12. **Dimensión tas (dim_sv_tas):** especifica la tensión arterial máxima del paciente.
13. **Dimensión pulso (dim_sv_pulso):** especifica el pulso del paciente.
14. **Dimensión signos vitales (dim_sv_temperatura):** especifica la temperatura corporal del paciente.
15. **Dimensión período del examen (dim_periodo_examen):** describe el período de tiempo en que fue realizado un examen de laboratorio
16. **Dimensión evento adverso (dim_evento_adverso):** describe el tipo, grado, causalidad y resultado del evento adverso presentado por el paciente.
17. **Dimensión hora (dim_hora):** describe la hora en que fueron tomados los signos vitales del paciente.

Dimensiones degeneradas:

1. **Dimensión paciente con carcinoma o metástasis estadio IV (pac_carcinomaIV_metastasis):** especifica si el paciente tiene carcinoma o metástasis con estadio IV.
2. **Dimensión paciente con estado general menor o igual a 2 (pac_estado_general_menor_2):** especifica si el paciente tiene estado general menor o igual a 2.
3. **Dimensión paciente con función renal conservada (pac_func_renal_conservada):** especifica si el paciente tiene función renal conservada.

4. **Dimensión paciente con edad entre 18 y 80 (pac_entre_18_y_80):** especifica si el paciente tiene edad entre 18 y 80 años.
5. **Dimensión paciente con hemoglobina mayor que 10 (pac_valor_hemoglobina):** especifica si el paciente presenta un valor de hemoglobina mayor que 10.
6. **Dimensión paciente con consentimiento informado (pac_consentimiento_informado):** especifica si el paciente dio su consentimiento para participar en el ensayo.
7. **Dimensión ocurrió evento adverso (ocurrio_evento_adverso):** especifica si el paciente sufrió eventos adversos.
8. **Dimensión interrumpió el estudio (interrumpio_estudio):** especifica si el paciente interrumpió el estudio.

Dimensiones Lentamente Cambiantes (SCD[9])

Las dimensiones pueden ser de dos tipos estáticas o cambiantes, las estáticas son aquellas cuya información no está propensa a cambios con el tiempo, mientras que las cambiantes son aquellas cuya información está propensa a cambios. Existen varios tipos de SCD las cuales permiten o no mantener en el almacén un registro histórico de los valores asociados a un identificador del sistema operacional. En sus inicios Ralph Kimball definió solo tres tipos: 1, 2 y 3; pero a través de los años han ido agregándose otros tipos de SCD al profundizarse su estudio y entendimiento. Los nuevos tipos incluidos fueron 0, 4 y 6, a continuación se describen cada uno estos (34):

- **Tipo 0 (no tiene en cuenta la gestión histórica):** no se tiene en cuenta la gestión de los cambios históricos por lo tanto nunca se cambia la información ni se sobrescribe.
- **Tipo 1 (sobrescribir):** no se guardan datos históricos, la nueva información sobrescribe siempre la antigua y la sobrescritura se realiza principalmente por errores de calidad de datos.
- **Tipo 2 (añadir fila):** toda la información se guarda en el almacén de datos. Cuando hay un cambio se crea una nueva entrada con la fecha y la llave subrogada propia y a partir de ese momento se usará este valor para futuras entradas y las antiguas usarán el valor anterior.

[9] **SCD:** *Slowly Changing Dimensions*

- **Tipo 3 (añadir columna):** toda la información se guarda en el almacén de datos. En este caso se crean nuevas columnas con los valores antiguos y los actuales son reemplazados con los nuevos.
- **Tipo 4 (tabla de historia separada):** es lo que se conoce habitualmente como tablas históricas. Existe una tabla con los datos actuales y otra con los antiguos o los cambios.
- **Tipo 6 (híbrido):** el método combina los tipos 1, 2 y 3; y se le denomina tipo 6 debido a la suma de los tres tipos que integra (1+2+3=6). Además, se añade una pareja adicional de columnas para indicar el rango de fechas al cual aplica cada fila en particular.

En la presente investigación se seleccionó el tipo 1 de SCD debido a que aunque la carga que se realizará es histórica y no incremental, una vez cargados los datos estos pueden ser sobrescritos debido a errores de calidad.

Matriz bus o matriz dimensional

La matriz bus se realiza para validar el correcto diseño del modelo de datos y permite además determinar el impacto que provocaría un cambio en la solución durante el desarrollo del sistema. Esta representa la relación entre un hecho y una dimensión de forma tal que las columnas de la matriz contienen los hechos y las filas las dimensiones. La intersección entre una fila y una columna especifica una relación entre un hecho y una dimensión. La matriz bus de los subsistemas de almacenamiento e integración del producto ACM 14F7 se muestra a continuación:

Tabla 3. Matriz bus

Dimensiones/hechos	**H1**	**H2**	**H3**	**H4**	**H5**	**H6**	**H7**
tiempo	x	x		x	x		x
edad	x						
peso	x						
estado _oms	x						
estadio	x						

tnm	x						
tecnica_analisis	x						
localiz_tumor	x	x					x
examen_lab	x			x			
hora		x	x				
causa_interrupcion					x		
evento_adverso						x	
sv_tad		x					
sv_tas		x					
sv_temperatura		x					
sv_pulso		x					
pac_entre_18_y_80	x					x	
pac_carcinomaIV_metastasis	x						
pac_estado_general_menor_2	x						
pac_func_renal_conservada	x						
pac_valor_hemoglobina	x						
pac_consentimiento_informado	x						

periodo_examen				x			
ocurrio_evento_adverso		x					x
interrumpio_estudio		x					x

Al realizar la matriz bus se determinó que existen dimensiones compartidas por varios hechos, sin embargo no existen hechos que se relacionen exactamente con las mismas dimensiones. Esto significa que queda descartado el solapamiento entre hechos.

Topologías

Una vez confeccionada la matriz bus y analizadas las relaciones existentes entre hechos y dimensiones debe determinarse la topología a utilizar para representar el modelo de datos. Estas pueden ser de tres tipos: esquema en estrella, constelación de hechos o copo de nieve. Estas tres topologías se muestran a continuación (6):

Esquema estrella: en este esquema existe un único elemento central (tabla de hechos) conectado radialmente con las tablas de dimensiones. Cada tupla de la tabla de hechos incluye las medidas consideradas y una referencia a cada dimensión.

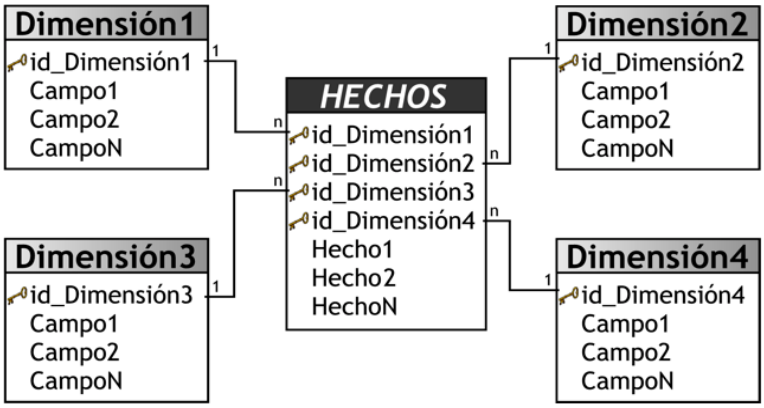


Fig 3. Vista esquema estrella.

Esquema copo de nieve: se deriva del esquema estrella, además las tablas de dimensiones se ramifican en más puntas y es capaz de soportar la jerarquía.

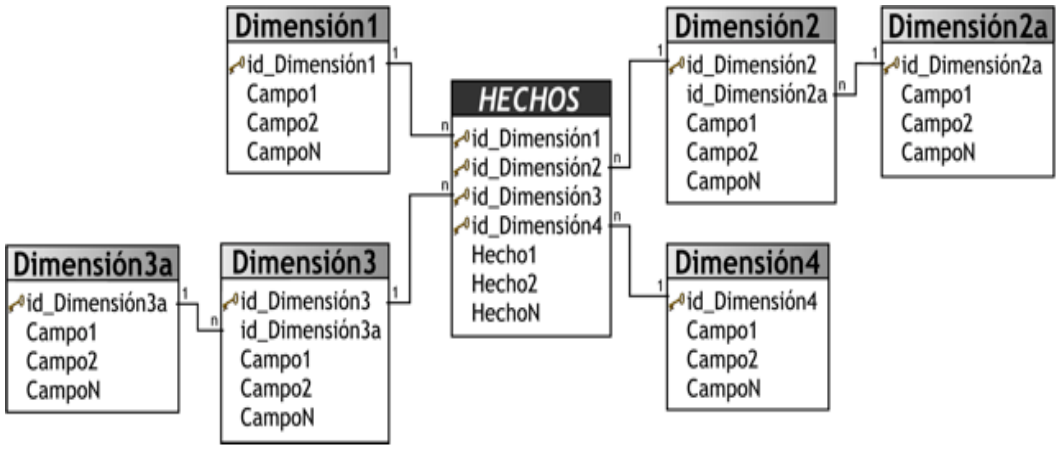


Fig 4. Vista esquema copo de nieve.

Esquema constelación de hechos: está compuesto por un conjunto de tablas de hechos que comparten algunas tablas de dimensiones. Comúnmente se observa esta representación como varios esquemas de estrella integrados.

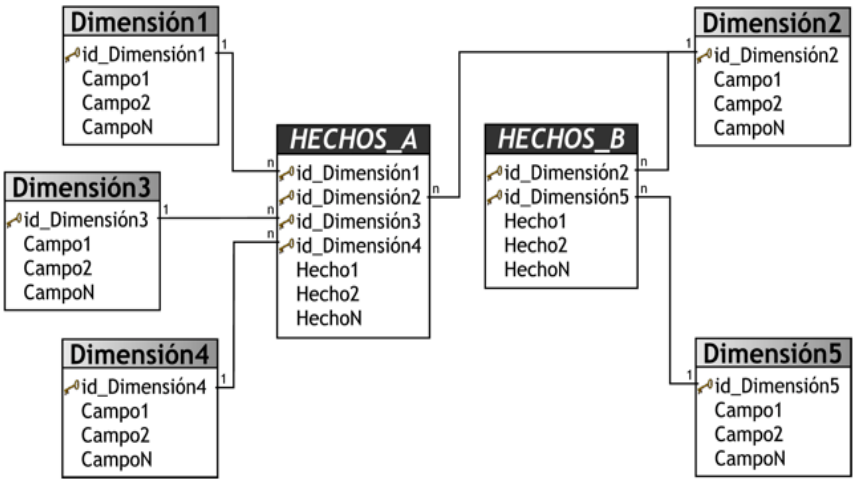


Fig 5. Vista esquema constelación de hechos.

Modelo de datos

Una vez definidos los hechos, medidas y dimensiones del negocio se procede a la creación del modelo de datos. A continuación se presenta un fragmento del modelo de datos diseñado donde se utiliza una topología constelación de hechos teniendo en cuenta que las tablas de hechos comparten algunas tablas de dimensiones, cuya reutilización optimiza el diseño del modelo de datos. Para una mejor comprensión

del mismo puede consultar el artefacto “Especificación del modelo de datos.doc” dentro del Expediente de Proyecto de los Subsistemas de almacenamiento e integración del producto ACM 14F7 y ver el Anexo 2.

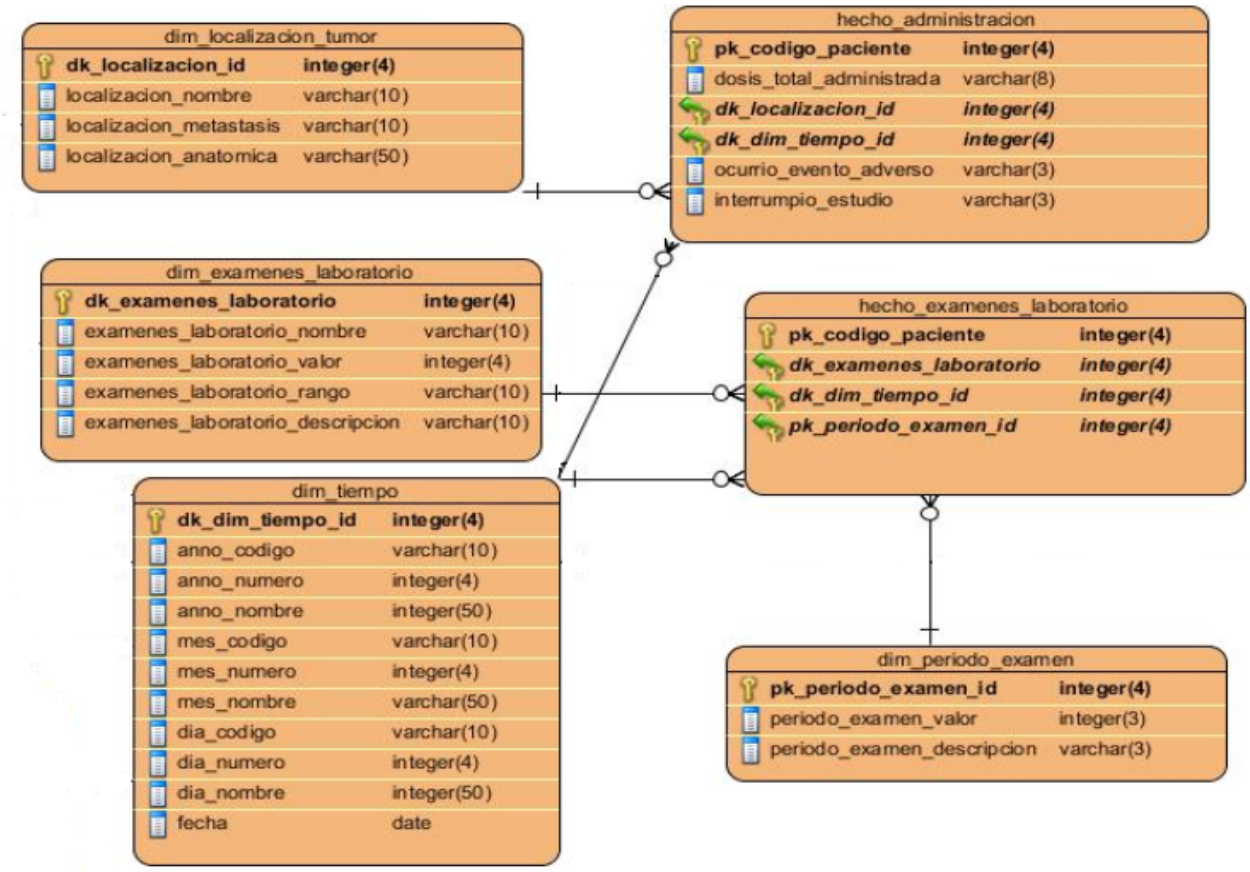


Fig 6. Fragmento del modelo de datos

2.6.2 Diseño del subsistema de integración

El diseño del subsistema de integración incluye el perfilado de datos, el diseño de las transformaciones y la extracción de los datos de la fuente. A través de estos elementos se obtienen un conjunto de datos listos para su correcto almacenamiento.

Perfilado de datos a la fuente de datos

El perfilado de datos puede describirse como un análisis de la fuente de datos que permite comprender su contenido, estructura y calidad. A través de este análisis se estudia la existencia de errores en los datos como pudieran ser los valores nulos, distintos y duplicados. Posteriormente se crean nuevas reglas del

negocio y transformación que contribuyan a la utilidad y aplicabilidad de los datos en el área del negocio donde se trabaja.

Al realizar el perfilado de datos a las fuentes se determinó que existen 4 tipos de datos*, integer, varchar, double y date* de los cuales predominan los tipos *integer y varchar*. El siguiente gráfico representa el comportamiento de los tipos de datos en la fuente:

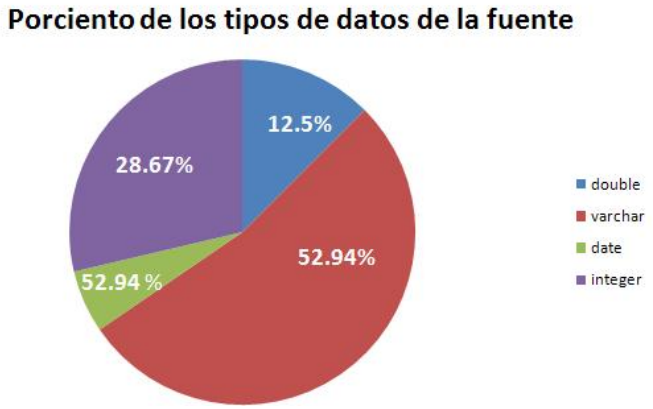


Fig 7. Gráfico de comportamiento de los tipos de datos en la fuente

También se analizó la calidad de los datos revelándose la no existencia de valores negativos o duplicados, que de un total de 136 campos hay 74 de ellos con valores nulos. Todo el análisis realizado a los datos se encuentra completamente en el artefacto:" DATEC_CIM_ACM 14F7_Perfilado de los Datos.doc".

Diccionario de datos

El diccionario de datos recoge la información de cada una de las variables contenidas en la fuente de datos, especificando el significado que tienen dentro del negocio y el rango de valores que pueden tomar. En el artefacto "DATEC_CIM_ACM 14F7_Diccionario de Datos.xls" dentro de Expediente de Proyecto de los Subsistemas de almacenamiento e integración del producto ACM 14F7 puede encontrarse íntegramente la especificación de cada una de las variables y los posibles valores que pueden tomar.

Diseño general de las transformaciones

Después de haber realizado el perfilado a la fuente de datos y analizado su contenido, estructura y calidad, se procede al diseño general de las transformaciones. Los diseños pueden variar al ser implementados ya que de existir problemas con los datos deben aplicarse diversas estrategias para solucionarlos. El diseño general de las transformaciones establece una serie de pasos que permiten a los

datos ser cargados en las tablas de la base de datos. En la figura 8 se muestra el diseño general de las transformaciones para la carga de dimensiones de los subsistemas de almacenamiento e integración del producto ACM 14F7. Como primer paso se establecen las variables de entorno que permitirán la conexión con la base de datos datamart_ACM14F7, luego se prosigue con la búsqueda de los ficheros fuentes, verificando la existencia de los mismos y en caso de no existir se genera un Excel de error. Si los ficheros fuentes existen se extraen los datos de los mismos. Posteriormente se verifica la existencia de valores nulos y cuando no existan se insertan directamente los datos en la base de datos. En caso de existir valores nulos se aplican estrategias de ETL y se prosigue a insertar los datos hacia la base de datos. Después se obtiene la información del sistema para generar los metadatos.

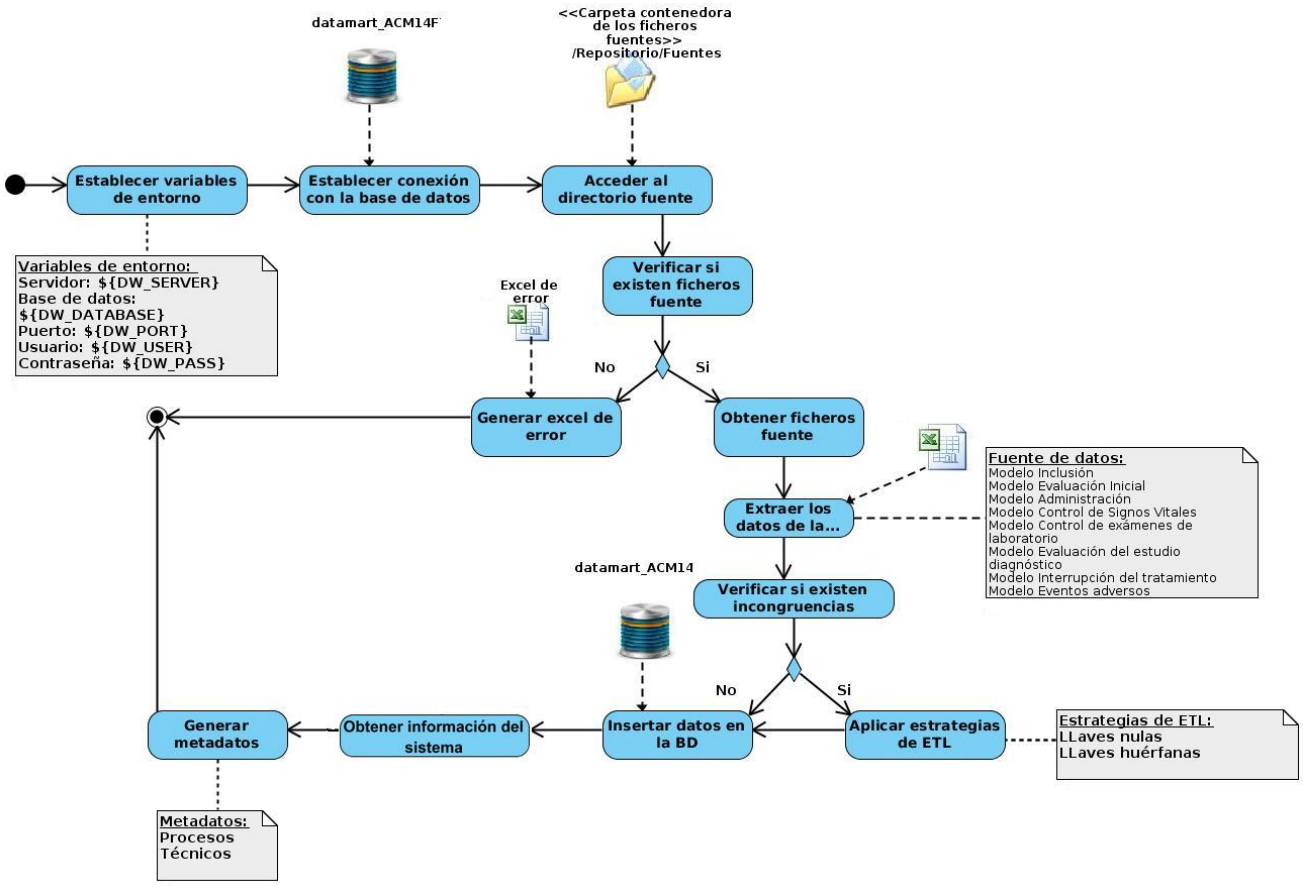


Fig 8. Diseño general de las transformaciones para la carga de dimensiones

El diseño general para la carga de los hechos de los subsistemas de almacenamiento e integración del producto ACM 14F7 es el siguiente: se realiza como primer paso el establecimiento de las variables de

entorno que permitirán la conexión con la base de datos datamart_ACM14F7, luego se prosigue con la búsqueda de los ficheros fuentes, donde se verifica la existencia de los mismos y en caso de no existir se genera un Excel de error. Una vez confirmada la existencia de los ficheros se extraen los datos que serán cargados para su posterior análisis. Se verifica la existencia de valores nulos y en caso de existir se realizan estrategias para corregirlos. Cuando no existan valores nulos se verifica la existencia de llaves nulas, si existen se transforman las mismas para realizar a través de ellas la búsqueda en la dimensión. En caso de no existir llaves nulas, se buscan las llaves dimensionales directamente y se insertan en la tabla correspondiente al esquema mart_cim. Luego se obtiene la información del sistema para generar los metadatos. La figura 9 muestra el diseño para la carga de los hechos:

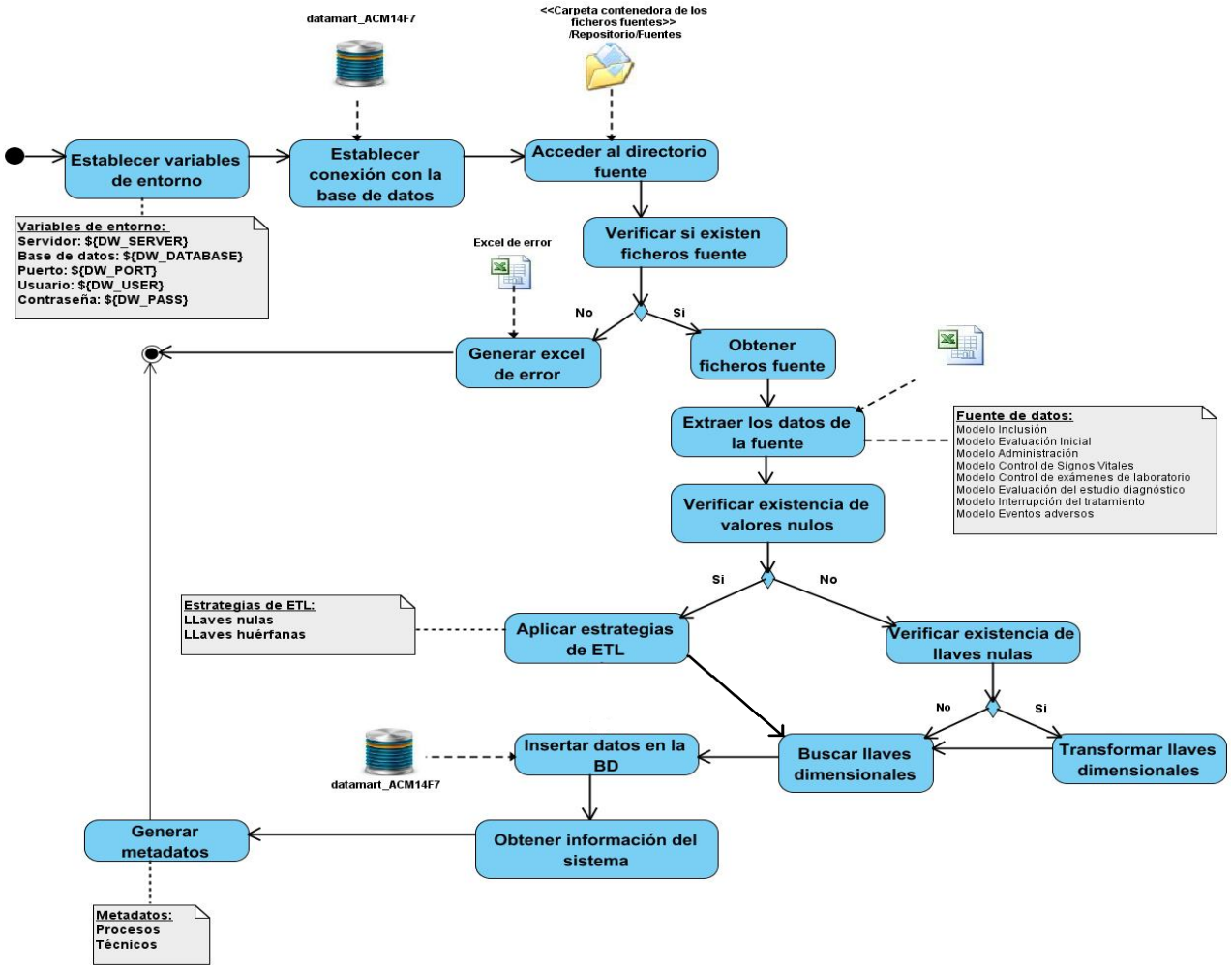


Fig 9. Diseño general de las transformaciones para la carga de hechos

2.7 Política de respaldo y recuperación

Para garantizar la persistencia de la información se establece una política de seguridad basada en las copias de seguridad. La misma contribuirá a evitar las consecuencias tanto monetarias como prestigiosas que podría desencadenar la pérdida de la información para una entidad. Debido a que el sistema posee una carga histórica, que solo se cargará una vez y no posee carga incremental, se realizará una copia de respaldo de toda la información almacenada en la base de datos, incluyendo las transformaciones y trabajos realizados con el objetivo de contar con un método de recuperación anti contingencias. También se propone salvar la información en algún dispositivo de almacenamiento, ya que es de vital importancia mantener su disponibilidad y seguridad en todo momento.

2.8 Esquema de seguridad

Es fundamental que un sistema cuente con un mecanismo de protección contra aquellas violaciones de seguridad que atenten contra la integridad, confidencialidad y disponibilidad de la información almacenada. Es por esto que para garantizar la seguridad en el subsistema de almacenamiento se definen los siguientes roles:

Tabla 4. Roles y permisos de acceso a la base de datos.

Roles	**Permisos**
Administrador de bases de datos	Realiza la administración de la base de datos relacional. Posee todos los permisos de administración y otorga permisos a los diferentes usuarios.
Administrador de ETL	Realiza los procesos de ETL sobre los datos y tiene permisos de lectura y escritura sobre los esquemas pertenecientes a los subsistemas de almacenamiento e integración del producto ACM 14F7.

La seguridad en el subsistema de integración se garantiza a nivel de sistema operativo, ya que permite asignar permisos de lectura y/o escritura a los archivos para determinados usuarios. De esta forma, puede restringirse el acceso de un archivo específico a uno o varios usuarios. Con esta medida se logra proteger

las carpetas que contienen las transformaciones y trabajos que posibilitan el desarrollo de los procesos de integración del producto.

Conclusiones

Una vez abordados los elementos básicos del análisis y diseño de los subsistemas de almacenamiento e integración del producto ACM 14F7 se puede concluir que:

- Fueron identificados 17 requisitos de información, 2 requisitos funcionales, 6 requisitos no funcionales y 11 reglas del negocio las cuales han sido aplicadas durante el diseño de los subsistemas de almacenamiento e integración.
- A través del diseño del modelo de datos y utilizando una topología constelación de hechos fueron identificadas 7 tablas de hechos y 17 tablas de dimensiones que permitieron la creación de la estructura física de almacenamiento y posteriormente el correcto funcionamiento del sistema.
- Una vez realizado el perfilado de datos se determinó el estado y la calidad de la fuente de datos.
- El diseño de las transformaciones proporcionó un acercamiento a las acciones que deben realizarse sobre los datos para lograr su estandarización y correcto almacenamiento.
- Las políticas de respaldo y los esquemas de seguridad definidos contribuyen a mantener la seguridad e integridad de los datos almacenados.

Capítulo 3: Implementación y prueba de los subsistemas de almacenamiento e integración del producto ACM 14F7

El capítulo está dirigido a la implementación de los subsistemas de almacenamiento e integración del producto ACM 14F7. El epígrafe tiene como propósito brindar una mayor comprensión de las estrategias y procedimientos utilizados. La implementación del subsistema de almacenamiento incluye aspectos como los estándares de codificación utilizados y la construcción del modelo físico, mientras que el subsistema de integración incluye la implementación de trabajos, transformaciones y metadatos. También se realiza la validación de los subsistemas de almacenamiento e integración a través de pruebas como las listas de chequeo y los casos de prueba. Las comprobaciones realizadas permiten encontrar y corregir los errores existentes, contribuyendo a lograr una aplicación con mayor calidad.

3.1 Implementación del subsistema de almacenamiento

El proceso de implementación del subsistema de almacenamiento incluye estándares de codificación de las estructuras, para facilitar la comprensión de los nombres definidos en cada uno de los esquemas. Una vez definidos estos estándares se procede a desarrollar la estructura física de almacenamiento.

3.1.1 Estándares de codificación

Con los estándares de codificación se propone el establecimiento de un patrón que permita normalizar los términos utilizados, fomentando un mejor entendimiento de las estructuras por los desarrolladores del sistema.

En la solución se propone que los nombres de las tablas de dimensiones sean antecedidos por el prefijo “dim”, y el caracter “_”, quedando como sigue: dim_edad. Para las tablas de hechos se define el prefijo “hech”, seguido por el carácter “_” y luego por el nombre de la tabla, ejemplo: hech_evento_adverso. Las llaves primarias de las dimensiones se nombran de la forma “dk_dim_dimension_id”.

El nombre de las transformaciones comienzan con las letras “trans”, luego el caracter especial “_” y finalmente el nombre de la misma, ejemplo trans_dim_edad. Igualmente se nombraron los trabajos, se anteponen las letras “trab” seguido del caracter “_” y luego el nombre de este, ejemplo trab_general. Por su parte los metadatos están conformados por las letras “md” y seguido por el caracter “_”, y luego su nombre, ejemplo md_carga_historica.

Una vez terminado el proceso de estandarización de los nombres, se encuentra organizada la nomenclatura utilizada para las tablas y atributos dentro de la base de datos y de las estructuras del subsistema de integración, por lo que se procede a la implementación de las estructuras físicas.

3.1.2 Implementación del modelo de datos físico

El modelo de datos físico representa un conjunto de entidades que describen las estructuras presentes en los datos. Este modelo es generado a partir del modelo lógico de datos, el cual ha sido representado en el capítulo anterior. En la base de datos se encuentran los datos organizados en estructuras lógicas que facilitan la correcta manipulación de los mismos. Estas estructuras son denominadas esquemas y tablas.

La base de datos “datamart_ACM14F7” cuenta con 27 tablas, de las cuales 17 son de dimensiones, 7 son de hechos y las restantes son de metadatos. A continuación se mencionan los esquemas utilizados en la solución:

- El esquema “mart_cim” recoge las tablas de hechos y las dimensiones propias de la solución.
- El esquema “metadatos” recoge las trazas de la ejecución de las transformaciones y trabajos, además contiene información de la carga histórica y los errores.

3.2 Implementación del subsistema de integración

Para llevar a cabo la implementación del subsistema de integración se deben realizar todos los procesos de ETL sobre los datos .Para ejecutar de forma correcta estos procesos se analizará a detalle la fuente, identificando los principales problemas existentes.

Una vez estudiada la fuente de datos se prosigue a extraer la información contenida en ella. Después se obtienen los campos relevantes para los subsistemas de almacenamiento e integración del producto ACM 14F7 teniendo en cuenta el modelo de datos realizado y las necesidades de información establecidas por el cliente. Luego se realiza la transformación y limpieza de los datos, corrigiéndose errores como los datos incorrectos y las entradas duplicadas. Finalmente se procede a la carga, donde se toman los datos transformados y se insertan en la base de datos. Es importante analizar el uso de los subsistemas propuestos por Kimball para lograr el correcto funcionamiento de los procesos de integración. A continuación se mencionan y describen los subsistemas utilizados en la implementación de la aplicación.

- **Perfilado de datos:** a través del perfilado se verificó la calidad de los datos de la fuente y se definieron nuevas reglas de transformación para solucionar problemas en los mismos.

- **Sistema de extracción:** permitió la extracción de los datos para luego ser transformados y cargados en la base de datos.
- **Subsistema de transformación:** este subsistema contribuyó a realizar el mapeo de valores, el cambio de los tipos de datos en algunos campos, la búsqueda de información en flujos de datos y el filtrado de valores.
- **Subsistema de carga:** permitió realizar la carga de los datos a las tablas de dimensiones y hechos de los subsistemas de almacenamiento e integración del producto ACM 14F7.
- **Llave subrogada:** posibilitó crear claves subrogadas independientes para cada tabla.
- **SCD:** implementó la lógica para crear atributos de variabilidad lenta a lo largo del tiempo. En la presente investigación se utilizó el tipo 1 de SCD para poder sobrescribir datos con errores de calidad.
- **Sistema de backup:** se realizaron copias de respaldo de los procesos ETL.
- **Seguridad:** se gestionó el acceso a los procesos de ETL y metadatos.
- **Repositorio de metadatos:** se capturaron los metadatos de los procesos de ETL y de los aspectos técnicos.
- **Programador de trabajos:** permitió gestionar los trabajos, los cuales se encargan de la ejecución de las transformaciones en un orden específico y atendiendo a la periodicidad definida para la carga de la información.

3.2.1 Implementación de las transformaciones

Las transformaciones se implementan a través de pasos, los mismos están conectados entre sí mediante saltos. Estos últimos representan el elemento más simple dentro de las transformaciones y es a través de ellos que viaja la información entre los distintos pasos de forma tal que la salida de un paso constituye la entrada del otro. Por su parte los pasos se agrupan por categorías dependiendo de la función que desempeñan dentro de la trasformación. Para cada paso existe una ventana específica donde se configuran los elementos sobre los cuales se trabajará y el comportamiento que se espera de los mismos.

A continuación se presentan dos ejemplos de las transformaciones realizadas para las dimensiones y los hechos respectivamente:

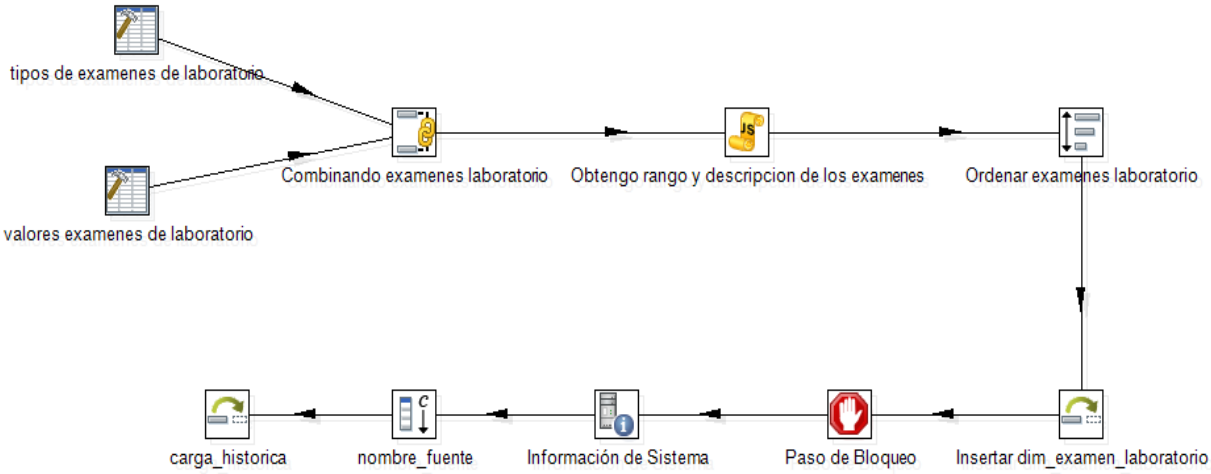


Fig 10. Transformación para cargar la dimensión exámenes de laboratorio.

En la figura 10 se hace referencia a la carga de la dimensión exámenes de laboratorio, donde en primer lugar se definen los nombres de los diferentes exámenes y los resultados de cada uno de ellos. Luego se combinan los exámenes con los resultados para obtener todas las combinaciones posibles. Además se obtiene la descripción y rango de los resultados y se ordenan. Una vez realizadas las modificaciones sobre los datos se procede a insertarlos en la base de datos. Posteriormente se obtiene la información del sistema que va a permitir la generación de los metadatos los cuales serán insertados también en la base de datos.

La figura 11 muestra un ejemplo de la carga del hecho exámenes de laboratorio donde luego de extraer los datos de la fuente se obtienen los campos necesarios y se le aplican a estos un conjunto de transformaciones como la normalización de las filas, la unión de flujos y la eliminación de valores nulos. Después se buscan los identificadores de cada dimensión asociada al hecho y se procede a su inserción en la base de datos. Seguidamente se obtiene la información del sistema, se generan los metadatos, y se insertan de igual forma en la base de datos.

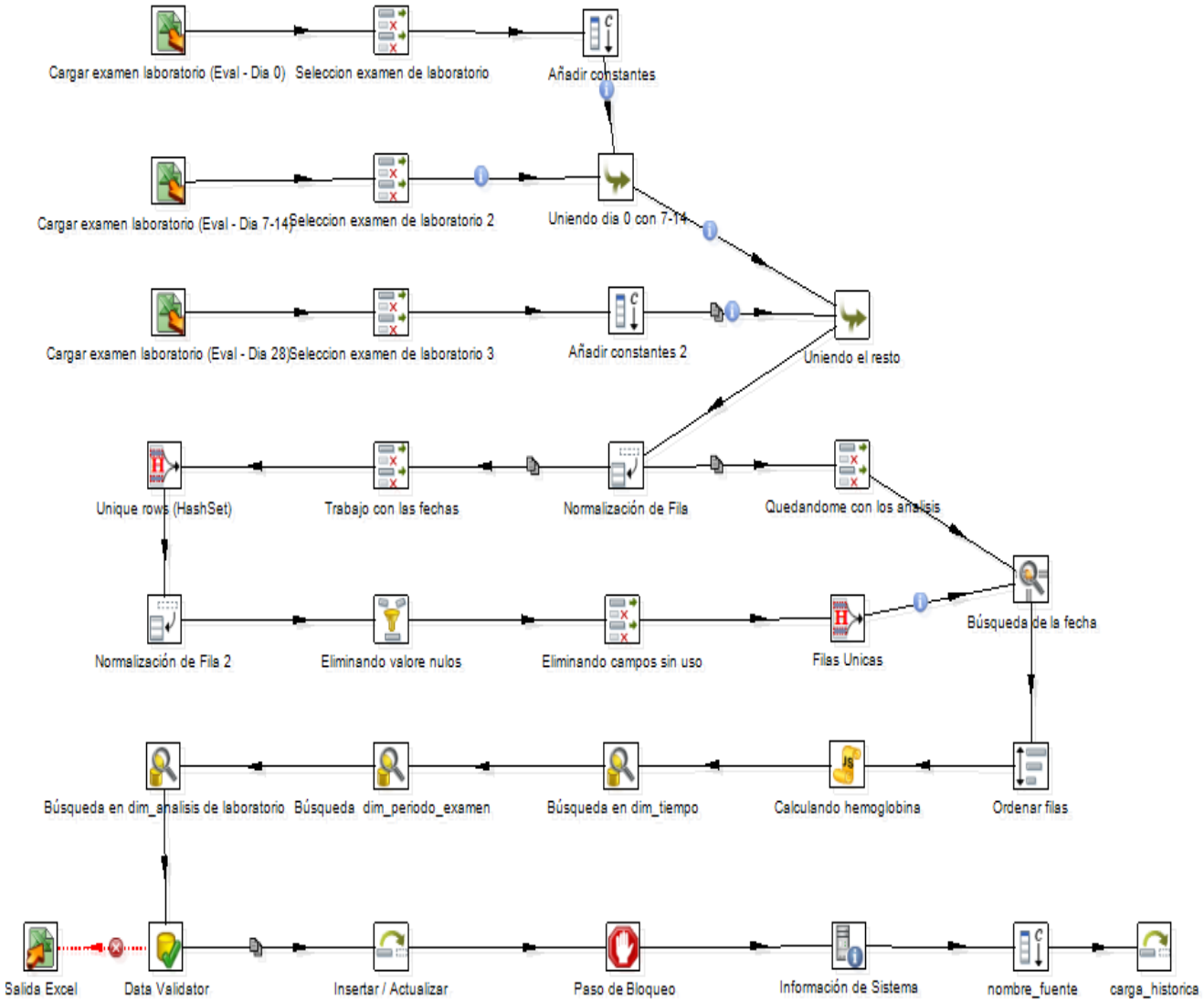


Fig 11.Transformación para cargar el hecho exámenes de laboratorio.

3.2.2 Implementación de los trabajos

Los trabajos en el contexto de los procesos de integración de datos son un conjunto de tareas que se realizan con el objetivo de ejecutar una acción determinada. Estos permiten ejecutar varias transformaciones o trabajos previamente diseñados, definiendo una secuencia lógica para su ejecución, mediante el uso de pasos definidos, que no están disponibles en las transformaciones.

En la figura 12 se muestra el trabajo general para cargar los hechos y las dimensiones. El proceso comienza con el establecimiento de la conexión y su validación. En el caso de que existan fallos en la conexión se termina el trabajo, sino se procede a la carga de las dimensiones y posteriormente a la de los hechos.

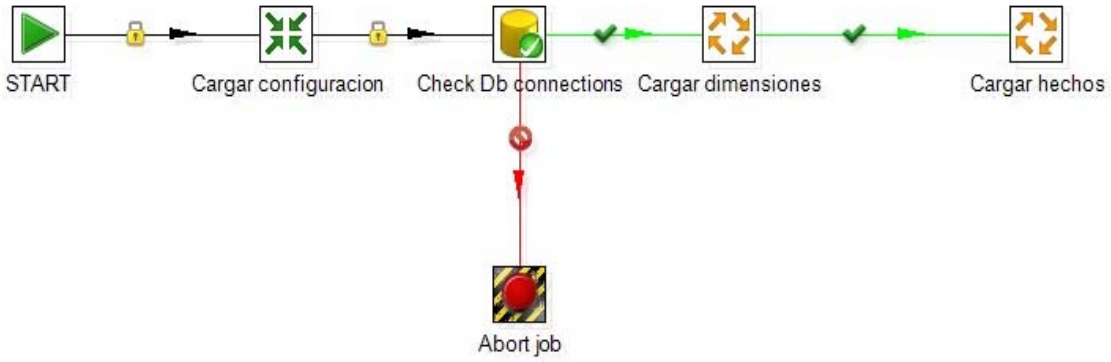


Fig 12. Trabajo general

3.2.3 Gestión de los metadatos

Los metadatos son datos escritos sobre los propios datos que ayudan a su vez a identificar, describir y localizar recursos digitales. Constituyen una información estructurada que describe y/o permite encontrar, gestionar, controlar y entender o preservar otra información (35). Los metadatos pueden agruparse en varias categorías (36):

- 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 se interrelacionan sus componentes.
- Metadatos de proceso: permiten obtener información de los procesos en que se ejecutan.
- Metadatos de negocio: posibilita obtener los datos y la información referente a los aspectos del negocio.

En la investigación se utilizaron los metadatos de proceso para obtener la información correspondiente a los procesos de las transformaciones y los trabajos. También fueron usados los metadatos técnicos para mostrar los resultados de la integración de datos y así poder guardar la información del nombre de la fuente, el nombre del destino, la fecha de ejecución y el nombre de la transformación.

3.3 Aplicación de pruebas

Las pruebas constituyen un elemento clave para verificar la calidad del software, ya que estas permiten identificar errores o deficiencias en el producto desarrollado. Existe una amplia gama de pruebas que pueden ser aplicadas a los subsistemas de almacenamiento e integración del producto ACM 14F7 que están enfocadas en el objetivo de obtener un producto con calidad.

Se evaluó la calidad del producto a través de pruebas definidas por el Centro Nacional de Calidad de Software (CALISOFT), las cuales son llevadas a la práctica por DATEC con el fin de crear un estándar para comprobar que el producto cumpla con las especificaciones del negocio. Las pruebas realizadas se muestran a continuación.

3.3.1 Pruebas Unitarias

Las pruebas unitarias se enfocan en un programa o un componente que desempeña una función específica, que puede ser probada y se asegura que funcione tal y como lo define la especificación del programa. Estas son realizadas a cada uno de los subsistemas de la solución por separado, para verificar su correcto funcionamiento. Las pruebas unitarias fueron realizadas al subsistema de almacenamiento y al subsistema de integración de forma independiente. Estas se realizaron en dos iteraciones, las No Conformidades (NC) detectadas en cada subsistema, en la primera iteración se muestran a continuación.

Subsistema de almacenamiento:

NC_1: Existen RN que no están clasificadas correctamente.

NC_2: Las RN de la categoría visualización no formarán parte de la solución.

NC_3: Existen problemas en la identificación de los niveles y las jerarquías de las dimensiones.

Subsistema de integración:

NC_4: No ha sido confeccionado el artefacto registro de sistemas fuente**.**

NC_5: En el diseño general de las transformaciones se deben especificar los nombres de los ficheros a cargar, además del nombre de la base de datos.

NC_6: Realizar ajustes en algunos componentes de manera que permitan optimizar las transformaciones.

Las NC identificadas en la primera iteración fueron resueltas inmediatamente por lo que en la segunda iteración realizada no se encontraron NC:

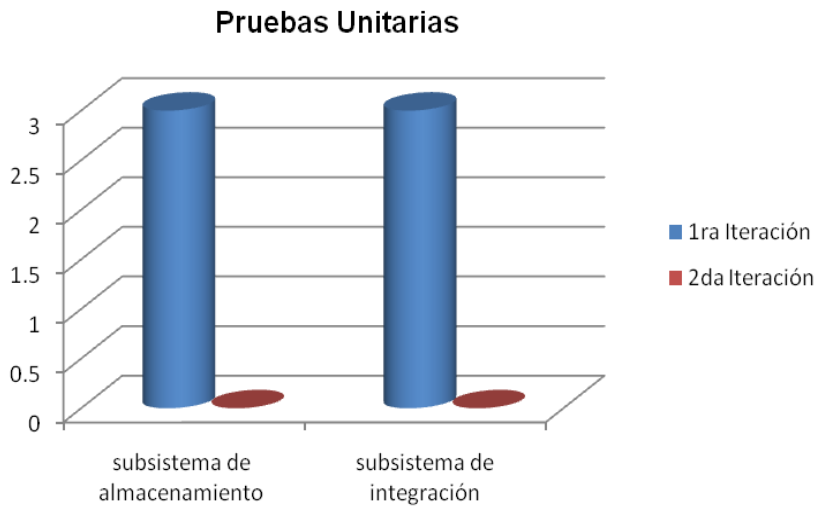


Fig. 13: Realización de las pruebas unitarias.

3.3.2 Pruebas de Integración

Las pruebas de integración se encargan de identificar errores introducidos por la combinación de programas o componentes probados unitariamente con anterioridad. Este tipo de pruebas también verifican que las especificaciones de diseño sean alcanzadas y que la integración entre los subsistemas de la solución sea correcta.

En la presente investigación se realizaron un conjunto de consultas a la base de datos y se determinó que sus respuestas coincidían con los valores de los datos provenientes de la fuente. Esto reflejó que todos los datos fueron cargados completamente y de forma correcta.

La estrategia definida para realizar las pruebas de integración incluye la confección de casos de prueba que resultan de gran ayuda para demostrar el correcto funcionamiento del sistema.

3.4 Herramientas para la aplicación de las pruebas

Para la aplicación de las pruebas fueron utilizadas las herramientas casos de pruebas y listas de chequeo las cuales se abordan a continuación.

3.4.1 Casos de prueba

Los casos de prueba están diseñados para determinar el grado de cumplimiento de los requisitos de una aplicación y permiten identificar posibles fallos de implementación. Son la base para diseñar y ejecutar los procedimientos de pruebas ya que muestran una secuencia ordenada de eventos al describir flujos básicos, flujos alternos, precondiciones y postcondiciones. Si los casos de prueba no son correctos, la calidad del sistema se pone en duda y las pruebas dejan de ser confiables.

En los subsistemas de almacenamiento e integración del producto ACM 14F7 fueron diseñados 7 casos de prueba en correspondencia a cada uno de los CUI identificados, los mismos serán recogidos de forma íntegra en el artefacto: “DATEC_CIM_ACM14F7_Casos_de_prueba_de_integracion.xls”. Para llevar a cabo un caso de prueba se realiza una consulta en lenguaje SQL a la base de datos. A continuación se muestra la consulta correspondiente al caso de prueba asociado al CUI Almacenar información del modelo administración 14F7, refiriéndose a la cantidad de pacientes que recibieron dos dosis del producto, presentaron eventos adversos y tienen una cantidad de metástasis igual a uno:

```
SELECT
  COUNT(DISTINCT hech_administracion.pk_codigo_paciente)
FROM
  public.hech_administracion,
  public.dim_localizacion_tumor
WHERE
  dim_localizacion_tumor.dk_dim_localizacion_id = hech_administracion.dk_dim_localizacion_id
  AND hech_administracion.dosis_total_administrada ='2' AND hech_administracion.ocurrio_evento_adverso ='Presenta'
  AND dim_localizacion_tumor.localizacion_cant_metastasis='1';
```

Fig 14. Caso de prueba basado en el CUI Almacenar información del modelo administración 14F7.

Una vez ejecutada la consulta se observó que la respuesta obtenida coincidía con los datos provenientes de la fuente, por lo que el resultado de la prueba es satisfactorio. De la misma manera se procedió con el caso de prueba relacionado al CUI Almacenar información del modelo control de exámenes de laboratorio obteniéndose también un resultado satisfactorio, el cual se muestra a continuación:

```
SELECT
  COUNT (DISTINCT hech_examenes_laboratorio.pk_codigo_paciente)
FROM
  public.dim_periodo_examen,
  public.dim_examenes_laboratorio,
  public.hech_examenes_laboratorio,
  public.dim_tiempo
WHERE
  dim_periodo_examen.dk_dim_periodo_examen_id = hech_examenes_laboratorio.dk_dim_periodo_examen_id AND
  dim_examenes_laboratorio.dk_dim_examenes_laboratorio = hech_examenes_laboratorio.dk_dim_examenes_laboratorio AND
  dim_tiempo.dk_dim_tiempo_id = hech_examenes_laboratorio.dk_dim_tiempo_id AND
  dim_periodo_examen.periodo_examen_valor = 28
  AND dim_examenes_laboratorio.examenes_laboratorio_nombre ='Hemoglobina'
  AND dim_examenes_laboratorio.examenes_laboratorio_descripcion ='Bajo';
```

Fig 15.Caso de prueba basado en el CUI Almacenar información del modelo control de exámenes de laboratorio.

Además se definieron 8 casos de prueba los cuales están relacionados con las reglas de transformación, con la finalidad de comprobar si luego de ejecutada cada una de las transformaciones, los datos cargados en la base de datos son los esperados. A continuación se muestra el caso de prueba para la regla de transformación: Cuando el resultado de los exámenes de laboratorio aparezca nulo se pondrá No Especificado.

Tabla 5. Caso de prueba para una regla de transformación

Nombre variable	examen_laboratorio					
Escenario	hech_examenes_laboratorio					
Regla de transformación	Debe ser numérico, entero y definido por un rango					
Valor de entrada	Estado del dato	Resultado esperado	Respuesta del Sistema	Resultado Real	Comentario	Resultado de la Prueba
	No válido	No especificado	El sistema transforma el dato aplicando la regla de transformación.	No especificado	La prueba se realizó sin ocurrir errores	Satisfactorio
11.5	Válido	11.5	El sistema comprueba que el valor se encuentre en un rango y lo adiciona satisfactoriamente	11.5	La prueba se realizó sin ocurrir errores	Satisfactorio

3.4.2 Listas de chequeo

Las listas de chequeo verifican a través de una serie de preguntas, en forma de cuestionario, el grado de cumplimiento de determinadas reglas establecidas para los procesos de desarrollo del sistema. También permiten medir la calidad de los artefactos de los procesos de ETL generados durante la realización del

producto. En estas se analizan principalmente varias pautas como la estructura del documento, la semántica y los indicadores a evaluar durante la etapa de desarrollo. Las listas de chequeo están compuestas por los siguientes elementos:

Peso: define si el indicador a evaluar es crítico o no. El mismo se describe con una C si es crítico.

Indicadores a evaluar: los indicadores a evaluar son: Estructura del documento, Semántica del documento e Indicadores definidos específicamente para el artefacto.

Evaluación: es la forma de evaluar el indicador en cuestión. El mismo se evalúa de uno en caso de que exista alguna dificultad sobre el indicador y de cero, en caso de que el indicador revisado no presente problemas.

N.P. (No Procede): se usa para especificar que no es necesario evaluar el indicador en ese caso.

Cantidad de elementos afectados (CEA): especifica la cantidad de errores encontrados sobre el mismo indicador.

Comentario: especifica los señalamientos o sugerencias que quiera incluir la persona que aplica la lista de chequeo. Pueden o no existir señalamientos o sugerencias.

En la presente investigación se le aplicará las listas de chequeo a los artefactos que se generan en los procesos de ETL (“Registro del Sistema Fuente (RSF)”, “Perfilado de Datos (PD)”, “Diccionario de Datos (DD)” y “Mapa Lógico de Datos (MLD)”), ya que estos constituyen la base de la solución.

A continuación se muestra una tabla que recoge la cantidad de indicadores evaluados en los diferentes artefactos a los que fueron aplicadas las listas de chequeo. Además se muestra el total de indicadores evaluados, los indicadores críticos y las NC encontradas en cada uno.

Tabla 6. Aplicación de las listas de chequeo a los artefactos de ETL.

Secciones	**RSF**	**PD**	**DD**	**MLD**
estructura	9	8	9	5
indicadores	1	1	1	1
semántica	3	3	3	3

total de indicadores	13	12	13	9
indicadores críticos	5	5	5	5
no conformidades	1	1	1	1

En la figura siguiente se muestra un gráfico que refleja el comportamiento de los indicadores definidos en la aplicación de las listas de chequeo a los 4 artefactos de ETL mencionados anteriormente.

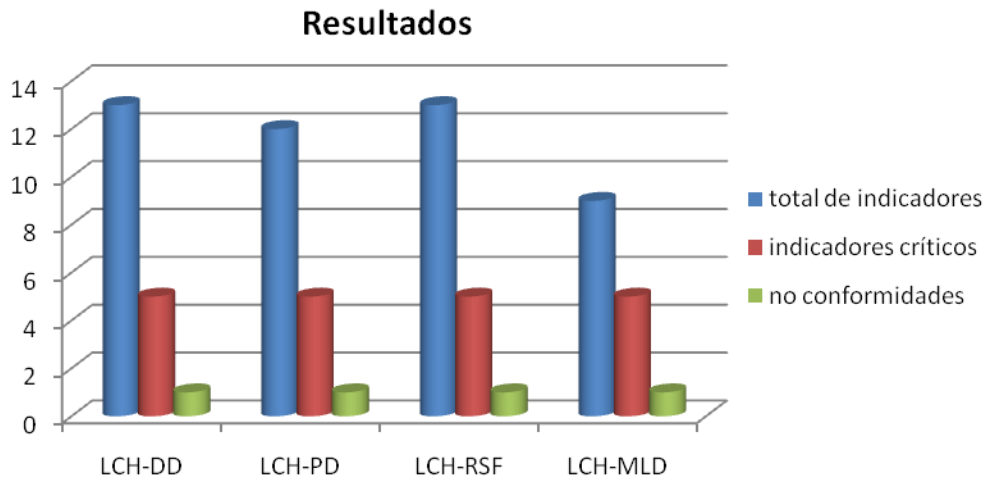


Fig 16. Aplicación de las listas de chequeo

3.5 Calidad de datos

El proceso de calidad de datos es muy importante, ya que permite comprobar que la información cargada no posea errores. Una vez concluido el proceso de ETL se realiza el perfilado de datos nuevamente, pero esta vez a la base de datos para corroborar que la información almacenada cumpla con la calidad requerida.

3.5.1 Perfilado de datos a la base de datos datamart_ACM14F7

El proceso de perfilado de datos permitió realizar un inventario sobre la información almacenada y así determinar la existencia de errores como valores duplicados, nulos o fuera de rango. Con el uso de la herramienta DataCleaner se obtuvieron resultados satisfactorios, quedando evidenciado que la carga de

todos los datos se realizó correctamente. El siguiente gráfico muestra los resultados obtenidos mediante el perfilado de datos realizado a la base de datos datamart_ACM14F7:

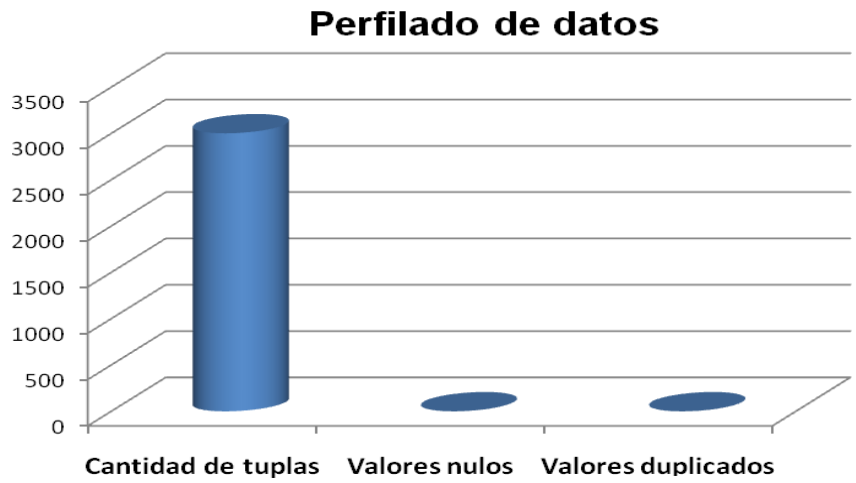


Fig 17. Resultado del perfilado de datos realizado a la base de datos datamart_ACM14F7.

Conclusiones

Luego de haber realizado la implementación y prueba de los subsistemas de almacenamiento e integración del producto ACM 14F7 puede concluirse que:

- Fueron implementados completamente los subsistemas que componen la solución: subsistema de almacenamiento y subsistema de integración.
- La estructura de los subsistemas contará con dos esquemas: mart_cim y metadatos lo cuales permitirán la correcta integración y almacenamiento de los datos.
- Además se realizaron las transformaciones pertinentes en los procesos de ETL obteniéndose 7 transformaciones para los hechos, 17 para las dimensiones y 3 para los trabajos, lo cual permitió eliminar las inconsistencias existentes y el almacenamiento de toda la información necesaria.
- La realización de pruebas de unitarias, de integración y de calidad contribuyó a la verificación de la calidad del producto final, corrigiéndose exitosamente las deficiencias encontradas.

Conclusiones generales

Una vez concluida la investigación puede afirmarse que se dio cumplimiento al objetivo general, completándose cada una de las tareas definidas. El estudio de los principales conceptos relacionados con los AD y los MD proporcionó la base para la elaboración del presente trabajo, cuyo resultado fue los "Subsistemas de almacenamiento e integración del producto ACM 14F7 para los ensayos clínicos del CIM". A partir del logro del resultado se arribó a las siguientes conclusiones:

- El análisis realizado de las metodologías utilizadas en el desarrollo de los subsistemas de almacenamiento e integración, garantizó que la Metodología para el Desarrollo de Almacenes de Datos en DATEC guiara el proceso de desarrollo a través de cada etapa del ciclo de vida, permitiendo la documentación de cada una de ellas.
- Las herramientas Visual Paradigm for UML en su versión 8.0, PostgreSQL 9.1 como SGBD, PgAdmin III 1.14.1 como administrador de base de datos, DataCleaner 1.5.4 para el perfilado y limpieza de los datos y el PDI 4.4.0 para la implementación del proceso de ETL contribuyeron al diseño, integración y almacenamiento de los subsistemas implementados.
- La tecnología de almacenamiento de datos ROLAP contribuyó a la construcción de la solución propuesta.
- Los requisitos y las reglas del negocio identificadas en el análisis y diseño de la solución dieron respuesta a las necesidades del cliente y posibilitaron la confección de los artefactos necesarios para la posterior etapa de implementación.
- La implementación de los esquemas y las transformaciones realizadas a los datos posibilitaron la integración de los mismos y su posterior almacenamiento.
- Las pruebas unitarias, de integración y de calidad efectuadas permitieron comprobar la funcionalidad del sistema y la calidad del producto a partir de los requisitos establecidos, obteniéndose resultados satisfactorios en cada una de ellas.

Recomendaciones

En la investigación se recomienda aplicar técnicas de minería de datos a la base de datos de los EC del producto ACM 14F7, que permitan detectar patrones repetitivos, tendencias o reglas que expliquen el comportamiento de la información almacenada.

Referencias Bibliográficas

1. CENTRO DE INMUNOLOGÍA MOLECULAR. Información General [en línea] (2013). Disponible en: <http://www.cim.sld.cu/>

2. CHAN, Dra. Margaret. Cáncer. Organización Mundial de la Salud [en línea] (2013). Disponible en: <http://www.who.int/topics/cancer/es/>

3. GONZÁLEZ HERNÁNDEZ, Marlien y DE LAOSA SOCARRÁS, Jenely. Sistema de Manejo de Datos de Ensayos Clínicos Cubano. Módulo Validación: Diseño del submódulo "Derivación de las variables del Cuaderno de Recogida de Datos". Tesis (Ingeniero en Ciencias Informáticas) Ciudad Habana. Universidad de las Ciencias Informáticas, 2008.

4. CENTRO NACIONAL COORDINADOR DE ENSAYOS CLÍNICOS (CENCEC). Experiencias [en línea] (2013). Disponible en: <http://www.cencec.sld.cu/pgs/resultados.htm>

5. INMON, Bill. Building The Data Warehouse. Canadá : John Wiley & Sons, Inc, 1996. pág 31.

6. KIMBALL, Ralph, ROSS, Margy. The Data Warehouse Toolkit. Canada : John Wiley & Sons, Inc, 1996. pág. 310.

7. Arley, Ricardo Chinchilla. Mercado de datos: conceptos y metodologías de desarrollo. [En línea] 2010. http://www.kerwa.ucr.ac.cr/bitstream/handle/10669/619/TM%2024-3%20art%206.pdf?sequence=1.

8. Ponniah, Paulraj.Data Warehousing Fundamentals. EUA: JOHN WILEY & SONS, INC, 2001. págs. 364-368. 0-471-22162-7.

9. HIDALGO LÓPEZ, Leydis. Mercado de datos para la Unidad Central de Cooperación Médica. Universidad de las Ciencias Informáticas, La Habana, 2012. [Consultado: el 25 de noviembre del 2013].

10. Ponniah, Paulraj.Data Warehousing Fundamentals. EUA: JOHN WILEY & SONS, INC, 2001. págs. 364-368. 0-471-22162-7.

11. ACOSTA MÉNDEZ, Geidy. Mercado de datos para una Dirección de Salud en Cuba. XV Convención y Feria Internacional Informática 2013. [Consultado: el 25 de noviembre del 2013].

12. SÁNCHEZ GALLARDO, Yisel de Lisy, PÉREZ REBOLLO, Rolando y WILKINSON BRITO, Bárbara.Diseño del mercado de datos CIMAVAX EGF para el almacén de datos del Centro de Inmunología.

13. WREMBEL, Robert y CONCILIA, Christian. El modelo de Hechos dimensionales [en línea] (2010). [Consulta: 9 de diciembre 2013]. Disponible en: <http://www.monografias.com/trabajos57/data-warehouse-sql/data-warehouse-sql2.shtml>

14. Sistemas de Información Orientados a la Toma de Decisiones:el enfoque multidimensional. Dapena Bosquet, Isabel. Muñoz San Roque, Antonio. Sánchez Miralles, Álvaro. Madrid, España: s.n., 2005.

15. El modelo multidimensional. [Online] [Cited: 11 10, 2013.] elvex.ugr.es/idbis/db/docs/intro/F Modelo multidimensional.pdf.

16. Kimball Ralph: "FactTables and Dimension Tables". [En línea]. Disponible en: http://www.intelligententerprise.com/030101/602warehouse1_1.jhtml.

17. MSDN. Tablas de hechos. [En línea] [Citado el: 2 de abril de 2014.] http://msdn.microsoft.com/es-es/library/ms244679%28v=vs.80%29.aspx.

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

19. GONZÁLEZ HERNÁNDEZ, Yanisbel. Propuesta de metodología para el desarrollo de almacenes de datos en DATEC. La Habana, 2011. [Consultado: el 9 de febrero del 2014].

20. ZEPEDA SÁNCHEZ, Leopoldo Zenaido. Universidad Politécnica de Valencia. Departamento de Sistemas Informáticos y Computación, 2008. [Consultado: el 22 de febrero del 2014]. Disponible en: <http://riunet.upv.es/bitstream/handle/10251/2506/tesisUPV2841.pdf>

21. BERNABEU, R.D. Hefesto: Metodología propia para la construcción de un Data Warehouse, 2007. [Consultado: el 9 de febrero del 2014]. Disponible en: <http://www.dataprix.com/es/hefesto-metodologia-propia-para-la-construccion-un-data-warehouse>

22. SOFTWARE.COM.AR. [Online] Targetware. [Cited: 11 6, 2014.] http://www.software.com.ar/visual-paradigm-para-uml.html.

23. EcuRed. Sistema Gestor de Base de Datos. [Consultado: el 15 de noviembre del 2013]. Disponible en:<http://www.ecured.cu/index.php/Sistema_Gestor_de_Base_de_Datos>

24. Martínez, Rafael. PostgreSQL. [En línea] 2 de octubre de 2010. [Citado el: 15 de noviembre de 2013] http://www.postgresql.org.es/sobre_postgresql.

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

26. PostgreSQL. PostgreSQL. [En línea] [Citado el: 5 de diciembre de 2012.] http://postgresql-dbms.blogspot.com/p/limitaciones-puntos-de-recuperacion.html.

27. Argentina, Grupo de Usuarios POSTGRES de. ARPUG. [Online] [Cited: 20 noviembre, 2013.] http://www.arpug.com.ar/trac/wiki/PgAdmin

28. ETL-Tools.Info. [Online] [Cited: diciembre 4, 2013.] http://etl-tools.info/es/bi/proceso_etl.htm.

29. Integración y Calidad de Datos. [Online] 7 17, 2008. [Cited: noviembre 14, 2013.] http://integracionycalidad.blogspot.com/2008/07/migraciones-fusiones-y-adquisiciones.html.

30. DataCleaner. [En línea] 2012. [Citado el: 6 de 2 de 2014.] http://datacleaner.org/

31. Ventajas de PENTAHO. [Consultado: el 7 de 2 del 2014]. Disponible en: https://sites.google.com/site/pentahounicah/ventajas-de-pentaho

32. Facultad de Ciencias Exactas y Naturales y Agrimiensura. [Online] [Cited: marzo 1, 2014.] http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/OLAPMonog.pdf.

33. RIVAS, Antonio. OLAP, MOLAP, ROLAP. Aprendiendo Business Intelligence, 2011. [Consultado: el 3 de marzo del 2014].Disponible en: <http://www.bi.dev42.es/2011/02/23/olap-molap-rolap>

34. Diaz, Josep Curto. Introducción al Business Intelligence. Editorial UOC, 2012.

35. Librería Digital. Metadata resources. [Consultado: el 8 de abril del 2014]. Disponible en:

<http://archive.ifla.org/II/metadata.htm>

36. MURTHA BACA, ed. Introduction to Metadata: Pathways to Digital Information. Los Angeles: Getty Research Institute, 2001. [Consultado: el 8 de abril del 2014]. Disponible en:

<http://getty.edu/research/publications/electronic_publications/intrometadata/>

Bibliografía

1. ACOSTA MÉNDEZ, Geidy. Mercado de datos para una Dirección de Salud en Cuba. XV Convención y Feria Internacional Informática 2013. [Consultado: el 25 de noviembre del 2013].

2. Argentina, Grupo de Usuarios POSTGRES de. ARPUG. [Online] [Cited: 20 noviembre, 2013.] http://www.arpug.com.ar/trac/wiki/PgAdmin.

3. Arley, Ricardo Chinchilla. Mercado de datos: conceptos y metodologías de desarrollo. [En línea] 2010. http://www.kerwa.ucr.ac.cr/bitstream/handle/10669/619/TM%2024-3%20art%206.pdf?sequence=1.

4. BERNABEU, R.D. Hefesto: Metodología propia para la construcción de un Data Warehouse, 2007. [Consultado: el 9 de febrero del 2014]. Disponible en: <http://www.dataprix.com/es/hefesto-metodologia-propia-para-la-construccion-un-data-warehouse>

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

6. CENTRO DE INMUNOLOGÍA MOLECULAR. Información General [en línea] (2013). Disponible en: <http://www.cim.sld.cu/>

7. CENTRO NACIONAL COORDINADOR DE ENSAYOS CLÍNICOS (CENCEC). Experiencias [en línea] (2013). Disponible en: <http://www.cencec.sld.cu/pgs/resultados.htm>

8. CHAN, Dra. Margaret. Cáncer. Organización Mundial de la Salud [en línea] (2013). Disponible en: <http://www.who.int/topics/cancer/es/>

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

10. Chuc-Durán, Diana Graciela. Introducción a los Datawarehouses. Introducción a los Datawarehouses. [En línea] Junio de 2007 [Citado el: 9 de Noviembre de 2013.] http://www.publicaciones.ujat.mx/publicaciones/revista_dacb/Acervo/v6n1OL/v6n1a5-ol/index.html.

11. DataCleaner. [En línea] 2012. [Citado el: 6 de 2 de 2014.] http://datacleaner.org/

12. DATAMART. [Online] [Cited: 11 5, 2013.] http://datamart.wikispaces.com/wiki/changes.

13. Diaz, Josep Curto. Introducción al Business Intelligence. Editorial UOC, 2012

14. EcuRed. Sistema Gestor de Base de Datos. [Consultado: el 15 de noviembre del 2013]. Disponible en:<http://www.ecured.cu/index.php/Sistema_Gestor_de_Base_de_Datos>

15. El modelo multidimensional. [Online] [Cited: 11 3, 2014.] elvex.ugr.es/idbis/db/docs/intro/F Modelo multidimensional.pdf.

16. ETL-Tools.Info. [Online] [Cited: diciembre 4, 2013.] http://etl-tools.info/es/bi/proceso_etl.htm.

17. Facultad de Ciencias Exactas y Naturales y Agrimiensura. [Online] [Cited: marzo 1, 2014.] http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/OLAPMonog.pdf.

18. GONZÁLEZ HERNÁNDEZ, Marlien y DE LAOSA SOCARRÁS, Jenely. Sistema de Manejo de Datos de Ensayos Clínicos Cubano. Módulo Validación: Diseño del submódulo "Derivación de las variables del Cuaderno de Recogida de Datos". Tesis (Ingeniero en Ciencias Informáticas) Ciudad Habana. Universidad de las Ciencias Informáticas, 2008.

19. GONZÁLEZ HERNÁNDEZ, Yanisbel. Propuesta de metodología para el desarrollo de almacenes de datos en DATEC. La Habana, 2011. [Consultado: el 9 de febrero del 2014].

20. HIDALGO LÓPEZ, Leydis. Mercado de datos para la Unidad Central de Cooperación Médica. Universidad de las Ciencias Informáticas, La Habana, 2012. [Consultado: el 1 de mayo del 2014].

21. INMON, Bill. Building The Data Warehouse. Canadá : John Wiley & Sons, Inc, 1996. pág 31.

22. Integración y Calidad de Datos. [Online] 7 17, 2008. [Cited: noviembre 14, 2013.] http://integracionycalidad.blogspot.com/2008/07/migraciones-fusiones-y-adquisiciones.html.

23. KIMBALL, Ralph, ROSS, Margy. The Data Warehouse Toolkit. Canada : John Wiley & Sons, Inc, 1996. pág. 310.

24. Librería Digital. Metadata resources. [Consultado: el 8 de abril del 2014]. Disponible en:

<http://archive.ifla.org/II/metadata.htm>.

25. Martínez, Rafael. PostgreSQL. [En línea] 2 de octubre de 2010. [Citado el: 15 de noviembre de 2013] http://www.postgresql.org.es/sobre_postgresql.

26. MURTHA BACA, ed. Introduction to Metadata: Pathways to Digital Information. Los Angeles: Getty Research Institute, 2001. [Consultado: el 8 de abril del 2014]. Disponible en:

<http://getty.edu/research/publications/electronic_publications/intrometadata/>

27. Ponniah, Paulraj.Data Warehousing Fundamentals. EUA: JOHN WILEY & SONS, INC, 2001. págs. 364-368. 0-471-22162-7.

28. PostgreSQL. PostgreSQL. [En línea] [Citado el: 5 de diciembre de 2012.] http://postgresql-dbms.blogspot.com/p/limitaciones-puntos-de-recuperacion.html.

29. Portal Web de la Oficina Nacional de Estadistica de la Republica Dominicana. Portal Web de la Oficina Nacional de Estadistica de la Republica Dominicana. [En línea] 11 de Febrero de 2010. [Citado el: 6 de Noviembre de 2013.] http://www.one.gob.do/index.php?module=articles&func=display&aid=1377.

30. RIVAS, Antonio. OLAP, MOLAP, ROLAP. Aprendiendo Business Intelligence, 2011. [Consultado: el 3 de marzo del 2014].Disponible en: <http://www.bi.dev42.es/2011/02/23/olap-molap-rolap>

31. Rolando Alfredo Hernández León Sayda Coello González EL PARADIGMA CUANTITATIVO DE LA INVESTIGACIÓN CIENTIFICA [Book]. - La Habana : [s.n.], 2002. - pp. 82-95.

32. SÁNCHEZ GALLARDO, Yisel de Lisy, PÉREZ REBOLLO, Rolando y WILKINSON BRITO, Bárbara.Diseño del mercado de datos CIMAVAX EGF para el almacén de datos del Centro de Inmunología.

33. Sistemas de Información Orientados a la Toma de Decisiones:el enfoque multidimensional. Dapena Bosquet, Isabel. Muñoz San Roque, Antonio. Sánchez Miralles, Álvaro. Madrid, España: s.n., 2005.

34. SOFTWARE.COM.AR. [Online] Targetware. [Cited: 11 6, 2014.] http://www.software.com.ar/visual-paradigm-para-uml.html.

35. Ventajas de PENTAHO. [Consultado: el 7 de 2 del 2014]. Disponible en: https://sites.google.com/site/pentahounicah/ventajas-de-pentaho

36. WREMBEL, Robert y CONCILIA, Christian. El modelo de Hechos dimensionales [en línea] (2010). [Consulta: 9 de marzo 2014]. Disponible en: <http://www.monografias.com/trabajos57/data-warehouse-sql/data-warehouse-sql2.shtml>

37. ZEPEDA SÁNCHEZ, Leopoldo Zenaido. Universidad Politécnica de Valencia. Departamento de Sistemas Informáticos y Computación, 2008. [Consultado: el 22 de febrero del 2014]. Disponible en: <http://riunet.upv.es/bitstream/handle/10251/2506/tesisUPV2841.pdf>

Anexos

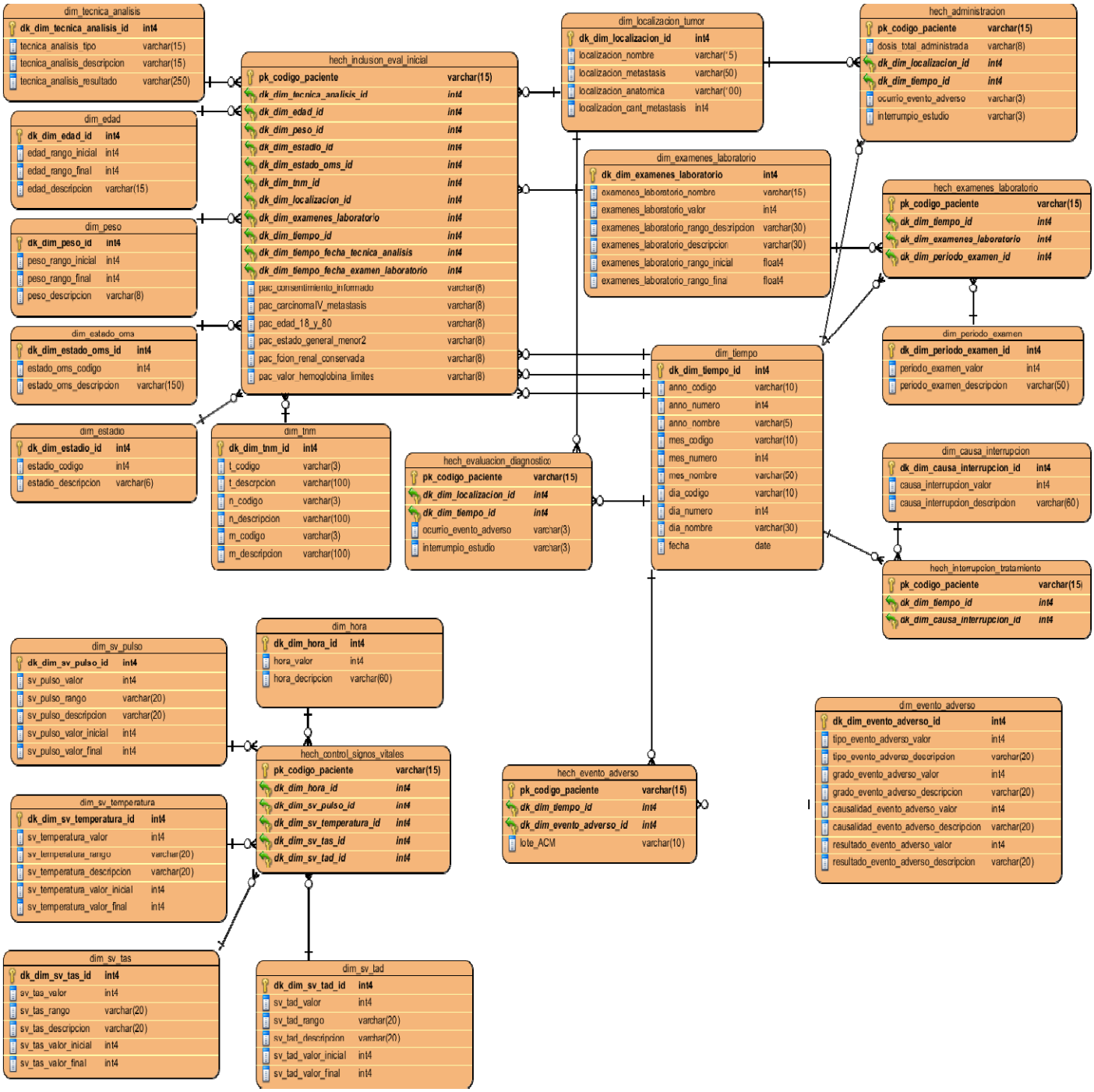
Anexo 1

Capacidad funcional del paciente según grados de la OMS

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

Anexo 2

Modelo de datos



Glosario de términos

Cáncer de mama: consiste en un crecimiento anormal y rápido de células cancerosas en el tejido mamario.

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

Biopsia: procedimiento diagnóstico que consiste en la extracción de una muestra de tejido obtenida por medio de métodos invasivos para examinarla al microscopio.

Metástasis: consiste en la aparición de un tumor canceroso a distancia del original en otros tejidos que se ha diseminado a través del torrente sanguíneo o del sistema linfático.