Universidad de las Ciencias Informáticas

FACULTAD 6

Título: Subsistemas de almacenamiento e integración LeukoCIM para el almacén de datos de los Ensayos Clínicos del Centro de Inmunología Molecular

Trabajo de Diploma para optar por el título de

Ingeniero en Ciencias Informáticas.

Autora: Yanet Cardoso García

Tutores: Ing. Rayko Emilio Torres Cruz

Ing. Luilly Díaz Montero

La Habana, junio de 2013

"Año 55 de la Revolución"

"Estar preparado es importante, saber esperar lo es aún más, pero aprovechar el momento adecuado es la clave de la vida."

Arthur Schnitzler

Declaración de autoría

Yo Yanet Cardoso García declaro ser la autora del trabajo de diploma con título: Subsistemas de almacenamiento e integración LeukoCIM para el almacén de datos de los Ensayos Clínicos del Centro de Inmunología Molecular y admito los derechos patrimoniales del mismo con carácter exclusivo a la Universidad de las Ciencias Informáticas, específicamente a la Facultad 6.

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

_____

Firma de la autora

Yanet Cardoso García

_____

Firma del tutor

Rayko Emilio Torres Cruz

_____

Firma del tutor

Luilly Díaz Montero

Datos de contacto

TUTORES:

Ing. Rayko Emilio Torres Cruz
Universidad de las Ciencias Informáticas,
Ciudad de la Habana, Cuba
E-mail: retorres@uci.cu

Ing. Luilly Díaz Montero
Universidad de las Ciencias Informáticas,
Ciudad de la Habana, Cuba
E-mail: lmontero@uci.cu

Dedicatoria

A toda mi familia, que siempre han esperado lo mejor de mí, especialmente a mi mamá Odalys, a mi papá Felipe, a mi hermana Yaquelín y a mi sobrinita Lili.

A la Revolución Cubana, sin la cual no hubiera sido posible este acontecimiento, en especial a su máximo líder Fidel Castro Ruz.

Resumen

Las Tecnologías de la Información y las Comunicaciones (TIC) son un factor de vital importancia en la transformación de la sociedad, especialmente para áreas del conocimiento como la medicina. La presente investigación surge por la necesidad de gestionar toda la información que generan los ensayos clínicos del producto LeukoCIM, donde se recopilan los datos relacionados con los eventos adversos, la inclusión y evaluación de los pacientes que participaron en dichos ensayos. Se pretende crear un repositorio donde la información se encuentre centralizada, estandarizada y accesible para su consulta y para la aplicación de técnicas de minería de datos. Con este fin se realiza un estudio de las tecnologías que posibilitan almacenar gran cantidad de datos, así como la metodología para el desarrollo de almacenes de datos y se definieron las herramientas que permiten la construcción del sistema. Se plasma el análisis, diseño e implementación de los subsistemas de almacenamiento e integración, así como la aplicación de las listas de chequeos y los casos de prueba, con el objetivo de obtener un producto que cumpla con las necesidades del cliente. Finalmente se obtuvieron los Subsistemas de almacenamiento e integración LeukoCIM, que forman parte del almacén de datos de los ensayos clínicos del Centro de Inmunología Molecular (CIM), el cual permite mantener disponible la información histórica para su análisis.

Palabras claves: almacén de datos, Centro de Inmunología Molecular, subsistemas de almacenamiento e integración LeukoCIM.

Índice de contenido

Índice de figuras ..... *V*
Índice de tablas ..... *VI*
Introducción *1*
Capítulo 1. Fundamentos teóricos *5*
Introducción 5
1.1 Centro de Inmunología Molecular. Ensayos Clínicos 5
1.1.1 LeukoCIM 5
1.2 Almacenes de Datos 7
1.2.1 Características de los almacenes de datos 7
1.2.2 Ventajas y desventajas de los almacenes de datos 8
1.3 Mercado de datos 8
1.3.1 Características de los mercados de datos 9
1.4 Subsistemas de almacenamiento e integración LeukoCIM 9
1.5 Sistemas de almacenamiento de datos 10
1.5.1 Procesamiento Analítico en Línea Relacional (ROLAP) 10
1.5.2 Procesamiento Analítico en Línea Multidimensional (MOLAP) 11
1.5.3 Procesamiento Analítico en Línea Híbrido (HOLAP) 12
1.6 Metodología de desarrollo 13
1.6.1 Propuesta de Metodología para el Desarrollo de Almacenes de Datos 14
1.7 Herramienta de modelado 15
1.7.1 Visual Paradigm 8.0 15
1.8 Sistema Gestor de Base de Datos 16
1.8.1 PostgreSQL 9.1 16
1.8.2 Administrador de Base de Datos. PgAdmin III 1.14.0 16
1.9 Herramientas para la integración de datos 17
1.9.1 Pentaho Data Integration 4.2.1 18
1.9.2 DataCleaner 1.5.3 19
Conclusiones del capítulo 19
Capítulo 2. Análisis y Diseño *21*
Introducción 21
2.1 Necesidades del negocio 21
2.2 Especificación de los requisitos 22
2.2.1 Requisitos de información 22
2.2.2 Requisitos funcionales 23
2.2.3 Requisitos no funcionales 24
2.3 Reglas del negocio 24
2.4 Casos de uso del sistema 26

2.4.1 Actores del sistema....26
2.4.2 Casos de uso de información....26
2.4.3 Casos de uso funcionales....27
2.4.4 Diagrama de casos de uso....30
2.5 Definición de la arquitectura base de un mercado de datos....31
2.6 Diseño de la solución....32
2.6.1 Diseño del subsistema de almacenamiento....32
2.6.2 Diseño del subsistema de integración....40
2.7 Política de respaldo y recuperación....44
2.8 Esquema de seguridad....44
Conclusiones del capítulo....45
Capítulo 3: Implementación y Prueba....46
Introducción....46
3.1 Implementación del subsistema de almacenamiento....46
3.1.1 Estándares de codificación....46
3.1.2 Implementación del modelo de datos físico....47
3.2 Implementación del subsistema de integración....49
3.2.1 Implementación de las transformaciones....51
3.2.2 Implementación de los trabajos....53
3.2.3 Gestión de los metadatos....54
3.3 Pruebas....55
3.3.2 Pruebas unitarias....56
3.3.3 Pruebas de integración....57
3.4 Herramientas para la aplicación de las pruebas....57
3.4.1 Casos de prueba....57
3.4.2 Listas de chequeo....60
3.5 Calidad de datos....62
3.5.1 Perfilado de datos....62
3.5.2 Auditoría de datos....63
Conclusiones del capítulo....63
Conclusiones generales....64
Recomendaciones....65
Referencias bibliográficas....66
Bibliografía....68

Índice de figuras

Figura 1: Procesamiento Analítico en Línea Relacional (ROLAP). 11

Figura 2: Procesamiento Analítico en Línea Multidimensional (MOLAP). 11

Figura 3: Enfoque ascendente o bottom-up definido por Ralph Kimbal. 13

Figura 4: Enfoque descendente o top-down definido por Bill Inmon. 13

Figura 5: Ciclo de vida de la Metodología de desarrollo de Almacenes de datos. 14

Figura 6: Diagrama de CUS. 30

Figura 7: Arquitectura del sistema. 31

Figura 8: Esquemas del modelo dimensional. 38

Figura 9: Modelo de datos. 39

Figura 10: Distribución de los tipos de datos. 41

Figura 11: Diseño de la transformación para la carga de los hechos. 43

Figura 12: Estructura física de la base de datos. 49

Figura 13: Transformación para cargar la dimensión inclusión. 51

Figura 14: Transformación para cargar el hecho de la inclusión de niños y adultos. 52

Figura 15: Trabajo general. 54

Figura 16: Metadatos de proceso para las transformaciones 55

Figura 17: Resultados de la aplicación de las listas de chequeo a los artefactos. 61

Figura 18: Resultados del perfilado de datos realizado a la base de datos leukocim_dwh. 62

Índice de tablas

Tabla 1: Descripción del CUF: Realizar la extracción de los datos. 27

Tabla 2: Descripción del CUF: Realizar la transformación y carga de los datos. 29

Tabla 3: Matriz dimensional. 36

Tabla 4: Diccionario de datos. 41

Tabla 5: Roles y permisos de acceso a la base de datos. 44

Tabla 6: Esquemas y tablas de la aplicación. 47

Tabla 7: Caso de prueba para una regla de transformación. 59

Tabla 8: Aplicación de las listas de chequeo a los artefactos de ETL. 61

Introducción

En la sociedad actual el gran avance tecnológico se debe fundamentalmente al desarrollo de las Tecnologías de la Información y las Comunicaciones (TIC), destacando el papel que juegan las ciencias informáticas. Ambas han marcado un cambio profundo en la forma en que las personas interactúan, trayendo consigo mejoras en el procesamiento y gestión de la información. Las TICs sirven de apoyo a ramas de la sociedad entre las que se pueden mencionar la salud, la educación y la biotecnología, contribuyendo significativamente a la toma de decisiones.

Cuba ha encontrado en la industria del *software* una vía para incrementar su desarrollo económico, por tanto no está ajeno a los cambios científicos-técnicos que acontecen en el mundo. La inserción en campos como la genética y la biotecnología son un factor de gran importancia, ya que representan una línea estratégica en el desarrollo científico de un país pequeño y desprovisto de recursos. [1]

Es por tal razón que se ha desarrollado una poderosa infraestructura para la industria biotecnológica, cuya génesis fue la creación en 1965 del Centro Nacional de Investigaciones Científicas (CNIC), el cual generó múltiples avances en la ciencia cubana y destacados líderes científicos en el campo de la biomedicina.

La principal zona de desarrollo, que agrupa a más de 42 instituciones científicas, de investigación, producción y comercialización, conocido como el Polo Científico del Oeste de La Habana, es similar a un parque tecnológico[1]. [1] Este cuenta con tres de los centros más modernos y conocidos a nivel mundial, los cuales son el Instituto de Investigación Carlos J. Finlay, el Centro de Ingeniería Genética y Biotecnología (CIGB) y el Centro de Inmunología Molecular (CIM).

Este último es una moderna organización tanto en términos de su equipamiento como de recursos humanos. Entre sus principales actividades se destaca la utilización y desarrollo de la biología molecular con el fin de estudiar diversos tipos de cáncer, así como la elaboración de vacunas contra esta enfermedad y anticuerpos monoclonales[2]. El CIM realiza en hospitales altamente especializados Ensayos Clínicos (EC) para el diagnóstico de tumores por imágenes y el tratamiento del cáncer de diferentes orígenes y otras enfermedades del sistema inmune.

1 Los parques científicos y tecnológicos son espacios e instalaciones de gran calidad donde se estimula y gestiona el flujo de conocimiento y tecnología entre universidades e instituciones de investigación, empresas y mercados.

2 Sustancia producida en el organismo animal por la presencia de un antígeno, contra cuya acción reacciona específicamente.

Se encarga además de gestionar, almacenar y analizar toda la información recogida en estos ensayos siempre que se aplique un producto determinado. Uno de los fármacos con mayores resultados es el LeukoCIM, el cual generó una gran cantidad de datos que se encuentran almacenados en un sistema llamado EpiData. Este programa arroja ficheros en formato de hojas de cálculo Excel, lo que resulta difícil para los especialistas realizar análisis certeros que contribuyan con las decisiones que la entidad debe tomar.

La información que generan los ensayos no está debidamente gestionada ni estandarizada, siendo de vital importancia la informatización de la misma, tarea que requiere de personal altamente calificado. En este sentido el Centro de Tecnología de Gestión de Datos (DATEC) de la Universidad de las Ciencias Informáticas (UCI) cuenta con un departamento especializado en el desarrollo de soluciones de Inteligencia de Negocios y Almacenes de Datos. El mismo se rige por un conjunto de políticas y procedimientos que le permiten organizar grandes volúmenes de información y actualmente está a cargo de varios proyectos, uno de ellos relacionado con los EC que se llevan a cabo por el CIM.

El volumen de información almacenada en esta institución se ha incrementado considerablemente, debido al gran cúmulo de datos que genera cada uno de los EC del producto LeukoCIM. De esta forma torna complicado el proceso de manejo de la información por parte de los directivos de la institución, ya que la gestionan manualmente y además dificultan la realización de reportes y análisis estadísticos complejos, así como las consultas y la determinación de los indicadores relacionados con los EC de este producto.

Teniendo en cuenta lo anteriormente, se plantea como **problema de la investigación:** ¿Cómo estandarizar los datos del producto LeukoCIM para su almacenamiento de forma homogénea?

Determinando como **objeto de estudio:** los almacenes de datos, enmarcado en el **campo de acción:** subsistemas de almacenamiento e integración LeukoCIM para el Centro de Inmunología Molecular.

Para solucionar el problema de la investigación se identifica como **objetivo general:** Desarrollar los subsistemas de almacenamiento e integración LeukoCIM para el almacén de datos de los Ensayos Clínicos del Centro de Inmunología Molecular, que permite el almacenamiento homogéneo de la información, del cual se derivan los siguientes **objetivos específicos:**

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

3. Realizar la implementación de los subsistemas de almacenamiento e integración LeukoCIM.
4. Realizar pruebas a los subsistemas de almacenamiento e integración LeukoCIM.

Para dar cumplimiento a los objetivos planteados se proponen las siguientes **tareas de la investigación:**

1. Caracterización de las metodologías, herramientas y tecnologías a utilizar en el desarrollo de almacenes de datos.
2. Realización del levantamiento de los requisitos.
3. Descripción de los casos de uso de los subsistemas de almacenamiento e integración LeukoCIM.
4. Definición de la arquitectura de los subsistemas de almacenamiento e integración LeukoCIM.
5. Realización del diseño del subsistema de almacenamiento.
6. Realización del diseño del subsistema de integración.
7. Implementación del subsistema de almacenamiento.
8. Implementación del subsistema de integración.
9. Aplicación de las listas de chequeo.
10. Aplicación de los casos de prueba.

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

Capítulo 1. Fundamentos teóricos

En este capítulo se abordan definiciones sobre los almacenes de datos y la gestión de los EC del producto LeukoCIM en el CIM. Asimismo, se realiza un estudio bibliográfico de las metodologías, herramientas y tecnologías que se utilizan para el desarrollo de un almacén de datos. Luego se caracterizan y seleccionan para realizar los subsistemas de almacenamiento e integración LeukoCIM.

Capítulo 2. Análisis y diseño

En este capítulo se realiza el análisis para comprender el negocio, del que se definieron los requisitos de información, los funcionales y los no funcionales, así como las reglas del negocio. Se conforma el diagrama de casos de uso y la arquitectura de los subsistemas de almacenamiento e integración

LeukoCIM. Por otra parte se diseñan ambos subsistemas y se describe el esquema de seguridad y la política de respaldo y recuperación.

Capítulo 3. Implementación y pruebas

En este capítulo se realiza la implementación y validación de los subsistemas de almacenamiento e integración LeukoCIM. La implementación del primero comprende la definición de los estándares de codificación y la construcción del modelo físico. Por su parte, en el subsistema de integración de datos se implementan las transformaciones y los trabajos, así como los metadatos. Se exponen las pruebas realizadas a los subsistemas, así como los resultados obtenidos en cada una de ellas. Dichas pruebas son realizadas para garantizar el cumplimiento de las exigencias del cliente y la calidad del producto.

Capítulo 1. Fundamentos teóricos

Introducción

Actualmente el CIM maneja gran cantidad de información, la cual se deriva de los EC que se realizan. Esto dificulta a especialistas y directivos consultar y analizar la información, ya que requieren de mucho tiempo, lo que repercute directamente en la eficiencia de los procesos que allí se llevan a cabo.

En el presente capítulo se realiza un estudio sobre dicha institución e información relacionada con los EC que se generan del producto LeukoCIM. Por otra parte se abordan definiciones y características sobre los almacenes y mercados de datos, así como sus ventajas y desventajas. Se explica la metodología de desarrollo a utilizar y las herramientas seleccionadas para la realización de la investigación.

1.1 Centro de Inmunología Molecular. Ensayos Clínicos

El CIM es un centro que desde sus inicios se ha dedicado al desarrollo de biomoléculas y otros fármacos para el tratamiento de diferentes enfermedades relacionadas con el sistema inmunológico, principalmente el cáncer. Para ello, se realizan los análisis pertinentes con el objetivo de probar el funcionamiento de los fármacos y luego se aplican en pacientes que padecen diferentes enfermedades. Para la aprobación, prueba e introducción de dichos productos en el mercado se realizan EC que consiste en *"... cualquier investigación en seres humanos dirigida a descubrir o verificar los efectos clínicos, farmacológicos u otros efectos farmacodinámicos de un producto en investigación...".* [2]

Al terminar el 2011, el CIM desarrolló 60 EC, 20 de los cuales son multinacionales. [3] Actualmente comercializa cuatro vacunas para el tratamiento del cáncer, cuyas producciones se exhibieron en la XXX Feria Internacional de La Habana. El doctor Normando Iznaga-Escobar, físico nuclear con un Doctorado en Farmacología Clínica, explicó que esos fármacos se dividen en los llamados de soporte, que son los que buscan disminuir los efectos nocivos de la radio y quimioterapia, y los específicos dirigidos a blancos tumorales, que son los que tienen una mayor respuesta antitumoral. [4]

1.1.1 LeukoCIM

LeukoCIM es un fármaco de soporte, recomendado para mejorar la respuesta del organismo de los pacientes con cáncer. Su empleo en el sistema nacional de salud cubano fue aprobado en el 2002. En abril del 2004 las autoridades sanitarias realizaron pruebas para valorar la efectividad y seguridad de este

medicamento y por ahora han obtenido excelentes resultados que muestran que este producto puede aumentar la efectividad de los sueros citostáticos[3]. A finales del año 2005 incluían a 373 pacientes adultos y 129 niños en hospitales de 13 provincias de la isla. Luego del año 2006 hasta el 2009 se realizaron los ensayos a pacientes que presentan Síndrome de Inmunodeficiencia Adquirida (SIDA) y neutropenia. [5]

Según la especialista Patricia Piedra, el fármaco es una solución inyectable para uso subcutáneo o endovenoso, tiene particular importancia en el tratamiento de dolencias como la leucemia linfoblástica[4] aguda en niños, y ayuda a evitar posibles infecciones bacterianas en pacientes con tumoraciones malignas, que ven reducidas sus defensas tras la aplicación de la quimioterapia. Este medicamento lo fabrican el CIM y el Centro Nacional de Biopreparados (BIOCEN) y sustituye a un producto similar, el *Neupogen*, de factura estadounidense que Cuba importaba para trasplantes de médula ósea, oncología pediátrica y otras enfermedades graves.

El BIOCEN plantea que el producto, cuyo nombre comercial es ior® LeukoCIM, es utilizado además en los trasplantes de médula ósea y para el tratamiento de personas infectadas con el Virus de Inmunodeficiencia Humana (VIH). En el año 2006 se produjeron por BIOCEN más de 23 mil dosis del medicamento, el cual es descrito como un factor estimulante de colonias granulosíticas[5].

Según el CIMAB S.A que es el representante exclusivo del CIM plantea que el producto ior® LeukoCIM regula la producción de neutrófilos[6] en la médula ósea y actúa sobre la proliferación, diferenciación y otras funciones celulares. Tiene un valor terapéutico potencial en pacientes con neutropenia iatrogénica, pacientes sometidos a quimioterapia y/o transplante de médula ósea y en aquellos pacientes con enfermedades relacionadas con anomalías en los neutrófilos como son la anemia aplástica, neutropenia cíclica o agranulocitocis congénita. [6]

Con la utilización de este producto se han realizado cuatro ensayos clínicos a pacientes niños, adultos, que presentan el SIDA y neutropenia. En dichos ensayos se recoge información confidencial de los pacientes, comenzando con datos demográficos en la inclusión así como los eventos adversos que han sufrido, el tratamiento indicado y administrado, entre otros. A partir de la recogida de esta información se

3 Fármaco que frena la multiplicación celular.

4 Cáncer de la sangre y la médula ósea.

5 Constituye un regulador fisiológico de la función y producción de granulocitos(células defensivas).

6 Su función es la defensa del organismo contra las infecciones bacterianas.

generan varios ficheros en disímiles formatos, los cuales se necesitan en una única estructura para ser analizados y consultados con rapidez.

Las empresas hoy día necesitan herramientas que le permitan actuar de manera correcta en las operaciones que realizan y en las decisiones que deben tomar. Estas deben ser rápidas y basadas en buenos cimientos, por lo que se necesita de hechos y cifras que deben ser manejados y analizados en el menor tiempo posible debido a que la competencia en los negocios crece rápidamente. Contar con conocimientos correctos significa tener respuestas correctas y trazarse estrategias en beneficio de la empresa. Uno de los mecanismos que permiten resolver el problema del análisis y manejo de la información son los almacenes de datos.

1.2 Almacenes de datos

Los almacenes de datos tienen como objetivo organizar y permitir la consulta de la información con inmediatez. Estos son muy necesarios en el entorno empresarial para tomar mejores decisiones. La definición de almacenes de datos ha sido planteada por varios autores, estas definiciones enfatizan el uso de la información dentro de una empresa para apoyar a la toma de decisiones.

El concepto de almacenes de datos proviene de la palabra en inglés *Data Warehouse*, cuyo padre del término es William H. Inmon, el cual plantea *"…un almacén de datos es una colección de datos orientado por temas, integrado, no volátil y variable en el tiempo que apoya las decisiones administrativas"*. [7]

Otro de los autores reconocidos en este ámbito es Ralph Kimball, expresa que un almacén de datos es *"...una copia de datos transaccionales, específicamente estructurados para la consulta y el análisis"*. [8]

En la presente investigación se concluye que un almacén de datos es la unión de varios mercados de datos (subconjunto del almacén) que contienen información de las principales áreas de la empresa, obteniéndola de una o varias fuentes para su posterior análisis y que persistirá en el tiempo.

1.2.1 Características de los almacenes de datos

Como bien propone la definición de Inmon [7] los almacenes de datos constan de cuatro características principales.

Orientado por temas: el almacén de datos está orientado por las principales áreas o temáticas de la empresa. Los datos se organizan por temas para facilitar su acceso y entendimiento por parte de los usuarios finales.

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

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

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

1.2.2 Ventajas y desventajas de los almacenes de datos

El uso de los almacenes de datos en empresas que manejan grandes volúmenes de información es vital ya que es una vía para tomar decisiones de forma clara, concisa y rápida. A continuación se proponen algunas ventajas y desventajas que poseen los almacenes de datos.

Poseen la ventaja de permitir el acceso rápido a una gran cantidad de datos en línea, debido a la capacidad que tienen para almacenar, estructurar y hacerlos disponibles. Además facilitan el funcionamiento de las aplicaciones de los sistemas de apoyo a la decisión tales como informes de tendencia e informes de excepción. Integran datos históricos sobre la actividad de la organización en un único repositorio, permitiendo analizar los datos del negocio desde la perspectiva de su evolución en el tiempo e identificar nuevas oportunidades, que posibiliten tomar decisiones estratégicas.

Como desventaja se encuentra que el ciclo de vida puede suponer altos costos, no suele ser estático y los costos de mantenimiento son elevados. Los almacenes de datos se pueden quedar obsoletos relativamente pronto si no se actualizan con frecuencia, por lo que en ocasiones no devuelven la información del todo óptima, y supone una pérdida para la organización. Se tiene el riesgo de fracasar en la construcción del sistema, al subestimar los costes de captura y preparación de los datos, así como cambios frecuentes en los requisitos de los usuarios.

1.3 Mercado de datos

Un mercado de datos, también conocido como *Data Mart,* almacena la información de un área o departamento específico dentro del negocio. “Un mercado de datos es una solución que, compartiendo

tecnología con el almacén de datos (pero con contenidos específicos, volumen de datos más limitado y un alcance histórico menor), permite dar soporte a una empresa pequeña, un departamento o área de negocio de una empresa grande". [8]

Con el objetivo de facilitar la construcción y utilización de un almacén de datos se crean los mercados de datos, que representan un subconjunto del almacén de datos, referentes a los requisitos de un departamento o área de negocio concreto.

1.3.1 Características de los mercados de datos

- ✓ Según las necesidades de los usuarios el diseño del mercado de datos se realiza siguiendo una estructura consistente.
- ✓ La información histórica que posee es mínima comparada con la información histórica que guardan los almacenes de datos.
- ✓ Presentan mayor nivel de detalle, por lo que contienen el grado de granularidad necesaria.
- ✓ A la hora de construirlo presenta costes adicionales en *hardware*, *software* y accesos de red.
- ✓ Debido a que hay grupos de usuarios que solo acceden a un subconjunto preciso de datos, se hace más fácil el acceso a las herramientas de consulta y divide los datos para controlar mejores accesos. [8]

1.4 Subsistemas de almacenamiento e integración LeukoCIM

Para gestionar los EC del producto LeukoCIM realizados por el CIM es necesario usar una tecnología de las antes expuestas. Un mercado de datos consta de tres etapas, las cuales son: análisis y diseño; extracción, transformación y carga de los datos (*Extraction, Transformation and Loading*, ETL por sus siglas en inglés) e inteligencia de negocio. Se ha determinado desarrollar dos subsistemas que darán respuesta al problema planteado, uno es el de almacenamiento que estará en correspondencia con la primera etapa, en la cual se estudia el negocio y se elaboran los requisitos de información, es decir las necesidades del cliente. Por otra parte en la segunda etapa, el subsistema de integración se encargará una vez de modelado los datos, extraerlos de la fuente, estandarizarlos, integrarlos y por último cargarlos en la base de datos. De esta forma se garantiza que en posteriores investigaciones se le apliquen técnicas de minería de datos.

La forma en que se almacenan los datos es elemental para garantizar el rendimiento de las consultas y el procesamiento en general, con este fin se utilizan los sistemas de almacenamiento.

1.5 Sistemas de almacenamiento de datos

Existen dos tipos de almacenamiento, uno es el Procesamiento de Transacciones en Línea (*On-Line Transactional Processing,* OLTP por sus siglas en inglés) y el otro Procesamiento Analítico en Línea (*On-Line Analytical Processing,* OLAP por sus siglas en inglés).

Los sistemas OLTP son bases de datos orientadas al procesamiento de transacciones, al que se le pueden realizar operaciones de inserción, modificación y borrado de datos. Este proceso es típico en bases de datos operacionales. [9]

Por su parte los sistemas OLAP son bases de datos orientadas al procesamiento analítico, en el que el acceso a los datos es de solo lectura y la operación más común es la consulta, con pocas inserciones, actualizaciones o eliminaciones. Este sistema permite acceder a grandes volúmenes de datos, de los que se puede extraer información útil. Este proceso es típico en los mercados de datos. [9]

Para el desarrollo de la investigación se utilizó este último sistema porque ofrece gran ventaja sobre los sistemas OLTP ya que los datos se estructuran según las áreas del negocio y los formatos de los mismos están integrados de manera uniforme en toda la organización, lo que favorece en la rapidez del tiempo de respuesta de las consultas. Proporciona análisis estadísticos y permite ver la información en determinadas vistas y dimensiones. Dentro de este sistema se pueden encontrar tres tipos de almacenamiento: el relacional, el multidimensional y el híbrido.

1.5.1 Procesamiento Analítico en Línea Relacional (ROLAP)

El tipo de almacenamiento de datos ROLAP guarda los datos en una base de datos relacional. Este utiliza una arquitectura de tres niveles (Figura 1 [25]). La base de datos relacional maneja el almacenamiento de los datos, el motor OLAP proporciona la funcionalidad analítica, y una herramienta especializada es empleada para el nivel de presentación.

Como parte de su función se encuentran los usuarios finales que ejecutan su análisis multidimensional, a través del motor OLAP, el cual transforma los datos del lenguaje de consulta estructurado (*Structured Query Language,* SQL por sus siglas en inglés) ejecutadas en las bases de datos relacionales y los resultados son devueltos a los usuarios.

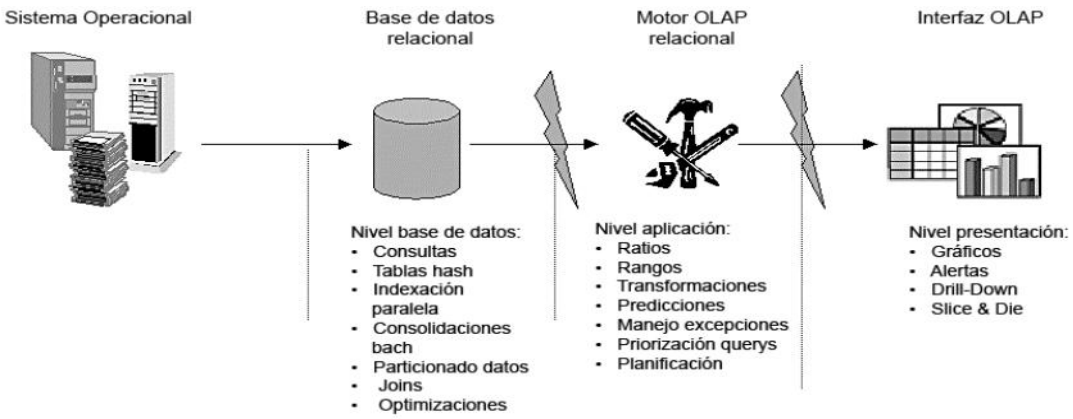

Figura 1: Procesamiento Analítico en Línea Relacional (ROLAP).

La arquitectura ROLAP es capaz de usar datos pre-calculados (si estos están disponibles), o de generar dinámicamente los resultados desde la información elemental. Esta arquitectura accede directamente a los datos del almacén y soporta técnicas de optimización para acelerar las consultas como tablas particionadas, soporte a la desnormalización e índices. [10]

1.5.2 Procesamiento Analítico en Línea Multidimensional (MOLAP)

El tipo de almacenamiento de datos MOLAP usa una base de datos multidimensional. Este utiliza una arquitectura de dos niveles: la base de datos multidimensional y el motor analítico (Figura 2 [25]). El primer nivel es el encargado del manejo, acceso y obtención de los datos. El nivel de aplicación es el responsable de la ejecución de las consultas OLAP. El nivel de presentación se integra con el de aplicación y proporciona una interfaz a través de la cual los usuarios finales visualizan los análisis OLAP. [10]

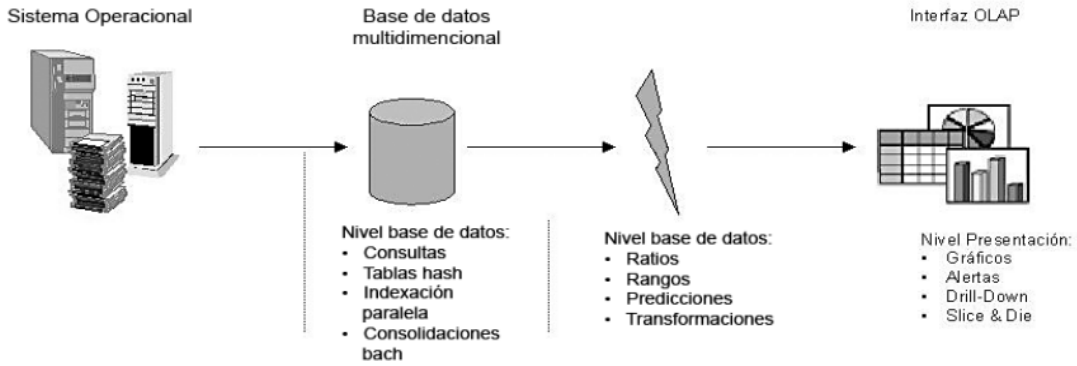

Figura 2: Procesamiento Analítico en Línea Multidimensional (MOLAP).

La información procedente de los sistemas transaccionales se carga en el sistema MOLAP. Una vez cargados los datos en la base de datos multidimensional, se realiza una serie de cálculos para obtener datos agregados a través de las dimensiones del negocio, poblando la estructura de la base de datos multidimensional.

Luego de completar esta estructura, se generan índices y se emplean algoritmos de tablas *hash*[7] para mejorar los tiempos de acceso de las consultas. Una vez que el proceso de poblado ha finalizado, la base de datos multidimensional está lista para su uso. Los usuarios solicitan informes a través de la interfaz y la lógica de aplicación de la base de datos multidimensional obtiene los datos.

1.5.3 Procesamiento Analítico en Línea Híbrido (HOLAP)

Se han desarrollado soluciones de OLAP híbridas que combinan el uso de las arquitecturas ROLAP y MOLAP. En una solución con HOLAP, los registros detallados (los volúmenes mayores) se mantienen en la base de datos relacional, mientras que los agregados lo hacen en un almacén MOLAP independiente. [11]

En los últimos años se han producido debates alrededor de los tipos de almacenamiento MOLAP y ROLAP. Por lo general, las implementaciones de MOLAP presentan mejor rendimiento que la tecnología relacional; sin embargo, tienen problemas de escalabilidad, por ejemplo, la adición de dimensiones a un esquema ya existente. Por otra parte, las implementaciones de ROLAP son más escalables y atractivas debido a que aprovechan las inversiones efectuadas en tecnología de base de datos relacionales. [11]

Una vez analizados los diferentes tipos de almacenamiento se selecciona ROLAP ya que ahorra espacio de almacenamiento y es útil cuando se trabaja con amplios conjuntos de datos. Su ventaja principal reside en la posibilidad de utilizar una tecnología ampliamente extendida y utilizada para la gestión de datos, los sistemas relacionales. [12] Además se tuvo en cuenta que el Sistema Gestor de Base de Datos (SGBD) PostgreSQL soporta el almacenamiento relacional, no así 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.

7 Hash es una función computable mediante un algoritmo, que tiene como entrada un conjunto de elementos y los mapea en un rango de salida finito.

1.6 Metodología de desarrollo

Existen diversas metodologías que pretenden dar un acercamiento a una propuesta ideal para el desarrollo de un almacén de datos. Cada autor la orienta a la optimización del rendimiento y a su visión de los principales procesos que se deben tener en cuenta para construir un almacén de datos flexible y dinámico, pero también es necesario conocer el contexto en que se encuentre la empresa y los objetivos que persigan para hacer una buena elección.

Para realizar un almacén de datos se encuentran dos grandes enfoques desde el punto de vista arquitectónico, el ascendente o *bottom-up* definido por Ralph Kimball (Figura 3) y el descendente o *top-down* planteado por Bill Inmon (Figura 4).

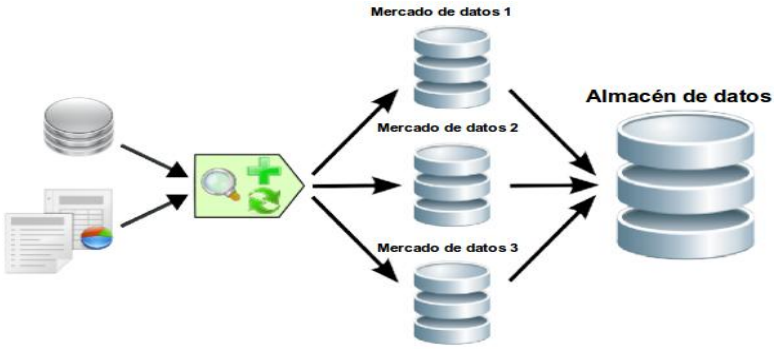

Figura 3: Enfoque ascendente o bottom-up definido por Ralph Kimbal.

El enfoque ascendente consiste en construir primeramente los mercados de datos de la entidad en cuestión, para luego unirlos y conformar el almacén de datos. En el enfoque descendente los mercados se crearán después de haber terminado el almacén completo de la organización, por lo que es necesario conocer todos los aspectos de la empresa para poder realizar el mismo, ello trae consigo varios cambios a la hora de mejorar los procesos.

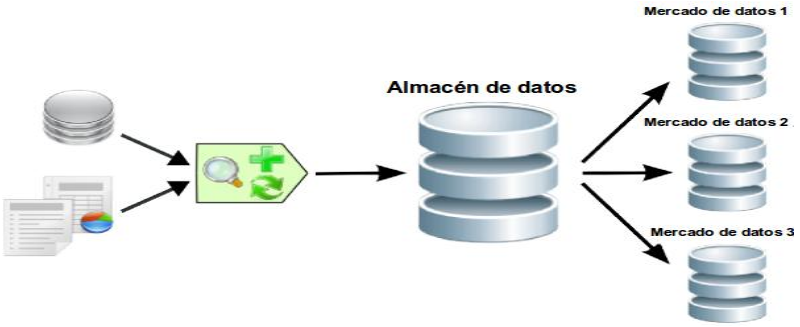

Figura 4: Enfoque descendente o top-down definido por Bill Inmon.

1.6.1 Propuesta de Metodología para el Desarrollo de Almacenes de Datos

Para guiar el proceso de desarrollo de los subsistemas de almacenamiento e integración LeukoCIM se selecciona la Propuesta de Metodología para el Desarrollo de Almacenes de Datos (Figura 5), que es una adaptación del ciclo de vida de la metodología de Kimball, la cual se encuentra fundamentada bajo los siguientes elementos: [13]

- ✓ Es una metodología madura y reconocida por los usuarios dedicados al tema.
- ✓ Permite crear los conceptos de hechos y dimensiones, lo que apoya en gran medida el desarrollo y el proceso de toma de decisiones.
- ✓ Propone la construcción del almacén de datos a través de la construcción de los mercados de datos departamentales, lo que constituye una buena estrategia y coincide con la división lógica de las empresas, entidades, organismos, entre otras instituciones.
- ✓ Permite ir presentando resultados parciales a los clientes en cortos plazos.
- ✓ Existe abundante documentación sobre la misma y se puede consultar a través de los servicios que brindan el grupo creador de la metodología.

Para complemento del modelo antes mencionado y por las características de trabajo de la universidad se tomó lo planteado por Leopoldo Zenaido Zepeda en su tesis de doctorado, incluir los casos de uso para guiar el proceso de desarrollo [14], y así lograr una mayor correlación con las tendencias y normas de la UCI. Se agrega además una etapa de prueba que permite comprobar la calidad de los productos que se desarrollan.

Figura 5: Ciclo de vida de la Metodología de desarrollo de Almacenes de datos.

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

1.7 Herramienta de modelado

Las herramientas de modelado se utilizan para representar los elementos claves del proceso de manera que sea posible alcanzar una mejor comprensión del mismo. [15] Las herramientas de Ingeniería Asistida por Computadora (*Computer Aided Software Engineering,* CASE por sus siglas en inglés), tienen como objetivo incrementar la productividad y calidad de los productos de software, mejorar la planificación del proyecto, así como reducir el tiempo y costo de su desarrollo. Existen varias herramientas CASE, entre las que se encuentran: ER Estudio, Rational Rose y Visual Paradigm.

1.7.1 Visual Paradigm 8.0

Visual Paradigm es una herramienta CASE de diseño que soporta los principales estándares de la industria donde se encuentra el Lenguaje de Modelado Unificado (*Unified Modeling Language*, UML por sus siglas en inglés). Ofrece un conjunto completo de herramientas para el desarrollo del *software*, necesarias para la captura de requisitos, *software* de planificación, planificación de controles, modelado de clases y de datos. [16]

En su versión 8.0 brinda la posibilidad de modelar numerosos tipos de diagramas de clases, permite la compatibilidad entre ediciones, la generación de código desde diagramas y la documentación asociada a cada etapa del proceso de desarrollo. Además presenta licencia gratuita y comercial, por lo que todas las personas pueden trabajar con dicha herramienta.

Se decidió utilizar Visual Paradigm 8.0 para UML porque es una herramienta CASE profesional que soporta todo el ciclo de vida del desarrollo de *software*. Además, brinda la posibilidad de modelar todo tipo de diagramas de clases, admite la compatibilidad entre ediciones, la documentación asociada a cada etapa del proceso de desarrollo y generar script para diferentes sistemas gestores de base de datos. Los desarrolladores lo utilizan para facilitar el modelado simultáneo, almacenar los archivos de proyectos y hacer un seguimiento de los cambios.

1.8 Sistema Gestor de Base de Datos

Un Sistema Gestor de Bases de Datos (DataBase Management System, en inglés DBMS) es un software que permite definir bases de datos, estructurar los datos que serán almacenados y la búsqueda de los mismos, ya sea de forma interactiva o a través de otras tecnologías. Es una herramienta efectiva que permite a varios usuarios acceder a la información al mismo tiempo. Brindan facilidades y un grupo de funciones con el objetivo de garantizar la confidencialidad, calidad, seguridad e integridad de los datos que contienen, así como el acceso fácil y relativamente rápido de los mismos. Dentro de los SGBD se pueden encontrar el PostgreSQL, Oracle, Sybase, MySQL, entre otros.

1.8.1 PostgreSQL 9.1

PostgreSQL es un SGBD objeto-relacional, distribuido bajo la licencia *Berkeley Software Distribution* (BSD), desarrollado en la Universidad de California, en el Departamento de Ciencias de la Computación de Berkeley. Brinda un control de concurrencia multi-versión (*Multiversion concurrency control*, MVCC por sus siglas en inglés) que permite trabajar con grandes volúmenes de datos, lo que beneficia en la construcción de los subsistemas de almacenamiento e integración LeukoCIM. La versión 9.1 es estable y segura, además de proporcionar la elaboración de consultas en SQL de forma gráfica.

Es el SGBD de código abierto más potente en el mercado y en sus últimas versiones ha mejorado varios factores que comparado con otros SGDB exponía ciertas desventajas. PostgreSQL utiliza un modelo cliente/servidor y usa multiprocesos en vez de multihilos para garantizar la estabilidad del sistema. Un fallo en uno de los procesos no afectará el resto y el sistema continuará funcionando. [17]

1.8.2 Administrador de Base de Datos. PgAdmin III 1.14.0

Un administrador de base de datos (*DataBase Administrator*, DBA por sus siglas en inglés) es fundamental para el desarrollo de una investigación de este tipo ya que son los responsables de la integridad y disponibilidad de los datos. Como parte de su función crean y configuran bases de datos relacionales. También se encargan de llevar a cabo el diseño de la distribución de los mismos y las soluciones de almacenamiento, el despliegue y monitorización de servidores de bases de datos. Algo muy importante y que a todas las instituciones les interesa es que garantizan la seguridad de las bases de datos, incluyendo *backups* y recuperación de desastres. A continuación se abordará sobre el Administrador de Base de Datos a utilizar en la construcción de los subsistemas.

PgAdmin III es una aplicación gráfica para administrar el SGBD PostgreSQL, siendo la más completa y popular con licencia *Open Source*[8]. [18] Este software fue diseñado para responder a las necesidades de todos los usuarios, desde la escritura de simples consultas SQL a la elaboración de bases de datos complejas. La interfaz gráfica es compatible con todas las características de PostgreSQL y facilita la administración. La aplicación también incluye un editor de la sintaxis SQL. PgAdmin III es capaz de gestionar versiones a partir de la PostgreSQL 7.3 ejecutándose en cualquier plataforma, así como versiones comerciales de PostgreSQL como *Pervasive Postgres*, *EnterpriseDB*, *Mammoth Replicator* y *SRA PowerGres*.

1.9 Herramientas para la integración de datos

El proceso de ETL es uno de los más importantes en la construcción de un mercado de datos, ya que es necesario que la información contenida represente la realidad del negocio, sea confiable y esté disponible en el momento que los usuarios y la organización la necesiten. Por la importancia que tienen estos procesos, es preciso contar con herramientas que permitan reducir tiempo y costos.

El desarrollo y la diversificación de las herramientas para los procesos de ETL, actualmente son crecientes, y se refleja en la amplia variedad de herramientas tanto comerciales como de código abierto. Entre las herramientas comerciales más conocidas se encuentran *Oracle Data Integration* (ODI) y *Oracle Warehouse Builder* (OWB), productos dentro de la familia de *Oracle Fusion Middleware*, las cuales hacen factible la optimización de inteligencia de negocios, almacén de datos y gestión de datos. OWB se fusiona con ODI para crear un sistema unificado de datos [19], siendo además la herramienta líder del sector en el diseño de los procesos de ETL. No se emplea en la investigación ya que esta necesita una herramienta que reconozca todas las fuentes de datos definidas en el negocio, así como que sea multiplataforma y de código abierto.

Después de realizar un estudio de las herramientas de código abierto para la integración de los datos, como son *Jaspersoft, Talend Open Studio, DataStage, Informatic Power Center* y *Pentaho Data Integration*, se ha determinado utilizar esta última en su versión 4.2.1, por las potencialidades que presenta, y *DataCleaner* en su versión 1.5.3, para el perfilado de los datos.

8 Open Source es el término con el que se conoce al software distribuido y desarrollado libremente.

1.9.1 Pentaho Data Integration 4.2.1

Pentaho Data Integration (PDI, también conocida como *Kettle*) es una de las soluciones más extendidas y mejor valoradas en el mercado, que reúne un conjunto de componentes que permiten modelar y ejecutar transformaciones sobre flujos de datos. Posee capacidades de integración de datos, entorno de diseño gráfico intuitivo y rico y una arquitectura altamente escalable, proporciona la solución ideal para cualquier tipo de integración de datos, análisis de negocio o proyectos con grandes capacidades de datos. [20]

Es una herramienta multiplataforma, lo que permite ejecutarla en cualquier sistema operativo. Está basada en dos tipos de objetos; las transformaciones que contienen una colección de pasos en un proceso de ETL y los trabajos que poseen una colección de transformaciones y/o trabajos. Además brinda soporte para metadatos, así como funciones que permiten operar con los campos en el flujo de datos, renombrando, calculando campos en función de otros, correlacionando valores y realizando búsquedas auxiliares en bases de datos. Ofrece soporte para operaciones de dimensiones lentamente cambiantes, permite ejecutar código *Java Script* dentro de las transformaciones e incorpora un evaluador de expresiones regulares. Hace uso de las tecnologías estándar *Java* y XML (*eXtensible Markub Language*).

Además esta herramienta es fácil de usar, brinda la posibilidad de copiar y leer del mismo fichero en paralelo, permitiendo maximizar la capacidad de entrada/salida en el entorno ETL. Añade un *debugger*[9] integrado diseñado para mejorar la productividad del desarrollador, ya que se pueden agregar puntos de ruptura condicionales en la ejecución de las transformaciones, dando la posibilidad de pausar y resumir la ejecución de la transformación, así como especificar el número de filas que se van a usar en las ejecuciones de prueba. Además, se pueden añadir registros personalizados. Cuenta con una gran comunidad de usuarios.

Presenta algunas desventajas ya que no cuenta con un componente de calidad de datos, no automatiza el proceso de separación y redistribución de datos para el procesamiento paralelo, además para realizar búsquedas de mayores volúmenes necesita utilizar una base de datos de búsqueda donde se ejecutan un gran número de sentencias SQL que frenan el rendimiento de la ETL.

En esta investigación el enfoque está dirigido a las herramientas de código abierto, pues aporta la posibilidad de manejar directamente el código fuente, modificarlo o adaptarlo a las necesidades individuales.

9 Es un programa usado para probar y depurar (eliminar los errores) de otros programas.

1.9.2 DataCleaner 1.5.3

El perfilado de los datos es una de las primeras tareas a realizar en el proceso de calidad de datos y consiste en realizar un análisis inicial sobre los datos de las fuentes, con el propósito de empezar a conocer su estructura, formato y nivel de calidad.

DataCleaner es una aplicación de código abierto para el perfilado, la validación y comparación de los datos, además permite mantener las bases de datos ordenadas. Estas actividades ayudan a administrar y supervisar la calidad de los datos con el fin de garantizar que la información sea útil y aplicable a su situación de negocio.

Según su creador Kasper Sorense, el sistema requiere *Java Runtime Enviroment* 5.0 o una versión superior y controladores de JDBC (*Java Data Base Connection*). La misma permite la evaluación del nivel de calidad de los datos contenidos en el sistema de información. Es una aplicación fácil de usar que genera sofisticados informes y gráficos que permiten a los usuarios determinar el nivel de calidad de los datos. Es utilizada, además, para identificar y analizar la estructura del origen de los mismos y combinar resultados y gráficos, creando vistas fáciles de interpretar para evaluar su calidad. [21].

Los perfiles de datos se utilizan para calcular y analizar diversas medidas importantes, basadas en los valores de los datos. La validación ofrece un resultado que puede ser interpretado para conocer la calidad de los datos. Esta herramienta soporta acceso de lectura a muchos tipos de almacenes de datos incluyendo: bases de datos compatibles con JDBC, hojas de cálculo, archivos XML, *Open Office Base* (ODB), entre otros.

Conclusiones del capítulo

Una vez concluido el estudio de las diferentes tecnologías para dar solución al problema planteado, se puede aseverar que los almacenes de datos constituyen un elemento primordial en la gestión de grandes volúmenes de información, auxiliándose en los mercados de datos para informaciones clasificadas específicamente. Se abordaron aspectos relacionados con esta tecnología y su relación con los subsistemas de almacenamiento e integración LeukoCIM, así mismo se opta por esta tecnología ya que constituye una solución viable en cuanto al almacenamiento y estructura de los EC del CIM.

Para la implementación de la solución se concluye que:

- La Propuesta de Metodología para el Desarrollo de Almacenes de Datos permite guiar el proceso de desarrollo de los subsistemas de almacenamiento e integración LeukoCIM por las diferentes etapas del ciclo de vida.

- Las herramientas que ayudarán a construir el subsistema de almacenamiento son: Visual Paradigm en su versión 8.0 que posibilita la generación de los diagramas necesarios para modelar el funcionamiento del sistema propuesto. Así como el SGBD PostgreSQL 9.1 y el administrador de base de datos PgAdmin III 1.14.0.

- Se decide emplear el sistema de almacenamiento ROLAP, ya que el SGBD PostgreSQL 9.1 soporta la administración de los grandes volúmenes de datos que utiliza el sistema.

- Para los procesos de integración se decide utilizar el DataCleaner 1.5.3 para el perfilado de los datos y el Pentaho Data Integration 4.2.1 para extraer, transformar y cargar los datos.

Capítulo 2. Análisis y Diseño

Introducción

En este capítulo se tienen en cuenta las necesidades del negocio, así como las reglas del mismo. Se especifican los requisitos de información, los funcionales y los no funcionales. Se conforma el diagrama de casos de uso del sistema. Por otra parte se define la arquitectura base de los subsistemas de almacenamiento e integración LeukoCIM y se diseñan ambos. Se construye la matriz bus, el modelo de datos donde se determinan: las dimensiones, los hechos y las medidas. Por último se realiza el esquema de seguridad y se describe la política de respaldo y recuperación, definiendo los roles y permisos.

2.1 Necesidades del negocio

Las necesidades del negocio están en correspondencia con lo que el cliente necesita, en este caso el CIM. Para cumplir con lo planteado por los especialistas de este centro es muy importante realizar un análisis previo para comprender los procesos de toma de decisiones.

Para identificar las necesidades de la organización se le efectuaron entrevistas al cliente, donde se identificaron diversos problemas con la gestión de los datos relacionados con los ensayos clínicos realizados a pacientes que presentan cáncer, con la utilización del producto LeukoCIM.

Como parte de las necesidades de información se acordó almacenar toda la información contenida de los diferentes ensayos, por lo que se establecieron tres grupos:

- ✓ **Leuko-Niños y Leuko-Adultos:** ensayos realizado a pacientes niños y adultos, contiene los datos de inclusión y los eventos adversos presentados por estos.
- ✓ **Leuko-SIDA:** ensayo realizado a pacientes que presentan el SIDA. Contiene los datos asociados con la inclusión, los eventos adversos, el tratamiento concomitante con el producto en cuestión y la evaluación final.
- ✓ **Leuko-Neutropenia:** ensayo realizado a pacientes que presentan neutropenia, contiene los eventos adversos y la evaluación realizada.

En el ámbito del diseño de un mercado de datos, los grupos de información se corresponden con los temas de análisis que contienen toda la información solicitada por los especialistas.

2.2 Especificación de los requisitos

El levantamiento de los requisitos consiste en identificar las necesidades de información de la organización, las características y cualidades que debe poseer el sistema. En las soluciones de almacenes de datos se identifican tres tipos de requisitos los cuales son: requisitos de información, requisitos funcionales y no funcionales.

2.2.1 Requisitos de información

Los requisitos de información (RI) describen la información y los datos que el sistema debe almacenar para satisfacer a los clientes. Estos se definen a partir de las necesidades de información identificadas en el negocio, que permitan el análisis del comportamiento de los indicadores a medir según los objetivos y metas de la organización.

Los RI identificados durante el proceso de análisis fueron clasificados por el tipo de información. Estos se encuentran de forma íntegra dentro del Expediente de Proyecto de los Subsistemas de almacenamiento e integración LeukoCIM, en el artefacto "DATEC_CIM_0113_Especificación de Requisitos de Software.doc". El objetivo fundamental de dichos requisitos es mantener disponible la información referente a los ensayos. A continuación se relacionan los correspondientes al tema de análisis "Leuko-SIDA":

RI1. Obtener la cantidad de pacientes con SIDA que sufrieron eventos adversos atendiendo al tipo de evento adverso, grado y causalidad del evento.

RI2. Obtener la cantidad de pacientes con SIDA que se incluyeron al ensayo atendiendo al hospital, provincia, indicación y si es incluido en el tratamiento.

RI3. Obtener la cantidad de pacientes con SIDA que se incluyeron al ensayo atendiendo a la fecha de inclusión, edad sexo, raza, peso y dosis de LeukoCIM indicada.

RI4. Obtener la cantidad de pacientes con SIDA que se incluyeron al ensayo atendiendo a las enfermedades marcadoras.

RI5. Obtener la cantidad de pacientes con SIDA que se incluyeron al ensayo atendiendo a si recibe QT/RT (Quimioterapia o Radioterapia), ciclo de QT/RT, tratamiento antirretroviral.

RI6. Obtener la cantidad de pacientes con SIDA que se incluyeron al ensayo atendiendo al tipo de neutropenia, el grado y su tratamiento.

RI7. Obtener la cantidad de pacientes con SIDA que se incluyeron al ensayo atendiendo a los exámenes de laboratorio.

RI8. Obtener la cantidad de pacientes con SIDA que se incluyeron al ensayo atendiendo a si se evalúo, interrumpió el tratamiento y si presenta tratamiento concomitante.

RI9. Obtener la cantidad de pacientes con SIDA que se incluyeron al ensayo atendiendo a si ocurrió algún evento adverso y el tratamiento al mismo.

RI10. Obtener la cantidad de pacientes con SIDA que se evaluaron atendiendo a si asistieron a la evaluación, fecha de la evaluación y número de dosis de LeukoCIM administradas.

RI11. Obtener la cantidad de pacientes con SIDA que se evaluaron atendiendo a los exámenes de laboratorio.

RI12. Obtener la cantidad de pacientes con SIDA que se evaluaron atendiendo al régimen de administración del tratamiento y si presenta tratamiento concomitante.

RI13. Obtener la cantidad de pacientes con SIDA que se evaluaron atendiendo a si recibirá el próximo ciclo de QT/RT, tipo de neutropenia y tratamiento.

RI14. Obtener la cantidad de pacientes con SIDA que se evaluaron atendiendo a si ocurrió algún evento adverso y si interrumpió el tratamiento.

RI15. Obtener la cantidad de pacientes con SIDA que presentaban tratamiento concomitante atendiendo al tratamiento, razón, fecha de inicio y fecha de fin del tratamiento.

2.2.2 Requisitos funcionales

Los requisitos funcionales (RF) describen las funcionalidades que el equipo de desarrollo debe construir. Estos requisitos incluyen las funcionalidades que deben implementarse en los tres subsistemas que se desarrollan en soluciones de almacenes de datos.

La actual investigación incluye dos subsistemas, por lo que se especifican las funcionalidades referentes a estos. Atendiendo a que los ensayos clínicos del producto LeukoCIM son cerrados, no se especifica la persistencia de la información ni las vistas integradas, por tanto no hay requisitos relacionados con el subsistema de almacenamiento. Los requisitos asociados al subsistema de integración de datos se describen a continuación:

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 (RNF) describen las propiedades y cualidades que debe tener la solución. Representan las características del producto. Estos requisitos se clasifican en un grupo de categorías que dependen de las características del negocio. Se definen siete RNF para los Subsistemas de almacenamiento e integración LeukoCIM, los cuales se encuentran descritos en el Expediente de Proyecto de los Subsistemas de almacenamiento e integración LeukoCIM, específicamente en el artefacto "DATEC_CIM_0113_Especificación de Requisitos de Software.doc". A continuación se mencionan algunos de ellos.

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

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.

RNF3: Lograr la homogeneidad de la estructura de los elementos definidos en el almacén. Las estructuras del almacén de datos deben tener un nombre estándar teniendo en cuenta el tipo de estructura que sea. Se realizaron convenciones de nombrado con el objetivo de manejar un vocabulario común en todo el almacén de datos, permitiendo un entendimiento claro y conciso por parte de los desarrolladores.

Finalmente se identificaron 23 requisitos de información, dos funcionales y siete no funcionales, que permitirán construir los subsistemas de almacenamiento e integración LeukoCIM.

2.3 Reglas del negocio

Las reglas del negocio describen políticas que deben cumplirse o condiciones que deben satisfacerse para cumplir con los aspectos del negocio.

El proceso de especificación implica que hay que identificarlas dentro del negocio, evaluar si son relevantes dentro del campo de acción que se está modelando e implementarlas en la propuesta de solución.

La metodología llevada a cabo presenta cuatro tipos de reglas del negocio, que se pueden encontrar en su totalidad en el Expediente de Proyecto de los Subsistemas de almacenamiento e integración LeukoCIM, específicamente en el artefacto "DATEC_CIM_Reglas del Negocio.doc". Estas se relacionan 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.

RN4. Las fechas serán de tipo date en formato (aaaa/mm/dd).

Reglas de transformación

RN5. Los valores del sexo tomarán el valor de 1 para Masculino y 2 para Femenino.

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

RN7. Los valores del grado del evento adverso según la Organización Mundial de la Salud (OMS) serán: 1 para Normal, 2 para Ligero, 3 para Moderado, 4 para Severo, 5 Que amenaza la vida, 6 Que causa la muerte y 7 para Muy severo.

RN8. Los valores de causalidad del evento adverso serán: 1 Definitiva, 2 Muy probable, 3 Probable, 4 Posible, 5 No relacionado, 6 Desconocido y 7 Remota.

RN9. La dimensión provincia contendrá el código del hospital proveniente de la fuente acompañada de la descripción tomada de los protocolos.

RN10. La dimensión hospital contendrá el código del hospital proveniente de la fuente junto con la descripción tomada de los protocolos.

RN11. Los valores de la indicación serán: 1 profilaxis primaria, 2 profilaxis secundaria y 3 tratamiento.

RN12. El valor de la ocurrencia de algún evento adverso será: 1 Sí y 2 No.

RN13. El valor de inclusión al tratamiento será: 1 Secundaria a inferior, 2 QT/RT y 3 causa no determinada.

RN14. El valor para el tipo de neutropenia será: 1 Febril, 2 Afebril y 3 No.

RN15. El valor para el grado de neutropenia cuando no se conozca se pondrá 0.

RN16. El valor del régimen de administración será: 1 Ambulatorio y 2 Ingresado y -1 No especificado.

RN17. Cuando no aparezca valor o es negativo en los exámenes de laboratorio se tomará como No disponible.

RN18. En las dimensiones numéricas cuando aparezca nulo, se sustituirá con la moda, media, mediana o promedio.

RN19. En las dimensiones de cadenas cuando aparezca nulo, se pondrá No especificado.

2.4 Casos de uso del sistema

Un caso de uso (CU) constituye una secuencia de interacciones que se desarrollan entre los actores y un sistema, en respuesta a un evento que inicia un actor sobre el propio sistema. Los requisitos de información y los requisitos funcionales son agrupados en casos de uso de información y casos de uso funcionales respectivamente, los cuales aportan una idea enfocada al correcto funcionamiento del sistema.

2.4.1 Actores del sistema

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

Administrador ETL: es el responsable de llevar a cabo los procesos de extracción, transformación y carga de los datos del sistema fuente.

2.4.2 Casos de uso de información

Los casos de uso de información (CUI) agrupan un conjunto de requisitos de información teniendo en cuenta los conceptos del negocio que manejan, fundamentalmente por temas de análisis. La presente investigación cuenta con seis CUI agrupados por el tipo de información.

CUI1. Mantener disponible la información de los eventos adversos ocasionados por la aplicación del producto LeukoCIM a pacientes niños, adultos, con SIDA y neutropenia.

CUI2. Mantener disponible la información del tratamiento concomitante con el producto LeukoCIM a pacientes que presentan SIDA.

CUI3. Mantener disponible la información de la inclusión a la aplicación del producto LeukoCIM de pacientes niños y adultos.

CUI4. Mantener disponible la información de la inclusión a la aplicación del producto LeukoCIM de pacientes con SIDA.

CUI5. Mantener disponible la información de la evaluación luego de la aplicación del producto LeukoCIM a pacientes con SIDA.

CUI6. Mantener disponible la información de la evaluación luego de la aplicación del producto LeukoCIM a pacientes con neutropenia.

La descripción de los CUI se encuentran en el artefacto "DATEC_CIM_0114_Especificación de casos de uso.doc" dentro del Expediente de Proyecto de los Subsistemas de almacenamiento e integración LeukoCIM.

2.4.3 Casos de uso funcionales

Los casos de uso funcionales (CUF) agrupan los requisitos funcionales definidos para los diferentes subsistemas que componen la solución. La investigación cuenta con dos CUF.
CUF1: Realizar la extracción de los datos.
CUF2: Realizar la transformación y carga de los datos.

A continuación se muestra la especificación de los casos de uso funcionales.

Tabla 1: Descripción del CUF: Realizar la extracción de los datos.

Objetivo:	Realizar la extracción de los datos.
Actores:	Administrador ETL.

Resumen	El caso de uso inicia cuando el actor selecciona los datos a extraer. Se extraen los mismos de la fuente. Finaliza cuando los datos se encuentran en la base de datos.
Complejidad:	Media
Prioridad:	Media
Precondiciones:	Disponibilidad de las fuentes.
Poscondiciones:	Los datos de la fuente correspondiente han sido extraídos y almacenados en la base de datos.

Flujo Normal de Eventos	
Acción del Actor	**Respuesta del sistema**
1. El administrador de ETL realiza la conexión a la fuente correspondiente.	2. Responde a la solicitud de conexión.
3. El administrador de ETL selecciona la estructura o archivo a extraer.	
4. El administrador de ETL realiza la extracción de los datos.	5. Ejecuta la extracción de los datos. Finaliza el caso de uso.

Flujos Alternos	
Acción del Actor	**Respuesta del sistema**
	2.1. No responde a solicitud de conexión.
	2.2. Notifica el error al administrador de ETL. Vuelve al paso 1 del Flujo Normal de Eventos.
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: Descripción del CUF: Realizar la transformación y carga de los datos.

Objetivo:	Realizar la transformación y carga de los datos.
Actores:	Administrador ETL.
Resumen	El caso de uso inicia cuando el actor desea realizar la transformación y carga de los datos. Finaliza cuando los datos son insertados satisfactoriamente en la base de datos.
Complejidad:	Media
Prioridad:	Media
Precondiciones:	Los datos se encontraron correctamente extraídos de la fuente y las estructuras del mercado de datos se encontraron disponibles para su uso. En la base de datos debe existir una estructura para almacenar la información.
Poscondiciones	Los datos han sido transformados y cargados en el mercado de datos.
Flujo Normal de Eventos	
Acción del Actor	**Respuesta del sistema**
1. El administrador de ETL selecciona las estructuras de la fuente que desea transformar.	
2. El administrador de ETL carga los datos seleccionados en memoria.	
3. El administrador de ETL aplica las transformaciones pertinentes y genera datos de auditoría.	
4. El administrador de ETL carga los datos en el	5. Ejecuta la consulta. Finaliza el caso de

mercado de datos.	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

Los diagramas de casos de uso del sistema (CUS) son artefactos narrativos que describen, bajo la forma de acciones y reacciones, el comportamiento del sistema desde el punto de vista del usuario. Es un proceso que da un resultado de valor para un actor determinado y una secuencia de actividades a automatizar. En la presente investigación se determinaron ocho CUS (Figura 6).

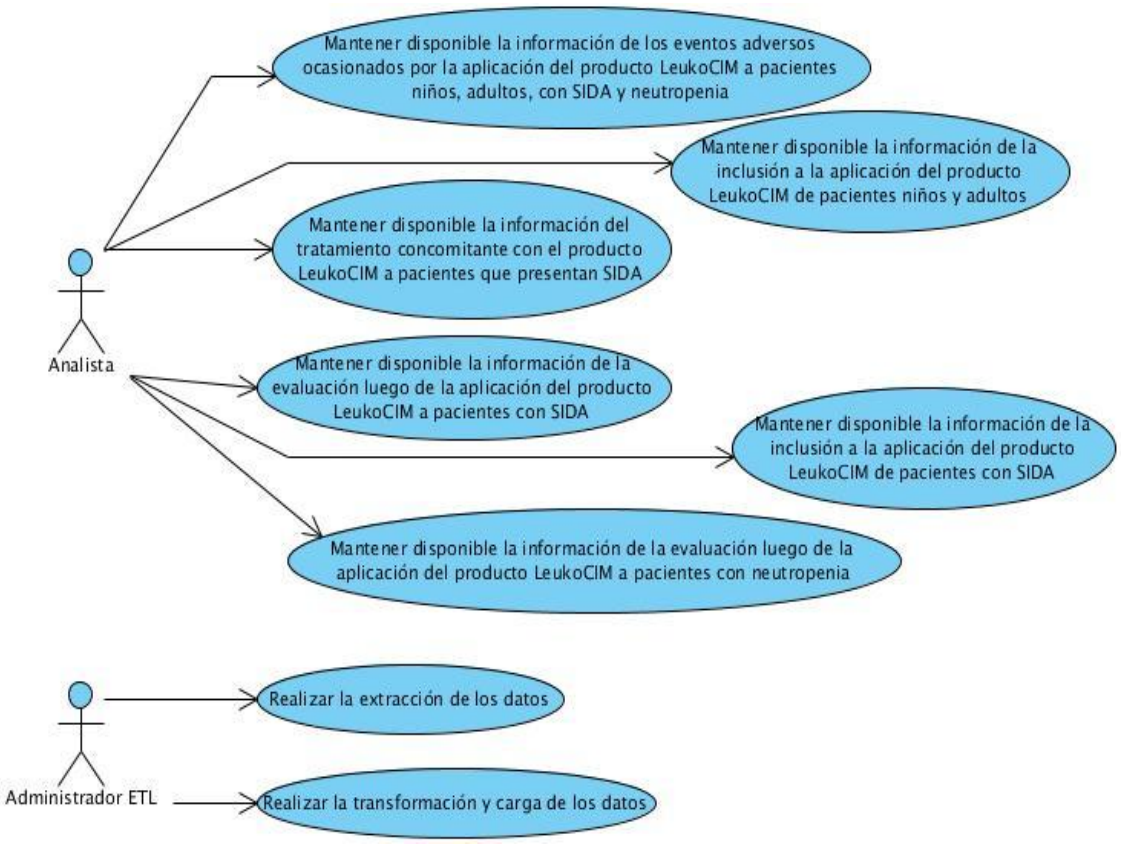

Figura 6: Diagrama de CUS.

2.5 Definición de la arquitectura base de un mercado de datos

La arquitectura base de un mercado de datos define los aspectos arquitectónicos de la solución. Esta presenta nueve vistas, agrupadas en tres áreas principales, en este caso se tendrán en cuenta las vistas de la arquitectura del sistema.

Vistas de Arquitectura de Sistema

- ✓ **Vista conceptual del negocio:** describe el marco estructural de la arquitectura del negocio, donde se identifican los conceptos de información que serán objeto de análisis para la organización. Es una adaptación de la Vista de procesos, considerando que para el desarrollo de los almacenes de datos es fundamental tener una vista integrada de los conceptos de información que maneja la organización y que son de interés para su análisis.
- ✓ **Vista de Sistema:** Se definen los subsistemas, componentes y paquetes que conforman la solución.
- ✓ **Vista de Datos:** Se definen los escenarios y patrones aplicables relacionados con el manejo de los datos y los metadatos asociados a los procesos de integración.
- ✓ **Vista de Integración:** Se describe como se integran los diferentes componentes y subsistemas que componen el producto final
- ✓ **Vista de Presentación:** Se define los elementos de la Arquitectura de Información de la solución.

La arquitectura definida para el desarrollo de la solución consta de dos subsistemas: almacenamiento e integración, ubicados en dos niveles y la fuente de datos (Figura 7).

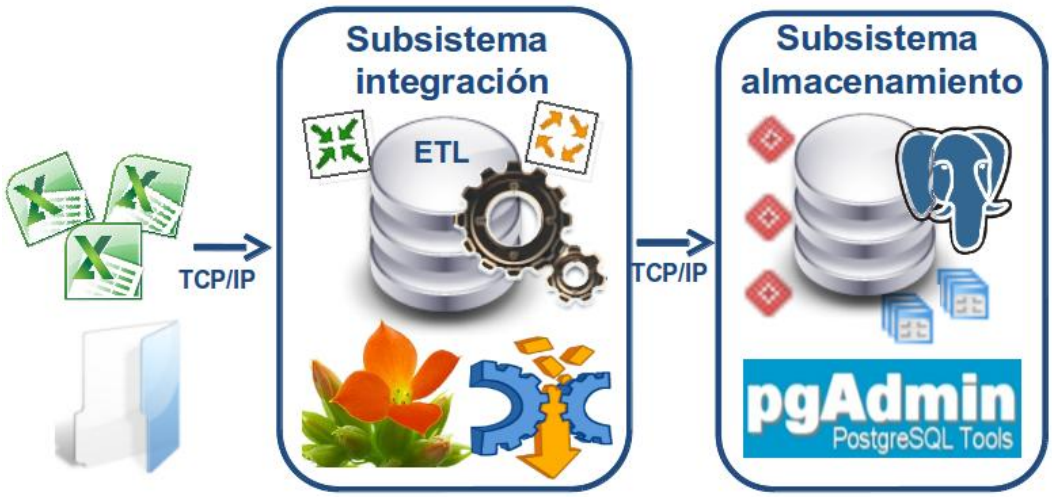

Figura 7: Arquitectura del sistema.

En el primer nivel se encuentra el subsistema de integración, el cual se abastece de las diferentes fuentes de datos y se encarga de llevar a cabo los procesos que extraen, integran y transforman la información para su almacenamiento. Los usuarios que acceden a este subsistema son los encargados de la administración de dichos procesos.
Por su parte, el subsistema de almacenamiento recibe la información manipulada durante la extracción, transformación y carga, finalmente se almacena en una base de datos soportada por el SGBD PostgreSQL y administrada por los usuarios autorizados mediante la herramienta PgAdmin III.

2.6 Diseño de la solución

El diseño es un refinamiento del análisis que tiene en cuenta los requisitos no funcionales y cómo cumple el sistema sus objetivos. El diseño debe ser suficientemente claro para que el sistema pueda ser implementado sin ambigüedades. Se modela el sistema incluyendo la arquitectura, para que soporte los requisitos funcionales y no funcionales, así como las restricciones que se le suponen.

La solución propuesta abarca dos subsistemas relacionados con el producto LeukoCIM. El subsistema de almacenamiento es el encargado de contener toda la información correspondiente a los ensayos clínicos realizados. Por otra parte el subsistema de integración se encarga de extraer los datos del sistema fuente, estandarizarlos, limpiarlos y cargarlos en el almacén.

2.6.1 Diseño del subsistema de almacenamiento

Para realizar el diseño del subsistema de almacenamiento es de vital importancia identificar las dimensiones y los hechos con sus medidas asociadas, ya que son el punto de partida. Se debe definir además, una política de respaldo y recuperación que garantice la integridad de los datos almacenados.

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

2.6.1.1 Dimensiones

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

Constituyen las perspectivas de análisis de la información y presentan entre sus características principales la definición de jerarquías entre sus atributos, cuyo objetivo es modelar explícitamente la forma en que se puede consolidar el proceso de análisis de la información. En el desarrollo de la solución se identificaron 24 dimensiones:

1. **Dimensión ensayo** (dim_ensayo): contiene el nombre de los cuatro ensayos realizados del producto LeukoCIM.
2. **Dimensión sexo** (dim_sexo): contiene el sexo del paciente.
3. **Dimensión edad** (dim_edad): contiene la edad del paciente en años.
4. **Dimensión peso** (dim_peso): posee el peso del paciente en kilogramos (Kg).
5. **Dimensión talla** (dim_talla): contiene la talla del paciente en centímetros (cm).
6. **Dimensión raza** (dim_raza). contiene la raza del paciente.
7. **Dimensión enfermedad de base** (dim_enfermedad_base): describe la enfermedad con la que el paciente comenzó el ensayo clínico.
8. **Dimensión tratamiento** (dim_tratamiento): describe los medicamentos aplicados al paciente.
9. **Dimensión razón** (dim_razon): describe la razón por la cual el paciente recibe el tratamiento.
10. **Dimensión hospital** (dim_hospital): describe el hospital donde se atiende el paciente.
11. **Dimensión provincia** (dim_provincia): describe la provincia donde el paciente se atiende.
12. **Dimensión tiempo** (dim_tiempo): describe el tiempo para enmarcar la información almacenada y organiza la información atendiendo al momento en que fue captada.
13. **Dimensión tipo de eventos adversos** (dim_tipo_evento_adverso): Describe el tipo de evento adverso que sufrió el paciente durante el ensayo.
14. **Dimensión grado del evento adverso** (dim_grado_evento_adverso): Describe el grado de intensidad de los eventos adversos según la OMS.
15. **Dimensión causalidad evento adverso** (dim_causalidad_evento_adverso): Describe la relación de intensidad y causalidad de los eventos adversos.

16. **Dimensión examen de laboratorio** (dim_examen_laboratorio): describe los distintos exámenes que se le realizan al paciente para evaluar la seguridad.
17. **Dimensión régimen de administración** (dim_reg_admin): describe si el paciente recibe el tratamiento de forma hospitalizada o ambulatoria.
18. **Dimensión ciclo de QT/RT** (dim_ciclo): describe la cantidad de ciclos de QT/RT que el paciente se ha proporcionado.
19. **Dimensión tipo de tratamiento** (dim_indicacion): describe el tipo de tratamiento que lleva el paciente.
20. **Dimensión dosis de leuko administrada** (dim_dosis_leuko_admin): describe la cantidad de dosis de LeukoCIM administradas por el paciente.
21. **Dimensión dosis de leuko indicada** (dim_dosis_leuko_ind): describe la cantidad de dosis de LeukoCIM indicadas al paciente.
22. **Dimensión enfermedad marcadora del SIDA** (dim_enfermedad_marcadora): describe la enfermedad marcadora que presenta el paciente con SIDA.
23. **Dimensión inclusión** (dim_inclusion_trat): describe la causa por la que el paciente fue incluido al ensayo.
24. **Dimensión neutropenia** (dim_neutropenia): describe el tipo de neutropenia, así como el grado y el tratamiento a la misma.

2.6.1.2 Dimensiones degeneradas

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. Esto sucede cuando un campo posee el mismo nivel de granularidad que los datos almacenados en una tabla de hechos, y que por lo tanto no se pueden realizar agrupaciones o sumarizaciones a través del mismo. Un ejemplo de este tipo de dimensión en la investigación es “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. Por tanto se utiliza este tipo de dimensiones, incluyendo los campos en las tablas de

hechos con el objetivo de eliminar la duplicación de los datos y simplificar las consultas. Hay presentes un total de nueve dimensiones degeneradas.

2.6.1.3 Hechos

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

Las tablas de hechos diseñadas no almacenan las medidas numéricas porque son hechos unitarios y estos contienen las llaves asociadas a cada una de las dimensiones con que se relacionan. Para el desarrollo de la solución se identificaron seis hechos:

1. **Hecho eventos adversos de ensayos** (hech_ninno_adulto_sida_neutrop_ea): contiene la cantidad de pacientes que presentaron eventos adversos en la realización de los cuatro ensayos clínicos del producto LeukoCIM.
2. **Hecho tratamiento concomitante de SIDA** (hech_sida_trat_concom): contiene la cantidad de pacientes que contaron con un tratamiento junto con el LeukoCIM y que estuvieron inmersos en el ensayo clínico realizado a pacientes con SIDA.
3. **Hecho inclusión de niños y adultos** (hech_ninno_adulto_inclus): contiene la cantidad de pacientes que se incluyeron en el ensayo clínico del producto LeukoCIM realizado a niños y adultos.
4. **Hecho evaluación de neutropenia** (hech_neutrop_eval): contiene la cantidad de pacientes que se evaluaron en el ensayo clínico del producto LeukoCIM realizado a pacientes con neutropenia.
5. **Hecho evaluación de SIDA** (hech_sida_eval): contiene la cantidad de pacientes que se evaluaron en el ensayo clínico del producto LeukoCIM realizado a pacientes con SIDA.
6. **Hecho inclusión de SIDA** (hech_sida_inclus): contiene la cantidad de pacientes que se incluyeron en el ensayo clínico del producto LeukoCIM realizado a pacientes con SIDA.

2.6.1.4 Matriz bus o matriz dimensional

La matriz bus especifica la relación entre los hechos y las dimensiones, donde las columnas contienen las tablas de hechos y las filas las dimensiones, la intersección de una fila con una columna especifica si hay relación entre una tabla de hechos y una dimensión.

Tabla 3: Matriz dimensional.

Dimensiones	Hechos					
	H1	**H2**	**H3**	**H4**	**H5**	**H6**
ensayo	X		X			
tipo evento adverso	X					
grado evento adverso	X					
causalidad evento adverso	X					
tratamiento		X	X			
razón		X				
hospital			X	X		X
provincia			X	X		X
edad			X			
sexo			X	X		X
peso			X	X		X
talla			X			
raza			X	X		X
indicación			X	X		X
enfermedad base			X	X		
dosis leuko indicada				X		X
dosis leuko administrada				X	X	
examen laboratorio				X	X	X
régimen administración					X	

neutropenia					X	X
ciclo						X
inclusión tratamiento						X
enfermedad marcadora						X
tiempo	X	X	X	X	X	X

Leyenda:

H1: Hecho eventos adversos de ensayos.

H2: Tratamiento concomitante de SIDA.

H3: Inclusión de niños y adultos.

H4: Evaluación de neutropenia.

H5: Evaluación de SIDA.

H6: Inclusión de SIDA.

Realizar la matriz dimensional permite conocer que los seis hechos definidos para el modelo de datos, comparten dimensiones, sin embargo no existen dos o más hechos que se relacionen con exactamente las mismas dimensiones. Esto indica que pudo verificarse la inexistencia de solapamiento entre hechos.

2.6.1.5 Topología

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

Esquema en estrella: está formado por una tabla de hechos y una tabla para cada dimensión. (Ver Figura 8 a).

Esquema copo de nieve: es una variante del esquema en estrella, donde las dimensiones son normalizadas en dependencia de la jerarquía existente, pero la estructura de una única tabla de hechos se mantiene. (Ver Figura 8 b).

Constelación de hechos: está compuesto por diversos esquemas de estrella, con la particularidad de que varias tablas de hechos comparten algunas tablas de dimensiones. (Ver Figura 8 c).

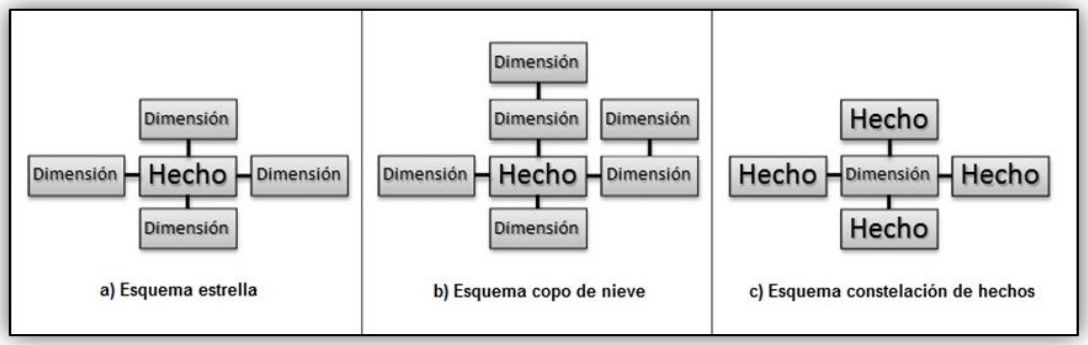

Figura 8: Esquemas del modelo dimensional.

2.6.1.1 Modelo de datos

Una vez definidas dentro del negocio las dimensiones, hechos y medidas se procede a la estructuración del modelo dimensional.

En la Figura 9 se muestra el modelo de datos diseñado para el desarrollo de los subsistemas de almacenamiento e integración LeukoCIM, donde se evidencia el uso de la topología **constelación de hechos**, atendiendo a que existen varias tablas de hechos que comparten algunas de sus dimensiones.

Para una mayor comprensión del modelo de datos puede consultarlo en el artefacto "DATEC_CIM_Especificación del modelo de datos.doc" dentro del Expediente de Proyecto de los Subsistemas de almacenamiento e integración LeukoCIM.

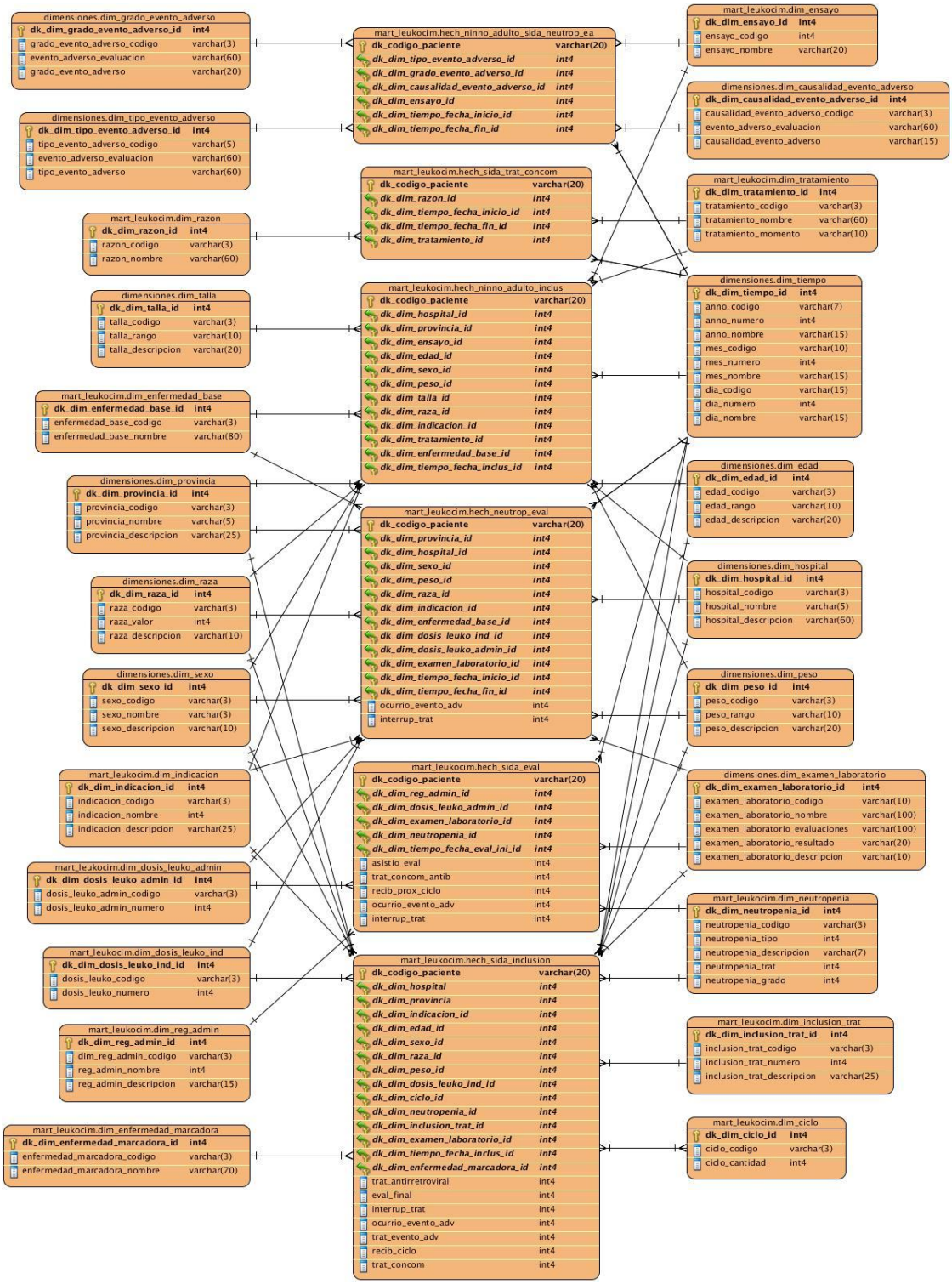

Figura 9: Modelo de datos.

2.6.2 Diseño del subsistema de integración

El subsistema de integración comprende el perfilado de los datos y la extracción de los mismos desde los sistemas fuentes, los cuales sufren un conjunto de transformaciones. Conforma estos de manera que fuentes separadas puedan ser aprovechadas conjuntamente, y finalmente hace su entrega en un formato listo para el almacenamiento, de forma tal, que permita a los desarrolladores construir la capa de presentación de la aplicación y los especialistas puedan tomar decisiones o realizarle técnicas de minería de datos a los que están almacenados. El perfilado, el diccionario de datos y el diseño de las transformaciones constituyen elementos esenciales para lograr el diseño del subsistema de integración.

2.6.2.1 Perfilado de datos

El perfilado de datos es el análisis que se le realiza, lo que permite comprender su contenido, estructura, calidad y dependencias. Se realiza con el objetivo de conocer el estado en que se encuentran los datos que próximamente se extraerán de las fuentes, así como administrar y supervisar la calidad de los mismos. Se verifica la existencia de valores nulos, distintos y duplicados, permitiendo definir nuevas reglas del negocio que pasarían a ser las reglas de trasformación aplicadas durante la implementación del subsistema de integración. De esta manera se garantiza que la información sea útil y aplicable a la situación del negocio.

Al realizar el perfilado de los datos a las fuentes de los ensayos clínicos del producto LeukoCIM se identificó que los tipos de datos son en su mayoría enteros y cadenas (Figura 10), además existe la presencia de fechas desde el año 2004 hasta el 2009 con el formato aaaa/mm/dd.

Se analizaron todos los campos de la fuente, detectando diversos errores descritos en el artefacto "DATEC_CIM_Perfil de los Datos.doc" en el Expediente de Proyecto de los Subsistemas de almacenamiento e integración LeukoCIM. También se encuentra reflejado de los diferentes campos la longitud de la cadena, la cantidad de valores nulos o negativos y los valores mínimos y máximos. Por ejemplo en el ensayo del SIDA en el Excel "EvalFinal.xls" existen 76 campos negativos y cuatro nulos, en "TTOCONC.xls" 103 campos nulos, en 'Inclusion.xls" 92 negativos y 100 nulos, lo que permitió crear nuevas RN que serán solucionadas en la implementación de las transformaciones.

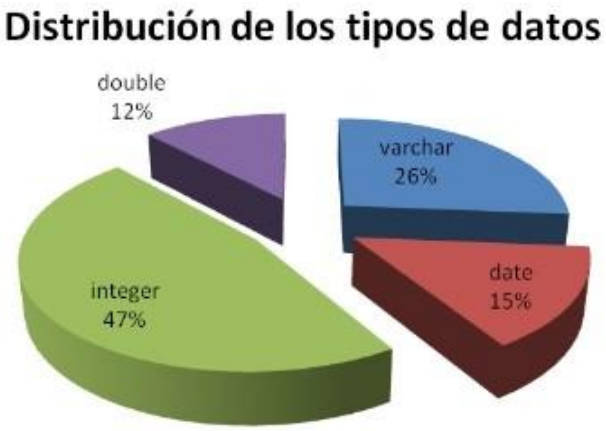

Figura 10: Distribución de los tipos de datos.

2.6.2.2 Diccionario de datos

En el diccionario de los datos se describen cada una de las variables contenidas en la fuente de datos, especificando el significado que tienen en el negocio y los posibles valores que pueden tomar. A continuación se muestra el nombre de las variables y su descripción.

Tabla 4: Diccionario de datos.

Nombre de las variables	Descripción de las variables
tipo evento adverso	Contiene el tipo de evento adverso presentado por los pacientes
grado evento adverso	Contiene el grado de los eventos adversos de los pacientes
causalidad evento adverso	Contiene la causa de los eventos adversos de los pacientes
hospital	Se refiere al hospital donde se atiende el paciente
provincia	Se refiere a la provincia donde se encuentra el hospital
sexo	Se refiere al sexo de los pacientes
peso	Se refiere al peso de los pacientes en kg
raza	Se refiere a la raza de los pacientes
indicación	Se refiere al tipo de tratamiento aplicado a los pacientes
tiempo	Se refiere al día, mes y año de realización de los ensayos
edad	Se refiere a la cantidad de años de los pacientes

tratamiento	Se refiere a los medicamento que se le aplicarán a los pacientes
enfermedad de base	Se refiere a la enfermedad con que los pacientes comenzaron el ensayo
talla	Se refiere a la talla de los pacientes expresada en cm
examen de laboratorio	Se refiere a los exámenes de laboratorio realizados a los pacientes
dosis leuko indicada	Es la dosis indicada del producto LeukoCIM a los pacientes
dosis leuko administrada	Es la dosis administrada por los pacientes
razón	Es la razón de uso del medicamento orientado
neutropenia	Es el tipo de neutropenia, contiene el grado y si es tratada
ciclo de QT/RT	Es la cantidad de ciclos de QT/RT que el paciente ha realizado
inclusión al tratamiento	Es la causa por la que los pacientes se incluyen a los ensayos
régimen de administración	Dice el régimen de administración del tratamiento de los pacientes
enfermedad marcadora	Dice la enfermedad marcadora del SIDA que presentan los pacientes

En el artefacto "DATEC_CIM_Diccionario de Datos.xls" dentro de Expediente de Proyecto de los Subsistemas de almacenamiento e integración LeukoCIM se encuentra cada una de las variables descritas con los posibles valores que pueden tomar.

2.6.2.3 Diseño general de las transformaciones

Una vez que se conoce la estructura, contenido y fiabilidad de los datos, se procede a realizar el diseño de las transformaciones. Estos diseños pueden variar a la hora de la implementación de las mismas, porque durante el proceso de desarrollo suelen surgir situaciones con los datos y se llevan a cabo diversas estrategias para resolverlas.

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

En el diseño general para implementar cada una de ellas lo primero que se realiza es la extracción de los datos desde la fuente (Excel), obteniéndose diferentes indicadores. Posteriormente se procede a la limpieza y transformación de los datos, aplicando las reglas del negocio. Luego de cargar las dimensiones se buscan los identificadores de las mismas. Se validan los tipos de datos con el objetivo que tengan la calidad requerida, si existen problemas con la validación se almacenan en un Excel de errores; de lo contrario se cargan los datos en la tabla de hechos correspondiente. Inmediatamente se obtiene la información necesaria del sistema para generar los metadatos técnicos, que guardará el nombre del fichero fuente, el nombre de la transformación, el nombre del destino o el hecho y la fecha de ejecución de la transformación. Por último se inserta o actualiza dicha información en la tabla md_carga_histórica. En la Figura 11 se puede apreciar el diseño general para la carga de los hechos de los Subsistemas de almacenamiento e integración LeukoCIM.

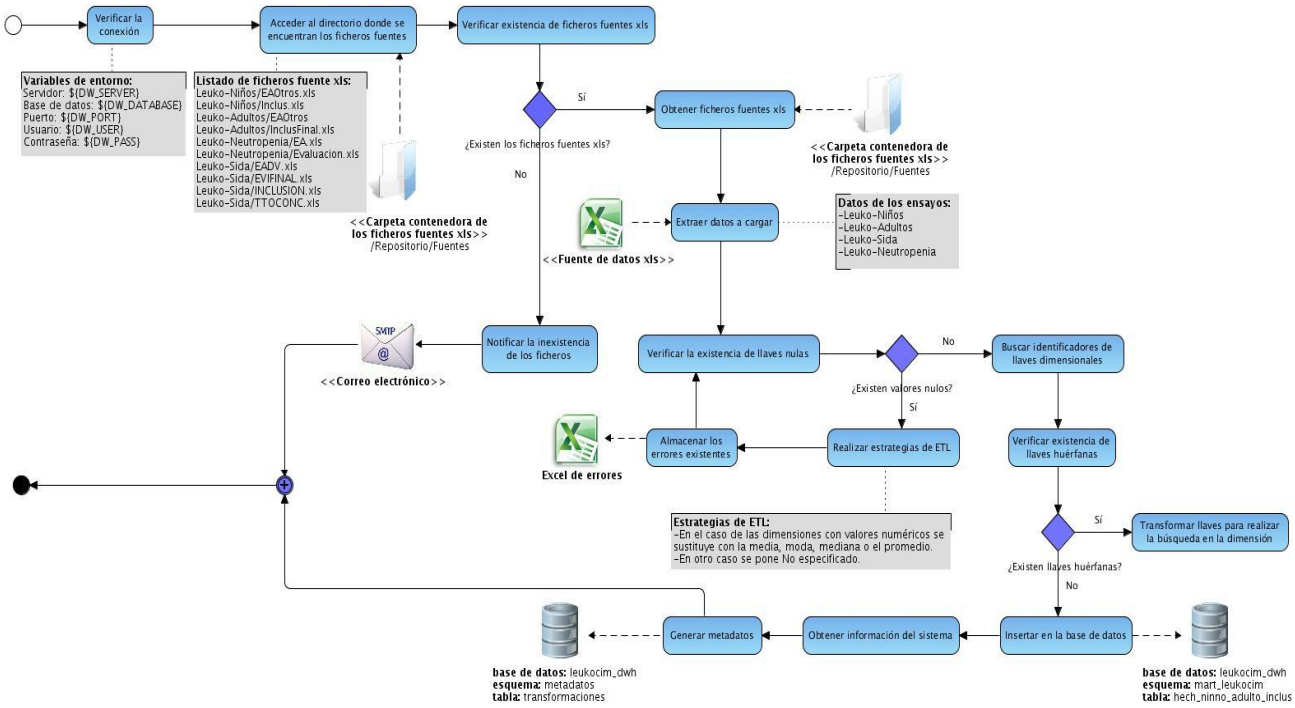

Figura 11: Diseño de la transformación para la carga de los hechos.

2.7 Política de respaldo y recuperación

Con el objetivo de garantizar la persistencia de la información, se establece una política de respaldo y recuperación basada fundamentalmente en las copias de seguridad. Debido a que el sistema posee una carga histórica, que solo se cargará una vez y no posee carga incremental, se realizarán copias de seguridad a la base de datos para salvaguardar los mismos. De igual forma en caso de perder la información se pueden volver a ejecutar las tranformaciones y se obtiene nuevamente la base de datos poblada. También se propone salvar la información en algún dispositivo de almacenamiento, ya que es de gran importancia mantener su disponibilidad y seguridad.

2.8 Esquema de seguridad

Es de gran importancia para un sistema de información contar con un mecanismo de protección contra aquellas acciones que puedan afectar la integridad de los datos almacenados. Por tal motivo, para el acceso a los Subsistemas de almacenamiento e integración LeukoCIM es necesario definir los roles que tendrán acceso a la base de datos.

Tabla 5: Roles y permisos de acceso a la base de datos.

Roles	Permisos
Administrador BD	Realiza la administración de la base de datos relacional que contiene todos los esquemas del almacén. Posee todos los permisos de administración y otorga permisos a los diferentes usuarios.
Administrador 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 LeukoCIM.

En el proceso de ETL la seguridad se garantiza con la opción que ofrece el sistema operativo, al marcar el atributo de solo lectura en las propiedades de la carpeta donde se almacenan todos los datos de las fuentes, las transformaciones y los trabajos.

Conclusiones del capítulo

Una vez abordados los principales elementos relacionados con los artefactos generados durante la etapa de análisis y diseño de los subsistemas de almacenamiento e integración, se puede concluir que:

- Se realizó un estudio de las necesidades de información del cliente o promotor del producto LeukoCIM, permitió la identificación de 23 requisitos de información, dos funcionales, siete no funcionales y 21 reglas del negocio, las cuales han sido aplicadas durante el diseño de los subsistemas.
- Mediante el diseño del modelo de datos fueron identificadas 24 tablas dimensionales y seis tablas de hechos, que garantizan el correcto funcionamiento del sistema.
- El perfilado de datos realizado a las diferentes fuentes de información permitió obtener una noción del estado de los sistemas fuentes, así como el establecimiento de nuevas reglas del negocio aplicables durante el proceso de transformación.
- El diseño de las transformaciones para la carga de las dimensiones y los hechos constituye una aproximación a los pasos que se deben realizar para lograr la estandarización de la información y su almacenamiento.
- Las políticas de recuperación y respaldo establecidas contribuyen a mantener la integridad de los datos almacenados.

Capítulo 3: Implementación y Prueba

Introducción

En este capítulo se realiza la implementación de los Subsistemas de almacenamiento e integración LeukoCIM. Este proceso se efectúa sobre las tablas del modelo de datos definido en el capítulo anterior. La implementación del subsistema de almacenamiento comprende la definición de los estándares de codificación y la construcción del modelo físico. Por su parte, en el subsistema de integración de datos se implementan las transformaciones y los trabajos, así como los metadatos. Se exponen las pruebas realizadas a los subsistemas y los resultados obtenidos en cada una de ellas. Dichas pruebas son realizadas para garantizar el cumplimiento de las exigencias del cliente y la calidad del producto.

3.1 Implementación del subsistema de almacenamiento

Una vez diseñado el modelo dimensional siguiendo una estandarización de los nombres, se dio lugar al modelo físico, permitiendo describir el almacenamiento de los datos y la relación entre las tablas. Además fueron creados los esquemas, así como las tablas correspondientes a cada uno de ellos.

3.1.1 Estándares de codificación

Los estándares de codificación pretenden organizar la forma en que se denominan las estructuras con el objetivo de lograr un patrón que tribute a la correcta normalización de los términos utilizados. Esta codificación está más bien dirigida a los desarrolladores, para que exista un vocabulario común en todo el almacén de datos, que permita un entendimiento claro y conciso.

En la solución propuesta se mantiene la misma nomenclatura atendiendo a la clasificación de las estructuras, teniendo en cuenta si la misma es una dimensión, un hecho, una transformación, un metadatos o un trabajo. Si la tabla es una dimensión, el nombre estaría compuesto por las letras “dim” separadas del nombre de la misma por el caracter “_”, ejemplo dim_sexo. En caso de ser una tabla de hecho, se le antepone las letras “hech” e igualmente se separa del nombre del hecho por el caracter “_”, ejemplo hech_sida_inclusion.

En el caso de los atributos de las dimensiones se siguió la misma estrategia para cada una de ellas. Las llaves primarias de las dimensiones fueron denominadas de la forma “dk_dim_dimension_id”. Si el atributo fuera un código del negocio se le especificó “dimension_codigo“. Se procedió de igual forma para el

nombre y la descripción: “dimension_nombre” y “dimension_descripcion” respectivamente. Las medidas contienen las letras “cant”, el caracter “_” y luego se especifica lo que se va a contar, ejemplo cant_pacientes.

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

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

3.1.2 Implementación del modelo de datos físico

El modelo de datos físico constituye una colección integrada de entidades que describen las estructuras de los datos, las restricciones de integridad y las operaciones de manipulación de los mismos. Dicho modelo se genera a partir del modelo lógico dimensional, mostrado 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.

En la presente investigación se definieron tres esquemas. El esquema dimensiones contiene las dimensiones compartidas con el resto de los mercados del almacén del CIM, el esquema mart_leukocim contiene las tablas de dimensiones y hechos propias del producto LeukoCIM y el esquema metadatos recoge las trazas de la ejecución de las transformaciones y los trabajos, así como la información de la carga histórica. La solución cuenta con 33 tablas, divididas en 24 tablas de dimensiones, seis tablas de hechos y tres tablas de metadatos. En la Tabla 6 se muestran las tablas asociadas a cada uno de los esquemas, mientras que en la Figura 12 aparece la estructura física de la base de datos leukocim_dwh.

Tabla 6: Esquemas y tablas de la aplicación.

Esquemas	Tablas

dimensiones	dim_edad	dim_provincia
	dim_tipo_evento_adverso	dim_raza
	dim_examen_laboratorio	dim_sexo
	dim_hospital	dim_talla
	dim_peso	dim_tiempo
	dim_causalidad_evento_adverso	dim_grado_evento_adverso
mart_leukocim	dim_ciclo	dim_razon
	dim_indicacion	dim_reg_admin
	dim_dosis_leuko_admin	dim_tratamiento
	dim_dosis_leuko_ind	dim_ensayo
	dim_enfermedad_base	hech_ninno_adulto_inclus
	dim_enfermedad_marcadora	hech_neutrop_eval
	dim_neutropenia	hech_sida_eval
	dim_inclusion_trat	hech_sida_inclusion
	hech_ninno_adulto_sida_neutrop_ea	hech_sida_trat_concom
metadatos	md_trabajos	md_transformacion
	md_carga_histórica	

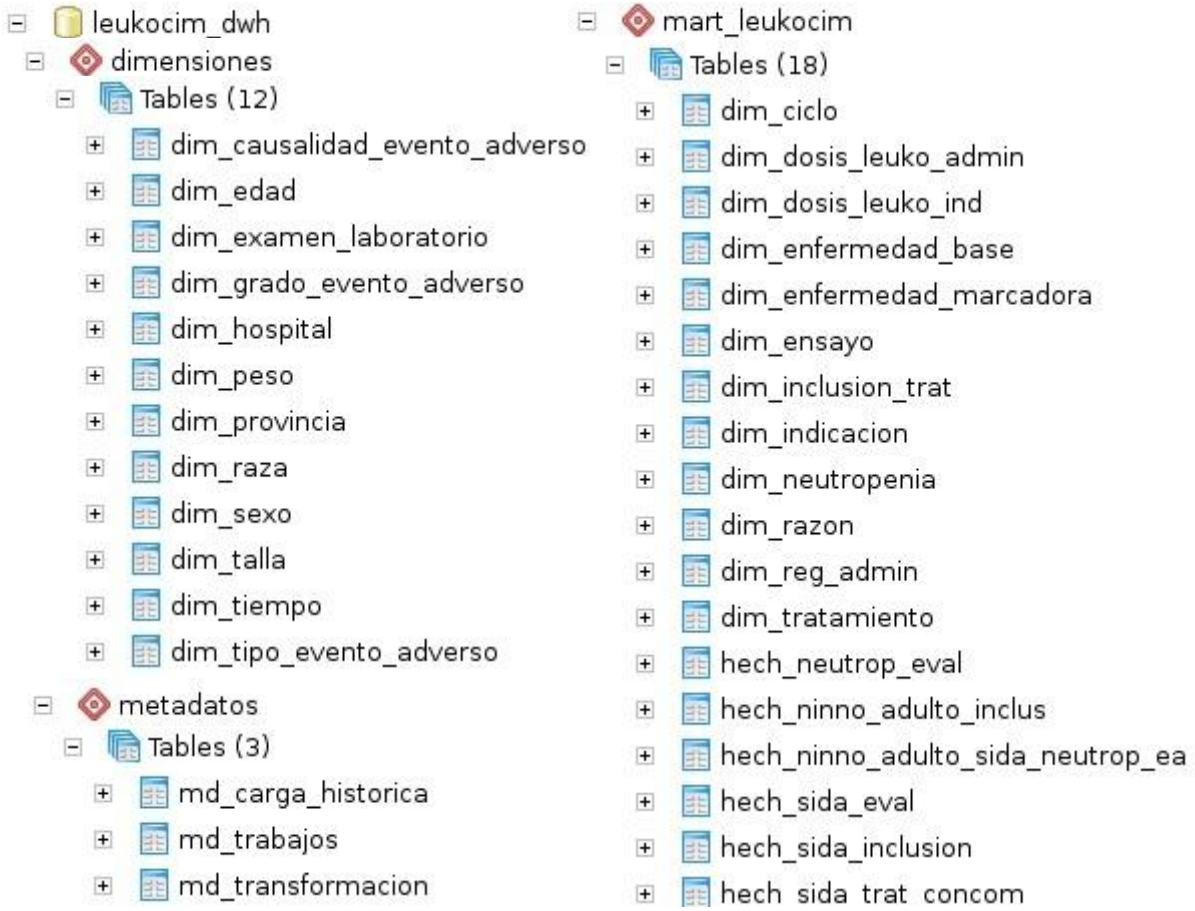

Figura 12: Estructura física de la base de datos.

3.2 Implementación del subsistema de integración

La implementación del subsistema de integración implica que se lleven a cabo los procesos de ETL. Para el desarrollo del presente trabajo se utiliza como estrategia de integración de datos: Extracción, Transformación y Carga. Se decide la utilización de esta estrategia pues da soporte a distintos orígenes de datos desde base de datos relacionales y no relaciones hasta archivos XML, archivos en formatos Excel y DBF, entre otro gran conjunto de ficheros.

El proceso de ETL brinda la posibilidad de enviar datos a otros *software* en tiempo real, tiene una amplia capacidad de transformación, desde una sencilla conversión de un tipo de dato, cálculos simples, hasta transformaciones complejas como agregaciones y sumarizaciones. Posee una amplia visualización de flujos, ya que se puede observar gráficamente cual es el recorrido del flujo de información desde que los

mismos son leídos de los sistemas fuentes hasta que son escritos en las distintas tablas de hechos y dimensiones y posee funcionalidades avanzadas que permiten limpiar los datos de origen evitando cargar información errónea en el almacén de datos.

La primera parte del proceso de ETL consiste en extraer la información desde los sistemas fuentes, en este caso ficheros en formato de hojas de cálculo Excel. Se seleccionan los campos relevantes para los Subsistemas de almacenamiento e integración LeukoCIM, teniendo en cuenta el modelo de datos realizado.

La segunda parte es la de transformación y limpieza, las cuales proveen una información lista para ser cargada en la base de datos. Con esta última se detectan los datos incorrectos, además de las entradas duplicadas y con las transformaciones se combinan y ordenan los datos.

Finalmente se procede a la carga, donde se toman los datos de la fase de transformación para cargarlos en el sistema destino, que sería la base de datos leukocim_dwh. Esta fase puede contener varias acciones pues en los almacenes generalmente se mantiene un historial de la carga de la información, cuyo objetivo es no actualizar los datos almacenados con anterioridad, para disponer de un historial de uno o más valores a lo largo del tiempo.

Es primordial tener en cuenta 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:** permitió explorar los datos para verificar su calidad y el cumplimiento de los estándares conforme a los requisitos especificados por el cliente. Mediante este subsistema fueron definidas nuevas reglas de transformación.
- ✓ **Subsistema de extracción:** permitió la extracción de los datos desde la fuente origen para su transformación y posterior carga. Para ello se tuvo en cuenta la información relacionada con cada uno de los hechos y dimensiones.
- ✓ **Subsistema de transformación:** permitió realizar transformaciones como el mapeo de valores, el cambio de datos en algunos campos, la búsqueda de información en flujos de datos, el filtrado de valores, entre otras.
- ✓ **Subsistema de carga:** permitió realizar la carga de los datos a las tablas de dimensiones y hechos de los Subsistemas de almacenamiento e integración LeukoCIM.

- ✓ **Rastreo de eventos de errores:** posibilita la captura de los errores que proporcionan información valiosa sobre la calidad de los datos y permiten la mejora de los mismos.
- ✓ **Llave subrogada:** permite crear llaves subrogadas independientes para cada tabla.
- ✓ **Programador de trabajos:** permite 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.
- ✓ **Repositorio de metadatos:** captura los metadatos de los procesos de ETL, de los datos de negocio y de los aspectos técnicos.

3.2.1 Implementación de las transformaciones

Las transformaciones están compuestas por pasos enlazados entre sí a través de los saltos. Los pasos constituyen el elemento más pequeño dentro de las transformaciones, y a través de los saltos fluye la información entre los diferentes pasos.

En la presente investigación se realizó un flujo de transformación para la carga de cada una de las tablas pertenecientes al esquema mart_leukocim. Para las dimensiones, la transformación se realizó a partir de la carga de los indicadores de cada uno de los ficheros de la fuente de datos, estos contienen los campos que son necesarios para poblar la base de datos, para luego elaborar las transformaciones de los hechos.

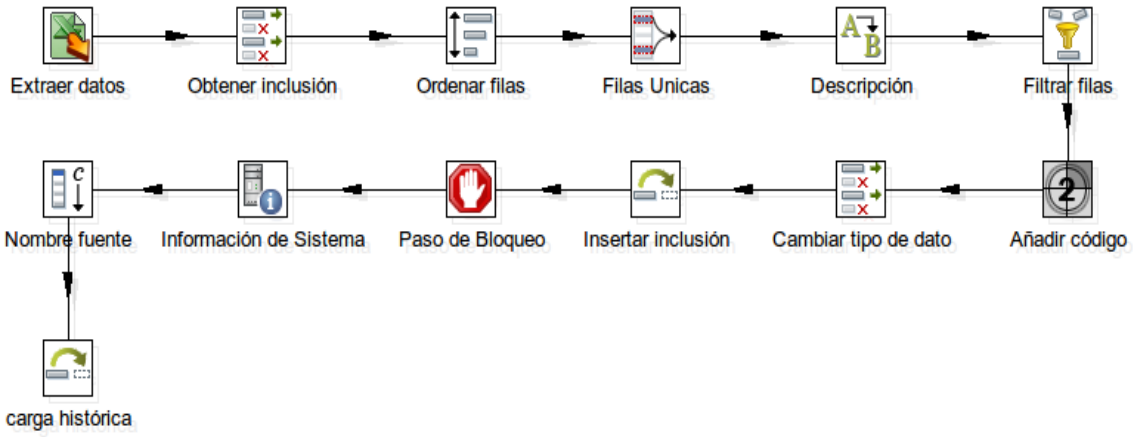

Figura 13: Transformación para cargar la dimensión inclusión.

Para realizar la carga de las dimensiones (Figura 13), el primer paso que se realiza es la extracción de los datos de las fuentes (Excel de los ensayos clínicos). Luego se obtuvo el campo específico para cargar. Se unieron los campos iguales y se ordenaron las filas, pues es necesario que los campos por los que se van

a agrupar estén ordenados. Posteriormente se utilizó el componente “Filas únicas” para seleccionar uno de los valores que se encuentran repetidos en el campo. Luego se añadió un código, que constituirá la llave subrogada de la dimensión y se le cambió el tipo de dato. El componente “Mapeo de Valores” permite entre otras cosas, especificar según el nombre del campo origen, los valores que tomará el campo destino. Posteriormente se inserta en la base de datos. Inmediatamente se obtiene la información necesaria del sistema para generar los metadatos técnicos, que guardará el nombre del fichero fuente, el nombre de la transformación, el nombre del destino o el hecho y la fecha de ejecución de la transformación. Por último se inserta o actualiza dicha información en la tabla md_carga_histórica. Se crean metadatos de procesos, para obtener información de la transformación, haciendo clic derecho sobre la misma y seleccionando “*Transformation Settings*” o con las teclas “*Ctrl + T*” y aparece una ventana donde se especifica la información que se desea obtener; en este caso se obtuvo las líneas de entrada, leídas, actualizadas, de salida y los errores encontrados durante la ejecución, entre otros.

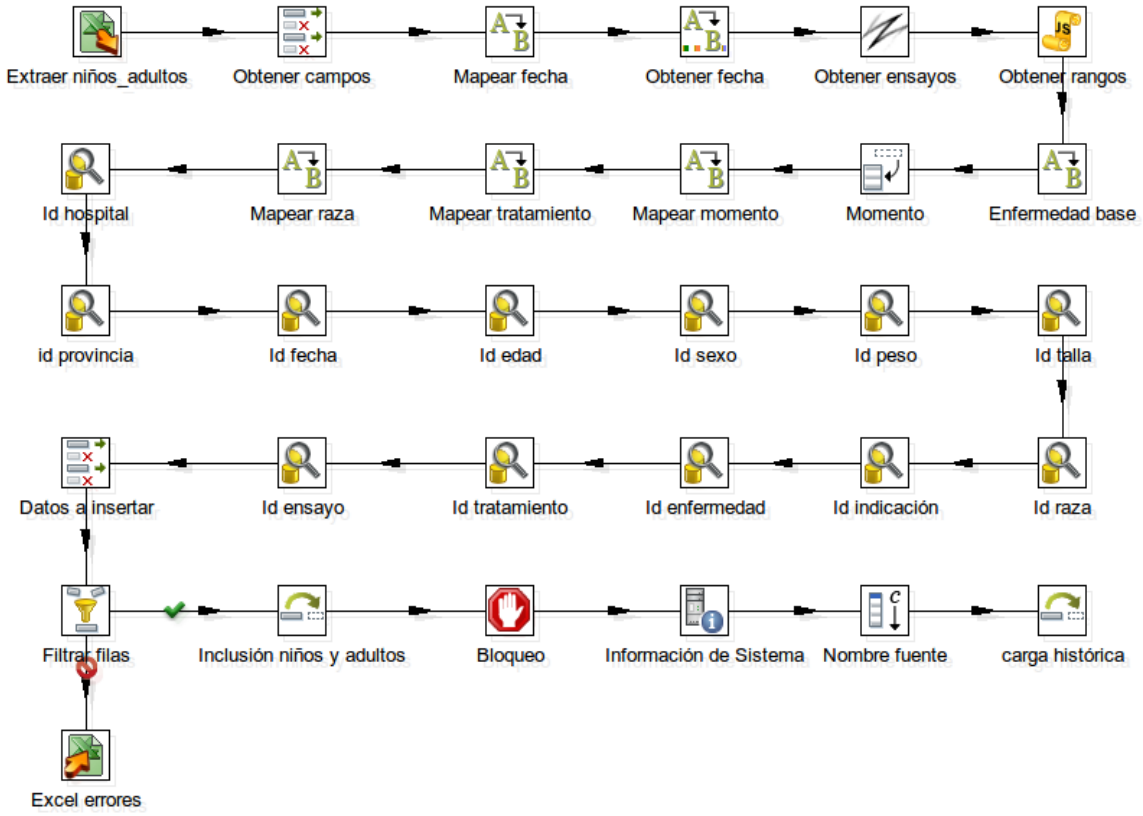

Figura 14: Transformación para cargar el hecho de la inclusión de niños y adultos.

Para realizar la extracción de los datos correspondientes a cada una de las tablas de hechos de la solución, se accede a los ficheros de la fuente, de donde son extraídos los campos necesarios atendiendo a las dimensiones con que se relaciona cada hecho y se procede a realizar las transformaciones pertinentes. Luego se procede a buscar la llave dimensional a partir de los datos que vienen de la fuente, para finalmente insertarlos en el esquema mart_leukocim en las tablas de hechos. Posteriormente de igual forma que en las dimensiones se procede a obtener información del sistema y del proceso de ETL para cargarlos en el esquema de metadatos en sus tablas correspondientes. La Figura 14 muestra la transformación correspondiente al hecho hech_ninno_adulto_inclus.

Los datos que sean almacenados en el Excel de errores recibirán tratamiento luego de definir con el cliente si constituyen información de gran importancia y cuáles reglas del negocio y transformación deben ser aplicadas para su carga a los Subsistemas de almacenamiento e integración LeukoCIM.

3.2.2 Implementación de los trabajos

En el contexto de integración de datos, el término trabajo o *job* en inglés, es un conjunto de tareas que se realizan con el objetivo de ejecutar una acción determinada. La implementación de un trabajo define una secuencia lógica para la ejecución de las transformaciones, mediante el uso de pasos definidos, los cuales son diferentes a los disponibles en las transformaciones. Es posible ejecutar una o varias transformaciones de las que se hayan diseñado y organizar una secuencia de ejecución para ellas. Los trabajos se encuentran en un nivel superior a las transformaciones.

En la Figura 15 se evidencia el trabajo general para la carga de las dimensiones y hechos del esquema mart_leukocim, porque las tablas del esquema dimensiones ya están cargadas. El cual primeramente se conecta a la base de datos y verifica la conexión, si está conectado comienza con la carga, de lo contrario termina su ejecución. La carga comienza con las transformaciones de las dimensiones compartidas por los hechos y luego carga cada uno de los trabajos de los hechos, que estos contienen las dimensiones propias y la transformación del hecho en sí.

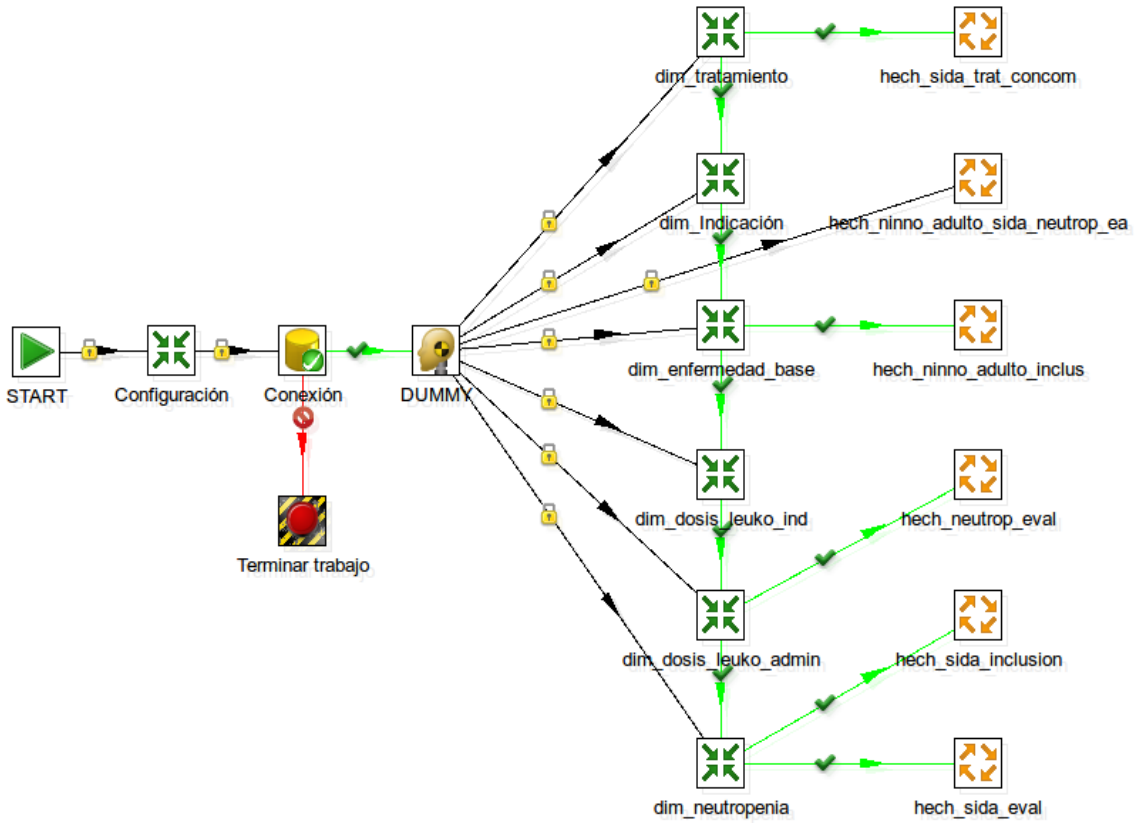

Figura 15: Trabajo general.

De esta forma se garantiza cargar los datos que no contengan errores, pues si existe algún problema con alguna dimensión en particular, los hechos que no están relacionados con esta podrán cargarse sin problemas, es decir que no existen dependencias entre ellos.

3.2.3 Gestión de los metadatos

Los metadatos son datos que ayudan a identificar, describir y localizar recursos digitales. Son información estructurada que describe y/o permite encontrar, gestionar, controlar y entender o preservar otra información; o sea son datos sobre los propios datos. Estos pueden ser agrupados en tres categorías:

- ✓ **Metadatos técnicos:** están relacionados con la función de un sistema o el modo en que se interrelacionan sus componentes.
- ✓ **Metadatos del negocio:** posibilita obtener los datos y la información referente a los aspectos del negocio, como son los datos provenientes de la fuente.

- ✓ **Metadatos de proceso:** permiten obtener información de los procesos que se ejecuten. Es la presentación de las estadísticas sobre los resultados de la ejecución del proceso de ETL, incluyendo medidas tales como filas cargadas con éxito, filas rechazadas y la cantidad de tiempo de carga; es muy importante en el proceso de limpieza de metadatos. [23]

En la investigación se utilizaron los metadatos de proceso para obtener la información correspondiente a los procesos de las transformaciones y los trabajos referentes a los subprocesos de ETL. Estos se almacenaron en dos tablas respectivamente, guardando en cada una de ellas el nombre, el estado, las líneas leídas, escritas, actualizadas, líneas de entrada y salida, los errores durante la ejecución, así como la fecha de carga, entre otras. En la Figura 16 se muestra un ejemplo de los metadatos de este tipo, realizado a las transformaciones para la carga de los hechos.

transname character varying(255)	status charact	lines_read bigint	lines_written bigint	lines_updated bigint	lines_input bigint	lines_output bigint	lines bigin	errors bigint
trans_hech_sida_trat_concom	end	269	269	0	269	0	0	0
trans_hech_ninno_adulto_sida_n	end	1498	1498	0	1498	0	0	0
trans_hech_ninno_adulto_inclus	end	2900	2900	0	2900	0	0	0
trans_hech_neutrop_eval	end	4515	4515	0	4515	0	0	0
trans_hech_sida_eval	end	357	357	0	357	0	0	0
trans_hech_sida_inclusion	end	8211	8211	0	8211	0	0	0

Figura 16: Metadatos de proceso para las transformaciones .

También se realizaron metadatos técnicos para guardar la información del nombre de la fuente, el nombre del destino, la fecha de ejecución y el nombre de la transformación.

3.3 Pruebas

Las pruebas constituyen un elemento crítico para la calidad del software. De acuerdo a la IEEE una prueba se define como: *Actividad en la cual un sistema o componente es ejecutado bajo condiciones específicas, se observan o almacenan los resultados y se realiza una evaluación de algún aspecto del sistema o componente*. [24] Una prueba se considera exitosa si encuentra alguna deficiencia en el *software*. Para obtener diferentes tipos de errores en el sistema se hace necesario aplicar un amplio conjunto de pruebas.

Existen diferentes tipos de pruebas que pueden ser aplicadas a un mercado de datos con el objetivo de obtener un producto con calidad. Estas pueden aplicarse de forma dirigida a componentes de *software* o

al sistema, con la finalidad de medir el grado en que el producto cumple con los requerimientos definidos. El proceso de pruebas comienza con la planificación de las mismas, seguidamente la ejecución, control y como paso final la evaluación.

Se evaluó la calidad de los Subsistemas de almacenamiento e integración LeukoCIM. Para ello se utilizó pruebas definidas por el Centro de Calidad para Aplicaciones Informáticas (CALISOFT) y lo pone en práctica DATEC con el fin de crear un estándar para comprobar que el producto cumpla con las especificaciones del negocio.

3.3.2 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 pruebas fueron realizadas fundamentalmente por los especialistas del departamento Almacenes de datos. Estuvieron centradas en la evaluación de cada subsistema por separado, verificando el correcto funcionamiento de los mismos, lo que facilita llegar a las pruebas de integración con un alto nivel de seguridad.

Una vez concluida la etapa de análisis y diseño se comprobó el subsistema de almacenamiento, donde se encontraron dos no conformidades (NC), las mismas quedaron resueltas.

NC1: Especificar el nivel de acceso a los pedidos de información.

NC2: Los casos de uso no están acorde a la representación del modelo de datos.

En la etapa de implementación se le realizaron pruebas unitarias al subsistema de integración de datos, detectando cinco no conformidades, solucionadas en su totalidad.

NC1: Desglosar y modificar los nombres de las actividades en el diseño de los procesos de integración de datos.

NC2: El diseño del proceso de integración no se corresponde con la implementación del problema modelado.

NC3: Incluir en el diseño el tratamiento de llaves nulas antes de realizar el de llaves huérfanas.

NC4: Revisar la estrategia de gestión de metadatos tanto en el diseño como en la implementación.

NC5: Optimizar la utilización de componentes en la implementación de las transformaciones.

3.3.3 Pruebas de integración

Dentro de los sistemas informáticos algunos componentes individuales son combinados con otros para asegurar que la comunicación, enlaces y los datos compartidos ocurran apropiadamente. El objetivo fundamental de las pruebas de integración es identificar errores introducidos por la combinación de programas o componentes probados unitariamente, además, verificar que las especificaciones de diseño sean alcanzadas. No son verdaderamente pruebas de sistema debido a que los componentes no se encuentran implementados en el ambiente operativo.

En la presente investigación se comprobó que los datos que están en la fuente compuesta por 10 ficheros en formato de hojas de cálculo Excel fueron cargados en su totalidad. Se le realizó consultas a la base de datos y se evidenció que la cantidad de pacientes que aparecen es igual.

La estrategia definida para realizar estas pruebas incluye la confección de casos de prueba que resultan de gran importancia para demostrar la funcionalidad del mismo estas se relacionan a continuación.

3.4 Herramientas para la aplicación de las pruebas

Dentro de las herramientas utilizadas para aplicar los distintos tipos de pruebas se tienen las listas de chequeo y los casos de prueba.

3.4.1 Casos de prueba

Los casos de prueba son utilizados para identificar posibles fallos de implementación y comprobar el grado de cumplimiento de los requisitos especificados para el sistema. En los Subsistemas de almacenamiento e integración LeukoCIM fueron diseñados doce casos de prueba asociados a dos RI por cada CUI identificados en la etapa de análisis, con el fin de comprobar que estén almacenadas las variables correspondientes.

Estos son esenciales para todas las actividades de pruebas porque son la base para diseñar y ejecutar los procedimientos de pruebas. Reflejan trazabilidad con los CU, ya que estos 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.

Al realizarle consultas a la base de datos se procedió de la siguiente forma:

Para el CUI5 se tomó como RI Obtener la cantidad de pacientes con SIDA que se evaluaron atendiendo a si recibirá el próximo ciclo de Quimioterapia/Radioterapia (QT/RT), tipo de neutropenia y tratamiento.

Para comprobar que los datos cargados coinciden con los que aparecen en la fuente se efectuó la siguiente consulta en lenguaje SQL.

```
SELECT COUNT(DISTINCT dk_codigo_paciente)
FROM
 mart_leukocim.hech_sida_eval,
 mart_leukocim.dim_neutropenia
WHERE
 dim_neutropenia.dk_dim_neutropenia_id =
hech_sida_eval.dk_dim_neutropenia_id AND
 hech_sida_eval.recib_prox_ciclo = 1 AND
 dim_neutropenia.neutropenia_descripcion = 'No';
```

Luego de contar la cantidad de pacientes que cumplían con las condiciones, se obtuvo como resultado 30. Se revisó el archivo en formato Excel “Leuko-Sida/EVFINAL.xls” y se le aplicaron los filtros apropiados para obtener finalmente 30 pacientes que presentan sida y recibirán el próximo ciclo de QT/RT y no presentan neutropenia; por lo que el resultado de la prueba es Satisfactorio.

Se trabajó de la misma manera para el CUI6 a través del RI Obtener la cantidad de pacientes con neutropenia que se evaluaron atendiendo a la dosis de LeukoCIM indicada, administrada, fecha de inicio y fin de la evaluación.

```
SELECT COUNT(DISTINCT dk_codigo_paciente)
FROM
mart_leukocim.hech_neutrop_eval,
mart_leukocim.dim_dosis_leuko_admin,
mart_leukocim.dim_dosis_leuko_ind
WHERE
dim_dosis_leuko_admin.dk_dim_dosis_leuko_admin_id =
hech_neutrop_eval.dk_dim_dosis_leuko_admin_id AND
```

dim_dosis_leuko_ind.dk_dim_dosis_leuko_ind_id =
hech_neutrop_eval.dk_dim_dosis_leuko_ind_id AND
dim_dosis_leuko_admin.dosis_leuko_admin_numero = 8 AND
dim_dosis_leuko_ind.dosis_leuko_numero = 300;

La cual arrojó como resultado que 21 pacientes presentan neutropenia y cuentan con ocho dosis de LeukoCIM administradas y 300 dosis de LeukoCIM indicadas, el cual coincide con lo que aparece en el archivo en formato de hojas de cálculo Excel “Leuko-Neutropenia/Evaluación.xls”; por lo que se concluye que el resultado del caso de prueba de integración en cuestión es Satisfactorio.

De igual forma se procedió para cada uno de los CUI, seleccionando dos RI. Para observar el resto de las pruebas realizadas, puede acceder al Expediente de Proyecto de los Subsistemas de almacenamiento e integración LeukoCIM en el artefacto “DATEC_CIM_Casos de prueba de integración.xls”.

Se definieron además 16 casos de prueba relacionados con las reglas de transformación, con la finalidad de comprobar si luego de ejecutada cada transformación, los datos cargados son los esperados. Como ejemplo de este proceso se presenta el Caso de prueba para la regla: Los valores del sexo tomarán el valor de 1 para Masculino y 2 para Femenino.

Tabla 7: Caso de prueba para una regla de transformación.

Caso de prueba						
Nombre variable	sexo					
Escenario	hech_ninno_adulto_inclus					
Regla de transformación	Los valores del sexo tomarán el valor de 1 para Masculino y 2 para Femenino.					
Valor de entrada	**Estado del dato**	**Resultado esperado**	**Respuesta del sistema**	**Resultado real**	**Comentario**	**Resultado de la prueba**
Femenino	No válido	2	Transforma el dato aplicando la regla	Femenino	La prueba se realizó sin ocurrir errores	Satisfac-torio

1	Válido	1	Agrega el dato correctamente	Masculino	La prueba se realizó sin ocurrir errores	Satisfac-torio

3.4.2 Listas de chequeo

La lista de chequeo consta de una serie de preguntas, en forma de cuestionario, mediante el cual se verifica el grado de cumplimiento de determinadas reglas establecidas para los procesos de desarrollo del sistema, además de medir la calidad de los artefactos de los procesos de ETL generados durante la realización del producto. Esta evaluación se desarrolla a través del análisis de un grupo de indicadores, distribuidos en tres secciones fundamentales:

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

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

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

Las listas de chequeo se le aplicaron a los artefactos de ETL “Registro del sistema fuente (RSF)”, “Perfilado de datos (PD)”, “Diccionario de datos (DD)” y “Mapa lógico de datos (MLD)”. Seguidamente se

muestra una tabla en la que se encuentran los principales aspectos que fueron evaluados y los resultados que arrojó la aplicación de los mismos.

Tabla 8: 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 indicadores	13	12	13	9
Indicadores críticos	5	5	5	5
No conformidades	0	1	1	1

En la Figura 17 se encuentra un resumen de los resultados obtenidos después de la aplicación de las Listas de Chequeo a los cuatro artefactos mencionados anteriormente.

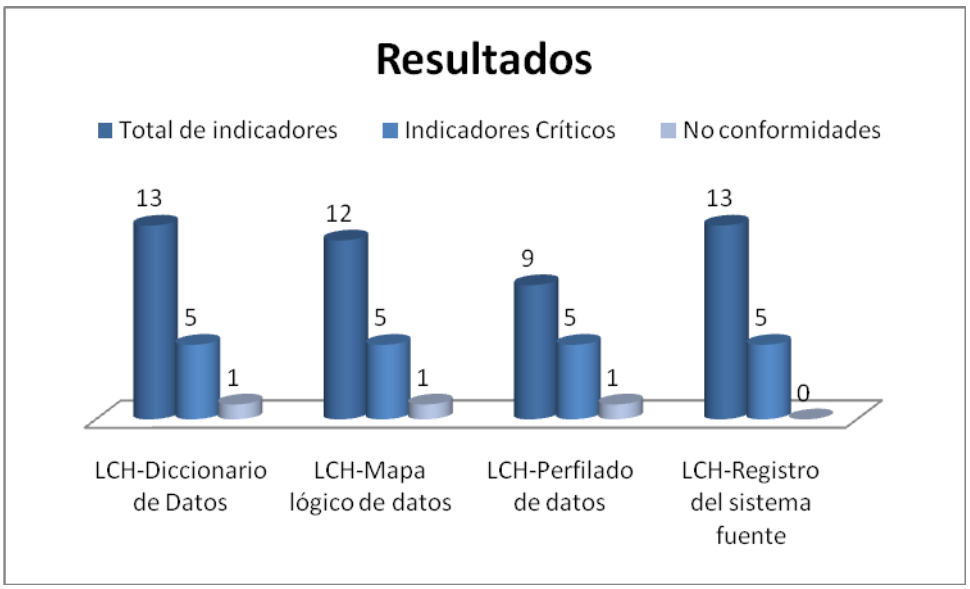

Figura 17: Resultados de la aplicación de las listas de chequeo a los artefactos.

3.5 Calidad de datos

Garantizar que los datos cargados tengan calidad es de vital importancia en el desarrollo de los subsistemas, debido a que de esta forma se puede afirmar que los datos no contienen errores. Este proceso de validación se realiza mediante el perfilado de los mismos luego de concluir con el proceso de ETL.

3.5.1 Perfilado de datos

El proceso de perfilado de datos permitió obtener estadísticas sobre los datos para de esta forma corregir errores como son valores duplicados, nulos o escritos incorrectamente. Luego de culminar el proceso de integración y carga de los datos se lleva a cabo este proceso para comprobar que los datos cargados no posean errores. Con el uso de la herramienta DataCleaner se obtuvieron resultados positivos respecto a los datos cargados.

Los reportes arrojados por este proceso indican que la carga de los datos correspondientes a cada uno de los hechos se realizó correctamente. No fueron almacenados valores vacíos ni nulos y los hechos contienen únicamente valores enteros, exceptuando el código del paciente que es cargado directamente de la fuente. El siguiente gráfico (Figura 18) muestra los resultados correspondientes al perfilado de datos realizado a la base de datos leukocim_dwh, que contiene todos los datos cargados de la fuente.

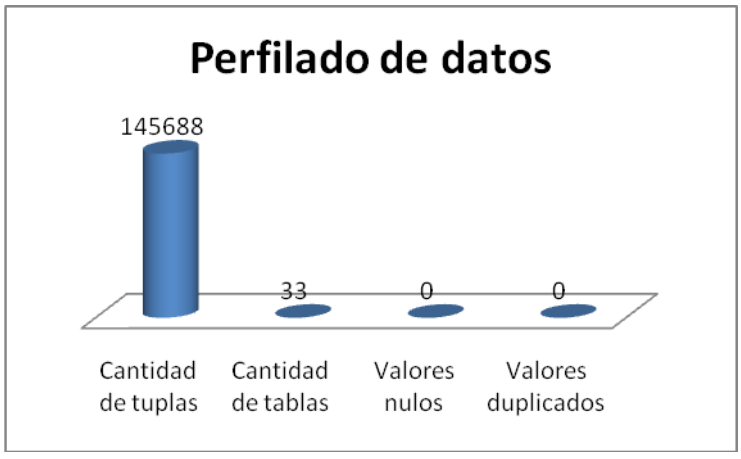

Figura 18: Resultados del perfilado de datos realizado a la base de datos leukocim_dwh.

3.5.2 Auditoría de datos

La auditoría de datos es el proceso de gestionar cómo los datos se ajustan a los propósitos definidos por la organización. Se establecen políticas para gestionar los criterios de datos para la organización. No es suficiente con actuar sino que se debe vigilar. [23]

A través de la realización de auditoría a los datos se obtiene conocimiento relacionado con la confiabilidad de los mismos, así como la información asociada a la ejecución de las transformaciones, como son: nombre de la transformación, fecha y hora de ejecución, cantidad de elementos de entrada, cantidad de elementos de salida, número de errores, entre otros elementos. La estrategia definida para auditar los datos almacenados se basa fundamentalmente en el uso de las tablas de metadatos implementadas, que proveen la información necesaria para comprobar y garantizar que los datos cargados sean confiables.

Conclusiones del capítulo

En el presente capítulo se abordó sobre la implementación y pruebas realizadas a los Subsistemas de almacenamiento e integración LeukoCIM, las cuales concluyen el proceso de construcción y validación, arrojando las siguientes conclusiones:

- Se implementaron los dos subsistemas que componen la aplicación: almacenamiento e integración, teniendo como resultado la disponibilidad de la información.
- La estructura física de los Subsistemas de almacenamiento e integración LeukoCIM está conformada por tres esquemas: dimensiones, mart_leukocim y metadatos, posibilitando la correcta integración de los datos a la base de datos.
- Se realizaron seis transformaciones para la carga de los hechos, 12 para las dimensiones y siete para los trabajos, que posibilitaron la carga de los datos de la fuente.
- Con la aplicación de las listas de chequeo a los artefactos de ETL, los 12 casos de prueba de integración basados en los casos de uso y los 16 aplicados a las reglas de transformación, se logró probar que los Subsistemas de almacenamiento e integración cumplen con los requisitos identificados.

Conclusiones generales

La investigación realizada cumple los objetivos planteados inicialmente mediante el desarrollo de los Subsistemas de almacenamiento e integración LeukoCIM, arribando a las siguientes conclusiones:

- El estudio de las metodologías y herramientas realizado, garantizó que la metodología seleccionada guiara el proceso de desarrollo de los Subsistemas de almacenamiento e integración LeukoCIM a través de cada etapa del ciclo de vida y la selección de las herramientas y tecnologías para la construcción de la solución propuesta cumplen con la política de migración a software libre, dando soporte a las necesidades del equipo de desarrollo.
- En el análisis y diseño de los Subsistemas de almacenamiento e integración LeukoCIM se identificaron 23 requisitos de información, dos funcionales, siete no funcionales y 21 reglas del negocio, logrando cumplir con las necesidades del cliente.
- La implementación de los tres esquemas, las seis transformaciones para la carga de los hechos, 12 para las dimensiones y siete para los trabajos, permitió la integración de los datos históricos y su almacenamiento.
- Las pruebas efectuadas durante las distintas etapas de desarrollo permitieron comprobar la funcionalidad del sistema a partir de los requisitos establecidos. Los resultados obtenidos durante las últimas pruebas realizadas fueron satisfactorios, validando el cumplimiento de los objetivos propuestos.

Recomendaciones

Para el presente trabajo de diploma se recomienda:

- Realizarle técnicas de minería de datos a la base de datos de los ensayos clínicos del producto LeukoCIM, que permitan detectar patrones de comportamiento sobre la información almacenada.

Referencias bibliográficas

1. **Díaz-Balart, Fidel Castro.** *Ciencia, Tecnología y Sociedad: Hacia Un Desarrollo Sostenible En La Era de La Globalización.* Cuba: Editorial Científico- Técnica, 2003. pág. 218. ISBN: 9789590105289.
2. **Ministerio de Salud Pública.** Centro para el control estatal de la calidad de los medicamentos. Regulación No. 45-2007. [En línea] 2009. [Citado el: 4 de noviembre de 2012] http://www.bvv.sld.cu.
3. Cubadebate. [En línea] 7 de diciembre de 2012. [Citado el: 5 de noviembre de 2012] http://www.cubadebate.cu/noticias/2011/12/07/investigacion-multinacional-probarafarmacos-cubanos-contra-el-cancer/.
4. Cuba comercializa cuatro vacunas contra el cáncer. [En línea] 6 de noviembre de 2012. [Citado el: 20 de noviembre de 2012] http://spanish.peopledaily.com.cn/92121/8006277.html.
5. Erosky Consumer. [En línea] 6 de septiembre de 2005. [Citado el: 5 de noviembre de 2012] http://www.consumer.es/web/es/salud/2005/09/06/145048.php.
6. **CIMAB S.A.** CIMAB S.A. [En línea] [Citado el: 4 de noviembre de 2012] http://www.cimabsa.com/index.php?action=producto&id=3.
7. **Inmon, William H.** *Building the Data Warehouse.* Fourt Edition. Indianapolis: Wiley Publishing, 2005. ISBN-13: 978-0-7645-9944-6.
8. **Kimball, Ralph y Ross, Margy.** *The Data Warehouse Toolkit:The Complete Guide to Dimensional Modeling.* New York: John Wiley & Sons, Inc., 2002. ISBN 0-471-20024-7.
9. **Sinnexus.** Sinnexus. Business Intelligence. [En línea] 2007. [Citado el: 16 de noviembre de 2012] http://sinnexus.es/business_intelligence/olap_vs_oltp.aspx.
10. **Tamayo, Marysol y Moreno, Fransisco J.** *Análisis del modelo de almacenamiento MOLAP frente al modelo de almacenamiento ROLAP.* No. 3, Colombia: s.n., diciembre de 2006, Ingeniería e Investigación, Vol. 26. ISBN: 0120-5609.
11. **Ibarzabal, J.** *Estrategia de reporting.* 2003.
12. **Becker, Shirley.** *Data Warehousing and Web Engineering*, idea Group publishing, 2003. pág 20. ISBN 1931777020.
13. **González Hernández, Yanisbel.** *Metodología para el Desarrollo de Soluciones de Almacenes de Datos e Inteligencia de Negocios en DATEC*. Ciudad de La Habana: s.n., 2010.

14. **Zepeda Sánchez, Leopoldo Z.** *Metodología para el Diseño Conceptual de Almacenes de Datos.* Universidad Politécnica de Valencia. España: s.n., 2008. Tesis doctoral.
15. **Pressman, Roger S.** *Ingeniería del Software. Un enfoque práctico.* Quinta Edición. 2005.
16. Visual Paradigm. *Boost Productivity with Innovative and Intuitive Technologies*. [En línea] 18 de junio de 2012. [Citado el: 14 de noviembre de 2012] http://www.visualparadigm.com/product/vpuml/.
17. **Martínez, Rafael.** PostgreSQL. [En línea] 2 de octubre de 2010. [Citado el: 15 de noviembre de 2012] http://www.postgresql.org.es/sobre_postgresql.
18. pgAdmin. PostgreSQL Tools. [En línea] 2012. [Citado el: 15 de noviembre de 2012] http://www.pgadmin.org.
19. **Yglesias, Rodolfo.** Oracle vs Oracle. [En línea] septiembre de 2008. [Citado el: 17 de noviembre de 2012] http://www.oracle.com/technology/global/lades/documentation/collaterals/BI-Whitepaper-Rodolfo-Yglesias.pdf.
20. **Pentaho Corporation.** Pentaho. *Powerful Analytics Made Easy.* [En línea] 2012. [Citado el: 18 de noviembre de 2012] http://www.pentaho.com/explore/pentaho-data-integration/.
21. **Gravitar.** Gravitar. [En línea] 2012. [Citado el: 18 de noviembre de 2012].
22. **Kimball, Ralph and Caserta, Joe.** The Data Warehouse ETL Toolkit. Canadá: Wiley Publishing, 2004.
23. **Medina Mustelier, Doris.** *Técnicas de Extracción, Transformación y Carga de Datos del Sistema de Información Nacional de Seguridad Ciudadana en la República Bolivariana de Venezuela*, Marzo 2009.
24. **IEEE-SA.** [En Línea] 2013. [Citado el: 19 de mayo de 2013]. http://standards.ieee.org/develop/project/29119-1.html.
25. **Fernández, Carlos.** DATAPRIX. *OLAP, ROLAP y MOLAP.* [En Línea] 2011. [Citado el: 24 de mayo de 2013]. http://dataprix.com/olap-rolap-molap.

Bibliografía

- **Díaz-Balart, Fidel Castro.** *Ciencia, Tecnología y Sociedad: Hacia Un Desarrollo Sostenible En La Era de La Globalización.* Cuba: Editorial Científico- Técnica, 2003. pág. 218. ISBN: 9789590105289.
- **Ministerio de Salud Pública.** Centro para el control estatal de la calidad de los medicamentos. Regulación No. 45-2007. [En línea] 2009. [Citado el: 4 de noviembre de 2012] http://www.bvv.sld.cu.
- Cubadebate. [En línea] 7 de diciembre de 2012. [Citado el: 5 de noviembre de 2012] http://www.cubadebate.cu/noticias/2011/12/07/investigacion-multinacional-probarafarmacos-cubanos-contra-el-cancer/.
- Cuba comercializa cuatro vacunas contra el cáncer. [En línea] 6 de noviembre de 2012. [Citado el: 20 de noviembre de 2012] http://spanish.peopledaily.com.cn/92121/8006277.html.
- Erosky Consumer. [En línea] 6 de septiembre de 2005. [Citado el: 5 de noviembre de 2012] http://www.consumer.es/web/es/salud/2005/09/06/145048.php.
- **CIMAB S.A.** CIMAB S.A. [En línea] [Citado el: 4 de noviembre de 2012] http://www.cimabsa.com/index.php?action=producto&id=3.
- **Inmon, William H.** *Building the Data Warehouse.* Fourt Edition. Indianapolis: Wiley Publishing, 2005. ISBN-13: 978-0-7645-9944-6.
- **Kimball, Ralph y Ross, Margy.** *The Data Warehouse Toolkit:The Complete Guide to Dimensional Modeling.* New York: John Wiley & Sons, Inc., 2002. ISBN 0-471-20024-7.
- **Sinnexus.** Sinnexus. Business Intelligence. [En línea] 2007. [Citado el: 16 de noviembre de 2012] http://sinnexus.es/business_intelligence/olap_vs_oltp.aspx.
- **Tamayo, Marysol y Moreno, Fransisco J.** *Análisis del modelo de almacenamiento MOLAP frente al modelo de almacenamiento ROLAP.* No. 3, Colombia: s.n., diciembre de 2006, Ingeniería e Investigación, Vol. 26. ISBN: 0120-5609.
- **Ibarzabal, J.** *Estrategia de reporting.* 2003.
- **Becker, Shirley.** *Data Warehousing and Web Engineering*, idea Group publishing, 2003. pág 20. ISBN 1931777020.
- **González Hernández, Yanisbel.** *Metodología para el Desarrollo de Soluciones de Almacenes de Datos e Inteligencia de Negocios en DATEC*. Ciudad de La Habana: s.n., 2010.

- **Zepeda Sánchez, Leopoldo Z.** *Metodología para el Diseño Conceptual de Almacenes de Datos.* Universidad Politécnica de Valencia. España: s.n., 2008. Tesis doctoral.
- **Pressman, Roger S.** *Ingeniería del Software. Un enfoque práctico.* Quinta Edición. 2005.
- Visual Paradigm. *Boost Productivity with Innovative and Intuitive Technologies*. [En línea] 18 de junio de 2012. [Citado el: 14 de noviembre de 2012] http://www.visualparadigm.com/product/vpuml/.
- **Martínez, Rafael.** PostgreSQL. [En línea] 2 de octubre de 2010. [Citado el: 15 de noviembre de 2012] http://www.postgresql.org.es/sobre_postgresql.
- pgAdmin. PostgreSQL Tools. [En línea] 2012. [Citado el: 15 de noviembre de 2012] http://www.pgadmin.org.
- **Yglesias, Rodolfo.** Oracle vs Oracle. [En línea] septiembre de 2008. [Citado el: 17 de noviembre de 2012] http://www.oracle.com/technology/global/lades/documentation/collaterals/BI-Whitepaper-Rodolfo-Yglesias.pdf.
- **Pentaho Corporation.** Pentaho. *Powerful Analytics Made Easy.* [En línea] 2012. [Citado el: 18 de noviembre de 2012] http://www.pentaho.com/explore/pentaho-data-integration/.
- **Gravitar.** Gravitar. [En línea] 2012. [Citado el: 18 de noviembre de 2012].
- **Kimball, Ralph and Caserta, Joe.** The Data Warehouse ETL Toolkit. Canadá: Wiley Publishing, 2004.
- **Medina Mustelier, Doris.** *Técnicas de Extracción, Transformación y Carga de Datos del Sistema de Información Nacional de Seguridad Ciudadana en la República Bolivariana de Venezuela*, Marzo 2009.
- **IEEE-SA.** [En Línea] 2013. [Citado el: 19 de mayo de 2013]. http://standards.ieee.org/develop/project/29119-1.html.
- **Fernández, Carlos.** DATAPRIX. *OLAP, ROLAP y MOLAP.* [En Línea] 2011. [Citado el: 24 de mayo de 2013]. http://dataprix.com/olap-rolap-molap.
- Postgres SQL Cuba. [En Línea] 2011. http://postgresql.uci.cu.
- **Levin, Jonathan.** [En Línea] [Citado el: 20 de marzo de 2008] http://mysqlbarbeque.blogspot.com.
- **Hernández, Griselda A.** Lenguaje Unificado de Modelado. [En línea] 14 de abril de 2010. [Citado el: 10 de noviembre de 2012] http://utec-uml.blogspot.com/2010/04/ventajas-y-desventajas-de-almacenes-de.html.