Universidad de las Ciencias Informáticas

Facultad 6

Título: Subsistemas de almacenamiento e integración del producto egf 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

Autoras:

Yoanna Jiménez Osés

Rebeca Rivera Osaba

Tutores:

Ing. Arodys Eugenio Dominguez Vaillant

Ing. José Leandro González González

Ciudad de la Habana, junio, 2013

"Año 55 de la Revolución"

Frases

"En la tierra hacen falta personas que trabajen más y critiquen menos, que construyan más y destruyan menos, que prometan menos y resuelvan más, que esperen recibir menos y dar más, que digan mejor ahora que mañana"

Ernesto Guevara de la Serna

"La vida es una gran sala de espera"

Alejandro Sanz

Declaración de Autoría

Declaramos ser autoras del presente trabajo “Subsistemas de almacenamiento e integración del producto egf para el almacén de datos de los Ensayos Clínicos del Centro de Inmunología Molecular” y reconocemos a la Universidad de las Ciencias Informáticas (UCI) los derechos patrimoniales de la misma, con carácter exclusivo.

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

Rebeca Rivera Osaba

_____

Firma de la Autora

Yoanna Jiménez Osés

_____

Firma de la Autora

Ing. Arodys E. Dominguez Vaillant

_____

Firma del Tutor

Ing. José Leandro González González

_____

Firma del Tutor

Datos de contacto

Tutores:

Ing. Arodys E. Domínguez Vaillant

Especialidad de graduación: Ingeniería en Ciencias Informáticas

Categoría docente: Asistente

Años de experiencia en el tema: 3 años

Años de graduado: 6 años

Correo Electrónico: adominguez@uci.cu

Ing. José Leandro González González

Especialidad de graduación: Ingeniería en Ciencias Informáticas

Categoría docente: Instructor

Años de experiencia en el tema: 2 Años

Años de graduado: 5 Años

Correo Electrónico: jlgonzalez@uci.cu

Dedicatoria

Les dedico el presente trabajo a toda mi familia por el gran apoyo que me han brindado, en especial a mi papá y a mis abuelos que son mi razón de ser.

Yoanna

A mi mamá por ser mi ejemplo a seguir, por haber confiado en mí y apoyarme siempre en todo. Por explicarme las cosas difíciles de una manera entendible y darme todo su amor y comprensión.

Rebeca

Agradecimientos

Agradecer a todas las personas que han sido mi apoyo y guía en todos los momentos es la parte más difícil, porque son muchos los que de una forma u otra han hecho algo por mí para salir adelante tanto en mi vida profesional como personal. Trataré de ser imparcial.

Primeramente quiero agradecerle a una persona muy especial, mi papá, por apoyarme en todo, por ser mi guía y estar siempre a mi lado en los buenos y malos momentos, por sus buenos consejos, por comprenderme, por ser exigente cuando tiene que serlo, por depositar en mí toda su confianza, porque sin él no hubiera sido capaz de realizar este sueño.

A mis abuelitos que son mi vida, gracias por cuidarme, por brindarme su cariño en los momentos que más los he necesitado, por ser la fuerza que me impulsa a seguir adelante.

A mi mejor amiga Gleibis que para mí es la hermana que siempre he querido tener, gracias amiga por tus consejos, por tu preocupación, por estar conmigo en las buenas y en las malas, por compartir todos estos años juntas, por tu amistad que para mí es el mayor regalo, siempre voy a llevarte en mi corazón adonde quiera que valla.

A mi novio Rey por ayudarme a salir adelante, por cuidarme, por brindarme su amor y apoyo, por darme tanta felicidad, por su paciencia, por soportarme como soy.

Agradezco a Mirna, Reinaldo y Lili por dejarme formar parte de su familia y por toda la confianza que han depositado en mí durante todo este tiempo.

A mi madrastra la China por ser como una madre para mí, estoy sumamente agradecida por toda tu dedicación y preocupación, todo tu cariño y apoyo han sido muy importantes para mí, gracias por formar parte de mi vida.

A mis tíos y tías, Grisel, Luisito, Pancho, Milagro, Magdalena por estar siempre pendientes de mis pasos y animarme a lograr este sueño que hoy se hace realidad.

A mis vecinos Olga, Jorge, Monguita a todos gracias por su preocupación y cariño, a Gloria por sus buenos consejos.

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

A mis amigos, por ser parte de mi vida, de mis momentos tristes y alegres, por apoyarme, por siempre estar ahí, en especial a Luis Angel , Lianet, Anneris, Yasmany, Kiel, Gema, Abelito, René, Tahimi, Yami a mis compañeros de aula y a todos los que hoy no pueden estar aquí conmigo. Los quiero.

A mis profesores, que compartieron conmigo sus conocimientos para convertirme en la persona que hoy soy, por su tiempo, dedicación y por su amistad. A Elio, Esley, José Carlos, Reinier, Fifi, Susana.

A mis tutores Arodys y José Leandro por dedicarnos su tiempo, por sus consejos.

Al tribunal y el oponente por sus correcciones, sus señalamientos, su sabiduría y su esfuerzo por ayudarnos a mejorar nuestro trabajo.

A la Revolución y a nuestro comandante Fidel por haber creado esta universidad de la que fui parte.

A Dios, por brindarme la oportunidad de vivir, por permitirme disfrutar cada momento de mi vida y guiarme por el camino que has trazado para mí.

Yoanna

Agradecimientos

A Dios por haberme dado la oportunidad de formar parte de esta vida.

A mi mamá Leonarda por ser mi guía, por haber creído en mí y haber inculcado en mí el hábito de la superación. Gracias por tus enseñanzas y por tus regaños; por ser mi ejemplo a seguir y por ser una mujer admirable. Por cuidarme y estar siempre pendiente de mí ante cualquier enfermedad, por preocuparte porque todo estuviera bien y por llamarme todos los días, por ayudarme siempre en todo lo que necesité y darme palabras de aliento en los momentos más difíciles de la carrera. Gracias por todo, porque sin ti hoy no estuviera aquí.

A Roxana por ser una gran hermana, por la forma en la que nos queremos y por lo distintas que somos, por darme todo tu apoyo y por disfrutar conmigo todas las cosas lindas que hemos vivido juntas.

A mi tía tata por haber pasado gran parte de mi niñez a su lado, por darme todo su cariño y soportar tantas malcriadeces.

A mi tía Margot por enseñarme su paciencia y por empujarme la comida cada vez que no quería, por hacerme ese jugo de mango que tanto me gustaba cada vez que llegaba de la escuela. Hoy no puede estar aquí pero sé que desde allá arriba está gozando junto conmigo de este momento.

A mis primas Sandra e Isca por ayudarme siempre en todo y por sacarme de tantos apuros.

A toda mi familia, por ser tan unida, y enseñarme que no existe mejor dicha que esa.

A Ayito por ser una persona tan especial en mi vida, por ayudarme en todos estos años en mi carrera, por tener tanta paciencia cuando quería por momentos tirarme por la escalera del edificio. Por darme todo su amor y comprensión y por enseñarme que las cosas hay que lucharlas hasta el final.

A Cristy por dejarme formar parte de su familia, por su apoyo y dedicación en todo momento.

A mis amigas Moraima, Dayana y Sobeida por aconsejarme en los momentos que más lo necesité, por estar presente en los buenos y malos momentos y por compartir juntas tantas cosas lindas.

A mi amiga Normita por haber sido tan alegre y tan entusiasta hasta el último momento de su vida, por enseñarme que la vida hay que vivirla con intensidad, porque al final del día no sabemos dónde estaremos.

A mi amigo Osmín por ayudarme a levantar mi ánimo con sus correos en los momentos más difíciles.

A Romy, Yohana, Lianet, Tahimy y Karla por compartir momentos juntas en estos años de la carrera, por estar siempre en el momento justo en que las necesite, por eso y por todo, gracias.

A todos aquellos que conocí a los largo de la carrera, a los que están aquí presentes y a los que no, por haber compartido estos años juntos y haber pasado momentos inolvidables.

A los profesores que de una forma u otra nos ayudaron, en especial al profesor Esley por su paciencia y dedicación, Elio por ayudarme tanto con P2, a Ada por estar siempre cuando la necesitamos y nunca decirnos que no y a nuestros tutores.

Al tribunal y oponente por sus correcciones y sus señalamientos que tanto nos ayudaron.

Y por último, pero no menos importante a mi compañera de tesis Yoanna, por soportarnos cuando teníamos la soga al cuello, por los malos ratos que pasamos en el laboratorio hasta las tantas, pero también por los momentos buenos, por siempre estar preocupada con todo, y porque juntas hicimos posible este gran sueño.

Rebeca

Resumen

La presente investigación surge con el objetivo de darle solución a una serie de problemas identificados en el Centro de Inmunología Molecular (CIM), en cuanto a la gestión de la información de los Ensayos Clínicos (EC) que son aplicados a pacientes con cierto tipo de enfermedades cancerígenas. Se pretende crear un repositorio que integre la información relacionada con dichos temas, con el propósito de facilitar a los especialistas realizar análisis y consultar la misma para obtener resultados sobre los datos que son arrojados después de aplicada la vacuna egf a los pacientes que padecen de cáncer de pulmón. Para su posterior desarrollo se documentan la metodología, herramientas y tecnologías que son utilizadas en estos procesos. Se realiza el análisis, diseño e implementación de los subsistemas de almacenamiento e integración, haciendo uso de las herramientas de la Suite de Pentaho, así como la validación del mismo a través de pruebas para probar el correcto funcionamiento de dicha solución. Finalmente se obtuvo como resultado un mercado de datos poblado con toda la información integrada de los EC.

Palabras claves

Integración de la información, mercado de datos.

Índice

Introducción 1

Capítulo 1: Fundamentación teórica de los almacenes de datos 4

1.1 Introducción 4

1.2 Ensayos Clínicos 4

1.2.1 Gestión de Ensayos Clínicos en el Centro de Inmunología Molecular 4

1.3 Almacenes de Datos 5

1.3.1 Características de los Almacenes de Datos 6

1.3.2 Ventajas y desventajas de los Almacenes de Datos 7

1.3.3 Actualidad de los Almacenes de Datos 7

1.4 Mercado de Datos 8

1.5 Modelo Multidimensional 9

1.6 Fases de construcción de un Mercado de Datos 9

1.6.1 Análisis y Diseño 10

1.7 Metodologías para el desarrollo de un Almacén de Datos 10

1.7.1 Propuesta de Metodología para el Desarrollo de Almacenes de Datos en DATEC 11

1.8 Herramientas de modelado 13

1.8.1 Visual Paradigm for UML 8.0 14

1.9 Sistema Gestor de Base de Datos 14

1.9.1 PostgreSQL 9.1 14

1.9.2 PgAdmin III 1.14.0 15

1.10 Herramientas para los procesos de Extracción, Transformación y Carga 16

1.10.1 DataCleaner 1.5.4 16

1.10.2 Pentaho Data Integration 4.2.1 16

Conclusiones del Capítulo 17
Capítulo 2: Análisis y diseño de los subsistemas de almacenamiento e integración. 18
2.1 Introducción 18
2.2 Análisis del negocio 18
2.2.1 Necesidades del negocio 18
2.3 Especificación de requisitos 19
2.3.1 Requisitos de información 19
2.3.2 Requisitos funcionales 20
2.3.3 Requisitos no funcionales 20
2.4 Reglas del negocio 20
2.5 Casos de uso del sistema 21
2.5.1 Actores del sistema 21
2.5.2 Casos de uso de información 21
2.5.3 Casos de uso funcionales 22
2.5.4 Diagrama de casos de uso 23
2.6 Definición de la arquitectura 24
2.7 Diseño de los subsistemas de almacenamiento e integración 25
2.7.1 Diseño del subsistema de almacenamiento 26
2.7.2 Dimensiones 26
2.7.3 Dimensiones degeneradas 27
2.7.4 Tablas de hechos 27
2.7.5 Matriz buz o matriz dimensional 28
2.7.6 Esquemas multidimensionales 31
2.7.7 Modelo de datos 33
2.8 Diseño del subsistema de integración 35

2.8.1 Estrategias de calidad de datos 35
2.8.2 Diseño de los procesos de integración 36
2.9 Política de respaldo y recuperación 39
1.10 Esquema de seguridad 39
1.10.1 Seguridad en el subsistema de almacenamiento 39
1.10.2 Seguridad en el subsistema de integración 40
Conclusiones del capítulo 40
Capítulo 3: Implementación y validación de los subsistemas de almacenamiento e integración. 41
3.1 Introducción 41
3.2 Estrategias generales de integración 41
3.3 Implementación del subsistema de almacenamiento 42
3.3.1 Estándares de codificación 42
3.3.2 Implementación del modelo de datos físico 43
3.4 Implementación del subsistema de integración 43
3.4.1 Implementación de las transformaciones 45
3.4.2 Implementación de los trabajos 46
3.4.3 Gestión del cambio lento en las dimensiones 47
3.5 Pruebas aplicadas a los subsistemas del producto egf 48
3.5.1 Pruebas unitarias 49
3.5.2 Pruebas de integración 50
3.6 Herramientas de pruebas para validar los subsistemas del producto egf 50
3.6.1 Casos de prueba 50
3.6.2 Listas de chequeo 52
3.7 Calidad de datos 53
3.7.1 Perfilado de los datos 53

3.7.2 Auditoría de datos 54
Conclusiones del Capítulo 54
Conclusiones Generales 55
Recomendaciones 56
Referencias Bibliográficas 57
Bibliografía 60
Anexos 63
Anexo 1: Modelo de Datos 63
Glosario de Términos 64

Introducción

El avance vertiginoso de la ciencia y la técnica ha demostrado la necesidad del hombre de aplicar métodos científicos para transformar el medio que lo rodea de acuerdo a sus necesidades. La tecnología existente hoy en día, es un medio indispensable para el desarrollo de la sociedad, aplicada en diferentes sectores en todo el mundo, la cual busca darle respuesta a los problemas y necesidades existentes en la sociedad, empleando para ello recursos que se encuentren disponibles dentro de la misma.

Actualmente en la esfera de la salud se incorporan nuevas técnicas para su informatización en aras de mejorar la calidad de los servicios que se le brindan a la población. Por lo cual, en este sector se evidencia la necesidad de la vinculación con las Tecnologías de la Información y las Telecomunicaciones (TICs) en las cuales se integra todo lo relacionado con la computación, las telecomunicaciones y técnicas para el procesamiento de datos.

En Cuba, con el triunfo de la Revolución, se crearon programas encaminados a resolver las necesidades de la población, desarrollando centros educacionales e instituciones con el objetivo de proveer al Sistema Nacional de Salud (SNS) técnicas y procedimientos para mejorar la calidad del procesamiento de los datos técnicos, biomédicos y administrativos, y brindar información confiable que sustente la toma de decisiones. Uno de estos centros creados es la Universidad de las Ciencias Informáticas (UCI), la cual tiene como principal misión formar profesionales comprometidos con su Patria y altamente calificados en la rama de la Informática y servir de soporte a la industria cubana de la informática. (1) La misma cuenta con el Centro de Tecnologías de Gestión de Datos (DATEC), integrado por cuatro líneas de desarrollo, teniendo como principal misión crear bienes y servicios informáticos relacionados con la gestión de datos, área del conocimiento que agrupa tanto a los sistemas de información, como a los denominados sistemas de inteligencia empresarial o de negocios, cuyo propósito fundamental es apoyar el proceso de toma de decisiones. (2) Dentro de estas líneas se encuentra Almacenes de Datos, la cual realiza soluciones informáticas con el objetivo de apoyar diferentes áreas dentro y fuera de la UCI. Una de estas áreas es la de los EC del CIM el cual maneja toda la información relacionada con los EC aplicados a cada paciente para su tratamiento. Los datos se recogen mediante un sistema informático que permite la gestión de archivos: EpiData, que son exportados en disímiles formatos (Text, Excel, Stata, SPSS, SAS, Access), por lo que se vuelve muy engorroso a la hora de confeccionar reportes, analizar, consultar la información recopilada y presentar los indicadores relacionados con dichos ensayos. Los EC llevan asociados una

gran cantidad de documentación necesaria para cumplir con las buenas prácticas clínicas, por lo que el volumen de información almacenada en esta área del CIM se ha incrementado considerablemente y la gestión de la misma se realiza manualmente, resultando difícil para los especialistas realizar análisis certeros que contribuyan con las decisiones que la entidad debe tomar.

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

Dicha investigación tiene como **objeto de estudio:** los Almacenes de Datos. Delimitándose como **campo de acción:** los subsistemas de almacenamiento e integración del producto egf para el Almacén de Datos de los EC del CIM.

Para darle solución al problema planteado, se determina como **objetivo general:** desarrollar los subsistemas de almacenamiento e integración del producto egf para el Almacén de Datos de los EC del CIM que permita el almacenamiento homogéneo de la información.

Para darle solución al objetivo general se identificaron los siguientes **objetivos específicos**:

1. Fundamentar 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.
3. Realizar la implementación y pruebas de los subsistemas de almacenamiento e integración.

Se propone la realización de las siguientes **tareas de la investigación** para dar cumplimiento a los objetivos específicos planteados:

1. Caracterización de la metodología, herramientas y tecnologías a utilizar en el desarrollo de almacenes de datos.
2. Levantamiento de requisitos para definir las necesidades del cliente.
3. Descripción de los casos de usos de los subsistemas de almacenamiento e integración.
4. Definición de la arquitectura de los subsistemas de almacenamiento e integración.
5. Diseño del subsistema de almacenamiento.
6. 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 para la calidad y auditoría de los datos.

Estructura del documento

El presente documento está estructurado como se muestra a continuación: resumen, introducción, tres capítulos; los cuales ayudarán a lograr un mejor entendimiento acerca de los temas referentes a los mercados y almacenes de datos para guiar el desarrollo del negocio, conclusiones generales, recomendaciones, referencias bibliográficas, bibliografía, glosario de términos y anexos.

Capítulo 1: Fundamentación teórica de los almacenes de datos.

En este capítulo se realiza un análisis de los principales elementos relacionados con los Almacenes de Datos (AD) y Mercados de Datos (MD). Se reflejan las principales características, ventajas, desventajas y otros conceptos de vital importancia que ayudan en el desarrollo de los mismos. También se realiza un estudio de la metodología y herramientas empleadas para la solución. Se describen las diferentes fases para la construcción de los MD y AD y la forma de representar los datos conociendo las definiciones de los hechos, dimensiones y medidas que son aplicadas para lograr un mejor entendimiento de elementos que intervienen en el negocio.

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

En este capítulo se realiza un análisis del negocio, con el objetivo de identificar los principales aspectos de relevancia para la organización, definiéndose los requisitos de información del sistema, agrupados en casos de uso de acuerdo a los criterios existentes. Se definen las reglas del negocio y se identifican las dimensiones, hechos y medidas para poder diseñar el modelo de datos. Se realiza el diseño de los subsistemas de almacenamiento e integración y el diseño del esquema de seguridad.

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

En este capítulo el proceso de desarrollo comprende la implementación de cada uno de los subsistemas definidos: el de integración y almacenamiento. Estos dos procesos permiten almacenar la información en las tablas correspondientes definidas en el modelo de datos. Además se describen las técnicas de carga de las dimensiones, tablas de hechos y trabajos correspondientes. Se detallan las pruebas a realizar, para la validación de la solución mediante herramientas de pruebas.

Capítulo 1: Fundamentación teórica de los almacenes de datos

1.1 Introducción

En este capítulo se realiza el análisis de los principales elementos para el desarrollo de la solución propuesta, se brindará una reseña relacionada con las características de los EC del producto egf y el impacto que ha tenido sobre pacientes que sufren cáncer de pulmón. Se abordarán múltiples elementos teóricos relacionados con los AD y MD; reflejándose las principales características, ventajas, desventajas, actualidad de los mismos y otros conceptos de vital importancia que ayudan en su desarrollo. Además se realiza un estudio y selección de las herramientas a utilizar en la solución para las diferentes etapas por las que transita, la metodología adoptada para guiar el proceso de desarrollo y a su vez reconocer las diferentes actividades, roles, procesos y características que posee la misma. También se describen las diferentes fases para la construcción de los AD y MD para lograr un mejor entendimiento de los procesos que son implementados.

1.2 Ensayos Clínicos

Los EC son estudios de investigación que permiten evaluar la eficacia y seguridad de tratamientos médicos en la confección de nuevos fármacos a través de la aplicación a seres humanos. Estos estudios amplían las posibilidades de mejorías en los involucrados, proceso que es llevado a cabo con plena aprobación del paciente.

Los EC se caracterizan por los siguientes aspectos:

- ✓ Son estudios experimentales, que se llevan a cabo en seres humanos. A diferencia del estudio observacional, el investigador interviene en el curso normal de los acontecimientos, de forma que se condiciona el tratamiento que reciben los pacientes.
- ✓ Son siempre prospectivos, se planifican y se realizan, siguiendo la evolución de los sujetos de investigación a lo largo del tiempo.
- ✓ Se emplea una intervención o se administra un tratamiento que normalmente difiere del habitual (sustancia no autorizada como especialidad farmacéutica o en condiciones de uso distintas a las autorizadas) y por ello puede no aportar beneficio al sujeto. Esta posibilidad cierta de no beneficio, tiene connotaciones legales y éticas relacionadas con la protección de los pacientes. (3)

1.2.1 Gestión de Ensayos Clínicos en el Centro de Inmunología Molecular

El CIM tiene como principal misión la búsqueda de nuevos productos para el diagnóstico y tratamiento del cáncer. Dicho centro actualmente uno de los productos que desarrolla es la vacuna CIMAvax EGF, cuyo tratamiento es utilizado en enfermedades de cáncer de pulmón. Esta vacuna disminuye o castra el factor de crecimiento epidérmico en el paciente sin bajar sus niveles a cero, de forma que las funciones vitales que requieren del mismo sigan ocurriendo, pero el tumor que demanda mayor cantidad no crece o al menos se estabiliza. No puede afirmarse la curación del individuo porque se reduzca el tumor, pero la experiencia de estudio demuestra que cuando el cáncer no se extiende durante un largo período, la persona se encuentra en una etapa estable de la enfermedad y puede vivir por mucho tiempo. (4)

En tres años se han tratado más de 700 pacientes en 65 policlínicos de todas las provincias. Los datos del ensayo clínico fase IV en los policlínicos han confirmado el perfil de seguridad y eficacia de la vacuna obtenidos en los ensayos fase III. (5) Hasta la actualidad el tratamiento ha sido aplicado a más de 2.000 cubanos con muy buenos resultados. (6)

Para llevar el control de estos estudios, los especialistas realizan la gestión de sus ensayos de forma tradicional a través de los cuadernos de recogida de datos (CRD). Luego son llevados al CIM, lugar donde se realiza el proceso de digitalización de la información almacenada en los cuadernos mediante el sistema EpiData; utilizado para la recopilación de la información; generándose reportes en diferentes formatos (Text, Excel, Stata, SPSS y SAS). A este sistema acceden diferentes especialistas que trabajan con la información almacenada, resultándole un poco difícil el análisis y consulta sobre los EC al no encontrarse integrada. Debido a esta situación se corre el riesgo de perder información valiosa para la entidad, se dificulta la búsqueda de alternativas para encontrar estabilidad en los resultados arrojados en los EC y el tiempo consumido en la misma interfiere en el impacto que pueda tener en la sociedad. Por todas estas razones la solución propuesta está enmarcada en la implementación de los procesos de extracción, transformación y carga sobre la información referente a los EC.

1.3 Almacenes de Datos

Actualmente las empresas dedican una parte importante de su tiempo y de sus recursos económicos y humanos a la obtención, procesamiento y proyección de la información para el desarrollo de sus actividades cotidianas. La información es un recurso vital para toda organización, y el buen manejo de esta puede significar la diferencia entre el éxito o el fracaso para todos los proyectos que se emprendan dentro del negocio. Muchas de estas entidades presentan un gran reto, debido al excesivo volumen de

información que maneja, la aparición de diversos formatos, influyendo esto en la demora de respuesta hacia los usuarios que dependan de la misma.

Con el pasar de los años la información es considerada un fenómeno que crece a raíz de las actividades generadas por las entidades, donde cada una depende de la otra y apoyan diferentes áreas dentro de la institución. Esto trae consigo un aumento considerable de tiempo y acumulación de datos a procesar, por lo que se hace necesaria la gestión de toda esta información para tener un control más amplio sobre el funcionamiento de todas las actividades de la organización. Con el propósito de apoyar el proceso de toma de decisiones, los AD (también conocidos como DataWarehouse (DWH)), permiten consultar, analizar y presentar la información que sea requerida por los especialistas, para darle respuesta a peticiones realizadas por el usuario y contribuir así a la toma de decisiones por parte de la gerencia de la entidad o la empresa.

Existen diversas tendencias y formas de conceptualizar a los AD y aunque se diferencian en algunos aspectos, todos estos conceptos giran sobre un mismo eje central. A continuación se enuncian dos de ellas: "una colección de datos, orientados a hechos relevantes del negocio, integrados, no volátiles y variantes en el tiempo, para el apoyo a la toma de decisiones administrativas", la cual fue expresada por William H. Inmon. (7)

Se puede citar lo planteado por otro reconocido autor como Ralph Kimball, considerado el principal promotor del enfoque dimensional para el diseño de AD, este lo define como: "…una copia de datos transaccionales, específicamente estructurados para la consulta y el análisis". (8)

Todos los autores de investigaciones referentes a AD coinciden en sus trabajos en los mismos argumentos. Brindan una percepción cualitativa sobre el tema, desde su punto de vista expresándolo de forma diferente.

De esta manera se puede concluir que los AD son estructuras que se definen en función de temas específicos, la información histórica debe estar integrada y robusta ante los cambios que puedan afectar a la organización. Su objetivo principal, es apoyar el proceso de toma de decisiones empresariales.

1.3.1 Características de los Almacenes de Datos

Los AD reúnen características especiales, las cuales se mencionan a continuación:

- ✓ **Integrado:** como los datos almacenados provienen de fuentes diferentes deben integrarse en una estructura consistente que elimine las inconsistencias existentes en los mismos.
- ✓ **Temático:** los datos se organizan por temas para facilitar su acceso y entendimiento por parte de los usuarios finales.

- ✓ **Histórico:** los datos se organizan y almacenan en jerarquías en el tiempo, lo que permite análisis comparativos de estados actuales y de períodos anteriores.
- ✓ **No volátil:** el almacén de datos puede ser leído pero no modificado. Es decir, se incorporan los últimos valores que tomarán las distintas variables contenidas en él, sin ningún tipo de acción sobre los valores que ya existían. (9)

1.3.2 Ventajas y desventajas de los Almacenes de Datos

Un AD aporta facilidad e inmediatez en el manejo de la información. También poseen inconvenientes que dificultan su óptimo rendimiento. A continuación se muestran un conjunto de ventajas y desventajas de los mismos:

Ventajas

- ✓ Alto retorno de inversión. La inversión que se realiza para una correcta implantación de un sistema de AD conlleva un coste muy elevado, sin embargo el retorno de la inversión es garantizado en gran medida.
- ✓ Como consecuencia de la ventaja anterior se puede conseguir una ventaja competitiva debido a una buena toma de decisiones gracias al AD implantado.
- ✓ Los AD integran datos de múltiples sistemas incompatibles y proporcionan una base de datos clasificada por temas e histórica. La transformación de los datos orientados a las aplicaciones de información, permite a dichos responsables realizar análisis más precisos y consistentes.
- ✓ Beneficios en costes, tiempos y productividad de la empresa. Ayuda a obtener mejores tiempos de respuesta y supone una mejora en los procesos de producción. (10)

Desventajas

- ✓ Subestimación del tiempo requerido para extraer, limpiar y cargar los datos en el almacén.
- ✓ Los almacenes de datos se pueden quedar obsoletos relativamente pronto si los usuarios incrementan sus necesidades.
- ✓ Problemas con los sistemas de origen de los datos.
- ✓ Pueden suponer altos gastos. Además de los gastos de mantenimiento que son muy elevados.
- ✓ Debido a que están estrechamente relacionadas con los sistemas operativos se han de tener en cuenta cuales son las funcionalidades que pueden aprovecharse. Como por ejemplo, la utilización de gran cantidad de espacio en disco. (10)

1.3.3 Actualidad de los Almacenes de Datos

La situación actual, en la que cada vez es más alto el nivel competitivo entre las empresas, provoca que las organizaciones recurran con más frecuencia a las nuevas tecnologías para conseguir estrategias de gestión que les reporten la información que necesitan. Poco a poco las empresas fueron almacenando un gran número de información en diferentes fuentes de datos (archivos, documentos de texto, bases de datos, etc.), y los directivos de las empresas se dieron cuenta de que ésta podría ser útil, pues reflejaba la mayoría de las operaciones diarias del negocio. Hoy en día muchas de estas empresas utilizan AD para mejorar la calidad en el almacenamiento de la información, lo cual ayuda a la toma de decisiones de dichas empresas. Entre ellas se pueden mencionar: Walmart, Procter & Gamble, Whirlpool, 3M, Coca Cola, Walt Disney, Tv Azteca, Banorte, Banco de México, Banamex, Nike, Baxter, GNP, Warner Lambert. INFOMEDIA. (11)

En Santo Domingo, República Dominicana el 11 de febrero de 2010 la Oficina Nacional de Estadística e Información (ONEI) puso en línea el Almacén Central de Datos del Sistema Estadístico Nacional: señaló que el AD apoyará la recolección, procesamiento, análisis y difusión de informaciones estadísticas y ofrecerá servicios de información basados en modelos analíticos y de minería de datos. (12)

Cuba a pesar de ser un país en vías de desarrollo no se encuentra rezagada respecto a este tema, ya que existen varias empresas que cuentan con AD con el fin de estandarizar los datos de las mismas para futuros análisis. Ejemplo de estas son: Almacén de Datos para la Gestión Contable de la Empresa de Proyectos de Arquitectura e Ingeniería (EMPAI): el trabajo consiste en el diseño e implementación de una herramienta que permita gestionar y organizar homogéneamente la información relevante actual e histórica sobre los indicadores de eficiencia de la gestión contable de la EMPAI de Matanzas, utilizando las posibilidades que brindan los AD. (13) También se encuentra el almacén comercial de la Corporación CIMEX.

En la UCI el centro DATEC tiene asociado el departamento de Almacenes de Datos, el cual ha brindado ayuda a centros de desarrollo del país en cuanto a este tema se refiere. Se han desplegado AD en algunas organizaciones como: el CIM, la Oficina Nacional de Estadísticas e Información (ONEI) y la Comisión Nacional Electoral.

1.4 Mercado de Datos

Un MD también conocido como Datamart, es una base de datos departamental, especializada en el almacenamiento de los datos de un área de negocio específica. Se caracteriza por disponer de una estructura óptima de datos para analizar la información al detalle desde todas las perspectivas que afecten

a los procesos de dicho departamento. Un MD puede ser alimentado desde los datos de un AD o integrar por sí mismo un compendio de distintas fuentes de información. (14)

Características

- ✓ Según las necesidades de los usuarios el diseño del MD se realiza siguiendo una estructura consistente.
- ✓ La información histórica que posee es mínima.
- ✓ Contiene el grado de granularidad necesaria.
- ✓ 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. (14)

1.5 Modelo Multidimensional

El modelado multidimensional es una técnica de diseño lógico que presenta los datos de un modo estandarizado, es intuitivo para los usuarios proporcionando acceso a la información. (15) Los datos se almacenan como hechos, medidas y dimensiones. A continuación se enunciarán los conceptos básicos para lograr un entendimiento referente al tema:

Hecho: evento específico que constituye la unidad fundamental de análisis de datos para la toma de decisiones. (16) Representa una actividad objeto de análisis, actividad que está caracterizada por un conjunto de dimensiones. (17)

Medidas: valores cuantitativos que almacenan las métricas del negocio. Están representados por columnas numéricas en las tablas de hechos. (16)

Dimensión: una dimensión puede tener múltiples niveles de agrupación. Están descritas por un conjunto de atributos que se organizan en jerarquías. Las jerarquías entre los atributos de una dimensión representan una organización de los valores de la dimensión, que van a permitir calcular las medidas de la actividad a distintos niveles de detalle. (17)

1.6 Fases de construcción de un Mercado de Datos

Para la implementación de los subsistemas de almacenamiento e integración del producto egf, es necesario conocer las fases por las cuales necesitaría transitar para así satisfacer las necesidades de la entidad. Estas se dividen en tres etapas elementales: **análisis y diseño** para un mejor entendimiento del negocio, **extracción, transformación y carga** para la integración de los datos e **inteligencia de negocio**

para analizar el comportamiento de los datos de cierta organización. De estas tres fases la investigación se centrará en la dos primeras, atendiendo a las necesidades del cliente.

1.6.1 Análisis y Diseño

En esta etapa se realiza un estudio preliminar del negocio identificándose las necesidades y principales requisitos de información que el cliente requiere sean implementados, sirviendo de base para obtener conocimiento del negocio; sus prioridades, cómo trabajan con la información y con qué frecuencia consultan la misma, todos estos elementos contribuyen en la implementación del AD.

Extracción, Transformación y Carga

Es la etapa siguiente del análisis y diseño. Surge con la necesidad de la integración de los datos que intervienen en el negocio, los cuales son utilizados como fuentes para su posterior composición. A continuación se definen los procesos pertenecientes a dicha etapa:

Extracción: los formatos de las fuentes normalmente se encuentran en bases de datos relacionales o ficheros planos, pero pueden incluir bases de datos no relacionales u otras estructuras diferentes. La fase de extracción convierte los datos de los diferentes sistemas a un formato preparado para iniciar el proceso de transformación.

Transformación: la fase de transformación aplica una serie de reglas de negocio o funciones sobre los datos extraídos para convertirlos en datos que puedan ser cargados.

Carga: la fase de carga es el momento en el cual los datos de la fase anterior son cargados en el destino, dependiendo de los requerimientos de la organización. (15)

1.7 Metodologías para el desarrollo de un Almacén de Datos

Las metodologías de desarrollo de software son un conjunto de procedimientos, técnicas y ayudas a la documentación para el desarrollo de productos de software. Se van indicando paso a paso todas las actividades a realizar para lograr el producto informático deseado, indicando además qué personas deben participar en el desarrollo de las actividades y qué papel deben tener. Además detallan la información que se debe producir como resultado de una actividad. (18) Desarrollar software de buena calidad, depende en gran medida de la correcta selección de métodos y herramientas que contribuyan al éxito del mismo. Con el propósito de guiar estos procesos han surgido diferentes metodologías de desarrollo, las cuales a su vez definen un conjunto de pasos que permiten planificar, definir, seleccionar, diseñar y conducir el proceso de desarrollo de software.

Para la confección de un AD existen varias metodologías, elegir la correcta depende de las necesidades u objetivos que persiga la empresa para emplear alguna. Estas diferentes metodologías se pueden englobar dentro de dos grandes bloques: top-down y bottom-up que se corresponden con las metodologías propuestas por Bill Inmon y Ralph Kimball respectivamente. Estos autores merecen una especial atención porque, en muchos aspectos, se consideran los precursores del AD y sus opiniones son muy valoradas en la industria.

La metodología descendente **(top-down)** que define Inmon se utiliza cuando la tecnología y los problemas del negocio se conocen de antemano. Se trata de un método sistémico, que minimiza los problemas de integración, pero es costoso, debido a la gran cantidad de datos y su poca flexibilidad.

En este método se formula un resumen del sistema, sin especificar detalles. Propone construir primero el AD y a partir de este los MD, plantea la creación de un repositorio de datos corporativo como fuente de información consolidada, persistente, histórica y de calidad.

Después, cada parte nueva se redefine, cada vez con mayor detalle, hasta que la especificación completa es lo suficientemente detallada para validar el modelo. Este modelo se diseña con frecuencia con la ayuda de "cajas negras" que hacen más fácil cumplir requerimientos, aunque estas cajas negras no expliquen en detalle los componentes individuales.

Kimball defiende por tanto una metodología ascendente **(bottom-up)** a la hora de diseñar un AD. Es una metodología rápida que se basa en experimentos y prototipos. Es un método flexible que permite a la organización ir más lejos con menores costos. La idea es construir MD independientes para evaluar las ventajas del nuevo sistema a medida que se vaya avanzando. En él, las partes individuales se diseñan con detalle y luego se enlazan para formar componentes más grandes, que a su vez se enlazan hasta que se forma el sistema completo. (19)

1.7.1 Propuesta de Metodología para el Desarrollo de Almacenes de Datos en DATEC

Para definir la metodología a usar en los subsistemas de almacenamiento e integración del producto egf para el área de EC, se optó por la propuesta de metodología para el desarrollo de AD en DATEC, la cual toma como base la metodología Ciclo de vida de Kimball y lo propuesto en la Tesis de Doctorado de Leopoldo Zenaido Zepeda, incluyendo los casos de uso para guiar el proceso de desarrollo. Además se incluye una etapa de prueba para fortalecer la salida del producto. Para obtener una mayor calidad en los productos que se desarrollan se incorpora el Expediente de Proyecto, creado por el Centro Nacional de Calidad de Software (CALISOF), encargado de certificar la calidad de los productos desarrollados por la UCI. Ver figura 1.

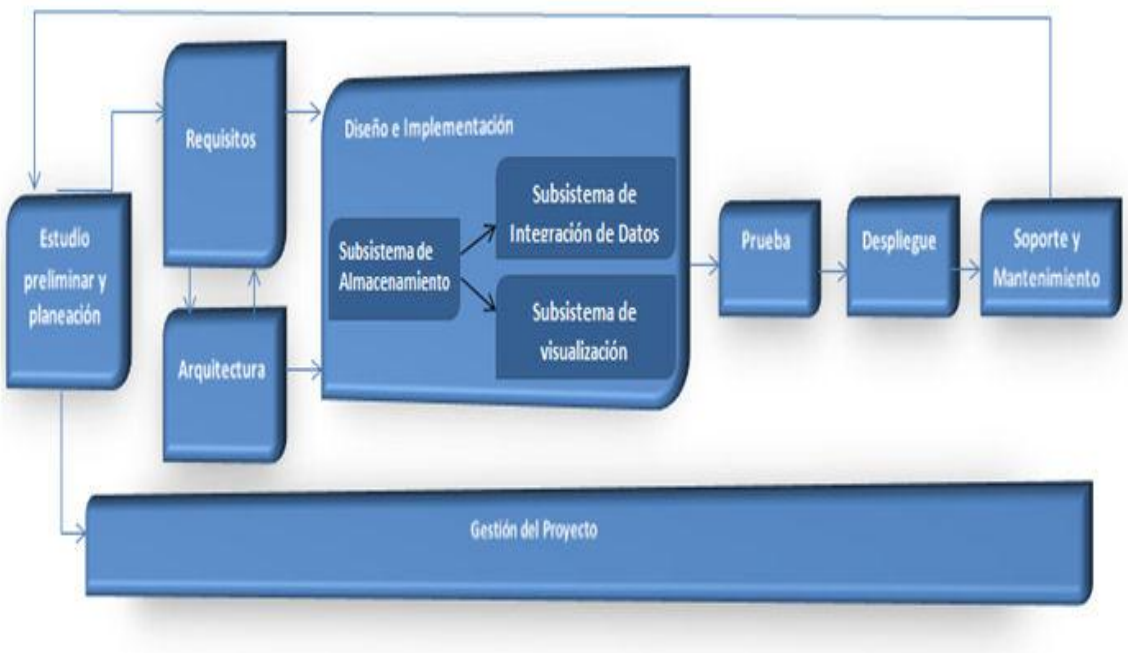


Figura 1: Metodología de Desarrollo de AD

La metodología de desarrollo está dividida en ocho fases:

- ✓ **Estudio preliminar o planeación:** se realiza un estudio minucioso en la entidad cliente. Esto incluye un diagnóstico integral de la organización, con el fin de determinar qué es lo que se desea construir y qué condiciones existen para el desarrollo y montaje de la misma. Además se llevan a cabo las tareas de planeación del proyecto.
- ✓ **Requisitos:** se realiza el proceso entrevistas al cliente para determinar los requisitos de información. Se hace un levantamiento detallado de las fuentes de datos para validar la disponibilidad de la información. Además se definen los requisitos funcionales y no funcionales de la solución y se hace el análisis de los requisitos que dan paso al diseño e implementación.
- ✓ **Arquitectura:** se definen las vistas arquitectónicas de la solución, aspectos como: los subsistemas y componentes, la seguridad, la comunicación y la tecnología a utilizar.
- ✓ **Diseño e Implementación:** se define el diseño de las estructuras de almacenamiento de datos, se diseñan los procesos de integración de datos como: el mapa lógico de datos, los cubos OLAP para la presentación de la información, así como el diseño gráfico de la aplicación definido por el cliente. Después se implementan cada uno de los subsistemas (repositorio de datos, integración de datos, presentación de datos).

- ✓ **Prueba:** se realizan las pruebas que validan la calidad del producto, comenzando por las Pruebas de Unidad, las Pruebas de Integración y Sistema, hasta llegar a las Pruebas de Aceptación con el cliente final. Esta fase no es la única en la que se realizan pruebas durante el desarrollo del proyecto, en todas las fases hay actividades de aseguramiento de la calidad.
- ✓ **Despliegue:** consta de dos etapas, despliegue piloto, donde se configuran los servidores necesarios y se instalan las herramientas según la arquitectura definida, se carga una muestra de los datos en un ambiente controlado, con el fin de mostrarle al cliente final el sistema en funcionamiento. Una vez aceptada la solución por el cliente, se realiza la carga histórica de los datos, puede ser en el mismo entorno que el despliegue piloto u otro, todo depende de las condiciones que establezca el cliente. Además se realiza la capacitación y transferencia tecnológica de la solución a los clientes. El resultado fundamental es la solución desplegada en el entorno real y en correcto funcionamiento.
- ✓ **Soporte y Mantenimiento:** comienza cuando la solución está implantada y en explotación, y se ejecuta según el contrato firmado y las condiciones de soporte establecidas. Puede realizarse a través de variados servicios, que pueden ser soporte en línea, vía telefónica, web, correo u otros y el acompañamiento al cliente. Además se realizan las tareas de manteniendo de la aplicación tan necesarias para este tipo de desarrollo y que garantiza el adecuado funcionamiento y crecimiento del AD.
- ✓ **Gestión del proyecto:** esta fase se ejecuta a lo largo de todo el ciclo de vida del proyecto. Es aquí donde se controla, gestiona y chequea todo el desarrollo, los gastos, las utilidades, los recursos, las adquisiciones, los planes y cronogramas entre otras actividades relacionadas con la gestión de proyectos. Esta fase es la columna vertebrar del proyecto y si no se ejecuta de forma continua y correcta el proyecto puede fracasar. (20)

En la investigación se hace uso de las cinco primeras fases definidas en la metodología propuesta por DATEC para el desarrollo de AD. Las restantes serán llevadas a cabo por los especialistas del proyecto.

1.8 Herramientas de modelado

A las herramientas de modelado se les define como herramientas CASE (Computer Aided Software Engineering). Se les puede definir como un conjunto de programas y ayudas que dan asistencia a los analistas, ingenieros de software y desarrolladores, durante todos los pasos de desarrollo de un Software. (21)

1.8.1 Visual Paradigm for UML 8.0

Visual Paradigm es una de las herramientas UML CASE del mercado, fácil de usar, con soporte multiplataforma y que proporciona excelentes facilidades de interoperabilidad con otras aplicaciones. (21)
Propicia un conjunto de ayudas para el desarrollo de programas informáticos, desde la planificación, pasando por el análisis y el diseño, hasta la generación del código fuente de los programas y la documentación.

Se caracteriza por:

- ✓ Disponibilidad en múltiples plataformas (Windows, Linux).
- ✓ Diseño centrado en casos de uso y enfocado al negocio que generan un software de mayor calidad.
- ✓ Uso de un lenguaje estándar común a todo el equipo de desarrollo que facilita la comunicación.
- ✓ Capacidades de ingeniería directa e inversa.
- ✓ Modelo y código que permanece sincronizado en todo el ciclo de desarrollo. (22)
- ✓ Diagramas de flujo de datos.
- ✓ Transformación de diagramas de Entidad-Relación en tablas de base de datos.

Se decidió usar Visual Paradigm for UML en su versión 8.0 debido a todas las características mencionadas anteriormente. Además es una herramienta CASE profesional que soporta todo el ciclo de vida de desarrollo de software. Es utilizada en los centros de desarrollo de la UCI para modelar los procesos a través de diferentes tipos de diagramas. Posee una licencia comercial y la UCI cuenta con la misma para su uso.

1.9 Sistema Gestor de Base de Datos

Los Sistemas Gestores de Base de Datos (SGBD) relacionales son una herramienta efectiva que permite a varios usuarios acceder a los datos al mismo tiempo. Brindan facilidades eficientes y un grupo de funciones con el objetivo de garantizar la confidencialidad, la calidad, la seguridad y la integridad de los datos que contienen, así como un acceso fácil y eficiente a los mismos. (23)

1.9.1 PostgreSQL 9.1

Se decidió seleccionar como SGBD PostgreSQL en su versión 9.1. Es un gestor de base de datos de código abierto. Además se puede usar o modificar de forma gratuita, siendo de libre licencia, y encontrarse bajo los términos de licencia de código abierto: Berkeley Software Distribution (BSD). Por sus características técnicas este SGBD es uno de los más potentes y robustos del mercado. Su desarrollo comenzó hace más de 16 años, y durante este tiempo, estabilidad, potencia, robustez, facilidad de

administración e implementación de estándares han sido las características que más se han tenido en cuenta durante su desarrollo. PostgreSQL funciona muy bien con grandes cantidades de datos y una alta concurrencia de usuarios accediendo a la vez al sistema. (24)

Entre las principales características por lo cual se seleccionó a PostgreSQL como SGBD se encuentran las siguientes:

- ✓ Disponible para casi todos los principales sistemas operativos: Linux, Unix, BSDs, Mac OS, Beos, Windows, etc.
- ✓ Altamente adaptable a las necesidades del cliente.
- ✓ Soporte nativo para los lenguajes más populares del medio: PHP, C, C++, Perl, Python.
- ✓ Soporte de todas las características de una base de datos profesional (triggers, funciones, secuencias, relaciones, reglas, tipos de datos definidos por usuarios, vistas, vistas materializadas, etc.).
- ✓ Extensiones para alta disponibilidad, nuevos tipos de índices, datos espaciales, minería de datos, etc.
- ✓ Utilidades para análisis y optimización de Querys. (25)
- ✓ Numerosos tipos de datos y posibilidad de definir nuevos tipos.
- ✓ Llaves primarias y foráneas.
- ✓ Herencia de tablas (Inheritance). (24)

1.9.2 PgAdmin III 1.14.0

Es necesario utilizar una herramienta que gestione el SGBD seleccionado y para ello se utilizará PgAdminIII en su versión 1.14.0, el cual posee una licencia Open Source. PgAdmin III está diseñado para responder a las necesidades de todos los usuarios, desde escribir consultas SQL simples hasta desarrollar bases de datos complejas. El interfaz gráfico soporta todas las características de PostgreSQL y facilita enormemente la administración. La aplicación también incluye un editor SQL con resaltado de sintaxis, un editor de código de la parte del servidor, un agente para lanzar scripts programados, soporte para el motor de replicación Slony-I. (26)

1.10 Herramientas para los procesos de Extracción, Transformación y Carga

1.10.1 DataCleaner 1.5.4

En el proceso de la calidad de los datos existe una tarea muy importante a realizar: el perfilado de los datos, el cual se basa en hacer un análisis de los datos de la fuente, para estar al tanto de cómo es su estructura y su nivel de calidad.

DataCleaner es una aplicación Open Source para analizar, determinar comportamientos, transformar y limpiar datos. Todas estas actividades son utilizadas para mantener la calidad de los datos, lo cual es esencial para analizar cualquier negocio.

DataCleaner puede leer datos desde diferentes clases de archivos (XLS, CSV, TXT) así como también desde diferentes motores de base de datos. Funciona bajo una plataforma Java del tipo Web por lo que es compatible con todos los sistemas operativos que cuenten con la máquina virtual de Java instalada. (27)

Además de esta herramienta utilizada en el perfilado de las fuentes de datos, también sirvieron de apoyo para la generación de gráficas el Excel, posibilitando una mejor interpretación de los datos, así como el Access para el filtrado de mínimos y máximos valores numéricos de las diferentes tablas que conforman la fuente.

1.10.2 Pentaho Data Integration 4.2.1

Pentaho Data Integration también conocida como Kettle es una herramienta que permite implementar los procesos de extracción, transformación y carga de datos (ETL), la misma es de código abierto compuesta por cuatro componentes fundamentales: SPOON para el diseño gráfico de las transformaciones, PAN para la ejecución de los trabajos y las transformaciones, CHEF para el diseño de la carga de datos y KITCHEN para la ejecución de los trabajos Batch diseñados con CHEF. El uso de Kettle permite evitar grandes cargas de trabajo manual frecuentemente difícil de mantener y de desplegar. Se observa que ha sido diseñado para cubrir las necesidades en la integración de datos. (28)

Características básicas:

- ✓ Entorno gráfico de desarrollo.
- ✓ Uso de tecnologías estándar: Java, XML, JavaScript.
- ✓ Multiplataforma: Windows, Macintosh, Linux.
- ✓ Basado en dos tipos de objetos: Transformaciones (colección de pasos en un proceso ETL) y trabajos (colección de transformaciones). (29)

Conclusiones del Capítulo

Con el fin de llevar a cabo los procesos de integración de las fuentes de datos para cada uno de los EC que se gestionan en el CIM, se realizó un estudio sobre las diferentes bibliografías relacionadas con los temas de AD y MD con el objetivo de adoptar la metodología a utilizar en la investigación. Se analizó un conjunto de herramientas con el propósito de seleccionar las más acordes con las necesidades del MD que abarcarán los procesos de desarrollo del mismo. Se arribaron a las siguientes conclusiones:

- ✓ Se decidió utilizar como metodología de desarrollo para la construcción de los subsistemas de almacenamiento e integración del producto egf del área de los EC, la propuesta por el departamento DATEC; para guiar el proceso de desarrollo a través de todo el ciclo de vida.
- ✓ Se decidió utilizar como herramienta de modelado Visual Paradigm for UML en su versión 8.0 para la generación de los diagramas que forman parte de la solución.
- ✓ Se decidió emplear como SGBD PostgreSQL 9.1 y el PgAdmin1.14.0 para la administración de la información.
- ✓ Las herramientas seleccionadas para el proceso de ETL fueron DataCleaner 1.5.4 para el perfilado de los datos y Pentaho Data Integration 4.2.1 para implementar los procesos de integración.

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

2.1 Introducción

En este capítulo se realiza un análisis del negocio, con el objetivo de identificar los principales aspectos de relevancia para la organización, definiéndose los requisitos de información del sistema, agrupados en casos de uso de acuerdo al tipo de información. Se definen las reglas del negocio en conjunto con el cliente para poder adaptar la solución a las necesidades que son requeridas. Se identifican las dimensiones, hechos y medidas para poder diseñar el modelo de datos, con el propósito de lograr un mejor entendimiento a través de la representación gráfica de los elementos mencionados anteriormente. Se realiza el diseño de los dos subsistemas: integración y almacenamiento que conforman la solución a desarrollar y se diseña el esquema de seguridad para evitar la manipulación de la información por personas no autorizadas.

2.2 Análisis del negocio

2.2.1 Necesidades del negocio

El objetivo principal de un sistema a desarrollar se centra en la identificación de las necesidades del negocio. Una fase muy importante es la identificación de los principales elementos a tener en cuenta para satisfacer las necesidades de la entidad.

La investigación tiene como objetivo darle solución a una serie de problemas identificados en el CIM en cuanto a la gestión de la información sobre los EC que son aplicados a pacientes que padecen de cáncer de pulmón. Por lo que resulta de vital importancia la recogida de los resultados arrojados del producto egf, para facilitar un análisis de la información referente al negocio por parte de los especialistas y se definen las bases para realizar un correcto diseño de los subsistemas de almacenamiento e integración del producto egf.

Luego de varias entrevistas con los especialistas se decide clasificar la información en once grupos:

- ✓ EC 019.
- ✓ EC 025.
- ✓ EC 033.
- ✓ EC 041.
- ✓ EC 062.
- ✓ Título de anticuerpo para los EC 033 y 041.
- ✓ Forma vacunal para el EC 041.

- ✓ Salida del ensayo para el EC 041.
- ✓ Anomalías para el EC 062.
- ✓ Salida del ensayo para el EC 062.
- ✓ Título de anticuerpo para el EC 062.

Estos se hacen corresponder con los temas de información obtenidos, los cuales satisfacen las necesidades del usuario final y facilitan los procesos de ETL a la hora de cargar los datos.

2.3 Especificación de requisitos

El análisis de requisitos constituye una de las fases más importantes en la construcción de un MD. Los requisitos dentro del negocio constituyen la descripción de los servicios que ha de ofrecer el sistema y las restricciones asociadas a su funcionamiento. En dicha fase son definidos los requisitos de información, funcionales y no funcionales del sistema, partiendo de las necesidades de los clientes.

2.3.1 Requisitos de información

Los requisitos de información (RI) describen qué información debe almacenar el sistema para satisfacer las necesidades de los clientes. Identifican los conceptos relevantes sobre los que se debe almacenar la información y los datos específicos que son de interés para los especialistas. Los RI identificados durante el proceso de análisis fueron clasificados según los temas de información definidos. Estos se encargan de mantener disponible la información referente a los EC del producto egf.

A continuación se mencionan los requisitos correspondientes a los temas de información de la “Forma vacunal para el EC 041” y el “EC 019”. Los demás se encuentran detallados en el Expediente de Proyecto, en el artefacto: “DATEC_CIM_EGF_20_02_2013_0113_Especificación de Requisitos de Software”.

RI1. Obtener la cantidad de pacientes del EC 041 por forma vacunal.

RI2. Obtener la cantidad de pacientes del EC 019 por edad, talla, peso, raza, sexo, índice de Karnofsky, estadio, hospital, localización del tumor, examen físico y diagnósticos.

RI3. Obtener la cantidad de pacientes del EC 019 por salida del ensayo, causa de fallecimiento, reacciones adversas, signos vitales, examen de laboratorio, respuesta inmunológica y respuesta clínica.

RI4. Obtener la cantidad de pacientes del EC 019 por criterios de inclusión, radioterapia, quimioterapia, quirúrgico, tratamiento previo, enfermedad asociada, tratamiento concomitante, RX, US, TAC y gammagrafía.

2.3.2 Requisitos funcionales

Los requisitos funcionales (RF) representan las condiciones o capacidades que el sistema debe cumplir. Estos incluyen las funcionalidades que deben implementarse en los subsistemas de almacenamiento e integración del producto egf. Los mismos se describen a continuación:

RF1. Extraer datos de la fuente.

RF2. Realizar la transformación y carga de los datos.

2.3.3 Requisitos no funcionales

Los requisitos no funcionales (RNF) describen las características, cualidades o propiedades que el producto debe tener. Definen las propiedades y restricciones del sistema. Después de un previo encuentro con el cliente se definieron siete requisitos no funcionales. Estos son descritos en el artefacto: "DATEC_CIM_EGF_20_02_2013_0113_Especificación de Requisitos de Software". A continuación se mencionan algunos de ellos.

RNF1. Utilizar la herramienta de integración de datos definida durante la investigación: Para el proceso de integración de datos se usará la herramienta Pentaho Data Integration en su versión 4.2.1.

RNF2. Utilizar el Sistema Gestor de Base de Datos definido durante la investigación: El SGBD que se utilizará es PostgreSQL en su versión 9.1 y como interfaz de administración de dicho gestor PgAdmin III en su versión 1.14.0.

2.4 Reglas del negocio

Las reglas de negocios (RN) definen y controlan la estructura, el funcionamiento y la estrategia de una organización a través de políticas, restricciones y normas que son establecidas y aplicadas en el negocio. Estas pueden estar formalmente definidas en manuales de procedimientos, contratos, acuerdos, o pueden existir como conocimiento y experiencia que tengan los involucrados en el negocio. Durante el análisis fueron identificadas las siguientes RN:

Reglas de variables:

RN1. Para acceder a la fuente de datos del producto egf se necesita una contraseña de acceso proporcionada por el cliente para ser manejadas por el equipo de desarrollo.

Reglas de Almacenamiento

RN2. Las dimensiones de tipo varchar tendrán como máximo una longitud de 200 caracteres.

RN3. Las dimensiones de tipo entero tendrán como máximo una longitud de 4 caracteres.

Reglas de Transformación

RN4. En los EC 019, 025, 033, 041 y 062 en causa de fallecimiento se sustituyen en Cual Causa de

Muerte la causa de la muerte en caso de que Otra Causa de Muerte sea verdadera.

RN5. En los EC 019, 025 y 033 el Diagnóstico Citológico si está vacío se sustituye por: No disponible.

RN6. En el EC 025 el Diagnóstico Clínico si está vacío se sustituye por el que más se repite.

RN7. En los EC 019 y 025 el examen físico si es cero: No disponible, 1: Positivo y 2: Negativo.

RN8. La forma vacunal del EC 025 se sustituye por: si es 1: Hidróxido de Alúmina y si es 2: Montanide 51, si se encuentra vacío se sustituye por la que más se repite.

RN9. La forma vacunal del EC 033 se sustituye por: si es Alúmina: Alúmina Hidróxido, si es Montanide: Montanide ISA 51 y si se encuentra vacío se sustituye por la que más se repite.

Las demás RN identificadas se describen en el artefacto: "DATEC_CIM_EGF_20_02_2013_Reglas de negocio y transformación".

2.5 Casos de uso del sistema

Durante la fase de análisis y diseño son definidos los casos de uso del sistema (CUS). Los casos de uso (CU) son "fragmentos" de funcionalidad que el sistema ofrece. Los RI y RF identificados son agrupados en casos de uso de información (CUI) y casos de uso funcionales (CUF) respectivamente.

2.5.1 Actores del sistema

Actor	Objetivo
Analista	Responsable de inicializar los CU relacionados con la información de los datos de los subsistemas de almacenamiento e integración para los EC del producto egf para su análisis.
Administrador ETL	Responsable de llevar a cabo los procesos de extracción, transformación y carga de los datos del sistema fuente.

Tabla 1: Actores del sistema

2.5.2 Casos de uso de información

Los CUI van a estar agrupados según el tipo de información que aporten al negocio, lo que sería por temas de información. A continuación se mencionan los CU definidos para los subsistemas de almacenamiento e integración del producto egf:

CUI1. Mantener disponible la información del producto egf de los pacientes del EC 019.

CUI2. Mantener disponible la información del producto egf de los pacientes del EC 025.

CUI3. Mantener disponible la información del producto egf de los pacientes del EC 033.

CUI4. Mantener disponible la información del producto egf de los pacientes del EC 041.

CUI5. Mantener disponible la información del producto egf de los pacientes del EC 062.

CUI6. Mantener disponible la información del producto egf de los pacientes de los títulos de anticuerpo de los EC 033 y EC 041.

CUI7. Mantener disponible la información del producto egf de los pacientes de la forma vacunal del EC 041.

CUI8. Mantener disponible la información del producto egf de los pacientes de la salida del ensayo del EC 041.

CUI9. Mantener disponible la información del producto egf de los pacientes por las anomalías del EC 062.

CUI10. Mantener disponible la información del producto egf de los pacientes de la salida del ensayo del EC 062.

CUI11. Mantener disponible la información del producto egf de los pacientes de los títulos de anticuerpo para el EC 062.

2.5.3 Casos de uso funcionales

Los CUF identificados en el negocio se basan en los RF para la ejecución de las operaciones de ETL que serían aplicadas a las fuentes de datos relacionadas con los EC. A continuación se mencionan los CUF definidos para los subsistemas de almacenamiento e integración del producto egf:

CUF1. Realizar la extracción de los datos: se realiza la extracción de los datos necesarios de las diferentes fuentes.

CUF2. Realizar la transformación y carga de los datos: se realiza la transformación y carga de los datos necesarios para la construcción del MD.

A continuación en la tabla se describe el CUF: Realizar la extracción de los datos (Consultar el artefacto: “DATEC_CIM_EGF_20_02_2013_0114_Especificación de casos de uso”).

Caso de Uso:	Realizar extracción de los datos.
Tipo:	Funcional.
Actores:	Administrador ETL.
Resumen	El caso de uso inicia cuando el actor selecciona los datos a extraer. Se extraen los datos de la fuente y finaliza el caso de uso.
Precondiciones:	Disponibilidad de las fuentes.
Referencias	RF1

Prioridad	Media	
Flujo Normal de Eventos		
Acción del Actor		**Respuesta del sistema**
1. El administrador de ETL realiza la conexión a la fuente.		2. Responde a la solicitud de conexión.
3. El administrador de ETL selecciona el archivo a extraer.		
4. El administrador de ETL realiza la extracción de los datos.		5. Ejecuta la extracción de los datos. 6. 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.
Pos condiciones	Los datos de la fuente correspondiente han sido extraídos de la fuente de datos y almacenados en la base datos Producto_egf.	

Tabla 2: Realizar la extracción de los datos

2.5.4 Diagrama de casos de uso

Para lograr un mejor entendimiento del funcionamiento del sistema, se realiza una representación gráfica de los actores que intervienen en el negocio y la relación existente de los CU con los mismos para proporcionar una visión general del modelo de CU. En el diagrama de casos de uso (DCU), los CUI fueron agrupados por temas de información. En la figura 2 se muestra el diagrama de CUS correspondiente a la presente investigación.

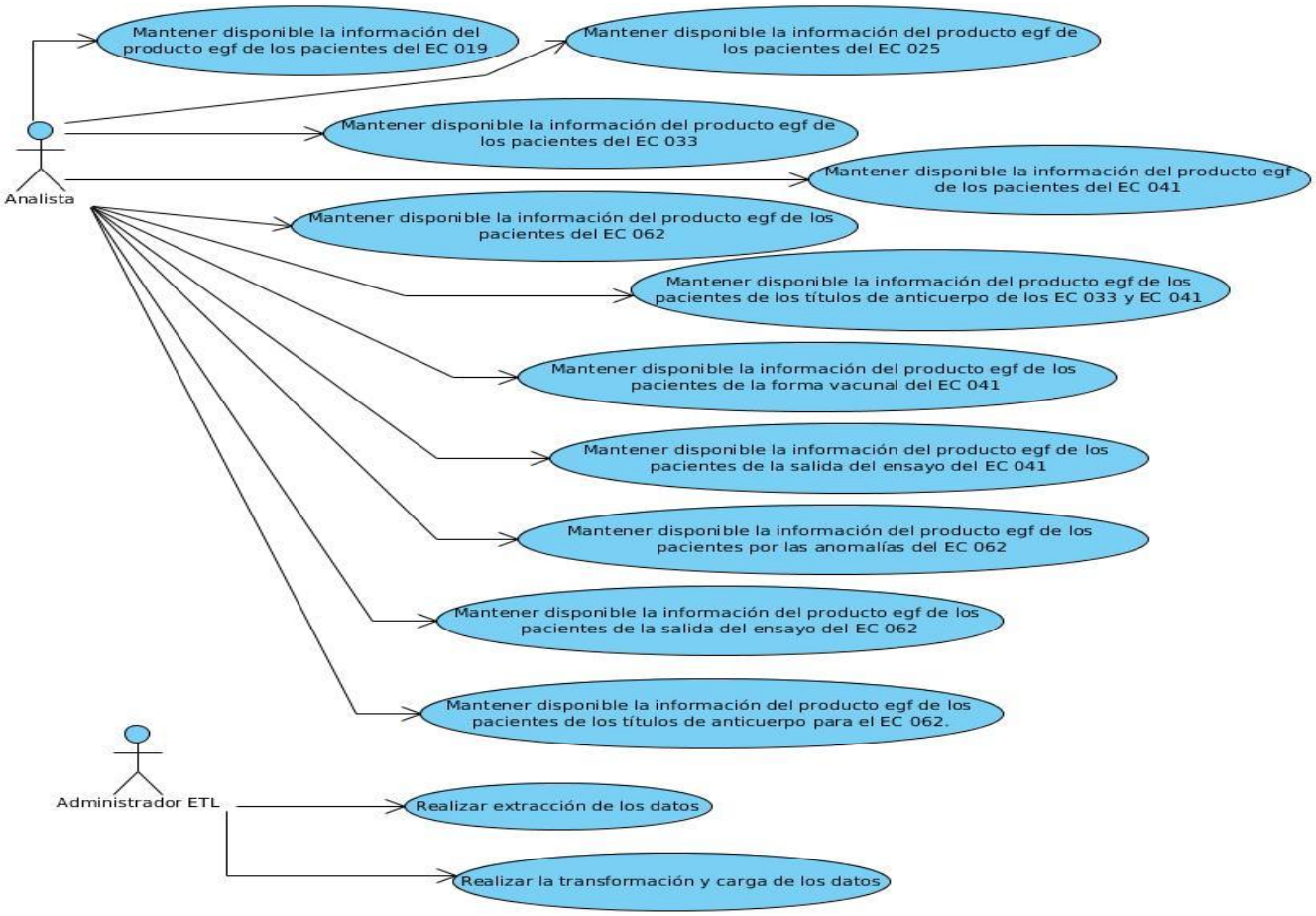


Figura 2: Diagrama de CUS

2.6 Definición de la arquitectura

En la arquitectura se debe tener en cuenta la forma de representar los datos globales en el negocio, los requisitos del sistema y las restricciones a las que está sujeto, las reglas y diseño a considerar, así como la presentación al usuario final. Todo esto con el objetivo de lograr flexibilidad ante cambios que necesite la organización que implemente MD.

La arquitectura lógica de este tipo de sistema consta de cuatro niveles:

1. **Fuente de datos:** contiene los datos fuente.
2. **Diseño del subsistema de integración de datos:** incluye los procesos que permiten que los datos sean extraídos de las fuentes, transformados e integrados en la fuente destino.
3. **Diseño del subsistema de almacenamiento:** base de datos que contiene las tablas de dimensiones y hechos cargadas a través de los procesos de ETL.

4. **Diseño del subsistema de visualización de información:** comprende las interfaces orientadas a usuarios que extraen información para la toma de decisiones.

La arquitectura definida para el desarrollo de la solución está conformada por dos subsistemas: integración y almacenamiento. Ver figura 3.

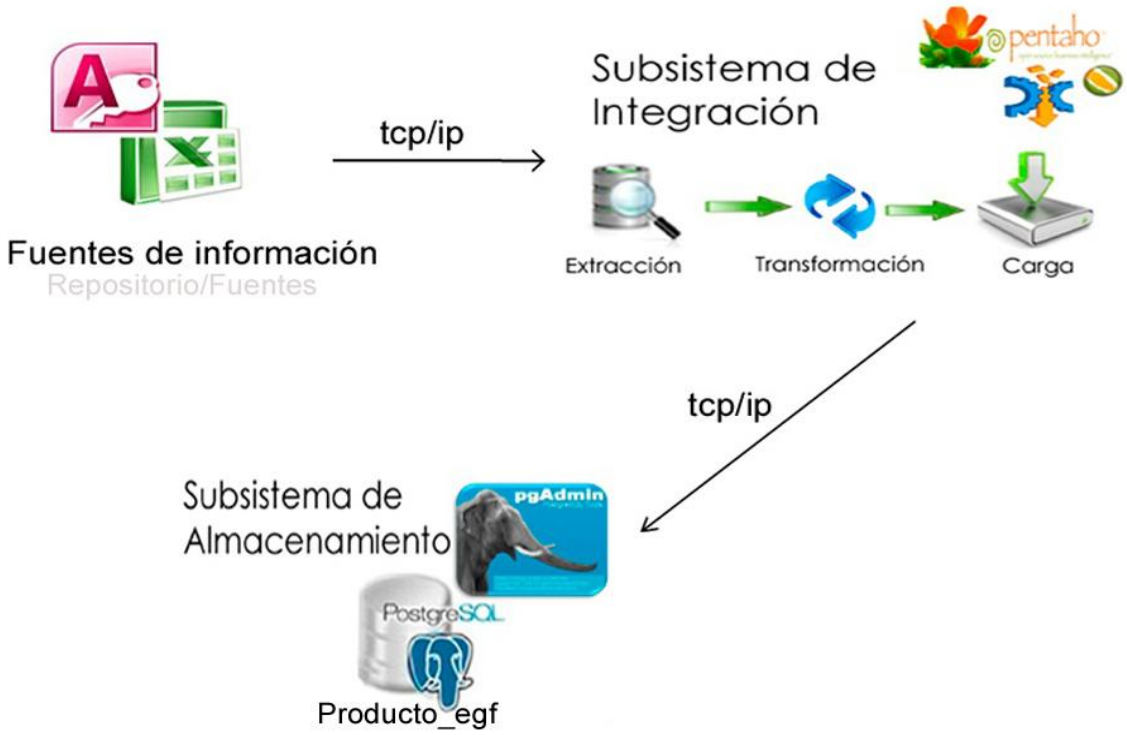


Figura 3: Diseño de la arquitectura

El subsistema de integración se abastece de las fuentes de información. Este es el encargado de realizar los procesos de extracción, transformación y carga de los datos hacia el subsistema de almacenamiento.

El subsistema de almacenamiento se alimenta de los datos que son cargados en los procesos de ETL. La información es almacenada en una base de datos relacional en el SGBD PostgreSQL y se administra a través del PgAdmin III.

2.7 Diseño de los subsistemas de almacenamiento e integración

La solución propuesta consiste en la implementación de los subsistemas de almacenamiento e integración del producto egf para apoyar a la toma de decisiones sobre la información de los EC manejadas en el CIM. Permite a los especialistas trabajar sobre fuentes estandarizadas que proporcionan mayor impacto

en los resultados de sus investigaciones, donde el análisis, comparaciones y búsquedas de alternativas para probar soluciones desarrolladas por ellos sería más accesible.

2.7.1 Diseño del subsistema de almacenamiento

Al diseñar la solución propuesta descrita anteriormente es necesario tener en cuenta la identificación de las dimensiones y tablas de hechos con sus medidas asociadas correspondientes.

2.7.2 Dimensiones

Como parte del diseño se identificaron las dimensiones candidatas que luego de analizarlas minuciosamente formarán parte de la solución. Las dimensiones representan las perspectivas de análisis de la información, incluyen los diferentes atributos que se pueden analizar, que además se estructuran de forma jerárquica, conforme a diferentes niveles de detalle. Cada una de las dimensiones tiene una clave que identifica cada uno de los registros que la conforman. En la solución se incluyen las llaves subrogadas debido a que habitualmente no es posible utilizar la clave del negocio como clave principal. Una clave subrogada es un identificador único que es asignado a cada fila de la tabla de dimensiones, en definitiva, será su clave principal. Son siempre de tipo numérico y asignado de forma secuencial. (14)

A continuación se describen algunas dimensiones, la restantes pueden ser consultadas en el artefacto: “DATEC_CIM_EGF_20_02_2013_Especificación del modelo de datos”:

- ✓ **Dimensión raza** (dim_raza): define el conjunto de valores bajo los cuales puede clasificarse la información correspondiente a la raza.
- ✓ **Dimensión sexo** (dim_sexo): define el conjunto de valores bajo los cuales puede clasificarse la información correspondiente al sexo.
- ✓ **Dimensión hospital** (dim_hospital): define el conjunto de valores bajo los cuales puede clasificarse la información correspondiente a los hospitales.
- ✓ **Dimensión edad** (dim_edad): define los valores que puede tomar la edad del paciente.
- ✓ **Dimensión talla** (dim_talla): define los valores que puede tomar la talla del paciente.
- ✓ **Dimensión estadio** (dim_estadio): define el conjunto de valores bajo los cuales puede clasificarse la información correspondiente al estadio.
- ✓ **Dimensión estado OMS** (dim_estado_OMS): define el conjunto de valores bajo los cuales puede clasificarse la información correspondiente al estado OMS.
- ✓ **Dimensión índice de karnofsky** (dim_karnofsky): define los valores que puede tomar el índice de karnofsky.

- ✓ **Dimensión examen de laboratorio** (dim_examen_laboratorio): define el conjunto de valores que puede tomar el examen de laboratorio.
- ✓ **Dimensión forma vacunal EC 041** (dim_forma_vacunal_EC041): define el conjunto de valores que puede tomar la forma vacunal, así como el lote y la dosis presentes en la misma.

2.7.3 Dimensiones degeneradas

Las dimensiones degeneradas constituyen un campo de dato que será almacenado en las tablas de hechos, en lugar de ser definidas como una dimensión. Este tipo de dimensiones son utilizadas en la solución, las mismas fueron incluidas en algunas tablas de hechos con el objetivo de reducir la duplicación de datos, simplificar las consultas y optimizar el diseño del modelo de datos. En el hecho del EC 019 las dimensiones criterios de inclusión, radioterapia, quimioterapia, quirúrgico, tratamiento previo, enfermedad asociada, tratamiento concomitante, Rayos x, tomografía axial computarizada (TAC), ultrasonido (US) y gammagrafía son dimensiones degeneradas definidas dentro del hecho. Estas representan un campo de dato de tipo boolean, no tienen asociados atributos, por lo que no se hace necesario definirles una nueva tabla.

2.7.4 Tablas de hechos

Las tablas de hechos diseñadas para la solución propuesta almacenan las llaves asociadas de cada una de las dimensiones, algunas métricas y un código de paciente para aquellos hechos que no presentan medidas numéricas, que a partir de ello permite obtener la cantidad de los mismos. Los hechos que no contienen medidas fueron concebidos así debido a las características presentadas en el negocio a la hora de agrupar la información de cada paciente. En el desarrollo de los subsistemas de almacenamiento e integración del producto egf fueron identificados once hechos, los cuales se describen a continuación:

- ✓ **Hecho EC019** (hech_EC019): se almacena toda la información concerniente a los pacientes del EC 019.
- ✓ **Hecho EC025** (hech_EC025): se almacena toda la información concerniente a los pacientes del EC 025.
- ✓ **Hecho EC033** (hech_EC033): se almacena toda la información concerniente a los pacientes del EC 033.
- ✓ **Hecho EC041** (hech_EC041): se almacena toda la información concerniente a los pacientes del EC 041.
- ✓ **Hecho EC062** (hech_EC062): se almacena toda la información concerniente a los pacientes del EC 062.

- ✓ **Hecho título de anticuerpo para EC 033 y 041** (hech_titulo_anticuerpo033_041): se almacena toda la información concerniente a los títulos de anticuerpos de los pacientes en los EC 033 y 041.
- ✓ **Hecho forma vacunal EC 041** (hech_forma_vacunal_EC041): se almacena toda la información concerniente a la forma vacunal de los pacientes del EC 041. Se define como medida numérica cantidad de pacientes.
- ✓ **Hecho salida de ensayo EC 041** (hech_salida_ensayo_EC041): se almacena toda la información concerniente de la salida de los pacientes del EC041. Se define como medida numérica cantidad de pacientes.
- ✓ **Hecho anomalías EC 062** (hech_anomalias_EC062): se almacena toda la información concerniente sobre anomalías presentes en el paciente del EC 062.
- ✓ **Hecho salida de ensayo EC 062** (hech_salida_ensayo_EC062): se almacena toda la información concerniente sobre la salida de los pacientes del EC062. Se define como medida numérica cantidad de pacientes.
- ✓ **Hecho título de anticuerpo para EC 062** (hech_titulo_anticuerpo_EC062): se almacena toda la información concerniente a los títulos de anticuerpos de los pacientes del EC 062.

2.7.5 Matriz buz o matriz dimensional

Para validar el correcto diseño del modelo de datos se realiza la matriz dimensional. Son representadas las relaciones existentes entre las dimensiones y las tablas de hechos, así como las intercepciones entre ellos: representados con una x. A través de la misma se puede apreciar la existencia o no de solapamiento. Ver tabla 3.

Dimensiones	Hechos										
	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11
Edad	x	x	x		x			x			
Talla	x	x	x		x			x			
Peso	x	x	x		x			x			
Raza	x	x	x		x			x			

Sexo	x	x	x		x			x			
Hospital	x	x	x	x	x	x	x	x	x	x	x
Diagnóstico	x	x	x								
Estadio	x	x	x		x			x			
Índice de Karnofsky	x	x						x			
Localización Tumor	x	x	x		x			x			
Examen Físico	x	x									
Evaluación de signos vitales	x	x	x					x			
Localización Metástasis		x	x					x			
Estado OMS		x	x		x						
TNM			x		x			x			
Medicamento			x								
Forma vacunal EC033			x								
Título de anticuerpo				x							x
Forma vacunal EC041					x	x	x				

Provincia								x	x	x	x
Localización L1								x			
No Vacunación								x			
Tipo evento adverso								x			
Causalidad evento adverso								x			
Grado evento adverso								x			
Tiempo	x	x	x	x	x	x	x	x	x	x	x
Causa salida EC062										x	
Anomalías			x		x			x	x		
Forma Vacunal		x									
Causa fallecimiento	x	x	x				x			x	
Salida de Ensayo	x	x	x				x				
Respuesta clínica	x	x	x								

Respuesta inmunológica	**x**										
Examen de laboratorio	**x**	**x**	**x**					**x**	**x**		
RA	**x**		**x**		**x**						

Tabla 3: Matriz dimensional

Leyenda

H1: hech_EC019

H2: hech_EC025

H3: hech_EC033

H4: hech_titulo_anticuerpo033_041

H5: hech_EC041

H6: hech_forma_vacunal_EC041

H7: hech_salida_ensayo_EC041

H8: hech_EC062

H9: hech_anomalias_EC062

H10: hech_salida_ensayo_EC062

H11: hech_titulo_anticuerpo_EC062

La realización de la matriz dimensional permitió dar a conocer que de los once hechos definidos para el diseño del modelo de datos, todos tienen dimensiones asociadas y no existen solapamiento entre ellos. El solapamiento se refiere a que no pueden existir dos o más hechos que compartan exactamente las mismas dimensiones. Para poder solucionar este tipo de problemas una alternativa aplicada en la solución fue la unión de dos hechos que se solapaban en el diseño, esta estrategia eliminó el problema detectado.

2.7.6 Esquemas multidimensionales

El modelo multidimensional contiene tres variantes de modelación, determinado por la complejidad del sistema, estas son:

Esquema de estrella: es el más sencillo de los esquemas de almacenamiento de datos, es ideal por su simplicidad y velocidad para ser usado en análisis multidimensionales como los modelos de datos. Permite acceder a los datos agregados. Consiste en estructurar la información en procesos, vistas y métricas. El mismo está conformado por una tabla de hecho que se relaciona con las tablas de dimensiones identificadas en el negocio. En este las dimensiones no se normalizan, con ello se logra minimizar el número de uniones e incrementar el rendimiento de las consultas. En la figura 4 se muestra la estructura que posee este tipo de modelo. (19)

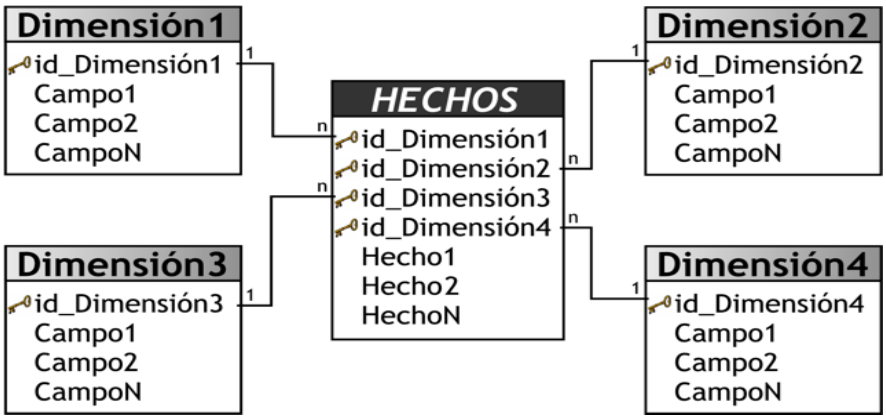


Figura 4: Esquema de estrella

Esquema de copo de nieve: también conocido como snowflake es un esquema de representación derivado del esquema en estrella, en el que las tablas de dimensión se normalizan en múltiples tablas. Por esta razón, las tablas de hechos dejan de ser la única tabla del esquema que se relaciona con otras tablas. El único argumento a favor de los esquemas en copo de nieve es que al estar normalizadas las tablas de dimensiones, se evita la redundancia de datos y con ello se ahorra espacio. (19) En la figura 5 se muestra la estructura que posee este tipo de modelo.

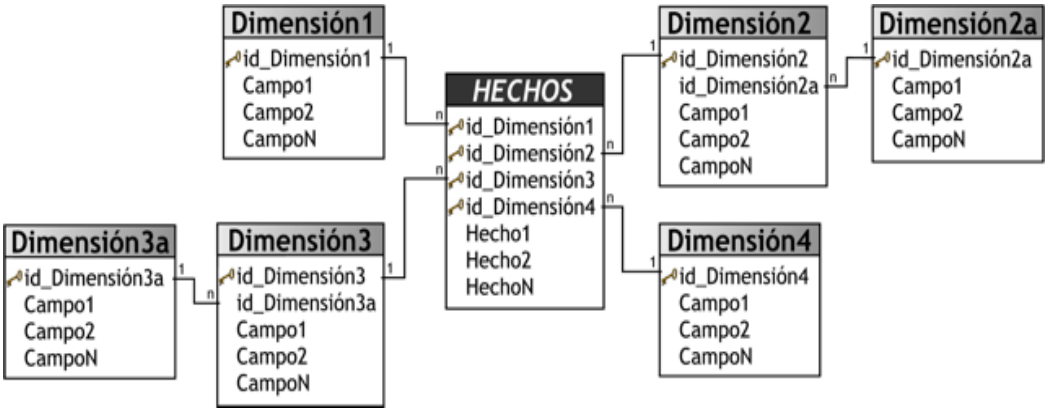


Figura 5: Esquema de copo de nieve

Esquema de constelación de hechos: es un conjunto de tablas de hechos que comparten algunas tablas de dimensiones. (30) Contribuye a la reutilización de las tablas de dimensiones, por lo que una misma tabla de dimensión puede ser utilizada para varias tablas de hechos. Un esquema de constelación es una combinación de un esquema de estrella y un esquema de copo de nieve. Ver figura 6.

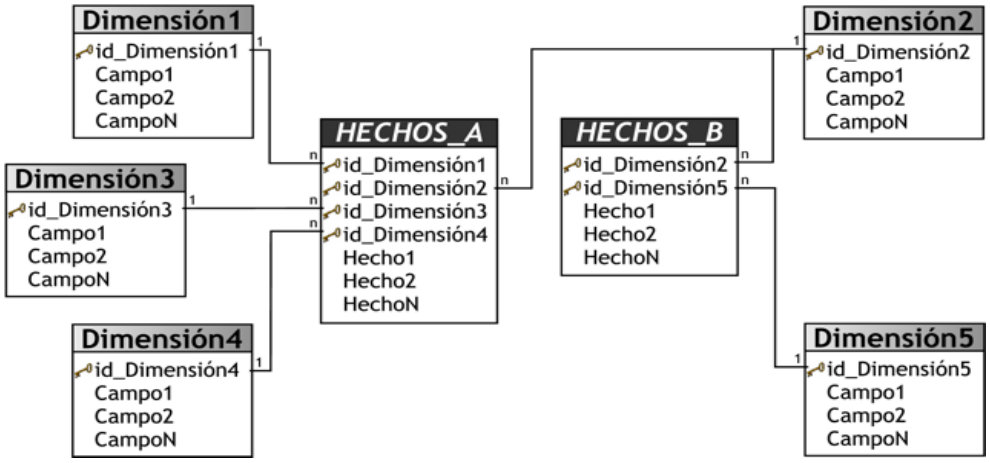


Figura 6: Esquema constelación de hechos

Atendiendo a los esquemas mencionados anteriormente, la confección de la matriz dimensional y las particularidades de las dimensiones identificadas en el negocio, se seleccionó el esquema constelación de hechos, debido a que este es el más adecuado teniendo en cuenta su diseño.

2.7.7 Modelo de datos

En el negocio, luego de definidas las dimensiones y medidas se procede a la realización del modelo dimensional. A continuación en la figura 7 se muestra un fragmento del modelo de datos para el desarrollo de la solución, donde se evidencia el uso de la topología constelación de hechos:

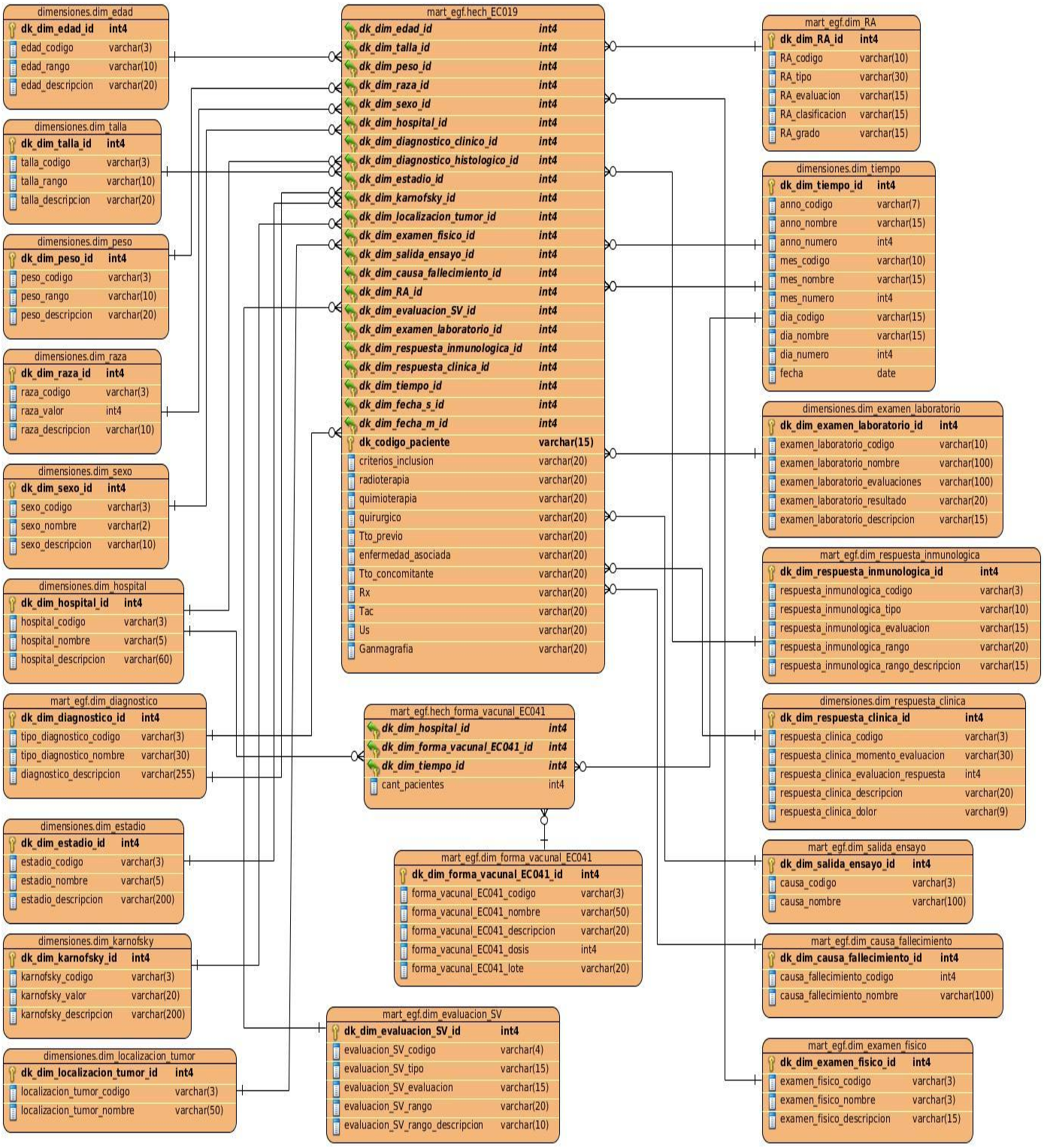


Figura 7: Modelo de datos

2.8 Diseño del subsistema de integración

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

2.8.1 Estrategias de calidad de datos

Existen diferentes estrategias de calidad de datos que son utilizadas para inspeccionar y encontrar errores provenientes de las fuentes de datos. Algunas de estas son: perfilado de datos, Matching, identificación de duplicados mediante técnicas de lógica difusa, mejora de datos, limpieza de datos, procesos de calidad de datos, entre otros. A continuación se describe la estrategia de calidad de datos empleada.

Perfilado de los datos

El perfilado de los datos, no es más que el análisis de los datos para entender su contenido, estructura, calidad y dependencias. Se realiza con el objetivo de identificar el estado actual de los datos en las fuentes y así poder establecer estrategias para corregir, eliminar o tratar los errores identificados. Se verifica le existencia de campos nulos, vacíos, distintos y duplicados para definir nuevas reglas de negocio, que pasarían a ser reglas de transformación. Para obtener más información referente con el perfilado, puede ser consultado el artefacto: “DATEC_CIM_EGF_22_02_2013_Perfil_de_los_Datos”.

Al realizar el perfilado de los datos a las fuentes de los EC del producto egf, fueron obtenidos los siguientes resultados:

- ✓ Los tipos de datos encontrados fueron: double 28%, date 13%, integer 22%, varchar 7% y boolean 31%. De los cuales el tipo de dato más usado es el boolean. Ver figura 8.

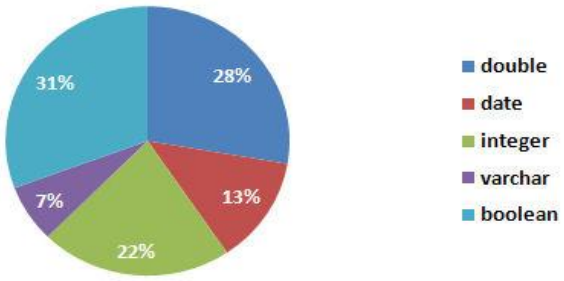


Figura 8: Porcentaje de los tipos de datos

- ✓ Los datos de tipo fecha encontrados tienen formatos distintos: dd/mm/yyyy y dd-mmm-yy. Para solucionar este problema se decidió como máscara de fecha a utilizar: dd/mm/yyyy.
- ✓ Cuando las fechas tengan valores nulos van a ser sustituidos por 00/00/0000. Ver figura 9.

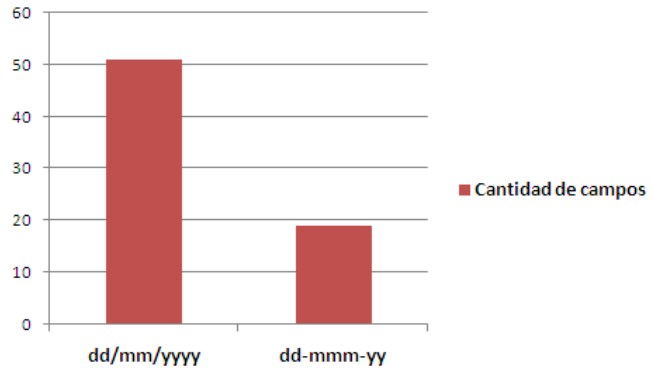


Figura 9: Formato fecha

- ✓ Los rangos de tipo fecha máximos y mínimos encontrados fueron: 1996, 1997,1998, 1999, 2000, 2001, 2002, 2003 y 2004. Este análisis permite conocer los años involucrados y la cantidad de ocurrencias que existen por cada uno de ellos. A la hora de definir la dimensión de tipo fecha esta debe incluir datos comprendidos entre los años 1996 y 2004. Ver figura 10.

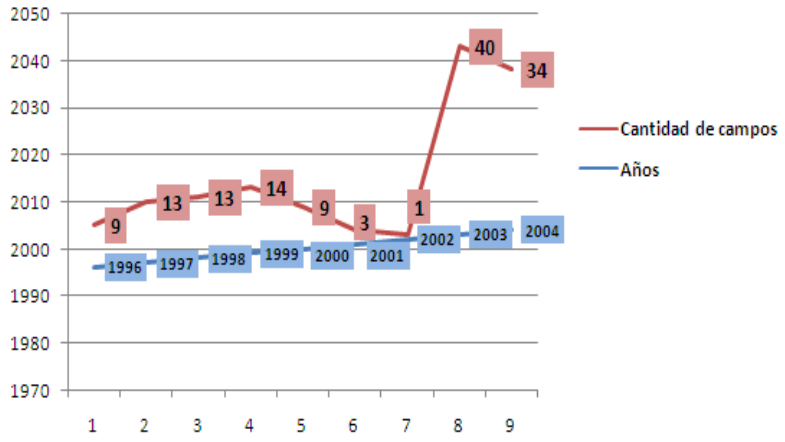


Figura 10: Valores máximos y mínimos de tipo fecha

2.8.2 Diseño de los procesos de integración

El diseño general para la carga de las dimensiones de los subsistemas de almacenamiento e integración del producto egf se describe a continuación: realizándose como primer paso la búsqueda de los ficheros fuentes, donde se verifica la existencia de los mismos y en caso de no existir se envía hacia un Excel de error. Si los ficheros existen se extraen los datos que se cargarán, se verifica si son de formato xls o mdb,

para su posterior análisis. En caso de ser un Excel se inserta la dimensión en la tabla correspondiente a los esquemas mart_egf o dimensiones. Si el archivo es Access, se verifica la existencia de valores nulos, si existen estos valores se realizan estrategias para corregirlos, poniéndose como ejemplo en el campo: número de vacunación en caso de no existir el valor se sustituye por no disponible. El cliente es quien define las estrategias para corregir errores detectados a la hora de insertar los datos en la base de datos, definiendo qué valores necesita que aparezcan en caso de encontrarse campos vacíos. Cuando no existan valores nulos se inserta la dimensión en la tabla correspondiente al esquema mart_egf o dimensiones. Luego se obtiene la información del sistema para generar los metadatos, en este caso técnicos, donde se almacena la fecha de inicio y fin de la carga, la dirección ip de la computadora donde se ejecutó la transformación y el nombre de la misma. La figura 11 muestra el diseño realizado para la carga de las dimensiones:

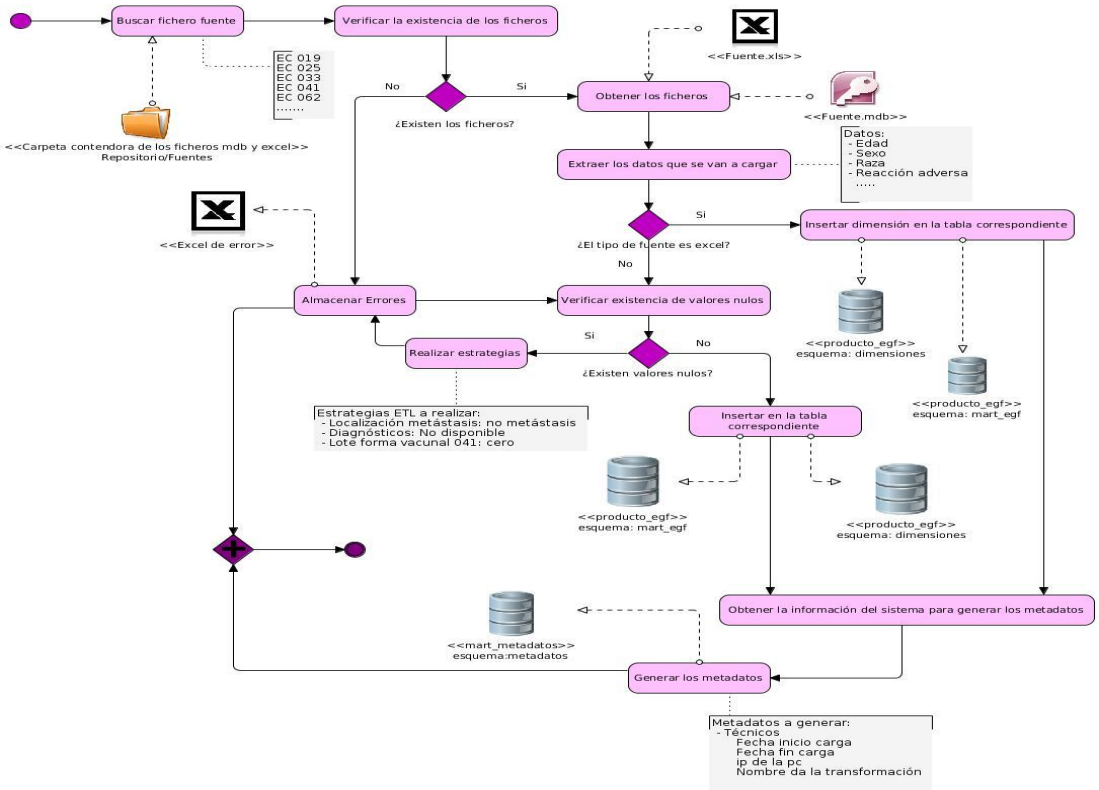


Figura 11: Diseño de la transformación para cargar dimensiones

El diseño general para la carga de los hechos de los subsistemas de almacenamiento e integración del producto egf es el siguiente: se realiza como primer paso la búsqueda de los ficheros fuentes, donde se verifica la existencia de los mismos y en caso de no existir se envía hacia un Excel de error. En caso de

existir dichos ficheros se extraen los datos que serán cargados para su posterior análisis. Se verifica la existencia de valores nulos, si existen, se realizan estrategias para corregirlos, poniéndose como ejemplo en el campo: metástasis en caso de no existir el valor se sustituye por no metástasis. El cliente es quien define las estrategias para corregir errores detectados a la hora de insertar los datos en la base de datos, definiendo qué valores necesita que aparezcan en caso de encontrarse campos vacíos. Cuando no existan valores nulos se verifica la existencia de llaves nulas, si existen se envía hacia un Excel de error. Luego se transforman estas llaves para realizar la búsqueda en la dimensión. En caso de no existir llaves nulas, se buscan las llaves dimensionales y se inserta en la tabla correspondiente al esquema mart_egf. Luego se obtiene la información del sistema para generar dos tipos de metadatos: técnicos, donde serán almacenados la fecha de inicio y fin de la carga, la dirección ip de la computadora donde se ejecutó la transformación y el nombre de la misma y de procesos que contendrá las líneas leídas, líneas escritas, los errores, entre otros datos que brindan información acerca de la transformación que se cargó. La figura 12 muestra el diseño para la carga de los hechos:

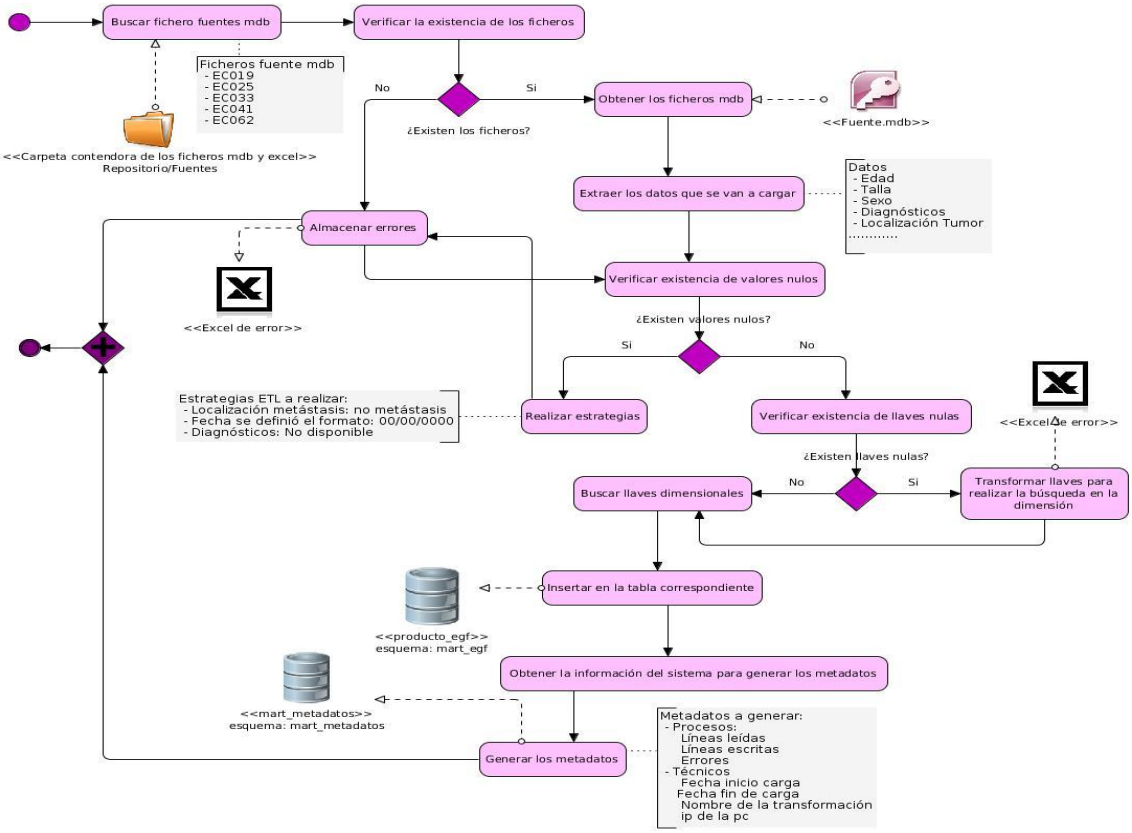


Figura 12: Diseño de la transformación para cargar los hechos

2.9 Política de respaldo y recuperación

La información es un recurso extremadamente valioso para cualquier entidad. La pérdida de la misma traería consecuencias tanto monetarias como prestigiosas para la organización. Las tecnologías no están exentas de fallas o errores, y los respaldos de información son utilizados como un plan de contingencia en caso de que una falla o error se presente.

Con el objetivo de lograr la persistencia y seguridad de la información que se maneja en el CIM y por la importancia que la misma provee a los especialistas de esa empresa, se establece una política de respaldo y recuperación la cual estaría comprendida dentro de dos elementos esenciales:

1. **Verificación de copias de seguridad:** debe verificarse que exista una copia de toda la información almacenada. Los datos asociados a esta información son los once hechos con sus dimensiones respectivas identificados en los subsistemas de almacenamiento e integración del producto egf.
2. **Backup existente:** se cuenta con un backup que contiene toda la información relacionada con los datos del negocio. Se pueden realizar copias de la información en otros medios de almacenamiento como DVD y memoria flash.

1.10 Esquema de seguridad

Es de vital importancia implementar un mecanismo de protección contra aquellas acciones que puedan afectar la integridad, confidencialidad y disponibilidad de los datos almacenados en un sistema de información.

1.10.1 Seguridad en el subsistema de almacenamiento

Para lograr seguridad en el subsistema de almacenamiento es necesario definir los roles que realizan acciones sobre los datos en la BD. Ver tabla 4.

Roles	Permisos
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 de los ensayos clínicos para el producto egf.

Tabla 4: Seguridad en el subsistema de almacenamiento

1.10.2 Seguridad en el subsistema de integración

La seguridad en el subsistema de integración se garantiza a través del sistema operativo que en este caso sería GNU/Linux. Dicho sistema operativo es el encargado de asignar permisos de acceso a los archivos para algunos usuarios que necesiten realizar análisis sobre la información. De esta manera se puede restringir que los datos de las fuentes, las transformaciones y los trabajos no sean modificados, eliminados o ejecutados, al marcar las propiedades de lectura sobre la carpeta que contengan los ficheros que permiten el desarrollo de los procesos de ETL.

Conclusiones del capítulo

Después de analizar las etapas correspondientes al análisis y diseño de los subsistemas de almacenamiento e integración, se arribaron a las siguientes conclusiones:

- ✓ Luego de las entrevistas realizadas al cliente en conjunto con los protocolos y las fuentes de datos proporcionadas, fueron identificados 18 RI, 2 RF y 7 RNF del sistema.
- ✓ Fueron identificadas 48 RN, mediante las características presentadas en las fuentes de datos y sobre el trabajo que realiza la organización con el producto egf.
- ✓ Fueron identificadas 35 tablas de dimensiones y 11 tablas de hechos, diseñadas en el modelo de datos para garantizar el correcto funcionamiento del sistema.
- ✓ El perfilado de datos realizado a las fuentes permitió conocer el estado de la información y definir reglas aplicables en el proceso de transformación.
- ✓ Se definieron como subsistemas a implementar en el diseño de la arquitectura para la solución propuesta: integración y almacenamiento.

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

3.1 Introducción

Una vez realizado el diseño de los subsistemas de almacenamiento e integración del producto egf, se procede a la implementación de los subsistemas definidos. Estos dos procesos permiten integrar y almacenar la información en las tablas correspondientes diseñadas en el modelo de datos.

Se especifican los nomencladores a utilizar para reconocer los datos dentro de los subsistemas, utilizando un lenguaje estándar y aplicable a cada uno de ellos para lograr un entendimiento de los datos en el negocio. Se definen los esquemas correspondientes para el almacenamiento de la información. Además se describen las técnicas de carga de las dimensiones, tablas de hechos y trabajos correspondientes. Se detallan las pruebas a realizar para la validación de la solución mediante herramientas de pruebas.

3.2 Estrategias generales de integración

La integración se puede enfocar de formas diferentes dependiendo de la idea de "Integración" que se tenga en cuenta. Existen cuatro estrategias de integración descritas a continuación:

- ✓ **Replicación de datos:** es una técnica de integración basada en la creación y mantenimiento de múltiples copias de una misma base de datos. En la mayoría de las implementaciones de replicación, un servidor mantiene la copia primaria de la base de datos y servidores adicionales mantienen las copias esclavas de la misma.
- ✓ **Extracción, Transformación y Carga de Datos (ETL):** extrae la información de un sistema fuente, transforma esos datos para satisfacer los requisitos del negocio y carga el resultado en el sistema destino. Tanto la fuente como el destino son generalmente base de datos y archivos. Esta técnica se encarga de la integración de datos, no de aplicaciones, y obtiene los datos directamente de la base de datos.
- ✓ **Integración de Información Empresarial (EII):** es un mecanismo de transformación y acceso a datos transparentes y optimizados para suministrar una única interfaz a lo largo de los datos de las organizaciones. Este tipo de solución consiste en crear un intermediario que contenga los directorios de la base de datos y que a su vez sirva de canal de consulta y representación de la información recuperada. Teniendo en cuenta estos aspectos, para la integración de datos a Tiempo Real la técnica EII constituye una buena alternativa, sin embargo no es factible para la integración de aplicaciones.

- ✓ **Integración de Aplicaciones Empresariales (EAI):** este proceso tiene dentro de sus principales objetivos proporcionar acceso transparente a la amplia gama de aplicaciones que existen en una organización. Las características más importantes de esta tecnología es que se utiliza para la integración de Aplicaciones – a – Aplicaciones y proporciona un enfoque de integración orientado a proceso basado en mensajes XML. Es generalmente utilizada para el procesamiento de transacciones de negocio operacional en tiempo real. (31)

La estrategia de integración a utilizar en la solución para los procesos de integración fue ETL. Esta permite la extracción de los datos del sistema fuente con el fin de transformarlos y cargarlos hacia las tablas correspondientes en la base de datos con el propósito de contribuir a la toma de decisiones sobre la información que es almacenada en el área de los EC del CIM.

3.3 Implementación del subsistema de almacenamiento

La implementación del subsistema de almacenamiento incluye estándares de codificación de las estructuras, para facilitar la comprensión de los nombres definidos en cada uno de los esquemas.

3.3.1 Estándares de codificación

Con el objetivo de estandarizar las estructuras de los subsistemas de almacenamiento e integración del producto egf, se define un patrón que conduzca a la correcta normalización de los términos utilizados. Estos facilitan a los desarrolladores entender a través de un lenguaje estándar las estructuras empleadas.

Se propone mantener la misma nomenclatura atendiendo a la clasificación de las diferentes estructuras, teniendo en cuenta si una de ellas es una tabla de hecho o una dimensión. Si la tabla resulta ser una dimensión, al nombre de la misma le preceden las letras “dim” separadas del nombre de la dimensión por el carácter “_”, ejemplo “dim_sexo”. En caso de ser una tabla de hecho, como prefijo se ubican las letras “hech”, igualmente separadas del nombre de la tabla de hechos por el carácter “_”, ejemplo “hech_EC019”.

Para los atributos de las dimensiones se siguió la misma política. En el caso de las llaves primarias de las dimensiones se les denominó “dk_dim_dimension_id”. En caso de que el atributo sea el código del negocio se le especificó como “dimension_codigo”, igualmente para los nombres, descripciones u otros atributos: “dimension_nombre”, “dimension_descripcion” respectivamente. De manera general los atributos fueron nombrados como “dimension_atributo”. Las medidas fueron definidas de la siguiente forma “cant_medida”, por ejemplo “cant_pacientes”.

Con este proceso se ha logrado estandarizar la nomenclatura a utilizar para cada una de las tablas del negocio, atributos y medidas dentro de la base de datos. Luego se procede a la implementación de las estructuras físicas.

3.3.2 Implementación del modelo de datos físico

El modelo de datos físico constituye una colección integrada de entidades que describen las estructuras de los datos. Dicho modelo se genera a partir del modelo lógico dimensional; conteniendo las relaciones entre las tablas de hechos y dimensiones. La solución cuenta con 49 tablas, divididas en 18 tablas de dimensiones, 28 tablas de hechos y dimensiones y 3 tablas de metadatos. Se define la utilización de tres esquemas:

El esquema “dimensiones” contiene las dimensiones compartidas con los demás MD del CIM, del almacén central.

El esquema “mart_egf” contiene las tablas de dimensiones y hechos propias de los subsistemas de almacenamiento e integración del producto egf.

El esquema “metadatos” definido con el propósito de almacenar la información correspondiente a las ejecuciones de los trabajos y las transformaciones. Ver figura 13.

Figura 13: Esquemas

3.4 Implementación del subsistema de integración

La implementación del subsistema de integración se divide en tres etapas fundamentales Extracción, Transformación y Carga. Para llevar a cabo dichos procesos es necesario haber realizado un análisis a las fuentes para identificar sus principales problemas. Dicha información se encuentra en diversos ficheros de formato mdb que son generados a través del sistema de archivos EpiData.

Primeramente se extraen de los datos de las fuentes donde se encuentran contenidos, seleccionando los datos que aportan información significativa al negocio, teniendo en cuenta el modelo de datos realizado anteriormente. Luego son implementadas las fases de transformación y limpieza de los datos para cargarlos hacia la BD. La limpieza posibilita detectar datos con errores, entradas duplicadas de información y a través de las transformaciones se combinan y ordenan los datos.

Se describen a continuación algunos de los subsistemas identificados por Kimball que se utilizan en el desarrollo de la solución propuesta:

- ✓ **Perfilado de datos:** permitió analizar los datos para verificar su calidad y el cumplimiento de algunos estándares conforme a los requisitos especificados por el cliente. A través del uso de este subsistema fueron definidas nuevas reglas de transformación.
- ✓ **Subsistema de extracción:** posibilitó la extracción de los datos desde la fuente de origen para su transformación y posterior carga.
- ✓ **Subsistema de transformación:** permitió realizar transformaciones como el mapeo de valores, el cambio de tipo de dato en algunos campos, el filtrado de valores, entre otras.
- ✓ **Subsistema de carga:** permitió realizar la carga de los datos hacia las tablas de dimensiones y hechos de los subsistemas de almacenamiento e integración del producto egf.
- ✓ **Llave subrogada:** permite crear claves subrogadas independientes para cada tabla de dimensión.
- ✓ **Rastreo de eventos de errores:** permite capturar errores que proporcionan información valiosa sobre la calidad de los datos y posibilita mejorarlos.
- ✓ **Cambio lento de las dimensiones (SCD):** implementa la lógica para crear atributos de variabilidad lenta a lo largo del tiempo. Fue utilizado el tipo 1 de SCD para brindar tratamiento al cambio de la información asociada a las dimensiones.
- ✓ **Programador de trabajos:** gestionan los trabajos, se encargan de la ejecución de las transformaciones atendiendo un orden específico y a la periodicidad definida para la carga de la información.
- ✓ **Repositorio de metadatos:** captura los metadatos asociado a los procesos de ETL, de los datos del negocio. (32)

Los metadatos representan información descriptiva sobre los datos y otras estructuras, tales como objetos, reglas de negocio, y los procesos que manipulan los datos. Estos pueden ser agrupados en tres categorías:

- ✓ **Metadatos técnicos:** se aplican para describir el funcionamiento de un sistema o el modo en que se relacionan sus componentes. Enfocado a los diseñadores, desarrolladores y administradores durante el desarrollo y mantenimiento. Proporcionan un rastro detallado de las actividades y objetos del AD que los encargados de su mantenimiento pueden utilizar para construir nuevos objetos y mantener los existentes.
- ✓ **Metadatos de proceso:** representan las estadísticas sobre los resultados de la ejecución del propio proceso de ETL, incluyendo medidas tales como filas cargadas con éxito, las filas rechazadas y la cantidad de tiempo de carga, particularmente es importante en el proceso de limpieza de metadatos.
- ✓ **Metadatos de negocio:** permite obtener los datos y la información que describe el negocio, usualmente son los datos fuentes. Incluyen descripciones de datos que no están relacionadas a implementaciones de software, por ejemplo, el nombre del negocio, las reglas del negocio en relación a otros datos y el dueño de la definición. (8)

En la solución se decidió utilizar los metadatos técnicos y de procesos con el propósito de conocer la trazabilidad referente sobre la ejecución de las transformaciones de los hechos, dimensiones y trabajos.

3.4.1 Implementación de las transformaciones

El proceso de transformación se compone de pasos entrelazados entre sí a través de saltos. Los pasos representan el elemento más pequeño dentro de las transformaciones, y a través de los saltos fluye la información que es obtenida de los sistemas fuentes. Los saltos constituyen el elemento a través del cual fluye la información entre los diferentes pasos (siempre es la salida de un paso y la entrada de otro). Los pasos están agrupados por categorías y cada uno de ellos está diseñado para cumplir una función determinada. Cada paso tiene una ventana de configuración específica, donde se determinan los elementos a tratar y su forma de comportamiento. (33)

Para realizar la extracción de los datos provenientes de las fuentes a cada una de las tablas de hechos relacionados con la información de los EC del producto egf, se procede de la siguiente manera:

Se accede a los datos de la fuente, de donde son extraídos los campos relacionados con las dimensiones que se relacionan con las tablas de hechos, se verifica la existencia de los ficheros que contiene la información a extraer y se procede a realizar las transformaciones necesarias para cada caso. Al final son insertados los datos en la tabla correspondiente en la base de datos y se genera la información para obtener los metadatos en el esquema determinado. La figura 14 muestra la transformación realizada al hecho hech_forma_vacunal_EC041:

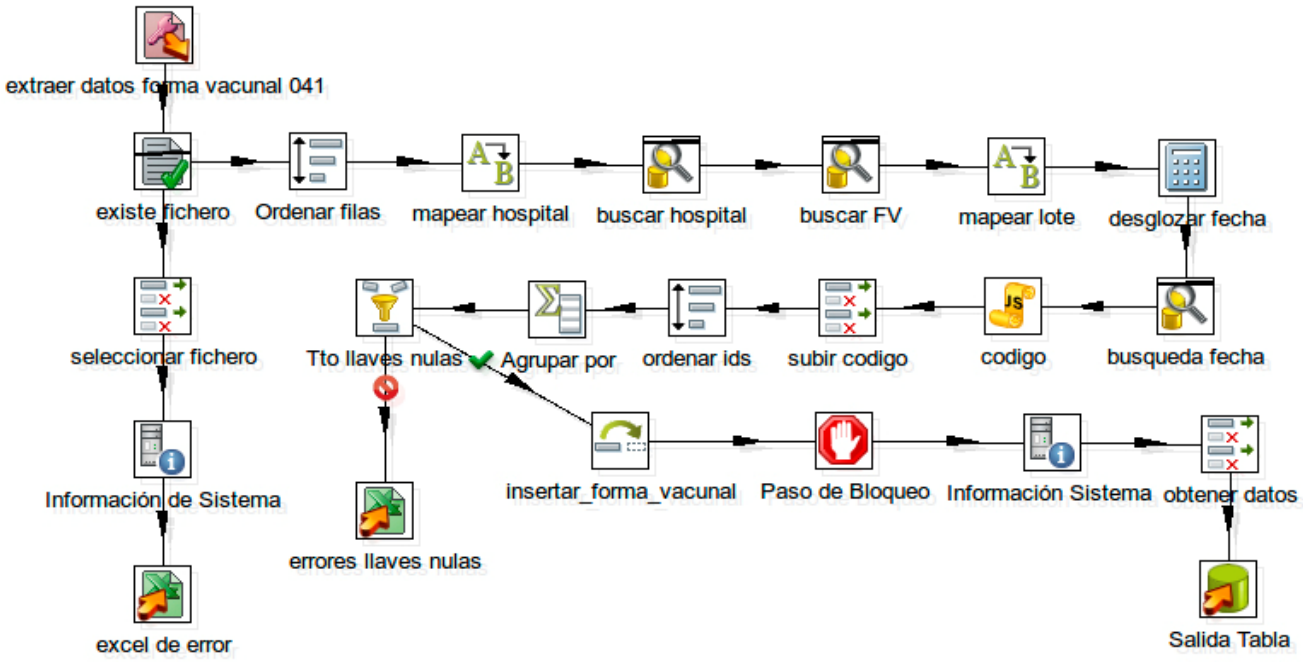


Figura 14: Transformación hech_forma_vacunal_EC041

Los datos almacenados en la tabla de error recibirán un tratamiento luego de entrevistarse con el cliente para consultar si la información tiene un significado relevante en el negocio. Luego se definen las reglas de negocio y transformación que serían aplicadas para su carga en la base de datos.

3.4.2 Implementación de los trabajos

Una vez finalizadas las transformaciones requeridas para la carga de los datos, se llevó a cabo la implementación de los trabajos. Los trabajos o los "Job" representan un conjunto de tareas con el objetivo de realizar una acción determinada. Permiten ejecutar una o varias transformaciones, siguiendo una secuencia lógica de pasos.

Los saltos o hops entre los componentes de un Job indican el orden de ejecución de cada uno de ellos (no empezando la ejecución del elemento siguiente hasta que el anterior no haya concluido). (33)

En la investigación fueron implementados 13 trabajos, uno destinado para la carga de las dimensiones comunes entre los hechos, otro por cada hecho con sus dimensiones propias y uno general para ejecutar los trabajos anteriores. En la figura 15 se muestra el trabajo general.

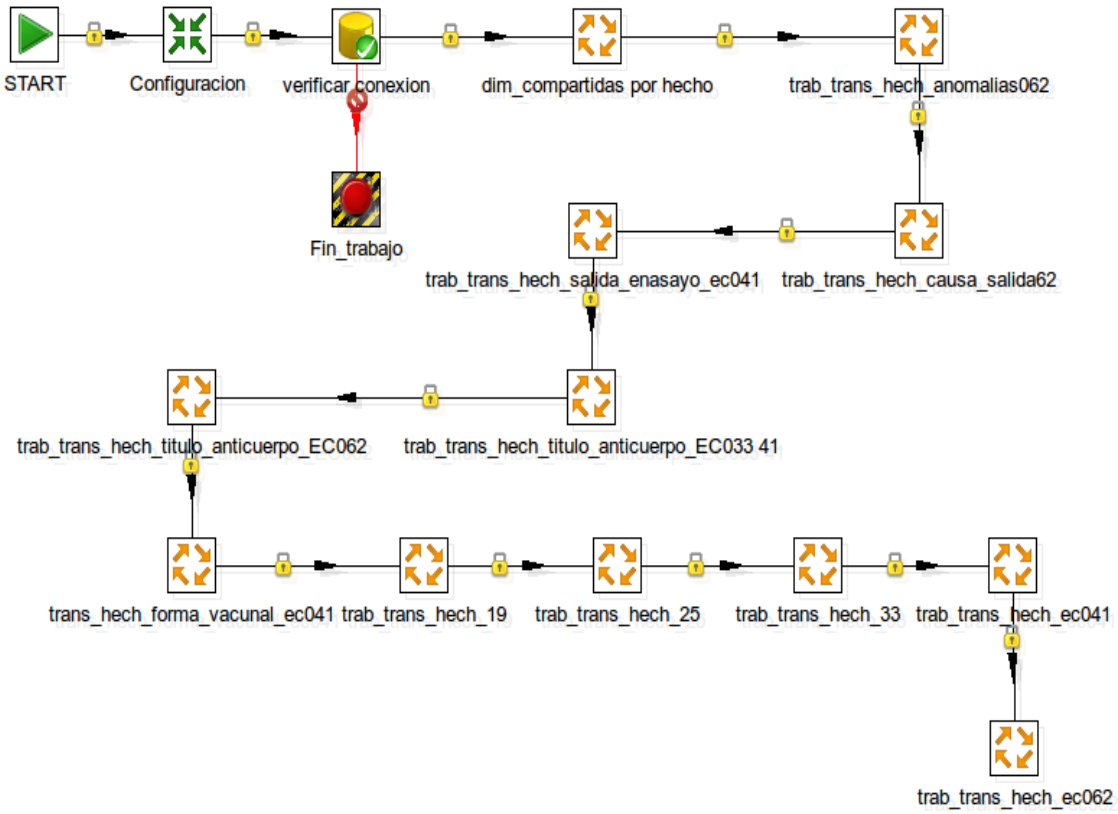


Figura 15: Trabajo general

3.4.3 Gestión del cambio lento en las dimensiones

Las dimensiones lentamente cambiantes o SCD (Slowly Changing Dimensions) son dimensiones en las cuales sus datos tienden a modificarse a través del tiempo, ya sea de forma ocasional o constante, o implique a un solo registro o la tabla completa. Cuando ocurren estos cambios, se puede optar por seguir alguna de estas dos grandes opciones: registrar el historial de cambios o reemplazar los valores que sean necesarios.

Inicialmente Ralph Kimball planteó tres estrategias a seguir cuando se tratan de las SCD: tipo 1, tipo 2 y tipo 3. A través de los años, los estudios realizados han logrado profundizar más sobre estas definiciones, permitiendo incluir varios tipos de SCD conocidos como los: tipo 0, tipo 4 y tipo 6. (34)

- ✓ **Tipo 0 (no tiene en cuenta la gestión histórica):** no se realiza ningún esfuerzo para lidiar con los problemas del cambio de la dimensión. Nunca se cambia la información ni se sobrescribe.

- ✓ **Tipo 1 (sobrescribir):** sobrescribe la información antigua por la nueva y es utilizado generalmente cuando se necesita corregir algún error de datos en las dimensiones.
- ✓ **Tipo 2 (añadir fila):** esta estrategia requiere que se agreguen algunas filas adicionales a la tabla de dimensión para que almacenen el historial de cambios. Cuando ocurra algún cambio en los valores de los registros, se añadirá una nueva fila y se deberá completar los datos referidos al historial de cambios.
- ✓ **Tipo 3 (añadir columna):** esta estrategia requiere que se agregue a la tabla de dimensión una columna adicional por cada columna de cuyos valores se desean mantener un historial de cambios.
- ✓ **Tipo 4 (tabla de historia separada):** esta técnica se utiliza en combinación con alguna otra y su función básica es almacenar en una tabla adicional los detalles de cambios históricos realizados en una tabla de dimensión. Dicha tabla histórica indicará qué tipo de operación se ha realizado (Insert, Update, Delete), sobre qué campo y en qué fecha. El objetivo de mantener esta tabla es para poder contar con un detalle de todos los cambios, analizarlos y poder tomar decisiones acerca de cuál técnica SCD podría aplicarse mejor.
- ✓ **Tipo 6 (híbrido):** este tipo combina los tipos 1, 2 y 3. Se denomina SCD Tipo "6", porque integra la suma de los tres tipos (6 = 1 + 2 +3). A parte de integrar estas tres estrategias, también añade una pareja adicional de columnas para indicar el rango de fechas al cual aplica cada fila en particular.

En los subsistemas de almacenamiento e integración del producto egf la estrategia de SCD empleada es el tipo 1. Se implementó este tipo de SCD aunque la carga sea histórica. Dicha estrategia permite después que son cargados los datos corregir errores que sean identificados en los procesos de transformación en caso de que fuera necesario. Por la importancia que presenta la información referente a los EC del producto egf es necesario que los errores que se detecten sean corregidos con el cliente, el mismo conoce la magnitud y contenido de los datos para facilitar la búsqueda de soluciones frente a cualquier tipo de problema.

3.5 Pruebas aplicadas a los subsistemas del producto egf

Resulta de gran importancia obtener un producto con calidad, para esto se requiere de la utilización de metodologías y procedimientos estándares para el desarrollo de los requisitos, el análisis, el diseño, la implementación y, finalmente, las pruebas del software, que son el elemento fundamental para el logro de la calidad de cualquier sistema. El único instrumento adecuado para determinar el estado de calidad de un producto de software, es el proceso de pruebas. Estas se realizan con el objetivo de medir el grado de

cumplimiento de los requisitos implementados. Básicamente es una fase en el desarrollo que consiste en probar las aplicaciones construidas.

Las pruebas desempeñan un papel fundamental garantizando que el producto salga al mercado con la calidad requerida, proporcionan elevar la reputación del equipo de desarrollo y principalmente la completa satisfacción del cliente. Las pruebas realizadas para validar el funcionamiento del sistema son las que se describen a continuación.

3.5.1 Pruebas unitarias

Las pruebas unitarias permiten probar el correcto funcionamiento de un componente o subsistema específico. Estas pruebas son desarrolladas por los propios desarrolladores durante la implementación de la solución. (20)

Luego de concluida la etapa de implementación fueron aplicadas las pruebas unitarias a los subsistemas de almacenamiento e integración del producto egf. Se detectaron 10 no conformidades (NC) que fueron resueltas rápidamente.

En el subsistema de almacenamiento se identificaron dos **NC** con complejidad Alta.

NC1. No se corresponden los casos de uso con la representación del modelo.

NC2. Las jerarquías y niveles de las dimensiones no se encuentran bien definidas.

En el subsistema de integración se identificaron ocho **NC** con complejidad Media:

NC1. No se especificaron correctamente los nombres de los ficheros a cargar en los diseños del proceso de integración de datos.

NC2. Las estrategias de gestión de metadatos tanto en el diseño, como en la implementación no se encuentran definidas.

NC3. Los estereotipos del diagrama de flujo para representar el diseño de los procesos de integración de datos (decisiones y sincronización) deben aplicarse correctamente.

NC4. Los nombres de los repositorios de carpetas en el diseño no se encuentran bien definidos.

NC5. Las relaciones entre las actividades y los objetos en el diseño de los procesos de integración de datos no están bien representadas.

NC6. Los nombres de las actividades en los diseños del proceso de integración de datos tienen que reflejar su verdadero objetivo.

NC7. No se optimizó la implementación de las transformaciones.

NC8. No se generalizó el tratamiento de llaves nulas y llaves huérfanas.

3.5.2 Pruebas de integración

Las pruebas de integración permiten verificar la correcta integración de los componentes y subsistemas que conforman la solución. Pone a prueba la vista arquitectónica del sistema definida en una infraestructura de desarrollo. Estas pruebas son ejecutadas por los arquitectos de software. (20)

Para validar la solución a través de los casos de pruebas de integración se pudo comprobar que los EC del producto egf fueron cargados satisfactoriamente. Las consultas realizadas a la base de datos en cuanto a sus resultados coinciden todos con los datos de los EC.

3.6 Herramientas de pruebas para validar los subsistemas del producto egf

3.6.1 Casos de prueba

Los casos de pruebas permiten verificar la calidad de un producto de software. Estos son utilizados para identificar posibles fallos en la implementación y comprobar el funcionamiento correcto de los requisitos implementados. Para los subsistemas de almacenamiento e integración del producto egf fueron implementados 11 casos de prueba de integración. Estos se diseñaron para que fueran un caso de prueba por RI por cada CUI.

La estrategia implementada para validar la solución se realizará a través de consultas ejecutadas a la base de datos que contiene la información almacenada. Permiten comprobar mediante los RI recogidos y agrupado por CUI, si los datos provenientes de la fuente se corresponden con los resultados arrojados por las consultas. Estos dos procesos tienen que ser compatibles para demostrar si fue satisfactoria la inserción de los datos. Esta consulta fue realizada para el RI: Obtener la cantidad de pacientes del ensayo clínico 019 por salida del ensayo, causa fallecimiento, reacciones adversas, signos vitales, examen de laboratorio, respuesta inmunológica, y respuesta clínica.

A continuación se muestra un ejemplo realizado y los resultados arrojados.

```
SELECT
COUNT (DISTINCT dk_codigo_paciente)
FROM
mart_egf.hech_ec019,
mart_egf.dim_causa_fallecimiento,
mart_egf.dim_ra,
mart_egf.dim_salida_ensayo,
mart_egf.dim_evaluacion_sv,
dimensiones.dim_examen_laboratorio,
```

```
mart_egf.dim_respuesta_inmunologica,
dimensiones.dim_respuesta_clinica
WHERE
hech_ec019.dk_dim_ra_id = dim_ra.dk_dim_ra_id AND
hech_ec019.dk_dim_fallecimiento_id = dim_causa_fallecimiento.dk_dim_causa_fallecimiento_id
AND
hech_ec019.dk_dim_salida_ensayo_id = dim_salida_ensayo.dk_dim_salida_ensayo_id AND
hech_ec019.dk_dim_evaluacion_sv_id = dim_evaluacion_sv.dk_dim_evaluacion_sv_id AND
hech_ec019.dk_dim_examen_laboratorio_id =
dim_examen_laboratorio.dk_dim_examen_laboratorio_id AND
hech_ec019.dk_dim_respuesta_inmunologica_id =
dim_respuesta_inmunologica.dk_dim_respuesta_inmunologica_id AND
hech_ec019.dk_dim_respuesta_clinica_id = dim_respuesta_clinica.dk_dim_respuesta_clinica_id
AND
dim_causa_fallecimiento.causa_nombre = 'Metastasis' AND
dim_ra.ra_clasificacion = 'Local' AND
dim_evaluacion_sv.evaluacion_sv_tipo = 'TAS' AND
dim_evaluacion_sv.evaluacion_sv_rango_descripcion = 'Normal';
```

Los datos obtenidos de la ejecución de la consulta fueron: 2 pacientes que fallecieron en el ensayo por metástasis, la reacción adversa presentada fue Local, el tipo de evaluación para los signos vitales se corresponde con la tensión arterial sistólica (TAS) dentro del rango normal, estos coinciden con los encontrados en el archivo de formato mdb; se puede concluir que el resultado del caso de prueba de integración se realizó satisfactoriamente. El requisito de información no arrojó ningún error al ser verificado.

Para los demás CUI se procedió de la misma manera, se les aplicó estas pruebas a un RI por cada uno de ellos. (Consultar el artefacto: “DATEC_CIM_EGF_Casos_de_prueba_de_integración”).

Se definieron también 36 casos de pruebas en correspondencia con las reglas de transformación definidas en el negocio. Se persigue con estas pruebas comprobar si las reglas aplicadas a las variables definidas en el negocio toman los valores que le fueron asignados. En este proceso se realiza el Caso de Prueba para la regla: El sexo estará definido por los siguientes valores: cuando sea Femenino: se mostrará un 1,

cuando sea Masculino se mostrará un 2 y en caso de que entre un valor vacío se mostrará: ND. (Consultar el artefacto: “DATEC_CIM_EGF_24_05_2013_Casos de prueba”). Ver figura 16.

Caso de prueba						
Nombre variable	sexo					
Escenario	hech_ec 019, hech_ec025, hech_ec033, hech_ec041, hech_ec062.					
Regla de	Los valores del sexo se sustituyeron por: si es 1 Femenino, si es 2 Masculino y si es 3 ND.					
Valor de entrada	**Estado del dato**	**Resultado esperado**	**Respuesta del Sistema**	**Resultado Real**	**Comentario**	**Resultado de la Prueba**
F1	No Válido	Femenino	El sistema transforma el dato aplicando la regla de transformación.	Femenino	La prueba se realizó sin ocurrir errores	Satisfactorio
2	Válido	Masculino	El sistema agrega el dato satisfactoriamente.	Masculino	La prueba se realizó sin ocurrir errores	Satisfactorio

Figura 16: Caso de prueba RN

3.6.2 Listas de chequeo

Las listas de chequeos están conformadas por una serie de preguntas, por la cual se verifica el grado de cumplimiento de las reglas establecidas. Se utilizará con el fin de medir una serie de indicadores implicados en la creación de la capa de integración, además de medir la calidad de los artefactos y documentos generados durante la realización del producto.

Los indicadores a evaluar se encuentran distribuidos en tres secciones fundamentales:

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

Luego de ser aplicadas las listas de chequeo a cada uno de los artefactos de ETL: “Registro del sistema fuente”, “Perfilado de datos”, “Diccionario de datos” y “Mapa lógico de datos”, fueron arrojados los siguientes resultados mostrados en forma de gráfica. En la figura 17 se observa el comportamiento de 47 indicadores, de los cuales 20 son críticos y fueron encontradas 4 NC. (Consultar Expediente de Proyecto).

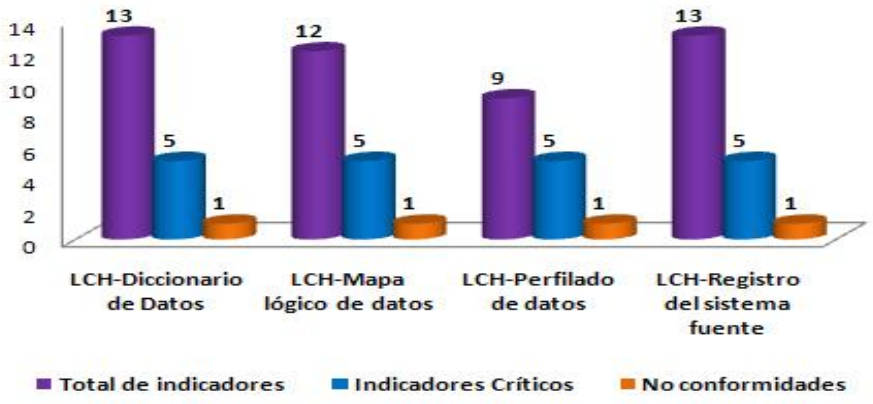


Figura 17: Comportamiento de indicadores en las listas de chequeos

3.7 Calidad de datos

El concepto de "calidad de datos" se asocia con mucha frecuencia a los sistemas de información, generando grandes volúmenes de datos y se corre el riesgo de encontrar datos incompletos e inconsistentes. En el marco de MD, estos problemas son identificados cuando se realiza la integración de distintas fuentes de información. Se describen a continuación distintos procesos que son realizados para comprobar la calidad de la información almacenada en los subsistemas de almacenamiento e integración del producto egf. (35)

3.7.1 Perfilado de los datos

El perfilado de datos es el proceso que permite entender la estructura, contenido y estado de los datos. Fue utilizado el PgAdmin III en su versión 1.14.0 en el perfilado realizado a la base de datos relacional que almacena la información de los EC del producto egf. Los resultados obtenidos durante este proceso indican que la carga de los datos a la base de datos a través de las tablas de dimensiones y hechos se realizó correctamente. Fueron analizadas 1060889 tuplas, las cuales no contienen valores nulos, ni vacíos. Ver figura 18.

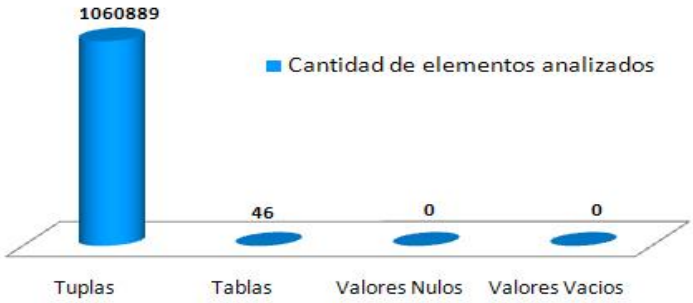


Figura 18: Resultado del perfilado de datos

3.7.2 Auditoría de datos

La realización de auditoría de los datos permite conocer el estado en que se encuentran los datos almacenados en el subsistema de almacenamiento, así como la información referente a los procesos de ejecución de las transformaciones en cuanto a: el nombre de la transformación, la dirección IP donde fue ejecutada, la fecha inicial y final de ejecución, cantidad de datos leídos, número de errores y otros elementos. La estrategia definida para auditar los datos almacenados consiste en la implementación de tablas de metadatos para guardar la información de los procesos mencionados anteriormente. Estas tablas proporcionan las trazas de todos los pasos que siguen las transformaciones antes y después de ser insertadas en la base de datos. Además sirven para comprobar si se realizó con éxito la carga de los datos hacia el destino, garantizando la confiabilidad de los procesos realizados.

Conclusiones del Capítulo

En el capítulo se abordó sobre la implementación y validación de los subsistemas de almacenamiento e integración del producto egf, arribándose a las siguientes conclusiones:

- ✓ Se definió utilizar como estrategia de integración ETL, permitiendo la extracción de los datos del sistema fuente con el fin de transformarlos y cargarlos a los subsistemas de almacenamiento e integración del producto egf.
- ✓ Fueron implementados los dos subsistemas que componen la solución: almacenamiento e integración, teniendo como resultado la disponibilidad de la información para ayudar al proceso de toma de decisiones de la organización.
- ✓ Los estándares de codificación definidos permitieron estandarizar la nomenclatura a utilizar para cada una de las tablas del negocio, atributos y medidas dentro de la base de datos.
- ✓ Se definieron tres esquemas a utilizar: para las dimensiones compartidas con los demás mercados del CIM del almacén central, para las tablas de dimensiones y hechos propias de los subsistemas de almacenamiento e integración del producto egf y otro para los metadatos con el objetivo de almacenar la información de las ejecuciones pertenecientes a los trabajos y las transformaciones.
- ✓ Los metadatos utilizados fueron: técnicos y de procesos, para obtener información referente sobre los procesos de las transformaciones y los trabajos.
- ✓ Las pruebas utilizadas en la validación de la solución fueron guiadas por la metodología utilizada: pruebas unitarias y de integración.

Conclusiones Generales

El desarrollo de la solución cumple con los objetivos planteados en el inicio de la investigación logrando los siguientes resultados:

- ✓ Se obtuvo un MD poblado con toda la información integrada de los ensayos clínicos del producto egf.
- ✓ Se hizo uso de la metodología definida por el departamento de Almacenes de Datos: Propuesta de Metodología para el Desarrollo de Almacenes de Datos en DATEC, para guiar el proceso de desarrollo de los subsistemas de almacenamiento e integración del producto egf.
- ✓ Las herramientas seleccionadas y aplicadas durante todas las fases de desarrollo, contribuyeron al diseño, integración y almacenamiento de los subsistemas implementados.
- ✓ Las pruebas realizadas permitieron validar el funcionamiento de los procesos implementados a partir de los requisitos establecidos por el cliente.

Recomendaciones

Para el presente trabajo de diploma con el propósito de mejorar la solución se recomienda:

- ✓ Aplicar a los subsistemas de almacenamiento e integración del producto egf técnicas de minería de datos para una mejor interpretación de la información y un rápido procesamiento de la misma.

Referencias Bibliográficas

1. Universidad de las Ciencias Informáticas. [En línea] [Citado el: 13 de septiembre de 2012.] http://www.uci.cu.

2. DATEC. [En línea] [Citado el: 15 de septiembre de 2012.] http://gespro.datec.prod.uci.cu.

3. Revista de Nefrología. [En línea] [Citado el: 11 de noviembre de 2012.] http://www.revistanefrologia.com/revistas/P1-E246/P1-E246-S136-A3328.pdf.

4. Periódico Juventud Rebelde. [En línea] [Citado el: 11 de noviembre de 2012.] http://www.juventudrebelde.cu/ciencia-tecnica/2012-05-07/vacuna-cubana-contra-cancer-de-pulmon-en-fase-de-ensayos-clinicos-2/.

5. Revista. [En línea] [Citado el: 10 de noviembre de 2012.] http://bvs.sld.cu/revistas/spu/vol38_5_12/sup11512.htm.

6. Ciencia de Cuba. [En línea] [Citado el: 10 de noviembre de 2012.] http://cienciadecuba.wordpress.com/2011/12/06/prueban-vacuna-contra-cancer-de-pulmon-en-cuba/#more-1587.

7. Inmon, William H. *Using the DataWarehouse.*

8. Kimball, Ralph y Ross, Margy. *The Data Warehouse Toolkit.*

9. Inmon, Bill. *Building the Data Warehouse.*

10. Torres Torrillas, Francisco José Lucas y otros. [En línea] [Citado el: 12 de noviembre de 2012.] http://alarcos.inf-cr.uclm.es/doc/bbddavanzadas/08-09/FUNCIONALIDAD%204.pdf..

11. Chuc-Durán, Diana Graciela. Introducción a los Datawarehouses. [En línea] [Citado el: 13 de noviembre de 2012.] http://www.publicaciones.ujat.mx/publicaciones/revista_dacb/Acervo/v6n1OL/v6n1a5-ol/index.html..

12. ONEI. [En línea] [Citado el: 15 de noviembre de 2012.] http://www.one.gob.do/index.php?module=articles&func=display&aid=1377...

13. Redalyc. [En línea] [Citado el: 16 de noviembre de 2012.] http://redalyc.uaemex.mx/src/inicio/ArtPdfRed.jsp?iCve=193915954004..

14. Kimball, Ralph y Ross, Margy. *The Data Warehouse Toolkit: the Complete Guide to Dimensional Modelling. Second Edition.*

15. DATA WAREHOUSE PARA LA GESTIÓN DE LISTA DE ESPERA SANITARIA. [En línea] [Citado el: 16 de noviembre de 2012.] http://oa.upm.es/1095/1/PFC_ITZIAR_ANGOITIA_ESPINOSA.pdf..

16. *Modelo Multidimensional.*

17. Zepeda, Leopoldo Zenaido. *Metodología para el Diseño Conceptual de Almacenes de Datos.* Valencia : s.n.

18. Informática Aplicada a la Gestión Pública. Facultad Derecho UMU. [En línea] 13 de octubre de 2011. [Citado el: 12 de noviembre de 2012.] http://www.um.es/docencia/barzana/IAGP/IAGP2-Metodologias-de-desarrollo.html..

19. *ANÁLISIS Y DISEÑO DE UN DATA MART PARA EL SEGUIMIENTO ACADÉMICO DE ALUMNOS EN UN ENTORNO UNIVERSITARIO.* Rodríguez Sanz, Miguel. Madrid : s.n., 2010.

20. González Hernández, Yanisbel, Límia Navarro, Alberto y otros. Propuesta de Metodología para el Desarrollo de Almacenes de Datos en DATEC. La Habana : s.n., 2012.

21. EcuRed. [En línea] [Citado el: 21 de noviembre de 2012.] http://www.ecured.cu/index.php/CASE.

22. EcuRed. [En línea] [Citado el: 22 de noviembre de 2012.] http://www.ecured.cu/index.php/Visual_Paradigm.

23. EcuRed. [En línea] [Citado el: 22 de noviembre de 2012.] http://www.ecured.cu/index.php/Sistema_Gestor_de_Base_de_Datos.

24. PostgreSql. [En línea] [Citado el: 15 de noviembre de 2012.] http://www.postgresql.org.es/sobre_postgresql.

25. PostgreSql. [En línea] [Citado el: 16 de noviembre de 2012.] http://www.postgresql.org.pe/articles/introduccion_a_postgresql.pdf.

26. Guia-ubuntu. [En línea] [Citado el: 12 de noviembre de 2012.] http://www.guia-ubuntu.com/index.php?title=PgAdmin_III..

27. 0x27. [En línea] [Citado el: 12 de noviembre de 2012.] http://0x27.com.ar/2011/05/limpiar-datos-con-datacleaner-2/.

28. EcuRed. [En línea] [Citado el: 12 de noviembre de 2012.] http://www.ecured.cu/index.php/Pentaho_Data_Integration.

29. Gravitar. [En línea] [Citado el: 12 de noviembre de 2012.] http://www.gravitar.biz/index.php/herramientas-bi/pentaho/caracteristicas-pentaho/.

30. Almacenes de datos. [En línea] [Citado el: 20 de noviembre de 2012.] http://informatica.uv.es/iiguia/DBD/Teoria/data-warehouses.pdf.

31. Azán Basallo, Yasser, Díaz Estrada, Anay y González Gómez, Salvador. Una experiencia en integración de aplicaciones empresariales. La Habana : s.n., 2010.

32. bi.social. [En línea] [Citado el: 29 de abril de 2013.] http://bi.social.uoc.edu/smc/blog/34-subsistemas-etl-de-kimball..

33. Churriwifi. [En línea] [Citado el: 25 de abril de 2013.] http://churriwifi.wordpress.com/2010/05/10/16-3-construccion-procesos-etl-utilizando-kettle-pentaho-data-integration/..

34. Introducción al Business Intelligence. [En línea] [Citado el: 23 de abril de 2013.] http://books.google.com.cu.

35. Medina Mustelier, Doris. Técnicas de Extracción, Transformación y Carga de Datos del Sistema de Información Nacional de Seguridad Ciudadana en la Sistema de Información Nacional de Seguridad Ciudadana en la. Cuba. La habana : s.n.

Bibliografía

1. 0x27. [En línea] [Citado el: 12 de noviembre de 2012.] http://0x27.com.ar/2011/05/limpiar-datos-con-datacleaner-2/.

2. Almacenes de datos. [En línea] [Citado el: 20 de noviembre de 2012.] http://informatica.uv.es/iiguia/DBD/Teoria/data-warehouses.pdf.

3. bi.social. [En línea] [Citado el: 29 de abril de 2013.] http://bi.social.uoc.edu/smc/blog/34-subsistemas-etl-de-kimball.

4. Churriwifi. [En línea] [Citado el: 25 de abril de 2013.] http://churriwifi.wordpress.com/2010/05/10/16-3-construccion-procesos-etl-utilizando-kettle-pentaho-data-integration/..

5. Ciencia de Cuba. [En línea] [Citado el: 10 de noviembre de 2012.] http://cienciadecuba.wordpress.com/2011/12/06/prueban-vacuna-contra-cancer-de-pulmon-en-cuba/#more-1587.

6. DATA WAREHOUSE PARA LA GESTIÓN DE LISTA DE ESPERA SANITARIA. [En línea] [Citado el: 16 de noviembre de 2012.] http://oa.upm.es/1095/1/PFC_ITZIAR_ANGOITIA_ESPINOSA.pdf.

7. DATEC. [En línea] [Citado el: 15 de septiembre de 2012.] http://gespro.datec.prod.uci.cu.

8. EcuRed. [En línea] [Citado el: 21 de noviembre de 2012.] http://www.ecured.cu/index.php/CASE.

9. EcuRed. [En línea] [Citado el: 22 de noviembre de 2012.] http://www.ecured.cu/index.php/Visual_Paradigm.

10. EcuRed. [En línea] [Citado el: 22 de noviembre de 2012.] http://www.ecured.cu/index.php/Sistema_Gestor_de_Base_de_Datos.

11. EcuRed. [En línea] [Citado el: 12 de noviembre de 2012.] http://www.ecured.cu/index.php/Pentaho_Data_Integration.

12. Gravitar. [En línea] [Citado el: 12 de noviembre de 2012.] http://www.gravitar.biz/index.php/herramientas-bi/pentaho/caracteristicas-pentaho/.

13. Guia-ubuntu. [En línea] [Citado el: 12 de noviembre de 2012.] http://www.guia-ubuntu.com/index.php?title=PgAdmin_III..

14. Informática Aplicada a la Gestión Pública. Facultad Derecho UMU. [En línea] 13 de octubre de 2011. [Citado el: 12 de noviembre de 2012.] http://www.um.es/docencia/barzana/IAGP/IAGP2-Metodologias-de-desarrollo.html.

15. Introducción al Business Intelligence. [En línea] [Citado el: 23 de abril de 2013.] http://books.google.com.cu.

16. *Modelo Multidimensional.*

17. ONEI. [En línea] [Citado el: 15 de noviembre de 2012.] http://www.one.gob.do/index.php?module=articles&func=display&aid=1377...

18. Periódico Juventud Rebelde. [En línea] [Citado el: 11 de noviembre de 2012.] http://www.juventudrebelde.cu/ciencia-tecnica/2012-05-07/vacuna-cubana-contra-cancer-de-pulmon-en-fase-de-ensayos-clinicos-2/.

19. PostgreSql. [En línea] [Citado el: 15 de noviembre de 2012.] http://www.postgresql.org.es/sobre_postgresql.

20. PostgreSql. [En línea] [Citado el: 16 de noviembre de 2012.] http://www.postgresql.org.pe/articles/introduccion_a_postgresql.pdf.

21. Redalyc. [En línea] [Citado el: 16 de noviembre de 2012.] http://redalyc.uaemex.mx/src/inicio/ArtPdfRed.jsp?iCve=193915954004..

22. Revista. [En línea] [Citado el: 10 de noviembre de 2012.] http://bvs.sld.cu/revistas/spu/vol38_5_12/sup11512.htm.

23. Revista de Nefrología. [En línea] [Citado el: 11 de noviembre de 2012.] http://www.revistanefrologia.com/revistas/P1-E246/P1-E246-S136-A3328.pdf.

24. Universidad de las Ciencias Informáticas. [En línea] [Citado el: 13 de septiembre de 2012.] http://www.uci.cu.

25. Azán Basallo, Yasser, Díaz Estrada, Anay y González Gómez, Salvador. Una experiencia en integración de aplicaciones empresariales. La Habana : s.n., 2010.

26. Chuc-Durán, Diana Graciela. Introducción a los Datawarehouses. [En línea] [Citado el: 13 de noviembre de 2012.] http://www.publicaciones.ujat.mx/publicaciones/revista_dacb/Acervo/v6n1OL/v6n1a5-ol/index.html.

27. González Hernández, Yanisbel, Límia Navarro, Alberto y otros. Propuesta de Metodología para el Desarrollo de Almacenes de Datos en DATEC. La Habana : s.n., 2012.

28. Inmon, Bill. *Building the Data Warehouse.*

29. Inmon, William H. *Using the DataWarehouse.*

30. Kimball, Ralph y Ross, Margy. *The Data Warehouse Toolkit.*

31. —. *The Data Warehouse Toolkit: the Complete Guide to Dimensional Modelling. Second Edition.*

32. Medina Mustelier, Doris. Técnicas de Extracción, Transformación y Carga de Datos del Sistema de Información Nacional de Seguridad Ciudadana en la Sistema de Información Nacional de Seguridad Ciudadana en la. Cuba. La habana : s.n.

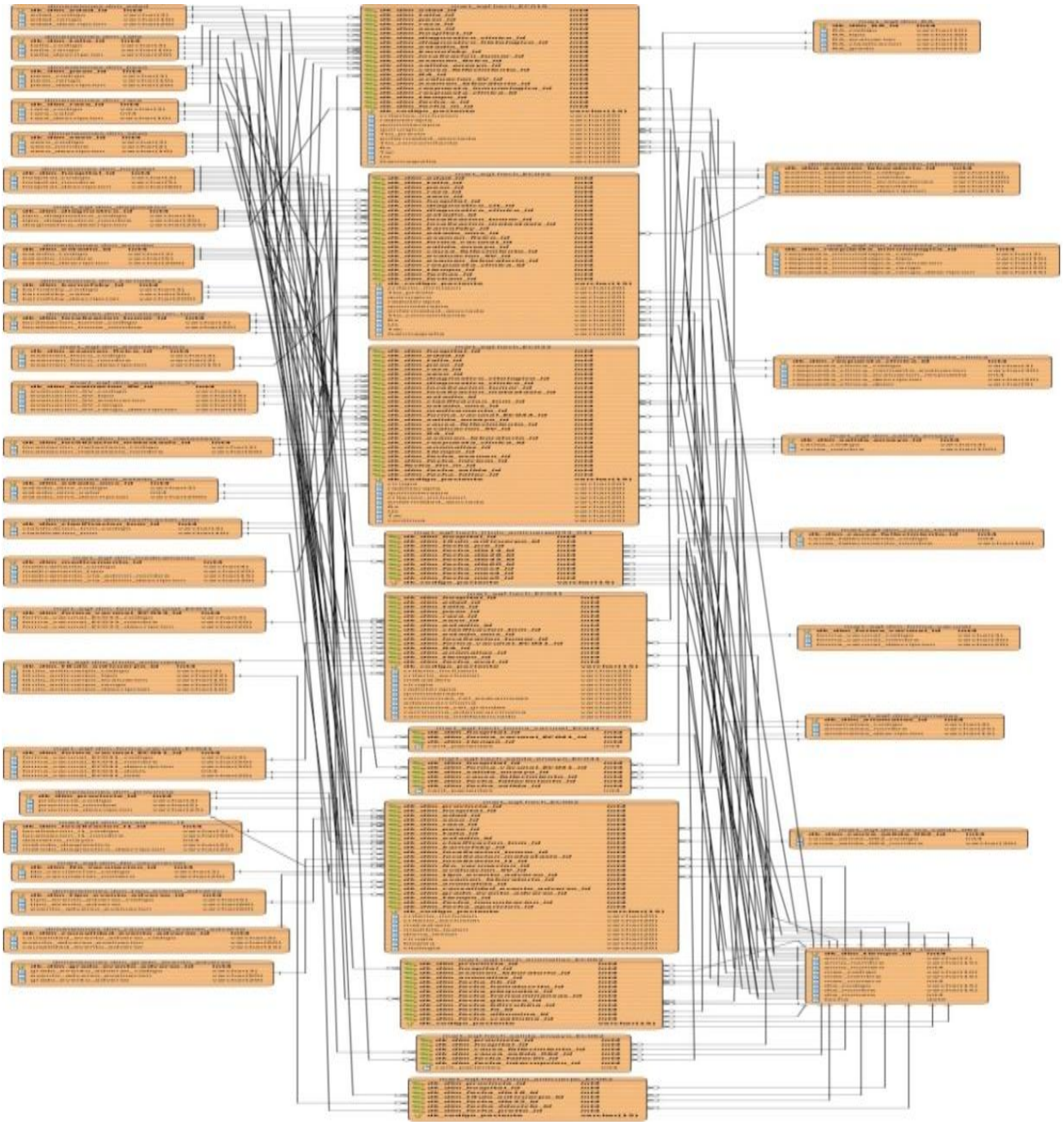
33. *ANÁLISIS Y DISEÑO DE UN DATA MART PARA EL SEGUIMIENTO ACADÉMICO DE ALUMNOS EN UN ENTORNO UNIVERSITARIO.* Rodríguez Sanz, Miguel. Madrid : s.n., 2010.

34. Torres Torrillas, Francisco José Lucas y otros. [En línea] [Citado el: 12 de noviembre de 2012.] http://alarcos.inf-cr.uclm.es/doc/bbddavanzadas/08-09/FUNCIONALIDAD%204.pdf.

35. Zepeda, Leopoldo Zenaido. *Metodología para el Diseño Conceptual de Almacenes de Datos.* Valencia : s.n.

Anexos

Anexo 1: Modelo de Datos



Glosario de Términos

Almacén de datos: repositorio lógico de datos que permite el acceso y la manipulación flexible de grandes volúmenes de información procedente de diferentes fuentes.

CU: caso de uso.

Dimensión: perspectiva mediante la cual se puede llevar a cabo el análisis sobre el hecho.

ETL (Extracción, Transformación y Carga): extrae los datos desde los sistemas fuentes, los transforma y luego los carga al almacén o mercado de datos.

Hecho: representa la ocurrencia de un proceso específico en el interior de la organización.

Jerarquía: organiza los niveles dentro de una dimensión, donde cada uno de ellos representa el total agregado de los datos del nivel inferior.

Medidas: constituye un valor o indicador de análisis del hecho.

Mercado de datos: implementación de un almacén con alcance limitado a un departamento.

RN: regla de negocio.

SGBD (Sistema Gestor de Bases de Datos): colección de programas cuyo principal objetivo es servir de interfaz entre la base de datos, el usuario y las aplicaciones.

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