Universidad de las Ciencias Informáticas

Facultad 6

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

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

Autores:

Dayatni Batista Rondón

Yaima Carmona Acosta

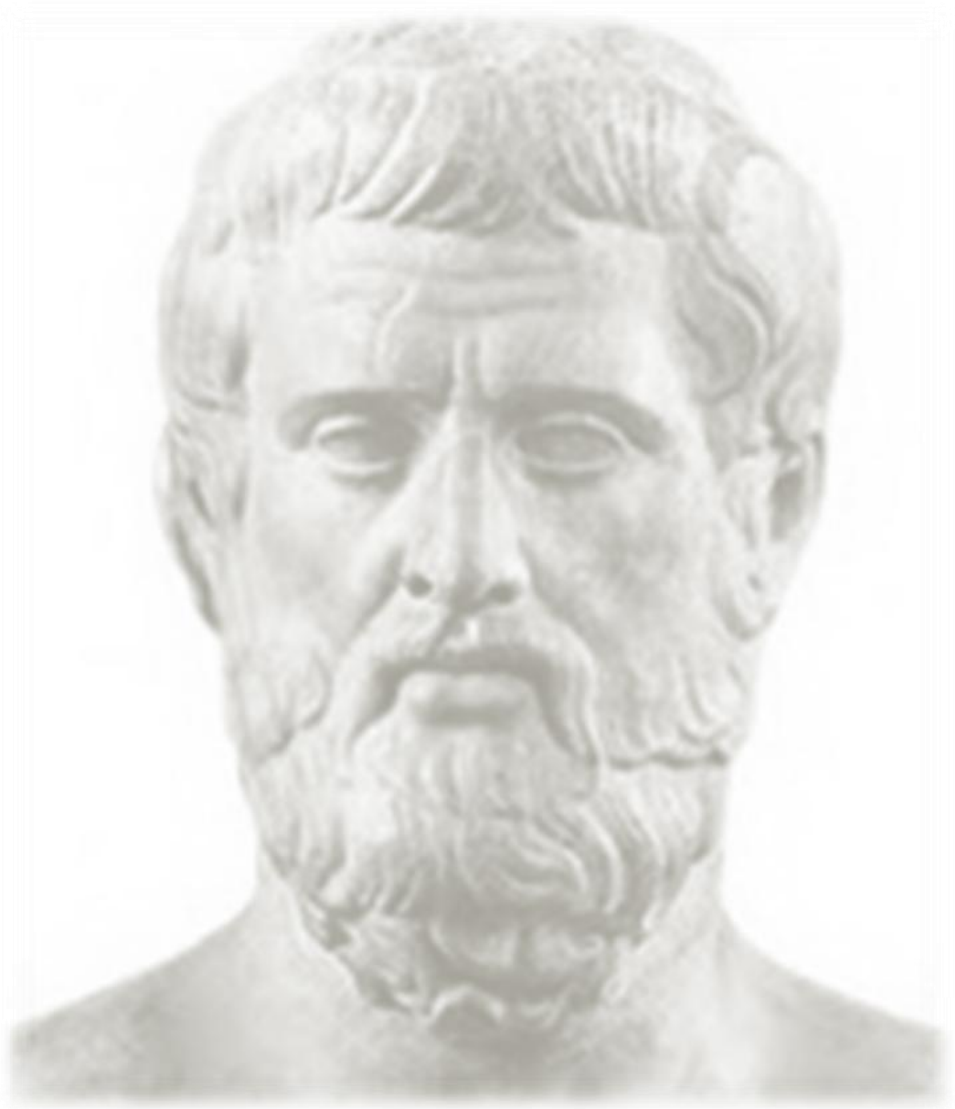
Tutores:

Ing. Yoendris Lacoste Ricardo

Ing. Ernesto Dueñas Rodríguez

La Habana, junio, 2013

"Año 55 de la Revolución"



Somos lo que hacemos día a día. De modo que la excelencia no es un acto sino un hábito.

Aristóteles

Declaración de Autoría

Declaramos ser autoras del presente trabajo de diplomas y reconocemos a la Universidad de las Ciencias Informáticas (UCI) los derechos patrimoniales de la misma, con carácter exclusivo.

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

Dayatni Batista Rondón

_____

Firma del Autor

Yaima Carmona Acosta

_____

Firma del Autor

Ing. Ernesto Dueñas Rodríguez

_____

Firma del Tutor

Ing. Yoendris Lacoste Ricardo

_____

Firma del Tutor

Datos de contacto

Tutor: Ing. Ernesto Dueñas Rodríguez.
Especialidad de graduación: Ingeniero en Ciencias Informáticas.
Situación laboral: Profesor, Instructor Facultad 6
Años de graduado: 1
Correo Electrónico: eduenas@uci.cu
Institución a la que pertenece: Universidad de las Ciencias Informáticas (UCI).
Dirección: Carretera San Antonio de los Baños, Torrens, Municipio Boyeros, La Habana, Cuba, Código postal 19370.

Tutor: Ing. Yoendris Lacoste Ricardo.
Especialidad de graduación: Ingeniero en Ciencias Informáticas.
Situación laboral: Profesor, Instructor Facultad 6.
Años de graduado: 5
Correo Electrónico: ylacoste@uci.cu
Institución a la que pertenece: Universidad de las Ciencias Informáticas (UCI).
Dirección: Carretera San Antonio de los Baños, Torrens, Municipio Boyeros, La Habana, Cuba, Código postal 19370.

Agradecimientos

Dayatni:
Primero que todo quiero agradecerle a diosito por permitirme estar aquí hoy compartiendo este momento tan importante de mi vida con las personas que más quiero. Agradecerle a mi mamá por ser lo más lindo de mi vida, mi inspiración y mi fuerza. Es por ella que hoy estoy aquí y quiero que sepas que estoy muy orgullosa de ti. Gracias por convertirme en la persona que soy. A mi papito lindo que aunque sigue peleonero se que se siente muy orgulloso de mi. Gracias mi papito lindo por todo, por tus regaños, tus sonrisas tu apoyo. Los quiero mucho a los dos y es para ustedes que luché por llegar hasta aquí. Les agradezco mucho a mis hermanos por siempre preocuparse por mí y darme mucho apoyo y quererme tanto, mi hermanito lindo que aunque siempre está molesto conmigo se cuanto me quiere, a mi hermanita Dalin que se lo importante que soy en su vida y ella sabe como lo es para la mía, mi hermanita bebe Evelyn y a Elenia gracias. A mis abuelitos que tantos los quiero en fin le doy gracias a dios por darme la familia que tengo que se dan todo por mí. Agradecerle a una de las personas más especiales de mi vida, darte las gracias mi nene por regalarme 3 años de tu vida, por ser tan paciente conmigo, por estar junto a mí en las buenas y en las malas, gracias por ser mi amigo, mi novio, mi todo. Gracias por darme ánimo cuando pensé que no podía, gracias por hacerme tan feliz en todo este tiempo, gracias por que se que me has dado lo mejor de ti. Te amo con mi vida. A mis suegros en especial a María que tanto la quiero y tanto me aconsejó siempre, es una súper mujer le mando muchos besitos. Al cuñado de Santiago que más quiero a Rafelin, Mary y las niñas y por su puesto Ivan a Débora, les doy gracias por todo el cariño que me han dado en este tiempo.

Quiero a agradecer a mi amigas que tanto las quiero y tanto me han apoyado, a Dayami mi súper amiga que aunque hoy no está aquí conmigo se que está pidiendo para que todo me salga bien. Te quiero mucho mi niña. A Nancy que se que aunque ha pasado mucho tiempo desde que me dejo solita aquí en la UCI nunca se olvidado de mi y tanta fuerza me dio siempre, gracias mi amiga por acordarte y tenerme siempre presente. A Danay que se que tanto me quiere, gracias mi amiga quiero que sepas que te quiero mucho y que nunca me voy olvidar de todo lo que hiciste por mí, de lo momentos buenos y malos que vivimos juntas. A Ana que a pesar de todo sé que me quiere mucho, gracias por todo mi amiga de verdad. A Margarita gracias por todo taty por las malas noches con el documento, por los días que tuvimos que aguantar a los pesados. A mis amigas Leisy y Yamila que las quiero mucho pero por cosas de la vida no están hoy aquí conmigo pero siempre las tengo presente. A mi súper Gordi que ella sabe cuánto la quiero, gracias por tus consejos y tus loqueras. En fin que mis amigas son unas de mis prioridades y si en algún momento fue un poco pesada discúlpenme que ustedes saben que las quiero mucho y que siempre estarán en mi corazón. A Liena, taty muchas gracias por todo, por hacerme reír incluso cuando lo único que quería era llorar o salir corriendo, gracias por ayudarme cuando lo necesité, gracias por todo. A Delia gracias taty por todo, por los momentos buenos y malos que pasamos juntas.

Agradecimientos

Muy importante también agradecerle a mis super pesados que aunque sé que los trato fuerte como es normal en mi, con ustedes me esmero muchísimo, pero nada que los quiero cantidad solo que a ustedes hay que tenerlos controladitos. A mi Ruanito pesado que tanto quiero, a Yan, Victor, Ramiro, Jose, Alain, Dausbert y por ultimo pero no menos importante al insoportable de Darien, que se que aunque siempre estamos discutiendo lo quiero mucho y sé que el sentimiento es mutuo, gracias por todo cariño. Muy importante a mis amigotes Pabliti y Migue dios abe cuanto los quiero y cuán importante han sido en mi vida, gracias por ayudarme siempre incondicionalmente y estar ahí siempre para mi sin poner peros sin pensarlo, nunca me voy a olvidar de ustedes y de todas las cosas que hicieron por mí en estos 5 años de carrera. A Mastrapo gracias por tus chistes y tus cochinadas, gracias por tu amistad incondicional. A Yaniel gracias por todo taty por tu jodederas y tu amistad.

Muchas gracias a mis dos dúos de tesis a Yaima y en especial a Viltre por ayudarme tanto con ese ETL que por poco nos vuelve locos. Gracias taty por convertirte en un súper amigo y apoyarme en todo momento, es por eso que hoy tengo que reconocer que eres alfa, es así. Vaya que BASTA YAA nunca voy a olvidar todo lo que hiciste por mí. Yaima gracias por aguantar mis arranques y ser tan paciente conmigo, sabes que a pesar de los desacuerdos entre nosotras me gustó mucho que fueras mi dúo. Muchas gracias Claudia por estar ahí junto a nosotros haciendo tesis, ayudarme a relajarme y coger las cosas con más calma. Quiero que sepas que en mi tienes a una amiga y que voy a estar ahí para cuando te haga falta. A mis tutores y en especial a Ernesto que tanto nos ayudó, nos aconsejo y corrió con nosotras, preguntando, para que todo saliera bien. Nunca voy a olvidar que todas las llamadas por teléfono para pregunta saber que habíamos hecho.

Muchas gracias a Denia y Jorge Luis por estar aquí conmigo. Denia eres un ejemplo para mí y bueno Jorge Luis ese chucho que me das para que hablar. Fue bueno pasar tiempo con ustedes. Nunca los voy a olvidar y si un día pasan por Las Tunas pasen a verme por favor. Al Chino gracias por estar aquí este día y por hacer tan feliz a mi Nancita. A Rebeca y a Pablo por preocuparse siempre por mí, darme mucho ánimo.

Gracias a los miembros del tribunal, por ayudarnos tantos y sabemos que todas las críticas siempre fueron constructivas. A mi oponente que tanto trabajo le dimos con el documento pero bueno todo salió bien. Gracias a todas las personas que estudiaron conmigo a Mirita, Elizabeth, Marianna, Anabel, Yanet, Liena, Keimer en fin muchas gracias a toda la gente que he conocido en estos 5 años de universidad y que todos inconscientemente contribuyeron a que yo llegara hoy hasta aquí con estos resultados. Gracias por todo y quiero que sepan que nunca me voy a olvidar de ustedes.

(Yaima) Yo agradezco:

A mi padre por ser el principal impulsor en este sueño que es tanto mío como de él, por darme las fuerzas necesarias para seguir adelante, por ayudarme en los momentos más difíciles y siempre estar ahí cuando más lo necesito, por ser exigente con mis estudios y siempre estar al tanto de todo lo que me sucede, por ser mi ejemplo a seguir, mi consultor mi amigo y sobre todo un padre al que amo con todo mi corazón y del cual siempre estaré orgullosa.

A mi madre por ser tan especial, tan cariñosa, por siempre estar a mi lado, darme su apoyo incondicional, por ser mi mejor amiga, quererme tanto, por haber pasado momentos difíciles y siempre estar además ahí para mí. Pos ser además como la hermana grande que nunca tuve y porque con su paciencia y carisma me ha hecho una mejor persona. Por todas estas cosas te agradezco mamita querida.

A mi abuela maría que fue y es como una madre para mí, porque mucho de lo que soy se lo debo a ella, por ser esa persona especial que todos tenemos y a quien amamos con toda el alma, porque sé que aunque no esté físicamente, estoy segura que me cuida y hoy estará muy feliz por mí.

A mis abuelos Eloy, Doris y Leoncio por ser tan cariñosos y darme todo su amor y comprensión en estos años y a pesar de ser la nieta más vieja siempre me ven como la niña de la casa. A mis tíos Ivan y Roli por ser tan buenos y quererme tanto, por darme apoyo incondicional y a pesar de no estar tan cerca sentir que siempre están ahí. A mi tía Maiby por ser tan cómica y tan especial por conocerme tanto como mis padres y ser yo su negrita linda.

A Jorge que en estos últimos años estuvo conmigo a cada momento, apoyándome y haciéndome feliz lo cual le agradezco desde lo más profundo de mi alma, te quiero mucho.

A mis amigas a las cuales quiero con todo mi corazón, que en estos 5 años siempre han estado ahí para mí y yo para ellas en especial a Liena, Yudith, Lily y Gleivis, gracias por todos esos momentos tan especiales que nunca podre olvidar las I love you.

A mis amigos Leonel y Keke los cuales aprecio y quiero como si fueran mis hermanos, muchas gracias por estar ahí y ser tan especiales.

No pueden faltar mis grandes amigas de la infancia Neysi y Yusle a las cuales quiero con todo mi corazón, a mi amiga Claudia ella sabe que la quiero muchísimo aunque no esté aquí.

A todos los profesores que contribuyeron en mi formación como ingeniera, especialmente a los del tribunal: Maricel, Nara, Mario y nuestro oponente Rayco; por ser ejemplos a seguir y por los consejos brindados.

A nuestros tutores Ernesto y Lacoste por la ayuda brindada para que todo saliera bien. En especial a Ernesto por su comprensión, preocupación y dedicación gracias por todo, te quiero mucho.

A Viltre por toda su ayuda, a su novia Claudia por estar ahí y darnos su apoyo y a mi dúo de tesis Dayatni a la cual le he tomado mucho aprecio en todo este tiempo a pesar de nuestros desacuerdos.

Resumen

La presente investigación surge como parte de la cooperación que existe entre la Universidad de las Ciencias Informáticas y el Centro de Inmunología Molecular. Este último tiene entre sus tareas gestionar, almacenar y analizar toda la información que se recoge en el área de ensayos clínicos. El objetivo del Trabajo de diploma es desarrollar los subsistemas de almacenamiento e integración del mercado de datos Nimotuzumab para el almacén de datos de los ensayos clínicos del Centro de Inmunología Molecular permitiendo el almacenamiento homogéneo de la información que se maneja en este centro. Para esto se hizo un estudio preliminar y se seleccionaron la metodología, las herramientas y tecnologías utilizadas en el desarrollo. Además, se realizó el análisis y diseño e implementación y prueba de los subsistemas de almacenamiento e integración, obteniéndose un mercado de datos poblado, el cual permitirá realizar análisis estadísticos históricos de los principales indicadores en el área de ensayos.

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

Índice de contenido

Introducción 1
CAPÍTULO 1: Fundamento teórico....5
1.1 ¿Qué es un Ensayo Clínico? 5
1.2 Almacenes de datos 6
1.3 Arquitectura de los almacenes de datos 8
1.4 Mercados de Datos....9
1.5 Modelo dimensional....10
1.6 Tipologías de esquema....10
1.7 Integración de datos 12
1.7.1 Integración de Aplicaciones Empresariales....12
1.7.2 Integración de Información Empresarial....13
1.7.3 Extracción, transformación y carga 13
1.8 Metodologías para el desarrollo de almacenes de datos....14
1.8.1 Propuesta de Metodología para el desarrollo de soluciones de almacenes de datos que utiliza DATEC 15
1.8.2 Fases del ciclo de vida....16
1.9 Herramientas para el desarrollo de soluciones de almacenes de datos que utiliza DATEC 18
1.9.1 Herramienta de Modelado 18
1.9.2 Sistema Gestor de Base de datos....19
1.9.3 Herramienta para el perfilado de datos 21
1.9.4 Herramienta para la integración de datos 22
CAPÍTULO 2: Análisis y diseño de los subsistemas de almacenamiento e integración del MD Nimotuzumab....24
2.1 Análisis del negocio 24
2.2 Reglas del negocio 24
2.3 Especificación de requisitos....26
2.3.1 Requisitos de información....26
2.3.2 Requisitos funcionales....27
2.3.3 Requisitos no funcionales 27
2.4 Diagrama de caso de uso del sistema 27

2.4.1 Especificación de casos de uso 28
2.5 Definición de la arquitectura base del mercado 30
2.6 Diseño del mercado de datos Nimotuzumab 31
2.6.1 Diseño del subsistema de almacenamiento 31
2.6.2 Diseño del subsistema de integración 36
2.6.3 Política de respaldo y recuperación 38
2.6.4 Esquema de seguridad en el subsistema de almacenamiento 38
2.6.5 Esquema de seguridad en el subsistema de integración 38
CAPÍTULO 3: Implementación y prueba de los subsistemas de almacenamiento e integración del MD Nimotuzumab 40
3.1 Implementación del subsistema de almacenamiento 40
3.1.1 Implementación del modelo de datos físico 41
3.2 Implementación del subsistema de integración de datos 41
3.2.1 Implementación de las transformaciones 42
3.2.2 Gestión del cambio en las dimensiones 44
3.2.3 Gestión de los metadatos 45
3.3 Pruebas aplicadas 46
3.3.1 Pruebas unitarias 46
3.3.2 Pruebas de integración 47
3.3.3 Pruebas de rendimiento 47
3.3.4 Herramientas de validación 47
3.4 Resultados de las pruebas 48
Conclusiones generales 52
Recomendaciones 53
Referencias bibliográficas 54
Bibliografía 56
Anexos 58
Glosario de términos 66

Índice de Figuras

Figura 1. Diagrama de casos de uso del sistema. Relación de los actores con los casos de uso. 28
Figura 2. Arquitectura de la solución. Consta de tres niveles: las fuentes de datos y los subsistemas de integración y almacenamiento.....30
Figura 3. Modelo de datos dimensional. Representa la relación entre los hechos y las dimensiones. Evidencia una topología constelación de hechos. 35
Figura 4. Perfilado del ensayo 046. 36
Figura 5. Diseño del subsistema de integración. 37
Figura 6. Esquemas de la BD (dimenciones, mart_hr3, metadatos). 41
Figura 7. Transformación de la dimensión respuesta global. 42
Figura 8. Transformación del hecho respuesta antitumoral. 43
Figura 9. Transformación del trabajo general del mercado de datos Nimotuzumab. 44
Figura 10. Modelo V. Permite realizar pruebas a lo largo del ciclo de desarrollo del software. 46
Figura 11.Resultado general de las pruebas unitarias realizadas. 49
Figura 12. Consulta realizada a la BD. La que devuelve la cantidad de pacientes que presentan respuesta global igual a cuatro. 50
Figura 13. Resultado del perfilado de los datos al hecho hech_respuesta_antitumoral. 51

Índice de Tablas

Tabla 1. Descripción del Caso de Uso: Extraer datos. 28
Tabla 2. Matriz bus del MD Nimotuzumab.....34
Tabla 3. Permisos de los roles en la BD. 38

Introducción

Desde el surgimiento de la humanidad el hombre ha tenido la necesidad de investigar y buscar soluciones a los males que lo afectan, principalmente epidemias y enfermedades. En la actualidad una de las enfermedades que más daño y muerte ha traído consigo es el cáncer.

Existen varios centros de investigación a nivel mundial que se dedican a la búsqueda de nuevos medicamentos que ayuden al tratamiento del cáncer y así lograr una mejor calidad de vida. En su gran mayoría estos laboratorios e instituciones se encuentran en los países más desarrollados. Cuba no está exenta de contar con centros de investigaciones con estas características, destacándose el Centro de Inmunología Molecular (CIM) el que tiene como objetivo principal la búsqueda de nuevos productos para el diagnóstico y tratamiento del cáncer y enfermedades relacionadas con el sistema inmune. Esta institución realiza, en hospitales altamente especializados, Ensayos Clínicos (EC) con el objetivo de comprobar la seguridad y eficacia de los productos que allí se desarrollan (1).

Entre los EC que se han aplicado hasta el momento se encuentran los del producto Nimotuzumab también conocido como hr3. Este presenta un protocolo, documento que establece la relación de ser del estudio, sus objetivos, diseño, métodos y el análisis previsto de sus resultados, así como las condiciones bajo las que se realizará y desarrollará el estudio.

Para la realización del proceso de digitalización de los EC, los especialistas diseñan varios modelos mediante el sistema EpiData; lo que trae consigo que los diseños de los modelos en los cuadernos de recogida de datos no sean iguales. Las variables con las que se tratan los mismos términos médicos son distintas y se almacenan de manera diferente. Además, el uso del sistema EpiData hace que la forma de generar los informes no se mantenga de manera uniforme, ya que la información se puede exportar en distintos formatos (Text, Excel, Access).

Debido al gran cúmulo de datos que se genera al aplicar cada uno de los EC, el volumen de información almacenada en esta área del CIM ha aumentado considerablemente. Información que al no encontrarse integrada ni estandarizada torna engorroso el proceso de manejo de los datos, por parte de los especialistas de la institución, dificultando así, la realización de complejos análisis estadísticos dentro de un mismo EC o entre diferentes EC. En ocasiones, el análisis de la información se realiza manualmente, corriendo el riesgo que se pierda información útil y valiosa con el transcurso de los años. Todas estas

deficiencias provocan que resulte difícil realizar análisis certeros que contribuyan positivamente con las decisiones que el CIM debe tomar.
Por la situación antes expuesta se plantea como **problema de la investigación**: ¿Cómo lograr la estandarización de los datos del producto Nimotuzumab del CIM para su almacenamiento de forma homogénea?

Definiendo como **objeto de estudio**: los almacenes de datos, centrando su **campo de acción** en los subsistemas de integración y almacenamiento del Mercado de Datos (MD) Nimotuzumab.

Para darle respuesta al problema planteado se determina como **objetivo general**: desarrollar los subsistemas de almacenamiento e integración del MD Nimotuzumab para el Almacén de Datos (AD) de los EC del CIM que permita al almacenamiento homogéneo de la información.

En correspondencia con el objetivo general se plantean los siguientes **objetivos específicos**:

1. Fundamentar teóricamente la selección de la metodología, herramientas y tecnologías a utilizar.
2. Realizar el análisis y diseño de los subsistemas de almacenamiento e integración del MD Nimotuzumab.
3. Realizar la implementación y prueba de los subsistemas de almacenamiento e integración del MD Nimotuzumab.

Para cumplir el objetivo y darle una solución al problema planteado se realizaron las siguientes **tareas de la investigación**:

1. Estudio de soluciones existentes.
2. Selección de la tecnología de almacenamiento de datos.
3. Caracterización y selección de la metodología, herramientas y tecnologías a utilizar en el desarrollo de almacenes de datos.
4. Levantamiento de requerimientos para definir las necesidades del cliente. Descripción de los casos de uso del MD para determinar cada una de las funcionalidades del sistema.

5. Definición de la arquitectura del MD identificando los subsistemas fundamentales que componen la solución.
6. Definición de los hechos, las medidas y las dimensiones del MD para identificar los elementos que forman parte del modelo lógico de datos.
7. Diseño del modelo lógico de datos para determinar los elementos que conforman el modelo físico.
8. Realización del perfilado de los datos de la fuente de datos para lograr una mejor calidad de la información.
9. Diseño del subsistema de integración para definir cómo se realizará la carga de las dimensiones y los hechos al MD.
10. Implementación del modelo físico de datos para garantizar la disponibilidad de las estructuras de la Base de Datos (BD).
11. Implementación del subsistema de integración para poblar el MD.
12. Aplicación de las listas de chequeo para garantizar la correcta implementación de los subsistemas de almacenamiento e integración.
13. Aplicación de los casos de prueba para avalar la disponibilidad de cada uno de los elementos del MD.

Como **resultado esperado** se obtendrá el MD poblado Nimotuzumab.

Estructura del Trabajo de diploma

El presente Trabajo de Diploma está estructurado como se muestra a continuación: resumen, introducción, tres capítulos, conclusiones, recomendaciones, referencias bibliográficas, bibliografía, anexos y glosario de términos.

Capítulo 1: Fundamento teórico.

En este capítulo se definen conceptos fundamentales sobre los AD y los MD, con sus principales características, metas y elementos que los componen, se hace referencia al desarrollo de los procesos de integración de datos. Se fundamenta la selección de la metodología, las herramientas y tecnologías que serán utilizadas.

Capítulo 2: Análisis y diseño de los subsistemas de almacenamiento e integración del MD Nimotuzumab.
En este capítulo se hará un estudio preliminar del negocio con el objetivo de definir: reglas del negocio, casos de uso y su descripción, identificación de dimensiones, hechos y medidas. Se realiza el análisis del sistema a desarrollar, con el propósito de refinar y estructurar los requisitos obtenidos con anterioridad para facilitar su comprensión, preparación, modificación y mantenimiento.

Capítulo 3: Implementación y prueba del subsistema de almacenamiento e integración del MD Nimotuzumab.
En este capítulo se hace referencia a la implementación de la solución, abordando específicamente cómo se realiza la misma en el subsistema de almacenamiento y en el subsistema de integración, teniendo en cuenta los requisitos y necesidades del negocio. Una vez terminado de implementar los subsistemas de almacenamiento e integración de la solución, se dará paso a la realización de las pruebas, para determinar que la solución cuente con la calidad requerida.

CAPÍTULO 1: Fundamento teórico

Introducción

En el siguiente capítulo se definen y exponen conceptos relacionados con la gestión de los EC en el CIM, los AD y los procesos de integración de datos. Toda esta información será tratada y consultada a lo largo del desarrollo de la investigación para un mejor entendimiento de la solución planteada. Además se fundamenta la selección de la metodología, tecnologías y herramientas a utilizar.

1.1 ¿Qué es un Ensayo Clínico?

Según la Organización Mundial de la Salud (por sus siglas en español OMS) un EC es cualquier estudio de investigación que asigna de manera prospectiva participantes humanos o grupos de humanos a una o más intervenciones sanitarias a fin de evaluar los efectos en los resultados sanitarios. Estos estudios pueden ser abiertos o cerrados, los abiertos son aquellos en los que la investigación todavía está en curso, es decir aún no ha terminado y los cerrados son en los que las investigaciones ya han sido concluidas (2).

Los EC están compuestos por cuatro fases fundamentales estas son (3):

- La fase I: incluye los primeros estudios que se realizan en seres humanos, que pretenden demostrar la seguridad del compuesto y orientar hacia la pauta de administración más adecuada para estudios posteriores. Se puede decir que se trata de estudios de farmacología humana.

- La fase II: tiene como objetivo proporcionar información preliminar sobre la eficacia del producto y establecer la relación dosis-respuesta.

- Los EC de fase III: evalúan la eficacia y seguridad del tratamiento experimental en las condiciones de uso habituales y con respecto a las alternativas terapéuticas disponibles para la indicación estudiada. Se trata de estudios terapéuticos de confirmación.

- La fase IV: se realiza después de la comercialización del fármaco para estudiar condiciones de uso distintas de las autorizadas, como nuevas indicaciones, la efectividad y seguridad en la utilización clínica diaria.

Ensayos clínicos relacionados al producto Nimotuzumab

En el CIM se han realizado hasta la actualidad una gran cantidad de EC a los productos que se han desarrollado. Uno de los productos es el Nimotuzumab, a continuación se dará una explicación para así comprender la importancia y el uso de este producto (4):

Este producto es un antitumoral, bloquea la unión del Factor de Crecimiento Epidémico (EGF) al Receptor del Factor de Crecimiento Epidémico (EGFR), inhibe la activación de la actividad tirosina-quinasa del receptor inducida por la unión del EGF y el crecimiento tumoral en modelos de tumores humanos xenoinjertados en ratones inmunodeficientes y atímicos.

El Nimotuzumab recibe alternativamente las denominaciones comerciales de CIMAher®, TheraCIM®-hR3, Theraloc®, BIOMAb®-EGFR, YMB 1000 y VECTHIX®, y es conocido comúnmente como hR3. Se investiga la seguridad y eficacia del Nimotuzumab en: cáncer de mama, esófago, cáncer de cabeza y cuello, próstata, cáncer de hígado, páncreas, pulmón (células no pequeñas), cáncer de cuello de útero y glioma en pacientes pediátricos y adultos. Entre los estudios realizados a este producto se encuentran: EC035 hR3 cabeza y cuello Farmacodinamia, EC040 hR3 cabeza y cuello, EC046 hR3 cabeza y cuello FI, EC053 hR3 Glioma FII, EC055 hR3 cabeza y cuello FII, EC069 hR3 Glioma FIII, EC070 hR3 Mama F1, EC075 hR3 Esófago FII, EC079 hR3 Meta Cerebral FII, hR3 Glioma FI y hR3-Re Cerebro FI.

Debido al gran cúmulo de datos que generan los EC aplicados a los medicamentos que se producen en el CIM y las tecnologías que se utilizan para almacenar la información, surge la necesidad de que se implemente una aplicación que permita almacenar y estandarizar toda esta información. Por lo que se decide realizar un AD que estará conformado por varios MD, que estos almacenaran la información relacionada con los EC aplicados a un producto en específico.

1.2 Almacenes de datos

El término de AD fue definido por primera vez por Bill Inmon, lo definió como (5):

“una colección de datos orientados por temas, integrados, variables en el tiempo y no volátiles para el apoyo de la toma de decisiones” (5).

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

Por su parte, Ralph Kimball quien es una de las personalidades más influyentes en el área, propone otra definición al catalogarlo como: "…una copia de datos transaccionales, específicamente estructurados para la consulta y el análisis, es la unión de todos los MD de una entidad" (6).

Con estos dos grandes conceptos se puede decir que los AD son una colección de datos orientados por temas, integrados, no volátiles y variables en el tiempo, que ayuda al proceso de toma de decisiones de la empresa u organización, favoreciendo así el análisis de los datos.

Principales ventajas de los almacenes de datos:

- Transforma datos orientados a las aplicaciones, en información orientada a la toma de decisiones.
- Permite un análisis inmediato de las actividades de la empresa.
- Capacidad de analizar y explorar las diferentes áreas de trabajo.
- Facilidades en la gestión y análisis de recursos.
- Acelera las consultas al reducir la cantidad de datos a recorrer.
- Permite el acceso a los datos mediante un gran número de herramientas del mercado, logrando así la independencia de estas.
- Centraliza y transforma la información de manera homogénea.

Principal desventaja de los almacenes de datos:

- Para la construcción de este tipo de proyecto es necesaria una gran inversión.

1.3 Arquitectura de los almacenes de datos

La arquitectura de un AD se suele representar como varias capas a través de las cuales circulan datos, de modo que la información de una se obtiene a partir de los datos de la capa previa. Al definir la arquitectura para un AD hay que dejar claro cuáles son los bloques funcionales que se corresponden con un sistema de información completo que utiliza un AD (7):

- **Nivel operacional**: contiene datos primitivos (operacionales) que están siendo permanentemente actualizados, usados por los sistemas operacionales tradicionales que realizan operaciones transaccionales.

- **Almacén de datos**: contiene datos primitivos correspondientes a sucesivas cargas del AD y algunos datos derivados. Los datos derivados son generados a partir de los primitivos al aplicarles algún tipo de procesamiento (resúmenes).

- **Nivel departamental**: contiene exclusivamente datos derivados. Cada departamento de la empresa determina su nivel departamental con información de interés a dicho nivel. Va a ser el blanco de salida sobre el cual los datos en el almacén son organizados y almacenados para las consultas directas por los usuarios finales, los desarrolladores de reportes y otras aplicaciones.

- **Nivel individual**: contiene pocos datos, resultado de aplicar heurísticas, procesos estadísticos, a los datos contenidos en el nivel anterior. El nivel individual es el objetivo final de un AD. Desde este nivel accederá el usuario final y se podrán plantear diferentes hipótesis, así como navegar a través de los datos contenidos en el AD.

A partir de esta arquitectura se considera que el desarrollo de un AD se puede estructurar en un marco integrado por tres niveles los cuales son (7):

- **Conceptual**: define el AD desde un punto de vista conceptual, o sea, desde el mayor nivel de abstracción, contiene únicamente los objetos y relaciones más importantes.

- **Lógico**: abarca aspectos lógicos del diseño del AD, como la definición de las tablas, claves, procesos de integración y otros.

- **Físico**: define los aspectos físicos del AD, como el almacenamiento de las estructuras lógicas o la configuración de los servidores que mantienen el AD.

1.4 Mercados de Datos

Un MD es una BD departamental, especializada en el almacenamiento de los datos de un área de negocio específica. Se caracteriza por disponer la estructura óptima de datos para analizar la información al detalle desde todas las perspectivas que afecten a los procesos de dicho departamento (8).

Principales ventajas de la utilización de los mercados de datos (8)

- Acelera las consultas al reducir la cantidad de datos a recorrer.
- Estructura los datos para su adecuado acceso por una herramienta.
- Centraliza y transforma la información de manera homogénea.

Con lo planteado anteriormente se puede llegar a la conclusión de que los MD son subconjuntos del AD, diseñados para satisfacer necesidades de un área de negocio específica por lo que el uso efectivo de estos son un factor importante en el funcionamiento del AD y a su vez en el éxito del proyecto.

Los MD están compuestos por tres subsistemas los cuales son (8):

- **Subsistema de almacenamiento:** es donde se crea el modelo multidimensional que contiene las tablas de hechos, de dimensiones y las relaciones que existen entre ellas. Además de almacenar la información en dichas tablas.
- **Subsistema de integración:** es el encargado de limpiar la información, así como estandarizarla e integrarla preparándola para la carga de datos.
- **Subsistema de visualización:** tiene como objetivo organizar los reportes por áreas de análisis facilitando al usuario una búsqueda rápida de la información.

Por las características de esta investigación solo se desarrollaron los subsistemas de almacenamiento e integración.

1.5 Modelo dimensional

El modelo dimensional divide el mundo de los datos en dos conjuntos: las tablas de hecho y de dimensiones. A continuación se explican cada una de ellas (9):

Tabla de hechos: es la tabla central en un esquema multidimensional. Es en ella donde se almacenan las mediciones numéricas del negocio. Estas medidas se hacen sobre el grano o unidad básica de la tabla.

Tabla de dimensiones: se conoce como dimensión a la característica de un hecho que permite su análisis posterior en el proceso de toma de decisiones y brinda una perspectiva adicional a un hecho dado. Son agrupaciones lógicas de atributos con un significado común y atómico.

Medidas: atributo numérico de un hecho que representan el comportamiento del negocio relativo a las dimensiones.

1.6 Tipologías de esquema

La tipología de esquema no es más que la forma en la cual se va a estructurar el depósito de datos. Es importante definir que tipología se empleará, ya que esta decisión afecta considerablemente la elaboración de los modelos dimensional y físico. Generalmente se utiliza la que se adapte mejor a los requerimientos y necesidades del cliente.

Esquema estrella: su estructura es una tabla central y un conjunto de tablas que la atienden radialmente, de este se deriva su nombre, del hecho que su diagrama forma una estrella, con puntos radiales desde el centro. El esquema en estrella es el más simple de interpretar y optimiza los tiempos de respuesta ante las consultas de los usuarios. Este modelo es soportado por casi todas las herramientas de consulta y análisis. Es necesario destacar que este es un esquema totalmente desnormalizado.

A continuación se muestran algunas características de este modelo, que ayudarán a comprender mejor sus ventajas (10):

- Posee los mejores tiempos de respuesta.
- Su diseño es fácilmente modificable.
- Existe paralelismo entre su diseño y la forma en que los usuarios visualizan y manipulan los datos.
- Simplifica el análisis.
- Facilita la interacción con herramientas de consulta y análisis.

Esquema copo de nieve (Snowflake): es un esquema derivado del esquema de estrella, las tablas de dimensiones se ramifican en más puntas. Este modelo es más cercano a un modelo de entidad relación, que al modelo en estrella, debido a que sus tablas de dimensiones están normalizadas (11).

Uno de los motivos principales de utilizar este tipo de modelo, es la posibilidad de segregar los datos de las tablas de dimensiones y proveer un esquema que sustente los requisitos de diseño. Otra razón es que es muy flexible y puede implementarse después de que se haya desarrollado un esquema en estrella.

Se pueden definir las siguientes características de este tipo de modelo (11):

- Posee mayor complejidad en su estructura.
- Hace una mejor utilización del espacio.
- Es muy útil en tablas de dimensiones de muchas tuplas.
- Las tablas de dimensiones están normalizadas, por lo que requiere menos esfuerzo de diseño.
- Puede desarrollar clases de jerarquías fuera de las tablas de dimensiones, que permiten realizar análisis de lo general a lo detallado y viceversa.

A pesar de todas las características y ventajas que trae aparejada la implementación del esquema copo de nieve, existen dos grandes inconvenientes:

- Si se poseen múltiples tablas de dimensiones, cada una de ellas con varias jerarquías, se creará un número de tablas bastante considerable, que pueden llegar al punto de ser inmanejables.
- Al existir muchas uniones y relaciones entre tablas, el desempeño puede verse reducido.

La existencia de las diferentes jerarquías de dimensiones debe estar bien fundamentada, ya que de otro modo las consultas demorarán más tiempo en devolver los resultados, debido a que se deben realizar las uniones entre las tablas (11).

Constelación de hechos: la constelación de hechos es un conjunto de tablas de hechos que comparten algunas tablas de dimensiones. Se puede decir que está compuesto por una serie de esquemas en estrella. Su diseño y cualidades son muy similares a las del esquema en estrella, pero posee una serie de diferencias con el mismo, que son precisamente las que lo destacan y caracterizan (12).

En este trabajo de diploma la tipología que se utilizó es la constelación de hechos ya que permite tener más de una tabla de hechos. Se podrán analizar más aspectos claves del negocio con un mínimo esfuerzo adicional de diseño. Contribuye a la reutilización de las tablas de dimensiones, pues una misma tabla de dimensión puede utilizarse para varias tablas de hechos.

1.7 Integración de datos

La mayoría de las organizaciones cuentan con aplicaciones que se encargan del manejo de los datos. Con el pasar de los años el volumen de información aumenta, dificultando la realización de un análisis detallado de la misma. Es por ello que el objetivo fundamental de la integración de datos, es permitir el desarrollo de nuevas aplicaciones que analicen información de múltiples fuentes y finalmente mostrar una vista unificada de esta información.

De acuerdo con lo descrito por Laura Haas (14), la integración de datos es un proceso con cuatro tareas principales: la comprensión, la normalización, especificación y ejecución.

Comprensión: la primera tarea de la integración de datos es analizar y entender la fuente. Durante esta tarea, el integrador puede buscar relaciones entre los datos y sus significados.
Normalización: esta tarea aprovecha el trabajo de la tarea comprensión para determinar el mejor método de integración, la forma de limpiar o reparar los datos.
Especificación: en esta tarea se producen los artefactos que controlarán la ejecución. Las técnicas y tecnologías utilizadas para obtener las especificaciones, están íntimamente vinculadas a la elección del motor de ejecución. La especificación es parte de la configuración de un motor de integración para hacerla deseada.
Ejecución: aquí es donde realmente sucede la integración de datos y puede ser lograda a través de la materialización, federación o indexación.

En la actualidad existen varias tecnologías para la integración de datos, a continuación se presentan las principales.

1.7.1 Integración de Aplicaciones Empresariales

Integración de Aplicaciones Empresariales. (Enterprise Application Integration por sus siglas en inglés EAI), engloba las metodologías, procesos, herramientas y tecnologías usadas para conectar sistemas, datos y procesos de una entidad o de un conjunto de entidades. Cuando la conexión es entre sistemas,

datos o procesos de distintas entidades se suele hablar de Business to Business integration o B2Bi. Por lo que se puede decir que utiliza una comunicación punto a punto (15).

1.7.2 Integración de Información Empresarial

La Integración de Información Empresarial (Enterprise Information Integration por sus siglas en inglés EII), es la integración de datos a partir de múltiples sistemas en un formato de representación unificado, coherente y preciso, orientado a la manipulación de datos y la navegación, donde estos se mantienen en los sistemas de información. Así que los datos son agregados, reestructurados, re-etiquetados (si es necesario) y presentados a un usuario. Por lo general, el resultado de este enfoque es prácticamente un sistema distribuido, heterogéneo, integrado de información (15).

1.7.3 Extracción, transformación y carga

Extracción, transformación y carga de datos (por sus siglas en ingles ETL) extraen datos de las diversas fuentes que se requieran. Los transforman para resolver posibles problemas de inconsistencias entre los mismos y finalmente después de haberlos depurado, se procede a su carga en el depósito de datos (15).

Extracción: consiste en sustraer los datos brutos desde las fuentes de origen, integrando en una misma metodología de negocios, toda la información empresarial proveniente de diferentes fuentes. Por lo general los datos brutos se ubican en BD relacionales o ficheros planos, igual pueden incluir BD no relacionales y diferentes estructuras de datos.
Transformación: una vez terminada la fase de extracción se realiza un chequeo que verifique si los datos cumplen con las pautas estipuladas, en caso de que los datos no cumplan, se aplican una serie de procedimientos donde estos quedarían listos para ser cargados.
Carga: la fase de carga interactúa directamente con la BD destino, debido a que los datos son incluidos en el sistema dependiendo de los requerimientos de la organización. En algunas BD, se sobrescribe la información antigua con los nuevos datos, pero en el caso de los AD, estos mantienen un historial de los registros de manera que se pueda hacer una auditoría de los mismos y disponer de un rastro del comportamiento de un determinado valor a lo largo del tiempo.

En el siguiente trabajo se ha decido escoger como técnica de integración de datos ETL debido a las inconsistencias que existen en los datos y al gran volumen de información con la que se trabaja, por lo que se hace necesario realizar complejas combinaciones de transformaciones.

1.8 Metodologías para el desarrollo de almacenes de datos

Una metodología es el conjunto de métodos por los cuales se rige una investigación científica. Por lo tanto lo que hace la metodología es estudiar los métodos para luego determinar la más adecuada a aplicar o sistematizar en una investigación.

No existe una única metodología en la cual basarse para la construcción de un AD, sino que dependiendo del contexto que se encuentre la empresa y los objetivos que persiga, se puede emplear una u otra, estas metodologías se engloban dentro de dos grandes enfoques: ascendente (top-down) y descendente (bottom-up) que se corresponden con las metodologías propuestas por Bill Inmon y Ralph Kimball respectivamente.

La principal diferencia que existe entre estas metodologías es la forma de enfrentar el problema. El enfoque de Inmon es un enfoque descendente, por lo que al ser construidos descendentemente los MD se nutren del AD. Es un método sistémico, que minimiza los problemas de integración, pero es costoso, debido a la gran cantidad de datos y su poca flexibilidad. Por el contrario, el enfoque que propone Kimball es ascendente, pues al final el AD no es más que la unión de los diferentes MD. Esta característica le hace más flexible y sencillo de implementar, pues se puede construir un MD como primer elemento del sistema y luego ir añadiendo otros que comparten las dimensiones ya definidas o incluyen otras nuevas. Este enfoque parte de los requisitos del negocio, mientras que el enfoque descendente propone la validación de los requisitos una vez que se tiene el sistema (16).

Ciclo de vida de la metodología propuesta por Ralph Kimball

El ciclo de vida Kimball comienza con una planificación de proyecto, donde se define el alcance, se identifican y programan las tareas, se planifica el uso de los recursos, conformando con todo esto el plan de proyecto. En la segunda etapa se definen los requerimientos del negocio. Luego de esto el proyecto se enfoca en tres líneas concurrentes: tecnología, datos y aplicaciones de inteligencia de negocio (BI). El ciclo de vida culmina con el despliegue y mantenimiento del producto.

Para definir la metodología de desarrollo a utilizar en el Departamento de AD de DATEC, se tomó como base la Metodología de Kimball por los siguientes elementos:

- Crea los conceptos de hechos y dimensiones.

- Propone ir construyendo el AD a través de la construcción de los MD departamentales, lo cual coincide con la división lógica de las empresas, entidades u organismos. Constituye una buena estrategia pues permite ir presentando resultados parciales a los clientes en cortos plazos.
- Existe abundante documentación sobre la misma y se puede consultar la web a través de los servicios que brindan el grupo creador de la metodología.

A pesar de todas las ventajas que ofrece la utilización de la Metodología de Kimball, esta no es totalmente adaptable a las características del centro y de la producción en la UCI, por lo que solo se decidió utilizarla como guía en el proceso de confección de una metodología de desarrollo para el Departamento de AD. Entre sus principales desventajas se encuentran:

- No tiene definido un criterio que permita estimar los costos de desarrollo de un AD, basándose en las características de la construcción del mismo.
- Presenta un grupo de roles, pero no explica claramente cuáles son las competencias y responsabilidades de cada uno dentro del proyecto. Por la cantidad de roles que propone se necesita de grupos grandes para su desarrollo.
- Está estructurada para el desarrollo de proyectos – productos, donde un proyecto desarrolla un producto determinado.
- No establece el análisis de diferentes criterios de diseño en el levantamiento de requisitos que permita la construcción más adecuada del almacén, teniendo en cuenta las metas de la organización, las necesidades de los usuarios y la disponibilidad de las fuentes primitivas.

Por tales motivos se definió una metodología que permite mitigar las desventajas identificadas en la Metodología de Kimball y ajustada a las condiciones y características de producción de DATEC y de la UCI.

1.8.1 Propuesta de Metodología para el desarrollo de soluciones de almacenes de datos que utiliza DATEC

La propuesta de metodología para el desarrollo de soluciones de AD que utiliza DATEC, se basa en el ciclo de vida Kimball y en la propuesta realizada por Leopoldo Zenaido Zepeda Sánchez en su tesis de doctorado, en la cual plantea incluir los casos de uso para guiar el proceso de desarrollo. Está adaptada a las necesidades de la UCI y cubre las fases por las que pasa la construcción de un AD. Las

particularidades que presenta este modelo de adaptación es la identificación de requisitos de información y a su vez la trazabilidad que tienen estos en todo el ciclo de desarrollo del MD. También la inclusión de una etapa de pruebas que fortalece en gran medida la calidad con que se despliegue la solución propuesta. Además, ajusta las fases, actividades y artefactos a la propuesta de programa de mejoras que lleva a cabo el Centro Nacional de Calidad de Software (CALISOFT) con el objetivo de alcanzar el nivel dos del Modelo de Integración de Capacidades de Madurez (conocido como Capability Maturity Model Integration CMMI) en los centros productivos de la UCI (16).

Definición del ciclo de vida de la metodología

En una metodología de desarrollo pueden unificarse varios modelos de desarrollo, por lo que después de analizar los modelos utilizados en la industria de software y según las características de desarrollo de DATEC, se decidió utilizar el Modelo Incremental y el Modelo de Desarrollo Rápido de Aplicaciones (DRA). El ciclo de vida de la metodología deberá estar organizado por fases, algunas de ellas podrán ser implementadas de forma paralela según el componente que se está desarrollando, los componentes se integran a medida que avanza el desarrollo de la solución. Esto permite agilizar la producción, reduciendo los tiempos de desarrollo (16).

1.8.2 Fases del ciclo de vida

Las fases se definieron teniendo en cuenta las propuestas por la metodología de Kimball, los procesos y actividades presentados anteriormente y las características del desarrollo de proyectos de software en la UCI. Finalmente se obtuvieron ocho fases que se describen a continuación.

Estudio preliminar y planeación: se realiza un estudio minucioso en la entidad cliente. Esto incluye un diagnóstico de información, de datos y de infraestructura tecnológica, con el fin de determinar lo que se desea construir y lo que existe para el desarrollo y montaje de la misma. También se llevan a cabo las tareas de planeación del proyecto, se definen los objetivos, el alcance preliminar, los costos estimados, los recursos necesarios, y otras series de actividades.

Levantamiento de requisitos: se realiza en tres direcciones, 1ra. Identificación de las metas y objetivos de la organización, 2da. Identificación de las necesidades de información de los clientes y las reglas de negocio; y 3ra. Haciendo un levantamiento detallado de cada una de las fuentes de datos a integrar para validar la disponibilidad de la información.

Arquitectura: se define la arquitectura de la solución, aspectos como, la seguridad del sistema, la comunicación entre los subsistemas, la tecnología a utilizar, hardware y software, entre otros aspectos de gran importancia. Vale aclarar que esta fase puede desarrollarse en paralelo con la fase de levantamiento de requisitos, siempre y cuando los resultados del diagnóstico tecnológico realizado durante la fase de estudio preliminar dejen bien definidas las características técnicas de la organización y el cliente sepa lo que desea.

Diseño e Implantación: se define el diseño de las estructuras de almacenamiento, se diseñan los procesos de integración de datos como, las reglas de extracción, transformación y carga, se diseñan los cubos para la presentación de los datos, así como el diseño visual 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: aquí se realizan varias pruebas, comenzando por las pruebas de unidad llevadas a cabo por los propios desarrolladores de cada uno de los grupos, luego las pruebas de integración y sistema, hasta llegar a las pruebas de aceptación con el cliente final.

Despliegue: consta de dos etapas, la primera es un despliegue piloto, donde se configuran los servidores necesarios y se instalan las herramientas según la arquitectura definida, se cargan una muestra de los datos en un ambiente controlado, con el fin de demostrarle al cliente final que la solución funciona. Una vez aceptada la solución por el cliente, se realiza la carga histórica de los datos. Es aquí el momento más idóneo para llevar a cabo la capacitación y transferencia tecnológica a los clientes.

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, hasta el acompañamiento junto al cliente.

Gestión y administración 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 y administración de proyecto. En la ejecución de cada fase participan los grupos de trabajo por lo que los roles del proyecto están distribuidos por diferentes grupos (16).

Debido a que en el trabajo de diploma solo se implementarán los subsistemas de almacenamiento e integración solo se llegará hasta la fase de prueba. Las restantes fases serán desarrolladas por los especialistas vinculados al proyecto.

1.9 Herramientas para el desarrollo de soluciones de almacenes de datos que utiliza DATEC

El flujo de datos desde los sistemas fuentes hacia el MD en este tipo de soluciones, se logra mediante el uso de herramientas, la cuales permitirán realizar un correcto desarrollo a lo largo del trabajo.

1.9.1 Herramienta de Modelado

Las herramientas de Ingeniería de Software Asistida por Ordenador (por sus siglas en ingles CASE) son de gran ayuda para el desarrollo de software. Cuentan con un grupo de programas que utilizan las personas que intervienen en el desarrollo de los mismos, con el objetivo de agilizar y facilitar el trabajo. Estas herramientas proveen métodos, técnicas y utilidades que ayudan al perfeccionamiento del desarrollo de sistemas de información, de forma total o parcial.

Visual Paradigm 8.0

Visual Paradigm para el Lenguaje Unificado de Modelado (por sus siglas en ingles UML) es una herramienta profesional que soporta el ciclo de vida completo del desarrollo de software: análisis y diseño orientados a objetos, construcción, pruebas y despliegue. Este software ayuda a una rápida construcción de aplicaciones de calidad, mejores y a un menor coste. Permite dibujar todos los tipos de diagramas de clases, código inverso, generar código desde diagramas y generar documentación. Proporciona además abundantes tutoriales, demostraciones interactivas y proyectos UML. Es multiplataforma por lo permite su uso en cualquier sistema operativo. Brinda una integración con sistemas de control de versiones que almacenan centralmente los artefactos y realizan un seguimiento de los cambios realizados sobre un proyecto. Está disponible en varias ediciones, cada una destinada a distintas necesidades: empresarial, profesional, comunidad, estándar, modelador y personal (19).

Se integra con las siguientes herramientas Java:

- Eclipse/IBM WebSphere
- JBuilder
- NetBeans IDE
- Oracle JDeveloper
- BEA Weblogic(20)

Debido a las características que se describieron anteriormente se utilizó el Visual Paradigm en su versión 8.0 como herramienta de modelado.

1.9.2 Sistema Gestor de Base de datos

Un Sistema Gestor de BD (DBMS por sus siglas en inglés) está conformado por programas que permiten crear y dar soporte a una BD, asegurando su integridad, confidencialidad y seguridad (17). Ejemplos de gestores de BD:

- PostgreSQL.
- SQLite.
- DB2 Express-C.
- Apache Derby.
- Microsoft SQL Server.
- Sybase ASE Express Edition.
- Oracle Express Edition 10.

PostgreSQL 9.1:

PostgreSQL es un sistema de gestor de BD objeto-relacional, con su código fuente disponible libremente, es una aplicación de código abierto y en el mercado es una de las más potentes. PostgreSQL utiliza un modelo cliente/servidor y usa multiprocesos en vez de multihilos para garantizar la estabilidad del sistema. De ocurrir algún tipo de fallo en algunos de los procesos no afectaría a los demás procesos y el sistema continuará funcionando correctamente (17).

Principales ventajas

- Diseñado para ambientes de alto volumen.
- Multiplataforma.
- Extensible.
- Ahorros considerables en costos de operación.
- Mejor soporte que los proveedores comerciales.
- Tamaño de BD desmedido.
- Es una herramienta libre.

Debido a las características que se describieron anteriormente se utilizó el PostgreSQL en su versión 9.1 como sistema gestor de BD.

PgAdmin3 1.14

PgAdmin es uno de los más populares administradores de BD por sus características en el mundo. Esto se debe en gran medida a que es una aplicación de código abierto y multiplataforma (19). PgAdmin está diseñado para responder a las necesidades de todos los usuarios, desde la escritura de consultas SQL para crear BD complejas. Su interfaz gráfica soporta todas las características de PostgreSQL y facilita la administración.

La conexión con el servidor se puede hacer a través de TCP/ IP o Unix Domain Sockets. No necesita controladores adicionales para comunicarse con el servidor de BD. PgAdmin es desarrollado por una comunidad de expertos de PostgreSQL en todo el mundo y está disponible en más de una docena de idiomas. Es Software Libre publicado bajo la licencia de PostgreSQL (18).

Debido a las características que se describieron anteriormente se utilizó el PgAdmin en su versión 1.14 como administrador de BD.

Tipos de almacenamiento OLAP

Los sistemas de Procesamiento Analítico en Línea (por sus siglas en ingles OLAP) son BD orientadas al procesamiento analítico. Este análisis suele implicar, generalmente, la lectura de grandes cantidades de datos para llegar a extraer algún tipo de información útil: tendencias de ventas, patrones de comportamiento de los consumidores, elaboración de informes complejos. Existen diferentes tipos de almacenamientos OLAP a continuación se explican cada uno de ellos.

Procesamiento Analítico Multidimensional en línea (MOLAP)

La arquitectura MOLAP usa BD multidimensionales para proporcionar el análisis, su principal premisa es que el OLAP está mejor implantado, almacenando los datos multidimensionalmente. Utiliza una arquitectura de dos niveles: las BD multidimensionales y el motor analítico donde la BD multidimensional es la encargada del manejo, acceso y obtención del dato. El nivel de aplicación es el responsable de la ejecución de los requerimientos OLAP. El nivel de presentación se integra con el de aplicación y proporciona una interfaz a través de la cual los usuarios finales visualizan los análisis OLAP. Una arquitectura cliente/servidor permite a varios usuarios acceder a la misma BD multidimensional.

Procesamiento Analítico Relacional en Línea (ROLAP)

La arquitectura ROLAP, accede a los datos almacenados en un AD para proporcionar los análisis OLAP. La premisa de los sistemas ROLAP es que las capacidades OLAP se soportan mejor contra las bases de datos relacionales. La arquitectura ROLAP es capaz de usar datos pre calculados si estos están disponibles, o de generar dinámicamente los resultados desde los datos elementales si es preciso. Esta arquitectura accede directamente a los datos del AD, y soporta técnicas de optimización de accesos para acelerar las consultas.

Procesamiento Analítico Híbrido en Línea (HOLAP)

Combina las arquitecturas ROLAP y MOLAP para brindar una solución con las mejores características de ambas: desempeño superior y gran escalabilidad. Un tipo de HOLAP mantiene los registros de detalle (los volúmenes más grandes) en la BD relacional, mientras que mantiene las agregaciones en un almacén MOLAP separado (13).

Después de realizarse un análisis de los tipos de almacenamiento de datos mencionados anteriormente, se llegó a la conclusión de que el más adecuado a utilizar en la solución es ROLAP, debido a que el gestor de BD escogido no soporta un modelo multidimensional.

1.9.3 Herramienta para el perfilado de datos

El perfilado de datos es un proceso que se lleva a cabo antes y después de los procesos de integración de datos. El primer perfilado se realiza con el objetivo de conocer la calidad de los datos provenientes de los sistemas fuentes, además brinda una noción de las posibles reglas del negocio y transformaciones en el ETL. El segundo perfilado de datos se realiza una vez terminados los procesos de integración de datos para comprobar que los datos cargados en el mercado tengan la calidad requerida y que no existan inconsistencias en los mismos.

DataCleaner 1.5.3

DataCleaner es una aplicación Open Source para el análisis, perfilado, transformación y limpieza de datos. Estas actividades ayudan a administrar y controlar la calidad de los datos. DataCleaner es la alternativa gratuita al software de gestión de datos maestros, metodologías, almacenamiento de datos proyectos, la investigación estadística, la preparación para el ETL y más actividades. Dentro de las principales características por lo cual se decidió utilizar dicha herramienta se encuentran las siguientes:

- Es una aplicación de código abierto.
- Es muy fácil de utilizar.

- Genera sofisticados informes y gráficos que permiten a los usuarios determinar el nivel de calidad de los datos e identificar y analizar la estructura del origen de los mismos.

Se considera una alternativa libre para la metodología de administración de la información, para proyectos de AD, búsquedas estadísticas, para actividades de preparación de ETL entre otros (20).

Debido a las características que se describieron anteriormente se utilizó el DataCleaner en su versión 1.5.3 para el perfilado de los datos.

1.9.4 Herramienta para la integración de datos

Las herramientas para la integración de datos agilizan el proceso de ETL, garantizan la homogeneidad y calidad de la información. Entre las más conocidas se encuentran:

- Talend Data Integration.
- Pentaho Data Integration.

Pentaho Data Integration en su versión 4.2.1.

Es una herramienta libre, muy potente, antigua y una de las más utilizadas por los usuarios, considerándola la más completa por la gran cantidad de conectores que posee y la posibilidad de crear flujos de trabajo integrados con transformaciones de datos de manera muy sencilla y funcional (21). Dentro de las principales características de dicha herramienta se encuentran las siguientes:

- Posee un entorno gráfico de desarrollo.
- Uso de tecnologías estándar: Java, XML, JavaScript.
- Es fácil de instalar y de configurar.
- Es multiplataforma: Windows, Macintosh, Linux.
- Basado en dos tipos de objetos: transformaciones (colección de pasos en un proceso ETL) y trabajos (colección de transformaciones).
- Es un software de código abierto.
- Sin costes de licencia.

Está formado por un conjunto de herramientas, cada una con un propósito específico:

Spoon: herramienta gráfica que permite el diseño de las transformaciones y trabajos. Incluye opciones para pre visualizar y testear los elementos desarrollados. Es la principal herramienta de trabajo de Pentaho Data Integration y con la que se construyen y validan los procesos de ETL.

PAN: herramienta que permite la ejecución de las transformaciones diseñadas en Spoon (bien desde un

fichero o desde el repositorio). Permite desde la línea de comandos preparar la ejecución mediante scripts.

CHEF: para crear trabajos.

Kitchen: permite ejecutar los trabajos batch diseñados con CHEF.

Principales ventajas del Pentaho Data Integration:

- Tiene una interfaz gráfica con indicadores de las transformaciones.
- Es una aplicación implementada en Java con algunas características avanzadas en JavaScript.
- Basada en metadatos.
- Como soporte se encuentran los foros de Pentaho y la comunidad Pentaho.
- Soporta Oracle, DB2.SQL Server, Sybase así como MySQL y Postgres.
- Soporta la arquitectura de procesamiento en paralelo para distribuir las tareas de ETL a través de múltiples servidores, basados en dos tipos de objetos: transformaciones y trabajos (21).

Debido a las características que se describieron anteriormente se utilizó el Pentaho Data Integration en su versión 4.2.1 para la integración de datos.

Conclusiones del capítulo

En este capítulo se muestra una panorámica general del proceso de desarrollo de un AD, así como una caracterización de las metodologías, herramientas y tecnologías que se utilizaron en este trabajo de diploma. Luego de esta investigación se arribaron a las siguientes conclusiones:

- Se seleccionó como metodología la propuesta de metodología para el desarrollo de soluciones de AD que utiliza DATEC.
- Se escogió como herramienta de modelado al Visual Paradigm en su versión 8.0 para el modelado del diseño e implementación de la solución.
- Se estableció el uso del PostgreSQL en su versión 9.1 como gestor de BD y como cliente para la administración de la BD el PgAdmin3 en su versión 1.14.
- Se seleccionó el DataCleaner en su versión 1.5.3 para el perfilado de los datos, permitiendo arribar a conclusiones sobre el estado y calidad de la información almacenada en el sistema fuente.
- Se seleccionó el Pentaho Data Integration en su versión 4.2.1 como herramienta para la integración de datos.

CAPÍTULO 2: Análisis y diseño de los subsistemas de almacenamiento e integración del MD Nimotuzumab.

Introducción

En este capítulo se realiza un análisis del negocio, con el propósito de comprender los principales aspectos de relevancia para el CIM. Se especifican las necesidades de información, Reglas del Negocio (RN), Requisitos Funcionales (RF), Requisitos No Funcionales (RNF), Requisitos de Información (RI) y los Casos de Uso del Sistema (CUS). Además se realiza el diseño de los subsistemas de almacenamiento e integración.

2.1 Análisis del negocio

El CIM realiza actividad científica y productiva obteniendo resultados relevantes en la búsqueda de medicamentos que contribuyan al tratamiento del cáncer. Debido a la gran cantidad de información que se procesa en dicho centro, esta es agrupada en diversos departamentos o áreas, siendo uno de estos los de EC. Uno de los productos desarrollados en esta institución es el hr3. La información que han generado algunos de los ensayos aplicados al mismo es almacenada en ficheros excel, exportados mediante el sistema EpiData. Durante la realización de dichos estudios se suministra a los pacientes el medicamento, comprobando mediante exámenes físicos y de laboratorio, las reacciones ante el tratamiento. Se les evalúa cada cierto tiempo los signos vitales. Además, se registran las respuestas, las lesiones y los eventos adversos presentados por los pacientes durante el proceso. Una vez concluidos dichos ensayos se recogen los datos de los pacientes que salieron o fallecieron durante los EC.

Debido a las necesidades de información existentes se almacenarán en el MD referentes a los siguientes EC:

- Los EC para el tratamiento del cáncer de cabeza y cuello (035, 040, 046 y 055).
- Los EC para el tratamiento del cáncer de glioma (053, 069 y glioma fase1).

2.2 Reglas del negocio

La función fundamental de las RN es definir políticas o condiciones que deben cumplirse para alcanzar el correcto cumplimiento de los objetivos del sistema, por lo que regulan aspectos del negocio. A continuación se muestran las 15 RN definidas:

RN1. Cuando no aparece la edad se toma la diferencia que existe entre la fecha actual y la fecha de nacimiento.

RN2. En el caso de que un paciente no cumpla con el modelo de inclusión se determina que no forma parte del ensayo.

RN3. Para los tratamientos previos se determinó que si el paciente tiene al menos uno de los tres (quimioterapia, radioterapia o cirugía) ya se asume que el paciente tuvo algún tratamiento previo.

RN4. Si un paciente está en un grupo de tratamiento donde se definió en el protocolo que recibe placebo (ningún producto) o es control, se define que el nivel de dosis es 0.

RN5. Si el nivel de dosis es 0 y presenta eventos adversos, la causalidad del evento adverso es no relacionada.

RN6. Cuando el ensayo se aplica a pacientes que se encuentran en fase I aparece el nivel de dosis recibido; sin embargo, si se está en fase II, lo que aparece es el grupo de tratamiento y se debe buscar en el protocolo el nivel de dosis de cada grupo.

RN7. El sexo se recoge en la mayoría de las BD como 1 y 2. Se determinó cambiar estos valores por 1(femenino) y 2(masculino).

RN8. La raza se recoge en la mayoría de las BD como 1, 2, 3 y 4. Se determinó cambiar estos valores por 1(blanca), 2(negra), 3(mestiza) y 4(amarilla).

RN9. Los únicos tratamientos previos que se tendrán en cuenta para el análisis serán quimioterapia, radioterapia y cirugía. Se determinó que en los ensayos donde no se hayan recogido, ese campo aparecerá como ¨No procede¨.

RN10. En el caso de las perspectivas que no se les encuentren relación con los datos fuentes en algunas de las BD porque no se recogieron en ese ensayo (Ej.: raza, causalidad, edad) se registra en el MD como Desconocido.

RN11. Los identificadores de los indicadores y de las dimensiones no pueden tomar valores nulos, ni repetidos.

RN12. En algunos ensayos el hospital y la provincia en algunos ensayos aparecen codificados en números (1-LI, 2-INOR), por lo que se decide almacenarlos como (LI, INOR).

RN13. Los hospitales (CH, INNOR, HHA) se almacenarán como (CHR, INOR y HA).

RN14. En ocasiones los resultados de los exámenes de laboratorio aparece nulo en la fuente por lo que se decide que el rango que tomará es No disponible.

RN15. Si la causa de salida o de fallecimiento tiene un valor nulo en la fuente de datos se almacenarán como "Conclusión del ensayo" y "No falleció durante el ensayo" respectivamente.

2.3 Especificación de requisitos

Los requisitos son las necesidades del producto así como un punto clave en el desarrollo de las aplicaciones informáticas. Un gran número de proyectos naufragan debido a una mala definición de los mismos, por eso es necesario que sean especificados por escrito como todo contrato o acuerdo entre dos partes y posible de probar o verificar (22).

2.3.1 Requisitos de información

Los requisitos de información son las principales informaciones que deben estar disponibles al realizar los análisis sobre los datos. Con el objetivo de mantener disponible la información de los EC realizados en el CIM al producto hr3 para los tipos de cáncer de cabeza, cuello y glioma. A continuación se muestran los ocho RI identificados:

RI1. Obtener la cantidad de pacientes incluidos y evaluados en los EC del producto hr3 atendiendo a la edad, raza, peso, talla, sexo, hospital, provincia, Ecog, Karnofsky, estadío, grado de diferenciación, tiempo y tratamientos previos.

RI2. Obtener la cantidad de pacientes que presentaron eventos adversos en los EC del producto hr3 atendiendo al hospital, tiempo, tipo, grado de severidad y causalidad de los eventos adversos.

RI3. Obtener la cantidad de pacientes que salieron de los EC del producto hr3 atendiendo a hospital, grupo de tratamiento, tiempo, causa de salida y causa de fallecimiento.

RI4. Obtener la cantidad de pacientes que presentaron lesiones en los EC del producto hr3 atendiendo al hospital, tiempo, provincia, localización de la lesión y diagnóstico de la lesión.

RI5. Obtener la cantidad de pacientes que se le realizaron exámenes físicos en los EC del producto hr3 atendiendo al hospital, tiempo, provincia y exámenes físicos.

RI6. Obtener la cantidad de pacientes que presentaron respuestas ante los tratamientos de los EC del producto hr3 por hospital, provincia, tiempo y respuesta global.

RI7. Obtener la cantidad de pacientes que se le monitorearon los signos vitales durante los EC del producto hr3 atendiendo al hospital, tiempo, provincia, nivel de dosis, signos vitales.

RI8. Obtener la cantidad de pacientes a los que se le realizaron exámenes de laboratorio en los EC del producto hr3 atendiendo al hospital, provincia, tiempo y examen de laboratorio.

2.3.2 Requisitos funcionales

Los RF son capacidades o condiciones que el sistema debe cumplir y su proceso de obtención es importante en el desarrollo del MD, de la calidad de los mismos depende el éxito del producto (24). En el caso de esta investigación solo incluye dos subsistemas (almacenamiento e integración). A continuación se muestran los dos RF definidos:

RF1. Realizar la extracción de los datos.

RF2. Realizar transformación y carga de los datos.

2.3.3 Requisitos no funcionales

Los RNF son las propiedades o cualidades que el producto debe tener, especifican los criterios que pueden usarse para juzgar la operación de un sistema en lugar de sus comportamientos específicos, ya que éstos corresponden a los RF. Por tanto, se refieren a todos los requisitos que ni describen información a guardar, ni funciones a realizar (23). Para el desarrollo del MD se identificaron cuatro RNF. A continuación se muestran los mismos:

Requisitos de restricciones de diseño

RNF1. Utilizar el Sistema Gestor de BD definido durante la investigación. El gestor de BD que se utilizará es PostgreSQL y como interfaz de administración de dicho gestor PgAdmin.

RNF2. 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.

RNF3. Utilizar el DataCleaner 1.5.3 con ayuda del Excel para el perfilado de los datos.

Requisitos de soporte

RNF4. Lograr la homogeneidad de la estructura de los elementos definidos en el almacén.

2.4 Diagrama de caso de uso del sistema

El diagrama de CUS es una representación de todos los actores del mismo, los CU y las relaciones que existen entre ellos (24). A continuación se muestra el diagrama de CUS en el cual de identificaron dos

actores, el actor Analista que inicializa ocho CU de información y el actor Administrador de ETL que inicializa dos CU funcionales (Figura 1).

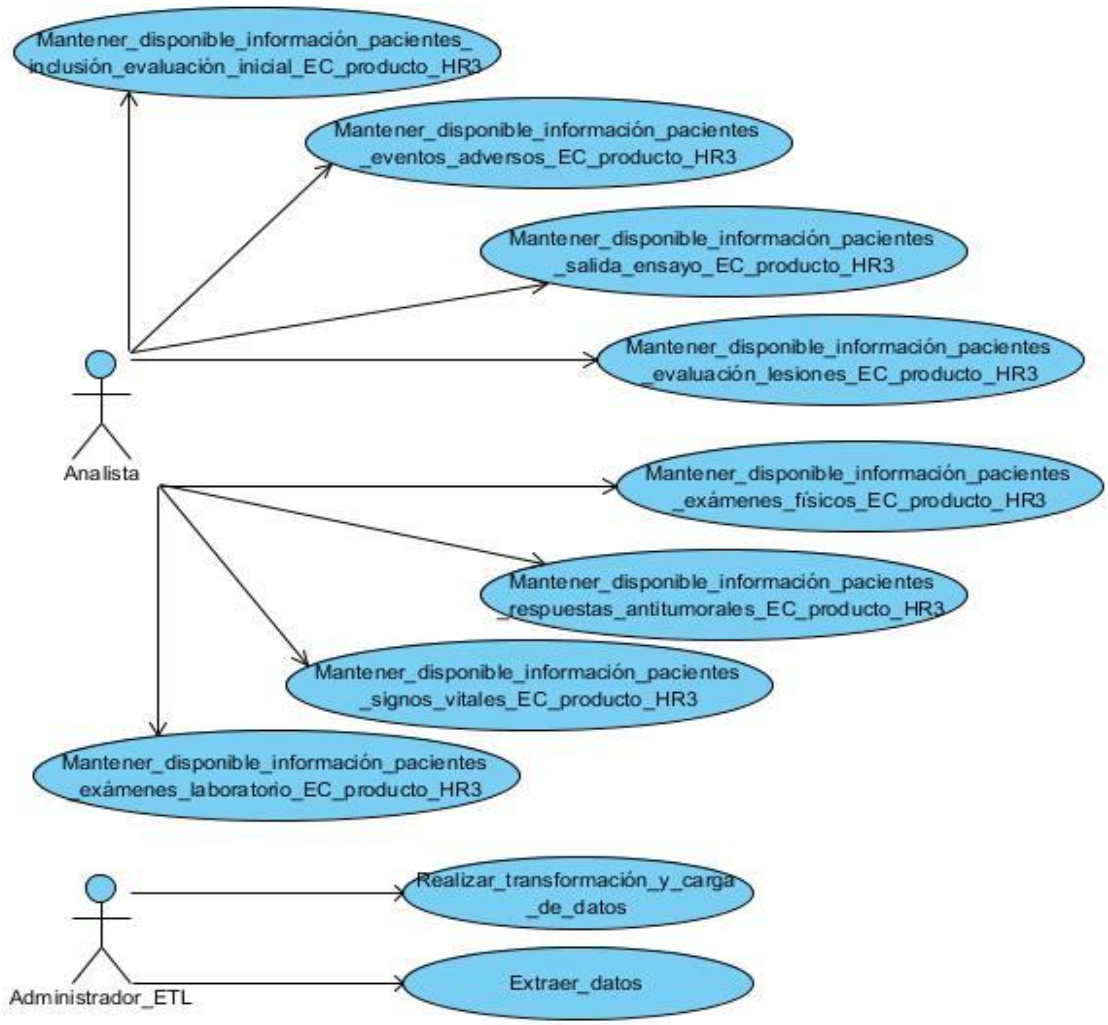


Figura 1. Diagrama de casos de uso del sistema. Relación de los actores con los casos de uso.

2.4.1 Especificación de casos de uso

CUF1. Extraer datos.

Tabla 1. Descripción del Caso de Uso: Extraer datos.

Objetivo	Extraer datos
Actores	Administrador de ETL

Resumen	El CU inicia cuando el actor desea realizar la extracción de los datos correspondientes a las fuentes de información. Se extraen los datos de la fuente. El CU finaliza una vez que los datos seleccionados por el actor son extraídos.	
Complejidad	Media	
Prioridad	Media	
Precondiciones	Disponibilidad de la fuente.	
Postcondiciones	Los datos seleccionados de las fuentes de información quedan extraídos y disponibles para transformar.	
Flujo básico		
	Actor	Sistema
1	Ejecuta la transformación	
2		Realiza la conexión a la fuente de información correspondiente.
3		Chequea la fecha de los datos.
4		Verifica el control de las extracciones.
5		Se realizar la extracción de los datos. Finaliza el CU.
Flujo alterno		
2ª.No responde a la solicitud de conexión		
Actor		Sistema
		Notifica el error al Administrador de ETL a través de un mensaje. Vuelve al paso 1 del flujo normal.
5ª.No se extrajeron los datos		
		Aborta la ejecución del proceso. Finaliza el CU.

Prototipo de interfaz:		
Relaciones	CU Incluidos	No aplica.
	CU Extendidos	No aplica.
Requisitos no funcionales	Sección: "3.2 Requisitos no funcionales" del documento: "0113_Especificación de requisitos de software".	
Asuntos pendientes	[Posibles mejoras al CU.]	

2.5 Definición de la arquitectura base del mercado

El proceso de desarrollo de software requiere la definición previa de una arquitectura. Esta representa la guía para el diseño y construcción del MD, así como marco de la organización para apoyar la integración de las tecnologías.

Subsistema de integración: comprende un conjunto de herramientas para la limpieza, extracción, transformación y carga de los datos hacia el MD.

Subsistema de almacenamiento: se representa a través de una BD relacional que contiene las tablas de dimensiones y hechos cargadas a través de los procesos de ETL. A continuación se muestra la arquitectura de los subsistemas de almacenamiento e integración del MD Nimotuzumab (Figura 2).

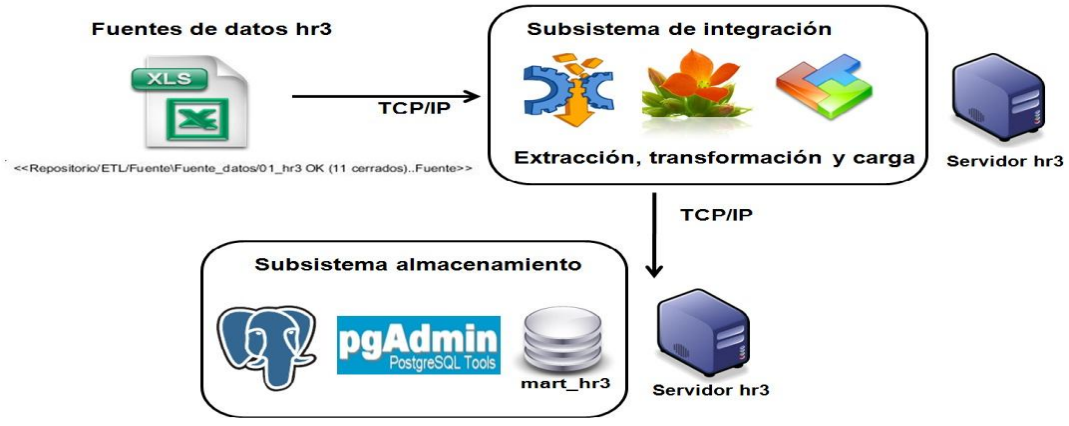


Figura 2. Arquitectura de la solución. Consta de tres niveles: las fuentes de datos y los subsistemas de integración y almacenamiento.

2.6 Diseño del mercado de datos Nimotuzumab

El diseño del MD se realizó con el objetivo de lograr una base y guía para realizar la implementación de los subsistemas de almacenamiento e integración.

2.6.1 Diseño del subsistema de almacenamiento

Para el desarrollo y el correcto funcionamiento del MD, en esta etapa se realiza el modelo dimensional el cual contiene las tablas de hechos identificadas en el negocio, las dimensiones seleccionadas para la solución y las relaciones que existen entre estas.

En el MD Nimotuzumab se identificaron 26 dimensiones a continuación se muestran cada una de ellas y sus descripciones:

Dimensión **dim_sexo**: dimensión que describe el sexo de los pacientes que se encuentran incluidos en cada uno de los EC del CIM.

Dimensión **dim_provincia**: esta dimensión describe la provincia de los pacientes que se encuentran incluidos en cada uno de los EC del CIM.

Dimensión **dim_tiempo**: esta dimensión describe los valores bajo los cuales puede clasificarse la información atendiendo a los años en lo que se almacena información.

Dimensión **dim_peso**: esta dimensión describe el peso de los pacientes que se encuentran incluidos en cada uno de los EC del CIM.

Dimensión **dim_talla**: esta dimensión describe la provincia de los pacientes que se encuentran incluidos en cada uno de los EC del CIM.

Dimensión **dim_raza**: esta dimensión describe la raza de los pacientes que se encuentran incluidos en cada uno de los EC del CIM.

Dimensión **dim_edad**: esta dimensión describe la edad de los pacientes que se encuentran incluidos en cada uno de los EC del CIM.

Dimensión **dim_hospital**: esta dimensión describe el hospital donde se tratan los pacientes que se encuentran incluidos en cada uno de los EC del CIM.

Dimensión **dim_ecog**: esta dimensión describe el grado de Ecog de los pacientes que se encuentran incluidos en cada uno de los EC del CIM.

Dimensión **dim_karnofsky**: esta dimensión describe el grado de Karnofsky de los pacientes que se encuentran incluidos en cada uno de los EC del CIM.

Dimensión **dim_grado_diferenciacion**: esta dimensión describe el grado de diferenciación de los pacientes que se encuentran incluidos en cada uno de los EC del CIM.

Dimensión **dim_respuesta_global**: esta dimensión describe la respuesta clínica de los pacientes que se encuentran incluidos en cada uno de los EC del CIM.

Dimensión **dim_signos_vitales**: esta dimensión describe los signos vitales que presentan los pacientes que se encuentran incluidos en cada uno de los EC del CIM.

Dimensión **dim_tratamientos_previos**: esta dimensión describe los tratamientos previos de los pacientes que se encuentran incluidos en cada uno de los EC del CIM.

Dimensión **dim_nivel_dosis:** esta dimensión describe el nivel de dosis administrada a los pacientes que se encuentran incluidos en cada uno de los EC del CIM.

Dimensión **dim_grupo_tratamiento**: esta dimensión describe el grupo de tratamiento al que pertenecen los pacientes que se encuentran incluidos en cada uno de los EC del CIM.

Dimensión **dim_examen_fisico:** esta dimensión describe los exámenes físicos realizados a los pacientes que se encuentran incluidos en cada uno de los EC del CIM.

Dimensión **dim_examen_laboratorio**: esta dimensión describe los exámenes de laboratorio realizados a los pacientes que se encuentran incluidos en cada uno de los EC del CIM.

Dimensión **dim_tipo_evento_adverso**: esta dimensión describe los tipos de eventos adversos que presentan los pacientes que se encuentran incluidos en cada uno de los EC del CIM.

Dimensión **dim_grado_evento_adverso**: esta dimensión describe el grado de los eventos adversos que presentan los pacientes que se encuentran incluidos en cada uno de los EC del CIM.

Dimensión **dim_causalidad_evento_adverso**: esta dimensión describe la causalidad de los eventos adversos que presentan los pacientes que se encuentran incluidos en cada uno de los EC del CIM.

Dimensión **dim_causa_fallecimiento**: esta dimensión describe la causa por la cual fallece los pacientes que se encuentran incluidos en cada uno de los EC del CIM.

Dimensión **dim_causa_salida**: esta dimensión describe la causa por la cual salen de los EC los pacientes.

Dimensión **dim_estadio**: esta dimensión describe el estadio en el que se encuentra el cáncer de los pacientes que se encuentran incluidos en cada uno de los EC del CIM.

Dimensión **dim_localizacion_lesion**: esta dimensión describe la localización de la lesión de los pacientes que se encuentran incluidos en cada uno de los EC del CIM.

Dimensión **dim_diagnostico_lesion:** esta dimensión describe el diagnóstico de la lesión de los pacientes que se encuentran incluidos en cada uno de los EC del CIM.

En el MD Nimotuzumab se identificaron ocho hechos a continuación se muestran cada uno de ellos y sus descripciones:

Hecho **hech_inclusion_evaluacion_inicial:** en este hecho se almacena toda la información relacionada con los datos de los pacientes en la inclusión y evaluación Inicial de los EC del CIM.

Hecho **hech_examenes_fisicos:** en este hecho se almacena toda la información relacionada con los exámenes físicos de los pacientes de los EC del CIM.

Hecho **hech_respuesta_antitumoral:** en este hecho se almacena toda la información relacionada con respuesta clínica que presentan los pacientes de los EC del CIM.

Hecho **hech_examenes_laboratorio:** en este hecho se almacena toda la información relacionada con los exámenes de laboratorio de los pacientes de los EC del CIM.

Hecho **hech_signos_vitales:** en este hecho se almacena toda la información relacionada con los signos vitales que presentan los pacientes de los EC del CIM.

Hecho **hech_evaluacion_lesiones:** en este hecho se almacena toda la información relacionada con las evaluaciones que se le realizan a los pacientes de los EC del CIM.

Hecho **hech_eventos_adversos:** en este hecho se almacena toda la información relacionada con los eventos adversos que presentan los pacientes de los EC del CIM.

Hecho **hech_salida_ensayo:** en este hecho se almacena toda la información relacionada con la salida de los pacientes de los EC del CIM.

Matriz bus

La matriz bus representa la relación existente entre las tablas de dimensiones y las tablas de hechos que con forman el MD Nimotuzumab. Constituye una validación del análisis realizado evitando que exista solapamiento entre los hechos. Las columnas de la matriz representan los hechos identificados en el mercado y las filas las dimensiones utilizadas. Las celdas marcadas con una X indican que la fila de dimensión está relacionada con la columna del hecho. A continuación se muestra la matriz bus realizada (Tabla 2).

Tabla 2. Matriz bus del MD Nimotuzumab.

Dimensiones/Hechos	hech_ inclusion_evaluacion_inicial	hech_examenes_fisicos	hech_respuesta_antitumoral	hech_examenes_laboratorio	hech_signos_vitales	hech_evaluacio_lesiones	hech_eventos_adversos	hech_salida_ensayo
dim_provincia	X	X	X	X	X	X		
dim_tiempo	X	X	X	X	X	X	X	X
dim_peso	X							
dim_talla	X							
dim_raza	X							
dim_edad	X							
dim_hospital	X	X	X	X	X	X	X	X
dim_sexo	X							
dim_ecog	X							
dim_karnosfky	X							
dim_grado_diferenciacion	X							
dim_respuesta_global			X					
dim_signos_vitales					X			
dim_tratamientos_previos	X							
dim_nivel_dosis					X			
dim_grupo_tratamiento								X
dim_examen_fisico		X						
dim_examen_laboratorio				X				
dim_tipo_evento_adverso							X	
dim_grado_evento_adverso							X	
dim_causalidad_evento_adverso							X	
dim_causa_fallecimiento								X
dim_causa_salida								X
dim_estadio	X							
dim_localizacion_lesion						X		
dim_diagnostico_lesion						X		

Modelo de datos

El principal objetivo de realizar el modelo dimensional es para que los datos del negocio queden representados en una estructura lógica, evidenciando las relaciones que existen entre las tablas de hechos y las dimensiones. Este modelo cuenta con ocho tablas de hechos que están relacionados con lo medido en los EC y 26 tablas de dimensiones que recogen la información de los pacientes (Figura 3).

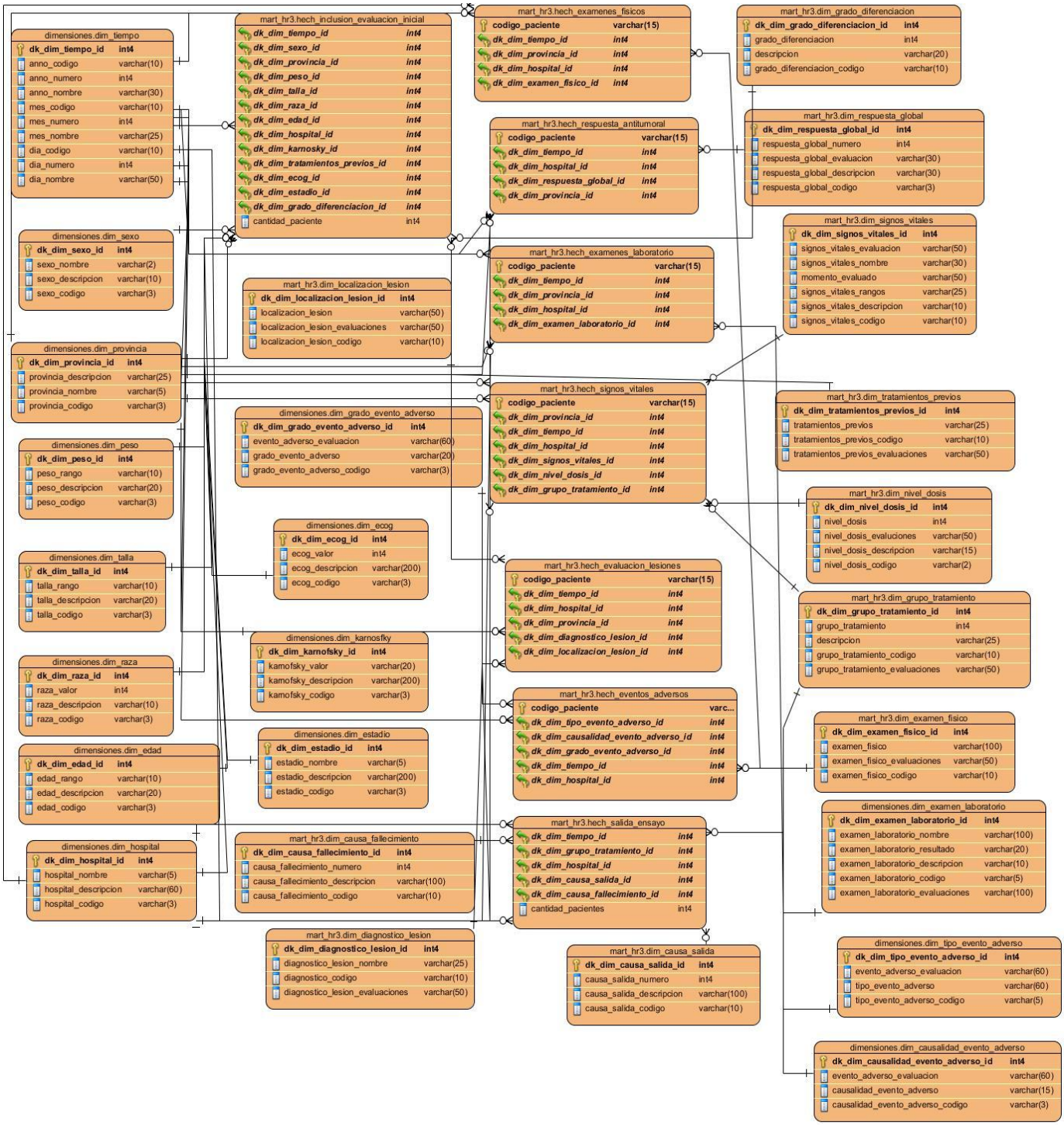


Figura 3. Modelo de datos dimensional. Representa la relación entre los hechos y las dimensiones. Evidencia una topología constelación de hechos.

2.6.2 Diseño del subsistema de integración

Diccionario de datos

El diccionario de datos sirve de apoyo para los procesos de ETL, debido a que este contiene la información necesaria para un correcto entendimiento de los sistemas fuentes, enfocado principalmente en la documentación de las variables de la fuente de datos. En él se describen cada una de estas variables especificando el significado que tienen en el negocio y los posibles valores que pueden tomar. Todo esto es recogido en el artefacto Diccionario de datos que se encuentra en el Expediente de Proyecto.

Perfilado de datos

El perfilado de los datos es el proceso que se encarga de analizar las fuentes de datos con el objetivo de saber el estado en que se encuentran los datos de la fuente, para así poder definir las transformaciones que se le deben realizar a los mismos. Esta información es recogida en el artefacto Perfil de los datos del Expediente de Proyecto. A continuación se muestra el resultado del perfilado de los datos realizado a los ficheros del ensayo clínico 046 y una descripción del mismo.

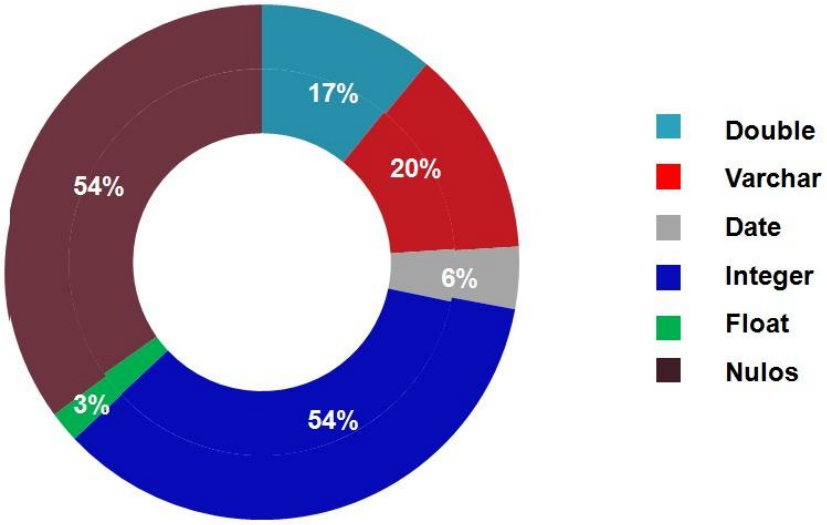


Figura 4. Perfilado del ensayo 046.

La figura 4 muestra el gráfico de los resultados obtenidos en el perfilado de los datos de los ficheros Excel del EC 046. Existen los tipos de datos que se pueden apreciar y su existencia en la fuente expresadas en %. De ese total el 54% independientemente del tipo, tienen valor nulo.

Diseño general de las transformaciones

Las transformaciones en los procesos de ETL son una colección de pasos, que al ser ejecutadas permiten que los datos sean cargados correctamente. Después de haber realizado el perfilado de los datos se definieron las transformaciones que se llevarán a cabo para poblar el MD del producto Nimotuzumab. A continuación se muestra el diseño de las transformaciones y una breve descripción de la misma.

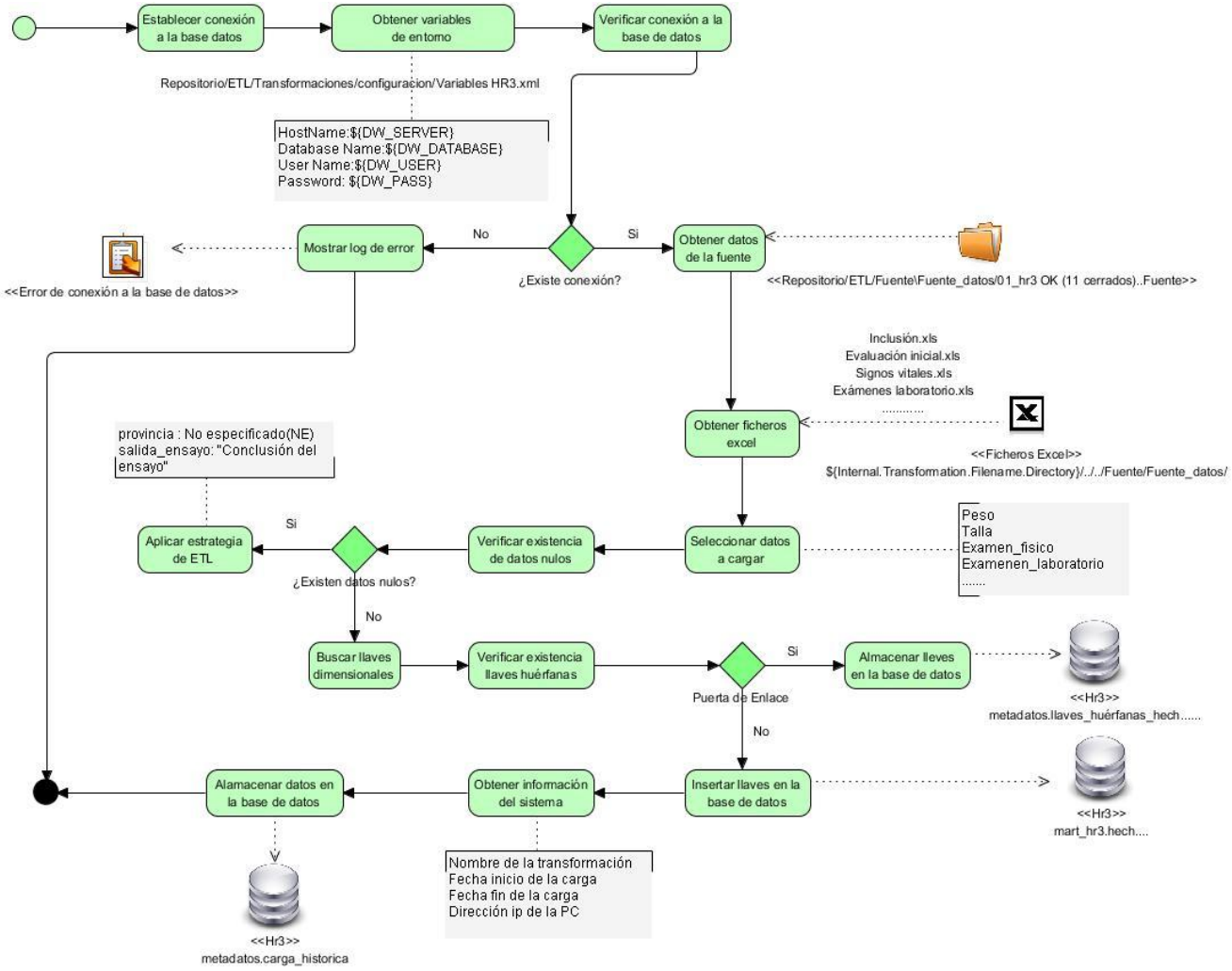


Figura 5. Diseño del subsistema de integración.

La figura muestra el diseño de las ETL realizado. Inicialmente se realiza la conexión a la BD, obteniendo las variables de entorno. Luego se verifica la conexión, en caso que no se logre un resultado exitoso se lanza un mensaje de error y finaliza la transformación. En caso contrario se obtienen y seleccionan los ficheros y datos a cargar. A continuación se verifica la existencia de valores nulos. Si existen se aplica una

estrategia de ETL, sino, se procede a realizar la búsqueda de llaves dimensionales, y se verifica la existencia de llaves huérfanas, si es positiva la respuesta, se insertan los valores en la tabla de metadatos correspondiente, si es negativa se insertar los datos en la tabla de hecho en la BD mart_hr3. Luego se obtienen los datos del sistema y se almacenan los datos de carga histórica.

2.6.3 Política de respaldo y recuperación

En el MD del producto Nimotuzumab de los EC del CIM se utiliza una política de respaldo y recuperación, midiéndose en el siguiente aspecto:

- **Backups existentes:** Actualmente existe un Backup donde se tiene la información relacionada con los datos de los EC del producto Nimotuzumab. Del cual se han realizado copias en varios dispositivos de almacenamiento.

2.6.4 Esquema de seguridad en el subsistema de almacenamiento

El esquema de seguridad en el subsistema de almacenamiento estará respaldado por los niveles de acceso al sistema, específicamente por los roles, regida fundamentalmente por los permisos y roles que los usuarios tienen a la hora de interactuar con la BD. A continuación se muestra como quedó el esquema de seguridad en este subsistema.

Tabla 3. Permisos de los roles en la BD.

Roles	Permisos
Aministrador_ETL	Permisos de lectura y escritura sobre los esquemas pertenecientes a los Subsistemas de almacenamiento e integración Nimotuzumab.

2.6.5 Esquema de seguridad en el subsistema de integración

La seguridad en el subsistema de integración se garantiza a través del Sistema Operativo (SO) que en este caso es GNU/Linux. Dicho SO es el que se encarga de asignar los permisos de acceso a los archivos, para los usuarios que necesiten realizar análisis sobre la información. De esta forma 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

En este capítulo se detallaron los aspectos principales dentro del proceso de análisis y diseño de la solución. Concluyéndose los siguientes resultados:

- Fueron identificadas 15 reglas del negocio, apoyándonos en las características presentadas en las fuentes de datos y en el resultado arrojado por el perfilado de los datos.
- Se definieron ocho requisitos de información, tres requisitos funcionales, cuatro requisitos no funcionales, ocho casos de uso de información y dos casos de uso funcionales.
- La arquitectura base del MD definida, permitió identificar los elementos y subsistemas que están implicados en el desarrollo de la solución.
- Se identificaron en el modelo de datos 26 dimensiones y ocho tablas de hechos, para garantizar el correcto funcionamiento del sistema.
- El perfilado de datos realizado a la fuente de información permitió conocer el estado de los sistemas fuentes, así como el establecimiento de nuevas reglas del negocio aplicables durante el proceso de transformación.
- El diseño del subsistema de integración constituye una aproximación a los pasos que se deben realizar para lograr la estandarización de la información y su almacenamiento.
- Las políticas de respaldo y recuperación establecidas contribuyen a mantener la integridad de los datos almacenados.

CAPÍTULO 3: Implementación y prueba de los subsistemas de almacenamiento e integración del MD Nimotuzumab.

Introducción

En este capítulo se contiene todos los elementos referentes a la implementación, abordando específicamente como se realiza la implementación del subsistema de almacenamiento y del subsistema de integración. También se incluirán los resultados obtenidos en las pruebas realizadas.

3.1 Implementación del subsistema de almacenamiento

En la implementación del subsistema de almacenamiento se definen los estándares de codificación de las estructuras, lo que facilita su comprensión. Se desarrolla además la estructura física del MD.

Estándares de codificación

Con el propósito de lograr un entendimiento entre todas las partes implicadas en un proyecto, son utilizados los estándares de codificación. 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 hechos 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_raza”. En caso de ser una tabla de hechos, como prefijo se ubican las letras “hech”, igualmente separadas del nombre de la tabla de hechos por el carácter “_”, ejemplo “hech_respuesta_antitumoral”.

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

Con este proceso se ha logrado estandarizar la nomenclatura a utilizar para cada una de las tablas (dimensiones y hechos) del negocio, atributos y medidas dentro de la BD.

3.1.1 Implementación del modelo de datos físico

Los esquemas en una BD representan una forma de organizar la información contenida en la misma. Dentro de los esquemas se pueden encontrar funciones, operadores y tipos de datos que facilitarán su implementación. En el presente trabajo se definieron tres esquemas los cuales serán explicados a continuación:

- Esquema “dimensiones” contiene las dimensiones compartidas con los demás mercados del CIM.
- Esquema “mart_hr3” contiene las tablas de dimensiones y tablas de hechos propias de los subsistemas de almacenamiento e integración del producto Nimotuzumab.
- Esquema “metadatos” contienen los metadatos técnicos y de procesos con el objetivo de almacenar las ejecuciones pertenecientes a los trabajos y las transformaciones.

A continuación se muestran los esquemas en la BD, donde el esquema de dimensiones cuenta con 15 tablas las cuales son compartidas en el AD de EC del CIM, el esquema mart_hr3 que contienen 19 tablas propias del mercado en las que ocho son tablas de hechos y 11 son de dimensiones y el esquema metadatos en el cual existen diez de los cuales nueve son metadatos técnicos y uno es de proceso (Figura 6).

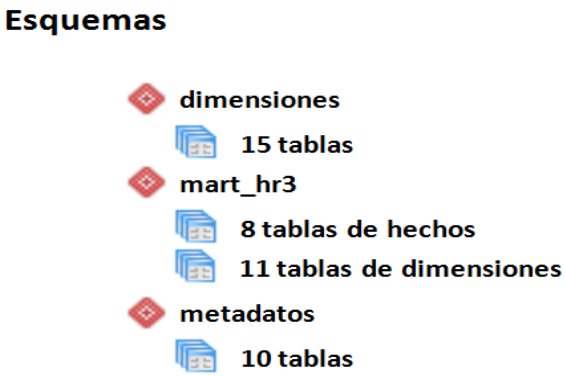


Figura 6. Esquemas de la BD (dimenciones, mart_hr3, metadatos).

3.2 Implementación del subsistema de integración de datos

El proceso de integración de los datos consta de tres etapas fundamentales relacionadas entre sí: extracción, transformación y carga de los datos. Para identificar y corregir los problemas se realiza la limpieza de los datos que permite llenar valores ausentes y corregir errores. Una vez que los datos son transformados se cargan, poblando las dimensiones y los hechos que conforman la estructura del subsistema de almacenamiento del MD Nimotuzumab.

3.2.1 Implementación de las transformaciones

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

En la presente investigación se realizó un flujo de transformación para la carga de cada una de las tablas pertenecientes al esquema mart_hr3. Para las dimensiones, la transformación se realizó a partir de la carga de los indicadores de cada uno de los modelos de la fuente de datos, estos contienen los campos que son necesarios para poblar la BD, para luego pasar a las transformaciones de los hechos.

Implementación de las dimensiones

La siguiente figura es un ejemplo de transformación para la dimensión respuesta global (Figura 7), A continuación se describe la misma: inicialmente se generan los valores que puede tomar la respuesta, así como los momentos en los que pueden presentar los pacientes dicha respuesta. Se añade una constante con valor igual a uno con el objetivo de realizar una unión por clave. Se incluye una secuencia para el código y luego se seleccionan e insertan los datos en el mercado.

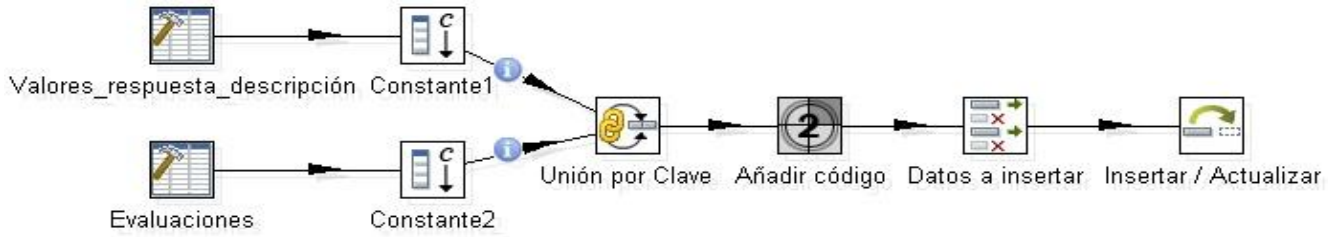


Figura 7. Transformación de la dimensión respuesta global.

Implementación de los hechos

La siguiente figura evidencia los pasos de la transformación del hecho respuesta antitumoral (Figura 8). A continuación se describe la misma: inicialmente se realiza la extracción de los datos. Luego se unen los flujos y se realizan una serie de transformaciones. Seguidamente hace una búsqueda de las llaves dimensionales y se insertan en la BD en caso de que existan, si no, se almacenan las llaves en la tabla hech_respuesta_antitimural. Finalmente se guardan los datos de carga histórica.

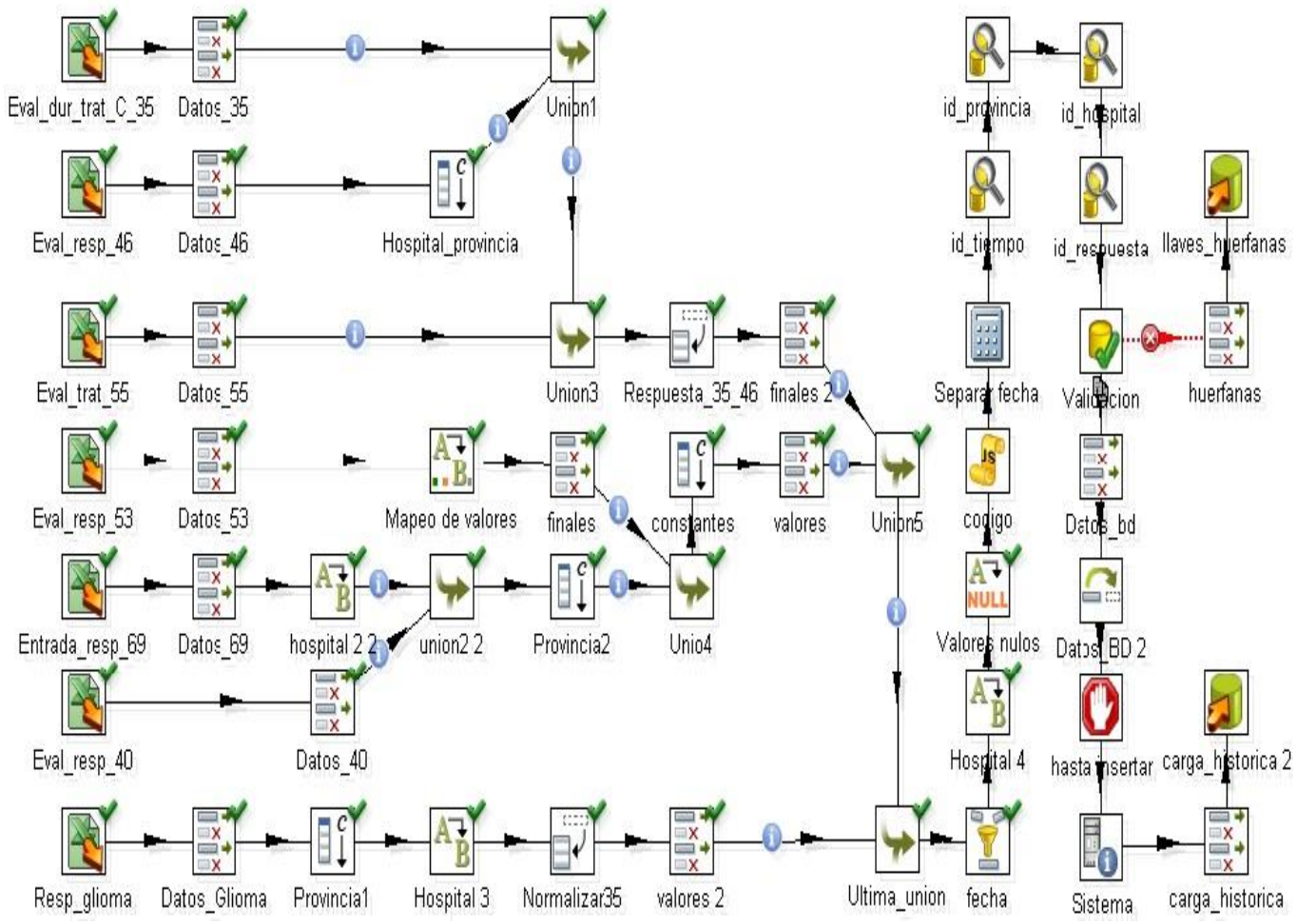


Figura 8. Transformación del hecho respuesta antitumoral.

Implementación de los trabajos

Un trabajo o job es similar a un proceso: conjunto de tareas con el objetivo de realizar una acción determinada. Estos permiten ejecutar varias transformaciones o trabajos previamente diseñados y organizar una secuencia de ejecución de estos. Los trabajos se encuentran en un nivel superior de las transformaciones. A continuación se muestra el ejemplo del trabajo general en el que se inicia la carga (Figura 9). Luego se verifica la conexión, en caso de que exista, se ejecuta la transformación correspondiente a la carga del hech_eventos_adversos. Por último se ejecutan los trabajos que posibilitan la carga de los restantes hechos y sus dimensiones asociadas.

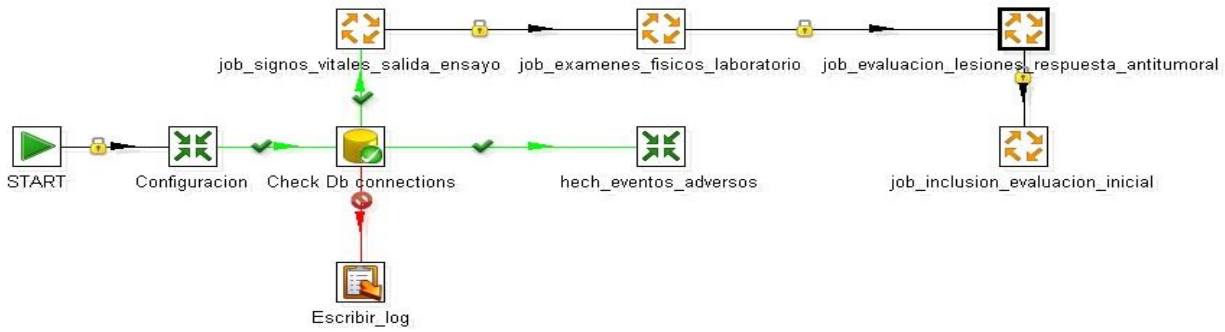


Figura 9. Transformación del trabajo general del mercado de datos Nimotuzumab.

3.2.2 Gestión del cambio en las dimensiones

Las dimensiones lentamente cambiantes son dimensiones en las cuales los datos tienden a modificarse a través del tiempo, ya sea de forma ocasional, constante, que implique a un solo registro o a la tabla completa. Al ocurrir estos cambios, se puede optar por registrar el historial de cambios o reemplazar los valores que sean necesarios (25).

Estrategias a seguir cuando se tratan las dimensiones lentamente cambiantes (25):

SCD Tipo 0: es un enfoque pasivo, es decir no se hace nada al respecto. Los valores permanecen como estaba la dimensión cuando los registros fueron creados.

SCD Tipo 1: es básico y sencillo de implementar. En este caso cuando un registro presente un cambio en alguno de los valores de sus campos, se procede simplemente a actualizar el dato en cuestión, sobrescribiendo el antiguo. Usualmente este tipo es utilizado en casos en donde la información histórica no sea importante de mantener, tal como sucede, por ejemplo, cuando se debe modificar el valor de un registro porque tiene errores ortográficos.

SCD Tipo 2: en este método se inserta un nuevo registro cada vez que existe un cambio en la dimensión. Se agrega un campo de versión u opcionalmente se agregan dos columnas para capturar la fecha de inicio y final de ese valor. Con este método se puede relacionar fácilmente el período de tiempo para el cual es válido cierto dato en la dimensión. Esta técnica permite guardar ilimitada información de cambios.

SCD Tipo 3: este método da seguimiento al cambio agregando nuevas columnas. Una columna mantendría el dato actual, y otra el dato nuevo por el que se quiere cambiar el actual, así como una columna de fecha efectiva del cambio. Este enfoque sólo puede mantener un cambio histórico, a diferencia del Tipo 2 que puede mantener cambios ilimitados en la historia.

SCD Tipo 4: este método mantiene una tabla histórica para todos los cambios y una tabla con el valor actual de la dimensión. Esta tabla histórica indicará, por ejemplo, que tipo de operación se ha realizado (Insertar, Modificar, Eliminar), sobre qué campo y en qué fecha. El objetivo de mantener esta tabla es el de contar con un detalle de todos los cambios, para luego analizarlos y poder tomar decisiones acerca de cuál técnica SCD podría aplicarse mejor.
SCD Tipo 6: este método, conocido también como Híbrido, es una combinación de los Tipos 1,2 y 3 (1 + 2 + 3 = 6). El enfoque es usar una Dimensión Tipo 1 (escribiendo el dato actual), pero agregar un par adicional de columnas con las fechas de validez (Tipo 2).

En el presente Trabajo de diploma se utilizó como estrategia a seguir tipo 0 debido a que la carga que se realiza no es incremental ya que los EC del producto Nimotuzumab están cerrados lo que quiere decir que no se crearán más valores de los que ya se tienen en la fuente de datos.

3.2.3 Gestión de los metadatos

Los metadatos son datos sobre los datos, o información descriptiva sobre los mismos y otras estructuras, como objetos, reglas de negocio y procesos que manipulan los datos. Esta información puede ser sobre cuando se creó un archivo, quién lo creó, cuando fue actualizado la última vez, su tamaño y su extensión, entre otros. Dado que los metadatos han sido utilizados en varios campos, existen modelos especializados y aceptados en su agrupación para especificar los tipos de metadatos.
Según Ralph Kimball, los metadatos se pueden dividir en tres categorías:

- Metadatos técnicos: se usan a menudo por un personal más técnico, tal como los desarrolladores. Incluye temas como las definiciones de tablas y tipos de datos. Estos objetos son utilizados frecuentemente durante el diseño de la aplicación y el proceso de desarrollo.
- Metadatos del negocio: permiten definir los términos en el lenguaje cotidiano, sin reparos a la implementación técnica.
- Metadatos de proceso: se refieren a los metadatos generados y capturados cuando se ejecuta un proceso. Permite que los administradores gestionen su sistema y aseguran que los procesos funcionen sin problemas. Si hay un problema con alguno de ellos, los metadatos operacionales también ayudan a los administradores a identificar y localizar los problemas.

En la presente investigación se utilizaron los metadatos técnicos y de procesos permitiendo así mantener una información sobre los datos u otras estructuras.

3.3 Pruebas aplicadas

La creación de un software es un proceso complicado que puede o no tener errores. Debido a esto es necesario que su desarrollo esté acompañado de una actividad que permita identificar posibles fallos en la implementación, calidad o usabilidad del mismo, para que una vez probado el sistema, el mismo se despliegue con la calidad requerida.

Para poner en marcha la calidad de la solución se utilizó el Modelo V, el mismo fue definido por CALISOFT y se utiliza en DATEC con el fin de crear un estándar de comprobación para que los productos cumplan con las especificaciones del negocio y así garantizar la calidad del producto final. A continuación se muestra una representación gráfica del ciclo de vida del software propuesto en el modelo V donde se puede observar la punta de la V que es la codificación es decir la implementación de los subsistemas de almacenamiento e integración del MD Nimotuzumab, a la izquierda del mismo se puede detallar las etapas de desarrollo del software y a la derecha de este las pruebas correspondientes a cada etapa. (Figura 10).

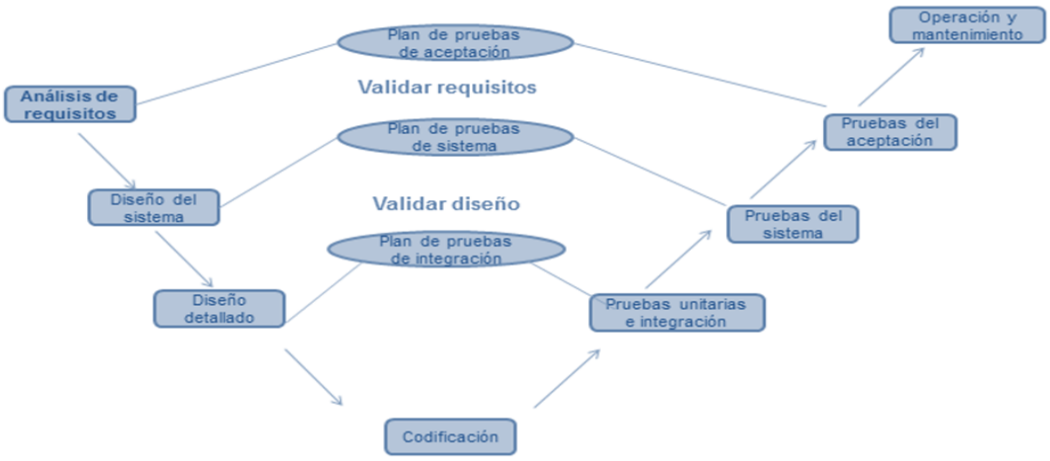


Figura 10. Modelo V. Permite realizar pruebas a lo largo del ciclo de desarrollo del software.

3.3.1 Pruebas unitarias

Este tipo de prueba fueron las primeras aplicadas y se realizan sobre cada uno de los módulos del sistema de manera independiente. Su objetivo fue comprobar cada uno de los módulos por separados, para confirmar su correcto funcionamiento (26).

3.3.2 Pruebas de integración

Este tipo de prueba fue la segunda en llevarse a cabo, se realizó la integración de cada uno de los módulos y componentes del sistema, para verificar su correcto funcionamiento. Ya que por separados los módulos del sistema pueden funcionar correctamente pero es necesario probarlos conjuntamente ya que al ser integrados, un subsistema puede tener un efecto adverso o no deseado sobre el otro (26).

3.3.3 Pruebas de rendimiento

Son las pruebas que se realizan, para determinar lo rápido que realizan las consultas en la BD. También puede servir para validar y verificar otros atributos de la calidad del sistema, tales como la escalabilidad, fiabilidad y uso de los recursos (26).

3.3.4 Herramientas de validación

Dentro de las herramientas utilizadas para aplicar los distintos tipos de pruebas a los subsistemas de almacenamiento e integración del MD Nimotuzumab se realizaron los casos de prueba y las listas de chequeo.

Casos de prueba

Los casos de prueba permiten la verificación de la calidad del software, son usados para la identificación de posibles fallos de implementación y para la comprobación del grado de cumplimiento de las especificaciones iniciales del sistema. En los subsistemas de almacenamiento e integración del MD Nimotuzumab se aplicaron ocho casos de prueba, basados en caso de uso, los cuales responden a los CUI en los que se encuentran agrupados los RI del MD.

Listas de chequeo

En la UCI se ha establecido por CALISOF estándares para definir un modelo propio de aseguramiento de la calidad de software. Entre los mecanismos más usados en la evaluación y control de los productos que se desarrollan en la universidad son las listas de chequeo (ver Anexo 1). Estas son un conjunto o listado de preguntas sobre un aspecto determinado, en forma de cuestionario, que sirven para verificar el grado de cumplimiento de determinadas reglas. Cada una de estas preguntas tiene asociada diversos aspectos los cuales posibilitan determinar el grado de cumplimiento y disponibilidad del indicador evaluado.
La lista de chequeo se divide en tres secciones fundamentales:

Estructura del documento: abarca todos los aspectos definidos por el expediente de proyecto o el formato establecido por el proyecto.

Indicadores definidos: abarca todos los indicadores a evaluar durante la etapa de desarrollo del mercado.

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

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

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

Indicadores a evaluar: son los indicadores a evaluar en las secciones Estructura del documento, Semántica del documento e Indicadores definidos por la etapa.

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

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

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

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

3.4 Resultados de las pruebas

Aplicadas las pruebas descritas anteriormente a los subsistemas de almacenamiento e integración del MD Nimotuzumab, se arrojaron los siguientes resultados.

Pruebas unitarias

Las pruebas unitarias se aplicaron y fueron identificadas tres no conformidades (NC) en el subsistema de almacenamiento, siete NC en el subsistema de integración y dos NC encontradas en las listas de chequeo (Figura 11) las mismas son detalladas a continuación.

NC identificadas en el subsistema de almacenamiento:

NC1. Realizar control de versiones en todo el expediente de proyecto.

NC2. Definir los niveles de acceso para los pedidos de información.

NC3. Los casos de uso no están de acorde con la representación del modelo.

NC identificadas en el subsistema de integración:

NC4. Incluir variables que faltan y se identifican en el negocio en el diccionario de datos.

NC5. Describir mejor los resultados del perfilado que reflejen la realidad del estado de las fuentes de datos.

NC6. Establecer correctamente la relación entre las llaves dimensionales y las secuencias en BD que es realmente su fuente de datos en el mapa lógico.

NC7. Incluir los tipos de SCD que si le aplicará a cada dimensión en el mapa lógico.

NC8. Realizar correcto diseño de ETL.

NC9. Optimizar la utilización de componentes en la implementación de las transformaciones.

NC10. Utilizar la normalización y la desnormalización para resolver los problemas que se presentan en el negocio.

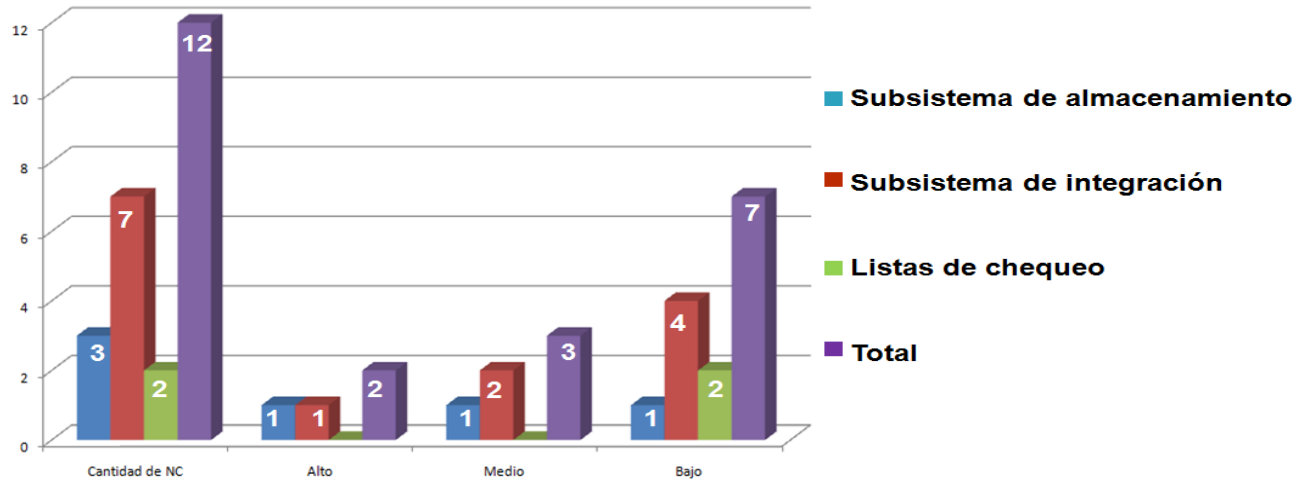


Figura 11.Resultado general de las pruebas unitarias realizadas**.**

Pruebas de integración

En este tipo de prueba se realizaron consultas a la BD, obteniéndose resultados satisfactorios en cada una de estas. A continuación se muestra un ejemplo de una de las consultas realizadas a la BD (figura 12).

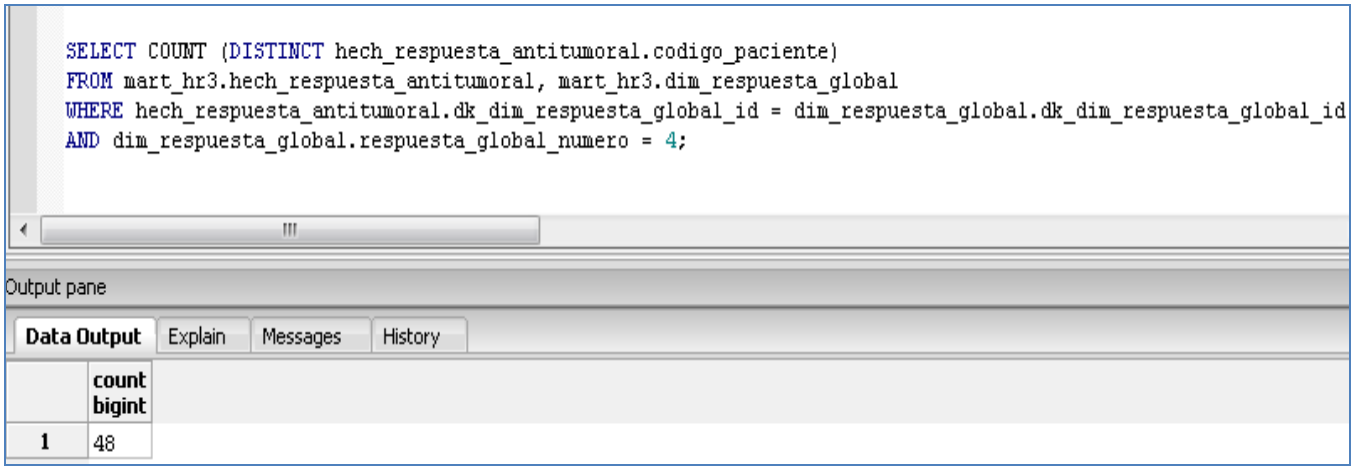


Figura 12. Consulta realizada a la BD. La que devuelve la cantidad de pacientes que presentan respuesta global igual a cuatro.

Calidad de datos

La calidad de los datos es un aspecto importante a tener en cuenta para el desarrollo del MD, debido a que evita que la información almacenada posea errores. Por tanto en la presente investigación, después de haber realizado el proceso de integración de datos, se realizó el Perfilado de los datos a los valores que fueron cargados, con el objetivo de verificar que éstos tuviesen la calidad requerida. El proceso de perfilado permite obtener, estadísticas e información sobre los datos, que posibilitan corregir problemas tales como: valores escritos incorrectamente, duplicados o ausentes.

Los resultados arrojados por este proceso indican que la carga de los datos correspondientes a cada uno de los hechos se realizó correctamente. No fueron almacenados valores vacíos ni nulos. A continuación se muestra los resultados obtenidos al realizar el perfilado de los datos al hecho hech_respuesta_antitumoral en el que se obtuvo un total de 423 tuplas, cinco campos y ningún valor nulo ni vacío, por lo que los resultados fueron satisfactorios (Figura 13).

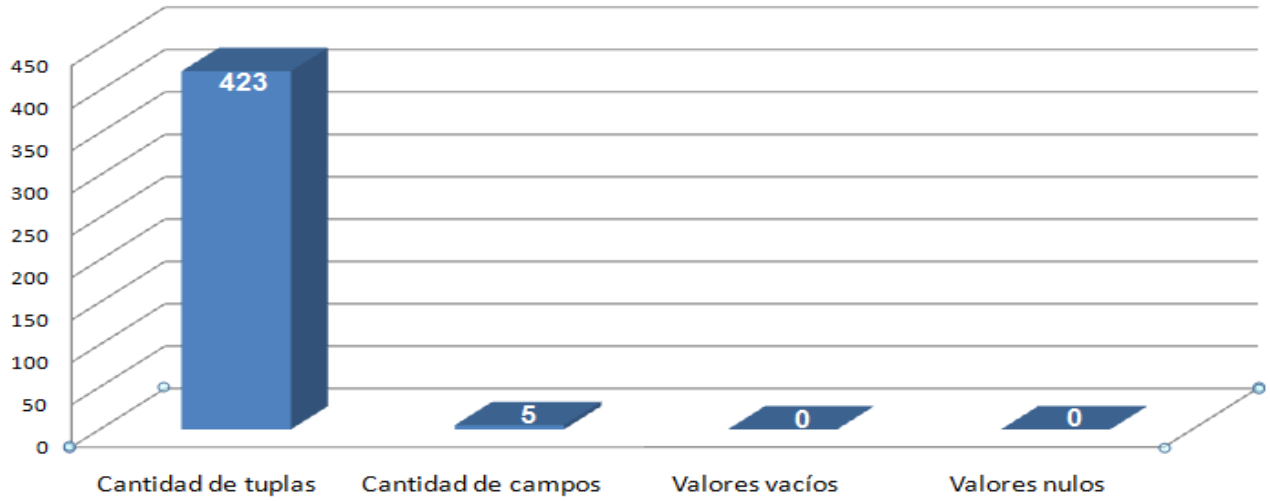


Figura 13. Resultado del perfilado de los datos al hecho hech_respuesta_antitumoral.

Conclusiones

En este presente capítulo se abordó sobre la implementación y validación de los subsistemas de almacenamiento e integración del MD Nimotuzumab arribándose a las siguientes conclusiones:

- Fueron implementados los dos subsistemas que componen la solución almacenamiento e integración, obteniendo como resultado la disponibilidad de la información.
- Se definieron los estándares de codificación con el fin de lograr un entendimiento entre todas las partes implicadas en un proyecto, además de los tres esquemas de dimensiones compartidas, el almacén central y los metadatos.
- Se implementó como estrategia de las dimensiones lentamente cambiante el tipo 0.
- Se obtuvieron resultados satisfactorios al aplicarse las pruebas a los subsistemas de almacenamiento e integración del MD Nimotuzumab del AD de los EC del CIM.

Conclusiones generales

Luego del desarrollo del presente trabajo de investigación de detecto que en el CIM, específicamente en el área de EC se encontraron grandes dificultades a la hora de gestionar la información y tomar decisiones. Para solucionar este problema se desarrollaron los subsistemas de almacenamiento e integración del producto Nimotuzumab en el que se arribaron a las siguientes conclusiones:

- La metodología seleccionada guió el proceso de desarrollo del MD a través de cada etapa del ciclo de vida. Las herramientas y tecnologías seleccionadas dieron soporte a las necesidades del equipo de desarrollo.
- El análisis y diseño de los subsistemas de almacenamiento e integración generó los artefactos necesarios para la posterior etapa de implementación.
- La implementación de los subsistemas de almacenamiento e integración permitió la integración de los datos históricos y su almacenamiento.
- Las pruebas efectuadas durante las distintas etapas de desarrollo permitieron comprobar la funcionalidad del sistema a partir de los requisitos establecidos. Los resultados obtenidos durante las últimas pruebas realizadas fueron satisfactorios, validando el cumplimiento de los objetivos propuestos.

Recomendaciones

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

- Aplicar al MD Nimotuzumab, alguna de las técnicas de minerías de datos con el objetivo de facilitar a los especialistas del CIM búsquedas rápidas de la información.

Referencias bibliográficas

1. Centro de Inmunología Molecular. [En línea] [Consultado el: 1 de noviembre de 2012.] Disponible en: http://www.cim.co.cu/productos_comerciales.php.
2. ¿Qué es un ensayo clínico? [En línea] [Consultado el: 1 de noviembre de 2012.] Disponible en: http://www.who.int/ictrp/es/.
3. ABAD SANTOS, FRANCISCO, MARTÍNEZ SANCHO, ESTHER y GÁLVEZ MÚGICA, MARÍAANGELES.LAS DISTINTAS FASES DEL ENSAYO CLINICO.
4. Nimotuzumab y sus aplicaciones. [En línea] [Consultado el: 2 de noviembre de 2012.] Disponible en: http://www.cimab-sa.com/index.php?action=producto&id=1.
5. Inmon., W. H.Building the Data Warehouse, 1992.
6. Ross, Kimball Ralph y Margy.The Data Warehouse Toolkit. New York : Wiley Computer Publishing Second Edition, 2002.
7. Almacenes de datos. [En línea] [Consultado el: 6 de noviembre de 2012.] Disponible en: http://www.ecured.cu/index.php/Almac%C3%A9n_de_Datos.
8. Kinnear, Thomas C y Taylor, James Ronald. Investigación de mercados: en enfoque aplicado.
9. Wolff, Carmen Gloria. Modelamiento multidimensional.
10. Dataprix. Tipología de estrella. [En línea] [Consultado el: 3 de diciembre de 2012.] Disponible en: http://www.dataprix.com/data-warehousing-y-metodologia-hefesto/arquitectura-del-data-warehouse/34-datawarehouse-manager#x1-500003.4.5.1.
11. Dataprix. Tipología de estrella. [En línea] [Consultado el: 3 de diciembre de 2012.] Disponible en: http://www.dataprix.com/data-warehousing-y-metodologia-hefesto/arquitectura-del-data-warehouse/34-datawarehouse-manager#x1-510003.4.5.2.
12. Dataprix. Esquema constelación de hechos. [En línea] [Consultado el: 3 de diciembre de 2012.] Disponible en: http://www.dataprix.com/data-warehousing-y-metodologia-hefesto/arquitectura-del-data-warehouse/34-datawarehouse-manager#x1-520003.4.5.3.
13. Dataprix. Tipos de almacenamientos. [En línea] [Consultado el: 4 de diciembre de 2012.] Disponible en: http://www.dataprix.com/data-warehousing-y-metodologia-hefesto/arquitectura-del-data-warehouse/34-datawarehouse-manager#x1-550003.4.7.1.
14. Haas, Laura.The theory and practice of information integration.

15. Group, IBM Software.EII -ETL -EAI.What, Why, and How! s.l : Comparación de las técnicas de integracion de datos.
16. González Hernández, Yanisbel. PROPUESTA DE METODOLOGIA PARA EL DASARROLLO DE ALMACENES DE DATOS EN DATEC. Ciudad de La Habana Septiembre, 2011.: Universidad de las Ciencias Informática.
17. PostgresSQL. [En línea] [Consultado el: 28 de noviembre de 2012.] Disponible en: http://www.postgresql.org.es/sobre_postgresql.
18. Pgadmin. [En línea] [Consultado el: 28 de noviembre de 2012.] Disponible en: http://www.pgadmin.org/.
19. Freedownloadmanager. [En línea] [Consultado el: 28 de noviembre de 2012.] Disponible en: http://www.freedownloadmanager.org/es/downloads/Paradigma_Visual_para_UML_%5Bcuenta_de_Plataforma_de_Java_14715_p.
20. DataCleaner. [En línea] [Consultado el: 28 de noviembre de 2012.] Disponible en: http://datacleaner.eobjects.org/.
21. .Pentaho Data Integration. [En línea] [Consultado el: 28 de noviembre de 2012.] Disponible en: http://www.pentaho.com/index.htlm.
22. Pressman, Roger S. Ingeniería de Software.
23. Pressman, Roger S. Ingeniería del software: un enfoque práctico MacGraw-Hill.
24. Addison Wesley Longman, Upper Saddle River. Un acercamiento a través de los casos de uso. 1992.
25. Diaz, Josep Curto; Introducción al business intelligence (2012). Barcelona: UOC 2010.
26. Ricardo Mansilla. Control de calidad de software. Pruebas de software. Nov,2009.http://www.slideshare.net/cliceduca/pruebas-de-software-2420588.

Bibliografía

1. Almacenes de datos. [En línea] [Consultado el: 6 de noviembre de 2012.] Disponible en: http://www.ecured.cu/index.php/Almac%C3%A9n_de_Datos.
2. Bernabeu, Dario. Data Warehouse Manager. [En línea] [Consultado el: 28 de noviembre de 2012.] Disponible en: http://www.dataprix.com/data-warehousing-y-metodologia-hefesto/arquitectura-del-data-warehouse.
3. CLEANER, D. DATA CLEANER [Consultado el: 15 de noviembre de 2012.] Disponible en: http://datacleaner.eobjects.org/.
4. DATACLEANER. DataCleaner. nº [Consultado el: 28 de noviembre de 2012.] Disponible en: http://datacleaner.eobjects.org.
5. Díaz Morales, Themis Patricia y Bermúdez Rodríguez, José Salvador. Diseño de un Datawarehouse para los EC que se gestionan en el Centro de Inmunología. La Habana. 2010.
6. Escobar Domínguez, René Rayde. Tesis de Diploma: Análisis, Diseño e Implementación del Almacén de Datos para el Control del Consumo Energético en la Oficina Nacional de estadística. La Habana: Universidad de las Ciencias Informáticas.
7. FERNÁNDEZ, L. P. y HERNÁNDEZ, A. R. Diseño de un Almacén de Datos para apoyar la toma de decisiones en el Centro de Informática Médica. Universidad de las Ciencias Informáticas, 2010.
8. García Molina, Jesús. M. José Ortín. Moros,Begoña, Joaquín Nicolás. Toval, Ambrosio. De los Procesos Del Negocio a los Casos de Uso. Grupo de Investigación de Ingeniería del Software2. Departamento de Informática y Sistemas. Facultad de Informática. Universidad de Murcia http://www.willydev.net/descargas/willydev_modeladodenegocio.pdf.
9. GEIGER, C.; GALEMMO, N., et al. Mastering Data Warehouse Design. Relational and Dimentional Techniques.
10. González Hernández, Yanisbel. PROPUESTA DE METODOLOGIA PARA EL DASARROLLO DE ALMACENES DE DATOS EN DATEC. Ciudad de La Habana Septiembre, 2011.: Universidad de las Ciencias Informática.
11. Group, IBM Software.EII -ETL -EAI.What, Why, and How! s.l : Comparación de las técnicas de integracion de datos.
12. Haas, Laura.The theory and practice of information integration.
13. Inmon, Willlian H. Building the Data Warehouse, Technical Publish Group. 1992.

14. JACOBSON, Ivar; BOOCH, Grady; RUMBAUGH, James.“El Proceso Unificado de Desarrollo de Software”.2000. Addison Wesley. Capítulos 7, 8 páginas 125-163, 187-202.28.
15. Kimball, Ralph y Kimball, Margy Ross. The Data Warehouse Toolkit: the Complete Guide to Dimensional Modeling. New York: John Wiley&Sons, 2002.
16. Kinnear, Thomas C y Taylor, James Ronald. Investigación de mercados: en enfoque aplicado.
17. Limia Navarro, Alberto, y otros. Metodología para el desarrollo de soluciones de almacenes de datos e inteligencia de negocios en DATEC. La Habana: Universidad de las Ciencias Informáticas, 2011.
18. Luviano Nava, JesusCristhian. Reporte de Investigación. 2011, 5.2 Herramientas Case. 0-807-0027. 2011.
19. MARTINEZ., R. Lanzamiento de PostgreSQL 9.1 [Consultado el: 25 de noviembre de 2012]. Disponible en: http://www.postgresql.org.es.
20. Mendoza Pacheco, Henry Jesus. ¿Qué es un Sistema Gestor de Bases de Datos o SGBD? [En línea] 2010. [Consultado el: 20 de diciembre de 2012.]. Disponible en: http://www.monografias.com/trabajos56/sistemas-bases-de-datos/sistemas-bases-de-datos.shtml.
21. MESA, V. B. Extracción, transformación y carga del mercado de datos Racotumumab para el almacén de datos del Centro de Inmunología Molecular. Universidad del as Ciencias Informáticas, 2011.
22. MILANÉS, Y. R. y JUAN, Y. I. S. Mercado de Datos Agricultura, ganadería y silvicultura para el Sistema de Información de Gobierno. Universidad de las Ciencias Informáticas, 2011.
-paradigm.com/product.
23. Reingart, Mariano. ArPUG, Grupo de usuarios PostgreSQL de Argentina. [En línea] [Consultado el: 11 de Noviembre de 2012.] Disponible en: http://www.arpug.com.ar/trac/wiki/PgAdmin.
24. Ricardo Mansilla. Control de calidad de software. Pruebas de software. Nov,2009.http://www.slideshare.net/cliceduca/pruebas-de-software-2420588.
25. Ricardo Mansilla. Control de calidad de software. Pruebas de software. Nov, 2009. http://www.slideshare.net/cliceduca/pruebas-de-software- 2420588.
26. Sherman Wood, JasperSoft. Pentaho. [En línea] Abril de 2007. [Citado el: 15 de Noviembre de 2012.] http://mondrian.pentaho.com/documentation/workbench.php.
27. VISUAL-PARADING.COM. Visual Parading para UML [Consultado el: 28 de noviembre de 2012.] Disponible en: http://www.google.com/http://www.visual

Anexos

Anexo 1: Listas de chequeo.

Registro sistema fuente.

Estructura del documento					
Peso	**Indicadores a evaluar**	**Eval**	**(NP)**	**Cantidad de elementos afectados**	**Comentarios**
crítico	1. ¿El entregable contiene las secciones obligatorias de la plantilla estándar definida para el expediente de proyecto? (ver expediente de proyecto)	0			
crítico	2. ¿El alcance del proyecto describe correctamente los datos de las dimensiones y hechos del mercado de datos?	0			
crítico	3. ¿El objetivo expresa correctamente el propósito del documento?	0			
	4. ¿Se hace un uso adecuado del control del documento?	0			
	5. ¿En la sección de acrónimos se definen todos los acrónimos utilizados en el documento?	0			
	6. ¿Queda establecida en el entregable el sistema fuente de los datos?	0			

	7. ¿Queda definido en el entregable los datos de contacto dentro de la empresa con la que se trabaja?	0			
	8. ¿Se especifica el sistema gestor en el que se encuentra disponible la fuente de datos?		NP		
	9. ¿Se especifica en el entregable las propiedades de la fuente de datos (Tamaño, cantidad de usuarios, cantidad de transacciones diarias, accesibilidad y disponibilidad)?	0			
Indicadores definidos en el desarrollo					
Peso	**Indicadores a evaluar**	**Eval**	**(NP)**	**Cantidad de elementos afectados**	**Comentarios**
	1. ¿Se utilizó un lenguaje cuyas sentencias son expresables mediante una sintaxis bien definida?	0			
Semántica del documento					
Peso	**Indicadores a evaluar**	**Eval**	**(NP)**	**Cantidad de elementos afectados**	**Comentarios**
crítico	1. ¿Se han identificado errores ortográficos en los entregables?	0			
crítico	2. ¿Se entiende claramente lo que se ha especificado en el documento?	0			
	3. ¿El número de página que aparece en el índice coincide con el contenido que se refleja realmente en dicha página?	0			

Perfilado de datos

Estructura del documento					
Peso	**Indicadores a evaluar**	**Eval**	**(NP)**	**Cantidad de elementos afectados**	**Comentarios**
crítico	1.¿El entregable contiene las secciones obligatorias de la plantilla estándar definida para el expediente de proyecto? (ver expediente de proyecto)	0			
crítico	1. ¿El alcance del proyecto describe correctamente los datos de las dimensiones y hechos del mercado de datos?	0			
crítico	3. ¿El objetivo expresa correctamente el propósito del documento?	0			
	4. ¿Se hace un uso adecuado del control del documento?	0			
	5. ¿En la sección de acrónimos se definen todos los acrónimos utilizados en el documento?	0			
	6. ¿Se identifican correctamente todos los objetos del negocio?	0			
	7. ¿Queda reflejado correctamente en el entregable la definición de todas las tablas de la fuente de datos?	0			
	8. ¿Existe calidad en los datos de la fuente?	0			
Indicadores definidos en el desarrollo					
Peso	**Indicadores a evaluar**	**Eval**	**(NP)**	**Cantidad de elementos afectados**	**Comentarios**

	1. ¿Se utilizó un lenguaje cuyas sentencias son expresables mediante una sintaxis bien definida?	0			
Semántica del documento					
Peso	**Indicadores a evaluar**	**Eval**	**(NP)**	**Cantidad de elementos afectados**	**Comentarios**
crítico	5. ¿Se han identificado errores ortográficos en los entregables?	0		0	
crítico	6. ¿Se entiende claramente lo que se ha especificado en el documento?	0			
	7. ¿El número de página que aparece en el índice coincide con el contenido que se refleja realmente en dicha página?	0			

Mapa lógico de datos

Estructura del documento					
Peso	**Indicadores a evaluar**	**Eval**	**(NP)**	**Cantidad de elementos afectados**	**Comentarios**
crítico	1. ¿El alcance del proyecto describe correctamente los datos de las dimensiones y hechos del mercado de datos?	0			
crítico	2. ¿El objetivo expresa correctamente el propósito del documento?	0			

crítico	3. ¿Existe una adecuada correspondencia entre el origen del los datos y los atributos del mercado?	0			
	4. ¿Se hace un uso adecuado del control del documento?	0			
	5. ¿En la sección de acrónimos se definen todos los acrónimos utilizados en el documento?	1			
Indicadores definidos en el desarrollo					
Peso	**Indicadores a evaluar**	**Eval**	**(NP)**	**Cantidad de elementos afectados**	**Comentarios**
	1. ¿Se utilizó un lenguaje cuyas sentencias son expresables mediante una sintaxis bien definida?	0			
Semántica del documento					
Peso	**Indicadores a evaluar**	**Eval**	**(NP)**	**Cantidad de elementos afectados**	**Comentarios**
crítico	1. ¿Se han identificado errores ortográficos en el documento?	0		0	
crítico	2. ¿Se entiende claramente lo que se ha especificado en el documento?	0			
	3. ¿El número de página que aparece en el índice coincide con el contenido		NP		

	que se refleja realmente en dicha página?				

Diccionario de datos

Estructura del documento					
Peso	**Indicadores a evaluar**	**Eval**	**(NP)**	**Cantidad de elementos afectados**	**Comentarios**
crítico	1¿El entregable contiene las secciones obligatorias de la plantilla estándar definida para el expediente de proyecto?	0			
crítico	2¿El alcance del proyecto describe correctamente los datos de las dimensiones y hechos del mercado de datos?	0			
crítico	3¿El objetivo expresa correctamente el propósito del documento?	0			
	4¿Se hace un uso adecuado del control del documento?	0			
	5¿En la sección de acrónimos se definen todos los acrónimos utilizados en el documento?	0			
	6¿En el entregable, la definición de las variables se hace correctamente?	0			

	7¿Existe una adecuada correspondencia entre las variables definidas y las descripciones que tienen estas variables?	0			
	8¿En el entregable se crea una hoja por cada variable definida?	0			
	9¿Queda registrado en el entregable todos los posibles valores que van a tener las variables definidas?	1			
Indicadores definidos en el desarrollo					
Peso	**Indicadores a evaluar**	**Eval**	**(NP)**	**Cantidad de elementos afectados**	**Comentarios**
	1. ¿Se utilizó un lenguaje cuyas sentencias son expresables mediante una sintaxis bien definida?	0			
Semántica del documento					
Peso	**Indicadores a evaluar**	**Eval**	**(NP)**	**Cantidad de elementos afectados**	**Comentarios**
crítico	1. ¿Se ha identificado errores ortográficos en el entregable?	0			
crítico	2. ¿Se entiende claramente lo que se ha especificado en el documento?	0			

	3. ¿El número de página que aparece en el índice coincide con el contenido que se refleja realmente en dicha página?		NP		

Glosario de términos

AD: Almacén de datos.

Base de datos relacional: es una BD que cumple con el modelo relacional, el cual es el modelo más utilizado en la actualidad para implementar bases de datos ya planificadas. Permiten establecer relaciones entre los datos (que están guardados en tablas), y a través de ellas relacionar los datos de ambas tablas, de ahí proviene el nombre de: "Modelo Relacional".

BD: Bases de Datos.

CIM: Centro de Inmunología Molecular.

CUS: Casos de Uso del Sistema.

DATEC: Centro de Tecnologías de Gestión de Datos.

EC: Ensayos Clínicos.

ETL: Extracción, Transformación y Carga.

MD: Mercado de datos.

OLAP: Procesamiento analítico en línea.

RF: Requisitos Funcionales.

RI: Requisitos de Información.

RN: Reglas del Negocio.

RNF: Requisitos No Funcionales.

SGBD: Sistema Gestor de Base de Datos.

SO: Sistema Operativo.

TCP/IP: son las siglas de Protocolo de Control de Transmisión/Protocolo de Internet (por el inglés Transmission Control Protocol/Internet Protocol), un sistema de protocolos que hacen posibles servicios Telnet, FTP, E-mail, y otros, entre ordenadores que no pertenecen a la misma red. TCP garantiza la entrega de datos y que los paquetes sean entregados en el mismo orden en el cual fueron enviados.

UCI: Universidad de las Ciencias Informáticas.

UML: Lenguaje de modelado visual consistente en el cual se expresan los resultados de numerosas metodologías de orientación a objetos existentes.