Universidad de las Ciencias Informáticas

Facultad 6

Título: Extracción, transformación y carga del mercado de datos CIMAvax EGF para el almacén de datos del Centro de Inmunología Molecular.

Trabajo de Diploma para optar por el título de

Ingeniero en Ciencias Informáticas

Autor: Rolando Pérez Rebollo

Tutoras: Ing. Yisel de Lisy Sánchez Gallardo

MSc. Bárbara Wilkinson Brito

Co-tutores: Ing. Themis Patricia Díaz Morales

Ing. José Salvador Bermúdez Rodríguez

Junio de 2011

Donde yo encuentro poesía mayor es en los libros de ciencia, en la vida del mundo, en el orden del mundo (…) y en la unidad del universo, que encierra tantas cosas diferentes, y es todo uno.

José Martí

Declaración de autoría

Declaro ser autor de la presente tesis y reconozco a la Universidad de las Ciencias Informáticas los derechos patrimoniales de la misma, con carácter exclusivo.

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

Rolando Pérez Rebollo

_____

Firma del Autor

Ing. Yisel de Lisy Sánchez Gallardo

_____

Firma de la Tutora

Ing. Themis Patricia Díaz Morales

_____

Firma de la Co-tutora

MsC. Bárbara Wilkinson Brito

_____

Firma de la Tutora

Ing. José Salvador Bermúdez Rodríguez

_____

Firma del Co-tutor

Tutoras: Tutora: Ing. Yisel de Lisy Sánchez Gallardo

Especialidad de graduación: Ingeniería en Ciencias Informáticas

Años de experiencia en el tema: 0

Años de graduado: 1

Correo Electrónico: Yisels@cim.sld.cu

Tutora: MSc. Bárbara Wilkinson Brito

Especialidad de graduación: Licenciada en Ciencias Farmacéuticas

Categoría docente:

Categoría Científica: Máster

Años de experiencia en el tema: 0

Años de graduado: 10

Correo Electrónico: wilkinson@cim.sld.cu

Cotutores: Cotutora: Ing. Themis Patricia Díaz Morales

Especialidad de graduación: Ingeniería en Ciencias Informáticas

Años de experiencia en el tema: 0

Años de graduado: 1

Correo Electrónico: tpdiaz@uci.cu

Cotutor: Ing. José Salvador Bermúdez Rodríguez

Especialidad de graduación: Ingeniería en Ciencias Informáticas

Categoría docente: Instructor en Adiestramiento

Categoría Científica: Ingeniero

Años de experiencia en el tema: 0

Años de graduado: 1

Correo Electrónico: jsbermudez@uci.cu

-A Fidel Castro y a la Revolución por devolvernos la capacidad de soñar. Por la oportunidad de realizar nuestros sueños.

- A mi eterno compañero de juegos y aventuras, a mi guía e inspiración, a mi héroe: al hombre que me parió.

- A mi mejor amiga y cómplice, mi soporte y refugio seguro, compañera de conciertos en noches de apagón. A mi Mariana Grajales: a la mujer de cuyas entrañas jamás saldré.

Al a los integrantes del tribunal: "las yanelis" y lesley.

A mi oponente Jose Leandro -

A mis tutores y cotutores: Yisel, Baby, Themis y Salvador.

- A todos los profesores que contribuyeron a nuestra formación como ingenieros, especialmente a Pacheco, Norkis, Nara y Lesley.

-A aquellos que me soportan: Guille, Os, Yasel, Albe, Betty, Raido, el Isa, mrrico, "la negra" y la gente del apto.

-A todas las trabajadoras del departamento de Manejo de Datos del Centro de Inmunología Molecular cuyo monumental trabajo hizo más llevadero el nuestro.

- Al Padre, porque alejados de ti nada podemos hacer.

-A los que aman y fundan.

-A los que hacen más y critican menos.

-A los que ya no les alcanza la fe, pero continúan.

-A Panchito.

La presente investigación se desarrolla en el marco de la colaboración existente entre la Universidad de las Ciencias Informáticas y el Centro de Inmunología Molecular. Este último ha conducido, desde 2001, varios Ensayos Clínicos del producto CIMAvax EGF en pacientes con cáncer. Como resultado del presente Trabajo de Diploma, a través de la realización del análisis y diseño y el proceso de extracción, transformación y carga, se obtuvo un mercado de datos poblado con información generada durante estos Ensayos Clínicos que responde a necesidades específicas del Departamento de Manejo de Datos. Dicho mercado de datos constituye un importante paso en la estandarización e integración de los datos de la institución. La metodología empleada en el desarrollo de la investigación tiene en cuenta las peculiaridades de la gestión de la información en el Centro de Inmunología Molecular y de los Ensayos Clínicos en cuestión; la misma ha sido formulada en investigaciones previas. Para la consecución exitosa de la investigación fueron empleadas las siguientes herramientas: Visual Paradigm, PostgreSQL, PgAdmin III y Pentaho Data Integration.

Palabras claves:

Mercado de Datos, Almacén de Datos, Análisis y Diseño, Extracción, Transformación, Carga, Centro de Inmunología Molecular, Ensayos Clínicos.

INTRODUCCIÓN

A pesar de su relativa juventud es la informática una de la las ciencias con mayor perspectiva de desarrollo y campos de aplicación en la era moderna. No podía ser de otra manera en un mundo globalizado donde la información se ha convertido en recurso de primera relevancia.

La necesidad de almacenar la información procedente del universo digital, estimada en 2010, se cifra en 1,2 Zettabyte. Entre los registros más gruesos se encuentran entidades como la Agencia Central de Inteligencia (CIA), YouTube y el Centro Mundial de Datos sobre el Clima, por solo citar algunos.

La actividad científica médica, particularmente la académica e investigativa, genera grandes volúmenes de datos susceptibles de ser digitalizados dadas la evidentes ventajas que ello implica. La lucha contra el cáncer, padecimiento que ha pasado a ser la primera causa de muerte a nivel mundial en 2010, y más específicamente la información generada por el procedimiento previo a la aprobación que sufren los fármacos anti cancerígenos, es un área con perspectivas para el almacenamiento digital.

De acuerdo con la Organización Mundial de la Salud (OMS):" Cáncer es un término genérico para un grupo de más de 100 enfermedades que pueden afectar a cualquier parte del organismo. […] Una de las características que define el cáncer es la generación rápida de células anormales que crecen más allá de sus límites normales y pueden invadir zonas adyacentes del organismo o diseminarse a otros órganos en un proceso que da lugar a la formación de las llamadas metástasis".[1]

En Cuba el cáncer es la segunda causa de muerte desde 1970 y la de mayor repercusión en la esperanza de vida al nacer. El Centro de Inmunología Molecular (CIM) fue fundado en diciembre de 1994 y sus productos son comercializados en alrededor de medio centenar de países. Su principal misión consiste en: "obtener y producir nuevos biofármacos destinados al tratamiento del cáncer y otras enfermedades crónicas no transmisibles e introducirlos en la Salud Pública cubana".[2]

Uno de los sellos distintivos del centro es que en él se lleva a cabo el ciclo completo de un producto, desde la fase de investigación y desarrollo, hasta la producción industrial, control de la calidad y comercialización. Como parte de este proceso realizan Ensayos Clínicos (EC) para el diagnóstico de tumores por imágenes y tratamiento de cáncer de diferentes orígenes.

Los EC no son más que: "Cualquier investigación en sujetos humanos dirigida a descubrir o verificar los efectos clínicos, farmacológicos u otros efectos farmacodinámicos de un producto(s) en investigación y/o a estudiar la absorción, distribución, metabolismo y excreción del producto en investigación y/o a identificar

cualquier reacción adversa al producto(s) en investigación con el objeto de determinar su seguridad y/o eficacia. Los términos ensayo clínico y estudio clínico son sinónimos".[3]

Desde junio de 2010 hasta la fecha el CIM ha registrado más de una docena de nuevos EC.

Entre los biofármacos desarrollados en el CIM se encuentra el denominado comercialmente CIMAvax EGF (EGF), sobre el que se realizaron una serie de EC. Dichos ensayos generaron un volumen considerable de información disponible actualmente en diversos formatos digitales.

La gestión de la información generada durante los EC en el CIM tiene la particularidad de ser regida por el sistema informático Epidata, este genera reportes en diferentes formatos: SPSS, Excel, SASS y Text. Así como también almacena información en otros formatos: .rec, .qes, .eix, .chk, .bak y .not. Es preciso hacer notar que el mismo no interactúa con sistemas gestores de bases de datos.

Otras características importantes son que a este sistema acceden diferentes especialistas y que los datos están dispersos en varios modelos que no presentan igual estructura de diseño. Al no encontrarse integrados, se torna engorroso el proceso para el manejo de la información por parte de los directivos de la institución, lo que dificulta la realización de análisis estadísticos complejos inter e intra-EC; ello aumenta el riesgo de que se pierda información útil con el paso del tiempo y disminuye la efectividad en el tratamiento de los datos.

Todos estos factores se conjugan para que se identifique como **problema de la investigación**: ¿cómo estandarizar los datos del producto CIMAvax EGF que permita almacenarlos de forma homogénea y que facilite su posterior análisis?

La investigación tiene como **objeto de estudio** los mercados de datos, enmarcado en el **campo de acción** el proceso de extracción, transformación y carga del mercado de datos CIMAvax EGF.

El **objetivo general** es desarrollar la extracción, transformación y carga del mercado de datos CIMAvax EGF para el almacén de datos del Centro de Inmunología Molecular, que contribuya al almacenamiento homogéneo de la información y garantice su posterior análisis.

En correspondencia con el objetivo general, se plantean como **objetivos específicos**:

- ✓ Realizar análisis y diseño del mercado de datos CIMAvax EGF.
- ✓ Implementar la extracción, transformación y carga de los datos del mercado de datos CIMAvax EGF.

- ✓ Realizar pruebas al mercado de los datos poblado.

Para el cumplimiento de estos objetivos se realizarán esencialmente las siguientes **tareas investigativas**:

- ✓ Caracterización de las metodologías, herramientas y tecnologías a utilizar en el desarrollo de almacenes de datos para profundizar el nivel de comprensión y dominio sobre las mismas.
- ✓ Identificación de las necesidades del negocio, elemento crucial para la fase de diseño que permite definir las expectativas del usuario respecto a la información en el mercado de datos.
- ✓ Definición del modelo conceptual con el que se visualizan los indicadores y perspectivas a tener en cuenta para la construcción del mercado de datos CIMAvax EGF de acuerdo con la especificación de necesidades.
- ✓ Evaluación del modelo conceptual para el mercado de datos CIMAvax EGF.
- ✓ Realización del modelo lógico para el mercado de datos CIMAvax EGF. Se obtiene una estructura de datos asociada a un Sistema de Gestión de Bases de Datos.
- ✓ Diseño de la arquitectura de integración para el mercado de datos CIMAvax EGF. Se obtiene una primera vista de la estructura interna del almacén de datos y la estrategia arquitectónica.
- ✓ Elaboración del Modelo Físico del s*taging area.* que sirve de soporte al proceso de extracción, transformación y carga del mercado de datos.
- ✓ Elaboración del Modelo Físico para el mercado de datos CIMAvax EGF a partir del cual se obtiene el script de la base de datos para el gestor seleccionado.
- ✓ Extracción de la información deseada a partir de los datos almacenados en fuentes externas.
- ✓ Realización de las mínimas transformaciones sobre los datos para su posterior carga en el s*taging area*.
- ✓ Realización de la carga de datos transformados hacia el s*taging area*.
- ✓ Creación de la base de datos para el mercado de datos CIMAvax EGF.
- ✓ Extracción de la información a partir de los datos almacenados en el *staging area*.
- ✓ Realización de las transformaciones sobre los datos para que puedan ser cargados en el mercado de datos CIMAvax EGF.

- ✓ Realización de la carga de datos transformados hacia el mercado de datos CIMAvax EGF.
- ✓ Aplicación de la listas de chequeo.

El Trabajo de Diploma está estructurado de la siguiente manera: introducción, tres capítulos, conclusiones, recomendaciones, referencias bibliográficas, bibliografía y anexos.

Capítulo 1: Fundamentos teóricos sobre el desarrollo de un mercado de datos

En este capítulo se exponen definiciones y conceptos esenciales sobre los almacenes de datos y los mercados de datos y la gestión de los Ensayos Clínicos en el Centro de Inmunología Molecular. Además se presentan algunas de las metodologías aplicadas en el mundo para la implementación de los mercados de datos. Por último se justifica el empleo en la presente investigación de las diversas herramientas y tecnologías.

Capítulo 2: Análisis y diseño del mercado de datos para los Ensayos Clínicos del producto CIMAvax EGF

En este capítulo se aborda la etapa de análisis y diseño de un mercado de datos para los Ensayos Clínicos del producto CIMAvax EGF, que se gestionan en el Centro de Inmunología Molecular. El principal objetivo del análisis y diseño consiste en definir con exactitud las características del mercado de datos. Se presentan además los ajustes realizados al procedimiento establecido por la institución, necesarios para la adaptación a los requerimientos de la investigación

Capítulo 3: Proceso de integración y pruebas del mercado de datos para los Ensayos Clínicos del producto CIMAvax EGF

Este capítulo aborda la realización del proceso de extracción, transformación y carga del mercado de datos CIMAvax EGF, etapa vital en la construcción de un mercado de datos. Cubre la extracción de datos desde las fuentes; el aseguramiento de su calidad y consistencia; la reagrupación de manera tal que datos de orígenes distintos puedan usarse de manera transparente y la realización de las transformaciones pertinentes para la exitosa carga del mercado de datos así como su evaluación.

Capítulo 1: Fundamentos Teóricos sobre el Desarrollo de un Mercado de Datos

1.1 Introducción

En este capítulo se exponen definiciones y conceptos esenciales sobre los almacenes de datos y los mercados de datos y la gestión de los Ensayos Clínicos en el Centro de Inmunología Molecular. Además se presentan algunas de las metodologías aplicadas en el mundo para la implementación de los mercados de datos. Por último se justifica el empleo en la presente investigación de las diversas herramientas y tecnologías.

1.2 Manejo de datos en los Ensayos Clínicos realizados por el Centro de Inmunología Molecular

Como bien se ha señalado el CIM es responsable no solo del desarrollo y producción sino también de la comercialización de sus productos. La legislación internacional regula la salida o no al mercado de cada medicamento acorde con estándares de eficacia y seguridad. Es aquí donde entran en escena los EC. La adecuada conservación y análisis de los datos registrados de los EC es de vital importancia para la institución por motivos de seguridad y de posibilitar auditorías e inspecciones.

El eslabón primario en este proceso lo constituye el Cuaderno de Recogida de Datos (CRD), que se llena por los especialistas en contacto directo con el paciente. Los CRD pueden contener cientos de variables en dependencia del alcance y envergadura del estudio clínico; desde información referida a mediciones cuantitativas del estado de un paciente (presión sanguínea, hemoglobina), hasta imágenes que caracterizan el tamaño del tumor, así como mensajes intercambiados entre los investigadores que conducen los estudios. Dado su propósito general el CRD no se acoge a modelo estándar alguno. Una vez en el CIM, se procede a digitalizar la información recogida a través del empleo del EpiData: programa de código abierto para la gestión de documentación que genera reportes en los formatos Text, dBase III, Excel, Stata, SPSS y SAS. Debido a la naturaleza y el volumen de los datos monitoreados, y la cantidad de personal involucrado en el proceso, la gestión de la información clínica se convierte en un verdadero problema.

De los EC conducidos por el CIM es objetivo de esta investigación desarrollar el proceso de integración a los relacionados con el producto EGF aplicado en las siguientes localizaciones:

- ✓ Pulmón
- ✓ Próstata

- ✓ Estómago
- ✓ Colon

En correspondencia con las cuales se tienen las siguientes bases de datos:

- ✓ PI tumores sólidos 019
- ✓ PII pulmón 025
- ✓ PIII pulmón 033
- ✓ PIV pulmón 041
- ✓ PV pulmón 062
- ✓ FII pulmón 056
- ✓ FII próstata 077
- ✓ FIII pulmón 081
- ✓ FIII pulmón VQV 111

1.3 Almacén de datos y mercado de datos

1.3.1 Revisión conceptual

Dos de los autores de más prestigio en el tema defienden enfoques distintos con respecto al concepto de almacén de datos.

De acuerdo con Bill Inmon un almacén de datos no es más que: “una colección de datos orientados al tema, integrados, no volátiles e históricos, cuyo objetivo es el de servir de apoyo en el proceso de toma de decisiones gerenciales”.[4]

En cambio Ralph Kimball define un almacén de datos como “una copia de las transacciones de datos específicamente estructurada para la consulta y el análisis”.[5]

Un mercado de datos, por otra parte, es: “un subconjunto del almacén de datos. En su forma más simple, un mercado de datos presenta los datos de un proceso de negocio único. Este proceso se encuentra entre los límites de las funciones de la organización”.[6] Se caracterizan por disponer de la estructura óptima de datos para analizar la información al detalle desde todas las perspectivas que afecten a los procesos correspondientes.

1.3.2 Principales características

Inmon definió las siguientes características:

- ✓ **Temático:** los datos en la base de datos 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 base de datos contiene los datos de todos los sistemas operacionales de la organización, y dichos datos deben ser consistentes.
- ✓ **No volátil:** la información no se modifica ni se elimina, una vez almacenado un dato, éste se convierte en información de sólo lectura, y se mantiene para futuras consultas.
- ✓ **Histórico:** los cambios producidos en los datos a lo largo del tiempo quedan registrados para que los informes que se puedan generar reflejen esas variaciones.

1.3.3 Modelo multidimensional

Debido a su orientación analítica los almacenes de datos imponen un procesamiento y pensamiento distinto, ello se sustenta a través de un modelamiento de bases de datos propio: modelo multidimensional.

Para mejor comprensión del modelo multidimensional es preciso dominar tres conceptos fundamentales: hechos, dimensiones y medidas.

- ✓ **Hecho:** es una operación que se realiza en el negocio en un tiempo determinado y que es objeto de análisis para la toma de decisiones.
- ✓ **Medida:** es una propiedad de un hecho (casi siempre numérica), que es usada para su análisis.
- ✓ **Dimensión:** es una característica de un hecho que permite su análisis posterior, en el proceso de toma de decisiones.

De acuerdo con R. Kimball el modelo multidimensional “contiene la misma información que un modelo normalizado pero empaqueta los datos de acuerdo con un formato cuyas metas de diseño son la comprensibilidad del usuario final, el rendimiento ante las consultas y la resistencia al cambio”.[7]

1.3.4 Comparación entre los sistemas tradicionales de bases de datos y los almacenes de datos

En la **Tabla 1** se pueden observar algunas de las principales diferencias entre los sistemas tradicionales y los almacenes de datos.

SISTEMA TRADICIONAL	ALMACÉN DE DATOS
Predomina la actualización	Predomina la consulta
La actividad más importante es de tipo operativo	La actividad más importante es el análisis y la decisión estratégica
Predomina el proceso puntual	Predomina el proceso
Mayor importancia a la estabilidad	Mayor importancia al dinamismo
Datos en general desagregados	Datos en distintos niveles de de detalle y agregación
Importancia del dato actual	Importancia del dato
Importancia del tiempo de respuesta de la transacción	Importancia de la respuesta masiva
Estructura relacional	Visión multidimensional
Usuarios de perfiles medios o bajos	Usuarios de perfiles altos
Explotación de la información relacionada con la operatividad del negocio	Explotación de toda la información interna y externa relacionada con el negocio

TABLA 1: SISTEMA TRADICIONAL VS ALMACÉN DE DATOS[8]

1.4 Etapa de análisis y diseño

En esta etapa se lleva a cabo el estudio del negocio para definir las necesidades y requerimientos del mismo. También se confecciona el modelo conceptual, y derivado de este, el modelo lógico.

1.4.1 Modelo conceptual

Modelo conceptual: “Modelo visual de un sistema que ilustra las interconexiones de los componentes del modelo”.[9]

Se trata de un esquema de alto nivel de la estructura de los datos de un sistema, ajeno a su posterior implementación en una base de datos. Para la presentación de la información se introducen los conceptos de indicador y perspectiva. Refleja además el nivel de granularidad especificado por los usuarios.

Los **indicadores** son valores numéricos y representan lo que se desea analizar concretamente, por ejemplo: promedios, cantidades, sumatorias y fórmulas. Encuentran su expresión en el modelo dimensional en el concepto de medidas (ver **Epíg. 1.3.3**).

Por su parte, las **perspectivas** se refieren a los objetos a través de los cuales obtienen sentido los indicadores. Una perspectiva bastente común, dada la variabilidad en esta de los almacén de datos es el tiempo. El concepto asociado al de perspectiva en el modelo dimensional es el de dimensión (ver **Epíg. 1.3.3**).

El nivel de detalle con el que se desea almacenar la información se conoce como **granularidad** e influye directamente en el modelo conceptual. El nivel de granularidad (fina, media o gruesa) determina la profundidad del análisis que se podrá realizar sobre el mercado o almacén de datos una vez implementado.

1.4.2 Modelo lógico

Modelo lógico: representa una estructura de datos asociada a un Sistema de Gestión de Bases de Datos (SGBD). Tiene en cuenta los siguientes aspectos para su modelacion: tipología de esquema, tablas de hechos, tablas de dimensiones, uniones, granularidad y jerarquías.

Los tipologías de esquema más comunes son:

- ✓ **En estrella:** “estructura relacional de base de datos en la que los datos son guardados en una única tabla de hechos en el centro del esquema que se relaciona con las tablas de dimensiones”.[10]
- ✓ **En copo de nieve:** “una extensión del esquema en estrella donde una o más dimensiones están definidas por múltiples tablas. En un esquema en estrella únicamente las tablas de dimensiones primarias están unidas a la tabla de hechos. Las tablas de dimensionales adicionales están unidas a las tablas primarias”. [11]
- ✓ **Constelaciones de hechos:** “contiene múltiples tablas de hechos que comparten relación con muchas tablas dimensionales”. [12]

La **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 la granularidad definida en el modelo conceptual.

Las **tablas de dimensiones** contienen el detalle de los valores que se encuentran asociados a la tabla de hechos.

Independientemente del esquema, se pueden realizar las **uniones** necesarias entre tablas de dimensiones y entre tablas de hechos.

Las **jerarquías** son grupos de atributos a nivel de dimensiones que mantienen una prioridad predispuesta.

El ejemplo en la **Fig. 1** muestra como se relacionarían varios de los conceptos abordados anteriormente.

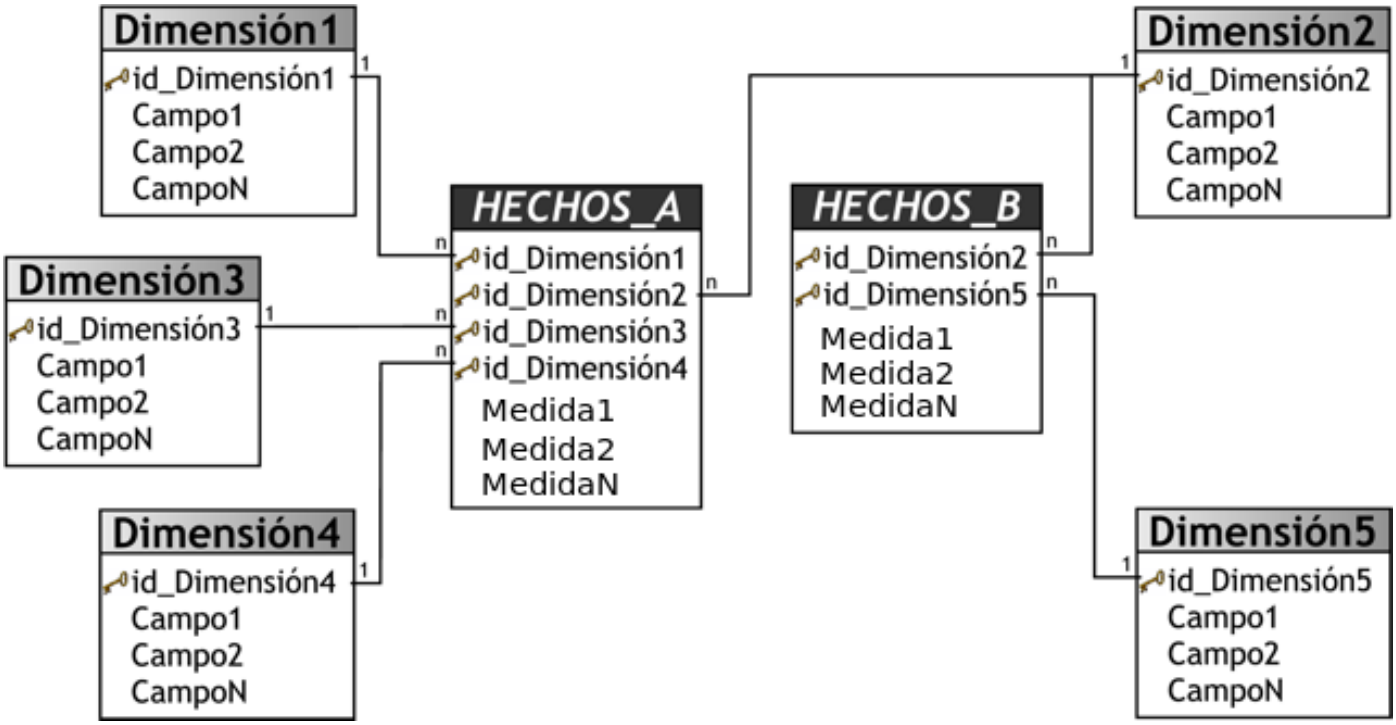

FIG. 1: VISTA TIPOLOGÍA DE ESQUEMA CONSTELACIÓN DE HECHOS

1.4.3 Evaluación del diseño

La mayoría de las inconformidades que hoy se encuentran con respecto al mal funcionamiento o pobre rendimiento de los mercados de datos están directamente ligados a un diseño superficial e ineficiente. El mercado de datos por sí solo no garantiza el éxito de las operaciones, responde a las cualidades y capacidades que su diseño le permitan. De ahí la necesidad de comprobar su idoneidad.

En esta investigación se propondrá, de acuerdo con el estándar establecido en el CIM, una lista de chequeo para la evaluación del diseño del mercado de datos.

1.5 Proceso de extracción, transformación y carga

El proceso de extracción, transformación y carga (ETL por sus siglas en inglés) es el encargado de trasladar los datos desde los sistemas de información transaccionales en línea (OLTP por sus siglas en inglés) hacia el mercado o almacén de datos. Representa un porcentaje de esfuerzo en torno al 80% en el desarrollo de la solución.

- ✓ **Extracción:** Los datos son extraídos desde los diferentes sistemas fuentes y llevados a un formato que haga posible la transformación.
- ✓ **Transformación:** su objetivo es asegurar la consistencia, compatibilidad y congruencia de los datos a cargar. Se realizarán para ello todas las transformaciones necesarias para alcanzar la integración de los datos.
- ✓ **Carga:** La carga de datos persigue la organización y actualización de los datos y los metadatos en el mercado o almacén de datos.

1.5.1 Algunos conceptos importantes

- ✓ **OLTP:** "la descripción original de todas las actividades y los sistemas asociados a la introducción de datos de forma fiable en una base de datos".[13] Los OLTP más comunes son: las hojas de cálculo, archivos de textos, hipertextos, BD transaccionales, entre otros.
- ✓ ***Staging area:*** "es al mismo tiempo un área de almacenamiento y procesamiento de datos, [...] abarca todo lo que se encuentre entre los sistemas operaciones origen y el área de presentación de datos". [14] Se trata así de evitar afectar estos últimos además de reducir la posibilidad de ocurrencia de errores.
- ✓ **Metadatos:** "son una forma de abstracción que describe la estructura y contenido del almacén de datos".[15] Constituyen un modelo de abstracción para describir la estructura y contenidos de un mercado o almacén de datos.

1.5.2 Evaluación del proceso de extracción, transformación y carga

De acuerdo con el estándar establecido en el CIM para tal procedimiento se elaborará una lista de chequeo que tendrá en cuenta los principales aspectos que tributan a una eficaz implementación de un mercado de datos.

1.6 Metodologías Kimball e Inmon para el desarrollo de almacenes de datos

Para definir la arquitectura y guiar el desarrollo de un almacén de datos existen varias alternativas de distinto grado de aprobación en el ámbito internacional.

Ralph Kimball, reconocida autoridad en el tema, determinó que un almacén de datos no es más que la unión de todos los mercados de datos de una entidad. Defiende por tanto una metodología ascendente (bottom-up) a la hora de diseñar el almacén de datos.

Inmon, quien es considerado el padre de los almacenes de datos, propone una metodología descendente (top-down), ya que de esta forma se considerarán mejor todos los datos corporativos. En esta metodología los mercados de datos se crearán después de haber terminado el almacén de datos de la institución.

La metodología Kimball es aceptada en todo el mundo como la más efectiva para desarrollar una solución de construcción de almacén de datos. Es más flexible y sencilla de implementar, ya que a partir de un mercado de datos como primer elemento del sistema de análisis, se pueden añadir otros que comparten dimensiones u otras nuevas.

La metodología Inmon por el contrario tiene como centro de su arquitectura al almacén de datos corporativo a partir del cual se alimentan los mercados de datos. Tiene un enfoque global muy complejo que en la práctica implica un ciclo de desarrollo mucho más largo.

1.7 Metodología para el desarrollo de un mercado de datos para los EC que se gestionan en el CIM

La metodología de desarrollo a seguir durante esta investigación queda establecida por el trabajo previo realizado entre el departamento de Manejo de Datos del CIM y la UCI. En la actualidad se encuentra en fase de despliegue el mercado de datos sobre los EC realizados al producto Nimotuzumab implementado bajo este estándar que ha adoptado la entidad. Dicho estándar tomó los elementos que más se adecuaban a la particular situación del CIM de algunas de las metodologías más utilizadas en el mundo como el Ciclo de vida Kimball, el Data Warehouse Engineering Process, entre otras un tanto menos conocidas como es el caso de Hefesto y del Desarrollo de Almacenes de Datos dirigidos por modelos.

Los epígrafes **1.7.1** y **1.7.2** pretenden proporcionar una idea general de cómo aborda esta metodología las etapas que de acuerdo con el alcance de la presente investigación son de interés.

1.7.1 Procedimiento para la etapa de análisis y diseño

Durante el estudio del procedimiento a seguir en la presente etapa, y luego de la consulta con los propios autores del mismo así como con expertos en la materia, se detectó que era necesario realizarle algunas modificaciones. Dichas modificaciones no alteran el procedimiento en sí esencialmente, mas buscan ganar en claridad y precisión para lograr una mejor correspondencia entre la denominación del flujo de trabajo y las actividades que en él se realizan.

Básicamente se trató de renombrar el segundo flujo de actividades a **Especificación de requerimientos de información** debido a que la denominación anterior daba una idea imprecisa de lo que realmente se hace en este flujo de trabajo, más acorde con el nuevo título. En el caso del último flujo de actividades se decidió englobar las actividades de **Diseñar tablas de hechos**, **Diseñar tablas de dimensiones**, **Realizar Uniones** y **Determinar jerarquías** en una nombrada **Confeccionar Modelo lógico** porque se consideró que la separación de tales tareas resultaba forzada.

Como se puede apreciar en la **Fig. 2** el procedimiento para la etapa de análisis y diseño cuenta con cuatro flujos desglosados a su vez en una serie de actividades.

- ✓ **Análisis del negocio:** se lleva a cabo la definición y el análisis de requisitos lo que proporciona al diseñador una mejor visión de los procesos de toma de decisiones y los objetivos del usuario. Para el logro de estos objetivos se utilizan técnicas como: entrevistas, cuestionarios y observación. Entregable: Análisis del negocio.
- ✓ **Especificación de requisitos de información:** se describen las necesidades de la organización a través del análisis de los objetivos y de las variables informacionales identificadas en el paso anterior y se redactan a manera de requisitos. Es en este flujo que se confecciona el modelo conceptual. Entregable: Especificación de las necesidades.
- ✓ **Desarrollo del modelo conceptual:** se desarrolla el modelo conceptual, se determina el cálculo de los indicadores y se establece la correspondencia entre las perspectivas del modelo conceptual y los diccionarios de datos (DD). Luego se define el nivel de granularidad. Finalmente, se evaluarán tanto el modelo conceptual como los DD. Entregable: Desarrollo del modelo conceptual.

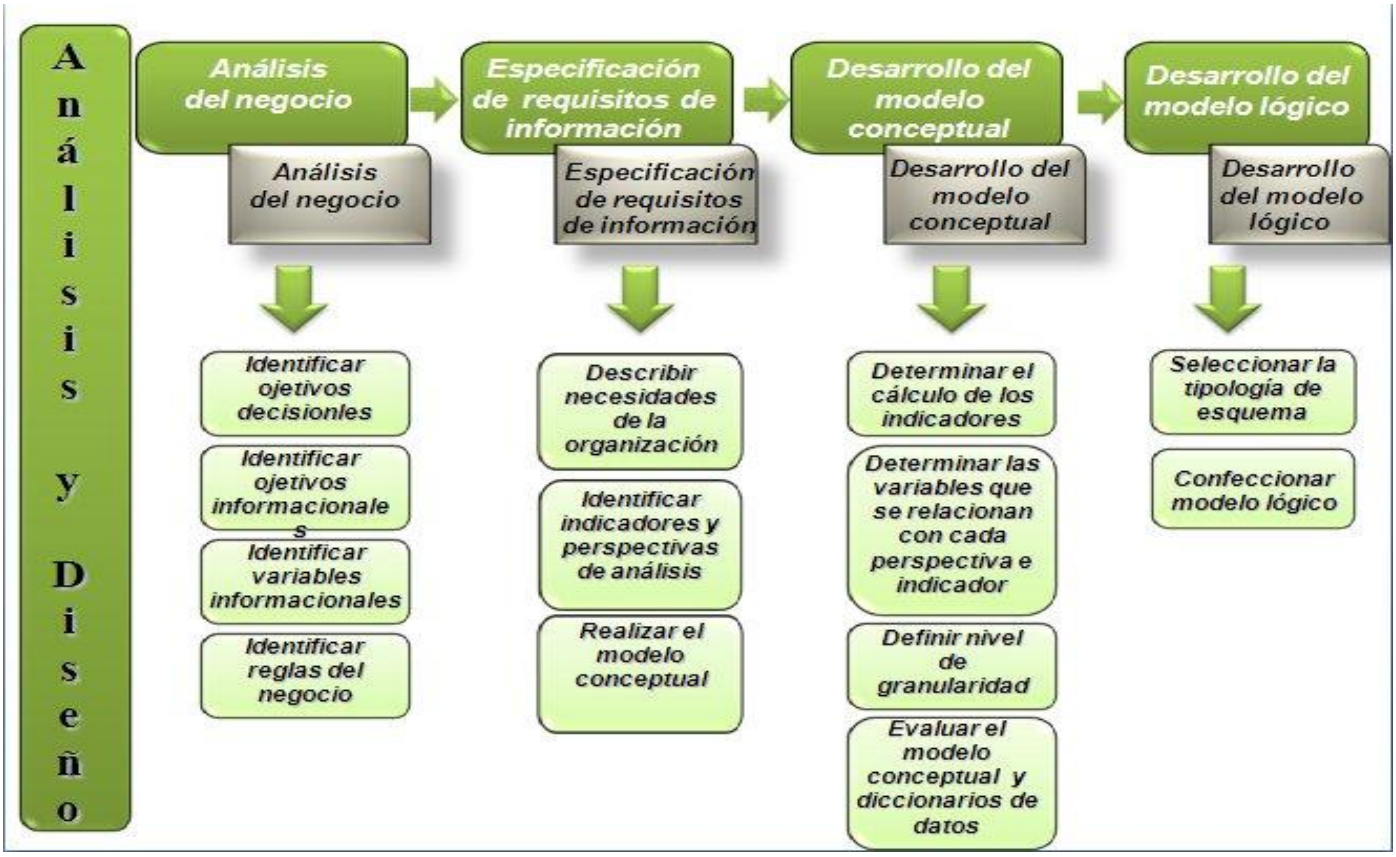

FIG. 2: PROCEDIMIENTO PARA LA ETAPA DE ANÁLISIS Y DISEÑO

- ✓ **Desarrollo del modelo lógico:** se confecciona el modelo lógico de la estructura del mercado de datos. Para ello se parte del modelo conceptual que ya ha sido creado y se respeta el tipo de modelo, así como las tablas de dimensiones y de hechos. Finalmente, se realizarán las uniones pertinentes entre las tablas y se determinarán las jerarquías. Entregable: Desarrollo del modelo lógico.

Para un estudio más profundo del procedimiento original remitirse a la Bibliografía.

1.7.2 Procedimiento para la etapa de extracción, transformación y carga

A raíz de la consulta con un experto fueron detectadas algunas irregularidades en el diseño original del procedimiento de ETL, por lo que se hizo necesario realizar una revisión más profunda como resultado de la cual se obtuvo el procedimiento que se muestra en la **Fig. 3**.

Se suprimió el flujo de trabajo **Selección de herramientas informáticas** por considerarse una actividad previa incluso a la etapa de análisis y diseño del mercado de datos. El flujo de trabajo **Extracción, transformación y carga** se presenta ahora en una forma más acorde con la definición del procedimiento que debe ser aplicable a la generalidad de situaciones de negocio del CIM.

El procedimiento para la etapa de ETL cuenta con cuatro flujos principales de actividades los cuales se abordarán de manera muy breve seguidamente.

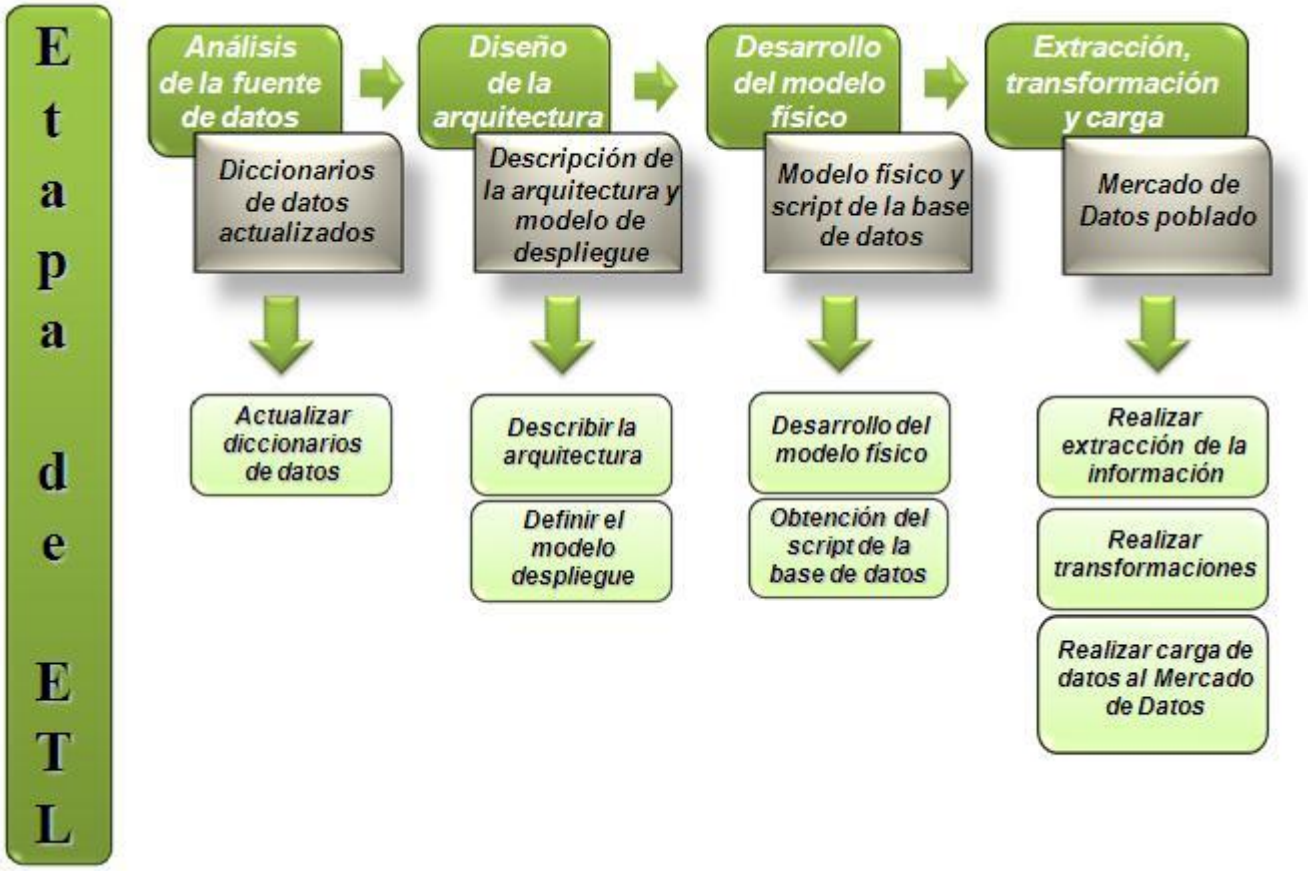

FIG. 3: PROCEDIMIENTO PARA LA ETAPA DE ETL

- ✓ **Análisis de las fuentes de datos:** se actualizan los DD al establecerse la correspondencia de todas las perspectivas con sus datos fuentes. A partir de este artefacto se define el tipo y formato de cada variable, así como las aplicaciones, bases de datos o cualquier archivo donde resida información útil para el proceso de ETL.

- ✓ **Diseño de la arquitectura:** se define la estrategia arquitectónica a usar; además se define si se hará uso de *staging area* y metadatos. Se deben especificar los requerimientos de hardware necesarios para asumir la implementación del mercado de datos así como la estructura del despliegue del mercado de datos. Entregable: Descripción de la arquitectura y modelo de despliegue.
- ✓ **Desarrollo del Modelo Físico:** en este paso es donde se convierte el modelo lógico, que se obtuvo a partir de las necesidades del negocio, a modelo físico. En él se especifican los tipos de datos de las variables que fueron definidos anteriormente y la cardinalidad entre las tablas. Además se genera el script de la base de datos para el gestor seleccionado. Entregable: Modelo Físico y script de la base de datos.
- ✓ **Extracción, transformación y carga:** este proceso puede repetirse o ajustarse el orden de las acciones de acuerdo con la arquitectura definida. La extracción de datos se lleva a cabo a partir del DD con la herramienta seleccionada. El proceso de transformación es regido por las reglas del negocio identificadas en el flujo de trabajo **Análisis del negocio** así como transformaciones adicionales. Una vez definidas se procede a la configuración e implementación de las mismas en la herramienta a utilizar. Luego los datos son almacenados en el mercado de datos con lo que se culmina la carga. Entregables: Reglas de transformación y mercado de datos poblado.

Para un estudio más profundo del procedimiento original remitirse a la Bibliografía.

1.8 Lenguaje Unificado de Modelado

En el desarrollo de esta investigación se decide emplear el Lenguaje Unificado de Modelado (UML por sus siglas en inglés) en su versión 2.0.

UML “es un lenguaje de modelado visual que se usa para especificar, visualizar, construir y documentar artefactos de un sistema de software”.[16]

El uso de los diagramas UML puede facilitar considerablemente la descripción y mejor comprensión de complejos diseños de software. Es el lenguaje de modelado más extendido y conocido y cuenta con el respaldo del Object Managment Group desde 1997. UML 2.0 ha sido la la primera gran revisión del lenguaje.

1.9 Herramientas de modelado

El número de herramientas para automatizar los procesos de Ingeniería de Software es cada día mayor. Las herramientas CASE (Computer Aided Software Engineering) facilitan enormemente el proceso de desarrollo de programas de computadoras.

Entre las herramientas CASE que soportan UML se encuentran:

- ✓ Rational Rose
- ✓ Visual Paradigm
- ✓ Borland Together

Debido a su independencia del sistema operativo, probada calidad, soporte para la generación de scripts de base de datos para distintos SGBD, integración con sistemas de control de versiones lo que posibilitaría el seguimiento de los cambios, facilidad de instalación y utilización, compatibilidad con otras versiones así como la experiencia previa del equipo de desarrollo, se decide emplear Visual Paradigm, versión 6.4.

1.10 Técnicas de captura de requisitos

Para lograr la correcta identificación de los requisitos del sistema se hace necesario el uso de métodos específicos. Se trata de un proceso que se torna muy variable dado el número de personas involucradas de distinta formación profesional.

Durante esta investigación se utilizaron simultáneamente las entrevistas, tormenta de ideas, observación y discusiones como principales técnicas.

1.11 Herramienta para el proceso de extracción, transformación y carga

Kettle o Pentaho Data Integration (PDI), es un proyecto de software libre compuesto por las herramientas SPOON, PAN, CHEF y KITCHEN que facilitan el proceso de ETL.

- ✓ **SPOON:** permite diseñar de forma gráfica las transformaciones ETL.
- ✓ **PAN:** ejecuta un conjunto de transformaciones diseñadas con SPOON.
- ✓ **CHEF:** permite diseñar la carga de datos incluyendo un control de estado de los trabajos.
- ✓ **KITCHEN:** permite ejecutar los trabajos *en lotes* diseñados con CHEF.

Se trata de una herramienta con un alto nivel de automatización y muy versátil.

Entre sus ventajas se encuentran:

- ✓ Se ejecuta sobre plataformas Windows, Unix y Linux.
- ✓ Tiene una interfaz gráfica con indicadores de las trasformaciones.
- ✓ Es una aplicación implementada en Java con algunas características avanzadas en JavaScript.
- ✓ Ofrece una licencia pública GNU Lesser General Public License versión 2.1 (LGPLv2.1).
- ✓ Soporte para el trabajo con metadatos.
- ✓ Cuenta con una activa comunidad de usuarios regida por los preceptos del software de código abierto. El sitio oficial de PDI cuenta con casi 30 mil miembros registrados que colaboran continuamente en los foros habilitados a tal efecto.
- ✓ 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.

1.12 Sistema de Gestión de Bases de Datos

El SGBD proporciona los procedimientos necesarios para la gestión de los datos. Permite la eliminación y modificación de los registros, la interacción con otras bases de datos y la generación de reportes.

Objetivos de los SGBD:

- ✓ Evitar la redundancia de los datos.
- ✓ Mejorar los mecanismos de seguridad de los datos y la privacidad.
- ✓ Asegurar la independencia de los programas y los datos, es decir, la posibilidad de modificar la estructura de la base de datos sin necesidad de modificar los programas de las aplicaciones que manejan esos datos.
- ✓ Mantener la integridad de los datos a través de las validaciones necesarias cuando se realicen modificaciones en la base de datos.

El CIM se encuentra en la etapa de migración hacia software libre de su infraestructura informática, entre ellos la plataforma PostgreSQL, otra razón para elegir este como SGBD en su versión estable 8.4.

1.12.1 PostgreSQL

"PostgreSQL es un sistema de gestión de bases de datos relacional orientado a objetos basado en POSTGRES, versión 4.2, desarrollado en la Universidad de California en el Departamento de Ciencias Computacionales de Berkeley".[17] Como muchos otros proyectos de código abierto es mantenido por una amplia comunidad de desarrolladores y organizaciones comerciales. PostgreSQL se acoge a una arquitectura que sigue el modelo cliente/servidor. El proceso servidor gestiona los scripts de la base de datos así como las configuraciones internas, además maneja las consultas a la base de datos de las aplicaciones clientes para ejecutar alguna operación.

Ventajas de PostgreSQL:

- ✓ Drivers: Open DataBase Connectivity (ODBC), Java Database Connectivity (JDBC) entre otros. JDBC es "un standard que permite a los programas escritos en Java interactuar con cualquier base de datos que disponga del controlador correspondiente".[18] JDBC es la interfaz empleada de forma native por PDI.
- ✓ Soporta: *triggers*, procedimientos almacenados, funciones, secuencias, relaciones, reglas, tipos de datos definidos por el usuario, vistas y vistas materializadas.
- ✓ Soporte de protocolo de comunicación encriptado Socket Secure Layer (SSL):"es un sistema de protocolos de carácter general diseñado en 1994 por la empresa Nestcape Communcations Corporation, y está basado en la aplicación conjunta de Criptografía Simétrica, Criptografía Asimétrica (de llave pública), certificados digitales y firmas digitales para conseguir un canal o medio seguro de comunicación a través de Internet".[19]
- ✓ Máximo de bases de datos: ilimitado.
- ✓ Máximo de tamaño de tabla: 32 TeraByte.
- ✓ Máximo de tamaño de registro: 1.6 TeraByte.
- ✓ Máximo de tamaño de campo: 1 GigaByte.
- ✓ Máximo de registros por tabla: ilimitado.
- ✓ Máximo de campos por tabla: 250 a 1600 (depende de los tipos usados).
- ✓ Máximo de índices por tabla: ilimitado.

- ✓ Número de lenguajes en los que se puede programar funciones: aproximadamente diez (pl/pgsql, pl/java, pl/perl, pl/python, tcl, pl/php, C, C++ y Ruby).

1.13 Conclusiones parciales

Este capítulo constituye un acercamiento al contexto tecnológico de los mercados de datos. En la primera parte se realizó una revisión de algunos conceptos generales relacionados con el tema, para dar paso luego a aspectos de carácter más específico directamente vinculados al desarrollo de la presente investigación.

Como fruto del análisis realizado en los epígrafes previos de este capítulo se decide emplear para la definición de la arquitectura, así como para las etapas de análisis y diseño y la extracción, transformación y carga del mercado de datos, la propuesta establecida por el CIM en investigaciones previas con las modificaciones anteriormente acotadas. Aunque existen metodologías de probada efectividad estas tienen un carácter general mientras que se cuenta con un procedimiento totalmente dirigido al proceso de manejo de datos del CIM que ha recibido un aval científico y está inspirado precisamente en varias de las más enraizadas metodologías a nivel internacional.

Para la realización de los modelos y diagramas propios del proceso de ingeniería de software se utilizará como lenguaje de modelado UML y como herramienta de modelado el Visual Paradigm 6.4, cuyas ventajas ya han sido expuestas. Para la gestión de las bases de datos se decide usar PostgreSQL 8.4 y como interfaz para la administración del mismo pgAdmin III. Entre los SGBD PostgreSQL se distingue por la atención que le brinda al mantenimiento de la integridad de datos y su fiabilidad al trabajar con altos volúmenes de carga, ambos factores determinantes si se tiene en cuenta que se pretende desarrollar un mercado de datos que almacenará información médica, por lo que el nivel de tolerancia a la pérdida de datos es cero.

Se definieron como técnicas de captura de requisitos durante la etapa de análisis y diseño las entrevistas, tormentas de ideas y la observación. Todo ello acorde con las características y posibilidades de intercambio del personal que trabaja en la investigación.

Para dar soporte al proceso de ETL se eligió el Pentaho Data Integration o Kettle. Los principales factores que influyeron en esta elección están dados por su carácter multiplataforma, el completo respaldo brindado al proceso ETL y la experiencia previa en su utilización en el entorno UCI.

La evaluación del procedimiento se llevará a cabo a través de la técnica de lista de chequeo para constatar el desempeño de la solución propuesta.

Capitulo 2: Análisis y Diseño del Mercado de Datos CIMAvax EGF

2.1 Introducción

En este capítulo se aborda la etapa de análisis y diseño de un mercado de datos para los Ensayos Clínicos del producto CIMAvax EGF que se gestionan en el Centro de Inmunología Molecular. El principal objetivo del análisis y diseño consiste en definir con exactitud las características del mercado de datos. Las especificaciones generales del procedimiento aplicado en esta etapa pueden encontrarse en el sub epígrafe **1.7.1 Procedimiento para la etapa de análisis y diseño**.

2.2 Aplicación del procedimiento elegido para la etapa de análisis y diseño

Los sub epígrafes a continuación presentan los resultados obtenidos de cada flujo de actividad correspondiente a la etapa de análisis y diseño.

2.2.1 Análisis del negocio

En este flujo para la definición de los requisitos se realizó una aproximación basada en objetivos.

Identificar objetivos decisionales

Se identificaron dos objetivos decisionales de la organización a partir de los objetivos estratégicos relacionados a la gestión de los Ensayos Clínicos conducidos en el Centro de Inmunología Molecular:

- ✓ Verificar la eficacia del producto **EGF** en los pacientes a los cuáles se aplica.
- ✓ Verificar la seguridad del producto **EGF** en los pacientes a los cuáles se aplica.

Identificar objetivos informacionales

Se identificaron 31 objetivos informacionales que son los que están relacionados con la información necesaria para alcanzar los objetivos decisionales identificados en el paso anterior:

Para **verificar la eficacia del producto EGF**, es necesario conocer:

- ✓ Relación entre la evaluación de la respuesta y el esquema de tratamiento recibido.
- ✓ Relación entre la evaluación de la respuesta y el número de dosis recibidas.
- ✓ Relación entre la evaluación de la respuesta y el título máximo de cada paciente.
- ✓ Relación entre la evaluación de la respuesta y la concentración mínima del EGF [EGF].

- ✓ Relación entre la evaluación de la respuesta y los pacientes GAR y PAR.
- ✓ Relación entre el tiempo de supervivencia y el esquema de tratamiento recibido.
- ✓ Relación entre el tiempo de supervivencia y el número de dosis recibidas.
- ✓ Relación entre el tiempo de supervivencia y el título máximo de cada paciente.
- ✓ Relación entre el tiempo de supervivencia y la [EGF] mínima.
- ✓ Relación entre el tiempo de supervivencia y los pacientes GAR y PAR.
- ✓ Relación entre el tiempo de progresión y el esquema de tratamiento recibido.
- ✓ Relación entre el tiempo de progresión y el número de dosis recibidas.
- ✓ Relación entre el tiempo de progresión y el título máximo.
- ✓ Relación entre el tiempo de progresión y la [EGF] mínima por paciente.
- ✓ Relación entre el tiempo de progresión y los pacientes GAR y PAR.

Para **verificar la seguridad del producto EGF**, es necesario conocer:

- ✓ Cuáles son los tipos de eventos adversos más frecuentes en cada esquema de tratamiento.
- ✓ Cuáles son los tipos de eventos adversos más frecuentes por número de dosis recibidas.
- ✓ Cuáles son los tipos de eventos adversos más frecuentes en los pacientes GAR y PAR
- ✓ Cuántos pacientes presentaron eventos adversos por esquema de tratamiento.
- ✓ Cuántos pacientes presentaron eventos adversos por número de dosis recibidas.
- ✓ Relación entre intensidad de los eventos adversos y esquema de tratamiento.
- ✓ Relación entre intensidad de los eventos adversos y número de dosis recibidas.
- ✓ Relación entre causalidad de los eventos adversos y esquema de tratamiento.
- ✓ Relación entre causalidad de los eventos adversos y número de dosis recibidas.
- ✓ Relación del producto con el medicamento concomitante.
- ✓ Relación entre seriedad de los eventos adversos y esquema de tratamiento.

- ✓ Cuáles son los tipos de eventos serios (relacionados y no relacionados) en los pacientes GAR y PAR.
- ✓ Relación del tiempo de supervivencia y los eventos serios relacionados y no relacionados.
- ✓ Relación entre seriedad de los eventos adversos y número de dosis recibidas.
- ✓ Tiempo (días/horas) entre la aparición de los EA y la administración del producto.

Para cada objetivo identificado se debe tener en cuenta el tiempo, la localización, el número del ensayo y los datos demográficos.

Identificar variables informacionales

Se identificaron las siguientes variables informacionales a partir de los objetivos informacionales identificados en el paso anterior.

Generales

Localización: es el lugar donde se manifiesta la enfermedad. Estas pueden ser:

- ✓ pulmón
- ✓ próstata
- ✓ estómago
- ✓ colon

Ensayo: Está en correspondencia con la localización. Estos son:

- ✓ PI tumores sólidos 019
- ✓ PII pulmón 025
- ✓ PIII pulmón 033
- ✓ PIV pulmón 041
- ✓ PV pulmón 062
- ✓ FII pulmón 056
- ✓ FII próstata 077
- ✓ FIII pulmón 081
- ✓ FIII pulmón VQV 111

Tratamientos previos: tratamiento(s) que ha recibido el paciente antes del ensayo. Los valores contemplados son quimioterapia, radioterapia y cirugía.

Número de dosis: cantidad de dosis que se le suministra al paciente de un determinado tratamiento.

Datos demográficos: sexo, edad, raza, estadio, T, N, M, talla, peso, clasificación anatomopatológica, grado de diferenciación, ECOG (Karnofsky) y tratamientos previos.

- ✓ Sexo: Indica el género del paciente.
- ✓ Edad: Indica la edad del paciente. En el caso de que no aparezca esta variable es calculable a partir de la fecha de inclusión y la fecha de nacimiento.
- ✓ Raza: Indica la raza del paciente. Se contemplan los valores blanca, negra o mestiza.
- ✓ Estadío: gravedad de la enfermedad al diagnosticar el paciente. Se contemplan los valores I, II, III o IV.
- ✓ T: tamaño del tumor. Se contemplan los valores X, 0, 1, 2, 3 ó 4.
- ✓ N: número de ganglios infiltrados. Se contemplan los valores X, 0, 1, 2 ó 3.
- ✓ M: si tiene metástasis a distancia. Se contemplan los valores 0, 1 ó X.
- ✓ Talla: Indica la talla del paciente en centímetros (cm).
- ✓ Peso: Indica el peso del paciente en kilogramos (kg).
- ✓ Clasificación anatomopatológica: clasificación del tumor al diagnosticar al paciente.
- ✓ Grado de diferenciación: forma parte de la clasificación anatomopatológica. Se contemplan los valores: moderadamente diferenciado, pobremente diferenciado, poco diferenciado o indiferenciado.
- ✓ ECOG (Karnofsky): estado general del paciente cuando se diagnostica.

Esquema de tratamiento**:**

EC	Esquema de tratamiento (dosis/ Proteína transportadora, adyuvante y forma de tratamiento)	
Grupos	1	2
PI T. sólidos 019	50µg/ TT Alúmina	50µg/ P64 Alúmina
PII pulmón 025	50µg/ P64 Alúmina	50µg/ P64 Montanide
PIII pulmón 033	50µg/ P64 Alúmina	50µg/ P64 Montanide
PIV pulmón 041	71µg/ P64 Alúmina	142µg/ P64 Alúmina
PV pulmón 062	200µg/ P64 Montanide, VQV	

FII pulmón 056	50µg/ P64 Montanide, QV	control (no recibe EGF)
FII próstata 077	200µg/ P64 Montanide, VQV	Control (no recibe EGF)
FIII pulmón 081	200µg/ P64 Montanide, QV	Control (no recibe EGF)
FIII pulmón VQV 111	200µg/ P64 Montanide, VQV	Control (no recibe EGF)

TABLA 2: ESQUEMA DE TRATAMIENTO

Título máximo: la mayor respuesta al anticuerpo de EGF que tuvo el paciente.

Mínima [EGF]: mínima concentración de EGF en sangre.

GAR / PAR: Buenos Respondedores al Anticuerpo/ Pobres respondedores al Anticuerpo. Se calcula a partir del Título Máximo.

Para el proceso de eficacia

Respuestas: respuestas de los pacientes ya sea al tratamiento oncoespecífico (respuesta al tratamiento previo) o al tratamiento recibido dentro del ensayo. Los tipos de respuestas contemplados son:

- ✓ RC: respuesta completa
- ✓ RP: respuesta parcial
- ✓ P: progresión
- ✓ EE: enfermedad estable
- ✓ MP: muerte precoz

Tiempo de Supervivencia: variable calculable a partir de la diferencia entre la fecha de la última consulta realizada o la última visita o el último contacto con el paciente o la fecha de fallecido y la fecha de inclusión.

Tiempo de progresión: variable calculable a partir de la diferencia entre la fecha de la primera vez que se observa la respuesta “progresión” y la fecha de inclusión.

Para el proceso de seguridad

Eventos adversos (EA): son las reacciones negativas que manifiestan los pacientes durante el ensayo. Algunos tipos de eventos adversos son:

- ✓ Cefalea, dolor de cabeza o migraña
- ✓ Fiebre
- ✓ Diarrea
- ✓ Vómitos

- ✓ Anemia (hemoglobina baja)
- ✓ Aumento de la fosfatasa alcalina

Intensidad: refleja la intensidad del evento adverso: Se contemplan los valores: leve, moderado, severo o, muy severo.

Causalidad: especifica si el evento adverso está relacionado o no con el tratamiento. Las clasificaciones son:

- ✓ Relacionado: posible, probable, muy probable y definitiva.
- ✓ No relacionado: improbable, no relacionado, desconocido.

Seriedad: es una definición regulatoria. Determina conductas a seguir.

- ✓ Serio: cualquier EA que a cualquier dosis de un producto tiene las siguientes consecuencias: Muerte, Riesgo de muerte, Hospitalización, Prolongación de la hospitalización, Incapacidad persistente o significativa y Anomalía congénita/defecto de nacimiento.
- ✓ No serio: no provoca estas consecuencias.

Tiempo de aparición del EA: fecha en que comienza a manifestarse el evento adverso.

Fecha de administración del producto: fecha en que el paciente es inmunizado con EGF.

Identificar reglas del negocio

Las reglas del negocio abarcan las políticas, estándares, operaciones, definiciones y restricciones propias de la organización. Son además, de importancia vital para el alcance de los objetivos y normalmente dan lugar a las reglas de transformación. A continuación se relacionan algunas de las reglas del negocio detectadas.

- ✓ De las diferentes respuestas al tratamiento registradas para un mismo paciente se tomará la mejor respuesta en orden descendente de acuerdo con la siguiente escala de prioridad: respuesta completa, respuesta parcial, progresión, enfermedad estable y muerte precoz.
- ✓ En el ensayo de localización próstata no se recogió el sexo de los pacientes, este es evidentemente masculino.
- ✓ Los eventos adversos de aquellos pacientes cuyo grupo de tratamiento es control tienen una causalidad no relacionada.

- ✓ Aquellas perspectivas que no se hayan tenido en cuenta en algún ensayo tomarán el valor: No se recogió en este ensayo.
- ✓ El valor de la variable talla debe presentarse en cm.
- ✓ El valor de la variable peso debe presentarse en kg.
- ✓ Los valores contemplados para la variable tratamientos previos son quimioterapia, radioterapia y cirugía.

2.2.2 Especificación de requisitos de información

Este flujo tiene como entrada principal los objetivos y variables identificados durante el **Análisis del negocio**. A continuación se presentan los resultados para cada una de las actividades del presente flujo.

Identificar las necesidades de la organización

Las necesidades de la organización identificadas fueron las siguientes:

Para el proceso de eficacia

- ✓ Se desea conocer la cantidad de respuestas por tipos de respuestas para un esquema de tratamiento determinado, lo que se traduce en**: "Cantidad de respuestas por tipos de respuestas para un esquema de tratamiento".**
- ✓ Se desea conocer la cantidad de respuestas por tipos de respuestas para una cantidad determinada de número de dosis, lo que se traduce en**: "Cantidad de respuestas por tipos de respuestas para un número de dosis".**
- ✓ Se desea conocer la cantidad de respuestas por tipos de respuestas para un título máximo determinado, lo que se traduce en: **"Cantidad de respuestas por tipos de respuestas para un título máximo".**
- ✓ Se desea conocer la cantidad de respuestas por tipos de respuestas para una concentración mínima de EGF, [EGF], dada, lo que se traduce en: **"Cantidad de respuestas por tipos de respuestas para una [EGF]".**
- ✓ Se desea conocer la cantidad de respuestas por tipos de respuestas para los pacientes GAR y los pacientes PAR, lo que se traduce en: **"Cantidad de respuestas por tipos de respuestas para los pacientes GAR y los pacientes PAR".**

- ✓ Se desea conocer cuál fue el tiempo de supervivencia de los pacientes que se enfrentaron a un esquema de tratamiento determinado, lo que se traduce en: **"Tiempo de supervivencia para un esquema de tratamiento"**.
- ✓ Se desea conocer cuál fue el tiempo de supervivencia de los pacientes que se enfrentaron a una determinada cantidad de dosis, lo que se traduce en**: "Tiempo de supervivencia para un número de dosis".**
- ✓ Se desea conocer cuál fue el tiempo de supervivencia de los pacientes con un título máximo determinado, lo que se traduce en**: "Tiempo de supervivencia para un título máximo".**
- ✓ Se desea conocer cuál fue el tiempo de supervivencia de los pacientes con una [EGF] determinada, lo que se traduce en**: "Tiempo de supervivencia para una [EGF]".**
- ✓ Se desea conocer cuál fue el tiempo de supervivencia de los pacientes GAR y los pacientes PAR, lo que se traduce en**: "Tiempo de supervivencia de los pacientes GAR y los pacientes PAR".**
- ✓ Se desea conocer el tiempo de progresión de los pacientes con un esquema de tratamiento determinado, lo que se traduce en: **"Tiempo de progresión para un esquema de tratamiento".**
- ✓ Se desea conocer el tiempo de progresión de los pacientes con un número de dosis determinado, lo que se traduce en: **"Tiempo de progresión para un número de dosis".**
- ✓ Se desea conocer el tiempo de progresión de los pacientes con un título máximo determinado, lo que se traduce en: **"Tiempo de progresión para un título máximo".**
- ✓ Se desea conocer el tiempo de progresión de los pacientes con una [EGF] determinada, lo que se traduce en: **"Tiempo de progresión para una [EGF]".**
- ✓ Se desea conocer el tiempo de progresión de los pacientes PAR y de los pacientes GAR, lo que se traduce en: **"Tiempo de progresión de los pacientes GAR y de los pacientes PAR".**

Para el proceso de seguridad

- ✓ Se desea conocer cuántos eventos adversos hubo por tipos, por intensidad y por causalidad de los eventos adversos según el esquema de tratamiento, lo que se traduce en**: "Cantidad de eventos adversos por tipos, por intensidad y por causalidad de los eventos adversos según el esquema de tratamiento".**

- ✓ Se desea conocer cuántos eventos adversos hubo por tipos, por intensidad y por causalidad de los eventos adversos según el número de dosis, lo que se traduce en**: "Cantidad de eventos adversos por tipos, por intensidad y por causalidad de los eventos adversos según el número de dosis".**
- ✓ Se desea conocer cuántos eventos adversos hubo por tipos, por intensidad y por causalidad de los eventos adversos para los pacientes GAR y para los pacientes PAR, lo que se traduce en**: "Cantidad de eventos adversos por tipos, por intensidad y por causalidad de los eventos adversos para los pacientes GAR y para los pacientes PAR".**
- ✓ Se desea conocer cuántos pacientes tuvieron eventos adversos ante un esquema de tratamiento determinado, lo que se traduce en**: "Cantidad de pacientes con eventos adversos por esquema de tratamiento".**
- ✓ Se desea conocer cuántos pacientes tuvieron eventos adversos ante un número de dosis, lo que se traduce en**: "Cantidad de pacientes con eventos adversos por número de dosis recibida".**
- ✓ Se desea conocer la cantidad de pacientes por nivel de intensidad de los eventos adversos para un esquema de tratamiento determinado, lo que se traduce en: **"Cantidad de pacientes por nivel de intensidad de los eventos adversos para un esquema de tratamiento".**
- ✓ Se desea conocer la cantidad de pacientes por nivel de intensidad de los eventos adversos para un número de dosis determinado, lo que se traduce en: **"Cantidad de pacientes por nivel de intensidad de los eventos adversos para un número de dosis".**
- ✓ Se desea conocer la cantidad de eventos adversos por clasificación de causalidad para un esquema de tratamiento determinado, lo que se traduce en: **"Cantidad de de eventos adversos por clasificación de causalidad para un esquema de tratamiento".**
- ✓ Se desea conocer la cantidad de eventos adversos por clasificación de causalidad para un número de dosis determinado, lo que se traduce en: **"Cantidad de de eventos adversos por clasificación de causalidad para un número de dosis".**
- ✓ Se desea conocer la cantidad de pacientes para cada medicamento concomitante, lo que se traduce como: **"Cantidad de pacientes por medicamento concomitante".**

- ✓ Se desea conocer la cantidad de pacientes por grado de seriedad de los eventos adversos para un esquema de tratamiento determinado, lo que se traduce en: **"Cantidad de pacientes por grado de seriedad de los eventos adversos para un esquema de tratamiento".**
- ✓ Se desean conocer los eventos adversos serios relacionados y no relacionados de los pacientes GAR y los pacientes PAR, lo que se traduce en: **"Eventos adversos serios relacionados y no relacionados de los pacientes GAR y los pacientes PAR".**
- ✓ Se desea conocer el tiempo de supervivencia de los pacientes aquejados de eventos adversos serios relacionados y de los pacientes aquejados de eventos adversos serios no relacionados, lo que se traduce en: **"Tiempo de supervivencia de los pacientes aquejados de eventos adversos serios relacionados y de los pacientes aquejados de eventos adversos serios no relacionados".**
- ✓ Se desea conocer la cantidad de pacientes por grado de seriedad de los eventos adversos para un número de dosis determinado, lo que se traduce en: **"Cantidad de pacientes por grado de seriedad de los eventos adversos para un número de dosis".**
- ✓ Se desea conocer el tiempo transcurrido entre la aparición del evento adverso y la administración del producto a los pacientes, lo que se traduce en: **"Tiempo transcurrido entre la aparición del evento adverso y la administración del producto para cada paciente".**

Para cada una de estas necesidades se debe tener en cuenta el **tiempo**, la **localización**, el **número del ensayo** y los **datos demográficos** de los pacientes.

Identificar indicadores y perspectivas del análisis

Los conceptos de indicadores y perspectivas fueron abordados en el sub epígrafe **1.4.1 Modelo conceptual**.

Indicadores

Para el proceso de eficacia

- ✓ Cantidad de pacientes
- ✓ Tiempo de supervivencia
- ✓ Tiempo de progresión

Para el proceso de seguridad

- ✓ Cantidad de eventos adversos
- ✓ Cantidad de pacientes

Perspectivas

Para el proceso de eficacia

- ✓ Tipos de respuestas
- ✓ Esquema de tratamiento
- ✓ Número de dosis
- ✓ Título máximo
- ✓ [EGF]
- ✓ Pacientes GAR
- ✓ Pacientes PAR

Para el proceso de seguridad

- ✓ Tipos de eventos adversos
- ✓ Intensidad de los eventos adversos
- ✓ Causalidad de los eventos adversos
- ✓ Seriedad de los eventos adversos
- ✓ Esquema de tratamiento
- ✓ Número de dosis
- ✓ Pacientes GAR
- ✓ Pacientes PAR
- ✓ Medicamento concomitante
- ✓ Fecha aparición del evento adverso
- ✓ Fecha administración del producto

Ya que es una característica de los mercados de datos el ser variables en el tiempo se decide incluir el tiempo como una perspectiva más. También se tienen en cuenta localización, el número del ensayo y los datos demográficos.

Realizar el modelo conceptual

Como puede apreciarse en las **Fig. 5** y en la **Fig. 6**, el modelo conceptual permite comprender, sin poseer profundos conocimientos previos, cuáles serán los resultados que se obtendrán, cuáles serán las variables que se utilizarán para analizarlos y cuál es la relación que existe entre ellos.

Para el proceso de eficacia

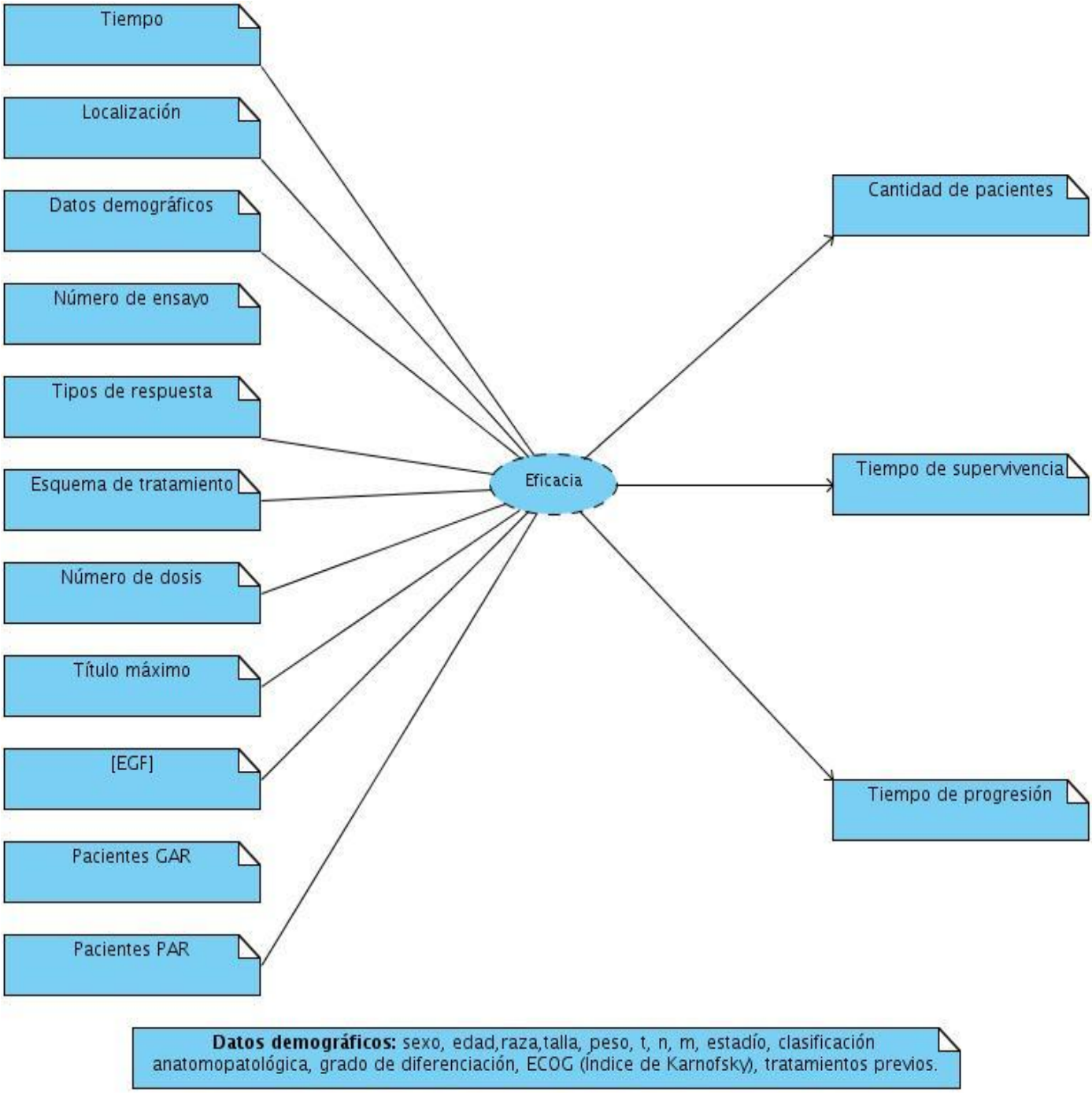

FIG. 4: MODELO CONCEPTUAL PARA EFICACIA

Para el proceso de seguridad

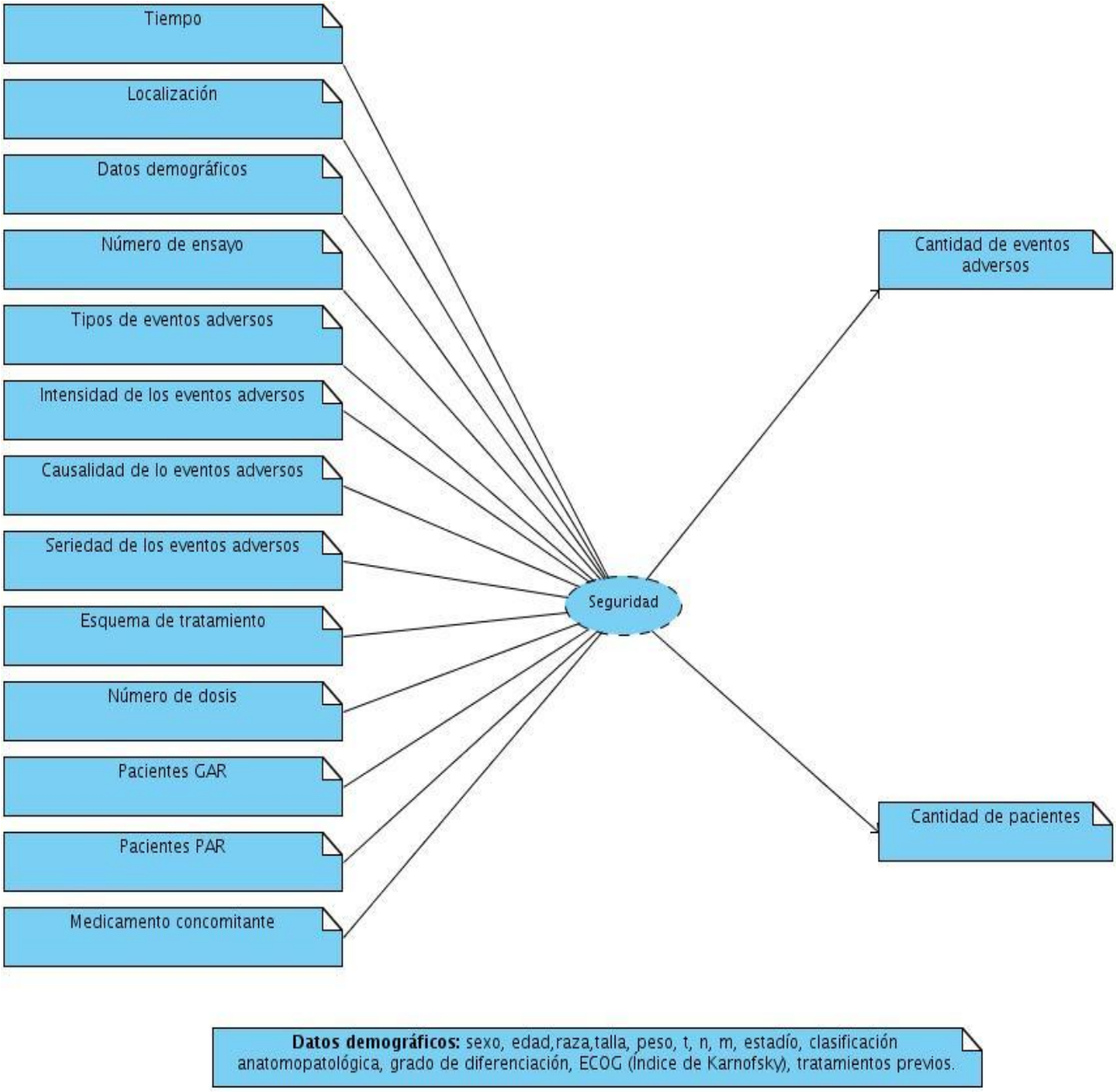

FIG. 5: MODELO CONCEPTUAL PARA SEGURIDAD

2.2.3 Desarrollo del modelo conceptual

Este flujo tiene como entrada principal las perspectivas e indicadores identificados durante el la **Especificación de requisitos de información**. A continuación se presentan los resultados para cada una de las actividades del presente flujo.

Determinar el cálculo de los indicadores

Los indicadores se calcularán de la siguiente forma:

Para el proceso de eficacia

- ✓ La **"Cantidad de pacientes"** representa la cantidad de pacientes con una determinada respuesta. Se obtiene al sumar cada uno de los pacientes que cumplan con los criterios de análisis.
- ✓ El **"Tiempo de supervivencia"** indica el tiempo transcurrido desde el diagnóstico de la enfermedad hasta el deceso.
- ✓ El **"Tiempo de progresión"** indica el tiempo transcurrido desde la inclusión en el ensayo hasta la fecha en la que se observa por primera vez como respuesta al tratamiento **Progresión (P)**.

Para el proceso de seguridad

- ✓ La **"Cantidad de pacientes"** representa la cantidad de pacientes con eventos adversos. Se obtiene al sumar cada uno de los pacientes que cumplan con los criterios de análisis.
- ✓ La **"Cantidad de eventos adversos"** representa la cantidad de eventos adversos por un tipo, una intensidad y una causalidad de evento adverso determinadas. Se obtiene al sumar cada uno de los eventos adversos que cumplan con los criterios de análisis.

Identificar las variables que se relacionan con cada perspectiva

Luego de un arduo trabajo de revisión de las fuentes de datos proporcionadas por el CIM se localizaron las perspectivas identificadas en el proceso de especificación de necesidades, a partir de las cuales se confeccionan los diccionarios de datos. En la **Tabla 3** se presenta la sección dedicada a los datos demográficos de uno de los DD confeccionados, el resto se adjunta al artefacto **Desarrollo del modelo conceptual**.

Perspectiva	Significado en el negocio	Modelo (s) donde se encuentra	Correspondencia con variable (s) de los datos fuentes	Posibles valores
Datos demográficos				
Sexo	Indica el género del paciente	Modelo1	Columna H, IdSexo	Masculino Femenino
Edad	Indica la edad del paciente	Modelo1	Columna G, Edad	Valor numérico entero positivo.
Raza	Indica la raza del paciente	Modelo1	Columna I, IdColorPiel	Blanca Negra Mestiza Amarilla
Estadío	Indica el estadío en que se encuentra el paciente	Modelo1	Columna G, Estadio	III IV
T	Indica el tamaño del tumor del paciente	Modelo2	Columna H, T	Valor numérico entero entre 0 y 4, incluyendo los extremos. X cuando no se conoce el valor.
N	Indica el número de ganglios que tiene el paciente en el tumor	Modelo2	Columna I, N	Valor numérico entero entre 0 y 2, incluyendo los extremos. X cuando no se conoce el valor.
M	Indica si el paciente tiene metástasis o no a alguna distancia del tumor	Modelo2	Columna J, M	0 1 X cuando no se conoce el valor.

Clasificación anatomopatoló-gica	Indica la clasificación del tumor al diagnosticar al paciente	Modelo2	Columna BH, newClasifAnatomopatologica	Cadena de caracteres.
Grado de diferenciación	Forma parte de la clasificación anatomopato-lógica	Modelo2	Columna BI, newGradDiferenciacion	MODERADAMENTE DIFERENCIADO POBREMENTE DIFERENCIADO POCO DIFERENCIADO INDIFERENCIADO
ECOG (Karnofsky)	Indica el estado general del paciente cuando se diagnostica	Modelo2	Columna K, EstadoGeneral	Valor numérico entero entre 0 y 2, incluyendo los extremos.
Tratamientos previos	Indica el tratamiento que ha recibido el paciente antes del ensayo	De acuerdo con el protocolo establecido para este ensayo el paciente no debe afrontar tratamiento previo alguno.		

TABLA 3: DICCIONARIO DE DATOS PULMÓN 062

Definir nivel de granularidad

El nivel de granularidad escogido es a nivel de paciente de acuerdo con la especificación del CIM.

Evaluar el modelo conceptual y diccionarios de datos

Posteriormente a la actualización de los diccionarios de datos se llevó a cabo una revisión tanto de este como del modelo conceptual. En la misma participaron compañeros del departamento de Manejo de Datos de Ensayos Clínicos del CIM. Se obtuvieron resultados satisfactorios y para tener constancia de ello el cliente emitió un aval de aceptación.

2.2.4 Desarrollo del modelo lógico

Este flujo tiene como entrada principal los entregables del **Desarrollo del modelo conceptual**. A continuación se presentan los resultados para cada una de las actividades del presente flujo.

Seleccionar la tipología de esquema

A partir de la especificación de las necesidades del cliente que permitió definir indicadores y perspectivas; y las bondades de las distintas tipologías multidimensionales estudiadas se escogió el esquema constelación de hechos.

Tablas de dimensiones

Cada una de las perspectivas identificadas en el modelo conceptual tributa a la creación de una o varias tablas de dimensión en el modelo lógico.

A continuación se muestran las tablas de dimensiones asociadas a la perspectiva identificada como datos demográficos:

FIG. 6: TABLA ASOCIADA A LA DIMENSIÓN SEXO

FIG. 7: TABLA ASOCIADA A LA DIMENSIÓN RAZA

FIG. 8: TABLA ASOCIADA A LA DIMENSIÓN EDAD

FIG. 9: TABLA ASOCIADA A LA DIMENSIÓN TALLA

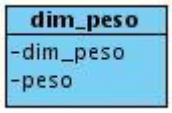

FIG. 10: TABLA ASOCIADA A LA DIMENSIÓN PESO

FIG. 11: TABLA ASOCIADA A LA DIMENSIÓN TAMAÑO

FIG. 12: TABLA ASOCIADA A LA DIMENSIÓN NÚMERO DE GANGLIOS

FIG. 13: TABLA ASOCIADA A LA DIMENSIÓN METÁSTASIS

FIG. 14: TABLA ASOCIADA A LA DIMENSIÓN ESTADIO

FIG. 15: TABLA ASOCIADA A LA DIMENSION ECOG

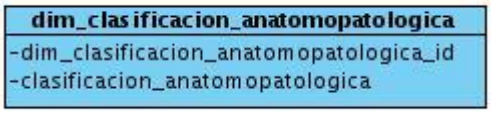

FIG. 16: TABLA ASOCIADA A LA DIMENSIÓN CLASIFICACIÓN ANATOMOPATOLÓGICA

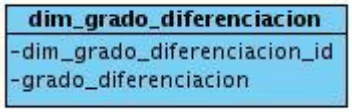

FIG. 17: TABLA ASOCIADA A LA DIMENSIÓN GRADO DE DIFERENCIACIÓN

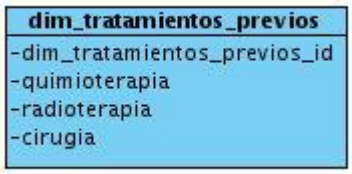

FIG. 18: TABLA ASOCIADA A LA DIMENSIÓN TRATAMIENTOS PREVIOS

La cantidad de tablas de dimensiones creadas es de 27.

Tablas de hechos

Las tablas de hechos pueden observarse en las **Fig. 19 y Fig. 20.**

Jerarquías

A partir de que la ejecución de los ensayos observa un ciclo semestral se decide añadir esta jerarquía a la tabla **dim_tiempo** (ver **Fig. 21**).

FIG. 19: TABLA DEL HECHO EFICACIA

FIG. 20: TABLA DEL HECHO SEGURIDAD

FIG. 21: TABLA ASOCIADA A LA DIMENSIÓN TIEMPO

Confeccionar modelo lógico

Primeramente se realizaron las uniones pertinentes entre las tablas de dimensiones y las tablas de hechos correspondientes. El análisis más detenido de este modelo lógico tentativo provocó que pasaran a conformar una sola tabla denominada **dim_codigo_ensayo** las tablas **dim_codigo_ensayo** y **dim_localización** porque ambas tienen como objetivo la identificación de los ensayos. El modelo lógico definitivo se muestra en la **Fig. 22**.

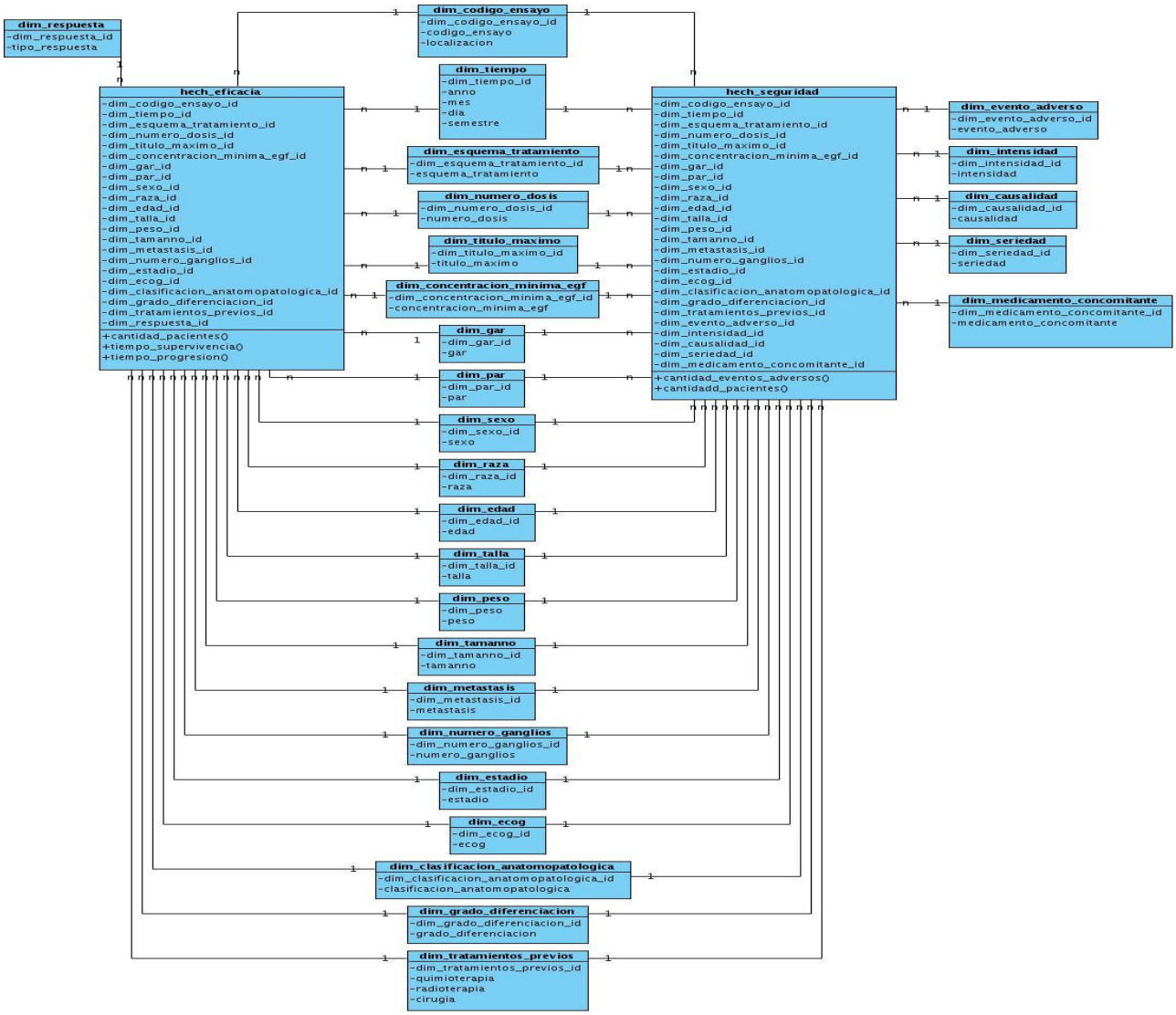

FIG. 22: MODELO LÓGICO

2.3 Conclusiones parciales

Como resultado del trabajo realizado se han identificado los dos objetivos decisionales a satisfacer por el CIM con el desarrollo del mercado de datos CIMAvax EGF. En correspondencia con los cuales se han identificado además 31 objetivos informacionales que involucran igual cantidad de variables. Se detectaron el momento han sido detectadas 29 reglas del negocio. Se diseñaron dos modelos conceptuales que relacionan cuatro indicadores y 14 perspectivas de análisis. A partir del estudio realizado sobre las fuentes de datos proporcionadas por el CIM se confeccionaron nueve diccionarios de datos. Esta labor permitió ubicar cada una de las variables previamente seleccionadas.

La definición del nivel de granularidad a nivel de paciente, el empleo de la tipología de esquema constelación de hechos, el diseño de 27 tablas de dimensiones y dos tablas de hechos permitieron confeccionar el modelo lógico.

De esta manera quedan sentadas las bases para emprender la etapa de ETL.

Capítulo 3: Proceso de Integración y Pruebas del Mercado de Datos CIMAvax EGF

3.1 Introducción

El proceso de extracción, transformación y carga es considerado vital en la construcción de un mercado de datos. La extracción de datos desde las fuentes; el aseguramiento de su calidad y consistencia; la reagrupación de manera tal que datos de orígenes distintos puedan usarse de manera transparente al usuario y la realización de las transformaciones pertinentes para la exitosa carga del mercado de datos, acarrean el mayor peso desde el punto de vista temporal para la investigación.

La descripción general del procedimiento aplicado para la realización del proceso de ETL del mercado de datos CIMAvax EGF puede encontrarse en el epígrafe **1.7.2 Procedimiento para la etapa de extracción, transformación y carga**.

3.2 Aplicación del procedimiento para la etapa de extracción, transformación y carga

Los sub epígrafes a continuación presentan los resultados obtenidos de cada flujo de actividad correspondiente a la etapa de ETL.

3.2.1 Análisis de la fuente de datos

Para llevar a cabo esta actividad se cuenta con los diccionarios de datos elaborados en la etapa de análisis y diseño durante la realización del modelo conceptual del mercado de datos. Se procede a agregar para cada campo de los mismos el formato de la fuente y el tipo de dato asociado con el que se corresponde el valoren en las fuentes de datos. En la **Tabla 4** se muestra el resultado de este paso para la sección Datos Demográficos del diccionario de datos FII Pulmón 056.

Perspectivas e indicadores	Significado en el negocio	Modelo(s) donde se encuentra	Correspondencia con variable(s) de los datos fuentes	Posibles valores	Tipo de dato	Formato de la fuente
Sexo	Indica el sexo del paciente	Tabla1	Columna H, Sexo	1: Masculino 2: Femenino	String	xls
Edad	Indica la edad del paciente	Tabla1	Columna G, Edad	Valor numérico entero positivo.	Integer	xls
Raza	Indica la raza del paciente	Tabla1	Columna I, ColorPiel	Blanca Negra Mestiza Amarilla	String	xls

Estadío	Indica el estadío en que se encuentra el paciente	Tabla2	Columna L, EstadioEnfermedad	Valor numérico entero entre 7(IIIB) y 8(IV), incluyendo los extremos.	Integer	xls
Clasificación anatomopato-lógica	Indica la clasificación del tumor al diagnosticar al paciente	Tabla2	Columna AR, ResultadoHistológico	Cadena de caracteres.	String	xls
Grado de diferenciación	Forma parte de la clasificación anatomopato-lógica	Tabla2	Columna AR, ResultadoHistológico	Cadena de caracteres.	String	xls
ECOG (Karnofsky)	Indica el estado general del paciente cuando se diagnostica	Tabla2	Columna K, EstadoGralOMS	Valor numérico entero entre 0 y 2, incluyendo los extremos.	Integer	xls
Tratamientos previos	Indica el tratamiento que ha recibido el paciente antes del ensayo	Tabla2	Columna P, CirugiaPrevia Columna R, RadioterapiaPrevia Columna T, QuimioterapiaPrevia	*VERDADERO en caso de haber realizado el tratamiento con anterioridad, FALSO en caso de que no lo haya hecho.	Boolean	xls

TABLA 4: SECCIÓN DATOS DEMOGRÁFICOS DICCIONARIO DE DATOS FII PULMÓN 056

3.2.2 Diseño de la arquitectura del mercado de datos CIMAvax EGF

A continuación se muestran los resultados de la ejecución de las actividades correspondientes al flujo de trabajo en cuestión.

Describir la arquitectura

El factor determinante en la elección de la arquitectura propuesta (ver **Fig. 23**) ha sido el alto nivel de complejidad y disgregación de las fuentes de datos lo que hizo preciso recurrir al *staging area* o área temporal de tráfico de datos como soporte al proceso de implementación del mercado de datos.

La descripción general del proceso quedaría como sigue: la información es extraída desde los sistemas operacionales, mayoritariamente ficheros .xls y .rec y es sometida a un primer proceso de ETL, que comprende labores de selección, perfilado y limpieza de datos, al cargarse hacia el *staging area.* Una de las principales ventajas del empleo del *staging area* reside en que el grado de alteración de las fuentes

originales se reduce considerablemente. Seguidamente se procede a la carga de las tablas de dimensiones y las tablas de hechos en el segundo proceso de ETL, que culmina con el mercado de datos poblado. La información técnica generada por la herramienta empleada durante el proceso ETL es registrada lo que proporciona una vista administrativa de metadados disponible para ser consultada por especialistas.

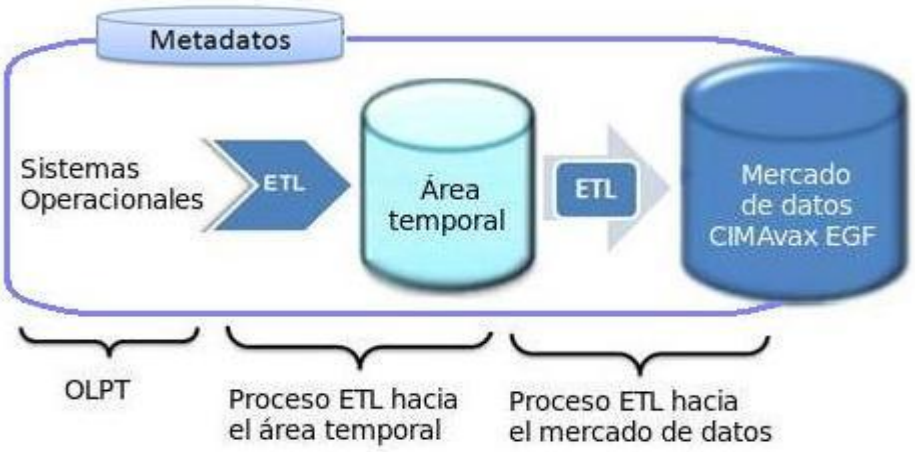

FIG. 23: ARQUITECTURA DEL MERCADO DE DATOS CIMAVAX EGF

Definir el modelo de despliegue

Como se puede apreciar en la **Fig. 24** el modelo de despliegue constituye un acercamiento a la forma de distribución física de los diferentes componentes del proceso de ETL. Se cuenta con un servidor donde se alojan las fuentes de datos de los EC del CIM. Al mismo se conecta, a través del protocolo TCP/IP, un ordenador personal en el que se encuentra instalado PDI este utiliza el protocolo JDBC para conectarse a los servidores de bases de datos, ya sea al *staging area,* o al mercado de datos.

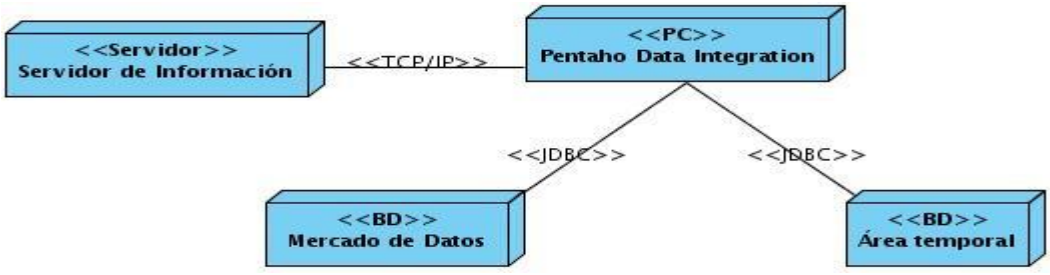

FIG. 24: MODELO DE DESPLIEGUE MERCADO DE DATOS CIMAVAX EGF

Como requerimientos críticos, acorde con el ambiente de trabajo se estipula que el ordenador sobre el que se ejecuta PDI deberá contar con un procesador Dual Intel Xeon 3GHz o equivalente, y RAM de hasta 4 GB. Debe garantizarse, además, la permanente disponibilidad de los servidores de bases de datos que deberán contar con al menos 40 GB de disco duro.

3.2.4 Desarrollo del modelo físico

Para dar cumplimento a las actividades de este flujo se confeccionaron 2 modelos físicos. El primero (ver **Fig. 25**) surge de la necesidad de dar soporte a la arquitectura que se escogió y modela el *staging area*; el segundo generado a partir del modelo lógico definido en la etapa de análisis y diseño de acuerdo con los requerimientos del negocio; el segundo (ver **Fig. 26**).

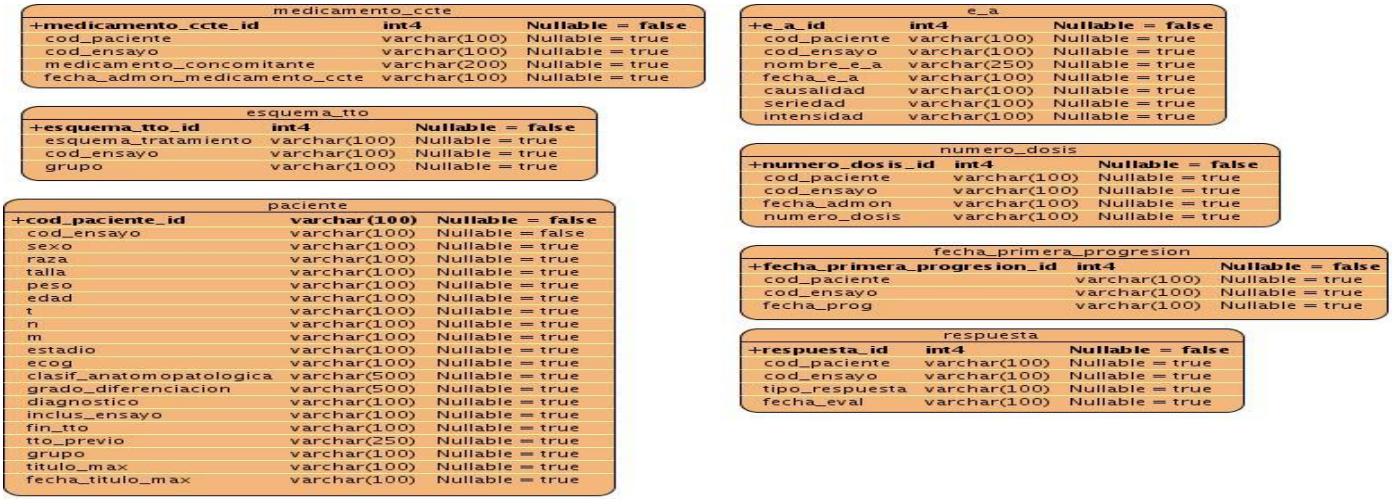

FIG. 25: MODELO FÍSICO DEL *STAGING AREA*

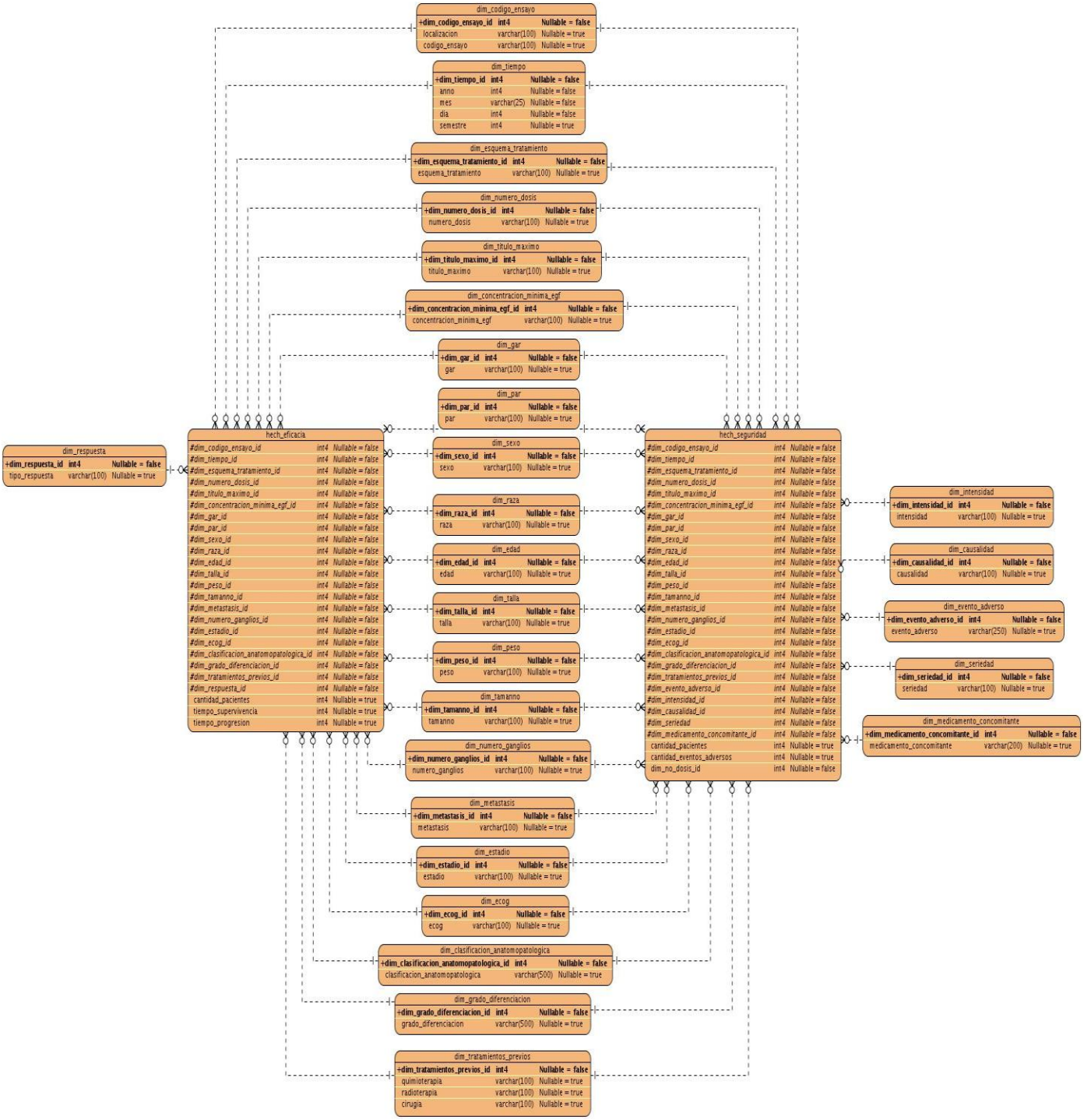

FIG. 26: MODELO FÍSICO DEL MERCADO DE DATOS CIMAVAX EGF

3.2.5 Extracción, transformación y carga de los datos al mercado de datos

La culminación exitosa del proceso de ETL depende en gran medida del cumplimiento de las buenas prácticas. A continuación se presentan, brevemente, algunas de las convenciones que se aplicaron en la elaboración de la presente solución.

Claves subrogadas

El empleo de las llaves subrogadas o sustitutas se justifica por la mejora del tiempo de respuesta de la base de datos, ya que la indexación de valores numéricos enteros resulta más eficiente que la de otros tipos de datos.

Propiedad	Valor
Nombre	dim_evento_adverso_id
Posición	1
Tipo de Dato	integer
Defecto	nextval('dim_evento_adverso_dim_evento_adverso_id_seq'::regclass)
Secuencia	dim_evento_adverso_dim_evento_adverso_id_seq
¿No Nulo?	Si
¿Clave Primaria?	Si

FIG. 27: CLAVE SUBROGADA DIM_EVENTO_ADVERSO_ID

Tratamiento de claves nulas

Para evitar la ocurrencia de claves nulas en las tablas de hechos, lo que provocaría problemas de integridad referencial, los valores nulos son sustituidos por valores preestablecidos en el paso hacia el *staging area*, ejemplo: NO REGISTRADO. En la dimensión correspondiente se ha añadido una fila con el mismo valor. De esta manera al encontrarse una clave nula se redirige a la fila destinada a tal efecto en la dimensión y es la clave de esta la que se añade en la tabla de hechos.

Metadatos

Para la satisfacción de este requerimiento se ha hecho empleo de una de las funcionalidades de la herramienta Spoon que permite configurar un repositorio de los registros generados durante la ejecución de las transformaciones. Evidentemente esta información será útil únicamente para personal familiarizado con los procesos de ETL.

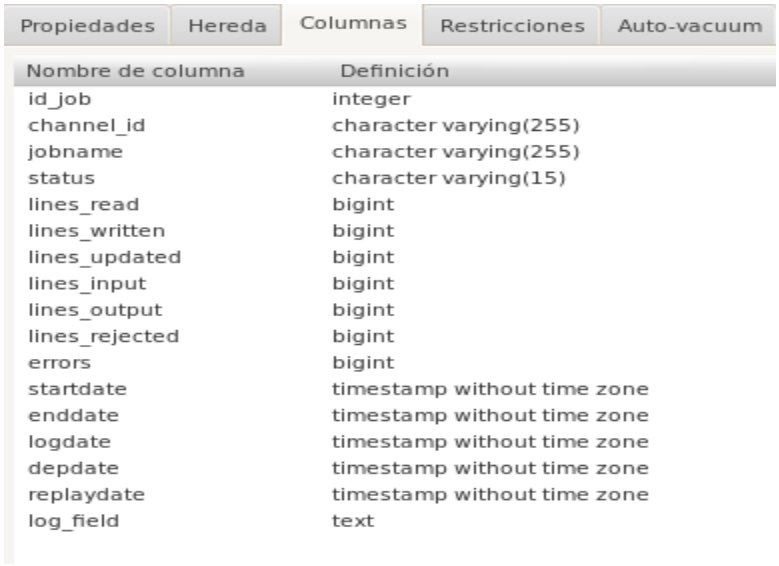

FIG. 28: VISTA DE TABLA EGF_LOG

Seguridad

Se creó el usuario "**kettle**" que es propietario de las bases de datos involucradas en el proceso de ETL y es el único autorizado a realizar operaciones sobre estas. "**kettle**" puede autenticarse únicamente desde el ordenador que ejecuta PDI.

```
TYPE  DATABASE    USER        CIDR-ADDRESS          METHOD
host    egfmart     kettle      192.168.4.6           md5
```

FIG. 29: VISTA CONFIGURACIÓN FICHERO PG_HBA.CONF

Para garantizar la seguridad del ambiente de desarrollo las transformaciones diseñadas desde el Spoon son almacenadas en un catálogo. Existen dos niveles de acceso al catálogo el nivel de administrador con todos los permisos habilitados y el nivel de invitado que tiene permisos de solo lectura (ver **Fig. 30**).

Definición de las transformaciones para el *staging area*

Las transformaciones definidas para esta etapa persiguen fundamentalmente la extracción de los datos de interés y el aseguramiento de la calidad y consistencia de los mismos. A continuación se ofrece una relación y breve descripción del empleo de algunas de estas.

FIG. 30: VISTA USUARIOS DE CATÁLOGO

Selección y nomenclatura de variables: Los ficheros .xls usados para el almacenamiento de los datos contienen el registro de hasta cientos de variables. Fue preciso seleccionar aquellas que son de interés para el mercado de datos. Por otra parte la denominación de algunas variables no se corresponde con su significado para el negocio. Para dar solución a esta problemática se utiliza el paso Selecciona/Renombra valores (ver **Fig.31**) del **Spoon**.

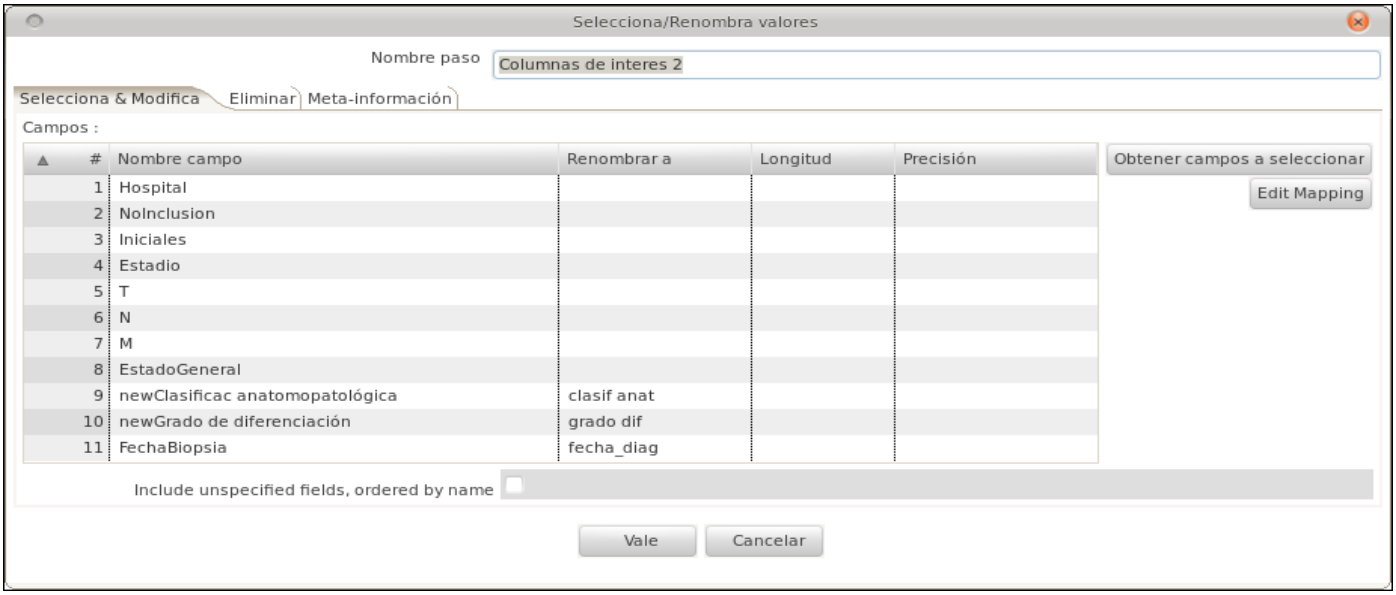

FIG. 31: VISTA DEL PASO SELECIONA/RENOMBRA VALORES

Tratamiento de los String y date: al realizar el perfilado de datos se detectaron una serie de irregularidades en el almacenamiento de ambos tipos de datos por lo que se hizo necesario la aplicación de un estándar. Para ellos se emplearon entre otros los pasos Operaciones sobre cadenas y Remplazar en una cadena (ver **Fig. 32**)**.**

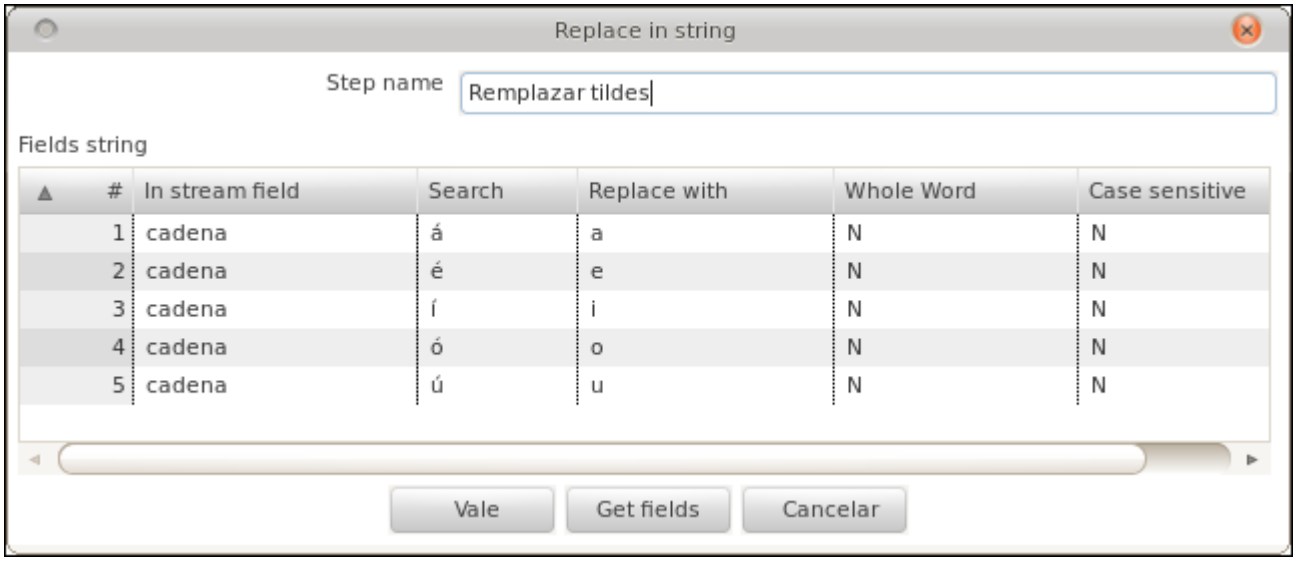

FIG. 32: VISTA DEL PASO REMPLAZAR EN UNA CADENA

Mapeo de valores: ocurre comúnmente que se utiliza un valor sustituto que es preciso mapear hacia el staging area. Por ejemplo en las fuentes la clasificación raza amarilla se almacenó pero los especialistas han considerado oportuno sumar las escasas ocurrencias de esta a la clasificación mestiza (ver **Fig. 33**).

FIG. 33: VISTA DEL PASO MAPEO DE VALORES

Carga de los datos hacia el staging area

Una vez hechas las transformaciones pertinentes se realiza el proceso de carga hacia el *staging area* con el empleo de la herramienta Kettle de la suite PDI. La base de datos del *staging area* corre sobre un servidor PostgreSQL y los detalles de la conexión desde el Kettle pueden ser observados en la **Fig. 35** a continuación.

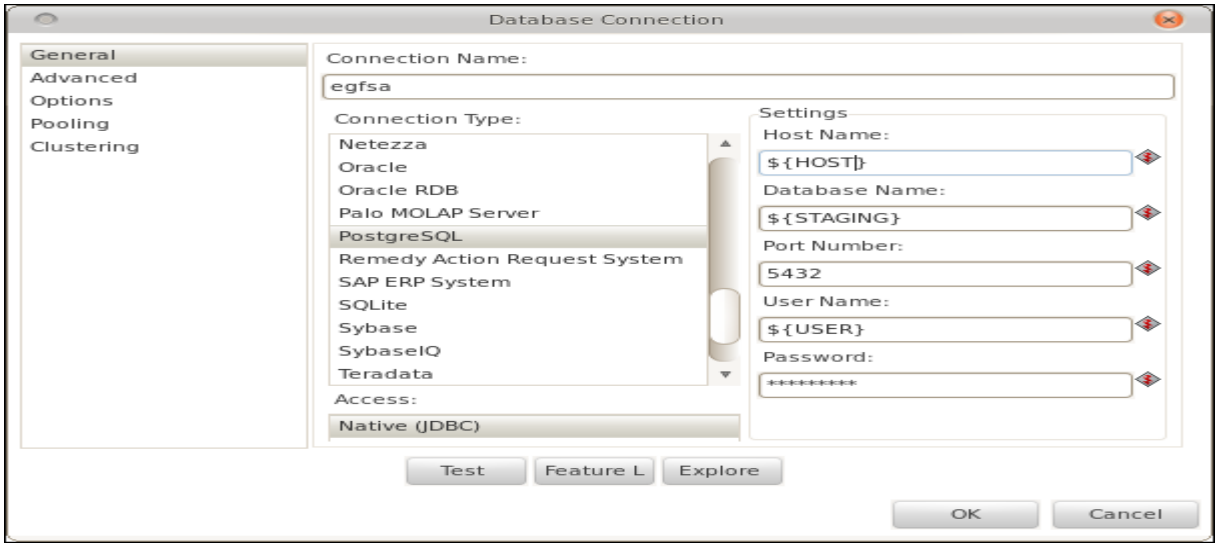

FIG. 34: CONFIGURACIÓN DE LA CONEXIÓN A LA BASE DE DATOS DEL *STAGING AREA*

En la **Fig. 35** se muestra la transformación realizada para la carga hacia la tabla paciente del *staging area* para el ensayo PIV Pulmón 062**.**

Definición de las transformaciones para el mercado de datos

Las transformaciones realizadas en esta etapa responden básicamente a las reglas del negocio previamente identificadas; así como particularidades del modelo dimensional y buenas prácticas en el diseño y construcción de mercados de datos.

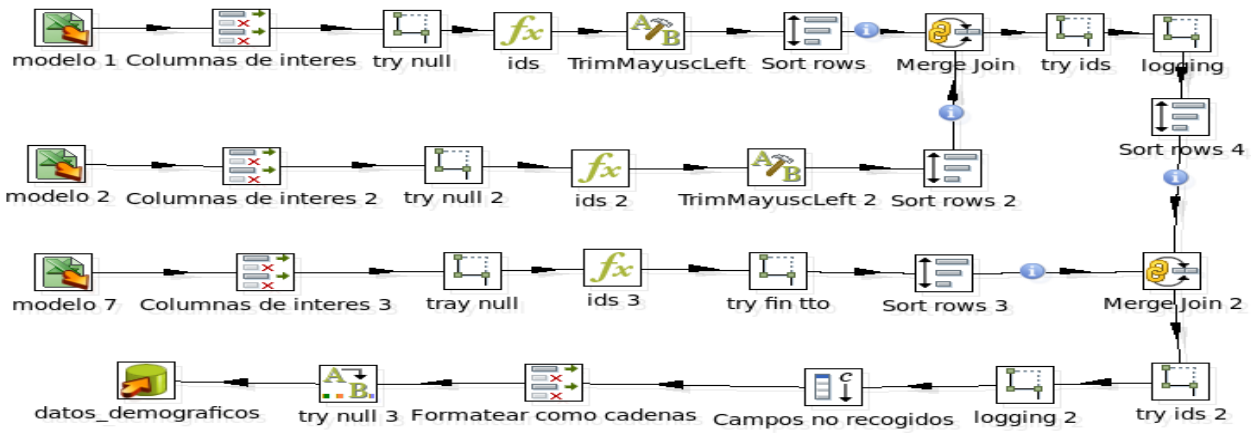

FIG. 35: VISTA TRANSFORMACION PACIENTE

Añadir valor constante: a través de este se tratan aquellas variables que por la naturaleza del ensayo tienen un valor constante, como puede ser el caso del sexo, o no han sido registradas (ver **Fig. 36**).

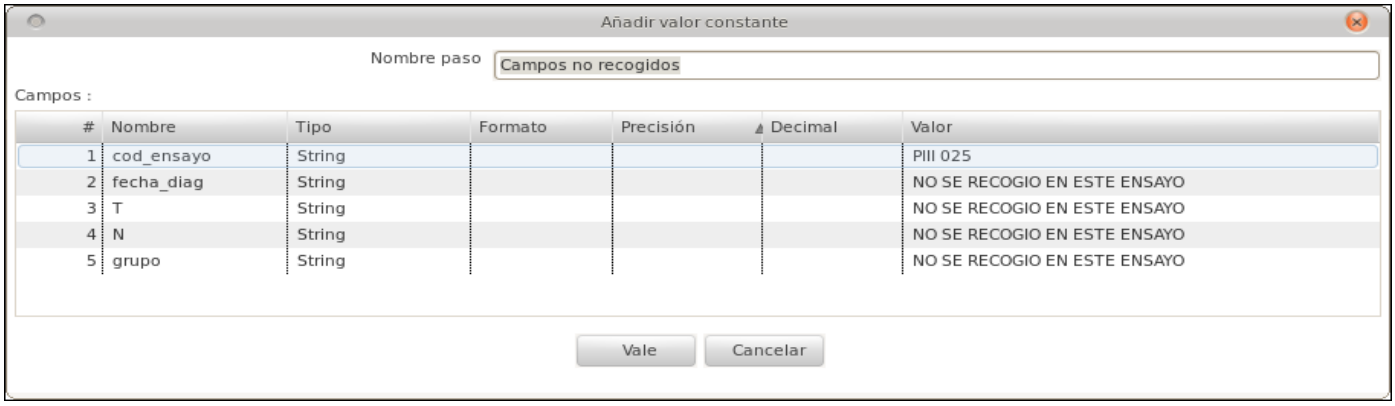

FIG. 36: VISTA PASO AÑADIR VALOR CONSTANTE

Fórmula: brinda una serie de funciones lógicas, matemáticas, de comparación y para operar con cadenas de texto, entre otras. En la **Fig. 37** se puede apreciar su empleo en la obtención de la diferencia entre dos fechas útil para el cálculo de la edad. También se utilizó para la conformación del código de identificación de los pacientes que se obtiene de la concatenación de tres campos.

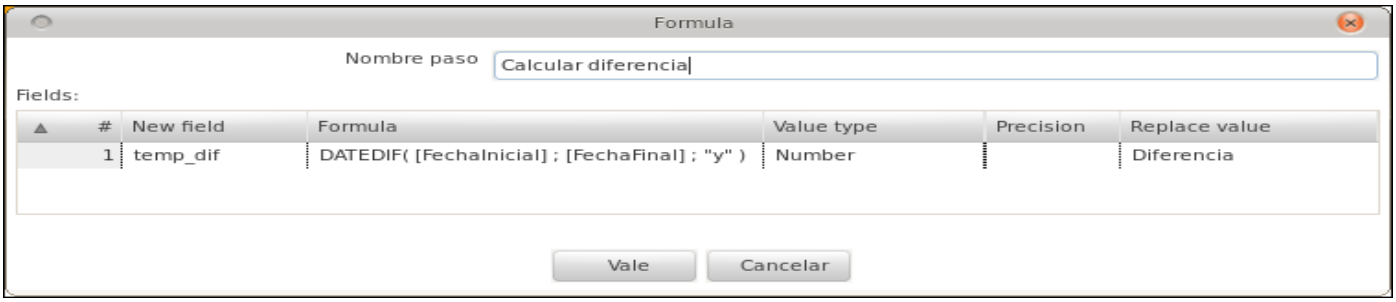

FIG. 37: VISTA PASO FÓRMULA CALCULAR DIFERENCIA

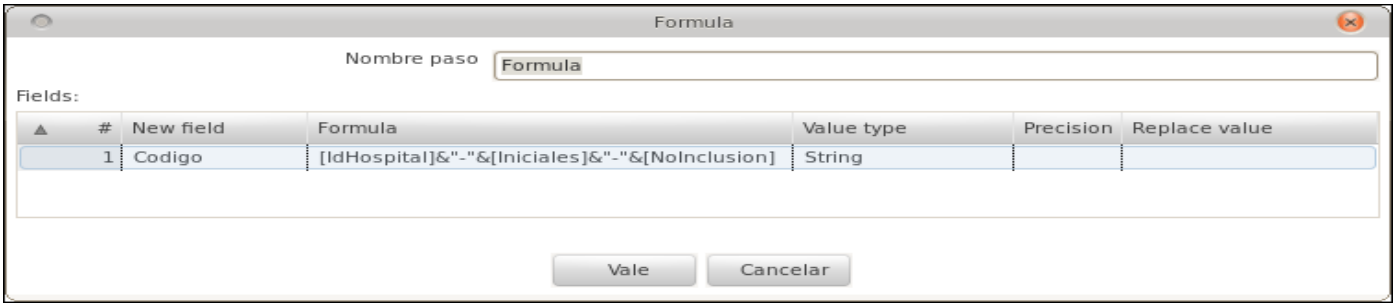

FIG. 38: VISTA PASO FÓRMULA CODIGO

Carga de los datos hacia el mercado de datos

Primeramente se ejecutan las transformaciones correspondientes a la carga hacia las dimensiones del mercado de datos y posterior a ello se realiza el llenado de las tablas de hechos. A continuación se muestra la transformación que se encarga de poblar la dimensión tiempo (**Fig. 49**).

FIG. 39: VISTA DIMENSION TIEMPO

El programa **Kitchen** permite ejecutar los trabajos diseñados gráficamente desde el **Spoon**. Los trabajos o jobs permiten el control de la secuencia de ejecución del proceso de transformaciones. Seguidamente se muestra el trabajo que guía la carga hacia el mercado de datos CIMAvax EGF.

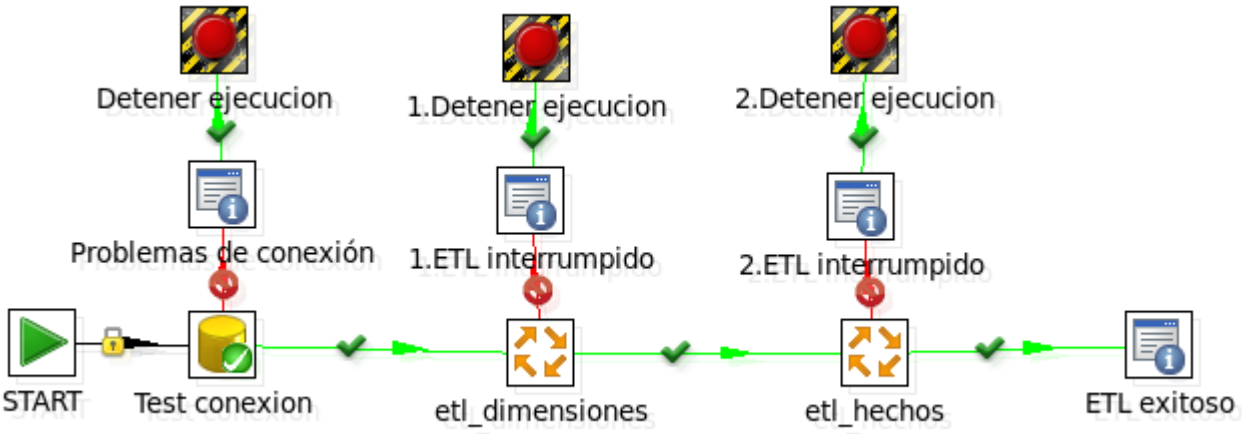

FIG. 40: JOB RECTOR DEL PROCESO DE ETL

El proceso de ETL para el mercado de datos CIMAvax EGF tiene la particularidad de no precisar de periodicidad en su actualización. De los nueve EC registrados en este ensayo seis se encuentran cerrados, lo que significa que no serán añadidos nuevos datos. En el caso de los están en curso aún la carga de los mismos se efectuará en la medida en que se termine la recogida de los datos desde los centros hospitalarios involucrados.

El curso natural del proceso de ETL comienza con la comprobación de la disponibilidad del servidor PostgreSQL que almacena las bases de datos del *staging area* y el mercado de datos CIMAvax EGF. Luego de esto procede a llamar el job que se encarga de coordinar el ETL de las dimensiones del mercado de datos y este al que guía el proceso de carga de las tablas de hechos. Es preciso tener en cuenta que ambos jobs contienen la secuencia de ejecución de las transformaciones tanto desde las fuentes originales al staging area como hacia el mercado de datos. Al finalizar el job exitosamente se muestra un mensaje de confirmación. En el caso de ocurrir alguna excepción, el paso en cuestión envía un mensaje de error.

3.3 Validación del proceso ETL

Se propone la validación del proceso a través de las siguientes listas de chequeo a los hitos del proceso ETL. El objetivo trazado es verificar y evaluar la fiabilidad de los datos cargados:

- ✓ Lista de chequeo del Modelo de datos.
- ✓ Lista de chequeo de los diccionarios de datos.
- ✓ Lista de chequeo de Registro de Sistemas Fuentes.

En la **Tabla 5** se pueden observar varias secciones de la lista de chequeo de los diccionarios de datos. A continuación se presenta una breve descripción de los componentes de la misma.

- ✓ Peso: Define si el indicador a evaluar es crítico o no.
- ✓ Evaluación (Eval): Es la forma de evaluar el indicador en cuestión. El mismo se evalúa de 1 en caso de mal y 0 en caso que elemento revisado no presente errores.
- ✓ Cantidad de elementos afectados: Especifica la cantidad de errores encontrados sobre el mismo indicador.
- ✓ Comentario: Especifica los señalamientos o sugerencias que quiera incluir la persona que aplica la lista de chequeo.

Estructura del documento					
Peso	Indicadores a Evaluar	Eval	(NP)	Cantidad de elementos afectados	Comentarios
	1. ¿Está el documento acorde a la plantilla estándar del proyecto o del expediente de proyecto?	0			Los DD forman parte del artefacto Desarrollo del modelo conceptual CIMAvax EGF
crítico	2. ¿Contiene las secciones obligatorias definidas en el expediente? (Ver Expediente de Proyecto)	0			
crítico	3. ¿Está especificado el nombre de la fuente de datos a la cual se le realiza el artefacto?	0			
	4. ¿Entre la referencia del documento se encuentre el Mapa Lógico de Datos?	0			
	5. ¿Se define el objetivo de la organización?	0			Se define el objetivo para la etapa correspondiente del proceso de implementación

Elementos definidos por la procedimiento					
Peso	Indicadores a Evaluar	Eval	(NP)	Cantidad de elementos afectados	Comentarios
Variable					
crítico	¿Las variables tienen al menos un valor?	0			
crítico	¿Cada una de las variables contiene su descripción y los valores posibles a tener?	0			
crítico	¿De cada uno de las variables están especificados todos los valores que toman, sus identificadores y sus significados?	0			En el caso de las variables cuyo rango de valores posibles es muy amplio se especifican los extremos únicamente.
	¿Todas las variables que se describen el documento están definidas en el mapa lógico de datos?	1			Una variable en el modelo de datos depende en ocasiones de varias de las relacionadas en los DD.

TABLA 5: VISTA DE SECCIONES DE LA LISTA DE CHEQUEO DE LOS DD

Como resultado del proceso revisión de las listas de chequeo fueron detectadas seis no conformidades relacionadas mayormente con la definición del rango de valores válidos para una variable, el tipo de datos asociado en el SGBD y la ubicación en las fuentes de datos. Dichas irregularidades fueron corregidas con celeridad.

3.4 Conclusiones parciales

Fue necesario realizar una reelaboración parcial del procedimiento utilizado en el CIM debido a las particularidades de la presente investigación

En este punto de la investigación se definió la arquitectura del mercado de datos. Además, se realizó la actualización de los diccionarios de datos con información necesaria para el proceso de ETL. Se confeccionaron dos modelos físicos; uno para el mercado de datos y otro para el *staging area* de acuerdo con la arquitectura propuesta.

Se realizó la configuración y carga del *staging area* con el objetivo de preservar el estado de las fuentes originales de datos. Como parte de este proceso se realizó la limpieza de datos y algunas transformaciones básicas. Finalmente, se realizó el proceso de ETL que permitió poblar el mercado de datos CIMAvax EGF y su posterior validación.

Conclusiones

Fruto de la continuidad de la colaboración entre el CIM y la UCI se realizó el proceso de ETL de un mercado de datos que tributa al almacenamiento homogéneo y estandarización de la información de los EC realizados sobre el producto CIMAvax EGF. Constituyeron hitos importantes en la realización de esta investigación:

A través de la aplicación del procedimiento de análisis y diseño se logró obtener la estructura del mercado de datos CIMAvax EGF.

La realización del proceso de ETL y su validación permitió la integración al mercado de datos de datos limpios, consistentes y estandarizados.

Recomendaciones

Con el propósito de extender los resultados de la presente propuesta de solución se recomienda:

- ✓ Realizar los procesos de Inteligencia de Negocio al mercado de datos CIMAvax EGF.
- ✓ Implementación de una capa de metadatos destinada a los usuarios finales.

Referencias Bibliográficas

[1] Salud, Organización Mundial de la. OMS. [En línea] [Citado el: 15 de 1 de 2011.] http://www.who.int/mediacentre/factsheets/fs297/es/index.html.

[2] Portal Centro de Inmunología Molecular. [En línea] [Citado: 2 de noviembre de 2010.] http://www.cim.co.cu/.

[3] MINSAP, CECMED. *Buenas Prácticas Clínicas en Cuba.* Ciudad Habana : s.n., 2000.

[4] INMON, Bill. *Building The Data Warehouse.* Canadá : John Wiley & Sons, Inc, 1996. pág 31.

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

[6] KIMBALL, Ralph, ROSS, Margy. *The Data Warehouse Toolkit*. Canada : John Wiley & Sons, Inc, 2002. pág. 10.

[7] KIMBALL, Ralph, ROSS, Margy. *The Data Warehouse Toolkit*. Canada : John Wiley & Sons, Inc, 2002. pág. 12.

[8] DATAPRIX. [En línea] [Citado el: 22 de diciembre de 2010.] http://www.dataprix.com/qu-es-un-data-warehouse.

[9] Espacio Logopédico. [En línea] [Citado el: 15 de diciembre de 2010.] http://www.espaciologopedico.com/recursos/glosariodet.php?Id=309.

[10] TermWiki. [En línea] [Citado el: 10 de enero de 2011.] http://www.termwiki.com/EN:star_schema.

[11] TermWiki. [En línea] [Citado el: 10 de enero de 2011.]

http://www.termwiki.com/EN:snowflake_schema.

[12] Datawarehouse4u.Info. [En línea] [Citado el: 10 de enero de 2011.]

http://datawarehouse4u.info/Data-warehouse-schema-architecture-fact-constellation-schema.html.

[13] KIMBALL, Ralph, ROSS, Margy. *The Data Warehouse Toolkit*. Canada : John Wiley & Sons, Inc, 2002. pág. 408.

[14] KIMBALL, Ralph, ROSS, Margy. *The Data Warehouse Toolkit*. Canada : John Wiley & Sons, Inc, 2002. pág. 8.

[15] HAWKINS, Humphries. *Data Warehousing Architecture and Implementation*. O'Reilly & Associates, Inc.pág 248.

[16]. **RUMBAUGH, James**. *El Lenguaje unificado de modelado. Manual de referencia.* Madrid : Educación Pearson, 2000.pág. 3.

[17] THE POSTGRESQL GLOBAL DEVELOPMENT GROUP. *PosgreSQL 8.4.1 Documentation.* University of California, 2009. pág. 14

[18] GILMORE, W. Jason, TREAT, Robert H. *Beginning PHP and PostgreSQL 8*. New York. Apress, 2006. pág 556.

[19] GARCIA FEAL, Miguel, CHOUCIÑO FERREIRO, Jose Luis. *Seguridad en Internet. SSL.* Madrid. pág. 11.

Bibliografía

Bases de datos. [En línea] [Consultado el: 30 de octubre de 2010.] http://basesdedatoss.blogspot.com/2008/12/que-son-las-bases-de-datos-relacional.html

Cubadebate, contra el Terrorismo Mediático. [En línea] [Consultado el: 10 de enero de 2011.] http://www.cubadebate.cu/noticias/2010/11/07/el-cancer-segunda-causa-de-muerte-en-cuba/

DÍAZ MORALES, Ing. Themis Patricia, BERMÚDEZ RODRÍGUEZ, Ing. José Salvador. *Diseño de un Almacén de datos para los Ensayos Clínicos que se gestionan en el Centro de Inmunología Molecular.* Tesis (Ingeniero en Ciencias Informáticas) Ciudad Habana. Universidad de las Ciencias Informáticas, 2010. [En línea] http://bibliodoc.uci.cu/TD/TD_02946_10.pdf

GONZÁLEZ HERNÁNDEZ, Delly Lien. EUMEDNET Enciclopedia y Biblioteca Virtual. [En línea] 2006. [Citado el: 1 de noviembre de 2010.]

MuyPymes. [En línea] [Consultado el: 30 de octubre de 2010.] http://muypymes.com/tecnologia/software/4826-las-diez-mayores-bases-de-datos-del-mundo.html

Portal Centro de Inmunología Molecular. [En línea] [Consultado el: 2 de noviembre de 2010.] http://www.cim.co.cu/invest_desa.asp

Primera Edición. El Diario de Misiones. [En línea] [Consultado el: 10 de enero de 2011.] http://www.primeraedicionweb.com.ar/nota/digital/23800/el-cancer-es-la-causa-de-muerte-con-mayor-impacto-economico-a-nivel-mundial.html

Registro Público Cubano de Ensayos Clínicos. [En línea] [Consultado el: 2 de noviembre de 2010.] http://registroclinico.sld.cu/ensayos-publicados/ensayos-por-promotores/cim

RODRÍGUEZ SOTOLONGO, Ing. Javier, PERALTA GÓNGORA, Ing. Yohan Orlando. *Implementación del proceso de extracción, transformación y carga de un Almacén de datos para los Ensayos Clínicos del Centro de Inmunología Molecular*. Tesis (Ingeniero en Ciencias Informáticas) Ciudad Habana. Universidad de las Ciencias Informáticas, 2010. [En línea] http://bibliodoc.uci.cu/TD/TD_02968_10.pdf

Vitadelia. Vida sana y nutrición. [En línea] [Consultado el: 10 de enero de 2011.] http://www.vitadelia.com/miscelanea/4-de-febrero-dia-mundial-del-cancer-cifras-en-la-ue

Object Management Group. [En línea] [Consultado el: 10 de noviembre de 2010.] http://www.omg.org/