Universidad de las Ciencias Informáticas

Facultad 6

Título: Implementación del proceso de extracción, transformación y carga de un Datawarehouse para los Ensayos Clínicos del Centro de Inmunología Molecular

Trabajo de Diploma para optar por el título de

Ingeniero en Ciencias Informáticas

Autores:

Javier Rodríguez Sotolongo

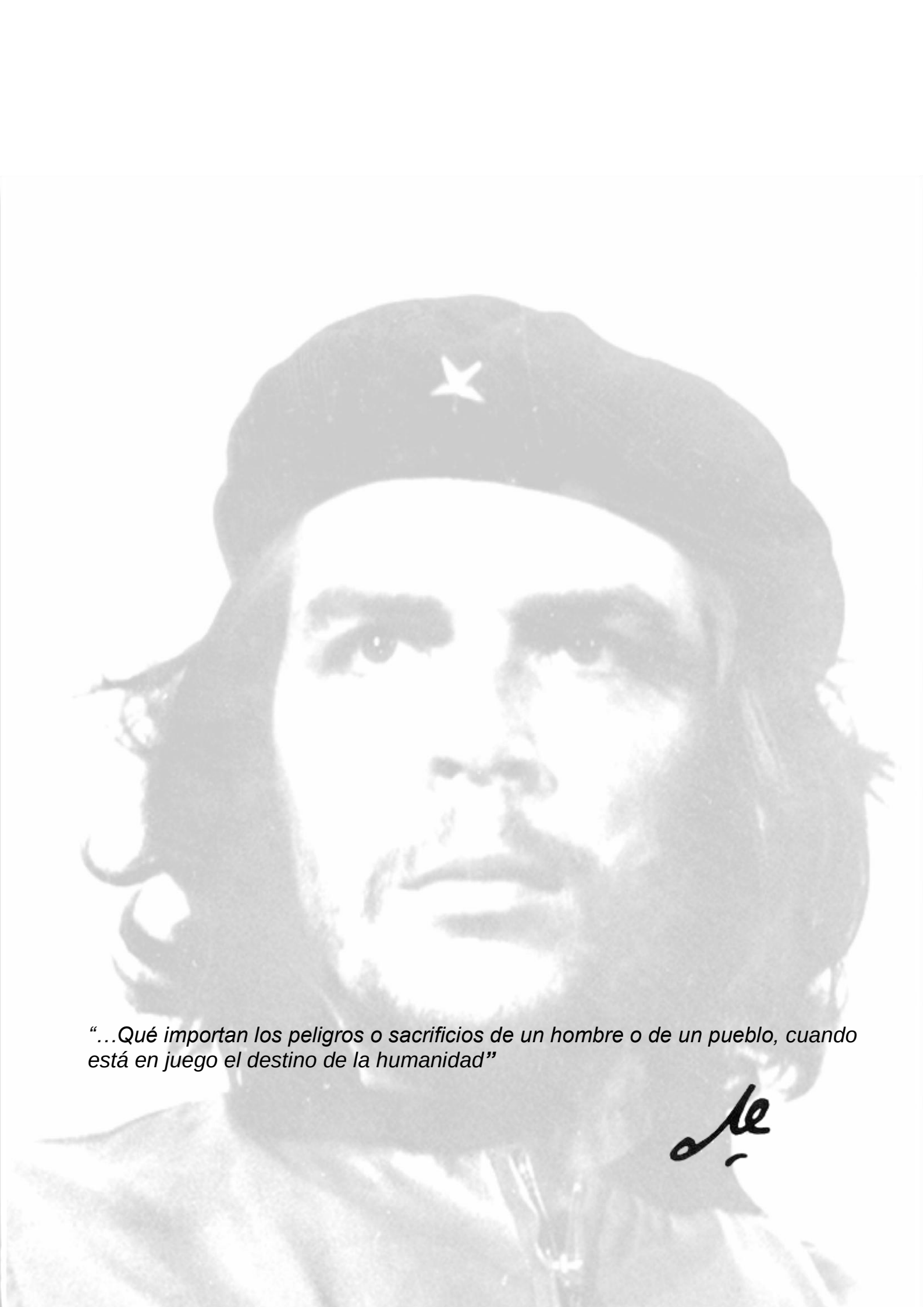
Yohan Orlando Peralta Góngora

Tutores:

Ing. Anthony Rafael Sotolongo León

Ing. Martha Denia Hernández Ramírez

CIUDAD DE LA HABANA, JUNIO 2010.



“…Qué importan los peligros o sacrificios de un hombre o de un pueblo, cuando está en juego el destino de la humanidad”

DECLARACIÓN DE AUTORÍA

Declaramos ser autores de la presente tesis y reconocemos 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 _____.

Javier Rodríguez Sotolongo

_____

Firma del Autor

Yohan Orlando Peralta Góngora

_____

Firma del Autor

Ing. Martha Denia Hernández Ramírez

_____

Firma del Tutor

Ing. Anthony Rafael Sotolongo León

_____

Firma del Tutor

Tutores:

Tutor: Ing. Martha Denia Hernández Ramírez
Especialidad de graduación: Ingeniería en Ciencias Informáticas
Categoría docente: Instructor en Adiestramiento
Categoría Científica: no
Años de experiencia en el tema: 0
Años de graduado: 1
Correo Electrónico: mdhernandez@uci.cu

Tutor: Ing. Anthony Rafael Sotolongo León
Especialidad de graduación: Ingeniería Informática
Categoría docente: Instructor
Categoría Científica: no
Años de experiencia en el tema: 0
Años de graduado: 3
Correo Electrónico: asotolongo@uci.cu

Agradecemos a Fidel, a la Revolución y a la Universidad de las Ciencias Informáticas por darnos la oportunidad de formarnos y hacer de nuestros sueños una realidad.

A nuestros padres, por su apoyo moral y espiritual en todo momento.

A nuestros tutores, que con su seriedad científica debidamente matizada de criticidad y humor, hicieron de esta experiencia académica algo muy satisfactorio.

A los profesores que han compartido sus experiencias durante estos años. Siempre les estaremos agradecidos.

A todos los que de una forma u otra, nos han ayudado durante la realización de esta investigación: amigos, compañeros, en fin gracias a todos.

RESUMEN

De manera proporcional al paso de los años, cada uno de los estudios sobre los Ensayos Clínicos (EC) que han tenido lugar en el Centro de Inmunología Molecular (CIM), ha hecho posible el almacenamiento de información de forma creciente. El volumen de datos recogidos ha permitido satisfacer las necesidades diarias en dicho centro, pero ha superado las capacidades humanas para analizar y transformar la información en conocimiento útil, que apoye la toma de decisiones en la práctica clínica, a la hora de satisfacer la demanda de información biomédica.

Por la importancia que tiene la optimización de los resultados de los EC en el CIM, se decidió el desarrollo de un Datawarehouse (DWH por sus siglas en inglés), por lo que el presente trabajo aborda la implementación del proceso de extracción, transformación y carga (ETL por sus siglas en inglés), teniendo como precedente el diseño del DWH, resultado de una investigación. Para esto se hizo un estudio bibliográfico sobre las tecnologías actuales y los conceptos fundamentales relacionados con el tema, además se propone un conjunto de actividades para llevar a cabo este proceso. Con el objetivo de evaluar el proceso de ETL, se propone una lista de chequeo ya que no existe en la bibliografía consultada una técnica de evaluación para esta etapa.

PALABRAS CLAVES:
Ensayos Clínicos, Centro de Inmunología Molecular, Datawarehouse, ETL

ÍNDICE

INTRODUCCIÓN

En la actualidad existen diversos centros de investigaciones científicas dedicados a la creación de fármacos para la prevención y/o tratamiento de las enfermedades que hasta hoy no tienen cura, como el cáncer y el Virus de Inmunodeficiencia Humana (VIH), entre ellos el Centro Nacional de Investigaciones Científicas de París (CNRS por sus siglas en inglés) y el Centro de Investigación e Información de Medicamentos y Tóxicos de Panamá. Para la realización de cada uno de estos medicamentos se crean los EC. "*Un EC es la investigación efectuada en seres humanos, con el fin de determinar o confirmar los efectos clínicos, farmacológicos, y/o demás efectos farmacodinámicos, y/o de detectar las reacciones adversas, y/o de estudiar la absorción, distribución, metabolismo y eliminación de uno o varios medicamentos en investigación con el fin de determinar su inocuidad y/o su eficacia*" (1).

En el mundo los principales productores de medicamentos y fármacos se encuentran en países desarrollados como son los Estados Unidos (E.E.U.U.), Canadá y Francia, entre otros. Estos medicamentos no son accesibles para muchos países tercermundistas debido a sus altos precios y restricciones. Por estas razones algunos países del tercer mundo, tales como Panamá y Cuba se inclinaron hacia este ámbito, con el fin de crear nuevos medicamentos y abastecer la demanda de los mismos, de manera tal que los más pobres puedan adquirirlos. En nuestro país existen varios centros que investigan en este campo de la biotecnología, dentro de los cuales se encuentra el CIM. "*El CIM fue inaugurado el 5 de Diciembre de 1994 en el oeste de la Habana con el objetivo de obtener y producir nuevos biofármacos, destinados al tratamiento del cáncer y otras enfermedades crónicas no transmisibles e introducirlos en la Salud Pública Cubana, además de hacer la actividad científica y productiva económicamente sostenible y realizar aportes importantes a la economía del país*" (2).

"En este centro actualmente se conducen más de 30 EC de tratamiento de cáncer en Cuba y más de 15 en el exterior, incluyendo ensayos en varios países altamente industrializados y acuerdos comerciales con más de 47 países" (3). Para realizar el proceso de digitalización de estos estudios los especialistas diseñan varios modelos mediante el sistema EpiData; lo que significa que los diseños de los modelos en los Cuadernos de Recogida de Datos (CRD) no sean iguales, incluso las variables con las que se tratan los mismos términos médicos son distintas y se almacenan de manera diferente. Además, el uso del sistema EpiData hace que la forma de generar los reportes no se mantenga de manera uniforme, ya que la información se puede exportar en distintos formatos (Text, dBase III, Excel, Stata, SPSS y SAS). Cada vez que se generan dichos reportes generalmente se crean 12 ficheros por cada EC. La realización de cada uno de estos estudios genera una gran cantidad de datos debido al

alto número de personas que involucran. Lo que significa que el análisis de esta documentación se convierte en un verdadero problema a la hora de satisfacer la demanda de información biomédica para la toma de decisiones en la práctica clínica; así como la necesidad de involucrar una gran cantidad de personas y tiempo para este estudio, lo que trae como consecuencia que una vez terminada de procesar dicha información y de generar los reportes respectivos, muchas de las decisiones ya son obsoletas.

La situación anteriormente descrita demuestra que la información que genera cada uno de los EC que se gestionan en el CIM no se encuentra integrada, lo que atenta contra el mejor desempeño del análisis estadístico para la toma de decisiones y por ende que se dificulte la manera de predecir hacia dónde dirigir cada uno de estos estudios en dicho centro.

Por la importancia que tiene obtener mejores resultados de los EC en el CIM, así como para incrementar el avance de la medicina en Cuba; se hace necesario resolver la integración de toda la información que se maneja en este centro, ayudando de esta forma a los especialistas en sus análisis y permitiéndoles un acceso directo a toda la información. Para esto existe un Trabajo de Diploma precedente que está orientado al modelado de la integración de los datos, por lo que se plantea como **problema científico**: ¿cómo lograr la integración de los datos de los Ensayos Clínicos que se gestionan en el Centro de Inmunología Molecular?

La investigación tiene como **objeto de estudio** el proceso de desarrollo de Datawarehouse, enmarcado en el **campo de acción** proceso de extracción, transformación y carga de un Datawarehouse para los Ensayos Clínicos que se gestionan en el Centro de Inmunología Molecular.

El **objetivo general** de este trabajo es implementar el proceso de extracción, transformación y carga de un Datawarehouse para contribuir a un mejor análisis de los datos de los Ensayos Clínicos del Centro de Inmunología Molecular.

En correspondencia con ello se plantean como **objetivos específicos:**

- Elaborar un procedimiento para implementar el proceso de extracción, transformación y carga de un Datawarehouse que integre los datos de los Ensayos Clínicos que se gestionan en el Centro de Inmunología Molecular.
- Diseñar la arquitectura del Datawarehouse para los Ensayos Clínicos del Centro de Inmunología Molecular.

- Implementar el proceso de extracción, transformación y carga de los datos de los Ensayos Clínicos del Centro de Inmunología Molecular al Datawarehouse.
- Elaborar una lista de chequeo para evaluar el proceso de extracción, transformación y carga.
- Evaluar el proceso de extracción, transformación y carga realizado a través de la aplicación de la lista de chequeo.

Para dar cumplimiento a los objetivos específicos se definen las siguientes **tareas de la investigación:**

- Estudio, análisis y selección de una de las metodologías de arquitectura de Datawarehouse.
- Estudio, análisis y selección de lo mejor y más aplicable de cada una de las metodologías de desarrollo existentes para el proceso de extracción, transformación y carga de un Datawarehouse.
- Estudio y selección de las herramientas informáticas existentes en el mundo para realizar el proceso de extracción, transformación y carga.
- Estudio y selección del gestor de base de datos a utilizar para el almacenamiento del Datawarehouse.
- Elaboración de un procedimiento para implementar el proceso de extracción, transformación y carga.
- Validación del procedimiento propuesto mediante el método Delphi.
- Diseño de la arquitectura del Datawarehouse.
- Elaboración del Modelo Físico del *Staging area*.
- Elaboración del Modelo Físico del Data Mart.
- 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 *Staging area*.
- Realización de la carga de datos transformados hacia el *Staging area*.
- Creación de la base de datos del Datawarehouse.
- 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 Datawarehouse.
- Realización de la carga de datos transformados hacia el Datawarehouse.
- Elaboración de una lista de chequeo para evaluar la implementación del proceso de extracción, transformación y carga.

- Evaluación de la implementación del proceso de extracción, transformación y carga a partir de la aplicación de la lista de chequeo.

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

En el **Capítulo 1: Fundamentos teóricos sobre el proceso de extracción, transformación y carga**
Se hace un análisis del estado del arte del objeto de estudio, se investiga acerca del proceso de ETL, herramientas informáticas para la integración de los datos, sistemas gestores de base de datos y se fundamentan las metodologías existentes para la arquitectura y desarrollo del proceso de ETL.

En el **Capítulo 2: Implementación del proceso de extracción, transformación y carga**
Se presenta un procedimiento a partir de las características del negocio y los aspectos más significativos de cada metodología analizada; teniendo en cuenta la metodología de arquitectura propuesta por Kimball. Además se presentan los resultados de la implementación del proceso de ETL para lograr un DWH orientado a los EC del CIM.

En el **Capítulo 3: Evaluación de la implementación del proceso de extracción, transformación y carga**
Teniendo en cuenta que en la bibliografía consultada no se encontró una manera de evaluar la implementación del proceso de ETL, se propone elaborar una lista de chequeo que cumpla con dicho fin, para evaluar posteriormente la implementación del proceso de ETL a partir de la aplicación de la lista de chequeo elaborada.

CAPÍTULO 1: FUNDAMENTOS TEÓRICOS SOBRE EL PROCESO DE EXTRACCIÓN, TRANSFORMACIÓN Y CARGA

Introducción

En este capítulo se presenta el estado de la gestión de los EC del CIM, así como la dificultad que trae el mismo a la hora de realizar un análisis para la toma de decisiones en posteriores estudios. Se analizan las herramientas, tecnologías, arquitecturas y metodologías aplicables al proceso de ETL para la migración de los datos hacia un DWH para el CIM. También se aborda la manera de evaluar este proceso mundialmente.

1.1 Proceso de digitalización de los Ensayos Clínicos en el Centro de Inmunología Molecular

El CIM tiene como principal misión la búsqueda de nuevos productos para el diagnóstico y tratamiento del cáncer. Las líneas básicas de su investigación están concentradas en la inmunoterapia del cáncer, especialmente en el desarrollo de vacunas moleculares, ingeniería de anticuerpos, ingeniería celular, bioinformática y regulación de la respuesta inmune. Este centro realiza EC en hospitales cubanos altamente especializados en el diagnóstico de tumores por imágenes y tratamiento de cáncer de diferentes orígenes. Para llevar el control de estos estudios los especialistas diseñan los CRD, en los cuales se recoge toda la información relacionada con el paciente durante su tratamiento. Una vez culminado dicho estudio se envían los cuadernos para el CIM, lugar donde se realiza el proceso de digitalización de la información almacenada en los cuadernos mediante el sistema EpiData, generándose reportes en diferentes formatos (Text, dBase III, Excel, Stata, SPSS y SAS). Generalmente cuando se concluye este proceso se generan aproximadamente 12 ficheros por cada EC aplicado a una localización determinada del cuerpo humano. Un ejemplo es el producto Nimotuzumab (conocido por su nombre comercial como hR3) perteneciente a un EC que fue aplicado en pacientes con cáncer de mama, esófago, cabeza y cuello, próstata, hígado, páncreas, pulmón, cuello de útero y glioma en pacientes pediátricos y adultos, del cual se obtuvo un total aproximado a los 120 ficheros de datos.

La situación descrita anteriormente hace que se dificulte cada día más la manera de realizar los análisis estadísticos de los distintos EC en el CIM; corriéndose el riesgo de que se pierda información útil al no contar con una forma ágil que facilite la integración de los datos y contribuya a elevar la efectividad del tratamiento de la información. Por todas estas razones este trabajo está enmarcado en implementar el proceso de extracción, transformación y carga de un DWH, con el objetivo de contribuir

a un mejor análisis de los datos de los EC del CIM. Es válido mencionar que todo este proceso de integración se le realizará únicamente a las diez Bases de Datos (BD) de los EC pertenecientes al producto hR3, aplicado específicamente en las siguientes localizaciones del cuerpo humano:

- Cabeza y Cuello
- Glioma
- Esófago
- Metástasis cerebral
- Tumores sólidos
- Mama

En correspondencia con estas localizaciones se tienen las siguientes BD:

- *C y C 040*
- *C y C 046*
- *C y C 055*
- *C y C 076*
- *Glioma 053*
- *Glioma 069*
- *Esófago 075*
- *Metacerebral 079*
- *T.Sólidos 035*
- *Mama 070*

1.2 Datawarehouse

Un DWH es una BD corporativa que se caracteriza por integrar y depurar información desde una o varias fuentes de datos. De esta manera se facilita el análisis de la información desde infinidad de perspectivas y con grandes velocidades de respuesta. Por tales razones se podría ver también a un almacén de datos como una enciclopedia especializada en los temas que afectan el quehacer de la empresa.

1.2.1 Metodologías de arquitectura

Para desarrollar la arquitectura de un DWH existen dos estrategias metodológicas, la de Bill Inmon y la de Ralph Kimball. Teniendo en cuenta que ambas son muy usadas para el desarrollo de almacenes de datos, se propone realizar un estudio de estas dos vertientes, con el fin de seleccionar la más apropiada.

1.2.1.1 Definición de Bill Inmon

El término DWH fue acuñado por primera vez por Bill Inmon, quien define un DWH en términos de las características del repositorio de datos. *"Un DWH es una colección de datos orientada a un determinado ámbito (empresa, organización), integrado, temático, histórico y no volátil"* (4).

***Integrado*:** se refiere a que los datos almacenados en el DWH deben integrarse en una estructura consistente, por lo que las inconsistencias existentes entre los diversos sistemas operacionales deben ser eliminadas. La información suele estructurarse también en distintos niveles de detalles para adecuarse a las distintas necesidades de los usuarios.

***Temático*:** con esto se plantea que sólo los datos necesarios para el proceso de generación del conocimiento del negocio se integran desde el entorno operacional. Los datos se organizan por temas para facilitar su acceso y entendimiento por parte de los usuarios finales.

***Histórico*:** esto significa que el tiempo es parte implícita de la información contenida en un DWH. En los sistemas operacionales, los datos siempre reflejan el estado de la actividad del negocio en el momento presente. Por el contrario, la información almacenada en el DWH sirve, entre otras cosas, para realizar análisis de tendencias. Por lo tanto, el DWH se carga con los distintos valores que toma una variable en el tiempo para permitir comparaciones.

***No volátil*:** con esto se refiere a que el almacén de información de un DWH existe para ser leído, pero no modificado. La información es por tanto permanente, significando la actualización del DWH la incorporación de los últimos valores que tomaron las distintas variables contenidas en él, sin ningún tipo de acción sobre lo que ya existía.

Inmon define una metodología descendente (top-down) a la hora de diseñar un almacén de datos, ya que de esta forma se considerarán mejor todos los datos corporativos. En esta metodología los Data Marts (datos de un área de negocio específica) se crean después de haber terminado el DWH completo de la organización.

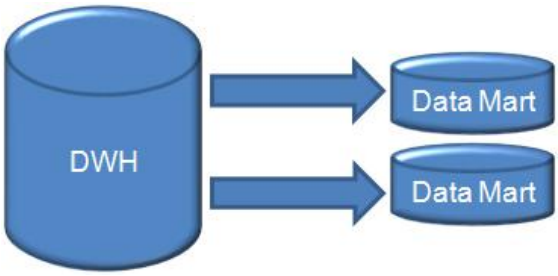


Figura 1. Arquitectura de Bill Inmon, componentes

1.2.1.2 Definición de Ralph Kimball

Kimball es otro conocido autor en el tema de los DWH, define un almacén de datos como: *"una copia de las transacciones de datos específicamente estructurada para la consulta y el análisis"* (5). También fue Kimball quien determinó que un DWH no era más que: *"la unión de todos los Data Marts de una entidad"* (5). Defiende por tanto una metodología ascendente (bottom-up) a la hora de diseñar un almacén de datos.

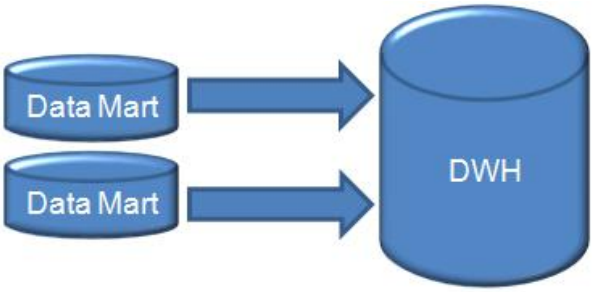


Figura 2. Arquitectura de Ralph Kimball, componentes

Después de haber estudiado y analizado ambas metodologías de arquitectura, no se puede decir que existe una aproximación equivocada entre estas dos vertientes, pues solamente representan dos formas distintas de implementar un DWH. Lo que se puede afirmar es que Inmon propone una metodología descendente, por lo que resulta más compleja y que se requiera más tiempo para su implementación, es por esto que en muchos casos fracasa por falta de paciencia y de compromiso. Sin embargo Kimball brinda una estrategia ascendente, de fácil comprensión y rápida de implementar por etapas, debido a estas razones se decide utilizar la arquitectura de DWH propuesta por Kimball.

1.3 Metodologías para el proceso de extracción, transformación y carga

El proceso de ETL está comprendido dentro del desarrollo de un DWH, por lo que no se encuentra en la literatura una metodología específica para realizar dicho proceso. En la presente investigación

solamente se tuvieron en cuenta aquellas metodologías de desarrollo de DWH que hacen referencia dentro de su desarrollo al proceso de ETL.

1.3.1 Ciclo de vida Kimball

El ciclo de vida Kimball propone cinco fases para su desarrollo: planeamiento, requerimientos, análisis y diseño, construcción y como última, la fase de implementación. Dentro de la fase de construcción se diseña y desarrolla el proceso de ETL, siendo esta la etapa de mayor importancia dentro del ciclo de vida Kimball; debido a que los datos en bruto son extraídos de los sistemas operacionales y transformados en información significativa para el negocio. En esta etapa los procesos de ETL deben ser diseñados mucho antes que cualquier dato sea extraído de la fuente y se verifica continuamente la calidad de los datos de entrada.

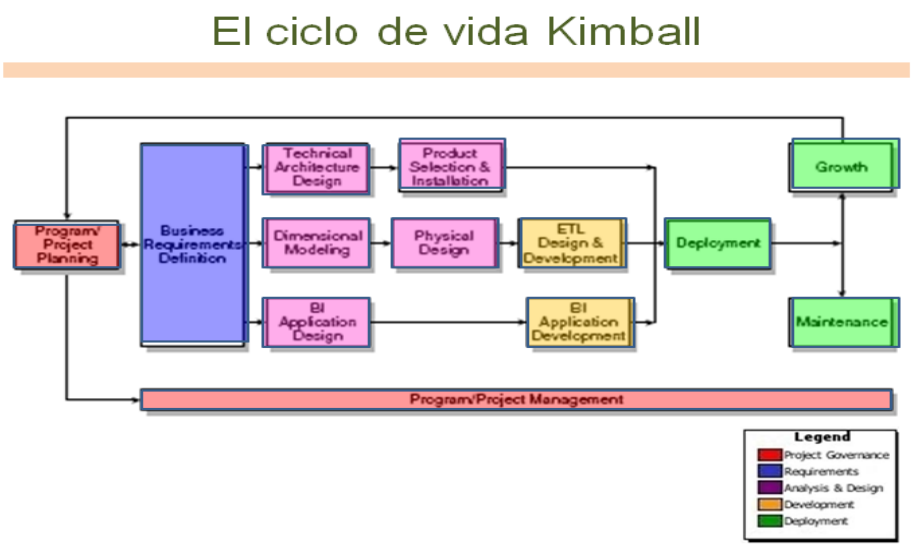


Figura 3. Ciclo de vida Kimball

1.3.2 Metodología Hefesto

Hefesto es una metodología propia, cuya propuesta está fundamentada en una muy amplia investigación, comparación de metodologías existentes y experiencias propias en procesos de confección de almacenes de datos. La metodología HEFESTO, puede ser embebida en cualquier ciclo de vida de cualquier metodología de desarrollo de software.

La metodología Hefesto puede resumirse a través del siguiente gráfico:

Figura 4. Metodología Hefesto, pasos

En la etapa número cuatro (procesos ETL) de la metodología de Hefesto se definen los procesos de extracción, transformación y carga de los datos fuentes, que poblarán y actualizarán el DWH. En esta etapa antes de realizar la carga de datos, es conveniente efectuar una limpieza de los mismos para evitar valores faltantes y anómalos. Al generar los ETL, se debe tener en cuenta cuál es la información que se desea almacenar en el depósito de datos; para ello se pueden establecer condiciones adicionales y restricciones. Estas condiciones deben ser analizadas y llevadas a cabo con mucha prudencia para evitar pérdidas de datos importantes.

Esta metodología plantea que cuando se trabaja con un esquema constelación, hay que tener presente que varias tablas de dimensiones serán compartidas con diferentes tablas de hechos, ya que puede darse el caso de que algunas restricciones aplicadas sobre una tabla de dimensión en particular para analizar una tabla de hechos, se puedan contraponer con otras restricciones o condiciones de análisis de otras tablas de hechos. Primero se cargarán los datos de las dimensiones y luego los de las tablas de hechos, teniendo en cuenta siempre, la correcta correspondencia entre cada elemento. En el caso en que se esté utilizando un esquema copo de nieve, cada vez que existan jerarquías de dimensiones, se comenzarán cargando las tablas de dimensiones del nivel más general al más detallado (6).

1.3.3 Selección de los mejores elementos expuestos en la metodología Hefesto y el ciclo de vida Kimball

Teniendo en cuenta que cada una de las metodologías consultadas no definen una manera de cómo realizar el proceso de ETL, optar por alguna de ellas no facilitará la implementación de dicho proceso. Por esta razón se propone emplear lo mejor de cada metodología ya que las mismas no dicen cómo, pero si destacan aspectos importantes a considerar en el momento de implementar el proceso de ETL. A continuación se relacionan los aspectos tomados de la metodología Hefesto y el ciclo de vida Kimball:

De la metodología Hefesto se tomaron seis aspectos que se deben tener en cuenta para implementar el proceso de ETL, de los cuales tres resultaron críticos (no se pueden obviar):

1. Antes de realizar la carga de datos, es conveniente efectuar una limpieza de los mismos, para evitar valores faltantes y anómalos.
2. Cuando se trabaja con un esquema constelación, hay que tener presente que varias tablas de dimensiones serán compartidas con diferentes tablas de hechos, ya que puede darse el caso de que algunas restricciones aplicadas sobre una tabla de dimensión en particular para analizar una tabla de hechos, se puedan contraponer con otras restricciones o condiciones de análisis de otras tablas de hechos.
3. Primero se cargarán los datos de las dimensiones y luego los de las tablas de hechos, teniendo en cuenta siempre, la correcta correspondencia entre cada elemento.
4. En el caso en que se esté utilizando un esquema copo de nieve, cada vez que existan jerarquías de dimensiones, se comenzarán cargando las tablas de dimensiones del nivel más general al más detallado.
5. Cuando se haya cargado en su totalidad el DWH, se deben establecer sus políticas de actualización o refresco de datos.
6. Debe tenerse en cuenta, que no siempre la clave primaria del OLTP, se corresponde con la clave primaria de la tabla de dimensión relacionada. Es recomendable manejar un sistema de claves en el DWH totalmente diferente al de los OLTP, ya que si estos últimos son recodificados, el almacén quedaría inconsistente y debería ser poblado nuevamente en su totalidad.

Del ciclo de vida Kimball se tomaron siete condiciones que se deben tener en cuenta para implementar el proceso de ETL, de las cuales dos resultaron de carácter crítico:

7. La arquitectura definida debe responder a las necesidades del proyecto.

8. La arquitectura debe soportar el incremento del proyecto.
9. Con el fin de no afectar en gran medida a los sistemas de origen de datos en cuanto al consumo de recursos; la carga de los datos hacia *Staging area* se realiza lo más rápido posible. Por esta razón se deben utilizar el menor número de transformaciones para poblar el *Staging area*.
10. El Modelo Físico se obtiene a partir del Modelo Lógico.
11. La implementación del proceso de ETL tiene que cumplir con la arquitectura definida.
12. Se debe tener en cuenta los formatos fuentes y tipos de datos de las perspectivas de análisis antes de iniciar el proceso de ETL.
13. La extracción de los datos se debe realizar a partir de las fuentes de datos.

Después de haber realizado un estudio donde se seleccionaron los aspectos más importantes expuestos en la metodología Hefesto y el ciclo de vida Kimball; se concluyó obteniendo un listado de 13 aspectos (cinco críticos) que deben tenerse en cuenta en la implementación del proceso de ETL.

1.4 Extracción, transformación y carga

Para poder extraer los datos desde los OLTP, para luego manipularlos, integrarlos, transformarlos y posteriormente cargar los resultados obtenidos en el DWH, es necesario contar con algún proceso que se encargue de ello. Precisamente, el proceso de ETL será el que cumplirá con tal fin, el mismo tiene como precedente una etapa de diseño que servirá como punto de partida para su implementación.

El proceso de ETL se encarga de extraer los datos desde las diversas fuentes que se requieran, los transforman para resolver posibles problemas de inconsistencias entre los mismos y finalmente, después de haberlos depurado se procede a su carga en el depósito de datos. En síntesis, las funciones específicas de los ETL son tres: la extracción, transformación y carga.

1.4.1 Extracción

La primera parte del proceso ETL consiste en extraer los datos desde los sistemas de origen. La mayoría de los proyectos de almacenamiento de datos fusionan datos provenientes de diferentes sistemas de origen. Cada sistema separado puede usar una organización diferente de los datos o formatos distintos. Los formatos de las fuentes normalmente se encuentran en BD relacionales o ficheros planos, pero pueden incluir BD no relacionales u otras estructuras diferentes. La extracción convierte los datos a un formato preparado para iniciar el proceso de transformación.

1.4.2 Transformación

Esta fase es la encargada de convertir aquellos datos inconsistentes en un conjunto de datos compatibles y congruentes, para que puedan ser cargados en el DWH. Estas acciones se llevan a cabo, debido a que pueden existir diferentes fuentes de información, y es vital conciliar un formato y forma única, definiendo estándares, para que todos los datos que ingresarán al DWH estén integrados.

1.4.3 Carga

Este proceso es el responsable de cargar la estructura de datos del DWH con aquellos datos que han sido transformados y que residen en el almacenamiento intermedio y aquellos datos de los OLTP que tienen correspondencia directa con el depósito de datos. Se debe tener en cuenta, que antes de mover los datos al almacén de datos, es necesario analizarlos con el propósito de asegurar su calidad, ya que es un factor clave que no debe dejarse de lado.

1.4.4 Conceptos relacionados con el proceso de extracción, transformación y carga

1.4.4.1 OLTP

Procesamiento de transacciones en línea (OLTP por sus siglas en inglés), representa toda aquella información transaccional que genera la empresa en su accionar diario, además, de las fuentes externas de las que puede llegar a disponer. Entre los OLTP más habituales que pueden existir en cualquier organización se encuentran: las hojas de cálculo, archivos de textos, hipertextos, BD transaccionales, entre otros.

1.4.4.2 *Staging area*

El *Staging area* es un área temporal donde se recogen los datos que se necesitan de los sistemas origen. Se recogen los datos estrictamente necesarios para las cargas, y se aplica el mínimo de transformaciones a los mismos. No se aplican restricciones de integridad ni se utilizan claves, los datos se tratan como si las tablas fueran ficheros planos. De esta manera se minimiza la afectación a los sistemas origen, haciendo la carga lo más rápida posible y se reduce también al mínimo la posibilidad de error.

1.4.4.3 Data Mart

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

1.4.4.4 Metadatos

Los metadatos son datos que describen o dan información de otros datos, que en este caso, existen en la arquitectura del DWH. Brindan información de localización, estructura y significado de los datos, básicamente mapean los mismos. El concepto de metadatos es análogo al uso de índices para localizar objetos en lugar de datos. La gran ventaja que trae aparejada el DWH en relación con los metadatos es que el usuario puede gestionarlos, exportarlos, importarlos, realizarles mantenimiento e interactuar con ellos, ya sea manual o automáticamente (7).

1.5 Herramientas para el proceso de extracción, transformación y carga

Debido a la gran importancia que tiene la reducción de los costos y la independencia con respecto a los proveedores, se decidió utilizar herramientas de software libre para desarrollar el proceso de ETL.

1.5.1 Kettle

Conocido actualmente como Pentaho Data Integration, es un proyecto belga de código abierto, ahora adoptado por Pentaho BI, que incluye un grupo de herramientas para realizar el proceso de ETL. Uno de sus objetivos es que dicho proceso sea más fácil de generar, mantener y desplegar. Kettle está compuesto por cuatro herramientas: SPOON, PAN, CHEF y KITCHEN.

- **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 *batch* diseñados con CHEF.

El uso de esta herramienta permite evitar grandes cargas de trabajo manual frecuentemente difícil de mantener y de desplegar. Además, es una herramienta que permite definir transformaciones de forma gráfica, interconectando bloques que tienen diversas funciones. Es extremadamente versátil, ya que se tienen bloques que permiten leer y escribir de cualquier BD, fichero Excel, Access y otros que permiten operar con los campos renombrando, calculando campos en función de otros, mapeando valores, realizando búsquedas auxiliares en BD y normalizando los datos de distintas filas en una sola. Las transformaciones que se hacen con el Kettle se guardan en un fichero *ktr* que luego puede ser ejecutado mediante líneas de comandos o un fichero *batch* (8).

Ventajas de Kettle:

- Funciona en 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 púbica GPL.
- Basada en metadatos.
- Como soporte se encuentran los foros de Pentaho y la comunidad Pentaho.
- Soporta Oracle, DB2.SQL Server, Sybase así como MySQL y Postgres. También soporta la conectividad con SAP.
- Con respecto a la escalabilidad, soporta la arquitectura de procesamiento en paralelo para distribuir las tareas de ETL a través de múltiples servidores.

Basado en dos tipos de objetos: transformaciones (colección de pasos en un proceso de ETL) y trabajos (colección de transformaciones).

1.5.2 Talend

Talend Data Integration, es una herramienta de ETL de código abierto, que efectúa operaciones tales como alimentación de un DWH, sincronización de BD, transformación y verificación de la integridad de los datos. Su interfaz gráfica de usuario está basada completamente en Eclipse RCP (Rich Client Platform) e incluye numerosos componentes para procesos de modelado de negocios, así como implementaciones técnicas para extracción de información, transformación y mapeo del flujo de datos. Las funciones generales de Talend incluyen modelador de negocios, diseñador del trabajo y administrador de metadatos (9).

Esta herramienta se ejecuta sobre las plataformas Windows, Unix, Linux. El código fuente Java/Eclipse está disponible para su descarga y personalización. Para su soporte cuenta con la wiki de Talend, foro de Talend y un rastreador de errores. La comunidad y el foro proporcionan un lugar donde obtener asistencia y soporte gratuito. Aunque este foro no tiene ningún tipo de garantía, es supervisado por el personal de Talend para mantener su calidad.

Talend a pesar de tener un gran número de características a su favor también cuenta con una serie de inconveniencias. Una de ellas es que necesita un controlador JDBC para acceder a las fuentes de datos, además al estar financiado por una firma de capital privado, existe un mínimo riesgo de que si se deja de inyectar capital puede provocarse la paralización en las mejoras del producto y por ende la no compatibilidad con nuevas versiones de BD. Entre otros de los aspectos negativos se podría destacar el hecho de que esta herramienta no cuenta con productos complementarios de calidad de

datos y gestión de metadatos; así como la falta de un proceso automático de separación y redistribución de datos, lo cual puede generar cuellos de botella.

1.5.3 Scriptella

Es una herramienta de ejecución de scripts desarrollada en Java. Su objetivo fundamental es la simplicidad, ya que no requiere que el usuario tenga que aprender otro complejo lenguaje basado en XML para usarlo. Permite el uso de SQL u otro lenguaje de scripting adecuado para la fuente de datos y para llevar a cabo las transformaciones necesarias.

1.5.4 Octopus

Es una herramienta de ETL basada en Java y soporta únicamente fuentes de datos que vengan con el manejador JDBC, aunque incluye driver especiales que permiten la conectividad con archivos CSV, XML, MS-SQL y archivos de propietarios. Octopus utiliza archivos XML para cargar los trabajos, así como para definir los parámetros de las transformaciones.

Luego de un estudio del estado de las herramientas Kettle, Talend, Scriptella y Octopus, se ha seleccionado la versión 3.1 de Kettle para el desarrollo del proceso de ETL, dado que es una herramienta libre muy potente, así como una de las más antiguas y utilizadas por los usuarios. Producto a esto tiene gran soporte técnico y los usuarios comparten muchos consejos y trucos en los foros. Además, Kettle es la más completa de las vistas por la gran cantidad de conectores que posee y la posibilidad de crear flujos de trabajo integrados con transformaciones de datos de manera muy sencilla y funcional.

1.6 Sistema gestor de base de datos

Los Sistemas Gestores de Bases de Datos (DBMS por sus siglas en inglés) son un elemento clave dentro del mundo de la información ya que contienen las rutinas necesarias para el manejo de los datos: dígase definición, construcción y manipulación de los mismos. Permiten la eliminación y actualización de registros, la combinación con otras BD y la generación de informes impresos.

Objetivos de los DBMS:

- 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 BD (esquema) sin necesidad de modificar los programas de las aplicaciones que manejan esos datos.

- Mantener la integridad de los datos realizando las validaciones necesarias cuando se realicen modificaciones en la BD.
- Mejorar la eficacia de acceso a los datos, en especial en el caso de consultas imprevistas.

Para decidir cuál DBMS utilizar, se ha tenido en cuenta que el CIM es una de las entidades que se encuentran en el país en franca migración hacia la independencia tecnológica. Actualmente está exportando todos sus dispositivos de almacenamiento hacia la plataforma PostgreSQL, por lo que queda seleccionado como DBMS PostgreSQL. La versión que se utilizará es la 8.3.7 por ser lo suficientemente estable y segura.

1.6.1 PostgreSQL

PostgreSQL es un sistema gestor de base de datos objeto-relacional libre, liberado bajo la licencia BSD (del inglés Berkeley Software Distribution). Como muchos otros proyectos Open Source, el desarrollo de PostgreSQL no es manejado por una sola compañía sino que es dirigido por una comunidad de desarrolladores y organizaciones comerciales las cuales trabajan en su desarrollo, dicha comunidad es denominada el PostgreSQL Grupo Global de Desarrollo (PGDG), sus siglas en inglés se definen como: PostgreSQL Global Development Group. PostgreSQL ha tenido una larga evolución, comenzando con el proyecto Ingres en la Universidad de Berkeley. Este proyecto, liderado por Michael Stonebraker, fue uno de los primeros intentos en implementar un motor de base de datos relacional (10).

En cuanto a la arquitectura de la herramienta, PostgreSQL utiliza un modelo cliente/servidor. Una sesión de este gestor consiste en los siguientes procesos:

Proceso servidor. Administra los archivos de la base de datos, acepta conexiones a las BD de aplicaciones clientes y realiza acciones sobre las BD por solicitud de los clientes.

Aplicaciones cliente. Permite realizar operaciones sobre las BD. Existen diversos tipos de aplicaciones clientes: un cliente puede ser una herramienta basada en texto, herramienta gráfica, un servidor web que acceda a la BD para sus páginas web, o herramientas de administración de las BD.

El servidor y clientes pueden estar ejecutándose en diferentes equipos, por lo tanto PostgreSQL permite la comunicación entre estos procesos a través de TCP/IP. El servidor soporta múltiples conexiones concurrentes de clientes. Para este propósito ejecuta nuevos procesos para cada conexión. Este proceso es transparente para los usuarios (11).

Ventajas encontradas en PostgreSQL:

- Soporta lenguajes: PHP, C, C++, Perl y Python.
- Drivers: ODBC, JDBC y.Net.
- Soporta: *triggers*, procedimientos almacenados, funciones, secuencias, relaciones, reglas, tipos de datos definidos por el usuario, vistas y vistas materializadas.
- Soporte de tipos de datos de SQL92, SQL99 y SQL2003.
- Soporte de protocolo de comunicación encriptado por SSL.
- Máximo de bases de datos: ilimitado.
- Máximo de tamaño de tabla: 32 TB.
- Máximo de tamaño de registro: 1.6 TB.
- Máximo de tamaño de campo: 1 GB.
- 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 10 (pl/pgsql, pl/java, pl/perl, pl/python, tcl, pl/php, C, C++ y Ruby).

1.7 Herramientas CASE de modelado con UML

Las herramientas CASE de modelado con UML permiten aplicar la metodología de análisis y diseño orientado a objetos, así como abstraerse del código fuente, en un nivel donde la arquitectura y el diseño se tornan más obvios, más fáciles de entender y modificar. Es importante resaltar que UML es un lenguaje de modelado para especificar o describir métodos o procesos. Dentro de las herramientas CASE que utilizan UML para su modelado se encuentran Rational Rose y Visual Paradigm. Teniendo en cuenta que en el Trabajo de Diploma precedente se utilizó la herramienta Visual Paradigm para UML en su versión 6.1 (versión 3.1 de la suite del producto), se decide emplear la misma para el modelado del proceso de ETL, manteniendo una uniformidad en cuanto a la herramienta de modelado.

1.7.1 Visual Paradigm

"Es una herramienta UML profesional que soporta el ciclo de vida completo del desarrollo de software: análisis y diseño orientados a objetos, construcción, pruebas y despliegue. El software de modelado UML ayuda a una más rápida construcción de aplicaciones de calidad, mejores y a un menor coste. Permite dibujar todos los tipos de diagramas de clases, código inverso, generar código desde diagramas y generar documentación. La herramienta UML CASE también proporciona abundantes tutoriales de UML, demostraciones interactivas de UML y proyectos UML" (12). Es una tecnología que

está disponible en varios idiomas, en conjunto con esto, es fácil de instalar y fácil de actualizar. Por último, Visual Paradigm admite compatibilidad con las demás versiones. De manera general, esta herramienta de modelado ofrece:

- Entorno de creación de diagramas para UML 2.0.
- Diseño centrado en casos de uso y enfocado al negocio que genera un software de mayor calidad.
- Uso de un lenguaje estándar común a todo el equipo de desarrollo que facilita la comunicación.
- Disponibilidad de múltiples versiones, para cada necesidad.
- Disponibilidad en múltiples plataformas.

1.8 Evaluación del proceso de extracción, transformación y carga

Después de haber realizado una búsqueda y estudio de cómo evaluar el proceso de ETL, no se ha encontrado una manera formal ni un procedimiento específico en la bibliografía consultada que diga cómo evaluar el mismo; aunque no se debe descartar que algunas de las metodologías consultadas describen elementos que se deben cumplir a la hora de desarrollar el proceso de ETL. Por estas razones se propone crear una lista de chequeo a partir del listado obtenido en el acápite 1.3.3, para evaluar la implementación del proceso de ETL.

Conclusiones

En este capítulo se demostró la necesidad de aplicar el proceso de ETL, con el fin de integrar los datos generados por cada uno de los EC que se gestionan en el CIM. Para el posterior desarrollo de dicho proceso, se realizó un estudio sobre las diferentes arquitecturas y metodologías de DWH, además se analizó un conjunto de herramientas informáticas para la implementación del proceso de ETL. Luego de esta investigación se arribaron a las siguientes conclusiones:

- Se decidió utilizar como metodología de arquitectura la propuesta por Kimball, siendo esta la que más se ajusta a las características del proyecto.
- Se propone tomar lo mejor y más aplicable de cada metodología de desarrollo analizada con el objetivo de elaborar un procedimiento para implementar el proceso de ETL.
- Se decidió utilizar PostgreSQL como DBMS en su versión 8.3.
- Se decidió utilizar la herramienta informática Kettle en su versión 3.1, de la suite Pentaho Data Integration.
- Se decidió utilizar UML como lenguaje de modelado en su versión 2.0.
- Se decidió utilizar Visual Paradigm en su versión 6.1 (versión 3.1 de la suite del producto) como herramienta de modelado.

- Se propone elaborar y aplicar una lista de chequeo con el propósito de evaluar la implementación del proceso de ETL.

CAPÍTULO 2: IMPLEMENTACIÓN DEL PROCESO DE EXTRACCIÓN, TRANSFORMACIÓN Y CARGA

Introducción

Teniendo en cuenta que no existe una metodología propia para la implementación del proceso de ETL, sino que los pasos para el desarrollo de esta etapa se embeben dentro del proceso de implementación de un DWH y que estas han sido creadas para el mundo empresarial; y además por las características que presentan los datos del CIM, se decidió elaborar un procedimiento para guiar el proceso y luego implementarlo. Una vez elaborado e implementado dicho procedimiento se valida el mismo a través del método Delphi. Vale destacar que para la definición de este proceso de ETL, se tiene como precedente una investigación metodológica orientada al modelado de la integración de los datos.

2.1 Procedimiento para la implementación del proceso de extracción, transformación y carga de un Datawarehouse

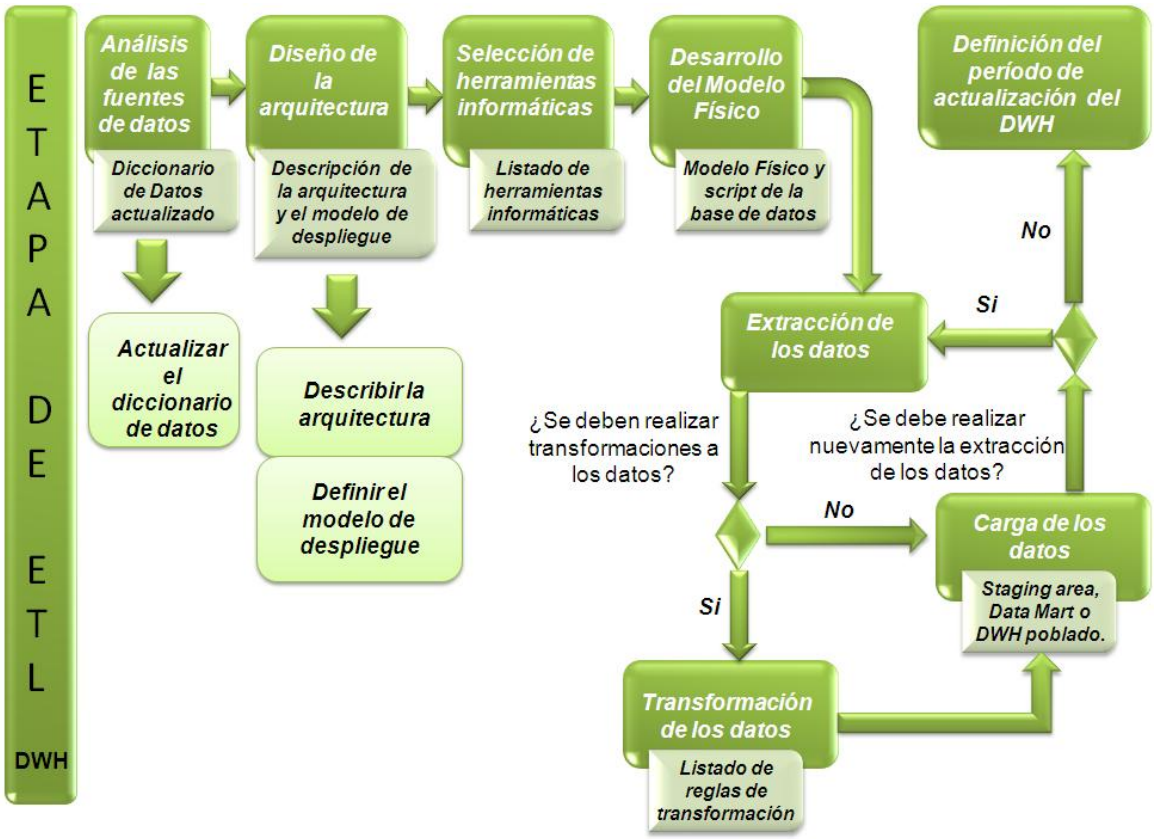


Figura 5. Procedimiento para implementar el proceso de ETL

2.1.1 Análisis de las fuentes de datos

Para la realización de este paso se deben revisar los datos fuentes con los que cuenta la institución teniendo en cuenta las necesidades del negocio de los usuarios, especificándose las variables que se corresponden con las perspectivas de análisis de la organización. En dicho estudio se genera un *Diccionario de Datos* (DD) donde se establece la correspondencia de todas las perspectivas con sus datos fuentes, significado para el negocio, modelo donde se encuentran y los posibles valores que pueden tomar. A partir de este artefacto se define el tipo de dato de cada variable que va a intervenir en la extracción de los datos y se especifica el formato en que se encuentran las fuentes de datos que van a alimentar el DWH; quedando definidas todas las aplicaciones, BD, o cualquier archivo donde resida información útil para el proceso de ETL.

En este paso se determinan los tipos de datos de cada variable y se define el formato de la fuente de datos de donde se obtuvo esta; actualizándose el Diccionario de Datos (ver tabla 1).

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

Tabla 1. Estructura del Diccionario de Datos actualizado

2.1.2 Diseño de la arquitectura

Esta etapa hace referencia al proceso de elegir y diseñar la estructura interna del DWH, basándose en que la misma estará formada por diversos componentes que interactúan entre sí y que cumplen una función específica dentro del sistema. Debe definirse la estrategia arquitectónica a usar, ya sea Inmon (descendente) o Kimball (ascendente); 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 DWH así como la estructura del despliegue del DWH.

Entregable: Descripción de la arquitectura y modelo de despliegue.

2.1.3 Selección de las herramientas informáticas para implementar el proceso de extracción, transformación y carga

Existen varios software que facilitan realizar la compleja actividad de extraer datos desde diferentes fuentes, para luego integrarlos, filtrarlos y depurarlos. Para saber cuál de ellos es conveniente utilizar es necesario hacer un estudio de sus características, teniendo en cuenta una serie de aspectos importantes relacionados con el negocio. Algunos de estos aspectos podrían ser: la capacidad de la

herramienta para reconocer todas las fuentes de datos definidas en el negocio y el soporte para el modelo de despliegue definido en el paso anterior.

Además se debe tener en cuenta el DBMS a utilizar según los requerimientos de hardware definidos en el paso anterior y el modelado de los datos.

Entregable: Listado de herramientas informáticas.

2.1.4 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 BD teniendo en cuenta el gestor seleccionado en el paso anterior.

Entregable: Modelo Físico y script de la BD.

2.1.5 Proceso de extracción, transformación y carga

En esta fase se realiza el proceso de extracción, transformación y carga de los datos hacia el DWH haciendo uso de la herramienta seleccionada. Es válido destacar que este proceso puede repetirse o ajustarse el orden de las acciones, teniendo en cuenta la arquitectura definida anteriormente.

2.1.5.1 Extracción de datos

A partir del DD actualizado en el paso uno, se procede a configurar e implementar el proceso de extracción con la herramienta previamente seleccionada. Una vez concluido este paso los datos quedarán convertidos en un formato preparado para iniciar el proceso de transformación.

2.1.5.2 Definición de las transformaciones

En este paso se tiene como entrada la *especificación de las reglas del negocio* y la *especificación de los indicadores* que se obtienen del análisis con el cliente; los cuales se convierten en transformaciones necesarias sobre los datos extraídos. Es válido aclarar que pueden surgir transformaciones adicionales para facilitar el proceso de ETL, sin afectar el negocio. Luego de haber definido todas las transformaciones, se inicia la configuración e implementación de las mismas en la herramienta con la que se va a trabajar.

Entregable: Reglas de transformación.

2.1.5.3 Carga de los datos

Una vez que los datos han pasado por todo el proceso de adecuación, sólo queda almacenarlos en el DWH. Para la realización de este paso se debe configurar e implementar la carga mediante la herramienta seleccionada.

Entregable: DWH poblado.

2.1.6 Definición del período de actualización del Datawarehouse

En este paso es donde se define la periodicidad de carga de los datos hacia el DWH teniendo en cuenta las necesidades del cliente. Luego se configura en la herramienta el período de actualización definido.

2.2 Implementación del procedimiento propuesto

La implementación del proceso de ETL se realizará a partir del procedimiento propuesto anteriormente. Para ello se tendrá en cuenta el desarrollo de cada uno de los puntos abordados en el mismo, así como su orden lógico.

2.2.1 Análisis de las fuentes de datos del Centro de Inmunología Molecular

Para realizar el análisis de las fuentes de datos del CIM se cuenta con un DD generado por un estudio metodológico precedente. A continuación se realiza la actualización de este artefacto, especificándose los formatos fuentes y tipos de datos que se generan por cada una de las perspectivas e indicadores del análisis. Es válido aclarar que en este paso se actualiza únicamente la dimensión datos demográficos del EC hR3 *C y C 040* (ver tabla 2), el resto de los DD se encuentran en el artefacto *Actualización del Diccionario de Datos* dentro del expediente de proyecto.

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	Nos dice el sexo del paciente	Modelo 1. Datos Demográficos	Sexo (v11)	1: Femenino 2: Masculino	String	xls
Edad	Nos dice la edad del paciente	Modelo 1. Datos Demográficos	Edad (v6)	Entre 18 y 80	Integer	xls
Raza	Nos dice la raza del paciente	Modelo 1. Datos Demográficos	Raza (v10)	1: Blanca 2: Negra 3: Mestiza	String	xls

Estadío	Nos dice el estadío en que se encuentra el paciente	Modelo 1. Datos Demográficos	Estadío (v18)	1: I 2: II 3: III 4: IV	Integer	xls
Clasificación anatomopato-lógica	Nos dice la clasificación del tumor al diagnosticar al paciente	Modelo 1. Datos Demográficos	V21a V21b	Texto hasta 80 caracteres. Estas variables deben concatenarse	String	xls
Grado de diferenciación	Forma parte de la clasificación anatomopato-lógica	Modelo 1. Datos Demográficos	V21a V21b	1: Bien Diferenciado 2: Moderado 3: Poco Diferenciado o Indiferencia-do	String	xls
ECOG (Karnofsky)	Nos dice el estado general del paciente cuando se diagnostica	Modelo 1. Datos Demográficos	Estado (OMS) (v20)	De 0-2 según la escala de ECOG	Integer	xls
Tratamientos previos	Nos dice el tratamiento que ha recibido el paciente antes del ensayo	Modelo 1. Datos Demográficos	Cirugía (v22) Radioterapia (v24) Quimioterapia (v26)	1: True 2: False	Boolean	xls

Tabla 2. Datos demográficos del paciente

2.2.2 Diseño de la arquitectura del Datawarehouse

La arquitectura del DWH (ver figura 6) se diseñó teniendo en cuenta la estrategia ascendente propuesta por Kimball, donde cada Data Mart representa a un producto del CIM y el DWH integra los resultados de análisis para cualquier localización y producto a partir de los distintos Data Marts.

Según la arquitectura diseñada, los datos son extraídos desde aplicaciones, BD o cualquier archivo donde resida información útil para el proceso de ETL. Una vez analizadas las fuentes de datos, se procede a extraer los datos estrictamente necesarios para la carga de los mismos dentro del *Staging area*, aplicándoles el mínimo de transformaciones posibles. Posteriormente los datos son integrados,

transformados y limpiados, para luego ser cargados en los distintos Data Marts que representan a cada uno de los esquemas del DWH. De esta forma se deja abierta la arquitectura con el objetivo de dar soporte a la aparición de un nuevo Data Mart. Todo este proceso descrito anteriormente se almacena en los metadatos a través de la herramienta informática Kettle.

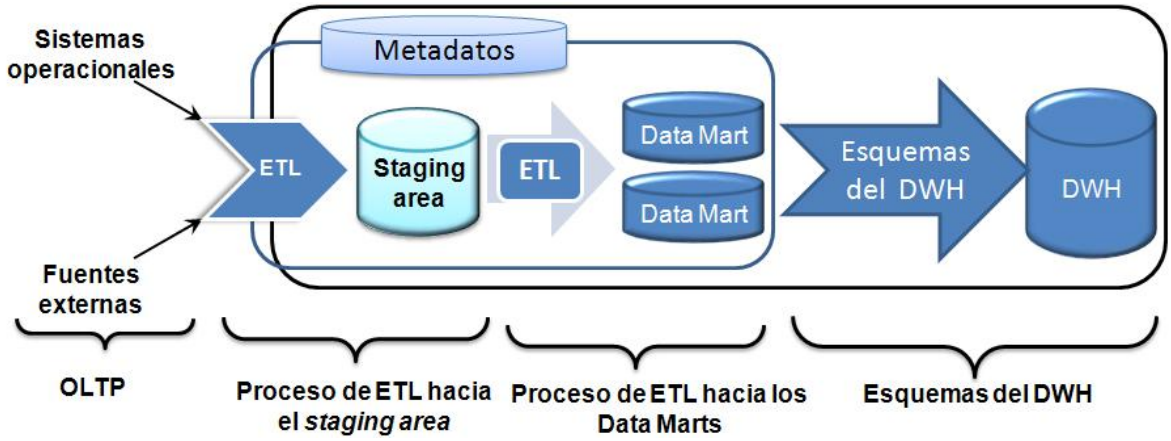


Figura 6. Arquitectura de DWH para el CIM

2.2.2.1 Modelo de despliegue

El modelo de despliegue (ver figura 7) define la estructura física de los procesos de ETL usados en la carga de los datos hacia el DWH. El mismo está compuesto por un servidor que contiene la información o datos de los EC del CIM. A este servidor se le conecta una computadora personal (PC por sus siglas en inglés) a través del protocolo TCP/IP, que es la encargada de ejecutar los procesos de ETL. Una vez realizadas las transformaciones, se carga toda la información procesada en el servidor de BD que contiene el DWH del CIM mediante el protocolo JDBC.

Los servidores deben tener alta disponibilidad en los períodos de actualización del DWH. Para la PC encargada de realizar los procesos de ETL se necesita un rendimiento adecuado, garantizado por al menos un procesador Dual Intel Xeon 3 GHz o similar y RAM suficiente (2 GB a 4 GB) y para el servidor de BD se requiere de uno a dos TB (unidad de medida de almacenamiento de datos) disponibles, pues el volumen de información es bastante grande y perdura en el tiempo.

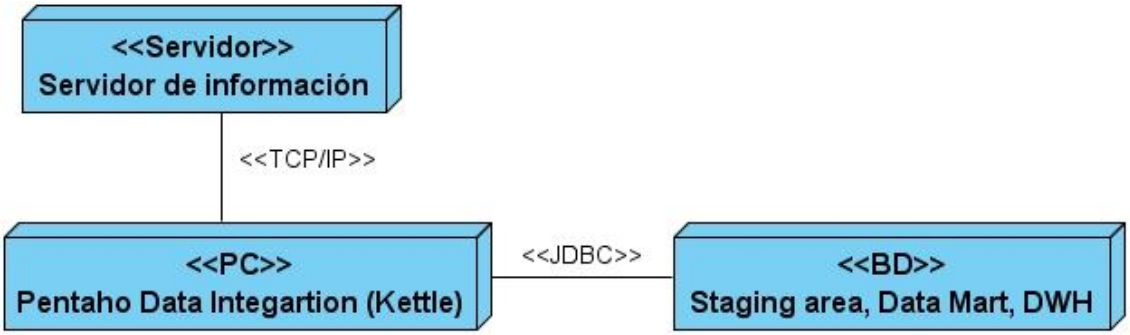


Figura 7. Modelo de despliegue

2.2.3 Selección de las herramientas informáticas para realizar el proceso de extracción, transformación y carga

Para la implementación y configuración del proceso de ETL se decidió optar por la herramienta informática Kettle en su versión 3.1 y para la gestión de los datos dentro del DWH el DBMS PostgreSQL en su versión 8.3.

2.2.4 Elaboración del Modelo Físico y script de la base de datos

En este paso se obtuvieron dos Modelos Físicos, el primero (ver figura 8) fue generado a partir del Modelo Lógico (ver anexo 3) obtenido de las necesidades del negocio y el segundo (ver figura 9) representa al *Staging area* que surgió para dar soporte a la arquitectura definida (13).

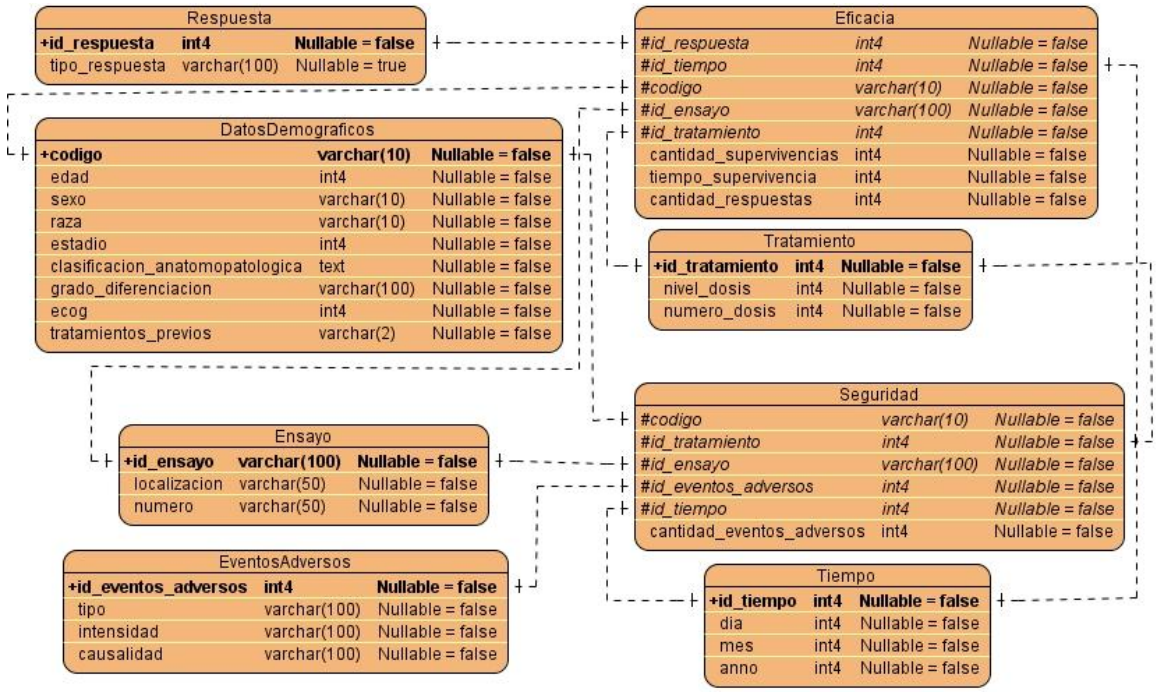


Figura 8. Modelo Físico del Data Mart

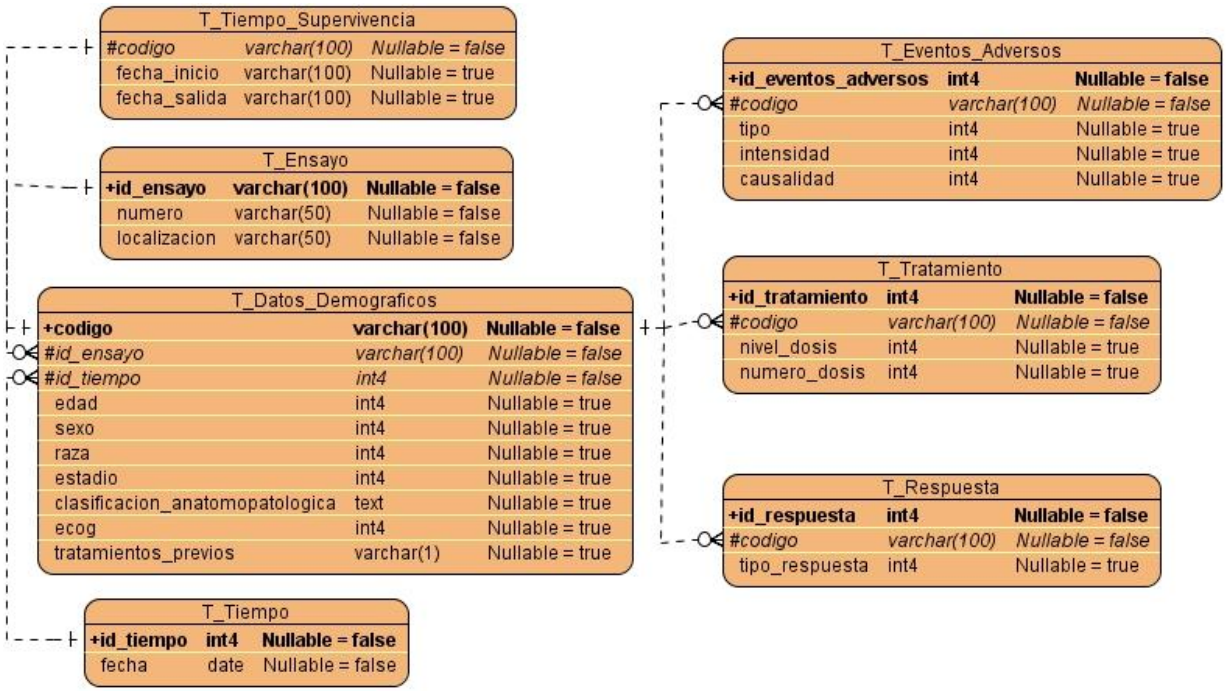


Figura 9. Modelo Físico del *Staging area*

2.2.5 Implementación del proceso de extracción, transformación y carga

Para poblar el Data Mart del producto hR3 con los datos fuentes que se obtuvieron en la actualización del DD, se hizo necesario realizar el proceso de ETL dos veces, cumpliendo de esta forma con la arquitectura definida. Primero se pasó toda la información necesaria contenida en los ficheros xls hacia el *Staging area*, realizando el menor número de transformaciones posibles con el propósito de no afectar a los sistemas origen; logrando de esta manera que la carga fuera lo más rápido posible. Una vez concluida esta primera etapa se insertaron los datos del *Staging area* dentro del Data Mart del producto hR3, donde se realizaron todas las transformaciones necesarias para que los datos quedaran lo más integrado posible.

2.2.5.1 Extracción de las fuentes de datos del Centro de Inmunología Molecular

Se procedió a configurar e implementar en el Kettle el proceso de extracción de los datos fuentes definidos en el DD. Debido a que todas las fuentes de información encontradas eran ficheros xls, se hizo necesario configurar el paso de Entrada Excel. En la configuración del mismo se definió la dirección o directorio donde se encontraban los ficheros, así como las hojas y los campos que participaron en la extracción de los datos. En la figura 10 se muestra un ejemplo de cómo se configuró la extracción de los datos demográficos del producto hR3 cabeza y cuello 040.

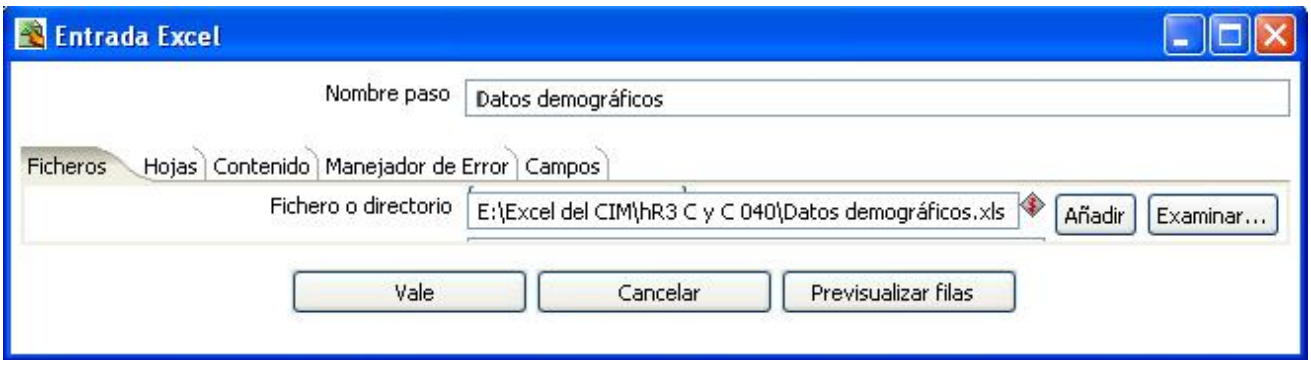


Figura 10. Extracción de los datos demográficos del producto hR3 C y C 040

Una vez concluida la configuración del paso *Entrada Excel*, se implementaron y configuraron las transformaciones.

2.2.5.2 Definición de transformaciones para el *Staging area*

En este paso se hizo uso del menor número de transformaciones posibles, con el fin de minimizar las afectaciones en los sistemas origen de información, en cuanto al consumo de recursos computacionales. Por tales razones, se definieron solamente cuatro transformaciones que serán usadas antes de cargar los datos en el Staging area.

***Definir si tuvo tratamientos previos*:** los únicos tratamientos previos que se tendrán en cuenta para el análisis serán quimioterapia, radioterapia y cirugía. En correspondencia con ello si el paciente participó en al menos uno de ellos, la variable tratamientos previos toma valor verdadero, en caso contrario se almacena como falso.

***Cambio de variable*:** se refiere a llevar todos los nombres de variables de las fuentes de datos a un lenguaje con más significado para el negocio, por ejemplo: la variable v6 es utilizada en los modelos Excel para hacer referencia a la edad del paciente, aplicándole esta transformación quedaría con el nombre edad (ver figura 11).

Figura 11. Transformación cambio de variable

***Concatenación de valores*:** esta transformación hace referencia a unir en una misma variable, el contenido de todas aquellas que se necesitan para completar su significado. Por ejemplo: la clasificación anatomopatológica de un paciente se almacena en las variables V21a y V21b, luego de ser aplicada dicha transformación la clasificación anatomopatológica estaría formada por la concatenación de V21a y V21b, completando de esta forma una misma idea en una sola variable (ver figura 12).

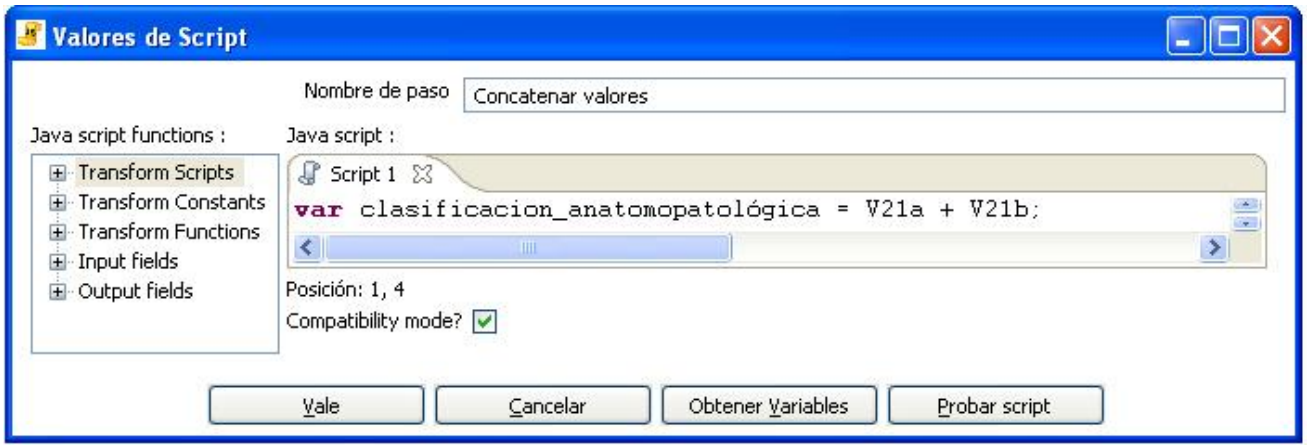


Figura 12. Transformación concatenar valores

***Creación del código que identifica al paciente*:** con esta regla se crea un código único para el paciente que lo distingue de los demás en cualquier EC. Al aplicar esta transformación el código

quedaría formado por las iniciales del paciente, un signo de resta, el número de inclusión, otro signo de resta y la localización del ensayo (ver figura 13).

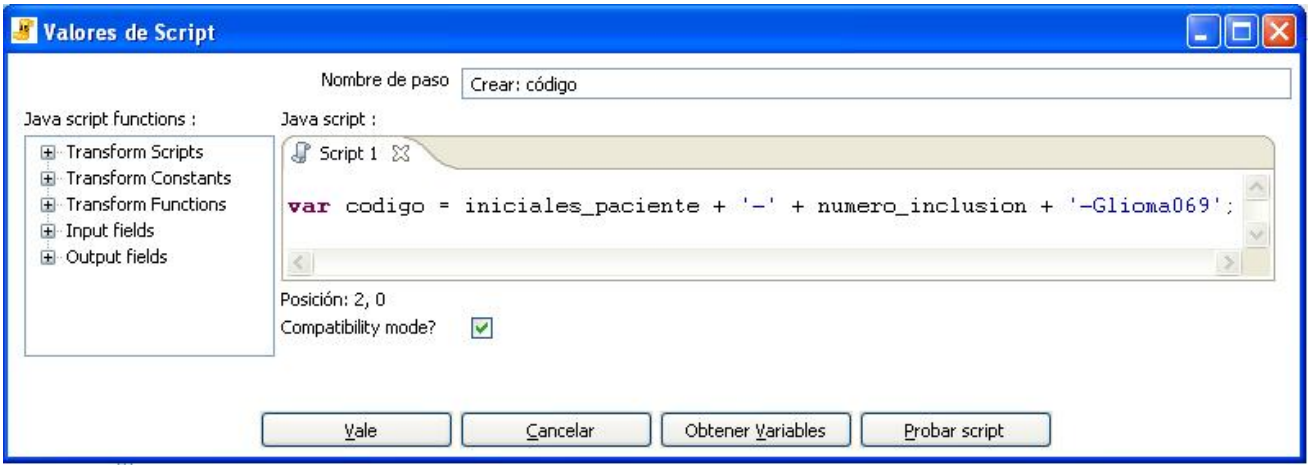


Figura 13. Creación del código del paciente

2.2.5.3 Carga de los datos hacia el *Staging area*

Una vez realizadas las transformaciones anteriores, se procedió a configurar e implementar en el Kettle el proceso de carga hacia el *Staging area*. Con el fin de dar cumplimiento a este paso se creó una conexión, donde se especificaron los parámetros nombre de la BD, puerto, tipo de conexión, usuario y contraseña (ver figura 14).

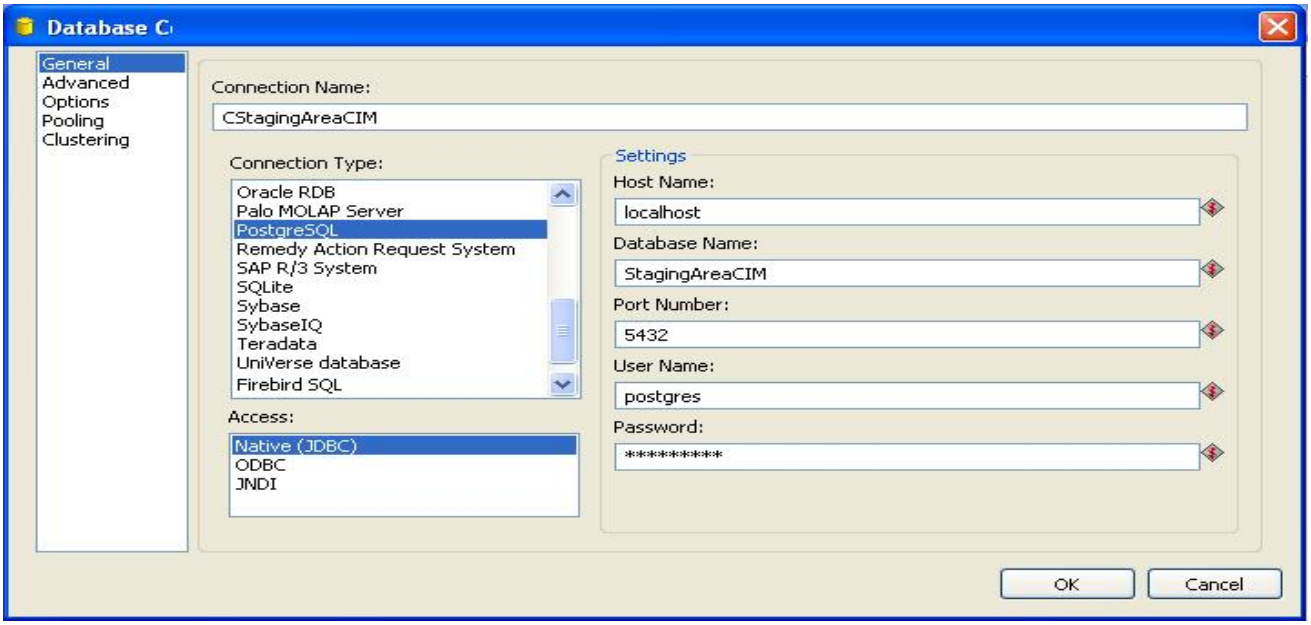


Figura 14. Conexión con base de datos PostgreSQL

Después de crear la conexión con el *Staging area* se configuró el paso de salida hacia la tabla destino, donde se especificó la conexión que se utilizaría y la tabla que recibiría los datos (ver figura 15).

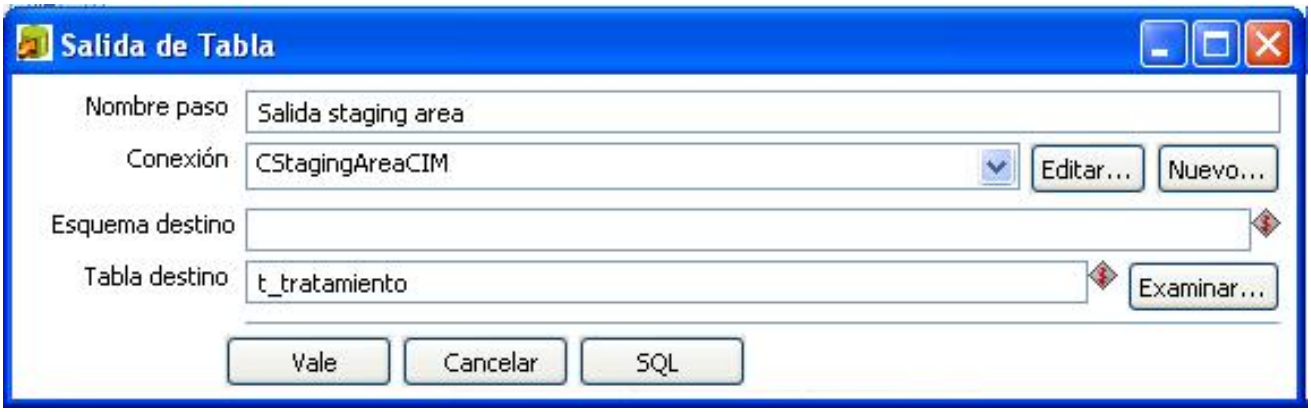


Figura 15. Salida de los datos hacia el Staging area

Una vez terminada la configuración e implementación de todo el proceso de ETL hacia el *Staging area*, quedan creadas las condiciones para poblar dicha BD con la información del producto hR3 del CIM. En la siguiente tabla se muestran los resultados obtenidos después de haber realizado la extracción (E), transformación (T) y carga (L) de los datos por cada una de las dimensiones de las diez BD del producto hR3.

Base de datos	Dimensiones del *Staging area*													
	Datos demográficos		Eventos adversos		Tiempo de supervivencia		Ensayo clínico		Tratamiento		Respuesta		Tiempo	
	E	L	E	L	E	L	E	L	E	L	E	L	E	L
	T		T		T		T		T		T		T	
C y C 040	14	14	838	838	28	14	1	1	76	76	40	40	0	1
	3		2		2		0		2		2		0	
C y C 046	30	10	186	186	20	10	1	1	56	56	40	40	0	1
	3		2		2		0		2		2		0	
C y C 055	208	103	183	183	181	73	1	1	285	285	1098	1098	0	1
	3		2		2		0		2		2		0	
C y C 076	20	10	100	100	20	10	1	1	40	40	44	11	0	1
	3		2		2		0		2		2		0	
Glioma 069	29	29	559	559	46	17	1	1	335	335	177	177	0	1
	3		2		2		0		2		2		0	

Base de datos	Dimensiones del *Staging area*															
	Datos demográficos		Eventos adversos		Tiempo de supervivencia		Ensayo clínico		Tratamiento		Respuesta		Tiempo	
	E	L	E	L	E	L	E	L	E	L	E	L	E	L
	T		T		T		T		T		T		T	
Glioma 053	29	29	43	43	46	16	1	1	335	335	177	177	0	1
	3		2		2		0		2		2		0	
Metace-rebral 079	54	27	190	190	54	27	1	1	116	116	32	32	0	1
	3		2		2		0		2		2		0	
T.Sólidos 035	12	12	108	108	22	10	1	1	54	41	36	36	0	1
	3		2		2		0		2		2		0	
Mama 070	25	12	77	77	24	11	1	1	119	119	36	36	0	1
	3		2		2		0		2		2		0	
Esófago 075	116	52	400	400	107	43	1	1	187	124	81	81	0	1
	3		2		2		0		2		2		0	

Tabla 3. Proceso de ETL hacia el *Staging area*

2.2.5.4 Extracción de los datos contenidos en el *Staging area*

Al tener poblado completamente el Staging area, se procedió a realizar una extracción de los datos contenidos en el mismo mediante el paso Entrada Tabla. Para configurar este paso se creó una conexión al Staging area y se especificó la(s) tabla(s) de la(s) cual(es) se extraerían los datos mediante una consulta SQL. En la figura 16 se muestra un ejemplo de cómo quedaría la extracción de los datos demográficos de un paciente.

Figura 16. Extracción de los datos del Staging area.

2.2.5.5 Definición de transformaciones para el Data Mart del producto hR3

En este paso se tiene como entrada una *especificación de reglas del negocio* (ver anexo 1), las cuales se convierten en transformaciones necesarias para poder implementar el proceso de ETL correctamente. Además, en este paso se definieron todas las transformaciones que se necesitaron para que los datos quedaran lo más integrado posible.

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

***Nivel de dosis igual cero*:** con esta transformación si el nivel de dosis aplicado a un paciente es cero y el mismo presenta eventos adversos, la causalidad del evento adverso se toma como no relacionada.

***Intensidad normal*:** cuando la intensidad del evento adverso es normal entonces no se incluye como evento adverso.

***Codificar valores libres*:** se refiere a realizar un cambio en los valores de las variables fuentes, creando de esta forma valores legibles para los usuarios finales. Por ejemplo: la variable v3 almacena los valores 1 y 2 para referirse al sexo del paciente, una vez aplicada esta transformación se sustituyeron estos valores por masculino y femenino en este mismo orden. Esta misma regla es aplicada a la variable raza, pero en este caso se sustituyen los valores numéricos 1, 2, 3 y 4 por blanca, negra, mestiza y amarilla respectivamente. Para configurar esta transformación en el Kettle se utilizó el paso mapeo de valores (ver figura 17).

Figura 17. Mapeo de valores.

***Traducir códigos*:** esta regla es aplicada cuando una misma variable no almacena los mismos valores para referirse a un resultado. Por ejemplo: en la variable v4 se registran los eventos adversos del

paciente almacenándose en algunas fuentes *cefalea* y en otras como *migraña* para referirse a un mismo término médico (dolor de cabeza). Una vez aplicada dicha transformación la variable v4 sólo admitirá cefalea para referirse a este término médico.

***Calcular edad*:** esta transformación es aplicada cuando la edad de un paciente no está registrada y la misma se puede obtener a partir de la resta de la fecha de fallecimiento y nacimiento de esta persona. Para la realización de este cálculo se ha escrito una función en código JavaScript.

***Calcular tiempo de supervivencia*:** se refiere a obtener el tiempo de supervivencia de una persona en días mediante la resta de la fecha de inclusión y la fecha de fallecimiento. Para la realización de este cálculo se ha escrito una función en código JavaScript.

***Asignar sexo según ensayo*:** esta transformación es aplicada cuando el sexo del paciente no fue registrado debido a que el mismo se podía obtener por la localización del ensayo. Un ejemplo es el caso del EC *Mama 070*, que fue aplicado a pacientes de sexo femenino. Por tales razones quedan con sexo femenino los pacientes que intervinieron en dicho ensayo (ver figura 18).

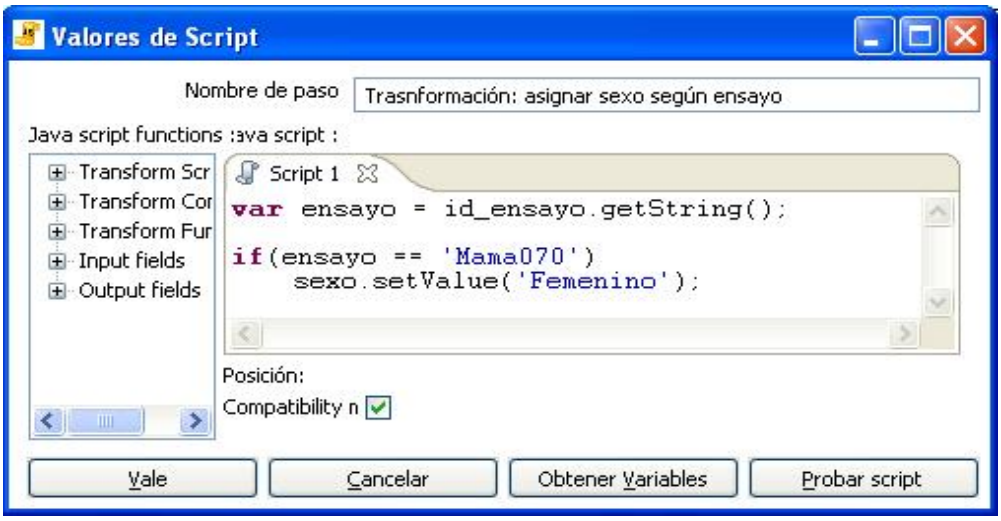


Figura 18. Asignar sexo según ensayo

***Valores nulos*:** en el caso de las variables que no contengan valor (raza, causalidad, grado de diferenciación) se almacenan en el DWH como missing (ver figura 19).

Figura 19. Eliminar valores nulos en la variable raza

***Generar valores constantes*:** esta transformación es utilizada para generar valores constantes en algunas variables nulas. Por ejemplo: en la BD *Metacerebral 079* no se recogió el estadío del paciente, pero el cliente determinó asignarle un valor constante igual a cuatro (ver figura 20).

Figura 20. Generar valores constantes

***Generar valores aleatorios*:** esta transformación hace referencia a generar valores aleatorios en algunas variables. Por ejemplo: en la BD *C y C 046*, se generan estados aleatorios de ECOG entre cero y dos (ver figura 21).

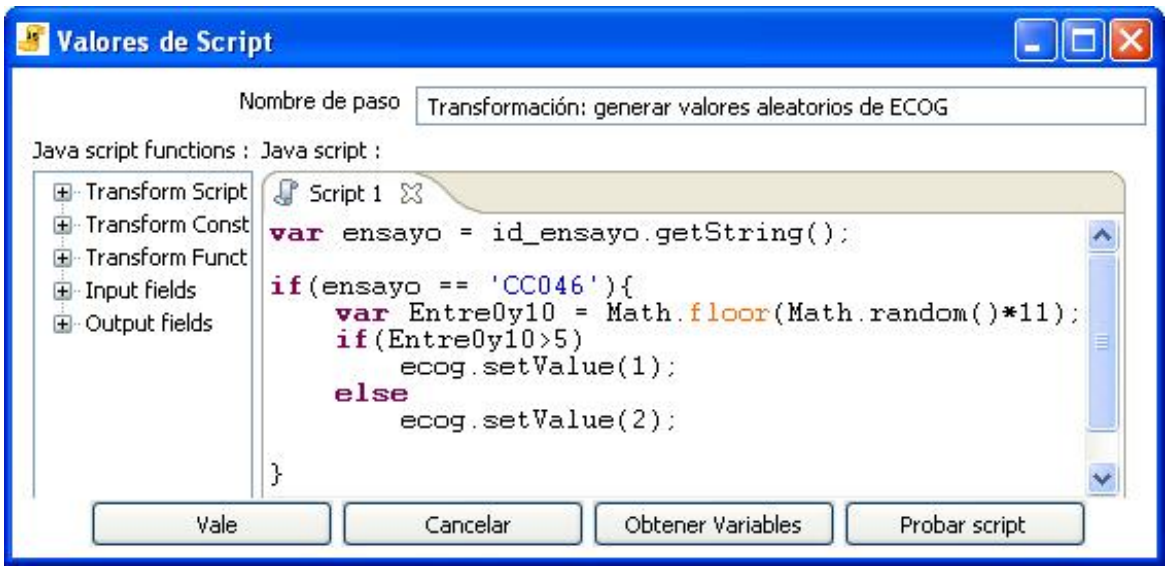


Figura 21. Generar valores aleatorios de la variable *ECOG*

***Obtener grado de diferenciación*:** el grado de diferenciación forma parte de la clasificación anatomopatológica y a los especialistas del negocio les hace falta tenerlo separado, ya que es otra perspectiva de análisis. Por esta razón, se determinó extraer el grado de diferenciación a partir de los posibles valores que pueda tomar la clasificación anatomopatológica (ver figura 22).

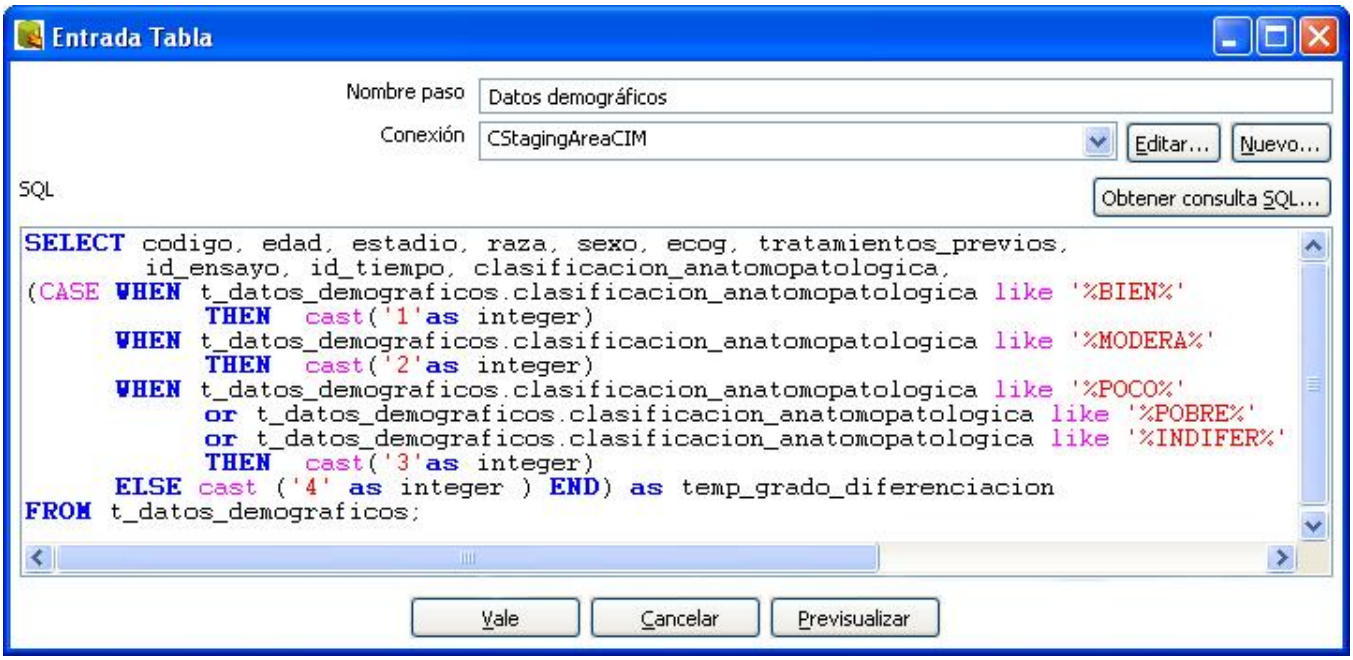


Figura 22. Obtener grado de diferenciación

2.2.5.6 Carga de los datos hacia el Data Mart del producto hR3

La carga de los datos hacia el Data Mart se realizó después de haber terminado con todas las transformaciones. Para la configuración de esta carga se hizo uso del paso Salida Tabla, para ello se creó una conexión hacia el Data Mart del producto hR3 y luego se especificó en el paso la tabla que recibiría los datos.

Una vez terminada la configuración e implementación de todo el proceso de ETL hacia el Data Mart, quedan creadas las condiciones para poblar la BD del producto hR3 con los datos contenidos en el *Staging area*. En la siguiente tabla se muestran los resultados obtenidos después de haber realizado la extracción (E), transformación (T) y carga (L) de los datos hacia el Data Mart por cada una de sus seis Dimensiones y sus dos tablas Hechos.

Dimensiones del Data Mart												**Hechos del Data Mart**			
Datos demográficos		**Eventos adversos**		**Ensayo clínico**		**Tratamiento**		**Respuesta**		**Tiempo**		**Eficacia**		**Seguridad**	
E	**L**	**E**	**L**	**E**	**L**	**E**	**L**	**E**	**L**	**E**	**L**	**E**	**L**	**E**	**L**
T		**T**		**T**		**T**		**T**		**T**		**T**		**T**	
298	298	2641	1094	10	10	1658	261	1800	192	1	1	186	159	967	967
11		3		0		3		2		0		4		2	

Tabla 4. Proceso de ETL hacia el Data Mart del producto hR3.

2.2.5.7 Definición del período de actualización de los datos

Kitchen: es el programa encargado de ejecutar los trabajos, diseñados en Spoon en extensión XML o en un repositorio de BD. Usualmente los trabajos son programados por lotes para que sean ejecutados en intervalos regulares de tiempo (14).

Para definir el período con el cual se actualizarán los datos se crea un trabajo donde lo primero que se hace es agregar el componente *start*, que es el que se programa para que se ejecute el trabajo en el momento que sea necesario; puede ser un intervalo de tiempo determinado en días, semanas o meses, en este caso se programó para que se ejecutara cada seis meses, ya que fue el tiempo estimado por los especialistas del CIM. Después de eso se verifica que tanto el servidor del DWH, como el servidor donde reside la información de los EC del CIM estén funcionando, para esto es necesario encuestar las dos computadoras (ver pasos 1 y 2 en la figura 23), y si alguna de ellas no

está funcionando se manda un mensaje al usuario y se aborta la ejecución del trabajo (ver pasos 1.1, 1.2, 2.1 y 2.2 en la figura 23). Si están en funcionamiento entonces se ejecutan las transformaciones en el orden que se hayan determinado, en caso de que alguna de ellas presente algún problema se desvía y manda un mensaje de error para notificar que la transformación no se ha podido ejecutar (ver paso 3.1 y 3.2 en la figura 23); en caso contrario, cuando todas se hayan ejecutado también se le comunica al usuario que las transformaciones se ejecutaron perfectamente (ver paso 4 en la figura 23). La siguiente imagen (ver figura 23) muestra el trabajo que se encarga de realizar lo anteriormente expuesto.

Figura 23. Trabajo que inicia el proceso de ETL

2.3 Validación del procedimiento propuesto

Posterior a la implementación del proceso de ETL, se valida el procedimiento propuesto a través del método Delphi. Para ello se seleccionó un panel de expertos auxiliándose de dos cuestionarios, en uno de ellos se verifica el nivel de conocimiento de los expertos que componen el panel y en el otro se comprueba el nivel de cumplimiento del procedimiento (ver anexo 4). Obteniéndose como resultado, luego de la aplicación de este método, la evaluación de Muy Adecuado. Todo este procedimiento quedó registrado en el documento Validación del procedimiento, almacenado en el expediente de proyecto. En el mismo se explican detalladamente los pasos realizados para dar cumplimiento a cada uno de los requerimientos indispensables en la validación del procedimiento.

Conclusiones

En este capítulo se elaboró un procedimiento para implementar el proceso de ETL, el mismo fue aplicado para integrar los datos del producto hR3 del CIM. Después de haber aplicado este procedimiento, se obtuvieron los siguientes resultados:

- Se diseñó la arquitectura del DWH.
- Se creó el Modelo Físico del Data Mart y el del *Staging area* para dar soporte a la arquitectura definida.
- Se crearon 18 transformaciones a partir del listado de las reglas del negocio (ver anexo 1).
- Se validó de Muy Adecuado el procedimiento propuesto a través del método de Delphi.
- Se pobló el DWH con los datos del producto hR3 del CIM.

CAPÍTULO 3: EVALUACIÓN DE LA IMPLEMENTACIÓN DEL PROCESO DE EXTRACCIÓN, TRANSFORMACIÓN Y CARGA

Introducción

En este capítulo se elabora una lista de chequeo para evaluar la implementación del proceso de ETL del DWH, debido a que en la bibliografía consultada no se encontró alguna forma de evaluación. Para ello se realiza primeramente una introducción al concepto en cuestión y luego se elabora dicha lista de chequeo a partir de algunos aspectos importantes encontrados en el ciclo de vida Kimball y en la metodología Hefesto.

3.1 ¿Qué es una lista de chequeo?

Se entiende por lista de chequeo a un listado de preguntas, en forma de cuestionario que sirve para verificar el grado de cumplimiento de determinadas reglas establecidas a priori con un fin determinado. Son un instrumento que contiene criterios o indicadores a partir de los cuales se miden y evalúan las características del objeto, comprobando si cumple con los atributos establecidos. Se utilizan básicamente en la práctica de la investigación que forma parte de un proceso de evaluación. En síntesis, la lista de chequeo es una herramienta confiable y manipulable, que permite registrar, clasificar y organizar todo tipo de elementos para una evaluación.

3.2 Elaboración de la lista de chequeo

Para elaborar la lista de chequeo, se tuvieron en cuenta los elementos de evaluación que no deben faltar una vez que se realice el proceso de ETL del DWH; estos elementos se encuentran referenciados en el Epígrafe 1.8 del Capítulo 1 de la presente investigación. La lista de chequeo contiene diferentes indicadores a evaluar los cuales se encuentran distribuidos en tres secciones fundamentales:

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

Elementos que forman parte de la estructura de la lista de chequeo:

- **Peso:** define si el indicador a evaluar es crítico o no.

- **Indicadores a evaluar:** son los indicadores a evaluar en las secciones **Estructura del documento, Semántica del documento e Indicadores definidos por la etapa**; estos últimos dependen de los elementos de evaluación definidos para la etapa del proceso de ETL del DWH. A un elemento de validación puede responder uno o varios indicadores.
- **Eval. (Evaluación):** es la forma de evaluar el indicador en cuestión. El mismo se evalúa de 1 en caso de que exista alguna dificultad sobre el indicador y 0 en caso de que el indicador revisado no presente problemas.
- **NP (No Procede):** se usa para especificar que el indicador no es necesario evaluarlo en ese caso.
- **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. Pueden o no existir señalamientos o sugerencias.

Una vez aplicada la lista de chequeo se detectan los indicadores evaluados de mal y con el objetivo de darles solución se especifican en una tabla de no conformidades, la cual presenta la siguiente estructura:

- **No.**: es un número consecutivo e indica la cantidad de no conformidades identificadas.
- **Elemento de evaluación**: se refiere a un número que identifica al elemento de evaluación para el cual se corresponden los indicadores identificados.
- **NC (No Conformidad)**: especifica la NC a la que se refiere.
- **Fase correspondiente**: especifica la fase del procedimiento a la que corresponde la NC encontrada.
- **Significación**: especifica si la NC es o no significativa, dependiendo si el indicador es o no crítico.
- **Recomendación**: especifica si la NC es una recomendación, es decir que no es de obligatorio cumplimiento que se solucione por parte de los especialistas técnicos.
- **Estado NC**: especifica el estado de solución en que se encuentra la NC, puede ser Pendiente o Solucionada.
- **Respuesta del equipo de desarrollo**: si es necesario se especifica la respuesta que le da el equipo de desarrollo a la NC.

Evaluación del proceso de ETL

Se aborta la revisión del proceso de ETL revisado si:

- Existen al menos dos indicadores críticos evaluados de mal en la sección Indicadores evaluados por la etapa, que posee la lista de chequeo.
- Más del 50 % de los indicadores a evaluar, están evaluados de mal.
- Se mantienen las NC de una revisión a otra.

Se evalúa de regular la calidad del proceso de ETL revisado si no cumple los criterios para ser abortado e:

- Incumple con los indicadores críticos a evaluar de las secciones **Estructura del documento** y **Semántica del documento**, que posee la lista de chequeo.
- Existe al menos un indicador crítico evaluado de mal.
- Existen al menos cinco indicadores no críticos evaluados de mal de la sección **Indicadores evaluados por la etapa**, que posee la lista de chequeo.

El proceso de ETL es evaluado de bien si no cumple con ninguno de los criterios anteriores y:

- No existe ningún indicador crítico evaluado de mal.
- Si la cantidad de indicadores no críticos evaluados de mal de la sección **Indicadores evaluados por la etapa**, que posee la lista de chequeo, no es mayor que cuatro.

3.3 Evaluación de la etapa de extracción, transformación y carga a través de la lista de chequeo

Lista de chequeo para evaluar el proceso de ETL realizado.

Estructura del documento					
Peso	**Indicadores a evaluar**	**Eval**	**(NP)**	**Cantidad de elementos afectados**	**Comentarios**
crítico	1. ¿Los entregables contienen las secciones obligatorias de la plantilla estándar definidas para un expediente de proyecto? (Portada, Control de Versiones, Reglas de Confidencialidad, Tabla de Contenidos y Contenido) (Ver Expediente de Proyecto)	0		0	
Indicadores definidos por la etapa					
Peso	**Indicadores a evaluar**	**Eval**	**(NP)**	**Cantidad de**	**Comentarios**

				elementos afectados	
	1. ¿La arquitectura satisface las necesidades del proyecto?	0		0	
	2. ¿La arquitectura soporta el incremento del proyecto?	0		0	
	3. ¿Se utilizó el menor número de transformaciones posibles al cargar los datos hacia el *Staging area*?	0		0	
crítico	4. ¿Se creó el Modelo Físico a partir del Modelo Lógico?	0		0	
crítico	5. ¿Cumple la implementación del proceso de ETL con la arquitectura definida?	0		0	
	6. ¿Se tuvo en cuenta los formatos fuentes y tipos de datos de las perspectivas de análisis?	0		0	
	7. ¿La extracción de los datos se realiza a partir de las fuentes de datos?	1		13	
	8. ¿Se realiza una limpieza de los datos antes de realizar la carga de los mismos?	0		0	
crítico	9. ¿No afectan las restricciones aplicadas sobre una dimensión para analizar una tabla hecho en específico, a otra(s) tabla(s) hecho(s) que comparten esta misma dimensión?	0		0	
crítico	10. ¿Se cargan primero los datos de	0		0	

	las tablas de dimensiones y luego los de las tablas de hechos?				
crítico	11. ¿De utilizar un esquema copo de nieve, se comienza cargando las tablas de dimensiones del nivel más general al más detallado?		x		
	12. ¿Se establecen las políticas de actualización y refresco de los datos?	0		0	
	13. ¿Se crearon claves en el DWH diferentes a las claves de los OLTP?	0		0	
Semántica del documento					
Peso	**Indicadores a evaluar**	**Eval**	**(NP)**	**Cantidad de elementos afectados**	**Comentarios**
crítico	1. ¿Se han identificado errores ortográficos en los entregables o en los modelos diseñados?	0		0	
crítico	2. ¿Se entiende claramente lo que se ha especificado en el documento?	0		0	
	3. ¿El número de página que aparece en el índice coincide con el contenido que se refleja realmente en dicha página?	0		0	

Tabla 5. Lista de chequeo

Durante la revisión de la implementación del proceso de ETL mediante la aplicación de la lista de chequeo elaborada; se identificó un indicador con dificultades del cual se generó una NC. Esta fue tratada durante la fase de implementación del proceso de ETL por parte de los especialistas técnicos, a través de la creación de transformaciones a partir de las reglas del negocio. De forma general la implementación fue evaluada de **Bien**, ya que no hubo ningún indicador crítico evaluado de mal, no existieron problemas con los formatos de las plantillas, no se encontraron errores ortográficos en los documentos revisados y fue solucionada la NC generada (ver tabla 6). La siguiente figura, representa el comportamiento de los indicadores en las diferentes secciones de la lista de chequeo, luego de evaluar la implementación del proceso de ETL.

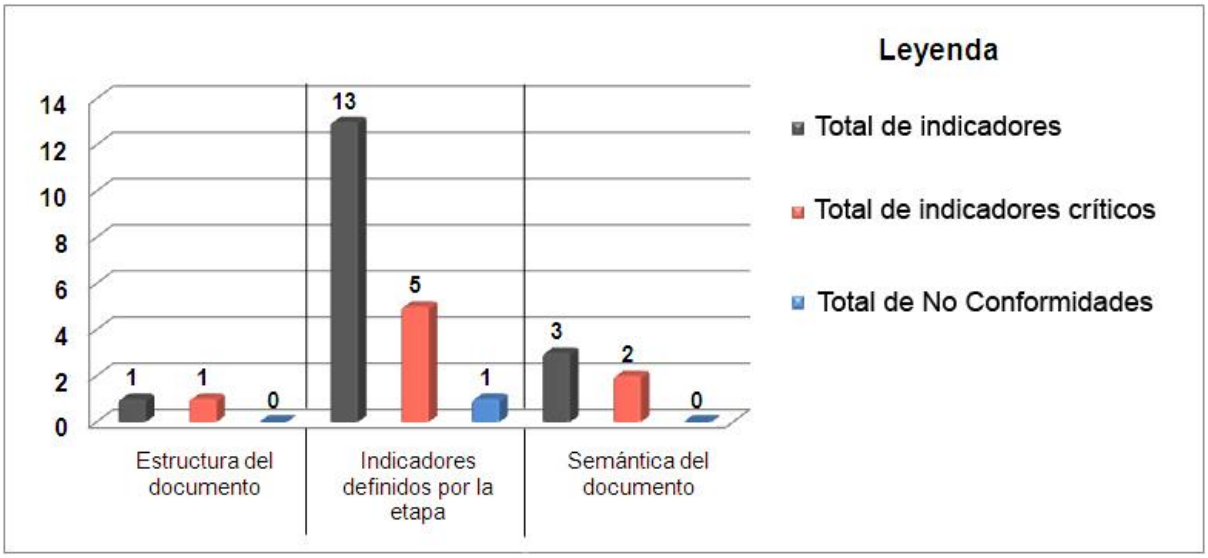


Figura 24. Comportamiento de los indicadores por secciones

No	Elemento de evaluación	No conformidad	Fase correspondiente	Significación	Recomendaciones	Estado NC	Respuesta del equipo de desarrollo
1	13	No se definieron 13 de las correspondencias que deben existir entre las variables de las tablas, BD o modelos determinados con sus perspectivas	Implementación del proceso de ETL.	No significativa		Resuelta	No se definieron porque no se encontraron los valores de las perspectivas en los modelos, debido a que cuando se condujo el ensayo no se recogieron estos datos. No obstante se decidió para cada caso darle solución a través de la creación de transformaciones a partir de las reglas del negocio.

Tabla 6. Listado de No Conformidades

Conclusiones

En este capítulo se confeccionó una lista de chequeo para evaluar la implementación del proceso de ETL del DWH de los EC del CIM. Posterior a la evaluación del proceso de ETL a través de la lista de chequeo se alcanzaron los resultados que se relacionan a continuación:

- Se identificaron 17 indicadores necesarios e imprescindibles para la evaluación final de la etapa de implementación del proceso de ETL, distribuidos en tres secciones: Estructura del documento, Indicadores definidos en la etapa, Semántica del documento; de ellos 8 indicadores con peso crítico.
- Se evaluó un indicador no crítico con dificultades en la sección Indicadores definidos por la etapa, lo cual no afectó la evaluación final de la etapa de implementación del proceso de ETL.
- Se lograron resultados satisfactorios en el procedimiento y la implementación del proceso de ETL realizado, obteniéndose una evaluación final de Bien.

CONCLUSIONES

Como consecuencia de la investigación desarrollada en el CIM acerca de la necesidad de integrar la información que se gestiona en este centro; se realizó la implementación del proceso de ETL de un DWH, lográndose los siguientes resultados:

- Se elaboró un procedimiento para implementar el proceso de ETL de un DWH para el CIM.
- Se diseñó la arquitectura del DWH para los EC del CIM.
- Se realizó la implementación del proceso de ETL, quedando poblado un DWH con los datos pertenecientes al producto hR3 de los EC del CIM.
- Se elaboró una lista de chequeo para evaluar el proceso de ETL.
- Se evaluó de Bien la implementación del proceso de ETL a través de la aplicación de la lista de chequeo elaborada.

Con este Trabajo de Diploma se logró obtener un DWH para el CIM, que integra toda la información de los EC pertenecientes al producto hR3, facilitando el análisis estadístico para la toma de decisiones con respecto hacia dónde dirigir cada uno de los estudios en dicho centro.

RECOMENDACIONES

Se recomienda:

- Crear nuevos Data Mart para otros productos del CIM.
- Implementar el proceso de ETL en el sistema alasClínicas, haciendo uso del procedimiento propuesto.
- Utilizar la lista de chequeo para evaluar el proceso en otras áreas de conocimiento.
- Realizar el análisis de los datos integrados en el DWH del CIM.

REFERENCIAS BIBLIOGRÁFICAS

1. **Limones, Cristina.** Reumatología, I. F. R. Ensayos Clínicos. [En línea] 9 de 12 de 2009. [Citado el: 11 de 12 de 2009.] http://www.institutferran.org/ensayos_cl%C3%ADnicos.htm.

2. **Pérez, Rolando.** Molecular, C. D. I. Portal del CIM. [En línea] 199-. [Citado el: 4 de 12 de 2009.] http://www.cim.sld.cu.

3. **González Hernández, Delly Lien.** EUMEDNET Enciclopedia y Biblioteca Virtual. [En línea] 2006. [Citado el: 13 de 11 de 2009.] http://www.eumed.net/libros/2009a/514/Caracterizacion%20del%20Centro%20de%20Inmunologia%20Molecular.htm.

4. **Inmon, Bill.** Building The Data Warehouse. 4a. ed. Canadá : Wiley Publishing, Inc, 2005. ISBN: 9780764599446.

5. **Kimball, Ralph y Ross, Margy.** The Data Warehouse Lifecycle Toolkit. 2a. ed. Canadá : Wiley Publishing, Inc, 2002. ISBN: 0471200247.

6. **Bernabeu, Ricardo.** HEFESTO: Metodología propia para la Construcción de un DWH. Córdoba, Argentina : s.n., 2009.

7. **Ponniah, Paulraj.** Data Warehousing Fundamentals. 1a. New York : John Wiley, 2001. ISBN: 0471412546.

8. **Pentaho.** Kettle Pentaho Data Integration. [En línea] 2009. [Citado el: 3 de 2 de 2010.] http://kettle.pentaho.org/.

9. Talend. [En línea] 2010. [Citado el: 07 de 02 de 2010.] http://es.talend.com/index.php.

10. **The PostgreSQL Global Development Group.** PostgreSQL. [En línea] 2005. [Citado el: 27 de 1 de 2010.] http://www.postgresql.org/docs/8.0/interactive/index.html.

11. **Boyd, Emily y Kilani, Omar.** PostgreSQL. About. [En línea] 2010. [Citado el: 5 de 02 de 2010.] http://www.postgresql.org/about/.

12. **FDM User Choice.** Visual Paradigm for UML. [En línea] 5 de 3 de 2007. [Citado el: 18 de 3 de 2010.] http://www.freedownloadmanager.org/es/downloads/Paradigma_Visual_para_UML%20%28M%C3%8D%29_14720_p/.

13. **Díaz Morales, Themis Patricia y Bermúdez Rodríguez, José Salvador.** Diseño de un Datawarehouse para los Ensayos Clínicos del Centro de Inmunología Molecular. 2010.

14. **Pelegrín Tamayo, Neyaris y Casaña Vinagera, Virgen.** Proceso de la migración de datos hacia un Data Warehouse para el módulo Análisis Químico del proyecto LIMS Control de Calidad. 2009.

BIBLIOGRAFÍA

- Bernabeu, Ricardo. HEFESTO: Metodología propia para la Construcción de un DWH. Córdoba, Argentina: s.n., 2009. [Consultada el: 4 de 12 de 2009.]
- Boyd, Emily y Kilani, Omar. PostgreSQL. About. [En línea] 2010. [Consultada el: 5 de 02 de 2010.]. Disponible en: http://www.postgresql.org/about/
- Calvo, Jorge Mario. 2005. ACIS. BI al alcance de todos. [En línea] 30 de 11 de 2005. [Consultada el: 27 de 01 de 2010.] http://www.acis.org.co/index.php?id=622
- Celma, M. Almacenes de Datos (Data Warehouse). 2000. 32 p.
- Diaz, M. and Y. Mestre. Gestcon Mart, Data Mart para la gestión del conocimiento. Ciudad de la Habana, Cuba, Universidad de las Ciencias Informáticas, Julio 2007 ,119. p.
- FDM User Choice. Visual Paradigm for UML. [En línea] 5 de 3 de 2007. [Consultada el: 18 de 3 de 2010.]. Disponible en: http://www.freedownloadmanager.org/es/downloads/Paradigma_Visual_para_UML%20%28M%C3%8D%29 14720_p
- Inmon, Bill. Building The Data Warehouse. 4a. ed. Canadá: Wiley Publishing, Inc, 2005. ISBN: 9780764599446. [Consultada el: 25 de 01 de 2009.]
- Juárez Giménez, Joan Carles. Fuentes de información biomédica. [Consultada el: 10 de octubre del 2009]. Disponible en: http://www.cedimcat.info/html/es/dir2471/doc26734.html
- Kimball, Ralph y Ross, Margy. The Data Warehouse Lifecycle Toolkit. 2a. ed. Canadá: Wiley Publishing, Inc, 2002. ISBN: 0471200247. [Consultada el: 25 de 01 de 2009.]
- Limones, Cristina. Reumatología, I. F. R. Ensayos Clínicos. [En línea] 9 de 12 de 2009. [Consultada el: 11 de 12 de 2009.]. Disponible en: http://www.institutferran.org/ensayos_cl%C3%ADnicos.htm
- Mundo Business Intelligence. Herramienta ETL (...o Mundo ETL). [En línea] [Consultada el: 30 de 01 de 2010.] http://mundobi.wordpress.com/2007/06/24/herramientas-etl-%E2%80%A6o-mundo-etl/
- Pecos, Daniel. PostGreSQL vs. MySQL. [En línea] [Consultada el: 28 de 01 de 2010.] Disponible en: http://www.netpecos.org/docs/mysql_postgres/index.html
- Pentaho. Kettle Pentaho Data Integration. [En línea] 2009. [Consultada el: 3 de 2 de 2010.]. Disponible en: http://kettle.pentaho.org/
- Pérez, Rolando. Molecular, C. D. I. Portal del CIM. [En línea] 199-. [Consultada el: 4 de 12 de 2009.]. Disponible en: http://www.cim.sld.cu

- pgAdmin . PostgreSQL Tool. [En línea] [Consultada el: 28 de 01 de 2010.] Disponible en: http://www.pgadmin.org
- Ponniah, Paulraj. Data Warehousing Fundamentals. 1a. New York: John Wiley, 2001. ISBN: 0471412546.
- Portada sobre la plataforma Pentaho Open Source Business Intelligence . La plataforma Pentaho Open Source Business Intelligence. [En línea] [Consultada el: 28 de 01 de 2010.] Disponible en: http://pentaho.almacen-datos.com
- Talend. [En línea] 2010. [Consultada el: 07 de 02 de 2010.]. Disponible en: http://es.talend.com/index.php
- The PostgreSQL Global Development Group. PostgreSQL. [En línea] 2005. [Consultada el: 27 de 01 de 2010.]. Disponible en: http://www.postgresql.org/docs/8.0/interactive/index.html
- Zorrilla, Marta. 2007. Data Warehouse y OLAP. Universidad de Cantabria: s.n., 2007.

ANEXOS

Anexo # 1: Reglas del Negocio

En el estudio precedente se identificaron las siguientes Reglas del Negocio (13):

- Si un paciente está en un grupo de tratamiento donde se definió en el protocolo que recibe placebo (ningún producto) o es control, se define que el nivel de dosis es 0.
- Si el nivel de dosis es 0 y presenta eventos adversos, la causalidad del evento adverso es no relacionada.
- La variable que identifica a un paciente en las BD es una concatenación entre sus iniciales y el número de inclusión en el ensayo.
- La variable que identifica a un paciente en las BD de gliomas (*Glioma 053* y *Glioma 069*) no puede ser solamente una concatenación entre sus iniciales y el número de inclusión en el ensayo pues hay pacientes que forman parte del estudio en ambos ensayos. Se determinó que la variable que identifica a los pacientes en estas BD fuera una concatenación entre sus iniciales, el número de inclusión en el ensayo y la localización (Ejemplo: *JSB-9-Glioma053*, *JSB-9-Glioma069*).
- Cuando la localización del tumor es mama (BD *Mama 070*) se asume que el sexo del paciente es femenino.
- Cuando el ensayo se aplica a pacientes que se encuentran en fase I aparece el nivel de dosis recibido; sin embargo, si se está en fase II, lo que aparece es el grupo de tratamiento y se debe buscar en el protocolo el nivel de dosis de cada grupo.
- Si la intensidad del evento adverso es normal entonces no se incluye como evento adverso.
- En el caso de las perspectivas que no se les encuentren relación con los datos fuentes en algunas de las BD porque no se recogieron en ese ensayo (Ejemplo: raza, causalidad, grado de diferenciación) se registra en el DWH como *missing*.
- Cuando no aparece la edad se toma la diferencia que existe entre la fecha actual y la fecha de nacimiento.
- En el caso de la BD *C y C 046*, se generan los estados generales aleatorios de ECOG 0, 1 ó 2 para que los pacientes cumplan con los criterios de inclusión en el ensayo.
- En el caso de la BD *Metacerebral 079* el cliente determinó que el estadío para todos los pacientes era fase IV.
- La clasificación anatomopatológica está recogida en las bases de datos en diferentes variables. Se determinó concatenar estas variables para obtener esta clasificación.

- El grado de diferenciación forma parte de la clasificación anatomopatológica y a los especialistas del negocio les hace falta tenerlo separado pues es otra perspectiva de análisis. Se determinó extraerlo de la clasificación anatomopatológica según los valores que puede tomar esta perspectiva.
- El sexo se recoge en la mayoría de las bases de datos como 1 y 2. Se determinó cambiar estos valores por masculino y femenino respectivamente.
- La raza se recoge en la mayoría de las bases de datos como 1, 2, 3 y 4. Se determinó cambiar estos valores por blanca, negra, mestiza y amarilla respectivamente.
- Los únicos tratamientos previos que se tendrán en cuenta para el análisis serán quimioterapia, radioterapia y cirugía. Se determinó que en los ensayos donde no se hayan recogido, ese campo aparecerá como *missing*.
- Para los tipos de respuestas se determinó que independientemente de las respuestas que pueda tener un mismo paciente a los diferentes eventos adversos, se selecciona la mejor respuesta (la mínima).
- Para el número de dosis se determinó que independientemente de la cantidad de dosis que reciba, se selecciona la última recibida (número máximo de dosis).
- Para los tratamientos previos se determinó que si el paciente tiene al menos uno de los tres (quimioterapia, radioterapia o cirugía) ya se asume que el paciente tuvo algún tratamiento previo.
- El estadío para el caso de la BD *Mama 070* se recogió como IIIa y IIIb. Se determinó que todos los pacientes incluidos en este ensayo tenían estadío III.
- En el caso de la causalidad e intensidad de los eventos adversos se determinó recogerlas como mismo está establecido en la identificación de las variables informacionales.
- En las BD *T.Sólidos 035*, *Metacerebral 079*, *Glioma 053* y *Glioma 069* no se recogió ECOG sino Karnofsky. Se determinó establecer la correspondencia que existe entre ECOG y Karnofsky a través de la escala evaluación de la capacidad funcional (ver anexo 2).

Anexo # 2: Evaluación de la capacidad funcional

Escala de ECOG		Escala de Karnofsky	
Grado	Descripciones	Porcentaje	Descripciones
0	Capaz de llevar a cabo una actividad física normal sin restricciones.	100	Normal. No presenta síntomas o signos de enfermedad.
		90	Capaz de una actividad normal, ligeros síntomas o signos de enfermedad
1	Sintomático, pero ambulatorio. Restricción en actividades físicas vigorosas, pero ambulatorio y capaz de hacer trabajos ligeros o de naturaleza sedentaria.	80	Actividad normal con esfuerzo. Algunos síntomas o signos de enfermedad.
		70	Puede cuidar de sí mismo. Incapaz de desarrollar una actividad o trabajo activo normales.
2	En cama menos del 50 % del tiempo. Ambulatorio y capaz de valerse por sí mismo, pero incapaz de trabajar. Más del 50 % del tiempo fuera de la cama.	60	Precisa ocasionalmente asistencia, pero es capaz de atender por sí mismo a la mayor parte de sus propias necesidades.
		50	Requiere asistencia y frecuentes cuidados médicos.
3	Capaz de realizar sus cuidados personales, pero más del 50 % del tiempo confinado a la cama o silla.	40	Incapacidad. Encamado. Requiere asistencia y cuidados especiales.
		30	Grave incapacidad. Estado severo. La muerte no es inminente. Requiere hospitalización.
4	Completamente incapaz de realizar ningún esfuerzo, confinado totalmente a la cama.	20	Estado grave. Intenso tratamiento de sostén. Requiere hospitalización Requiere hospitalización.
		10	Estado muy grave (moribundo). Proceso fatal que progresa rápidamente.
5	Muerto	0	Muerto

Anexo # 3: Modelo Lógico

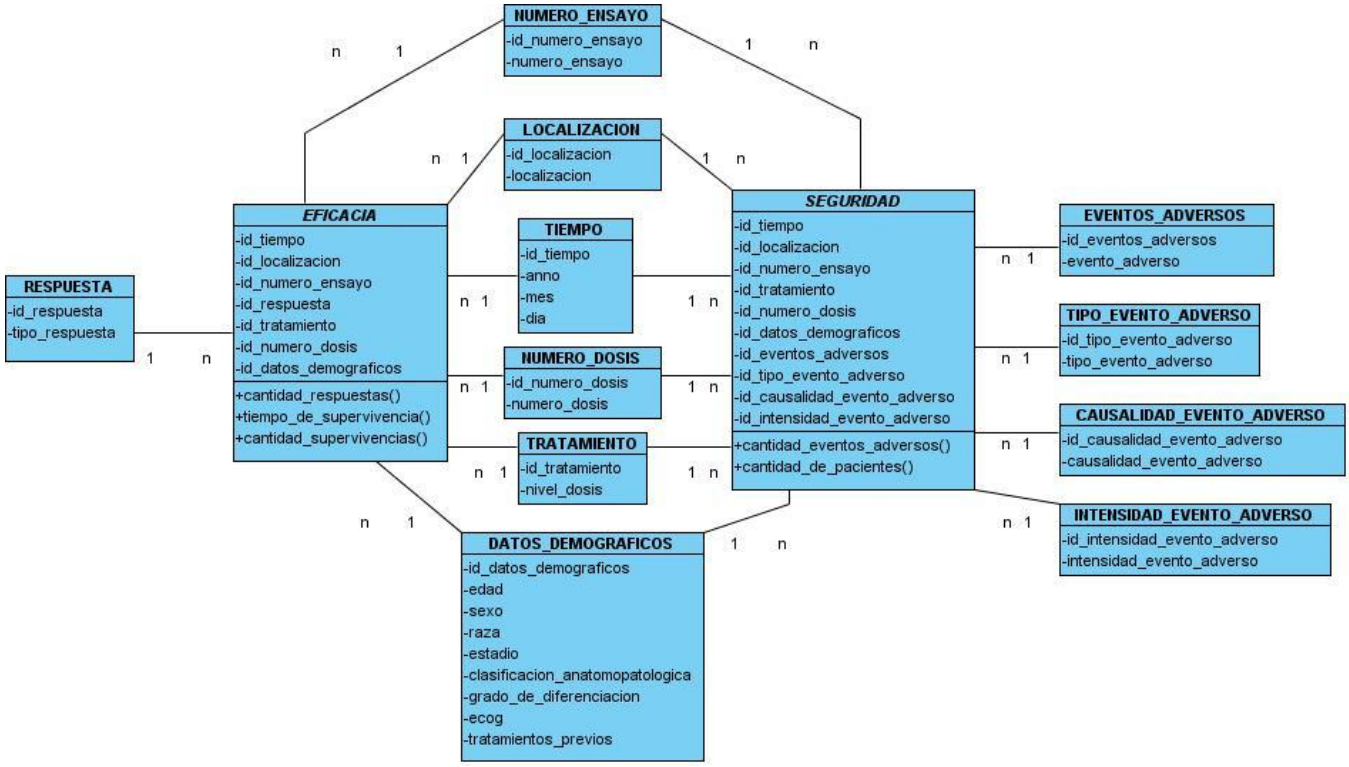


Figura 25. Modelo Lógico

Anexo # 4: Cuestionario para validar la propuesta de procedimiento

La presente encuesta forma parte de la aplicación del método de valoración de expertos. Con este fin se solicita su valiosa colaboración para validar si las etapas, actividades así como los artefactos de entrada y salida que se propusieron son correctos, para lograr este objetivo se ha elaborado un conjunto de preguntas que permiten medir la efectividad del modelo. De antemano se le asegura que nadie podrá saber quién es el encuestado y además se garantiza que sus opiniones se tendrán en cuenta para la posterior aplicación del procedimiento para la etapa de ETL de un DWH para los EC que se gestionan en el CIM.

Valore el grado de factibilidad de cada pregunta de acuerdo a la siguiente escala:
Muy Adecuado (C1)
Bastante Adecuado (C2)
Adecuado (C3)
Poco Adecuado (C4)
No Adecuado (C5)

Preguntas	Criterio del experto				
	C1	C2	C3	C4	C5
1. La utilidad de un procedimiento para aplicarlo a la etapa de ETL de un DWH con el objetivo de lograr la integración de los datos de los EC que se gestionan en el CIM es:					
2. Los siguientes aspectos forman parte del procedimiento. Categorice cada uno de ellos:					
2.1 Las etapas propuestas.					
2.2 Los pasos dentro de cada una de las etapas.					
2.3 Los artefactos propuestos para cada una de las etapas.					
3. La creación de transformaciones a partir de las reglas del negocio, para lograr la integración de los datos de los EC que se gestionan en el CIM es:					

4. La actualización del diccionario de datos en el procedimiento, para establecer la correspondencia entre las perspectivas, tipos de variables de los datos fuentes y el formato en que se encuentra la información es:					
5. El diseño de la arquitectura y la definición del modelo de despliegue en el procedimiento son:					
6. La correspondencia que existe entre la elaboración y aplicación del procedimiento propuesto es:					
7. En sentido general emita su criterio acerca de los aportes que pueda tener la aplicación del procedimiento para los EC que se gestionan en el CIM, a la hora de adicionar la información relacionada con otros productos al almacén de datos.					

GLOSARIO DE TÉRMINOS

A continuación se presentan los términos que podrían resultar de difícil comprensión, nuevos al lector o de diversos significados dependiendo del contexto que se analiza. Esta sección tiene como objetivo facilitar la comprensión del contenido expuesto en el documento.

Batch: Es un archivo de procesamiento por lotes.

Data Mart: son subconjuntos de datos con el propósito de ayudar a que un área específica dentro del negocio pueda tomar mejores decisiones.

DBMS: Sistema gestor de bases de datos.

Dimensión: característica de un hecho que permite su análisis posterior en el proceso de toma de decisiones y brinda una perspectiva adicional a un hecho dado.

Disparadores: *triggers*.

Hecho: operación que se realiza en el negocio la cual está estrechamente relacionada con el tiempo y es objeto de análisis para la toma de decisiones.

Indicador: generalmente es un valor numérico y representan lo que se desea analizar concretamente.

JDBC: Protocolo de conexión de Java a base de datos (del inglés Java Data Base Connectivity).

Mapeo: Proceso de convertir los datos que son transmitidos en un formato por el remitente, al formato de datos que puede ser aceptado por el receptor.

Metadatos: datos acerca de los datos que describen los contenidos del DWH.

NC: No Conformidad.

OLTP: Procesamiento de transacciones en línea (del inglés On Line Transaction Processing).

Perspectiva: se refiere a un objeto mediante el cual se quiere examinar un indicador, con el fin de responder a una pregunta planteada.

Staging area: Área de almacenamiento temporal.

TCP/IP: Protocolo de internet.