Universidad de las Ciencias Informáticas

FACULTAD 6

Titulo: "Proceso de la migración de datos hacia un Data Warehouse para el módulo Análisis Químico del proyecto LIMS Control de Calidad."

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

Autores: Neyaris Pelegrín Tamayo

Virgen Casaña Vinagera

TUTOR: MsC. Maypher Román Durán

Co-tutor: Lic. Roberto Acosta González

Junio 2009

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 _____.

_____ _____

Autor: Neyaris Pelegrín Tamayo Autor: Virgen Casaña Vinagera

_____ _____

Tutor: MsC. Maypher Román Durán Co-tutor: Lic. Roberto Acosta González

Uno debería guardarse contra aquellos que sermonean habitualmente a los jóvenes con la importancia del éxito como principal propósito en la vida. El estímulo más importante para el trabajo, en la escuela y en la vida, es el placer de trabajar, el placer de sus resultados, y el conocimiento del valor del resultado para la comunidad.

Albert Einstein

Opinión del tutor:

Agradecimientos

Este trabajo pudo ser escrito gracias al apoyo, colaboración y paciencia de muchas personas. Por lo mismo es probable que olvidemos mencionar a algunos de ellos en estas líneas, a ellos les pedimos disculpas.

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

A nuestros padres, amantes de sus hijos, y con esperanza desmedida de la utilidad y significado de la formación académica, por su apoyo moral y espiritual en todo momento.

A nuestro tutor y cotutor MsC. Maypher Román Durán e Ing. Roberto Acosta González respectivamente, que con su seriedad científica debidamente matizada de criticidad y humor, hicieron de esta experiencia académica algo muy satisfactorio.

En general queremos agradecer a nuestros amigos Alexander Rodríguez Torres, Yosvanis Sánchez Corales, Vladimir Urquia Cordero y al dúo de tesis del DW, por ser tan atentos, por su ayuda y su empeño en la realización de este trabajo.

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

A nuestras amigas, Yadira, Yailin, Rodaisy y Lorena, gracias por su ayuda, por su cariño, por su amistad.

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

Con cariño y admiración gracias a todos.

Dedicatoria Neyaris

A mí adorada madre Nilda Tamayo, que siempre me incitó a luchar por el futuro, por todo ese amor sin límites, por su apoyo, dedicación, sacrificio y por ser guía y ejemplo para mí. No hay una sola línea de este trabajo donde no estés presente. Nunca me alcanzará el tiempo para hacer por ti lo que has hecho por mí. Siempre has confiado ciegamente en todo lo que he hecho y hacer que te sientas orgullosa ha sido la meta desde que tuve conciencia de ello. Eres la mejor madre del mundo y sabes que nunca te fallaré.

A mí querida abuela Nina, por confiar en mí, por su dedicación, su amor, su aliento y sus consejos.

A Dagoberto Alvarado por ser un ejemplo a seguir, por su experiencia, por ayudarme y enseñarme, por ser sobre todas las cosas un padre para mí.

A mi amigo Alexander Rodríguez, por siempre estar ahí en las buenas y en las malas, por su paciencia, por su ayuda y apoyo incondicional en todos estos años de carrera, sin ti, todo habría sido más difícil, gracias.

A mis hermanos Reinier, Yuri ya mi amiga Lisandra, por esa forma tan especial de quererme y brindarme su apoyo, por siempre estar ahí para mí, los quiero mucho.

A todos los miembros de mi familia, por darme ánimo en todo momento.

A mis compañeros de aulas, a mis amigos y a todos los que de una forma u otra hicieron mi estancia más placentera en esta universidad.

A mi amiga y compañera de tesis Virgen Casaña Vinagera gracias por confiar en mí, por tu paciencia y por darme la posibilidad de trabajar contigo. Ha sido un placer. Gracias.

Dedicatoria Virgen

A mis queridos e idolatrados padres Cristina Vinagera Gerez y Rafael Juan Casaña Fuentes, muchas gracias por su amor, su comprensión, su confianza y por ser la luz que me guió durante todos estos años. Espero que se sientan orgullosos de mi cómo mismo me he sentido yo de ustedes todos estos años, gracias por haberme educado con esos principios sin los cuales hoy no podría ser lo que soy, por darme las fuerzas necesarias para luchar y así poder realizar mis sueños, por todo el sacrificio y por lo que han puesto de su parte para que todo esto sea posible, por su gran apoyo incondicional, este trabajo es de ustedes, un millón de gracias y les prometo que no los defraudaré.

A mi hermanita querida, Yoliannis Casaña Vinagera que siempre ha estado a mi lado, sin importar lo difícil que haya sido el momento, y si hoy estoy aquí en gran medida te lo debo a ti. Gracias Yoli por apoyarme, por los sacrificios hechos, por los consejos dados, por el gran ejemplo que has sido para mí y por compartir mi alegría.

A mi adorada abuela por su amor desmedido, por su fe en mi, por demostrarme que siempre hay que seguir hacia adelante, gracias por todo.

A mis tíos Guillermo y Yaumara Miur Vinagera por siempre estar ahí en las buenas y en las malas. ¿Cómo no agradecerles todo lo que han hecho por mí y por mi familia?

Al resto de mi familia por confiar en mí, por estar al tanto y darme ánimo.

A mis grandes amigos Yosvani Sánchez Corales y a Alexander Rodríguez Torres, por aguantarme durante tanto tiempo, por ayudarme con la tesis, por estar ahí cada vez que lo necesite sin importar para lo que fuera y porque siempre trataban de levantarme el ánimo. Muchas gracias a los dos.

A todos mis amigos y conocidos, a mis profesores, a mi piquete de la secundaria y en especial a mis compañeros de la Vocacional, protagonistas de que ahora yo esté aquí. Ustedes son los mejores amigos que he tenido en mi vida. No importa lo lejos que estemos.

A mis compañeros de grupo y a todas las maravillosas personas que he conocido a lo largo de estos cinco años, gracias a ustedes seré cada día mejor. Los recordare siempre.

A mi amiga y compañera de tesis Neyaris Pelegrín Tamayo, por su persistencia, su afán, su preocupación y por confiar en mí para la realización de este trabajo. Gracias.

Resumen

Con la informatización de la sociedad y dentro de ésta las empresas, ha crecido a nivel mundial la capacidad de generación y almacenamiento de la información, que no puede ser analizada por los métodos tradicionales existentes, mientras mayor es la capacidad para almacenar más y más datos mayor es la incapacidad para extraer información realmente útil de éstos en las empresas. Sin embargo, lo que constituye un valioso recurso para todos, se ha tornado en el gran problema de principios de siglo, manejar de forma óptima grandes volúmenes de información. Debido a todo lo mencionado se ve la necesidad de la implementación de un *Data Warehouse (DW)* para el proyecto Lims Control de Calidad y una de las etapas más importantes del en el desarrollo de un Data Warehouse es el proceso de extracción, transformación y carga de los datos (ETL).

Palabras Claves:

Data Warehouse

Extracción

Transformación

Carga

Datos

Volumen

Almacén

Información

Integración

Acceso

Índice:

Índice de Figuras:

Índice de Tablas:

Introducción

Mucho se ha hablado de la Era de la Información y sus ventajas; con las nuevas posibilidades se acortan las distancias y crecen los beneficios para quienes tienen acceso al gran caudal de datos. Sin embargo, lo que constituye un valioso recurso para todos, se ha tornado en el gran problema de principios de siglo. ¿Cómo *manejar de forma óptima grandes volúmenes de información?*

Un Data Warehouse provee dos beneficios empresariales reales: Integración y Acceso de datos. Un Data Warehouse elimina una gran cantidad de datos inútiles y no deseados, como también el procesamiento desde el ambiente operacional clásico. (1) Soporta el procesamiento informático al proveer una plataforma sólida, a partir de los datos históricos para hacer el análisis. Facilita la integración de sistemas de aplicación no integrados. Organiza y almacena los datos que se necesitan para el procesamiento analítico, informático sobre una amplia perspectiva de tiempo.

Se plantea que deben conformarse orientados hacia materias o temas (por ejemplo, clientes o productos), con datos perfectamente integrados y coherentes con respecto al nombre de las variables, los formatos de los campos, la medida de los atributos y la codificación de estructuras, con información histórica para comparar datos en distintos períodos de tiempo e identificar tendencias. (2) Toda esta información, una vez incorporada al Data Warehouse debe mantenerse, en general, invariable, cargándose pocas veces en el tiempo y no permitiendo actualizaciones frecuentes.

Unos de los componentes fundamentales y quizás el más crítico es el proceso de integración de datos en un repositorio que permita almacenar la información ya consolidada para ser explotada por herramientas de análisis.

Detrás de la arquitectura de un Data Warehouse existe un conjunto básico de procesos de suma importancia para el mismo, entre los cuales se pueden mencionar algunos elementales como:

- El proceso de ***extracción***, que consiste en estudiar y entender los datos fuente, tomando aquellos que son de utilidad para el *almacén.*
- El proceso de ***transformar*** a una forma presentable y de valor para los usuarios

- La ***carga*** de los datos en el Data Warehouse

De estos procesos, es muy importante para las empresas prestarle atención a la **transformación de datos**, donde se incluyen operaciones de corrección de errores, resolución de problemas de dominio, borrado de campos que no son de interés, generación de claves, agregación de información, etc. (3) La transformación de datos es necesaria porque no siempre los datos están en la forma más adecuada para poder aplicar los métodos que hacen falta para la tarea que se ha de llevar a cabo y el modelo que se quiere obtener. La fase de extracción, transformación y carga de datos, aunque parezca sencilla, conlleva aproximadamente el 70% del esfuerzo en los proyectos de un Data Warehouse.

Una vez extraídos los datos, se integran y almacenan en un Data Warehouse. La meta de un almacén de datos es integrar aplicaciones a nivel de datos. El dato extraído de los sistemas operacionales se procesa, transforma y ubica de acuerdo a un esquema similar a un modelo entidad/relación. (3)

La universidad, enfrenta un gran reto ante la era del conocimiento donde existe un mercado internacional muy dinámico y competitivo, y la calidad se impone como un requisito indispensable. En los proyectos productivos de la Universidad de las Ciencias Informáticas, existen muchas dificultades respecto a la puesta en práctica de un Data Warehouse y la utilización de herramientas para el proceso ETL.

Al realizar un estudio relacionado con las dificultades presentes, se manifiesta la poca preparación del los integrantes del los equipos de desarrollo para realizar las diferentes fases que componen el proceso ETL. Los estudiantes responsables de estos procesos no se seleccionan de acuerdo a sus cualidades, expectativas y habilidades. El tiempo que se le dedica a estos procesos es muy poco y no se le da un tratamiento adecuado. Estas causas provocan una definición inadecuada del alcance del sistema que en algunas ocasiones visualizan elementos innecesarios, los usuarios en ocasiones no están seguros de sus necesidades.

El objetivo fundamental del proceso de extracción, transformación y carga de los datos, es ayudar al usuario en el entendimiento del pasado y mostrarle los elementos para la planeación del futuro de corto, mediano y largo plazo. Este proceso ayuda a resolver muchos problemas y aporta elementos valiosos en la toma de decisiones al personal encargado de las mismas, los cuales pueden ser analistas, especialistas o

directores, y en forma inmediata, por ejemplo, consultas en línea, sin necesidad de que el usuario final solicite que se elaboren y ejecuten procesos especiales. Por lo tanto, es importante asegurar que los datos del almacén sean adecuados, suficientes y seleccionados de acuerdo a las necesidades del usuario final. (4)

Descripción del problema

El Centro de Ingeniería Genética y Biotecnología (CIGB) es uno de los pilares fundamentales en el desarrollo de la biotecnología cubana. Enfocado a perfeccionar el sistema de salud, asume la responsabilidad de apoyar directamente el progreso económico y social del país. Actualmente se lleva a cabo el proceso de informatización de esta institución que espera aprovechar al máximo las múltiples ventajas que ofrecen las TIC.

El CIGB manipula una amplia gama de datos de todos los procesos que realiza. Se hace necesario llevar un control estricto de la información que resulta importante para un buen desempeño de la institución. Como una solución informática que maneje una cantidad significativa de datos surgen a finales del pasado siglo los Sistemas de Información. Con el objetivo de apoyar las actividades de una empresa estos sistemas aprueban la entrada de datos, el almacenamiento, el procesamiento de grandes volúmenes de información y permiten generar una salida.

Continuar con los paradigmas de épocas anteriores ya no resulta factible. El desarrollo actual exige a las empresas que produzcan con la calidad requerida para poder entrar en la competencia del mercado mundial. Bajo la premisa de garantizar un producto eficaz se propone incorporar al entorno del CIGB los beneficios y mejoras de los Sistemas de Gestión de Información de Laboratorios (LIMS, del inglés Laboratory Information Management System). Estas aplicaciones informáticas manejan variada y abundante información y resultan imprescindibles para la industria moderna. Además permiten almacenar, calcular y gestionar datos de distintas formas, aumentando la productividad y la eficiencia en las actividades realizadas en los laboratorios.

Todos los productos desarrollados en este centro llevan el sello de la calidad que los distingue y los hace únicos entre los de su tipo en el mundo. La Dirección de Calidad, destinada a garantizar la eficiencia, eficacia y seguridad en sus producciones, está constituida por la Dirección de Aseguramiento de la Calidad y la Dirección de Control de la Calidad. (5)

La Dirección de Aseguramiento de la Calidad garantiza que se lleven a cabo las acciones planificadas y sistemáticas que son necesarias para proporcionar la confianza de que sus productos y servicios satisfacen los requisitos de calidad establecidos, verifica el cumplimiento de las Buenas Prácticas de Producción (BPP), Buenas Prácticas de Laboratorio (BPL) y Buenas Prácticas Clínicas (BPC). La dirección está compuesta por dos departamentos y cuatro grupos de trabajo que se mencionan a continuación:

- Departamento de Inspección y Auditoría
- Departamento de Mejoramiento e Ingeniería de Calidad
- Grupo de Administración y Costos
- Grupo de Liberación
- Grupo de Documentación
- Grupo de Metrología

La Dirección de Control de la Calidad lleva a cabo una serie de funciones relacionadas con el muestreo, los ensayos, las especificaciones y la evaluación de la calidad de los productos que se desarrollan. Para obtener buenos resultados en el trabajo realizado cuenta con la ayuda de dos departamentos con sus grupos de trabajos específicos y 3 grupos de trabajo, a continuación se hace referencia a los mismos:

- Departamento Biológico
    - Grupo de Microbiología
    - Grupo de Ensayos Biológicos I
    - Grupo de Inmunoquímica
    - Grupo de Ensayos Biológicos II
    - Grupo de Biología Molecular
- Departamento Físico Químico
    - Grupo de Aseguramiento Analítico
    - Grupo de Cromatografía y Electroforesis.
    - Grupo de Análisis Químico
    - Grupo de Sistemas Críticos

- Grupo de Estudios de Estabilidad y Materiales de Referencia
- Grupo de Administración y Costo
- Grupo de Liberación Analítica

Los Sistemas de Gestión de Información de los Laboratorios (LIMS) proporcionan un conjunto de herramientas basadas en sistemas informáticos, que permiten gestionar, evaluar y almacenar información. (5)

Para facilitar el almacenamiento de estos importantes datos y ayudar al funcionamiento del Centro de Ingeniería Genética y Biotecnología de La Habana (CIGB) se hará la migración de los datos hacia un DWH, del módulo Análisis Químico utilizando procesos ETL.

Hay que ser cuidadoso para no generar falsos resultados, porque los clientes son cada vez más exigentes al adquirir los productos y esto hace casi imprescindible la necesidad de llevar a cabo un desarrollo correcto de los mismos. Además, en muchas ocasiones la implementación del sistema se comienza sin una previa validación de los objetivos planteados en el diseño, causas que afectan en la calidad del producto y que conllevan a retrasos y por ende incumplimientos en la fecha de entrega.

Hay una premisa en el mundo del negocio que plantea que el futuro pertenece a quienes puedan verlo y llegar a él primero. (4) Por tanto, es muy importante que el proyecto Lims Control de Calidad cuente con un Data Warehouse. Las instituciones que tengan automatizados todos o parcialmente sus procesos y cuenten con información acumulada sobre los mismos, se les recomienda la implementación de un Data Warehouse, ya que éste permite no solo comprender lo que está pasando, sino predecir lo que va a suceder.

Para realizar la migración del gran volumen de datos es necesario contar con técnicas de extracción, transformación y carga de datos (ETL).

Por lo que se plantea el siguiente problema a resolver:

¿Cómo ejecutar el proceso de migración de los datos del proyecto Lims Control de Calidad hacia el Data Warehouse de dicho proyecto?

El objeto de estudio:

El proceso de migración e integración de datos en proyectos informáticos.

El campo de acción:

Técnicas de extracción, transformación y carga de datos en el proceso de migración de los datos del proyecto Lims Control de Calidad de la facultad 6 de la Universidad de las Ciencias Informáticas (UCI).

Objetivo general:

Realizar la migración de los datos que se encuentran en la base de datos del proyecto para un Data Warehouse realizando un proceso extracción, transformación y carga de datos (ETL).

Objetivos específicos:

1. Realizar el diseño e implementación de los procedimientos de extracción de los datos desde los sistemas de información originales (conocido como OLTP, On Line Transaccional Processing).
2. Realizar el diseño e implementación de los procedimientos de transformación de los datos.
3. Realizar el diseño e implementación de los procedimientos de carga de los datos hacia el Data Warehouse.

Tareas investigativas:

1. Realizar análisis sobre las técnicas de extracción, transformación y carga de datos.
2. Estudiar posibles tecnologías a utilizar para el proceso de extracción, transformación y carga de datos.
3. Analizar los OLTP.
4. Realizar el diseño e implementación el proceso de extracción, trasformación y carga de los datos hacia un Data Warehouse.

Idea a defender:

Con el empleo de técnicas de extracción, transformación y carga de datos realizar el proceso de migración de los datos desde la base de datos del proyecto Lims Control de Calidad hacia el Data Warehouse de dicho proyecto.

Estructura de la tesis:

La tesis, quedará estructurada de la siguiente manera introducción, tres capítulos, conclusiones, recomendaciones, bibliografía, anexos y glosario de términos.

Capitulo 1: Fundamentación teórica.

Se realiza la fundamentación teórica de la investigación donde se incluye el estado del arte a nivel internacional, nacional y de la universidad; las técnicas, las tecnologías y la metodología que pueden servir de apoyo para la solución del problema. Además se abordan temas y conceptos relacionados con Extracción, Transformación y Carga de Datos para el proceso de migración para un Data Warehouse.

Capítulo 2: Análisis de los OLTP.

En este capítulo y teniendo en cuenta que ya se han detallado claramente las características generales del proceso de Extracción, Transformación y Carga de datos, se definirán y describirán todos los componentes que intervienen en su arquitectura o ambiente. Se hará una selección y argumentación de los requisitos funcionales del sistema propuesto .Además de analizar los OLTP para luego diseñar e implementa los procesos ETL.

Capítulo 3: Diseño e Implementación del proceso ETL.

En este último capítulo además de realizar el diseño e implementación de las transformación de datos se da algunos puntos de vista particulares, se mencionaran cuáles son las oportunidades que existen para trabajar en proyectos de este tipo, y qué es lo que se espera de los grandes almacenes de datos en los próximos años.

CAPÍTULO 1: Fundamentación Teórica

En este capítulo después de un estudio realizado en el tema del desarrollo de la tecnología, se brinda la posibilidad de conocer y hacer un análisis de los conceptos y elementos que son relevantes para la elaboración de un proceso de migración de los datos hacia un DW del proyecto Lims Control de Calidad para el Centro de Ingeniería Genética y Biotecnología de la Habana.

Se proporciona un conocimiento sobre el origen del proceso ETL, además de cómo ha evolucionado en las distintas etapas. También se abordará sobre la situación que tiene actualmente las herramientas ETL en el mundo, en Cuba y en la UCI (Universidad de las Ciencias Informáticas) y las tendencias, técnicas, tecnologías y metodologías que brindaron apoyo en la solución del problema.

1.1 ETL (Extracción, Transformación y Carga).

Los Data Warehouse surgen con la promesa del manejo y control de la información, los mismos aseguran una vista única de los datos, los cuales pueden provenir de diferentes fuentes, en este caso de diferentes Bases de Datos. Esto beneficia mucho a los usuarios finales pues no tienen necesidad de aprender a utilizar varios sistemas de acceso y manipulación de datos, con uno solo basta.

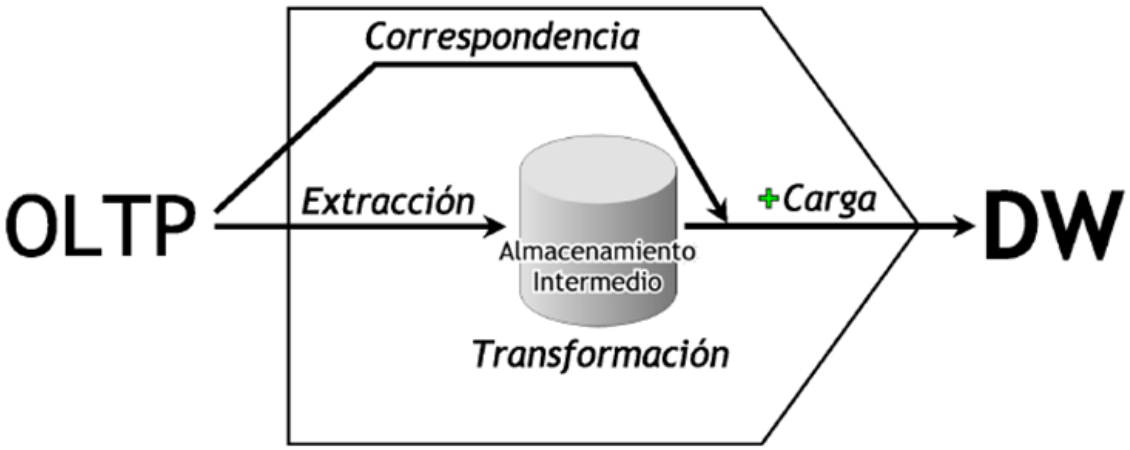
Esta nueva forma de almacenar datos es muy importante en el desarrollo de la sociedad y la tecnología, su uso es muy ventajoso en las empresas y se ha extendido a compañías o negocios, o sea a toda organización de producción o de servicios.

Todas las empresas que tienen automatizados, todos o parcialmente sus procesos o algún tipo de información acumulada que sea importante para su funcionamiento y desarrollo, están tratando de contar con la implementación de un Data Warehouse pues los sistemas relacionales, por mucho que se les trate de mejorar, perfeccionar y adaptar a las nuevas necesidades que están presentando los usuarios, no son los que resuelven los problemas que van surgiendo a medida que pasa el tiempo.

Como uno de los factores más importantes en la realización de un Data Warehouse está el proceso ETL, este constituye el proceso que permite a las organizaciones mover datos desde múltiples fuentes, reformatearlos, limpiarlos y cargarlos en otra Base de Datos o Data Warehouse para analizarlo y apoyar un proceso de negocio.

1.2 PROCESO ETL.

A continuación, se explicará en síntesis el accionar del proceso ETL y cuál es la relación existente entre sus diversas funciones. En la siguiente figura se puede apreciar mejor lo antes descrito.



Figuras 1: Proceso ETL. (6)

Los pasos que se siguen son:

- Se extraen los datos relevantes desde los OLTP.
- Estos datos se depositan en un almacenamiento intermedio.
- Se integran y transforman los datos para evitar inconsistencias.
- Finalmente los datos depurados son cargados desde el almacenamiento intermedio hasta el DW. Si existiesen correspondencias directas entre datos de los OLTP y el DW, se procede también a su respectiva carga. (6)

El proceso de ETL (Extracción, Transformación y Carga) forma parte de la Inteligencia Empresarial (Business Intelligence), también llamado “Gestión de los Datos” (Data Management).

La idea consiste en que una aplicación ETL lea los datos primarios de unas bases de datos de sistemas principales, realice transformación, validación, el proceso

cualitativo, filtración y al final escriba datos en un nuevo almacén de datos donde los mismos estarán disponibles para ser analizados posteriormente por los usuarios.

La primera fase para la realización de un Data Warehouse es un proceso de ETL que es sumamente necesaria y aunque es la menos glamorosa, es fundamental para su éxito.

La fase de ETL (Extracción, Transformación y Carga) es:

- **crítica** porque el resto de las fases del proyecto se alimentan de ella y no pueden comenzar hasta que esta haya concluido satisfactoriamente.
- una **fuente potencial de costes inesperados** dado que, si bien no debería absorber más allá del 60% del tiempo de ejecución del proyecto, no es extraño que llegue a acaparar el 90% de él y que, además, acabe provocando retrasos importantes en su ejecución.
- **específica** ya que la información que se extrae de las bases de datos tiene que ajustarse a los criterios de contenido, calidad y formato a los que los responsables de los sistemas de información de las empresas no están habituados.
- **difícil** pues conlleva el extraer e integrar datos de fuentes muy diversas y plataformas muchas veces heterogéneas y acceder a información contenida en sistemas que no están concebidos ni diseñados para las exigencias de un proceso masivo de análisis de datos.
- **multidisciplinar** a causa de que no se trata de una actividad meramente técnica, implica la adquisición en un plazo de tiempo corto, de la visión de negocio necesaria para comprender el valor funcional de la información, además, de los diferentes puntos de mira de los distintos usuarios finales. (3)

1.2.1 Función de 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, pueden ser Bases de Datos o de otras fuentes existentes. Los formatos de las fuentes normalmente se encuentran en Base de Datos Relacionales o ficheros planos, pero pueden incluir Bases de Datos no relacionales u otras estructuras diferentes. La extracción convierte los datos a un formato preparado para iniciar el proceso de transformación.

Una parte del proceso de extracción es la de analizar los datos extraídos, de lo que resulta un chequeo que verifica si los datos cumplen la pauta o estructura que se esperaba, de no ser así, los datos son rechazados.

Un aspecto importante que se debe exigir a la tarea de extracción es que ésta produzca un impacto mínimo en el sistema origen. Debido a esto se han tomado diferentes medidas, por ejemplo, si los datos a extraer son muchos, el sistema de origen se podría ralentizar e incluso colapsar, provocando que éste no pueda utilizarse con normalidad para su uso cotidiano, por esta razón, en sistemas grandes las operaciones de extracción suelen programarse en horarios o días donde este impacto sea nulo o mínimo.

1.2.2 Función de Transformación

La fase de transformación aplica una serie de reglas de negocio o funciones sobre los datos extraídos para convertirlos en datos que serán cargados. Algunas fuentes de datos requerirán alguna pequeña manipulación de los datos.

No obstante en otros casos pueden ser necesarias aplicar algunas de las siguientes transformaciones:

- Seleccionar sólo ciertas columnas para su carga (Ej. que las columnas con valores nulos no se carguen).
- Traducir códigos (Ej. Si la fuente almacena una "H" para Hombre y "M" para Mujer pero el destino tiene que guardar "1" para Hombre y "2" para Mujer).
- Codificar valores libres (Ej. convertir "Hombre" en "H" o "Sr" en "1").
- Obtener nuevos valores calculados (Ej. total _ venta = cantidad * precio).
- Unir datos de múltiples fuentes (Ej. búsquedas, combinaciones, etc.).
- Calcular totales de múltiples filas de datos (Ej. ventas totales de cada región).
- Generación de campos clave en el destino.
- Transponer o pivotar (girando múltiples columnas en filas o viceversa).
- Dividir una columna en varias (Ej. columna "Nombre: García, Miguel"; pasar a dos columnas "Nombre: Miguel" y "Apellido: García").
- La aplicación de cualquier forma, simple o compleja, de validación de datos, y la consiguiente aplicación de la acción que en cada caso se requiera:

1. Datos OK: Entregar datos a la siguiente etapa (Carga).
2. Datos erróneos: Ejecutar políticas de tratamiento de excepciones (Ej. Rechazar el registro completo, dar al campo erróneo un valor nulo o un valor *centinela*). (7)

1.2.3 Función de Limpieza de Datos

Su objetivo principal es el de realizar distintos tipos de acciones contra el mayor número de datos erróneos, inconsistentes e irrelevantes.

Las acciones más típicas que se pueden llevar a cabo al encontrarse con Datos Anómalos (Outliers) son:

- Ignorarlos.
- Eliminar la columna.
- Filtrar la columna.
- Filtrar la fila errónea, ya que a veces su origen, se debe a casos especiales.
- Reemplazar el valor.
- Discretizar los valores de las columnas. Por ejemplo de 1 a 2, poner “bajo”; de 3 a 7, “óptimo”; de 8 a 10, “alto”. Para que los outliers caigan en “bajo” o en “alto” sin mayores problemas.

Las acciones que suelen efectuarse contra Datos Faltantes (Missing Values) son:

- Ignorarlos.
- Eliminar la columna.
- Filtrar la columna.
- Filtrar la fila errónea, ya que a veces su origen, se debe a casos especiales.
- Reemplazar el valor.
- Esperar hasta que los datos faltantes estén disponibles.

Un punto muy importante que se debe tener en cuenta al elegir alguna acción, es el de identificar el por qué de la anomalía, para luego actuar en consecuencia, con el fin de

evitar que se repitan, agregándole de esta manera más valor a los datos de la organización. Se puede dar que en algunos casos, los valores faltantes sean inexistentes, ya que por ejemplo, un nuevo asociado o cliente, no poseerá consumo medio del último año.

1.2.4 Función de Carga

Este proceso es el responsable de cargar la estructura de datos del DW con:

- Aquellos datos que han sido transformados y que residen en el almacenamiento intermedio.
- Aquellos datos de los OLTP que tienen correspondencia directa con el depósito de datos.

Dependiendo de los requerimientos de la organización, este proceso puede abarcar una amplia variedad de acciones diferentes. En algunos casos las bases de datos sobrescriben la información antigua con nuevos datos.

Los Data Warehouse mantienen un historial de los registros de manera que se pueda hacer una auditoría de los mismos y disponer de un rastro de toda la historia de un valor a lo largo del tiempo.

Existen dos formas básicas de desarrollar el proceso de carga:

- Acumulación simple: La acumulación simple es la más sencilla y común, y consiste en realizar un resumen de todas las transacciones comprendidas en el período de tiempo seleccionado y transportar el resultado como una única transacción hacia el Data Warehouse, almacenando un valor calculado que consistirá típicamente en un sumatorio o un promedio de la magnitud considerada.
- Rolling: El proceso de Rolling por su parte, se aplica en los casos en que se opta por mantener varios niveles de granularidad. Para ello se almacena información resumida a distintos niveles, correspondientes a distintas agrupaciones de la unidad de tiempo o diferentes niveles jerárquicos en alguna o varias de las dimensiones de la magnitud almacenada (Ej. Totales diarios, totales semanales, totales mensuales, etc.). (7)

La fase de carga interactúa directamente con la base de datos de destino. Al realizar esta operación se aplicarán todas las restricciones y triggers (disparadores) que se

hayan definido en ésta (Ej. valores únicos, integridad referencial, campos obligatorios, rangos de valores). Estas restricciones y triggers (si están bien definidos) contribuyen a que se garantice la calidad de los datos en el proceso ETL (Extracción, Transformación y Carga), y deben de tomarse en cuenta.

1.3 Tareas del ETL

Los ETL, son los encargados de realizar dos tareas bien definidas:

- Carga Inicial (Initial Load).
- Actualización, mantenimiento o refresco periódico (siempre teniendo en cuenta un intervalo de tiempo predefinido para tal operación). (8)

La carga inicial, se refiere precisamente a la primera carga de datos que se le realizará al DW. Por lo general esta tarea consume un tiempo bastante considerable, ya que se deben insertar registros que han sido generados aproximadamente, y en casos ideales, durante más de cinco años.

Los mantenimientos periódicos mueven pequeños volúmenes de datos, y su frecuencia está dada en función del gránulo del DW y los requerimientos del usuario. El objetivo de esta tarea es añadir al depósito aquellos datos nuevos que se fueron generando desde el último refresco.

Antes de realizar una nueva actualización, es necesario identificar si se han producido cambios en las fuentes originales de los datos recogidos, desde la fecha del último mantenimiento, a fin de no atentar contra la consistencia del DW.

Para efectuar esta operación, se pueden realizar las siguientes acciones:

- Cotejar las instancias de los OLTP involucrados.
- Utilizar disparadores en los OLTP.
- Recurrir a Marcas de Tiempo (Time Stamp), en los registros de los OLTP.
- Hacer uso de técnicas mixtas.

Si este control consume demasiado tiempo y esfuerzo, o simplemente no puede llevarse a cabo por algún motivo en particular, existe la posibilidad de cargar el DW desde cero, este proceso se denomina Carga Total (Full Load). (8)

1.4 Herramientas ETL a nivel mundial.

Las herramientas ETL (Extracción, Transformación y Carga) son muy utilizadas en el mundo entero.

Por ejemplo en Córdoba (Argentina), dentro de los proyectos de Business Intelligence. El componente ETL es el que soporta esta característica y define un workflow de tareas a realizar, posee funcionalidades de notificaciones, eventos y hasta en algunos casos procesos extras al mundo de la Inteligencia de Negocios (BI por sus siglas en inglés Business Intelligence) como hacer réplica y/o uniformidad de bases de datos. BI es uno de los componentes fundamentales y quizás el más crítico es el proceso de integración de datos en un repositorio que permita almacenar la información ya consolidada para ser explotada por herramientas de análisis. (9)

Existen componentes ETL que proveen las herramientas BI y hasta empresas que se dedican exclusivamente a desarrollar este tipo de aplicaciones. En los últimos 5 años ha crecido el segmento de herramientas ETL más del 50 % en lo que se refiere a las ventas de licencias de productos y de servicios asociados a nivel mundial.

1.5 La utilización de los ELT se reflejan en:

- Data Warehouse: Casi el 80% del uso está relacionado con esta arquitectura. (Incluyendo Data Marts o BD (base de datos) que dan soporte a un proyecto.
- Tareas de Bases de datos: También se utilizan para consolidar, migrar y sincronizar bases de datos operativas.
- CDI: El Customer Data Integration es una función que permite unificar y homogenizar la cartera de clientes de las grandes corporaciones, disueltas y repetidas en diferentes fuentes.
- Gobierno: Algunos gobiernos consolidan información transaccional de todos los procesos magnéticos, electrónicos en un ODS (Almacén Operacional de Datos) para controlar en línea ciertas operaciones. (9)

El grupo Forrester ofrece una definición sobre el mercado ETL y analiza porqué es un mercado a seguir de cerca, examina el tamaño de éste durante un periodo determinado de tiempo, así como su evolución futura, ofrece un ranking comparativo de los principales vendedores de ETL según su cuota de mercado, y hasta brinda una serie de recomendaciones sobre posibles oportunidades del mercado en el sector.

El banco canadiense CIBC utilizará la solución Enterprise ETL Server de SAS Institute para obtener la información proveniente de diferentes fuentes y estandarizarla para su utilización por parte de las personas responsables de Préstamos e Hipotecas.

1.6 El uso de las ETL en Cuba.

En las empresas cubanas la entrada y avance de tecnologías de la información y las comunicaciones en lo que se refiere a conocimiento y desarrollo, tiene como objetivo fundamental el uso racional de los recursos, tratar de obtener la mayor productividad posible y los productos cada vez con una mejor calidad.

Por eso se ha pensado desde ya, en el uso de los almacenes de datos por ejemplo el Data Warehouse de la Corporación CIMEX está en tránsito de convertirse en uno de los compañeros inseparables de los comerciales y analistas de la corporación, ambientado con una interfaz agradable y asequible a través del ambiente Web, pone a disposición de todos los niveles, con las restricciones propias de acceso a la información que cada nivel exige, un mundo de posibilidades al alcance de sólo unos clics.

La Corporación CIMEX se dedica fundamentalmente a la exportación e importación de mercancías. Forman parte de ella un conjunto de empresas que se encuentran enfocadas en diversos negocios, como la red de Comercio Minorista y la Dirección de Logística, esta última dedicada al Comercio Mayorista. El sistema Data Warehouse comercial de la Corporación CIMEX centra su atención en la actividad del comercio, principalmente en la gestión de inventario, permitiendo una gestión de compra–venta eficiente, con una finalidad fundamental: *"Disminuir los costos, sin afectar al cliente, permitiendo prestaciones eficientes y con la calidad requerida, aumentando las ganancias o utilidades de las empresas".* (10)

En la CUJAE, centro de estudios universitarios que también abarca la rama de la informática, Sergio Luján Mora desarrollo un trabajo de tesis teórica y muy profunda de los Data Warehouse la cual ha servido como ejemplo para los seguidores de su trabajo. Sergio Luján Mora se encuentra hoy como profesor en la Universidad de Las Ciencias Informáticas (UCI).

1.7 Las ETL en la UCI, una nueva vía de desarrollo.

La Universidad de las Ciencias Informáticas está enfrascada en lograr que los proyectos cada día se hagan con una mayor calidad, y está tomando medidas para que esto se haga realidad. De ahí que hay que ganar experiencia en el uso de herramientas ETL para poderlas utilizar después en el diseño de grandes almacenes de datos.

Debido a la escasa práctica y uso de las mismas se ha decidido diseñar una que se pueda utilizar en el proyecto Lims Control de Calidad de la facultad 6, aunque anteriormente había surgido el tema del Data Warehouse en la facultad 2, específicamente en el proyecto SINSEC.

Es importante destacar también, que la facultad 4 no ha querido quedarse atrás en esta nueva fase de desarrollo tecnológico, por lo que estudiantes pertenecientes a la misma, desarrollaron un Data Mart para la gestión del conocimiento, algo que ha influido mucho en los nuevos seguidores del avance informático, ya que estos le pueden ofrecer apoyo y experiencia en su trabajo.

1.8 Las tendencias, técnicas, tecnologías y metodologías que brindaron apoyo a la solución del problema.

En la construcción de un Data Warehouse el mayor impacto que pueden presentar las personas que se dedican a su realización, está en el giro que se da en cuanto al aprendizaje, además de las muchas destrezas que se deben aprender a utilizar, incluyéndole a esto los conceptos y la estructura del Data Warehouse en sí.

Se introducen además una gran cantidad de tecnologías como: las de Extracción, Transformación y Carga (ETL), el Acceso a Datos, el Catálogo de Metadatos, la implementación de un Sistemas de Soporte a Decisiones (DSS) y Sistema de Información Ejecutiva (EIS), o sea que se cambia la manera en que se usa la tecnología existente. Las responsabilidades de soporte, las demandas de recursos y las expectativas que se van creando a medida del desarrollo del mismo, son los efectos de los nuevos cambios.

La construcción del Sistema ETL es responsabilidad del equipo de desarrollo del almacén de datos, se construye uno para cada almacén de datos y representa más o menos el 70 % del esfuerzo. Se pueden utilizar en su construcción herramientas del mercado o programas que han sido diseñados específicamente para esto.

La herramienta ETL seleccionada debe permitir lograr los resultados deseados en un tiempo relativamente menor que la forma tradicional de codificar y mantener los objetos del Data Warehouse.

1.8.1 Algunos aspectos a considerar para seleccionarla herramienta ETL más adecuada.

- Fácil de usar y comprensible desde el punto de vista del mantenimiento y el desarrollo de la perspectiva.
- El proceso ETL debe integrarse con el proceso de negocio.
- Debe soportar el procesamiento de grandes volúmenes de datos.
- Poder extraer datos desde distintas fuentes heterogéneas.
- Puede ser necesario que la herramienta ETL (Extracción, Transformación y Carga) soporte procesamiento paralelo.

Debe poseer un amplio espectro de conectividad y la habilidad de estandarizar los datos tomados desde diversas fuentes, que pueden estar incluso almacenadas en bases de datos soportadas sobre una plataforma diferente.

1.8.2 Herramientas Propuestas.

Periódicamente, se importan datos al almacén de datos de los distintos sistemas de la entidad y de otros sistemas de software relacionados con el negocio para la transformación posterior. Es práctica común normalizar los datos antes de combinarlos en el almacén de datos mediante herramientas de extracción, transformación y carga (ETL). Estas herramientas leen los datos primarios (a menudo bases de datos OLTP de un negocio), realizan el proceso de transformación al almacén de datos (filtración, adaptación, cambios de formato) y escriben en el almacén.

1.8.2.1 Pentaho Data Integration.

La corporación Pentaho es el patrocinador primario y propietario del proyecto Pentaho BI, el cual es una iniciativa en curso por la comunidad de Open Source que provee organizaciones con mejores soluciones para las necesidades de BI de su empresa.

Se puede describir BI, como un concepto que integra por un lado el almacenamiento y por el otro el procesamiento de grandes cantidades de datos, con el principal objetivo de transformarlos en conocimiento y en decisiones en tiempo real, a través de un sencillo análisis y exploración. La definición antes expuesta puede representarse a través de la siguiente fórmula:

Datos +Análisis =Conocimiento

Existe una definición muy popular acerca de BI, que dice: “Inteligencia de Negocios es el proceso de convertir datos en conocimiento y el conocimiento en acción, para la toma de decisiones”. (9)

El proyecto Pentaho BI abarca diferentes aéreas de reporte entre las que se encuentra:

Pentaho Data Integration:

Esta surge debido a que muchas de las organizaciones existentes tienen información disponible en aplicaciones y base de datos separados. Por esto Pentaho Data Integration brinda la posibilidad de abrir, limpiar e integrar esta valiosa información y ponerla en manos del usuario.

También provee una consistencia, una sola versión de todos los recursos de información, que es uno de los más grandes desafíos para las organizaciones hoy en día. (11)

Pentaho Data Integration permite una poderosa extracción, transformación y carga utilizando la herramienta Kettle.

1.8.2.1.1 Kettle

El uso de la herramienta Kettle permite evitar grandes cargas de trabajo manual frecuentemente difícil de mantener y de desplegar.

Kettle, es una herramienta que permite definir transformaciones de forma gráfica, interconectando bloques que tienen diversas funciones. Es tremendamente versátil, ya que se tienen bloques que permiten leer y escribir de cualquier base de datos, fichero Excel o CVS, Access, etc. y otros que permiten operar con los campos renombrando, normalizando, calculando campos en función de otros, mapeando valores, realizando búsquedas auxiliares en bases de datos, normalizando/desnomalizando los datos de distintas filas en una sola, etc. Las transformaciones que se hacen con el Kettle se guardan en un fichero ***ktr*** que luego puede ser ejecutado desde línea de comandos o un fichero batch. Además de ser open source y sin costes de licencia.

Ventajas de Kettle

- Funciona en Windows, Unix y Linux.
- Tiene una interfaz grafica con indicadores de las trasformaciones.

- Es una aplicación escrita en Java con algunas características avanzadas en JavaScript.
- Ofrece una licencia púbica GPL.
- Basado en metadatos.
- Como soporte se encuentra los foros de Pentaho y la comunidad Pentaho.
- Soporta Oracle, DB2.SQL Server; Sybase así como MySQL, Postgres; HYpersonic.También soporta la conectividad con SAP.
- Con respecto a las escalabilidad, soporta la arquitectura de procesamiento en paralelo para distribuir las tares de ETL a través de múltiples servidores.

Basado en dos tipos de objetos: Transformaciones (colección de pasos en un proceso ETL) y trabajos (colección de transformaciones). (12)

1.8.2.2 Talend Open.

Talend Open tiene una visión más amplia de integración de datos y cree que todas las fuentes de datos deben interpretar.

Ofrece capacidades avanzadas que mejoran realmente la productividad de la integración de los trabajos y de la escalabilidad para proporcionar una ejecución óptima. (13)

La aplicación Business Modeler de Talend Open Studio aprovecha una propuesta descendente (de lo general a lo particular) que permite a los grupos de interés de las líneas de negocio involucrarse en los procesos de integración.

El Business Modeler ofrece una visualización no técnica del business workflow muy fácil. Incluyen sistemas y procesos en la organización hoy pero también para el futuro. Sistemas, conexiones, necesidades son creados utilizando un workflow estándar a través de una toolbox gráfica y intuitiva. (14)

1.8.2.3 Octopus

Es una herramienta de ETL basada en Java y que se puede conectar a cualquier fuente JDBC y realizar la transformación que se encuentra definida en un archivo XML. Esto permite aumentar la interoperatividad simultánea con diferentes bases de datos entre las que se encuentran: MSSQL, Oracle, DB2, QED, JDBC-ODBC con Excel y Access, MySQL, CSV-files, XML-files. (15)

1.9 Base de Datos

Desde el punto de vista de BI de la base de datos de trabajo cotidiano se puede extraer conocimiento. El uso de las bases de datos transaccionales sirve para varios cometidos: primero se mantiene el trabajo transaccional diario de los sistemas de información originales (conocido como OLTP, On Line Transaccional Processing) y en segundo lugar se hace análisis de los datos en tiempo real sobre la misma base de datos (conocido como OLAP, On-Line Analytical Processing). (12)

PostgreSQL

PostgreSQL está teniendo gran aceptación por la adopción del lenguaje SLQ, aumentando su compatibilidad con otros productos comerciales y manteniendo la robustez y consistencia propia del Postgres. Otra mejora importante que posee es la existencia de los ejecutables para Windows.

Características

Las principales características de este gestor de bases de datos son las siguientes:

- Disparadores (*triggers*): Un *disparador* o *triggers* se define en una acción específica basada en algo ocurrente dentro de la base de datos.
- Vistas.
- Integridad transaccional.
- Herencia de tablas.
- Tipos de datos y operaciones geométricas.

Lo mejor de PostgreSQL

Las características positivas que posee este gestor son:

1. Posee una gran ***escalabilidad***. Es capaz de ajustarse al número de CPU y a la cantidad de memoria que posee el sistema de forma óptima, haciéndole capaz de soportar una mayor cantidad de peticiones simultáneas de manera correcta (en algunos benchmarks se dice que ha llegado a soportar el triple de carga de lo que soporta MySQL).
2. Implementa el uso de ***rollback's, subconsultas y transacciones***, haciendo su funcionamiento mucho más eficaz, y ofreciendo soluciones en campos.

3. Tiene la capacidad de comprobar la ***integridad referencial***, así como también la de almacenar procedimientos en la propia base de datos, equiparándolo con los gestores de bases de datos de alto nivel, como puede ser Oracle. (16)

Funciones

Son bloques de código que se ejecutan en el servidor. Pueden ser escritos en varios lenguajes, con la potencia que cada uno de ellos da, desde las operaciones básicas de programación, tales como bifurcaciones y bucles, hasta las complejidades de la programación orientada a objetos o la programación funcional.

Los disparadores son funciones enlazadas a operaciones sobre los datos. Algunos de los lenguajes que se pueden usar son los siguientes:

- Un lenguaje propio llamado [PL/PgSQL [1] [2](similar al PL/SQL de Oracle).
- C.
- C++.
- Gambas.
- Java PL/Java web.
- PL/Perl.
- plPHP.
- PL/Python.
- PL/Ruby.
- PL/sh.
- PL/Tcl.
- PL/Scheme.

PostgresSQL soporta funciones que retornan "filas", donde la salida puede tratarse como un conjunto de valores que pueden ser tratados igual a una fila retornada por una consulta.

Las funciones pueden ser definidas para ejecutarse con los derechos del usuario ejecutor o con los derechos de un usuario previamente definido. El concepto de funciones, en otros SGBD, son muchas veces referidas como "procedimientos almacenados" (stored procedures en inglés).

1.10 PgAdmin III.

PgAdmin3 que sirve para un entorno de escritorio visual.

Esta aplicación se puede utilizar en Linux, FreeBSD, OpenSUSE, Solaris, Mac OSX y plataformas de Windows para gestionar PostgreSQL 7.3 y encima se ejecutan en cualquier plataforma, así como comerciales y las versiones derivadas de PostgreSQL como EnterpriseDB, Mammoth PostgreSQL, y Bizgres Greenplum base de datos. (17)

PgAdmin III está diseñado para responder a las necesidades de todos los usuarios, desde simples consultas SQL escrito a la elaboración de bases de datos complejos. La aplicación también incluye una sintaxis SQL editor, un servidor del lado del editor de código, una de SQL / lote / shell programación de agente de empleo, el apoyo a la I-Slony motor de replicación y mucho más. La conexión con el servidor puede hacerse utilizando TCP / IP o Unix Domain Sockets (en * nix plataformas),

PgAdmin III es desarrollado por una comunidad de PostgreSQL expertos de todo el mundo y está disponible en más de una docena de idiomas. Es Software Libre liberado bajo la BSD / Licencia Artística.

1.11 Metodologías de apoyo.

1.11.1 Data Warehouse Engeneering Process (DWEP).

Como existe una gran variedad de modelos utilizados en las fases de diseño de los DW, se desarrolló un método que proporciona guías de diseño para crear y transformar modelos del almacén datos, el Data Warehouse Engineering Process *(DWEP)*, propuesto en la tesis de Sergio Luján Mora.

Es un método orientado a objetos, independiente de cualquier implementación específica, ya sea relacional, multidimensional, orientado a objetos, etc. permite la representación de todas las etapas del diseño de un Data Warehouse, está basada en UML (Lenguaje Unificado de Modelado) (OMG 2003) y RUP (Proceso Unificado de Desarrollo de Software) . (18)

1.11.2 HEFESTO

HEFESTO es una metodología, 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 idea principal, es comprender cada paso que se realizará, para no caer en el tedio de tener que seguir un método al pie de la letra sin saber exactamente qué se está haciendo, ni por qué.

La construcción e implementación de un DW puede adaptarse muy bien a cualquier ciclo de vida de desarrollo de software, con la salvedad de que para algunas fases en particular, las acciones que se han de realizar serán muy diferentes. Lo que se debe tener muy en cuenta, es no entrar en la utilización de metodologías que requieran fases extensas de reunión de requerimientos y análisis, fases de desarrollo monolítico que conlleve demasiado tiempo y fases de despliegue muy largas. Lo que se busca, es entregar una primera implementación que satisfaga una parte de las necesidades, para demostrar las ventajas del DW y motivar a los usuarios.

> La metodología HEFESTO, puede ser embebida en cualquier ciclo de vida que cumpla con la condición antes declarada.

Características:

Esta metodología cuenta con las siguientes características:

- Los objetivos y resultados esperados en cada fase se distinguen fácilmente y son sencillos de comprender.
- Se basa en los requerimientos del usuario, por lo cual su estructura es capaz de adaptarse con facilidad y rapidez ante los cambios en el negocio.
- Reduce la resistencia al cambio, ya que involucra al usuario final en cada etapa para que tome decisiones respecto al comportamiento y funciones del DW.
- Utiliza modelos conceptuales y lógicos, los cuales son sencillos de interpretar y analizar.
- Es independiente del tipo de ciclo de vida que se emplee para contener la metodología.
- Es independiente de las herramientas que se utilicen para su implementación.
- Es independiente de las estructuras físicas que contengan el DW y de su respectiva distribución.

- Cuando se culmina con una fase, los resultados obtenidos se convierten en el punto de partida para llevar a cabo el paso siguiente.
- Se aplica tanto para DM como para DW (6).

Descripción

La metodología HEFESTO puede resumirse a través de los siguientes pasos:

1. **Análisis de Requerimientos**.

a. Identificar preguntas

b. Identificar indicadores y perspectivas de análisis.

c. Modelo Conceptual.

2. **Análisis de los OLTP**

a. Establecer correspondencia con los requerimientos.
b. Seleccionar los campos que integrarán cada perspectiva. Nivel de granularidad.

3. **Elaboración del Modelo Lógico de la Estructura del DW.**

a. Diseñar tablas de dimensiones.
b. Diseñar tablas de hechos.
c. Realizar uniones.
d. Determinar jerarquías.

4. **Proceso ETL, Limpieza de Datos y Sentencias SQL**.

Después de estudiar la metodología HEFESTO y analizar todas las características y ventajas que brinda se decidió utilizar esta metodología para traducir el esquema conceptual de un DW a un esquema lógico relacional

1.12 Principales tipos de paralelismos que se pueden implementar en las aplicaciones ETL.

Actualmente muchos desarrolladores de herramientas ETL están extendiendo sus productos para agrupar nuevos requisitos de usuarios. Lo que ha dado como resultado

una nueva generación de productos ETL llamados Plataformas de Integración de Datos. Estos productos extienden las herramientas ETL con una variedad de nuevas capacidades, incluyendo limpieza de datos, perfilando los datos, captura avanzada de datos, actualizaciones incrementales y un host de nuevas fuentes y destinos.

Un desarrollo reciente en el software ETL es la aplicación de procesamiento paralelo. Esto ha permitido desarrollar una serie de métodos para mejorar el rendimiento general de los procesos ETL cuando se trata de grandes volúmenes:

- De datos**:** Consiste en dividir un único archivo secuencial en pequeños archivos de datos para proporcionar acceso paralelo.
- De segmentación (pipeline): Permitir el funcionamiento simultáneo de varios componentes en el mismo flujo de datos. Un ejemplo de ello sería buscar un valor en el registro número 1 a la vez que se suman dos campos en el registro número 2.
- De componente: Consiste en el funcionamiento simultáneo de múltiples procesos en diferentes flujos de datos en el mismo puesto de trabajo.

Estos tres tipos de paralelismo no son excluyentes, sino que pueden ser combinados para realizar una misma operación ETL. (7)

Conclusiones

En este capítulo se ha demostrado la necesidad del proceso extracción, transformación y carga de datos para un Data Warehouse que facilite el intercambio, manejo, procesamiento y almacenamiento de información de una forma más óptima y fácil de manejar para el Sistema de Gestión de Información de los Laboratorios (LIMS). Se ha seleccionado la metodología HEFESTO y dentro del Pentaho Business Intelligence Open Suite la herramienta Spoon, Kettle.

CAPÍTULO 2: Análisis de los OLTP

En este capítulo y teniendo en cuenta que ya se han detallado claramente las características generales del proceso extracción, transformación y carga de datos, se definirán y describirán todos los componentes que intervienen en su arquitectura o ambiente. Se hará una selección y argumentación de los requisitos. Además de analizar los OLTP para luego diseñar e implementa los procesos ETL.

El proceso de toma de decisiones en toda empresa, independientemente de la envergadura no es tarea fácil, porque cualquier cambio es riesgoso, si bien deben aceptarlos cuando se establecen, también deben saber que esos cambios pueden ser muy beneficiosos. Para ello se maneja un gran volumen de datos que interpretados con inteligencia, se convierten en una información muy valiosa, hoy en día existen herramientas automáticas muy poderosas, con un alto grado de confiabilidad para tomar la decisión acertada. Y de eso se tratan los negocios, saber tomar la decisión correcta en el momento oportuno. Aquí es donde aparece el término, Business Intelligence (BI) o Inteligencia de Negocio pues por medio de dicha información se pueden generar escenarios, pronósticos y reportes que apoyen a la toma de decisiones, lo que se traduce en una ventaja competitiva. (4)

Los sistemas BI son una categoría de software que ya tiene un grado de madurez en el mercado, brindan una amplia gama de capacidades funcionales. En general este tipo de soluciones procesa gran cantidad de datos para generar información relevante y disminuir la incertidumbre en la toma de decisiones de negocio.

Tradicionalmente, el tener que elegir una plataforma y una solución de BI es un proceso que consume mucho tiempo y dinero. Pero las soluciones de BI de código abierto reducen notablemente el costo.

2.1 Spoon, Kettle o Pentaho Data Integration.

Para darle solución a lo antes mencionado y por todas las ventajas que brida se decidió utilizar la siguiente herramienta: Spoon, Kettle o Pentaho Data Integration.

Pentaho es una plataforma que brinda distintas soluciones a nivel de inteligencia de negocios. Una de las herramientas de **Pentaho** es **Kettle** que se va a utilizar para el proceso de ETL.

Para hacer un proceso de ETL en Kettle, hay disponibles transformaciones y trabajos. Una transformación es el proceso de limpieza en sí, y un trabajo define la secuencia en la que se desea ejecutar las trasformación y los trabajos.

Una vez que una transformación es creada, se interpreta, no es necesario compilarla antes. Kettle permite verificar que la transformación se haya definido correctamente.

Dispone de distintos *stages* para cada acción que se quiera realizar.

Incluye cuatro herramientas:

- **Spoon:** para diseñar transformaciones ETL usando el entorno gráfico (un componente de Pentaho Data Integration) para crear unas transformaciones simples.
- **PAN:** forma parte de Pentaho Data Integration (Kettle ETL) y es una herramienta que permite ejecutar las transformaciones Spoon desde la línea de comando. Las transformaciones spoon pueden ser ejecutadas como los **ficheros XML** (con la extensión **ktr** – que viene de Kettle transformación) ó también desde el **repositorio** guardado en la base de datos.
- **CHEF:** para crear trabajos
- **Kitchen:** para ejecutar trabajos (11)

2.1.1 Repositorio de Kettle.

Un repositorio de Kettle (*Kettle Repository*) es un conjunto de tablas de base de datos que son accesibles por los clientes de Kettle (*Spoon*, *Chef*, *Pan* y *Kitchen*) para almacenar y recibir transformaciones, trabajos, conexiones de base de datos, etc.

Un poco más en detalle, un repositorio de Kettle es una base de datos que contiene las tablas necesarias para poder almacenar todos los "*objetos*" que se crean, transformaciones, trabajos, conexiones estarán disponibles de forma centralizada, facilitando la reutilización y gestión de cambios.

El repositorio es independiente de la base de datos, Kettle sólo necesita que existan las tablas definidas en el esquema relacional. Kettle no soporta oficialmente ningún otro método de acceso para los repositorios que no sea *Native (JDBC)* y las pruebas que se han realizado demuestran que al menos con ODBC falla. No obstante, se puede acceder a casi cualquier base de datos con JDBC desde Kettle.

Para ver como se crear un repositorio ir al **Anexo A.**

2.1.1.1 Explorador de repositorio en Kettle.

El "*Repository Explorer*" o Explorador de repositorio es la herramienta dentro de Spoon que permite ver y modificar los objetos almacenados en el repositorio al que se está conectado.

De forma sencilla en una vista de árbol se puede ver los diferentes objetos que es posible almacenar en un repositorio (Conexiones, esquemas de partición, Slave Servers, Clusteres, transformaciones, trabajos, usuarios y perfiles).

Para ver el explorador de este trabajo ir al **Anexo B**.

En resumen: **Kettle hace fácil la construcción, actualización y mantenimiento de almacenes de datos.**

Para seguir con la línea arquitectónica del proyecto Lims Control de Calidad se utilizó:

- **PostgreSQL** como SGBD además de que dispone de todas las características de una base de datos de nivel empresarial, como es requerido para un sistema BI.
- **PgAdmin III** que sirve para un entorno de escritorio visual, es el más popular código abierto de administración y plataforma de desarrollo para PostgreSQL, la más avanzada base de datos de Open Source en el mundo.

2.2 Empresa analizada: Centro de Ingeniería Genética y Biotecnología de La Habana (CIGB)

Se realizaron análisis a los usuarios en busca de sus necesidades de información, pero las mismas abarcaban casi todas las actividades de la empresa, por lo cual se les pidió que escogieran el proceso que considerasen más importante en las actividades

diarias de la misma y que estuviese soportado de alguna manera por algún OLTP Procesamiento de Transacciones En Línea (OnLine Transaction Processing). (19) El proceso elegido fue Análisis Químico.

Antes de comenzar con el primer paso, es importante describir las características principales del grupo de análisis químico al cual se le aplicará la metodología HEFESTO, así se podrá tener como base un ámbito predefinido y se comprenderá mejor cada decisión que se tome con respecto a la extracción, transformación y carga de los datos.

Además, este análisis ayudará a conocer el funcionamiento, lo que permitirá examinar e interpretar de forma óptima las necesidades de información del mismo, como así también apoyará a una mejor construcción y adaptación del depósito de datos.

El **Grupo de Análisis Químico** realiza diferentes técnicas físico-químicas y bioquímicas para la determinación de impurezas y purezas de las diferentes proteínas que se producen o investigan en el Centro. También se lleva el control analítico de los reactivos y componentes críticos que son utilizados como materia prima en la producción de los productos farmacéuticos en el CIGB. Todas estas técnicas se realizan bajo un estricto cumplimiento de las Buenas Prácticas de Laboratorio garantizando así resultados confiables, lo cual ha sido demostrado en cada inspección realizada por inspectores internos, por la entidad nacional regulatoria y por organizaciones internacionales del nivel de la Organización Mundial de la Salud (OMS).En el grupo se realizan alrededor de 25 técnicas analíticas diferentes, las cuales se encuentran validadas.

El Sistema de Gestión de Información de los Laboratorios (LIMS) se encarga de la automatización de los principales Departamentos del Centro de Ingeniería Genética y Biotecnología (CIGB) lo que genera el almacenamiento de grandes volúmenes de información dificultando el trabajo del personal a la hora de realizar alguna consulta o buscar alguna información de utilidad.

2.3 Requisitos

Frederick P. Brooks, dice "La parte más difícil de construir un sistema es precisamente saber qué construir. Ninguna otra parte del trabajo conceptual es tan difícil como establecer los requisitos técnicos detallados, incluyendo todas las interfaces con gente, máquinas y otros sistemas. Ninguna otra parte del trabajo afecta tanto el sistema si es hecha mal. Ninguna es tan difícil de corregir más adelante. Entonces, la

tarea más importante que el ingeniero de software hace para el cliente es la extracción iterativa y el refinamiento de los requisitos del producto." (20)

El análisis de los requerimientos de los diferentes usuarios, es el punto de partida de esta metodología, ya que ellos son los que deben, en cierto modo, guiar la investigación hacia un desarrollo que refleje claramente lo que se espera del depósito de datos, en relación a sus funciones y cualidades.

El objetivo principal de esta fase, es obtener e identificar las necesidades de información clave de alto nivel, que es esencial para llevar a cabo las metas y estrategias de la empresa, y que facilitará una eficaz toma de decisiones.

Debe tenerse en cuenta que dicha información, es la que proveerá el soporte para desarrollar los pasos sucesivos, por lo cual, es muy importante que se preste especial atención al relevar los datos.

Una forma de asegurarse de que se ha realizado un buen análisis, es que el resultado del mismo debe hacer explícitos los objetivos estratégicos planteados por la empresa que se está estudiando.

A continuación, se procedió a identificar que era lo que les interesaba conocer acerca de este proceso y cuáles eran las variables o perspectivas que debían tenerse en cuenta para poder tomar decisiones. (21)

La institución desea resolver los siguientes requerimientos:

1. Se desea conocer la cantidad de lotes que fueron recibidos por el centro en un período de tiempo.
2. Se desea conocer la cantidad de ensayos realizados por lote en un período de tiempo.
3. Se desea conocer la cantidad de ensayos repetidos por lote y en un período de tiempo.
4. Se desea conocer las principales causas de repetición de ensayos en un período determinado de tiempo.
5. Se desea conocer dado un período de tiempo y lote, el promedio de ensayos conformes (satisfactorios).
6. Se desea conocer dado un período de tiempo y lote, promedio de ensayos no conformes.

7. Se desea conocer la cantidad de técnicas realizadas por lote en un período de tiempo.
8. Se desea una lista con todas las técnicas utilizadas por un lote determinado en un determinado período de tiempo (con el número de veces incluido).
9. Se desea conocer la cantidad de ensayos conformes por técnica en un período de tiempo.
10. Se desea conocer la cantidad de ensayos no conformes por técnica en un período de tiempo.
11. Se desea conocer mes más satisfactorio en cuanto a ensayos conformes.
12. Se desea conocer año más satisfactorio en cuanto a ensayos conformes.

Identificar indicadores y perspectivas de análisis

1. **"**Unidades recibidas de un producto determinado en un tiempo determinado"

Indicador **Perspectivas**

2. **"**Unidades recibidas de un producto determinado dado un lote en un tiempo determinado"

Indicador **Perspectivas**

3. **"**Unidades recibidas de un producto determinado de un origen dado en un tiempo determinado"

Indicador **Perspectivas**

4. **"**Cantidad recibida de lotes en un tiempo determinado"

Indicador **Perspectivas**

5. **"**Cantidad realizada de ensayos por lote en un tiempo determinado"

Indicador **Perspectivas**

6. **“**Cantidad repetida de ensayos por lote en un tiempo determinado”

Indicador **Perspectivas**

7. **“**Cantidad de causas de repetición de ensayos en un tiempo determinado”

8. **“**Promedio de ensayos conformes dado un lote y un tiempo determinado”

Indicador **Perspectivas**

9. **“**Promedio de ensayos no conformes dado un lote y un tiempo determinado”

Indicador **Perspectivas**

10. **“**Cantidad de técnicas realizadas por lote en un tiempo determinado”

Indicador **Perspectivas**

11. **“**Cantidad de técnicas utilizadas por lote en un tiempo determinado”

Indicador **Perspectivas**

12. **“**Cantidad de ensayos conformes por técnica en un tiempo determinado”

Indicador **Perspectivas**

13. **“**Cantidad de ensayos no conformes por técnica en un tiempo determinado”

Indicador **Perspectivas**

Los indicadores son:

- Unidades recibidas
- Cantidad recibida
- Cantidad realizada de ensayos

- Cantidad repetida de ensayos
- Promedio de ensayos conformes
- Promedio de ensayos no conformes
- Cantidad de técnicas utilizadas
- Cantidad de técnicas realizadas
- Cantidad de ensayos conformes
- Cantidad de ensayos no conformes

Las perspectivas de análisis son:

- Producto
- Lote
- Origen
- Tiempo
- Ensayos
- Técnicas

2.4 Análisis de los OLTP

Un punto importante que debe tenerse muy en cuenta es que la información, debe estar soportada de alguna manera por algún OLTP (On Line Transaction Processing), representa toda aquella información transaccional que genera la empresa en su accionar diario, además, de las fuentes externas con las que puede llegar a disponer.
Entre los OLTP más habituales que pueden existir en cualquier organización se encuentran:

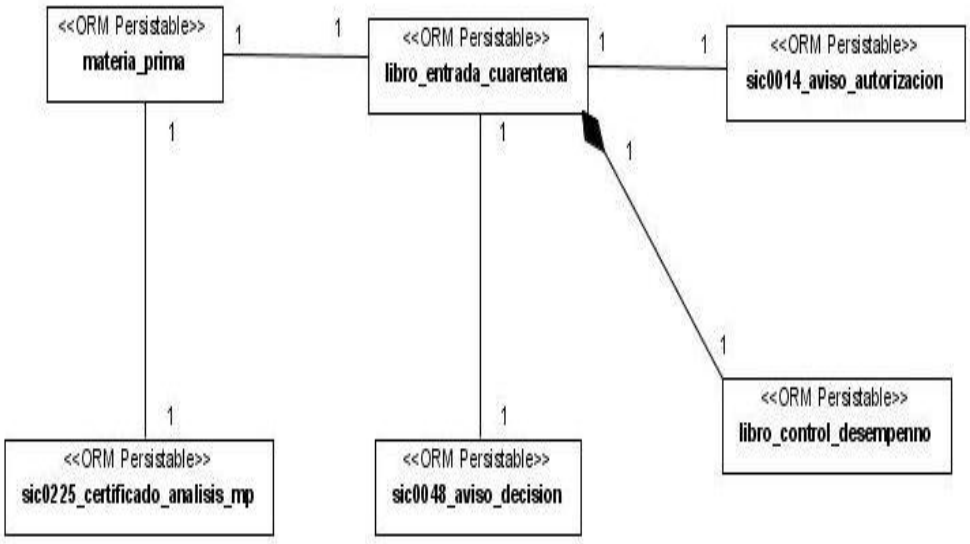
- Archivos de textos.
- Hipertextos.
- Hojas de cálculos.
- Informes semanales, mensuales, anuales, etc.
- Bases de datos transaccionales.

Una vez establecidos en una organización unos sistemas, OLTP, se necesitara poder valorar y analizar el estado de cada proceso. A la hora de realizar estas evaluaciones aparece la primera dificultad: La cantidad de datos generados por las empresas.

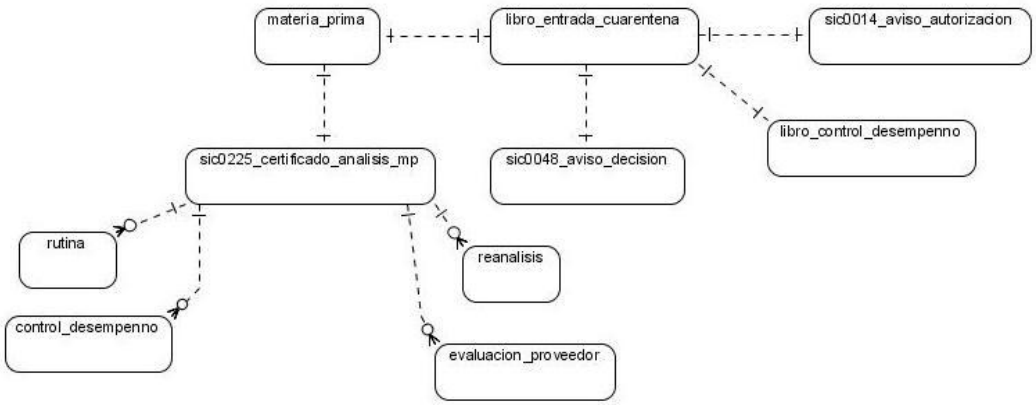
Por pequeña que sea la empresa, a través de los OLTP, los datos almacenados llegan a tomar dimensiones inmanejables por los sistemas de análisis convencionales, tales

como la auditoría contable, o los resultados de estos análisis no son rápidos ni precisos, al requerir un tiempo de procesado que impide la decisión y la identificación de oportunidades de negocio.

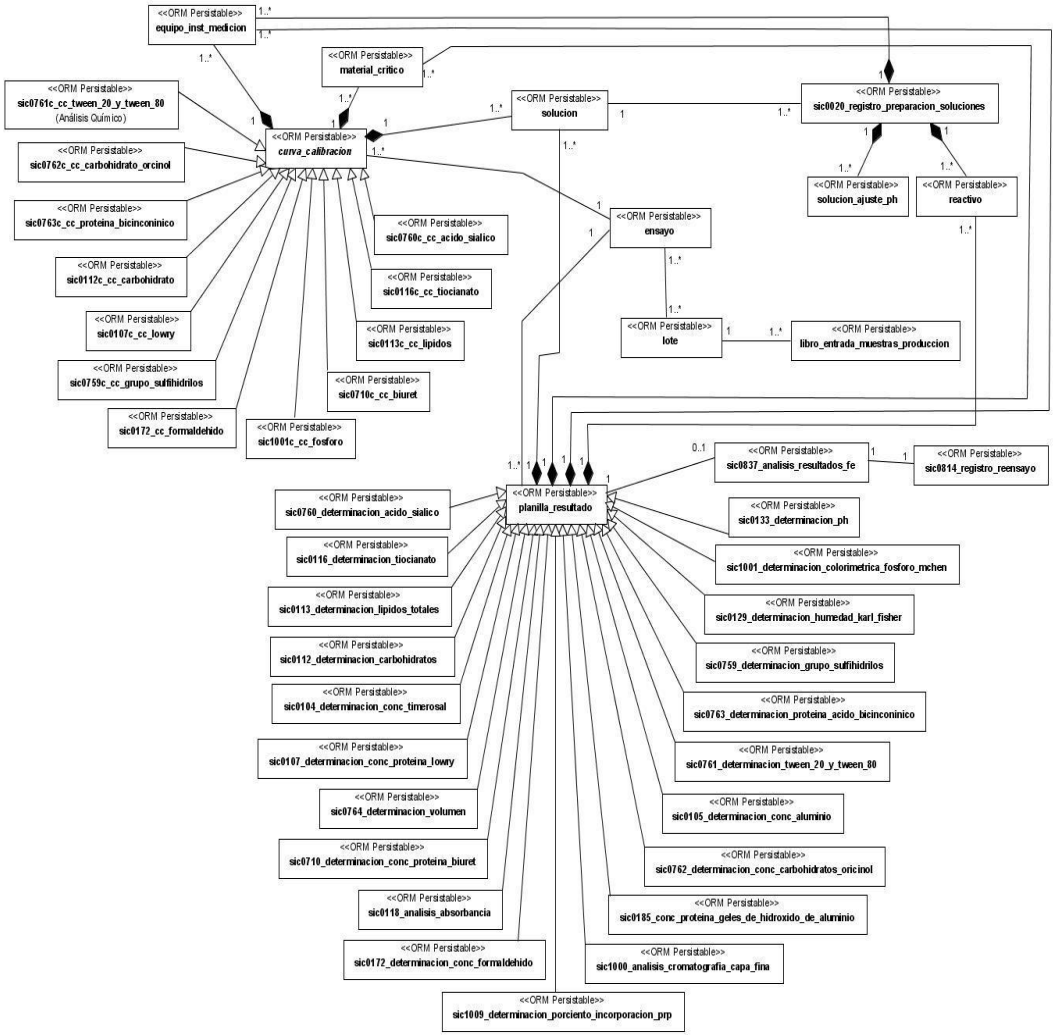
2.4.1 Diagrama de Clases Persistentes.



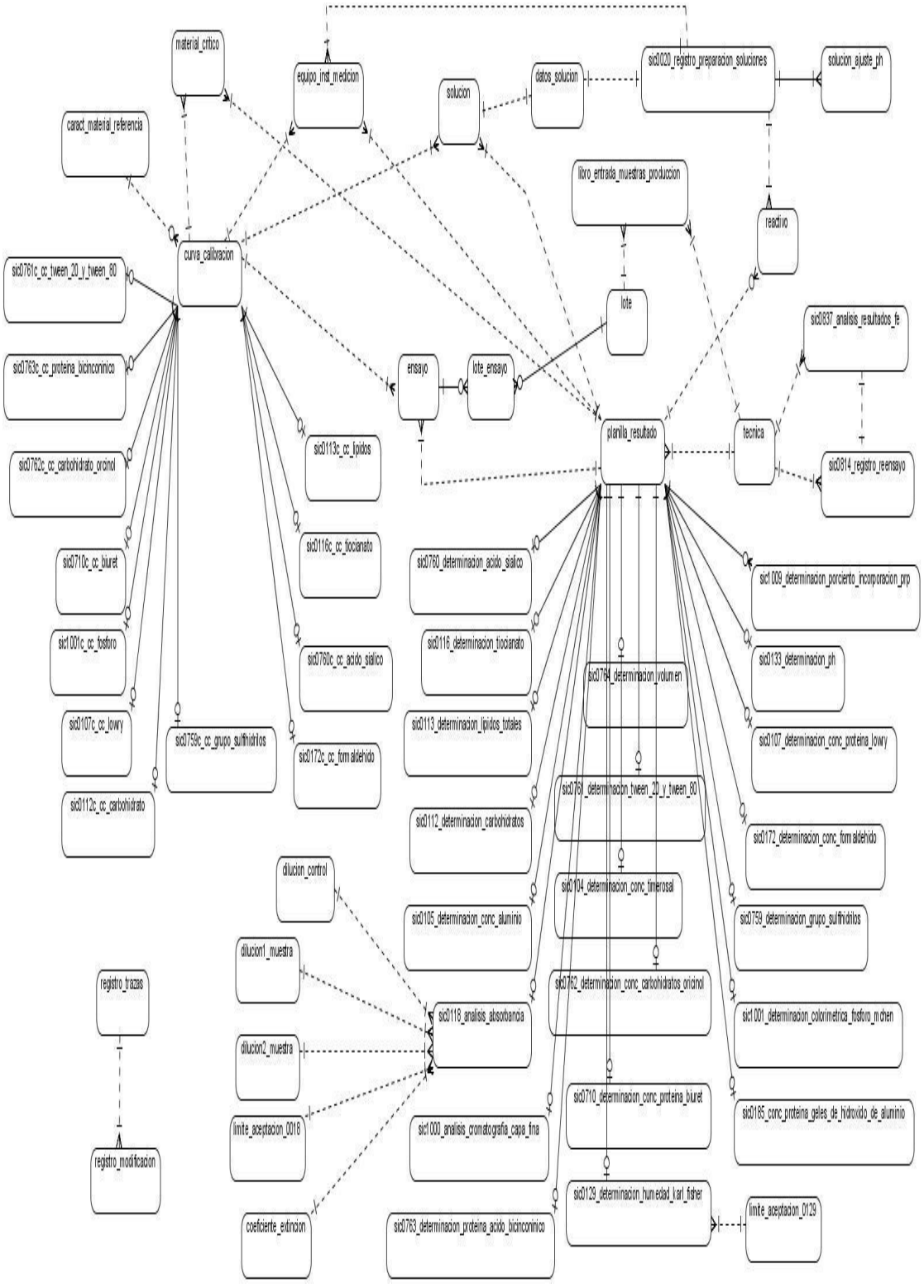
Figuras 2: Diagrama de Clases Persistentes.



Figuras 3: Modelo Físico del Proceso de Determinar el Estado de las Materias Primas.



Figuras 4: Diagrama de Clases Persistentes que agrupa las Clases del Proceso de Determinación de Purezas e Impurezas de las Proteínas.



Figuras 5: Modelo Físico del Proceso de Determinación del Grado de Purezas e impurezas de las Proteínas.

Descripción de las tablas

A continuación se describen las tablas más significativas la Base de Datos del módulo Análisis Químico de la cual se van a extraer los datos y hacer las trasformaciones necesarias para cargarlas en el DHW. (19)

Tabla 1: Material Crítico.

Nombre: material critico		
Descripción: Tabla que recoge los datos de los materiales críticos utilizados en las Curvas de Calibración		
Atributo	**Tipo**	**Descripción**
id_material	serial	Valor entero de autoincremento, llave primaria
id_curva	integer	Llave primaria foránea que toma de la entidad curva_calibracion
id_planilla_resultado	integer	Llave primaria foránea que toma de la entidad planilla_resultado
nombre_material	varchar(50)	Nombre del material crítico utilizado en la Curva de Calibración
no_parte	integer	Número único que define características del material
cant_requerida	integer	Cantidad de material requerida
ppo_referencia	varchar (11)	Código del PPO de referencia

Tabla 2: Equipo_inst_medicion.

Nombre: equipo_inst_medicion		
Descripción: Tabla que recoge los datos de los equipos en instrumentos de medición utilizados en las Curvas de Calibración		
Atributo	**Tipo**	**Descripción**
id_equipo_inst_medicion	serial	Valor entero de autoincremento, llave primaria
id_curva	integer	Llave primaria foránea que toma de la

		entidad curva_calibracion
id_planilla_resultado	integer	Llave primaria foránea que toma de la entidad planilla_resultado
no_folio_sic0020	integer	Llave primaria foránea que toma de la entidad sic0020_registro_preparacion_soluciones
equipamiento	varchar(50)	Nombre del equipo o instrumento
no_identificacion	varchar(15)	Número único de identificación del equipo o instrumento
fecha_venc_equipos	Date	fecha de vencimiento de la calibración

Tabla 3: Solución.

Nombre: solución		
Descripción: Tabla que recoge los datos de las soluciones		
Atributo	**Tipo**	**Descripción**
id_solucion	serial	Valor entero de autoincremento, llave primaria
id_curva	integer	Llave primaria foránea que toma de la entidad curva_calibracion
id_planilla_resultado	integer	Llave primaria foránea que toma de la entidad planilla_resultado
Id_informacion	integer	Llave primaria foránea que toma de la entidad datos_solucion
nombre_solucion	varchar(25)	Nombre de la solución utilizada
no_parte	integer	Número único que define características de la solución
no_lote	varchar(15)	Lote de la solución
fecha_venc_solucion	Date	fecha de vencimiento de la solución

Tabla 4: Curva_Calibración.

Nombre: curva_calibracion

Descripción: Tabla que recoge los datos de las Curvas de Calibración aplicadas a las soluciones.		
Atributo	**Tipo**	**Descripción**
id_curva	serial	Valor entero de autoincremento, llave primaria
id_caracteristica_mr	integer	Llave foránea que toma de la entidad caract_material_referencia
otros_mr	varchar(25)	Otras características, puede ser null
nombre_mr	varchar(25)	Nombre del Material de Referencia a utilizar
lote_mr	varchar(15)	Lote del Material de Referencia a emplear
repetir_ensayo	bool	Puede ser Sí o No
causa_repeticion	varchar(255)	Descripción de la causa de la repetición, puede ser null
desviacion	bool	Puede ser Sí o No
observaciones	varchar(255)	Puede ser null
pasa_prueba	bool	Puede ser Sí o No
fecha _pasar_prueba	Date	fecha que pasó la prueba
rango_aceptacion_pendiente_max	float	Rango de aceptación máximo de la pendiente
rango_aceptacion_pendiente_min	float	Rango de aceptación mínimo de la pendiente
rango_aceptacion_intercepto_max	float	Rango de aceptación máximo del Intercepto
rango_aceptacion_intercepto_min	float	Rango de aceptación mínimo del Intercepto
cumple_pendiente	bool	Puede ser Sí o No
cumple_intercepto	bool	Puede ser Sí o No

realizado_por	varchar(25)	Nombre de la persona que realizó la Curva
revisado_por	varchar(25)	Nombre de la persona que revisó la Curva
fecha _realizacion	Date	fecha de realización de la Curva
terminado	bool	Puede ser Sí o No

La clase *curva_calibracion* descrita anteriormente cuenta con 12 especializaciones, las cuales se presentan a continuación, *sic0763c_cc_proteina_bicinconinico, sic0761c_cc_tween_20_y_tween_80, Sic0760c_cc_acido_siálico, sic0112c_cc_carbohidrato, sic0762c_cc_carbohidrato_orcinol, Sic0113c_cc_lipidos, sic0759c_cc_grupo_sulfihidrilos, sic0710c_cc_biuret* y *sic0116c_cc_tiocianato; sic0172_cc_formaldehido, sic1001c_cc_fosforo, y sic0107c_cc_lowry.*

Tabla 5: Planilla_Resultado.

Nombre: planilla_resultado		
Descripción: Tabla que recoge los datos de las Planillas de Resultados		
Atributo	**Tipo**	**Descripción**
id_planilla	serial	Valor entero de autoincremento, llave primaria
id_tecnica	integer	Llave foránea que toma de la entidad técnica
muestra_de	varchar(25)	Nombre del Producto
no_lote	varchar(10)	Lote de la muestra
no_entrada_lab	integer	Número de entrada al laboratorio
fecha _recepcion	Date	fecha de recepción de la muestra en el laboratorio
ensayo_inicial	boolean	Puede ser Sí o No

Repeticion	boolean	Puede ser Sí o No
no_valido	boolean	Puede ser Sí o No
no_cumple_m	boolean	Puede ser Sí o No
repetir_ensayo	boolean	Puede ser Sí o No
causa_repeticion	varchar(255)	Causa de repetición del ensayo, puede ser null
Desviaciones	boolean	Puede ser Sí o No
lista_desviaciones	varchar(255)	Lista de desviaciones, puede ser null
Observaciones	varchar(255)	Observaciones del resultado, puede ser null
realizado_por	varchar(25)	Nombre de la persona que realizó la determinación
revisado_por	varchar(25)	Nombre de la persona que revisó la determinación
recibido_por	varchar(25)	Nombre de la persona que recibió la determinación
fecha _realizacion	Date	fecha de realización de la determinación
Terminado	boolean	Puede ser Sí o No

La clase *planilla_resultado* descrita anteriormente cuenta con 21 especializaciones, de las cuales 17 se presentan a continuación, *sic0112_detrminacion_carbohidratos, sic0113_determinacion_lipidos_totales, sic0185_conc_proteina_geles_de_hidroxido_de_aluminio, sic0760_determinacion_acido_sialico, Sic0763_determinacion_proteina_acido_bicinconinico, sic1001_determinacion_colorimetrica_fosforo_chen, sic0759_determinacion_grupo_sulfihidrilos, sic0762_determinacion_conc_carbohidratos_oricinol, Sic0105_determinacion_conc_aluminio, sic0710_determinacion_conc_proteina_biuret, Sic1000_analisis_cromatografia_capa_fina, sic0104_determinacion_conc_timerosal,*

Sic0764_determinacion_volumen, sic0133_determinacion_ph, sic0172_determinacion_conc_formaldehido, sic0118_analisis_absorbancia .Las cinco restantes se pueden observar en el Anexo 5, ellas son *sic0116_determinacion_tiocianato, sic1009_determinacion_porciento_incorporacion_prp, Sic0761_determinacion_tween_20_y_tween_80, sic0129_determinacion_humedad_karl_fisher,* y *sic0107_determinacion_conc_proteina_lowry.*

Tabla 6: Lote.

Nombre: lote		
Descripción: Tabla que recoge los datos de los lotes		
Atributo	**Tipo**	**Descripción**
Lote	varchar(15)	Lote del producto, llave primaria
nombre_producto	varchar(25)	Nombre del producto
nombre_etapa	varchar(25)	Nombre de la etapa
origen_producto	varchar(25)	Origen del producto
lote_origen	varchar(15)	Lote de Origen del producto
fecha_fab_lote	Date	fecha de fabricación del lote

Tabla 7: Caract_Material_Referencia.

Nombre: caract_material_referencia		
Descripción: Tabla que recoge los datos de las características de los materiales de referencias utilizados en las Curvas de Calibración		
Atributo	**Tipo**	**Descripción**
id_caracteristica	serial	Valor entero de autoincremento, llave primaria
nombre_material	varchar(25)	Nombre del material de referencia

Tabla 8: Registro_Preparación_Soluciones.

Nombre: registro_preparacion_soluciones

Descripción: Tabla que recoge los datos del Registro de Preparación de Soluciones		
Atributo	**Tipo**	**Descripción**
no_folio	varchar(15)	Número de Folio del registro, llave primaria
id_solucion	integer	Llave foránea que toma de la entidad solución
no_lote	varchar(15)	Lote de la solución
vt	float	Volumen total de la solución
fecha _prep	Date	fecha de preparación de la solución
nombre_equipo	varchar(25)	Nombre de los equipos utilizados
codigo_equipo	varchar(10)	Código de los equipos utilizados
fecha _venc_calib	Date	fecha de vencimiento de la calibración
solvente	varchar(25)	Nombre del solvente en que se disuelve el reactivo
otro_solvente	varchar(25)	Otros solventes en caso de que haya
ph_solvente	integer	PH del solvente
conductividad_solvente	float	Conductividad del solvente
silice_solvente	float	Contenido de sílice del solvente
ph_final_solucion	integer	PH final de la solución
conductividad_solucion	float	Conductividad de la solución
esterilizacion	boolean	Puede ser Sí o No
filtracion	varchar(25)	Nombre de la filtración realizada
otras_f	varchar(25)	Otras filtraciones
vapor_saturado	float	Vapor saturado, puede ser null
tiempo_e	integer	Tiempo de esterilización, puede ser null
no_frascos	integer	Cantidad de frascos
volumen_frasco	float	Volumen por frasco

temp_almacenamiento	float	Temperatura de almacenamiento
fecha _vencimiento_s	Date	fecha de vencimiento de la solución
pruebas_aceptacion	varchar(255)	Pruebas de aceptación de la solución, puede ser null
limite	float	Límite de aceptación de la solución, puede ser null
valor_obtenido	float	Valor obtenido en las pruebas, puede ser null
realizado_por	varchar(25)	Nombre de la persona que hizo el registro
revisado_por	varchar(25)	Nombre de la persona que revisó el registro
fecha _realizacion	Date	fecha de realización del registro
terminado	boolean	Puede ser Sí o No

Tabla 9: Libro_Entrada_Muestras_Producción.

Nombre: libro_entrada_muestras_produccion		
Descripción: Tabla que recoge los datos del Libro de Entrada de Muestras de Producción		
Atributo	**Tipo**	**Descripción**
id_libro	serial	Valor entero de autoincremento, llave primaria
Id_lote	integer	Llave foránea que la toma de la entidad lote, lote de la muestra
Id_tecnica	integer	Llave foránea que la toma de la entidad técnica
Anno	integer	Año de entrada de la muestra
no_entrada	integer	Número de entrada de la muestra, valor entero de autoincremento
fecha _entrada	Date	fecha de entrada de la muestra

Prod	varchar(25)	Nombre del Producto
Ident	varchar(25)	Identificación del tipo de muestra
nombre_entrega_m	varchar(25)	Nombre de quien entrega la muestra
nombre_recibe_m	varchar(25)	Nombre de quien recibe la muestra
Vol	float	Volumen de la muestra
cond_almac	varchar(10)	Condiciones de almacenamiento
nombre_realiza	varchar(25)	Nombre de quien entrega el resultado de la muestra
fecha _realiza	Date	fecha de la realización de la planilla
nombre_recibe	varchar(25)	Nombre de quien recibe el registro de la muestra
fecha _entrega_p	Date	fecha de entrega de la planilla
Observaciones	varchar(255)	Puede ser null

Tabla 10: Registro_Reensayo.

Nombre: registro_reensayo		
Descripción: Tabla que recoge los datos de Registros de Reensayos		
Atributo	**Tipo**	**Descripción**
folio_registro	varchar(15)	Número de Folio del registro, llave primaria
id_sic0837	integer	Llave foránea que la toma de la entidad sic0837_analisis_resultados_fe
id_tecnica	integer	Llave foránea que la toma de la entidad técnica

Fecha	Date	fecha del reensayo
folio_sic0837	varchar(15)	Número de Folio del SIC0837
Causa	varchar(255)	Causa del reensayo
Analista	varchar(25)	Nombre del analista del laboratorio que realizó en reensayo
Supervisor	varchar(25)	Nombre del supervisor del reensayo

Tabla 11: Técnica.

Nombre: técnica		
Descripción: Tabla que recoge los nombres de las técnicas que se aplican en el laboratorio con su PPO correspondiente		
Atributo	**Tipo**	**Descripción**
id_tecnica	serial	Valor entero de autoincremento, llave primaria
nombre_tecnica	varchar(50)	Nombre de la Técnica
ppo	varchar(25)	Número de PPO correspondiente

Tabla 12: Ensayo.

Nombre: ensayo		
Descripción: Tabla que recoge los datos de los ensayos		
Atributo	**Tipo**	**Descripción**
id_ensayo	serial	Valor entero de autoincremento, llave primaria
id_planilla_resultado	integer	Llave foránea que la toma de la entidad planilla_resultado
id_curva	integer	Llave foránea que la toma de la entidad curva_calibracion

Tabla 13: Datos_Solución.

Nombre: datos_solucion		
Descripción: Tabla que recoge los datos de las soluciones		
Atributo	**Tipo**	**Descripción**
id_sol	serial	Valor entero de autoincremento, llave primaria
nombre_solucion	varchar(255)	Nombre de la solución
np_solucion	integer	Número de Parte correspondiente a cada solución

Tabla 14: Libro_Control_Desempeño.

Nombre: libro_control_desempenno		
Descripción: Tabla que recoge los datos del Libro de Control de Desempeño		
Atributo	**Tipo**	**Descripción**
id_libroc	serial	Valor entero de autoincremento, llave primaria
id_libro_entrada_cuarentena	integer	Llave foránea que la toma de la entidad libro_entrada_cuarentena
Anno	integer	Año actual de producción
lote_cigb	varchar(15)	Número de Lote de la materia prima cuando llega al laboratorio
no_parte	integer	Número único que define características de la materia prima
Proveedor	varchar(25)	Nombre del proveedor de la materia prima
Frecuencia	integer	Número de veces que se analiza la materia prima

Tabla 15: Libro_Entrada_Cuarentena.

Nombre: libro_entrada_cuarentena		
Descripción: Tabla que recoge los datos del Libro de Entrada de Cuarentena		
Atributo	**Tipo**	**Descripción**
id_libro	serial	Valor entero de autoincremento, llave primaria
anno	integer	Año en que se realiza el registro
folio	varchar(15)	Número de folio de la Materia Prima
fecha _entrada	Date	fecha de entrada de la Materia Prima
materia_prima	varchar(25)	Nombre de la Materia Prima
lote_cigb	varchar(15)	Número de Lote con que la Materia Prima llega al laboratorio
proveedor	varchar(25)	Nombre del proveedor de la Materia Prima
lote_proveedor	varchar(15)	Número de Lote del proveedor de la Materia Prima
no_envase	integer	Cantidad de envases que tiene el lote
u_m	varchar(5)	Unidad de medida de los envases
fecha _venc	Date	fecha de vencimiento de la Materia Prima
recibe	varchar(25)	Nombre de quien recibe la Planilla
decision	varchar(25)	Enuncia que decisión se va a tomar
fecha _dec	Date	fecha de la decisión

Tabla 16: Registro_Trazas.

Nombre: registro_trazas		
Descripción: Tabla utilizada para mantener el registro de las trazas de modificación de los datos de las planillas		
Atributo	**Tipo**	**Descripción**
idtraza	serial	Valor entero de autoincremento, llave primaria

usuario	varchar(15)	Nombre del usuario que está haciendo la modificación
tabla_modificada	varchar(50)	Nombre del documento modificado
fecha_modificacion	timestamp	Fecha de modificación del documento.
id_tupla_modificada	varchar(50)	Identificador de la tupla modificada

Tabla 17: Registro_Modificación.

Nombre: registro_modificacion		
Descripción: Tabla utilizada para mantener el registro de las trazas de modificación de los datos de las planillas		
Atributo	**Tipo**	**Descripción**
idmod	serial	Valor entero de autoincremento, llave primaria
id_traza	integer	Llave foránea que toma de la entidad registro_trazas
valor_nuevo	varchar(255)	Valor nuevo de la modificación
valor_viejo	varchar(255)	Valor viejo
campo_modificado	varchar(50)	Nombre del campo que se modificó

Tabla 18: Ctrl_Desempeño.

Nombre: ctrl_desempenno		
Descripción: Tabla que recoge datos del control de desempeño del SIC0225		
Atributo	**Tipo**	**Descripción**
id_ctrol_desempenno	serial	Valor entero de autoincremento, llave primaria
no_folio_sic0225	varchar(15)	Llave foránea que toma de la entidad sic0225_certificado_analisis_mp
analisis_realizados	varchar(255)	Análisis realizados para el control de desempeño

limite_aceptacion	varchar(255)	Límite de Aceptación
resultados	varchar(255)	Resultados obtenidos

Tabla 19: Materia_Prima.

Nombre: materia_prima		
Descripción: Tabla que recoge datos de las Materias Primas		
Atributo	**Tipo**	**Descripción**
id_materia_prima	serial	Valor entero de autoincremento, llave primaria
no_folio_sic0225	integer	Llave foránea que la toma de la entidad sic0225_certificado_analisis_mp
id_libro_entrada_cuarentena	integer	Llave foránea que la toma de la entidad libro_entrada_cuarentena

2.5 Diseño del DW (Estrella)

Después de haber analizado detenidamente los sistemas operacionales de los que se van a extraer los datos y estudiado todas las necesidades del CIGB específicamente el módulo Análisis Químico del proyecto Lims Control de Calidad el DW quedo diseñado en esquema de estrella.

El esquema en estrella, consiste en estructurar la información en procesos, vistas y métricas recordando a una estrella. Es decir, se tendrá una visión multidimensional de un proceso que se mide a través de unas métricas.

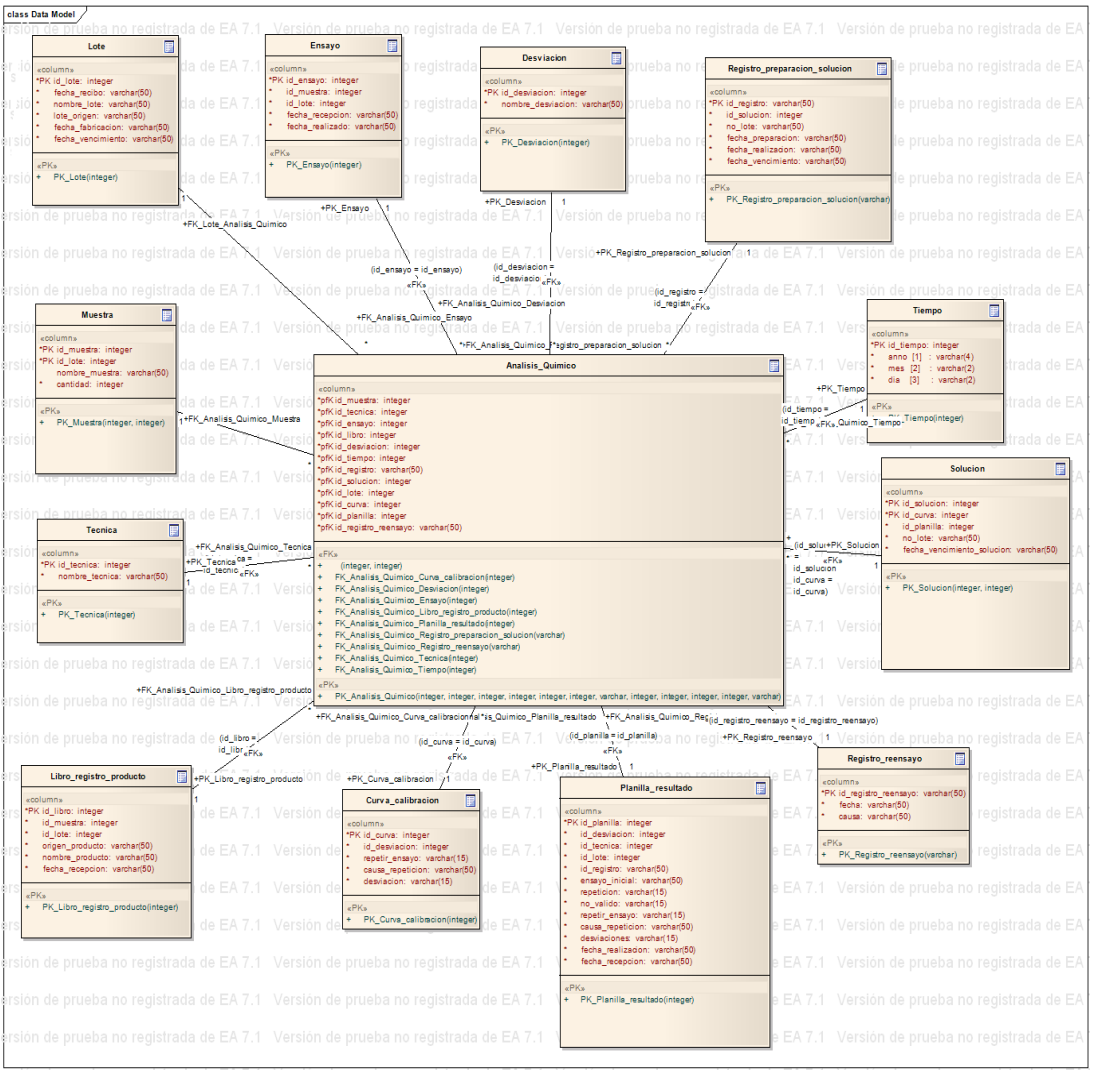
A nivel de diseño, consiste en una tabla de hecho y una o varias tablas de dimensión por cada dimensión de análisis que participa de la descripción de ese hecho. En la tabla de hecho se encuentran los atributos destinados a medir (cuantificar) el hecho: sus métricas.

En el esquema en estrella la tabla de hechos es la única tabla del esquema que tiene múltiples joins que la conectan con otras tablas (*foreign keys* hacia otras tablas). El

resto de tablas del esquema (tablas de dimensión) únicamente hacen joins con esta tabla de hechos.

Como se puede ver en la siguiente figura las tablas de dimensiones están ligadas a la tabla Hecho, por relaciones. La integridad referencial es llevada a cabo por la creación de llaves foráneas en la tabla hecho, que a su vez forman parte de la llave principal de esta tabla. Es importante destacar que las jerarquías completas son guardadas en una sola tabla dimensión.

Las tablas de dimensión se encuentran además totalmente desnormalizadas, es decir, toda la información referente a una dimensión se almacena en la misma tabla.



Figuras 6: Diseño del Esquema en Estrella del Data Warehouse. (21)

Conclusiones del Capítulo

En este capítulo se escogieron las herramientas que se usaron en el desarrollo del proceso de ETL, las mismas son muy eficientes en la actualidad y más teniendo en cuenta la meta del país de migrar a software libre. Se identificaron los indicadores resultantes de los interrogativos y sus respectivas perspectivas de análisis, mediante las cuales se construirá el modelo conceptual de datos del DW. Se analizaron los OLTP para señalar las correspondencias con los datos fuentes y seleccionar los campos de estudio de cada perspectiva.

CAPÍTULO 3: Diseño e Implementación del Proceso ETL

En este último capítulo además de dar los principales elementos para realizar el diseño e implementación del proceso ETL del Módulo de Análisis Químico, se darán algunos puntos de vista particulares, se mencionaran cuáles son las oportunidades que existen para trabajar en proyectos de este tipo y qué es lo que se espera de los grandes almacenes de datos en los próximos años.

3.1 Procesos ETL, limpieza de datos y sentencias SQL.

Después de analizar detenidamente la base de datos del módulo de Análisis Químico y haber diseñado el DW determinado cuales son las principales necesidades, se puede empezar a realizar el proceso ETL. Se diseñan e implementan las trasformaciones cuidadosamente teniendo siempre en cuenta que es lo que se tiene y cuáles son las fuentes de las que se van a extraer y como se van a mostrar.

3.2 Diseño e implementación de las trasformaciones para el módulo de Análisis Químico.

Al realizar de **extracción** se extraen datos de las diferentes fuentes, se convierten los datos a un formato estándar, con el cual se inicia el proceso de transformación. Una parte intrínseca del proceso de extracción es la de analizar los datos extraídos, de lo cual resulta un chequeo que verifica si los datos cumplen con las pautas estipuladas y se adaptan al formato estándar diseñado. De no ser así, los datos son rechazados.

En este caso los datos operacionales residen en un SGBD Relacional, por lo que proceso de extracción se puede reducir a, por ejemplo, consultas en SQL o rutinas programadas. En cambio, si se encontraran en un sistema propietario o fuentes externas, ya sean textuales, hipertextuales, hojas de cálculos, etc, la obtención de los

mismos puede ser un tanto más dificultoso, debido a que, se tendrán que realizar cambios de formato y/o volcado de información a partir de alguna herramienta específica.

Una vez que los datos son seleccionados y extraídos, se guardan en un **almacenamiento intermedio**, lo cual permite, entre otras ventajas:

- Manipular los datos sin interrumpir ni paralizar los OLTP, ni tampoco el DW.
- Almacenar y gestionar los metadatos que se generarán en los procesos ETL.
- Facilitar la integración de las diversas fuentes, internas y externas.

Luego en la fase de **trasformación** se aplican una serie de procedimientos de negocios sobre los datos extraídos, con el objeto de convertirlos en datos aptos para ser cargados realizando algunas transformaciones específicas en caso de que sea necesario. Estas transformaciones se pueden realizar en múltiples lenguajes de script. No siempre los datos están en la forma más adecuada para poder aplicar los métodos que hacen falta para la tarea que se ha de llevar a cabo y el modelo que se quiere obtener.

Los casos más comunes en los que se deberá realizar integración, son los siguientes:

- **Codificación**.
- Medida de atributos.
- **Convenciones de nombramiento.**
- Fuentes múltiples.
- **Seleccionar solo ciertas columnas para su carga.**

Por último en la fase de **carga** es cuando lo datos, de la fase anterior, son incluidos en el sistema de destino. Dependiendo de los requerimientos de la organización, este proceso puede abarcar una amplia variedad de acciones diferentes, en este caso el DW mantiene un historial de los registros de manera que se pueda hacer una auditoria de los mismos y disponer de un rastro del comportamiento de un determinado valor a lo largo del tiempo.

Una diferencia entre la carga inicial de los datos y la actualización de datos es que la verificación de la integridad referencial debe realizarse incrementalmente en la actualización antes que los datos sean cargados en el DW y queden disponibles para los usuarios.

Al generar los ETL, se debe tener en cuenta cual 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.

3.3 Preparación de los datos.

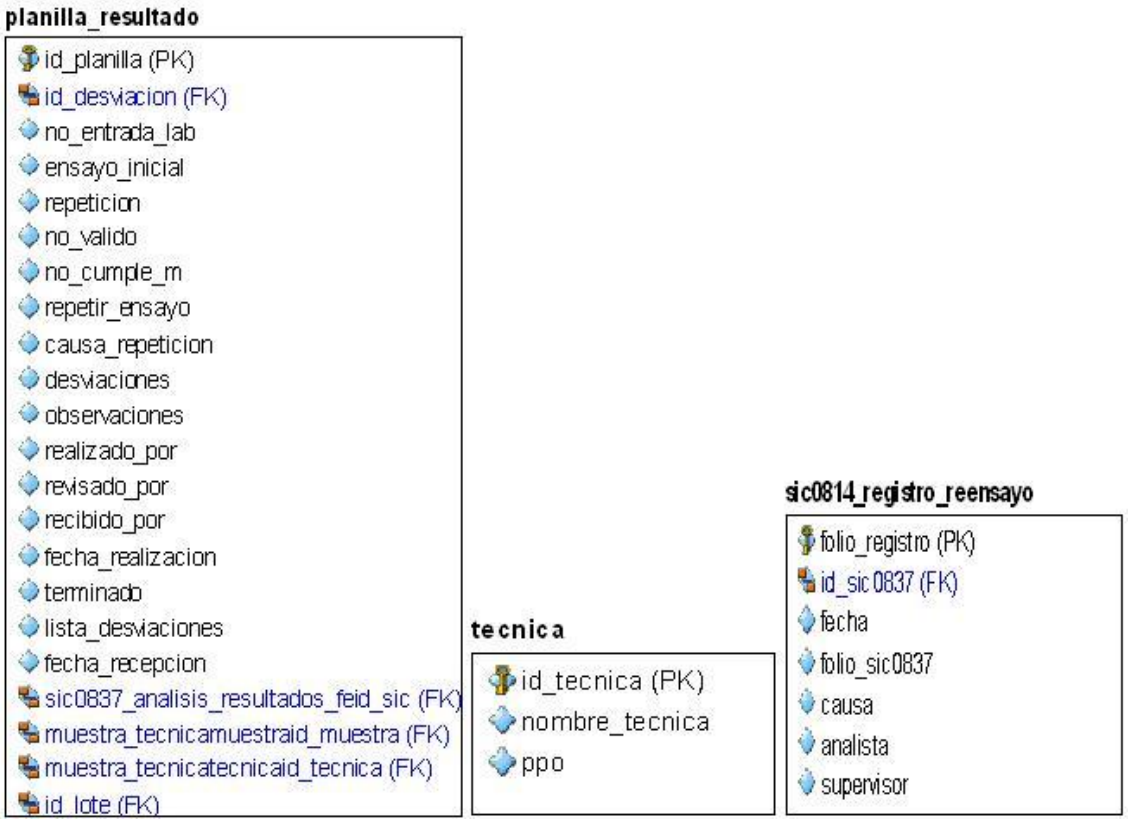
En este punto hay que asegurarse de:

- Que los datos tengan la calidad suficiente: es decir, que no contengan errores, redundancias o que presenten otro tipo de problemas.
- Que los datos sean los necesarios.
- Que estén en la forma adecuada: muchos métodos de construcción de modelos requieren que los datos estén en un formato determinado que no ha de coincidir necesariamente con el que está almacenado. (3)

Las **tablas de dimensiones** son elementos que contienen atributos (o campos) que se utilizan para restringir y agrupar los datos almacenados en una tabla de hechos cuando se realizan consultas sobre dicho datos en un entorno de almacén de datos.

Estos datos sobre dimensiones son parámetros de los que dependen otros datos que serán objeto de estudio y análisis y que están contenidos en la tabla de hechos. Las tablas de dimensiones ayudan a realizar ese estudio/análisis aportando información sobre los datos de la tabla de hechos, por lo que puede decirse que en un cubo OLAP, la tabla de hechos contiene los datos de interés y las tablas de dimensiones contienen metadatos sobre dichos hechos.

Planilla de Resultado: Tabla dimensión que recoge los datos de las Planillas de Resultados



Figuras 7: Tablas Planilla_Resultado, Técnica y SIC0814_Registro_Reensayo de la BD del Módulo Análisis Químico de Proyecto Lims Control de Calidad.

En esta transformación se quiere pasar los datos necesarios que están en las tablas planilla_ resultado técnica y sic0814_Registro_Reensayo de la bd_analisis_quimico para la dimensión Planilla de resultado del DW. Pare realizar esta trasformaciones se utilizan una serie de componentes que brinda la herramienta Kettle.

Los componentes se relacionan por vínculos que también se llaman **Hops** en Spoon y básicamente indican el sentido del flujo de los datos en una transformación. La manera más fácil para crear un Hop es arrastrar una línea entre dos objetos manteniendo la tecla SHIFT pulsada.

Selecciona/renombra valores se utiliza cuando se necesitas hacer cierta transformación una y otra vez, se puede convertir la parte repetitiva en un mapeo. Un mapeo es una transformación que:

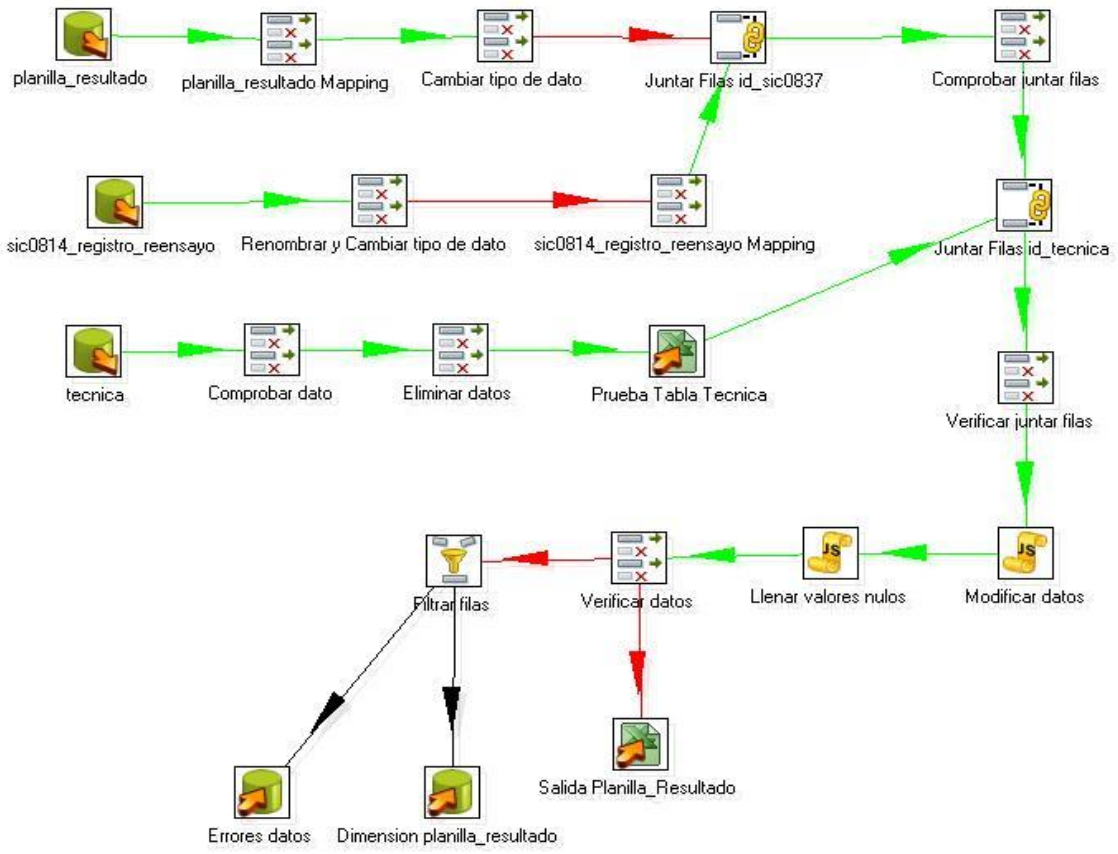
- especifica como la entrada va a utilizar un paso de Mapeo de Entrada.
- especifica como se transforman los campos de entrada: los campos que se añaden y eliminan.

Entre los componentes mas utilizado están el valor **Java Script** este paso que permite realizar cálculos complejos utilizando el lenguaje javascript. En este caso se utilizo para modificar datos, llenar y eliminar campos vacios.

Juntar filas es el componente que se utiliza para realizar la unión entre las diferentes tablas de la base datos.

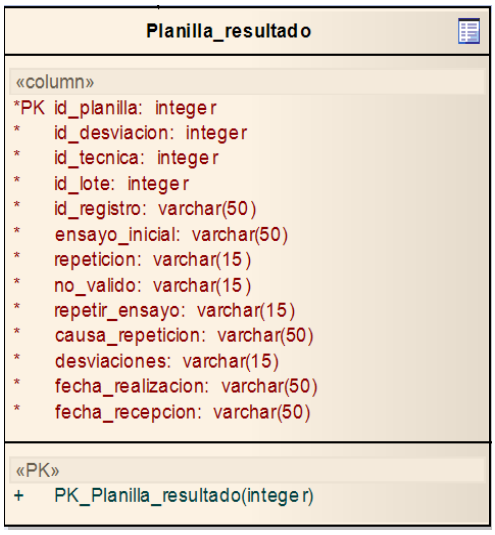
Salida Excel este paso es muy importante ya que en él se puede verificar cuales son los datos que se van almacenar antes de ser cargados en el DW. Para configurar este paso se abre una ventana con las propiedades del fichero, hay que poner el nombre, indicar el directorio donde guardarlo y si es necesario ajustar cualquier otra opción relacionada con el archivo.

Filtrar Filas este tipo de paso permite filtrar filas basado en condiciones y comparaciones. Una vez se haya conectado este paso a uno anterior (uno o más están recibiendo entrada), puedes utilizar las áreas de “<campo>”, “=” y “<valor>” para construir una condición.



Figuras 8: Proceso ETL para la Tabla Planilla_Resultado.

Por último se realiza la carga hacia la dimensión Planilla_resultado del DW y si hay datos que no se pueden cargar porque presentan algún error pasan directamente a la tabla errores de datos que está en almacenamiento intermedio creada para guardar todos los datos que no pueden ser cargados en el DW y luego analizarlos.



Figuras 9: Dimensión Planilla_Resultado del Data Warehouse.

Una vez que los datos han sido cargados en la base de datos del DW, se verifica la integridad referencial entre las dimensiones y la tabla principal con el objetivo de asegurar que todos los registros de una tabla estén relacionados correctamente con registros de otras tablas. Se debe verificar también que cada registro de la tabla principal se relacione con un registro en la tabla de la dimensión que será usada con esa tabla principal.

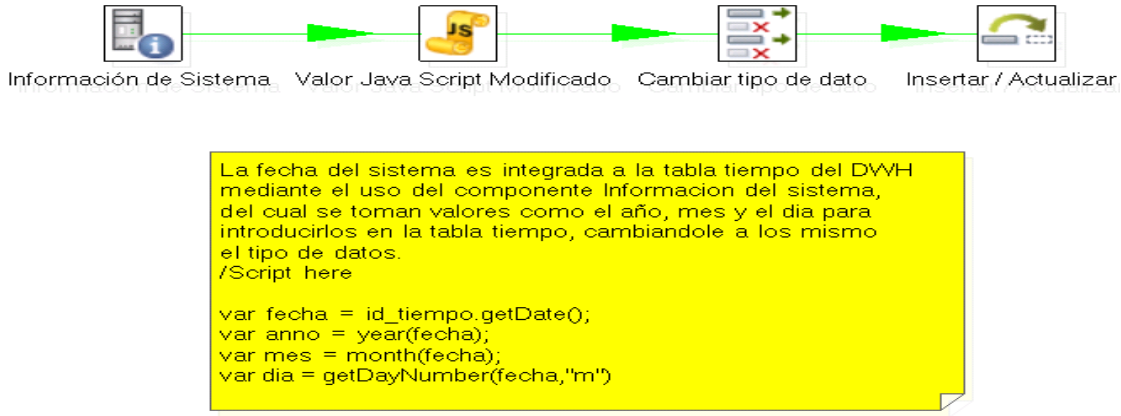
La integración de los datos en el orden contrario, no es necesaria, o sea, no es necesario para cada record perteneciente a la tabla de una dimensión relacionarse con un record en la tabla principal.

Tiempo: Dimensión que es agregada al DW.
En cualquier DW se pueden encontrar varios cubos con sus tablas de hechos repletas de registros sobre alguna variable de interés para el negocio que debe ser estudiada. Como ya se ha comentado, cada tabla de hechos estará rodeada de varias tablas de dimensiones, según que parámetros sirvan mejor para realizar el análisis de los hechos que se quieren estudiar. Un parámetro que casi con toda probabilidad será común a todos los cubos es el **tiempo**, ya que lo habitual es almacenar los hechos

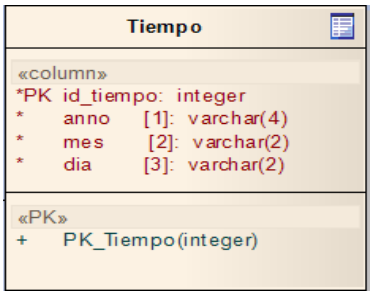
conforme van ocurriendo a lo largo del tiempo, obteniéndose así una serie temporal de la variable a estudiar.

Es importante tener en cuenta que el tiempo no es solo una secuencia cronológica representada de forma numérica, sino que posee fechas especiales que inciden notablemente en las actividades de la organización. Esto se debe a que los usuarios podrán por ejemplo analizar las actividades realizadas teniendo en cuenta el día de la semana en que se produjeron, quincena, mes, trimestre, semestre, año, etc.

Existen muchas maneras de diseñar esta tabla, y en adición a ello no es una tarea sencilla de llevar a cabo. Por estas razones se considera una buena práctica evaluar con cuidado la temporalidad de los datos, la forma en que trabaja la organización, los resultados que se esperan obtener del almacén de datos relacionados con una unidad de tiempo y la flexibilidad que se desea obtener de dicha tabla.



Figuras 10: Proceso ETL para la Tabla Tiempo.



Figuras 11: Dimensión Tiempo del Data Warehouse.

El Proceso ETL del resto de las dimensiones que integran el DW se pueden ver en el Anexo C.

3.4 Finalizando las transformaciones (trabajos).

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

Los trabajos en general controlan el flujo del trabajo, llamadas a más trabajos o transformaciones, generar log, finalizar la ejecución o Mail Obtener Mails y enviar emails. Gestor/Gestor Remoto de Ficheros Diferentes acciones de sistema, crear, comparar, mover, zip, unzip, ftp, scp.

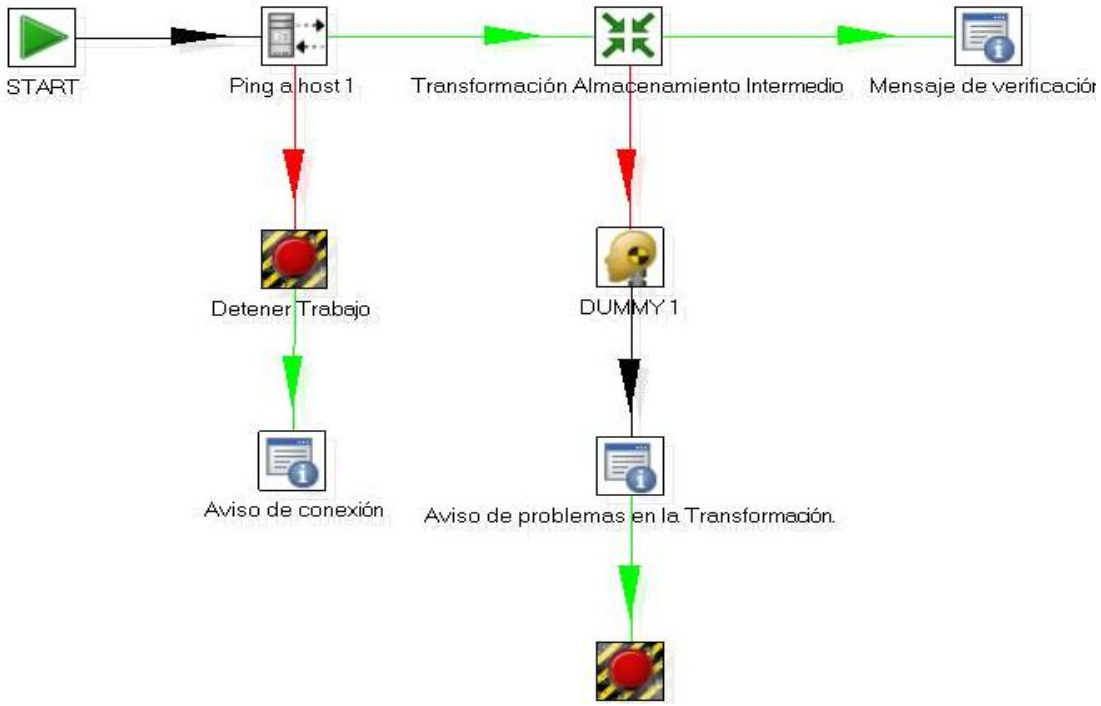
Comprueban si existen ficheros y datos en la base de datos, esperar a una cierta condición, o Scripting Shell, SQL, JavaScript XML Validador XML, XSL, XSD, transformación XSL o Volcados Masivos Importar y Exportar grandes cantidades de datos de diferentes base de datos. (12)

Para la ejecución de los trabajos, 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 mes.

Después de eso se verifica que el servidor donde serán guardados los datos esté funcionando, para esto es necesario encuestar a la computadora y si no está funcionando se manda un mensaje al usuario y se aborta la ejecución del trabajo.

Si está en funcionamiento entonces también se le informa al usuario y 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, en caso contrario, cuando todas se hayan ejecutado también se le comunica al usuario que las transformaciones se ejecutaron perfectamente.

La siguiente figura muestra el trabajo para ejecutar el almacenamiento intermedio que guarda todos los datos seleccionados en la fase de extracción.

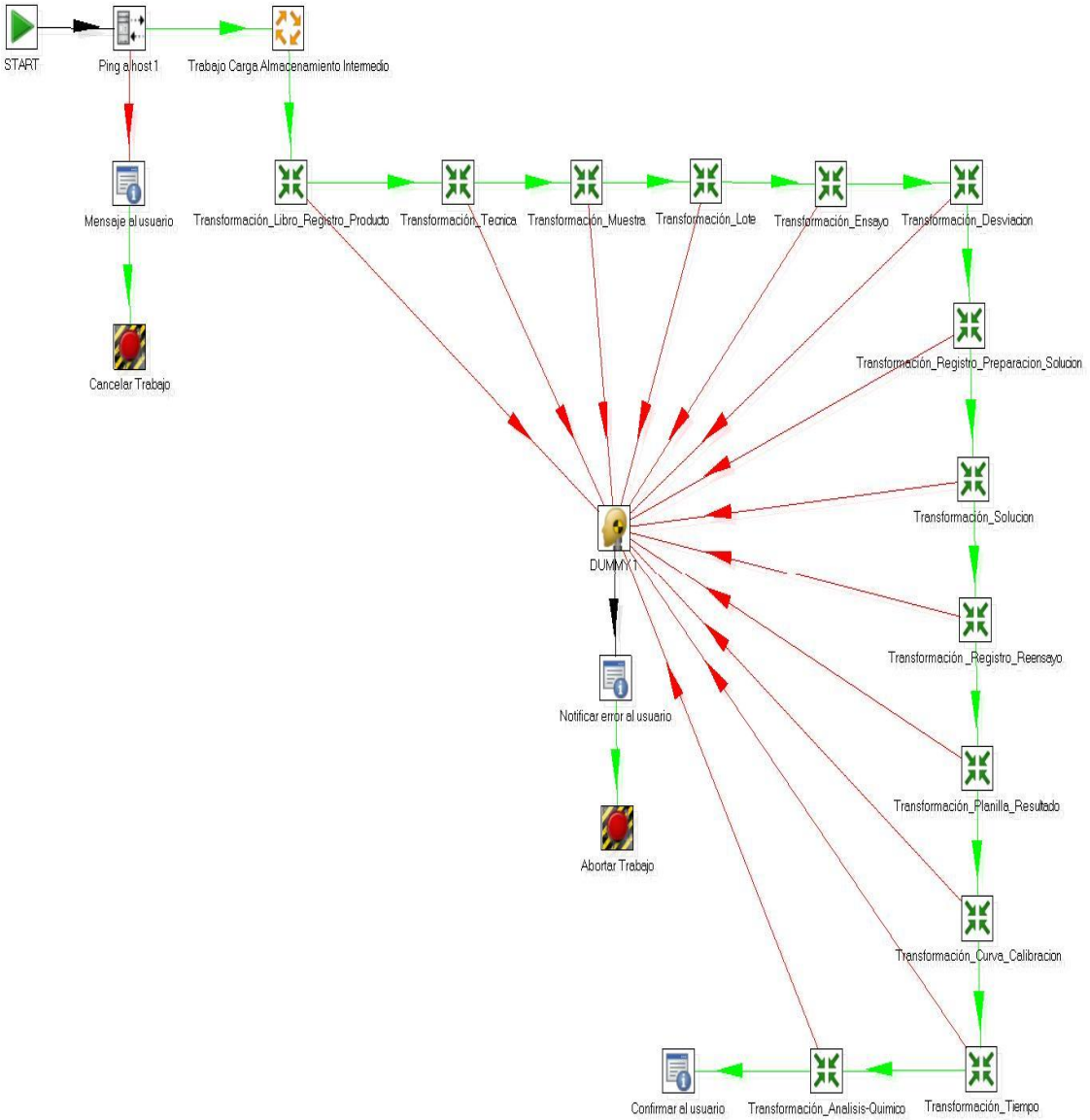


Figuras 12: Trabajo Almacenamiento Intermedio.

Para la ejecución del trabajo final hay que tener en cuenta que la tabla tiempo es una de las últimas en ejecutarse debido a que no es como las demás tablas, pues cuenta con un id_tiempo que es autoincremental y por lo tanto cuando se realiza su transformación, solo se le pasan los datos del año, mes y día en que se ejecute, estos datos son extraídos del sistema ya que la herramienta te permite realizar pasos como estos.

Cuando se termina de ejecutar esta transformación solo queda realizar la de la tabla Análisis_Químico que es la tabla de hechos, o sea en ella se encuentran todos los id de las dimensiones antes mencionadas y de ahí la importancia de que la misma sea la última en ejecutarse, pues no se puede llenar hasta que todas las dimensiones no estén llenas.

Es importante destacar que cada vez que se realiza cualquiera de las transformaciones que se encuentran dentro de este trabajo, si alguna de ellas presenta algún problema que impida su ejecución, el usuario es avisado mediante un mensaje inmediatamente.



Figuras 13: Trabajo Final.

3.5 Los Retos del ETL.

Existen numerosos desafíos para implementar procesos ETL eficaces y fiables.

- Los volúmenes de datos crecen de forma exponencial, y los procesos ETL tienen que procesar grandes cantidades de datos granulares (productos vendidos, llamadas telefónicas, transacciones bancarias…). Algunos sistemas de BI se actualizan simplemente de manera incremental, mientras que otros requieren una recarga completa en cada iteración.

- Los procesos ETL necesitan una extensa conectividad a las aplicaciones en paquetes (ERP, CRM, etc.), bases de datos, mainframes, archivos, servicios Web, etc.
- Las estructuras y aplicaciones de Inteligencia de negocio incluyen los almacenes de datos históricos generales e individuales y las aplicaciones OLAP, para el análisis, notificación, cuadros de mando operacionales, tácticos (dashboarding), estratégicos (scorecarding), etc. Todas estas estructuras objetivo tienen **requisitos diferentes de transformación de datos,** y distintas latencias.
- Las transformaciones implicadas en los procesos ETL pueden ser **muy complejas**. Los datos necesitan agregarse, analizarse, computarse, procesarse estadísticamente, etc. También se necesitan transformaciones específicas a BI, como Slowly Changing Dimensions.

Mientras que la Inteligencia de Negocio tiende hacia una **puntualidad real**, los almacenes de datos generales e individuales se tienen que actualizar más a menudo, ya que las ventanas de tiempo de carga se reducen. (13)

3.6 Oportunidades.

El espectacular crecimiento de los datos almacenados por organizaciones y empresas constituye una de las tendencias más destacadas en el panorama actual de las nuevas tecnologías. Esta situación requiere medidas para garantizar un rendimiento y una protección de los datos que se puedan implementarse en las mejores condiciones cuando quienes participan en el proceso completo de toma de decisiones son conscientes de su importancia.

Las herramientas ETL apoyan también el intercambio de información entre sistemas empresariales, ayudando a crear organizaciones en tiempo real, un entorno que maximiza la agilidad de la organización y en el que la información más actualizada está disponible para aquellos que más la necesitan. Además de una mayor demanda de información actualizada, relevante y detallada, el crecimiento del tamaño de los Data Warehouse está aumentando la necesidad por parte de las organizaciones de instalar herramientas ETL con mayores capacidades. Las bases de datos de varios terabytes de tamaño son ya moneda común. Mientras el volumen de la información empresarial aumenta y sus fuentes se multiplican, mientras las “batch windows” disminuyen o incluso desaparecen en un mundo de acceso las 24 horas al día y los

siete días de la semana, las organizaciones buscan soluciones de integración que sean capaces de consolidar rápidamente toda su información en bruto y convertirla en datos útiles que puedan ser utilizados para el análisis, la toma de decisiones o en la demanda de información por parte de usuarios finales.

3.7 Comentarios Finales

Se ha visto, a lo largo de este trabajo, qué es un *Data Warehouse,* qué *son los procesos ETL* y cuáles son sus principales aplicaciones, entre otras cosas. Se cuenta, en la actualidad con herramientas muy poderosas que se están introduciendo cada vez más en el mundo empresarial y científico, no obstante, queda mucho camino por recorrer, se habla de tecnología, de oportunidades, de cómo ha evolucionado dicha tecnología y lo que se espera de ella, sin embargo, su campo de aplicación sigue siendo un tanto reducido en países de los llamados tercer mundo y en algunas otras áreas diferentes a las empresariales.

La pregunta ahora sería ¿Las computadoras ahora pensarán por los humanos?, una posible respuesta sería que se debe trabajar en que ellas piensen, es decir, que descubran lo que se quiere que descubran, que sirvan para lo que se desea que sirvan, el uso de tecnologías como ETL conlleva muchas facilidades hacia la vida de los seres humanos, las computadoras se han hecho para facilitar la vida cotidiana no para complicarla y *los procesos* ETL representa una de estas facilidades, la cual por supuesto evolucionará.

Conclusiones del Capítulo

En este capítulo se diseñaron e implementaron los procesos ETL para el módulo Análisis Químico del proyecto Sistemas de Gestión de Información de Laboratorios (Lims). También se mencionaron numerosos desafíos para implementar procesos ETL fiables. Además de dar algunos puntos de vistas particulares sobre el trabajo realizado.

Conclusiones

Como resultado de este trabajo se realizó un análisis sobre las técnicas de extracción, transformación y carga de datos dando una visión de cómo resolver el problema, lo que permitió un mayor avance en la concepción e implementación de la solución.

Con el estudio sobre posibles tecnologías a utilizar para el proceso ETL se seleccionaron cuidadosamente las herramientas que responden a las posibilidades tecnológicas de la institución.

Se analizaron los OLTP para señalar las correspondencias con los datos fuentes y seleccionar los campos de estudio de cada perspectiva.

Se realizó el diseño e implementación de 13 trasformaciones y 2 trabajos, utilizando la herramienta Kettle y la metodología HEFFESTO que demostrando ser la solución idónea para realizar la migración de los datos hacia un Data Warehouse.

Recomendaciones

- Lograr la integración con el resto de los módulos del proyecto LIMS Control de Calidad para el CIGB.

Referencias Bibliográficas

1. **Vidal, LV y Monteagudo, MV.** *Estudio Teórico_Conceptual sobre Data Warehouse.* Ciudad de la Habana : s.n., 2000.

2. SQLMax Connections . *Data Warehousing.* [En línea] 2001. [Citado el: 04 de 05 de 2009.] http://www.sqlmax.com/dataw1.asp.

3. **Kimball, Ralph.** *El Juego de herramientas del Almacén de Datos.* s.l. : John Wiley & Sons., 1996.

4. **Rizo Rizo, Emma, y otros.** *Importancia de la utilización de un Data Warehouse (DW) en las.* 2007.

5. **Sánchez García, Alberto y Puig Pinto, Jorge Carlos.** *Sistema para la Gestión de la Información de Laboratorios de la Dirección de Calidad del Centro de Ingeniería Genética y Biotecnología: Implementación del Módulo Análisis Químico.* 2009.

6. **Bernabeu, Ricardo Dario.** *DATA WAREHOUSING:Investigación y Sistematización de Conceptos – HEFESTO:Metodología propia para la Construcción de un Data Warehouse.* Córdoba, Argentina : s.n., 2007.

7. Wikipedia, la enciclopedia libre. *ETL - Wikipedia, la enciclopedia libre.htm.* [En línea] 21 de 04 de 2009. [Citado el: 04 de 05 de 2009.] http://es.wikipedia.org/wiki/ETL.

8. **Inmon, Bill.** *Building the Data Warehouse* . Indianapolis : Wiley Publishing , 2005. 4ta.

9. Mundo Business Intelligence. *Herramienta ETL (...o Mundo ETL).* [En línea] [Citado el: 04 de 05 de 2009.] http://mundobi.wordpress.com/2007/06/24/herramientas-etl-%E2%80%A6o-mundo-etl/.

10. **Calvo, Jorge Mario.** ACIS. *BI al alcance de todos.* [En línea] 30 de 11 de 2005. [Citado el: 05 de 05 de 2009.] http://www.acis.org.co/index.php?id=622.
Sistema DataWarehouse Comercial de la Corporación CIMEX.

11. Todo BI <business intelligence>. *Pentaho: la solucion Open Source Business Intelligence* . [En línea] 21 de 05 de 2006. [Citado el: 05 de 05 de 2009.] http://todobi.blogspot.com/2006/05/pentaho-la-solucion-open-source.html.

12. Todo BI . *Kettle: ETL para Pentaho.* [En línea] 05 de 04 de 2006. [Citado el: 05 de 05 de 2009.] http://todobi.blogspot.com/2006/04/kettle-etl-para-pentaho.html.
14. Talend Open. *Talend Open Studio.* [En línea] [Citado el: 05 de 05 de 2009.] http://es.talend.com/products-data-integration/talend-open-studio.php#feature.

13. Talend Open Data Solutions. *ETL para Analytics.* [En línea] 2008. [Citado el: 04 de 05 de 2009.] http://es.talend.com/solutions-data-integration/etl-for-analytics.php.

14. Talend Open. *Talend Open Studio.* [En línea] [Citado el: 05 de 05 de 2009.] http://es.talend.com/products-data-integration/talend-open-studio.php#feature.

15. Portada sobre la plataforma Pentaho Open Source Business Intelligence . *La plataforma Pentaho Open Source Business Intelligence* . [En línea] [Citado el: 05 de 05 de 2009.] http://pentaho.almacen-datos.com/.

16. **Pecos, Daniel.** *PostGreSQL vs. MySQL.* [En línea] [Citado el: 04 de 05 de 2009.] http://www.netpecos.org/docs/mysql_postgres/index.html.

17. pgAdmin . *PostgreSQL Tool.* [En línea] [Citado el: 05 de 05 de 2009.] http://www.pgadmin.org/.

18. **Luján-Mora, Sergio.** *Data Warehouse Desing with UML.* Universidad de Alicante : s.n., 2005.

19. **Rodríguez Ruiz, Alfredo y Vega Calcines, Anabel.** *LIMS de Calidad del Centro de Ingeniería Genética y Biotecnologia* . Ciudad de la Habana : s.n., 2008.

20. **Zorrilla, Marta.** *Data Warehouse y OLAP.* Universidad de Cantabria : s.n., 2007.

21. **Suárez Giro, Milagros de la Caridad y González Sardina, Jose Raúl.** *Realizar el Diseño e Implementación de un Data Warehouse para el proyecto LIMS CONTROL DE CALIDAD.* Ciudad de la Habana : s.n., 2009.

Bibliografía

Bernabeu, Ricardo Dario. 2007. DATA WAREHOUSING:Investigación y Sistematización de Conceptos – HEFESTO:Metodología propia para la Construcción de un Data Warehouse. Córdoba, Argentina : s.n., 2007.

Calvo, Jorge Mario. 2005. ACIS. BI al alcance de todos. [En línea] 30 de 11 de 2005. [Citado el: 05 de 05 de 2009.] 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.

GONZÁLEZ, A. DESARROLLO METODOLÓGICO PARA LA IMPLEMENTACIÓN DEL PROCESO DE DATAWAREHOUSING EN CUBAENERGIA 2008, 3: 16.

HERNÁNDEZ, R. and S. COELLO. El paradigma cuantitativo de la investigación científica Ciudad de la Habana, Universidad de las Ciencias Informáticas, Noviembre del 2002, 114 . p.

HOWARD, P. El caso de ETL de código abierto 23 de diciembre del 2005.

HURTADO, C. Repositorios (data warehouses) OLAP. Tuluá, Universidad de Chile, Abril del 2005, 35 . p.

ICFLORES, B. Repositorio de objetos en kettle 21 de mayo del 2007, 7. p.

INMON, W. Building the Data Warehouse, Fourth Edition. Wiley Publishing, Inc., 2005. 543 p. 4. 13; 978-0-7645-9944-6

Kimball, Ralph. 1996. El Juego de herramientas del Almacén de Datos. s.l. : John Wiley & Sons., 1996.

Luján-Mora, Sergio. 2005. Data Warehouse Desing with UML. Universidad de Alicante : s.n., 2005.

MARRERO, I. La inteligencia de negocios desde la perspectiva cubana: retos y tendencias. Cuba, Universidad de la Habana, 2008. 73 p.

MORA, S. L.-. Data Warehouse Desing with UML, junio del 2005. 318. p.

Mundo Business Intelligence. Herramienta ETL (...o Mundo ETL). [En línea] [Citado el: 04 de 05 de 2009.] http://mundobi.wordpress.com/2007/06/24/herramientas-etl-%E2%80%A6o-mundo-etl/.

Pecos, Daniel. PostGreSQL vs. MySQL. [En línea] [Citado el: 04 de 05 de 2009.] http://www.netpecos.org/docs/mysql_postgres/index.html.
pgAdmin . PostgreSQL Tool. [En línea] [Citado el: 05 de 05 de 2009.] http://www.pgadmin.org/.

PERALTA, V. Un caso de estudio sobre diseño lógico de Data Warehouse, 2008, 6: 19.

Portada sobre la plataforma Pentaho Open Source Business Intelligence . La plataforma Pentaho Open Source Business Intelligence . [En línea] [Citado el: 05 de 05 de 2009.] http://pentaho.almacen-datos.com/.

Rizo Rizo, Emma, y otros. 2007. Importancia de la utilización de un Data Warehouse (DW) en las. 2007.

Rodríguez Ruiz, Alfredo y Vega Calcines, Anabel. 2008. LIMS de Calidad del Centro de Ingeniería Genética y. Ciudad de la Habana : s.n., 2008.

SALINAS, A. Introducción a Pentaho. Business Intelligence, Data Mining, Data Warehouse, PENTAHO, Tecnología y Negocios, 12 de marzo del 2008: 2.

Sánchez García, Alberto y Puig Pinto, Jorge Carlos. 2009. Sistema para la Gestión de la Información de Laboratorios de la Dirección de Calidad del Centro de Ingeniería Genética y Biotecnología: Implementación del Módulo Análisis Químico.

Sistema DataWarehouse Comercial de la Corporación CIMEX.

SERRANO, M. Data Warehouse. Almacenamiento y Recuperación de la Información, 2007: 21.

SQLMax Connections . Data Warehousing. [En línea] 2001. [Citado el: 04 de 05 de 2009.] http://www.sqlmax.com/dataw1.asp.

Suárez Giro, Milagros de la Caridad y González Sardina, Jose Raúl. 2009. Realizar el Diseño e Implementación de un Data Warehouse para el proyecto LIMS CONTROL DE CALIDAD. Ciudad de la Habana : s.n., 2009.

Talend Open. Talend Open Studio. [En línea] [Citado el: 05 de 05 de 2009.] http://es.talend.com/products-data-integration/talend-open-studio.php#feature.

2008. Talend Open Data Solutions. ETL para Analytics. [En línea] 2008. [Citado el: 04 de 05 de 2009.] http://es.talend.com/solutions-data-integration/etl-for-analytics.php.

2006. Todo BI . Kettle: ETL para Pentaho. [En línea] 05 de 04 de 2006. [Citado el: 05 de 05 de 2009.] http://todobi.blogspot.com/2006/04/kettle-etl-para-pentaho.html.

2006. Todo BI <business intelligence>. Pentaho: la solucion Open Source Business Intelligence . [En línea] 21 de 05 de 2006. [Citado el: 05 de 05 de 2009.] http://todobi.blogspot.com/2006/05/pentaho-la-solucion-open-source.html.

VEGA, A. and A. RODRÍGUEZ. LIMS de Calidad del Centro de Ingeniería Genética y Biotecnología: Desarrollo de la Base de Datos del Módulo Liberación Analítica, Universidad de las Ciencias Informáticas, 24 de junio del 2008. 118. p.

Vidal, LV y Monteagudo, MV. 2000. Estudio Teórico_Conceptual sobre Data Warehouse. Ciudad de la Habana : s.n., 2000.

2009. Wikipedia, la enciclopedia libre. ETL - Wikipedia, la enciclopedia libre.htm. [En línea] 21 de 04 de 2009. [Citado el: 04 de 05 de 2009.] http://es.wikipedia.org/wiki/ETL.

Zorrilla, Marta. 2007. Data Warehouse y OLAP. Universidad de Cantabria : s.n., 2007.

SQLMax Connections . *Data Warehousing.* [En línea] 2001. [Citado el: 04 de 05 de 2009.] http://www.sqlmax.com/dataw1.asp

Wikipedia, la enciclopedia libre. *ETL - Wikipedia, la enciclopedia libre.htm.* [En línea] 21 de 04 de 2009. [Citado el: 04 de 05 de 2009.] http://es.wikipedia.org/wiki/ETL

Mundo Business Intelligence. *Herramienta ETL (...o Mundo ETL).* [En línea] [Citado el: 04 de 05 de 2009.] http://mundobi.wordpress.com/2007/06/24/herramientas-etl-%E2%80%A6o-mundo-etl/

Calvo, Jorge Mario. ACIS. *BI al alcance de todos.* [En línea] 30 de 11 de 2005. [Citado el: 05 de 05 de 2009.] http://www.acis.org.co/index.php?id=622

Todo BI <business intelligence>. *Pentaho: la solucion Open Source Business Intelligence .* [En línea] 21 de 05 de 2006. [Citado el: 05 de 05 de 2009.] http://todobi.blogspot.com/2006/05/pentaho-la-solucion-open-source.html

Todo BI . *Kettle: ETL para Pentaho.* [En línea] 05 de 04 de 2006. [Citado el: 05 de 05 de 2009.] http://todobi.blogspot.com/2006/04/kettle-etl-para-pentaho.html.
14. Talend Open. *Talend Open Studio.* [En línea] [Citado el: 05 de 05 de 2009.] http://es.talend.com/products-data-integration/talend-open-studio.php#feature

Talend Open Data Solutions. *ETL para Analytics.* [En línea] 2008. [Citado el: 04 de 05 de 2009.] http://es.talend.com/solutions-data-integration/etl-for-analytics.php

Talend Open. *Talend Open Studio.* [En línea] [Citado el: 05 de 05 de 2009.] http://es.talend.com/products-data-integration/talend-open-studio.php#feature

Portada sobre la plataforma Pentaho Open Source Business Intelligence . *La plataforma Pentaho Open Source Business Intelligence .* [En línea] [Citado el: 05 de 05 de 2009.] http://pentaho.almacen-datos.com/

Pecos, Daniel. *PostGreSQL vs. MySQL.* [En línea] [Citado el: 04 de 05 de 2009.] http://www.netpecos.org/docs/mysql_postgres/index.html

pgAdmin . *PostgreSQL Tool.* [En línea] [Citado el: 05 de 05 de 2009.] http://www.pgadmin.org

Anexos

Anexo A

Para crear un repositorio de datos se uso Spoon que permite crear y actualizar los repositorios de Kettle desde un interfaz grafico. Es sencillo, en la primera ventana de Spoon (si no se a deshabilitado su aparición) se puede seleccionar el repositorio que se desea utilizar, al mismo tiempo también permite las opciones de editar, actualizar (migrar a la nueva versión) o borrar el repositorio.

Para crear un nuevo repositorio, se pulsa el botón "*Nuevo*", y se crea una conexión a la base de datos donde se almacenará.

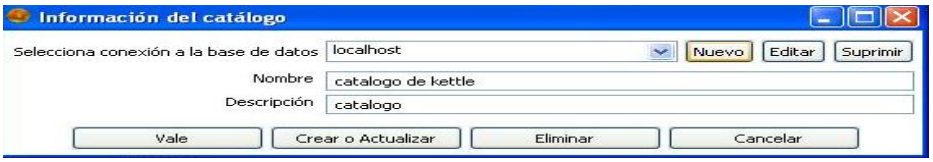


Figura: 1 Repositorio de Kettle.

Para crear una nueva conexión, es importante tener en cuenta dos cosas:

1. El tipo de conexión debe ser *Native (JDBC)*.
2. El usuario de acceso a la base de datos tienen que tener permisos para crear tablas al menos durante el proceso de creación del repositorio, sino obviamente no podrá crear las tablas del repositorio.

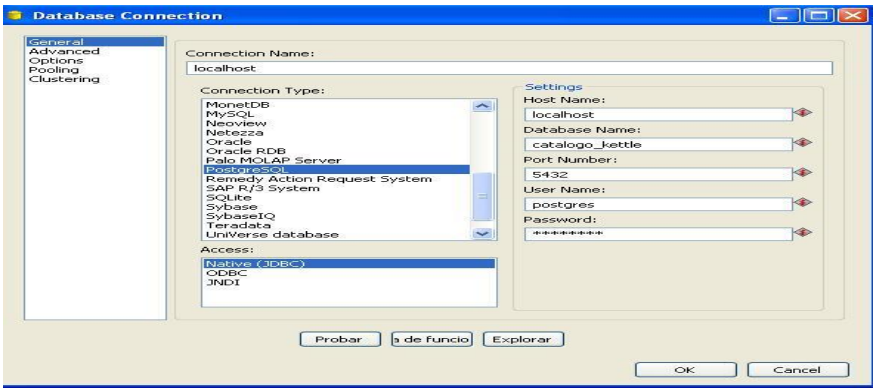


Figura: 2 Conexión a la base de datos para crear el repositorio.

Una vez configurada la conexión se dará un nombre y una descripción al repositorio, si todo es correcto, Spoon creará automáticamente todas las tablas y datos necesarios para el repositorio en la base de datos. Al mismo tiempo, toda la configuración de acceso creada quedará almacenada en el fichero "*repositories.xml*" dentro del directorio "*.Kettle*" (este a su vez dentro del directorio personal del usuario).

Una vez creado el repositorio, se necesita también un usuario y un password para poder conectarse. El primer usuario/password en un repositorio recién creado es admin/admin. El password puede ser (o mejor dicho, debería ser) cambiado después con el explorador de repositorio.

ANEXO B

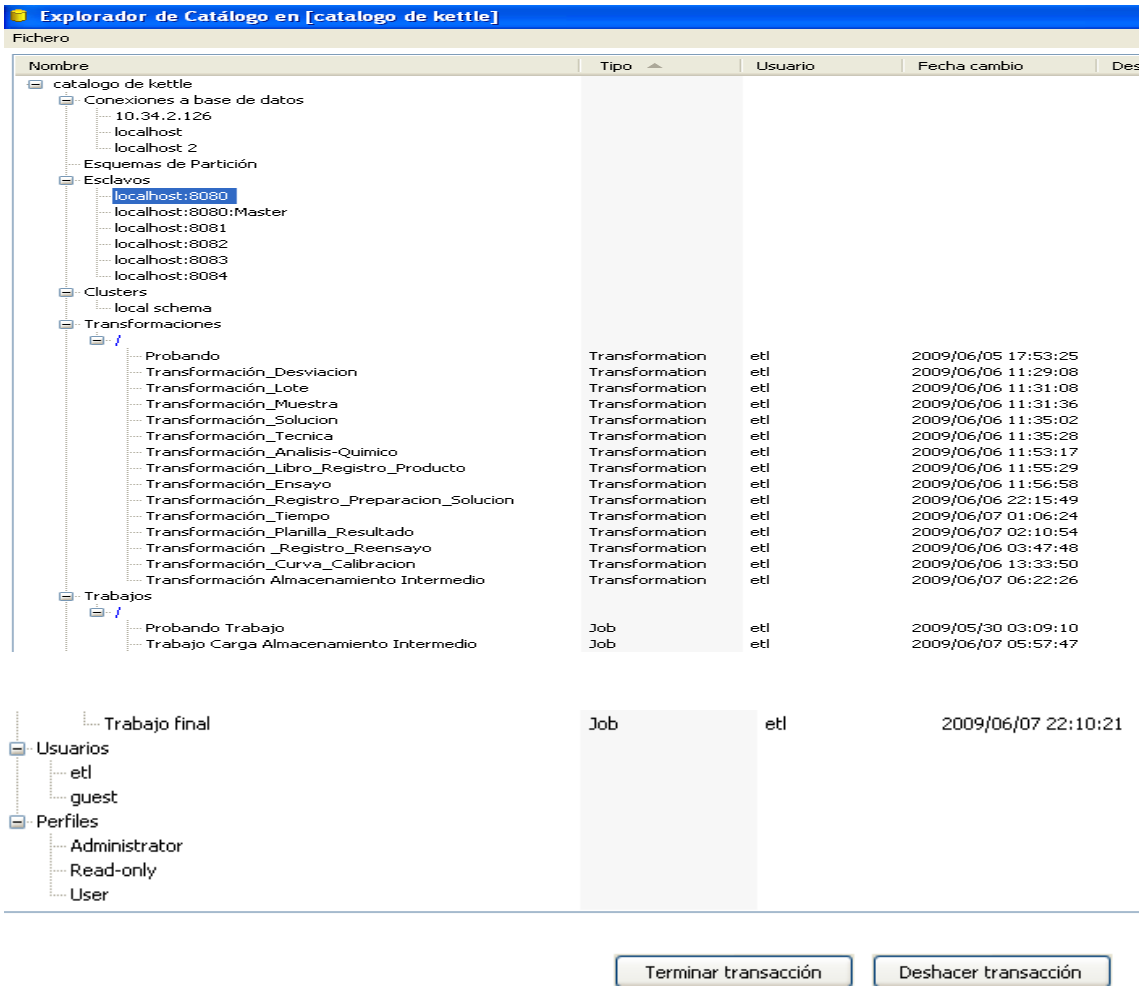


Figura: 3 Explorador de Kettle

ANEXO C

1. **Lote:** Es la entidad que recoge la información referente al lote de las muestras.

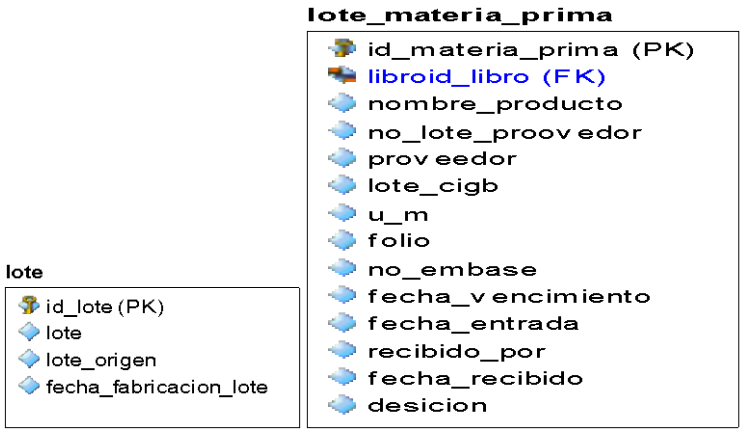


Figura 4: Tablas Lote y lote_materia_prima de la BD del Módulo Análisis Químico del proyecto Lims Control de la Calidad.

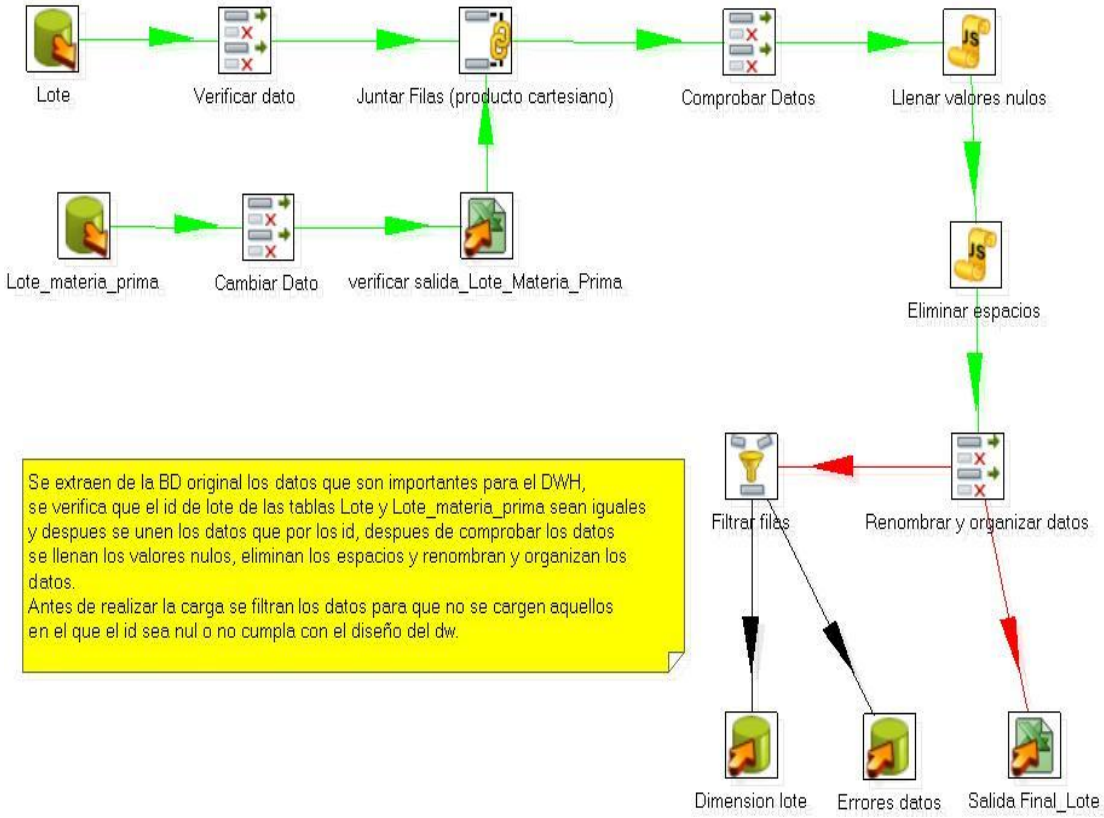


Figura 5: Proceso ETL para la tabla Lote.

Finalmente en la tabla lote que es la entidad que recoge la información referente al lote de las muestras del DW quedaron estos datos que se cargaron satisfactoriamente.

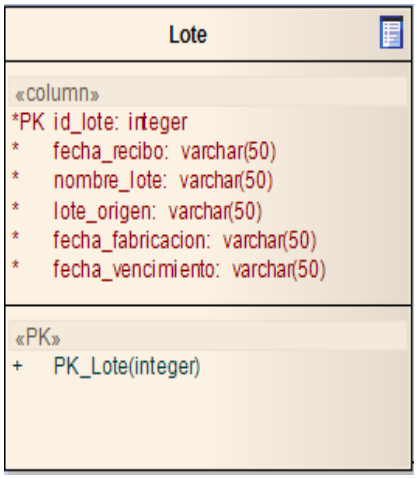


Figura 6: Dimensión Lote del DW.

2. **Ensayo:** Tabla que guarda los datos generales de todos los SIC de ensayos de los laboratorio del CIGB.

Figura 7: Tabla sic_ensayos de la BD del Módulo Análisis Químico del proyecto Lims Control de la Calidad.

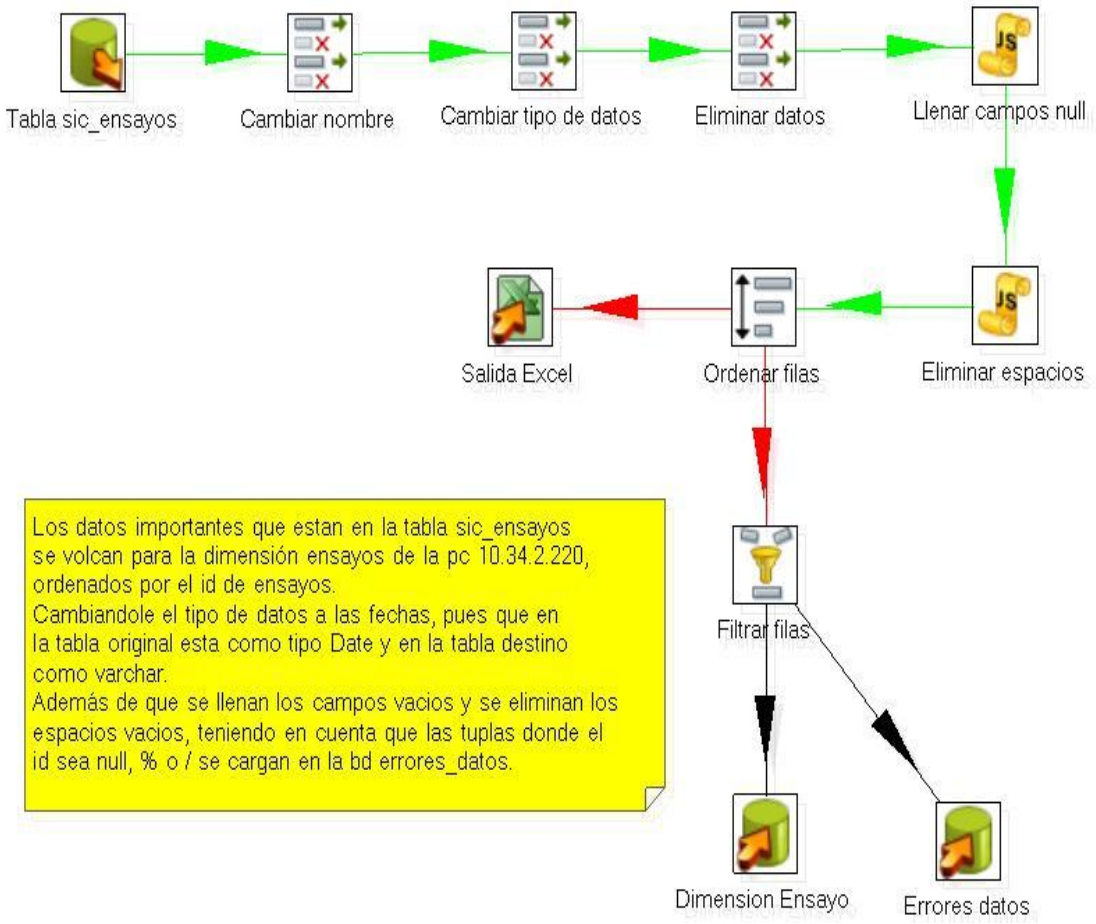


Figura 8: Proceso ETL para la tabla Ensayo.

Se cargaron los siguientes datos en la tabla Ensayo del DW que guarda los datos generales de todos los sic de ensayos de los laboratorios del CIGB.

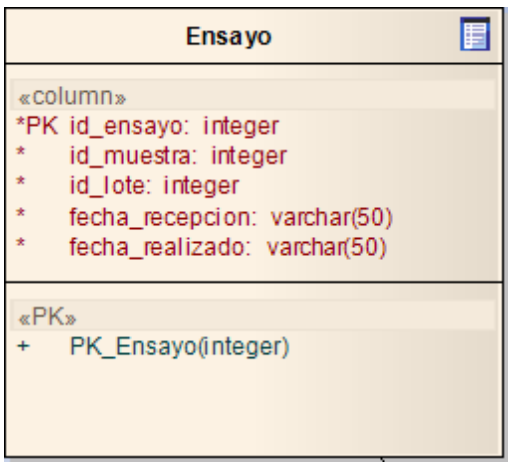


Figura 9: Dimensión Ensayo del DW.

3. **Solución:** Tabla que recoge los datos de las soluciones.

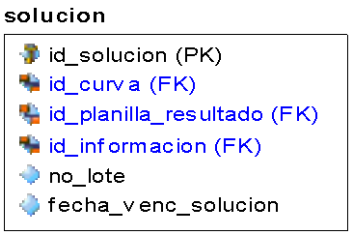


Figura 10: Tabla solucion de la BD del Módulo Análisis Químico del proyecto Lims Control de la Calidad.

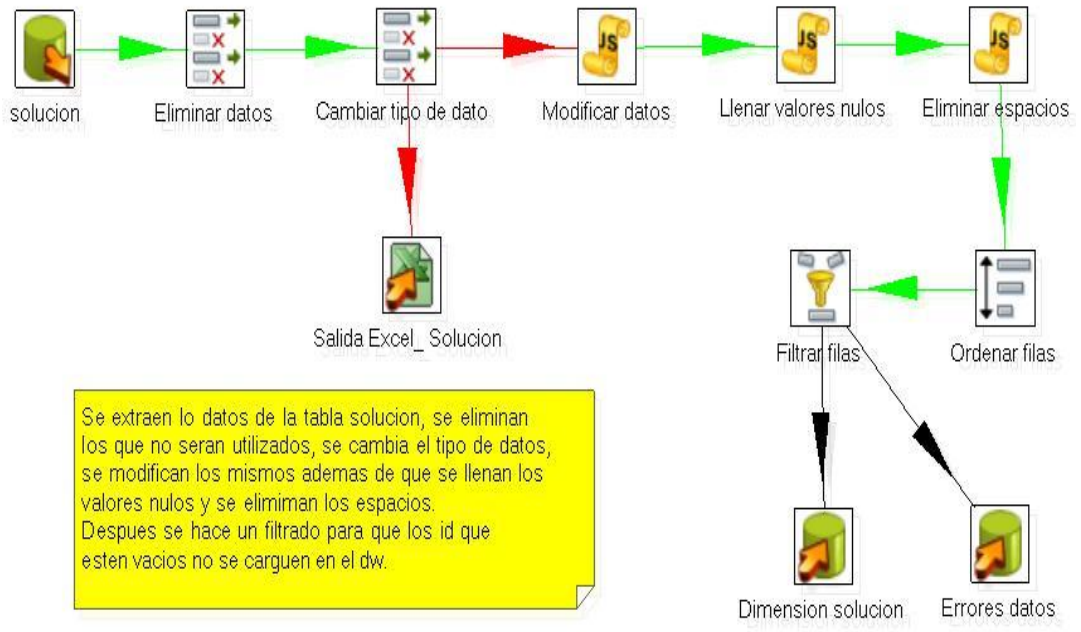


Figura 11: Proceso ETL para la tabla Solución.

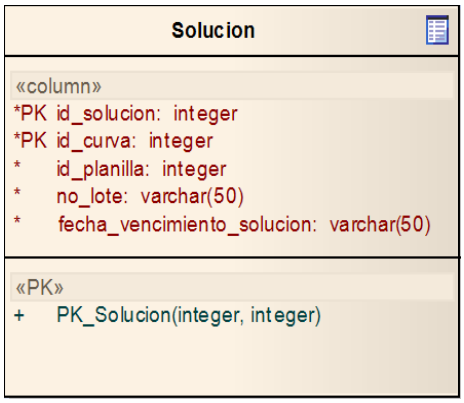


Figura 12: Dimensión Solución del DW.

4. **Muestra:** Representa los datos de la muestra a registrar en el libro

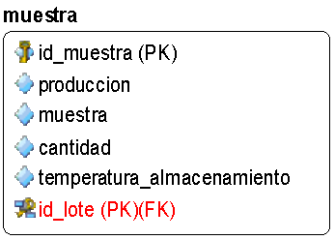


Figura 13: Tabla muestra de la BD del Módulo Análisis Químico del proyecto Lims Control de la Calidad.

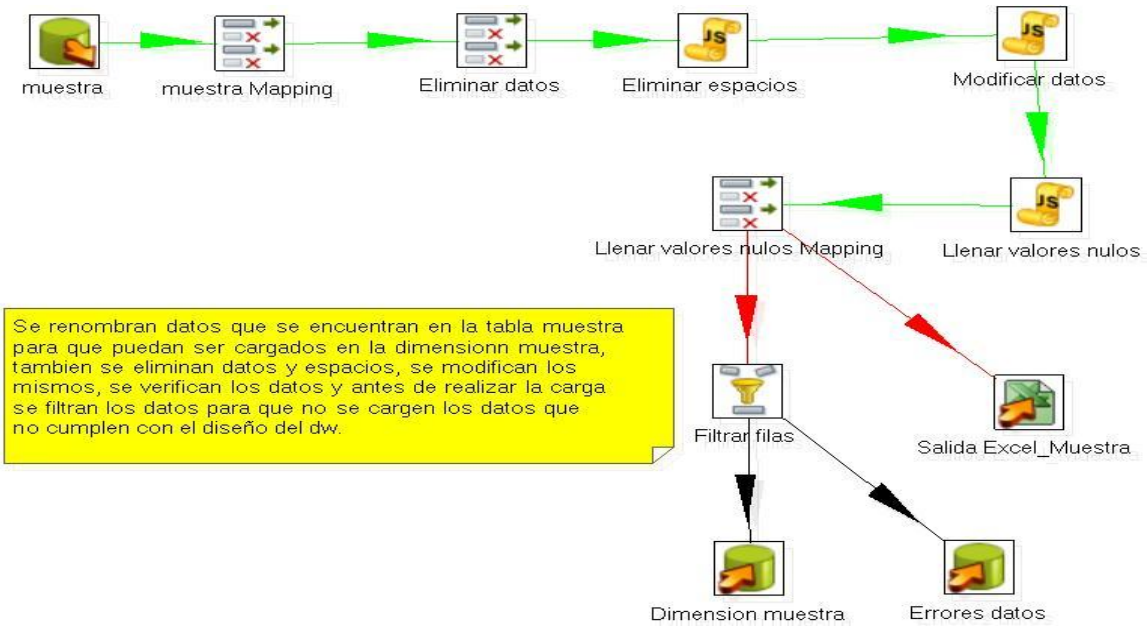


Figura 14: Proceso ETL para tabla la Muestra.

Muestra: Representa los datos de la muestra al registrarse en el libro.

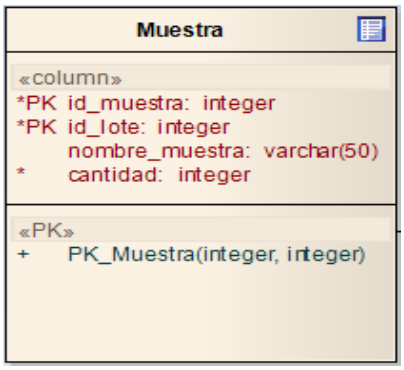


Figura 15: Dimensión Muestra del DW.

5. Técnica: Tabla que recoge los nombres de las técnicas que se aplican en el laboratorio.

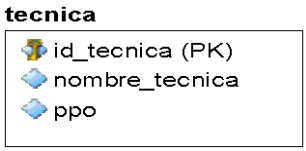


Figura 16: Tabla técnica de la BD del Módulo Análisis Químico del proyecto Lims Control de la Calidad.

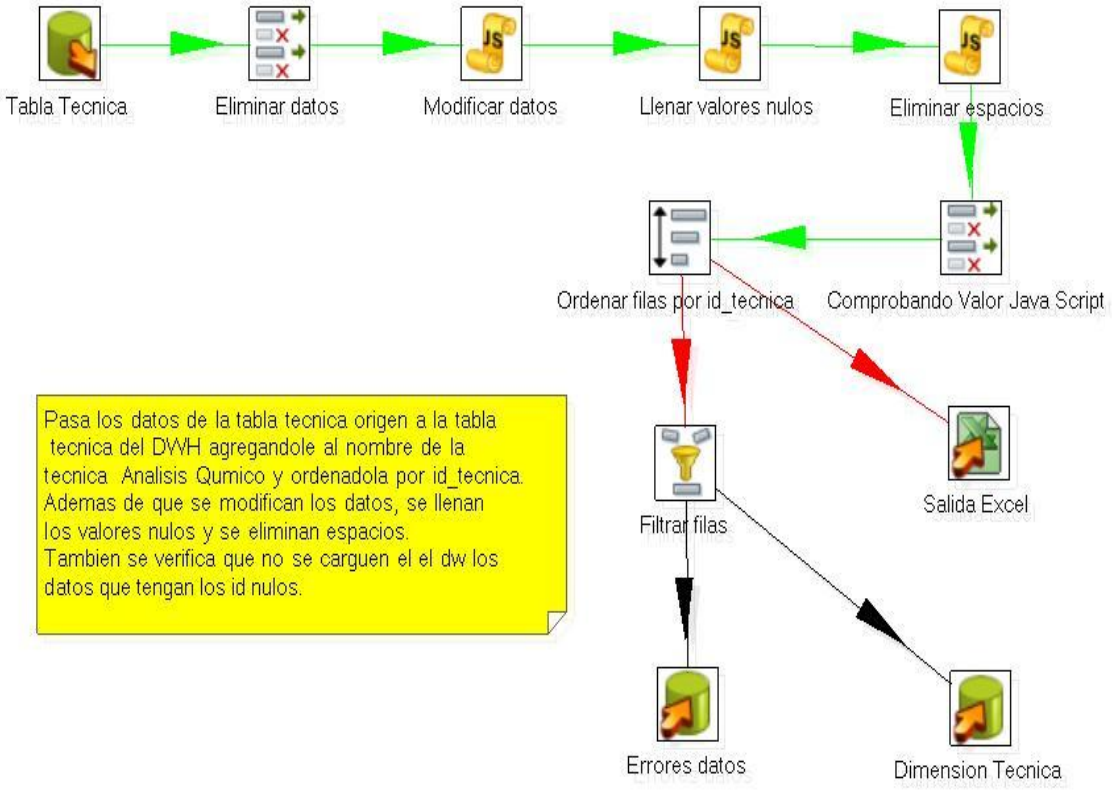


Figura 17: Proceso ETL para la tabla Técnica.

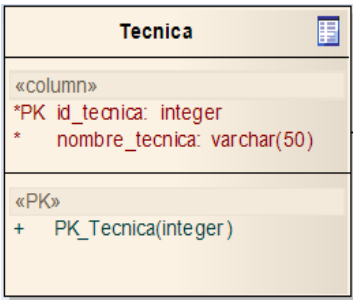


Figura 18: Dimensión Técnica del DW

6. Curva de Calibración.

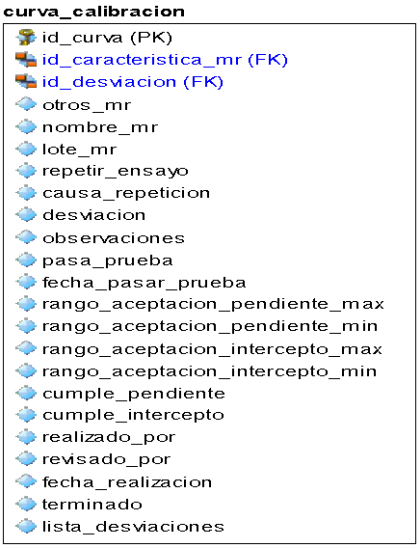


Figura 19: Tabla curva_calibración de la BD del Módulo Análisis Químico del proyecto Lims Control de la Calidad.

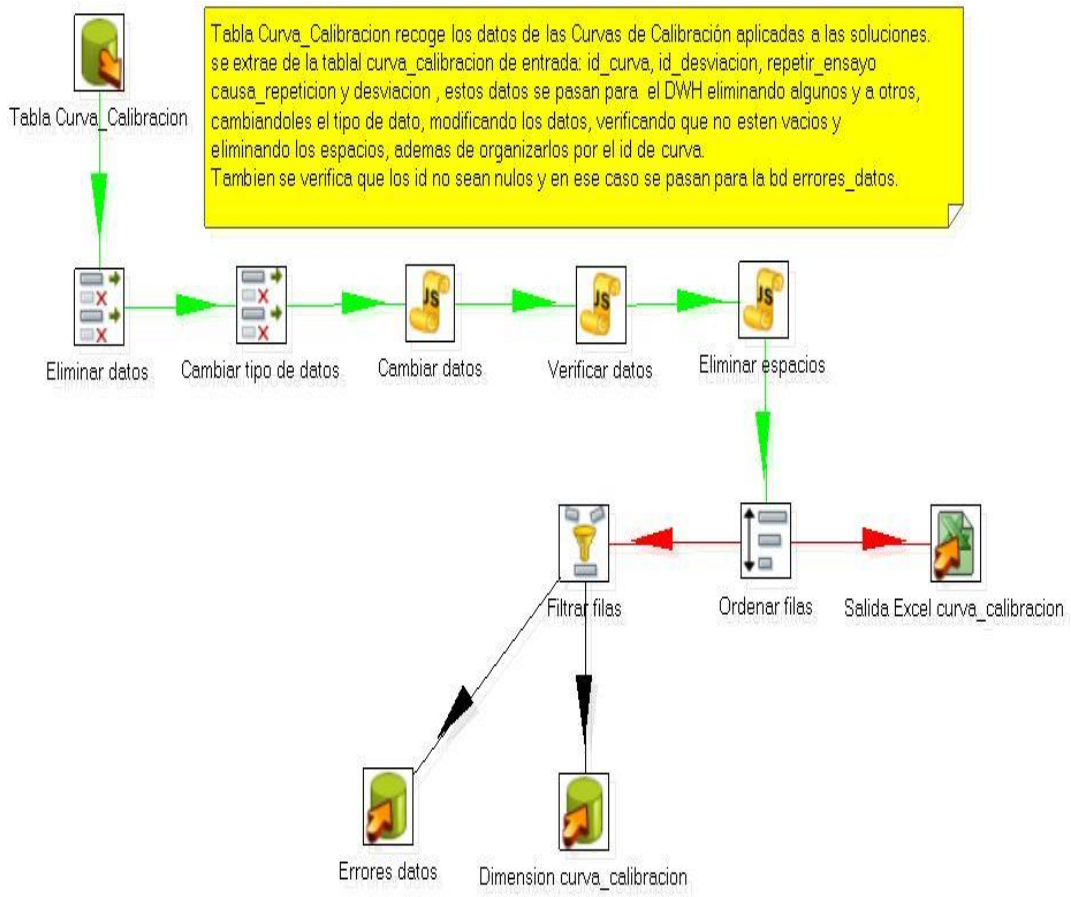


Figura 20: Proceso ETL para la tabla Curva de Calibración.

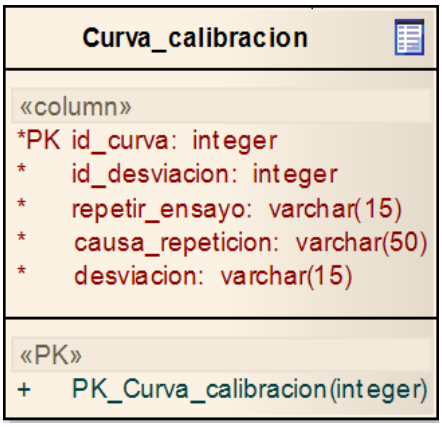


Figura 21: Dimensión Curva_calibracion del DW

7. **Libro de Registro del Producto:** Tabla dimensión que recoge todos los datos del producto.

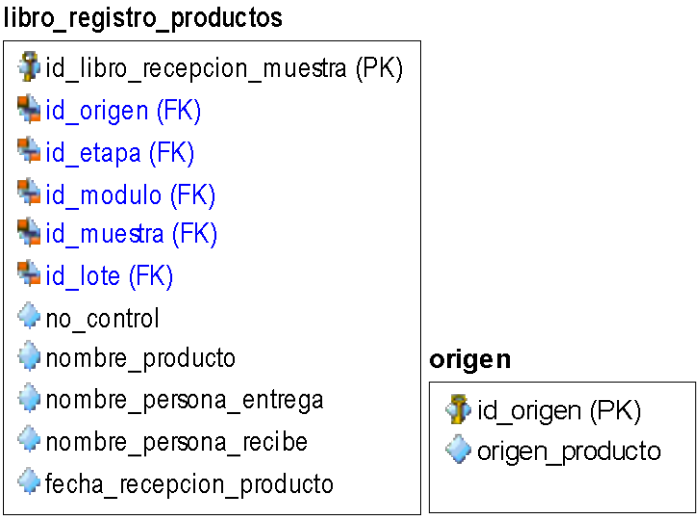


Figura 22: Tablas libro_registro-productos y origen de la BD del Módulo Análisis Químico del proyecto Lims Control de la Calidad.

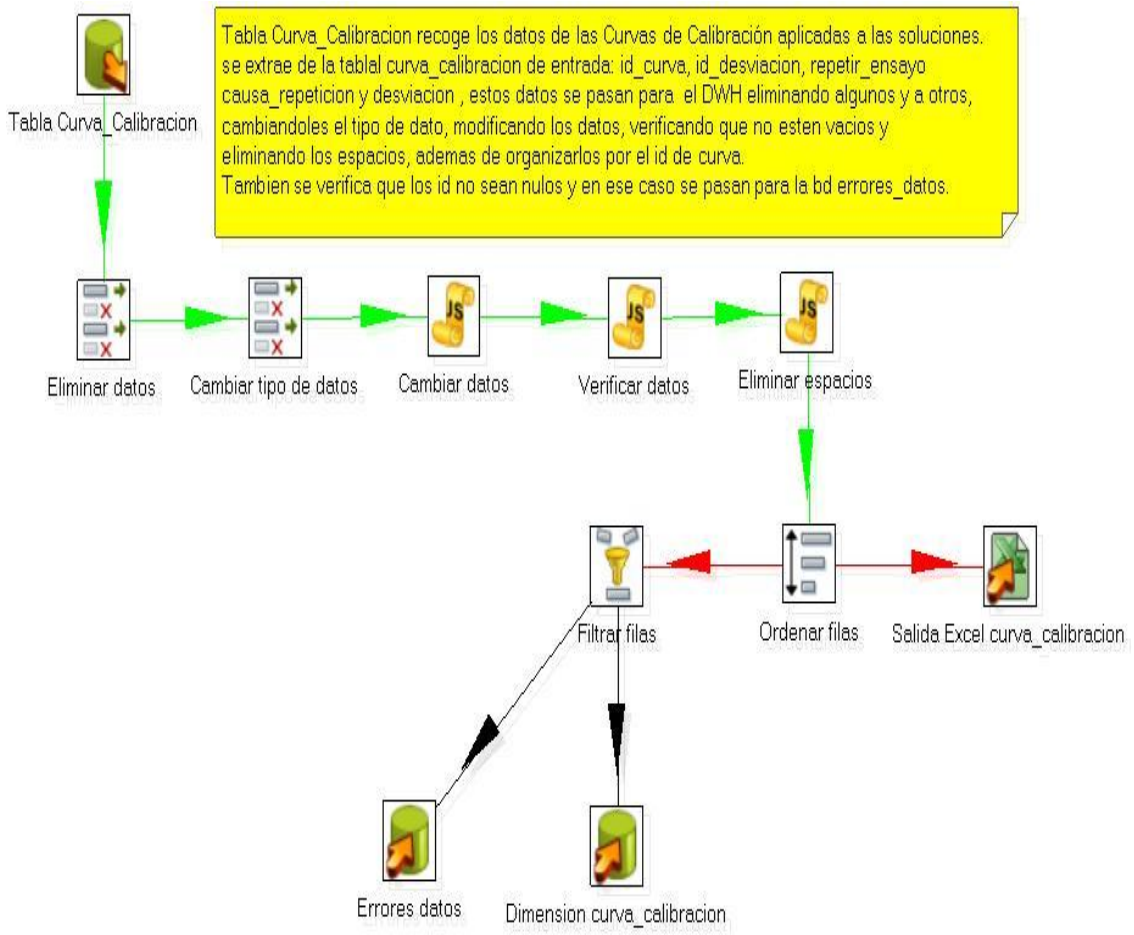


Figura 23: Proceso ETL para la tabla Libro de Registro del Producto.

Para realizar la carga de la tabla Libro_registro_producto se extraen los datos de las tablas de la figura 22 quedando solo en el DW los datos que se muestran a continuación.

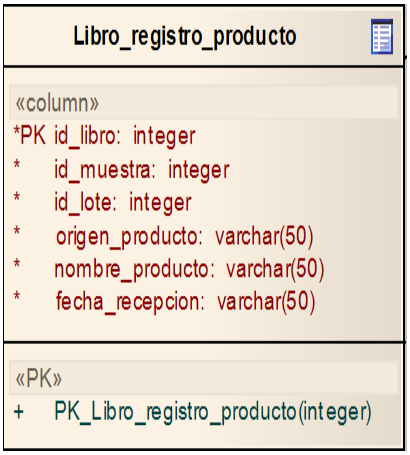


Figura 24: Dimensión Libro_registro_producto del DW

8. Registro de Preparación de Soluciones.

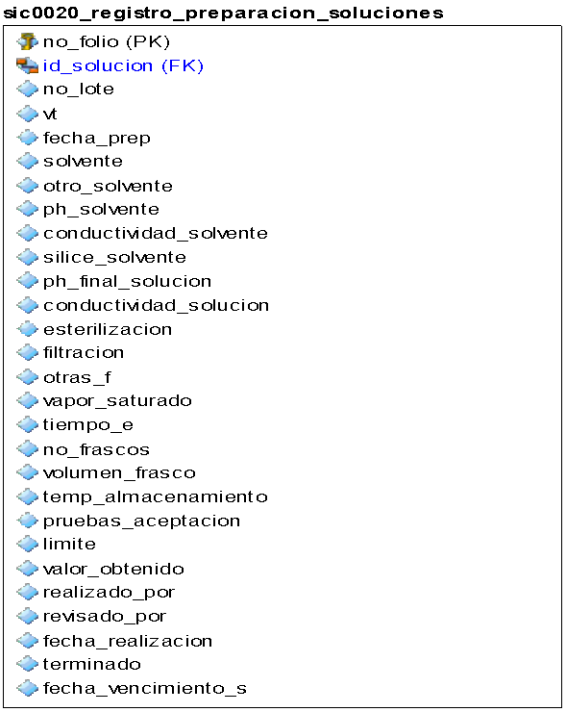


Figura 25:Sic0020_registro_preparacion_soluciones de la BD del Módulo Análisis Químico del proyecto Lims Control de la Calidad.

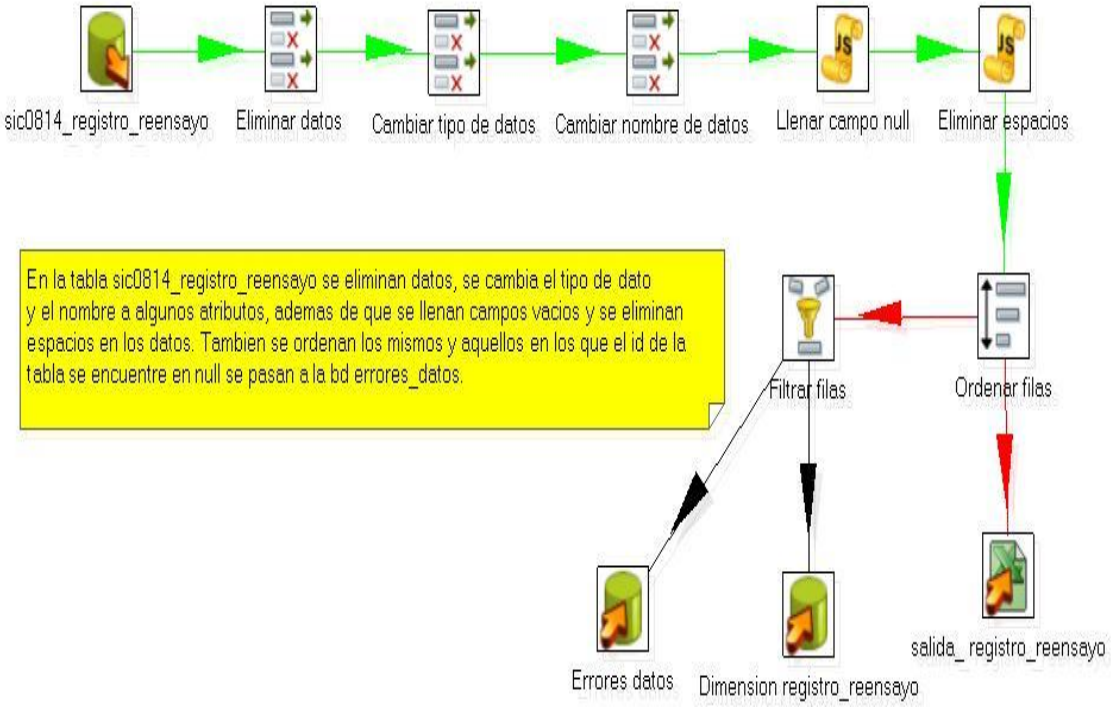


Figura 26: Proceso ETL para la tabla Registro de Preparación de Soluciones.

Para cargar la siguiente tabla se extraen los datos y se le cambia el nombre y el tipo de dato de algunos atributos de la tabla origen también se eliminan otros que no serán utilizados en la tabla destino Registro_preparacion_solucion en el DW.

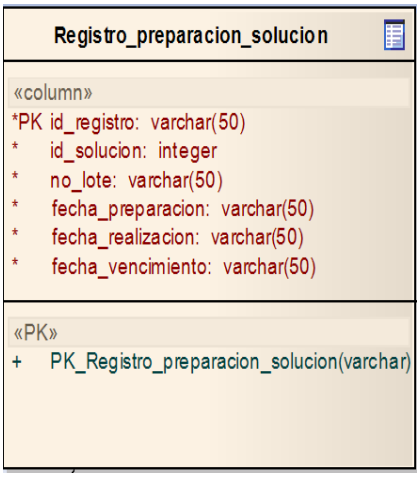


Figura 27: Dimensión Registro_preparacion_solucion del DW

9. **Planilla de Resultado:** Tabla dimensión que recoge los datos de las Planillas de Resultados

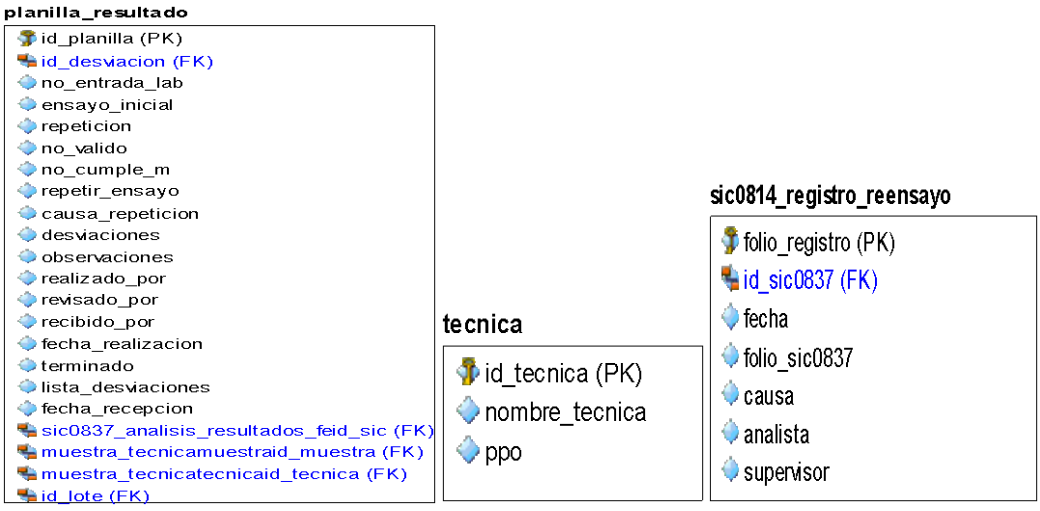


Figura 28: Tabla planilla_resultado, técnica y sic0814_registro_reensayo de la BD del Módulo Análisis Químico del proyecto Lims Control de la Calidad.

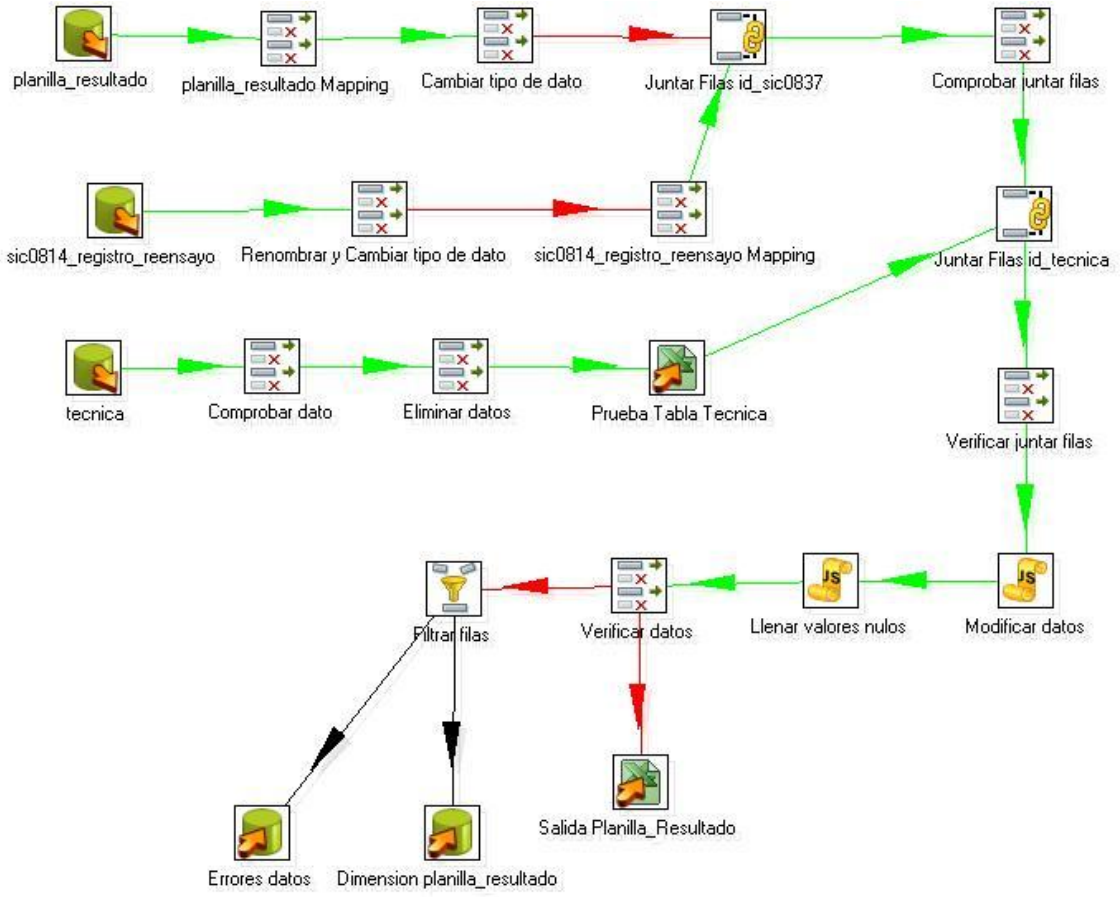


Figura 29: Proceso ETL para la tabla Planilla de Resultados.

Para realizar la carga de la tabla Planilla_resultado se extraen los datos de las tablas planilla_resultado, técnica y sic0814_registro_reensayo quedando solo los datos que se muestran a continuación.

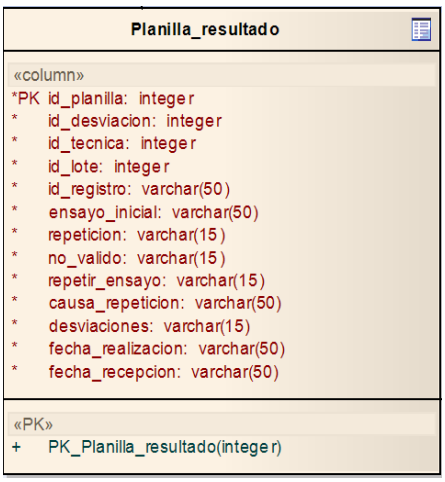


Figura 30: Dimensión Planilla_resultado del DW.

10. Registro Reensayo.

Figura 31: Tabla sic0814_registro_reensayo de la BD del Módulo Análisis Químico del proyecto Lims Control de la Calidad.

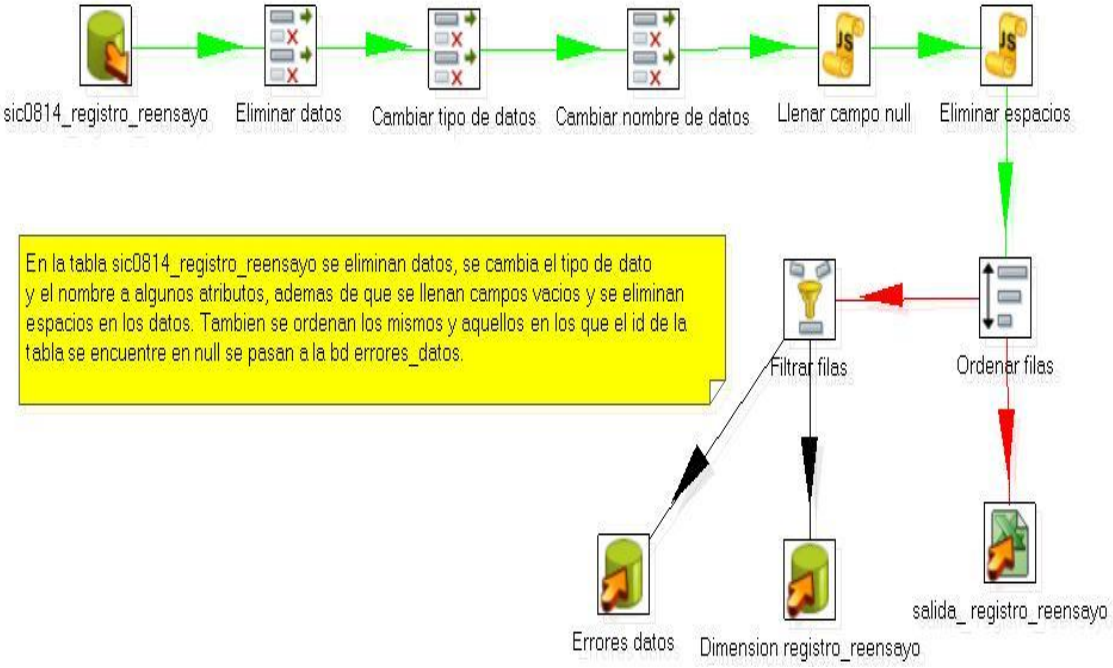


Figura 32: Proceso ETL para la tabla Registro de Reensayo.

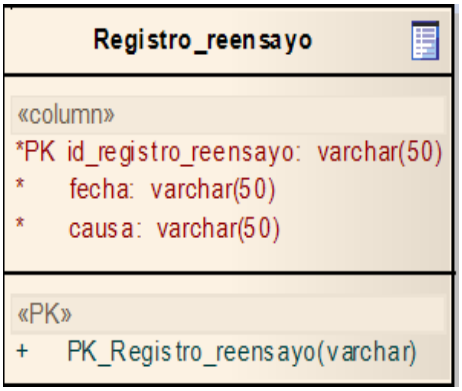


Figura 33: Dimensión Registro_reensayo del DW.

11. Desviación

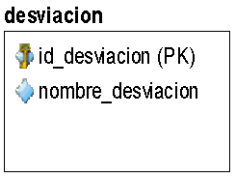


Figura 34: Tabla desviación de la BD del Módulo Análisis Químico del proyecto Lims Control de la Calidad.

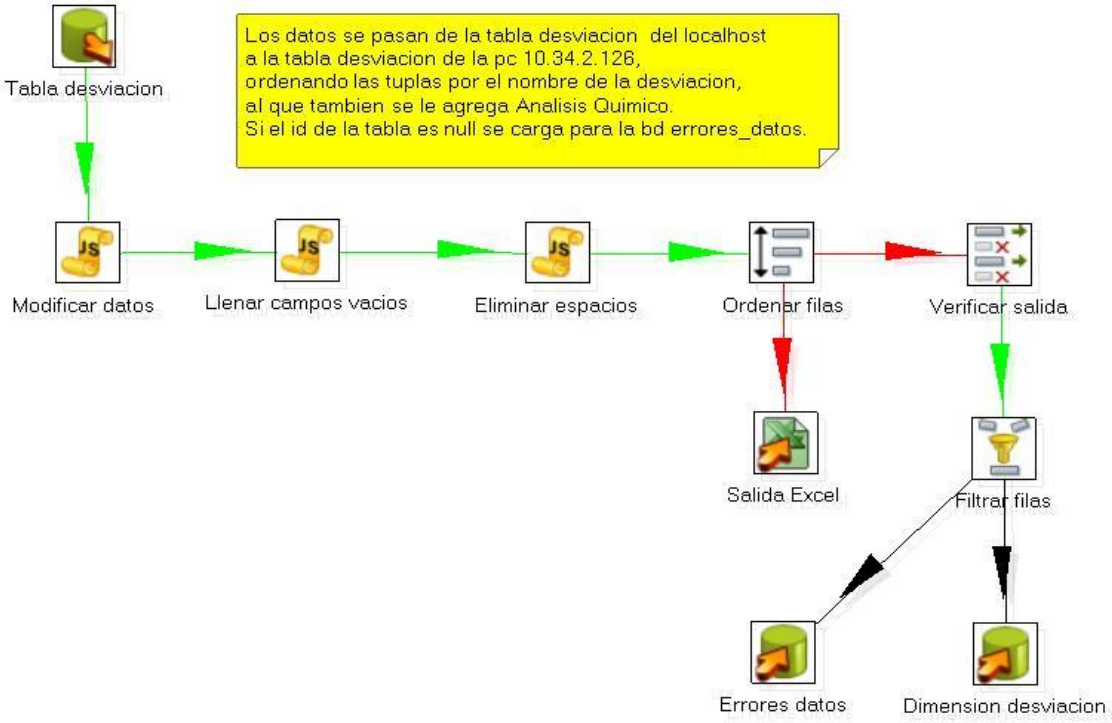


Figura 35: Proceso ETL para la tabla Desviación.

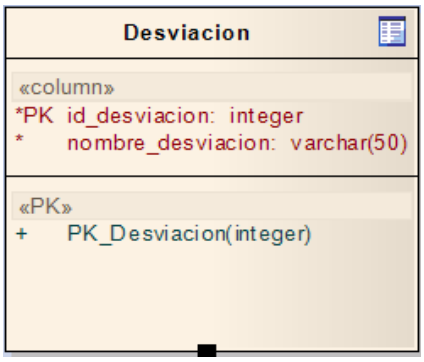


Figura 36: Dimensión Desviación del DW.

12. Tabla tiempo.

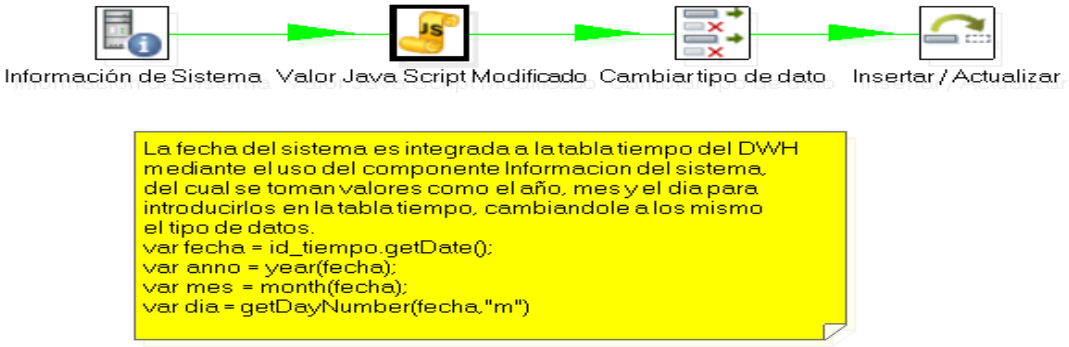


Figura 37: Proceso ETL para la tabla Tiempo.

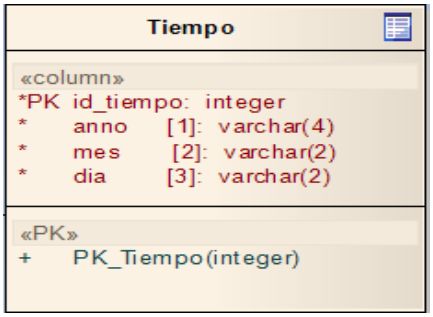


Figura 38: Dimensión Tiempo del DW.

13. Análisis Químico es la tabla hecho del DW que esta diseñado en forma de estrella. La integridad referencial es llevada a cabo por la creación de llaves foráneas en la tabla Hecho, que a su vez forman parte de la llave principal de la esta tabla. Es importante destacar que las jerarquías completas son guardadas en una sola tabla dimensión.

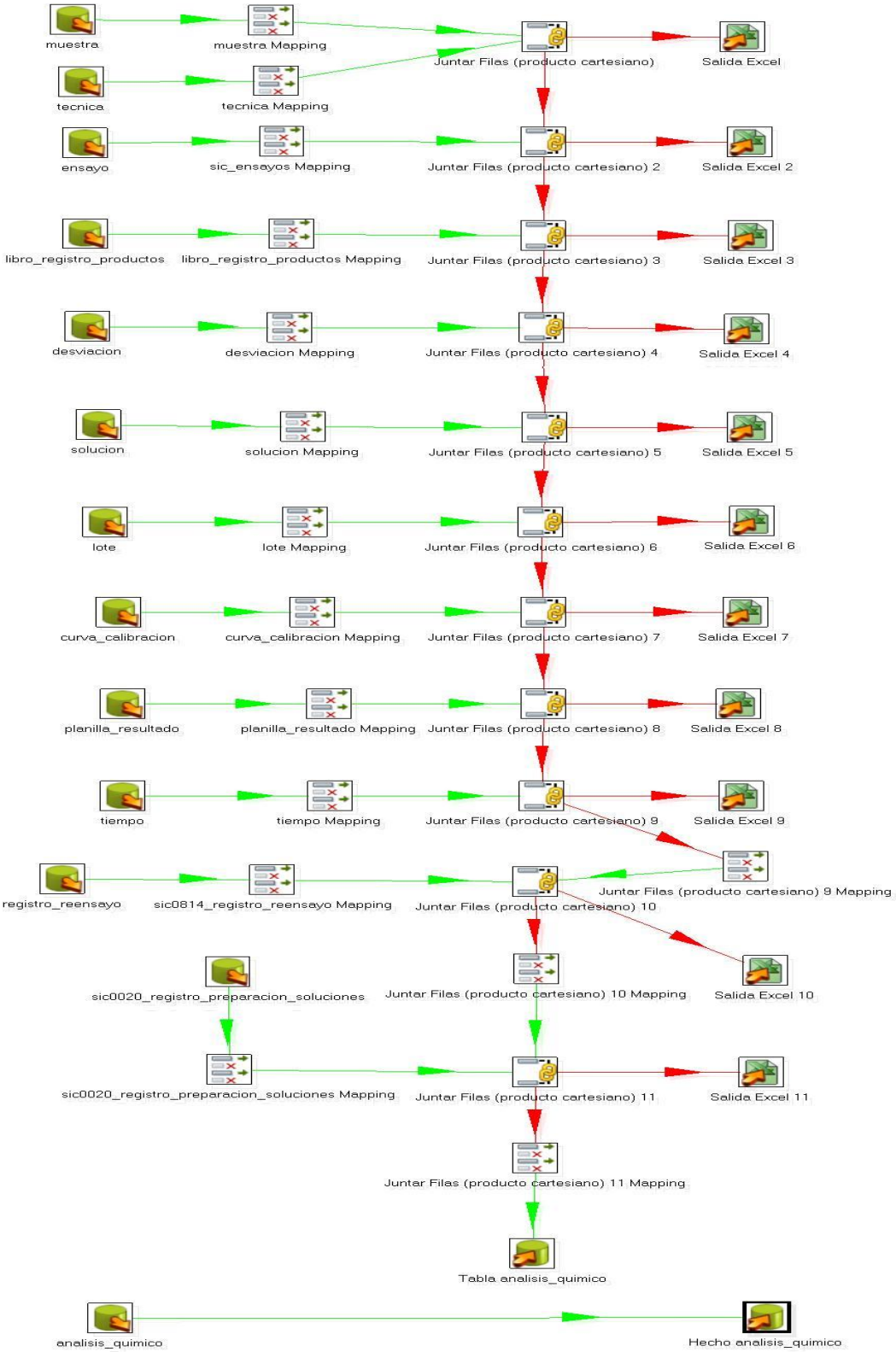
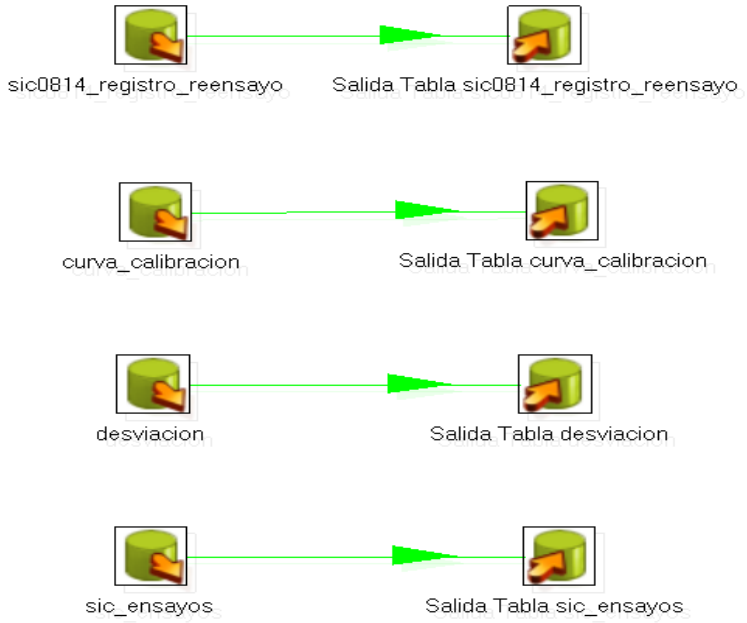


Figura 39: Proceso ETL para la tabla Análisis Químico.

Figura 40: Análisis Químico: Tabla hecho del DW.

ANEXO D



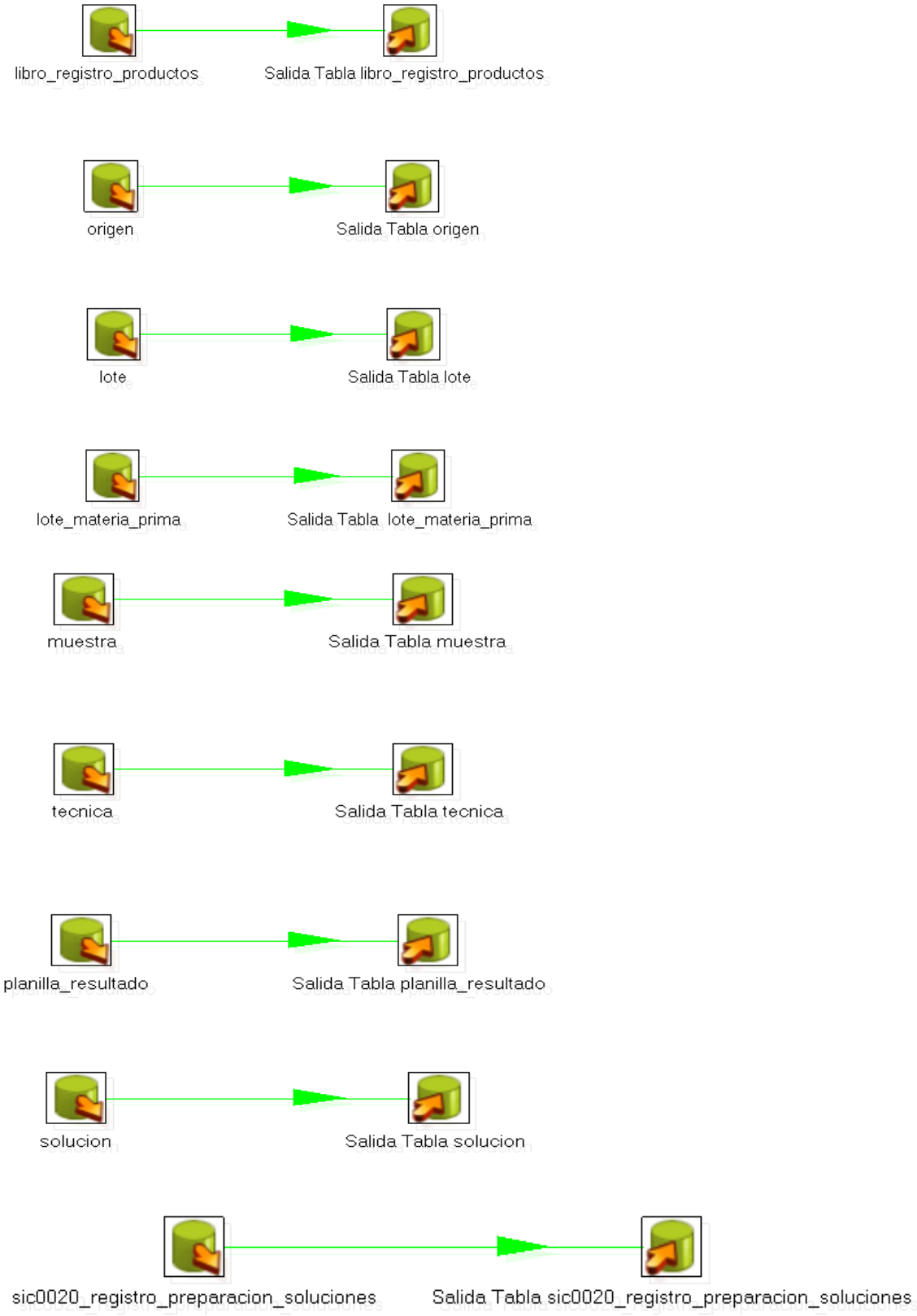


Figura 41: Transformación Almacenamiento Intermedio.

Glosario de Términos

OLTP: Procesamiento de Transacciones En Línea (del inglés de **O**n**L**ine **T**ransaction **P**rocessing).

ETL: Extracción, Transformación y Carga de Datos.

DW: Data Warehouse. Almacén de datos que reúne la información histórica generada por todos los distintos departamentos de una organización, orientada a consultas complejas y de alto rendimiento.

CIGB: Centro de Ingeniería Genética y Biotecnología.

TIC: Las tecnologías de la información y la comunicación.

LIMS: Sistemas de Gestión de Información de Laboratorios (del inglés Laboratory Information Management System).

BPP: Buenas Prácticas de Producción.

BPC: Buenas Prácticas Clínicas.

BPL: Buenas Prácticas de Laboratorio.

UCI: Universidad de las Ciencias Informáticas.

DSS: Sistemas de Soporte a Decisiones.

BI: Inteligencia de Negocios (del inglés Business Intelligence). Conjunto de metodologías y tecnologías orientadas a potenciar la gestión inteligente de la empresa que permitan a los equipos directivos controlar los negocios.

SGBD: Sistema de Gestión de Base de Datos.

BD: Base de Datos.

Disparadores*:* *triggers.*

OLAP: Online Analitical Processing, análisis on-line de información, análisis de datos con los que se trabaja día a día.

ROLAP (*Relational* OLAP): los datos son almacenados y recuperados de una base de datos relacional.

Metadatos: datos acerca de datos que describen los contenidos del almacén de datos.
DM: Data Mart (es una versión especial de almacén de datos (Data Warehouse). Son subconjuntos de datos con el propósito de ayudar a que un área específica dentro del negocio pueda tomar mejores decisiones)
JDBC: Conectividad de la base de datos de Java (Java Database Connectivity)
ODBC: Estándar de acceso a bases de datos que utilizan los sistemas Microsoft (Open DataBase Connectivity)