UNIVERSIDAD DE LAS CIENCIAS INFORMÁTICAS

Facultad # 6

TÍTULO: "REALIZAR EL DISEÑO E IMPLEMENTACIÓN DE UN DATA WAREHOUSE PARA EL PROYECTO LIMS CONTROL DE CALIDAD"

TRABAJO DE DIPLOMA PARA OBTAR POR EL TÍTULO INGENIERO EN CIENCIAS INFORMÁTICAS

Autores:

Milagros de la Caridad Suárez Giró

José Raúl González Sardina

Tutores:

Msc. Maypher Román Durán

Ing. Alfredo Rodríguez Ruíz

Junio de 2009

DECLARACIÓN DE AUTORÍA

Declaramos que somos los únicos autores de este trabajo y autorizamos a la Universidad de las Ciencias Informáticas a hacer uso del mismo en su beneficio.

Para que así conste firmamos la presente a los :

____ días del mes de _____ del año _____.

Autora: Milagros de la Caridad Suárez Giró. Autor: José Raúl González Sardina.

_____ _____

Tutor: Msc. Maypher Román Durán

_____

Tutor: Ing. Alfredo Rodríguez Ruíz

_____

...si supiese qué es lo que estoy haciendo,

no lo llamaría INVESTIGACIÓN...

Albert Einstein

Agradecer en primer lugar a esta universidad que sin ella no hubiéramos sido lo que hoy en día somos.

A nuestro tutor Maypher que nos ha ayudado en cada momento de la preparación de esta tesis.

A todos lo que de alguna forma han ayudado a la elaboración de esta tesis, a los profesores Roberto Acosta González y Alejandro Catalá Aguirre de la Facultad 7, así como, el estudiante Vladimir Urquía de la misma Facultad que han sido de gran ayuda.

A nuestra Revolución por darnos la oportunidad de estudiar en la UCI.

A nuestros padres por darnos todo su apoyo para que nuestro sueño se hiciera realidad.

A nuestros compañeros que nos han apoyado en todo el transcurso de nuestra carrera y al dúo ETL que hemos trabajado juntos para tener lo que hoy tenemos.

A mi padre por ser aquella persona que me dio todo su apoyo para que yo pudiera transitar por esta universidad, por estar ahí cuando lo necesite y por ser quien soy hoy en día.

A mi abuela por ser la razón de mí existir y darme todo su empeño para que alcanzara mi meta.

A todos los que nos han brindado su apoyo incondicionalmente.

A todas las personas que de una forma u otra han estado a mi lado por el paso de la UCI.

Milagros.

A mi mamá, a mi papá por ser los verdaderos artífices de todo logro que pueda alcanzar.

A mis hermanas por servir de apoyo y soporte.

A todos los que están y los que no, que de alguna manera me llevaron a ser lo que ahora soy.

José Raúl.

RESUMEN

Con el desarrollo de la informatización a nivel mundial ha aumentado la capacidad de generación y almacenamiento de la información, lo cual ha llevado a aplicar nuevos métodos para almacenar toda la información que aumenta cada día más, pues por los métodos tradicionales surge la incapacidad de extraer la información necesaria y útil para el desarrollo en cada empresa así como para la toma de decisiones en cada una de estas. Es así como surge Data Warehouse (DW). Dado el gran volumen de información que se maneja en el proyecto LIMS Control de Calidad, este se ha dado a la tarea de implementar un Data Warehouse para dar una visión más amplia y generar un expediente completo, teniendo en cuenta que los datos no se modifican ni destruyen sino que son utilizados para futuras consultas que brinden el estado del comportamiento del proyecto, convirtiéndose éste en una poderosa herramienta especializada en el manejo de datos históricos.

PALABRAS CLAVES

- Data Warehouse
- Almacenes de Datos
- Proceso Analítico en Línea
- Hefesto
- Mondrian
- Sistemas Operacionales

ÍNDICE

INTRODUCCIÓN

Con el gran desarrollo de la Informatización, crecen las ventajas para quienes tienen acceso a grandes volúmenes de datos, pero a su vez esto trae un gran problema ¿cómo manejar de manera exitosa estos volúmenes?.

Hoy en día existen diversos tipos de sistemas de soporte para la toma de decisiones, pero el que ha tenido más auge a escala mundial en las grandes instituciones sin duda ha sido el DW o almacenes de datos, convirtiéndose en el centro de atención de las organizaciones, puesto que provee un ambiente para hacer un mejor uso de la información administrada por diversas aplicaciones operacionales.

Para solucionar dicha problemática surge el almacén de datos, que en la rama de la informática, no es más que una colección de datos orientada a un determinado ámbito (empresa, organización, etc.), integrado, no volátil y variable en el tiempo, que ayuda a la toma de decisiones en la entidad en la que se utiliza. Se trata, sobre todo, de un expediente completo de una organización, más allá de la información transaccional y operacional, almacenada en una base de datos diseñada para favorecer el análisis y la divulgación eficiente de datos.

En el caso de las empresas cubanas la introducción de tecnologías de la información y las comunicaciones en el manejo del conocimiento tiene como objetivo fundamental el uso más racional de los recursos, el logro de una mayor productividad y la obtención de los productos con una mayor calidad. Aunque no se cuenta con gran experiencia en el uso de los Data Warehouse si existen empresas que han tenido experiencias en su uso principalmente en el turismo.

En la Universidad de las Ciencias Informáticas existe mucha experiencia en el uso e implementación de un Data Warehouse, en la Facultad 2 específicamente en el proyecto SINSEC , también existen 4 Data Mart para la gestión del conocimiento y en la Facultad 7 están desarrollando una tesis con el objetivo de implementar un Data Warehouse para el control de Recursos Humanos de la Salud. Dado el gran volumen de información que se tiene almacenado en la Base de Datos del Proyecto LIMS Control de Calidad perteneciente a la Facultad 6 surge la necesidad de desarrollar un Data Warehouse para evaluar y planificar el comportamiento y alcance del proyecto.

En Cuba se ha llevado a cabo un intenso proceso inversionista y de formación de personal, que permite disponer en la actualidad de un complejo e integrado sistema de investigación-producción en la esfera de la Biotecnología aplicada a diferentes ramas de la sociedad.

Con la inversión intensiva y el personal que entrena han sido desarrollados actualmente procesos para conceder el acceso a un amplio sistema de producción e investigación complejo e integrado en el campo de Biotecnología aplicada a las ramas (sucursales) diferentes de sociedad. En este contexto, el Centro Cubano para la Ingeniería Genética y la Biotecnología, es una institución de un desarrollo dinámico que ha alcanzado un nivel alto de investigación, desarrollo, producción y comercialización de productos biológicos obtenidos por los métodos de biotecnología moderna.

El Centro para la Ingeniería Genética y la Biotecnología (CIGB) juega un papel integrante en el campo de Biotecnología cubana, con altas capacidades técnicas. Esto también asume la responsabilidad de directamente contribuir al desarrollo económico y social del país.

Su funcionamiento es proyectado hacia la investigación para generar el conocimiento y desarrollo de nuevos productos, servicios y actividades comerciales, basadas en un sistema de calidad que asegura la satisfacción de sus clientes que tienen en cuenta la dimensión ambiental. Su impacto está destinado a la salud humana, la producción agrícola, la acuicultura y el ambiente.

Misión del CIGB.

Es totalmente la conciencia sobre los riesgos de Ingeniería genética para el ser humano y el Ambiente si no es usado con la responsabilidad. Son confidentes del empleo de Ingeniería genética como un camino revolucionario para producir los nuevos productos que ayudarían a solucionar los problemas corrientes científicos y tecnológicos del género humano. Su trabajo principalmente es enfocado (concentrado) para alcanzar este objetivo de un punto de vista responsable y ético, evaluando cada riesgo y después de todas las medidas de biosafety que son solicitadas según sus legislaciones.

Debido a la gran importancia que tiene el CIGB para al desarrollo del país y el gran volumen de información se le realiza al módulo (Análisis Químico) la propuesta de la implementación de un Data Warehouse para el manejo eficiente de los datos.

Problema Científico:

¿Cómo solucionar el creciente proceso de registro, procesamiento y análisis de información en el tiempo, para el proyecto LIMS Control de Calidad del Centro de Ingeniería Genética y Biotecnología, que permita a los directivos de dicho centro el análisis de la información para sus procesos de gestión interna?

Objeto de estudio:

Realización del Diseño e Implementación de un Data Warehouse en la Universidad de las Ciencias Informáticas.

Campo de acción:

Realización del Diseño e Implementación de un Data Warehouse para el proyecto LIMS Control de Calidad de la Facultad 6 de la Universidad de las Ciencias Informáticas.

Objetivo General:

Realizar el Diseño e Implementación de un Data Warehouse para el proyecto LIMS Control de Calidad.

Objetivos específicos:

- Realizar el Diseño de las tablas que conformen el Data Warehouse.
- Realizar la Implementación de las tablas que conformen el Data Warehouse.
- Implementar los procedimientos de visualización de los datos a los usuarios finales del Data Warehouse.

Tareas investigativas:

- Realización de estudios sobre las técnicas y prácticas empleadas en el desarrollo de un Data Warehouse.
- Estudio de posibles tecnologías a utilizar durante el proceso de desarrollo de un Data Warehouse.
- Validación de la propuesta.

Idea a defender:

Con la creación de Data Warehouse se mejorará el proceso de análisis de los datos transformándolos en información útil para usuarios del Centro de Ingeniería Genética y Biotecnología, que evalúen aspectos generales de los ensayos registrados y almacenados.

Hipótesis:

Si el proyecto LIMS Control de Calidad cuenta con un Data Warehouse para obtener estadísticas generales de las operaciones realizadas por el sistema, los usuarios de alto rango serán capaces de tomar mejores decisiones a nivel estratégico por contar con una herramienta que les permita reutilizar de manera eficiente la información almacenada.

CAPÍTULO 1 .FUNDAMENTACIÓN TEÓRICA

En este capítulo vamos a tratar todo lo relacionado al estado del arte, se abordará sobre la metodología utilizada, de las aplicaciones que han servido para la elaboración del DW, se exponen las características y los métodos más utilizados de los almacenes de datos, así como, sus cualidades, ventajas, desventajas, arquitectura del DW , así como, las herramientas de consultas y análisis .

1.1 CARACTERÍSTICAS DEL DATA WAREHOUSE

Varias han sido las definiciones acerca del término de un Data Warehouse:

Bill Inmon, fue uno de los primeros autores en escribir sobre el tema en términos de las características del almacén de datos:

Orientado a temas

Los datos en la base de datos están organizados de manera que todos los elementos de datos relativos al mismo evento u objeto del mundo real queden unidos entre sí.

Esta definición del DW clasifica la información en base a los aspectos que son de interés para la empresa. Dicha clasificación afecta el diseño y la implementación de los datos encontrados en el almacén de datos, debido a que la estructura del mismo difiere considerablemente a la de los clásicos procesos operacionales orientados a las aplicaciones.

En síntesis, la ventaja de contar con procesos orientados a la aplicación, está fundamentada en la alta accesibilidad de los datos, lo que implica un elevado desempeño y velocidad en la ejecución de consultas, ya que las mismas están predeterminadas; mientras que en el DW para satisfacer esta ventaja se requiere que la información este desnormalizada, es decir, con redundancia, duplicidad de los datos y que la misma esté dimensionada, para evitar tener que recorrer toda la base de datos cuando se necesite realizar algún análisis determinado, sino que simplemente la consulta sea enfocada por vectores y variables que permitan localizar los datos de manera rápida y eficaz, para poder de esta manera satisfacer una alta demanda de complejos exámenes en un mínimo tiempo de respuesta. [1]

Variante en el tiempo

Los cambios producidos en los datos a lo largo del tiempo quedan registrados para que los informes que se puedan generar reflejen esas variaciones.

Debido al gran volumen de información que se manejará en el DW, cuando se le realiza una consulta, los resultados deseados demorarán en originarse. Este espacio de tiempo que se produce desde la búsqueda de datos hasta su consecución es del todo normal en este ambiente y es, precisamente por ello, que la información que se encuentra dentro del depósito de datos se denomina de tiempo variable.

El intervalo de tiempo y periodicidad de los datos debe definirse de acuerdo a la necesidad y requisitos de los usuarios.

Es elemental aclarar, que el almacenamiento de datos históricos, es lo que permite al DW desarrollar pronósticos y análisis de tendencias y patrones, a partir de una base estadística de información, ya que las instantáneas son actualizadas de acuerdo con las actividades del negocio. [1]

No volátil

La información no se modifica ni se elimina, una vez almacenado un dato, éste se convierte en información de sólo lectura, y se mantiene para futuras consultas.

La información es útil para el análisis y la toma de decisiones solo cuando es estable. Los datos operacionales varían momento a momento, en cambio, los datos una vez que entran en el DW no cambian.

La actualización, o sea, insertar, eliminar y modificar, se hace de forma muy habitual en el ambiente operacional sobre una base, registro por registro, en cambio en el depósito de datos la manipulación básica de los datos es mucho más simple, debido a que solo existen dos tipos de operaciones: la carga de datos y el acceso a los mismos.

Por esta razón es que en el DW no se requieren mecanismos de control de la concurrencia y recuperación. [1]

Integrado

La base de datos contiene los datos de todos los sistemas operacionales de la organización, y dichos datos deben ser consistentes.

La integración implica que todos los datos de diversas fuentes que son producidos por distintos departamentos, secciones y aplicaciones, tanto internos como externos, deben ser consolidados en una instancia antes de ser agregados al DW. A este proceso se lo conoce como Extracción, Transformación y Carga de Datos (ETL). [1]

Inmon defiende una metodología descendente (top-down) a la hora de diseñar un almacén de datos, ya que de esta forma se considerarán mejor todos los datos corporativos.

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

Las definiciones anteriores se centran en los datos en sí mismos. Sin embargo, los medios para obtener y analizar esos datos, para extraerlos, transformarlos y cargarlos, así como las diferentes formas para realizar la gestión de datos son componentes esenciales de un almacén de datos.

Muchas referencias a un almacén de datos utilizan esta definición más amplia. Por lo tanto, en esta definición se incluyen herramientas para la inteligencia empresarial, herramientas para extraer, transformar y cargar datos (ETL) en el almacén de datos, y herramientas para gestionar y recuperar los metadatos.

1.2 MÉTODOS MÁS USADOS EN LA CONSTRUCCIÓN DE DATA WAREHOUSE

Los modelos propuestos por William H. Inmon y Ralph Kimball para llevar a cabo el diseño de un Data Warehouse son los más aplicados en la actualidad, coincidiendo en que un Data Mart o un almacén de datos independiente no satisface las necesidades que tienen las compañías a escala corporativa de acceder inmediatamente y con facilidad a sus datos, pero sus criterios difieren en cuanto al modelo de datos y a las arquitecturas.

El término Data Mart es usado para designar a los almacenes de datos cuyo ámbito es más reducido, normalmente un departamento o área específica dentro de la empresa, es definido por Ralph Kimball como bodegas de datos con información de interés particular para un determinado sector de la empresa y aunque su enfoque sea para una sola perspectiva departamental, no lo exime de tener que seguir los lineamientos generales de implementación que posee el Data Warehouse (KIMBALL 1996).

Ralph Kimball propone como modelo de datos al modelo dimensional, el más popular en las soluciones que se implementan de manera práctica, el cual facilita a los usuarios finales las consultas y el análisis. Se caracteriza por ser sencillo de crear, extremadamente estable en presencia de cambios, además de mostrarse muy intuitivo y comprensible; el autor sugiere el uso de este modelo de datos para el desarrollo de los Data Marts y del Data Warehouse (KIMBALL and CASERTA 2004).

También William H. Inmon reconoce al modelo dimensional como el mejor para el desarrollo de los Data Marts por las ventajas brindadas, pero propone la construcción del Data Warehouse basado en el modelo entidad relación. La idea de Inmon se basa en que el modelo entidad relación es mucho más rico y adaptable que el dimensional (INMON, WILLIAM 2002).

En cuanto a la arquitectura William H. Inmon en su libro “Building the Data Warehouse” (INMON, WILLIAN H. 2005) plantea que la construcción del Data Warehouse no debe ser sustituida por la implementación de varios Data Marts.

Resaltando que la excusa para no desarrollar un almacén de datos la mayoría de las veces es por no contar con un gran presupuesto, la sustitución de este por los Data Marts trae desventajas puesto que están diseñados para un área particular de la empresa, lo que trae consigo diferencias entre las estructuras de datos de los mismos, que al integrarlos en el Data Warehouse algunos no serán reusables, ni flexibles, ni útiles para la reconciliación que se necesita. Inmon manifiesta que el proceso de construcción del Data Warehouse parte de los sistemas operacionales existentes, creándose áreas de diferentes temas, cuando existan una cierta cantidad de estas, el Data Warehouse inicia el proceso de población de las áreas de una manera integrada, una vez concluido se comienza a dar respuestas a las inquietudes de los usuarios; empezando así el florecimiento del nivel departamental a medida que se tienen más datos en el Data Warehouse y es en este punto del desarrollo cuando se centra la atención en las cuestiones de los diferentes departamentos, para definir y crear los almacenes de datos departamentales, los Data Marts.

Ralph Kimball en desacuerdo con la arquitectura propuesta por William H. Inmon resalta en su libro “The Data Warehouse ETL Toolkit, Practical Techniques for Extracting, Cleaning, Conforming and Delivering Data” (KIMBALL and CASERTA 2004), que los Data Marts están basados en los datos de la fuente y no en la visión departamental, en otras palabras que el Data Mart es sólo una parte de un producto orientado a la compañía, los cuales deben consistir en una continua pirámide de estructuras dimensionales idénticas, comenzando siempre con los datos atómicos. Plantea también que la idea de construcción de un Data Warehouse centralizado no es realista, siendo más real construirlo en un

ambiente descentralizado e incremental, porque las empresas están en constante cambio, adquiriendo nuevas fuentes de datos y necesitando nuevas perspectivas, propone además, centrarse en trazar estrategias adaptables e incrementales basándose en una idealista visión de controlar toda la información antes de construir el Data Warehouse. Por esta razón manifiesta que el proceso de construcción de un almacén de datos parte de los sistemas operacionales existentes, creando los diferentes Data Marts basados en la información de dichas fuentes, para luego de tenerlos desarrollados y funcionales se comience con la construcción del Data Warehouse basado en la información que éstos contienen.

En la actualidad este método es el más usado, gracias a las diferentes ventajas que proporciona, permitiendo a las empresas acometer los proyectos de manera separada y de esta forma reducir los efectos negativos que tendría fracasar en un intento por construir un Data Warehouse.

1.3 CUALIDADES DEL DATA WAREHOUSE

Una de las primeras cualidades que se puede mencionar del DW, es que maneja un gran volumen de datos, debido a que consolida en su estructura la información recolectada durante años, proveniente de diversas fuentes, en un solo lugar centralizado. Es por esta razón que el depósito puede ser soportado y mantenido sobre diversos medios de almacenamiento.

Además, el almacén de datos presenta la información sumarizada y agregada desde múltiples versiones, y maneja información histórica. Organiza y almacena los datos que se necesitan para el procesamiento analítico e informático, con el propósito de responder a preguntas de negocios y brindarles a los usuarios finales una interface amigable, comprensible y fácil de utilizar, para que los mismos puedan tomar decisiones sobre los datos sin tener que poseer demasiados conocimientos informáticos. El DW permite un acceso más directo, es decir, la información gira en torno al negocio, y es por ello que también los usuarios pueden sentirse cómodos al explorar los datos y encontrar relaciones complejas entre los mismos.

El DW no es solo datos, sino un conjunto de herramientas para consultar, analizar y presentar información, que permiten obtener o realizar análisis, reportes, extracción y explotación de los datos, con alta performance, para transformar dichos datos en información valiosa para la organización.

Con respecto a las tecnologías empleadas, en un almacén de datos se pueden encontrar las siguientes:

- Arquitectura cliente/servidor.
- Técnicas avanzadas para replicar, refrescar y actualizar datos.
- Software front-end, para acceso y análisis de datos.
- Herramientas para extraer, transformar y cargar datos en el depósito, desde múltiples fuentes muy heterogéneas.
- Sistema de Gestión de Base de Datos (SGBD).

Cabe destacar, que todas las cualidades expuestas anteriormente, son imposibles de saldar en un típico ambiente operacional, y esto es una de las razones de ser del DW. [1]

1.4 VENTAJAS DEL DATA WAREHOUSE

- Transforma datos orientados a las aplicaciones en información orientada a la toma de decisiones.
- Integra y consolida diferentes fuentes de datos y departamentos empresariales, que anteriormente formaban islas, en una única plataforma sólida y centralizada.
- Provee la capacidad de analizar y explotar las diferentes áreas de trabajo y de realizar un análisis inmediato de las mismas.
- Permite reaccionar rápidamente a los cambios del mercado.
- Aumenta la competitividad en el mercado.
- Elimina la producción y el procesamiento de datos que no son utilizados ni necesarios, producto de aplicaciones mal diseñadas o ya no utilizadas.
- Mejora la entrega de información, es decir, información completa, correcta, consistente, oportuna y accesible. Información que los usuarios necesitan, en el momento adecuado y en el formato apropiado.
- Logra un impacto positivo sobre los procesos empresariales. Cuando los usuarios tienen acceso a una mejor calidad de información, la empresa puede lograr por sí misma: aprovechar el enorme valor potencial de sus recursos de información y transformarlo en valor verdadero; eliminar los retardos de los procesos empresariales que resultan de información incorrecta, inconsistente y/o inexistente; integrar y optimizar procesos a través del uso compartido e integrado de las fuentes de información; permitir al usuario adquirir mayor confianza acerca de sus propias decisiones y de las del resto, y lograr así, un mayor entendimiento de los impactos ocasionados.
- Aumento de la competitividad de los encargados de tomar decisiones.

- Los usuarios pueden acceder directamente a la información en línea, lo que contribuye a su capacidad para operar con mayor efectividad en las tareas rutinarias o no. Además, pueden tener a su disposición una gran cantidad de valiosa información multidimensional, presentada coherentemente como fuente única, confiable y disponible en sus estaciones de trabajo. Así mismo, los usuarios tienen la facilidad de contar con herramientas que les son familiares para manipular y evaluar la información obtenida en el DW, tales como: hojas de cálculo, procesadores de texto, software de análisis de datos, software de análisis estadístico, reportes, etc.
- Permite la toma de decisiones estratégicas y tácticas. [1]

1.5 DESVENTAJAS DEL DATA WAREHOUSE

- Requiere una gran inversión, debido a que su correcta construcción no es tarea sencilla y consume muchos recursos, además, su misma implementación implica desde la adquisición de herramientas de consulta y análisis, hasta la capacitación de los usuarios.
- Existe resistencia al cambio por parte de los usuarios.
- Los beneficios del almacén de datos son apreciados en el mediano y largo plazo. Este punto deriva del anterior, y básicamente se refiere a que no todos los usuarios confiarán en el DW en una primera instancia, pero sí lo harán una vez que comprueben su efectividad y ventajas. Además, su correcta utilización surge de la propia experiencia.
- Si se incluyen datos propios y confidenciales de clientes, proveedores, etc, el depósito de datos atentará contra la privacidad de los mismos, ya que cualquier usuario podrá tener acceso a ellos.
- Infravaloración de los recursos necesarios para la captura, carga y almacenamiento de los datos.
- Infravaloración del esfuerzo necesario para su diseño y creación.
- Incremento continuo de los requerimientos del usuario. [1]

El almacenamiento de los datos no debe usarse con datos de uso actual. Los almacenes de datos contienen a menudo grandes cantidades de información que se subdividen a veces en unidades lógicas más pequeñas dependiendo del subsistema de la entidad del que procedan o para el que sea necesario. Los Data Warehouse surgen con la promesa del manejo y control de la información. Ellos

aseguran una vista única de los datos, que pueden provenir de diversas fuentes. Gracias a esto, los usuarios finales no se ven en la necesidad de aprender y utilizar múltiples sistemas de acceso y manipulación de los datos. Un almacén de datos facilita la comprensión de los datos, transformándolos en información útil, teniendo como bandera el apoyo a la Toma de Decisiones. Permite no solo comprender lo que está pasando, sino predecir lo que va a suceder.

El Objetivo del Data Warehouse es integrar datos corporativos, residentes en bases de datos operacionales de la organización, en un único repositorio sobre el cual los usuarios puedan realizar consultas o informes y hacer análisis de datos.

La tecnología de almacenes de datos integra las técnicas de bases de datos y las técnicas de análisis de datos.

1.6 METODOLOGÍA HEFESTO

HEFESTO es una metodología propia, cuya propuesta está fundamentada 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 Data Warehouse 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 Data Warehouse y motivar a los usuarios.

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

Con el fin de que se llegue a una total comprensión de cada paso o etapa, se ha realizado una detallada explicación de todo el ciclo de vida del DW, reflejándose todos los resultados obtenidos en cada paso de la metodología utilizada.[1]

1.7 CARACTERÍSTICAS DE LA METODOLOGÍA HEFESTO

- 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 Data Warehouse.
- 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 Data Warehouse 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 Data Mart como para Data Warehouse. [1]

1.8 DESCRIPCIÓN DE LA METODOLOGÍA HEFESTO

Figura 1: Descripción de la Metodología Hefesto

1.9 APLICACIONES DEL DATA WAREHOUSE

Para el diseño e implementación del Data Warehouse se usaron las siguientes herramientas de modelado:

Herramientas de modelado

- Enterprise Architect 7.1
- Pentaho Schema Workbench
- Apache-tomcat-5.5.12
- Mondrian

Enterprise Architect 7.1

Enterprise Architect (EA) combina el poder de la última especificación UML 2.1 con alto rendimiento, interfaz intuitiva, para traer modelado avanzado al escritorio, y para el equipo completo de desarrollo e implementación. Con un gran conjunto de características y un valor sin igual para el dinero.

Alta capacidad – Características finales superiores a un precio justo.

EA es una herramienta comprensible de diseño y análisis UML, cubriendo el desarrollo de software desde el paso de los requerimientos a través de las etapas del análisis, modelos de diseño, pruebas y mantenimiento. Es una herramienta multiusuario, basada en Windows, diseñada para ayudar a construir software robusto y fácil de mantener.

Velocidad, estabilidad y buen rendimiento.

El Lenguaje Unificado de Modelado provee beneficios significativos para ayudar a construir modelos de sistemas de software rigurosos y donde es posible mantener la trazabilidad de manera consistente. Enterprise Architect soporta este proceso en un ambiente fácil de usar, rápido y flexible.

Trazabilidad de extremo a extremo.

EA provee trazabilidad completa desde el análisis de requerimientos hasta los artefactos de análisis y diseño, a través de la implementación y el despliegue. Combinados con la ubicación de recursos y tareas incorporados, los equipos de Administradores de Proyectos y Calidad están equipados con la información que ellos necesitan para ayudarles a entregar proyectos en tiempo.

Construido sobre las bases de UML 2.1.

Las bases de EA están construidas sobre la especificación de UML 2.1 pero no se detiene ahí. Usa Perfiles UML para extender el dominio de modelado, mientras que la Validación del Modelo asegura integridad. Soporte para los 13 diagramas de UML 2 y más.

Diagramas Estructurales:

• Clase

• Objeto

• Compuesto
• Paquete
• Componente
• Despliegue Diagramas de Comportamiento:
• Casos de Uso
• Comunicación
• Secuencia
• Descripción de la Interacción
• Actividad
• Estado

• Tiempo Extendidos:
• Análisis (actividad simple)
• Personalizado (para requisitos, cambios)

EA ayuda a administrar la complejidad con herramientas para rastrear las dependencias, soporte para modelos muy grandes, control de versiones con proveedores CVS o SCC, Líneas Base por cada punto del tiempo, la utilidad de comparar (diff) para seguir los cambios del modelo, interfaz intuitiva y de alto rendimiento con vista de proyecto como un “explorador”.

EA provee una generación poderosa de documentos y herramientas de reporte con un editor de plantilla completo WYSIWYG. Genera reportes detallados y complejos de EA con la información que usted necesita en el formato que su compañía o cliente demanda.

EA soporta generación e ingeniería inversa de código fuente para muchos lenguajes populares, incluyendo C++, C#, Java, Delphi, VB.Net, Visual Basic y PHP. También hay Add-ins gratis para CORBA y Python disponibles. Con un editor de código fuente con “resaltador de sintaxis” incorporado, EA permite navegar y explorar su modelo de código fuente en el mismo ambiente. Para aquellos que trabajan en Eclipse o Visual Studio.Net, Sparx Systems también vende puentes livianos para estas IDE’s, permitiéndole modelar en EA y saltar directamente al código fuente en su editor preferido. Las plantillas de generación de código le permiten personalizar el código fuente generado a las especificaciones de su compañía.

EA ayuda a visualizar sus aplicaciones soportando ingeniería inversa de un amplio rango de lenguajes de desarrollo de software y esquemas de repositorios de base de datos. Soporta transformaciones de

Arquitectura avanzada dirigida por Modelos (MDA) usando plantillas de transformaciones de desarrollo y fáciles de usar. Con transformaciones incorporadas para DDL, C#, Java, EJB y XSD, Ud. Puede rápidamente desarrollar soluciones complejas desde los simples "modelos independientes de plataforma" (MIP) que son el objetivo en "modelos específicos de plataforma" (MEP). Un MIP se puede usar para generar y sincronizar múltiples MIP's – proveyendo un aumento de productividad significativo. [2]

Apache-tomcat-5.5.12

Tomcat (también llamado **Jakarta Tomcat** o **Apache Tomcat**) funciona como un contenedor de servlets (objetos que corren dentro del contexto de un servidor de aplicaciones y extienden su funcionalidad) desarrollado bajo el proyecto Jakarta en la Apache Software Foundation. Tomcat implementa las especificaciones de los servlets y de Java Server Pages (JSP) de Sun Microsystems. Se le considera un servidor de aplicaciones. Éste funciona con cualquier servidor web con soporte para servlets y JSPs.

Tomcat incluye el compilador Jasper, que compila JSPs convirtiéndolas en servlets. El motor de servlets del Tomcat a menudo se presenta en combinación con el servidor web Apache.

Tomcat puede, asimismo, funcionar como servidor web por sí mismo. Opera de tal manera en entornos de desarrollo poco exigentes en términos de velocidad y de manejo de transacciones. Dado que fue escrito en Java, funciona en cualquier sistema operativo que disponga de la máquina virtual. Lo desarrollan y lo mantienen miembros de la Apache Software Foundation y voluntarios independientes. Los usuarios disponen de libre acceso a su código fuente y a su forma binaria en los términos establecidos en la Apache Software Licence. [3]

Pentaho Schema Workbench

Es una interfaz de diseñador que permite crear y probar en las herramientas OLAP esquemas de cubos dimensionales visualmente. Pentaho Schema Workbench cuenta con un motor de consultas especializado en procesar consultas MDX lo que asegura al usuario que todas las consultas que posteriormente serán exportadas a la herramienta OLAP para la visualización de éstas están libres de posibles errores y cumplen con las necesidades del usuario. Estos archivos de esquema que se generan son XML que son creados en una estructura específica usada por el motor Mondrian. Estos modelos de XML son estructuras parecidas a un cubo que utilizan una tabla HECHO existente y varias

tablas DIMENSIÓN encontradas en su RDBMS. Esto no requiere que un cubo real físico sea construido o mantenido; sólo que el modelo de metadata sea creado. [4]

El servidor OLAP Mondrian

Mondrian es una de las aplicaciones más importantes de la plataforma Pentaho BI. Mondrian es un servidor OLAP open source que gestiona comunicación entre una aplicación OLAP (escrita en Java) y la base de datos con los datos fuente. Mondrian actúa como "JDBC para OLAP".

Se ha escogido como servidor OLAP precisamente por ser conjuntamente con Pentaho Schema Workbench suites de Pentaho, por lo que son perfectamente compatibles, de la misma manera el trabajo con Apache-tomcat-5.5.12 obtiene resultados sumamente satisfactorios. Cuenta con un paquete de opciones bastante amplio que la convierten en una herramienta muy atractiva:

- Navegador OLAP.
- Editor MDX.
- Configuración de las Tablas OLAP.
- Graficador.
- Exportador de las consultas a PDF y Excel.

Sistema de Gestor de Base de Datos (SGBD)

POSTGRESQL es un sistema de gestión de base de datos relacional orientada a objetos de software libre, publicado bajo la licencia BSD. Como muchos otros proyectos open source, el desarrollo de PostgreSQL no es manejado por una sola compañía sino que es dirigido por una comunidad de desarrolladores y organizaciones comerciales las cuales trabajan en su desarrollo. Dicha comunidad es denominada el PGDG. [5]

Características

Alta concurrencia

Mediante un sistema denominado MVCC (Acceso concurrente multiversión, por sus siglas en inglés) PostgreSQL permite que mientras un proceso escribe en una tabla, otros accedan a la misma tabla sin necesidad de bloqueos. Cada usuario obtiene una visión consistente de lo último a lo que se le hizo commit. Esta estrategia es superior al uso de bloqueos por tabla o por filas común en otras bases, eliminando la necesidad del uso de bloqueos explícitos.

Amplia variedad de tipos nativos

PostgreSQL provee nativamente soporte para:

- Números de precisión arbitraria.
- Texto de largo ilimitado.
- Figuras geométricas (con una variedad de funciones asociadas)
- Direcciones IP (Ipv4 e Ipv6).
- Bloques de direcciones estilo CIDR.
- Direcciones MAC.
- Arrays.

Adicionalmente los usuarios pueden crear sus propios tipos de datos, los que pueden ser por completo indizables gracias a la infraestructura GiST de PostgreSQL. Algunos ejemplos son los tipos de datos GIS creados por el proyecto PostGIS.

Otras características

- Claves ajenas también denominadas Llaves ajenas o Claves Foráneas (foreign keys).
- Disparadores (triggers): Un disparador o trigger se define en una acción específica basada en algo ocurrente dentro de la base de datos. En PostgreSQL esto significa la ejecución de un procedimiento almacenado basado en una determinada acción sobre una tabla específica. Ahora todos los disparadores se definen por seis características:

    - El nombre del trigger o disparador.
    - El momento en que el disparador debe arrancar.
    - El evento del disparador deberá activarse sobre...
    - La tabla donde el disparador se activara .
    - La frecuencia de la ejecución .

- La función que podría ser llamada.

Entonces combinando estas seis características, PostgreSQL le permitirá crear una amplia funcionalidad a través de su sistema de activación de disparadores (triggers).

- Vistas.
- Integridad transaccional.
- Herencia de tablas.
- Tipos de datos y operaciones geométricas.

Funciones

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 (triggers en inglés) son funciones enlazadas a operaciones sobre los datos.

1.10 SISTEMAS OPERACIONALES VS DWH

Sistemas Operacionales (OLTP)	Almacenes de Datos (DWH)
- almacena datos actuales	- almacena datos históricos
- almacena datos de detalle	- almacena datos de detalle y datos agregados a distintos niveles
- los datos son dinámicos (actualizables)	- los datos son estáticos
- los procesos (transacciones) son repetitivos	- los procesos no son previsibles
- soporta decisiones diarias	- soporta decisiones estratégicas
- dedicado al procesamiento de transacciones	- dedicado al análisis de datos
- orientado a los procesos de la organización	- orientado a la información relevante

Figura 2: Diferencia entre Sistemas operacionales y Data Warehouse.

Diferencia entre OLTP y Data Warehouse

Los sistemas tradicionales de transacciones y las aplicaciones de los Data Warehouse son polos opuestos en cuanto a sus requerimientos de diseño y sus características de operación. Es importante comprender perfectamente estas diferencias para evitar caer en el diseño de un Data Warehouse como si fuera una aplicación de transacciones en línea (OLTP).

Las aplicaciones de OLTP están organizadas para ejecutar las transacciones para los cuales fueron hechos, como por ejemplo: mover dinero entre cuentas, un cargo o abono, una devolución de inventario, etc. Por otro lado, un DW está organizado en base a conceptos, como por ejemplo: clientes, facturas, productos, etc.

Otra diferencia radica en el número de usuarios. Normalmente, el número de usuarios de un Data Warehouse es menor al de un OLTP. Es común encontrar que los sistemas transaccionales son accesados por cientos de usuarios simultáneamente, mientras que los DW sólo por decenas. Los sistemas de OLTP realizan cientos de transacciones por segundo mientras que una sola consulta de un Data Warehouse puede tomar minutos. Otro factor es que frecuentemente los sistemas transaccionales son menores en tamaño a los almacenes de datos, esto es debido a que un DW puede estar formado por información de varios OLTP´s.

El OLTP normalmente está formado por un número mayor de tablas, cada una con pocas columnas, mientras que en un Data Warehouse el número de tablas es menor, pero cada una de éstas tiende a ser mayor en número de columnas.

Los OLTP son continuamente actualizados por los sistemas operacionales del día a día, mientras que los Data Warehouse son actualizados de manera periódica. Las estructuras de los OLTP son muy estables, rara vez cambian, mientras las de los Almacenes de Datos sufren cambios constantes derivados de su evolución. Esto se debe a que los tipos de consultas a los cuales están sujetos son muy variados y es imposible preverlos todos de antemano. [6]

1.11 HERRAMIENTAS DE CONSULTAS Y ANÁLISIS

Las herramientas de consulta y análisis son sistemas que permiten al usuario realizar la exploración de datos del DW. Básicamente constituyen el nexo entre el depósito de datos y los usuarios. [1]
Existen varios tipos de herramientas de consulta y análisis en dependencia de los requerimientos del negocio como tal y las necesidades del usuario. Entre ellas se encuentran:

- Reportes y Consultas
- OLAP
- Data Mining
- EIS

1.11.1 REPORTES Y CONSULTAS

Se han desarrollado varias herramientas para la producción de consultas y reportes, que ofrecen a los usuarios, a través de pantallas gráficas intuitivas, la posibilidad de generar informes avanzados y detallados del área de interés del negocio que se este analizando. El usuario solo debe seguir una serie de simples pasos, como por ejemplo seleccionar opciones de un menú, presionar tal o cual botón para especificar los elementos de datos, sus condiciones, criterios de agrupación y demás atributos que se consideren significativos. [1]

1.11.2 PROCESO ANALÍTICO EN LÍNEA (OLAP)

El procesamiento analítico en línea OLAP (On Line Analytic Processing), es el componente más poderosa de los DW, ya que es el motor de consultas especializado del Almacén de Datos.

Las herramientas OLAP requieren que los datos estén organizados dentro del depósito en forma multidimensional, por lo cual es que utilizan los cubos multidimensionales.

Además, a través de este tipo de herramientas, se puede analizar el negocio desde diferentes escenarios históricos, y proyectar como se ha venido comportando y evolucionando en un ambiente multidimensional, o sea, mediante la combinación de diferentes perspectivas, temas de interés o dimensiones. Esto permite deducir tendencias, por medio del descubrimiento de relaciones entre las perspectivas que a simple vista no se podrían encontrar sencillamente.

Además de las características ya descritas, se pueden enumerar las siguientes:

- Permite recolectar y organizar la información analítica necesaria para los usuarios y disponer de ella en diversos formatos, tales como tablas, gráficos, reportes, etc.
- Soporta análisis complejos de grandes volúmenes de datos.
- Complementa las actividades de otras herramientas que requieran procesamiento analítico en línea.
- Presenta al usuario una visión multidimensional de los datos (matricial) para cada tema de interés del negocio.
- Es transparente al tipo de tecnología que soporta el DW, ya sea ROLAP, MOLAP o HOLAP.

- Permite definir de forma flexible las dimensiones que se quieren analizar, sus restricciones, jerarquías y combinaciones.
- No tiene limitaciones con respecto al número máximo de dimensiones permitidas.
- Permite a los usuarios, analizar la información basándose en más criterios que un análisis de forma tradicional.
- Al contar con muestras grandes, se pueden explorar mejor los datos en busca de respuestas.
- Permiten realizar agregaciones y combinaciones de los datos de maneras complejas y específicas, con el fin de realizar análisis más estratégicos. [1]

1.11.2.1 TIPOS DE OLAP

1.11.2.1.1 MOLAP

El objetivo de los sistemas MOLAP (Multidimentional On Line Analytic Processing) es almacenar físicamente los datos en estructuras multidimensionales de manera que la representación externa y la interna coincidan. Para ello, se dispone de estructuras de almacenamiento específicas (Arrays) y técnicas de compactación de datos que favorecen el rendimiento del depósito de datos.

Las principales características de MOLAP son:

- Posee tecnología optimizada para consultas y análisis, basada en el modelo multidimensional.
- Cuenta con un motor especializado.
- Provee herramientas limitadas y propietarias.
- No es adecuada para muchas dimensiones.
- Construye y almacena datos en estructuras multidimensionales.
- Almacenamiento en estructura multidimensional
  (Analysis Services).
- Mayor rapidez de respuestas.

Las herramientas de consulta y exploración OLAP, presentan estas estructuras multidimensionales.

1.11.2.1.2 ROLAP

Este tipo de organización física se implementa sobre tecnología relacional, pero disponen de algunas facilidades para mejorar el rendimiento.

Es decir, ROLAP (Relational On Line Analytic Processing) cuenta con todos los beneficios de una SGBD Relacional a los cuales se les provee extensiones y herramientas para poder utilizarlo como un Sistema Gestor de DW.[1]

Entre las características más importantes y sobresalientes de ROLAP, se encuentran las siguientes:

- Almacena la información en una base de datos relacional.
- Posee tres capas lógicas: de almacenamiento, de análisis y de presentación.
- Utiliza índices de mapas de bits.
- Utiliza índices de Join.
- Posee técnicas de particionamiento de datos.
- Posee optimizadores de consultas.
- Cuenta con extensiones del SQL (drill-up, drill-down, etc).
- Almacenamiento en base de datos relacional.
- Para grandes volúmenes de datos.

El almacén de datos se organiza a través de una base de datos multidimensional, sin embargo, puede ser soportado por un SGBD Relacional. Para lograr esto se utilizan los diferentes esquemas, en estrella, copo de nieve y constelación, los cuales transformarán el modelo multidimensional y permitirán que pueda ser gestionado por un SGDB Relacional, ya que solo se almacenarán tablas.

1.11.2.1.3 HOLAP

HOLAP (Hybrid On Line Analytic Processing) constituye un sistema híbrido entre MOLAP y ROLAP, que combina estas dos implementaciones para almacenar algunos datos en un motor relacional y otros en una base de datos multidimensional.

Híbrido

- Respuestas rápidas y gran cantidad de datos en origen.

Debido a que todas las herramientas OLAP requieren que los datos estén organizado dentro del depósito en forma multidimensional, para realizar los cubos multidimensionales se decidió escoger el sistema MOLAP por las facilidades que brinda de almacenar físicamente los datos en estructuras multidimensionales, y como del diseño del DW se desprendieron pocas tablas dimensiones es sistema perfecto en la generación de respuestas rápidas.[1]

1.11.3 DATA MINING

Esta herramienta constituye una poderosa tecnología con un gran potencial que ayuda y brinda soporte a los usuarios, con el fin de permitirles analizar y extraer conocimientos ocultos y predecibles a partir de los datos almacenados en un DW o en un OLTP. Claro que es deseable que la fuente de información sea un DW, por todas las ventajas que aporta. [1]

Los sistemas Data Mining se desarrollan bajo lenguajes de última generación basados en la Inteligencia Artificial y utilizan métodos matemáticos tales como:

- Redes Neuronales.
- Sistemas Expertos.
- Programación Genética.
- Árboles de Decisión.

1.11.3.1 REDES NEURONALES

Se utilizan para construir modelos predictivos no lineales que aprenden a través de entrenamiento y que semejan la estructura de una red neuronal biológica. Una red neuronal es un modelo computacional con un conjunto de propiedades específicas, como la habilidad de adaptarse o aprender, generalizar u organizar la información, todo ello basado en un procesamiento eminentemente paralelo. Por ejemplo, las redes neuronales pueden emplearse para:

- Resolver problemas en dominios complejos con variables continuas y categóricas.
- Modelizar relaciones no lineales.
- Clasificar y predecir resultados. [1]

1.11.3.2 SISTEMAS EXPERTOS

Un sistema experto, puede definirse como un sistema informático (hardware y software) que simula a los expertos humanos en un área de especialización dada. La principal ventaja de estos sistemas es que un usuario con poca experiencia puede resolver problemas que requieren el conocimiento de un experto en el tema. Por ejemplo, los sistemas expertos pueden utilizarse para:

- Realizar transacciones bancarias a través de cajeros automáticos.
- Controlar y regular el flujo de tráfico en las calles y en los ferrocarriles, mediante la operación automática de semáforos.
- Resolver complicados problemas de planificación en los cuales intervienen muchas variables.
- Descubrir relaciones entre diversos conjuntos de variables. [1]

1.11.3.3 PROGRAMACIÓN GENÉTICA

El principal objetivo de la programación genética es lograr que las computadoras aprendan a resolver problemas sin ser explícitamente programadas para ello, generando de esta manera soluciones a partir de la inducción de los programas. El verdadero valor de esta inducción está fundamentado en que todos los problemas se pueden expresar como un programa de computadora. Por ejemplo, la programación genética se utiliza para:

- Resolver problemas, para los cuales es difícil y no natural tratar de especificar o restringir con anticipación el tamaño y forma de una solución eventual.
- Analizar sistemas que actúan sobre condiciones inestables en ambientes cambiantes.
- Generar de manera automática programas que solucionen problemas planteados. [1]

1.11.3.4 ÁRBOLES DE DECISIÓN

Son estructuras de forma de árbol que representan conjuntos de decisiones. Estas decisiones generan reglas para la clasificación de un conjunto de datos, las cuales explican el comportamiento de una variable con relación a otras, y pueden traducirse fácilmente en reglas de negocio. Son utilizados con finalidad predictiva y de clasificación. Por ejemplo, los árboles de decisión pueden emplearse para:

- Optimizar respuestas de campañas.
- Identificar clientes potenciales.
- Realizar evaluación de riesgos.

1.11.4 DETECCIÓN DE DESVIACIÓN

Analiza una serie de datos similares, y cuando encuentra un elemento que no coincide con el resto lo considera una desviación. Usualmente para la detección de la desviación en base de datos grandes se utiliza la información explícita externa a los datos, así como las limitaciones de integridad o modelos predefinidos. En un método lineal, al contrario, se enfoca el problema desde el interior de los datos, empleando la redundancia implícita de los mismos. Por ejemplo, la detección de desviación puede utilizarse para:

- Descubrir excepciones a modelos establecidos.
- Delimitar grupos que cumplan con condiciones preestablecidas.

1.11.5 EIS

EIS (Executive Information System) proporciona medios sencillos para consultar, analizar y acceder a la información de estado del negocio. Además, pone a disposición facilidades para que el usuario pueda conseguir los datos buscados rápidamente, empleando el menor tiempo posible para comprender el uso de la herramienta.
Usualmente, EIS se utiliza para analizar las métricas e indicadores de performance y desempeño del negocio o área de interés, a través de la presentación de vistas con datos simplificados, altamente consolidados, mayormente estáticos y preferentemente gráficos.

Debido a que el procesamiento analítico en línea brinda rápidas respuestas a complejas preguntas, para la interpretación de la tomas de decisiones, se ha decidido tomar a este como la herramienta de consulta y análisis para la interpretación del negocio.

1.12 TECNOLOGÍAS OLTP vs. OLAP

	OLTP	OLAP
Usuario	Profesional de TI	Analista de Información
Función	Operaciones diarias	Apoyo a la decisión
Diseño de BD	Orientada a la aplicación (Basado en EE-R)	Orientado al tema/negocio (estrella, Copos de nieve,...)
Datos	Actuales, Aislados	Históricos, Consolidados
Vistas	Detallados, Planos, Relac.	Agregados, Multidimensional
Destino/utilización	Estructuradas, repetitivas	Ad-Hoc
Unidades de trabajo	Transacciones simples	Consultas complejas
Acceso	Lectura/escritura	Lectura mayoritariamente
# Registros accedidos	Decenas	Millones
# Usuarios	"Miles"	"Centenares"
Tamaño de la BD	100 MB-GB	100 GB-TB
Medidas de rendimiento	Cantidad de transacciones	Cantidad de consultas, Respuesta

Figura 3: Diferencia entre OLTP y OLAP.

Las bases de datos multidimensionales implican tres variantes posibles de modelamiento, que permiten realizar consultas de soporte de decisión:

1.13 VARIANTES DE MODELACIÓN DE BASE DE DATOS

- Esquema en estrella (Star Scheme).
- Esquema copo de nieve (Snowflake Scheme).

- Esquema constelación o copo de estrellas (Starflake Scheme).

Estos esquemas pueden ser implementados de diversas maneras, que, independientemente al tipo de arquitectura, requieren que toda la estructura de datos este desnormalizada o semidesnormalizada, para evitar desarrollar uniones complejas para acceder a la información, con el fin de agilizar la ejecución de consultas. Los diferentes tipos de implementación son los siguientes:

- Relacional – ROLAP.
- Multidimensional – MOLAP.
- Híbrido – HOLAP.

Esquema en estrella

El esquema en estrella, consta de una tabla de hechos central y de varias tablas de dimensiones relacionadas a esta, a través de sus respectivas claves.

Características del Esquema en estrella

- Posee los mejores tiempos de respuesta.
- Su diseño es fácilmente modificable.
- Existe paralelismo entre su diseño y la forma en que los usuarios visualizan y manipulan los datos.
- Simplifica el análisis.
- Facilita la interacción con herramientas de consulta y análisis. [1]

En la figura 4 se puede apreciar un esquema en estrella estándar:

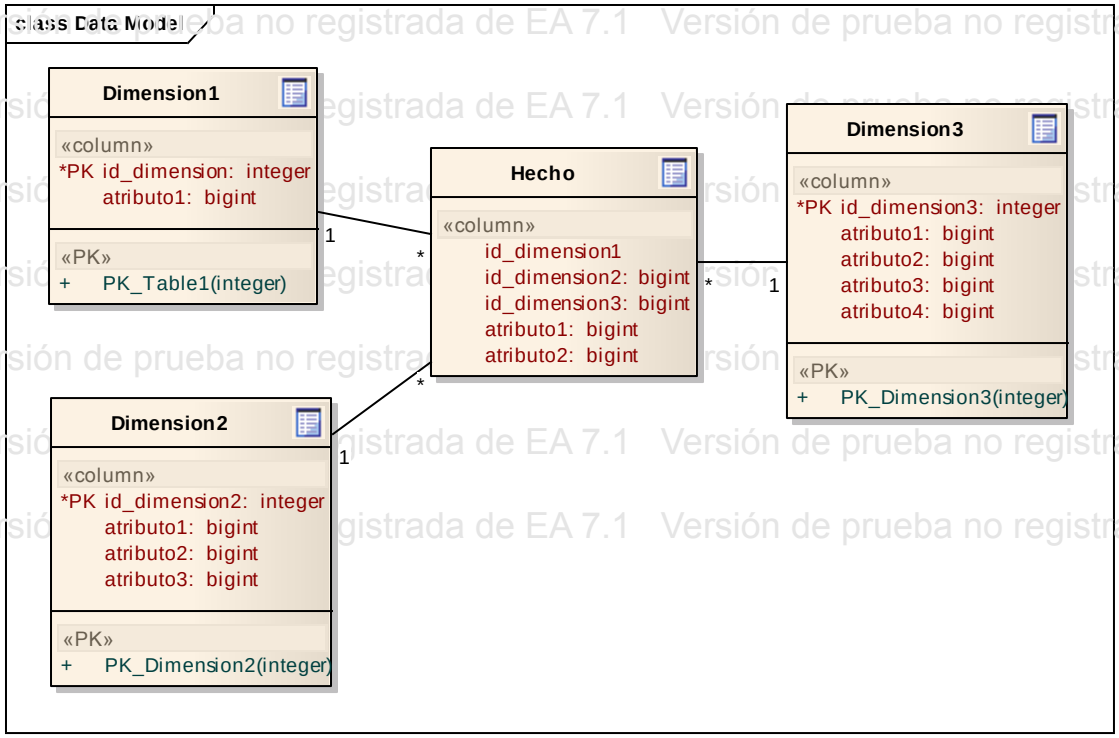


Figura 4: Esquema Estrella.

Esquema Copo de Nieve

Este esquema representa una extensión del modelo en estrella cuando las dimensiones se organizan en jerarquías de dimensiones.[1]

Se pueden definir las siguientes características de este tipo de modelo:

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

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

- Si se poseen múltiples dimensiones, cada una de ellas con varias jerarquías, se creará un número de dimensiones bastante considerable, que pueden llegar al punto de ser inmanejables.
- Al existir muchas uniones y relaciones entre tablas, el desempeño puede verse reducido.[1]

En la figura 5 se puede apreciar un esquema de Copo de Nieve estándar:

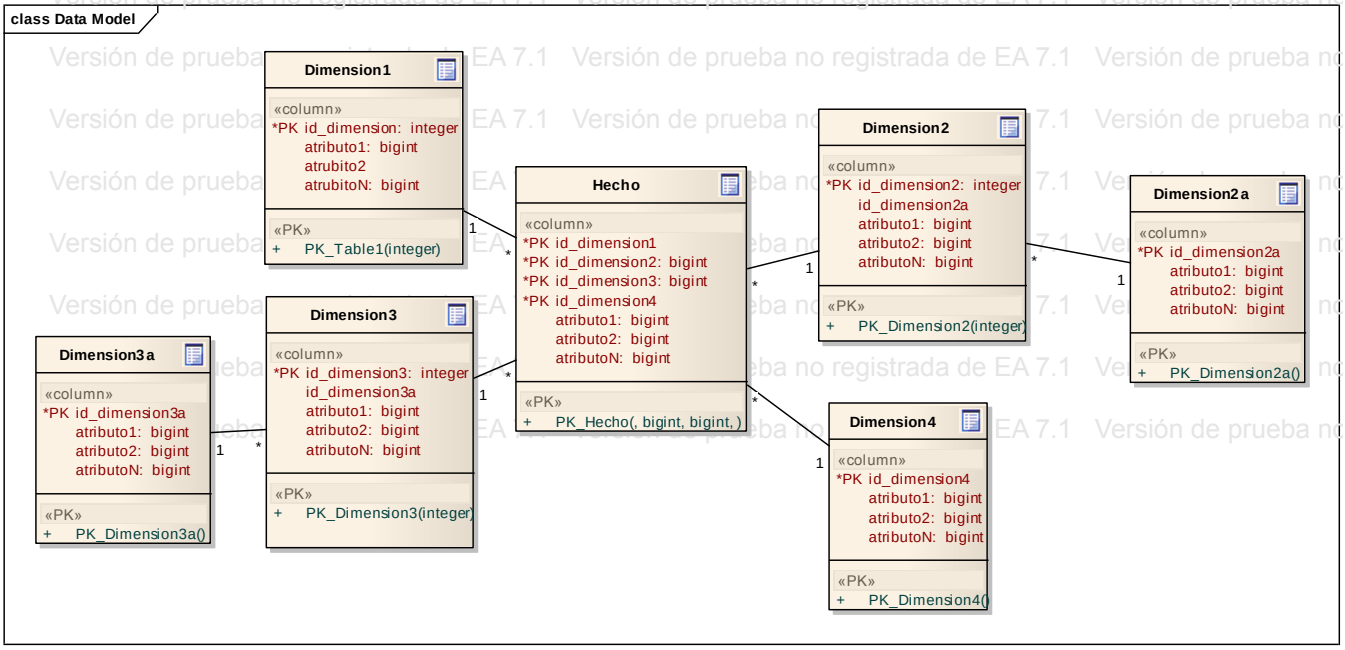


Figura 5: Esquema Copo de Nieve.

Esquema Constelación

Este modelo está compuesto por una serie de esquemas en estrella, y está formado por una tabla de hechos principal y por una o más tablas de hechos auxiliares ,las cuales pueden ser sumarizaciones de la principal. Dichas tablas yacen en el centro del modelo y están relacionadas con sus respectivas tablas de dimensiones.

No es necesario que las diferentes tablas de hechos compartan las mismas tablas de dimensiones, ya que, las tablas de hechos auxiliares pueden vincularse con solo algunas de las tablas de dimensiones asignadas a la tabla de hechos principal, y también pueden hacerlo con nuevas tablas de dimensiones.

Su diseño y cualidad es muy similar a la del esquema en estrella, pero posee una serie de diferencias con el mismo. Entre ellas se pueden mencionar:

- Permite tener más de una tabla de hechos, por lo cual se podrán analizar más aspectos claves del negocio con un mínimo esfuerzo adicional de diseño.
- Contribuye a la reutilización de dimensiones, ya que una misma dimensión puede utilizarse para varias tablas de hechos.
- No es soportado por todas las herramientas de consulta y análisis.[1]

En la figura 6 se puede apreciar un esquema de Constelación estándar:

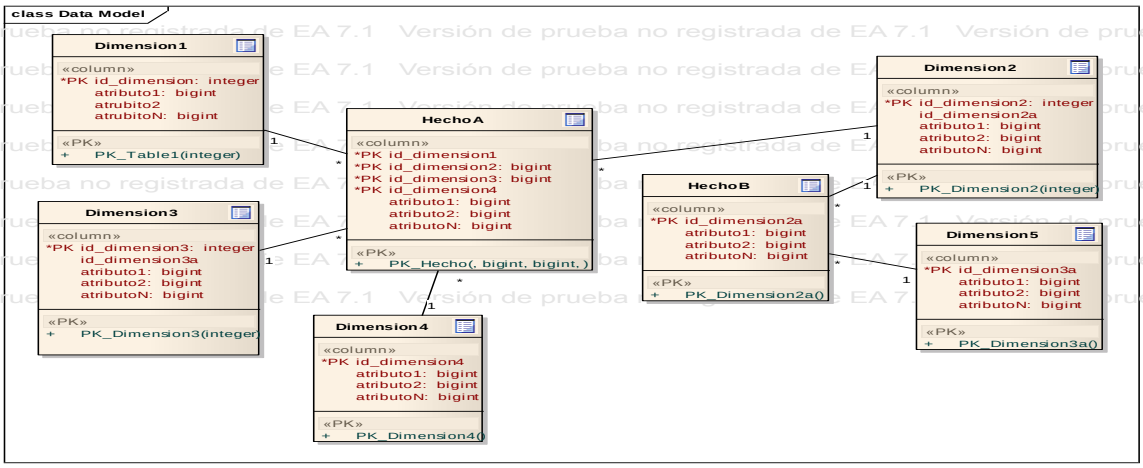


Figura 6: Esquema Constelación.

El esquema que se utilizará será en estrella, debido a sus características, ventajas y diferencias con los otros esquemas, contiene una estructura de depósito de datos que se adapta mejor a los requerimientos y necesidades del usuario. Al contar con baja cantidad de dimensiones ofrece respuestas más rápidas a preguntas más complejas. Por lo que cumple con todos los requisitos necesarios para el desarrollo del Data Warehouse en cuestión.

1.14 RESULTADOS DE LA ARQUITECTURA

Los principales resultados del desarrollo de la arquitectura DW incluyen:

- El modelo de datos fuente.
- El modelo de datos conceptual DW.
- Arquitectura tecnológica DW.
- Estándares y procedimientos DW.
- El plan de implementación incremental para el DW.

Los modelos de datos proveen una estructura para identificar, nombrar, describir y asociar los componentes de una base de datos. En general se necesitan modelos de datos para los datos fuente como para los datos seleccionados para existir en el DW.

Los estándares DW son una parte importante de la arquitectura DW. Sin estándares, oportunidades para rehusar no son posibles y hay riesgos de que partes del desarrollo no ganen trabajando juntos.

El plan de implementación DW es la parte de la arquitectura de DW que identifica los incrementos del DW y describe la secuencia de desarrollo de estos incrementos.

Las implementaciones de los depósitos de datos varían entre sí de forma considerable, teniendo en cuenta las herramientas de software que se empleen, los modelos que se utilicen, recursos disponibles, SGBD que lo soporten, herramientas de análisis y consulta, entre otros.

Al construir un DW, es fundamental que los usuarios del mismo participen activamente durante todo su desarrollo, debido a que son ellos los que conocen en profundidad su negocio y saben cuáles son los resultados que se desean obtener. Además, es precisamente en base a la utilización que se le dé, que el depósito de datos madurará y se adaptará a las situaciones cambiantes por las que atraviese la empresa.

Los usuarios, al trabajar junto a los desarrolladores y analistas podrán comprender más en profundidad sus propios sistemas operacionales, con todo lo que esto implica. Con la implementación del DW, los procesos de toma de decisiones serán optimizados, al obtener información correcta al instante en que se necesita, evitando pérdidas de tiempo y anomalías en los datos. Al contar con esta información, los usuarios tendrán más confianza en las decisiones que tomarán y en adición a ello, poseerán una base sustentable para justificarlas.

Usualmente, los DW integrarán fuentes de datos de diversas áreas y sectores de la empresa, esto tendrá como beneficio contar con una sola fuente de información centralizada y común para todos los usuarios. Esto posibilitará que en las diferentes áreas se compartan los mismos datos, lo cual conducirá a un mayor entendimiento, comunicación, confianza y cooperación entre las mismas.

El depósito de datos introducirá nuevos conceptos tecnológicos y de inteligencia de negocios, lo cual requerirá que se aprendan nuevas técnicas, herramientas, métodos, destrezas, formas de trabajar, etc.

1.15 LENGUAJE MDX

El lenguaje MDX (expresiones multidimensionales) es un lenguaje de secuencias de comandos basado en instrucciones que se utiliza para definir, manipular y recuperar datos de objetos multidimensionales de SQL Server 2005 Analysis Services (SSAS). El lenguaje MDX proporciona instrucciones del lenguaje de definición de datos (DDL); instrucciones del lenguaje de manipulación de datos (DML); instrucciones del lenguaje de secuencias de comandos para administrar el ámbito, el contexto y el control de flujo en secuencias de comandos MDX; operadores y funciones para la manipulación de datos recuperados a partir de objetos multidimensionales; y la posibilidad de ampliar el lenguaje MDX con funciones definidas por el usuario.

1.16 MONDRIAN

Mondrian, ahora rebautizado como Pentaho Analysis Services, es el motor OLAP integrado en la suite de Business Intelligence Open Source Pentaho. Es un proyecto Open Source, licenciado bajo la Mozilla Public License (MPL). Esta licencia es una de las “Business Friendly” lo cual implica que es de las menos restrictivas para su uso desde la mayor parte de los puntos de vista (al igual que el resto de la suite de Pentaho), permitiendo Modificar, Embeber, Modularizar, el software sin restricciones; dejando al parecer de la organización el aporte o no de los cambios realizados al proyecto.

1.17 ¿CÓMO FUNCIONA MONDRIAN?

Visión Global

Mondrian es un motor ROLAP con caché, lo cual lo sitúa cerca del concepto de Hybrid OLAP. ROLAP significa que en Mondrian no residen datos (salvo en la caché) sino que estos residen en una Sistema de Gestión de Bases de Datos externo.

Es en esta base de datos en la que residen las tablas que conforman la información multidimensional con la que Mondrian trabaja. MOLAP es el nombre que reciben los motores OLAP en los que los datos residen en una estructura dimensional.

Mondrian se encarga de recibir consultas dimensionales (lenguaje MDX) y devolver los datos de un cubo, sólo que este cubo no es algo físico sino un conjunto de metadatos que definen como se han de “mapear” estas consultas que tratan conceptos dimensionales a sentencias SQL ya tratando con conceptos relacionales que obtengan de la base de datos la información necesario para satisfacer la consulta dimensional.

Algunas de las ventajas de este modelo son:

El no tener que generar cubos estáticos ahorrando que cuesta generarlos y la memoria que ocupan.

- La posibilidad de utilizar siempre los datos residentes en la base de datos, de forma que se trabaja con datos actualizados. Muy útil en entorno de BI Operacional.
- Pese a que tradicionalmente los sistemas MOLAP tienen una cierta ventaja de rendimiento, la aproximación híbrida de Mondrian, el uso de caché y de tablas agregadas, hace que se puedan obtener muy buenos rendimientos con él, sin perder las ventajas del modelo ROLAP clásico. Es muy importante aprovechar bien las ventajas de la base de datos donde residen las tablas.

CONCLUSIONES DEL CAPÍTULO

Se ha brindado una visión teórica de las principales características e importancia de los Data Warehouse en el manejo y uso eficiente de un gran volumen de información, así como la presentación de la metodología Hefesto. De las herramientas de consulta se utilizó la OLAP por permitir inferir información del comportamiento del negocio y ser una tecnología de software para la ejecución de las consultas. Se seleccionó de las variantes de modelación de las de las Base de Datos el esquema en Estrella por sus mejores tiempos de respuestas ante las necesidades a los clientes y adaptarse perfectamente al caso de estudio.

CAPÍTULO 2. ANÁLISIS Y DISEÑO

En este capítulo se mostrarán los pasos a seguir para la elaboración del diseño del Data Warehouse con el análisis de la metodología. Con la aplicación de la metodología se llevará a la construcción de un modelo conceptual a partir de los requerimientos obtenidos de las entrevistas con los clientes, seguido de los demás pasos de la metodología para llegar así a la elaboración del Data Warehouse.

2.1 PASO 1: ANÁLISIS DE REQUERIMIENTOS

Para la realización del almacén de datos se han seleccionado 12 tablas de la Base de Datos del módulo Análisis Químico del proyecto LIMS Control de Calidad del CIGB.

2.1.1 IDENTIFICAR PREGUNTAS

En las entrevistas con los clientes se indagó cuáles eran sus necesidades, los resultados que esperaban y los reportes que considerasen más importantes donde se viesen reflejadas las actividades con mayor peso dentro del Centro, y que estuviese soportado de alguna manera por algún OLTP.

A continuación, se procedió a identificar qué 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 basadas en ellas.

Se les preguntó cuáles eran según ellos, los indicadores que representan de mejor modo el proceso y qué sería exactamente lo que se desea analizar del mismo, así como cuáles serían las variables o perspectivas desde las cuales se consultarán dichos indicadores.

Resultado de las entrevistas:

- Se desea conocer cuántas unidades de un producto o muestra determinada que el Centro recibió en un tiempo determinado.
- Se desea conocer cuántas unidades de un producto o muestra fueron recibidas por el Centro dado un determinado lote en un tiempo determinado.
- Se desea conocer cuántas unidades de un producto o muestra fueron recibidas por el Centro provenientes de un origen determinado en un tiempo determinado.
- Se desea conocer la cantidad de lotes fueron recibidos por el Centro en un periodo de tiempo.

- Se desea conocer la cantidad de ensayos realizados por lote en un periodo de tiempo.
- Se desea conocer la cantidad de ensayos repetidos por lote y en un periodo de tiempo.
- Se desea conocer las principales causas de repetición de ensayos en un periodo determinado de tiempo.
- Se desea conocer dado un periodo de tiempo y lote, promedio de ensayos conformes (satisfactorios).
- Se desea conocer dado un periodo de tiempo y lote, promedio de ensayos no conformes.
- Se desea conocer la cantidad de técnicas realizadas por lote en un periodo de tiempo.
- Se desea una lista con todas las técnicas utilizadas por un lote determinado en un determinado periodo de tiempo (con el número de veces incluido) .
- Se desea conocer la cantidad de ensayos conformes por técnica en un periodo de tiempo.
- Se desea conocer la cantidad de ensayos no conformes por técnica en un periodo de tiempo.
- Se desea conocer mes más satisfactorios en cuanto a ensayos conformes.
- Se desea conocer año más satisfactorios en cuanto a ensayos conformes.

2.1.2 IDENTIFICAR INDICADORES Y PERSPECTIVAS DE ANÁLISIS

Luego de haber identificado las preguntas según las necesidades de los clientes se procede a determinar los indicadores y perspectivas.

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

Indicador **Perspectivas**

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

Indicador **Perspectivas**

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

Indicador **Perspectivas**

"Cantidad recibida de lotes en un tiempo determinado".

Indicador **Perspectivas**

"Cantidad realizada de ensayos por lote en un tiempo determinado"

Indicador **Perspectivas**

"Cantidad repetida de ensayos por lote en un tiempo determinado"

Indicador **Perspectivas**

"Cantidad de causas de repetición de ensayos en un tiempo determinado"

Indicador **Perspectivas**

"Promedio de ensayos conformes dado un lote y un tiempo determinado"

Indicador **Perspectivas**

"Promedio de ensayos no conformes dado un lote y un tiempo determinado"

Indicador **Perspectivas**

"Cantidad de técnicas realizadas por lote en un tiempo determinado"

Indicador **Perspectivas**

"Cantidad de técnicas utilizadas por lote en un tiempo determinado"

Indicador **Perspectivas**

"Cantidad de ensayos conformes por técnica en un tiempo determinado"

Indicador **Perspectivas**

"Cantidad de ensayos no conformes por técnica en un tiempo determinado"

Indicador **Perspectivas**

Los indicadores que se identificaron fueron:

- 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 identificadas son:

- Lote.
- Planilla Resultado.
- Tiempo.
- Ensayos.
- Técnicas.
- Desviación.
- Curva Calibración.
- Solución.
- Muestra.
- Libro Registro Producto.
- Registro reensayo.
- Registro de Preparación Solución.

2.1.3 MODELO CONCEPTUAL

Una vez identificadas las perspectivas y los indicadores se procede a la confección del modelo conceptual, en el cual se reflejan los indicadores(derecha de la Figura 7) y perspectivas(izquierda de la Figura 7) alcanzándose un elevado nivel de detalle en los datos y permitiendo ser interpretado por el usuario con mayor claridad, proporcionando así una idea precisa de que es lo que se quiere y donde se encuentra ubicada.

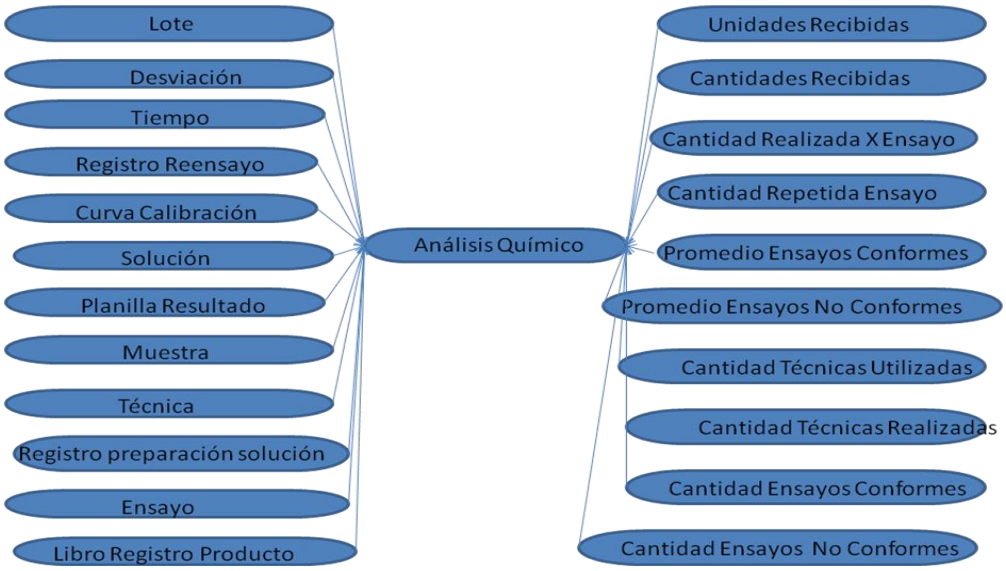


Figura 7: Modelo Conceptual.

2.2 PASO 2: ANÁLISIS DE LOS OLTP

2.2.1 ESTABLECER CORRESPONDENCIAS CON LOS REQUERIMIENTOS

El objetivo de este análisis, es el de examinar los OLTP disponibles que contengan la Información requerida, como así también sus características, para poder identificar las Correspondencias entre el modelo conceptual y las fuentes de datos.

En el caso de los indicadores, deben explicitarse como se calcularán, y más aún si son fórmulas u operaciones complejas. La idea es, que todos los elementos del modelo conceptual estén correspondidos en los OLTP.

Cálculo de los Indicadores

- ”Unidades Recibidas” representa las unidades que se han recibido de un producto en particular.
- ”Cantidades Recibidas” representa la cantidad de muestras recibidas en el lote.
- ”Cantidad Ensayos” representa la cantidad de ensayos realizados, y se obtiene del conteo de los identificadores almacenados de los ensayos.

- ”Cantidad Ensayos Repetidos” representa la cantidad de ensayos repetidos, y se obtiene del conteo de los identificadores almacenados y que en tabla “planilla_resultado” en el campo “repetición” tienen como resultado “S”.
- ”Promedio Ensayos Conformes” representa el promedio de ensayos conformes y se obtiene sumando los ensayos conformes (aquellos ensayos que en la tabla “planilla_resultado” en el campo “repetición” tienen como resultado “N”) obtenidos en un lote en particular y dividiéndolo por la cantidad de ensayos realizados en ese lote.
- ”Promedio Ensayos No Conformes” representa el promedio de ensayos conformes y se obtiene sumando los ensayos conformes (aquellos ensayos que en la tabla “planilla_resultado” en el campo “repetición” tienen como resultado “S”) obtenidos en un lote en particular y dividiéndolo por la cantidad de ensayos realizados en ese lote.
- ”Cantidad Técnica Utilizada” representa la cantidad de técnicas que se utilizan en las muestras y se obtiene con la sumatoria de las técnicas que tengan diferentes identificadores.
- ”Cantidad Técnica Realizada” representa la cantidad de técnicas que se hacen con las muestras y se obtiene con la suma de todas las técnicas que se realizan sin importar que se repita alguna.
- ”Cantidad Ensayos Conformes” representa la cantidad de ensayos satisfactorios por técnicas que se realicen.
- ”Cantidad Ensayos No Conformes” representa la cantidad de ensayos fallidos por técnicas que se realicen.

2.2.2 SELECCIONAR LOS CAMPOS QUE INTEGRARÁN CADA PERSPECTIVA. NIVEL DE GRANULARIDAD

Una vez que se han establecido las relaciones con los OLTP, se examinarán y seleccionarán los campos que contendrá cada perspectiva, ya que será a través de estos por los que se manipularán y filtrarán los indicadores.

Luego de exponer frente al usuario los datos existentes, explicando su significado, valores posibles y características, se decidieron cuales son los que considera relevantes para consultar los indicadores y cuáles no.

Finalmente, y con el fin de graficar los resultados obtenidos, se ampliará el modelo conceptual expuesto anteriormente, colocando bajo cada perspectiva los campos o atributos elegidos. [1]

Una vez que se recolecta toda la información pertinente y se consulta con los usuarios cuáles eran los datos que consideraban de interés para analizar los indicadores ya expuestos, los resultados obtenidos fueron los siguientes:

- En la perspectiva “Muestra” se tendrán en cuenta: el identificador de la muestra, el identificador del lote, el nombre de la muestra y la cantidad de muestra de la tabla “Muestra”.
- En la perspectiva “Lote” se tendrán en cuenta: el identificador del lote, la fecha de recibo, el nombre del lote, el origen del lote, la fecha de fabricación y la fecha de vencimiento de la tabla “Lote”.
- En la perspectiva “Ensayo” se tendrán en cuenta: el identificador del ensayo, el identificador de la muestra, el identificador del lote, la fecha de recepción y la fecha que se realizó de la tabla “Ensayos”.
- En la perspectiva “Tecnica” se tendrán en cuenta: el identificador de la técnica y el nombre de esta de la tabla “Tecnica”.
- En la perspectiva “Tiempo”, se tendrán en cuenta: el identificador del tiempo y se seleccionaron los campos “Año”, “Mes” y “Dia”.
- En la perspectiva “Libro_Registro_Producto” se tendrán en cuenta: el identificador del libro, el identificador de la muestra, el identificador del lote, el origen del producto, el nombre del producto, y la fecha de recepción de la tabla “Libro_Registro_Producto”.
- En la perspectiva “Desviacion” se tendrán en cuenta: el identificador de la desviación y el nombre de la Desviación de la tabla “Desviacion”.
- En la perspectiva “Registro_preparacion_solucion” se tendrán en cuenta: el identificador del registro, el identificador de solución, el número del lote, la fecha de preparación, la fecha de realización y la fecha de vencimiento de la tabla “Registro_Preparacion_Solucion”.
- En la perspectiva “Solucion” se tendrán en cuenta: el identificador de la solución, el identificador de la curva de calibración, el identificador de la planilla de resultado, el nombre del lote y la fecha de vencimiento de la solución de la tabla “Solucion”.
- En la perspectiva “Curva_Calibracion” se tendrán en cuenta: el identificador de la curva de calibración, el identificador de la desviacion, la repeticion de ensayos, la causa de la repetición y la desviación de la tabla “Curva_Calibracion”.
- En la perspectiva “Registro_Reensayo” se tendrán en cuenta: el identificador del registro de reensayo, la fecha y la causa de la tabla “Registro_Reensayo”.

- En la perspectiva “Planilla_Resultado” se tendrán en cuenta: el identificador de la planilla de resultado, el identificador de la desviación, el identificador de la técnica, el identificador del lote, el identificador del registro, el ensayo inicial, la repetición, lo no válido, la repetición del ensayo, causa de la repetición, , las desviaciones, la fecha de realización, así como, la fecha de recepción de la tabla “Planilla_Resultado”.

2.3 PASO 3: ELABORACIÓN DEL MODELO LÓGICO DE LA ESTRUCTURA DEL DW

2.3.1 MODELO DE LA BASE DE DATOS

Las bases de datos multidimensionales, proveen una estructura que permite tener acceso flexible a los datos, para explorar y analizar sus relaciones, y resultados consiguientes. [1]

Estas se pueden visualizar como un cubo multidimensional, en donde las variables asociadas existen a lo largo de varios ejes o dimensiones, y la intersección de las mismas representa la medida, indicador o el hecho que se está evaluando.

2.3.2 CUBO MULTIDIMENSIONAL

Un cubo multidimensional o hipercubo, representa o convierte los datos planos que se encuentran en filas y columnas, en una matriz de N dimensiones.

En los ejes de la matriz se encuentran los criterios o dimensiones, mediante los cuales se analizará el negocio y en los cruces o intersecciones, residen los valores o hechos que se desean consultar. [1]

2.3.3 DISEÑAR TABLAS DE DIMENSIONES

TABLAS DE DIMENSIONES

Las tablas de dimensiones definen como están los datos organizados lógicamente y proveen el medio para analizar el contexto del negocio. Representan los ejes del cubo, y los aspectos de interés, mediante los cuales el usuario podrá filtrar y manipular la información almacenada en la tabla de hechos. [1]

Este paso, se aplicará por igual a todos los tipos de esquemas lógicos. Lo primero que se hará será crear las dimensiones del mismo, para ello se tomará cada perspectiva con sus atributos relacionados y se les realizará el siguiente proceso:

- Se elegirá un nombre que identifique la dimensión.

- Se añadirá un campo que represente su clave principal.
- Se redefinirán los nombres de los atributos si es que no son lo bastante explicativos.

A continuación, se mostrará como quedarán las tablas de dimensiones:

Perspectiva “Muestra”:

- La nueva dimensión tendrá el nombre “Muestra”.
- Se le agregará una clave principal con el nombre “id_muestra”
- Se le agregará también una clave con el identificador del lote con el nombre ”id_lote”.
- Se le agregará el nombre de la muestra con el nombre ”nombre_muestra”.
- Se le agregará la cantidad de la muestra con el nombre ”cantidad”.

Se puede apreciar el resultado de estas operaciones en la siguiente gráfica:

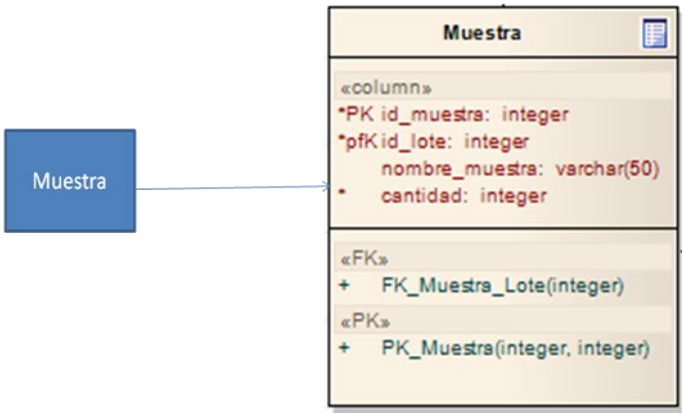


Figura 8: Diseño de la tabla de dimensión muestra.

Perspectiva “Lote”:

• La nueva dimensión tendrá el nombre “Lote”.

• Se le agregará una clave principal con el nombre “id_lote”.

• Se le agregará la fecha de recibo con el nombre “fecha_recibo”.

• Se le agregará el nombre del lote con el nombre “nombre_lote”.

• Se le agregará el origen del lote con el nombre “lote_origen”.

• Se le agregará la fecha de fabricación con el nombre “fecha_fabricacion”.

• Se le agregará la fecha de vencimiento con el nombre “fecha_vencimiento”.

Se puede apreciar el resultado de estas operaciones en la siguiente gráfica:

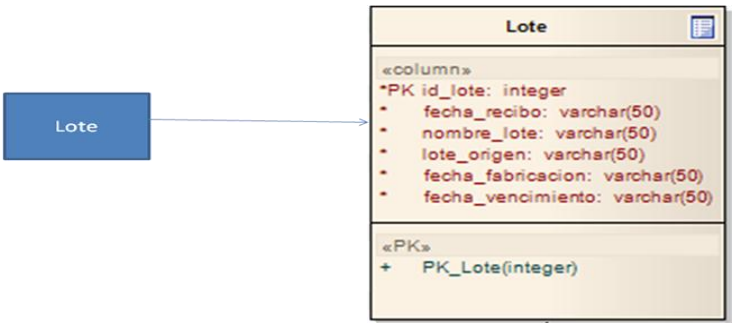


Figura 9: Diseño de la tabla de dimensión lote.

Perspectiva “Técnica”:

• La nueva dimensión tendrá el nombre “Técnica”.

• Se le agregará una clave principal con el nombre “id_tecnica”.

• Se le agregará el nombre de la técnica con el nombre “nombre_tecnica”.

Se puede apreciar el resultado de estas operaciones en la siguiente gráfica:

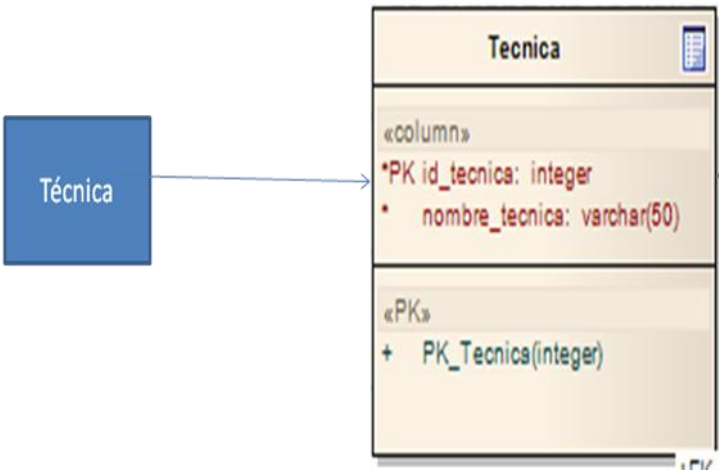


Figura 10: Diseño de la tabla de dimensión técnica.

Perspectiva “Libro_registro_producto”:

• La nueva dimensión tendrá el nombre “Libro_registro_producto”.

• Se le agregará una clave principal con el nombre “id_libro”.

• Se le agregará el identificador de la muestra con el nombre “id_muestra”.

• Se le agregará el identificador del lote con el nombre “id_lote”.

• Se le agregará el origen del producto con el nombre “origen_producto”.

• Se le agregará el nombre del producto con el nombre “nombre_producto”.

• Se le agregará la fecha de recepción con el nombre “fecha_recepcion”.

Se puede apreciar el resultado de estas operaciones en la siguiente gráfica:

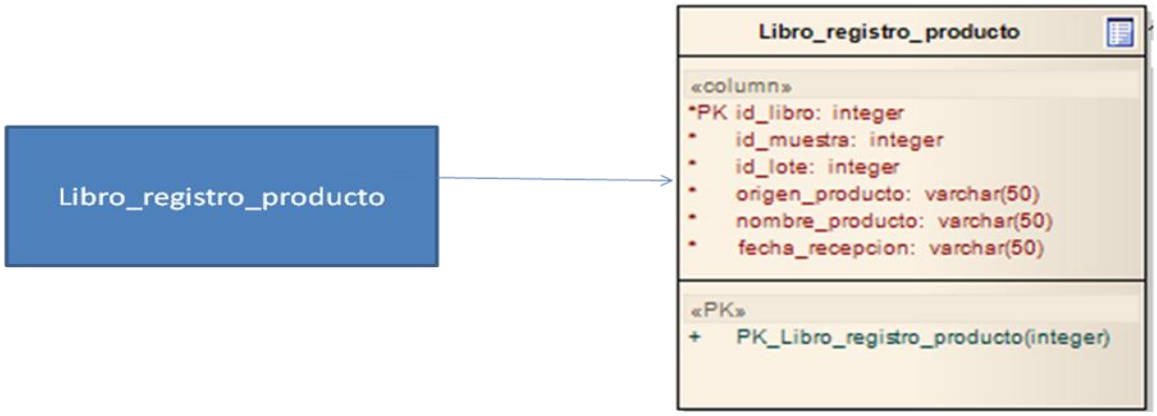


Figura 11: Diseño de la tabla de dimensión Libro_Registro_Producto.

Perspectiva “Ensayo”:

• La nueva dimensión tendrá el nombre “Ensayo”.

• Se le agregará una clave principal con el nombre “id _ensayo”.

• Se le agregará el identificador de la muestra con el nombre “id_muestra”.

• Se le agregará el identificador del lote con el nombre “id_lote”.

• Se le agregará la fecha de recepción con el nombre “fecha_recepcion”.

• Se le agregará la fecha que se realizó con el nombre “fecha_realizado”.

Se puede apreciar el resultado de estas operaciones en la siguiente gráfica:

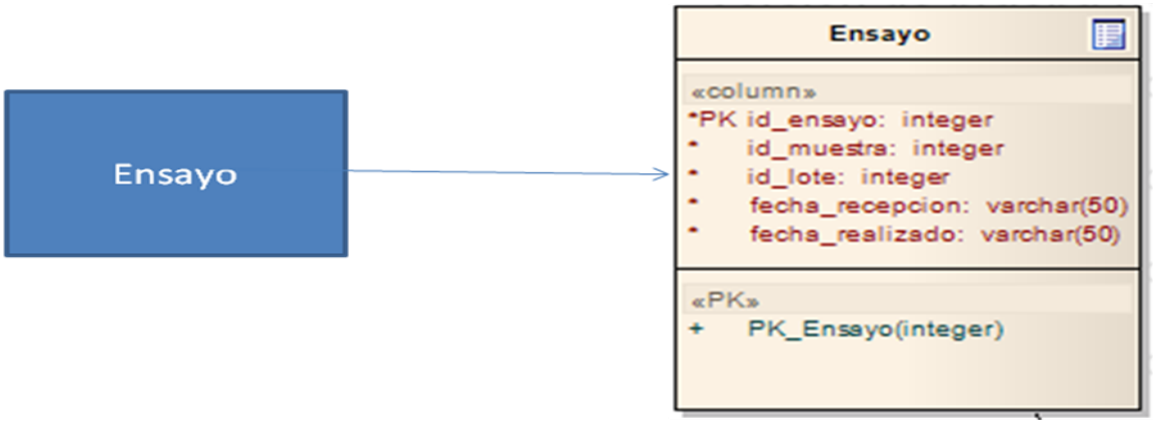


Figura 12: Diseño de la tabla de dimensión ensayo.

Perspectiva “Desviación”:

• La nueva dimensión tendrá el nombre “Desviación”.

• Se le agregará una clave principal con el nombre “id_desviacion”.

• Se le agregará el nombre de la desviación con el nombre “nombre_desviacion”.

Se puede apreciar el resultado de estas operaciones en la siguiente gráfica:

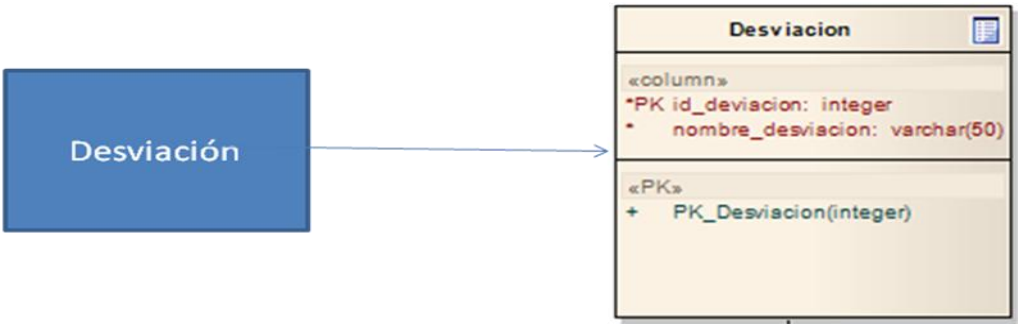


Figura 13: Diseño de la tabla de dimensión desviación.

Perspectiva “Registro_preparacion_solucion”:

• La nueva dimensión tendrá el nombre “Registro_preparacion_solucion”.

• Se le agregará una clave principal con el nombre “id_registro”.

- Se le agregará el identificador de la solución con el nombre “id_solucion”.
- Se le agregará número del lote con el nombre “no_lote”.
- Se le agregará la fecha de preparación con el nombre “fecha_preparacion”.
- Se le agregará la fecha que se realizó con el nombre “fecha_realizacion”.
- Se le agregará la fecha de vencimiento con el nombre “fecha_vencimiento”.

Se puede apreciar el resultado de estas operaciones en la siguiente gráfica:

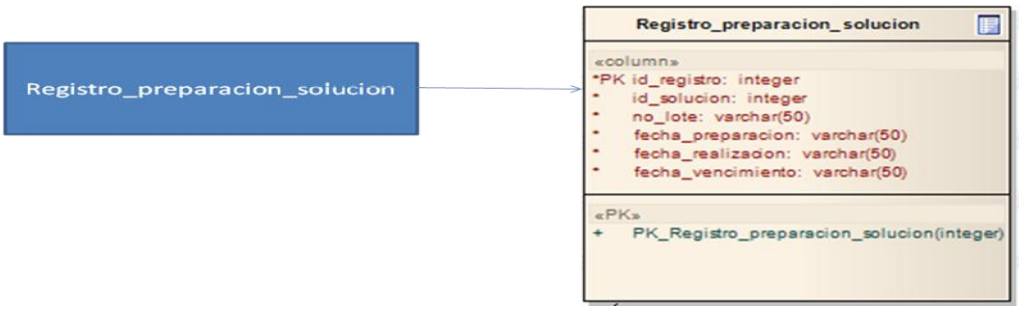


Figura 14: Diseño de la tabla de dimensión Registro_Preparacion_Solucion.

Perspectiva “Tiempo”:

- La nueva dimensión tendrá el nombre “Tiempo”.
- Se le agregará una clave principal con el nombre “id_tiempo”.
- Se le agregará el año con el nombre “anno”.
- Se le agregará el mes con el nombre “mes”.
- Se le agregará el día con el nombre “dia”.

Se puede apreciar el resultado de estas operaciones en la siguiente gráfica:

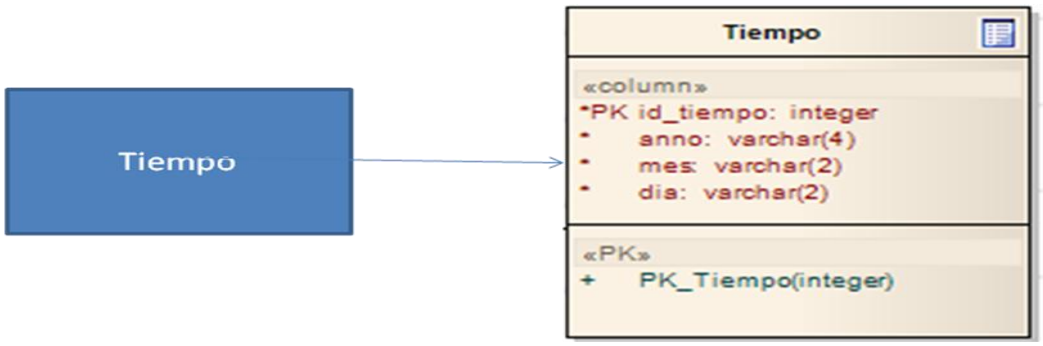


Figura 15: Diseño de la tabla de dimensión tiempo.

Perspectiva “Solución”:

- La nueva dimensión tendrá el nombre “Solución”.
- Se le agregará una clave principal con el nombre “id_solucion”.
- Se le agregará el identificador de la curva de calibración con el nombre “id_curva”.
- Se le agregará el identificador de la planilla de resultado con el nombre “id_planilla”.
- Se le agregará el número del lote con el nombre “no_lote”.
- Se le agregará la fecha en que vence la solución con el nombre “fecha_vencimiento_solucion”.

Se puede apreciar el resultado de estas operaciones en la siguiente gráfica:

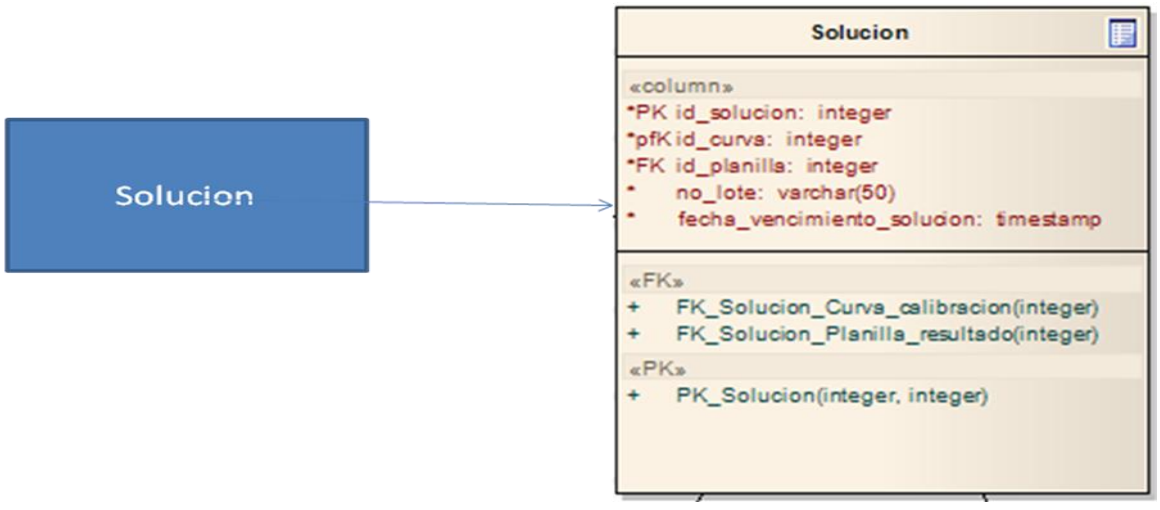


Figura 16: Diseño de la tabla de dimensión solución.

Perspectiva “Curva_Calibracion”:

- La nueva dimensión tendrá el nombre “Curva_calibracion”.
- Se le agregará una clave principal con el nombre “id_curva”.
- Se le agregará el identificador de la desviación con el nombre “id_desviacion”.
- Se le agregará el ensayo repetido con el nombre “repetir_ensayo”.
- Se le agregará la causa de la repetición del ensayo con el nombre “causa_repeticion”.
- Se le agregará la desviación con el nombre “desviacion”.

Se puede apreciar el resultado de estas operaciones en la siguiente gráfica:

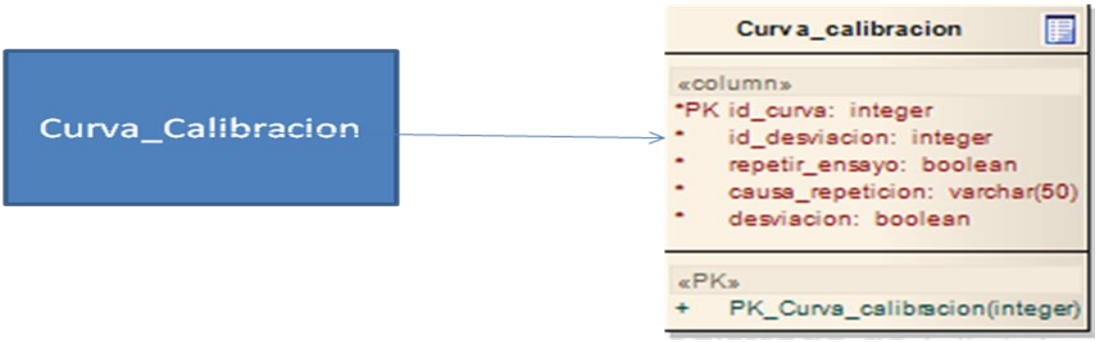


FIGURA 17: DISEÑO DE LA TABLA DE DIMENSIÓN CURVA_CALIBRACION.

Perspectiva “Registro_reensayo”:

- La nueva dimensión tendrá el nombre “Registro_reensayo”.
- Se le agregará una clave principal con el nombre “id_registro_reensayo”.
- Se le agregará la fecha con el nombre “fecha”.
- Se le agregará la causa con el nombre “causa”.

Se puede apreciar el resultado de estas operaciones en la siguiente gráfica:

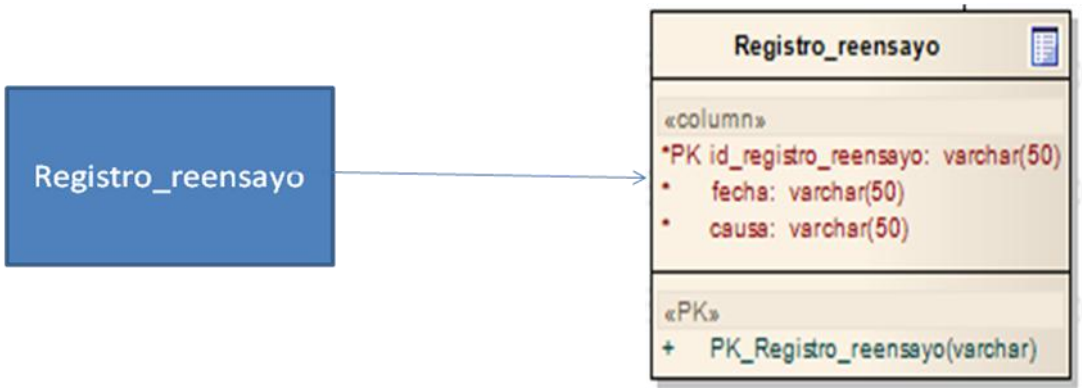


Figura 18: Diseño de la tabla de dimensión registro_reensayo.

Perspectiva “Planilla_resultado”:

- La nueva dimensión tendrá el nombre “Planilla_resultado”.
- Se le agregará una clave principal con el nombre “id_planilla”.
- Se le agregará el identificador de la desviación con el nombre “id_desviacion”.
- Se le agregará el identificador de la técnica con el nombre “id_tecnica”.
- Se le agregará el identificador del lote con el nombre “id_lote”.
- Se le agregará el identificador del registro con el nombre “id_registro”.
- Se le agregará ensayo inicial con el nombre “ensayo_inicial”.
- Se le agregará la repetición con el nombre “repeticion”.
- Se le agregará lo no válido con el nombre “no_valido”.
- Se le agregará la repetición de ensayo con el nombre “repetir_ensayo”.
- Se le agregará la causa de la repetición con el nombre “causa_repeticion”.
- Se le agregará la desviación con el nombre “desviaciones”.
- Se le agregará la fecha que se realizó con el nombre “fecha_realizacion”.
- Se le agregará la fecha de recepción con el nombre “fecha_recepcion”.

• Se le agregará el identificador del registro de reensayo con el nombre “id_registro_reensayo”.

Se puede apreciar el resultado de estas operaciones en la siguiente gráfica:

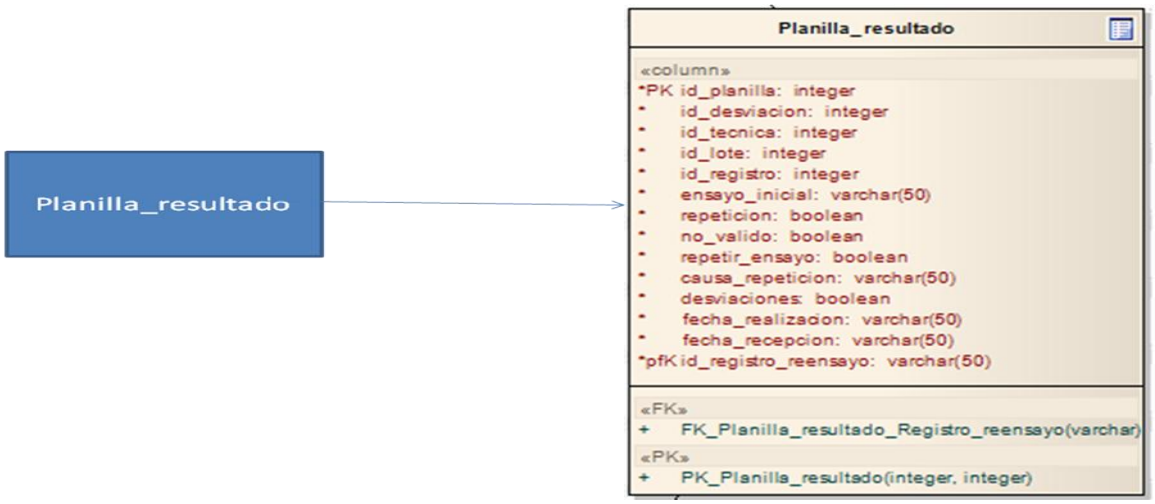


Figura 19: Diseño de la tabla de dimensión planilla_resultado.

2.3.4 DISEÑAR TABLAS DE HECHOS

TABLAS DE HECHOS

Las tablas de hechos contienen los hechos, medidas o indicadores que serán utilizados por los analistas de negocio para apoyar el proceso de toma de decisiones. Los hechos son datos instantáneos en el tiempo, que son filtrados, agrupados y explorados a través de condiciones definidas en las tablas de dimensiones. [1]

En este paso, se definirán las tablas de hechos, que son las que contendrán los indicadores de estudio.

Para el esquema en Estrella se realizará lo siguiente:

•Se le asigna un nombre a la tabla hecho, el cual representará el negocio enfocado

• Se le asignará una clave primaria que será la combinación de todos las llaves primarias de todas las dimensiones que se usarán para realizar las consultas.

• Se renombrarán los hechos o indicadores si es que no llegasen a ser lo suficientemente explícitos.

Las preguntas realizadas por el usuario son examinadas a través de indicadores y dimensiones porque las perspectivas fueron convertidas en dimensiones.

A continuación, se confeccionará la tabla de hechos:

- La tabla de hechos tendrá el nombre “Analisis_Quimico”.
- Su clave principal será la combinación de las claves principales de las dimensiones antes definidas: “id_muestra”, “id_tecnica”, “id_ensayo”, “id_libro”, “id_desviacion”, “id_Tiempo”. “id_registro”, “id_solucion”, “id_lote”, “id_curva”, “id_planilla”, e “id_registro_reensayo” .

En el gráfico siguiente se puede apreciar mejor este paso:

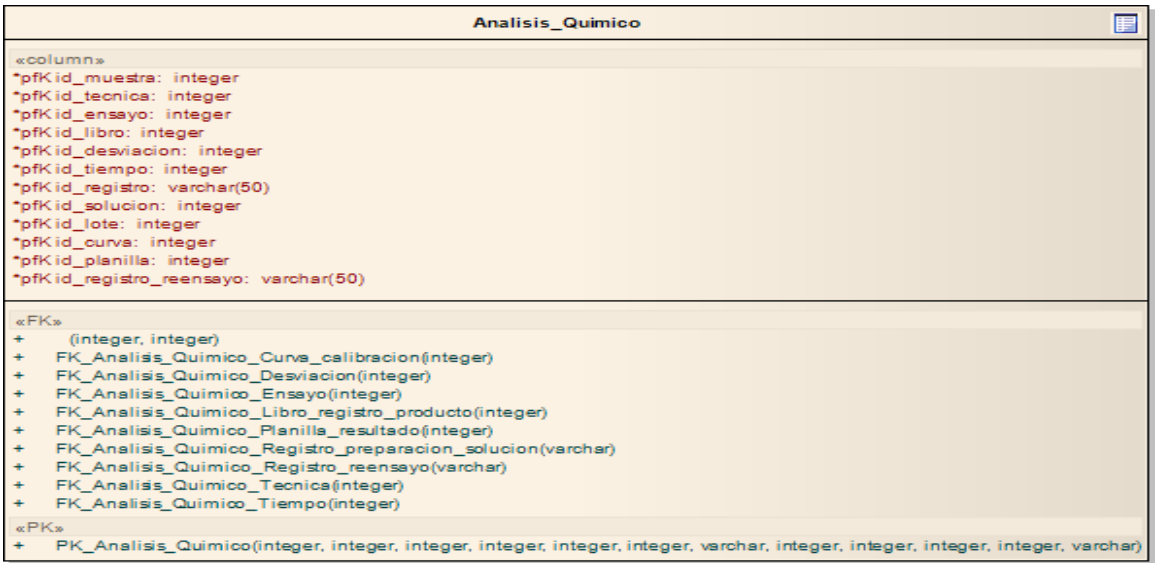


Figura 20: Diseño de la tabla de dimensión Analisis_Quimico.

2.3.5 REALIZAR UNIONES

Para todos los tipos de esquemas, se realizarán las uniones correspondientes entre sus tablas de dimensiones y sus tablas de hechos.

Se realizarán las uniones pertinentes, de acuerdo corresponda:

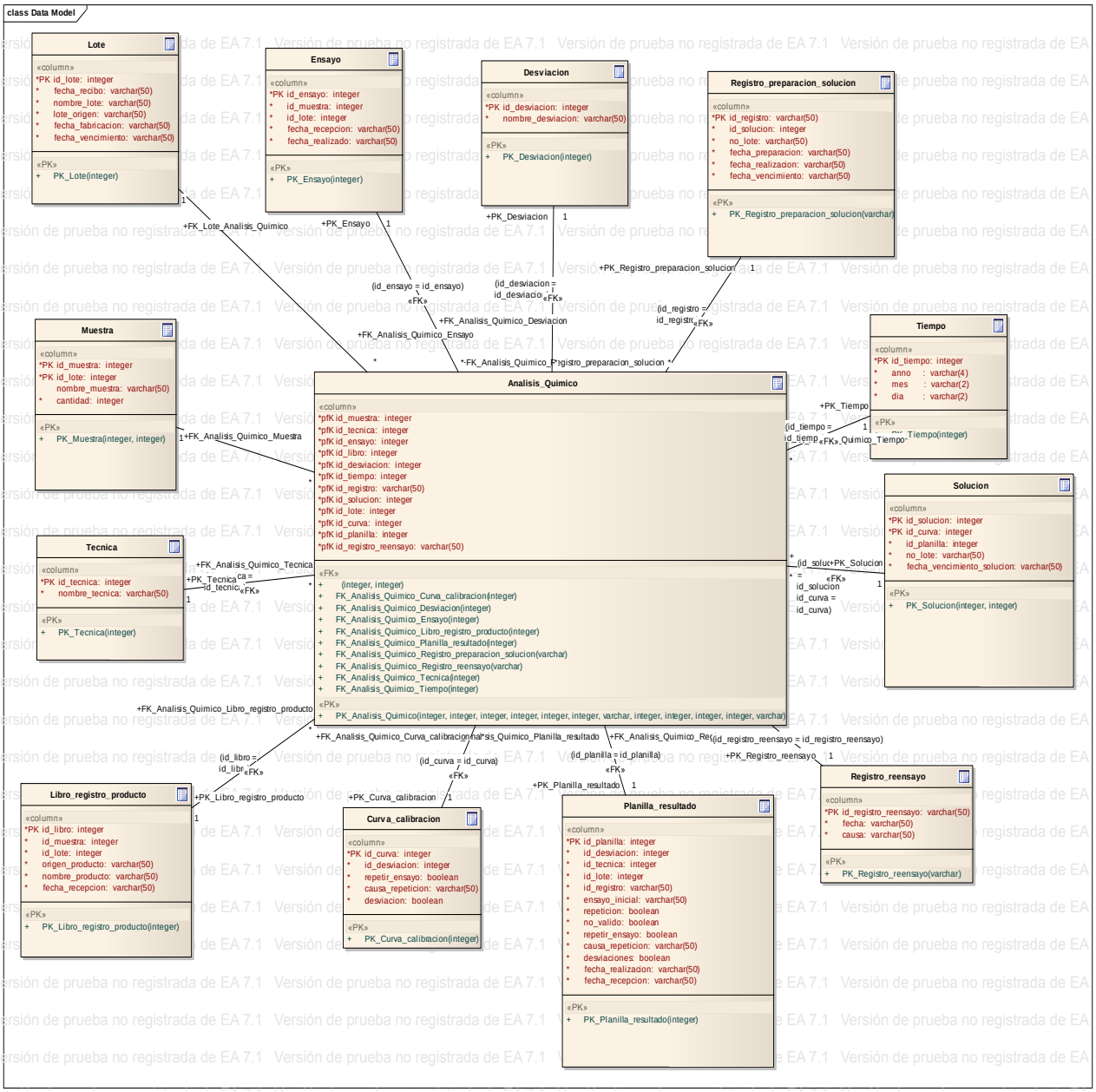


Figura 21: Diseño del Data Warehouse

2.3.6 DETERMINAR JERARQUÍA

A nivel de dimensiones es posible definir jerarquías, las cuales son grupos de atributos que siguen un orden preestablecido. Una jerarquía implica una organización de niveles dentro de una dimensión, con cada nivel representando el total agregado de los datos del nivel inferior. Las jerarquías definen cómo

los datos son sumarizados desde los niveles más bajos hacia los más altos. Una dimensión típica soporta una o más jerarquías naturales.

Una jerarquía puede pero no exige contener todos los valores existentes en la dimensión. Para representar las jerarquías en el modelo lógico, se deberán colocar los atributos pertenecientes a las jerarquías en sus respectivas tablas, en orden descendente y acompañado con un número ordinal encerrado entre corchetes.

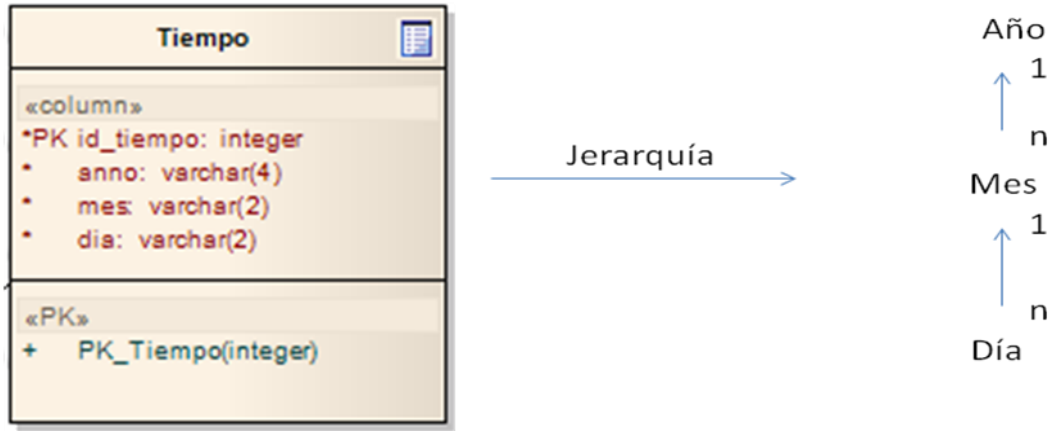


Figura 22: Diseño de la jerarquía de la tabla tiempo.

Entonces, para representar esta jerarquía, la tabla “Tiempo” debe quedar como sigue:

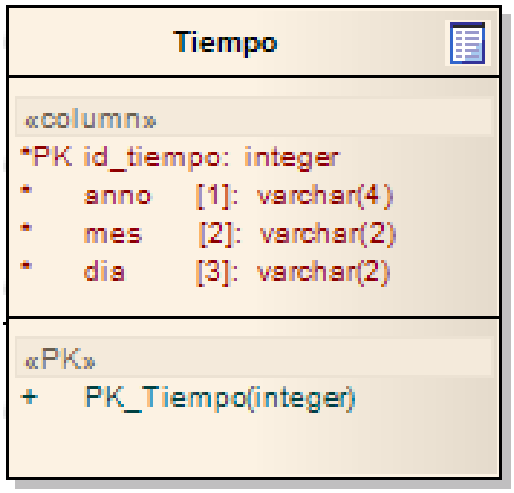


Figura 23: Diseño de la jerarquía final de la tabla tiempo.

En la figura que se presentará a continuación, se puede apreciar el esquema lógico del DW resultante, tras haber definido sus jerarquías correspondientes.

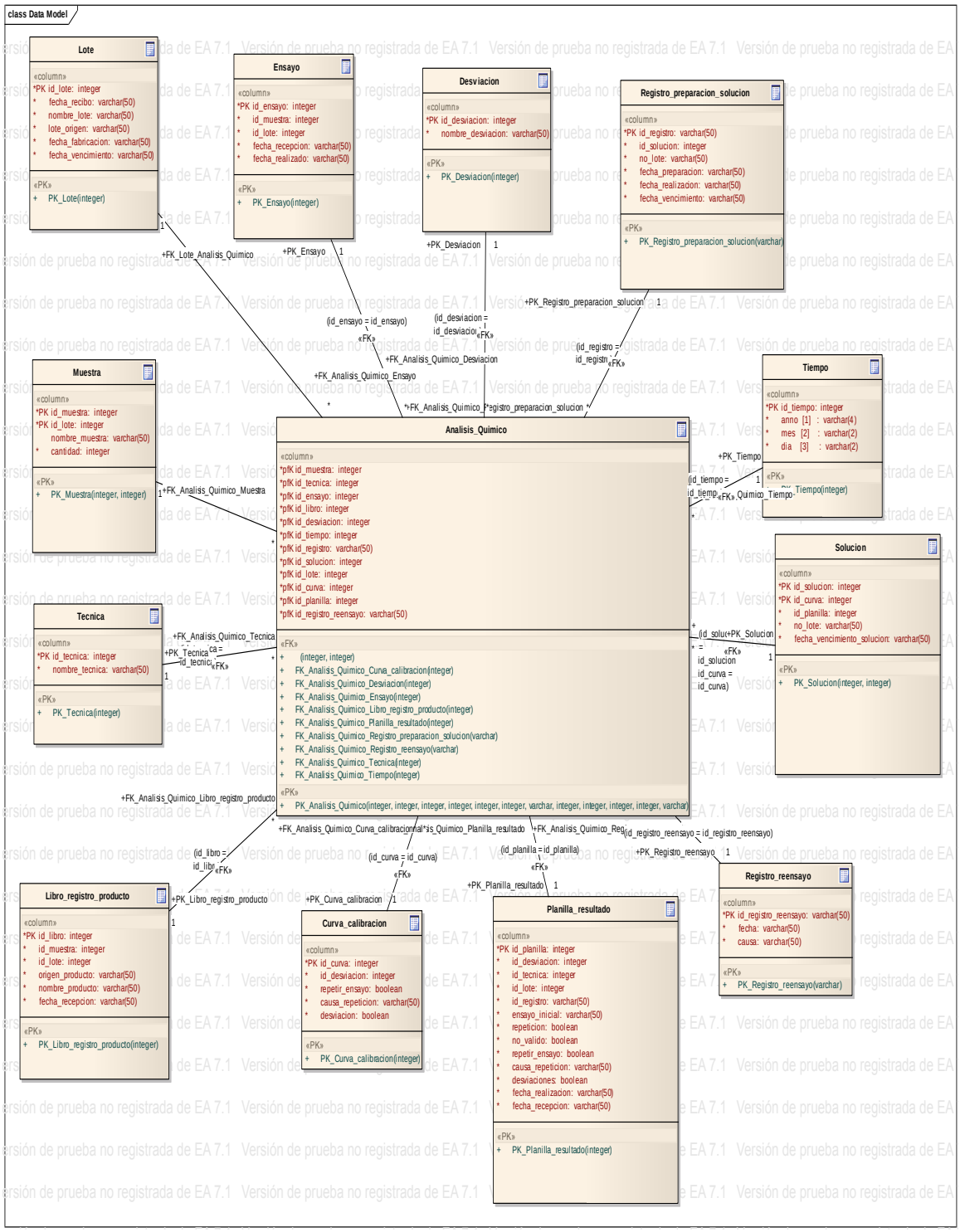


Figura 24: Diseño del Data Warehouse con la jerarquía.

Una vez realizado todo el proceso de análisis y diseño del DW y haber obtenido el diseño final del mismo, este debe ser surtido con la información que se desprende de la Extracción, Transformación y Limpieza de datos que se le realiza a la Base de Datos con la cual se está trabajando, todo este proceso es llevado a cabo por el Trabajo de Diploma “Proceso de la migración de datos hacia un Data Warehouse para el módulo Análisis Químico del proyecto LIMS Control de Calidad”.

Conclusiones del Capítulo

Con al análisis de la Base de Datos del módulo Análisis Químico del Proyecto LIMS Control de Calidad, el seguimiento de los pasos de la Metodología Hefesto y atendiendo a los requerimientos del cliente se han elaborado todas las dimensiones (con los atributos necesarios para la toma de decisiones) conjuntamente con la tabla hecho, quedando como resultado el diseño final del Data Warehouse.

CAPÍTULO 3. IMPLEMENTACIÓN DEL DATA WAREHOUSE

Al iniciar un proyecto Data Warehouse no se debe olvidar establecer un marco de referencia de construcción del DW. Podemos distinguir en dicha construcción dos etapas principales: la definición de una arquitectura DW y la construcción de los incrementos DW. Se explicará el trabajo con las diferentes herramientas y los resultados obtenidos.

Definición de la arquitectura.

3.1 ARQUITECTURA DW:

Arquitectura DW establece el marco de trabajo, estándares y procedimientos para el DW a un nivel empresarial. Los objetivos de las actividades de la arquitectura son simples, integrar al DW las necesidades de información empresarial.

Para la implementación del Data Warehouse se ha decidido realizar los cubos multidimensionales en la herramienta de modelado “Pentaho Schema Workbench”, ya se ha explicado las razones del uso de esta herramienta, y es precisamente en ella donde se harán las consultas que den respuesta a las exigencias de los clientes y que por último serán probadas y visualizadas en la herramienta OLAP “Mondrian”.

A continuación se muestra con una explicación clara y en detalle el trabajo con la herramienta Pentaho Schema Workbench.

Cuando se crean las dimensiones estas deben tener el mismo nombre que poseen en la Base de Datos, luego se le pone el identificador correspondiente, en este caso se nombra “id_muestra” de la tabla “muestra” que forma parte del DW. Se adicionaron a esta herramienta 12 dimensiones para la elaboración del Data Warehouse. Se puede apreciar que aparecen todos los identificadores que posee la tabla Hecho Análisis Químico. Este mismo procedimiento se le aplica a todas las restantes dimensiones. (Ver Figura 25).

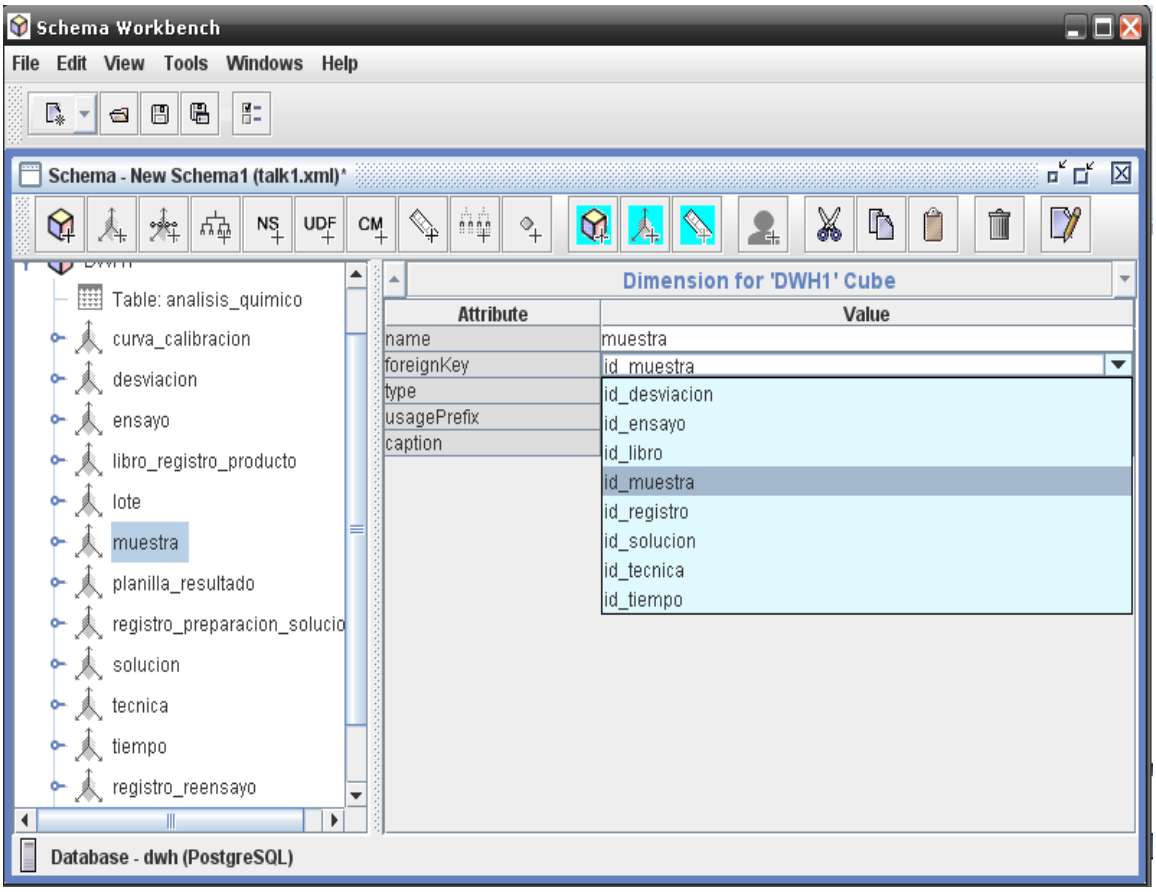


Figura 25: Realización de las dimensiones.

Para la confección del DW se utilizaron las siguientes medidas:

Medidas	Descripción
cant_muestra	Esta es una medida que toma el valor de los identificadores de la dimensión “muestra” y los cuenta según su ocurrencia en la tabla Hecho “Análisis Químico”, obteniéndose como resultado la cantidad total de muestras.
cant_lote	Esta es una medida que toma el valor de los identificadores de la dimensión “lote” y los cuenta según su ocurrencia en la tabla Hecho “Análisis Químico”, obteniéndose como resultado la cantidad total de lotes.
cant_ensayo	Esta es una medida que toma el valor de los identificadores de la dimensión “ensayo” y los

	cuenta según su ocurrencia en la tabla Hecho "Análisis Químico", obteniéndose como resultado la cantidad total de ensayos.
causa_lote	Esta es una medida que toma el valor de los identificadores de la dimensión "lote" y cuenta aquellos que tengan diferente identificador en su ocurrencia en la tabla Hecho "Análisis Químico.
cant_tecnica	Esta es una medida que toma el valor de los identificadores de la dimensión "tecnica" y los cuenta según su ocurrencia en la tabla Hecho "Análisis Químico", obteniéndose como resultado la cantidad total de técnicas.
cant_ensayos_conformes	Esta es una medida que toma el valor de los identificadores de la dimensión "ensayo" y los cuenta según su ocurrencia en la tabla Hecho "Análisis Químico", obteniéndose como resultado la cantidad total de ensayos conformes.

Tabla 1: Medidas y Descripción

Seguidamente se pasa a elaborar las consultas, las cuales después de realizadas se ejecutan para ver si se obtienen los resultados esperados o presente error por algún motivo, ya sea, porque la base de datos este vacía y no pueda ofrecer valores o porque la consulta esté mal elaborada. Se elaboraron 13 consultas que dan respuesta a los todos los requerimientos de los clientes.

Aquí se muestra un ejemplo de una consulta que se genera en la herramienta Pentaho Schema Workbench para luego obtener los datos de los indicadores en la herramienta Mondrian que es precisamente la herramienta OLAP que se utiliza para testear los resultados:

```
SELECT {[MEASURES].[CANTIDAD_MUESTRAS]} ON COLUMNS,
{[MUESTRA].[NOMBRE_MUESTRA].MEMBERS} ON ROWS
FROM DWH1
```

Después de haber obtenido los resultados de la consulta donde se muestran los valores numéricos, la herramienta brinda la posibilidad de graficar estos resultados. A continuación se muestra la configuración de la grafica.

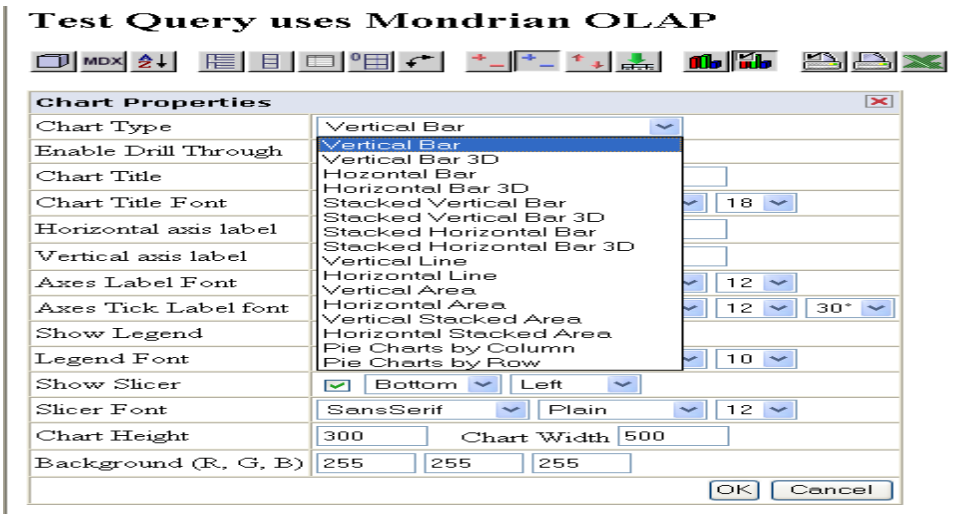


Figura 26: Graficar Resultados.

La grafica quedaría de la siguiente manera:

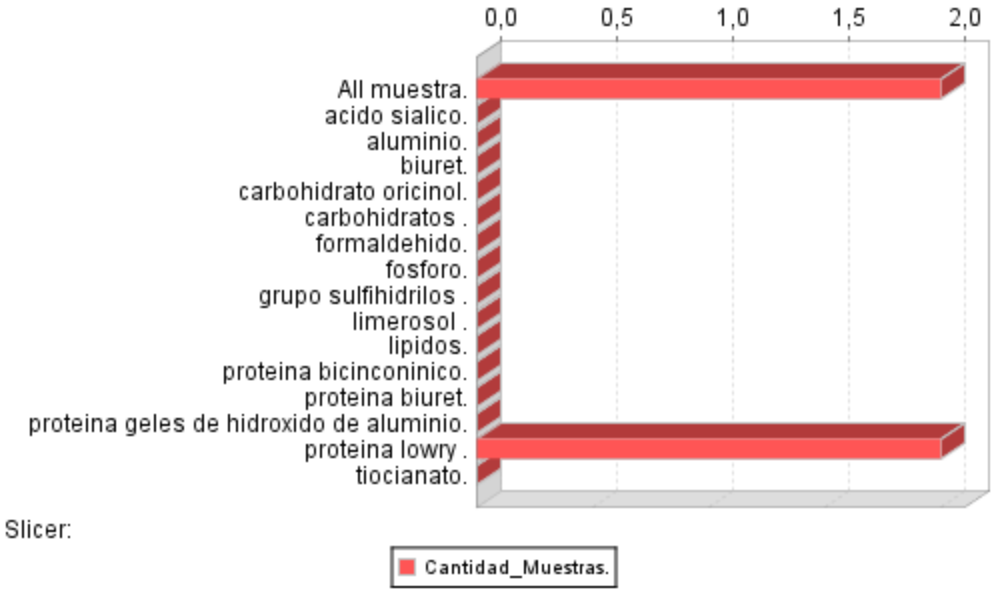


Figura 27: Gráfica.

Conclusiones del Capítulo

Con la utilización de las herramientas escogidas “Pentaho Schema Workbench” y “Mondrian” se ha realizado la perfecta implementación de las consultas que dan respuesta a los requerimientos exigidos

por el cliente, mostrando en datos numéricos y gráficas el resultado de dichas consultas, y permitiendo la generación de cada uno de los reportes en PDF y páginas EXCEL, contando así el usuario con información necesaria y útil para la toma de decisiones.

CONCLUSIONES GENERALES

Con todos los resultados obtenidos se puede llegar a la conclusión que:

- Siguiendo los pasos de la metodología HEFESTO fueron identificadas 12 perspectivas, 10 indicadores, se decidió a utilizar el esquema Estrella para el modelado del DW.
- Con el estudio del Data Warehouse se ha realizado el diseño del mismo para el proyecto LIMS Control de Calidad, el cual cuenta con 12 tablas de dimensiones y una tabla de hecho.
- Se realizó la implementación de éste con la herramienta Pentaho Schema Workbench, en la cual se generó el diseño multidimensional utilizándose 6 medidas y 13 consultas para dar respuesta a los requerimientos del cliente.
- Luego de haber procesado toda la implementación los procedimientos se visualizan con la herramienta Mondrian para el uso de los usuarios finales del Data Warehouse.
- Por lo que se logró el correcto diseño e implementación del Data Warehouse para el proyecto LIMS Control de Calidad que facilitará la toma de decisiones en el Centro de Ingeniería Genética y Biotecnología.

RECOMENDACIONES

Hay una premisa en el mundo del negocio que plantea que el futuro pertenece a quienes pueden verlo y llegar a él primero. Por tanto, es una recomendación a todas las empresas que tengan automatizados todos o parcialmente sus procesos y cuenten con información acumulada sobre los mismos, la implementación de un Data Warehouse, ya que éste permite no solo comprender lo que está pasando, sino predecir lo que va a suceder.

Se propone que a este trabajo de diploma se le haga una aplicación para el posterior uso del Data Warehouse en los Proyectos, tanto de la facultad como a nivel universitario.

REFERENCIAS BIBLIOGRÁFICAS

[1] -Data Warehousing: Investigación y Sistematización de Conceptos –Hefesto: Metodología propia para la Construcción de un Data Warehouse. Ing. Bernabue, Ricardo Dario, Córdoba, Argentina. Miércoles 07 de Noviembre del 2007.

[2] HTTP://WWW.TARINGA.NET/POSTS/DOWNLOADS/1877926/ENTERPRISE-ARCHITECT-7_1.HTML

[3] HTTP://DIS.UM.ES/~LOPEZQUESADA/DOCUMENTOS/IES_0506/RAL_0506/DOC/PRAC6UT1.PDF

[4] HTTP://MONDRIAN.PENTAHO.ORG/DOCUMENTATION/WORKBENCH.PHP

[5] HTTP://ES.WIKIPEDIA.ORG/WIKI/POSTGRESQL

[6] -WWW.AUTENTIA.COM FECHA DE CREACIÓN DEL TUTORIAL: 2007-10-30

TUTORIALES EN ADICTOSALTRABAJO: JAVA, J2EE, VISUAL C+...

HTTP://WWW.ADICTOSALTRABAJO.COM/TUTORIALES/TUTORIALES

BIBLIOGRAFÍA

Data Warehousing: Investigación y Sistematización de Conceptos –Hefesto: Metodología propia para la Construcción de un Data Warehouse. Ing. Bernabue, Ricardo Dario, Córdoba, Argentina. [En línea] Miércoles 07 de Noviembre del 2007.

HTTP://WWW.TARINGA.NET/POSTS/DOWNLOADS/1877926/ENTERPRISE-ARCHITECT-7_1.HTML [EN LÍNEA]

HTTP://MONDRIAN.PENTAHO.ORG/DOCUMENTATION/WORKBENCH.PHP . [En línea]

HTTP://ES.WIKIPEDIA.ORG/WIKI/POSTGRESQL . [En línea]

FECHA DE CREACIÓN DEL TUTORIAL: 2007-10-30 TUTORIALES EN ADICTOSALTRABAJO: JAVA, J2EE, VISUAL C+... [En línea] 2007 *WWW.AUTENTIA.COM*

HTTP://WWW.ADICTOSALTRABAJO.COM/TUTORIALES/TUTORIALES . [En línea]

Inmon, B. Building Data Warehouse. [En línea] 2002

Kimball, R. The Data Warehouse ETL Toolkit. [En línea] 2002

Kimball, R. The Data Warehouse Toolkit. [En línea] 2002

Bernabue, I. R. D. Data Warehousing-Hefesto. Córdoba, Argentina. [En línea] 2007

Ayala, A. P. Inteligencia de Negocios: Una Propuesta para su Desarrollo en las Organizaciones. Mexico. [En línea] 2006

Pujol, J. C. Introducción a Data Warehousing. Habana, Cuba. [En línea]

Ralph Kimball Associates and R.Kimball University. R.Kimball. [En línea] http://www.ralph.kimball.com/

Inmon.com LLC. B. Inmon. [En línea] http://www.billinmon.com/

OLAP Council. [En línea] http://www.olapcouncil.org/

Hyperion Solutions. Antiguo Arbor Software. E. Cood. [En línea] http://www.hyperion.com/

The Data Warehouse Institute. [En línea] http://www.dw-institute.com/

The Standford University Database group. [En línea] http://www-db.standford.edu/

The Database Group. University of Erlanger-Nunberg. [En línea] http://www6.informatik.uni-erlangen.de/

Building A Data Warehouse for Decision Support. Poe, V., Klauer, P., Brobst, S. Ed. Prentice-Hall. [En línea] 1998

OLAP Solutions. .Thomsen. E. Ed. Jhon-Wiley. [En línea] 2002

ANEXOS

Descripción de las tablas

Se describen a continuación algunas de las tablas más significativas de la Base de Datos del módulo Análisis Químico donde se van a seguir los pasos de la metodología Hefesto para realizar el diseño del DW.

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
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

Anexo 1: Tabla que recoge los datos de los lotes.

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

Anexo 2: Tabla que recoge los datos de las Planillas de Resultados.

Nombre: solucion		
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

Anexo 3: Tabla que recoge los datos de las soluciones.

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

Anexo 4: Tabla que recoge los datos de las Curvas de Calibración aplicadas a las soluciones.

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

Anexo 5: Tabla que recoge los nombres de las técnicas que se aplican en el laboratorio con su PPO correspondiente.

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

Anexo 6: Tabla que recoge los datos de los ensayos.

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

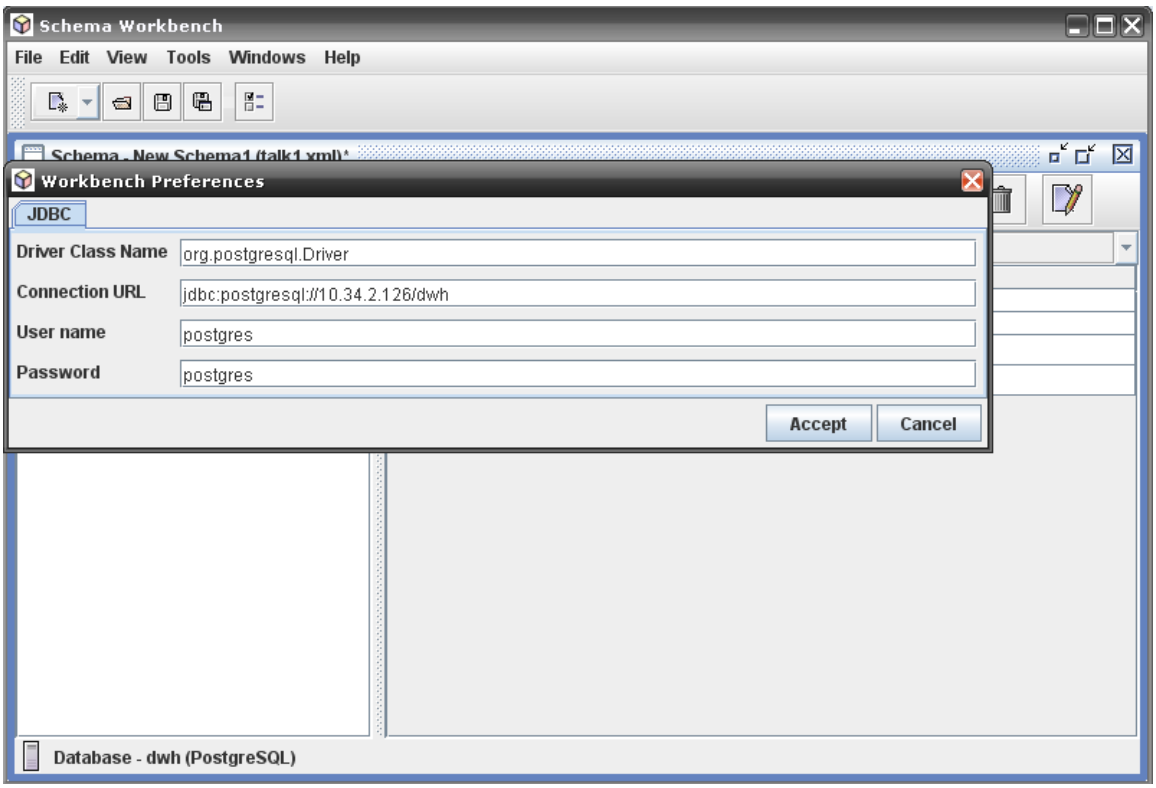
revisado_por	varchar(25)	Nombre de la persona que revisó el registro
fecha _realizacion	Date	fecha de realización del registro
terminado	boolean	Puede ser Sí o No

Anexo 7: Tabla que recoge los datos del Registro de Preparación de Soluciones.

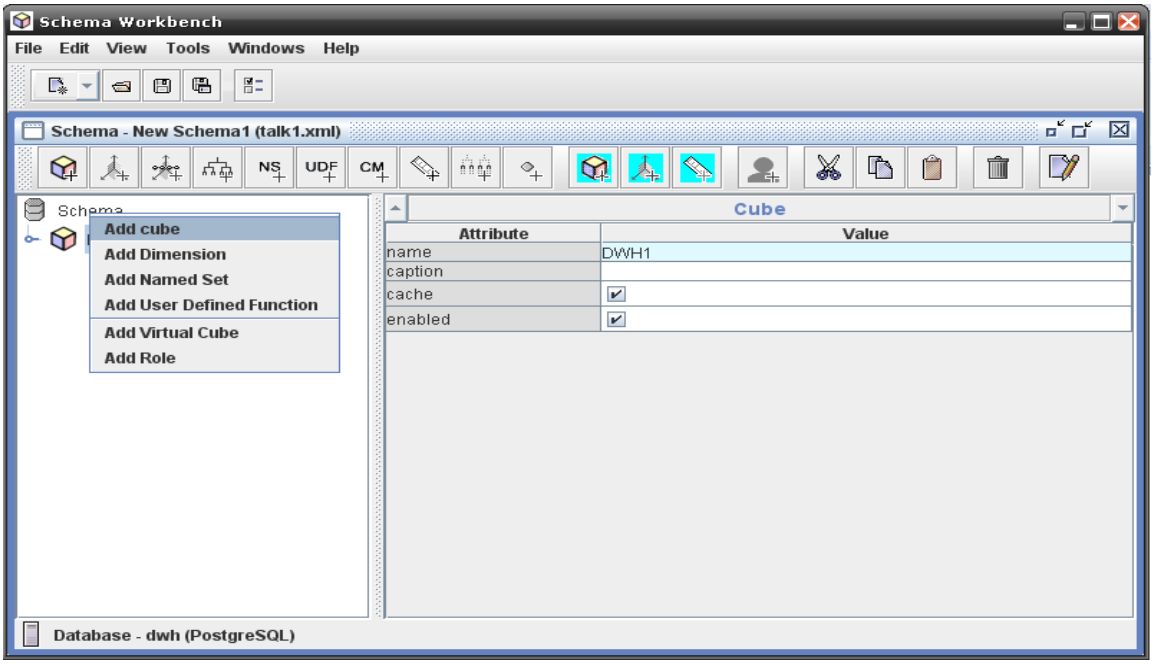
Nombre: registro_reensayo		
Descripción: Tabla que recoge los datos de Registros de Reensayos		
Atributo	**Tipo**	**Descripción**
folio_registro	varchar(15)	Número de Folio del registro, llave primaria
id_sic0837	integer	Llave foránea que la toma de la entidad sic0837_analisis_resultados_fe
id_tecnica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

Anexo 8: Tabla que recoge los datos de Registros de Reensayos.

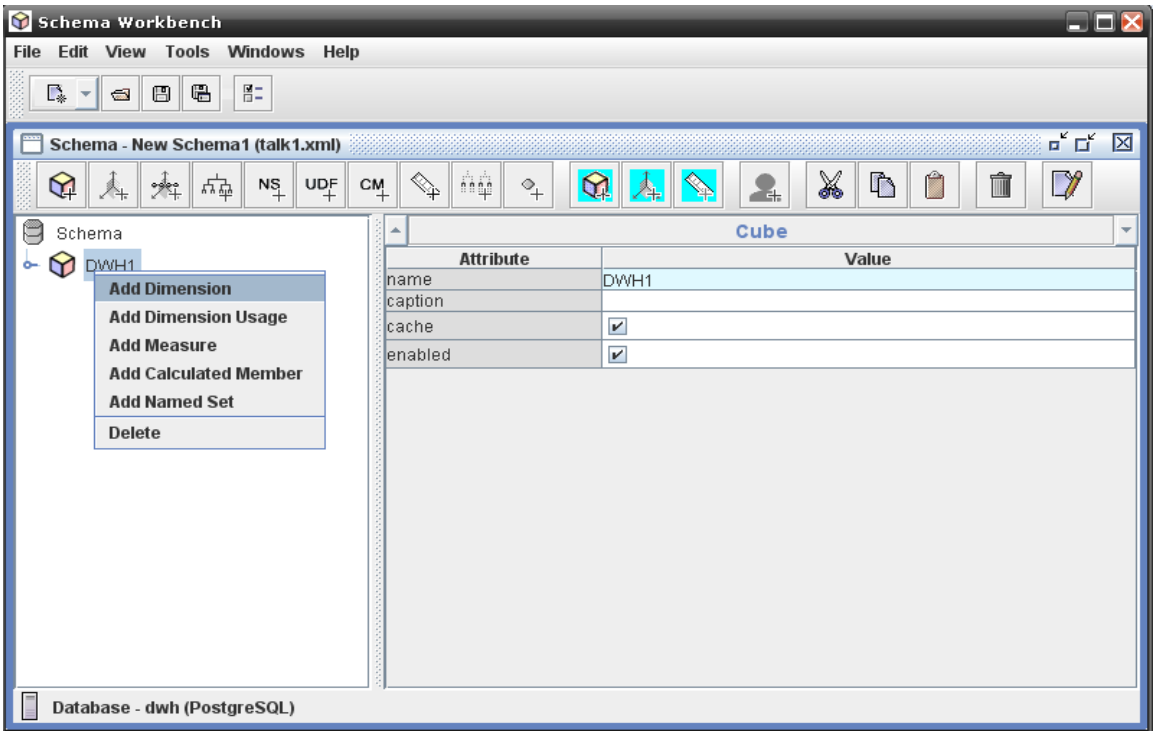
Conexión de la herramienta con la Base de Datos donde está montado el Data WareHouse, el anexo 9 detalla la conexión.



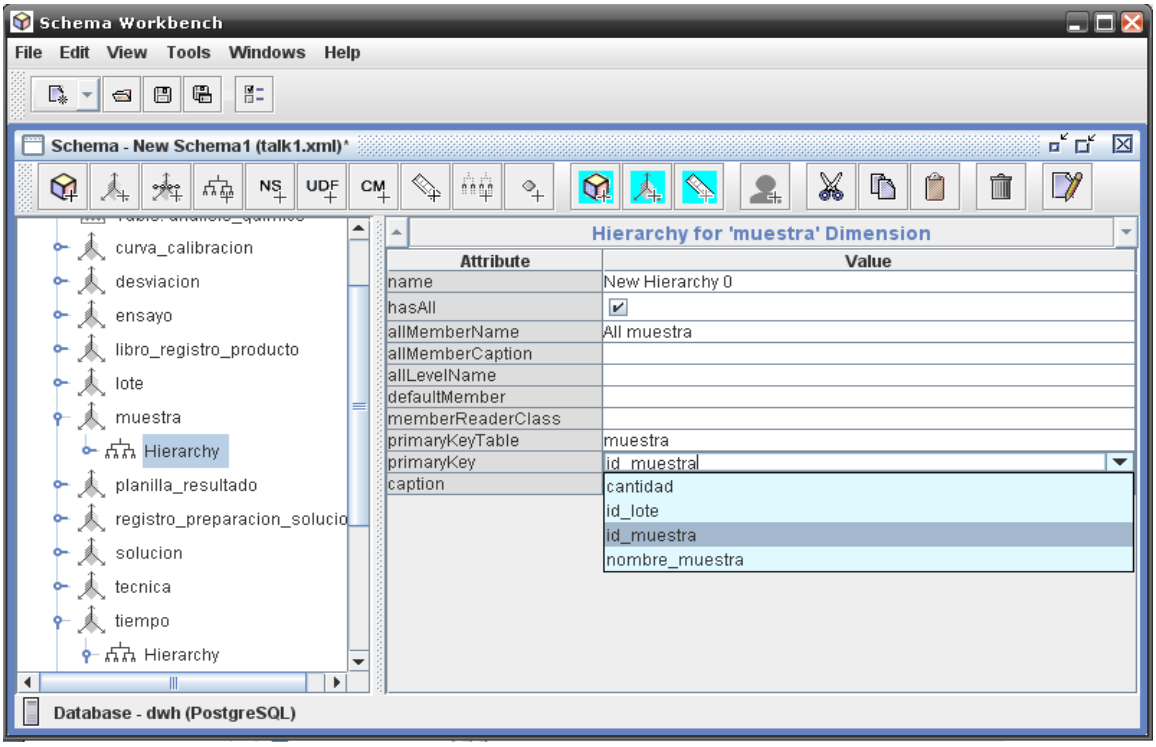
Anexo 9: Conexión de la herramienta Pentaho Schema Workbench con la base de datos.



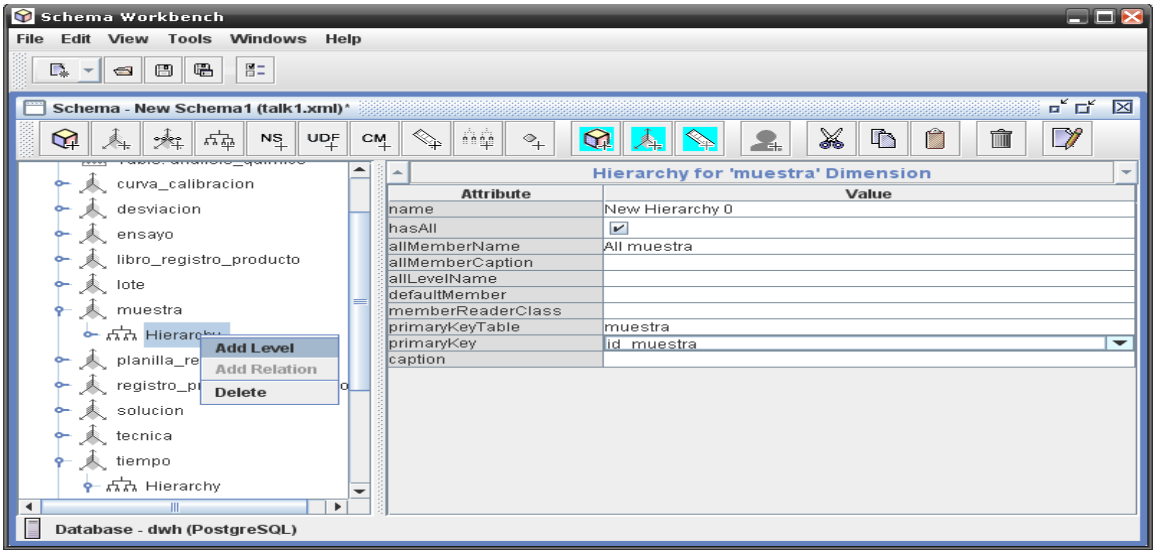
Anexo 10: Realización del cubo.



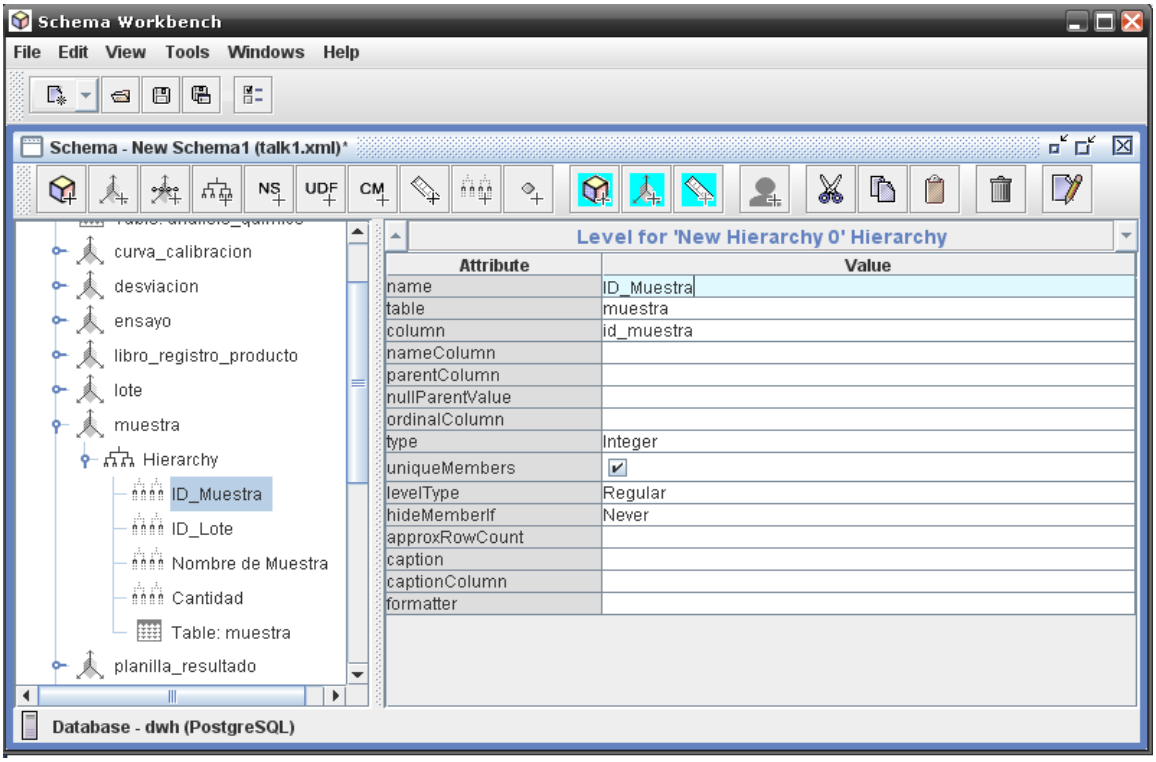
Anexo 11: Realización de la dimensión.



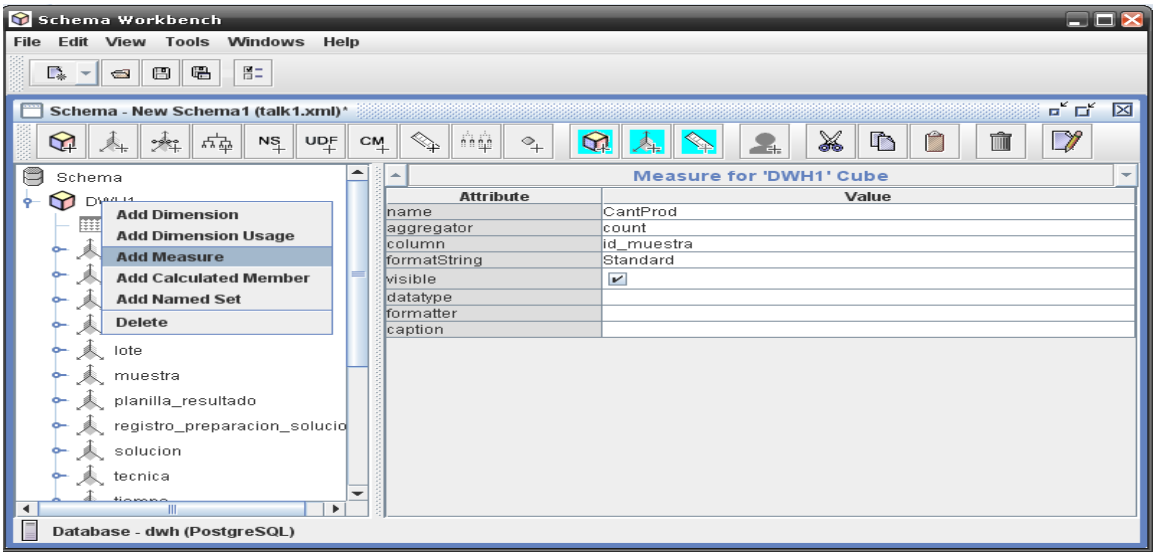
Anexo 12: Diseño de las jerarquía.



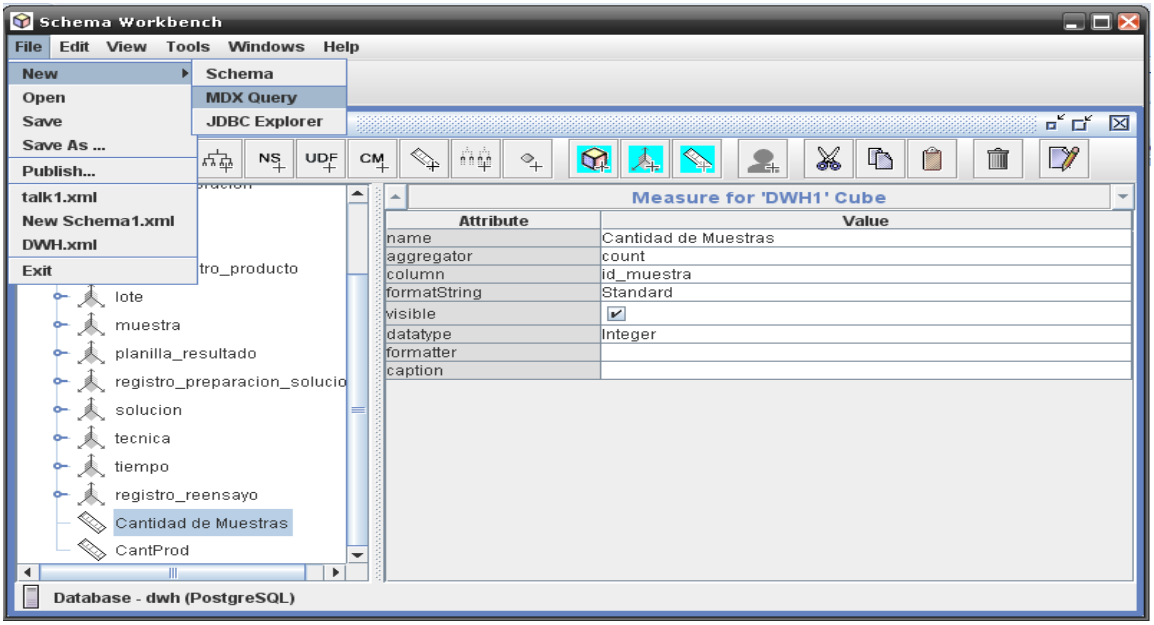
Anexo 13: Diseño de los niveles de las jerarquías.



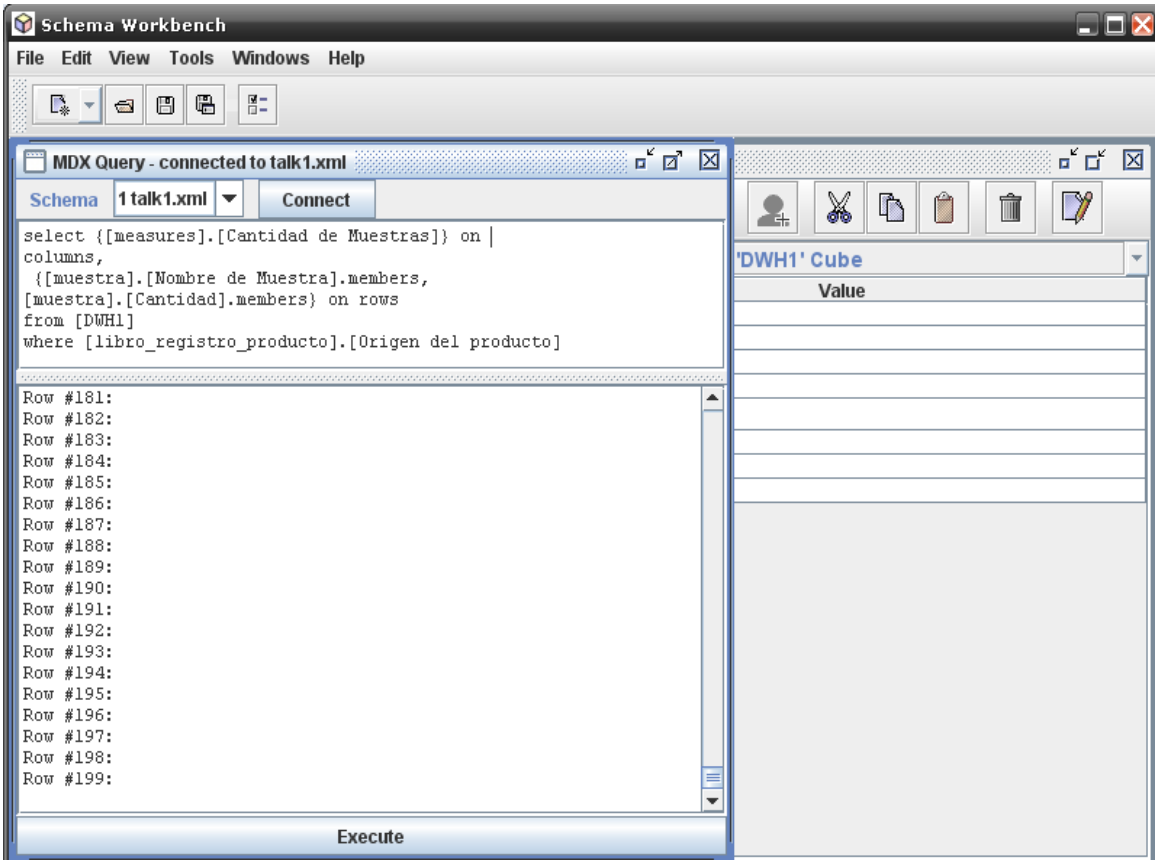
Anexo 14: Diseño de los niveles de las jerarquías



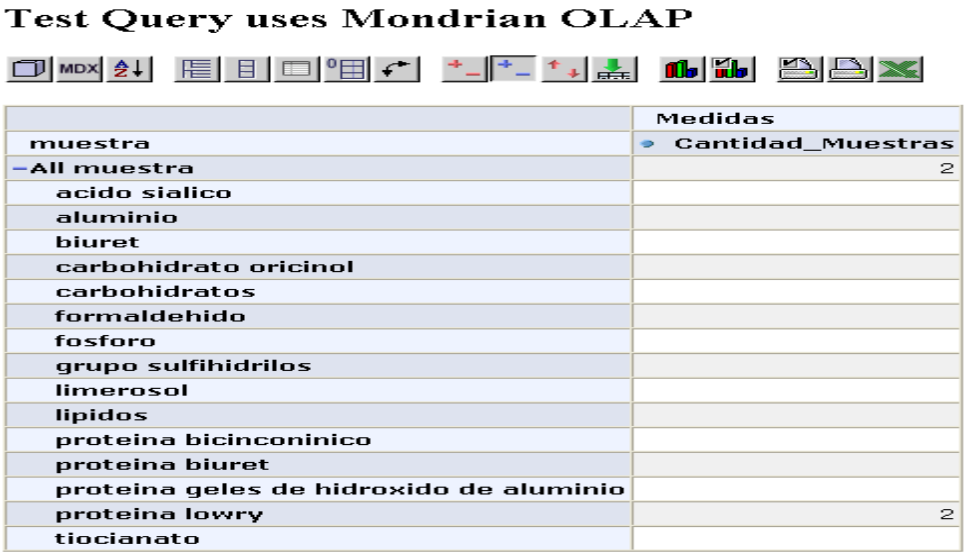
Anexo 15: Diseño de la medida.



Anexo 16: Adicionar consulta.

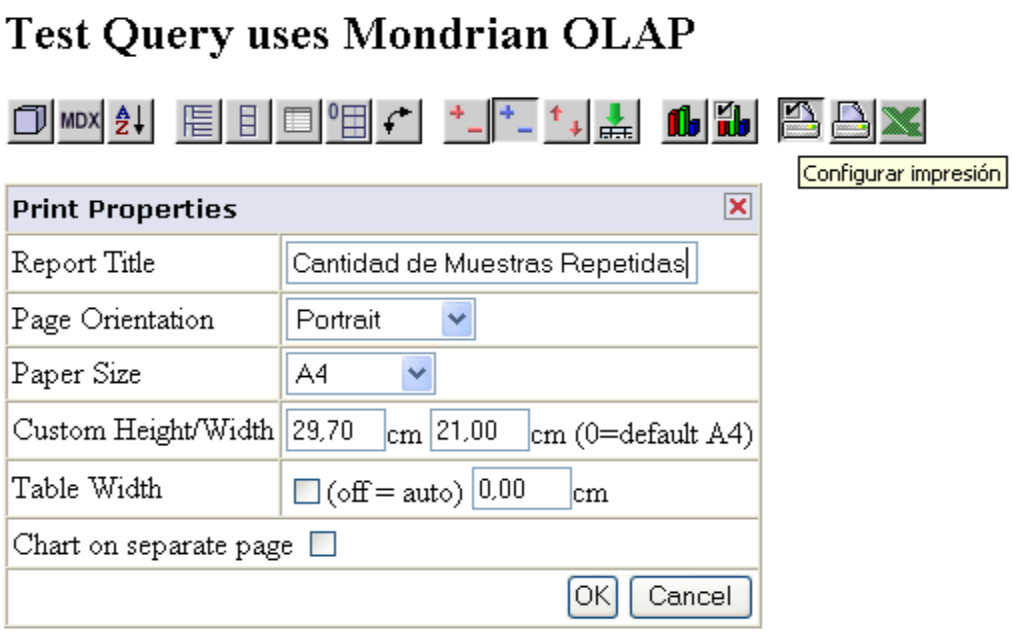


Anexo 17: Ejemplo de consulta en ejecución.

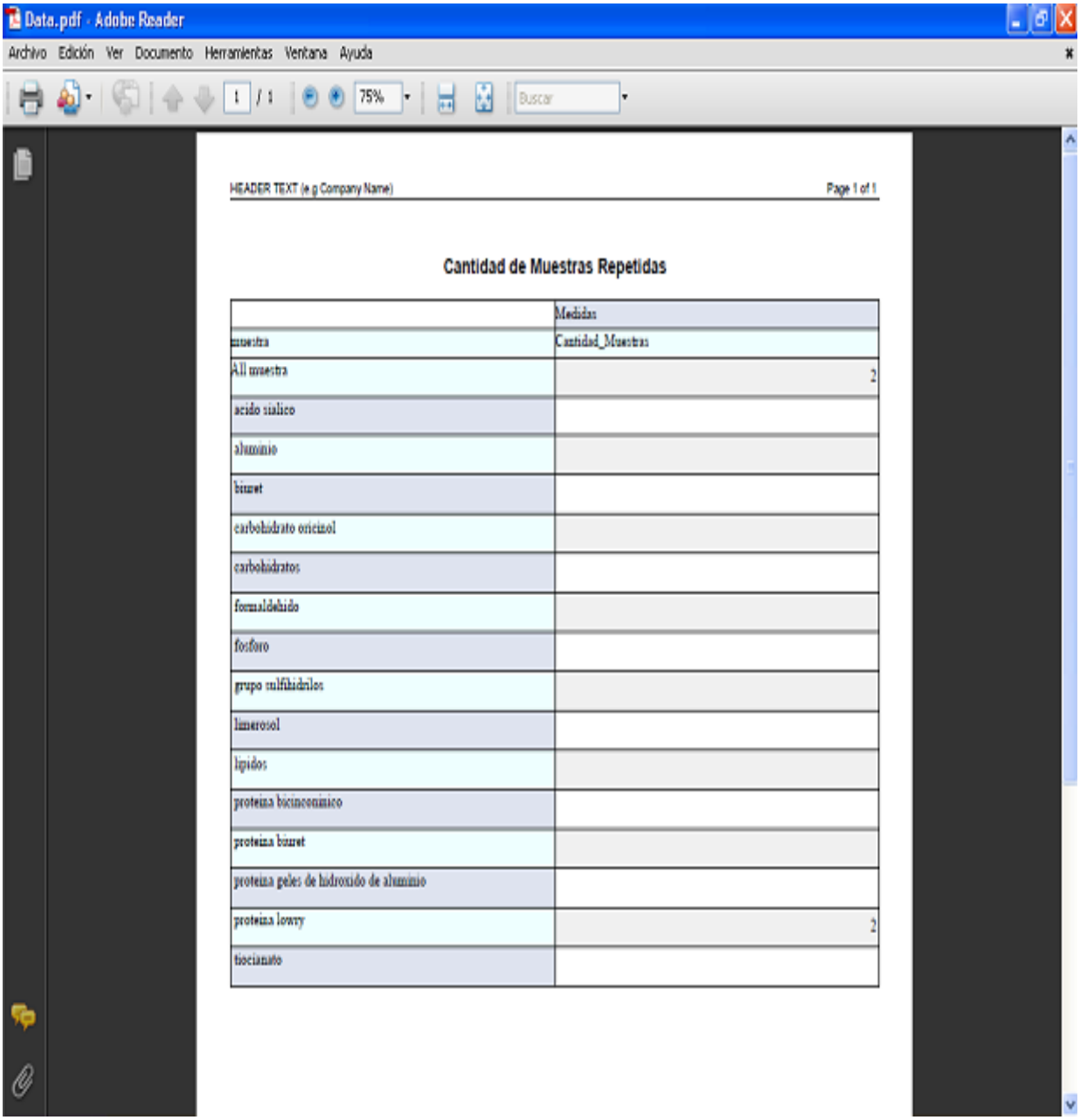


	Medidas
muestra	Cantidad_Muestras
−All muestra	2
acido sialico	
aluminio	
biuret	
carbohidrato oricinol	
carbohidratos	
formaldehido	
fosforo	
grupo sulfihidrilos	
limerosol	
lipidos	
proteina bicinconinico	
proteina biuret	
proteina geles de hidroxido de aluminio	
proteina lowry	2
tiocianato	

Anexo 18: Resultado una consulta visualizada en la herramienta OLAP Mondrian.



Anexo 19: Configuración de Impresión.



Anexo 20: Resultado de la configuración de impresión en un PDF generado.

GLOSARIO DE TÉRMINOS

***DW* --** DATA WAREHOUSE.

OLAP-- PROCESAMIENTO ANALÍTICO EN LÍNEA.

MOLAP-- PROCESAMIENTO ANALÍTICO EN LÍNEA MULTIDIMENSIONAL.

ROLAP-- PROCESAMIENTO ANALÍTICO EN LÍNEA RELACIONAL.

HOLAP PROCESAMIENTO ANALÍTICO EN LÍNEA HÍBRIDO.

OLTP-- SISTEMAS OPERACIONALES.

ETL -- EXTRACCIÓN , TRANSFORMACIÓN Y CARGA.

CIGB-- CENTRO DE INGENIERÍA GENÉTICA Y BIOTECNOLOGÍA.

SGBD-- SISTEMA DE GESTIÓN DE BASE DE DATOS.

EA-- ENTERPRISE ARCHITECT.

CVS-- CONCURRENT VERSIONS SYSTEM (SISTEMA DE CONTROL DE VERSIONES).

SCC-- SOURCE CODE CONTROL.

WYSIWYG-- WHAT YOU SEE IS WHAT YOU GET *(LO QUE VES ES LO QUE OBTIENES)*.

MDA --ARQUITECTURA AVANZADA DIRIGIDA POR MODELOS.

MIP-- MODELOS INDEPENDIENTES DE PLATAFORMA.

PGDG --COMUNIDAD DE DESARROLLADORES Y ORGANIZACIONES COMERCIALES.

MVCC-- ACCESO CONCURRENTE MULTIVERSIÓN.

MDX-- LENGUAJE DE EXPRESIONES MULTIDIMENSIONALES.

MPL-- MOZILLA PUBLIC LICENSE.

EIS--Executive Information System (Sistema de Información Ejecutiva).

JSP-- Java Server Pages.