Universidad de las Ciencias Informáticas

Modelo de descripción de arquitectura de almacenes de datos para ensayos clínicos del Centro de Inmunología Molecular.

Trabajo para optar por la Maestría en Informática Aplicada

Autor: Ing.Anthony Rafael Sotolongo León

Tutor: Msc Maikel Yelandi Leyva Vázquez

Ciudad de la Habana

Junio de 2010

Todo hombre debe decidir una vez en la vida

si se lanza a triunfar arriesgándolo todo

o se sienta a contemplar el paso de los triunfadores

Benedetti

Agradecimientos

Quisiera agradecer a todas las personas que han contribuido con su ayuda directa e indirecta con la realización de esta investigación.

A todos los profesores que de una forma u otra participaron en mi formación como Máster en informática aplicada, además de los que pusieron su granito de arena para la realización de este documento. A mis tesistas de este curso que llevaron a cabo una tarea bien difícil, a mi compañera de trabajo Martha. Y a mi tutor que me ha orientado en el momento justo.

También agradezco a mi novia y mi suegra que ayudaron en la redacción de este documento.

Dedicatoria

Este trabajo está dedicado a ***mi mamá, mi papá y a mi hermano*** *que han seguido mi formación desde que era estudiante hasta ahora que soy un profesional, desde los inicios me han dado todo su apoyo y orientación para llegar a esta meta.*

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

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

_____

Ing Anthony Rafael Sotolongo León

_____

M.Sc. Maikel Yelandi Leyva Vázquez

Resumen

La descripción correcta y detallada de la arquitectura de los sistemas informáticos es muy importante para lograr éxito en el desarrollo de los mismos. Los almacenes de datos como solución informática que apoya la toma de decisiones en las entidades que los implementan también necesitan una descripción de la arquitectura detallada. Ralph Kimball propone los aspectos a tenerse en cuenta para la descripción pero no expone con que realizarla. Existen modelos específicos para describir la arquitectura como es el 4+1 vistas de Kruchten o el meta-modelo Common Warehouse Metamodel (CWM) pero no se ajustan a las necesidades de descripción que requiere un almacén de datos que integre información de ensayos clínicos del Centro de Inmunología Molecular (CIM). En esta investigación se propone un modelo basado en vistas para describir la arquitectura de almacenes de datos que se ajustan a las necesidades del Centro de Inmunología Molecular siguiendo el marco de referencia de Kimball y usando como lenguaje de modelado UML 2.0, se realiza también la aplicación del modelo y se detalla la implementación del Data Mart del producto hR3.

.

PALABRAS CLAVES: Arquitectura de Almacenes de datos, Vistas arquitectónicas, Ensayos Clínicos

ÍNDICE

INTRODUCCIÓN 3

CAPÍTULO 1: MARCO TEÓRICO 9

1.1. Almacenes de Datos 9
1.1.1. Definición de almacenes de datos 9
1.1.2. Conceptos importantes dentro de almacenes de datos 10
1.1.3. Ventajas y Desventajas de los almacenes de datos 16
1.1.5 Ejemplo de aplicaciones de los almacenes de datos. 16
1.2 Gestión de los Ensayos Clínicos en el CIM 17
1.3 Arquitectura de software. 18
1.4 Arquitectura de almacenes de datos. 19
1.5 Modelos de descripción de arquitectura. 22
1.5.1 El modelo 4+1 vistas de Krutchen 22
1.5.2 CWM 23
1.6 Lenguajes para el modelado de la arquitectura 24
1.7 Conclusiones parciales 26
CAPÍTULO 2: DESCRIPCIÓN DEL MODELO 27

2.1 Modelo para describir arquitectura en almacenes de datos. 27
2.1.1 Propuestas de vistas para la descripción de arquitectura de almacenes de datos 27
2.3 Conclusiones parciales 36
CAPITULO 3: COMPARACION Y APLICACIÓN DEL MODELO 37

3.1 Comparación y evaluación del modelo. 37
3.2 Propuesta de arquitectura para almacenes de datos de ensayos clínicos del Centro de Inmunología Molecular 39
3.2.1Características fundamentales que debe poseer la arquitectura del almacén de datos. 39
3.2.2Descripción de la arquitectura. 39
3.3 Implementación del almacén de datos de ensayos clínicos del producto hR3 del CIM. 50
3.4 Conclusiones parciales 57
CONCLUSIONES 58

RECOMENDACIONES 59

REFERENCIAS BIBLIOGRÁFICAS 60

BIBLIOGRAFÍA 61

ANEXO 1 64

GLOSARIO DE TÉRMINOS 65

INTRODUCCIÓN

En la actualidad la humanidad enfrenta diversas enfermedades que constituyen la causa de muerte de millones de personas, el número de decesos por enfermedades crónicas es muy alto en el todo el mundo. Pero a la par de esta situación los científicos están realizando investigaciones e inventando medicamentos para combatirlas.

En Cuba se funda de manera oficial el 5 de diciembre de 1994 el Centro de Inmunología Molecular (CIM) con el motivo de estudiar el comportamiento de estas enfermedades y crear biofármacos que puedan combatirlas, y de esta forma lograr un mejoramiento en la calidad de vida del pueblo. El CIM tiene como principal misión obtener y producir nuevos medicamentos destinados al tratamiento del cáncer y otras enfermedades crónicas no transmisibles e introducirlos en la Salud Pública cubana. Hacer la actividad científica y productiva económicamente sostenible y realizar aportes importantes a la economía del país [1].
Entre los productos que se desarrollan en el centro se puede mencionar el anticuerpo monoclonal anti CD3 para el tratamiento de pacientes con rechazo del trasplante de órganos, Eritropoyetina humana recombinante para el tratamiento de la anemia, Factor Estimulante de Colonias granulocíticas para el tratamiento de la Neutropenia, anticuerpo monoclonal “humanizado” que reconoce el receptor del Factor de Crecimiento Epidérmico (EGF-R) para el tratamiento del cáncer, así como otros anticuerpos para el estudio *in vivo* por inmunogammagrafía de pacientes con cáncer.

Lo que le ha dado un prestigio a nivel mundial en la producción de fármacos para combatir enfermedades cancerígenas y poniendo a Cuba en la vanguardia de la investigación farmacéutica a nivel mundial.
Las líneas principales de investigación están concentradas en la inmunoterapia del cáncer, especialmente en el desarrollo de "vacunas moleculares", ingeniería de anticuerpos, ingeniería celular, bioinformática y regulación de la respuesta inmune. Para el estudio y aprobación de los fármacos fabricados se realiza en varios hospitales del país y algunos en el exterior varios Ensayos Clínicos (EC).

Un Ensayo Clínico, es *un tipo de estudio clínico en el que se evalúan nuevos fármacos o tratamientos médicos a través de su aplicación a seres humanos [10].* Por tanto es un estudio experimental, analítico, prospectivo, controlado y con tamaños muestréales suficientes. Los cuales presentan un protocolo, documento que establece la razón de ser del estudio, sus objetivos, diseño, métodos y el análisis previsto de sus resultados, así como las condiciones bajo las que se realizará y desarrollará el estudio, también debe contemplar el acceso a los datos por la importancia y sensibilidad de los mismos de modo que este bien descrito como es la manipulación de esa información en todas las gestiones que se realicen sobre esos

datos. Por estas razones, los Ensayos Clínicos llevan asociados una documentación muy amplia, y esta información debe ser almacenada como mínimo 15 años posteriores al cierre del estudio.

Esta información es recogida en los Cuadernos de Recogida de Datos (CRD), *formulario diseñado para anotar las variables recogidas durante un ensayo clínico [2],* este formulario es diseñado de modo diferente en dependencia de lo que se quiera lograr, por tanto no tienen porque coincidir las variables, la información recogida puede llegar a ser de hasta 1000 variables por cada CRD.

Para comprender, analizar y tomar decisiones de toda esta información relacionada con los productos, el CIM almacena los datos de cada ensayo clínico por separado teniendo en cuenta que este análisis requiere la interrelación de las estadísticas de cada ensayo. Este es el punto crítico y se hace necesario integrar información de las diferentes fuentes disponibles de cada ensayo clínico relacionado. Para esto se ha trazado una estrategia de utilizar almacenes de datos por todas las ventajas que tienen los mismos para la toma de decisiones. Pero la entidad no cuenta con una arquitectura ni soporte tecnológico para la implementación del mismo. Siendo la arquitectura un punto clave en el desarrollo de los almacenes de datos, la definición de la misma cuenta con tres procesos fundamentales: diseño, descripción y evaluación. La arquitectura sino está bien descrita pierde su verdadera utilidad. Una excelente descripción de la arquitectura debe representar a todas las estructuras del sistema así como la interacción entre sus partes.
Ralph Kimball investigador destacado en la materia de almacenes de datos propone un marco de referencia donde expone los criterios a tenerse en cuenta para realizar la descripción de la arquitectura de un almacén de datos pero a su vez no define con que realizar esta descripción.

Una alternativa viable para describir la arquitectura son las vistas o modelos de vistas las cuales son las más utilizadas por la comunidad de arquitectos de software. Aun cuando existe un consenso general acerca de la necesidad de representar la arquitectura utilizando diferentes vistas, tal consenso desaparece cuando hay que definir cuáles son esas vistas. El estándar IEEE-1471-2000 [3] define de alguna manera los puntos de vista que deben tenerse en consideración para describir la arquitectura, pero a un nivel de abstracción tan alto que las vistas pueden seleccionarse o definirse siguiendo criterios muy diferentes en dependencia de las necesidades. Ejemplos de modelos que definen puntos de vista explícitos son el modelo de 4+1 vistas de Kruchten, los modelos de Sowa y Zachman y el modelo propuesto por Hofmeister, Soni y Nord. De todos estos modelos, el que ha conseguido mayor aceptación es el modelo de 4 + 1 vistas de Krutchen. [4].

Los almacenes de datos de ensayos clínicos tienen características diferentes a los sistemas que comúnmente se modelan con estas vistas. Incluso el modelo de 4+1 vistas no describe en su totalidad la

arquitectura necesaria para un almacén de datos de ensayos clínicos. Pues en estos sistemas la arquitectura gira alrededor de los requisitos y en los almacenes gira alrededor de los datos. También existe el Common Warehouse Metamodel (CWM) que no es más que un meta-modelo para concebir almacenes de datos, el mismo no realiza la descripción basada en vistas sino por escenarios tales como: ETL, OLAP, Cuestionarios, Administración y Herramientas. Pero este meta-modelo ignora los aspectos relacionados con el hardware, tema esencial para la descripción de la arquitectura propuesto por Kimball y tampoco contempla del todo el acceso a datos.

Después de analizar la problemática, se identifica el **problema científico:** Dificultad para describir la arquitectura de sistemas de información basada en almacenes de datos para ensayos clínicos del CIM.

Se plantea como **objeto de estudio** arquitectura de los almacenes de datos y **campo de acción** proceso de descripción de arquitectura en almacenes de datos.

Se identifica como **objetivo general**: Definir un modelo para la descripción de la arquitectura de sistemas de información basada en de almacenes de datos para ensayos clínicos, que contribuya a la construcción de almacenes de datos en el Centro de Inmunología Molecular.

Este objetivo se desglosa en los siguientes **objetivos específicos**:

- ✓ Elaborar el marco teórico de la investigación.
- ✓ Definir el modelo para describir la arquitectura de almacenes de datos en ensayos clínicos.
- ✓ Aplicar el modelo propuesto en un proyecto real de ensayos clínicos.

Para cumplir los objetivos se plantean las **tareas de la investigación** siguientes:

- ✓ Revisión bibliográfica sobre almacenes de datos y arquitectura de los mismos.
- ✓ Estudio de los modelos basados en vistas para modelar arquitectura de sistemas.
- ✓ Entrevistas con los especialistas del centro de inmunología molecular.
- ✓ Definición de vistas para modelar arquitectura de almacenes de datos.
- ✓ Comparación con los modelos existentes de descripción de arquitectura.
- ✓ Modelación de una arquitectura que integre los datos de los ensayos clínicos que se gestionan en el Centro de Inmunología Molecular.
- ✓ Implementación de la arquitectura modelada.

Hipótesis:

Si se define un modelo para describir la arquitectura de sistemas de información basada en almacenes de datos de ensayos clínicos entonces se facilitará la descripción de los almacenes de datos en el Centro de Inmunología Molecular.

Novedad y aporte teórico:

- Concepción de un modelo para el proceso de descripción de arquitectura de almacenes de datos de ensayos clínicos de CIM basado en UML 2.0
- Desarrollo de una vista que describe el acceso a datos por usuarios.

Novedad y aporte práctico:

- Arquitectura para almacenes de datos de ensayos clínicos del CIM
- Soporte tecnológico para almacenes de datos de ensayos clínicos del CIM
- La implementación de un almacén de datos que contiene la información de los ensayos clínicos del producto hR3 de manera integrada y estandarizada.

Importancia y actualidad de la investigación

La importancia que tiene el desarrollo de esta investigación, está aras de resolver los problemas integración de datos de ensayos clínicos en un almacén de datos además de la gestión de todo el análisis que llevan los mismos pues la información asociada al proceso es muy importante en la validación de los fármacos desarrollados por el CIM.

El almacén de datos soportará cualquier tipo de ensayo clínico desarrollado en el centro pues se diseñará capaz de adicionarle en cualquier momento datos de diversos ensayos. Lo que quiere decir que quedará garantizado su incremento.

Impacto socioeconómico

Esta investigación tiene un impacto tecnológico-económico en Cuba, dada su influencia directa en el costo elevado de estos procesos de gestión de ensayos clínicos en el mercado. El resultado esperado con esta arquitectura, incide directamente en uno de los problemas que más afectan a los centros cubanos actuales del polo investigativo: el tiempo. Estos estudios de los ensayos clínicos normalmente lleva meses integrarlos

y contar con un modelo para describir la arquitectura para un almacén de datos representa un paso importante en el mejoramiento de la competitividad de la industria cubana de la fabricación de fármacos en el mercado mundial.

Durante el proceso de investigación se utilizaron fundamentalmente los siguientes **métodos**:
Inductivo-Deductivo: Posibilita el trabajo con las concepciones generales revisadas en la bibliografía, llevadas a casos particulares y elaboración de las conclusiones.

El análisis y la síntesis: El análisis permite descomponer el problema en diversas partes y cualidades, permite la división mental del todo en sus múltiples relaciones y componentes. La síntesis permite la unión entre las partes previamente analizadas y posibilita descubrir las relaciones esenciales y características generales entre ellas.

Observación: Se utiliza con el objetivo de analizar el comportamiento y utilización de la temática de estudio con el fin de constatar la situación actual de empleo.

La modelación: Es justamente el método mediante el cual se crean abstracciones con vistas a explicar la realidad. El modelo como sustituto del objeto de investigación se muestra como algo semejante a él, donde existe una correspondencia objetiva entre el modelo y el objeto.

Como técnica de recopilación de la información se utilizará **la entrevista** con la cual se obtendrá información valiosa sobre los principales procesos que se realizan acerca de tema que se investiga.

La investigación que a continuación se presenta se estructura en tres capítulos fundamentales:

El **Capítulo 1** contiene temas relacionados con el concepto de almacenes de datos, sus áreas de aplicación. Además, se presentan temas de arquitectura de software específicamente en almacenes de datos, se describen los modelos más utilizados para describir arquitectura. También se detalla el proceso de gestión de ensayos clínicos en el CIM.

El **Capítulo 2** se detalla el modelo de vistas para describir la arquitectura de almacenes de datos basándose en el marco de referencia de Kimball utilizando el estándar IEEE 1471.

El **Capítulo 3** se hace una comparación con algunos modelos existentes. Se describe el modelo propuesto aplicado en la arquitectura de almacenes de datos de ensayos clínicos del Centro de Inmunología Molecular. Y se detallan aspectos de la implementación del DM del producto hR3.

CAPÍTULO 1: MARCO TEÓRICO

1.1. Almacenes de Datos

1.1.1. Definición de almacenes de datos

Con la informatización de la sociedad, ha crecido la capacidad de generación y almacenamiento de la información, mientras mayor es la capacidad para almacenar, mayor es la incapacidad para extraer información realmente útil de éstos en las empresas o entidades. Mucha información importante queda normalmente sin descubrir puesto que el cúmulo de esta es enorme. Muchas son las ventajas que trae consigo esta nueva era de la Información en la cual aumentan las posibilidades de quienes tienen posibilidad de almacenar y analizar los datos. Sin embargo, esto puede tornarse engorroso y lento pues en ocasiones no se encuentra la mejor forma de explotación de los datos. Para resolver esta disyuntiva surge un término llamado almacén de datos.

Durante la revisión bibliográfica sobre la definición de almacenes de datos se encuentra que existen varios especialistas que emiten sus criterios los cuales se muestran a continuación se decidió analizar las definiciones de Inmon y Kimball por considerar que son los principales autores referentes a los almacenes de datos:

Listado de definiciones

1. Colección de datos orientada al negocio, integrada, variante en el tiempo y no volátil para el soporte del proceso de toma de decisiones de la gerencia. [5]
2. Colección de datos orientados a temas, integrada, variante en el tiempo, no volátil, que añade la geografía del dato. [6]

Luego de un análisis de las definiciones anteriores se considera lo siguiente:
Características principales de almacenes de datos:

- Orientado por temas: orientado a un tema específico de la empresa. Ejemplo; compras, ventas, etc.

- Integrados: Datos integrados de distintas fuentes.
- No volátiles: los datos almacenados no se modifican.
- Variables en el tiempo: datos relativos a un período de tiempo y se incrementan periódicamente

1.1.2. Conceptos importantes dentro de almacenes de datos

Modelo dimensional

El modelado dimensional es una técnica de diseño lógico que busca presentar los datos en un estándar y facilitar una recuperación adecuada de estos. Los datos son almacenados como hechos y dimensiones en un modelo de datos relacional. Para poder entender la definición presentada así como el modelo multidimensional, se deben comprender tres conceptos fundamentales: hechos, dimensiones y atributos.
Se llama hecho a una operación que se realiza en el negocio la cual está estrechamente relacionada con el tiempo y es objeto de análisis para la toma de decisiones. También puede verse como un valor numérico que representa una actividad específica casi siempre con cifras que se suman entre sí.
Se conoce como dimensión a la característica de un hecho que permite su análisis posterior en el proceso de toma de decisiones y brinda una perspectiva adicional a un hecho dado. Son agrupaciones lógicas de atributos con un significado común y atómico [7]. Generalmente se puede expresar su representación por los siguientes esquemas:

El esquema estrella (figura 1): deriva su nombre del hecho que su diagrama forma una estrella, con puntos radiales desde el centro. El centro de la estrella consiste de una tabla de hechos y las puntas de la estrella son las tablas dimensiones [8].

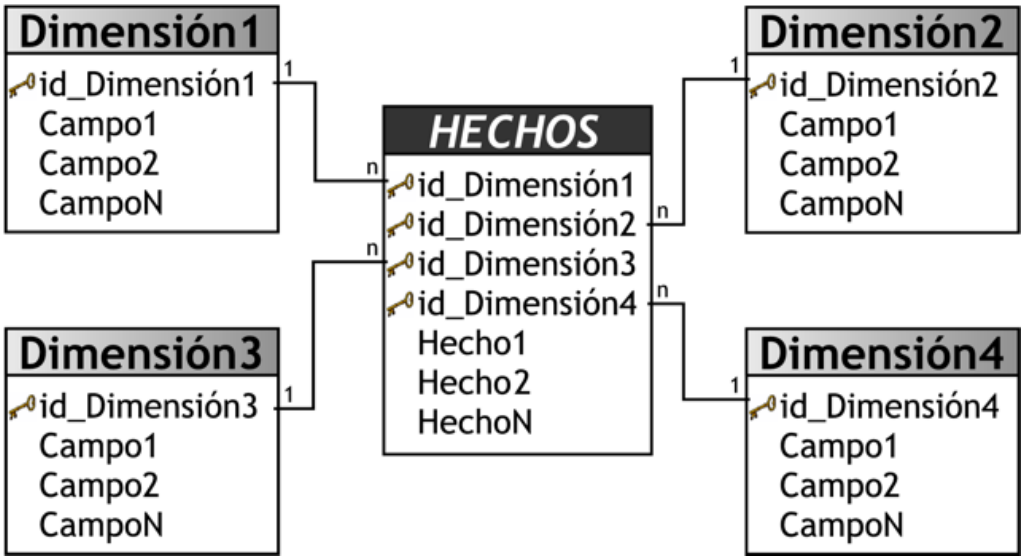

Figura 1. Esquema de estrella

Esquema de copo de nieve (figura 2): una estructura algo más compleja que el esquema en estrella. Se da cuando alguna de las dimensiones se implementa con más de una tabla de datos es decir una especie de normalización de las dimensiones [8]

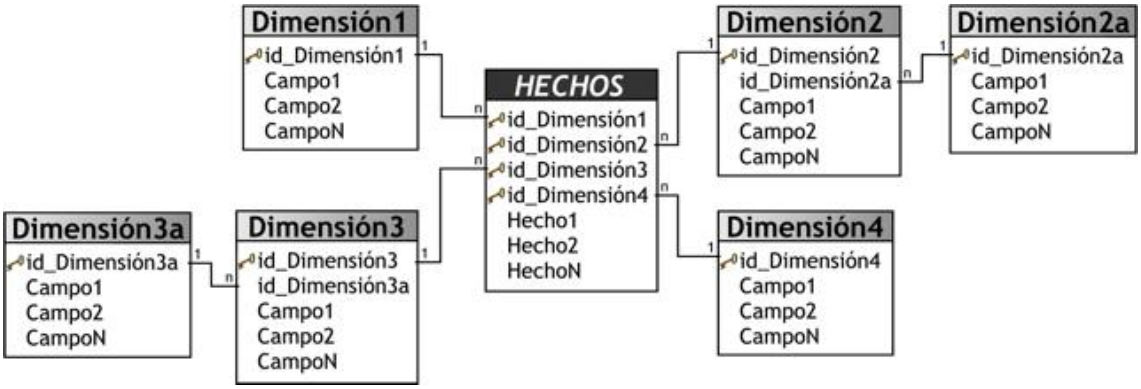

Figura 2. Esquema de Copo de nieve

Esquema de constelación de hechos (figura 3): Este esquema es más complejo que los anteriores debido al facto de que contiene múltiples tablas de hechos que comparten varias dimensiones [8].

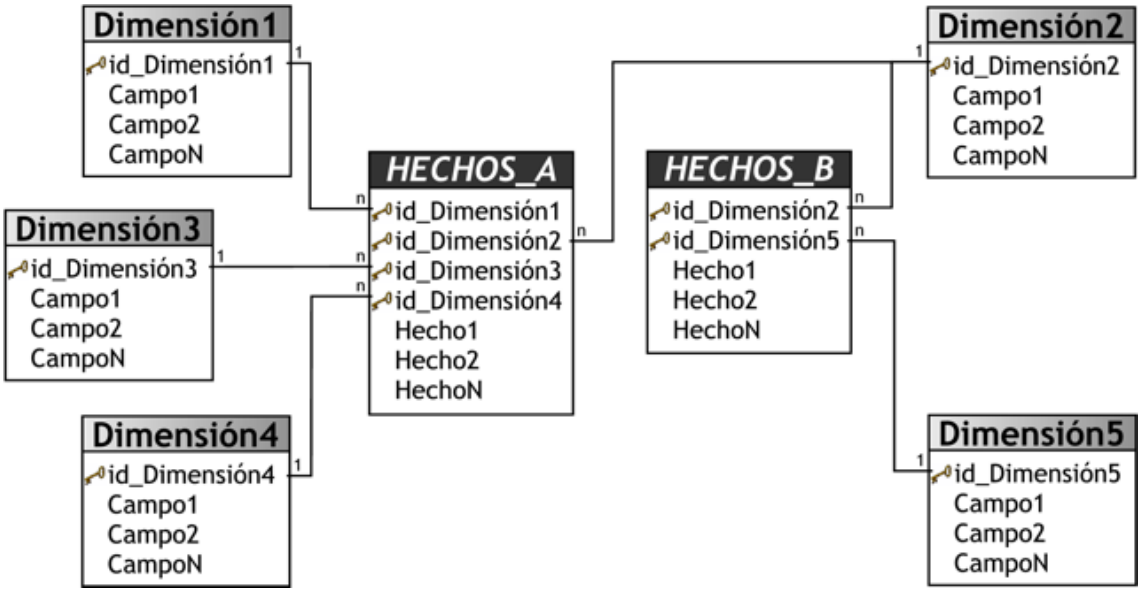

Figura 3.Constelación de hechos

Data Mart (DM)

Es un subconjunto de los datos de un almacén de datos, normalmente en la forma de información resumida que soporta los requerimientos de un departamento o función de negocio particular [8].

Existen varios enfoques para construir DM. Un enfoque es construir un almacén de datos corporativo que puede ser usado directamente por los usuarios y proveer los datos para otros DM (enfoque descendente, Inmon). Otro enfoque es construir varios DM con una vista para la virtual integración en un almacén, y el enfoque final es construir la infraestructura para un almacén corporativo mientras al mismo tiempo se construye uno o más DM para satisfacer las necesidades inmediatas del negocio (enfoque ascendente, Kimball).

Diferencias entre DM y Datawarehouse (DWH)

- El DM se centra solamente en los requerimientos de usuarios asociados con un departamento o función de negocio
- Los DM normalmente no contienen datos operacionales detallados a diferencia de almacén.
- Debido a que los DM contienen menos información comparados con los almacenes, los DM son más fácilmente entendibles y navegables.

Stagingarea

El *stagingarea* es un área temporal donde se recogen los datos que se necesitan de los sistemas operacionales de origen. Se recogen los datos estrictamente necesarios para las cargas, y se aplica el mínimo de transformaciones a los mismos. No se aplican restricciones de integridad ni se utilizan claves, los datos se tratan como si las tablas fueran ficheros planos. De esta manera se minimiza la afectación a los sistemas origen, haciendo la carga lo más rápida posible y se reduce también al mínimo la posibilidad de error. [8]

Metadatos

Los metadatos son datos que describen o dan información de otros datos, que en este caso, existen en la arquitectura del DWH. Brindan información de localización, estructura y significado de los datos, básicamente mapean los mismos. El concepto de metadatos es análogo al uso de índices para localizar objetos en lugar de datos. La gran ventaja que trae aparejada el DWH en relación con los metadatos es que el usuario puede gestionarlos, exportarlos, importarlos, realizarles mantenimiento e interactuar con ellos, ya sea manual o automáticamente [8]

OLAP

OLAP (*Online Transaction Processing*, OLTP por sus siglas en inglés) es la tecnología que permite que la información sea vista multidimensionalmente, a través de cubos con categorías descriptivas (dimensiones) y valores cuantitativos (medidas).

Permite un uso más eficaz de los almacenes de datos para el análisis en línea, lo que proporciona respuestas rápidas a consultas analíticas, complejas e iterativas. Los modelos de datos multidimensionales de OLAP organizan y resumen grandes cantidades de datos para que puedan ser evaluados con rapidez y variedad, usando herramientas gráficas. La respuesta a una consulta realizada sobre datos históricos, a menudo suele conducir a consultas posteriores en las que el analista busca respuestas más concretas o explora posibilidades. Los sistemas OLAP proporcionan la velocidad y la flexibilidad necesarias para dar apoyo al analista en tiempo real [8].

Proporciona muchas ventajas a los usuarios que realizan análisis; por ejemplo:

- ✓ Un modelo de datos intuitivo y multidimensional que facilita la selección, recorrido y exploración de los datos.

- ✓ Un lenguaje analítico de consulta que proporciona la capacidad de explorar las complejas relaciones existentes entre los datos empresariales.
- ✓ Un pre cálculo de los datos consultados con más frecuencia que permite una rápida respuesta a las consultas ad hoc (*ad hoc query*). El uso de estas consultas, a diferencias de las generadas mediante el Lenguaje de Consultas Estructurado (*Structural Query Languaje*, SQL por sus siglas en inglés), implica que el sistema permita al usuario personalizar una consulta en tiempo real.
- ✓ Soporta análisis complejos de grandes volúmenes de datos.
- ✓ Permite a los usuarios, analizar la información, basándose en más criterios que en un análisis de forma tradicional.
- ✓ Al contar con muestras grandes, se pueden explorar mejor los datos en busca de respuestas.
- ✓ Se puede analizar el negocio desde diferentes escenarios históricos, y proyectar cómo se ha venido comportando y evolucionando en un ambiente multidimensional, o sea, mediante la combinación de diferentes perspectivas, temas de interés o dimensiones.

Se pueden definir varios sistemas OLAP, dependiendo de las técnicas que se utilicen a la hora de obtener los datos y la forma en la que están estructurados y almacenados. Para implementar estos sistemas OLAP se pueden utilizar herramientas informáticas que permitirán realizar análisis de los datos, entre los principales sistemas OLAP se tienen [9]:

- ✓ **ROLAP (R**elational **O**n**l**ine **A**nalytical **P**rocess, por sus siglas en inglés**):** implementación OLAP que almacena los datos en un motor relacional. La arquitectura está compuesta por un servidor de de datos relacional y el motor OLAP, que se encuentra en un servidor dedicado. La principal ventaja de esta arquitectura es que permite el análisis de una enorme cantidad de datos usando motores relacionales.
- ✓ **MOLAP (*M*ultidimensional *Onl*ine *A*nalytical *P*rocessing,** por sus siglas en inglés**):** esta implementación OLAP almacena los datos en una base de datos multidimensional.
- ✓ **HOLAP (H**ybrid **O**n**l**ine **A**nalytical **P**rocess, por sus siglas en inglés**)**: es una combinación de ROLAP y MOLAP, que son otras posibles implementaciones de OLAP. HOLAP permite almacenar una parte de los datos como en un sistema MOLAP y el resto como en uno ROLAP.

Puede verse como los sistemas OLAP no tienen la misma utilidad que los OLTP, pues cada uno es para fines diferentes y la selección de cada uno va en dependencia de las necesidades de negocio que se está implementando.

Extracción transformación y carga.

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

En síntesis, las funciones específicas de los ETL son tres: la extracción, transformación y carga.

Extracción

El primer paso del proceso ETL consiste en extraer los datos desde los sistemas fuentes. Generalmente se agrupan datos de diferentes sistemas de origen. Cada sistema separado puede tener una organización diferente de los datos o formatos distintos. Los formatos de las fuentes normalmente se encuentran en bases de datos relacionales o ficheros, pero pueden incluir bases de datos no relacionales u otras estructuras diferentes. La extracción deja los datos preparados para las transformaciones.

Transformación

El segundo paso del proceso ETL consiste en convertir aquellos datos inconsistentes en un conjunto de datos compatibles y estandarizados, para que puedan ser cargados en el almacén de datos. Estas transformaciones se ejecutan con el objetivo de llevar todo a un único modo, definiendo estándares, para que todos los datos que ingresarán al almacén estén integrados y estandarizados.

Carga

El tercer paso del proceso ETL es el responsable de cargar los datos al almacén una vez que han sido transformados y que normalmente se encuentran en el stagingarea y en ocasiones de las fuentes pues tienen correspondencia directa con el almacén. Se debe ser muy riguroso con estos datos antes de cargarlos definitivamente pues deben tener la mejor calidad, ya que este es un factor clave, que no debe pasarse por alto.

Este proceso de ETL es de vital importancia en la construcción de un almacén de datos pues es el que le da vida al almacén y debe realizarse con la mayor atención posible porque no deben entrar al almacén datos inconsistentes y corruptos pues una vez en el almacén no se deben eliminar y cambiar.

1.1.3. Ventajas y Desventajas de los almacenes de datos

Como toda nueva tecnología está sujeta a ventajas y desventajas. Las cuales se especifican a continuación:

Ventajas:

- Posibilita la integración los datos de varias fuentes de datos a una única plataforma y formato.
- Mejora la calidad de la información de los datos de la entidad.
- Proporciona el análisis de los datos que realmente son útiles en temas determinados en la entidad.
- Ayuda a la toma de decisiones, basada en historial de la entidad.
- Posibilita una mejor visión de determinadas áreas de la entidad.

Desventajas:

- Incremento continuo de los requerimientos del usuario.
- Su construcción es compleja, consume recursos y tiempo.

Como puede verse las ventajas que traen consigo los almacenes son de gran utilidad y pueden ser muy provechosos a las entidades que los implementen.

1.1.5 Ejemplo de aplicaciones de los almacenes de datos.

En este acápite, se pretende dar una panorámica de cómo se comporta, tanto en Cuba como en el mundo, la utilización de los almacenes de datos. Existen muchas empresas que dependen del uso de los almacenes de datos para llevar a cabo un negocio de mercado. La compañía WalMart, considerada la empresa más grande a nivel mundial cuenta con el DWH más voluminoso y poblado del mundo; el cual usa para tomar decisiones acerca de todos los procesos que realizan en el mercado internacional, elevar su economía y mantenerse en competencia respecto a otras compañías. Igualmente, Twentieth Century Fox utiliza la información relacionada con las películas que se proyectan en distintos lugares de los Estados Unidos para predecir qué actores, argumentos y

filmes serán populares, con el objetivo de ganar audiencia en sus producciones. El uso del DWH no sólo se aplica en áreas dentro del negocio, es aplicable al 100% de las áreas fuera de este. Se evidencia su utilización en Venezuela, aplicado al tema de la seguridad ciudadana; así como en hospitales de Perú para la sectorización de pacientes en el consumo de medicamentos. Cuba no se encuentra alejada de la aplicación de esta herramienta para la toma de decisiones. La empresa CIMEX, destacada por el crecimiento constante y la estabilidad financiera, tanto dentro como fuera del país utiliza un DWH para la gestión de inventarios. Además, en la Universidad de las Ciencias Informáticas se ha desarrollado un almacén de datos para la toma de decisiones en cuanto al consumo energético.

1.2 Gestión de los Ensayos Clínicos en el CIM

Uno de los fármacos que se desarrollan en el CIM y sobre el cual se realizan los EC es el Nimotuzumab, anticuerpo monoclonal recombinante contra el Receptor del Factor de Crecimiento Epidérmico (EGFR, siglas en inglés). Nimotuzumab es un medicamento que se utiliza como monoterapia o en combinación con radioterapia y/o quimioterapia para el tratamiento de cáncer de origen epitelial y gliomas. Nimotuzumab recibe alternativamente las denominaciones comerciales de CIMAher®, TheraCIM®-hR3, Theraloc®, BIOMAb®-EGFR, YMB 1000 y VECTHIX®. Internacionalmente, se encuentra en Canadá, la India, China y Alemania; de ahí la importancia que representa para el CIM garantizar la seguridad y eficacia de este producto.

El programa de desarrollo clínico de este fármaco, incluye 33 ensayos clínicos adicionales que están en curso y tres programas de acceso expandido, para investigar la seguridad y eficacia de Nimotuzumab en cáncer de mama, esófago, cáncer de cabeza y cuello, próstata, cáncer de hígado, páncreas, pulmón (células no pequeñas), cáncer de cuello de útero y glioma en pacientes pediátricos y adultos. Hasta la fecha, Nimotuzumab ha sido administrado a más de 1000 pacientes en ensayos clínicos terminados y en más de 5000 pacientes con estudios en curso.

El proceso de monitorización de toda la información clínica (cuyo objetivo es controlar la calidad y uniformidad de los datos recogidos) se lleva a cabo mediante visitas periódicas a cada uno de los sitios que participan en esta actividad (que se refiere a todas las provincias involucradas). Luego de llenar los Cuadernos de Recogida de Datos (CRD) en los hospitales, estos se llevan al CIM y se guardan en bases

de datos electrónicas para facilitar los resultados estadísticos que de ellos se derivan. En el CIM, la gestión de los EC se realiza a través del sistema EPIDATA, utilizado para la recopilación de la información. Este sistema genera reportes en diferentes formatos: SPSS, Excel, SASS y Text. Igualmente almacena información en otros formatos: .rec, .qes, .eix, .chk, .bak y .not. Otra característica importante es que a este sistema acceden diferentes especialistas, que la información está disgregada en diferentes modelos y que estos no presentan igual estructura de diseño. Al no encontrarse integradas ni estandarizadas, se torna engorroso el proceso para el manejo de la información por parte de los directivos de la institución, dificultando la realización de análisis estadísticos complejos dentro de un mismo EC o entre diferentes EC; corriendo el riesgo de que se pierda información útil y valiosa con el decursar del tiempo e impidiendo una adecuada integración de los datos que contribuya a elevar la efectividad del tratamiento de la información.

Los principales procesos de análisis de los datos están centrados en la eficacia y la seguridad de los productos. Cabe destacar la importancia que se le da a estos datos en el CIM pues en ellos está la documentación de los medicamentos que puede salvar las vida de cientos de personas e información sensible de los pacientes que se someten al estudio, por ese motivo siempre están bajo estricto monitoreo y control del estado y acceso de los datos por los diferentes especialistas y aplicaciones que interactúan con estos.

1.3 Arquitectura de software.

La definición de arquitectura de software es tan variada que el número de conceptos alcanza las docenas las cuales puedes consultarse en el sitio del SEI dedicado al tema. Algunas de las más reconocidas se listan a continuación [11]:

- Garlan y Perry: "…es el nivel del diseño del software donde se definen la estructura y propiedades globales del sistema".
- Shaw y Garlan: "…se centra en aquellos aspectos del diseño y desarrollo que no pueden tratarse de forma adecuada dentro de los módulos que forman el sistema."
- Clements: "…es, a grandes rasgos, una vista del sistema que incluye los componentes principales del mismo, la conducta de esos componentes según se le percibe desde el resto del sistema y las formas en que los componentes interactúan y se coordinan para alcanzar la misión del sistema. La vista arquitectónica es una vista abstracta, aportando el más alto nivel

de comprensión y la supresión o diferimiento del detalle inherente a la mayor parte de las abstracciones."

Analizando la variedad de definiciones anteriores se puede ver que de modo general todos los autores plantean que la arquitectura de software es la estructura global del sistema y las relaciones entre sus partes. También refieren que una arquitectura de software bien definida permitirá que un exitoso desarrollo del sistema.

1.4 Arquitectura de almacenes de datos.

La arquitectura de los almacenes de datos es una forma de representar la estructura global de los datos, la comunicación, los procesos y la presentación del usuario final, puede verse en la figura 4. Además es típica e incluye lo siguiente:

- ***Datos operacionales***. Origen de datos o datos fuentes de donde se extraerán los datos primarios.
- ***Extracción de datos***. Extracción de datos operacionales con el fin de formar parte del Almacén de Datos.
- ***Transformación de datos***. Procesos de cambios que ocurren sobre los datos que poblarán el almacén.
- ***Carga de datos***. Proceso de poblado del almacén con los datos extraídos y transformados.
- ***Almacén***. Almacenamiento físico de datos.
- ***Herramienta de acceso***. Herramientas que proveen acceso a los datos.

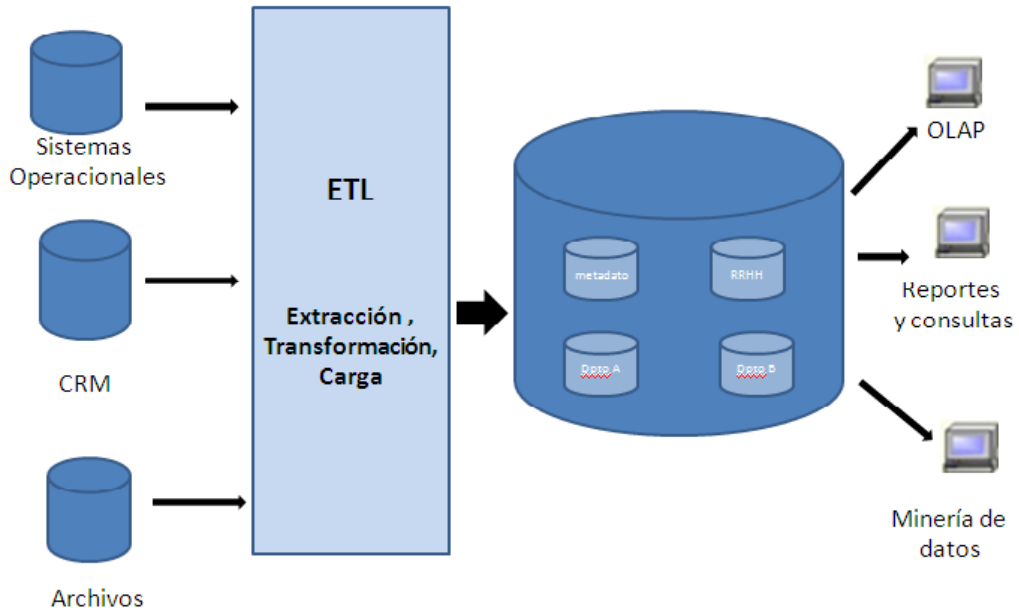

Figura 4.Arquitectura básica de un almacén de datos

Marco de arquitectura de Kimball

Kimball plantea un marco de referencia de arquitectura para almacenes de datos el cual está bien descrito en su libro *The Data Warehouse Lifecycle Toolkit*

En este marco de referencia se proponen tres columnas fundamentales en la arquitectura de los almacenes de datos:

1. La columna de datos (¿el qué?):

   Describe qué datos se van a analizar, las características, cuál va a ser su estructura dentro del almacén, especificando el modelo dimensional, además de los modelos lógico y físico.

2. La columna técnica (¿el cómo?):

   Esta área abarca el flujo de los datos. Es decir cómo se van a extraer los datos de las fuentes y ubicarlos en un lugar accesible, describiendo las transformaciones que van a

sufrir los mismos. Además de cuáles son los estándares y productos que se necesitan para acceder y realizar la carga de los datos y qué tipo de análisis se les va realizar, también qué estándares se van a usar para mostrar la información y cómo se va a ver la misma.

Se divide en dos partes fundamentales back room y front room. El back room es la parte encargada de acceder a las fuentes de orígenes y colocar los datos en el almacén y front room es el responsable de hacer accesible los datos a los usuarios finales.

3. La columna de infraestructura (¿el dónde?):

   Describe dónde se van a almacenar los datos físicamente, tiene en cuenta las plataformas y los servidores y las ubicaciones físicas de los componentes.

Marco de arquitectura de Luján

Luján plantea un marco de referencia de arquitectura para almacenes de datos el cual está bien descrito en su tesis de doctorado.

En este marco de referencia se proponen tres niveles fundamentales en la arquitectura de los almacenes de datos [12]:

Nivel 1 Conceptual: define el almacén de datos desde el punto de vista conceptual, es decir, desde el mayor nivel de abstracción y contiene únicamente los objetos y relaciones más importantes.

Nivel 2 Lógico: Abarca aspectos lógicos, como la definición de las tablas y claves, la definición de los procesos ETL, entre otros.

Nivel 3 Físico: define los aspectos físicos del almacén de datos como el almacenamiento de las estructuras lógicas, en diferentes discos, o la configuración de servidores de base de datos que mantienen los datos del almacén.

Se decide utilizar para definir y describir la arquitectura del almacén de datos el marco de referencia de Kimball pues plantea una serie de aspectos muy importantes de esta área, así mismo hace un énfasis mayor en las herramientas a utilizar y además está bien documentada, contando con un amplio respaldo y

apoyo de la comunidad de desarrollo de almacenes de datos. También hay que tener en cuenta que no se especifica cómo realizar la descripción de la misma, es decir que artefactos generar para el apoyo a la descripción.

1.5 Modelos de descripción de arquitectura.

Existen varios modelos de descripción de arquitectura donde se definen ciertas vistas y diagramas correspondientes. Los mismos han sido asimilados por los diseñadores de sistemas para la documentar la arquitectura, a continuación se realiza una descripción de los modelos más utilizados:

1.5.1 El modelo 4+1 vistas de Krutchen

El modelo 4+1 define cuatro estructuras o *vistas* que deben ser consideradas en la documentación de todo sistema [4]:

Vista Física. Describe la relación entre el *software* y el *hardware* y sus aspectos de distribución. Básicamente es la descripción de cómo el *software* se ha *trasladado* al *hardware* del sistema. En esta vista se identifican los procesadores del sistema, su interconexión y los componentes *software* que debe ejecutar cada uno de ellos.

Vista Lógica. Describe los requisitos funcionales del sistema. En un sistema orientado a objetos esta vista sería la descomposición del sistema en clases y objetos, la descripción de sus relaciones y la definición de las interfaces proporcionadas por los mismos. De forma más general, esta vista describe los servicios que deben proporcionar los componentes del sistema, bien a otros componentes, bien aun usuario externo.

Vista de Procesos. Describe el comportamiento dinámico del sistema y sus procesos. Cubre aspectos tales como el paralelismo, la concurrencia y la sincronización de tareas. Krutchen enfatiza que estos aspectos capturan los requisitos no funcionales del sistema y menciona explícitamente el rendimiento, que considera no funcional, la disponibilidad, la integridad y la tolerancia a fallos. Este enfoque no es del todo compartido por otros autores, que argumentan que dichos requisitos están presentes en todas las vistas y que ciertos requisitos no funcionales están más directamente asociados a otras vistas. De hecho, Krutchen asocia la reusabilidad y la portabilidad a la vista de desarrollo.

Vista de Desarrollo. Describe la organización estática del *software* en su entorno de desarrollo. El *software* se agrupa en módulos o subsistemas, organizados en diferentes librerías *software*, y cada módulo ofrece a los demás una interfaz bien definida. En esta vista se definen las relaciones de uso (importación y exportación de servicios) entre los diferentes módulos del sistema. La vista de desarrollo tiene una extraordinaria importancia en la planificación del proyecto, ya que es la más habitualmente usada para la asignación de trabajo a los diferentes equipos de desarrollo.

Vista de escenarios. Los escenarios, representan casos de usos significativos del sistema y son la argamasa que une a las diferentes vistas del sistema, ya que muestran al sistema como un todo. La vista de escenarios es redundante o, si se prefiere complementaria, respecto del resto (de ahí el "+1") y tiene dos propósitos:

- Servir de herramienta para descubrir elementos arquitectónicos (componentes y relaciones).
- Validar la arquitectura. Los mismos casos de uso que se han usado para el diseño, aunque más elaborados, sirven de trazas para validar el sistema en sus diferentes fases de desarrollo hasta llegar al producto final.

Este modelo ha sido de gran aceptación por los arquitectos de sistemas enterprise donde existe una gran cantidad de usuarios interactuando y de operaciones de todo tipo, además la lógica de negocio está bien definida. Pero analizando el tema de almacenes de datos y los ensayos clínicos en el CIM un aspecto importante son el flujo de los datos y acceso de los datos de los mismos por las diferentes partes del sistema, aspecto mencionado en el epígrafe 1.2 y este modelo no lo contempla este punto de vista.

1.5.2 CWM

El CWM es un meta-modelo destinado a la concepción de un almacén de datos desarrollado por el Object Management Group (OMG) para lograr una estandarización e intercambio de la información de los almacenes de datos. Este meta-modelo está basado en las notaciones MOF, UML, XMI y XML. No realiza su descripción a través de vistas sino de escenarios los cuales se listan a continuación:

ETL: describe el proceso de extracción, transformación y carga del almacén.
OLAP: describe el proceso de definición de los cubos y sus características.

Cuestionarios: describe los principales cuestionarios y reportes del almacén.
Administración: describe como debe ser el control y monitoreo del estado del almacén.
Herramientas: describe las herramientas a utilizar.

Utiliza para la representación principalmente diagramas de clases y de paquetes para realizar agrupamientos.

Este meta-modelo aunque no es basado en vistas, ha sido utilizado por los diseñadores de almacenes de datos pero no tiene en cuenta aspectos relacionados con el hardware, aspecto fundamental para la descripción de la arquitectura según Kimball, ni tampoco tiene en cuenta las características de cómo va a ser el acceso a los datos por los usuarios.

1.6 Lenguajes para el modelado de la arquitectura

En un esfuerzo para estandarizar las notaciones y procesos a utilizar para describir arquitecturas son el lenguaje de descripción de arquitectura (ADL, por sus siglas en inglés) y el Lenguaje Unificado de Modelado (UML, por sus siglas en inglés, Unified Modeling Language)

ADL

El surgimiento de este lenguaje data de principios de los 90 y hasta la fecha se han materializado diversas propuestas para describir y razonar en términos de arquitectura de software; muchas de ellas han asumido la forma de lenguajes de descripción arquitectónicos. Los ADL se utilizan para describir un sistema en componentes y conectores, especificando de qué manera estos elementos se combinan para formar configuraciones y definiendo familias de arquitecturas o estilos. Existen ADLs de propósito general y otros de dominio específicos. Algunos de estos son : Aesop, ArTek, C2, Darwin, LILEANNA, MetaH, Rapide, SADL, UniCon, Weaves y Wright[13], entre otros

Características de los ADLs [13]

- Composición: Permiten la representación del sistema como composición de una serie de partes.

- Configuración: La descripción de la arquitectura es independiente de la de los componentes que formen parte del sistema.
- Abstracción: Describen los roles o papeles abstractos que juegan los componentes dentro de la arquitectura.
- Flexibilidad: Permiten la definición de nuevas formas de interacción entre componentes.
- Reutilización: Permiten la reutilización tanto de los componentes como de la propia arquitectura.
- Heterogeneidad: Permiten combinar descripciones heterogéneas.
- Análisis: Permiten diversas formas de análisis de la arquitectura y de los sistemas desarrollados a partir de ella.

UML

Es un lenguaje gráfico para visualizar, especificar, construir y documentar un sistema de software. Es el más conocido y utilizado en la actualidad. En las primeras versiones de UML no tuvieron notaciones para componentes y conectores que contribuyan a describir arquitectura de software. UML 2.0 fue liberado en el 2003 y se ocupa de muchos asuntos que se necesitan en un lenguaje de descripción de arquitectura, cumple con la mayoría de las características los ADL, hace grandes adelantos para convertirse en una notación para describir arquitecturas. RUP promueve el uso de UML 2.0 para modelar la arquitectura de software lo que le da un gran impulso para ser la notificación estándar de descripción de arquitectura de software , ante la variedad de ADL los cuales difieren unos de otros y en ocasiones son difíciles de comprender , por tanto se decide utilizar UML 2.0 como lenguaje de modelado.

Perfiles de UML

Los Perfiles UML son una posibilidad que proporciona el propio UML para ampliar su sintaxis y su semántica con el objetivo expresar los conceptos específicos de un determinado dominio de aplicación, El hecho de que UML haya sido un lenguaje diseñado de propósito general nos brinda una gran flexibilidad y expresividad a la hora de modelar sistemas. Pero en ocasiones, es más aconsejable utilizar algún lenguaje más específico para modelar y representar los conceptos de ciertos dominios particulares.
Existen dos posibilidades para definir lenguajes específicos de un dominio [14]:

- Definir un nuevo lenguaje.

- Extender el propio UML

Pero decidirse por una alternativa u otra trae sus inconvenientes si bien con la primera se puede lograr un modelado a la medida del dominio esto puede traer consigo que las herramientas CASE no puedan utilizarlas, de lo contrario con la segunda, se facilitaría su implementación, que sería ajustar y extender UML a la terminología de otros modelos ya existentes. En el caso de esta investigación si no se puede describir un dominio específico con UML, el autor propone extender UML a una terminología ya existente que se ajuste a las necesidades del dominio a analizar siempre sea posible , esto traería consigo que se pueda modelar en las herramientas CASE existentes.

1.7 Conclusiones parciales

En este capítulo se abordaron e identificaron temas centrales acerca de los almacenes de datos, como sus principales características, ventajas, desventajas y conceptos fundamentales, entre otros. Se describió la arquitectura de un almacén así como se seleccionó el lenguaje de modelado UML 2.0 para describir la arquitectura. Se detallaron los ensayos clínicos así como su gestión en el centro de inmunología molecular. También se evidenció las deficiencias de los modelos existentes para describir la arquitectura de un almacén de datos para ensayos clínicos de CIM.

CAPÍTULO 2: DESCRIPCIÓN DEL MODELO

2.1 Modelo para describir arquitectura en almacenes de datos.

Analizando el marco de referencia para describir arquitectura que propone Kimball, descrito en el capítulo anterior en el epígrafe 1.4 y teniendo en cuenta que no se especifica como describirla, además de las deficiencias de los modelos existentes para la descripción de los almacenes de datos descritos en el epígrafe 1.5. Se propone definir un modelo para la descripción arquitectónica del almacén de datos basadas en UML 2.0 siguiendo los detalles de Kimball.

2.1.1 Propuestas de vistas para la descripción de arquitectura de almacenes de datos

Siguiendo el estándar IEEE-1471-2000 que define una serie de pasos para determinar las vistas de la arquitectura para que la misma posea una buena documentación. Se precisan los siguientes aspectos [3]:

Stakeholders: Son las personas involucradas en el desarrollo de la solución, ya sea directa o indirecta.
Necesidades (Concerns): Incumbencias que tiene cada stakeholder con la solución.
Punto de vista (Viewpoints): Un punto de vista determina los lenguajes (anotaciones, modelos, etc.) que se usaran para describir intentando satisfacer las necesidades de los stakeholders.
View (Vista): Es la representación de una arquitectura con respecto a un punto de vista particular.
Modelos: Son los diagramas (elementos y sus relaciones) que se utilizan para comunicar una vista particular.
Descripción de la arquitectura: Es la documentación de la arquitectura propiamente dicha, que está organizada.

En la figura 5 se muestras la relación entre los aspectos mencionados anteriormente.

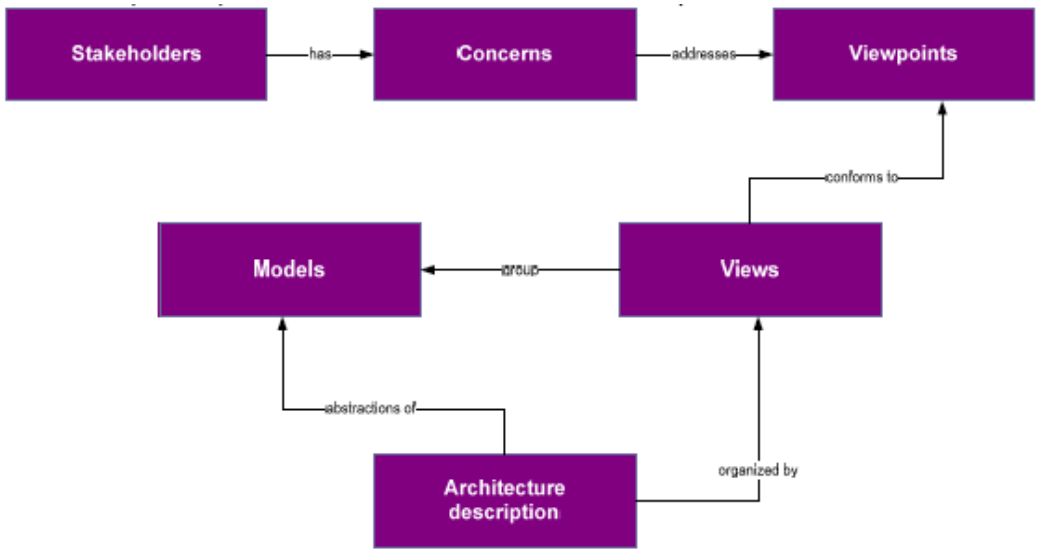

Figura 5. Relación de aspectos de descripción de arquitectura [3]

.

Definición de Stakeholders: expertos del CIM, diseñadores de base de datos, integradores de datos, arquitectos, especialista de hardware y configuración.

Necesidades:

Stakeholders	Necesidades
Expertos del CIM (Usuarios)	Necesita conocer cómo van a fluir los datos por la solución, cómo va a ser el acceso a los mismos por los diferentes componentes y cómo los usuarios los van a manipular, la cantidad de recursos de cómputos y su estructura, además de los productos necesarios, la relación entre ellos y su ubicación en el hardware.
Diseñadores de base de datos	Necesitan conocer las características de los datos a almacenar y las estructura de los mismos.
Integradores de datos	Necesitan conocer el flujo de los datos y las herramientas, así como la estructura de los datos
Arquitectos	Necesitan saber las responsabilidades de cada parte de la solución, su alcance, sus dependencias y maneras de comunicación (protocolos), así como

	la ubicación lógica de cada componente, cómo va a ser el acceso a los datos por los diferentes componentes y cómo los usuarios los van a manipular.
Especialistas de hardware	Necesitan conocer la disposición de hardware que se requiere, así como la distribución de los componentes por hardware.

Tabla 1. Necesidades de los stakeholder

Puntos de vista: Se propone que los puntos de vistas sean tres, haciéndolos coincidir con las columnas del marco de referencia de Kimball.

Punto de vista	**Punto de vista de datos**: corresponde con la columna de datos (¿El qué?). Describirá la estructura de los datos.
Stakeholders	Expertos del CIM, diseñadores de base datos, integradores de datos.
Necesidades	Características de los datos y estructura.
Lenguaje de descripción	Diagramas UML.
Vistas	Vista de datos
Diagramas a representar	Diagrama de modelo relacional de datos.

Tabla 2. Punto de vista de datos

Punto de vista	**Punto de vista técnico:** corresponde con la columna técnica (¿El cómo?). Describirá la parte técnica de la solución, la relación entre sus componentes y su ubicación lógica.
Stakeholders	Expertos del CIM, diseñadores de base datos, integradores de datos, arquitectos.
Necesidades	Flujo de los datos, componentes y productos necesarios y ubicación lógica de cada uno.

	Conocer cómo van a fluir los datos por la solución, cómo va a ser el acceso a los mismos por los diferentes componentes y cómo los usuarios los van a manipular
Lenguaje de descripción	Diagramas UML.
Vistas	Vista de flujo de datos, Vista de de implementación y Vista lógica de implementación, Vista de acceso a datos.
Diagramas a representar	Diagrama de flujo de datos, diagrama de componentes, diagrama de paquetes, diagramas de acceso a datos por usuarios y diagramas de acceso a datos por componentes.

Tabla 3. Punto de vista de técnico

Punto de vista	**Punto de vista de infraestructura:** corresponde con la columna de infraestructura (¿El dónde?). Responde a la distribución física de componentes y las necesidades de hardware.
Stakeholders	Expertos del CIM, Especialistas de hardware, arquitectos.
Necesidades	Ubicación de los componentes y necesidades de medios de cómputo.
Lenguaje de descripción	Diagramas UML.
Vistas	Vista de despliegue, Vista de despliegue por componentes.
Diagramas a representar	Diagrama de despliegue, diagrama de distribución física de los componentes.

Tabla 4. Punto de vista de infraestructura

Vistas y modelos:

La vista que conforma el punto de vista de datos**,** describe la estructura que van a tener las base de datos necesarias ya sea el staging área, o los DM que van a conformar el almacén, puede describirse de modo dimensional si se necesita, representando el esquema seleccionado ya sea de estrella, copo de nieve o constelación de hechos, también sería útil describir cómo están los datos fuentes pero no se considera que sea necesario u obligatorio hacerlo, se hace necesario extender UML al dominio de modelado de datos, pues se ajusta mejor al dominio, utilizando el diagrama relacional de datos. En la figura 6 se muestra un ejemplo.

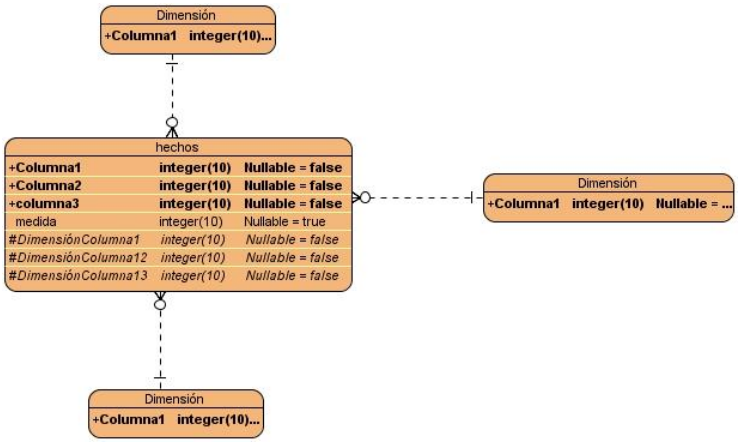

Figura 6. Diagrama relacional

Las vistas que conforman el punto vista técnico, describen los flujos de los datos del almacén (vista de flujo de datos), los componentes que van a conformar el almacén así como las relaciones entre ellos (vista de implementación), también incluye la localización lógica de los componentes lo que significa la ubicación de los mismos en las diferentes capas lógicas (vista lógica de implementación) y además la descripción de cómo se va a realizar el acceso a los datos por los diferentes componentes y usuarios (Vista de acceso a datos por usuarios y componentes). Se hace necesario extender UML para modelar el flujo de datos y acceso a datos pues en este no existe dichos diagramas en UML, se decide utilizar estereotipos de los diagramas de actividades, componentes y caso de uso, pues se ajusta mejor al dominio. Utilizando los diagramas de: diagrama de flujo de datos, diagramas de componentes, diagramas de acceso a datos por

usuario y componentes. Se recomienda que se haga la descripción separada por partes de la columna técnica (back room y front room). A continuación se muestran ejemplos de los diferentes diagramas, diagrama de flujo de datos (figura 7), diagrama de componentes respectivamente (figura 8), diagrama de paquetes (figura 9), diagramas de acceso a datos por usuarios (figura 10) y por componentes (figura 11). Con la realización de los dos últimos diagramas se puede obtener también un análisis cuantitativo de la cantidad de operaciones que se ejecutan sobre determinados datos lo cual puede determinar la prioridad de los mismos y así tenerlos en cuenta a la hora de asignación de recursos y configuración de los servidores donde van a residir los mismos.

Figura 7. Diagrama de flujo de datos (vista de flujo de datos)

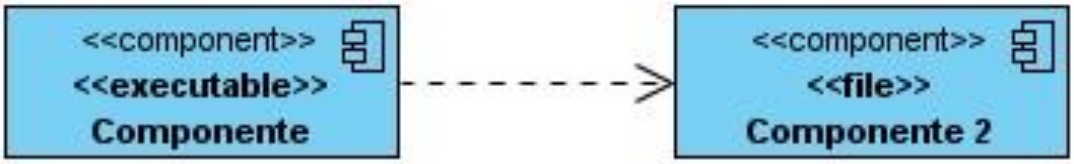

Figura 8. Diagrama de componentes (vista de implementación)

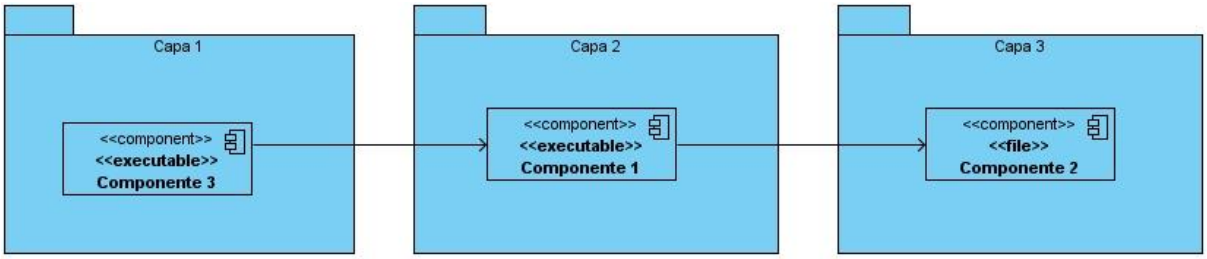

Figura 9. Diagrama de paquetes de componentes (vista lógica de implementación)

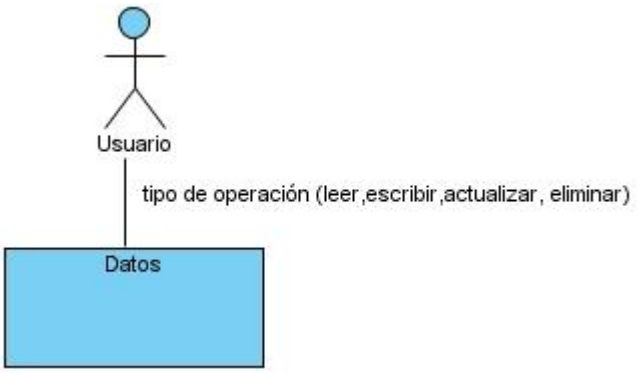

Figura 10. Diagrama de acceso a datos por usuarios (vista de acceso a datos por usuarios)

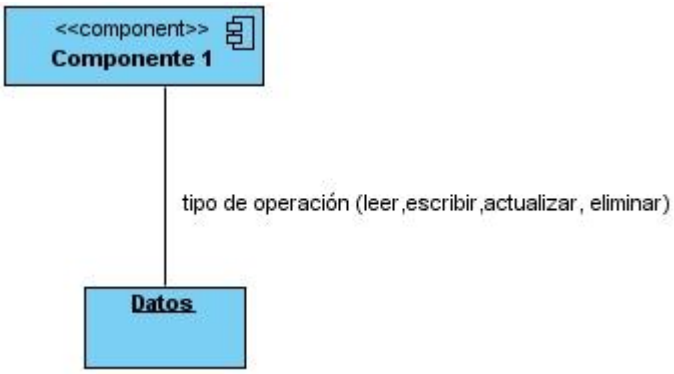

Figura 11. Diagrama de acceso a datos por componentes (vista de acceso a datos por componentes)

Las vistas que conforman el punto vista de infraestructura, describen dónde se van a almacenar los datos físicamente, tiene en cuenta las plataformas (vista de despliegue) y los servidores y las ubicaciones físicas de los componentes (vista de despliegue por componentes). Utilizando el diagrama de UML: diagrama de

despliegue. Se muestran ejemplos: diagrama de despliegue (figura 12) y distribución física de los componentes (figura 13):

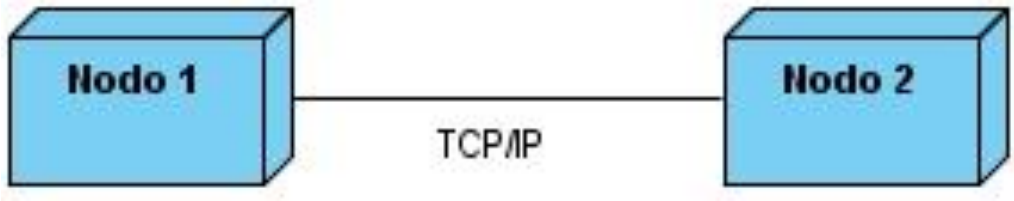

Figura 12. Diagrama de despliegue (vista de despliegue)

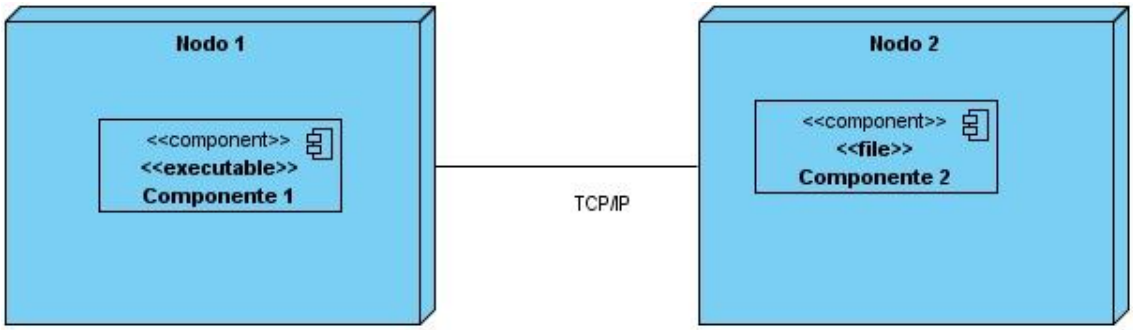

Figura 13. Distribución física de los componentes (vista de despliegue por componentes)

Cabe destacar que debe hacerse una descripción textual de los diagramas pues esto permite una mejor documentación de la arquitectura. En la figura 14 se muestra un resumen de los diagramas UML que se deben utilizar para describir la arquitectura de un almacén de datos por las vistas propuestas.

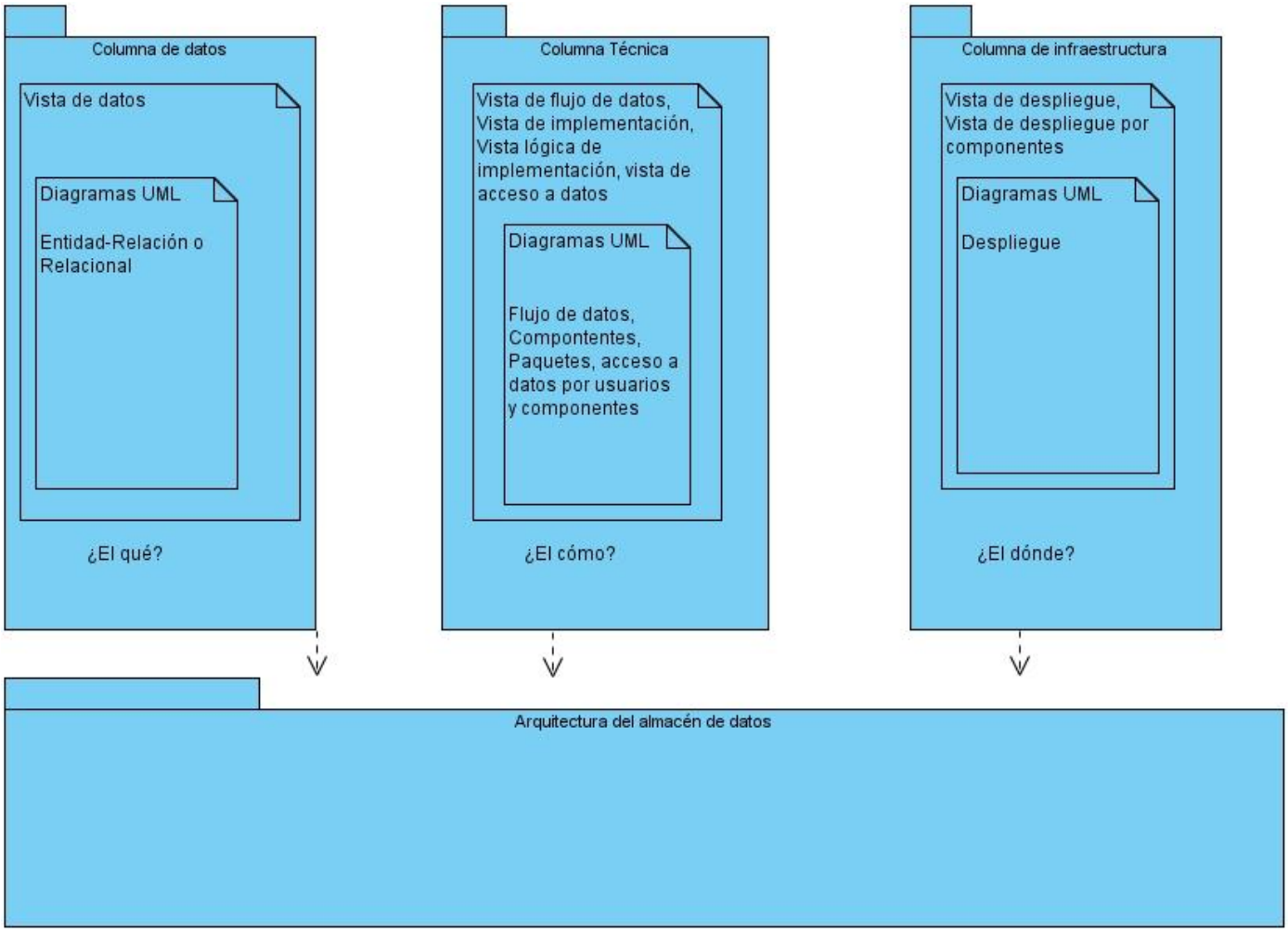

Figura 14. Resumen de los diagramas

2.3 Conclusiones parciales

En este capítulo se definieron las vistas necesarias para la descripción de la arquitectura del almacén de datos, basándose en el marco de referencia de descripción de arquitectura planteado por Kimball y utilizando UML 2.0 como lenguaje de modelado, describiéndose cada una de las vistas así como los diagramas necesarios, siguiendo el estándar IEEE-1471-2000.

CAPITULO 3: COMPARACION Y APLICACIÓN DEL MODELO

3.1 Comparación y evaluación del modelo.

A continuación se realiza una comparación de los modelos de vistas descritos en el epígrafe 1.4 y el modelo propuesto para describir arquitectura de almacenes de datos de EC.

Diagramas	***El modelo 4+1 de Krutchen***	***CWM***	***Modelo para describir almacenes de datos de Ensayos clínicos.***
Diagrama de clases	**x**	**x**	
Diagrama de entidad- relación o Relacional	**x**		**x**
Diagrama de componentes	**x**		**x**
Diagrama de despliegue	**x**		**x**
Diagrama de casos de uso	**x**		
Diagrama de secuencia	**x**		
Diagrama de paquetes	**x**	**x**	**x**
Diagrama de flujo de datos			**x**
Diagrama de acceso a datos de usuarios			**x**
Diagrama de acceso a datos de componentes			**x**

Tabla 5. Comparación de los modelos de vistas

Existen varios diagramas en común con ambos modelos y a su vez se tiene dos diagramas en particular que son el de flujo de datos y los de acceso a datos por usuarios y componentes, los cual describen el flujo y acceso a los datos, aspecto importante en los almacenes de datos de ensayos clínicos, y contribuye a una descripción más completa de la arquitectura del los almacenes, tema que no describen los modelos existentes.

El estándar IEEE-1471-2000 describe que se obtienen las vistas y modelos para describir la arquitectura a partir de las necesidades de los stakeholder, es decir una vez satisfechas las necesidades con las vistas y diagramas entonces son suficientes para la describir la arquitectura. En nuestro caso se determinaron 3 puntos de vista y dentro de estos se determinaron 7 vistas arquitectónicas, y a su vez dentro de estos se determinaron 7 modelos. A continuación se muestra la correspondencia de las necesidades de los stakeholder con las vistas y me demuestra que todos fueron abarcados.

1. Necesita conocer cómo van a fluir los datos por la solución,
2. Cómo va a ser el acceso a los mismos por los diferentes componentes
3. Cómo los usuarios los van a manipular,
4. Además de los productos necesarios, la relación entre ellos.
5. Necesitan conocer las características de los datos a almacenar y las estructura de los mismos.
6. Necesitan saber las responsabilidades de cada parte de la solución, su alcance, sus dependencias y maneras de comunicación (protocolos)
7. Necesitan saber la ubicación lógica de componente.
8. Necesitan conocer la disposición de hardware que se requiere.
9. Así como la distribución de los componentes por hardware.

Vitas	Necesidades
Vista de datos	5
Vista de flujo de datos	1
Vista de implementación	4,6
Vista lógica de implementación	4, 6,7
Vista de acceso a datos por usuarios	3
Vista de acceso a datos por componentes	2
Vista de despliegue	8
Vista de despliegue por componentes	9

Tabla 6: correspondencia de las necesidades con las vistas.

3.2 Propuesta de arquitectura para almacenes de datos de ensayos clínicos del Centro de Inmunología Molecular

3.2.1 Características fundamentales que debe poseer la arquitectura del almacén de datos.

Luego de entrevistas con los especialistas del CIM donde se abordaron temas generales que debe poseer la arquitectura del almacén de datos para los ensayos clínicos, se llegó a la conclusión que se necesita que el almacén de datos cumpla con las siguientes características y requerimientos técnicos:

- Debe ser sobre software libre.
- Debe soportar análisis OLAP.
- Debe mostrar la información sobre tecnología Web.
- Debe ofrecer distintos tipos de reportes útiles para el usuario.
- Debe permitir una fácil integración de los componentes.
- Debe admitir la incorporación de otros ensayos.

3.2.2 Descripción de la arquitectura.

Visual Paradigm

Visual Paradigm para UML versión 6.1 es una herramienta de modelado que soporta UML 2.0,así como es muy eficiente para dibujar diagramas y generar códigos en varios lenguajes de programación. Además, permite una integración con sistemas de control de versiones que almacenan centralmente los artefactos y realizan un seguimiento de los cambios realizados sobre un proyecto. Los desarrolladores lo utilizan para facilitar el modelado simultáneo, almacenar los archivos de proyectos y hacer un seguimiento de los cambios, en la presente investigación se utilizará dicha herramienta para dibujar los diagramas pertinentes.

Una vez definidas las características y requerimientos técnicos del almacén en el epígrafe anterior así como la herramienta para modelar en UML se describe la arquitectura. Siguiendo las vistas propuestas para describir la arquitectura del almacén de datos.

Vista de datos:

- Se analizarán datos de ensayos clínicos que tienen como características su diversidad de fuentes y formatos incluso siendo del mismo ensayo pues los modelos no cuentan con el mismo modo de recoger la información, los mismos se encuentra en archivos *Excel (.xls)*. Se decide utilizar un staging área para facilitar el proceso de ETL. Como modelo dimensional se propone usar el modelo de estrella por las ventajas que tiene el mismo:
    - Posee mejores tiempos de respuesta.
    - Su diseño es fácilmente modificable.
    - Simplifica el análisis.

O constelación de hechos que es una composición de varios modelos de estrella. A continuación en las figuras 15 y 16 se muestran las vistas de modelo físico:

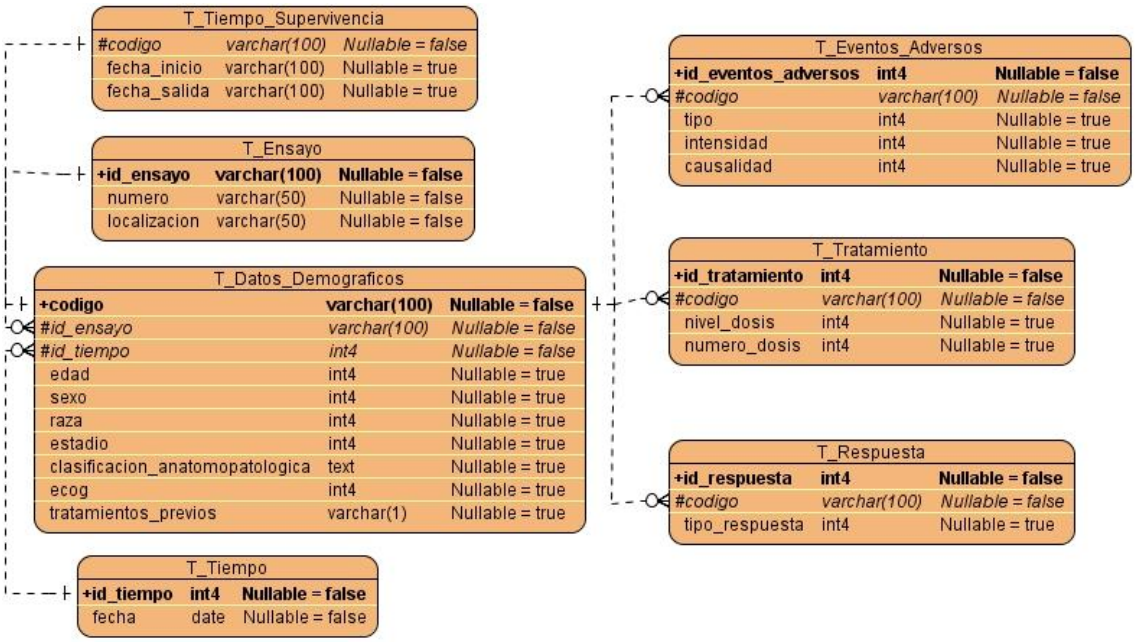

figura 15. Diagrama relacional del stagin área

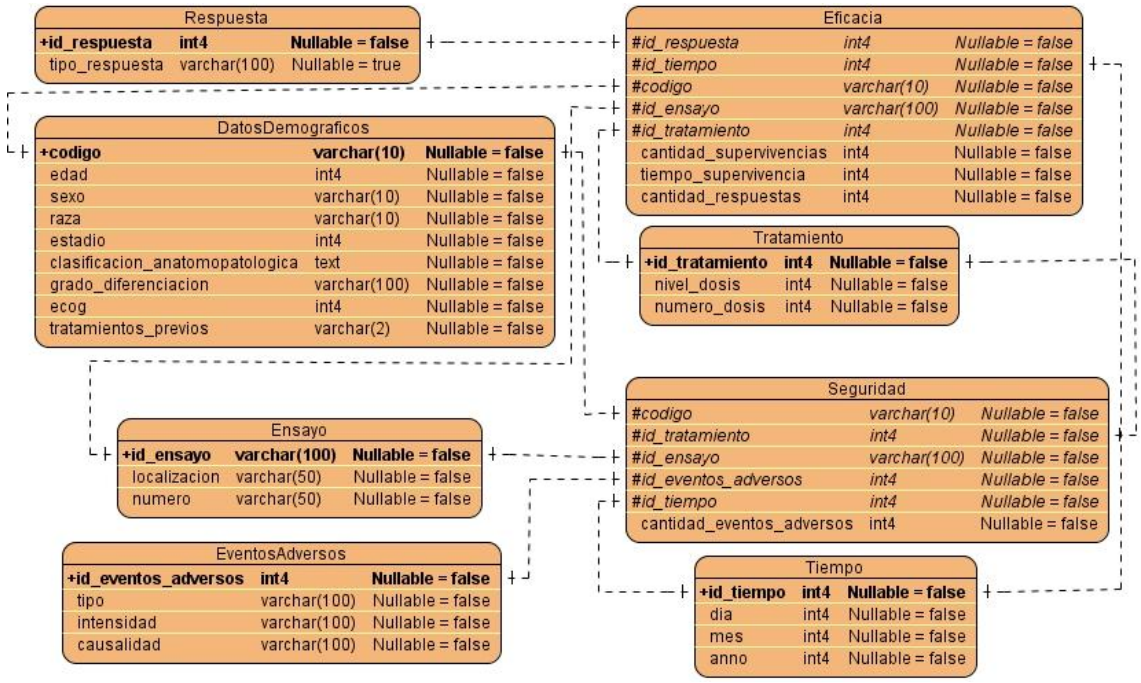

figura 16. Diagrama relacional del DM del producto hR3

Vista de flujo de datos, Vista de implementación, Vista lógica de implementación, Vista de acceso a datos por usuarios y vistas de acceso a datos por componentes:

Back Room: Se determina utilizar la estrategia ascendente propuesta por Kimball, donde cada DM representa a un producto al cual se le está aplicando un ensayo clínico y un DWH que reúne puntos en común de los distintos DM.

Los datos son extraídos desde bases de datos o cualquier archivo donde resida información útil para el proceso de ETL. Una vez analizadas las fuentes de datos se procede a extraer los datos estrictamente necesarios para la carga de los mismos dentro del *staging área*. Posteriormente los datos son integrados, transformados y limpiados, para luego ser cargados en los distintos DM, luego y por último se les aplica otro proceso de ETL hacia el DWH. De esta forma se deja abierta la arquitectura con el objetivo de dar soporte a la aparición de un nuevo producto el cual formará un nuevo DM. Todo este proceso descrito anteriormente se almacena en los metadatos a través de la herramienta informática Kettle 3.1 de la suite de Pentaho encargada de la ETL. Para almacenar la información el SGBD PostgreSQL 8.3 que es un

gestor relacional y libre, además de ser el gestor de base de datos de código abierto más potente en la actualidad, cuenta con la mayoría de las características y funcionalidades de gestores que se encuentran en el sector privativo. Para lograr la replicación de los datos con el objetivo de lograr mayor disponibilidad y seguridad de respaldo de los mismos se propone usar un mecanismo de replicación en este caso se propone Slony-I, el cual brinda replicación maestro/esclavo y trae como ventaja que tiene un servidor primario que manda las actualizaciones en tiempo real a la base de datos esclava y en caso de que falle el servidor primario, continúa trabajando con la base de datos esclava.

En las figuras 16 y 17 se pueden ver los diagramas de flujo de datos, diagrama de componentes que describen el Back Room:

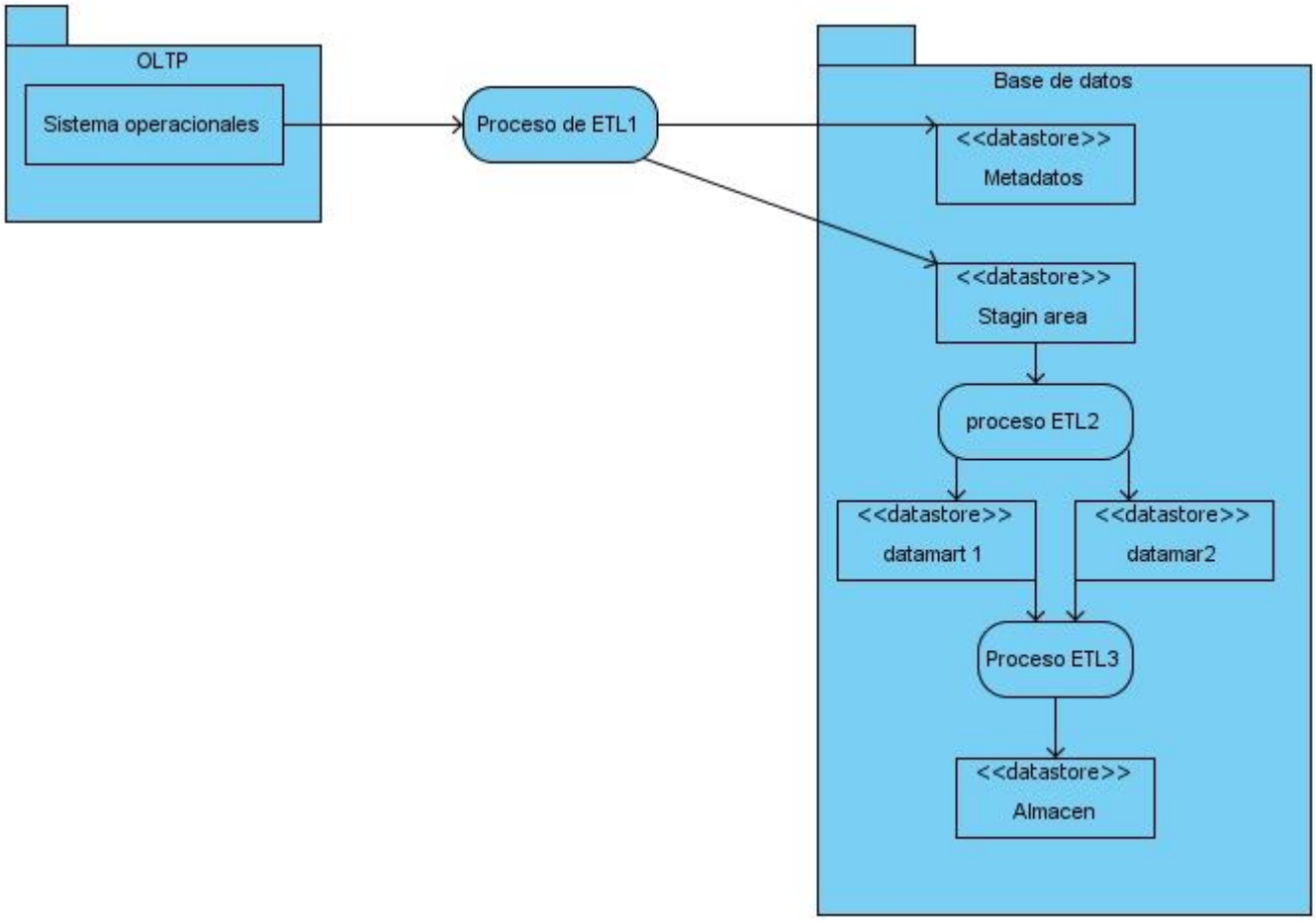

figura 16. Diagrama de flujo de datos del back Room

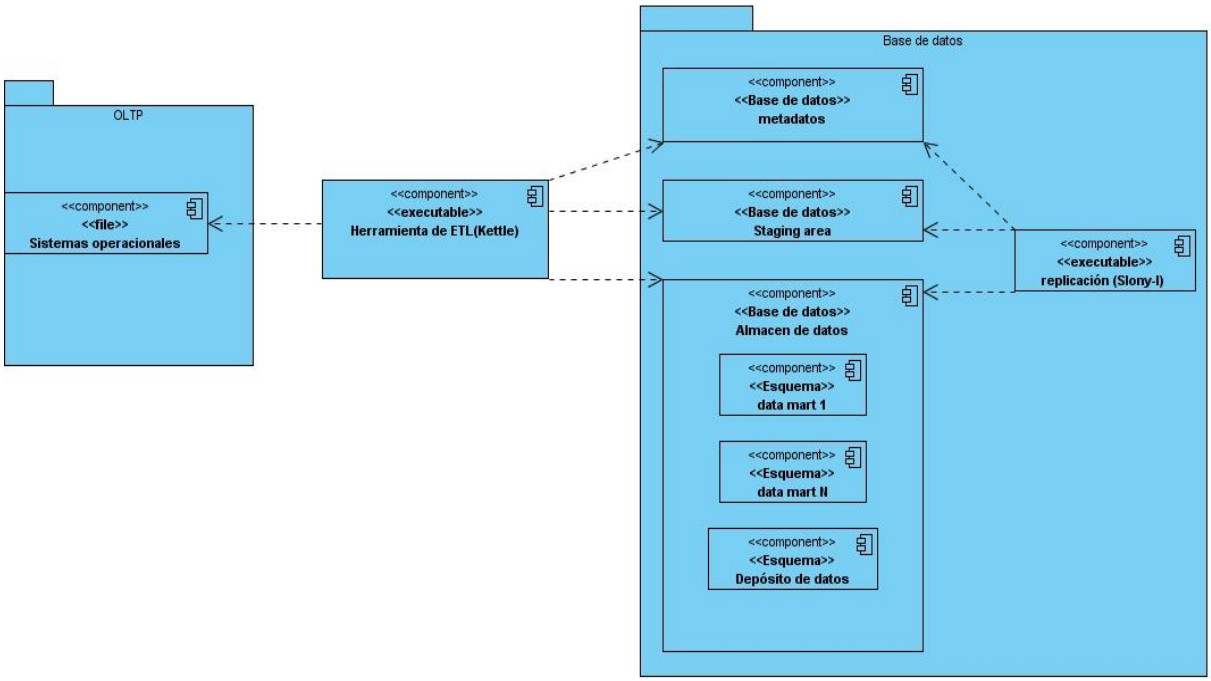

Figura 17. Diagrama de componentes del back room

Front Room: Se propone utilizar como motor OLAP el Mondrian 3.0.4 de la suite de Pentaho, el cual está basado en el tipo de análisis ROLAP; gestiona la comunicación entre una aplicación OLAP (escrita en Java) y la base de datos con los datos fuente, actuando como JDBC (Java Database Connectivity) para OLAP. Cabe destacar que este es capaz de comunicarse con bases de datos PostgresSQL. Para mostrar la información se utilizará JPivot 3.0 de la suite de Pentaho que es un cliente web para OLAP realizado en Java/Swing, donde los usuarios pueden realizar navegaciones típicas de OLAP como taladrar, rebanar y cortar en cubitos, permitiendo recoger datos de un servidor Mondrian y realizar operaciones sobre estos, el mismo cuenta con 16 tipos de gráficas donde están las más comunes como son los de barra, línea, pie, entre otros también destacar la fácil integración de ambos.

Se decide utilizar Apache Tomcat como servidor web debido a que este se ha convertido en la implementación de referencia para las especificaciones de *Servlet y JSP*; esto último es de gran importancia para la ejecución del servidor Mondrian, y el cliente de OLAP JPivot. Además de ser servidor libre.

También se requiere de la máquina virtual de java (JVM, por sus siglas en inglés) la cual es la pieza principal de las aplicaciones en java, sobre la cual se ejecutaría todo lo anterior descrito.

En las figuras 18 y 19 se pueden ver los diagramas de flujo de datos y diagrama de componentes que describen el Front Room

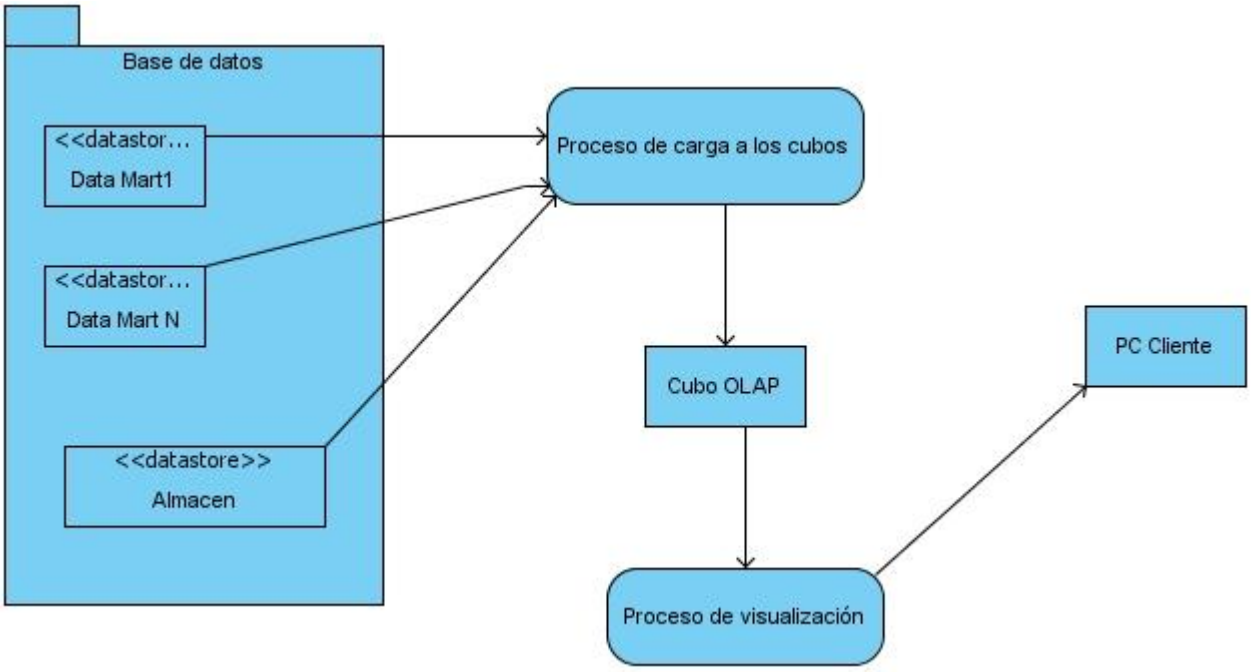

Figura 18. Diagrama de flujo de datos del front room

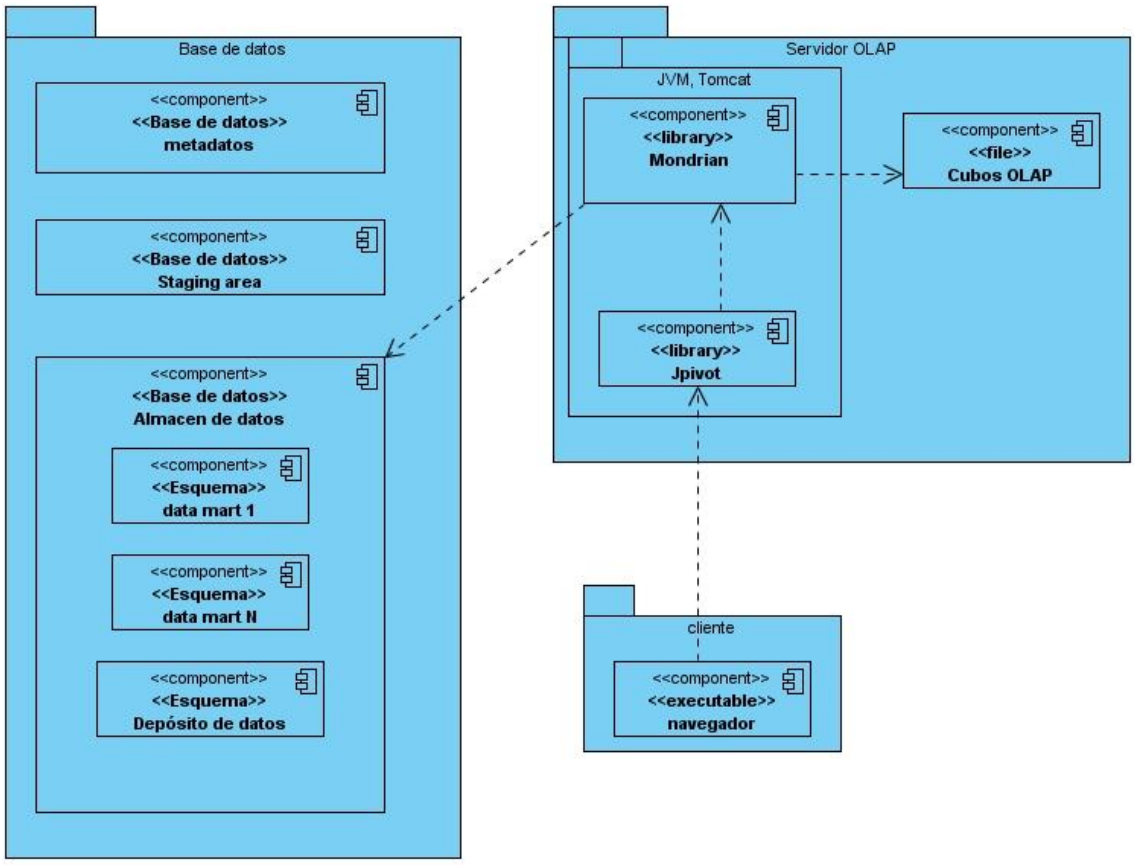

Figura 19. Diagrama de componentes del Front Room

Se propone un desglose en tres capas: **capa de integración**, **capa de análisis** y **capa de visualización**.

En la **capa de integración** es donde se extraen los datos de las fuentes externas, los cuales pasan por un proceso de ETL para estandarizarlos y eliminar posibles errores y luego pasaran a poblar los DMs y/o datawarehouse.

La **capa de análisis** es una capa intermedia entre la integración y la visualización que se encarga de realizar las consultas a los datos (DM y/o datawarehouse) y enviarlos para que sean visibles a los usuarios.

Por último la **capa de visualización** es la que se encarga de mostrar los resultados al usuario final para que este pueda interpretar fácilmente.

En la figura 20 se puede ver el diagrama de componentes por capas de la arquitectura.

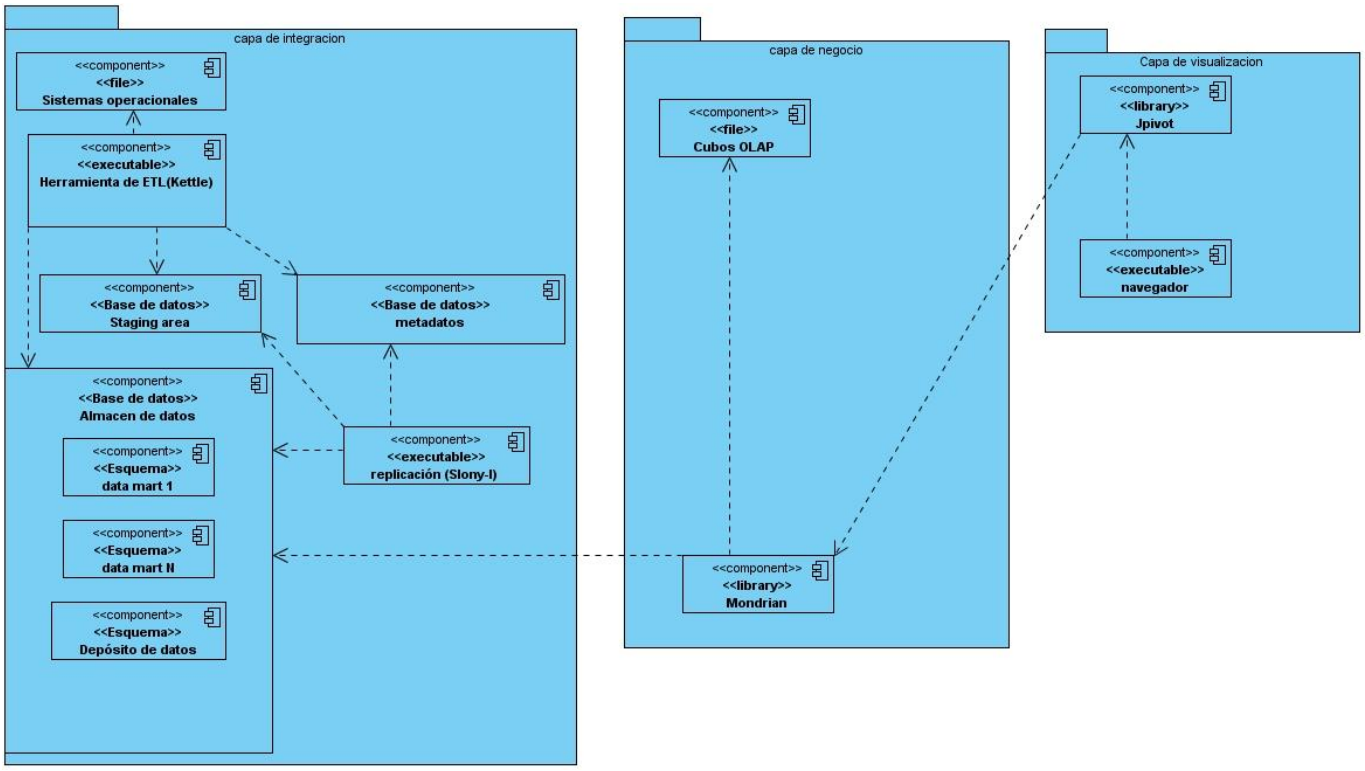

Figura 20. Diagrama de componentes por capas

A continuación se muestran los diagramas de acceso a datos por los usuarios y por los componentes.

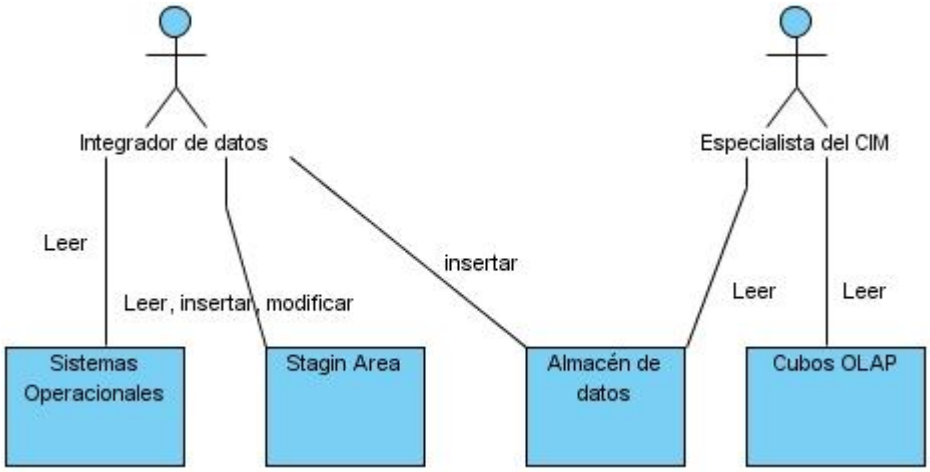

Figura 21. Diagrama de acceso a datos por usuarios

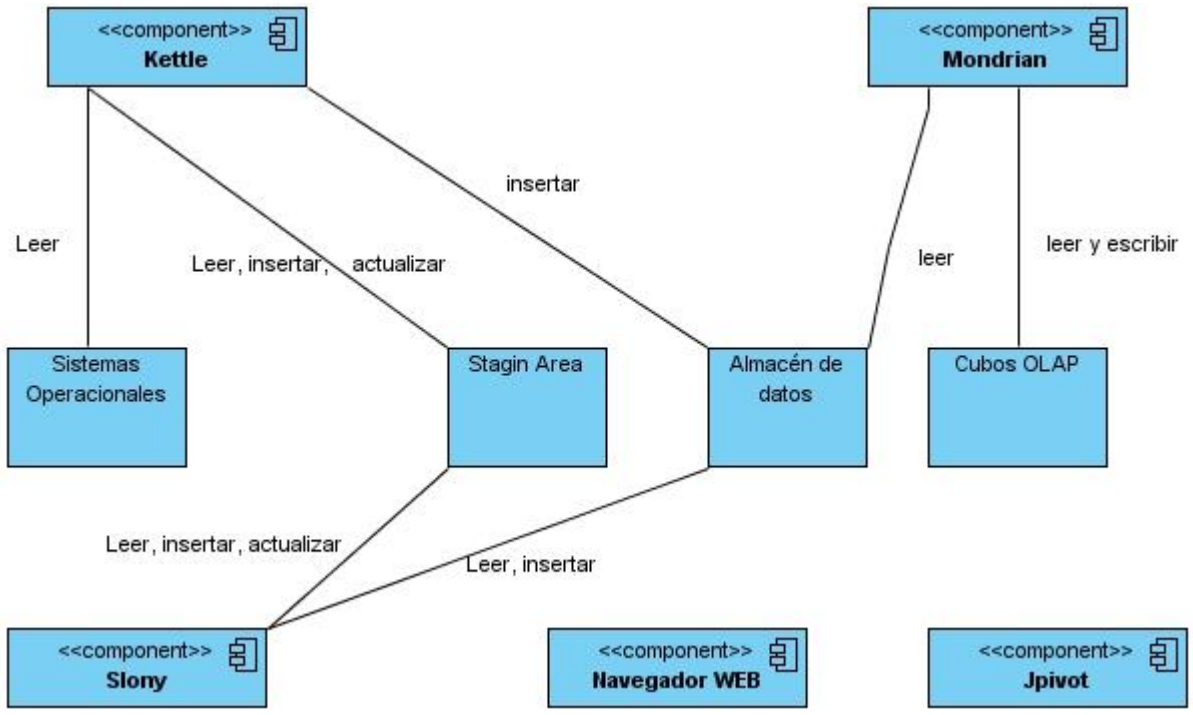

Figura 22. Diagrama de acceso a datos por componentes

Realizando un análisis cuantitativo de la cantidad de operaciones sobre los datos se puede observar que los datos que van a tener más accesos son los que están en el almacén de datos, pues va ser accedidos por tres componentes y dos usuarios diferentes.

Vista de despliegue y Vista de despliegue por componentes:

Se propone una estructura centralizada donde el almacén de datos se encuentra en un único servidor. Y sea una sola base de datos aprovechando la oportunidad que brinda PostgreSQL de esquemas y ubicar en cada uno un DM y el almacén, donde el nodo de replicación tendría la misma información en otro servidor. El motor OLAP Mondrian y el cliente Jpivot estarían en un servidor de aplicaciones que se conectaría al almacén para extraer sus datos. Además la información de la fuentes de datos estaría en un servidor de información al cual se conectaría el servidor de transformación donde se ejecutaría la herramienta de ETL (*pueden estar en la misma estación las fuentes y la herramienta de ETL, depende de la disponibilidad que se tenga*).

En la figuras 21 y 22 se pueden ver los diagramas de despliegues y la distribución física de los componentes de la arquitectura propuesta.

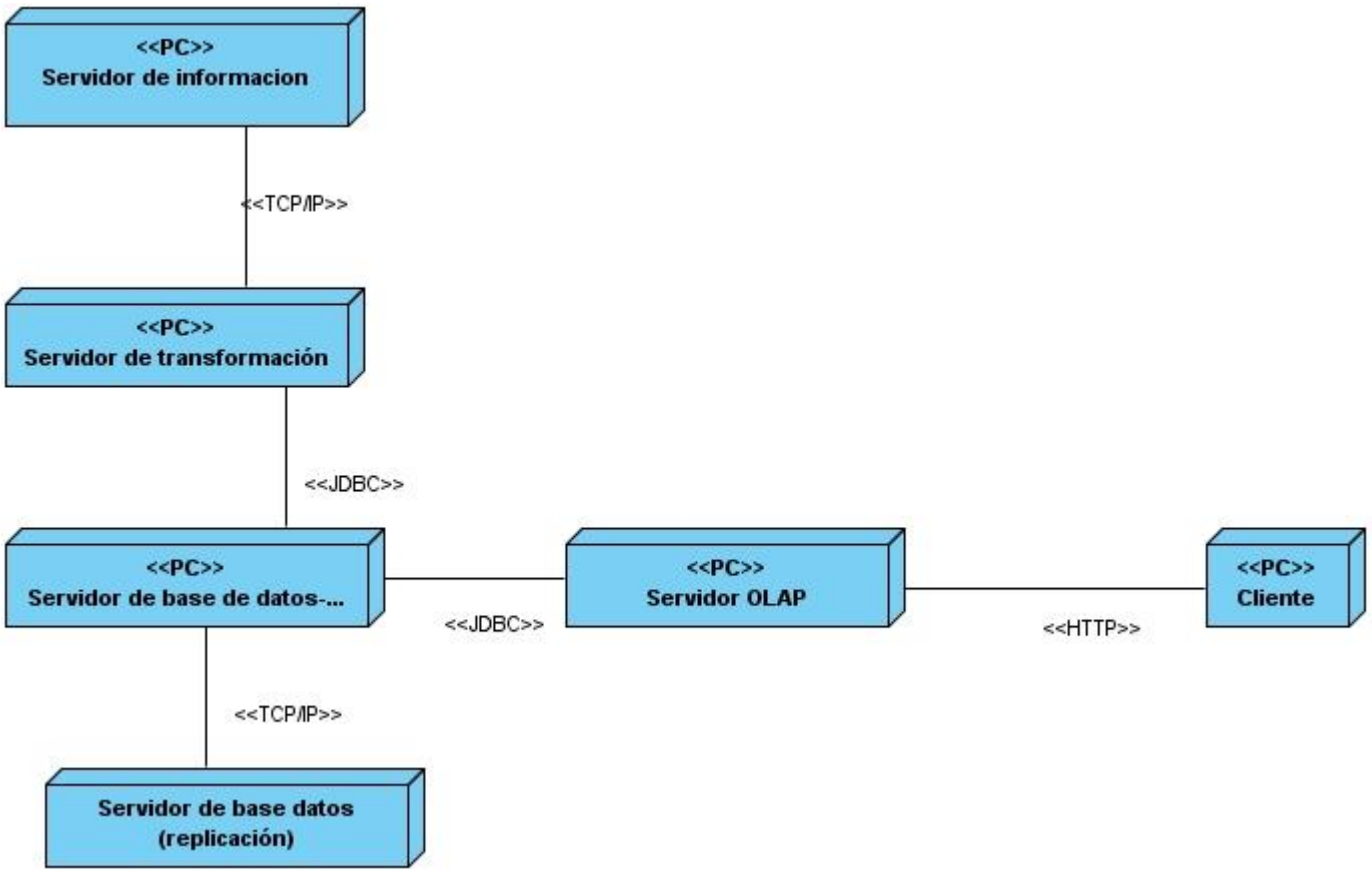

Figura 21. Diagrama de despliegue

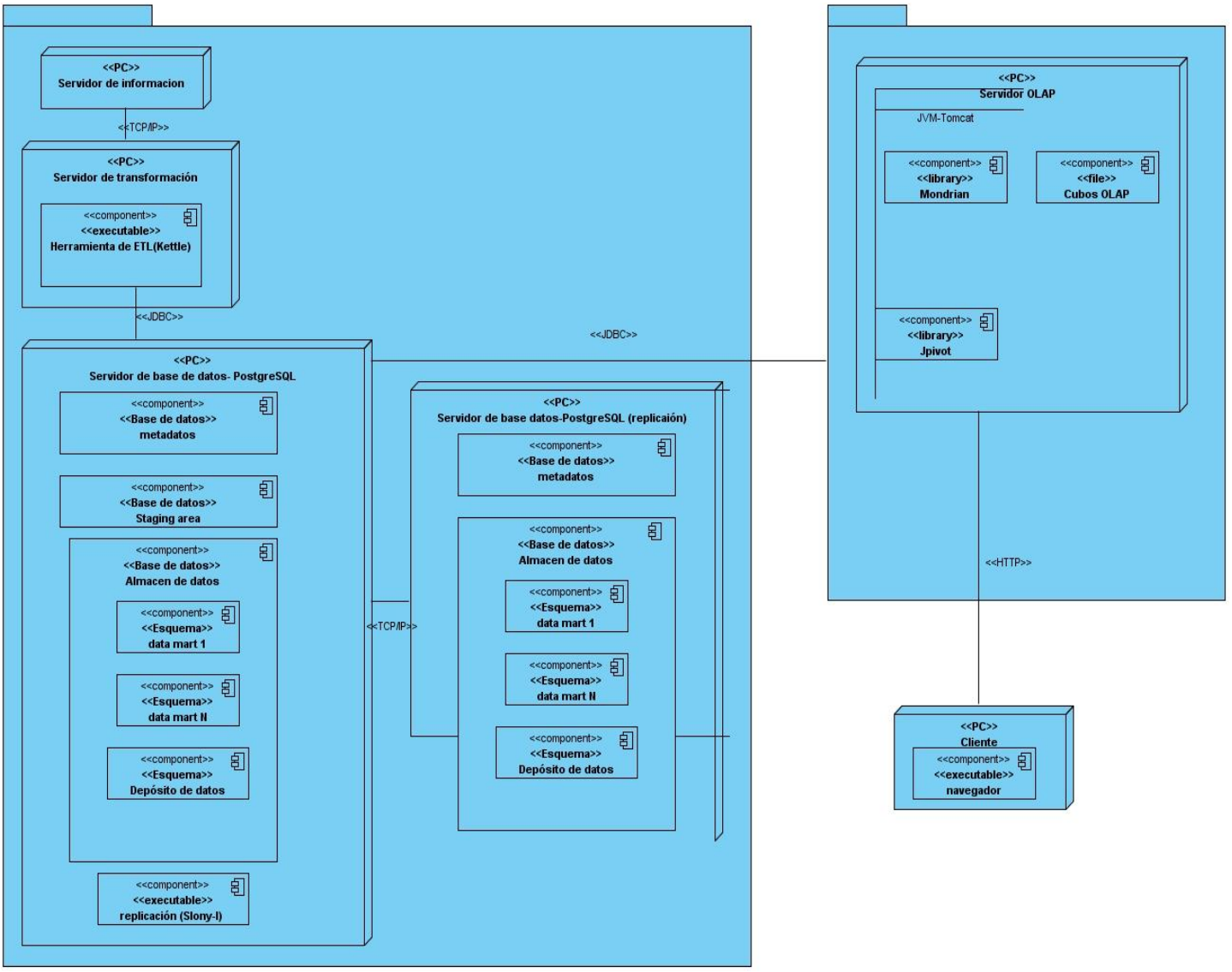

Figura 22. Diagrama de distribución física

Luego de la descripción de la arquitectura se puede ver como se cumplen todas las características y requerimientos técnicos descritas por los clientes.

- Debe ser sobre software libre: Todas la herramientas propuestas son sobre software libre(Kettle 3.1, PostgreSQL 8.3, slony-I, Mondrian 3.0, Jpivot 3.0)
- Debe soportar análisis OLAP: soporta tecnología OLAP mediante el motor OLAP Mondrian 3.0 y la visualiza mendiante Jpivot 3.0
- Debe mostrar la información sobre tecnología Web: la información puede ser vista por la web por el Jpivot 3.0
- Debe ofrecer distintos tipos de reportes útiles para el usuario: el Jpivot ofrece 16 tipos de gráficas, los cuales son los más comunes.
- Debe permitir una fácil integración de los componentes: todos los componentes se ejecutan sobre tecnología java y pertenecen a la misma suite en este caso Pentaho.
- Debe permitir la incorporación de otros ensayos: se propone usar para cada ensayo un esquema nuevo de PostgreSQL dentro de la misma base de datos.

Además en el anexo 1 se puede ver el aval de la entidad donde expresa que está conforme y de acuerdo con la arquitectura propuesta.

3.3 Implementación del almacén de datos de ensayos clínicos del producto hR3 del CIM.

Una vez determinadas las vistas arquitectónicas que describen el almacén se procedió a la implementación del mismo.

Puede verse en la figura 23 las bases de datos del Staging área y las del producto hR3 en el servidor de PostgreSQL 8.3.

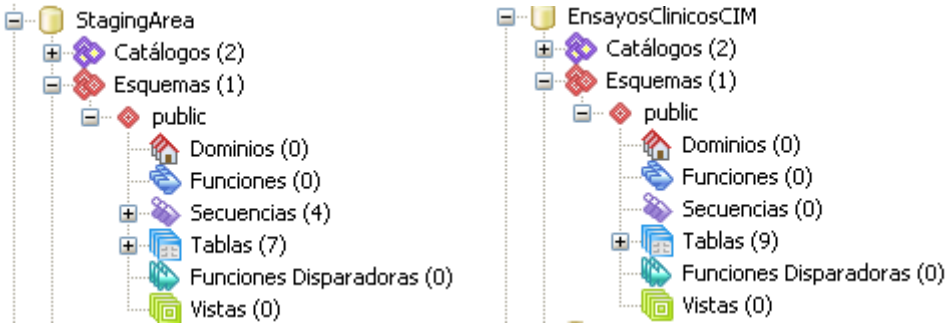

Figura 23. Bases de datos StaginArea y EnsayosCinicosCIM

Implementación del proceso de extracción, transformación y carga

Para poblar el DM del producto hR3 con los datos fuentes, se hizo necesario realizar el proceso de ETL dos veces. Primero se pasó toda la información necesaria contenida en los ficheros xls hacia el *Staging área*, realizando algunas transformaciones necesarias. Una vez concluida esta primera etapa se insertaron los datos del *Staging área* dentro del DM del producto hR3, donde se realizaron todas las transformaciones necesarias para que los datos quedaran lo más integrados posibles.

En la siguiente tabla se muestran los resultados de la extracción (E), transformación (T) y carga (L) hacia el Staging área de 10 ensayos clínicos del producto hR3.

Base de datos	**Dimensiones del *Staging área***													
	Datos demográficos		**Eventos adversos**		**Tiempo de supervivencia**		**Ensayo clínico**		**Tratamiento**		**Respuesta**		**Tiempo**	
	E	**L**	**E**	**L**	**E**	**L**	**E**	**L**	**E**	**L**	**E**	**L**	**E**	**L**
	T		**T**		**T**		**T**		**T**		**T**		**T**	
C y C 040	14	14	838	838	28	14	1	1	76	76	40	40	0	1
	3		2		2		0		2		2		0	
C y C 046	30	10	186	186	20	10	1	1	56	56	40	40	0	1
	3		2		2		0		2		2		0	

Base de datos	Dimensiones del *Staging área*													
	Datos demográficos		Eventos adversos		Tiempo de supervivencia		Ensayo clínico		Tratamiento		Respuesta		Tiempo	
	E	L	E	L	E	L	E	L	E	L	E	L	E	L
	T		T		T		T		T		T		T	
C y C 055	208	103	183	183	181	73	1	1	285	285	1098	1098	0	1
	3		2		2		0		2		2		0	
C y C 076	20	10	100	100	20	10	1	1	40	40	44	11	0	1
	3		2		2		0		2		2		0	
Glioma 069	29	29	559	559	46	17	1	1	335	335	177	177	0	1
	3		2		2		0		2		2		0	
Glioma 053	29	29	43	43	46	16	1	1	335	335	177	177	0	1
	3		2		2		0		2		2		0	
Metace-rebral079	54	27	190	190	54	27	1	1	116	116	32	32	0	1
	3		2		2		0		2		2		0	
T.Sólidos 035	12	12	108	108	22	10	1	1	54	41	36	36	0	1
	3		2		2		0		2		2		0	
Mama 070	25	12	77	77	24	11	1	1	119	119	36	36	0	1
	3		2		2		0		2		2		0	
Esófago 075	116	52	400	400	107	43	1	1	187	124	81	81	0	1
	3		2		2		0		2		2		0	

Tabla 11. Datos de ETL hacia el Staging área

Luego se procedió a la extracción (E), transformación (T) y carga (L) de los datos hacia en DM y los resultados se muestran a continuación:

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

Tabla 12. Datos de ETL hacia el DM de hR3

Definición de los Cubos OLAP

Una vez el DM poblado se realizó la instalación de apache Tomcat como servidor de aplicaciones, el motor OLAP en este caso el Mondrian 3.0 y la interfaz web Jpivot para la visualización de los datos. Para poder mostrar los datos se deben concebir los cubos OLAP necesarios en este caso se crearon dos cubos de datos correspondientes a la eficacia y seguridad del producto hR3.En la figura 24 se muestra la definición de los mismos.

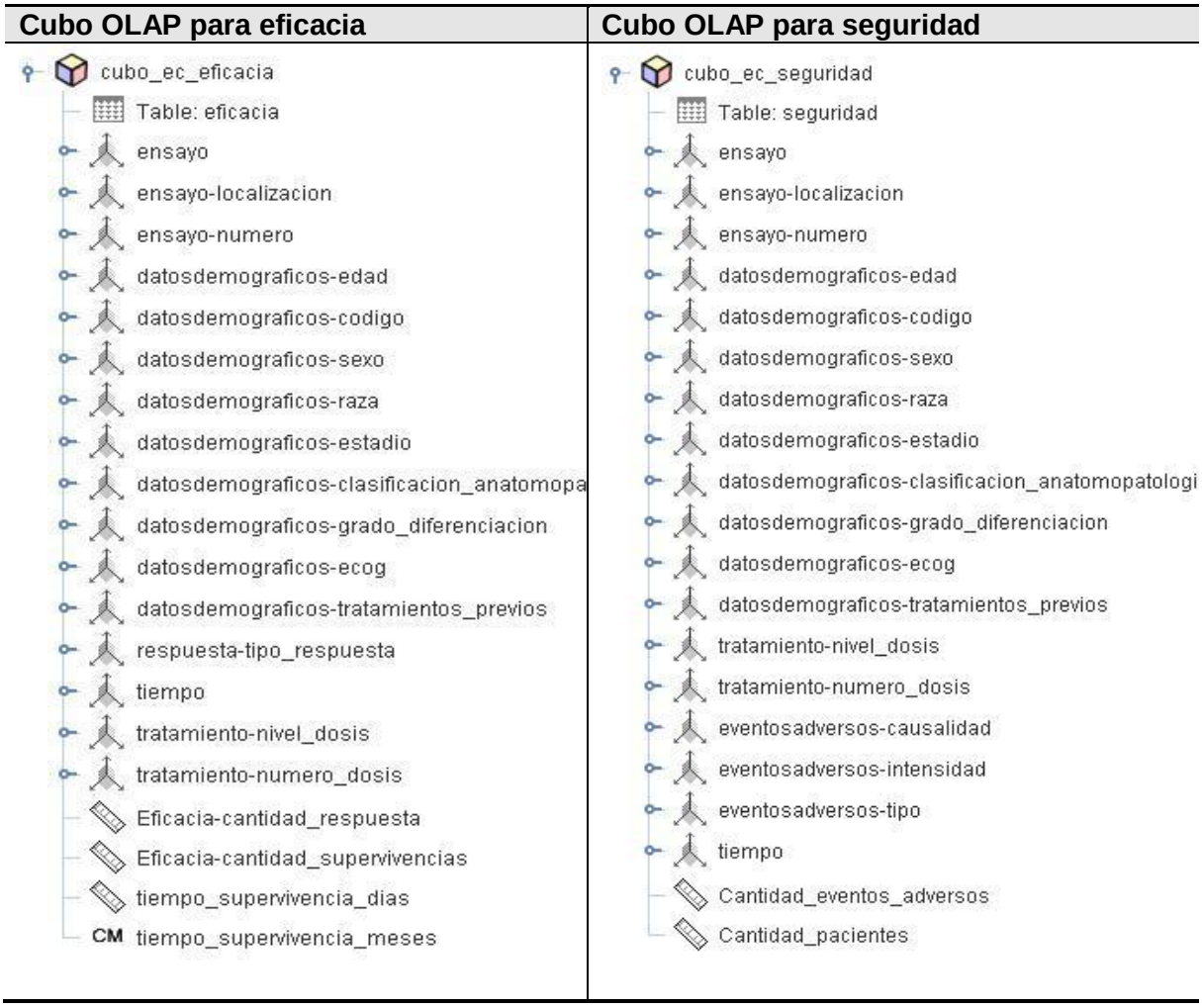

Figura 24. Cubos OLAP del producto hR3

La interfaz web para mostrar los datos

Una vez definidos los cubos OLAP se procede a conectar la interfaz web con los mismos para mostrar la información. En la figura 25 se muestra la interfaz de usuario desarrollada para permitir la navegabilidad de los usuarios por los reportes de análisis del DWH poblado para el producto hR3. Además de permite la autenticación de los usuarios, para usuarios no autorizados no puedan acceder a los informes.

Figura 25. Interfaz web para acceder al almacén de datos

La interfaz de usuario visualizador de reportes (Figura 26) permite mostrar el resultado del análisis realizado pues a través de la misma es posible ver los reportes y analizar la información mediante tablas de datos o gráficas de diversos tipos (barras, pastel, líneas); permitiendo además, desplegar el cubo de información, y modificar el reporte creando uno totalmente nuevo. Esto se logra configurando archivos del Jpivot donde se les especifica la ubicación de los cubos y el servidor de base datos donde están los datos del almacén .Por otra parte, esta interfaz brinda la posibilidad de cambiar el tipo de gráfica, imprimir el reporte o salvarlo en un archivo de formato PDF o XLS. Estas operaciones, entre otras, se pueden realizar accediendo mediante una serie de íconos, que se encuentran en la parte superior izquierda.

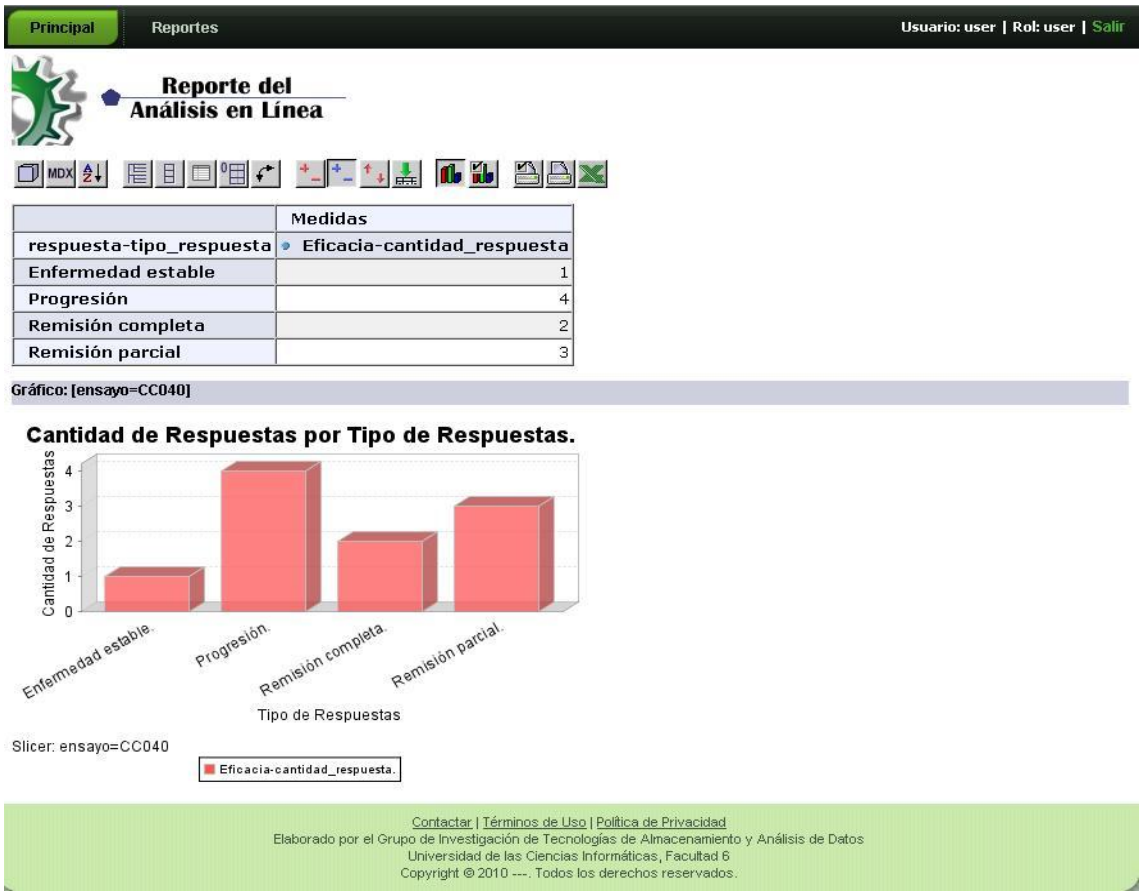

Figura 26 interfaz Jpivot

3.4 Conclusiones parciales

En este capítulo se realizó una comparación con los modelos más usados existentes para describir arquitectura y se demostró que las vistas del modelo satisfacen las necesidades de los clientes del CIM. Se describieron las características que debe cumplir la arquitectura del almacén de datos para ensayos clínicos del CIM. Se aplicaron las vistas propuestas para la descripción de la arquitectura almacenes de datos de los ensayos clínicos del CIM, se detalló la implementación del almacén de datos de ensayos clínicos del producto hR3 del CIM.

CONCLUSIONES

En la presente investigación con la elaboración del marco teórico se evidenció las deficiencias de los modelos existentes para describir la arquitectura de un almacén de datos para ensayos clínicos. Como propuesta de modelo para describir la arquitectura de almacenes de datos para ensayos clínicos se determinaron las vistas necesarias. Se logró aplicar el modelo propuesto mediante la descripción de la arquitectura para almacenes de datos de ensayos clínicos, se realizó una comparación con los modelos existentes y se detalló la implementación del DM del producto hR3 del CIM así como se realizo la evaluación de la arquitectura propuesta.

RECOMENDACIONES

Como recomendaciones de la presente investigación se propone:

- Agregar varios productos de ensayos clínicos del CIM sobre la arquitectura propuesta.
- Extender el modelo de descripción de arquitectura basado en vistas a otros dominios de almacenes de datos.

REFERENCIAS BIBLIOGRÁFICAS

1. Centro de Imunologia Molecular. [En línea] [Citado el: 15 de marzo de 2010.] www.cim.sld.cu.
2. **AZNAR-SALATTI, J.S.***Diseño de protocolos de un estudio clínico: las denominadas "case report forms" o cuadernos de recogida de datos.* 2007.
3. **Hilliard, Rich.***All About IEEE Std 1471.* 2007.
4. **Zaragoza, Francisco José Ortiz.***Arquitectura de Referencia para Unidades de Control de Robots de Servicio Teleoperados.* Cartagena : s.n., 2005.
5. **Inmon, Bill.***Building The Data Warehouse.* s.l. : Wiley Publishing, Inc, 2005.
6. **Ralph, Kimball**. *THE DATA WAREHOUSE STAGING TOOLKIT.* 2004.
7. Mónica Carreño león, Jesús Sandoval Gringas, José Torres Jimenez. Construcción de una bodega de datos para el proceso de autorización de gastos médicos. [En línea] [Citado el: 20 de mayo de 2010.] http://creaweb.ei.uvigo.es/creaweb/Asignaturas/PSI/sw/Art022.pdf.
8. Datawarehouse manager. [En línea] [Citado el: 25 de mayo de 2010.] http://www.dataprix.com/datawarehouse-manager.
9. **Paloma Sánchez López y Isabel Criado Gómez,** Ministerio de Trabajo y Asuntos Sociales. OLAP, ROLAP, MOLAP. *OLAP, ROLAP, MOLAP.* [En línea] Empresa Consultora Externa, NorSistemas, 2008. [Citado el: 10 de Febrero del 2010.] http://www.csae.map.es/csi/silice/DW2251.html.
10. *Clinical Trials.* [En línea] [Citado el: 8 de mayo de 2010.] http://clinicaltrials.gov/ct/info/whatis#whatis.
11. Sitio oficial del SEI . [En línea] [Citado el: 15 de marzo de 2010.] http://www.sei.cmu.edu/architecture/start/community.cfm.
12. **Luján, Sergio.***Data Warehouse Desing whit UML.* Alicante : s.n., 2005.
13. **Reynoso, Carlos Billy**. *De Lenguajes de descripción arquitectónica de.* Buenos Aires : s.n., 2004.
14. **Vallecillo, Lidia Fuentes y Antonio.** Una Introducción a los Perfiles UML. [En línea] [Citado el: 1 de junio de 2010.] http://www.lcc.uma.es/~av/Publicaciones/04/UMLProfiles-Novatica04.pdf.
15. **Medina, Cesar Julio Bustacara.** Evaluación de Arquitecturas de Software. [En línea] [Citado el: 30 de marzo de 2010.] http://sophia.javeriana.edu.co/~cbustaca/Arquitectura%20Software/Clases/Conceptos/Presentaciones/Analisis_AS.pdf.

BIBLIOGRAFÍA

1. Arquitectura del Data Warehouse. [En línea] [Citado el: 26 de mayo de 2010.] http://www.dataprix.com/ca/node/564.
2. **AZNAR-SALATTI, J.S.***Diseño de protocolos de un estudio clínico: las denominadas "case report forms" o cuadernos de recogida de datos.* 2007.
3. Centro de Imunologia Molecular. [En línea] [Citado el: 15 de marzo de 2010.] www.cim.sld.cu.
4. **Chaudhuri S, Dayal U.***An overview of data warehouse and OLAP technology.* 1997.
5. **Clements, Paul C.***Active Reviews for Intermediate Designs .* 2000.
6. *Clinical Trials.* [En línea] [Citado el: 8 de mayo de 2010.] http://clinicaltrials.gov/ct/info/whatis#whatis.
7. Datawarehouse manager. [En línea] [Citado el: 25 de mayo de 2010.] http://www.dataprix.com/datawarehouse-manager.
8. **Framework, The Open Group Architecture.***TOGAF Version 9.* 2009.
9. **Group, Object Management.***Common Warehouse Metamodel (CWM) Specification.* 2003.
10. **Gustavo A. Brey, Santiago Blanco.***Arquitectura de Proyectos de IT.* [En línea] [Citado el: 29 de mayo de 2010.] http://apit.wdfiles.com/local--files/start/04_apit_apunte_arq_de_sw_comunicacion.pdf.
11. **Hilliard, Rich.***All About IEEE Std 1471.* 2007.
12. **Hofmeister, C.** "*Applied Software Architecture".* Addison Wesley,2000.
13. IEEE recommended practice for architectural description of software-intensive systems. Technical report, 2000.
14. **Inmon, Bill.***Building The Data Warehouse.* s.l. : Wiley Publishing, Inc, 2005.
15. **Inmon, Bill.***Buiildiing the Data Warehouse:Gettiing Started.* 2005
16. **Ivan Tapia, Maria Ruiz, Edgar Ruiz.***Una metodologia para sectorizar pacientes en el consumo de medicamentos aplicando datamart y dataminig en un hospital.* 2007.
17. **Juan Trujillo, Emilio Soler.***Desarrollo de almacenes de datos dirigido por modelos.* 2005.
18. **Kruchten, Philippe**. *Planos Arquitectónicos: El Modelo de "4+1" Vistas de la Arquitectura del Software*. 1995.
19. **López, Asnioby Hernández.***ALMACENES DE DATOS APLICADA A LA SEGURIDAD CIUDADANA .* Ciudad de la Habana : s.n., 2009.
20. **Luján, Sergio.***Data Warehouse Desing whit UML.* Alicante : s.n., 2005.
21. **Mallach.***Decision Support and DataWarehouse System.* 2000.
22. **Medina, Cesar Julio Bustacara.** Evaluación de Arquitecturas de Software. [En línea] [Citado el: 30 de marzo de 2010.] http://sophia.javeriana.edu.co/~cbustaca/Arquitectura%20Software/Clases/Conceptos/Presentaciones/Analisis_AS.pdf.
23. Modelamiento multidimensional. [En línea] [Citado el: 10 de marzo de 2010.] http://www.inf.udec.cl/~revista/ediciones/edicion4/modmulti.PDF.

24. **Mónica Carreño león, Jesús Sandoval Gringas, José Torres Jimenez**. Construcción de una bodega de datos para el proceso de autorización de gastos médicos. [En línea] [Citado el: 20 de mayo de 2010.] http://creaweb.ei.uvigo.es/creaweb/Asignaturas/PSI/sw/Art022.pdf.
25. OMG-UML. [En línea] [Citado el: 1 de junio de 2010.] http://www.omg.org/technology/documents/profile_catalog.htm.
26. **P. Clements, D. Garlan, L. Bass, J. Sta_ord, R. Nord, J. Ivers, and R. Little,** "*Documenting software architectures: views and beyond"*. Pearson Education, 2002.
27. **P. Clements and R. Kazman**. *"Software Architecture in Practices"*. Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA, 2003.
28. **P. Clements, R. Kazman, and M. Klein**. "*Evaluating software architectures: methods and case studies"*. Addison-Wesley Reading, MA, 2002.
29. **Paloma Sánchez López y Isabel Criado Gómez,** Ministerio de Trabajo y Asuntos Sociales. OLAP, ROLAP, MOLAP. *OLAP, ROLAP, MOLAP.* [En línea] Empresa Consultora Externa, NorSistemas, 2008. [Citado el: 10 de Febrero del 2010.] http://www.csae.map.es/csi/silice/DW2251.html.
30. **R. Kazman, L. Bass, M. Webb, and G. Abowd**. "*SAAM: A method for analyzing the properties of software architectures"*. In Proceedings of the 16th international conference on Software engineering,. IEEE Computer Society Press, 1994.
31. **R. Kazman, M. Klein, and P. Clements.** *"ATAM: Method for architecture evaluation".* CMU/SEI, 2000.
32. **Ralph, Kimball**. "The Data Warehouse ETL Toolkit". s.l. : Wiley, 2004.
33. **Ralph, Kimball**. *THE DATA WAREHOUSE STAGING TOOLKIT.* 2004.
34. **Reynoso, Carlos Billy**. *De Lenguajes de descripción arquitectónica de.* Buenos Aires : s.n., 2004.
35. **Reynoso, Carlos Billy.***Introducción a la Arquitectura de Software.* BUENOS AIRES : s.n., 2004.
36. **Rick Kazman, Paul Clements, Mark Klein.***Evaluating Software Architectures. Methods and case studies. .* s.l. : Addison Wesley, 2001.
37. **Rivera, J. D.** *Utilización de información histórica para decisiones empresariales.* Santa Fé de Bogotá D.C. 2005
38. **Ross, Ralph Kimball y Margy.***The Data Warehouse Lifecycle Toolki.* s.l. : Wiley Publishing, Inc, 2002.
39. **Sánchez, Leopoldo Zenaido Zepeda.***Metodología para el Diseño Conceptual de Almacenes de Datos.* Valencia : s.n., 2008.
40. Sitio oficial de Pentaho**.** [En línea] 2009. [Citado el: 3 de mayo de 2010.] http://pentaho.org/.
41. Sitio oficial del SEI . [En línea] [Citado el: 15 de marzo de 2010.] http://www.sei.cmu.edu/architecture/start/community.cfm.
42. Sitio oficial de UML [En línea] [Citado el: 15 de marzo de 2010.] http://uml.org/
43. Sitio Oficial de Visual Paradimg [En línea] [Citado el: 15 de mayo de 2010.] http://www.visual-paradigm.com/

44. **Thomsen, Erik.** *OLAP Solutions. Building Multidimensional Information Systems.* NEW YORK , CHICHESTER , WEINHEIM, BRISBANE, SINGAPORE, TORONTO: Robert Ipsen, 2002.
45. **Vallecillo, Lidia Fuentes y Antonio.** Una Introducción a los Perfiles UML. [En línea] [Citado el: 1 de junio de 2010.] http://www.lcc.uma.es/~av/Publicaciones/04/UMLProfiles-Novatica04.pdf.
46. **Zaragoza, Francisco José Ortiz.***Arquitectura de Referencia para Unidades de Control de Robots de Servicio Teleoperados.* Cartagena : s.n., 2005.
47. **Zayas, Carlos Álvarez de.***METODOLOGIA DE LA INVESTIGACIÓN CIENTIFICA.* Santiago de Cuba : s.n., 1995.

ANEXO 1

Centro de Inmunología Molecular

Aval de aceptación

Ciudad de La Habana, junio de 2010

A través de la presente se certifica que el equipo de trabajo conformado por: Ing. Martha D. Hernández Ramírez, Ing. Anthony R. Sotolongo León, Themis P. Díaz Morales, José S. Bermúdez Rodríguez, Javier Rodríguez Sotolongo, Yohan O. Peralta Góngora, Yailín Simón Mir, Yoander Iñiguez Bermúdez; desarrolló un almacén de datos que permitió la integración de la información que se gestiona en los Ensayos Clínicos relacionados con los estudios realizados sobre el fármaco nimotuzumab (hR3). Esto permitirá un mejor análisis de los datos en el proceso de toma de decisiones en el departamento de Buenas Prácticas Clínicas de la dirección de investigaciones clínicas perteneciente al Centro de Inmunología Molecular (CIM).

El trabajo realizado consistió en proponer un procedimiento para cada una de las etapas de desarrollo de un almacén de datos, así como su estructura arquitectónica y soporte tecnológico. Cada uno de estos procedimientos servirá de guía a los especialistas del centro para la incorporación de la información relacionada con los estudios clínicos de los productos elaborados en el CIM. Asimismo, tanto la estructura arquitectónica como el soporte tecnológico se ajustan a las necesidades del centro y soporta el incremento de información perteneciente a los estudios realizados hacia el almacén de datos.

Se considera **satisfactorio** y de **vital importancia** el trabajo realizado por el equipo de desarrollo en aras de facilitar el análisis de los datos y con ello la toma de decisiones en el departamento de la dirección de investigaciones clínicas. Para que así conste se emite este aval de aceptación por parte de los especialistas del área de la institución.

Centro de Inmunología Molecular
DIRECCIÓN INVESTIGACIONES CLÍNICAS

Firma
Carmen E. Viada
Jefa del Grupo. Manejos de Datos Clínicos
Dir. Investigaciones Clínicas

Firma
Patricia Lorenzo-Luaces Alvarez
Jefa del Dpto. Buenas Prácticas Clínicas
Dir. Investigaciones Clínicas

GLOSARIO DE TÉRMINOS

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

CIM: Centro de Inmunología Molecular. Centro biotecnológico y científico - investigativo asociado a la producción. Creado en Cuba el 5 de diciembre de 1994.

DWH o almacén de datos: Datawarehouse. Almacén de datos que reúne la información histórica generada por todos los distintos departamentos de una organización, orientada a consultas complejas y de alto rendimiento.

DM: Es un subconjunto de los datos de un almacén de datos, en la forma de información resumida que soporta los requerimientos de un departamento o función de negocio particular.

EC: Ensayos Clínicos. Cualquier investigación en seres humanos dirigida a descubrir o verificar los efectos clínicos, farmacológicos u otros efectos fármacodinámicos de un producto en investigación.

Herramientas CASE: (ComputerAided Software Engineering): Constituyen un conjunto de ayudas para el desarrollo de programas informáticos, modelando los mismos.

Nimotuzumab (Hr3): medicamento que se utiliza como monoterapia o en combinación con radioterapia y/o quimioterapia para el tratamiento de cáncer de origen epitelial y gliomas.

Staging área: Es un área temporal de datos.

Stakeholders: Son las personas involucradas en el desarrollo de la solución, ya sea directa o indirecta.

UML: (por sus siglas en inglés, Unified Modeling Language) : lenguaje de modelado encargado de describir notaciones y procesos de los sistemas.