1(III), 15-20, 2012

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

Anthony Rafael Sotolongo León
Correo electrónico:asotolongo@uci.cu

Martha Denia Hernández Ramírez
Correo electrónico:mdhhernandez@uci.cu
Universidad de Ciencias Informática (UCI), La Habana, Cuba

Artículo Original

Resumen
La descripción correcta y detallada de la arquitectura de los sistemas informáticos resulta 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 necesita una descripción detallada de la arquitectura. Ralph Kimball propone los aspectos a tener en cuenta para la descripción pero no expone con qué realizarla. Existen modelos específicos para describir la arquitectura como es el 4+1 vistas de Kruchten o el metamodelo 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 este artículo se propone un modelo basado en vistas para describir la arquitectura de almacenes de datos que se ajusta a las necesidades del Centro de Inmunología Molecular siguiendo el marco de referencia de Kimball y usando como lenguaje de modelado UML 2.0.

Palabras clave: arquitectura de almacenes de datos, vistas arquitectónicas, ensayos clínicos

Recibido: 10 de noviembre del 2011 Aprobado: 18 de diciembre del 2011

INTRODUCCIÓN

En la actualidad, la humanidad enfrenta diversas enfermedades que constituyen causa de muerte de millones de personas. El número de decesos por enfermedades crónicas es muy alto en el mundo; pero a la par de esta situación los científicos realizan investigaciones y crean medicamentos para combatirlas.

En Cuba se funda de manera oficial el 5 de diciembre de 1994 el Centro de Inmunología Molecular (CIM) con el propósito de estudiar el comportamiento de estas enfermedades y crear biofármacos que puedan combatirlas, para 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 y ponerlos al servicio de la salud pública cubana; también se encarga de hacer la actividad científica y productiva económicamente sostenible y realizar aportes importantes a la economía del país [1]. El CIM ha desarrollado además varios productos como el anticuerpo monoclonal anti CD3 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.

Para el estudio y aprobación de los fármacos elaborados se realiza en varios hospitales del país y en algunos en el exterior, diferentes 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 [2]. Por tanto, es un estudio experimental, analítico, prospectivo, controlado y con tamaños muestreales suficientes. Estos 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 debe quedar bien descrita la manipulación de esa información en todas las gestiones que se realicen con esos datos.

La información es recogida en los cuadernos de recogida de datos (CRD), formulario diseñado para anotar las variables obtenidas durante un ensayo clínico [3], este formulario es diseñado de modo diferente en dependencia de lo que se quiera lograr, por tanto no tienen que coincidir las variables, con 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 (AD) por todas las ventajas que tienen los mismos para la toma de decisiones. Pero la entidad no dispone de una arquitectura ni soporte tecnológico para su implementación. 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 si no 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. Cabe destacar la importancia que se le da a estos datos en el CIM, pues en ellos se encuentra 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, así como se regula el acceso de los datos por los diferentes especialistas y aplicaciones que interactúan con estos.

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 qué realizar esta descripción.

Una alternativa viable para describir la arquitectura de software son las vistas o modelos de vistas utilizados por la comunidad de arquitectos de software. Aún cuando existe un consenso general acerca de la necesidad de representar la arquitectura utilizando diferentes vistas, cuando dicho consenso desaparece hay que definir cuáles son esas vistas. El estándar IEEE-1471-2000 [4] 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. [5]

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 los almacenes giran alrededor de los datos. Además, existe el Common Warehouse Metamodel (CWM)[6], que aunque no es un modelo para describir arquitectura, permite concebir almacenes de datos e intercambiar información sobre los mismos, este metamodelo no realiza la descripción basada en vistas sino por escenarios tales como: ETL, OLAP, cuestionarios, administración y herramientas. Con este metamodelo se ignoran 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: *¿Cómo describir la arquitectura de sistemas de información basada en almacenes de datos para ensayos clínicos del CIM*?

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

MATERIALES Y MÉTODOS

Arquitectura de almacenes de datos

Los AD, según Inmon [7], son un conjunto de datos con las siguientes características: orientado a temas, variante en el tiempo, no volátil e integrado. Otra definición muy aceptada es la que presenta Kimball [8], y plantea que es: "…una copia de las transacciones de datos específicamente estructurada para la consulta y el análisis…". Ambos autores plantean que es un conjunto de datos con fines de análisis de un tema específico.

La arquitectura de los AD 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 1. Además, es típica e incluye lo siguiente [9]:

- Datos operacionales: Origen de datos o datos fuentes de donde se extraerán los datos primarios.
- Extracción, transformación y carga (ETL de sus siglas en inglés): 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.

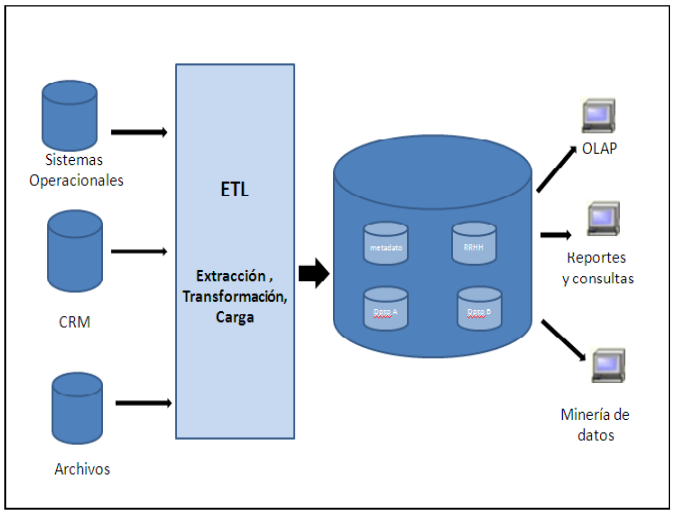

Fig. 1. Arquitectura básica de un almacén de datos.

Marco de descripción de arquitectura de almacenes de datos

El marco de descripción de arquitectura es el punto de vista sobre el cual se debe hacer la misma. En Luján [10] se plantea un marco de referencia de arquitectura para AD. En el mismo se proponen tres niveles fundamentales en la arquitectura:

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.

Otro marco de referencia es el planteado por Kimball[11]. En él 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.

Se decide utilizar para definir y describir la arquitectura del almacén de datos el marco de referencia de Kimball [11] pues plantea una serie de aspectos muy importantes de esta área, además, este marco hace un énfasis mayor en las herramientas a utilizar además de estar bien documentado, 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, no expresa qué artefactos generar para el apoyo a la descripción.

Modelos de descripción de arquitectura

Existen varios modelos donde se definen ciertas vistas y diagramas; correspondientes a esta investigación se estudian el modelo 4+1 vistas de Krutchen [5] y el CWM [6]. El primero es de gran aceptación por los arquitectos de sistemas clásicos de gestión donde hay una gran cantidad de usuarios interactuando y operaciones de todo tipo, además, la lógica de negocio está bien definida. Analizando el tema de AD donde todo fluye alrededor de los datos y los ensayos clínicos en el CIM, en el cual un aspecto importante es el flujo y acceso a los datos por las diferentes partes del sistema, cuestiones que no las contempla el 4+1 vistas. El CWM aunque no es exactamente para describir arquitectura, ha sido utilizado por los diseñadores de AD para intercambiar información y describir aspectos de los mismos, y no tiene en cuenta puntos relacionados con el hardware, tema 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. A partir de esto, se elaborará un modelo para la descripción de la arquitectura de un AD que pueda ser aplicado a la gestión de los EC en el CIM.

Lenguajes para el modelado de la arquitectura

En un esfuerzo para estandarizar las notaciones y procesos a utilizar para describir arquitecturas se define el lenguaje de descripción de arquitectura [12] (ADL por sus siglas en inglés, Architecture Description Language) y el lenguaje unificado de modelado [13] (UML por sus siglas en inglés, Unified Modeling Language).

El UML 2.0 fue liberado en el 2003 y cumple con la mayoría de las características de los ADL: además, hace grandes adelantos para convertirse en una notación para describir arquitecturas. La metodología 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 , por otro lado los ADL difieren unos de otros y en ocasiones son difíciles de comprender , por tanto se decide utilizar UML 2.0 como lenguaje de modelado.

Los perfiles UML son una posibilidad que proporciona el propio UML para ampliar su sintaxis y su semántica con el objetivo de expresar los conceptos específicos de un determinado dominio de aplicación, El hecho de que UML haya sido un lenguaje diseñado de propósito general 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]

1. Definir un nuevo lenguaje.
2. Extender el propio UML

Decidirse por una alternativa u otra trae sus inconvenientes si bien con la primera se puede lograr un modelado a la medida del dominio lo que puede traer consigo que las herramientas CASE no puedan ser utilizadas, 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, los autores proponen extender UML a una terminología ya existente que se ajuste a las necesidades del dominio a analizar siempre que sea posible, esto traería consigo que se pueda modelar en las herramientas CASE existentes.

RESULTADOS Y DISCUSIÓN

Propuesta de modelo de descripción de arquitectura de AD

Siguiendo el estándar IEEE-1471-2000 se definen una serie de pasos para determinar las vistas de la arquitectura para que la misma posea una buena documentación. Este estándar queda resumido en la figura 2.

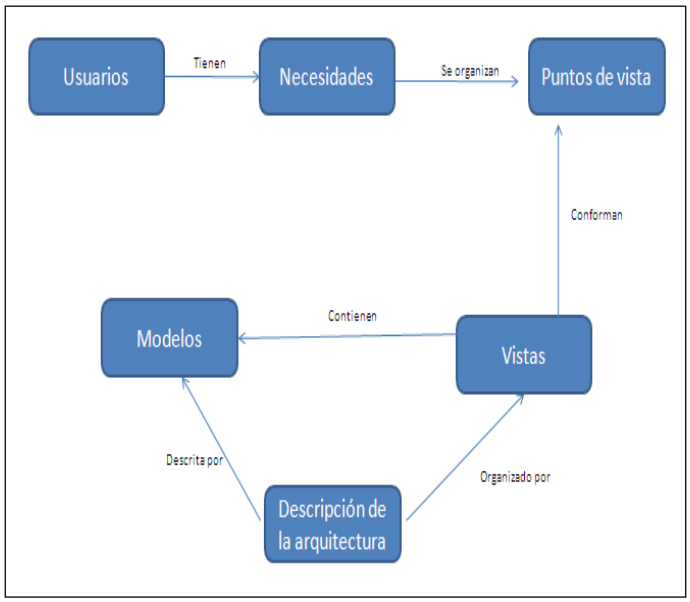

Fig. 2. Estándar IEEE 1471-2000.

Usuarios: Expertos del CIM, diseñadores de base de datos, integradores de datos, arquitectos, especialistas de hardware. [15]

Necesidades:

- Conocer cómo van a fluir los datos por la solución,
- Cómo va a ser el acceso a los mismos por los diferentes componentes
- Cómo los usuarios los van a manipular,
- Además de los productos necesarios, la relación entre ellos.
- Conocer las características de los datos a almacenar y las estructura de los mismos.
- Saber las responsabilidades de cada parte de la solución, su alcance, sus dependencias y la maneras de comunicación (protocolos de seguridad).
- Saber la ubicación lógica de componente.
- Disposición de hardware que se requiere.
- Distribución de los componentes por hardware.

Puntos de vistas: Punto de vista de datos, punto de vista técnico y punto de vista de infraestructura. [15]

- Punto de vista de datos: corresponde con la columna de datos (¿el qué?). Describirá la estructura de los datos.
- 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.
- 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.

Vistas y modelos: [15]

- Punto de vista de datos: Vista de datos.
- Punto de vista técnico: Vista de flujo de datos, vista de de implementación y vista lógica de implementación, vista de acceso a datos.
- Punto de vista de infraestructura: vista de despliegue, Vista de despliegue por componentes.

En la figura 3 se puede observar un resumen del modelo y sus diagramas. [15]

El estándar UML 2.0 no contempla diagramas UML para acceso a datos por usuarios y componentes, por lo cual se realiza una extensión y se definen dos diagramas nuevos UML:

- Diagramas de acceso a datos por usuarios y componentes.

En las figuras 4 y 5 se muestran los componentes y su relación en los respectivos diagramas

En la tabla 1 se realiza una comparación de los modelos descritos anteriormente y el modelo propuesto para describir arquitectura de almacenes de datos de EC en el CIM.

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

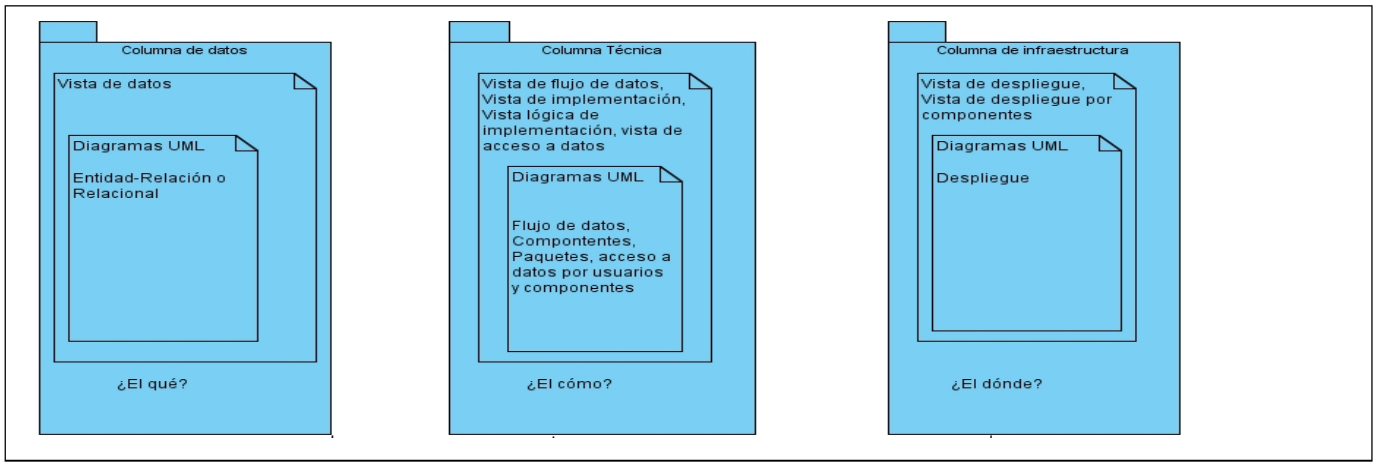

Fig. 3. Resumen del modelo.

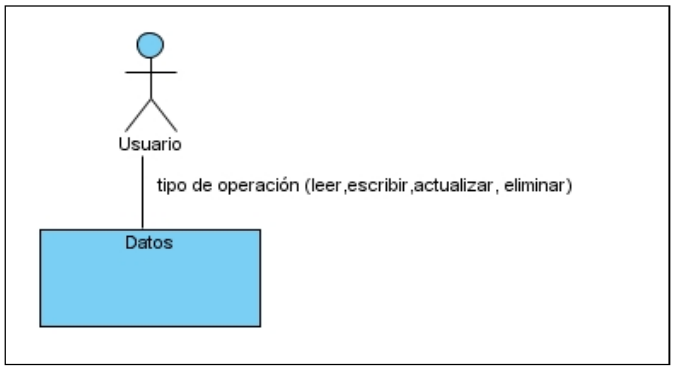

Fig. 4. Diagrama de acceso a datos por usuarios.

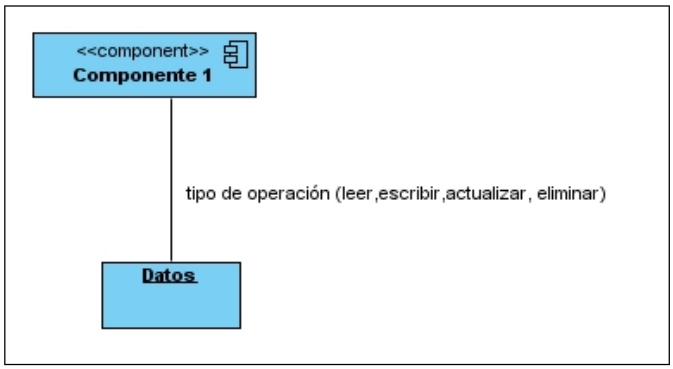

Fig. 5. Diagrama de acceso a datos por componentes.

Tabla 1
Comparación de los modelos

Diagramas	El modelo 4+1 de Krutchen	CWM	Modelo para describir almacenes de datos de EC en el CIM
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

El modelo fue aplicado en el CIM para describir la arquitectura del AD del producto Nimotuzumab (hR3) que fue explicado por A. Sotolongo [15] y permitió facilitar la implementación del almacén y entregar a los especialistas del CIM los datos históricos integrados del producto, para realizar análisis de los mismos.

CONCLUSIONES

En el presente artículo se evidenciaron las deficiencias de los modelos existentes para describir la arquitectura de un almacén de datos para ensayos clínicos en una institución especializada como es el Centro de Inmunología Molecular de La Habana.

Basado en el estudio presentado, se propone la utilización de un modelo especializado para describir la arquitectura de almacenes de datos necesaria para el control de los ensayos clínicos fundamentado en el estándar IEEE 1471 y utilizando UML 2.0 compuesto por tres puntos de vistas y siete diagramas, el cual cubre las necesidades y demandas de los sistemas de almacenes de datos del CIM de La Habana.

RECONOCIMIENTOS

Los autores agradecen a los trabajadores del departamento de Informática y Ensayos Clínicos del Centro de Inmunología Molecular, que hicieron posible la realización de este trabajo.

REFERENCIAS

1. CIM. Página de Presentación [En línea] [Citado el: 15 de marzo de 2010.]http://www.cim.co.cu
2. Clinical Trials. Understanding Clinical Trials. [En línea] [Citado el: 8 de mayo de 2010.] Http://clinicaltrials.gov/ct/info/whatis#whatis.
3. **AZNAR-SALATTI, J. S.** "Diseño de protocolos de un estudio clínico: las denominadas case report forms o cuadernos de recogida de datos", *Revista: Jano*, 2007, registro 591.
4. **HILLIARD, Rich.** *All About IEEE Std 1471*. 2007.
5. **ZARAGOZA ORTIZ, Francisco José.** Arquitectura de Referencia para Unidades de Control de Robots de Servicio Teleoperados. Cartagena, 2005.
6. Group, Object Management.Common Warehouse Metamodel (CWM) Specification. 2003.
7. **INMON, Bill.** *Building the Data Warehouse*. s.l.: Wiley Publishing, Inc, 2002.p 428 , ISBN 0-471-08130-2.
8. **KIMBALL, Ralph.** *The Kimball Group Reader: Relentlessly Practical Tools for Data Warehousing and Business Intelligence*, Wiley, 2010, p. 744, ISBN 978-0470563106
9. **SEN, A. and SINHA, A. P.** A *Comparison of Datawarehousing Methodologies, Communication of the ACM*, pp. 79-84, 2005.
10. **LUJÁN, Sergio.** "Data Warehouse Desing whit UML. Tesis doctoral, Alicante, s.n., 2005.
11. **KIMBALL, Ralph.** *The Data Warehouse Lifecycle Toolkit*, 2nd. Edition, Wiley, 2008, p. 636, ISBN 978-0470149775.
12. **REYNOSO BILLY, Carlos .** De *lenguajes de descripción arquitectónica,* Buenos Aires, s.n., 2004.
13. **OMG-UML**. Catalog Of UML Profile Specifications, [En línea] [Citado el: 1 de junio de 2010.] http://www.omg.org/technology/documents/profile_catalog.htm.
14. **VALLECILLO, Lidia y FUENTES, 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. **SOTOLONGO, A.** "Modelo de descripción de arquitectura de almacenes de datos para ensayos clínicos del Centro de Inmunología Molecular". Tesis de maestría, La Habana. Universidad de las Ciencias Informáticas, 2010.

AUTORES

Anthony Rafael Sotolongo León
Ingeniero Informático, Máster en Ciencias, Profesor Asistente, Universidad de las Ciencias Informáticas (UCI), La Habana, Cuba.

Martha Denia Hernández Ramírez
Ingeniera Informática, Máster en Ciencias, Instructora, Universidad de las Ciencias Informáticas (UCI), La Habana, Cuba.

Description Model of Warehouse Architecture for Clinical Test at the Molecular Immunology Centre

Abstract

Accurate and detailed description of the architecture of computer systems is very important to achieve success in their development. As informatic solutions, data warehouses and software support decision-making in institutions that need to implement a detailed description of the architecture. Ralph Kimball proposes the aspects to be considered of the description and explains how it is done. There are specific models used to describe the architecture such as Kruchten 4 +1 views of meta-model or the Common Warehouse Metamodel (CWM) however these models do not meet the need of the description that requires a data warehouse that integrates information from clinical trials of the Molecular Immunology Centre (CIM). In this paper we propose a model for describing the data warehouse architecture that fits the needs of the Molecular Immunology Center following the Kimball framework and using as UML 2.0 modeling language.

Key words: datawarehouse architecture, architecture views, clinical trial