Incorporación de Seguridad en el Modelado Conceptual de Procesos Extracción-Transformación-Carga

Rodolfo H. Villarroel⁽¹⁾, Yessica M. Gómez⁽²⁾ y Constanza B. Krause⁽¹⁾

(1) Pontificia Universidad Católica de Valparaíso, Escuela de Ingeniería Informática, Facultad de Ingeniería, Avenida Brasil 2241, Valparaíso-Chile (e-mail: rodolfo.villarroel@ucv.cl; constanza.krause.l@mail.pucv.cl)

(2) Universidad Católica del Maule, Escuela de Ingeniería Civil Informática, Facultad de Ciencias de la Ingeniería, Avenida San Miguel 3605, Talca-Chile (e-mail: jgomez@ucm.cl)

Recibido Feb. 04, 2013; Aceptado Mar. 15, 2013; Versión final recibida Abr. 10, 2013

Resumen

Este artículo presenta un análisis comparativo de propuestas de modelado conceptual de procesos Extracción-Transformación-Carga (ETL) y propuestas que incorporan seguridad en almacenes de datos y procesos ETL. El método utilizado considera las características de las propuestas y sus principales actividades en el modelado de procesos ETL. También considera propuestas que incorporan seguridad en el modelado de almacenes de datos y procesos. Ante la carencia de propuestas existentes, se concluye que es necesaria la incorporación de seguridad en el modelado de procesos ETL para contribuir en la obtención de almacenes de datos seguros.

Palabras clave: procesos ETL, almacenes de datos, seguridad, modelado conceptual

Incorporation of Security in the Conceptual Modeling of Extraction-Transformation-Loading Processes

Abstract

This paper presents a comparative analysis of proposals for the conceptual modeling of Extraction-Transformation-Loading processes (ETL) and proposals that incorporate security in data warehouses and ETL processes. The method used considers the characteristics of the proposals and their principal activities in the ETL process modeling. Furthermore, considers proposals that incorporate security in the modeling of data warehouses and ETL processes. Considering the lack of existing proposals, it is concluded that is necessary to incorporate security in the ETL processes modeling to assist in obtaining secure data warehouses.

Keywords: ETL processes, data warehouses, security, conceptual modeling

INTRODUCCIÓN

La seguridad de la información se define como la protección de la información computacional frente a consultas no autorizadas, modificaciones inapropiadas o la falta de disponibilidad de un servicio en un momento dado. Por lo tanto, se tienen los siguientes componentes: a) Confidencialidad: Prevenir/detectar/impedir la revelación impropia de la información, b) Integridad: Prevenir/detectar/impedir la modificación indebida de la información, c) Disponibilidad: Prevenir/detectar/impedir la denegación de acceso a los servicios proporcionados por el sistema.

Los componentes mencionados deberían estar presentes en todo desarrollo de sistemas de información. Sin embargo, sólo se considera una vez que el sistema se entrega para su utilización, por lo tanto, es poco frecuente que los desarrolladores se preocupen en etapas anteriores, tales como el análisis y diseño. Las soluciones se enfocan principalmente a proporcionar defensas de seguridad en vez de solucionar una de las causas fundamentales de problemas de seguridad, que se refiere a un buen diseño de software. Diversos trabajos se refieren a la importancia de la seguridad en el proceso de desarrollo de software. En Ghosh et al. (2002) se afirma que la seguridad debería influir en todos los aspectos del diseño, la implementación y las pruebas de software. En Franco et al. (2012) se propone una metodología para la detección de vulnerabilidades en redes de datos, esta metodología presenta un enfoque práctico y conceptual para la detección y erradicación de vulnerabilidades. Hall y Chapman (2002) proponen ideas de cómo construir sistemas correctos que no sólo cumplan los requisitos normales sino también los de seguridad. Estas ideas se basan en la utilización de diversas técnicas formales de representación de requisitos, y en un fuerte análisis para la corrección de cada etapa. No obstante, se descuida la seguridad de bases de datos y almacenes de datos en relación con el diseño, enfocando la seguridad de los datos desde un punto de vista criptográfico, es decir, haciendo uso de métodos y técnicas con el objeto principal de cifrar y/o proteger un mensaje o archivo por medio de un algoritmo, usando una o más claves. Chung et al. (2000) también enfatizan la integración de los requisitos de seguridad en el diseño, ofreciendo a los diseñadores modelos que especifican aspectos de seguridad, pero sin abordar temas específicos de bases de datos y almacenes de datos.

En la literatura se pueden encontrar diversas iniciativas para incluir seguridad en almacenes de datos. Muchas de ellas se enfocan a aspectos específicos relacionados con el control de acceso, la seguridad multinivel, sus aplicaciones en bases de datos federadas, y aplicaciones con herramientas comerciales. Sin embargo, no se han considerado los aspectos de seguridad desde las primeras fases de diseño, teniendo en cuenta los requisitos de los usuarios (Fernández-Medina et al., 2007, Rodríguez et al., 2011). En la literatura relevante podemos encontrar muchas iniciativas para incluir seguridad en almacenes de datos (Blanco et al., 2010, Soler et al., 2008, Fernández-Medina et al., 2007), sin embargo, es escasa la literatura que considera los aspectos de seguridad en procesos ETL.

Existe un trabajo previo relacionado con el tema de seguridad en almacenes de datos, pero no con procesos ETL. La investigación de Villarroel et al. (2005) y Villarroel et al. (2006) establece las bases para proveer un modelo conceptual para el diseño de almacenes de datos seguros. Esta propuesta anterior corresponde a la alineación del enfoque de modelado conceptual de almacenes de datos seguros con el enfoque MDA (Model Driven Architecture) (Kleppe et al., 2003). El modelo conceptual de almacenes de datos seguros corresponde a un modelo independiente de plataforma generado a partir de un modelo de diseño basado en UML y a un modelo de seguridad denominado Modelo de Control de Acceso y Auditoría. La disciplina de modelado conceptual está logrando una mayor relevancia gracias a MDA, que promueve el uso de modelos en todos los pasos de un proyecto de desarrollo de software (Nelson et al., 2012). Además, Moody (2005) destaca la importancia de la calidad de los modelos conceptuales, indicando que un modelo conceptual de poca calidad puede incrementar el esfuerzo de desarrollo o resultar en un sistema que no satisface a los usuarios.

Como Kimball señala, "Nuestro trabajo como diseñadores de almacenes de datos es comenzar con las fuentes de datos existentes que se han utilizado" (Kimball y Caserta, 2004). Un sistema ETL sólido, bien diseñado y documentado se necesita para el éxito de un proyecto de almacenes de datos (El-Sappagh et al., 2011, Mrunalini et al., 2013). Sin embargo, poco esfuerzo se ha dedicado a proponer modelos conceptuales y modelos lógicos para definir formalmente estos procesos ETL. Aunque el área de procesos ETL es muy importante, ésta tiene poca investigación debido a la dificultad y carencia de un modelo formal para representar los escenarios ETL (El-Sappagh et al., 2011, Simitsis y Vassiliadis, 2008).

El mejor avance en esta línea de investigación ha sido realizado por el Laboratorio de Sistemas de Bases de Datos y Conocimiento de la Universidad Técnica Nacional de Atenas. En particular, ellos propusieron un modelo conceptual que provee su propia notación gráfica que permite al diseñador definir formalmente la mayoría de los problemas técnicos usuales relacionados con procesos ETL (Vassiliadis et al., 2002).

Además, este enfoque presenta una herramienta ETL llamada ARKTOS como un marco de trabajo fácil para el diseño y mantenimiento de estos procesos ETL (Vassiliadis et al., 2001). Sin embargo, no está considerada la seguridad en estas propuestas.

Los trabajos más recientes están asociados a un estudio sistemático de propuestas de modelado conceptual de procesos ETL, que reconoce la necesidad de un mayor esfuerzo para cubrir el vacío relacionado al modelado de procesos ETL (Muñoz et al., 2011), y a presentar un modelo de simulación para la extracción de datos seguros en procesos de Extracción, Transformación y Carga (Mrunalini et al, 2007). Este último artículo trata el tema de seguridad mediante la simulación de escenarios con restricciones de seguridad. Además, indican que el escenario es más complicado en el caso de procesos ETL debido a su heterogeneidad y naturaleza distribuida. Posteriormente, Mrunalini et al. (2013) automatiza y simula el proceso de evaluación de la seguridad en un sistema ETL y Anand (2012) presenta una visión general sobre la implementación física de un flujo de trabajo ETL seguro. Se visualiza, por lo tanto, la necesidad de profundizar en la investigación respecto al modelado conceptual de procesos ETL, y en la incorporación de seguridad en el modelado de procesos ETL. Para esto, se realizará una comparativa de propuestas de modelado de procesos ETL y el análisis de propuestas que incorporen seguridad en el modelado de almacenes de datos y en procesos ETL.

La estructura de este artículo es como sigue. En la sección siguiente se describen conceptos asociados a almacenes de datos y procesos ETL. Luego, se presentan las propuestas de modelado conceptual de procesos ETL. Sigue una sección con la comparativa de las propuestas mencionadas. Posteriormente, hay una sección que analiza las propuestas que incorporan seguridad en el modelado de almacenes de datos y procesos ETL. Finalmente, una sección de conclusiones y referencias.

ALMACENES DE DATOS Y PROCESOS ETL

Para autores como Ralph Kimball los Almacenes de Datos son “una copia de los datos transaccionales estructurados específicamente para consultas y análisis” (Kimball y Caserta, 2004), mientras que Bill Inmon define el término Almacén de Datos como: “una colección de datos orientados por temas, integrados, variables en el tiempo y no volátiles para el apoyo de la toma de decisiones” (Inmon, 2005). Los Almacenes de Datos son integradores, ya que su contenido proviene de diversas fuentes de datos como: Sistemas heredados, Archivos de Textos, Base de Datos Relacionales, ERP, entre otras posibilidades. La forma de lograr esta integración es a través del uso y desarrollo de los Procesos ETL. Estos procesos son los encargados de la extracción de los datos desde sus fuentes de origen, de transformarlos a la información deseada, de lograr la limpieza necesaria en aquellos que lo requieran y finalmente cargar al Almacén de Datos deseado, el que será utilizado con alguna finalidad como análisis en un área de las ventas de una corporación o el estudio de tendencias de alguna consultora (Muñoz et al., 2011).

Los procesos ETL cargan periódicamente los datos desde las fuentes a los almacenes de datos, siempre procurando registrar la variación que tenga el dato a lo largo del tiempo, lo cual determina que los almacenes de datos solamente sean utilizados para la lectura de la información y no como una base de datos donde se modifiquen los datos. Desde esta perspectiva, los procesos ETL son trascendentales ya que cumplen con la importante labor de recoger los datos y posicionarlos en el lugar que será la fuente del conocimiento para quienes realicen análisis sobre ellos. Las personas que construyan el ETL, y la tecnología anexa para la visualización de los datos, serán los responsables de generar toda la información necesaria para su posterior análisis. Es por esto que se considera que el diseño y mantenimiento de los procesos ETL son factores clave en el éxito de algún proyecto que involucre el uso de Almacenes de Datos (March y Hevner, 2007).

Antes de empezar cualquier proyecto de desarrollo que involucre la implementación de un Proceso ETL se debe tomar en cuenta el diseño y los alcances que tendrá este proceso, que como en todos los proyectos informáticos, tienen directa relación con la consolidación de la información en un Almacén de Datos. Como el objetivo de las estructuras multidimensionales es el análisis de la información, resulta importante que antes de iniciar la elaboración de un proceso ETL se entiendan las necesidades que debe solventar el análisis, así saber qué elementos formarán parte del Proceso ETL, cuáles serán las fuentes desde donde se obtendrán los datos, qué transformaciones y limpiezas deben ser llevadas a cabo, las características de la estructura que soportarán los datos para ser consultados, cómo se mostrará la información y cómo podrá ser consultada.

Las funcionalidades propias de la Extracción, Transformación y Carga son (Vassiliadis y Simitsis, 2002): a) La identificación de la información relevante desde el lado de las fuentes, b) La extracción de esa información, c) La personalización e integración de la información proveniente de múltiples fuentes con múltiples formatos. d) La limpieza del conjunto de los datos resultantes, sobre la base de datos base y las

reglas del negocio, e) La propagación de los datos al Almacén de Datos. Por lo tanto, se pueden identificar tres tareas importantes dentro del Proceso ETL: a) La extracción de los datos desde diferentes fuentes, b) La propagación de los datos al DSA, donde éstos son transformados y limpiados, c) La carga de los datos al Almacén de Datos. Cada una de estas tareas, determina la estructura básica de un Proceso ETL, lo cual estructura de manera clara cuales son las actividades necesarias de cualquier Proceso ETL.

Para la construcción de un Almacén de Datos, una de las actividades que más consume tiempo, es la elaboración del Proceso ETL, y a pesar de ser una parte importante de toda construcción de un Almacén de Datos, existe poca investigación con respecto a éste, debido principalmente a que no existe un modelo formal para la representación de las actividades del Proceso ETL que permitan que los datos de entrada de diferentes fuentes estén en un formato adecuado para la carga en el Almacén de Datos (El-Sappagh et al., 2011). Una carencia mayor se da respecto a la investigación sobre la incorporación de seguridad en procesos ETL. En este sentido, Mrunali et al (2013) indican que los actuales enfoques para el modelado conceptual de ETL no consideran los aspectos de seguridad en el modelado conceptual. Además, Patil et al. (2012) destaca que es importante identificar a los posibles agentes responsables de la fuga de datos cuando se externalizan los datos de los almacenes de datos y de procesos ETL.

PROPUESTAS EXISTENTES DE MODELADO CONCEPTUAL DE PROCESOS ETL

La importancia que tienen los Procesos ETL no guarda relación con el número de investigaciones llevadas a cabo en torno a este tema. Pero existen esfuerzos que, desde diferentes perspectivas, tratan el tema de los Procesos ETL y de esta manera se enriquece el conocimiento en torno a este tema.

La simbología que identifica a cada propuesta en las Tablas 1 y 2 es la siguiente:

(a) Una Metodología para el Modelado Conceptual de Procesos ETL (Simitsis y Vassiliadis, 2003)

La Metodología propuesta por Simitsis y Vassiliadis para el desarrollo de Procesos de ETL consiste en el desarrollo del modelo por medio de una serie de pasos que llevarán a la especificación de atributos, es decir se busca establecer la estructura y contenidos de las fuentes de datos existentes y su relación con el Almacén de Datos. Los pasos que esta Metodología indica son: a) La identificación de los Almacenes de Datos apropiados, b) La identificación de Candidatos y Candidatos Activos, c) La relación entre los proveedores y consumidores, d) La descripción del diagrama con las restricciones en tiempo de ejecución.

(b) Modelado Conceptual de Procesos ETL (Vassiliadis et al., 2002)

El Modelo Conceptual ideado por Vassiliadis, Skiadopulos y Sellis se enfoca en el desarrollo de las actividades de ETL, presentando las diferentes capas de instanciación y generalización. El Modelo establece una representación gráfica que define la semántica de un Proceso ETL, además se definen elementos para el Modelado como: Concepto, Atributo, Transformación, Restricción ETL. A todo lo anterior, se agrega una paleta de actividades ETL de uso frecuente, que se personaliza para la búsqueda de relaciones entre los atributos y actividades.

(c) Un Enfoque Basado en UML para el Modelado de Procesos ETL en Almacenes de Datos (Trujillo y Luján-Mora, 2003)

Los autores Trujillo y Luján-Mora identifican la posibilidad que al momento del diseño del proceso ETL se puede descomponer en un conjunto de procesos más sencillos, de esta manera se facilita el diseño y mantenimiento. Aquí se propone el uso de UML para el modelado de las actividades del proceso ETL.

Se desarrolla una paleta de actividades comunes en un proceso ETL tales como: Agregación (Aggregation), Conversión (Conversion), Filtrar (Filter), Incorrecto (Incorrect), Unir (Join), Cargador (Loader), Registro (Log), Integrar (Merge), Sustituto (Surrogate), Envoltorio (Wrapper). También se propone el uso de paquetes de manera de agrupar los elementos a utilizar en procesos de ETL.

(d) Diseño de Procesos ETL Usando Tecnologías de Web Semántica (Skoutas y Simitsis, 2006)

Los autores Skoutas y Simitsis proponen un Modelo Conceptual de Procesos ETL utilizando tecnologías de Web Semántica, donde su creación fue a través del uso de Lenguaje de Ontología Web. Se describen las transformaciones apropiadas para la integración de las fuentes de datos y la carga de los datos al Almacén de Datos. Otro aspecto es la definición de un conjunto de operadores comúnmente utilizados en los procesos ETL, asociados a filtrado, unión, agregación y función.

(e) *Diagramas de Mapeo de Datos para el Diseño de Almacenes de Datos con UML* (Luján-Mora et. al., 2004)

En este artículo los autores Luján-Mora, Vassiliadis y Trujillo proponen realizar el Modelado Conceptual enfocado a solventar la manipulación de los datos en su ámbito de la granularidad. Se propone el diagrama de mapeo de datos (data mapping), donde se presentan las reglas de transformación y sus niveles son: a) El nivel 1 contiene el esquema del Almacén de Datos, b) En el nivel 2 se detallan las relaciones existentes entre las tablas de origen y las del Almacén de Datos, c) En el nivel 3 se detallan las transformaciones, d) En el nivel 4 se presenta la relación entre los atributos del diagrama de mapeo de datos.

COMPARATIVA DE PROPUESTAS DE MODELADO CONCEPTUAL DE PROCESOS ETL

Al observar las distintas opciones desarrolladas sobre el Modelado de Procesos ETL, es necesario identificar cuáles son las ventajas y desventajas de cada propuesta. Por otro lado, pueden existir elementos, que al momento de unir, completan el desarrollo de mejores propuestas para el modelado de Procesos ETL.

Para realizar la comparación entre las propuestas se procedió en primer lugar a identificar las principales características de cada una de ellas, de esta manera poder observar que elementos contienen comúnmente y cuales son propios de cada propuesta. En segundo lugar, al ser el ETL un Proceso con diversas actividades, se consideró como otro elemento la selección de actividades habituales en el desarrollo de un ETL, de esta manera observar si las propuestas las contemplan para un mejor desarrollo de Procesos.

Las Tablas 1 y 2 concuerdan en varias de las características y actividades analizadas por Muñoz et al. (2011). Sin embargo, no se consideró Actividades como característica debido a que todas las propuestas la cumplen. Además, se consideró incorporar como característica: Metamodelo, porque permitirá crear modelos de manera automática ajustándose a un estándar, y considera como característica la Granularidad de los elementos del modelo.

Se puede observar en la Tabla 1 cada una de las características en las respectivas propuestas:

Tabla 1: Características de las Propuestas asociadas a Procesos ETL

Característica	Propuestas				
	(a)	(b)	(c)	(d)	(e)
Metodología	X				
Notación Propia		X			
Paleta de Actividades		X	X		
Metamodelo		X			
Granularidad			X		X
UML	X		X		X
Ontología				X	

Otro punto a tomar en cuenta dentro de la comparativa, es con respecto a las principales actividades consideradas dentro de cada artículo, cada uno de ellos, en mayor o menor grado se puntualiza sobre actividades, las cuales son parte importante dentro de cualquier proceso ETL. Las actividades consideradas son: Unir (Join), Filtrar (Filter), Conversión (Conversion), Cargar (Load), Seleccionar (Select), Registro (Log), Agregación (Aggregation), Integrar (Merge), Concatenar (Concatenate), Sustituto (Surrogate).

A continuación, se detalla en la Tabla 2 las actividades consideradas en cada una de las propuestas:

Tabla 2: Principales Actividades de Procesos ETL

Actividades	Propuestas				
	(a)	(b)	(c)	(d)	(e)
Unir	X	X	X	X	
Filtrar	X	X	X	X	X
Conversión	X			X	
Cargar	X				
Seleccionar				X	
Registro	X	X			
Agregación	X	X	X	X	X
Integrar	X				
Concatenar				X	
Sustituto	X	X	X		

Un proceso ETL requiere del uso de muchas actividades para lograr llevar los datos desde las fuentes al Almacén de Datos, de esta manera de las 10 actividades seleccionadas se subdividió en 3 grupos principales las propuestas:

Contemplan 3 o menos actividades: la propuesta de Luján-Mora, Vassiliadis y Trujillo establece una forma de realizar el Modelado Conceptual enfocado en la manipulación de los datos. Las actividades contempladas en este modelo son Filtrar y Agregación, lo cual es muy poco pensando que son actividades habituales en un proceso ETL, por lo que no es lo suficientemente consistente para la construcción de un proceso ETL.

Contemplan entre 4 y 7 actividades: aquí son 3 las propuestas que abarcan este intervalo de actividades "Modelado Conceptual de Procesos ETL", "Un Enfoque Basado en UML para el Modelado de Procesos ETL en Almacenes de Datos" y "Diseño de Procesos ETL Usando Tecnologías de Web Semántica". En estas propuestas se logra abarcar un mayor número de actividades, lo que permite un mejor desarrollo de un proceso ETL, a pesar de no abarcar todas las actividades consideradas.

Contemplan más de 7 actividades: la única propuesta capaz de abarcar un mayor número de actividades fue "Una Metodología para el Modelado Conceptual de Procesos ETL", lo que permite la elaboración de más acabado, ya que se estiman a lo menos, las actividades más frecuentes dentro del proceso.

Por medio de la clasificación en relación al número de actividades consideradas, dentro de un grupo de mecanismos más habituales para desarrollar un proceso ETL, se observa que la propuesta "Una Metodología para el Modelado Conceptual de Procesos ETL" abarca el mayor número de actividades, por lo tanto se logra una mayor generalización para el desarrollo de un proceso ETL. Por otro lado, es importante ver qué cualidades son importantes y que apoyen al Modelado de un Proceso ETL, de esta manera se puede determinar qué aspectos son indispensables para el Modelado, los cuales son: Actividades, Uso de lenguaje, y Definición de metodología. Desde esta perspectiva el punto de actividades todas las propuestas contemplan las actividades, en mayor o menor grado, lo que permite un cierto nivel necesario para el Modelado de Proceso ETL.

Otra de las características consideradas como indispensables para el Modelado de procesos ETL es el uso de lenguaje que permite una mayor universalidad en su empleo, desde esta perspectiva tres propuestas utilizan un lenguaje como es UML, "Una Metodología para el Modelado Conceptual de Procesos ETL", "Un Enfoque Basado en UML para el Modelado de Procesos ETL en Almacenes de Datos" y "Diagramas de Mapeo de Datos para el Diseño de Almacenes de Datos con UML". UML es un lenguaje específico para el modelado y de amplio uso a nivel informático.

Finalmente la única propuesta que establece una clara metodología para la elaboración de un proceso ETL es "Una Metodología para el Modelado Conceptual de Procesos ETL". La metodología conceptual que plantean los autores establece cuatro pasos para la elaboración del proceso.

PROPUESTAS DE SEGURIDAD EN ALMACENES DE DATOS Y PROCESOS ETL

A nivel de investigación, existen diferentes propuestas que consideran la seguridad como un importante elemento dentro del desarrollo de Almacenes de Datos, pero son escasísimas aquellas que hablen sobre la seguridad en los procesos ETL, es por esto que se consideraron propuestas que hablen sobre aspectos de seguridad a nivel de almacenes de datos y también de procesos ETL.

En términos generales, se presentan siete propuestas, cuatro de ellas abarcan ampliamente el tema de seguridad en lo que se refiere a los Almacenes de Datos, sin embargo, solamente tres propuestas consideran el tema a nivel específico de procesos ETL.

Las propuestas son las siguientes:

Aplicación de QVT al Desarrollo de Almacenes de Datos: Un caso de Estudio (Soler et al., 2007).

Se presenta el empleo de estándar MDA (Model-Driven Architecture) en el modelado seguro de los Almacenes de Datos, permitiendo la obtención del esquema lógico a partir del modelo conceptual multidimensional. Además, se utilizó el lenguaje QVT (Query/View/Transformation) en el desarrollo de un almacén de datos seguro usando un caso de estudio.

Un Modelo de Diseño de Seguridad para Almacenes de Datos e Implementación Semiautomática con OLS10g (Villarroel et al., 2005)

Se establece un enfoque para el diseño de almacenes de datos seguros por medio de las siguientes actividades: a) la definición del enfoque en base a la Arquitectura Dirigida por Modelos (MDA) y a la Seguridad Dirigida por Modelos (MDS), b) el diseño de un Modelo de Control de Acceso y Auditoría (ACA) para el modelado multidimensional. A través del estándar MDA se obtiene la especificación de un Modelo Independiente de Plataforma (PIM), donde este PIM puede ser transformado en un Modelo Específico de Plataforma (PSM), para que finalmente el PSM pueda ser transformado a un código específico.

Construcción de un Esquema de Estrella Seguro en Almacenes de Datos mediante una Extensión del Paquete Relacional desde CWM (Soler et al., 2008)

Se busca el alineamiento del diseño de los almacenes de datos con la Arquitectura Dirigida por Modelos (MDA), de este modo se pretende obtener independencia e interoperabilidad en los sistemas diseñados gracias a la posibilidad de transformaciones entre diferentes modelos. Se presentan las siguientes extensiones en la propuesta: Metamodelo (en Análisis de Requerimientos), Metamodelo UML (en Modelo multidimensional seguro) y Paquete Relacional desde el metamodelo CWM (Modelo Relacional con elementos de seguridad).

Desarrollo de Almacenes de Datos Seguros con una Extensión de UML (Fernández-Medina et al., 2007)

Se realiza una extensión de UML para incorporar seguridad en el modelado conceptual de almacenes de datos. Los estereotipos, valores etiquetados y restricciones forman parte de la extensión de UML que permite un modelado de almacenes de datos seguros.

Se entregan los detalles necesarios para implementar la seguridad, por medio de la aproximación conceptual, en plataformas comerciales como Oracle 10g, permitiendo la aplicación de la seguridad de la información en las etapas del diseño en proyectos de la vida real.

Modelado de Transformación, Limpieza y Carga Seguras en Procesos ETL con UML 2.0 (Mrunalini et al., 2007)

Se propone un enfoque usando UML para llevar a cabo el modelado conceptual de la transformación, limpieza y carga segura en los procesos ETL. Incluyen la seguridad en el análisis de requerimientos, para ello utilizan casos de uso con seguridad para considerar los requerimientos. En la propuesta que incorpora seguridad se utilizan las características de UML 2.0 para modelar las características complejas, implementar políticas de seguridad y discutir la validez del modelo usando un caso de estudio de una aplicación bancaria. Sin embargo, a la propuesta le falta formalidad, debido a que el modelo es solamente explicado con la ayuda de un caso de estudio.

Modelo de Procesos ETL Seguro: Una Evaluación de la Seguridad en Diferentes Fases de ETL (Mrunalini et al., 2013)

Se propone un marco de trabajo para procesos ETL seguros, para ello automatizan y simulan el proceso de evaluación de la seguridad en el sistema ETL. Como resultado de esta propuesta es la estimación del mínimo número de requisitos de seguridad que deben ser considerados para garantizar la seguridad, lo cual ayuda a evitar implementación innecesaria. El marco de trabajo se compone de las siguientes etapas: Selección de la fase del proceso ETL con la cual se trabajará, Establecer políticas de seguridad para la fase seleccionada, Desarrollar una metodología de seguridad para alcanzar la política de seguridad definida, Desarrollar un modelo UML seguro para modelar los requerimientos, incluyendo los requisitos de seguridad, Validar con métricas de seguridad. Este marco de trabajo es muy interesante, sin embargo, la profundización de cada una de las etapas solamente la realiza mostrando su aplicación a un caso práctico.

Una Visión General de Implementación Física de un Flujo de Trabajo ETL Seguro (Anand, 2012)

Se propone una visión general respecto a la implementación de un flujo de trabajo ETL seguro. El flujo de trabajo se representa como un grafo dirigido cuyos nodos son actividades y conjunto de registros (recordsets). Se presentan los diferentes tipos de actividades ETL basadas en las interrelaciones de sus entradas y salidas. Posteriormente, muestra el procedimiento principal para un proceso de extracción de datos seguro. Esta propuesta es muy general, que no considera el modelado conceptual de procesos ETL, sino la implementación física de un flujo de trabajo ETL.

No se realiza un análisis comparativo porque las cuatro propuestas relacionadas a almacenes de datos son parecidas, son propuestas que han sido ampliadas o consideran aspectos más específicos (como por ejemplo, el uso de QVT, la generación del modelo lógico con aspectos de seguridad, etc.). En el caso de la seguridad en procesos ETL, solamente una de las tres propuestas considera el modelado conceptual, debido a la escasez de este tipo de propuestas.

CONCLUSIONES

Se puede señalar que a partir de la literatura existente queda demostrada la importancia del proceso ETL y la necesidad de que ésta sea planificada y modelada correctamente para facilitar esfuerzos de definición y redefinición de actividades. Cabe señalar que no existen muchas propuestas de modelado conceptual de procesos ETL y la mayoría de estas son extensiones de los mismos autores que presentaron propuestas previas.

Respecto a la seguridad de procesos ETL se puede indicar que la literatura es escasa, lo que nos indica que es un tema que se está recién trabajando aun cuando la gran mayoría de los autores coinciden que el proceso ETL es la etapa más crítica para la calidad de la información de los almacenes de datos..

Respecto a la incorporación de seguridad en procesos ETL a través del modelado conceptual no se han encontrado propuestas formales que traten directamente este tema, por lo tanto, como trabajo futuro, se pretende desarrollar un modelo conceptual para procesos ETL con su propia notación gráfica, tomando en cuenta los aspectos de seguridad.

REFERENCIAS

Anand, N., An Overview on Physical Implementation of Secure ETL Workflow. *Journal of Global Research in Computer Science*, 43-45, (2012).

Chung, L., Nixon, B., Yu, E. y Mylopoulos, J., *Non-functional requirements in software engineering*. Boston/Dordrecht/London, Kluwer Academic Publishers, (2000).

Blanco, C., García-Rodríguez, I., Fernandez-Merdina, E., Trujillo, J., y Piattini, M., Defining and transforming security rules in an MDA approach for DWs. *International Journal Business and Data Mining*. Vol 5, No. 2, 116-133, (2010).

El-Sappagh, S. H. A., Hendawi, A. M. A. y El Bastawissy, A. H., A proposed model for Data Warehouses ETL processes. *Journal of King Saud University*, 91-104, (2011).

Fernández-Medina, E., Trujillo, J., Villarroel, R. y Piattini, M., Developing Secure data warehouses with a UML extension, *Information Systems*, 32(6): 826-856, (2007).

Franco, D., Perea, J. y Puello, P., Metodología para la Detección de Vulnerabilidades en Redes de Datos. *Información Tecnológica*. Vol. 23, No. 3, 113-120, (2012).

Ghosh, A., Howell, C. y Whittaker, J., Building Software Securely from the Ground Up. *IEEE Software*. 19(1): 14-16, (2002).

Hall, A. y Chapman, R., Correctness by Construction: Developing a Commercial Secure System. *IEEE Software*. 19(1): 18-25, (2002).

Inmon, W. H., *Building the Data Warehouse*, Fourth Edition. Wiley Publishing, USA, (2005).

Kimball, R. y Caserta, J., *The data Warehouse ETL Toolkit*. Wiley Publishing, USA, (2004).

Kleppe, A., Warmer, J. y Bast, W., MDA explained. *The Model Driven Architecture: Practice and Promise*, Addison Wesley, (2003).

Luján-Mora, S., Vassiliadis, P. y Trujillo, J., Data Mapping Diagrams for Data Warehouse Design with UML. *23rd International Conference on Conceptual Modeling (ER)*, Shanghai, China, 191-204, (2004).

March, S. y Hevner, A., Integrated decision support systems: A data warehousing perspective, *Decision Support Systems*, 1031-1043, (2007).

- Moody, D., Theoretical and practical issues in evaluating the quality of conceptual models: current state and future directions. *Data & Knowledge Engineering*, Elsevier. Vol. 55, 243-276, (2005).
- Mrunalini, M., Geetha, D. E., Suresh, T.V. y Rajani, K., Modeling of Secure Transformation, Cleaning and Loading in ETL Processes using UML 2.0, *International Conference on Advances in Computer Vision and Information Technology*, Aurangabad, India. (2007).
- Mrunalini, M., Suresh, T.V. y Rajani, K., Secure ETL Process Model: An Assessment of Security in Different Phases of ETL. *International Journal of Software Engineering*. Vol. 6 No. 1, (2013).
- Muñoz, L., Mazón, J. N. y Trujillo, J., ETL Process Modeling Conceptual for Data Warehouses: A Systematic Mapping Study. *IEEE Latin America Transactions*, Vol. 9, No. 3, 360-365, (2011).
- Nelson, H. J., Poels, G., Genero, M., y Piattini, M., A conceptual modeling quality framework. *Software Quality Journal*. Vol 20, 201-228, (2012).
- Patil, P., Chavan, N., Rao, S. y Patil, S. V., Building of a Secure Data Warehouse by Enhancing the ETL Processes for Data Leakage. *IJCA Proceedings on International Conference and workshop on Emerging Trends in Technology (ICWET 2012)*, 18-23, (2012).
- Rodríguez, A., Fernández-Medina, E., Trujillo, J., y Piattini, M., Secure business process model specification through a UML 2.0 activity diagram profile. *Decision Support Systems*. Vol 51, 446-465, (2011).
- Simitsis, A. y Vassiliadis, P., A Methodology for the Conceptual Modeling of ETL Processes, 15th Conference on Advanced Information Systems Engineering (CAiSE '03), Klagenfurt/Velden, Austria, (2003).
- Simitsis, A., y Vassiliadis, P., A method for the mapping of conceptual designs to logical blueprints for ETL processes. *Decision Support Systems*. Vol 45, 22-40, (2008).
- Skoutas, D. y Simitsis, A., Designing ETL processes using semantic web technologies, 9th ACM International Workshop on Data Warehousing and OLAP (DOLAP '06), Arlington, Virginia, USA, 67-74, (2006).
- Soler, E., Trujillo, J., Fernández-Medina, E. y Piattini, M., Aplicación de QVT al Desarrollo de Almacenes de Datos Seguros: Un Caso de Estudio, X Conferencia Iberoamericana de Ingeniería de Software Cibse '07, Isla de Margarita, Venezuela, 209-222, (2007).
- Soler, E., Trujillo, J., Fernández-Medina, E. y Piattini, M., Building a Secure Star-Schema in Data Warehouses by an Extension of the Relational Package from CWM, *Computer Standards & Interfaces*. Vol 30, No. 6, 341-350, (2008).
- Trujillo, J. y Lujan-Mora, S., UML Based Approach for Modeling ETL Processes in Data Warehouses. *22nd International Conference on Conceptual Modeling*, Chicago, IL, USA, 307-320, (2003).
- Vassiliadis, P., Vagena, Z., Skiadopoulos, S., Karayannidis, N., y Sellis, T., ARKTOS: Towards the modeling, design, control and execution of ETL processes. *Information Systems*, 537-561, (2001).
- Vassiliadis, P., Simitsis, A. y Skiadopoulos, S., Conceptual Modeling for ETL Processes, 5th ACM International Workshop on Data Warehousing and OLAP (DOLAP '02), McLean, Virginia, USA, 14-21, (2002).
- Villarroel, R., Fernández-Medina, E., Trujillo, J. y Piattini, M., Un Modelo de Diseño de Seguridad para Almacenes de Datos e Implementación Semiautomática con OLS10g, VIII Iberoamerican Workshop on Requirements Engineering and Software Environments IDEAS 2005, Valparaíso, Chile, 243-254, (2005).
- Villarroel, R., Fernández-Medina, E., Trujillo, J. y Piattini, M., A UML 2.0/OCL Extension for Designing Secure Data Warehouses. *Journal of Research and Practice in Information Technology (JRPIT)*. 38(1): 31-44, (2006).