Técnicas de modelado de procesos de ETL: una revisión de alternativas y su aplicación en un proyecto de desarrollo de una solución de BI

ETL Processes modeling techniques: an alternatives review and its application in a BI solution development project

Alexander Bustamante Martínez[1], Ernesto Amaru Galvis Lista[2], Luis Carlos Gómez Flórez [3]
Ingeniería de Sistemas e Informática, Universidad Industrial de Santander, Bucaramanga, Colombia
alex.bustamante.martinez@gmail.com
egalvis@unimagdalena.edu.co
lcgomezf@uis.edu.co

***Resumen*—La tarea de un diseñador de procesos de ETL involucra: (1) analizar las fuentes de datos existentes para encontrar la semántica oculta en ellas y (2) diseñar el flujo de trabajo que extraiga los datos desde las fuentes, repare sus inconsistencias, los transforme en un formato deseado, y, finalmente, los inserte en la bodega de datos. Con el propósito de facilitar esta tarea, se han desarrollado diferentes técnicas, dos categorías que sobresalen son: (a) Las inspiradas en los diagramas de flujo y de procesos y (b) las inspiradas en el paradigma de programación orientada a objetos (POO) y los diagramas de UML. En el presente artículo se expone un par de alternativas halladas en la literatura y se ilustra la técnica utilizada en el proyecto "Desarrollo de una solución de inteligencia de negocios para apoyar a la toma de decisiones en el Proyecto Círculos de Aprendizaje", explicando el porqué de su elección y cómo se usó.**

***Palabras clave*— Bodegas de Datos, Inteligencia de Negocios, Zona de Preparación de Datos, Proceso de ETL.**

***Abstract*—The task of a designer ETL process involves: (1) analyzing existing data sources to find the semantics hidden in them and (2) design workflow that extracts data from sources, repair its inconsistencies, transforms it into a desired format, and finally inserted into the data warehouse. In order to facilitate this task, different techniques have been developed, two categories that stand out are: (a) inspired flow diagrams and process and (b) those based on the paradigm of object-oriented programming (OOP) and diagrams in UML. This article presents a couple of alternatives found in the literature and illustrates the technique used in the project "Development of a business intelligence solution to support decision making in the Learning Circles Project," explaining why of their choice and how it was used**

***Key Word* — Business Intelligence, Data Warehouse, Data Staging Area, ETL Process.**

I. INTRODUCCIÓN

El proceso de extracción, transformación y carga – ETL (Extraction, Transformation and Load) es una de las actividades técnicas más críticas en el desarrollo de soluciones de inteligencia de negocios – BI (Business Intelligence) [1][2]. Hace parte del componente de integración y, de su implementación adecuada dependen la integridad, uniformidad, consistencia y disponibilidad de los datos utilizados en el componente de análisis de una solución de BI. Su función es extraer, limpiar, transformar, resumir, y formatear los datos que se almacenarán en la bodega de datos de la solución de BI [3][4][5].

La construcción del ETL puede dividirse en tres subprocesos o componentes: componente de extracción, componente de transformación y componente de carga. En la Tabla 1 se presenta la descripción de cada uno de estos componentes identificando los elementos objetivo, las operaciones realizadas, y los resultados esperados.

[1] MSc(C) Ingeniería de Sistemas e Informática – UIS; Grupo de Investigación en Sistemas y tecnología de la Información – STI.

[2] PhD(C) Ingeniería de Sistemas – Universidad Nacional; Profesor Asistente, Universidad del Magdalena; Grupo de Investigación y Desarrollo en Organizaciones, Sistemas y Computación – GIDOSC.

[3] MSc Ingeniería de Sistemas – UIS; Profesor Asociado, Universidad Industrial de Santander; Grupo de Investigación en Sistemas y Tecnología de la Información – STI.

Fecha de Recepción: 23 de Enero de 2012
Fecha de Aceptación: 30 de Abril de 2013

Adicionalmente, el incremento en el uso de soluciones de inteligencia de negocios [6], implica un incremento en la construcción de procesos de ETL y una necesidad por garantizar su previsibilidad, extensibilidad, y adaptabilidad. Esto último, ha llevado al desarrollo de técnicas que permiten modelarlo de forma que esta especificación de sus características sea útil al momento de construirlo, probarlo y desplegarlo.

A continuación, se describen dos alternativas encontradas en la literatura, se ilustra el uso de una de ellas en un proyecto de aplicación, y finalmente se presentan las conclusiones obtenidas y las referencias

Componente	Elementos Objetivos (entrada)	Operaciones realizadas (proceso)	Resultado de la tarea (salida)
Extracción	Fuentes de datos, sistemas transaccionales, hojas de cálculo, archivos de texto.	Selección	Datos crudos(cargados en memoria)
Transformación	Datos crudos(cargados en memoria)	Limpieza, transformación, personalización, realización de cálculos y aplicación de funciones de agregación.	Datos formateados, estructurados y resumidos de acuerdo a las necesidades(aún en memoria)
Carga	Datos formateados, estructurados y resumidos de acuerdo a las necesidades(aún en memoria)	Inserción	Datos formateados, estructurados y resumidos con persistencia en el DW

Tabla 1. Descripción de los componentes del proceso de ETL.

II. TÉCNICAS PARA MODELAR EL PROCESO DE ETL

El modelado del proceso de ETL, al igual que el de cualquier objeto computacional, puede representarse utilizando tres niveles de abstracción: conceptual, lógico y físico. En la **¡Error! No se encuentra el origen de la referencia.**, se puede apreciar una breve descripción y comparación de los mismos.

De igual manera, y utilizando lo descrito en [7], las técnicas para modelar se pueden clasificar en cuatro categorías: modelado dimensional, extensiones del modelado E/R estándar, modelado basado en UML y modelado Sui-generis.

Complementando lo mencionado, existen dos grupos de técnicas que sobresalen: (a) Las inspiradas en los diagramas de flujo y de procesos y (b) las inspiradas en el paradigma de programación orientada a objetos (POO) y los diagramas de UML. En el proyecto [8] se vivió la disyuntiva de escoger la que se ajustara a los requerimientos establecidos.

A. ALTERNATIVA UNO: ENFOQUE ORIENTADO AL MODELADO DE PROCESOS

.

En esta categoría hay diferentes técnicas [9][10][1][11]. Para el proyecto se analizó [10], debido a que agrupa un conjunto de técnicas elaboradas en este enfoque. La investigación cubre dos tipos de modelado: el conceptual y el lógico. Para el modelado conceptual, plantea identificar las fuentes, los atributos y las transformaciones a realizar. Además, se proponen los iconos de la **¡Error! No se encuentra el origen de la referencia.** para realizar este tipo de modelado. En la Tabla 3, se indica el uso pensado para cada icono.

Tipos	Niveles de detalle	Dependiente de la plataforma	Objeto y conceptos	Ejemplo
Conceptual	Bajo	No	Fuentes, atributos, transformaciones y estructura de destino.	La información de los clientes, está en el sistema OLTP; debe calcularse la edad de los clientes previo a su inserción en el DW.
Lógico	Medio	No	Tablas de origen, dimensiones, tablas de hechos, Atributos, operaciones (aritméticas, lógicas y etc.)	La información del cliente debe calcularse así: extraer el año de nacimiento de la tabla ABC y retárselo al actual).Para luego insertarlo en el atributo edad de la dimensión cliente.
Físico	alto	Si	Tablas de origen, dimensiones, tablas de hechos Atributos, tipos de datos, precisión, restricciones, índices, entre muchas más.	La edad es representada por un entero de un byte, y para poder restársela al año actual, el valor extraído de ABC debe transformarse a tipo entero de un byte, luego restarlo al año que se extraerá de la fecha del sistema, posteriormente el valor se insertará en el atributo edad de la dimensión cliente, el cual está indexado.

Tabla 2. Tipos de modelado de ETL

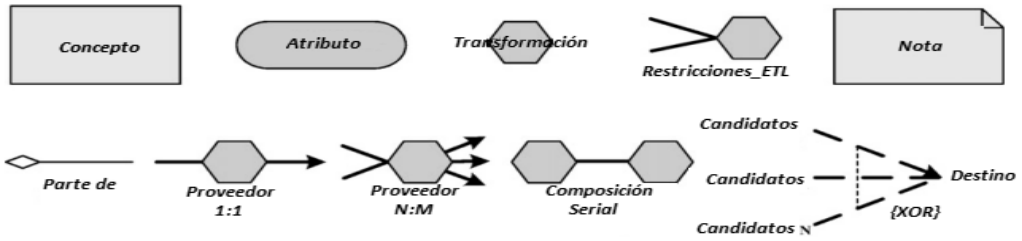


Figura 1. Elementos utilizados para el modelado conceptual**.**

Figura 2. Elementos utilizados para el modelado lógico.

Respecto del modelado lógico, y debido a que este se ocupa del flujo de datos necesario para garantizar que la bodega de datos sea poblada con éxito, se plantea considerar otros elementos como el conjunto de valores admitidos por un atributo, los parámetros necesarios para que las funciones ejecuten su trabajo, entre otros. Para hacer esto posible, se proporciona otro conjunto de elementos (ver **¡Error! No se encuentra el origen de la referencia.**).

Adicionalmente, en esta propuesta también se expone un método, para mapear el diseño conceptual al lógico, ilustrando como cada elemento del modelado conceptual puede representarse en elementos lógicos, minimizando la complejidad y la perdida de información.

Elemento	Descripción
Concepto	Representa una entidad en la base de datos de origen, o en la bodega de datos.
Atributo	Es el nodo más granular de información.
Transformación	Son abstracciones que representan partes, rutinas completas ejecutan dos tareas esencialmente: (a) filtrar, y (b) transformar datos
Restricción ETL	Es utilizada para indicar que los datos deben cumplir un conjunto de requerimientos.
Nota	Son usadas para capturas comentarios extras que los diseñadores quieren realizar durante la fase de diseño. Además de explicar la semántica de las funciones.
Relación-Parte	Un concepto está compuesto por un conjunto de atributos
Proveedor 1:1/N:M	Mapea un conjunto de atributos de entrada, a un conjunto de atributo de salida a través de una transformación relevante.
Composición Serial	Se usa cuando necesitamos combinar varias relaciones en un único proveedor.
Candidatos	Captura el concepto de que un Concepto de la bodega de datos, puede ser poblado por más que un Concepto de la fuente de datos

Tabla 3. Descripción de los elementos utilizados en el modelado conceptual

B. ALTERNATIVA DOS: ENFOQUE ORIENTADO A OBJETOS

En este enfoque se analiza la alternativa presentada en [7], en donde, se plantea usar el diagrama de clases de UML. El autor expone que con la adopción del diagrama de clases de UML, no solo se logra cubrir la especificación de conceptos (clases), atributos (atributos) y funciones (métodos), sino que al ser UML reconocido como una buena práctica para el desarrollo de software, logra acortar la curva de aprendizaje necesaria en la asimilación de una nueva tecnología. Esta técnica hace uso de estereotipos (ver Tabla 4) para cualificar el comportamiento general de la clase.

.

Estereotipo	Descripción	Icono
Agregación.	Realizar agregaciones, con los datos, basados en algún criterio.	
Conversión	Cambiar sea: el tipo de dato, el formato u obtener un nuevo dato, derivado de otro existente.	
Filtro	Filtrar los dados por algún criterio.	
Incorrecto	Marcar y re-direccionar datos incorrectos para la operación.	
Join	Unir dos conjuntos de datos, tomando como referencia algún(os) atributos.	
Cargador	Insertar los datos en la estructura de datos de destino.	
Log	Registrar información sobre el proceso.	

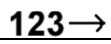
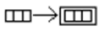
Estereotipo	Descripción	Icono
Combinación	Integrar dos o más conjuntos de datos, estos deben tener atributos compatibles.	
Delegados	Genera llaves delegadas únicas,	
Wrapper	Transforma una fuente de datos, en un conjunto de datos en memoria, similar a la fuente.	
Tabla	Representa una tabla de la base de datos, se de origen o de destino.	
Nota	Permite realizar anotaciones que ayudan a clarificar las operaciones realizadas.	

Tabla 4. Descripción de los estereotipos.

III. ELECCIóN DE LA TÉCNICA

Para la elección de la técnica de modelado se determinó que ésta debería: 1) tener una representación visual del proceso; 2) facilitar la documentación, y 3) ser comprendida, tanto por el diseñador, como por el programador. Igualmente, se consideraron factores como el tiempo de aprendizaje, pues el desarrollo solicitado debía ser rápido y no se podía adicionar más personal o cambiar la tecnología a utilizar. Cabe destacar que los miembros del equipo (un diseñador, un analista, un desarrollador y un Ingeniero de Software) tenían experiencia en el uso de UML como lenguaje para elaborar modelos y especificar procesos, lo cual facilitó el proceso.

Luego de evaluar las alternativas, utilizando criterios como el tiempo que demandaría al equipo de desarrollo su aprendizaje, la facilidad para describir las operaciones, y demás características descritas en la **¡Error! No se encuentra el origen de la referencia.**. El equipo se inclinó por el enfoque de modelado orientado a objetos.

Criterios	Modelado Orientado a Objetos	Modelado de Procesos
Tiempo de aprendizaje requerido	-	+
Facilidad para describir las operaciones	+	-
Encapsulamiento	+	-
Identificación de recursos empleados	+	-
Legibilidad	+	-
Facilidad de documentación	+	-

Tabla 5. Comparación entre las técnicas.

IV. APLICACIÓN DE LA TÉCNICA DE MODELADO EN UN PROYECTO DE DESARRROLLO DE UNA SOLUCIÓN DE BI

La aplicación de la técnica seleccionada en un proyecto de desarrolló tomó como contexto la ejecución de proyecto educativo y social cuyo propósito es garantizar que la población infantil en condición de vulnerabilidad y desplazamiento tenga acceso a la educación básica. Este proyecto, auspiciado por el Ministerio de Educación Nacional de Colombia – MEN, implementa modelos educativos flexibles como el denominado "Círculos de Aprendizaje" – CA [12][13][14]. El objetivo principal del modelo CA es lograr la vinculación al sistema de educación formal de niños, niñas y adolescentes NNA con edades entre 5 y 16 años, que han sido afectados por problemas de desplazamiento forzado y vulnerabilidad.

En el caso específico de la Universidad del Magdalena el equipo del proyecto CA – PCA se encuentra conformado por más de 200 personas distribuidas en todos los departamentos del caribe colombiano. La dinámica del PCA requiere que para cada uno de los NNA pertenecientes a los CA se registre información de tipo personal, familiar, académica y psicosocial. Esta información se procesa con el fin de generar reportes orientados al seguimiento y desempeño de los NNA, así como el cumplimiento de los objetivos del proyecto [12]. En este sentido, la ejecución del PCA presenta dos retos concernientes a la administración de la información: 1) la gestión del proyecto y el procesamiento de las transacciones en línea, considerando que el proyecto se opera en siete departamentos de la costa Caribe; y 2) la generación de consultas y reportes, a partir de la información registrada, que permitan apoyar la toma de decisiones. Tales retos motivaron el desarrollo de una solución de BI que permitiera apoyar el proceso de toma de decisiones y de gestión del PCA operado por la Universidad del Magdalena. En el desarrollo de la solución de BI, fue necesario modelar el proceso de extracción, transformación y carga de la solución, para tener una representación visual del proceso, llevar una documentación de los modelos diseñados y evitar los retrasos en la ejecución del proyecto.

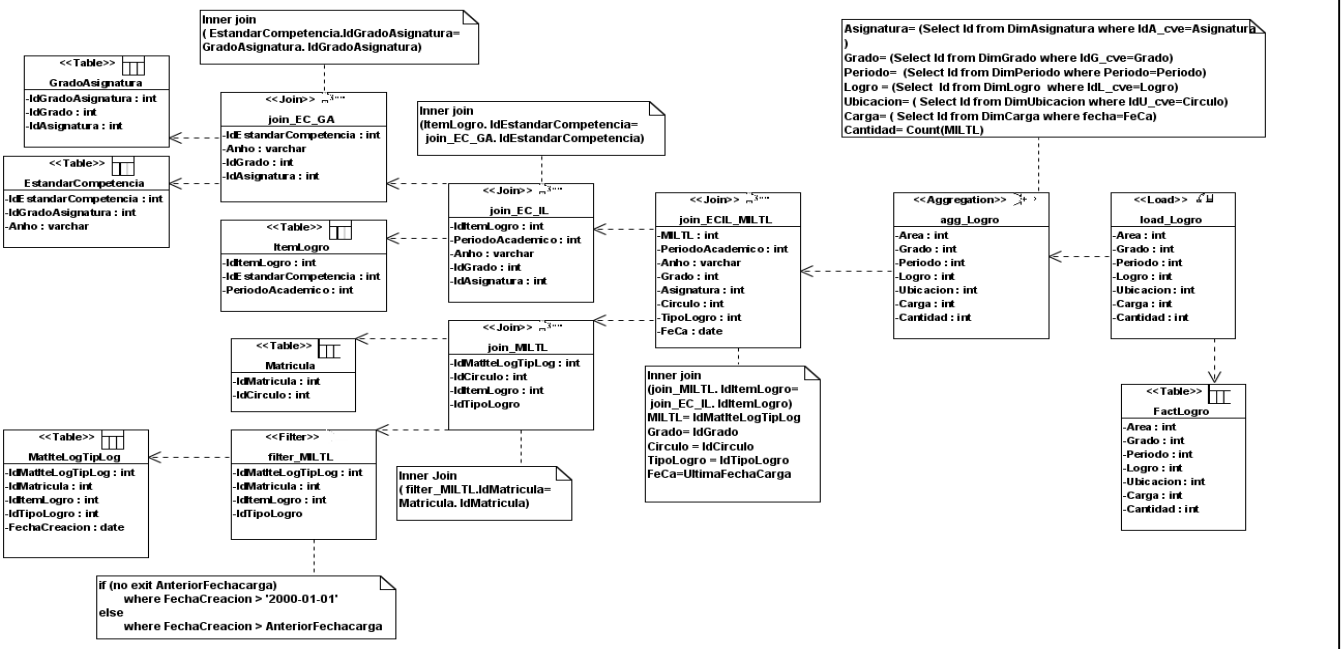


Figura 3. Diseño del proceso de ETL, para la tabla de hechos D2R

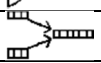
.

El modelado total del proceso de ETL se cubrió con el diseño de 44 diagramas que especificaban el flujo de datos desde la fuente, un sistema OLTP denominado INTRANET, hasta la bodega de datos compuesta por 55 tablas. En la Tabla 6 se presenta el detalle de la composición de la bodega de datos.

Tipo	Cantidad	Descripción
Hechos	12	contiene las medidas a analizar, como deserción, inasistencia y rotación
Dimensiones	43	Almacenan las variables que se utilizan para analizar las medidas (Sexo, Edad, Grupo Poblacional, etc.)

Tabla 6. Descripción de tablas de la bodega de datos.

Al aplicar la técnica de modelado seleccionada, el uso de los estereotipos facilitó la comprensión del funcionamiento de cada elemento dentro de la tarea principal, cargar datos limpios y formateados y resumidos en la bodega de datos. La Tabla 7, muestra la distribución y frecuencia de uso por cada estereotipo en el caso de un componente de seguimiento psicosocial.

Mecanismo ETL (Estereotipo)	Icono	Cantidad
Agregación.		9
Conversión		1
Filtro		8
Join		20

Mecanismo ETL (Estereotipo)	Icono	Cantidad
Cargador		48
Tabla		88
Nota		37

Tabla 7. Frecuencia pos estereotipo en el componente carpeta integral.

La representación visual del proceso de ETL se alcanzó con la elaboración de los diagramas que se presentan en Figura 3. Además, con la implementación de los diagramas diseñados usando en la herramienta Microsoft Sql Server Integratión Services(SSIS), se cumplió con el criterio de facilitar la traducción de los diseños en rutinas por parte de los programadores, (ver Figura 4). Igualmente, el proceso de documentación se facilitó pues, una vez creados los diagramas de clases que modelan el proceso, se documentaban las clases que intervenían, y esta documentación se convirtió en la documentación del proceso. Por último, es importante destacar que el proceso de diseñar, implementar y documentar, fue realizado iterativamente.

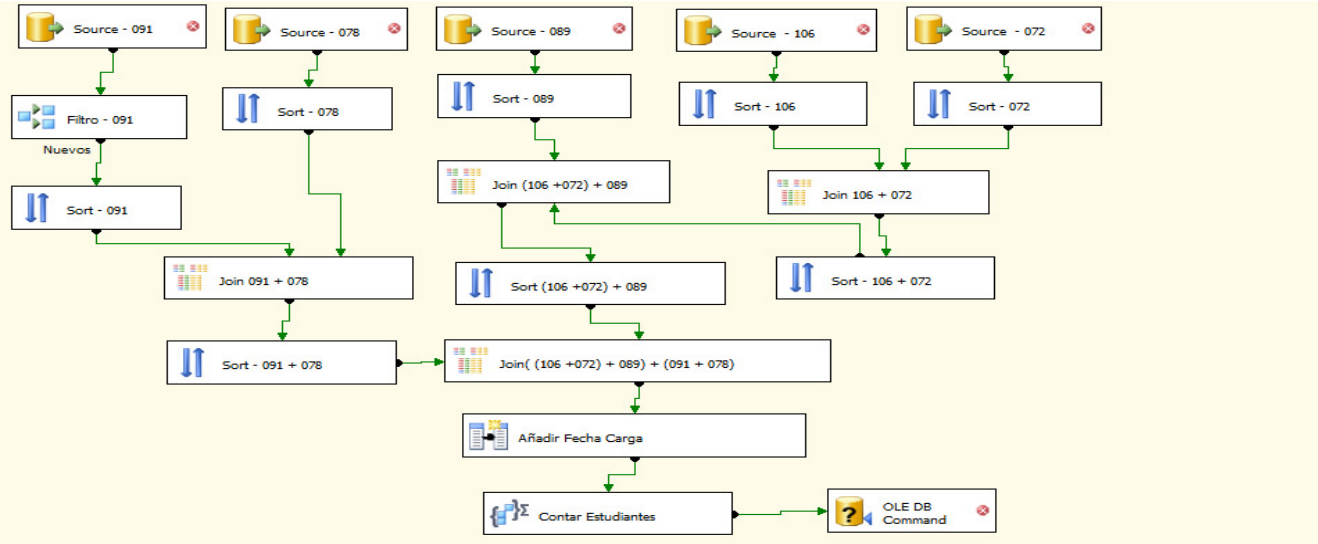


Figura 4. Diseño del proceso de ETL, para la tabla de hechos D2R.

V. CONCLUSIONES

Independiente de la pieza de software que se construya, el diseño es una actividad esencial, si se quiere reducir la incertidumbre inherente a la implementación y propiciar su comprensión, extensibilidad y mantenibilidad. Por lo cual el diseño de los procesos de ETL debe ser una tarea indispensable en el proceso de desarrollo de una solución de Inteligencia de negocios, si se quiere que ésta tenga las propiedades mencionadas.

Pesé a la existencia de técnicas para realizar el modelado del proceso de ETL, no existe una que haya sido considerada como una buena práctica, y menos un estándar. Muestra de ello, es que aunque la alternativa seleccionada en el proyecto se adaptó a las necesidades, está lejos de ser una técnica adecuada para modelar el flujo de datos presente en los procesos de ETL.

Los procedimientos de ETL por su naturaleza, son procesos con entrada, operaciones, salidas y que consumen recursos. Realizar su diseño utilizando algún lenguaje para el modelado de procesos de negocios como BPEL o una notación como BPMN puede considerarse adecuado. Pero, estos lenguajes en su forma básica se abstraen de la descripción y el comportamiento específico de este proceso, por lo que cual una notación útil combinaría esta forma de hacer modelado con características del diagrama de clase de UML.

Aplicar diferentes enfoques para el modelado de procesos de ETL en proyectos a nivel de pregrado, permitirá conocer más acerca de la aceptación y utilidad de las técnicas existentes, para así conocer la ventaja de cada enfoque, y poder emitir una tesis acerca de este tema.

Desarrollar herramientas CASE que faciliten el uso de las técnicas. Lo anterior, con el objeto de crear un ambiente propicio para la realización de la tarea de diseño dentro de proyectos de Inteligencia de Negocios. La herramienta debe permitir diseñar los procesos de ETL, generar el código y documentar el proceso desde un mismo entorno. Además de la integración con herramientas de implementación de procesos de ETL, como SSIS.

REFERENCIAS

[1] A. Simitsis y P. Vassiliadis, «A method for the mapping of conceptual designs to logical blueprints for ETL processes», Decision Support Systems, vol. 45, nº. 1, págs. 22-40, Abr. 2008.

[2] C. Howson, Successful Business Intelligence: Secrets to Making BI a Killer App, 1st ed. McGraw-Hill Osborne Media, 2007.

[3] R. Kimball y J. Caserta, The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleanin, 1st ed. Wiley, 2004.

[4] L. T. Moss y S. Atre, Business Intelligence Roadmap: The Complete Project Lifecycle for Decision-Support Applications. Addison-Wesley Professional, 2003.

[5] E. Turban, R. Sharda, D. Delen, y D. King, Business

Intelligence (2nd Edition), 2nd ed. Prentice Hall, 2010.

[6] «Gartner Forecasts Global Business Intelligence Market to Grow 9.7 Percent in 2011». [Online]. Available: http://www.gartner.com/it/page.jsp?id=1553215. [Accessed: 06-Jun-2011].

[7] S. Luján-Mora y J. Trujillo, «A Data Warehouse Engineering Process», in Advances in Information Systems, vol. 3261, T. Yakhno, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, págs. 14-23.

[8] L. Benavides, A. Bustamante, y A. García, «Desarrollo de una solución de Inteligencia De Negocios para apoyar a la toma de decisiones en el Proyecto Círculos De Aprendizaje», Universidad del Magdalena, 2009.

[9] Z. El Akkaoui y E. Zimanyi, «Defining ETL worfklows using BPMN and BPEL», in Proceeding of the ACM twelfth international workshop on Data warehousing and OLAP, New York, NY, USA, 2009, pág. 41–48.

[10] P. Vassiliadis, A. Simitsis, P. Georgantas, M. Terrovitis, y S. Skiadopoulos, «A generic and customizable framework for the design of ETL scenarios», Information Systems, vol. 30, nº. 7, págs. 492-525, Nov. 2005.

[11] P. Vassiliadis, A. Simitsis, y S. Skiadopoulos, «Conceptual modeling for ETL processes», in Proceedings of the 5th ACM international workshop on Data Warehousing and OLAP, New York, NY, USA, 2002, pág. 14–21.

[12] «Alcance del Proyecto - Circulos de Aprendizaje Unimagdalena». [Online]. Available: http://vicextension.unimagdalena.edu.co/Circulos/Portal/proyecto/alcance.aspx. [Accessed: 23-Ago-2011].

[13] Ministerio de Educación Nacional, «Lineamientos de política para la atención educativa a la población afectada porla violencia». .

[14] Ministerio de Educación Nacional, «PLAN NACIONAL DE DESARROLLO EDUCATIVOINFORME DE GESTIÓN JUNIO 2008 A NOVIEMBRE DE 2009». .