Definición y validación de medidas para procesos ETL en almacenes de datos

Lilia Muñoz[1], Jesús Pardillo[2], José-Norberto Mazón[2], y Juan Trujillo[2]

[1] Grupo de Investigación Lucentia. Departamento de Sistemas de Información, Control y Evaluación de Recursos Informáticos, Universidad Tecnológica de Panamá
Panamá
lilia.munoz@utp.ac.pa

[2] Grupo de Investigación Lucentia, Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante
España
jesuspv,jnmazon,jtrujillo@dlsi.ua.es

Resumen. In data warehousing, ETL (Extract, Transform, and Load) processes are in charge of extracting the data from data sources that will be contained in the data warehouse. Due to their relevance, the quality of these processes should be formally assessed from early stages of development, in order to avoid making bad decisions as a result of incorrect data. In this paper, a set of measures is presented to evaluate the structural complexity of ETL process models at conceptual level. Moreover, this study is accompanied by one controlled experiment whose aim is the empirical validation of the proposed measures. The use of these measures can aid designers to predict the effort associated with the maintenance tasks of ETL processes. This proposal is based on UML (*Unified Modeling Language*) activity diagrams for modeling ETL processes, and on the FMESP (*Framework for the Modeling and Evaluation of Software Processes*) framework for the validation of the measures.

Palabras clave: Procesos ETL, validación, medidas, almacenes de datos

1 Introducción

En los años '90s, Inmon [8] definió el término Almacén de Datos (AD) como: *una colección de datos orientados por temas, integrados, variables en el tiempo y no volátiles para el apoyo de la toma de decisiones*. Un AD es "integrado", porque los datos que se introducen en el almacén se obtienen de una variedad de fuentes de datos (sistemas heredados, bases de datos relacionales, ficheros COBOL, etc.). Para lograr la integración de esa variedad de fuentes, se utilizan los procesos ETL. Dichos procesos son los responsables de la extracción de los datos a partir de las diversas fuentes de datos heterogéneas, de la transformación de estos (conversión, limpieza, etc.), y de su carga en el AD. Se reconoce ampliamente que el diseño y mantenimiento de los procesos ETL son factores claves en el éxito de proyectos de AD [10, 26].

Por su parte, un proceso ETL es extremadamente complejo, propenso a errores y consume mucho tiempo [24]. Se ha argumentado ampliamente, en la literatura, que los procesos ETL son costosos y que son una de las partes más importantes del desarrollo de un AD [8, 29]. En [23], se reporta que los costos de herramientas ETL se estiman en al menos la tercera parte de los gastos del presupuesto de un AD. Por estas razones, así como por el alto coste de adquisición y mantenimiento, muchas organizaciones prefieren desarrollar sus propios procesos ETL.

En los últimos años se han definido varias propuestas para el modelado conceptual y lógico de procesos ETL para AD [13, 25, 27, 29, 28, 30]. Dada la gran complejidad de los procesos ETL, estas propuestas tratan sobre el diseño, mantenimiento y gestión de los procesos ETL. Por otro lado, un mal diseño de los procesos ETL, pueden suponer errores graves en la demora de la carga en el AD. Por su parte, los esfuerzos en obtener la calidad de los ADs se han orientado a la calidad del producto final, no así a la calidad del modelado de procesos ETL que es parte fundamental del mismo. En la literatura más relevante sobre este tema se han encontrado algunas aproximaciones [1, 21, 30, 31] donde se han presentado medidas de calidad para procesos ETL. Sin embargo, en ninguna de estas aproximaciones se han validado formal o empíricamente las medidas propuestas.

Con el fin de cubrir esta carencia, en este artículo se describe un conjunto de medidas para evaluar la mantenibilidad del modelado conceptual de procesos ETL, partiendo de la hipótesis de que la baja mantenibilidad (facilidad de mantenimiento) de un modelo de procesos ETL influye negativamente en su calidad, y por lo tanto, puede tener repercusión en el desarrollo de AD (más costosos en recursos y tiempo). Las medidas son aplicables a una aproximación de modelado conceptual de procesos ETL basada en diagramas de actividad de UML [13]. Mediante la utilización de diagramas de actividad UML conseguimos una mejor representación de los procesos ETL, permitiendo representar los aspectos dinámicos dichos procesos. Por su parte, la propuesta para modelado conceptual de procesos ETL presentada en [13], se enmarca dentro de una propuesta global para acometer el desarrollo de AD con MDA [15], en este marco de trabajo se han desarrollado las diferentes capas de la arquitectura de un AD con MDA [11, 12, 17], lo que ha permitido el desarrollo del AD de manera sistemática y automatizada.

Para la definición de las medidas se utilizó el marco FMESP (*Framework for the Modeling and Evaluation of Software Processes*) [7]. FMESP consiste en un marco de modelado y medición del proceso del software, en el que los modelos se representan en base a SPEM (*Software Process Engineering Metamodel Specification*) [14]. Sin embargo, dada su generalidad y flexibildad lo hemos adaptado para la evaluación de modelos de procesos ETL a nivel conceptual. El resto del artículo se estructura como sigue: En la sección 2 se describe el trabajo relacionado. Las medidas propuestas definidas para modelos de procesos a nivel conceptual se presentan en la sección 3 y a continuación en la sección 4 se realiza la validación teórica, en la sección 5 se presenta la validación empírica de las medidas propuestas. Finalmente, en la sección 6 se aportan las conclusiones y trabajo futuro.

2 Trabajos relacionados

La calidad de un AD vendrá influenciada por la calidad del SGBD, por la calidad de los modelos de datos y por la calidad de los procesos ETL [20]. El trabajo relacionado lo enmarcamos en estos dos últimos aspectos.

Calidad de modelos de datos. Con respecto a las medidas de calidad para modelos de datos podemos hacer referencia a los trabajos presentados en [3, 16, 18–20]. Estas propuestas son buenas aproximaciones a las medidas de ADs; sin embargo, no son completas, ya que no son parte de un modelo de calidad que permita a los diseñadores su uso de una forma sistemática y objetiva. Por otro lado, podemos encontrar la propuesta de un modelo de calidad de un AD, llamado DWQ (*Data Warehouse Quality*) [9] que intenta asegurar la calidad de los datos almacenados para mejorar el uso del AD; esta aproximación evalúa algunas dimensiones de calidad, centrándose en la calidad de los datos. No obstante, la arquitectura de un AD es compleja, por lo que consideramos que la calidad de los ADs debe abarcar todas las capas que componen el almacén, y en ninguna de estas propuestas están considerados los aspectos de calidad de procesos ETL.

Calidad de procesos ETL. En la literatura consultada existen algunas aproximaciones sobre

medidas de calidad para procesos ETL. En [30], los autores modelan un escenario ETL como un grafo e introducen la importancia de las medidas en los nodos del grafo. Los mismos autores definen, en [31] una colección de medidas, usadas para evaluar los flujos de trabajo de los procesos ETL. Por su parte, en el campo de la integración de datos para procesos ETL, las medidas se pueden ubicar en las siguientes categorías [21]: tipos de datos, conformidad del domino de datos, características estadísticas de un conjunto de datos (valor máximo, valor mínimo, etc.) y relaciones referenciales. En [1], se presenta una propuesta para verificar la consistencia de los datos que son cargados en el AD. Se calcula la *Entropía de Shannon* [22] en particiones producidas por un conjunto de atributos, demostrando que estos valores pueden ser utilizados como señales de problemas en el proceso de extracción de los datos. Sin embargo, en ninguna de estas aproximaciones se han validado formal y empíricamente las medidas propuestas, de tal manera que podamos contar con parámetros objetivos para elegir entre ciertos modelos de procesos ETL. Por lo tanto, en este artículo proponemos y validamos teórica y empíricamente medidas para evaluar la mantenibilidad de procesos ETL a nivel conceptual.

3 Medición de la calidad de procesos ETL

Antes de definir las medidas se debe tener claro el objetivo que se quiere conseguir de una forma precisa. Las propiedades estructurales de un modelo, como la complejidad estructural, influyen sobre la complejidad cognitiva [4], entendible como la carga mental que se produce en las personas que tienen que tratar con el modelo. Una elevada complejidad cognitiva puede generar diferentes problemas tales como baja mantenibilidad y alta propensión a errores [4]. Para la definición del objetivo de esta propuesta se utilizó la plantilla GQM (*Goal Question Metric*) [2]:

- *Analizar* medidas relacionadas con la complejidad estructural de modelos de procesos ETL en almacenes de datos
- *Con el propósito de* evaluarlas
- *Con respecto* a su capacidad de ser usadas como indicadores de la mantenibilidad de dichos modelos,
- *Desde el punto de vista* de los diseñadores de procesos ETL
- *En el contexto de* estudiantes de Licenciatura en Desarrollo de Software.

Considerando la importancia de los procesos ETL mencionada anteriormente y en que un mal diseño de éstos puede suponer graves errores en la demora de la carga en el AD, nuestro principal interés es evaluar la mantenibilidad de los modelos conceptuales de procesos ETL representados con diagramas de actividad de UML, mediante medidas, nuestra aproximación adapta la propuesta de FMESP al modelado de procesos ETL presentado en [13], obteniendo con ello dos niveles de abstracción para la definición de las medidas: a nivel del modelado de procesos ETL y a nivel de actividad de procesos ETL. En este caso, las medidas propuestas son a nivel de proceso y serán validadas de manera teórica y empírica.

3.1 Medidas de procesos ETL modelados a nivel conceptual

Las medidas a nivel de modelo de procesos ETL han sido definidas con el objetivo de evaluar la complejidad estructural de procesos ETL a nivel conceptual. Para ello, el método de medición utilizado ha sido el conteo del número de elementos más significativos del modelo de procesos ETL (actividades, acciones, etc.) y el número de relaciones significativas entre los elementos (flujos de entrada y/o salida). Los términos utilizados en este artículo en cuanto a la medición de modelos de procesos de ETL, se basa en la *Ontología de la Medición del Software definida* por García et al. [6]. De este modo, el conjunto de medidas definidas han sido agrupadas en dos categorías:

– *Medidas base*, que consisten principalmente en contar elementos significativos del modelo de procesos ETL y de las cuales se ha definido un total de 11 medidas base en función de los elementos que componen el modelado de procesos ETL propuesto en [13].
– *Medidas derivadas*, definidas a partir de las medidas base, permiten conocer los principales ratios existentes entre los diferentes elementos del modelo. Este grupo está compuesto por 4 medidas.

Las medidas y definiciones se muestran en la Tabla 1, las primeras 11 medidas son medidas base y las 4 restantes son medidas derivadas. Las medidas derivadas son obtenidas a partir de las medidas base y representan los ratios existentes entre los diferentes elementos del modelo de procesos ETL.

Tabla 1. Medidas para modelos de procesos ETL a nivel conceptual

Medida	Definición
NAP	Número de actividades en un proceso ETL
NEE	Número elementos de entradas en el modelo de procesos ETL
NES	Número elementos de salidas en el modelo de procesos ETL
NFES	Número de flujos entrantes y salientes en el modelo de procesos ETL
NET	Número de eventos de tiempo en el modelo de procesos ETL
NEM	Número de elementos Merge en el modelo de procesos ETL
NEF	Número de elementos Fork en el modelo de procesos ETL
NEJ	Número de elementos Join en el modelo de procesos ETL
NOSEF	Número de objetos de salida de un elemento Fork en el modelo de procesos ETL
NOEEJ	Número de objetos de entrada de un elemento Join en el modelo de procesos ETL
NODS	Número de objetos DataStore en el modelo de procesos ETL
NTEES	Número total de elementos de entrada y salida en el modelo de procesos ETL $NTEES = NEE + NES$
RFESA	Ratio de flujos entrantes y salientes entre cada actividad en el modelo de procesos ETL $RFESA = \frac{NFES}{NAP}$
RDInEE	Ratio de dependencias de elementos de entrada en el modelo de proceso ETL $RDInEE = \frac{NEE}{NTEES}$
RDOutES	Ratio de dependencias de elementos de salida en el modelo de proceso ETL $RDOutES = \frac{NES}{NTEES}$

4 Validación teórica

En la actualidad diferentes acercamientos han sido propuestos para validar teóricamente las medidas software. Este tipo de validación es importante porque es un prerequisito para demostrar la utilidad de una medida, propósito principal de una validación empírica. Estas aproximaciones se han orientado básicamente hacia dos tendencias: aproximaciones axiomáticas y aproximaciones basadas en la teoría de la medición, las cuales son ampliamente reconocidas por la comunidad científica. Para este trabajo se utiliza la aproximación axiomática. El marco seleccionado han sido el de Briand et al. [4] como marco basado en aproximaciones axiomáticas.

4.1 Marco formal de Briand et al. [4]

Briand et al. [4] destacan la necesidad de definir los conceptos más relevantes utilizados en la medición de productos software de manera no ambigua. Y una manera es definir precisamente que propiedades matemáticas caracterizan estos conceptos. Briand et al. [4], proponen

un marco matemático genérico, el cual no se instancia en algún artefacto de software en particular, y es riguroso, porque está basado sobre conceptos matemáticos precisos.

En este marco formal se clasifican las medidas a partir de varios conceptos de medición importantes como son: *tamaño, complejidad, longitud, cohesión y acoplamiento*. A la hora de realizar la definición de sistema, módulo, elemento, y relación de acuerdo a este marco es importante considerar que las medidas se definen sobre modelos de procesos ETL compuestos por actividades, elementos y flujos de objetos. En función de las características de cada medida y de los elementos del modelo de procesos ETL sobre los que se aplican, se definirá para su validación de acuerdo a este marco lo que se entiende por sistema, módulo, elemento y relación. Dada la limitación de espacio, el proceso se ilustra con la medida **NFES**.

Validación de la medida NFES

Consideramos que un *modelo de procesos ETL* (sistema) está compuesto por un conjunto de *actividades* (módulos) y un conjunto de elementos que corresponden a los *flujos de objetos* (elementos). Según Briand et al. [4], la complejidad de un sistema se caracteriza por las siguientes propiedades:

- **No Negatividad (Propiedad 1)**. La complejidad de un modelo de procesos ETL es no negativa. Por la definición dada es imposible que la medida **NFES** tome valores negativos.
- **Valor Nulo (Propiedad 2)**. La complejidad de un modelo de procesos ETL es nula si el conjunto de los flujos de objetos es el conjunto vacío. Si no hay flujos de objetos, **NFES** será igual a cero.
- **Simetría (Propiedad 3)**. La complejidad de un modelo de proceso ETL no debe depender de la convención escogida para representar los flujos de objetos entre sus elementos. Por la propia definición de **NFES**, la dirección de los flujos de objetos no afecta a su valor.
- **Monotonicidad de Módulos (Propiedad 4)**. La complejidad de un modelo de procesos ETL no es menor que la suma de las complejidades de cualesquiera de dos de sus actividades sin flujos en común. El hecho de unir dos actividades, sin flujos comunes, no hará que el valor de **NFES** disminuya.
- **Aditividad de Módulos disjuntos (Propiedad 5)**. La complejidad de un modelo de procesos ETL compuesto por dos actividades disjuntas es igual a la suma de las complejidades de los dos actividades. El número de flujos, entrantes y salientes, entre cada par de actividades de un proceso ETL, será la suma del número de flujos entre cada par de actividades (la suma de las complejidades de sus actividades). Por tanto, **NFES** es una medida de *complejidad*.

Para las demás medidas se ha seguido el mismo procedimiento de validación. Por restricciones de espacio, en la Tabla 2, se presenta un resumen de la validación de las otras medidas propuestas para el modelo de procesos ETL. Además, podemos decir que las medidas derivadas **NTEES, RFESA, RDInEE, RDOutES** son medidas válidas al ser definidas a partir de una función de cálculo sobre medidas base válidas.

5 Validación empírica

El experimento que se ha llevado a cabo con el objetivo de establecer qué medidas son útiles para evaluar la entendibilidad y mantenibilidad de los modelos de procesos ETL.

5.1 Selección del contexto

Para alcanzar resultados más generales el experimento debería ser realizado con proyectos reales y con profesionales experimentados, sin embargo esto es muy costoso en tiempo,

Tabla 2. Medidas definidas a nivel de modelo de proceso ETL y propiedades que satisfacen en el marco de Briand et al. [4]

Propiedades	**NAP, NEF, NEM,NET, NOSEF, NEJ,NODS, NOEEJ**	**NEE, NES**	**NFES**
No negatividad	✓	✓	✓
Valor nulo	✓	✓	✓
Aditividad de módulo	✓		
Monotonicidad no creciente para componentes no conectados			
Módulo disjuntos			
Aditividad de módulos disjuntos		✓	✓
Monotonicidad		✓	
Fusión de módulos		✓	
Monotonicidad de módulo			✓
Simetría			✓
Módulos cohesivos			
No Negatividad y Normalización			
	Tamaño	**Acomplamiento**	**Complejidad**

dinero y esfuerzo; por lo tanto se torna difícil de llevar a cabo. Para reducir costos se ejecutan proyectos reales con estudiantes avanzados y egresados recientemente, en un entorno experimental controlado [32].

5.2 Material

El material experimental han sido 10 modelos de procesos ETL desarrollados a partir de estándares (p.e. UML), los cuales presentaban diferentes grados de complejidad, obtenidos variando los valores de las medidas. El propósito de estos modelos es determinar la influencia de la complejidad en los diferentes sujetos. Cada modelo es entregado a los sujetos acompañado de un formulario para recolectar los datos.

5.3 Selección de los sujetos

Para la selección de los sujetos del experimento se uso la técnica de muestreo por conveniencia (no probabilística). Los sujetos tenían amplios conocimientos de modelado (UML, bases de datos, etc.), pero no tenían conocimientos previos acerca del modelado conceptual de procesos ETL con diagramas de actividad de UML. Sin embargo, a todos se les impartió una sesión de entrenamiento, sin que con ello fueran conscientes de los aspectos que se pretendían evaluar. Cada sujeto recibio 10 modelos de procesos ETL, para cada modelo se elaboraron dos cuestionarios: uno relativo a la *entendibilidad* del modelo en el cual se pedía responder (Si o No) a cinco cuestiones y otro relativo a la *modificabilidad* en el que se propuso una serie de cuatro modificaciones a realizar (ver anexo).

5.4 Variables e hipótesis

Para el experimento la variable independiente es la complejidad estructural de los modelos conceptuales de procesos ETL. En el caso de las variables dependientes son dos: la subcaracterística de la usabilidad: *entendibilidad* y la subcaracterística de la mantenibilidad:

modificabilidad de los procesos ETL. Por su parte, las variables dependientes fueron medidas a través de los tiempos de respuesta empleados por los sujetos para llevar a cabo las tareas requeridas. Por otro lado, las hipótesis planteadas de acuerdo a nuestro objetivo de investigación son las siguientes:

Hipótesis de entendibilidad:

- H_{0e}: No hay una correlación significativa entre las medidas de complejidad estructural y el tiempo de entendibilidad.
- H_{1e}: Hay una correlación significativa entre las medidas de complejidad estructural y el tiempo de entendibilidad.

Hipótesis de modificabilidad

- H_{0m}: No hay una correlación significativa entre las medidas de complejidad estructural y el tiempo de modificabilidad.
- H_{1m}: Hay una correlación significativa entre las medidas de complejidad estructural y el tiempo de modificabilidad.

5.5 Análisis de los resultados

El conjunto de datos recolectados sin procesar se encuentra alojado en el sitio web: *http* : *//www.lucentia.es/index.php/ETL_Process_Modelling*. Para su análisis se empleo el software SPSS versión 15 que es una herramienta estadística adecuada a efectos de potenciar el tratamiento del experimento. A partir del resumen de los datos se realizó por tanto su análisis estadístico. Éste resumen estaba compuesto por los valores de las medidas para cada uno de los modelos y por los promedios en los tiempos de entedibilidad y de modificabilidad. En primera instancia, para corroborar si la distribución de los datos obtenidos era normal, se aplicó el test de *Kolmogorov-Smirnov*. La prueba de *Kolmogorov-Smirnov* (también prueba K-S) se basa en la idea de comparar la función de distribución acumulada de los datos observados con la de una distribución normal, midiendo la máxima distancia entre ambas curvas. Como resultado de ello, se obtuvo que la distribución era no normal, por lo que se decidió utilizar un test estadístico no paramétrico como el coeficiente de correlación de *Spearman*. El coeficiente de correlación de *Spearman* es una versión no paramétrica del coeficiente de correlación de *Pearson*, que se basa en los rangos de los datos en lugar de hacerlo en los valores reales [5]. Resulta apropiado para datos ordinales, o los de intervalo que no satisfagan el supuesto de normalidad. Para nuestro experimento se utilizó un nivel de significancia 0.05 ($\alpha = 0.05$) lo cual indica la probabilidad de rechazar la hipótesis nula cuando es cierta, es decir, el nivel de confianza es del 95%. Para una muestra de tamaño 10 y un α =0.05 el umbral de *Spearman* para aceptar H_{0e} y H_{0m} es **0,6320**. Por su parte, cada una de las medidas fue correlacionada separadamente con los tiempos de entendibilidad y modificabilidad.

En la Tabla 3 se muestran los resultados del análisis del experimento, para los tiempos de entendibilidad y modificabilidad. A partir de estos resultados para el experimento, se puede observar que existe una correlación (rechazando la hipótesis H_{0e}) entre los tiempos de entendibilidad y las medidas NES (Número de elementos de salidas en el modelo de procesos ETL), NFES (Número de flujos entrantes y salientes en el modelo de procesos ETL), NEM (Número de elementos Merge en el modelo de procesos ETL) y NTEES (Número total de elementos de entrada y salida en el modelo de procesos ETL). Con respecto al tiempo empleado por los sujetos para realizar las modificaciones de los diagramas, el análisis de correlación evidenció que existe correlación con las medidas NES (Número de elementos de salidas en el modelo de procesos ETL), NEM (Número de elementos Merge en el modelo de procesos ETL) y NTEES (Número total de elementos de entrada y salida en el modelo de procesos ETL).

Tabla 3. Resultados de la correlación de Spearman para los tiempos de entendibilidad y modificabilidad

Medida	Tiempo de Entendibilidad	Tiempo de Modificabilidad
NAP	- 0.022 p=0.952	0.155 p=0.670
NEE	0.205 p=0.569	0.277 p=0.438
NES	$0.637^{(*)}$ p=0.048	$0.705^{(*)}$ p=0.0.23
NFES	0.618 p=0.057	$0.634^{(*)}$p=0.049
NET	0.145 p=0.690	0.034 p=0.927
NEM	$0.728^{(*)}$ p=0.041	$0.812^{(*)}$ p=0.014
NEF	- 0.531 p=0.175	-0.659 p=0.016
NEJ	-0.018 p=0.934	0.143 p=0.787
NOSEF	-0.257 p=0.540	0.353 p=391
NOEEJ	0.224 p=0.562	0.080 p=0.838
NODS	0.495 p=0.146	0.457 p=0.189
NTEES	$0.718^{(*)}$ p=0.019	$0.776^{(*)}$ p=0.008
RFESA	0.508 p=0.134	0.452 p=0.190
RDInEE	-0.070 p=0.848	-0.249 p=488
RDOutES	0.215 p=0.550	0.259 p=0.471

6 Conclusiones y Trabajos futuros

En este trabajo se han presentado la definición y validación de medidas para modelos de procesos ETL en ADs. Estas medidas fueron validadas tanto teórica como empíricamente. Para la validación teórica se utilizó el marco de aproximaciones axiomáticas de Briand et al. [4]. En el marco de la validación empírica se desarrollo un experimento. Dicho experimento se ha centrado en el estudio de la relación entre las medidas propuestas y la entendibilidad y la modificabilidad de los modelos de procesos ETL. Para lo cual, se han considerado los tiempos que invirtieron los sujetos en realizar los ejercicios relacionados con la subcaracterística de la usabilidad: *entendibilidad* y la subcaracterística de la mantenibilidad: *modificabilidad*.

Como resultado de este estudio, podemos concluir que las medidas de NES, NFES, NEM y NTEES son buenos indicadores de mantenimiento. Estas medidas proveen información sobre la mantenibilidad de los modelos de procesos ETL. A mayor mantenibilidad de los modelos de procesos ETL se pueden lograr beneficios significativos en el desarrollo de los ADs en los siguiente aspectos: i)Más facilidad para resolver los cambios en los modelos, ii) Reducción en los costos y en los esfuerzos necesarios para realizar los cambios en los modelos.

Aunque los resultados obtenidos en este experimento son representativos, no podemos considerarlos como resultados concluyentes. Es necesario replicar el experimento y realizar nuevos para poder confirmar estos resultados. Como trabajo futuro inmediato se realizaran réplicas para confirmar los resultados obtenidos, estas réplicas serán desarrolladas con profesionales. Por otro lado, se pretende llevar a cabo casos de estudio usando modelos de procesos ETL reales de empresas.

Agradecimientos

Este trabajo es soportado por los proyectos ESPIA (TIN2007-67078) del Ministerio de Educación y Ciencia de España y QUASIMODO (PAC08-0157-0668) de la Consejería de Educación y Ciencia de Castilla-La Mancha, España. Lilia Muñoz dispone de una beca de la Secretaria Nacional de Ciencia, Tecnología e Innovación (SENACYT) y el Instituto para la Formación y Aprovechamiento de Recursos Humanos (IFARHU), de la República de

Panamá, Jesús Pardillo y José-Norberto Mazón disponen de becas AP2006-00332 y AP2005-1360 respectivamente, del Ministerio de Educación y Ciencia de España.

Anexo

Tabla 4. Ejemplo de cuestionario para la subcaracterística de entendibilidad.

Cuestionario 1

Anotar la hora de inicio (indique hh:mm:ss) ————

1. Contestar las siguientes preguntas:

1. ——¿Se puede llevar a cabo la actividad *SurrogateSales* sin realizar previamente la actividad *SummedSales*?.
2. ——¿Sería la carga del total de ventas diarias de boletos en línea un producto de salida de la actividad *SalesLoader*?.
3. ——¿El proceso ETL finaliza una vez que se llevaron a cabo las actividades *ProductLoader,SalesLoader y TimeLoader*?.
4. ——¿Podría llevarse a cabo el proceso si no se realizara la operación *SummedSales*?.
5. ——¿Deben de agruparse los atributos para poder completar la actividad *SummedSales*.?

2. Según su criterio valore la COMPLEJIDAD del modelo de procesos ETL:

Muy Simple	Algo Simple	Normal	Algo Complejo	Muy Complejo

Anotar la hora de finalización (indique hh:mm:ss) ————

Referencias

1. M. Balta, V. Felea. Using Shannon Entropy in ETL Processes. IEEE Computer Society. ISBN:0-7695-3078-8 pp. 151-156, 2007.
2. V. Basili, and H. Rombach. The TAME Project: Towards Improvement-Oriented Software Environments, IEEE Transactions on Software Engineering, pp. 758-773. 1988.
3. G. Berenguer,R. Romero, J. Trujillo, M. Serrano, M. Piattini. A Set of Quality Indicators and Their Corresponding Metrics for Conceptual Models of Data Warehouses. In: DaWaK 2005: 95-104.
4. L. Briand, S. Morasca, and V. Basili. Property-Based Software Engineering Measurement. IEEE Transactions on Software Engineering, (1996). 22(1), pp. 68-86.
5. G. Corder, D. Foreman. Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach", Wiley (2009).
6. F. García, M. Bertoa, C. Calero et al. Towards a Consistent Terminology for Software Measurement. Information and Software Technology, (2006). 48(8) 631-644.

7. F. García, M. Piattini, F. Ruiz, G. Canfora, C. Visaggio. FMESP: Framework for the modeling and evaluation of software processes. Journal of Systems Architecture 52(11): 627-639 (2006).
8. W. Inmon. Building the Data Warehouse. QED Press/John Wiley, 1992.
9. M. Jarke, M. Lenzerini, Y. Vassiliou, and P. Vassiliadis. Fundamentals of Data Warehouses, second edition ed: Springer-Verlag., 2002.
10. S. March, A. Hevner. Integrated decision support systems: A data warehousing perspective, Decision Support Systems, Volume 43, Issue 3, 2007, pp. 1031-1043.
11. J-N. Mazón, J. Trujillo, and J. Lechtenbörger. Reconciling requirement-driven data warehouses with data sources via multidimensional normal forms. Data & Knowledge Engineering, 2007. 63(3): p.725-751.
12. J-N. Mazón, and J. Trujillo. An MDA approach for the development of data warehouses. Decision Support Systems, 2008. 45(1): p. 41-58.
13. L. Muñoz, J-N. Mazón, J. Pardillo, J. Trujillo. Modelling ETL Processes of Data Warehouses with UML Activity Diagrams pp. 44-53 LNCS 5333, Monterrey, (Mexico), November 9-14, 2008.
14. OMG,"Software Process Engineering Metamodel Specification", adopted specification, version 1.0. Object Management Group, Inc., April, 2008.
15. OMG: MDA Guide (draft version 2). http://www.omg.org/docs/omg/03-06-01.pdf(2003)
16. G. Papastefanatos, P. Vassiliadis, A. Simitsis, Y. Vassiliou. Design Metrics for Data Warehouse Evolution. In: ER 2008: 440-454
17. J. Pardillo, J-N. Mazón, J. Trujillo. Model-Driven Metadata for OLAP Cubes from the Conceptual Modelling of Data Warehouses. In: DaWaK 2008: 13-22.
18. N. Prat, and S. Cherfi. Multidimensional Schemas Quality Assessment. In: CAiSE Workshops 2003.
19. M. Serrano, C. Calero, J. Trujillo, S. Luján, M. Piattini. Towards a Metrics Suite for Conceptual Models of Datawarehouses. Software Audit and Metrics 2004: 105-117
20. M. Serrano, C. Calero, J. Trujillo, M. Piattini. Metrics for data warehouse conceptual models understandability. Information & Software Technology 49(8): 851-870 (2007)
21. F. Shah. Data integration strategies for reliable information delivery. DM Review Magazine, November 2005.
22. C. Shannon. A Mathematical Theory of Communication. The Bell System Technical Journal, Vol. 27, pp. 379-423, 623-656, July, October, 1948
23. C. Shilakes, J. Tylman. Enterprise Information Portals.Enterprise Software Team http://sagemaker.com/ company/downloads /eip/indepth.pdf
24. Simitsis, A., Vassiliadis, P., and Sellis, T. State Space Optimization of ETL Workflows. IEEE Trans. Knowl. Data Eng., 17(10)1404-1419, 2005.
25. A. Simitsis, P. Vassiliadis. A Methodology for the Conceptual Modeling of ETL Processes, in CAiSE Workshops. 2003.
26. M. Solomon. Ensuring A Successful Data Warehouse Initiative. Information Systems Management, 22:1, (2005) 26-36.
27. J. Trujillo, & S. Luján. A UML Based Approach for Modeling ETL Processes in Data Warehouses. In: 22nd International Conference on Conceptual Modeling, (ER'03),Chicago (USA), 2003. 307-320.
28. P. Vassiliadis, Z. Vagena, S. Skiadopulos, N. Karayannidis, T. Sellis. ARKTOS: towards the modeling, design, control and excution of ETL processes. Information Systems 26(8), (2001) 537-561.
29. P. Vassiliadis, A. Simitsis, S. Skiadopoulos. Conceptual modeling for ETL processes, In: ACM Fifth International Workshop on Data Warehousing and OLAP (DOLAP'02), Virginia (USA), 2002.
30. P. Vassiliadis, A. Simitsis, S. Skiadopoulos. Modeling ETL Activities as Graphs. In: DMDW 2002: 52-61.
31. P. Vassiliadis, A. Simitsis, M. Terrovitis, S. Skiadopoulos. Blueprints and Measures for ETL Workflows. In: ER 2005: 385-400.
32. C. Wohlin, P. Runeson, M. Höst, M. Ohlson, B. Regnell, A. Wesslén. Experimentation in Software Engineering: An Introduction. Kluwer Academic Publishers. 2000.