

A Family of Experiments to Validate Measures for an ETL

En: Journal Information and Software Technology, Elsevier. Vol. 52, N.º 11; pp 1188-1203. November, 2010

Autores: Lilia Muñoz¹, Jose-Norberto Mazón, Juan Trujillo

¹Facultad de Ingeniería de Sistemas Computacionales, universidad Tecnológica de Panamá

Contacto: lilia.munoz@utp.ac.pa

Resumen: en los almacenes de datos, los procesos ETL (extracción, transformación y carga) son los encargados de extraer los datos de las fuentes de datos que se contienen en el almacén de datos. Su diseño y el mantenimiento es por lo tanto uno de los pilares de cualquier proyecto de desarrollo del almacén de datos. Debido a su importancia, la calidad de estos procesos debería evaluarse formalmente a principios del desarrollo con el fin de evitar llenar el almacén de datos con datos incorrectos. Con este fin, se presenta un conjunto de medidas con las que se evalúan la complejidad estructural de los modelos de procesos ETL a nivel conceptual. Este estudio, además, acompañado de una validación teóricos y una familia de experimentos cuyo objetivo es teórica y empíricamente validar las medidas propuestas. El uso de estas medidas puede ayudar a los diseñadores a predecir el esfuerzo asociado con las tareas de mantenimiento de los procesos ETL. Esta propuesta se basa en UML (Unified Modeling Language) diagramas de actividad para el modelado de procesos de ETL, y en el (Marco para el Modelado y Evaluación de los Procesos de Software) Marco FMESP para la validación de las medidas.

Palabras claves: procesos ETL; validación de la medida; datawarehouse.

Abstract: In data warehousing, ETL (Extract, Transform, and Load) processes are in charge of extracting the data from the data sources that will be contained in the data warehouse. Their design and maintenance is thus a cornerstone in any data warehouse development project. Due to their relevance, the quality of these processes should be formally assessed early in the development in order to avoid populating the data warehouse with incorrect data. To this end, this paper presents a

set of measures with which to evaluate the structural complexity of ETL process models at the conceptual level. This study is, moreover, accompanied by a theoretical validate and family experiments whose aim is to theoretical and empirically validate 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 modelling ETL processes, and on the FMESP (Framework for the Modeling and Evaluation of Software Processes) framework for the validation of the measures.

Keywords: ETL processes; measure validation; datawarehouse.