Information Systems ■ (■■■■) ■■■–■■■

Contents lists available at ScienceDirect

Information Systems

journal homepage: www.elsevier.com/locate/infosys

Data generator for evaluating ETL process quality

Vasileios Theodorou*, Petar Jovanovic, Alberto Abelló, Emona Nakuçi

Universitat Politècnica de Catalunya, BarcelonaTech, Barcelona, Spain

ARTICLE INFO

Keywords:
Data generator
ETL
Process quality

ABSTRACT

Obtaining the right set of data for evaluating the fulfillment of different quality factors in the extract-transform-load (ETL) process design is rather challenging. First, the real data might be out of reach due to different privacy constraints, while manually providing a synthetic set of data is known as a labor-intensive task that needs to take various combinations of process parameters into account. More importantly, having a single dataset usually does not represent the evolution of data throughout the complete process lifespan, hence missing the plethora of possible test cases. To facilitate such demanding task, in this paper we propose an automatic data generator (i.e., *Bijoux*). Starting from a given ETL process model, *Bijoux* extracts the semantics of data transformations, analyzes the constraints they imply over input data, and automatically generates testing datasets. *Bijoux* is highly modular and configurable to enable end-users to generate datasets for a variety of interesting test scenarios (e.g., evaluating specific parts of an input ETL process design, with different input dataset sizes, different distributions of data, and different operation selectivities). We have developed a running prototype that implements the functionality of our data generation framework and here we report our experimental findings showing the effectiveness and scalability of our approach.

1. Introduction

Data-intensive processes constitute a crucial part of complex business intelligence (BI) systems responsible for delivering information to satisfy the needs of different end users. Besides delivering the right information to end users, data-intensive processes must also satisfy various quality standards to ensure that the data delivery is done in the most efficient way, whilst the delivered data are of certain quality level. The quality level is usually agreed beforehand in the form of service-level agreements (SLAs) or business-level objects (BLOs).

In order to guarantee the fulfillment of the agreed quality standards (e.g., data quality, performance, reliability, recoverability; see [1–3]), an extensive set of experiments over the designed process must be performed to test the behavior of the process in a plethora of possible execution scenarios. Essentially, the properties of input data (e.g., value distribution, cleanness, and consistency) play a major role in evaluating the resulting quality characteristics of a data-intensive process. Furthermore, to obtain the finest level of granularity of process metrics, quantitative analysis techniques for business processes ;(e.g., [4]) propose analyzing the quality characteristics at the level of individual activities and resources. Moreover, one of the most popular techniques for quantitative analysis of process models is process simulation [4], which assumes creating large number of hypothetical process instances that will simulate the execution of the process flow for different scenarios. In the case of data-intensive processes, the simulation should be additionally accompanied by a sample of input data (i.e., *work item* in the language of [4]) created for simulating a specific scenario.

* Corresponding author.
E-mail addresses: vasileios@essi.upc.edu (V. Theodorou), petar@essi.upc.edu (P. Jovanovic), aabello@essi.upc.edu (A. Abelló), emona.nakuci@est.fib.upc.edu (E. Nakuçi).

http://dx.doi.org/10.1016/j.is.2016.04.005
0306-4379/

Please cite this article as: V. Theodorou, et al., Data generator for evaluating ETL process quality, Information Systems (2016), http://dx.doi.org/10.1016/j.is.2016.04.005