Real-Time Data Quality Monitoring System for Data Cleansing

Cihan Varol, Sam Houston State University, USA
Henry Neumann, Sam Houston State University, USA

ABSTRACT
To assist business intelligence companies dealing with data preparation problems, different approaches have been developed to handle the dirty data. However, these data cleansing approaches do not have real-time monitoring capabilities. Therefore, business intelligence companies and their clients are not able to predict the final outcome before running all business process. This yields an extra cost for the company if the data are highly corrupted. Therefore, to reduce cost for these types of businesses, the authors design a framework that monitors the quality attributes during the data cleansing process. Moreover, the system provides feedback to the user and allows the user to restructure the workflow based on quality attributes. The main concept of the framework is based on client-server architecture that uses multithreading to allow real-time monitoring of the process. A child thread is dedicated to run and another is dedicated to monitor the processes and give feedback to the user. The real-time monitoring system not only displays the cleansing process done on the data set, but also estimates the risk propagation probabilities in the data cleansing process. De-duplication elimination, address normalization, spelling correction for personal names, and non-ASCII character removal techniques are employed.

Keywords: Business Intelligence, Business Process, Data Cleansing, Data Quality, Data Quality Monitoring, Risk Assessment

INTRODUCTION
Today, business intelligence companies are collecting large amounts of data from a number of sources. In such an environment, the quality of the data can be affected by a number of different causes that result in unnecessary expenditure for the companies. For example, the Data Warehousing Institute estimates that low-quality customer data cost U.S. businesses about $611 billion a year in excess postage alone (Eckerson, 2002). In a recent example, a pizza chain sending an offer through the mail to the top 20% of its customers missed its target by $0.5M because of bad customer data (Dravis, 2009). The cost of poor-data quality is not always measured in dollars. In 1986, NASA space shuttle Challenger’s solid rocket booster joint seals burst, leading to an explosion that killed seven people. NASA used a flawed decision-
making process to approve the launch of the 
shuttle, which was caused by incomplete and 
misleading information (Rogers, 1986).

As information has become one of the 
most important resources in an organization, 
data and data quality is receiving increased at-
tention as an important and maturing field of 
management information systems. The Total 
Data Quality Management (TDQM) approach 
for systematically managing the data quality 
in organizations is an important paradigm in 
the information and data quality area (Wang, 
1998). In 2002, the Massachusetts Institute of 
Technology launched the Information Qual-
ity Program (MITIQ) where researchers are 
developing and testing new knowledge in the 
data quality field as well as developing data 
quality benchmarking standards. The principles 
that have been driving the data quality field 
for more than 15 years are reflected in Wang 
et al. (1993), Madnick et al. (2009), Strong et 
al. (1997), and Kahn et al. (2002).

Organizations are increasingly interested 
in understanding and monitoring the quality of 
their information through data quality metrics 
and scorecards (Talburt & Campbell, 2006). 
In many of these organizations, data admin-
istrators (DA) are responsible for exploring the 
relationships among values across data sets 
(profiling), combining data residing in differ-
ent sources and providing users with a unified 
view of these data (integrating), parsing and 
standardizing (cleansing), and monitoring of the 
data. Employing only the data administrators 
for intelligent business process can lead to the 
following problems (Varol & Bayrak, 2008):

- The outcome can be error-prone;
- Different selections may be provided for 
  the same job by different DAs;
- A DA may not know to reuse past solutions 
  developed by other DAs;
- The process is labor-intensive. It can take 
  a significant amount of time to produce 
  results.

Problems with the quality of data are 
driving the development of data quality tools 
that are designed to support and simplify the 
data cleansing process. Although there are 
a few open-source data quality tools available, 
a majority of them are created by commercial 
companies in order to address the customers’ 
needs (see Goasdoue et al., 2007; Barateirio & 
Galhardas, 2005, for an exclusive list). These 
commercial business process tools are based on 
workflow structures, where a number of differ-
ent functions work consecutively or in parallel 
one after another. Most of these tools are capable 
of profiling, integrating, and cleansing the data.

Data cleansing is one of the business 
intelligence practices conducted by variety 
of companies. These business intelligence 
companies charge a fee for each cleansing 
technique applied to the data set. However, 
clients would like to assess the quality of the 
original data and the possible outcome before 
allocating large amounts of money for cleansing 
purposes. Moreover, these tools lack real-time 
data process monitoring capabilities. In other 
words, the tools do not reflect the results of 
each cleansing process in real-time. Ideally, 
they should provide real-time checking against 
established business rules and detect when the 
data exceed the pre-set limits. They should also 
provide capabilities to recognize immediately 
and correct issues before the quality of the 
data declines.

In detail, being able to track the data cleans-
ing process in real-time will have these advan-
tages (Cardoso, 2004; Bethem et al., 2002):

- More timely resolution of data
- Reduce subjectivity in data quality 
  interpretations
- Monitoring the system from business in-
  telligence perspective: while fulfilling the 
  customer expectations, the designed model 
  must be constantly monitored throughout 
  its life cycle to assure that both the initial 
  business requirements and the targeted 
  objectives are satisfied. When undesired 
  metrics are identified or threshold values 
  are reached, the real-time monitoring sys-
  tem allows for adaptations of new strategies 
  or the abortion of the process.
Business Intelligence in the Bayou: Recovering Costs in the Wake of Hurricane Katrina
Gregory Smith, Thilini Ariyachandra and Mark Frolick (2012). Organizational Applications of Business Intelligence Management: Emerging Trends (pp. 29-37).
www.igi-global.com/chapter/business-intelligence-bayou/63964?camid=4v1a

Bender Decomposition
www.igi-global.com/chapter/bender-decomposition/107235?camid=4v1a