Chapter 5
The Challenges of Data Cleansing with Data Warehouses

Nigel McKelvey
Letterkenny Institute of Technology, Ireland

Kevin Curran
Ulster University, UK

Luke Toland
Letterkenny Institute of Technology, Ireland

ABSTRACT

Data cleansing is a long standing problem which every organisation that incorporates a form of data processing or data mining must undertake. It is essential in improving the quality and reliability of data. This paper presents the necessary methods needed to process data at a high quality. It also classifies common problems which organisations face when cleansing data from a source or multiple sources while evaluating methods which aid in this process. The different challenges faced at schema-level and instance-level are also outlined and how they can be overcome. Currently there are tools which provide data cleansing, but are limited due to the uniqueness of every data source and data warehouse. Outlined are the limitations of these tools and how human interaction (self-programming) may be needed to ensure vital data is not lost. We also discuss the importance of maintaining and removing data which has been stored for several years and may no longer have any value.

1. INTRODUCTION

Processing and analysing data has become increasingly important to organisations in recent years. As companies are growing and adapting, the ability to retrieve current and correct data is of key importance. Data cleansing, cleaning or scrubbing is the process of identifying and removing or modifying incorrect entries or inconsistencies in a dataset to improve the overall quality (Rahm et al, 2000). Data

DOI: 10.4018/978-1-5225-0182-4.ch005

warehousing is the concept of storing data in a relational database which is designed for query and analysis rather than transaction processes (Docs.oracle.com, 2014). It is also referred to as an organisation's "single source of truth". It is designed to provide management with a large amount of data from multiple sources within the organisation, which is vital in strategic decision making. For data to be stored in a data warehouse, it is crucial that it is cleansed. This process becomes more difficult as retrieving data from multiple sources increases the amount of "dirty data" and may also introduce an inconsistency in the way in which the data is represented.

Figure 1 describes the typical flow and layout of a data warehouse. Extraction, Transformation and Loading is the process reliable for the initial loading and refreshing the contents of the data warehouse. The probability of this data being incomplete or incorrect is quite high as it has been retrieved from multiple sources, therefore the data is processed through a number of methods, which include instance extraction and transformation, instance matching and integration, filtering and aggregation. Data cleansing is normally performed in a separate area before data is loaded into the data warehouse. The sheer volume of data being processed means that writing a successful tool to complete this task is very difficult.

2. DATA QUALITY

Data auditing is the first step in the data cleansing process. Its purpose is to process through the data and outline any data anomalies that are found (Muller et al, 2003). Using statistical and parsing methods, this process derives information such as value range, frequency of values, variance, uniqueness, occurrence of null values, typical string patterns, also detecting any functional dependencies and association rules in the complete data collection (Muller et all, 2003). Data quality refers to the standard, reliability and

Figure 1. Data warehouse model