Chapter 13

Data Profiling and Data Quality Metric Measurement as a Proactive Input into the Operation of Business Intelligence Systems

Scott Delaney
Hydro Tasmania, Australia

ABSTRACT

Business intelligence systems have reached business critical status within many companies. It is not uncommon for such systems to be central to the decision-making effectiveness of these enterprises. However, the processes used to load data into these systems often do not exhibit a level of robustness in line with their criticality to the organisation. The processes of loading business intelligence systems with data are subject to compromised execution, delays, or failures as a result of changes in the source system data. These ETL processes are not designed to recognise nor deal with such shifts in data shape. This chapter proposes the use of data profiling techniques as a means of early discovery of issues and changes within the source system data and examines how this knowledge can be applied to guard against reductions in the decision making capability and effectiveness of an organisation caused by interruptions to business intelligence system availability or compromised data quality. It does so by examining issues such as where profiling can be best be applied to get appropriate benefit and value, the techniques of establishing profiling, and the types of actions that may be taken once the results of profiling are available. The chapter describes components able to be drawn together to provide a system of control that can be applied around a business intelligence system to enhance the quality of organisational decision making through monitoring the characteristics of arriving data and taking action when values are materially different than those expected.

DOI: 10.4018/978-1-4666-4892-0.ch013

INTRODUCTION

Business intelligence and decision support systems are becoming increasingly important in the operation of the modern day corporation with many achieving business or mission critical status (Thierauf, 2001). However, despite their criticality, many of these systems often have fragility in the extract transform and load (ETL) processes that ingest data into data warehouses and similar repositories (Kimball & Ross, 2010). When such processes encounter problems the decision making ability of an organisation is compromised to some degree, either through reduced speed / agility, quality or some combination of the two. It is not uncommon for such circumstances to trigger expensive investigations to uncover the root cause of the problem before designing and implementing a means of rectification. Such circumstances often arise as the result of shifts in the characteristics of the data being loaded in to the business intelligence system.

This paper demonstrates the business case for, and value of, a non-traditional use for data profiling as a solution to this problem. Specifically, an application of these techniques which is able to assist in the early detection of issues within source data with the potential to cause problems when entering the ETL process, thereby protecting against unplanned downtime. In so doing they contribute to the improvement of the overall decision making agility of the enterprise by removing impediments to an organisation's decision making ability, as well as delivering operational cost savings and avoidance.

Techniques allowing practitioners to initiate and carry out such an undertaking in their own organisations will be introduced within the paper. Whilst it is specifically aimed at introducing business intelligence practitioners to methods which they could apply it will also help data and information governance practitioners understand how they can leverage their knowledge to benefit the discipline of business intelligence.

MAIN FOCUS OF THE CHAPTER

Problem Statement

Business intelligence and decision support systems often have fragility in the extract transform and load (ETL) processes that are used to bring data into the data repositories underpinning business intelligence systems. When such processes encounter problems the decision making ability of an organisation is compromised to some degree, either through reduced speed / agility, quality or some combination of the two. It is not uncommon for such circumstances to trigger expensive investigations to uncover the root cause of the problem before designing and implementing a means of rectification. In essence these are reactive data profiling, discovery and data quality rule establishment undertakings. Often such circumstances arise as the result of subtle (or not so subtle) shifts in the profile of the data arriving at the warehouse, but they can also occur as a result of incomplete or immature understanding of the data or the business rules at the time of the design and original implementation of the ETL processes.

A Victim of Our Own Prudence

It can be the very measures put in place to protect the integrity of the business intelligence system, and the data repository below it, that cause the financial (and other) impacts to the business that are described in this chapter. Data warehouses are often protected by their designers and builders by introducing checks and validations at either or both of the database tier or ETL layer. The system analysis and requirements gathering exercises which occurred during the project to bring the business intelligence system in to being will have provided the designers with much information about the data within the source systems which will supply the future data warehouse. It is considered good practice (Kimball, et al, 2008) to ensure that rules and checks provide coverage

Related Content