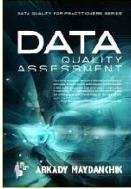


Data Quality Rules: Rules for Event Histories

By
Arkady Maydanchik



Editor's Note: This article is the fourth in a series of excerpts from the book **Data Quality Assessment** by Arkady Maydanchik. The book presents systematic, step-by-step instructions for identifying, warehousing, and analyzing data errors. Part 1 of the series appears in our July 2007 issue. Readers will find a link to the book at our IAIDQ Bibliography page at <http://iaidq.org/main/bibliography.shtml>.

INTRODUCTION

Time is arguably the most important aspect of our life. We are surrounded by calendars and watches, and rare is the activity that does not involve time. When my son entered elementary school, his life became a collection of timestamps and time intervals: school schedule, soccer schedule, play date, time to do homework, TV time, time to play video games, time to go to bed, number of days until Christmas and to the next vacation, and even the number of years left to accumulate college funds. And it stays that way for an entire life, except for rare Hawaii vacations.

This phenomenon stays true in the databases we build. Much of the data is time-stamped, and the absolute majority of the database entities contain histories. In the previous article (*Data Quality Rules: Rules for Historical Data* in the January 2008 issue of the IAIDQ newsletter) we discussed data quality rules for simple value histories. In this article we will discuss event histories.

Car accidents, doctor appointments, employee reviews, and pay raises are all examples of events. Event histories are more complex than value histories. First, events often apply to several objects. For instance, a doctor's appointment involves two individuals – the doctor and the patient. Secondly, events sometimes occupy a period rather than a point in time. Thus, recording a doctor's appointment requires appointment scheduled time and duration. Finally, events are often described with several event-specific attributes. For example, the doctor's appointment can be prophylactic, scheduled, or due to an emergency. It can further be an initial or a follow-up visit, and it will often result in a diagnosis.

In the practice of data quality assessment, the task of validating event histories often occupies the bulk of the project and yields numerous errors. Rules that are specific to event histories can be classified into event dependencies, event conditions, and event-specific attribute constraints.

EVENT DEPENDENCIES

Various events often affect the same objects and therefore may be interdependent. Data quality rules can use these dependencies to validate the event histories. The simplest rule of this kind restricts frequency of the events. For example, patients may be expected to visit the dentist at least every six months for regular checkups. While the length of time between particular visits will differ, it has to be no longer than six months.

Sometimes event frequency can be defined as a function of other data. For example, an airplane is required to undergo extensive maintenance after a certain number of flights. Here, frequency of maintenance events is not a function of time but of another data attribute. Assuming good safety procedures, a greater than required number of flights between maintenance events is a likely indication of a missing record in the event history.

A constraint can also be placed on the number of events per unit of time. For example, a doctor may not be able to see more than 15 patients in a normal workday. A higher number of doctor visits will likely indicate that some of the records in the event history have errors in the name of the doctor or the date of the visit.

The most complex rules apply to situations where events are tied by a cause-and-effect relation. For example, mounting a dental crown will involve several visits to the dentist. The nature, spacing, and duration of the visits are related. Relationships of this kind can get quite complex with the timing and nature of the next event being a function of the outcome of the previous event. For instance, a diagnosis made during the first appointment will influence following appointments.

EVENT CONDITIONS

Events of many kinds do not occur at random but rather only happen under certain unique circumstances. Event conditions verify these circumstances. Consider a typical new car maintenance program. It includes several visits to the dealership for scheduled maintenance activities. These activities may include engine oil change, wheel alignment, tire rotation, and break pad replacement. For each activity, there is a desired frequency.

My new car has a great gadget that reminds me when each of the activities is due. It does it in a beautiful voice, but in no uncertain terms. A typical announcement is “Your tires are due for rotation. Driving the car may be VERY unsafe. Please, make a legal U-turn and proceed to the nearest dealership at a speed of no more than 15 miles an hour.”

Since I do not appreciate this kind of life-threatening circumstance, I decided to visit the dealership before maintenance was due. Unfortunately, for obvious business reasons, the dealership would not do the maintenance before it is due. My only option was to wait for the next announcement and find my way to the nearest dealership at the speed of 15 miles an hour.

On a more serious note, this constraint is an example of *event conditions* – a condition that must be satisfied for an event to take place. Each specific car maintenance event has pre-conditions based on the car make and model, the age of the car, car mileage, and the time since the last event of same type.

All of these conditions can be implemented in a data quality rule (or rules) and used to validate car maintenance event histories in an auto dealership database.

EVENT-SPECIFIC ATTRIBUTE CONSTRAINTS

Events themselves are often complex entities, each with numerous attributes. Consider automobile accidents. The record of each accident must be accompanied by much data – involved cars and their post-accident condition, involved drivers and their accident accounts, police officers and their observations, witnesses and their view of events. The list of data elements can be quite long, and the data may be stored simply in extra attributes of the event table or in additional dependent entities.

Event-specific attribute constraints enforce that all attributes relevant to the event are present. The exact form of these constraints may depend on the nature of the event and its specific characteristics. For instance, a collision must involve two or more cars with two or more drivers, each driver matched to a car. Having two drivers steering the same car is a recipe for collision, or more likely an indication of a data error.

It gets even more exciting when different events may have different attributes. For instance, collision events have somewhat different attributes than hit-and-run events. The former embroils two or more cars, each with a driver; the latter usually involves a single car with no identified driver. Thus the name “event-specific attribute constraints” has two connotations – both the attributes and the constraints are event-specific.

SUMMARY

The critical importance of the event histories in the database makes them the primary target in any data quality assessment project. At the same time, event constraints are rarely enforced by databases, and erroneous data in this area proliferate. In most cases event dependencies can only be found by extensive analysis of the nature of events. Business users will provide key input here. While the data quality rules for event histories are rather complex to design and implement, they are crucial to data quality assessment since they usually identify numerous critical data errors.

About the Author



Arkady Maydanchik is a recognized practitioner, author, and educator in the field of data quality and information integration.

For 10 years, he led Arkidata Corporation, which used his methodology for data quality assessment and data cleansing to provide services to numerous organizations. Arkady is a frequent speaker at The Data Warehousing Institute conferences, IAIDQ conferences, and various data quality seminars.

He has written numerous articles and is author of *Data Quality Assessment*. Arkady also teaches *Practical Skills for Data Quality* courses through his company, Data Quality Group LLC (www.dataqualitygroup.com), and through TDWI (www.tdwi.org/Education/Courses/index.aspx).

He can be reached via email at arkadym@dataqualitygroup.com

© 2007 Arkady Maydanchik