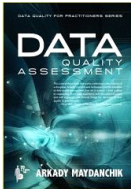


Data Quality Rules: Rules for Historical Data

By
Arkady Maydanchik



Editor's Note: This article is the third 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. You will find Part 1 in our July 2007 issue and Part 2 in our September 2007 issue.

Visit our IAIDQ Bibliography page at <http://iaidq.org/main/bibliography.shtml>.

INTRODUCTION

Most real world objects change over time. Newborn babies grow into playful toddlers, love-stricken teenagers, busy adults, and finally wise matriarchs and patriarchs. Employee positions change over time, their skills increase, and so hopefully do their salaries. Stock markets fluctuate, product sales ebb and flow, corporate profits vary, empires rise and fall, and even celestial bodies move about in an infinite dance of time. We use the term *time-dependent attribute* to designate any object characteristic that changes over time.

The databases charged with the task of tracking various object attributes inevitably have to contend with this time-dependency of the data. Historical data comprise the majority of data in both operational systems and data warehouses. They are also most error-prone. There is always a chance that we'd miss parts of the history during data collection, or incorrectly timestamp the collected records. Also, historical data often spend years inside databases and undergo many transformations, providing plenty of opportunity for data corruption and decay. This combination of abundance, critical importance, and error-affinity of the historical data makes them the primary target in any data quality assessment project.

The good news is that historical data also offer great opportunities for validation. Both the timestamps and values of time-dependent attributes usually follow predictable patterns that can be checked using data quality rules.

TIMESTAMP CONSTRAINTS

Timestamp constraints validate that all required, desired, or expected measurements are recorded and that all timestamps are valid.

A *currency rule* enforces the desired "freshness" of the historical data. Currency rules are usually expressed in the form of constraints on the effective date of the most recent record in the history. For example, the currency rule for annual employee compensation history requires the most recent record for each employee to match the last complete calendar year.

A *retention rule* enforces the desired depth of the historical data. Retention rules are usually expressed in the form of constraints on the overall duration or the number of

records in the history. Retention rules often reflect common retention policies and regulations requiring data to be stored for a certain period of time before it can be discarded. For instance, all tax-related information may need to be stored for seven years pending possibility of an audit. Further, a bank may be required to keep data of all customer transactions for several years.

Values of some attributes are most meaningful when accumulated over a period of time. We refer to any series of cumulative time-period measurements as accumulator history. For instance, product sales history might be a collection of the last 20 quarterly sales totals. Employee compensation history may include annual compensation for the last five calendar years. Accumulator histories are typically subject to these additional constraints:

- A **granularity rule** requires all measurement periods in an accumulator history to have the same size. In the product sales example, it is a calendar quarter; for the employee compensation example, it is a year.
- A **continuity rule** prohibits gaps and overlaps in accumulator histories. Continuity rules require that the beginning date of each measurement period immediately follows the end date of the previous period.

The aforementioned rules enforce that historical data cover the entire desired space of time. However, this does not yet guarantee that the data is complete and accurate. More advanced rules are necessary to identify possibly missing historical records or to find records with incorrect timestamps. All such rules are based on validation of more complex patterns in historical data.

A **timestamp pattern rule** requires all timestamps to fall into a certain prescribed date interval, such as every March or every other Wednesday or between the first and fifth of each month. Occasionally the pattern takes the form of minimum or maximum length of time between measurements. For example, participants in a medical study may be required to take blood pressure readings at least once a week. While the length of time between particular measurements will differ, it has to be no longer than seven days.

Timestamp patterns are common to many historical data. However, finding the pattern can be a challenge. Extensive data profiling and analysis is the only reliable solution. A useful profiling technique is to collect counts of records by calendar year, month, day, day of week, or any other regular time interval.

For example, frequencies of records for each calendar month (year and day of the record does not matter) will tell if the records have effective dates spread randomly over the year or if they follow some pattern.

VALUE CONSTRAINTS

Value histories for time-dependent attributes usually also follow systematic patterns. A **value pattern rule** utilizes these patterns to predict reasonable ranges of values for each measurement and identify likely outliers. Value pattern rules can restrict direction, magnitude, or volatility of change in data values.

The simplest value pattern rules restrict the **direction** in value changes from measurement to measurement. This is by far the most common rule type. Electric meter measurements, total number of copies of my book sold to-date, and the values of many other common attributes always grow or at least remain the same.

A slightly more complex form of a value pattern rule restricts the *magnitude* of value changes. It is usually expressed as a maximum (and occasionally minimum) allowed change per unit of time. For instance, a person's height changes might be restricted to six inches per year. This does not mean that values from measurement to measurement may not change by more than six inches, but rather that the change cannot exceed six inches times the length of the interval in years.

The magnitude-of-change constraints work well for attributes whose values are rather stationary. This does not apply to many real world attributes. For instance, regular pay raises rarely exceed 10-15%, but raises for employees promoted to a new position routinely reach 20-30% or even more. Since the majority of employees experience a promotion at least once in their career, we could not use magnitude-of-change constraint for pay rate histories.

However, pay rates still do not change arbitrarily. Normal behavior of pay rate history for an employee of most companies is a steady increase over the years. Sudden increase in pay rate followed by a drop signals an error in the data (or the end to the dot-com bubble). The value pattern rule that can identify such errors must look for spikes and drops in consecutive values. Here we do not restrict individual value change, but rather do not permit an increase to be followed by a decrease and vice versa. In other words, the rule restricts *volatility* of value changes. Rules of this type are applicable to many data histories.

SUMMARY

Historical data comprise the majority of data in both operational systems and data warehouses. The abundance, critical importance, and error-affinity of the historical data make them the primary target in any data quality assessment project. In this article we discussed data quality rules for the common time-dependent attributes. In the future articles of this series we will address more advanced data categories, such as event histories and state-dependent data.

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