

Faster or Better: The Ultimate Choice

The Data Needs

We live in the Information Age, meaning that information is our most important and valuable resource. We want **more** data and we want each piece of data to serve more purposes. Data warehousing is a great example. We have amazing technology in place to store, manipulate, and analyze unthinkable volumes of data. But the data does not magically materialize in the data warehouse. And so, data warehouses routinely get data from scores of source systems through numerous data interfaces.

As if the desire for more data were not enough, we want to get it **faster**. Monthly batch feeds are the long-forgotten past. Nightly feeds are the norm, and real-time data propagation is often promoted as the way of the future.

Of course, we do not want just any data, but rather we need high-quality data. When the data we get is incomplete, inaccurate, or misunderstood, the warm and powerful Gulfstream turns into a turbulent Maelstrom that turns the data from a great asset into an even greater liability.

The Problem

The problem is that in the real world the adjectives “more” and “faster” are typically in conflict with the equally important “better”. Asking for more data coming faster from more source systems is akin to asking a doctor to see more patients with different health problems while spending less time with each patient.

When confronted with the need to deliver high quality data from source to target systems at the breakneck speed, the data interface designers fall back on a simple philosophy that “the data quality must be managed at the source.” This paradigm, while widely heralded by the data quality professionals is unfortunately misunderstood.

It is certainly true that given a known data problem the best course of action is to perform root cause analysis and fix the problem at its root. This way we do not just reactively cleanse individual erroneous data elements, but rather proactively prevent all future problems of the same kind before they actually occur.

Regardless of the ideal, however, it is not practical or possible to ensure data quality at the source and guarantee that all data coming via interfaces to downstream systems is accurate. There are several important reasons.

1. New problems find their way into the source systems despite the best proactive efforts. The data gathering processes and business processes behind the data are just too complex and fluid to be fully controlled. Further, even with an adequate data quality management program in place, it takes time to identify new problems.

Thus, when the interfaces pick up the data too soon the bad data come across before the problems are identified.

2. In reality, most source systems lack adequate controls. Sometimes it is due to ignorance, but often it is a financial decision. It is often deemed that existing data quality is adequate for the purposes for which the data is used within the source system and investing in data quality improvement has a negative ROI. Of course, such calculations ignore the impact of source data quality on downstream systems such as data warehouses, but it is the reality with which we must contend.
3. Data warehouses obtain data from multiple source systems. Oftentimes, the data coming from each source seems accurate when looked at independently from the other sources. It is only when data from multiple sources is put together that the inconsistencies can be discovered.
4. The structure and meaning of the data in the source systems change regularly to accommodate new data collection processes or as a result of system upgrades. Such changes may undergo regression testing within the system, but in most cases no mechanism is in place to test the implications on the downstream interfaces. As a result, the source data remains accurate but is incorrectly processed on the way to the downstream systems, creating an avalanche of data problems.

The Solution

In the world where scores of systems exchange huge volumes of data at breakneck speed, it is unwise to completely relegate data quality management to the source systems. Monitoring data quality in each interface is a necessary part of any data integration solution!

There are different types of data quality monitors. Error monitors look for individual erroneous data elements. Change monitors look for unexpected changes in data structure and meaning. The level of sophistication within these monitors may vary from simple attribute level screens to advanced data quality rules and statistical monitors. More comprehensive monitors will catch more data problems.

Of course, monitoring data quality in data interfaces is not free. Advanced monitors require greater investment of time and money. Most importantly, the more frequent the data feeds the less opportunity for data quality monitoring is afforded and vice versa. The ultimate decision is a financial trade-off. It requires analysis of the ramifications of bad data on the downstream systems, understanding the limitations of data quality monitoring for each level of feed frequency, and making a conscious comparison of the costs of data inaccuracy with the costs of data latency.

One thing is certain. If we continue to exchange more data at ever increasing speeds with disregard of its implications to data quality management, the next epoch will be known as Disinformation Age.