What is the desirable "upper bound" to data quality?

Tom Breur
February 2010

Introduction

A common argument holds that every company has some "ideal" or "optimal" level of data quality that they should pursue. The underlying assumption is that trying to drive down defect rates further is prohibitively costly. This would become unsustainable because you would need to spend more money than is justified by the reduction in losses. That all sounds very sensible.

Another line of reasoning that we frequently encounter is that data quality levels need to be tuned to their intended use. Extensive cleaning operations that make sense for the census, could be costly overkill for a "quick and dirty" survey. Herzog et al (2007): "How accurate do our data need to be?"

Regardless of what an *acceptable* level of data quality is, errors are almost always attributable to some breakdown in (primary) business processes. This is what Michael Hammer referred to in Reengineering the Corporation (1994): "seemingly small data quality issues are, in reality, important indications of broken business processes."

And even if you are happy to settle for the current level of data quality, you can still learn a lot from studying the errors. This is quite powerful in identifying where your business processes are breaking down. Because every error means a loss, and losses it the corporate bottom line. The business intelligence (BI) function can provide tremendous value by pointing to process errors, and demonstrating where glitches occur in your value chain.

Of course different companies have different cost structures. How much did it cost to acquire the data? What are the decisions being made that hinge on this information? What are the costs/benefits from accurate versus invalid outcomes? Might there be alternative sources of data? How many? How costly are they to acquire? Etcetera.

 © 2010 Tom Breur

All this begs the question what levels of data quality one should strive for, given the situation you are in. It is almost always possible to improve data quality, but at what cost? Does better decision making merit that expense? In other words, how much can you afford to spend in order to nudge up data quality? Is there a "best possible" level of data quality given the incremental expenses required for improvement?

Driving down defect rates

While working on a Lean office transformation, we ran an assessment of costs and losses resulting from sloppy data entry. This was a credit card business where losses on a defaulting card easily run in the thousands. Accuracy targets for data entry were set in the high 90's, which was considered very stringent. On top of that, a timeline was given of two years in which accuracy levels needed to go up, stepwise, for one key metric from 96% to 98%.

Here's what happened. New accuracy targets were included in people's performance objectives. To enable them meeting those goals a data quality scorecard was put in place. After all, you can't manage what you don't measure. The scorecard allowed close monitoring of (individual and group) performance.

Initially, the 96% target was (very) difficult to meet. So there was some disbelief how 98% was *ever* to be attained. Consensus was that higher accuracy levels would be prohibitively expensive to attain, more specifically because of man-hours required. The costs of checking initial data entry and double checking the fixes were simply too high. Also, very "careful" data entry was thought to slow down processing and thus lower back-office productivity.

Interestingly, as an aside, the fault rates on fixes proved to be not all that much lower than the average initial fault rate. And that finding ran firmly counter to intuition. You'd imagine that when people fix an error, they'd make sure to get it right. And they were definitely aware of the importance of striving for zero defects. Yet fixes had to be refixed remarkably often. And then you need to check the refix, of course. And the fix of the refix, etc. …

Continuous improvement

It turns out that the sheer fact that teams (and individuals, upon request) got consistent and detailed feedback on their performance resulted in a remarkable jump in quality. The data quality scorecard provided OLAP functionality, which made it straightforward to analyze in great detail when

 © 2010 Tom Breur

(remarkable) changes in quality had taken place. Both jumps and drops in accuracy percentages precipitated insight, this way. The problem wasn't low data quality *in general*, but rather that it was *sometimes* very low.

In one case, a drop turned out to be attributable to a new hire that had not been properly instructed. Training materials were reviewed, and underscoring the importance of data quality as part of the induction rogram was emphasized once more. Further analysis showed that new hires tend to have somewhat erratic (variable) fault rates in general. The source of their difficulties had to do (in part) with peculiarities in the data entry application.

These IT issues were noted, which eventually led to a change request for the software. But more importantly, the insight triggered continuous monitoring of "new hire performance", to ensure that any such swings would be promptly noted in the future. The power of analytic insight lies in spurring such innovations.

In another case, an improvement was associated with a change in forms. Form design may not be the sexiest job, but it sure has a huge impact on data quality. This holds for paper and web forms alike. Sometimes you may change a form to make it easier for customers, and that is fine (it probably improves response). Make sure you *also* track data quality errors (or ambiguous consumer responses), to ensure it "works" for your back-office as well.

Any time there was some notable change, it wasn't always clear why. But this would later be subject to in-depth analysis in a Kaizen. Kaizens are used to have a small (and empowered) team of workers (*not* external experts or consultants) investigate where process improvement is possible. And how. Interesting spikes (or drops) in data quality would be investigated in Kaizens, and the team then would implement lessons learned themselves.

Taking data quality to a "new" level

Continuous process improvement, of which Kaizens are but one of the instruments, implies that you never ever stop improving. Whenever you remove one obstacle to zero defect processing, you move on to the next, and so on. And every time you take the next step, you leave from an even higher plateau.

Schedule improvement. Don't *wait* for incidents. It's a trap to only scheduling Kaizens when business is slow. These are not "luxury tasks" that you engage in during downtime. An important requirement to make this possible is

 © 2010 Tom Breur

planning your workload. Only if you can predict how much work will come in, can you match this with availability of staff.

When you're constantly swinging between firefighting and slow business, planning for Kaizens is simply too hard. And if you're repeatedly "surprised" by sudden influxes of work because marketing forgot to tell you which campaigns they were planning, opening up that communication channel might well be an excellent Kaizen to start with.

Another important ingredient of Lean processing is to make switching between tasks easy and inexpensive (Womack & Jones, 2003). The more your staff can flex between different streams of work (potentially requiring different skills), the better you can *adapt* to fluctuations in work coming in. You want to be adept at predicting, and superb at adapting.

Each time you embrace innovations from Kaizen teams, you break through new barriers, making room to adopt yet new working practices. As long as you make sure that changes you implement are *sustainable*, there is no upper limit. There is always room for further improvement.

Conclusion

Investments in data quality improvement need to be justified. Only if the quality of decision making improves sufficiently (and represents genuine value) is there merit in spending money. This mechanism demonstrates that marginal improvement needs to outweigh incremental expenditures.

Because of the inherent trade-off between cost of data quality improvement and the value of better decision making, it's only natural to assume some "optimum" data quality level must exist. The implication is that *further* improvement probably costs even more. But this is only partly true.

For one thing, fixing data quality without removing the root cause, surely will become exceedingly expensive. The more problems you fix, the more you enable new sources of error to arise. That dynamic is clearly not sustainable.

Another key point is that data quality errors are typically a sign of some broken business process. When making a business case for data quality improvement, don't just factor in "pure" data quality costs, but also costs of your primary process breaking down.

One client we worked with discovered a "bad" (invalid) count of customer accounts. Many accounts were opened, and activated, but then no money was ever transferred to actually start *using* these accounts. Losses all

 © 2010 Tom Breur

around, with zero value being generated. This was discovered very late. In fact, so late, that following up with these customers, who had been at the very least warm prospects, was now too late. Removing accounts to correct the count was one thing, but the value of lost prospects quite another!

Every time you drive out a root cause for data quality errors, you reach a new, and higher plateau. As long as changes in working practices are sustainable and improve efficiency (which is typical for data quality errors), you are excellently placed to continue on that path of continuous improvement. There is no inherent barrier to working even *more* efficiently.

Ensure you always adopt sustainable working practices. Avoid cutting corners in order to reach progress "faster." The reasoning seems to go: Yes, we know that is how we *should* do things, but for now we'll skip those steps." There's your limit. And that limit now becomes your self inflicted "upper bound."

References

Thomas Herzog, Fritz Scheuren & William Winkler (2007) Data Quality and Record Linking Techniques. ISBN# 0387695028

Michael Hammer (1994) Reengineering the Corporation. ISBN# 1857880560

James Womack & Daniel Jones (2003) Lean Thinking. ISBN# 0743249275

 © 2010 Tom Breur