Data Quality: When Is Good Enough, Enough?

ABSTRACT

One of the major problem-facing practitioners of Data Quality is to know:

1. the Cost Benefit Ratio of Quality, when the knee of the curve has been reached and the cost of a minor increase in Quality becomes excessively expensive and no longer worth the effort
2. What level of Quality is actually needed for the Business Functions at hand? Different levels of materiality exist in the commercial world and the cost to exceed that level of quality may not in fact be worth the effort.
3. That absolute Data Quality, while the Holy Grail of the field, is like the Holy Grail unachievable.

There are several factors at work that, make achieving 100% Data Quality a laudable but unachievable goal, they are:

- Law of Data Uncertainty - this is similar to the Heisenberg Uncertainty, the act of measuring something changes it. We may know the correctness of a data element at a specific time, for a specific domain but not over time for all domains. Mainly because we:
  - Cannot know with certainty that the data has not changed at some time in the future from the time we certified its correctness. For example, we can say that on 10/1/xx the closing price of IBM on the NY Exchange is $50 but we do not know if on 10/2/xx the price will change. Which it often does, corrections are common.
  - That the meaning of a data element changes by the perspective of the person defining it. Closing Price means any of the following depending on who you are:
    - Bid
    - Asked
    - Settled
    - after market trade as of time
- Perspective is everything - the meaning of data changes based on your perspective and who and what you are doing. In fact, there are many Data Stewards for the same data element and they are all providing definitions of the element using a language that is by its nature ambiguous. There are many definitions and nuances for the same word and we are using words to describe the meaning. Accuracy and precision of meaning decreases as the exponential sum of the ambiguity of the language being used to define it. Even if one takes the alternative of mathematical definition, it is a length between 2 and 4 inches, there is an inherent ambiguity since our ability to measure something is dictated by what level of accuracy and precision we wish to include. It is not 3 inches; it is 3.0001 inches plus or minus .001 inch.

In this presentation, we will address this issue and discuss practical and cost beneficial ways to address them.

BIOGRAPHY

Phil Teplitzky
Independent Consultant

Mr. Teplitzky has more than thirty-three years of experience in the Information Technology business both as a senior level Consultant and as a C level executive. Mr. Teplitzky has over his career started two successful consulting companies and been a senior executive at SunGard, SHL SystemHouse and Coopers & Lybrand. Mr. Teplitzky while at SHL SystemHouse grew the business from Six (6) million to a Hundred and fifty million ($150,000,000) run rate in 18 months. He also served as National IT Director for Coopers & Lybrand with responsibility for Data Architectures, Systems Development Life Cycles and Quality Engineering. This included auditing engagements and proposals.

Mr. Teplitzky has material experience in the area of Project Management, Project Management Office, Enterprise Risk Management, Compliance and Regulatory Reporting as well as the establishment of System of Internal Control. The experience is as a Consultant, a member of Audit teams while at Peat Marwick Mitchell & CO as well as a CIO and CTO. Mr. Teplitzky has been instrumental in developing Data awareness both as a frequent lecturer on Data Base topics and as Editor of several Data Base Related journals. Mr. Teplitzky is a founding Vendor member of the International DB2 Users Group and a member of the New York Chapter of DAMA.
In addition Mr. Teplitzky extensive experience in Data Modeling, the use of multiple data base technologies and related tools. Mr. Teplitzky was responsible for creating several Data Base Life Cycles methodologies and integrating them into the overall systems development process. Mr. Teplitzky was served as the consultant to the Joint Task Force of the AICPA, IIA and CICA on Data Base Audit and Control and was instrumental in the Development of the generally accepted Audit and Control standards for Data Bases.

Data Quality
When is Good Enough, Enough

When it is good enough!

Goals & Objectives of Presentation

- Increase your understanding that there is no absolutes when it comes to Data Quality
- To understand that:
  - there only shades of Gray
  - That the level of Quality is dependent upon what we define RIGHT to be
  - Higher levels of Quality cost more $$$
- We may never know or be able to achieve 100% correctness – ***live with it I do!***

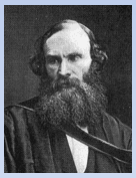
HPSquared LLC

2

Perspective & Context

To measure is to know."

"If you can not measure it, you can not improve it.“



Lord Kelvin

HPSquared

3

To Quote

- ***Quality is everyone's responsibility***

  W. Edwards Deming

- ***Quality is the result of a carefully constructed cultural environment. It has to be the fabric of the organization, not part of the fabric***

  Philip Crosby

HPSquared

4

To Whit

- To know it we have to:
  - be able to Measure it,
  - Understand it, and
  - It is my responsibility
  - And its cultural

Still does not tell me what it is!

5

Quality

- **Is data quality like porn?**
- **I know it when I see it!**

  1964, Justice Potter Stewart of the United States Supreme Court
- *Or is there an actual definition and set of metrics?*

6

Data Quality

A Working definition of Data Quality:

The data is accurate and precise enough to do what I need to do!

What I need to do, knowing what has to be done is the Problem

And what needs to be done varies by person and function

Phil Teplitzky 2010

HPSquared LLC

7

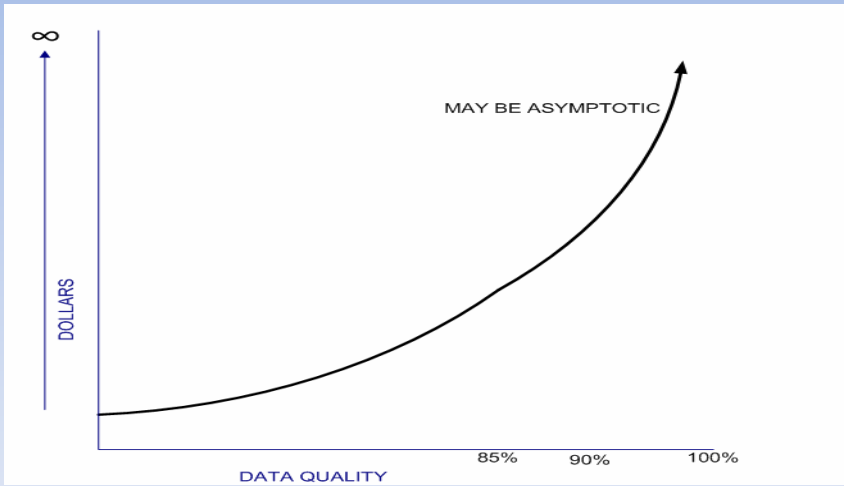
Dimensions of Quality

Dimension	Definition
Accessibility	The extent to which data is available or easily and quickly retrievable
Appropriate Amount of Data	The extent to which the volume of data is appropriate to the task at hand
Believability	The extent to which the data is regarded as true and credible
Completeness	The extent to which data is to missing and is of sufficient breadth and depth for the task at hand
Concise	The extent to which data is compactly represented
Consistent Representation	The extent to which data is presented in the same format
Ease of Manipulation	The extent to which data is easy to manipulate and apply to different tasks
Free of errors	The extent to which data is correct and reliable
Interpretability	The extent to which data is in appropriate languages symbols and units and the definitions are clear
Objectivity	The extent to which data is unbiased and unprejudiced and impartial
Relevancy	The extent to which the data is applicable and helpful for the task at hand
Reputation	The extent to which data is highly regarded and in terms of its source or content
Security	The extent to whcih access to data is restricted appropriately to maintain its security
Timeliness	The extent to which the data is sufficiently up to date for the task at hand
Understandability	The extent the data is easily comprehended
Value Added	The extent to which data is beneficial and provides advantages from its use

HPSquared LLC

8

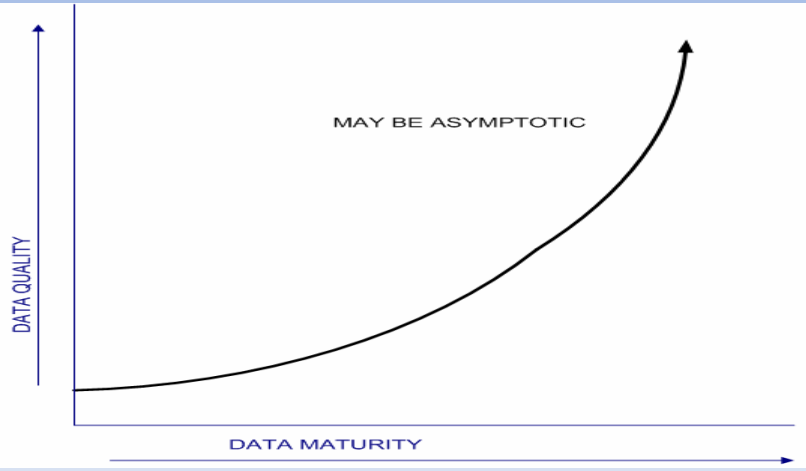
Financial Implications



HPSquared

9

Data Maturity vs. Data Quality



HPSquared

10

Applying Teplitzky's Law of Presentations

So what does this all mean?

1. That data is relative
2. That achieving 100% quality may be impossible or else **really really** expensive
3. That is cannot be achieved by itself but only within the context of an overall data maturity framework

HPSquared

11

Some other less Obvious Considerations

- If Quality is a function of our definition of Correct
- And if our definitions are based on language, which by its very nature is imprecise (see the multiple definitions of common English words) then how accurate, complete and precise can the definitions of the characteristics and attributes used to create the correct definitions?
- You can only be as accurate and precise as the definitions you provide

HPSquared

12

Ah there is the rub!

- Data Quality observes the rules of Quantum Mechanics more than the rules of Newtonian Mechanics!
- It depends on the observer, and it can only be **statistically measured**!!!!
- By and large the only thing we can say for certain is- that if there is an error in Data Quality then there are more!
- If you have not found any – you just have not looked hard enough!

HPSquared

13

Conclusions

- Data Quality like Zeno's Paradox

You can never get to 100% but you are close enough for all practical purposes!

So be a good Quality Engineer and define for good ENOUGH !

AT LEAST FOR WHAT YOUR PURPOSE IT

HPSquared

14