A GENERALIZED APPROACH TO DATA SUPPLY CHAIN MANAGEMENT – BALANCING DATA VALUE AND DATA DEBT

Roberto Maranca[1] and Michele Staiano[2]

[1]Data Excellence – Schneider Electrics, UK
[2]Department of Industrial Engineering – University of Napoli Federico II, Italy

ABSTRACT

The "data supply chains" (DSCs), which are connecting the point where physical information is digitized to the point where the data is consumed, are getting longer and more convoluted. Although plenty of frameworks have emerged in the recent past, none of them, in the authors' opinion, have so far provided a robust set of formalised "how to", that would connect a "well built" DSC to a higher likelihood to achieve the expected value. This paper aims at demonstrating: (i) a generalized model of the DSC in its constituent parts (source, target, process, controls), and (ii) a quantification methodology that would link the underlying current quality as well as the legacy "bad data" to the cost or effort of attaining the desired value. Such approach offers a practical and scalable model enabling to restructure at its foundation some practices of data management priming them for the digital challenges of the future.

KEYWORDS

Data Management, Data Supply Chain, Quality, Complexity, Value.

1. INTRODUCTION

In an increasingly digitised world "Data" is becoming crucial for solving the essential challenges that mankind faces in its way forward. The insight or knowledge that derives from the analysis of the digital representation of reality is more and more required in a world whose complexity and interdependencies grow exponentially.

The management of information has become a key constituency for enterprises, and over the years it has pressurized them into developing complete capability made of people, processes, tools and, obviously, data to operationalise its undertaking. The DSCs are clearly an integral part of the creation of value in human activities, in some cases the most important one, and yet the canonical approach to data in private or public enterprises, though innovating at speed with Data Science (sometimes with inflated expectations), is handling such chains in somewhat artisan fashion.

The principal objective for the exploitation of data is to monitor certain activities from a revenue, performance or compliance point of view, and to optimize the input parameters of such activities to pursue an enterprise's strategic objectives of growth, cost containment and risk management, and also more recently of social responsibilities. However, as the catalyst for the set-up of a data capability varied in time and kind from the more appealing (e.g., digital marketing) to the non-negotiable (e.g., supervisory reporting), the consequence is that within the same company, different areas matured their approach to data at different pace and following different models. In

David C. Wyld et al. (Eds): CMC, NCO, SOFT, CDKP, MLT, ICAITA - 2021
pp. 73-79, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111105

this multi-speed and siloed approach to data, superimposed generations of technologies, processes and procedures have created convoluted (and surprisingly unchartered) internal avenues of distribution and consumption of data; in this scenario, the combined effect of continuous business changes (e.g. mergers, product development, regulations, leadership turnover) and the decreasing ability to respond to such changes with robust simplifications, owing to the increasingly complex enterprise setting, have been feeding each other creating a *chaotic environment*, in which the proverbial flapping of a butterfly's wings can generate unforeseen and very costly consequences. Facing a looming complexity tipping point of the ever more interdependent DSCs, one has just to look at the increased amount of "data breaches" or "data leaks" or "data flops" or "algorithmic failures" to quantify how close the above mentioned complexity tipping point is. Thus, while the data supply chains, DSCs, are getting longer and longer to fuel digital transformations that are coalescing larger and larger ecosystems of functions, intermediaries, partners, and of course third parties (i.e., customers, prospects, accounts), it has emerged a greater awareness of the need to know the what, the where, the who and the how of the enterprise's data, as a risk reduction factor for those unintended consequences. The "Enterprise Metadata Management" discipline – as the ability of collect, organize, relate and take advantage of a set of descriptors of the data used in the enterprise – has greatly increased its presence in the data stacks and has been overtime significantly extended by the raise of the Semantic of Data, already indispensable for the World Wide Web interoperability. However, the authors are hereby going to demonstrate that a further formalization of the DSCs, that connects the semantic approach to a generalised value base and a quantitative model, can provide the basis for the creation of a stronger causality between the assessment of the *quality* of the information *flowing* in a DSC and the predictable and reliable attainment of the intended value.

2. Generalised Data Supply Chain Model

The simplest model underpinning a data supply chain can be described as a single sequential path (see also Fig.1) that comprises:

i. a **point of consumption C** where a set of information $\mathbf{D}_i$ – the data elements, $i = 1, \ldots, n$ – is output and *used* (consumed) by an *agent* $\mathbf{A}_j$, the *data consumers* $j = 1, \ldots, k$, to deliver a tangible or intangible value $\mathbf{V}_{ij}$

ii. **a source S** where the set $\mathbf{D}_i$ is extracted with a process $\mathbf{P}_{ie}$ in conformity with a set of requirements $\mathbf{R}_{ij}$, such process is commonly known as *ETL*, Extract Transformation and Load

iii. a quality process $\mathbf{P}_{iq}$ that produces a set of $\mathbf{Q}_{ij}$ measurements for $\mathbf{D}_i$, based on quality requirements imposed by $\mathbf{A}_j$

iv. a visualization process $\mathbf{P}_{iv}$ that allows $\mathbf{A}_j$ to *consume* $\mathbf{D}_i$ and $\mathbf{Q}_{ij}$

v. a set of tolerances $\mathbf{L}_{ij}$ for each $\mathbf{Q}_{ij}$ imposed by $\mathbf{A}_j$ on the basis of which $\mathbf{D}_i$ is accepted or rejected.

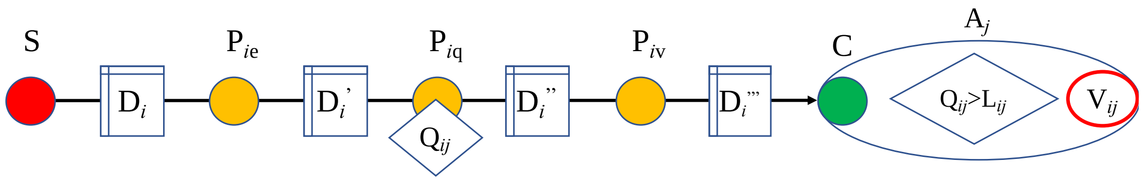

Figure 1. The simplest model for a DSC.

The flow of $\mathbf{D}_i$ is assumed to have a certain cadence **T**, which may vary from a *quasi-real time* – i.e., any time a new set $\mathbf{D}_i$ is available in **S** – to once a day (overnight batches) to *on demand* when $\mathbf{A}_j$ is making a request for the set.

When the flow is established for the first time, it is very likely that the $\mathbf{Q}_{ij}$ could be quite inferior to the level of acceptability, so an "uplift" of quality would be required. Thus, it is useful to express such uplift in a quantitative manner as an amount, $\mathbf{M}_{ij}$, proportional to the measured gap according to the formula $\mathbf{M}_{ij}\_= \mu(\mathbf{Q}_{ij} - \mathbf{L}_{ij})$, where μ is an "issue fixing cost" function. As the sum of all $\mathbf{M}_{ij}$ for all the data sets $\boldsymbol{D}_i$ will constitute the theoretical amount that an enterprise would need to pay to unlock for all the data consumers $\mathbf{A}_j$ the expected $\mathbf{V}_{ij}$ values, we would like to call this the enterprise's ***Data Debt***.

2.1. Simple Data Supply Chain Example

Let's demonstrate with a practical example inspired to a real business situation how the data debt comes into play. A Customer Relationship Management tool is capturing and managing sales opportunities; the enterprise at a time T_0 has got 1000 sales opportunities. An opportunity data set is transferred to Operational Data Store, and it is there consumed by a Sales Director to fine tune their pricing strategy. The pricing strategy has got the obvious intent to increase sales and revenue adjusting the list price for certain specific customers and it is thus based on an internal customer classification. There are 10 different customer classes defined by the enterprise and they are represented by a data element called *customer type*, which would therefore is expected to assume a value between 1 and 10. As the customer type is essential to the action that would derive value from data, it is useful to assign to it the status of *critical data element (CDE)* within the data set. So according to our model above we have (note that, being this simple DSC built against the needs of just one data consumer **A**, in the following we have dropped the second index for the sake of conciseness in the notation):

- **S = *CRM***
- **C = *ODS***
- **A = *Sales Director***
- **V = *Opportunity*(T_1) – *Opportunity*(T_0) > 0**
- **$\mathbf{D}_1$ = *customer type for each opportunity***
- **$\mathbf{P}_{1e}$ = *Extracts Opportunity dataset from S***
- **$\mathbf{P}_{1q}$ = *Execute the Q_1 rule on all the customer types contained in D_1***
- ***q_1 = value output of the Q_1 rule***
- **$\mathbf{P}_{1v}$** = ***Displays for A_1 the list of customer types and the Q_1 result***
- **$\mathbf{L}_1$ = *acceptance level is 1000***

For the sake of simplicity, let consider that in this case **A** will be able to achieve its objective if the output value form the sole rule $\mathbf{Q}_1$ is able to satisfy the requisite $\mathbf{L}_1$:

$$\mathbf{Q}_1 \triangleq \{q_1 = \textbf{Count of all records in } \mathbf{D}_1\textbf{: } \text{'customer type'} \in \mathbf{[1,2,3,4,5,6,7,8,9,10]}\}$$
$$\textbf{is greater or equal to } \mathbf{L}_1$$

However, having checked the 1000 customer types in $\mathbf{D}_1$, it is found that only 800 are valid customer types, so $\mathbf{Q}_1 = 800$. It is important to note that in this case one is not checking whether the customer type is "accurate", i.e., it is exact customer type given the customer is referred to, but the rule only checks whether the customer type is valid, therefore the pricing A_1 will implement would be consistent with the pricing policies but not necessarily yielding the expected result if the customer had been mislabelled with the *wrong* customer type. In any case, for this case, the Data Debt would be $\mathbf{M_1} = \mu(200)$ as 200 are the *issues* affecting the data set as per the rule $\mathbf{Q_1}$; since in the vast majority of cases the cost to fix a single issue can be expressed as a function of time spent by an employee to access and amend the single Customer Type. Let's say that per acquired experience and for the sake of the exercise, the enterprise expects the typical Data Steward to take 30mins to fix one customer type, with a typical hourly rate of 30€/hr, the function $\mu(\cdot)$ is reduced to constant coefficient m_1 that, for our example, yields:

$$\mathbf{M_1} = 200 \times m_1 = 200 \times (0.5\text{h} \times 30€/\text{h}) = 3000€$$

Thus, whatever expectation **A** had of the **value** generated by pricing an opportunity based on customer type, they should add 3000€ of data debt to their cost benefit analysis.

Although this is an extremely simplified case under almost "aseptic" laboratory conditions, the $\mathbf{M_1}$ still constitute a powerful quantification of a *cost* hurdle the DSC has to overcome to start positively to contribute to the bottom-line of the company. Furthermore, as finance and operation functions are getting more proficient in detailing their costs at activity level (as per activity-based costing, ABC), there is an ideal synergy in including data debt considerations in those frameworks.

3. Connecting Data Supply Chain Costs and Value Chains

On the other hand, the more the downstream value creation mechanism is known and the finer the $\mathbf{L}_{ij}$ requirements can be set to optimize the acceptable reduction of $\mathbf{V}_i$ in presence of a greater $\mathbf{M}_i$ carried over: in fact, once the model of a DSC has been defined and the different $\mathbf{M}_{ij}$ have been calculated, the logical next step is to optimize the efforts in data debt reduction to unlock value faster. To this end, let us slightly modify a chart commonly used in stock analysis, a cost/value chart, to look at the relationship between the cost to carry out the establishing of the DSC and improving its data standards, and the value seen from the perspective of the agent $\mathbf{A}_j$.

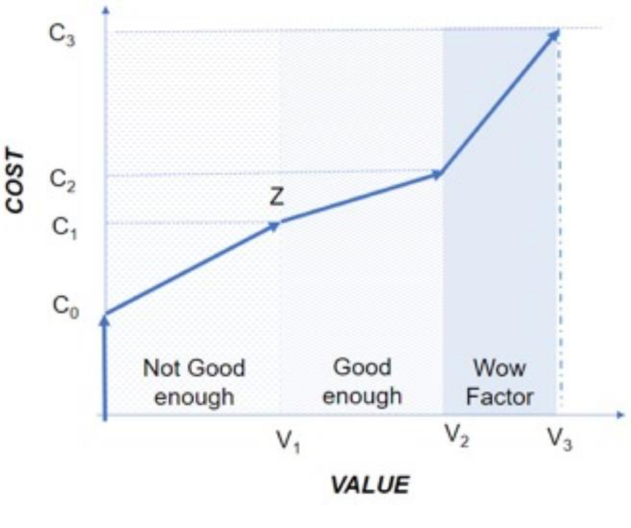

Figure 2. The modified stock chart.

A first initial cost will be required to set up the infrastructure, the process, and the organization to operate the DSC; let denote that cost C_0 and assume it constant for now (i.e., there are no running costs). Obviously, for the agent no value is available at this point in time ($V_0 = 0$), then we could assume that data is starting to flow in the DSC and, especially if working with agile methodologies, very well suiting data, some initial value could be measured, in fact some call this phase “Proof of Value” (POV). In reality, and as per the model, until the point where all the *essential* $\boldsymbol{L_{ij}}$ are satisfied is reached, the DSC is not *good enough* to be *operational* – i.e., to be used in a live business environment to generate value. The point Z, which it is in effect the MVP (minimum viable product) for the DSC, it is identified now as the point where V_1 is achieved at a cost C_1, with C_1 the cost incurred to set up the DSC and to repay the data debt linked to the *minimum consumption requirements.*

What are then these minimum requirements? The expression of $\mathbf{Q}_{ij}$ rules is commonly done using a taxonomy of data quality dimensions as a reference (e.g., completeness, validity, etc.), however rather than picking one it is preferrable to further classify such dimensions in a value generating optic that could marry much more the agent **A**’s view of quality. So in line with the assumption that data should be more and more treated as a product, we could borrow an approach to quality based on customer satisfaction, and the Kano’s model comes handy for simplifying the approach to define the minimum consumption requirement and selecting the $\mathbf{Q}_{ij}$ rules that are instrumental in reaching that point/level. Using the Kano definition *of must-be, one dimensional and attractive* the total cost $C_{1,}$ proportional to the sum of data debts, $\Sigma\mathbf{M}_{ij}$, in the $\boldsymbol{D_i}$, can now be expressed as:

$$C_T = C_0 + C_1 + C_2 + C_3$$

But if we now assume that the three costs are in fact associated to the debt to repay for the fulfilment of must be quality level (= reaching the minimal fitness for consumption), the saturation of one-dimensional quality (= attaining the expected capability) and the achievement of an attractive level (= hitting unspoken needs about data so increasing its value), respectively, then in terms of data debt the formula could be written as:

$$C_T = C_0 + \sum_{i=1}^{n} (M_i^m + M_i^o + M_i^a)$$

which, in the case of MVP as the one where the must be rules about basic/critical requirements ought to be satisfied, becomes:

$$C_T = C_0 + \sum_{i=1}^{n} M_i^m$$

Remarkably the objective of achieving value at the lowest cost can be now visualized geometrically as the reduction of the angle that the segment ending in the point Z measures with the Value axis. The geometry is indeed highlighting a proportion between partial derivatives:

$$\partial C/\partial V \propto \partial C/\partial M$$

where a decrease of Data debt will produce to a proportional increase of the value:

$$dV \propto -dM$$

More pragmatically, from the formula above it is easy to gather that a lower C_1 is achieved with a higher maturity of the Data capabilities of the enterprise. Specifically, this entiles:

- **Lower set up costs (C_0)**:

  a. **Agile Architecture** achieves efficient and rapid instantiation of the DSC
  b. **Robust Delivery Methodology** increases firs time right outcome and minimize resource waste

c. **Data Productization** creates reusable information products to quickly enable consumption

- **Reduced data debt (ΣM*i*)**:
    a. **Active Data Monitoring** capitalizes on previous data debt reduction exercise to keep the target achieved
    b. **Robust Change Control** provides sustainability, so that endogenous (e.g. org changes) or exogenous (e.g. acquisitions) changes are not adversely affecting the quality and integrity of the data

It is worth to highlight, and can be proven, that the additional efforts required to move from *must be* quality (the one related to a minimum viable product) to satisfy *one dimensional* needs are usually comparatively less costly than the former (as the less steep segment between V_1 and V_2 depicts) at least until they cross the line of the *unspoken* needs. In the proposed model these circumstances depend, accordingly to the Kano's theory of attractive quality, on the different quality dimensions that matter in a path toward data excellence. Once the V_2 point is reached by *saturating* the quality standards of the data set, additional unexpected value could only be supplied by increasing quality in a fashion not previously envisioned by the customer themselves, i.e., by capturing a deeper understanding of **A**'s *value chains* to be able to reflect it in the data supply ones. Practically speaking, that would imply that a previously not supplied data element is identified to be beneficial to increase the value, thus changing the DSC structure, and although that would require extra cost for the provider, it could be presented to the agent **A** as a value adding service and afforded in delta value sharing model, that in turn would reinforce the data consumer trust and satisfaction.

4. CONCLUSIONS

The case for a formalized modelling of Data Supply Chains has been introduced, its aim is to create a modular approach that could tackle the complexity tipping point of modern digital enterprises. A causality linkage between the desired outcome of the Data Consumer and the underline status quo of the available data has also been introduced. The concept of Data Debt has been defined as a versatile quantity to gauge the benefit deriving from the DSC itself. A simple example of a practical application of the concept has been provided, drawing a parallel between a quality appreciation model (Kano's) and an optimized approach to converge to minimum value from DSC in an accelerated fashion. Most importantly the introduction of a Cost/Value model has allowed to firmly correlate the quantification of data debt to existing nomenclature of phases (POC, MPV, etc.) adopted in the development of DSCs, phases which are now identified to specific level of debt reductions.

ACKNOWLEDGEMENTS

The authors are grateful to the organizers of a talk addressed to the students of the Master in Data Science administered by University of Naples Federico II for incentivising them to further develop the key ideas behind this paper.

REFERENCES

[1] Ballou, D., Wang, R., Pazer, H., & Tayi, G. K. (1998) "Modeling information manufacturing systems to determine information product quality", *Manage. Sci.,* Vol. 44, No. 4, pp462–484.

[2] Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009) "Methodologies for data quality assessment and improvement", *ACM Comput. Surv.*, Vol. 41, No. 3, pp1-52.

[3] Erto, P., Vanacore, A., & Staiano, M. (2011). "A service quality map based on Kano's theory of attractive quality", *TQM Journal,* Vol. 23, No. 2, pp196-215.

[4] Houhamdi, Z., & Athamena, B. (2019) “Impacts of information quality on decision-making”, *Glob2l Bus. Econ. Rev.*, Vol. 21, No. 1, pp26–42.

[5] Kano, N., Seraku, N., Takahashi, F., & Tsjui, S., (1984) “Attractive quality and must-be quality”, *Hinshitsu*, Vol. 14, No. 2, pp147-56.

[6] Li, A., Zhang, L., Qian, J., Xiao, X., Li, X.-Y., & Xie, Y. (2019) “TODQA: Efficient Task-Oriented Data Quality Assessment”, *Proceedings of 2019 15th International Conference on Mobile Ad-Hoc and Sensor Networks (MSN). IEEE*, pp81-88

[7] Petrakos, G., Conversano, C., Farmakis, G., Mola, F., Siciliano, R., & Stavropoulos, P. (2004). “New ways of specifying data edits”, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, Vol. 167, No. 2, pp249–274.

[8] Silvola, R., Harkonen, J., Vilppola, O., Kropsu-Vehkapera, H., & Haapasalo, H. (2016) “Data quality assessment and improvement”, *Int. J. Bus. Inf. Syst.*, Vol. 22, No. 1, pp62–81.

[9] Batini, C. & Scannapieca M.(2006) *Data Quality – Concepts, Methodologies and Techniques*. Springer, Berlin, Heidelberg.

AUTHORS

Roberto Maranca, during more than 25 years of experience in the world of IT and Data, has spent most his working life with General Electric in their Capital Division, where, since 2014 as Chief Data Officer for their International Unit he has implemented Data Governance, Data Quality and Advanced Analytics, spanning from supporting risk model validation to enabling divestitures and leading their regulatory reporting initiatives. After a year as Group Chief Data Officer at Lloyds Banking Group, shaping a new Data Strategy and dividing his time between the BCBS 239 and GDPR programs, he has joined Schneider Electric as Data Excellence VP delving into the cultural and methodological aspects of becoming a data driven company. Roberto Maranca has got a Master’s Degree in Aeronautical Engineering from Federico II Naples University.

Michele Staiano is a senior researcher in Statistics, Technology and Analysis of Data, STAD, in the Department of Industrial Engineering at University of Napoli Federico II. He has lectured probability, statistics, fundamentals of reliability and innovation courses, and recently developed statistical protocols for Multi-Scale Integrated Analysis of Societal and Ecosystem Metabolism which contributes to integrated research and whole system-level evaluation of sustainability of energy, water, land, food and economies. Currently he is serving as tutor for many students with backgrounds in engineering as well as economics, aiming at helping them to develop actionable data science applications. Michele Staiano holds a Master’s Degree in Aeronautical Engineering and a PhD in Computational Statistics and Applications, both received from Federico II Naples University.