Importance of Data Quality in Existing systems

[1]Ms. Iqbaldeep Kaur, [1]Ms.Nafiza Mann, [1]Ms. Tanisha, [1]Ms.Gurmeen, [2]Ms.Deepi
Associate Professor, Assistant Professor, Assistant Professor, Assistant Professor, Assistant Professor
[1]Department of CSE, [2]Department of ECE, Chandigarh Engineering College, Landran, Punjab, India

Abstract - Existing systems are facing ever more different challenges in managing their systems as developing technologies bring both added complexities as well as opportunities to the way they handle their work. Underpinning this ever-increasing vitality is the importance of having quality data to provide information to make those important system-wide decisions. Numerous studies suggest that many organizations are not paying enough attention to their data and that a major cause of this is their failure to measure its quality and value and/or evaluate the costs of having poor data. Existing databases have a number of particular features that can be described and should be addressed in quality management programme. A professional can make significant contributions to ongoing data quality management and should be alert to data quality issues since they are significant actors. This study provides a mechanism for quantifying data problems, costing potential solutions and monitoring the on-going costs and benefits, to assist them in improving and then sustaining the quality of their data.

I. INTRODUCTION

Data quality is crucial to organizational success due to the increasing amounts and diversity of data processed by organizations. Poor data quality is estimated to cost a company 10–20 % of its revenue. Data quality management is a major concern across organizations and is predicted to gain further importance in the light of increasing amounts and diversity of data, improved analysis capabilities, and business process integration. However, it is not possible to systematically assess costs that are caused by poor data quality since they depend on the context in which the data is used as well as on the impact of direct and hid- den costs of operational and strategic activities and decisions. To assess and sustainably improve data quality within organizations, process-driven data quality management (PDDQM) techniques should be applied. PDDQM aims at redesigning processes that create or modify data. Hence, data and data quality should be taken into attention in the context of the business processes they are processed in. Quality improvement in system development ranks high among the priorities of information systems(IS) managers today. On the other hand IS units are ordered to develop application systems that enable organizations to effectively use information technology. On the other hand, these IS units are facing difficulties in delivering systems that meet user needs in a timely and cost effective manner. A major underlying element of this apparent 'indifference' is that many organizations miss to value either the quality of the data they hold, or the cost of having poor and inaccurate data. If an organization is not able to access the quality of its data how can it determine its value in relation to the corporate decision making process?

Software Quality assurance research has emphasized software quality characteristics, software metrics, and quality control techniques and tools. Key software quality dimensions are portability, reliability, efficiency, human engineering and maintainability which have been identified and defined. A variety of metrics have also been developed and validated for specific software quality characteristics.

II. RESEARCH APPROACH

This on-going investigation is trying to build upon the work of these studies and to develop a specific cost/benefit framework to enable *individual* organizations to: a) analyze the costs of low quality data (consequential costs); b) determine the costs of improving/assuring data quality (investment costs) and c) access 'other' benefits of having quality data. The intended outcomes are to provide mechanisms to: d) identify and analyze the data quality issues; e) build a strong business case to promote improvements, where applicable; f) implement improvement processes; g) establish the on-going monitoring of the quality of the data. It is intended that the outcomes of this study will provide organizations with the opportunity to build this framework within their procedures and systems, both operational and financial, so that the processes will become a permanent integrated management and financial control mechanism to add real value, rather than an occasional one-off ad hoc 'data clean up' exercise. In this way the organization is able to take real 'ownership' of its data.

In this paper one has scoped the problem and based the discussion on reviewing relevant literature, feedback from a related case study, together with one's own experiences from having worked with major organizations related to the quality of organizational data, from which the proposed framework summarized above has been developed. The intention is to conduct a research investigation with a number of Small and Medium-Sized Enterprises (SMEs) to test and refine the proposed framework.

III. PREVIOUS WORK

The lack of quality in the information being provided by data warehouse can lead to bad strategic decisions. Thus, information quality in data warehouse needs to be assured

which further depends on presentation quality, data quality and data model quality (both physical and logical model).

Following on from the themes of English and Loshin, Eppler and Helfert proposed a model which divides data quality costs into two major categories relating to those costs incurred as a result of low quality data and the consequential costs of improving or assuring ongoing data quality. Each category then consists of subordinate categories relating to the direct and indirect costs of poor data and the prevention, detention and repair costs associated with data quality improvement processes as shown in Table 1. Each subordinate category is then further subdivided into six quality costs elements and seven cost improvement elements.

Data quality costs	Costs caused by low data quality	Direct costs	Verification costs
			Re-entry costs
			Compensation costs
		Indirect costs	Costs based on lower reputation
			Costs based on wrong decisions or actions
			Sunk investment costs
	Costs of improving or assuring data quality	Prevention costs	Training costs
			Monitoring costs
			Standard development and deployment costs
		Detection costs	Analysis costs
			Reporting costs
		Repair costs	Repair planning costs
			Repair implementation costs

Table 1: "A data quality cost taxonomy".

Haug, Zachariassen, and van Liempd in Table 2 provide examples of various types of costs, direct (tangible) and hidden (intangible) from both an organizational and strategic perspective.

	Effects of poor quality data on operational tasks	Effects of poor quality data on strategic decisions
Hidden costs	E.g. long lead times, data being registered multiple times, employee dissatisfaction, etc.	E.g. focus on wrong customer segments, poor overall production planning, poor price policies, etc.
Direct costs	E.g. manufacturing errors, wrong deliveries, payment errors, etc.	E.g. few sales, low efficiency, problems in keeping delivery times, etc.

Table 2: Types of Data Quality Costs

Case Study related to Research problem

This article has provided illustrations from the literature to highlight examples of the costs of poor data quality and potential benefits of related improvement programmes. A further example of the effects of such an initiative may be seen from a recent study conducted with a large quasi-public sector organization which has again highlighted the brunt of poor data quality. The organization faced various problems relating to data quality whilst providing its services. The study conducted in the form of focus groups, highlighted a number of key themes relating to data quality.

The main themes identified are as follows. Firstly, in the discussion among the workforce, it was noted that data and information governance were of low priority. Employees' awareness of data governance issues and the associated responsibilities were low; the communication channels used to highlight and promote data quality issues were either non-existent or curbed. Secondly, there was not any formal mechanism or a procedure to report data problems. However one of the positive aspects of the discussion was that the senior management were aware of the data quality issues and the pressures of compliance and were highly keen in improving the current practices and procedures, but the existing organizational culture and the remains of its public sector heritage made their task harder and less effective.

Each of the six focus groups, comprising practitioners from a similar function or department, was asked to undertake separate individual projects to investigate areas within the groups' sphere of influence of actual/potential information risk which were there. Each of the identified risks was summarized (numbered 1-6) and sub-divided further into their more detailed elements and identified (lettered a-d) as appropriate. While it is not possible to measure the above risks and issues in strict monetary terms, an evaluation matrix has been developed based upon a) the level of risk (high, medium, low) and b) the related organizational decision making level (strategic, tactical, operational). Each of the sub-risks (analyzed by major risk 1-6 and detailed risk a-d) was then evaluated as to its potential risk level (high, medium, low) and to which organizational level it related (strategic, tactical, and operational).

Development of a Data Quality Cost/Benefit Framework

The initial research objectives are to develop a framework to enable organizations to: a) analyze the costs of low quality data (consequential costs); b) determine the costs of improving/assuring data quality (investment costs) and c) evaluate 'other' benefits of having quality data and thereby provide them with mechanisms to: d) identify and analyze the data quality issues; e) build a strong business case to promote improvements, where applicable; f) implement improvement processes; g) establish the on-going monitoring of the quality of the data

The outcomes from the above case study, following on from the review of the relevant literature, together with one's own experiences from having worked with major organizations related to the quality of organizational data, an initial conceptual framework has been developed which attempts to embrace the requirements and outcomes of the research objectives as shown in the Initial Data Quality Evaluation Framework Figure 1 below.

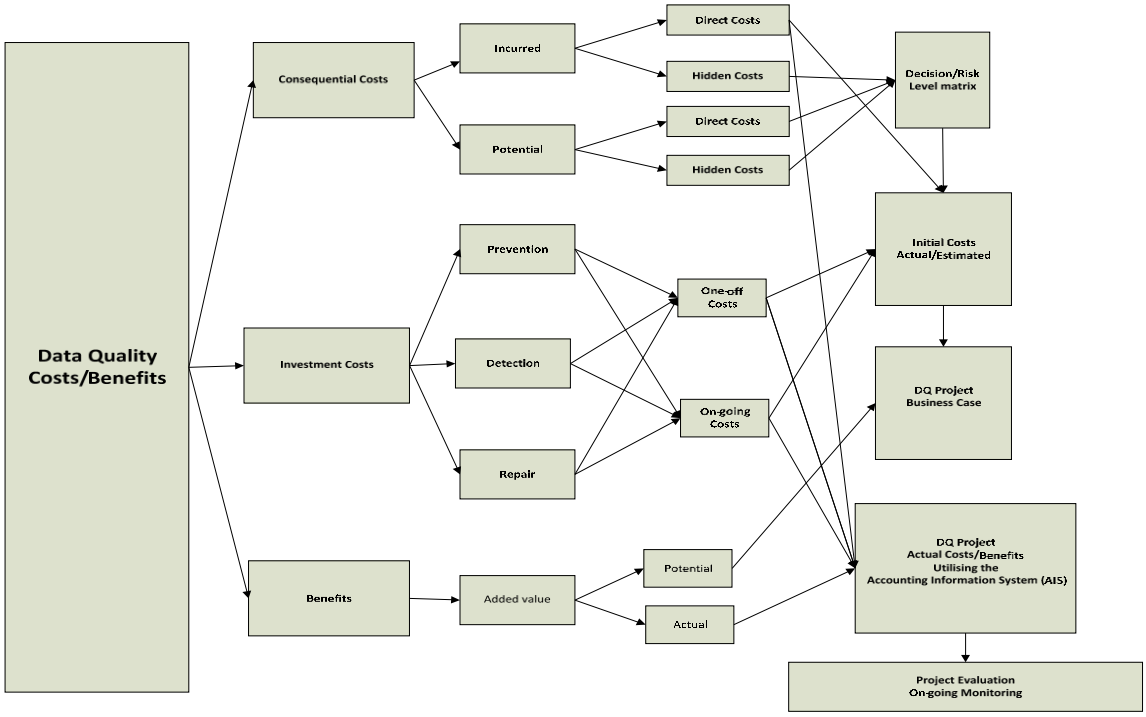

Figure 1: Initial Data Quality Evaluation Framework

The costs and benefits are broken down into three sections:

- Consequential costs- the costs resulting from having low quality data, analyzed between:
  - Costs already incurred
  - Potential costs, which could occur in the future
  - The above costs are then analyzed further into:
    - Direct costs which can be quantified (incurred) or estimated (Potential)
    - Hidden costs which are intangible and difficult to estimate in monetary terms

- Investment costs- the actual costs relating to the process of improving the data, analyzed by
  - Detection costs
  - Repair costs
  - Prevention costs
  - The above costs are then analyzed further into:
    - One-off costs as part of the initial programme
    - On-going costs which will be incurred into the future in order to sustain the programme

- Benefits, those outcomes of the improvement programme which add real value to the business as against reducing costs, analyzed between those that have:
  - Potential to occur
  - Those that actually occur in the future which require to be captured

The essence of this study is to provide organizations with the ability to capture, analyzed and appraise all of the above consequences and outcomes as effectively as possible, initially to ascertain the size of the problem; to prepare a business case for an improvement programme if this is applicable; supervise the actual initial improvement process if this is implemented and then to supervise the subsequent events into the future ,utilizing the organization's existing Accounting Information System (AIS) to induce whether progress if any is being made on an on-going basis. It is recommended that a number of analytical tools be employed to analyze each of the components described above, which can be linked together to provide a composed evaluation process. While it is acknowledged that this is a

'working concept' at this time the process provides an initial robust framework on which to base the initial research.

The Evaluation and Monitoring Process

The process of the cost/benefit evaluation, the monitoring and the ultimate overall project evaluation corresponds to the *right hand portion* of The Initial Data Quality Evaluation Framework: Figure 1 above. *(box 1)* provides a format to analyze, evaluate and prioritize the 'hidden' intangible consequential costs and risks. The actual direct cost together with the estimated one-off and on-going improvement (investment) costs can be evaluated more easily within some form of database/spreadsheet *(box 2*). The outcomes of these two evaluation, together with any estimated potential additional value added benefits can be combined to form the basis of analyzing and subsequently building a valid business case to initiate improvements *(box 3).* Whilst the 'hidden' costs may not be assessed strictly in monetary terms, the matrix can provide a means of evaluating the potential risks, their impact and chances of occurrence which can affect the overall business case decision.

The monitoring of costs and benefits is required if an organization is to manage and control any form of project or programme. Failure to do so is a common source of project failure. It is suggested that an organization can use the analysis and reporting features of its Accounting Information System (AIS) to identify those direct consequential and investment costs and tangible benefits over periods of time. Within a typical AIS the 'general ledger' 'collects' and analyses all types of transactions (costs, revenues, income, assets and liabilities) by way of the 'chart of accounts' and is also able to relate the transactions to a specific business, factory, department, function, location, employee etc. by a designated 'cost or profit centre' *(box 4).* Modern systems have additional features by which transactions can be analyzed usually in the form of 'dimensions'. It is advised that a specific 'dimension' be set up and allocated to each transaction relating to the data quality project whether consequential and investment costs or added value benefits. In this way all actual transactions relating to the data quality programme can be determined by the designated dimension code and subsequently evaluated by type of transaction (cost/benefit) and by location (factory, department etc.) via the AIS reporting structure.

The aftereffect of the business case will provide projections, forecasts, targets, milestones over time, against which the organization can determine the project's actual on-going performance from the AIS general ledger reporting as the necessary part of the Project Evaluation *(box 5*). The intention is that the above framework will be built into an organization's procedures and systems, both operational and financial so that the processes will become persistent business activities rather than occasional one off ad-hoc exercises.

Further Research

It is contended that this project has real possibility to make considerable growth towards achieving the initial research objectives as detailed at the beginning of this section. Further research is needed and to this aim the idea is to conduct a research investigation with a number of Small and medium enterprise (SMEs) to test and refine the proposed framework. At this stage SMEs are considered to be the most applicable type of organization to approach as they come out to be more accessible and provide a wider scope for cooperation than larger organizations.

Concluding Remarks

This study does not intend to provide solutions as to how organizations may improve the quality of their data or to implement changes to sustain such improvements. Rather it states that practical improvement programmes and real process change, cannot take place successfully without some form of assessment and auditing to establish, 'where one is starting from', 'where one wants to go' and 'where one is now' within the overall process. This study therefore is an endeavor to help organizations in making that journey, thereby taking real 'ownership' of its data, by providing a mechanism for quantifying data problems, costing potential solutions and monitoring costs and benefits via an integrated management and financial control mechanism to add real value to its operations.

IV. REFERENCES

[1]. Redman, T. C. (1995) Improve Data Quality for Competitive Advantage. Sloan Management Review, Winter: 99-107.

[2]. Redman, T. C. (1998) The Impact of Poor Data Quality on the Typical Enterprise. Communications of the ACM, 41(2): 79-82.

[3]. English, L. P. (1999) Improving Data Warehouse and Business Information Quality- Methods for Reducing Costs and Increasing Profits. New York: Wiley Computer Publishing: 518.

[4]. Loshin, D. (2001) The Cost of Poor Data Quality. DM Review Magazine, June.

[5]. Haug, A., Zachariassen, F., & van Liempd, D. (2011). The cost of poor data quality. Journal of Industrial Engineering and Management, 4(2), 168-193.

[6]. Eppler M, Helfert M (2004) A classification and analysis of data quality costs. 9th MIT International Conference on Information Quality, November 5-6, Boston, USA

[7]. O'Brien, T, Sukumar, A and Helfert, (2013) The Value of Good Data- A Quality Perspective. The International Conference of Enterprise Information Systems, Angers, France

[8]. Lee YW, Strong DM (2003) Knowing-why about data processes and data quality. Journal of Management

Information Systems 20(3):13–39
[9].Lee YW (2006) Journey to data quality. MIT Press, Cambridge
[10].Ofner MH, Otto B, Österle H (2012) Integrating a data quality perspective into business process management. Business Process Management Journal 18(6):1036– 1067
[11]. Gosain A, Singh J. (2014) Quality Metrics for Data Warehouse Multidimensional Models with Focus on Dimension Hierarchy Sharing. Advances in Intelligent Informatics, Springer; 429-443
[12]. Kumar M, Gosain A, Singh Y. (2013)Empirical validation of structural metrics for predicting understandability of conceptual schemas for data warehouse. International Journal of System Assurance Engineering & Management, Springer, Vol 5, Issue 3; 291-306
[13]. Ali K.B, Gosain A, (2012) Predicting the Quality of Object-Oriented Multidimensional (OOMD) Model of Data Warehouse using Decision Tree Technique, IJSER, Volume 3,Issue 8; 816-820
[14]. Gupta R, Gosain A. (2010) Analysis of data warehouse quality metrics using LR, International conference on Information and Communication Technologies;
[15]. Gupta R, Gosain A. (2012) Validating data warehouse quality metrics using PCA. In: Kannan, R., Andres, F. (eds.) ICDEM 2010. LNCS, vol. 6411, Springer, Heidelberg; 170-172
[16]. Gosain Anjana,Heena(2015),Literature Review of Data model Quality metrics of Data Warehouse, International Conference on Intelligent Computing, Communication & Convergence, Procedia Computer Science 48 ; 236 – 243
[17]. Brien Tony O'(2015)'Accounting' for data quality in enterprise systems, Conference on Enterprise Information Systems / International Conference on Project Management / Conference on Health and Social Care Information Systems and Technologies,Centeris / Projman / Hcist 2015 October 7-9, 2015, Procedia Computer Science 64 ; 442 – 449