International Journal of Research in Social Sciences
Vol. 9 Issue 4, April 2019,
ISSN: 2249-2496 Impact Factor: 7.081
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell’s Directories of Publishing Opportunities, U.S.A

A CONCEPTUAL STUDY ON DATA QUALITY AND ITS IMPACT ON BUSINESS DECISION; A TECHNO-BUSINESS LEADERSHIP PERSPECTIVE

ProfDr.C.Karthikeyan[*]

AsstProfKrishna[**]

Ms Anna Benjamin[***]

Abstract: Data Quality (DQ) is a niche area required for the integrity of the data management by covering gaps of data issues. This is one of the key functions that aid data governance by monitoring data to find exceptions undiscovered by current data management operations. Data Quality checks may be defined at attribute level to have full control on its remediation steps.DQ checks and business rules may easily overlap if an organization is not attentive of its DQ scope. Business teams should understand the DQ scope thoroughly in order to avoid overlap. Data quality checks are redundant if **business logic** covers the same functionality and fulfills the same purpose as DQ. The DQ scope of an organization should be defined in DQ strategy and well implemented. Some data quality checks may be translated into business rules after repeated instances of exceptions in the past.

Key Words: Data; Quality; Logic ; Validity; Accuracy ; Consistency; Remediation; Completeness ; High Impact.

[*] Director and Professor, Management, Studies, T.John College, Bangalore, Affiliated to Bangalore, University, Bangalore, Karnataka, and Accredited by NAAC "A", and Approved by AICTE, New Delhi

[**] Asst Professor, Management Studies,T.John College, Bangalore, Affiliated to Bangalore University,Bangalore, Karnataka, and Accredited by NAAC "A", and Approved by AICTE, New Delhi

[***] Asst Prof, T.John Institute of Management Science, Bangalore, Affiliated to Bangalore University,Approved by AICTE, New Delhi.

Objectives:

(i) To understand the different perceptions and technological developments in data quality issues.

(ii) To learn the different process and metrics governing the data quality management issues.

(iii) To understand the latest technology support and its application in upkeep of data quality.

(iv) To know the application developments happening around data quality management

Review of related Literature:

Fisher et al. (2011) found a strong correlation between accuracy, believability, objectivity, and reputation of data. “The high correlation indicates that the data consumers consider these four dimensions to be intrinsic in nature”. The quality of the data is intrinsic when the quality of the data is directly knowable of the data.

Batini & Scannapieco (2006) emphasize that there are two kinds of data accuracy. One, syntactic accuracy considers the closeness of a value to a definition domain.

Wang & Strong (1996) noted that companies are focusing too much on accuracy as the only data quality dimension. The authors suggest considering a much broader conceptualization of data quality.

Fisher et al. (2011) talk about multiple factors determining this dimension of data quality. One’s knowledge, experience, and the degree of uncertainty in related data are known to be the influencing elements on believability.

Wang & Strong (1996) noted, both data and data sources can build reputation. 2.1.2 Contextual Data Quality This category includes relevancy, completeness, value-added, timeliness, and amount of data (Fisher et al., 2011).

Wang & Strong (1996) brought up that the value-added dimension of data quality can be understood as data that adds value to a company’s operations and, thus, gives the organization a competitive edge. Timeliness refers to how old data is.

Fisher et al. (2011) found some data are affected by age, whereas other data are not. As an example, the authors refer to George Washington, who was the first president of the United States. This information is unaffected by age. Incorrect decisions are often the result of financial decisions that are based on old data. The quantity of information is a serious issue in evaluating data quality. A study on the use of graphs to aid decisions and a phenomenon called information overload was once conducted by Chan (2001).

Wang & Strong (1996) describe representational consistency as data that is continuously presented in the same format, consistently represented and formatted, as well as compatible with data that was presented previously. The scholars list clarity and readability as synonyms for the understandability of data. Attributes comprising the dimension of consistency are as follows: aesthetically pleasing, well-formatted, well-organized, and represented compactly.

Fisher et al. (2011) emphasize that there is a fine line between having troubles excerpting the essential point of an expression that is too long and having problems remembering what an acronym or short expression stands for when shortening long expressions. This could lead to errors in decision-making and, thus, it is suggested that data analysts work with users in determining the ideal version of data presentation. In addition, different users should be involved at different times.

Fisher et al. (2011) illustrate some examples of different data quality dimensions. For instance, high data quality in terms of being accurate means that if there is an inventory database showing that 79 parts are in stock, then there should also be exactly the same amount of items in the stockroom.

Sedera & Gable (2004), enterprise systems success is dependent upon attributes within the dimensions of system quality, information quality, individual impact, and organizational impact.
Sedera & Gable present the following attributes for information quality: Availability, usability, understandability, relevance, format, and conciseness. Moreover, system accuracy is mentioned to belong to the category system quality.

Wang & Strong's (1996) data quality framework will be followed, since most research efforts have been undertaken into this direction.

Eppler & Muenzenmayer (2002) came up with a conceptual framework for information quality in the website context. They generally distinguish between content quality and media quality. For content quality, they further distinguish between relevant information and sound information.

Batini & Scannapieco (2006) talk about research areas that are being discussed in relation to data quality: Statistics: Making predictions and formulating decisions in different sets of contexts even if there is inaccurate data available is possible due to the development of a wide variety of methods and models in this field. Statistical methods help to measure and improve data quality.

Dasu & Johnson(2003) Management information systems: This research area is probably the most relevant for this Master's thesis. Data and knowledge in operational and decision business processes are resources that are gaining in value and importance.

Madnick et al. (2009) note that there are technical and nontechnical issues that may cause data and information quality problems: "Organizations have increasingly invested in technology to collect, store, and process vast quantities of data. Issues surrounding the quality of data and information that cause these difficulties range in nature from the technical (e.g., integration of data from disparate sources) to the nontechnical (e.g., lack of a cohesive strategy across an organization ensuring the right stakeholders have the right information in the right format at the right place and time)."

Tee et al. (2007) show in their article that can be found in the Accounting and Finance Journal. The scholars examined factors that influence the level of data quality in an organization. Senior managers as well as general users were sampled through interviews and surveys in a target organization.

Petter, DeLone & McLean (2008). The scholars point out that there are six major dimensions that are known to have an influence on the successful usage of information systems: system quality, information quality, service quality, use, user satisfaction, and net benefits.

Wang & Strong's (1996) framework of data quality dimensions. For example, understandability and user friendliness of a system are two attributes of system quality. These might be closely related to ease of understanding as well as interpretability of data in Wang & Strong's framework.

Sedera & Gable (2004) argued that overall productivity of an organization has an impact on the success of enterprise systems, whereas Fisher et al. (2011: 4) summarize that data quality in organizations has an influence on productivity. In a distributed project setting, the quality of aggregate project-status data that needs to be sent between organizations can be a major problem in a lot of companies.

Cao & Zhu (2013) view it from a different perspective and talk about data quality problems in ERP-enabled manufacturing. Data Quality Organizational Performance Enterprise Systems Success 11 Changes in the Bill of Materials (BOM) require adjustments in calculating materials required, and in generating product, purchase, as well as work orders. The scholars found that adjustments of these data were especially difficult if the Bill of Materials had to be changed frequently.

McNaull et al. (2012), assisted living technologies use artificial intelligence and automated reasoning to understand the behavior of people who need care due to chronic diseases, and people who need health and social care provision due to their age. Inherently, ambient intelligence-based systems, or Ambient Assisted Living (AAL) technologies make it possible for people to extend the time they live at home by providing feedback to users and carrying out particular actions based on patterns that these systems are able to observe.

Curé (2012) emphasizes the importance of high data quality in drug databases which are often exploited by health care systems and services. Poor data quality, e.g. the inaccuracy of drug contraindications, can have a severe negative impact on a patient's health condition. The author notes that data quality should be ensured in terms of data completeness and soundness.

Olivier Curé presents special technologies to represent hierarchical structures of pharmacology information (e.g. the technology of the Semantic Web). Moreover, SPARQL is presented in the article as a query language for resolving issues of conditional dependencies (CINDs – conditional inclusion dependencies) for these graph-oriented structures.

Quality Pipino, Lee & Wang (2002) tried to answer the question of how good a company's data quality is. The authors describe principles to develop data quality metrics that are useful for measuring data quality. The core of their study was the presentation of three functional forms for developing objective data quality metrics.

Embury et al. (2009) talk about variability of data quality in query-able data repositories. Data with low quality can be useful, but only if data consumers are aware of the data quality problems. Quality measures computed by the information provided have been used to incorporate quality constraints into database queries. The authors describe the possibility of embedding data quality constraints into a query. These constraints should describe the consumer's data quality requirements. The problem that the research team attempted to address was that poor data quality is a consequence of information providers who define quality constraints. Their idea was to increase the level of data quality by incorporating quality constraints into database queries whereas users define quality such that domain-specific notions of quality can be embedded.

Heinrich & Klier (2011) propose a novel method for assessing data currency (one dimension of data quality, as mentioned earlier). Data currency is an important aspect of data quality management. In terms of quality in information systems, the authors distinguish between quality of design and quality of conformance. The latter is essential for this study, since it refers to "the degree of correspondence between the data values stored in a database and the corresponding real world counterparts”. As an example, data values stored in the database might not be up-to-date and, thus, lack quality of conformance. In other words, these data sets do not correspond with their real world counterparts.

Berti-Équille et al. (2011) propose a novel approach for measuring and investigating information quality. The scholars developed a model which can be transversally applied by users, designers, and developers. In their study, the quality of customer information at a French electricity company and patient records at a French medial institute were analyzed to create a multidimensional notion of multidimensional information exploration.

Ballou and Pazer contributed a pioneering work for information quality research and they are the authors with the greatest impact on the study of information quality in information system research. The work of Delone and Mclean (1992) reviewed information system literature and is cited by subsequent information quality research.

Delone and Mclean have done a significant work that further derives information quality from information systems. Figure 4 also has shown that extensive information quality research was produced around 1995.

Keller and Staelin (1987) investigate how decision effectiveness is affected by both information quality and information quantity. By employing social interaction and decision aids.

Sage (1991) imply that the major purpose of social interaction and decision aids is to enhance in-16 formation quality, and through this, the quality of decision-making. Based on crisis decision environments and decision aids.

Belardo and Pazer (1995) propose a model to present the relationship between information quality and decision quality. Considering decision strategy and decision costs.
Ballou and Pazer (1995) analyse the trade-off between two information quality dimensions (accuracy and timeliness) in decision-making. Taking task complexity and decision strategy into account.

Chengalur-Smith et al (1999) show that including information regarding information quality can impact upon the decision-making process.

Fisher et al (2003) develop an experiment to address the utility of information quality information in decision-making. In a dynamic decision environment.

Shankaranarayan (2003) proposes a virtual business environment to address the role of information quality management in dynamic decision environments. From the perspective of task complexity and information quality categorisation.

Jung and Olfman (2005) find that contextual information quality significantly contributes to decision performance. From the above literature, we can observe that various factors are considered when researchers study the effects of information quality on decision-making.

Tsichritzis and Lochovsky (1982) propose that data is based on a modelling of the real world. In this thesis, we define data as model of the real-world facts and are typically represented by in the database. In information quality research, data is often considered to be the raw material for information manufacturing.

Gilmore (1974) defines quality as conformance to specifications. This definition is relatively straightforward and frequently used in manufacturing industries. It facilitates measurement and increases measuring efficiency. Organisations can determine the quality of products by measuring how well the product conforms to an established specification. Also, the measuring procedure can be automatically implemented. However it fails to capture the customer’s view on product performance. To compensate for the disadvantage of this definition,

Gronroos (1983) defines quality as meeting and/or exceeding customer’s expectations. This definition is especially prevalent in marketing research and the service industries. Following this definition, researchers posit that it is the customer who is the ultimate judge of the quality of a product/service. Thus organisations can make a quick response to market changes. However, it is difficult to measure the extent to which a product/service meets and/or exceeds the customer's expectation.

Juran (1974) introduces the definition of quality as fitness for use, which is used to measure the extent to which a product successfully serves its intended use. As this is the definition widely used in information quality research, we adopt fitness for use as the definition of quality in our research.

Wang and Strong (1996)'s definition is widely accepted by subsequent research. Their definition stems from user perspective and can directly capture the user's opinion on the information. However, definitions from the information perspective are practical for objective assessment. As we differentiate information products from raw data, we consider the quality of information products from the user perspective and the quality of raw data from the information perspective.

Cushing (1974) proposed a mathematical approach to detect and prevent data errors in accounting internal control systems.

Hilton (1979) provided an illustrative analysis to show the effect of information accuracy and timeliness on accounting. In the mean time, psychology and management researchers also began to consider the concept of information quality.

Streufert (1973) used an experimental methodology to investigate the effects of information relevance on decision-making.

Zmud (1978) carried out an empirical study to derive certain information dimensions - relevant, accurate, factual, quantity, reliable, timely, arrangement, readable and reasonable. Most of these dimensions are confirmed as information quality dimensions by subsequent research (e.g. Wang and Strong 1996, Lee et al. 2002).

Olson and Lucas (1982) used appearance and accuracy to measure data quality in office automation information systems.

Morey (1982) considered information quality to be 23 information accuracy and proposed three information accuracy measures in the context of information systems.

O' Reilly (1982) investigated the effects of information quality on the use of information sources. In his study, information quality is measured by accessibility, accuracy, specificity, timeliness, relevance and amount of information.

Ballou and Pazer (1985) considered accuracy, completeness, timeliness and consistency in the measurement of information quality in multi-input and multi-output information systems.

Laudon (1986) identified completeness, accuracy and ambiguity as information quality dimensions for criminal-record systems. Observing the works above, it is found that different information quality dimensions can be derived from different contexts. Additionally, in this phase different assessment methodologies were proposed.

Paradice and Fuerst (1991) developed a quantitative measure to formulate the error rate of stored records in information systems.

O' Reilly (1982) proposed 18 questions with which users could determine information quality. One prominent work in this period is by Agmon and Ahituv (1987). They proposed an approach to assess information quality consisting of three elements: (1) internal assessment, to determine the intrinsic characteristics of data using widely accepted criteria, (2) assessment from the users' perspective, which focuses on the expectations and requirements of the user, and (3) assessment by comparison between database and reality.

Garvin (1988) pointed out three types of information quality problems: biased information, outdated information, and massaged information. Biased information means the content of the information is inaccurate or distorted in the transformation process. Outdated information is information that is not sufficiently up to date for the task. Massaged information refers to different representations of the same information.

Lesca and Lesca (1995) classified information quality problems into the product and process views. The product view focuses on the deficiencies of the information itself, such as incompleteness and inconsistency, whilst the process view concentrates on the problems that are caused in the information production and distribution process.

Wang and Strong (1996) propose three approaches to study information quality: the intuitive, theoretical and empirical approaches. We adopt these approaches in our analysis of the identification of information quality dimensions. The intuitive approach derives information quality dimensions from the researchers' experience or from the requirements of particular cases. In this approach, information quality dimensions are identified according to specific application contexts. For example, O'Reilly (1982) used accessibility, accuracy, specificity, timeliness, relevance, and the amount of information to assess information quality in the context of decision-making.

Ballou and Pazer (1985) employed accuracy, timeliness, completeness and consistency to model information quality deficiencies in multi-input, multi-output information systems. The theoretical approach generates information quality dimensions on the basis of data deficiencies in the data manufacturing process.

Wand and Wang (1996) used an ontological approach to derive information quality dimensions by observing inconsistencies between the real-world system and the information system. The empirical approach provides information quality dimensions by focusing on whether the information is fit for use by information consumers. For example,

Kahn et al. (2002) selected 16 information quality dimensions for delivering high quality information to information consumers. From the discussion above, we found that varying sets of information quality dimensions can be identified using different approaches.

Ballou and Pazer (1985) defined completeness as a situation in which all values for a certain variable are recorded. The theoretical approach defines information quality dimensions from the real-world 32 perspective.

Wand and Wang (1996) defined completeness as the ability of an information system to represent every meaningful state of the represented real world system. The empirical approach defines information quality dimensions from the user's perspective.

Wang and Strong (1996) defined completeness as the extent to which data are of sufficient breath, depth, and scope for the task at hand. We describe the approaches and perspectives for defining information quality dimensions.

Wang and Strong (1996) proposed a hierarchical framework that consists of four information quality categories: intrinsic information quality, contextual, representational and accessibility.

Wand and Wang (1996) used an ontological approach to derive information quality dimensions and categorised them by internal view and external view. Internal view is use-independent and contains a set of information quality dimensions that are comparable across applications. External view is concerned with the use and effect of information systems, which represent the real-world system.

Naumann and Rolker (2000) organised information quality dimensions with three main factors that influence information quality: the perception of the user, the information itself, and the process of accessing information. These three factors can be considered as subject, object and process.

Helfert (2001) classified information quality dimensions by employing semiotics and two elements of quality, which are quality of design and quality of conformance. Semiotics comprises three levels: syntactic, semantic and pragmatic. The syntactic level considers the basic representation of information. The semantic level focuses on information related to real world objects. Finally the pragmatic level deals with information processes and information users.

Kahn et al. (2002) developed a two-by-two conceptual model to describe information quality dimensions. Whilst the two rows are product quality and service quality, the two columns are conformance to specifications and meeting and exceeding consumer expectations. Therefore information quality dimensions are located in four quadrants: sound, dependable, useful, and usable.

Bovee et al. (2003) presented a categorisation of information quality dimensions with the sequence of using information. The sequence includes the following four steps: obtaining the information (accessibility), understanding the information (interpretability), connecting the information with the given context (relevance), and assuring the information is free from error (integrity).

Bovee et al. (2003) Sequence of using data Accessibility Interpretability Relevance Integrity Classifications of information quality dimensions Using the above classifications and widely accepted information quality dimensions with characteristics of each classification.

Introduction: **Data quality** refers to the condition of a set of values of qualitative or quantitative variables. There are many definitions of data quality but data is generally considered high quality if it is "fit for intended uses in operations, decision making and planning." Alternatively, data is deemed of high quality if it correctly represents the real-world construct to which it refers. Furthermore, apart from these definitions, as data volume increases, the question of internal data consistency becomes significant, regardless of fitness for use for any particular external purpose. People's views on data quality can often be in disagreement, even when discussing the same set of data used for the same purpose. Data cleansing may be required in order to ensure data quality. "Data Quality: High-impact Strategies". See also the glossary of data quality terms. Degree of excellence exhibited by the data in relation to the portrayal of the actual scenario. The state of completeness, validity, consistency, timeliness and accuracy that makes data appropriate for a specific use. The totality of features and characteristics of data that bears on its ability to satisfy a given purpose; the sum of the degrees of excellence for factors related to data. The processes and technologies involved in ensuring the conformance of data values to business requirements and acceptance criteria. Complete, standards based, consistent, accurate and time stamped.If the ISO 9000:2015 definition of quality is applied, data quality can be defined as the degree to which a set of characteristics of data fulfills requirements. Examples of characteristics are: completeness, validity, accuracy, consistency, availability and timeliness. Requirements are defined as the need or expectation that is stated, generally implied or obligatory. Before the rise of the inexpensive computer data storage, massive mainframe computers were used to maintain name and address data for delivery services. This was so that mail could be properly routed to its destination. The mainframes used business rules to correct common misspellings and typographical errors in name and address data, as well as to track customers who had moved, died, gone to prison, married, divorced, or experienced other life-changing events. Government agencies began to make postal data available to a few service companies to cross-reference customer data with the National Change of Address registry (NCOA). This technology saved large companies millions of dollars in comparison to manual correction of customer data. Large companies saved on postage, as bills and direct marketing materials made their way to the intended customer more accurately. Initially sold as a service, data quality moved inside the walls of corporations, as low-cost and powerful server technology became available. Companies with an emphasis on marketing often focused their quality efforts on name and address information, but data quality is recognized as an important property of all types of data. Principles of data quality can be applied to supply chain data, transactional data, and nearly every other category of data found. For example, making supply chain data conform to a certain standard has value to an organization by: 1) avoiding overstocking of similar but slightly different stock; 2) avoiding false stock-out; 3) improving the understanding of vendor purchases to negotiate volume discounts; and 4) avoiding logistics costs in stocking and shipping parts across a large organization. For companies with significant research efforts, data quality can include developing protocols for research methods, reducing measurement error, bounds checking of data, cross tabulation, modeling and outlier detection, verifying data integrity, etc. There are a number of theoretical frameworks for understanding data quality. A systems-theoretical approach influenced by American pragmatism expands the definition of data quality to include information quality, and emphasizes the inclusiveness of the fundamental dimensions of accuracy and precision on the basis of the theory of science (Ivanov, 1972). One framework, dubbed "Zero Defect Data" (Hansen, 1991) adapts the principles of statistical process control to data quality. Another framework seeks to integrate the product perspective (conformance to specifications) and the service perspective (meeting consumers' expectations) (Kahn et al. 2002). Another framework is based in semioticsto evaluate the quality of the form, meaning and use of the data (Price and Shanks, 2004). One highly theoretical approach analyzes the ontological nature of information systems to define data quality rigorously (Wand and Wang, 1996).A considerable amount of data quality research involves investigating and describing various categories of desirable attributes (or dimensions) of data. These dimensions commonly include accuracy, correctness, urrency, completeness and relevance. Nearly 200 such terms have been identified and there is little agreement in their nature (are these concepts, goals or criteria?), their definitions or measures (Wang et al., 1993). Software engineers may recognize this as a similar problem to "ilities".MIT has a Total Data Quality Management program, led by Professor Richard Wang, which produces a large number of publications and hosts a significant international conference in this field (International Conference on Information Quality, ICIQ). This program grew out of the work done by Hansen on the "Zero Defect Data" framework (Hansen, 1991).In practice, data quality is a concern for professionals involved with a wide range of information systems, ranging from data warehousing and business intelligence to customer relationship management and supply chain management. Most data quality tools offer a series of tools for improving data, which may include some or all of the following: **Data profiling** - initially assessing the data to understand its quality challenges. **Data standardization** - a business rules engine that ensures that data conforms to quality rules. **Geo coding** - for name and address data. Matching or Linking - a way to compare data so that similar, but slightly different records can be aligned. Matching may use "fuzzy logic" to find duplicates in the data. It often recognizes that "Bob" and "Robert" may be the same individual. It might be able to manage "house holding", or finding links between spouses at the same address, for example. Finally, it often can build a "best of breed" record, taking the best components from multiple data sources and building a single super-record. **Monitoring** - keeping track of data quality over time and reporting variations in the quality of data. Software can also auto-correct the variations based on pre-defined business rules. **Batch and Real time** - Once the data is initially cleansed (batch), companies often want to build the processes into enterprise applications to keep it clean. There are several well-known authors and self-styled experts, with Larry English perhaps the most popular guru. In addition, IQ International - the International Association for Information and Data Quality was established in 2004 to provide a focal point for professionals and researchers in this field.ISO 8000 is an international standard for data quality. **Data quality assurance;** Data quality assurance is the process of data profiling to discover inconsistencies and other anomalies in the data, as well as performing data cleansing activities (e.g. removing outliers, missing data interpolation) to improve the data quality. These activities can be undertaken as part of data warehousing or as part of the database administration of an existing piece of application software.

Objective: (i) ;To understand the different perceptions and technological developments in data quality issues.

Data quality control; Data quality control is the process of controlling the usage of data with known quality measurements for an application or a process. This process is usually done after a Data Quality Assurance (QA) process, which consists of discovery of data inconsistency and correction. Data QA processes provides following information to Data Quality Control (QC):Severity of inconsistency, Incompleteness, Accuracy, Precision, Missing / Unknown. The Data QC process uses the information from the QA process to decide to use the data for analysis or in an application or business process. For example, if a Data QC process finds that the data contains too many errors or inconsistencies, then it prevents that data from being used for its intended process which could cause disruption. For example, providing invalid measurements from several sensors to the automatic pilot feature on an aircraft could cause it to crash. Thus, establishing data QC process provides the protection of usage of data control and establishes safe information usage.

Optimum use of data quality; Data Quality (DQ) is a niche area required for the integrity of the data management by covering gaps of data issues. This is one of the key functions that aid data governance by monitoring data to find exceptions undiscovered by current data management operations. Data Quality checks may be defined at attribute level to have full control on its remediation steps. DQ checks and business rules may easily overlap if an organization is not attentive of its DQ scope. Business teams should understand the DQ scope thoroughly in order to avoid overlap. Data quality checks are redundant if **business logic** covers the same functionality and fulfills the same purpose as DQ. The DQ scope of an organization should be defined in DQ strategy and well implemented. Some data quality checks may be translated into business rules after repeated instances of exceptions in the past. Below are a few areas of data flows that may need perennial DQ checks:

Completeness and **precision** DQ checks on all data may be performed at the point of entry for each mandatory attribute from each source system. Few attribute values are created way after the initial creation of the transaction; in such cases, administering these checks becomes tricky and should be done immediately after the defined event of that attribute's source and the transaction's other core attribute conditions are met.All data having attributes referring to *Reference Data* in the organization may be validated against the set of well-defined valid values of Reference Data to discover new or discrepant values through the **validity** DQ check. Results may be used to update *Reference Data* administered under *Master Data Management (MDM)*.All data sourced from a *third party* to organization's internal teams may undergo **accuracy** (DQ) check against the third party data. These DQ check results are valuable when administered on data that made multiple hops after the point of entry of that data but before that data becomes authorized or stored for enterprise intelligence. All data columns that refer to *Master Data* may be validated for its **consistency** check. A DQ check administered on the data at the point of entry discovers new data for the MDM process, but a DQ check administered after the point of entry discovers the failure (not exceptions) of consistency. As data transforms, multiple timestamps and the positions of that timestamps are captured and may be compared against each other and its leeway to validate its value, decay, operational significance against a defined SLA (service level agreement). This **timeliness** DQ check can be utilized to decrease data value decay rate and optimize the policies of data movement timeline. In an organization complex logic is usually segregated into simpler logic across multiple processes. **Reasonableness** DQ checks on such complex logic yielding to a logical result within a specific range of values or static interrelationships (aggregated business rules) may be validated to discover complicated but crucial business processes and outliers of the data, its drift from BAU (business as usual) expectations, and may provide possible exceptions eventually resulting into data issues. This check may be a simple generic aggregation rule engulfed by large chunk of data or it can be a complicated logic on a group of attributes of a transaction pertaining to the core business of the organization. This DQ check requires high degree of business knowledge and acumen. Discovery of reasonableness issues may aid for policy and strategy changes by either business or data governance or both. **Conformity** checks and **integrity checks** need not covered in all business needs, it's strictly under the database architecture's discretion. There are many places in the data movement where DQ checks may not be required. For instance, DQ check for completeness and precision on not–null columns is redundant for the data sourced from database. Similarly, data should be validated for its accuracy with respect to time when the data is stitched across disparate sources. However, that is a business rule and should not be in the DQ scope. Regretfully, from a software development perspective, Data Quality is often seen as a non functional requirement. And as such, key data quality checks/processes are not factored into the final software solution. Within Healthcare, wearable technologies or Body Area Networks, generate large volumes of data. The level of detail required to ensure data quality is extremely high and is often under estimated. This is also true for the vast majority of health apps, EHRs and other health related software solutions. However, some open source tools exist that examine data quality. The primary reason for this, stems from the extra cost involved is added a higher degree of rigor within the software architecture.

Objective; (ii) ;To learn the different process and metrics governing the data quality management issues.

Understanding Data Quality Management; Today, more than ever, organizations realize the importance of data quality. By ensuring that quality data is stored in your data warehouse or business intelligence application, you also ensure the quality of information for dependent applications and analytics. Oracle Warehouse Builder offers a set of features that assist you in creating data systems that provide high quality information to your business users. You can implement a quality process that assesses, designs, transforms, and monitors quality. Within these phases, you will use specific functionality from Warehouse Builder to create improved quality information.

About the Data Quality Management Process; Quality data is crucial to decision-making and planning. The aim of building a data warehouse is to have an integrated, single source of data that can be used to make business decisions. Since the data is usually sourced from a number of disparate systems, it is important to ensure that the data is standardized and cleansed before loading into the data warehouse. Warehouse Builder provides functionality that enables you to effectively manage data quality by assessing, transforming, and monitoring your data. The benefits of using Warehouse Builder for data management are as follows: Provides an end-to-end data quality solution. Enables you to include data quality and data profiling as an integral part of your data integration process. Stores metadata regarding the quality of your data alongside your data. Both the data and the metadata are stored in the design repository. Automatically generates the mappings that you can use to correct data. These mappings are based on the business rules that you choose to apply to your data. and decision you make on how to correct data.

Phases in the Data Quality Life Cycle

Ensuring data quality involves the following phases:

- Quality Assessment
- Quality Design
- Quality Transformation
- Quality Monitoring

Providing Quality Information

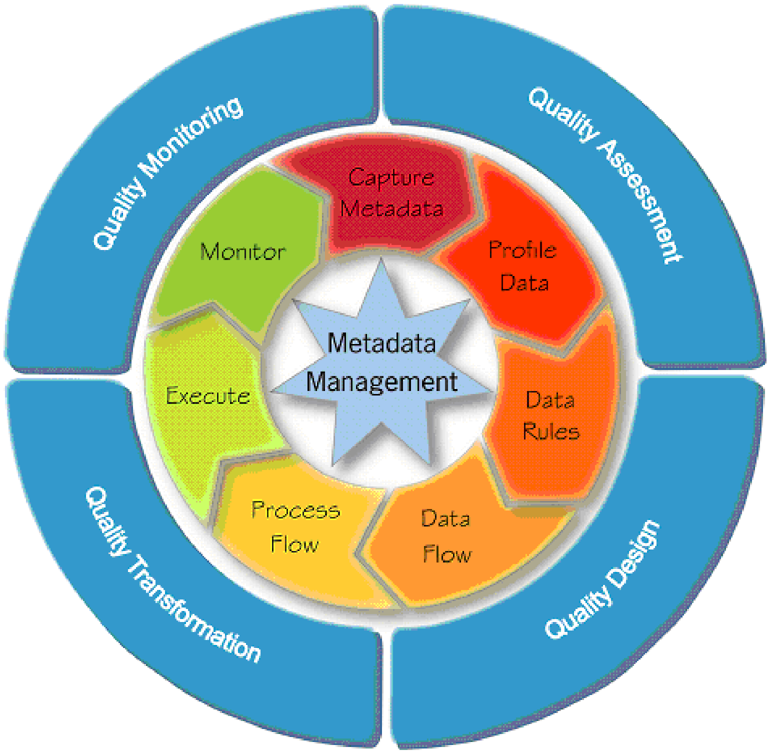

Figure: 1; Description of Phases Involved in Providing Quality Information; Source: Wikipedia.org

Quality Assessment; In the quality assessment phase, you determine the quality of the source data. The first step in this phase is to load the source data, which could be stored in different sources, into Warehouse Builder. You can import metadata and data from both Oracle and non-Oracle sources. After you load the source data, you use data profiling to assess its quality. Data profiling is the process of uncovering data anomalies, inconsistencies, and redundancies by analyzing the content, structure, and relationships within the data. The analysis and data discovery techniques form the basis for data monitoring. **Quality Design;** The quality design phase consists designing your quality processes. You can specify the legal data within a data object or legal relationships between data objects using data rules. You also correct and augment your data. You can use data quality operators to correct and augment data. As part of the quality design phase, you also design the transformations that ensure data quality. These transformations could be mappings that are generated by Warehouse Builder as a result of data profiling or mappings you create. **Quality Transformation;** The quality transformation phase consists of running the correction mappings that are used to correct the source data. **Quality Monitoring;** Data monitoring is the process of examining your data over time and alerting you when the data violates any business rules that are set. **About Data Profiling;** Data profiling is the first step for any organization to improve information quality and provide better decisions. It is a robust data analysis method available in Warehouse Builder that you can use to discover and measure defects in your data before you start working with it. Because of its integration with the ETL features in Warehouse Builder and other data quality features, such as data rules and built-in cleansing algorithms, you can also generate data cleansing and schema correction. This enables you to automatically correct any inconsistencies, redundancies, and inaccuracies in both the data and metadata. Data profiling enables you to discover many important things about your data. Some common findings include the following: A domain of valid product codes. A range of product discounts. Columns that hold the pattern of an e-mail address. A one-to-many relationship between columns. Anomalies and outliers within columns. Relations between tables even if they are not documented in the database.

Uses of Data Profiling

Using the data profiling functionality in Warehouse Builder enables you to: Profile data from any source or combination of sources that Warehouse Builder can access. Explore data profiling results in tabular or graphical format. Drill down into the actual data related to any profiling result. Derive data rules, either manually or automatically, based on the data profiling results. Attach any data rule to a target object and select an action to perform if the rule fails. Create a data auditor from a data rule to continue monitoring the quality of data being loaded into an object. Derive quality indices such as six-sigma valuations. Profile or test any data rules you want to verify before putting in place.

Types of Data Profiling; Following the selection of data objects, determine the aspects of your data that you want to profile and analyze. Data profiling offers three main types of analysis: attribute analysis, functional dependency, and referential analysis. You can also create custom profiling processes using data rules, allowing you to validate custom rules against the actual data and get a score of their accuracy.

Attribute Analysis; Attribute analysis seeks to discover both general and detailed information about the structure and content of data stored within a given column or attribute. Attribute analysis looks for information about patterns, domains, data types, and unique values. Pattern analysis attempts to discover patterns and common types of records by analyzing the string of data stored in the attribute. It identifies the percentages of your data that comply with a certain regular expression format pattern found in the attribute. Using these pattern results, you can create data rules and constraints to help clean up current data problems. Some commonly identified patterns include dates, e-mail addresses, phone numbers, and social security numbers. Domain analysis identifies a domain or set of commonly used values within the attribute by capturing the most frequently occurring values. For example, the Status column in the Customers table is profiled and the results reveal that 90% of the values are among the following: "MARRIED", "SINGLE", "DIVORCED". Further analysis and drilling down into the data reveal that the other 10% contains misspelled versions of these words with few exceptions. Configuration of the profiling determines when something is qualified as a domain, so review the configuration before accepting domain values. You can then let Warehouse Builder derive a rule that requires the data stored in this attribute to be one of the three values that were qualified as a domain. Data type analysis enables you to discover information about the data types found in the attribute. This type of analysis reveals metrics such as minimum and maximum character length values as well as scale and precision ranges. In some cases, the database column is of data type VARCHAR2, but the values in this column are all numbers. Then you may want to ensure that you only load numbers. Using data type analysis, you can have Warehouse Builder derive a rule that requires all data stored within an attribute to be of the same data type. Unique key analysis provides information to assist you in determining whether or not an attribute is a unique key. It does this by looking at the percentages of distinct values that occur in the attribute. You might determine that attributes with a minimum of 70% distinct values should be flagged for unique key analysis. For example, using unique key analysis you could discover that 95% of the values in the EMP_ID column are unique. Further analysis of the other 5% reveals that most of these values are either duplicates or nulls. You could then derive a rule that requires that all entries into the EMP_ID column be unique and not null.

Functional Dependency; Functional dependency analysis reveals information about column relationships. This enables you to search for things such as one attribute determining another attribute within an object. Referential analysis attempts to detect aspects of your data objects that refer to other objects. The purpose behind this type of analysis is to provide insight into how the object you are profiling is related or connected to other objects. Because you are comparing two objects in this type of analysis, one is often referred to as the parent object and the other as the child object. Some of the common things detected include orphans, childless objects, redundant objects, and joins. Orphans are values that are found in the child object, but not found in the parent object. Childless objects are values that are found in the parent object, but not found in the child object. Redundant attributes are values that exist in both the parent and child objects.

Data Rule Profiling; In addition to attribute analysis, functional dependency, and referential analysis, Warehouse Builder offers data rule profiling. Data rule profiling enables you to create rules to search for profile parameters within or between objects. This is very powerful as it enables you to validate rules that apparently exist and are defined by the business users. By creating a data rule, and then profiling with this rule you can verify if the data actually complies with the rule, and whether or not the rule needs amending or the data needs cleansing.

How to Perform Data Profiling; Data profiling is, by definition, a resource-intensive process that requires forethought and planning. It analyzes data and columns and performs much iteration to detect defects and anomalies in your data. So it warrants at least some forethought and planning in order to be as effective as possible. Before beginning data profiling, you should first identify the data objects that you want to target. Instead of profiling everything, choose objects that are deemed crucial. You should not select an entire source system for profiling at the same time. Not only is it a waste of resources, but it is also often unnecessary. Select areas of your data where quality is essential and has the largest fiscal impact. For example, you have a data source that contains five tables: Customers, Regions, Orders, Products, and Promotions. You decide that the two most important tables with respect to data quality are Customers and Orders. The Customers table is known to contain many duplicate and erroneous entries that cost your company money on wasted marketing efforts. The Orders table is known to contain data about orders in an incorrect format. In this case, you would select only these two tables for data profiling. After you have chosen the object you want to profile, use the following steps to guide you through the profiling process: **Import or Select the Metadata;** Data profiling requires the profiled objects to be present in the project in which you are performing data profiling. Ensure that these objects are either imported into this project or created in it. Also ensure that the data is loaded into the objects. Having the data loaded is essential to data profiling. Also, because data profiling uses mappings to run the profiling, you must ensure that all locations that you are using are registered. Data profiling attempts to register your locations. If, for some reason, data profiling cannot register your locations, you will need to explicitly register the locations before you begin profiling.**Create a Data Profile;** After your system is set up, you can create a data profile. A data profile is a metadata object in the Warehouse Builder repository and you create in the navigation tree. It contains the definitions and settings necessary for profiling objects. It includes the set of data objects you want profiled, the settings controlling the profiling operations, the results returned after you profile the data, and correction information (if you decide to use these corrections). **Profile the Data;** After you have created a data profile, you can open it in the Data Profile Editor to profile the data or review profile results from a previous run.

Data profiling is achieved by performing deep scans of the selected objects. This can be a time-consuming process, depending on the number of objects and type of profiling you are running. However, profiling is run as an asynchronous job, and the client can be closed during this process. You will see the job running in the job monitor and Warehouse Builder prompts you when the job is complete. The results are generated and can be viewed from the Data Profile Editor as soon as they are available. You can, and should, configure the profile before running it if there are specific types of analysis you do, or do not, want to run. Configuration of the profile and its objects is possible at the following levels: the entire profile (all the objects it contains),an individual object (for example, a table),a single column in a table. For example, if you know you only have one problematic column in a table and you already know that most of the records should conform to values within a certain domain, then you can focus your profiling resources on domain discovery and analysis. By narrowing down the type of profiling necessary, you use fewer resources and obtain the results faster. **View Profile Results and Derive Data Rules;** The profiling results contain a variety of analytical and statistical information about the data profiled. You can immediately drill down into anomalies and view the data that caused them. You can then determine what data must be corrected. Based on your decisions, you can derive data rules. Data rules are used to ensure that only values compliant with the data rules are allowed within a data object. Data rules will form the basis for correcting or removing data if you decide to cleanse the data. You can also use data rules to report on non-compliant data. After you have derived data rules from the profiling results, you can create the schema and mapping corrections. The schema correction creates scripts that can be used to create a corrected set of source data objects with the derived data rules applied. The mapping correction creates new correction mappings to take your data from the source objects and load them into new objects.

Define and Edit Data Rules Manually; Data rules can be derived or manually created. Before and after you have created the corrections, you can define additional data rules manually.

Objective (iii); To understand the latest technology support and its application in upkeep of data quality.

Generate, Deploy, and Execute; Finally, you can generate, deploy, and execute the correction mappings and data rules. After you run the correction mappings with the data rules, your data is corrected. The derived data rules remain attached to the objects in the corrected schema for optional use in data monitors.**Six Sigma;** Warehouse Builder provides Six Sigma results embedded within the other data profiling results to provide a standardized approach to data quality. **What is Six Sigma?**Six Sigma is a methodology that attempts to standardize the concept of quality in business processes. It achieves this goal by statistically analyzing the performance of business processes. The goal of Six Sigma is to improve the performance of these processes by identifying the defects, understanding them, and eliminating the variables that cause these defects. Six Sigma metrics give a quantitative number for the number of defects in each 1,000,000 opportunities. The term "opportunities" can be interpreted as the number of records. The perfect score is 6.0. The score of 6.0 is achieved when there are only 3.4 defects in each 1,000,000 opportunities. The score is calculated using the following formula: Defects Per Million Opportunities (DPMO) = (Total Defects / Total Opportunities) * 1,000,000. Defects (%) = (Total Defects / Total Opportunities)* 100%. Yield (%) = 100 - %Defects. Process Sigma = NORMSINV(1-((Total Defects) / (Total Opportunities))) + 1.5where NORMSINV is the inverse of the standard normal cumulative distribution.

Six Sigma Metrics for Data Profiling; Six Sigma metrics are also provided for data profiling in Warehouse Builder. When you perform data profiling, the number of defects and anomalies discovered are shown as Six Sigma metrics. For example, if data profiling finds that a table has a row relationship with a second table, the number of records in the first table that do not adhere to this row-relationship can be described using the Six Sigma metric. Six Sigma metrics are calculated for the following measures in the Data Profile Editor:

- **Aggregation:** For each column, the number of null values (defects) to the total number of rows in the table (opportunities).
- **Data Types:** For each column, the number of values that do not comply with the documented data type (defects) to the total number of rows in the table (opportunities).
- **Data Types:** For each column, the number of values that do not comply with the documented length (defects) to the total number of rows in the table (opportunities).
- **Data Types:** For each column, the number of values that do not comply with the documented scale (defects) to the total number of rows in the table (opportunities).

- **Data Types:** For each column, the number of values that do not comply with the documented precision (defects) to the total number of rows in the table (opportunities).
- **Patterns:** For each column, the number of values that do not comply with the common format (defects) to the total number of rows in the table (opportunities).
- **Domains:** For each column, the number of values that do not comply with the documented domain (defects) to the total number of rows in the table (opportunities).
- **Referential:** For each relationship, the number of values that do not comply with the documented foreign key (defects) to the total number of rows in the table (opportunities).
- **Referential:** For each column, the number of values that are redundant (defects) to the total number of rows in the table (opportunities).
- **Unique Key:** For each unique key, the number of values that do not comply with the documented unique key (defects) to the total number of rows in the table (opportunities).
- **Unique Key:** For each foreign key, the number of rows that are childless (defects) to the total number of rows in the table (opportunities).
- **Data Rule:** For each data rule applied to the data profile, the number of rows that fail the data rule to the number of rows in the table.

About Data Quality; Warehouse Builder enables you to automatically create correction mappings based on the results of data profiling. On top of these automated corrections that make use of the underlying Warehouse Builder architecture for data quality, you can create your own data quality mappings. Warehouse Builder provides functionality that enables you to ensure data quality.

About the Match-Merge Operator

Match-Merge is a data quality operator that identifies matching records and merges them into a single record. Master data management working on various systems will make use of this operator to ensure that records are created and matched with a master record. You define the business rules that the Match-Merge operator uses to identify records that refer to the same data.

About the Name and Address Operator

Warehouse Builder enables you to perform name and address cleansing on data using the Name and Address operator. The Name and Address operator identifies and corrects errors and inconsistencies in name and address source data by comparing input data to the data libraries supplied by third-party name and address cleansing software vendors. You can purchase the data libraries directly from these vendors. The errors and inconsistencies corrected by the Name and Address operator include variations in address formats, use of abbreviations, misspellings, outdated information, inconsistent data, and transposed names. The operator fixes these errors and inconsistencies by: Parsing the name and address input data into individual elements. Standardizing name and address data, using standardized versions of nicknames and business names and standard abbreviations of address components, as approved by the postal service of the appropriate country. Standardized versions of names and addresses facilitate matching and house holding, and ultimately help you obtain a single view of your customer. Correcting address information such as street names and city names. Filtering out incorrect or undeliverable addresses can lead to savings on marketing campaigns. Augmenting names and addresses with additional data such as gender, ZIP+4, country code, apartment identification, or business and consumer identification. You can use this and other augmented address information, such as census geo coding, for marketing campaigns that are based on geographical location. Augmenting addresses with geographic information facilitates geography-specific marketing initiatives, such as marketing only to customers in large metropolitan areas (for example, within an *n*-mile radius from large cities); marketing only to customers served by a company's stores (within an *x*-mile radius from these stores). Oracle Spatial, an option with Oracle Database, and Oracle Locator, packaged with Oracle Database, are two products that you can use with this feature. The Name and Address operator also enables you to generate postal reports for countries that support address correction and postal matching. Postal reports often qualify you for mailing discounts. For more information, see "About Postal Reporting".

Example: Correcting Address Information; This example follows a record through a mapping using the Name and Address operator. This mapping also uses a Splitter operator to demonstrate a highly recommended data quality error handling technique. **Example Input;** The data contains a nickname, a last name, and part of a mailing address, but it lacks the customer's full name, complete street address, and the state in which he lives. The data also lacks geographic information such as latitude and longitude, which can be used to calculate distances for truckload shipping. **Example Steps;** This example uses a mapping with a Name and Address operator to cleanse name and address records, followed by a Splitter operator to load the records into separate targets depending on whether they were successfully parsed. This section explains the general steps required to design such a mapping. **Example: Australia Post AMAS Certification;** The Address Matching Approval System (AMAS) was developed by Australia Post to improve the quality of addressing. It provides a standard by which to test and measure the ability of address-matching software to:

- Correct and match addresses against the Postal Address File (PAF)
- Append a unique Delivery Point Identifier (DPID) to each address record, which is a step toward bar coding mail.

AMAS allows companies to develop address matching software which:

- Prepares addresses for barcode creation
- Ensures quality addressing
- Enables qualification for discounts on Pre-Sort letter lodgments

Pre-Sort Letter Service prices are conditional upon customers using AMAS Approved Software with Delivery Point Identifiers (DPIDs) being current against the latest version of the PAF. A declaration that the mail was prepared appropriately must be made when using the Presort Lodgment Document, available from post offices.

About Data Rules

Data rules are definitions for valid data values and relationships that can be created in Warehouse Builder. They determine legal data within a table or legal relationships between tables. Data rules help ensure data quality. They can be applied to tables, views, dimensions, cubes, materialized views, and external tables. Data rules are used in many situations including data profiling, data and schema cleansing, and data auditing. The metadata for a data rule is stored in the repository. To use a data rule, you apply the data rule to a data object. For example, you create a data rule called gender rule that specifies that valid values are 'M' and 'F'. You can apply this data rule to the emp_gender column of the Employees table. Applying the data rule ensures that the values stored for the emp_gender column are either 'M' or 'F'. You can view the details of the data rule bindings on the Data Rule tab of the Data Object Editor for the Employees table. There are two ways to create a data rule. A data rule can be derived from the results of data profiling, or it can be created using the Data Rule Wizard. For more information about data rules, see "Using Data Rules".

About Quality Monitoring; Quality monitoring builds on your initial data profiling and data quality initiatives. It enables you to monitor the quality of your data over time. You can define the business rules to which your data should adhere. To monitor data using Warehouse Builder you need to create data auditors. Data auditors ensure that your data complies with the business rules you defined. You can define the business rules that your data should adhere to using a feature called data rules.

Objective (iv) :To know the application developments happening around data quality management

About Data Auditors; Data auditors are processes that validate data against a set of data rules to determine which records comply and which do not. Data auditors gather statistical metrics on how well the data in a system complies with a rule, and they can off-load defective data into auditing and error tables. Data auditors have thresholds that allow you to create logic based on the fact that too many non-compliant records can divert the process flow into an error or notification stream. Based on this threshold, the process can choose actions. Also the audit result can be captured and stores for analysis purposes. Data auditors can be deployed and executed ad-hoc, but they are typically run to monitor the quality of the data in an operational environment like a data warehouse or ERP system and, therefore, can be added to a process flow and scheduled. When executed, the data auditor sets several output values. One of these output values is called the audit result. If the audit result is 0, then there were no errors. If the audit result is 1, at least one error occurred. Data auditors also set the actual measured values such as Error Percent and Six Sigma values. For more information about using data auditors, see "Using Data Auditors". Data auditors are a very important tool in ensuring data quality levels are up to the standards set by the users of the system. It also helps determine spikes in bad data allowing events to the tied to these spikes. The ultimate goal of data quality management is not to create subjective notions of what “high quality” data is, the ultimate goal is to increase return on investment (ROI) for those business segments that depend upon data. From customer relationship management (CRM), to supply chain management (SCM), to enterprise resource planning (ERP), the benefits of effective DQM can have a wide organizational reach. With quality data at their disposal, organizations can form data warehouses for the purposes of examining trends and establishing future-facing strategies. Industry wide, the profoundly positive ROI on quality data is well understood. According to recent big data surveys by Accenture, 92% of executives managing with big data are satisfied with the results, and 89% rate data as “very” or “extremely” important, as it will “revolutionize operations the same way the internet did.”As you can see, the leaders of big business clearly understand the importance of quality data to build business dashboards – it is the strategy for the present and the future. Without further adieu, let’s take a look at the five essential pillars of DQM.

Pillar #1 – The People

Technology is only as efficient as the individuals who implement it. We may function within a technologically advanced business society, but human oversight and process implementation have not yet been rendered obsolete. Therefore, there are several DQM roles that need to be filled, including:

DQM Program Manager: The program manager role should be filled by a high level leader who accepts the responsibility of general oversight for business intelligence initiatives. He/she should also oversee the management of the daily activities involving data scope, project budget and program implementation. The program manager should lead the vision for quality data and ROI. **Organization Change Manager:** The change manager does exactly what the title suggests, organize. He/she assists the organization by providing understanding and insight into advanced data technology solutions. As quality issues are often highlighted through BI dashboard software, the change manager plays an important role in the visualization of data quality.**Business/Data Analyst:** The business analyst is all about the “meat and potatoes” of the business. This individual defines the data quality needs from an organizational perspective. These needs are then quantified into data models for acquisition and delivery. This person (or group of individuals) ensures that the theory behind data quality is communicated to the development team.

Pillar #2 – Data Profiling

Data profiling is an essential process in the data quality management lifecycle. It is the process of reviewing data in detail, comparing and contrasting the data to its own metadata, running statistical models and then reporting the quality of the data.This process is initiated for the purpose of developing insight into existing data, with the goal of comparing it to data quality goals. It helps businesses develop a starting point in the DQM process, and sets the standard for how to improve data quality. Imperative to this step are the data quality metrics of complete and accurate data. **Accurate Data:** looking for disproportionate numbers. **Complete Date:** defining data body and ensuring that all data points are whole.

Pillar #3 – Defining Data Quality

The third pillar of data quality management is quality itself. The information obtained in the second pillar is expanded upon to ensure that compromised data points are identified. “Quality rules” should be created and defined based on business goals and requirements. These are the business/technical rules with which data must comply in order to be considered viable. Business requirements are likely to take a front seat in this pillar, as critical data elements are to depend upon industry. The development of quality rules is essential to the success of any DQM process, as the rules will detect and prevent compromised data from infecting the health of the whole set. Much like antibodies detecting and correcting viruses within our bodies, data quality rules will correct inconsistencies among valuable data. When teamed together with business intelligence solutions, data quality rules can be key in predicting trends and reporting analytics.

Pillar #4 – Data Reporting

DQM reporting is the process of removing and recording all compromising/data quality exceptions. This pillar should be designed to occur as a natural process of data rule enforcement. Once exceptions have been identified and captured, they should be aggregated to modeling’s that tally the number of data points in breach of the rules. The captured data points should be modeled and defined based on specific characteristics (e.g., by rule, by date, by source, etc.). Once this data is tallied, it can be applied to a business intelligence solution to report on the state of data quality and the exceptions that exist within a dashboard. If possible, automated and "on-demand" technology solutions should be implemented as well, so dashboard insights can appear in real time. Reporting and monitoring are the cruces of data quality management ROI, as they provide visibility into the state of data at any moment in real time. By allowing businesses to identify the location and domiciles of data exceptions, teams of data specialists can begin to strategize remediation processes. Knowledge of where to begin engaging in proactive data adjustments will help businesses move one step closer to recovering their part of the $600 billion lost each year to low-quality data.

Pillar #5 – Data Repair

Date repair is the two-step process of determining how best data should be remediated, and the most efficient manner in which to implement the change. The most important aspect of data remediation is the performance of a "root cause" examination to determine why, where and how the data defect originated. Once this examination has been implemented, the remediation plan should begin. Data processes that depended upon the previously defected data will likely need to be re-initiated, especially if their functioning was at risk or compromised by the defected data. These processes could include reports, campaigns or financial documentation. It is also here, in the remediation phase, where data quality rules should again be reviewed. The review process will help determine if the rules need to be adjusted or updated, and it will help begin the process of data evolution.Once data is deemed of high quality, critical business processes and functions should run more efficiently and accurately, with a higher ROI and lower costs.

What is Data Quality Metrics?

Data quality metrics are key in determining the overall health of an organization. As we have demonstrated, low-quality data can impact productivity, bottom line and overall ROI. In order for organizations to follow the general pattern of the 5 Pillars of DQM, data metrics must be of a high quality, and clearly defined. While data analysis can be quite complex, there are a few basic measurements that all key DQM stakeholders should be aware of. They include: **Accuracy:** This metric makes reference to business transactions or status changes as they happen in real-time.

Accuracy should be measured through source documentation (i.e., from the business interactions), but if not available, then through confirmation techniques of an independent nature. It will indicate whether date is void of significant errors. **Completeness:** As a data quality metric, completeness means determining whether or not each data entry is a “full” data entry. All available data entry fields must be complete, and sets of data records should not be missing any pertinent information. Completeness will indicate if there is enough information to draw conclusions. **Integrity:** Also known as data validation, integrity refers to the structural testing of data to ensure that the data complies with procedures. This means there are no unintended data errors, and it corresponds to its appropriate designation (e.g., date, month and year).

Findings and Conclusions:

Data quality management is a process that wears many hats and comes in various forms. One of those forms can be to help your organization win back its share of the $600 billion American businesses annually lose due to low-quality data. Big data resources are rich in information, and ROI is high. Follow and customize the 5 Pillars to meet your businesses, and you will be one step further on the path to digital age (and big data) innovation.

References:

1. Redman, Thomas C. (30 December 2013). Data Driven: Profiting from Your Most Important Business Asset. Harvard Business Press. ISBN 978-1-4221-6364-1.

2. "What is data scrubbing (data cleansing)? - Definition from WhatIs.com".

3. "Data Quality: High-impact Strategies - What You Need to Know: Definitions, Adoptions, Impact, Benefits, Maturity, Vendors". Retrieved 5 February 2013.

4. "IAIDQ--glossary".

5. Government of British Columbia

6. "ISTA Con - Innovations in Software Technologies and Automation.".

7. Anonymous (23 December 2014). "Data Quality".

8. "Liability and Leverage - A Case for Data Quality".

9. "Address Management for Mail-Order and Retail".

10. http://ribbs.usps.gov/move_update/documents/tech_guides/PUB363.pdf

11. E. Curry, A. Freitas, and S. O'Riáin, "The Role of Community-Driven Data Curation for Enterprises," in Linking Enterprise Data, D. Wood, Ed. Boston, MA: Springer US, 2010, pp. 25-47.

12. "ISO/TS 8000-1:2011 Data quality -- Part 1: Overview". International Organization for Standardization. Retrieved 8 December 2016.

13. "Can you trust the quality of your data?".

14. "What is Data Cleansing? - Experian Data Quality". 13 February 2015.

15. "Lecture 23 Data Quality Concepts Tutorial – Data Warehousing". Watch Free Video Training Online. Retrieved 8 December 2016.

16. O'donoghue, John, and John Herbert. "Data management within mHealth environments: Patient sensors, mobile devices, and databases." Journal of Data and Information Quality (JDIQ) 4.1 (2012): 5.

17. Huser, Vojtech; DeFalco, Frank J; Schuemie, Martijn; Ryan, Patrick B; Shang, Ning; Velez, Mark; Park, Rae Woong; Boyce, Richard D; Duke, Jon; Khare, Ritu; Utidjian, Levon; Bailey, Charles (30 November 2016). "Multisite Evaluation of a Data Quality Tool for Patient-Level Clinical Datasets". eGEMs (Generating Evidence & Methods to improve patient outcomes). 4 (1). doi:10.13063/2327-9214.1239.

18. "IQ International - the International Association for Information and Data Quality". IQ International website. Retrieved 2016-08-05.

19. Baškarada, S; Koronios, A (2014). "A Critical Success Factors Framework for Information Quality Management". Information Systems Management. 31 (4): 1–20.doi:10.1080/10580530.2014.958023.

20. Baamann, Katharina, "Data Quality Aspects of RevenuAssurance", Article

21. Eckerson, W. (2002) "Data Warehousing Special Report: Data quality and the bottom line", Article

22. Ivanov, K. (1972) "Quality-control of information: On the concept of accuracy of information in data banks and in management information systems". The University of Stockholm and The Royal Institute of Technology. Doctoral dissertation.

23. Hansen, M. (1991) Zero Defect Data, MIT. Masters thesis [1]

24. Kahn, B., Strong, D., Wang, R. (2002) "Information Quality Benchmarks: Product and Service Performance," Communications of the ACM, April 2002. pp. 184–192. Article

25. Price, R. and Shanks, G. (2004) A Semiotic Information Quality Framework, Proc. IFIP International Conference on Decision Support Systems (DSS2004): Decision Support in an Uncertain and Complex World, Prato. Article

26. Redman, T. C. (2008) Data Driven: Profiting From Our Most Important Business Asset

27. Wand, Y. and Wang, R. (1996) "Anchoring Data Quality Dimensions in Ontological Foundations," Communications of the ACM, November 1996. pp. 86–95. Article

28. Wang, R., Kon, H. & Madnick, S. (1993), Data Quality Requirements Analysis and Modelling, Ninth International Conference of Data Engineering, Vienna, Austria. Article

29. Fournel Michel, Accroitre la qualité et la valeur des données de vos clients, éditions Publibook, 2007. ISBN 978-2-7483-3847-8.

30. Daniel F., Casati F., Palpanas T., Chayka O., Cappiello C. (2008) "Enabling Better Decisions through Quality-aware Reports", International Conference on Information Quality (ICIQ), MIT. Article

31. Jack E. Olson (2003), "Data Quality: The Accuracy dimension"Morgan Kaufmann Publishers

32. Woodall P., Oberhofer M., and Borek A. (2014), "A Classification of Data Quality Assessment and Improvement Methods". International Journal of Information Quality 3 (4), 298–321. doi:10.1504/ijiq.2014.068656.

33. Woodall, P., Borek, A., and Parlikad, A. (2013), "Data Quality Assessment: The Hybrid Approach." Information & Management 50 (7), 369–382.

34. Applebaum, R., & Phillips, P. (August 1990). Assuring quality of in-home care: The other challenge for long-term care. The Gerontologist, 30(4), 444-450.

35. Applebaum, R., Mollica, R., & Tilly, J. (Winter 1997-Winter 1998). Assuring homecare quality: A case study of state strategies. Generations, 21(4), 57-63.

36. Applebaum, R., Regan, S., & Woodruff, L. (1993). Assuring the quality of in-home suuportive services: An evolving challenge. Home Health Care Services Quartlerly, 14(2/3).

37. Benjamin, A. E. (November 2001-December 2001). Consumer-directed services at home: A new model for persons with disabilities.

38. Health Affairs, 20(6), 80-95. Brook, R. H., & McGlynn, E. A. (September 1996). Measuring quality of care. The New England Journal of Medicine, 335(13), 966-970.

39. Brook, R. H., McGlynn, E. A., & Shekelle, P. G. (2000). Defining an d measuring quality of care: a perspective from US reserachers. International Journal for Quality in Health Care, 12(4), 281-295. Campbell, J. (1998). Assessment of outcomes: Consumerism, outcomes, and satisfaction: A review of the literature In R. W. Manderscheid, & M. J. Henderson (Eds.), Mental health, United States, 1998 (pp. 11-28).

40. Rockville, MD: U.S. Department of Health and Human Services, Substance Abuse and Mental Health Services Administration, Center for Mental Health Services.

41. Campbell, S. M., Roland, M. O., & Buetow, S. A. (2000). Defining quality of care. Social Science & Medicine, 51, 1611-1625.

42. Capitman, J., Abrahams, R., & Ritter, G. (1997). Measuring the adequacy of home care for frail elders. The Gerontologist, 37(3), 303-313. Centers for Medicare and Medicaid Servic es. (2002). Outcome-based quality improvement (OBQI): Implementation manual. Baltimore, MD: HCFA.

43. Chasin, M. R. (October 1996). Improving the quality of care. The New England Journal of Medicine, 335(14), 1060-1063.

44. Demakis, J. G., McQuenn, L., Kizer, K. W., & Feussner, J. R. (2000). Quality enhancement research initiative (QUERI): A collaboration between research and clinical practice . Medical Care, 38(6), I17-I25. Edmund S. Muskie School of Public Service - 1/10/02 40

45. Donabedian, A. (1980). The definition of quality and approaches to its assessment: Explorations in quality assessment and monitoring. Ann Arbor, MI: Health Administration Press.

46. Donabedian, A. (1996). Explorations in quality assessment and monitoring. Volume I: The definition of quality and approaches ti its assessment. Ann Arbor, MI: Health Administration Press.

47. Feder, J., Komisar, H. L., & Niefeld, M. (May 2000-June 2000). Long-term care in the united states: An overview. Health Affairs, 19(3), 40-56. Foundation for Accountability . (1997). Reporting Quality Information to Consumers. Portland, OR: The Foundation for Accountability.

48. Fries, B. (December 21 , 2001). Personal communication. University of Michigan.

49. Geron, S. M., Smith, K., Tennstedt, S., Jette, A., Chasssler, D., & Kasten, L. (2000). The home care satisfaction measure: A client-centered approach to assesing the satisfaction of frail older adults with home care services. Journal of Gerontology, Social Sciences, 55B(5), S259-S270.

50. Gold, M., Sparer, M., & Chu, K. (Fall 1996). Medicaid Managed Care: Lessons from five States. Health Affairs, 15(3), 154-165.

51. Greene, A., Hawes, C., Wood, M., & Woodsong, C. (Winter 1997-Winter 1998). How do family members define quality in assisted living facilities? Generations, 21(4), 34-36.

52. Hawes, C., Greene, A., Wood, M., & Woodsong, C. (1997). Family members' views: What is quality in assisted living facilities providing care to people with dementia? Washington, DC: Alzheimer's Association. Hawes, C., Mor, V., Phillips, C. D., Fries, B. E., Morris, J. N., Steele-Friedlob, E. et al. (August 1997). The OBRA-87 nursing home regulations and implementation of the Resident Assessment Instrument: Effects on process quality. Journal of the American Geriatrics Society, 45(8), 977-985. Hawes, C., Mor, V., Wildfire, J., Iannocchione, V., Lux, L., Green, R., Greene, A., Wilcox, V., Spore, D., &

53. Trends and issues in the Medicaid 1915 (c) waiver program. Health Care Finance Review, 20(4), 139-160.

54. Mollica, R. (2000). State assisted living policy: 2000. Portland, ME: National Academy for State Health Policy.

55. Moos, R. H., & Lemke, S. (1980). Assessing the physical and architectural features of sheltered care settings. Journal of Gerontology, 35(4), 571-583. Moos, R. H., &

56. Lemke, S. (1996). Evaluating residental facilities: The multiphasic environmental assessment procedure. Thousand Oaks, California: sage Publications, Inc. Morris, J. N., Bernabei, R., Ikegami, N., Gilgen, R. et al. (1999).

57. RAI-Home Care (RAI-HC) assessment manual for version 2.0. Washington, D.C.: interRAI Corporation. Morris, J. N., Murphy, K., & Nonemaker, S. (1995). Resident assessment instrument (RAI): User's manual. Health Care Financing Administration.

58. Mukamel, D. B. (1997). Risk-adjusted outcome measures and quality of care in nursing homes. Medical Care, 35(4), 367-385.

59. Madsen, R. W., Popejoy Lori L., Hicks, L. L. et al. (2000). Initial field testing of an instrument to measure: Observable indicators of nursing home care quality. Journal of Nursing Care Quality, 14(3), 1-12.

60. Rantz, M. J., Mehr, D. R., Popejoy Lori L., Zwygart-Stauffacher, M., Hicks, L. L., Grando, V. T. et al. (1998). Nursing home care quality: A multidimensional theoretical model. Journal of Nursing Care Quality, 12(3), 30-46.

61. Riley, P., Fortsinsky, R. H., & Coburn, A. F. (Summer 1992). Developing consumer-centered quality assurance strategies for home care. Journal of Case Management, 1(2), 39-48. Robison, J. (August , 2001). Project Director, Personal communication.

62. Sainfort, F., Ramsay, J. D., & Monato, H. Jr. (March 1995). Conceptual and methodological sources of variation in the measurement of nursing facility quality: An evaluation of 24 models and an empirical study. Medical Care Research and Review, 52(1), 60-87.

63. Tilly, J., Wiener, J., & Cuellar A. (2000). Consumer-directed home and community services programs in five countries: Policy issues for older people and goverment.

64. Washington, D.C.: The Urban Institute. United States. General Accounting Office. (1996). Medicaid long term care: State use of assessment instruments in care planning. (GAO/PEMD-96-4).

65. Washington, D.C.: GAO. Wagner, E. H., Austin, B. T., Davis, C., Hindmarsh, M., Schafer, J., & Bonomi, A. (November 2001-December 2001). Improving chronic illness care: Translating evidence into action. Health Affairs, 20(6), 64-78. Wagner, E. H., Austin, B. T., & Von Korff, M. (1996). Organizing care for patients with chronic illness. Milbank Quarterly, 74(4), 511-542.

66. Zimmerman, D. R. (Winter 1997-Winter 1998). The power of information: Using resident assessment data to assure and improve the quality of nursing home care. Generations, 21(4), 52-56. Zimmerman, D. R., Karon, S. L., Arling, G., Clark, B. R., Collins, T., Ross, R. et al. (Summer 1995). Development and testing of nursing home quality indicators. Health Care Financing