Developing a Data Quality Scorecard that Measures Data Quality in a Data Warehouse

A thesis submitted for the degree of Doctor of Philosophy

By

Aderibigbe Grillo

College of Engineering, Design and Physical Sciences

Brunel University

November 2018

Brunel
University
London

ABSTRACT

The main purpose of this thesis is to develop a data quality scorecard (DQS) that aligns the data quality needs of the Data warehouse stakeholder group with selected data quality dimensions. To comprehend the research domain, a general and systematic literature review (SLR) was carried out, after which the research scope was established. Using Design Science Research (DSR) as the methodology to structure the research, three iterations were carried out to achieve the research aim highlighted in this thesis. In the first iteration, as DSR was used as a paradigm, the artefact was build from the results of the general and systematic literature review conduct. A data quality scorecard (DQS) was conceptualised. The result of the SLR and the recommendations for designing an effective scorecard provided the input for the development of the DQS. Using a System Usability Scale (SUS), to validate the usability of the DQS, the results of the first iteration suggest that the DW stakeholders found the DQS useful. The second iteration was conducted to further evaluate the DQS through a run through in the FMCG domain and then conducting a semi-structured interview. The thematic analysis of the semi-structured interviews demonstrated that the stakeholder's participants‘ found the DQS to be transparent; an additional reporting tool; Integrates; easy to use; consistent; and increases confidence in the data. However, the timeliness data dimension was found to be redundant, necessitating a modification to the DQS. The third iteration was conducted with similar steps as the second iteration but with the modified DQS in the oil and gas domain. The results from the third iteration suggest that DQS is a useful tool that is easy to use on a daily basis. The research contributes to theory by demonstrating a novel approach to DQS design This was achieved by ensuring the design of the DQS aligns with the data quality concern areas of the DW stakeholders and the data quality dimensions. Further, this research lay a good foundation for the future by establishing a DQS model that can be used as a base for further development.

This thesis is dedicated to the Grillo Family

Table of Contents

List of Figures

List of tables

Acknowledgement

Firstly, I would like to thank God for seeing me through this journey. Also a big thank you to my family for bearing with me during the rough times and….. moods.

I would like to thank my Supervisor Dr Alan Serrano for his supervision and tremendous support, for his endless encouragement and guidance throughout the duration of this thesis. Dr Serrano‘s contributions and consistent feedback got me to the finish line. I am indeed appreciative of the level of enthusiasm over the so many years I have known him. Dr Serrano was not just my research advisor, he is a friend.

I am also grateful to the head of the department – Professor Tracy Hall for creating a terrific research atmosphere in the department of Computer Science.

I would like to say a big thank you to all academic and non-academic staff for all the support.

A big thank you also to all my fellow researchers at St. Johns…… thanks for the stimulating environment. It’s done now.

Chapter 1: Introduction

1.1 Overview

This chapter introduces the research into the data warehouse domain, with particular emphasis on data quality issues within the domain. The importance of data quality and the perception of quality was then highlighted. Afterwards, the research aims and objectives are presented. The methodology adopted to provide the structure, and the logical flow of this research is described. The layout of this thesis is then presented in section 1.5 below.

Chapter one is structured as follows: **Section 1.2** highlights the research background which includes the research problem and motivation and scope. **Section 1.3** Presents the research problem. **Section 1.4** highlights the research aim and objectives. **Section 1.5** describes the methodology adopted for this thesis while **Section 1.6** presents the thesis layout.

1.2 Research background

Data quality measurement within a data warehouse has emerged as a very important and popular area of research especially with the advent of 'Big Data' within the last few years. Big data has changed the way corporations view data and the intrinsic value they can generate from harnessing the data. Data quality dimensions enable corporations to further analyse data based on focused areas.

The data quality dimensions can be logically ordered and classified in a hierarchy to facilitate technology implementation, governance, the definition of compliance, reporting and operational processes (Batini and Scannapieco, 2016). Rules related to various dimensions can be used in varied organisational data aspects. Some dimensions are intrinsic to values that comprise a set of data. More complex data quality norms are an outcome of logical relations that exist at a record level than at the set of data followed by the level of application. From a different standpoint, there are other kinds of rules that can be used to encapsulate rule compliance of information with business policy. The expectations of data quality for analytical or operational silos of data are specified, master data expectations of data quality are organised within defined dimensions of data quality to simplify their measurement/validation and specification.

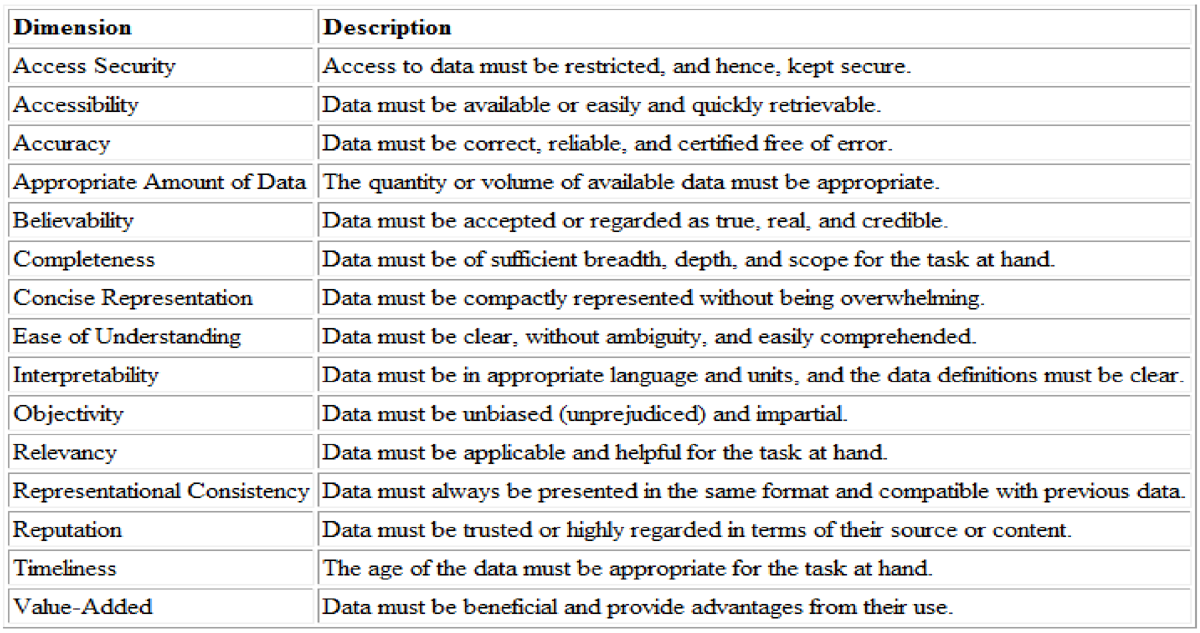
This research project looks to respond to a future where big data can be harnessed reliably without the perception of the data quality in a questionable state. Considering the quantum of big data available today, this is most likely the norm for most organisations, as a starting point. This research aims to identify the areas where data is susceptible to data quality loss at various points within the development of a data warehouse and then develop a scorecard that will measure the data quality of data at various critical points before the data is consumed by the end user.

Data Quality Dimension -DQD

Data quality, as presented in most of the literature, is a multidimensional concept (Firmani et al., 2016). The research community has identified various dimensions. Data quality can be measured according to various parameters. Previous work provides different classifications of the data quality dimensions (Batini et al., 2016; Cai and Zhu, 2015). Most frequently mentioned dimensions are accuracy, completeness, consistency, and timeliness. The choice of these dimensions is primarily based on intuitive understanding (Jarke et al., 2013) industrial experience (Liaw et al., 2013) or literature review (Khanam et al., 2016). However, the literature review shows that there is no general agreement on data quality dimensions (Lin, 2012). For example, accuracy, which in some data quality literature is included as a critical dimension. However, there is no uniquely accepted definition of what it means precisely.

Khanam et al., (2016) characterise accuracy as “the correctness of the output information.” Wang et al., (2014), describe accuracy as “the recorded value conforms to the actual value.”

The notion of data quality depends on the actual use of data. What may be considered consistent data in one case may not be sufficient in another case, for example, analysis of the unit of measure in a company may require data in units of pounds, whereas auditing requires units of measure in kilograms. This relativity of quality presents a problem. The quality of the data generated by an information system depends on the design of the system. The real use of the data is outside of the designer’s control. Thus, it is essential to provide a design-oriented definition of data quality that will reflect the intended use of the information



Dimension	Description
Access Security	Access to data must be restricted, and hence, kept secure.
Accessibility	Data must be available or easily and quickly retrievable.
Accuracy	Data must be correct, reliable, and certified free of error.
Appropriate Amount of Data	The quantity or volume of available data must be appropriate.
Believability	Data must be accepted or regarded as true, real, and credible.
Completeness	Data must be of sufficient breadth, depth, and scope for the task at hand.
Concise Representation	Data must be compactly represented without being overwhelming.
Ease of Understanding	Data must be clear, without ambiguity, and easily comprehended.
Interpretability	Data must be in appropriate language and units, and the data definitions must be clear.
Objectivity	Data must be unbiased (unprejudiced) and impartial.
Relevancy	Data must be applicable and helpful for the task at hand.
Representational Consistency	Data must always be presented in the same format and compatible with previous data.
Reputation	Data must be trusted or highly regarded in terms of their source or content.
Timeliness	The age of the data must be appropriate for the task at hand.
Value-Added	Data must be beneficial and provide advantages from their use.

Figure 1: Data Quality Dimensions and description (Batini and Scannapieco, 2016)

Batini and Scannapieco (2016) discussed how to construct specific data quality dimensions. His group first gathered 179 data quality attributes, from the data quality literature, from researchers and consumers. They used factor analysis to collapse their list of attributes into fifteen data quality dimensions shown above in figure 1.

Data warehouse Roles – The Stakeholder Groups

The data warehouse stakeholder group are responsible for the entire chain of activities of the data warehouse process. The data quality management process is handled by this group. They play a very pivotal part in the development of this thesis. Four types of stakeholders have been identified as the primary groups within the data chain of a data warehouse (Fuber, 2015). They are;

(1) Data producers: Data producer collects the raw data from multiple source systems which are essential for the input of the data warehouse. Data producer is the one who is responsible for the quality of input into the source systems. In other words, data producers are those who create or collect raw data.

(2) Data custodians: These are the group of people responsible for collecting the information from the data producers and then transforming the data into useful information for the use of data consumers by entering it into the data warehouse. Data custodians provide resources for the data consumers by collecting, entering, updating and storing the data in the data warehouse. The data custodian's primary responsibility is to design, develop and operate the data warehouse.

(3) Data Managers: Data managers are responsible for setting up the right standards and policies related to protecting and managing the day to day usage of the data warehouse. The Data manager group is responsible for managing the day-to-day activities of the data. The primary responsibility of the data manager is to ensure that data custodians are fulfilling their responsibility correctly and also to ensure the security of the entire data warehouse.

(4) Data Consumers: The data consumer is a group or individual who uses the data or information. Data consumers use the set of data for analysis, query, and reporting. In other words, data consumers are the individuals or group of people who use the data in the data warehouses for various purposes. Data consumers are associated with the processes of data utilisation, and also they may involve in additional processes like data integration and aggregation.

The data quality perspective of all stakeholder groups varies based on their requirements (Fuber, 2015). For example, the consumer group is more interested in data that is fit for use, i.e. information, while the producer group is more interested in the raw data and not necessarily the collection of the data into information.

Data Quality Challenges

Literature into data quality improvement has shown that two main approaches are available, mainly the technological approach, and the data-driven approach. The methods by which technological processes are altered can be labelled "technology-driven" approaches to improvement. Those methods used to alter the fundamental processes can be labelled "data-driven" approaches (McAfee et al., 2012).

Figure 2 below shows that there are two basic categories of improvement opportunities, those that require the alteration of fundamental processes and those that require alterations of technological processes.

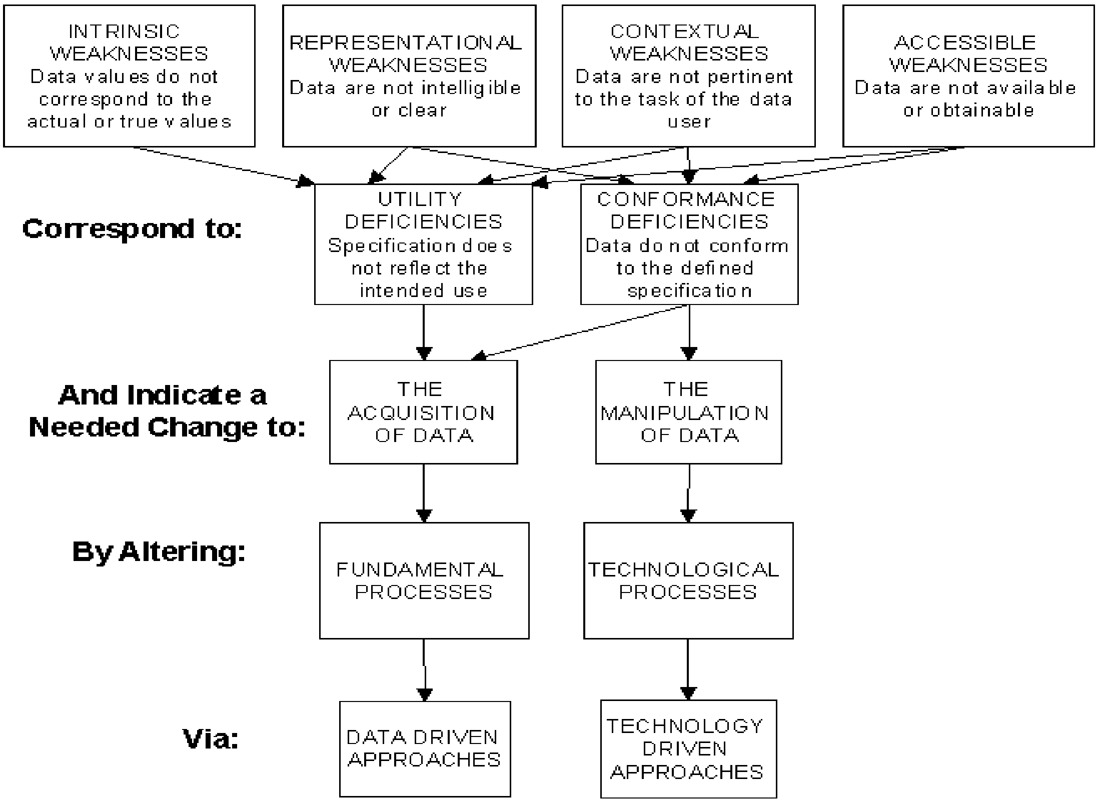

INTRINSIC WEAKNESSES Data values do not correspond to the actual or true values
REPRESENTATIONAL WEAKNESSES Data are not intelligible or clear
CONTEXTUAL WEAKNESSES Data are not pertinent to the task of the data user
ACCESSIBLE WEAKNESSES Data are not available or obtainable
Correspond to:
UTILITY DEFICIENCIES Specification does not reflect the intended use
CONFORMANCE DEFICIENCIES Data do not conform to the defined specification
And Indicate a Needed Change to:
THE ACQUISITION OF DATA
THE MANIPULATION OF DATA
By Altering:
FUNDAMENTAL PROCESSES
TECHNOLOGICAL PROCESSES
Via:
DATA DRIVEN APPROACHES
TECHNOLOGY DRIVEN APPROACHES

Figure 2: Data Quality approaches (Wang et al., 2011)

This research covers the measurement of data quality in a data warehouse. Previous studies in the data quality domain have highlighted four main challenges that will have an impact on this research project. These challenges can be summarised as follows;

(i) How can a data quality tool ensure that data is clean before being used in a data warehouse system

(ii) Awareness of the quality of data being used in the development of a data warehouse system

(iii) What are the best areas to concentrate efforts for the most significant improvement in data quality in a data warehouse system development?

(iv) What types of improvement efforts will remedy the most significant problems?

In order to overcome the limitations of existing data quality measurement/assessment tools, this thesis attempts to align the varied data quality needs of the data warehouse stakeholders with the most commonly used DQD. Subjective data quality assessments reflect the needs and experiences of stakeholders. If the stakeholders assess the quality of data as inferior, their behaviour will be influenced by this assessment (Wang et al., 2011).

Objective assessments can be task-independent or dependent. Task-independent data quality metrics reflect the state of the data without the contextual knowledge of the application software and can be applied to any data set, regardless of the tasks at hand. Task-dependent metrics, which include the organisation's business rules, government regulations, and constraints provided by the database administrator, are developed in specific application contexts.

1.3 The Research Problem

With the advent of 'Big Data' within the last decade, corporations have changed the way they view data and the intrinsic value they can generate from harnessing the data. 'Big Data' is used to describe a massive volume of both structured and unstructured data. (Gandomi and Haider, 2015)

However, 'Big Data' also arises with many challenges, such as difficulties in data capture, data storage, data analysis and data visualization. (Chen and Zhang, 2014)

Furthermore, the massive volume of data now being managed in the data warehouse has led to the misalignment of the data quality dimensions (DQD), and the varied data quality needs/concern areas of the DW stakeholders. This misalignment has led to a persistence in the problem of data quality. (Batini et al., 2008; Chennel et al., 2000; Sebastian-Coleman, 2012)

This study proposes the development of a data quality scorecard (DQS) that measures the quality of data in the data warehouse. The DQS intents at aligning the DQD with the needs/concern areas of the data warehouse stakeholders. The most commonly used DQD from the literature findings are; 1) Completeness, 2) Validity, 3) Accuracy, 4) Consistency, 5) Timeliness and, 6) Integrity. The roles involved in an effective data warehouse data quality management according to literature are (i) data producers; (2) data custodians; (3) data managers; and (4) data consumers in the data quality chain of a data warehouse.

In section 1.4 below, the aim and objectives are presented, and the methodology that will provide the structure for the research is subsequently explained.

From the literature, it can be seen that the concept of Data quality management is pivotal to the development of an efficient data warehouse.

1.4 Research Aim and Objective

Data cleansing has been discussed and numerous research about data quality exist, however, little research is found on the relevance and impact of these approaches with a specific requirement to the roles of the data warehouse stakeholder groups and the data quality dimensions. Therefore, the specific aim of this thesis is to develop a role-based scorecard to Measure Data Quality in a Data warehouse.

The scorecard will be developed using specific stakeholder requirements as the measurement parameters. This framework uses the identified stakeholders; (i) data producers; (2) data custodians; (3) data managers; and (4) data consumers in the data quality chain of a data warehouse.

The following objectives were established to achieve the aim of the research.

1. To identify and understand data quality issues, data warehouse roles, stakeholder groups and data dimensions in the data warehouse domain, so as to set the right scope.
2. To investigate and explore the use of scorecards within the data warehouse domain, and formulate a conceptual framework for the development of a data quality scorecard
3. To develop and validate the conceptual data quality scorecard
4. To evaluate the data quality scorecard using two techniques (1) Case Study –live run through in 2 iterations; (2) semi-structured interviews

1.5 Research Methodology

Design science research (DSR) was selected as the desirable research methodology for the execution of this research project after analysis of other methodology. The researcher explored two other alternative methodologies: action research and applied research. Design science research was chosen as the best fit for this research project as it provides the required structure to enable the artefact created to be evaluated in iterative cycles.

The DSR process is a sequence of activities that produces an innovative product (i.e., the design artefact). The evaluation of the artefact then provides a continuous feedback of information and a better understanding of the problem to enhance both the quality of the product and the design process. This build-and-evaluate loop is typically iterated a few times before the final design artefact is generated (Markus et al., 2002). Once the problems of the thesis are assessed, the design science research addresses the research through building and evaluation of artefacts that are designed to meet the issues or of the hypothesis.

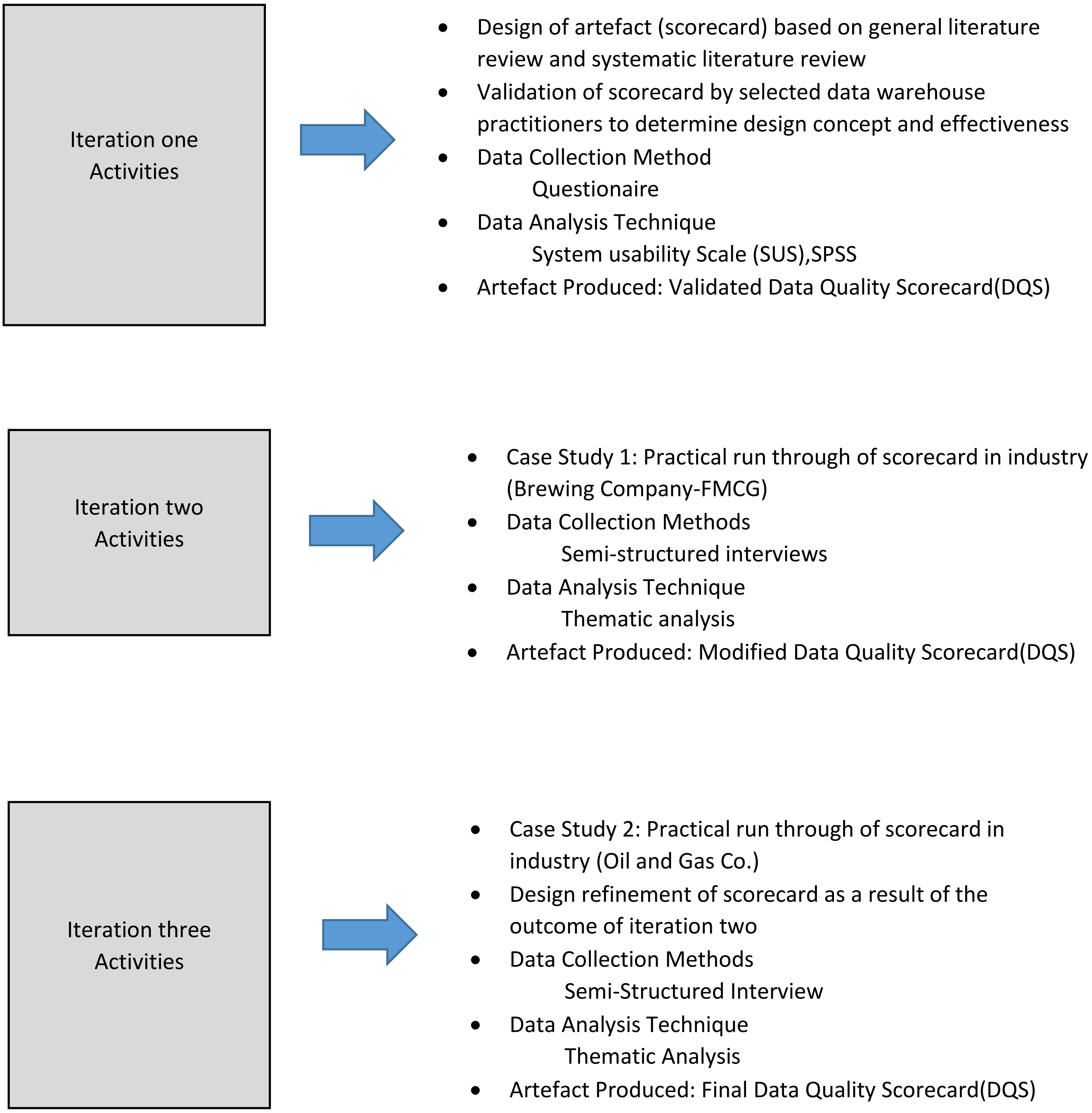
The four phases of the DSR methodology were included in the research design and implementation process to meet the objectives of the thesis as stated above. These phases were used in three DSR iterations reported in this thesis. The phases are; (1) Problem awareness (2) Suggestion (3) Development; and (4) Evaluation.

The first iteration is used to develop the data quality framework. The conceptual model of the framework and scorecard is designed based on the results of the general and systematic literature review conducted. The limitations of data quality as discovered during the literature review and systematic literature review were used to drive the artefact development process. The development of scenarios to thoroughly test the identified gaps was carried out. Scenarios were developed based on the six most relevant data quality dimensions according to the literature review conducted. An initial evaluation was then carried out by performing a qualitative interview with the identified stakeholder groups within the data warehouse domain.

The second iteration is conducted as a case study in a brewery company. The artefact produced from the first iteration: the data quality framework was deployed within the brewery organisation. The representatives of the identified four stakeholder groups (i) Data Producers (ii) Data Managers (iii) Data custodians (iv) Data took part in the exercise. An overt and unsystematic observation technique was performed to gather relevant data. In addition to the observation carried out, a semi-structured interview was conducted. A key advantage of using semi-structured interviews is that it allows the researcher to ask additional questions to gain further clarity on the data obtained during observation and the interview itself. The data from the interview was transcribed and was then analysed using an analytic technique called thematic analysis.

In the third iteration of this research, a case study was conducted in an oil and gas company. The results of the analysis performed in the second iteration suggest that not all the data dimensions are required for effective and efficient data quality management.

Figure 3: below summarizes the activities performed within each of the iterations.



Iteration one Activities

- Design of artefact (scorecard) based on general literature review and systematic literature review
- Validation of scorecard by selected data warehouse practitioners to determine design concept and effectiveness
- Data Collection Method
  Questionaire
- Data Analysis Technique
  System usability Scale (SUS),SPSS
- Artefact Produced: Validated Data Quality Scorecard(DQS)

Iteration two Activities

- Case Study 1: Practical run through of scorecard in industry (Brewing Company-FMCG)
- Data Collection Methods
  Semi-structured interviews
- Data Analysis Technique
  Thematic analysis
- Artefact Produced: Modified Data Quality Scorecard(DQS)

Iteration three Activities

- Case Study 2: Practical run through of scorecard in industry (Oil and Gas Co.)
- Design refinement of scorecard as a result of the outcome of iteration two
- Data Collection Methods
  Semi-Structured Interview
- Data Analysis Technique
  Thematic Analysis
- Artefact Produced: Final Data Quality Scorecard(DQS)

Figure 3: Summary of Iteration activities

1.6 Thesis Layout

Chapter 1: This chapter introduces the research into the data warehouse domain, with particular emphasis on data quality issues within the domain. The importance of data quality and the perception of quality was then highlighted. Afterwards, the research aims and objectives are presented. The methodology adopted to provide the structure, and logical flow of this research is discussed and presented.

Chapter 2: Presents the research methodology that will be followed throughout the conduct of this research to provide the structure and guide for the thesis. It comprehensively describes the research methods used in designing and testing the proposed data quality scorecard, lists the research assumptions and details the research design. This chapter also presents the possible alternative research methods and techniques that could have been used while also addressing the rationale for the selected research methodology for this research project.

Chapter 3: In this Chapter, the researcher reviewed literature related to the aim and objectives of the research, and also conducted a systematic literature review of the design of scorecards. This chapter is organised in the sections as described below:

1. The first section highlighted the main strengths and limitations of current approaches to data quality management.
2. In the second section, the state of the art is explored, issues, challenges, data quality tools and techniques are highlighted and investigated
3. In this section, literature related to data warehouse roles, stakeholder groups and data quality dimensions were reviewed.
4. This section highlights the main strengths and limitations of scorecards through a systematic review of the relevant literature.

Chapter 4: This chapter is divided into two main sections, in the first section, the data quality measurement scorecard was developed based on the gaps identified in the literature and systematic literature review carried out, this formed the first iterative cycle of the DSR methodology. The development of the artefact followed the DSR methodology process. In the second section, using SUS, a questionnaire was conducted on a random selection of data warehouse stakeholders from three different companies to gather qualitative data on the design and suitability for operational use of the developed scorecard. The results of the questionnaire conducted were analysed using the SUS technique and SPSS, and are presented in this chapter.

Chapter 5: The second DSR iteration is presented here. The designed scorecard as developed and presented in the previous chapter was deployed to a brewing company, as a case study to evaluate the benefits of the scorecard to measure data quality. Observation and semi-structured interviews were carried out by the various data warehouse stakeholder groups to ascertain the functionality of the scorecard. The data collected by these techniques were then analysed using thematic analysis. The results of the analysis led to the refinement of the artefact.

Chapter 6: The research presented in this Chapter continues from work carried out in the previous chapter. The artefact has been refined based on the outcome of the second iteration, and this chapter presents the evaluation of the artefact after undergoing refinement. Following the guidelines from Hevner et al., (2004), the designed artefact has to be rigorously tested and evaluated. The area of concern from the previous iteration from the data warehouse stakeholder group about the scorecard, i.e., the time dimension, prompted a re-design of the scorecard, which was retested in this iteration using three techniques; case study, observation and Semi-structured interviews.

Chapter 7: presents the overall findings and contributions of the research to both theory and practise. The chapter describes how the research objectives were accomplished; this chapter also outlines the drawbacks of the overall research and then presents the concluding remarks and potential areas for further research.

Chapter 2: Research Design

2.1 Introduction

This chapter explicates the research method that will be followed throughout the conduct of this research. It comprehensively describes the research method used in designing and testing the proposed data quality scorecard, and details the research design. This chapter also presents the possible alternative research methods that could have been used while also addressing the rationale for the selected research methods, techniques and tools in this research project.

The chapter is organised as follows: **Section 2.2** Introduces Design research and details the rationale for the selected methodology. Definitions are provided for the terms used in the context of this study, paradigm, technique, method and methodology to avoid conflicting interpretations. **Section 2.3** describes the methods and techniques followed to accomplish the aim of this research. **Section 2.4** highlights the practical application of design science to this research and presents the iterations conducted. **Section 2.5** Provides the summary of the chapter.

2.2 Design Science Research (DSR) Paradigm

The research methodology majorly functions to guide the researcher, and to provide structured steps to ensure the research project follows a defined path (Peffers et al., 2007).

Design science research (DSR) was selected as the desirable research methodology for the execution of this research project. The researcher explored two other alternative methodologies: action research and applied research. Design science research was chosen as the best fit for this research project as it provides the required structure to enable the artefact created to be evaluated in iterative cycles.

The DSR process is a sequence of expert activities that produces an innovative product (i.e., the design artefact). The evaluation of the artefact then provides a continuous feedback of information and a better understanding of the problem to enhance both the quality of the product and the design process. This build-and-evaluate loop is typically iterated a number of times before the final design artefact is generated (Markus et al. 2002). Once the problems of the thesis are assessed, the design science research addresses the research through building and evaluation of artefacts that are designed to meet the issues or the needs of the hypothesis.

Terminology Definition: Methodology, Paradigms, Methods and Techniques

Mutchnick and Berg (2015) state that the term research methodology defines the general aims and approach of a research project. The choice of research method chosen is dependant on the researcher's output requirements and research scope. The research methodology development follows logically from the research paradigm. Correspondingly, Mouton (2013) states that a research methodology is also referred to as the series of instructions and guidelines to be followed in representing the research issue. The research methodology directs the efforts of research by applying the context within which it is organised and offers the link between research activities and research objectives. Research methods are acquired from methodological paradigm such as either quantitative or qualitative which fits a specific research issue.

The table below shows the research paradigms and the methods and techniques for each approach

Paradigm	**Ontology**	**Epistemology**	**Theoretical Perspective**	**Methodology**	**Method**
Positivism	There is a single reality or truth (more realism)	Realism can be measured, and hence the focus is on reliable and valid tools to obtain that.	Positivism Post-positivism	Experimental research Survey research	Usually quantitative could include: Sampling Measurement and scaling Statistical analysis Questionnaire Focus group Interview
Constructivist/ interpretive	There is no single reality or truth. Reality is created by individuals in groups (less realist).	Therefore, realism needs to be interpreted. It is used to discover the underlying meaning of events and activities.	Interpretivism (reality needs to be interpreted) Phenomenology Symbolic interactionism Hermeneutics Critical inquiry Feminism	Ethnography Grounded theory Phenomenological research Heuristic inquiry Action research Discourse Analysis Feminist standpoint research	Usually qualitative could include; Qualitative interview Observation Participant Non-participant Case study Life history Narrative Theme identification
Pragmatism	Reality is continuously renegotiated, debated, interpreted in light of its usefulness in new, unpredictable situations.	The best method is one that solves problems. Finding out is the means, change is the underlying aim.	Deweyan pragmatism *Research through design*	Mixed methods Design-based research Action research	Combination of the above and more, such as data mining expert review, usability testing, physical prototype
Subjectivist	Realism is what we perceive to be real	All knowledge is purely a matter of perspective	Postmodernism Structuralism Post-structuralism	Discourse theory Archaeology Genealogy Deconstruction	Autoethnography Semiotics Literary analysis Pastiche Intertextuality
Critical	Realities are constructed entities that are under the constant internal influence	Reality and knowledge is both socially constructed and influenced by power relations from within society	Marxism Queer theory feminism	Critical discourse analysis, critical ethnography Action research Ideology Critique	Ideological review Civil actions interviews focus groups, open-questionnaires, open-ended observations, and journals

Table 1: Paradigms, methodologies and Methods

Design Science Methodology Process steps

Fettke et al., (2010) describe the breakdown of DSR into structured phases. The five DSR phases include (1) Awareness of the problem, (2) Suggestion, (3) Development, (4) Evaluation, and (5) Conclusion. Figure 4: below shows the DSR methodology process steps

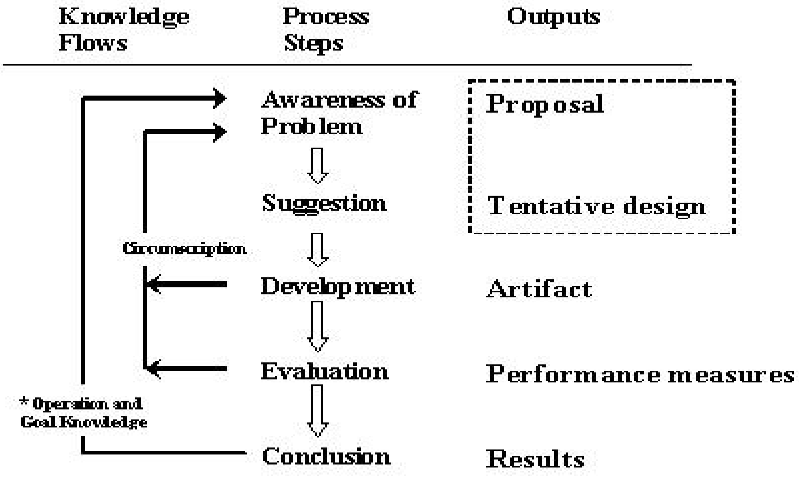


Figure 4: Design Science Research Methodology (Hevner et al., 2007)

The DSR phases are further explained below:

Awareness of Problem:

The vital problem awareness may exist from several sources such as in a reference discipline or new industry developments. In a related subject, reading may also offer the chance for new observations to the field of the researcher. This phase output is a formal, informal or proposal for efforts of further research (Brancheau et al., 2014). The literature review pointed out that organisations are facing the challenges of data quality, which results either from mergers and acquisitions with other companies, consolidation of different systems, from the upgrade of systems or from an unwillingness to simplify the data storage architecture. The literature reviewed also showed that the lack of proper engagement at various levels within the data warehouse roles involved in the process meant that where data quality checks are available, it was ineffective, as the awareness of these tests were not shared by all involved in the process.

Suggestion:

DeLone and McLean (2012) explained that behind the plan the suggestion phase succeeds and is linked intimately with it as the marked line around Tentative and Proposal Design which is the suggestion phase output represents. Indeed in any design science, formal research plan such as an industry sponsor or one to be made to NSF, an experimental design and likely the prototype based performance on that model would be the significant section of the plan. Moreover, if after an interesting problem assumption an interim model does not represent itself to the researcher, the proposal will be permitted. Similarly, Mingers (2015) described that suggestion is an important innovative step wherein the new functionality is envisaged based on the novel form of either unique or occurring and subsisting element. The action has been indicated as implementing non-repeatability into the method of DS research and creativity of human is a poorly understood cognitive process. However, the effect has essential similarities in entire methods of research for instance in the creativity of positivist research is innate in a leap from a strong desire about phenomena of an organisation to the proper development of construct which that expresses the aspects and adequate RD for their measurement.

Development:

In this phase, the tentative Design is implemented. For implementation, the techniques will differ relying on the artefact to be built. An algorithm may need formal proof construction. An expert system incorporating novel considerations about human cognition in the interest area will necessitate the development of software probably using high-level tool or package. The implementation can be pedestrian and would not involve innovation beyond the state-of-practice for given artefact, and the novelty is mainly in the design not in the artefact construction (Hevner et al., 2007)

Evaluation:

Ulrich (2016) suggests that the artefact is evaluated once constructed according to a principle which is often absolute and made explicit frequently in the phase of awareness of the problem. Deflections from expectations, both qualitative and quantitative are noted carefully and must be explained tentatively. That is, the phase of evaluation consists of an analytic subdivision in which the hypotheses are made about the artefact behaviour. This phase implements an epistemic fluidity which is in sheer contrast to the severe explanation of positivist stance. The analysis either contradicts or confirms a hypothesis at a similar point in positivist research. Peffers et al., (2013) mentioned that significantly save for some assumptions of future work as may be represented by experimental outcomes the efforts of the study is over. By contrast, events are getting interesting for the researcher of design science. Rarely in DS research, are starting hypothesis regarding behaviour wholly borne out. Instead, the outcomes of the evaluation phase and extra information achieved in construction and performing of the artefact are integrated together and offer to another suggestion round. The explanatory hypotheses, which are somewhat extensive are rarely discarded but are modified rather be in confer with new findings. This represents a new design, frequently preceded by new directions for library research suggested by theoretical performance deviations. The researchers of design science seem to share the conception of Allen Newell from theories of cognitive science as robust and complicate nomological networks. This notion has been noticed by science philosophers in several communities (Agarwal and Lucas, 2015), and performing from it Newell describes that theories are not similar to clay pigeons, to be exploded to bits with falsification of the Popperian shotgun. Preferably they must be treated like doctoral students. One rectifies them when they make mistakes and is promising they can amend their fault behaviour and go on to be more productive and useful (Newell, 2012).

Conclusion:

This phase is the last part of a particular effort of research. Typically, it is the outcome of fulfilling, i.e. still there are declinations in the artefact behaviour from the re-examined hypothetical findings, the results are declared better enough. Not only are the outcomes of the effort written up and consolidated at this phase, but the knowledge achieved in the effort is frequently classified as either fact of a firm theory, that has been studied and can be applied repeatably, or behaviour which can be invoked repeatably, or as loose ends anomalous behaviour which refuses description and serves well as a further research subject (Carlsson, 2013).

Design Science Research Outputs:

In this section a comprehensive perspective which interprets the levels and kinds of knowledge which can be derived from DSR while assuring judgment on whether more broad DSR aims must be held within any particular community of research. Baskerville (2008) compares natural science research with DSR and forms four general outputs for DSR such as models, methods, instantiations and constructs. Birkhofer (2011) described that the design methodology key is predicting a better solution for every situation of design whether it be in architecture, technology or industrial design. Design Methodology forces the brainstorming usage to motivate creative ideas and cooperating thinking to perform through every approach and reach a better solution. Attaining the wishes and requirements of the end user is the most challenging concern. Design methodology also offers basic research methods such as testing and analysis.

The below figure shows the different outputs of DS research which are categorised by abstraction levels:

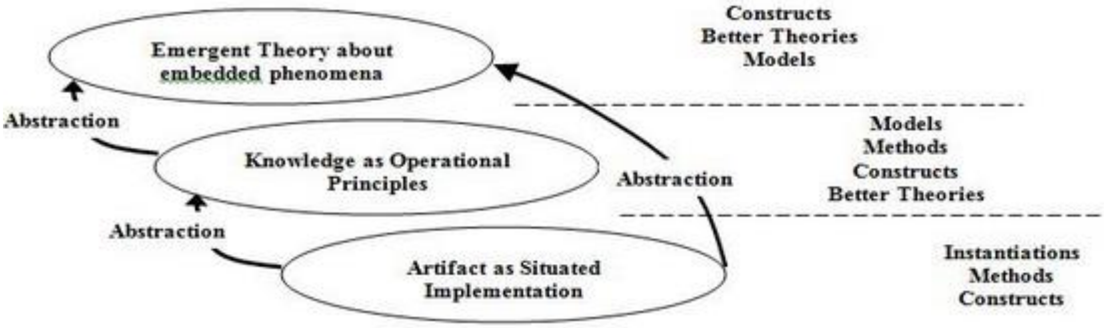


Figure 5: Design Research Outputs (Birkhofer, 2011)

Each of these outputs is further explained below::

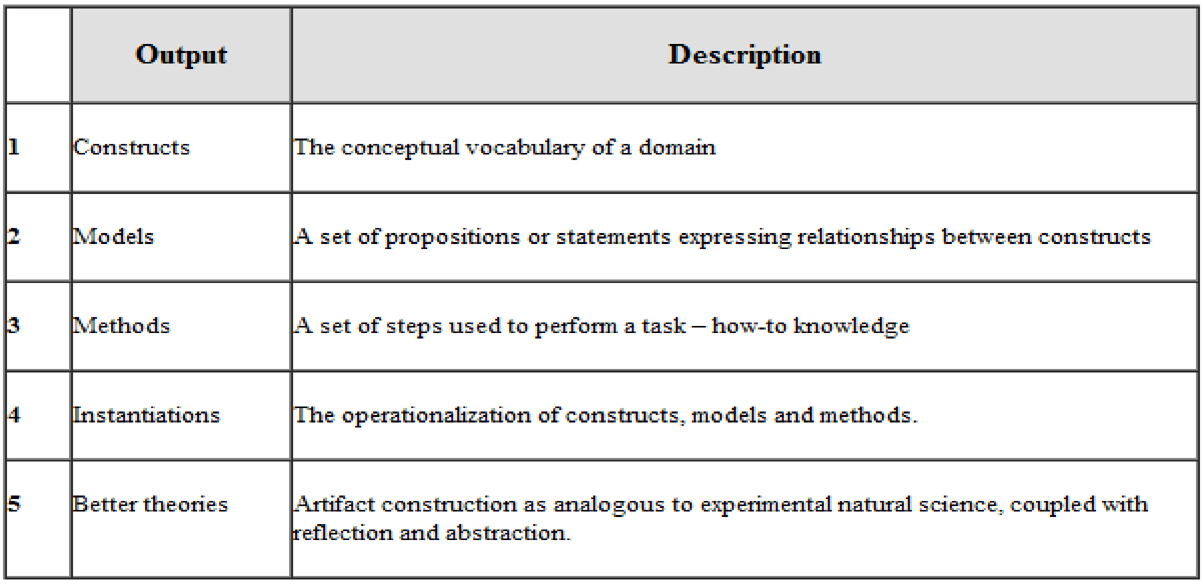
According to Baskerville et al., (2009), a model is a series of statements or propositions representing constructs relationships. March and Smith (2008) recognise models with the solution and problem statements. They are suggestions for how things should be or are. From the theories of natural science, models vary mainly in intent that physical science has a traditional concentration of truth whereas DS research concentrates more on utility. Thus a model is denoted regarding what it does, and a theory explained in construct relationships terms. However, an argument can often be extended to what can be completed with a series of entities and implicit knowledge, and proposed relationships are usually denoted as a theoretical statement of why or how the results exist (Kuechler and Vaishnavi, 2011).

Goldkuhl (2014) explained that a method is a series of stages, i.e. guidelines or an algorithm used to operate an activity. Methods are targeted supervised schedules for estimating constructs so that the solution statement model is comprehended. Implicit in a DS research method is the solution and problem statement represented in construct vocabulary. A more efficient way of enclosing final results sometimes or even primarily a natural or previously gained outcome is valued since the DS research axiology stresses problem-solving. From a DS research, Smith and March's explication efforts is an instantiation output which expresses models, methods and constructs. It is the artefact realisation in surroundings. Emphasizing the DS research proactive nature, they mention that sometimes an instantiation foregoes an entire renunciation of the theories or models and conceptual vocabulary which it provides. The researcher emphasises this further by denoting the instance of aeronautical engineering, and it is unlikely the understanding would ever have existed in the absence of performing artefacts (Kuechler and Vaishnavi, 2011).

The final output of the design science research is the constructs which are the conceptual vocabulary of a solution/problem domain (Friedman, 2013). During the definition of the problem, constructs emerge and are elegant throughout the design cycle. Since a working design, i.e. artefact involves a considerable amount of entities and their relationships, the series of construct for DS research procedures may be more significant than the regular series of detailed methods.

Baskerville et al., (2011) in a continuous integrated effort to deliver DS research have given an account of their list of DS research results. Whole but one of these can be represented directly in Smith and March’s list. Their fifth output is ethical theories which are essential and merits inclusion in their general list of outputs of DS research. DS research can contribute to theory building or good arguments in at least two varied ways, both of which may be denoted as similar to preliminary scientific examination in the sense of natural science. First the artefacts methodological building is a theorising object for several communities, the phase of construction of DS research efforts can be the method of an experimental proof or an exploration of the empirical approach or both. Similarly, Samuel-Ojo et al., (2010) mentioned that secondly, the artefact could express relationships between its elements. However, if the artefacts elements relationships are smaller than wholly understood and if the link is made more applicable than past during either the evaluation or construction artefact phase, then the elements understanding has been developed primarily elaborating or falsifying on earlier theorised relationships. For some kinds of research, the building of artefact is valued highly accurately for its integration of theory. Walls et al., (2014) expand the importance of theory building of construction and design in the particular context of information systems.

The below table summarises the outputs which can be acquired from the efforts of design science research:



	Output	Description
1	Constructs	The conceptual vocabulary of a domain
2	Models	A set of propositions or statements expressing relationships between constructs
3	Methods	A set of steps used to perform a task – how-to knowledge
4	Instantiations	The operationalization of constructs, models and methods.
5	Better theories	Artifact construction as analogous to experimental natural science, coupled with reflection and abstraction.

Table 2: Design Science Research Outputs (Baskerville et al., 2011)

2.3 Research Methods and Techniques

The researcher adopted various research techniques/methods during the execution of this research project due to the multidisciplinary nature of information system research. However, due to the requirement of this research project to target a particular stakeholder group in the data warehouse domain, a mixed approach of both quantitative and qualitative methods was followed throughout the research. Critical and interpretive researchers mainly analyse qualitative data. Hence, the researcher took an interpretive philosophical stance due to the mainly qualitative interview driven hypothesis derived from the initial research model. Researchers often adopt qualitative studies majorly when the research requires the researcher to observe the participant’s behaviour (Richards and Morse, 2015).

Qualitative studies are not based on numbers in most cases but are designed to show a focused target audience's range of behaviour. In-depth studies of small groups of people are usually used to guide and support the construct of the hypothesis. The data sources for such studies are typically focused interviews, case studies, observations and theme identification. The primary methods used in this research were case studies, semi-structured interviews and observation. These methods are explained below with emphasis on how it achieves the aim of the research.

Case Study

The second and third iteration of this research employed the use of a case study to perform a qualitative investigation into the measurement of the impact of the deployed DQS in two different organisations.

A case study is defined as a multifaceted investigation using qualitative research methods (Feagin et al., 1991). Using a case study for the second and third iteration of this research has enabled the researcher to examine and observe the direct interaction between the various stakeholder groups within the data warehouse domain. The researcher was also able to grasp the totality of the complexity of the stakeholder roles within the two diverse organisations studied, and how this affects their approach to data quality management.

Case studies tend to lead the researcher to the suggestion of new interpretations and the re-examination of earlier preconceived concepts in innovative new ways (Yin, 1984).

Semi-Structured Interviews

Semi-structured interviews termed SSI in this research, are a widely accepted way of gathering qualitative data. Semi-structured interviews allow the researcher to ask clarification questions to gain clarity about answers given to question. This is key as generalisations and ambiguity can be immediately resolved. The three iterations conducted in this research makes use of interviews to gather qualitative data at various points (These are discussed in more details in subsequent Chapters).

Arthur and Nazroo (2003) emphasis on the significance of careful preparation for interviews, and particularly the planning of a "topic guide". The primary focus of planning the semi-structured interview for this research was on identifying relevant questions to ask in the interview that corresponds to the appropriate stakeholder group within the data warehouse domain. For example, asking a question on the ease of loading data into a data warehouse would not be a suitable or relevant question to ask the data manager stakeholder group.

Arthur and Nazroo went ahead to advise that planning the topic guide should be done within a frame comprising of the following;

(i) Introduction

(ii) Opening questions;

(iii) Core in-depth questions; and

(iv) Closure.

Legard et al., (2003) states the importance of building trust with the participants, noting that the interviewer is a "research instrument", but also pointed out the researcher needs "a degree of humility, the ability to be recipients of the participant's wisdom without needing to compete by demonstrating their own."

Observation

As mentioned earlier, two case studies were conducted in this research to execute the aims and objectives of this study, as part of the data gathering method, an observatory study was conducted. The form of observation to be performed depends on the type of research being carried out and can vary based on the researcher's approach (Flick, 2009).

Flick (2009, p.222) proposes five dimensions on which observational studies may vary:

- Covert vs overt: to what extent are participants aware of being observed?
- Non---participant vs participant: to what extent does the observer become part of the situation being observed?
- Systematic vs unsystematic: how structured are the observation notes that are kept
- Natural vs controlled context: how realistic is the environment in which observation takes place?
- Self---observation vs observation of others: how much attention is paid to the researcher's reflexive self-observation in data gathering?

The researcher combined elements from all five dimensions in this study for both case studies, however in most cases, the observation was overt and unsystematic. Flick (2009, p.223) also identifies seven phases of planning an observational research:

- Selection of setting(s) for observation;

- Determining what is to be documented in each observation;
- Training of observers (The researcher being the sole observer in this study)
- Descriptive observations to gain an overview of the context;
- Focused observations on the aspects of the context that are of interest;
- Selective observations of central aspects of the context;
- Finish when theoretical saturation has been reached –i.e. when nothing further is being learned about the context.

The broader idea of careful preparation for a study and recognition that the nature of observations will evolve is essential and was duly put into consideration by the researcher.

Willig (2008, p.28) highlights the nature of data gathering, with particular emphasis on the importance of keeping detailed notes, such as direct quotations from participants and "concrete descriptions of the setting, people and events involved". These are referred to as "substantive notes", which may be supplemented by "methodological notes" – based on the method applied in the research – and "analytical notes", which links and forms the beginning stage of the analysis of the collected data. Willig (2008) also argues that the gathering and analysis of data are somewhat integrated. The Data analysis techniques are discussed in the next section below.

Analysing Collected Data

Chapter 4 presents the first empirical study carried out in the course of this research project. An initial DQS was built based on the gaps identified from (i) literature review and (ii) systematic literature review. The researcher then used a SUS questionnaire to collect data. The system usability scale (SUS) is a one-dimensional scale which consists of 10 questionnaire items that evaluate the subjective perception of the stakeholder's usage of the system regardless of their personal interpretations (Brooke, 1996). SUS is an industry-accepted scale for measuring the subjective views of the users of the system. It utilises a five-point Likert scale with anchors for "strongly agree" to "strongly disagree ".

The first step in scoring a SUS is to determine each item's score contribution, which ranges from 0 (being a poor score) to 4 (a good score). For odd-numbered items, the score contribution is the scale position minus 1, while For even-numbered items, the score contribution is 5 minus the scale position. The overall SUS score is derived by multiplying the sum of the item score contributions by 2.5, the result produces a score that can range from 0 (very poor perceived usability) to 100 (excellent perceived usability).

SPSS software package (IBM SPSS Statistics 20) was used for the data analysis of the system usability scale (SUS). To measure how closely related the internal consistency of the set of items are as a group, Cronbach‛s alpha was calculated (Tavakol and Dennick, 2011).

The first case study was carried out in a brewery company with an explicit aim to validate the artefact- DQS. Semi-structured interviews were used to collect qualitative data from the stakeholders on their views of the applied DQS to their current processes. The received data were analysed using thematic analysis. (Details of the intricacies of the case study, data collection and analysis are presented in Chapter 5).

The second case study was on an Oil and Gas company. The DQS was refined as a result of the data analysis performed in the first case study; the data quality dimensions were reduced to five, as the dimension for timeliness was found to be redundant by the stakeholders from the first evaluation.

2.4 Practical Application of DSR in this Research

In this section, a detailed description is given of how the DSR phases were adopted to execute the aims of this research project. The issues identified in the literature review and questionnaires conducted was used to formulate a purposeful artefact that was devised by the recognised four major stakeholder groups within the data warehouse domain; data producer, data custodian, data manager and data consumer. The literature reviewed also highlighted six crucial dimensions of data quality within the data warehouse domain; these dimensions were deployed for the development of the artefact. The artefact used in this thesis was developed using the DSR framework as described in earlier sections. The framework provided the structure and guidelines for the development of the artefact. Three research methods; case study, semi-structured interviews, questionnaires were adopted to validate the artefact in this research based on ideas extrapolated from various research paradigms.

A three iteration plan was designed for this research project. The first iteration was aimed at developing the DQS, based on literature review and validated using a system usability scale (SUS). The focus of the second iteration was to deploy and evaluate the functionality of the designed DQS in industry, testing at a Brewing company was done to achieve this. The results of the initial live test required refinement to be made to the scorecard from six dimensions of data quality to five dimensions. The third iteration focused on the effectiveness of the refined DQS by testing it in an Oil and Gas Company. The artefact was utilised to attain the desired ends, and the results were communicated efficiently to both, technology-oriented and management-oriented audience. While the details of the research methodology are mentioned in this Chapter, the intricacies of the interviews, the development of artefacts and iteration cycles are explained in Chapter 4 and subsequent Chapters in this thesis.

2.4.1 First DSR Iteration Cycle

The first iteration is used to develop the data quality scorecard. The conceptual model of the scorecard is designed based on the results of the general and systematic literature review conducted. The limitations of data quality as discovered during the literature review and systematic literature review will be used to drive the process. The development of scenarios to thoroughly test the identified gaps was carried out. Scenarios were developed based on the six most relevant data quality dimensions according to the literature review conducted. An initial evaluation was then carried out by performing a system usability scale (SUS) with the identified stakeholder groups within the data warehouse domain. The DSR phases are expanded on below; while Figure 1: shows the DSR phases, methods and artefacts produced during the first iteration.

Phase 1- Problem Awareness

With the advent of 'Big Data' and with companies now spending more time and money harnessing and using data from all areas of their business, data quality and the assurance of quality is now a significant area of study for the academic and business community. Big data has changed the way corporations view data and the intrinsic value they can generate from harnessing the data. Data quality dimensions enable companies to further analyse data based on focused areas. It is therefore essential to consider having an efficient and flexible system of ensuring data quality is being met.

It is clear as a result of the literature reviewed that organisations are facing the challenges of data quality, which results either from mergers and acquisitions with other companies, consolidation of different systems, from the upgrade of systems or from an unwillingness to simplify the data storage architecture. The literature reviewed also showed that the lack of proper engagement at various levels within the data warehouse stakeholder groups meant that where data quality checks are available, it was ineffective, as the awareness of these tests were not shared at all levels.

The awareness of the problems for this iteration was identified using the following sources:

(i) A literature review of the issues and state of the art of data warehouse and data quality dimensions.

(ii) A literature review of the roles involved in the data warehouse domain.

(iii) A systematic literature review of the design of a scorecard

Phase 2 – Suggestion

In this phase, the results of the literature review conducted on data warehouse, and the systematic literature review on scorecards and data quality dimensions provided the researcher with a conceptual framework of ideas to enable the development of the data quality scorecard. The theoretical framework suggests that when designing the data quality frame, input from the following stakeholder groups within the domain is critical; (1) Data Producers; (2) Data Managers; (3) Data Custodians; (4) Data Consumers. The following data dimensions are considered as the most important by the data warehouse stakeholders: (i) Completeness; (ii) Validity; (iii) Accuracy; (iv) Timeliness; (v) Consistency; (vi) Integrity

This research identified scorecards as an efficient way to measure results and performance.

Phase 3 – Design

In this phase, the proposed data quality framework was developed based on the limitations identified in the general and systematic literature review carried out. The constructs within the model include (i) Data quality dimensions (ii) Data quality concern areas (iii) Data warehouse stakeholders. As stated by Hevner et al., 2007, the novelty is mainly in the design not in the artefact construction.

The DQS is designed for use as either a writable electronic form using a portable document format (PDF). PDF is a widely accepted file format used for presenting and exchanging data reliably, or as a web-based DQS. The web-centric scorecard was developed using HTML 5, CSS 5, PHP, JavaScript and MySQL database management system. The web version of the DQS is currently hosted on a private network from a commercial ISP, the researcher chose this particular service provider because it guarantees a 99.95% uptime and for its overall reliability.

The web-centric scorecard can also be hosted privately by companies on their intranet so as to secure the scorecard solely on their private network.

Phase 4 - Evaluation

The researcher then validated the DQS using a SUS questionnaire to collect data. The system usability scale (SUS) is a one-dimensional scale which consists of 10 questionnaire items that evaluate the subjective perception of the stakeholder's usage of the system regardless of their personal interpretations (Brooke, 1996). SPSS software package (IBM SPSS Statistics 20) was used for the data analysis of the system usability scale (SUS).

To measure how closely related the internal consistency of the set of items is as a group, Cronbach‘s alpha was calculated (Gliem and Gliem, 2003). The Cronbach‘s alpha of 0.782 was calculated. This shows larger values than the acknowledged level of 0.7. Nunnaly (1978) has indicated 0.7 to be an acceptable reliability coefficient. Hence, making this analysis statistically adequate. Details of the SUS are presented in Chapter four of this thesis.

2.4.2 Second DSR Iteration Cycle

The second iteration is conducted as a case study in a brewery company. The artefact produced from the first iteration: the data quality scorecard was deployed within the brewery organisation. The representatives of the identified four stakeholder groups (i) Data Producers (ii) Data Managers (iii) Data custodians (iv) Data consumers; took part in the exercise. A semi-structured interview was conducted to collect data after a run through in the brewery company. A key advantage of using semi-structured interviews is that it allows the researcher to ask additional questions to gain further clarity on the data obtained during observation and the interview itself. The data from the interview was transcribed and was then analysed using an analytic technique called thematic analysis. More details on the initial theme and final themes generated using the thematic analysis carried out are presented in Chapters 5 and 6 of this thesis.

Phase 1- Problem Awareness

The problem awareness for the second iteration was drawn from the evaluation results from the first iteration. After the SUS carried out in the first iteration, the results suggest that the participants found the DQS useful. However, to further evaluate the DQS, extensive consultation with the data warehouse stakeholders from various organisations will be necessary to ensure the robustness and fit for the operational use of the designed data quality scorecard.

Phase 2 – Suggestion

The suggestion is based on the literature review as well as the systematic review of the use and characteristics of scorecards and data quality dimensions. A conceptual framework to guide the evaluation of the scorecard is proposed. The conceptual framework suggests that when designing an effective data quality scorecard, the data warehouse stakeholders; (i), Data Producers, (ii) Data Custodians, (iii) Data Managers, and (iv) Data Producers need to be involved in the evaluation of any data quality solution. The design of the DQS should consider the data quality dimensions for effective data quality management.

While data quality scorecards have been empirically validated in other research contexts, this study proposes the need to extend the use of scorecards by including the DW stakeholder groups and data quality dimensions as a construct to efficiently formulate a model that best measures the users requirements.

Phase 3 – Design

The design uses the exact data quality scorecard designed in the first iteration as the artefact. No change was made to the artefact as this iteration is mainly for the evaluation of the designed artefact.

Phase 4 – Evaluation

Two evaluation techniques were adopted in the second iteration for testing the data quality scorecard. Firstly, a case study was conducted. The case study is a run through of the DQS in industry. Participants within the data warehouse stakeholder groups were asked to use the DQS to measure the quality of data at various stages within the data warehouse. A semi-structured interview was then conducted to collect the views of the stakeholders on the use of the DQS. Using thematic analysis technique to analyse the data, the results suggest the DQS was useful in assisting the data quality stakeholders in measuring data quality in various susceptible areas.

The findings from the analysis show that not all the selected data quality dimensions are needed for efficient use of the data quality scorecard. The data quality dimension for timeliness was found to be redundant. (Details of the evaluation conducted for this iteration are presented in Chapter 5)

2.4.3 Third DSR Iteration Cycle

In the third iteration of this research, a case study was conducted in an oil and gas company. The results of the analysis in the second iteration suggest that not all the data dimensions are required for effective and efficient data quality management. The phases in the DSR methodology as it applies to this iteration is explained below, and a diagrammatic representation of the DSR phases, methods, techniques and data sources of this iteration is shown in figure 3 below.

Phase 1- Problem Awareness

After the thematic analysis carried out on the semi-structured interview conducted in the second iteration, the results suggest that not all the data dimensions are relevant for data quality measurements, and including all six as part of the DQS might reduce the usage of the DSQ. For example, the timeliness dimension was seen as not particularly relevant as all loaded data already had an audit time string before being loaded into the data warehouse. Therefore, in the final DQS, the data quality dimension for 'time' was excluded.

Phase 2 – Suggestion

The suggestion is based on the literature review as well as the systematic review of the use and characteristics of scorecards and data quality dimensions. A conceptual framework to guide the evaluation of the refined scorecard is designed. The results of the thematic analysis conducted in the previous iteration suggest that the timeliness dimension can be removed from the scorecard design.

Phase 3 – Design

In this phase, the requirement analysis for the refinement of the data quality scorecard was conducted. The data quality scorecard was modified excluding the data quality dimension for timeliness. As explained earlier, the timeliness dimension was not seen by the stakeholder group as being a critical dimension for the measurement of data quality as all data is timestamped within the data warehouse already.

Phase 4 - Evaluation

Two evaluation techniques were adopted for the third iteration for evaluating the data quality scorecard. The first evaluation technique adopted a case study, which was a run through of the modified DQS in the Oil and Gas domain. The main aim of the case study is to ascertain if the stakeholders see the data quality scorecard as a useful and efficient data quality measurement tool.

A semi-structured interview was conducted to collect the views of the stakeholders on the data quality dimensions selected for the scorecard. The responses of the stakeholders were analysed using thematic analysis.

The findings from the analysis show that the modified DQS is an efficient data quality measurement tool. (Details of the evaluation conducted for this iteration are presented in Chapter 6)

2.5 Summary

This research methodology design Chapter discusses the research paradigms, research methods, techniques and methodology used to conduct the overall research thesis. the literature reviewed showed that organisations are facing the challenges of data quality, mainly as a result of either mergers and acquisitions with other companies, consolidation of different systems, or from an unwillingness to simplify the data storage architecture. The literature reviewed also showed that the lack of proper engagement at various levels within the data warehouse stakeholder roles meant that where data quality checks are available, it was ineffective, as the awareness of these tests were not shared at all levels.

The rationale for selecting the DSR methodology as the underlining guide for the execution of this project is discussed, then, a detailed discussion of the research techniques and data analysis approach used was discussed and presented. A three iteration design plan was developed for this research using the various phases of the DSR methodology. In the next Chapter, a general and systematic literature review on the research domain is conducted to highlight and justify the scope of the overall research.

Chapter 3: Data warehouse, Data Quality Dimensions and Scorecards Literature

3.1 Introduction

This chapter explores the complexities of data quality assessment in the data warehouse domain. The importance of the data warehouse stakeholder group in managing quality is explored in the literature.

The chapter is organised as follows. **Section 3.2** provides an analysis of the data warehouse domain and the segments within the domain. Literature themes and direction are covered in **Section 3.3**. The current state of the art of the domain is then presented. **Section 3.4** presents the data warehouse stakeholder roles. **Section 3.5** covers the design factors for an effective DQS; **Section 3.6** presents the systematic literature review carried out on scorecards. **Section 3.7** summarises the chapter and provides an introduction to the next chapter.

3.2 The Data Warehouse Domain

A data warehouse is a system for organising, gathering, sharing, and managing historical data. It consists of user data as the data exists from operational systems that acquire and use data within the context of that systems need (Laberge, 2011). The term data warehouse is always used to define a data warehouse system and at times about the repository of the data warehouse. Data warehouse repository will be used when referring to a vast number of database or its design which are tools of the data warehouse system. It is expected to have appropriate information in the appropriate place at the appropriate time with the appropriate cost to support the appropriate decision. Data warehousing has become an essential strategy to combine heterogeneous sources of data and to enhance online analytical processing. However, Golfarelli and Rizzi, (2017) affirm to the significant challenges faced in data analysis that is exemplified by the limitations on the size of data that is handled by the warehouse were the end-user analytical applications constitute the last stage in the analysis. Besides, the limits on the data size were considered to be arbitrary. A data warehouse is integrated because it denotes a unified view over several information systems. A data warehouse is non-volatile because data warehouses are enclosed, into which new data is loaded in huge loads but where data that has been entered once is not updated later on. A data warehouse is subject oriented because it is arranged around the central subject areas of an organisation such as products and customers. Finally, a data warehouse is a time-variant because the data warehouse consists of historical data with a time horizon of many years. Thus, the term data warehousing defines the entire methods, tools, concepts, and technologies.

Ponniah (2011) states that the primary attributes of a data warehouse are; 1) makes decision support transactions applicable without obstructing operational systems; 2) offers a combined an entire view of an organization; 3) denotes an interactive and flexible strategic information source; 4) makes the present and historical information of an organization simple possible for strategic decision making, and 5) Renders consistent information about an organization. Figure 6 below shows a typical data warehouse environment:

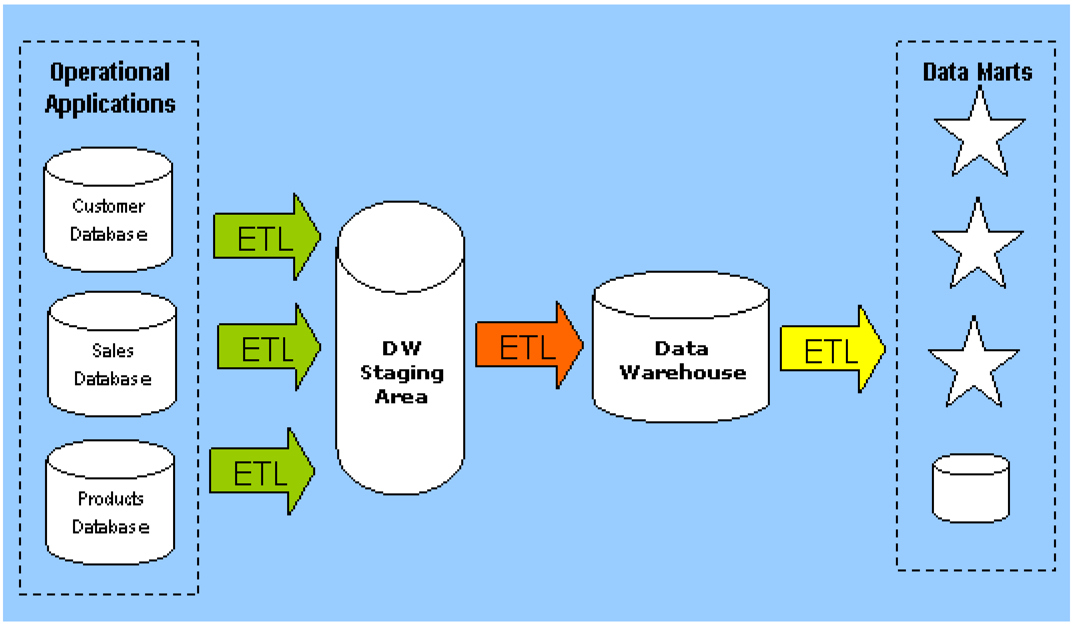


Figure 6: Data Warehouse Environment (Meyer et al., 2016)

Data within the data warehouse as seen in the environment shown above, needs to be staged during the ETL process, that is, to extract data from the source system and bring it, collectively, into the data warehouse. Such a process ensures that the quality of the data does not degrade; however, a data warehouse is defined by several intricacies that necessitates its effectiveness in a given environment. Segarra et al., (2016) asserted the use of scorecards in data warehousing to consider all the critical operational measures. The authors further suggest the improvements that could be achieved when the scorecard is employed as a strategy for control bias compared to the traditional measurement systems. The balanced scorecard was utilised to adopt techniques to identify the required data to be stored in the data warehouse and considers several strategies to achieve these goals.

3.2.1 Data Quality

Quality is referred to as the fitness for need and must not only consist of the intrinsic data characteristics itself but also assessments of data of users (Sloan et al., 2015). Thus, signifying that within a given data warehouse, the required data must be useful to the customers of data and support their similar practices at work. Two varied levels of data quality may be referred to data such as content (data) and structure (metadata). The data warehouse quality of a structure is referred to as the conceptual model quality that is the basis for the data warehouse design. Quality metadata is essential for all stakeholders in the process of data warehousing so that they understand what the data warehouse consists and how to access data in the data warehouse. Quality data is essential so that the data warehouse users can understand and assess data readily in the data warehouse and use the data efficiently in their tasks of decision making. Most of the data quality work contains a list of possible dimensions of data quality (Wang et al., 2011). The occurrence of data does not assure that all decisions and functions of management can be undertaken smoothly. The absolute data quality definition is that it is about worse data, i.e. the data is incorrect or invalid or missing in some context. A more comprehensive definition is that data quality is gained when an organisation uses data that is timely, comprehensive, relevant, understandable, and consistent. The first step to the improvement of data quality is to understand the critical dimensions of data quality. To be interpretable, and processable efficiently and effectively, data has to fulfil a group of quality criteria. Data fulfilling those criteria of quality is referred to be of higher quality. Affluent attempts have been made to refer to data quality and to recognise its dimensions.

To evaluate the objectives regarding data quality, Al Za'noun and Wilson (2015) asserted the use of scorecards in assessing the quality of data. The impact of improvement in data quality, and their financial viability, the procedure involved in the collection of data, analysed, presented, and further use in an optimal manner. These findings indicate that an appropriate technique is necessary for improving data quality that must be complete, accurate, and aggregated into formats that can be readily interpreted and used.

3.2.2 Dimensions of Data Quality

Data quality consist of dimensions such as reliability, accuracy, timeliness, usefulness, importance, precision, understandability, and conciseness. These dimensions are always referred vaguely, overlapping, not soundly and ambiguous based in theory.

According to the study conducted by Wang and Strong (1996), accessibility is perceived to be a component of the data quality rather than a separate entity. Guo et al., (2013) considered data accuracy along with source validity as an essential component of data quality in relations to internet-of-things applications. Li et al., (2012) posited using the currency, validity and availability data quality dimensions in pervasive applications wherein, the dimensions of currency and validity are intricately related to the timelessness and accuracy dimensions. In the context of data warehouses, the quality of data is of utmost importance; trustworthiness is a crucial component of user engagement and in sustaining the functioning of a data warehouse.

The dimensions of data quality are a feature or aspect of information and a way to categorise data quality and information requirements (Diggins et al., 2015). The data quality dimensions are used to measure, define and handle the information and data quality. Figure 10 below shows the dimensions of data quality:

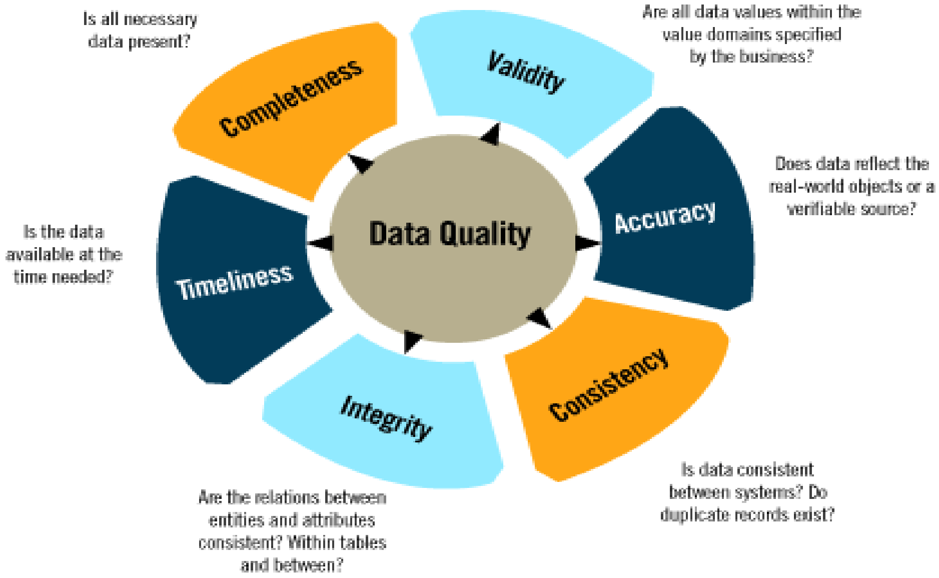

Is all necessary data present?
Completeness
Validity
Are all data values within the value domains specified by the business?
Data Quality
Does data reflect the real-world objects or a verifiable source?
Accuracy
Is the data available at the time needed?
Timeliness
Consistency
Integrity
Are the relations between entities and attributes consistent? Within tables and between?
Is data consistent between systems? Do duplicate records exist?

Figure 7: Dimensions of Data Quality (informatica.com, 2015)

The above-shown dimensions of data quality, are discussed below:

Data Accuracy:

Data accuracy is the measure to which the data appropriately reflects an event or real-world object being explained (Cai and Zhu, 2015). This accuracy denotes that most spatial phenomena observations are only assumed to the evaluation of actual value. The variations between actual and observed values represent observations accuracy. There are two kinds of accuracy; these are an attribute and positional accuracy. Positional accuracy is the regarded deviation in geographic place of an object from its real ground position. There are two tools of positional accuracy. These are absolute and relative accuracy. Relative accuracy concerns the map features positioning similar to one another. Absolute accuracy concerns data elements accuracy concerning coordinate scheme. Relative accuracy is of a more significant concern than absolute accuracy (Winkler, 2004). Therefore, Krenzelok et al., (2014) examined a paradigm for determining the accuracy of the data by considering relevant metrics based on the frequency of the process being studied such as scoreboard or checklists. They propose an iterative technique to be employed in a checklist comprising the progress of the data stored in the data warehouse and further outcome of the data to be utilised. Hence, data accuracy verifies for the actual representation of the world values. For instance, the bank balance in a customer account is the real value that the customer deserves from a given bank. Any inaccuracy in the existing data can lead up to operational, analytical woes (Singh and Signh, 2010).

Data Completeness:

According to Hazen et al., (2014), data completeness refers to the expected availability of the data. Even if the data is not complete, it may be sufficient enough to satisfy the user. Helfert (2014) described that completeness is the extent to which expected data attributes are offered. Completeness can be referred to as the degree to which data are of adequate depth, scope, and breadth of the activity at hand. There are three kinds of completeness. They are column completeness, schema completeness and population completeness. Column completeness is the missing value function in the table's column. Schema completeness is referred to as the measure to which attribute and entities are not missing from the schema. Population completeness numbers for calculating missing values concerning the reference population. If focusing on a particular model of data a more precise completeness characterisation can be given. For example, data of a customer is assumed as complete if all contact details, addresses and other information of customers are available

Data Consistency:

Data consistency means that data across the organisation must be similar to each other (Wang et al., 2015). The dimension of consistency catches the semantic norms violation referred to a collection of data components. Constraints of integrity are semantic norms instantiation concerning relational theory. Constraints of integrity are properties that must be fulfilled by all database schema instances. Data can be accurate, but it will be still inconsistent. Instances of inconsistency data are a credit card is inactive and cancelled, but the status of card billing shows due. Data is inconsistent, when it is common in large organisation domain but not universal across the organisation. The consistency of data means that the data across the enterprise should be correlated and synchronised with one another without providing any different data (Hazen et al., 2014).

Furthermore, the studies conducted by Collins et al., (2015) indicates that data consistency and completeness will be obtained by the collection of data based on the interaction of individuals based on the applications. This research lacked the consistency of data due to the absence of a reference mode leading to further complexity in the implementation and organisation of the data. Besides, the absence of a predefined reference model in the given data warehouse, led to condition was extensive manual effort was required to verify the completeness and consistency of the data. Therefore, for better data completeness and consistency a functional design is necessary to manage a substantial quality of data is required in addition to the consequences of the downstream process, techniques such as scoreboard or checklist could be used to manage the data.

Data Timeliness:

Timeliness is the measure to which data is adequately up to date for the activity at hand (Jarke, 2015). The data timeliness is exceptionally essential. The measurement of timeliness denotes that not only data are present but are also in time for particular usage. Therefore, available measurement includes currency measurement and a check if data are possible before the scheduled time of usage. Several difficult metrics can be referred for evaluation of time similar dimensions. The timeliness relies on expectations of users. Examples of data timeliness are the company’s financial statements are published one month after the end of the year.

Data Reliability:

Juran (2015) described that data reliability must reflect consistent and stable processes of data collection across gathering points and over time whether using computer-based or manual systems or an integration of both the concepts. Stakeholders and managers must be confident that progress towards performance goals reflects actual alterations rather than differences in methods or approaches of data collection. Appropriate relationship linkages among records are significant else it might introduce unnecessary duplication throughout the systems (Hazen et al., 2014).

Data Validity:

The validity of data must be recorded and used in agreement with general needs including proper application of any definitions or norms. This will assure consistency between periods and with familiar organisations. Where appropriate data is used for actual data absence, organisations must assume how well this data is capable of fulfilling the intended need. Data Validity depicts the correctness and reasonableness of data (Cai and Zhu, 2015).

3.2.3 Data Quality Model Foundations

A variation between the internal and external opinions of an information system is made with an initiation (Wand and Weber, 2015). The external view is related to the effect and use of an information system. It represents the justification and needs for the system and its deployment in the organisation an information system is assumed given that is a black box with the essential functionality to denote a real-world system in the external view. The perspective of external view is adopted by the researchers who are intrigued by the occurrences such as, the processes identified by the stakeholders to define the requirements of the information, the informal and formal power shifts that occur among the users when a given organisation implements the information system to attain a competitive edge.

Conversely, the internal view represents the operation and construction essential to meet the needed functionality, given a group of needs which considers the external view. Researchers interested in the aspect of internal view would focus on the intricacies of the different screen that might empower the users, the structure of data and processes that can enhance the functionality of a given system and varied hardware platforms that are required to meet response times. Other intricacies such as the construction of system consist of implementation and design; the operation of the system consists of tasks included in generating data such as data entry, data capture, data delivery and data maintenance. For simplicity, perfect implementation is considered because; an erroneous implementation is similar to an incorrect design with proper implementation for research purpose. Thus, the researcher's analysis focuses on the internal view and is aligned with data production and system design.

Wang et al., (2015) described that both the internal and external view have two essential conclusions. First, since the internal view is user-substantive, it helps the group of definitions of different dimensions of data quality that can be compared across various applications. Hence, these dimensions can be looked as intrinsic to data. Second, such different perspective of look can be perceived to lead the information system design with specific objectives of data quality. The variation between the internal and external views must not be defined within the process of successive systems development. Instead, it represents to implement designer, having no power of needs of users should take the needs as given at any time during development. It is possible that the system users and designers will collaborate in an iterative process of design as required.

In recent years, it was noticed that Balanced Scorecard (BSC) is an emerging technique utilised towards the designing, implementation and as a performance measurement tool, based on the organisation and managerial practices including operating practices that are not cost driven. However, the study conducted by Emami and Doolen (2015) asserts the reasons for not employing BSC even though it is an active method of assuring the quality of the data. They went on to suggest that several internal and external changes in the industry limit this technique from achieving the targeted level of performance measurement. Besides, in comparison to the conventional models BSC were found to have a customised performance measurement system based on organisational strategies. Presently, approaches based on the application of the BSC as a management and measurement technique have been rising dramatically.

3.2.4 Data Quality in Data warehouses

Data quality offered is difficult for the success of initiatives in data warehousing (Krogstie et al., 2015). There is substantial proof that several organisations have essential issues of data quality and these issues have consistent economic and social influences. Thus, it is imperative to maintain quality standards of data in a data warehouse to avoid such shortcomings in its functionality. Thus, by the thought mentioned above, various organisations have enhanced the functions of the data warehouse to reduce the cost based on the data provided to support a focus on entire processes of business and to gain more significant calculated ROI (McFadden, 2016). In the success of initiatives in data warehousing, the primary factor is the data quality offered. Therefore, it is vital that the quality of data be understood and that assurance procedure of data quality are established and developed. While several organisations are aware of the data quality importance for their capability to rival in the marketplace successfully, industry and the surveys of research represent that the organisations are experiencing data quality issues increasingly and that these have consistent social and economic impacts (Wang et al., 2015). There has been a lack of structures and methods for evaluating, improving, and measuring the quality of data and small discussion of the organisational, economic and management data quality aspects. Varied dimensions of data quality have been studied by researchers, in cognisant with which, some structures have been improved that put forth essential concepts for the understanding of data quality (Shanks and Darke, 2014), and helps methodical approaches to develop processes of data quality within organisations. Various groups of stakeholders have also been recognised by the consumption, maintenance, and generation of data. However, despite the relationship between data quality and stakeholders, limited studies have been conducted that primarily focus on the concerns of the stakeholders and the need for data quality.

Data warehouses are viewed as a means of offering infrastructure of data management for decision support systems, management support systems and executive information systems (Gartner Group, 2015). A data warehouse is a group of databases enhanced to offer information to decision makers and managers through some combined hardware and software surroundings that are optimised for extraction rather than for transaction throughput and update integrity. Efficient decision making in business relies on better and poor data quality which can be unsuccessful and sometimes expensive. The primary factor in data warehousing success is the data quality offered. Thus, to maintain the data quality, a given organisation must implement procedures to protect the data quality and to understand the notion to ensure optimisation of data quality within an organisation.

According to Wang (2013) still, several organisations do not have timely, useful and accurate data which they need for decision making and efficient operations despite their expenditure on IT. The issues of data quality are spreading widely in practice and can have essential economic and social impacts. Before the problems involved in handling quality of data can be denoted, it is essential to understand what data quality means first. They further asserted that organisations must treat information as a product that can enhance the customer base, without ridiculing the productivity of the organisation. Thus, the maintenance of data quality is of utmost importance to a given organisation. Drachsler and Greller (2016) considered several issues to process standardisation in the checklist and asserted the requirement of thought and planning in the creation of a checklist. A broadly used concept in the data quality domain is fitness for need. This must encompass not only the intrinsic data characteristics itself but also assessments by data users about data quality (Wu et al., 2015). In a data warehouse, data must be useful and usable for data customers and support their practices of work.

There is no proper consensus of what forms the final group of dimensions of data quality (Olson, 2013), although completeness, currency and data accuracy are assumed essentially. Some structures have been developed which structure and arrange essential data quality concepts. The authors arrange dimensions of data into four main types. They are: contextual, intrinsic, and representational and accessibility. Figure 11 below shows the hierarchy of data quality issues:

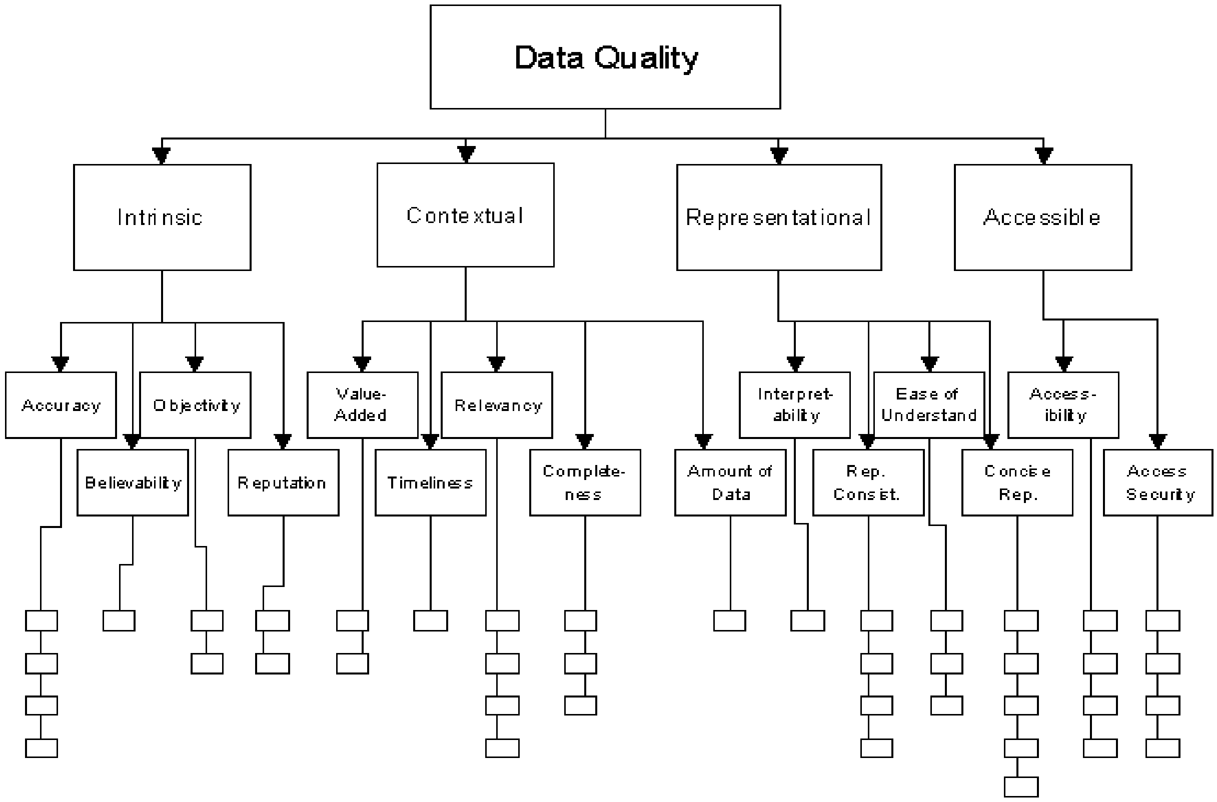


Figure 8: Hierarchy of Data Quality issues

Delone and McLean (2012) described that on one side the data quality is of considerably subjective and must be treated ideally or variedly for every user. At the same time, the aims of quality of involved stakeholders are significantly varied in nature. They can neither be achieved nor assessed directly but needs a critical prediction, measurement, and configuration techniques, always an interactive process form. Moreover, the reasons for reachability, data deficiencies and non-availability issues are objective definitely and rely mostly on implementation and definition of the information system. They further highlighted the critical measures that can be taken to effectivity collate data within a warehouse, such as, system quality, information quality, use, user satisfaction, organisational impact as well as individual impact. Forza (2015) explained that furthermore, the data quality prediction for every user must be based on factors of objective quality that are compared and calculated to expectations of users'. The question that emerges is how to arrange the evolution, design, and administration of data warehouse in such a way that all varied and sometimes opposing, user quality needs can be satisfied simultaneously. As several users and data warehouse systems complexity do not allow to attain every user's total quality, another query is how to organise these needs to satisfy them concerning their significance. Typically, this issue is described by the data warehouses physical design where the issue is to predict a group of materialised opinions that rearrange response of user requests and the maintenance cost of the global data warehouse at the same time. The below figure shows the data warehouse quality factors:

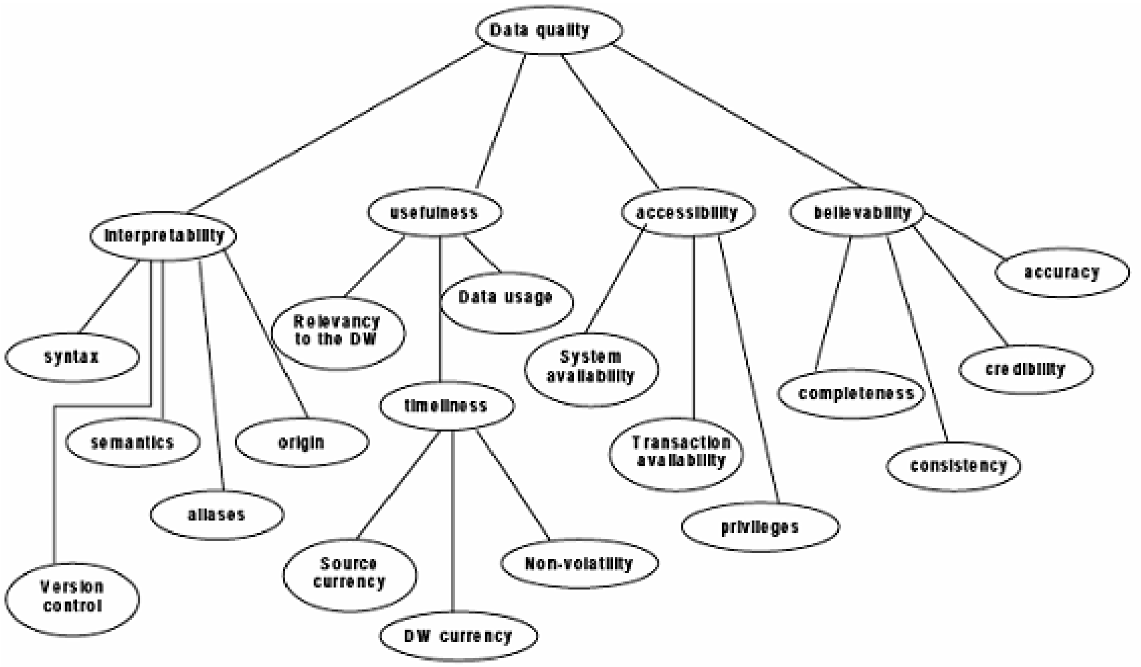


Figure 9: Data warehouse quality factors

Galliers (2013) explained that it must be used to make a clear-cut definition of central concepts in these quality management issues of the data warehouse. The data warehouse processes and data interpretability relies heavily on the design process, i.e., the data description level and the warehouse processes) and the languages and models expressive power which are used. Both the systems and data architecture (i.e. where every piece of information situates and what the system architecture is) are part of the dimension of interpretability. The process of integration is similar to dimensions of interpretability by trying to generate minimal schemata. Similarly, Batini and Scannapieco (2016) described that furthermore, procedures like multidimensional aggregation and optimisation of the query rely on data's interpretability and the warehouse processes. According to Hill et al., (2016), data warehousing is a science that will continuously evolve. Various designs, processes that are introduced have a significant influence on the orientation of the data within the data warehouse. Thus, it is vital that continuous and consistent changes in the hardware and software technology must be pursued which can influence the capabilities of the data warehouse. Data warehousing systems have become an essential component of information technology architecture. A flexible enterprise data warehouse strategy can yield significant benefits for an extended period.

The accessibility quality dimension relies on the type of data sources and the design of the data and the processes of the warehouse. The kind of views stored in a warehouse, the querying processes and update policy are all impacting information's accessibility. Optimization of the query is similar to the dimension of accessibility since sooner the queries are responded higher the availability of transaction is. The data extraction from the sources is also impacting the data warehouse availability. Accordingly, one of the principal aims of the policy of update propagation must be to gain essential data warehouses availability. The warehouse evolution, update policies and the type of data sources are all impacting timeliness and accordingly data usefulness. Furthermore, the dimension of timeliness influences the design of the data warehouse and information querying stored in the warehouse. The data warehouse believability is influenced obviously by the data believability in the sources. Furthermore, the desired believability level impacts the design of processes and views of the warehouse. Consequently, the integration of source must take into account the data believability, whereas the design process of the data warehouse must also take into account the processes believability. The data warehouses processes validation is another problem similar to each activity in the surroundings of a data warehouse and specifically with the design process.

Bouzeghoub et al., (2016) described that within the data warehouse, unnecessary information could be employed from the optimisation of a query, aggregation, and customisation processes to acquire information quicker. Also, the problems are replications resonate with the activities. Finally, aspects of quality impact many data warehouse design factors. For example, the needed space of storage can be impacted by the value and amount of required quality indicators (believability, time indicators etc.). Furthermore, issues like query optimisation improvement through the usage of quality indicators, incomplete information modelling of the sources of data in the data warehouse, the adverse effects schema reduction evolution has on quality of data. The expansion of models of data warehouse and languages, to make better use of quality information has to be reviewed by the data managers.

Pandey (2014) aimed to analyse the issues about the topic of data quality within a data warehouse. In the study conducted, Pandey (2014) highlighted the data quality issues at the data sources, data profiling, data staging and data modelling. These stages highlight the intricacies of the functioning of data in a warehouse and the problems incurred in all the mentioned stages. Furthermore, Pandey (2014) suggested various strategies that can help in the reduction of data redundancies, leading up to project redundancies. He posited that maintaining and creating enterprise architecture (EA) is a crucial aspect that can lead to the optimisation of data and ensure maximum quality. He further posited that in addition to applying enterprise-wide data quality disciplines, creating an enterprise data model, and documenting metadata, the data quality group should develop their data quality improvement process.

Data Quality and Capability Maturity Model Levels:

Various maturity models have been developed, but there are only a few that have gained global acceptance. Capability Maturity Models is one such model that has become a standard for rating software developments. The CMM is a framework that describes the critical elements of an efficient software process and presents an evolutionary improvement path from an ad-hoc, immature process to a mature, disciplined one. However, it has been criticised due to difficulty in its implementation. Furthermore, Calvanese (2014) states that capability Maturity Model supports organisations ponder on their present operating processes giving a denotation of maturity level, and decides the quality of their business process. The Capability Maturity Model of Data Warehousing offers the researcher with metrics to rank the efforts of organisations data warehousing. Capability Maturity Model consists of 5 primary levels such as:

Level 0 – Not Accomplished

1st Level – Carried out Informally

2nd Level – Scheduled and Tracked

3rd Level – Well-Defined

4th Level – Controlled Quantitatively

5th Level – Improving Continuously

Capability Maturity Model assumed only 1st level to 5th level.

The below figure shows the Capability Maturity Model Levels for a data warehouse:

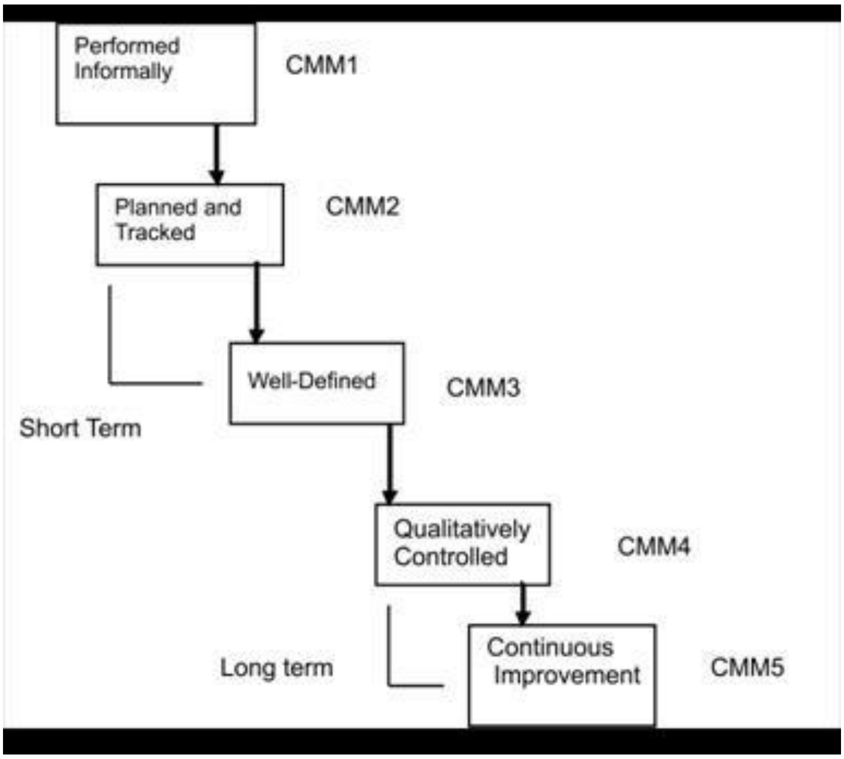


Figure 10: Capability Maturity Model Levels for data warehouse (Calvanese, 2014)

Level 0: Not Accomplished

Chaudhuri and Dayal (2013) described that if any organisation has not constructed a data warehouse or has attempted but failed to construct a data warehouse, then that organisation is at Capability Maturity Model Level 0.

1st Level: Carried Out Informally

Paulk (2014) described that consistent tracking and scheduling of tasks of data warehousing are missing at Capability Maturity Model 1st level. Additionally, the projects of data warehousing are performed complexly, with little standards and sharing/reuse. As an outcome, a single team will construct it is DW in one manner, and another team will construct its DW in a wholly different manner. The organisation is unaware of the problems similar to data quality and mainly treats the quality problems on an ad-hoc basis at this level. Some of the issues in this level 1 are that:

- Initiatives for data quality are chaotic and ad-hoc.
- No formal process and structure of data quality in place.
- Related problems of data quality are not taken into account and are operated as one-off conditions.

Organizations that are at Capability Maturity Model 1st level in data warehousing, typically invest a considerable amount of money. It is essential to understand that investing money in applications of data warehousing will not move the organisation past Capability Maturity Model 1st Level unless it is invested wisely. In fact, the most expensive implementations of data warehousing locate at 1st Level. Unfortunately, a massive number of prominent government organisations and Fortune 500 organisations has DW's that are at Capability Maturity Model 1st level (Howard, 2014).

2nd Level: Scheduled and Tracked

From 1st Level, in this 2nd level, there will be a small improvement. Here organisation admits the following such as 1) significant issues are managed as and when they surface, and 2) issues of data quality are admitted.

3rd Level: Well-Defined

According to Miller (2015), the jump from 2nd level to 3rd level is most critical for a government entity or prominent organisations. At 3rd level, best practices of information technology are performed and documented throughout the organisation. Additionally, deliverables of information technology are transferable and repeatable across the organisation. Here the organisations grant the following such as:

- Quality assessments are completed.
- Initiatives of data quality are moved forward across the organisation.
- Improvement process of data quality is started.
- Process gaps are recognised.

4th Level: Controlled Quantitatively

Hsu et al., (2012) described that the 4th level organisation have implemented measurable process aims for every defined process of data warehousing. These measurements are gathered and analysed quantitatively. At this level, organisations can initiate to find future implementation performance of information technology. Here the organisations admit the following such as:

- All groups of business are involved.
- Groups of data quality are formed.
- Management takes responsibility and ownership.

On the whole, at this stage, the efforts of data warehousing are successful consistently, and an organisation can initiate to forecast the future performance of these efforts accurately. Occurring efforts of the data warehouse are developing data quality and worth to the business.

5th Level: Improving Continuously

Reingruber and Gregory (2014) described that at 5th level, organisations have a qualitative and quantitative understanding of every data warehousing and information technology process. At this level, an organisation understands how every information technology process is similar to the overall goals and strategies of the business of the corporation. For instance, each programmer must understand how every Structured Query Language line would assist the organisation in attaining its strategic aims. At this level, decidedly fewer data levels, technology and process redundancy occur, and the redundancy that does occur is understood and documented. Thus, investments in data warehousing are becoming optimised.

3.2.5 Data Quality Tools

Various studies, as mentioned above, have postulated that data quality is a multi-dimensional concept. The concerned organisations must deal with both, the subjective perceptions of individuals involve with data and the objective notions by the dataset. Thus, various tools must be implemented to assure the quality of the data. The tools of data quality are used in DW to suit the data and assure that specific data from the warehouse, thus developing its usability (Pipino et al., 2012). The tools of data quality are possible to develop the data quality at many stages. Cleansing tools can be essential in enhancing several of the tasks in the developing process of a data warehouse that is involved in correction, data cleansing - parsing, matching, standardising, householding and transformation. Several of the tools focus on predicting data patterns, auditing data, and comparing data to norms of business. Data loading and data extraction tools are possible to convert data from one platform to another platform and form the Data Warehouse. Some of the tools of data quality are described below:

Statistical Analysis System (SAS):

Redman (2012) described that statistical Analysis System data integration offers unique surroundings that combine data quality seamlessly within the process of data integration taking users from norms creation and profiling through controlling and performing outcomes. Organizations can exchange and integrate different data, analyse values, eliminate inaccuracies, standardise universal values, and cleanse worse data to create reliable and consistent information.

Oracle 10g Warehouse builder:

The Oracle database has several features that make it well applicable to data warehousing, including vast databases support, embedded multidimensional online analytical processing engine and developed summary management. Present Oracle versions are existing with built-in ETL (extraction, transformation, and load) features, and it is available to build Oracle data warehouse using SQL*Plus and to use these features (Kahn and Strong, 2014).

SAP business intelligence:

Firth and Wang (2013) state that SAP BI is an organisation complete, class, combined and open solution that supplies actionable insights. SAP business intelligence supports the needs of decision-making of the whole organisation regardless of access methods data sources. SAP business intelligence offers data warehousing, data acquisition, scheduling structure, online analytical processing, dashboards, business intelligence tools and analytical applications; with pre-designed content using better models of practice. SAP business intelligence has well-documented and open interfaces and application programming interfaces along with entire functionality to combine unstructured and structured, heterogeneous data and transforms data into information and assure information is supplied at the appropriate time to appropriate individual in appropriate format to help decision making of business.

3.2.6 State of the art of Data quality

As described by Siddiqqa et al., (2016), several tools of the data warehouse in some manner impact the quality of the data warehouse. However, only a few of them deals explicitly with data quality. The data quality in a data warehouse is affected by three factors. They are:

- Data quality inserted into the data warehouse
- Schema design of a data warehouse
- Data manipulation in the data warehouse.

All these mentioned components impact a specific and distinct aspect of a data warehouse.

Data delivery systems high quality is essential to offer proper access for data customers. Data tagging is an essential way of offering information on data quality to data customers. A degree for usefulness and usability are mostly hugely subjective and consists of time length from last update (for time-sensitive data), stakeholder's beliefs surveys about data warehouse data (subjective rating of stakeholder) and influence on their results and processes of decision making.

Calero et al., (2014) state that in Information systems these approaches study data characteristics regarding real implementation and design concepts such as values, attribute, and entities. Such approaches can be referred to as data-centric as they concentrate on the values and framework of data in a system. They have two significant shortcomings although pragmatic. First, they do not obtain dimensions of data quality from fundamental principles. Second, since these approaches depend on particular concepts of data design, they consider the design briefly should be known before the requirements of data quality can be denoted. Thus, they do not help preceding data quality needs specification. This study analyses data quality in terms that are not data-centric yet are aligned towards system-design. Notably, the researchers suggest severe dimensions of data quality definitions by protecting them in foundations of ontology, and they have revealed such dimensions can offer guidance to systems designers on issues of data quality.

Jarke (2012) described that researchers base their approach on the information systems notion is to offer an application domain representation also referred to as the real-world system perceived by the user. The researchers derived a group of dimensions of data quality from different kinds of representation deficiencies. Thus, in their user's approach views perform as a standard against which data quality is referred.

3.2.7 Data quality Issues in Data warehouse:

Data quality issues are prevalent and a significant concern within the realm of data warehouses which have a significant social and economic impact (Wand and Wang, 1996). Data in a data warehouse is obtained from multiple sources, and hence, there are various changes that can ridicule the quality of the data, as it is difficult to decipher the nuances in the mentioned context. The quality of data can get compromised on the prospect when data is received, entered, integrated, maintained and processed. The given data is influenced by various processes that initiated the advent of data within the respective data environment, the majority of which impacts the quality. All the phases mentioned, in one way or another, impact the quality of the data in a data warehouse. Despite the preventive measures, there exists a certain percentage of data that is not optimised for its quality. There are various ways wherein the data quality problems can occur (Informatica, 2006), such as, inefficient handling of data procedures and processes; failure to stick on to defined procedures and process; failure to adhere to data entry and maintenance procedures; prevalent errors in the process of migration from one system to another; external data that does not correlated with the standards set by the company. Thus, data quality issues can take place at any stage of the realm of data warehousing, that is, in data sources, integration and profiling, data staging in ELT and database modelling. The following pointers mention the various data quality issues that can be of existence within a data warehouse and significantly limit and influence the quality that must be maintained efficiently. The figure below shows the areas of data quality can be compromised in the data warehouse.

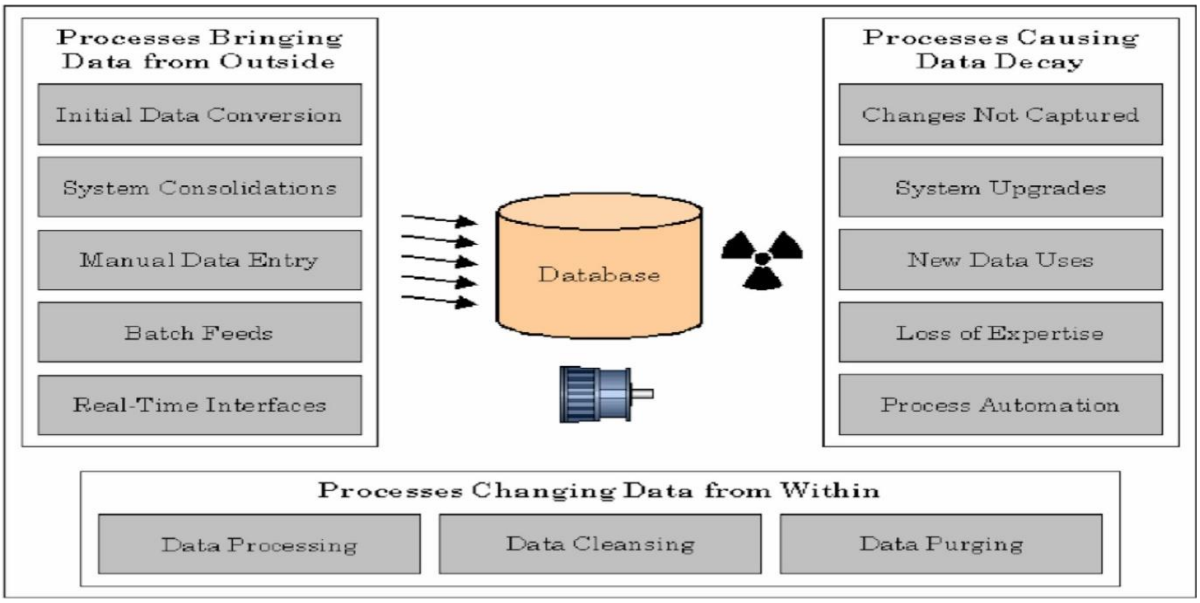


Figure 11: Potential Data quality issue areas (Madrick et al., 2009)

Entry Quality:

Entry quality is the most straightforward issue to recognise but is always most critical to correct (Smith, 2013). In most cases, entry problems are caused by a human being entering data into a system. The issue may be a willful or typo determination such as offering an incorrect address or phone number. Recognizing these missing data is completed merely with simple queries or profiling components. The price of entry issues relies on the user. If an email address or phone number is used only for purposes of information, then the cost of its absence is probably less. Instead, if a phone number is used for promoting and motivating new sales, then the opportunity cost may be essential over the main percentage of records. At the source address, data quality can be critical. If data was sourced from a 3rd party there is usually small the organisation can do. Likewise, applications that offer internal data sources might be costly and old to modify. To conclude, in the simplest form, entry quality relates to whether the information enters the system correctly at the point of origin (Singh and Singh, 2010).

Process Quality:

Usually, a process quality problem exists as data is forwarded to an organisation (Erdmann, 2014). They may become obvious from a system crash, technical existence, or any lost file from combined systems. These problems are usually difficult to recognise, especially if data has made several transformations on the way to its destination. Usually, process quality can be set properly once the source of the issue is recognised. Appropriate verification and quality control at every touch-point along the path can support assure that issues are rooted out, but these verifications are always absent in processes of legacy.

Identification Quality:

Das et al. (2011) described that the issues of identification quality outcomes from a fault to identify the relationship between two things. For example, two common products with different stock keeping units are concluded to be similar incorrectly. Identification quality may have essentially related prices such as mailing the same document more than once. The processes of data quality can hugely avoid this issue by identifying duplicates, matching records, and placing a belief score or resemblance of records. Uncertainly scored records can be judged and reviewed by a data steward. Still, the outcomes are never absolute and deciding appropriate business norms for matching can involve error and trial.

Integration Quality:

Integration quality can represent huge challenges for big organisations (Scime et al., 2012). They went further to say "Integration quality issues can exist because information is separated by departmental or system boundaries". While the wish to have integrated information may seem evident, the reality is that it is not often obvious. Business users who are usually working with one group of data may not be aware that other data occurs or may not understand its value. Scime Master data management enhances the procedure of recognising records from several systems that define a common entity. Then the records are combined into an individual master record. The data warehouse permits the details of operations similar to that entity to be combined so that its relationships and behaviours across systems can be analysed and assessed.

Usage quality:

Usage quality becomes an issue during the development of a data warehouse when the actual implementors of the data warehouse lack access to documentation of legacy source or experts of subject matter (Milea et al., 2013). Without sufficient guidance, data developers are left to guess the use and meaning of specific data tools. Another scenario exists in organisations where users are given the components to write their questions or create their reports. Improper usage may be critical to quantify and predict cost. Robust metadata, thorough documentation and training of users are using and should be constructed into any new initiative but achieving support for metadata's post-implementation project can be critical. Again, this is where the program of data governance should be implemented and adequate effort made to document and recognise data definitions and corporate systems. This metadata can be inserted into processes and systems as it becomes a part of the organisation's culture to do so. This may be more realistic and efficient than a huge bang approach to metadata

Ageing Quality:

According to Welty and Fikes (2016), Aging quality's most challenging aspect is deciding at which point the information is no longer valid. In most cases, such decisions somewhat differ and arbitrary by usage. For instance, maintaining the previous address of a customer for more than five years is impossible. At the same time, managing customer address information for a homeowner's insurance claim may be essential and even needed by law. Such decisions required to be made by owners of the business and the norms should be followed by the solution. Several master data management solutions offer a platform for establishing ageing norms and survivorship.

Organizational Quality:

Similar to entry quality organisational quality is simple to predict and sometimes very critical to solving (Karp et al., 2013). It shares much in similar to integration and process quality but is less a technical issue than a systematic issue that exists in big organisations. Organizations problems exist when for instance, various departments try to combine their evaluations to finance. Financial reporting system takes into account related information which may be varied than how the organisation markets the products or make its customers. These business norms may be holed in several code layers throughout several systems.

Point for Consideration

Corcho and Gómez-Pérez (2015) described that the issues of data quality are difficult for the data warehousing project success. In this study, the data manager had a bright and comprehensive data quality issues understanding and had achieved to some extent in communicating knowledge to other stakeholders. The structure explained in this study offers an essential part in scheduling the establishment of data quality issues awareness among familiar stakeholders. Knowing which dimensions of data quality are essential for specific stakeholders will support data quality management within organisations. In the process of data warehousing, all stakeholders required to understand how they can develop the data quality and therefore higher the perceived data quality level. Data consumers required to be regularly surveyed for their opinions on data quality which they are using. Developed data warehouse usage will lead to feedback on issues of data quality that is required to be addressed.

usage.

Data quality compliance structures

Finally, Moody and Kortink (2014) described that data producers suggested that several issues of data quality specifically the fundamental issues of data quality could be fixed appropriately at the data source. In several cases this was in the files and databases of legacy system applications within the organisation and due to issues, such as coding and interpretation errors at subsequent data transformation or starting data entry. Reward frameworks for compliance with data quality are offered for those consumers who operate starting entry of data must be implemented. The resulting development in intrinsic quality of data will flow through the data warehouse. A data producer also suggests that some data quality issues source was critical remarkably to predict. Documenting process maps and information flows for the information system of an organisation would support data producers.

3.3 Stakeholders and Data Quality Goals

Quality goals must be identified and structured (Jarke et al., 2013). There is a big deal of work similar to the dimension of data quality. Various dimensions of data quality that have been referred are, accuracy (conformity of the stored with real value), consistency (uniform representation of data), timeless (the recorded value is updated) and completeness (no missing information). The data quality definition is modelled through the definition of contextual, intrinsic, representational, accessibility and contextual data aspects. Other factors such as availability, credibility, validation, and traceability are established. In software engineering, many hierarchies of goals of quality factors have been describing including the GE model. ISO 9126 represents six fundamental factors that are refined further to an overall 21 quality factors. In a similar presentation of these three models is provided and the SATC software quality model is described along with metrics for all their software quality dimensions. A structured overview of these strategies and problems embedded in a repository structure has been explained. It is suggested that the data quality dimensions establishment can be gained systematically in 2 possible ways. The first is the use of a scientifically grounded approach to gain a rigorous definition. The second way is to implement dimensions of data quality is the pragmatic approach. A combination of both of these approaches is followed. Using the above-described quality factors nominated in data and software engineering the major stakeholder groups are linked with these factors involved in the projects of data warehouse thus deriving prototypical aim hierarchies for each of these user roles (McClanahan, 2014).

Naiman and Ouksel (2015) described that usually, the decision maker employs an Online Analytical Processing query component to get responses of interest. A decision maker is interested in the stored data quality, their ease of querying them and timeliness through online analytical processing components. The data warehouse administrator requires facilities such as metadata accessibility, timeliness of data knowledge and error reporting to predict reasons and alterations for them or issues in the stored information. The data warehouse designers require measuring the schemata quality of the surroundings of the data warehouse and the metadata quality as well. Furthermore, the designer of a data warehouse requires standards of software evaluation to check the packages of software that are being assumed for buying. The Data warehouse tools programmers can make better use of implementation standards of software to achieve an estimate their work. Metadata reporting can also alleviate their work because they can eliminate errors similar to schema information. Based on this analysis the varied roles represent a diverse gathering of quality dimensions which a quality model must be capable of denoting in a meaningful and consistent way. In the following the quality dimensions of 3 stakeholders have summarised the decision maker, the programmer, and the data warehouse administrator (Radan, 2014). The below figure shows the design and administration quality dimensions:

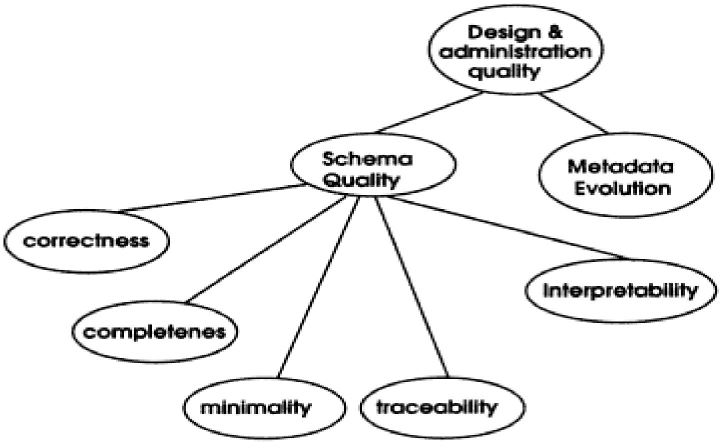


Figure 12: Design and Administration Quality Dimensions (Jarke et al., (2013)

Design and Administration Quality:

Jarke et al., (2013) described that the design and administration quality could be analysed into more brief dimensions as depicted in the above figure. The schema quality defines the capability of the model or schema to denote the information efficiently and adequately. The correctness dimension is concerned with appropriate entities comprehension of the real world the schemata of sources and the user requirements. The completeness dimension is concerned with the security of all problematic knowledge in the schema of the data warehouse. The minimalist dimension represents the degree to which undesired redundancy is eliminated during the process of source integration. The traceability dimension is concerned with the fact that all needs of designers, users, managers, and administrators should be traceable to the schema of the data warehouse. The interpretability dimension assures that all tools of the data warehouse are well explained to be easily administered. The metadata evolution dimension is concerned with the way the schema expands during the operation of the data warehouse.

Data usage quality:

Calero et al. (2011) state that since databases and data warehouses are built to be questioned the most basic procedure of the warehouse is the querying and usage of its data. The below figure shows the quality dimensions hierarchy similar to data usage:

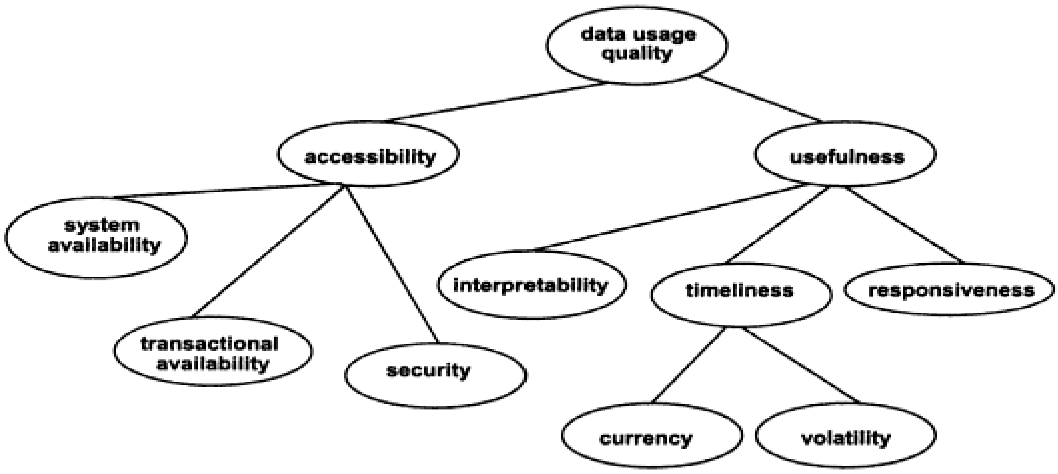


Figure 13: Data usage quality dimensions (Helfert, 2012**)**

Hammergren (2014) explained that the accessibility dimension is similar to the availability of accessing data for querying. The security dimensions explain the authorisation policy and charters every user has for the data querying. System availability explains the percentage of time the data warehouse or source is available. The transactional availability dimensions explain the percentage of time the warehouse information or the source is possible due to the absence of update processes which write lock the data. The usefulness dimension explains the temporal features of data as well as the system’s responsiveness. The responsiveness is concerned with the interaction of a user process. The currency dimension explained when the information was entered in sources and the data warehouse. The volatility dimension explains the period for which the information is possible in the real world. The interpretability dimension explains the extent to which the data warehouse is efficiently modelled in the information repository including the data lineage query.

Training in the usage and content of the data warehouse is essential:

Similarly, Motro and Smets (2015) described that in data quality training for all stakeholders in the process of data warehousing is essential for developing data quality. On one side training must be offered to enhance a comprehensive understanding of the data quality scope as referred by the structure used in this study. More real training must be offered to specific stakeholders in the data quality chain which is most similar to them. Primarily, data consumers must be trained in data warehouse content to higher the usage of the data warehouse. Data custodians and data providers required to be aware of significant issues of data quality and strategies for denoting those issues. The programs of training must be organised be pro-actively rather than being demand-driven. In this study, programs of training were offered on data consumer's request. While the active data consumers were aware of issues of data quality and contents of the data warehouse, a more proactive training program would motivate widespread data warehouse

Data warehousing stages susceptible to issues of data quality:

The below figure shows the data warehousing stages susceptible to issues of data quality:

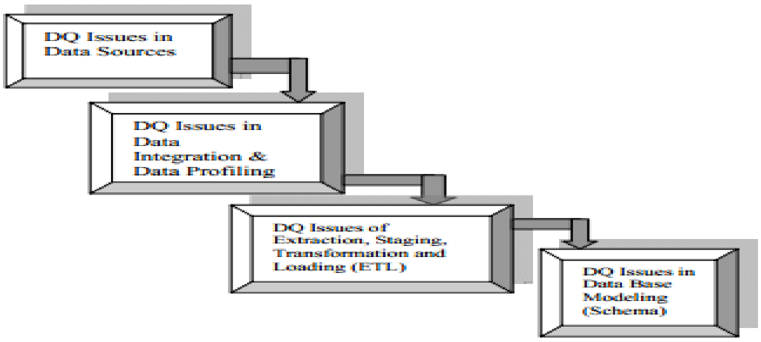

DQ Issues in Data Sources
DQ Issues in Data Integration & Data Profiling
DQ Issues of Extraction, Staging, Transformation and Loading (ETL)
DQ Issues in Data Base Modeling (Schema)

Figure 14: DW stages susceptible to issues of data quality (Wand et al., 2013)

In the above figure, the stages of data warehousing susceptible to data quality issues are:

- Data Source

- Data Profiling and Data Integration
- Extract, transform and load and data staging
- Database scheme (Modelling)

Data is impacted by numerous processes that bring data into the data environment, the majority of which compromises on its quality. All these phases of data warehousing are responsible for data quality in a data warehouse. According to Singh and Singh (2010), during the extraction phase, the issues that can reduce the quality of the data can be, heterogeneous data sources having own storage methods, imperfect schema level definition, insufficient source data analysis and undocumented alterations. Moreover, during the transformation phase, data quality issues could be insufficient source data analysis, the application of business rules that impact the quality of the data, unhandled null values in ETL process, inaccurate conditional statements, and undocumented alterations. Moreover, during the loading phase, the issues faced are inclusive of lack of periodical refreshments of integrated data, incorrect mapping of data leading, lack of error reporting, validation and metadata updates, and inappropriate handling of rerun strategies. Lastly, the issues faced during database modelling are incomplete analysis for schema design, late arriving and multivalued dimension caused data quality issues and delayed identification of slowly changing dimensions (Singh and Singh, 2010). Thus, there are various problems that impact the quality of the data, and thus should be avoided to maintain data quality in a data warehouse to ensure productivity.

Rudra A and Yeo E (2014) described that data Quality could be settled relying upon how data is entered, combined, gained, processed, loaded, and maintained (Extracted, Transformed, and cleansed). Data is influenced by several procedures that bring data into data surroundings most of which cause its quality to some extent. All these data warehousing phases are responsible for data warehouse data quality. Despite all the efforts, still there occurs specific % of worse data. This residual worse data must be reported, representing the reasons for data cleansing failure for the same. The issues of data quality exist in several varied ways. The most similar issues include:

- Failure to stick on to maintenance and data entry processes;
- Poor data handling processes and procedures;
- Obstacles in the process of migration from one system to another system; and
- Third party and external data that may not fix with the organisation's data standards or may be of uncertain quality (Dung, 2014).

The considerations undertaken are that the issues of data quality can emerge at any data warehousing stage, i.e. in data profiling and integration, data sources, in Extract, Transform and Load, in data staging and database modelling. The database modelling is describing available stages which are susceptible to getting issues of data quality.

Schema design of data warehouse:

Gupta (2012) described that the data warehouse schema design is responsible for a meaningful, correct, and complete combination of sources. If the process of design fails to include all the needed information in the schema of the data warehouse, then the data may be incomplete or even unambiguous. If the source data semantics is misinterpreted or if different sources are not integrated appropriately then the data warehouse will consist of incorrect data. Also, if the design process does not recognise the needed constraints of integrity the data warehouse may store incorrect or meaningless information. The design process causes all the dimensions of quality of meaningfulness, completeness, correctness, and unambiguousness. The data warehouse schema design is a complicated process involving the analysis of needs, available data analysis, schema integration and extraction and other general database design steps. The components which may assist in this process belong to the following classifications such as 1) Data Modelling; 2) Management of metadata; 3) CASE tools; 4) Data reverse engineering; 5) Database design; and 6) schema integration.

Schema design of data warehouse:

Gyssens and Lakshmanan (2014) mentioned that in a data warehouse data are usually managed by a database management system and cannot be updated by users. The most similar manipulations are multidimensional data and aggregations reorganisation which are undertaken by the database management system. This means that the data quality is secured inside a data warehouse and it is caused hardly by the manipulation processes. In several cases, only components used to manipulate data in a data warehouse belong to the following classifications such as 1) Multidimensional DBMS, and 2) General purpose DBMS. However, the technique of scoreboard or checklist could be further implemented for the real-time scenario and can be updated on a regular basis.

Blaschka et al. (2015) posited that to understand the multi-dimensional aspect of data quality among various group of stakeholders, it is essential to comprehend varied groups and develop ways for its enhancement in the quality The structure for understanding data quality needs of stakeholders' integrates the property, stakeholder, measure and improvement strategy concepts from structure for understanding data quality with the classification and the concepts of data quality dimensions from data quality structure. Four stakeholders are responsible for the handling and usage of data within the data warehouse; these four stakeholders have been considered for this thesis as well, namely, data managers, data producers, data custodians and data consumers. The data quality dimension and property concepts are universal although the property is not referred. The data quality dimension definition is a group of attributes of data quality that denotes an individual aspect or construct of data quality is acquired. Improvement strategies are the procedures used to acquire dimensions of data quality, and a measure is a systematic way of estimating dimensions of data quality. The structure tools and their interrelationships are shown in the below figure:

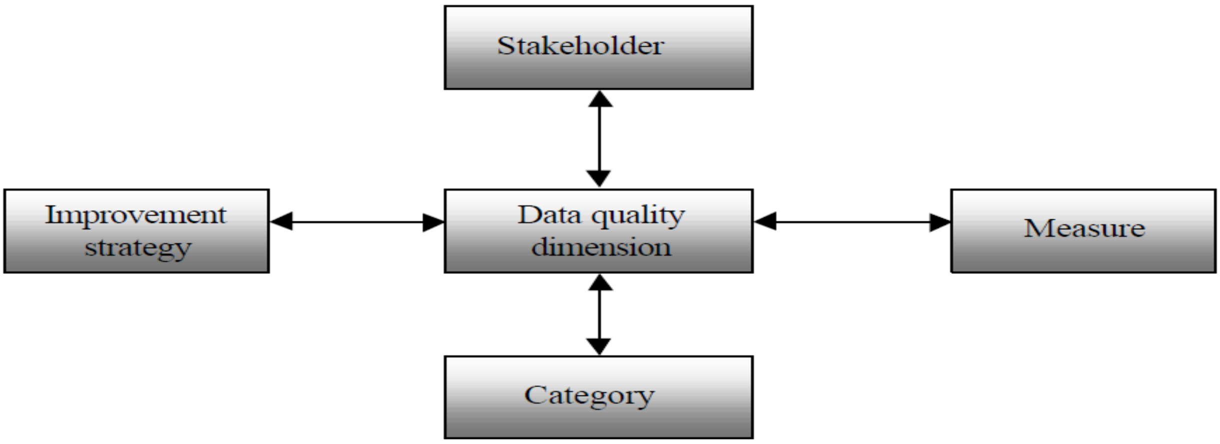
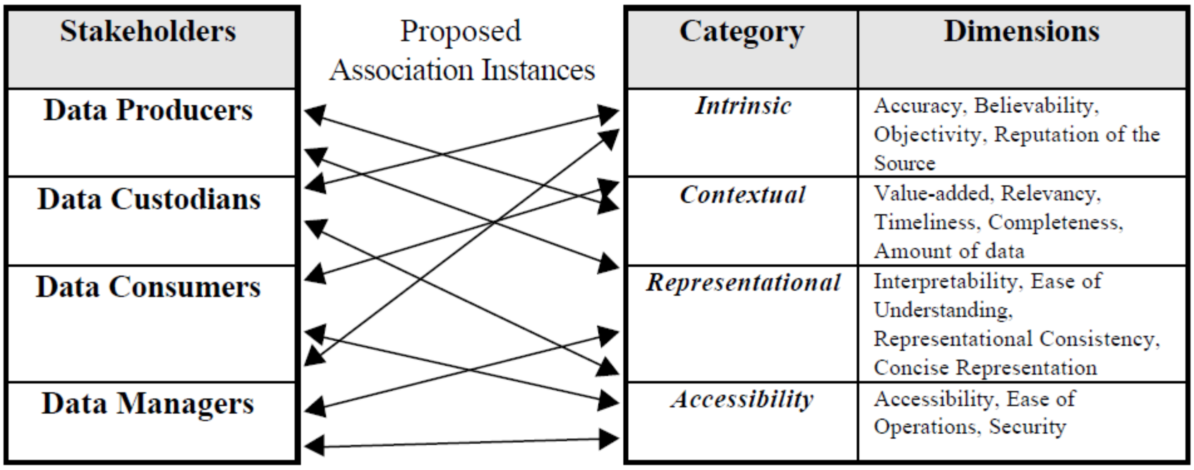


Figure 15: Structure of understanding relationships between stakeholder groups and data quality dimensions in data warehouse surroundings (Wang et al., 2013)

Chaudhuri and Dayal (2013) described that varied stakeholders might have varied data quality perspectives. The structure relates stakeholders to dimensions of data quality to help recognition of varying needs of data quality. Dimensions are similar to measures to enhance data quality evaluation by stakeholders, and improvement strategies offer for recognition of tasks that will provide higher quality data regarding particular dimensions. The classifications offer a means of arranging several dimensions of data quality that have been recognised into a categorised scheme that is an essential characteristic of data quality that is essential for data consumers. The structure offers a group of concepts that can be used as a basis for experimental studies of the stakeholder's quality improvement processes and data quality needs in practice. The structure's three tools are category; stakeholder and the dimensions of data quality were used to estimate different stakeholder's data quality needs nature in data warehouse surroundings. Representing the possible data warehouse and data quality particular relationship examples between specific kinds of stakeholders in the structure and specific dimension categories in the framework were explained. These were estimated which investigates different data quality dimensions perceptions of stakeholders which is essential for customers (Madnick

2014). The below figure shows the instances of the relationship between stakeholder kinds of data quality

dimensions and classifications:



Category	Dimensions
Intrinsic	Accuracy, Believability, Objectivity, Reputation of the Source
Contextual	Value-added, Relevancy, Timeliness, Completeness, Amount of data
Representational	Interpretability, Ease of Understanding, Representational Consistency, Concise Representation
Accessibility	Accessibility, Ease of Operations, Security

Figure 16: Relationship between stakeholder kinds of data quality dimensions and classifications (Cipriano, 2015)

According to Weir et al. (2013), it was assumed likely that data producers would acquire representational and contextual classifications of data quality to be essential in specific ease of understanding and amount of data as they are applicable for processing and producing data in specific conditions. It was assumed that data custodians would be interested in the classifications of intrinsic and accessibility specifically in dimensions of security and accuracy. They apply to offer and handling the resources for accessing, processing, and storing data. To assure data warehouse usage data must be accessible, complete and accurate. Data consumers were assumed likely to understand the contextual and accessibility classifications as most significant species in the dimensions of completeness, accessibility, and timeliness as they are dependent on a data warehouse for the deliverance of accessible, complete and timely data to support other activities and their decision making. Incomplete data is the primary factor in the complaints of data consumers' about data insufficiencies. Data managers were assumed likely to acquire the accessibility, intrinsic and representational classifications as essential specifically in dimensions of interpretability, security, and accuracy. They are possible for handling data warehouse operation including the representation, accuracy, and security of the data.

Garcia-Molina (2015) described that the project team of data warehouse includes a data warehouse system supervisor, project manager i.e. data manager, responsible for data in data warehouse i.e. data custodian, many analysts of data production responsible for transforming and sourcing data i.e. data producers, and managers of business solutions responsible for communicating with business units using data warehouse i.e. data consumers. Within the data warehouse project team data quality was an essential assumption and several principles of data quality principles were improved such as 1) in the data warehouse data is allocated a business area liable for its quality; 2) issues of data quality must be solved in source systems where applicable; and 3) in the data warehouse data is published after being verified by the liable business area.

3.3.1 Stakeholder Data Quality Perception

Despite the studies conducted in the context of data quality, only the study conducted by Shanks and Drake (1998) researched the perception of the stakeholders. Different stakeholders have different perceptions of data quality and act accordingly. Despite their subjective perceptions and preferences towards some aspect of data quality, they all attest towards the importance of maintaining data quality.

Data Producers:

Eppler and Wittig (2014) described that the organisation's information system data was sourced and was analysed carefully before data warehouse uploading. Referential integrity and consistency of data representation were viewed as very essential. It was predicted that the data consistency had an essential effect on the perceived data quality. If the data customers predict that their data has consistency issues, then there is an actual hazard which they will lose faith in the data warehouse. The believability and accuracy of data were also viewed as essential. All data's of Data producers are system produced, and they must be alert to assure that data's believability, and accuracy remains high. Usually, data producers are the ones who create or collect data for the data warehouse. Thus, for them, the following aspects of data quality are of utmost importance, namely, concise representation, accuracy, believability, and relevancy.

Data producers were anxious that there were no programs of incentives in place at the organisation to develop the data quality at its source. It was indicated that a reward framework for agreement of data quality would improve higher the reputation and believability of source data and therefore higher the data quality perception in the data warehouse. Data customers information needs were gathered carefully to assure that data loaded into the data warehouse was common. Therefore, data producers had real unity with the representational and intrinsic classifications of data quality and a weaker unity with the classification of contextual data quality.

Data Custodian:

Pipino et al. (2012) described that the data warehouse design was related to data custodian, being exchanged between the stored data volume, and gaining the desired granularity level for drill-down questions. Another concern area was assuring that the reputation and accuracy of data were assumed high. Database integrity and edit checks constraints were occupied to ensure consistency of data which remains high. Timeliness and relevancy were considered as significant data quality dimensions. The data needs to be sound contextually to fulfil the needs of diverse departments within the organisation. Therefore, the data custodian had real unity with the contextual and intrinsic classifications of data quality.

Data custodians are those who design, develop and operate the data warehouse. For them, accuracy, relevancy, reputation, and timelessness constitute to be the essential dimensions for data quality.

Data Consumers:

Typically, data consumers are business analysts who need accurate and brief data to make sound and proper decisions of business (Lee, 2014). In general, data consumers agreed that their data quality perception was familiar to the activities which they required to execute. An obstacle to examining data was perceived as an essential concern. Data accessibility consists of both the requirement for simple access to needed data and the ease with which data can be operated, i.e. formatted and aggregated. The classification of representational data quality was viewed as very essential. Data consumers could demand training about using data warehouse at any time. The training concerned about the data warehouse usage and content. There were no appropriate programs provided periodically throughout the organisation. Therefore, data producers had real unity with the accessibility, representational and contextual classifications of data quality.

Those who consume the data for their work activities are called data consumers. They focus on various dimensions of data quality, such as accessibility, relevancy, timelessness, access to security, accuracy and representational consistency.

Data Managers:

Hull R and Zhou G (2015) described that the data manager had a general opinion of data quality. Relevance was viewed as the most essential data quality dimension and after that precision and well-classified data which reduces subjectivity on the part of those who gather the data. The data manager denoted that if the classification of data is not represented in a meaningful way, then the data aggregation mechanisms use loses their appeal. The problem of accessibility consists of ease of access to standard data, ease of calculation of that data and sufficient training. Completeness was also viewed as essential. All the critical values must exist to assure the needs of data consumers are attained. The high data quality perception by data consumers was viewed as very essential. If data consumers do not assume data to be of higher quality, then they will use less data warehouse if at all. Due to his extensive data quality perspective, the data manager had real unity with all four classifications of data quality such as accessibility, intrinsic, representational, and contextual classifications of data quality.

Data managers are entitled to manage all the processes that take place in the data warehouse. Aspects such as accessibility, completeness, accuracy, relevancy, concise representation are of utmost importance within the realm of data quality for data managers.

3.3.2 Stakeholder Data quality Concern Areas

Syntactic Data Quality:

Syntactic data quality concerns data's structure. The aim for syntactic data quality is consistency where the values of data for specific elements of data in the data warehouse surroundings use a consistent symbolic representation (Ballou et al., 2012; Wang et al., 2014). This may be within an individual data file where all the values of data must conform to strict universal definitions of data type or between files of data where values required to be consistent to allow comparison and consolidation (Mattison, 2014). Specifically, consistency is essential in schemes of coding throughout an organisation, for instance, part codes, region codes and customer codes.

This means to ensure consistency is to have a formal and well-defined syntax for entire elements of data. The development strategies consist of corporate data model development with syntax norms for elements of data having an organisation-wide perspective. This is especially essential in data warehouse surroundings where similar syntax norms are essential for cleansing, sourcing, and loading legacy systems data (Inmon and Hackathorn, 2014). A 2nd development strategy is to have automatic syntax verifying at the entry of data or to have producers of human data well beached in the norms of syntax. A consistency measure is to represent the ratio of several inconsistent values of data to the total number of values of data for every element of data in the data warehouse. Thus, to conclude, syntactic data quality emphasises the structure of symbols and focuses on form rather than content. The syntax is inclusive of valid syntactic categories and the rules that govern their formation. If the syntax is defined, then the symbolic forms can be converted to other symbolic forms. The goal for syntactic data quality is consistency wherein the data values for particular data elements in the data warehouse use a consistent symbolic representation (Ballou et al., 1996). Consistency is of the utmost essence when defining coding schemes in any given organisation (Shanks and Darke, 1998).

Semantic Data Quality:

Semantic data quality concerns the data meaning. The semantic quality aims are accuracy and comprehensiveness (Ding et al., 2015). Comprehensiveness is concerned with the extent to which for every similar state in the actual world system there is a value of data in the data warehouse. Accuracy is concerned with how well the values of data in data warehouse correspond to the real-world state. As every stakeholder may have varied prior experience and knowledge, varied stakeholders may have varied opinions on accuracy and comprehensiveness of data warehouse.

The essential properties to assure accuracy and comprehensiveness are consistency which is the aim of syntactic data quality and the dimensions of data quality. These dimensions are referred by analysing mappings between data warehouse symbols and understanding of stakeholder of actual world events and structure. The dimensions are meaningful, complete, correct and unambiguous. The mapping between the values of data and actual world system in a data warehouse must be comprehensive for data in the data warehouse to be absolute. No two real-world system states should be mapped into similar values of data of data warehouse for data to be unambiguous. There should be no values of data in a data warehouse which cannot be mapped to a state of actual world system for meaningful data. The states of the real-world system must not be mapped onto incorrect values of data in the data warehouse for correct data. According to Ding et al. (2015) the development strategies to gain an accurate and comprehensive data warehouse consist of training the data producers in the significance of accurate and comprehensive data. Another essential strategy is to reduce the number of transcriptions of data and transformations of data from when the data is obtained first until it is preserved in the data warehouse. Measures for ambiguity, correctness, completeness, and meaningfulness consists of surveying samples of the population and comparing values of data in real-world system states data warehouse.

Pragmatic Data Quality:

Hopfgartner and Jose (2010) described that pragmatic data quality concerns the data usage. The realistic quality goals are usefulness and usability. Usefulness is the measure to which the data helps stakeholder in fulfilling their activities within an organisation's social context. Usability is the extent to which every stakeholder is capable of using and accessing the data warehouse data efficiently. Usefulness and usability will differ among varied stakeholders due to their varied interpretations of data values meaning and varied activities in nature. The properties essential to assure usefulness and usability consists of accuracy, consistency, and completeness (the semantic and syntactic goals of data quality), ease of understanding, timeliness, reputation, conciseness, and accessibility. For the activity at hand, timeliness is the degree to which information is up to date. Accessibility is the ease with which data can be manipulated and extracted. Ease of understanding defines the measure to which data warehouse data is understood by stakeholders. Reputation is the degree to which data is regarded as highly regarding its content, source, and credibility. The table below shows the Properties, goals, measures, and improvement strategies summary:

Level of Semiotic	**Goal**	**Property**	**Improvement Strategy**	**Measure**
Semantic	Accurate and comprehensive	Meaningful, correct, complete and unambiguous	Reduce data transcriptions and transformations, Data producers training	Errors percentage in population or data sample
Syntactic	Consistent	Well defines syntax	Syntax verifying, data producers training and corporate data model	Inconsistent values of data percentage
Pragmatic	Useful and usable	Concise, reputable, timely, understood and accessed easily	Visualization and description, data consumers, data tagging, delivery systems of high-quality data	User surveys, time of update and effect on decision-making results and processes

Table 3: Measures and improvement strategies summary Shanks and Darke (2014)

In subsequent sections, the researcher conducts a literature review to find out from literature the recommendations for designing an effective data quality scorecard. The recommendations are then subjected to a systematic literature review. The intricacies of the review are presented in sections 3.5 and 3.6 below.

3.4 Designing an effective Data Quality Scorecard – DQS

In this section, the researcher discusses factors that need to be considered when designing a DQS. Previous studies by (Kaplan and Norton, 2001; Lawrie and Cobbold, 2004; Chang, 2006; Brace, 2008; Coe and Letza, 2014) made various recommendations for the design of an effective scorecard, the following five factors have been considered from the recommendations by the researcher for the design of the DQS; 1) use simplicity, 2) time efficiency, 3) electronic/online features, 4) non-technical language and a 5) Intuitive approach.

1) **Use simplicity:** Luther et al., (2015) recommends a scorecard design that is simple and straightforward to use. They further state that a correlation exists between the frequency of use of a scorecard and the simplicity of design. They argue that the use of colours, drop down menus and graphs aids the design simplicity. To aid the simplicity of usage and immediately grasp the users understanding, Swain, (2015) advocates the use of 'traffic light' indicators. The figure below shows the traffic light indicator used by Swain. The data warehouse stakeholder pragmatic requirements must be considered in the design of the DQS.

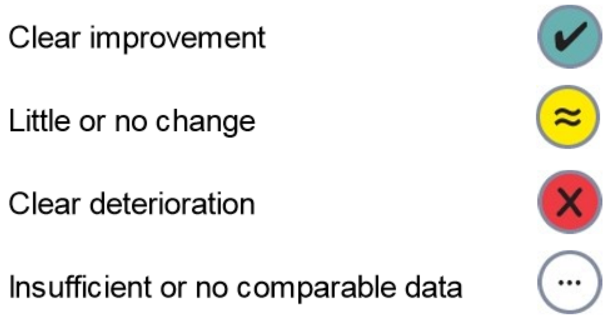


Figure 17: Traffic light measurement indicator. (Swain, 2015)

However, literature highlights some disadvantages in the use of the traffic light measurement indicator (Anderson 2002; Sun and Li 2004).

The possibility to lose some information is higher as the measurement is anchored on a three-point scale. The traffic light scoring scale is essentially a three-point scale, which according to literature, does not maximize discriminability. With more options, it would be possible to better discriminate on the measured concept (Allen and Seaman, 2007).

2) **Time efficiency:** The design of a DQS needs to consider the time constraints of the various DW stakeholder groups. Mislevy et al., (2017) argues that the less time the users have to spend on completing a scorecard/assessment tool increases the future usage of the tool by the user.
3) **Electronic/Online Features:** Keyes, 2016 argues that for modern scorecards to be effective with a higher ratio of usage, an electronic and or a web-based version needs to be built into the design. She went on to state that one of the major criteria in designing a scorecard should be electronic and web-based deployment to the users. Therefore, in the current era of technology-driven media, having a none electronic and web-centric design would reduce the adopt of the DQS.
4) **Non-technical language:** The design of a scorecard should be such that all members of the stakeholder group should not need an added technological training before usage. Mislevy, (2017) states that design patterns should be non-technical and should focus on the skills and abilities of the target users. One of the important points raised by the author is that the users do not necessarily need to be educated on the technical workings of a data warehouse, the terminologies are best suited to the system administrators and technical consultants.
5) **Intuitive approach:** Kahtri et al., (2000) surveyed senior managers of companies representing a computer, banking, and utility industries in the United States and found that intuitive processes are used often in organizational decision making. While intuitive decisions might be a departure from the norm for the majority of scientist,

studies have shown that an element of intuition needs to be planned into the design. The authors also found out from the analysis of their survey that the use of intuitive synthesis was found to be positively associated with organizational performance mostly in an unstable environment.

3.5 Existing Scorecard Design – Systematic Literature Review

Background

This section presents the results of a systematic literature review (SLR) of publications regarding data quality scorecards/assessment. The systematic review was carried out for the following reasons: 1) To summarise the existing evidence concerning the use of scorecards as a useful data quality measurement tool. 2) To identify limitations in current research in the area of scorecard designs and the usage of data quality dimensions, to suggest areas for further investigation. 3) To provide a background to build on for the design of the DQS in subsequent chapters. The systematic review was carried out on papers published from 2007 to 2017. The rationale for choosing this time-range is due to the explosion of 'big data' experienced between this period as well as concerns for proper measurement of the quality of data within the domain. The analysis period was between June 2017 and August 2017.

Research Questions - SLR

The research questions to be addressed by this SLR are:

- What evidence exists concerning the use of scorecards as a useful data quality measurement tool within the timeframe selected?
- What are the limitations of current research?

Search Process

The search process is a manual search of specific conference proceedings and journal papers within the data quality domain between 2007-2017.

The list below details the databases that were searched;

1. **IEEE/IET** is a publisher of computer science and information systems research articles with the highest quality. This database was selected because it contains quality technical literature in engineering and technology which have been published since 1998.
2. **ScienceDirect** is a source for journal articles by millions of researchers with over 10 million journals and articles across a significant number of areas.
3. **Google Scholar** is considered the foremost database for researchers. Google Scholar, provides comprehensive access to conferences, journals, white papers, and books in academia and the industry. Google Scholar serves as an intermediary between the original source and the researcher.

Inclusion criteria

The inclusion criteria used to select the publications analysed are;

1. The publication needs to mention data quality. Some of the scorecards mentioned are not solely for the data quality/data warehouse domain, but their methods of execution are similar.
2. The publication needs to explain how the scorecard works functionally. This is essential as the paper will be evaluated against the recommendations detailed above, hence. the operational procedures must be known.
3. The publication needs to mention the targeted DW stakeholder groups. This is critical as the scorecard is aimed solely for use by this group. The four stakeholder groups, i.e

data producers, data custodians, data managers and the consumer's group, all have varied data quality needs.

4. The publication must either be a journal or conference article.

Exclusion Criteria

The following types of papers were excluded

• Informal literature surveys

• Papers not subject to peer-review.

The Search Results

Research into data quality scorecard/assessment design is vast, with researchers focusing on varied aspects within the domain. The search term initially used by the researcher was; 'data quality scorecard design'. The initial search returned a total number of 22,136 results, broken down as follows:

- Google Scholar – 17,900
- ScienceDirect - 2,903
- IEEE - 4

Fifteen papers appeared in multiple databases, all duplicates were removed. A total of 20 papers was selected from Google Scholar for review, with 14 included.

Ten papers were selected from ScienceDirect, six were selected for review, four were included. Five publications were found using the same search term on IEEE but were all found to be none relevant. The search term was modified to 'data quality assessment design' which returned 932 results, of which 8 were reviewed, and were all excluded based on the inclusion criteria.

Likewise, the search term ‘data quality assessment design' was used on Google Scholar and ScienceDirect databases which returned 1,310,00 and 349,332 results respectively. The researcher however discovered after reviewing a number of papers that the initial search carried are all included in the modified search. A total number of 18 articles were included in the systematic literature review.

3.5.1 Systematic Literature Review Analysis

This section presents the analysis of the systematic literature review on designing a data quality scorecard. The data will be tabulated (ordered alphabetically by the first author name) to show the basic information about each study.

The table will be reviewed to answer the research questions and evaluated based on the guidelines recommended for designing an effective DQS. (See section 3.5). The recommendations are represented in the analysis as stated below;

Recommendation	Denoted As
Use simplicity	A
Time efficiency	B
Electronic/online features	C
Non-technical language	D
Intuitive approach	E

Table 4: Representation of Recommendation in analysis

In Table 5, the number 1 denotes YES, i.e. the reviewed article complies with the recommended guideline, while the number 2 denotes NO, i.e. there was no evidence any of the recommended guidelines were followed. Furthermore, to answer the research questions, the following will be done;

Question 1: What evidence exists concerning the use of scorecards as a useful data quality measurement tool within the timeframe selected – A count of the total relevant papers will be presented

Question 2: What are the limitations of current research? – A detailed limitation analysis of the system in relation to the DW stakeholders and the design recommendations is presented in section 3.5.4 below.

Table 5 below shows the findings from the systematic literature review conducted

Table 5: Reviewed Data Quality measurement/assessment systems

Author	Date	Topic Area	A	B	C	D	E	Description of study	Evaluation	Findings
Brockman et al.,	2008	Data quality Score Evaluation	2	1	1	2	2	Describes general methods for improved quality scores and accurate automate detection and apply them to data	None	The quality score derived from an algorithm .
Cai L, Zhu Y	2013	Challenges of Data Quality and Data Quality Assessment	1	1	1	2	2	Comprehensive analysis and research of quality standards and quality assessment methods for big data	None	The paper constructs a dynamic assessment process for data quality
Batini et al.,	2009	Defining methodologies to improve the quality of data	2	1	1	1	2	Techniques to assess and improve the quality of data	None	Addresses data quality dimensions but does not meet all the required needs of the stakeholders
Kauffman et al.,	2009	Assessment of data quality	1	1	2	2	2	Development of a quality metrics to assess reproducibility, identify apparent outlier arrays and compute measures of signal	None	The tool handles most current technologies and is amenable to use in automated analysis or for automatic report generation
Acosta et al.,	2013	Quality assessment methodology	1	1	2	1	2	Implementation of a quality assessment methodology for Linked Data that leverages the wisdom of the crowds in different ways	None	Amenable to a specific form
Bergdahl et al.,	2007	Systematic assessment of data quality	2	1	1	1	2	Data Quality Assessment Methods and Tools (DatQAM) aims at facilitating a systematic implementation of	None	Targets only the Data manager group

								data quality assessment		
Devillers et al.,	2007	Data quality information analysis tools	1	1	1	2	2	Design of a data quality tool that can manage heterogeneous data quality information	None	Combines concepts from GIS and Business Intelligence
Mendes et al.,	2012	Quality assessment methods	1	1	2	2	2	Framework for flexibly expressing quality assessment methods	None	Linked Data Integration Framework (LDIF), which handles Data Access, Schema Mapping and Identity Resolution,
Weiskopf NG, Weng C,	2013	Dimensions of data quality assessment	1	1	1	2	2	Methods and dimensions of data quality assessment in the context of electronic health record (EHR) data reuse for research.	None	Links data quality dimensions to defined quality assessment.
Corso-Radu et al.,	2007	Data quality monitoring	2	2	2	2	2	DQM involves automated analysis of monitoring data through user-defined algorithms	None	An automated system with no online user participation
Kontoknontas et al.,	2014	Quality assessment methods	2	2	2	2	2	Present a methodology for assessing the quality of linked data resources, based on a formalization of bad smells and data quality problems.	None	Datasets are of varying **quality** ranging from extensively curated datasets to extracted **data**
Fan and Geerts,	2013	Data Quality Management	2	2	2	2	2	Dependency theory for capturing data inconsistencies	None	Promotes a uniform logical framework for dealing with quality issues, based on data quality rules.
Madrick et al.,	2009	Data and Information Quality management	2	2	1	1	2	Introduces a framework to characterize the research along two dimensions: *topics* and *methods*	None	Awareness of data quality issues. MIT has made a huge investment in the last decade on DQ research.

Vetrò et al.,	2016	Data quality measurement framework	1	1	1	2	2	Development of a framework of indicators to measure the quality of Open Government Data on a series of data quality dimensions	None	The framework of indicators provides accurate data quality measures
Watts et al.,	2009	Data quality assessment	2	2	1	1	2	A theoretical model for understanding users' contextual information quality assessment processes	None	Data quality model that measures the user's subjective data requirement
Otto et al.,	2007	Data Quality Management	2	2	2	2	2	Framework for Corporate Data Quality Management	None	Tailored to the Data Manager stakeholder group.

3.5.2 Limitations of Existing DQS

The systematic literature review conducted has shown that considerable research exists in the area of data quality assessment, data quality management and monitoring. However, the analysis in Table 4 suggests very minimal research exists when it comes to designing a DQS that is specific to the needs of the DW stakeholder groups. Previous studies have shown that the data quality needs of the various DW stakeholder groups are varied (Chaudhuri and Dayal, 2013). Meeting the data quality needs of the DW stakeholders requires a data quality measurement system that is designed to align their requirements with the data quality dimensions.

Many of the existing data quality measurement tools are designed as automated systems, with predefined parameters, leaving out the interface with the stakeholders. For example, the data quality measurement tool designed by (Corso-Radu et al., 2007) uses automated predefined parameters to measure data quality. The researchers did not consider the data quality perception and transparency attributes of the DW stakeholders.

Due to these limitations, the problem of data quality still persists.

In Chapter 4, this research proposes a conceptual DW stakeholder focused data quality scorecard (DQS), designed based on the identified data quality dimensions (See Section xx), the data concern areas of the stakeholder groups (See Section xx), and the recommendations for designing an effective data quality scorecard (See section 3.5

3.6 Summary

In this Chapter, a literature review of research into the data warehouse domain was carried out. The complexities of data quality measurement were explored. Literature has shown that data quality is a multidimensional concept. Frequently described dimensions are consistency, accuracy, timeliness, and completeness. The option of these dimensions is mainly based on literature review, intuitive understanding, or industrial experience. The data quality produced by an information system relies on system design. Thus, it is essential to offer design-oriented data quality definition that will represent the intended information usage. In quantitative positions, the purpose of estimating data has long been identified. Similar work can be described as design-oriented or theory-based. Literature has shown that a full agreement on the various dimensions does not exist within the academic community. Data quality assessment and management is a vast area of study as shown by literature; however little literature exists in the area of data quality measurement using a customizable scorecard matrix. This research project can move on and build on existing knowledge practically.

Chapter 4: DQS Model Development and Validation – Iteration I

4.1 Overview

In this chapter, a data quality scorecard (DQS) is developed based on the findings from the literature. In chapter 3, a comprehensive analysis of the existing knowledge in the domain of conceptual data quality dimensions and scorecard design was carried out, this was used as the basis for the design of the DQS. Although considerable research has been done in the data quality domain; not much work has been done with the alignment of the varied needs of the stakeholder groups with the DQD in the data warehouse domain.

This chapter is divided into two parts. Section 4.2 introduces the development of a proposed data quality dimensional based scorecard for measuring data quality in the data warehouse aligned to the needs of the data warehouse stakeholder groups. The development of the scorecard is taken from general literature and the systematic literature review conducted in chapter 3. Section 4.3 presents the mechanics of the scorecard. In the second part of this chapter, we conducted an empirical research to validate the effectiveness of the developed DQS with a system usability scale. The second part of the chapter is structured as follows; section 4.4 introduces the data collection method adopted; section 4.5 presents the results of the analysis of the data that was carried out. In section 4.6 the researcher discusses and reflects on the results, while the chapter is summarised in section 4.7.

4.2 DQS Model Development

How good is a company's data quality? Answering this question requires usable data quality metrics (Pipino et al., 2002). The literature review from the previous chapter suggests that data quality issues fall into one of the following concern areas:

Syntactic data quality concerns the data’s structure. The aim for syntactic data quality is consistency where the values of data for specific elements of data in the data warehouse surroundings use a consistent symbolic representation.

Semantic data quality concerns the data meaning. The semantic quality aims are accuracy and comprehensiveness. Comprehensiveness is concerned with the extent to which for every similar state in the actual world system there is a value of data in the data warehouse. Accuracy is concerned with how well the values of data in data warehouse correspond to the real-world state.

Pragmatic data quality concerns the data usage. The pragmatic quality goals are usefulness and usability. Usefulness is the measure to which the data helps stakeholder in fulfilling their activities within an organisation’s social context. Usability is the extent to which every stakeholder is capable to use and access the data warehouse data effectively. Table 6 below shows the selected DQD based on the findings from literature to be used for this study.

Selected Dimensions	Description
Completeness	Completeness is a data quality dimensions which ensures that there are no missing values for the given tuples or attributes in the system. In other words, completeness can be achieved when all the values for certain attributes are entered.

Validity	The data is invalid if it does not have the data items within the pre-specified value attributes. Validity measures the degree to which the tuple has valid data items. In other words, it defines the reasonableness and correctness
Accuracy	Accuracy defines the accuracy of data in the data warehouse. In other words, it measures the degree to which the data warehouse has correct or accurate data items. Accuracy can be achieved when entered value in the data warehouse is in conformity with original or actual value. The accuracy of data can be characterized as the percentage of real-world objects without any data errors such as out of range values, misspellings, etc.
Timeliness	Timeliness is used to measure the age of the data in the data warehouse. Generally, timeliness can be achieved when the value entered in the data warehouse is not out of date. Timeliness is the degree of the extent to which the data is up-to-date for specific purposes at hand. Timeliness measures the time elapsed between when the data was created and updated
Consistency	Consistency is used to measure the degree to which the data in the data warehouse adheres to a pre-defined constraint. Consistency manifest the data degree to which the data satisfies the integrity constraints
Integrity	Integrity can be defined as the practices of the one-time process when the data gets loaded into the data warehouse. Integrity is used to check whether the data is true or not.

Table 6: Selected DQD

The concern areas as detailed from literature can be mapped to the DQD. Table 7 below maps the data concern areas to the relevant DQD.

Data Quality Concern Areas	**Data Quality Dimensions (DQD)**
Syntactic	Consistency
Semantic	Accuracy, Completeness and Timeliness
Pragmatic	Integrity and Validity

Table 7: Mapping of concern areas-DQD

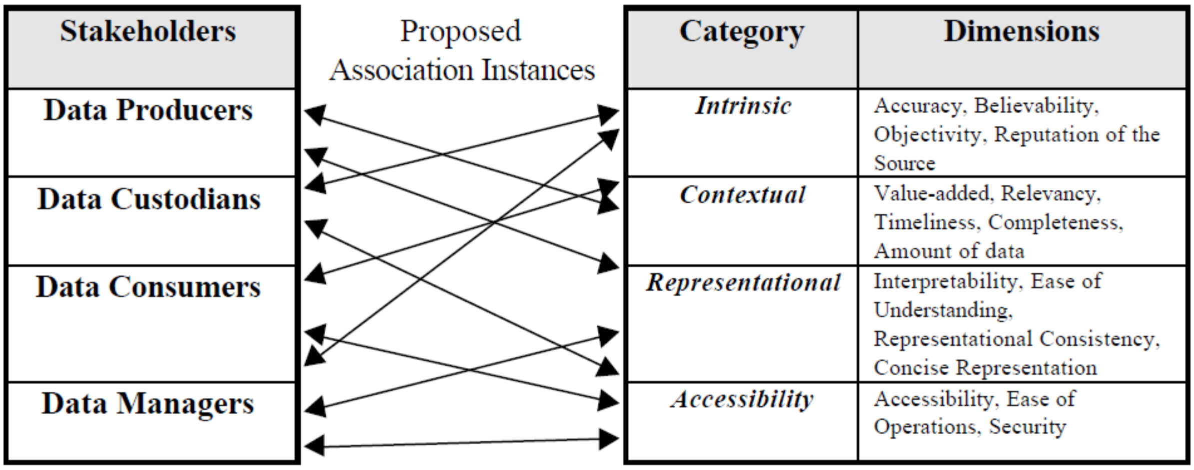
Furthermore, the dimensions of data quality as identified from the literature are a feature or aspect of information and a way to categorize data quality and information requirements. The data quality dimensions are used to measure, define and handle the information and data quality. Remarkably, the findings suggest a misalignment between the data warehouse stakeholder groups, data susceptible areas and the data quality dimensions. This misalignment shows to some extent why data quality issues persist in organisations. Table 8 below provides an alignment map based on the findings from the literature.

Data Quality Stakeholders	**Data Quality Concern Areas**	**Data Quality Dimensions (DQD)**
Data Producer	Syntactic	Consistency
Data Custodian		
Data Manager	Semantic	Accuracy, Completeness and Timeliness
Data Consumer	Pragmatic	Integrity and Validity

Table 8: Alignment of Stakeholders-DQ concern areas-DQD

In this research context, DQS focuses on providing an efficient data quality measure by ensuring an alignment between the DQD with the needs of the individual stakeholder groups. As shown in table xx above. While the literature has shown that previous studies have laid more emphasis on the design of a more system based rather than an interactive data quality assessment tool, the current research argues that the stakeholder group are more likely to use the DQS if it is targeted to their data quality concerns as detailed above. Therefore, the DQS is perceived to be a useful data quality measurement tool.

The DQS is designed based on an extension to the research carried out by Cipriano, 2015. The figure below shows a mapping of the DQD to the requirements of each stakeholder group.



Category	Dimensions
Intrinsic	Accuracy, Believability, Objectivity, Reputation of the Source
Contextual	Value-added, Relevancy, Timeliness, Completeness, Amount of data
Representational	Interpretability, Ease of Understanding, Representational Consistency, Concise Representation
Accessibility	Accessibility, Ease of Operations, Security

Figure 18: Relationship between stakeholder, data quality dimensions and classifications (Cipriano, 2015)

Figure 19 below shows the conceptual framework of the DQS.

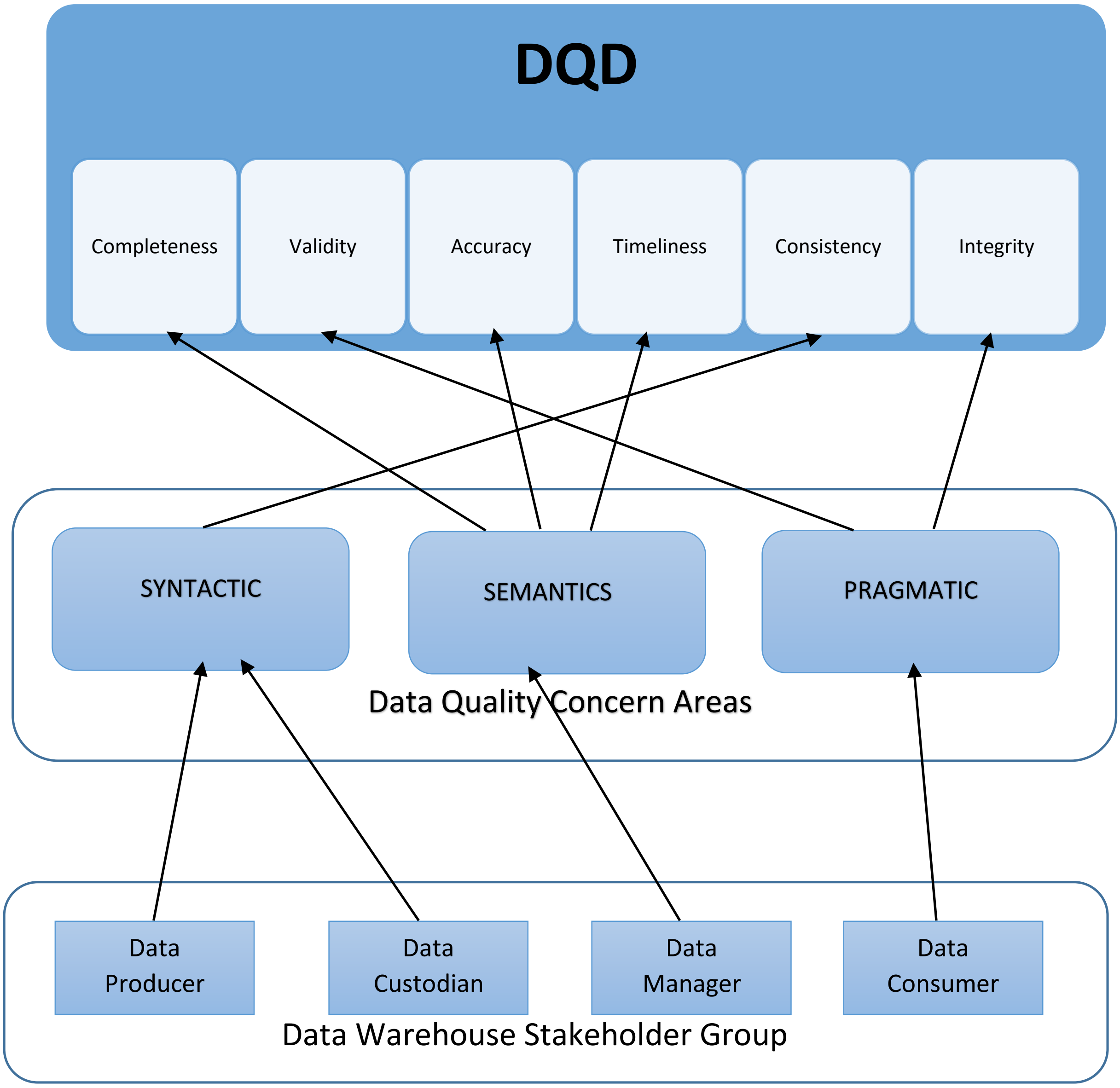


Figure 19: Proposed DQS framework

The development of the DQS from literature takes the research of Cipriano, 2015, and studies that were done by (Madnick et al., 2009; Blake, 2010; Clement et al., 2011; Odera-Kwach et al., 2011; Naiem et al., 2014). The Construction details for the DQS framework is justified by the systematic literature review conducted, the previous studies conducted (Acosta et al., 2013; Fan and Geerts, 2013; Weiskopf NG & Weng C, 2012; Mendes et al., 2012; Kauffman et al., 2009; Batini et al., 2009) provided the researcher with grounding for the development of the proposed DQS framework presented in figure 19 above.

4.3 Scorecard Mechanics – Electronic and Web-centric DQS Development

In Chapter 3, a systematic literature review was carried out to investigate the current state-of-the-art of scorecards. Previously, studies by (Kaplan and Norton, 2001; Lawrie and Cobbold, 2004; Chang, 2006; Brace, 2008; Coe and Letza, 2014) recommended that in the implementation and design of a scorecard metric, the following factors need to be considered; use simplicity, time efficiency, electronic/online features, non-technical language and a less intuitive approach. Therefore, these factors were incorporated into the design of the DQS.

The DQS is designed for use as either a writable electronic form using a portable document format (PDF). PDF is a widely accepted file format used for presenting and exchanging data reliably. The web-centric scorecard was developed using HTML 5, CSS 5, PHP, JavaScript and MySQL database management system. The web version of the DQS is currently hosted on a private network from a commercial ISP, the researcher chose this particular service provider because it guarantees a 99.95% uptime and for its overall reliability.

The web-centric scorecard can also be hosted privately by companies on their intranet so as to secure the scorecard solely on their private network.

HyperText Markup Language (HTML 5)

HTML is a Markup language used for organising and presenting information on the World Wide Web (Anthes, 2012). HTML5 was used as the underlying programming language due to its capability to animate text, graphics and image content as well as continuous media. Besides, the proliferation of various technology devices coupled with the variety of browsers significantly motivated the adoption of HTML5 for developing the web-centric scorecard. HTML5 is compatible with the majority of internet browsers and is compatible with flexible programming. HTML5 presents a new security model that is not only easy to use but is also used regularly by several APIs. HTML5 has the features of being able to communicate securely and seamlessly across domains productively. The figure below shows the front page of the designed web-based scorecard.

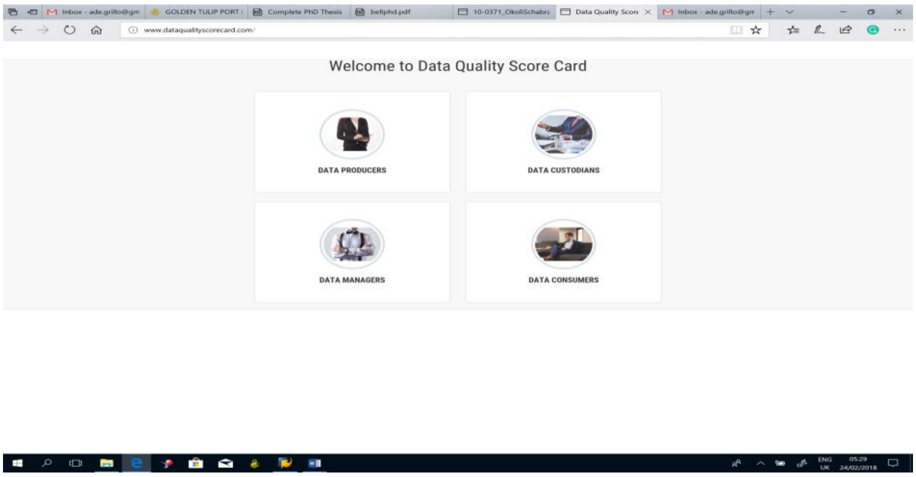

www.dataqualityscorecard.com/
Welcome to Data Quality Score Card
DATA PRODUCERS
DATA CUSTODIANS
DATA MANAGERS
DATA CONSUMERS

Figure 20: Data Quality Scorecard front page

Cascading style sheets (CSS 3)

CSS is a style sheet language that gives appearance changes to HTML. CSS version 3 was used to design the web-based data quality scorecard for two main reasons; 1) to ensure compatibility with most web browsers, 2) to ensure a clean and warm feel of the scorecard. Similar to HTML, CSS based applications are industry standard and are compatible with a large number of web browsers. Vital to any website‘s success is the usability factor and as such utilising a language that is standard for website creation will reduce the amount of change management required for users. With the use of the improved version 3 of CSS, the researcher was able to design the page layout and presentation of the scorecard efficiently and with the ability to easily resolve common design problems with fewer lines of code. The figure below shows a simple code extract using CSS 3.

```
body {
font: normal 100% "trebuchet ms", Arial, Helvetica, sans-serif;
}
a {
color: #000000;
}
A: visited {
color: #005177;
}
a:hover {
color: #005177;
}
```

Figure 21: CSS code structure

Hypertext Preprocessor (PHP)

PHP is an HTML-embedded Web scripting language. PHP is a free proficient server-side scripting language for creating dynamic and interactive Web applications. PHP is largely integrated with HTML elements. PHP Web-scripting language is compatible with most of the major Web servers. PHP was used in the design mechanics of the scorecard to embed code fragments in with the HTML pages. PHP also functioned as the link between the scorecards Web pages and its MySQL databases. PHP is a leader in the web development domain. The researcher was motivated to use PHP over other tools like CGI for the development of the web-based scorecard because of the simplicity of the tool.

JavaScript

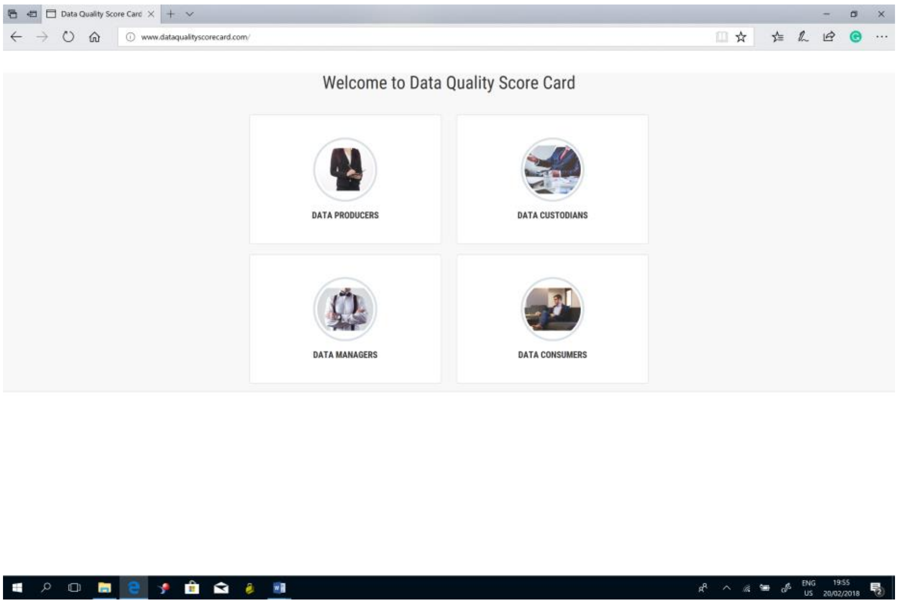
JavaScript is a dynamic programming language used as a part of web pages; it allows client-side scripts to interact with the user and make dynamic pages. The inputs made by the stakeholders are seamlessly validated before sending the page off to the server; thus, saving server traffic resulting in fewer loads on the server. JavaScript also allows immediate feedback to the visitors without waiting for a page reload. Moreover, JavaScript helped to enhance the interactivity of the scorecard; through interfaces that created dynamic movements when a user places a mouse cursor over active areas on the website.

MySQL

MySQL is a relational database management system integrated with PHP to store user data. The rationale for using MySQL because it is platform independent. Although it can be utilised in a wide range of applications, MySQL is most often associated with web applications and is a vital element of an open-source enterprise stack called XAMPP.

4.3.1 Walkthrough of the Web-centric DQS

The web version of the DQS is currently published at www.dataqualityscorecard.com, this is currently the only way to access the DQS. Further access avenues are being developed for offline use. The DQS is designed and features the four identified stakeholder groups within the DW domain. The figure below shows the landing page of the DQS.

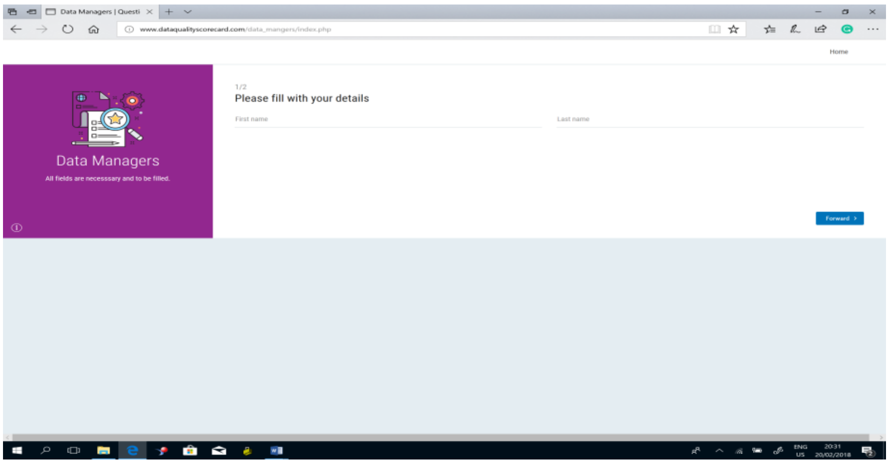


Welcome to Data Quality Score Card

Figure 22: DQS Landing page

Using the Webcentric DQS

A login is required by each stakeholder that uses the system. Currently, authentication is not required to use the system, only the name of the stakeholder is required. A separate login page is presented for each stakeholder group. Figure xx below shows the login page of the DQS.



Data Managers

All fields are necesssary and to be filled.

1/2

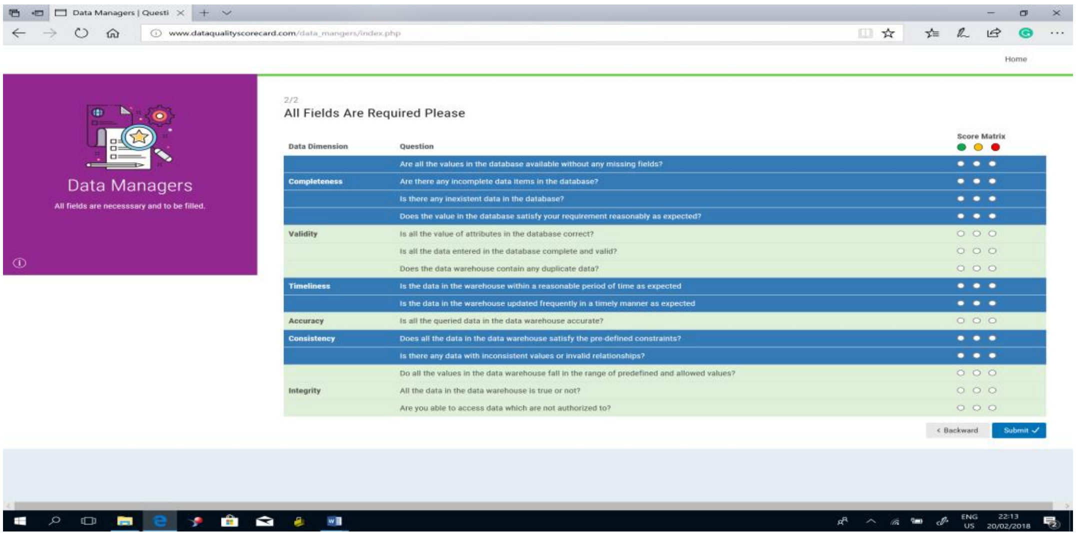
Please fill with your details

First name

Last name

Figure 23: Login page after selection of stakeholder group

Each stakeholder is then presented with a set of questions related closely to their concern areas as identified from the literature. The selection made by each stakeholder is then stored and available for review/reporting.



Data Managers

All fields are necesssary and to be filled.

2/2

All Fields Are Required Please

Data Dimension	Question	Score Matrix		
		🟢	🟡	🔴
	Are all the values in the database available without any missing fields?	○	○	○
Completeness	Are there any incomplete data items in the database?	○	○	○
	Is there any inexistent data in the database?	○	○	○
	Does the value in the database satisfy your requirement reasonably as expected?	○	○	○
Validity	Is all the value of attributes in the database correct?	○	○	○
	Is all the data entered in the database complete and valid?	○	○	○
	Does the data warehouse contain any duplicate data?	○	○	○
Timeliness	Is the data in the warehouse within a reasonable period of time as expected	○	○	○
	Is the data in the warehouse updated frequently in a timely manner as expected	○	○	○
Accuracy	Is all the queried data in the data warehouse accurate?	○	○	○
Consistency	Does all the data in the data warehouse satisfy the pre-defined constraints?	○	○	○
	Is there any data with inconsistent values or invalid relationships?	○	○	○
	Do all the values in the data warehouse fall in the range of predefined and allowed values?	○	○	○
Integrity	All the data in the data warehouse is true or not?	○	○	○
	Are you able to access data which are not authorized to?	○	○	○

Figure 24: DQS Questions aligned with DQD and stakeholder requirements

4.4 DQS Validation

To validate the usability of the DQS as designed, a system usability scale was adopted. The SUS (Brooke, 1996) is a very popular standardized questionnaire for the assessment of system usability. In a study of unpublished industrial usability studies, it was found that the SUS accounted for 43% of questionnaire usage. (Sauro and Lewis, 2009).

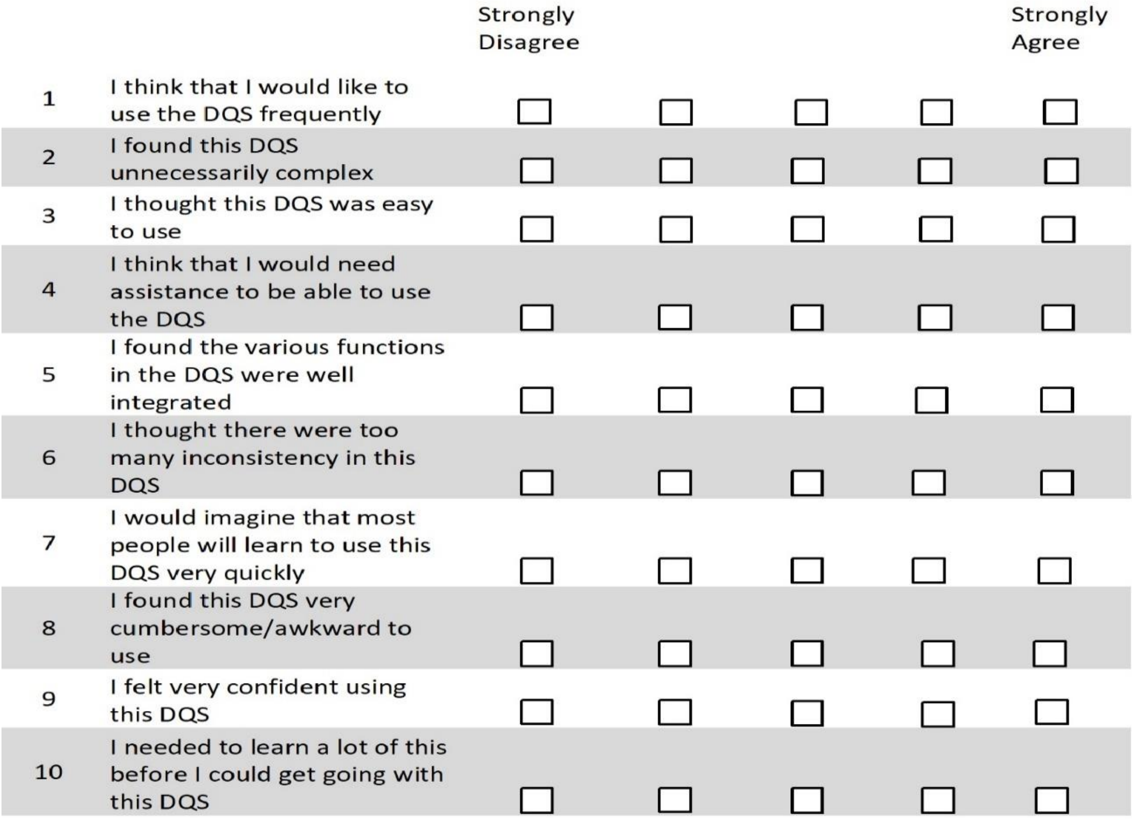
4.4.1 Data Collection Techniques

The usability questionnaire items used were adopted from Brooke et al, (1996), and slightly modified to suit the research context. The modified SUS was designed to validate the usefulness of the DQS and possibly gather inputs for further improvement of the scorecard.

4.4.2 SUS Questionnaire Design

The standard questionnaire items (Brooke, 1996) of the system usability scale (SUS) was adopted with some slight modification. The SUS is a one-dimensional scale which consists of 10 questionnaire items that evaluate the subjective perception of the stakeholder's usage of the system regardless of their personal interpretations. SUS is an industry-accepted scale for measuring the subjective views of the users of the system. It utilises a five-point Likert scale with anchors for "strongly agree" to "strongly disagree ". It is a two-tone questionnaire in which odd-numbered items have a positive tone and even-numbered items have a negative tone. The SUS has also been shown to have acceptable levels of concurrent validity (Bangor et al., 2008).

The figure below shows the questionnaire items



		Strongly Disagree				Strongly Agree
1	I think that I would like to use the DQS frequently	□	□	□	□	□
2	I found this DQS unnecessarily complex	□	□	□	□	□
3	I thought this DQS was easy to use	□	□	□	□	□
4	I think that I would need assistance to be able to use the DQS	□	□	□	□	□
5	I found the various functions in the DQS were well integrated	□	□	□	□	□
6	I thought there were too many inconsistency in this DQS	□	□	□	□	□
7	I would imagine that most people will learn to use this DQS very quickly	□	□	□	□	□
8	I found this DQS very cumbersome/awkward to use	□	□	□	□	□
9	I felt very confident using this DQS	□	□	□	□	□
10	I needed to learn a lot of this before I could get going with this DQS	□	□	□	□	□

Figure 25: SUS questionnaire items

4.4.3 Participants

According to Boud (1995), all assessment including self-assessment comprises two main elements: making decisions about the standards of performance expected and then making judgments about the quality of the performance in relation to these standards. Using a self-assessment role process to gather and review data quality concern areas about the domain ensures an increased involvement in the process of assessing strengths and areas in need of improvement, identify discrepancies of performance between the data and the roles, and to conduct a more constructive evaluation of the data quality needs specific to each of the roles in the domain. The self-assessment process for each role was introduced with a clear rationale and guidelines for each stage of the process.

The participants were initially contacted to get their buy-in to participate in the DQS evaluation.

Two companies in the oil and gas and one in the fast-moving consumer goods (FMCG) domains took part in this study. Research suggests that sample sizes of at least 12 – 14 participant (Sheng et al., 2010) are needed to achieve statistically reliable results. A total of 10 data producers, 8 data custodians, 3 data managers and 15 data consumers participated in this study (Total of 36 participants). The demography of the participants is shown in table 9 below.

Measure	Item	Frequency	Percentage (%)
Stakeholder Roles	Data Custodians	8	22.2
	Data Producers	10	27.8
	Data Managers	3	8.3
	Data Consumers	15	41.7
Age	21-30	7	19.4
	30-45	18	50
	45 and Over	11	30.6
Years in Industry	0-3 years	4	11.1
	3-7 years	9	25
	7 years +	23	63,9

Table 9: Demography of stakeholder Roles

4.4.4 Procedure

The stakeholders were given the link to the scorecard a day before the questionnaire. Thereafter, they were asked to complete the system usability scale (SUS) items. The main objective of this evaluation is to gauge the usability of the DQS in practice.

4.4.5 SUS Evaluation Results

The first step in scoring a SUS is to determine each item's score contribution, which ranges from 0 (being a poor score) to 4 (a good score). For odd-numbered items, the score contribution is the scale position minus 1, while For even-numbered items, the score contribution is 5 minus the scale position. The overall SUS score is derived by multiplying the sum of the item score contributions by 2.5, the result produces a score that can range from 0 (very poor perceived usability) to 100 (excellent perceived usability).

SPSS software package (IBM SPSS Statistics 20) was used for the data analysis of the system usability scale (SUS).

To measure how closely related the internal consistency of the set of items is as a group, Cronbach‘s alpha was calculated (Bonett and Wright, 2015). The Cronbach‘s alpha of 0.790

was calculated. This shows larger values than the acknowledged level of 0.7. Nunnaly (1978) has indicated 0.7 to be an acceptable reliability coefficient. Hence, making this analysis statistically adequate.

The SUS score from each stakeholder is presented in table 10 below. Furthermore, to ensure the reliability and validity of the questions, the mean and standard deviation of the questions are calculated and presented in table 11 below.

Stakeholder ID	SUS Score	Stakeholder ID	SUS Score
1	95	19	82.5
2	87.5	20	92.5
3	80	21	87.5
4	77.5	22	72.5
5	95	23	87.5
6	95	24	87.5
7	92.5	25	87.5
8	87.5	26	87.5
9	82.5	27	95
10	60	28	92.5
11	92.5	29	87.5
12	77.5	30	82.5
13	90	31	87.5
14	92.5	32	50
15	92.5	33	77.5
16	92.5	34	90
17	82.5	35	85
18	85	36	87.5

Table 10: Stakeholders SUS Score

Question	Mean	Deviation	Respondents
Q1	3.31	0.668	36
Q2	3.5	0.609	36
Q3	3.03	0.774	36
Q4	3.5	0.655	36
Q5	3.03	0.774	36
Q6	3.64	0.543	36
Q7	3.44	0.607	36
Q8	3.67	0.586	36
Q9	3.58	0.500	36
Q10	3.5	0.655	36

Table 11: Mean and Standard Deviation for each Question

The average stakeholders‘ subjective satisfaction of the DQS was significantly high (85.5 out of 100) (Brooke et al., 1996). The majority of the stakeholders‘ perceived the DQS to be useful, easy to use and created an awareness platform for the quality of data in the data warehouse.

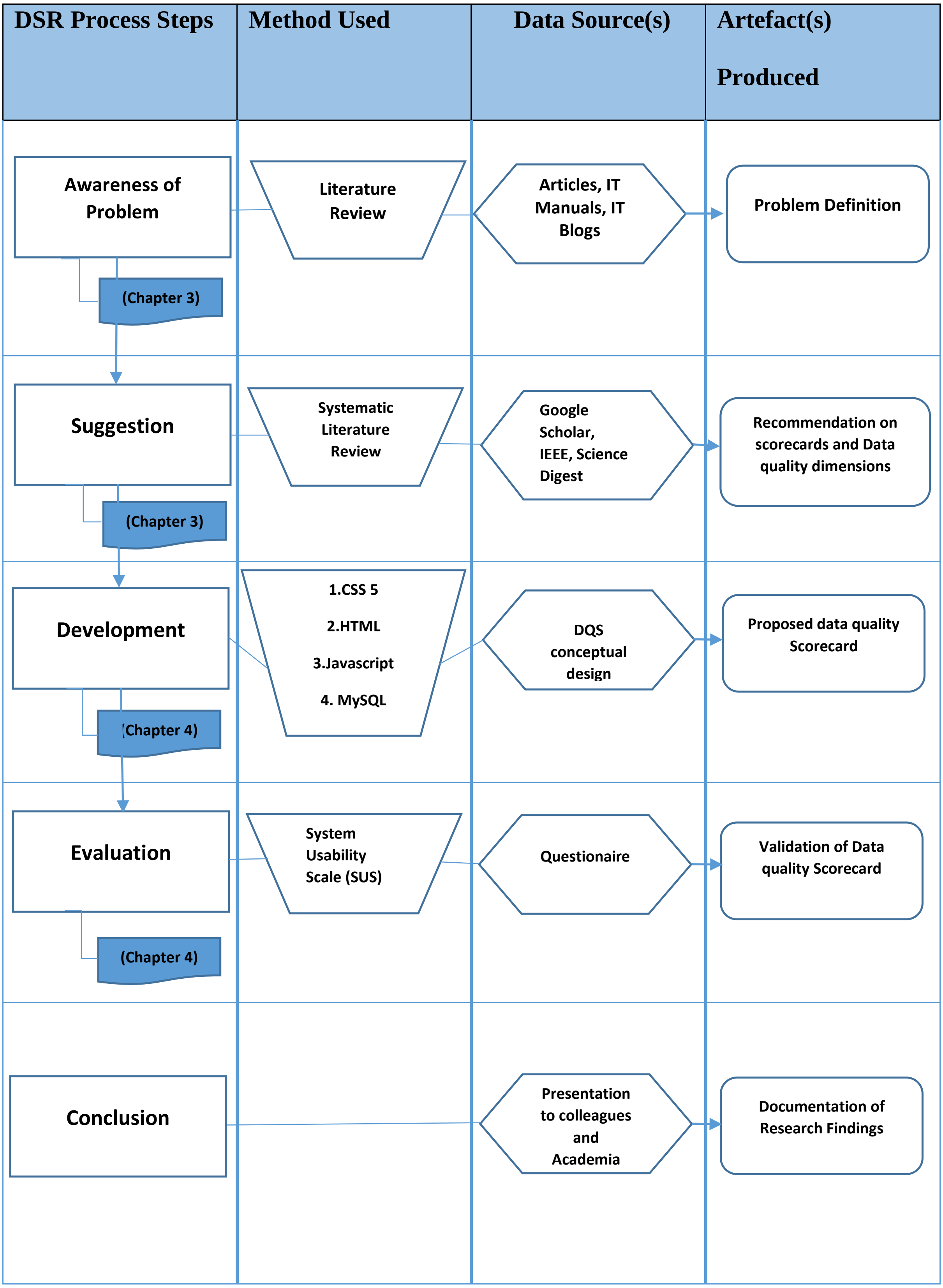
4.4.6 Analysis of results

The study attempts to validate the stakeholders‘ subjective satisfaction with DQS and its potential use as a data quality measurement. To accomplish these objectives, a SUS study was employed. Preliminary SUS results show that the stakeholders find the DQS as a useful tool based on the average usability score of 85.5 that was achieved. The stakeholders indicated a few areas that needed some improvement, but these changes were cosmetic in nature and were immediately addressed. In subsequent Chapters, a run through and further evaluation of the DQS will be carried out in two domains.

4.5 Summary

In this chapter, a data quality scorecard (DQS) is developed based on the findings from the literature. In chapter 3, a comprehensive analysis of the existing knowledge in the domain of data quality dimensions and scorecard design was carried out, this was used as the basis for the design of the DQS. Although considerable research has been done in the data quality domain; not much work has been done with the alignment of the varied needs of the stakeholder groups with the DQD in the data warehouse domain. A system usability scale (SUS) was used to validate the effectiveness of the DQS. This study has shown that the stakeholder group find the DQS useful and effective based on the high average usability score of 85.5 that was achieved. In the next chapter, the DQS is further evaluated in the FMCG domain.

Figure 26 below summarises the first DSR iteration with the methods and techniques used.



DSR Process Steps	Method Used	Data Source(s)	Artefact(s) Produced
Awareness of Problem (Chapter 3)	Literature Review	Articles, IT Manuals, IT Blogs	Problem Definition
Suggestion (Chapter 3)	Systematic Literature Review	Google Scholar, IEEE, Science Digest	Recommendation on scorecards and Data quality dimensions
Development [Chapter 4)	1.CSS 5 2.HTML 3.Javascript 4. MySQL	DQS conceptual design	Proposed data quality Scorecard
Evaluation (Chapter 4)	System Usability Scale (SUS)	Questionaire	Validation of Data quality Scorecard
Conclusion		Presentation to colleagues and Academia	Documentation of Research Findings

Figure 26: Summary of results of the first DSR Iteration cycle

Chapter 5: DQS Evaluation-Iteration II

5.1 Introduction

This chapter aims to report on the run through of the data quality scorecard in a live environment and then evaluate the designed scorecard (Artefact) as developed and presented in the previous chapter. Qualitative methods were employed for this evaluation, which includes: (1) Case study and (2) Semi-structured interview. The chapter is structured as follows: **Section 5.2** describes the run through of the scorecard to a specific case study. The case study involves a brewery in the Fast Moving Consumer Goods (FMCG) domain. In this section, the researcher will briefly present the architecture of the organisation's data warehouse, the problems they are currently facing and the way the DQS was applied to attempt to resolve the respective issues. **Section 5.3** describes the semi-structured interview conducted after the run through, the procedures, results and further discussions. The chapter concludes with a summary in **Section 5.4**.

5.2 About Brewing Company Ltd

Brewing Ltd is a brewing company which is headquartered in Europe. Brewing Ltd is one of the leading companies in the world, whose principal activity is the production, distribution, and sale of beer and soft drinks. Brewing Ltd is part of the Brewing Company Group, which also includes a host of other subsidiaries. The Company employs over 125,000 people and runs sixty-seven plants in forty countries. The R&D (research and development) activities are an integral part of the Brewing Company group to strengthen the position of an international brand, notably Brewing Ltd combined with the development of various regional brands. The primary objective of Brewing Ltd is to create a world-class team which includes highly qualified, motivated, ambitious, and open-minded people, and also oriented to constant development. The Company has substantial emphasis mainly on the development of a transparent organisation structure that reflects the responsibilities of all staff members.

Brewing Ltd is a GloCal organisation, being part of a GloCal organisation, it has to find the right balance between working closely together at the global level, and by allowing several strong initiatives and local brands to flourish. The Company always creates significant value locally in each marketplace by becoming close to the consumers and customers, and at the same time, significant value is created by using the strengths of being part of the group, sharing best practices, taking advantage of the scale, and centralising and standardising processes and functions across borders. The value creation strategy of the Brewing Ltd promotes growth and efficiency and also improves their practices. Finding the right GloCal balance in the matrix structure is one of the keys that will make Brewing Ltd achieve its ambition and success. In this study, the company details, investors, media, careers, markets, contact and CSR (corporate social responsibility) are the things that do not change frequently. Therefore, this study intends to conduct a case study based on the products of Brewery Ltd.

5.2.1 The Data Quality Problem at Brewery Ltd

The value of data changes usually correlates with the growth of a company. The major problem is that it needs to be regularly updated, so that, end users can be able to get better product development input. In general, even one data missing can have the ability to generate a big problem for the company. Due to the complexity of the data warehouse, it is more or less unfeasible to maintain the entire database by a single role, the more efficient way is to break the entire workload into several parts to manage it. Data Manager, Data Producer, and Data Custodian play significant roles to perform this job safely. The segregation of these roles is essential, and one which is currently not enforced by the system at Brewery Ltd.

- Data Producer: -

The data producer at Brewery Ltd collects the raw data from several source systems and then forwards it to the Data Custodian. In general, the data producer is the person that makes contact with several operating units and then collects the entire data systematically. At first, the data producer collects the region type in a separate file from others. The product type and the other attributes cannot be attached along with it, and this is because the other attributes can be changed based on a daily manner. Therefore, it needs to be regularly updated.

- Data Custodian: -

The data Custodian at Brewery Ltd is the person that collects the information from the data producer and then transfers the data into the database. The data custodian operates the database server for the company. An SAP BI system is used for the data warehousing. It was stressed by the company that it is essential for the data custodian to use an efficient technological system to maintain the data in the data warehouse. SAP, in this case, is the market leader, and most widely used ERP system worldwide. The benefit chart of the SAP BI system given is detailed below:

SAP BI Benefit Chart
Reduce or eliminate data movement
Fewer copies of the data
In-memory performance to provide answers in seconds, not in hours
Reduced latency which means current data is addressed, not old data

Access data across the enterprise
Unmatched federation of the data without centralising it
Advanced analytics for mining the non-traditional data
Petabytes of the historical data storage
Extensive Hadoop and no-SQL support
Streaming analytics
Analytics and data management from device to the enterprise
Innovate with entirely new applications that leverage the cultivated storage of the big data
Modernize data warehouse infrastructure along with the dynamic cloud

Figure 27: Brewery Ltd. SAP benefit Chart

- Data Manager:

The Data Manager in Brewery Ltd is the person responsible for maintaining the entire data in the data warehouse. The Data Manager's principal responsibility is to check the security of Brewery Ltd data warehouse database. So, it is vital to maintaining some specific guidelines during the testing process.

Data manager should concentrate on the important points below:

- Whether the new data sources will need any audit restrictions and new security to be implemented?

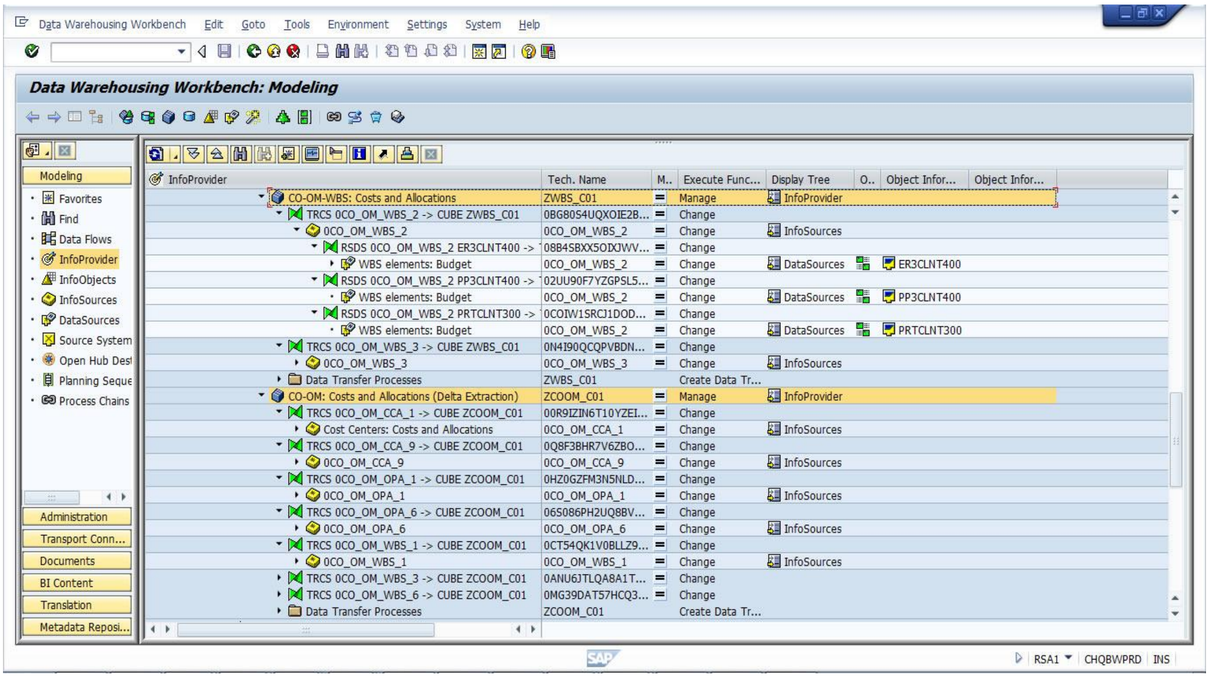
- Whether the new users added who have restricted access to data is already available?

Encryption: Data encryption is an important process that is required for transferring it into the data warehouse. So, it is essential to store the data by using any unique encryption algorithm. The administrative section is restricted to the end users, and so it is allowed only for the Company's administrator to access it. At the same time, the finance manager can only get access to the finance database. It was observed that every section of the company has separate admin roles for different admin functionality. Very robust roles and authorisation matrix are in place to ensure only authorised data is being accessed.

- Data Consumer*:*

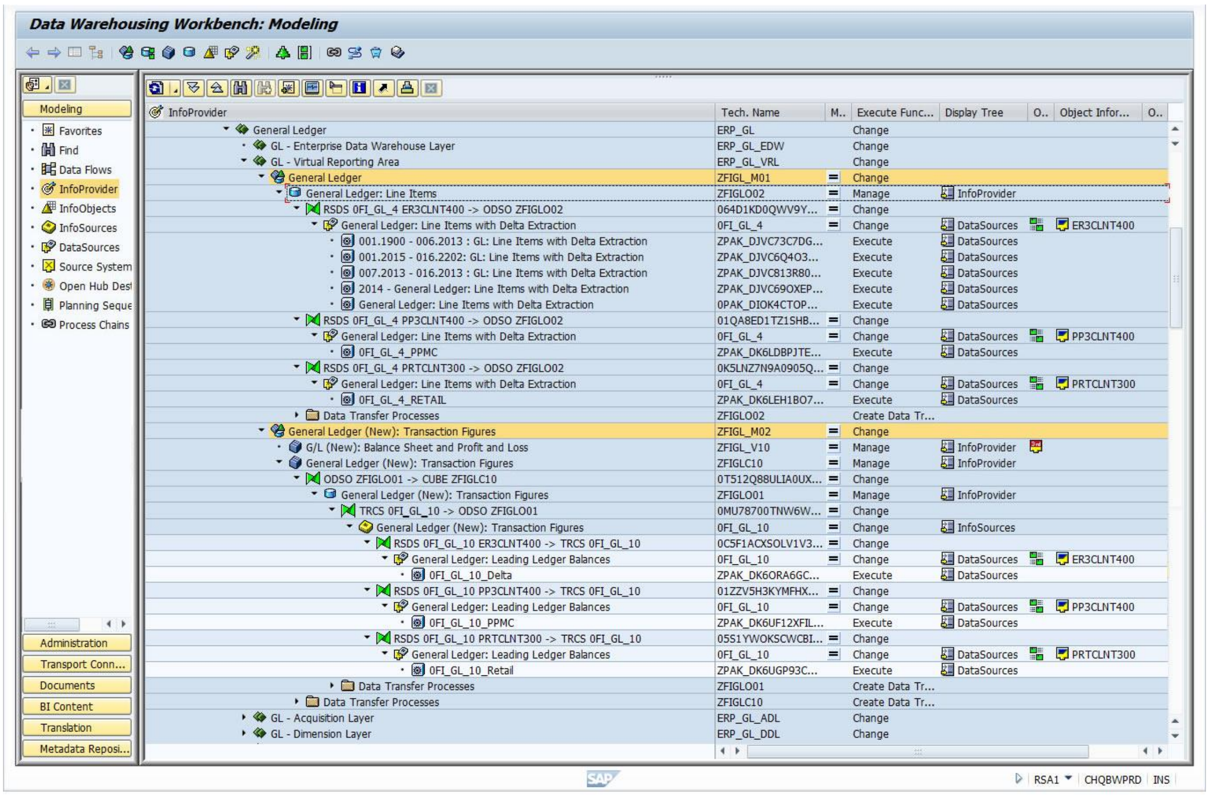
One of the vital issues in Brewery Ltd is the end user security and authorisation. The EEM (End User Experience Monitoring) tool, provided by SAP, is used by the Company to stimulate the behaviour of users who have rights to access the central servers at various locations and also to run the business processes. As like the administrator, it is possible for data consumers to monitor the availability of the systems and also the performance of the connections from the perspective of the end user in real time.

The legacy system which represents a data run without the use of the framework was used as a basis for comparison. The figures 28 and 29 below shows the data warehouse modelling workbench at brewery Limited.



InfoProvider	Tech. Name	M..	Execute Func...	Display Tree	O..	Object Infor...	Object Infor...
CO-OM-WBS: Costs and Allocations	ZWBS_C01	=	Manage	InfoProvider			
TRCS 0CO_OM_WBS_2 -> CUBE ZWBS_C01	0BG80S4UQXOIE2B...	=	Change				
0CO_OM_WBS_2	0CO_OM_WBS_2	=	Change	InfoSources			
RSDS 0CO_OM_WBS_2 ER3CLNT400 ->	08B4SBXX5OIXJWV...	=	Change				
WBS elements: Budget	0CO_OM_WBS_2	=	Change	DataSources		ER3CLNT400	
RSDS 0CO_OM_WBS_2 PP3CLNT400 ->	02UU90F7YZGPSL5...	=	Change				
WBS elements: Budget	0CO_OM_WBS_2	=	Change	DataSources		PP3CLNT400	
RSDS 0CO_OM_WBS_2 PRTCLNT300 ->	0COIW1SRCJ1DOD...	=	Change				
WBS elements: Budget	0CO_OM_WBS_2	=	Change	DataSources		PRTCLNT300	
TRCS 0CO_OM_WBS_3 -> CUBE ZWBS_C01	0N4I90QCQPVBDN...	=	Change				
0CO_OM_WBS_3	0CO_OM_WBS_3	=	Change	InfoSources			
Data Transfer Processes	ZWBS_C01		Create Data Tr...				
CO-OM: Costs and Allocations (Delta Extraction)	ZCOOM_C01	=	Manage	InfoProvider			
TRCS 0CO_OM_CCA_1 -> CUBE ZCOOM_C01	00R9IZIN6T10YZEI...	=	Change				
Cost Centers: Costs and Allocations	0CO_OM_CCA_1	=	Change	InfoSources			
TRCS 0CO_OM_CCA_9 -> CUBE ZCOOM_C01	0Q8F3BHR7V6ZBO...	=	Change				
0CO_OM_CCA_9	0CO_OM_CCA_9	=	Change	InfoSources			
TRCS 0CO_OM_OPA_1 -> CUBE ZCOOM_C01	0HZ0GZFM3N5NLD...	=	Change				
0CO_OM_OPA_1	0CO_OM_OPA_1	=	Change	InfoSources			
TRCS 0CO_OM_OPA_6 -> CUBE ZCOOM_C01	06S086PH2UQ8BV...	=	Change				
0CO_OM_OPA_6	0CO_OM_OPA_6	=	Change	InfoSources			
TRCS 0CO_OM_WBS_1 -> CUBE ZCOOM_C01	0CT54QK1V0BLLZ9...	=	Change				
0CO_OM_WBS_1	0CO_OM_WBS_1	=	Change	InfoSources			
TRCS 0CO_OM_WBS_3 -> CUBE ZCOOM_C01	0ANU6JTLQA8A1T...	=	Change				
TRCS 0CO_OM_WBS_6 -> CUBE ZCOOM_C01	0MG39DAT57HCQ3...	=	Change				
Data Transfer Processes	ZCOOM_C01		Create Data Tr...				

Figure 28: Brewery Ltd.’s SAP data warehouse Modelling workbench 1



InfoProvider	Tech. Name	M..	Execute Func...	Display Tree	O..	Object Infor...	O..
General Ledger	ERP_GL		Change				
GL - Enterprise Data Warehouse Layer	ERP_GL_EDW		Change				
GL - Virtual Reporting Area	ERP_GL_VRL		Change				
General Ledger	ZFIGL_M01	=	Change				
General Ledger: Line Items	ZFIGLO02	=	Manage	InfoProvider			
RSDS 0FI_GL_4 ER3CLNT400 -> ODSO ZFIGLO02	064D1KD0QWV9Y...	=	Change				
General Ledger: Line Items with Delta Extraction	0FI_GL_4	=	Change	DataSources		ER3CLNT400	
001.1900 - 006.2013 : GL: Line Items with Delta Extraction	ZPAK_DJVC73C7DG...		Execute	DataSources			
001.2015 - 016.2202: GL: Line Items with Delta Extraction	ZPAK_DJVC6Q4O3...		Execute	DataSources			
007.2013 - 016.2013 : GL: Line Items with Delta Extraction	ZPAK_DJVC813R80...		Execute	DataSources			
2014 - General Ledger: Line Items with Delta Extraction	ZPAK_DJVC69OXEP...		Execute	DataSources			
General Ledger: Line Items with Delta Extraction	0PAK_DIOK4CTOP...		Execute	DataSources			
RSDS 0FI_GL_4 PP3CLNT400 -> ODSO ZFIGLO02	01QA8ED1TZ1SHB...	=	Change				
General Ledger: Line Items with Delta Extraction	0FI_GL_4	=	Change	DataSources		PP3CLNT400	
0FI_GL_4_PPMC	ZPAK_DK6LDBPJTE...		Execute	DataSources			
RSDS 0FI_GL_4 PRTCLNT300 -> ODSO ZFIGLO02	0K5LNZ7N9A0905Q...	=	Change				
General Ledger: Line Items with Delta Extraction	0FI_GL_4	=	Change	DataSources		PRTCLNT300	
0FI_GL_4_RETAIL	ZPAK_DK6LEH1BO7...		Execute	DataSources			
Data Transfer Processes	ZFIGLO02		Create Data Tr...				
General Ledger (New): Transaction Figures	ZFIGL_M02	=	Change				
G/L (New): Balance Sheet and Profit and Loss	ZFIGL_V10	=	Manage	InfoProvider			
General Ledger (New): Transaction Figures	ZFIGLC10	=	Manage	InfoProvider			
ODSO ZFIGLO01 -> CUBE ZFIGLC10	0T512Q88ULIA0UX...	=	Change				
General Ledger (New): Transaction Figures	ZFIGLO01	=	Manage	InfoProvider			
TRCS 0FI_GL_10 -> ODSO ZFIGLO01	0MU78700TNW6W...	=	Change				
General Ledger (New): Transaction Figures	0FI_GL_10	=	Change	InfoSources			
RSDS 0FI_GL_10 ER3CLNT400 -> TRCS 0FI_GL_10	0C5F1ACXSOLV1V3...	=	Change				
General Ledger: Leading Ledger Balances	0FI_GL_10	=	Change	DataSources		ER3CLNT400	
0FI_GL_10_Delta	ZPAK_DK6ORA6GC...		Execute	DataSources			
RSDS 0FI_GL_10 PP3CLNT400 -> TRCS 0FI_GL_10	01ZZV5H3KYMFHX...	=	Change				
General Ledger: Leading Ledger Balances	0FI_GL_10	=	Change	DataSources		PP3CLNT400	
0FI_GL_10_PPMC	ZPAK_DK6UF12XFIL...		Execute	DataSources			
RSDS 0FI_GL_10 PRTCLNT300 -> TRCS 0FI_GL_10	05S1YWOKSCWCBI...	=	Change				
General Ledger: Leading Ledger Balances	0FI_GL_10	=	Change	DataSources		PRTCLNT300	
0FI_GL_10_Retail	ZPAK_DK6UGP93C...		Execute	DataSources			
Data Transfer Processes	ZFIGLO01		Create Data Tr...				
Data Transfer Processes	ZFIGLC10		Create Data Tr...				
GL - Acquisition Layer	ERP_GL_ADL		Change				
GL - Dimension Layer	ERP_GL_DDL		Change				

Figure 29: Brewery Ltd.’s SAP data warehouse Modelling workbench 2

Selected Data Dimension: -

- *Completeness:* - Raw data from the operation house has been taken entirely and then it is encrypted by using an encryption process before transferring it into the data warehouse.
- *Validity:*-It is essential to update the database up to date and also with a valid set of data when the new product comes on the market.
- *Accuracy: The company stated* that the accuracy of the data in the data warehouse is of utmost importance. Here, it was observed that all the data in the data warehouse went through various manual quality gates to check for the accuracy of the raw data obtained directly from the factory.
- *Timeliness:* -Timeliness is used to measure the age of the data in the data warehouse. Thus, the timeline of the data is always dependent on other factors. The product of the Company changes the season. New products, new flavours and beer strength are frequently developed/changed to coincide with the present season. Apart from these, the price of the products also has seasonal variations.
- *Consistency:*-In Brewery Ltd., it was observed that brand blending and consistency in flavour is essential to the organisation. This particular area was excluded from our research as authorisation was not given to analysing these specific data.
- *Integrity:* - Data Custodian is one of the persons, who is responsible for managing the database. Integrity is mainly used to check whether the data is valid or not. Thus, it is essential for data managers to check whether the transferred data is valid or not. Apart from these, they also need to check the efficiency.

The testing of the deployed framework showed that both requirements were achieved. In particular, the researcher observed that there were products where the total number of requested changes did not match the number calculated for the legacy system. Thus, the following quality goal was set: “Achieve 100% data consistency for the data warehouse views”. The results are shown in the scorecard below

5.2.2 Web-centric DQS-Iteration 2

The stakeholders were then directed to the website located at www.dataqualityscorecard.com to fill out the form about the quality of the data that was just entered or viewed into/from the data warehouse. The figures below show the login screen and the DQS from each of the stakeholder group.

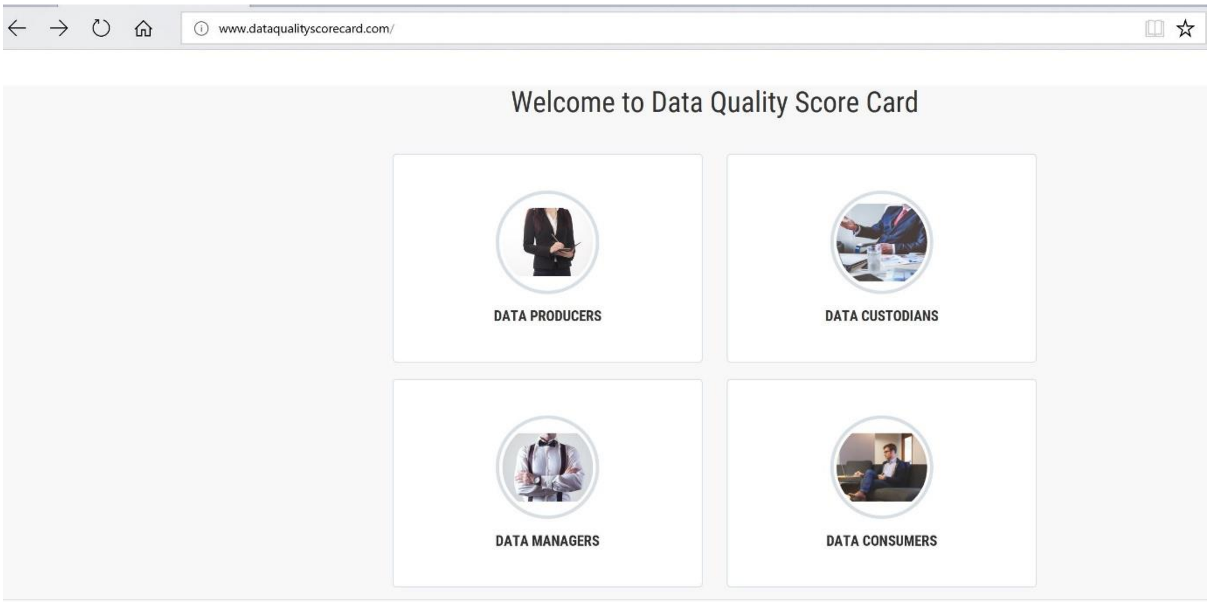
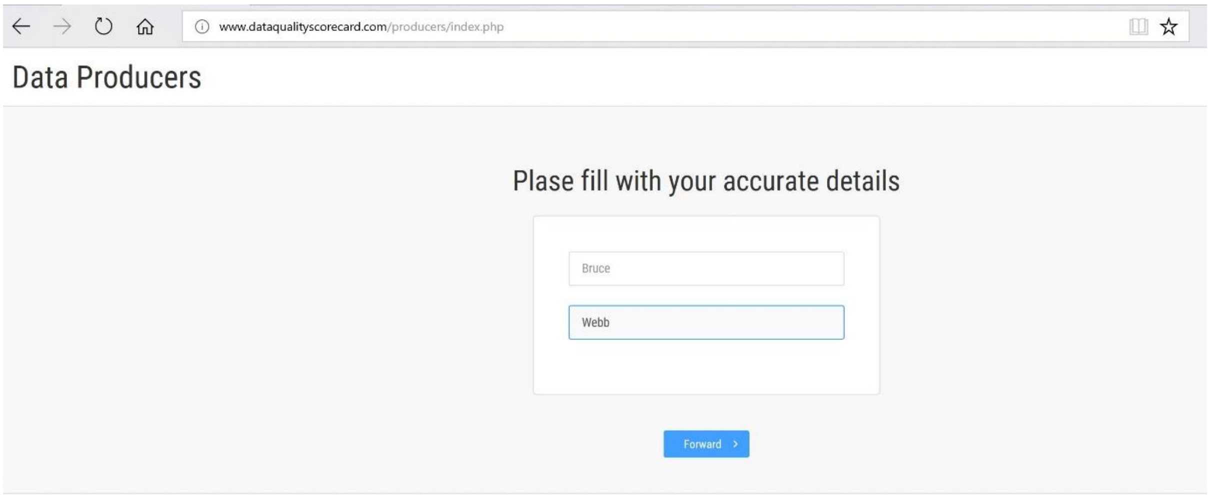

www.dataqualityscorecard.com/
Welcome to Data Quality Score Card
DATA PRODUCERS
DATA CUSTODIANS
DATA MANAGERS
DATA CONSUMERS

Figure 30: DQS Website Login screen

All stakeholders will be required to enter their names after selecting the group they belong to in the organization. The stakeholder's details will be validated internally using the relevant authorization metric in the organisation. Figure 31 below shows a stakeholder specific screen



Data Producers

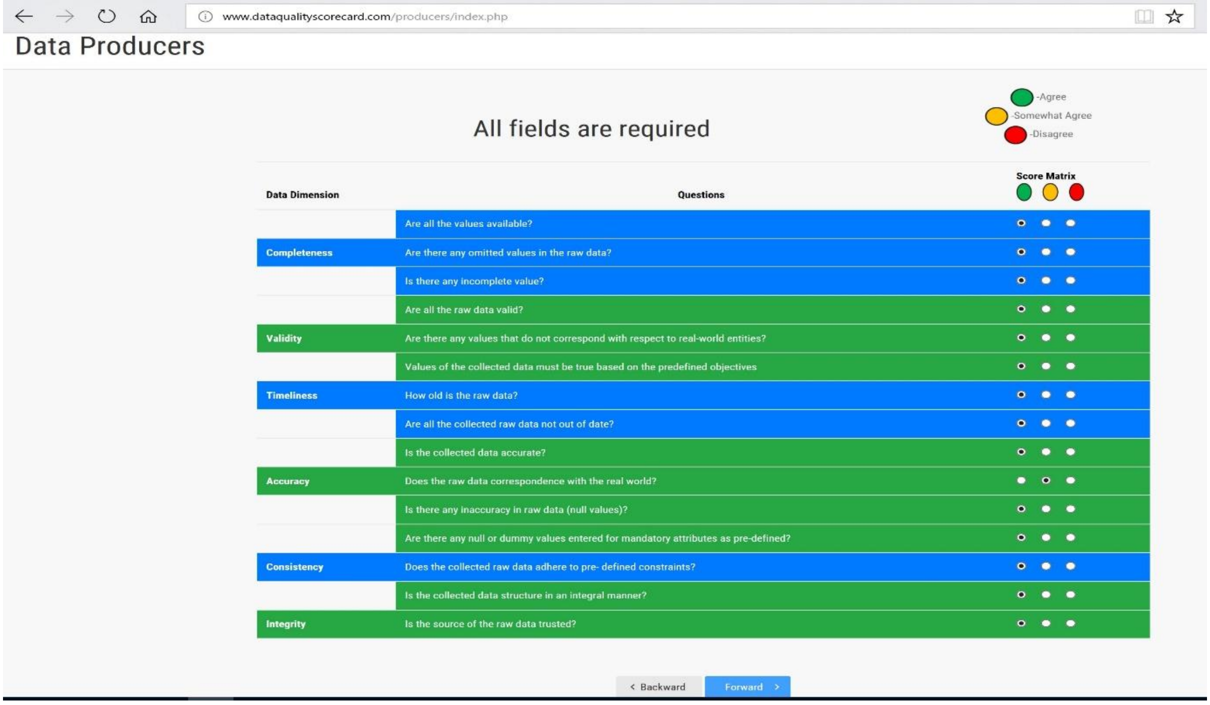
Plase fill with your accurate details

Bruce

Webb

Figure 31: Stakeholder specific login screen

The stakeholders are then presented with a series of questions that reflect their concern areas. The stakeholder concern areas according to literature are varied and reflect the needs of the various groups. The figures below depict the quality score attributed to each concern area by the stakeholder groups.



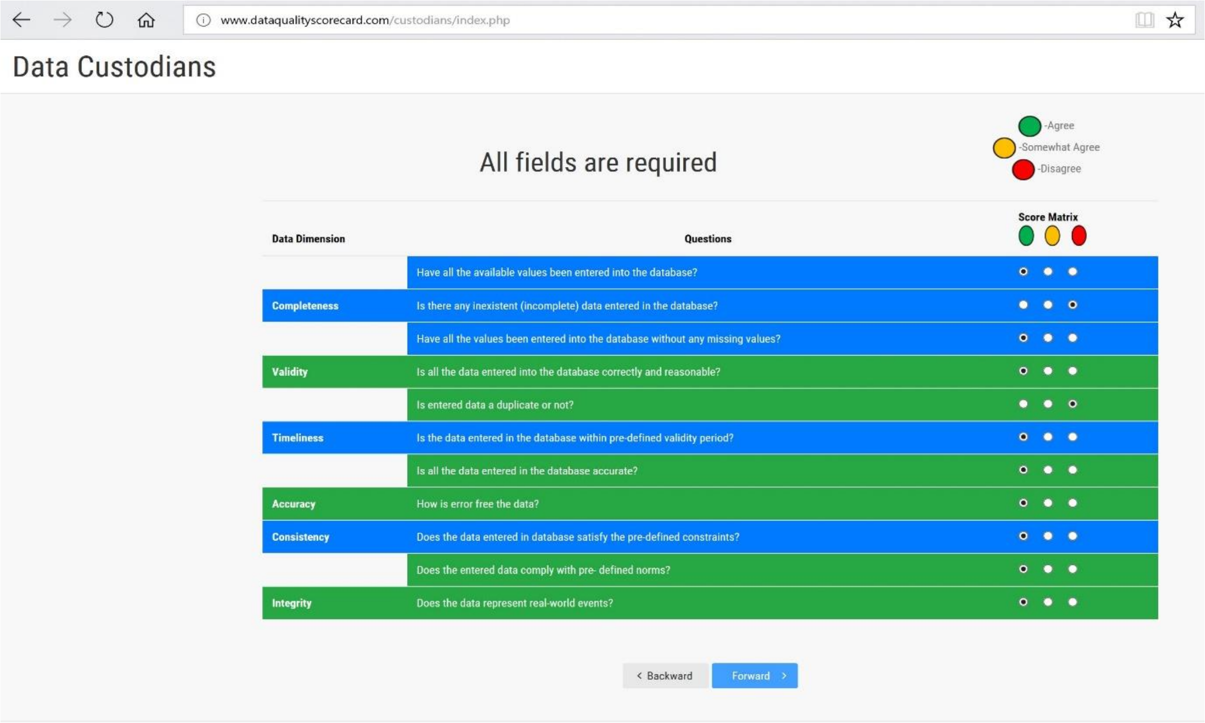
Data Producers

All fields are required

🟢 -Agree
🟡 -Somewhat Agree
🔴 -Disagree

Data Dimension	Questions	Score Matrix		
		🟢	🟡	🔴
	Are all the values available?	◉	○	○
Completeness	Are there any omitted values in the raw data?	◉	○	○
	Is there any incomplete value?	◉	○	○
	Are all the raw data valid?	◉	○	○
Validity	Are there any values that do not correspond with respect to real-world entities?	◉	○	○
	Values of the collected data must be true based on the predefined objectives	◉	○	○
Timeliness	How old is the raw data?	◉	○	○
	Are all the collected raw data not out of date?	◉	○	○
	Is the collected data accurate?	◉	○	○
Accuracy	Does the raw data correspondence with the real world?	○	◉	○
	Is there any inaccuracy in raw data (null values)?	◉	○	○
	Are there any null or dummy values entered for mandatory attributes as pre-defined?	◉	○	○
Consistency	Does the collected raw data adhere to pre- defined constraints?	◉	○	○
	Is the collected data structure in an integral manner?	◉	○	○
Integrity	Is the source of the raw data trusted?	◉	○	○

Figure 32: Data Producer DQS



www.dataqualityscorecard.com/custodians/index.php

Data Custodians

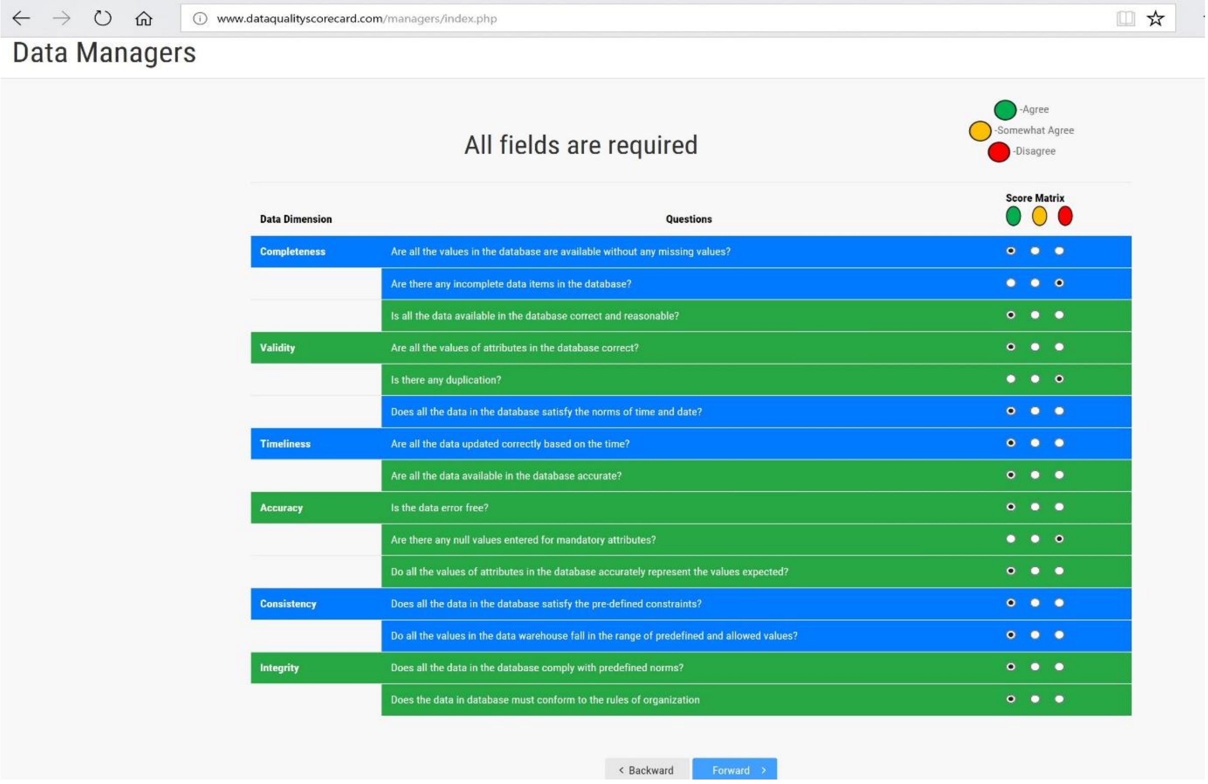
All fields are required

🟢 -Agree
🟡 -Somewhat Agree
🔴 -Disagree

Data Dimension	Questions	Score Matrix		
		🟢	🟡	🔴
	Have all the available values been entered into the database?	◉	○	○
Completeness	Is there any inexistent (incomplete) data entered in the database?	○	○	◉
	Have all the values been entered into the database without any missing values?	◉	○	○
Validity	Is all the data entered into the database correctly and reasonable?	◉	○	○
	Is entered data a duplicate or not?	○	○	◉
Timeliness	Is the data entered in the database within pre-defined validity period?	◉	○	○
	Is all the data entered in the database accurate?	◉	○	○
Accuracy	How is error free the data?	◉	○	○
Consistency	Does the data entered in database satisfy the pre-defined constraints?	◉	○	○
	Does the entered data comply with pre- defined norms?	◉	○	○
Integrity	Does the data represent real-world events?	◉	○	○

‹ Backward Forward ›

Figure 33: Data Custodian DQS



www.dataqualityscorecard.com/managers/index.php

Data Managers

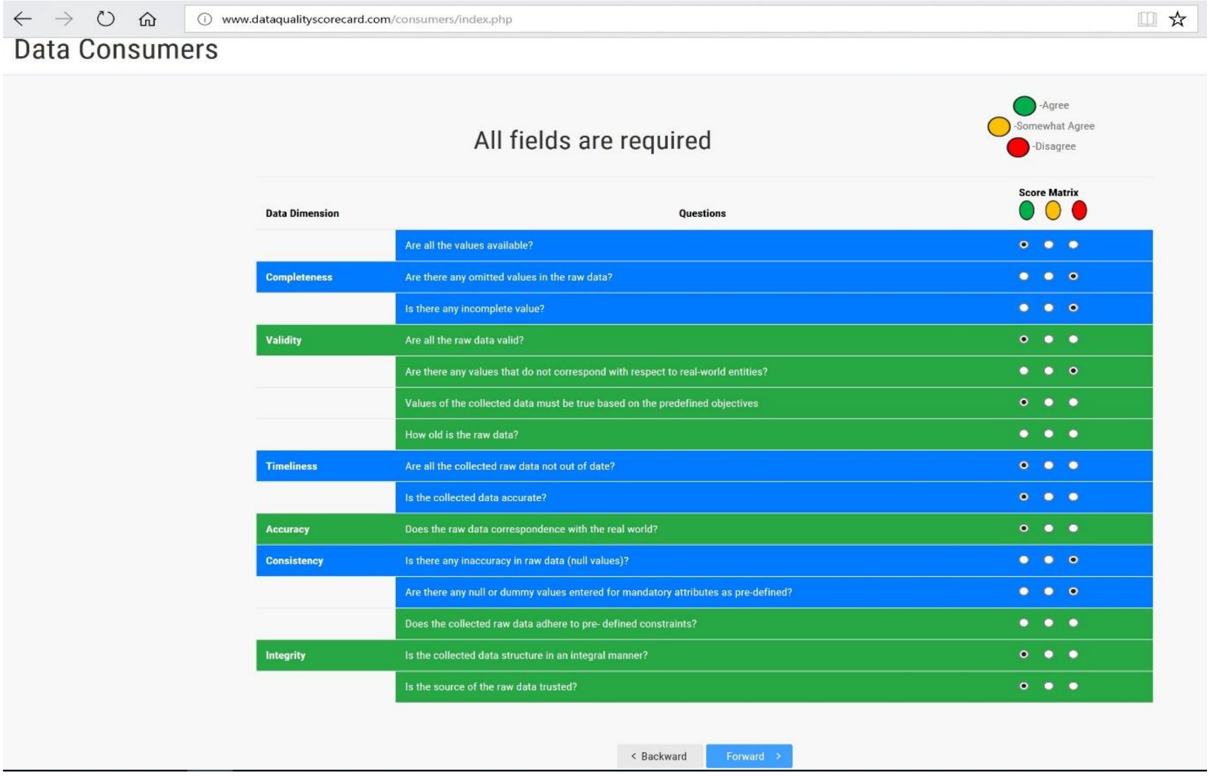
All fields are required

🟢 -Agree
🟡 -Somewhat Agree
🔴 -Disagree

Data Dimension	Questions	Score Matrix		
		🟢	🟡	🔴
Completeness	Are all the values in the database are available without any missing values?	◉	○	○
	Are there any incomplete data items in the database?	○	○	◉
	Is all the data available in the database correct and reasonable?	◉	○	○
Validity	Are all the values of attributes in the database correct?	◉	○	○
	Is there any duplication?	○	○	◉
	Does all the data in the database satisfy the norms of time and date?	◉	○	○
Timeliness	Are all the data updated correctly based on the time?	◉	○	○
	Are all the data available in the database accurate?	◉	○	○
Accuracy	Is the data error free?	◉	○	○
	Are there any null values entered for mandatory attributes?	○	○	◉
	Do all the values of attributes in the database accurately represent the values expected?	◉	○	○
Consistency	Does all the data in the database satisfy the pre-defined constraints?	◉	○	○
	Do all the values in the data warehouse fall in the range of predefined and allowed values?	◉	○	○
Integrity	Does all the data in the database comply with predefined norms?	◉	○	○
	Does the data in database must conform to the rules of organization	◉	○	○

‹ Backward Forward ›

Figure 34: Data Manager DQS


www.dataqualityscorecard.com/consumers/index.php

Data Consumers

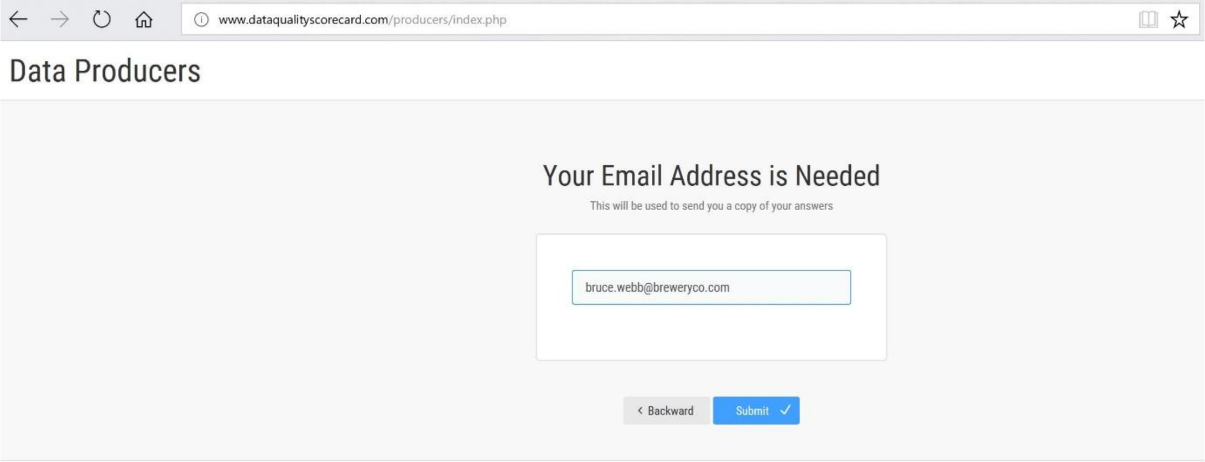
All fields are required

🟢 -Agree
🟡 -Somewhat Agree
🔴 -Disagree

Data Dimension	Questions	Score Matrix		
		🟢	🟡	🔴
	Are all the values available?	●	○	○
Completeness	Are there any omitted values in the raw data?	○	○	●
	Is there any incomplete value?	○	○	●
Validity	Are all the raw data valid?	●	○	○
	Are there any values that do not correspond with respect to real-world entities?	○	○	●
	Values of the collected data must be true based on the predefined objectives	●	○	○
	How old is the raw data?	○	○	○
Timeliness	Are all the collected raw data not out of date?	●	○	○
	Is the collected data accurate?	●	○	○
Accuracy	Does the raw data correspondence with the real world?	●	○	○
Consistency	Is there any inaccuracy in raw data (null values)?	○	○	●
	Are there any null or dummy values entered for mandatory attributes as pre-defined?	○	○	●
	Does the collected raw data adhere to pre- defined constraints?	○	○	○
Integrity	Is the collected data structure in an integral manner?	●	○	○
	Is the source of the raw data trusted?	●	○	○

Figure 35: Data Consumer DQS

The stakeholders are then required to enter their email address for the DQS to be stored. The, however, can be made an optional step by the various organizations. The figure below shows the email screen


www.dataqualityscorecard.com/producers/index.php

Data Producers

Your Email Address is Needed

This will be used to send you a copy of your answers

bruce.webb@breweryco.com

Figure 36: DQS Email screen

The DQS results can be viewed by the individual stakeholders or as a collective. The reports can be viewed by login into www.dataqualityscorecard.com/collected_data. The figures below show the report delineated by stakeholder groups and then by individuals.

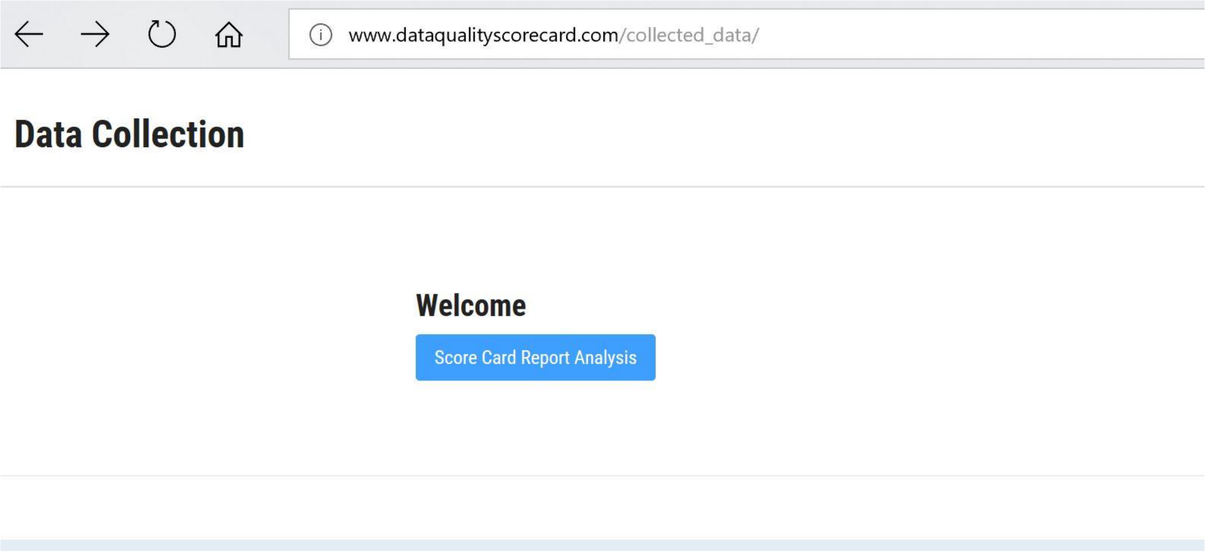
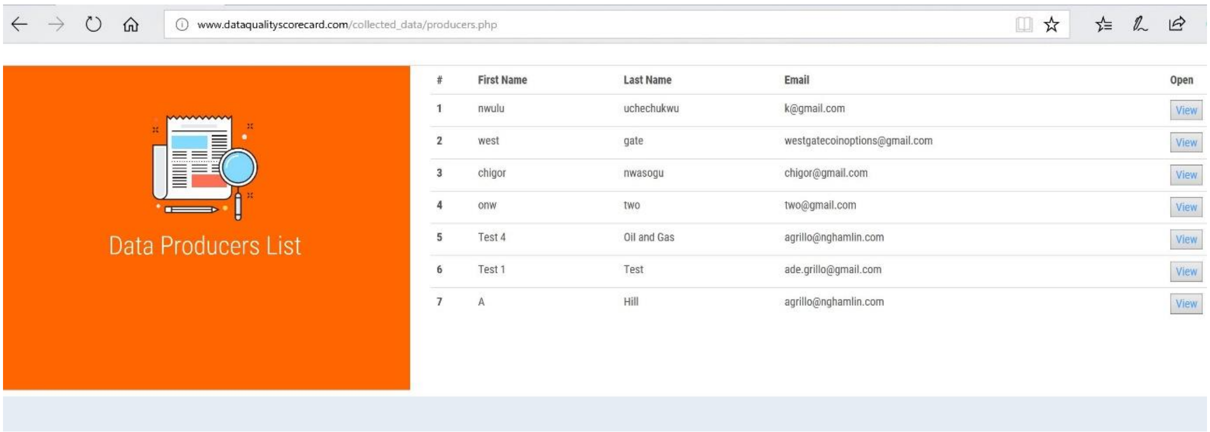

www.dataqualityscorecard.com/collected_data/
Data Collection
Welcome
Score Card Report Analysis

Figure 37: DQS report area

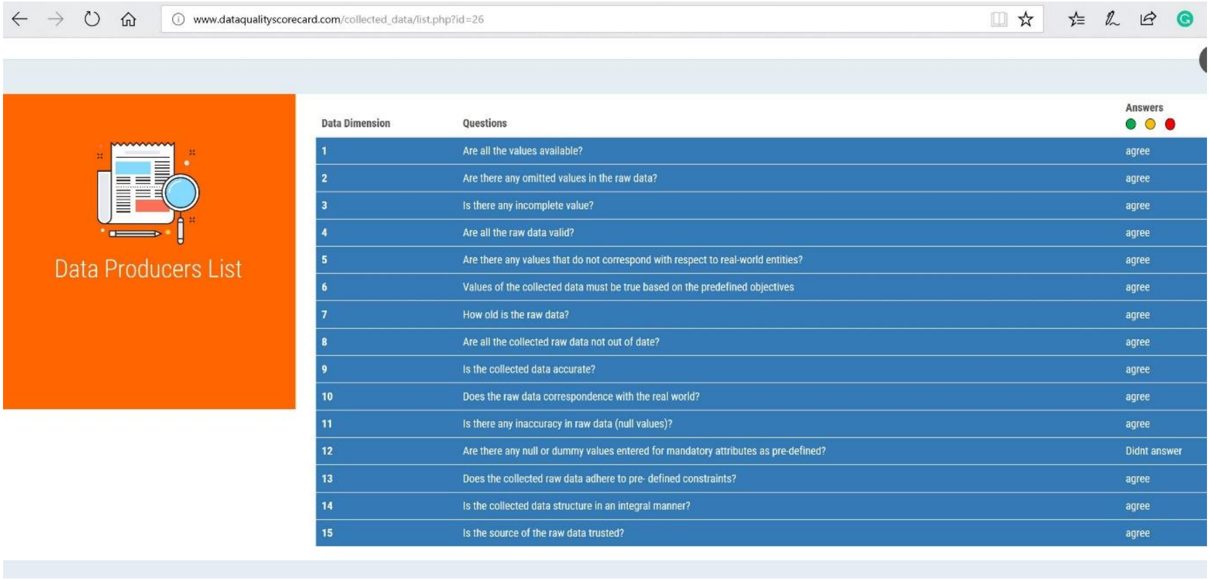
A list of all entries made by the individuals that belong to the stakeholder group selected then displays.



#	First Name	Last Name	Email	Open
1	nwulu	uchechukwu	k@gmail.com	View
2	west	gate	westgatecoinoptions@gmail.com	View
3	chigor	nwasogu	chigor@gmail.com	View
4	onw	two	two@gmail.com	View
5	Test 4	Oil and Gas	agrillo@nghamlin.com	View
6	Test 1	Test	ade.grillo@gmail.com	View
7	A	Hill	agrillo@nghamlin.com	View

Figure 38: below shows the list of reports.

Individual reports can then be displayed by selecting the required stakeholder. Figure 41 below shows an individual stakeholder report.



Data Dimension	Questions	Answers ● ● ●
1	Are all the values available?	agree
2	Are there any omitted values in the raw data?	agree
3	Is there any incomplete value?	agree
4	Are all the raw data valid?	agree
5	Are there any values that do not correspond with respect to real-world entities?	agree
6	Values of the collected data must be true based on the predefined objectives	agree
7	How old is the raw data?	agree
8	Are all the collected raw data not out of date?	agree
9	Is the collected data accurate?	agree
10	Does the raw data correspondence with the real world?	agree
11	Is there any inaccuracy in raw data (null values)?	agree
12	Are there any null or dummy values entered for mandatory attributes as pre-defined?	Didnt answer
13	Does the collected raw data adhere to pre- defined constraints?	agree
14	Is the collected data structure in an integral manner?	agree
15	Is the source of the raw data trusted?	agree

Figure 39: Individual Stakeholder DQS report

Challenges:

Accurately collecting the raw data is the primary challenge. End-user security is considered as most important and to solve problems related to end-user security; here warehouse database is integrated along with the SAP diagnostics agent.

Factors Beyond Results

Trust is a critical factor in the organisation. Employees are trusted and dedicated to their jobs. However, Data Managers, Producers, Custodians and Consumers are required to sign an NDA (Non-Disclosure Agreement) with the company.

Lessons Learned

- Raw data selection and availability of more comprehensive range and sources.
- SAP provides a useful tool to copy the database instead using the live system.
- Agreement procedure with the management of Brewery Ltd., for authorisation, should have been done at a very early stage.

5.3 Evaluation of DQS

After the run through of the DQS, the participants were then interviewed to get a first-hand view of the effectiveness of the DQS. The interview was recorded using a dictation machine and later transcribed for further analysis. The details of the participants and the intricacies of the analysis are detailed in subsequent chapters below.

5.3.1 Participants

The semi-structured interview was run with participants from each stakeholder group. A total of 20 participants across the data warehouse stakeholder groups were involved in the interview. The primary purpose of the semi-structured interview was to get verbal feedback on the usage of the data measurement scorecard. The participants were also involved in the case study.

A summary of the demography of the participants is presented in Table 12 below;

Measure	**Item**	**Frequency**	**Percentage (%)**
Stakeholder Role	Data Custodians	5	25
	Data Producers	3	15
	Data Managers	2	10
	Data Consumers	10	50
Age	21-30	4	20
	30-45	13	65
	45 and Over	3	15
Years in Industry	0-3 years	2	10
	3-7 years	3	15
	7 years +	15	75

Table 12: Semi-structured Interview Demographics

5.3.2 Procedure

Each participant was briefed on the nature of the evaluation, the required participants' gave their consent for the interview. The researcher informed the stakeholders that the purpose of the interview was to get their feedback of the DQS to aid in the further refinement of the DQS.

5.3.3 Data Collection Mechanism

The data collection was carried out through a qualitative data collection approach. Immediately after the run through was carried out, participants from the stakeholder groups were interviewed to get their views on the effectiveness of the DQS. The semi-structured interview conducted is detailed in section 5.3.5 below.

The primary purpose of the results presented is to answer the following questions;

1. Does the use of the DQS have a significant effect on the perception of the data warehouse stakeholders on the quality of data within their data warehouse?
2. Does the use of the DQS give an accurate information about the quality of data within the data warehouse?

The interview questions were analysed using a qualitative data analysis technique called thematic analysis. The intricacies of the analysis technic and results are presented in sections 5.3.6, 5.3.7 and 5.3.8 respectively.

5.3.4 Semi-Structured Interviews

The semi-structured interviews allowed the participants to provide feedback about the usage of the scorecard. Five participants from each stakeholder group took part in the interview. Arthur and Nazroo (2003) underscore the importance of careful preparation for interviews, and particularly the preparation of a “topic guide”. Their primary focus is on categorising topics to cover rather than specific questions to ask in the interview. It can be useful to have prepared essential questions ‘verbatim’, not because the question should then be asked rigidly as prepared, but because it identifies one way of asking it, which is mainly valuable if the interviewer has a memory lapse during the interview. Arthur and Nazroo advocate planning the topic guide within a frame comprising the following:

- Introduction;
- Opening questions;
- Core in---depth questions; and
- Closure.

This planning corresponds to the stages of an interview process as described by Legard et al., (2003), who present two views on in---depth interviewing. One starts from the premise that knowledge is ‘given.’ and that the researcher’s task is to dig it out; although the term was not used, this corresponds to a positivist approach. The other approach is a constructivist one: that knowledge is created and shared ground is reached through the conversation between the interviewer and interviewee. Legard et al. (p.143) emphasise the importance of establishing a relationship, noting that the interviewer is a “research instrument”, but also that researchers need “a degree of humility, the ability to be recipients of the participant’s wisdom without needing to compete by demonstrating their own”. After that, some clarification was sort from a number of the participants on what they meant by particular words or phrases.

5.3.5 Results of Evaluation

A thematic analysis technique was deployed to analyse the qualitative data items. A mix of inductive and deductive approaches was adopted. The deductive thematic analysis is an approach driven by a researcher‘s analytical or theoretical interests, while an inductive thematic analysis approach is mainly data-driven, and is based wholly on the participant's responses. The choice of using a mixed approach is motivated by the quest to avoid research bias by allowing the opportunity to identify potential new factors.

The following process was followed based on Braun and Clarke (2006) six phases of thematic analysis (pp.87---88): to conduct the analysis.

1. The researcher got familiarised with the data: after transcribing the raw data from the dictation device, it was meticulously reviewed for accuracy.
2. Initial code generation: features of the data were systematically coded about the theoretical model
3. Searching for themes: The initial themes generated from the data transcribed from the raw interview data is presented in figure 40 below;
4. Reviewing themes: The themes were reviewed, and their interrelationships were accessed. After that, strongly related themes were combined to represent a single theme as seen in Figure 41.
5. Defining and naming themes: refining the themes and the overall narrative iteratively.
6. Producing the report: which will, in turn, require a further level of reflection

The initial themes generated from the data transcribed from the interview data is presented in figure 40 below;

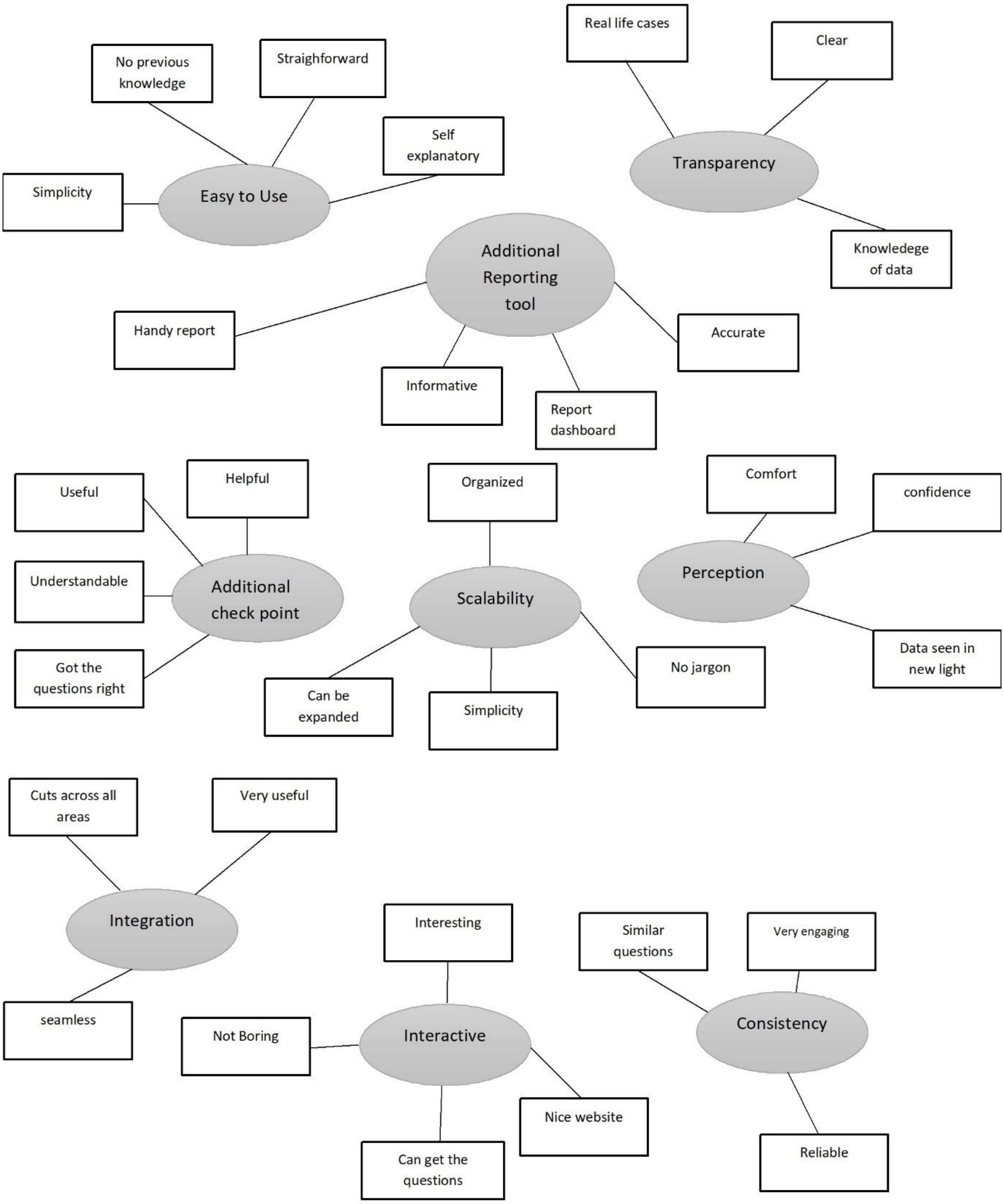


Figure 40: Thematic map of initial 9 central themes

The initial themes seen in figure 40 above were then reviewed, with closely related and overlapping themes collapsed into single themes to form the final thematic map shown in figure 41 below

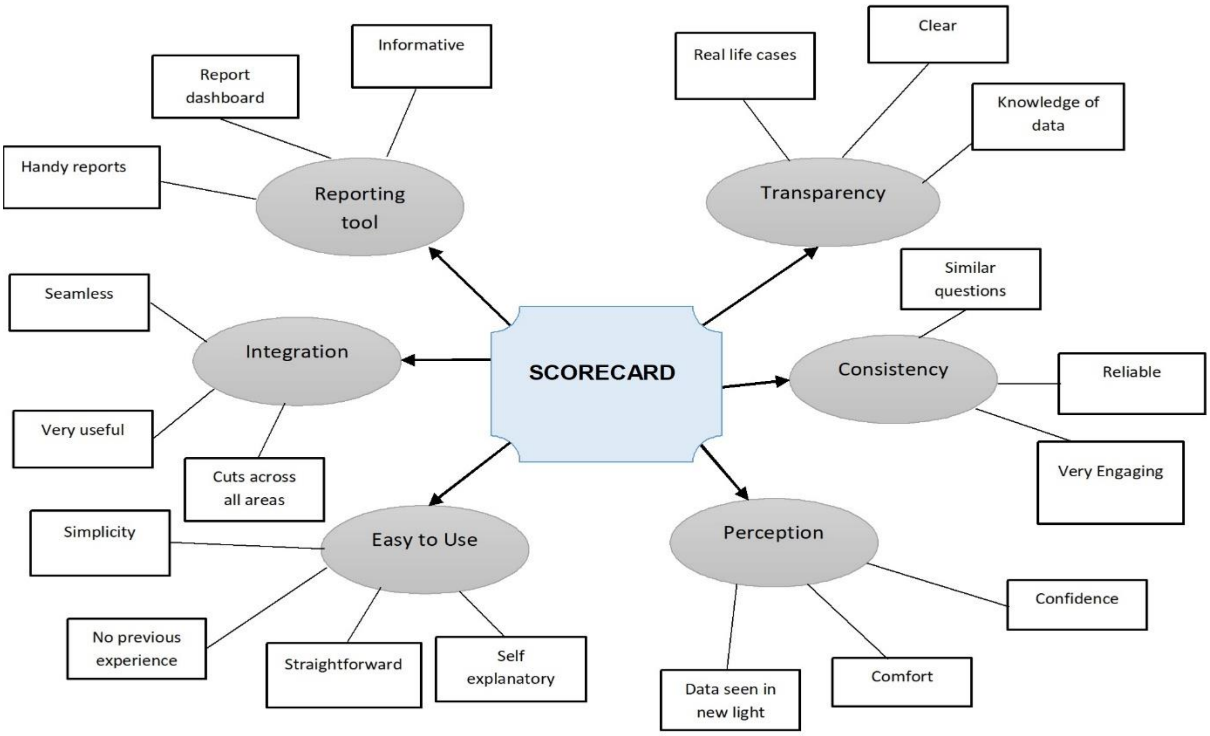


Figure 41: Final thematic map with 6 main themes

Table 13 below shows the results of the final themes and sub-themes. The responses from the participants are also shown below.

Table 13: Summary of results – Final and sub-themes

High-level themes	Sub-themes	Response from Participants
Additional Reporting Tool	(a) Extra reporting layer (b) Real business scenarios (c) Real-time analysis	**Role-Data Manager:** The scorecard provides an additional layer of reporting that gives excellent information about the data. **Role-Data Custodian:** Scenario questions represent real business concepts and are also changeable. **Role-Data Custodian:** Yes, it does provide real-time data analysis, quite a quick way of checking the quality of the database. **Role-Data Producer:** I particularly like the way the questions are relevant to our business operations.
Integration	(a) Handy tool (b) Database add-on (c) Works well with other reports	**Role-Data Manager:** Handy tool to have in addition to our other data quality measurement tools. **Role-Data Producer:** Scenario questions represent real business concepts and are also changeable. **Role-Data Producer:** Will work well with our other reports. **Role-Data Custodian:** I get the idea, and I think it's very useful. **Role-Data Manager:** I can see this integrating well with our reporting landscape. **Role-Data Manager:** Can serve as an add-on to our database.

Easy to Use	(a) Not time-consuming (b) Simple traffic light design (c) No previous knowledge required (d) Interactive and engaging	**Role-Data Consumer:** This is very interest and not time-consuming. **Role-Data Consumer:** Simple traffic light design for scoring works well and easy to understand. **Role-Data Consumer:** So easy to use, no previous knowledge required to fill the scorecard, I like that. **Role-Data Producer:** I like the concept, very interactive and engaging questions. **Role-Data Consumer:** Erm… Yes, I do like it. **Role-Data Consumer:** Very simple, straightforward. **Role-Data Consumer:** I will keep using it for scoring the data. **Role-Data Consumer:** I like the traffic lights, similar to our SAP early watch report.
Transparency	(a) End to end information (b) Information trail (c) Relevant questions	**Role-Data Manager:** Gives an end to end information of the data from inception to usage. **Role-Data Custodian:** The information trail is fantastic; I can see what everybody else thinks about the data. **Role-Data Consumer:** Very relevant questions that relate to what I do, I, however, changed one of the questions. The option to change the questions is good. **Role-Data Consumer:** I like how I can use it to view my colleague's views on the data.

Consistency	(a) Data dimensions (b) Similar questions (c) Look and feel (d) Repeatable steps	**Role-Data Producer:** The arrangement of the scorecard by data dimensions is a very good idea. **Role-Data Producer:** The look and feel of the scorecard are consistent with our approach to the measurement of data quality, I like it. **Role-Data Manager:** Data quality dimensions gives a good overall report of the data, not sure the timeliness dimension is necessary though since all our data has a timestamp. **Role-Data Custodian:** The questions are well organized, I'm sure we will find it useful. **Role-Data Manager:** Cool idea, but timeliness dimension not too important for us. **Role-Data Producer:** I really like the concept…. **Role-Data Custodian:** I like it, but the timeliness part probably needs to be removed. **Role-Data Consumer:** A very interesting and simple tool.

Perception	(a) Gives comfort (b) Provides all stakeholder viewpoint (c) Snapshot of data integrity (d) More confidence to use data	**Role-Data Consumer:** I'm more comfortable using the data with the results of the scorecard. **Role-Data Consumer:** Gives me comfort to know my colleagues have rated the data quality already. **Role-Data Manager:** I like the idea of having all the data stakeholders rating the portion of the scorecard that relates to their area. **Role-Data Consumer:** Yea, I do have more confidence using the data now than I did before the scorecard measurement. **Role-Data Producer:** Good initial reference tool to have. **Role-Data Custodian:** Provides a good snapshot of the expected quality of the data in the data warehouse. **Role-Data Manager:** The tool will give an added comfort to the quality of our data usage. **Role-Data Consumer:** My comfort level is definitely higher with the use of the scorecard.

5.3.6 Analysis of Results

This study empirically evaluated the data quality scorecard, using a qualitative approach. In Chapter 4 of this thesis, the Scorecard was designed to measure the quality of data at various stages within the development and usage of a data warehouse. To accomplish the evaluation of the scorecard, a case study was conducted that enabled the various stakeholders to use the scorecard within their data warehouse development environment, after the case study, semi-structured interviews were employed to get the views and perception of the participants. A total of 20 participants took part in the case study and semi-structured interview sessions.

Overall, each participant spent approximately one hour to participate in the study. The results of the completed scorecards from the case study showed significant added value from the usage of the scorecards, as the participants perceived the data quality scorecard as an advantageous tool.

Additionally, the participants gave verbal feedback on their viewpoints about the scorecard after the case study. Their responses supported the results of the completed scorecards from the case study. Most of the participants acknowledged that the scorecard increased their comfort levels about the data in the data warehouse. Also, the results from the semi-structured interview show that not only is the scorecard simple and straightforward to use, but it increased their perception of the quality of their data.

The findings from the thematic analysis carried out suggest that the scorecard demonstrates the following 6 key themes; (a) Additional reporting tool; (b) Integration; (c) Easy to use; (d) Transparency; (e) Consistency; and (f) Perception. Section 5.3.8 below, presents the analysis of the final themes identified in relation to the data quality scorecard.

5.3.7 Discussion

In this section, the final 6 identified themes are discussed. The themes are as follows: (a) Additional reporting tool; (b) Integration; (c) Easy to use; (d) Transparency; (e) Consistency; and (f) Perception.

(a) Additional reporting tool

Some of the participants pointed out that the scorecard could also be used as a reporting tool to augment their primary reporting suite of tools. Even though the data quality scorecard is not primarily designed as a reporting tool, but a measurement tool, it was interesting to note the multi-use of the data quality scorecard. According to some feedback from participants;

"The scorecard provides an additional layer of reporting that gives excellent information about the data"

In the literature review presented in Chapter 3, recall the researcher mentioned that the scorecard designs are usually not interactive and are mostly system driven with hardly any user interface with the software. From the literature review, it was clear that organisations predominantly focus on tools that measure the capabilities of data quality within their data warehouse, rather than a measurement tool that is interactive and can be changed to focus on specific data concern areas by the various stakeholders.

One of the participants mentioned;

"Scenario questions represent real business concepts and are also changeable"

The ability to change the questions within the scorecard is perceived as an added advantage by most participants from the data producer stakeholder group. A number of users from the data consumer stakeholder group agree that the scorecard provides them with a quick report of the data being used. According to the statements of the participants;

“Yes, it does provide real-time data analysis, quite a quick way of checking the quality of the database”

(b) Integration

Integration brings together all areas of the process into a long chain of connected activities. According to a number of statements of study participants;

“I can see this integrating well with our reporting landscape”.

“will work well with our other reports”

A key point to note in the first statement is – integration. Integration is a well-known concept in data warehouse designs. The integration of all areas within the data warehouse and associated tools like a data quality scorecard is paramount to the success of the data warehouse. In Chapter 3, one of the limitations of existing data quality measurement tools identified in the literature is – lack of proper integration. The design of the scorecard as an online repository using industry standard web development tools (see Chapter 4 for scorecard mechanics) allows for a seamless integration of the data scorecard results to any database. Other participants found the data quality scorecard somewhat relatable to their current data quality efforts, and see the scorecard as a valuable additional tool.

According to the participant's statements;

"Handy tool to have in addition to our other data quality measurement tools ".

“I get the idea, and I think it’s very useful”

From the participants‘ statements, it‘s can be deduced that the data quality scorecard would help in the companies data quality management efforts. By also using industry standard development tools, the integration of the score from the data quality scorecard can be fed back into the data warehouse, which according to one of the participants, “it will work well with other reports”.

(c) Easy to use

There is substantial evidence from the participant's responses to support how easy it is to use the data quality scorecard tool. In the context of this research, easy to use describes the extent to which the stakeholders found the ease of usage of the scorecard on a day to day basis. As discussed in Chapter 4, the mechanics of the web-centric data quality scorecard is such that ease of use was paramount in the design.

According to the statements of the participants‘;

"Simple traffic light design for scoring works well and easy to understand"

"So easy to use, no previous knowledge required to fill the scorecard, I like that"

The design of the data quality scorecard was very carefully scripted to avoid the use of too technical jargons often linked to the setup of data warehouses. The researcher ensured simplicity in the design as well as the tasks to appeal to users regardless of the number of years they have been in the industry or their computing background. Almost all users in the consumer stakeholder group who are not data warehouse professionals found the scorecard easy to use.

According to the statements of the participants from the consumer stakeholder group‘;

"Erm… Yes, I do like it"

"Very simple, straightforward"

"Simple traffic light design for scoring works well and easy to understand"

The sub-themes provides evidence that the data quality scorecard tool is engaging, works well and easy to use enough that the participants are willing to use the scorecard regularly. In the theoretical model developed and validated in Chapter 4, there is enough evidence to suggest that the simplicity (easy to use) of the data quality scorecard tool, determines the regular usage of the tool by the stakeholders.

(d) Transparency

Practically all the stakeholders showed a tendency to share the same views on the transparency of the data within the data warehouse. The source and validity of the data were paramount to most of the stakeholder groups, with the data custodian group showing a lot more concern in this area. The data dimension for validity and accuracy were seen as very essential, this also supports the theoretic design. The researcher expected that the data dimension for accuracy and validity would excite the participants as the validated theoretical model in Chapter 4 suggests, however, the rate at which almost all participants were willing to share their views on transparency, came as a welcomed surprise.

According to statements from the participants‘;

"Gives an end to end information of the data from inception to usage"

"The information trail is fantastic; I can see what everybody else thinks about the data"

“I like how I can use it to view my colleague's views about the data”

In chapter 4, validity and accuracy were empirically validated as a construct in the data quality dimension which impacts the data quality in a data warehouse.

It can be deduced from the participant's statements above that no matter how well a data quality measurement tool is designed, many of the stakeholders may not use it unless the tool exhibits a level of transparency that is obvious to the users. The data quality scorecard has the features for users to report on the scoring of other stakeholders about the same data (see scorecard mechanics in chapter 4). Through this feature, stakeholders from other groups can seamlessly have a transparent view of the data value chain.

This research considers transparency (data dimensions: Validity and accuracy) as a vital attribute of the data quality scorecard.

(e) Consistency

The research findings show that the participants recognise the consistency in the approach of the data quality scorecard. According to quotes from some participants;

"The arrangement of the scorecard by data dimensions is a very good idea"

"The look and feel of the scorecard is consistent with our approach to the measurement of data quality, I like it"

"The questions are well organized, I'm sure we will find it useful"

In the literature review conducted in Chapter 3 of this research, it was identified that syntactic and semantic data quality is seen as a very important attribute in the design of a data quality measurement tool.

Syntactic data quality concerns data's structure. The aim for syntactic data quality is consistency where the values of data for specific elements of data in the data warehouse surroundings use a consistent symbolic representation (Ballou et al., 2012; Wang et al., 2014).

Semantic data quality concerns the data meaning. The semantic quality aims are accuracy and comprehensiveness (Ding et al., 2015). Comprehensiveness is concerned with the extent to which for every similar state in the actual world system there is a value of data in the data warehouse. Accuracy is concerned with how well the values of data in data warehouse correspond to the real world state. As every stakeholder may have varied prior experience and knowledge, varied stakeholders may have varied opinions on accuracy and comprehensiveness of data warehouse.

Arguably, the findings suggest that the data quality scorecard meets the required attributes for a data quality measurement tool. According to statements from some participants below that support this position';

"The questions are well organized, I'm sure we will find it useful"

"Cool idea, but timeliness dimension not too important for us"

"I really like the concept…."

(f) Perception

Perception describes the level to which the usage of the data quality scorecard improved the stakeholder's confidence in the accuracy of the data being used. Recall in Chapter 3 in the literature review, we established that the perception of the quality of data in the data warehouse is paramount for the usage of the data for decision making by the stakeholders. The study identified sub-themes such as; (i) Gives comfort (ii) Provides all stakeholder viewpoint (iii) Snapshot of data integrity (iv) More confidence to use data

Jarke (2012) described that the researchers base their approach on the information systems notion is to offer an application domain representation also referred to as the real-world system perceived by the user.

(i) Gives comfort:

The findings show that the data quality scorecard gave the participants, especially those from the consumer stakeholder group, an added comfort level. According to the participants‘ statements;

"I'm more comfortable using the data with the results of the scorecard "

"Gives me comfort to know my colleagues have rated the data quality already"

"My comfort level is definitely higher with the use of the scorecard"

(ii) Provides all stakeholder viewpoint:

The study shows that almost all the participants found the aspect of the data quality scorecard that enables the participants to view the scoring of other stakeholders a novel idea.

According to the participants' statements;

"Gives me comfort to know my colleagues have rated the data quality already"

"I like the idea of having all data stakeholders rating the portion of the scorecard that relates to their area"

(iii) Snapshot of data integrity:

The findings also show that the data quality scorecard provided the stakeholders with a quick data integrity reference report, which stirred the participants into seeking further information about the data. According to the participants' statements;

"Provides a good snapshot of the expected quality of the data in the data warehouse"

"Good initial reference tool to have"

(iv) More confidence to use data:

The study shows that the stakeholders in practically all four groups were more confident about their data after the use of the data quality scorecard. According to the participants' statements;

"My comfort level is definitely higher with the use of the scorecard"

"I'm more comfortable using the data with the results of the scorecard "

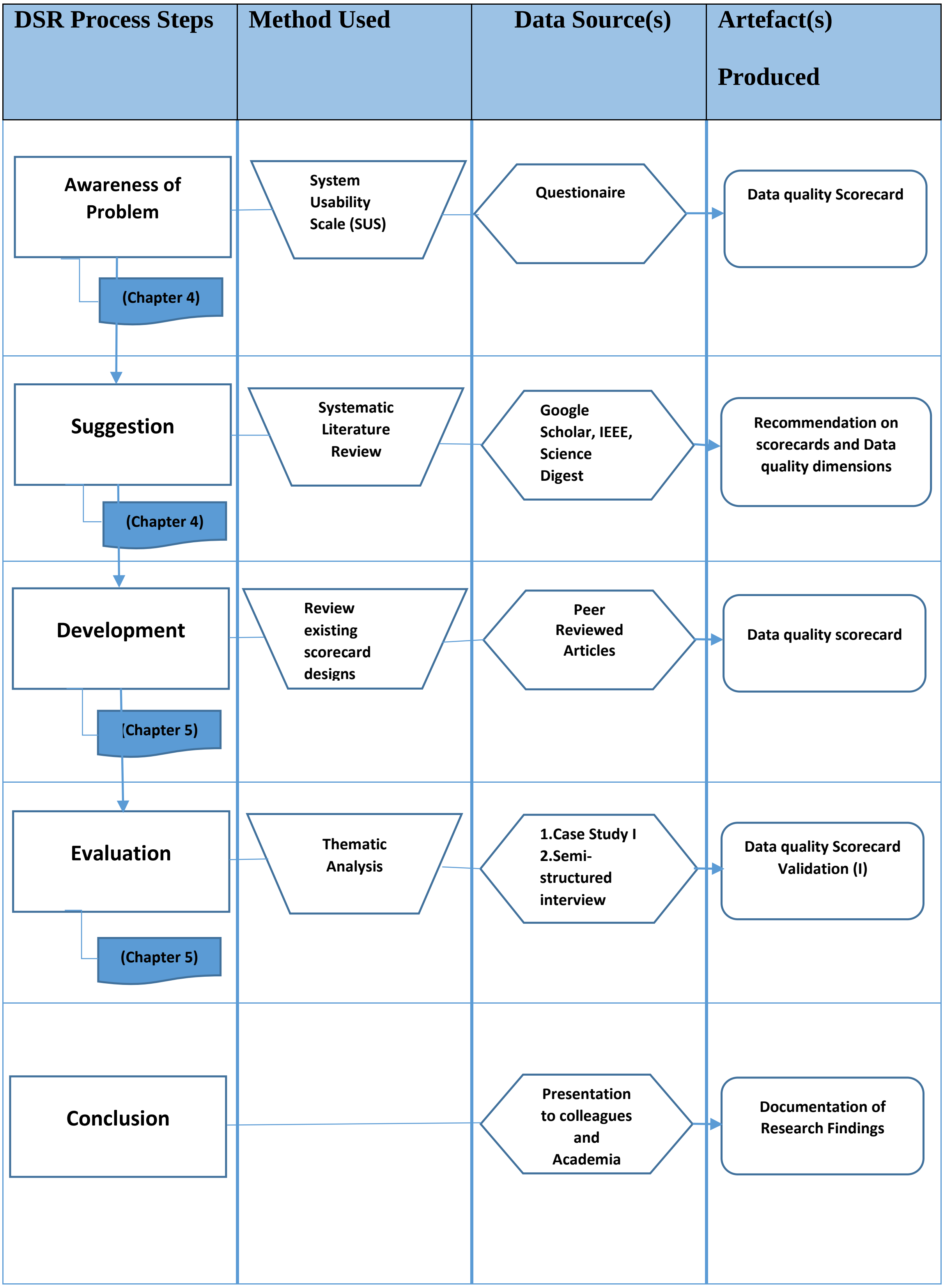
"Yea, I do have more confidence using the data now than I did before the scorecard measurement"

The sub-themes provides evidence that the data quality scorecard was able to provide the stakeholders with a positive perception of the data that in essence increased the confidence in their data

5.4 Summary

In this chapter, the researcher applied the DQS as presented in Chapter 4 in the FMCG domain, to measure the quality of data in the data warehouse within the organisation. A run through of the scorecard was carried out by the stakeholder. After the run through of the DQS, the participants were then interviewed to get a first-hand view of the effectiveness of the DQS. The interview was recorded using a dictation machine and later transcribed for further analysis. The DQS was discovered to provide an improvement in a number of the selected data dimensions tested. However, it was noted by a number of stakeholders that the timeliness dimension is redundant as a timestamp is standard in their data warehouse. Arguably, the findings suggest that the data quality scorecard meets the required attributes for a data quality measurement tool. According to statements from most participants that support this position. In the next Chapter, the modified DQS is evaluated in the Oil and Gas domain.

Figure 42 below summarises the second DSR iteration with the methods and techniques used.



DSR Process Steps	Method Used	Data Source(s)	Artefact(s) Produced
Awareness of Problem (Chapter 4)	System Usability Scale (SUS)	Questionaire	Data quality Scorecard
Suggestion (Chapter 4)	Systematic Literature Review	Google Scholar, IEEE, Science Digest	Recommendation on scorecards and Data quality dimensions
Development (Chapter 5)	Review existing scorecard designs	Peer Reviewed Articles	Data quality scorecard
Evaluation (Chapter 5)	Thematic Analysis	1.Case Study I 2.Semi-structured interview	Data quality Scorecard Validation (I)
Conclusion		Presentation to colleagues and Academia	Documentation of Research Findings

Figure 42: Result of second DSR Iteration cycle

Chapter 6: DQS Evaluation - Iteration III

6.1 Introduction

In the previous chapter, the researcher evaluated the DQS through a live run through in a company in the FMCG domain. A semi-structured interview was conducted thereafter to get the views of the DW stakeholder group. The semi-structured interview data was then analysed for themes using a technique called thematic analysis. The results from the previous chapter suggested that the removal of timeliness dimension would enhance the usage of the scorecard as most data warehouses have a standard timestamp on all data. Hence, the DQS was modified.

This chapter reports on the run through of the modified data quality scorecard in a live environment and then evaluate the usefulness of the scorecard by conducting a semi-structured interview with the relevant stakeholders. Qualitative methods were employed for this evaluation, which includes: (1) Case study and (2) Semi-structured interview. The chapter is structured as follows: **Section 6.2** describes the second run through of the scorecard to a specific case study. The case study involves an organisation in the Oil and Gas (O&G) sector. In this section, the researcher will briefly present the architecture of the organisation's data warehouse, the problems they are currently facing and the application of the DQS **Section 6.3** describe the evaluation process, the semi-structured interview conducted and the procedures, results and further discussions. The chapter concludes with a summary in **Section 6.4**.

6.2 About company Oil and Gas Ltd. (O&G Ltd.)

O&G Ltd. is a worldwide group of petrochemical and energy firms. The parent firm is located in Europe, which is a multinational oil and gas firm with a worldwide presence. O&G Ltd. is one of the biggest firms on the globe in 2014 in revenue terms. The strategy of the Company is to produce and ensure sustained profitable development, remains to drive forward with their investment program to deliver sustainable development and offer competitive gains to shareholders while helping to meet global demand for energy in a reliable way. O&G Ltd. focuses on mining for new oil and gas reserves in the upstream oil and gas sector, evolving leading projects where their know-how and technology adds value to the holders of the resource. Similarly, in the downstream oil and gas sector, their emphasis remains on supporting the generation of revenue from their existing assets and selective investments in developed markets. As a worldwide energy firm, the Company sets greater ethical behaviours and performance standards. They are judged by how they perform and their status is upheld by how they live up to their core values namely respect, honesty and integrity for people. The general business principles of O&G Ltd.'s code of ethics and code of conduct helps all employee's act in line with these values and comply with the entire similar regulations and legislation. Their major aim is to meet the energy needs of society in ways that are social, environmentally and economically essential now and in future. The major aim of the Company is to employ responsible, efficient standards and tools in the oil and gas industry to achieve sustainable growth of resources of energy.

6.2.1 The Data Quality Problem at Oil and Gas Ltd:

Oil and Gas Ltd is a huge company and it is important to maintain a big database with proper security. One unchecked step may harm the whole system. Oil and Gas Ltd has a database virtualization deployed to increase the effectiveness of query response time. After the initial consultation with the stakeholder group within the company, the following areas were identified as susceptible to quality issues. The data quality framework will be tested against these quality checkpoints, and an estimated value improvement scorecard tabulated. The susceptible areas as identified are:

1. During interface with data sources
2. During data integration and profiling
3. During extraction, transformation and loading (ETL)
4. During base data modelling (Schema design)

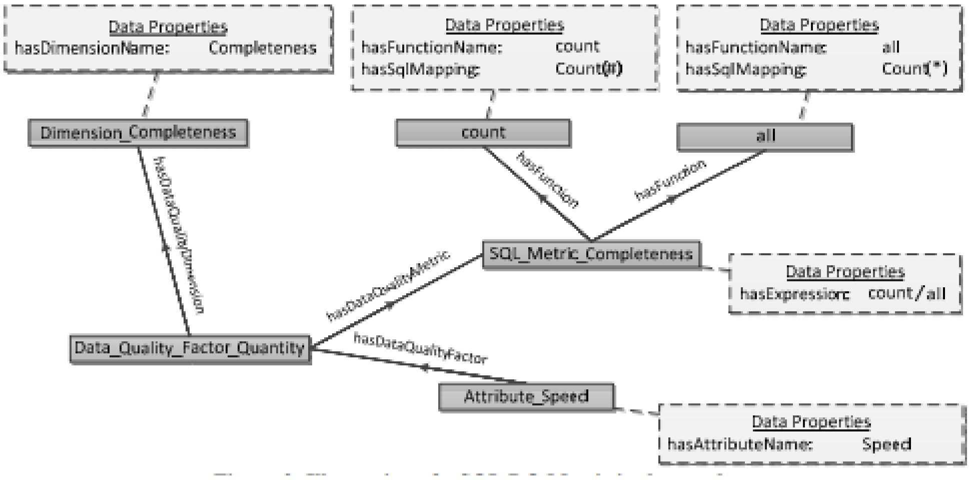


Figure 43: Structured Query Language Data Quality Metric **(**Jeung H et al., 2010)

Steps Taken to Address the Problem

The data quality scorecard as a first step will be deployed to ascertain the data quality degradation based on the scorecard matrix. The stakeholder group will test the framework with customizable parameters using the defined data quality dimensions.

The Stakeholder Group at Oil and Gas Ltd

Data Custodian:

The data custodian is liable for implementing safeguards of data storage and assuring data availability. The role of the custodian is to assist the user of business and if something is incorrect it is the liability of custodian to manage with this promptly. The data custodians must be appointed or nominated for specific sets of data over which they have control. For such data, the custodians are generally those groups of individuals who have the major responsibility for managing the data that is their function of the job includes the regular data updates. The data custodian duties involve the following namely: 1) administer controls of user access; 2) implement controls matching classification and information; 3) supervise security of data for violations; 4) back up information to secure from loss; 5) assure the integrity of data through controls of processing; and 6) be feasible to resolve any issues. The data custodians mainly determine what data is gathered, released and aggregated to the public. In this case, the data Custodian uses the SAP technology for composites and database maintenance. The virtualization technology is used in this case to operate this large database and it is regarded as a more efficient way.

Data producer:

Data producer refers the metadata and it is possible that more than one data producer refers to similar metadata and similar data producer refers greater than one metadata. In several cases, data producers are rapid but not essentially exact. Any time a data producer is paid solely based on how many new customers, new accounts or new policies can be entered and they will figure out how to game the system. Data producers always do not have visibility into the influence of their decisions on data customers. The data producer is responsible for data input quality into the source systems. In this case, the data producer forwards the raw data to Data Custodian.

Virtualization is helpful in this case to reduce the cost, to reduce buildup o heat, to redeploy rapidly, easier backup, easy testing, rapid recovery of disaster, easy cloud migration and moreover, it is much advantageous.

Data Manager:

Data manager coordinates the data partitioning process and replication across varied sites. The data manager performs as an intermediary between the business warehouse objects in metadata and the technical data storage in associated database systems. Data manager enhances tools to access to business warehouse system. Data manager manage any data and the data must be prepared and received by the extractors of respective source systems. In this case, the data manager checks the transferred data and verify it whether it is true or not. The responsibility of the data manager is to take care of security. SAP Hana provides the best solution for optimization and security of data warehouse. In the world of enterprise data warehouse SAP, HANA is a disruptive technique truly as it has redefined technical possibilities. By providing magnitude improvement orders in reporting performance while reducing data redundancy and staging simultaneously, HANA largely pays for itself by simply lowering maintenance and development effort. However, its maximum value will be realized by those who captured the chance to leverage its performance in non-traditional ways. The businesses that use SAP HANA as a stepping stone to implementing predictive analytics create federated data warehouses or enhance real-time reporting will reap the largest advantages from this technique. Those with less lofty ambitions will still handle to perform similar things they perform today just with larger efficacy and much-developed performance. Regardless of the objective of business the first step towards accomplishing outcomes generally initiates with a decision to migrate to SAP HANA. The data is encrypted on the server with a strong encryption algorithm and only admin can access sensitive data. Untrusted participation in Oil and Gas Ltd database is limited.

Paramount Data Quality Dimension at Oil and Gas Ltd:

Completeness: The extent to which data is populated based on rules of business that state when data is needed to be populated with a value. A much complex norm might state that a collateral record is needed if and only if a record of loan is present and the type of loan needs to be collateral. In this case, the data is complete and there is no missing value during the testing of the Oil and Gas Ltd warehouse.

Validity: The extent to which the data conforms to the rules of business for acceptable content. This can involve pattern, format, type of data, domain, range and valid value list. In this case, the Oil and Gas Ltd operated with hard technique and so it is not altering frequently.

Accuracy: The extent to which data corresponds to known correct values in the real world as offered by an established or recognized truth source. In several cases, accuracy is estimated by how the values agree with an identified source of proper data. There are varied sources of correct information namely a record of the database, a similar corroborative data values set from another table, dynamically evaluated values or the outcome of the manual process. Accuracy is quite challenging to supervise not just because one needs a secondary source for corroboration because real-world data may alter over time. In this case, the data accuracy can be interrupted during a disaster.

Timeliness: The extent to which alterations to data are feasible within the timeline needed by the business. For instance, the alteration to an allotted airline seat must be reflected on the website in real time. In this case, the data is valid for a long time.

Consistency: Consistency defines to values of data set is consistent with values in another set of data. Consistency specifies that the values of data drawn from separate sets of data must not conflict with each other. The idea of consistency with a set of predefined limitations can be much complicated. The energy sector has more priority. Especially natural gas.

Integrity: A measure of the validity, existence, content, structure and other basic features of the data. All other quality dimensions build on what is learned in the fundamentals of data integrity. This dimension involves basic data quality measures such as fill rate/completeness, frequency distributions, validity, lists of values, ranges, patterns, referential integrity and minimum and maximum values. It is the duty of Data Manager to check data is true or not. In this step, the company can be assured that data already checked by data Manager. So, no need to worry.

Challenges:

Collecting the raw data accurately is the major challenge. This study has taken product data from Oil and Gas Ltd for their analysis. Virtualization of big data is most typical and so SAP technology is considered for this project. For the database, the ABAP Programming is a huge support which matches with SAP technology. The SAP managed system & solution manager are linked with the services of the web. At last, end-user security is considered as most important and to solve problems related to end-user security, here warehouse database is integrated along with the SAP diagnostics agent.

Factors Beyond Results:

Trust is the essential factor for any organization. Employees are dedicated and trusted in Oil and Gas Ltd for their jobs. If anyone in Oil and Gas Ltd leak the database to other organization it may cause a huge amount of loss for Oil and Gas Ltd. Data Custodian and Data Manager must sign a Non-Disclosure Agreement with Oil and Gas Ltd.

Lessons Learned:

1) The RAW data is selected and major priority is provided for them from the big resource.
2) SAP offers usefulness in the maintenance of data warehouse.\
3) Composites (CISCO) is used as the data virtualization technology.
4) The legal agreement procedure is signed by the employees of Oil and Gas Ltd.

The maintenance of the data warehouse is a huge task in Oil and Gas Ltd. The infrastructure of the data warehouse is regularly updated with the latest techniques. SAP offers support for security for data. One incorrect step can be a reason for the huge loss in the database. Carefulness and accuracy are essential in this case. Some of the steps used to maintain the data warehouse successfully in future involves: 1) reduction of power use; 2) extend the safety; 3) advanced virtualization; 4) combination with more than one data warehouse; 5) enhance the physical server stability; 6) supervising the conditions of data warehouse to get alert when important; and 7) powerful security.

Reduction of power use:

Reducing the use of power is essential for the architecture of the data warehouse. It lowers the overall cost of the firm. The company should view for a platform that is easy and rapid to maintain and greater capacity servers save time.

Extend the safety:

The solutions to the data warehouse must be resilient and reliable. Efficiency and transparency is the key to managing big data warehouse.

Advanced Virtualization:-

Hadoop big data analysis and data virtualization technology are very useful to maintain big data warehouses. It helps to manage data safely as well as to make and track a better report of analytics.

The combination of more than one data warehouse:

When more than one enterprise data warehouse is built the integration is known as a distributed data warehouse. A data warehouse is virtually any database comprising data from more than one source gathered for the purpose of offering management data.

Supervising the conditions of the data warehouse to get alert when important:

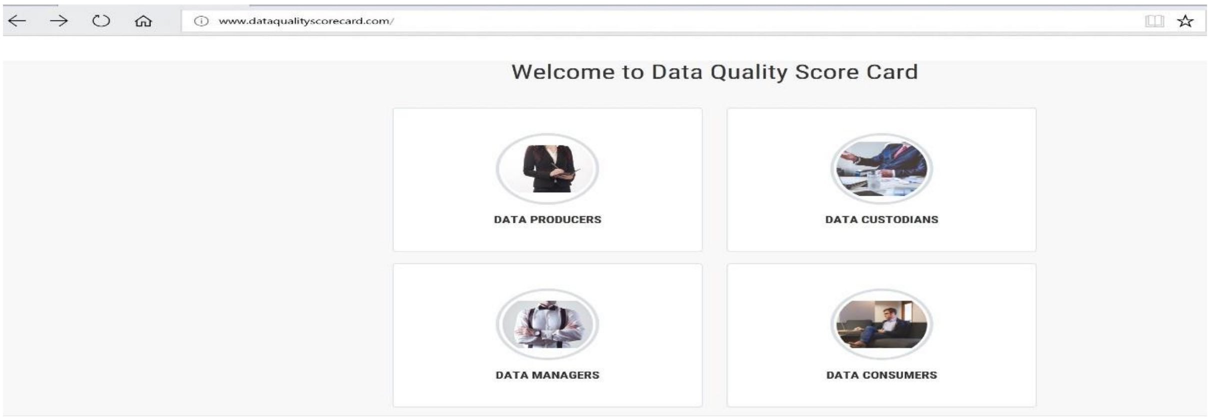
The scope of monitoring activity in a data warehouse expands over several functions and features. Unless data warehouse monitoring takes place in a formalized way desired outcomes cannot be accomplished. The outcomes of monitoring provide the data required to plan for developing and to evolve performance.

Powerful security:

The aim of every data warehouse is to make available to all concerned the information they require and too much security may have the consequence that users do not have access to all data that is essential to perform their job.

6.2.2 Web-centric DQS-Iteration III

The stakeholders were then directed to the website located at www.dataqualityscorecard.com to fill out the form about the quality of the data that was just entered or viewed into/from the data warehouse. The figures below show the login screen and the DQS from each of the stakeholder group.

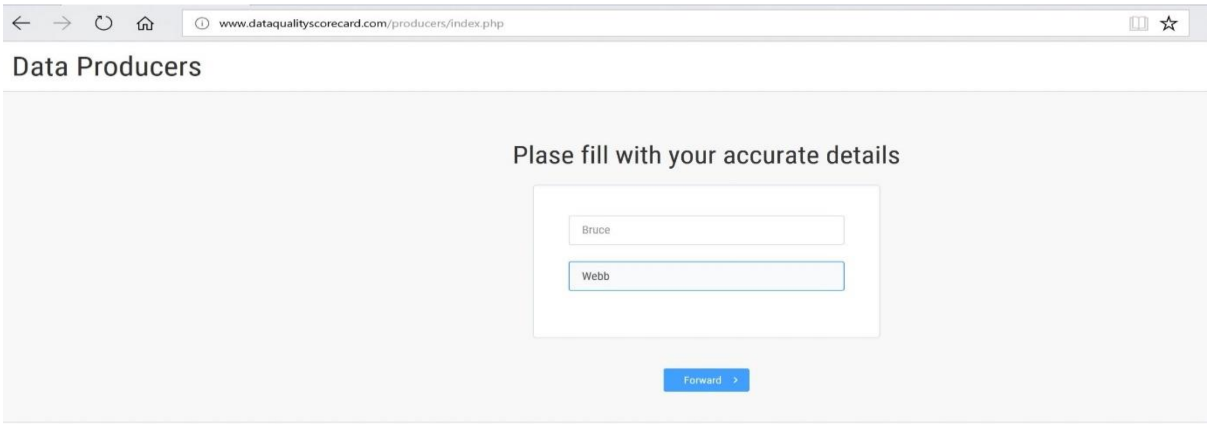


www.dataqualityscorecard.com/

Welcome to Data Quality Score Card

Figure 44: DQS Login Screen

All stakeholders will be required to enter their names after selecting the group they belong to in the organization. The stakeholder's details will be validated internally using the relevant authorization metric in the organisation. Figure 45 below shows a stakeholder specific screen


www.dataqualityscorecard.com/producers/index.php

Data Producers

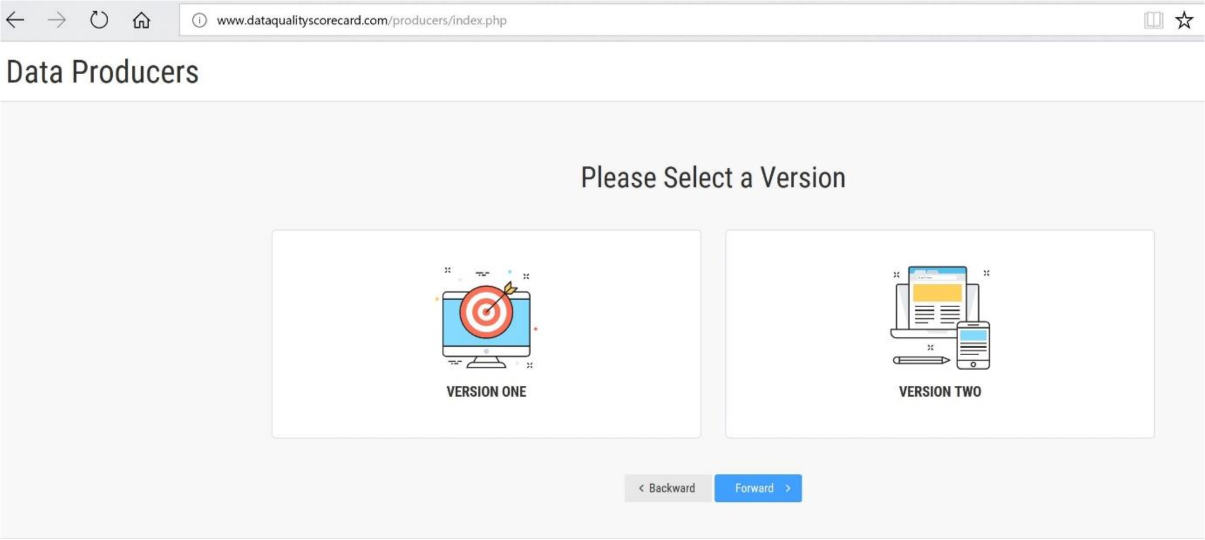
Plase fill with your accurate details

Bruce

Webb

Figure 45: Stakeholder specific login screen

Immediately after login, the stakeholder will be presented with a version selection screen as shown in figure 46 below. The versions represent the modification that has been carried out as a result of the previous evaluation. The stakeholders at Oil and Gas Ltd were advised to use the newer version 2. Figure 46 below shows the version selection screen

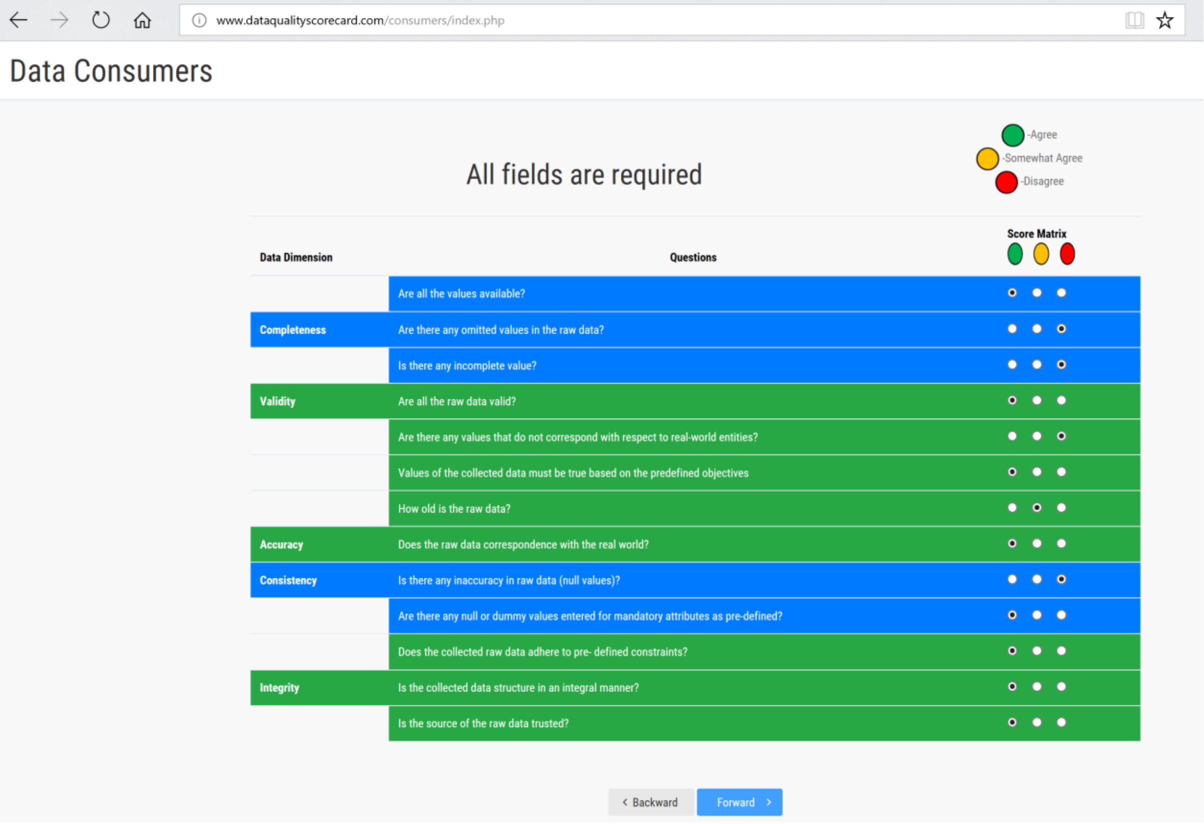

www.dataqualityscorecard.com/producers/index.php

Data Producers

Please Select a Version

Figure 46: DQS version selection screen

The stakeholders are then presented with a series of questions that reflect their concern areas. The stakeholder concern areas according to literature are varied and reflect the needs of the various groups. The DQS has been modified to account for the recommendations from the previous evaluation. The figures below depict the quality score attributed to each concern area by the stakeholder groups.



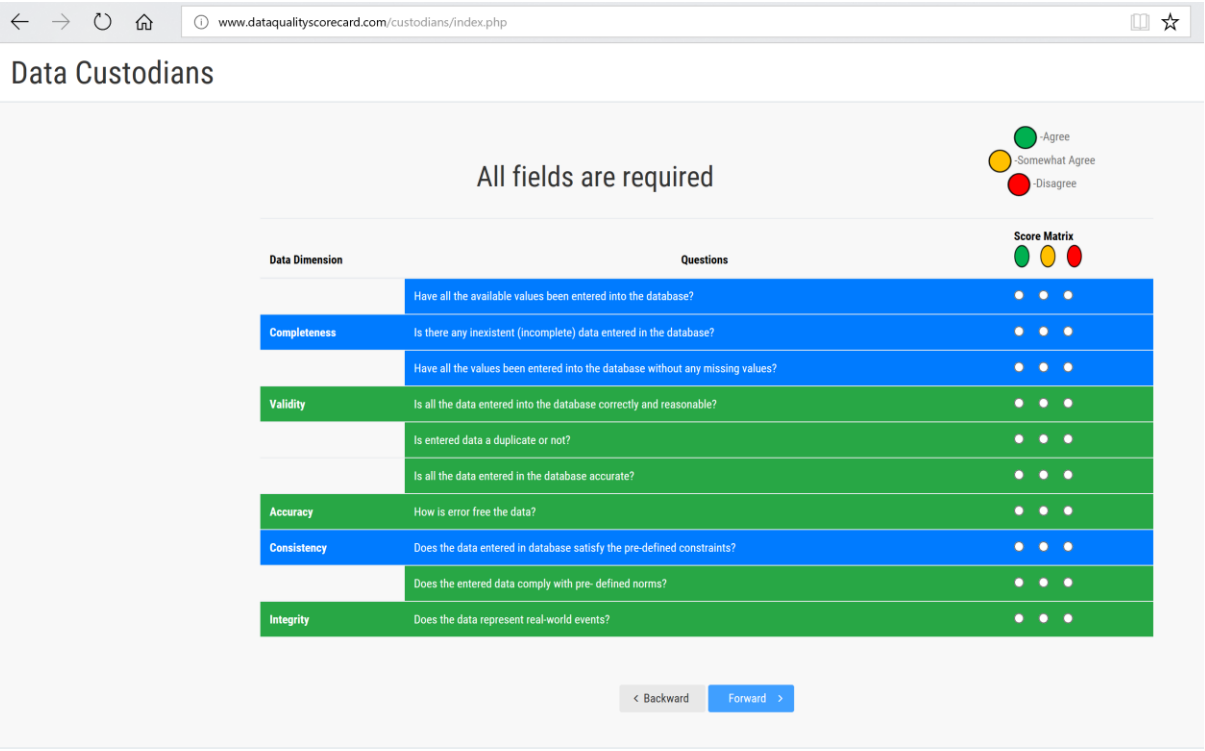
Data Consumers

All fields are required

🟢 -Agree
🟡 -Somewhat Agree
🔴 -Disagree

Data Dimension	Questions	Score Matrix		
		🟢	🟡	🔴
	Are all the values available?	●	○	○
Completeness	Are there any omitted values in the raw data?	○	○	●
	Is there any incomplete value?	○	○	●
Validity	Are all the raw data valid?	●	○	○
	Are there any values that do not correspond with respect to real-world entities?	○	○	●
	Values of the collected data must be true based on the predefined objectives	●	○	○
	How old is the raw data?	○	●	○
Accuracy	Does the raw data correspondence with the real world?	●	○	○
Consistency	Is there any inaccuracy in raw data (null values)?	○	○	●
	Are there any null or dummy values entered for mandatory attributes as pre-defined?	●	○	○
	Does the collected raw data adhere to pre- defined constraints?	●	○	○
Integrity	Is the collected data structure in an integral manner?	●	○	○
	Is the source of the raw data trusted?	●	○	○

Figure 47: Data Consumer DQS v2



www.dataqualityscorecard.com/custodians/index.php

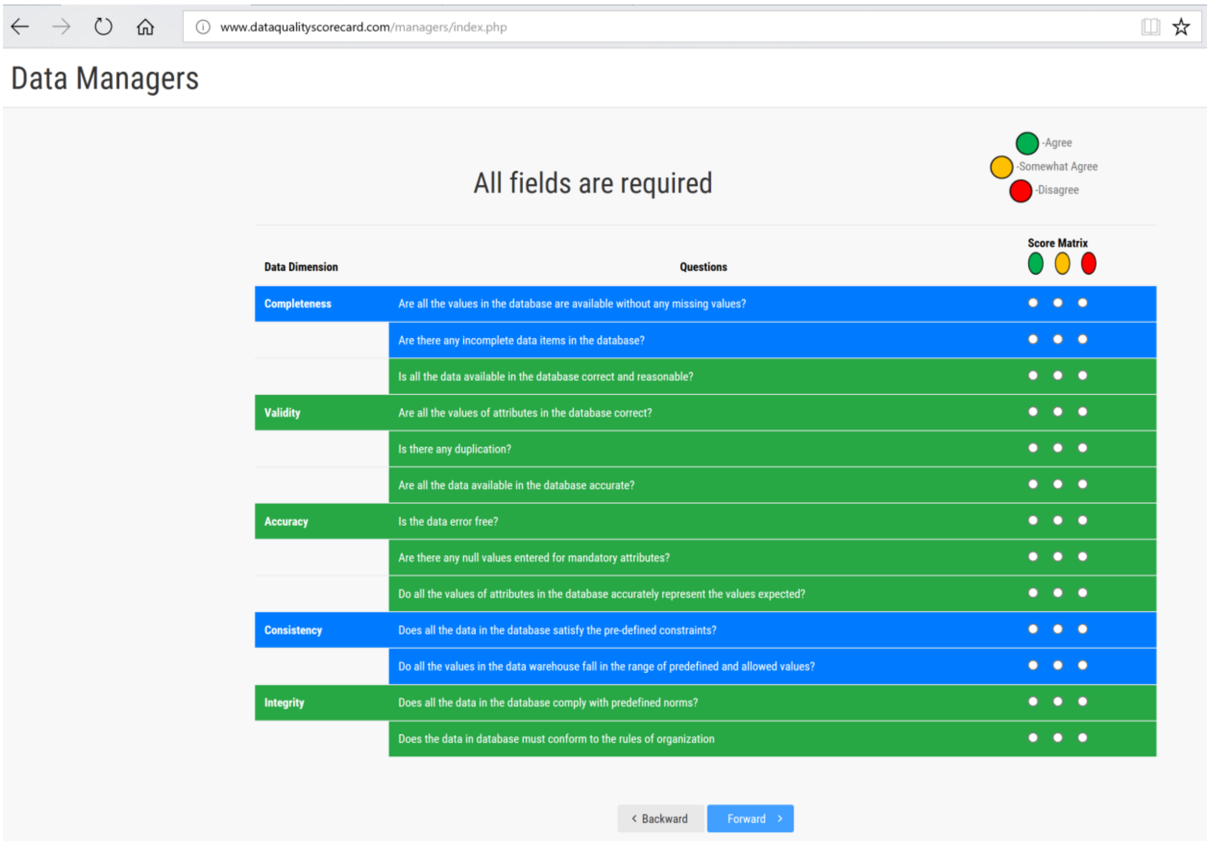
Data Custodians

All fields are required

🟢 -Agree
🟡 -Somewhat Agree
🔴 -Disagree

Data Dimension	Questions	Score Matrix 🟢 🟡 🔴
	Have all the available values been entered into the database?	○ ○ ○
Completeness	Is there any inexistent (incomplete) data entered in the database?	○ ○ ○
	Have all the values been entered into the database without any missing values?	○ ○ ○
Validity	Is all the data entered into the database correctly and reasonable?	○ ○ ○
	Is entered data a duplicate or not?	○ ○ ○
	Is all the data entered in the database accurate?	○ ○ ○
Accuracy	How is error free the data?	○ ○ ○
Consistency	Does the data entered in database satisfy the pre-defined constraints?	○ ○ ○
	Does the entered data comply with pre- defined norms?	○ ○ ○
Integrity	Does the data represent real-world events?	○ ○ ○

Figure 48: Data Custodian DQS v2



www.dataqualityscorecard.com/managers/index.php

Data Managers

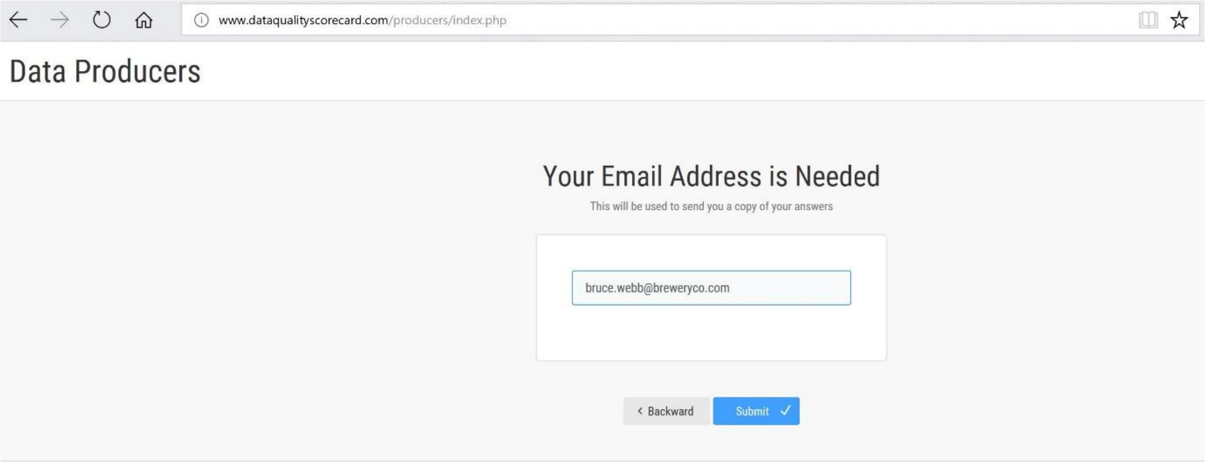
All fields are required

🟢 -Agree
🟡 -Somewhat Agree
🔴 -Disagree

Data Dimension	Questions	Score Matrix 🟢 🟡 🔴
Completeness	Are all the values in the database are available without any missing values?	○ ○ ○
	Are there any incomplete data items in the database?	○ ○ ○
	Is all the data available in the database correct and reasonable?	○ ○ ○
Validity	Are all the values of attributes in the database correct?	○ ○ ○
	Is there any duplication?	○ ○ ○
	Are all the data available in the database accurate?	○ ○ ○
Accuracy	Is the data error free?	○ ○ ○
	Are there any null values entered for mandatory attributes?	○ ○ ○
	Do all the values of attributes in the database accurately represent the values expected?	○ ○ ○
Consistency	Does all the data in the database satisfy the pre-defined constraints?	○ ○ ○
	Do all the values in the data warehouse fall in the range of predefined and allowed values?	○ ○ ○
Integrity	Does all the data in the database comply with predefined norms?	○ ○ ○
	Does the data in database must conform to the rules of organization	○ ○ ○

Figure 49: Data Manager DQS v2

The stakeholders are then required to enter their email address for the DQS to be stored. The email can be made an optional step by the various organizations if required. The figure below shows the email screen


www.dataqualityscorecard.com/producers/index.php

Data Producers

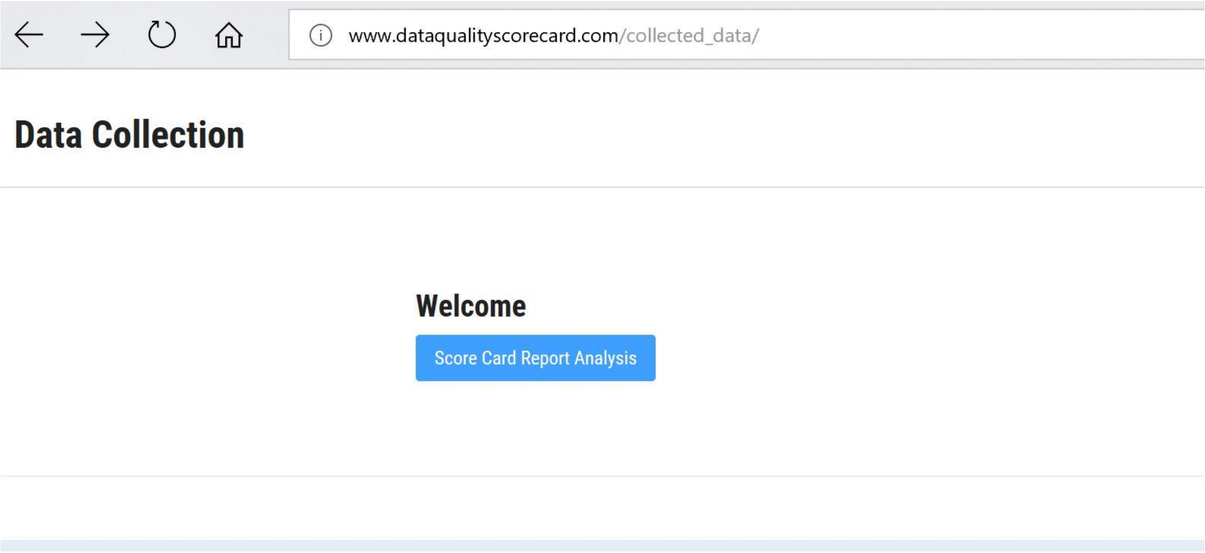
Your Email Address is Needed

This will be used to send you a copy of your answers

bruce.webb@breweryco.com

Figure 50: DQS Email screen v2

The DQS results can be viewed by the individual stakeholders or as a collective. The reports can be viewed by login into www.dataqualityscorecard.com/collected_data. The figures below show the report delineated by stakeholder groups and then by individuals.

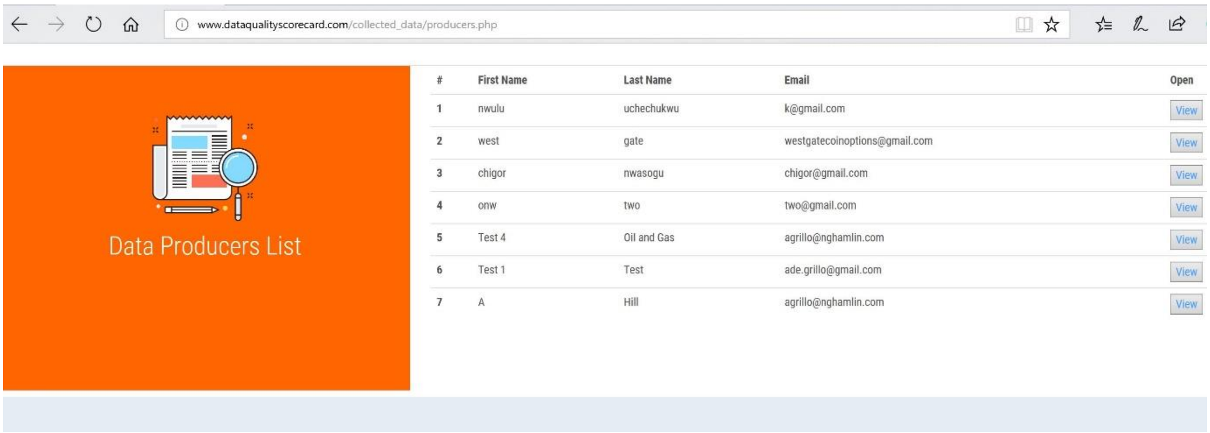

www.dataqualityscorecard.com/collected_data/

Data Collection

Welcome

Figure 51: DQS report area

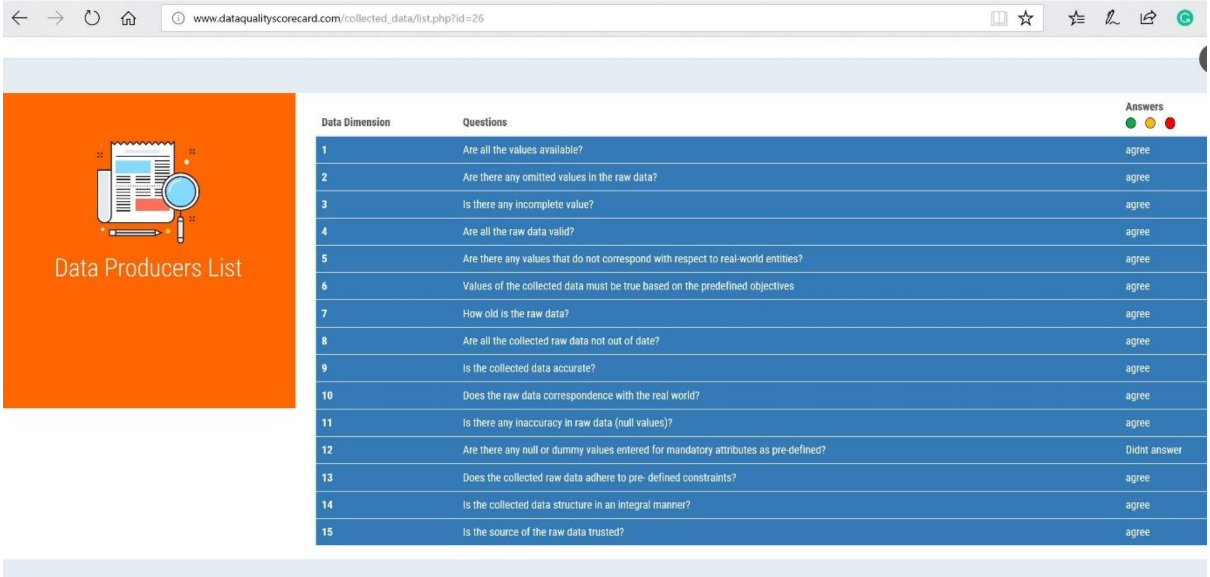
A list of all entries made by the individuals that belong to the stakeholder group is then displayed.



#	First Name	Last Name	Email	Open
1	nwulu	uchechukwu	k@gmail.com	View
2	west	gate	westgatecoinoptions@gmail.com	View
3	chigor	nwasogu	chigor@gmail.com	View
4	onw	two	two@gmail.com	View
5	Test 4	Oil and Gas	agrillo@nghamlin.com	View
6	Test 1	Test	ade.grillo@gmail.com	View
7	A	Hill	agrillo@nghamlin.com	View

Figure 52: Stakeholder list of reports.

Individual reports can then be displayed by selecting the required stakeholder. Figure 53 below shows an individual stakeholder report.



Data Dimension	Questions	Answers ● ● ●
1	Are all the values available?	agree
2	Are there any omitted values in the raw data?	agree
3	Is there any incomplete value?	agree
4	Are all the raw data valid?	agree
5	Are there any values that do not correspond with respect to real-world entities?	agree
6	Values of the collected data must be true based on the predefined objectives	agree
7	How old is the raw data?	agree
8	Are all the collected raw data not out of date?	agree
9	Is the collected data accurate?	agree
10	Does the raw data correspondence with the real world?	agree
11	Is there any inaccuracy in raw data (null values)?	agree
12	Are there any null or dummy values entered for mandatory attributes as pre-defined?	Didnt answer
13	Does the collected raw data adhere to pre- defined constraints?	agree
14	Is the collected data structure in an integral manner?	agree
15	Is the source of the raw data trusted?	agree

Figure 53: Individual Stakeholder DQS report

6.3 Evaluation of DQS

After the run through of the modified DQS, the participants were then interviewed to get a first-hand view of the effectiveness of the DQS. The interview was recorded using a dictation machine and later transcribed for further analysis. The details of the participants and the intricacies of the analysis are detailed in subsequent chapters below.

6.3.1 Participants

The semi-structured interview was run with participants from each stakeholder group. A total of 15 participants across the data warehouse stakeholder groups were involved in the interview. The primary purpose of the semi-structured interview was to get verbal feedback on the usage of the data measurement scorecard. The participants were also involved in the case study.

A summary of the demography of the participants is presented in Table 14 below;

Measure	**Item**	**Frequency**	**Percentage (%)**
Stakeholder Role	Data Custodians	3	20
	Data Producers	3	20
	Data Managers	2	13.3
	Data Consumers	7	46.7
Age	21-30	2	13.3
	30-45	3	20
	45 and Over	10	66.7
Years in Industry	0-3 years	1	6.7
	3-7 years	3	20
	7 years +	11	73.3

Table 14: Semi-structured Interview Demographics

6.3.2 Procedure

Each participant was briefed on the nature of the evaluation, the required participants' gave their consent for the interview. The researcher informed the stakeholders that the purpose of the interview was to get their feedback of the DQS to aid in the further refinement of the DQS.

6.3.3 Data Collection Mechanism

The data collection was carried out through a qualitative data collection approach. Immediately after the run through was carried out, participants from the stakeholder groups were interviewed to get their views on the effectiveness of the DQS. The semi-structured interview conducted is detailed in section 6.3.4 below.

The primary purpose of the results presented is to answer the following questions;

3. Does the use of the DQS have a significant effect on the perception of the data warehouse stakeholders on the quality of data within their data warehouse?
4. Does the use of the DQS give an accurate information about the quality of data within the data warehouse?

The interview questions were analysed using a qualitative data analysis technique called thematic analysis. The intricacies of the analysis technic and results are presented in sections 6.3.5, 6.3.6 and 6.3.7 respectively.

6.3.4 Semi-Structured Interviews

The semi-structured interviews allowed the participants to provide feedback about the usage of the scorecard. Five participants from each stakeholder group took part in the interview. Arthur and Nazroo (2003) underscore the importance of careful preparation for interviews, and particularly the preparation of a “topic guide”. Their primary focus is on categorising topics to cover rather than specific questions to ask in the interview. It can be useful to have prepared essential questions ‘verbatim’, not because the question should then be asked rigidly as prepared, but because it identifies one way of asking it, which is mainly valuable if the interviewer has a memory lapse during the interview. Arthur and Nazroo advocate planning the topic guide within a frame comprising the following:

- Introduction;
- Opening questions;
- Core in---depth questions; and
- Closure.

This planning corresponds to the stages of an interview process as described by Legard et al. (2003), who present two views on in---depth interviewing. One starts from the premise that knowledge is ‘given.’ and that the researcher’s task is to dig it out; although the term was not used, this corresponds to a positivist approach. The other approach is a constructivist one: that knowledge is created and shared ground is reached through the conversation between the interviewer and interviewee. Legard et al. (p.143) emphasise the importance of establishing a relationship, noting that the interviewer is a “research instrument”, but also that researchers need “a degree of humility, the ability to be recipients of the participant’s wisdom without needing to compete by demonstrating their own”. After that, some clarification was sort from a number of the participants on what they meant by particular words or phrases.

6.3.5 Results of Evaluation

A thematic analysis technique was deployed to analyse the qualitative data items. A mix of inductive and deductive approaches was adopted. The deductive thematic analysis is an approach driven by a researcher‘s analytical or theoretical interests, while an inductive thematic analysis approach is mainly data-driven, and is based wholly on the participant's responses. The choice of using a mixed approach is motivated by the quest to avoid research bias by allowing the opportunity to identify potential new factors.

The following process was followed based on Braun and Clarke (2006) six phases of thematic analysis (pp.87---88): to conduct the analysis.

7. The researcher got familiarised with the data: after transcribing the raw data from the dictation device, it was meticulously reviewed for accuracy.
8. Initial code generation: features of the data were systematically coded about the theoretical model
9. Searching for themes: The initial themes generated from the data transcribed from the raw interview data is presented in figure 54 below;
10. Reviewing themes: The themes were reviewed, and their interrelationships were accessed. After that, strongly related themes were combined to represent a single theme as seen in Figure 55.
11. Defining and naming themes: refining the themes and the overall narrative iteratively.
12. Producing the report: which will, in turn, require a further level of reflection

The initial themes generated from the data transcribed from the interview data is presented in figure 54 below;

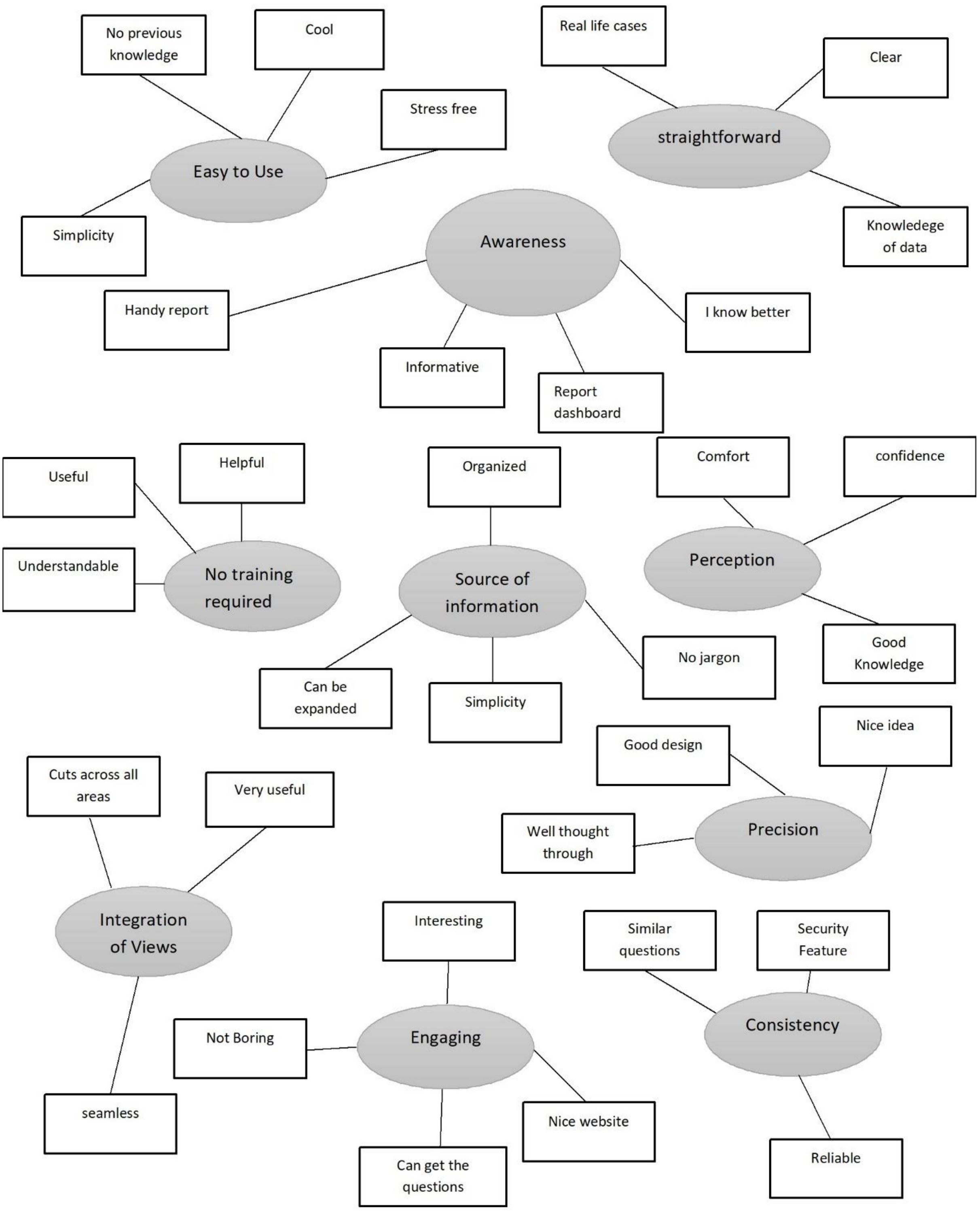


Figure 54: Thematic map of initial 10 central themes

The initial themes seen in figure 54 above were then reviewed, with closely related and overlapping themes collapsed into single themes to form the final thematic map shown in figure 55 below

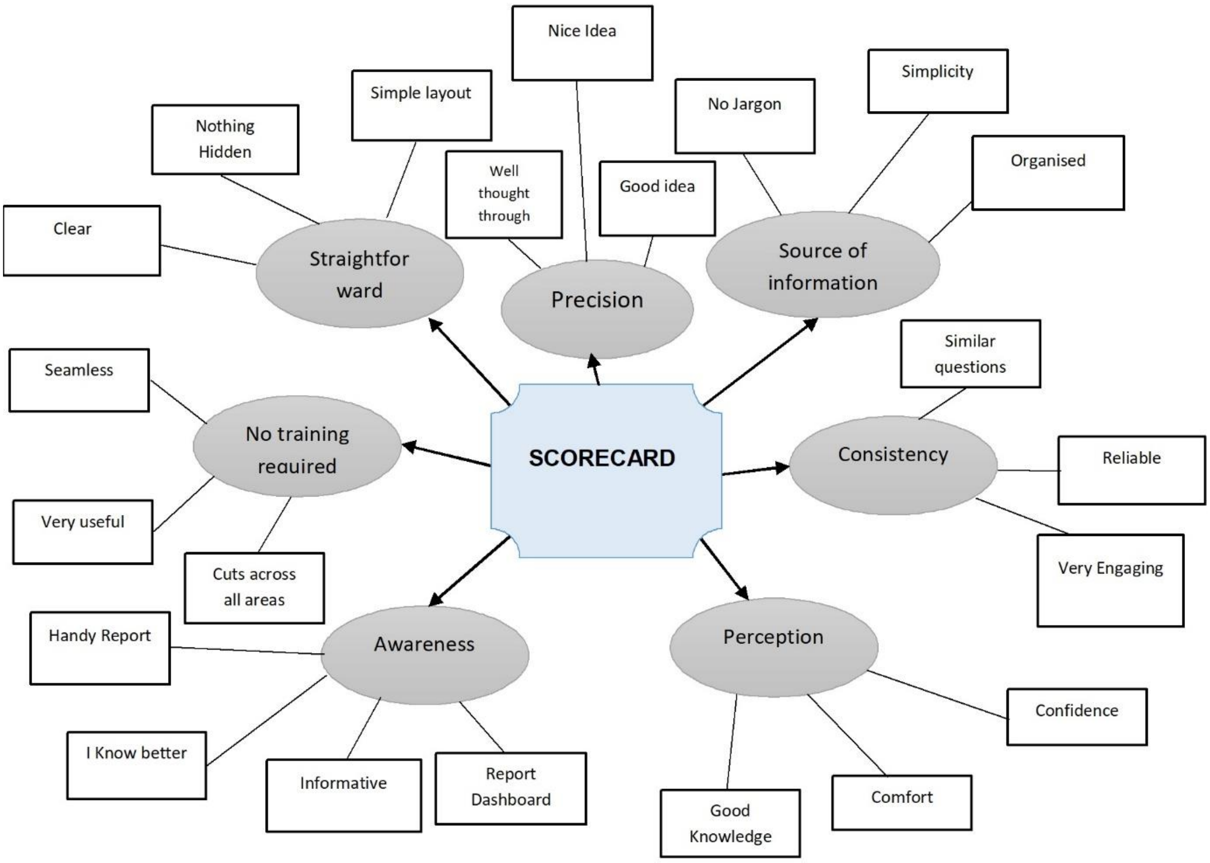


Figure 55: Final thematic map with 7 main themes

Table 15 below shows the results of the final themes and sub-themes. The responses from the participants are also shown below.

Table 15: Summary of results – Final and sub-themes

High-level themes	Sub-themes	Response from Participants
Straightforward	1. Clear 2. Nothing hidden 3. Simple Layout	**Role-Data Consumer:** The scorecard is clear and easy to use. **Role-Data Consumer:** The questions are straightforward. **Role-Data Manager:** Everything is clear, nothing hidden. **Role-Data Producer:** Very simple and easy layout.
Precision	1. Good Knowledge 2. Comfort 3. Confidence	**Role-Data Custodian:** Very precise gives good knowledge of the scenario. **Role-Data Consumer:** Comfortable using the scorecard. **Role-Data Producer:** I have confidence in using the scorecard, looks impressive. **Role-Data Consumer:** Nice looking website.
Source of Information	1. No jargon 2. simplicity 3. Organised	**Role-Data Consumer:** This is very interest and no technical language. **Role-Data Custodian:** Simple design. **Role-Data Manager:** Looks well organised, all groups go to separate areas. **Role-Data Manager:** Nice, well done. **Role-Data Consumer:** Very simple, straightforward.

		Role-Data Consumer: I like the traffic lights, very engaging.
No training required	1. Seamless 2. Cuts across all areas 3. Very useful	**Role-Data Manager:** No training at all required. **Role-Data Producer:** Cuts across all areas, no need to train anybody. **Role-Data Manager:** Very relevant and useful questions. **Role-Data Custodian:** The reports are very useful and handy as a reference.
Consistency	1. Reliable 2. Similar questions 3. Very engaging	**Role-Data Manager:** The questions are very consistent and similar to what we are used to. **Role-Data Custodian:** It looks like a reliable platform that we can use on a daily basis. **Role-Data Manager:** Very engaging platform, it's worth looking at further. **Role-Data Manager:** The pattern of the scorecard seems clear enough to see it as a reliable tool for future use.

Perception	1. Comfort 2. Good knowledge 3. Confidence	**Role-Data Manager:** If I can get this report daily about the views of all the users about the data warehouse, I'll be very comfortable using the scorecard regularly. **Role-Data Producer:** This tool provides excellent knowledge of the views of all concerned with the data warehouse, makes you feel confident about what's going in the data warehouse" **Role-Data Consumer:** Really, really nice tool.
Awareness	1. Handy Report 2. I know better 3. Informative 4. Report Dashboard	**Role-Data Manager:** I don't need to keep wondering what's going on, I know better now, I can easily view the report to see what others are saying. **Role-Data Consumer:** Quite informative will assist as an input to our operational reporting" **Role-Data Producer:** Provides a good and useable dashboard that shows the views of the data across all users. **Role-Data Custodian:** Handy reporting tool that we can build on for other uses.

6.3.6 Analysis of Results

This study empirically evaluated the modified data quality scorecard, using a qualitative approach. In Chapter 5 of this thesis, the Scorecard was evaluated in a brewery company in the FMCG domain. As a result of the thematic analysis carried out in the previous evaluation, a modification of the DQS to remove the timeliness data dimension became necessary, as the aim of the scorecard is to ensure the data quality dimensions align with the data quality needs of the stakeholder groups. Recall in chapter 3, the literature shows that aligning the needs of the DW stakeholder groups with the data quality dimension will result in improved data quality. To accomplish the evaluation of the modified scorecard, a run through was conducted that enabled the various stakeholders to use the scorecard within their data warehouse development environment, after the run through, semi-structured interviews were employed to get the views and perception of the participants. A total of 15 participants across the four stakeholder groups took part in the run through and semi-structured interview sessions.

Overall, each participant spent approximately thirty minutes to participate in the study. The results of the completed semi-structured interview showed that the stakeholders felt the usage of the DQS gave them an improved confidence in the data, as they were able to view all other stakeholder views about the data loaded into the data warehouse.

The responses of the participants of the second evaluation supported the results of the previous evaluation. As most of the participants acknowledged that the DQS increased their confidence and comfort levels about the data in the data warehouse. Also, the results from the semi-structured interview show that not only is the scorecard simple and straightforward to use, but it increased their perception of the quality of their data.

The findings from the thematic analysis carried out in the second evaluation suggest that the DQS demonstrates the following 7 key themes; (a) No training required; (b) Straightforward; (c) Precision; (d) Source of information; (e) Consistency; and (f) Perception; and (g) Awareness. Section 6.3.7 below, presents the discussion of the analysis of the final themes identified in relation to the data quality scorecard.

6.3.7 Discussion

In this section, the final 6 identified themes are discussed. The themes are as follows: (a) Straightforward; (b) Precision; (c) Source of information; (d) No training required; (e) Consistency; (f) Perception; and (g) Awareness.

(a) Straightforward

The results of this research point to the DQS being straightforward and easy to use the tool. This is corroborated by some of the comments from the participants;

"The scorecard is clear and easy to use".

"The questions are straightforward"

"Very simple, straightforward"

"I like the traffic lights, very engaging"

As part of the design consideration for the DQS as detailed in Chapter 4, ensuring simplicity and ease of use was a major attribute in the design of the DQS. Design simplicity and straightforwardness has been achieved based on the comments from the participants in this research. A number of the stakeholders also commented on how clear they felt the DQS process was;

"Everything is clear, nothing hidden"

(b) Precision

The precise and focused aspect of the design of the DQS was commented on by the participants. According to the responses of the participants;

"Very precise gives good knowledge of the scenario"

The design of the DQS is predicated on aligning the data quality concerns of the DW stakeholders with the most widely used data quality dimensions. The literature review conducted in Chapter 3, highlighted the issue of none alignment of DW stakeholder concerns/needs with the most commonly used data quality dimensions. The main focus of this thesis is to design a DQS that measures the quality of data in a data warehouse that specifically focuses on the concern areas of the DW stakeholders. Based on the comments of the participants in this thesis, it can be inferred that the DQS meets the alignment requirement of the DW stakeholder group. Other comments from the stakeholders confirm how confident the stakeholders are with using the DQS;

"Comfortable using the scorecard"

"I have confidence in using the scorecard, looks impressive"

(c) Source of Information

According to the statements of the participants, the DQS served as a source of information about the data. One of the participants also commented on how well organised and technical language free the DQS is. The report that the DQS generates can be used as reports for further analysis according to some of the stakeholders;

"The reports are very useful and handy as a reference"

"This is very interest and no technical language"

"Looks well organised, all groups go to separate areas"

Reflectively, the value of the source of information could be drawn from its ability to stimulate the interest of the data manager and data consumer stakeholder groups. One of the limitations of existing data quality assessment/measurement tools as identified in the literature presented in Chapter 3, is the lack of information across the data warehouse chain. The “source of information” quality of the DQS is very useful as it could drive the regular usage of the DQS by certain stakeholder groups just for the reporting aspect.

(d) No training required

A number of participants commented on how no training is required to use the DQS. They found the ease of use of the DQS as very important. According to some feedback from participants;

"No training at all required"

"Cuts across all areas, no need to train anybody"

In the systematic literature review presented in Chapter 3, one of the recommendations for designing an effective data quality scorecard is simplicity. The design should be simple and easy for the user to use with minimum to no training requirement (Lawrie and Cobbold, 2004). Thus, it can be inferred that the 'no training required' comments by the stakeholders are somewhat as a result of the simple design of the DQS.

One of the participants mentioned;

"Very simple and easy layout"

(e) Consistency

The findings from this thesis have shown that the participants see the DQS as consistent in design. According to quotes from some participants;

"The questions are very consistent and similar to what we are used to"

"It looks like a reliable platform that we can use on a daily basis"

Recall in Chapter 3, that part of the limitations with current data quality tools is the none alignment with the concerns/needs of the stakeholders. One of the major constructs of the DQS is Consistency. Syntactic data quality concerns as identified in the literature, is a major concern for the stakeholder group, especially the data producer stakeholder group. Syntactic data quality concerns data's structure. The aim for syntactic data quality is consistency where the values of data for specific elements of data in the data warehouse surroundings use a consistent symbolic representation (Ballou et al., 2012; Wang et al., 2014). The participants also commented on the platform of the DQS, some of their comments are below;

"Very engaging platform, it's worth looking at further"

"The pattern of the scorecard seems clear enough, looks like a reliable tool for future use"

Arguably, the findings suggest that the DQS provides a consistent and reliable platform. According to statements from some participants as shown above.

(f) Perception

As defined in the previous evaluation, perception describes the level to which the usage of the DQS improved the stakeholder's viewpoint of the quality of the data being used. Recall in Chapter 3 in the literature review, we established that the perception of the quality of data in the data warehouse is paramount for the usage of the data for decision making by the stakeholders. If the stakeholders perceive the data in the data warehouse of being of inferior quality, they most likely will not use it (Etemadpor and Motta, 2015). The study identified sub-themes such as; (i) Gives comfort (ii) Good knowledge (iii) Confidence

The findings show that the data quality scorecard gave the participants, especially those from the consumer stakeholder group, an added comfort level. The findings also suggest an increased level of confidence in the data as a result of the use of the DQS. This also supports the findings from the previous evaluation. According to the participants' statements;

“If I can get this report daily about the views of all the users about the data warehouse, I’ll be very comfortable using the scorecard regularly”

“This tool provides excellent knowledge of the views of all concerned with the data warehouse, makes you feel confident about what’s going in the data warehouse”

“Really, really nice tool”

The sub-themes provides evidence that the DQS was able to provide the stakeholders with a positive perception of the data, that in essence increased the confidence in their data. This finding supports the evaluation result from the previous

(g) Awareness

All the participants acknowledged that the DQS stirred their awareness about the quality of their data in the data warehouse. According to some feedback from participants;

"I don’t need to keep wondering what’s going on, I know better now, I can easily view the report to see what others are saying”

”Quite informative will assist as an input to our operational reporting”

“Handy reporting tool that we can build on for other uses”

In the literature review presented in Chapter 3, we mentioned that the DW stakeholder group awareness of the quality of their data was not being given adequate attention. More than often, the consequence of unawareness leads to wasted time in getting information from other stakeholders and ultimately could lead to incorrect decisions being made.

One of the participants mentioned;

”Provides a good and useable dashboard that shows the views of the data across all users”

Arguably, the findings suggest that the need to make the DW stakeholders aware of the quality of the data in the data warehouse increases the likelihood of the data being of good quality. As it could be argued that once the various stakeholders are aware of the deficiency of the data in the data warehouse, they will initiate a data correction process.

6.4 Learning, Reflections and Outcome

The design of the final artefact was reached through an iterative cycle of learning, designing, more learning and re-designing. Reflecting on my experience of writing this thesis, I have come to appreciate the DSR methodology as I was able to synthesize the design of the artefact through learning at every point through the iterative cycles. The planning of the thesis was one of the most difficult aspects of this study, as I tried to match my ideas for what I wanted to write about with relevant topics and information that actually existed in literature.

The structure and guidelines provided by DSR enabled a more focused and repeatable process. The design process was central to learning and has deepened my appreciation for this thesis, and will give me a framework to build upon. This study has taken me into different but challenging experiences. While parts of those experiences had been straightforward and exciting, others I have gone through with much difficulties and frustrations, however, I was able to derive specific lessons for myself, yet able to contribute to the body of knowledge in the domain. The contribution to practise and theory of this thesis is mostly in the area of providing a new approach and awareness to role-based data quality.

6.5 Summary

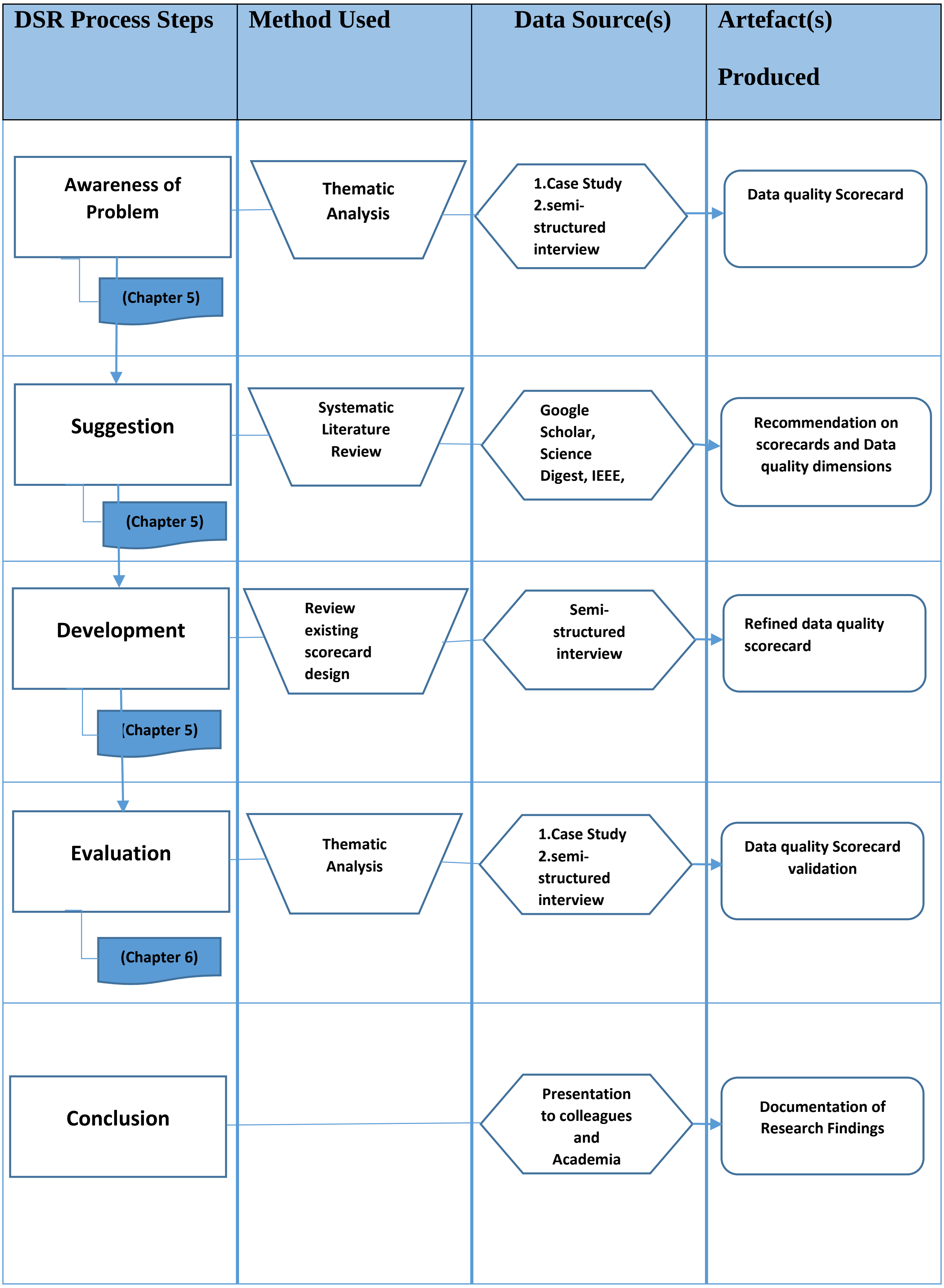
The research presented in this Chapter continues from the work carried out in the previous chapter. The DQS was modified based on the outcome of the second iteration and evaluated again in this chapter. The objective of the third evaluation is to find out whether or not the modified DQS is a useful tool.

The objective was achieved by conducting a semi-structured interview with 15 participants to prompt feedback on the use of the DQS and access whether or not their responses support the attributes of the DQS.

The stakeholders‘ verbal responses from the semi-structured interviews were analysed using thematic analysis and 7 main themes were collected. The main themes identified were; (a) Straightforward; (b) Precision; (c) Source of information; (d) No training required; (e) Consistency; (f) Perception; and (g) Awareness. The themes were discussed, and the findings suggest the DQS is a useful tool.

Chapter 7 will focus on the contributions and the implications of the research work reported in this thesis, as well as a reflection and potential areas for further studies.

Figure 56 below summarises the third DSR iteration with the methods and techniques used.



DSR Process Steps	Method Used	Data Source(s)	Artefact(s) Produced
Awareness of Problem (Chapter 5)	Thematic Analysis	1.Case Study 2.semi-structured interview	Data quality Scorecard
Suggestion (Chapter 5)	Systematic Literature Review	Google Scholar, Science Digest, IEEE,	Recommendation on scorecards and Data quality dimensions
Development (Chapter 5)	Review existing scorecard design	Semi-structured interview	Refined data quality scorecard
Evaluation (Chapter 6)	Thematic Analysis	1.Case Study 2.semi-structured interview	Data quality Scorecard validation
Conclusion		Presentation to colleagues and Academia	Documentation of Research Findings

Figure 56: Summary of results of the third DSR Iteration cycle

Chapter 7: Conclusions and Further Research

7.1 Overview

This Chapter presents the overall conclusion of the thesis. The chapter starts by reiterating the objectives of the thesis set in Chapter 1 and then presents the interconnectivity of the chapters. The theoretical and practical research contributions are then discussed. The research limitations are then highlighted with a summary of how the research methodology mitigated the identified limitations. The researcher then presents some concluding remarks, highlighting some interesting areas of the thesis that may require further research for the future. Finally, the researcher reflects on the PhD journey.

7.2 Research Summary

The overall aim of this thesis was to develop a data quality scorecard to measure the quality of data in a data warehouse and to provide the stakeholders with an integrated data quality awareness platform. The research into data quality is a very popular subject area, with researchers focusing largely on big data. However, the misalignment of the data quality needs of the DW stakeholders has meant a persistence in the problem of data quality. This research attempts to provide a solution to this problem by developing a DW stakeholder group focused DQS that provides the stakeholders with an integrated data quality awareness platform, and also gives an indication of the quality of data in the data warehouse. In order to ensure the objectives have been met, a revisit of the initially formulated objectives is discussed in line with the completed research activities. The following objectives helped in accomplishing the research aim.

Objective 1: To identify and understand data quality issues, data warehouse roles/stakeholder groups and data dimensions in the data warehouse domain, so as to set the right scope.

In Chapter 3, the data quality issues faced by data warehouse users were explored. The literature review carried identified the following susceptible areas as the main weak points for data warehouse quality issues:

The susceptible areas as identified are:

1. During interface with data sources
2. During data integration and profiling
3. During extraction, transformation and loading (ETL)
4. During base data modelling (Schema design)

The literature also shows that the susceptible areas also link directly to the identified data warehouse roles. The data warehouse roles are the identified group of stakeholders responsible for the entire chain of activities within a data warehouse, the identified stakeholder roles as shown in the literature are:

Data producers: Data producer collects the raw data from multiple source systems which are essential for the input of the data warehouse. Data producer is the one who is responsible for the quality of input into the source systems. In other words, data producers are those who create or collect raw data.

Data custodians: These are the group of people responsible for collecting the information from the data producers and then transforming the data into useful information for the use of data consumers by entering it into the data warehouse. Data custodians provide resources for the data consumers by collecting, entering, updating and storing the data in the data warehouse.

The data custodian's primary responsibility is to design, develop and operate the data warehouse.

Data Managers: Data managers are responsible for setting up the right standards and policies related to protecting and managing the day to day usage of the data warehouse. The Data manager group is responsible for managing the day-to-day activities of the data. The primary responsibility of the data manager is to ensure that data custodians are fulfilling their responsibility correctly and also to ensure the security of the entire data warehouse.

Data Consumers: The data consumer is the group or individual who uses the data or information. Data consumers use the set of data for analysis, query, and reporting. In other words, data consumers are the individuals or group of people who use the data in the data warehouses for various purposes. Data consumers are associated with the processes of data utilisation, and also they may involve in additional processes like data integration and aggregation.

Furthermore, the literature review shows that data quality issues fall into one of the following concern areas:

Syntactic data quality concerns the data's structure. The aim for syntactic data quality is consistency where the values of data for specific elements of data in the data warehouse surroundings use a consistent symbolic representation.

Semantic data quality concerns the data meaning. The semantic quality aims are accuracy and comprehensiveness. Comprehensiveness is concerned with the extent to which for every similar state in the actual world system there is a value of data in the data warehouse. Accuracy is concerned with how well the values of data in data warehouse correspond to the real-world state.

Pragmatic data quality concerns the data usage. The pragmatic quality goals are usefulness and usability. Usefulness is the measure to which the data helps stakeholder in fulfilling their activities within an organisation's social context. Usability is the extent to which every stakeholder is capable to use and access the data warehouse data effectively.

Furthermore, the dimensions of data quality as identified from the literature are a feature or aspect of information and a way to categorize data quality and information requirements. The data quality dimensions are used to measure, define and handle the information and data quality. The main dimensions of data quality as identified from the literature are shown in table xx below:

Data Dimension	**Definition**
Completeness	All required data is in inclusive state
Timeliness	Up to date and presented in real time
Integrity	Data is integral, protected from deliberate manipulation
Validity	Data is appropriate and within usage parameters
Accuracy	Data is within the scope of intended use
Consistency	Data collated and collected in a reliable and consistent manner

Table 16: Main Data Quality Dimensions

Remarkably, the findings suggest a misalignment between the data warehouse stakeholder groups, data susceptible areas and the data quality dimensions. This misalignment shows to some extent why tackling data quality issues persist in organisations. The table below shows the alignment required based on literature findings

Data Quality Stakeholders	**Data Quality Concern Areas**	**Data Quality Dimensions (DQD)**
Data Producer	Syntactic	Consistency
Data Custodian		
Data Manager	Semantic	Accuracy, Completeness and Timeliness
Data Consumer	Pragmatic	Integrity and Validity

Table 17: Stakeholders-DQ concern areas-DQD

Objective 2: To investigate and explore the use and limitation of scorecards within the data warehouse domain, and formulate a conceptual framework for the development of a DQS.

In Chapter 3, a systematic literature review was carried out to investigate the state of the art of data quality scorecards/assessment in the literature. Previous studies by (Brockman et al.,2008; Batini et al., 2009; Kauffman et al., 2009; Cai and Zhu, 2013), recommended that in the design of a scorecard, the factors that need to be considered include; use simplicity, time efficiency, electronic/online features, non-technical language and a less intuitive approach. The findings from 18 systematically reviewed papers suggest not much attention have been given to the development of interactive data warehouse stakeholder role focused data quality measurement tools grounded in theoretical and pragmatic paradigms.

Moreover, existing data quality measurement tools are relatively generic in their approach are limited by their technological design in dealing with the interactive requirements of the various data warehouse stakeholder groups. Hence, a conceptual framework was developed for a data quality measurement scorecard specific to the various stakeholder groups, as part of the contribution to the data warehouse research domain.

Objective 3: To develop and validate the conceptual DQS

In chapter 4, the conceptual data quality scorecard was developed based on ideas from the literature. The design of the scorecard used concepts from the systematic literature review conducted in Chapter 3. The architecture of the data quality scorecard model was developed around the identified needs of the data warehouse stakeholder groups. In chapter 3, the data quality susceptible areas were identified, and also the most commonly used data quality dimensions. The study also shows that the data quality needs of each stakeholder group are varied, and might not necessarily be the same prone area. The questions in the scorecard were derived based on the opinions of each stakeholder group as shown in Chapter 3 of this thesis.

The web-centric scorecard was developed using HTML 5, CSS 5, PHP, JavaScript and MySQL database management system. A web domain (www.dataqualityscorecard) was procured to host the scorecard. The mechanics of the scorecard is designed for organizations to be able to manage the usage of the scorecard on their intranet or the security of the web-based scorecard can be ensured by the use of HyperText Transport Protocol Secure (HTTPS) through a Secure Socket Layer (SSL) encryption.

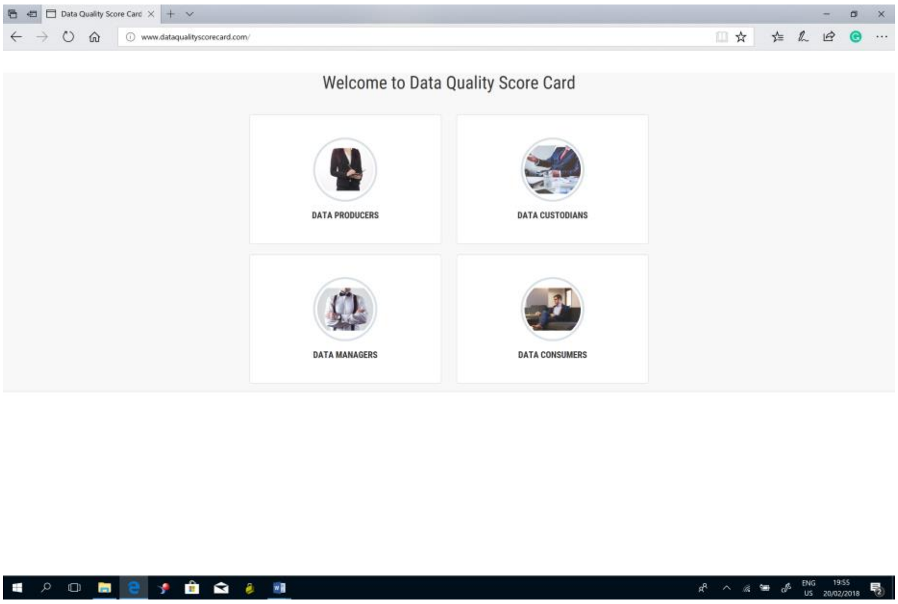
After the development of the scorecard, an initial validation was carried out. Based on the attributes of the scorecard, the researcher's interpretive viewpoint and the need to target stakeholders within the data warehouse domain only, the researcher chose a qualitative method for validation. The researcher sent the web link to selected data warehouse stakeholders in four companies for validation. The four companies were chosen based on the researcher's prior knowledge of the companies data warehouse. The researcher then booked a suitable time with the stakeholders and conducted a semi-structured interview with the stakeholders at there offices.

Semi-structured interviews termed SSI in this research, are a widely accepted way of gathering qualitative data. Semi-structured interviews allow the researcher to ask clarification questions to gain clarity about answers given to question. This is key as generalisations and ambiguity can be immediately resolved. The results were analysed using a data analysis technique called thematic analysis. The key themes identified from the analysis were; Awareness of Quality, Straightforward process, No training Required, Source of Information, Simple Layout, Integrates stakeholder views and overall perception. The themes as identified from the analysis, best describes the major attributes of the data quality scorecard from the perspective of the participants. Furthermore, the study shows, from the responses of the participants that the data quality scorecard is usable and an effective tool in an overall data quality management approach.

Objective 4: To evaluate the DQS using two techniques (1) Case Study –live run through in 2 iterations; (2) semi-structured interviews

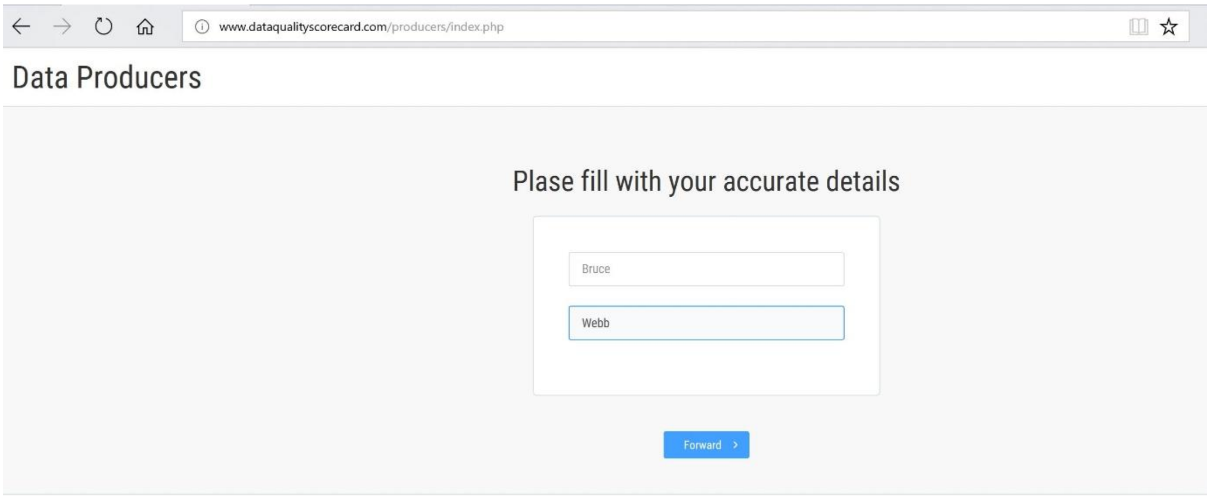
Chapters 5 and 6 reports on the evaluation of the usability and effectiveness of the data quality scorecard relative to the overall aim of this research using two case studies, i.e a practical run-through in two different companies, and then conducting semi-structured interviews to capture the views of the stakeholders. To ensure the participants understand the usage of the web-based scorecard, a briefing and short demo was conducted by the researcher to the various stakeholder groups, this was done to enhance the rigour and quality of the evaluation as teething issues like internet browser settings and firewall/connectivity issues were resolved before the start of the actual evaluation.

The participants for the study were drawn from each of the four stakeholder groups as detailed in Chapters 5 and 6. The selected participants from each of the stakeholder group went about their day to day data warehouse activities within the company, but with an added process of logging into the web-based data quality scorecard to give a quality score to the data, they have just worked on. A domain name was created for the data quality scorecard, (www.dataqualityscorecard.com) and currently hosted on a public network. Figure 60 and 61 below is the login screen based on stakeholder roles



Welcome to Data Quality Score Card

Figure: 57: Data quality scorecard stakeholder group selection screen



Data Producers

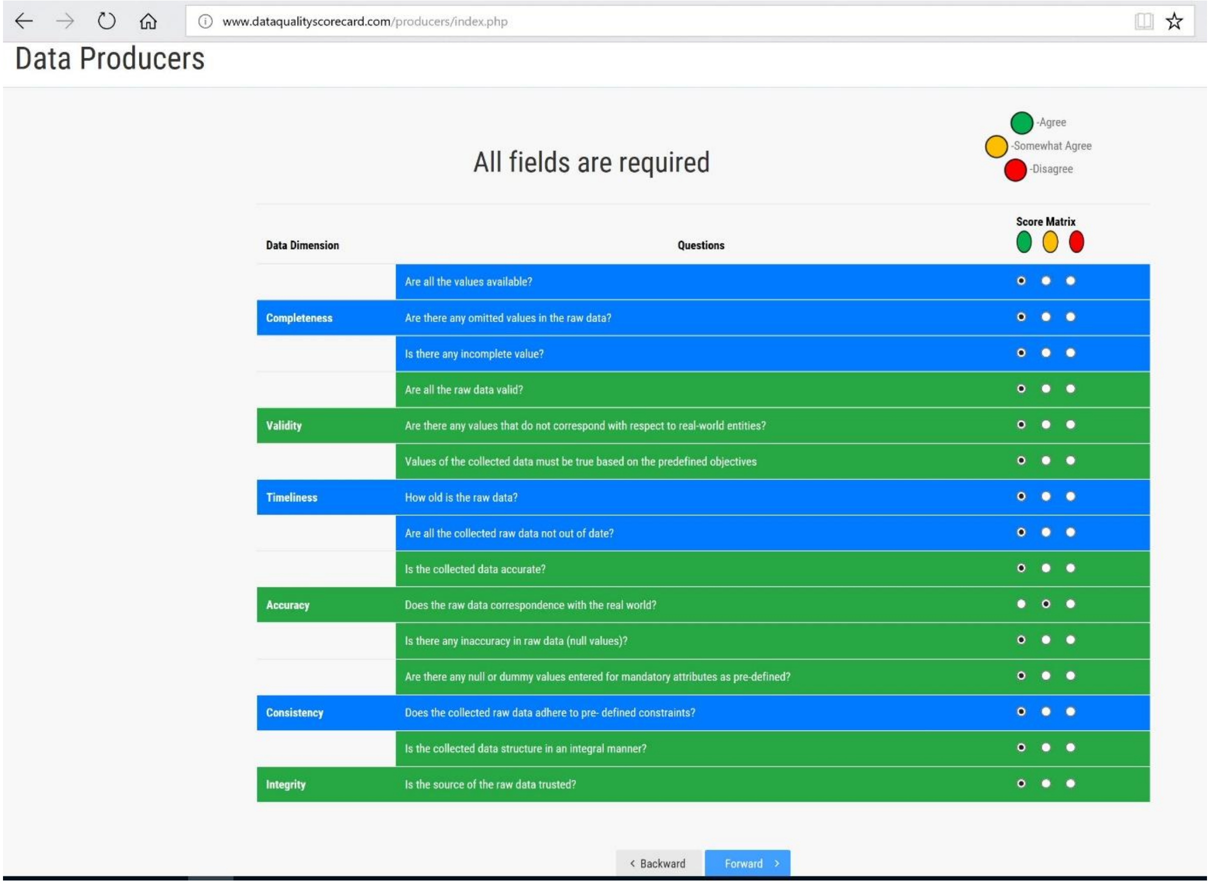
Plase fill with your accurate details

Bruce

Webb

Figure 58: Log on screen based on initial role selection

The data quality scorecard is then presented after the participant enters their name. The selection made by each participant is then stored and available for review/reporting. Figure 59 below shows the web-based scorecard.



Data Producers

All fields are required

🟢 -Agree
🟡 -Somewhat Agree
🔴 -Disagree

Data Dimension	Questions	Score Matrix 🟢	🟡	🔴
Completeness	Are all the values available?	◉	○	○
	Are there any omitted values in the raw data?	◉	○	○
	Is there any incomplete value?	◉	○	○
Validity	Are all the raw data valid?	◉	○	○
	Are there any values that do not correspond with respect to real-world entities?	◉	○	○
	Values of the collected data must be true based on the predefined objectives	◉	○	○
Timeliness	How old is the raw data?	◉	○	○
	Are all the collected raw data not out of date?	◉	○	○
Accuracy	Is the collected data accurate?	◉	○	○
	Does the raw data correspondence with the real world?	○	◉	○
	Is there any inaccuracy in raw data (null values)?	◉	○	○
	Are there any null or dummy values entered for mandatory attributes as pre-defined?	◉	○	○
Consistency	Does the collected raw data adhere to pre- defined constraints?	◉	○	○
	Is the collected data structure in an integral manner?	◉	○	○
Integrity	Is the source of the raw data trusted?	◉	○	○

Figure 59: Web-based scorecard

The semi-structured interview was conducted after the scorecard exercise had been completed by all stakeholders. Open-ended questions were asked to get the views of the stakeholders of the usage and effectiveness of the scorecard. The data collected during the interview was recorded using a dictation machine and later transcribed. The transcribed data was then analysed using thematic analysis to find common themes. In chapter 5, the main themes identified are; Additional Reporting Tool, Transparency, Consistency, Integration, Easy to Use, and Perception. The identified theme was closely aligned with the data quality dimensions. However, based on the responses of some of the participants, the data quality dimension for timeliness was seen as redundant and not adding any value to the scorecard. The scorecard was modified with the dimension for Timeliness removed. In Chapter 6, the second practical run-through of the scorecard was conducted in the oil and gas domain. The modified scorecard was evaluated with the same methods used in Chapter 5. Figure 60 below shows the version selection screen for the modified DQS

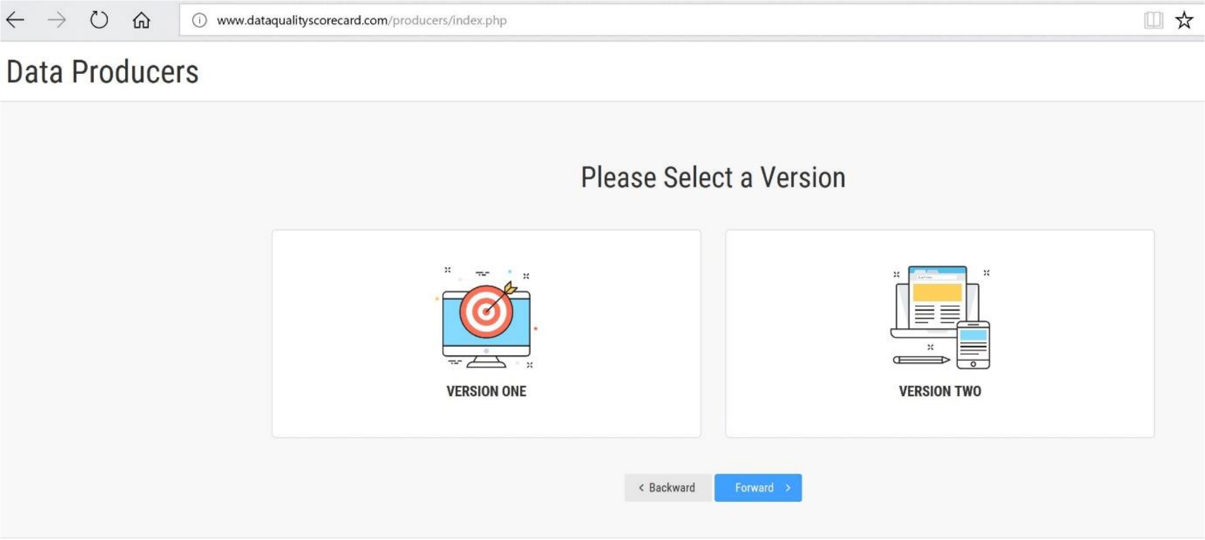


Figure 60: DQS v2 Screen

The responses from the participants suggest that the DQS is not only useful as a data quality measurement tool but can also be used as an additional reporting tool. In section 7.4, the contribution of this research to theory and practice are explained.

7.3 Research Contribution

In this section, the research contributions are discussed in comparison to the challenges of the research domain as underscored by literature.

This research project presents a data quality scorecard that measures the quality of data in a data warehouse and provides the DW stakeholders with an integrated data quality awareness platform. From the literature, one of the challenges of data quality faced by organizations is the volume of data to manage especially with the advent of big data, which results either from the consolidation of divergent systems due to mergers and acquisitions or from an upgrade of systems or from a willingness to simply change current data storage architecture.

The literature review conducted also pointed out that organizational factors, personnel management and technological mechanisms effectively influence the capability to manage the quality of data in a data warehouse. These factors may translate into severe consequences at an organisational level in terms of legal reporting requirement or using the data for decision support purposes. Some of the factors influencing data quality as discussed in Chapter 3 are data cleansing techniques, data storage, system architecture, organisational culture, customer focus, internal control systems, teamwork communication, employee relations, training, performance rewards and evaluation, the culture of the organization and quality management of information supplier. However, the overreaching data quality problem in the domain according to literature is the misalignment of the data quality needs/concerns of the data stakeholders with the data dimension. Literature has shown that this misalignment is what leads to data quality issues as described above. According to the results of the participants from this study, the various stakeholder groups unawareness of the entire chain of activities has also contributed to the perception of low data quality in the data warehouse. This study in Chapter 4 presents the validated DQS that was developed based on literature, which measures the quality of data from selected data quality dimensions in a data warehouse by incorporating a role-based data warehouse stakeholder group approach. One of the remarkable findings from the SUS with a high score of 85.5, demonstrates the enormous value of the DQS to manage the awareness of the quality of the data in the data warehouse by providing an integrated report platform that stores the views of the data from all the stakeholder group. This approach to the best of our knowledge is unique when compared to existing work in data quality management (Madnick et al., 2009; Blake, 2010; Clement et al., 2011; Odera-Kwach et al., 2011; Naiem et al., 2014). Furthermore, the design of the DQS incorporates a stakeholder role-based attribute that is used as part of the data quality scoring parameter that can be customized to meet the needs of the organization. Contributions made by design research methodology must be clear and verifiable in the area of the design artefact (Hevner et al., 2004).

In sections 7.2.1 and 7.2.2 below, the contribution made by this study to practice and theory is presented.

7.3.1 Contribution to Practice

The contributions of this research to industry practices would be discussed by clarifying the alignment between the data quality dimensions and the data warehouse stakeholder roles, as both attributes were adopted to conceptualise the data quality scorecard presented in chapter 4. The various discourse in the literature about data quality management points to a misalignment of the stakeholder roles with their varied data quality needs. However, results from the qualitative study expose how severely unaware the various stakeholders are about the quality requirements of other stakeholder groups they do not belong to. The significance of the awareness attributes of the data quality scorecard for the data warehouse stakeholders helps to further alleviate the primary research problem – misalignment of stakeholder roles with the data quality dimension. The following are the specific contributions to practice

- **Data Quality Awareness:** The contribution to the general awareness of the quality of data in the data warehouse was clearly observed to have increased through the use of the DQS in both organizations. The results of the thematic analysis of the semi-structured interview show that the awareness attributes of the scorecard are seen as very important and useful. The results from the participants in this study about the data quality awareness of the data in their data warehouse, especially the data consumer group was found to be very low, however, the use of the scorecard provided the stakeholder group with improved knowledge of the process, as the perspective of each stakeholder group is transparent and available to all review and analyse. This according to the results provided the stakeholder groups with more confidence in the information extracted from their data warehouse.

- Integration of DQS with reports: The introduction of a structured user-defined process that provides a data quality gate at various points susceptible to data quality degradation was brought about by the use of the DQS. Literature shows that the integration of all areas within a data warehouse and associated tools like a data quality scorecard is paramount to the assurance of data quality in a data warehouse. In Chapter 3, one of the limitations of existing data quality measurement tools identified is – lack of proper integration. The design of the scorecard as an online repository using industry standard web development tools (see Chapter 4 for scorecard mechanics) allows for a seamless integration of the data scorecard results to other reporting tools and database.

- Historical Scorecard Benchmark: Every time data is received by the producer, a scorecard for the data is determined based on the agreed user defined scorecard matrix before the data is entered into the data warehouse, this process is also followed for all other stakeholder groups at various stages of the data management process. The user-defined scorecard results can be saved and used to provide historical data quality maps for data coming from various sources. This can be used as a determinant as to who to use as a producer or as a check to ensure extra quality gates are provided for specific regions because of the historical bad/good quality data score.

7.3.2 Contribution to Theory

The outcome of this thesis provides new intuitions on the stakeholder's perception of data quality in their data warehouse. Also, the research shows the potential factors that create a positive perception and how this could influence the usage of the data quality scorecard.

User Perception of data quality

One of the significant values drawn from this research is the positive perception elements of the stakeholders that characterised the transparency features of the scorecard. Results from the qualitative study suggest that the stakeholders find the scorecard to increase their perception of the quality of data. According to comments from the stakeholders;

"My comfort level is definitely higher with the use of the scorecard"

"I'm more comfortable using the data with the results of the scorecard "

"Yea, I do have more confidence using the data now than I did before the scorecard measurement"

In this research context, perception describes the level to which the usage of the data quality scorecard improved the stakeholder's confidence in the quality of the data being used. Jarke (2012) describes how researchers base their approach on the information systems notion is to offer an application domain representation also referred to as the real-world system perceived by the user. In the literature review, we established that the perception of the quality of data in the data warehouse is a paramount factor for the usage of the data for decision making by the stakeholders. Two key findings from the qualitative study (Easy to use and Transparency) provide an indication of the positive precepts of the scorecard. Easy to use describes the extent to which the stakeholders found the ease of usage of the scorecard on a day to day basis.

The transparency aspect of the scorecard also influenced its positive perception. In the context of this research, transparency describes how the stakeholders find the information availability of all users of the scorecard. From the qualitative study, the awareness of the source and validity of the data were paramount to the perception of the data by most of the stakeholder groups, with the data custodian group showing a lot more concern in this area.

From this discourse, it is, therefore, logical to infer that the positive perception experience of the stakeholder groups with the easy to use and transparency features of the scorecard.

Furthermore, the results of the qualitative study suggest that no matter how well a data quality measurement tool is designed, many of the stakeholders may not use it unless the tool exhibits a level of transparency that increases the perception of the stakeholders.

Unawareness About data quality susceptible areas

One of the remarkable findings from the qualitative study using semi-structured interviews demonstrates the unawareness of some of the stakeholder groups, most especially the consumer group to the areas most susceptible to data quality degradation. Based on the literature on the data quality domain, the identified susceptible areas are: 1) During interface with data sources 2) During data integration and profiling 3) During extraction, transformation and loading (ETL) 4) During base data modelling (Schema design). The research initially presumed that all stakeholders would have a good to a satisfactory level of awareness on the susceptible areas that affect their data quality requirements. However, results from the qualitative study expose how unaware and uninformed some of the stakeholder groups are about the chain of activities that ensure the management of data quality. According to comments from the stakeholders;

”Quite informative will assist as an input to our operational reporting”

“Handy reporting tool that we can build on for other uses”

"I don’t need to keep wondering what’s going on, I know better now, I can easily view the report to see what others are saying”

The DQS provides a contribution to theory by designing a focused stakeholder-based strategy to measure the quality of data in a data warehouse using the most commonly used data quality dimensions. This approach to the best of our knowledge is unique when compared to existing work in data quality management (Madnick et al., 2009; Blake, 2010; Clement et al., 2011; Odera-Kwach et al., 2011; Naiem et al., 2014) In that the data quality scorecard incorporates a stakeholder role-based attribute that is used as part of the data quality scoring parameter that can be customized to meet the needs of the organization. Contributions made by design research methodology must be clear and verifiable in the area of the design artefact (Hevner et al., 2004). This was achieved by the artefact created as a result of the work done in Chapter 4 of this study, and the iterations presented in Chapters 5 & 6.

Finally, based on the availability of stakeholder group level reports, organisations could holistically measure the data quality levels of the entire stakeholder group and access how varied stakeholder data quality needs are being achieved in comparison to the day-to-day business reports generated. Such a possibility could be accomplished by integrating and correlating the data quality scorecard report with the general reporting tool of the organization.

7.4 Reflection of Research Methodology

The DSR process is a sequence of activities that produces an innovative product i.e., the design artefact. (Hevner et al., 2004). The evaluation of the artefact then provides a continuous feedback of information and a better understanding of the problem to enhance both the quality of the product and the design process. This build-and-evaluate loop is typically iterated a few times before the final design artefact is generated (Markus et al. 2002). Once the problems of the thesis are assessed, the design science research addresses the research through building and evaluation of artefacts that are designed to meet the issues or of the hypothesis.

The four phases of the DSR methodology were included in the research design and implementation process to meet the objectives of the thesis as stated above. These phases were used in three DSR iterations reported in this thesis. The phases are; (1) Problem awareness (2) Suggestion (3) Development; and (4) Evaluation.

The first iteration is used to develop the data quality framework. The conceptual model of the framework and scorecard is designed based on the results of the general and systematic literature review conducted. The limitations of data quality as discovered during the literature review and systematic literature review were used to drive the artefact development process. The development of scenarios to thoroughly test the identified gaps was carried out. Scenarios were developed based on the six most relevant data quality dimensions according to the literature review conducted. An initial validation was then carried using a system usability scale (SUS) with the identified stakeholder groups within the data warehouse domain.

The second iteration is conducted as a case study in a brewery company. The artefact produced from the first iteration: the validated DQS was deployed within the brewery organisation. The representatives of the identified four stakeholder groups (i) Data Producers (ii) Data Managers (iii) Data custodians (iv) Data Consumers, took part in the exercise. A semi-structured interview was conducted to collect the data. A key advantage of using semi-structured interviews is that it allows the researcher to ask additional questions to gain further clarity on the data obtained during the interview. The data from the interview was transcribed and was then analysed using an analytic technique called thematic analysis.

In the third iteration of this research, a case study was conducted in an oil and gas company. The results of the analysis performed in the second iteration suggest that not all the data dimensions are required for effective and efficient data quality management, hence the DQS was modified and a version two created.

7.5 Research Limitations

This research focuses on the following data quality dimensions due to their importance and being ‘commonly used’, according to the literature review conducted.

Data Dimension	Definition
Completeness	All required data is in inclusive state
Timeliness	Up to date and presented in real time
Integrity	Data is integral, protected from deliberate manipulation
Validity	Data is appropriate and within usage parameters
Accuracy	Data is within the scope of intended use
Consistency	Data collated and collected in a reliable and consistent manner

Table 18: Selected Data Quality Dimensions

However, it was observed during testing the evaluation of the DQS that other dimensions are also important to the stakeholders. A useful addition would be to add more dimensions, and also include the ability to have a user-defined selectable data dimension. The first step to the improvement of data quality is to understand the key dimensions of data quality within various organizational domains, as this study has shown that the data quality needs of the stakeholders in the FCMG domain divers from the Oil and Gas domain. The security dimension explains the authorization policy every user has for data querying. This dimension was requested for by one of the data managers in the oil and gas company and was commented as a very important dimension. To be interpretable and processable in an efficient and effective manner, data has to fulfil a group of quality criteria’s. These criteria are defined to meet specific stakeholder groups. Affluent attempts have been made to refer to data quality and to recognize its dimensions. The lists of additional data quality consist of dimensions such as reliability, usefulness, importance, precision, and conciseness.

7.6 Future Work

Future research in data quality should focus on the evolution of DQ/IS over time. (Blake 2010). In order to provide a significant contribution to theory, more testing is required in more domains. A rigorous testing regime would highlight other areas for improvement of the DQS. Automation of some of the process steps would be an area to focus future development of the DQS. It was observed during the testing of the DQS that some checklist steps were monotonous. The data entry process steps attributed to the data producers, in particular, can be researched upon further to provide for automation. Future development would look into improving the deployment approach of the DQS.

7.7 Personal Reflection

As I reflect over the years invested in doing this research work, I could simply say that every aspect of the study has taken me into different but challenging experiences. I experienced a huge shift in my sense of identity and in my view of what a PhD is all about. I knew nothing about scholarly writing before I started this program. Coming from the business environment, I wrote good business documents, plans, and reports. Scholarly writing is completely different – it took time to acquire just the basic skills. I still have a lot to learn about writing arguments and critiquing others' work. The writing task: In retrospect, the shaping of my thesis was slow (and sometimes frustrating) process, but I was always encouraged by the people around me, particularly my supervisor. Indeed, I found that it was important to have time to think and work alone, but also to have time to share thoughts and develop ideas with other people. It is not uncommon to feel that the PhD thesis is an insurmountable task that will never end. From my experience, the key to completing such a big project is perseverance, hard work, good time management, constant chats with your supervisor and prayer.

This is by far the most difficult thing I have done, the rigour, the late nights, the data….but just like Alan would say ‘if it is easy, everyone would have a PhD’.

References

Abmann U, Zschaler S and Wagner G 2006, Ontologies for Software Engineering and Software

Adamson C and Venerable M 1998, Data Warehouse Design Solutions, John Wiley & Sons, Inc., New York.

Agarwal R and Lucas H C 2005, The information systems identity crisis: focusing on high Visibility and high-impact research, MIS Quarterly, 293, 381-398.

Agrawal R, Gupta A and Sarawagi S 1997, Modeling multidimensional databases, Proceedings of the 13th International Conference on Data Engineering, IEEE, pp. 232–243.

Alan, F. K., Sanil, A. P., Sacks, J., et al. 2001 Workshop Report: Affiliates Workshop on Data Quality, North Carolina: NISS.

Alexander, J. E., & Tate, M. A. Web wisdom: How to evaluate and create information on the web, Mahwah, NJ: Erlbaum

Arens Y, Chee C Y, Hsu C and Knoblock C A 1993, Retrieving and integrating data from Multiple

Baader F, Calvanese D, McGuinness DL, Nardi D and Patel-Schneider PF 2003, The Description Logic Handbook: Theory, Implementation, and Applications, Cambridge University Press, Cambridge.

Ballard C 1998, Data Modeling Techniques for Data Warehousing, SG24-2238-00, IBM Red Book, ISBN number 0738402451.

Ballou D P and Pazer H 2003, Modeling Completeness versus Consistency Tradeoffs in Information Decision Contexts, IEEE Transactions and Data Engineering,

Ballou D P, Wang R Y, Pazer H and Tayi GK 1998, Modeling Information Manufacturing

Systems to Determine Information Product Quality Management Science, vol. 44, no. 4.

Baskerville R 2008, "What Design Science is Not." European Journal of Information Systems 175: 441-443.

Baskerville R, Pries-Heje J and Venable J 2009, Soft Design Science Methodology, in proceedings of Design Science Research in Information Systems and Technology DESRIST, Philadelphia.

Batini C and Scannapieca M 2006, Data Quality: Concepts, Methodologies and Techniques. Springer, Germany.

Batini, Ceri and Navathe 1992, Conceptual Database Design: An Entity-Relationship The approach, The Benjamin/Cummings Publishing Company, Inc., USA.

Bechhofer S, Horrocks I, Goble C and Stevens R 2001, Oil-ED: A Reasonable Ontology Editor

Bellatreche L, Xuan G, Pierra D N and Dehainsala H 2006, Contribution of ontology-based data modelling to automatic integration of electronic catalogues within engineering databases, Computers in Industry Journal.

Berger P and Luckman T 1966, The Social Construction of Reality: a treatise in the sociology of knowledge, Garden City, Doubleday, New York.

Besterfield D H, Besterfield-Michna C, Besterfield G and Besterfield-Sacre M 1995, Total Quality Management, Prentice Hall, New York.

Blaschka M, Sapia C, Hofling G and Dinter B 1998, Finding your way through multidimensional data models. Proceedings of 9th International, Workshop On Database and Expert Systems Applications DEXA, Vienna, Austria.

Blyth B 2006, Independent, Transparent, Externally Audited: The ISO Approach to Survey

Böhnlein M and Vom Ende A U 1999, Deriving initial data warehouse structures from the conceptual data models of the underlying operational information systems, Proceedings of 2nd International Workshop on Data Warehousing and OLAP, ACM, pp. 15–21.

Bonifati A, Cattaneo F, Ceri S, Fuggetta A and Paraboschi S 2001, Designing data marts for data warehouses, ACM Trans. Software Engineering Method, 10 4, p 452–483.

Bouzeghoub M, Fabret F, Llirbat F, Matulovic M and Simon E 1997, Designing data warehouse refreshment system, DWQ Technical report.

Bouzeghoub M, Fabret F, Matulovic M and Simon E 1998, Data Warehouse Refreshment: A Design Perspective from Quality Requirements, Technical Report D8.5, DWQ Consortium, available at http://www.dbnet.ece.ntua.gr/~dwq/.

Brancheau J, Janz B and Wetherbe J 1996, Key Issues in Information Systems Management: 1994-95 SIM Delphi Results,. MIS Quarterly 20:2, pp. 225-242.

Brickley D and Guha R V 2004, RDF Vocabulary Description Language 1.0: RDF Schema, W3C Recommendation, accessed on 21st July 2012 from

Bunge M 1984, Philosophical Inputs and Outputs of Technology, History and Philosophy of Technology, G. Bugliarello and D. Donner. Urbana, IL, University of Illinois Press**:** 263-281.

Calero C, Piattini M and Genero M 2001, Metrics for controlling database complexity, Chapter III in Developing quality complex database systems: practices, techniques and technologies. Becker ed, Idea Group Publishing, UK.

Calvanese D, De Giacomo G, Lenzerini M, Nardi D and Rosati R 1998, Information integration, Conceptual modelling and reasoning support, Proceedings of the 6th International Conference on Cooperative Information Systems CoopIS-98, pp. 280-291.

Campbell R 1997, Making Information Organization Universal, Database Web Advisor, Oct 1997, Vol. 15 n10, pp 62.

Cao, J. J., Diao, X. C., Wang, T., et al. 2010 Research on Some Basic Problems in Data Quality Control. Microcomputer Information 09, pp 12–14.

Cappiello, C., Francalanci, C., & Pernici, B. 2004 Data quality assessment from user‘s perspective. Procedures of the 2004 International Workshop on Information Quality in Information Systems, New York: ACM, pp 78–73.

Carlsson S A 2003, Advancing information systems evaluation research: a critical realist

The approach, Electronic Journal of Information Systems Evaluation, 62, 11-20.

Carroll J and Kellogg W 1989, Artifact as Theory Nexus: Hermeneutics Meets Theory-Based Design, In Proceedings of CHI '89, ACM Press.

Carroll J J and Roo J D 2004, OWL Web Ontology Language Test Cases, W3C The recommendation, accessed on 21st July 2012 from http://www.w3.org/TR/2004/REC-owl-test-20040210/.

Chaudhuri S and Dayal U 1997, An Overview of Data Warehousing and OLAP Technology, SIGMOD Record, Vol. 26, No. 1.

Chaudhuri S and Dayal U 1997, An overview of Data Warehousing and OLAP Technology, SIGMOD Record 261, New Delhi.

Chen W. and Hirschheim R 2004, A paradigmatic and methodological examination of information systems research, Information Systems Journal, 143, 197-235.

Chong Q, Marwadi A, Supekar K and Lee Y 2003, Ontology-based metadata management in medical domains. J of Res and Prac in Inform Tech, 352: p 139-54.

Cipriano F 1995, The Impact of Information Systems on Quality Performance: An Empirical Study, International Journal of Operations and Production Management, Vol. 156, pp. 69-83.

Cole R, Purao S, Rossi M and Sein M K 2005, Being proactive: where action research meets design research, Proceedings of the Twenty-Sixth International Conference on Information Systems, 325-336.

Consistent terminology for software measurement, Inf. & Software Technol. 48 8, p 631–644.

Construction. International Journal of Human−Computer Studies, 46,6:707−727.

Corcho O and Gómez-Pérez A 2000, “A Roadmap to Ontology Specification Languages,” 12th

Crosby, P. B. 1988 Quality is Free: The Art of Making Quality Certain, New York: McGraw-Hill.

Das A, Wu W and McGuinness D L 2001, Industrial Strength Ontology Management, Stanford Knowledge Systems Laboratory Technical Report KSL-01-09 2001. In the Proceedings of the International Semantic Web Working Symposium, Stanford, CA.

Data Application Environment Construction and Service of Chinese Academy of Sciences 2009

Data Quality Evaluation Method and Index System. Retrieved October 30, 2013, from the World

Wide Web: http://www.csdb.cn/upload/101205/1012052021536150.pdf

De Vries M 1993, Design Methodology and relationships with science: introduction, In M J De Vries, N Cross and D P Grant Eds., Design Methodology and Relationships with Science.

Decker, Erdmann S, Fensel M, Studer D and Ontobroker R 1999, Ontology-Based Access to Distributed and Semi-Structured Information. In R. Meersman et al. eds.: Semantic Issues in Multimedia Systems, Proceedings of DS−8. Kluwer Academic Publisher, Boston, p 351−369.

Delone W H and McLean E R 1992, Information Systems Success: The Quest for the Dependent Variable, Information Systems Research, 3:1, pp.60-95.

DeLone, W H and McLean E R 1992, Information Systems Success: The Quest for the e Dependent Variable, Information Systems Research 3:1, pp. 60-95.

Demchenko, Y., Grosso, P., de Laat, C., et al. 2013 Addressing Big Data Issues in Scientific Data Infrastructure. Procedures of the 2013 International Conference on Collaboration Technologies and Systems, California: ACM, pp 48–55.

Ding L, Finin T, Joshi A, Pan R, Cost RS, Peng Y, Reddivari P, Doshi V and Sachs J 2004, Swoogle: a search and metadata engine for the semantic web, In Proceedings of the thirteenth ACM international conference on Information and knowledge management, p 652-659.

Ding Z, Peng Y and Pan R 2006, Bayes OWL: Uncertainty modelling in semantic web Ontologies. Soft Computing in Ontologies and Semantic Web, pages 3{29, Springer Verlag, Germany.

Dori D, Feldman R and Sturm A 2005, Transforming an operational system model to a data warehouse model: a survey of techniques, IEEE International Conference on Software-Science, Technology and Engineering SwSTE 2005, IEEE Computer Society, pp. 47–56.

Du Plooy G M 2001, Communication Research: Techniques, Methods and Applications, Juta and Company Limited, Lansdowne.

Dung P M 1996, “Integrating Data from Possibly Inconsistent Databases,” Proc . Cooperative Information System, pp. 58-65.

Dyke TP V, Kappelman L A and Prybutok V R 1997, "Measuring Information Systems Service Quality: Concerns about the Use of the SERVQUAL Questionnaire", Management Information Systems Quarterly 212, pp. 195-208.

Elmasri R and Navathe SB 2000, Fundamentals of Database Systems, Addison-Wesley New Jersey.

English L 2001, 10 years of Information Quality Advances: What Next? Information Management Magazine.

English, L 1999, Improving Data Warehouse and Business Information Quality, New York: John Wiley & Sons, Inc.

Eppler M J 2001 "The Concept of Information Quality: An Interdisciplinary Evaluation of Recent Information Quality Frameworks", Studies in Communication Sciences 1, pp. 167-182.

Eppler M J and Wittig D 2000, Conceptualizing Information Quality: A Review of Information Quality Frameworks from the Last Ten Years, in Klein, B. D., Rossin, D. F. ed.: Proceedings of the Conference on Information Quality, Cambridge, MA, p. 83-96.

Evbuonwan N F O, Sivaloganathan S and Jebb A 1996, A survey of design philosophies, models, methods and systems. Proceedings Institute of Mechanical Engineers, 210, p 301-320.

Farquhar A, Fikes R and Rice J 1997, The Ontolingua Server: A tool for collaborative ontology

Feng, Z. Y., Guo, X. H., Zeng, D. J., et al. 2013 On the research frontiers of business management in the context of Big Data. Journal of Management Sciences in China 1601, pp 1–9.

Fensel D 2000, The semantic web and its languages, IEEE Computer Society 15, 6 November /December, p 67−73.

Fettke P, Houy C and Loos P 2010, "On the Relevance of Design Knowledge for Design-

Oriented Business and Information Systems Engineering", Business and Information Systems Engineering 26: 347-358.

Fonseca F, Egenhofer M, Agouris, P and Camara G 2002, Using Ontologies for Integrated Geographic Information Systems, Transactions in GIS, −6:3 in print.

For the Semantic Web. Lecture Notes in Computer Science, SAGE, London.

Forza C 1995, Quality information systems and quality management: A reference model and associated measures for empirical research, Industrial Management and Data Systems, Vol. 952, pp. 6-14.

Friedman, K 2003, "Theory construction in design research: criteria: approaches, and methods," Design Studies 24:6, pp. 507-522.

Fulcher A J and Hills P 1996, Towards a strategic framework for design research. Journal of Engineering Design, 7, 1, 183-193.

Galliers RD 1993, Research Issues in Information Systems, Journal of Information Technology, Vol. 8 n2, pp. 92-98.

García F, Bertoa M F, Calero C, Vallecillo A, Ruiz F, Piattini M and Genero M 2006, Towards a

Gardner SP 2005, Ontologies and semantic data integration. Vol. 10. Elsevier, UK, p 1001-1007.

Gartner Group 1997, Four Myths about Data Warehouses, Gartner Group Research Note.

General Administration of Quality Supervision 2008 Inspection and Quarantine of the People's

The Republic of China. Quality management systems-Fundamentals and vocabulary GB/T19000—2008/ISO9000:2005, Beijing.

Giorgini P, Rizzi S and Garzetti M 2005, Goal-oriented requirement analysis for data warehouse design, in DOLAP, pp. 47–56.

Goldkuhl G 2004, "Design Theories in Information Systems - A Need for Multi-Grounding" Journal of Information Technology Theory and Application 62: 59-72.

Golfarelli M 2009, Data Warehouse Design, Tata McGraw Hill, New York.

Golfarelli M and Rizzi S 2009, Data Warehouse Design, Modern Principles and Methodologies, McGraw-Hill, New York.

Golfarelli M, Maio D and Rizzi S 1998, “Conceptual design of data warehouses from E/R schemes,” Proc. 32th HICSS.

Golfarelli M, Maio D and Rizzi S 1998, The dimensional fact model: a conceptual model for data warehouses, International Journal Coop. Information System, 7 2–3, p 215–247.

Gregg D, Kulkarni U and Vinze A 2001, "Understanding the Philosophical Underpinnings of Software Engineering Research in Information Systems," Information Systems Frontiers, 32: 169-183.

Gregor S 2006, "The Nature of Theory in Information Systems", MISQ 303: 611-642.

Gregor S and Jones D 2007, "The Anatomy of a Design Theory", Journal of the Association for Information Systems JAIS 85: Article 19.

Gruber T R 1993, A Translation Approach to Portable Ontology Specifications, Knowledge Acquisition, 5:199−220.

Gruber T R 1995, Toward principles for the design of ontologies used for knowledge sharing. Presented at the Padua workshop on Formal Ontology, March 1993, later published in International Journal of Human-Computer Studies, Vol. 43, Issues 4-5, pp. 907-928.

Guarino N 1998, Formal Ontology in Information Systems, In N. Guarino ed. Formal Ontology in Information Systems. Proceedings of FOIS’98, Trento, Italy, 6−8 June 1998. IOS Press, Amsterdam, p 3−15.

Guarino, N 1998, “Formal Ontology and Information Systems,” in Formal Ontology in Information Systems, N. Guarino, Ed. Amsterdam, Netherlands: IOS Press.

Guba E and Lincoln Y 1994, Competing Paradigms in Qualitative Research, The Handbook of Qualitative Research, N. Denzin and Y. Lincoln. Thousand Oaks, CA, Sage: 105-117.

Gupta H 1997, Selection of Views to Materialize in a Data Warehouse, Proceedings of International Conference on Database Theory ICDT, Delphi, Greece.

Gupta V R 1997, Introduction to Data Warehousing, System Services Corporation, Chicago, Illinois.

Gyssens M and Lakshmanan LV S 1997, A Foundation for Multi-Dimensional Databases, Proceedings of the 23rd International Conference on Very Large Databases VLDB, Athens.

Hainaut J L 1991, Entity-Generating schema transformations for Entity-Relationship models, SAGE, London p 643 – 670.

Hammer J, Garcia-Molina H, Widom J, Labio W and Zhuge Y 1995, The Stanford Data Warehousing Project. Data Engineering, Special Issue Materialized Views on Data Warehousing, 182, pp. 41-48.

Hammergren T 1996, Data Warehousing Building the Corporate Knowledge Base. International Thomson Computer Press, Milford, USA.

Heflin J 2004, OWL Web Ontology Language Use Cases and Requirements, accessed on

Heflin J and Hendler J 2000, Dynamic Ontologies on the web. Proceedings of the Seventeenth National Conference on Artificial Intelligence AAAI/MIT Press, p 443–449.

Helfert M 2001, Managing and Measuring Data Quality in Data Warehousing. In: Proceedings of the World Multiconference on Systemics, Cybernetics and Informatics, Orlando, FL, pp. 55-65.

Hendry, R. 2004. “Are Realism and Instrumentalism Methodologically Different?” Online working paper, Department of Philosophy, University of Durham, UK, last accessed 25th August 2012 available at http://hypatia.ss.uci.edu/lps/psa2k/realism-and-instrumentalism.pdf.

Hevner A, March S, Park J and Ram S 2004, "Design Science in Information Systems Research", MIS Quarterly 281: p 75-105.

Hevner A.R., March S.T., Park J., Ram S. 2004 Design science in informatresearch. MIS Quarterly Vol. 28 No. 1, pp. 75-105/March 2004.

Holstrom J, Ketokivi M and Hameri A 2009, "Bridging Practice and Theory: A design Science Approach" Decision Sciences 401: 65-87.

Hopfgartner F and Jose J 2010, Semantic user modelling for personal news video retrieval. Advances in Multimedia Modeling, p 336-346. Springer Verlag, Germany.

Howard P 2004, Data Quality Products: An Evaluation and Comparison, Bloor Research,

Miller H 1996, The Multiple Dimensions of Information Quality, Information Systems Management, Spring, Vol. 132, pp. 79-82. Golfarelli, M. and Rizzi, S., 2018. From Star Schemas to Big Data: 20+ Years of Data Warehouse Research. In A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years (pp. 93-107). Springer International Publishing.

Hsu C, Babin G, Bouziane M, Cheung W, Rattner L and Yee L 1992, "Metadatabase Modeling for Enterprise Information Integration," Journal of Systems Integration, vol. 2, no. 1, pp. 5-37.

Huang J, Lee YW and Wang R Y 1999, Quality Information and Knowledge, Prentice Hall, Upper Saddle River, New Jersey.

Huang KT, Lee YW and Wang R Y 1999, Quality Information and Knowledge, Prentice Hall PTR, Upper Saddle River, New Jersey.

Hull R and Zhou G 1996, A Framework for Supporting Data Integration Using the Materialized and Virtual Approaches, In Proc. of the ACM SIGMOD Intl. Conf. on Management of Data, pages 481–492.

Hüsemann B, Lechtenbörger J and Vossen G 2000, Conceptual data warehouse modelling Proc. of 2nd Int. Workshop on Design and Management of Data Warehouses, CEUR-WS.org, p. 6.

Iivari J 2003: The IS core — VII towards information systems as a science of meta-

artefacts, Communications of the Association for Information Systems, 12, Article 37, 568-581.

Iivari, J. 2007. A Paradigmatic Analysis of Information Systems As a Design Science.

Scandinavian Journal of Information Systems, 2007, 192, p. 39-64.

Information sources, Journal of Intelligent and Cooperative Information Systems, 2 2,
p 127-158.

Inmon W and Hackathorn R 1994, Using The Data Warehouse, John Wiley & Sons, New

York.

International Conf. Knowledge Eng. and Knowledge Management, Lecture Notes in Artificial Intelligence, Springer-Verlag, Berlin, pp. 80–96.

Jarke M and Pohl K 1992, Information systems quality and quality information systems, Proceedings IFIP 8.2 Working Conference, Minneapolis.

Jarke M and Vassiliou Y 1997, Data Warehouse Quality: A Review of the DWQ Project. In Proc. of the 2nd Intl. Conf on Information Quality Cambridge, Mass., pages 98–112.

Jarke M and Vassiliou Y 1997, Foundations of Data Warehouse Quality – A Review of the DWQ-Project, in Strong, D. M., Kahn, B. K. ed.: Proceedings of the 2nd International Conference on Information Quality, Cambridge, MA, pp. 299-313

Jarvinen P 2007, "Action Research is Similar to Design Science", Quantity and Quality 41: 37-54.

Jensen M R, Holmgren T and Pedersen 2004, Discovering multidimensional structure in relational data, 6th Int. Conf. on Data Warehousing and Knowledge Discovery, volume 3181 of LNCS, Springer, 2004, pp. 138–148.

Jeusfeld MA, Quix C and Jarke M 1998, Design and Analysis of Quality Information for Data Warehouses, In Proc. of the 17th International Conference on the Entity Relationship

Jones J C 1992, Design Methods, John Wiley & Sons, Canada**.**

Juran J M 1998, How to think about Quality, In Juran J M, Godfrey A B.ed.: Juran’s quality handbook, 5th ed., McGraw-Hill, New York, pp 2.1-2.18

Kahn B K and Strong DM 1998, 'Product and Service Performance Model for information Quality: an Update', in Conference on Information Quality, Cambridge, MA, USA, pp. 102-115.

Kashyap V and Sheth AP 1996, Semantic and Schematic Similarities between Database Objects: A Context-Based Approach, VLDB Journal 54: p 276-304.

Katal, A., Wazid, M., & Goudar, R. 2013 Big Data: Issues, Challenges, Tools and Good Practices. Procedures of the 2013 Sixth International Conference on Contemporary Computing, Noida: IEEE, pp 404–409.

Katerattanakul, P., & Siau, K. 1999 Measuring information quality of websites: Development of an instrument. Procedures of the 20th International Conference on Information Systems, North Carolina: ACM, pp 279–285.

Kimball R 1996, Slowly Changing Dimensions, The Data Warehouse Architect, DBMS Magazine, available at http://www.dbmsmag.com.

Kimball R 1996, The Data Warehouse Toolkit, John Wiley & Sons, Inc, New York.

Kimball R 1996, The Data Warehousing Toolkit, John Wiley, New York.

Klein, A., & Lehner, W. (2009). Representing data quality in sensor data streaming Journal of Data and Information Quality (JDIQ), 1(2), 10.

Knight, S., & Burn, J. 2005 Developing a Framework for Assessing Information Quality on the World Wide Web. Information Science Journal 18, pp 159–171.

Kortnik M A R and Moody D L 1999, From Entities to Stars, Snowflakes, Clusters, Constellations and Galaxies: A Methodology for Data Warehouse Design, 18th International Conference on Conceptual Modeling, Industrial Track Proceedings.

Krogstie J, Lindland O I and Sindre G 1995a, 'Defining Quality Aspects for Conceptual Models', in an IFIP8.1 working conference on Information Systems Concepts ISCO3: Towards a Consolidation of Views, eds., Falkenberg E D, Hesse W & Olive A, Marburg, Germany, pp. 216-231.

Kuechler W and Vaishnavi V 2008, "On Theory Development in Design Science Research: Anatomy of a Research Project." European Journal of Information Systems 175: 1-23.

Kuechler W and Vaishnavi V 2011, "A Framework for Theory Development in Design Science Research: Multiple Perspectives" under the third review for Journal of the Association for Information Systems JAIS.

Kuhn T 1996, The Structure of Scientific Revolutions, University of Chicago Press, Chicago.

Laberge R 2011, The Data Warehouse Mentor, Tata McGraw Hill, New Delhi, page 18.

Labio W J and Garcia-Molina H 1996, Efficient Snapshot Differential Algorithms for Data Warehousing. In Proc. of the 22nd Intl. Conf. on Very Large Data Bases, pages 63–74.

Labio W, Quass D and Adelberg B 1997, Physical Database Design for Data Warehouses, Thirteen International Conference on Data Engineering, IEEE Computer Society, Birmingham, 277-288.

Lakatos I 1978, The Methodology of Scientific Research Programmes John Worral and Gregory Currie, Eds., Cambridge University Press, Cambridge.

Land, F 1992, The Information Systems Domain. Information Systems Research — Issues, Methods and Practical Guidelines. R. Galliers, Ed. Blackwell Scientific Publications, Oxford, England.

Lassila O and Swick R R 1999, Resource Description Framework RDF Model and Syntax Specification, World Wide Web Consortium Recommendation, accessed on 21st July 2012 from http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/.

Lechtenbörger J and Vossen G 2003, Multidimensional normal forms for data warehouse design, Information System, 28 5, p 415–434.

Lechterborger J L 2001, Data Warehouse Schema Design, IOS Press, Germany, p 23.

Lee A and Hubona G 2009, "A Scientific Basis for Rigor in Information Systems Research" MIS Quarterly 332: 237-262.

Lee Y W 2004, Crafting rules: context reflective data quality problem-solving. Journal of Management Information Systems, vol 20, no. 3, pp. 92-119.

Lee YW, Pipino LL, Strong, D M and Wang R Y 2004, Process-embedded data integrity, Journal of Database Management. vol. 15, pp. 87-103.

Lee, Yang W, Strong DM, Kahn B K and Wang R Y 2002, "AIMQ: a Methodology for Information Quality Assessment", Information and Management, 40, pp. 133-146.

Lehner W, Albrecht J and Wedekind H 1998, Normal Forms for Multidimensional Databases.

Lenz H and Shoshani A 1997, Summarizability in OLAP and statistical databases, Proceedings of 9th International Conference on Scientific and Statistical Database Management, IEEE, pp. 132–143.

Lenzerini M, Vassiliadis P, Jarke M and Vassiliou0 Y 203, Fundamentals of Data

Warehouses, Springer, New York, p 1.

Li, G. J., & Chen, X. Q. 2012 Research Status and Scientific Thinking of Big Data. Bulletin of Chinese Academy of Sciences 2706, pp 648–657.

Li, J. Z., & Liu, X. M. 2013 An Important Aspect of Big Data: Data Usability. Journal of Computer Research and Development 506, pp 1147–1162.

Liepins G E and Uppuluri V R R 1990, Data Quaky Control: Theory and Pragmatics, D B Owen, vol. 112., John Wiley & Sons, New York.

Ligouditianos S, Sellis T, Theodoratos D and Vassiliou Y 1999, DWQ project Heuristic Algorithms for Designing a Data Warehouse with SPJ Views, Proc. DaWaK '99, Florence, Italy.

Liu L and Chi L 2002, Evolutionary Data quality, In Proceedings of the 7th International Conference on Information Quality ICIQ 02, Boston.

Lotman, Yuri 1990, Universe of the Mind: A Semiotic Theory of Culture, Bloomington, IN: Indiana University Press.

Luján-Mora S, Trujillo J and Song I Y 2006, A UML profile for multidimensional modelling in data warehouses, Data Knowledge Eng., 59 3, p 725–769.

Madnick S 1997, Database in the Internet age. Database Programming and Design, 10, 1, 28–33.

Maedche, A and Staab S 2001, Ontology learning for the Semantic Web, Vol. 16 2001 p 72-79.

March S and Smith G 1995, "Design and Natural Science Research on Information Technology" Decision Support Systems 15, 251 - 266.

Marjanovic O and Orlowska M 1999, On Modelling and Verification of Temporal Constraints

Martin R 2008, Data Warehouse 100 Success Secrets - 100 Most Asked Questions on Data Warehouse Design, Projects, Business Intelligence, Architecture, Software and Models, Lulu.com, UK, p 139**.**

Mattison R 1996, Data Warehousing: Strategies, Technologies and Techniques McGraw Hill, New York.

Maturana H and Varela F 1987, The Tree of Knowledge: The Biological Roots of Human Understanding, New Science Library, Boston.

Mazón J N, Trujillo J and Lechtenbörger J 2007, Reconciling requirement-driven data warehouses with data sources via multidimensional normal forms, Data Knowledge Engineering, 63 3, p 725–751.

McClanahan D 1996, Making Sense Of Enterprise Data, Data Based Advisor, Data Based Solutions Inc, Nov 1996, Vol. 14 n11, pp. 764, Available [Online Database - Computer Database ASAP].

McFadden F 1996, Data Warehouse for EIS: Some Issues and Impacts, Proc. 29th Annual Hawaii International Conference on System Sciences, Hawaii.

McGilvray, D. 2008 Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information, California: Morgan Kaufmann.

McGilvray, D. 2010 Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information, Beijing: Publishing House of Electronics Industry.

McKay J and Marshall P 2005, A Review of Design Science in Information Systems, Australian Conference on Information Systems, Sydney.

Meng, X. F., & Ci, X. 2013 Big Data Management: Concepts, Techniques and Challenges. Journal of Computer Research and Development 501, pp 146–169.

Milea V, Frasincar F, Kaymak U and Dioia T 2007, An OWL-based approach towards representing time in web information systems, In The 4th International Workshop on Web Information Systems Modeling Workshop WISM 2007, pages 791–802, Tapir, Academic Press.

Mingers J 2001, "Combining IS Research Methods: Towards a Pluralist Methodology Information Systems Research 123: 240-259.

Mingers J and Willcocks L 2004, Social Theory and Philosophy for Information Systems Research, Wiley, Chichester.

Mingers J C 1995, Information and Meaning: foundations for an intersubjective account, Information Systems Journal, 5, pp 285-306.

Moody D and Kortink M 2000, From enterprise models to dimensional models: a methodology for data warehouse and data mart design, Proc. of 2nd Int. Workshop on Design and Management of Data Warehouses, CEUR-WS.org.

Moody D L and Kortink M. A.R 2000, From enterprise models to dimensional models: a methodology for data warehouse and data mart design, in DMDW, p. 5.

Metro A and Smets P 1996, "Uncertainty Management in Information Systems: from Needs to Solutions", Norwell, MA: Kluwer Academic Publishers, pp. 480.

Mouton J 1996, Understanding Social Research, Van Schaik Publishers, Pretoria.

Mueller J 2000, Transformation operativer Daten zur Nutzung im Data Warehouse, Dt. Univ.-Verlag / Gabler, Wiesbaden.

Mutchnick R and Berg B 1996, Research Methods for the Social Sciences: Practice and Applications, Simon & Chuster Company, USA, p 7.

Naiman C F and Ouksel A M 1995, "A Classification of Semantic Conflicts in Heterogeneous Database Systems", Journal of Organizational Computing, Vol. 5.

Nature 2008 Big Data. Retrieved November 5, 2013, from the World Wide Web: http://www.nature.com/news/specials/bigdata/index.html

Naumann F and Rolker C 2000, Assessment Methods for Information Quality Criteria, In: Proceedings of the 2000 Conference on Information Quality, Cambridge, MA 1999, pp 148-162.

Newell, A 1990, Unified Theories of Cognition, Cambridge, Mass, Harvard University Press, USA.

Niehaves B 2007. "On Epistemological Diversity in Design Science - New Vistas for a Design-Oriented IS Research?" in the proceedings of ICIS 2007, Montreal.

Noy N F and Mc-Guinness DL 2001, Ontology Development 101: A Guide to Creating

Your First Ontology, Available at http://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html.

Nunamaker J, Chen M and Purdin T 1991, “System Development in Information Systems Research”, Journal of Management Information Systems, 7:3, pp. 89 – 106.

Olson J E 2003, Data Quality: The Accuracy Dimension, Morgan Kaufmann, San Francisco.

Orlikowski W and Iacono C 2001, "Desperately Seeking the "IT" in IT Research - A Call to Theorizing the IT Artifact”, Information Systems Research 122: 121-134.

Orr K 1998, Data quality and systems theory, Communications of the ACM, 412, p 66-71.

Owen C 1997, "Design Research: Building the Knowledge Base" Journal of the Japanese Society for the Science of Design 52: 36-45.

Patel-Schneider PF, Hayes P and Horrocks I 2004, OWL Web Ontology Language Semantics and Abstract Syntax, Editors. W3C Recommendation accessed on 21st July 2012 from http://www.w3.org/TR/2004/REC-owl-semantics-20040210/.

Paulk M 1999, “Using the Software CMM with Good Judgment,” ASQ Software Quality Professional, vol. 1, no. 3, pp 19-29

Peffers K, Gengler C and Tuunanen, T 2003, Extending Critical Success Factors Methodology to Facilitate Broadly Participative Information Systems Planning, Journal of Management Information Systems, 20, 1, 51-85.

Pitt L F, Watson R.T and Kavan C B 1997, 'Measuring Information Systems Service Quality: Concerns for a Complete Canvas', Management Information Systems Quarterly, pp. 209-221.

Pitt, L F, Watson R.T and Kavan C B 1997, "Measuring Information Systems Service Quality: Concerns for a Complete Canvas", Management Information Systems Quarterly, pp. 209-221.

Ponniah P 2011, Data Warehousing Fundamentals for It Professionals, John Wiley & Sons, New York.

Pries-Heje J, Baskerville R and Venable J 2008, "Strategies for Design Science Research

Evaluation," 16th European Conference on Information Systems ECIS, Galway, Ireland, pp. 255-266.

Proceedings of 10th International Conference on Scientific and Statistical Database Management SSDBM, Capri, Italy.

Production Workflows, Knowledge and Information Systems 12.

Purao S 2002, "Design Research in the Technology of Information Systems:

Truth or Dare." GSU Department of CIS Working Paper, Atlanta.

Radan N 1996, Warehouses and the Web, Information Week, 579, pp. 80-86.

Redman T 1998, The Impact of Poor Data Quality on the Typical Enterprise, Communications of the ACM, 41, 79-82.

Redman T C 1996, Data Quality for the Information Age, Artech House, New Delhi.

Reich, Y 1994, The Study of Design Methodology. Journal of Mechanical Design, 117, 2, p 211-214.

Reingruber MC and Gregory W W 1994, Data Modeling Handbook-A Best-Practice Approach to Building Quality Data Models, Wiley-QED, pp. 293, 334.

Rizzi S, Abelló A, Lechtenbörger J and Trujillo J 2006, Research in data wareho Modelling and design: dead or alive?, in I.-Y. Song, P. Vassiliadis Eds., DOLAP, ACM, pp. 3–10.

Romero O and Abelló A 2007, Automating multidimensional design from ontologies, in: DOLAP, pp. 1–8.

Romero O and Abelló A 2007, Automating multidimensional design from Ontologies, Proceedings of ACM10th International Workshop on Data Warehousing and OLAP, ACM, pp. 1–8.

Rossi M and Sein M K 2003, "Design Research workshop: A proactive research approach," 26th Information Systems Research Seminar in Scandinavia, Haikko, Finland.

Rudra A and Yeo E 1999, "Key Issues in Achieving Data Quality and Consistency in Data Warehousing among Large Organizations in Australia", Proceedings of the 32nd Hawaii International Conference on System Sciences.

Rudra A and Yeo E 1999, Key Issues in Achieving Data Quality and Consistency in Data

Warehousing among Large Organizations in Australia, HICSS-32 Hawaii International Conference on System Sciences, Maui, Hawaii.

Russell S and Norvig P 1995, Artificial Intelligence: A Modern Approach, Prentice Hall, Englewood Cliffs, New Jersey.

Samuel-Ojo O, Shimabukuro D, Chatterjee S, Muthui M, Babineau T, Prasertsilp, P., Ewais, S., and Young, M. 2010 "Meta-analysis of Design Science Research within the IS Community: Trends, Patterns, and Outcomes," in Global Perspectives on Design Science Research, R. Winter, L. Zhao and S. Aier eds., Springer, Berlin, pp. 124-138.

Schiefer J, List B and Bruckner RM 2002, A holistic approach to managing requirements of data warehouse systems, 8th Americas Conference on Information Systems AMCIS 2002, pp. 77–87

Science 2011 Special online collection: Dealing with data. Retrieved November 5, 2013, from the World Wide Web: http://www.sciencemag.org/site/special/data/

Scime A and Kerschberg L 2000, "Web Sifter: An Ontology-Based Personalizable Search Agent for the Web," presented at International Conference on Digital Libraries: Research and Practice, Kyoto Japan.

Scime A and Kerschberg L 2001, "WebSifter: An Ontological Web-Mining Agent for E-Business," presented at the IFIP 2.6 Working Conference on Data Semantics DS-9, Hong Kong, China.

Sebeok and Thomas A 1994, An Introduction to Semiotics, SAGE, London.

Serrano M, Calero C and Piattini M 2002, Validating metrics for data warehouses, IEE Proceedings SOFTWARE 149, p 161–166.

Serrano M, Calero C, Trujillo J, Lujan S and Piattini M 2004, Empirical validation of metrics for data warehouses, 4th ASERC Workshop on Quantitative and Soft Computing Based Software Engineering QSSE 2004, Banff, Alberta Canada.

Shankaranarayan G, Ziad M and Wang R Y 2003, Managing data quality in dynamic decision environments: an information product approach. Journal of Data Management, vol 14, no. 4, pp. 14-32.

Shankaranarayanan, G., Ziad, M., & Wang, R. Y. 2012 Preliminary Study on Data Quality

Assessment of Socialized Media. China Science and Technology Resources 442, pp 72–79.

Shanks G and Darke P 1998b, Understanding Data Quality in a Data Warehouse: A Semiotic Approach, Proc. of the 1998 Conference on Information Quality, Boston, Massachusetts.

Shanks, G., & Corbitt, B. 1999 Understanding data quality: Social and cultural aspects. Procedures of the 10th Australasian Conference on Information Systems, Wellington: MCB University Press Ltd., pp 785–797.

Silberschatz, A., Korth, H., & Sudarshan, S. 2006 Database System Concepts, Beijing: Higher Education Press.

Silverston L, Inmon W H and Graziano K 1997, The Data Model Resource Book, John Wiley & Sons, Inc., New York.

Simon H 1996, The Sciences of the Artificial, 3rd Edition, MA, MIT Press, Cambridge

Si-Saıd S and Prat N 2003, Multidimensional Schemas Quality: Assessing and Balancing Analyzability and Simplicity. in: MA P Jeusfeld O, Ed., ER 2003 Workshops, pp. 140–151.

Smith M, Welty C and McGuinness D 2003, OWL Web Ontology Language Guide accessed on 21st July 2012 from http://www.w3.org/TR/2003/WD-owl-guide-20030331/.

Son S, Weitzel T and Laurent F 2005, "Designing a process-oriented framework for IT performance management systems," The Electronic Journal of Information Systems Evaluation 8:3, pp. 219-228.

Song, M., & Qin, Z. 2007 Reviews of Foreign Studies on Data Quality Management. Journal of Information 2, pp 7–9.

Stamper R 1992, Signs, Organizations, Norms and Information Systems, Proceedings 3rd

Australian Conference on Information Systems, Wollongong.

Starlab 2003, Systems Technology and Applications Research Laboratory home page, Faculty of Sciences, Department of Computer Science, Vrije Universiteit Brussel. Available at: http://www.starlab.vub.ac.be/default.htm.

Staudt Lerner B and Nico Habermann A 1990, Beyond Schema Evolution to Database Reorganization, ECOOP/OOPSLA 1990 Proceedings.

Sure Y, Erdmann M, Angele J., Staab S., Studer R and Wenke D 2002, OntoEdit: Collaborative Ontology Development for the Semantic Web. In Proceedings of the 1st International Semantic Web Conference −ISWC2002, Springer, LNCS.

Swartout B, Patil R, Knight K and Russ T 1996, Toward distributed use of large-scale Ontologies, In Proceedings of the Tenth Knowledge Acquisition for Knowledge−Based Systems Workshop, KAW '96 November 9−14, Banff, Alberta, Canada.

Takeda H, Veerkamp P, Tomiyama T and Yoshikawa H 1990, "Modeling Design Processes," AI Magazine_Winter: p 37-48.

Tayi G K and Ballou D P 1998, Examining Data Quality, In Communications of the ACM, 412, p 54-57.

Technology, Springer, Ch. Ontologies, Metamodels, and the Model-Driven Paradigm, Pages 249–273.

Theodoratos D and Sellis T 1998, Data Warehouse Schema and Instance Design, In Proc. of the 17th Intl. Conf. on Conceptual Modeling, Springer LNCS 1507, p 363-376.

Theodoratos D and Sellis T 1999, DWQ project Designing Data Warehouses, SAGE, London.

Theodoratos D, Ligoudistianos S and Sellis T 1999, DWQ project Designing the Global Data Warehouse with SPJ Views, Proceedings CAISE '99, Heidelberg, Germany.

Thwaites, Tony, Davies L and Mules W 2002, Introducing Cultural and Media Studies: A Semiotic Approach, Palgrave Macmillan, London.

Trujillo J, Palomar M, Gomez J and Song I Y 2001, Designing Data Warehouses with OO

Conceptual Models.,IEEE Computer, Special issue on Data Warehouses 34, p 66–75.

Tryfona N, Busborg F and Christiansen J 1999, Star ER: A Conceptual Model for Data Warehouse Design, ACM 2nd International Workshop on Data Warehousing and OLAP DOLAP' 99, ACM, Missouri USA, pp. 3–8.

Ulrich F 2006, Towards a Pluralistic Conception of Research Methods in Information Systems Research, Tel Aviv University, Department of Management, Research http://www.icb.unidue.de/fileadmin/ICB/research/research_reports/ICBReport07.pdf.

Vaishnavi V, Buchanan G and Kuechler W 1997, "A Data/Knowledge Paradigm for the Modelling and Design of Operations Support Systems", IEEE Transactions on Knowledge and Data Engineering, Vol. 9, No. 2, March-April 1997, pp. 275 – 291.

Van Aken J E 2004, Management research based on the paradigm of design sciences: the quest for field-tested and grounded technological rules, Journal of Management Studies, 412, 219-246.

Van Solingen R and Berghout E 1999, The Goal/Question/Metric Method: A Practical Guide for Quality Improvement of Software Development, McGraw-Hill, New York.

Varela F 1988, "Structural Coupling and the Origin of Meaning in a Simple Cellular Automata", The Semiotics of Cellular Communication in the Immune System, E. Scaraz, F. Celada, N. Michenson and T. Tada, Springer Verlag, New York.

Walls J, Widmeyer G and El Sawy O 1992, "Building an Information System Design Theory for Vigilant EIS" Information Systems Research **3**1, 36 - 59.

Walls J, Widmeyer G and El Sawy O 2004, "Assessing Information System Design Theory in Perspective: How Useful was our 1992 Initial Rendition." Journal of Information Technology Theory and Application 62: 43-58.

Wand Y and Wang R 1996, Anchoring Data Quality Dimensions in Ontological Foundations, Communications of the ACM, 39, p 86-95.

Wand Y and Weber R 1995, On the deep structure of information systems, Journal of

Information Systems, pp. 203–223.

Wang R 1998, A Product Perspective on Total Data Quality Management, the ACM, 41:2, 58-65.

Wang R Y, Kon H B and Madnick S E 1993, Data Quality Requirements Analysis and Modelling, In Proc. of 9 the International Conference on Data Engineering, pp. 670-677, IEEE Computer Society, Vienna, Austria.

Wang R Y, Lee YW, Pipino L L and Strong DM 1998, Manage your information as a the Sloan Management Review, pp. 95-105.

Wang R Y, Reddy MP and Kon H B 1995, Toward quality data: An attribute-based approach. Decision Support System, pp 349–372.

Wang R Y, Strong D and Guarascio L M 1994, Beyond Accuracy: What Data Quality Means to Data Consumers, Technical Report TDQM-94-10, Total Data Quality Management Research Program, MIT Sloan School of Management, Cambridge.

Wang R Y, Ziad M and Lee Y 2001, Data Quality, Kluwer Academic Publishers, Boston.

Wang Y and Strong DM 1996, Beyond Accuracy: What Data Quality Means to Data Consumers, Journal of Management Information Systems, 12, 5-34.

Wang, H., & Zhu, W. M. 2007 Quality of Audit Data: A Perspective of Evidence. Journal of Nanjing University Natural Sciences 431, pp 29–34.

Wang, J. L., Li, H., & Wang, Q. 2010 Research on ISO 8000 Series Standards for Data Quality. Standard Science 12, pp 44–46.

Wang, R. Y., & Strong, D. M. 1996 Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems 124, pp 5–33.

Wang, R., & Storey, V. 1995 Framework for Analysis of Quality Research. IEEE Transactions on Knowledge and Data Engineering 14, pp 623–637.

Wang, Y. F., Zhang, C. Z., Zhang, B. B., et al. 2007 A Survey of Data Cleaning. New Technology of Library and Information Service 12, pp 50–56.

Wayne W. E. 2004 “Data Quality and the Bottom Line: Achieving Business Success through a Commitment to High-Quality Data “, The Data warehouse Institute TDWI report, available at www.dw-institute.com.

Weber S, Beck, R and Gregory R 2011, Combining Design Science and Design Research

Perspectives - Findings of Three Prototyping Projects. in proceedings of HICSS 2011.

Weir R, Peng T and Jon K 2003, Best Practice for Implementing a Data warehouse: A Review of Strategic Alignment, DMDW.

Welty C and Fikes R 2006, A reusable ontology for fluents in OWL. In The 4th International Conference on Formal Ontology in Information Systems FOIS 2006, p 226–336, IOS Press.

Widom J 1995, Research Problems in Data Warehousing. In Proc. of the 4th Intl. Conf. on Information and Knowledge Management, p 25–30, 1995.

Wiener J, Gupta H, Labio W, Zhuge, Garcia-Molina H and Widom J 1996, A System The prototype for Warehouse View Maintenance. In Workshop on Materialized Views: Techniques and Applications.

Winkler W E 2004, Methods for Evaluating and Creating Data Quality, Information Systems, vol. 29, no. 7.

Winter R 2008, "Design Science Research in Europe." European Journal of Information Systems 175: 470-475.

Zhou G, Hull R, King R and Franchitti J C 1995, Supporting Data Integration and Warehousing Using H20, Data Engineering, 182:29–40, 1995.

Zhu, X., & Gauch, S. 2000 Incorporating quality metrics in centralized/distributed information retrieval on the World Wide Web. Procedures of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Athens: ACM, pp 288–295.

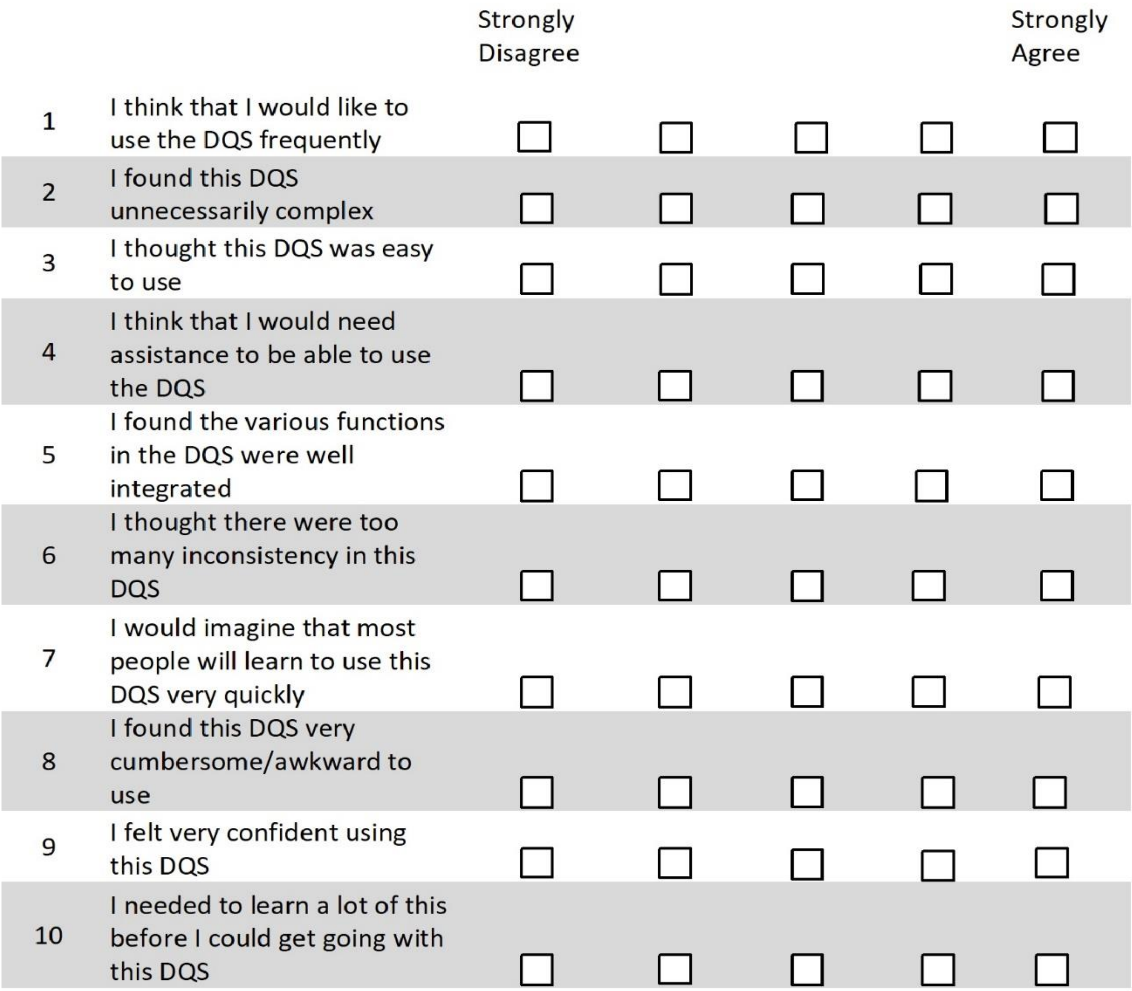
Zhu, Y. Y., & Xiong, Y. 2009 Datalogy and Data Science, Shanghai: Fudan University Press.

Zhuge Y, Garcia-Molina H, Hammer J and Widom J 1995, View Maintenance in a Warehousing Environment. In Proc. of the ACM SIGMOD Intl. Conf. on Management of Data, pages 316–327.

Zong, W., & Wu, F.2013 The Challenge of Data Quality in the Big Data Age. Journal of Xi'an Jiaotong University Social Sciences 335, pp 38–43.

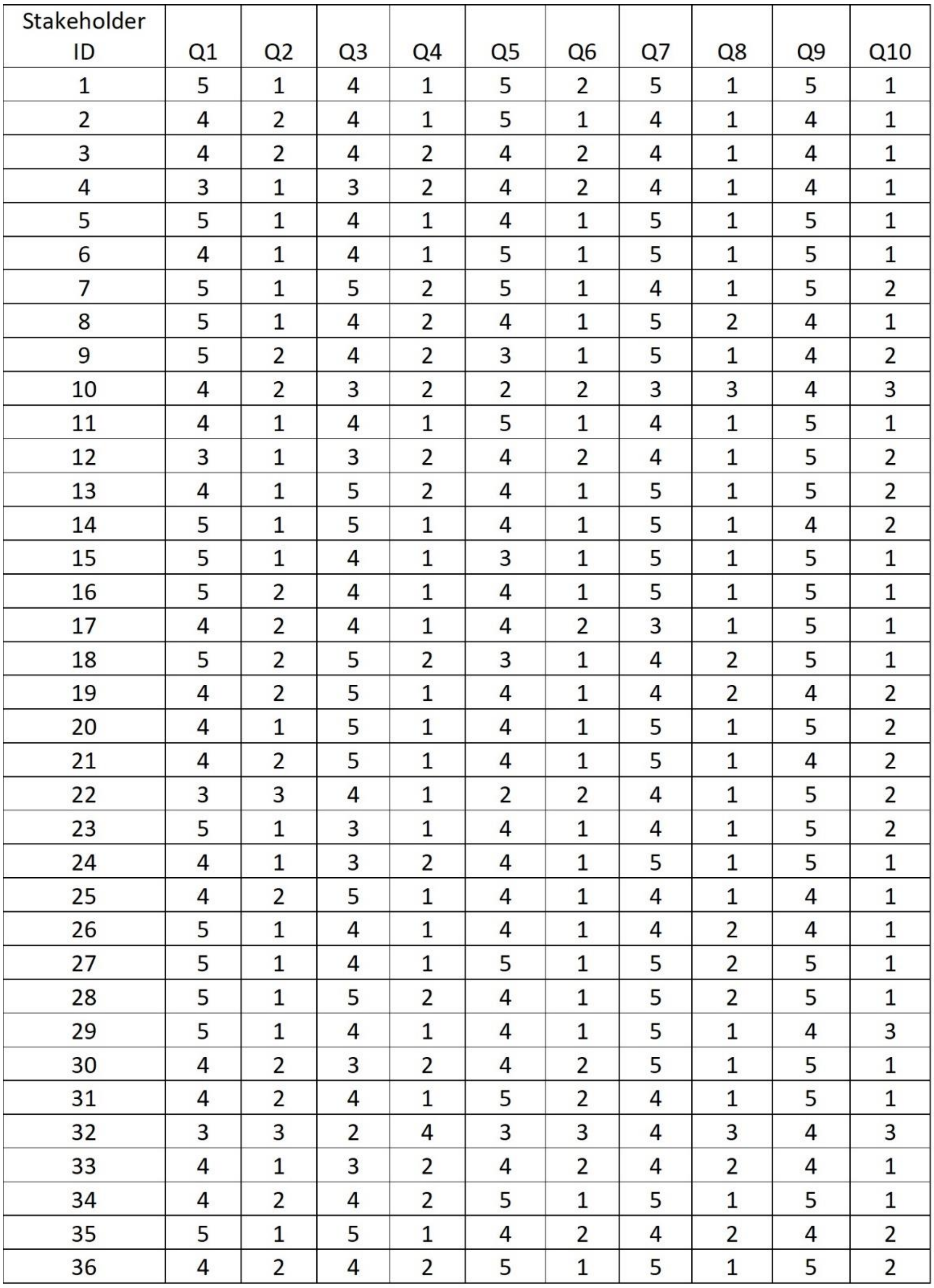
Appendix

Appendix A: Questionaire items



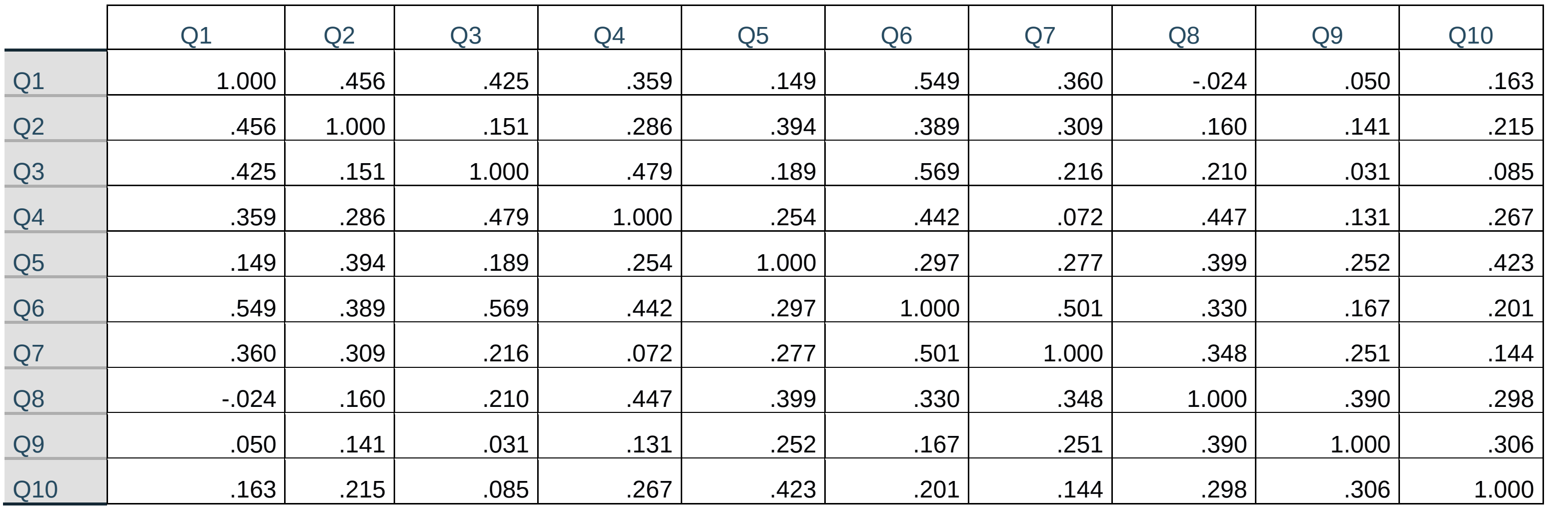
		Strongly Disagree				Strongly Agree
1	I think that I would like to use the DQS frequently	□	□	□	□	□
2	I found this DQS unnecessarily complex	□	□	□	□	□
3	I thought this DQS was easy to use	□	□	□	□	□
4	I think that I would need assistance to be able to use the DQS	□	□	□	□	□
5	I found the various functions in the DQS were well integrated	□	□	□	□	□
6	I thought there were too many inconsistency in this DQS	□	□	□	□	□
7	I would imagine that most people will learn to use this DQS very quickly	□	□	□	□	□
8	I found this DQS very cumbersome/awkward to use	□	□	□	□	□
9	I felt very confident using this DQS	□	□	□	□	□
10	I needed to learn a lot of this before I could get going with this DQS	□	□	□	□	□

Appendix B: SUS Raw Questionaire Results



Stakeholder ID	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
1	5	1	4	1	5	2	5	1	5	1
2	4	2	4	1	5	1	4	1	4	1
3	4	2	4	2	4	2	4	1	4	1
4	3	1	3	2	4	2	4	1	4	1
5	5	1	4	1	4	1	5	1	5	1
6	4	1	4	1	5	1	5	1	5	1
7	5	1	5	2	5	1	4	1	5	2
8	5	1	4	2	4	1	5	2	4	1
9	5	2	4	2	3	1	5	1	4	2
10	4	2	3	2	2	2	3	3	4	3
11	4	1	4	1	5	1	4	1	5	1
12	3	1	3	2	4	2	4	1	5	2
13	4	1	5	2	4	1	5	1	5	2
14	5	1	5	1	4	1	5	1	4	2
15	5	1	4	1	3	1	5	1	5	1
16	5	2	4	1	4	1	5	1	5	1
17	4	2	4	1	4	2	3	1	5	1
18	5	2	5	2	3	1	4	2	5	1
19	4	2	5	1	4	1	4	2	4	2
20	4	1	5	1	4	1	5	1	5	2
21	4	2	5	1	4	1	5	1	4	2
22	3	3	4	1	2	2	4	1	5	2
23	5	1	3	1	4	1	4	1	5	2
24	4	1	3	2	4	1	5	1	5	1
25	4	2	5	1	4	1	4	1	4	1
26	5	1	4	1	4	1	4	2	4	1
27	5	1	4	1	5	1	5	2	5	1
28	5	1	5	2	4	1	5	2	5	1
29	5	1	4	1	4	1	5	1	4	3
30	4	2	3	2	4	2	5	1	5	1
31	4	2	4	1	5	2	4	1	5	1
32	3	3	2	4	3	3	4	3	4	3
33	4	1	3	2	4	2	4	2	4	1
34	4	2	4	2	5	1	5	1	5	1
35	5	1	5	1	4	2	4	2	4	2
36	4	2	4	2	5	1	5	1	5	2

Appendix C: Inter-Question Correlation Matrix



	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
Q1	1.000	.456	.425	.359	.149	.549	.360	-.024	.050	.163
Q2	.456	1.000	.151	.286	.394	.389	.309	.160	.141	.215
Q3	.425	.151	1.000	.479	.189	.569	.216	.210	.031	.085
Q4	.359	.286	.479	1.000	.254	.442	.072	.447	.131	.267
Q5	.149	.394	.189	.254	1.000	.297	.277	.399	.252	.423
Q6	.549	.389	.569	.442	.297	1.000	.501	.330	.167	.201
Q7	.360	.309	.216	.072	.277	.501	1.000	.348	.251	.144
Q8	-.024	.160	.210	.447	.399	.330	.348	1.000	.390	.298
Q9	.050	.141	.031	.131	.252	.167	.251	.390	1.000	.306
Q10	.163	.215	.085	.267	.423	.201	.144	.298	.306	1.000

Appendix D: Qualitative Data Items – Brewery Ltd

Participant ID	Participants Responses	Initial themes
1	The scorecard provides an additional layer of reporting that gives excellent information about the data	Additional reporting tool Informative Report dashboard
2	Yes, it does provide real-time data analysis, quite a quick way of checking the quality of the database	Informative Very useful Real life cases Additional checkpoint
3	I can see this integrating well with our reporting landscape will work well with our other reports	Integration Can be expanded Additional reporting tools

4	Handy tool to have in addition to our other data quality measurement tools. I get the idea, and I think it’s very useful	Handy tool Very useful
5	Simple traffic light design for scoring works well and easy to understand So easy to use, no previous knowledge required to fill the scorecard, I like that	Simplicity Easy to understand No previous knowledge
6	Erm… Yes, I do like it Very simple, straightforward	Simplicity straightforward
7	Gives an end to end information of the data from inception to usage	Informative Cuts across all areas
8	The information trail is fantastic. I can see what everybody else thinks about the data	Information trail Informative Awareness
9	I like how I can use it to view my colleague's views about the data	Informative Awareness
10	The arrangement of the scorecard by data dimensions is a very good idea	organised
11	The look and feel of the scorecard is consistent with our approach to the measurement of data quality, I like it	Consistent Easy to use interesting

12	The questions are well organised, I’m sure we will find it useful	organised
13	Cool idea, but timeliness dimension not too important for us I really like the concept	Clear Useful
14	Provides a good snapshot of the expected quality of the data in the data warehouse. Good initial reference tool to have	Additional reporting tool Useful Report dashboard
15	Gives me comfort to know my colleagues have rated the data quality already. I like the idea of having all data stakeholders rating the portion of the scorecard that relates to their area	Comfort Transparent Consistent
16	Provides a good snapshot of the expected quality of the data in the data warehouse. Good initial reference tool to have	Snapshot Initial reference
17	My comfort level is definitely higher with the use of the scorecard. It makes the job of ensuring everyone is ok with the data load easier I guess	Comfort Useful
18	Yea, I do have more confidence using the data now than I did before using	Confidence Informative

	the scorecard, shows what knowing a bit more about the data can do.	Knowledge of data
19	I can see other uses for this scorecard, I believe we should be able to develop the report generated better and include it in our report dashboard	Data was seen in a new light Report dashboard Can be expanded
20	Very nice tool, not sure I can see the need for the timeliness section… I do like it though. I particularly like that we can start using it immediately in our data warehouse.	Handy tool Immediate usage

Appendix E: Qualitative Data Items – Oil and Gas Ltd

Participant ID	Participants Responses	Initial themes
1	The scorecard is clear and easy to use, the questions are straightforward	Clear Easy to use straightforward
2	Everything is clear and quite easy to get to, it’s all self-explanatory, nothing hidden, Very simple and easy layout	Clear Nothing hidden Self-explanatory Simple

3	Very precise and straight to the point, I really do like the simplicity and the way it separates the scenario questions based on our roles. It gives some useful knowledge of the data, Comfortable using the scorecard	Very precise Simplicity Useful comfortable
4	I have confidence in using the scorecard, looks impressive and something we can really use. The report is also useful	Confidence Useful
5	Nice looking website, This is very interest and no technical language used, very simple website	Very interesting Simple
6	Looks well organised, all groups go to separate areas is a good idea, it will be good also if the reports can be generated by the roles. A very interesting and simple tool	Organised Interesting Simple

7	No training at all required, Cuts across all areas, no need to train anybody	No training required Cuts across all areas
8	The reports are very useful and handy as a reference. I like the scorecard, can actually use it on a daily.	Very useful Handy reference
9	Yea, I do have more confidence using the data now than I did before the scorecard measurement	Confidence
10	The tool will give an added comfort to the quality of our data usage	Comfort
11	very useful tool to have, wonder why we didn't have this before now. Having everyone involved in the data warehouse reviewing the data in a single place is a good idea	Very useful integration
12	really like the concept, as I mentioned earlier, the scorecard can definitely play	Like the concept

	a role here as part of our data quality management	
13	We need this scorecard to be on our intranet rather than in the public domain, apart from that I find it very useful when used with our data quality control procedure	Very useful
14	It's erm..very simple and functional. The website looks good too, not too busy, quite efficient	Simplicity
15	The report function is very useful, provides a good idea of what others think about the data. The dimensions are ok, but the scenario questions might need changing from time to time	Very useful