THE DESIGN OF A DATA QUALITY ASSESSMENT TOOL FOR WAREHOUSING DATA

Bart Lesschen
b.s.lesschen@student.utwente.nl
S2145545

Supervisors University of Twente:
Dr. G. Sedrakyan
J.P.S. Piest

Supervisor Bricklog:
H. Benneker

Management Summary

This research has been performed at Bricklog B.V. in Apeldoorn and Enschede. Bricklog is a company founded in 2015 in Apeldoorn. Since September 2022 they also have an office in Enschede. Currently, Bricklog is helping Small and Medium-sized enterprises(SMEs) become data-driven in the transport sector. Since almost every transport company also has a warehouse Bricklog wants to expand to the warehousing sector as well.

Bricklog wants to help its customers gain useful insights for their warehouse. To gain useful insights sufficient enough data quality is needed. For this sufficient data quality, Bricklog needs a generic solution which they can implement easily for their customers. Since this research is focused on the design of an IT artefact the Design Science Research Methodology(DSRM) was used. To perform this research the main research question was stated:

MRQ: Can a generic approach to assess the data quality of warehousing data with the use of a data quality assessment instrument be designed for Bricklog to help their customers improve their data?

First, the problems in the sector have to be understood. After identifying multiple problems a problem cluster was determined. From the problem cluster was the core problem identified which is that there is no data quality filter for warehousing data. To solve this problem a data quality assessment tool is designed & developed.

To perform this research a Systematic Literature Review(SLR) has been performed to identify possible data quality frameworks or theories applicable to this research. Following this research, it became clear that no theory is directly applicable to this research. However, data quality dimensions were derived from theory which can be used. These data quality dimensions are completeness, accuracy, uniqueness and consistency. These data quality dimensions apply to the key concepts of data in warehousing. These key concepts are mutations, locations and stock.

With this gathered knowledge the data quality assessment tool is developed in Power BI. To create a generic solution a generic data model is developed. This data model is split into two, a static model and a dynamic model. From here the tool is also split into 2 dashboards. Each dashboard has 2 data quality check pages. The dynamic dashboard is solely focused on mutations and the static dashboard is focused on locations and stock.

After the development, the data quality assessment tool is validated by an expert panel consisting of four representatives from Bricklog. This is done by employing a survey. The survey followed the concepts of the Technology Acceptance Model (TAM) to gain insights into the perceived ease of use and perceived usefulness of the data quality assessment tool.

The expert panel accepted the data quality assessment tool however, this is no guarantee that the tool will be useful in practice since it was only possible to validate internally which is the main limitation of this research. For Bricklog it is advised to implement this data quality assessment tool at a customer of Bricklog with caution.

Acknowledgements

Dear reader,

This is the final version of my bachelor thesis, which I completed during my last year in the Industrial Engineering and Management (IEM) program at the University of Twente. The research was carried out in collaboration with Bricklog, a company based in Apeldoorn. I thoroughly enjoyed being part of the Bricklog team and taking part in various activities, both work-related and social. Learning Power BI was particularly beneficial, not only for my thesis but also for my future career. I would like to thank Hubert Benneker, my supervisor at Bricklog, for his guidance in ensuring my thesis had practical value, rather than being purely theoretical.

I would also like to thank my supervisors from University of Twente for their guidance and patience. It has been a long process so thank you Gayane Sedrakyan and Sebastian Piest to help me with the academic parts of my thesis.

Last but not least I want to thank my friends and family for their supportive actions.

Kind Regards,

Bart Lesschen

May 2024

Table of Contents

List of Figures

List of Tables

1. Introduction

This first chapter introduces this bachelor assignment. In section 1.1 the bachelor assignment will be introduced. In section 1.2 the company Bricklog is described. In section 1.3 the industry and digitalization will be explained. In section 1.4 the problem context is introduced. In section 1.5 the research design and the used methodology. In section 1.6 the main research question is introduced along with research questions for every step of the methodology. In section 1.7 the practical and scientific contribution is explained. In section 1.8 the scope of Bricklog is determined. In section 1.9 the thesis structure will be explained.

1.1 Introduction

During the previous months, this bachelor assignment has been performed in collaboration with Bricklog B.V. Bricklog is a company that specializes in advising transport companies and helping them become data-driven. The goal of becoming data-driven is to make business decisions and decide on their strategy based on their historical daily data. Together with the staff, a bachelor assignment has been created, which will help Bricklog improve the data quality of the warehouses of the transport companies. Because Bricklog is a data club the goal is to provide a general solution which can be easily implemented at all their customers.

1.2 Bricklog

Bricklog was founded in 2015 and is based in Apeldoorn and since September 2023 also in Enschede. Bricklog has 15 full-time employees and Bricklog offers the opportunity for internships and graduation projects for a small number of students every six months.

Bricklog identified a pressing need for change within the transport and logistics sector. His goal is to help transport companies change. In the beginning, the company was providing solely consultancy however, that was not working as expected because the companies it tried helping had a lot of issues with the data quality. For instance, data quality issues such as a trip spanning 80,000 kilometres that is unrealistic significantly and negatively impact the solutions proposed by Bricklog. Bricklog wanted to help the other company change to be data-driven. Data-driven will be used to support decision-making, "*Data-driven decision making (DDD) refers to the practice of basing decisions on the analysis of data rather than purely on intuition."* (Provost & Fawcett, 2013) Thus, over the years Bricklog transformed from a sole consultancy company to a data club which helps companies become data-driven and from there the data can be used to help improve strategies, dashboards and so on. In addition to assisting companies in adopting a data-driven approach, Bricklog facilitates the practical utilization of data and fosters trust in its reliability for informed business decision-making. This shift represents a significant transformation for transport companies, emphasizing the pivotal role of change management integrated with data technology

The main customers of Bricklog are Small and Medium-sized Enterprises (SMEs) within the transport and logistics sector. However, Bricklog also takes on challenges in other sectors such as the waste industry and the insurance industry. Next to giving consultancy and helping companies become data-driven, Bricklog tries to stay ahead of the market and predict the demand from the sector. Almost every transport company also has a warehouse to store goods or products before transporting them to another location. Bricklog predicts that insights for the warehouse will soon be in demand from their customers.

1.3 Industry Context

"Firms of all sizes, across all sectors are increasingly equipping their staff with digital tools." (*The Digital Transformation of SMEs*, n.d.). Digital tools bring many benefits for firms and they can help

SMEs integrate in global markets. Digitalization helps firms to generate data and analyse their operations to increase performance. Early Evidence suggests that SMEs intensified their use of digital technologies due to COVID-19. Due to this accelerated use of digital technologies many SMEs have not had the time to plan their use of digital systems and select the right tools. For these SMEs the transition is going on and not yet complete and this comes with risks for security and quality. “the logistics business becomes complex due to globalization and ever changing market and consumer behaviour” (Andiyappillai, 2020). With the use of Warehouse Management Systems(WMS) businesses can capture the right data as much as possible and also analyse this data extensively to become more efficient in performance. Analysing this data can give insights into order and business profile, hardware/software configuration, labour productivity, warehouse space optimization and much more. The next step for the industry is the implementation of Smart warehouses. Smart warehouses use Internet of Things(IoT) Technology, Augmented Reality(AR), big data analytics, digital twin and machine learning (van Geest et al., 2021). However, before these smart warehouses can be realized for SMEs analytical tools are needed to analyse the data from the WMS.

1.4 Problem Context

In the transport sector, every company has been collecting data for multiple years but most companies do not know what to do with that data and they do not have the time to figure it out. Furthermore, the data is not of good quality, e.g. in the data, important inputs are missing or wrong. Thus, even when a company has time or knowledge to build a Business Intelligence (BI) report the insights from that report are almost always not valid. For example, if a company wants to calculate how many CO^2 emissions a certain truck has emitted they can calculate this with their data. However, if a trip has recorded 7000 kilometres instead of the 70 kilometres the trip actually was, this will calculate a much larger number of CO^2 emissions than reality. On the other hand, the kilometres could have been empty in the data and this would mean that the number of CO^2 emissions is smaller than reality. Thus, these mistakes in the data prevent companies from creating useful BI reports and supporting their decision-making.

Bricklog already has a working process to help companies become data-driven for the transport part of the client. They have improved and tweaked this process over the years so it is a working process and has clear goals before taking the next step in the process. The main problem at the beginning of the process is that the customer's database contains numerous errors. Bricklog refrains from cleaning the data on behalf of the client to avoid setting the expectation that Bricklog will continue to provide such services in the future. Their customer has to learn to work with data and the goal is to change the mindset of the client and help them provide clean data so a big part of that process includes change management. Their data has to be of a certain quality before taking the next step until they can get valid insights into their company and decide their strategy.

Currently, Bricklog is mostly focused on the transport side of the customer. However, a transport company almost always has a warehouse and the problems such as missing inputs or wrong inputs in the data, are comparable to the transport side. The warehouses are not being used efficiently which causes longer times for tasks to complete and can be also costly. Efficiency in warehouses can be defined in multiple ways with the use of multiple metrics e.g. order lead time and order picking time, lead time from order placement to order shipment and lead time to pick an order respectively (Staudt et al., 2015). The type of metrics used to determine efficiency depends on the type of warehouse and the objectives of the warehouse manager. On top of pallets, products are stored in the warehouse, with the use of forklifts or other equipment pallets can be easily moved inside a warehouse. However, pallets are being moved constantly additionally, certain warehouses lack a

strategic framework altogether. This issue can be solved by looking at the data and determining a strategy for a concrete case. However, the available size of data can be huge. Furthermore, similar to the transport domain this data contains numerous mistakes and requires time or knowledge to gather valid insights from that data can be lacking. This is the case with almost all clients of Bricklog. Therefore, Bricklog will start a new process with new goals and targets to help their clients also become data-driven on the warehouse side of the company. Bricklog does not aim for another long-term transformation process for the warehouse side of their clients. **Thus, the action problem for this bachelor assignment is to help Bricklog improve the data quality of their customers up to a certain quality of their potential customers and ensure the customers keep improving.** The objective is not about reaching that certain quality, but rather ensuring that the customer keeps on improving their data quality compared to previous situations.

1.5 Research Design

The research design of this research is described in the following section. This study is designed according to the Design Science Research Methodology (DSRM), which consists of six steps (Peffers et al., 2007). The main deliverable of this bachelor assignment is a data quality assessment tool for warehousing built in a dashboard style. DSRM has been used to create an IT artefact. DSRM implies designing and building novel artefacts iteratively, in each iteration going through (re)evaluation loops. Due to the time frame, one iteration of DSRM is executed. The artefact is evaluated by experts of Bricklog through a survey. The survey follows the Technology Acceptance Model(TAM)(Chuttur, 2009)**.** The Design Science Research Methodology consists of the following 6 steps:

1. Problem identification and motivation: In this phase, the problem is identified and motivated by what value the solution will add.
2. Define the objectives of a solution: In this phase, the objectives are derived from the problem identification.
3. Design and Development: In this phase, the artefact will be designed and possibly developed with a prototype.
4. Demonstration: In this phase, the artefact is demonstrated if it functions as it is intended with synthetic data.
5. Evaluation: In this phase, the designed artefact is evaluated and checked if it meets the first 2 phases of the methodology. It is possible to iterate between Phase 3 and Phase 5 depending on modification needs.
6. Communication: In the last phase the artefact is presented to the stakeholders and other relevant audiences.

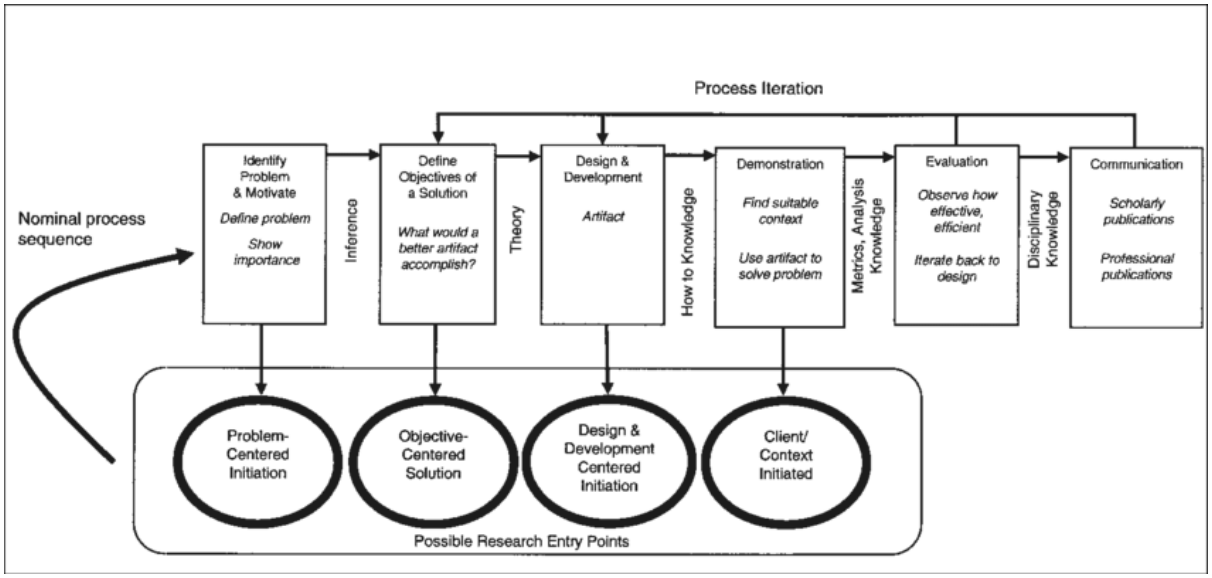

Figure 1 - Design Science Research Methodology (Peffers et al., 2007)

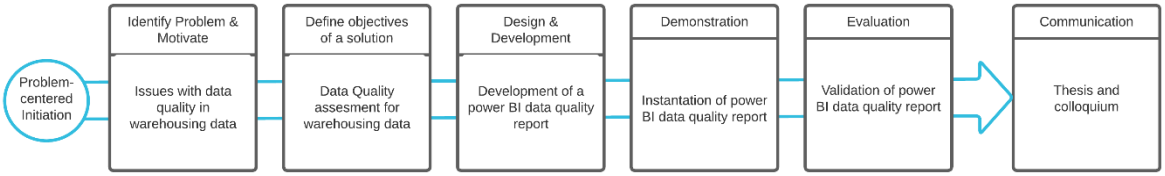

Figure 2 - Iteration of the DSRM

Bricklog is a consultancy company that helps companies become data-driven. In the first stage with a customer, they create basic dashboards to already give some insight based on the data from that customer. In this phase, Bricklog also assesses the data quality of the customer for the transport side. In a later phase they will help the customer improve the data quality by looking at their process but to do that the data quality first has to be assessed which is the scope of this research.

1.6 Research Questions

The structure of this thesis is determined by 6 research questions. The Main Research Question relates to the research objective:

MRQ: Can a generic approach to assess the data quality of warehousing data with the use of a data quality assessment instrument be designed for Bricklog to help their customers improve their data?

To improve data quality an assessment of the data quality has to be made. This is a knowledge gap. To fill this knowledge gap a knowledge question has been formulated:

KQ: What theories or models are available for data quality assessment to warehousing data within the scope of Bricklog?

To answer this KQ a systematic literature review has been carried out. The goal of this systematic literature review is to find an existing data quality theory or framework which is within the scope of Bricklog.

Furthermore, for the 6 steps of the DSRM, a research question has been developed to structure the research according to the DSRM and all within the scope of Bricklog. Step 4 and 5 have been combined since the demonstration and evaluation is within the same step since this research is only internally validated.

Phase of DSRM	Research Question	Chapter
Problem identification and motivation	RQ1: *What is the core problem which leads to data quality errors in warehousing data?*	2
Define the objectives of the solution	RQ2: *What are the objectives of the solution?*	2
Design and Development	RQ3: *What metrics from data quality assessment frameworks are relevant for data quality assessment instruments?*	5
Demonstration and Evaluation	RQ4: *Can the developed data quality assessment instruments be used in practice?*	6

1.7 Contributions

The practical contribution of this bachelor assignment will be a data quality assessment instrument for Bricklog for warehousing data. This will be made in Power BI since that is the program that Bricklog works with. This tool includes a dashboard style and has multiple pages to address multiple parts of the warehousing data.

The scientific contribution of this research is the design and development of a novel bottom-up approach to data quality assessment for warehousing data. With this research, it is shown that dimensions from data quality frameworks can be used to develop a data quality assessment tool.

The last contribution includes recommendations for further research. For Bricklog this could be interesting as well as for other researchers.

1.8 Thesis Outline

This thesis consists of the following chapters. The contents of each chapter are briefly described in this section.

Chapter 2, the first general context is given about warehousing. Furthermore, in this chapter, the problem is further analysed and the core problem is identified. From here the objectives of the solution are determined.

Chapter 3 contains the systematic literature view. With this literature review data quality relevant assessment theories are found and applied to the warehousing context.

Chapter 4 explains the DSRM in more detail and applies it to this specific research.

Chapter 5 contains the development of the data quality assessment tool. The theories of Chapter 3 are applied in the development of the dashboard.

Chapter 6 is about the validation of the data quality assessment tool. The tool is validated by an expert panel consisting of four representatives. A survey was conducted following the TAM.

Chapter 7 is the final chapter which contains the conclusion, a summary of the findings, the limitations, the scientific contribution of this thesis and recommendations for future research.

2. Context Analysis

In Chapter 2 general context for warehousing and the core problem of the warehousing sector regarding data quality is identified. In section 2.1 general information about different types of warehouses is explained. In section 2.2 different types of pallets are explained. In section 2.3.1 Bulk storage is explained and in section 2.3.2 Rack storage is explained. In section 2.4 Location names are explained. In section 2.5 the problem cluster has been determined and from this problem cluster in section 2.6, the core problem is identified. In section 2.7 the gap between the norm and reality is determined. And in section 2.8 the objectives of this research and solution are determined.

2.1 Warehouses

There exist different types of warehouses. Bricklog differentiates warehouses into 3 types of warehouses. Non scanned warehouse a scanned warehouse and a data-driven warehouse. The focus of this research is on scanned warehouses. A non-scanned warehouse won't have any data so those cannot be included in this research and a data-driven warehouse is what Bricklog wants to achieve with their customers.

In the following section background information about warehousing is explained. This information is essential to understand Data Quality Errors in warehousing data.

2.2 Blok and Euro Pallet

Pallets are the backbone of warehousing. A pallet is "*a portable platform for handling, storing, or moving materials and packages (as in warehouses, factories, or vehicles)*" (*Pallet Definition & Meaning - Merriam-Webster*, n.d.). There are all types of pallets but the 2 most used pallets warehousing are Blok and Euro pallets. The difference is the size of the pallet.

*Table 1 (*Pallet Afmetingen | Europallets, Blokpallets, Etc. | Palletcentrale*, n.d.)*

Type Pallet	**Size**
EuroPallet	80 x 120 cm
BlokPallet	100 x 120 cm

As can be seen from Table 1 the size differs by 20 centimetres. Due to this 20 cm difference, 4 Euro pallets can fit in the same space compared to only 3 Blok pallets in that same space. This difference is important when storing pallets.

2.3.1 Bulk Storage

A way of storing pallets in a warehouse is the use of Bulk Storage places. In bulk storage pallets are stacked on top of each other without shelves or any other kind of equipment. This is a cheap way of storing pallets since no equipment is needed. However, pallets cannot be reached easily if the pallet is on the bottom of the stack. Furthermore, there is a maximum amount of pallets that can be stacked on top of each other. The maximum capacity differs between Euro and Blok pallets. According to Bricklog, the maximum capacity for euro pallets in Block storage is: bottom layer 5 pallets, second layer 5 pallets third layer 4 pallets and top layer 2 pallets which is in total 16 pallets as maximum. For Blok Pallets is bottom layer 4 pallets, the second layer is 4 pallets, the third layer is 2 pallets and the top layer is 1 pallet which is in total 11 pallets.

2.3.2 Rack Storage

Another way of storing pallets in a warehouse is with the use of racks. Compared to Block storage this is a way of storing pallets with the use of equipment and pallets can be accessed since they are not stacked on each other. A rack is often 360 cm wide so in that space there could be 4 Euro Pallets

or 3 Blok Pallets. The height of a rack could differ for every warehouse but in practice, most warehouses do not use racks with higher than 7 layers.

2.4 Location Name

Most warehouses use the same logic for their location names. This logic is explained to help understand the subjects of the data in the artefact. Furthermore, this logic shows the distinction between rack and block storage by only looking at the location name.

First of all a warehouse consists of multiple halls, inside a hall there are multiple lanes with pallets. These lanes have been divided into rows. For rack stacking inside these rows, there is a level and within that level, there is a place for a pallet. Block storage does not have a level but does have a place, unfortunately in practice this place is never used in the data. Thus, the difference between the location names is the extra number for the level.

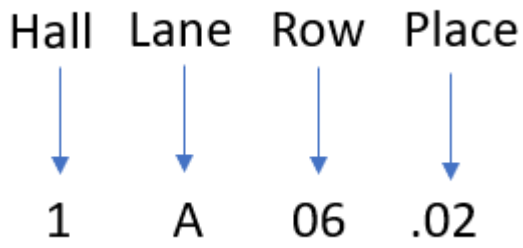

Figure 3 - Block Storage location name example

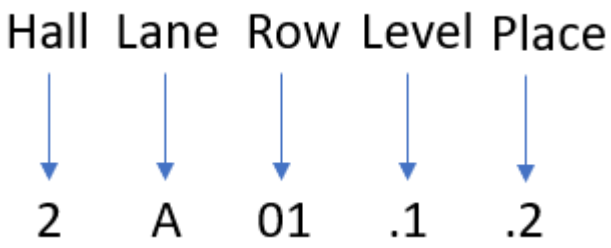

Figure 4 - Rack Storage location name example

2.5 Key Concepts

To understand warehousing the key concepts of warehousing are explained in this section

First of all, mutations are the backbone of a warehouse every incoming and outgoing pallet containing products is recorded through a mutation. Even movements of pallets within the warehouse are recorded in the mutation database. If there is sufficient enough data quality of mutations useful insights into the efficient use of the warehouse can be derived.

Second of all, locations are the place to store the pallets inside the warehouse. Good data quality of locations means that the data actually represents the current warehouse. This overcomes that employees go to the wrong location to pick up a product this saves time which reduces costs. With this data, decisions can also be made about optimal location or pallet placement.

The last of the key concepts of warehousing is the stock. The stock is what is actually inside the warehouse. Most warehouses keep track of their stock and use a database to store information

about the stock once every day. Good data quality stock ensures that the data actually represents the current amount of stock inside the warehouse.

2.6 Problem Cluster

In Chapter 1, the action problem has been described. To solve this action problem the core problem has to be identified. To achieve this a problem cluster has been created. To deduct the core problem from the problem cluster, the chain of problems will be followed until there is a problem with no direct cause. (Heerkens & van Winden, 2016)

In Figure 5 the problem cluster is presented. The action problem is highlighted in blue. With the use of the problem cluster, the core problem can be identified and it can be seen what other problems the action problem causes. Starting from the top, the 2 main problems of the sector include the missing or incorrectly constructed mandatory CO^2 reports and the inefficient use of the warehouse. The mandatory CO^2 reports that companies need to deliver to the government are missing or incorrectly constructed this is caused by the Key Performance Indicators (KPI) not displaying correctly. The inefficient use of the warehouse problem is also by the incorrect KPIs, if the KPIs were correct a strategy could be chosen and the warehouse would be used more efficiently. Another cause for the inefficient use of the warehouse is the high workload. If employees need to work very hard they often do not have the time to think about and do tasks more efficiently. This high workload is caused by staff shortages. The incorrect displaying of KPIs is caused by issues with the data quality and because different systems often do not communicate with each other. Due to the latter cause, the probability of duplicate or empty values is higher, and even the possibility exists that a lot of information is missing because of it. The bad data quality is caused by a lot of factors such as a wrong data entry in the system, wrong converting of a value by the system and a lot more and this differs at every warehouse. But one thing the warehouses have in common is that none of them has a data quality filter.

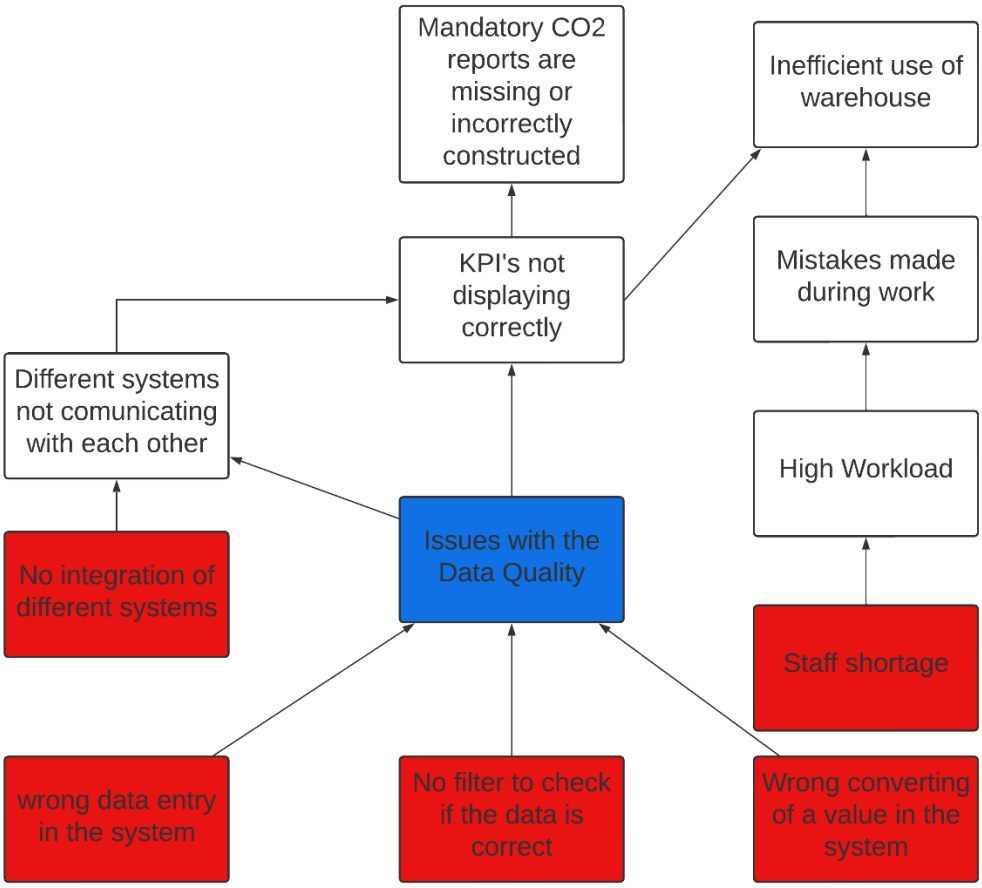

Figure 5 - Problem Cluster

2.7 Core Problem

As can be deducted from figure 1 there are 5 possibilities for a core problem these 5 problems are highlighted in red. To decide which has a higher priority is estimated by which of the problems yields the biggest impact while having the lowest cost(Heerkens & van Winden, 2016).

The first possible core problem is that there is no integration of different systems. Even if there is an integration of the systems, this could result in data mistakes. The systems could overwrite each other or create duplicate data. However, after consulting with experts from Bricklog they want a generic solution and creating a generic solution for integration of multiple systems is not possible within the time for this research.

The second possible core problem is the staff shortage. However, this problem does not directly influence the action problem. Furthermore, staff shortage is also not a generic solution since every warehouse could have different problems with personnel. This means that staff shortage is also not the core problem for this research.

The third possible core problem is the wrong conversion of a value in the system. To solve this problem research is needed into the specific system to locate the problem. However as mentioned before a generic solution is needed and different warehouses use different systems thus, this problem is not the core problem.

The next possible core problem is a wrong data entry in the system. This problem directly influences the action problem. However, a wrong data entry in the system is a human mistake. Solving these

human mistakes is difficult because most warehouses use a different system thus, preventing these mistakes is difficult. Furthermore, there is no clear overview of when or how these mistakes occurred. Thus, wrong data entry in the system is not the core problem of this research.

Instead of preventing a data entry, it is possible to locate the mistakes. This can be done by a data quality filter however, there is no data quality filter for warehousing data. With this data quality filter mistakes in the data can be located and it is even possible to show when these mistakes occurred. With this filter, data cleaning will be easier because the mistakes are already located and it is possible to hopefully prevent these mistakes in the future.

2.8 Gap Norm and Reality

Before solving the core problem, it is important to find and understand the difference between the norm and reality. This difference is the problem and it will be important that de provided solution can fill this gap between the norm and reality.

However, in reality there is not a strict norm for a certain percentage which is enough to take the next step. The norm stated by Bricklog is that data quality has to improve significantly and constantly before taking the next steps in the process. With this constantly improving Bricklog notices that their customer is active with improving the data quality and the company probably understands the importance of data quality and understands the impact of data quality errors.

2.9 Research Objective

This research was initiated by Bricklog. The action problem of this research has already been mentioned thus, the main objective of this research is:
The research objective is to design a generic tool in Power BI that can assess the data quality of warehousing data. To achieve this research objective the following requirements for the Data Quality Assessment Tool are determined with the experts of Bricklog:
Functional requirements:

1. Asses the data quality of Mutations.
2. Asses the data quality of Stock and Locations.
3. Generic solution applicable to multiple customers of Bricklog
4. Intuitive usability for end users
5. In the style of Bricklog

Technical requirements:

1. Built in Power BI
2. Load time does not exceed 2 seconds

2.10 Conclusion

In this chapter, context information about warehousing is given and the problem is further analysed. For this analysis, a problem cluster has been derived from the problems and the core problem could be identified. This answers the first research question:

RQ1: what is the core problem for data quality errors in warehousing data?"

Which is that there is no data quality filter available for warehousing data. After the analysis of the core problem, the objectives of the solution have been determined together with the experts of the Bricklog. This answers the second research question:

RQ2: What are the objectives of the solution?

The research objective is to design a generic tool in Power BI that can assess the data quality of warehousing data. This tool must assess the data quality of mutations, stock and locations.

3. Systematic Literature Review

In chapter 3 the systematic literature review has been carried out to gather information about data quality frameworks. In section 3.1 the knowledge question is introduced. From this knowledge question, the research goal has been derived in section 3.2. In section 3.3 the key concepts for the SLR are determined. In section 3.4 the criteria for literature have been determined. In Section 3.5 the search terms are explained. In section 3.6 the search log is shown. In section 3.7 the different Theories and studies are explained. And in section 3.8 the different studies are integrated into the scope of Bricklog and warehousing.

3.1 Knowledge Question

To answer the knowledge problems and design a solution for the problem a sufficient theoretical background is needed. This theoretical background consists of selecting the theoretical perspective and discussing relevant theories that are available for assessing data quality. A systematic Literature Review(SLR) has been used to identify the theories that are most relevant to this research.

In the following section, the model for this research is explained. This model will explain the main variables together with a data quality assessment theory. This will result in a theoretical model which is used for the development of the data quality assessment tool.

Before the theoretical model can be created a data quality assessment theory or framework is needed. To find such a theory a literature study has been performed with the use of a SLR with the following research question.

"What theories or models are available for data quality assessment?"

To broaden the scope of this research warehousing data is left out of this research. However, the articles are selected if they apply to this research.

3.2 Research Goal

The goal of this research is to find a theory, model or framework for data quality assessment in warehousing data. If this consists the theoretical model can be on this theory. If it does not exist this gap should be filled and a theoretical model should be created with the use of existing theories. But before that could be done an overview of existing theories with assessment based on usefulness is needed.

3.3 Key Concepts

Determining key concepts in the research question is the first step in the SLR. These concepts are:

- Theories or models. This has been chosen as a key concept since a theory or model is needed for the research
- Data quality. This is a key concept because this is the concept we want to measure
- Assessment. This is a key concept because the goal is to find a way to measure the data quality

3.4 Criteria

Nr.	Inclusion Criteria	Reason
1	Include literature which revolves around data quality assessment, measurement or framework.	The goal of this research is to find a data quality measurement or framework.

2	Include literature which revolves around data quality and warehouses	If there is information already written about data quality in warehousing it would probably be very useful information
	Exclusion Criteria	**Reason**
1	Articles solely on data warehousing	Data warehousing is a whole other subject than warehouse data. So this will not include useful information
2	Articles before 1990	Articles earlier than 1990 will most likely be outdated and not contain useful information

Table 2 - Criteria

3.5 Search Terms

Before creating the search strings it is important to provide multiple terms for the key constructs. These terms are divided into related, broader and narrower terms. With these terms divided it would save time during the literature search to broaden or narrow the search string to get the desired amount of articles.

Constructs	**Related Terms**	**Broader Terms**	**Narrower Terms**
Theories	“Framework”	“methodology” “model*”	
Data quality	“quality dimension”		“quality factors” “quality framework” “quality challenges”
Assessment	“calculat*” “measur*”	“improvement”	“framework”

Table 3 - Search Terms

3.6 Search Log

Table 4 will show the search log of the SLR

Date	**Database**	**Search String**	**Number of hits**
27/10/2022	Scopus	(Theor* OR model* OR Framework?) AND(“Data quality” OR “data quality dimension”) AND (Assessment? OR calculat* OR measur*)	5863 results. Search string is too broad so no article selected.
27/10/2022	Scopus	((theor* OR model* OR framework?) AND ("Data quality assessment" OR "data quality dimension") AND (assessment? OR calculat* OR measur*))	340 results. 8 selected after ordering on both relevance and Cited by (most)
27/10/2022	Scopus	framework AND ("Data quality assessment" OR "data quality measurement")	319 results. 5 selected

27/10/2022	Scopus	framework AND ("Data quality assessment" OR "data quality measurement") AND Warehouse	7 results. 1 selected 1 article selected by drill down.
31/10/2021	EBSCO	(theory or model or framework) AND (data quality) AND (assessment or measurement)	991 results. 1 selected
31/10/2021	EBSCO	Data quality dimension OR data quality measurement	335 results. 5 selected
31/10/2021	EBSCO	Total selected	21
31/10/2021	EBSCO	Removed duplicates	2
07/11/2021		Removed for exclusion criteria	1
07/11/2021		Removed after complete reading	12
07/10/2021		Total selected for review	6

Table 4 - Search Log

Year	Author	Document Type	Data quality dimensions	Data quality assessment	Data quality measurement	Description key findings
2013	Laura Sebastian-Coleman	Book	x	x		An in-depth explanation of data quality dimensions and how to use them but does not include measurement of the data quality dimensions(Sebastian-Coleman, 2013)
2019	Faisal Saeed et al.	Journal Article and review	x	x		Review of problems in big data. Explained with data quality dimensions and an assessment process but no measurement(Salih et al., 2019)
2016	Antonio Vetro et al.	Journal Article and review	x	x	x	A review of open government data explains all the necessary data quality dimensions and why they are needed. From there explains how to assess the dimensions with corresponding measurements. But it is open data so not 1 on 1 applicable but does give clear insights for the project. (Vetrò et al., 2016)
2003	Jack E. Olson	Book				Old book. Gave clear insights into the problems but were not up to date anymore and should have put the exclusion criteria on newer books(Olson, 2003)
2018	Natasha, Mimic et al.	Journal Article	x	x	x	explained data quality dimensions but not very in-depth as well as data quality assessment and from there also included how to calculate the mentioned dimensions (Micic et al., 2018).
2021	Kashif Ali & Satirenjit Kaur Johl	Journal Article and review				Introduced TQM total quality management a data quality framework. For this research, it is not possible to totally change a process so a framework this will be difficult to implement has interesting topics. (Ali & Johl, 2021)

Table 5 - Key Findings

3.7 Integration of Theory

The aim of this literature research was to gain insights into available models or theories for data quality assessment. Different theories exist in the literature regarding data quality assessment. However, much of the current literature on data quality pays particular attention to data quality dimensions and data quality measurements. “the word dimension is used to identify aspects of data that can be measured and through which data’s quality can be described and quantified”(Sebastian-Coleman, 2013). Data quality measurement represents the actual calculation of the data quality dimension.

The studies outlined in the table below provide multiple data quality dimensions. Each study includes multiple data quality dimensions, however, the most common and relevant ones that emerged consistently across all studies were found to be the metrics presented in Table 6. These dimensions were further validated with the assistance of experts from Bricklog, ensuring the practical usefulness of the data quality dimensions.

Data quality Dimension	Definition	Measurement	Reference
Completeness	Degree to which there is no missing or insufficient data		(Salih et al., 2019)
	implies having all the necessary or appropriate parts; being entire, finished, total.		(Sebastian-Coleman, 2013)
	Indicates the percentage of complete cells in a dataset. It means the cells that are not empty and have a meaningful value assigned (i.e. a value coherent with the domain of the column)	$(1 - \frac{\text{Number of incomplete cells}}{\text{Number of cells}}) * 100$	(Vetrò et al., 2016)
	Indicates the percentage of complete rows in a dataset. It means the rows that don't have any incomplete cell	$(1 - \frac{\text{Number of incomplete rows}}{\text{Number of rows}}) * 100$	(Vetrò et al., 2016)
	describes if all required rows, based on frequency and start and end time, are accounted for in the output file.	$\frac{\text{Total Non Missing Values}}{\text{Number of Expected Values}}$	(Micic et al., 2018)
Consistency	Consistency can be regarded as the absence of variety or change.		(Sebastian-Coleman, 2013)
	Describes the structure of the values in the data	$\frac{\text{structural consistent values}}{\text{Total Values}}$	(Micic et al., 2018)
Accuracy	Indicates the percentage cells in a dataset that has correct values according to the domain and the type of information of the dataset.	$(1 - \frac{\text{Number of cells with errors}}{\text{Number of cells}}) * 100$	(Vetrò et al., 2016)
Uniqueness	is defined by time records not being duplicated	$\frac{\text{number of unique Rows}}{\text{total Rows}}$	(Micic et al., 2018)

Table 6 - data quality dimensions

The average of these dimensions is the overall data quality.

$$data\ quality = \frac{\sum_{i=1}^{n} D_n}{\text{Total number of Dimensions}}\ where\ D_n = \ \ data\ quality\ dimensions$$

3.8 Key Variables

The key variables for this research are the 4 dimensions determined in the previous section which are: completeness, accuracy, consistency and uniqueness. For completeness, there is no arguing if a value is missing or not but for accuracy, it is important to know when a value is incorrect. However, the criteria if a value is incorrect will be different for every customer of Bricklog. So, it is important to work with thresholds. The best way to implement these thresholds is in the Power BI Dashboard so they can be easily changed.

In the previous section, it has been explained how to calculate the 4 data quality dimensions but since this research is only focused on warehousing data it is important to limit the scope. These key variables can be applied to mutations, locations and stock.

3.9 Conclusion

The knowledge question is answered by the SLR, which is part of the second phase of the DSRM.

"KQ: What theories or models are available for data quality assessment to warehousing data within the scope of Bricklog?"

From the literature, six articles have been selected for review. Together these studies provide important insights into data quality assessment. These studies introduced data quality dimensions and data quality measurements. Four data quality dimensions have been selected which will be used for the assessment of the data quality for warehousing data. These data quality dimensions are Completeness, Uniqueness, Consistency and Accuracy. These data quality dimensions were not selected from one framework but these dimensions were mentioned in every framework about data quality. Due to the scope of Bricklog, it was not possible to apply a particular framework since every framework is focused on the whole process of data and Bricklog is for this research only focused on the assessment of data quality and not the improvement or a change in the process.

4. Methodology

In chapter 4 the Design Science Research Methodology is explained in more detail and applied to this research. In Section 4.1 the Problem identification is described. In section 4.2 the objectives of the solution are described. In section 4.3 the design and development phase is explained. In Section 4.4 the demonstration phase is explained. In Section 4.5 the Evaluation phase is explained. In section 4.6 the Communication phase is explained.

The 6 phases of the Design Science Research Methodology (DSRM) are explained in more detail in the following section as to what value it added to the study(De Sordi, 2021). The DSRM has a problem-centred initiation thus, the DSRM started with problem identification. This research performed one iteration of the DSRM.

4.1 Problem Identification and Motivation

The problem identification has already been introduced in the previous section of this project plan. The action problem of this bachelor assignment is the improvement of data quality of warehousing data. If the data quality of warehousing databases can be improved with the use of the artefact, companies can actually use their data to improve their warehousing strategies and decision-making.

4.2 Define the Objectives of a Solution

The main objective of the research is to help Bricklog help their customers improve the data quality of warehousing data. The solution proposed is to design a data quality assessment tool in Power BI. The objective of this artefact is to show where the mistakes are in the dataset. With this tool, the customers of Bricklog can see where their mistakes happen, solve them and prevent them in the future. Currently, there is no knowledge of what is considered a mistake in data and what theories are available for data quality assessment. To solve this a Systematic Literature Review(SLR) has been carried out to answer the following question:

"What theories or models are available for data quality assessment?"

4.3 Design and Development

In this phase, the design and development of the artefact is conducted. The results of the SLR are used for the calculation of the KPIs of the dashboard. With the use of these metrics, the data quality of mutations, stock and locations can be determined. During this phase, it is important to keep the dashboard simple and easily useable and use the design style of Bricklog.

4.4 Demonstration

In this phase, the artefact is demonstrated. During this phase, it is important to know to whom the demonstration is meant. For employees of Bricklog demonstrating the artefact a more technical side can be highlighted. But if the demonstration is meant for a Warehouse manager, for example, a less technical side will be highlighted but a more in-depth explanation about data quality will be needed.

4.5. Evaluation

In this phase, the artefact is evaluated. This phase is based on the previous phase. It is important to evaluate if the artefact actually solves the problem. This is done by comparing the objectives of the solution to the results of the demonstration. This artefact is evaluated with the use of a questionnaire. The questions of the questionnaire are based on the Technology Acceptance Model(TAM) (Chuttur, 2009). The questions are divided into three categories; perceived usefulness, perceived ease of use and general. Each category has 2 questions except for the general questions

which consist of 3 questions. Due to time limits, this artefact is not evaluated but it is validated by an expert panel instead.

4.6 Communication

The last phase of the DSRM includes communication with researchers and other relevant audiences. The first part of the communication phase will be this thesis which will follow the structure of the DSRM. With this thesis, other researchers could take this knowledge and build upon it. The second part of the communication phase will be the colloquium given to a relevant audience.

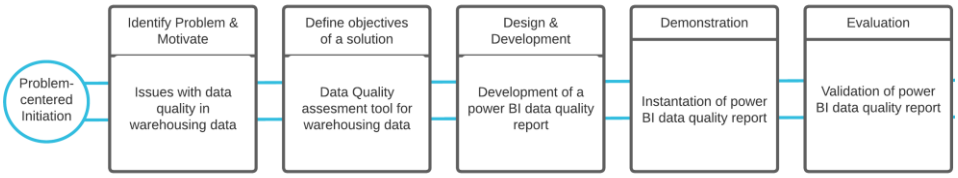

Figure 6 - Iteration of the DSRM

5. Development of the artefact

In chapter 5 the steps and choices are explained during the development of the dashboard. In section 5.1 the dataset is introduced. In section 5.2 the static dataset is explained. In section 5.3 the dynamic dataset is explained. In section 5.4 the method for data gathering is explained. In section 5.5 the data quality dimensions are applied to the development of the dashboard. In section 5.6 the first data quality page is explained and in section 5.7 the second data quality page is explained. In section 5.8 the first dynamic data quality page is introduced and in section 5.9 the second dynamic data quality page.

5.1 Data model

Before the development of the dashboard in Power BI, a data model and a dataset is needed, since Bricklog currently lacks a data model and dataset. To create a generic solution it is important to design a data model that includes a perspective from all different types of warehouses and all different types of data storages. This research does not include information about databases since every customer of Bricklog has a different type of database. This research does not entail database information. Instead, a generic data model has been designed together with experts of Bricklog. This data model is split into a dynamic dataset and a static dataset. The dynamic dataset covers all the dynamic events. An example of such an event can be the movement of a pallet from Location A to Location B. The static dataset covers the events that do not have a movement, for instance, the location of the pallet at the end of the day. An important argument to split the dataset is the refresh rate of the different datasets. The static dataset will only be refreshed at the end of the day and the dynamic dataset can be refreshed more often to increase the accuracy of the data during the day. Furthermore, this distinction between datasets will also increase the performance of the dashboard.

5.4 Data Gathering

For this research, synthetic data is used since there is not a customer with warehousing data for Bricklog yet. Synthetic data is crucial for building a Power BI dashboard from scratch. It not only helps in designing the dashboard but also allows for creating a demo to validate its performance and features. This synthetic data has been generated by an expert of Bricklog. This synthetic data has a limited amount of data generated for the stock history, mutations and essential reference tables. Because this data is generated there are limited errors in the data which might limit the demonstration of the data quality dashboards since there are limited data quality errors. For future use of the tool data of actual customers is used. However, Not every company uses the same Warehousing Management System(WMS) thus every company records their data differently and with these two datasets it is able to load data from any kind of WMS. This is done on the Azure Data Factory(ADF) of Bricklog and the data engineers of Bricklog will connect the different kinds of WMS to the ADF and from there the same Data model is used so this generic solution is applicable to any customer of Bricklog with a warehouse.

5.5 Dashboard

In the following section, the design of the 2 Power BI dashboards will be explained. Similar to the dataset, there is a distinction between static and dynamic for the dashboards. There is 1 dashboard for the dynamic side of warehousing data and 1 dashboard for the static side of warehousing data. Each dashboard consists of 3 pages, the first 2 pages display the data quality errors and checks while the third page serves as an export feature, enabling users to export data quality errors to Excel. With these 2 dashboards, it will be possible to assess the data quality of locations, mutations and stock which are part of the research objective. The dashboard is built in Power BI as stated before because this is the required tool from Bricklog to build the dashboard in.

For both sides, the same approach is used. The first report will look at the completeness of the data. And the second report will go more in-depth at data quality errors while also providing some insights next to only showing data quality errors. These insights are there to give warehouse managers an incentive to look at the dashboard instead of only acknowledging that their data is not of good quality.

From the theory, 4 data quality dimensions have been derived which can be found in table 6.

$$completeness = \frac{\text{missing values}}{\text{Number of total values}}$$

$$accuracy = \frac{\text{number of correct values}}{\text{Number of total values}}$$

$$consistency = \frac{\text{structural consistent values}}{\text{Total Values}}$$

$$uniqueness = \frac{\text{number of unique values}}{\text{total values}}$$

However, when loading data into Power BI there is already a filter on data consistency such that the data quality dimension will not be used in the dashboard. The other 3 data quality dimensions have to be rewritten to work properly Power BI dashboard. Power BI works with data entry on a row level. Therefore, one data entry is a row which contains data and Bricklog wants the data quality in a percentage thus that also changes the formulas.

The possibility exists that a data entry misses multiple values which could lead to a negative value of completeness. Thus the following formula has been used to determine the completeness of the dataset.

$$completeness = 1 - \frac{\text{rows with 1 or multiple missing values}}{\text{Number of total rows}}$$

The accuracy uses the same logic as completeness to prevent negative values.

$$accuracy = 1 - \frac{\text{rows with 1 or multiple incorrect values}}{\text{Number of total rows}}$$

Uniqueness will only be changed into a percentage so that will become:

$$uniqueness = 1 - \frac{\text{number of unique values}}{\text{total values}}$$

5.6 Static Data Quality Sheet 1

For the static data quality sheets, the main focus will be the stock history. The first part will be the check on completeness and accuracy. This will be done on Customer, Unit and Amount. The user of the dashboard can filter some settings which include which hall, kind of storage, date filter and a customizable threshold for the amount of products on a pallet, to their own preference and can close the filter settings to use more of the screen. Furthermore, the overall Data quality can be seen in the left top corner and the number of Pallets in this time frame. Next to this is the amount of data quality errors per week. For this Data quality sheet completeness and accuracy are used thus, the DQ calculation used is:

$$Data\ Quality = \frac{Completeness + Accuracy}{2}$$

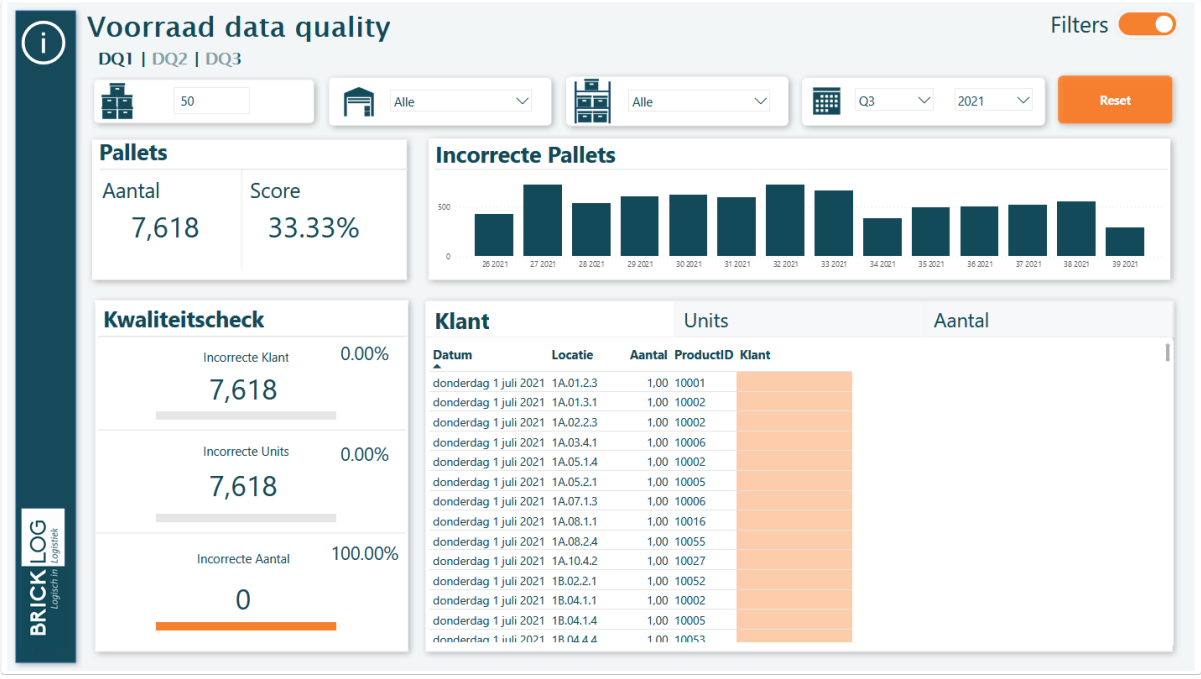

Figure 7 - Static Data Quality Sheet 1

5.7 Static Data Quality Sheet 2

For this second data quality sheet the focus is on locations. For now, the focus is on Bulk and rack locations. As mentioned in Chapter 2 locations have a maximum number of stock which is logical to store in a location. With the date filter, a day is selected to look thoroughly through the selected date. There are 3 buttons which filter the bulk, rack and missing locations. For the bulk locations, the threshold amount of pallets in the same location can be changed in the filters.
The Rack Locations button shows a table with every rack location with more than 1 stock number. The missing locations button shows every pallet stored without a location or with a location used which is not in the locations table.

The DQ calculation used is:

$$Data\ Quality = 1 - \frac{\text{bulk+rack locations with more than treshold}}{\text{Number of Locations with data}}$$

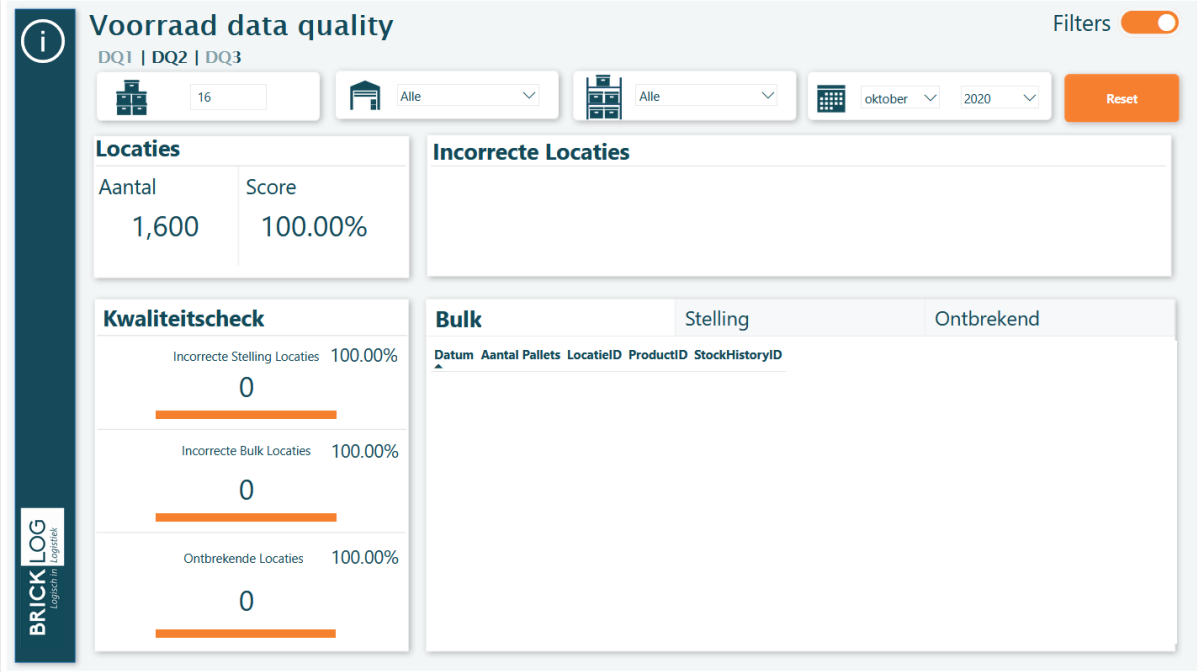

Figure 8 - Static Data Quality Sheet 2

5.8 Dynamic Data Quality Sheet 1

Just like the first static data quality sheet, the first check is on completeness and accuracy. This time the check is on Location, Unit and Customer. There are 5 buttons which can be used to filter on the incorrect start location, end location, start and end location, unit and customer. The user of the dashboard can filter the dashboard on type of mutation, hall, type of storage and date and can hide the filters to create more space on the screen. In the top left corner the amount of mutations and the Data quality score is shown. Next, the amount of incorrect mutations per week is shown.

Furthermore, for this data quality sheet completeness and accuracy have been used so that the data quality calculation used is:

$$Data\ Quality = \frac{Completeness + Accuracy}{2}$$

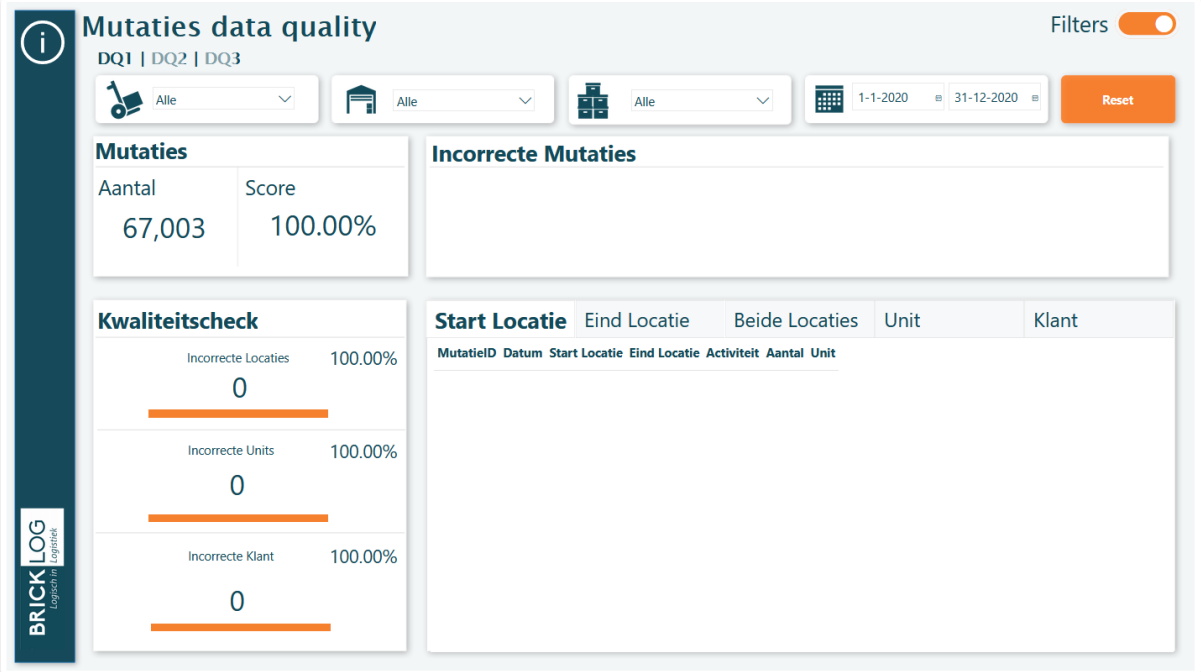

Figure 9 - Dynamic Data Quality Sheet 1

5.9 Dynamic Data Quality Sheet 2

For this data quality sheet, there is no overall data quality score. This sheet is pallet-focused and every pallet has a DQ score. On the left a table is shown with PalletID and the amount of mutations for the specific pallet and DQ score. In the settings the threshold for the number of mutations can be changed now it is at 6 so every pallet with 6 or more mutations is shown. Furthermore, in the settings the maximum amount of products on a pallet can be changed and the Hall, type of storage and date.

If the user clicks on a palletID on the left the user can use the table on the right to filter the mutations for this Pallet on incorrect start location, end location, amount of products, Unit and Customer.

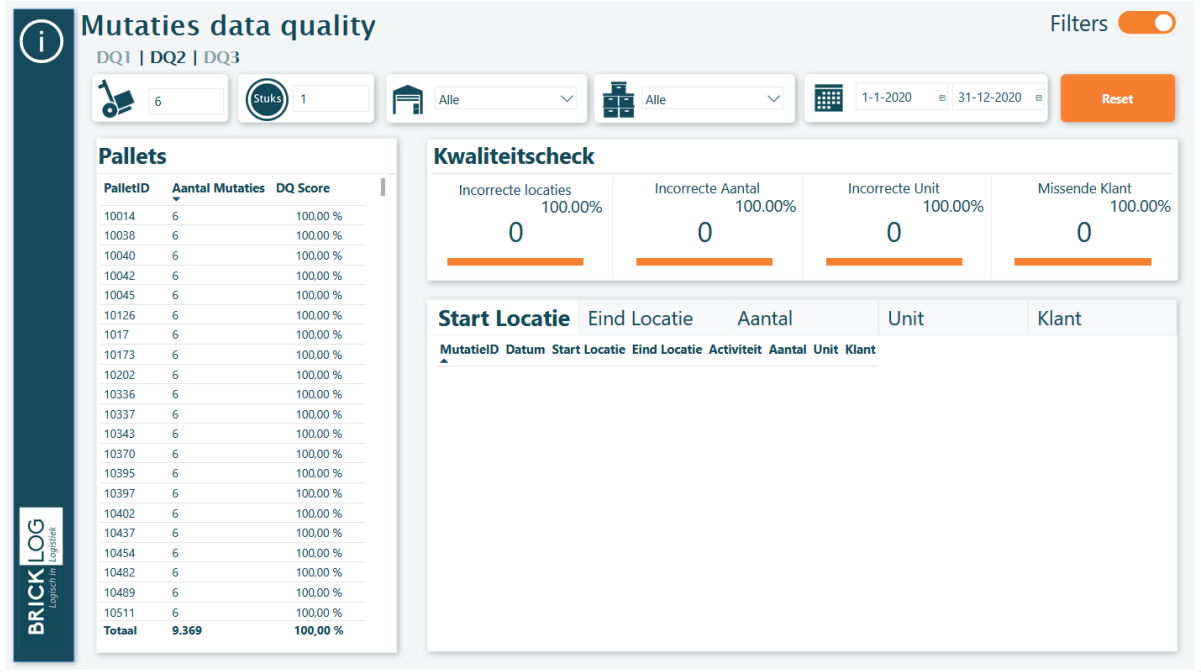

Figure 10 - Dynamic data quality sheet 2

5.10 Extra Features

To increase the usability experience of the data quality assessment tool extra features are added. First of all every sheet of the tool has an information button which when pressed shows information about that page. It explains the different filters which can be changed and what the numbers on the page mean.

And there is a third page for the static and dynamic. This third page is an export page which can be used to export all the mistakes in to an Excel file. This export can be used to easily trace the mistakes in the dataset and solve them.

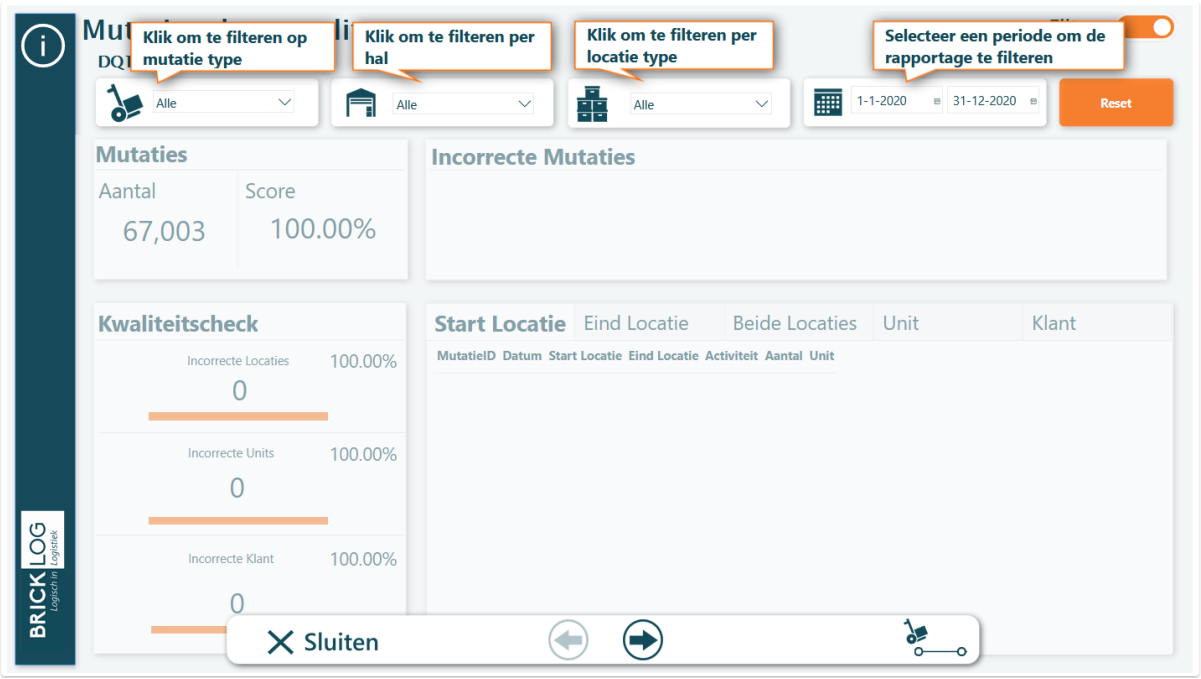

Figure 11 - Information shown on the page

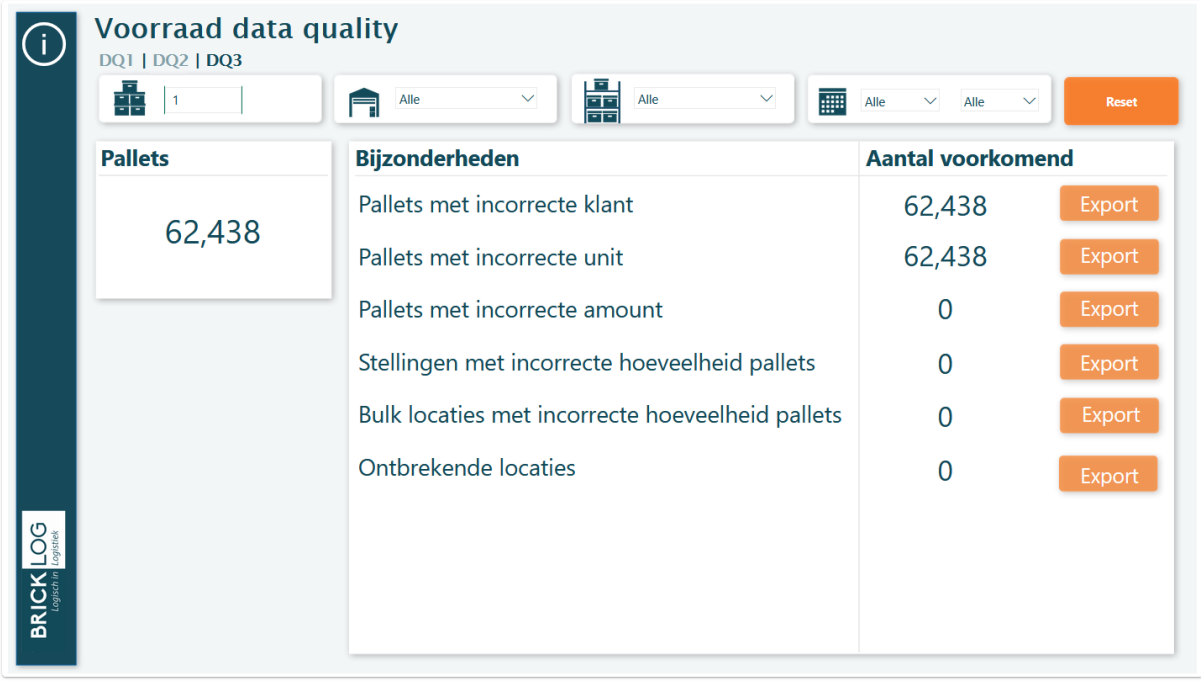

Figure 12 - Export page

5.11 Usage of the Artefact

To effectively use the data quality assessment tool Bricklog and the customer should work together. The data quality assessment tool should be used to manage policy regarding data quality. After Bricklog has collected the data from the WMS to their data factory the data quality assessment tool should be used to assess the data quality. Accordingly, it should be used to check if it is due to human mistakes with activities in the warehousing or errors from the usage of the different systems this should result in a change in or the creation of policies regarding data quality.

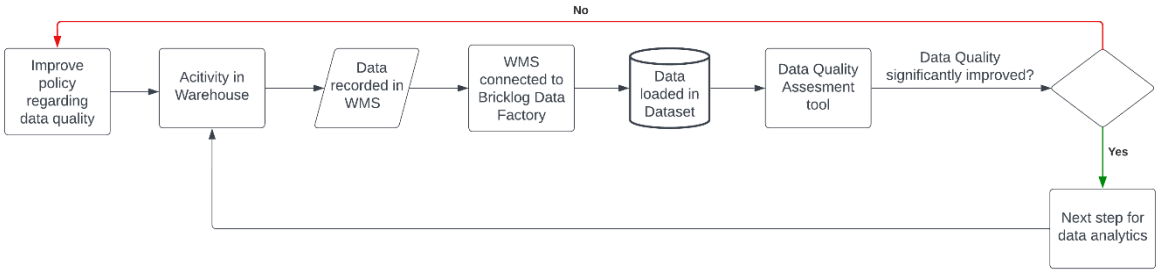

Figure 13 - Simplified process

5.12 Conclusion

In this chapter, the design and development of the data quality assessment tool has been done. The first step was to design a dataset for the static and dynamic parts of the warehouse. This has been done together with the experts of Bricklog. From here 2 dashboards were developed each with 2 pages. The design of these dashboards has been done with the data quality dimensions of completeness, accuracy and uniqueness. This answers the third research question

"RQ3: what metrics from data quality assessment frameworks are relevant for data quality assessment tool?"

6. Validation

In Chapter 6 the validation process is addressed which is the fifth phase of the DSRM. In Section 6.1 the validation process is explained with the use of the TAM model. In Section 6.2 the validation results from the survey are determined.

6.1 Validation Process

"The goal of validation is to predict how an artefact will interact with its context, without actually observing an implemented artefact in a real-world context" (Peffers et al., 2007) This artefact has not been used in a real-world context due to time limits so it is important to validate this dashboard. One of the options to validate this research is through an expert panel. *"The design of an artefact is submitted to a panel of experts, who imagine how such an artefact will interact with problem contexts imagined by them and then predict what effects they think this would have."* (Peffers et al., 2007)

To validate the dashboard a questionnaire has been formulated. The questions are based on the Technology Acceptance Model. (TAM)

Perceived Usefulness: *The degree to which an individual believes that using a particular system would enhance his or her job performance.* (Chuttur, 2009)

Perceived ease of use: *The degree to which an individual believes that using a particular system would be free of physical and mental effort.* (Chuttur, 2009)

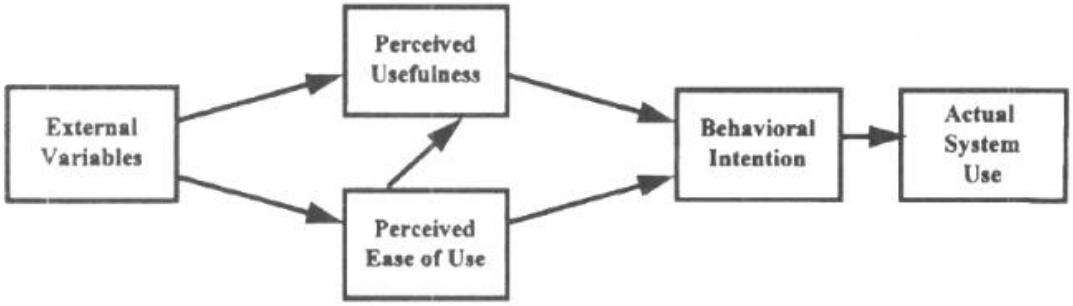

Figure 14 - Technology Acceptance Model

3 general multiple-choice questions about the dashboard were formulated. After the general questions, 4 questions related to TAM were formulated. After these 7 multiple choice questions 2 open questions were formulated to create the possibility to give feedback on the dashboard. The multiple-choice questions were formulated based on the Likert scale. Which is a 5-point scale from strongly disagree to strongly agree with a particular statement.

Questions		Answers			
Name					
Position					
How long have you been working with dashboard design and development?					
Date					
Likert Scale	Strongly disagree	Disagree	Neutral	Agree	Strongly Agree
General					
How well-organized is the dashboard in terms of making it easy to find the information users need?					
How easy are the data visualizations to understand and interpret?					
Does the dashboard provide users with the ability to customize the information displayed to meet their specific needs?					
Perceived Ease of Use					
Overall, how easy is it to use the dashboard					
Do you find the dashboard's user interface intuitive and easy to navigate?					
Perceived Usefulness					
How satisfied are you with the overall usefulness of the dashboard?					
To what extent do you believe that the dashboard is useful for future customers of Bricklog?					
General Feedback					
Are there any additional features or functionality that users would find useful?					
Are there any errors or bugs in the dashboard that need to be addressed?					

Table 7 - Validation survey

6.2 Validation Results

Here, the validation results of the dashboard are presented. The results are described in terms of the technology acceptance model which is used to validate the dashboard.

This dashboard is part of the product development process of Bricklog. Right now, Bricklog mainly serves transport companies and does not have any customers in the warehousing sector yet. Therefore, there is no customer to validate the results of the dashboard. Instead, the dashboard is validated by a panel of experts from Bricklog. It was presented to four representatives from Bricklog who evaluated the dashboard and filled in the questionnaire. This panel consists of four experts which have the most experience with dashboard design in Bricklog and thus were the most relevant for this study. The decision was made to also prioritize expertise over a larger sample size, although it could have been possible to use a sample size larger than four.

On the other hand, the limited number of experts in this study can undermine its validity for a use in a broader context. With such a small sample size, there is a risk that the dashboard may not perform as anticipated in real-world scenarios. Additionally, bias could to some extend skew the results, as all panel experts are accustomed to working with dashboards. Moreover, important insights about the dashboard might be overlooked due to the limited range of expert opinions.

	General	Perceived ease of use	Perceived usefulness
Respondent	4	4,5	4,5
Respondent	3,67	4,5	5
Respondent	4,33	4,5	4
Respondent	4,33	4	4,5
Average	**4,1**	**4,4**	**4,5**

Table 8 - Validation results

The dashboard is validated on general use, perceived ease of use and perceived usefulness. In the following sections, the results will be discussed for each subject.

General the score of the general questions is 4,1 out of 5. The score is derived from questions 1, 2 and 3. The dashboard is overall easy to navigate and most data visualizations are easy to interpret. However, some data visualizations could cause some confusion for customers. Furthermore, customization of the dashboard is not possible except for editing the thresholds of some values.

Perceived ease of use The score for the perceived ease of use is 4,4 out of 5. The score is derived from questions 4 and 5. The dashboard is easy to use and intuitive to navigate. The information overlay page increases the usability of the dashboard and is considered a substantial benefit. However, this could improve more with additional functionalities such as buttons which get a glow when the cursor hovers over the button.

Perceived usefulness The score for perceived usefulness is 4,5 out of 5. The score is derived from questions 6 and 7. The dashboard will probably be useful for new customers of Bricklog. The dashboard will help customers increase their data quality and this will help them to gain accurate insights from the data from their warehouse. This could be improved with more storytelling to help customers gain more useful information from the dashboard.

Overall, the dashboard was accepted by the expert panel with an overall score of 4,3 out of 5. The dashboard is perceived as useful and will probably be used by future customers of Bricklog. There were some general comments for additional features which can be implemented in the future and 1 bug occurred but that was solved right away.

6.3 Conclusion

In this chapter, the dashboard has been validated. This was done by a survey for an expert panel consisting of four representatives of Bricklog. To answer the 4th research question, a survey was developed based on the Technology Acceptance Model. The average score is 4,3 out of 5 thus, the tool is expected to be useable as intended in practice, however, this result could be biased and unreliable due to the small sample size of only 4 experts.

RQ4: Can the developed data quality assessment tool be used in practice?

7. Conclusion and recommendations

In chapter 7 the results of this research are determined and future research recommendations are outlined. In section 7.1 a summary of the results of the research questions is provided. In section 7.2 the scientific contribution of this research is explained. In section 7.3 the limitations of this research are determined. And in section 7.4 future research recommendations are provided.

7.1 Conclusion

The main research question of this thesis is: *"MRQ: Can a generic approach to assess the data quality of warehousing data with the use of a data quality assessment instrument be designed for Bricklog to help their customers improve their data?"*

This research question has been answered by the design of a data quality assessment tool in Power BI and the validation of this dashboard. The tool can be used to measure the data quality of mutations, locations and stock. The dashboard is built upon a generic data model which allows it to be used by multiple future customers with ease. With the use of thresholds, these customers can use the tool according to their preferences.

The dashboard has been validated by a panel of experts that confirmed its usefulness and useability. This validation suggests that it can be used for future customers. However, the dashboard has only been internally validated which does not guarantee that it will be used in the future. Furthermore, there are limitations discussed further in this study which allows for opportunities for future research.

7.2 Research Questions

Here, a summary of the findings of the research questions is provided.

RQ1: What is the core problem which leads to data quality errors in warehousing data?

Multiple causes for data quality errors exist. First of all, staff shortage could lead to high workload which could to mistakes during work and could lead to data quality errors. Moreover, the interaction between different systems could result in data quality errors, such as inconsistencies arising from storing data in varying formats. However, investigating each system for every customer of Bricklog is a costly and time-consuming process. Bricklog requested a generic solution therefore the core problem should be generic as well which is that none of the customers has a data quality filter. Thus, the core problem which leads to data quality errors in warehousing data is that there exists no data quality filter for warehousing data.

RQ2: What are the objectives of the solution?

The solution should be generic and Bricklog should be able to help their customers improve their data quality with the use of the solution. The solution should be provided through instruments that are compatible with the company products, technologies and infrastructures, e.g. a power BI dashboard.

KQ: What theories or models are available for data quality assessment to warehousing data within the scope of Bricklog?

To answer this knowledge question a Systematic Literature Review (SLR) has been carried out. The goal was to find an existing data quality theory or framework which applies to warehousing data within the scope of Bricklog. Unfortunately, there was no data quality framework immediately applicable within the scope of Bricklog. Every framework was focused on the whole process instead

of only assessing the data quality. However, these data quality frameworks did provide data quality dimensions and the 4 most important dimensions have been used to asses data quality. These dimensions were determined with the assistance of experts from Bricklog to ensure the practical usefulness of the data quality dimensions. These were completeness, accuracy, consistency and uniqueness.

RQ3: What metrics from data quality assessment frameworks are relevant for the data quality assessment tool?

The 4 dimensions from the data quality assessment frameworks were used in the development of the dashboard. These dimensions have been used to assess the data quality of the most important subjects of warehousing which are mutations, locations and stock, these subjects were predefined by Bricklog. Together these have led to 2 dashboards with each 2 pages within the dashboard.

RQ4: Can the developed data quality assessment tool be used in practice?

To answer this question the dashboard has been internally validated by an expert panel consisting of 4 representatives of Bricklog. Overall, the dashboard has been accepted by the expert panel but, there are possibilities for improvements to make the dashboard easier to use. However, the acceptance of the expert panel does not guarantee that the dashboard will be used since it is only internally validated with a limited sample size. The reason behind the size is the prioritization of expertise most relevant for the outcomes of this study over a larger sample size. To know if the dashboard will be accepted the recommendation is to implement it at a warehouse customer of Bricklog and validate it externally.

7.3 Contribution to Knowledge

In the introduction of this thesis, a knowledge gap was identified. During the systematic literature review, it was identified that multiple data quality frameworks exist, however, none of the frameworks solely prioritize data quality assessment; rather, they entail the entire process. This study developed a practical solution to data quality assessment. This study evaluated aspects of the practical use of data quality dimensions for example with the use of data quality dimensions the data quality of mutations inside a warehouse can be assessed. This assessment can be applied to different sectors and other datasets and not only to warehousing data.

7.4 Limitations

The main limitation of this study is the external validity. This research was conducted as Research and Development therefore, there was no customer to validate the usefulness of the data quality assessment tool. It is a realistic possibility that the results are biased based on the expectations of Bricklog which could differ from the reality. Another limitation is that the dashboard is only validated by experts who have been working within dashboard development for multiple years. This may lead to bias in the results for the perceived ease of use.

The next limitation of this study is the questionnaire. The questionnaire is based on the Technology Acceptance Model (TAM). The categories of the questions are divided into 3 categories; general, perceived usefulness and perceived ease of use. These categories validate the perceived ease of use and perceived usefulness of the dashboard however, they do not validate the objectives of this study. This could mean that this study does not meet the requirements determined by Bricklog. However, together with the experts of Bricklog, it has been decided that a data quality assessment tool is the output of this study thus, if the dashboard is perceived useful the output of this study will also be useful. The panel of experts from Bricklog found the dashboard to be perceived as useful,

indirectly validating the objectives. However, direct validation of these objectives with refined metrics is recommended as a future research direction.

Another limitation is that this tool does not guarantee perfect data quality after using the tool extensively. It assesses the most important subjects of warehousing but not every column from the dataset is assessed, because of the size of the dataset, it is impractical to assess each column individually. Furthermore, text values for columns such as 'UnitID' are only checked on completeness it was not possible to check if the filled value is correct.

7.5 Future Research

The limitations above provide possibilities for future research. First of all, for a future study, it is possible to implement this tool at a warehouse. With this study causes for data quality errors can be determined and solutions can be provided. This study could show that the tool is useful in practice. This study will probably be more focused on change management to help the warehouse staff understand their data mistakes and how to correct them, rather than just on the technical aspects.

Another future study could expand the data quality assessment tool to assess and determine more data quality errors for example in the reference tables or provide a solution to determine if a text value is correct. Further expansion could be an addition of a prediction if a value is missing. This prediction could prove helpful in reducing the need for manual labour. However, this is only useful when the causes of the data quality errors are already determined. Because solving the problem with a prediction is not sustainable compared to solving the problem at the root of the problem.

If data quality is of high quality warehouses could look into advanced tools to improve the efficiency of their warehouse. Future research can create a 3D model of the warehouse and see how this can be used to improve efficiency. This could be expanded with the use of simulation.

This study highlighted the practical implications of data quality frameworks and developed a data quality assessment tool. It laid the groundwork for a standardized approach to evaluating the accuracy and consistency of data. This research can be further explored to develop a generic data quality assessment framework applicable across various industries and applications.

Bibliography

Ali, K., & Johl, S. K. (2021). Soft and hard TQM practices: future research agenda for industry 4.0. *Https://Doi-Org.Ezproxy2.Utwente.Nl/10.1080/14783363.2021.1985448*, *33*(13–14), 1625–1655. https://doi.org/10.1080/14783363.2021.1985448

Andiyappillai, N. (2020). ISSN : 2249-0868 Foundation of Computer Science FCS. *International Journal of Applied Information Systems (IJAIS)*, *12*(35). www.ijais.org

Chuttur, M. (2009). Overview of the Technology Acceptance Model: Origins, Developments and Future Directions. *All Sprouts Content*, *9*(37). https://aisel.aisnet.org/sprouts_all/290

De Sordi, J. O. (2021). Design Science Research Method. *Design Science Research Methodology*, 59–78. https://doi.org/10.1007/978-3-030-82156-2_5

Heerkens, H., & van Winden, A. (2016). *Solving Managerial Problems Systematically 1 e edition* (1st ed.). Noordhoff.

Micic, N., Neagu, D., Campean, F., & Zadeh, E. H. (2018). Towards a Data Quality Framework for Heterogeneous Data. *Proceedings - 2017 IEEE International Conference on Internet of Things, IEEE Green Computing and Communications, IEEE Cyber, Physical and Social Computing, IEEE Smart Data, IThings-GreenCom-CPSCom-SmartData 2017*, *2018-January*, 155–162. https://doi.org/10.1109/ITHINGS-GREENCOM-CPSCOM-SMARTDATA.2017.28

Olson, J. E. (2003). Data Quality: The Accuracy Dimension. *Data Quality: The Accuracy Dimension*, 1–293. https://doi.org/10.1016/B978-1-55860-891-7.X5000-8

Pallet Afmetingen | Europallets, blokpallets, etc. | Palletcentrale. (n.d.). Retrieved February 21, 2023, from https://palletcentrale.nl/kennisbank/pallet-formaten-welke-afmetingen-hebben-pallets/

Pallet Definition & Meaning - Merriam-Webster. (n.d.). Retrieved February 21, 2023, from https://www.merriam-webster.com/dictionary/pallet

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, *24*(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Provost, F., & Fawcett, T. (2013). Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Big Data*, *1*(1), 51–59. https://doi.org/10.1089/BIG.2013.1508/ASSET/IMAGES/LARGE/FIGURE1.JPEG

Salih, F. I., Ismail, S. A., Hamed, M. M., Mohd Yusop, O., Azmi, A., & Mohd Azmi, N. F. (2019). Data quality issues in big data: A review. *Advances in Intelligent Systems and Computing*, *843*, 105–116. https://doi.org/10.1007/978-3-319-99007-1_11

Sebastian-Coleman, L. (2013). Measuring data quality for ongoing improvement: A data quality assessment framework. *Measuring Data Quality for Ongoing Improvement: A Data Quality Assessment Framework*, 1–324. https://doi.org/10.1016/C2011-0-07321-0

Staudt, F. H., Alpan, G., di Mascolo, M., & Rodriguez, C. M. T. (2015). Warehouse performance measurement: a literature review. *Http://Dx.Doi.Org.Ezproxy2.Utwente.Nl/10.1080/00207543.2015.1030466*, *53*(18), 5524–5544. https://doi.org/10.1080/00207543.2015.1030466

The Digital Transformation of SMEs. (n.d.). https://doi.org/10.1787/bdb9256a-en

van Geest, M., Tekinerdogan, B., & Catal, C. (2021). Design of a reference architecture for developing smart warehouses in industry 4.0. *Computers in Industry*, *124*, 103343. https://doi.org/10.1016/J.COMPIND.2020.103343

Vetrò, A., Canova, L., Torchiano, M., Minotas, C. O., Iemma, R., & Morando, F. (2016). Open data quality measurement framework: Definition and application to Open Government Data. *Government Information Quarterly*, *33*(2), 325–337. https://doi.org/10.1016/J.GIQ.2016.02.001

Appendix A

Validation Survey Results

Questions		Answers			
Name **Respondent 1**					
Position **Lead BI**					
How long have you been working with dashboard design and development? **3-years**					
Date **7-7-23**					
Likert Scale	Strongly disagree	Disagree	Neutral	Agree	Strongly Agree
General					
How well-organized is the dashboard in terms of making it easy to find the information users need?				**x**	
How easy are the data visualizations to understand and interpret?					**x**
Does the dashboard provide users with the ability to customize the information displayed to meet their specific needs?				**x**	
Perceived Ease of Use					
Overall, how easy is it to use the dashboard				**x**	
Do you find the dashboard's user interface intuitive and easy to navigate?					**x**
Perceived Usefulness					
How satisfied are you with the overall usefulness of the dashboard?				**x**	
To what extent do you believe that the dashboard is useful for future customers of Bricklog?				**x**	
General Feedback					
Are there any additional features or functionality that users would find useful?	More storytelling and call-to-actions.				
Are there any errors or bugs in the dashboard that need to be addressed?	No				

Questions		Answers			
Name **Respondent 2**					
Position **Data Engineer/BI Specialist**					
How long have you been working with dashboard design and development? **2 years**					
Date **07/07/2023**					
Likert Scale	Strongly disagree	Disagree	Neutral	Agree	Strongly Agree
General					
How well-organized is the dashboard in terms of making it easy to find the information users need?				X, the 100% and the orange line graph can bring some confusion to a customer.	
How easy are the data visualizations to understand and interpret?				x	
Does the dashboard provide users with the ability to customize the information displayed to meet their specific needs?			X, then can customize filters. But the dashboard is meant for users to customize it like they want.		
Perceived Ease of Use					
Overall, how easy is it to use the dashboard				x	
Do you find the dashboard's user interface intuitive and easy to navigate?					x
Perceived Usefulness					
How satisfied are you with the overall usefulness of the dashboard?					x
To what extent do you believe that the dashboard is useful for future customers of Bricklog?					x
General Feedback					
Are there any additional features or functionality that users would find useful?	Every button in the dashboard should get a 'glow' function when you hover over them (i, DQ1, DQ2, DQ3). After				

	clicking the reset button, the filters that are cleaned should show the most common values
Are there any errors or bugs in the dashboard that need to be addressed?	The unit page shows an error.

Questions		Answers			
Name **Respondent 3**					
Position					
How long have you been working with dashboard design and development? **1,5 years**					
Date **4-7-2023**					
Likert Scale	Strongly disagree	Disagree	Neutral	Agree	Strongly Agree
General					
How well-organized is the dashboard in terms of making it easy to find the information users need?					x
How easy are the data visualizations to understand and interpret?				x	
Does the dashboard provide users with the ability to customize the information displayed to meet their specific needs?				x	
Perceived Ease of Use					
Overall, how easy is it to use the dashboard				x	
Do you find the dashboard's user interface intuitive and easy to navigate?				x	
Perceived Usefulness					
How satisfied are you with the overall usefulness of the dashboard?					x
To what extent do you believe that the dashboard is useful for future customers of Bricklog?				x	
General Feedback					
Are there any additional features or functionality that users would find useful?	Make sure that the tables are filled horizontally, because there are some blank spaces at the moment.				
Are there any errors or bugs in the dashboard that need to be addressed?	-				

Questions	Answers				
Name					
Position **Owner**					
How long have you been working with dashboard design and development? >**5 years**					
Date **05/07/2023**					
Likert Scale	Strongly disagree	Disagree	Neutral	Agree	Strongly Agree
General					
How well-organized is the dashboard in terms of making it easy to find the information users need?					X
How easy are the data visualizations to understand and interpret?				X	
Does the dashboard provide users with the ability to customize the information displayed to meet their specific needs?			X		
Perceived Ease of Use					
Overall, how easy is it to use the dashboard					X
Do you find the dashboard's user interface intuitive and easy to navigate?				X	
Perceived Usefulness					
How satisfied are you with the overall usefulness of the dashboard?				X	
To what extent do you believe that the dashboard is useful for future customers of Bricklog?					x
General Feedback					
Are there any additional features or functionality that users would find useful?	A check if the locations from the mutations match the locations from the Stock History				
Are there any errors or bugs in the dashboard that need to be addressed?	Not relevant, this is POC for new productline.				